U.S. DEPARTMENT OF JUSTICE
OFFICE OF JUSTICE PROGRAMS
NATIONAL INSTITUTE OF JUSTICE

PROJECT REVIEW FORM

GRANT/CONTRACT #: 90-IJ-CX-0038

GRANT/CONTRACT TITLE:
Analysis of DNA Typing Data for Forensic Applications

GRANTEE/CONTRACTOR NAME AND ADDRESS:
University of Texas Health Science Center at Houston
P.O. Box 20036
Houston, Texas 77225

PROJECT DIRECTOR AND ADDRESS:
Stephen P. Daiger, Ph.D.
(Same as Grantee)

FUNDING LEVEL THIS PHASE: $290,962
TOTAL LEVEL: Same

PROJECT PERIOD THIS PHASE: 6/1/90-5/31/92
TOTAL PERIOD: 6/1/90-12/31/92

PRODUCTS	TITLE AND AUTHOR	DATE SUBMITTED
1.	"Analysis of DNA Typing Data for Forensic Applications," by Stephen P. Daiger, Eric Boerwinkle, Ranajit Chakraborty	3/15/95
2.		
3.		
4.		

PROJECT MONITOR		FROM	TO
1.	Richard M. Rau	6/1/90	3/19/92
2.	Richard S. Laymon	3/19/92	5/27/94
3.	Richard M. Rau	5/31/94	Present

OUTSIDE REVIEWERS	NAME AND TITLE
1.	
2	
3	

STAFF REVIEWERS			
1	Dr. Richard M. Rau	3	
2		4.	

PROJECT REVIEWER: Richard M. Rau — DATE: 4/12/95

DIVISION DIRECTOR: — DATE:

OFFICE DIRECTOR: David G. Boyd — DATE: 4/12/95

(CONTINUED ON REVERSE)

I. FINDINGS AND SUBSTANTIVE QUALITY

Grant Manager's Assessment Report

Provide a narrative assessment *not to exceed 200 words* describing the following: problem addressed and major objectives, accomplishments, activities undertaken, principal findings and documents produced. This report will be entered into the Grant Profile System (PROFILE). For further clarification of the requirements, see chapter 7 of the effective edition of OJP HB 4500.2.

The goals of this study were to collect and analyze data sets of DNA typing results in order to address questions regarding the power and reliability of methods for calculating the probability of DNA matches. The study found that different types of DNA markers, such as VNTRs, STRs, and microsatellites, show different modes of evolution but the differences are too slight to affect forensic calculations. The study developed novel measures of genetic distance to evaluate the differences between populations. That is, these measures demonstrated that any biases in measuring population frequencies are either too slight to affect forensic calculations or are always conservative, and benefit a defendant. The study established the minimal size of a reliable data sample (approximately 100 to 150 individuals) and evaluated the occurrence, and consequences, of null or overlapping VNTR alleles. Thus, most current data sets used in forensic cases are of adequate size. Further, the study developed a technically-correct method for determining confidence intervals in calculating the probability of DNA matches (in order to deal with limited sample size); a method that is more reliable than the ceiling principal. Finally, this study demonstrated that the presence of population substructure does not pose any problem in providing conservative forensic calculations.

160143

THE UNIVERSITY OF TEXAS
HOUSTON
HEALTH SCIENCE CENTER

Human Genetics Center
School of Public Health

Summary Report: Grant 90-IJ-CX-0038

"Analysis of DNA Typing Data for Forensic Applications"

National Institute of Justice
June 1990 - December 1992

Investigators: Stephen P. Daiger, PhD, Professor
Eric Boerwinkle, PhD, Professor
Ranajit Chakraborty, PhD, Professor

I. Goals and Objectives

The overall goals of this research project were to collect and analyze data sets of DNA typing results in order to address questions regarding the power and reliability of methods for calculating the probability of DNA matches. The types of markers evaluated were VNTRs (variable number of tandem repeats) detected by Southern gel analysis and microsatellites polymorphisms detected by PCR methods. The analytic issues included the extent of variation between ethnic groups, deviations from expected equilibrium conditions, the consequences of sampling error and the consequences of population substructure. In addition, we developed a model VNTR marker for laboratory testing of these issues.

The specific aims of the project were to answer five related questions, i) how should forensic DNA data be recorded and disseminated, ii) how should data sets from different laboratories be analyzed and compared, iii) how different are the data from different ethnic groups and what are the forensic implications of population substructure, iv) how and why do allele frequencies differ from expected equilibrium values (and what are the consequences), and v) using a model VNTR system, what is the origin and evolution of VNTRs and what does this imply about their application to forensic testing?

During the course of the project we collected VNTR data from 32 ethnic groups (including American Blacks, American Caucasians, Hispanics and American Indians). Each of these data sets included from 4 to 6 VNTR markers tested in at least 100 individuals. The data were collected from participating members of TWGDAM (the Technical Working Group on DNA Analysis and Methods), established by the FBI. In a related effort we also collected and distributed a bibliography of scientific articles on forensic applications of DNA fingerprinting, with over 900 references [43]. A summary of the data sets and a copy of the bibliography are attached.

As explained in the following sections, our analyses of actual and simulated forensic data answered many of the theoretical criticisms of DNA match calculations. First, we demonstrated that different types of DNA markers, such as VNTRs, short tandem repeats and microsatellites, show different modes of evolution but that the differences are too slight to affect forensic calculations [6,22,23,28,63,64] . Second, we developed novel measures of genetic distance to evaluate the differences between populations [27,37]. These measures demonstrated that any biases in measuring population frequencies are either too slight to affect forensic calculations or are always conservative, that is, to the benefit of a defendant. Third, we established the minimal size of a reliable data sample (approximately 100 to 150 individuals) and evaluated the occurrence, and consequences, of null or overlapping VNTR alleles [21]. These results show that most current data sets used in forensic cases are of adequate size. Further, to deal with limited sample size, we developed a technically-correct method for determining confidence intervals [65]. This new method is more reliable than the ceiling principal. Also, the existence of null or overlapping alleles shows that simplistic measures of equilibrium are not applicable to VNTRs and that, in fact, most alleles and loci, within and across populations, are in equilibrium. (Hence methods using Hardy Weinberg equilibrium and the product rule are appropriate and reliable [26,30].) Finally, we developed several measures of population substructure and applied them to actual forensic data sets [25,30,39]. These analyses demonstrate that the presence of population substructure does not pose any problem in providing conservative forensic calculations.

In addition to the major analytic components of this project, we also conducted laboratory studies of a model VNTR marker [60,63]. This marker, a VNTR near the gene for apolipoprotein B on human chromosome 2, is a powerful, technically-reliable DNA marker for DNA typing. By better understanding the origin and evolution of this locus, we have contributed to its utility in DNA typing.

II. VNTR Allele Frequency Distributions Under the Stepwise Mutation Model: A Computer Simulation Approach for Forensic Analysis

Variable number of tandem repeats (VNTRs) are a class of highly informative and widely dispersed genetic markers. Despite their wide application in biological science, little is known about their mutational mechanisms or population dynamics. The objective of this work was to investigate four summary measures of VNTR allele frequency distributions: number of alleles, number of modes, range in allele size, and heterozygosity, using computer simulations of the one-step stepwise mutation model (SMM). These summary measures are directly related to forensic calculations using VNTR markers

We estimated these measures and their probability distributions for a wide range of mutation rates and compared the simulation results with predictions from analytical expectations under the SMM [14]. The average number of alleles, however, was larger in the simulations than the analytical expectation of the SMM. We then compared our simulation expectations with actual data reported in the forensic literature. We used the sample size and observed heterozygosity to determine i) the expected value, ii) the 5th and 95th percentiles for the other three summary measures, iii) allelic size range, iv) number of modes, and v) number of alleles.

The loci analyzed were classified into three groups based on the size of the repeat unit: microsatellites (1-2 base pair [bp] repeat units), short tandem repeats (STR) (3-5 bp repeat units), and minisatellites (15-70 bp repeat units). In general, STR loci were most similar to the simulation results under the SMM for the three summary measures (number of alleles, number of modes and range in allele size), followed by the microsatellite loci and then by the minisatellite loci, which showed deviations in the direction of the infinite allele model (IAM). Based on these differences, we hypothesize that these three classes of loci are subject to different mutational forces. Hence their spread through human populations and their properties related to forensic applications (e.g., number of alleles, frequencies and heterozygosity) will differ significantly.

III. A Novel Measure of Genetic Distance for Highly Polymorphic Tandem Repeat Loci

Genetic distance is a measure of the relatedness between two populations or species, and is useful for reconstructing the historic relationships among groups [37]. Moreover, it is directly related to issues regarding population substructuring, a major concern in forensic applications of DNA markers [30].

The usefulness of a genetic distance measure in phylogenetic reconstruction is dependent on the linearity of the measure with respect to time since divergence [37]. The standard measures of genetic distance in use today were developed for analysis of protein isoelectric focusing and antigenic polymorphisms, systems that generally show limited variability. Hypervariable tandem repeat DNA loci, genetic markers which have a larger number of alleles and a higher level of heterozygosity than traditional genetic markers, are being used increasingly to address evolutionary questions.

Traditional measures of genetic distance are non-linear at the higher levels of heterozygosity observed for these hypervariable loci. In addition, these traditional measures do not take into account what is known of the mutational mechanisms of hypervariable loci. There is good empirical and observational evidence that the generation of variability at loci of two classes of hypervariable tandem repeat loci, microsatellites and short tandem repeats, is *via* a stepwise mutation mechanism [50,51,64]. Using computer simulation, we showed that the new measure, D_{sw1}, conforms to linearity much better than do existing measures of genetic distance. We also showed that the variance of the new measure is similar in magnitude to existing genetic distance measures. In addition, an example of the application D_{sw} to an evolutionary analysis of human populations was published.

IV. Characteristics of Genotype Frequency Distributions at VNTR Loci

As mentioned in earlier sections, the molecular mechanisms through which new alleles are generated at the variable number of tandem repeat (VNTR) loci are not known precisely. In spite of this, it is well known that the levels of polymorphism at such loci are extremely large even in small inbred populations, and that new alleles are produced by mechanisms different from point mutations which lead to simple nucleotide substitutions. The fact of large numbers of alleles, and consequently large numbers of possible genotypes, at each locus necessarily

requires formal statistical procedures to study the characteristics of genotype distributions within populations.

Our efforts to develop such statistical procedures were directed toward development of statistical algorithms to study properties of genotype frequencies at VNTR loci. Evaluation of genotype frequencies at such loci is important i) to place appropriate statistical weight on DNA matches in forensic applications, ii) to determine parentage of individuals, and iii) to determine the validity of any stated relationships between individuals. Our research shows that each genotype is usually so rare in a population, that its frequency cannot be predicted reliably by simply counting its occurrence in a sample [21]. This is so because rare genotypes are either not found in a database or, even when found, they occur too infrequently to provide a reliable estimate of frequency [23,28,36,46,50,51].

Therefore, first, we determined the sample size requirements of DNA typing databases that will allow one to estimate genotype frequencies with adequate accuracy [21]. We concluded that in preparing DNA typing databases, 100 to 150 individuals per population are adequate to estimate all common allele frequencies precisely [21]. With such sample sizes, a threshold allele frequency can be predicted to give conservative estimates of all rare allele frequencies. In other words, the process of rebinning, a protocol practiced by the forensic community, has a scientific basis which is rooted in the population genetic characteristics of the VNTR loci. Sample size requirements are even less stringent for STR loci, or for polymarker loci, since the allelic variation is less extensive than for minisatellite VNTR loci.

Second, we showed that the classic, traditional approach to predicting genotype frequencies from allele frequencies (using the Hardy-Weinberg principle) is applicable to VNTR loci also [26,28]. As a step in reaching this conclusion, we developed computer algorithms to test the applicability of the product rule to VNTR databases. Applications of such computer algorithms to current DNA typing databases indicate that VNTR alleles within loci indeed combine at random, and there is no allelic association between loci that are physically distant from each other (i.e., when they are on different chromosomes, or far apart on the same chromosome, such as more distant than 7 cM from each other). In other words, genotype frequencies at a VNTR locus can be reliably predicted by multiplying the frequencies of its constituent alleles. Furthermore, the frequency of a multi-locus genotype (a combined DNA profile based on multiple probes) can be predicted by simply multiplying individual-locus genotype frequencies [39].

Third, we showed that in the RFLP protocol for DNA typing, there are inherent limitations, in that alleles of nearly equal size may not be resolved from each other (technically known as coalescence of alleles), and that unusually small (or large) alleles may not be distinguishable on Southern gels (technically called null or non-detectable alleles). Using experimental data, we showed that these situations *do* occur for several loci. As a consequence, when such situations occur in databases, not all single-banded DNA alleles are truly "homozygous". Interpretation of genotype frequencies must accommodate these inherent limitations [26,30].

We showed that both nondetectibility as well as coalescence would generate an apparent excess homozygosity, and this, by itself, cannot be regarded as evidence of allelic non-independence within and across loci [30]. We also developed statistical methods to account for such events in testing Hardy-Weinberg and linkage equilibrium, and applied these methods to

all current DNA typing databases. As a consequence of these investigations, we demonstrated that when these limitations are ignored, and allele frequencies are estimated by simple gene counts (treating all single-banded profiles as true homozygotes), the resultant allele frequency estimates are conservative (in the sense that they tend to be larger than their true values in the population) [39].

V. Population Substructure and Its Effect of VNTR Genotype Frequencies

It is well known that all natural populations are inherently substructured, since individuals do not mate at random [25]. Our efforts, therefore, focused on evaluating the effect of population substructuring on genotype frequency estimation at VNTR loci.

First, we showed that the classic measures of F_{ST} and G_{ST} (co-efficient of gene diversity) are applicable to VNTR loci, and that they can be estimated from population data [25].

Second, we showed that allele frequency differences between populations at VNTR loci are dominated by the large differences between alleles that are rare in all populations. Common alleles are shared by all populations; only the rare ones exhibit statistically significant differences in frequencies across populations [27]. Moreover, frequency differences across populations are such that inter-ethnic differences within major racial groups are much smaller than inter-racial differences [35,36]. In other words, by showing frequency differences across broad racial groups, we can capture almost all (if not the total) effects of population substructuring [35]. Coupled with this, when we recognized that homozygosity cannot be unequivocally asserted for all DNA typing protocols (because of coalescence and nondetectability), we recommended that a strict Hardy-Weinberg rule should not be applied to predict the frequency of a single-band DNA profile. Instead, a frequency of $2p$, should be used, where for p we can substitute the gene count estimate of the allele that represents the observed band in the profile [39]. We showed that this is sufficient to account for the effect of population substructure on genotype frequencies. The genotype frequencies resulting from this procedure, and their multiplication across loci, will then produce DNA profile frequencies which are expected to be conservative, even if population substructure exists in the database [39].

Predictions of genotype frequencies are made from samples and, hence, such predictions always have aa associated sampling error. Therefore, quantification of such sampling errors, or evaluation of confidence intervals for single- or multi-locus genotype frequencies, is an important task. This issue has been discussed in various contexts within the DNA forensic literature. The recommended procedure to account for sampling error, as discussed in the NRC report of 1992 (the ceiling principle), is erroneous. We developed the appropriate method by giving a confidence interval for single- and multi-locus genotype frequencies based on the theory of sampling for categorical data [65]. We recommend the use of this procedure.

Application of the method for determining confidence intervals shows that for databases where the sample size per locus is approximately 200 individual, the lower and upper bounds of the confidence intervals can differ by as much as 10-fold, so that an order of magnitude difference between 4- or 5-locus DNA profile frequencies may not be statistically significant [65].

VI. The Apolipoprotein B VNTR: A Paradigm for Forensic Applications of VNTR Markers

VNTRs are powerful markers for forsensic applications of DNA typing, but a better understanding of their genetic basis is essential for full acceptance and accurate population analysis. One useful model system to study the origin and evolution of VNTRs is the polymorphic repeat locus found beyond the 3' end of the human apolipoprotein (apo) B gene. A central aim of this project was to use the apo B 3' VNTR marker as a model for such studies.

The apo B gene is located at 2p24 and encodes two large hydrophobic proteins, apo B-100 and apo B-48. The apo B-48 mRNA is derived from the apo B-100 mRNA by means of a unique mRNA editing mechanism. Apo B-100 and B-48 are major protein components of low density lipoproteins and very low density lipoproteins and chylomicrons. Besides serving a structural role in these lipoproteins, apo B-100 is a ligand for the LDL receptor and is required for the receptor-mediated removal of LDL from plasma. The apo B gene is composed of 29 exons and 28 introns and spans 43 kb.

Approximately 500 bp downstream of the first polyadenylation signal in the apo B gene is the apo B 3' VNTR, which has an AT-rich dimeric repeat unit of 30 bp. This locus was first typed by Southern blotting methods. Using this technique, only five alleles were discernable. In 1989, two groups simultaneously developed protocols for the typing of the apo B 3' VNTR using PCR and high resolution agarose or polyacrylamide gel electrophoresis. Using PCR typing methods it is possible to detect differences in allele sizes of only one repeat unit (15 base pairs each) and it is not uncommon to find more than 10 allele in a population. Thus this marker is an excellent candidate for DNA typing for forensic purposes.

The apo B locus has been typed extensively in human population [28,46]. In addition, the apo VNTR has been identified as the 3' matrix-attachment region of the human apo B chromosomal region. Because of the importance of the apo B gene product in determining variation in lipid and lipoprotein levels, much effort has gone into characterizing antigenic and DNA variation at this locus [63]. Many genetic polymorphisms in the apo B gene have been reported and typed in samples from several different ethnic groups. Therefore the apo B gene has been well studied, and the apo B VNTR can be readily and accurately typed and can be amplified across a wide range of primate species.

VII. Origins and Early Evolution of the Apolipoprotein B 3' VNTR

One way to define the frequency and mechanisms of change that ultimately define the VNTR in and among human populations is to identify and isolate new mutations in human pedigrees by virtue of the existence of an allele in a child that does not appear in his or her mother or father. An evaluation of these new mutations is also relevant to the use of the marker in paternity testing. However, this process will yield an incomplete picture because of the small number of mutations that can be observed directly. In contrast, an evolutionary approach to this problem enables one to examine the full spectrum of mutational events.

An evolutionary approach was taken to distinguish between two hypotheses for the origin of VNTR loci, array transposition or initial duplication [63]. In the transpositional model, an array of repeats moves from one location in the genome to another, forming a new VNTR locus. In the initial duplication model, the array of repeats forms *in situ* following an early duplication event. Growth of the apo B 3' VNTR is evident in the primate lineage ranging from simple sequences in New World monkeys to highly polymorphic repeat arrays in chimpanzees and humans.

Molecular analysis of flanking direct repeats in primates identified a duplication between New World and Old World monkeys, and the VNTR repeats resulted from further amplification of two juxtaposed direct repeats. We conclude that the apo B VNTR did not arise by insertion of multiple tandem repeats, but originated *in situ* as the result of an initial duplication event.

VII. References

1. **E. BOERWINKLE**, C Leffert, H Hobbs (1991). Analysis of the apolipoprotein(a) gene structure in two populations with different distributions of plasma lipoprotein(a). Proc. 8th Int. Cong. of Human Genetics, Washington, DC. A#67.
2. **E BOERWINKLE** (1992). Genetics of plasma lipoprotein(a) concentrations. Current Opinion in Lipidology 3:128-136.
3. **E BOERWINKLE**, S-H Chen, S Visvikis, CL Hanis, G Siest, L Chan (1991). Signal peptide-length variation in human apolipoprotein B gene: Molecular characteristics and association with plasma glucose levels. Diabetes 40:1539-1544.
4. **E BOERWINKLE**, CC Leffert, J Lin, C Lackner, G Chiesa, HH Hobbs (1992). Apolipoprotein(a) gene accounts for greater than 90% of the variation in plasma lipoprotein(a) concentrations. J. Clin. Invest. 90:52-60.
5. B Budowle, AM Giusti, **R CHAKRABORTY** (1990). Discretized allelic data for VNTR locus by amplified fragment length polymorphism (AMP-FLP) analysis. Amer. J. Hum. Genet. 47:A#0502.
6. B Budowle, **R CHAKRABORTY**, AM Giusti, AJ Eisenberg, RC Allen (1991). Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution PAGE. Am. J. Hum. Genet. 48:137-144.
7. RM Cerda-Flores, GK Kshatriya, SA Barton, CH Leal-Garza, R Garza-Chapa, WJ Schull, **R CHAKRABORTY** (1991). Genetic structure of the populations migrating from San Luis Potosi and Zacatecas to Nuevo León in Mexico. Hum. Biol. 63:309-327.
8. RM Cerda-Flores, GK Kshatriya, TK Bertin, D Hewett-Emmett, CL Hanis, **R CHAKRABORTY** (1992). Gene diversity and estimation of genetic admixture among Mexican-Americans of Starr County, Texas. Ann. of Hum. Biol. 19:347-360.
9. **R CHAKRABORTY** (1990). Genetic profile of cosmopolitan populations: Effects of hidden subdivision. Anthrop. Anz. 48:313-331.
10. **R CHAKRABORTY** (1990). Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. Am.J. Hum. Genet. 47:87-94.
11. **R CHAKRABORTY** (1991). Book review: DNA Technology and Forensic Science, Banbury Report 32, (J Ballantyne and J Witkowski, eds.) Cold Spring Harbor Press, New York. Amer. J. Hum. Genet. 48:173-174.

12. **R CHAKRABORTY** (1991). Book review: Genetic Data Analysis. BS Weir, Sinauer Assoc. Mol. Biol. Evol. 8:396-397.
13. **R CHAKRABORTY** (1992). "Commentaries" on DNA typing and its court use. Professional Ethics Report V2:3-4.
14. **R CHAKRABORTY** (1992). Generalized occupancy problem and its applications in population genetics. "Impact of Genetics Variation on Individuals, Families and Populations" (CF Sing, CL Hanis, eds) Oxford University Press, New York pp. 179-192.
15. **R CHAKRABORTY** (1991). Impact of molecular genetics in studying origin of human populations. Archivos de Biologia y Medicina Experimentales 24:R98
16. **R CHAKRABORTY** (1991). Letters to the Editor: Inclusion of data on relatives for estimation of allele frequencies. Amer. J. Hum. Genet. 49:242-243.
17. **R CHAKRABORTY** (1991). Statistical interpertation of DNA typing data. Amer. J. Hum. Genet. 49:895-897.
18. **R CHAKRABORTY** (1991). Population genetics of hypervariable loci. Proc. 8th Int. Cong. of Human Genetics, Washington, DC 49:A252.
19. **R CHAKRABORTY** (1992). Book review: Convergent Issues in Genetics and Demography. (JA Adams, A Hermalin, D Lam, PE Smouse, eds.) Oxford Univ. Press, New York. Amer. J. Hum. Biol. 4:421-428.
20. **R CHAKRABORTY** (1992). Letters to the Editor: Multiple alleles and estimation of genetic parameters: Computational equations showing involvement of all alleles. Genetics 130:231-234.
21. **R CHAKRABORTY** (1992). Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Hum. Biol. 64:141-159.
22. **R CHAKRABORTY, E BOERWINKLE** (1990). Population genetics of VNTR polymorphism in humans. Amer. J. Hum. Genet. 47:A0504.
23. **R CHAKRABORTY, SP DAIGER** (1991). Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah. Hum. Biol. 63:571-587.
24. **R CHAKRABORTY, SP DAIGER, E BOERWINKLE** (1991). Patterns of genetic variation within and between populations detected by PCR-based VNTR polymorphisms. Proc. In. Seminar of the Forensic App. of PCR Technology, FBI Academy, Quantico, VA. Crime Lab Digest 18:148-152.
25. **R CHAKRABORTY**, H Danker-Hopfe (1991). Analysis of population structure: A comparative study of different estimators of Wright's fixation indices. Handbook of Statistics, 8:203-254.
26. **R CHAKRABORTY**, M de Andrade, **SP DAIGER**, B Budowle (1992). Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Ann. Hum. Genet 56:45-57.
27. **R CHAKRABORTY**, R Deka, L Jin, RE Ferrell (1992). Allele sharing at six VNTR loci and genetic distances among three ethnically defined human populations. Amer. J. Hum. Giol. 4:387-397.
28. **R CHAKRABORTY**, M Fornage, R Gueguen, **E BOERWINKLE** (1991). Population genetics of hypervariable loci: Analysis of PCR based VNTR polymorphism within a population. "DNA Fingerprinting: Approaches and Applications", (T Burke, G Dolf, AJ Jeffreys, R Wolff, eds.), Birkhauser-Verlag, Bern., pp. 127-143.
29. **R CHAKRABORTY**, L Jin (1992). Formal statistics of DNA fingerprinting data and relatedness between individuals. Amer. J. Hum. Genet. 51:A46.

30. **R CHAKRABORTY**, L Jin (1992). Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. Hum. Genet. 88:267-272.
31. **R CHAKRABORTY**, L Jin with 43 co-authors (1992). Multiple origins for phenylketonuria in Europe. Amer. J. Hum. Genet. 51:1355-1365.
32. **R CHAKRABORTY**, I Kamboh, RE Ferrell (1991). 'Unique' alleles in admixed populations: A strategy for determining 'hereditary' population differences of disease frequencies. Ethnicity & Disease 1:245-256.
33. **R CHAKRABORTY**, MI Kamboh, M Nwankwo, RE Ferrell (1992). Caucasian genes in American blacks: New data. Amer. J. Hum. Genet. 50:145-155.
34. **R CHAKRABORTY**, MI Kamboh, RE Ferrell (1992). Letter to the Editor: Response to Issues in estimating Caucasian admixture in American blacks. Reply to Reed. Amer. J. Hum. Genet. 51:680-681.
35. **R CHAKRABORTY**, KK Kidd (1991). The utility of DNA typing in forensic work. Science 254:1735-1739.
36. **R CHAKRABORTY**, KK Kidd (1992). Letter to the Editor: Forensic DNA typing: response. Science 255:1053.
37. **R CHAKRABORTY**, CR Rao (1991). Measurement of genetic variation for evolutionary studies. Handbook of Statistics 8:271-316.
38. **R CHAKRABORTY**, MR Srinivasan (1992). A modified "best maximum likelihood" estimator of line regression with errors in both variables: an application for estimating genetic admixture. Biometrical J. 5:567-576.
39. **R CHAKRABORTY**, MR Srinivasan, L Jin, M de Andrade (1992). Effects of population subdivision and allele frequency differences on interpretation of DNA typing data for human identification. Proc. 2nd Intl. Symp. of Hum. Id., Promega Corp., Madison WI pp 205-222.
40. **R CHAKRABORTY**, KM Weiss (1991). Genetic variation of the mitochondrial DNA genome in American Indians is at mutation-drift equilibrium. Amer. J. Phy. Anth. 86:497-506.
41. P Clemens, RG Fenwick, JS Chamberlain, RA Gibbs, M de Andrade, **R CHAKRABORTY**, CT Caskey (1991). A rapid and informative assay for linkage analysis and prenatal diagnosis in Duchenne muscular dystrophy families using CA polymorphisms in a deletion-prone region of dystrophin. Amer. J. Hum. Genet 49:A978.
42. PR Clemens, RG Fenwick, JS Chamerlain, RA Gibbs, M de Andrade, **R CHAKRABORTY**, CT Caskey (1991). Carrier detection and prenatal diagnosis in Duchenne and Becker Muscular Dystrophy families, using dinucleotide repeat polymorphisms. Amer. J. Hum. Genet. 49:951-960.
43. **SP DAIGER** (1991). Issues in DNA fingerprinting for forensic purposes. State Bar of Texas Professional Development Program, J1-J41.
44. **SP DAIGER** (1991). Letter to the Editor. DNA Fingerprinting. Amer. J. Hum. Genet. 49:897.
45. M de Andrade, **R CHAKRABORTY**, RP Clemens, CT Caskey (1991). Linkage disequilibria among CA polymorphisms in the human dysrophin gene. Amer. J. Hum. Genet. 49:A983.
46. R Deka, **R CHAKRABORTY**, S DeCroo, F Rothhammer, SA Barton, RE Ferrell (1992). Characteristics of polymorphism at a VNTR locus 3' to the apolipoprotein b gene in five human populations. Amer. J. Hum. Genet. 51:1325-1333.

47. R Deka, **R CHAKRABORTY**, RE Ferrell (1990). Population genetics of human hypervariable loci. Amer. J. Hum. Genet. 47:A0512.
48. R Deka, **R CHAKRABORTY**, RE Ferrell (1991). A population genetic study of six VNTR loci in three ethnically defined populations. Genomics 11:83-92.
49. R Deka, **R CHAKRABORTY**, RE Ferrell (1991). Allele sharing and genetic distance at VNTR loci among three ethnic groups. Proc. 8th Int. Cong. of Human Genetics, Washington DC 49:A2837.
50. A Edwards, HA Hammond, **R CHAKRABORTY**, CT Caskey (1991). DNA typing with trimeric and tetrameric tandem repeats: polymorphic loci, detection systems, and population genetics. Proceedings of the 2nd Int'l. Symposium on Hum. Id. Promega Corp. Madison, WI pp. 31-52.
51. A Edwards, HA Hammond L Jin, CT Caskey, **R CHAKRABORTY** (1992). Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12:241-253.
52. J Ely, **R CHAKRABORTY**, R Deka, RE Ferrell (1991). Comparison of VNTR polymorphisms among human and chimpanzee. Amer. J. Hum. Genet. 49:A2615.
53. J Ely, R Deka, **R CHAKRABORTY**, RE Ferrell (1992). Comparison of five tandem repeat loci between humans and chimpanzees. Genomics 14:692-698.
54. M Fornage, G Siest, **E BOERWINKLE** (1991). Frequency distribution of a (TG)*n*(AG)*m* microsatellite reflects the mechanisms of production of new alleles. Proc. 8th Int'l. Cong. of Hum. Genetics, Washington, DC 49:A2800.
55. M Fornage, L Chan, G Siest, **E BOERWINKLE** (1992). Allele frequency distribution of the (TG)*n*(AG)*m* microsatellite in the apolipoprotein C-II gene. Genomics 12:63-68.
56. HA Hammond, A Edwards, L Jin, **R CHAKRABORTY**, CT Caskey (1991). Studies of multilocus genotype data validate the use of DNA typing with polymorphic trimeric and tetrameric tandem repeats for personal identification. Amer. J. Hum. Genet. 49:A2506.
57. L Jin, **R CHAKRABORTY** (1992). Population dynamics of DNA fingerprinting patterns within and between populations. Amer. J. Hum. Genet. 51:A603.
58. L Jin, **R CHAKRABORTY**, HA Hammond, CT Caskey (1991). Polymorphisms at short tandem repeat (STR) loci within and between four ethnic populations of Texas. Proc. 8th Int'l. Cong. of Hum. Genet. Washington, DC 49:A65.
59. MI Kamboh, **R CHAKRABORTY**, RE Ferrell (1990). Caucasian genes in the American blacks: New data. Amer. J. Hum. Genet. 47:A0540.
60. C Lackner, **E BOERWINKLE**, CC Leffert, T Rahmig, HH Hobbs (1991). Molecular basis of apolipoprotein(a) isoform size heterogeneity as revealed by pulsed-field gel electrophoresis. J. Clin. Invest. 87:2153-2161.
61. CR Rao, **R CHAKRABORTY** (1991). Handbook of Statistics, Vol. 8: Statistical Methods in Biological and Medical Sciences. Elsevier Science Publishing Company, Inc., New York.
62. M Shriver, **SP DAIGER, R CHAKRABORTY, E BOERWINKLE** (1991). Multimodal distribution of length variation in VNTR loci detected using PCR. Proc. Int'l. Seminar on the Forensic App. of PCR Technology. Crime Laboratory Digest 18:144-147.
63. MD Shriver, JE Hixson, **SP DAIGER, E BOERWINKLE** (1991). Birth of a VNTR: Size and sequence comparison of the apo B 3' VNTR in humans and non-humans primates. Proc. 8th Int'l. Cong. of Hum. Genet. Washington, DC 49:A2631.

64. MD Shriver, L Jin, **R CHAKRABORTY, E BOERWINKLE** (1992). Computer simulations of the stepwise mutation model and VNTR allele frequency distributions. Amer. J. Hum. Genet. 51:A623.

65. MR Srinivasan, **SP DAIGER, R CHAKRABORTY** (1992). Interval estimation of multilocus genotype frequencies and its forensic implications. Amer. J. Hum. Genet. 51:A628.

DNA Fingerprint Bibliography

February 24, 1995
Dr. Stephen P. Daiger, Dr. Ranajit Chakraborty, Dr. Eric Boerwinkle
Supported by NIJ Grant 90-IJ-CX-0038

1.
Auth: Acton, Ronald T.//Harmon, Leigh//Go, Rodney C. P.//Budowle, Bruce
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA.
Ttl1: Comparison of VNTR allele frequencies in white and black populations
Type: book chapter
Area: technical or scientific
Ttl2: Proceedings for the International Symposium on Human Identification
Plac: Madison, WI
Publ: Promega Corp.
Date: 1990
Page: 5-20

2.
Auth: Acton, Ronald T.//Hodge, T. W.//Go, Rodney C.//Peiper, S. C.//Harman, Leigh
Affl: University of Alabama at Birmingham, Birmingham, Alabama
Ttl1: A stratified approach to forensic identification by DNA analysis: a comparison of traditional genetic systems and restriction fragment length polymorphisms (RFLPs)
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 273-275

3.
Auth: Adams, D. E.//Presley, L. A.//Deadman, H. A.//Lynch, A. G.
Affl: FBI Laboratory, Quantico, VA.
Ttl1: DNA analysis in the FBI laboratory
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 173-177

4.
Auth: Adams, Dwight E.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Validation of the FBI procedure for DNA analysis: a summary
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1988
Volm: 15(4)
Page: 106-108

5.
Auth: Adams, Dwight E.//Presley, Lawrence A.//Baumstark, Anne L.//Hensley, Kathy W.//Hill, Alice L.//Anoe, Kim S.//et al
Affl: FBI Laboratory, Quantico, VA.
Ttl1: Deoxyribonucleic acid (DNA) analysis by restriction fragment length polymorphisms of blood and other body fluid stains subjected to contamination and environmental insults
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: September 1991
Volm: 36(5)
Page: 1284-1298

6.
Auth: Adema, Alison Priske
Ttl1: DNA fingerprinting evidence: the road to admissibility in California
Type: journal article
Area: legal
Ttl2: San Diego Law Review
Plac: California
Date: March-April 1989
Volm: 26(2)
Page: 377-415

7.
Auth: Adkison, Linda R.
Ttl1: DNA fingerprinting: a scientific perspective.
Type: journal article
Area: legal
Ttl2: Mercer Law Review
Plac: United States
Date: Spring 1991
Volm: 42(3)
Page: 1099-1111

8.
Auth: Akane, A.//Matsubara, K.//Shiono, H.//Yamada, M.//Nakagome, Y.
Ttl1: Diagnosis of twin zygosity by hypervariable RFLP markers
Type: journal article
Area: technical or scientific
Ttl2: American Journal of Medical Genetics
Date: 1991
Volm: 41
Page: 96-98

9.
Auth: Akane, A[illegible]ushi//Matsubara, Kazuo//Shiona, Hiroshi//Yuasa, Isao//Yokota, Shin Ichi//Yamada, Masao//Nakagome Yasuo
Affl: Shimane Medical Univeristy, Izumo; Tottori Univ., School of Medicine, Yonago; National Children's Medical Research Center, Taishido, Setagaya, Japan
Ttl1: Paternity Testing: blood group systems and DNA analysis by variable number of tandem repeat markers.
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: September 1990
Volm: 35(5)
Page: 1217-1225

10.
Auth: Akasaka, H.//Nonomura, S.
Affl: Metropolitan Police Department, Tokyo, Japan
Ttl1: Application of DNA fingerprint analysis with minisatellite DNA probes to individual identification
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 211

11.
Auth: Aldhous, Peter
Affl: Writer - Nature Magazine
Ttl1: Challenge to British forensic database
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: January 16, 1992
Volm: 355
Page: 191

12.
Auth: Aldhous, Peter
Affl: Writer - Nature Magazine
Ttl1: Geneticists attack NRC report as scientifically flawed
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: February 5, 1993
Volm: 259
Page: 755-756

13.
Auth: Aldrous, Peter
Affl: Writer - Nature Magazine
Ttl1: Congress reviews DNA testing
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: June 27, 1991
Volm: 351
Page: 684

14.
Auth: Ali, Sher//Muller, C. R.//Epplen, Jorg T.
Affl: Dept. of Molecular Biochemistry, Beckman Research Institute, Duarte, CA.
Ttl1: DNA fingerprinting by oligonucleotide probes specific for simple repeats
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: November 1986
Volm: 74(3)
Page: 239-243

15.
Auth: Ali, Sher//Wallace, R. Bruce
Affl: Dept. of Molecular Biochemistry, Beckman Research Institute, Duarte, CA.
Ttl1: Intrinsic polymorphism of variable number tandem repeat loci in the human genome
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: September 12, 1988
Volm: 16(17)
Page: 8487-8496

16.
Auth: Allard, J.
Ttl1: Murder in South London - a novel use of DNA profiling
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science Society
Date: 1992
Volm: 32
Page: 49-58

17.
Auth: Allen, Robert C.//Graves, George M.//Budowle, Bruce
Affl: Medical Univ. of South Carolina, Charleston, SC; FBI Laboratory, VA
Ttl1: Effects of viscosity, pore size, ionic strength and power application on the relative mobility of DNA fragments separated in polyacrylamide gels with discontinuous buffers
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 249-252

18.
Auth: Allen, Robert C.//Graves, George//Budowle, Bruce
Affl: Dept. of Pathology and Laboratory Medicine, Medical University of South Carolina, Children's Hospital, Charleston, S.C. 29425
Ttl1: Polymerase chain reaction amplification products separated on rehydratable polyacrylamide gels and stained with silver
Type: journal article
Area: technical or scientific
Ttl2: Biotechniques
Date: July-August 1989
Volm: 7(7)
Page: 736-744

19.
Type: catalog
Area: technical or scientific
BkAu: American Type Culture Collection
Ttl2: ATCC/NIH Repository Catalog of Human and Mouse DNA Probes and Libraries, Fifth Edition, 1991
Plac: 12301 Parklawn Drive, Rockville, MD 20852-1776
Publ: American Type Culture Collection
Date: 1991
Volm: ISBN 0-930009-42-8
Page: i-viii,1-364

20.
Type: periodical
Area: technical or scientific
BkAu: Amos, Bill Pemberton ,. Josephine
Ttl2: Fingerprint News
Date: 1989-present
Srce: Dept. of Genetics, Downing Street, Cambridge, CB2 3EH, U.K.

21.
Auth: Anderson, Alun
Affl: Writer - Nature Magazine
Ttl1: DNA fingerprinting on trial
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: December 21-28, 1989
Volm: 342(6252)
Page: 844

22.
Auth: Anderson, Alun
Affl: Writer - Nature Magazine
Ttl1: Forensic tests proved innocent
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: August 9, 1990
Volm: 346
Page: 499

23.
Auth: Anderson, Alun
Affl: Writer - Nature Magazine
Ttl1: Judge backs technique
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: August 24, 1989
Volm: 340(6235)
Page: 582

24.
Auth: Anderson, Alun
Affl: Writer - Nature Magazine
Ttl1: New technique on trial
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: June 8, 1989
Volm: 339(6224)
Page: 408

25.
Auth: Anderson, Cerisse
Ttl1: DNA fingerprint factor in rape trial: Queens defendant accused in series of attacks
Type: journal article
Area: legal
Ttl2: New York Law Journal
Plac: New York (State)
Date: September 21, 1988
Volm: 200(57)
Page: 1 (col 3)

26.
Auth: Anderson, Cerisse
Ttl1: DNA test accepted in judging paternity: surrogate's ruling reportedly state's first in civil action centering on genetic fingerprints
Type: journal article
Area: legal
Ttl2: New York Law Journal
Plac: New York (State)
Date: August 11, 1988
Volm: 200(29)
Page: 1 (col 3)

27.
Auth: Anderson, Christopher
Affl: Writer - Nature Magazine
Ttl1: Academy approves, critics still cry foul
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: April 16, 1992
Volm: 356
Page: 552

28.
Auth: Anderson, Christopher
Affl: Writer - Nature Magazine
Ttl1: Conflict concerns disrupt panels, cloud testimony
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: February 27, 1992
Volm: 355(6363)
Page: 753-754

29.
Auth: Anderson, Christopher
Affl: Writer - Nature magazine
Ttl1: Courts reject DNA fingerprinting, citing controversy after NAS report
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: October 1, 1992
Volm: 359
Page: 349

30.
Auth: Anderson, Christopher
Affl: Writer - Nature Magazine
Ttl1: DNA fingerprinting discord
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: December 19-26, 1991
Volm: 354
Page: 500

31.
Auth: Anderson, Christopher
Affl: Writer Nature Magazine
Ttl1: FBI attached strings to its DNA database
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: June 25, 1992
Volm: 357
Page: 618

32.
Auth: Anderson, Christopher
Affl: Writer - Nature magazine
Ttl1: FBI gives in on genetics
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: February 20, 1992
Volm: 355
Page: 663

33.
Auth: Annas, G. J.
Affl: Law Medicine & Ethics Program, Boston University Schools of Medicine and Public Health
Ttl1: DNA fingerprinting in the twilight zone
Type: journal article
Area: technical or scientific
Ttl2: Hasting Century Report
Date: March-April 1990
Volm: 20(2)
Page: 35-37

34.
Auth: Annas, J. D. ,. M. P. H. ,. George J.
Affl: Law Medicine & Ethics Program, Boston University Schools of Medicine and Public Health
Ttl1: Setting standards for the use of DNA-typing results in the courtroom-the state of the art
Type: journal article
Area: technical or scientific
Ttl2: The New England Journal of Medicine
Date: June 11, 1992
Volm: 326(24)
Page: 1641-1644

35.
Auth: Arnemann, J.//Schmidtke, J.//Epplen, J. T.//Kuhn, H. J.//Kaumanns, W.
Affl: Institute of Human Genetics, Bottingen, West Germany
Ttl1: DNA fingerprinting for paternity and maternity in "group-O" Cayo Santiago derived rhesus monkeys at the German Primate Center, Puerto Rico: results of a pilot study
Type: journal article
Area: technical or scientific
Ttl2: Health Science Journal
Date: April 1989
Volm: 8(1)
Page: 181-184

36.
Auth: ASHG Ad Hoc Committee on Individual Identification by DNA analysis
Affl: Social Issues Committee
Ttl1: Individual identification by DNA analysis: points to consider
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1990
Volm: 46
Page: 631-634

37.
Auth: Austad, Steven N.
Affl: Dept. of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: February 28, 1992
Volm: 255
Page: 1051

38.
Auth: author not given
Ttl1: Admission of DNA fingerprints prompts queries
Type: journal article
Area: legal
Ttl2: National Law Journal
Case: State v. Andrews
Plac: Florida; states
Date: January 18, 1988
Volm: 10(19)
Page: 42 (col 1)

39.
Auth: author not given
Ttl1: "Amp-le" evidence: a case for PCR fragment analysis in DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Biosystems Reporter
Page: 8, 10

40.
Auth: author not given
Ttl1: Decision on the admissibility of DNA identification tests
Type: legal decision
Area: legal
Case: The People of the State of New York against Joseph Castro
Cort: Supreme Court of the State of New York - Gerald Sheindlin J.S.C.
Plac: Bronx, New York
Date: August 14, 1989
Volm: Indictment No. 1508/87
Srce: County of Bronx: Criminal Term Part 28

41.
Auth: author not given
Ttl1: DNA fingerprints now possible from small amounts of evidence
Type: journal article
Area: legal
Ttl2: Criminal Justice Newsletter
Plac: United States
Date: June 1, 1988
Volm: 19(11)
Page: 6(1)

42.
Auth: author not given
Ttl1: DNA fingerprinting upheld in first appellate-level challenge
Type: journal article
Area: legal
Ttl2: Criminal Justice Newsletter
Plac: United States
Date: December 1, 1988
Volm: 19(23)
Page: 3-4

43.
Auth: author not given
Ttl1: DNA probes control immigration
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: January 16, 1986
Volm: 319
Page: 171

44.
Auth: author not given
Type: book
Area: technical or scientific
Ttl2: DNA Technology in Forensic Science
Plac: Washington, D.C.
Publ: National Academy Press
Date: April 16, 1992
Page: i-xiv, S.1-8.14

45.
Auth: author not given
Ttl1: DNA testing on the increase
Type: journal article
Area: legal
Ttl2: Solicitors Journal
Plac: Great Britain
Date: November 27, 1987
Volm: 131(48)
Page: 1596

46.
Auth: author not given
Ttl1: Fingerprint trials
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: June 15, 1989
Volm: 339
Page: 491-492

47.
Auth: author not given
Ttl1: Genetic testing found reliable
Type: newspaper
Area: lay press
Ttl2: The Houston Post
Date: Wednesday, April 15, 1992
Page: A-1, A-13

48.
Auth: author not given
Type: database
Area: technical or scientific
BkAu: author not applicable
Ttl2: Genome Data Base (GDB)
Publ: Johns Hopkins University
Srce: GDB/OMIM, Welch Medical Library, 1830 E. Monument St., Baltimore, MD 21205; 301-955-7058

49.
Auth: author not given
Ttl1: Guidelines for a quality assurance program for DNA analysis (The)
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: April 1991
Volm: 18
Page: 44

50.
Auth: author not given
Ttl1: Independence of forensic science
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: March 14, 1991
Volm: 350
Page: 95

51.
Auth: author not given
Ttl1: Opinion and Order
Type: legal decision
Area: legal
Case: United States of America v. Randolph B. Jakobetz
Cort: United States District Court for the District of Vermont - Franklin S. Billings, Jr. Chief Judge
Date: September 20, 1990
Volm: Criminal No. 89-65
Page: 1-35
Srce: District Court

52.
Auth: author not given
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Procedures for the detection of restriction fragment length polymorphisms in human DNA
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: December 4, 1989

53.
Auth: author not given
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA.
Ttl1: Statement of the working group on statistical standards for DNA analysis
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: July 1990
Volm: 17(3)
Page: 53-58

54.
Auth: author not given
Ttl1: United States approval for DNA fingerprinting
Type: letter to editor
Area: technical or scientific
Ttl2: The Lancet
Date: May 9, 1992
Volm: 339(8802)
Page: 1165-1166

55.
Auth: Avise, J. C.//Bowen, B. W.//Lamb, T.
Affl: Department of Genetics, University of Georgia, Athens 30602
Ttl1: DNA fingerprinting from hypervariable mitochondrial genotypes
Type: journal article
Area: technical or scientific
Ttl2: Molecular Biology Evolution
Date: May 1989
Volm: 6(3)
Page: 258-269

56.
Auth: Azuma, C.//Kamiura, S.//Nobunaga, T.//Negoro, T.//Saji, F.//Tanizawa, O.
Affl: Dept. of Ob. and Gyn., Osaka University Medical School, Osaka, Japan
Ttl1: Zygosity determination of multiple pregnancy by deoxyribonucleic acid fingerprints
Type: journal article
Area: technical or scientific
Ttl2: American Journal Obstet. Gynecology
Date: 1989
Volm: 160
Page: 734-736

57.
Auth: Baechtel, F. Samuel
Affl: FBI Laboratory, Quantico, VA.
Ttl1: The extraction, purification and quantification of DNA
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 25-28

58.
Auth: Baechtel, F. Samuel
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA.
Ttl1: A primer on the methods used in the typing of DNA
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: 1988
Volm: 15 Supp. 1
Page: 3-9

59.
Auth: Baechtel, F. Samuel
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA.
Ttl1: Recovery of DNA from human biological specimens
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: 1988
Volm: 15
Page: 95-96

60.
Auth: Baechtel, F. S.//Monson, K. L.//Forsen, G. E.//Budowle, B.//Kearney, J. J.
Affl: Forensic Science Research and Training Section, FBI Laboratory, Quantico, VA.
Ttl1: Tracking the violent criminal offender through DNA typing profiles - a national database system concept
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 356-360

61.
Auth: Baird, M. L.
Affl: Lifecodes Corporation, Valhalla, N.Y.
Ttl1: Analysis of forensic DNA samples by single locus VNTR probes
Type: book chapter
Area: technical or scientific
BkAu: Farley, Mark A.//Harrrington, James J.
Ttl2: Forensic DNA Technology
Date: 1991
Page: 39-49

62.
Auth: Baird, M.//Balazs, I.//Giusti, A.//Miyazaki, L.//Nicholas, L.//Wexler, K.//Kanter, E.//Glassberg, J.//Allen, F.//Rubinstein, P.//et al
Affl: Lifecodes Corporation, Valhalla, N.Y. 10595
Ttl1: Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: October 1986
Volm: 39(4)
Page: 489-501

63.
Auth: Baird, M.//Giusti, A.//Meade, E.//Clyne, M.//Shaler, R.//Benn, P.//Glassberg, J.//Balazs, I.
Affl: Lifecodes Corporation, Valhalla, N.Y. 10595
Ttl1: The application of DNA-Print(tm) for the estimation of paternity
Type: book chapter
Area: technical or scientific
Ttl2: Advances in Forsensic Haemogenetics 2
Plac: New York, N.Y.
Publ: Springer-Verlag
Date: 1987
Volm: 2
Page: 354-358

64.
Auth: Baird, M.//Giusti, A.//Meade, E.//Clyne, M.//Shaler, R.//Benn, P.//Glassberg, J.//Balazs, I.
Affl: Lifecodes Corporation, Valhalla, N.Y. 10595
Ttl1: The application of DNA-Print(tm) for identification from forensic biological materials
Type: book chapter
Area: technical or scientific
Ttl2: Advances in Forensic Haemogenetics 2
Plac: New York, N.Y.
Publ: Springer-Verlag
Date: 1987
Volm: 2
Page: 396-402

65.
Auth: Baird, Michael
Affl: Lifecodes Corporation, Vallhalla, N. Y.
Ttl1: DNA testing - is forensic DNA testing reliable?
Type: journal article
Area: technical or scientific
Ttl2: ABA Journal
Date: September 1990
Page: 34

66.
Auth: Baird, Michael L.
Affl: Lifecodes Corporation, Valhalla, N.Y.
Ttl1: Quality control and quality assurance
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 175-190

67.
Auth: Baird, Michael L.//Balazs, Ivan//McElfresh, Kevin
Affl: Lifecodes Corporation, Valhalla, New York
Ttl1: Examination of forensic biological evidence by DNA print analysis
Type: book chapter
Area: technical or scientific
BkAu: Lee, Henry C.//Gaensslen, R. E.
Ttl2: DNA and Other Polymorphisms in Forensic Science
Date: 1990
Page: 61-75

68.
Auth: Balazs, Ivan//Baird, Michael//Clyne, Mindy//Meade, Ellie
Affl: Lifecodes Corporation, Valhalla, N.Y. 10595
Ttl1: Human population genetic studies of five hypervariable DNA loci
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: February 1989
Volm: 44(2)
Page: 182-190

69.
Auth: Balazs, Ivan//Neuweiler, John//Gunn, Peter//Kidd, Judith//Kidd, Kenneth K.//Kuhl, Joy//Mingjun, Liu
Affl: Lifecodes Corp.; Genetic Technologies Crows Nest; Yale Univ.;
Ttl1: Human population genetic studies using hypervariable loci. I. Analysis of Assamese, Australian, Cambodian, Caucasian, CHinese adn Melanesian populations
Type: journal article
Area: technical or scientific
Ttl2: Genetics
Date: May 1992
Volm: 131(1)
Page: 191-198

70.
Auth: Balding, David J.//Torney, David C.
Affl: University London, London, UK; Los Alamos National Lab., Los Alamos, NM
Ttl1: Statistical analysis of DNA fingerprint data for ordered clone physical mapping of human chromosomes
Type: journal article
Area: technical or scientific
Ttl2: Bulletin Mathematical Biology
Date: 1991
Volm: 53(6)
Page: 853-879

71.
Auth: Ballantyne, Jack
Affl: Suffolk County Crime Laboratory, Hauppauge, N.Y.
Ttl1: DNA technology in a county forensic science laboratory setting
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA technology and forensic Science
Date: 1989
Page: 213-216

72.
Type: book
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Plac: Box 100, Cold Spring Harbor, New York 11724
Publ: Cold Spring Harbor Laboratory Press
Date: 1989
Volm: ISNB 0-87969-232-4
Page: i-xvii, 1-352

73.
Auth: Banks, Peter
Ttl1: The trials of DNA fingerprinting: science takes the stand
Type: journal article
Area: technical or scientific
Ttl2: Journal of NIH Research
Date: November 1990
Volm: 2
Page: 75-77

74.
Auth: Bar, W.//Hummel, K.
Affl: University of Zurich, Zurichbergstrasse, Zurich, Switzerland; Institut fur Blutgruppenserologie, Postfach, Freiburg, Germany
Ttl1: DNA Fingerprinting: its application in forensic case work
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 349-355

75.
Auth: Bar, W.//Kratzer, A.
Affl: Institute of Forensic Medicine, Zurich, Switzerland
Ttl1: Assessment of disputed identity of blood alcohol samples using DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Z Rechtsmed
Date: 1989
Volm: 102(4)
Page: 263-270

76.
Auth: Bar, W[illegible]r//Kratzer, Adelgunde//Machler, Marco//Schmid, Werner
Affl: Institute of Forensic Medicine, Univ. of Zurich, Switzerland
Ttl1: Postmortem stability of DNA
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: October 1988
Volm: 39(1)
Page: 59-70

77.
Auth: Barinaga, Marcia
Ttl1: DNA fingerprinting database to finger criminals
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: January 21, 1988
Volm: 331(6153)
Page: 203

78.
Auth: Barinaga, Marcia
Ttl1: Pitfalls come to light
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: May 11, 1989
Volm: 339(6220)
Page: 89

79.
Auth: Barr, Jessie Jo
Ttl1: The use of DNA typing in criminal prosecutions: a flawless partnership of law and science?
Type: journal article
Area: legal
Ttl2: New York Law School Law Review
Date: 1989
Volm: 34
Page: 485

80.
Auth: Bashinski, J.
Affl: California Department of Justice, Oakland, California
Ttl1: Managing the implementation and use of DNA typing in the crime labatory
Type: book chapter
Area: technical or scientific
BkAu: Farley, Mark A.//Harrington, James J.
Ttl2: Forensic DNA Technology
Date: 1991
Page: 201-235

81.
Auth: Bashinski, Jan S.
Affl: Oakland Police Department, Oakland, CA.
Ttl1: Laboratory standards: accreditation, training and certification of staff in the forensic context
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 159-173

82.
Auth: Beardsley, Tim
Ttl1: Pointing Fingers - DNA identification is called into question
Type: journal article
Area: technical or scientific
Ttl2: Scientific American
Date: March 1992
Page: 26-27

83.
Auth: Beeler, Lourel//Wiebe, William R.
Ttl1: DNA identification tests and the courts
Type: journal article
Area: legal
Ttl2: Washington Law Review
Date: October 1988
Volm: 63
Page: 903

84.
Auth: Bell, Graeme I.//Selby, Mark J.//Rutter, William J.
Affl: Dept. of Biochemistry & Biophysics, Univ. of California, San Francisco,CA.
Ttl1: The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: January 7-13, 1982
Volm: 295(5844)
Page: 31-35

85.
Auth: Bellamy, R. J.//Inglehearn, C. F.//Jalili, I. K.//Jeffreys, A. J.//Bhattacharya, S. S.
Affl: University of Newcastle-upon-Tyne, UK; St. John's Ophthalmic Hospital, Jerusalem, Israel; Dept. of Genetics, Leicester Univ., Leicester, UK
Ttl1: Increased band sharing in DNA fingerprints of an inbred human population
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: 1991
Volm: 87
Page: 341-347

86.
Auth: Bellamy, Richard//Inglehearn, Chris//Lester, Doug//Hardcastle, Alison//Bhattacharya, Shomi
Affl: Dept. of Human Genetics, Newcastle Univ., University of Newcastle-upon-Tyne, UK
Ttl1: Better fingerprinting with PCR
Type: journal article
Area: technical or scientific
Ttl2: Trends in Genetics
Date: February 1990
Volm: 6(2)
Page: 32

87.
Auth: Bentzen, P.//Harris, A. S.//Wright, J. M.
Affl: Marine Gene Probe Laboratory and Dept. of Biology, Dalhousie University, Halifax, Nova Scotia, Canada
Ttl1: Cloning of hypervariable minisatellite and simple sequence microsatellite repeats for DNA fingerprinting of important aquacultural species of salmonids and tilapia
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 243-262

88.
Auth: Berry, Donald A.
Affl: School of Statistics, University of Minnesota, Minneapolis, MN. 55455
Ttl1: DNA Fingerprinting: What does it prove?
Type: journal article
Area: technical or scientific
Ttl2: Chance: New Directions for Statistics and Computing
Publ: Springer-Verlag New York, Inc.
Date: 1990
Page: 15-25

89.
Auth: Berry, Donald A.
Affl: School of Statistics, University of Minnesota, Minneapolis, MN 55455
Ttl1: Inferences using DNA profiling in forensic indentification and paternity cases
Type: journal article
Area: technical or scientific
Ttl2: Stat. Science
Date: July 1991

90.
Auth: Berry, Donald A.
Affl: School of Statistics, University of Minnesota, Minneapolis, MN 55455
Ttl1: Statistical inference in crime investigations using DNA profiling: single locus probes
Type: journal article
Area: technical or scientific
Ttl2: Journal Royal Statistical Society Series C
Date: 1992
Volm: 41
Page: 499-531

91.
Auth: Bever, Robert A.//DeGuglielmo, Michael//Staud, Rick W.//Kelly, Charles M.//Foster, R. Scott
Affl: Genetic Design, Inc., Greensboro, NC 27409
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: February 28, 1992
Volm: 255
Page: 1050-1051

92.
Auth: Bigbee, David//Tanton, Richard L.//Ferrara, Paul B.
Affl: FBI Laboratory; Palm Beach Sheriff's Dept.; Bureau of Forensic Science
Ttl1: Implementation of DNA analysis in American crime laboratories
Type: journal article
Area: technical or scientific
Ttl2: The Police Chief
Date: October 1989
Page: 86-89

93.
Auth: Black, W. J.
Affl: Department of Microbiology, Stanford University Medical Center, CA. 94305
Ttl1: Drug products of recombinant DNA technology
Type: journal article
Area: technical or scientific
Ttl2: American Journal Hospital Pharmacy
Date: September 1989
Volm: 46(9)
Page: 1834-1844

94.
Auth: Blair, C. Thomas
Ttl1: Recent developments in the admissibility of DNA fingerprint evidence
Type: journal article
Area: legal
Ttl2: Recent developments in the admissibility of DNA fingerprint evidence
Case: Spencer v. Commonwealth
Cort: 385 S.E.2d 850 (VA. 1989)
Plac: Virginia
Date: May 1990
Volm: 76(4)
Page: 853-876

95.
Auth: Blake, E. T.//Paabo, S.//Stolorow, M. D.
Affl: Forensic Science Assoc., Richmond; Dept. of Biochemistry, Univ. of CA.; Berkeley; Cellmark Diagnostic, Germantown, MA
Ttl1: DNA amplification and typing from aged biological evidence
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 267-268

96.
Auth: Blake, [redacted]ard T.
Affl: Forensic Science Associates, Richmond, CA.
Ttl1: Scientific and legal issues raised by DNA analysis
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 109-115

97.
Auth: Blake, Edward T.//Higuchi, Russell//von Beroldingen, Cecilia//Sensabaugh, George F.
Affl: Forensic Science Associates; Cetus Corp.; U.C. Berkeley
Ttl1: DNA extraction and evaluation
Type: unpublished article
Area: technical or scientific
Ttl2: Seventy-First Semi-Annual Seminar of the California Association of Criminalists
Plac: Berkeley, CA
Date: May, 1988
Page: 1-18

98.
Auth: Blanchetot, A.
Affl: Dept. of Biochemistry, University of Saskatchewan, Sakatoon, Sask, Canada
Ttl1: Genetic variability of satellite sequence in the dipteran Musca domestica
Type: book chapter
Area: technical or scientific
BkAu: Burek, T.//Dolf, G.//Jeffreys, A. J. Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 106-112

99.
Auth: Bockel, Barbara//Nurnberg, Peter//Krawczak, Michael
Affl: Institut fur Humangenetik, Gottingen; Institut fur Medizinische Genetick, Charite, Berlin; Abteilung Humangenetik, Medizinische Hochschule, Hannover
Ttl1: Likelihoods of multilocus DNA fingerprints in extended families
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: August 1992
Volm: 51
Page: 554-561

100.
Auth: Boerwinkle, Eric//Xiong, Weijun J.//Fourest, Eric//Chan, Lawrence
Affl: Center of Demographic and Population Genetics, University of Texas Health Science Center, Houston, Tx. 77225
Ttl1: Rapid typinng of tandemly repeated hypervariable loci by the polymerase chain reaction: application to the apolipoprotein B 3' hypervariable region
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy Science USA
Date: January 1989
Volm: 86(1)
Page: 212-216

101.
Auth: Boggs, Danny J. (Honorable)
Affl: United States Court of Appeals for the Sixth Circuit, Louisville, Kentucky
Ttl1: Reactions and judicial perspective
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 347-352

102.
Auth: Boltz, E. M.//Harnett, P.//Leary, J.//Houghton, R.//Kefford, R. F.//Friedlander, M. L.
Affl: Department of Medicine, University of Sydney Westmead Centre, NSW, Australia
Ttl1: journal article
Type: technical or scientific
BkAu: British Journal Cancer
Date: July 1990
Volm: 62(1)
Page: 23-27

103.
Auth: Booth, F.//Tilzer, L.//Moreno, R.
Affl: Kansas City Police Department; Kansas University Medical Center, Kansas City
Ttl1: Extraction of genomic DNA from stains without ethanol precipitation
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 231-232

104.
Auth: Bowcock, A. M.//Bucci, C.//Hebert, J. M.//Kidd, J. R.//Kidd, K. K.//Friedlaender, J. S.//Cavalli-Sforza, L. L.
Affl: Stanford Univ.; Yale Univ. School of Medicine; Temple University
Ttl1: Study of 47 DNA markers in five populations from four continents
Type: journal article
Area: technical or scientific
Ttl2: Gene Geography
Date: 1987
Volm: 1
Page: 47-64

105.
Auth: Bragg, [redacted] Nakamura, Y.//Jones, C.//White, R.
Affl: The Howard Hughes Medical Institute, Univ. of Utah Medical School, Salt Lake City, UT; Eleanor Roosevelt Institute for Cancer Research, Denver,CO.
Ttl1: Isolation and mapping of a polymorphic DNA sequence (cTBQ7) on chromosome 10 [D10S28]
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: December 9, 1988
Volm: 16(23)
Page: 11395

106.
Auth: Brdicka, R.
Ttl1: DNA analysis--a new tool for the identification of individuals
Type: journal article
Area: technical or scientific
Ttl2: Casopis Lekaru Ceskkych
Date: June 16, 1989
Volm: 238(25)
Page: 787-789

107.
Auth: Brenig, B.//Brem, G.
Affl: Dept. of Molecular Animal Breeding, Ludwig-Maximilians-University of Munich, Veterinarstrasse, Munchen, Germany
Ttl1: Human VNTR sequences in Porcine HTF-Islands
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 39-49

108.
Auth: Brenner, Sydney//Livak, Kenneth J.
Affl: Medical Research Council Molecular Genetics Unit, Cambridge, England; E.I. DuPont de Nemours & Company, Wilmington, DE.
Ttl1: DNA fingerprinting by sampled sequencing
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy Science USA
Date: November 1989
Volm: 86(22)
Page: 8902-8906

109.
Auth: Brinkmann, Bernd
Affl: Institute of Fornsic Medicine, Munster, Germany
Ttl1: Population studies on selected AMP-FLPs and their use in the investigation of mixtures of body fluids
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 153-155

110.
Auth: Brocas, H.//Georges, M.//Christophe, D.//Monsieur, R.//Lequarre, [redacted] S.//Wassart, G.
Affl: Institut de Recherche Interdisciplinaire, University Libre de Bruxelles
Ttl1: A family of hypervariable minisatellites detected by means of a sequence derived from phage M13
Type: journal article
Area: technical or scientific
Ttl2: C R Academy Science III
Date: 1987
Volm: 304(3)
Page: 67-69

111.
Auth: Brookfield, John
Affl: Dept. of Genetics, University of Nottingham, Nottingham, UK
Ttl1: Interpreting DNA fingerprints
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: April 9, 1992
Volm: 356
Page: 483

112.
Auth: Brookfield, John
Affl: Dept. of Genetics, University of Nottingham, Nottingham, UK
Ttl1: Law and probabilities
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: January 16, 1992
Volm: 355
Page: 207

113.
Auth: Bruford, M. W.//Burke, T.
Affl: University of Leicester, UK
Ttl1: Hypervariable DNA markers and their applications in the chicken
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 230-242

114.
Auth: Buckleton, J.//Walsh, K. A. J.//Triggs, C. M.
Ttl1: A continuous model for interpreting the positions of bands in DNA locus--specific work.
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science Society
Date: 1991
Volm: 31
Page: 3 or 353

115.
Auth: Budowle, Bruce
Affl: FBI Laboratory, Quantico, VA.
Ttl1: AMP-FLPs: genetic markers for forensic identification
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 134-136

116.
Auth: Budowle, Bruce
Affl: FBI Laboratory, Quantico, VA.
Ttl1: A protocol for RFLP analysis of forensic biospecimens
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 57-62

117.
Auth: Budowle, Bruce
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA.
Ttl1: Reply to Green
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: February 1992
Volm: 50
Page: 441-443

118.
Auth: Budowle, Bruce
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: The RFLP technique
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1988
Volm: 15(4)
Page: 97-98

119.
Auth: Budowle, Bruce//Adams, D. E.//Allen, R. C.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Fragment length polymorphisms for forensic science application
Type: book chapter
Area: technical or scientific
BkAu: Karam//Chao//Warr
Ttl2: Methods in Nucleic Acids
Plac: Boca Ratan, FL
Publ: CRC Press
Date: 1991
Page: 181-202

120.
Auth: Budowle, Bruce//Adams, Dwight E.//Comey, Catherine T.//Merril, C. R.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Mitochondrial DNA: a possible genetic material suitable for forensic analysis
Type: book chapter
Area: technical or scientific
BkAu: Lee, Henry C.//Gaensslen, R. E.
Ttl2: Advances in Forensic Sciences: DNA and Other Polymorphisms in Forensic Science
Date: 1990
Page: 76-97

121.
Auth: Budowle, Bruce//Allen, R. C.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Electrophoresis reliability: I. The contaminant issue
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: November 1987
Volm: 32(6)
Page: 1537-1550

122.
Auth: Budowle, Bruce//Baechtel, F. Samuel//Giusti, Alan M.//Monson, Keith L.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Applying highly polymorphic VNTR loci genetic markers to identity testing
Type: journal article
Area: technical or scientific
Ttl2: Clinical Biochemistry
Date: 1990
Volm: 23
Page: 287-293

123.
Auth: Budowle, Bruce//Baechtel, F. Samuel//Giusti, Alan M.//Monson, Keith L.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Data for forensic matching criteria for VNTR profiles
Type: book chapter
Area: technical or scientific
Ttl2: Proceedings for the International Symposium on Human Identification
Plac: Madison, WI
Publ: Promega Corp.
Date: 1990
Page: 103-115

124.
Auth: Budowle, Bruce//Baechtel, F. Samuel
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Modifications to improve the effectiveness of restriction fragment length polymorphism typing
Type: journal article
Area: technical or scientific
Ttl2: Applied and Theoretical Electrophoresis
Date: 1990
Volm: 1
Page: 181-187

125.
Auth: Budowle, Bruce//Baechtel, F. Samuel//Adams, Dwight E.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Validation with regard to environmental insults of the RFLP procedure for forensic purposes
Type: book chapter
Area: technical or scientific
BkAu: Farley, Mark A.//Harrington, James J.
Ttl2: Forensic DNA Technology
Date: 1991
Page: 83-91

126.
Auth: Budowle, Bruce//Chakraborty, Ranajit//Giusti, Alan M.//Eisenberg, Arthur//-/Allen, Robert C.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Analysis of the VNTR locus D1S80 by PCR followed by high resolution polyacrylamide gel electrophoresis
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: January 1991
Volm: 48(1)
Page: 137-144

127.
Auth: Budowle, Bruce//Deadman, Harold A.//Murch, Randall S.//Baechtel, F. Samuel
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: An introduction to the methods of DNA analysis under investigation in the FBI laboratory
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: January 1988
Volm: 15(1)
Page: 8-21

128.
Auth: Budowle, Bruce//Giusti, Alan M.//Chakraborty, Ranajit
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA.; Univ. of Texas Graduate School of Biomedical Sciences, Houston,Tx
Ttl1: Discretized allelic data for VNTR locus by amplified fragment length polymorphism (AMP-FLP) analysis
Type: abstract
Area: technical or scientific
Ttl2: 41st American Society of Human Genetics
Plac: Cincinnati, OH.
Date: October 16-20, 1990
Volm: 47(3) Supplement
Page: A129

129.
Auth: Budowle, Bruce//Giusti, Alan M.//Waye, John S.//Baechtel, F. Samuel//Fourney, R. M.//Adams, Dwight E.//Presley, L. E.//Deadman, Harold A.//Monson, Keith L.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA; Royal Canadian Mounted Police, Ottawa, Ontario; FBI, Washington,DC
Ttl1: Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic comparisons
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 48
Page: 841-855

130.
Auth: Budowle, Bruce//Monson, Keith L.//Anoe, Kim S.//Baechtel, F. Samuel//Bergman, Dan L.//Buel, Eric//Campbell, Priscilla et al
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: A preliminary report on binned general population data on six VNTR loci in caucasians, blacks and hispanics from the United States
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: January 1991
Volm: 18(1)
Page: 9-26

131.
Auth: Budowle, Bruce//Monson, Keith L.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: A statistical approach for VNTR analysis
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of an International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Volm: 121-126

132.
Auth: Budowle, Bruce//Stafford, J.
Affl: Forensic SCience Research and Training Section, FBI Academy, Quantico, VA.
Ttl1: Response to expert report by F. L. Hartl
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Case: United States v. Yee
Date: July 1991
Volm: 18
Page: 101

133.
Auth: Budowle, Bruce//Stafford, J.
Affl: Forensic Science Reseach and Training Section, FBI Academy, Quantico, Va.
Ttl1: Response to "Population genetic problems in the forensic use of DNA profiles"
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Case: United States v. Yee
Date: July 1991
Volm: 18
Page: 109

134.
Auth: Budowle, Bruce//Waye, John S.//Shutler, Gary G.//Baechtel, F. Samuel
Affl: Forensic Science Research and Training Center, FBI Academy, Quantico, VA
Ttl1: HAE III--a suitable restriction endonuclease for restriction fragment length polymorphism analysis of biological evidence samples
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: May 1990
Volm: 35(3)
Page: 530-536

135.
Auth: Buffery, C.//Catterick, T.//Greenhalgh, M.//Jones, S.//Russell, J. R.
Affl: Metropolitan Police Forensic Science Laboratory, London, UK
Ttl1: Assessment of a video system for scanning DNA autoradiographs
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: 1991
Volm: 49
Page: 17-20

136.
Auth: Bugawan, Teodorica L.//Saiki, Randall K.//Levenson, Corey H.//Watson, Robert M.//Erlich, Henry A.
Affl: Cetus Corporation, 1400 Fifty-Third Street, Emeryville, CA. 94608
Ttl1: The use of non-radioactive oligonucleotide probes to analyze enzymatically amplified DNA for prenatal diagnosis and forensic HLA typing
Type: journal article
Area: technical or scientific
Ttl2: Bio/Technology
Date: August 1988
Volm: 6
Page: 943-947

137.
Auth: Buitkamp, J.//Ammer, H.//Geldermann, H.
Ttl1: DNA fingerprinting in domestic animals
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 169-174

138.
Auth: Buoncristiani, M.//von Beroldingen, C.//Sensabaugh, G. F.
Affl: University of California, Berkeley, CA.
Ttl1: Effects of UV damage on DNA amplification by the polymerase chain reaction
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 259

139.
Auth: Burk, Dan L.
Ttl1: DNA fingerprinting: possibilities and pitfalls of a new technique
Type: journal article
Area: legal
Ttl2: Jurimetrics Journal of Law, Science and Technology
Plac: United States
Date: Summer, 1988
Volm: 28(4)
Page: 455-471

140.
Auth: Burke, T.//Bruford, M. W.
Affl: Department of Zoology, University of Leicester, UK
Ttl1: DNA fingerprinting in birds
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: May 14-20, 1987
Volm: 327(6118)
Page: 149-152

141.
Auth: Burke, T.//Hanotte, O.//Bruford, M. W.//Cairns, E.
Affl: Univ. of Leicester, UK; Universite de Mons-Hainaut, Mons, Belgium
Ttl1: Multilocus and single locus minisatellite analaysis in population biological studies
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 154-168

142.
Type: book
Area: general reference
BkAu: Burke, Terry//Dolf, Gaudenz//Jeffreys, Alec J.//Wolff, Roger
Ttl2: DNA Fingerprinting: Approaches and Applications
Plac: P. O. Box 133, 4010 Basel, Switzerland
Publ: Birkhauser Verlag
Date: 1991
Volm: ISBN 3-7643-2562-3; ISBN 0-8176-2562-3
Page: v-x, 1-398

143.
Auth: Burke, Tod W.//Rowe, Walter F.
Affl: St. Anselm College, Mass.; The George Washington Univ., Washington,D.C.
Ttl1: DNA analysis: The challenge for police
Type: journal article
Area: technical or scientific
Ttl2: The Police Chief
Date: October 1989
Page: 92-94

144.
Auth: Byers, Brad R.
Ttl1: DNA fingerprinting and the criminal defendant: guilty or innocent? Only his molecular biologist knows for sure
Type: journal article
Area: legal
Ttl2: Ohio Northern University Law Review
Plac: United States
Date: Winter 1989
Volm: 16(1)
Page: 57-79

145.
Auth: Capon, Daniel J.//Chen, Ellson Y.//Levinson, Arthur D.//Seeburg, Peter H.//Goeddel, David V.
Affl: Genetech, Inc., 4600 Point San Bruno Blvd, So. San Francisco, CA.
Ttl1: Complete nucleotide sequences of the T24 human bladder carcinoma oncogene and its normal homologue
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: March 3, 1983
Volm: 302(5903)
Page: 33-37

146.
Auth: Caskey, C. Thomas
Affl: Institute for Molecular Genetics, Baylor College of Medicine, Houston, Tx.
Ttl1: Book review: "Forensic DNA Technology"
Type: book review
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1992
Volm: 50
Page: 1143-1144

147.
Auth: Caskey, C. Thomas
Affl: Institute for Molecular Genetics and Howard Hughes Medical Institute; Baylor College of Medicine, Houston, Tx.
Ttl1: Comments on DNA-based forensic analysis
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: October 1991
Volm: 49(4)
Page: 893-895

148.
Auth: Caskey, C. Thomas
Affl: Dept. of Medical Genetics, Baylor College of Medicine, Houston, Tx.
Ttl1: New aid to human gene mapping
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: March 7, 1985
Volm: 314
Page: 19

149.
Auth: Caskey, C. Thomas
Affl: Institute for Molecular Genetics, Baylor College of Medicine, Houston, Tx.
Type: letter of response
Area: technical or scientific
Ttl2: Professional Ethics Report
Plac: 1333 H Street, NW, Washington, DC 20005
Publ: American Association for the Advancement of Science
Date: Spring 1992
Volm: V(2)
Page: 4

150.
Auth: Caskey, C. Thomas
Affl: Institute for Molecular Genetics, Baylor College of Medicine, Houston, Tx.
Ttl1: untitled
Type: letter to editor
Area: technical or scientific
Ttl2: The New York Times
Plac: New York City, N.Y.
Date: February 9, 1990

151.
Auth: Caskey, C. Thomas//Edwards, A.//Hammond, H. A.
Affl: Baylor College of Medicine, Houston, TX.
Ttl1: DNA: The history and future use in forensic analysis
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 3-9

152.
Auth: Caskey, C. Thomas//Hammond, Holly
Affl: Baylor College of Medicine, Houston, Tx.
Ttl1: DNA-based identification: disease and criminals
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 127-142

153.
Type: book
Area: general reference
BkAu: Cavalli-Sforza, L. L.//Bodmer, W. F.
Ttl2: The Genetics of Human Populations
Plac: San Francisco, CA
Publ: W. H. Freeman and Company
Date: 1971
Volm: ISBN 0-7167-0681-4
Page: i-xvi,1-965

154.
Auth: Cavalli-Sforza, Luigi Luca
Affl: Department of Genetics, Stanford University, Stanford, CA.
Ttl1: A.2 Measures of variability: the variance and standard deviation
Type: book chapter
Area: technical or scientific
BkAu: Crow, James F.//Kimura, Motoo
Ttl2: An Introduction to Population Genetics Theory
Plac: 49 East 33rd Street, New York, N.Y. 10016
Publ: Harper & Row, Publishers, Inc.
Date: 1970
Page: 482-486

155.
Auth: Cavalli-Sforza, Luigi Luca
Affl: Department of Genetics, Stanford University, Stanford, CA
Ttl1: Randomly mating populations (2.6 Two Loci)
Type: book chapter
Area: technical or scientific
BkAu: Crow, James E.//Kimura, Motoo
Ttl2: An Introduction to Population Genetics Theory
Plac: 49 East 33rd Street, New York, N.Y. 10016
Publ: Harper & Row, Publishers, Inc.
Date: 1970
Page: 47-56

156.
Auth: Cavalli-Sforza, Luigi L.//Kidd, J. R.//Kidd, K. K.//Bucci, C.//Bowcock, A. M.//Hewlett, B. S.//Freidlaender, J. S.
Affl: Department of Genetics, Stanford University Medical Center, CA 04305
Ttl1: DNA markers and genetic variation in the human species
Type: journal
Area: technical or scientific
Ttl2: Cold Spring Harbor Symposium Quant Biology
Date: 1986
Volm: 51 Pt. 1
Page: 411-417

157.
Auth: Cawood, A. H.
Affl: ICI Diagnostics, Germantown, MD 20874
Ttl1: DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Clinical Chemistry
Date: September 1989
Volm: 35(9)
Page: 1832-1837

158.
Auth: Cawood, A.//Webb, M.//Riley, J.//Mead, R.//Markham, A. F.//Smith, J. C.
Affl: Cellmark Diagnostics, Abingdon, England; OX14 IDY and ICI Diagnostics, Cheshire, England
Ttl1: Characterization of MS51. A single locus VNTR probe which provides a quality control procedure to demonstrate complete digestion with restriction enzymes
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 237

159.
Auth: Cerda-Flores, R. M.//Kshatriya, G. K.//Barton, S. A.//Leal-Garza, C.//Garza--Chapa, R.//Schul, W. J.//Chakraborty, R.
Affl: Unidad de Investigacion Bimedica del Noreste, Mexico; Nat'l Inst. of Health and Family Welfare, India; University of Texas, USA
Ttl1: Genetic structure of the immigrant populations of San Luis Potosi and Zacatecas to Nuevo Leon in Mexico.
Type: journal article
Area: technical or scientific
Ttl2: Human Biology
Date: June 1991
Volm: 63(3)
Page: 309-327

160.
Auth: Cerde-Flores, R. M.//Kshatriya, G. K.//Bertin, T. K.//Hewett-Emmett, D./-/Hanis, C. L.//Chakraborty, R.
Affl: Unidad de Investigacion Biomedica del Noreste, Mexico; National Institute of Health and Family Welfare, India; Univ. of Texas, USA
Ttl1: Gene diversity and estimation of genetic admixture among Mexican-American of Starr County, Texas
Type: journal article
Area: technical or scientific
Ttl2: Annal Human Biology
Date: 1992
Volm: 19
Page: 347-360

161.
Auth: Chakrab[redacted]y, Ranajit
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Book review: "DNA Technology and Forensic Science"
Type: book review
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 48
Page: 173-174

162.
Auth: Chakraborty, Ranajit
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Book review: "Genetic Data Analysis"
Type: book review
Area: technical or scientific
Ttl2: Molecular Biology Evolution
Date: 1991
Volm: 8(3)
Page: 396-397

163.
Auth: Chakraborty, Ranajit
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Generalized occupancy problem and its applications in population genetics
Type: book chapter
Area: technical or scientific
BkAu: Sing, C. F.//Hanis, C. L.
Ttl2: Impact of Genetic Variation on Individuals, Families and Populations
Plac: New York
Publ: Oxford University Press
Date: 1991
Page: (in press)

164.
Auth: Chakraborty, Ranajit
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Genetic profile of cosmopolitan populations: Effects of hidden subdivision
Type: journal article
Area: technical or scientific
Ttl2: Anthrop Anzeiger
Date: December 1990
Volm: 48
Page: 313-331

165.
Auth: Chakraborty, Ranajit
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Letters to the Editor: Multiple alleles and estimation of genetic parameters: computational equations showing involvement of all alleles.
Type: journal article
Area: technical or scientific
Ttl2: Genetics
Date: January 1992
Volm: 130(1)
Page: 231-234

166.
Auth: Chakraborty, Ranajit
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: NRC Report on DNA Typing
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: May 21, 1993
Volm: 26
Page: 1059

167.
Auth: Chakraborty, Ranajit
Affl: Univ. of Texas Health Science Center, Graduate School of Biomedical Sciences, Houston, Tx. 77225-0334
Type: letter of response
Area: technical or scientific
Ttl2: Professional Ethics Report
Plac: 1333 H Street, NW, Washington, DC 20005
Publ: American Association for the Advancement of Science
Date: Spring 1992
Volm: V(2)
Page: 3-4

168.
Auth: Chakraborty, Ranajit
Affl: University of Texas Graduate School of Biomedical Sciences
Ttl1: Sample size requirements for addressing the population genetic issues of forensic use of DNA typing
Type: journal article
Area: technical or scientific
Ttl2: Human Biology
Date: April 1992
Volm: 64(2)
Page: 141-159

169.
Auth: Chakraborty, Ranajit
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Statistical Interpretation of DNA typing data
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 49
Page: 895-897

170.
Auth: Chakraborty, Ranajit//Boerwinkle, Eric
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Population genetics of VNTR polymorphisms in human
Type: journal article
Area: technical or scientific
Ttl2: American Journal of Human Genetics
Date: 1990
Volm: 47
Page: A129

171.
Auth: Chakraborty, Ranajit//Daiger, Stephen P.//Boerwinkle, Eric
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Patterns of genetic variation within and between populations detected by PCR-based VNTR polymorphisms.
Type: journal article
Area: technical or scientific
Ttl2: Proceedings Int. Seminar of the Forensic Application of PCR Technology
Date: 1992
Volm: 18
Page: 148-152

172.
Auth: Chakraborty, Ranajit//Daiger, Stephen P.
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah
Type: journal article
Area: technical or scientific
Ttl2: Human Biology
Date: October 1991
Volm: 63(5)
Page: 571-587

173.
Auth: Chakraborty, Ranajit//de Andrade, M.//Daiger, Stephen P.//Budowle, Bruce
Affl: Univ. of Texas Graduate School of Biomedical Sciences, Houston, Tx.; FBI Academy, Quantico, VA.
Ttl1: Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications
Type: journal article
Area: technical or scientific
Ttl2: Annal Humum Genetics
Date: 1992
Volm: 56
Page: 45-47

174.
Auth: Chakraborty, Ranajit//Deka, Ranjan//Jin, Li//Ferrell, Robert E.
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.; University of Pittsburgh, PA.
Ttl1: Allele sharing at six VNTR loci and genetic distances among three ethnically defined human population.
Type: journal article
Area: technical or scientific
Ttl2: American Journal of Human Biology
Date: 1992
Volm: 4
Page: 387-397

175.
Auth: Chakraborty, Ranajit//Fornage, M.//Gueguen, R.//Boerwinkle, Eric
Affl: Univ. of Texas Graduate School of Biomedical Sciences, Houston, Tx.; Center of Preventive Medicine, Nancy, France
Ttl1: Population genetics of hypervariable loci: analysis of PCR based VNTR polymorphism within a population
Type: book chapter
Area: technical or scientific
BkAu: Burek, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 127-143

176.
Auth: Chakraborty, Ranajit//Jin, Li
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Heterozygote deficiency, population substructure and their implications in DNA fingerprinting.
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: 1992
Volm: 88(3)
Page: 267-272

177.
Auth: Chakraborty, Ranajit//Kamboh, Mohammad I.//Nwankwo, M.//Ferrell, Robert E.
Affl: University of Texas Graduate School of Biomedical Sciences; University of Pittsburgh, PA.; University of Benin, Benin City, Nigeria
Ttl1: Caucasian genes in the American Blacks: new data
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: January 1992
Volm: 50(1)
Page: 145-155

178.
Auth: Chakraborty, Ranajit//Kidd, Kenneth K.
Affl: Univ. of Texas Graduate School of Biomedical Sciences, Houston, Tx.; Dept. of Genetics, Yale University School of Medicine, New Haven, Ct.
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: February 28, 1992
Volm: 255
Page: 1053

179.
Auth: Chakraborty, Ranajit//Kidd, Kenneth
Affl: University of Texas Health Science Center, Houston, Tx.; Yale University
Ttl1: The utility of DNA typing in forensic work
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: December 20, 1991
Volm: 254
Page: 1735-1739

180.
Auth: Chakrab[redacted], Ranajit//Rao, C. R.
Affl: Univ. of Texas Graduate School of Biomedical Sciences, Houston, Tx., Dept. of Mathematics and Statistics, Univ. of Pittsburgh, Pittsburgh, PA.
Ttl1: Measurement of genetic variation for evolutionary studies
Type: book chapter
Area: technical or scientific
BkAu: Chakraborty, Rananjit//Rao, C. R.
Ttl2: Handbook of Statistics 8: Statistical Methods for Biological and Medical Sciences
Plac: New York
Publ: North-Holland
Date: 1991
Volm: 8
Page: 271-316

181.
Auth: Charrow, Robert P.
Affl: Attorney - Crowell & Moring in Washington, D. C.
Ttl1: Forensic evidence in the courtroom is really probability theory in the jury box
Type: journal article
Area: technical or scientific
Ttl2: Journal of NIH Research
Date: June 1991
Volm: 3
Page: 96-97

182.
Auth: Chen, Phil//Hayward, Nicholas K.//Kidson, Chev//Ellem, Kay A.
Affl: Queensland Institute of Medical Research, Herston, Brisbane, Australia
Ttl1: Conditions for generating well-resolved human DNA fingerprints using M13 phage DNA
Type: technical or scientific
Area: journal article
Ttl2: Nucleic Acids Research
Date: February 25, 1990
Volm: 18(4)
Page: 1065

183.
Auth: Chen, Philip//Hurst, Terence//Khoo, Soo Keat
Affl: University of Queensland, Brisbane, Australia
Ttl1: Detection of somatic DNA alterations in ovarian cancer by DNA fingerprint analysis
Type: journal article
Area: technical or scientific
Ttl2: Cancer
Date: March 15, 1991
Volm: 67
Page: 1551-1555

184.
Auth: Cherfas, Jeremy
Ttl1: Ancient DNA: still busy after death
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: September 10, 1991
Volm: 253
Page: 1354-1356

185.
Auth: Chiafari, F. A.//Wenk, R. E.
Affl: Baltimore Rh Typing Laboratory, Inc., Baltimore, MD.
Ttl1: Parentage analysis by endonuclease shattering of hypervariable DNA
Type: journal article
Area: technical or scientific
Ttl2: Transfusion
Date: September 1990
Volm: 30(7)
Page: 648-650

186.
Auth: Chorazy, Paula A.//Edlind, Thomas D.
Affl: Dept. of Microbiology and Immunology, Medical College of Pennsylvania, PA
Ttl1: Artifactual bands associated with alkaline transfer
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: 1990
Volm: 18(10)
Page: 3101

187.
Auth: Clarke, George W.
Affl: Office of the District Attorney, San Diego, CA.
Ttl1: Legal issues pertinent to typing of DNA
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 195-199

188.
Auth: Clayborn, Charity Lynn
Ttl1: Evidence - criminal law - evidence of DNA fingerprinting admitted for identification purposes in rape trial
Type: journal article
Area: legal
Ttl2: University of Arkansas at Little Rock Law Journal
Case: Andrews v. State
Cort: 533 So. 2d 842 (Fla. Dis. Court App. 1988)
Plac: Florida
Date: Summer 1989
Volm: 12(3)
Page: 543-556

189.
Auth: Cleveland, Don W.
Affl: Dept. of Biological Chemistry, Johns Hopkins University, Baltimore, MD
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: February 28, 1992
Volm: 255
Page: 1050

190.
Auth: Cobb, Kim
Affl: Houston Chronicle
Ttl1: DNA fingerprinting foolproof test - New tool to ID criminals may soon be widely used
Type: newspaper
Area: lay press
Ttl2: Houston Chronicle
Publ: Houston, Texas
Date: August 7, 1989
Page: 7B-8B

191.
Auth: Cobb, Kim
Affl: Writer - Houston Chronicle
Ttl1: DNA vs blood test for proving paternity
Type: newspaper
Area: lay press
Ttl2: Houston Chronicle
Plac: Houston, Texas
Date: August 7, 1989
Page: 8B

192.
Auth: Cohen, Joel E.
Affl: Rockfeller University, New York, NY
Ttl1: The ceiling principle is not always conservative in assigning genotype frequencies for forensic DNA testing
Type: Letter to the Editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: November 1992
Volm: 51(5)
Page: 1165-1168

193.
Auth: Cohen, Joel E.
Affl: Rockefeller University, New York, N.Y.
Ttl1: DNA fingerprinting: what (really) are the odds?
Type: journal article
Area: technical or scientific
Ttl2: Chance, New Directions for Statistics and Computing
Date: 1990
Volm: 3(3)
Page: 26-32

194.
Auth: Cohen, Joel E.
Affl: Rockefeller University, New York, NY 10021-6399
Ttl1: DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation of heterogeneity and band number variability
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: February 1990
Volm: 46(2)
Page: 358-368

195.
Auth: Cohen, Joel E.//Lynch, Michael//Taylor, Charles E.
Affl: Rockfeller University; University of Oregon; University of California
Ttl1: Forensic DNA Tests and Hardy-Weinberg Equilibrium
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: August 30, 1991
Volm: 253
Page: 1037-1041

196.
Auth: Coleman, H. C.//MacClaren, D. C.//van Dijk, K. W.//Lotshaw, C. J.//Milner, E. C. B.
Affl: GeneLex Corp.; Washington State Patrol; Virginia Mason Research Center, Seattle, Washington
Ttl1: Application of a new DNA probe to forensic analysis
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 277-278

197.
Auth: Comey, C. Thomas
Affl: FBI Laboratory, Quantico, VA.
Ttl1: DNA sequencing
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 103-111

198.
Auth: Comey, [redacted]herine T.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: The use of DNA amplification in the analysis of forensic evidence
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1988
Volm: 15(4)
Page: 99-103

199.
Auth: Comey, Catherine T.
Affl: FBI Laboratory, Quantico, VA
Ttl1: Validation of the HLA DQa typing system
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 129-131

200.
Auth: Comey, Catherine Theisen//Budowle, Bruce
Affl: Forensic Science Research and Training Center, FBI Academy, Quantico, VA
Ttl1: Validation studies on the analysis of the HLA-HQa locus using the polymerase chain reaction
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: November 1991
Volm: 36(6)
Page: 1633-1648

201.
Auth: Comey, Catherine Theisen//Jung, Janet M.//Budowle, Bruce
Affl: Forensic Science Research and Training Center, FBI Academy, Quantico, VA
Ttl1: Use of formamide to improve amplification of HLA DQa sequences
Type: journal article
Area: technical or scientific
Ttl2: BioTechniques
Date: January 1991
Volm: 10(1)
Page: 60-61

202.
Auth: Coppieters, W.//van DeWeghe, A.//Depicker, A.//Bouquet, Y.//van Zeveren, A.
Affl: Department of Animal Genetics and Breeding, Faculty of Veterinary Medicine, State University of Ghent, Belgium
Ttl1: A hypervariable pig DNA fragment
Type: journal article
Area: technical or scientific
Ttl2: Animal Genetics
Date: 1990
Volm: 21(1)
Page: 29-38

[redacted].
Auth: Coquoz, R.
Affl: Institut de Police Scientifique et de Crimonologie, Lausanne, Switzerland
Ttl1: The use of non-isotopic detection methods in DNA fingerprinting
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 207-208

204.
Auth: Cotton, Robin W.//Anderson, Mariane B.//Herrin Jr., George//Corey, Amy C.//Sheridan, Kathleen T.//Tonelli, Lois A.//Washowski, Clare A.//Garner, Daniel D.
Affl: Cellmark Diagnostics, Germantown, Maryland
Ttl1: Current case experience with single locus hypervariable probes
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 191-206

205.
Auth: Cotton, Robin W.//Forman, Lisa//Word, Charlotte J.
Affl: Cellmark Diagnostics, Germantown, MD
Ttl1: Research on DNA typing validated in the literature
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: October 1991
Volm: 49
Page: 898-899

206.
Auth: Crocker, Laura H.
Ttl1: DNA typing: novel scientific evidence in the military courts
Type: journal article
Area: legal
Ttl2: Air Force Law Review
Date: 1990
Volm: 32
Page: 227

207.
Auth: Daiger, Stephen P.
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: DNA fingerprinting
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: October 1991
Volm: 49(4)
Page: 897

208.
Auth: Daiger, Stephen D.
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Issues in DNA fingerprinting for forensic purposes
Type: unpublished document
Area: technical or scientific
Date: April 11, 1991
Srce: State Bar of Texas Professional Development Program

209.
Auth: Daiger, Stephen P.
Affl: Univ. of Texas Health Science Center, Graduate School of Biomedical Sciences, Houston, Tx. 77225-0334
Type: letter of response
Area: technical or scientific
Ttl2: Professional Ethics Report
Plac: 1333 H Street, NW, Washington, DC 20005
Publ: American Association for the Advancement of Science
Date: Spring 1992
Volm: V(2)
Page: 6

210.
Auth: Dallapiccola, Bruno//Novelli, Giuseppe//Spinella, Aldo
Affl: IInd University of Rome, Via Ramazzini; Central Criminal Police, III Scientific Division, Viale Aeronautica, Rome, Italy
Ttl1: PCR DNA typing for forensics
Type: letter to editor
Area: technical or scientific
Ttl2: Nature
Date: November 21, 1991
Volm: 354
Page: 179

211.
Auth: Dallas, John F.
Affl: Department of Genetics, University of Nottingham Medica School, Queens Medical Centre, Nottingham, UK
Ttl1: Detection of DNA "fingerprints" of cultivated rice by hybridization with a human minisatellite DNA probe
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy Science USA
Date: September 1988
Volm: 85(18)
Page: 6831-6835

212.
Auth: Daly, A.//Kellam, P.//Berry, S. T.//Chojecki, A. J. S.//Barnes, S. R.
Affl: I.C.I. Seeds, Jealott's Hill Research Station, Bracknell, Berkshire, UK
Ttl1: The isolation and characterisation of plant sequences homologous to human hypervariable minisatellites
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 330-341

213.
Auth: Damore, Michael
Ttl1: DNA fingerprinting: what every criminal lawyer should know
Type: journal article
Area: legal
Ttl2: Criminal Law Bulletin
Plac: United States
Date: March-April 1991
Volm: 27(2)
Page: 114-133

214.
Auth: Das, Aparup
Affl: Centre of Advanced Study in Zoology, Banaras Hindu University, Varansasi, India
Ttl1: DNA fingerprinting in India
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: April 4, 1991
Volm: 350
Page: 387

215.
Auth: Davies, A.
Ttl1: The use of DNA profiling and behavioural science in the investigation of sex offences
Type: journal article
Area: technical or scientific
Ttl2: Med. Science Law
Date: 1991
Volm: 31
Page: 95

216.
Auth: Davies, Elizabeth W.
Ttl1: The Family Law Reform Act 1987 and DNA fingerprinting
Type: journal article
Area: legal
Ttl2: Family Law
Plac: Great Britain
Date: June 1988
Volm: 18
Page: 221-223

217.
Auth: D'Dustachio, Peter
Affl: Dept. of Biochemistry, New York University Medical Center, NY, NY
Ttl1: Interpreting DNA fingerprints
Type: letter to editor
Area: technical or scientific
Ttl2: Nature
Date: April 9, 1992
Volm: 356
Page: 483

218.
Auth: de Gouyo[redacted] Beatrice//Julier, Cecile//Avner, P.//Georges, Michel//Lathrop, Mark
Affl: Centre d'Etude du Polymorphism Humain; Institut Pasteur; Paris, France; Genmark Inc., Salt Lake, Utah
Ttl1: Human variable number of tandem repeat probes as a source of polymorphic markers in experimental animals
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 85-94

219.
Auth: DeBenedictis, Don J.
Ttl1: DNA report raises concerns
Type: journal article
Area: technical or scientific
Ttl2: ABA Journal
Date: July 1992
Page: 20

220.
Auth: Debenham, P. G.
Affl: Cellmark Diagnostics, Blacklands Way, Abingdon, Oxforshire, UK
Ttl1: DNA Fingerprinting; a biotechnology in business
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G. Jeffreys ,. A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 342-348

221.
Auth: Debenham, Paul G.
Affl: Cellmark Diagnostics, Abingdon Business Park, Abingdon, Oxon, UK
Ttl1: DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Journal Pathology
Date: 1991
Volm: 164
Page: 101-106

222.
Auth: Debenham, Paul
Affl: International Scientific Services, Cellmark Diagnostics, Abingdon, Oxfordshire, UK
Ttl1: The use of genetic markers for personal identification and the analysis of family relationships
Type: journal article
Area: technical or scientific
Ttl2: Ciba Foundation Symposium
Date: 1990
Volm: 149
Page: 37-43 (discussion 43-47)

[redacted].
Auth: Decorte, R.//Cassiman, J. J.
Affl: Center for Human Genetics, University of Leuven, Campus Gasthuisberg, Leuven, Belgium
Ttl1: Detection of amplified VNTR alleles by direct chemiluminescence: application to the genetic indentification of biological samples in forensic cases
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 371-390

224.
Auth: De-Gorgey, Andrea
Ttl1: The advent of DNA databanks: Implications for information privacy
Type: journal article
Area: technical or scientific
Ttl2: American Journal of Law & Medicine
Volm: XVI #3
Page: 381-398

225.
Auth: DeJohn, D.//Woetdijk, B. M.//Kluin-Nelemans, J. C.//van Ommen, G. J.//Kluin, P. M.
Affl: Laboratory of Pathology, University Medical Centre, Leiden, The Netherlands
Ttl1: Somatic changes in B-lymphoproliferative disorders (B-LPD) detected by DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: British Journal Cancer
Date: December 1988
Volm: 58(6)
Page: 773-775

226.
Auth: Deka, Ranjan//Chakraborty, Ranajit//Ferrell, Robert E.
Affl: University of Pittsburgh, PA; University of Texas, Houston, Tx.
Ttl1: Allele sharing and genetic distance at VNTR loci among three ethnic groups.
Type: journal article
Area: technical or scientific
Ttl2: 8th International Congress of Human Genetics
Plac: Washington, D. C.
Date: October 1991
Volm: 49(4)
Page: 497

227.
Auth: Deka, R[illegible]an//Chakraborty, Ranajit//Ferrell, Robert E.
Affl: University of Pittsburgh, PA; University of Texas Health Science Center, Houston, Tx.
Ttl1: A population genetic study of six VNTR loci in three ethnically defined populations
Type: journal article
Area: technical or scientific
Ttl2: Genomics
Date: September 1991
Volm: 11(1)
Page: 83-92

228.
Auth: Devlin, B.//Krontiris, Theodore//Risch, Neil
Affl: Depts. Epidemiology & Public Health, & Genetics, Yale University, New Haven; Dept. of Medicine, Tufts University School of Medicine, New England Medical Center Hospital, Boston
Ttl1: Population genetics of the HRASI minisatellite locus
Type: journal article
Area: technical or scientific
Ttl2: American Jouranl Human Genetics
Date: December 1993
Volm: 53(6)
Page: 1298-1305

229.
Auth: Devlin, B.//Risch, Neil//Roeder, Kathryn
Affl: Yale University, New Haven, CT.
Ttl1: Estimation of allele frequencies for VNTR loci
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 48
Page: 662-676

230.
Auth: Devlin, B.//Risch, Neil
Affl: Depts. of Epidemiology & Public Health and Human Gentics, Yale University, New Haven, Conn.
Ttl1: Ethnic differentiation at VNTR loci, with special reference to forensic applications
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: August 1992
Volm: 51
Page: 534-548

231.
Auth: Devlin, B.//Risch, N.//Roeder, K.
Ttl1: Forensic inference from DNA fingerprints
Type: journal article
Area: technical or scientific
Ttl2: Journal American Statistical Society
Date: 1992
Volm: 87
Page: 337-350

232.
Auth: Devlin, B.//Risch, Neil//Roeder, Kathryn
Affl: Department of Epidemiology and Public Health, Yale University, New Haven, Ct. 06510
Ttl1: No excess of homozygosity at loci used for DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: September 21, 1990
Volm: 249(4975)
Page: 1416-1420

233.
Auth: Devlin, B.//Risch, Neil
Affl: Dept. of Epidemiology & Public Health and Human Genetics, Yale University, New Haven, Conn.
Ttl1: A note of Hardy-Weinberg Equilibrium of VNTR data by using the Federal Bureau of Investigation's fixed-bin method
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: August 1992
Volm: 51
Page: 549-553

234.
Auth: Devlin, B.//Risch, Neil//Roeder, Kathryn
Affl: School of Medicine & Dept. of Statistic, Yale University
Ttl1: NRC Report on DNA Typing
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: May 21, 1993
Volm: 260
Page: 1057

235.
Auth: Devlin, B.//Risch, Neil
Affl: Dept. of Epidemiology and Public Health; Dept. of Genetics, Yale University, New Haven
Ttl1: Physical properties of VNTR data, and their impact on a test of allelic independence
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: August 1993
Volm: 53
Page: 324-329

236.
Auth: Devlin, B.//Risch, Neil//Roeder, Kathryn
Affl: School of Medicine; Dept. of Statistics, Yale University, New Haven, CT
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: August 30, 1991
Volm: 253
Page: 1039-1041

237.
Auth: Devlin, B.//Risch, Neil//Roeder, Kathryn
Affl: Yale University, New Haven, CT.
Ttl1: Statistical evaluation of DNA fingerprinting: A critique of the NRC's report
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: February 5, 1993
Volm: 259
Page: 748-749, 837

238.
Auth: Devor, E. J.//Ivanovich, A. K.//Hickok, J. M.//Todd, R. D.
Affl: Department of Psychiatry, Washington University School of Medicine, St. Louis, MO 63110
Ttl1: A rapid method for confirming cell line identity: DNA "fingerprinting" with a minisatellite probe from M13 bacteriophage
Type: journal article
Area: technical or scientific
Ttl2: Biotechniques
Date: March 1988
Volm: 6(3)
Page: 200,202

239.
Auth: Dickson, David
Affl: Nature magazine writer
Ttl1: Academy under fire over plans for new study of DNA statistics....
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: January 13, 1994
Volm: 367(6459)
Page: 101-102

240.
Auth: Dixson, A. F.//Hastie, N.//Patel, I.//Jeffreys, A. J.
Affl: MRC Reproductive Biology Unit, Centre for Reproductive Biology, Edinburgh, UK
Ttl1: DNA "fingerprinting" of captive family groups of common marmosets (Callithrix jacchus)
Type: journal article
Area: technical or scientific
Ttl2: Folia Primatologica (Basel)
Date: 1988
Volm: 51(1)
Page: 52-55

241.
Auth: Dodd, Barbara E.
Affl: Professor of Blood Group Serology, London Hospital Medical College, London, UK
Ttl1: DNA fingerprinting in matters of family and crime
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: December 12-18, 1985
Volm: 318(6046)
Page: 506-507

242.
Auth: Dodd, Barbara E.
Ttl1: DNA fingerprinting in matters of family and crime
Type: journal article
Area: legal
Ttl2: Medicine, Science and the Law
Plac: Great Britain
Date: January 1986
Volm: 26(1)
Page: 5-7

243.
Auth: Dolf, G.//Glowatzki, M. L.//Gaillard, C.
Affl: Institute of Animal Breeding, University of Berne, Berne, Switzerland
Ttl1: Searching for genetic markers for hereditary diseases in cattle by means of DNA fingerprinting
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 109-112

244.
Auth: Dougherty, John Caleb
Ttl1: Beyond People v. Castro: a new standard of admissibility for DNA fingerprinting
Type: journal article
Area: legal
Ttl2: Journal of Contemporary Health Law and Policy
Case: People v. Castro, 545 N. Y. S. 2d 985 (N Y. App Div 1989)
Plac: New York (State)
Date: Spring 1991
Volm: 7
Page: 269-306

245.
Auth: Dover, Gabriel A.
Affl: Department of Genetics, University of Cambridge, Cambridge, UK
Ttl1: Mapping frozen accidents
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: April 26, 1990
Volm: 344(6269)
Page: 812-813

246.
Auth: Dover, [redacted]iel A.
Affl: Department of Genetics, University of Cambridge, Cambridge, UK
Ttl1: Victims or perpetrators of DNA turnover?
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: November 23, 1989
Volm: 342(6248)
Page: 347-348

247.
Type: book
Area: technical or scientific
BkAu: Easteal//McLeod//Reed
Ttl2: DNA profiling: principles, pitfalls, and potential
Plac: 5301 Tacony Street, Box 330, Philadelphia, PA. 19137
Publ: Gordon and Breach Science Publishers/Harwood Academic Publishers
Volm: 3-7186-5190-4

248.
Auth: Edman, Jeffrey C.//Evans-Holm, Martha E.//Marich, Jim E.//Ruth, Jerry L.
Affl: Department of Laboratory Medicine, University of California, San Francisco, CA 94143-0134
Ttl1: Rapid DNA fingerprinting using alkaline phosphatase-conjugated oligo-nucleotides
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: July 11, 1988
Volm: 16(13)
Page: 6235

249.
Auth: Edwards, Al//Civitello, Andrew//Hammond, Holly A.//Caskey, C. Thomas
Affl: Institute for Molecular Genetics and Howard Hughes Medical Institute, Baylor College of Medicine, Houston, Tx.
Ttl1: DNA typing and genetic mapping with trimeric and tetrameric tandem repeats
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 49
Page: 746-756

250.
Auth: Edwards, Al//Hammond, Holly A.//Caskey, C. Thomas//Chakraborty, Ranajit
Affl: Howard Hughes Medical Institute; Baylor College of Medicine; University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups
Type: journal article
Area: technical or scientific
Ttl2: Genomics
Date: February 1992
Volm: 12(2)
Page: 241-253

[redacted].
Auth: Eisenberg, Marcia T.//Chimera, Joseph A.
Affl: Roche Biomedical Laboratories, Research Triangle Park, N. C.
Ttl1: The use of AMP-FLPs in identity testing
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 183-186

252.
Auth: Elder, J. K.//Southern, E. M.
Affl: Western General Hospital; Univ. of Edinburgh, Edinburgh, U.K.
Ttl1: Measurement of DNA length by gel electrophoresis II: comparison of methods for relating mobility to fragment length
Type: journal article
Area: technical or scientific
Ttl2: Analytical Biochemistry
Date: 1983
Volm: 128
Page: 227-231

253.
Auth: Elliot, James C.//Fourney, Ronald M.
Affl: Central Forensic Laboratory, Royal Canadian Mounted Police, Ottawa, Ontario, Canada
Ttl1: Evaluation of the amplified VNTR probes COL2A1 and pMCT118 in Canadian caucasians and native Indians
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 197

254.
Auth: Elliott, J. C.//Fourney, R. M.//Budowle, B.//Aubin, R. A.
Affl: Royal Canadian Mounted Police; FBI Academy; Health & Welfare Canada
Ttl1: Quantitative Reproduction of DNA Typing Minisatellites REsolved on Ultrathin Silver-Stained Polyacrylamide Gels with X-ray Duplicating Film
Type: journal article
Area: technical or scientific
Ttl2: Biotechniques
Date: 1993
Volm: 14(5)
Page: 702-704

255.
Auth: Elliott, Janet
Affl: Writer - The Houston Post
Ttl1: Rapist gets maximum in case using DNA match
Type: newspaper
Area: lay press
Ttl2: The Houston Post
Plac: Houston, Texas
Date: November 8, 1988
Page: A-3

256.
Auth: Epplen, [redacted]T.//Ammer, H.//Epplen, C.//Kammerbauer, C.//Mitreiter, R.//Roewer, L.//Schwaiger, W.//Steimle, V.//Zischler, H.//Albert, E.//et al
Affl: Max-Planck-Institute for Psychiatry, Martinsried; Kinderpoliklinik University, Munchen, Germany; et al
Ttl1: Oligonucleotide fingerprinting using simple repeat motifs: a convenient, ubiquitously applicable method to detect hypervariability for multiple purposes
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 50-69

257.
Auth: Epplen, Jorg T.
Affl: Max-Planck-Institut fur Psychiatrie, Am Klopferspitz 18a, Federal Republic of Germany
Ttl1: DNA-fingerprinting-A short note on mutation rates (Reply)
Type: letter to editor
Area: technical or scientific
Ttl2: Human Genetics
Date: 1991
Volm: 87
Page: 633

258.
Auth: Epplen, Jorg T.
Affl: Junior Research Unit, Max-Planck-Institut fur Psychiatrie, Munchen, FRG
Ttl1: On simple repeated GATCA sequences in animal genomes: a critical reappraisal
Type: journal article
Area: technical or scientific
Ttl2: Journal Heredity
Date: November-December 1988
Volm: 79(6)
Page: 409-417

259.
Auth: Epstein, Charles J.
Affl: Editor, American Journal Human Genetics
Ttl1: Editorial: The forsensic application of molecular genetics-the Journal's responsibilities
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 49
Page: 697-698

260.
Auth: Epstein, Charles J.
Affl: Editor - American Journal of Human Genetics
Type: letter of response
Area: technical or scientific
Ttl2: Professional Ethics Report
Plac: 1333 H Street, NW, Washington, DC 20005
Publ: American Association for the Advancement of Science
Date: Spring 1992
Volm: V(2)
Page: 4-5

261.
Auth: Epstein, Edna Selan
Ttl1: The problem with DNA tests: cross-examining the DNA fingerprint
Type: journal article
Area: legal
Ttl2: Illinois Bar Journal
Plac: Illinois
Date: August 1990
Volm: 78(8)
Page: 392(6)

262.
Auth: Erickson, Deborah
Ttl1: Do DNA fingerprints protect the innocent?
Type: journal article
Area: technical or scientific
Ttl2: Scientific American
Date: August 1991
Page: 18

263.
Auth: Erlich, H. A.//Higuchi, R.//von Beroldingen, C.//Blake, E.
Affl: Dept. of Human Genetics; Dept. of Public Health; Forensic Science Assoc.
Ttl1: The use of the polymerase chain reaction for genetic typing in forensic samples
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 93-101

264.
Auth: Erlich, Henry
Affl: Cetus Corporation, Emeryville, CA.
Ttl1: The application of PCR amplification to casework analysis
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 127-128

265.
Type: book
Area: technical or scientific
BkAu: Erlich, Henry A.
Ttl2: PCR Technology - Principles and Applications for DNA Amplification
Plac: 15 East 26th Street, New York, NY 10010
Publ: Stockton Press
Date: 1989
Volm: ISBN 0-333-48948-9
Page: v-x, 1-23

266.
Auth: Erlich, Henry A.//Gelfand, David//Sninsky, John J.
Affl: Dept. of Human Genetics, Cetus Corp, Emeryville, CA.
Ttl1: Recent advances in the polymerase chain reaction
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: June 21, 1991
Volm: 252
Page: 1643-1651

267.
Auth: Erlich, Henry A.//Sheldon, Edward L.//Horn, Glenn
Affl: Department of Human Genetics, Cetus Corp., Emeryville, CA.
Ttl1: HLA typing using DNA probes
Type: journal article
Area: technical or scientific
Ttl2: Biotechnology
Date: November 1986
Volm: 4
Page: 975-979

268.
Auth: Eubanks, William G.
Affl: FBI Laboratory, Washington, D.C.
Ttl1: Cost, implementation and training for DNA analysis
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 189-193

269.
Auth: Eubanks, William G.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Expenses associated with DNA typing methods
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: 1988
Volm: 15 (Supplement)
Page: 10-11

270.
Auth: Eubanks, William G.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: FBI laboratory DNA evidence examination policy
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1988
Volm: 15(4)
Page: 114

271.
Auth: Evett, I. W.//Scranage, J.//Pinchin, R.
Ttl1: Efficient retrieval from DNA databases - based on the 2nd European DNA profilin group collaborative experiment
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: 1992
Volm: 53
Page: 45-50

272.
Auth: Evett, Ian W.
Affl: Home Office Forensic Science Service, Aldermaston, Reading, Berks, UK
Ttl1: Analysis of DNA multilocus profiles in a paternity case in which the child's profile may be partial
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Science Society
Date: 1990
Volm: 30
Page: 293

273.
Auth: Evett, Ian W.
Affl: The Forensic Science Service, Central Research and Support Establishment, Aldermaston, Reading, Berkshire, UK
Ttl1: Trivial error
Type: letter to editor
Area: technical or scientific
Ttl2: Nature
Date: November 14, 1991
Volm: 354
Page: 114

274.
Auth: Evett, Ian W.//Buffery, C.//Willot, G.//Stoney, D.
Affl: Home Office Forensic Science Service, Aldermaston, Reading, Berks, UK
Ttl1: A guide to interpreting single locus profiles of DNA mixtures in forensic cases
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science Society
Date: 1991
Volm: 31
Page: 41

275.
Auth: Evett, [redacted] W.//Gill, Peter
Ttl1: A discussion of the robustness of methods for assessing the evidential value of DNA single locus profiles in crime investigations
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 226-230

276.
Auth: Evett, Ian W.//Pinchin, R.
Affl: Home Office Forensic Science Service, Aldermaston, Reading, Berks, UK
Ttl1: DNA single locus profiles: tests for the robustness of statistical procedures within the context of forensic science
Type: journal article
Area: technical or scientific
Ttl2: International Journal of Legal Medicine
Date: 1991
Volm: 104
Page: 267

277.
Auth: Evett, Ian W.//Werrett, D. J.//Gill, Peter//Buckleton, J. S.
Affl: Home Office Forensic Science Service, Aldermaston, Reading, Berks, UK; Mount Albert Research Centre, Auckland, New Zealand
Ttl1: DNA fingerprinting on trial
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: August 10, 1989
Volm: 340(6233)
Page: 435

278.
Auth: Evett, Ian W.//Werrett, David J.//Buckleton, J. S.
Affl: Home Office Forensic Science Service, Aldermaston, Reading, Brks, UK
Ttl1: Paternity calculations from DNA multilocus profiles
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Science Society
Date: 1989
Volm: 29
Page: 249

279.
Auth: Evett, Ian W.//Werrett, David J.//Smith, A. F. M.
Affl: Home Office Forensic Science Service, Aldermaston, Reading, Brks, UK
Ttl1: Probabilistic analysis of DNA profiles
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Science Society
Date: 1989
Volm: 29
Page: 191

[redacted].
Auth: Executive Committee of the International Society for Forensic Haemogenetics
Affl: International Society for Forensic Haemogenetics
Ttl1: Recommendations of the Society for Forensic Haemogenetics concerning DNA polymorphisms
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: 1989
Volm: 43
Page: 109-111

281.
Auth: Fadda, S.//Pelotti, S.//Pappalardo, G.
Ttl1: A common disinfectant used in condom processing inhibits endonuclease digestion of sperm DNA
Type: journal article
Area: technical or scientific
Ttl2: International Journal Legal Medicine
Date: 1991
Volm: 104
Page: 281

282.
Auth: Farber, C. M.//Georges, M.//DeBock, G.//Verhest, A.//Simon, P.//Verschraegen-Spae, M.//Vassart, G.
Affl: Department of Immunology, Cliniques Universitaires Erasme, Brussels, Belgium
Ttl1: Demonstration of spontaneous XX/XY chimerism by DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: May 1989
Volm: 82(2)
Page: 197-198

283.
Type: book
Area: technical or scientific
BkAu: Farley, Mark A.//Harrington, James J.
Ttl2: Forensic DNA Technology
Plac: 121 South Main Street, Chelsea, Michigan 48118
Publ: Lewis Publishers, Inc.
Date: 1991
Volm: ISBN 0-87371-265-X
Page: i-xvi, 1-235

284.
Auth: Farr, C. J.//Goodfellow, P. N.
Affl: ICRF Laboratories, Lincoln's Inn Fields, London, UK
Ttl1: New variations on the theme
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: November 21, 1991
Volm: 354
Page: 184

285.
Auth: Ferrara, Paul B.
Affl: Virginia Bureau of Forensic Sciences, Richmond, VA.
Ttl1: DNA analysis in the Virginia Bureau of Forensic Science
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 183-187

286.
Auth: Ferrie, Richard M.//Smith, Hilary//Downes, Evelyn A.//McKechnie, Douglas//-/Little, Stephen
Affl: Cellmark Diagnostics, Blacklands Way, Abingdon Business Park, Abingdon,UK
Ttl1: Repeat unit multipriming and hybridization--A novel method for the production of DNA fingerprints using minisatellite probes
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: 1991
Volm: 19(9)
Page: 2505

287.
Auth: Fey, Martin F.
Affl: Institut fur Medizinische Onkologie der Universitat, Inselspital Bern
Ttl1: DNA fingerprints and hypervariable regions: genetic marker with many application potentials in medicine and biology
Type: journal article
Area: technical or scientific
Ttl2: Schweizerische Medizinische Wochenschrift
Date: June 19, 1989
Volm: 119(23)
Page: 815-825

288.
Auth: Fey, Martin F.//Tobler, Andreas
Affl: University of Berne, Tiefenauspital; Central Heamatology Laboratory, Inselspital, Berne, Switzerland
Ttl1: Assessment of DNA 'fingerprinting' as a method for validating the identity of cancer cell lines maintained in long term culture
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: 1991
Volm: 19(12)
Page: 3464

289.
Auth: Fey, Martin F.//Wells, R. A.//Wainscoat, J. S.//Thein, S. L.
Affl: Department of Haematology, John Radcliffe Hospital, Headington, Oxford, UK
Ttl1: Assessment of clonality in gastrointestinal cancer by DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Journal Clinical Investigation
Date: November 1988
Volm: 82(5)
Page: 1532-1537

290.
Auth: Fiori, Angelo//Pascali, Vincenzo L.
Affl: Dept. of Forensic Medicine, Catholic Univ. of Sacred Heart, Roma, Italy
Ttl1: Forensic use of PCR in Italy
Type: letter to editor
Area: technical or scientific
Ttl2: Nature
Date: April 9, 1992
Volm: 356
Page: 471

291.
Auth: Flint, J.//Boyce, A. J.//Martinson, J. J.//Clegg, J. B.
Affl: Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, UK
Ttl1: Population bottlenecks in Polynesia revealed by minisatellites
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: October 1989
Volm: 83(3)
Page: 257-263

292.
Auth: Forbes, K. J.//Bruce, K. D.//Jordens, J. Z.//Ball, A.//Pennington, T. H.
Affl: University Aberdeen, Dept. Med Microbiology, Aberdeen, Scotland
Ttl1: Rapid methods in bacterial DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Journal General Microbiology
Date: September 1991
Volm: 137(P9)
Page: 2051-2058

293.
Auth: Ford, Simon//Thompson, William C.
Affl: University of California at Irvine
Ttl1: A question of identity. Some reasonable doubts about DNA fingerprints
Type: journal article
Area: technical or scientific
Ttl2: The Sciences
Date: Jan/Feb 1990
Page: 37-43

294.
Ttl1: Forensic laboratories report progress in implementation of HLA-DQa typing
Type: newsletter
Area: technical or scientific
BkAu: Perkin Elmer Cetus
Ttl2: Forensic Forum - Updates on PCR in Casework and Research
Date: August 1991
Volm: FF-1
Page: 1-4

295.
Auth: Fornage, Myriam//Chan, L.//Siest, G.//Boerwinkle, Eric
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Allele frequency distribution of the (TG)n(AG)m microsatellite in the apolipoprotein C-II gene.
Type: journal article
Area: technical or scientific
Ttl2: Genomics
Date: January 1992
Volm: 12(1)
Page: 63-68

296.
Auth: Fornage, Myriam//Siest, G.//Boerwinkle, Eric
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Frequency distribution of a (TG)n(AG)m microsatellite reflects the mechanisms of production of new alleles.
Type: journal article
Area: technical or scientific
Ttl2: 8th International Congress of Human Genetics
Plac: Washington, D. C.
Date: October 1991
Volm: 49(4)
Page: 491

297.
Auth: Fourney, R. M.//Shutler, G. G.//Monteith, N.//Bishop, L.//Gaudette, B.//Waye, J. S.
Affl: Central Forensic Laboratory, Royal Canadian Mounted Police, Ottawa, Ontario, Canada
Ttl1: DAN typing in the Royal Canadian Mounted Police
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 137-146

298.
Auth: Fowler, J. Craig S.//Burgoyne, Leigh A.//Scott, Andrew C.//Harding, Harry W. J.
Affl: Dept. of Services and Supply, Adelaide; Flinders University of South Australia, Bedford Park, Australia
Ttl1: Repetitive deoxyribonucleic acid (DNA) and human genome variation - a concise review relevant to forensic biology
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: September 1988
Volm: 33(5)
Page: 1111-1126

299.
Auth: Fowler, J. C. S.//Harrington, C. S.//Dunaiski, V.//Williams, K. E.//Lienert, K.
Affl: State Forensic Science Laboratory, Adelaide, South Australia, Australia
Ttl1: Avoiding errors in PCR analysis: dual typing of the HLA DQa locus by ASO probing and restriction mapping
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 169-172

300.
Auth: Fowler, Susan J.//Gill, Peter//Werrett, David J.//Higgs, Douglas R.
Affl: Central Research Establishment, Home Office Forensic Science Service, Aldermaston, Reading, Berks, UK
Ttl1: Individual specific DNA fingerprints from a hypervariable region probe: alpha-globin 3'HVR
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: June 1988
Volm: 79(2)
Page: 142-146

301.
Auth: Francisco J. Ayala//Bert Black
Affl: University of California, Irvine; Weinberg and Green Law Firm, Baltimore.
Ttl1: Science and the Courts
Type: journal article
Area: technical or scientific
Ttl2: American Scientist
Date: May-June 1993
Volm: 81
Page: 230-239

302.
Type: newslett[redacted]
Area: technical or scientific
BkAu: Frankel, Mark S.
Ttl2: Professional Ethics Report
Plac: 1333 H Street, NW, Washington, DC 20005
Publ: American Association for the Advancement of Science
Date: Spring 1992
Volm: V(2)
Page: 1-8

303.
Auth: Fukushami, Hirofumi//Hasekurea, Hayato//Nagai, Kozo
Affl: Shinshu University School of Medicine, Matsumoto; Tokyo Medical College, Tokyo, Japan
Ttl1: Identification of male bloodstains by dot hybridization of human Y chromosome-specific deoxyribonucleic acid (DNA) probe
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science
Date: May 1988
Volm: 33(3)
Page: 621-627

304.
Auth: Gaensslen, R. E.//Berka, Karen M.//Ruano, Gualberto//Pagliaro, Elaine M.//Lee, Henry C.
Affl: Univ. of New Haven Forensic Science Laboratories, West Haven; Conn. State Police Forensic Science Laboratory, Meriden; Yale University Medical School, New Haven, Conn.
Ttl1: PCR amplification of X and Y chromosome and single- and multi-copy control sequences in DNA from blood, bone and other tissues - PCR determination of sex and species and useful controls for PCR reactions in forensic tests
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 198

305.
Auth: Gaensslen, R. E.//Berka, Karen M.//Grosso, Dina A.//Ruano, Gualberto//Phil, M.//Pagliaro, Elaine M.//Messina, Deborah//Lee Henry C.
Affl: Univ. of New Haven, West Haven, CT.; Connecticut State Police Forensic Science Laboratory, Meriden, CT.
Ttl1: A polymerase chain reaction (PCR) method for sex and species determination with novel controls for deoxyribonucleic acid (DNA) template length
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: January 1992
Volm: 37(1)
Page: 6-20

[redacted].
Auth: Galbraith, David A.//Boag, Peter T.//Gibbs, H. Lisle//White, Brad N.
Affl: Population Genetics Molecular Laboratory, Dept. of Biology, Queen's University, Kingston, Ontario, Canada
Ttl1: Sizing bands on autoradiograms: A study of precision for scoring DNA fingerprints
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 210-220

307.
Auth: Gardner, Laurence
Ttl1: One prosecutor's wretched experience
Type: letter
Area: legal
Ttl2: Los Angeles Daily Journal
Plac: Maine
Date: March 14, 1990
Volm: 103(53)
Page: 6 (col 3)

308.
Auth: Garfield, Eugene
Ttl1: DNA fingerprinting: a powerful law-enforcement tool with serious social implications
Type: commentary
Area: technical or scientific
Ttl2: The Scientist
Date: May 29, 1989
Page: 10

309.
Auth: Gasparini, P.//Martinelli, G. Trabetti ,. E. /Ambrosetti, A.//Benedetti, F.//Pignatti, P. R.
Affl: Istituto di Scienze Biologiche, Universita di Verona, Italia
Ttl1: Bone marrow transplantation monitoring by DNA analysis
Type: journal article
Area: technical or scientific
Ttl2: Bone Marrow Transplant
Date: December 1989
Volm: 4 (Supplement 4)
Page: 157-159

310.
Auth: Gasparine, P.//Trabetti, E.//Savoia, A.//Marigo, M.//Pignatti, P. F.
Affl: Istituto di Scienze Biologiche, Universita di Verona, Italia
Ttl1: Frequency distribution of the allelles of several variable number of tandem repeat DNA polymorphisms in the Italian population
Type: journal article
Area: technical or scientific
Ttl2: Human Heredity
Date: 1990
Volm: 40(2)
Page: 61-68

311.
Auth: Gatti, R[illegible]hard A.//Nakamura, Yusuke//Nussmeier, Marianne//Susi, Ellen//Shan, Wei//Grody, Wayne W.
Affl: Department of Pathology, UCLA School of Medicine, 90024
Ttl1: Informativeness of VNTR genetic markers for detecting chimerism after bone marrow transplantation
Type: journal article
Area: technical or scientific
Ttl2: Disease Markers
Date: April-June 1989
Volm: 7(2)
Page: 105-112

312.
Auth: Gaudette, Barry D.
Affl: Royal Canadian Mounted Police, Ottawa, Ontario, Canada
Ttl1: Forensic DNA analysis in the Royal Canadian Mounted Police
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 229-232

313.
Auth: Gazit, E.//Kelt, R.//Shani, S.//Mozer, M.
Affl: Tissue Typing Laboratory, Chaim Sheba Medical Center, Tel Hashomer
Ttl1: DNA fingerprinting in paternity testing
Type: journal article
Area: technical or scientific
Ttl2: Harefuah
Date: February 1, 1990
Volm: 118(3)
Page: 129-133

314.
Auth: Gazit, Esther//Gazit, Ephraim
Affl: Goldschleger School of Dental Medicine, Tel Aviv University, Israel
Ttl1: DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Israel Journal Medical Science
Date: March 1990
Volm: 26(3)
Page: 158-162

315.
Auth: Geberth, Lt Cmdr (Ret). NYPD, Vernon J.
Affl: President of P.H.I. Investigatiave Consultants, Inc.
Ttl1: DNA print identification test provides crucial evidence in lust murder case
Type: journal article
Area: technical or scientific
Ttl2: Law and Order
Date: July 1988
Page: 22-28

316.
Auth: Geisser, Seymour
Affl: School of Statistics, Univ. of Minnesota, Twin City Campus
Type: letter of response
Area: technical or scientific
Ttl2: Professional Ethics Report
Plac: 1333 H Street, NW, Washington, DC 20005
Publ: American Association for the Advancement of Science
Date: Spring 1992
Volm: V(2)
Page: 5-6

317.
Auth: Geisser, Seymour
Affl: University of Minnesota, Minneapolis, MN. 55455
Ttl1: Some remarks of DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Chance: New Directions for Statistics and Computing
Date: 1990
Volm: 3(3)
Page: 8-9

318.
Auth: Geisser, Seymour
Affl: School of Statistics, University of Minnesota, Mineapolis, MN.
Ttl1: Some statistical issues in Medicine and Forensics
Type: journal article
Area: technical or scientific
Ttl2: Journal of the American Statistical Association
Date: September 1992
Volm: 87(419)
Page: 607-614

319.
Auth: Geisser, Seymour//Johnson, Wesley
Affl: School of Statistics, Univ. of Minnesota, Minneapolis & Univ. of CA. Davis
Ttl1: Testing independence of fragment lenghts within VNTR loci
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: November 1993
Volm: 53(5)
Page: 1103-1106

320.
Auth: Georges, M.//Lequarre, A. S.//Castelli, M.//Hanset, R.//Vassart, G.
Affl: Chaire de Genetique, Faculte de Medecine Veterinaire, University de Liege, Bruxelles, Belgium
Ttl1: DNA fingerprinting in domestic animals using four different minisatellite probes
Type: journal article
Area: technical or scientific
Ttl2: Cytogenetics and Cell Genetics
Date: 1988
Volm: 47(3)
Page: 127-131

321.
Auth: Georges, Michel//Cochaux, Pascale//Lequarre, Anne Sophie//Young, Michael W.//Vassart, Gilbert
Affl: Chaire de Genetique, Faculte de Medecine Veterinaire, University de Liege, Belgium
Ttl1: DNA fingerprinting in man using a mouse probe related to part of the Drosophila Per gene
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: September 11, 1987
Volm: 15(17)
Page: 7193

322.
Auth: Georges, Michel//Hilbert, Pascale//Lequarre, Anne Sophie//Leclerc, Veronique//-//Hanset, Roger//Vassart, Gilbert
Affl: Department of Genetics, Faculty of Veterinary Medicine, University of Liege, Bruxelles, Belgium
Ttl1: Use of DNA bar codes to resolve a canine paternity dispute
Type: journal article
Area: technical or scientific
Ttl2: Journal American Veterinary Medicine Association
Date: November 1, 1988
Volm: 193(9)
Page: 1095-1098

323.
Auth: Georges, Michel//Lathrop, Mark//Hilbert, Pascale//Marcotte, Anne//Schwers, Anne//Swillens, Stephane//Vassart, Gilbert//Hanset, Roger
Affl: Faculte de Medecine Veterinaire, Universite de Liege, Belgium; CEPH, Paris
Ttl1: On the use of DNA fingerprints for linkage studies in cattle
Type: journal article
Area: technical or scientific
Ttl2: Genomics
Date: March 1990
Volm: 6(3)
Page: 461-474

324.
Auth: Gerard, Catherine//Christophe, Daniel//Compere, Thierry//Vassart, Gilbert
Affl: Institut de Recherche Interdisciplinaire, Faculte de Medecine University Libre de Bruxelles, Belgium
Ttl1: The poly (purine) poly (pyrimidine) sequence in the 5' end of the thyroglobulin gene used as a probe, identifies a DNA fingerprint in man
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: July 25, 1990
Volm: 18(14)
Page: 4297

325.
Auth: Giannelli, Paul C.
Ttl1: Criminal discovery, scientific evidence and DNA
Type: journal article
Area: legal
Ttl2: Vand. Law Review
Date: May 1991
Volm: 44
Page: 791

326.
Auth: Gilbert, Dennis A.//Lehman, Niles//O'Brien, Stephen J.//Wayne, Robert K.
Affl: Biological Carcinogenesis and Development Program, Program Resources Incorporated, NCI-FCRF, Maryland 21701
Ttl1: Genetic fingerprinting reflects population differentiation in the California Channel Island fox
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: April 19, 1990
Volm: 344(6268)
Page: 764-767

327.
Auth: Gilbert, Dennis A.//Reid, Yvonne A.//Gail, Mitchell H.//Pee, David//White, Christine//Hay, Robert J. /. O'Brien, Stephen J.
Affl: Laboratory of Viral Carcinogenesis, National Cancer Institute, Frederick, MD 21701
Ttl1: Application of DNA fingerprints for cell-line individualization
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: September 1990
Volm: 47(3)
Page: 499-514

328.
Auth: Gill, P.//Evett, I. W.//Woodroffe, S.//Lygo, J. E.//Millican, E.//Webster, M.
Ttl1: Databases, quality control and interpretation of DNA profiling in the Home Office Forensic Science Service
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 204-209

329.
Auth: Gill, P[illegible]
Affl: Home Office Forensic Science Service, Aldermaston, Reading, Berkshire UK
Ttl1: A new method for sex determination of forensic samples using a recombinant DNA probe
Type: journal article
Area: technical or scientific
Ttl2: Electrophoresis
Date: 1987
Volm: 8
Page: 35-38

330.
Auth: Gill, Peter//Jeffreys, Alec J.//Werrett, David J.
Affl: Central Research Establishment, Home Office Forensic Science Service, Aldermaston, Reading, Berkshire, UK
Ttl1: Forensic application of DNA 'fingerprints'
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: December 12-18, 1985
Volm: 318(6046)
Page: 577-579

331.
Auth: Gill, Peter//Lygo, Joan E.//Fowler, Susan J.//Werrett, David J.
Affl: Home Office Forensic Science Service, Aldermaston, Reading, Berkshire UK
Ttl1: An evaluation of DNA fingerprinting for forensic purposes
Type: journal article
Area: technical or scientific
Ttl2: Electrophoresis
Date: 1987
Volm: 8
Page: 38

332.
Auth: Gill, Peter//Werrett, David J.
Affl: Central Research Establishment, Home Office Forensic Science Service, Reading, Berkshire, UK
Ttl1: Exclusion of a man charged with murder by DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: October-November 1987
Volm: 35(2-3)
Page: 145-148

333.
Auth: Gill, Peter//Woodroffe, Susan//Lygo, Joan E.//Millican, Emma S.
Affl: Home Office Forensic Science Service, Aldermaston, Reading, Berks, UK
Ttl1: Population genetics of four hypervariable loci
Type: journal article
Area: technical or scientific
Ttl2: International Journal of Legal Medicine
Date: 1991
Volm: 104
Page: 221

334.
Auth: Gill, Peter//Woodroffe, S. Bar ,. W.//Brinkman, D.//Carracedo, A.//Eriksen, B.//Jones, S.//Kloostermann, A. D.//et al
Ttl1: A report of an international collaborative experiment to demonstrate the uniformity obtainable using DNA profiling techniques
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: 1992
Volm: 53
Page: 29-43

335.
Auth: Giusti, Alan//Baird, Michael//Pasquale, Sam//Balazs, Ivan//Glassberg, Jeffrey
Affl: Lifecodes Corp., Elmsford, NY
Ttl1: Application of deoxyribonucleic acid (DNA) polymorphisms to the analysis of DNA recovered from sperm
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: April 1986
Volm: 31(2)
Page: 409-417

336.
Auth: Giusti, Alan//Baird, Michael//Pasquale, Sam//Balazs, Ivan//Glassberg, Jeffrey
Affl: Lifecodes Corp, Elmsford, N.Y.; Rutgers Medical School, New Brunswick, NJ
Ttl1: Application of deoxyribonucleic acid (DNA) polymorphisms to the analysis of DNA recovered from sperm
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: April 1986
Volm: 31(2)
Page: 409-417

337.
Auth: Gjertson, David W.//Mickey, M. Ray//Hopfield, Judy//Takenouchi, Toshinao/-/Terasaki, Paul
Affl: University of California at Los Angeles Tissue Typing Laboratory, Los Angeles, CA.
Ttl1: Calculation of probability of paternity using DNA sequences
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: December 1988
Volm: 48(6)
Page: 860-869

338.
Auth: Gold, Stephen
Ttl1: DNA explosion
Type: journal article
Area: legal
Ttl2: New Law Journal
Plac: Great Britain
Date: November 27, 1987
Volm: 137(6333)
Page: 1104(1)

339.
Auth: Goldberg, Stephanie B.
Ttl1: A new day for DNA? Despite widespread acceptance, admissibility standards vary
Type: journal article
Area: technical or scientific
Ttl2: Trends in the Law
Date: April 1992
Page: 84-85

340.
Auth: Gonzales, Juan Martinez
Affl: Attorney and Counselor at Law, Beeville, Texas
Ttl1: Attacking forensic DNA profiling evidence for lack of validation
Type: journal article
Area: technical or scientific
Ttl2: The National Lawyers ~ Guild Practitioner
Date: Spring, 1989
Volm: 46(2)
Page: 51-55

341.
Auth: Gordon, Judith M.
Ttl1: DNA identification tests--on the way toward judicial acceptance
Type: journal article
Area: legal
Ttl2: J. Suffolk Acad. Law
Date: 1989
Volm: 6
Page: 1

342.
Auth: Green, Philip
Affl: Genetic Department, Washington University School of Medicine, St. Louis
Ttl1: Population genetic issues in DNA fingerprinting
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal of Human Genetics
Date: 1992
Volm: 50
Page: 440-441

343.
Auth: Green, Philip//Lander, Eric S.
Affl: Washington Univ. School of Medicine, St. Louis, MO.; Massachusetts Inst. of Technology, Cambridge, MA.
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: August 30, 1991
Volm: 253
Page: 1038-1039

344.
Auth: Greenberg, Jonathan
Ttl1: DNA fingerprinting: a guide for defense counsel
Type: journal article
Area: legal
Ttl2: Army Law
Date: November 1989
Volm: 27-50-203
Page: 16

345.
Auth: Grimberg, J.//Nawoschik, S.//Belluscio, L.//McKee, R.//Turck, A.//Eisengerg, A.
Affl: Lifecodes Corporation, Valhalla, NY 10595
Ttl1: A simple and efficient non-organic procedure for the isolation of genomic DNA from blood
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: October 25, 1989
Volm: 17(20)
Page: 8390

346.
Auth: Grody, W. W.//Gatti, R. A.//Naeim, F.
Affl: Department of Pathology, UCLA School of Medicine
Ttl1: Diagnostic molecular pathology
Type: journal article
Area: technical or scientific
Ttl2: Modern Pathology
Date: November 1989
Volm: 2(6)
Page: 553-568

347.
Auth: Groner, Peter
Affl: Writer - Chicago Tribune
Ttl1: DNA testing provides better breeding of endangered species
Type: newspaper
Area: lay press
Ttl2: Houston Chronicle
Plac: Houston, Texas
Date: August 7, 1989
Page: 8B

348.
Auth: Grossma[redacted]. I.
Affl: Wayne State University School of Medicine
Ttl1: Gel electrophoresis of DNA
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 37-46

349.
Auth: Grover, Adrienne M.
Ttl1: A new twist in the double helix: admissibility of DNA fingerprinting in California
Type: journal article
Area: legal
Ttl2: Santa Clara Computer and High-Technology Law Journal
Plac: California
Date: June 1989
Volm: 5(2)
Page: 469-496

350.
Auth: Gustafson, Sarah//Proper, Jacqueline A.//Bowie, E. J. Walter//Sommer, Steve S.
Affl: Dept. of Biocheminstry and Molecular Biology, Guggenheim; Rochester, MN.
Ttl1: Parameters affecting the yield of DNA from human blood
Type: journal article
Area: technical or scientific
Ttl2: Analytical Biochemistry
Date: 1987
Volm: 165
Page: 294-299

351.
Auth: Gyllensten, Ulf B.//Erlich, Henry A.
Affl: Dept. of Human Genetics, Cetus Corporation, Emeryville, Ca.
Ttl1: Ancient roots for polymorphism at the HLA-DQa locus in primates
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy Science USA
Date: December 1989
Volm: 86
Page: 9986-9990

352.
Auth: Gyllensten, Ulf B.//Erlich, Henry A.
Affl: University of Uppsala, Sweden; Cetus Corp. Emeryville, CA
Ttl1: Evolution of HLA class-II polymorphism in primates: the DQa locus
Type: journal article
Area: technical or scientific
Ttl2: Immunological Research
Date: 1990
Volm: 9
Page: 223-233

353.
Auth: Gyllensten, Ulf B.//Erlich, Henry A.
Affl: Cetus Corporation, Emeryville; University of California, Berkeley; CA
Ttl1: Generation of single-stranded DNA by the polymerase chain reaction and its application to direct sequencing of the HLA-DQa locus
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy Science, USA
Date: October 1988
Volm: 85
Page: 7652-7656

354.
Auth: Gyllensten, Ulf B.//Jakobsson, Sven//Temrin, Hans//Wilson, Allan C.
Affl: Department of Biochemistry, University of California, Berkeley 94720
Ttl1: Nucleotide sequence and genomic organization of bird minisatellites
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: March 25, 1989
Volm: 17(6)
Page: 2203-2214

355.
Auth: H, S. J.
Ttl1: DNA fingerprint ruling
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: November 15-19, 1987
Volm: 330(6145)
Page: 197

356.
Auth: Haberfeld, A.//Hillel, J.
Ttl1: Development of DNA fingerprint probes: an approach and its application
Type: journal article
Area: technical or scientific
Ttl2: Animal Biotechnology
Date: 1991
Volm: 2
Page: 61-73

357.
Auth: Hagelberg, Erika//Gray, Ian C.//Jeffreys, Alec J.
Affl: University of Oxford; University of Leicester, UK
Ttl1: Identification of the skeletal remains of a murder victim by DNA analysis
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: August 1, 1991
Volm: 352
Page: 427-429

358.
Auth: Hager, Philip
Affl: Times legal affairs writer
Ttl1: DNA on trial as evidence
Type: newspaper article
Area: lay press
Ttl2: Los Angeles Times
Plac: Los Angeles, California
Date: Wednesday, March 27, 1991
Page: 1, A15

359.
Auth: Hagerman, Paul J.
Affl: University of Colorado Health Sciences Center, Denver, CO
Ttl1: DNA typing in the forensic arena
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1990
Volm: 47
Page: 876-877

360.
Auth: Haglund, William D.//Reay, Donald T.//Tepper, Shelly L.
Affl: King County Medical Examiner's Office, Department of Public Health, Seattle, WA
Ttl1: Identification of decomposed human remains by deoxyribonucleic acid (DNA) profiling
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: May 1990
Volm: 35(3)
Page: 724-729

361.
Auth: Hall, Andrew
Ttl1: DNA fingerprints - black box or black hole?
Type: journal article
Area: legal
Ttl2: New Law Journal
Plac: Great Britain
Date: February 16, 1990
Volm: 140(6443)
Page: 203(3)

362.
Auth: Hanner, Jane E.
Ttl1: DNA fingerprinting: evidence of the future
Type: journal article
Area: legal
Ttl2: Kentucky Law Journal
Plac: United States
Date: Winter 1991
Volm: 79(2)
Page: 415-438

363.
Auth: Hanotte, O.//Burke, T.//Armour, J. A. L.//Jeffreys, A. J.
Affl: Univ. of Leicester, UK; Universite de Mons-Hainaut, Service de Biochimie Moleculaire, Mons, Belgium
Ttl1: Cloning, characterization and evolution of Indian Peafowl Pavo cristatus minisatellite loci
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 193-216

364.
Auth: Hanson, Eric
Affl: Houston Chronicle writer
Ttl1: Police say DNA lab will cut down costs - HPD to test suspect genetics in-house
Type: newspaper
Area: lay press
Ttl2: Houston Chronicle
Publ: Houston, Texas
Date: January 3, 1991
Page: 20A

365.
Auth: Harding, H. W.//Ross, A. M.//Fowler, J. C.//Miller. I.
Affl: Foresnic Science Centre, Adelaide; Forensic Science Technology Int'l Pty. Ltd., The Levels, South Australia
Ttl1: Tracktel: an electrophoretic pattern image processing system for the forensic laboratory
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-3, 1989
Page: 247

366.
Auth: Harding, H. W.//Ross, A. M.//Fowler, J. C.//McInnes, J. L.
Affl: Forensic Science Centre; University of Adelaide, Adelaide, South Australia
Ttl1: The use of Polystat 3 probe for forensic testing
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 213-214

367.
Auth: Harding, [redacted]//Gebeyehu, G.//Beebe, R.//Simms, D.//Klevan, L.
Affl: Life Technologies Inc., Gaithersburg, Maryland
Ttl1: Rapid isolation of DNA from body fluids using a novel nucleic acid capture reagent
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 225-227

368.
Auth: Hareuveni, M.//Tsarfaty, I.//Zaretsky, J.//Kotkes, P.//Horev, J.//Zrihan, S.//Weiss, M.//Green, S.//Lathe, R.//Keydar, I.//et al
Affl: Department of Microbiology, Faculty of Life Sciences, Tel Aviv University, Israel
Ttl1: A transcribed gene, containing a variable number of tandem repeats, codes for a human epithelial tumor antigen. cDNA cloning, expression of the transfected gene and over-expression in breast cancer tissue
Type: journal article
Area: technical or scientific
Ttl2: European Journal Biochemistry
Date: May 20, 1990
Volm: 189(3)
Page: 475-486

369.
Auth: Harmon, R.
Affl: Alameda County District Attorney's Office, Oakland, CA.
Ttl1: General admissibility considerations for DNA typing evidence: let's learn from the past and let the scientists decide this time around
Type: book chapter
Area: technical or scientific
BkAu: Farley, Mark A.//Harrington, James J.
Ttl2: Forensic DNA Technology
Date: 1991
Page: 153-180

370.
Auth: Harmon, Rockne
Affl: Alameda County District Attorney's Office, Oakland, CA.
Ttl1: DNA fingerprinting critics titillate rather than inform
Type: journal article
Area: legal
Ttl2: Los Angeles Daily Journal
Plac: United States
Date: March 14, 1990
Volm: 103(53)
Page: 6 (col 4)

371.
Auth: Harmon, Rockne P.
Affl: Alameda County District Attorney's Office, Oakland CA.
Ttl1: The Frye test: considerations for DNA identification techniques
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 89-107

372.
Auth: Harmon, Rockne P.
Affl: Alameda County District Attorney's Office, Oakland
Ttl1: Please leave law to the lawyers
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 49
Page: 891-892

373.
Auth: Harmon, Rockne
Affl: Alameda County District Attorney's Office, Oakland, CA.
Ttl1: Recent experiences aside, how should future genetic identification techniques be reviewed for admissibility?
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 190

374.
Auth: Hartl, Daniel L.
Affl: Washington University School of Medicine, St. Louis, Missouri
Ttl1: Expert's report in the case of United States vs Yee, et al
Type: unpublished document
Area: technical or scientific
Case: United States vs. Yee et al.
Page: 1-22

375.
Auth: Hayward, Nicholas//Chen, Philip//Nancarrow, Derek//Kearsley, John//Smith, Peter//Kidson, Chev//Ellem, Kay
Affl: Queensland Institute of Medical Research, Herston, Brisbane, Australia
Ttl1: Detection of somatic mutations in tumours of diverse types by DNA fingerprinting with M13 phage DNA
Type: journal article
Area: technical or scientific
Ttl2: International Journal Cancer
Date: April 15, 1990
Volm: 45(4)
Page: 687-690

376.
Auth: He, Zho[redacted]ian//Jiang, Xean Hua//Lu, Shi Hui//Wang, Guo Lin//Zhu, Yu Wen/-/Shen, Yan//Gao, Qing Sheng//Liu, Jing Zhong//Wu, Guan Yu
Affl: Liaoning Criminal Scientific and Technical Research Institute, Shenyang, Liaoning; Chinese Academy of Medical Science, Beijing People's Republic of China
Ttl1: A study of sex identification of trace, dried bloodstains using a Y chromosome-specific deoxyribonucleic acid (DNA) probe
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: March 1989
Volm: 34(2)
Page: 346-351

377.
Auth: Heery, D. M.//Gannon, F.//Powell, R.
Affl: Dept. of Microbiology, University College, Galway, Republic of Ireland
Ttl1: A simple method for subcloning DNA fragments from gel slices
Type: journal article
Area: technical or scientific
Ttl2: Trends in Genetics
Date: June 1990
Volm: 6(6)
Page: 173

378.
Auth: Hegele, R. A.
Affl: Department of Human Genetics, Howard Hughes Medical Institute, University of Utah, Salt Lake City
Ttl1: Molecular forensics: applications, implications and limitations
Type: journal article
Area: technical or scientific
Ttl2: Cancer Med Association Journal
Date: October 1, 1989
Volm: 141(7)
Page: 668-672

379.
Auth: Helminen, Paivi//Ehnholm, Christian//Lokki, Marja Liisa//Jeffreys, Alec/-/Peltonen, Leena
Affl: Univ. of Helsinki, Finland; Univ. of Leicester, UK
Ttl1: Application of DNA "fingerprints" to paternity determinations
Type: journal article
Area: technical or scientific
Ttl2: The Lancet
Date: March 12, 1988
Volm: 1(8585)
Page: 574-576

380.
Auth: Helminen, Paivi//Johnsson, Vivian//Ehnholm. Christian//Peltonen, Leena
Affl: Lab. of Molecular Genetics; Lab. of Forensic Serology, National Public Health Institute, Mannerheimintie 166, SF-00300 Helsinki, Finland
Ttl1: Proving paternity of children with deceased fathers
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: 1991
Volm: 87
Page: 657-660

381.
Auth: Helmuth, Rhea//Fildes, Nicola//Blake, Edward//Luce, Michael C.//Chimera, J.//Madej, Roberta//Gorodezky, C.//Stoneking, Mark//Schmill, N.//et al
Affl: Department of Human Genetics, Cetus Corproation, Emeryville, CA 94608
Ttl1: HLA-DQ alpha allele and genotype frequencies in various human populations, determined by using enzymatic amplification and oligonucleotide probes
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: September 1990
Volm: 47(3)
Page: 515-523

382.
Auth: Henke, L.//Cheef, S.//Zakrzewska, M.//Henke, J.
Ttl1: BamHI polymorphism of locus D2S44 in a West German population as revealed by VNTR probe YNH24
Type: journal article
Area: technical or scientific
Ttl2: International Journal of Legal Medicine
Date: 1990
Volm: 104
Page: 33

383.
Auth: Henke, L.//Cleef, S.//Zakrzewska, M.//Henke, J.
Affl: Institut f. Blutgrupperforschung, Otto-Hahn-Str. Dusseldorf, Germany
Ttl1: Population genetic data determined for five different single locus mini-satellite probes
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 144-153

384.
Auth: Hernandez, J. L.//Weir, B. S.
Affl: Dept. of Statistics, North Carolina State University, Raleigh, N.C.
Ttl1: A disequilibrium coefficient approach to Hardy-Weinberg testing
Type: journal article
Area: technical or scientific
Ttl2: Biometrics
Date: March 1989
Volm: 45
Page: 53-70

385.
Auth: Herrin J[illegible] G.//Cotton, R. W.//Corey, A. C.//David, K.//McNeil, T. A.//Rubenstein, K. R.//Tonelli, L. A.//Garner, D. D.
Affl: Cellmark Diagnostics, Germantown, Maryland
Ttl1: Case examples using differential extraction procedures
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 233

386.
Auth: Herrin, George Jr
Affl: Division of Forensic Sciences, Georgia Bureau of Investigation, Decatur
Ttl1: Probability of matching RFLP patterns from unrelated individuals
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: June, 1993
Volm: 52
Page: 491-497

387.
Auth: Herrin Jr., George//Forman, Lisa//Garner, Daniel D.
Affl: Cellmark Diagnostics Division of Imperial Chemical Industries Ltd., Germantown, Maryland
Ttl1: The use of Jeffreys' mutlilocus and single locus DNA probes in forensic analysis
Type: book chapter
Area: technical or scientific
BkAu: Lee, Henry C.//Gaensslen, R. E.
Ttl2: DNA and Other Polymorphisms in Forensic Science
Date: 1990
Page: 45-60

388.
Auth: Hibbs, Mary
Ttl1: Applications of DNA fingerprinting - truth will out
Type: journal article
Area: legal
Ttl2: New Law Journal
Plac: Great Britain
Date: May 5, 1989
Volm: 139(6406)
Page: 619(3)

389.
Auth: Hicks, John W.
Affl: Assistant Director, FBI, Quantico, VA.
Ttl1: DNA typing: a unique weapon against crime
Type: journal article
Area: technical or scientific
Ttl2: The Scientist
Date: January 23, 1989
Page: 12-13

390.
Auth: Hicks, John W.
Affl: FBI Laboratory, Washington, D.C.
Ttl1: FBI program for the forensic application of DNA technology
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 209-212

391.
Auth: Hicks, John H.
Affl: US Department of Justice, FBI, Washington, D.C. 20535
Ttl1: FBI's case for genetics
Type: letter to editor
Area: technical or scientific
Ttl2: Nature
Date: June 4, 1992
Volm: 357
Page: 355

392.
Auth: Hicks, John W.
Affl: Assistant Director-FBI Laboratory
Ttl1: Message from the Assistant Director in charge of the FBI laboratory
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: v-vii

393.
Type: book
Area: technical or scientific
BkAu: Hicks, John W. (Assistance Director in Charge)
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Plac: Washington, D.C. 20535
Publ: Superintendent of Documents, U. S. Government Printing Office
Date: June 19-23, 1989
Volm: ISBN 0-932115-10-1
Page: i-ix, 1-282

394.
Auth: Hicks, John W.
Affl: Assistant Director in Charge, Laboratory Division, FBI, Quantico, VA.
Type: letter of response
Area: technical or scientific
Ttl2: Professional Ethics Report
Plac: 1333 H Street, NW, Washington, DC 20005
Publ: American Association for the Advancement of Science
Date: Spring 1992
Volm: V(2)
Page: 6-7

395.
Auth: Hicks, John W.
Affl: FBI Laboratory
Ttl1: Statement of John W. Hicks, Deputy Assistant Director, for presentation before the House Committee
Type: unpublished document
Area: technical or scientific
Plac: House Committee on the Judiciary Subcommittee on Civil and Constitutional Rights, Honorable Don Edwards, Chairman
Date: March 22, 1989
Page: 1-13

396.
Auth: Hicks, John W.
Affl: FBI Laboratory, Washington, D.C.
Ttl1: Summary of the International Symposium on the Forensic Aspects of DNA Analysis
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceeds of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 201-204

397.
Auth: Hicks, John W.
Affl: Laboratory Division, FBI, Washington, D.C.
Ttl1: Understanding the DNA Proficiency Testing Act of 1991
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: January 1991
Volm: 18(1)
Page: 1-8

398.
Auth: Higashiguchi, T.//Serikawa, T.//Kuramoto, T.//Mori, M.//Yamada, J.
Ttl1: Identification of inbred strains of rats by DNA fingerprints using enchanced chemiluminescence
Type: journal article
Area: technical or scientific
Ttl2: Transplant. Proc.
Date: 1990
Volm: 22
Page: 2564-2565

399.
Auth: Higuchi, Russell
Affl: Department of Human Genetics, Cetus Corporation, Emeryville, CA.
Ttl1: Human error in forensic DNA typing
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 48
Page: 1215-1216

400.
Auth: Higuchi, Russell//Blake, Edward T.
Affl: Cetus Corporation, Emeryville, CA.; Forensic Sciences Assoc., Richmond, CA.
Ttl1: Applications of the polymerase chain reaction in forensic science
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 265-281

401.
Auth: Higuchi, Russell//von Beroldingen, Cecilia H.//Sensabaugh, George F.//Erlich, Henry A.
Affl: Dept. of Human Genetics, Cetus Corp., Emeryville; Forensic Sciences Program, University of California, Berkeley, CA.
Ttl1: DNA typing from single hairs
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: April 1988
Volm: 332(6164)
Page: 543-546

402.
Auth: Hill, William G.
Affl: Univ. of Edinburgh, West Mains Road, Edinburgh, UK
Ttl1: DNA fingerprint analysis in immigration test-cases
Type: letter to editor
Area: technical or scientific
Ttl2: Nature
Date: July 17-23, 1986
Volm: 322(6076)
Page: 290-291

403.
Auth: Hill, William G.
Affl: University of Edinburgh, West Mains Road, Edinburgh, UK
Ttl1: DNA fingerprints applied to animal and bird population
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: May 14-20, 1987
Volm: 327(6118)
Page: 98-99

404.
Auth: Hillel, J.//Gal O.//Schaap, T.//Haberfeld, A.//Plotsky, Y.//Marks, H.//Siegel, P. B.//Dunnington, E. A.//Cahaner, A.
Affl: The Hebrew University; Hebrew University Medical Center; University of Georgia; Virginia Polytechnic Institute and State University
Ttl1: Genetic factors accountable for line specific DNA fingerprint bands in quail
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 263-273

405.
Auth: Hillel, [redacted] Plotzy ,. Y.//Haberfeld, A.//Lavi, U.//Cahaner, A.//Jeffreys, A. J.
Affl: Department of Genetics, Faculty of Agriculture, Hebrew University of Jerusalem, Rehovot, Israel
Ttl1: DNA fingerprints of poultry
Type: journal article
Area: technical or scientific
Ttl2: Animal Genetics
Date: 1989
Volm: 20(2)
Page: 145-155

406.
Auth: Hillel, J.//Schaap, T.//Haberfeld, A.//Jeffreys, A. J.//Plotzky Y.//Cahaner, A.//Lavi, U.
Affl: Department of Genetics, Faculty of Agriculture, Hebrew University of Jerusalem, Rehovot, Israel
Ttl1: DNA fingerprints applied to gene introgression in breeding programs
Type: journal article
Area: technical or scientific
Ttl2: Genetics
Date: March 1990
Volm: 124(3)
Page: 783-789

407.
Auth: Hochmeister, Manfred//Borer, Urs
Affl: Department of Forensic Medicine, Bern, Switzerland
Ttl1: Practical applications of PCR-based typing of DNA
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 166-168

408.
Auth: Hochmeister, Manfred N.//Budowle, Bruce//Baechtel, F. Samuel
Affl: Forensic Science Research and Training Section, FBI Acedemy, Quantico, VA
Ttl1: Effects of presumptive test reagents on the ability to obtain RFLP patterns from human blood and semen stains
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: May 1991
Volm: 36(3)
Page: 656-661

409.
Auth: Hochmeister, Manfred N.//Budowle, Bruce//Borer, Urs V.//Eggmann, [redacted]rs//Comey, Catherine T.//Dirnhofer, Richard
Affl: Univ. of Bern, Bern, Switzerland; FBI Academy, Quantico, VA.
Ttl1: Typing of deoxyribonucleic acid (DNA) extracted from compact bone from human remains
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: November 1991
Volm: 36(6)
Page: 1649-1661

410.
Auth: Hoeffel, Janet C.
Ttl1: The dark side of DNA profiling: unreliable scientific evidence meets the criminal defendant
Type: journal article
Area: legal
Ttl2: Stan. Law Review
Date: January 1990
Volm: 42
Page: 465

411.
Auth: Hoeffel, Janet C.
Ttl1: DNA fingerprinting flaws belie claims of infallibility
Type: journal article
Area: legal
Ttl2: Los Angeles Daily Journal
Plac: United States
Date: March 14, 1990
Volm: 103(53)
Page: 6 (col 3)

412.
Auth: Hoelzel, A. Rus//Amos, William
Affl: Dept. of Genetics, Univ. of Cambridge, Downing Street, Cambridge, UK
Ttl1: DNA fingerprinting and 'scientific' whaling
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: May 26, 1988
Volm: 333(6171)
Page: 305

413.
Auth: Holm, Thomas//Terry, Christi//Georges, Michel
Affl: Genmark, Inc., Salt Lake, UT
Ttl1: In vitro amplification of a set of VNTR loci for forensic science
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 187-189

414.
Auth: Holtz, [redacted]/Olek, K.//Higuchi, M.
Ttl1: Forensic application of DNA fingerprints
Type: journal article
Area: technical or scientific
Ttl2: Beitrage Gerichtl Med
Date: 1987
Volm: 45
Page: 5-10

415.
Auth: Honma, M.//Ishiyama, I.
Affl: Department of Forensic Medicine, Faculty of Medicine, University of Toyko, Japan
Ttl1: DNA fingerprints: its importance in forensic medicine (1) Application to paternity testing by minisatellite DNA probes
Type: journal article
Area: technical or scientific
Ttl2: Nippon Hoigaku Zasshi
Date: June 1987
Volm: 41(3)
Page: 236-241

416.
Auth: Honma, M.//Ishiyama, I.
Affl: Department of Forensic Medicine, Faculty of Medicine, University of Toyko, Japan
Ttl1: Variability of DNA fingerprint in a Japanese population
Type: journal article
Area: technical or scientific
Ttl2: Nippon Hoigaku Zasshi
Date: April 1989
Volm: 43(2)
Page: 128-133

417.
Auth: Honma, Masamitsu//Ishiyama, Ikuo
Affl: Depart. of Forensic Medicine, Faculty of Medicine, Univ. of Toyko, Japan
Ttl1: Application of DNA fingerprinting to parentage and extended family relationship testing
Type: journal article
Area: technical or scientific
Ttl2: Human Heredity
Date: 1990
Volm: 40
Page: 356-362

418.
Auth: Honma, Masamitsu//Ishiyama, Ikuo
Affl: Department of Forensic Medicine, Faculty of Medicine, University of Tokyo, Japan
Ttl1: Probability of paternity in paternity testing using the DNA fingerprint procedure
Type: journal article
Area: technical or scientific
Ttl2: Human Heredity
Date: 1989
Volm: 39(3)
Page: 165-169

419.
Auth: Honma, Masamitsu//Yoshii, Tomio//Ishiyama, Ikuo//Mitani, Kohnosuke//Kominami, Ryo//Muramatsu, Masami
Affl: Department of Forensic Medicine, Faculty of Medicine, University of Tokyo, Japan
Ttl1: Individual identification from semen by the deoxyribonucleic acid (DNA) fingerprint technique
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: January 1989
Volm: 34(1)
Page: 222-227

420.
Auth: Hood, L.//Delahunty, C.//Nickerson, D.
Affl: California Institute of Technology, Pasadena, CA.
Ttl1: Automated DNA fingerprinting, polymorphic sequence tagged sites, and forensics
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 191

421.
Auth: Hooper, Celia
Ttl1: DNAs fingerprinting's first cases
Type: journal article
Area: technical or scientific
Ttl2: Journal of NIH Research
Date: March 1992
Volm: 4
Page: 81-87

422.
Auth: Hooper, Celia
Ttl1: Rancor precedes National Academy of Science's DNA fingerprinting report
Type: journal article
Area: technical or scientific
Ttl2: Journal of NIH Research
Date: March 1992
Volm: 4
Page: 76-80

423.
Auth: Hopkins[redacted]//Morten, J. E. N.//Smith, J. C.//Markham, A. F.
Ttl1: Development of methods for the analysis of DNA extracted from forensic samples (The)
Type: journal article
Area: technical or scientific
Ttl2: Journal Methods in Cell & Biology
Date: October 1989
Volm: 1
Page: 96

424.
Auth: Horn, George T.//Richards, Brenda//Klinger, Katherine W.
Affl: Department of Human Genetics, Integrated Genetics, Farmington, MA 01701
Ttl1: Amplification of a highly polymorphic VNTR segment by the polymerase chain reaction
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: March 11, 1989
Volm: 17(5)
Page: 2140

425.
Auth: Horn, Glenn T.//Bugawan, Teodorica L.//Long, Christopher M.//Erlich, Henry A.
Affl: Dept. of Human Genetics; Microbial Genetics, Cetus Corp, Emeryville, CA.
Ttl1: Allelic sequence variation of the HLA-DQ loci: relationship to serology and to insulin-dependent diabetes susceptibility
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy Science, USA
Date: August, 1988
Volm: 85
Page: 6012-6016

426.
Auth: Howard, B. H.
Affl: National Cancer Institute, Bethesda, Maryland
Ttl1: Restriction enzymes: basic properties and use in RFLP analysis
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 29-35

427.
Auth: Howlett, Rory
Affl: Assistant editor of Nature
Ttl1: DNA forensics and the FBI
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: September 21, 1989
Volm: 341(6239)
Page: 182-183

428.
Auth: Huang, Paul L.//Huang, Philip L.//Lee-Huang, Sylvia
Affl: Harvard Medical School, Massachusetts General Hospital, New York University Medical Center
Ttl1: DNA polymorphism and forensic identification
Type: book chapter
Area: technical or scientific
BkAu: Lee, Henry C.//Gaensslen, R. E.
Ttl2: DNA and Other Polymorphisms in Forensic Science
Date: 1990
Page: 1-25

429.
Auth: Huber, Peter
Ttl1: Junk science in the courtroom
Type: magazine article
Area: lay press
Ttl2: Forbes
Date: July 8, 1991
Page: 68-72

430.
Auth: Huey, Bing//Hall, Jeff
Affl: School of Public Health, University of California, Berkeley 94720
Ttl1: Hypervariable DNA fingerprinting in Escherichia coli: minisatellite probe from bacteriophage M13
Type: journal article
Area: technical or scientific
Ttl2: Journal Bacteriology
Date: May 1989
Volm: 171(5)
Page: 2528-2532

431.
Auth: Huey, Bing//Hall, Jeff M.//King, M. C.
Affl: School of Public Health, University of California, Berkeley 94720
Ttl1: A VNTR polymorphism, D1S110, at 1q21-1q31
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: April 11, 1990
Volm: 18(7)
Page: 1928

432.
Auth: Hunkapiller, Michael
Affl: Research and Development Applied Biosystems, Inc., Foster City, CA.
Ttl1: Automated DNA sequencing: technical characteristics affecting its use in forensic science
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 305-310

433.
Auth: Hupe, Robert
Ttl1: The development of DNA fingerprint use in courts of law
Type: journal article
Area: legal
Ttl2: Southwestern University Law Review
Plac: United States
Date: Summer 1990
Volm: 19(3)
Page: 1045-1065

434.
Auth: Hymer, A.
Affl: Alameda County Public Defender's Office, Oakland, California
Ttl1: DNA testing in criminal cases: a defense perspective
Type: book chapter
Area: technical or scientific
BkAu: Farley, Mark A.//Harrington, James J.
Ttl2: Forensic DNA Technology
Date: 1991
Page: 181-199

435.
Auth: Imwinkelried, Edward J.
Affl: Professor of law at the University of California, Davis, CA.
Ttl1: Court rules work to exclude valid scientific testimony
Type: journal article
Area: technical or scientific
Ttl2: The Scientist
Date: October 29, 1990
Page: 15, 17

436.
Auth: Imwinkelried, Edward J.
Affl: Professor of Law at the University of California, Davis, CA.
Ttl1: The debate in the DNA cases over the foundation for the admission of scientific evidence: the importance of human error as a cause of forensic misanalysis
Type: journal article
Area: legal
Ttl2: Washington University Law Q.
Date: 1991
Volm: 69
Page: 19

437.
Auth: Imwinkelried, Edward J.
Affl: University of California, Davis, CA
Ttl1: Recent developments in forensics science: more good news
Type: journal article
Area: technical or scientific
Ttl2: The Champion
Date: August 1990
Page: 8-13

438.
Auth: Ip, Nancy Y.//Nicholas, Leric//Baum, Howard//Balazs, Ivan
Affl: Lifecodes Corporation, Valhalla, NY 10595
Ttl1: Discovery of a novel multilocus DNA polymorphism DNF24
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: June 12, 1989
Volm: 17(11)
Page: 4427

439.
Auth: Ip, Nancy Y.//van de Stadt, I.//Loewy, Zvi G.//Leary, Susan//Grzeschik, Karl Heinz//Balazs, Ivan
Affl: Lifecodes Corporation, Valhalla, NY 10595
Ttl1: Identification and characterization of a hypervariable region (D18S27) on chromosome 18
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: October 25, 1989
Volm: 17(20)
Page: 8404

440.
Auth: Ishiyama, I.//Yoshii, T.//Honma, M.//Mukaida, M.//Yamaguchi, T.
Affl: Department of Forensic Medicine, University of Tokyo, Japan
Ttl1: DNA fingerprints: the importance in forensic medicine. II. The significance of examining DNA polymorphisms of placental tissues for the purpose of paternity determination during the early states within the first trimester
Type: journal article
Area: technical or scientific
Ttl2: Zeitschrift fur Rechtsmedizin
Date: 1988
Volm: 99(4)
Page: 241-248

441.
Auth: Ivanov, P. L.
Ttl1: DNA fingerprinting: hypervariable loci and genetic marking (a mini-review)
Type: journal article
Area: technical or scientific
Ttl2: Molekulyarnaya Biologiya (Mosk)
Date: March-April 1989
Volm: 23(2)
Page: 342-347

442.
Auth: Ivanov, P. L.//Gurtovaia, S. V.//Verbovaia, L. V.//Boldesku, N. G.//Plaksin, V. O.//Ryskov, A. P.
Ttl1: Genomic "dactyloscopy" in the expertise of disputed paternity and the determination of biological relationship
Type: journal article
Area: technical or scientific
Ttl2: Sudebno Medicinskaya Expertiza (Forensic Medical Examinations)
Date: April-June 1990
Volm: 33(2)
Page: 36-38

443.
Auth: Ivanov, [redacted]L.//Semiokhina, A. F.//Ryskov, A. P.
Ttl1: Rat DNA fingerprinting of Rattus norvegicus: a new approach in genetic analysis
Type: journal article
Area: technical or scientific
Ttl2: Genetika
Date: February 1989
Volm: 25(2)
Page: 238-249

444.
Auth: Ivanov, P. L.//Verbovaya, L. V.//Gurtovaya, S. V.
Ttl1: Use of DNA printing for diagnosis of monozygotic twins
Type: journal article
Area: technical or scientific
Ttl2: Sud.-Med Eskpert
Date: 1991
Volm: 34
Page: 32

445.
Auth: J., K. S.
Ttl1: Cut-price fingerprints
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: July 20, 1989
Volm: 340
Page: 175

446.
Auth: Jabs, Ethylin W.//Goble, Corintha A.//Cutting, Garry R.
Affl: Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD 21205
Ttl1: Macromolecular organization of human centromeric regions reveals high frequency, polymorphic macro DNA repeats
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy Science USA
Date: Janaury 1989
Volm: 86(1)
Page: 202-206

447.
Auth: Jeanpierre, M.
Affl: Unite INSERM U129 et Laboratoire de Biochimie Genetique, Paris France
Ttl1: A rapid method for the purification of DNA from blood
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: 1876
Volm: 15(22)
Page: 9611

[redacted].
[redacted]th: Jeffreys, A. J.//Royle, N. J.//Patel, I.//Armour, J. A. L.//Mac[redacted], A.//Collick, A.//Gray, I. C.//Newmann, R.//Gibbs, M.//Crosier, M.//Hill, M.//Signer, E.//Monckton, D.
Affl: Dept. of Genetics, Univ. of Leicester, University, Road, Leicester, UK
Ttl1: Principles and recent advances in human DNA fingerprinting
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 1-19

449.
Auth: Jeffreys, A. J.//Wilson, V.//Wong, Z.//Royale, N.//Patel, I.//Kelly, R.//Clarkson, R.
Affl: Department of Genetics, University of Leicester, UK
Ttl1: Highly variable minisatellites and DNA fingerprints
Type: journal article
Area: technical or scientific
Ttl2: Biochemical Coc Symp
Date: 1987
Volm: 53
Page: 165-180

450.
Auth: Jeffreys, A. J.//Wilson, V.//Wong, Z.//Patel, I.//Royle, N.//Neumann, R.//Armour, J. A.//Kelley, R.//Collick, A.//Gray, I.//Gibbs, M.
Affl: University of Leicester, Leicester, UK
Ttl1: Multi locus and single locus minisatellite DNA probes in forensic medicine
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 77-86

451.
Auth: Jeffreys, Alec J.
Affl: Department of Genetics, University of Leicester, Leicester, UK
Ttl1: 1992 William Allan Award Address
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: July 1993
Volm: 53(1)
Page: 1-5

452.
Auth: Jeffreys, Alec J.
Affl: Department of Genetics, University of Leicester, Leicester, UK
Ttl1: Highly variable minisatellites and DNA fingerprints
Type: journal article
Area: technical or scientific
Ttl2: Biochemical Society Transactions
Date: June 1987
Volm: 15(3)
Page: 309-317

453.
Auth: Jeffreys, Alec J.//Brookfield, John F. Y.//Semeonoff, Robert
Affl: Department of Genetics, University of Leicester, UK
Ttl1: Positive identification of an immigration test-case using human DNA fingerprints
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: October 31-November 6, 1985
Volm: 317(6040)
Page: 818-819

454.
Auth: Jeffreys, Alec J.//MacLeod, Annette//Neumann, Rita//Povey, Susan//Royle, Nicole J.
Affl: Department of Genetics, University of Leicester, UK
Ttl1: "Major minisatellite loci" detected by minisatellite clones 33.6 and 33.15 correspond to the cognate loci D1S111 and D7S437
Type: journal article
Area: technical or scientific
Ttl2: Genomics
Date: July 1990
Volm: 7(3)
Page: 449-452

455.
Auth: Jeffreys, Alec J.//MacLeod, Annette//Tamaki, Keiji//Neil, David L.//Monckton, Darren G.
Affl: Dept. of Genetics, Univ. of Leicester, Leicester, UK
Ttl1: Minisatellite repeat coding as a digital approach to DNA typing
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: November 21, 1991
Volm: 354
Page: 204-209

456.
Auth: Jeffreys, Alec J.//Morton, D. B.
Affl: Department of Genetics, University of Leicester, UK
Ttl1: DNA fingerprints of dogs and cats
Type: journal article
Area: technical or scientific
Ttl2: Animal Genetics
Date: 1987
Volm: 18(1)
Page: 1-15

457.
Auth: Jeffreys, Alec J.//Royle, Nicola J.//Wilson, Victoria//Wong, Zi
Affl: Dept. of Genetics, University of Leicester, Leicester, UK
Ttl1: Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: March 17, 1988
Volm: 332
Page: 278-281

458.
Auth: Jeffreys, Alec J.//Turner, Michelle//Debenham, Paul
Affl: University of Leicester; Cellmark Diagnostics, England
Ttl1: The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships, determined from extensive casework
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 48
Page: 824-840

459.
Auth: Jeffreys, Alec J.//Wilson, Victoria//Newmann, Rita//Keyte, John
Affl: Department of Genetics, University of Leicester, UK
Ttl1: Amplification of human minisatellites by the polymerase chain reaction: toward DNA fingerprinting of single cells
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: December 9, 1988
Volm: 16(23)
Page: 10953-10971

460.
Auth: Jeffreys, Alec J.//Wilson, Victoria//Thein, Swee Lay//Weatherall, David J.//Ponder, Bruce A. J.
Affl: Department of Genetics, University of Leicester, UK
Ttl1: DNA 'fingerprints' and segregation analysis of multiple markers in human pedigrees
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: July 1986
Volm: 39(1)
Page: 11-24

461.
Auth: Jeffreys, Alec J.//Wilson, Victoria//Thein, Swee Lay
Affl: Department of Genetics, University of Leicester, UK
Ttl1: Hypervariable 'minisatellite' regions in human DNA
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: March 7-13, 1985
Volm: 314(6006)
Page: 67-73

462.
Auth: Jeffreys, Alec J.//Wilson, V.//Thein, S. L.
Affl: Department of Genetics, University of Leicester; John Radcliffe Hospital, Headington, Oxford UK
Ttl1: Individual specific 'fingerprints' of human DNA
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: July 4-10, 1985
Volm: 316(6023)
Page: 76-79

463.
Auth: Jeffreys, Alec J.//Wilson, Victoria//Kelly, Robert//Taylor, Benjamin A.//Bulfield, Grahame
Affl: Department of Genetics, University of Leicester, UK
Ttl1: Mouse DNA "fingerprints": analysis of chromosome localization and germ-line stability of hypervariable loci in recombinant inbred strains
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: April 10, 1987
Volm: 15(7)
Page: 2823-2836

464.
Auth: Jeffreys, Alec J.//Wong, Zilla//Wilson, Victoria//Patel, Ila//Neumann, Rita//Royle, Nicola//Armour, John A. L.
Affl: Department of Genetics, University of Leicester, Leicester, UK
Ttl1: Applications of multilocus and single-locus minisatellite DNA probes in forensic medicine
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 283-295

465.
Auth: Jennifer R. Slimowitz//Joel E. Cohen
Affl: Dept. of Mathematics, Duke Univ.; Rockefeller Univ. New York
Ttl1: Violations of the ceiling principle: Exact conditions and statistical evidence
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: August 1993
Volm: 53
Page: 314-323

466.
Auth: Jin, Li//Chakraborty, Ranajit//Hammond, Holly A.//Caskey, C. Thomas
Affl: University of Texas; Baylor College of Medicine, Houston, Tx.
Ttl1: Polymorphisms at short tandem repeat (STR) loci within and between four ethnic populations of Texas.
Type: journal article
Area: technical or scientific
Ttl2: 8th International Congress of Human Genetics
Date: October 1991
Volm: 49(4)
Page: 14

467.
Auth: Johnson, D. M.
Affl: FBI Technical Services, Washington, D.C.
Ttl1: The national crime information center: past, present and future
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 133-135

468.
Auth: Johnson, Kathy
Ttl1: UK immigration authorities may use DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: September 3-9, 1987
Volm: 329(6134)
Page: 5

469.
Auth: Jonakait, Randolph N.
Ttl1: When blood is their argument: Probabilities in criminal cases, genetic markers, and, once again, Bayes' theorem
Type: journal article
Area: legal
Ttl2: University of Illinois Law Review
Volm: 1983(2)
Page: 369-421

470.
Auth: Jones, [redacted].//Lessells, C. M.//Krebs, J. R.
Affl: Oxford University, Oxford; Sheffield University, Sheffield, UK
Ttl1: Helpers-at-the-nest in European bee-eaters (Merops apiaster): a genetic analysis
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 169-192

471.
Auth: Kahn, R.
Affl: Metro-Dade Police Department, Miami, FL.
Ttl1: DNA chemistry and genome organization: an introduction for the forensic scientist
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 11-18

472.
Auth: Kahn, R.
Affl: Biology/Serology Section, Crime Laboratory Bureau, Metro-Dade Police Department, Miami, FL.
Ttl1: An introduction of DNA structure and genome organization
Type: book chapter
Area: technical or scientific
BkAu: Farley, Mary A.//Harrington, James J.
Ttl2: Forensic DNA Technology
Date: 1991
Page: 25-38

473.
Auth: Kanter, Evan//Baird, Michael//Shaler, Robert//Balazs, Ivan
Affl: Lifecodes Corporation, Elmsford, NY
Ttl1: Analysis of restriction fragment length polymorphisms in deoxyribonucleic acid (DNA) recovered from dried bloodstains
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: April 1986
Volm: 31(2)
Page: 403-408

474.
Auth: Kasai, K.//Mukoyama, H.//Nakamura, Y.//White, R. L.//Sakai, I.
Affl: Howard Hughes Medical Institute Research Laboratories, Salt Lake City, Utah; Laboratory of Forensic Serology, Tokyo, Japan
Ttl1: Personal identification for Japanese using variable number of tandem repeat (VNTR) loci
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 239

475.
Auth: Kasai, K.//Nakamura, Y.//White, R. L.
Affl: University of Utah, Salt Lake City, UT; Nat'l Research Institute of Police Science, Tokyo, Japan
Ttl1: Amplification of VNTR locus by the polymerase chain reaction (PCR)
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium of the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 279

476.
Auth: Kashi, Y.//Tikochinsky, Y.//Genislav, E.//Iraqi, F.//Nave A.//Beckmann J. S.//Gruenbaum, Y.//Soller, M.
Affl: Department of Genetics, Hebrew University, Jerusalem, Israel
Ttl1: Large restriction fragments containing poly-TG are highly polymorphic in a variety of vertebrates
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: March 11, 1990
Volm: 18(5)
Page: 1129-1132

477.
Auth: Katzer, Michael
Affl: Albany County District Attorney's Office, Albany, N.Y.
Ttl1: Review of present cases
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 117-126

478.
Auth: Keable, [redacted]//Bourhis, J. H.//Brison, O.//Lehn, P.//Schenmetzler, C.//Devergie, A.//Gluckman, E.
Affl: Laboratorie d'Oncologie Moleculaire, UA 1158 CNRS, Institut Gustave Roussy, Villejuif, France
Ttl1: Long term study of chimaerism in bone marrow transplantation recipients for severe aplastic anaemia
Type: journal article
Area: technical or scientific
Ttl2: British Journal Haematology
Date: April 1989
Volm: 71(4)
Page: 525-533

479.
Auth: Kearney, James T.
Affl: FBI Laboratory, Quantico, VA.
Ttl1: Guidelines for a quality assurance program for DNA restriction fragment length polymorphism analysis
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: April-June 1989
Volm: 16(2)
Page: 40-59

480.
Auth: Kearney, James J.
Affl: Section Chief, Forensic Science Research and Training Center, Quantico, VA.
Ttl1: International seminar on the forensic applications of PCR technology
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 123-124

481.
Auth: Kearney, James J.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Summary of the first meeting of the technical working group of DNA analysis methods
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: 1988
Volm: 15
Page: 115

482.
Auth: Kelly, K. F.//Rankin, J. J.//Wink, R. C.
Ttl1: Method and applications of DNA fingerprinting: a guide for the non-scientist
Type: journal article
Area: legal
Ttl2: Criminal Law Review
Plac: Great Britain
Date: February 1987
Page: 105-110

[redacted].
Auth: Kelly, Robert//Bulfield, Grahame//Collick, Andrew//Gibbs, Mark//[redacted]effreys, Alec J.
Affl: Department of Genetics, University of Leicester, UK
Ttl1: Characterization of a highly unstable mouse minisatellite locus: evidence for somatic mutation during early development
Type: journal article
Area: technical or scientific
Ttl2: Genomics
Date: November 1989
Volm: 5(4)
Page: 844-856

484.
Auth: Kidd, J. R.//Black, F. L.//Weiss, K. M.//Balazs, I.//Kidd, K. K.
Ttl1: Studies of three Amerindian populations using nuclear DNA polymorphisms
Type: journal article
Area: technical or scientific
Ttl2: Human Biology
Date: December 1991
Volm: 63
Page: 775

485.
Auth: King, Julia
Ttl1: Disorder in the court when science takes the witness stand
Type: journal article
Area: technical or scientific
Ttl2: The Scientist
Date: October 29, 1990
Page: 15

486.
Auth: King, Mary Claire
Affl: School of Public Health, University of California, Berkeley, CA
Ttl1: Invited editorial: genetic testing of identity and relationship
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: February 1989
Volm: 44(2)
Page: 179-181

487.
Type: book
Area: technical or scientific
BkAu: Kirby, Lorne T.
Ttl2: DNA Fingerprinting - An Introduction
Plac: 15 East 26th Street, New York, N.Y. 10010
Publ: Stockton Press
Date: 1990
Volm: ISBN 0-935859-94-2
Page: vii-xvi, 1-365

488.
Auth: Kobayas[redacted] Ryo//Nakasuchi, Hiromitsu//Nakahori, Yutaka//Nakagome, Yasuo/-/Matsuzawa, Shigetaka
Affl: Juntendo University School of Medicine; National Children's Medical Center, Tokyo, Japan
Ttl1: Sex identification in fresh blood and dried bloodstains by a nonisotopic deoxyribonucleic acid (DNA) analyzing technique
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: May 1988
Volm: 33(3)
Page: 613-620

489.
Auth: Kolata, Gina
Affl: Writer - The New York Times
Ttl1: Chief says panel backs courts' use of a genetic test - Times account in error - Report urges strict standards but no moratorium on DNA fingerprinting for now
Type: newspaper
Area: lay press
Ttl2: The New York Times
Date: April 15, 1992
Volm: CXLI(48,937)
Page: A-1, A15-3

490.
Auth: Kolata, Gina
Affl: Writer - New York Times
Ttl1: Chief says panel backs courts' use of a genetic test
Type: newspaper
Area: lay press
Ttl2: The New York Times
Date: Wednesday, April 15, 1992
Volm: XCLI(48,937)
Page: A1, A23 Column 1

491.
Auth: Kolata, Gina
Affl: Writer - New York Times
Ttl1: Critic of 'Genetic Fingerprint' testing tells of pressure to withdraw paper
Type: newspaper
Area: lay press
Ttl2: The New York Times
Plac: New York
Date: December 20, 1991
Page: C18

492.
Auth: Kolata, Gina
Affl: Writer - New York Times
Ttl1: DNA Fingerprinting: Built-In Conflict
Type: newspaper
Area: lay press
Ttl2: The New York Times NATIONAL
Date: Friday, April 17, 1992
Page: A13

493.
Auth: Kolata, Gina
Affl: Writer - New York Times
Ttl1: F.B.I. defends genetic tests
Type: newspaper
Area: lay press
Ttl2: The New York Times
Plac: New York
Date: December 25, 1991
Page: A8

494.
Auth: Kolata, Gina
Affl: Writer - The New York Times
Ttl1: Some Scientists doubt the value of genetic fingerprint evidence
Type: newspaper
Area: lay press
Ttl2: The New York Times
Plac: New York
Date: January 29, 1990
Volm: CXXXIX(48,130)
Page: A1

495.
Auth: Kolata, Gina
Affl: Writer - New York Times
Ttl1: U.S. panel seeking restriction on use of DNA in courts - Labs' standards faulted - Judges are asked to bar genetic "fingerprint" until basis in science is stronger
Type: newspaper
Area: lay press
Ttl2: The New York Times
Date: Tuesday, April 14, 1992
Volm: CXLI(48,936)
Page: A1, C7 Column 1

496.
Auth: Konzak, K. C.//Reynolds, R.//von Beroldingen, C.//Buoncristiani, M./-/Sensabaugh, G. F.
Affl: University of California, Berkeley, CA.
Ttl1: Effects of DNA damage and degradation of RFLP analysis
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 255

497.
Auth: Koppel, [redacted]
Affl: ABC News, 47 W. 66th St., New York, NY 10023
Ttl1: DNA fingerprinting
Type: public transcript
Area: television
BkAu: Report from ABC New Correspondent - John Martin
Ttl2: Nightline
Date: August 15, 1989
Volm: Show #2147
Srce: Transcripts: Journal Graphics, Inc.

498.
Auth: Koshland Jr., Daniel E.
Affl: Editor - Science Magazine
Type: letter to responses
Area: technical or scientific
Ttl2: Science
Date: February 28, 1992
Volm: 255
Page: 1052-1053

499.
Auth: Koshland, Jr Daniel E.
Affl: Editor - Science Magazine
Ttl1: DNA Fingerprinting and Eyewitness Testimony
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: May 1, 1992
Volm: 256
Page: 593

500.
Auth: Kovacs, B. W.//Shahbahrami, B.//Comings, D. E.//Johnson, B. L.
Affl: City of Hope Medical Center, Duarte, CA., Los Angeles County Sheriff's Department, Los Angeles, CA.
Ttl1: Human identification with synthetic oligonucleotide probes
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 223-224

501.
Auth: Kraemer, Paul M.//Ratliff, Robert L.//Bartholdi, Marty F.//Brown, Nancy C.//Longmire, Jonathan L.
Affl: Life Sciences Division, Los Alamos Nat'l Lab., Los Alamos, NM 87545
Ttl1: Use of variable number of tandem repeat (VNTR) sequences for monitoring chromosomal instability
Type: journal article
Area: technical or scientific
Ttl2: Progress Nucleic Acid Research Molecular Biology
Date: 1989
Volm: 36
Page: 187-204

502.
Auth: Krawczak, Michael//Bockel, Barbara
Affl: Abteilung Humangenetick, Hannover; Institut fur Humangenetick, Gottingen, Federal Republic of Germany
Ttl1: DNA fingerprinting: a short note on mutation rates
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: 1991
Volm: 87
Page: 632-633

503.
Auth: Krawzak, Michael
Affl: Abteilung Humangenetik, Medizinische Hochschule, Hannover, FRG
Ttl1: DNA fingerprinting and mutations rates: reply to letter by Ritter
Type: letter to editor
Area: technical or scientific
Ttl2: Human Genetics
Date: May 1992
Volm: 89(3)
Page: 363-364

504.
Auth: Kriss, Jodi//Herrin, George//Forman, Lisa//Cotton, Robin
Affl: Cellmark Diagnostics, Germantown, MD 20876
Ttl1: Digestion conditions resulting in altered but site specificity for Hinf1
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: June 25, 1990
Volm: 18(12)
Page: 3665

505.
Auth: Kuhnlein, U.//Zadworny, D.//Dawe, Y.//Fairfull, R. W.//Gavora, J. S.
Affl: Department of Animal Science, Maedonald College of McGill University, Ste. Anne de Bellevue, Quebec, Canada
Ttl1: Assessment of inbreeding by DNA fingerprinting: development of a calibration curve using defined strains of chickens
Type: journal article
Area: technical or scientific
Ttl2: Genetics
Date: May 1990
Volm: 125(1)
Page: 161-165

506.
Auth: Kuhnlein, U.//Zadworny, D.//Gavora, J. S.//Fairfull, R. W.
Affl: MacDonald Campus of McGill University, Quebec; Animal Research Centre, Agriculture Canada, Ottawa, Canada
Ttl1: Identification of markers associated with quantitative trait loci in chickens by DNA fingerprinting
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 274-282

507.
Auth: Kwok, S.//Higuchi, R.
Ttl1: Avoiding false positives with PCR
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: May 18, 1989
Volm: 339
Page: 237-238

508.
Auth: Labaton, Stephen
Ttl1: DNA fingerprinting is facing showdown at Ohio hearing
Type: journal article
Area: legal
Ttl2: Chicago Daily Law Bulletin
Plac: Ohio
Date: June 22, 1990
Volm: 136(123)
Page: 3 (col 2)

509.
Auth: Lagoda, P. J.//Seitz, G.//Epplen, J. T.//Issinger, O. G.
Affl: Institut fur Humangenetik, Universitat des Saarlandes, Homburg, Federal Republic of Germany
Ttl1: Increased detectability of somatic changes in the DNA from human tumors after probing with "synthetic" and "genome-derived" hypervariable multilocus probes
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: December 1989
Volm: 84(1)
Page: 35-40

510.
Auth: Lander, Eric S.
Affl: Whitehead Institute for Biomedical Research, Nine Cambridge Center, Massachusetts 02142
Ttl1: DNA fingerprinting on trial
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: June 15, 1989
Volm: 339(6225)
Page: 501-505

511.
Auth: Lander, Eric S.
Affl: Whitehead Institute for Biomedical Research and Harvard University
Ttl1: Expert's report in People vs Castro
Type: unpublished document
Area: technical or scientific
Case: People vs Castro
Cort: Washington D.C.
Date: 1991
Page: 1-51

512.
Auth: Lander, Eric S.
Affl: Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA.
Ttl1: Invited Editorial: Research on DNA typing catching up with courtroom application
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 48
Page: 819-823

513.
Auth: Lander, Eric S.
Affl: Whitehead Institute for Biomedical Research and Dept. of Biology, Massachusetts Institute of Technology, Cambridge, MA.
Ttl1: Lander Reply
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 49
Page: 899-903

514.
Auth: Lander, Eric S.
Affl: Whitehead Institute for Biomedical Research, Cambridge, Massachusetts
Ttl1: Population genetic considerations in the forensic use of DNA typing
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 143-156

515.
Auth: Landers, T. A.
Affl: Life Technologies, Inc, Gaithersburg, Maryland
Ttl1: Strategies for labeling and detection of nucleic acid probes
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 63-75

516.
Auth: Landers, Terry A.
Affl: Life Technologies Incorporated, Gaithersbury, Maryland
Ttl1: Labelling and detection of nucleic acid probes
Type: book chapter
Area: technical or scientific
BkAu: Lee, Henry C.//Gaensslen, R. E.
Ttl2: DNA and Other Polymorphisms in Forensic Science
Date: 1990
Page: 114-134

517.
Auth: Lauer, Joyce//Shen, Che Kun James//Maniatis, Tom
Affl: Div. of Biology, California Institute of Technology, Pasadena, CA.
Ttl1: The chromosomal arrangement of human a-like globin genes: sequence homology and a-globin gene deletions
Type: journal article
Area: technical or scientific
Ttl2: Cell
Date: May 1980
Volm: 20
Page: 119-130

518.
Auth: Lawson, Leigh C.
Ttl1: DNA fingerprinting and its impact upon criminal law
Type: journal article
Area: legal
Ttl2: Mercer Law Review
Plac: United States
Date: Summer 1990
Volm: 41(4)
Page: 1453-1468

519.
Auth: Lawton, M. E.//Stringer, P.//Churton, M.
Affl: Chemistry Division, Auckland, New Zealand
Ttl1: DNA profiles from dental pulp
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 217-218

520.
Auth: Layton, D. M.//Mufti, G. J.
Affl: Dept. of Haematology, King's College School of Medicine and Dentistry, Denmark Hill, London, UK
Ttl1: Human cancer DNA fingerprint analysis
Type: letter to editor
Area: technical or scientific
Ttl2: British Journal Cancer
Date: September 1987
Volm: 56(3)
Page: 381

521.
Auth: Ledwith, B. J.//Storer, R. D.//Prahalada, S.//Manam, S.//Leander, K. R.//van Zwieten, M. J.//Nichols, W. W.//Bradley, M. O.
Affl: Merck Sharp and Dohme Research Laboratories, West Point, PA 19486
Ttl1: DNA fingerprinting of 7,12-dimethylbenz[a]anthracine-induced and spontaneous CD-1 mouse liver tumors
Type: journal article
Area: technical or scientific
Ttl2: Cancer Research
Date: September 1, 1990
Volm: 50(17)
Page: 5245-5249

522.
Auth: Ledwith, Brian J.//Manam, Sujata//Nichols, Warren W.//Bradley, Matthews O.
Affl: Merck Sharp and Dohme Research Laboratories, West Point, PA 19486
Ttl1: Preparation of synthetic tandem-repetitive probes for DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Biotechniques
Date: August 1990
Volm: 9(2)
Page: 149-152

523.
Auth: Lee, H. C.//Pagliaro, E. M.//Gaensslen, R. E.//Berka, K. M.//et al
Affl: Connecticut State Police, Meriden; Univ. of New Haven, West Haven, Conn.
Ttl1: DNA analysis in human bone tissue: RFLP typing
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Science Society
Date: October 1990
Volm: 31(2)
Page: 209

524.
Auth: Lee, H. C.//Pagliaro, E. M.//Gaensslen, R. E.//Keith, T.//Rose, S.
Affl: Connecticut State Police Forensic Science Lab.; Univ. of New Haven Forensic Sciences Program, Conn.; Collaborative Research, Inc., Bedford, Massachusetts
Ttl1: The effect of detergents on DNA in blood and semen
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 269-271

525.
Auth: Lee, H.//Ruano, G.//Pagliaro, E.//Berka, K. M.//Gaensslen, R. E.
Ttl1: DNA analysis in human bone and other speciments of forensic interest: PCR typing and testing
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Science Society
Date: 1991

526.
Type: book
Area: technical or scientific
BkAu: Lee, Henry C.//Gaensslen, Robert E.
Ttl2: Advances in Forensic Science: DNA and Other Polymorphisms in Forensic Science
Plac: United States
Publ: Year Book Medical Publishers, Inc.
Date: 1990
Volm: ISNB 0-8151-5348-1
Page: vii-xii, 1-278

527.
Auth: Lee, Henry C.//Gaensslen, Robert E.
Affl: Connecticut State Police Forensic Science Laboratory, Meriden, Conn.; University of New Haven Forensic Science Laboratories, West Haven, Conn.
Ttl1: Analysis of human bone DNA by PCR
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 156-159

528.
Auth: Lee, Henry C.//Gaensslen, Robert E.
Affl: Connecticut State Police, Meriden; Univ. of New Haven, West Haven, Conn.
Ttl1: The need for standardization of DNA analysis methods
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 217-222

529.
Auth: Lee, Henry C.//Pagliaro, Elaine M.//Berka, Karen M.//Folk, N. L.//et al.
Affl: Connecticut State Police, Meriden; Univ. of New Haven, West Haven, Conn.
Ttl1: Genetic markers in human bone: I. deoxyribonucleic acid (DNA) analysis
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: March 1991
Volm: 36(2)
Page: 320-330

530.
Auth: Levy, Gabor B.
Affl: Consulting editor, American Clinical Laboratory
Ttl1: Castro vs DNA typing
Type: letter to editor
Area: technical or scientific
Ttl2: International Scientific Communications
Date: May, June 1990
Page: 4, 6

531.
Type: book
Area: general reference
BkAu: Lewin, Benjamin
Ttl2: Genes IV
Plac: Walton Street, Oxford, OX2 6DP, U.K.
Publ: Oxford University Press
Date: 1990
Volm: ISBN 0-19-854268-2
Page: i-xxii, 1-857

532.
Auth: Lewin, Roger
Ttl1: DNA fingerprints in health and disease
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: August 1, 1986
Volm: 233(4763)
Page: 521-522

533.
Auth: Lewin, Roger
Ttl1: DNA typing on the witness stand
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: June 2, 1989
Volm: 244
Page: 1033-1035

534.
Auth: Lewin, Roger
Ttl1: DNA typing is called flawed
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: July 28, 1989
Volm: 245
Page: 355

535.
Auth: Lewin, Roger
Ttl1: Limits to DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: March 24, 1989
Volm: 243(4898)
Page: 1549-1551

536.
Auth: Lewis, [redacted]cia E.//Kouri, Richard E.//Latorra, David//Berka, Karen M.//Lee, Henry C.//Gaensslen, Robert E.
Affl: BIOS Corp., New Haven; University of New Haven, West Haven; Connecticut State Police Forensic Science Laboratory, Meriden, CT.
Ttl1: Restriction fragment length polymorphism DNA analysis by the FBI laboratory protocol using a simple, convenient hardward system
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: September 1990
Volm: 35(5)
Page: 1186-1190

537.
Auth: Lewis, Ricki
Affl: State University of New York at Albany
Ttl1: DNA fingerprints. Witness of the prosecution
Type: journal article
Area: technical or scientific
Ttl2: Discover
Date: June 1988
Page: 44-52

538.
Auth: Lewontin, R. C.
Affl: Harvard University, Cambridge, MA.
Ttl1: Book reviews - The dream of the human genome
Type: book review
Area: technical or scientific
Ttl2: The New York Review
Date: May 28, 1992
Page: 31-40

539.
Auth: Lewontin, Richard C.
Affl: Harvard School of Public Health
Ttl1: Population genetic problems in the forensic use of DNA profiles
Type: unpublished document
Area: technical or scientific
Cort: Court unknown
Page: 1-27

540.
Auth: Lewontin, Richard C.//Hartl, Daniel L.
Affl: Museum of Comparative Zoology at Harvard University; Washington University School of Medicine
Ttl1: Population genetics in forensic DNA typing
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: December 20, 1991
Volm: 254
Page: 1745-1750

541.
Auth: Lewontin, Richard C.//Hartl, Daniel
Affl: Harvard University, Cambridge, MA.; Washington University School of Medicine, St. Louis, MO.
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: February 28, 1992
Volm: 255
Page: 1054-1055

542.
Auth: Li, Honghua//Gyllensten, Ulf B.//Cui, Xiangfeng//Saiki, Randall K.//Erlich, Henry A.//Arnheim, Norman
Affl: University of Southern California, Los Angeles; Cetus Corp., CA
Ttl1: Amplification and analysis of DNA sequences in single human sperm and diploid cells
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: September 1988
Volm: 335(6189)
Page: 414-417

543.
Auth: Liebeschuetz, Joe
Ttl1: Statutory control of DNA fingerprinting in Indiana
Type: journal article
Area: legal
BkAu: Indiana Law Review
Plac: Indiana
Date: Winter, 1991
Volm: 25(1)
Page: 205-233

544.
Auth: Lienert, Kristin//Fowler, J. Craig S.
Affl: The Flinders University of South Australia , Bedford Park, S. Australia
Ttl1: Analysis of mixed human/microbial DNA samples: A validation study of two PCR AMP-FLP typing methods
Type: journal article
Area: technical or scientific
Ttl2: BioTechniques
Date: August 1992
Volm: 13(2)
Page: 276-281

545.
Auth: Lincoln, P. J.//Phillips, C. P.//Thomson, J. A.//Watts, P. H.//Wood, N. J.
Affl: London Hospital Medical College, London, England
Ttl1: The reproducibility of CRI probe L336 for identification and parentage analysis
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 209-210

546.
Auth: Lloyd, Marilyn A.//Fields, Mary J.//Thorgaard, Gary H.
Affl: Program in Genetics and Cell Biology, Washington State University, Pullman, WA. 99164-4350
Ttl1: BKm minisatellite sequences are not sex associated but reveal DNA fingerprint polymorphisms in rainbow trout
Type: journal article
Area: technical or scientific
Ttl2: Genome
Date: October 1989
Volm: 32(5)
Page: 865-868

547.
Auth: Logtenberg, H.//Bakker, E.
Affl: Dept. of Biological Criminalistics, Neatherlands Forensic Laboratory; Ryswyk; Dept. of Human Genetics, University of Leiden
Ttl1: The DNA fingerprint
Type: journal article
Area: technical or scientific
Ttl2: Endeavour
Date: 1988
Volm: 12(1)
Page: 28-33

548.
Auth: Long Jr., Robert R.
Ttl1: The DNA fingerprint: a guide to admissibility
Type: journal article
Area: legal
Ttl2: Army Lawyer
Plac: United States
Date: October 1988
Page: 36-45

549.
Auth: Longmire, J. L.//Ambrose, R. E.//Brown, N. C.//Cade, T. J.//Maechtle, T. L.//Seegar, W. S.//Ward, F. P.//White, C. M.
Affl: Los Alamos National Laboratory; United States Fish and Wildlife Service;Boise State Univ; Aberdeen Proving Grounds; Brigham Young Univ.
Ttl1: Use of sex linked minisatellite fragments to investigate genetic differentiation and migration of North American populations of the Peregrine Falcon (Falco peregrinus)
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 217-229

550.
Auth: Longmire, Jonathan L.//Kraemer, Paul M.//Brown, Nancy C.//Hardekopf, L. Cathy//Deaven, Larry L.
Affl: Life Sciences Division, Los Alamos National Laboratory, NM 87545
Ttl1: A new multi-locus DNA fingerprinting probe: pV47-2
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: March 25, 1990
Volm: 18(6)
Page: 1658

551.
Auth: Longobardi, JoAnn Marie
Ttl1: DNA fingerprinting and the need for a national data base
Type: journal article
Area: legal
Ttl2: Fordham Urban Law Journal
Plac: United States
Date: September-October 1989
Volm: 17(3)
Page: 323-357

552.
Auth: Ludes, B. P.//Mangin, P. D.//Malicer, D. J.//Chalumeau, A. N.//Chaumont, A. J.
Affl: Institut de Medecine Legale et de Medecine Sociale, Strasbourg, France
Ttl1: Parentage determination on aborted fetal material through deoxyribonucleic acid (DNA) profiling
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: July 1991
Volm: 36
Page: 1219-1223

553.
Auth: Lygo, Jo
Ttl1: Sharpening the focus
Type: journal article
Area: legal
Ttl2: New Law Journal
Plac: Great Britain
Date: April 5, 1991
Volm: 141(6498)
Page: 448(4)

554.
Auth: Lynch, M.
Affl: Dept. of Biology, University of Oregon, Eugene, Oregon 79403
Ttl1: Analysis of population genetic structure by DNA fingerprinting
Type: book chapter
Area: technical or scientific
BkAu: Burek, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 113-126

555.
Auth: Lynch, Michael
Affl: Department of Ecology, Ethology and Evolution, University of Illinois, Champaign 61820
Ttl1: Estimation of relatedness by DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Molecular Biology Evolution
Date: September 1988
Volm: 5(5)
Page: 584-599

556.
Auth: MacDonald, M. E.//Cheng, S. V.//Zimmer, M.//Haines, J. L.//Poustka, A./-/Allitto, B.//Smith, B.//Whaley, W. L.//Romano, D. M.//Jagadeesh, J.//et al
Affl: Molecular Neurogenetics Laboratory, Massachusetts General Hospital, Boston 02114
Ttl1: Clustering of multiallele DNA markers near the Huntington's disease gene
Type: journal article
Area: technical or scientific
Ttl2: Journal Clinical Investigation
Date: September 1989
Volm: 84(3)
Page: 1013-1016

557.
Auth: Macedo, A. M.//Medeiros, A. C.//Pena, S. D.
Affl: Department of Biochemistry, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Ttl1: A general method for efficient non-isotopic labeling of DNA probes cloned in M13 vectors: application to DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: June 12, 1989
Volm: 17(11)
Page: 4414

558.
Auth: Maher, Fred
Ttl1: DNA fingerprinting approved by court
Type: journal article
Area: legal
Ttl2: Pennsylvania Law Journal-Reporter
Plac: Pennsylvania
Date: July 3, 1989
Volm: 12(26)
Page: 1 (col 3)

559.
Auth: Maidment, Susan
Ttl1: DNA fingerprinting
Type: journal article
Area: legal
Ttl2: New Law Journal
Plac: Great Britain
Date: April 11, 1986
Volm: 13(6249)
Page: 326(1)

560.
Auth: Makeig, John
Affl: Houston Chronicle, Houston, Texas
Ttl1: DNA testing clears suspect in rape case
Type: newspaper
Area: lay press
Ttl2: Houston Chronicle
Plac: Houston, Texas
Date: October 30, 1991
Page: 16A

561.
Auth: Mangin, P. D.//Ludes, B. P.
Affl: Institute of Legal Medicine, Strasbourg, France
Ttl1: A forensic application of DNA typing. Paternity determination in a putrefied fetus
Type: journal article
Area: technical or scientific
Ttl2: American Journal Forensic Medical Pathology
Date: 1991
Volm: 12(2)
Page: 161-163

562.
Auth: Mangin, [redacted] D.//Ludes, B.//Lugnier, A.//Chaumont, A. J.
Ttl1: Les empreintes genetiques en medecine legale
Type: journal article
Area: technical or scientific
Ttl2: Journal Med. Leg. Droit. Med
Date: 1991
Volm: 34
Page: 2

563.
Auth: Markowicz, Karen R.//Tonelli, Lois A.//Anderson, Mariane B.//Green, David J.//Herrin, George L.//Cotton, Robin W.//et al
Ttl1: Use of deoxyribonucleic acid (DNA) fingerprints for identity determination: comparison with traditional paternity testing methods (part 2)
Type: journal article
Area: legal
Ttl2: Journal Forensic Sciences
Plac: United States
Date: November 1990
Volm: 35(6)
Page: 1270-1276

564.
Auth: Marks, Bruce S.
Ttl1: Dispute resolution in the space age: forensic applications of earth observation satellite data through adaptation of technical standards similar to DNA fingerprinting protocols
Type: journal article
Area: legal
Ttl2: Ohio State Journal of Dispute Resolution
Plac: United States
Date: Fall 1989
Volm: 5(1)
Page: 19-73

565.
Auth: Marx, Jean L.
Ttl1: DNA fingerprinting takes the witness stand
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: June 17, 1988
Volm: 240(4859)
Page: 1616-1618

566.
Auth: May, Celia A.//Wetton, Jon H.
Affl: University of Nottingham, Queens Medical Centre, Nottingham, UK
Ttl1: DNA fingerprinting by specific priming of concatenated oligonucleotides
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: 1991
Volm: 19(16)
Page: 4557

567.
Auth: Mayersak, J. S.
Ttl1: DNA fingerprinting problems for resolution
Type: journal article
Area: legal
Ttl2: Medical Trail Technique Quarterly
Plac: United States
Date: Summer 1990
Volm: 36(4)
Page: 441-449

568.
Auth: McCabe, Edward R. B.
Affl: Baylor College of Medicine, Houston, Tx.
Ttl1: Applications of DNA fingerprinting in pediatric practice
Type: journal article
Area: technical or scientific
Ttl2: The Journal of Pediatrics
Date: April 1992
Volm: 120(4-1)
Page: 499-509

569.
Auth: McCabe, Edward R. B.//Huang, Shu Zhen//Seltzer, William K.//Law, Martha L.
Affl: University of Colorado School of Medicine; Shanghai Children's Hospital, Peoples Republic of China; Eleanor Roosevelt Inst. for Cancer Research
Ttl1: DNA microextraction from dried blood spots on filter paper blotters: Potential applications to newborn screening
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: 1987
Volm: 75
Page: 213-216

570.
Auth: McElfresh, Kevin C.
Affl: Forensic and Paternity Laboratories, Lifecodes Corporation, Valhalla, NY
Ttl1: DNA fingerprinting
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: October 13, 1989
Volm: 246(4927)
Page: 192

571.
Auth: McGourty, Christine
Ttl1: Genetic tests made official by UK courts
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: June 8, 1989
Volm: 339(6224)
Page: 408

572.
Auth: McGourty, Christine
Affl: Writer - Nature Magazine
Ttl1: New York State leads on genetic fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: September 14, 1989
Volm: 341
Page: 90

573.
Auth: McGourty, Christine
Ttl1: Profiles bank on the way
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: June 1, 1989
Volm: 339(6223)
Page: 327

574.
Auth: McMahon, E. J.
Ttl1: Hearing opens on reliability of DNA fingerprinting test
Type: journal article
Area: legal
Ttl2: New York Law Journal
Plac: New York (State)
Date: February 24, 1988
Volm: 199(35)
Page: 1 (col 3)

575.
Auth: McNally, Lorah//Baird, Michael//McElfresh, Kevin//Eisenberg, Arthur//Balazs, Ivan
Affl: Lifecodes Corporation, Valhalla, NY 10595
Ttl1: Increased migration rate observed in DNA from evidentiary material precludes the use of sample mixing to resolve forensic cases of identity
Type: journal
Area: technical or scientific
Ttl2: Applied Theoretical Electrophoresis
Date: 1990

576.
Auth: McNally, Lorah//Shaler, Robert C.//Baird, Michael//Balazs, Ivan//Kobilinsky, Lawrence//De Forest, Peter
Affl: Lifecodes Corporation, Valhalla, NY
Ttl1: The effects of environment and substrata on deoxyribonucleic acid (DNA): the use of casework samples for New York City
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: September 1989
Volm: 34(5)
Page: 1070-1077

577.
Auth: McNally, Lorah//Shaler, Robert C.//Baird, Michael Balazs ,. Ivan//De Forest, Peter//Kobilinsky, Lawrence
Affl: Lifecodes Corporation, Valhalla, NY
Ttl1: Evaluation of deoxyribonucleic acid (DNA) isolated from human bloodstains exposed to ultraviolet light, heat, humidity and soil contamination
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: Sepember 1989
Volm: 34(5)
Page: 1059-1069

578.
Auth: Medeiros, Arnaldo C.//Macedo, Andrea M.//Pena, Sergio D. J.
Affl: Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Ttl1: M13 Bioprints: non-isotopic detection of individual-specific human DNA fingerprints with biotinylated M13 bacteriophage
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Plac: Brazil
Date: December 1989
Volm: 43(3)
Page: 275-280

579.
Auth: Medeiros, Arnaldo C.//Macedo, Andrea M.//Pena, Sergio D. J.
Affl: Department of Biochemistry, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Ttl1: A simple non-isotopic method for DNA fingerprinting with M13 phade
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: November 11, 1988
Volm: 16(21)
Page: 10394

580.
Auth: Merz, Beverly
Ttl1: DNA fingerprints come to court
Type: journal article
Area: technical or scientific
Ttl2: Journal American Medical Association
Date: April 15, 1988
Volm: 259(15)
Page: 2193-2194

581.
Auth: Michaud, Stephen G.
Ttl1: DNA detectives. Genetic 'fingerprinting' may herald a revolution in law enforcement
Type: newspaper article
Area: laypress
Ttl2: New York Times
Plac: New York
Date: November 21, 1988
Volm: Sunday

582.
Auth: Michaud, Stephen G.
Ttl1: DNA fingerprints spawn law enforement revolution
Type: journal article
Area: legal
Ttl2: Chicago Daily Law Bulletin
Plac: United States; Great Britain
Date: November 10, 1988
Volm: 134(221)
Page: 2 (col 3)

583.
Auth: Min, Gao L.//Hibbin, Jill//Arthur, Christopher//Apperley, Jane//Jeffreys, Alec//Goldman, John
Affl: MRC Leukemia Unit, Royal Postgraduate Medical School, London
Ttl1: Use of minisatellite DNA probes for recognition and characterization of relapse after allogeneic bone marrow transplantation
Type: journal article
Area: technical or scientific
Ttl2: British Journal Haematology
Date: February 1988
Volm: 68(2)
Page: 195-201

584.
Auth: Moenssens, Andre A.
Affl: Law professor at the University of Richmond, senior co-author of "Scientific Evidence in Criminal Cases" (3rd ed. 1986).
Ttl1: DNA evidence and its critics - how valid are the challenges?
Type: journal article
Area: technical or scientific
Ttl2: Jurimetrics Journal
Date: Fall, 1990
Page: IV80-IV101

585.
Auth: Monson, Keith L.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, Va
Ttl1: The FBI system for semiautomated analysis of DNA autoradiograms
Type: unpublished document
Ttl2: Available from the FBI
Plac: Quantico, Va

586.
Auth: Monson, Keith L.
Affl: Foresnic Science Research and Training Section, FBI Academy, Quantico, Va
Ttl1: Semiautomated analysis of DNA autoradiograms
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1988
Volm: 15(4)
Page: 104-105

587.
Auth: Monson, Keith L.//Budowle, Bruce
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: A system for semiautomated analysis of DNA autoradiograms
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of an International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 127-132

588.
Auth: Moreno, Ruben F.//Booth, Frank//Thomas, Stanley M.//Tilzer, Lowell L.
Affl: Kansas University Medical Center, Department of Pathology and Oncology, Kansas City
Ttl1: Enchanced conditions for DNA fingerprinting with biotinylated M13 bacteriophage
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: July 1990
Volm: 35(4)
Page: 831-837

589.
Auth: Morris, Jeffrey W.//Sanda, A. I.//Glassberg, Jeffrey
Affl: Memorial Medical Center; Rockefeller University; Nucleics Inc.
Ttl1: Biostatistical evaluation of evidence from continuous allele frequency distribution deoxyribonucleic acid (DNA) probes in reference to disputed paternity and identity
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Page: 1311-1317

590.
Auth: Morton, D. B.//Yaxley, R. E.//Patel, I.//Jeffreys, A. J.//Howes, S. J./-/Debenham, P. G.
Affl: Unit of Biomedical Services, University of Leicester
Ttl1: Use of DNA fingerprint analysis in identification of the sire
Type: journal article
Area: technical or scientific
Ttl2: Veterinary Record
Date: December 19-26, 1987
Volm: 121(25-26)
Page: 592-594

591.
Auth: Morton, [redacted]wton E.
Affl: Southampton General Hospital, Southampton, U.K.
Ttl1: Genetic structure of forensic populations
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy of Science USA
Date: April 1992
Volm: 89
Page: 2556-2560

592.
Auth: Moss, Debra Cassens
Ttl1: DNA the new fingerprints
Type: journal article
Area: legal
Ttl2: ABA Journal
Plac: United States
Date: May 1, 1988
Volm: 74
Page: 66(5)

593.
Auth: Moss, Debra Cassens
Ttl1: Free at last
Type: journal article
Area: legal
Ttl2: ABA Journal
Plac: Illinois
Date: October 1989
Volm: 75
Page: 19(1)

594.
Auth: Motulsky, Arno G.
Affl: Department of Medicine and of Genetics, University of Washington, Seattle, WA.
Ttl1: Societal problems of forensic use of DNA techology
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 3-12

595.
Auth: Mouri, Michael
Ttl1: The myth of the DNA fingerprint - is it for real?
Type: journal article
Area: legal
Ttl2: Medical Trial Technique Quarterly
Plac: United States
Date: Spring 1991
Volm: 37(3)
Page: 337-359

596.
Auth: Mudd, James L.
Affl: FBI Laboratory, Quantico, VA.
Ttl1: Laboratory safety
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 87-91

597.
Auth: Mudd, James L.//Presley, Lawrence A.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Quality control in DNA typing: a proposed protocol
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1988
Volm: 15(4)
Page: 109-113

598.
Auth: Mueller, L.
Affl: University of California, Department of Ecology and Evolutionary Biology, Irvine, CA.
Ttl1: Population genetics of hypervariable human DNA
Type: book chapter
Area: technical or scientific
BkAu: Farley, Mark A.//Harrington, James J.
Ttl2: Forensic DNA Technology
Date: 1991
Page: 51-62

599.
Auth: Mulhare, Patricia//McQuillen, Eleanore//Colline, Cheryl//Heintz, Nicholas//Howard, Phillip
Affl: University of Vermont College of Medicine; Vermont Dept. of Health, Burlington, Vermont
Ttl1: An unusual case of using DNA polymorphisms to determine partentage of human remains
Type: journal article
Area: technical or scientific
Ttl2: American Journal of Forensic Medicine and Pathology
Date: 1991
Volm: 12(2)
Page: 157-160

600.
Auth: Muller, [redacted]ver//Alonso, Angel//Buchmann, Albrecht//Bauer-Hofmann, Richard//Bock, Karl Walter//Schwarz, Michael
Affl: German Cancer Research Center, Im Neuenheimer Feld, Heidelberg, Germany
Ttl1: Detection of genomic alterations in carcinogen-induced mouse liver tumors by DNA fingerprint analysis
Type: journal article
Area: technical or scientific
Ttl2: Molecular Carcinogen
Date: 1990
Volm: 3(6)
Page: 330-334

601.
Auth: Mullis, K.//Faloona, F.//Scharf, S.//Saiki, R.//Horn, G.//Erlich, H.
Affl: Cetus Corporation, Dept. of Human Genetics, Emeryville, CA.
Ttl1: Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction
Type: journal article
Area: technical or scientific
Ttl2: Cold Spring Harbor Symposia on Quantitative Biology
Date: 1986
Volm: L1
Page: 263-273

602.
Auth: Mullis, Kary B.//Erlich, Henry A.//Arnheim, Norman//Horn, Glenn T.//Saiki, Randall K.//Scharf, Stephen J.
Affl: Cetus Corporation, Emeryville, CA.
Ttl1: Process for amplifying, detecting and/or cloning nucleic acid sequences
Type: Public document
Area: technical or scientific
Ttl2: United States Patent
Date: July 28, 1987
Volm: Patent #4,683,195
Page: Application #828,144

603.
Auth: Naito, E.//Dewa, K.//Yamanouchi, H.//Mitani, K.//Kominami, R.
Ttl1: DNA fingerprinting by means of a nonradioactive probe of sulfonated DNA
Type: journal article
Area: technical or scientific
Ttl2: Nippon Hoigaku Zasshi
Date: June 1989
Volm: 43(3)
Page: 243-245

604.
Auth: Nakamura, Y.//Culver, M.//Gill, J.//O'Connell, P.//Leppert, M.//Lathrop, G. M.//Lalouel, J. M.//White, R.
Affl: The Howard Hughes Medical Institute, University of Utah Medical School, Salt Lake City, UT.
Ttl1: Isolation and mapping of a polymorphic DNA sequence pMLJ14 on chromosome 14 [D14S13]
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: January 11, 1988
Volm: 16(1)
Page: 381

605.
Auth: Nakamura, Y.//Gillilan, S.//O'Connell, P.//Leppert, M.//Lathrop, G. M.//Lalouel, J. M.//White, R.
Affl: The Howard Hughes Medical Institute, University of Utah Medical School, Salt Lake City, UT
Ttl1: Isolation and mapping of a polymorphic DNA sequence pYNH24 on chromosome 2 (D2S44)
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: December 10, 1987
Volm: 15(23)
Page: 10073

606.
Auth: Nakamura, Y.//Leppert, M.//O'Connell, P.//Wolff, R.//Holm, T.//Culver, M.//Martin, C.//Fujimoto, E.//Hoff, M.//Kumlin, E.//White, R.
Affl: Howard Hughes Medical Institute, University of Utah School of Medicine, Salt Lake City, Utah 84132
Ttl1: Variable number of tandem repeat (VNTR) markers for human gene mapping
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: March 27, 1987
Volm: 235(4796)
Page: 1616-1622

607.
Auth: Nakamura, Yusuke//Carlson, Mary//Krapcho, Karen//Kanamori, Masao//White, Ray
Affl: Howard Hughes Medical Institute, University of Utah Health Sciences Center, Salt Lake City, Utah 84132
Ttl1: New approach for isolation of VNTR markers
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: December 1988
Volm: 43(6)
Page: 854-859

608.
Auth: Nanda, [illegible]ajit//Zischler, Hans//Epplen, Conny//Guttenbach, Martina//Schmid Michael
Affl: University of Wurzburg; Max-Planck-Institute for Psychiatry, Martinsried, Germany
Ttl1: Chromosomal organization of simple repeated DNA sequences used for DNA fingerprinting
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 193-203

609.
Auth: Nata, M.//Yokoi, T.//Aoki, Y.//Sagisaka, K.//Hiraiwa, K.//Takatori, T.
Ttl1: Paternity test with single locus DNA probes
Type: journal article
Area: technical or scientific
Ttl2: Japan Journal Legal Medicine
Date: 1991
Volm: 45
Page: 138

610.
Type: book
Area: technical or scientific
BkAu: National Research Council
Ttl2: DNA Technology in Forensic Science
Plac: Washington, D. C.
Publ: National Academy Press
Date: April 16, 1992
Volm: ISBN 0-309-04587-8
Page: i-xiv, S1-S27, 1-1/8-14

611.
Auth: Neel, James V.//Satoh, Chiyoko//Smouse, Peter//Asakawa, Jun ichi//Takahashi, Norio//Goriki, Kazuaki//Fujita, Mikio//Kageoka, Takeshi//Hazama, Ryuji
Affl: University of Michigan Medical School, Ann Arbor; Radiation Effects Research Foundation, Hiroshima
Ttl1: Protein variants in Hiroshima and Nagasaki: Tale of Two Cities
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1988
Volm: 43
Page: 870-893

612.
Auth: Nelkin, Dorothy
Affl: Department of Sociology, NY University, N.Y., N.Y.
Ttl1: The social meaning of biological tests
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 13-23

613.
Auth: Neufeld, Peter J.
Affl: 56 Thomas Street, N.Y., N.Y. 10013
Ttl1: Admissibility of new or novel scientific evidence in criminal cases
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 73-87

614.
Auth: Neufeld, Peter J.//Colman, Neville
Affl: Fordham University School of Law; Mount Sinai Medical Center, New York City
Ttl1: When science takes the witness stand
Type: journal article
Area: technical or scientific
Ttl2: Scientific American
Date: May 1990
Volm: 262(5)
Page: 46-53

615.
Auth: Neufeld, Peter J.//Scheck, Barry C.
Ttl1: No: less than meets the eye
Type: journal article
Area: technical or scientific
Ttl2: ABA Journal
Date: September 1990
Page: 35

616.
Auth: Neuweiler, John//Ruvolo, Vivian//Baum, Howard//Grzeschik, Karl Heinz//Balazs, Ivan
Affl: Lifecodes Corporation, Valhalla, NY 10595
Ttl1: Isolation and characterization of a hypervariable region (D4S163) on chromosome 4
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: February 11, 1990
Volm: 18(3)
Page: 691

617.
Auth: Newmark, Peter
Ttl1: Biotechnology: DNA fingerprints go commerical
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: May 8-14, 1986
Volm: 321(6066)
Page: 104

618.
Auth: Newmark, Peter
Ttl1: Dispute over who should do DNA fingerprinting in murder hunt
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: Janaury 8-14, 1987
Volm: 325(7000)
Page: 97

619.
Auth: Newmark, Peter
Ttl1: DNA fingerprinting at a price at ICI's UK laboratory
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: June 18-24, 1987
Volm: 327(6123)
Page: 548

620.
Auth: Newmark, Peter
Ttl1: DNA fingerprinting to be used for British immigrants?
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: February 18, 1988
Volm: 331(6157)
Page: 556

621.
Auth: Nichols, Richard A.//Balding, David J.
Affl: Queen Mary and Estfield College, University of London, U.K.
Ttl1: Effects of population structure on DNA fingerprint analysis in forensic science
Type: journal article
Area: technical or scientific
Ttl2: Heredity
Date: 1991
Volm: 66
Page: 297-302

622.
Auth: no author given
Affl: National AudioVisual Center
Ttl1: Focus on DNA Technology
Type: literature on video tapes
Area: technical or scientific
Plac: 8700 Edgeworth Drive, Capitol Heights, MD 20743
Volm: video tapes

623.
Auth: Noppinger, K.//Duncan, G.//Ferraro, D.//Watson, S.//Ban, J.
Affl: Broward County Sheriff's Crime Laboratory, DNA Unit, Ft. Lauderdale, FL.
Ttl1: Evaluation of DNA probe removal from nylon membrane
Type: journal article
Area: technical or scientific
Ttl2: Biotechniques
Date: October 1992
Volm: 13(4)
Page: 572-575

624.
Auth: Norman, Colin
Ttl1: Caution urged on DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: August 18, 1989
Volm: 245(4919)
Page: 699

625.
Auth: Norman, Colin
Ttl1: Maine case deals blow to DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: December 22, 1989
Volm: 246(4937)
Page: 1556-1558

626.
Auth: Norman, Jeffrey A.
Ttl1: DNA fingerprinting: is it ready for trial?
Type: journal article
Area: legal
Ttl2: University of Miami Law Review
Plac: United States
Date: September 1990
Volm: 45(1)
Page: 243-259

627.
Auth: Nurnberg, P.//Barth, I.//Fuhrmann, E.//Lenzner, C.//Losanova, T.//Peters, C.//Poche, H.//Thiel, G.
Ttl1: Monitoring genomic alterations with a panel of oligonucleotide probes specific for various simple repeat motifs
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 186-192

628.
Auth: Nurnberg, P.//Roewer, L.//Neitzel, H.//Sperling, K.//Popperl, A.//Hundrieser, J.//Poche, H.//Epplen, C.//Zischler, H.//Epplen, J. T.
Affl: Institut fur Medizinische Genetik des Bereichs Medizin der Charite, Humboldt Universitat, Berlin, German Democratic Republic
Ttl1: DNA fingerprinting with the oligonucleotide probe (CAC)5/(GTG)5: somatic stability and germline mutations
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: December 1989
Volm: 84(1)
Page: 75-78

629.
Auth: Nybom, H.
Affl: Swedish University of Agricultural Sciences, Fjakestadsvagen, Kristianstad, Sweden
Ttl1: Applications of DNA fingerprinting in plant breeding
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 294-311

630.
Auth: Odelberg, S. J.//Demers, D. B.//Westin, E. H.//Hossaini, A. A.
Affl: University of Utah Medical Center, Salt Lake City, Utah
Ttl1: Establishing paternity using minisatellite DNA probes when the putative father is unavailable for testing
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: July 1988
Volm: 33(4)
Page: 921-928

631.
Auth: Odelberg, S. J.//Plaetke, R.//Eldridge, J. R.//Ballard, L.//O'Connell, P.//Nakamura, Y.//Leppert, M.//Lalouel, J. M.//White, R.
Affl: Department of Human Genetics, University of Utah Medical Center, Salt Lake City, 84132
Ttl1: Characterization of eight VNTR loci by agarose gel electrophoresis
Type: journal article
Area: technical or scientific
Ttl2: Genomics
Date: November 1989
Volm: 5(4)
Page: 915-924

632.
Auth: Odelberg, Shannon J.//White, Ray
Affl: Department of Human Genetics, The University of Utah School of Medicine, Salt Lake City, Utah
Ttl1: Repetitive DNA: molecular structure, polymorphisms and forensic applications
Type: book chapter
Area: technical or scientific
BkAu: Lee, Henry C.//Gaensslen, R. E.
Ttl2: DNA and Other Polymorphisms in Forensic Science
Date: 1990
Page: 26-44

633.
Auth: Odelberg, Shannon J.//White Ray
Affl: University of Utah Medical Center, Salt Lake City, Utah
Ttl1: Tandemly repeated DNA and its application in forensic biology
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 257-263

634.
Auth: Office of Technology Assessment (OTA)
Affl: U.S. Congress
Ttl1: Genetic Witness: Forensic Uses of DNA Tests
Type: Report
Area: technical or scientific
Plac: U.S. Government Printing Office, Washington, DC
Publ: OTA Publications Office
Date: 1990
Volm: 052 003 01203 1

635.
Type: book
Area: technical or scientific
BkAu: Office of Technology Assessment, Congress of the United States
Ttl2: Genetic Witness: Forensic Uses of DNA Tests
Plac: Washington, D.C.
Publ: U.S. Government Printing Office
Date: July 1990
Volm: OTA BA 438
Page: i-v, 1-178

636.
Auth: Ogata, M.//Mattern, R.//Schneider, P. M.//Schacker, U.//Kaufmann, T.//Rittner, C.
Affl: Institute of Legal Medicine, Johannes Gutenberg University, Mainz, Federal Republic of Germany
Ttl1: Quantitative and qualitative analysis of DNA extracted from postmortem muscle tissues
Type: journal article
Area: technical or scientific
Ttl2: Zeitschrift fur Rechtsmedizin
Date: 1990
Volm: 103(6)
Page: 397-406

637.
Auth: Ohno, S.//Yomo, T.
Ttl1: The grammatical rule for all DNA: junk and coding sequences
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 103-108

638.
Auth: Pakkala, Seppo//Helminen, Paivi//Saarinen, Ulla M.//Alitalo, Riitta//Peltonen, Leena
Affl: Transplantation Laboratory, University of Helsinki, Finland
Ttl1: Differences in DNA fingerprints between remission and relapse in childhood acute lymphoblastic leukemia
Type: journal article
Area: technical or scientific
Ttl2: Leukemia Research
Date: 1988
Volm: 12(9)
Page: 757-762

639.
Auth: Pakkala, Seppo//Helminen, Paivi//Ruutu, Tapani//Saarinen, Ulla M.//Peltonen, Leena
Affl: Transplantation Laboratory, University of Helsinki, Finland
Ttl1: New molecular marker in AML: DNA fingerprint differences between leukemic phase and remission in acute myeloid leukemia
Type: journal article
Area: technical or scientific
Ttl2: Leukemia Research
Date: 1989
Volm: 13(19)
Page: 907-913

640.
Ttl1: Paper Symposium: DNA Fingerprinting
Type: Journal
Area: technical or scientific
BkAu: Epplen, Jorg T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 103-231

641.
Auth: Parkin, B. H.
Affl: Metropolitan Police Forensic Science Laboratory, London, England
Ttl1: DNA analysis in the metropolitan police forensic science laboratory
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 163-167

642.
Auth: Parkin, David T.
Affl: University of Nottingham, Nottingham, UK
Ttl1: Teach yourself fingerprinting. "DNA Fingerprinting: An Introduction"
Type: book review
Area: technical or scientific
Ttl2: Trends in Genetics
Date: September 1991
Volm: 7(9)
Page: 305-306

643.
Auth: Pascal, O.//Aubert, D.//Gilbert E.//Moisan J.P.
Affl: Ctr Hospital Reg & Univ. Nantes, Biol Molecular Laboratory, Nantes, FR.
Ttl1: Sexing of forensic samples using PCR
Type: journal article
Area: technical or scientific
Ttl2: International Journal Legal Medicine
Date: 1991
Volm: 104(4)
Page: 250

644.
Auth: Pascal, O.//Aubert, D.//Henderson, R.//Moisan J. P.
Affl: Molecular Biology Library, University Hospital of Nantes, Nantes, France
Ttl1: Sexing of forensic samples by PCR
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 177-178

645.
Auth: Pascali, V. L.//d'Aloja, E.//Dubosz, M.//Pescarmona, M.
Affl: Universita Cattolica del S. Cuore, Largo, F. Vito, Roma Italy
Ttl1: Estimating allele frequencies of hypervariable DNA systems
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: 1991
Volm: 51
Page: 273-280

646.
Auth: Pascali, V. L.//d'Aloja, E.//Dobosz, M.//Spinella, A.//Quaresima, A.
Affl: Universita Catolica del S. Cuore; Servizio di Polizia Scientifica, Rome, Italy
Ttl1: Identification of genomic digests from several somatic sources by multi-locus and locus-specific DNA profiles. A survey of some criminal cases worked out in the UCSC and SPS Laboratories
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 215-216

647.
Auth: Patton, Stephen M.
Ttl1: DNA fingerprinting
Type: journal article
Area: legal
Ttl2: Harvard Journal of Law & Technology
Case: People v. Castro 545
Cort: N.Y.S.2d 985 (App. Div. 1989)
Plac: New York (State)
Date: Spring 1990
Volm: 3
Page: 223-240

648.
Auth: Pearsall, Anthony
Ttl1: DNA printing: the unexamined "witness" in criminal trials
Type: journal article
Area: legal
Ttl2: California Law Review
Date: March 1, 1989
Volm: 77
Page: 665

649.
Auth: Pemberton, Josephine//Amos, Bill
Affl: Dept. of Genetics, University of Cambridge, Downing Street, Cambridge, UK.
Ttl1: DNA fingerprinting: a new dimension
Type: journal article
Area: technical or scientific
Ttl2: Trends In Genetics
Date: April 1990
Volm: 6(4)
Page: 101-103

650.
Auth: Pena, S. D. J.//Macedo, A. M.//Gontijo, N. F.//Medeiros, A. M.//Ribeiro, J. C. C.
Ttl1: DNA bioprints: simple nonisotopic DNA fingerprints with biotinylated probes
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoreis
Date: February 3, 1991
Volm: 12(2-3)
Page: 146-152

651.
Type: journal
Area: technical or scientific
BkAu: Perkin Elmer Cetus
Ttl2: PCR Bibliography
Plac: 761 Main Avenue, Norwalk, CT. 06859-0251
Publ: Perkin Elmer Cetus
Date: June 1990
Volm: 1(5)
Page: 1-128

652.
Auth: Peters, C.//Schneider, V.//Epplen, J. T.//Poche, H.
Affl: Institut fur Rechtsmedizin, Freie Universitat Berlin, Berlin, Max-Planck--Institut fur Psychiatrie, Martinsried, Germany
Ttl1: Individual-specific DNA fingerprinting in man using the oligonucleotide probe (GTG)5/(CAC)5
Type: journal article
Area: technical or scientific
Ttl2: European Journal of Clinical Chemistry and Clinical Biochemistry
Date: May 1991
Volm: 29(5)
Page: 321-325

653.
Auth: Peterson, Joseph L.
Affl: Department of Criminal Justice, University of Illinois at Chicago, Chicago, IL.
Ttl1: Impact of biological evidence on the adjudication of criminal cases: potential for DNA technology
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 55-70

654.
Auth: Petronzelli, Theresa//Schildkraut, Ira
Affl: New England Biolabs, 32 Tozer Road, Beverly, MA.01915
Ttl1: Altered specificity of restriction endonuclease HinfI
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: 1990
Volm: 18(12)
Page: 3666

655.
Auth: Petrosinelli, Jospeh G.
Ttl1: The admissibility of DNA typing: a new methodology
Type: journal article
Area: legal
Ttl2: Geo. Law Journal
Date: December 1990
Volm: 79
Page: 313

656.
Auth: Petrovich, Stephen C.
Ttl1: DNA typing: a rush to judgment
Type: journal article
Area: legal
Ttl2: Ga. Law Review
Date: 1990
Volm: 24
Page: 669

657.
Auth: Presley, L. A.//Adams, D. E.
Affl: FBI Laboratory, Washington, D.C.
Ttl1: Restriction fragment length polymorphism (RFLP) analysis of biological stain mixtures
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 219-222

658.
Auth: Prosniak, M. I.//Kartel, N. A.//Perebitiuk, A. N.//Limborskaia, S. A.//Ryskov, A. P.
Ttl1: Nonisotope variant of genomic fingerprinting based on the phage M13 DNA in kinship studies in man
Type: journal article
Area: technical or scientific
Ttl2: Genetika
Date: January 1990
Volm: 26(1)
Page: 134-137

659.
Auth: Proudfoot, Nicholas J.//Gil, Anna//Maniatis, Tom
Affl: University of Oxford, Oxford, England; Harvard University, Cambridge, Mass.
Ttl1: The structure of the human zeta-globin gene and a closely linked, nearly identical pseudogene
Type: journal article
Area: technical or scientific
Ttl2: Cell
Date: December 1982
Volm: 31(3)
Page: 553-563

660.
Auth: Purcell, W. Anthony
Ttl1: Criminal procedure -- match-game 1990's: the admissibility of DNA profiling
Type: journal article
Area: legal
Ttl2: Campbell Law Review
Case: State v. Pennington
Date: Spring 1991
Volm: 13
Page: 209

661.
Auth: Rand, S.//Wiegnand, P.//Brinkmann, B.
Ttl1: Problems associated with the DNA analysis of stains
Type: journal article
Area: technical or scientific
Ttl2: International Journal of Legal Medicine
Date: 1991
Volm: 104
Page: 293

662.
Auth: Rankin, John J.
Ttl1: DNA fingerprinting
Type: journal article
Area: legal
Ttl2: Journal of the Law Society of Scotland
Plac: Great Britain
Date: April 1988
Volm: 33(4)
Page: 124-126

663.
Auth: Rassman, K.//Schlotterer, C.//Tautz, D.
Ttl1: Isolation of simple sequence loci for use in polymerase chain reaction based DNA fingerprinting
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 113-118

664.
Auth: Rath, D[redacted]//Merril, C. R.
Affl: National Institute of Medical Health, Washington, D.C. 20032
Ttl1: Mitochondrial DNA and its forensic potential
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the Internatinal Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 113-120

665.
Auth: Rathbun, Emmet A.
Affl: National Crime Information Center, FBI, Washington, D.C.
Ttl1: The NCIC experience
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 327-333

666.
Auth: Reed, T. Edward
Affl: University of Toronto, Toronto, Canada
Ttl1: Caucasian genes in American Negroes. Measurement of non-African ancestry is difficult, but it is worthwhile for several genetic reasons.
Type: journal article
Area: technical or scientific
Ttl2: Science
Volm: 165
Page: 762-765

667.
Auth: Reeder, Dennis J.//Kline, Margaret C.
Affl: National Institute of Standards and Technology, Gaithersburg, MD
Ttl1: Questionable use of DNA standards for sizing PCR products in small-format gels
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 201

668.
Auth: Reeve, Husdon K.//Westneat, David F.//Noon, William A.//Sherman, Paul W.//Aquadro, Charles F.
Affl: Section of Neurobiology and Behavior, Cornell University, Ithaca, NY 14853
Ttl1: DNA fingerprinting reveals high levels of inbreeding in colonies of the eusocial naked mole-rat
Type: journal acticle
Area: technical or scientific
Ttl2: Proceedings National Academy of Sciences USA
Date: April 1990
Volm: 87(7)
Page: 2496-2500

669.
Auth: Reilly, Philip
Affl: University Affiliated Program Shriver Center for Mental Retardation, Inc., Waltham, Massachusetts
Ttl1: Reflections on the use of DNA forsensic science and privacy issues
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 43-54

670.
Auth: Renskers, Sally E.
Ttl1: Trial by certainty: implications of genetic "DNA fingerprints"
Type: journal article
Area: legal
Ttl2: Emory Law Journal
Plac: United States
Date: Winter 1990
Volm: 39(1)
Page: 309-346

671.
Auth: Reynolds, R.//von Beroldingen, C.//Sensabaugh, G. F.
Affl: University of California, Berkeley, CA.
Ttl1: Effects of DNA degradation on amplification by the polymerase chain reaction
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 257

672.
Auth: Reynolds, Rebecca
Affl: Cetus Corporation, Emeryville, CA.
Ttl1: The development of new PCR markers for the analysis of forensic evidence
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 132-133

673.
Auth: Reynolds, Rebecca//Sensabaugh, George//Blake, Edwards
Affl: School of Public Health, Univ. of Ca., Berkeley, CA.; Forensic Science Assoc. Richmond, Ca.
Ttl1: Analysis of genetic markers in forensic DNA samples using the polymerase chain reaction
Type: journal article
Area: technical or scientific
Ttl2: Analytical Chemistry
Date: January 1, 1991
Volm: 63(1)
Page: 2-15

674.
Auth: Richards, C. S.//Watkins, S. C.//Hoffman, E. P.//Schneider, N. R.//Milsark, I. W.//Katz, K. S.//Cook, J. D.//Kunkel, L. M.//Cortada, J. M.
Affl: GeneScreen, Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas 75207
Ttl1: Skewed X inactivation in a female MZ twin results in Duchenne muscular dystrophy
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: April 1990
Volm: 46(4)
Page: 672-681

675.
Auth: Richards, Ronald J.
Ttl1: DNA fingerprinting and paternity testing
Type: journal article
Area: legal
Ttl2: U.C. Davis Law Review
Plac: United States
Date: Winter 1989
Volm: 22(2)
Page: 609-651

676.
Auth: Ridge, S. A.//Worwood, M.
Affl: Dept. of Haematology, Univ. of Wales College of Medicine, Cardiff, UK
Ttl1: Comments on differences in DNA fingerprints between remission and relapse in childhood acute lymphoblastic leukemia
Type: letter to editor
Area: technical or scientific
Ttl2: Leukemia Research
Date: 1989
Volm: 13(6)
Page: 511-512

677.
Auth: Riley, John P.
Affl: Forensic Science Research and Training Section, FBI Academy, Quantico, VA
Ttl1: Radiation aspects of DNA analysis
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: 1988
Volm: 15 (supplement)
Page: 12

678.
Auth: Risch, Neil J.//Devlin, B.
Affl: Department of Genetics, Yale University, New Haven, CT 06510
Ttl1: On the probability of matching DNA fingerprints
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: February 7, 1992
Volm: 255(2)
Page: 717-720

679.
Auth: Ritter, H.
Affl: Institut fur Anthropologie und Humangenetik der Universitat, Wilhelmstrasse 27, W-7400 Tubingen, Fed. Republic of Germany
Ttl1: DNA fingerprinting: a further note on mutation rates
Type: letter to editor
Area: technical or scientific
Ttl2: Human Genetics
Date: May 1992
Volm: 89(3)
Page: 363

680.
Auth: Rittner, C.//Prager-Eberle, M.//Schneider, P. M.
Affl: Institute of Legal Medicine, Mainz, Germany
Ttl1: Application of HLA DQa typing in forensic casework
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 160-162

681.
Auth: Robertson, J.//Ziegle, J.//Krnoick, M.//Madden, D.//Budowle, B.
Affl: Forensic Science Research and Training Section, FBI Laboratory, Quantico, Va.; Applied Biosystems, Inc., Foster City, CA.
Ttl1: Genetic typing using automated electrophoresis and fluorescence detection
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 391-398

682.
Auth: Robertson, James M.//Kronick, Mel
Affl: Applied Biosystems, Incorporated, Foster City, CA.
Ttl1: Automating DNA fingerprinting - a multifluorophore approach to reduce chance in match calling
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 179-182

683.
Auth: Roberts, Leslie
Affl: Writer - Science Magazine
Ttl1: DNA fingerprinting: Academy reports
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: April 17, 1992
Volm: 256
Page: 300-301

684.
Auth: Roberts, Leslie
Affl: Writer - Science magazine
Ttl1: Fight erupts over DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: December 20, 1991
Volm: 254
Page: 1721-1723

685.
Auth: Roberts, Leslie
Affl: Writer-Science Magazine
Ttl1: Science in Court: A culture clash
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: August 7, 1992
Volm: 257
Page: 732-736

686.
Auth: Roewer, L.//Nurnberg, P.//Fuhrmann, E.//Rose, M.//Prokop, O.//Epplen J.T.
Affl: Institut fur Gerichtliche Medizin der Humboldt-Universitat zu Berlin
Ttl1: Stain analysis using oligonucleotide probes specific for simple repetitive DNA sequences
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: 1990
Volm: 47
Page: 59-70

687.
Auth: Roewer, L.//Rieb, O.//Prokop, O.
Ttl1: Hybridization and polymerase chain reaction amplification of simple repeated DNA sequences for the anlaysis of forensic stains
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 181-185

688.
Auth: Roewer, L.//Rose, M.//Semm, K.//Correns, A.//Epplen, J. T.
Ttl1: Typing of stored, hemolyzed blood samples by DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Arch Kriminol
Date: September-October 1989
Volm: 184(3-4)
Page: 103-107

689.
Auth: Ronningen, Kjersti S.//Spurkland, Anne//Markussen, Gunnar//Iwe, Thomas//-/Vartdal, Frode//Thorsby, Erik
Affl: Institute of Transplantation Immunology, The National Hospital, Oslo, Norway
Ttl1: Distribution of HLA class II alleles among Norwegian caucasians
Type: journal article
Area: technical or scientific
Ttl2: Human Immunology
Date: 1990
Volm: 29
Page: 275-281

690.
Auth: Rose, Michael//Nagai, Tatsuo//Sato, Miwako//Prokop, Eberhard
Affl: Institute of Forensic Medicine, School of Medicine (Charite), Humboldt University Berlin, German Democratic Republic
Ttl1: DNA fingerprinting in a random sample of Japanese population with the minisatellite probe MZ 1.3
Type: journal article
Area: technical or scientific
Ttl2: Experimental and Clinical Immunogenetics
Date: 1990
Volm: 7(2)
Page: 136-140

691.
Auth: Rose, Stanley D.//Keith, Tim P.
Affl: DNA Products and Services; Human Genetics Department, Bedford, Mass.
Ttl1: Standardization of systems: essential or desirable?
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 319-326

692.
Auth: Ross, Alastair M.//Harding, Harry W. J.
Affl: Forensic Science Centre, Adelaide, South Australia
Ttl1: DNA typing and forensic science
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: 1989
Volm: 41
Page: 197-203

693.
Auth: Rowe, Walter F.
Affl: The George Washington University, Washington, D.C.
Ttl1: DNA Testing not ready for court?! A tale of two surveys
Type: letter to editor
Area: journal article
Ttl2: Journal of Forensic Sciences
Page: 803-805

694.
Type: book
Area: technical or scientific
BkAu: Roychoudhury, Arun K.//Nei, Masatoshi
Ttl2: Human Polymorphic Genes. World Distribution
Plac: 200 Madison Ave., New York, N.Y. 10016
Publ: Oxford University Press
Date: 1988
Volm: ISBN 0-19-505123-8
Page: i-viii,1-393

695.
Auth: Royle, Nicole J.//Clarkson, Richard E.//Wong, Zilla//Jeffreys, Alec J.
Affl: Department of Genetics, Univeristy of Leicester, UK
Ttl1: Clustering of hypervariable minisatellites in the proterminal regions of human autosomes
Type: journal article
Area: technical or scientific
Ttl2: Genomics
Date: November 1988
Volm: 3(4)
Page: 352-360

696.
Auth: Ruano, G.//Pagliaro, E. M.//Schwartz, T. R.//Lamy, K.//Messina, D.//Gaensslen, R. E.//Lee, H. C.
Affl: Conn. State Police Forensic Science Lab.; Yale Univ. School of Med.; Univ. of New Haven
Ttl1: Heat-soaked PCR: an efficient method for DNA amplification with applications to forensic analysis
Type: journal article
Area: technical or scientific
Ttl2: BioTechniques
Date: August 1992
Volm: 13(2)
Page: 266-274

697.
Auth: Ryskov, A. P.//Dzhincharadze, A. G.//Prosnik, M. I.//Ivanov, P. L./-/Lomborskaia, S. A.
Ttl1: Genomic fingerprints of organisms from different taxonomic groups: the use of phage M13 DNA as a hybridization probe
Type: journal article
Area: technical or scientific
Ttl2: Genetika
Date: Feburary 1988
Volm: 24(2)
Page: 227-238

698.
Auth: Ryskov, A. P.//Gaizov, T. K.//Alimov, A. M.//Romanova, E. A.
Ttl1: Genomic fingerprinting: new possibilities in determining the species identity of Brucella
Type: journal article
Area: technical or scientific
Ttl2: Genetika
Date: January 1990
Volm: 26(1)
Page: 130-133

699.
Auth: Ryskov, A. P.//Jincharadze, A. G.//Prosnyak, M. I.//Ivanov, P. L./-/Limborskaya, S. A.
Affl: Institute of Molecular Biology, USSR Academy of Sciences, Moscow
Ttl1: M13 phage DNA as a universal marker for DNA fingerprinting of animals, plants and microorganisms
Type: journal article
Area: technical or scientific
Ttl2: FEBS Letter
Date: June 20, 1988
Volm: 233(2)
Page: 388-392

700.
Auth: Ryskov, A. P.//Tokarshaia, O. N.//Verbovaia, L. V.//Dzhincharadze, A. G.//Gintsburg, A. L.
Ttl1: Genomic fingerprinting of microorganisms: its use as a hybridization probe of phage M13 DNA
Type: journal article
Area: technical or scientific
Ttl2: Genetika
Date: July 1988
Volm: 24(7)
Page: 1310-1313

701.
Auth: Saiki, Randall K.//Bugawan, Teodorica L.//Horn, Glenn T.//Mullis, Kary B.//Erlich, Henry A.
Affl: Dept. of Human Genetics, Cetus Corp., Emeryville, CA.
Ttl1: Analysis of enzymatically amplified B-globin and HLA-DQa DNA with allele-specific oligonucleotide probes
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: November 13, 1986
Volm: 324
Page: 163-166

702.
Auth: Saiki, Randall K.//Gelfand, David H.//Stoffel, Susanne//Scharf, Stephen J.//Higuchi, Russell//Horn, Glenn T.//Mullis, Kary B.//Erlich, Henry A.
Affl: Cetus Corporation, Department of Human Genetics, Emeryville, CA. 94608
Ttl1: Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: January 29, 1988
Volm: 239
Page: 487-491

703.
Auth: Saiki, Randall K.//Walsh, P. Sean//Levenson, Corey H.//Erlich, Henry A.
Affl: Depart. of Human Genetics and Chemistry, Cetus Corp., Emeryville, CA.
Ttl1: Genetic analysis of amplified DNA with immobilized sequence specific oligonucleotide probes
Type: journal article
Area: technical of scientific
Ttl2: Proceedings National Academy Science USA
Date: August 1989
Volm: 86
Page: 6230-6234

704.
Auth: Sajantila, A.//Helminen, P.//Syvanen, A. C.//Ehnholm, C.//Peltonen, L.
Affl: National Public Health Institute, Helsinki, Finland
Ttl1: Forensic applications of DNA amplification (PCR)
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 173-176

705.
Auth: Sajantila, A.//Helminen, P.//Ehnholm, C.//Peltonen, L.//Saukko, P.//Pakkala, S.
Affl: Nat'l Public Health Institute; Univ. of Oulu; Univ. of Helsinki, Helsinki, Finland
Ttl1: Identification of individuals with multi and single locus probes: the effect of genetic isolation, autolysis and somatic mutations
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.,
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 243-246

706.
Auth: Sajantila, Antti//Budowle, Bruce//Strom, Marjanne//Johnsson, Vivian//Lukka, Matti//Peltonen, Leena//Ehnholm, Christian
Affl: National Public Health Institute; Dept. of Forensic Medicine, Univ. of Helsinki; FBI Academy, Quantico, VA.
Ttl1: PCR amplification of alleles at the D1S80 locus: Comparison of a Finnish and a North American Caucasian Population sample, and forensic casework evaluation
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: April 1992
Volm: 50(4)
Page: 816-825

707.
Auth: Sajantila, Antti//Makkonen, Kaisa//Ehnholm, Christian//Peltonen, Leena
Affl: Department of Forensic Medicine, Helsinki, Finland
Ttl1: DNA profiling in a genetically isolated population using three hypervariable DNA markers
Type: journal article
Area: technical or scientific
Ttl2: Human Heredity
Volm: (in press)

708.
Auth: Sajantila, Antti//Strom, Marjanne//Budowle, Bruce//Tienari, P. J.//Ehnholm, Christian//Peltonen, Leena
Affl: National Public Health Institute, Mannerheimintie, Helsinki, Finland
Ttl1: The distribution of the HLA-DQalpha alleles and genotypes in the Finnish population as determined by the use of DNA amplification and allele specific oligonucleotides
Type: journal article
Area: technical or scientific
Ttl2: International Journal Legal Medicine
Date: 1991
Volm: 104(4)
Page: 181

709.
Auth: Saji, Fumitaka//Tokugawa, Yoshihiro//Kimura, Tadashi//Kamiura, Shoji//-/Nobunaga, Toshikatsu//Azuma, Chihiro//Tanizawa, Osamu
Affl: Department of Obstetrics and Gynecology, Osaka University Medical School, Fukushima, Fukushima-ku, Osaka, Japan
Ttl1: A new approach using DNA fingerprinting for the determination of androgenesis as a cause of hydatidiform mole
Type: journal article
Area: technical or scientific
Ttl2: Placenta
Date: July-August 1989
Volm: 10(4)
Page: 399-405

710.
Auth: Samani, [redacted]esh J.//Swales, John D.//Jeffreys, Alex J.//Morton, David B./-
/Naftilan, Alan J.//Lindpaintner, Klaus//Ganten, Detlev//Brammar, W. J.
Affl: Department of Medicine, University of Leicester, U.K.
Ttl1: DNA fingerprinting of spontaneously hypertensive and Wistar-Kyoto rats: implications for hypertension research
Type: journal article
Area: technical or scientific
Ttl2: Journal Hypertension
Date: October 1989
Volm: 7(10)
Page: 809-816

711.
Auth: Schacker, U.//Schneider, P. M.//Holtkamp, B.//Bohnke, E.//Fimmers, R./-
/Sonneborn, H. H.//Rittner, C.
Affl: Institute of Legal Medicine, Johannes Gutenberg University, Mainz, Federal Republic of Germany
Ttl1: Isolation of the DNA minisatellite probe MZ 1.3 and its application to DNA fingerprinting analysis
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: Feburary 1990
Volm: 44(2-3)
Page: 209-224

712.
Auth: Schafer, Renate//Zischler, Hans//Epplen, Jorg T.
Affl: Max-Planck-Institut fur Psychiatrie, Martinsried, Federal Republic of Germany
Ttl1: (CAC)5, a very informative oligonucleotide probe for DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: June 10, 1988
Volm: 16(11)
Page: 5196

713.
Auth: Schafer, Renata//Zischler, Hans//Epplen, Jorg T.
Affl: Max-Planck-Institute fur Psychiatrie, Martinsried, Federal Republic of Germany
Ttl1: DNA fingerprinting using non-radioactive oligonucleotide probes specific for simple repeats
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: October 11, 1988
Volm: 16(19)
Page: 9344

[redacted].
Auth: Schafer, Renate//Zischler, Hans//Birsner, Uli//Becker, Andrea//Epplen, Jorg T.
Affl: Max-Plank-Institut fur Psychiatrie, Martinsried, Federal Republic of Germany
Ttl1: Optimized oligonucleotide probes for DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Electrophoresis
Date: August 1988
Volm: 9(8)
Page: 369-374

715.
Auth: Scharf, Stephen J.//Horn, Glenn T.//Erlich, Henry A.
Affl: Cetus Corporation, Dept. of Human Genetics, Emeryville, CA.
Ttl1: Direct cloning and sequence analysis of enzymatically amplified genomic sequences
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: September 5, 1986
Volm: 233
Page: 1076-1078

716.
Auth: Scheck, Barry
Affl: Director, Clinical Education, Benjamin N. Cardozo School of Law
Type: letter of response
Area: technical or scientific
Ttl2: Professional Ethics Report
Plac: 1333 H Street, NW, Washington, DC 20005
Publ: American Association for the Advancement of Science
Date: Spring 1992
Volm: V(2)
Page: 7-8

717.
Auth: Scheithauer, R.//Weisser, H. J.
Ttl1: DNA profiling of bloodstains on linen pretreated with remedies used for cleaning and maintaining clothes
Type: journal article
Area: technical or scientific
Ttl2: International Journal Legal Medicine
Date: 1991
Volm: 104
Page: 273

718.
Auth: Schelling, C. P.//Clavadetscher, E.//Scharer, E.//Thomann, P. E.//Kuenzle, C. C.//Hubscher, U.
Affl: University of Zurich, Winterhurerstr, Zurich, Switzerland
Ttl1: Two dimensional DNA fingerprinting in animals
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 283-293

719.
Auth: Schmeck[illegible], Harold M.
Affl: New York Times
Ttl1: DNA evidence faulted in a Bronx murder case
Type: newspaper
Area: lay press
Ttl2: The New York Times
Plac: New York
Date: May 25, 1989
Page: 14

720.
Auth: Schmitter, H.//Herrmann, S.//Pflug, W.
Affl: Bundeskriminalamt; Polizeitechnische Untersuchungsstelle; Landeskriminalamt, Federal Republic of Germany
Ttl1: The use of DNA polymorphisms in the police laboratories of the Federal Republic of Germany
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 169-172

721.
Auth: Schneider, P. M.//Fimmers, R.//Woodroffe, S.//Werrett, D. J.//et al.
Affl: Institut of Legal Medicine, University of Mainz (F.R.G.)
Ttl1: Report of a European collaborative exercise comparing DNA typing results using a single locus VNTR probe
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: 1991
Volm: 49
Page: 1-15

722.
Auth: Schwartz, C. E.//Brown, A. M.//der Kaloustain, V. M.//McGill, J. J.//Saul, R. A.
Affl: Greenwood Genetic Center, S.C.; Montreal Children's Hospital, Montreal, Canada; Royal Children's Hospital, Australia
Ttl1: DNA fingerprinting: the utilization of minisatellite probes to detect a somatic mutation in the Proteus syndrome
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 95-105

723.
Auth: Schwartz, Ted R.//Schwartz, Elayne A.//Kobilinsky. Laura//McNally, Lorah/-/Kobilinsky, Lawrence
Affl: Lifecodes, Corp.; Union County Prosecutor's Office, Westfield,NJ; City University of New York, NY
Ttl1: Characterization of deoxyribonucleic acid (DNA) obtained from teeth subjected to various environmental conditions
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: July 1991
Volm: 36(4)
Page: 979-990

724.
Auth: Sealey, P. G.//Southern, E. M.
Type: book
Area: general reference
BkAu: Rickwood, D.//Hames
Ttl2: Gel Electrophoresis of DNA, In: Gel Electrophoresis of Nucleic Acids, A Practical Approach
Plac: Washington, D. C.
Publ: IRL Press
Date: 1987
Page: 39-76

725.
Auth: Sensabaugh, G.//Crim, D.//von Beroldingen, C. S.
Affl: School of Public Health, Univ. of CA., Berkeley; Oregon State Police Crime Detection Laboratory, Portland, Oregon
Ttl1: The polymerase chain reaction: application to the analysis of biological evidence
Type: book chapter
Area: technical or scientific
BkAu: Farley, Mark A.//Harrington, James J.
Ttl2: Forensic DNA Technology
Date: 1991
Page: 63-82

726.
Auth: Sensabaugh, George F.
Affl: University of California, Berkeley, CA.
Ttl1: Consequences of nucleotide misincorporation during the polymerase chain reaction
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 263-264

727.
Auth: Sensabaugh, George
Affl: School of Public Health, University of California, Berkeley, CA.
Ttl1: Forensic biology - is recombinant DNA technology in its future?
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science
Date: April 1986
Volm: 31(2)
Page: 393-396

728.
Auth: Shaler, Robert C.
Affl: Director of Forensic Training and Technical Support, Lifecodes Corp.
Ttl1: DNA-Print(tm) identification test
Type: open letter
Area: technical or scientific
Plac: Valhalla, NY
Date: August 21, 1989
Page: 1-23

729.
Auth: Shapiro, E. Donald
Ttl1: Dangers of DNA: it ain't just fingerprints
Type: journal article
Area: legal
Ttl2: New York Law Journal
Plac: United States
Date: January 23, 1990
Volm: 203(15)
Page: 1

730.
Auth: Shapiro, Martin M.
Affl: Department of Psychology, Emory University, Atlanta, GA.
Ttl1: Imprints on DNA fingerprints
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: September 12, 1991
Volm: 353
Page: 121-122

731.
Auth: Sheard, Brian
Affl: Metropolitan Police, Forensic Science Laboratory, London, UK
Ttl1: DNA profiling and the police
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: February 20, 1992
Volm: 355
Page: 667

732.
Auth: Sheridan, K. T.//Cotton, R. W.//Foster, I. M.
Affl: Cellmark Diagnostics, Germantown, Maryland; Baltimore County Health Department, Towson, Maryland
Ttl1: Effect of Nonoxynol-9 on recovery of DNA from sperm
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 235

733.
Auth: Sherman, Rorie
Affl: National Law Journal Staff Reporter
Ttl1: DNA is on trail yet again
Type: newspaper
Area: lay press
Ttl2: The National Law Journal
Date: March 16, 1992
Volm: 14(28)
Page: 1, 10

734.
Auth: Sherman, Rorie
Affl: National Law Journal staff reporter
Ttl1: DNA typing - NAS's final report is released
Type: journal article
Area: technical or scientific
Ttl2: The National Law Journal
Date: April 27, 1992
Page: 3, 35

735.
Auth: Sherman, Rorie
Affl: National Law Journal staff reporter
Ttl1: Genetic Testing Criticized - A draft report says DNA typing testimony should not be admitted.
Type: journal article
Area: legal
Ttl2: The National Law Journal
Date: April 20, 1992
Volm: 14(33)
Page: 1, 45-46

736.
Auth: Sherman, Rorie
Ttl1: Lawyers attacking test's reliability
Type: journal article
Area: legal
Ttl2: Commonwealth Law Bulletine
Plac: United States
Date: July 3, 1989
Volm: 11(43)
Page: 14

737.
Auth: Sherman [redacted]rie
Affl: National Law Journal staff reporter
Ttl1: Study endorses DNA evidence
Type: newspaper
Area: lay press
Ttl2: The National Law Journal
Date: August 13, 1990
Page: 3, 12

738.
Auth: Shibata, Darryl//Kurosu, Mitsuyasu//Noguchi, Thomas T.
Affl: Univ. of Southern California School of Medicine and Los Angeles County; Nippon Medical School, Toyoko;
Ttl1: Fixed human tissues: A resource for the identification of individuals
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: July 1991
Volm: 36(4)
Page: 1204-1212

739.
Auth: Shibata, Darryl//Namiki, Thomas//Higuchi, Russell
Affl: University of Southern California Medical Center, Los Angeles, CA; Cetus Corporation, Emeryville, CA
Ttl1: Identification of a mislabeled fixed specimen by DNA analysis
Type: journal article
Area: technical or scientific
Ttl2: American Journal of Surgical Pathology
Date: 1990
Volm: 14(11)
Page: 1076-1078

740.
Auth: Shriver, Mark D.//Daiger, Stephen P.//Chakraborty, Ranajit//Boerwinkle, Eric
Affl: University of Texas Graduate School of Biomedical Sciences, Houston, Tx.
Ttl1: Multimodal distribution of length variation in VNTR loci detected using PCR.
Type: journal article
Area: technical or scientific
Ttl2: Proceedings International Seminar on the Forensic App. of PCR Technology
Date: 1992
Volm: 18
Page: 144-147

741.
Auth: Signer, Ester//Kuenzle, Clive C.//Thomann, Peter E.//Hubscher, Ulrich
Affl: Department of Laboratory Animals, Univeristy of Zurich-Irchel, Switzerland
Ttl1: DNA fingerprinting: improved DNA extraction from small blood samples
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: August 11, 1988
Volm: 16(15)
Page: 7738

[redacted].
Auth: Signer, Esther//Kuenzle, Clive C.//Thomann, Peter E.//Hubscher, Ulrich
Affl: Department of Laboratory Animals, University of Zurich-Irchel, Switzerland
Ttl1: Modified gel electrophoresis for higher resolution of DNA fingerprints
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: August 11, 1988
Volm: 16(15)
Page: 7739

743.
Auth: Sinnock, Pomeroy//Sing, Charles F.
Affl: Univ. Michigan Medical School, Ann Arbor; Univ. of Maine, Orono
Ttl1: Analysis of multilocus genetic systems in Tecumseh, Michigan. I. Definition of the data set and tests for goodness-of-fit to expectations based on gene, gamete, and single locus phenotype frequencies
Type: journal article
Area: technical or scientific
Ttl2: American Journal of Human Genetics
Date: 1972
Volm: 24
Page: 381-392

744.
Auth: Slater, Nancy
Ttl1: DNA fingerprinting: Dick Tracy of the '90s
Type: journal artiicle
Area: legal
Ttl2: St. John's Journal of Legal Commentary
Plac: United States
Date: Fall 1989
Volm: 4(2)
Page: 183-203

745.
Auth: Smit, Vincent T.//Cornelisse, Cees J.//de Jong, Daphne//Dijkshoorn, Nel J.//Peters, Alex A.//Fleuren, Gert J.
Affl: Department of Pathology, University of Leiden, The Netherlands
Ttl1: Analysis of tumor heterogeneity in a patient with synchronously occurring female genital tract malignancies by DNA flow cytometry, DNA fingerprinting and immunohistochemistry
Type: journal article
Area: technical or scientific
Ttl2: Cancer
Date: September 15, 1988
Volm: 62(6)
Page: 1146-1152

746.
Auth: Smith, J. C.//Anwar, R.//Riley, J.//Jenner, D.//Markham, A. F.//Jeffreys, A. J.
Affl: ICI Diagnostics, Gadbrook Park, Rudheath, Northwich; University of Leicester, UK
Ttl1: Highly polymorphic minisatellite sequences: allele frequencies and mutation rates for five locus-specific probes in a caucasian population
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science Soc.
Date: January-February 1990
Volm: 30(1)
Page: 19-32

747.
Auth: Smith, J. C.//Newton, C. R.//Alves, A.//Anwar, R.//Jenner, D.//Markham, A. F.
Affl: ICI Diagnostics, Gadbrook Park, Rudheath, Northwich, Cheshire, UK
Ttl1: Highly polymorphic minisatellite DNA probes. Further evaluation for individual identification and paternity testing.
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science Soc
Date: January-February 1990
Volm: 30(1)
Page: 3-18

748.
Auth: Smith, Ted A.//Frasca, Deborah L.//Alevy, Martin C.//Budowle, Bruce
Affl: West Virginia Dept. of Public Safety, South Charleston, WV; Conn. Dept. of Health Services, Hartford, Conn.; FBI Laboratory, Quantico, VA.
Ttl1: PCR amplification of the VNTR polymorphism located 3' to the human Type II collagen gene
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 18(4)
Page: 203

749.
Auth: Southern, E. M.
Affl: Dept. of Zoology, University of Edinburgh, Edinburgh, Scotland
Ttl1: Detection of specific sequences among DNA fragments separated by gel electrophoresis
Type: journal article
Area: technical or scientific
Ttl2: Journal Molecular Biology
Date: 1975
Volm: 98
Page: 503-517

750.
Auth: Speth, C.//Epplen, J. T.//Oberbaumer, I.
Ttl1: DNA fingerprinting with oligonucleotides can differentiate cell lines derived from the same tumor
Type: journal article
Area: technical or scientific
Ttl2: In Vitro Cell Dev. Biol.
Date: 1991
Volm: 27A
Page: 646-650

751.
Auth: Stacey, G. N.//Bolton, B. J.//Doyle, A.
Affl: PHLS Centre for Applied Microbiology and Research, Porton Down, Salisbury, Wiltshire, UK
Ttl1: The quality control of cell banks using DNA fingerprinting
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 361-370

752.
Auth: Stacey, Glyn N.//Bolton, Bryan J.//Doyle, Alan
Affl: Public Health Laboratory Service for Applied Microbiology and Research, Division of Biologics, Porton Down, Salisbury, Wiltshire, SP4 0JG, UK
Ttl1: DNA fingerprinting transforms the art of cell authentication
Type: product review
Area: technical or scientific
Ttl2: Nature
Date: May 21, 1992
Volm: 357(6375)
Page: 261-262

753.
Auth: Stacy, John E.//Ims, Rolf A.//Stenseth, Nils C.//Jakobsen, Kjetill S.
Affl: University of Oslo, Norway
Ttl1: Fingerprint of diverse species with DNA probes generated from immobilized single-stranded DNA templates
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: 1991
Volm: 19(14)
Page: 4004

754.
Auth: Stafford, John H.
Affl: Legal Counsel Division, FBI, Washington, D.C.
Ttl1: Massachusetts rejects use of population statistics
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: January 1991
Volm: 18(1)
Page: 27-28

755.
Auth: Starrs, [redacted] E.
Ttl1: The forensic scientist and the open mind
Type: journal article
Area: technical or scientific
Ttl2: Journal Forensic Science Society
Date: 1991
Volm: 31
Page: 2

756.
Auth: Stenson, Suzanne Hickman
Ttl1: Admit it! DNA fingerprinting is reliable
Type: journal article
Area: legal
Ttl2: Houston Law Review
Plac: United States
Date: July 1989
Volm: 26(4)
Page: 677-706

757.
Auth: Stoney, D. A.
Affl: University of Illinois at Chicago, Chicago, Illinois
Ttl1: Description and evaluation of quantitative methods used to assess the strength of correspondence between many-banded patterns resulting from DNA typing
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 253

758.
Auth: Studer, R.//Kammerbauer, C.//Zischler, H.//Hinkkanen, A.
Ttl1: Highly instable (GATA)n containing sequences of the mouse during the cloning process
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 153-158

759.
Auth: Stuver, Willard C.//Kahn, Roger
Affl: Metro-Dade Police Department, Miami, FL.
Ttl1: DNA analysis in the Metro-Dade Police Department Crime Laboratory Bureau
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 179-181

760.
Auth: Stuver, Willard C.//Kahn, Roger
Affl: Metro-Dade Police Department, Miami, FL.
Ttl1: Establishing a DNA testing program in the Metro-Dade County crime laboratory
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 223-228

761.
Auth: Sullivan, K. M.//Pope, S.//Hopgood, R.//Gill, P.
Affl: Home Office Forensic Science Service, Central Research and Support Establishment, Aldermaston, Reading, UK
Ttl1: The use of fluorescence labelling technology in the forensic analysis of PCR products
Type: journal article
Area: technical or scientific
Ttl2: Crime Laboratory Digest
Date: October 1991
Volm: 163-165

762.
Auth: Sullivan, Patrick J.
Affl: Senior Attorney, Asst. Hennepin County Public Defender, Minneapolis, MN.
Ttl1: DNA fingerprint matches
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: June 26, 1992
Volm: 256
Page: 1743-1746

763.
Auth: Sussking, Alan//Eccles, Frances
Ttl1: DNA fngerprinting; implications for civil proceedings
Type: journal article
Area: legal
Ttl2: Journal of the Law Society of Scotland
Plac: Great Britain
Date: September 1988
Volm: 33(9)
Page: 324-325

764.
Auth: Swafford, Lori L.
Ttl1: Admissibility of DNA genetic profiling evidence in criminal proceedings: the case for caution
Type: journal article
Area: legal
Ttl2: Pepperdine Law Review
Date: December 1990
Volm: 18
Page: 123

765.
Auth: Swarner, ...//Reynolds, R.//Sensabaugh, G. F.
Affl: University of California, Berkeley, CA.
Ttl1: A comparative study of DNA extracted from seven postmortem tissues
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 261

766.
Auth: Takahashi, H.//Goto, S.//Kobayashi, M.//Takeuchi, S.
Affl: Department of Obstetrics and Gynecology, Niigata University School of Medicine
Ttl1: A method of differentiating between monozygotic and dizygotic twins
Type: journal article
Area: technical or scientific
Ttl2: Nippon Sanka Fujinka Gakkai Zasshi
Date: December 1989
Volm: 41(12)
Page: 2005-2009

767.
Auth: Taylor, Charles E.
Affl: Professor of Biology, UCLA, Los Angeles, CA.
Ttl1: Survey of Population Geneticists Concerning Methods of Calculating Matches in Forensic Applications of VNTR Loci
Area: technical or scientific
Date: July 9, 1991

768.
Auth: Taylor, Graham
Affl: Reg. DNA Laboratory, Leeds General Infirmary, Belmont Grove, Leeds, UK
Ttl1: DNA fingerprinting
Type: letter to editor
Area: technical or scientific
Ttl2: Nature
Date: August 31, 1989
Volm: 340(6236)
Page: 672

769.
Auth: Taylor, Judge
Affl: Alabama Court of Criminal Appeals
Ttl1: Appeal Court Decision
Type: legal document
Area: legal
Case: TJ Yelder vs State of Alabama
Cort: Alabama Courts of Criminal Appeals
Date: October 11, 1991
Page: 1,18-31
Srce: Alabama Court of Criminal Appeals, Judicial Department

770.
Auth: Templeton, Joe W.
Affl: Professor of Genetics and Veterinary Pathology, Texas A&M University
Ttl1: Canine DNA fingerprinting: can it identify breeds?
Type: journal article
Area: technical or scientific
Ttl2: Journal American Veterinary Medical Association
Date: May 1, 1990
Volm: 196(9)
Page: 1357,1359,1365

771.
Auth: Thein, S. L.//Jeffreys, A. J.//Gooi, H. C.//Cotter, F.//Flint, J.//O'Connor, N. T.//Weatherall, D. J.//Wainscoat, J. S.
Affl: MRC Molecular Haematology Unit, John Radcliffe Hospital, Oxford, UK
Ttl1: Detection of somatic changes in human cancer DNA by DNA fingerprint analysis
Type: journal article
Area: technical or scientific
Ttl2: British Journal Cancer
Date: April 1987
Volm: 55(4)
Page: 353-356

772.
Auth: Thein, S. L.//Jeffreys, A. J.//Blacklock, H. A.
Ttl1: Identification of post transplant cell population by DNA fingerprint analysis
Type: journal article
Area: technical or scientific
Ttl2: Lancet
Date: July 5, 1986
Volm: 2(8497)
Page: 37

773.
Auth: Thein, S. L.//Oscier, D. G.//Jeffreys, A. J.//Hesketh, C.//Pilkington, S.//Summers, C.//Fitchett, M.//Wainscoat, J. S.
Affl: Nuffield Department of Clinical Medicine, John Radcliffe Hospital, Oxford, UK
Ttl1: Detection of chromosomal 7 loss in myelodysplasia using an extremely polymorphic DNA probe
Type: journal article
Area: technical or scientific
Ttl2: British Journal Cancer
Date: February 1988
Volm: 57(2)
Page: 131-134

774.
Auth: Thompson, Mark
Ttl1: Authorities moving toward use of DNA fingerprinting
Type: journal article
Area: legal
Ttl2: Criminal Justice Newsletter
Plac: United States
Date: February 1, 1988
Volm: 19(3)
Page: 3-4

775.
Auth: Thompson, Mark
Ttl1: DNA wins in court; anticipating an age of genetic fingerprinting, police are planning their own high-tech labs
Type: journal article
Area: legal
Ttl2: California Lawyer
Plac: California
Date: October 1989
Volm: 9(10)
Page: 36(1)

776.
Auth: Thompson, Mark
Ttl1: DNA's troubled debut: genetic fingerprinting promises to revolutionize law enforcement in California - if the technology can survive its inaugural run
Type: journal article
Area: legal
Ttl2: California Lawyer
Plac: United States
Date: June 1988
Volm: 8(5)
Page: 36(7)

777.
Auth: Thompson, Mark
Ttl1: The myth of DNA fingerprints: the attorney general changed his mind about genetic evidence - just in time for a court challenge (CA.)
Type: journal article
Area: legal
Ttl2: California Lawyer
Plac: California
Date: April 1989
Volm: 9(4)
Page: 34(2)

778.
Auth: Thompson, William C.//Ford, Simon
Affl: University of California, Irvine
Ttl1: DNA typing - promising forensic technique needs additional validation
Type: journal article
Area: technical or scientific
Ttl2: Trial
Date: September 1988
Page: 56-6

779.
Auth: Thompson, William C.//Ford, Simon
Ttl1: DNA typing: acceptance and weight of the new genetic identification tests
Type: journal article
Area: legal
Ttl2: Va. Law Review
Date: February 1989
Volm: 75
Page: 45

780.
Auth: Thompson, William//Ford, Simon
Affl: Program in Social Ecology; Public Policy Research Organization, Univ. of California, Irving, California
Ttl1: The meaning of a match: sources of ambiguity in the interpretation of DNA prints
Type: book chapter
Area: technical or scientific
BkAu: Farley, Mark A.//Harrington, James J.
Ttl2: Forensic DNA Technology
Date: 1991
Page: 93-152

781.
Auth: Tilzer, L.//Moreno, R.//Booth, F.
Affl: Kansas University Medical Center, Kansas City Police Department, Kansas City
Ttl1: DNA fingerprinting with M13mp8 RF bacteriophage using nonradioactive methods
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 229

782.
Auth: Tilzer, Lowell L.//Moreno, Ruben F.//Booth, Frank//Wilbur, Stephen//Thomas, Stanley M.
Affl: Department of Pathology and Oncology, Kansas University Medical Center, Kansas City, 66103
Ttl1: DNA fingerprinting with nonradioactive gene probes
Type: journal article
Area: technical or scientific
Ttl2: Clinical Chemistry
Date: October 1989
Volm: 35(10)
Page: 2147

783.
Auth: Tonelli, Lois A.//Markowicz, Karen R.//Anderson, Mariane B.//Green, David J.//Herrin, George L.//Cotton, Robin W.//et al
Ttl1: Use of deoxyribonucleic acid (DNA) fingerprints for identity determination: comparison with traditional paternity testing methods (part 1)
Type: journal article
Area: legal
Ttl2: Journal Forensic Sciences
Plac: United States
Date: November 1990
Volm: 35(6)
Page: 1265-1269

784.
Auth: Track, [redacted]na K.//Ricciuti, Florence C.//Kidd, Kenneth K.
Affl: Howard Hughes Medical Institute Human Gene Mapping Library; Yale University School of Medicine, New Haven, Connecticut
Ttl1: Information of DNA polymorphisms in the human gene mapping library
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 335-345

785.
Auth: Trainor, George L.//Prober, James M.//Dam, Rudy J.
Affl: E.I. du Pont de Nemours & Co., Inc., Wilmington, Delaware
Ttl1: Fluorescence detection in nucleic acid analysis
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 297-304

786.
Auth: Troyer, D.//Howard, D.//Leipold, H. W.//Smith, J. E.
Ttl1: A human minisatellite sequence reveals DNA polymorphism in the equine species
Type: journal article
Area: technical or scientific
Ttl2: Zentralbl Veterinarmed 'A:
Date: February 1989
Volm: 36(2)
Page: 81-83

787.
Auth: Turner, Bruce J.//Elder Jr., John F.//Laughlin, Thomas F.//Davis, William P.
Affl: Department of Biology, Virginia Polytechnic Institute and State University, Blacksburg, VA. 24061
Ttl1: Genetic variation in clonal vertebrates detected by simple sequence DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy Science USA
Date: August 1990
Volm: 87(15)
Page: 5653-5657

788.
Auth: Ubell, Earl
Ttl1: Whodunit? Quick, Check the genes!
Type: newspaper article
Area: technical or scientific
Ttl2: Parade Magazine
Date: March 31, 1991
Page: 10-13

[redacted].
Auth: Uitterlinden, Andre G.//Slagboom, Eline P.//Mullaart, Erick//Meulenbelt, Ingrid//Vijg, Jan
Affl: Medscand Ingeny; TNO Institute for Experimental Gerontology, Rijswijk, The Neatherlands
Ttl1: Genome scanning by two dimensional DNA typing: the use of repetitive DNA sequences for rapid mapping of genetic traits
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 119-134

790.
Auth: Uitterlinden, Andre G.//Slagboom, P. Elaine//Knook, Dick L.//Vijg, Jan
Affl: Department of Molecular Biology, TNO Institute for Experimental Gerontology, Rijswijk, The Netherlands
Ttl1: Two dimensional DNA fingerprinting of human individuals
Type: journal article
Area: technical or scientific
Ttl2: Proceedings Natinal Academy Sciences USA
Date: April 1989
Volm: 86(8)
Page: 2742-2746

791.
Auth: Upcroft, P.//Mitchell, R.//Boreham, P. F.
Affl: Queensland Institute of Medical Research, Herston, Brisbane, Australia
Ttl1: DNA fingerprinting of the intestinal parasite Giardia Duodenalis with the M13 phage genome
Type: journal article
Area: technical or scientific
Ttl2: International Journal Parasitology
Date: May 1990
Volm: 20(3)
Page: 319-323

792.
Auth: van Brunt, Jennifer
Ttl1: Are DNA fingerprints admissible in court?
Type: journal article
Area: technical or scientific
Ttl2: Biotechnology
Date: November 1988
Page: 1271

793.
Auth: Van Eede[illegible]. H.//Henke, L.//Fimmers, R.//Henke, J.//de Lange, G. D.
Affl: Central Laboratory of the Netherlands Red Cross Blood Transfusion Service
Ttl1: Size calculation of restriction enzyme HaeIII-generated fragments detected by probe YNH24 by comparison of data from two laboratories: the generation of fragment--size frequencies.
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: January-February 1991
Volm: 49(1)
Page: 21-31

794.
Auth: van Helden, Paul D.//Wiid, Ian J.//Albrecht, Carl F.//Theron, Elize//Thornley, Alan L.//Hoal-van Helden, Eileen G.
Affl: Department of Medical Biochemistry, University of Stellenbosch Medical School, Tygerberg, South Africa
Ttl1: Cross-contamination of human eosophageal squamous carcinoma cell lines detected by DNA fingerprint analysis
Type: journal article
Area: technical or scientific
Ttl2: Cancer Research
Date: October 15, 1988
Volm: 48(20)
Page: 5660-5662

795.
Auth: Vandenberghe, Antoon//Van Den Broeck, Marleen//Peeters, Inge
Affl: Dept. of Biochemistry, University of Antwerp, Antwerp
Ttl1: Paternal Matation of YNH24 (D2S44), a probe frequently used in paternity testing
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: June 1993
Volm: 52
Page: 206-207

796.
Auth: Vassart, Gilbert//Georges, Michel//Monsieur, Rita//Brocas, Huguette//Lequarre, Anne S.//Christophe, Daniel
Affl: Institut de Recherche Interdisciplinaire, University Libre de Bruxelles, Campus Erasme
Ttl1: A sequence in M13 phage detects hypervariable minisatellites in human and animal DNA
Type: journal article
Area: technical or scientific
Ttl2: Science
Date: February 6, 1987
Volm: 235(4789)
Page: 683-684

797.
Auth: Veggeberg, S.
Ttl1: Report validating DNA fingerprint method could hasten growth in biotech - a recent document from the NRC is seen as a boost for entrepeneurs whose ventures center on the controversial technique
Type: journal article
Area: technical or scientific
Ttl2: Scientist
Date: 1992
Volm: 6
Page: 1

798.
Auth: Verbovaya, Lilya V.//Ivanov, Pavel L.
Affl: U.S.S.R. Academy of Sciences; Russia Soviet Federative Socialist Republic Ministry of Health, Moscow, U.S.S.R.
Ttl1: "Sexing" deoxyribonucleic acid (DNA) on DNA fingerprint gel: an internal control for DNA fingerprint evidence
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Plac: Soviet Union
Date: July 1991
Volm: 36(4)
Page: 991-998

799.
Auth: Vergnaud, G.//Mariat, D.//Zoroastro, M.//Lauthier, V.
Ttl1: Detection of single and multiple polymorphic loci by synthetic tandem repeats of short oligonucleotides
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 134-140

800.
Type: book
Area: general reference
BkAu: Vogel, Friedrich//Motulsky, Arno G.
Ttl2: Human Genetics. Second Edition.
Plac: Berlin, Germany
Publ: Springer-Verlag
Date: 1986
Volm: ISBN 3-540-16411-1
Page: i-xxxiv,1-807

801.
Auth: von Beroldingen, C. H.//Roby, R.//Sensabaugh, G. F.//Walsh, S.
Affl: University of California, Berkeley, CA.; Dept. of Human Genetics, Cetus Corp., Emeryville, CA.
Ttl1: DNA in hair
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 265-266

802.
Auth: Wainscoat, J. S.//Pilkington, S.//Peto, T. E. A.//Bell, J. I.//Higgs, D. R.
Affl: John Radcliff Hospital, Headington, Oxford, UK; Stanford University School of Medicine, Stanford, CA.
Ttl1: Allele-specific DNA identity patterns
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: 1987
Volm: 75
Page: 384-387

803.
Auth: Walsh, P. Sean//Fildes, Nicola//Louie, Alan S.//Higuchi, Russell
Affl: Cetus Corporation, Emeryville, CA.
Ttl1: Report of the blind trial of the Cetus AmpliType HLA-DQa forensic deoxyribonucleic acid (DNA) amplification and typing kit
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: September 1991
Volm: 36(5)
Page: 1551-1556

804.
Auth: Walsh, P. Sean//Metzger, David A.//Higuchi, Russell
Affl: Cetus Corporation and Illinois State Police
Ttl1: Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material
Type: journal article
Area: technical or scientific
Ttl2: BioTechniques
Date: April 1991
Volm: 10(4)
Page: 506-513

805.
Auth: Walsh, S.//Higuchi, R.//Blake, E.
Affl: Cetus Corporation; Forensic Science Assoc.; California
Ttl1: PCR inhibition and bloodstains
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 281-282

806.
Auth: Washio, Keiko//Misawa, Shogo//Ueda, Shintaroh
Affl: Department of Anthropology, Faculty of Science, University of Tokyo, Hongo, Japan
Ttl1: Probe walking: development of novel probes for DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: October 1989
Volm: 83(3)
Page: 223-226

807.
Auth: Washio, Keiko//Ueda, Shintaroh//Misawa, Shogo
Affl: University of Tsukuba, Ibaraki; University of Tokyo; Japan
Ttl1: Effects of cytosin methylation at restriction sites on deoxyribonucleic acid (DNA) typing
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: November 1990
Volm: 35(6)
Page: 1277-1283

808.
Auth: Waye, J. S.//Shutler, G. G.//Monteith, N.//Bishop, L.//Fourney, R. M.
Affl: Royal Canadian Mounted Police, Ottawa, Canada
Ttl1: DNA typing using a panel of VNTR probes
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 241

809.
Auth: Waye, J[redacted]S.
Affl: Royal Canadian Mounted Police, Ottawa, Ontario, Canada
Ttl1: Cloning and recombinant DNA technologies for the development of hybridization probes
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 47-56

810.
Auth: Waye, John S.
Affl: Central Forensic Laboratory, Royal Canadian Mounted Police, Ottawa, Ontario, Canada
Ttl1: Discussion of "Repetitive deoxyribonucleic acid and human genome variation - a concise review relevant to forensic biology"
Type: letter to editor
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: November 1989
Volm: 34(6)
Page: 1296-1299

811.
Auth: Waye, John S.//Fourney, R. M.
Affl: Royal Canadian Mounted Police, Ottawa, Ontario, Canada
Ttl1: Agarose gel electrophoresis of linear genomic DNA in the presence of ethidium bromide: band shifting and implications for forensic identity testing
Type: journal article
Area: technical or scientific
Ttl2: Applied and Theoretical Electrophoresis
Date: 1990
Volm: 1
Page: 193

812.
Auth: Waye, John S.//Fourney, Ron M.
Affl: Central Forensic Laboratory, Royal Canadian Mounted Police, Ottawa, Ontario
Ttl1: Identification of complex DNA polymorphisms based on variable number of tandem repeats (VNTR) and restriction site polymorphism
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: Feburary 1990
Volm: 84(3)
Page: 223-227

813.
Auth: Waye, John S.//Michaud, Denis//Bowen, John H.//Fourney, Ron M.
Affl: McMaster University, Hamilton; Central Forensic Lab., Royal Canadian Mounted Police, Ottawa, Ontario, Canada
Ttl1: Sensitive and specific quantification of human genomic deoxyribonucleic acid (DNA) in forensic science speciments: casework examples
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: July 1991
Volm: 36
Page: 1198-1203

814.
Auth: Waye, John S.//Presley, Lawrence A.//Budowle, Bruce//Shutler, Gary G./-/Fourney, Ron M.
Affl: Central Forensic Laboratory, Royal Canadian Mounted Police
Ttl1: A simple and sensitive method for quantifying human genomic DNA in forensic specimen extracts
Type: journal article
Area: technical or scientific
Ttl2: BioTechniques
Date: September 1989
Volm: 7(8)
Page: 852-855

815.
Auth: Weir, B. S.
Affl: Program in Statistical Genetics, Dept. of Statistics North Carolina State University, Raleigh
Ttl1: Book review "DNA on trail: Genetic identification and criminal justice.
Type: book review
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: November 1993
Volm: 53(5)
Page: 1158-1160

816.
Auth: Weir, B. S.
Affl: Program in Statistic, Dept. of Statistics, North Carolina State University, Raleigh
Ttl1: Independence of VNTR alleles defined as fixed bins
Type: journal article
Area: technical or scientific
Ttl2: Genetics
Date: April 1992
Volm: 130
Page: 873-887

817.
Auth: Weir, B
Affl: Dept. of Statistics, North Carolina State University, Raleigh, N.C.
Ttl1: Independence of VNTR alleles defined as floating bins
Type: journal article
Area: technical or scientific
Ttl2: American Jouran Human Genetics
Date: November 1992
Volm: 51(5)
Page: 992-997

818.
Auth: Weir, B. S.
Affl: Program in Statistical Genetics, Dept. of Statistics, North Carolina State University, Raleigh
Ttl1: Independence tests for VNTR alleles defined as quantile bins
Type: journal article
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: November 1993
Volm: 53(5)
Page: 1107-1113

819.
Auth: Weir, B. S.//Evett, I. W.
Affl: North Carolina University, Raleigh; Home Office Forensic Science Service, Aldermaston, Reading, England
Ttl1: Whose DNA?
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: April 1992
Volm: 50(4)
Page: 869

820.
Auth: Weir, Bruce S.
Affl: Dept. of Statistics, North Carolina State Univ. Raleigh, NC
Ttl1: Forensic Population Genetics and the National Research Council (NRC)
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: June 1993
Volm: 52
Page: 437-440

821.
Auth: Weir, Bruce S.
Affl: Dept. of Statistics, North Carolina State University, Raleigh, N. C.
Type: book
Area: general reference
Ttl2: Genetic Data Analysis
Plac: Sunderland, Mass.
Publ: Sinauer Associates
Date: 1991
Page: 377 + xii

822.
Auth: Weir, Bruce S.
Affl: Dept. of Statistics, North Carolina State University, Raleigh, N.C.
Ttl1: Inferences about linkage disequilibrium
Type: journal article
Area: technical or scientific
Ttl2: Biometrics
Date: March 1979
Volm: 35
Page: 235-254

823.
Auth: Weir, Bruce S.
Affl: Dept. of Statistics, North Carolina State Univ, Raleigh, NC
Ttl1: Population genetics in the forensic DNA debate
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy Science USA
Date: December 1992
Volm: 89
Page: 11654-11659

824.
Auth: Weir, Bruce S.
Affl: Dept. of Statistics, North Carolina State University, Raleigh, NC
Ttl1: Population genetics in the forensic DNA debate
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy Science USA
Date: December 1992
Volm: 89
Page: 11654-11659

825.
Auth: Weir, Bruce S.
Ttl1: State of Oregon vs. Herbert Jackson Futch
Type: unpublished document
Area: technical or scientific
Case: State of Oregon vs. Herbert Jackson Futch
Date: November 7, 1990
Page: 1-7

826.
Auth: Weir, Bruce S.//Evett, I. W.
Affl: Dept. of Statistics, North Carolina State Univ.; Home Office Forensic Science Service, Aldermaston, Reading, England
Ttl1: Reply to Lewontin
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: June 1993
Volm: 52
Page: 206

827.
Auth: Weising, [redacted]//Beyermann, B.//Ramser, J.//Kahl, G.
Ttl1: Plant DNA fingerprinting with radioactive and digoxigenated oligonucleotide probes complementary to simple repetitive DNA sequences
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 159-168

828.
Auth: Weising, K.//Ramser, J.//Kaemmer, D.//Kahl, G.//Epplen, J. T.
Affl: Johann Wolfgang Goethe-Universitat, Frankfurt; Max-Planck-Institut fur Psychiatrie, Martinsried, Germany
Ttl1: Oligonucleotide fingerprinting in plants and fungi
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 312-329

829.
Auth: Weitzel, Jeffrey N.//Hows, Jill M.//Jeffreys, Alec J.//Min, Gao L.//Goldman, John M.
Affl: Department of Haematology, Hammersmith Hospital, London
Ttl1: Use of a hypervariable minisatellite DNA probe (33.15) for evaluating engraftment two or more years after bone marrow transplantation for aplastic anaemia
Type: journal article
Area: technical or scientific
Ttl2: British Journal Haematology
Date: September 1988
Volm: 70(1)
Page: 91-97

830.
Auth: Wells, Richard A.//Green, Philip//Reeders, Stephen T.
Affl: MRC Molecular Haematology Unit, University of Oxford, John Radcliffe Hospital, Headington, UK
Ttl1: Simultaneous genetic mapping of multiple human minisatellite sequences using DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Genomics
Date: November 1989
Volm: 5(4)
Page: 761-772

831.
Auth: Wells, Richard A.//Wonke, B.//Thein, Swee Lay
Affl: MRC Molecular Haematology Unit, University of Oxford, John Radcliffe Hospital, Headington, UK
Ttl1: Prediction of consanguinity using human DNA fingerprints
Type: journal article
Area: technical or scientific
Ttl2: Journal of Medical Genetics
Date: October 1988
Volm: 25(10)
Page: 660-662

832.
Auth: Werrett, D. J.//Gill, P. D.//Evett, I. W.//Lygo, J. E.//Sullivan, K. M./-/Buckleton, J.
Affl: Home Offfice Forensic Science Service, Aldermaston, UK
Ttl1: DNA analysis in home office laboratories: its introduction, immediate future and statistical assessment
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 147-162

833.
Auth: Werrett, David J.//Lygo, Joan E.//Sutton, John G.
Affl: Home Office Forensic Science Service, Aldermaston, Reading, Berkshire, UK
Ttl1: The introduction of DNA analysis into home office forensic science laboratories in England and Wales
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 233-241

834.
Auth: Westin, Alan F.
Affl: Columbia University, N.Y., N.Y.
Ttl1: A privacy analysis of the use of DNA techniques as evidence in courtroom proceedings
Type: book chapter
Area: technical or scientific
BkAu: Ballantyne, Jack//Sensabaugh, George//Witkowski, Jan
Ttl2: Banbury Report 32: DNA Technology and Forensic Science
Date: 1989
Page: 25-42

835.
Auth: Westnea[redacted]avid F.//Noon, William A.//Reeve, Hudson K.//Aquadro, Charles F.
Affl: Section of Genetics and Development, Cornell University, Ithaca, NY 14853
Ttl1: Improved hybridization conditions for DNA fingerprints probed with M13
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: May 11, 1988
Volm: 16(9)
Page: 4161

836.
Auth: Westwood, Sara A.//Werrett, David J.
Affl: Home Office Forensic Science Service, Aldermaston, Reading, Berkshire, UK
Ttl1: An evaluation of the polymerase chain reaction method for forensic applications
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: 1990
Volm: 45
Page: 201-215

837.
Auth: Wetton, Jon H.//Carter, Royston E.//Parkin, David T.//Walters, David
Affl: Department of Genetics, Medical School, Queen's Medical Centre, Nottingham, UK
Ttl1: Demographic study of a wild house sparrow population by DNA fingerprinting
Type: journal article
Area: technical or scientific
Ttl2: Nature
Date: May 14-20, 1987
Volm: 327(6118)
Page: 147-149

838.
Auth: White, John J.//Neuwirth, Harry//Miller, C. Dennis//Schneider, Edward L.
Affl: Gerontology Research Center, National Institute on Aging, Baltimore, MD; University of California, UCLA School of Medicine, Los Angeles, CA
Ttl1: DNA alterations in prostatic adenocarcinoma and benign prostatic hyperplasia: detection by DNA fingerprint analyses
Type: journal article
Area: technical or scientific
Ttl2: Mutation Research
Date: January 1990
Volm: 237(1)
Page: 37-43

839.
Auth: White, R.//Nakamura, Y.//Kasai, K.//Odelberg, S.//Wolff, R.
Affl: University of Utah School of Medicine, Salt Lake City, UT
Ttl1: DNA markers in forensic applications
Type: book chapter
Area: technical or scientific
BkAu: Hicks, John W.
Ttl2: Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis
Date: June 19-23, 1989
Page: 19-24

840.
Auth: White, Robin M.
Ttl1: Developments in DNA fingerprinting and the law
Type: journal article
Area: legal
Ttl2: Scolag
Plac: Scotland
Date: February 1990
Volm: 161
Page: 24-27

841.
Auth: White, Robin M.//Greenwood, Jeremy
Ttl1: DNA fingerprinting and the law
Type: journal article
Area: legal
Ttl2: Modern Law Review
Plac: Great Britain
Date: March 1988
Volm: 51(2)
Page: 145-155

842.
Auth: White, Thomas J.//Arnheim, Norman//Erlich, Henry A.
Affl: Diagnostic Research, Hoffmann-LaRoche, Inc.; Univ. of Southern California; Cetus Corporation, Emerville; CA
Ttl1: The polymerase chain reaction
Type: journal article
Area: technical or scientific
Ttl2: Trends in Genetics
Date: June 1989
Volm: 5(6)
Page: 185-189

843.
Auth: Williams, Charles L.
Ttl1: DNA fingerprinting: a revolutionary technique in forensic science and its probable effects on criminal evidentiary law
Type: journal article
Area: legal
Ttl2: Drake Law Review
Plac: United States
Date: Fall 1987
Volm: 37(1)
Page: 1-32

844.
Auth: Wills, Christopher
Affl: Dept. of Biology, University of California, San Diego, La Jolla, CA. 92093
Ttl1: Forensic DNA typing
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: February 28, 1992
Volm: 255
Page: 1050

845.
Auth: Winkler [redacted], John K.
Affl: Lifecodes Corporation, Saw Mill River Road, Valhalla, N.Y.
Ttl1: DNA Fingerprinting
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: March 2, 1990
Volm: 247(4946)
Page: 1018-1019

846. Auth: Witkowski, J. A.
Affl: The Banbury Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
Ttl1: Milestones in the development of DNA technology
Type: book chapter
Area: technical or scientific
BkAu: Farley, Mark A.//Harrington, James J.
Ttl2: Forensic DNA Technology
Date: 1991
Page: 1-23

847.
Auth: Witkowski, Jan
Affl: Banbury Center, Cold Spring Harbor Laboratory
Ttl1: Fingerprint of the future?
Type: journal article
Area: technical or scientific
Ttl2: The NEB Transcript
Date: December 1989
Volm: 2(1)
Page: 2-8

848.
Auth: Witt, Michal//Erickson, Robert P.
Affl: University of Michigan Medical School, Ann Arbor, MI.
Ttl1: Determination of the sex of origin of blood and bloodstains using recombinant DNA techniques
Type: book chapter
Area: technical or scientific
BkAu: Lee, Henry C.//Gaensslen, R. E.
Ttl2: DNA and Other Polymorphisms in Forensic Science
Date: 1990
Page: 98-113

849.
Auth: Wolfes, Rudiger//Mathe, Judith//Seitz, Alfred
Affl: Institut fur Zoologie I, Johannes Gutenberg Universitat Mainz, Germany
Ttl1: Forensics of birds of prey by DNA fingerpriting with 32P-labeled oligonucleotide probes
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 175-180

850.
Auth: Wolff, R.//Nakamura, Y.//Odelberg, S.//Shiang, R.//White, R.
Affl: Univ. of Utah Medical Center, Salt Lake Utah; Cancer Institute, Tokyo; University of Iowa City;
Ttl1: Generation of variability of VNTR loci in human DNA
Type: book chapter
Area: technical or scientific
BkAu: Burke, T.//Dolf, G.//Jeffreys, A. J.//Wolff, R.
Ttl2: DNA Fingerprinting: Approaches and Applications
Date: 1991
Page: 20-38

851.
Auth: Wolff, Roger K.//Nakamura, Yusuke//White, Ray
Affl: Department of Cellular, Viral and Molecular Biology, University of Utah, Salt Lake City, Utah 84132
Ttl1: Molecular characterization of a spontaneously generated new allele at a VNTR locus: No exchange of flanking DNA sequence
Type: journal article
Area: technical or scientific
Ttl2: Genomics
Date: November 1988
Volm: 3(4)
Page: 347-351

852.
Auth: Wolff, Roger K.//Plaetke, Rosemarie//Jeffreys, Alec J.//White, Ray
Affl: Department of Cellular, Viral and Molecular Biology, University of Utah, Salt Lake City, Utah 84132
Ttl1: Unequal crossing over between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci
Type: journal article
Area: technical or scientific
Ttl2: Genomics
Date: August 1989
Volm: 5(2)
Page: 382-384

853.
Auth: Wong, Zilla//Wilson, Victoria//Patel, I.//Povey, S.//Jeffreys, Alec J.
Affl: Department of Genetics, University of Leicester, UK
Ttl1: Characterization of a panel of highly variable minisatellites cloned from human DNA
Type: journal article
Area: technical or scientific
Ttl2: Annals of Human Genetics
Date: October 1987
Volm: 51 (Pt 4)
Page: 269-288

854.
Auth: Wong, Z[redacted]//Wilson, Victoria//Jeffreys, Alec J.//Thein, Swee Lay
Affl: Department of Genetics, University of Leicester, UK
Ttl1: Cloning a selected fragment from a human DNA fingerprint: isolation of an extremely polymorphic minisatellite
Type: journal article
Area: technical or scientific
Ttl2: Nucleic Acids Research
Date: June 11, 1986
Volm: 14(11)
Page: 4605-4616

855.
Auth: Wood, N. A. P.//Clay, T. M.//Didwell, J. L.
Ttl1: HLA-DR/Dw matching by PCR fingerprinting: The origin of PCR fingerprints and further applications
Type: journal article
Area: technical or scientific
Ttl2: Eur. J. Immunogenetic
Date: 1991
Volm: 18
Page: 147-153

856.
Auth: Wooley, James R.
Affl: Organized Crime Strike Force Division, Northern District of Ohio, Cleveland
Ttl1: A response to Lander: the courtroom perspective
Type: letter to editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: 1991
Volm: 49
Page: 892-893

857.
Auth: Wooley, James//Harmon, Rockne P.
Affl: Asst. US Attorney, Dept. of Justice Strike Force, Cleveland; Senior Deputy District Attorney, Alameda County, CA.
Ttl1: The forensic DNA brouhaha: science or debate?
Type: Letter to the editor
Area: technical or scientific
Ttl2: American Journal Human Genetics
Date: November 1992
Volm: 51(5)
Page: 1164-1165

858.
Auth: Wyman, Arlene R.//White Ray
Affl: Dept. of Microbiology, Univ. of Mass. Medical School, Worcester, Mass.
Ttl1: A highly polymorphic locus in human DNA
Type: journal article
Area: technical or scientific
Ttl2: Proceedings National Academy of Science, USA
Date: November 1980
Volm: 77(11)
Page: 6754-6758

[redacted].
Auth: Xiang, K. S.
Ttl1: A study of DNA fingerprinting in China
Type: journal article
Area: technical or scientific
Ttl2: Chung Hua I Hsueh Tsa Chih
Date: October 1989
Volm: 69(10)
Page: 569-572, 40

860.
Auth: Yarbrough, Lynwood R.
Affl: Dept. of Biochemistry and Molecular Biology, University of Kansas Medical Center, Kansas City, KS. 66160-7421
Type: letter to editor
Area: technical or scientific
Ttl2: Science
Date: February 28, 1992
Volm: 255
Page: 1052

861.
Auth: Yassouridis, A.//Epplen, J. T.
Ttl1: On paternity determination from multilocus DNA profiles
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 221-225

862.
Auth: Yaxley, Ron
Ttl1: DNA fingerprinting
Type: journal article
Area: legal
Ttl2: Commonwealth Law Bulletin
Plac: Great Britain
Date: April 1989
Volm: 15(2)
Page: 614-619

863.
Auth: Yokoi, T.//Odaira, T.//Nata, M.//Aoki, Y.//Sagisaka, K.
Ttl1: Application of single-locus hypervariable region DNA probes to deficiency cases in paternity testing
Type: journal article
Area: technical or scientific
Ttl2: International Journal of Legal Medicine
Date: 1991
Volm: 104
Page: 117

864.
Auth: Yokoi, [redacted]Odaira, T.//Nata, M.//Sagisaka, K.
Ttl1: Investigation of paternity establishing without the putative father using hypervariable DNA probes
Type: journal article
Area: technical or scientific
Ttl2: Japan Journal Human Genet.
Date: 1990
Volm: 35
Page: 235

865.
Auth: Yokoi, Tsuyoshi//Nata, Masayuki//Odaira, Toru//Sagisaka, Kaoru
Affl: Department of Forensic Medicine, Tohoku University School of Medicine, Aoba-ku, Sendai, Japan
Ttl1: Hypervariable polymorphic VNTR loci for parentage testing and individual identification
Type: journal article
Area: technical or scientific
Ttl2: Japan Journal Human Genetics
Date: June 1990
Volm: 35(2)
Page: 179-188

866.
Auth: Yokoi, Tsuyoshi//Sagisaka, Kaoru
Affl: Dept. of Forensic Medicine, Tohoku Univ. School of Medicine, Sendai, Japan
Ttl1: Sex determination of blood stains recombinant DNA probe: comparison with radioactive and non-radioactive labeling
Type: journal article
Area: technical or scientific
Ttl2: Forensic Science International
Date: 1989
Volm: 41
Page: 117-124

867.
Auth: Zhang, Xiao Wei//Lan, Lin//Huo, Zheng Yi//Duan, Bing Zhang Z.//Koblinsky, Lawrence
Affl: Beijing Forensic Science Institute, Beijing, People's Republic of China; City University of New York, N.Y.
Ttl1: Restriction fragment length polymorphism analysis of forensic science casework in the People's Republic of China
Type: journal article
Area: technical or scientific
Ttl2: Journal of Forensic Sciences
Date: March 1991
Volm: 36(2)
Page: 531-536

[redacted].
[redacted]th: Zischler, H.//Hinkkanen, A.//Studer, R.
Ttl1: Oligonucleotide fingerprinting with (CAC)5: nonradiocactive in gel hybridization and isolation of individual hypervariable loci
Type: journal chapter
Area: technical or scientific
BkAu: Epplen, J. T.
Ttl2: Electrophoresis
Date: February 3, 1991
Volm: 12(2-3)
Page: 141-145

869.
Auth: Zischler, Hans//Nanda, Indrajit//Schafer, Renate//Schmid, Michael//Epplen, Jorg T.
Affl: Max-Planck-Institut fur Psychiatrie, Martinsried, Federal Republic of Germany
Ttl1: Digoxigenated oligonucleotide probes specific for simple repeats in DNA fingerprinting and hybridization in situ
Type: journal article
Area: technical or scientific
Ttl2: Human Genetics
Date: June 1989
Volm: 82(3)
Page: 227-233

870.
Auth: Zurer, P.
Ttl1: DNA fingerprinting - use upheld, but strict standards urged
Type: journal article
Area: technical or scientific
Ttl2: Chemical and Engineering News
Date: 1992
Volm: 70
Page: 4-5

871.
Auth: Zurer, P.
Ttl1: FBI director defends use of DNA profiling
Type: journal article
Area: technical or scientific
Ttl2: Chemical and Engineering News
Date: 1992
Volm: 70
Page: 7-8

Publications on DNA Fingerprinting Based on Research Supported by NIJ Grant 92-IJ-CX-K024, "Analysis of DNA Typing Data for Forensic Applications"

Stephen P. Daiger, PhD; Ranajit Chakraborty, PhD; Eric Boerwinkle, PhD
Graduate School of Biomedical Sciences
The University of Texas Health Science Center at Houston

Period covered: June 1990 - December 1992

REFERENCES

1. **E. BOERWINKLE**, C Leffert, H Hobbs (1991). Analysis of the apolipoprotein(a) gene structure in two populations with different distributions of plasma lipoprotein(a). Proc. 8th Int. Cong. of Human Genetics, Washington, DC. A#67.
2. **E BOERWINKLE** (1992). Genetics of plasma lipoprotein(a) concentrations. Current Opinion in Lipidology 3:128-136.
3. **E BOERWINKLE**, S-H Chen, S Visvikis, CL Hanis, G Siest, L Chan (1991). Signal peptide-length variation in human apolipoprotein B gene: Molecular characteristics and association with plasma glucose levels. Diabetes 40:1539-1544.
4. **E BOERWINKLE**, CC Leffert, J Lin, C Lackner, G Chiesa, HH Hobbs (1992). Apolipoprotein(a) gene accounts for greater than 90% of the variation in plasma lipoprotein(a) concentrations. J. Clin. Invest. 90:52-60.
5. B Budowle, AM Giusti, **R CHAKRABORTY** (1990). Discretized allelic data for VNTR locus by amplified fragment length polymorphism (AMP-FLP) analysis. Amer. J. Hum. Genet. 47:A#0502.
6. B Budowle, **R CHAKRABORTY**, AM Giusti, AJ Eisenberg, RC Allen (1991). Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution PAGE. Am. J. Hum. Genet. 48:137-144.
7. RM Cerda-Flores, GK Kshatriya, SA Barton, CH Leal-Garza, R Garza-Chapa, WJ Schull, **R CHAKRABORTY** (1991). Genetic structure of the populations migrating from San Luis Potosi and Zacatecas to Nuevo León in Mexico. Hum. Biol. 63:309-327.
8. RM Cerda-Flores, GK Kshatriya, TK Bertin, D Hewett-Emmett, CL Hanis, **R CHAKRABORTY** (1992). Gene diversity and estimation of genetic admixture among Mexican-Americans of Starr County, Texas. Ann. of Hum. Biol. 19:347-360.
9. **R CHAKRABORTY** (1990). Genetic profile of cosmopolitan populations: Effects of hidden subdivision. Anthrop. Anz. 48:313-331.
10. **R CHAKRABORTY** (1990). Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. Am.J. Hum. Genet. 47:87-94.
11. **R CHAKRABORTY** (1991). Book review: DNA Technology and Forensic Science, Banbury Report 32, (J Ballantyne and J Witkowski, eds.) Cold Spring Harbor Press, New York. Amer. J. Hum. Genet. 48:173-174.
12. **R CHAKRABORTY** (1991). Book review: Genetic Data Analysis. BS Weir, Sinauer Assoc. Mol. Biol. Evol. 8:396-397.
13. **R CHAKRABORTY** (1992). "Commentaries" on DNA typing and its court use. Professional Ethics Report V2:3-4.
14. **R CHAKRABORTY** (1992). Generalized occupancy problem and its applications in population genetics. "Impact of Genetics Variation on Individuals, Families and Populations" (CF Sing, CL Hanis, eds) Oxford University Press, New York pp. 179-192.

15. **R CHAKRABORTY** (1991). Impact of molecular genetics in studying origin of human populations. Archivos de Biologia y Medicina Experimentales 24:R98
16. **R CHAKRABORTY** (1991). Letters to the Editor: Inclusion of data on relatives for estimation of allele frequencies. Amer. J. Hum. Genet. 49:242-243.
17. **R CHAKRABORTY** (1991). Statistical interpertation of DNA typing data. Amer. J. Hum. Genet. 49:895-897.
18. **R CHAKRABORTY** (1991). Population genetics of hypervariable loci. Proc. 8th Int. Cong. of Human Genetics, Washington, DC 49:A252.
19. **R CHAKRABORTY** (1992). Book review: Convergent Issues in Genetics and Demography. (JA Adams, A Hermalin, D Lam, PE Smouse, eds.) Oxford Univ. Press, New York. Amer. J. Hum. Biol. 4:421-428.
20. **R CHAKRABORTY** (1992). Letters to the Editor: Multiple alleles and estimation of genetic parameters: Computational equations showing involvement of all alleles. Genetics 130:231-234.
21. **R CHAKRABORTY** (1992). Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Hum. Biol. 64:141-159.
22. **R CHAKRABORTY, E BOERWINKLE** (1990). Population genetics of VNTR polymorphism in humans. Amer. J. Hum. Genet. 47:A0504.
23. **R CHAKRABORTY, SP DAIGER** (1991). Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah. Hum. Biol. 63:571-587.
24. **R CHAKRABORTY, SP DAIGER, E BOERWINKLE** (1991). Patterns of genetic variation within and between populations detected by PCR-based VNTR polymorphisms. Proc. In. Seminar of the Forensic App. of PCR Technology, FBI Academy, Quantico, VA. Crime Lab Digest 18:148-152.
25. **R CHAKRABORTY**, H Danker-Hopfe (1991). Analysis of population structure: A comparative study of different estimators of Wright's fixation indices. Handbook of Statistics, 8:203-254.
26. **R CHAKRABORTY**, M de Andrade, **SP DAIGER**, B Budowle (1992). Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Ann. Hum. Genet 56:45-57.
27. **R CHAKRABORTY**, R Deka, L Jin, RE Ferrell (1992). Allele sharing at six VNTR loci and genetic distances among three ethnically defined human populations. Amer. J. Hum. Giol. 4:387-397.
28. **R CHAKRABORTY**, M Fornage, R Gueguen, **E BOERWINKLE** (1991). Population genetics of hypervariable loci: Analysis of PCR based VNTR polymorphism within a population. "DNA Fingerprinting: Approaches and Applications", (T Burke, G Dolf, AJ Jeffreys, R Wolff, eds.), Birkhauser-Verlag, Bern., pp. 127-143.
29. **R CHAKRABORTY**, L Jin (1992). Formal statistics of DNA fingerprinting data and relatedness between individuals. Amer. J. Hum. Genet. 51:A46.
30. **R CHAKRABORTY**, L Jin (1992). Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. Hum. Genet. 88:267-272.
31. **R CHAKRABORTY**, L Jin with 43 co-authors (1992). Multiple origins for phenylketonuria in Europe. Amer. J. Hum. Genet. 51:1355-1365.
32. **R CHAKRABORTY**, I Kamboh, RE Ferrell (1991). 'Unique' alleles in admixed populations: A strategy for determining 'hereditary' population differences of disease frequencies. Ethnicity & Disease 1:245-256.
33. **R CHAKRABORTY**, MI Kamboh, M Nwankwo, RE Ferrell (1992). Caucasian genes in American blacks: New data. Amer. J. Hum. Genet. 50:145-155.
34. **R CHAKRABORTY**, MI Kamboh, RE Ferrell (1992). Letter to the Editor: Response to Issues in estimating Caucasian admixture in American blacks. Reply to Reed. Amer. J. Hum. Genet. 51:680-681.
35. **R CHAKRABORTY**, KK Kidd (1991). The utility of DNA typing in forensic work. Science 254:1735-1739.
36. **R CHAKRABORTY**, KK Kidd (1992). Letter to the Editor: Forensic DNA typing: response. Science 255:1053.

37. **R CHAKRABORTY**, CR Rao (1991). Measurement of genetic variation for evolutionary studies. Handbook of Statistics 8:271-316.
38. **R CHAKRABORTY**, MR Srinivasan (1992). A modified "best maximum likelihood" estimator of line regression with errors in both variables: an application for estimating genetic admixture. Biometrical J. 5:567-576.
39. **R CHAKRABORTY**, MR Srinivasan, L Jin, M de Andrade (1992). Effects of population subdivision and allele frequency differences on interpretation of DNA typing data for human identification. Proc. 2nd Intl. Symp. of Hum. Id., Promega Corp., Madison WI pp 205-222.
40. **R CHAKRABORTY**, KM Weiss (1991). Genetic variation of the mitochondrial DNA genome in American Indians is at mutation-drift equilibrium. Amer. J. Phy. Anth. 86:497-506.
41. P Clemens, RG Fenwick, JS Chamberlain, RA Gibbs, M de Andrade, **R CHAKRABORTY**, CT Caskey (1991). A rapid and informative assay for linkage analysis and prenatal diagnosis in Duchenne muscular dystrophy families using CA polymorphisms in a deletion-prone region of dystrophin. Amer. J. Hum. Genet 49:A978.
42. PR Clemens, RG Fenwick, JS Chamerlain, RA Gibbs, M de Andrade, **R CHAKRABORTY**, CT Caskey (1991). Carrier detection and prenatal diagnosis in Duchenne and Becker Muscular Dystrophy families, using dinucleotide repeat polymorphisms. Amer. J. Hum. Genet. 49:951-960.
43. **SP DAIGER** (1991). Issues in DNA fingerprinting for forensic purposes. State Bar of Texas Professional Development Program, J1-J41.
44. **SP DAIGER** (1991). Letter to the Editor. DNA Fingerprinting. Amer. J. Hum. Genet. 49:897.
45. M de Andrade, **R CHAKRABORTY**, RP Clemens, CT Caskey (1991). Linkage disequilibria among CA polymorphisms in the human dysrophin gene. Amer. J. Hum. Genet. 49:A983.
46. R Deka, **R CHAKRABORTY**, S DeCroo, F Rothhammer, SA Barton, RE Ferrell (1992). Characteristics of polymorphism at a VNTR locus 3' to the apolipoprotein b gene in five human populations. Amer. J. Hum. Genet. 51:1325-1333.
47. R Deka, **R CHAKRABORTY**, RE Ferrell (1990). Population genetics of human hypervariable loci. Amer. J. Hum. Genet. 47:A0512.
48. R Deka, **R CHAKRABORTY**, RE Ferrell (1991). A population genetic study of six VNTR loci in three ethnically defined populations. Genomics 11:83-92.
49. R Deka, **R CHAKRABORTY**, RE Ferrell (1991). Allele sharing and genetic distance at VNTR loci among three ethnic groups. Proc. 8th Int. Cong. of Human Genetics, Washington DC 49:A2837.
50. A Edwards, HA Hammond, **R CHAKRABORTY**, CT Caskey (1991). DNA typing with trimeric and tetrameric tandem repeats: polymorphic loci, detection systems, and population genetics. Proceedings of the 2nd Int'l. Symposium on Hum. Id. Promega Corp. Madison, WI pp. 31-52.
51. A Edwards, HA Hammond L Jin, CT Caskey, **R CHAKRABORTY** (1992). Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12:241-253.
52. J Ely, **R CHAKRABORTY**, R Deka, RE Ferrell (1991). Comparison of VNTR polymorphisms among human and chimpanzee. Amer. J. Hum. Genet. 49:A2615.
53. J Ely, R Deka, **R CHAKRABORTY**, RE Ferrell (1992). Comparison of five tandem repeat loci between humans and chimpanzees. Genomics 14:692-698.
54. M Fornage, G Siest, **E BOERWINKLE** (1991). Frequency distribution of a (TG)*n*(AG)*m* microsatellite reflects the mechanisms of production of new alleles. Proc. 8th Int'l. Cong. of Hum. Genetics, Washington, DC 49:A2800.
55. M Fornage, L Chan, G Siest, **E BOERWINKLE** (1992). Allele frequency distribution of the (TG)*n*(AG)*m* microsatellite in the apolipoprotein C-II gene. Genomics 12:63-68.
56. HA Hammond, A Edwards, L Jin, **R CHAKRABORTY**, CT Caskey (1991). Studies of multilocus genotype data validate the use of DNA typing with polymorphic trimeric and tetrameric tandem repeats for personal identification. Amer. J. Hum. Genet. 49:A2506.
57. L Jin, **R CHAKRABORTY** (1992). Population dynamics of DNA fingerprinting patterns within and between populations. Amer. J. Hum. Genet. 51:A603.
58. L Jin, **R CHAKRABORTY**, HA Hammond, CT Caskey (1991). Polymorphisms at short tandem

repeat (STR) loci within and between four ethnic populations of Texas. Proc. 8th Int'l. Cong. of Hum. Genet. Washington, DC 49:A65.

59. MI Kamboh, **R CHAKRABORTY**, RE Ferrell (1990). Caucasian genes in the American blacks: New data. Amer. J. Hum. Genet. 47:A0540.
60. C Lackner, **E BOERWINKLE**, CC Leffert, T Rahmig, HH Hobbs (1991). Molecular basis of apolipoprotein(a) isoform size heterogeneity as revealed by pulsed-field gel electrophoresis. J. Clin. Invest. 87:2153-2161.
61. CR Rao, **R CHAKRABORTY** (1991). Handbook of Statistics, Vol. 8: Statistical Methods in Biological and Medical Sciences. Elsevier Science Publishing Company, Inc., New York.
62. M Shriver, **SP DAIGER, R CHAKRABORTY, E BOERWINKLE** (1991). Multimodal distribution of length variation in VNTR loci detected using PCR. Proc. Int'l. Seminar on the Forensic App. of PCR Technology. Crime Laboratory Digest 18:144-147.
63. MD Shriver, JE Hixson, **SP DAIGER, E BOERWINKLE** (1991). Birth of a VNTR: Size and sequence comparison of the apo B 3' VNTR in humans and non-humans primates. Proc. 8th Int'l. Cong. of Hum. Genet. Washington, DC 49:A2631.
64. MD Shriver, L Jin, **R CHAKRABORTY, E BOERWINKLE** (1992). Computer simulations of the stepwise mutation model and VNTR allele frequency distributions. Amer. J. Hum. Genet. 51:A623.
65. MR Srinivasan, **SP DAIGER, R CHAKRABORTY** (1992). Interval estimation of multilocus genotype frequencies and its forensic implications. Amer. J. Hum. Genet. 51:A628.

Austin Feb 92 because not new

Session 16: Population Genetics

63

Detecting loci for oligogenic traits by linkage analysis. B.K. Suarez*, P. Van Eerdewegh, C.L. Hampe. Washington University School of Medicine, St. Louis, MO, USA.

Recent successes in mapping a wide variety of simple diseases of unknown etiology have engendered the hope that similar techniques will prove equally effective for complex phenotypes. The basic assumption is that the complex phenotype is not determined by too many loci. In this analysis we investigate two questions: what is the power to detect a contributing oligogenic locus using information on a linked marker; and, once such a locus has been detected, what is the probability that the same locus will be detected in an independent replication study?

In order to answer these questions, we have carried out a Monte Carlo simulation in families with a "CEPH-like" structure wherein between 2 to 10 independent loci contribute to a quasi-continuous threshold trait. Models with equal and unequal genic effects have been investigated using three different statistical methods (the "mean statistic" (t_2) of Blackwelder and Elston, the affected pedigree method of Weeks and Lange, and an admittedly misspecified LOD score approach suggested by Risch). Results are reported for the case where a study consists of 100 ascertained multiplex pedigrees. With this sample size, three conclusions emerge: 1) The ability to detect a contributing locus improves as the threshold becomes more deviant, 2) When more than about six loci determine liability, fewer than 50% may be detected and, 3) under such circumstances, attempts to replicate a previously reported (true) linkage claim are likely to fail.

64

A new measure of similarity of DNA fingerprints. C.C. Li*, D.E. Weeks, R.E. Ferrell, and A. Chakravarti. University of Pittsburgh, Pittsburgh, PA.

Given the DNA fingerprints of two individuals (x and y), let n_x and n_y represent the number of bands of x and y, respectively, and n_{xy} be the number of bands shared by both individuals. The degree of similarity between these two individuals is traditionally measured by $S = 2n_{xy}/(n_x + n_y)$. However, if this usual S-measure is used to calculate the similarity between random pairs of individuals in a randomly mating population, the result does not agree with the following calculation: Given x has a band-i (with probability p_i), the conditional probability that y will also have band-i is $p_i^2 + 2p_i(1 - p_i) = p_i(2 - p_i)$, so the frequency of sharing band-i is $p_i^2(2 - p_i)$. Summing over all alleles yields the total frequency $2\Sigma p_i^2 - \Sigma p_i^3$. This disagreement may be eliminated by proposing a new measure of similarity:

$$S = \frac{1}{2}\left[\frac{n_{xy}}{n_x} + \frac{n_{xy}}{n_y}\right] = \frac{n_{xy}}{2}\left[\frac{1}{n_x} + \frac{1}{n_y}\right],$$

which is the arithmetic average of the two fractions n_{xy}/n_x and n_{xy}/n_y. This is equivalent to dividing n_{xy} by the harmonic mean of n_x and n_y, instead of by the arithmetic mean as in the traditional S-measure. The similarity between random pairs calculated using our new S agrees with that obtained by other methods.

65

Polymorphisms at short tandem repeat (STR) loci within and between four ethnic populations of Texas, USA. L. Jin*(1), R.Chakraborty(1), H.A.Hammond (2) and C.T.Caskey(2). (1) Genetics Centers, Grad. Sch. of Biomed. Sciences, Univ. of Texas, (2) Inst. Mol. Genet., Baylor College of Medicine, Houston, Texas, USA.

Using PCR-based protocols genetic polymorphisms at four trimeric and tetrameric STR loci (Tho-1, Rena-4, Hprt-b and Fab-p) are detected in each of the four ethnic groups, Orientals, Blacks, Hispanics and Caucasians of Texas, USA. This data characterizes genetic variation within and between these populations at STR loci. Allele numbers range between 4 and 8; and heterozygosities (H) are between 36 and 80% per locus. Blacks show the highest heterozygosity and largest number of alleles, although these differences are not substantial. The Rena-4 locus is the least variable (H = 36-76%) and the Tho-1 locus is the most variable (H = 76-77%). Permutation tests show that the genotype frequencies are at Hardy-Weinberg equilibrium and allelic associations among loci are not significant within each population. The Blacks, Hispanics and Caucasians share larger proportions (85%) of alleles, while the Orientals share approximately 74% of their alleles with the other three groups. The alleles which are not found in common among the four groups always have small (< 2%) frequencies. The Hispanics are closest to the Caucasians (Nei's standard D = 0.03) and even the largest distance (0.20 between Blacks and Orientals) is not appreciable for such hypervariable loci. These population genetic characteristics substantiate the previous claim that such STR loci are extremely useful for forensic and paternity determination purposes. (Research supported by NIJ grants 90-IJ-CX- 0037 and 90-IJ-CX-0038).

66

Genetic relationship between DNA polymorphisms in the apolipoprotein B (apo B) gene in individuals of European and South Asian origin - population-genetic aspects of the apo B gene VNTR site. E. Renges *, R. Peacock, A. Dunning, P. Talmud and S. Humphries. Arterial Diseases Research Group, CX Sunley Research Centre, London, UK.

We have investigated the genetic relationship between apo B gene polymorphisms: insertion/deletion (Leu_{14}-Ala-Leu_{14}), XbaI (Thr_{2488}), EcoRI (Glu → Lys_{3158}), Asn/Ser_{3311} and VNTR in samples of South Asian (Indian) and Swedish individuals. The frequency distribution at all these sites was found to be significantly different between the Indian and the Swedish sample (deletion allele: 0.20 v 0.31, $p<0.025$, X+ (presence of XbaI site): 0.29 v 0.55, $p<0.001$, R- (absence of EcoRI site): 0.11 v 0.19, $p<0.05$, Ser_{3311}: 0.46 v 0.18, $p<0.001$). Although the distribution of alleles at the VNTR site was bimodal in both populations, in Indians, the most common allele was a 35-repeat unit allele, whilst in the Swedish sample, and in all other reports from Caucasian samples the 37-repeat unit allele was the most frequent. Furthermore, four new alleles at the apo B gene VNTR site were discovered in South Asians, two (17 and 15 repeat unit alleles) are well outside the bimodal distribution indicating a possible trimodal distribution. Strong linkage disequilibrium and allelic association was detected between alleles at the VNTR and the EcoRI site in both samples. Together with the bimodal distribution, this supports a mechanism for generation of new alleles at the apo B VNTR locus that does not involve unequal crossing-over as a major factor but favours a mechanism of replication slippage and deletion/insertion of repeat units.

67

Analysis of the apolipoprotein(a) gene structure in two populations with different distributions of plasma lipoprotein(a). Eric Boerwinkle (1), Carla Leffert (2), and Helen Hobbs (2). (1) The University of Texas Health Science Center in Houston, Texas and (2) The University of Texas Southwestern Medical School, Dallas, Texas.

In Caucasians, elevated levels of lipoprotein(a) [Lp(a)] are associated with coronary atherosclerosis. African-Americans tend to have higher plasma concentrations of Lp(a) and yet do not have a higher incidence of heart disease. We compared plasma levels of Lp(a) in 105 African-Americans (mean=37.5 mg/dl) to 102 Caucasians (mean=19.3 mg/dl). Not only were levels in African-Americans significantly higher ($p=.011$) but their distribution was also more Gaussian. Prior studies in Caucasians have shown that the plasma Lp(a) levels are inversely related to the number of kringle 4-encoding repeats in the apolipoprotein(a) [apo(a)] gene as determined by pulse-field gel electrophoresis and a kringle 4 specific probe. We are investigating the contribution of molecular variation in the apolipoprotein(a) [apo(a)] gene to differences in lp(a) levels between Caucasians and African-Americans. In Caucasians, 19 alleles (48 to 190 kilobases) representing size variation in the apo(a) gene were found. Eighteen of the 19 alleles were also present in the African-Americans; absent was one rare allele. There were no significant differences in apo(a) allele frequencies between groups ($p=.31$). We conclude that the higher plasma concentrations of Lp(a) in the African-American population are not due to differences in the numbers of kringle 4 repeats in the apo(a) gene. To determine if the observed differences in plasma Lp(a) concentrations are due to sequence differences at the apo(a) locus or other loci, family studies are being performed.

Lost in Feb 92 because not new

Session 16: Population Genetics

63

Detecting loci for oligogenic traits by linkage analysis. B.K. Suarez*, P. Van Eerdewegh, C.L. Hampe. Washington University School of Medicine, St. Louis, MO, USA.

Recent successes in mapping a wide variety of simple diseases of unknown etiology have engendered the hope that similar techniques will prove equally effective for complex phenotypes. The basic assumption is that the complex phenotype is not determined by too many loci. In this analysis we investigate two questions: what is the power to detect a contributing oligogenic locus using information on a linked marker; and, once such a locus has been detected, what is the probability that the same locus will be detected in an independent replication study?

In order to answer these questions, we have carried out a Monte Carlo simulation in families with a "CEPH-like" structure wherein between 2 to 10 independent loci contribute to a quasi-continuous threshold trait. Models with equal and unequal genic effects have been investigated using three different statistical methods (the "mean statistic" (t_2) of Blackwelder and Elston, the affected pedigree method of Weeks and Lange, and an admittedly misspecified LOD score approach suggested by Risch). Results are reported for the case where a study consists of 100 ascertained multiplex pedigrees. With this sample size, three conclusions emerge: 1) The ability to detect a contributing locus improves as the threshold becomes more deviant, 2) When more than about six loci determine liability, fewer than 50% may be detected and, 3) under such circumstances, attempts to replicate a previously reported (true) linkage claim are likely to fail.

64

A new measure of similarity of DNA fingerprints. C.C. Li*, D.E. Weeks, R.E. Ferrell, and A. Chakravarti. University of Pittsburgh, Pittsburgh, PA.

Given the DNA fingerprints of two individuals (x and y), let n_x and n_y represent the number of bands of x and y, respectively, and n_{xy} be the number of bands shared by both individuals. The degree of similarity between these two individuals is traditionally measured by $S = 2n_{xy}/(n_x + n_y)$. However, if this usual S-measure is used to calculate the similarity between random pairs of individuals in a randomly mating population, the result does not agree with the following calculation: Given x has a band-i (with probability p_i), the conditional probability that y will also have band-i is $p_i^2 + 2p_i(1 - p_i) = p_i(2 - p_i)$, so the frequency of sharing band-i is $p_i^2(2 - p_i)$. Summing over all alleles yields the total frequency $2\Sigma p_i^2 - \Sigma p_i^3$. This disagreement may be eliminated by proposing a new measure of similarity:

$$S = \frac{1}{2}\left[\frac{n_{xy}}{n_x} + \frac{n_{xy}}{n_y}\right] = \frac{n_{xy}}{2}\left[\frac{1}{n_x} + \frac{1}{n_y}\right]$$

which is the arithmetic average of the two fractions n_{xy}/n_x and n_{xy}/n_y. This is equivalent to dividing n_{xy} by the harmonic mean of n_x and n_y, instead of by the arithmetic mean as in the traditional S-measure. The similarity between random pairs calculated using our new S agrees with that obtained by other methods.

65

Polymorphisms at short tandem repeat (STR) loci within and between four ethnic populations of Texas, USA. L.Jin*(1), R.Chakraborty(1), H.A.Hammond (2) and C.T.Caskey(2). (1) Genetics Centers, Grad. Sch. of Biomed. Sciences, Univ. of Texas, (2) Inst. Mol. Genet., Baylor College of Medicine, Houston, Texas, USA.

Using PCR-based protocols genetic polymorphisms at four trimeric and tetrameric STR loci (Tho-1, Rena-4, Hprt-b and Fab-p) are detected in each of the four ethnic groups, Orientals, Blacks, Hispanics and Caucasians of Texas, USA. This data characterizes genetic variation within and between these populations at STR loci. Allele numbers range between 4 and 8; and heterozygosities (H) are between 36 and 80% per locus. Blacks show the highest heterozygosity and largest number of alleles, although these differences are not substantial. The Rena-4 locus is the least variable (H = 36-76%) and the Tho-1 locus is the most variable (H = 76-77%). Permutation tests show that the genotype frequencies are at Hardy-Weinberg equilibrium and allelic associations among loci are not significant within each population. The Blacks, Hispanics and Caucasians share larger proportions (85%) of alleles, while the Orientals share approximately 74% of their alleles with the other three groups. The alleles which are not found in common among the four groups always have small (< 2%) frequencies. The Hispanics are closest to the Caucasians (Nei's standard D = 0.03) and even the largest distance (0.20 between Blacks and Orientals) is not appreciable for such hypervariable loci. These population genetic characteristics substantiate the previous claim that such STR loci are extremely useful for forensic and paternity determination purposes. (Research supported by NIJ grants 90-IJ-CX- 0037 and 90-IJ-CX-0038).

66

Genetic relationship between DNA polymorphisms in the apolipoprotein B (apo B) gene in individuals of European and South Asian origin - population-genetic aspects of the apo B gene VNTR site. E. Renges *, R. Peacock, A. Dunning, P. Talmud and S. Humphries. Arterial Diseases Research Group, CX Sunley Research Centre, London, UK.

We have investigated the genetic relationship between apo B gene polymorphisms: insertion/deletion (Leu_{4154}-Ala-Leu_{4155}), XbaI (Thr_{3444}), EcoRI (Glu → Lys_{4154}), Asn/Ser_{4311} and VNTR in samples of South Asian (Indian) and Swedish individuals. The frequency distribution at all these sites was found to be significantly different between the Indian and the Swedish sample (deletion allele: 0.20 v 0.31, $p<0.025$; X+ (presence of XbaI site): 0.29 v 0.55, $p<0.001$, R- (absence of EcoRI site): 0.11 v 0.19, $p<0.05$, Ser_{4311}: 0.46 v 0.18, $p<0.001$). Although the distribution of alleles at the VNTR site was bimodal in both populations, in Indians, the most common allele was a 35-repeat unit allele, whilst in the Swedish sample, and in all other reports from Caucasian samples the 37-repeat unit allele was the most frequent. Furthermore, four new alleles at the apo B gene VNTR site were discovered in South Asians, two (17 and 15 repeat unit alleles) are well outside the bimodal distribution indicating a possible trimodal distribution. Strong linkage disequilibrium and allelic association was detected between alleles at the VNTR and the EcoRI site in both samples. Together with the bimodal distribution, this supports a mechanism for generation of new alleles at the apo B VNTR locus that does not involve unequal crossing-over as a major factor but favours a mechanism of replication slippage and deletion/insertion of repeat units.

67

Analysis of the apolipoprotein(a) gene structure in two populations with different distributions of plasma lipoprotein(a). Eric Boerwinkle*(1), Carla Leffert (2), and Helen Hobbs (2). (1) The University of Texas Health Science Center in Houston, Texas and (2) The University of Texas Southwestern Medical School, Dallas, Texas.

In Caucasians, elevated levels of lipoprotein(a) [Lp(a)] are associated with coronary atherosclerosis. African-Americans tend to have higher plasma concentrations of Lp(a) and yet do not have a higher incidence of heart disease. We compared plasma levels of Lp(a) in 105 African-Americans (mean=37.5 mg/dl) to 102 Caucasians (mean=19.3 mg/dl). Not only were levels in African-Americans significantly higher ($p=.011$) but their distribution was also more Gaussian. Prior studies in Caucasians have shown that the plasma Lp(a) levels are inversely related to the number of kringle 4-encoding repeats in the apolipoprotein(a) [apo(a)] gene as determined by pulse-field gel electrophoresis and a kringle 4 specific probe. We are investigating the contribution of molecular variation in the apolipoprotein(a) [apo(a)] gene to differences in lp(a) levels between Caucasians and African-Americans. In Caucasians, 19 alleles (48 to 190 kilobases) representing size variation in the apo(a) gene were found. Eighteen of the 19 alleles were also present in the African-Americans; absent was one rare allele. There were no significant differences in apo(a) allele frequencies between groups ($p=.31$). We conclude that the higher plasma concentrations of Lp(a) in the African-American population are not due to differences in the numbers of kringle 4 repeats in the apo(a) gene. To determine if the observed differences in plasma Lp(a) concentrations are due to sequence differences at the apo(a) locus or other loci, family studies are being performed.

Reprinted from

Current Opinion in LIPIDOLOGY

Volume 3, 1992

CS
CURRENT
SCIENCE

Genetics of plasma lipoprotein (a) concentrations

Eric Boerwinkle

Center for Demographic and Populations Genetics, The University of Texas Health Science Center in Houston, Houston, USA

Lipoprotein (a) [Lp(a)] is a low-density-like lipoprotein with the addition of a Lp(a)-specific protein, apolipoprotein (a) [apo(a)]. Plasma concentrations of Lp(a) are a risk factor for premature coronary heart disease (CHD). Genetic studies have been key in developing an understanding of Lp(a) and its association with CHD. Quantitative genetic studies have demonstrated that plasma Lp(a) concentrations are largely genetically determined. Using length variation in the apo(a) gene and gene product, it has been shown that the apo(a) gene has a large effect on Lp(a) levels, and in fact variation in this one gene is largely responsible for the high heritability of plasma Lp(a) concentrations. The data presented in this review, represent the most complete picture to date of the genetic architecture of a major CHD risk factor.

Current Opinion in Lipidology 1992, 3:128–136

Introduction

In 1988, coronary heart disease (CHD) accounted for 35.3% of all deaths in the USA [1]. Unlike the inborn errors of metabolism (for example cystic fibrosis) CHD is the result of lifelong interactions among numerous environmental and genetic factors. Therefore, no single gene is responsible for the high frequency of CHD. Rather, there are many genes, each making a relatively small contribution to the overall disease liability in the population. Genetic studies of CHD are also complicated by the fact that there may be different causes in different individuals, families and populations; i.e. there are multiple paths to the same disease endpoint. The multifactorial nature of CHD has slowed progress in understanding its etiology; however, progress has been best realized when researchers have focused on one or only a few facets of this complex system. Since its discovery by Berg [2] in 1963, lipoprotein(a) [Lp(a)] has been at the forefront of research on the genetics of CHD and its risk factors.

Lp(a) is a cholesteryl ester-rich plasma lipoprotein composed of two components: an LDL-like particle to which is attached a single large glycoprotein, apolipoprotein (a) [apo(a)]. Lp(a) is an important risk factor for atherosclerotic CHD. High plasma levels of Lp(a) are positively associated with the development of premature atherosclerosis and other vascular diseases. The mechanism by which Lp(a) contributes to the atherosclerotic process is unknown, although it is probably mediated through its close homology to the plasma zymogen plasminogen [3]. Lp(a) is distinct from other CHD risk factors associated with intermediary metabolism in that plasma concentrations of Lp(a) vary over a very wide range among individuals but are extremely stable within a given individual [4]. Furthermore, many physiological, pharmacological and environmental factors that effect the levels of other plasma lipoproteins have no effect on Lp(a) concentration.

Our goals in studying the genetics of common chronic diseases such as CHD include a better understanding of their etiology and improved prediction. Such an understanding will better position society and medicine for preventative, rather than pharmacological, surgical or mere palliative measures. Throughout its relatively brief history, human genetics has led the way in studies of Lp(a) and its association with CHD. Knowledge accumulating about the genetics of plasma Lp(a) concentrations and the role of this lipoprotein in atherosclerosis will at least partially unveil the mechanisms of CHD initiation and progression, and will facilitate early identification of individuals at increased risk.

Structure of lipoprotein (a)

A schematic diagram of a Lp(a) particle is shown in Fig. 1. In many respects, the Lp(a) lipoprotein particle resembles a LDL particle with the addition of a single

Abbreviations

apo—apolipoprotein; **CHD**—coronary heart disease; **FH**—familial hypercholesterolemia; **Lp(a)**—lipoprotein (a); **mRNA**—messenger RNA.

Current Science Ltd ISSN 0957-9672

molecule of apo(a). The physicochemical properties of Lp(a) have been determined using a variety of techniques [5], and Table 1 compares these properties between Lp(a) and LDL. As in LDL, apolipoprotein (apo)B_{100} is present in Lp(a) [6] and, unlike LDL, is linked to apo(a) by a disulfide bond [7]. Lp(a) has a higher density and overall molecular weight than LDL, characteristics that make ultracentrifugation and gel filtration chromatography useful tools for its purification. The electrophoretic mobility of Lp(a) in agarose is to the pre-β position, which is similar to VLDL, and its buoyant density is characteristic of HDL_2. The protein and lipid composition of Lp(a) differs only slightly with that of LDL, with a protein: lipid ratio in Lp(a) and LDL of 1 : 2.2 and 1 : 3.5, respectively. Total lipid content represents approximately 69% in Lp(a) and 79% in LDL, but the amount of free cholesterol relative to total lipids is approximately the same in both particles (11.5 and 11.8%, respectively).

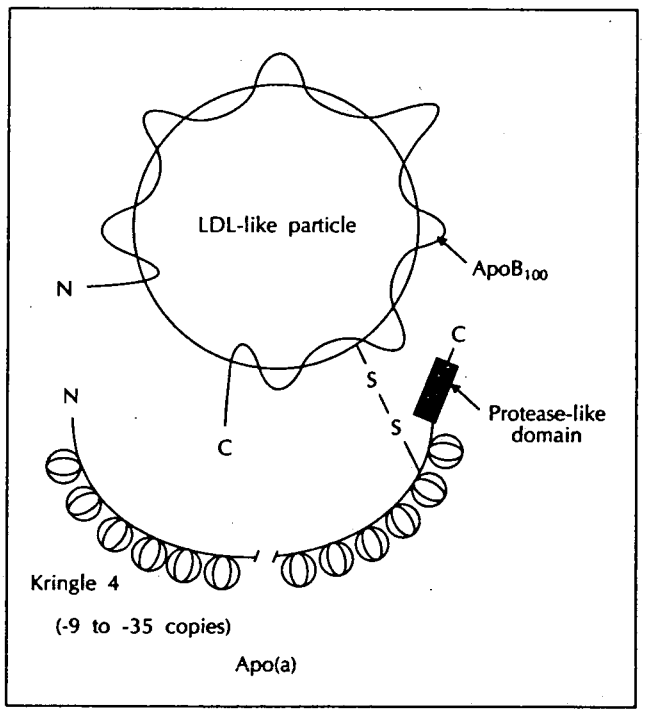


Fig. 1. Schematic diagram of an Lp(a) lipoprotein particle.

Data concerning the site of synthesis of Lp(a) in humans are scarce. The liver is thought to be involved in the production of Lp(a) because apo(a) messenger RNA (mRNA) has been detected by northern blot analysis in liver [8,9•] and because liver damage is associated with reduced Lp(a) levels [10]. However, it is unclear whether apo(a) synthesized in hepatocytes is secreted for later production of Lp(a) in the interstitia or in circulation, or whether the hepatocytes secrete Lp(a) directly. Rainwater and Lanford [11] reported that cultured primary baboon hepatocytes synthesized an Lp(a) particle identical to plasma Lp(a). In addition, the isoform type of the secreted Lp(a) was identical to the observed isoforms from the hepatocyte donor animal. To investigate the tissue expression of apo(a) mRNA in rhesus monkey, Tomlinson *et al.* [12] analyzed 12 different tissues and detected apo(a) mRNA only in liver, brain, and testes. However, no apoB mRNA was present in the latter two tissues suggesting that apo(a) may function in some tissues independently of the Lp(a) particle.

Table 1. Physical and chemical properties of Lp(a) and LDL particles.

	Lp(a)	LDL
Electrophoretic mobility	pre-β	β
Buoyant density (g/ml)	1.05–1.08	1.03–1.06
Molecular weight (× 10^6)	3.1–3.8	2.5–3.1
Apolipoproteins	$apoB_{100}$, apo(a)	$apoB_{100}$
% Composition as protein	26.0–35.7	20.7
% Composition as lipid	64.3–74.0	79.3
% Composition as cholesterol	33.6–45.4	52.3
(free cholesterol)	(7.6–10.2)	(9.4)

Data from Fless *et al.* [5] and Gaubatz *et al.* [7]

Partial protein sequencing of purified plasma apo(a) revealed remarkable sequence similarity with that of plasminogen [13], an observation which led to the cloning of the apo(a) complementary DNA by McLean *et al.* [8]. The mature apo(a) mRNA encodes a 4529-amino-acid protein which includes a 19 bp signal peptide, multiple copies of a kringle-4-like domain, one copy of a region homologous to the plasminogen kringle 5, and a protease domain. Complementary DNA sequences of the protease domain in apo(a) and plasminogen share 94% homology. This striking structural similarity with plasminogen led to the suggestion that Lp(a) might have prothrombic action. However, because of two important amino acid substitutions in the activation site (Arg–Val mutated to Ser–Ile), apo(a), unlike plasminogen, cannot be transformed into an active fibrinolytic agent. Miles *et al.* [3] demonstrated that Lp(a) was able to bind to plasminogen receptors distributed on peripheral blood cells and vascular endothelial cells, and postulated that Lp(a) competes with plasminogen for these receptors. Thus, by interfering with plasminogen binding, Lp(a) would prevent plasmin generation. Consequently, thrombolysis would be inhibited and thrombosis promoted. However, at this time we cannot rule out the possibility that the homology between apo(a) and plasminogen may be functionally coincidental and the mechanism for the association between Lp(a) and CHD is by some as-yet-unidentified path.

Lipoprotein (a) as a risk factor for atherosclerosis

The association between elevated plasma Lp(a) concentrations and the occurrence of CHD has been reported by numerous investigators employing a broad range of study designs, disease definitions, and analytical methods. A positive but moderate relationship between plasma Lp(a) concentrations and the occurrence of CHD has been shown clearly. The results of several cross-sectional epidemiological studies establishing Lp(a) as a risk factor for CHD were reviewed by Morrisett *et al.* [14]. Such a review will not be repeated here. Rather, two studies which help cultivate an appreciation for the complex relationship between Lp(a) and CHD are discussed below.

To our knowledge only one prospective study has examined the relationship between plasma Lp(a) concentrations and CHD. Rosengren *et al.* [15•] followed 776 50-year-old Swedish men for 6 years (for a total of 4656 follow-up years) and noted all CHD events and deaths due to CHD (26 in this study). Each CHD case was matched with four controls who did not suffer an identified CHD event. In this sample, the CHD-positive group had significantly higher Lp(a) levels at baseline (mean = 27.8 mg/dl) compared with the control group (mean = 17.3 mg/dl). In addition, those in the upper one-fifth of the population Lp(a) distribution had twice as much CHD than those in the lower four-fifths. These data underscore the role of Lp(a) as a risk factor for CHD in a well-designed prospective case–control study. Because Lp(a) levels were measured before the onset of acute disease, these results counter those that argue that the association between CHD and Lp(a) is a consequence of the potential role of Lp(a) as an acute phase protein [16]. However, it should be kept in mind that the scope of the study was quite narrow. For example, it only investigated the role of Lp(a) as a risk factor for CHD in men of one age. In fact, a review of the literature indicates a dearth of data documenting a role for Lp(a) in CHD in women or other groups. In addition, the control group in this study was not necessarily disease-free. There is a desperate need for a well-designed population-based prospective case–control study using angiographically documented atherosclerotic vascular disease cases and disease-free controls.

Lp(a), in its role as a risk factor for CHD, does not act alone. Armstrong *et al.* [17] reported that the association between elevated plasma Lp(a) levels and CHD was dependent on plasma LDL cholesterol levels. Individuals with both high Lp(a) and high LDL cholesterol had a sixfold increased risk of disease, whereas those with only one or the other risk factor had a 1.5- to twofold increased risk. These data suggest that there is interaction, either direct or indirect, between these two lipoprotein particles at the metabolic or cellular levels which increases CHD risk.

Plasma Lp(a) concentrations are also associated with other disease states such as cerebrovascular disease [18], renal disease [19], and diabetes mellitus. It is now well accepted that atherosclerotic vascular disease is the most common complication of diabetes [20]. Although Lp(a) represents a potent marker for and physiological element in the formation of atherosclerosis in non-diabetics, it has not been shown to be a risk factor for CHD in diabetics. Recent data suggest a relationship between diabetes and serum Lp(a) levels. Two independent studies [21,22] have reported high serum levels of Lp(a) in poorly controlled insulin-dependent diabetics compared with non-diabetic individuals. Interestingly, Lp(a) levels were dramatically decreased with improved metabolic control. A study by Joven and Vilella [23] showed that Lp(a) levels were not increased in well-controlled non-insulin-dependent diabetics and the authors suggest a predominant role of exogenous insulin in the regulation of Lp(a) levels in diabetic patients. Levitsky *et al.* [24] examined the relationship between Lp(a) levels, diabetes, and glycemic control among groups of white and black non-diabetic and insulin-dependent diabetic children. They found that circulating Lp(a) levels were increased in hyperglycemic whites, but no such relationship was observed in blacks. Although our understanding of the relationship between glycemic control and plasma Lp(a) concentrations is meager, Lp(a) should be considered as a factor contributing to CHD in this already-at-risk group.

The genetics of lipoprotein (a)

Biometrical genetic analyses

Traditional biometrical genetic analyses follow a logical series of progressively more focused questions about the role of genes influencing a quantitative trait. First, is there significant familial aggregation for the trait of interest and is this aggregation due to shared genes rather than shared environmental effects? The latter point is particularly important when investigating the risk factors for CHD as they are usually influenced by a wide variety of environmental factors. Second, is there evidence that one gene has a large influence on the quantitative phenotype? Such 'major gene' effects are usually detected using a method known as complex segregation analysis [25]. Finally, if there is a significant major gene effect, can it be localized in the human genome using linkage analysis? Each of these questions has been addressed for plasma Lp(a) concentrations.

The genetic nature of the Lp(a) phenotype was established shortly after its discovery [26]. Since that time, numerous studies have firmly established that Lp(a) levels are strongly influenced by genetic factors [27–30]. Based on the similarity of plasma Lp(a) concentration among related individuals, estimates of the polygenic heritability of Lp(a) typically range between 70% and 95%. In other words, between 70% and 95% of the interindividual variation in plasma Lp(a) levels is attributable to genetic differences among individuals. No other major CHD risk factor is influenced by genes to a greater extent. Such a high degree of genetic determination raises the question as to the nature of the responsible loci. Identifying and characterizing the loci underlying this high heritability is an intense and fruitful field of investigation.

Using complex segregation analysis, two studies have determined that there is a single gene with a large effect on plasma Lp(a) concentrations. Morton *et al.* [31] investigated the segregation of plasma Lp(a) levels among members of 227 Japanese-American families, and reported the existence of a dominant major gene with residual polygenic effects. In addition, some evidence for a third allele was indicated. Formal analysis incorporating three alleles affecting plasma Lp(a) concentrations was reported by Hasstedt and Williams [32] using a single large Utah pedigree. In this pedigree, a full 73% of the variance of plasma Lp(a) concentrations was attributable to the effects of the major gene with three alleles. An additional 26% was caused by the effects of polygenes. Drayna *et al.* [33] and Weitkamp *et al.* [34] reported that the major gene influencing plasma Lp(a) concentrations was genetically linked to a marker near the apo(a) structural gene. It is important to keep in mind that in all these biometrical genetic studies DNA variation in known genes was not measured or directly assessed. Rather, genetic effects were inferred by the pattern of aggregation or segregation of the trait among family members.

A measured genotype approach

In 1986, Boerwinkle *et al.* [35] outlined a measured genotype approach whereby physiologically important variability in a candidate gene or gene product is directly employed in the genetic analysis of CHD and its risk factors. For plasma Lp(a) concentrations, the primary goal of the measured genotype approach has been to establish the role of the apo(a) gene in determining interindividual variation of Lp(a) levels, the extent of familial aggregation of Lp(a) levels and the segregation of Lp(a) levels in pedigrees.

Using sodium dodecylsulfate polyacrylamide gel electrophoresis followed by immunoblotting with a poly- or monoclonal human anti-apo(a) antibody, several apo(a) isoforms are distinguishable in human plasma [36,37]. The isoforms range in apparent molecular weight between 400 and 700 kD, and plasma from a single individual shows only one or two major isoforms. Utermann *et al.* [37] described six isoforms designated (from smallest to largest) F, B, S1, S2, S3, and S4. Because of the limited sensitivity of the western blot assay, individuals with very low plasma Lp(a) levels exhibit no apo(a) isoforms. Therefore, an operational null allele was also postulated. In a sample of 473 individuals from the Tyrol region of Austria, no visible bands were observed in approximately 25% of the individuals [38]. Considering only the observed single band types, the B, S1, S2, S3, and S4 isoforms occurred at a frequency of 2%, 6%, 26%, 25%, and 41%, respectively. In addition, the distribution of plasma Lp(a) concentrations was significantly different among Lp(a) isoform phenotypes [37,38]. The molecular weights of the apo(a) isoforms were, in general, inversely correlated with plasma levels of circulating Lp(a). Again considering only the observed single band types, the B isoform individuals had the highest average Lp(a) level (59 mg/dl), followed by the S1 individuals (28 mg/dl), the S2 individuals (24 mg/dl), the S3 individuals (12 mg/dl), and the S4 individuals (7.5 mg/dl). Boerwinkle *et al.* [38] determined that approximately 42% of the variation in plasma Lp(a) levels was attributable to the genetically determined Lp(a) protein isoforms. These studies clearly established that the apo(a) gene itself was an important determinant of plasma Lp(a) concentrations. Variation in the apo(a) gene and protein is responsible for the unmeasured major gene effect described by Morton *et al.* [31] and Hasstedt and Williams [32]. In addition there is some correspondence of frequency and effects between the measured apo(a) isoform described by Utermann *et al.* [37] and the major gene alleles inferred by Hasstedt and Williams [32]. Similar frequencies and effects can be imagined between the null apo(a) isoform and the low allele detected by segregation analysis, the S3 and S4 isoforms and the medium allele, and the S1 and S2 isoforms and the high allele.

Sandholzer *et al.* [39•] recently characterized the frequency and effects of the apo(a) protein isoforms in seven ethnic groups (Tyrolean, Icelandics, Hungarian, Maly, Chinese, Indian, and Black Sudanese). The distribution of plasma Lp(a) concentrations was significantly different among these groups, with the Chinese having the lowest (7.0 mg/dl) and the Sudanese the highest (46 mg/dl) Lp(a) levels. Even though the frequency of the apo(a) size isoforms were significantly different among the seven populations, their effects on plasma Lp(a) concentrations were not different. The authors conclude that the observed differences in the distribution of plasma Lp(a) concentrations could not be accounted for by differences in the frequencies or effects of the apo(a) protein isoforms.

Seed *et al.* [40] studied the association of Lp(a) concentrations and apo(a) isoforms with CHD in patients with heterozygous familial hypercholesterolemia (FH). Heterozygotes for defects in the LDL receptor gene leading to familial hypercholesterolemia have elevated LDL cholesterol levels and are at greatly increased risk of premature CHD. However, not all these individuals have diagnosed CHD; there is considerable variability in both the age of onset and the severity of disease [41]. In a sample of 115 FH heterozygotes, plasma Lp(a) concentration was the best predictor of angiographically documented CHD. The median Lp(a) level in patients with CHD was 57 mg/dl whereas it was only 18 mg/dl in those without disease. In addition, there was an increased frequency of the apo(a) isoforms associated with elevated Lp(a) levels in FH patients with CHD (Table 2). For example, the S2 isoform, which is typically associated with elevated Lp(a) levels (see below), was found more frequently in the patients with CHD, and the S4 isoform, which is typically associated with reduced Lp(a) levels, was found more frequently in the patients without CHD. Seed *et al.* [40] conclude that Lp(a) is a genetically determined risk factor for atherosclerosis in individuals with elevated serum LDL cholesterol levels, and that the Lp(a) system is at least partially responsible for the variability in expression of LDL receptor mutations. The results from Seed *et al.* [40] have been supported by a study of 120 FH heterozygotes by Wiklund *et al.* [42]. In addition Wiklund *et al.* show that cholesterol-lowering therapy in

the FH heterozygotes had no effect in reducing plasma Lp(a) concentrations. However, these results have not been supported by Mbewu *et al.* [43] who show that in a sample of 60 FH heterozygotes, those with CHD did not have significantly higher plasma Lp(a) concentrations than those without CHD.

Table 2. Apolipoprotein (a) isoform frequencies in FH patients with and without CHD.

	Frequency	
Isoform	CHD (n = 54)	No CHD (n = 55)
B	0.009	0.009
S1	0.048	0.000
S2	0.325	0.117
S3	0.169	0.194
S4	0.154	0.270
'null'	0.294	0.410

Data from Seed *et al.* [40].

Considerable insight into the role of the apo(a) gene in determining plasma Lp(a) concentrations has been obtained using the apo(a) protein isoforms. However, these analyses were compromised by the fact that the immunoblotting technique used to type the Lp(a) isoforms was not sensitive enough to detect those associated with low plasma Lp(a) concentrations. In general, and considering the limits of the sensitivity of the western blot assay, the apo(a) isoforms segregated in families as a Mendelian trait [37]. However, strict Mendelian segregation of the Lp(a) protein isoforms has been questioned [44]. In addition and undoubtedly because of the threshold of sensitivity of the western blot assay, the observed phenotype frequencies were significantly different from those expected under Hardy-Weinberg equilibrium. Clearly, a more sensitive and specific method of typing the apo(a) polymorphism was needed.

McLean *et al.* [8] proposed that the size differences observed in the apo(a) protein were due to variable numbers of kringle 4 repeats in the apo(a) gene. Various studies have provided experimental data which confirm this initial proposal [9•,45,46,47••]. Lackner *et al.* [47••] identified a single *Kpn*I restriction fragment containing most, if not all, the kringle-4-encoding sequences of the apo(a) gene. Using restriction enzymes that cut at a single place in the kringle-4-encoding region (for example *Pvu*II) it was inferred that a single kringle 4-encoding domain is approximately 5.5 kb in length. In contrast, *Kpn*I digested genomic DNA yielded bands ranging from 40 to 200 kb in length. Using carefully controlled pulsed-field gel electrophoresis conditions, 19 different-sized apo(a) fragments were observed in a sample of 102 Caucasian Americans. The 19 alleles formed a ladder with adjacent bands usually differing by 5 to 6 kb in length. No individuals showed more than two bands and the bands were inherited in a manner consistent with autosomal codominant segregation. The frequency distribution of apo(a) alleles observed in this sample is shown in Table 3. Alleles 12–15 were the most common, making up 50% of the total number of alleles. In general, smaller apo(a) alleles were not as frequent as larger alleles, giving the distribution a negatively skewed appearance. The observed apo(a) genotype frequencies agreed with Hardy-Weinberg expectations. This latter result is important in light of the lack of fit of the Lp(a) isoform frequencies to Hardy-Weinberg [38], and it indicates that the pulsed field typing technique is sensitive enough to detect all apo(a) length alleles. Size variation in the apo(a) *Kpn*I fragment indicates that the number of kringle-4-encoding domains in the apo(a) gene ranges from approximately nine to 35 copies. Therefore, apo(a) is a highly polymorphic gene with at least 19 alleles differing because of a variable number of tandemly repeated kringle-4-encoding intron-exon domains. A larger population-based survey of apo(a) gene length is likely to reveal that more than 19 alleles exist, as the sample used by Lackner *et al.* [47••] was relatively small and some adjacent bands differed by more than 5.5 kb in length. Indeed, using high resolution sodium dodecylsulfate agarose electrophoresis Kamboh *et al.* [48] observed 23 different protein isoforms in a sample of 270 Caucasian individuals.

Table 3. Apo(a) allele frequencies as determined by pulsed field gel electrophoresis.

Allele size*	Frequency
1	0.005
2	0.015
3	0.010
4	0.029
5	0.044
6	0.039
7	0.029
8	0.059
9	0.044
10	0.069
11	0.049
12	0.137
13	0.118
14	0.152
15	0.093
16	0.044
17	0.025
18	0.034
19	0.005

Data from Lackner *et al.* [45••]. *Sizes are given in relative units from smallest to largest.

Lackner *et al.* [47••] also report an inverse relationship between the size of the apo(a) allele and plasma Lp(a) levels. Figure 2 shows a three-dimensional graph of the size of the apo(a) alleles as measured by pulsed-field gel electrophoresis and the level of plasma Lp(a). Individuals with at least one small apo(a) allele tend to have relatively high plasma Lp(a) levels; whereas individuals with large apo(a) alleles have reduced Lp(a) levels. The rank correlation between the length of the apo(a) alleles and plasma Lp(a) levels was −0.4. As is evident in Figure 2, there are several notable exceptions to this general trend. For example, there is one 14/17 heterozygote with an Lp(a) concentration of approximately 40 mg/dl. This concentration is far greater than one would have predicted based on the apo(a) genotype alone. The factors responsible for these discrepancies may be environmental in nature or the result of other genetic determinants. In addition, if there are other genetic loci influencing Lp(a) levels they may be within the apo(a) gene alongside the length variation or coded for by other genes.

In order to quantitate the contribution of the entire apo(a) gene to plasma Lp(a) concentrations we have analyzed the segregation of the apo(a) gene and Lp(a) levels in several families [47••]. Examination of the family data indicate that in addition to the number of kringle 4 repeats there are other sites in the apo(a) gene which influence plasma Lp(a) concentrations. Table 4 gives the apo(a) types and Lp(a) levels in siblings with identical and discordant apo(a) types. Sibling pairs with identical apo(a) types were highly concordant, whereas sibling pairs in the same families who were discordant for their apo(a) types often had very different Lp(a) levels. In a sample of 73 sibling pairs that share both apo(a) alleles identical by descent, the correlation of plasma Lp(a) concentrations was very high (0.95; Boerwinkle E and Hobbs HH, unpublished data). In contrast, in a sample of 52 sibling pairs with no apo(a) alleles identical by descent, the correlation coefficient of plasma Lp(a) concentrations was low and negative (−0.23). The high degree of similarity of Lp(a) levels among sibling pairs with identical apo(a) genotypes suggests that factors at the apo(a) gene other than the number of kringle 4 repeats are contributing to plasma Lp(a) levels. In other words, apo(a) alleles of the same length (as determined by pulsed-field gel electrophoresis) do not necessarily have the same DNA sequence. There may be sequence differences in the apo(a) promoter that influence the transcription of the gene or in other sequences that alter the transport and metabolism of the Lp(a) particle. The identity of these sequences is currently unknown.

A synthesis

Genetic architecture is defined as the number and type (e.g. structural or regulatory) of polymorphic genes af-

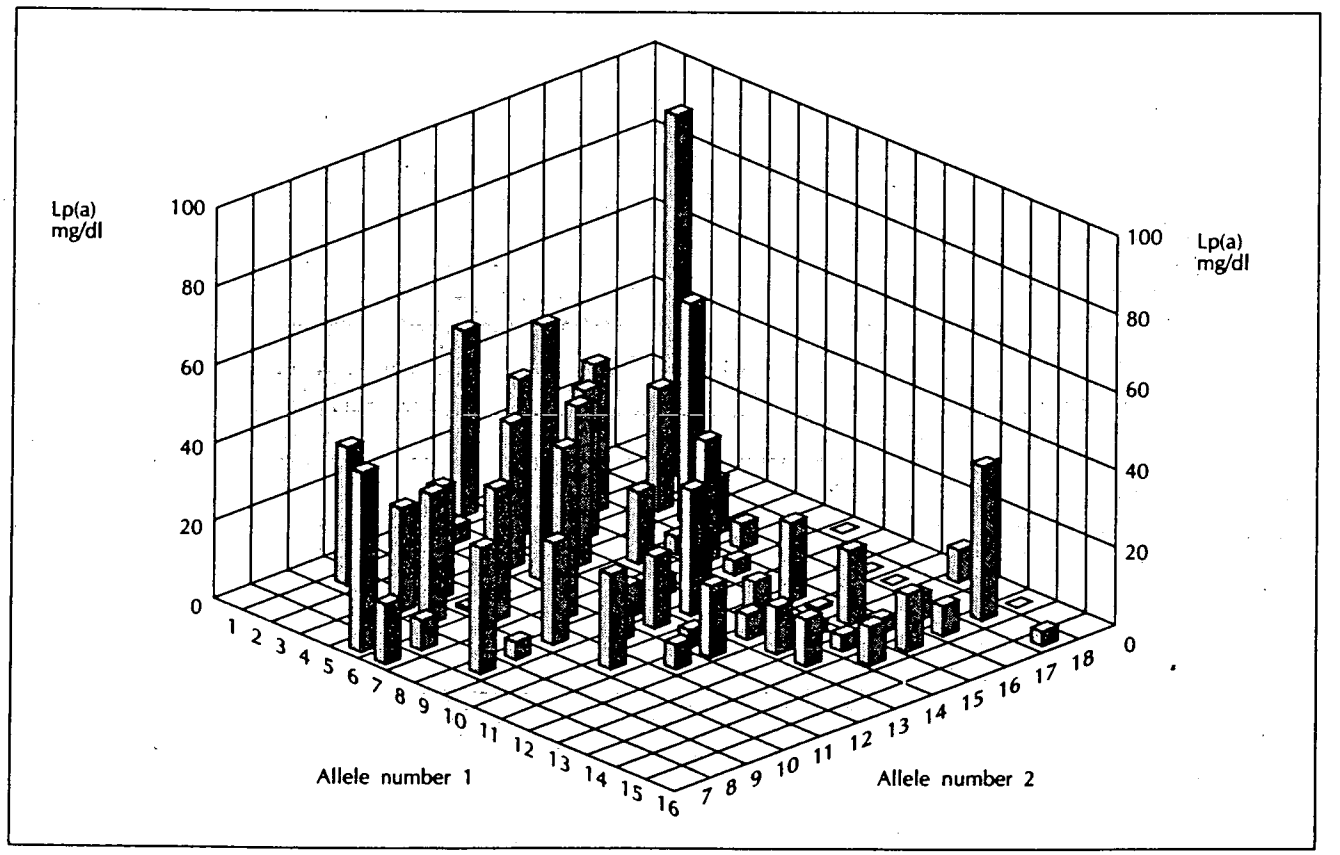


Fig. 2. Association between apo(a) length genotypes and plasma Lp(a) levels. Each cartesian coordinate represents an apo(a) genotype and Lp(a) levels are plotted in the vertical axis.

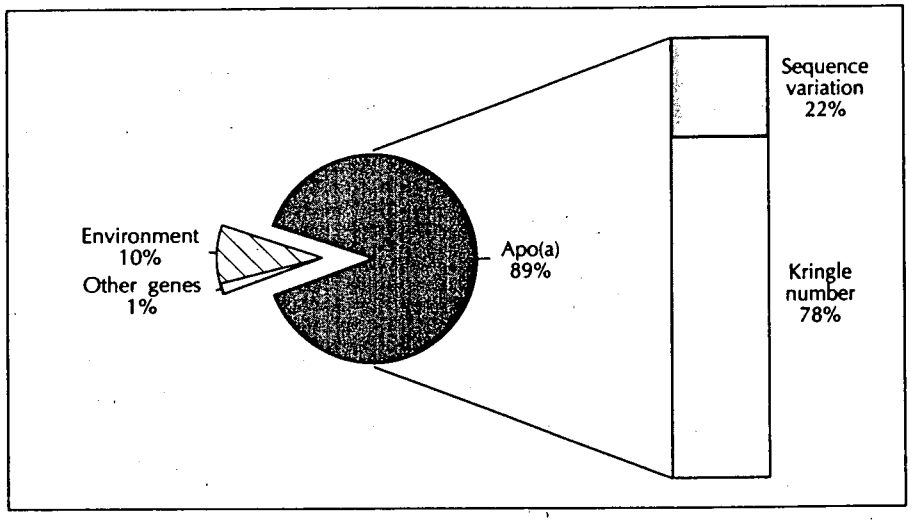


Fig. 3. Summary of the genetic architecture of plasma Lp(a) concentrations. The entire pie represents the interindividual variance of plasma Lp(a) concentrations. Approximately 90% of this variance is attributable to genetic factors, and virtually all of this is due to effects of the apo(a) gene. Other genes, such as the LDL receptor gene [52], have a slight effect on Lp(a) concentrations. The remaining 10% is due to environmental factors. The effects of the apo(a) gene may be partitioned into two components: length and sequence variation.

Table 4. Comparison of Lp(a) levels among sibling pairs.

	Concordant siblings		Other siblings	
Family	Apo(a) type	Lp(a) (mg/dl)	Apo(a) type	Lp(a) (mg/dl)
1	4/10	41;42	10/14;6/14	< 1;16
2	9/16	7;9	5/9;4/5	50;75
3	12/14	6;9	8/12	< 1
4	12/15	1;1	14/15,7/12	< 1;28
5	14/17	1;1	14/15,15/17	2;3
6	6/16	5;6	15/15	6
7	2/18	< 1; < 1	15/18	< 1
8	11/15	< 1; < 1	15/17,8/11	< 1;7
9	14/15	< 1; < 1	15/18	< 1

Data from Lackner *et al.* [45••].

fecting a trait, the number of alleles at each locus, the frequencies of the alleles, and the size of their effects [49,50]. Our understanding of the genetics of plasma Lp(a) concentrations has evolved in parallel with the ability to measure accurately both the quantity of Lp(a) in plasma and characteristics of the apo(a) gene. The genetic architecture of plasma Lp(a) concentrations has been defined at three levels: polygenic heritability, role of *in toto* apo(a) gene variation, and effects of length variation in the apo(a) gene (Fig. 3). As already mentioned, a number of investigators have concluded that the Lp(a) phenotype and Lp(a) levels are highly genetically determined. A consensus value for the polygenic heritability of plasma Lp(a) concentrations is approximately 90%; most studies report estimates close to this value. Hasstedt and Williams [32] conclude on the basis of segregation analysis of a large Utah pedigree that the majority of this heritability is attributable to the effects of a single major gene, and this gene is linked to the structural gene for plasminogen [33, 34]. This result takes on added importance with the knowledge that the structural gene for apo(a), the unique protein component of Lp(a), is closely linked to plasminogen [51]. Utermann *et al.* [37] described several electrophoretically separable isoforms in Lp(a), and Boerwinkle *et al.* [38] estimated that these isoforms accounted for 42% of the interindividual variation in plasma Lp(a) concentrations. Utermann *et al.* [52] demonstrated that mutations in the LDL receptor gene also influence plasma Lp(a) levels. However, these are likely to have only a small effect in the general population because their frequency is rare. The identification of a *KpnI* restriction fragment containing most if not all the kringle-4-encoding domain of the apo(a) gene [47••] and the demonstration that this fragment was highly polymorphic in length provided a powerful tool in the genetic analysis of plasma Lp(a) concentrations. By examining the degree of resemblance of sibling pairs sharing both alleles identical by descent as compared with siblings with no apo(a) alleles identical by descent, we can infer that virtually all the variability of plasma Lp(a) levels is due to variation in the apo(a) gene. Most, but not all, of this can be attributed to differing number of kringle 4 repeats. However, cis-acting sequence variation in this gene also affects plasma Lp(a) concentrations. These data represent the most complete picture to date of the genetic architecture of a major CHD risk factor.

Acknowledgements

This work was carried out in part through the assistance of the grants NIH HL-40613, NIJ 90-IJ-CX-0038, and Research Career Development Award. Eric Boerwinkle is an Established Investigator of the American Heart Association. Eric Boerwinkle would also like to thank Drs Gerd Utermann and Helen Hobbs for their collaboration in this work.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. NATIONAL CENTER FOR HEALTH STATISTICS: **Vital Statistics of the United States, 1988, Vol II. Mortality.** Public Health Service, Washington DC, 1990.

2. BERG K: **A New Serum Type System in Man: The Lp System.** *Acta Pathol Microbiol Scand* 1963, 59:369–382.

3. MILES LA, FLESS GM, LEVIN EG, SCANU AM, PLOW EF: **A Potential Basis for the Thrombotic Risks Associated with Lipoprotein (a).** *Nature* 1989, 339:301–305.

4. ALBERS JJ, ADOLPHSON JL, HAZARD WR: **Radio-Immunoassay of Human Plasma Lp(a) Lipoprotein.** *J Lipid Res* 1977, 18:331–338.

5. FLESS GM, ZUM MALLEN ME, SCANU AM: **Physiological Properties of Apolipoprotein(a) and Lipoprotein(a-) Derived from the Dissociation of Human Plasma Lipoprotein (a).** *J Biol Chem* 1986, 261:8712–8718.

6. UTERMANN G, WEBER W: **Protein Composition of Lp(a) Lipoprotein for Human Plasma.** *FEBS Lett* 1983, 154:357–361.

7. GAUBATZ JW, HEIDEMAN C, GOTTO AM, MORRISETT JD, DAHLEN GH: **Human Plasma Lipoprotein (a): Structural Properties.** *J Biol Chem* 1983, 258:4582–4589.

8. MCLEAN JW, TOMLINSON JE, KUANG WJ, EATON DL, CHEN EY, FLESS GM, SCANU AM, LAWN RM: **cDNA Sequence of Human Apolipoprotein(a) is Homologous to Plasminogen.** *Nature* 1987, 330:132–137.

9. • KOSCHINSKY M, BEISIEGEL U, HENNE-BRUNS D, EATON DL, LAWN RM: **Apolipoprotein(a) Size Heterogeneity is Related to Variable Number of Repeat Sequences in its mRNA.** *Biochemistry* 1990, 29:640–644.

This paper reports data providing direct and definitive evidence that the size polymorphism in apo(a) was attributable to length variation in the apo(a) transcript.

10. KOSTNER GM: **Apolipoproteins and Lipoproteins of Human Plasma: Significance in Health and in Disease.** *Adv Lipid Res* 1983, 20:1–43.

11. RAINWATER DL, LANFORD RE: **Production of Lipoprotein(a) by Primary Baboon hepatocytes.** *Biochim Biophys Acta* 1989, 1003:30–35.

12. TOMLINSON JE, MCLEAN JW, LAWN R: **Rhesus Monkey Apolipoprotein(a).** *J Biol Chem* 1989, 264:5957–5965.

13. EATON DL, FLESS GM, KOHR WJ, MCLEAN JW, XU QT, MILLER CG, LAWN RM, SCANU AM: **Partial Amino Aacid Sequence of Apolipoprotein (a) Shows that it is Homologous to Plasminogen.** *Proc Natl Acad Sci U S A* 1987, 84:3224–3228.

14. MORRISETT JD, GUYTON JR, GAUBATZ JW, GOTTO AM: **Lipoprotein(a): Structure, Metabolism, and Epidemiology.** In *Plasma Lipoproteins.* Edited by Gotto AM Jr. Amsterdam: Elsevier; 1987:129–152.

15. • ROSENGREN A, WILHELMSEN L, ERIKSSON E, RISBERG B, WEDEL H: **Lipoprotein(a) and Coronary Heart Disease: A Prospective Case–Control Study in a General Population Sample of Middle Aged Men.** *BMJ* 1990, 301:1248–1251.

To my knowledge, this paper reports the results of the only longitudinal prospective study demonstrating that plasma Lp(a) concentrations are a risk factor for CHD.

16. MAEDA S, ABE A, SEISHIMA K, MAKINO K, NOMA A, KAWADE M: **Transient Changes of Serum Lipoprotein(a) as an Acute Phase Protein.** *Atherosclerosis* 1989, 78:145–150.

17. ARMSTRONG VW, CREMER P, EBERLE E, MANKE A, SCHULZE F, WIELAND H, KREUZER H, SEIDEL D: **The Association Between Serum Lp(a) and Angiographically Assessed Coronary Atherosclerosis.** *Atherosclerosis* 1986, 62:249–257.

18. ZENKER G, KOLTRINGER P, BONE G, NIEDERKORN K, PFEIFFER K, JURGENS G: **Lipoprotein(a) as a Strong Indicator for Cerebrovascular Disease.** *Stroke* 1986, 17:942–945.

19. PARRA HJ, MEZDOUR H, CACHERA C, DRACON M, TACQUET A, FRUCHART JC: **Lp(a) Lipoprotein in Patients with Chronic Renal Failure Treated by Hemodialysis.** *Clin Chem* 1987, 33:721–721.

20. KROLEWSKI AS, WARRAM JH, RAND LI, KHAN CR: **Epidemiological Approach to the Etiology of Type I Diabetes Mellitus and its Complications.** *N Engl J Med* 1987, 317:1390–1398.

21. HAFFNER SM, TUTTLE KR, RAINWATER DL: **Decrease of Lipoprotein (a) with Improved Glycemic Control in IDDM Subjects.** *Diabetes Care* 1991, 14:302–307.

22. BRUCKERT E, DAVIDOFF P, GRIMALDI A, TRUFFERT J, GIRAL P, DOUMITH R, THERVET F, DE GENNES JL: **Increased Serum Levels of Lipoprotein(a) in Diabetes Mellitus and their Reduction with Glycemic Control.** *JAMA* 1990, 263:35–36.

23. JOVEN J, VILELLA E: **Serum Levels of Lipoprotein(a) In Patients with Well Controlled Non-Insulin-Dependent Diabetes Mellitus.** *JAMA* 1991, 265:1113–1114.

24. LEVITSKY LL, SCANU AM, GOULD SH: **Lipoprotein (a) Levels in Black and White Children and Adolescents with IDDM.** *Diabetes Care* 1991, 14:283–287.

25. LALOUEL JM, RAO DC, MORTON NE, ELSTON RC: **A Unified Model for Complex Segregation Analysis.** *Am J Hum Genet* 1983, 35:816–826.

26. BERG K, MOHR J: **Genetics of the Lp System.** *Acta Genet* 1963, 13:349–360.

27. ALBERS JJ, WAHL P, HAZZARD WR: **Quantitative Genetic Studies of the Human Plasma Lp(a) Lipoprotein.** *Biochem Genet* 1974, ii:475–486.

28. SING CF, SCHULTZ JS, SHREFFLER DC: **The Genetics of the Lp Antigen. II. A Family Study and Proposed Models of Genetic Control.** *Ann Hum Genet* 1974, 38:47–56.

29. MORTON NE, GULBRANDSEN CL, RHOADS GG, KAGAN A: **The Lp Lipoprotein in Japanese.** *Clin Genet* 1978, 14:207–212.

30. ISELIUS L, DAHLEN G, DEFAIRE U, LUNDMAN T: **Complex Segregation Analysis of the Lp(a)/pre-$_1$-lipoprotein Trait.** *Clin Genet* 1981, 20:147–151.

31. MORTON NE, BERG K, DAHLEN G, FERRELL RE, RHOADS GG: **Genetics of the Lp Lipoprotein in Japanese-Americans.** *Genet Epidemiol* 1985 2:113–121.

32. HASSTEDT SJ, WILLIAMS RR: **Three Alleles For Quantitative Lp(a).** *Genet Epidemiol* 1986, 3:53–55.

33. DRAYNA DT, HEGELE RA, HASS PE, EMI M, WU LL, EATON DL, LAWN RM, WILLIAMS RR, WHITE RL, LALOUEL J-M: **Genetic Linkage Between Lipoprotein(a) Phenotype and a DNA Polymorphism in the Plasminogen Gene.** *Genomics* 1988, 3:230–236.

34. WEITKAMP LR, GUTTORMSEN SA, SCHULTZ JS: **Linkage Between the Loci for the Lp(a) Lipoprotein (LP) and Plasminogen (PLG).** *Hum Genet* 1988, 79:80–82.

35. BOERWINKLE E, CHAKRABORTY R, SING CF: **The Use of Measured Genotype Information in the Analysis of Quantitative Phenotypes in Man. I. Models and Analytical Methods.** *Ann Hum Genet* 1986, 50:181–194.

36. FLESS GM, ROLIH CA, SCANU AM: **Heterogeneity of Human Plasma Lipoprotein (a).** *J Biol Chem* 1984, 259:11470–11478.

37. UTERMANN G, MENZEL HJ, KRAFT HG, DUBA C, KEMMLER HG, SEITZ C: **Lp(a) Glycoprotein Phenotypes: Inheritance and Relation to Lp(a) Lipoprotein Concentrations in Plasma.** *J Clin Invest* 1987, 80:458–465.

38. BOERWINKLE E, MENZEL HG, KRAFT HG, UTERMANN G: Genetics of the Quantitative Lp(a) Lipoprotein Trait. III. Contribution of Lp(a) Glycoprotein Phenotypes to Normal Lipid Variation. *Hum Genet* 1989, 82:73–78.

39. • SANDHOLZER C, HALLMAN DM, SAHA N, SIGURDSSON G, LACKNER C, CSASZAR A, BOERWINKLE E, UTERMANN G: Effects of the Apolipoprotein(a) Size Polymorphism of the Lipoprotein(a) Concentration in 7 Ethnic Groups. *Hum Genet* 1991, 86:607–614.

This paper demonstrates that apo(a) is polymorphic in multiple ethnic and racial groups. However, the estimated allele frequencies are significantly different among the populations studied. In addition, the influence of the apo(a) length variation on plasma Lp(a) levels is present in all groups.

40. SEED M, HOPPICHLER F, REAVELEY D, MCCARTHY S, THOMPSON GR, BOERWINKLE E, UTERMANN G: Relation of Serum Lipoprotein(a) Concentration and Apolipoprotein(a) Phenotype to Coronary Heart Disease in Patients with Familial Hypercholesterolemia. *N Engl J Med* 1990, 322:1494–1499.

41. THOMPSON GR, SEED M, NITHYNANTHAN S, MCCARTHY S, THOROGOOD M: Genotypic and Phenotypic Variation in Familial Hypercholesterolemia. *Arteriosclerosis* 1989, 9:175–180.

42. WIKLUND O, ANGELIN B, OLOFSSON S-O, ERIKSSON M, FAGER G, BERLUND L, BONDJERS: Apolipoprotein(a) and Ischemic Heart Disease in Familial Hypercholesterolemia. *Lancet* 1990, 335:1360–1363.

43. MBEWU AD, BHATNAGAR D, DURRINGTON PN, HUNT L, ISHOLA M, ARROL S, MACKNESS M, LOCKLEY P, MILLER JP: Serum Lipoprotein(a) in Patients Heterozygous for Familial Hypercholesterolemia, Their Relatives, and Unrelated Control Populations. *Arteriosclerosis Thromb* 1991, 11:940–946.

44. GAUBATZ JW, GHANEM KI, GUEVARA J, NAVA ML, PATSCH W, MORRISETT JD: Polymorphic Forms of Human Apolipoprotein(a): Inheritance and Relationship of their Molecular Weights to Plasma Levels of Lipoprotein(a). *J Lipid Res* 1990, 31:603–613.

45. HIXSON JE, BRITTEN ML, MANIS GS, RAINWATER DL: Apolipoprotein(a) (Apo(a)) Glycoprotein Isoforms Result from Size Differences in Apo(a) mRNA in Baboons. *J Biol Chem* 1989, 264:6013–6016.

46. GAVISH D, AZROLAN N, BRESLOW JL: Plasma Lp(a) Concentration is Inversely Correlated with the Ratio of Kringle IV/Kringle V Encoding Domains in the Apo(a) Gene. *J Clin Invest* 1989, 84:2021–2027.

47. •• LACKNER C, BOERWINKLE E, LEFFERT CC, RAHMIG T, HOBBS HH: Molecular Basis of Apolipoprotein(a) Isoform Size Heterogeneity as Revealed by Pulsed-field Gel Electrophoresis. *J Clin Invest* 1991, 87:2077–2086.

The authors have identified a single *KpnI* restriction fragment containing most, if not all, of the kringle-4-encoding sequences of the apo(a) gene. Using pulsed-field gel electrophoresis conditions, 19 different sized apo(a) DNA fragments ranging in size between 40 and 200 kb were observed in a sample of 102 Caucasian Americans. This size variation in the apo(a) *KpnI* fragment indicates that the number of kringle-4-encoding domains in the apo(a) gene ranges from approximately nine to 35 copies. Apo(a) is a highly polymorphic gene with at least 19 alleles differing because of a variable number of tandemly repeated kringle-4-encoding intron-exon domains. The authors also report an inverse relationship between the size of the apo(a) allele and plasma Lp(a) levels.

48. KAMBOH MI, FERRELL RE, KOTTKE BA: Expressed Hypervariable Polymorphism of Apolipoprotein(a). *Am J Hum Genet* 1991, 49:1063–1074.

49. BOERWINKLE E, SING CF: The Use of Measured Genotype Information in the Analysis of Quantitative Phenotypes in Man. III. Simulataneous Estimation of the Frequencies and Effects of the Apolipoprotein E Polymorphism and Residual Polygenic Effects on Cholesterol, Betalipoprotein, and Triglyceride Levels. *Ann Hum Genet* 1987, 51:211–226.

50. SING CF, MOLL PP: Genetics of Atherosclerosis. *Annu Rev Genet* 1990, 24:171–187.

51. FRANK SL, KLISAK I, SPARKES RS, MOHANDAS T, TOMLINSON JE, MCLEAN JW, LAWN RM, LUSIS AJ: The Apolipoprotein(a) Gene Resides on Human Chromosome 6q26-27, in Close Proximity to the Homologous Gene for Plasminogen. *Hum Genet* 1988, 79:352–356.

52. UTERMANN G, HOPPICHLER F, DIEPLINGER H, SEED M, THOMPSON G, BOERWINKLE E: Defect in the Low Density Lipoprotein Receptor Gene Affect Lipoprotein (a) Levels: Multiplicative Interaction of Two Gene Loci Associated with Premature Atherosclerosis. *Proc Natl Acad Sci U S A* 1989, 86:4171–4174.

E. Boerwinkle, Center for Demographic and Population Genetics, The University of Texas Health Science Center in Houston, PO Box 2033, Houston, TX 77225, USA.

Reprinted from

Current Opinion in LIPIDOLOGY

Volume 3, 1992

CS
CURRENT
SCIENCE

Genetics of plasma lipoprotein (a) concentrations

Eric Boerwinkle

Center for Demographic and Populations Genetics, The University of Texas Health Science Center in Houston, Houston, USA

Lipoprotein (a) [Lp(a)] is a low-density-like lipoprotein with the addition of a Lp(a)-specific protein, apolipoprotein (a) [apo(a)]. Plasma concentrations of Lp(a) are a risk factor for premature coronary heart disease (CHD). Genetic studies have been key in developing an understanding of Lp(a) and its association with CHD. Quantitative genetic studies have demonstrated that plasma Lp(a) concentrations are largely genetically determined. Using length variation in the apo(a) gene and gene product, it has been shown that the apo(a) gene has a large effect on Lp(a) levels, and in fact variation in this one gene is largely responsible for the high heritability of plasma Lp(a) concentrations. The data presented in this review, represent the most complete picture to date of the genetic architecture of a major CHD risk factor.

Current Opinion in Lipidology 1992, **3**:128–136

Introduction

In 1988, coronary heart disease (CHD) accounted for 35.3% of all deaths in the USA [1]. Unlike the inborn errors of metabolism (for example cystic fibrosis) CHD is the result of lifelong interactions among numerous environmental and genetic factors. Therefore, no single gene is responsible for the high frequency of CHD. Rather, there are many genes, each making a relatively small contribution to the overall disease liability in the population. Genetic studies of CHD are also complicated by the fact that there may be different causes in different individuals, families and populations; i.e. there are multiple paths to the same disease endpoint. The multifactorial nature of CHD has slowed progress in understanding its etiology; however, progress has been best realized when researchers have focused on one or only a few facets of this complex system. Since its discovery by Berg [2] in 1963, lipoprotein(a) [Lp(a)] has been at the forefront of research on the genetics of CHD and its risk factors.

Lp(a) is a cholesteryl ester-rich plasma lipoprotein composed of two components: an LDL-like particle to which is attached a single large glycoprotein, apolipoprotein (a) [apo(a)]. Lp(a) is an important risk factor for atherosclerotic CHD. High plasma levels of Lp(a) are positively associated with the development of premature atherosclerosis and other vascular diseases. The mechanism by which Lp(a) contributes to the atherosclerotic process is unknown, although it is probably mediated through its close homology to the plasma zymogen plasminogen [3]. Lp(a) is distinct from other CHD risk factors associated with intermediary metabolism in that plasma concentrations of Lp(a) vary over a very wide range among individuals but are extremely stable within a given individual [4]. Furthermore, many physiological, pharmacological and environmental factors that effect the levels of other plasma lipoproteins have no effect on Lp(a) concentration.

Our goals in studying the genetics of common chronic diseases such as CHD include a better understanding of their etiology and improved prediction. Such an understanding will better position society and medicine for preventative, rather than pharmacological, surgical or mere palliative measures. Throughout its relatively brief history, human genetics has led the way in studies of Lp(a) and its association with CHD. Knowledge accumulating about the genetics of plasma Lp(a) concentrations and the role of this lipoprotein in atherosclerosis will at least partially unveil the mechanisms of CHD initiation and progression, and will facilitate early identification of individuals at increased risk.

Structure of lipoprotein (a)

A schematic diagram of a Lp(a) particle is shown in Fig. 1. In many respects, the Lp(a) lipoprotein particle resembles a LDL particle with the addition of a single

Abbreviations

apo—apolipoprotein; **CHD**—coronary heart disease; **FH**—familial hypercholesterolemia; **Lp(a)**—lipoprotein (a); **mRNA**—messenger RNA.

Current Science Ltd ISSN 0957-9672

molecule of apo(a). The physicochemical properties of Lp(a) have been determined using a variety of techniques [5], and Table 1 compares these properties between Lp(a) and LDL. As in LDL, apolipoprotein (apo)B_{100} is present in Lp(a) [6] and, unlike LDL, is linked to apo(a) by a disulfide bond [7]. Lp(a) has a higher density and overall molecular weight than LDL, characteristics that make ultracentrifugation and gel filtration chromatography useful tools for its purification. The electrophoretic mobility of Lp(a) in agarose is to the pre-β position, which is similar to VLDL, and its buoyant density is characteristic of HDL_2. The protein and lipid composition of Lp(a) differs only slightly with that of LDL, with a protein: lipid ratio in Lp(a) and LDL of 1 : 2.2 and 1 : 3.5, respectively. Total lipid content represents approximately 69% in Lp(a) and 79% in LDL, but the amount of free cholesterol relative to total lipids is approximately the same in both particles (11.5 and 11.8%, respectively).

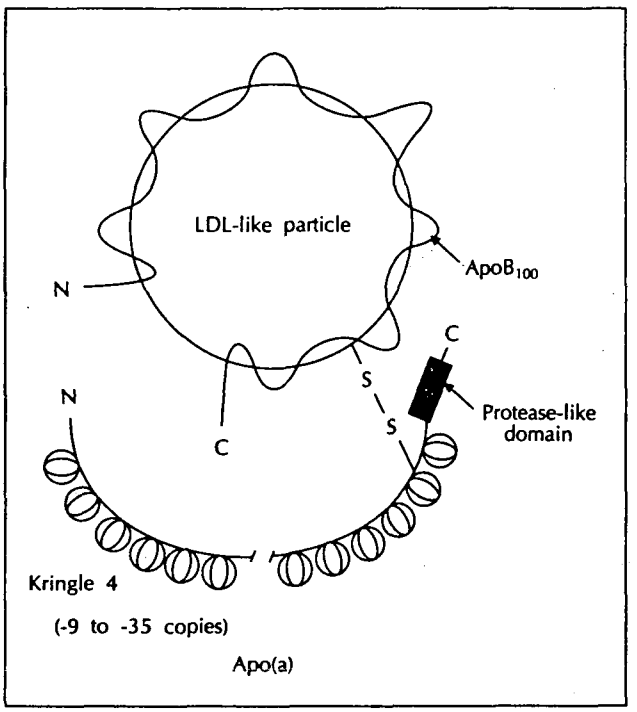


Fig. 1. Schematic diagram of an Lp(a) lipoprotein particle.

Data concerning the site of synthesis of Lp(a) in humans are scarce. The liver is thought to be involved in the production of Lp(a) because apo(a) messenger RNA (mRNA) has been detected by northern blot analysis in liver [8,9•] and because liver damage is associated with reduced Lp(a) levels [10]. However, it is unclear whether apo(a) synthesized in hepatocytes is secreted for later production of Lp(a) in the interstitia or in circulation, or whether the hepatocytes secrete Lp(a) directly. Rainwater and Lanford [11] reported that cultured primary baboon hepatocytes synthesized an Lp(a) particle identical to plasma Lp(a). In addition, the isoform type of the

Table 1. Physical and chemical properties of Lp(a) and LDL particles.

	Lp(a)	LDL
Electrophoretic mobility	pre-β	β
Buoyant density (g/ml)	1.05–1.08	1.03–1.06
Molecular weight ($\times 10^6$)	3.1–3.8	2.5–3.1
Apolipoproteins	$apoB_{100}$, apo(a)	$apoB_{100}$
% Composition as protein	26.0–35.7	20.7
% Composition as lipid	64.3–74.0	79.3
% Composition as cholesterol	33.6–45.4	52.3
(free cholesterol)	(7.6–10.2)	(9.4)

Data from Fless *et al.* [5] and Gaubatz *et al.* [7]

secreted Lp(a) was identical to the observed isoforms from the hepatocyte donor animal. To investigate the tissue expression of apo(a) mRNA in rhesus monkey, Tomlinson *et al.* [12] analyzed 12 different tissues and detected apo(a) mRNA only in liver, brain, and testes. However, no apoB mRNA was present in the latter two tissues suggesting that apo(a) may function in some tissues independently of the Lp(a) particle.

Partial protein sequencing of purified plasma apo(a) revealed remarkable sequence similarity with that of plasminogen [13], an observation which led to the cloning of the apo(a) complementary DNA by McLean *et al.* [8]. The mature apo(a) mRNA encodes a 4529-amino-acid protein which includes a 19 bp signal peptide, multiple copies of a kringle-4-like domain, one copy of a region homologous to the plasminogen kringle 5, and a protease domain. Complementary DNA sequences of the protease domain in apo(a) and plasminogen share 94% homology. This striking structural similarity with plasminogen led to the suggestion that Lp(a) might have prothrombic action. However, because of two important amino acid substitutions in the activation site (Arg–Val mutated to Ser–Ile), apo(a), unlike plasminogen, cannot be transformed into an active fibrinolytic agent. Miles *et al.* [3] demonstrated that Lp(a) was able to bind to plasminogen receptors distributed on peripheral blood cells and vascular endothelial cells, and postulated that Lp(a) competes with plasminogen for these receptors. Thus, by interfering with plasminogen binding, Lp(a) would prevent plasmin generation. Consequently, thrombolysis would be inhibited and thrombosis promoted. However, at this time we cannot rule out the possibility that the homology between apo(a) and plasminogen may be functionally coincidental and the mechanism for the association between Lp(a) and CHD is by some as-yet-unidentified path.

Lipoprotein (a) as a risk factor for atherosclerosis

The association between elevated plasma Lp(a) concentrations and the occurrence of CHD has been reported by numerous investigators employing a broad range of study designs, disease definitions, and analytical methods. A positive but moderate relationship between plasma Lp(a) concentrations and the occurrence of CHD has been shown clearly. The results of several cross-sectional epidemiological studies establishing Lp(a) as a risk factor for CHD were reviewed by Morrisett *et al.* [14]. Such a review will not be repeated here. Rather, two studies which help cultivate an appreciation for the complex relationship between Lp(a) and CHD are discussed below.

To our knowledge only one prospective study has examined the relationship between plasma Lp(a) concentrations and CHD. Rosengren *et al.* [15•] followed 776 50-year-old Swedish men for 6 years (for a total of 4656 follow-up years) and noted all CHD events and deaths due to CHD (26 in this study). Each CHD case was matched with four controls who did not suffer an identified CHD event. In this sample, the CHD-positive group had significantly higher Lp(a) levels at baseline (mean = 27.8 mg/dl) compared with the control group (mean = 17.3 mg/dl). In addition, those in the upper one-fifth of the population Lp(a) distribution had twice as much CHD than those in the lower four-fifths. These data underscore the role of Lp(a) as a risk factor for CHD in a well-designed prospective case–control study. Because Lp(a) levels were measured before the onset of acute disease, these results counter those that argue that the association between CHD and Lp(a) is a consequence of the potential role of Lp(a) as an acute phase protein [16]. However, it should be kept in mind that the scope of the study was quite narrow. For example, it only investigated the role of Lp(a) as a risk factor for CHD in men of one age. In fact, a review of the literature indicates a dearth of data documenting a role for Lp(a) in CHD in women or other groups. In addition, the control group in this study was not necessarily disease-free. There is a desperate need for a well-designed population-based prospective case–control study using angiographically documented atherosclerotic vascular disease cases and disease-free controls.

Lp(a), in its role as a risk factor for CHD, does not act alone. Armstrong *et al.* [17] reported that the association between elevated plasma Lp(a) levels and CHD was dependent on plasma LDL cholesterol levels. Individuals with both high Lp(a) and high LDL cholesterol had a sixfold increased risk of disease, whereas those with only one or the other risk factor had a 1.5- to twofold increased risk. These data suggest that there is interaction, either direct or indirect, between these two lipoprotein particles at the metabolic or cellular levels which increases CHD risk.

Plasma Lp(a) concentrations are also associated with other disease states such as cerebrovascular disease [18], renal disease [19], and diabetes mellitus. It is now well accepted that atherosclerotic vascular disease is the most common complication of diabetes [20]. Although Lp(a) represents a potent marker for and physiological element in the formation of atherosclerosis in non-diabetics, it has not been shown to be a risk factor for CHD in diabetics. Recent data suggest a relationship between diabetes and serum Lp(a) levels. Two independent studies [21,22] have reported high serum levels of Lp(a) in poorly controlled insulin-dependent diabetics compared with non-diabetic individuals. Interestingly, Lp(a) levels were dramatically decreased with improved metabolic control. A study by Joven and Vilella [23] showed that Lp(a) levels were not increased in well-controlled non-insulin-dependent diabetics and the authors suggest a predominant role of exogenous insulin in the regulation of Lp(a) levels in diabetic patients. Levitsky *et al.* [24] examined the relationship between Lp(a) levels, diabetes, and glycemic control among groups of white and black non-diabetic and insulin-dependent diabetic children. They found that circulating Lp(a) levels were increased in hyperglycemic whites, but no such relationship was observed in blacks. Although our understanding of the relationship between glycemic control and plasma Lp(a) concentrations is meager, Lp(a) should be considered as a factor contributing to CHD in this already-at-risk group.

The genetics of lipoprotein (a)

Biometrical genetic analyses

Traditional biometrical genetic analyses follow a logical series of progressively more focused questions about the role of genes influencing a quantitative trait. First, is there significant familial aggregation for the trait of interest and is this aggregation due to shared genes rather than shared environmental effects? The latter point is particularly important when investigating the risk factors for CHD as they are usually influenced by a wide variety of environmental factors. Second, is there evidence that one gene has a large influence on the quantitative phenotype? Such 'major gene' effects are usually detected using a method known as complex segregation analysis [25]. Finally, if there is a significant major gene effect, can it be localized in the human genome using linkage analysis? Each of these questions has been addressed for plasma Lp(a) concentrations.

The genetic nature of the Lp(a) phenotype was established shortly after its discovery [26]. Since that time, numerous studies have firmly established that Lp(a) levels are strongly influenced by genetic factors [27–30]. Based on the similarity of plasma Lp(a) concentration among related individuals, estimates of the polygenic heritability of Lp(a) typically range between 70% and 95%. In other words, between 70% and 95% of the interindividual variation in plasma Lp(a) levels is attributable to genetic differences among individuals. No other major CHD risk factor is influenced by genes to a greater extent. Such a high degree of genetic determination raises the question as to the nature of the responsible loci. Identifying and characterizing the loci underlying this high heritability is an intense and fruitful field of investigation.

Using complex segregation analysis, two studies have determined that there is a single gene with a large effect on plasma Lp(a) concentrations. Morton *et al.* [31] investigated the segregation of plasma Lp(a) levels among members of 227 Japanese-American families, and reported the existence of a dominant major gene with residual polygenic effects. In addition, some evidence for a third allele was indicated. Formal analysis incorporating three alleles affecting plasma Lp(a) concentrations was reported by Hasstedt and Williams [32] using a single large Utah pedigree. In this pedigree, a full 73% of the variance of plasma Lp(a) concentrations was attributable to the effects of the major gene with three alleles. An additional 26% was caused by the effects of polygenes. Drayna *et al.* [33] and Weitkamp *et al.* [34] reported that the major gene influencing plasma Lp(a) concentrations was genetically linked to a marker near the apo(a) structural gene. It is important to keep in mind that in all these biometrical genetic studies DNA variation in known genes was not measured or directly assessed. Rather, genetic effects were inferred by the pattern of aggregation or segregation of the trait among family members.

A measured genotype approach

In 1986, Boerwinkle *et al.* [35] outlined a measured genotype approach whereby physiologically important variability in a candidate gene or gene product is directly employed in the genetic analysis of CHD and its risk factors. For plasma Lp(a) concentrations, the primary goal of the measured genotype approach has been to establish the role of the apo(a) gene in determining interindividual variation of Lp(a) levels, the extent of familial aggregation of Lp(a) levels and the segregation of Lp(a) levels in pedigrees.

Using sodium dodecylsulfate polyacrylamide gel electrophoresis followed by immunoblotting with a poly- or monoclonal human anti-apo(a) antibody, several apo(a) isoforms are distinguishable in human plasma [36,37]. The isoforms range in apparent molecular weight between 400 and 700 kD, and plasma from a single individual shows only one or two major isoforms. Utermann *et al.* [37] described six isoforms designated (from smallest to largest) F, B, S1, S2, S3, and S4. Because of the limited sensitivity of the western blot assay, individuals with very low plasma Lp(a) levels exhibit no apo(a) isoforms. Therefore, an operational null allele was also postulated. In a sample of 473 individuals from the Tyrol region of Austria, no visible bands were observed in approximately 25% of the individuals [38]. Considering only the observed single band types, the B, S1, S2, S3, and S4 isoforms occurred at a frequency of 2%, 6%, 26%, 25%, and 41%, respectively. In addition, the distribution of plasma Lp(a) concentrations was significantly different among Lp(a) isoform phenotypes [37,38]. The molecular weights of the apo(a) isoforms were, in general, inversely correlated with plasma levels of circulating Lp(a). Again considering only the observed single band types, the B isoform individuals had the highest average Lp(a) level (59 mg/dl), followed by the S1 individuals (28 mg/dl), the S2 individuals (24 mg/dl), the S3 individuals (12 mg/dl), and the S4 individuals (7.5 mg/dl). Boerwinkle *et al.* [38] determined that approximately 42% of the variation in plasma Lp(a) levels was attributable to the genetically determined Lp(a) protein isoforms. These studies clearly established that the apo(a) gene itself was an important determinant of plasma Lp(a) concentrations. Variation in the apo(a) gene and protein is responsible for the unmeasured major gene effect described by Morton *et al.* [31] and Hasstedt and Williams [32]. In addition there is some correspondence of frequency and effects between the measured apo(a) isoform described by Utermann *et al.* [37] and the major gene alleles inferred by Hasstedt and Williams [32]. Similar frequencies and effects can be imagined between the null apo(a) isoform and the low allele detected by segregation analysis, the S3 and S4 isoforms and the medium allele, and the S1 and S2 isoforms and the high allele.

Sandholzer *et al.* [39•] recently characterized the frequency and effects of the apo(a) protein isoforms in seven ethnic groups (Tyrolean, Icelandics, Hungarian, Maly, Chinese, Indian, and Black Sudanese). The distribution of plasma Lp(a) concentrations was significantly different among these groups, with the Chinese having the lowest (7.0 mg/dl) and the Sudanese the highest (46 mg/dl) Lp(a) levels. Even though the frequency of the apo(a) size isoforms were significantly different among the seven populations, their effects on plasma Lp(a) concentrations were not different. The authors conclude that the observed differences in the distribution of plasma Lp(a) concentrations could not be accounted for by differences in the frequencies or effects of the apo(a) protein isoforms.

Seed *et al.* [40] studied the association of Lp(a) concentrations and apo(a) isoforms with CHD in patients with heterozygous familial hypercholesterolemia (FH). Heterozygotes for defects in the LDL receptor gene leading to familial hypercholesterolemia have elevated LDL cholesterol levels and are at greatly increased risk of premature CHD. However, not all these individuals have diagnosed CHD; there is considerable variability in both the age of onset and the severity of disease [41]. In a sample of 115 FH heterozygotes, plasma Lp(a) concentration was the best predictor of angiographically documented CHD. The median Lp(a) level in patients with CHD was 57 mg/dl whereas it was only 18 mg/dl in those without disease. In addition, there was an increased frequency of the apo(a) isoforms associated with elevated Lp(a) levels in FH patients with CHD (Table 2). For example, the S2 isoform, which is typically associated with elevated Lp(a) levels (see below), was found more frequently in the patients with CHD, and the S4 isoform, which is typically associated with reduced Lp(a) levels, was found more frequently in the patients without CHD. Seed *et al.* [40] conclude that Lp(a) is a genetically determined risk factor for atherosclerosis in individuals with elevated serum LDL cholesterol levels, and that the Lp(a) system is at least partially responsible for the variability in expression of LDL receptor mutations. The results from Seed *et al.* [40] have been supported by a study of 120 FH heterozygotes by Wiklund *et al.* [42]. In addition Wiklund *et al.* show that cholesterol-lowering therapy in

the FH heterozygotes had no effect in reducing plasma Lp(a) concentrations. However, these results have not been supported by Mbewu *et al.* [43] who show that in a sample of 60 FH heterozygotes, those with CHD did not have significantly higher plasma Lp(a) concentrations than those without CHD.

Table 2. Apolipoprotein (a) isoform frequencies in FH patients with and without CHD.

	Frequency	
Isoform	CHD (n = 54)	No CHD (n = 55)
B	0.009	0.009
S1	0.048	0.000
S2	0.325	0.117
S3	0.169	0.194
S4	0.154	0.270
'null'	0.294	0.410

Data from Seed *et al.* [40].

Considerable insight into the role of the apo(a) gene in determining plasma Lp(a) concentrations has been obtained using the apo(a) protein isoforms. However, these analyses were compromised by the fact that the immunoblotting technique used to type the Lp(a) isoforms was not sensitive enough to detect those associated with low plasma Lp(a) concentrations. In general, and considering the limits of the sensitivity of the western blot assay, the apo(a) isoforms segregated in families as a Mendelian trait [37]. However, strict Mendelian segregation of the Lp(a) protein isoforms has been questioned [44]. In addition and undoubtedly because of the threshold of sensitivity of the western blot assay, the observed phenotype frequencies were significantly different from those expected under Hardy-Weinberg equilibrium. Clearly, a more sensitive and specific method of typing the apo(a) polymorphism was needed.

McLean *et al.* [8] proposed that the size differences observed in the apo(a) protein were due to variable numbers of kringle 4 repeats in the apo(a) gene. Various studies have provided experimental data which confirm this initial proposal [9•,45,46,47••]. Lackner *et al.* [47••] identified a single *Kpn*I restriction fragment containing most, if not all, the kringle-4-encoding sequences of the apo(a) gene. Using restriction enzymes that cut at a single place in the kringle-4-encoding region (for example *Pvu*II) it was inferred that a single kringle 4-encoding domain is approximately 5.5 kb in length. In contrast, *Kpn*I digested genomic DNA yielded bands ranging from 40 to 200 kb in length. Using carefully controlled pulsed-field gel electrophoresis conditions, 19 different-sized apo(a) fragments were observed in a sample of 102 Caucasian Americans. The 19 alleles formed a ladder with adjacent bands usually differing by 5 to 6 kb in length. No individuals showed more than two bands and the bands were inherited in a manner consistent with autosomal codominant segregation. The frequency distribution of apo(a) alleles observed in this sample is shown in Table 3. Alleles 12–15 were the most common, making up 50% of the total number of alleles. In general, smaller apo(a) alleles were not as frequent as larger alleles, giving the distribution a negatively skewed appearance. The observed apo(a) genotype frequencies agreed with Hardy-Weinberg expectations. This latter result is important in light of the lack of fit of the Lp(a) isoform frequencies to Hardy-Weinberg [38], and it indicates that the pulsed field typing technique is sensitive enough to detect all apo(a) length alleles. Size variation in the apo(a) *Kpn*I fragment indicates that the number of kringle-4-encoding domains in the apo(a) gene ranges from approximately nine to 35 copies. Therefore, apo(a) is a highly polymorphic gene with at least 19 alleles differing because of a variable number of tandemly repeated kringle-4-encoding intron-exon domains. A larger population-based survey of apo(a) gene length is likely to reveal that more than 19 alleles exist, as the sample used by Lackner *et al.* [47••] was relatively small and some adjacent bands differed by more than 5.5 kb in length. Indeed, using high resolution sodium dodecylsulfate agarose electrophoresis Kamboh *et al.* [48] observed 23 different protein isoforms in a sample of 270 Caucasian individuals.

Table 3. Apo(a) allele frequencies as determined by pulsed field gel electrophoresis.

Allele size*	Frequency
1	0.005
2	0.015
3	0.010
4	0.029
5	0.044
6	0.039
7	0.029
8	0.059
9	0.044
10	0.069
11	0.049
12	0.137
13	0.118
14	0.152
15	0.093
16	0.044
17	0.025
18	0.034
19	0.005

Data from Lackner *et al.* [45••]. *Sizes are given in relative units from smallest to largest.

Lackner *et al.* [47••] also report an inverse relationship between the size of the apo(a) allele and plasma Lp(a) levels. Figure 2 shows a three-dimensional graph of the size of the apo(a) alleles as measured by pulsed-field gel electrophoresis and the level of plasma Lp(a). Individuals with at least one small apo(a) allele tend to have relatively high plasma Lp(a) levels; whereas individuals with large apo(a) alleles have reduced Lp(a) levels. The rank correlation between the length of the apo(a) alleles and plasma Lp(a) levels was −0.4. As is evident in Figure 2, there are several notable exceptions to this general trend. For example, there is one 14/17 heterozygote with an Lp(a) concentration of approximately 40 mg/dl. This concentration is far greater than one would have predicted based on the apo(a) genotype alone. The factors responsible for these discrepancies may be environmental in nature or the result of other genetic determinants. In addition, if there are other genetic loci influencing Lp(a) levels they may be within the apo(a) gene alongside the length variation or coded for by other genes.

In order to quantitate the contribution of the entire apo(a) gene to plasma Lp(a) concentrations we have analyzed the segregation of the apo(a) gene and Lp(a) levels in several families [47••]. Examination of the family data indicate that in addition to the number of kringle 4 repeats there are other sites in the apo(a) gene which influence plasma Lp(a) concentrations. Table 4 gives the apo(a) types and Lp(a) levels in siblings with identical and discordant apo(a) types. Sibling pairs with identical apo(a) types were highly concordant, whereas sibling pairs in the same families who were discordant for their apo(a) types often had very different Lp(a) levels. In a sample of 73 sibling pairs that share both apo(a) alleles identical by descent, the correlation of plasma Lp(a) concentrations was very high (0.95; Boerwinkle E and Hobbs HH, unpublished data). In contrast, in a sample of 52 sibling pairs with no apo(a) alleles identical by descent, the correlation coefficient of plasma Lp(a) concentrations was low and negative (−0.23). The high degree of similarity of Lp(a) levels among sibling pairs with identical apo(a) genotypes suggests that factors at the apo(a) gene other than the number of kringle 4 repeats are contributing to plasma Lp(a) levels. In other words, apo(a) alleles of the same length (as determined by pulsed-field gel electrophoresis) do not necessarily have the same DNA sequence. There may be sequence differences in the apo(a) promoter that influence the transcription of the gene or in other sequences that alter the transport and metabolism of the Lp(a) particle. The identity of these sequences is currently unknown.

A synthesis

Genetic architecture is defined as the number and type (e.g. structural or regulatory) of polymorphic genes af-

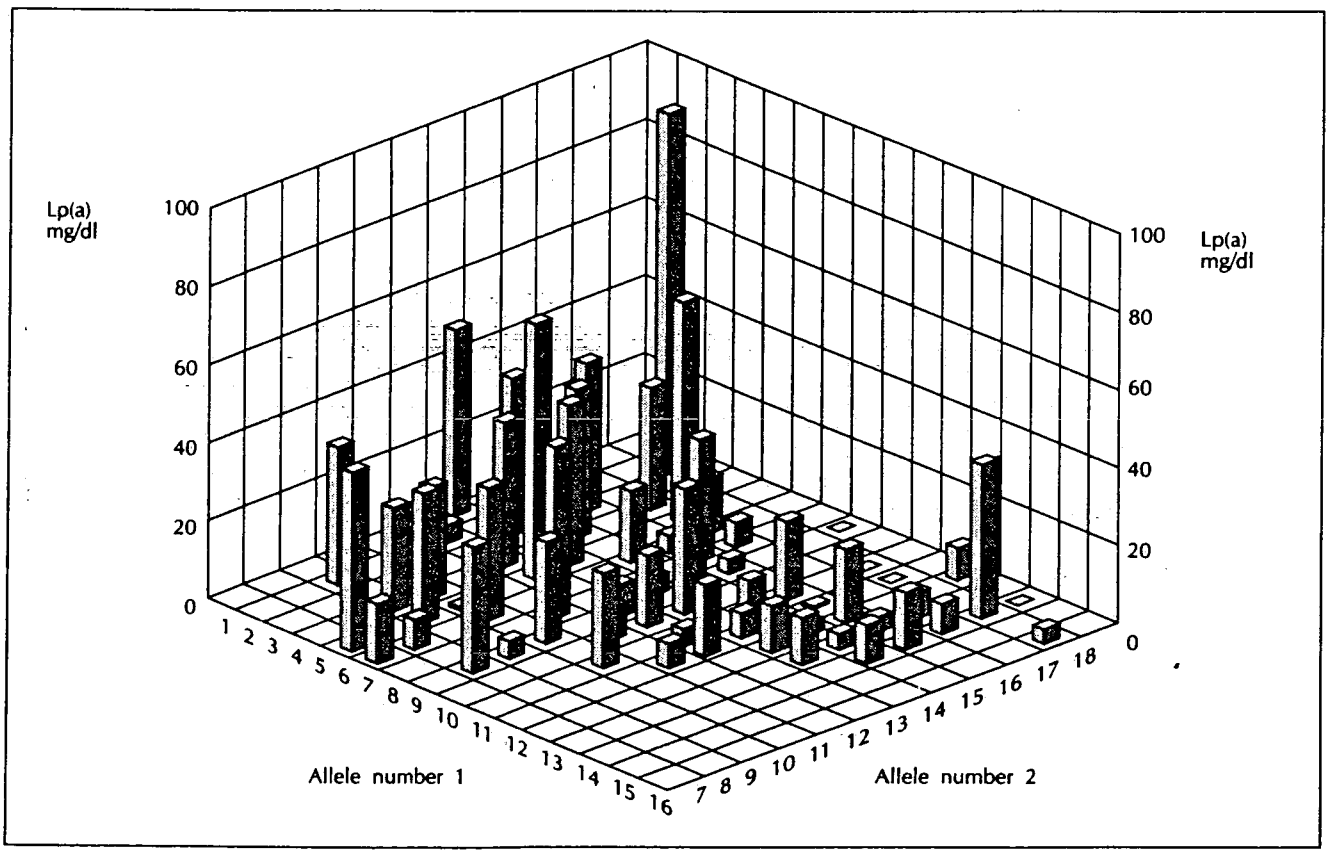


Fig. 2. Association between apo(a) length genotypes and plasma Lp(a) levels. Each cartesian coordinate represents an apo(a) genotype and Lp(a) levels are plotted in the vertical axis.

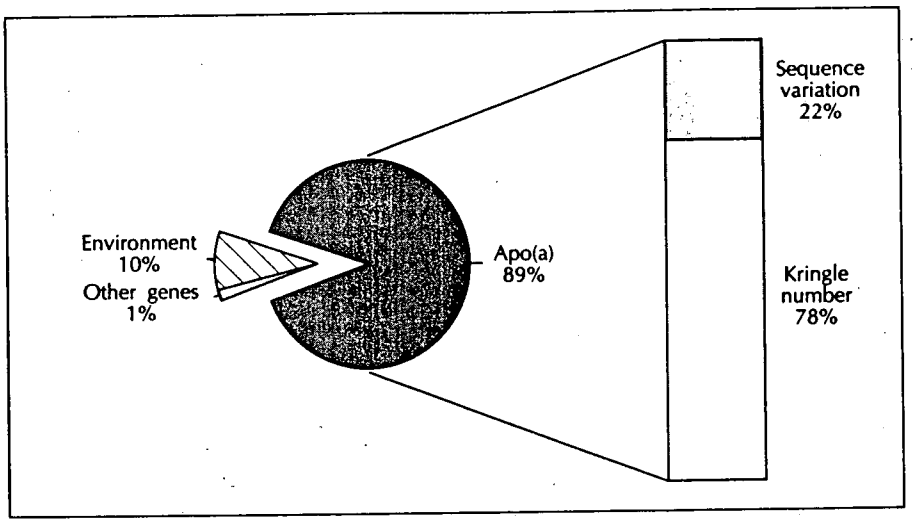


Fig. 3. Summary of the genetic architecture of plasma Lp(a) concentrations. The entire pie represents the interindividual variance of plasma Lp(a) concentrations. Approximately 90% of this variance is attributable to genetic factors, and virtually all of this is due to effects of the apo(a) gene. Other genes, such as the LDL receptor gene [52], have a slight effect on Lp(a) concentrations. The remaining 10% is due to environmental factors. The effects of the apo(a) gene may be partitioned into two components: length and sequence variation.

Table 4. Comparison of Lp(a) levels among sibling pairs.

	Concordant siblings		Other siblings	
Family	Apo(a) type	Lp(a) (mg/dl)	Apo(a) type	Lp(a) (mg/dl)
1	4/10	41;42	10/14;6/14	< 1;16
2	9/16	7;9	5/9;4/5	50;75
3	12/14	6;9	8/12	< 1
4	12/15	1;1	14/15,7/12	< 1;28
5	14/17	1;1	14/15,15/17	2;3
6	6/16	5;6	15/15	6
7	2/18	< 1; < 1	15/18	< 1
8	11/15	< 1; < 1	15/17,8/11	< 1;7
9	14/15	< 1; < 1	15/18	< 1

Data from Lackner *et al.* [45••].

fecting a trait, the number of alleles at each locus, the frequencies of the alleles, and the size of their effects [49,50]. Our understanding of the genetics of plasma Lp(a) concentrations has evolved in parallel with the ability to measure accurately both the quantity of Lp(a) in plasma and characteristics of the apo(a) gene. The genetic architecture of plasma Lp(a) concentrations has been defined at three levels: polygenic heritability, role of *in toto* apo(a) gene variation, and effects of length variation in the apo(a) gene (Fig. 3). As already mentioned, a number of investigators have concluded that the Lp(a) phenotype and Lp(a) levels are highly genetically determined. A consensus value for the polygenic heritability of plasma Lp(a) concentrations is approximately 90%; most studies report estimates close to this value. Hasstedt and Williams [32] conclude on the basis of segregation analysis of a large Utah pedigree that the majority of this heritability is attributable to the effects of a single major gene, and this gene is linked to the structural gene for plasminogen [33, 34]. This result takes on added importance with the knowledge that the structural gene for apo(a), the unique protein component of Lp(a), is closely linked to plasminogen [51]. Utermann *et al.* [37] described several electrophoretically separable isoforms in Lp(a), and Boerwinkle *et al.* [38] estimated that these isoforms accounted for 42% of the interindividual variation in plasma Lp(a) concentrations. Utermann *et al.* [52] demonstrated that mutations in the LDL receptor gene also influence plasma Lp(a) levels. However, these are likely to have only a small effect in the general population because their frequency is rare. The identification of a *KpnI* restriction fragment containing most if not all the kringle-4-encoding domain of the apo(a) gene [47••] and the demonstration that this fragment was highly polymorphic in length provided a powerful tool in the genetic analysis of plasma Lp(a) concentrations. By examining the degree of resemblance of sibling pairs sharing both alleles identical by descent as compared with siblings with no apo(a) alleles identical by descent, we can infer that virtually all the variability of plasma Lp(a) levels is due to variation in the apo(a) gene. Most, but not all, of this can be attributed to differing number of kringle 4 repeats. However, cis-acting sequence variation in this gene also affects plasma Lp(a) concentrations. These data represent the most complete picture to date of the genetic architecture of a major CHD risk factor.

Acknowledgements

This work was carried out in part through the assistance of the grants NIH HL-40613, NIJ 90-IJ-CX-0038, and Research Career Development Award. Eric Boerwinkle is an Established Investigator of the American Heart Association. Eric Boerwinkle would also like to thank Drs Gerd Utermann and Helen Hobbs for their collaboration in this work.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. NATIONAL CENTER FOR HEALTH STATISTICS: **Vital Statistics of the United States, 1988, Vol II. Mortality.** Public Health Service. Washington DC, 1990.

2. BERG K: **A New Serum Type System in Man: The Lp System.** *Acta Pathol Microbiol Scand* 1963, 59:369–382.

3. MILES LA, FLESS GM, LEVIN EG, SCANU AM, PLOW EF: **A Potential Basis for the Thrombotic Risks Associated with Lipoprotein (a).** *Nature* 1989, 339:301–305.

4. ALBERS JJ, ADOLPHSON JL, HAZARD WR: **Radio-Immunoassay of Human Plasma Lp(a) Lipoprotein.** *J Lipid Res* 1977, 18:331–338.

5. FLESS GM, ZUM MALLEN ME, SCANU AM: **Physiological Properties of Apolipoprotein(a) and Lipoprotein(a-) Derived from the Dissociation of Human Plasma Lipoprotein (a).** *J Biol Chem* 1986, 261:8712–8718.

6. UTERMANN G, WEBER W: **Protein Composition of Lp(a) Lipoprotein for Human Plasma.** *FEBS Lett* 1983, 154:357–361.

7. GAUBATZ JW, HEIDEMAN C, GOTTO AM, MORRISETT JD, DAHLEN GH: **Human Plasma Lipoprotein (a): Structural Properties.** *J Biol Chem* 1983, 258:4582–4589.

8. MCLEAN JW, TOMLINSON JE, KUANG WJ, EATON DL, CHEN EY, FLESS GM, SCANU AM, LAWN RM: **cDNA Sequence of Human Apolipoprotein(a) is Homologous to Plasminogen.** *Nature* 1987, 330:132–137.

9. • KOSCHINSKY M, BEISIEGEL U, HENNE-BRUNS D, EATON DL, LAWN RM: **Apolipoprotein(a) Size Heterogeneity is Related to Variable Number of Repeat Sequences in its mRNA.** *Biochemistry* 1990, 29:640–644.

This paper reports data providing direct and definitive evidence that the size polymorphism in apo(a) was attributable to length variation in the apo(a) transcript.

10. KOSTNER GM: **Apolipoproteins and Lipoproteins of Human Plasma: Significance in Health and in Disease.** *Adv Lipid Res* 1983, 20:1–43.

11. RAINWATER DL, LANFORD RE: **Production of Lipoprotein(a) by Primary Baboon hepatocytes.** *Biochim Biophys Acta* 1989, 1003:30–35.

12. TOMLINSON JE, MCLEAN JW, LAWN R: **Rhesus Monkey Apolipoprotein(a).** *J Biol Chem* 1989, 264:5957–5965.

13. EATON DL, FLESS GM, KOHR WJ, MCLEAN JW, XU QT, MILLER CG, LAWN RM, SCANU AM: **Partial Amino Aacid Sequence of Apolipoprotein (a) Shows that it is Homologous to Plasminogen.** *Proc Natl Acad Sci U S A* 1987, 84:3224–3228.

14. MORRISETT JD, GUYTON JR, GAUBATZ JW, GOTTO AM: **Lipoprotein(a): Structure, Metabolism, and Epidemiology.** In *Plasma Lipoproteins.* Edited by Gotto AM Jr. Amsterdam: Elsevier; 1987:129–152.

15. • ROSENGREN A, WILHELMSEN L, ERIKSSON E, RISBERG B, WEDEL H: **Lipoprotein(a) and Coronary Heart Disease: A Prospective Case–Control Study in a General Population Sample of Middle Aged Men.** *BMJ* 1990, 301:1248–1251.

To my knowledge, this paper reports the results of the only longitudinal prospective study demonstrating that plasma Lp(a) concentrations are a risk factor for CHD.

16. MAEDA S, ABE A, SEISHIMA K, MAKINO K, NOMA A, KAWADE M: **Transient Changes of Serum Lipoprotein(a) as an Acute Phase Protein.** *Atherosclerosis* 1989, 78:145–150.

17. ARMSTRONG VW, CREMER P, EBERLE E, MANKE A, SCHULZE F, WIELAND H, KREUZER H, SEIDEL D: **The Association Between Serum Lp(a) and Angiographically Assessed Coronary Atherosclerosis.** *Atherosclerosis* 1986, 62:249–257.

18. ZENKER G, KOLTRINGER P, BONE G, NIEDERKORN K, PFEIFFER K, JURGENS. G: **Lipoprotein(a) as a Strong Indicator for Cerebrovascular Disease.** *Stroke* 1986, 17:942–945.

19. PARRA HJ, MEZDOUR H, CACHERA C, DRACON M, TACQUET A, FRUCHART JC: **Lp(a) Lipoprotein in Patients with Chronic Renal Failure Treated by Hemodialysis.** *Clin Chem* 1987, 33:721–721.

20. KROLEWSKI AS, WARRAM JH, RAND LI, KHAN CR: **Epidemiological Approach to the Etiology of Type I Diabetes Mellitus and its Complications.** *N Engl J Med* 1987, 317:1390–1398.

21. HAFFNER SM, TUTTLE KR, RAINWATER DL: **Decrease of Lipoprotein (a) with Improved Glycemic Control in IDDM Subjects.** *Diabetes Care* 1991, 14:302–307.

22. BRUCKERT E, DAVIDOFF P, GRIMALDI A, TRUFFERT J, GIRAL P, DOUMITH R, THERVET F, DE GENNES JL: **Increased Serum Levels of Lipoprotein(a) in Diabetes Mellitus and their Reduction with Glycemic Control.** *JAMA* 1990, 263:35–36.

23. JOVEN J, VILELLA E: **Serum Levels of Lipoprotein(a) In Patients with Well Controlled Non-Insulin-Dependent Diabetes Mellitus.** *JAMA* 1991, 265:1113–1114.

24. LEVITSKY LL, SCANU AM, GOULD SH: **Lipoprotein (a) Levels in Black and White Children and Adolescents with IDDM.** *Diabetes Care* 1991, 14:283–287.

25. LALOUEL JM, RAO DC, MORTON NE, ELSTON RC: **A Unified Model for Complex Segregation Analysis.** *Am J Hum Genet* 1983, 35:816–826.

26. BERG K, MOHR J: **Genetics of the Lp System.** *Acta Genet* 1963, 13:349–360.

27. ALBERS JJ, WAHL P, HAZZARD WR: **Quantitative Genetic Studies of the Human Plasma Lp(a) Lipoprotein.** *Biochem Genet* 1974, ii:475–486.

28. SING CF, SCHULTZ JS, SHREFFLER DC: **The Genetics of the Lp Antigen. II. A Family Study and Proposed Models of Genetic Control.** *Ann Hum Genet* 1974, 38:47–56.

29. MORTON NE, GULBRANDSEN CL, RHOADS GG, KAGAN A: **The Lp Lipoprotein in Japanese.** *Clin Genet* 1978, 14:207–212.

30. ISELIUS L, DAHLEN G, DEFAIRE U, LUNDMAN T: **Complex Segregation Analysis of the Lp(a)/pre-$_1$-lipoprotein Trait.** *Clin Genet* 1981, 20:147–151.

31. MORTON NE, BERG K, DAHLEN G, FERRELL RE, RHOADS GG: **Genetics of the Lp Lipoprotein in Japanese-Americans.** *Genet Epidemiol* 1985 2:113–121.

32. HASSTEDT SJ, WILLIAMS RR: **Three Alleles For Quantitative Lp(a).** *Genet Epidemiol* 1986, 3:53–55.

33. DRAYNA DT, HEGELE RA, HASS PE, EMI M, WU LL, EATON DL, LAWN RM, WILLIAMS RR, WHITE RL, LALOUEL J-M: **Genetic Linkage Between Lipoprotein(a) Phenotype and a DNA Polymorphism in the Plasminogen Gene.** *Genomics* 1988, 3:230–236.

34. WEITKAMP LR, GUTTORMSEN SA, SCHULTZ JS: **Linkage Between the Loci for the Lp(a) Lipoprotein (LP) and Plasminogen (PLG).** *Hum Genet* 1988, 79:80–82.

35. BOERWINKLE E, CHAKRABORTY R, SING CF: **The Use of Measured Genotype Information in the Analysis of Quantitative Phenotypes in Man. I. Models and Analytical Methods.** *Ann Hum Genet* 1986, 50:181–194.

36. FLESS GM, ROLIH CA, SCANU AM: **Heterogeneity of Human Plasma Lipoprotein (a).** *J Biol Chem* 1984, 259:11470–11478.

37. UTERMANN G, MENZEL HJ, KRAFT HG, DUBA C, KEMMLER HG, SEITZ C: **Lp(a) Glycoprotein Phenotypes: Inheritance and Relation to Lp(a) Lipoprotein Concentrations in Plasma.** *J Clin Invest* 1987, 80:458–465.

38. BOERWINKLE E, MENZEL HG, KRAFT HG, UTERMANN G: Genetics of the Quantitative Lp(a) Lipoprotein Trait. III. Contribution of Lp(a) Glycoprotein Phenotypes to Normal Lipid Variation. *Hum Genet* 1989, 82:73–78.

39. • SANDHOLZER C, HALLMAN DM, SAHA N, SIGURDSSON G, LACKNER C, CSASZAR A, BOERWINKLE E, UTERMANN G: Effects of the Apolipoprotein(a) Size Polymorphism of the Lipoprotein(a) Concentration in 7 Ethnic Groups. *Hum Genet* 1991, 86:607–614.

This paper demonstrates that apo(a) is polymorphic in multiple ethnic and racial groups. However, the estimated allele frequencies are significantly different among the populations studied. In addition, the influence of the apo(a) length variation on plasma Lp(a) levels is present in all groups.

40. SEED M, HOPPICHLER F, REAVELEY D, MCCARTHY S, THOMPSON GR, BOERWINKLE E, UTERMANN G: Relation of Serum Lipoprotein(a) Concentration and Apolipoprotein(a) Phenotype to Coronary Heart Disease in Patients with Familial Hypercholesterolemia. *N Engl J Med* 1990, 322:1494–1499.

41. THOMPSON GR, SEED M, NITHYNANTHAN S, MCCARTHY S, THOROGOOD M: Genotypic and Phenotypic Variation in Familial Hypercholesterolemia. *Arteriosclerosis* 1989, 9:175–180.

42. WIKLUND O, ANGELIN B, OLOFSSON S-O, ERIKSSON M, FAGER G, BERLUND L, BONDJERS: Apolipoprotein(a) and Ischemic Heart Disease in Familial Hypercholesterolemia. *Lancet* 1990, 335:1360–1363.

43. MBEWU AD, BHATNAGAR D, DURRINGTON PN, HUNT L, ISHOLA M, ARROL S, MACKNESS M, LOCKLEY P, MILLER JP: Serum Lipoprotein(a) in Patients Heterozygous for Familial Hypercholesterolemia, Their Relatives, and Unrelated Control Populations. *Arteriosclerosis Thromb* 1991, 11:940–946.

44. GAUBATZ JW, GHANEM KI, GUEVARA J, NAVA ML, PATSCH W, MORRISETT JD: Polymorphic Forms of Human Apolipoprotein(a): Inheritance and Relationship of their Molecular Weights to Plasma Levels of Lipoprotein(a). *J Lipid Res* 1990, 31:603–613.

45. HIXSON JE, BRITTEN ML, MANIS GS, RAINWATER DL: Apolipoprotein(a) (Apo(a)) Glycoprotein Isoforms Result from Size Differences in Apo(a) mRNA in Baboons. *J Biol Chem* 1989, 264:6013–6016.

46. GAVISH D, AZROLAN N, BRESLOW JL: Plasma Lp(a) Concentration is Inversely Correlated with the Ratio of Kringle IV/Kringle V Encoding Domains in the Apo(a) Gene. *J Clin Invest* 1989, 84:2021–2027.

47. •• LACKNER C, BOERWINKLE E, LEFFERT CC, RAHMIG T, HOBBS HH: Molecular Basis of Apolipoprotein(a) Isoform Size Heterogeneity as Revealed by Pulsed-field Gel Electrophoresis. *J Clin Invest* 1991, 87:2077–2086.

The authors have identified a single *Kpn*I restriction fragment containing most, if not all, of the kringle-4-encoding sequences of the apo(a) gene. Using pulsed-field gel electrophoresis conditions, 19 different sized apo(a) DNA fragments ranging in size between 40 and 200 kb were observed in a sample of 102 Caucasian Americans. This size variation in the apo(a) *Kpn*I fragment indicates that the number of kringle-4-encoding domains in the apo(a) gene ranges from approximately nine to 35 copies. Apo(a) is a highly polymorphic gene with at least 19 alleles differing because of a variable number of tandemly repeated kringle-4-encoding intron-exon domains. The authors also report an inverse relationship between the size of the apo(a) allele and plasma Lp(a) levels.

48. KAMBOH MI, FERRELL RE, KOTTKE BA: Expressed Hypervariable Polymorphism of Apolipoprotein(a). *Am J Hum Genet* 1991, 49:1063–1074.

49. BOERWINKLE E, SING CF: The Use of Measured Genotype Information in the Analysis of Quantitative Phenotypes in Man. III. Simulataneous Estimation of the Frequencies and Effects of the Apolipoprotein E Polymorphism and Residual Polygenic Effects on Cholesterol, Betalipoprotein, and Triglyceride Levels. *Ann Hum Genet* 1987, 51:211–226.

50. SING CF, MOLL PP: Genetics of Atherosclerosis. *Annu Rev Genet* 1990, 24:171–187.

51. FRANK SL, KLISAK I, SPARKES RS, MOHANDAS T, TOMLINSON JE, MCLEAN JW, LAWN RM, LUSIS AJ: The Apolipoprotein(a) Gene Resides on Human Chromosome 6q26-27, in Close Proximity to the Homologous Gene for Plasminogen. *Hum Genet* 1988, 79:352–356.

52. UTERMANN G, HOPPICHLER F, DIEPLINGER H, SEED M, THOMPSON G, BOERWINKLE E: Defect in the Low Density Lipoprotein Receptor Gene Affect Lipoprotein (a) Levels: Multiplicative Interaction of Two Gene Loci Associated with Premature Atherosclerosis. *Proc Natl Acad Sci U S A* 1989, 86:4171–4174.

E. Boerwinkle, Center for Demographic and Population Genetics, The University of Texas Health Science Center in Houston, PO Box 2033, Houston, TX 77225, USA.

Reprinted from DIABETES, VOL. 40, NO. 11, NOVEMBER 1991

Signal Peptide–Length Variation in Human Apolipoprotein B Gene

Molecular Characteristics and Association with Plasma Glucose Levels

ERIC BOERWINKLE, SAN-HWAN CHEN, SOPHIA VISVIKIS, CRAIG L. HANIS, GERARD SIEST, AND LAWRENCE CHAN

We studied the molecular characteristics of three naturally occurring variants in the human apolipoprotein B (apoB) signal peptide, their frequencies in non-insulin-dependent diabetic and random populations, and their association with several measures of lipid and carbohydrate metabolism. In a random sample of 197 French whites, there were two common alleles, 5′βSP-24 and 5′βSP-27, with frequencies of 0.35 and 0.65, respectively. In a random sample of 181 Mexican Americans, there was an additional allele, 5′βSP-29, with a frequency of 0.03. DNA sequence analysis indicated that the signal peptide alleles consisted of the following: 5′βSP-29 encoded 29 amino acids in the signal peptide containing two copies of the sequence CTG GCG CTG encoding Leu-Ala-Leu and a consecutive run of eight Leu-encoding condons; 5′βSP-27 encoded 27 amino acids with a run of only six Leu condons; 5′βSP-24 encoded 24 amino acids and contained a single copy of CTG GCG CTG and a run of six Leu codons. In the sample of French whites, average apoAI and glucose levels were significantly different among signal peptide genotypes. 5′βSP-24/24 homozygotes had higher apoAI levels than the two other signal peptide genotypes (1.59 vs. 1.42 g/L, respectively). Heterozygous 5′βSP-24/27 individuals had the highest glucose levels. In the random sample of Mexican Americans, average glucose levels were also significantly different among signal peptide genotypes. However, the rank order of average glucose levels was not the same between the two samples. In the sample of Mexican Americans, glucose levels were significantly elevated (6.14 mM) in the 5′βSP-24/24 homozygotes relative to the other genotypes. Correspondingly, average glycosylated hemoglobin levels were increased and C-peptide levels were decreased in homozygous 5′βSP-24/24 individuals. There were no significant differences in the frequency of the three signal peptide alleles between the random sample of Mexican Americans and a sample of 203 non-insulin-dependent diabetic Mexican Americans. The cause of the association (e.g., chance, linkage-phase disequilibrium, causation) between signal peptide–length variation and plasma glucose levels is not known. *Diabetes* 40:1539–44, 1991

From the Center for Demographic and Population Genetics, The University of Texas Health Science Center at Houston; and Departments of Cell Biology and Medicine and the Diabetes and Endocrinology Research Center, Baylor College of Medicine, Houston, Texas; and the Center for Preventive Medicine, Vandoeuvre-les-Nancy, France.

Address correspondence and reprint requests to Dr. Eric Boerwinkle, Center for Demographic and Population Genetics, The University of Texas Health Science Center at Houston, Houston, TX 77025.

Received for publication 18 February 1991 and accepted in revised form 28 May 1991.

Non-insulin-dependent diabetes mellitus (NIDDM) is a heterogeneous disorder resulting from interactions of numerous genetic and environmental factors (1). Abnormalities in lipid metabolism including elevated plasma triglyceride levels are associated with NIDDM, and cardiovascular disease is the primary cause of death among diabetic individuals (2,3). Direct links between glucose and lipid metabolism are hypothesized to account for altered lipid profiles in diabetic individuals (2). Finding structural variation in genes whose products are involved in glucose and lipid metabolism, such as the apolipoprotein B (apoB) gene, figures prominently in strategies to elucidate the role of genes in determining cardiovascular disease risk among diabetic individuals. ApoB is the major apolipoprotein in postprandial chylomicrons, very-low-density lipoprotein (VLDL), intermediate-density lipoprotein, low-density lipoprotein (LDL), and lipoprotein(a) particles. Metabolism of these apoB-containing lipoprotein particles is altered in diabetic individuals (2,3).

Because of the potential association between apoB and cardiovascular complications of NIDDM, several laboratories have studied DNA polymorphisms in the apoB gene in diabetic populations (4). We have described polymorphic length variation in the signal peptide of apoB (5). Its location suggests that this sequence polymorphism may affect apoB synthesis and secretion. Eukaryotic signal peptides are im-

portant elements of most secretory proteins and are required for the translocation of the nascent polypeptide chain into the lumen of the rough endoplasmic reticulum (6–8). Sequence comparisons among known eukaryotic signal peptides revealed structural features that may be important both for the efficient transport of the peptide across the endoplasmic reticulum membrane and the accurate cleavage of the signal peptides from the mature protein by signal peptidase. Numerous signal peptide mutants have been produced in the laboratory to study structure-function relationships. However, naturally occurring mutants in higher organisms have not been described until now. To our knowledge, the apoB polymorphism that we described is the only such signal peptide–length polymorphism in humans (5). This study describes the molecular characteristics of three naturally occurring variants in the human apoB signal peptide, their frequency distribution in NIDDM and nondiabetic populations, and their association with altered lipid and glucose levels.

RESEARCH DESIGN AND METHODS

The frequencies of the apoB signal peptide alleles and their association with plasma lipid and carbohydrate levels were investigated in three groups of individuals. The first group consisted of 197 unrelated adults taking part in systematic health examinations at the Center for Preventive Medicine in Nancy, France. This sample was composed of 98 men and 99 women with an average age of 42 yr. The second group consisted of 181 unrelated Mexican-American adults from Starr County, Texas, a population with increased prevalence of NIDDM (9). This Mexican-American sample had 49 men and 132 women with an average age of 40 yr. In addition to these individuals, ascertained without regard for their disease status, we characterized the apoB signal peptide in 203 Mexican-American individuals with NIDDM and an average age of 53 yr.

Total plasma cholesterol in the French subjects was measured enzymatically on an SMA (Technicon, Tarrytown, NY) by the method of Allain et al. (10), and triglycerides and plasma glucose levels were measured enzymatically (11,12). In the samples from Mexican Americans, cholesterol and triglyceride levels were measured enzymatically with commercially available kits (Boehringer Mannheim, Indianapolis, IN). High-density lipoprotein cholesterol (HDL-chol) and its subfractions HDL_2-chol and HDL_3-chol were measured similarly after dextran sulfate–Mg^{2+} precipitation (13). VLDL-chol levels were estimated as one-fifth of triglyceride levels, and LDL-chol was estimated by the method of Friedwald et al. (14). Plasma insulin and C-peptide levels was measured with commercially available radioimmunoassays and monoclonal antibodies (Coat-a-Count, Diagnostic Products, Los Angeles, CA).

Blood was collected into EDTA vacutainers, buffy-coat separated, and frozen until genomic DNA was purified. Oligonucleotide primers for the polymerase chain reaction (PCR) were 22 and 23 nucleotides in length (15). The 5′ oligonucleotide was 5′-CAGCTGGCGATGGACCCGCCGA-3′ and the 3′ oligonucleotide was 5′-ACCGGCCC-TGGCGCCCGCCAGCA-3′. The PCR was performed in a final volume of 100 μl with ~0.5 μg genomic DNA. The reaction mixture used was recommended by the manufacturer with the addition of 10% DMSO. Before adding 1.5 U thermostable *Taq* polymerase (Perkin-ElmerCetus, Norwalk, CT), the reaction mixture was boiled for 6 min and then allowed to cool briefly. Denaturation was performed at 94°C for 1 min. Annealing and extension were done simultaneously for 1.5 min at 64°C. Amplified DNA was electrophoresed in 8% polyacrylamide gels at 90 V for 4–5 h. PCR products were directly visualized after ethidium bromide staining of the acrylamide gels.

To characterize the sequence variation responsible for the signal peptide polymorphism, PCR was performed with oligonucleotide primers that contained artificial 5′ *Eco*RI and 3′ *Bam*HI restriction sites. Gel purification products were subcloned into the plasmid pGEM-3Z. Multiple clones were sequenced directly with double-stranded DNA by the dideoxy chain–termination technique (16). To determine whether both signal peptide alleles were expressed, we extracted total RNA from cultured HepG2 cells, which synthesize apoB100 (17–19), by the guanidinium thiocyanate technique of Chirgwin et al. (20). PCR was performed on the RNA as described previously, except that the first-strand cDNA synthesis was performed at 42°C with reverse transcriptase. Subsequently, the standard PCR protocol was followed. PCR amplification products from the RNA were analyzed on a 3% agarose gel.

Routine statistical methods were used throughout. Allele frequencies were estimated by the gene-counting method. A contingency χ^2 test was used to test the homogeneity of genotype and allele frequencies between groups (21). Analysis of covariance was used to test the quality of phenotypic levels among apoB signal peptide genotypes (22). The covariates whose influence on variation of the lipid and glucose traits were considered include sex, age, height, weight, and body mass index. Variability of the lipid and glucose measures was expressed as the square root of the mean square error from the analyses of covariance. This measure is analogous to the common standard deviation and has the same units as the variable itself (e.g., mM). The distributions of plasma triglyceride and insulin levels were skewed positively. Therefore, the natural logarithm of these variables was also analyzed.

RESULTS

The electrophoresed and ethidium bromide–stained amplification products for each of the observed apoB signal peptide alleles are shown in Fig. 1. Individuals showed amplification products from one or two of three potential alleles. The alleles and their amplification products were named according to the number of amino acid residues in the apoB signal peptide, as determined by direct DNA sequencing (see below). In whites, there were two alleles that

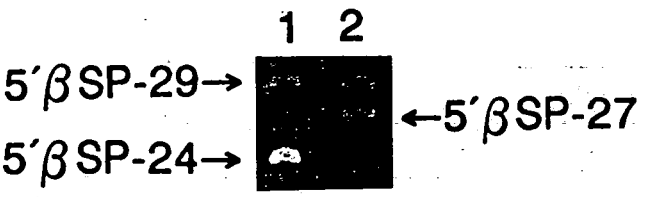


FIG. 1. Polymerase chain reaction products from each of the 3 apoB signal peptide alleles. *Lane 1*, heterozygous 5′βSP-24/29 individual. *Lane 2*, heterozygous 5′βSP-27/29 individual.

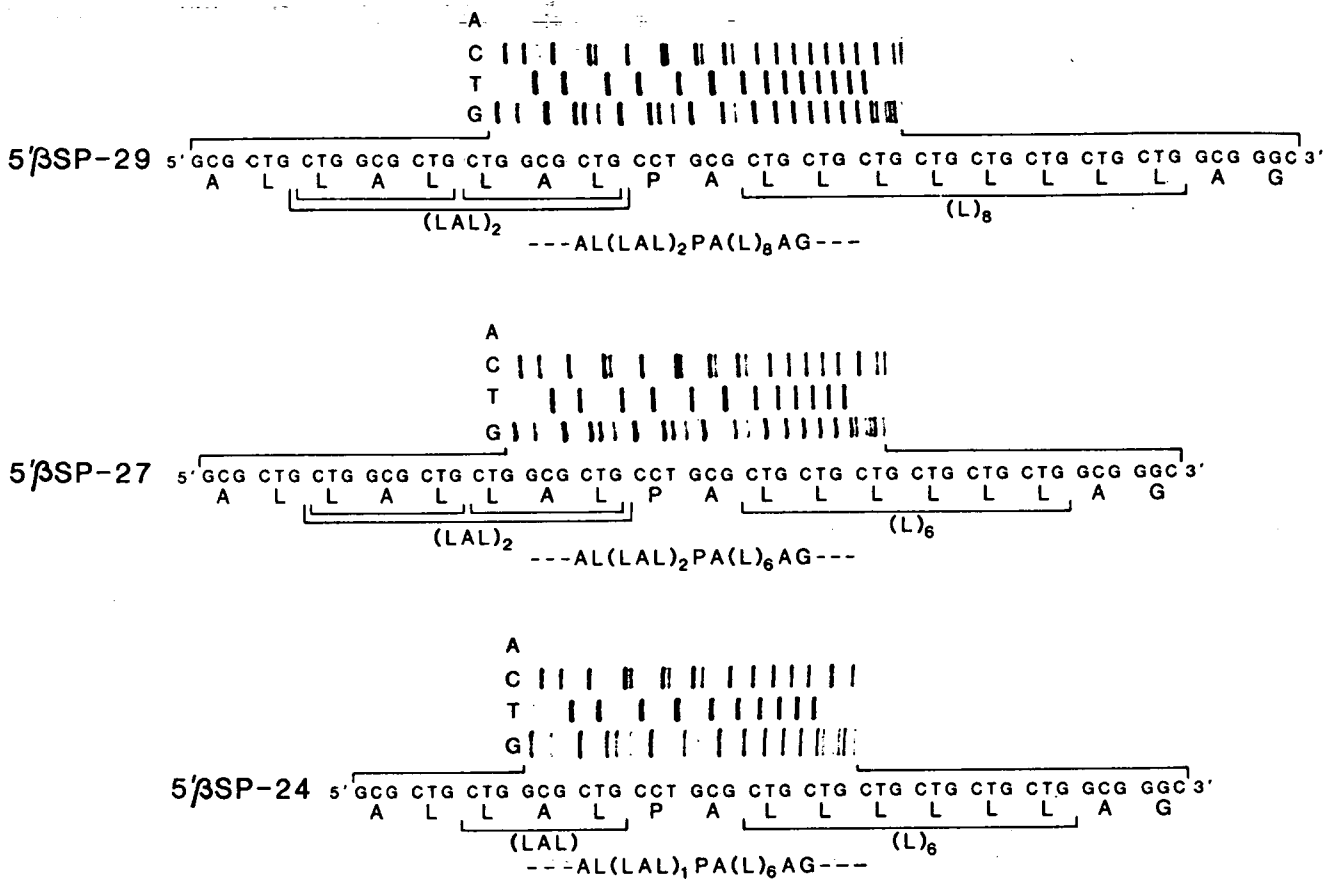


FIG. 2. Sequence analysis of 5′βSP-24, 5′βSP-27, and 5′βSP-29 polymerase chain reaction (PCR) products. PCR products from genomic DNA of individuals containing these alleles were produced and subcloned into plasmid pGEM3Z as described in METHODS. Multiple clones form each allele were sequenced by method of Sanger et al. (16). Representative autoradiograms of sequencing gels are shown. Note that in this region of sequence there is no adenine. A, alanine; G, glycine; L, leucine; P, proline.

differed by three amino acids (9 base pairs [bp]). The amplification product of 5′βSP-27, the larger allele of whites, was 93 bp and that of 5′βSP-24, the smaller allele, was 84 bp. Heterozygous individuals yielded PCR products of both sizes in approximately equal molar amounts. In Mexican Americans, a third allele was observed, 5′βSP-29, whose amplification product was 99 bp in length. The 5′βSP-29 allele was not observed in the sample of whites from Nancy, France.

Although the sizes of the PCR products were consistent with 24–, 27–, and 29–amino acid signal peptide alleles, we needed direct-sequence confirmation of the length and specific codons involved. DNA sequence analyses indicated that the signal peptide alleles consisted of the following (Fig. 2). The longest allele (5′βSP-29) encoded 29 amino acids in the signal peptide and contained two copies of the sequence CTG GCG CTG encoding Leu-Ala-Leu and a consecutive run of eight CTG codons encoding eight Leu residues. The medium-sized allele 5′βSP-27 encoded 27 amino acids and contained two copies of CTG GCG CTG but has a run of only six CTG codons. The shortest allele (5′βSP-24) encoded 24 amino acids and contained a single copy of CTG GCG CTG and a run of six CTG codons.

The PCR analysis were performed on genomic DNA. It is difficult from genomic DNA alone to determine whether both alleles are actually expressed in humans. The human hepatoma cell line HepG2 has been studied with respect to its apoB production and lipoprotein assembly (17–19). We analyzed the genomic DNA from this cell line and found it to be heterozygous for the 5′βSP-24 and 5′βSP-27 alleles. PCR analysis of RNA isolated from these cells indicated that both alleles are expressed at the mRNA level in roughly equal proportions (Fig. 3). Therefore, these alleles are probably also expressed in vivo. However, we have no information on the translational efficiency of the two mRNAs that are expressed or the early signal peptide processing within the cell. The absence of an association between the polymorphism and plasma apoB levels (see below) suggests that the biosynthetic rate of apoB100 is not affected by the difference in the length of the signal peptide. Because these subjects were fasting, no data exist concerning the role of apoB signal peptide–length variation in the synthesis and function of apoB48-containing postprandial lipoprotein particles.

The observed frequencies of each of the apoB signal peptide genotypes are given in Table 1. The apoB signal peptide was polymorphic, and the heterozygous 5′βSP-24/27 genotype was the most frequent type in each group. The 5′βSP-29 allele was found exclusively in the sample of Mexican Americans and was observed in both the random and NIDDM Mexican-American samples. No individuals homozygous for 5′βSP-29 allele were observed; this result is not

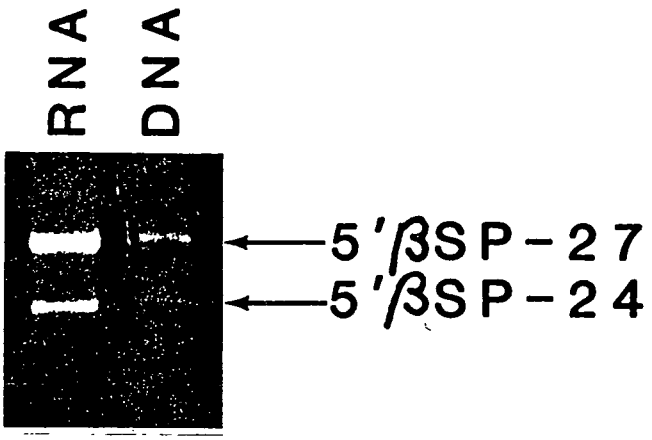


FIG. 3. Agarose gel analysis of apolipoprotein B polymerase chain reaction (PCR) products from HepG2 cells. PCR products amplified from signal peptide region of purified HepG2 RNA and DNA were fractionated on 3% agarose gel. *Double band* on *Right* indicates that HepG2 is heterozygous for 5′βSP-24 and 5′βSP-27; *double band* on left indicates that both alleles are expressed at RNA level.

surprising given the low frequency of this allele (Table 1). The 5′βSP-27 allele is the most frequent allele in each of the three groups. The relative frequencies of the 5′βSP-24 and 5′βSP-27 apoB signal peptide alleles in the French population were 0.355 and 0.645, respectively. The relative frequencies of the 5′βSP-24, -27, and -29 alleles in the Mexican-American population were 0.337, 0.630, and 0.033, respectively. The frequencies of the 5′βSP-24 and -27 alleles were not significantly different between the French and Mexican-American samples. The allele frequencies also were not significantly different between the sample of diabetic individuals and the random sample of Mexican Americans.

In the sample from Nancy, France, average cholesterol, LDL-chol, and apoB levels were not different among signal peptide genotypes (Table 2). Average apoAI and glucose levels and, to a lesser extent, triglyceride levels were different among signal peptide genotypes (Table 2). Individuals homozygous for the 5′βSP-24 allele had significantly higher plasma apoAI levels than those with only one or no 5′βSP-24 allele (1.59 vs. 1.42 g/L, respectively). Heterozygous individuals had average glucose levels that were significantly higher than the other two genotypes. Plasma triglyceride levels followed the same trend as apoAI levels; homozygous 5′βSP-24/24 individuals had higher levels than the other two genotypes.

To confirm and extend these suggestive findings, the sample of Mexican Americans was considered. Mexican Americans from Starr County, Texas, have an increased prevalence of NIDDM (9). Confirming the results obtained in the sample from Nancy, France, plasma cholesterol, LDL-chol, and apoB levels were not significantly different among signal peptide genotypes (Table 3). Due to the low frequency of the 5′βSP-29 allele, too few individuals were available for meaningful association studies. However, preliminary analyses indicated that this allele did not have a large effect on any of the laboratory measures (data not shown). In the random sample of Mexican Americans, average glucose levels were again significantly different among apoB signal peptide genotypes. However, the rank orders of average glucose levels were not the same in the sample of Mexican Americans and whites. In the Mexican Americans, average plasma glucose levels were elevated (6.14 mM) in the 5′βSP-24/24 homozygotes and lower in the other signal peptide genotypes. Consistent with this result for plasma glucose levels, and probably a consequence of it, blood levels of glycosylated hemoglobin, a monitor of integrated plasma glucose levels in the previous 3 mo, were also elevated in the homozygous 5′βSP-24/24 individuals. Plasma concentrations of C-peptide levels, which have a longer half-life than insulin concentrations and are an index of pancreatic insulin reserve, were different among the common genotypes. Homozygous 5′βSP-24/24 and heterozygous 5′βSP-24/27 individuals had elevated C-peptide levels, and homozygous 5′βSP-27/27 individuals had lower C-peptide levels. Insulin levels were not significantly different among genotypes.

TABLE 1
Observed frequencies of apolipoprotein B signal peptide variants

	Whites from Nancy, France (random)	Mexican Americans	
		Random	Non-insulin-dependent diabetes mellitus
Genotypes			
Total	1.0 (197)	1.0 (181)	1.0 (203)
5′ βSP-24/24	0.076 (15)	0.094 (17)	0.108 (22)
5′ βSP-24/27	0.558 (110)	0.464 (84)	0.517 (105)
5′ βSP-24/29	0	0.022 (4)	0.020 (4)
5′ βSP-27/27	0.365 (72)	0.376 (68)	0.315 (64)
5′ βSP-27/29	0	0.044 (8)	0.040 (8)
Alleles			
5′ βSP-24	0.355	0.337	0.377
5′ βSP-27	0.645	0.630	0.594
5′ βSP-29	0	0.033	0.029

n in parentheses.

DISCUSSION

ApoB is initially synthesized with an NH_2-terminal signal peptide sequence (23). This signal peptide is cleaved during transport of the polypeptide from the site of synthesis to the endoplasmic reticulum. Recently, it has been shown that apoB is translocated though the endoplasmic reticulum by a novel mechanism that maintains its solubility (24). Information specifying cellular or subcellular localization may also reside in the signal peptide sequence. Further studies on the role of apoB signal peptide variation on apoB synthesis and metabolism will shed light on the role of signal peptides in protein function and metabolism. As a cell that is heterozygous for the 5′βSP-24 and -27 alleles, the HepG2 cell should serve as a useful model to study the effect of the different-length alleles on apoB expression.

The nomenclature presented here for the signal peptide polymorphism (e.g., 5′βSP-24) indicates that the gene is apoB, specifies that the gene region is the signal peptide, and gives the length of the signal peptide in amino acids. We prefer this nomenclature over that suggested previously (5) because of its informativeness and flexibility and the fact that it is independent of the typing technique such as the position of the amplifying oligonucleotides. The previously

TABLE 2
Average lipid, apolipoprotein, glucose, and insulin levels in random sample of adults from Nancy, France

Variable	5′βSP-24/24 (n = 15)	5′βSP-24/27 (n = 110)	5′βSP-27/27 (n = 72)	Total (n = 197)	$\sqrt{MSE}$	P
Cholesterol (mM)	5.72	5.96	5.90	5.92	1.08	0.93
Low-density lipoprotein cholesterol (mM)	4.28	4.46	4.32	4.40	1.13	0.79
Triglycerides						
mM	1.63	1.27	1.11	1.24	0.93	0.08
ln mM	0.06	0.02	−0.03	0.01	0.21	0.12
High-density lipoprotein cholesterol (mM)	1.46	1.27	1.36	1.32	0.35	0.14
Apolipoprotein B (g/L)	1.23	1.20	1.14	1.18	0.29	0.29
Apolipoprotein A-I (g/L)	1.59	1.41	1.45	1.44	0.23	0.03
Glucose (mM)	5.23	5.65	5.38	5.52	0.76	0.05

$\sqrt{MSE}$, square root of mean square error from analysis of covariance; P, probability of equality of adjusted means.

suggested nomenclature referred to the 5′βSP-24 allele as a deletion and the 5′βSP-27 allele as an insertion (5). The nomenclature we use makes no assumption about the mechanism of origin of the signal peptide alleles. It also allows for the naming of additional signal peptide alleles should they be described later.

The apoB signal peptide polymorphism was associated with altered levels of several measures of lipid and carbohydrate metabolism (Tables 2 and 3). What are the possible links between genetic variation in apoB and plasma triglyceride and glucose levels? One possible site of action of this polymorphism is in the production and secretion of VLDL lipoproteins by the liver. In individuals with NIDDM, there is an overproduction of VLDL particles, possibly due to increased influx of free fatty acids and glucose or as a direct consequence of hyperinsulinemia (2). A caveat to this pathway of action is that we did not show an effect of the apoB signal peptide on plasma apoB levels. It is also possible that the apoB signal peptide polymorphism influences intestinal chylomicron production after a meal. These postprandial chylomicron particles contain apoB48 and not apoB100. Such effects on postprandial lipemia may be reflected by differences in fasting plasma glucose and triglyceride levels but not total apoB levels.

Plasma glucose was the only variable in which the association was significant in the random samples of whites and Mexican Americans. However, the direction of the association between the samples from Nancy, France, and Starr County, Texas, were not consistent. In the sample from Nancy, average glucose levels were elevated in the 5′βSP-24/27 heterozygotes relative to the two homozygous classes. In the sample from Starr County, average glucose levels followed a consistent trend; they were elevated in the 5′βSP-24/24 homozygote genotype and reduced in the 5′βSP-27/27 homozygote genotype. Xu et al. (25) reported that the apoB signal peptide genotypes were associated with plasma triglyceride levels in a sample of 106 individuals from North Karelia, Finland. In our study, the shorter 5′βSP-24 allele was only weakly associated with triglyceride levels in the sample from Nancy. Average triglyceride levels were not significantly different among genotypes in the random sample of Mexican Americans.

There are several possible reasons for these and other discrepancies. First, the significant results reported here and elsewhere may be due to chance, i.e., they represent statistical type I errors (26). Although possible, we believe that this is unlikely in light of the repeated significant associations with related measures (e.g., glucose, triglycerides, C-peptide) among studies and the repeated nonsignificant associations with other variables (total cholesterol, LDL-chol, apoB) in these same studies. Second, the association is real but different among populations. One cause of such a dif-

TABLE 3
Average lipid, apolipoprotein, glucose, and insulin levels in random sample of adults from Starr County, Texas

Variable	5′βSP-24/24 (n = 17)	5′βSP-24/27 (n = 84)	5′βSP-27/27 (n = 68)	Total (n = 181)	$\sqrt{MSE}$	P
Cholesterol (mM)	4.88	5.10	5.02	5.06	0.91	0.67
Low-density lipoprotein cholesterol (mM)	3.21	3.15	3.13	4.14	0.79	0.91
Triglycerides						
mM	1.28	1.69	1.44	1.59	0.99	0.17
ln mM	−2.46	−2.39	−2.47	−2.43	0.27	0.13
High-density lipoprotein cholesterol (mM)	1.15	1.15	1.25	1.20	0.27	0.08
Apolipoprotein B (g/L)	0.86	0.94	0.91	0.92	0.21	0.53
Apolipoprotein AI (g/L)	1.25	1.12	1.17	1.16	0.22	0.27
Glucose (mM)	6.14	5.63	5.34	5.59	1.02	0.01
Insulin						
pM	102.60	104.40	90.00	99.90	62.4	0.20
ln pM	2.89	2.91	2.87	2.88	0.27	0.24
C-peptide (nM)	1.00	1.09	0.91	1.01	0.53	0.05
HbA_{1c} (%)	6.60	5.90	6.10	6.10	0.92	0.05

$\sqrt{MSE}$, square root of mean square error from analysis of covariance; P, probability of equality of adjusted means.

ference could be that there are unidentified factors interacting with the apoB signal peptide, and these factors are different among populations. Finally, we favor the possibility that the signal peptide polymorphism did not directly cause the observed effect but rather was in linkage disequilibrium with a second locus with a direct effect on glucose and triglyceride metabolism. In addition, the magnitude and direction of this disequilibrium was different among populations. Regardless of the exact mechanisms, the apoB signal peptide should be included in any candidate-gene approach to the study of NIDDM and its complications.

ACKNOWLEDGMENTS

E.B. thanks Ms. Noosheen Behzadpour and other laboratory personnel for their valuable assistance. This work was supported by National Institutes of Health Grants HL-40613, (E.B.). HL-34823 (C.L.H), HL-27341, and DK-27685; NIH Grant 90-IJ-CX-0038; Baylor Diabetes and Endocrinology Research Center; and The March of Dimes Birth Defects Foundation (L.C.).

REFERENCES

1. Rossini AA, Mordes JP, Handler ES: Speculation on etiology of diabetes mellitus: tumbler hypothesis. *Diabetes* 37:257–61, 1988
2. Howard BV, Abbott WF, Beltz WF, Harper I, Fields RM, Grunoy SM, Taskinen RM: Integrated study of low density lipoprotein and very low density lipoprotein metabolism in non-insulin dependent diabetics. *Metabolism* 36:870–77, 1987
3. Betteridge DF: Lipoprotein metabolism. In *Recent Advances in Diabetes*. Nattrass M, Ed. New York, Churchill Livingstone, 1986, p. 91–107
4. Xiang K-S, Cox NJ, Huang P, Karam JH, Bell GI: Insulin-receptor and apolipoprotein genes contribute to development of NIDDM in Chinese Americans. *Diabetes* 38:17–23, 1989
5. Boerwinkle E, Chan L: A three codon insertion/deletion polymorphism in the signal peptide region of the human apolipoprotein B (apo B) gene directly typed by the polymerase chain reaction. *Nucleic Acids Res* 17:4003, 1989
6. Chan L, Bradley WA: Signal peptides: properties and interactions. In *Cellular Regulation of Secretion and Release*. Conn PM, Ed. New York, Academic 1986, p. 301–21
7. Benson SA, Hall MN, Silhavy TJ: Genetic analysis of protein export in *Escherichia coli* K12. *Annv Rev Biochem* 54:101–34, 1985
8. Randall LL, Hardy SJS: Unity in function in the absence of consensus in sequence: role of leader peptides in export. *Science* 243:1156–59, 1989
9. Hanis CL, Ferrell RE, Barton SA, Aguilar L, Garza-Ibarra A, Tulloch BR, Garcia CA, Schull WJ: Diabetes among Mexican Americans in Starr County, Texas. *Am J Epidemiol* 118:659–72, 1983
10. Allain CC, Poon LS, Chan CSG, Richmond W, Fu PC: Enzymatic determination of total serum cholesterol. *Clin Chem* 20:470–75, 1974
11. Steinmetz J, Panek E: Adaptation sur GSAII Greiner du dosage des triglycerides par voie entierement enzymatique: application a l'etude de leur conservation et de certaines interferences analytiques. *J Clin Chem Clin Biochem* 16:613–19, 1978
12. Trinder P: Determination of blood glucose using 6 aminophenazone as oxygen acceptor (Letter). *J Clin Pathol* 22:246, 1969
13. Warnick GR, Benerson J, Albers JJ: Dextran sulfate-Mg^{++} precipitation procedure for quantitation of high density lipoprotein cholesterol. *Clin Chem* 28:1379–88, 1972
14. Friedewald WT, Levy RI, Fredrickson DS: Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifugation. *Clin Chem* 18:499–502, 1972
15. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi RG, Horn TT, Mullis KB, Erlich HA: Primer-directed enzymatic amplification of DNA with thermostable DNA polymerase. *Science* 239:487–91, 1988
16. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain terminating inhibitors. *Proc Natl Acad Sci USA* 74:5403–67, 1977
17. Thrift RN, Forte, TM, Cahoon BE, Shore VG: Characterization of lipoproteins produced by the human liver cell line, Hep G2, under defined conditions. *J Lipid Res* 27:236–50, 1986
18. Bostrom K, Boren J, Wettesten M, Sjoberg A, Bondjers G, Wiklund O, Carlsson P, Olofsson S: Studies on the assembly of apo B-100–containing lipoproteins in Hep G2 cells. *J Biol Chem* 263:4434–42, 1988
19. Boren J, Wettesten M, Sjoberg A, Thorlin T, Bondjers G, Wiklund O, Olofsson S-O: The assembly and secretion of apoB-100–containing lipoproteins in HepG2 cells: evidence for different sites for protein synthesis and lipoprotein assembly. *J Biol Chem* 265:10556–64, 1990
20. Chirgwin JM, Przybyla AE, MacDonald RJ, Rutter WJ: Isolation of biochemically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* 18:5294–99, 1979
21. Sokal R, Rohlf FJ: *Biometry*. New York, Freeman, 1980
22. Neter J, Wasserman W: *Applied Linear Statistical Models*. Homewood, IL, Irwin, 1974
23. Yang C, Gu ZW, Weng S, Kim TW, Chen SH, Pownall HJ, Sharp PM, Liu SW, Li WH, Gotto AM, Chan L: Structure of apolipoprotein B-100 of human low density lipoproteins. *Arteriosclerosis* 9:96–108, 1089
24. Chuck SL, Yao Z, Blackhart BD, McCarthy BJ, Lingapp VR: New variation on the translocation of proteins during early biogenesis of apolipoprotein B. *Nature* (Lond) 346:382–85, 1990
25. Xu C-F, Tikkanen MJ, Huttunen JK, Pietinen P, Butler R, Humphries S, Talmud P: Apolipoprotein B signal peptide insertion/deletion polymorphism is associated with Ag epitopes and involved in the determination of serum triglyceride levels. *J Lipid Res* 31:1255–61, 1990
26. Larson RJ, Marx ML: An *Introduction to Mathematical Statistics and Its Applications*. Inglewood Cliffs, NJ, Prentice-Hall, 1981

Reprinted from DIABETES, VOL. 40, NO. 11, NOVEMBER 1991

Signal Peptide–Length Variation in Human Apolipoprotein B Gene

Molecular Characteristics and Association with Plasma Glucose Levels

ERIC BOERWINKLE, SAN-HWAN CHEN, SOPHIA VISVIKIS, CRAIG L. HANIS, GERARD SIEST, AND LAWRENCE CHAN

We studied the molecular characteristics of three naturally occurring variants in the human apolipoprotein B (apoB) signal peptide, their frequencies in non-insulin-dependent diabetic and random populations, and their association with several measures of lipid and carbohydrate metabolism. In a random sample of 197 French whites, there were two common alleles, 5′βSP-24 and 5′βSP-27, with frequencies of 0.35 and 0.65, respectively. In a random sample of 181 Mexican Americans, there was an additional allele, 5′βSP-29, with a frequency of 0.03. DNA sequence analysis indicated that the signal peptide alleles consisted of the following: 5′βSP-29 encoded 29 amino acids in the signal peptide containing two copies of the sequence CTG GCG CTG encoding Leu-Ala-Leu and a consecutive run of eight Leu-encoding condons; 5′βSP-27 encoded 27 amino acids with a run of only six Leu condons; 5′βSP-24 encoded 24 amino acids and contained a single copy of CTG GCG CTG and a run of six Leu codons. In the sample of French whites, average apoAI and glucose levels were significantly different among signal peptide genotypes. 5′βSP-24/24 homozygotes had higher apoAI levels than the two other signal peptide genotypes (1.59 vs. 1.42 g/L, respectively). Heterozygous 5′βSP-24/27 individuals had the highest glucose levels. In the random sample of Mexican Americans, average glucose levels were also significantly different among signal peptide genotypes. However, the rank order of average glucose levels was not the same between the two samples. In the sample of Mexican Americans, glucose levels were significantly elevated (6.14 mM) in the 5′βSP-24/24 homozygotes relative to the other genotypes. Correspondingly, average glycosylated hemoglobin levels were increased and C-peptide levels were decreased in homozygous 5′βSP-24/24 individuals. There were no significant differences in the frequency of the three signal peptide alleles between the random sample of Mexican Americans and a sample of 203 non-insulin-dependent diabetic Mexican Americans. The cause of the association (e.g., chance, linkage-phase disequilibrium, causation) between signal peptide–length variation and plasma glucose levels is not known. *Diabetes* 40:1539–44, 1991

Non-insulin-dependent diabetes mellitus (NIDDM) is a heterogeneous disorder resulting from interactions of numerous genetic and environmental factors (1). Abnormalities in lipid metabolism including elevated plasma triglyceride levels are associated with NIDDM, and cardiovascular disease is the primary cause of death among diabetic individuals (2,3). Direct links between glucose and lipid metabolism are hypothesized to account for altered lipid profiles in diabetic individuals (2). Finding structural variation in genes whose products are involved in glucose and lipid metabolism, such as the apolipoprotein B (apoB) gene, figures prominently in strategies to elucidate the role of genes in determining cardiovascular disease risk among diabetic individuals. ApoB is the major apolipoprotein in postprandial chylomicrons, very-low-density lipoprotein (VLDL), intermediate-density lipoprotein, low-density lipoprotein (LDL), and lipoprotein(a) particles. Metabolism of these apoB-containing lipoprotein particles is altered in diabetic individuals (2,3).

Because of the potential association between apoB and cardiovascular complications of NIDDM, several laboratories have studied DNA polymorphisms in the apoB gene in diabetic populations (4). We have described polymorphic length variation in the signal peptide of apoB (5). Its location suggests that this sequence polymorphism may affect apoB synthesis and secretion. Eukaryotic signal peptides are im-

From the Center for Demographic and Population Genetics, The University of Texas Health Science Center at Houston; and Departments of Cell Biology and Medicine and the Diabetes and Endocrinology Research Center, Baylor College of Medicine, Houston, Texas; and the Center for Preventive Medicine, Vandoeuvre-les-Nancy, France.

Address correspondence and reprint requests to Dr. Eric Boerwinkle, Center for Demographic and Population Genetics, The University of Texas Health Science Center at Houston, Houston, TX 77025.

Received for publication 18 February 1991 and accepted in revised form 28 May 1991.

portant elements of most secretory proteins and are required for the translocation of the nascent polypeptide chain into the lumen of the rough endoplasmic reticulum (6–8). Sequence comparisons among known eukaryotic signal peptides revealed structural features that may be important both for the efficient transport of the peptide across the endoplasmic reticulum membrane and the accurate cleavage of the signal peptides from the mature protein by signal peptidase. Numerous signal peptide mutants have been produced in the laboratory to study structure-function relationships. However, naturally occurring mutants in higher organisms have not been described until now. To our knowledge, the apoB polymorphism that we described is the only such signal peptide–length polymorphism in humans (5). This study describes the molecular characteristics of three naturally occurring variants in the human apoB signal peptide, their frequency distribution in NIDDM and nondiabetic populations, and their association with altered lipid and glucose levels.

RESEARCH DESIGN AND METHODS

The frequencies of the apoB signal peptide alleles and their association with plasma lipid and carbohydrate levels were investigated in three groups of individuals. The first group consisted of 197 unrelated adults taking part in systematic health examinations at the Center for Preventive Medicine in Nancy, France. This sample was composed of 98 men and 99 women with an average age of 42 yr. The second group consisted of 181 unrelated Mexican-American adults from Starr County, Texas, a population with increased prevalence of NIDDM (9). This Mexican-American sample had 49 men and 132 women with an average age of 40 yr. In addition to these individuals, ascertained without regard for their disease status, we characterized the apoB signal peptide in 203 Mexican-American individuals with NIDDM and an average age of 53 yr.

Total plasma cholesterol in the French subjects was measured enzymatically on an SMA (Technicon, Tarrytown, NY) by the method of Allain et al. (10), and triglycerides and plasma glucose levels were measured enzymatically (11,12). In the samples from Mexican Americans, cholesterol and triglyceride levels were measured enzymatically with commercially available kits (Boehringer Mannheim, Indianapolis, IN). High-density lipoprotein cholesterol (HDL-chol) and its subfractions HDL_2-chol and HDL_3-chol were measured similarly after dextran sulfate–Mg^{2+} precipitation (13). VLDL-chol levels were estimated as one-fifth of triglyceride levels, and LDL-chol was estimated by the method of Friedwald et al. (14). Plasma insulin and C-peptide levels was measured with commercially available radioimmunoassays and monoclonal antibodies (Coat-a-Count, Diagnostic Products, Los Angeles, CA).

Blood was collected into EDTA vacutainers, buffy-coat separated, and frozen until genomic DNA was purified. Oligonucleotide primers for the polymerase chain reaction (PCR) were 22 and 23 nucleotides in length (15). The 5′ oligonucleotide was 5′-CAGCTGGCGATGGACCCGCCGA-3′ and the 3′ oligonucleotide was 5′-ACCGGCCC-TGGCGCCCGCCAGCA-3′. The PCR was performed in a final volume of 100 μl with ~0.5 μg genomic DNA. The reaction mixture used was recommended by the manufacturer with the addition of 10% DMSO. Before adding 1.5 U thermostable *Taq* polymerase (Perkin-ElmerCetus, Norwalk, CT), the reaction mixture was boiled for 6 min and then allowed to cool briefly. Denaturation was performed at 94°C for 1 min. Annealing and extension were done simultaneously for 1.5 min at 64°C. Amplified DNA was electrophoresed in 8% polyacrylamide gels at 90 V for 4–5 h. PCR products were directly visualized after ethidium bromide staining of the acrylamide gels.

To characterize the sequence variation responsible for the signal peptide polymorphism, PCR was performed with oligonucleotide primers that contained artificial 5′ *Eco*RI and 3′ *Bam*HI restriction sites. Gel purification products were subcloned into the plasmid pGEM-3Z. Multiple clones were sequenced directly with double-stranded DNA by the dideoxy chain–termination technique (16). To determine whether both signal peptide alleles were expressed, we extracted total RNA from cultured HepG2 cells, which synthesize apoB100 (17–19), by the guanidinium thiocyanate technique of Chirgwin et al. (20). PCR was performed on the RNA as described previously, except that the first-strand cDNA synthesis was performed at 42°C with reverse transcriptase. Subsequently, the standard PCR protocol was followed. PCR amplification products from the RNA were analyzed on a 3% agarose gel.

Routine statistical methods were used throughout. Allele frequencies were estimated by the gene-counting method. A contingency χ^2 test was used to test the homogeneity of genotype and allele frequencies between groups (21). Analysis of covariance was used to test the quality of phenotypic levels among apoB signal peptide genotypes (22). The covariates whose influence on variation of the lipid and glucose traits were considered include sex, age, height, weight, and body mass index. Variability of the lipid and glucose measures was expressed as the square root of the mean square error from the analyses of covariance. This measure is analogous to the common standard deviation and has the same units as the variable itself (e.g., mM). The distributions of plasma triglyceride and insulin levels were skewed positively. Therefore, the natural logarithm of these variables was also analyzed.

RESULTS

The electrophoresed and ethidium bromide–stained amplification products for each of the observed apoB signal peptide alleles are shown in Fig. 1. Individuals showed amplification products from one or two of three potential alleles. The alleles and their amplification products were named according to the number of amino acid residues in the apoB signal peptide, as determined by direct DNA sequencing (see below). In whites, there were two alleles that

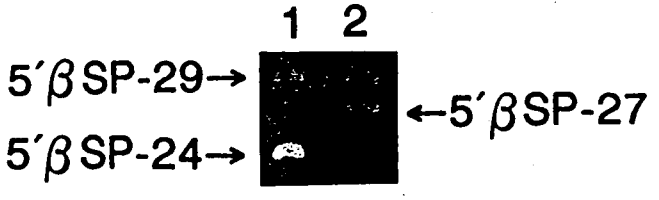


FIG. 1. Polymerase chain reaction products from each of the 3 apoB signal peptide alleles. *Lane 1*, heterozygous 5′βSP-24/29 individual. *Lane 2*, heterozygous 5′βSP-27/29 individual.

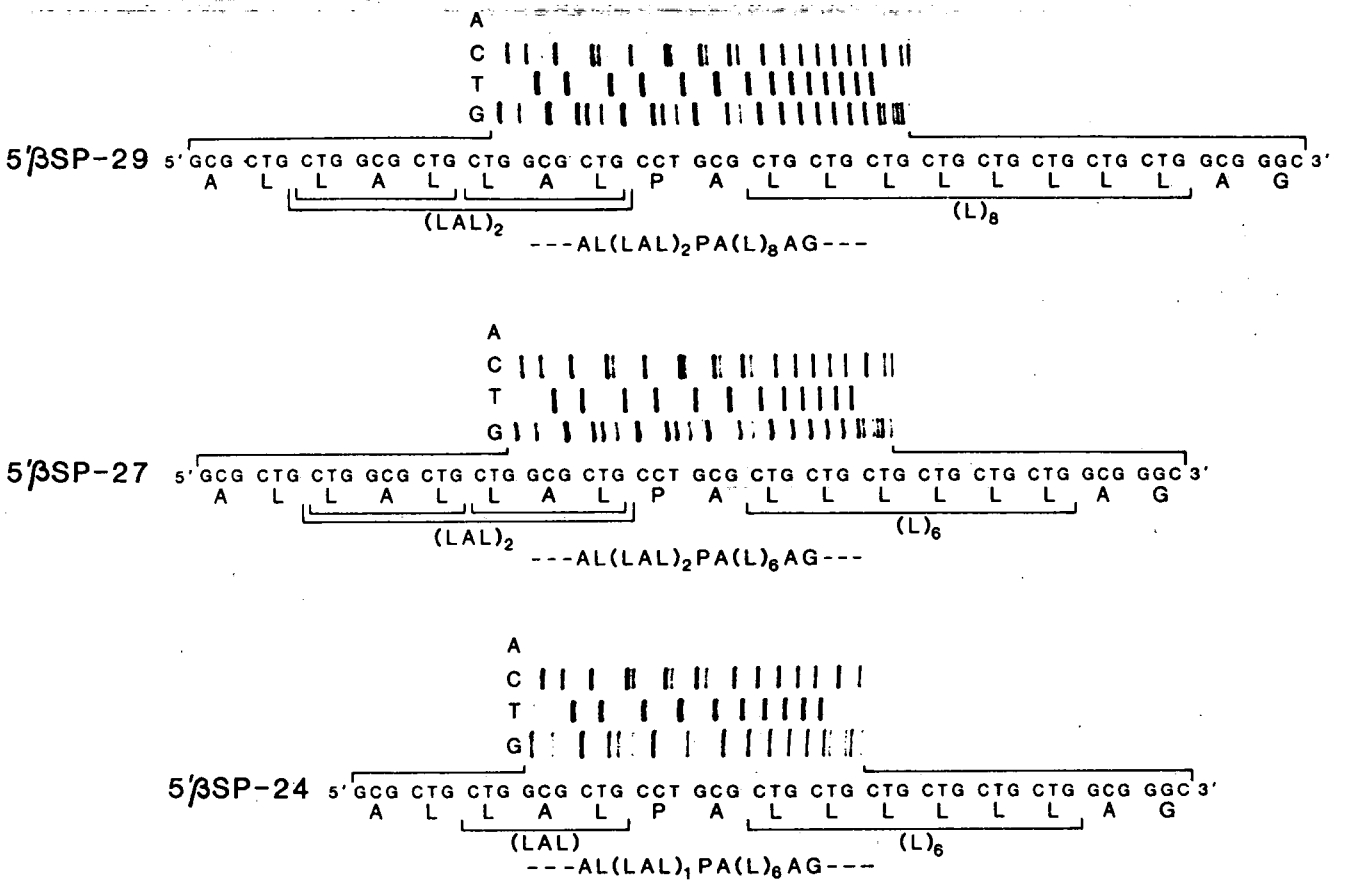


FIG. 2. Sequence analysis of 5′βSP-24, 5′βSP-27, and 5′βSP-29 polymerase chain reaction (PCR) products. PCR products from genomic DNA of individuals containing these alleles were produced and subcloned into plasmid pGEM3Z as described in METHODS. Multiple clones form each allele were sequenced by method of Sanger et al. (16). Representative autoradiograms of sequencing gels are shown. Note that in this region of sequence there is no adenine. A, alanine; G, glycine; L, leucine; P, proline.

differed by three amino acids (9 base pairs [bp]). The amplification product of 5′βSP-27, the larger allele of whites, was 93 bp and that of 5′βSP-24, the smaller allele, was 84 bp. Heterozygous individuals yielded PCR products of both sizes in approximately equal molar amounts. In Mexican Americans, a third allele was observed, 5′βSP-29, whose amplification product was 99 bp in length. The 5′βSP-29 allele was not observed in the sample of whites from Nancy, France.

Although the sizes of the PCR products were consistent with 24–, 27–, and 29–amino acid signal peptide alleles, we needed direct-sequence confirmation of the length and specific codons involved. DNA sequence analyses indicated that the signal peptide alleles consisted of the following (Fig. 2). The longest allele (5′βSP-29) encoded 29 amino acids in the signal peptide and contained two copies of the sequence CTG GCG CTG encoding Leu-Ala-Leu and a consecutive run of eight CTG codons encoding eight Leu residues. The medium-sized allele 5′βSP-27 encoded 27 amino acids and contained two copies of CTG GCG CTG but has a run of only six CTG codons. The shortest allele (5′βSP-24) encoded 24 amino acids and contained a single copy of CTG GCG CTG and a run of six CTG codons.

The PCR analysis were performed on genomic DNA. It is difficult from genomic DNA alone to determine whether both alleles are actually expressed in humans. The human hepatoma cell line HepG2 has been studied with respect to its apoB production and lipoprotein assembly (17–19). We analyzed the genomic DNA from this cell line and found it to be heterozygous for the 5′βSP-24 and 5′βSP-27 alleles. PCR analysis of RNA isolated from these cells indicated that both alleles are expressed at the mRNA level in roughly equal proportions (Fig. 3). Therefore, these alleles are probably also expressed in vivo. However, we have no information on the translational efficiency of the two mRNAs that are expressed or the early signal peptide processing within the cell. The absence of an association between the polymorphism and plasma apoB levels (see below) suggests that the biosynthetic rate of apoB100 is not affected by the difference in the length of the signal peptide. Because these subjects were fasting, no data exist concerning the role of apoB signal peptide–length variation in the synthesis and function of apoB48-containing postprandial lipoprotein particles.

The observed frequencies of each of the apoB signal peptide genotypes are given in Table 1. The apoB signal peptide was polymorphic, and the heterozygous 5′βSP-24/27 genotype was the most frequent type in each group. The 5′βSP-29 allele was found exclusively in the sample of Mexican Americans and was observed in both the random and NIDDM Mexican-American samples. No individuals homozygous for 5′βSP-29 allele were observed; this result is not

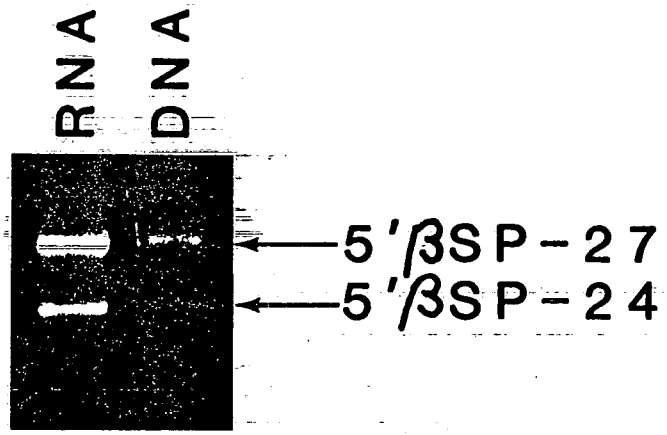


FIG. 3. Agarose gel analysis of apolipoprotein B polymerase chain reaction (PCR) products from HepG2 cells. PCR products amplified from signal peptide region of purified HepG2 RNA and DNA were fractionated on 3% agarose gel. *Double band* on *Right* indicates that HepG2 is heterozygous for 5′βSP-24 and 5′βSP-27; *double band* on left indicates that both alleles are expressed at RNA level.

surprising given the low frequency of this allele (Table 1). The 5′βSP-27 allele is the most frequent allele in each of the three groups. The relative frequencies of the 5′βSP-24 and 5′βSP-27 apoB signal peptide alleles in the French population were 0.355 and 0.645, respectively. The relative frequencies of the 5′βSP-24, -27, and -29 alleles in the Mexican-American population were 0.337, 0.630, and 0.033, respectively. The frequencies of the 5′βSP-24 and -27 alleles were not significantly different between the French and Mexican-American samples. The allele frequencies also were not significantly different between the sample of diabetic individuals and the random sample of Mexican Americans.

In the sample from Nancy, France, average cholesterol, LDL-chol, and apoB levels were not different among signal peptide genotypes (Table 2). Average apoAI and glucose levels and, to a lesser extent, triglyceride levels were different among signal peptide genotypes (Table 2). Individuals homozygous for the 5′βSP-24 allele had significantly higher plasma apoAI levels than those with only one or no 5′βSP-24 allele (1.59 vs. 1.42 g/L, respectively). Heterozygous individuals had average glucose levels that were significantly higher than the other two genotypes. Plasma triglyceride levels followed the same trend as apoAI levels; homozygous 5′βSP-24/24 individuals had higher levels than the other two genotypes.

To confirm and extend these suggestive findings, the sample of Mexican Americans was considered. Mexican Americans from Starr County, Texas, have an increased prevalence of NIDDM (9). Confirming the results obtained in the sample from Nancy, France, plasma cholesterol, LDL-chol, and apoB levels were not significantly different among signal peptide genotypes (Table 3). Due to the low frequency of the 5′βSP-29 allele, too few individuals were available for meaningful association studies. However, preliminary analyses indicated that this allele did not have a large effect on any of the laboratory measures (data not shown). In the random sample of Mexican Americans, average glucose levels were again significantly different among apoB signal peptide genotypes. However, the rank orders of average glucose levels were not the same in the sample of Mexican Americans and whites. In the Mexican Americans, average plasma glucose levels were elevated (6.14 mM) in the 5′βSP-24/24 homozygotes and lower in the other signal peptide genotypes. Consistent with this result for plasma glucose levels, and probably a consequence of it, blood levels of glycosylated hemoglobin, a monitor of integrated plasma glucose levels in the previous 3 mo, were also elevated in the homozygous 5′βSP-24/24 individuals. Plasma concentrations of C-peptide levels, which have a longer half-life than insulin concentrations and are an index of pancreatic insulin reserve, were different among the common genotypes. Homozygous 5′βSP-24/24 and heterozygous 5′βSP-24/27 individuals had elevated C-peptide levels, and homozygous 5′βSP-27/27 individuals had lower C-peptide levels. Insulin levels were not significantly different among genotypes.

TABLE 1
Observed frequencies of apolipoprotein B signal peptide variants

	Whites from Nancy, France (random)	Mexican Americans	
		Random	Non-insulin-dependent diabetes mellitus
Genotypes			
Total	1.0 (197)	1.0 (181)	1.0 (203)
5′ βSP-24/24	0.076 (15)	0.094 (17)	0.108 (22)
5′ βSP-24/27	0.558 (110)	0.464 (84)	0.517 (105)
5′ βSP-24/29	0	0.022 (4)	0.020 (4)
5′ βSP-27/27	0.365 (72)	0.376 (68)	0.315 (64)
5′ βSP-27/29	0	0.044 (8)	0.040 (8)
Alleles			
5′ βSP-24	0.355	0.337	0.377
5′ βSP-27	0.645	0.630	0.594
5′ βSP-29	0	0.033	0.029

n in parentheses.

DISCUSSION

ApoB is initially synthesized with an NH_2-terminal signal peptide sequence (23). This signal peptide is cleaved during transport of the polypeptide from the site of synthesis to the endoplasmic reticulum. Recently, it has been shown that apoB is translocated though the endoplasmic reticulum by a novel mechanism that maintains its solubility (24). Information specifying cellular or subcellular localization may also reside in the signal peptide sequence. Further studies on the role of apoB signal peptide variation on apoB synthesis and metabolism will shed light on the role of signal peptides in protein function and metabolism. As a cell that is heterozygous for the 5′βSP-24 and -27 alleles, the HepG2 cell should serve as a useful model to study the effect of the different-length alleles on apoB expression.

The nomenclature presented here for the signal peptide polymorphism (e.g., 5′βSP-24) indicates that the gene is apoB, specifies that the gene region is the signal peptide, and gives the length of the signal peptide in amino acids. We prefer this nomenclature over that suggested previously (5) because of its informativeness and flexibility and the fact that it is independent of the typing technique such as the position of the amplifying oligonucleotides. The previously

TABLE 2
Average lipid, apolipoprotein, glucose, and insulin levels in random sample of adults from Nancy, France

Variable	5′βSP-24/24 (n = 15)	5′βSP-24/27 (n = 110)	5′βSP-27/27 (n = 72)	Total (n = 197)	$\sqrt{MSE}$	P
Cholesterol (mM)	5.72	5.96	5.90	5.92	1.08	0.93
Low-density lipoprotein cholesterol (mM)	4.28	4.46	4.32	4.40	1.13	0.79
Triglycerides						
mM	1.63	1.27	1.11	1.24	0.93	0.08
ln mM	0.06	0.02	−0.03	0.01	0.21	0.12
High-density lipoprotein cholesterol (mM)	1.46	1.27	1.36	1.32	0.35	0.14
Apolipoprotein B (g/L)	1.23	1.20	1.14	1.18	0.29	0.29
Apolipoprotein A-I (g/L)	1.59	1.41	1.45	1.44	0.23	0.03
Glucose (mM)	5.23	5.65	5.38	5.52	0.76	0.05

$\sqrt{MSE}$, square root of mean square error from analysis of covariance; P, probability of equality of adjusted means.

suggested nomenclature referred to the 5′βSP-24 allele as a deletion and the 5′βSP-27 allele as an insertion (5). The nomenclature we use makes no assumption about the mechanism of origin of the signal peptide alleles. It also allows for the naming of additional signal peptide alleles should they be described later.

The apoB signal peptide polymorphism was associated with altered levels of several measures of lipid and carbohydrate metabolism (Tables 2 and 3). What are the possible links between genetic variation in apoB and plasma triglyceride and glucose levels? One possible site of action of this polymorphism is in the production and secretion of VLDL lipoproteins by the liver. In individuals with NIDDM, there is an overproduction of VLDL particles, possibly due to increased influx of free fatty acids and glucose or as a direct consequence of hyperinsulinemia (2). A caveat to this pathway of action is that we did not show an effect of the apoB signal peptide on plasma apoB levels. It is also possible that the apoB signal peptide polymorphism influences intestinal chylomicron production after a meal. These postprandial chylomicron particles contain apoB48 and not apoB100. Such effects on postprandial lipemia may be reflected by differences in fasting plasma glucose and triglyceride levels but not total apoB levels.

Plasma glucose was the only variable in which the association was significant in the random samples of whites and Mexican Americans. However, the direction of the association between the samples from Nancy, France, and Starr County, Texas, were not consistent. In the sample from Nancy, average glucose levels were elevated in the 5′βSP-24/27 heterozygotes relative to the two homozygous classes. In the sample from Starr County, average glucose levels followed a consistent trend; they were elevated in the 5′βSP-24/24 homozygote genotype and reduced in the 5′βSP-27/27 homozygote genotype. Xu et al. (25) reported that the apoB signal peptide genotypes were associated with plasma triglyceride levels in a sample of 106 individuals from North Karelia, Finland. In our study, the shorter 5′βSP-24 allele was only weakly associated with triglyceride levels in the sample from Nancy. Average triglyceride levels were not significantly different among genotypes in the random sample of Mexican Americans.

There are several possible reasons for these and other discrepancies. First, the significant results reported here and elsewhere may be due to chance, i.e., they represent statistical type I errors (26). Although possible, we believe that this is unlikely in light of the repeated significant associations with related measures (e.g., glucose, triglycerides, C-peptide) among studies and the repeated nonsignificant associations with other variables (total cholesterol, LDL-chol, apoB) in these same studies. Second, the association is real but different among populations. One cause of such a dif-

TABLE 3
Average lipid, apolipoprotein, glucose, and insulin levels in random sample of adults from Starr County, Texas

Variable	5′βSP-24/24 (n = 17)	5′βSP-24/27 (n = 84)	5′βSP-27/27 (n = 68)	Total (n = 181)	$\sqrt{MSE}$	P
Cholesterol (mM)	4.88	5.10	5.02	5.06	0.91	0.67
Low-density lipoprotein cholesterol (mM)	3.21	3.15	3.13	4.14	0.79	0.91
Triglycerides						
mM	1.28	1.69	1.44	1.59	0.99	0.17
ln mM	−2.46	−2.39	−2.47	−2.43	0.27	0.13
High-density lipoprotein cholesterol (mM)	1.15	1.15	1.25	1.20	0.27	0.08
Apolipoprotein B (g/L)	0.86	0.94	0.91	0.92	0.21	0.53
Apolipoprotein AI (g/L)	1.25	1.12	1.17	1.16	0.22	0.27
Glucose (mM)	6.14	5.63	5.34	5.59	1.02	0.01
Insulin						
pM	102.60	104.40	90.00	99.90	62.4	0.20
ln pM	2.89	2.91	2.87	2.88	0.27	0.24
C-peptide (nM)	1.00	1.09	0.91	1.01	0.53	0.05
HbA_{1c} (%)	6.60	5.90	6.10	6.10	0.92	0.05

$\sqrt{MSE}$, square root of mean square error from analysis of covariance; P, probability of equality of adjusted means.

ference could be that there are unidentified factors interacting with the apoB signal peptide, and these factors are different among populations. Finally, we favor the possibility that the signal peptide polymorphism did not directly cause the observed effect but rather was in linkage disequilibrium with a second locus with a direct effect on glucose and triglyceride metabolism. In addition, the magnitude and direction of this disequilibrium was different among populations. Regardless of the exact mechanisms, the apoB signal peptide should be included in any candidate-gene approach to the study of NIDDM and its complications.

ACKNOWLEDGMENTS

E.B. thanks Ms. Noosheen Behzadpour and other laboratory personnel for their valuable assistance. This work was supported by National Institutes of Health Grants HL-40613, (E.B.). HL-34823 (C.L.H), HL-27341, and DK-27685; NIH Grant 90-IJ-CX-0038; Baylor Diabetes and Endocrinology Research Center; and The March of Dimes Birth Defects Foundation (L.C.).

REFERENCES

1. Rossini AA, Mordes JP, Handler ES: Speculation on etiology of diabetes mellitus: tumbler hypothesis. *Diabetes* 37:257–61, 1988
2. Howard BV, Abbott WF, Beltz WF, Harper I, Fields RM, Grundy SM, Taskinen RM: Integrated study of low density lipoprotein and very low density lipoprotein metabolism in non-insulin dependent diabetics. *Metabolism* 36:870–77, 1987
3. Betteridge DF: Lipoprotein metabolism. In *Recent Advances in Diabetes*. Nattrass M, Ed. New York, Churchill Livingstone, 1986, p. 91–107
4. Xiang K-S, Cox NJ, Huang P, Karam JH, Bell GI: Insulin-receptor and apolipoprotein genes contribute to development of NIDDM in Chinese Americans. *Diabetes* 38:17–23, 1989
5. Boerwinkle E, Chan L: A three codon insertion/deletion polymorphism in the signal peptide region of the human apolipoprotein B (apo B) gene directly typed by the polymerase chain reaction. *Nucleic Acids Res* 17:4003, 1989
6. Chan L, Bradley WA: Signal peptides: properties and interactions. In *Cellular Regulation of Secretion and Release*. Conn PM, Ed. New York, Academic, 1986, p. 301–21
7. Benson SA, Hall MN, Silhavy TJ: Genetic analysis of protein export in *Eschericia coli* K12. *Annv Rev Biochem* 54:101–34, 1985
8. Randall LL, Hardy SJS: Unity in function in the absence of consensus in sequence: role of leader peptides in export. *Science* 243:1156–59, 1989
9. Hanis CL, Ferrell RE, Barton SA, Aguilar L, Garza-Ibarra A, Tulloch BR, Garcia CA, Schull WJ: Diabetes among Mexican Americans in Starr County, Texas. *Am J Epidemiol* 118:659–72, 1983
10. Allain CC, Poon LS, Chan CSG, Richmond W, Fu PC: Enzymatic determination of total serum cholesterol. *Clin Chem* 20:470–75, 1974
11. Steinmetz J, Panek E: Adaptation sur GSAII Greiner du dosage des triglycerides par voie entierement enzymatique: application a l'etude de leur conservation et de certaines interferences analytiques. *J Clin Chem Clin Biochem* 16:613–19, 1978
12. Trinder P: Determination of blood glucose using 6 aminophenazone as oxygen acceptor (Letter). *J Clin Pathol* 22:246, 1969
13. Warnick GR, Benerson J, Albers JJ: Dextran sulfate-Mg++ precipitation procedure for quantitation of high density lipoprotein cholesterol. *Clin Chem* 28:1379–88, 1972
14. Friedewald WT, Levy RI, Fredrickson DS: Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifugation. *Clin Chem* 18:499–502, 1972
15. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi RG, Horn TT, Mullis KB, Erlich HA: Primer-directed enzymatic amplification of DNA with thermostable DNA polymerase. *Science* 239:487–91, 1988
16. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain terminating inhibitors. *Proc Natl Acad Sci USA* 74:5403–67, 1977
17. Thrift RN, Forte, TM, Cahoon BE, Shore VG: Characterization of lipoproteins produced by the human liver cell line, Hep G2, under defined conditions. *J Lipid Res* 27:236–50, 1986
18. Bostrom K, Boren J, Wettesten M, Sjoberg A, Bondjers G, Wiklund O, Carlsson P, Olofsson S: Studies on the assembly of apo B-100–containing lipoproteins in Hep G2 cells. *J Biol Chem* 263:4434–42, 1988
19. Boren J, Wettesten M, Sjoberg A, Thorlin T, Bondjers G, Wiklund O, Olofsson S-O: The assembly and secretion of apoB-100–containing lipoproteins in HepG2 cells: evidence for different sites for protein synthesis and lipoprotein assembly. *J Biol Chem* 265:10556–64, 1990
20. Chirgwin JM, Przybyla AE, MacDonald RJ, Rutter WJ: Isolation of biochemically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* 18:5294–99, 1979
21. Sokal R, Rohlf FJ: *Biometry*. New York, Freeman, 1980
22. Neter J, Wasserman W: *Applied Linear Statistical Models*. Homewood, IL, Irwin, 1974
23. Yang C, Gu ZW, Weng S, Kim TW, Chen SH, Pownall HJ, Sharp PM, Liu SW, Li WH, Gotto AM, Chan L: Structure of apolipoprotein B-100 of human low density lipoproteins. *Arteriosclerosis* 9:96–108, 1089
24. Chuck SL, Yao Z, Blackhart BD, McCarthy BJ, Lingapp VR: New variation on the translocation of proteins during early biogenesis of apolipoprotein B. *Nature* (Lond) 346:382–85, 1990
25. Xu C-F, Tikkanen MJ, Huttunen JK, Pietinen P, Butler R, Humphries S, Talmud P: Apolipoprotein B signal peptide insertion deletion polymorphism is associated with Ag epitopes and involved in the determination of serum triglyceride levels. *J Lipid Res* 31:1255–61, 1990
26. Larson RJ, Marx ML: An *Introduction to Mathematical Statistics and Its Applications*. Inglewood Cliffs, NJ, Prentice-Hall, 1981

Apolipoprotein(a) Gene Accounts for Greater Than 90% of the Variation in Plasma Lipoprotein(a) Concentrations

Eric Boerwinkle,* Carla C. Leffert,‡ Jingping Lin,* Carolin Lackner,‡ Giulia Chiesa,‡ and Helen H. Hobbs‡
**Center for Demographic and Populations Genetics, University of Texas Health Science Center in Houston, Houston, Texas 77225; and ‡Departments of Internal Medicine and Molecular Genetics, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75235*

Abstract

Plasma lipoprotein(a) [Lp(a)], a low density lipoprotein particle with an attached apolipoprotein(a) [apo(a)], varies widely in concentration between individuals. These concentration differences are heritable and inversely related to the number of kringle 4 repeats in the apo(a) gene. To define the genetic determinants of plasma Lp(a) levels, plasma Lp(a) concentrations and apo(a) genotypes were examined in 48 nuclear Caucasian families. Apo(a) genotypes were determined using a newly developed pulsed-field gel electrophoresis method which distinguished 19 different genotypes at the apo(a) locus. The apo(a) gene itself was found to account for virtually all the genetic variability in plasma Lp(a) levels. This conclusion was reached by analyzing plasma Lp(a) levels in siblings who shared zero, one, or two apo(a) genes that were identical by descent (ibd). Siblings with both apo(a) alleles ibd (n = 72) have strikingly similar plasma Lp(a) levels (r = 0.95), whereas those who shared no apo(a) alleles (n = 52), had dissimilar concentrations (r = −0.23). The apo(a) gene was estimated to be responsible for 91% of the variance of plasma Lp(a) concentration. The number of kringle 4 repeats in the apo(a) gene accounted for 69% of the variation, and yet to be defined *cis*-acting sequences at the apo(a) locus accounted for the remaining 22% of the inter-individual variation in plasma Lp(a) levels. During the course of these studies we observed the de novo generation of a new apo(a) allele, an event that occurred once in 376 meioses. (*J. Clin. Invest.* 1992. 90:52–60.) Key words: apolipoprotein (a) • lipoprotein (a) • low density lipoprotein

Introduction

Lipoprotein(a) [Lp(a)][1] is a cholesterol ester-rich plasma lipoprotein comprising two attached components: a low density lipoprotein (LDL) particle and a single large glycoprotein, apolipoprotein(a) [apo(a)] (1–3). High plasma levels of Lp(a) are associated with the development of coronary atherosclerosis (4–6) and other vascular diseases (7). The mechanism by which Lp(a) expedites the atherosclerotic process is not known. Apo(a) strongly resembles plasminogen, and it may competitively interfere with plasminogen action in fibrinolysis (8, 9).

Plasma concentrations of Lp(a) vary over a wide range among individuals, but are remarkably stable in any given individual (10). Many physiological, pharmacological, and environmental factors that affect the levels of other plasma lipoproteins have no effect on the plasma concentration of Lp(a) (10). This lack of environmental and physiological influences suggests that plasma Lp(a) levels are largely genetically determined. Consistent with this formulation, early genetic studies suggested that the presence of Lp(a) in plasma was inherited as an autosomal dominant trait (11–13). When more sensitive immunoassays of plasma Lp(a) concentrations were used, it was found that plasma Lp(a) concentrations varied continuously among individuals (14), and the pattern of inheritance indicated that a major gene, as well as polygenic factors, contributed to plasma Lp(a) concentrations (15–17).

Fless et al. (18) and Utermann et al. (19) found that the apo(a) glycoprotein varied in size among individuals. In an important series of studies, Utermann and his colleagues demonstrated that the size of the apo(a) protein is inversely related to the level of plasma Lp(a), thus implicating the apo(a) gene as a major determinant of plasma Lp(a) concentrations (20–23). However, the immunoblotting technique used to type the apo(a) isoforms was not sensitive enough to detect low levels of apo(a) protein, and not all of the apo(a) isoforms were detected. As a result, the frequency distribution of the apo(a) isoforms failed to fit the expectations of Hardy-Weinberg equilibrium (22). In addition, when immunoblotting was employed to examine the segregation of the apo(a) isoforms in families, the results were frequently uninformative, and occasionally inconsistent (24). Further progress required the development of a technique that was more discriminating than immunoblotting in classifying apo(a) alleles.

A potential method to study this polymorphism was suggested by the findings of McLean et al. who discovered that the apo(a) cDNA contains multiple tandem copies of a sequence that encodes a cysteine-rich protein motif called a kringle. The repeated kringle in apo(a) is designated kringle 4 because it closely resembles the fourth kringle in plasminogen. McLean et al. proposed that the apo(a) isoforms are of different size because of variations in the numbers of kringle 4-encoding repeats in the apo(a) gene (25). This hypothesis was supported by studies of the apo(a) mRNA and gene structure (26–28). In attempt to devise a way to measure the size of the apo(a) gene in different individuals, we previously identified a large restriction fragment from the apo(a) gene which contains most, if not all, of the kringle 4–encoding sequences (29). The size of this fragment was too large to be examined by standard electrophoresis techniques. Accordingly, we used pulsed-field gel electrophoresis to size this large restriction fragment and 19 fragments of different length were identified. A total of 103 unrelated

Address reprint requests to Dr. Hobbs, Department of Molecular Genetics, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd., Dallas, TX 75235.

Received for publication 6 December 1991.

1. *Abbreviation used in this paper:* Lp(a), lipoprotein (a).

J. Clin. Invest.

0021-9738/92/07/0052/09 $2.00
Volume 90, July 1992, 52–60

Caucasians were evaluated and 94% were heterozygous for fragments of two different sizes. The length polymorphism was used as a genetic marker to analyze the segregation of the apo(a) gene in 12 Caucasian families. It was found that within a given family, sibling pairs with identical apo(a) genotypes tended to have very similar plasma Lp(a) levels (29). However, individuals with the same apo(a) genotypes who were members of different families often had significantly different plasma concentrations of Lp(a). Taken together, these observations suggest that the apo(a) gene is the major determinant of plasma Lp(a) levels and that cis-acting DNA sequences at or near the apo(a) locus, other than the number of kringle 4 repeats, contribute importantly to plasma Lp(a) concentrations.

In the current study, we analyzed the segregation of the apo(a) gene and Lp(a) levels in 48 Caucasian pedigrees to determine the contribution of the apo(a) gene (or closely linked loci) to the plasma concentrations of Lp(a). The genetic architecture (30) of plasma Lp(a) concentrations was defined at three levels: the polygenic heritability, the total genetic contribution of the apo(a) gene, and the effects of length variation in the apo(a) gene. In addition, we describe the de novo generation of a new apo(a) allele of different size within a family.

Methods

Subjects. Plasma Lp(a) concentrations were measured in a sample of 288 fasting individuals from 48 Caucasian American families living in the greater Dallas, Texas area. Families in which both parents and at least three children were available for sampling were selected for study. None of the families had evidence of a monogenic hyperlipidemia. In one family, F153, several members (who are denoted in Table II) had very low LDL-cholesterol levels, suggesting the possible existence of familial hypobetalipoproteinemia. For the random effects analysis of variance (see below), the families were augmented with a sample of 107 unrelated individuals. Preliminary findings on a subset of these unrelated individuals have been reported previously (29).

Phlebotomy was carried out after an overnight fast. A total of 30 ml of blood was collected from each individual in vacutainer tubes containing sodium-EDTA. The plasma was separated within one hour of collection by centrifugation at 2,000 *g* for 15 min at 4°C. Multiple 50 μl aliquots of plasma were stored at −70°C and Lp(a) levels were assayed within 4 wks.

Pulsed-field gel analysis of the apo(a) gene. A total of 15 ml of blood was maintained at room temperature prior to transfer to two LeucoPREP tubes (Becton, Dickinson & Co., Lincoln Park, NJ). Lymphocytes were isolated and embedded in agarose plugs as previously described (29). The agarose-cellular plugs were incubated twice with 40 U of KpnI in 170 μl of the buffer suggested by the manufacturer (New England Biolabs, Beverly, MA). The digested cellular-agarose plugs were subjected to pulsed-field gel electrophoresis in a vertical submarine gel apparatus with a transverse alternating field (Geneline I, Beckman Instruments, Inc., Fullerton, CA) using low-endosmosis coefficient agarose, TAFE buffer, and λ phage concatamer standards (Beckman Instruments, Inc.) as described by Lackner et al. (29). After 18 hours of electrophoresis, the gel was stained with ethidium bromide and photographed. The DNA was transferred and fixed to nylon membrane (Biotrans, ICN Biomedicals, Costa Mesa, CA). MP-1, a 1.5-kb PstI genomic fragment from the kringle 4-encoding region of the apo(a) gene (29) which had been subcloned into M13mp18, was used to generate a ^{32}P-radiolabeled single-stranded probe (31). The filter was incubated overnight at 42°C in hybridization solution containing 5 × 10^6 cpm/ml of the single-stranded apo(a)-specific probe. Hybridizations were carried out in a rotating incubator (model 310, Robbins Scientific Corp., Sunnyvale, CA). Filters were washed as described by Lackner et al. (29) and exposed to film.

Immunoblotting of plasma apo(a). An aliquot of frozen plasma (between 1 and 30 μl) containing 1 μg of Lp(a) was brought up to a total volume of 30 μl using phosphate-buffered saline. The sample was mixed with 20 μl of buffer A which contained 15% filtered SDS (wt/vol), 8 M urea, 5 mM dithiothreitol, and 62.5 mM Tris at pH 7.5 and with 50 μl buffer B (10% glycerol [vol/vol], 2.3% SDS [wt/vol], 0.025% bromophenol blue [wt/vol], 5% β-mercaptoethanol [vol/vol], and 50.0 mM Tris at pH 6.8). The samples were boiled for 10 min before loading onto a 3–7% gradient polyacrylamide gel with SDS. A total of 1 μg of purified LDL (molecular weight of apo B is ~ 513 kD) was used as a size standard. The electrophoresis, transfer to nitrocellulose, and hybridization conditions were exactly as preciously described except that IgG-1A^2, the apo(a)-specific antibody, was radiolabeled directly with ^{125}I to a specific activity of 5 × 10^6 cpm/ml (29). The filters were washed, dried and exposed to XAR-5 film (Eastman Kodak Co., Rochester, NY) at −70°C with an intensifying (Lightening Plus, Dupont Co., Wilmington, DE).

Plasma lipid and lipoprotein assays. Measurement of plasma Lp(a) concentrations were performed at GeneScreen, Dallas, TX, using a sensitive enzyme-linked immunosorbent sandwich assay (ELISA), as described (32). In this assay, Lp(a) was captured by a polyclonal rabbit anti-human Lp(a) antibody and then detected by a monoclonal anti-human Lp(a) antibody, IgG-1A^2. Plasma Lp(a) standards were obtained from Immuno, Vienna, Austria. Total cholesterol and triglyceride levels were measured enzymatically using commercially available kits (Boehringer Mannheim, Indianapolis, IN; Sigma Chemical Co., St. Louis, MO). Plasma lipoproteins were quantified in the laboratory of Dr. Scott Grundy (University of Texas Southwestern Medical Center) according to the procedures of the Lipid Research Clinic (33).

Statistical methods. The distribution of plasma Lp(a) concentration was positively skewed in these data, and thus all analyses were carried out both on the raw and square-root transformed data. For each analysis, the primary inferences were identical whether the raw or transformed data was used.

The contribution of unmeasured polygenic variation to the inter-individual variability of plasma Lp(a) concentrations ($\sigma_{Lp(a)}^2$) was assessed from the extent of familial aggregation of Lp(a) levels in the sample of pedigrees. The ratio of the polygenic variance component (σ_{pg}^2) to $\sigma_{Lp(a)}^2$ was estimated by maximum likelihood principles as implemented in the computer program PAP V3.0 (34).

Sibling-pair linkage methods were used to estimate the overall contribution of genetic variation in and around the apo(a) gene ($\sigma_{apo(a)}^2$) to plasma Lp(a) levels (35, 36). These methods are most frequently employed to detect linkage between a marker and a quantitative trait locus, but can also be used to define the overall contribution of a candidate gene to a quantitative phenotype. For each sibling pair, three new variables were considered: y_j, the squared difference of plasma Lp(a) concentrations in sibship j, f_{1j}, an indicator variable describing whether or not the jth sib pair shares only 1 allele identical by descent (ibd), and π_j, the proportion of alleles ibd in sibship j. π_j can take on the values 0, ½, or 1. $E(y_j)$ is the expected value of an individual's Lp(a) concentration. Assuming there is no recombination as would be the case for a candidate gene, Haseman and Elston (30) show that:

$$E(y_j) = \alpha + \beta\pi_j + \gamma f_{1j} \quad (1)$$

where

$$\alpha = 2\sigma_{apo(a)}^2 + \sigma_e^2 \quad (2)$$

$$\beta = -2\sigma_{apo(a)}^2 \quad (3)$$

$$\gamma = -\sigma_d^2 \quad (4)$$

In Eqs. 2–4, σ_e^2 is a residual variance component describing the effects of factors other than the apo(a) gene on Lp(a) levels, and σ_d^2 describes the dominance effects at the apo(a) locus on Lp(a) levels. An estimate of the overall contribution of the apo(a) gene to plasma Lp(a) concentrations can be made by examining the regression of the squared differ-

Table I. Correlations of Plasma Lp(a) Concentrations between Family Members

	n	Lp(a)	√Lp(a)
Spouses	48	0.17 [−0.12, 0.43]*	0.17 [−0.12, 0.43]
Parent-offspring	400	0.44§ [0.36, 0.52]	0.48§ [0.40, 0.55]
Midparent-offspring	200	0.59§ [0.49, 0.67]	0.61§ [0.51, 0.69]
Siblings (all)	284	0.28§ [0.16, 0.39]	0.32§ [0.21, 0.43]
Siblings sharing no alleles ibd	52	−0.23 [−0.47, 0.05]	−0.25 [−0.48, 0.02]
Siblings sharing one allele ibd	159	0.15 [−0.16, 0.30]	0.19‡ [0.04, 0.34]
Siblings sharing two alleles ibd	73	0.95§ [0.92, 0.97]	0.96§ [0.94, 0.97]

* 95% confidence interval. ‡ $P < 0.05$. § $P < 0.001$.

ence between the Lp(a) levels of siblings who share none, one, or all apo(a) alleles ibd. The regression analyses were performed both unweighted and weighted, as suggested by Amos et al. (36), with nearly identical results. Therefore, only the results of the unweighted analyses are presented. Even though the sibships were typically larger then size two, the above method has been shown to be valid when overlapping sibling pairs are analyzed as though they were independent (36).

The contribution of length variation in the apo(a) gene, as measured by pulsed-field gel electrophoresis, to Lp(a) concentrations, σ_{length}^2, was estimated using a random effects analysis of variance (37). A random effects or type II model was selected because of the large number of potential genotypes at the apo(a) locus (38).

Results

Plasma Lp(a) concentrations were measured in 288 individuals from 48 pedigrees. There was no significant effect of age, sex, or the concentration of other plasma lipoproteins on the plasma level of Lp(a), so these factors were not considered further in the family members (data not shown). There were significant correlations between the plasma Lp(a) levels of parents and offspring ($r = 0.44$), and siblings ($r = 0.28$), but not between spouses ($r = 0.17$) (Table 1). By using standard biometrical genetic analyses, it was estimated that 85% (±8%) of the inter-individual variance of Lp(a) concentrations was attributable to polygenic effects ($\sigma_{pg}^2/\sigma_{Lp(a)}^2$) (or 88% (±6.5) when the square-root of the plasma Lp(a) levels was used).

Pulsed-field gel electrophoresis and genomic blotting of KpnI digested-genomic DNA was performed to assess the size of the kringle 4–encoding region of the apo(a) alleles in each family member. 16 of the 19 previously described apo(a) alleles were observed in the sample, and their frequencies did not differ significantly from those previously described from the same population (29). In general, there was an inverse relationship between the size of the apo(a) allele and the plasma level of Lp(a). One way to illustrate this phenomenon is to examine the relationship between the plasma Lp(a) concentrations and the apo(a) allele size in the individuals who had one of the two most common alleles, *apo(a)14* or *apo(a)15* plus a different allele (Fig. 1). Individuals with one copy of *apo(a)14* or *apo(a)15* plus one copy of *apo(a)2-apo(a)4* tended to have high plasma Lp(a) levels (> 30 mg/dl). If the second allele was *apo(a)5-apo(a)7*, the Lp(a) levels were lower (15–30 mg/dl). If the second allele was larger than *apo(a)8*, the plasma concentrations of Lp(a) were low [< 10 mg/dl, excluding *apo(a)10*].

In the population as a whole, different apo(a) genotypes, as determined by pulsed-field gel electrophoresis, were associated with significantly different plasma levels of Lp(a) ($P < 0.001$ for both the raw and transformed data). A random effects analysis of variance (37) was used to determine the contribution of the length variation in the apo(a) gene to the distribution of plasma Lp(a) in 203 unrelated Caucasians. For the raw data, 69% of the variation in Lp(a) concentrations was attributable to inter-individual differences in the number of kringle-4 repeats. The square-root transformation had little effect on this value (66% vs. 69%).

Although length variation in the apo(a) gene had a profound influence on Lp(a) concentrations, there were several exceptions to the general trend. Fig. 2 shows two pedigrees in which an apo(a) allele of the same size, *apo(a)6*, segregates. In the two pedigrees this allele gives rise to very different plasma concentrations of Lp(a). In *A*, the *apo(a)6* allele of the father (*a*), is inherited by three of his offspring (*c*, *d*, and *f*). The father, as well as the three offspring, have modest plasma Lp(a) concentrations (6 mg/dl, and 7, 5, and 3 mg/dl, respectively). In the family shown in *B*, individual h, who is also heterozygous for an allele the size of *apo(a)6*, and has a high plasma Lp(a) concentration (51 mg/dl). Of her four children, only the second child (*j*) inherited *apo(a)6* and she is the only offspring with a comparable plasma level of Lp(a) (51 mg/dl). Therefore, in these two families, the same sized apo(a) allele (*apo(a)6*) segregated with very different plasma levels of Lp(a). This was true even though the other alleles at the apo(a) locus in the families were similar (*apo(a)13-apo(a)17*). These findings suggest that factors at the apo(a)

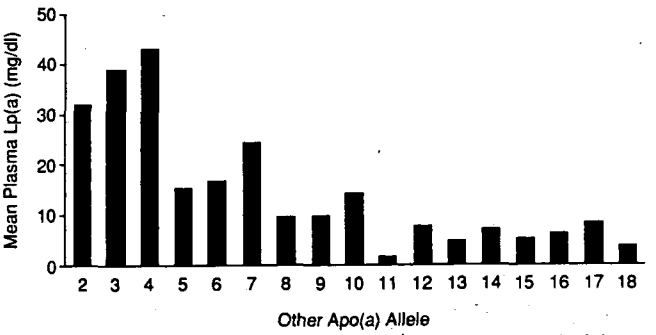


Figure 1. Lp(a) levels in individuals heterozygous for apo(a)14 or apo(a)15 allele. In the sample of 288 family members and 107 unrelated individuals, there were 194 individuals with either *apo(a)14* or *apo(a)15*. The average Lp(a) levels (*y*-axis) for individuals with each genotype are plotted against the other apo(a) allele (*x*-axis).

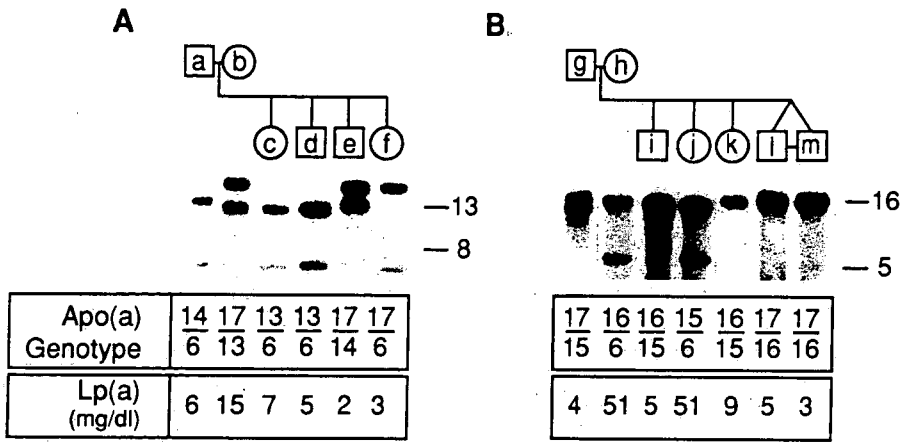


Figure 2. Genomic blot of the apo(a) gene from two unrelated families with *apo(a)6*. High molecular weight leukocyte DNA from members of two unrelated families was digested with KpnI, size-fractionated on a pulsed-field gel, transferred to a nylon membrane, and hybridized with a single-stranded apo(a)-specific probe (MP-1) as described in the Methods. The filter was exposed to Kodak XAR-5 film for 18 h with an intensifying screen. The plasma concentrations of Lp(a) were measured using an ELISA assay as described in the Methods. The *apo(a)6* allele segregates with a low (*A*) and high (*B*) plasma concentration of Lp(a) in two different pedigrees.

locus, in addition to the number of kringle 4 repeats, strongly influence the plasma Lp(a) concentration.

Another instance in which apo(a) alleles of the same size are associated with different amounts of circulating apo(a) protein, is shown in Fig. 3. In this family, the mother (b) is homozygous for *apo(a)12*, so all of the children (*c–f*) are heterozygous for that allele. Based on the genomic blot, it cannot be determined which of the two *apo(a)12* alleles each child inherited from their mother. However, analysis of the apo(a) protein isoforms reveals that three of the offspring (*c*, *d*, and *f*) have no detectable apo(a) protein corresponding to *apo(a)12*. Only offspring *e* has a band the same size as the isoform of the mother. This suggests that the mother is heterozygous and has one *apo(a)12* that produces no detectable circulating apo(a) protein which she gave to *c*, *d*, and *f* and another that is associated with the production of a moderate amount of apo(a) protein which she donated to offspring *e*.

To confirm that cis-acting sequences at the apo(a) locus are responsible for the observed differences in plasma Lp(a) concentrations in individuals with apo(a) alleles of the same size, the plasma Lp(a) concentrations were compared in sibling pairs who shared all, one, or no apo(a) alleles ibd. In 40 of the 48 families, all four parental apo(a) alleles could be differentiated using pulsed-field gel electrophoresis. In six families (including the one shown in Fig. 3), the length polymorphism was uninformative because one of the parents was apparently homozygous for the same sized apo(a) allele, and in two families one parent was not available for sampling; these eight families were not included in the sibling-pair analysis. The families were selected at random, so in some families all the plasma concentrations of Lp(a) were low (i.e., < 5 mg/dl) reflecting the highly skewed distribution of plasma Lp(a) levels in the Caucasian population.

In the 40 families in which the segregation of each parental allele could be distinguished, 72 sibling pairs shared both, 52 shared none, and 159 pairs shared one parental allele ibd. The apo(a) genotypes and plasma Lp(a) levels of these sibling pairs are given in Table II and can be compared to those of the other siblings and parents. The sibling pairs who had plasma levels of Lp(a) that were similar to each other and significantly different from the other siblings are denoted by an asterisk. In 24 families, at least one sibling pair had inherited identical apo(a) alleles and one sibling pair had no apo(a) alleles ibd (Table III). In 21 of the 24 sibling pairs (denoted by asterisks), the mean difference between plasma Lp(a) levels in the sibling pair who shared no apo(a) alleles ibd was twice that of the sibling pair who shared both apo(a) alleles ibd.

Fig. 4 shows the scatter plot of Lp(a) levels of siblings pairs that share (*A*) both or (*B*) no apo(a) alleles ibd. Lp(a) levels for the older (sibling 1) and younger (sibling 2) sibling are plotted on the horizontal and vertical axis, respectively. The correlation coefficient for Lp(a) levels between siblings who share both apo(a) alleles ibd was very high ($r = 0.95$), whereas there was a negative correlation ($r = -0.23$) between the Lp(a) concentration of siblings who share no apo(a) alleles ibd. Similar results were obtained for the square-root transformed data (Table I). Owing to the highly skewed distribution of Lp(a) in the population, many of the sibling pairs had very low Lp(a) levels. Therefore, the same comparison was made in the sibling pairs with apo(a) alleles ibd ($n = 48$) who had plasma Lp(a) levels over 5 mg/dl and the correlation coefficient remained very high ($r = 0.94$).

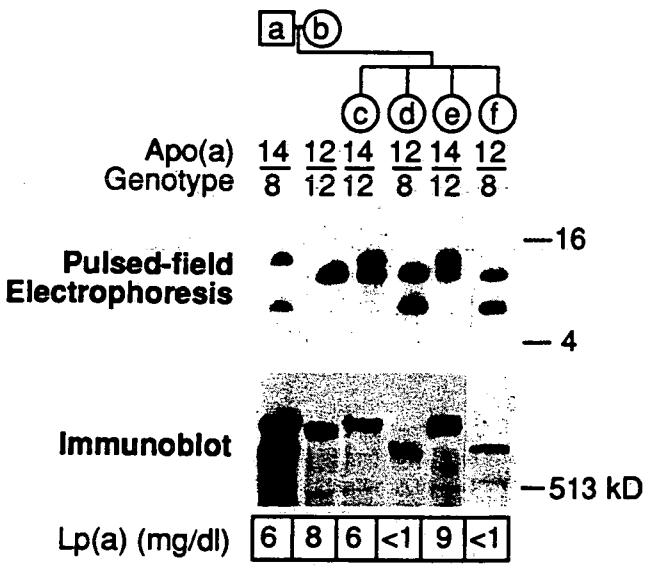


Figure 3. Genomic blot of apo(a) gene and immunoblot of apo(a) protein in a pedigree. The plasma Lp(a) concentrations were measured and the genomic blot and immunoblot was performed as described in Fig. 2 and in the Methods.

Table II. Apo(a) Genotypes and Lp(a) Levels in Sibling Pairs with Apo(a) Alleles ibd and Their Family Members

		Sib with identical apo(a) genotypes				Other Sibs		Parents (father, mother)	
	Family no.	Apo(a) genotype	Plasma Lp(a)			Apo(a) genotypes	Plasma Lp(a)	Apo(a) genotype	Lp(a)
			mg/dl				*mg/dl*		*mg/dl*
1.	F141	2/13	26		41	13/16, 15/16, 2/15	1, 6, 32	2/16, 13/15	36, 6
2.	F168	2/17	*40	*42	*55	12/17, 12/17	< 1, < 1	6/17, 2/12	40, 54
3.	F162	4/5	*44	*47	*58	5/14	21	5/15, 4/14	38, 19
4.	F24	4/10	*41		*42	10/14, 6/14	< 1, 16	4/14, 6/10	31, 62
5.	F154	4/10	58		34	12/15, 4/15, 4/15	14, 32, 43	10/15, 4/12	32, 52
6.	F161	4/13	*49		*52	13/13	10	4/13, 13/14	50, 9
7.	F158	4/14	*36		*48	8/14	10	4/8, 14/15	48, 7
8.	F156	4/15	*47		*49	10/15	7	14/15, 4/10	4, 42
9.	F135	4/15	*55	*56	*66	14/15, 14/15	3, 3	14/15, 4/14	4, 45
10.	F154	4/15	32		43	12/15, 4/10, 4/10	14, 34, 58	10/15, 4/12	32, 52
11.	F142	5/9	*49		*56	9/16	4	5/16, 9/10	72, 21
12.	F145	5/12	64		98	11/16, 11/16, 12/16, 5/11	< 1, 3, 8, 51	5/16, 11/12	47, 5
13.	F129	6/6	7		7	6/16, 6/16, 15/16	5, 6, 6	6/15, 6/16	4, 15
14.	F149	6/13	5		7	14/17, 6/17	2, 3	6/14, 13/17	6, 15
15.	F129	6/16	5		6	15/16, 6/6, 6/6	6, 7, 7	6/15, 6/16	4, 15
16.	F146	7/11	*28		*44	9/17, 9/11, 9/11	< 1, 5, 6	7/9, 11/17	22, 9
17.	F166	8/12	5		7	12/13, 13/15	17, 19	8/13, 12/15	9, 10
18.	F150	8/15	4	6	8	7/15	5	15/16, 7/8	5, 5
19.	F146	9/11	5		6	9/17, 7/11, 7/11	< 1, 28, 44	7/9, 11/17	22, 9
20.	F124	9/16	*7		*9	4/16, 5/9, 4/5	22, 50, 75	5/16, 4/9	54, 27
21.	F134	9/17	6	8	11	16/17, 9/18	12, 12	9/16, 17/18	15, 3
22.	F137	10/13	< 1		3	13/14, 13/14, 13/14	< 1, 1, 1	10/14, 13/15	< 1, 2
23.	F21	11/15	< 1		< 1	15/17, 8/11	< 1, 7	8/15, 11/17	5, < 1
24.	F143	11/15	1		2	14/15, 14/15, 14/15, 7/11	< 1, < 1, < 1, 17	7/15, 11/14	28, < 1
25.	F164	11/15	1	1	1	15/16	3	11/16, 8/15	2, 4
26.	F145	11/16	*< 1		*3	12/16, 5/11, 5/12, 5/12	8, 51, 64, 98	5/16, 11/12	47, 5
27.	F167	12/13	*15	*15	*21	12/17, 13/13	1, 5	13/17, 12/13	12, 1
28.	F138	12/14	1		5	13/15	1	14/15, 12/13	4, 2
29.	F152	12/14	*< 1		*< 1	8/13	16	12/13, 8/14	< 1, 30
30.	F160	12/14	*3		*6	10/15, 10/14	21, 34	10/12, 14/15	27, 10
31.	F125	12/15	1		1	14/15, 7/12	< 1, 28	12/14, 7/15	< 1, 32
32.	F168	12/17	*< 1		*< 1	2/17, 2/17, 2/17	40, 42, 55	6/17, 2/12	40, 54
33.	F126	12/18	< 1		< 1	15/18	< 1	12/15, 9/18	3, < 1
34.	F137	13/14	1	1	< 1	10/13, 10/13	< 1, 3	10/14, 13/15	< 1, 2
35.	F136	13/15	4		4	16/17, 15/17, 8/16	< 1, 3, 3	13/17, 15/16	1, 9
36.	F135	14/15	*3		*3	4/15, 4/15, 4/15	56, 56, 66	14/15, 4/14	4, 45
37.	F143	14/15	< 1	< 1	< 1	11/15, 11/15, 7/11	1, 2, 17	7/15, 11/14	28, < 1
38.	F131	14/15	< 1	< 1	4	15/18	< 1	14/18, 5/15	< 1, < 1
39.	F132	14/17	1	1	3	14/15, 15/17	2, 3	14/17, 14/15	8, < 1
40.	F165	14/18	8		8	15/18, 15/18	< 1, 9	14/15, 10/18	16, < 1
41.	F157	15/16	5		9	16/17, 16/17, 6/15	3, 5, 51	15/17, 6/16	4, 51
42.	F159	15/17	1	5	2	10/17	< 1	16/17, 10/15	< 1, 6
43.	F153	16/16	*1‡	*4‡	*5‡	7/16	27	7/16, 12/16	52, 1
44.	F157	16/17	3		5	15/16, 15/16, 6/15	5, 9, 51	15/17, 6/16	4, 51
45.	F165	15/18	< 1		9	14/18, 14/18	8, 8	14/15, 10/18	16, 1

* Sibling pairs with identical apo(a) genotypes who have Lp(a) levels which are significantly different from all the other siblings. ‡ These individuals have a plasma LDL-cholesterol concentration less than the 5th percentile when compared to age and sex-matched controls.

The overall contribution of the apo(a) gene to plasma Lp(a) concentrations was estimated by examining the regression of the squared difference of Lp(a) levels between siblings (y_j) based on the proportion of apo(a) alleles shared ibd (π_j). The dominance deviations at the apo(a) locus (γ) was not significantly different from zero ($\gamma = 1.27$ for untransformed Lp(a) levels) so was not considered in further analyses. The simple linear regression of the squared difference of Lp(a) levels between siblings on the proportion of apo(a) alleles shared ibd is graphically presented in Fig. 5. The average squared dif-

Table III. Lp(a) Levels in Sibling Pairs in the Same Family Who Share Both or No Apo(a) Alleles Identical by Descent

	Siblings sharing both apo(a) alleles			Siblings sharing no apo(a) alleles			
		Lp(a)		Genotype Lp(a)			
Family	Genotype	Sib1	Sib2	Sib1		Sib2	
141*	2/13	41	26	13/16	1	2/15	32
24*	4/10	42	41	4/10	42	6/14	16
154	4/10	58	34	12/15	14	4/10	34
154*	4/15	32	43	12/15	14	4/10	58
124*	4/16	22	36	5/9	50	4/16	22
145*	5/12	98	64	11/16	1	5/12	98
149*	6/13	7	5	6/13	7	14/17	2
146*	7/11	44	28	7/11	44	9/17	1
166*	8/12	5	7	8/12	5	13/15	19
146*	9/11	5	6	9/17	1	7/11	28
124*	9/16	9	7	9/16	9	4/5	75
134	9/17	11	8	9/18	12	16/17	12
21*	11/15	1	1	15/17	1	8/11	.7
143*	11/15	2	1	7/11	17	14/15	1
145*	11/16	1	3	12/16	8	5/11	51
138	12/14	1	5	13/15	1	12/14	1
160*	12/14	3	6	10/15	21	12/14	3
152*	12/14	1	1	12/14	1	8/13	16
125*	12/15	1	1	7/12	28	14/15	1
136*	13/15	4	4	13/15	4	8/16	3
143*	14/15	1	1	7/11	17	14/15	1
143*	14/15	1	1	7/11	17	14/15	1
157*	15/16	5	9	6/15	51	16/17	5
157*	16/17	5	3	6/15	51	16/17	3

* The difference in Lp(a) levels in the siblings sharing no apo(a) alleles is at least twice that of the siblings sharing both alleles.

ferences are 1248, 654, and 58 $(mg/dl)^2$ for those sibling pairs that share no, one, and both of their apo(a) alleles ibd, respectively. There was a heteroscedastic distribution of squared Lp(a) differences among the three groups, so weighted regression analysis was performed, as suggested by Amos et al. (36); the results were similar for the weighted and unweighted analyses (data not shown). The linear regression line that best fits these data was equal to $y_j = 1249.2 - 1190.6\pi_j$. These parameter estimates combined with algebraic manipulation of Eqs. 2 and 3 yield estimates of $\sigma_{apo(a)}^2$ and the residual variance component, σ_e^2. For the raw untransformed data, $\sigma_{apo(a)}^2$ and σ_e^2 were equal to 595.3 and 58.6, respectively. As a ratio, these results indicate that 91% of the variation of plasma Lp(a) concentrations among individuals was attributable to genetic variation in the apo(a) gene ($\sigma_{apo(a)}^2/\sigma_{apo(a)}^2 + \sigma_e^2$). For the square-root transformed data these values were 7.00%, 0.86%, and 89%, respectively.

Finally, given the extensive degree of size heterogeneity at the apo(a) locus, it would have been expected that new apo(a) alleles would be encountered if a sufficient number of meioses were analyzed. In this sample, a total of 376 meiosis were examined and a single apo(a) allele was found in an offspring that was not present in either parent (Fig. 6). The fourth child, individual *f*, has *apo(a)16* and *apo(a)9*. Clearly, he inherited *apo(a)16* from his mother, but his father does not have *apo(a)9*. Paternity testing was performed using 7 unlinked varying number of tandem repeat (VNTRs), and in each case, the genotype of individual *f* was consistent with individual f being the child of individual a (39). The calculated probability of individual a not being the true father was $< 1 \times 10^{-6}$ (data not shown). Therefore, a mutation must have occurred in a paternal gamete which resulted in the generation of an apo(a) allele of different size.

Discussion

In this article we have evaluated the segregation of the apo(a) gene and plasma Lp(a) levels in 48 Caucasian families and found that virtually all the inter-individual variation in plasma Lp(a) levels was attributable to the genomic region encoding the apo(a) glycoprotein. It had been clear from previous family studies that plasma Lp(a) levels are largely genetically determined; prior estimates of the heritability of plasma Lp(a) levels have ranged from 0.75 to 0.98 (15, 17, 40, 41) which is comparable to our estimate of 0.85 (±8%). Initially, Lp(a) could only be detected in the plasma of Lp(a) of a third of individuals, and yet when family studies were performed, the inheritance pattern suggested a single autosomal dominant gene (11–13, 42–44). When more sensitive radioimmunoas-

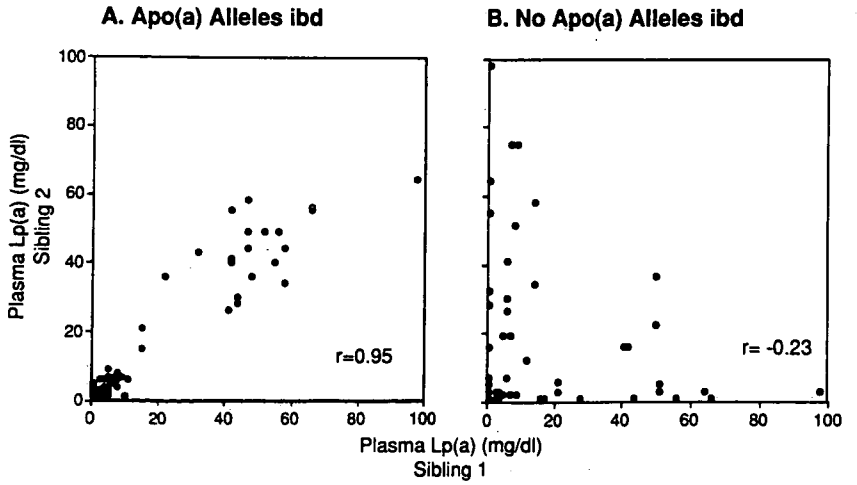


Figure 4. Scatter plot of Lp(a) levels for sibling pairs sharing (*A*) both ($n = 72$) or (*B*) no ($n = 52$) apo(a) alleles identical by descent.

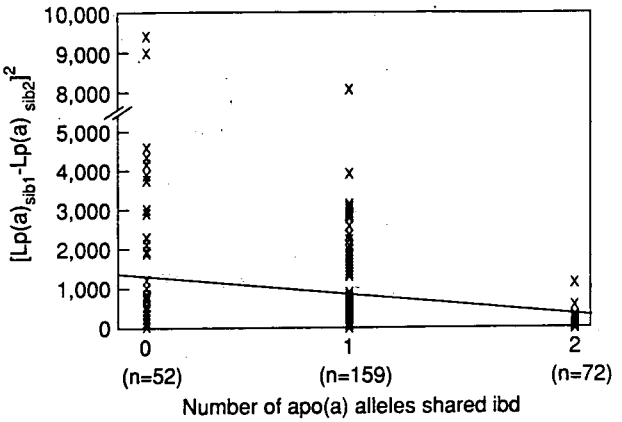


Figure 5. Squared difference between Lp(a) levels of siblings as a function of the proportion of apo(a) alleles shared identical by descent. The regression line for the squared difference on the proportion of apo(a) alleles shared identical by descent is given.

says were employed to measure plasma Lp(a) concentrations in families, there was evidence for a major gene, as well as polygenic factors, contributing to the plasma Lp(a) level (16, 17). In one large Caucasian pedigree, a major gene with three alleles was estimated to account for 73% of the variance in Lp(a) levels (16).

The first molecular clue that the apo(a) gene played a key role in the genetics of plasma Lp(a) concentrations, was the observation that the size of the apo(a) glycoprotein was inversely related to the plasma level of Lp(a) (19). Utermann and his colleagues estimated that differences in the size of the apo(a) glycoprotein accounted for 41% of the variation in inter-individual plasma Lp(a) levels (22). Further support for the apo(a) gene being the major gene influencing Lp(a) levels came from linkage analyses between segregation of plasma

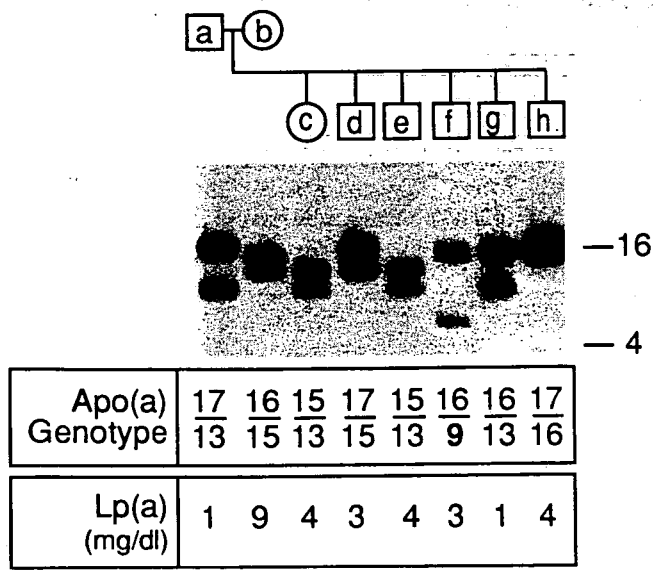


Figure 6. Genomic blot of kringle 4–encoding region of an apo(a) gene in a family in which there is generation of a new apo(a) allele of different length. Southern blotting of the kringle 4-encoding region of the apo(a) gene was performed as described in the Methods and Fig. 2. Individual *f* has inherited *apo(a) 16* from his mother, and *apo(a) 9* from his father. Paternity was confirmed by analysis of seven different VNTR sequences, as described in Methods.

Lp(a) concentrations and DNA sequences in the plasminogen gene which is closely linked to the apo(a) gene (45, 46).

The present study is distinguished from prior family studies by the fact that the apo(a) gene, rather than the expressed protein, was examined in relation to the level of Lp(a) in plasma. In prior studies the immunoblotting techniques used to examine apo(a) isoforms failed to detect protein products from all apo(a) alleles. Numerous exceptions to the inverse relationship between plasma Lp(a) levels and the size of the apo(a) protein were reported. It was suggested that these exceptions were due to the fact that not all apo(a) alleles were detected using the immunoblotting assay. In this study, apo(a) alleles associated with little or no production of apo(a) protein were included in the analysis. As a result, length variation within the kringle 4-encoding region of the apo(a) gene accounted for a greater proportion of the inter-individual variation in plasma Lp(a) concentrations than had been previously estimated (69% vs. 41% [22]).

The reason for the inverse correlation between the size of the apo(a) gene and the level of plasma Lp(a) is not known. Studies in primates have shown that there is not always a consistent relationship between the abundance of apo(a) mRNA, and its size, suggesting that differences in apo(a) gene transcription can not entirely account for this inverse relationship (26, 47). The size of the apo(a) mRNA transcript or glycoprotein may influence its rate of translation, or intracellular transport, respectively.

Alternatively, the observed inverse relationship may not be directly related to the number of kringle 4 repeats in the apo(a) gene, mRNA, or glycoprotein. The apo(a) alleles of different size might be in linkage disequilibrium with the actual sequences at the apo(a) locus that mediate the effect on plasma Lp(a) concentrations. The number of kringle 4 repeats in the apo(a) gene may not have a direct effect on plasma Lp(a) concentrations. In the marmoset monkey, for example, the plasma Lp(a) concentrations vary over a 100-fold range and yet there is only a single apo(a) isoform (48). In the current study, the contribution of the apo(a) gene was partitioned into two components to demonstrate that sequences at the apo(a) locus, other than the number of kringle 4 repeats, contribute importantly to plasma levels of Lp(a). If, however, the number of kringle 4 repeats in the apo(a) gene are in linkage disequilibrium with the actual sequences responsible for mediating the effect of the apo(a) gene or plasma level of Lp(a), then the contribution of the length polymorphism to the inter-individual variation in plasma Lp(a) levels has been overestimated.

Another possible cause for individuals with the same apo(a) genotypes having different plasma concentrations of Lp(a) is that alleles of the same size may differ in the composition of kringle 4 sequences. Not all the kringle sequences are identical. The first, as well as the last eight kringle repeats, differ from the common kringle 4 repeat (the so-called "A" repeat) by between 4 and 35 amino acids (25). Due to the frequent recombinational events involving this locus, is it highly likely that apo(a) alleles vary in their kringle 4 composition, as well as number. Subtle differences in the kringle 4 sequences may impact importantly on the synthesis, and/or degradation of Lp(a).

The length polymorphism in the apo(a) gene has a heterozygosity index comparable to that of number of tandem repeat (VNTR) loci employed in forensic and genetic linkage studies (39, 49, 50). The mutation rate at VNTR loci is several magni-

tudes higher than the usual bi-allelic DNA sequence polymorphisms (39, 51). Given the large number of different sized alleles at the apo(a) locus, a relatively high mutation rate was expected. Therefore, it was anticipated that mutations in the gene would be identified if a sufficient number of families were analyzed. We have observed one mutation of an apo(a) allele out of a total of 376 meioses, and this rate is of the same order of magnitude as the frequency of newly generated alleles for VNTR sequences (39, 51).

Most length polymorphisms in the human genome involve noncoding sequences. The coding regions of several mammalian genes have short tandem repeats (i.e., less the 50 basepairs) which are polymorphic in length (52–56). There are also examples of entire genes being tandemly repeated, as is the case with rDNA, 5S DNA, and the histone genes. The apo(a) length polymorphism is distinguished by the fact that the repeated sequence is large (5.5 kb) and contains both coding and noncoding sequences. The polyubiquitin gene (UbC), contains a large length polymorphism within its coding sequence, but all of the repeated sequences are contained in one exon, and each of the seven to nine repeats encodes the entire protein (57). The heterozygosity index of this length polymorphism is low (22%) compared to the apo(a) gene (94%). The extremely high degree of heterozygosity at the apo(a) locus may reflect the fact that it is under less selective pressure. A physiological function for this enigmatic protein has yet to be identified (2). Alternatively, there may be something intrinsic to the kringle 4-encoding sequences which make them more susceptible to recombinational events.

Mutations of repeated sequence domains result from either intrachromosomal or interchromosomal events. Initially, it was proposed that the mechanism primarily responsible for the high degree of size polymorphism in VNTRs was due to homologous recombination and unequal exchange during meiosis. However, molecular analysis of several new mutations revealed no exchange of flanking genetic markers, which suggests that intrachromosomal, rather than interchromosomal, events appear to be predominantly responsible (51. 58). Similarly, recent molecular analysis of tandem duplication within the Duchenne muscular dystrophy gene demonstrated that the recombinational events were due to intrachromosomal unequal exchange between sister chromatids rather than involving homologous chromosomes (59, 60). Efforts are now being directed to identify polymorphisms flanking the apo(a) gene to analyze the nature of the mutational event(s) responsible for the observed size heterogeneity at the apo(a) locus.

Acknowledgments

The authors thank Drs. Michael Brown and Joseph Goldstein for helpful discussions, Dr. Scott Grundy for assistance with the lipid and lipoprotein measures, and the laboratory of Dr. Steve Daiger at the University of Texas Health Science Center at Houston for markers which aided in the paternity analysis. Tommy Hyatt, Kathy Schueler, and Myriam Fornage provided excellent technical assistance.

This work was supported by grants from the Perot Family Foundation and grants, HL-47619, HL-20948, and HL-40613 from the National Institutes of Health, and 90-IJ-CS-0038 from the National Institute of Justice. E. Boerwinkle is a recipient of a Research Career Development Award and is an Established Investigator of the American Heart Association. H. H. Hobbs is an Established Investigator of the American Heart Association. Giulia Chiesa has a fellowship from the Italian Ministry of Scientific and Technological Research, and Carolin Lackner is a Schroedinger Scholar.

References

1. Berg, K. 1963. A new serum type system in man—the Lp system. *Acta Pathol. Microbiol. Scand.* 59:369–382.

2. Utermann, G. 1989. The mysteries of lipoprotein(a). *Science (Wash. DC).* 246:904–910.

3. Scanu, A. M., and G. M. Fless. 1990. Lipoprotein(a): heterogeneity and biological relevance. *J. Clin. Invest.* 85:1709–1715.

4. Rhoads, G. G., G. H. Dahlèn, K. Berg, N. E. Morton, and A. L. Dannenberg. 1986. Lp(a) lipoprotein as a risk factor for myocardial infarction. *JAMA (J. Am. Med. Assoc.)* 256:2540–2544.

5. Dahlen, G. H., J. R. Guyton, M. Attar, J. A. Farmer, J. A. Kautz, and A. M. Gotto, Jr. 1986. Association of levels of lipoprotein Lp(a), plasma lipids, and other lipoproteins with coronary artery disease documented by angiography. *Circulation.* 74:758–765.

6. Seed, M., F. Hoppichler, D. Reaveley, S. McCarthy, G. R. Thompson, E. Boerwinkle, and G. Utermann. 1990. Relation of serum lipoprotein(a) concentration and apolipoprotein(a) phenotype to coronary heart disease in patients with familial hypercholesterolemia. *N. Engl. J. Med.* 322:1494–1499.

7. Zenker, G., P. Költringer, G. Bonè, K. Niederkorn, K. Pfeiffer, and G. Jürgens. 1986. Lipoprotein(a) as a strong indicator for cerebrovascular disease. *Stroke.* 17:942–945.

8. Miles, L. A., G. M. Fless, E. G. Levin, A. M. Scanu, and E. F. Plow. 1989. A potential basis for the thrombotic risks associated with lipoprotein(a). *Nature (Lond.).* 339:301–303.

9. Edelberg, J. M., and S. V. Pizzo. 1991. Lipoprotein(a): the link between impaired fibrinolysis and atherosclerosis. *Fibrinolysis.* 5:135–143.

10. Albers, J. J., J. L. Adolphson, and W. R. Hazzard. 1977. Radio-immunoassay of human plasma Lp(a) lipoprotein. *J. Lipid Res.* 18:331–338.

11. Berg, K., and J. Mohr. 1963. Genetics of the Lp system. *Acta Genet.* 13:349–360.

12. Sing, C. F., J. S. Schultz, and D. C. Shreffler. 1974. The genetics of the Lp antigen. II. A family study and proposed models of genetic control. *Ann. Hum. Genet.* 38:47–56.

13. Iselius, L., G. H. Dahlèn, U. De Faire, and T. Lundman. 1981. Complex segregation analysis of the Lp(a)/pre-β1-lipoprotein trait. *Clin. Genet.* 20:147–151.

14. Albers, J. J., and W. R. Hazzard. 1974. Immunochemical quantification of human plasma Lp(a) lipoprotein. *Lipids.* 9:15–26.

15. Hasstedt, S. J., D. E. Wilson, C. Q. Edwards, W. N. Cannon, D. Carmelli, and R. R. Williams. 1983. The genetics of quantitative plasma Lp(a): analysis of a large pedigree. *Am. J. Med. Genet.* 16:179–188.

16. Morton, N. E., K. Berg, G. H. Dahlèn, R. E. Ferrell, and G. Rhoads. 1985. Genetics of the Lp Lipoprotein in Japanese-Americans. *Genet. Epidemiol.* 2:113–121.

17. Hasstedt, S. J., and R. R. Williams. 1986. Three alleles for quantitative Lp(a). *Genet. Epidemiol.* 3:53–55.

18. Fless, G. M., M. E. Zum Mallen, and A. M. Scanu. 1986. Physiological properties of apolipoprotein(a) and lipoprotein(a-) derived from the dissociation of human plasma lipoprotein(a). *J. Biol. Chem.* 261:8712–8718.

19. Utermann, G., H. J. Menzel, H. G. Kraft, H. C. Duba, H. G. Kemmler, and C. Seitz. 1987. Lp(a) glycoprotein phenotypes: inheritance and relation to Lp(a)-lipoprotein concentrations in plasma. *J. Clin. Invest.* 80:458–465.

20. Utermann, G., H. G. Kraft, H. J. Menzel, T. Hopferwieser, and C. Seitz. 1988. Genetics of the quantitative Lp(a) lipoprotein trait. I. Relation of Lp(a) glycoprotein phenotypes to Lp(a) lipoprotein concentrations in plasma. *Hum. Genet.* 78:41–46.

21. Utermann, G., C. Duba, and H. J. Menzel. 1988. Genetics of the quantitative Lp(a) lipoprotein trait. II. Inheritance of Lp(a) glycoprotein phenotypes. *Hum. Genet.* 78:47–50.

22. Boerwinkle, E., H. G. Menzel, H. G. Kraft, and G. Utermann. 1989. Genetics of the quantitative Lp(a) lipoprotein trait. III. Contribution of Lp(a) glycoprotein phenotypes to normal lipid variation. *Hum. Genet.* 82:73–78.

23. Sandholzer, C., D. M. Hallman, N. Saha, G. Sigurdsson, C. Lackner, A. Csàszàr, E. Boerwinkle, and G. Utermann. 1991. Effects of the apolipoprotein(a) size polymorphism on the lipoprotein(a) concentration in 7 ethnic groups. *Hum. Genet.* 86:607–614. (Abstr.)

24. Gaubatz, J. W., K. I. Ghanem, J. Guevara, Jr., M. L. Nava, W. Patsch, and J. D. Morrisett. 1990. Polymorphic forms of human apolipoprotein(a): Inheritance and relationship of their molecular weights to plasma levels of lipoprotein(a). *J. Lipid Res.* 31:603–613.

25. McLean, J. W., J. E. Tomlinson, W.-J. Kuang, D. L. Eaton, E. Y. Chen, G. M. Fless, A. M. Scanu, and R. M. Lawn. 1987. cDNA sequence of human apolipoprotein(a) is homologous to plasminogen. *Nature (Lond.).* 330:132–137.

26. Hixson, J. E., M. L. Britten, G. S. Manis, and D. L. Rainwater. 1989.

Apolipoprotein(a) (apo(a)) glycoprotein isoforms result from size differences in apo(a) mRNA in baboons. *J. Biol. Chem.* 264:6013–6016.

27. Gavish, D., N. Azrolan, and J. L. Breslow. 1989. Plasma Lp(a) concentration is inversely correlated with the ratio of kringle IV/kringle V encoding domains in the apo(a) gene. *J. Clin. Invest.* 84:2021–2027.

28. Koschinsky, M. L., U. Beisiegel, D. Henne-Bruns, D. L. Eaton, and R. M. Lawn. 1990. Apolipoprotein(a) size heterogeneity is related to variable number of repeat sequences in its mRNA. *Biochemistry* 29:640–644.

29. Lackner, C., E. Boerwinkle, C. C. Leffert, T. Rahmig, and H. H. Hobbs. 1991. Molecular basis of apolipoprotein(a) isoform size heterogeneity as revealed by pulsed-field gel electrophoresis. *J. Clin. Invest.* 87:2077–2086.

30. Boerwinkle, E., and C. F. Sing. 1987. The use of measured genotype information in the analysis of quantitative phenotypes in man. III. Simultaneous estimation of the frequency and effects of the apolipoprotein E polymorphism and residual polygenic effects on cholesterol, betalipoprotein and triglyceride levels. *Ann. Hum. Genet.* 51:211–226.

31. Church, G. M., and W. Gilbert. 1984. Genomic sequencing. *Proc. Natl. Acad. Sci. USA.* 81:1991–1995.

32. Menzel, H. J., H. Dieplinger, C. Lackner, F. Hoppichler, J. K. Lloyd, D. R. Muller, C. Labeur, P. J. Talmud, and G. Utermann. 1990. Abetalipoproteinemia with an apoB-100-lipoprotein(a) glycoprotein complex in plasma. *J. Biol. Chem.* 265:981–986.

33. Lipid Research Clinic Program. 1982. Lipid and Lipoprotein Analysis: Manual of Laboratory Operations. Department of Health, Education and Welfare Publ NIH/75-628 Government Printing Office, Washington, DC.

34. Hasstedt, S. J., and P. Cartwright. 1981. PAP: Pedigree Analysis Program. Technical Report 13. Department of Medical Biophysics and Computing, University of Utah, Salt Lake City, UT.

35. Haseman, J. K., and R. C. Elston. 1972. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* 2:2–19.

36. Amos, C. I., R. C. Elston, A. F. Wilson, and J. E. Bailey-Wilson. 1989. A more powerful robust sib-pair test of linkage for quantitative traits. *Genet. Epidemiol.* 6:435–449.

37. Scheffe, H. 1959. The Analysis of Variance. John Wiley & Sons, Inc., New York. 221–235.

38. Neter, J., W. Wasserman, and M. H. Kutner. 1990. Applied Statistical Models. R. D. Irwin, Inc., Homewood, IL. 660–661.

39. Jeffreys, A. J., N. J. Royle, V. Wilson, and Z. Wong. 1988. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature (Lond.).* 332:278–281.

40. Albers, J. J., P. Wahi, and W. R. Hazzard. 1974. Quantitative genetic studies of the human plasma Lp(a) lipoprotein. *Biochem. Genet.* 11:475–486.

41. Hewitt, D., J. Milner, C. Breckenridge, and G. Macguire. 1977. Heritability of "sinking" pre-beta lipoprotein level: a twin study. *Clin. Genet.* 11:224–226.

42. Postle, A. D., J. M. Darmady, and D. C. Siggers. 1978. Double pre-β lipoprotein in ischaemic heart disease. *Clin. Genet.* 13:233–236.

43. Schultz, J. S., D. C. Shreffler, and C. F. Sing. 1974. The genetics of the Lp antigen. I. Its quantitation and distribution in a sample population. *Ann. Hum. Genet.* 38:39–46.

44. Hewitt, D., J. Milner, A. R. G. Owen, W. C. Breckenridge, G. F. Macguire, G. J. L. Jones, and J. A. Little. 1982. The inheritance of sinking-pre-beta lipoprotein and its relation to the Lp(a) antigen. *Clin. Genet.* 21:301–308.

45. Drayna, D. T., R. A. Hegele, P. E. Hass, M. Emi, L. L. Wu, D. L. Eaton, R. M. Lawn, R. R. Williams, R. L. White, and J.-M. Lalouel. 1988. Genetic linkage between lipoprotein(a) phenotype and a DNA polymorphism in the plasminogen gene. *Genomics.* 3:230–236.

46. Weitkamp, L. R., S. A. Guttormsen, and J. S. Schultz. 1988. Linkage between the loci for the Lp(a) lipoprotein (LP) and plasminogen (PLG). *Hum. Genet.* 79:80–82.

47. Azrolan, N., D. Gavish, and J. L. Breslow. 1991. Plasma lipoprotein(a) concentration is controlled by apolipoprotein(a) (Apo(a)) protein size and the abundance of hepatic apo(a) mRNA in a cynomolgus monkey model. *J. Biol. Chem.* 266(21):13866–13872.

48. Guo, H.-C., J.-B. Michel, Y. Blouquit, and M. J. Chapman. 1991. Lipoprotein(a) and apolipoprotein(a) in a new world monkey, the common marmoset (callithrix jacchus): Association of variable plasma lipoprotein(a) levels with a single apolipoprotein(a) isoform. *Arterioscler. Thromb.* 11:1030–1041.

49. Weber, J. L. 1990. Informativeness of human $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ polymorphisms. *Genomics.* 7:524–530.

50. Boerwinkle, E., W. Xiong, E. Fourest, and L. Chan. 1989. Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: Application to the apolipoprotein B 3′ hypervariable region. *Proc. Natl. Acad. Sci. USA.* 86:212–216.

51. Jeffreys, A. J., R. Neumann, and V. Wilson. 1990. Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell.* 60:473–485.

52. Oberlè, I., F. Rousseau, D. Heitz, C. Kretz, D. Devys, A. Hanauer, J. Bouè, M. F. Bertheas, and J. L. Mandel. 1991. Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science (Wash. DC).* 252:1097–1102.

53. Azen, E., K. M. Lyons, T. McGonigal, N. L. Barrett, L. S. Clements, N. Maeda, E. F. Vanin, D. M. Carlson, and O. Smithies. 1984. Clones from the human gene complex coding for salivary proline-rich proteins. *Proc. Natl. Acad. Sci. USA.* 81:5561–5565.

54. Swallow, D. M., S. Gendler, B. Griffiths, G. Corney, J. Taylor-Papadimitriou, and M. E. Bramwell. 1987. The human tumour-associated epithelial mucins are coded by an expressed hypervariable gene locus PUM. *Nature (Lond.).* 328:82–84.

55. Boerwinkle, E. S.-H. Chen, S. Visvikis, C.-L. Hanis, G. Siest, and L. Chan. 1991. Signal peptide-length variation in human apolipoprotein B gene. *Diabetes.* 40:1539–1544.

56. McPhaul, M. J., M. Marcelli, W. D. Tiley, J. E. Griffin, R. F. Isidro-Gutierrez, and J. D. Wilson. 1991. Molecular basis of androgen resistance in a family with a qualitative abnormality of the androgen receptor and responsive to high-dose androgen therapy. *J. Clin. Invest.* 87:1413–1421.

57. Baker, R. T., and P. G. Board. 1989. Unequal crossover generates variation in ubiquitin coding unit number at the human UbC polyubiquitin locus. *Am. J. Hum. Genet.* 44:534–542.

58. Wolff, R. K., Y. Nakamura, and R. White. 1988. Molecular characterization of a spontaneously generated new allele at a VNTR locus: No exchange of flanking DNA sequence. *Genomics.* 3:347–351.

59. Hu, X., A. H. M. Burghes, D. E. Bulman, P. N. Ray, and R. G. Worton. 1989. Evidence for mutation by unequal sister chromatid exchange in the Duchenne muscular dystrophy gene. *Am. J. Hum. Genet.* 44:855–863.

60. Hu, X., P. N. Ray, and R. G. Worton. 1991. Mechanisms of tandem duplication in the Duchenne muscular dystrophy gene include both homologous and nonhomologous intrachromosomal recombination. *EMBO (Eur. Mol. Biol. Organ.) J.* 10:2471–2477.

Apolipoprotein(a) Gene Accounts for Greater Than 90% of the Variation in Plasma Lipoprotein(a) Concentrations

Eric Boerwinkle,* Carla C. Leffert,‡ Jingping Lin,* Carolin Lackner,‡ Giulia Chiesa,‡ and Helen H. Hobbs‡
**Center for Demographic and Populations Genetics, University of Texas Health Science Center in Houston, Houston, Texas 77225; and ‡Departments of Internal Medicine and Molecular Genetics, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75235*

Abstract

Plasma lipoprotein(a) [Lp(a)], a low density lipoprotein particle with an attached apolipoprotein(a) [apo(a)], varies widely in concentration between individuals. These concentration differences are heritable and inversely related to the number of kringle 4 repeats in the apo(a) gene. To define the genetic determinants of plasma Lp(a) levels, plasma Lp(a) concentrations and apo(a) genotypes were examined in 48 nuclear Caucasian families. Apo(a) genotypes were determined using a newly developed pulsed-field gel electrophoresis method which distinguished 19 different genotypes at the apo(a) locus. The apo(a) gene itself was found to account for virtually all the genetic variability in plasma Lp(a) levels. This conclusion was reached by analyzing plasma Lp(a) levels in siblings who shared zero, one, or two apo(a) genes that were identical by descent (ibd). Siblings with both apo(a) alleles ibd (n = 72) have strikingly similar plasma Lp(a) levels (r = 0.95), whereas those who shared no apo(a) alleles (n = 52), had dissimilar concentrations (r = −0.23). The apo(a) gene was estimated to be responsible for 91% of the variance of plasma Lp(a) concentration. The number of kringle 4 repeats in the apo(a) gene accounted for 69% of the variation, and yet to be defined *cis*-acting sequences at the apo(a) locus accounted for the remaining 22% of the inter-individual variation in plasma Lp(a) levels. During the course of these studies we observed the de novo generation of a new apo(a) allele, an event that occurred once in 376 meioses. (*J. Clin. Invest.* 1992. 90:52–60.) Key words: apolipoprotein (a) • lipoprotein (a) • low density lipoprotein

Introduction

Lipoprotein(a) [Lp(a)][1] is a cholesterol ester-rich plasma lipoprotein comprising two attached components: a low density lipoprotein (LDL) particle and a single large glycoprotein, apolipoprotein(a) [apo(a)] (1–3). High plasma levels of Lp(a) are associated with the development of coronary atherosclerosis (4–6) and other vascular diseases (7). The mechanism by which Lp(a) expedites the atherosclerotic process is not known. Apo(a) strongly resembles plasminogen, and it may competitively interfere with plasminogen action in fibrinolysis (8, 9).

Plasma concentrations of Lp(a) vary over a wide range among individuals, but are remarkably stable in any given individual (10). Many physiological, pharmacological, and environmental factors that affect the levels of other plasma lipoproteins have no effect on the plasma concentration of Lp(a) (10). This lack of environmental and physiological influences suggests that plasma Lp(a) levels are largely genetically determined. Consistent with this formulation, early genetic studies suggested that the presence of Lp(a) in plasma was inherited as an autosomal dominant trait (11–13). When more sensitive immunoassays of plasma Lp(a) concentrations were used, it was found that plasma Lp(a) concentrations varied continuously among individuals (14), and the pattern of inheritance indicated that a major gene, as well as polygenic factors, contributed to plasma Lp(a) concentrations (15–17).

Fless et al. (18) and Utermann et al. (19) found that the apo(a) glycoprotein varied in size among individuals. In an important series of studies, Utermann and his colleagues demonstrated that the size of the apo(a) protein is inversely related to the level of plasma Lp(a), thus implicating the apo(a) gene as a major determinant of plasma Lp(a) concentrations (20–23). However, the immunoblotting technique used to type the apo(a) isoforms was not sensitive enough to detect low levels of apo(a) protein, and not all of the apo(a) isoforms were detected. As a result, the frequency distribution of the apo(a) isoforms failed to fit the expectations of Hardy-Weinberg equilibrium (22). In addition, when immunoblotting was employed to examine the segregation of the apo(a) isoforms in families, the results were frequently uninformative, and occasionally inconsistent (24). Further progress required the development of a technique that was more discriminating than immunoblotting in classifying apo(a) alleles.

A potential method to study this polymorphism was suggested by the findings of McLean et al. who discovered that the apo(a) cDNA contains multiple tandem copies of a sequence that encodes a cysteine-rich protein motif called a kringle. The repeated kringle in apo(a) is designated kringle 4 because it closely resembles the fourth kringle in plasminogen. McLean et al. proposed that the apo(a) isoforms are of different size because of variations in the numbers of kringle 4-encoding repeats in the apo(a) gene (25). This hypothesis was supported by studies of the apo(a) mRNA and gene structure (26–28). In attempt to devise a way to measure the size of the apo(a) gene in different individuals, we previously identified a large restriction fragment from the apo(a) gene which contains most, if not all, of the kringle 4–encoding sequences (29). The size of this fragment was too large to be examined by standard electrophoresis techniques. Accordingly, we used pulsed-field gel electrophoresis to size this large restriction fragment and 19 fragments of different length were identified. A total of 103 unrelated

Address reprint requests to Dr. Hobbs, Department of Molecular Genetics, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd., Dallas, TX 75235.

Received for publication 6 December 1991.

1. *Abbreviation used in this paper:* Lp(a), lipoprotein (a).

J. Clin. Invest.

0021-9738/92/07/0052/09 $2.00
Volume 90, July 1992, 52–60

Caucasians were evaluated and 94% were heterozygous for fragments of two different sizes. The length polymorphism was used as a genetic marker to analyze the segregation of the apo(a) gene in 12 Caucasian families. It was found that within a given family, sibling pairs with identical apo(a) genotypes tended to have very similar plasma Lp(a) levels (29). However, individuals with the same apo(a) genotypes who were members of different families often had significantly different plasma concentrations of Lp(a). Taken together, these observations suggest that the apo(a) gene is the major determinant of plasma Lp(a) levels and that cis-acting DNA sequences at or near the apo(a) locus, other than the number of kringle 4 repeats, contribute importantly to plasma Lp(a) concentrations.

In the current study, we analyzed the segregation of the apo(a) gene and Lp(a) levels in 48 Caucasian pedigrees to determine the contribution of the apo(a) gene (or closely linked loci) to the plasma concentrations of Lp(a). The genetic architecture (30) of plasma Lp(a) concentrations was defined at three levels: the polygenic heritability, the total genetic contribution of the apo(a) gene, and the effects of length variation in the apo(a) gene. In addition, we describe the de novo generation of a new apo(a) allele of different size within a family.

Methods

Subjects. Plasma Lp(a) concentrations were measured in a sample of 288 fasting individuals from 48 Caucasian American families living in the greater Dallas, Texas area. Families in which both parents and at least three children were available for sampling were selected for study. None of the families had evidence of a monogenic hyperlipidemia. In one family, F153, several members (who are denoted in Table II) had very low LDL-cholesterol levels, suggesting the possible existence of familial hypobetalipoproteinemia. For the random effects analysis of variance (see below), the families were augmented with a sample of 107 unrelated individuals. Preliminary findings on a subset of these unrelated individuals have been reported previously (29).

Phlebotomy was carried out after an overnight fast. A total of 30 ml of blood was collected from each individual in vacutainer tubes containing sodium-EDTA. The plasma was separated within one hour of collection by centrifugation at 2,000 *g* for 15 min at 4°C. Multiple 50 μl aliquots of plasma were stored at −70°C and Lp(a) levels were assayed within 4 wks.

Pulsed-field gel analysis of the apo(a) gene. A total of 15 ml of blood was maintained at room temperature prior to transfer to two LeucoPREP tubes (Becton, Dickinson & Co., Lincoln Park, NJ). Lymphocytes were isolated and embedded in agarose plugs as previously described (29). The agarose-cellular plugs were incubated twice with 40 U of KpnI in 170 μl of the buffer suggested by the manufacturer (New England Biolabs, Beverly, MA). The digested cellular-agarose plugs were subjected to pulsed-field gel electrophoresis in a vertical submarine gel apparatus with a transverse alternating field (Geneline I, Beckman Instruments, Inc., Fullerton, CA) using low-endosmosis coefficient agarose, TAFE buffer, and λ phage concatamer standards (Beckman Instruments, Inc.) as described by Lackner et al. (29). After 18 hours of electrophoresis, the gel was stained with ethidium bromide and photographed. The DNA was transferred and fixed to nylon membrane (Biotrans, ICN Biomedicals, Costa Mesa, CA). MP-1, a 1.5-kb PstI genomic fragment from the kringle 4-encoding region of the apo(a) gene (29) which had been subcloned into M13mp18, was used to generate a ^{32}P-radiolabeled single-stranded probe (31). The filter was incubated overnight at 42°C in hybridization solution containing 5×10^6 cpm/ml of the single-stranded apo(a)-specific probe. Hybridizations were carried out in a rotating incubator (model 310, Robbins Scientific Corp., Sunnyvale, CA). Filters were washed as described by Lackner et al. (29) and exposed to film.

Immunoblotting of plasma apo(a). An aliquot of frozen plasma (between 1 and 30 μl) containing 1 μg of Lp(a) was brought up to a total volume of 30 μl using phosphate-buffered saline. The sample was mixed with 20 μl of buffer A which contained 15% filtered SDS (wt/vol), 8 M urea, 5 mM dithiothreitol, and 62.5 mM Tris at pH 7.5 and with 50 μl buffer B (10% glycerol [vol/vol], 2.3% SDS [wt/vol], 0.025% bromophenol blue [wt/vol], 5% β-mercaptoethanol [vol/vol], and 50.0 mM Tris at pH 6.8). The samples were boiled for 10 min before loading onto a 3–7% gradient polyacrylamide gel with SDS. A total of 1 μg of purified LDL (molecular weight of apo B is ~ 513 kD) was used as a size standard. The electrophoresis, transfer to nitrocellulose, and hybridization conditions were exactly as preciously described except that IgG-1A^2, the apo(a)-specific antibody, was radiolabeled directly with ^{125}I to a specific activity of 5×10^6 cpm/ml (29). The filters were washed, dried and exposed to XAR-5 film (Eastman Kodak Co., Rochester, NY) at −70°C with an intensifying (Lightening Plus, Dupont Co., Wilmington, DE).

Plasma lipid and lipoprotein assays. Measurement of plasma Lp(a) concentrations were performed at GeneScreen, Dallas, TX, using a sensitive enzyme-linked immunosorbent sandwich assay (ELISA), as described (32). In this assay, Lp(a) was captured by a polyclonal rabbit anti-human Lp(a) antibody and then detected by a monoclonal anti-human Lp(a) antibody, IgG-1A^2. Plasma Lp(a) standards were obtained from Immuno, Vienna, Austria. Total cholesterol and triglyceride levels were measured enzymatically using commercially available kits (Boehringer Mannheim, Indianapolis, IN; Sigma Chemical Co., St. Louis, MO). Plasma lipoproteins were quantified in the laboratory of Dr. Scott Grundy (University of Texas Southwestern Medical Center) according to the procedures of the Lipid Research Clinic (33).

Statistical methods. The distribution of plasma Lp(a) concentration was positively skewed in these data, and thus all analyses were carried out both on the raw and square-root transformed data. For each analysis, the primary inferences were identical whether the raw or transformed data was used.

The contribution of unmeasured polygenic variation to the inter-individual variability of plasma Lp(a) concentrations ($\sigma_{Lp(a)}^2$) was assessed from the extent of familial aggregation of Lp(a) levels in the sample of pedigrees. The ratio of the polygenic variance component (σ_{pg}^2) to $\sigma_{Lp(a)}^2$ was estimated by maximum likelihood principles as implemented in the computer program PAP V3.0 (34).

Sibling-pair linkage methods were used to estimate the overall contribution of genetic variation in and around the apo(a) gene ($\sigma_{apo(a)}^2$) to plasma Lp(a) levels (35, 36). These methods are most frequently employed to detect linkage between a marker and a quantitative trait locus, but can also be used to define the overall contribution of a candidate gene to a quantitative phenotype. For each sibling pair, three new variables were considered: y_j, the squared difference of plasma Lp(a) concentrations in sibship j, f_{1j}, an indicator variable describing whether or not the jth sib pair shares only 1 allele identical by descent (ibd), and π_j, the proportion of alleles ibd in sibship j. π_j can take on the values 0, ½, or 1. $E(y_j)$ is the expected value of an individual's Lp(a) concentration. Assuming there is no recombination as would be the case for a candidate gene, Haseman and Elston (30) show that:

$$E(y_j) = \alpha + \beta\pi_j + \gamma f_{1j} \quad (1)$$

where

$$\alpha = 2\sigma_{apo(a)}^2 + \sigma_e^2 \quad (2)$$

$$\beta = -2\sigma_{apo(a)}^2 \quad (3)$$

$$\gamma = -\sigma_d^2 \quad (4)$$

In Eqs. 2–4, σ_e^2 is a residual variance component describing the effects of factors other than the apo(a) gene on Lp(a) levels, and σ_d^2 describes the dominance effects at the apo(a) locus on Lp(a) levels. An estimate of the overall contribution of the apo(a) gene to plasma Lp(a) concentrations can be made by examining the regression of the squared differ-

Table 1. Correlations of Plasma Lp(a) Concentrations between Family Members

	n	Lp(a)	√Lp(a)
Spouses	48	0.17 [−0.12, 0.43]*	0.17 [−0.12, 0.43]
Parent-offspring	400	0.44§ [0.36, 0.52]	0.48§ [0.40, 0.55]
Midparent-offspring	200	0.59§ [0.49, 0.67]	0.61§ [0.51, 0.69]
Siblings (all)	284	0.28§ [0.16, 0.39]	0.32§ [0.21, 0.43]
Siblings sharing no alleles ibd	52	−0.23 [−0.47, 0.05]	−0.25 [−0.48, 0.02]
Siblings sharing one allele ibd	159	0.15 [−0.16, 0.30]	0.19‡ [0.04, 0.34]
Siblings sharing two alleles ibd	73	0.95§ [0.92, 0.97]	0.96§ [0.94, 0.97]

* 95% confidence interval. ‡ $P < 0.05$. § $P < 0.001$.

ence between the Lp(a) levels of siblings who share none, one, or all apo(a) alleles ibd. The regression analyses were performed both unweighted and weighted, as suggested by Amos et al. (36), with nearly identical results. Therefore, only the results of the unweighted analyses are presented. Even though the sibships were typically larger then size two, the above method has been shown to be valid when overlapping sibling pairs are analyzed as though they were independent (36).

The contribution of length variation in the apo(a) gene, as measured by pulsed-field gel electrophoresis, to Lp(a) concentrations, σ_{length}^2, was estimated using a random effects analysis of variance (37). A random effects or type II model was selected because of the large number of potential genotypes at the apo(a) locus (38).

Results

Plasma Lp(a) concentrations were measured in 288 individuals from 48 pedigrees. There was no significant effect of age, sex, or the concentration of other plasma lipoproteins on the plasma level of Lp(a), so these factors were not considered further in the family members (data not shown). There were significant correlations between the plasma Lp(a) levels of parents and offspring ($r = 0.44$), and siblings ($r = 0.28$), but not between spouses ($r = 0.17$) (Table 1). By using standard biometrical genetic analyses, it was estimated that 85% (±8%) of the inter-individual variance of Lp(a) concentrations was attributable to polygenic effects ($\sigma_{pg}^2/\sigma_{Lp(a)}^2$) (or 88% (±6.5) when the square-root of the plasma Lp(a) levels was used).

Pulsed-field gel electrophoresis and genomic blotting of KpnI digested-genomic DNA was performed to assess the size of the kringle 4–encoding region of the apo(a) alleles in each family member. 16 of the 19 previously described apo(a) alleles were observed in the sample, and their frequencies did not differ significantly from those previously described from the same population (29). In general, there was an inverse relationship between the size of the apo(a) allele and the plasma level of Lp(a). One way to illustrate this phenomenon is to examine the relationship between the plasma Lp(a) concentrations and the apo(a) allele size in the individuals who had one of the two most common alleles, *apo(a)14* or *apo(a)15* plus a different allele (Fig. 1). Individuals with one copy of *apo(a)14* or *apo(a)15* plus one copy of *apo(a)2-apo(a)4* tended to have high plasma Lp(a) levels (> 30 mg/dl). If the second allele was *apo(a)5-apo(a)7*, the Lp(a) levels were lower (15–30 mg/dl). If the second allele was larger than *apo(a)8*, the plasma concentrations of Lp(a) were low [< 10 mg/dl, excluding *apo(a)10*].

In the population as a whole, different apo(a) genotypes, as determined by pulsed-field gel electrophoresis, were associated with significantly different plasma levels of Lp(a) ($P < 0.001$ for both the raw and transformed data). A random effects analysis of variance (37) was used to determine the contribution of the length variation in the apo(a) gene to the distribution of plasma Lp(a) in 203 unrelated Caucasians. For the raw data, 69% of the variation in Lp(a) concentrations was attributable to inter-individual differences in the number of kringle-4 repeats. The square-root transformation had little effect on this value (66% vs. 69%).

Although length variation in the apo(a) gene had a profound influence on Lp(a) concentrations, there were several exceptions to the general trend. Fig. 2 shows two pedigrees in which an apo(a) allele of the same size, *apo(a)6*, segregates. In the two pedigrees this allele gives rise to very different plasma concentrations of Lp(a). In *A*, the *apo(a)6* allele of the father (*a*), is inherited by three of his offspring (*c*, *d*, and *f*). The father, as well as the three offspring, have modest plasma Lp(a) concentrations (6 mg/dl, and 7, 5, and 3 mg/dl, respectively). In the family shown in *B*, individual h, who is also heterozygous for an allele the size of *apo(a)6*, and has a high plasma Lp(a) concentration (51 mg/dl). Of her four children, only the second child (*j*) inherited *apo(a)6* and she is the only offspring with a comparable plasma level of Lp(a) (51 mg/dl). Therefore, in these two families, the same sized apo(a) allele (*apo(a)6*) segregated with very different plasma levels of Lp(a). This was true even though the other alleles at the apo(a) locus in the families were similar (*apo(a)13-apo(a)17*). These findings suggest that factors at the apo(a)

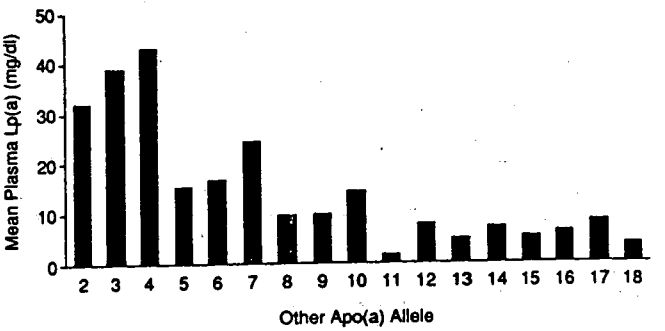


Figure 1. Lp(a) levels in individuals heterozygous for apo(a)14 or apo(a)15 allele. In the sample of 288 family members and 107 unrelated individuals, there were 194 individuals with either *apo(a)14* or *apo(a)15*. The average Lp(a) levels (*y*-axis) for individuals with each genotype are plotted against the other apo(a) allele (*x*-axis).

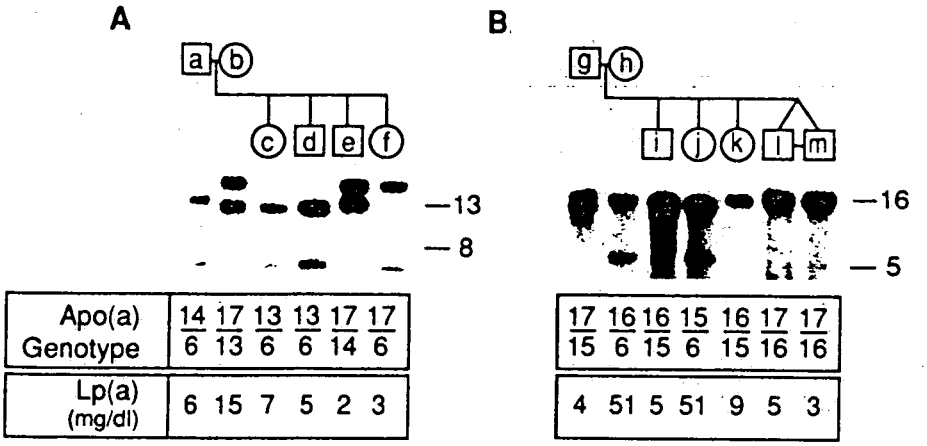


Figure 2. Genomic blot of the apo(a) gene from two unrelated families with *apo(a)6*. High molecular weight leukocyte DNA from members of two unrelated families was digested with KpnI, size-fractionated on a pulsed-field gel, transferred to a nylon membrane, and hybridized with a single-stranded apo(a)-specific probe (MP-1) as described in the Methods. The filter was exposed to Kodak XAR-5 film for 18 h with an intensifying screen. The plasma concentrations of Lp(a) were measured using an ELISA assay as described in the Methods. The *apo(a)6* allele segregates with a low (*A*) and high (*B*) plasma concentration of Lp(a) in two different pedigrees.

locus, in addition to the number of kringle 4 repeats, strongly influence the plasma Lp(a) concentration.

Another instance in which apo(a) alleles of the same size are associated with different amounts of circulating apo(a) protein, is shown in Fig. 3. In this family, the mother (b) is homozygous for *apo(a)12*, so all of the children (*c–f*) are heterozygous for that allele. Based on the genomic blot, it cannot be determined which of the two *apo(a)12* alleles each child inherited from their mother. However, analysis of the apo(a) protein isoforms reveals that three of the offspring (*c*, *d*, and *f*) have no detectable apo(a) protein corresponding to *apo(a)12*. Only offspring *e* has a band the same size as the isoform of the mother. This suggests that the mother is heterozygous and has one *apo(a)12* that produces no detectable circulating apo(a) protein which she gave to *c*, *d*, and *f* and another that is associated with the production of a moderate amount of apo(a) protein which she donated to offspring *e*.

To confirm that cis-acting sequences at the apo(a) locus are responsible for the observed differences in plasma Lp(a) concentrations in individuals with apo(a) alleles of the same size, the plasma Lp(a) concentrations were compared in sibling pairs who shared all, one, or no apo(a) alleles ibd. In 40 of the 48 families, all four parental apo(a) alleles could be differentiated using pulsed-field gel electrophoresis. In six families (including the one shown in Fig. 3), the length polymorphism was uninformative because one of the parents was apparently homozygous for the same sized apo(a) allele, and in two families one parent was not available for sampling; these eight families were not included in the sibling-pair analysis. The families were selected at random, so in some families all the plasma concentrations of Lp(a) were low (i.e., < 5 mg/dl) reflecting the highly skewed distribution of plasma Lp(a) levels in the Caucasian population.

In the 40 families in which the segregation of each parental allele could be distinguished, 72 sibling pairs shared both, 52 shared none, and 159 pairs shared one parental allele ibd. The apo(a) genotypes and plasma Lp(a) levels of these sibling pairs are given in Table II and can be compared to those of the other siblings and parents. The sibling pairs who had plasma levels of Lp(a) that were similar to each other and significantly different from the other siblings are denoted by an asterisk. In 24 families, at least one sibling pair had inherited identical apo(a) alleles and one sibling pair had no apo(a) alleles ibd (Table III). In 21 of the 24 sibling pairs (denoted by asterisks), the mean difference between plasma Lp(a) levels in the sibling pair who shared no apo(a) alleles ibd was twice that of the sibling pair who shared both apo(a) alleles ibd.

Fig. 4 shows the scatter plot of Lp(a) levels of siblings pairs that share (*A*) both or (*B*) no apo(a) alleles ibd. Lp(a) levels for the older (sibling 1) and younger (sibling 2) sibling are plotted on the horizontal and vertical axis, respectively. The correlation coefficient for Lp(a) levels between siblings who share both apo(a) alleles ibd was very high ($r = 0.95$), whereas there was a negative correlation ($r = -0.23$) between the Lp(a) concentration of siblings who share no apo(a) alleles ibd. Similar results were obtained for the square-root transformed data (Table I). Owing to the highly skewed distribution of Lp(a) in the population, many of the sibling pairs had very low Lp(a) levels. Therefore, the same comparison was made in the sibling pairs with apo(a) alleles ibd ($n = 48$) who had plasma Lp(a) levels over 5 mg/dl and the correlation coefficient remained very high ($r = 0.94$).

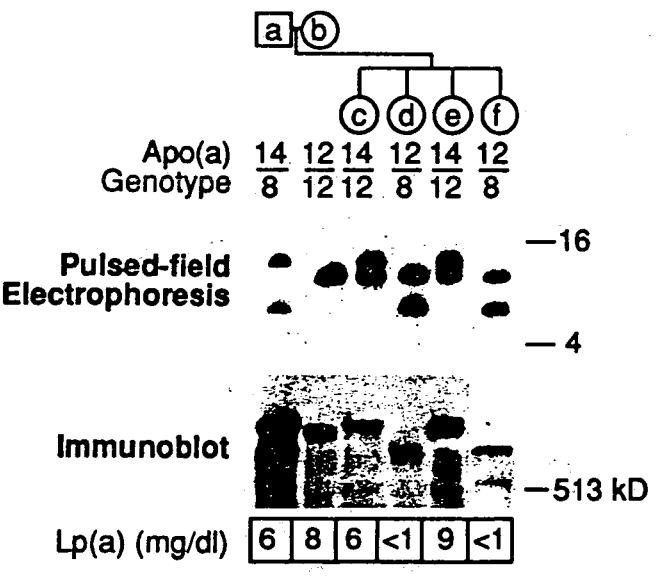


Figure 3. Genomic blot of apo(a) gene and immunoblot of apo(a) protein in a pedigree. The plasma Lp(a) concentrations were measured and the genomic blot and immunoblot was performed as described in Fig. 2 and in the Methods.

Table II. Apo(a) Genotypes and Lp(a) Levels in Sibling Pairs with Apo(a) Alleles ibd and Their Family Members

		Sib with identical apo(a) genotypes				Other Sibs		Parents (father, mother)	
	Family no.	Apo(a) genotype	Plasma Lp(a)			Apo(a) genotypes	Plasma Lp(a)	Apo(a) genotype	Lp(a)
			mg/dl				*mg/dl*		*mg/dl*
1.	F141	2/13	26		41	13/16, 15/16, 2/15	1, 6, 32	2/16, 13/15	36, 6
2.	F168	2/17	*40	*42	*55	12/17, 12/17	< 1, < 1	6/17, 2/12	40, 54
3.	F162	4/5	*44	*47	*58	5/14	21	5/15, 4/14	38, 19
4.	F24	4/10	*41		*42	10/14, 6/14	< 1, 16	4/14, 6/10	31, 62
5.	F154	4/10	58		34	12/15, 4/15, 4/15	14, 32, 43	10/15, 4/12	32, 52
6.	F161	4/13	*49		*52	13/13	10	4/13, 13/14	50, 9
7.	F158	4/14	*36		*48	8/14	10	4/8, 14/15	48, 7
8.	F156	4/15	*47		*49	10/15	7	14/15, 4/10	4, 42
9.	F135	4/15	*55	*56	*66	14/15, 14/15	3, 3	14/15, 4/14	4, 45
10.	F154	4/15	32		43	12/15, 4/10, 4/10	14, 34, 58	10/15, 4/12	32, 52
11.	F142	5/9	*49		*56	9/16	4	5/16, 9/10	72, 21
12.	F145	5/12	64		98	11/16, 11/16, 12/16, 5/11	< 1, 3, 8, 51	5/16, 11/12	47, 5
13.	F129	6/6	7		7	6/16, 6/16, 15/16	5, 6, 6	6/15, 6/16	4, 15
14.	F149	6/13	5		7	14/17, 6/17	2, 3	6/14, 13/17	6, 15
15.	F129	6/16	5		6	15/16, 6/6, 6/6	6, 7, 7	6/15, 6/16	4, 15
16.	F146	7/11	*28		*44	9/17, 9/11, 9/11	< 1, 5, 6	7/9, 11/17	22, 9
17.	F166	8/12	5		7	12/13, 13/15	17, 19	8/13, 12/15	9, 10
18.	F150	8/15	4	6	8	7/15	5	15/16, 7/8	5, 5
19.	F146	9/11	5		6	9/17, 7/11, 7/11	< 1, 28, 44	7/9, 11/17	22, 9
20.	F124	9/16	*7		*9	4/16, 5/9, 4/5	22, 50, 75	5/16, 4/9	54, 27
21.	F134	9/17	6	8	11	16/17, 9/18	12, 12	9/16, 17/18	15, 3
22.	F137	10/13	< 1		3	13/14, 13/14, 13/14	< 1, 1, 1	10/14, 13/15	< 1, 2
23.	F21	11/15	< 1		< 1	15/17, 8/11	< 1, 7	8/15, 11/17	5, < 1
24.	F143	11/15	1		2	14/15, 14/15, 14/15, 7/11	< 1, < 1, < 1, 17	7/15, 11/14	28, < 1
25.	F164	11/15	1	1	1	15/16	3	11/16, 8/15	2, 4
26.	F145	11/16	*< 1		*3	12/16, 5/11, 5/12, 5/12	8, 51, 64, 98	5/16, 11/12	47, 5
27.	F167	12/13	*15	*15	*21	12/17, 13/13	1, 5	13/17, 12/13	12, 1
28.	F138	12/14	1		5	13/15	1	14/15, 12/13	4, 2
29.	F152	12/14	*< 1		*< 1	8/13	16	12/13, 8/14	< 1, 30
30.	F160	12/14	*3		*6	10/15, 10/14	21, 34	10/12, 14/15	27, 10
31.	F125	12/15	1		1	14/15, 7/12	< 1, 28	12/14, 7/15	< 1, 32
32.	F168	12/17	*< 1		*< 1	2/17, 2/17, 2/17	40, 42, 55	6/17, 2/12	40, 54
33.	F126	12/18	< 1		< 1	15/18	< 1	12/15, 9/18	3, < 1
34.	F137	13/14	1	1	< 1	10/13, 10/13	< 1, 3	10/14, 13/15	< 1, 2
35.	F136	13/15	4		4	16/17, 15/17, 8/16	< 1, 3, 3	13/17, 15/16	1, 9
36.	F135	14/15	*3		*3	4/15, 4/15, 4/15	56, 56, 66	14/15, 4/14	4, 45
37.	F143	14/15	< 1	< 1	< 1	11/15, 11/15, 7/11	1, 2, 17	7/15, 11/14	28, < 1
38.	F131	14/15	< 1	< 1	4	15/18	< 1	14/18, 5/15	< 1, < 1
39.	F132	14/17	1	1	3	14/15, 15/17	2, 3	14/17, 14/15	8, < 1
40.	F165	14/18	8		8	15/18, 15/18	< 1, 9	14/15, 10/18	16, < 1
41.	F157	15/16	5		9	16/17, 16/17, 6/15	3, 5, 51	15/17, 6/16	4, 51
42.	F159	15/17	1	5	2	10/17	< 1	16/17, 10/15	< 1, 6
43.	F153	16/16	*1‡	*4‡	*5‡	7/16	27	7/16, 12/16	52, 1
44.	F157	16/17	3		5	15/16, 15/16, 6/15	5, 9, 51	15/17, 6/16	4, 51
45.	F165	15/18	< 1		9	14/18, 14/18	8, 8	14/15, 10/18	16, 1

* Sibling pairs with identical apo(a) genotypes who have Lp(a) levels which are significantly different from all the other siblings. ‡ These individuals have a plasma LDL-cholesterol concentration less than the 5th percentile when compared to age and sex-matched controls.

The overall contribution of the apo(a) gene to plasma Lp(a) concentrations was estimated by examining the regression of the squared difference of Lp(a) levels between siblings (y_j) based on the proportion of apo(a) alleles shared ibd (π_j). The dominance deviations at the apo(a) locus (γ) was not significantly different from zero ($\gamma = 1.27$ for untransformed Lp(a) levels) so was not considered in further analyses. The simple linear regression of the squared difference of Lp(a) levels between siblings on the proportion of apo(a) alleles shared ibd is graphically presented in Fig. 5. The average squared dif-

Table III. Lp(a) Levels in Sibling Pairs in the Same Family Who Share Both or No Apo(a) Alleles Identical by Descent

	Siblings sharing both apo(a) alleles			Siblings sharing no apo(a) alleles			
		Lp(a)		Genotype Lp(a)			
Family	Genotype	Sib1	Sib2	Sib1		Sib2	
141*	2/13	41	26	13/16	1	2/15	32
24*	4/10	42	41	4/10	42	6/14	16
154	4/10	58	34	12/15	14	4/10	34
154*	4/15	32	43	12/15	14	4/10	58
124*	4/16	22	36	5/9	50	4/16	22
145*	5/12	98	64	11/16	1	5/12	98
149*	6/13	7	5	6/13	7	14/17	2
146*	7/11	44	28	7/11	44	9/17	1
166*	8/12	5	7	8/12	5	13/15	19
146*	9/11	5	6	9/17	1	7/11	28
124*	9/16	9	7	9/16	9	4/5	75
134	9/17	11	8	9/18	12	16/17	12
21*	11/15	1	1	15/17	1	8/11	7
143*	11/15	2	1	7/11	17	14/15	1
145*	11/16	1	3	12/16	8	5/11	51
138	12/14	1	5	13/15	1	12/14	1
160*	12/14	3	6	10/15	21	12/14	3
152*	12/14	1	1	12/14	1	8/13	16
125*	12/15	1	1	7/12	28	14/15	1
136*	13/15	4	4	13/15	4	8/16	3
143*	14/15	1	1	7/11	17	14/15	1
143*	14/15	1	1	7/11	17	14/15	1
157*	15/16	5	9	6/15	51	16/17	5
157*	16/17	5	3	6/15	51	16/17	3

* The difference in Lp(a) levels in the siblings sharing no apo(a) alleles is at least twice that of the siblings sharing both alleles.

ferences are 1248, 654, and 58 $(mg/dl)^2$ for those sibling pairs that share no, one, and both of their apo(a) alleles ibd, respectively. There was a heteroscedastic distribution of squared Lp(a) differences among the three groups, so weighted regression analysis was performed, as suggested by Amos et al. (36); the results were similar for the weighted and unweighted analyses (data not shown). The linear regression line that best fits these data was equal to $y_j = 1249.2 - 1190.6\pi_j$. These parameter estimates combined with algebraic manipulation of Eqs. 2 and 3 yield estimates of $\sigma_{apo(a)}^2$ and the residual variance component, σ_e^2. For the raw untransformed data, $\sigma_{apo(a)}^2$ and σ_e^2 were equal to 595.3 and 58.6, respectively. As a ratio, these results indicate that 91% of the variation of plasma Lp(a) concentrations among individuals was attributable to genetic variation in the apo(a) gene ($\sigma_{apo(a)}^2/\sigma_{apo(a)}^2 + \sigma_e^2$). For the square-root transformed data these values were 7.00%, 0.86%, and 89%, respectively.

Finally, given the extensive degree of size heterogeneity at the apo(a) locus, it would have been expected that new apo(a) alleles would be encountered if a sufficient number of meioses were analyzed. In this sample, a total of 376 meiosis were examined and a single apo(a) allele was found in an offspring that was not present in either parent (Fig. 6). The fourth child, individual *f*, has *apo(a)16* and *apo(a)9*. Clearly, he inherited *apo(a)16* from his mother, but his father does not have *apo(a)9*. Paternity testing was performed using 7 unlinked varying number of tandem repeat (VNTRs), and in each case, the genotype of individual *f* was consistent with individual f being the child of individual a (39). The calculated probability of individual a not being the true father was $< 1 \times 10^{-6}$ (data not shown). Therefore, a mutation must have occurred in a paternal gamete which resulted in the generation of an apo(a) allele of different size.

Discussion

In this article we have evaluated the segregation of the apo(a) gene and plasma Lp(a) levels in 48 Caucasian families and found that virtually all the inter-individual variation in plasma Lp(a) levels was attributable to the genomic region encoding the apo(a) glycoprotein. It had been clear from previous family studies that plasma Lp(a) levels are largely genetically determined: prior estimates of the heritability of plasma Lp(a) levels have ranged from 0.75 to 0.98 (15, 17, 40, 41) which is comparable to our estimate of 0.85 (±8%). Initially, Lp(a) could only be detected in the plasma of Lp(a) of a third of individuals, and yet when family studies were performed, the inheritance pattern suggested a single autosomal dominant gene (11–13, 42–44). When more sensitive radioimmunoas-

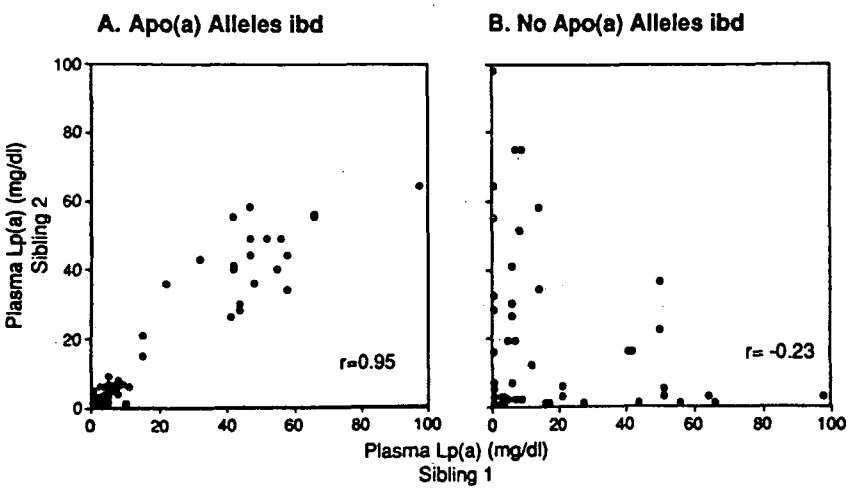


Figure 4. Scatter plot of Lp(a) levels for sibling pairs sharing (*A*) both ($n = 72$) or (*B*) no ($n = 52$) apo(a) alleles identical by descent.

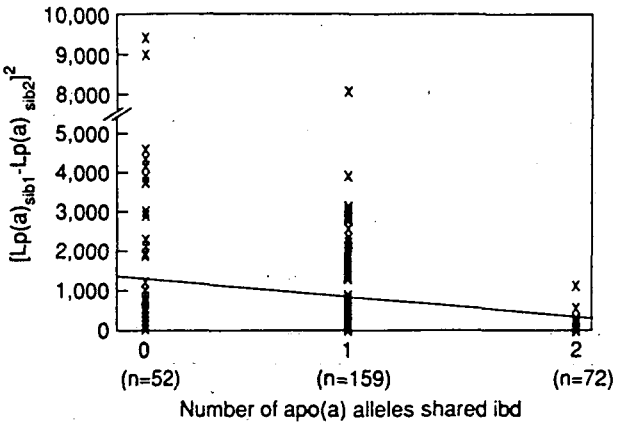


Figure 5. Squared difference between Lp(a) levels of siblings as a function of the proportion of apo(a) alleles shared identical by descent. The regression line for the squared difference on the proportion of apo(a) alleles shared identical by descent is given.

says were employed to measure plasma Lp(a) concentrations in families, there was evidence for a major gene, as well as polygenic factors, contributing to the plasma Lp(a) level (16, 17). In one large Caucasian pedigree, a major gene with three alleles was estimated to account for 73% of the variance in Lp(a) levels (16).

The first molecular clue that the apo(a) gene played a key role in the genetics of plasma Lp(a) concentrations, was the observation that the size of the apo(a) glycoprotein was inversely related to the plasma level of Lp(a) (19). Utermann and his colleagues estimated that differences in the size of the apo(a) glycoprotein accounted for 41% of the variation in inter-individual plasma Lp(a) levels (22). Further support for the apo(a) gene being the major gene influencing Lp(a) levels came from linkage analyses between segregation of plasma

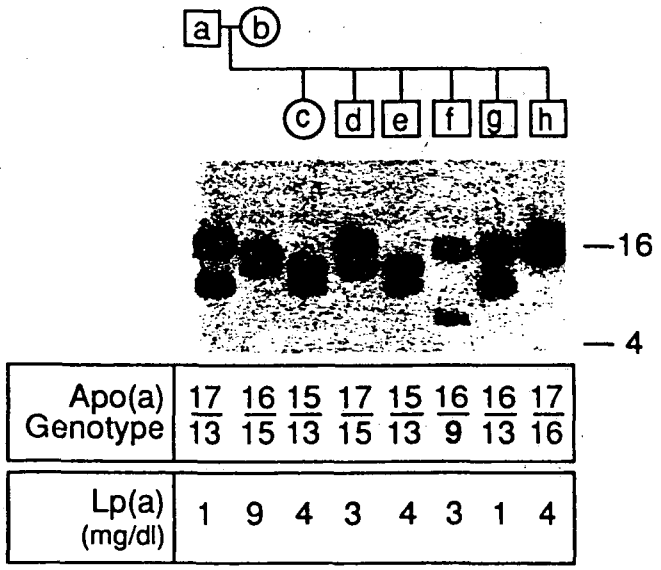


Figure 6. Genomic blot of kringle 4–encoding region of an apo(a) gene in a family in which there is generation of a new apo(a) allele of different length. Southern blotting of the kringle 4–encoding region of the apo(a) gene was performed as described in the Methods and Fig. 2. Individual *f* has inherited *apo(a) 16* from his mother, and *apo(a) 9* from his father. Paternity was confirmed by analysis of seven different VNTR sequences, as described in Methods.

Lp(a) concentrations and DNA sequences in the plasminogen gene which is closely linked to the apo(a) gene (45, 46).

The present study is distinguished from prior family studies by the fact that the apo(a) gene, rather than the expressed protein, was examined in relation to the level of Lp(a) in plasma. In prior studies the immunoblotting techniques used to examine apo(a) isoforms failed to detect protein products from all apo(a) alleles. Numerous exceptions to the inverse relationship between plasma Lp(a) levels and the size of the apo(a) protein were reported. It was suggested that these exceptions were due to the fact that not all apo(a) alleles were detected using the immunoblotting assay. In this study, apo(a) alleles associated with little or no production of apo(a) protein were included in the analysis. As a result, length variation within the kringle 4-encoding region of the apo(a) gene accounted for a greater proportion of the inter-individual variation in plasma Lp(a) concentrations than had been previously estimated (69% vs. 41% [22]).

The reason for the inverse correlation between the size of the apo(a) gene and the level of plasma Lp(a) is not known. Studies in primates have shown that there is not always a consistent relationship between the abundance of apo(a) mRNA, and its size, suggesting that differences in apo(a) gene transcription can not entirely account for this inverse relationship (26, 47). The size of the apo(a) mRNA transcript or glycoprotein may influence its rate of translation, or intracellular transport, respectively.

Alternatively, the observed inverse relationship may not be directly related to the number of kringle 4 repeats in the apo(a) gene, mRNA, or glycoprotein. The apo(a) alleles of different size might be in linkage disequilibrium with the actual sequences at the apo(a) locus that mediate the effect on plasma Lp(a) concentrations. The number of kringle 4 repeats in the apo(a) gene may not have a direct effect on plasma Lp(a) concentrations. In the marmoset monkey, for example, the plasma Lp(a) concentrations vary over a 100-fold range and yet there is only a single apo(a) isoform (48). In the current study, the contribution of the apo(a) gene was partitioned into two components to demonstrate that sequences at the apo(a) locus, other than the number of kringle 4 repeats, contribute importantly to plasma levels of Lp(a). If, however, the number of kringle 4 repeats in the apo(a) gene are in linkage disequilibrium with the actual sequences responsible for mediating the effect of the apo(a) gene or plasma level of Lp(a), then the contribution of the length polymorphism to the inter-individual variation in plasma Lp(a) levels has been overestimated.

Another possible cause for individuals with the same apo(a) genotypes having different plasma concentrations of Lp(a) is that alleles of the same size may differ in the composition of kringle 4 sequences. Not all the kringle sequences are identical. The first, as well as the last eight kringle repeats, differ from the common kringle 4 repeat (the so-called "A" repeat) by between 4 and 35 amino acids (25). Due to the frequent recombinational events involving this locus, is it highly likely that apo(a) alleles vary in their kringle 4 composition, as well as number. Subtle differences in the kringle 4 sequences may impact importantly on the synthesis, and/or degradation of Lp(a).

The length polymorphism in the apo(a) gene has a heterozygosity index comparable to that of number of tandem repeat (VNTR) loci employed in forensic and genetic linkage studies (39, 49, 50). The mutation rate at VNTR loci is several magni-

tudes higher than the usual bi-allelic DNA sequence polymorphisms (39, 51). Given the large number of different sized alleles at the apo(a) locus, a relatively high mutation rate was expected. Therefore, it was anticipated that mutations in the gene would be identified if a sufficient number of families were analyzed. We have observed one mutation of an apo(a) allele out of a total of 376 meioses, and this rate is of the same order of magnitude as the frequency of newly generated alleles for VNTR sequences (39, 51).

Most length polymorphisms in the human genome involve noncoding sequences. The coding regions of several mammalian genes have short tandem repeats (i.e., less the 50 basepairs) which are polymorphic in length (52–56). There are also examples of entire genes being tandemly repeated, as is the case with rDNA, 5S DNA, and the histone genes. The apo(a) length polymorphism is distinguished by the fact that the repeated sequence is large (5.5 kb) and contains both coding and noncoding sequences. The polyubiquitin gene (UbC), contains a large length polymorphism within its coding sequence, but all of the repeated sequences are contained in one exon, and each of the seven to nine repeats encodes the entire protein (57). The heterozygosity index of this length polymorphism is low (22%) compared to the apo(a) gene (94%). The extremely high degree of heterozygosity at the apo(a) locus may reflect the fact that it is under less selective pressure. A physiological function for this enigmatic protein has yet to be identified (2). Alternatively, there may be something intrinsic to the kringle 4-encoding sequences which make them more susceptible to recombinational events.

Mutations of repeated sequence domains result from either intrachromosomal or interchromosomal events. Initially, it was proposed that the mechanism primarily responsible for the high degree of size polymorphism in VNTRs was due to homologous recombination and unequal exchange during meiosis. However, molecular analysis of several new mutations revealed no exchange of flanking genetic markers, which suggests that intrachromosomal, rather than interchromosomal, events appear to be predominantly responsible (51, 58). Similarly, recent molecular analysis of tandem duplication within the Duchenne muscular dystrophy gene demonstrated that the recombinational events were due to intrachromosomal unequal exchange between sister chromatids rather than involving homologous chromosomes (59, 60). Efforts are now being directed to identify polymorphisms flanking the apo(a) gene to analyze the nature of the mutational event(s) responsible for the observed size heterogeneity at the apo(a) locus.

Acknowledgments

The authors thank Drs. Michael Brown and Joseph Goldstein for helpful discussions, Dr. Scott Grundy for assistance with the lipid and lipoprotein measures, and the laboratory of Dr. Steve Daiger at the University of Texas Health Science Center at Houston for markers which aided in the paternity analysis. Tommy Hyatt, Kathy Schueler, and Myriam Fornage provided excellent technical assistance.

This work was supported by grants from the Perot Family Foundation and grants, HL-47619, HL-20948, and HL-40613 from the National Institutes of Health, and 90-IJ-CS-0038 from the National Institute of Justice. E. Boerwinkle is a recipient of a Research Career Development Award and is an Established Investigator of the American Heart Association. H. H. Hobbs is an Established Investigator of the American Heart Association. Giulia Chiesa has a fellowship from the Italian Ministry of Scientific and Technological Research, and Carolin Lackner is a Schroedinger Scholar.

References

1. Berg, K. 1963. A new serum type system in man—the Lp system. *Acta Pathol. Microbiol. Scand.* 59:369–382.

2. Utermann, G. 1989. The mysteries of lipoprotein(a). *Science (Wash. DC).* 246:904–910.

3. Scanu, A. M., and G. M. Fless. 1990. Lipoprotein(a): heterogeneity and biological relevance. *J. Clin. Invest.* 85:1709–1715.

4. Rhoads, G. G., G. H. Dahlèn, K. Berg, N. E. Morton, and A. L. Dannenberg. 1986. Lp(a) lipoprotein as a risk factor for myocardial infarction. *JAMA (J. Am. Med. Assoc.)* 256:2540–2544.

5. Dahlen, G. H., J. R. Guyton, M. Attar, J. A. Farmer, J. A. Kautz, and A. M. Gotto, Jr. 1986. Association of levels of lipoprotein Lp(a), plasma lipids, and other lipoproteins with coronary artery disease documented by angiography. *Circulation.* 74:758–765.

6. Seed, M., F. Hoppichler, D. Reaveley, S. McCarthy, G. R. Thompson, E. Boerwinkle, and G. Utermann. 1990. Relation of serum lipoprotein(a) concentration and apolipoprotein(a) phenotype to coronary heart disease in patients with familial hypercholesterolemia. *N. Engl. J. Med.* 322:1494–1499.

7. Zenker, G., P. Költringer, G. Bonè, K. Niederkorn, K. Pfeiffer, and G. Jürgens. 1986. Lipoprotein(a) as a strong indicator for cerebrovascular disease. *Stroke.* 17:942–945.

8. Miles, L. A., G. M. Fless, E. G. Levin, A. M. Scanu, and E. F. Plow. 1989. A potential basis for the thrombotic risks associated with lipoprotein(a). *Nature (Lond.).* 339:301–303.

9. Edelberg, J. M., and S. V. Pizzo. 1991. Lipoprotein(a): the link between impaired fibrinolysis and atherosclerosis. *Fibrinolysis.* 5:135–143.

10. Albers, J. J., J. L. Adolphson, and W. R. Hazzard. 1977. Radio-immunoassay of human plasma Lp(a) lipoprotein. *J. Lipid Res.* 18:331–338.

11. Berg, K., and J. Mohr. 1963. Genetics of the Lp system. *Acta Genet.* 13:349–360.

12. Sing, C. F., J. S. Schultz, and D. C. Shreffler. 1974. The genetics of the Lp antigen. II. A family study and proposed models of genetic control. *Ann. Hum. Genet.* 38:47–56.

13. Iselius, L., G. H. Dahlèn, U. De Faire, and T. Lundman. 1981. Complex segregation analysis of the Lp(a)/pre-β1-lipoprotein trait. *Clin. Genet.* 20:147–151.

14. Albers, J. J., and W. R. Hazzard. 1974. Immunochemical quantification of human plasma Lp(a) lipoprotein. *Lipids.* 9:15–26.

15. Hasstedt, S. J., D. E. Wilson, C. Q. Edwards, W. N. Cannon, D. Carmelli, and R. R. Williams. 1983. The genetics of quantitative plasma Lp(a): analysis of a large pedigree. *Am. J. Med. Genet.* 16:179–188.

16. Morton, N. E., K. Berg, G. H. Dahlèn, R. E. Ferrell, and G. Rhoads. 1985. Genetics of the Lp Lipoprotein in Japanese-Americans. *Genet. Epidemiol.* 2:113–121.

17. Hasstedt, S. J., and R. R. Williams. 1986. Three alleles for quantitative Lp(a). *Genet. Epidemiol.* 3:53–55.

18. Fless, G. M., M. E. Zum Mallen, and A. M. Scanu. 1986. Physiological properties of apolipoprotein(a) and lipoprotein(a-) derived from the dissociation of human plasma lipoprotein(a). *J. Biol. Chem.* 261:8712–8718.

19. Utermann, G., H. J. Menzel, H. G. Kraft, H. C. Duba, H. G. Kemmler, and C. Seitz. 1987. Lp(a) glycoprotein phenotypes: inheritance and relation to Lp(a)-lipoprotein concentrations in plasma. *J. Clin. Invest.* 80:458–465.

20. Utermann, G., H. G. Kraft, H. J. Menzel, T. Hopferwieser, and C. Seitz. 1988. Genetics of the quantitative Lp(a) lipoprotein trait. I. Relation of Lp(a) glycoprotein phenotypes to Lp(a) lipoprotein concentrations in plasma. *Hum. Genet.* 78:41–46.

21. Utermann, G., C. Duba, and H. J. Menzel. 1988. Genetics of the quantitative Lp(a) lipoprotein trait. II. Inheritance of Lp(a) glycoprotein phenotypes. *Hum. Genet.* 78:47–50.

22. Boerwinkle, E., H. G. Menzel, H. G. Kraft, and G. Utermann. 1989. Genetics of the quantitative Lp(a) lipoprotein trait. III. Contribution of Lp(a) glycoprotein phenotypes to normal lipid variation. *Hum. Genet.* 82:73–78.

23. Sandholzer, C., D. M. Hallman, N. Saha, G. Sigurdsson, C. Lackner, A. Csàszàr, E. Boerwinkle, and G. Utermann. 1991. Effects of the apolipoprotein(a) size polymorphism on the lipoprotein(a) concentration in 7 ethnic groups. *Hum. Genet.* 86:607–614. (Abstr.)

24. Gaubatz, J. W., K. I. Ghanem, J. Guevara, Jr., M. L. Nava, W. Patsch, and J. D. Morrisett. 1990. Polymorphic forms of human apolipoprotein(a): Inheritance and relationship of their molecular weights to plasma levels of lipoprotein(a). *J. Lipid Res.* 31:603–613.

25. McLean, J. W., J. E. Tomlinson, W.-J. Kuang, D. L. Eaton, E. Y. Chen, G. M. Fless, A. M. Scanu, and R. M. Lawn. 1987. cDNA sequence of human apolipoprotein(a) is homologous to plasminogen. *Nature (Lond.).* 330:132–137.

26. Hixson, J. E., M. L. Britten, G. S. Manis, and D. L. Rainwater. 1989.

Apolipoprotein(a) (apo(a)) glycoprotein isoforms result from size differences in apo(a) mRNA in baboons. *J. Biol. Chem.* 264:6013–6016.

27. Gavish, D., N. Azrolan, and J. L. Breslow. 1989. Plasma Lp(a) concentration is inversely correlated with the ratio of kringle IV/kringle V encoding domains in the apo(a) gene. *J. Clin. Invest.* 84:2021–2027.

28. Koschinsky, M. L., U. Beisiegel, D. Henne-Bruns, D. L. Eaton, and R. M. Lawn. 1990. Apolipoprotein(a) size heterogeneity is related to variable number of repeat sequences in its mRNA. *Biochemistry* 29:640–644.

29. Lackner, C., E. Boerwinkle, C. C. Leffert, T. Rahmig, and H. H. Hobbs. 1991. Molecular basis of apolipoprotein(a) isoform size heterogeneity as revealed by pulsed-field gel electrophoresis. *J. Clin. Invest.* 87:2077–2086.

30. Boerwinkle, E., and C. F. Sing. 1987. The use of measured genotype information in the analysis of quantitative phenotypes in man. III. Simultaneous estimation of the frequency and effects of the apolipoprotein E polymorphism and residual polygenic effects on cholesterol, betalipoprotein and triglyceride levels. *Ann. Hum. Genet.* 51:211–226.

31. Church, G. M., and W. Gilbert. 1984. Genomic sequencing. *Proc. Natl. Acad. Sci. USA.* 81:1991–1995.

32. Menzel, H. J., H. Dieplinger, C. Lackner, F. Hoppichler, J. K. Lloyd, D. R. Muller, C. Labeur, P. J. Talmud, and G. Utermann. 1990. Abetalipoproteinemia with an apoB-100-lipoprotein(a) glycoprotein complex in plasma. *J. Biol. Chem.* 265:981–986.

33. Lipid Research Clinic Program. 1982. Lipid and Lipoprotein Analysis: Manual of Laboratory Operations. Department of Health, Education and Welfare Publ NIH/75-628 Government Printing Office, Washington, DC.

34. Hasstedt, S. J., and P. Cartwright. 1981. PAP: Pedigree Analysis Program. Technical Report 13. Department of Medical Biophysics and Computing, University of Utah, Salt Lake City, UT.

35. Haseman, J. K., and R. C. Elston. 1972. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* 2:2–19.

36. Amos, C. I., R. C. Elston, A. F. Wilson, and J. E. Bailey-Wilson. 1989. A more powerful robust sib-pair test of linkage for quantitative traits. *Genet. Epidemiol.* 6:435–449.

37. Scheffe, H. 1959. The Analysis of Variance. John Wiley & Sons, Inc., New York. 221–235.

38. Neter, J., W. Wasserman, and M. H. Kutner. 1990. Applied Statistical Models. R. D. Irwin, Inc., Homewood, IL. 660–661.

39. Jeffreys, A. J., N. J. Royle, V. Wilson, and Z. Wong. 1988. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature (Lond.).* 332:278–281.

40. Albers, J. J., P. Wahi, and W. R. Hazzard. 1974. Quantitative genetic studies of the human plasma Lp(a) lipoprotein. *Biochem. Genet.* 11:475–486.

41. Hewitt, D., J. Milner, C. Breckenridge, and G. Macguire. 1977. Heritability of "sinking" pre-beta lipoprotein level: a twin study. *Clin. Genet.* 11:224–226.

42. Postle, A. D., J. M. Darmady, and D. C. Siggers. 1978. Double pre-β lipoprotein in ischaemic heart disease. *Clin. Genet.* 13:233–236.

43. Schultz, J. S., D. C. Shreffler, and C. F. Sing. 1974. The genetics of the Lp antigen. I. Its quantitation and distribution in a sample population. *Ann. Hum. Genet.* 38:39–46.

44. Hewitt, D., J. Milner, A. R. G. Owen, W. C. Breckenridge, G. F. Macguire, G. J. L. Jones, and J. A. Little. 1982. The inheritance of sinking-pre-beta lipoprotein and its relation to the Lp(a) antigen. *Clin. Genet.* 21:301–308.

45. Drayna, D. T., R. A. Hegele, P. E. Hass, M. Emi, L. L. Wu, D. L. Eaton, R. M. Lawn, R. R. Williams, R. L. White, and J.-M. Lalouel. 1988. Genetic linkage between lipoprotein(a) phenotype and a DNA polymorphism in the plasminogen gene. *Genomics.* 3:230–236.

46. Weitkamp, L. R., S. A. Guttormsen, and J. S. Schultz. 1988. Linkage between the loci for the Lp(a) lipoprotein (LP) and plasminogen (PLG). *Hum. Genet.* 79:80–82.

47. Azrolan, N., D. Gavish, and J. L. Breslow. 1991. Plasma lipoprotein(a) concentration is controlled by apolipoprotein(a) (Apo(a)) protein size and the abundance of hepatic apo(a) mRNA in a cynomolgus monkey model. *J. Biol. Chem.* 266(21):13866–13872.

48. Guo, H.-C., J.-B. Michel, Y. Blouquit, and M. J. Chapman. 1991. Lipoprotein(a) and apolipoprotein(a) in a new world monkey, the common marmoset (callithrix jacchus): Association of variable plasma lipoprotein(a) levels with a single apolipoprotein(a) isoform. *Arterioscler. Thromb.* 11:1030–1041.

49. Weber, J. L. 1990. Informativeness of human $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ polymorphisms. *Genomics.* 7:524–530.

50. Boerwinkle, E., W. Xiong, E. Fourest, and L. Chan. 1989. Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: Application to the apolipoprotein B 3′ hypervariable region. *Proc. Natl. Acad. Sci. USA.* 86:212–216.

51. Jeffreys, A. J., R. Neumann, and V. Wilson. 1990. Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell.* 60:473–485.

52. Oberlè, I., F. Rousseau, D. Heitz, C. Kretz, D. Devys, A. Hanauer, J. Boué, M. F. Bertheas, and J. L. Mandel. 1991. Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science (Wash. DC).* 252:1097–1102.

53. Azen, E., K. M. Lyons, T. McGonigal, N. L. Barrett, L. S. Clements, N. Maeda, E. F. Vanin, D. M. Carlson, and O. Smithies. 1984. Clones from the human gene complex coding for salivary proline-rich proteins. *Proc. Natl. Acad. Sci. USA.* 81:5561–5565.

54. Swallow, D. M., S. Gendler, B. Griffiths, G. Corney, J. Taylor-Papadimitriou, and M. E. Bramwell. 1987. The human tumour-associated epithelial mucins are coded by an expressed hypervariable gene locus PUM. *Nature (Lond.).* 328:82–84.

55. Boerwinkle, E. S.-H. Chen, S. Visvikis, C. L. Hanis, G. Siest, and L. Chan. 1991. Signal peptide-length variation in human apolipoprotein B gene. *Diabetes.* 40:1539–1544.

56. McPhaul, M. J., M. Marcelli, W. D. Tiley, J. E. Griffin, R. F. Isidro-Gutierrez, and J. D. Wilson. 1991. Molecular basis of androgen resistance in a family with a qualitative abnormality of the androgen receptor and responsive to high-dose androgen therapy. *J. Clin. Invest.* 87:1413–1421.

57. Baker, R. T., and P. G. Board. 1989. Unequal crossover generates variation in ubiquitin coding unit number at the human UbC polyubiquitin locus. *Am. J. Hum. Genet.* 44:534–542.

58. Wolff, R. K., Y. Nakamura, and R. White. 1988. Molecular characterization of a spontaneously generated new allele at a VNTR locus: No exchange of flanking DNA sequence. *Genomics.* 3:347–351.

59. Hu, X., A. H. M. Burghes, D. E. Bulman, P. N. Ray, and R. G. Worton. 1989. Evidence for mutation by unequal sister chromatid exchange in the Duchenne muscular dystrophy gene. *Am. J. Hum. Genet.* 44:855–863.

60. Hu, X., P. N. Ray, and R. G. Worton. 1991. Mechanisms of tandem duplication in the Duchenne muscular dystrophy gene include both homologous and nonhomologous intrachromosomal recombination. *EMBO (Eur. Mol. Biol. Organ.) J.* 10:2471–2477.

Am J H Genetics 49: A129

0\, 1.275

uency of ΔF_{508} or non-ΔF_{508} haplotypes near the cystic sis (CF) locus in relation to ascertainment, ancestry and :al type among 120 Canadian CF families. J.A.Buchanan, M. /, B.Kerem, P.Durie, L.-C.Tsui, P.Ray and M.Buchwald. The Hospital ck Children (HSC) and University of Toronto, Toronto, Ont.,Canada.
ce the discovery of the CF gene and common associated mutation, ΔF_{508}, :s have been made to determine how this and other CF mutations are dis- ed among various populations, and to assess the related clinical signifi- :. Three groups of Canadian patients were asked about their ancestry. p A (49 families) comprised multiplex CF sibships who have been part of esearch base since 1985. Group B (16 families) was ascertained through reatic sufficient (PS) probands and group C (55 families) through the cular Diagnostic Laboratory at HSC. Half were from the HSC clinic and the from other Canadian referral centers. Patients were classified as PS or reatic insufficient (PI) and by whether they were born with meconium ileus . The 240 CF chromosomes were grouped according to the contributing ıts' ancestry: a.Northern European (n=77) b.French Canadian referred from outimi, Que.(n=22) c.other French Canadian (n=16) d.Central European 7) e.Southern European (n=26) or f.mixed/ unknown/other (n=82). all, 65% of CF chromosomes had ΔF_{508}, with the following in ancestral ps: a.74% b.45% c.69% d.71% e.50% f.66%. Analysis by clinical p showed 82% in MI, 71% in PI(without MI) and 34% in PS. Non-ΔF_{508} mosomes were then subdivided according to the major haplotype groups of m et al. (Science 245:1073,1989). Among non-CF chromosomes of all ic groups, and among CF-PS chromosomes, group II haplotypes ominated. In contrast, group III/IV/V dominated the non-ΔF_{508} CF mosomes of Northern Europeans. Among those from Chicoutimi and ng MI patients, group I haplotypes were more prevalent. Such analysis led e following observations: 1) Ancestry of these Canadian families reflected a h-south European gradient in ΔF_{508} frequency. 2) The proportion of ΔF_{508} ng those from Chicoutimi (but not other French Canadians) was relatively due to the prevalence of other CF mutation(s) on group I haplotype grounds. 3) Southern European ancestry was common among PS patients. F_{508} was most frequent among MI probands and least among PS. lotypes for non-ΔF_{508} chromosomes also were differentially distributed eg the 3 clincal groups. 5) Ascertainment through the diagnostic laboratory ded a somewhat higher frequency of MI and of ΔF_{508} than that from CF ics. 6) Linkage disequilibrium was still apparent for non-ΔF_{508} CF chromo- es, an observation difficult to reconcile with the current indication that there not her common CF mutations. Supported by Cystic Fibrosis ndati nada and US).

✓(0502) 1.276

Discretized allelic data for a VNTR locus by amplified fragment length polymorphism (AMP-FLP) analysis. B. Budowle and A. M. Giusti, FSRTC, FBI Academy, Quantico, VA and R. Chakraborty, Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston, TX. (Intro. by C. Comey).

Allelic data for the D1S80 locus was obtained by using the polymerase chain reaction and subsequent analysis with a high resolution, polyacrylamide gel electrophoresis technique and silver staining. Compared with restriction fragment length polymorphism (RFLP) analysis of variable number of tandem repeat (VNTR) loci by Southern blotting, this approach offers certain advantages: 1) discrete allele resolution, 2) minimal measurement error, 3) correct genotyping of single band VNTR patterns, 4) a nonisotopic assay, and 5) a permanent record of the electrophoretic separation. For a sample population analyzed by this approach for D1S80, the distribution of phenotypes is in agreement with expected values according to the Hardy-Weinberg equilibrium. Moreover, the observed number of alleles and the level of gene diversity are congruent with each other in accordance with the expectation of a mutation-drift equilibrium model for a single homogeneous random mating population. AMP-FLP analysis of VNTR loci may prove useful as models for population genetic issues for VNTR loci analyzed by RFLP typing via Southern blotting.

503) 1.277

stic fibrosis in the Basque Country: European igin of the F508 mutation. I Casals, V. nes, P Mañà, M Chillon, C Lázaro, C Vázquez d X Estivill. Molecular Genetics Department, ndació d'Investigació Sant Pau 08025 Barcelona
The analysis of the ΔF508 cystic fibrosis (CF) tation in diferent populations shows frequencies 90% in Danish, 75% in British and North erican, 50% in Spanish and Italian, and 30% in hkenazic populations. We have analysed the sque Country population for DNA polymorphism plotypes and for the F508 mutation. The rthplaces of parents and grandparents were aced, and CF chromosomes were classified by igin. 81% of CF chromosomes of Basque origin rry the Δ F508 mutation, and only 51% of non sque chromosomes have the mutation. The iqueness of the Basque population in respect to netic differences demonstrated at other genetic ci (ABO and HLA), and linguistic differences of e Basque language to the Indo-European nguages, suggests that the Basque population was ttled in their territory before the arrival, in rope, of the Indo-Europeans. The high frequency f the F508 mutation in the Basque Country ntradicts the hypothesis that the mutated CFTR ne arrived with the migrants from the Middle ast to the North-West of Europe, and suggests at the F508 mutation was already present within he inhabitants of the old Continent, before this igration, about 5.000 years ago. If this is the ase, the Indo-European migration diluted the he frequency of the mutation Δ F508 by bringing ther CF mutations into Europe.
ckno gments: "Fondo de Investigaciones anit de la Seguridad Social" 90E1254 and Inst Català de la Salut".

(0504) 11.4

Population genetics of VNTR polymorphism in humans. R. Chakraborty and E. Boerwinkle. Genetics Centers, University of Texas Graduate School of Biomedical Sciences, Houston, Texas, USA.

Several regions of the human genome have been identified that exhibit a high degree of polymorphism due to a variable number of tandemly repeated (VNTR) DNA sequences. While the utility of these VNTR loci have been well publicized, their population genetic characteristics are poorly understood because of: (1) the large number of rare alleles; (2) presumptive high rate of "mutation"; and (3) possibility of incomplete resolution of similar size alleles. Understanding the population genetic characteristics is necessary for optimal utilization of these highly informative loci for gene mapping and genetic identification purposes.

We are studying the population genetic characteristics of several VNTR and microsatellite loci in a sample of 600 individuals, in addition to the data available in the published literature. Because of the large number of alleles and a relatively moderate sample size, standard tests of Hardy-Weinberg equilibrium (HWE) and gametic disequilibrium are inadequate, and alternative tests are proposed. These conservative tests are based on the exact sampling distributions of numbers of observed homozygotes and heterozygotes in a finite sample. When the typing method is such that all alleles are distinguishable (e.g., PCR typing of the 3' apolipoprotein-B VNTR), the genotype distribution fits the predictions of HWE, and no disequilibria are observed among unlinked VNTR loci. In the presence of incomplete resolution of alleles (e.g., D2S44 VNTR) significant departures from equilibrium expectations are observed. In addition, when complete resolution of alleles and genotypes is achieved, the classic mutation-drift (infinite allele) model accounts for the large amount of allelic diversity. The lack of complete resolution causes conspicuous discrepencies between the observed and expected allele frequency profiles. We propose a new model of forward-backward "mutational" changes that represents the population dynamics of VNTR allelic diversity more adequately. Our results indicate that the laboratory techniques applied for typing VNTR alleles plays a large role in dictating the population genetic features of VNTR loci. Ignoring this aspect may result in a wrong inference about population structure, consequently handicapping the optimal utility of these loci. (Research supported by grants GM-41399 and HL-40613 from the US National Institutes of Health).

Am J H Genetics 49: A129

0.. 1.275

quency of ΔF_{508} or non-ΔF_{508} haplotypes near the cystic sis (CF) locus in relation to ascertainment, ancestry and cal type among 120 Canadian CF families. J.A.Buchanan, M. y, B.Kerem, P.Durie, L.-C.Tsui, P.Ray and M.Buchwald. The Hospital ick Children (HSC) and University of Toronto, Toronto, Ont.,Canada.
ice the discovery of the CF gene and common associated mutation, ΔF_{508}, ts have been made to determine how this and other CF mutations are dis- ted among various populations, and to assess the related clinical signifi- e. Three groups of Canadian patients were asked about their ancestry. p A (49 families) comprised multiplex CF sibships who have been part of esearch base since 1985. Group B (16 families) was ascertained through reatic sufficient (PS) probands and group C (55 families) through the cular Diagnostic Laboratory at HSC. Half were from the HSC clinic and the from other Canadian referral centers. Patients were classified as PS or reatic insufficient (PI) and by whether they were born with meconium ileus . The 240 CF chromosomes were grouped according to the contributing nts' ancestry: a.Northern European (n=77) b.French Canadian referred from outimi, Que.(n=22) c.other French Canadian (n=16) d.Central European 7) e.Southern European (n=26) or f.mixed/ unknown/other (n=82). rall, 65% of CF chromosomes had ΔF_{508}, with the following in ancestral ps: a.74% b.45% c.69% d.71% e.50% f.66%. Analysis by clinical p showed 82% in MI, 71% in PI(without MI) and 34% in PS. Non-ΔF_{508} mosomes were then subdivided according to the major haplotype groups of m et al. (Science 245:1073,1989). Among non-CF chromosomes of all ic groups, and among CF-PS chromosomes, group II haplotypes lominated. In contrast, group III/IV/V dominated the non-ΔF_{508} CF mosomes of Northern Europeans. Among those from Chicoutimi and ng MI patients, group I haplotypes were more prevalent. Such analysis led e following observations: 1) Ancestry of these Canadian families reflected a h-south European gradient in ΔF_{508} frequency. 2) The proportion of ΔF_{508} ng those from Chicoutimi (but not other French Canadians) was relatively , due to the prevalence of other CF mutation(s) on group I haplotype kgrounds. 3) Southern European ancestry was common among PS patients. ΔF_{508} was most frequent among MI probands and least among PS. lotypes for non-ΔF_{508} chromosomes also were differentially distributed ng the 3 clincal groups. 5) Ascertainment through the diagnostic laboratory ded a somewhat higher frequency of MI and of ΔF_{508} than that from CF ics. 6) Linkage disequilibrium was still apparent for non-ΔF_{508} CF chromo- es, an observation difficult to reconcile with the current indication that there y not b other common CF mutations. Supported by Cystic Fibrosis ndati nada and US).

✓(0502) 1.276

Discretized allelic data for a VNTR locus by amplified fragment length polymorphism (AMP-FLP) analysis. B. Budowle and A. M. Giusti, FSRTC, FBI Academy, Quantico, VA and R. Chakraborty, Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston, TX. (Intro. by C. Comey).

Allelic data for the D1S80 locus was obtained by using the polymerase chain reaction and subsequent analysis with a high resolution, polyacrylamide gel electrophoresis technique and silver staining. Compared with restriction fragment length polymorphism (RFLP) analysis of variable number of tandem repeat (VNTR) loci by Southern blotting, this approach offers certain advantages: 1) discrete allele resolution, 2) minimal measurement error, 3) correct genotyping of single band VNTR patterns, 4) a nonisotopic assay, and 5) a permanent record of the electrophoretic separation. For a sample population analyzed by this approach for D1S80, the distribution of phenotypes is in agreement with expected values according to the Hardy-Weinberg equilibrium. Moreover, the observed number of alleles and the level of gene diversity are congruent with each other in accordance with the expectation of a mutation-drift equilibrium model for a single homogeneous random mating population. AMP-FLP analysis of VNTR loci may prove useful as models for population genetic issues for VNTR loci analyzed by RFLP typing via Southern blotting.

503) 1.277

stic fibrosis in the Basque Country: European igin of the F508 mutation. I Casals, V. ines, P Maña, M Chillón, C Lázaro, C Vázquez d X Estivill. Molecular Genetics Department, ndació d'Investigació Sant Pau 08025 Barcelona
The analysis of the ΔF508 cystic fibrosis (CF) tation in diferent populations shows frequencies 90% in Danish, 75% in British and North erican, 50% in Spanish and Italian, and 30% in hkenazic populations. We have analysed the sque Country population for DNA polymorphism plotypes and for the F508 mutation. The rthplaces of parents and grandparents were aced, and CF chromosomes were classified by igin. 81% of CF chromosomes of Basque origin rry the Δ F508 mutation, and only 51% of non sque chromosomes have the mutation. The niqueness of the Basque population in respect to netic differences demonstrated at other genetic oci (ABO and HLA), and linguistic differences of e Basque language to the Indo-European anguages, suggests that the Basque population was ettled in their territory before the arrival, in urope, of the Indo-Europeans. The high frequency f the F508 mutation in the Basque Country ontradicts the hypothesis that the mutated CFTR ene arrived with the migrants from the Middle ast to the North-West of Europe, and suggests hat the F508 mutation was already present within he inhabitants of the old Continent, before this igration, about 5.000 years ago. If this is the ase, the Indo-European migration diluted the he frequency of the mutation Δ F508 by bringing ther CF mutations into Europe.
cknowledgments: "Fondo de Investigaciones anit de la Seguridad Social" 90E1254 and Inst Català de la Salut".

(0504) 11.4

Population genetics of VNTR polymorphism in humans. R. Chakraborty and E. Boerwinkle. Genetics Centers, University of Texas Graduate School of Biomedical Sciences, Houston, Texas, USA.

Several regions of the human genome have been identified that exhibit a high degree of polymorphism due to a variable number of tandemly repeated (VNTR) DNA sequences. While the utility of these VNTR loci have been well publicized, their population genetic characteristics are poorly understood because of: (1) the large number of rare alleles; (2) presumptive high rate of "mutation"; and (3) possibility of incomplete resolution of similar size alleles. Understanding the population genetic characteristics is necessary for optimal utilization of these highly informative loci for gene mapping and genetic identification purposes.

We are studying the population genetic characteristics of several VNTR and microsatellite loci in a sample of 600 individuals, in addition to the data available in the published literature. Because of the large number of alleles and a relatively moderate sample size, standard tests of Hardy-Weinberg equilibrium (HWE) and gametic disequilibrium are inadequate, and alternative tests are proposed. These conservative tests are based on the exact sampling distributions of numbers of observed homozygotes and heterozygotes in a finite sample. When the typing method is such that all alleles are distinguishable (e.g., PCR typing of the 3' apolipoprotein-B VNTR), the genotype distribution fits the predictions of HWE, and no disequilibria are observed among unlinked VNTR loci. In the presence of incomplete resolution of alleles (e.g., D2S44 VNTR) significant departures from equilibrium expectations are observed. In addition, when complete resolution of alleles and genotypes is achieved, the classic mutation-drift (infinite allele) model accounts for the large amount of allelic diversity. The lack of complete resolution causes conspicuous discrepancies between the observed and expected allele frequency profiles. We propose a new model of forward-backward "mutational" changes that represents the population dynamics of VNTR allelic diversity more adequately. Our results indicate that the laboratory techniques applied for typing VNTR alleles plays a large role in dictating the population genetic features of VNTR loci. Ignoring this aspect may result in a wrong inference about population structure, consequently handicapping the optimal utility of these loci. (Research supported by grants GM-41399 and HL-40613 from the US National Institutes of Health).

Am. J. Hum. Genet. 48:137–144, 1991

Analysis of the VNTR Locus DIS80 by the PCR Followed by High-Resolution PAGE

Bruce Budowle,* Ranajit Chakraborty,† Alan M. Giusti,* Arthur J. Eisenberg,‡ and Robert C. Allen§

*Forensic Science Research and Training Center, Laboratory Division, Federal Bureau of Investigation Academy, Quantico, VA; †Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston; ‡DNA Identity Laboratory, Department of Pathology, Texas College of Osteopathic Medicine, Fort Worth; and §Department of Pathology, Medical University of South Carolina, Charleston

Summary

Allelic data for the D1S80 locus was obtained by using the PCR and subsequent analysis with a high-resolution, horizontal PAGE technique and silver staining. Compared with RFLP analysis of VNTR loci by Southern blotting, the approach described in this paper offers certain advantages: (1) discrete allele resolution, (2) minimal measurement error, (3) correct genotyping of single-band VNTR patterns, (4) a nonisotopic assay, (5) a permanent record of the electrophoretic separation, and (6) reduced assay time. In a sample of 99 unrelated Caucasians, the D1S80 locus demonstrated a heterozygosity of 80.8% with 37 phenotypes and 16 alleles. The distribution of genotypes is in agreement with expected values according to the Hardy-Weinberg equilibrium. Furthermore, the observed number of alleles and the level of heterozygosity, obtained through the protocol described here, were congruent with each other in accordance with the expectation of a mutation-drift equilibrium model for a single, homogeneous, random-mating population. Therefore, the analysis of D1S80 and similar VNTR loci by amplified fragment length polymorphism (AMP-FLP) may prove useful as models for population genetic issues for VNTR loci analyzed by RFLP typing via Southern blotting.

Introduction

Identity tests, as performed in the fields of paternity and forensics, rely on the detection of genetic differences among individuals. At present, highly polymorphic loci whose alleles are the result of VNTRs are the most informative genetic markers for genetic characterization. Although extremely effective for VNTR analyses, the RFLP methodology via Southern blotting (Southern 1975) is time consuming and requires an isotopic assay to achieve the sensitivity necessary to detect VNTR alleles in samples containing as little as 10–50 ng of human DNA samples (Budowle and Baechtel 1990). Additionally, because of the inability of the RFLP technology to resolve discretely the alleles of most VNTR loci, statistical analyses that are different from those used for traditional genetic marker systems have been required (Budowle et al., in press).

The PCR (Saiki et al. 1985) offers a viable alternative to RFLP analysis of VNTR loci, particularly in situations where limited quantities of DNA are available. The use of the PCR can obviate the need for isotopic detection and reduce assay time and cost. With appropriate VNTR loci and high-resolution discontinuous buffer electrophoretic systems in polyacrylamide gels (Allen et al. 1989; Budowle and Allen 1990), amplification of specific DNA sequences by the PCR could prove useful for identity testing, population genetics, and disease susceptibility studies. In fact, the D17S30 (also designated D17S5) locus (Horn et al. 1989) and the 3′ hypervariable region of the apolipoprotein B gene (Boerwinkle et al. 1989; Ludwig et al. 1989) have been

Received May 15, 1990; revision received August 24, 1990.
Address for correspondence and reprints: Bruce Budowle, Forensic Science Research and Training Center, Federal Bureau of Investigation Academy, Quantico, VA 22135.

analyzed using the PCR and subsequent electrophoretic separation of the amplified fragments.

This paper describes the results of the analysis of PCR-amplified products of the VNTR locus D1S80 (Nakamura et al. 1988). The procedure resolves alleles of D1S80 into discrete entities, uses an inexpensive silver stain for detection, and provides a permanent record of the electrophoretic separation. With an analytical system that enables resolution of discrete alleles and therefore permits correct genotyping of VNTR profiles, it will now be possible to apply the conventional formula of the Hardy-Weinberg rule (i.e., χ^2 analysis on observed and expected genotype classes). This will allow for an evaluation of the goodness of fit of the genotype distributions of the particular VNTR locus for a sample population. Moreover, with the discrete resolution of alleles it will be possible to evaluate the appropriateness of classical population genetic models of allele frequency distributions at this locus to validate an assumption of genetic homogeneity of the population from which the sample is derived.

Material and Methods

Whole blood was obtained in EDTA Vacutainer tubes by venipuncture from 100 unrelated Caucasian donors at the FBI Academy. The DNA was extracted as described previously (Budowle and Baechtel 1990). Purified DNA from a two-generation (10 individuals) and a four-generation family (18 individuals) was provided by M. Skolnick (University of Utah, Salt Lake City, UT).

Amplification of D1S80 was achieved using the primers described by Kasai et al. (in press). The primers were 5′-GAAACTGGCCTCCAAACACTGCCCGCCG-3′ and 5′-GTCTTGTTGGAGATGCACGTGCCCCTT-GC-3′. Each sample that was amplified contained 100 ng DNA, 10 mM Tris-Cl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 0.01% gelation, 2.5 units of Amplitaq™ DNA polymerase (Perkin Elmer–Cetus), 1 µM of each primer, and 200 µM of each dNTP. The total volume of each sample was 50 µl. Each sample was overlaid with 50 µl of mineral oil. The PCR was carried out in a Perkin-Elmer Thermocycler for 25 cycles. Each cycle consisted of 1 min at 95°C for denaturation, 1 min at 65°C for primer annealing, and 8 min at 70°C for primer extension. After amplification, the mineral oil was removed and the samples were stored at either 4°C or −20°C prior to electrophoretic analysis.

Ultrathin-layer polyacrylamide gels (5% T, 3% C; 400 µm thick) were cast onto Gelbond (FMC, Rockland, ME) using the flap technique (Allen 1980). The gels were cross-linked with piperazine diacrylamide (Hochstrasser et al. 1988) (Bio-Rad Laboratories, Richmond, CA). All gels contained 7.1% glycerol and 33 mM Tris-sulfate buffer, pH 9.0. If rehydratable polyacrylamide gels were used (Allen et al. 1989; Budowle and Allen 1990), they were rehydrated in a solution containing 33 mM Tris-sulfate, pH 9.0, and 7.1% glycerol. The trailing ion, contained in 2% (wt/vol) agarose plugs, was 0.14 M Tris-borate, pH 9.0. Bromophenol blue (a final concentration of 0.01%) was added to the electrode buffer to serve as a dye marker for the discontinuous buffer boundary. The electrophoretic setup was similar to that described by Allen et al. (1989) and Budowle and Allen (1990). The distance between the edges of the agarose plugs was 10 cm. The amplified fragment length polymorphisms (AMP-FLPs) of D1S80 were absorbed into fiberglass applicator tabs (2.5 × 5.0 mm, Pharmacia-LKB, catalog no. 1850-901), lightly blotted, and applied to the gel surface 1 cm from the cathode. The conditions for electrophoretic separation were similar to those described previously for rehydratable polyacrylamide gels (Allen et al. 1989). Electrophoretic separation was stopped when the bromophenol blue dye front reached the anodal wick. Following electrophoresis, the gels were stained with silver, according to the conditions described in table 1, so the pattern could be visualized directly.

Hybridization analysis of the PCR-amplified products (or AMP-FLPs) subsequent to electrophoresis in the ultrathin-layer polyacrylamide gels was accomplished using a passive blotting procedure. After electrophoresis, the DNA in the gel was denatured by washing the gel in 0.4 M NaOH for 5 min. A nylon membrane (Zeta Probe, Bio-Rad Laboratories, Richmond, CA), prewetted in 0.4 M NaOH, was placed directly on the gel surface, and, subsequently, a blot pad (BRL, Gaithersburg, MD) was placed on the membrane. Transfer time was 1 h at ambient temperature. After transfer, the membrane was washed briefly in a solution containing 2 × SSC (20 × SSC = 1,753 g NaCl and 88.2 g sodium citrate/l, pH 7.0) (Maniatis et al. 1982) and 0.2 M Tris, pH 7.5. The membrane was blotted between two Whatmann 1 MM papers and baked in an oven at 80°C for 30 min. The membrane was wrapped in plastic wrap and stored at −20°C. The probe pMCT118 (for locus D1S80) was provided by Y. Nakamura and R. White (Howard Hughes Medical Institute, Salt Lake City, UT). Random primer labeling was accomplished according to the manufacturer's in-

Table 1

Protocol for Silver-Staining AMP-FLP Gels

Step	Time
Place gel in 10% ethanol solution	5 min
Oxidize gel in 1% nitric acid solution	3 min
Rinse gel in distilled water	A few seconds
Place gel in 0.012 M silver nitrate solution	20 min
Decant silver nitrate and rinse gel in distilled water	A few seconds
Reduce gel in a solution containing 0.28 M sodium carbonate (anhydrous) and 0.019% formalin; several changes of reducing solution may be necessary; the solution should be changed when it turns brown	Will depend on desired intensity of image; image develops before eye
Stop reduction process with 10% glacial acetic acid	2 min
Place gel in distilled water	2 min
Air dry gel for permanent record	. . .

structions contained within the BRL Random Primer DNA labeling system kit or according to the method of Feinberg and Vogelstein (1983, 1984). Hybridization and stringency washes were carried out according to the method of Budowle and Baechtel (1989). The labeled DNA duplex was detected by autoradiography using Kodak XAR film and Dupont Cronex Lightning Plus intensifying screens.

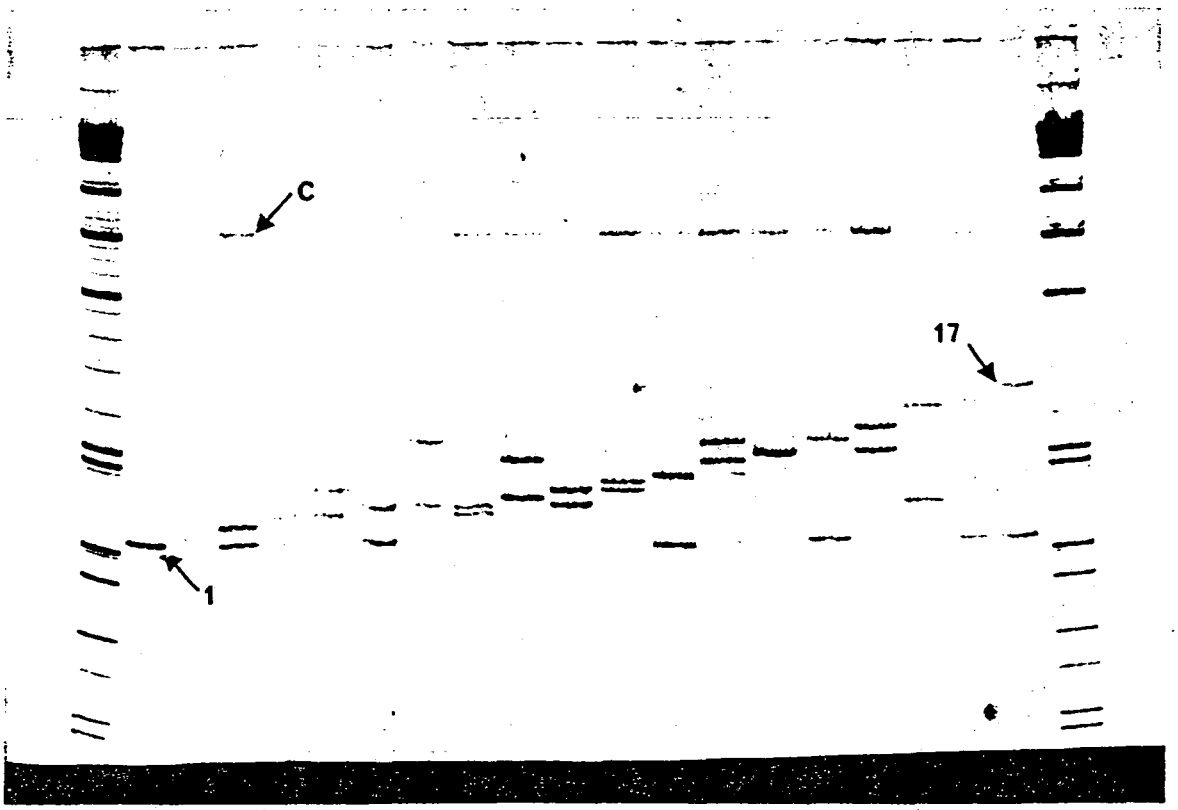


Figure 1 Silver-stained AMP-FLP gel displaying D1S80 profiles. The D1S80 types from left to right are 1-1, 2-1, 3-1, 4-1, 7-4, 5-1, 12-5, 5-4, 10-6, 7-5, 8-7, 9-1, 12-10, 11-11, 12-1, 13-11, 15-5, 16-1, and 17-1. C = constant band. The size standards are a combination of the 1-kb and 123-bp ladder (BRL, Gaithersburg, MD). The cathode is at the top.

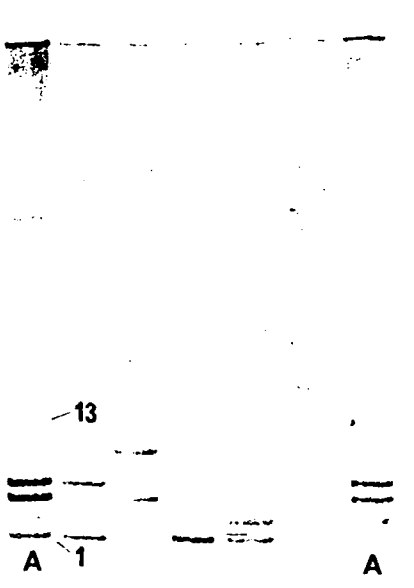


Figure 2 Silver-stained AMP-FLP gel displaying a ladder of composite alleles and D1S80 profiles. The ladder (A) is composed of alleles 1, 3, 5, 7, 9, 11 and 13. The D1S80 types from left to right are 7-1, 10-5, 1-1, 3-1, and 15-1. The cathode is at the top.

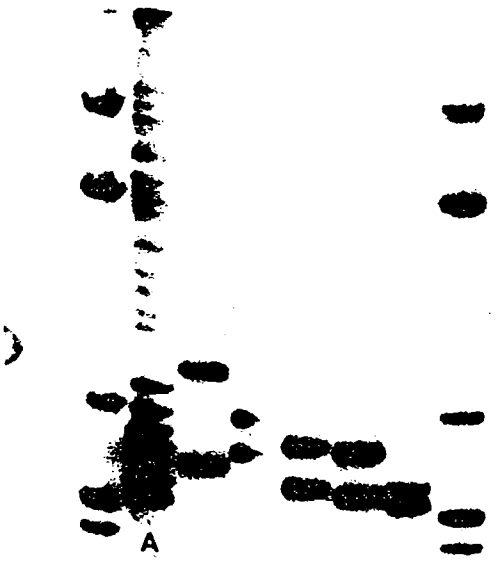


Figure 3 Autoradiogram of an AMP-FLP gel displaying the ladder of composite alleles (A) and D1S80 types. The pMCT118 probe was provided by Ray White and his colleagues (Howard Hughes Medical Institute, Salt Lake City, UT).

Results

Figure 1 shows that AMP-FLP analysis of D1S80 can be performed using the techniques described in this paper. Sixteen different alleles were observed in 99 unrelated Caucasians (one sample did not amplify by PCR). Each allele was completely resolved based on increments of the repeat unit of the VNTR locus. The length of the repeat unit has been determined to be 16 base pairs (Y. Nakamura, Howard Hughes Medical Institute, and T. Holm, GenMark, Salt Lake City, UT, personal communications). The alleles have been designated 1–17 (allele 14 has not yet been observed), where allele 1 is the smallest in length and allele 17 is the largest in length. (It should be noted that the allele designations are temporary and will eventually be based on the number of repeats.)

With this system the alleles can be designated specifically without determining base-pair size. Unknown samples can be compared with a "ladder" consisting of a composite of D1S80 alleles; thus, allele designations are much easier and measurement error is greatly reduced (fig. 2).

It should be noted that a constant (or monomorphic) band appears in each amplified sample (indicated by the arrow at point C in fig. 1). The band does not hybridize with the pMCT118 probe (fig. 3) and thus is a sequence unrelated to D1S80. However, shifts in the position of the constant band can indicate unaligned positioning of the sample tabs at the sample origin. Therefore, the constant band can serve an important function as an internal marker to minimize errors in AMP-FLP typing of D1S80.

The distributions of observable phenotypes and allelic frequencies for D1S80 in a Caucasian population

Table 2

Distribution of D1S80 Genotypes from 99 Unrelated Caucasians

Genotype	Number Observed
1-1	7
2-1	1
3-1	3
4-1	2
5-1	1
6-1	2
7-1	24
8-1	1
9-1	3
10-1	1
11-1	2
12-1	2
16-1	1
17-1	1
7-2	1
8-3	1
5-4	1
7-4	1
10-4	1
7-5	2
10-5	1
11-5	2
12-5	1
15-5	1
10-6	2
7-7	10
8-7	3
10-7	3
11-7	4
12-7	5
13-7	1
15-7	1
10-8	1
12-8	1
12-10	2
11-11	2
13-11	1

sample of 99 unrelated individuals are shown in tables 2 and 3, respectively. The observed heterozygosity is 80.8%. The distribution of the phenotypes is in Hardy-Weinberg equilibrium ($\chi^2 = 2.50$; df = 1; $.100 < P < .250$; Hardy-Weinberg formulation was calculated by comparing observed and expected genotypes; all classes with less than four events were pooled). Further, although not extremely informative (because of limited variation among family members), the two families demonstrated Mendelian inheritance of the AMP-FLP alleles (data not shown).

While hybridization assays do not appear to be necessary for routinely typing D1S80 AMP-FLP profiles, it may be desirable, at times, to use a probe to confirm that the AMP-FLPs truly represent the described locus or to increase the level of sensitivity of detection provided by silver staining. Figure 3 shows that the blotting approach can be used. As expected, after the PCR there should be more than adequate quantities of DNA for hybridization analysis. Although the patterns are weak, there can still be enough residual DNA left in the polyacrylamide gel for detection by silver staining (fig. 4).

Discussion

AMP-FLP analysis of the D1S80 locus offers advantages over the typing of other highly polymorphic VNTR loci by Southern blotting. With routine Southern blotting, the resolution of alleles that differ by one to a few repeat units may not possible. Therefore, the alleles from a sample population form a quasi-continuous distribution of allele sizes (Budowle et al., in press). However, the alleles associated with the D1S80 locus are resolved into discrete entities using the AMP-FLP analytical technique. This greatly reduces the chance of measurement error. In fact, typing of D1S80 AMP-FLP profiles now is similar to that used for conventional protein genetic marker systems. The efficiency of amplification or yield of PCR products is related to the length of the target site between the primers. For example, at the D17S30 (also designated D17S5) locus it was observed by Horn et al. (1989) that larger alleles could be amplified to a significantly less extent than smaller ones. However, all AMP-FLP D1S80 alleles examined to date are less than 700 bp in length. There is no apparent difference in band intensity between the largest (number 17) and the smallest (number 1) alleles (fig. 1). Thus, AMP-FLP analysis of D1S80 permits correct genotyping, not just phenotyping, of VNTR profiles. This is in contrast to the situation in RFLP analysis via Southern blotting, where correct genotyping may not always be possible. Larger DNA fragments, which contain more repeat sequences, are more readily detectable by hybridization assays than smaller fragments. Thus, some small-sized VNTR alleles may go undetected by RFLP analysis. Also, small-sized alleles can migrate off the end of the gel and therefore be undetectable (whereas, in AMP-FLP analysis, because of the versatility of the AMP-FLP gels, small-sized alleles need not migrate off the gel). Thus, single-band patterns, detected by RFLP analysis via Southern blotting, may or may not be true homozygotes (Budowle et al., in press).

Table 3

D1S80 Allele Frequencies from 99 Unrelated Caucasians

Allele	95% Lower Confidence Limit[a]	Point Estimate Frequency	95% Upper Confidence Limit[a]
1	.212	.293	.389
2	.002	.010	.055
3	.006	.020	.106
4	.008	.025	.078
5	.019	.045	.071
6	.006	.020	.071
7	.244	.328	.426
8	.013	.035	.093
9	.003	.015	.063
10	.025	.056	.120
11	.031	.066	.133
12	.025	.056	.120
13	.002	.010	.055
14	.000	.000	.047[b]
15	.002	.010	.055
16	.001	.005	.047
17	.001	.005	.047

[a] Confidence limits were calculated according to Goodman (1965).

[b] Value cannot be determined; therefore, an upper confidence limit of one observation (or .047 frequency) was used.

Since alleles are resolved discretely, there is little or no measurement error, and correct genotyping is permitted, the conventional formula of the Hardy-Weinberg rule can be applied to assess the goodness of fit of the distribution of genotypes for D1S80 ($\chi^2 = 2.50$, df = 1; $.100 < P < .250$; the Hardy-Weinberg formulation was calculated by comparing observed and expected genotypes [table 2]; all classes with less than four events were pooled). Therefore, it can be stated that the alleles associate randomly with each other at this locus and there is no detectable population heterogeneity. Additional evidence that the alleles associate randomly can be obtained from the allele frequency distributions. For example, using the allele frequency distribution shown in table 3, the expected number of distinct homozygote and heterozygote genotypes (and standard errors) under the Hardy-Weinberg model (see Chakraborty et al. [1988] for the theory) were computed. In the sample of 99 individuals, three homozygote genotypes and 34 different heterozygote genotypes (table 2) were observed, and the observed values are in close agreement with the expected values (3.37 ± 1.03 and 30.84 ± 3.56, respectively). In addition, using the theory described by Chakraborty et al. (1988) and Chakraborty (1990) the expected heterozygosity ($79.7\% \pm 8.9\%$) was obtained for the given observed number of alleles (16) in the sample. The expected value is almost identical to the observed heterozygosity (80.8%) in the sample. Alternatively, if the heterozygosity were fixed at the observed level, 13.3 ± 3.1 alleles would be expected for the 198 chromosomes sampled at this locus; this in turn shows that the observed number of alleles (16) is in close agreement with this expected value ($P = .236$, using the theory of Chakraborty [1990]). Since the classical Hardy-Weinberg test is known to lack adequate statistical power for detecting population heterogeneity, particularly when the number of alleles is as large as found in the context of VNTR studies (Ward and Singh 1970; Emigh 1980), these additional data provide further confidence that the VNTR polymorphism at the D1S80 locus in a Caucasian population satisfies the basic population genetic premises to make it useful for forensic applications. AMP-FLP analysis of D1S80 and similar VNTR loci may, therefore, be useful as models for population genetics issues for the more complex VNTR profiles detected by RFLP analysis via Southern blotting.

In conclusion, a simple, discontinuous, horizontal PAGE approach followed by silver-staining techniques was used to type AMP-FLPs from the D1S80 locus. Silver staining of AMP-FLP profiles offers an inexpensive, nonmutagenic assay that provides a permanent record of the actual electrophoretic separation (unlike ethidium

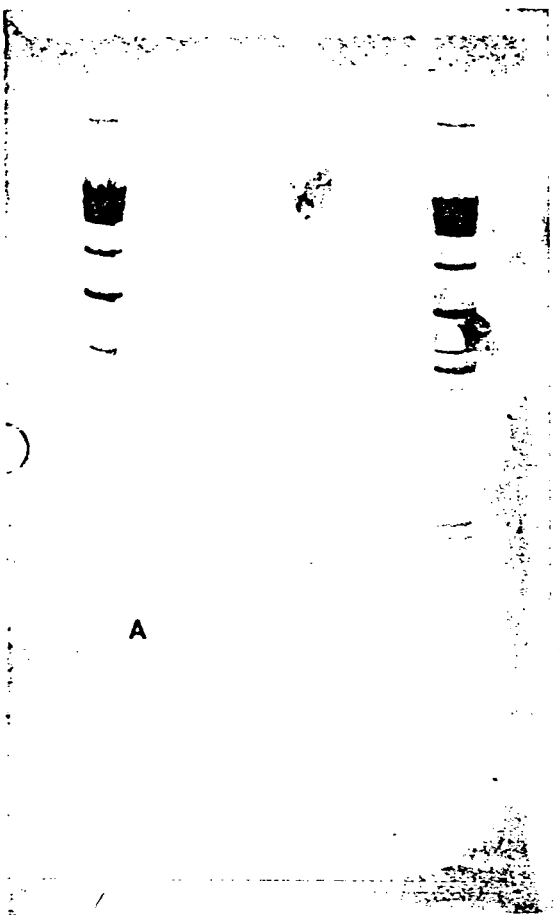


Figure 4 Silver-stained AMP-FLP gel used to generate the autoradiogram in fig. 3. Although the patterns are weak, there is sufficient residual DNA contained within the polyacrylamide gel for detection by silver staining.

bromide staining). The technique makes it possible to obtain discrete allelic data and to correctly genotype VNTR profiles for D1S80. Therefore, the conventional approaches for establishing goodness of fit of the phenotype distributions in population sample can be applied with confidence.

Acknowledgments

This is publication number 90–06 of the Laboratory Division of the Federal Bureau of Investigation. Names of commercial manufacturers are provided for identification only and inclusion does not imply endorsement by the Federal Bureau of Investigation. Parts of the statistical analyses were supported by grant GM41399 (to R.C.) from the National Institutes of Health.

References

Allen RC (1980) Rapid isoelectric focusing and detection of nanogram amounts of proteins from body tissues and fluids. Electrophoresis 1:32–37

Allen RC, Graves G, Budowle B (1989) Polymerase chain reaction amplification products separated on rehydratable polyacrylamide gels and stained with silver. BioTechniques 7: 736–744

Boerwinkle E, Xiong W, Fourest E, Chan L (1989) Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: application to the apolipoprotein B 3′ hypervariable region. Proc Natl Acad Sci USA 86: 212–216

Budowle B, Allen RC. Discontinuous polyacrylamide gel electrophoresis of DNA fragments. In: Mathew, C (ed) Methods in molecular biology—molecular biology in medicine, vol. 7. Humana, London (in press)

Budowle B, Baechtel FS (1990) Modifications to improve the effectiveness of restriction fragment length polymorphism typing. Appl Theor Electrophoresis 1:181–187

Budowle B, Giusti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, Presley LA, et al. Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic comparisons. Am J Hum Genet (in press)

Chakraborty R (1990) Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. Am J Hum Genet 47:87–94

Chakraborty R, Smouse PE, Neel JV (1988) Population amalgamation and genetic variation: observations on artificially agglomerated tribal populations of Central and South America. Am J Hum Genet 43:709–725

Emigh TH (1980) A comparison of tests for Hardy-Weinberg equilibrium. Biometrics 36:627–642

Feinberg AP, Vogelstein B (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 132:6–13

——(1984) Addendum. Anal Biochem 137:266–267

Goodman LA (1965) On simultaneous confidence intervals for multinomial distributions. Technometrics 7:247–254

Hochstrasser D, Patchornik A, Merril C (1988) Development of polyacrylamide gels that improve the separation of proteins and their detection by silver staining. Anal Biochem 173:412–423

Horn GT, Richards B, Klinger KW (1989) Amplification of a highly polymorphic VNTR segment by the polymerase chain reaction. Nucleic Acids Res 17:2140

Kasai K, Nakamura Y, White R. Amplification of a VNTR locus by the polymerase chain reaction (PCR). In: Proceedings of an international symposium on the forensic aspects of DNA analysis. Government Printing Office, Washington, DC (in press)

Ludwig EH, Friedl W, McCarthy BJ (1989) High-resolution analysis of a hypervariable region in the human apolipoprotein B gene. Am J Hum Genet 45:458–464

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp 438–454

Nakamura Y, Carlson M, Krapcho V, White R (1988) Isolation and mapping of a polymorphic DNA sequence (pMCT118) on chromosome 1p (D1S80). Nucleic Acids Res 16:9364

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N (1985) Enzymatic amplification of beta-globin genomic sequences and restriction analysis for diagnosis of sickle cell anemia. Science 230:1350–1354

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517

Ward RH, Sing CF (1970) A consideration of the power of the χ^2 test to detect inbreeding effects in natural populations. Am Nat 104:355–363

Am. J. Hum. Genet. 48:137–144, 1991

Analysis of the VNTR Locus DIS80 by the PCR Followed by High-Resolution PAGE

Bruce Budowle,* Ranajit Chakraborty,† Alan M. Giusti,* Arthur J. Eisenberg,‡ and Robert C. Allen§

*Forensic Science Research and Training Center, Laboratory Division, Federal Bureau of Investigation Academy, Quantico, VA; †Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston; ‡DNA Identity Laboratory, Department of Pathology, Texas College of Osteopathic Medicine, Fort Worth; and §Department of Pathology, Medical University of South Carolina, Charleston

Summary

Allelic data for the D1S80 locus was obtained by using the PCR and subsequent analysis with a high-resolution, horizontal PAGE technique and silver staining. Compared with RFLP analysis of VNTR loci by Southern blotting, the approach described in this paper offers certain advantages: (1) discrete allele resolution, (2) minimal measurement error, (3) correct genotyping of single-band VNTR patterns, (4) a nonisotopic assay, (5) a permanent record of the electrophoretic separation, and (6) reduced assay time. In a sample of 99 unrelated Caucasians, the D1S80 locus demonstrated a heterozygosity of 80.8% with 37 phenotypes and 16 alleles. The distribution of genotypes is in agreement with expected values according to the Hardy-Weinberg equilibrium. Furthermore, the observed number of alleles and the level of heterozygosity, obtained through the protocol described here, were congruent with each other in accordance with the expectation of a mutation-drift equilibrium model for a single, homogeneous, random-mating population. Therefore, the analysis of D1S80 and similar VNTR loci by amplified fragment length polymorphism (AMP-FLP) may prove useful as models for population genetic issues for VNTR loci analyzed by RFLP typing via Southern blotting.

Introduction

Identity tests, as performed in the fields of paternity and forensics, rely on the detection of genetic differences among individuals. At present, highly polymorphic loci whose alleles are the result of VNTRs are the most informative genetic markers for genetic characterization. Although extremely effective for VNTR analyses, the RFLP methodology via Southern blotting (Southern 1975) is time consuming and requires an isotopic assay to achieve the sensitivity necessary to detect VNTR alleles in samples containing as little as 10–50 ng of human DNA samples (Budowle and Baechtel 1990). Additionally, because of the inability of the RFLP technology to resolve discretely the alleles of most VNTR loci, statistical analyses that are different from those used for traditional genetic marker systems have been required (Budowle et al., in press).

The PCR (Saiki et al. 1985) offers a viable alternative to RFLP analysis of VNTR loci, particularly in situations where limited quantities of DNA are available. The use of the PCR can obviate the need for isotopic detection and reduce assay time and cost. With appropriate VNTR loci and high-resolution discontinuous buffer electrophoretic systems in polyacrylamide gels (Allen et al. 1989; Budowle and Allen 1990), amplification of specific DNA sequences by the PCR could prove useful for identity testing, population genetics, and disease susceptibility studies. In fact, the D17S30 (also designated D17S5) locus (Horn et al. 1989) and the 3′ hypervariable region of the apolipoprotein B gene (Boerwinkle et al. 1989; Ludwig et al. 1989) have been

Received May 15, 1990; revision received August 24, 1990.

Address for correspondence and reprints: Bruce Budowle, Forensic Science Research and Training Center, Federal Bureau of Investigation Academy, Quantico, VA 22135.

analyzed using the PCR and subsequent electrophoretic separation of the amplified fragments.

This paper describes the results of the analysis of PCR-amplified products of the VNTR locus D1S80 (Nakamura et al. 1988). The procedure resolves alleles of D1S80 into discrete entities, uses an inexpensive silver stain for detection, and provides a permanent record of the electrophoretic separation. With an analytical system that enables resolution of discrete alleles and therefore permits correct genotyping of VNTR profiles, it will now be possible to apply the conventional formula of the Hardy-Weinberg rule (i.e., χ^2 analysis on observed and expected genotype classes). This will allow for an evaluation of the goodness of fit of the genotype distributions of the particular VNTR locus for a sample population. Moreover, with the discrete resolution of alleles it will be possible to evaluate the appropriateness of classical population genetic models of allele frequency distributions at this locus to validate an assumption of genetic homogeneity of the population from which the sample is derived.

Material and Methods

Whole blood was obtained in EDTA Vacutainer tubes by venipuncture from 100 unrelated Caucasian donors at the FBI Academy. The DNA was extracted as described previously (Budowle and Baechtel 1990). Purified DNA from a two-generation (10 individuals) and a four-generation family (18 individuals) was provided by M. Skolnick (University of Utah, Salt Lake City, UT).

Amplification of D1S80 was achieved using the primers described by Kasai et al. (in press). The primers were 5′-GAAACTGGCCTCCAAACACTGCCCGCCG-3′ and 5′-GTCTTGTTGGAGATGCACGTGCCCCTT-GC-3′. Each sample that was amplified contained 100 ng DNA, 10 mM Tris-Cl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 0.01% gelation, 2.5 units of Amplitaq™ DNA polymerase (Perkin Elmer–Cetus), 1 μM of each primer, and 200 μM of each dNTP. The total volume of each sample was 50 μl. Each sample was overlaid with 50 μl of mineral oil. The PCR was carried out in a Perkin-Elmer Thermocycler for 25 cycles. Each cycle consisted of 1 min at 95°C for denaturation, 1 min at 65°C for primer annealing, and 8 min at 70°C for primer extension. After amplification, the mineral oil was removed and the samples were stored at either 4°C or −20°C prior to electrophoretic analysis.

Ultrathin-layer polyacrylamide gels (5% T, 3% C; 400 μm thick) were cast onto Gelbond (FMC, Rockland, ME) using the flap technique (Allen 1980). The gels were cross-linked with piperazine diacrylamide (Hochstrasser et al. 1988) (Bio-Rad Laboratories, Richmond, CA). All gels contained 7.1% glycerol and 33 mM Tris-sulfate buffer, pH 9.0. If rehydratable polyacrylamide gels were used (Allen et al. 1989; Budowle and Allen 1990), they were rehydrated in a solution containing 33 mM Tris-sulfate, pH 9.0, and 7.1% glycerol. The trailing ion, contained in 2% (wt/vol) agarose plugs, was 0.14 M Tris-borate, pH 9.0. Bromophenol blue (a final concentration of 0.01%) was added to the electrode buffer to serve as a dye marker for the discontinuous buffer boundary. The electrophoretic setup was similar to that described by Allen et al. (1989) and Budowle and Allen (1990). The distance between the edges of the agarose plugs was 10 cm. The amplified fragment length polymorphisms (AMP-FLPs) of D1S80 were absorbed into fiberglass applicator tabs (2.5 × 5.0 mm, Pharmacia-LKB, catalog no. 1850-901), lightly blotted, and applied to the gel surface 1 cm from the cathode. The conditions for electrophoretic separation were similar to those described previously for rehydratable polyacrylamide gels (Allen et al. 1989). Electrophoretic separation was stopped when the bromophenol blue dye front reached the anodal wick. Following electrophoresis, the gels were stained with silver, according to the conditions described in table 1, so the pattern could be visualized directly.

Hybridization analysis of the PCR-amplified products (or AMP-FLPs) subsequent to electrophoresis in the ultrathin-layer polyacrylamide gels was accomplished using a passive blotting procedure. After electrophoresis, the DNA in the gel was denatured by washing the gel in 0.4 M NaOH for 5 min. A nylon membrane (Zeta Probe, Bio-Rad Laboratories, Richmond, CA), prewetted in 0.4 M NaOH, was placed directly on the gel surface, and, subsequently, a blot pad (BRL, Gaithersburg, MD) was placed on the membrane. Transfer time was 1 h at ambient temperature. After transfer, the membrane was washed briefly in a solution containing 2 × SSC (20 × SSC = 1,753 g NaCl and 88.2 g sodium citrate/l, pH 7.0) (Maniatis et al. 1982) and 0.2 M Tris, pH 7.5. The membrane was blotted between two Whatmann 1 MM papers and baked in an oven at 80°C for 30 min. The membrane was wrapped in plastic wrap and stored at −20°C. The probe pMCT118 (for locus D1S80) was provided by Y. Nakamura and R. White (Howard Hughes Medical Institute, Salt Lake City, UT). Random primer labeling was accomplished according to the manufacturer's in-

Table I

Protocol for Silver-Staining AMP-FLP Gels

Step	Time
Place gel in 10% ethanol solution	5 min
Oxidize gel in 1% nitric acid solution	3 min
Rinse gel in distilled water	A few seconds
Place gel in 0.012 M silver nitrate solution	20 min
Decant silver nitrate and rinse gel in distilled water	A few seconds
Reduce gel in a solution containing 0.28 M sodium carbonate (anhydrous) and 0.019% formalin; several changes of reducing solution may be necessary; the solution should be changed when it turns brown	Will depend on desired intensity of image; image develops before eye
Stop reduction process with 10% glacial acetic acid	2 min
Place gel in distilled water	2 min
Air dry gel for permanent record	. . .

structions contained within the BRL Random Primer DNA labeling system kit or according to the method of Feinberg and Vogelstein (1983, 1984). Hybridization and stringency washes were carried out according to the method of Budowle and Baechtel (1989). The labeled DNA duplex was detected by autoradiography using Kodak XAR film and Dupont Cronex Lightning Plus intensifying screens.

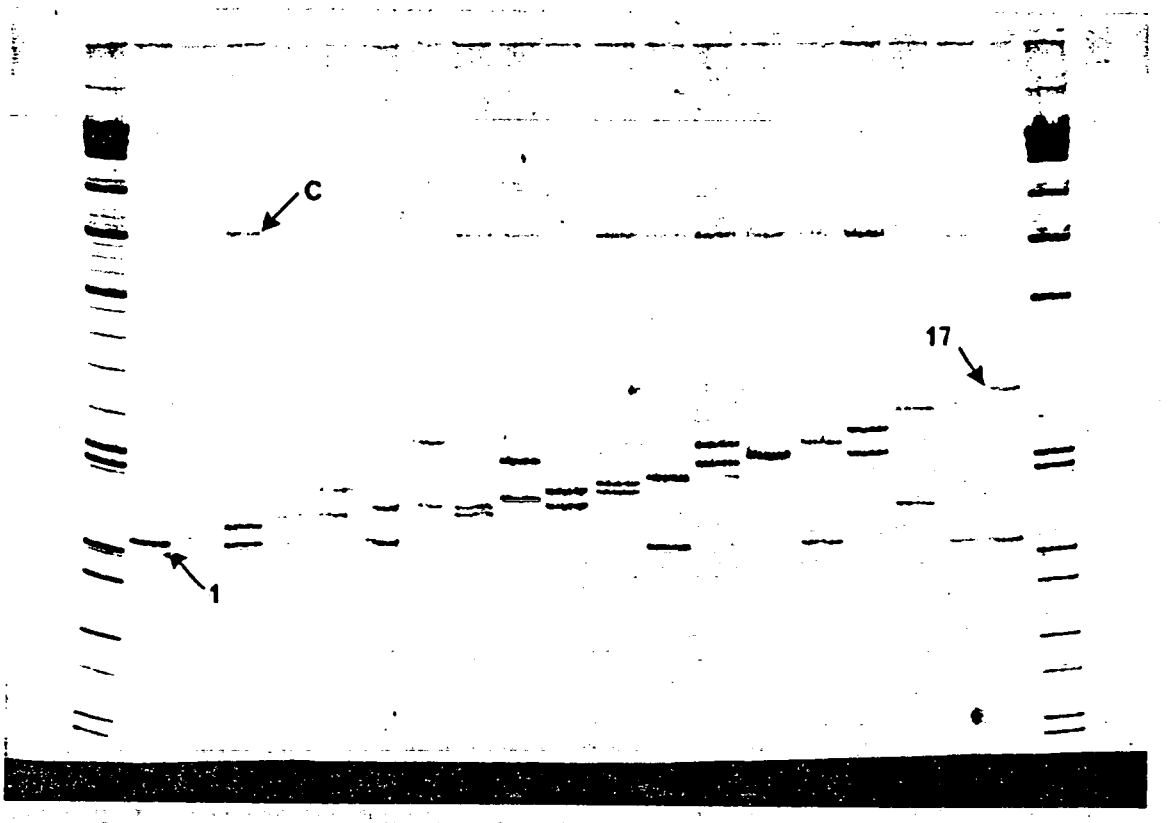


Figure 1 Silver-stained AMP-FLP gel displaying D1S80 profiles. The D1S80 types from left to right are 1-1, 2-1, 3-1, 4-1, 7-4, 5-1, 12-5, 5-4, 10-6, 7-5, 8-7, 9-1, 12-10, 11-11, 12-1, 13-11, 15-5, 16-1, and 17-1. C = constant band. The size standards are a combination of the 1-kb and 123-bp ladder (BRL, Gaithersburg, MD). The cathode is at the top.

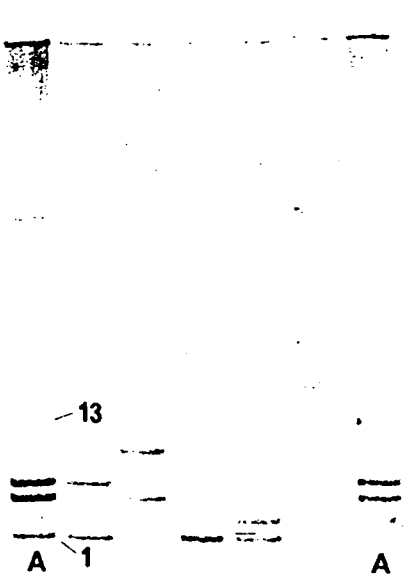


Figure 2 Silver-stained AMP-FLP gel displaying a ladder of composite alleles and D1S80 profiles. The ladder (A) is composed of alleles 1, 3, 5, 7, 9, 11 and 13. The D1S80 types from left to right are 7-1, 10-5, 1-1, 3-1, and 15-1. The cathode is at the top.

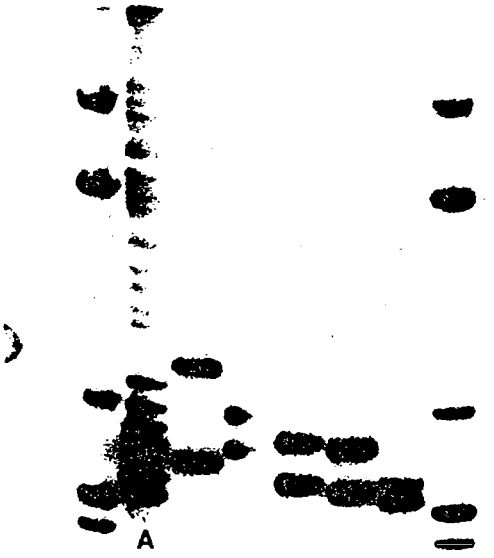


Figure 3 Autoradiogram of an AMP-FLP gel displaying the ladder of composite alleles (A) and D1S80 types. The pMCT118 probe was provided by Ray White and his colleagues (Howard Hughes Medical Institute, Salt Lake City, UT).

Results

Figure 1 shows that AMP-FLP analysis of D1S80 can be performed using the techniques described in this paper. Sixteen different alleles were observed in 99 unrelated Caucasians (one sample did not amplify by PCR). Each allele was completely resolved based on increments of the repeat unit of the VNTR locus. The length of the repeat unit has been determined to be 16 base pairs (Y. Nakamura, Howard Hughes Medical Institute, and T. Holm, GenMark, Salt Lake City, UT, personal communications). The alleles have been designated 1–17 (allele 14 has not yet been observed), where allele 1 is the smallest in length and allele 17 is the largest in length. (It should be noted that the allele designations are temporary and will eventually be based on the number of repeats.)

With this system the alleles can be designated specifically without determining base-pair size. Unknown samples can be compared with a "ladder" consisting of a composite of D1S80 alleles; thus, allele designations are much easier and measurement error is greatly reduced (fig. 2).

It should be noted that a constant (or monomorphic) band appears in each amplified sample (indicated by the arrow at point C in fig. 1). The band does not hybridize with the pMCT118 probe (fig. 3) and thus is a sequence unrelated to D1S80. However, shifts in the position of the constant band can indicate unaligned positioning of the sample tabs at the sample origin. Therefore, the constant band can serve an important function as an internal marker to minimize errors in AMP-FLP typing of D1S80.

The distributions of observable phenotypes and allelic frequencies for D1S80 in a Caucasian population

Table 2

Distribution of D1S80 Genotypes from 99 Unrelated Caucasians

Genotype	Number Observed
1-1	7
2-1	1
3-1	3
4-1	2
5-1	1
6-1	2
7-1	24
8-1	1
9-1	3
10-1	1
11-1	2
12-1	2
16-1	1
17-1	1
7-2	1
8-3	1
5-4	1
7-4	1
10-4	1
7-5	2
10-5	1
11-5	2
12-5	1
15-5	1
10-6	2
7-7	10
8-7	3
10-7	3
11-7	4
12-7	5
13-7	1
15-7	1
10-8	1
12-8	1
12-10	2
11-11	2
13-11	1

sample of 99 unrelated individuals are shown in tables 2 and 3, respectively. The observed heterozygosity is 80.8%. The distribution of the phenotypes is in Hardy-Weinberg equilibrium ($\chi^2 = 2.50$; df = 1; $.100 < P < .250$; Hardy-Weinberg formulation was calculated by comparing observed and expected genotypes; all classes with less than four events were pooled). Further, although not extremely informative (because of limited variation among family members), the two families demonstrated Mendelian inheritance of the AMP-FLP alleles (data not shown).

While hybridization assays do not appear to be necessary for routinely typing D1S80 AMP-FLP profiles, it may be desirable, at times, to use a probe to confirm that the AMP-FLPs truly represent the described locus or to increase the level of sensitivity of detection provided by silver staining. Figure 3 shows that the blotting approach can be used. As expected, after the PCR there should be more than adequate quantities of DNA for hybridization analysis. Although the patterns are weak, there can still be enough residual DNA left in the polyacrylamide gel for detection by silver staining (fig. 4).

Discussion

AMP-FLP analysis of the D1S80 locus offers advantages over the typing of other highly polymorphic VNTR loci by Southern blotting. With routine Southern blotting, the resolution of alleles that differ by one to a few repeat units may not possible. Therefore, the alleles from a sample population form a quasi-continuous distribution of allele sizes (Budowle et al., in press). However, the alleles associated with the D1S80 locus are resolved into discrete entities using the AMP-FLP analytical technique. This greatly reduces the chance of measurement error. In fact, typing of D1S80 AMP-FLP profiles now is similar to that used for conventional protein genetic marker systems. The efficiency of amplification or yield of PCR products is related to the length of the target site between the primers. For example, at the D17S30 (also designated D17S5) locus it was observed by Horn et al. (1989) that larger alleles could be amplified to a significantly less extent than smaller ones. However, all AMP-FLP D1S80 alleles examined to date are less than 700 bp in length. There is no apparent difference in band intensity between the largest (number 17) and the smallest (number 1) alleles (fig. 1). Thus, AMP-FLP analysis of D1S80 permits correct genotyping, not just phenotyping, of VNTR profiles. This is in contrast to the situation in RFLP analysis via Southern blotting, where correct genotyping may not always be possible. Larger DNA fragments, which contain more repeat sequences, are more readily detectable by hybridization assays than smaller fragments. Thus, some small-sized VNTR alleles may go undetected by RFLP analysis. Also, small-sized alleles can migrate off the end of the gel and therefore be undetectable (whereas, in AMP-FLP analysis, because of the versatility of the AMP-FLP gels, small-sized alleles need not migrate off the gel). Thus, single-band patterns, detected by RFLP analysis via Southern blotting, may or may not be true homozygotes (Budowle et al., in press).

Table 3

D1S80 Allele Frequencies from 99 Unrelated Caucasians

Allele	95% Lower Confidence Limit[a]	Point Estimate Frequency	95% Upper Confidence Limit[a]
1	.212	.293	.389
2	.002	.010	.055
3	.006	.020	.106
4	.008	.025	.078
5	.019	.045	.071
6	.006	.020	.071
7	.244	.328	.426
8	.013	.035	.093
9	.003	.015	.063
10	.025	.056	.120
11	.031	.066	.133
12	.025	.056	.120
13	.002	.010	.055
14	.000	.000	.047[b]
15	.002	.010	.055
16	.001	.005	.047
17	.001	.005	.047

[a] Confidence limits were calculated according to Goodman (1965).

[b] Value cannot be determined; therefore, an upper confidence limit of one observation (or .047 frequency) was used.

Since alleles are resolved discretely, there is little or no measurement error, and correct genotyping is permitted, the conventional formula of the Hardy-Weinberg rule can be applied to assess the goodness of fit of the distribution of genotypes for D1S80 ($\chi^2 = 2.50$, df = 1; $.100 < P < .250$; the Hardy-Weinberg formulation was calculated by comparing observed and expected genotypes [table 2]; all classes with less than four events were pooled). Therefore, it can be stated that the alleles associate randomly with each other at this locus and there is no detectable population heterogeneity. Additional evidence that the alleles associate randomly can be obtained from the allele frequency distributions. For example, using the allele frequency distribution shown in table 3, the expected number of distinct homozygote and heterozygote genotypes (and standard errors) under the Hardy-Weinberg model (see Chakraborty et al. [1988] for the theory) were computed. In the sample of 99 individuals, three homozygote genotypes and 34 different heterozygote genotypes (table 2) were observed, and the observed values are in close agreement with the expected values (3.37 ± 1.03 and 30.84 ± 3.56, respectively). In addition, using the theory described by Chakraborty et al. (1988) and Chakraborty (1990) the expected heterozygosity ($79.7\% \pm 8.9\%$) was obtained for the given observed number of alleles (16) in the sample. The expected value is almost identical to the observed heterozygosity (80.8%) in the sample. Alternatively, if the heterozygosity were fixed at the observed level, 13.3 ± 3.1 alleles would be expected for the 198 chromosomes sampled at this locus; this in turn shows that the observed number of alleles (16) is in close agreement with this expected value ($P = .236$, using the theory of Chakraborty [1990]). Since the classical Hardy-Weinberg test is known to lack adequate statistical power for detecting population heterogeneity, particularly when the number of alleles is as large as found in the context of VNTR studies (Ward and Singh 1970; Emigh 1980), these additional data provide further confidence that the VNTR polymorphism at the D1S80 locus in a Caucasian population satisfies the basic population genetic premises to make it useful for forensic applications. AMP-FLP analysis of D1S80 and similar VNTR loci may, therefore, be useful as models for population genetics issues for the more complex VNTR profiles detected by RFLP analysis via Southern blotting.

In conclusion, a simple, discontinuous, horizontal PAGE approach followed by silver-staining techniques was used to type AMP-FLPs from the D1S80 locus. Silver staining of AMP-FLP profiles offers an inexpensive, nonmutagenic assay that provides a permanent record of the actual electrophoretic separation (unlike ethidium

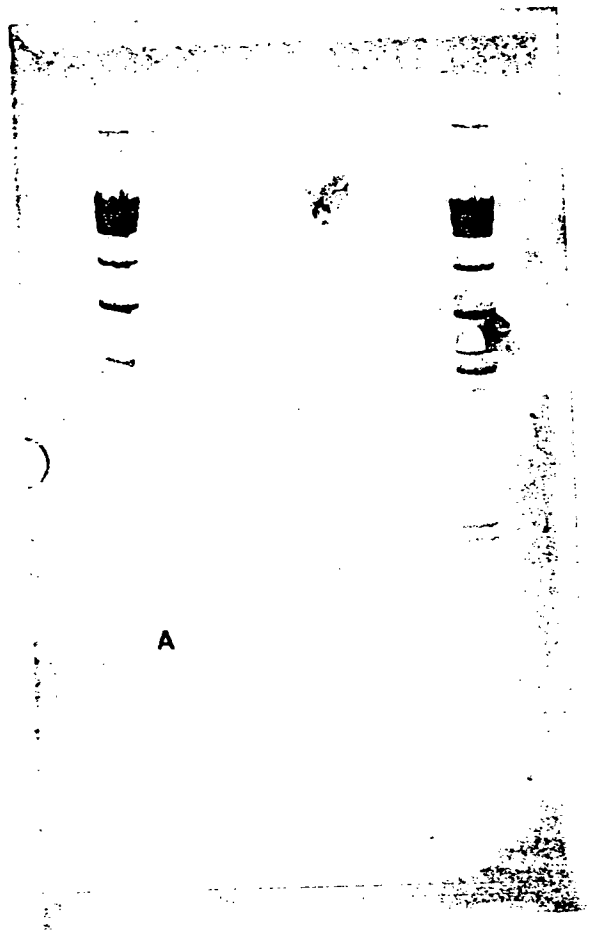


Figure 4 Silver-stained AMP-FLP gel used to generate the autoradiogram in fig. 3. Although the patterns are weak, there is sufficient residual DNA contained within the polyacrylamide gel for detection by silver staining.

bromide staining). The technique makes it possible to obtain discrete allelic data and to correctly genotype VNTR profiles for D1S80. Therefore, the conventional approaches for establishing goodness of fit of the phenotype distributions in population sample can be applied with confidence.

Acknowledgments

This is publication number 90–06 of the Laboratory Division of the Federal Bureau of Investigation. Names of commercial manufacturers are provided for identification only and inclusion does not imply endorsement by the Federal Bureau of Investigation. Parts of the statistical analyses were supported by grant GM41399 (to R.C.) from the National Institutes of Health.

References

Allen RC (1980) Rapid isoelectric focusing and detection of nanogram amounts of proteins from body tissues and fluids. Electrophoresis 1:32–37

Allen RC, Graves G, Budowle B (1989) Polymerase chain reaction amplification products separated on rehydratable polyacrylamide gels and stained with silver. BioTechniques 7: 736–744

Boerwinkle E, Xiong W, Fourest E, Chan L (1989) Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: application to the apolipoprotein B 3′ hypervariable region. Proc Natl Acad Sci USA 86: 212–216

Budowle B, Allen RC. Discontinuous polyacrylamide gel electrophoresis of DNA fragments. In: Mathew, C (ed) Methods in molecular biology—molecular biology in medicine, vol. 7. Humana, London (in press)

Budowle B, Baechtel FS (1990) Modifications to improve the effectiveness of restriction fragment length polymorphism typing. Appl Theor Electrophoresis 1:181–187

Budowle B, Giusti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, Presley LA, et al. Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic comparisons. Am J Hum Genet (in press)

Chakraborty R (1990) Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. Am J Hum Genet 47:87–94

Chakraborty R, Smouse PE, Neel JV (1988) Population amalgamation and genetic variation: observations on artificially agglomerated tribal populations of Central and South America. Am J Hum Genet 43:709–725

Emigh TH (1980) A comparison of tests for Hardy-Weinberg equilibrium. Biometrics 36:627–642

Feinberg AP, Vogelstein B (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 132:6–13

———(1984) Addendum. Anal Biochem 137:266–267

Goodman LA (1965) On simultaneous confidence intervals for multinomial distributions. Technometrics 7:247–254

Hochstrasser D, Patchornik A, Merril C (1988) Development of polyacrylamide gels that improve the separation of proteins and their detection by silver staining. Anal Biochem 173:412–423

Horn GT, Richards B, Klinger KW (1989) Amplification of a highly polymorphic VNTR segment by the polymerase chain reaction. Nucleic Acids Res 17:2140

Kasai K, Nakamura Y, White R. Amplification of a VNTR locus by the polymerase chain reaction (PCR). In: Proceedings of an international symposium on the forensic aspects of DNA analysis. Government Printing Office, Washington, DC (in press)

Ludwig EH, Friedl W, McCarthy BJ (1989) High-resolution analysis of a hypervariable region in the human apolipoprotein B gene. Am J Hum Genet 45:458–464

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp 438–454

Nakamura Y, Carlson M, Krapcho V, White R (1988) Isolation and mapping of a polymorphic DNA sequence (pMCT118) on chromosome 1p (D1S80). Nucleic Acids Res 16:9364

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N (1985) Enzymatic amplification of beta-globin genomic sequences and restriction analysis for diagnosis of sickle cell anemia. Science 230:1350–1354

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517

Ward RH, Sing CF (1970) A consideration of the power of the χ^2 test to detect inbreeding effects in natural populations. Am Nat 104:355–363

Genetic Structure of the Populations Migrating from San Luis Potosi and Zacatecas to Nuevo León in Mexico

RICARDO M. CERDA-FLORES,[1,2] GAUTAM K. KSHATRIYA,[2] SARA A. BARTON,[2] CARLOS H. LEAL-GARZA,[1] RAUL GARZA-CHAPA,[1] WILLIAM J. SCHULL,[2] AND RANAJIT CHAKRABORTY[2]

Abstract The Mexicans residing in the Monterrey metropolitan area in Nuevo León, Mexico, were grouped by generation and birthplace [Monterrey Metropolitan Area (MMA), San Luis Potosi (SLP), and Zacatecas (ZAC)] of the four grandparents to determine the extent of genetic variation within this population and the genetic differences, if any, between the natives living in the MMA and the immigrant populations from SLP and ZAC. Nine genetic marker systems were analyzed. The genetic distance analysis indicates that SLP and ZAC are similar to the MMA, irrespective of birthplace and generation. Gene diversity analysis (G_{ST}) suggests that more than 96% of the total gene diversity (H_T) can be attributed to individual variation within the population. The genetic admixture analysis suggests that the Mexicans of the MMA, SLP, and ZAC, stratified by birthplace and generation, have received a predominantly Spanish contribution (78.5%), followed by a Mexican Indian contribution (21.5%). Similarly, admixture analysis, conducted on the population of Nuevo León and stratified by generation, indicates a substantial contribution from the MMA (64.6%), followed by ZAC (22.1%) and SLP (13.3%). Finally, we demonstrate that there is no nonrandom association of alleles among the genetic marker systems (i.e., no evidence of gametic disequilibrium) despite the Mestizo origin of this population.

The state of Nuevo León in northeastern Mexico (Figure 1) has an area of 64,555 km^2, and in 1990 had a population of 4,492,500 inhabitants. The age distribution of this Mestizo population indicates that 72% of the total population is under 30 years, 23% is between 30 and 59 years, and only 5% exceeds 59 years of age (Dirección de Estadistica y Procesamiento de Datos del Gobierno de Nuevo León, 1977).

The Monterrey Metropolitan Area (MMA) is located in the central western section of the state of Nuevo León and has an area of 2118 km^2.

[1]Subjefatura de Investigacion Cientifica, Instituto Mexicano del Seguro Social, Unidad de Investigacion Biomedica del Noreste, Apartado Postal 020-E, 64720 Monterrey, Nuevo León, Mexico.
[2]Center for Demographic and Population Genetics, Graduate School of Biomedical Sciences, University of Texas, PO Box 20334, Houston, Texas 77225.

Human Biology, June 1991, Vol. 63, No. 3, pp. 309–327.

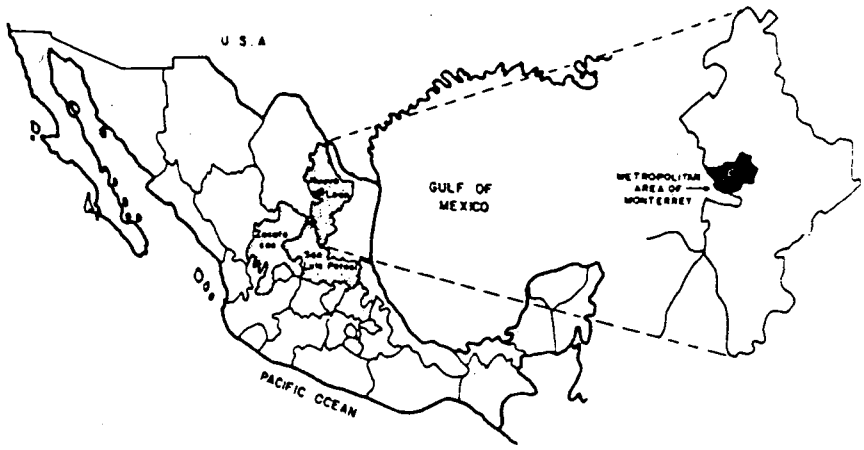


Figure 1. Location of the state of Nuevo León in Mexico.

In 1940 Nuevo León had 541,147 inhabitants, of whom 39% resided in the MMA, whereas in 1960 the population was 1,078,848 and 67% resided in the MMA. At present, nearly 80% of the state's population is concentrated in the MMA (Dirección de Estadistica y Procesamiento de Datos, 1977). This increase is principally due to the immigration of people from several southern states of Mexico to the MMA since 1910. For example, in 1960 alone 400,000 people migrated to the MMA mainly from the states of San Luis Potosi (28%), Coahuila (24%), Tamaulipas (16%), Zacatecas (11%), and the Federal District (3%) and some foreign countries, particularly the United States (Montemayor-Hernández 1971). The explanation for this immigration is that the MMA is an important industrial, commercial, and educational zone in Mexico that attracts people in search of jobs, livelihood, business, and education.

In the present study the Mexican population that resides in the MMA is grouped by generation and birthplace of the four grandparents. Our aims are (1) to study the genetic variation among the Mexican population of the MMA, stratified by generation and birthplace, and to determine whether the immigrant population groups, for example, San Luis Potosi (SLP) and Zacatecas (ZAC), are different from the native Mexicans whose four grandparents were born in the MMA; (2) to compute the contribution of the ancestral populations (i.e., Spanish and Indians) to the Mexicans of the MMA, stratified by generation and birthplace, and of the MMA, SLP, and ZAC populations to the population of Nuevo León, stratified by generation only; (3) to study the proportion of genes received from the ancestral populations; and (4)

to demonstrate whether there is any residual effect of population mixture on the nonrandom association of alleles.

Materials and Methods

From the 4680 people residing in the MMA who were randomly interviewed at the Instituto Mexicano del Seguro Social (IMSS), Monterrey, in 1985, 207 were selected whose 4 grandparents were born in the MMA (Cerda-Flores and Garza-Chapa 1989), 151 were selected whose grandparents were born in SLP, and 153 were selected whose grandparents were born in ZAC, thereby constituting a total sample size of 511 individuals. The frequencies of the phenotypes of the blood group systems ABO, Rh, MNSs, Duffy (Fy), Kidd (Jk), Lutheran (Lu), P, Lewis (Le), and Secretor (Se) were analyzed using commercial antisera and the microplate method described by Crawford et al. (1970) and Lapinski et al. (1978).

The data were subdivided by year of birth, and the number of generations was estimated assuming a generation time of 30 years. Accordingly, 13 human generations have elapsed since the MMA was colonized in 1596 (del Hoyo 1979). We consider the last three: (1) persons born between 1896 and 1925 (generation 11), (2) persons born between 1926 and 1955 (generation 12), and (3) persons born between 1956 and 1985 (generation 13).

The statistical analysis was conducted in six parts. First, the gene frequencies for different systems were computed using the maximum likelihood method (Reed and Schull 1968). Next, the genetic distances among the Mexicans of the MMA, stratified by generation and birthplace of the grandparents, were computed by Nei's standard genetic distance (Nei 1972), and their standard errors (SEs) were computed using the method of Nei and Roychoudhury (1974). The gene frequency data were further subjected to the pairwise chi-square statistic to determine the statistical significance of the genetic distance (Nei and Roychoudhury 1974). Third, the extent of genetic variation between the subpopulations of Mexicans in the MMA (by generation and birthplace) was studied using the nested gene diversity computer program (NEGST) developed by Chakraborty (1980) and Chakraborty et al. (1982). Next, the contribution (%) of Spanish and Mexican Indians to the population groups of the MMA and of the MMA, SLP, and ZAC groups to the population of Nuevo León, stratified by three generations, was calculated following procedures detailed by Chakraborty (1985, 1986) using dihybrid and trihybrid models.

To determine whether the proportions of genes received by the Mexican subpopulations of the MMA from their ancestral sources are signifi-

cantly different from each other, we next performed a regression analysis of heterozygosity on genetic distance, as proposed by Harpending and Ward (1982). The regression equation was subjected to a test of significance, following the method of Snedecor and Cochran (1967). Finally, the computation for nonrandom association of alleles among different genetic loci was conducted following the methods of Brown et al. (1980) and Chakraborty (1981, 1984) to examine whether any residual effects of admixture remain in the current MMA population that would make this population heterogeneous.

Gene frequency data on the ancestral populations were obtained from Hanis et al. (1991) (see appendix).

Results

Genetic Distance. Table 1 provides the allele frequency estimates for the analyzed loci. Based on the allele frequency, we first determined the genetic differences within the Mexicans of the MMA, stratified by generation and birthplace of the four grandparents. We estimated Nei's standard genetic distances for all pairs of populations and their SEs (Table 2). Table 2 also indicates average heterozygosity ($\bar{H}$) among the subpopulations. The subpopulations of the MMA, arranged by generation, have an $\bar{H}$ that varies between 47.3% (generation 12) and 51.7% (generation 11) and between 45.3% (ZAC) and 49.4% (MMA) when the populations are grouped according to the birthplace of the four grandparents. The genetic distance analysis (Table 3) does not reveal any significant level of differentiation, as reflected by the pairwise chi-square statistic (Nei and Roychoudhury 1974).

Gene Diversity Analysis. Table 4 presents the hierarchical gene diversity analysis (G_{ST}) among the subpopulations of the MMA. The total average gene diversity (H_T) is 47.9%. Over 96% of H_T, computed on the basis of the 9 loci, can be attributed to individual variation within the population. However, a small contribution to the total variability (2.97%) comes from the between-birthplace variation of the four grandparents, suggesting that there is some genetic variation among the Mexicans that reside in the MMA. The generation difference in the extent of genetic variation is even smaller (0.67% of the total) and probably not of any biologic significance.

Genetic Admixture Analysis. Table 5 presents the estimated values of admixture based on nine polymorphic genetic loci. In the present investigation we considered the Mexicans of the MMA to be the product of admixture of two parental populations having Spanish and Mexican

Indian ancestry. The allele frequencies of the ancestral populations are presented in the appendix. The results of fitting the dihybrid model show that the contribution from the Spanish ancestry to the Mexicans of the MMA, stratified by birthplace, varies from 66.21% in ZAC to 82.15% in the natives of the MMA. However, the Spanish contribution, when the data are stratified by generation, varies from 52.26% in generation 13 of ZAC to 95.8% in generation 11 of the natives of the MMA.

The overall contribution of the Spanish to the Mexicans of the MMA (total) is 78.46% and, when subdivided by generation, varies from 70.47% in generation 12 to 91.14% in generation 11. Based on these results, it can be inferred that the Mexicans of the MMA, stratified by generation and birthplace, have received a predominantly Spanish contribution, followed by a Mexican Indian contribution.

The MMA population increased from 44,808 in 1900 to 112,864 in 1921, to 382,021 in 1950, and to 1,242,558 in 1970 (Dirección de Estadistica y Procesamiento de Datos, 1977). Therefore we thought it would be interesting to examine, using a trihybrid model, the contribution (%) of the MMA and the immigrant populations of SLP and ZAC to the Nuevo León population, stratified by generation, to see whether the percentage of genetic contribution and the percentage of immigration in each generation are similar. We found that the genetic contribution of the MMA population is larger than the immigrant population in the three generations (Table 6). That genetic contributions of 56.6% in generation 12 and of 43.4% in the immigrant populations are found agrees with the historical account of Montemayor-Hernández (1971), who reported that, of the 400,000 people who entered the MMA, an industrial region, 156,000 (39%) were from SLP and ZAC. From these results it can be inferred that the population of Nuevo León has received predominantly an MMA contribution (64.5%), followed by contributions from ZAC (22.1%) and SLP (13.4%).

Heterozygosity and Genetic Distance. Table 7 shows the average heterozygosity (H_i) for nine genetic loci and the genetic distance (r_{ii}) for each subpopulation along with its interlocus standard error. The regression of heterozygosity on genetic distance is consistent with linearity. $\bar{H}$ in the population pooled by generation and birthplace of grandparents (48.6% and 48.1%) does not differ significantly from the regression coefficient b (49.6% and 48.0%). This indicates that the Mexican populations of the MMA, SLP, and ZAC are similar in the proportion of the genes that they have received from the ancestral populations. This finding is consistent with the similarity of admixture proportions estimated in the previous section.

Nonrandom Association among Genetic Loci. From the available genotype data on each individual, we defined the multilocus genotype for each

Table 1. Allele Frequencies among Mexicans of the State of Nuevo León Grouped by Generation and Birthplace of the Four Grandparents

	Birthplace of Generation 11				*Birthplace of Generation 12*				*Birthplace of Generation 13*				*Total*			
System	*MMA* (*n*=45)	*SLP* (*n*=15)	*ZAC* (*n*=19)	*Total* (*n*=79)	*MMA* (*n*=34)	*SLP* (*n*=44)	*ZAC* (*n*=72)	*Total* (*n*=150)	*MMA* (*n*=128)	*SLP* (*n*=92)	*ZAC* (*n*=62)	*Total* (*n*=282)	*MMA* (*n*=207)	*SLP* (*n*=151)	*ZAC* (*n*=153)	*Total* (*n*=511)
ABO																
A1	.212	.177	.054	.165	.143	.160	.126	.140	.143	.157	.130	.145	.158	.160	.118	.146
A2	.014	.000	.000	.008	.017	.013	.024	.020	.042	.045	.038	.042	.032	.031	.026	.030
B	.093	.067	.083	.085	.045	.011	.087	.055	.036	.044	.041	.040	.050	.037	.068	.051
O	.681	.756	.863	.742	.795	.816	.763	.785	.779	.754	.791	.773	.760	.772	.788	.773
Rh																
DCE	.000	.000	.104	.000	.079	.053	.102	.072	.033	.027	.090	.041	.000	.033	.103	.037
DCe	.240	.300	.290	.339	.201	.390	.377	.300	.216	.347	.330	.285	.263	.347	.328	.288
DcE	.128	.233	.265	.271	.042	.288	.155	.124	.291	.245	.241	.265	.247	.255	.198	.209
Dce	.207	.467	.340	.172	.260	.159	.366	.290	.228	.238	.184	.218	.188	.231	.268	.239
dCE	.053	.000	.000	.163	.000	.000	.000	.012	.015	.000	.000	.010	.060	.000	.000	.026
dCe	.062	.000	.001	.000	.176	.000	.000	.082	.040	.045	.056	.044	.034	.034	.036	.050
dcE	.030	.000	.000	.000	.128	.000	.000	.078	.000	.000	.000	.000	.007	.000	.000	.025
dce	.280	.000	.001	.055	.113	.110	.000	.041	.177	.098	.099	.137	.201	.100	.067	.126
MNSs																
MS	.337	.324	.453	.360	.309	.308	.302	.301	.324	.268	.358	.313	.323	.284	.343	.317
Ms	.297	.343	.153	.273	.396	.374	.303	.349	.320	.379	.312	.338	.329	.375	.288	.331
NS	.041	.209	.205	.114	.058	.101	.226	.155	.137	.124	.118	.129	.105	.126	.180	.134
Ns	.325	.124	.189	.253	.237	.217	.169	.195	.219	.229	.212	.220	.243	.215	.189	.218

Duffy																
Fy(a)	.459	.507	.276	.413	.455	.489	.416	.444	.401	.326	.372	.367	.422	.388	.379	.396
Fy(b)	.459	.386	.489	.447	.428	.350	.359	.370	.443	.614	.347	.464	.444	.503	.369	.433
Fy	.082	.107	.235	.140	.117	.161	.225	.186	.156	.060	.281	.169	.134	.109	.252	.171
Kidd																
Jk(a)	.423	.317	.205	.344	.431	.397	.237	.322	.235	.278	.260	.254	.302	.314	.242	.287
Jk(b)+Jk	.577	.683	.795	.656	.569	.603	.763	.678	.765	.722	.740	.746	.698	.686	.758	.713
Lutheran																
Lua	.144	.317	.173	.262	.109	.160	.065	.102	.099	.134	.093	.109	.110	.158	.089	.117
Lub	.856	.683	.827	.738	.891	.840	.935	.898	.901	.866	.907	.891	.890	.842	.911	.883
P																
P1	.506	.553	.771	.564	.657	.703	.627	.576	.549	.546	.492	.535	.555	.518	.580	.551
P2+p	.494	.447	.229	.436	.343	.297	.373	.424	.451	.454	.508	.465	.445	.482	.420	.449
Lewis																
Le	.553	.684	.622	.606	.564	.746	.663	.663	.575	.598	.655	.603	.569	.645	.654	.610
le	.447	.316	.378	.394	.436	.254	.337	.337	.425	.402	.345	.397	.431	.355	.346	.390
Secretor[a]																
Se	.646	1.00	.592	.671	.758	.737	1.00	.803	.657	.717	1.00	.696	.670	.739	.842	.720
se	.354	0.00	.408	.329	.242	.263	0.00	.197	.343	.283	0.00	.304	.330	.261	.158	.280
n	20	10	14	44	21	31	44	96	83	62	42	187	124	103	100	327

a. The *n* values for Secretor are different from those for all other systems.

Table 2. Standard Genetic Distances and Average Heterozygosity within the Population of Nuevo León Grouped by Generation and Birthplace of the Four Grandparents [a]

	Generation			*Birthplace of Four Grandparents*			
Population	*11*	*12*	*13*	*MMA*	*SLP*	*ZAC*	*Total*
Generation							
11	51.71 ± 4.38						
12	14.76 ± 9.52	47.30 ± 6.35					
13	8.10 ± 5.64	8.06 ± 4.82	48.06 ± 5.95				
Birthplace of four grandparents							
MMA	5.70 ± 6.05	10.83 ± 7.95	0.60 ± 1.01	49.38 ± 5.83			
SLP	4.81 ± 3.92	3.81 ± 3.74	0.50 ± 0.77	3.47 ± 2.41	48.58 ± 5.25		
ZAC	16.02 ± 7.49	1.70 ± 1.78	5.61 ± 3.98	11.82 ± 6.66	4.49 ± 2.01	45.29 ± 6.60	
Total	6.70 ± 5.58	3.34 ± 2.43	0.28 ± 0.44	0.89 ± 1.34	0.12 ± 0.69	3.81 ± 2.57	48.56 ± 5.96

a. Values on the diagonal are the average heterozygosities expressed in percentage (e.g., generation 11 by 11 = 51.7%); below the diagonal are standard genetic distances in 10^{-3} codon differences per locus. The computations are done with nine polymorphic loci (ABO, Rh, MNSs, Duffy, Kidd, Lutheran, P, Lewis, and Secretor).

Table 3. Test of Significance of Genetic Distances among Populations of the State of Nuevo León Grouped by Generation and Birthplace of the Four Grandparents Based on the Pairwise Chi-Square

	Generation 11				*Generation 12*				*Generation 13*				*Birthplace of Four Grandparents*		
Population	*MMA*	*SLP*	*ZAC*	*Total*	*MMA*	*SLP*	*ZAC*	*Total*	*MMA*	*SLP*	*ZAC*	*Total*	*MMA*	*SLP*	*ZAC*
Generation 11															
SLP	2.43														
ZAC	2.48	2.23													
Total	1.10	1.50	1.63												
Generation 12															
MMA	1.07	2.48	2.32	2.31											
SLP	1.42	1.68	1.53	0.91	1.48										
ZAC	3.22	1.04	2.81	2.11	2.87	1.87									
Total	1.38	1.31	1.36	1.30	0.55	0.84	0.83								
Generation 13															
MMA	0.89	1.94	1.12	0.99	1.68	0.84	1.58	0.81							
SLP	1.04	1.49	1.20	0.84	1.39	0.62	1.33	0.55	0.21						
ZAC	2.07	1.38	3.00	1.72	1.98	1.35	0.72	0.76	0.92	0.77					
Total	1.06	1.73	1.08	1.08	2.16	0.63	1.21	0.64	0.05	0.08	0.62				
Birthplace of four grandparents															
MMA	0.41	1.95	2.44	0.59	2.12	0.97	1.88	0.91	0.22	0.35	1.22	0.24			
SLP	1.15	1.21	1.27	0.85	1.64	0.40	1.17	0.51	0.29	0.06	0.63	0.12	0.39		
ZAC	1.80	1.33	1.04	1.32	1.75	0.76	0.51	0.37	0.57	0.37	0.39	0.36	0.83	0.35	
Total	0.65	1.49	1.11	0.92	1.12	0.59	1.06	0.33	0.16	0.13	0.62	0.09	0.20	0.14	0.34

Pairwise chi-square matrix for all loci: d.f. = 19; $p > 0.05$ nonsignificant.

Table 4. Gene Diversity Analysis of Allele Frequency Data from Populations of Nuevo León Grouped by Generations and Birthplace of the Four Grandparents

	Relative Gene Diversity (G_{ST}) (%)			
Locus	*Within Population*	*Between Birthplace of Four Grandparents*	*Between Generation within Birthplace of Four Grandparents*	*Total Gene Diversity (H_T)*
ABO	98.71	1.11	0.18	0.371
Rh	95.74	3.51	0.75	0.778
MNSs	98.13	1.52	0.36	0.724
Duffy	97.86	1.62	0.52	0.484
Kidd	96.85	2.40	0.75	0.427
Lutheran	96.10	2.05	1.85	0.246
P	96.66	2.08	1.25	0.480
Lewis	98.45	1.36	0.19	0.467
Secretor	85.34	13.92	0.74	0.332
Mean ± S.E.	96.36 ± 1.03	2.97 ± 0.99	0.67 ± 0.14	0.479 ± 0.058

individual for eight loci. Data on the secretor locus were excluded from this analysis because of sample size limitations (see Table 1). The number of loci with respect to which the individual was heterozygous was determined. This generated an observed distribution of the number of heterozygous loci across 511 individuals. Chakraborty (1981) provided a numerical algorithm to compute the expected distribution for such observations, assuming random association of alleles at the different loci. Table 8 shows the results for our sample. In general, the observed distribution agrees with the expected one ($\chi^2 = 11.47, p > 0.05$). In our data the mean number of heterozygous loci is 3.97 and the variance is 1.59. Their expected values (under the random association model) are 3.97 and 1.75, respectively. The 95% confidence limit of the variance is 1.55–1.96. Clearly, these values provide no evidence of nonrandom association of the alleles among the eight polymorphic loci in the total Mexican population that resides in the MMA.

Discussion and Conclusion

The results of genetic distance analysis between various Mexican subpopulations residing in the MMA indicate that these populations are similar to each other. The gene differentiation among the subpopulations suggests that overall the level of gene diversity (G_{ST}) is small and more than 96% of the total gene diversity (H_T) is accounted for by individual

Table 5. Contribution (%) from Spanish and Mexican Indian Gene Pools to the Population of Nuevo León Grouped by Generation and Birthplace of the Four Grandparents

Birthplace of the Four Grandparents	*Ancestral Population*	
	Spanish	*Mexican Indians*
MMA		
Generation 11	95.80 ± 7.28	4.20 ± 7.28
Generation 12	82.70 ± 6.35	17.30 ± 6.35
Generation 13	81.78 ± 5.75	18.22 ± 5.75
Total	82.15 ± 5.51	17.85 ± 5.51
SLP		
Generation 11	66.59 ± 11.27	33.41 ± 11.27
Generation 12	69.25 ± 3.34	30.75 ± 3.34
Generation 13	80.40 ± 5.19	19.60 ± 5.19
Total	75.64 ± 5.42	24.60 ± 5.42
ZAC		
Generation 11	94.21 ±7.12	5.79 ± 7.12
Generation 12	53.49 ± 4.31	46.51 ± 4.31
Generation 13	52.26 ± 3.36	47.74 ± 3.36
Total	66.21 ± 4.49	33.79 ± 4.49
Total		
Generation 11	91.14 ± 9.63	8.86 ± 9.63
Generation 12	70.47 ± 5.49	29.53 ± 5.49
Generation 13	78.48 ± 4.62	21.52 ± 4.62
Total	78.46 ± 5.56	21.54 ± 5.56

The computations are done with nine polymorphic loci (ABO, Rh, MNSs, Duffy, Kidd, Lutheran, P, Lewis, and Secretor).

Table 6. Contribution (%) from the MMA, SLP, and ZAC Gene Pools to the Population of Nuevo León Stratified by Generation

Generation	*Population*		
	MMA	SLP	ZAC
11	71.61 ± 6.24	9.10 ± 7.52	19.29 ± 8.33
12	56.58 ± 8.52	17.22 ± 11.48	26.20 ± 4.84
13	76.26 ± 6.62	16.14 ± 7.40	7.60 ± 1.62
Total	64.56 ± 8.70	13.37 ± 12.96	22.07 ± 4.77

The computations are done with nine polymorphic loci (ABO, Rh, MNSs, Duffy, Kidd, Lutheran, P, Lewis, and Secretor).

Table 7. Average Heterozygosity (H_i) and Genetic Distances from Centroid (r_{ii}) among the Populations of Nuevo León Grouped by Generation and Birthplace of the Four Grandparents Based on Nine Polymorphic Loci

Population Grouped by:	r_{ii}	H_i
Generation		
11	0.0322 ± 0.0174	0.5133 ± 0.0436
12	0.0117 ± 0.0045	0.4714 ± 0.0633
13	0.0034 ± 0.0009	0.4796 ± 0.0594
Birthplace of four grandparents		
MMA	0.0074 ± 0.0035	0.4925 ± 0.0582
SLP	0.0042 ± 0.0015	0.4841 ± 0.0523
ZAC	0.0126 ± 0.0051	0.4513 ± 0.0658

Regression analysis: $H_i = b(1 - r_{ii})$. H_i plotted against $1 - r_{ii}$ through the origin:
Generation: $t = 2.70$; d.f. = 1, $p > 0.05$.
Birthplace of four grandparents: $t = -1.34$; d.f. = 1, $p > 0.05$.
Regression coefficient through the origin:
Generation: $b = 0.496 \pm 0.029$, d.f. = 2.
Birthplace of four grandparents: $b = 0.480 \pm 0.020$, d.f. = 2.
Average heterozygosity in pooled population:
Generation: $\bar{H} = 0.486 \pm 0.058$.
Birthplace of four grandparents: $\bar{H} = 0.481 \pm 0.059$.

variation within the population. However, the Mexicans stratified by birthplace of the four grandparents do contribute a small fraction (2.97%) of the genetic variation to H_T, suggesting that the geographic isolation of the population may bring about genetic variation over time. The subdivision by generations, on the contrary, provided little contribution (0.67%) to H_T. The overall pattern of gene differentiation conforms with the parental affinities between the subpopulations in the MMA.

The results obtained from the dihybrid model showed 78.5% Spanish and 21.5% Mexican Indian ancestry. However, the Spanish component is more pronounced in the natives of the MMA and in generation 11 of the total Mexicans. Furthermore, we found that the population of Nuevo León, stratified by generation, has received predominantly an MMA contribution, followed by ZAC and SLP. These results are consistent with previous work on genetic admixture in the population of the state of Nuevo León (Garza-Chapa 1983; Cerda-Flores et al. 1987; Cerda-Flores and Garza-Chapa 1989). Also, it is interesting to note that the computations of genetic admixture are similar to those obtained in our previous work based on 17 polymorphic loci in Mexican-Americans of Texas (with Spanish and Amerindian contributions of 70.1% and 29.9%, respectively (Cerda-Flores, Kshatriya et al. 1991).

Table 8. Observed and Expected Distribution of the Number of Heterozygous Loci in the Mexicans of Nuevo León, Mexico

Number of Heterozygous Loci	*Number of Individuals*	
	Observed	*Expected*
0	2	1.15
1	11	12.54
2	38	53.63
3	125	118.47
4	168	150.16
5	117	113.01
6	43	49.49
7	6	11.49
8	1	1.07
Total	511	511.00
Mean	3.97	3.97
Variance	1.59	1.75

Goodness of fit $\chi^2_8 = 11.48$, $p > 0.05$.
95% Confidence interval for variance (1.55, 1.95).

There are, however, some differences between our estimates of ancestral population contribution in these gene pools in the MMA and those of Lisker et al. (1986) in Mexico City (university student population) and Crawford et al. (1979) and Crawford and Devor (1980) in the Tlaxcala valley and the state of Coahuila (populations with indigenous influences). As we have mentioned previously, we selected people who knew the age and birthplace of their four grandparents (MMA, SLP, and ZAC) from the Mestizo population of the MMA. This is a different approach from other investigators.

Historical evidence (Cossio 1925; Montemayor-Hernández 1971; del Hoyo 1979; Hernández-Garza 1973) indicates that, when the Spanish, Portuguese, and Arabs (Sephardic Jews) colonized Nuevo León in 1596, the native Indians were forced to migrate because of the increasing pressure from the colonizers, thus leaving the region primarily to the colonized populations. But these same historians do not mention the influence of black populations in any time period in the state of Nuevo León, only later Tlaxcaltecan Indian, French, German, and United States populations.

Other studies also report little or no influence from African populations (Garza-Chapa et al. 1982; Garza-Chapa 1983; Cerda-Flores et al. 1987; Cerda-Flores and Garza-Chapa 1988, 1989; Cerda-Flores, Arriaga-Rios et al. 1990; Cerda-Flores et al. 1991; González-Quiroga et al. 1990).

This does not rule out the possibility that the Mestizo population of the MMA is influenced by African genes; therefore we considered that this genetic influence could be from Arab-African admixture or from ancestors from the Gulf of Mexico. Sandler et al. (1979) provide evidence of mixture of Africans with Arabs 900 years ago; they studied the Duffy system distribution in an Arab (Jewish) population. This Arab population dominated Spain for 400 years before the colonization of Nuevo León (del Hoyo 1979). Alternatively, there could have been immigration of people whose grandparents were born on the coast of the Gulf of Mexico, where there was a major African influence; in 1610, 150,000 black slaves arrived at the Mexican coast (Lisker 1980).

In another study, González-Quiroga et al. (1985) found that 5 of 752 male neonates in a Mestizo population in the MMA were deficient in G6PD. From the two maternal grandparent birthplaces, we can explain the distribution of the 5 deficiencies: MMA (2/253), SLP (1/146), Coahuila (1/42), and Tamaulipas (1/35). González-Quiroga et al. (1985) compared the frequencies of Coahuila and Tamaulipas and found them to be similar to those described by Lisker et al. (1969) for coastal populations of the Gulf of Mexico, where the African gene was assumed to originate. Later, González-Quiroga et al. (1990) reported 13 of 829 male neonates with jaundice in a Mestizo population of the MMA; 10 of the 13 neonates had variant A−, and their maternal grandparents were from the coast of the Gulf of Mexico. These results are similar to those published by Lisker et al. (1969). The interesting point of this study was that the three other deficiences were found to be B− variant, and their maternal grandparents were born in the MMA, ZAC, and SLP. Therefore González-Quiroga et al. (1990) concluded that there was minimal African influence on the Mestizo population of the MMA. Cerda-Flores, Arriaga-Rios et al. (1990) studied selected Nuevo León populations in which all four grandparents were born in Nuevo León but outside the MMA; no G6PD deficiencies were found in 428 children. Therefore the population outside the MMA has a small or no African influence because these populations are not from the industrial zone where there is a continuous immigration process, as in the MMA.

One can argue that our estimates could have been affected by the choice of gene frequency data on the ancestral populations, ignoring the possibility of a third component, namely, the black contributions in these gene pools. We contend that this is not the case. Although an exact specification of the ancestral allele frequencies is always difficult in any admixture study, our choice of the ancestral allele frequencies has been demonstrated to be adequate for populations of Mexican origin (Chakraborty et al. 1986; Hanis et al. 1991). In the present analysis we entertained a trihybrid model of admixture, incorporating contributions from blacks in addition to those from the Spanish and Mexican Indians.

The details of such an analysis are not presented here because the black admixture turned out to be negative in most cases, suggesting little contribution of the black gene pool to these populations. Crawford et al. (1979) and Crawford and Devor (1980) also found a small ($4.2\% \pm 2.8\%$) contribution of blacks to the Chamizal population. There are also some minor allele frequency estimation errors in the article by Crawford et al. (1979). For example, their estimates of allele frequencies at the ABO locus (shown in Table 4 of their article) do not sum to 1, nor do the reported allele frequencies at the P locus agree with their maximum likelihood estimation. To what extent these discrepancies contribute to their black admixture component is not known. At any rate, our observation of a relatively more pronounced Spanish contribution and an absence of a black component may in part be explained by the medium and higher socioeconomic background of the study sample compared with the samples examined by Crawford et al. (1979), Crawford and Devor (1980), and Lisker et al. (1986).

The expected and observed distributions of the number of heterozygous loci indicate that there is no residual effect of such admixture on the nonrandomness of allelic associations at the polymorphic loci examined here. These findings suggest that the admixture occurred long enough ago that at present the Mexicans of the MMA are a homogeneous group.

On the basis of the genetic data presented here, we conclude that the Mexicans who reside in the MMA, stratified by generation and birthplace of the grandparents, are genetically similar. The findings of the genetic admixture analysis suggest that this population had a predominant influence from the Spanish and a lesser contribution from Mexican Indians. We also observe a differential contribution of the MMA, SLP, and ZAC to the population of Nuevo León, with the predominant contribution being from the MMA. Furthermore, the multilocus heterozygosity distribution suggests that the history of admixture is old enough that the present group of Mexicans in the MMA is homogeneous.

Acknowledgments We express our thanks to Medical Units 5, 25, and 26 of the Instituto Mexicano del Seguro Social in Monterrey, Nuevo León, for facilities to sample and interview the study participants. Statistical analyses were supported by the National Institutes of Health under US Public Health Service Research Grant GM 41399. Special thanks to M.H. Crawford and the three anonymous reviewers for their constructive criticisms and suggestions on an earlier draft.

Received 6 August 1990; revision received 19 October 1990.

Appendix: Allele Frequencies for 9 Genetic Loci

System	*Spanish*	*Mexican Indian*
A	0.310(1)	0.063(9)
B	0.067	0.003
O	0.623	0.934
DCE	0.048(2)	0.022(9)
DCe	0.418	0.626
DcE	0.090	0.330
Dce	0.049	0.000
dCE	0.002	0.000
dCe	0.011	0.000
dcE	0.001	0.000
dce	0.381	0.022
MS	0.243(3)	0.346(9)
Ms	0.311	0.444
NS	0.057	0.080
Ns	0.389	0.130
Fy(a)	0.365(4)	0.820(9)
Fy(b)+Fy	0.635	0.180
Jk(a)	0.537(5)	0.360(9)
Jk(b)+Jk	0.463	0.640
Lua	0.041(6)	0.000(10)
Lub	0.959	1.000
P1	0.540(7)	0.367(7)
P2+p	0.460	0.633
Le	0.699(8)	0.616(11)
le	0.301	0.384
Se	0.517(8)	1.000(11)
se	0.483	0.000

Adapted from Hanis et al. (1991).

Sources: The numbers in parentheses represent the source of the allele frequency data. (1) Mourant et al. (1976), Table 1.1. Weighted average of populations from Extremadura, Galicia and Leon. (2) Mourant et al. (1976), Table 4.13. Weighted average on non-Basques and Galicia (frequencies recomputed from phenotypic counts). (3) Mourant et al. (1976), Table 2.7, Lugo, Galicia, Spain. (4) Tills et al. (1983), Table 8.3.1, Barcelona, Spain. (5) Tills et al. (1983), Table 9.3.1, Barcelona, Spain. (6) Mourant et al. (1976), Table 5.1, Basques. (7) Crawford et al. (1974), Table 5. (8) Mourant et al. (1976), Table 7.7.4, Basques. (9) Niswander et al. (1970), Table 4. (10) Mourant et al. (1976), Table 5.1, Chiapas, Yucatan, Oaxaca, Veracruz. (11) Mourant et al. (1976), Table 7.7.3, Chiapas.

Literature Cited

Brown, A.H.D., M.W. Feldman, and E. Nevo. 1980. Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* 96:523–536.

Cerda-Flores, R.M., and R. Garza-Chapa. 1988. Cambios en las frecuencias de incompatibilidad para ABO y Rh(D) en tres generaciones de la poblacion del area metropolitana de Monterrey, Nuevo León, Mexico. *Arch. Invest. Med. (Mex.)* 19(1):79–89.

Cerda-Flores, R.M., and R. Garza-Chapa. 1989. Variation in the gene frequencies of three generations of humans from Monterrey, Nuevo León, Mexico. *Hum. Biol.* 61:249–261.

Cerda-Flores, R.M., E. Ramirez-Fernandez, and R. Garza-Chapa. 1987. Genetic admixture and distances between populations from Monterrey, Nuevo León, Mexico, and their putative ancestral populations. *Hum. Biol.* 59:31–49.

Cerda-Flores, R.M., V.A. Bautista-Pena, M.A. Rojas-Alvarado, and R. Garza-Chapa. 1991. Polimorfismo genetico en la poblacion de Cerralvo, Nuevo León. *Estud. Antropol. Biol. (Mex.)* (in press).

Cerda-Flores, R.M., G.K. Kshatriya, T.K. Bertin, D. Hewett-Emmett, C.L. Hanis, and R. Chakraborty. 1991. Gene diversity and estimation of genetic admixture among Mexican-Americans of Starr County, Texas. *Ann. Hum. Biol.* (in press).

Cerda-Flores, R.M., G. Arriaga-Rios, J. Munoz-Campos, V.A. Bautista-Pena, M.A. Rojas-Alvarado, G. Gonzalez-Quiroga, C.H. Leal-Garza, and R. Garza-Chapa. 1990. Frecuencia de la ceguera para los colores y de la deficiencia a la enzima glucosa-6-fosfato-deshidrogenasa en poblaciones no industrializadas del Estado de Nuevo León, Mexico. *Arch. Invest. Med. (Mex.)* 21(3).

Chakraborty, R. 1980. Gene diversity analysis in nested subdivided populations. *Genetics* 96:721–726.

Chakraborty, R. 1981. The distribution of the number of heterozygous loci in natural populations. *Genetics* 98:461–466.

Chakraborty, R. 1984. Detection of nonrandom association of alleles from the distribution of the number of heterozygous loci in a sample. *Genetics* 108:719–731.

Chakraborty, R. 1985. Gene identity in racial hybrids and estimation of admixture rates. In *Genetic Microdifferentiation: Human and Other Populations*, Y.R. Ahuja and J.V. Neel, eds. New Delhi: Indian Anthropological Association, 171–180.

Chakraborty, R. 1986. Gene admixture in human populations: Models and predictions. *Yrbk. Phys. Anthropol.* 29:1–43.

Chakraborty, R., M. Haag, N. Ryman, and G. Stahl. 1982. Hierarchical gene diversity analysis and its implication to brown trout population data. *Hereditas* 97:17–21.

Chakraborty, R., R.E. Ferrell, M.P. Stern, S.M. Haffner, H.P. Hazuda, and M. Rosenthal. 1986. Relationship of prevalence of non-insulin-dependent diabetes mellitus to Amerindian admixture in the Mexican Americans of San Antonio, Texas. *Genet. Epidemiol.* 3:435–454.

Cossio, D.A. 1925. *Historia de Nuevo León, evolucion politica y social*, S. Cantu Leal, ed. Monterrey, Nuevo León, Mexico.

Crawford, M.A., F.E. Gottman, and C.A. Gottman. 1970. Microplate system for routine use in blood bank laboratories. *Transfusion* 10:258–263.

Crawford, M.H., and E.J. Devor. 1980. Population structure and admixture in transplanted Tlaxcaltecan populations. *Am. J. Phys. Anthropol.* 52:485–490.

Crawford, M.H., D.D. Dykes, K. Skradski, and H.P. Polesky. 1979. Gene flow and genetic microdifferentiation of a transplanted Tlaxcaltecan Indian population: Saltillo. *Am. J. Phys. Anthropol.* 50:401–412.

Crawford, M.H., W.C. Leyshon, K. Brown, F. Lees, and L. Taylor. 1974. Human biology in Mexico. II. A comparison of blood groups, serum and red cell enzyme frequencies, and genetic distances of the Indian population of Mexico. *Am. J. Phys. Anthropol.* 41:251–268.

Del Hoyo, H. 1979. *Historia del Nuevo Reino de León (1577–1723)*, 2d ed. Mexico: Editorial Libros de Mexico.

Dirección de Estadistica y Procesamiento de Datos del Gobierno de Nuevo León. 1977. *Aspectos demograficos del Estado de Nuevo León.* Monterrey, Nuevo León, Mexico.

Garza-Chapa, R. 1983. Genetic distances for ABO and Rh(D) blood groups in the state of Nuevo León, Mexico. *Soc. Biol.* 30:24–31.

Garza-Chapa, R., C.H. Leal-Garza, and R.M. Cerda-Flores. 1982. Population genetics in the state of Nuevo León, Mexico. V. Frequencies of ABO, Rh(D), MN blood groups and other genetic traits. *Acta Anthropogenet.* 6(4):225–245.

González-Quiroga, G., M. Walle-Cardona, R. Ortiz-Jalomo, and R. Garza-Chapa. 1985. Deficiencia de deshidrogenosa de la glucosa-6-fosfato (G6PD) en varones neonatos en Monterrey, Nuevo León. *Rev. Med. IMSS (Mex.)* 23:247–250.

González-Quiroga, G., J.L. Ramirez-Del Rio, R. Ortiz-Jalomo, R.F. Garcia-Contreras, R.M. Cerda-Flores, B.D. Mata-Cardenas, and R. Garza-Chapa. 1990. Frecuencia relativa de la deficiencia de glucosa-6-fosfato-deshidrogenasa (G6PD) en neonatos ictericos del area metropolitana de Monterrey, Nuevo León. *Arch. Invest. Med. (Mex.)* 21(3).

Hanis, C.L., D. Hewett-Emmett, T.K. Bertin, and W.J. Schull. 1991. The origins of US Hispanics: Implications for diabetes. *Diabetes Care* (in press).

Harpending, H.C., and R.H. Ward. 1982. Chemical systematics and human evolution. In *Biochemical Aspects of Evolutionary Biology*, M. Nitecki, ed. Chicago, Ill.: University of Chicago Press, 213–256.

Hernández-Garza, T.L. 1973. *Breve historia de Nuevo León*, 3d ed. Mexico: Editorial Trillas.

Lapinski, F.J., K.M. Crowley, C.A. Merrit, and J.B. Henry. 1978. Use of microplate methods in paternity testing. *Am. J. Clin. Pathol.* 70:766–769.

Lisker, R. 1980. *Estructura Genética de la Poblacion Mexicana: Aspectos Medicos y Anthropologicos.* Mexico: Salvat.

Lisker, R., M. Cordova, and G. Zarate. 1969. Studies on several genetic hematological traits of the Mexican population. XVI. Hemoglobin S and G-6-PD deficiency in the East Coast. *Am. J. Phys. Anthropol.* 30:349.

Lisker, R., R. Perez-Briceño, J. Granados, V. Babinsky, J. De Rubens, S. Armendares, and L. Buentello. 1986. Gene frequencies and admixture estimates in a Mexico City population. *Am. J. Phys. Anthropol.* 71:203–207.

Montemayor-Hernández, A. 1971. *Historia de Monterrey.* Mexico: Asociacion de Editores de Monterrey, A.C.

Mourant, A.E., A.C. Kopec, and K. Domaniewska-Sobczak. 1976. *The Distribution of the Human Blood Groups and Other Polymorphisms.* London: Oxford University Press.

Nei, M. 1972. Genetic distance between populations. *Am. Natur.* 106:283–292.

Nei, M., and A.K. Roychoudhury. 1974. Sampling variances of heterozygosity and genetic distance. *Genetics* 76:379–390.

Niswander, J.D., K.S. Brown, B.Y. Iba, W.C. Leyshow, and P.L. Workman. 1970. Population studies on Southwestern Indian tribes. I. History, culture, and genetics of the Papago. *Am. J. Hum. Genet.* 22:7–23.

Reed, T.E., and W.J. Schull. 1968. A general maximum likelihood estimation program. *Am. J. Hum. Genet.* 20:579–580.

Sandler, S.G., C. Krovitz, R. Sharon, D. Hermoni, E. Ezckiel, and T. Cohen. 1979. The Duffy blood group system in Israeli Jews and Arabs. *Vox Sang.* 37:41–46.

Snedecor, G.W., and W.G. Cochran. 1967. *Statistical Methods*, 6th ed. Ames, Iowa: Iowa State University Press.

Tills, D., A.C. Kopec, and R.E. Tills. 1983. *The Distribution of the Human Blood Groups and Other Polymorphisms*, suppl. 1. Oxford: Oxford University Press.

Genetic Structure of the Populations Migrating from San Luis Potosi and Zacatecas to Nuevo León in Mexico

RICARDO M. CERDA-FLORES,[1,2] GAUTAM K. KSHATRIYA,[2] SARA A. BARTON,[2] CARLOS H. LEAL-GARZA,[1] RAUL GARZA-CHAPA,[1] WILLIAM J. SCHULL,[2] AND RANAJIT CHAKRABORTY[2]

Abstract The Mexicans residing in the Monterrey metropolitan area in Nuevo León, Mexico, were grouped by generation and birthplace [Monterrey Metropolitan Area (MMA), San Luis Potosi (SLP), and Zacatecas (ZAC)] of the four grandparents to determine the extent of genetic variation within this population and the genetic differences, if any, between the natives living in the MMA and the immigrant populations from SLP and ZAC. Nine genetic marker systems were analyzed. The genetic distance analysis indicates that SLP and ZAC are similar to the MMA, irrespective of birthplace and generation. Gene diversity analysis (G_{ST}) suggests that more than 96% of the total gene diversity (H_T) can be attributed to individual variation within the population. The genetic admixture analysis suggests that the Mexicans of the MMA, SLP, and ZAC, stratified by birthplace and generation, have received a predominantly Spanish contribution (78.5%), followed by a Mexican Indian contribution (21.5%). Similarly, admixture analysis, conducted on the population of Nuevo León and stratified by generation, indicates a substantial contribution from the MMA (64.6%), followed by ZAC (22.1%) and SLP (13.3%). Finally, we demonstrate that there is no nonrandom association of alleles among the genetic marker systems (i.e., no evidence of gametic disequilibrium) despite the Mestizo origin of this population.

The state of Nuevo León in northeastern Mexico (Figure 1) has an area of 64,555 km^2, and in 1990 had a population of 4,492,500 inhabitants. The age distribution of this Mestizo population indicates that 72% of the total population is under 30 years, 23% is between 30 and 59 years, and only 5% exceeds 59 years of age (Dirección de Estadistica y Procesamiento de Datos del Gobierno de Nuevo León, 1977).

The Monterrey Metropolitan Area (MMA) is located in the central western section of the state of Nuevo León and has an area of 2118 km^2.

[1]Subjefatura de Investigacion Cientifica, Instituto Mexicano del Seguro Social, Unidad de Investigacion Biomedica del Noreste, Apartado Postal 020-E, 64720 Monterrey, Nuevo León, Mexico.
[2]Center for Demographic and Population Genetics, Graduate School of Biomedical Sciences, University of Texas, PO Box 20334, Houston, Texas 77225.

Human Biology, June 1991, Vol. 63, No. 3, pp. 309-327.

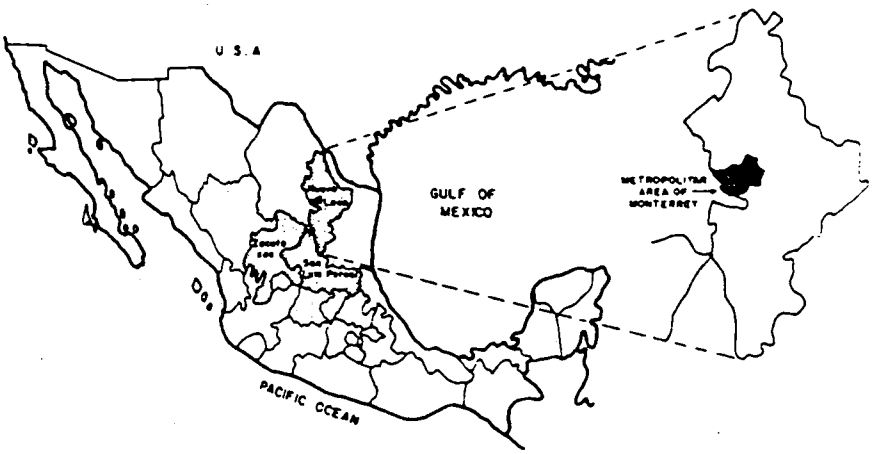


Figure 1. Location of the state of Nuevo León in Mexico.

In 1940 Nuevo León had 541,147 inhabitants, of whom 39% resided in the MMA, whereas in 1960 the population was 1,078,848 and 67% resided in the MMA. At present, nearly 80% of the state's population is concentrated in the MMA (Dirección de Estadistica y Procesamiento de Datos, 1977). This increase is principally due to the immigration of people from several southern states of Mexico to the MMA since 1910. For example, in 1960 alone 400,000 people migrated to the MMA mainly from the states of San Luis Potosi (28%), Coahuila (24%), Tamaulipas (16%), Zacatecas (11%), and the Federal District (3%) and some foreign countries, particularly the United States (Montemayor-Hernández 1971). The explanation for this immigration is that the MMA is an important industrial, commercial, and educational zone in Mexico that attracts people in search of jobs, livelihood, business, and education.

In the present study the Mexican population that resides in the MMA is grouped by generation and birthplace of the four grandparents. Our aims are (1) to study the genetic variation among the Mexican population of the MMA, stratified by generation and birthplace, and to determine whether the immigrant population groups, for example, San Luis Potosi (SLP) and Zacatecas (ZAC), are different from the native Mexicans whose four grandparents were born in the MMA; (2) to compute the contribution of the ancestral populations (i.e., Spanish and Indians) to the Mexicans of the MMA, stratified by generation and birthplace, and of the MMA, SLP, and ZAC populations to the population of Nuevo León, stratified by generation only; (3) to study the proportion of genes received from the ancestral populations; and (4)

to demonstrate whether there is any residual effect of population mixture on the nonrandom association of alleles.

Materials and Methods

From the 4680 people residing in the MMA who were randomly interviewed at the Instituto Mexicano del Seguro Social (IMSS), Monterrey, in 1985, 207 were selected whose 4 grandparents were born in the MMA (Cerda-Flores and Garza-Chapa 1989), 151 were selected whose grandparents were born in SLP, and 153 were selected whose grandparents were born in ZAC, thereby constituting a total sample size of 511 individuals. The frequencies of the phenotypes of the blood group systems ABO, Rh, MNSs, Duffy (Fy), Kidd (Jk), Lutheran (Lu), P, Lewis (Le), and Secretor (Se) were analyzed using commercial antisera and the microplate method described by Crawford et al. (1970) and Lapinski et al. (1978).

The data were subdivided by year of birth, and the number of generations was estimated assuming a generation time of 30 years. Accordingly, 13 human generations have elapsed since the MMA was colonized in 1596 (del Hoyo 1979). We consider the last three: (1) persons born between 1896 and 1925 (generation 11), (2) persons born between 1926 and 1955 (generation 12), and (3) persons born between 1956 and 1985 (generation 13).

The statistical analysis was conducted in six parts. First, the gene frequencies for different systems were computed using the maximum likelihood method (Reed and Schull 1968). Next, the genetic distances among the Mexicans of the MMA, stratified by generation and birthplace of the grandparents, were computed by Nei's standard genetic distance (Nei 1972), and their standard errors (SEs) were computed using the method of Nei and Roychoudhury (1974). The gene frequency data were further subjected to the pairwise chi-square statistic to determine the statistical significance of the genetic distance (Nei and Roychoudhury 1974). Third, the extent of genetic variation between the subpopulations of Mexicans in the MMA (by generation and birthplace) was studied using the nested gene diversity computer program (NEGST) developed by Chakraborty (1980) and Chakraborty et al. (1982). Next, the contribution (%) of Spanish and Mexican Indians to the population groups of the MMA and of the MMA, SLP, and ZAC groups to the population of Nuevo León, stratified by three generations, was calculated following procedures detailed by Chakraborty (1985, 1986) using dihybrid and trihybrid models.

To determine whether the proportions of genes received by the Mexican subpopulations of the MMA from their ancestral sources are signifi-

cantly different from each other, we next performed a regression analysis of heterozygosity on genetic distance, as proposed by Harpending and Ward (1982). The regression equation was subjected to a test of significance, following the method of Snedecor and Cochran (1967). Finally, the computation for nonrandom association of alleles among different genetic loci was conducted following the methods of Brown et al. (1980) and Chakraborty (1981, 1984) to examine whether any residual effects of admixture remain in the current MMA population that would make this population heterogeneous.

Gene frequency data on the ancestral populations were obtained from Hanis et al. (1991) (see appendix).

Results

Genetic Distance. Table 1 provides the allele frequency estimates for the analyzed loci. Based on the allele frequency, we first determined the genetic differences within the Mexicans of the MMA, stratified by generation and birthplace of the four grandparents. We estimated Nei's standard genetic distances for all pairs of populations and their SEs (Table 2). Table 2 also indicates average heterozygosity ($\bar{H}$) among the subpopulations. The subpopulations of the MMA, arranged by generation, have an $\bar{H}$ that varies between 47.3% (generation 12) and 51.7% (generation 11) and between 45.3% (ZAC) and 49.4% (MMA) when the populations are grouped according to the birthplace of the four grandparents. The genetic distance analysis (Table 3) does not reveal any significant level of differentiation, as reflected by the pairwise chi-square statistic (Nei and Roychoudhury 1974).

Gene Diversity Analysis. Table 4 presents the hierarchical gene diversity analysis (G_{ST}) among the subpopulations of the MMA. The total average gene diversity (H_T) is 47.9%. Over 96% of H_T, computed on the basis of the 9 loci, can be attributed to individual variation within the population. However, a small contribution to the total variability (2.97%) comes from the between-birthplace variation of the four grandparents, suggesting that there is some genetic variation among the Mexicans that reside in the MMA. The generation difference in the extent of genetic variation is even smaller (0.67% of the total) and probably not of any biologic significance.

Genetic Admixture Analysis. Table 5 presents the estimated values of admixture based on nine polymorphic genetic loci. In the present investigation we considered the Mexicans of the MMA to be the product of admixture of two parental populations having Spanish and Mexican

Indian ancestry. The allele frequencies of the ancestral populations are presented in the appendix. The results of fitting the dihybrid model show that the contribution from the Spanish ancestry to the Mexicans of the MMA, stratified by birthplace, varies from 66.21% in ZAC to 82.15% in the natives of the MMA. However, the Spanish contribution, when the data are stratified by generation, varies from 52.26% in generation 13 of ZAC to 95.8% in generation 11 of the natives of the MMA.

The overall contribution of the Spanish to the Mexicans of the MMA (total) is 78.46% and, when subdivided by generation, varies from 70.47% in generation 12 to 91.14% in generation 11. Based on these results, it can be inferred that the Mexicans of the MMA, stratified by generation and birthplace, have received a predominantly Spanish contribution, followed by a Mexican Indian contribution.

The MMA population increased from 44,808 in 1900 to 112,864 in 1921, to 382,021 in 1950, and to 1,242,558 in 1970 (Dirección de Estadistica y Procesamiento de Datos, 1977). Therefore we thought it would be interesting to examine, using a trihybrid model, the contribution (%) of the MMA and the immigrant populations of SLP and ZAC to the Nuevo León population, stratified by generation, to see whether the percentage of genetic contribution and the percentage of immigration in each generation are similar. We found that the genetic contribution of the MMA population is larger than the immigrant population in the three generations (Table 6). That genetic contributions of 56.6% in generation 12 and of 43.4% in the immigrant populations are found agrees with the historical account of Montemayor-Hernández (1971), who reported that, of the 400,000 people who entered the MMA, an industrial region, 156,000 (39%) were from SLP and ZAC. From these results it can be inferred that the population of Nuevo León has received predominantly an MMA contribution (64.5%), followed by contributions from ZAC (22.1%) and SLP (13.4%).

Heterozygosity and Genetic Distance. Table 7 shows the average heterozygosity (H_i) for nine genetic loci and the genetic distance (r_{ii}) for each subpopulation along with its interlocus standard error. The regression of heterozygosity on genetic distance is consistent with linearity. $\bar{H}$ in the population pooled by generation and birthplace of grandparents (48.6% and 48.1%) does not differ significantly from the regression coefficient b (49.6% and 48.0%). This indicates that the Mexican populations of the MMA, SLP, and ZAC are similar in the proportion of the genes that they have received from the ancestral populations. This finding is consistent with the similarity of admixture proportions estimated in the previous section.

Nonrandom Association among Genetic Loci. From the available genotype data on each individual, we defined the multilocus genotype for each

Table 1. Allele Frequencies among Mexicans of the State of Nuevo León Grouped by Generation and Birthplace of the Four Grandparents

	Birthplace of Generation 11				*Birthplace of Generation 12*				*Birthplace of Generation 13*				*Total*			
System	*MMA* (*n*=45)	*SLP* (*n*=15)	*ZAC* (*n*=19)	*Total* (*n*=79)	*MMA* (*n*=34)	*SLP* (*n*=44)	*ZAC* (*n*=72)	*Total* (*n*=150)	*MMA* (*n*=128)	*SLP* (*n*=92)	*ZAC* (*n*=62)	*Total* (*n*=282)	*MMA* (*n*=207)	*SLP* (*n*=151)	*ZAC* (*n*=153)	*Total* (*n*=511)
ABO																
A1	.212	.177	.054	.165	.143	.160	.126	.140	.143	.157	.130	.145	.158	.160	.118	.146
A2	.014	.000	.000	.008	.017	.013	.024	.020	.042	.045	.038	.042	.032	.031	.026	.030
B	.093	.067	.083	.085	.045	.011	.087	.055	.036	.044	.041	.040	.050	.037	.068	.051
O	.681	.756	.863	.742	.795	.816	.763	.785	.779	.754	.791	.773	.760	.772	.788	.773
Rh																
DCE	.000	.000	.104	.000	.079	.053	.102	.072	.033	.027	.090	.041	.000	.033	.103	.037
DCe	.240	.300	.290	.339	.201	.390	.377	.300	.216	.347	.330	.285	.263	.347	.328	.288
DcE	.128	.233	.265	.271	.042	.288	.155	.124	.291	.245	.241	.265	.247	.255	.198	.209
Dce	.207	.467	.340	.172	.260	.159	.366	.290	.228	.238	.184	.218	.188	.231	.268	.239
dCE	.053	.000	.000	.163	.000	.000	.000	.012	.015	.000	.000	.010	.060	.000	.000	.026
dCe	.062	.000	.001	.000	.176	.000	.000	.082	.040	.045	.056	.044	.034	.034	.036	.050
dcE	.030	.000	.000	.000	.128	.000	.000	.078	.000	.000	.000	.000	.007	.000	.000	.025
dce	.280	.000	.001	.055	.113	.110	.000	.041	.177	.098	.099	.137	.201	.100	.067	.126
MNSs																
MS	.337	.324	.453	.360	.309	.308	.302	.301	.324	.268	.358	.313	.323	.284	.343	.317
Ms	.297	.343	.153	.273	.396	.374	.303	.349	.320	.379	.312	.338	.329	.375	.288	.331
NS	.041	.209	.205	.114	.058	.101	.226	.155	.137	.124	.118	.129	.105	.126	.180	.134
Ns	.325	.124	.189	.253	.237	.217	.169	.195	.219	.229	.212	.220	.243	.215	.189	.218

Duffy																
Fy(a)	.459	.507	.276	.413	.455	.489	.416	.444	.401	.326	.372	.367	.422	.388	.379	.396
Fy(b)	.459	.386	.489	.447	.428	.350	.359	.370	.443	.614	.347	.464	.444	.503	.369	.433
Fy	.082	.107	.235	.140	.117	.161	.225	.186	.156	.060	.281	.169	.134	.109	.252	.171
Kidd																
Jk(a)	.423	.317	.205	.344	.431	.397	.237	.322	.235	.278	.260	.254	.302	.314	.242	.287
Jk(b)+Jk	.577	.683	.795	.656	.569	.603	.763	.678	.765	.722	.740	.746	.698	.686	.758	.713
Lutheran																
Lua	.144	.317	.173	.262	.109	.160	.065	.102	.099	.134	.093	.109	.110	.158	.089	.117
Lub	.856	.683	.827	.738	.891	.840	.935	.898	.901	.866	.907	.891	.890	.842	.911	.883
P																
P1	.506	.553	.771	.564	.657	.703	.627	.576	.549	.546	.492	.535	.555	.518	.580	.551
P2+p	.494	.447	.229	.436	.343	.297	.373	.424	.451	.454	.508	.465	.445	.482	.420	.449
Lewis																
Le	.553	.684	.622	.606	.564	.746	.663	.663	.575	.598	.655	.603	.569	.645	.654	.610
le	.447	.316	.378	.394	.436	.254	.337	.337	.425	.402	.345	.397	.431	.355	.346	.390
Secretor[a]																
Se	.646	1.00	.592	.671	.758	.737	1.00	.803	.657	.717	1.00	.696	.670	.739	.842	.720
se	.354	0.00	.408	.329	.242	.263	0.00	.197	.343	.283	0.00	.304	.330	.261	.158	.280
n	20	10	14	44	21	31	44	96	83	62	42	187	124	103	100	327

a. The *n* values for Secretor are different from those for all other systems.

Table 2. Standard Genetic Distances and Average Heterozygosity within the Population of Nuevo León Grouped by Generation and Birthplace of the Four Grandparents[a]

	Generation			*Birthplace of Four Grandparents*			
Population	*11*	*12*	*13*	*MMA*	*SLP*	*ZAC*	*Total*
Generation							
11	51.71 ± 4.38						
12	14.76 ± 9.52	47.30 ± 6.35					
13	8.10 ± 5.64	8.06 ± 4.82	48.06 ± 5.95				
Birthplace of four grandparents							
MMA	5.70 ± 6.05	10.83 ± 7.95	0.60 ± 1.01	49.38 ± 5.83			
SLP	4.81 ± 3.92	3.81 ± 3.74	0.50 ± 0.77	3.47 ± 2.41	48.58 ± 5.25		
ZAC	16.02 ± 7.49	1.70 ± 1.78	5.61 ± 3.98	11.82 ± 6.66	4.49 ± 2.01	45.29 ± 6.60	
Total	6.70 ± 5.58	3.34 ± 2.43	0.28 ± 0.44	0.89 ± 1.34	0.12 ± 0.69	3.81 ± 2.57	48.56 ± 5.96

a. Values on the diagonal are the average heterozygosities expressed in percentage (e.g., generation 11 by 11 = 51.7%); below the diagonal are standard genetic distances in 10^{-3} codon differences per locus. The computations are done with nine polymorphic loci (ABO, Rh, MNSs, Duffy, Kidd, Lutheran, P, Lewis, and Secretor).

Table 3. Test of Significance of Genetic Distances among Populations of the State of Nuevo León Grouped by Generation and Birthplace of the Four Grandparents Based on the Pairwise Chi-Square

	Generation 11				*Generation 12*				*Generation 13*				*Birthplace of Four Grandparents*		
Population	*MMA*	*SLP*	*ZAC*	*Total*	*MMA*	*SLP*	*ZAC*	*Total*	*MMA*	*SLP*	*ZAC*	*Total*	*MMA*	*SLP*	*ZAC*
Generation 11															
SLP	2.43														
ZAC	2.48	2.23													
Total	1.10	1.50	1.63												
Generation 12															
MMA	1.07	2.48	2.32	2.31											
SLP	1.42	1.68	1.53	0.91	1.48										
ZAC	3.22	1.04	2.81	2.11	2.87	1.87									
Total	1.38	1.31	1.36	1.30	0.55	0.84	0.83								
Generation 13															
MMA	0.89	1.94	1.12	0.99	1.68	0.84	1.58	0.81							
SLP	1.04	1.49	1.20	0.84	1.39	0.62	1.33	0.55	0.21						
ZAC	2.07	1.38	3.00	1.72	1.98	1.35	0.72	0.76	0.92	0.77					
Total	1.06	1.73	1.08	1.08	2.16	0.63	1.21	0.64	0.05	0.08	0.62				
Birthplace of four grandparents															
MMA	0.41	1.95	2.44	0.59	2.12	0.97	1.88	0.91	0.22	0.35	1.22	0.24			
SLP	1.15	1.21	1.27	0.85	1.64	0.40	1.17	0.51	0.29	0.06	0.63	0.12	0.39		
ZAC	1.80	1.33	1.04	1.32	1.75	0.76	0.51	0.37	0.57	0.37	0.39	0.36	0.83	0.35	
Total	0.65	1.49	1.11	0.92	1.12	0.59	1.06	0.33	0.16	0.13	0.62	0.09	0.20	0.14	0.34

Pairwise chi-square matrix for all loci: d.f. = 19; $p > 0.05$ nonsignificant.

Table 4. Gene Diversity Analysis of Allele Frequency Data from Populations of Nuevo León Grouped by Generations and Birthplace of the Four Grandparents

	Relative Gene Diversity (G_{ST}) (%)			
Locus	*Within Population*	*Between Birthplace of Four Grandparents*	*Between Generation within Birthplace of Four Grandparents*	*Total Gene Diversity (H_T)*
ABO	98.71	1.11	0.18	0.371
Rh	95.74	3.51	0.75	0.778
MNSs	98.13	1.52	0.36	0.724
Duffy	97.86	1.62	0.52	0.484
Kidd	96.85	2.40	0.75	0.427
Lutheran	96.10	2.05	1.85	0.246
P	96.66	2.08	1.25	0.480
Lewis	98.45	1.36	0.19	0.467
Secretor	85.34	13.92	0.74	0.332
Mean ± S.E.	96.36 ± 1.03	2.97 ± 0.99	0.67 ± 0.14	0.479 ± 0.058

individual for eight loci. Data on the secretor locus were excluded from this analysis because of sample size limitations (see Table 1). The number of loci with respect to which the individual was heterozygous was determined. This generated an observed distribution of the number of heterozygous loci across 511 individuals. Chakraborty (1981) provided a numerical algorithm to compute the expected distribution for such observations, assuming random association of alleles at the different loci. Table 8 shows the results for our sample. In general, the observed distribution agrees with the expected one ($\chi^2 = 11.47, p > 0.05$). In our data the mean number of heterozygous loci is 3.97 and the variance is 1.59. Their expected values (under the random association model) are 3.97 and 1.75, respectively. The 95% confidence limit of the variance is 1.55–1.96. Clearly, these values provide no evidence of nonrandom association of the alleles among the eight polymorphic loci in the total Mexican population that resides in the MMA.

Discussion and Conclusion

The results of genetic distance analysis between various Mexican subpopulations residing in the MMA indicate that these populations are similar to each other. The gene differentiation among the subpopulations suggests that overall the level of gene diversity (G_{ST}) is small and more than 96% of the total gene diversity (H_T) is accounted for by individual

Table 5. Contribution (%) from Spanish and Mexican Indian Gene Pools to the Population of Nuevo León Grouped by Generation and Birthplace of the Four Grandparents

Birthplace of the Four Grandparents	*Ancestral Population*	
	Spanish	*Mexican Indians*
MMA		
Generation 11	95.80 ± 7.28	4.20 ± 7.28
Generation 12	82.70 ± 6.35	17.30 ± 6.35
Generation 13	81.78 ± 5.75	18.22 ± 5.75
Total	82.15 ± 5.51	17.85 ± 5.51
SLP		
Generation 11	66.59 ± 11.27	33.41 ± 11.27
Generation 12	69.25 ± 3.34	30.75 ± 3.34
Generation 13	80.40 ± 5.19	19.60 ± 5.19
Total	75.64 ± 5.42	24.60 ± 5.42
ZAC		
Generation 11	94.21 ± 7.12	5.79 ± 7.12
Generation 12	53.49 ± 4.31	46.51 ± 4.31
Generation 13	52.26 ± 3.36	47.74 ± 3.36
Total	66.21 ± 4.49	33.79 ± 4.49
Total		
Generation 11	91.14 ± 9.63	8.86 ± 9.63
Generation 12	70.47 ± 5.49	29.53 ± 5.49
Generation 13	78.48 ± 4.62	21.52 ± 4.62
Total	78.46 ± 5.56	21.54 ± 5.56

The computations are done with nine polymorphic loci (ABO, Rh, MNSs, Duffy, Kidd, Lutheran, P, Lewis, and Secretor).

Table 6. Contribution (%) from the MMA, SLP, and ZAC Gene Pools to the Population of Nuevo León Stratified by Generation

Generation	*Population*		
	MMA	SLP	ZAC
11	71.61 ± 6.24	9.10 ± 7.52	19.29 ± 8.33
12	56.58 ± 8.52	17.22 ± 11.48	26.20 ± 4.84
13	76.26 ± 6.62	16.14 ± 7.40	7.60 ± 1.62
Total	64.56 ± 8.70	13.37 ± 12.96	22.07 ± 4.77

The computations are done with nine polymorphic loci (ABO, Rh, MNSs, Duffy, Kidd, Lutheran, P, Lewis, and Secretor).

Table 7. Average Heterozygosity (H_i) and Genetic Distances from Centroid (r_{ii}) among the Populations of Nuevo León Grouped by Generation and Birthplace of the Four Grandparents Based on Nine Polymorphic Loci

Population Grouped by:	r_{ii}	H_i
Generation		
11	0.0322 ± 0.0174	0.5133 ± 0.0436
12	0.0117 ± 0.0045	0.4714 ± 0.0633
13	0.0034 ± 0.0009	0.4796 ± 0.0594
Birthplace of four grandparents		
MMA	0.0074 ± 0.0035	0.4925 ± 0.0582
SLP	0.0042 ± 0.0015	0.4841 ± 0.0523
ZAC	0.0126 ± 0.0051	0.4513 ± 0.0658

Regression analysis: $H_i = b(1 - r_{ii})$. H_i plotted against $1 - r_{ii}$ through the origin:
Generation: $t = 2.70$; d.f. = 1, $p > 0.05$.
Birthplace of four grandparents: $t = -1.34$; d.f. = 1, $p > 0.05$.
Regression coefficient through the origin:
Generation: $b = 0.496 \pm 0.029$, d.f. = 2.
Birthplace of four grandparents: $b = 0.480 \pm 0.020$, d.f. = 2.
Average heterozygosity in pooled population:
Generation: $\bar{H} = 0.486 \pm 0.058$.
Birthplace of four grandparents: $\bar{H} = 0.481 \pm 0.059$.

variation within the population. However, the Mexicans stratified by birthplace of the four grandparents do contribute a small fraction (2.97%) of the genetic variation to H_T, suggesting that the geographic isolation of the population may bring about genetic variation over time. The subdivision by generations, on the contrary, provided little contribution (0.67%) to H_T. The overall pattern of gene differentiation conforms with the parental affinities between the subpopulations in the MMA.

The results obtained from the dihybrid model showed 78.5% Spanish and 21.5% Mexican Indian ancestry. However, the Spanish component is more pronounced in the natives of the MMA and in generation 11 of the total Mexicans. Furthermore, we found that the population of Nuevo León, stratified by generation, has received predominantly an MMA contribution, followed by ZAC and SLP. These results are consistent with previous work on genetic admixture in the population of the state of Nuevo León (Garza-Chapa 1983; Cerda-Flores et al. 1987; Cerda-Flores and Garza-Chapa 1989). Also, it is interesting to note that the computations of genetic admixture are similar to those obtained in our previous work based on 17 polymorphic loci in Mexican-Americans of Texas (with Spanish and Amerindian contributions of 70.1% and 29.9%, respectively (Cerda-Flores, Kshatriya et al. 1991).

Table 8. Observed and Expected Distribution of the Number of Heterozygous Loci in the Mexicans of Nuevo León, Mexico

Number of Heterozygous Loci	*Number of Individuals*	
	Observed	*Expected*
0	2	1.15
1	11	12.54
2	38	53.63
3	125	118.47
4	168	150.16
5	117	113.01
6	43	49.49
7	6	11.49
8	1	1.07
Total	511	511.00
Mean	3.97	3.97
Variance	1.59	1.75

Goodness of fit $\chi^2_8 = 11.48, p > 0.05$.
95% Confidence interval for variance (1.55, 1.95).

There are, however, some differences between our estimates of ancestral population contribution in these gene pools in the MMA and those of Lisker et al. (1986) in Mexico City (university student population) and Crawford et al. (1979) and Crawford and Devor (1980) in the Tlaxcala valley and the state of Coahuila (populations with indigenous influences). As we have mentioned previously, we selected people who knew the age and birthplace of their four grandparents (MMA, SLP, and ZAC) from the Mestizo population of the MMA. This is a different approach from other investigators.

Historical evidence (Cossio 1925; Montemayor-Hernández 1971; del Hoyo 1979; Hernández-Garza 1973) indicates that, when the Spanish, Portuguese, and Arabs (Sephardic Jews) colonized Nuevo León in 1596, the native Indians were forced to migrate because of the increasing pressure from the colonizers, thus leaving the region primarily to the colonized populations. But these same historians do not mention the influence of black populations in any time period in the state of Nuevo León, only later Tlaxcaltecan Indian, French, German, and United States populations.

Other studies also report little or no influence from African populations (Garza-Chapa et al. 1982; Garza-Chapa 1983; Cerda-Flores et al. 1987; Cerda-Flores and Garza-Chapa 1988, 1989; Cerda-Flores, Arriaga-Rios et al. 1990; Cerda-Flores et al. 1991; González-Quiroga et al. 1990).

This does not rule out the possibility that the Mestizo population of the MMA is influenced by African genes; therefore we considered that this genetic influence could be from Arab-African admixture or from ancestors from the Gulf of Mexico. Sandler et al. (1979) provide evidence of mixture of Africans with Arabs 900 years ago; they studied the Duffy system distribution in an Arab (Jewish) population. This Arab population dominated Spain for 400 years before the colonization of Nuevo León (del Hoyo 1979). Alternatively, there could have been immigration of people whose grandparents were born on the coast of the Gulf of Mexico, where there was a major African influence; in 1610, 150,000 black slaves arrived at the Mexican coast (Lisker 1980).

In another study, González-Quiroga et al. (1985) found that 5 of 752 male neonates in a Mestizo population in the MMA were deficient in G6PD. From the two maternal grandparent birthplaces, we can explain the distribution of the 5 deficiencies: MMA (2/253), SLP (1/146), Coahuila (1/42), and Tamaulipas (1/35). González-Quiroga et al. (1985) compared the frequencies of Coahuila and Tamaulipas and found them to be similar to those described by Lisker et al. (1969) for coastal populations of the Gulf of Mexico, where the African gene was assumed to originate. Later, González-Quiroga et al. (1990) reported 13 of 829 male neonates with jaundice in a Mestizo population of the MMA; 10 of the 13 neonates had variant A−, and their maternal grandparents were from the coast of the Gulf of Mexico. These results are similar to those published by Lisker et al. (1969). The interesting point of this study was that the three other deficiences were found to be B− variant, and their maternal grandparents were born in the MMA, ZAC, and SLP. Therefore González-Quiroga et al. (1990) concluded that there was minimal African influence on the Mestizo population of the MMA. Cerda-Flores, Arriaga-Rios et al. (1990) studied selected Nuevo León populations in which all four grandparents were born in Nuevo León but outside the MMA; no G6PD deficiencies were found in 428 children. Therefore the population outside the MMA has a small or no African influence because these populations are not from the industrial zone where there is a continuous immigration process, as in the MMA.

One can argue that our estimates could have been affected by the choice of gene frequency data on the ancestral populations, ignoring the possibility of a third component, namely, the black contributions in these gene pools. We contend that this is not the case. Although an exact specification of the ancestral allele frequencies is always difficult in any admixture study, our choice of the ancestral allele frequencies has been demonstrated to be adequate for populations of Mexican origin (Chakraborty et al. 1986; Hanis et al. 1991). In the present analysis we entertained a trihybrid model of admixture, incorporating contributions from blacks in addition to those from the Spanish and Mexican Indians.

The details of such an analysis are not presented here because the black admixture turned out to be negative in most cases, suggesting little contribution of the black gene pool to these populations. Crawford et al. (1979) and Crawford and Devor (1980) also found a small ($4.2\% \pm 2.8\%$) contribution of blacks to the Chamizal population. There are also some minor allele frequency estimation errors in the article by Crawford et al. (1979). For example, their estimates of allele frequencies at the ABO locus (shown in Table 4 of their article) do not sum to 1, nor do the reported allele frequencies at the P locus agree with their maximum likelihood estimation. To what extent these discrepancies contribute to their black admixture component is not known. At any rate, our observation of a relatively more pronounced Spanish contribution and an absence of a black component may in part be explained by the medium and higher socioeconomic background of the study sample compared with the samples examined by Crawford et al. (1979), Crawford and Devor (1980), and Lisker et al. (1986).

The expected and observed distributions of the number of heterozygous loci indicate that there is no residual effect of such admixture on the nonrandomness of allelic associations at the polymorphic loci examined here. These findings suggest that the admixture occurred long enough ago that at present the Mexicans of the MMA are a homogeneous group.

On the basis of the genetic data presented here, we conclude that the Mexicans who reside in the MMA, stratified by generation and birthplace of the grandparents, are genetically similar. The findings of the genetic admixture analysis suggest that this population had a predominant influence from the Spanish and a lesser contribution from Mexican Indians. We also observe a differential contribution of the MMA, SLP, and ZAC to the population of Nuevo León, with the predominant contribution being from the MMA. Furthermore, the multilocus heterozygosity distribution suggests that the history of admixture is old enough that the present group of Mexicans in the MMA is homogeneous.

Acknowledgments We express our thanks to Medical Units 5, 25, and 26 of the Instituto Mexicano del Seguro Social in Monterrey, Nuevo León, for facilities to sample and interview the study participants. Statistical analyses were supported by the National Institutes of Health under US Public Health Service Research Grant GM 41399. Special thanks to M.H. Crawford and the three anonymous reviewers for their constructive criticisms and suggestions on an earlier draft.

Received 6 August 1990; revision received 19 October 1990.

Appendix: Allele Frequencies for 9 Genetic Loci

System	*Spanish*	*Mexican Indian*
A	0.310(1)	0.063(9)
B	0.067	0.003
O	0.623	0.934
DCE	0.048(2)	0.022(9)
DCe	0.418	0.626
DcE	0.090	0.330
Dce	0.049	0.000
dCE	0.002	0.000
dCe	0.011	0.000
dcE	0.001	0.000
dce	0.381	0.022
MS	0.243(3)	0.346(9)
Ms	0.311	0.444
NS	0.057	0.080
Ns	0.389	0.130
Fy(a)	0.365(4)	0.820(9)
Fy(b)+Fy	0.635	0.180
Jk(a)	0.537(5)	0.360(9)
Jk(b)+Jk	0.463	0.640
Lua	0.041(6)	0.000(10)
Lub	0.959	1.000
P1	0.540(7)	0.367(7)
P2+p	0.460	0.633
Le	0.699(8)	0.616(11)
le	0.301	0.384
Se	0.517(8)	1.000(11)
se	0.483	0.000

Adapted from Hanis et al. (1991).

Sources: The numbers in parentheses represent the source of the allele frequency data. (1) Mourant et al. (1976), Table 1.1. Weighted average of populations from Extremadura, Galicia and Leon. (2) Mourant et al. (1976), Table 4.13. Weighted average on non-Basques and Galicia (frequencies recomputed from phenotypic counts). (3) Mourant et al. (1976), Table 2.7, Lugo, Galicia, Spain. (4) Tills et al. (1983), Table 8.3.1, Barcelona, Spain. (5) Tills et al. (1983), Table 9.3.1, Barcelona, Spain. (6) Mourant et al. (1976), Table 5.1, Basques. (7) Crawford et al. (1974), Table 5. (8) Mourant et al. (1976), Table 7.7.4, Basques. (9) Niswander et al. (1970), Table 4. (10) Mourant et al. (1976), Table 5.1, Chiapas, Yucatan, Oaxaca, Veracruz. (11) Mourant et al. (1976), Table 7.7.3, Chiapas.

Literature Cited

Brown, A.H.D., M.W. Feldman, and E. Nevo. 1980. Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* 96:523–536.

Cerda-Flores, R.M., and R. Garza-Chapa. 1988. Cambios en las frecuencias de incompatibilidad para ABO y Rh(D) en tres generaciones de la poblacion del area metropolitana de Monterrey, Nuevo León, Mexico. *Arch. Invest. Med. (Mex.)* 19(1):79–89.

Cerda-Flores, R.M., and R. Garza-Chapa. 1989. Variation in the gene frequencies of three generations of humans from Monterrey, Nuevo León, Mexico. *Hum. Biol.* 61:249–261.

Cerda-Flores, R.M., E. Ramirez-Fernandez, and R. Garza-Chapa. 1987. Genetic admixture and distances between populations from Monterrey, Nuevo León, Mexico, and their putative ancestral populations. *Hum. Biol.* 59:31–49.

Cerda-Flores, R.M., V.A. Bautista-Pena, M.A. Rojas-Alvarado, and R. Garza-Chapa. 1991. Polimorfismo genetico en la poblacion de Cerralvo, Nuevo León. *Estud. Antropol. Biol. (Mex.)* (in press).

Cerda-Flores, R.M., G.K. Kshatriya, T.K. Bertin, D. Hewett-Emmett, C.L. Hanis, and R. Chakraborty. 1991. Gene diversity and estimation of genetic admixture among Mexican-Americans of Starr County, Texas. *Ann. Hum. Biol.* (in press).

Cerda-Flores, R.M., G. Arriaga-Rios, J. Munoz-Campos, V.A. Bautista-Pena, M.A. Rojas-Alvarado, G. Gonzalez-Quiroga, C.H. Leal-Garza, and R. Garza-Chapa. 1990. Frecuencia de la ceguera para los colores y de la deficiencia a la enzima glucosa-6-fosfato-deshidrogenasa en poblaciones no industrializadas del Estado de Nuevo León, Mexico. *Arch. Invest. Med. (Mex.)* 21(3).

Chakraborty, R. 1980. Gene diversity analysis in nested subdivided populations. *Genetics* 96:721–726.

Chakraborty, R. 1981. The distribution of the number of heterozygous loci in natural populations. *Genetics* 98:461–466.

Chakraborty, R. 1984. Detection of nonrandom association of alleles from the distribution of the number of heterozygous loci in a sample. *Genetics* 108:719–731.

Chakraborty, R. 1985. Gene identity in racial hybrids and estimation of admixture rates. In *Genetic Microdifferentiation: Human and Other Populations*, Y.R. Ahuja and J.V. Neel, eds. New Delhi: Indian Anthropological Association, 171–180.

Chakraborty, R. 1986. Gene admixture in human populations: Models and predictions. *Yrbk. Phys. Anthropol.* 29:1–43.

Chakraborty, R., M. Haag, N. Ryman, and G. Stahl. 1982. Hierarchical gene diversity analysis and its implication to brown trout population data. *Hereditas* 97:17–21.

Chakraborty, R., R.E. Ferrell, M.P. Stern, S.M. Haffner, H.P. Hazuda, and M. Rosenthal. 1986. Relationship of prevalence of non-insulin-dependent diabetes mellitus to Amerindian admixture in the Mexican Americans of San Antonio, Texas. *Genet. Epidemiol.* 3:435–454.

Cossio, D.A. 1925. *Historia de Nuevo León, evolucion politica y social*, S. Cantu Leal, ed. Monterrey, Nuevo León, Mexico.

Crawford, M.A., F.E. Gottman, and C.A. Gottman. 1970. Microplate system for routine use in blood bank laboratories. *Transfusion* 10:258–263.

Crawford, M.H., and E.J. Devor. 1980. Population structure and admixture in transplanted Tlaxcaltecan populations. *Am. J. Phys. Anthropol.* 52:485–490.

Crawford, M.H., D.D. Dykes, K. Skradski, and H.P. Polesky. 1979. Gene flow and genetic microdifferentiation of a transplanted Tlaxcaltecan Indian population: Saltillo. *Am. J. Phys. Anthropol.* 50:401–412.

Crawford, M.H., W.C. Leyshon, K. Brown, F. Lees, and L. Taylor. 1974. Human biology in Mexico. II. A comparison of blood groups, serum and red cell enzyme frequencies, and genetic distances of the Indian population of Mexico. *Am. J. Phys. Anthropol.* 41:251–268.

Del Hoyo, H. 1979. *Historia del Nuevo Reino de León (1577–1723)*, 2d ed. Mexico: Editorial Libros de Mexico.

Dirección de Estadistica y Procesamiento de Datos del Gobierno de Nuevo León. 1977. *Aspectos demográficos del Estado de Nuevo León.* Monterrey, Nuevo León, Mexico.

Garza-Chapa, R. 1983. Genetic distances for ABO and Rh(D) blood groups in the state of Nuevo León, Mexico. *Soc. Biol.* 30:24–31.

Garza-Chapa, R., C.H. Leal-Garza, and R.M. Cerda-Flores. 1982. Population genetics in the state of Nuevo León, Mexico. V. Frequencies of ABO, Rh(D), MN blood groups and other genetic traits. *Acta Anthropogenet.* 6(4):225–245.

González-Quiroga, G., M. Walle-Cardona, R. Ortiz-Jalomo, and R. Garza-Chapa. 1985. Deficiencia de deshidrogenosa de la glucosa-6-fosfato (G6PD) en varones neonatos en Monterrey, Nuevo León. *Rev. Med. IMSS (Mex.)* 23:247–250.

González-Quiroga, G., J.L. Ramirez-Del Rio, R. Ortiz-Jalomo, R.F. Garcia-Contreras, R.M. Cerda-Flores, B.D. Mata-Cardenas, and R. Garza-Chapa. 1990. Frecuencia relativa de la deficiencia de glucosa-6-fosfato-deshidrogenasa (G6PD) en neonatos ictericos del area metropolitana de Monterrey, Nuevo León. *Arch. Invest. Med. (Mex.)* 21(3).

Hanis, C.L., D. Hewett-Emmett, T.K. Bertin, and W.J. Schull. 1991. The origins of US Hispanics: Implications for diabetes. *Diabetes Care* (in press).

Harpending, H.C., and R.H. Ward. 1982. Chemical systematics and human evolution. In *Biochemical Aspects of Evolutionary Biology*, M. Nitecki, ed. Chicago, Ill.: University of Chicago Press, 213–256.

Hernández-Garza, T.L. 1973. *Breve historia de Nuevo León*, 3d ed. Mexico: Editorial Trillas.

Lapinski, F.J., K.M. Crowley, C.A. Merrit, and J.B. Henry. 1978. Use of microplate methods in paternity testing. *Am. J. Clin. Pathol.* 70:766–769.

Lisker, R. 1980. *Estructura Genética de la Poblacion Mexicana: Aspectos Medicos y Anthropologicos.* Mexico: Salvat.

Lisker, R., M. Cordova, and G. Zarate. 1969. Studies on several genetic hematological traits of the Mexican population. XVI. Hemoglobin S and G-6-PD deficiency in the East Coast. *Am. J. Phys. Anthropol.* 30:349.

Lisker, R., R. Perez-Briceño, J. Granados, V. Babinsky, J. De Rubens, S. Armendares, and L. Buentello. 1986. Gene frequencies and admixture estimates in a Mexico City population. *Am. J. Phys. Anthropol.* 71:203–207.

Montemayor-Hernández, A. 1971. *Historia de Monterrey.* Mexico: Asociacion de Editores de Monterrey, A.C.

Mourant, A.E., A.C. Kopec, and K. Domaniewska-Sobczak. 1976. *The Distribution of the Human Blood Groups and Other Polymorphisms.* London: Oxford University Press.

Nei, M. 1972. Genetic distance between populations. *Am. Natur.* 106:283–292.

Nei, M., and A.K. Roychoudhury. 1974. Sampling variances of heterozygosity and genetic distance. *Genetics* 76:379–390.

Niswander, J.D., K.S. Brown, B.Y. Iba, W.C. Leyshow, and P.L. Workman. 1970. Population studies on Southwestern Indian tribes. I. History, culture, and genetics of the Papago. *Am. J. Hum. Genet.* 22:7–23.

Reed, T.E., and W.J. Schull. 1968. A general maximum likelihood estimation program. *Am. J. Hum. Genet.* 20:579–580.

Sandler, S.G., C. Krovitz, R. Sharon, D. Hermoni, E. Ezckiel, and T. Cohen. 1979. The Duffy blood group system in Israeli Jews and Arabs. *Vox Sang.* 37:41–46.

Snedecor, G.W., and W.G. Cochran. 1967. *Statistical Methods*, 6th ed. Ames, Iowa: Iowa State University Press.

Tills, D., A.C. Kopec, and R.E. Tills. 1983. *The Distribution of the Human Blood Groups and Other Polymorphisms*, suppl. 1. Oxford: Oxford University Press.

ANNALS OF HUMAN BIOLOGY, 1992, VOL. 19, NO. 4, 347–360

Gene diversity and estimation of genetic admixture among Mexican-Americans of Starr County, Texas

R. M. CERDA-FLORES†, G. K. KSHATRIYA‡, T. K. BERTIN§ D. HEWETT-EMMETT§, C. L. HANIS§ and R. CHAKRABORTY¶

† Subjefatura de Investigacion Cientifica, IMMS, Unidad de Investigacion Biomedica del Noreste, Monterrey, Nuevo León, México

‡ Department of Population Genetics, National Institute of Health and Family Welfare, New Delhi, India

§ Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston, Texas, USA

Received August 13 1990; revised April 24 1991

Summary. The Mexican-Americans of Starr County, Texas, classified by sex and birthplace, were studied to determine the extent of genetic variation and contributions from ancestral populations such as Spanish, Amerindian and West African. Using 21 genetic marker systems, genetic distance and diversity analyses indicate that subpopulations of Mexican-Americans in Starr County are similar, and that more than 99% of the total gene diversity (H_T) can be attributed to individual variation within the population. Genetic admixture analysis shows the predominant influence comes from the Spanish, a lesser contribution from Amerindians and a slight one from the West Africans. The contribution of the ancestral population to various subpopulations of the Mexican-Americans of Starr County is similar. The Mexican-Americans of Starr County are similar to the Mexican population from northeastern Mexico. The history of admixture is apparently old enough to have brought the entire Mexican-American gene pool to Hardy-Weinberg equilibrium. There is no non-random association of alleles among the genetic marker systems considered in the present study, in spite of the fact that this population is of admixed origin. These results, in aggregate, suggest genetic homogeneity of the Mexican-Americans of Starr County, Texas, and point towards the utility of this population for genetic and epidemiological studies.

I. Introduction

The first European contact with Mexico occurred in 1519 with the arrival of Hernan Cortez on the eastern coast of the Gulf of Mexico (del Hoyo 1979). In 1583 Don Luis Carvajal founded a new administrative division, 'Nuevo Reino de León', in northern Mexico. Geographically this area, from 1583 to 1824, was composed of the states of Texas, New Mexico, Nuevo León, Coahuila, Tamaulipas, Chihuahua, San Luis Potosi, Zacatecas, Durango, Nayarit and Sinaloa (Montemayor-Hernandez 1971). Later in 1832 the State of Texas dissociated itself from Mexico, and became a separate independent State (Republic of Texas); only to be incorporated into the United States in 1845 (Hernández-Garza 1973).

When the Spanish (31·7%), Portuguese and Sephardic Jews (68·3%), colonized the northeastern section of 'Nuevo Reino de León' (del Hoyo 1979), there was almost no admixture with the nomadic native populations of the region (estimated at 35,000 inhabitants in Nuevo León and 35,000 in Coahuila). The native Indians were forced to leave Nuevo León and Coahuila because of the increasing pressure from the colonizers, thereby abandoning the region principally to the Europeans. However, the Tlaxcaltecan Indians from Central Mexico migrated to this region as a result of an agreement with the colonizers. Subsequently, a considerable degree of admixing of the European and Tlaxcaltecan gene pools occurred (Cossio 1925, Montemayor-

¶ To whom correspondence should be addressed.

0301-4460/92 $3.00

Hernandez 1971, del Hoyo 1979, Hernandez-Garza 1973, Crawford, Leyshon, Brown, Less and Taylor 1974).

The Mexican population (8,740,439) that reside in United States (Mexican-Americans) are distributed principally in the States of Texas (31·5%), New Mexico (2·7%), Arizona (4·5%), Colorado (2·4%), California (41·6%) and Illinois (4·7%) (US Census 1980). Starr County, one of the 254 counties in Texas, is at least 93·8% Mexican-American. Its Mexican population originated principally from the States of Nuevo León and Coahuila before and after the independence of the State of Texas (Hernandez-Garza 1973). The majority of the Starr County population (27,249) is concentrated in the towns of Rio Grande City, Roma-Los Saenz and La Grulla (US Census 1980).

The contemporary gene pool of the Starr County population contains contributions of 61%, 31% and 8% from Spanish, Amerindian and Black, respectively, ancestral populations (Hanis, Chakraborty, Ferrell and Schull 1986, Hanis, Hewett-Emmett, Bertin and Schull 1991b). This study assesses the extent of heterogeneity of the genetic structure within the population classified by birthplace and sex. It is important to do so because of the known parallelism between the amount of Amerindian admixture and the prevalence of several common chronic diseases, including diabetes, gallbladder disease and obesity (Hanis *et al.* 1986; Weiss, Ferrell and Hanis 1984).

2. Materials and methods

Genetic data were collected as part of a larger investigation of the epidemiology of gallbladder disease, where 1004 randomly selected persons had complete physical examinations, during which blood specimens were obtained by venipuncture, between April 1985 and December 1986. Sampling protocols have been described previously (Hanis, Hewett-Emmett, Kubrusly, Maklad, Douglas, Mueller, Barton, Yoshimaru, Kubrusly, Gonzalez and Schull 1991). From the 1004 persons examined, 993 blood samples were analysed; those missing were due to insufficient sample, sample haemolysis or refusal to allow venous puncture. Venous blood samples were drawn into two 10 ml EDTA vacutainers (Becton-Dickinson, Rutherford, NJ). Cells and plasma were separated by centrifugation. Aliquots of plasma were transferred to screwtop vials and frozen within 2 h of collection in the field, and all sample fractions were deposited within 36 h of collection in laboratories in Houston. The samples were typed according to methods described by Ferrell, Bertin, Young, Barton, Murillo and Schull (1978), Ferrell, Chakraborty, Gershowitz, Laughlin and Schull (1981), Hanis, Douglas and Hewett-Emmett (1991a) and Itakura, Matsudate, Sakurai, Hashimoto, Ito, Kanno, Hirata and Nakamura (1986). The following erythrocyte enzymes and plasma proteins were examined: the polymorphic systems ABO, Rh, MNSs, Duffy (Fy), Kell (K), Kidd (Jk), adenylate kinase (AK), haptoglobin (Hp), phosphoglucomutase 1 (PGM1), glutamate pyruvate transaminase (GPT), glyoxalase I (GLO), phosphoglycolate phosphatase (PGP), esterase D (ESTD), acid phosphatase (ACPI), 6-phosphogluconate dehydrogenase (PGD), adenosine deaminase (ADA), group specific component (Gc), apolipoproteins E (APOE) and A-IV (APOAIV); and the monomorphic systems haemoglobin (Hb) and Kell-antigen (Kp). The study population, geographic area, and project are described in Hanis, Ferrell, Barton, Aguilar, Garza-Ibarra, Tulloch, Garcia and Schull (1983), Hanis, Ferrell, Tulloch and Schull (1985) and Hanis *et al.* (1990). The genetic marker data were subdivided by sex and into two groups by birthplace: (1) individuals born in Texas, and (2) individuals born in Mexico. Although the total population also includes Mexican-Americans born

outside Texas and Mexico, the sample size was too small (37 individuals) to provide much information.

The allele frequencies for different systems other than PGM1 were computed by the maximum-likelihood method (Reed and Schull 1968). Allele frequencies for PGM1 were calculated by gene counting. Genetic distances among Mexican-Americans of Starr County, classified by sex and birthplace, were computed by Nei's standard genetic distance (Nei 1972), and their standard errors (SE) by Nei and Roychoudhury's (1974) method. To determine the significance of the genetic distances among the different subpopulations the gene frequency data were compared pairwise by the chi-square statistic (Nei and Roychoudhury 1974). The extent of genetic variation between the subpopulations of Mexican-Americans was assessed using the nested gene diversity computer program (NEGST) developed by Chakraborty, Haag, Ryman and Ståhl (1982). The percentage contribution of ancestral populations to the hybrid populations (the present-day Mexican-Americans) was calculated by the method of Chakraborty (1985, 1986), each group being considered the product of the admixture of three parental populations, Spanish, Amerindian, and West African. To determine whether the proportions of genes received by the subpopulations from their ancestral sources are significantly different from each other, a regression analysis of heterozygosity on genetic distance (Harpending and Ward 1982) was carried out, and the significance of the regression equation was assessed by the method of Snedecor and Cochran (1967). Finally, examination of non-random association of alleles at different genetic loci by the methods of Brown, Feldman and Nevo (1980) and Chakraborty (1981) was intended to show whether any residual effects of admixture remain in the current population, and so make it heterogeneous.

Gene frequency data on the ancestral populations was obtained from the compilation of Mourant *et al.* (1976), the exact sources of which are available in Hanis *et al.* (1991b).

3. Results

3.1. Allele frequency

The allele frequency estimates for the 21 loci (table 1) were used for a goodness-of-fit chi-square test to determine whether the phenotype and genotype frequencies in the Mexican-Americans, and their sex and birthplace subgroups, depart from the Hardy-Weinberg proportions (table 2), omitting the cases where sample sizes were too small (< 10 individuals). The phenotype (genotype) frequencies for most of the loci are in reasonable agreement with their respective Hardy-Weinberg expectations.

Only 11 chi-squares values out of 189 are significant (at $p < 0{\cdot}05$). Some of these involve small observed frequencies (< 5 individuals) of specific phenotypes (e.g. KK phenotype of the Kell blood group in total; 2^+2^+, 2^-1^-, 2^-2^-, 2^+2^-, phenotype of PGM1 in Texas males). The overall pattern of phenotype (genotype) distributions at these 21 loci is in accordance with the Hardy-Weinberg expectations.

3.2. Genetic distance and heterozygosity

Genetic differences among the groups of Mexican-Americans classified by sex and birthplace, and then by birthplace only, were estimated using Nei's standard genetic distances among all pairs of populations and their respective standard errors (table 3). The average heterozygosity (H) among the subpopulations of the Mexican-Americans varies between 33·4% (Texas males) and 34·3% (Texas females), and is 34·1% overall. Since 90·0% of the 21 loci included in the present study are polymorphic, H may not

Table 1. Allele frequencies among Mexican-Americans of Starr County by sex and place of birth.

	Birthplace					
	Texas			Mexico		
System	Males	Females	Total	Males	Females	Total
ABO						
A1	0·153	0·170	0·164	0·141	0·136	0·138
A2	0·031	0·038	0·036	0·018	0·027	0·024
B	0·070	0·063	0·065	0·093	0·076	0·081
O	0·747	0·730	0·735	0·748	0·761	0·757
n	156	362	518	129	302	431
Rh						
DCE	0·006	0·015	0·012	0·011	0·022	0·019
DCe	0·440	0·433	0·435	0·414	0·461	0·447
DcE	0·187	0·168	0·173	0·204	0·167	0·178
Dce	0·077	0·059	0·065	0·050	0·057	0·055
dCE	0·000	0·000	0·000	0·000	0·000	0·000
dCe	0·000	0·000	0·000	0·000	0·000	0·000
dcE	0·000	0·008	0·006	0·000	0·000	0·000
dce	0·291	0·318	0·309	0·321	0·293	0·301
n	156	362	518	128	302	430
MNSs						
MS	0·297	0·294	0·296	0·280	0·296	0·291
Ms	0·383	0·382	0·381	0·349	0·344	0·345
NS	0·072	0·096	0·088	0·068	0·066	0·067
Ns	0·249	0·227	0·235	0·304	0·294	0·297
n	156	360	516	128	301	429
Duffy						
Fy (a)	0·463	0·399	0·417	0·396	0·474	0·450
Fy (b)	0·438	0·504	0·484	0·538	0·455	0·480
Fy	0·099	0·097	0·099	0·066	0·070	0·070
n	151	353	504	125	294	419
Kell						
K	0·010	0·014	0·014	0·016	0·003	0·007
k	0·990	0·986	0·986	0·984	0·997	0·993
n	156	362	518	129	302	431
Kidd						
Jka	0·532	0·504	0·513	0·395	0·454	0·436
Jkb	0·468	0·496	0·487	0·605	0·546	0·564
n	155	359	514	129	302	431
AK						
AK1	0·962	0·977	0·972	0·977	0·977	0·977
AK2	0·038	0·023	0·028	0·023	0·023	0·023
n	156	363	519	129	302	431
ADA						
ADA1	0·968	0·967	0·967	0·961	0·962	0·941
ADA2	0·032	0·033	0·033	0·039	0·038	0·059
n	156	364	520	129	302	451
ESTD						
ESD1	0·889	0·851	0·863	0·840	0·892	0·877
ESD2	0·111	0·149	0·138	0·160	0·108	0·123
n	157	363	520	128	302	430

Table 2. *Continued*

System	Birthplace					
	Texas			Mexico		
	Males	Females	Total	Males	Females	Total
ACP						
ACPa	0·229	0·256	0·248	0·295	0·276	0·282
ACPb	0·752	0·737	0·741	0·671	0·712	0·700
ACPc	0·019	0·007	0·011	0·035	0·012	0·019
n	157	363	520	129	302	431
GPT						
GPT1	0·539	0·458	0·483	0·480	0·455	0·463
GPT2	0·461	0·542	0·517	0·520	0·545	0·537
n	153	348	501	125	290	415
GLO						
GLO1	0·385	0·385	0·385	0·395	0·396	0·396
GLO2	0·615	0·615	0·615	0·605	0·604	0·604
n	157	364	521	129	302	431
PGD						
PGDA	0·974	0·983	0·981	0·992	0·980	0·984
PGDC	0·026	0·017	0·019	0·008	0·020	0·016
n	155	362	517	128	301	429
PGP						
PGP1	0·914	0·859	0·875	0·895	0·887	0·890
PGP2	0·073	0·136	0·117	0·101	0·108	0·106
PGP3	0·013	0·005	0·008	0·004	0·005	0·005
n	156	365	522	129	302	431
Hp						
Hp1	0·404	0·437	0·427	0·388	0·465	0·442
Hp2	0·596	0·563	0·573	0·612	0·535	0·558
n	156	365	522	129	299	428
Gc						
Gc1S	0·510	0·521	0·517	0·547	0·493	0·509
Gc1F	0·186	0·203	0·198	0·213	0·222	0·219
Gc2	0·304	0·277	0·285	0·240	0·285	0·272
n	153	365	518	129	300	429
APOE						
APOE2	0·050	0·033	0·038	0·052	0·032	0·038
APOE3	0·864	0·867	0·866	0·868	0·847	0·853
APOE4	0·086	0·101	0·096	0·080	0·121	0·109
n	151	353	504	125	298	423
APOIV						
ApoAIV1	0·884	0·938	0·918	0·926	0·943	0·939
ApoAIV2	0·105	0·062	0·076	0·074	0·057	0·061
ApoAIV3	0·012	0·000	0·006	0·000	0·000	0·000
n	43	121	165	34	123	157
PGM1						
PGM1+	0·621	0·575	0·589	0·601	0·553	0·567
PGM1−	0·197	0·216	0·211	0·201	0·229	0·221
PGM2+	0·115	0·137	0·130	0·132	0·150	0·144
PGM2−	0·067	0·071	0·070	0·066	0·068	0·067
n	157	365	522	129	301	430

Kell-Kp and HB loci were monomorphic for Kpb and Hb-a allele, respectively.

Table 2. Chi-square test for estimating Hardy-Weinberg equilibrium for the Mexican-Americans of Starr County, Texas by sex and birthplace.

	Texas			Mexico			Total		
System	Males	Females	Total	Males	Females	Total	Males	Females	Total
ABO									
χ^2	2·99	2·32	2·90	2·99	1·57	4·06	1·83	0·18	0·19
n	156	362	518	129	302	431	298	691	989
Rh									
χ^2	2·39	3·94	1·95	6·74	6·07	8·57	2·13	4·65	3·25
n	156	362	518	128	302	430	297	691	988
MNSs									
χ^2	6·57	2·15	2·99	1·58	0·70	0·54	4·35	1·52	2·54
n	156	360	516	128	301	429	297	687	984
Duffy									
χ^2	3·66	1·99	0·00	0·92	3·35	1·25	0·89	0·04	0·11
n	151	353	504	125	294	419	288	674	962
Kell									
χ^2	0·01	10·36†	8·90†	0·03	0·00	0·02	0·05	12·42†	6·44†
n	156	362	518	129	302	431	298	691	989
Kidd									
χ^2	0·00	0·07	0·04	0·09	0·00	0·04	0·21	0·12	0·00
n	155	359	514	129	302	431	297	688	985
AK									
χ^2	0·25	0·21	0·43	0·07	0·17	0·24	0·32	0·46	0·77
n	156	363	519	129	302	431	298	692	990
ADA									
χ^2	0·17	0·99	0·38	0·21	0·47	1·76	0·36	0·05	0·02
n	156	364	520	129	302	451	298	693	991
ESTD									
χ^2	0·59	0·16	0·00	0·22	0·09	0·04	0·01	0·46	0·25
n	157	363	520	128	302	430	298	692	990
ACP									
χ^2	3·25	3·57	0·61	1·95	2·03	2·12	3·69	0·41	2·84
n	157	363	520	129	302	431	299	692	991
GPT									
χ^2	1·29	0·39	0·00	0·01	0·93	0·57	0·03	0·03	0·02
n	153	348	501	125	290	415	291	665	956
GLO									
χ^2	0·19	0·72	0·91	1·36	0·00	0·48	1·71	0·36	1·48
n	157	364	521	129	302	431	299	692	991
PGD									
χ^2	0·11	0·10	0·20	0·01	0·12	0·12	0·09	0·23	0·32
n	155	362	517	128	301	429	296	690	986
PGP									
χ^2	0·38	0·66	1·21	0·21	0·15	0·83	0·58	2·16	2·89
n	156	365	522	129	302	431	299	694	993
Hp									
χ^2	0·31	0·24	0·02	0·05	4·76†	4·27†	0·00	2·55	2·02
n	156	365	522	129	299	428	299	691	990
Gc									
χ^2	0·38	3·35	3·05	0·12	1·20	0·53	0·40	2·88	2·28
n	153	365	518	129	300	429	295	692	987
APOE									
χ^2	1·31	1·53	0·21	0·44	1·76	0·33	1·32	1·66	0·35
n	151	353	504	125	298	423	287	676	963
APOIV									
χ^2	0·74	0·53	6·13	0·21	0·45	0·65	0·88	0·95	7·59†
n	43	121	165	34	123	157	78	250	329
PGM1									
χ^2	14·12†	10·72	18·74†	2·75	10·43	11·48	9·07	12·66†	20·97†
n	157	365	522	129	301	430	299	692	991

d.f.: ABO = 2; Rh = 10; MNSs = 5; ACP, PGP, Gc, APOE, APOIV = 3, Duffy, Kell, Kidd, AK, ADA, ESTD, GPT, GLO, PGD, Hp = 1; PGM1 = 6

† χ^2 significant at $p < 0{\cdot}05$.

Table 3. Standard genetic distances, average heterozygosity and Chi-square values among Mexican-Americans of Starr County by sex and birthplace†

	Males		Females		Total	
	Texas	Mexico	Texas	Mexico	Texas	Mexico
Males						
Texas	33·43 ±5·50	0·32	0·20	0·22		
Mexico	3·54 ±1·62	33·86 ±5·49	0·27	0·24		
Females						
Texas	1·67 ±0·63	2·18 ±1·00	34·34 ±5·53	0·16		
Mexico	2·18 ±0·81	2·20 ±0·77	1·35 ±0·49	33·90 ±5·59		
Total						
Texas					34·11 ±5·52	0·14
Mexico					1·13 ±0·51	34·14 ±5·52

† Figures on the diagonal are the average heterozygosities expressed in percentage (e.g. Texas males = 33·43%); below the diagonal are standard genetic distances in 10^{-3} codon differences per locus and above the diagonal are χ^2 values (polymorphic loci) with d.f. = 35, $p > 0{\cdot}05$.

All computations are based on 18 polymorphic loci (ABO, Rh, MNSs, Duffy, Kell, Kidd, AK, ADA, ESTD, ACP, GPT, GLO, PGD, PGP, Hp, Gc, APOE, and PGM1) and two monomorphic loci (Kp and Hb).

reflect the actual level of genetic variation generally found in human populations. The genetic distances show no significant differentiation, as examined pairwise by the chi-square statistic (Nei and Roychoudhury 1974).

3.3. *Gene diversity*

The total average gene diversity (H_T) of 34·1% (including the two monomorphic loci) and 37·7% (over the polymorphic loci only) mainly (over 99%) can be attributed to individual variation within the population. Only a small contribution to total variability (0·71%) comes from the between-birthplaces level of subdivision. The sex difference is even smaller (0·21% of the total).

3.4. *Genetic admixture*

Table 5 presents the estimated values of admixture based on 17 polymorphic genetic loci, fitting a trihybrid model using the ancestral frequencies shown in the appendix. There is little difference among the Mexican-American subgroups. The Spanish contribution varies from 55·9% for Texas males to 66·2% for Texas females, that from Amerindians varies from 27·6% in Texas females to 34·2% in Mexico females, and the African contribution varies from 5·9% in the Mexico total to 11·7% in Texas males. The Mexican-Americans of Starr County, Texas appear to be hybrid populations with Spanish, Amerindian and West African admixture, with a predominantly Spanish contribution followed by Amerindian and a small West African contribution.

Although the standard errors of these estimates are provided, no rigorous test of homogeneity is possible because of the unknown sampling distribution of the estimated admixture proportions to determine whether the gene-flow from outside is homogeneous. The procedure of Harpending and Ward (1982) was therefore applied. The genetic distance (r_{ii}) of the ith subpopulation from a hypothetical centroid of all subpopul-

Table 4. Gene diversity analysis of allele frequency data from Mexican-Americans in Starr County, Texas.

Locus	G_{ST}†			H_T‡
	Within population	Between birthplaces	Between sex within birthplaces	Total gene diversity
ABO	98·50	1·13	0·37	0·418
Rh	99·36	0·59	0·05	0·696
MNSs	98·31	1·40	0·29	0·699
Duffy	99·15	0·84	0·01	0·550
Kell	99·26	0·62	0·11	0·033
Kidd	99·14	0·81	0·05	0·497
AK	98·78	1·15	0·07	0·074
ADA	98·76	1·24	0·00	0·046
ESTD	99·27	0·65	0·08	0·239
ACP	98·67	1·13	0·20	0·395
GPT	99·46	0·15	0·38	0·500
GLO	99·98	0·02	0·00	0·474
PGD	99·49	0·40	0·10	0·030
PGP	96·07	1·88	2·04	0·211
Hp	99·65	0·18	0·17	0·485
Gc	99·04	0·82	0·13	0·592
APOE	99·23	0·65	0·12	0·261
PGM1	99·73	0·15	0·12	0·592
Mean§	99·08	0·71	0·21	0·377
s.e.	±0·17	±0·13	±0·07	±0·054

† Expressed as percentage of total.
‡ Absolute total gene diversity in the entire sample.
§ Excluding monomorphic loci (Kp and Hb), the absolute total gene (diversity per locus (H_T) including all 20 loci is 0·340±0·055.

Table 5. Percentage contribution from Spanish, Amerindian and West African gene pools to the contemporary Mexican-Americans of Starr County by sex and birthplace.

	Spanish	Amerindian	West African
Males			
Texas	55·95±2·74	32·33±2·36	11·72±1·24
Mexico	64·08±1·42	29·33±1·23	6·59±0·65
Total	58·62±2·47	31·69±2·13	9·68±1·12
Females			
Texas	66·25±0·21	27·57±0·19	6·17±0·10
Mexico	59·30±1·75	34·19±1·51	6·51±0·79
Total	63·77±0·66	30·16±0·57	6·08±0·30
Total			
Texas	63·25±0·94	28·88±0·81	7·88±0·43
Mexico	62·71±1·04	31·37±0·90	5·92±0·47
Total	62·30±1·19	30·55±1·03	7·16±1·00

The computations are done with 17 polymorphic loci (ABO, Rh, MNSs, Duffy, Kell, Kidd, AK, ADA, ESTD, ACP, GPT, GLO, PGD, PGP, Hp, Gc and PGM1). APOE locus data are not used for admixture estimation because allele frequencies at this locus are not available for the appropriate ancestral populations.

ations is related to the average heterozygosity (H_i) of the ith subpopulation. If gene-flow from outside is uniform, then $H_i = b(1 - r_{ii})$, with absolute value of b being equal to $\bar{H}$, the average heterozygosity in the pooled population. Using 18 genetic loci (excluding the two monomorphic Hb and Kp loci, and the APOA IV locus for which a large number of individuals were not typed), analysis of the regression (table 6) of heterozygosity on genetic distance shows it to be consistent with linearity. $\bar{H}$ in the pooled populations (34·06%) does not differ significantly from the regression coefficient (33·97%). The various subpopulations of the Mexican-American population of Starr County are therefore similar in the proportions of the genes they have received from the ancestral populations, which is consistent with the similarity of admixture proportions estimated in the previous section.

Table 6. Average heterozygosity (H_i) and genetic distances from a centroid (r_{ii}) among the Mexican-Americans of Starr County based on 18 polymorphic loci.

Population	$r_{ii} \pm$ SE	$H_i \pm$ SE
Males		
Texas	0·0038 ± 0·0010	0·3343 ± 0·0550
Mexico	0·0048 ± 0·0015	0·3386 ± 0·0549
Females		
Texas	0·0012 ± 0·0003	0·3432 ± 0·0553
Mexico	0·0016 ± 0·0004	0·3390 ± 0·0559

Regression analysis: $H_i = b(1 - r_{ii})$; H_i plotted against $1 - r_{ii}$ through the origins has $t = -0·901$; d.f. = 2, $p > 0·05$.

Regression coefficient through origin (b) = 0·3397 ± 0·0033.

Average heterozygosity in pooled population ($\bar{H}$) = 0·3406 ± 0·0554.

3.5. *Non-random association among genetic loci*

It is well known that the mixture of populations with disparate allele frequencies can produce non-random association of alleles at two or more unlinked loci (Li 1955, Nei and Li 1973; Chakraborty and Weiss 1988). Employing the procedure suggested by Brown *et al.* (1980), from the available genotype data on each individual, we defined a multi-locus genotype for each individual excluding the monomorphic Hb and Kp loci and the APOA IV genotype, for which too few data were available. With respect to the remaining 18 loci, the number of loci was determined for which each of 862 individuals was heterozygous. Comparing the observed distribution with that expected, under the assumption of random association of alleles at different loci using Chakraborty's (1981) algorithm we find (table 7) that the observed distribution is in fair agreement with the observed one (goodness-of-fit χ^2 with 9 d.f. = 19·96, $p > 0·01$). Since the expected distribution involves the observed data at least partially (locus-specific observed heterozygosity values), there are some technical difficulties for determining the degrees of freedom of the above goodness-of-fit statistic, detailed in Chakraborty (1984). However, Brown *et al.* (1980) showed that the expectations of mean and variance of the number of heterozygous loci can be written as functions of locus-specific heterozygosities under the assumption of random association of alleles, and the 95% confidence limit of the observed variance of the number of heterozygous loci can also

Table 7. Observed and expected distribution of the number of heterozygous loci in the Mexican-Americans of Starr County, Texas.

Number of heterozygous loci	Number of individuals	
	Observed	Expected
0–2	6	5·72
3	27	20·73
4	81	58·15
5	125	115·96
6	180	169·61
7	167	185·10
8	137	151·80
9	86	93·48
10	31	42·89
11–18	22	18·56
Total	862	862·00
Mean	6·630	6·836
Variance	3·540	3·331

Goodness-of-fit $\chi^2 = 19{\cdot}96$, d.f. $= 9$, $p > 0{\cdot}01$.
95% Confidence interval for variance (3·024, 3·638).

be calculated. In the Mexican-American data the mean number of heterozygous loci was 6·63 and the variance was 3·54. Their expected values (under random association) are 6·84 and 3·33, respectively. The 95% confidence limits of the variance are 3·02–3·64. Clearly, these suggest no evidence of non-random association of alleles among the 18 polymorphic loci in this population.

4. Discussion and conclusion

The results of genetic distance analysis between various subpopulations of Mexican-Americans indicate that they are similar to each other. Overall, the level of gene diversity (G_{ST}) is small and more than 99% of total gene diversity (H_T) is accounted for by individual variation within the population.

The admixture results are largely consistent with reports for other Mexican-American groups in the United States. Reed (1974), using the Rh blood group system, estimated $32{\cdot}0 \pm 5{\cdot}6$% Amerindian ancestry among the Mexican-Americans of California. Gottlieb and Kimberling's (1979) findings, from a small admixture study in Colorado, showed 60·0% Spanish contribution to the population, indicating a somewhat larger Amerindian component than seen in other studies. Population admixture estimates computed for the Mexican-Americans of San Antonio, Texas, based on skin reflectance (Relethford, Stern, Gaskill and Hazuda 1983), showed the Amerindian contribution to be 46·0%, 27·0% and 18·0%, respectively, among the Barrio, Transition and Suburban Mexican-American neighbourhoods. Based on gene frequency data on 18 genetic loci Chakraborty, Ferrell, Stern, Haffner, Hazuda and Rosenthal (1986) estimated 43·8%, 30·0% and 18·7% Amerindian ancestry, respectively, in the same three social classes.

The results obtained from the trihybrid model in the present investigation indicated around 62·0%, 30·0% and 8·0% contribution from the Spanish, Amerindian and West African gene pools, respectively. Although various studies show some regional as well as social class variation regarding the contribution of Amerindians to the Mexican-Americans of the United States, there was no remarkable heterogeneity in

genetic admixture among subgroups of the Starr County population. Our estimates of admixture are similar to those obtained by Garza-Chapa (1983), Cerda-Flores, Ramirez-Fernandez and Garza-Chapa (1987), Cerda-Flores and Garza-Chapa (1989) and Cerda-Flores, Kshatriya, Barton, Leal-Garza, Garza-Chapa, Schull and Chakraborty (1991) for the Mexicans of Nuevo León in northeastern Mexico. The chi-square statistic for the 21 genetic marker systems to fit the Hardy-Weinberg equilibrium indicates that intermixing is old enough to have eliminated any early non-random association of genes.

In summary, on the basis of the genetic data presented, we conclude that the Mexican-Americans of Starr County, Texas, classified by sex and birthplace, are not genetically distinguishable. The findings on genetic admixture indicate a predominant influence from the Spanish, and lesser contributions from Amerindians and West Africans. The history of admixture is apparently old enough to have brought the entire Mexican-American gene pool to Hardy-Weinberg equilibrium. The multi-locus heterozygosity distribution also supports the inference of genetic homogeneity.

These findings have a number of important implications with respect to the utility of such populations in anthropogenetic and epidemiological contexts. First, the demonstration of genetic homogeneity of the Mexican-Americans of Starr County, Texas, in spite of their admixed origin, suggests that this population is suitable for studying disease-marker associations in the search of candidate genes of complex diseases; such an association cannot possibly arise from population mixture alone (Chakraborty and Weiss 1988). Secondly, in spite of the polyphyletic origin of the Mexican-Americans, their multi-locus genotypic distribution satisfies the premises of random segregation of unlinked loci. Therefore, the probability of finding a specific multiple-locus genotype in such a population can be determined by the product rule of locus-specific genotype probabilities, contrary to the recent claim of Cohen (1990). The 862 individuals on which we had the 18-locus genotype data available constitute 862 different multiple-locus genotypes; i.e. no repeat of any multiple-locus genotype was observed in the sample. Based on the allelic frequencies the most probable multiple-locus genotype in this population would be encountered once in every 885,764 individuals. This shows that the discretized genotypic information in such a population is sufficient to determine the identity of individuals even when the population is of admixed origin.

Acknowledgements

This work was supported in part by the US Public Health Service Research Grants DK 34666, DK01748 and GM 41399 from the National Institutes of Health. We thank an anonymous reviewer for extensive editorial and other constructive suggestions on this paper.

References

Brown, A. H. D., Feldman, M. W., and Nevo, E., 1980, Multilocus structure of natural populations of Hordeum spontaneum. *Genetics*, 96, 523–536.

Cerda-Flores, R. M., and Garza-Chapa, R., 1989, Variation in the gene frequencies of three generations of humans from Monterrey, Nuevo León, Mexico. *Human Biology*, 61, 249–261.

Cerda-Flores, R. M., Ramirez-Fernandez, E., and Garza-Chapa, R., 1987, Genetic admixture and distances between populations from Monterrey, Nuevo León, Mexico and their putative ancestral populations. *Human Biology*, 59, 31–49.

Cerda-Flores, R. M., Kshatriya, G. K., Barton, S. A., Leal-Garza, C. H., Garza-Chapa, R., Schull, W. J., and Chakraborty, R., 1991, Genetic structure of the immigrant populations of San Luis Potosi and Zacatecas to Nuevo León in Mexico. *Human Biology*, 63, 309–327.

CHAKRABORTY, R., 1981, The distribution of the number of heterozygous loci in natural populations. *Genetics*, **98**, 461–466.

CHAKRABORTY, R., 1984, Detection of non-random association of alleles from the distribution of the number of heterozygous loci in a sample. *Genetics*, **108**, 719–731.

CHAKRABORTY, R., 1985, Gene identity in racial hybrids and estimation of admixture rates. In *Genetic Microdifferentiation-Human and Other Populations*, edited by Y. R. Ahuja and J. V. Neel (New Delhi: Indian Anthropological Association), pp. 171–180.

CHAKRABORTY, R., 1986, Gene admixture in human populations: Models and predications. *Yearbook of Physical Anthropology*, **29**, 1–43.

CHAKRABORTY, R., and WEISS, K. M., 1988, Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the Academy of Sciences, USA*, **85**, 9119–9123.

CHAKRABORTY, R., HAAG, M., RYMAN, N., and STÅHL, G., 1982, Hierarchical gene diversity analysis and its implication to brown trout population data. *Hereditas*, **97**, 17–21.

CHAKRABORTY, R., FERRELL, R. E., STERN, M. P., HAFFNER, S. M., HAZUDA, H. P., and ROSENTHAL, M., 1986, Relationship of prevalence of non-insulin dependent diabetes mellitus with Amerindian admixture in Mexican-Americans of San Antonio, Texas, *Genetic Epidemiology*, **3**, 435–454.

COHEN, J. E., 1990, DNA fingerprinting for forensic identification: Potential effects on data interpretation of subpopulation heterogeneity and band number variability. *American Journal of Human Genetics*, **46**, 358–368.

COSSIO, D. A., 1925, *Historia de Nuevo León, Evolucion Politica y Social*, edited by S. Cantu Leal (Mexico: Monterrey, Nuevo León).

CRAWFORD, M. H., LEYSHON, W. C., BROWN, K., LESS, F., and TAYLOR, L., 1974, Human biology in Mexico. II. A comparison of blood group, serum and red cell enzyme frequencies and genetic distances of the Indian populations of Mexico. *American Journal of Physical Anthropology*, **41**, 251–268.

DEL HOYO, H., 1979, *Historia del Nuevo Reino de León (1577–1723)* (Mexico: Editorial Libros de Mexico, S. A.), 2nd edn.

FERREL, R. E., BERTIN, T., YOUNG, R., BARTON, S. A., MURILLO, F., and SCHULL, W. J., 1978, The Aymará of Western Bolivia. IV. Gene frequencies of eight blood groups and 19 protein and erythrocyte enzyme systems. *American Journal of Human Genetics*. **30**, 539–549.

FERRELL, R. E., CHAKRABORTY, R., GERSHOWITZ, H., LAUGHLIN, W. S., and SCHULL, W. J., 1981, The St. Lawrence Island Eskimos. Genetic variation and genetic distance. *American Journal of Physical Anthropology*, **55**, 351–358.

GARZA-CHAPA, R., 1983, Genetic distances for ABO and Rh(D) blood groups in the State of Nuevo León, Mexico. *Social Biology*, **30**, 24–31.

GOTTLIEB, K., and KIMBERLING, W. J., 1979, Admixture estimates for the gene pool of Mexican-Americans in Colorado. *Abstracts, Forty-eighth Annual Meeting of the American Association of Physical Anthropologists*, San Francisco, California, p. 444.

HANIS, C. L., FERRELL, R. E., BARTON, S. A., AGUILAR, L., GARZA-IBARRA, A., TULLOCH, B. R., GARCIA, C. A., and SCHULL, W. J., 1983, Diabetes among Mexican Americans in Starr County, Texas. *American Journal of Epidemiology*, **118**, 659–672.

HANIS, C. L., FERRELL, R. E., TULLOCH, B. R., and SCHULL, W. J., 1985, Gallbladder disease epidemiology in Mexican Americans in Starr County. *American Journal of Epidemiology*, **122**, 820–829.

HANIS, C. L., CHAKRABORTY, R., FERRELL, R. E., and SCHULL, W. J., 1986, Individual admixture estimates: Disease associations and individual risk of diabetes and gallbladder disease among Mexican-Americans in Starr County, Texas. *American Journal of Physical Anthropology*, **70**, 433–441.

HANIS, C. L., HEWETT-EMMETT, D., KUBRUSLY, L. F., MAKLAD, M. N., DOUGLAS, T. C., MUELLER, W. H., BARTON, S. A., YOSHIMARU, H., KUBRUSLY, D. B., GONZALEZ, R., and SCHULL, W. J., 1991, An ultrasound survey of gallbladder disease among Mexican-Americans in Starr County, Texas: Associations with obesity, diabetes, hypertension, lipids, lipoproteins and apolipoproteins. (Submitted).

HANIS, C. L., DOUGLAS, T. C., and HEWETT-EMMETT, D., 1991a, Apolipoprotein A-IV protein polymorphism: Frequency and effects of lipids, lipoproteins and apolipoproteins among Mexican-Americans in Starr County, Texas. *Human Genetics*, **86**, 323–325.

HANIS, C. L., HEWETT-EMMETT, D., BERTIN, T. K., and SCHULL, W. J., 1991b, The origins of U.S. Hispanics: Implications for diabetes. *Diabetes Care*, **14**, 618–627.

HARPENDING, H. C., and WARD, R. H., 1982, Chemical systematics and human evolution. In *Biochemical Aspects of Evolutionary Biology*, edited by M. Nitecki (Chicago: University of Chicago Press), pp. 213–256.

HERNANDEZ-GARZA, T. L., 1973, *Breve Historia de Nuevo León* (Mexico: Editorial Trillas), 3rd edn.

ITAKURA, K., MATSUDATE, T., SAKURAI, T., HASHIMOTO, S., ITO, K., KANNO, H., HIRATA, M., and NAKAMURA, K., 1986, Single radial immunodiffusion of serum apolipoproteins C-II, C-III and E—pretreatment of samples with surfactant. *Clinica et Chimica Acta*, **161**, 275–282.

LI, C.C., 1955, *Population Genetics* (Chicago: University of Chicago Press).
MONTEMAYOR-HERNANDEZ, A., 1971, *Historia de Monterrey* (Mexico: Asociacion de Editores de Monterrey, AC).
MOURANT, A.E., KOPEC, A.C., and DOMANIEWSKA-SOBCZEK, K., 1976, *The Distribution of the Human Blood Groups and Other Polymorphisms* (London: Oxford University Press).
NEI, M., 1972, Genetic distance between populations. *American Naturalist*, 106, 283–292.
NEI, M., and LI, W.H., 1973, Linkage disequilibrium in subdivided populations. *Genetics*, 75, 213–219.
NEI, M., and ROYCHOUDHURY, A.K., 1974, Sampling variances of heterozygosity and genetic distance. *Genetics*, 76, 379–390.
REED, T.E., 1974, Ethnic classification of Mexican-Americans. *Science*, 185, 283.
REED, T.E., and SCHULL, W.J., 1968, A general maximum likelihood estimation program. *American Journal of Human Genetics*, 20, 579–580.
RELETHFORD, J.H., STERN, M.P., GASKILL, S.P., and HAZUDA, H.P., 1983, Social class, admixture, and skin color variation in Mexican-Americans and Anglo-Americans living in San Antonio, Texas. *American Journal of Physical Anthropology*, 61, 97–102.
SNEDECOR, G.W., and COCHRAN, W.G., 1967, *Statistical Methods* (Ames, IO: Iowa State University Press), 6th edn.
US BUREAU OF THE CENSUS, 1980 (1982), *General Population Characteristics: Texas. Census of Population*, Vol. 1, Part 45, PC80-1-B45 (Washington, DC: US Government Printing Office).
WEISS, K.M., FERRELL, R.E., and HANIS, C.L., 1984, A new world syndrome of metabolic disease with a genetic and evolutionary basis. *Yearbook of Physical Anthropology*, 27, 153–178.

Zusammenfassung. Die nach Geschlecht und Geburtsort klassifizierten Mexican-Americans von Starr County, Texas, wurden im Hinblick auf das Ausmaß ihrer genetischen Differenzierung und der Beiträge von ancestralen Populationen, wie Spanier, amerikanischer Indianer und Westafrikaner untersucht. Unter Berücksichtigung von 21 genetischen Markern zeigten genetische Abstandsanalysen und Gene Diversity-Analysen, daß Subpopulationen der Mexican-Americans in Starr County einander genetisch ähnlich sind und daß mehr als 99% der insgesamt beobachteten Gene Diversity (H_T) auf individuelle Variationen in der Population zurückgeführt werden können. Eine Analyse der genetischen Durchmischung belegt einen prädominanten Einfluß der Spanier, einen geringeren Beitrag der amerikanischen Indianer und einen geringfügigen der Westafrikaner. Der genetische Beitrag der ancestralen Populationen zu den verschiedenen Subpopulationen der Mexican-Americans in Starr County ist ähnlich. Die Mexican-Americans von Starr County sind den mexikanischen Populationen aus dem Nordosten Mexicos ähnlich. Der Prozeß der Durchmischung reicht historisch offensichtlich soweit zurück, daß die Zeit ausreichend war, um den gesamten Genpool der Mexican-Americans ins Hardy-Weinberg-Gleichgewicht zu bringen. Obwohl diese Population per Durchmischung enstanden ist, gibt es keine nichtzufälligen Assoziationen der Allele zwischen den genetischen Markersystemen, die in dieser Studie berücksichtigt wurden. Insgesamt belegen die Ergebnisse die genetische Homogenität der Mexican-Americans von Starr County, Texas, und weisen auf die Bedeutung dieser Population für genetische und epidemiologische Studien hin.

Résumé. Les mexicains-américains du comté de Starr au Texas, répertoriés par sexe et lieu de naissance, ont été étudiés afin de déterminer l'étendue de leur variation génétique ainsi que des contributions qu'ils ont recues de populations ancestrales telles qu'espagnole, amérindienne et ouest-africaine. Les distances génétiques et les analyses de diversité effectuées à partir de 21 marqueurs génétiques, indiquent que les sous-groupes de mexicains-américains du comté de Starr sont similaires et que plus de 99% de la diversité génétique totale (H_T) peuvent être attribués à des variations individuelles dans la population. Les analyses d'apports génétique montrent que l'influence dominante provient des espagnols, à un moindre degré des amérindiens et dans une faible mesure des ouest-africains. La contribution de la population ancestrale aux diverses sous-populations est similaire. Les mexicains-américains du comté de Starr sont semblables aux populations mexicaines du nord-est du Mexique. L'histoire du métissage est apparemment assez ancienne pour avoir permis l'atteinte de l'équilibre d'Hardy-Weinberg, par l'ensemble du patrimoine génétique mexicain-américain. Il n'y a pas d'association non-alléatoire d'allèles parmi les systèmes considérés dans cette étude, en dépit de l'origine composite de cette population. Ces résultats montrent l'homogénéité génétique des mexicains-américains du comté de Starr au Texas et indiquent l'utilité de cette population pour des études épidémiologiques et génétiques.

Appendix. Allele frequencies for 17 genetic loci in Mexican-Americans of Starr County and putative ancestral populations†

Allele	Mexican-American			Spanish	Amerindian	West African
	Males	Females	Total			
A	0·174	0·184	0·181	0·310	0·063	0·191
B	0·077	0·073	0·074	0·067	0·003	0·213
O	0·749	0·743	0·745	0·623	0·934	0·596
DCE	0·011	0·020	0·017	0·048	0·022	0·002
DCe	0·429	0·443	0·438	0·418	0·626	0·028
DcE	0·191	0·166	0·174	0·090	0·330	0·084
Dce	0·068	0·064	0·065	0·049	0·000	0·637
dCE	0·000	0·000	0·000	0·002	0·000	0·000
dCe	0·000	0·000	0·000	0·011	0·000	0·007
dcE	0·000	0·004	0·003	0·001	0·000	0·000
dce	0·302	0·303	0·303	0·380	0·022	0·242
MS	0·287	0·299	0·296	0·243	0·346	0·088
Ms	0·365	0·360	0·361	0·311	0·444	0·432
NS	0·067	0·080	0·075	0·057	0·080	0·138
Ns	0·282	0·262	0·268	0·389	0·130	0·342
Fy (a)	0·436	0·434	0·435	0·365	0·820	0·002
Fy (b)	0·484	0·482	0·483	0·635	0·180	0·000
Fy	0·080	0·084	0·083	0·000	0·000	0·998
K	0·013	0·010	0·011	0·038	0·000	0·003
k	0·987	0·990	0·989	0·962	1·000	0·997
Jka	0·468	0·480	0·476	0·537	0·360	0·433
Jkb	0·532	0·520	0·524	0·463	0·640	0·567
AK1	0·968	0·975	0·973	0·978	1·000	0·993
AK2	0·032	0·025	0·027	0·022	0·000	0·007
ADA1	0·966	0·966	0·966	0·951	1·000	1·000
ADA2	0·034	0·034	0·034	0·049	0·000	0·000
ESD1	0·867	0·868	0·868	0·879	0·716	0·938
ESD2	0·133	0·132	0·132	0·121	0·284	0·062
ACPa	0·254	0·265	0·262	0·298	0·203	0·180
ACPb	0·721	0·725	0·724	0·666	0·797	0·820
ACPc	0·025	0·009	0·014	0·036	0·000	0·000
GPT1	0·514	0·456	0·473	0·505	0·442	0·863
GPT2	0·486	0·544	0·527	0·495	0·558	0·137
GLO1	0·390	0·391	0·391	0·423	0·211	0·309
GLO2	0·610	0·609	0·609	0·577	0·789	0·691
PGDA	0·983	0·982	0·982	0·975	0·997	0·937
PGDC	0·017	0·018	0·018	0·025	0·003	0·063
PGP1	0·908	0·868	0·880	0·927	0·690	1·000
PGP2	0·082	0·126	0·113	0·047	0·310	0·000
PGP3	0·010	0·006	0·007	0·026	0·000	0·000
Hp1	0·396	0·448	0·432	0·429	0·452	0·654
Hp2	0·604	0·552	0·568	0·571	0·548	0·346
Gc1	0·725	0·720	0·721	0·641	0·859	0·916
Gc2	0·275	0·280	0·279	0·359	0·141	0·084
PGM1+	0·610	0·566	0·579	0·621	0·485	0·787
PGM1−	0·199	0·223	0·215	0·114	0·342	0·039
PGM2+	0·124	0·144	0·138	0·211	0·020	0·147
PGM2−	0·067	0·068	0·068	0·054	0·153	0·027

† The allele frequencies for the Mexican-Americans are from the present survey (summing over all individuals) and those for the ancestral populations are compiled from the literature (see Hanis *et al.* 1991b for exact sources from Mourant *et al.*'s 1976 compilation).

ANNALS OF HUMAN BIOLOGY, 1992, VOL. 19, NO. 4, 347-360

Gene diversity and estimation of genetic admixture among Mexican-Americans of Starr County, Texas

R. M. CERDA-FLORES†, G. K. KSHATRIYA‡, T. K. BERTIN§ D. HEWETT-EMMETT§, C. L. HANIS§ and R. CHAKRABORTY¶

† Subjefatura de Investigacion Cientifica, IMMS, Unidad de Investigacion Biomedica del Noreste, Monterrey, Nuevo León, México

‡ Department of Population Genetics, National Institute of Health and Family Welfare, New Delhi, India

§ Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston, Texas, USA

Received August 13 1990; revised April 24 1991

Summary. The Mexican-Americans of Starr County, Texas, classified by sex and birthplace, were studied to determine the extent of genetic variation and contributions from ancestral populations such as Spanish, Amerindian and West African. Using 21 genetic marker systems, genetic distance and diversity analyses indicate that subpopulations of Mexican-Americans in Starr County are similar, and that more than 99% of the total gene diversity (H_T) can be attributed to individual variation within the population. Genetic admixture analysis shows the predominant influence comes from the Spanish, a lesser contribution from Amerindians and a slight one from the West Africans. The contribution of the ancestral population to various subpopulations of the Mexican-Americans of Starr County is similar. The Mexican-Americans of Starr County are similar to the Mexican population from northeastern Mexico. The history of admixture is apparently old enough to have brought the entire Mexican-American gene pool to Hardy-Weinberg equilibrium. There is no non-random association of alleles among the genetic marker systems considered in the present study, in spite of the fact that this population is of admixed origin. These results, in aggregate, suggest genetic homogeneity of the Mexican-Americans of Starr County, Texas, and point towards the utility of this population for genetic and epidemiological studies.

I. Introduction

The first European contact with Mexico occurred in 1519 with the arrival of Hernan Cortez on the eastern coast of the Gulf of Mexico (del Hoyo 1979). In 1583 Don Luis Carvajal founded a new administrative division, 'Nuevo Reino de León', in northern Mexico. Geographically this area, from 1583 to 1824, was composed of the states of Texas, New Mexico, Nuevo León, Coahuila, Tamaulipas, Chihuahua, San Luis Potosi, Zacatecas, Durango, Nayarit and Sinaloa (Montemayor-Hernandez 1971). Later in 1832 the State of Texas dissociated itself from Mexico, and became a separate independent State (Republic of Texas); only to be incorporated into the United States in 1845 (Hernandez-Garza 1973).

When the Spanish (31·7%), Portuguese and Sephardic Jews (68·3%), colonized the northeastern section of 'Nuevo Reino de León' (del Hoyo 1979), there was almost no admixture with the nomadic native populations of the region (estimated at 35,000 inhabitants in Nuevo León and 35,000 in Coahuila). The native Indians were forced to leave Nuevo León and Coahuila because of the increasing pressure from the colonizers, thereby abandoning the region principally to the Europeans. However, the Tlaxcaltecan Indians from Central Mexico migrated to this region as a result of an agreement with the colonizers. Subsequently, a considerable degree of admixing of the European and Tlaxcaltecan gene pools occurred (Cossio 1925, Montemayor-

¶ To whom correspondence should be addressed.

Hernandez 1971, del Hoyo 1979, Hernandez-Garza 1973, Crawford, Leyshon, Brown, Less and Taylor 1974).

The Mexican population (8,740,439) that reside in United States (Mexican-Americans) are distributed principally in the States of Texas (31·5%), New Mexico (2·7%), Arizona (4·5%), Colorado (2·4%), California (41·6%) and Illinois (4·7%) (US Census 1980). Starr County, one of the 254 counties in Texas, is at least 93·8% Mexican-American. Its Mexican population originated principally from the States of Nuevo León and Coahuila before and after the independence of the State of Texas (Hernandez-Garza 1973). The majority of the Starr County population (27,249) is concentrated in the towns of Rio Grande City, Roma-Los Saenz and La Grulla (US Census 1980).

The contemporary gene pool of the Starr County population contains contributions of 61%, 31% and 8% from Spanish, Amerindian and Black, respectively, ancestral populations (Hanis, Chakraborty, Ferrell and Schull 1986, Hanis, Hewett-Emmett, Bertin and Schull 1991b). This study assesses the extent of heterogeneity of the genetic structure within the population classified by birthplace and sex. It is important to do so because of the known parallelism between the amount of Amerindian admixture and the prevalence of several common chronic diseases, including diabetes, gallbladder disease and obesity (Hanis *et al.* 1986; Weiss, Ferrell and Hanis 1984).

2. Materials and methods

Genetic data were collected as part of a larger investigation of the epidemiology of gallbladder disease, where 1004 randomly selected persons had complete physical examinations, during which blood specimens were obtained by venipuncture, between April 1985 and December 1986. Sampling protocols have been described previously (Hanis, Hewett-Emmett, Kubrusly, Maklad, Douglas, Mueller, Barton, Yoshimaru, Kubrusly, Gonzalez and Schull 1991). From the 1004 persons examined, 993 blood samples were analysed; those missing were due to insufficient sample, sample haemolysis or refusal to allow venous puncture. Venous blood samples were drawn into two 10 ml EDTA vacutainers (Becton-Dickinson, Rutherford, NJ). Cells and plasma were separated by centrifugation. Aliquots of plasma were transferred to screwtop vials and frozen within 2 h of collection in the field, and all sample fractions were deposited within 36 h of collection in laboratories in Houston. The samples were typed according to methods described by Ferrell, Bertin, Young, Barton, Murillo and Schull (1978), Ferrell, Chakraborty, Gershowitz, Laughlin and Schull (1981), Hanis, Douglas and Hewett-Emmett (1991a) and Itakura, Matsudate, Sakurai, Hashimoto, Ito, Kanno, Hirata and Nakamura (1986). The following erythrocyte enzymes and plasma proteins were examined: the polymorphic systems ABO, Rh, MNSs, Duffy (Fy), Kell (K), Kidd (Jk), adenylate kinase (AK), haptoglobin (Hp), phosphoglucomutase 1 (PGM1), glutamate pyruvate transaminase (GPT), glyoxalase I (GLO), phosphoglycolate phosphatase (PGP), esterase D (ESTD), acid phosphatase (ACPI), 6-phosphogluconate dehydrogenase (PGD), adenosine deaminase (ADA), group specific component (Gc), apolipoproteins E (APOE) and A-IV (APOAIV); and the monomorphic systems haemoglobin (Hb) and Kell-antigen (Kp). The study population, geographic area, and project are described in Hanis, Ferrell, Barton, Aguilar, Garza-Ibarra, Tulloch, Garcia and Schull (1983), Hanis, Ferrell, Tulloch and Schull (1985) and Hanis *et al.* (1990). The genetic marker data were subdivided by sex and into two groups by birthplace: (1) individuals born in Texas, and (2) individuals born in Mexico. Although the total population also includes Mexican-Americans born

outside Texas and Mexico, the sample size was too small (37 individuals) to provide much information.

The allele frequencies for different systems other than PGM1 were computed by the maximum-likelihood method (Reed and Schull 1968). Allele frequencies for PGM1 were calculated by gene counting. Genetic distances among Mexican-Americans of Starr County, classified by sex and birthplace, were computed by Nei's standard genetic distance (Nei 1972), and their standard errors (SE) by Nei and Roychoudhury's (1974) method. To determine the significance of the genetic distances among the different subpopulations the gene frequency data were compared pairwise by the chi-square statistic (Nei and Roychoudhury 1974). The extent of genetic variation between the subpopulations of Mexican-Americans was assessed using the nested gene diversity computer program (NEGST) developed by Chakraborty, Haag, Ryman and Ståhl (1982). The percentage contribution of ancestral populations to the hybrid populations (the present-day Mexican-Americans) was calculated by the method of Chakraborty (1985, 1986), each group being considered the product of the admixture of three parental populations, Spanish, Amerindian, and West African. To determine whether the proportions of genes received by the subpopulations from their ancestral sources are significantly different from each other, a regression analysis of heterozygosity on genetic distance (Harpending and Ward 1982) was carried out, and the significance of the regression equation was assessed by the method of Snedecor and Cochran (1967). Finally, examination of non-random association of alleles at different genetic loci by the methods of Brown, Feldman and Nevo (1980) and Chakraborty (1981) was intended to show whether any residual effects of admixture remain in the current population, and so make it heterogeneous.

Gene frequency data on the ancestral populations was obtained from the compilation of Mourant *et al.* (1976), the exact sources of which are available in Hanis *et al.* (1991b).

3. Results

3.1. Allele frequency

The allele frequency estimates for the 21 loci (table 1) were used for a goodness-of-fit chi-square test to determine whether the phenotype and genotype frequencies in the Mexican-Americans, and their sex and birthplace subgroups, depart from the Hardy-Weinberg proportions (table 2), omitting the cases where sample sizes were too small (< 10 individuals). The phenotype (genotype) frequencies for most of the loci are in reasonable agreement with their respective Hardy-Weinberg expectations.

Only 11 chi-squares values out of 189 are significant (at $p < 0{\cdot}05$). Some of these involve small observed frequencies (< 5 individuals) of specific phenotypes (e.g. KK phenotype of the Kell blood group in total; 2^+2^+, 2^-1^-, 2^-2^-, 2^+2^-, phenotype of PGM1 in Texas males). The overall pattern of phenotype (genotype) distributions at these 21 loci is in accordance with the Hardy-Weinberg expectations.

3.2. Genetic distance and heterozygosity

Genetic differences among the groups of Mexican-Americans classified by sex and birthplace, and then by birthplace only, were estimated using Nei's standard genetic distances among all pairs of populations and their respective standard errors (table 3). The average heterozygosity ($\bar{H}$) among the subpopulations of the Mexican-Americans varies between 33·4% (Texas males) and 34·3% (Texas females), and is 34·1% overall. Since 90·0% of the 21 loci included in the present study are polymorphic, $\bar{H}$ may not

Table 1. Allele frequencies among Mexican-Americans of Starr County by sex and place of birth.

	Birthplace					
	Texas			Mexico		
System	Males	Females	Total	Males	Females	Total
ABO						
A1	0·153	0·170	0·164	0·141	0·136	0·138
A2	0·031	0·038	0·036	0·018	0·027	0·024
B	0·070	0·063	0·065	0·093	0·076	0·081
O	0·747	0·730	0·735	0·748	0·761	0·757
n	156	362	518	129	302	431
Rh						
DCE	0·006	0·015	0·012	0·011	0·022	0·019
DCe	0·440	0·433	0·435	0·414	0·461	0·447
DcE	0·187	0·168	0·173	0·204	0·167	0·178
Dce	0·077	0·059	0·065	0·050	0·057	0·055
dCE	0·000	0·000	0·000	0·000	0·000	0·000
dCe	0·000	0·000	0·000	0·000	0·000	0·000
dcE	0·000	0·008	0·006	0·000	0·000	0·000
dce	0·291	0·318	0·309	0·321	0·293	0·301
n	156	362	518	128	302	430
MNSs						
MS	0·297	0·294	0·296	0·280	0·296	0·291
Ms	0·383	0·382	0·381	0·349	0·344	0·345
NS	0·072	0·096	0·088	0·068	0·066	0·067
Ns	0·249	0·227	0·235	0·304	0·294	0·297
n	156	360	516	128	301	429
Duffy						
Fy (a)	0·463	0·399	0·417	0·396	0·474	0·450
Fy (b)	0·438	0·504	0·484	0·538	0·455	0·480
Fy	0·099	0·097	0·099	0·066	0·070	0·070
n	151	353	504	125	294	419
Kell						
K	0·010	0·014	0·014	0·016	0·003	0·007
k	0·990	0·986	0·986	0·984	0·997	0·993
n	156	362	518	129	302	431
Kidd						
Jka	0·532	0·504	0·513	0·395	0·454	0·436
Jkb	0·468	0·496	0·487	0·605	0·546	0·564
n	155	359	514	129	302	431
AK						
AK1	0·962	0·977	0·972	0·977	0·977	0·977
AK2	0·038	0·023	0·028	0·023	0·023	0·023
n	156	363	519	129	302	431
ADA						
ADA1	0·968	0·967	0·967	0·961	0·962	0·941
ADA2	0·032	0·033	0·033	0·039	0·038	0·059
n	156	364	520	129	302	451
ESTD						
ESD1	0·889	0·851	0·863	0·840	0·892	0·877
ESD2	0·111	0·149	0·138	0·160	0·108	0·123
n	157	363	520	128	302	430

Table 2. *Continued*

System	Birthplace					
	Texas			Mexico		
	Males	Females	Total	Males	Females	Total
ACP						
ACPa	0·229	0·256	0·248	0·295	0·276	0·282
ACPb	0·752	0·737	0·741	0·671	0·712	0·700
ACPc	0·019	0·007	0·011	0·035	0·012	0·019
n	157	363	520	129	302	431
GPT						
GPT1	0·539	0·458	0·483	0·480	0·455	0·463
GPT2	0·461	0·542	0·517	0·520	0·545	0·537
n	153	348	501	125	290	415
GLO						
GLO1	0·385	0·385	0·385	0·395	0·396	0·396
GLO2	0·615	0·615	0·615	0·605	0·604	0·604
n	157	364	521	129	302	431
PGD						
PGDA	0·974	0·983	0·981	0·992	0·980	0·984
PGDC	0·026	0·017	0·019	0·008	0·020	0·016
n	155	362	517	128	301	429
PGP						
PGP1	0·914	0·859	0·875	0·895	0·887	0·890
PGP2	0·073	0·136	0·117	0·101	0·108	0·106
PGP3	0·013	0·005	0·008	0·004	0·005	0·005
n	156	365	522	129	302	431
Hp						
Hp1	0·404	0·437	0·427	0·388	0·465	0·442
Hp2	0·596	0·563	0·573	0·612	0·535	0·558
n	156	365	522	129	299	428
Gc						
Gc1S	0·510	0·521	0·517	0·547	0·493	0·509
Gc1F	0·186	0·203	0·198	0·213	0·222	0·219
Gc2	0·304	0·277	0·285	0·240	0·285	0·272
n	153	365	518	129	300	429
APOE						
APOE2	0·050	0·033	0·038	0·052	0·032	0·038
APOE3	0·864	0·867	0·866	0·868	0·847	0·853
APOE4	0·086	0·101	0·096	0·080	0·121	0·109
n	151	353	504	125	298	423
APOIV						
ApoAIV1	0·884	0·938	0·918	0·926	0·943	0·939
ApoAIV2	0·105	0·062	0·076	0·074	0·057	0·061
ApoAIV3	0·012	0·000	0·006	0·000	0·000	0·000
n	43	121	165	34	123	157
PGM1						
PGM1+	0·621	0·575	0·589	0·601	0·553	0·567
PGM1−	0·197	0·216	0·211	0·201	0·229	0·221
PGM2+	0·115	0·137	0·130	0·132	0·150	0·144
PGM2−	0·067	0·071	0·070	0·066	0·068	0·067
n	157	365	522	129	301	430

Kell-Kp and HB loci were monomorphic for Kpb and Hb-a allele, respectively.

Table 2. Chi-square test for estimating Hardy-Weinberg equilibrium for the Mexican-Americans of Starr County, Texas by sex and birthplace.

	Texas			Mexico			Total		
System	Males	Females	Total	Males	Females	Total	Males	Females	Total
ABO									
χ^2	2·99	2·32	2·90	2·99	1·57	4·06	1·83	0·18	0·19
n	156	362	518	129	302	431	298	691	989
Rh									
χ^2	2·39	3·94	1·95	6·74	6·07	8·57	2·13	4·65	3·25
n	156	362	518	128	302	430	297	691	988
MNSs									
χ^2	6·57	2·15	2·99	1·58	0·70	0·54	4·35	1·52	2·54
n	156	360	516	128	301	429	297	687	984
Duffy									
χ^2	3·66	1·99	0·00	0·92	3·35	1·25	0·89	0·04	0·11
n	151	353	504	125	294	419	288	674	962
Kell									
χ^2	0·01	10·36†	8·90†	0·03	0·00	0·02	0·05	12·42†	6·44†
n	156	362	518	129	302	431	298	691	989
Kidd									
χ^2	0·00	0·07	0·04	0·09	0·00	0·04	0·21	0·12	0·00
n	155	359	514	129	302	431	297	688	985
AK									
χ^2	0·25	0·21	0·43	0·07	0·17	0·24	0·32	0·46	0·77
n	156	363	519	129	302	431	298	692	990
ADA									
χ^2	0·17	0·99	0·38	0·21	0·47	1·76	0·36	0·05	0·02
n	156	364	520	129	302	451	298	693	991
ESTD									
χ^2	0·59	0·16	0·00	0·22	0·09	0·04	0·01	0·46	0·25
n	157	363	520	128	302	430	298	692	990
ACP									
χ^2	3·25	3·57	0·61	1·95	2·03	2·12	3·69	0·41	2·84
n	157	363	520	129	302	431	299	692	991
GPT									
χ^2	1·29	0·39	0·00	0·01	0·93	0·57	0·03	0·03	0·02
n	153	348	501	125	290	415	291	665	956
GLO									
χ^2	0·19	0·72	0·91	1·36	0·00	0·48	1·71	0·36	1·48
n	157	364	521	129	302	431	299	692	991
PGD									
χ^2	0·11	0·10	0·20	0·01	0·12	0·12	0·09	0·23	0·32
n	155	362	517	128	301	429	296	690	986
PGP									
χ^2	0·38	0·66	1·21	0·21	0·15	0·83	0·58	2·16	2·89
n	156	365	522	129	302	431	299	694	993
Hp									
χ^2	0·31	0·24	0·02	0·05	4·76†	4·27†	0·00	2·55	2·02
n	156	365	522	129	299	428	299	691	990
Gc									
χ^2	0·38	3·35	3·05	0·12	1·20	0·53	0·40	2·88	2·28
n	153	365	518	129	300	429	295	692	987
APOE									
χ^2	1·31	1·53	0·21	0·44	1·76	0·33	1·32	1·66	0·35
n	151	353	504	125	298	423	287	676	963
APOIV									
χ^2	0·74	0·53	6·13	0·21	0·45	0·65	0·88	0·95	7·59†
n	43	121	165	34	123	157	78	250	329
PGM1									
χ^2	14·12†	10·72	18·74†	2·75	10·43	11·48	9·07	12·66†	20·97†
n	157	365	522	129	301	430	299	692	991

d.f.: ABO = 2; Rh = 10; MNSs = 5; ACP, PGP, Gc, APOE, APOIV = 3, Duffy, Kell, Kidd, AK, ADA, ESTD, GPT, GLO, PGD, Hp = 1; PGM1 = 6

† χ^2 significant at $p < 0·05$.

Table 3. Standard genetic distances, average heterozygosity and Chi-square values among Mexican-Americans of Starr County by sex and birthplace†

	Males		Females		Total	
	Texas	Mexico	Texas	Mexico	Texas	Mexico
Males						
Texas	33·43 ±5·50	0·32	0·20	0·22		
Mexico	3·54 ±1·62	33·86 ±5·49	0·27	0·24		
Females						
Texas	1·67 ±0·63	2·18 ±1·00	34·34 ±5·53	0·16		
Mexico	2·18 ±0·81	2·20 ±0·77	1·35 ±0·49	33·90 ±5·59		
Total						
Texas					34·11 ±5·52	0·14
Mexico					1·13 ±0·51	34·14 ±5·52

† Figures on the diagonal are the average heterozygosities expressed in percentage (e.g. Texas males = 33·43%); below the diagonal are standard genetic distances in 10^{-3} codon differences per locus and above the diagonal are χ^2 values (polymorphic loci) with d.f. = 35, $p > 0{\cdot}05$.

All computations are based on 18 polymorphic loci (ABO, Rh, MNSs, Duffy, Kell, Kidd, AK, ADA, ESTD, ACP, GPT, GLO, PGD, PGP, Hp, Gc, APOE, and PGM1) and two monomorphic loci (Kp and Hb).

reflect the actual level of genetic variation generally found in human populations. The genetic distances show no significant differentiation, as examined pairwise by the chi-square statistic (Nei and Roychoudhury 1974).

3.3. *Gene diversity*

The total average gene diversity (H_T) of 34·1% (including the two monomorphic loci) and 37·7% (over the polymorphic loci only) mainly (over 99%) can be attributed to individual variation within the population. Only a small contribution to total variability (0·71%) comes from the between-birthplaces level of subdivision. The sex difference is even smaller (0·21% of the total).

3.4. *Genetic admixture*

Table 5 presents the estimated values of admixture based on 17 polymorphic genetic loci, fitting a trihybrid model using the ancestral frequencies shown in the appendix. There is little difference among the Mexican-American subgroups. The Spanish contribution varies from 55·9% for Texas males to 66·2% for Texas females, that from Amerindians varies from 27·6% in Texas females to 34·2% in Mexico females, and the African contribution varies from 5·9% in the Mexico total to 11·7% in Texas males. The Mexican-Americans of Starr County, Texas appear to be hybrid populations with Spanish, Amerindian and West African admixture, with a predominantly Spanish contribution followed by Amerindian and a small West African contribution.

Although the standard errors of these estimates are provided, no rigorous test of homogeneity is possible because of the unknown sampling distribution of the estimated admixture proportions to determine whether the gene-flow from outside is homogeneous. The procedure of Harpending and Ward (1982) was therefore applied. The genetic distance (r_{ii}) of the ith subpopulation from a hypothetical centroid of all subpopul-

Table 4. Gene diversity analysis of allele frequency data from Mexican-Americans in Starr County, Texas.

	G_{ST}†			H_T‡
Locus	Within population	Between birthplaces	Between sex within birthplaces	Total gene diversity
ABO	98·50	1·13	0·37	0·418
Rh	99·36	0·59	0·05	0·696
MNSs	98·31	1·40	0·29	0·699
Duffy	99·15	0·84	0·01	0·550
Kell	99·26	0·62	0·11	0·033
Kidd	99·14	0·81	0·05	0·497
AK	98·78	1·15	0·07	0·074
ADA	98·76	1·24	0·00	0·046
ESTD	99·27	0·65	0·08	0·239
ACP	98·67	1·13	0·20	0·395
GPT	99·46	0·15	0·38	0·500
GLO	99·98	0·02	0·00	0·474
PGD	99·49	0·40	0·10	0·030
PGP	96·07	1·88	2·04	0·211
Hp	99·65	0·18	0·17	0·485
Gc	99·04	0·82	0·13	0·592
APOE	99·23	0·65	0·12	0·261
PGM1	99·73	0·15	0·12	0·592
Mean§	99·08	0·71	0·21	0·377
s.e.	±0·17	±0·13	±0·07	±0·054

† Expressed as percentage of total.
‡ Absolute total gene diversity in the entire sample.
§ Excluding monomorphic loci (Kp and Hb), the absolute total gene (diversity per locus (H_T) including all 20 loci is 0·340±0·055.

Table 5. Percentage contribution from Spanish, Amerindian and West African gene pools to the contemporary Mexican-Americans of Starr County by sex and birthplace.

	Spanish	Amerindian	West African
Males			
Texas	55·95±2·74	32·33±2·36	11·72±1·24
Mexico	64·08±1·42	29·33±1·23	6·59±0·65
Total	58·62±2·47	31·69±2·13	9·68±1·12
Females			
Texas	66·25±0·21	27·57±0·19	6·17±0·10
Mexico	59·30±1·75	34·19±1·51	6·51±0·79
Total	63·77±0·66	30·16±0·57	6·08±0·30
Total			
Texas	63·25±0·94	28·88±0·81	7·88±0·43
Mexico	62·71±1·04	31·37±0·90	5·92±0·47
Total	62·30±1·19	30·55±1·03	7·16±1·00

The computations are done with 17 polymorphic loci (ABO, Rh, MNSs, Duffy, Kell, Kidd, AK, ADA, ESTD, ACP, GPT, GLO, PGD, PGP, Hp, Gc and PGM1). APOE locus data are not used for admixture estimation because allele frequencies at this locus are not available for the appropriate ancestral populations.

ations is related to the average heterozygosity (H_i) of the ith subpopulation. If gene-flow from outside is uniform, then $H_i = b(1 - r_{ii})$, with absolute value of b being equal to $\bar{H}$, the average heterozygosity in the pooled population. Using 18 genetic loci (excluding the two monomorphic Hb and Kp loci, and the APOA IV locus for which a large number of individuals were not typed), analysis of the regression (table 6) of heterozygosity on genetic distance shows it to be consistent with linearity. $\bar{H}$ in the pooled populations (34·06%) does not differ significantly from the regression coefficient (33·97%). The various subpopulations of the Mexican-American population of Starr County are therefore similar in the proportions of the genes they have received from the ancestral populations, which is consistent with the similarity of admixture proportions estimated in the previous section.

Table 6. Average heterozygosity (H_i) and genetic distances from a centroid (r_{ii}) among the Mexican-Americans of Starr County based on 18 polymorphic loci.

Population	$r_{ii} \pm$ SE	$H_i \pm$ SE
Males		
Texas	0·0038 ± 0·0010	0·3343 ± 0·0550
Mexico	0·0048 ± 0·0015	0·3386 ± 0·0549
Females		
Texas	0·0012 ± 0·0003	0·3432 ± 0·0553
Mexico	0·0016 ± 0·0004	0·3390 ± 0·0559

Regression analysis: $H_i = b(1 - r_{ii})$; H_i plotted against $1 - r_{ii}$ through the origins has $t = -0·901$; d.f. = 2, $p > 0·05$.

Regression coefficient through origin (b) = 0·3397 ± 0·0033.

Average heterozygosity in pooled population ($\bar{H}$) = 0·3406 ± 0·0554.

3.5. *Non-random association among genetic loci*

It is well known that the mixture of populations with disparate allele frequencies can produce non-random association of alleles at two or more unlinked loci (Li 1955, Nei and Li 1973; Chakraborty and Weiss 1988). Employing the procedure suggested by Brown *et al.* (1980), from the available genotype data on each individual, we defined a multi-locus genotype for each individual excluding the monomorphic Hb and Kp loci and the APOA IV genotype, for which too few data were available. With respect to the remaining 18 loci, the number of loci was determined for which each of 862 individuals was heterozygous. Comparing the observed distribution with that expected, under the assumption of random association of alleles at different loci using Chakraborty's (1981) algorithm we find (table 7) that the observed distribution is in fair agreement with the observed one (goodness-of-fit χ^2 with 9 d.f. = 19·96, $p > 0·01$). Since the expected distribution involves the observed data at least partially (locus-specific observed heterozygosity values), there are some technical difficulties for determining the degrees of freedom of the above goodness-of-fit statistic, detailed in Chakraborty (1984). However, Brown *et al.* (1980) showed that the expectations of mean and variance of the number of heterozygous loci can be written as functions of locus-specific heterozygosities under the assumption of random association of alleles, and the 95% confidence limit of the observed variance of the number of heterozygous loci can also

Table 7. Observed and expected distribution of the number of heterozygous loci in the Mexican-Americans of Starr County, Texas.

Number of heterozygous loci	Number of individuals	
	Observed	Expected
0–2	6	5·72
3	27	20·73
4	81	58·15
5	125	115·96
6	180	169·61
7	167	185·10
8	137	151·80
9	86	93·48
10	31	42·89
11–18	22	18·56
Total	862	862·00
Mean	6·630	6·836
Variance	3·540	3·331

Goodness-of-fit $\chi^2 = 19{\cdot}96$, d.f. $= 9$, $p > 0{\cdot}01$.
95% Confidence interval for variance (3·024, 3·638).

be calculated. In the Mexican-American data the mean number of heterozygous loci was 6·63 and the variance was 3·54. Their expected values (under random association) are 6·84 and 3·33, respectively. The 95% confidence limits of the variance are 3·02–3·64. Clearly, these suggest no evidence of non-random association of alleles among the 18 polymorphic loci in this population.

4. Discussion and conclusion

The results of genetic distance analysis between various subpopulations of Mexican-Americans indicate that they are similar to each other. Overall, the level of gene diversity (G_{ST}) is small and more than 99% of total gene diversity (H_T) is accounted for by individual variation within the population.

The admixture results are largely consistent with reports for other Mexican-American groups in the United States. Reed (1974), using the Rh blood group system, estimated 32·0 ± 5·6% Amerindian ancestry among the Mexican-Americans of California. Gottlieb and Kimberling's (1979) findings, from a small admixture study in Colorado, showed 60·0% Spanish contribution to the population, indicating a somewhat larger Amerindian component than seen in other studies. Population admixture estimates computed for the Mexican-Americans of San Antonio, Texas, based on skin reflectance (Relethford, Stern, Gaskill and Hazuda 1983), showed the Amerindian contribution to be 46·0%, 27·0% and 18·0%, respectively, among the Barrio, Transition and Suburban Mexican-American neighbourhoods. Based on gene frequency data on 18 genetic loci Chakraborty, Ferrell, Stern, Haffner, Hazuda and Rosenthal (1986) estimated 43·8%, 30·0% and 18·7% Amerindian ancestry, respectively, in the same three social classes.

The results obtained from the trihybrid model in the present investigation indicated around 62·0%, 30·0% and 8·0% contribution from the Spanish, Amerindian and West African gene pools, respectively. Although various studies show some regional as well as social class variation regarding the contribution of Amerindians to the Mexican-Americans of the United States, there was no remarkable heterogeneity in

genetic admixture among subgroups of the Starr County population. Our estimates of admixture are similar to those obtained by Garza-Chapa (1983), Cerda-Flores, Ramirez-Fernandez and Garza-Chapa (1987), Cerda-Flores and Garza-Chapa (1989) and Cerda-Flores, Kshatriya, Barton, Leal-Garza, Garza-Chapa, Schull and Chakraborty (1991) for the Mexicans of Nuevo León in northeastern Mexico. The chi-square statistic for the 21 genetic marker systems to fit the Hardy-Weinberg equilibrium indicates that intermixing is old enough to have eliminated any early non-random association of genes.

In summary, on the basis of the genetic data presented, we conclude that the Mexican-Americans of Starr County, Texas, classified by sex and birthplace, are not genetically distinguishable. The findings on genetic admixture indicate a predominant influence from the Spanish, and lesser contributions from Amerindians and West Africans. The history of admixture is apparently old enough to have brought the entire Mexican-American gene pool to Hardy-Weinberg equilibrium. The multi-locus heterozygosity distribution also supports the inference of genetic homogeneity.

These findings have a number of important implications with respect to the utility of such populations in anthropogenetic and epidemiological contexts. First, the demonstration of genetic homogeneity of the Mexican-Americans of Starr County, Texas, in spite of their admixed origin, suggests that this population is suitable for studying disease-marker associations in the search of candidate genes of complex diseases; such an association cannot possibly arise from population mixture alone (Chakraborty and Weiss 1988). Secondly, in spite of the polyphyletic origin of the Mexican-Americans, their multi-locus genotypic distribution satisfies the premises of random segregation of unlinked loci. Therefore, the probability of finding a specific multiple-locus genotype in such a population can be determined by the product rule of locus-specific genotype probabilities, contrary to the recent claim of Cohen (1990). The 862 individuals on which we had the 18-locus genotype data available constitute 862 different multiple-locus genotypes; i.e. no repeat of any multiple-locus genotype was observed in the sample. Based on the allelic frequencies the most probable multiple-locus genotype in this population would be encountered once in every 885,764 individuals. This shows that the discretized genotypic information in such a population is sufficient to determine the identity of individuals even when the population is of admixed origin.

Acknowledgements

This work was supported in part by the US Public Health Service Research Grants DK 34666, DK01748 and GM 41399 from the National Institutes of Health. We thank an anonymous reviewer for extensive editorial and other constructive suggestions on this paper.

References

Brown, A. H. D., Feldman, M. W., and Nevo, E., 1980, Multilocus structure of natural populations of Hordeum spontaneum. *Genetics*, **96**, 523–536.

Cerda-Flores, R. M., and Garza-Chapa, R., 1989, Variation in the gene frequencies of three generations of humans from Monterrey, Nuevo León, Mexico. *Human Biology*, **61**, 249–261.

Cerda-Flores, R. M., Ramirez-Fernandez, E., and Garza-Chapa, R., 1987, Genetic admixture and distances between populations from Monterrey, Nuevo León, Mexico and their putative ancestral populations. *Human Biology*, **59**, 31–49.

Cerda-Flores, R. M., Kshatriya, G. K., Barton, S. A., Leal-Garza, C. H., Garza-Chapa, R., Schull, W. J., and Chakraborty, R., 1991, Genetic structure of the immigrant populations of San Luis Potosi and Zacatecas to Nuevo León in Mexico. *Human Biology*, **63**, 309–327.

CHAKRABORTY, R., 1981, The distribution of the number of heterozygous loci in natural populations. *Genetics*, **98**, 461–466.

CHAKRABORTY, R., 1984, Detection of non-random association of alleles from the distribution of the number of heterozygous loci in a sample. *Genetics*, **108**, 719–731.

CHAKRABORTY, R., 1985, Gene identity in racial hybrids and estimation of admixture rates. In *Genetic Microdifferentiation-Human and Other Populations*, edited by Y. R. Ahuja and J. V. Neel (New Delhi: Indian Anthropological Association), pp. 171–180.

CHAKRABORTY, R., 1986, Gene admixture in human populations: Models and predications. *Yearbook of Physical Anthropology*, **29**, 1–43.

CHAKRABORTY, R., and WEISS, K. M., 1988, Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the Academy of Sciences, USA*, **85**, 9119–9123.

CHAKRABORTY, R., HAAG, M., RYMAN, N., and STÅHL, G., 1982, Hierarchical gene diversity analysis and its implication to brown trout population data. *Hereditas*, **97**, 17–21.

CHAKRABORTY, R., FERRELL, R. E., STERN, M. P., HAFFNER, S. M., HAZUDA, H. P., and ROSENTHAL, M., 1986, Relationship of prevalence of non-insulin dependent diabetes mellitus with Amerindian admixture in Mexican-Americans of San Antonio, Texas, *Genetic Epidemiology*, **3**, 435–454.

COHEN, J. E., 1990, DNA fingerprinting for forensic identification: Potential effects on data interpretation of subpopulation heterogeneity and band number variability. *American Journal of Human Genetics*, **46**, 358–368.

COSSIO, D. A., 1925, *Historia de Nuevo León, Evolucion Politica y Social*, edited by S. Cantu Leal (Mexico: Monterrey, Nuevo León).

CRAWFORD, M. H., LEYSHON, W. C., BROWN, K., LESS, F., and TAYLOR, L., 1974, Human biology in Mexico. II. A comparison of blood group, serum and red cell enzyme frequencies and genetic distances of the Indian populations of Mexico. *American Journal of Physical Anthropology*, **41**, 251–268.

DEL HOYO, H., 1979, *Historia del Nuevo Reino de León (1577–1723)* (Mexico: Editorial Libros de Mexico, S. A.), 2nd edn.

FERREL, R. E., BERTIN, T., YOUNG, R., BARTON, S. A., MURILLO, F., and SCHULL, W. J., 1978, The Aymará of Western Bolivia. IV. Gene frequencies of eight blood groups and 19 protein and erythrocyte enzyme systems. *American Journal of Human Genetics*, **30**, 539–549.

FERRELL, R. E., CHAKRABORTY, R., GERSHOWITZ, H., LAUGHLIN, W. S., and SCHULL, W. J., 1981, The St. Lawrence Island Eskimos. Genetic variation and genetic distance. *American Journal of Physical Anthropology*, **55**, 351–358.

GARZA-CHAPA, R., 1983, Genetic distances for ABO and Rh(D) blood groups in the State of Nuevo León, Mexico. *Social Biology*, **30**, 24–31.

GOTTLIEB, K., and KIMBERLING, W. J., 1979, Admixture estimates for the gene pool of Mexican-Americans in Colorado. *Abstracts, Forty-eighth Annual Meeting of the American Association of Physical Anthropologists*, San Francisco, California, p. 444.

HANIS, C. L., FERRELL, R. E., BARTON, S. A., AGUILAR, L., GARZA-IBARRA, A., TULLOCH, B. R., GARCIA, C. A., and SCHULL, W. J., 1983, Diabetes among Mexican Americans in Starr County, Texas. *American Journal of Epidemiology*, **118**, 659–672.

HANIS, C. L., FERRELL, R. E., TULLOCH, B. R., and SCHULL, W. J., 1985, Gallbladder disease epidemiology in Mexican Americans in Starr County. *American Journal of Epidemiology*, **122**, 820–829.

HANIS, C. L., CHAKRABORTY, R., FERRELL, R. E., and SCHULL, W. J., 1986, Individual admixture estimates: Disease associations and individual risk of diabetes and gallbladder disease among Mexican-Americans in Starr County, Texas. *American Journal of Physical Anthropology*, **70**, 433–441.

HANIS, C. L., HEWETT-EMMETT, D., KUBRUSLY, L. F., MAKLAD, M. N., DOUGLAS, T. C., MUELLER, W. H., BARTON, S. A., YOSHIMARU, H., KUBRUSLY, D. B., GONZALEZ, R., and SCHULL, W. J., 1991, An ultrasound survey of gallbladder disease among Mexican-Americans in Starr County, Texas: Associations with obesity, diabetes, hypertension, lipids, lipoproteins and apolipoproteins. (Submitted).

HANIS, C. L., DOUGLAS, T. C., and HEWETT-EMMETT, D., 1991a, Apolipoprotein A-IV protein polymorphism: Frequency and effects of lipids, lipoproteins and apolipoproteins among Mexican-Americans in Starr County, Texas. *Human Genetics*, **86**, 323–325.

HANIS, C. L., HEWETT-EMMETT, D., BERTIN, T. K., and SCHULL, W. J., 1991b, The origins of U.S. Hispanics: Implications for diabetes. *Diabetes Care*, **14**, 618–627.

HARPENDING, H. C., and WARD, R. H., 1982, Chemical systematics and human evolution. In *Biochemical Aspects of Evolutionary Biology*, edited by M. Nitecki (Chicago: University of Chicago Press), pp. 213–256.

HERNANDEZ-GARZA, T. L., 1973, *Breve Historia de Nuevo León* (Mexico: Editorial Trillas), 3rd edn.

ITAKURA, K., MATSUDATE, T., SAKURAI, T., HASHIMOTO, S., ITO, K., KANNO, H., HIRATA, M., and NAKAMURA, K., 1986, Single radial immunodiffusion of serum apolipoproteins C-II, C-III and E—pretreatment of samples with surfactant. *Clinica et Chimica Acta*, **161**, 275–282.

LI, C. C., 1955, *Population Genetics* (Chicago: University of Chicago Press).
MONTEMAYOR-HERNANDEZ, A., 1971, *Historia de Monterrey* (Mexico: Asociacion de Editores de Monterrey, AC).
MOURANT, A. E., KOPEC, A. C., and DOMANIEWSKA-SOBCZEK, K., 1976, *The Distribution of the Human Blood Groups and Other Polymorphisms* (London: Oxford University Press).
NEI, M., 1972, Genetic distance between populations. *American Naturalist*, **106**, 283–292.
NEI, M., and LI, W. H., 1973, Linkage disequilibrium in subdivided populations. *Genetics*, **75**, 213–219.
NEI, M., and ROYCHOUDHURY, A. K., 1974, Sampling variances of heterozygosity and genetic distance. *Genetics*, **76**, 379–390.
REED, T. E., 1974, Ethnic classification of Mexican-Americans. *Science*, **185**, 283.
REED, T. E., and SCHULL, W. J., 1968, A general maximum likelihood estimation program. *American Journal of Human Genetics*, **20**, 579–580.
RELETHFORD, J. H., STERN, M. P., GASKILL, S. P., and HAZUDA, H. P., 1983, Social class, admixture, and skin color variation in Mexican-Americans and Anglo-Americans living in San Antonio, Texas. *American Journal of Physical Anthropology*, **61**, 97–102.
SNEDECOR, G. W., and COCHRAN, W. G., 1967, *Statistical Methods* (Ames, IO: Iowa State University Press), 6th edn.
US BUREAU OF THE CENSUS, 1980 (1982), *General Population Characteristics: Texas. Census of Population*, Vol. 1, Part 45, PC80-1-B45 (Washington, DC: US Government Printing Office).
WEISS, K. M., FERRELL, R. E., and HANIS, C. L., 1984, A new world syndrome of metabolic disease with a genetic and evolutionary basis. *Yearbook of Physical Anthropology*, **27**, 153–178.

Zusammenfassung. Die nach Geschlecht und Geburtsort klassifizierten Mexican-Americans von Starr County, Texas, wurden im Hinblick auf das Ausmaß ihrer genetischen Differenzierung und der Beiträge von ancestralen Populationen, wie Spanier, amerikanischer Indianer und Westafrikaner untersucht. Unter Berücksichtigung von 21 genetischen Markern zeigten genetische Abstandsanalysen und Gene Diversity-Analysen, daß Subpopulationen der Mexican-Americans in Starr County einander genetisch ähnlich sind und daß mehr als 99% der insgesamt beobachteten Gene Diversity (H_T) auf individuelle Variationen in der Population zurückgeführt werden können. Eine Analyse der genetischen Durchmischung belegt einen prädominanten Einfluß der Spanier, einen geringeren Beitrag der amerikanischen Indianer und einen geringfügigen der Westafrikaner. Der genetische Beitrag der ancestralen Populationen zu den verschiedenen Subpopulationen der Mexican-Americans in Starr County ist ähnlich. Die Mexican-Americans von Starr County sind den mexikanischen Populationen aus dem Nordosten Mexicos ähnlich. Der Prozeß der Durchmischung reicht historisch offensichtlich soweit zurück, daß die Zeit ausreichend war, um den gesamten Genpool der Mexican-Americans ins Hardy-Weinberg-Gleichgewicht zu bringen. Obwohl diese Population per Durchmischung enstanden ist, gibt es keine nichtzufälligen Assoziationen der Allele zwischen den genetischen Markersystemen, die in dieser Studie berücksichtigt wurden. Insgesamt belegen die Ergebnisse die genetische Homogenität der Mexican-Americans von Starr County, Texas, und weisen auf die Bedeutung dieser Population gür genetische und epidemiologische Studien hin.

Résumé. Les mexicains-américains du comté de Starr au Texas, répertoriés par sexe et lieu de naissance, ont été étudiés afin de déterminer l'étendue de leur variation génétique ainsi que des contributions qu'ils ont recues de populations ancestrales telles qu'espagnole, amérindienne et ouest-africaine. Les distances génétiques et les analyses de diversité effectuées à partir de 21 marqueurs génétiques, indiquent que les sous-groupes de mexicains-américains du comté de Starr sont similaires et que plus de 99% de la diversité génétique totale (H_T) peuvent être attribués à des variations individuelles dans la population. Les analyses d'apports génétique montrent que l'influence dominante provient des espagnols, à un moindre degré des amérindiens et dans une faible mesure des ouest-africains. La contribution de la population ancestrale aux diverses sous-populations est similaire. Les mexicains-américains du comté de Starr sont semblables aux populations mexicaines du nord-est du Mexique. L'histoire du métissage est apparemment assez ancienne pour avoir permis l'atteinte de l'équilibre d'Hardy-Weinberg, par l'ensemble du patrimoine génétique mexicain-américain. Il n'y a pas d'association non-alléatoire d'allèles parmi les systèmes considérés dans cette étude, en dépit de l'origine composite de cette population. Ces résultats montrent l'homogénéité génétique des mexicains-américains du comté de Starr au Texas et indiquent l'utilité de cette population pour des études épidémiologiques et génétiques.

Appendix. Allele frequencies for 17 genetic loci in Mexican-Americans of Starr County and putative ancestral populations†

	Mexican-American					
Allele	Males	Females	Total	Spanish	Amerindian	West African
A	0·174	0·184	0·181	0·310	0·063	0·191
B	0·077	0·073	0·074	0·067	0·003	0·213
O	0·749	0·743	0·745	0·623	0·934	0·596
DCE	0·011	0·020	0·017	0·048	0·022	0·002
DCe	0·429	0·443	0·438	0·418	0·626	0·028
DcE	0·191	0·166	0·174	0·090	0·330	0·084
Dce	0·068	0·064	0·065	0·049	0·000	0·637
dCE	0·000	0·000	0·000	0·002	0·000	0·000
dCe	0·000	0·000	0·000	0·011	0·000	0·007
dcE	0·000	0·004	0·003	0·001	0·000	0·000
dce	0·302	0·303	0·303	0·380	0·022	0·242
MS	0·287	0·299	0·296	0·243	0·346	0·088
Ms	0·365	0·360	0·361	0·311	0·444	0·432
NS	0·067	0·080	0·075	0·057	0·080	0·138
Ns	0·282	0·262	0·268	0·389	0·130	0·342
Fy (a)	0·436	0·434	0·435	0·365	0·820	0·002
Fy (b)	0·484	0·482	0·483	0·635	0·180	0·000
Fy	0·080	0·084	0·083	0·000	0·000	0·998
K	0·013	0·010	0·011	0·038	0·000	0·003
k	0·987	0·990	0·989	0·962	1·000	0·997
Jka	0·468	0·480	0·476	0·537	0·360	0·433
Jkb	0·532	0·520	0·524	0·463	0·640	0·567
AK1	0·968	0·975	0·973	0·978	1·000	0·993
AK2	0·032	0·025	0·027	0·022	0·000	0·007
ADA1	0·966	0·966	0·966	0·951	1·000	1·000
ADA2	0·034	0·034	0·034	0·049	0·000	0·000
ESD1	0·867	0·868	0·868	0·879	0·716	0·938
ESD2	0·133	0·132	0·132	0·121	0·284	0·062
ACPa	0·254	0·265	0·262	0·298	0·203	0·180
ACPb	0·721	0·725	0·724	0·666	0·797	0·820
ACPc	0·025	0·009	0·014	0·036	0·000	0·000
GPT1	0·514	0·456	0·473	0·505	0·442	0·863
GPT2	0·486	0·544	0·527	0·495	0·558	0·137
GLO1	0·390	0·391	0·391	0·423	0·211	0·309
GLO2	0·610	0·609	0·609	0·577	0·789	0·691
PGDA	0·983	0·982	0·982	0·975	0·997	0·937
PGDC	0·017	0·018	0·018	0·025	0·003	0·063
PGP1	0·908	0·868	0·880	0·927	0·690	1·000
PGP2	0·082	0·126	0·113	0·047	0·310	0·000
PGP3	0·010	0·006	0·007	0·026	0·000	0·000
Hp1	0·396	0·448	0·432	0·429	0·452	0·654
Hp2	0·604	0·552	0·568	0·571	0·548	0·346
Gc1	0·725	0·720	0·721	0·641	0·859	0·916
Gc2	0·275	0·280	0·279	0·359	0·141	0·084
PGM1+	0·610	0·566	0·579	0·621	0·485	0·787
PGM1−	0·199	0·223	0·215	0·114	0·342	0·039
PGM2+	0·124	0·144	0·138	0·211	0·020	0·147
PGM2−	0·067	0·068	0·068	0·054	0·153	0·027

† The allele frequencies for the Mexican-Americans are from the present survey (summing over all individuals) and those for the ancestral populations are compiled from the literature (see Hanis *et al.* 1991b for exact sources from Mourant *et al.*'s 1976 compilation).

Anthrop. Anz.	Jg. 48	4	313 - 331	Stuttgart, Dezember 1990

Genetic profile of cosmopolitan populations: Effects of hidden subdivision

R. Chakraborty

Center for Demographic and Population Genetics, The University of Texas, Graduate School of Biomedical Sciences, Houston, TX, USA

With 2 figures and 10 tables in the text

Summary: Natural populations of many organisms exhibit excess of rare alleles in comparison with the predictions of the neutral mutation hypothesis. It has been shown before that either a population bottleneck or the presence of slightly deleterious mutations can explain this phenomenon. A third explanation is presented in this work, showing that hidden subdivision within a population can also lead to an excess of rare alleles in the total population when the expectations of the neutral model are based on the allele frequency profile of the entire population data.

With two examples (mitrochondrial DNA-morph distribution and isozyme allele frequency distributions), it is shown that most cosmopolitan human populations exhibit excess of rare as well as total allele counts, when these are compared with the expectations of the neutral mutation hypothesis. The mitochondrial data demonstrate that such excesses can be detected from genetic variation at a single locus as well, and this is not due to stochastic error of allele frequency distributions. Contrast of the present observations with the allele frequency profiles in agglomerated tribal populations from South and Central America shows that even when the neutral expectations hold for individual subpopulations, if all subpopulations are grouped into a single population, the pooled data exhibit an excess of total number of alleles that is mainly due to the excess of rare alleles. Therefore, a primary cause of the excess number of rare alleles could be the hidden subdivision, and the magnitude of the excess indicates the extent of substructuring. The two components of hidden subdivision are: 1) Number of subpopulations, and 2) the average genetic distance among them. The implications of this observation in estimating mutation rate are discussed indicating the difficulties of comparing mutation rates form different population surveys.

Zusammenfassung: Natürliche Populationen zahlreicher Organismen weisen einen Überschuß an seltenen Allelen auf, der nicht mit der Hypothese neutraler Mutationen in Übereinstimmung steht. Es wurde zur Erklärung dieses Phänomens bisher angenommen, daß entweder „Bottleneck-Effekte" oder schwach nachteilige Mutationen in diesem Zusammenhang eine Rolle spielen. In dieser Untersuchung wird eine dritte Erklärungsmöglichkeit aufgezeigt, indem dargestellt wird, daß eine nicht offen erkennbare Strukturierung einer Population zu einem Überschuß an seltenen Allelen in der Gesamtpopulation führen kann, und zwar dann, wenn die Erwartungen nach dem Modell neutraler Mutationen auf dem Allelenfrequenzprofil für die Gesamtpopulation basieren.

An zwei Beispielen (mitochondriale DNA-morph-Verteilung und Verteilung der Allelenfrequenzen von Isoenzymen) wird gezeigt, daß die meisten menschlichen Großbevölkerungen Überschüsse sowohl seltener Allele als auch der Allelzahlen insgesamt zeigen, wenn diese mit den Erwartungswerten nach der Hypothese neutraler Mutationen verglichen werden. So lassen

0003-5548/90/0048-0313 $ 4.75

die Daten für die mitochondriale DNA erkennen, daß ein solcher Überschuß ebenso anhand der genetischen Variabilität an einem einzigen Genlocus entdeckt werden kann, was nicht durch stochastische Fehler der Allelenfrequenzverteilungen bedingt ist. Beobachtungen an zusammengesetzten süd- und mittelamerikanischen Stammesbevölkerungen zeigen, daß die Hypothese neutraler Mutationen für die einzelnen Stammesbevölkerungen durchaus zutreffen kann. Wenn jedoch alle Subpopulationen zu einer Gesaamtpopulation zusammengefaßt werden, lassen die gepoolten Frequenzdaten einen Überschuß bezüglich der Gesamtallelenzahl erkennen, welcher hauptsächlich durch einen Überschuß an seltenen Allelen bedingt ist. Eine wesentliche Ursache hierfür ist offenbar in einer nicht offen erkennbaren Strukturierung der Gesamtbevölkerung zu sehen, und das Ausmaß des Überschusses reflektiert den Grad dieser Strukturierung. Die beiden Komponenten der verborgenen Bevölkerungsgliederung sind 1. die Zahl der Subpopulationen und 2. der durchschnittliche genetische Abstand zwischen ihnen. Die Bedeutung dieser Beobachtung für die Schätzung von Mutationsraten wird diskutiert, wobei auch auf die Schwierigkeiten hinsichtlich des Vergleichs von Mutationsraten eingegangen wird, die an verschiedenen Populationen ermittelt worden sind.

Introduction

In genetic analysis of population data, the genetic make-up of a population is generally studied in a variety of ways. The basic data for such analyses are frequencies of various alleles (or genotypes) at one or more loci, estimated from random samples drawn from a population. The population, in this context, is usually defined as a breeding unit, within which mating occurs with a well-specified pattern (generally assumed to be random).

Over the history of population genetics, the technique of detecting genetic variation has changed considerably. Initially, before the advent of serological techniques of blood grouping (Landsteiner & Levine 1928), morphological and physiological traits had been popular for studying genetic variations within and between populations. Soon after the discovery of blood group systems in humans, anthropological and human genetic studies started using these techniques extensively. As a result, morphological data had been now replaced by voluminous gene frequency surveys in various ethnic groups around the world (Mourant et al. 1976). The development of electrophoretic techniques introduced another set of traits which detect genetic variations based on charge and/or molecular size differences of protein-enzyme molecules. Since such changes are due to new mutations at the nucleotide level that are translated into mRNA during protein synthesis, the electrophoretically determined genetic variations were the first step of studying evolution at a molecular level. Genetic variation detected by electrophoresis, furthermore, does not depend upon the antigen-antibody specificity which is essential for the serological methods applied to detect the genetic variation at blood groups and immunological systems. Use of electrophoretic techniques, therefore, produced data not only in humans, but also on other organisms, ranging from *E. coli* (Milkman 1973) to primates (King & Wilson 1975, Bruce & Ayala 1979). It soon became apparent that genetic variation is widespread; its extent varies from organism to organism (Nei 1975), and for some organisms the extent of genetic variation vary widely over different geographic regions, depending upon other factors such as population size, reproductive isolation, and ecological conditions. These assertions have been even more firmly established by the recent molecular techniques of restriction fragment length polymorphisms (RFLPs), and nucleotide-sequencing technique, whereby the detection

of genetic variation is extended beyond the translated region of the DNA. Although these later techniques are far more powerful to study genetic variation at a molecular level, it should be noted that current knowledge of DNA polymorphism at a population level is still scanty compared to the electrophoretically determined polymorphisms (see Roychoudhury & Nei 1987 for a comparative compilation of the current data).

While such data convincingly established the ubiquity of genetic variation, there is still a question as to which evolutionary factors play a dominant role in maintaining such variations in natural populations. In other words, it is not certain if the main cause of extensive genetic polymorphism is natural selection, nor it is clear whether or not the genetic variation is being maintained by counteracting forces of mutation and random genetic drift (Lewontin 1974, Ayala 1976, Nei 1975, 1987, Kimura 1983). This controversy, called the selectionist-neutralist controversy, still remains unresolved for the reason that there are several features of data on genetic polymorphisms that cannot be rigorously explained by either of these hypotheses. For example, the observed average level of heterozygosity is generally too low in contrast with the predictions of the balancing selection hypothesis of genetic polymorphism (Nei & Graur 1984). At the same time there is a relative excess of the frequency of "rare" alleles in comparison with the expectation of the neutral mutation hypothesis, which is particularly noteworthy in many species, ranging from Drosophila to human (Ohta 1976, Chakraborty et al. 1980). This later observation led Ohta (1973) to propose that many of the new mutations that occur in nature are deleterious, but they are quickly eliminated from the population because of the negative selection against them. However, there are some "slightly" deleterious mutants which are not so quickly eliminated from the population because of the weakness of selection intensity against them. Negative selection against them, however, prevents their frequencies to attain intermediate or high level. As a result, these slightly deleterious mutations account for the relative excess of "rare" alleles in a population.

An alternative explanation of the excess of rare alleles is given by Nei et al. (1975), Chakraborty & Nei (1977), Maruyama & Fuerst (1984, 1985), and Watterson (1984) who proposed that in nature many populations (or species) are subject to drastic fluctuations of population sizes over time due to ecological and/or environmental changes. When a population goes through a sudden reduction of its size, the genetic variability in the gene pool is substantially reduced, and it takes a long time (of the order of the inverse of mutation rate, in units of generation length) to recover from the loss of genetic variation. On the other hand, the population size may return to the resource capacity of the population/species comparatively much earlier. This phenomenon, called "bottleneck", can easily explain the apparent excess of rare alleles under the premises of the neutral mutation hypothesis.

While both of these explanations are based on mathematical arguments that are difficult to refute, direct validations of "bottleneck" and/or "slightly" deleterious mutations are not available from data of natural populations. This is so because of the lack of past historical data on population sizes for many organisms, and the accurate measurement of selection coefficient for or against any specific allele is a difficult task (Lewontin & Cockerham 1959). In summary, it is not universally accepted that the population "bottleneck" is an wide-spread evolutionary phenomenon applicable for all species/populations in which an excess of rare alleles is observed (Goodnight 1987). In a similar vein, the hypothesis of slightly deleterious mutations has also its own caveat (Li 1978).

The purpose of this presentation is to show that there is another factor which can give rise to the same observation (excess of rare or total number of alleles compared to the expectations under the neutral mutation hypothesis). In an earlier publication (Cahkraborty et al. 1988) it is shown that when there is a hidden subdivision within a population, caused by microdifferentiation within a population, the allele frequency profile in the total population deviates from the neutral expectation in such a fashion that the total number of alleles exhibit an excess which mainly occurs through an excess of rare alleles. This presentation extends the above study demonstrating that hidden subdivision possibly are present in many national (cosmopolitan) human populations which can cause rare alleles to be observed in frequencies higher than the expectations of the mutation-drift model (neutral mutation hypothesis). This is shown first with the mitochondrial DNA (mtDNA) survey data from several oriental populations (Harihara et al. 1988), which implies that the effects of hidden subdivision may be detected even with data from a single locus, provided that it contains enough variability. Secondly it is shown that the summary observations from the three major cosmopolitan populations (Japanese, US Whites, and US Blacks), studied by Neel et al. (1988) and Mohrenweiser et al. (1987), also exhibit the same phenomenon when several isozyme loci are simultaneously considered. This indicates that the observations from the mtDNA data are not artifacts of stochastic errors of single-locus data. A recapitulation of Chakraborty et al.'s (1988) computations is presented to demonstrate that the relative excess of rare alleles present in a population, produced by the phenomenon of amalgamation, is determined by two factors: 1) the number of subpopulations hidden within the population studied; and 2) their genetic dissimilarities (average genetic distances among them). Therefore, the complexity of the population can be examined in terms of the observed level of excess number of alleles encountered in any given survey. Lastly, it is argued that since the excess mainly occurs through the excess of rare alleles, unless the above factor is critically taken into consideration, the indirect estimate of mutation rate per locus per generation may be overestimated from the frequencies of rare alleles, as proposed by Nei (1977), Neel & Rothman (1978), Chakraborty (1981), and others.

Genetic diversity at the mtDNA genome in some Oriental populations

The mitochondrial DNA (mtDNA) is particularly useful in evolutionary studies of ethnic origins of various human populations (e.g., Brown 1980, Denaro et al. 1981, Blanc et al. 1983, Johnson et al. 1983, Horai et al. 1987, Harihara et al. 1988) and in detecting DNA polymorphisms that existed before the geographic dispersal of the human species in the world (Cann et al. 1982, Cann & Wilson 1983, Cann et al. 1987). The mtDNA has a distinct advantage over nuclear DNA because the substantial sequence variation present in the mtDNA that can be detected by the restriction fragment length polymorphisms (RFLPs) is produced unequivocally by new mutations and no recombination is involved in the generation of the mtDNA-morphs that can be defined by using various restriction enzymes (Brown & Goodman 1979, Horai et al. 1986, Horai & Matsunaga 1986). The molecular sequence variation at the mtDNA genome has also provided evidence for founder effects in several human populations (Wallace et al. 1985).

The power of resolution of sequence variability at the mtDNA genome, however, varies substantially from laboratory to laboratory, depending upon the restriction enzymes used and detectability of fragment size differences. In comparing mtDNA-morph frequency differences among various populations, therefore, the uniformity of laboratory methods must be taken into account. For the present purpose, here we consider the data from a recent survey where mtDNA polymorphisms were detected using 13 restriction enzymes on the total DNA obtained from blood samples of five Asian populations; Japanese and Ainu of Northern Japan, Korean, Negrito (Aëta) of the Philippines, and Vedda of Sri Lanka (Harihara et al. 1988). In the total sample of 243 individuals 20 different mtDNA-morphs were detected from the combination of 28 different restriction enzyme morphs. Since the rate of nucleotide substitutions and the extent of nucleotide diversity at the mtDNA genome, by and large, follow the pattern of the predictions of the neutral mutation hypothesis (for a review see Nei 1987), it is interesting to ask whether the mtDNA-morph distributions observed in the survey of Harihara et al. (1988) are in accordance with the sampling theory of neutral mutations (Ewens 1972, Chakraborty & Griffiths 1982).

Estimation

In a survey of n genes from a steady-state population of effective size N_e, the expected number of alleles (morphs) that occur with r copies in the sample is given by

$$E(k_r) = \frac{\theta}{r} \cdot \frac{n!}{(n-r)!} \cdot \frac{\Gamma(n+\theta-r)}{\Gamma(n+\theta)} \tag{1}$$

for $r = 1, 2, \ldots$; where $\theta = 2N_{e(f)}v$, in which v is the mutation rate per generation (for mtDNA $v = m\mu$, where μ is the mutation rate per nucleotide site per generation, and m is the length of the mtDNA genome ≈ 16.5 kb in man) $N_{e(f)}$ is the effective female population size, and $\Gamma(.)$ is a Gamma function (Chakraborty & Griffiths 1982).

In addition, the expectation and variance of the total number of alleles (mtDNA-morphs) in a sample of size n are given by Ewens (1972)

$$E(k) = \theta \cdot \sum_{r=0}^{n-1} (\theta + r)^{-1} \tag{2}$$

and

$$V(k) = \theta \cdot \sum_{r=0}^{n-1} r / (\theta + r)^2 \tag{3}$$

in which θ is the same as defined in equation (1).

Obviously, evaluations of equations (1)–(3) require the knowledge of the composite parameter $\theta = 2N_{e(f)}v$, for which two alternatives are suggested in the literature. In the first, called the gene-diversity estimator, the observed mtDNA-morph frequency distribution (k_r; $r = 1, 2, \ldots$) is used to generate an unbiased estimator of the function $\theta/(1+\theta)$, the expected gene-diversity in the population. Nei (1978) has shown that

$$\hat{H} = \frac{n}{n-1}\left[1 - \sum_{r=1}^{n} r^2 k_r / n^2\right] \tag{4}$$

is an unbiased estimator of $\theta/(1+\theta)$; i.e., $E(\hat{H}) = \theta/(1+\theta)$. Therefore, a candidate estimator of θ is given by the gene-diversity estimator

$$\hat{\theta}_{GS} = \hat{H}/(1 - \hat{H}), \tag{5}$$

for which an approximate sampling variance formula is given by Chakraborty & Schwartz (1990) as

$$V(\hat{\theta}_{GS}) \approx \frac{2\theta(1+\theta)^2}{(2+\theta)(3+\theta)} - \frac{2(1+\theta)^3}{n} \tag{6}$$

Alternatively, θ can be estimated using the maximum likelihood (ML) method suggested by Ewens (1972), in which the ML-estimator of θ, denoted by $\hat{\theta}_{MLE}$, satisfies the equation

$$k = \hat{\theta}_{MLE} \cdot \sum_{r=0}^{n-1} (\hat{\theta}_{MLE} + r)^{-1} \tag{7}$$

whose sampling variance has the close form (see Chakraborty & Schwartz 1990 for a derivation)

$$V(\hat{\theta}_{MLE}) \approx \theta / [\sum_{r=0}^{n-1} r / (\theta + r)^2] \tag{8}$$

While the use of either of the above two alternative estimators of θ can be used to evaluate the expected number of alleles (mtDNA-morphs) with a specified number of copies in a sample (equation 1), from a pure statistical consideration one might be inclined to use the MLE, $\hat{\theta}_{MLE}$, because it is more efficient than $\hat{\theta}_{GS}$ (i.e., $\hat{\theta}_{MLE}$ has smaller sampling variance compared to $\hat{\theta}_{GS}$). However, as we shall see below, when the observed distribution of k_r is at discrepancy with the prediction of equation (1), $\hat{\theta}_{GS}$ is a more realistic estimator, because problems such as hidden subdivision affects $\hat{\theta}_{MLE}$ to deviate from the true value more substantially compared to $\hat{\theta}_{GS}$ (Chakraborty et al. 1988, Chakraborty & Schwartz 1990). It should also be stated that none of these estimators are unbiased estimator of θ, for which no formulation is available in the current literature.

For our purpose, we shall use both estimators ($\hat{\theta}_{GS}$ and $\hat{\theta}_{MLE}$) to demonstrate that certain features of the allele frequency distribution always detect hidden subdivision in a population irrespective of the choice of estimators.

Results

Table 1 shows the summary statistics of the mtDNA survey reported by Harihara et al. (1988) for each of the five Asian populations mentioned earlier, and for the pooled sample. In addition, this table also presents the two estimators of θ, ($\hat{\theta}_{MLE}$ from k through an iterative solution of equation 7, and $\hat{\theta}_{GS}$ from H using equation 5), along with their standard errors.

Two features of these estimates are notworthy. First, there is a direct positive association between the estimates of θ from k (i.e., $\hat{\theta}_{MLE}$) with the sample size, while this is not so for the estimate from H (i.e., $\hat{\theta}_{GS}$). This feature is parallel to the observations noted by Chakraborty & Schwartz (1990) in the context of analyzing the surname frequency distributions in England and Wales (Fox & Lasker 1983) and in Italy (Zei et al. 1983). This raises doubt as to whether the relative magnitude of the

Table 1. Summary statistics of mtDNA surveys from five Asian populations (adapted from Harihara et al. 1988).

Populations	n	k	$\hat{H}$	Estimate of $\theta = 2N_{e(f)}v$ from	
				$\hat{H}$ ($\hat{\theta}_{GS}$)	k ($\hat{\theta}_{MLE}$)
Japanese	74	11	0.400 ± 0.072	0.681 ± 0.202	3.341 ± 1.240
Ainu	48	6	0.231 ± 0.081	0.309 ± 0.136	1.588 ± 0.808
Korean	64	7	0.332 ± 0.075	0.509 ± 0.169	1.796 ± 0.832
Aëta	37	3	0.199 ± 0.086	0.257 ± 0.132	0.569 ± 0.434
Vedda	20	4	0.510 ± 0.104	1.159 ± 0.462	1.209 ± 0.808
Pooled	243	20	0.340 ± 0.040	0.517 ± 0.092	4.991 ± 1.307

n = Number of individuals sampled.
k = Observed number of different mtDNA-morphs in the sample.
$\hat{H}$ = Gene-diversity in the sample (computed by equation 4).

$\hat{\theta}_{MLE}$ values for these five populations truly reflect their differences of effective sizes, as it should, since the same mtDNA-genome has been examined in these surveys and hence the mutation rate component (v) should be the same, unless it varies across populations. Second, the estimate of θ from k for the pooled sample is much larger than those of the individual samples, while the pooled estimate of θ from $\hat{H}$ (pooled gene diversity in the entire sample) is within the range of the individual sample estimates. These two features suggest that the total number of mtDNA-morphs observed in the sample may not follow the prediction of equation (2), which is the premise of the maximum likelihood estimator of θ ($\hat{\theta}_{MLE}$, equation 5).

To check this assertion, we substituted the stimators of θ in equation (1) to get the expectation k_r for all values of r = 1, 2, ..., for each of the populations sampled by Harihara et al. (1988). These are shown in Tables 2 through 4 for the Japanese, the Ainu, and the Koreans, respectively; and in Table 5 for the pooled sample.

The standard errors of the estimates of k_r, shown in these tables are computed by the formula given in Chakraborty & Griffiths (1982). A comparison of the expected values of k_r with the observed frequencies show three features: 1) Within each population when θ is estimated from $\hat{H}$ (i.e., $\hat{\theta}_{GS}$), the observed mtDNA-morph distributions show excess of the total number of mtDNA-morphs compared to the neutral expectation, and this excess is mainly due to the excess of rare morphs (i.e., those occurring with few number of copies in the samples); 2) this phenomenon is more conspicuous in the pooled sample suggesting that perhaps the reason of the above observation is the fact that within each of the defined populations there is hidden subdivision; and 3) θ is estimated from the total number of observed mtDNA-morphs in the sample (i.e., $\hat{\theta}_{MLE}$), although the expectation and observed for the total number of morphs agree with each other (as it should, because of equation 7), there are excesses of the rare morphs, compensated by deficiencies in the number of common morphs in the sample, and this phenomenon is more prominent in the pooled sample. In other words, irrespective of the estimator used for θ, these data exhibit a departure from the predictions of the neutral allele theory, suggesting

Table 2. mtDNA-morph profile in the Japanese population (based on data of Harihara et al. 1988).

Number of copies	Frequencies		
		Expected based on θ from	
	Obs.	$\hat{H}$ ($\hat{\theta}_{GS}$)	k ($\hat{\theta}_{MLE}$)
1	6	0.684 ± 0.827	3.238 ± 1.798
2	2	0.343 ± 0.586	1.569 ± 1.251
3	1	0.230 ± 0.480	1.013 ± 1.004
4	1	0.173 ± 0.416	0.736 ± 0.855
57*	1	2.370 ± 1.556	4.444 ± 2.206
Total	11	3.800 ± 2.434	11.000 ± 2.693

* The expected for this category represents frequencies for 5 or more copies.

Table 3. mtDNA-morph profile in the Ainu population (based on data of Harihara et al. 1988).

Number of copies	Frequencies		
		Expected based on θ from	
	Obs.	$\hat{H}$ ($\hat{\theta}_{GS}$)	k ($\hat{\theta}_{MLE}$)
1	4	0.313 ± 0.560	1.569 ± 1.252
2	1	0.159 ± 0.399	0.775 ± 0.880
42*	1	1.770 ± 1.346	3.656 ± 2.013
Total	6	2.243 ± 1.067	6.000 ± 1.965

* The expected for this category represents frequencies for 3 or more copies.

Table 4. mtDNA-morph profile in the Korean population (based on data of Harihara et al. 1988).

Number of copies	Frequencies		
		Expected based on θ from	
	Obs.	$\hat{H}$ ($\hat{\theta}_{GS}$)	k ($\hat{\theta}_{MLE}$)
1	3	0.513 ± 0.716	1.774 ± 1.332
2	1	0.259 ± 0.508	0.876 ± 0.936
3	1	0.174 ± 0.416	0.576 ± 0.759
4	1	0.131 ± 0.362	0.427 ± 0.653
52*	1	2.018 ± 1.542	3.346 ± 1.937
Total	7	3.094 ± 1.363	7.000 ± 2.160

* The expected for this category represents frequencies for 5 or more copies.

Table 5. mtDNA-morph profile in the pooled Asian population (based on data of Harihara et al. 1988).

Number of copies	Frequencies		
		Expected based on θ from	
	Obs.	$\hat{H}$ ($\hat{\theta}_{GS}$)	k ($\hat{\theta}_{MLE}$)
1	8	0.519 ± 0.720	4.910 ± 2.216
2	6	0.260 ± 0.510	2.415 ± 1.554
3	1	0.174 ± 0.417	1.584 ± 1.258
4	1	0.130 ± 0.361	1.169 ± 1.080
5	2	0.105 ± 0.323	0.920 ± 0.958
9	1	0.059 ± 0.241	0.478 ± 0.690
197*	1	2.570 ± 1.806	8.525 ± 2.995
Total	20	3.815 ± 1.603	20.000 ± 3.818

* The expected for this category represents frequencies for 10 or more copies.

that the mtDNA-morph distributions in these Asian populations show excess of rare morphs, indicative of the presence of hidden subdivisions within each of the populations studied.

As a consequence of these results, if one uses $\hat{\theta}_{MLE}$ as a valid estimator for θ in such surveys, we further note that the expected gene diversity, $\hat{\theta}_{MLE}/(1 + \hat{\theta}_{MLE})$, becomes much larger than the observed. This is shown in Table 6 for each population sample, and for the pooled data; suggesting that if internal subdivision is the cause of the above-noted discrepancy, $\hat{\theta}_{MLE}$ may not be an appropriate estimator of θ. On the contrary, there are some circumstances under which even in the presence of hidden subdivision, the estimator $\hat{\theta}_{GS}$ (from observed gene diversity) may not be in too much error, as argued in Chakraborty et al. (1988) and Chakraborty & Schwartz (1990).

Table 6. Gene-diversity for the mtDNA genome for five Asian populations (from data of Harihara et al. 1988).

Population	n	Heterozygosity (H)	
		Obs. ± s.e.	Exp. (from $\hat{\theta}_{MLE}$)
Japanese	74	0.400 ± 0.072	0.770
Ainu	48	0.231 ± 0.081	0.609
Korean	64	0.332 ± 0.075	0.642
Aẽta	37	0.199 ± 0.086	0.363
Vedda	20	0.510 ± 0.104	0.547
Pooled	243	0.340 ± 0.040	0.833

Allele frequency profile at isozyme loci in three major races of man

While the above section demonstrates that the presence of hidden subdivision may be detected by studying data from single locus, one criticism of such an analysis is that the theory used in this context is known to have a large sampling variance, due to stochastic error accumulated during evolution (Li & Nei 1975, Nei 1978). Therefore, data from a single locus may easily cause deviation due to this artifact. In this section, therefore, we show that the observations noted above from the analysis of mtDNA data are not due to stochastic errors alone, which can be substantially reduced if a large number of loci are used together to perform similar analysis. This is done here using the isozyme surveys reported by Neel et al. (1988) and Mohrenweiser et al. (1987) who used uniform laboratory methods to detect isozyme variations in three cosmopolitan populations: US Whites and US Blacks from a survey of cord blood samples from new borns in Ann Arbor, Michigan; and Japanese from Hiroshima and Nagasaki, studied to examine the effect of radiation exposure during the atom-bomb exposure. Table 7 presents a summary of their findings, pertinent details of which can be found in the original reports.

Table 7. Summary statistics from isozyme surveys in the three major racial groups of humans (adapted from Neel et al. 1988, Mohrenweiser et al. 1987).

Statistics	Japanese	US White	US Black
No. of loci surveyed	32	51	51
No. of gene sampled/locus	29,272	4,435	374
Av. heterozygosity/locus (in %)	8.699	5.011	5.230
Av. no. of alleles/locus	5.531	2.608	1.667
Av. no. of singleton alleles/locus	2.031	0.843	0.196
Prop. of variant loci	0.875	0.667	0.471

From these summary statistics, we can use equations (5) and (7) to obtain the two alternative estimators of θ, and derive expectations for the number of alleles per locus for each specific allele frequency classes, substituting such parameter estimates in equation (1). Figs. 1a, 1b, and 1c and Table 8 show the contrast of the observed allele frequency distributions for all loci pooled together in these three populations separately. For brevity, we show only the predictions based on the estimator $\hat{\theta}_{GS}$, since the qualitative results are similar for the other estimator ($\hat{\theta}_{MLE}$) as well. These figures and Table 8 clearly show that within each of the three major racial groups of man, there is a conspicuous excess of total number of alleles compared to the neutral expectation, and the excess is mainly due an increase in the number of rare alleles.

Fig. 1. Observed (blank bars) and expected (black bars) allele frequency distributions from isozyme data in the three major racial groups of man. Panel (a) = Japanese; Panel (b) = US White; and Panel (c) = US Black. The raw data are given in Mohrenweiser et al. (1987) and Neel et al. (1988). The expectations are based on the parameter estimate $\hat{\theta}_{GS}$, and employs the theory of Chakraborty & Griffiths (1982).

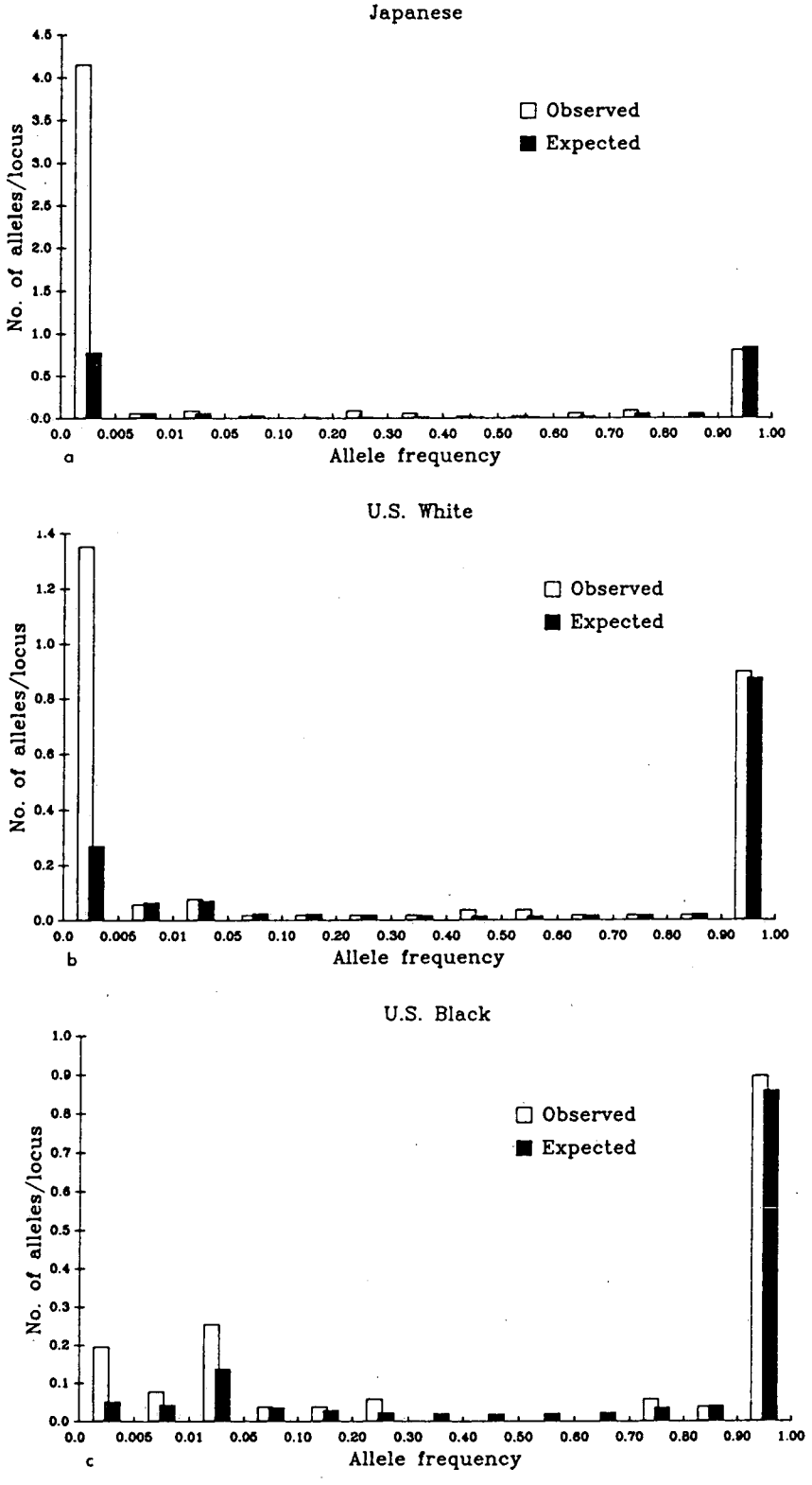

Japanese
Observed
Expected
No. of alleles/locus
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0 0.005 0.01 0.05 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
a
Allele frequency
U.S. White
Observed
Expected
No. of alleles/locus
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.005 0.01 0.05 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
b
Allele frequency
U.S. Black
Observed
Expected
No. of alleles/locus
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0 0.005 0.01 0.05 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
c
Allele frequency

Table 8. Allele frequency distributions per locus from isozyme data in the three major racial groups of man.

Allele frequency class	Japanese		US White		US Black	
	Obs.	Exp.*	Obs.	Exp.*	Obs.	Exp.*
0 -0.005	4.156	0.783	1.353	0.270	0.196	0.052
0.005-0.01	0.063	0.068	0.059	0.065	0.078	0.044
0.01 -0.05	0.094	0.063	0.078	0.072	0.255	0.138
0.05 -0.10	0.031	0.032	0.020	0.026	0.039	0.037
0.10 -0.20	0.0	0.019	0.020	0.024	0.039	0.030
0.20 -0.30	0.094	0.018	0.020	0.020	0.059	0.024
0.30 -0.40	0.063	0.018	0.020	0.018	0.0	0.022
0.40 -0.50	0.031	0.016	0.039	0.016	0.0	0.020
0.50 -0.60	0.031	0.016	0.039	0.016	0.0	0.022
0.60 -0.70	0.063	0.026	0.020	0.019	0.0	0.024
0.70 -0.80	0.094	0.058	0.020	0.020	0.059	0.037
0.80 -0.90	0.0	0.060	0.020	0.024	0.039	0.041
0.90 -1.00	0.813	0.843	0.902	0.879	0.902	0.864
Total	5.531	2.021	2.608	1.469	1.667	1.354

* The expected number of alleles in a gene frequency class is computed by the theory of Chakraborty & Griffiths (1982), using the estimate $\hat{\theta}_{GS}$ from the average gene-diversity. All expectations refer to the appropriate sample size of the relevant surveys (see Table 7).

The distributions shown in these three figures suggest that such excess can be detected even by a graphic display of the data. An objective test of significance of such departures (not attempted here) can be done using a result from the theory of Chakraborty & Griffiths (1982), who showed that the number of rare alleles (e.g., alleles with gene frequency in the range below 1 % or 5 %) approximately follows a Poisson distribution, and hence appropriate exact tests may be conducted using the expected frequency of such alleles as the estimated parameter of the Poisson distribution, with sample size being the number of loci involved in such calculations (e.g., the significance of the departure of the observed frequency of 4.156 alleles with gene frequency below 0.5 % per locus in the Japanese population compared to its expectation of 0.783 can be tested with a Poisson test assuming the parameter value of 0.783 and sample size 32, the number of loci).

In summary, these data show that the observed isozyme allele frequency profile in each of the three major races of man depart from the neutral predictions, exhibiting an excess of rare alleles, which is the main contributor of the apparent excess of total number of alleles per locus. In conjunction to this we also note that the proportion of variant loci (i.e., the loci with at least one variant allele observed) is also in excess of the expectation. These summary observations are shown in Table 9, particularly using the gene-diversity estimator, $\hat{\theta}_{GS}$.

As in the case of mtDNA data, we also note that when $\hat{\theta}_{MLE}$ is used as an estimator of θ, the predicted level of heterozygosity ($\hat{H}$) becomes too large compared to the observed in each of these three populations. This is shown in Table 10.

All of the findings of the single-locus analysis using mtDNA-morph distributions in the five Asian populations are comparable with the isozyme surveys in the major racial groups of man. In other words, the conclusions derived from the mtDNA survey are not due to an artifact of stochastic errors associated with single-locus data.

Table 9. Summary statistics of departures from the neutral model in the isozyme variation in three racial groups of man.

Statistics	Observed	Expectd ± s.e.
Japanese ($\hat{\theta}_{GS} = 0.095$; n = 29,272)		
Total no. of alleles/locus	5.531	2.021 ± 0.177
No. of singleton alleles/locus	2.031	0.095 ± 0.055
Proportion of variant loci	0.875	0.625 ± 0.086
US White ($\hat{\theta}_{GS} = 0.053$; n = 4,435)		
Total no. of alleles/locus	2.608	1.469 ± 0.095
No. of singleton alleles/locus	0.843	0.053 ± 0.032
Proportion of variant loci	0.667	0.358 ± 0.067
US Black ($\hat{\theta}_{GS} = 0.055$; n = 374)		
Total no. of alleles/locus	1.667	1.354 ± 0.083
No. of singleton alleles/locus	0.196	0.052 ± 0.032
Proportion of variant loci	0.471	0.279 ± 0.063

Table 10. Observed and expected gene-diversity from the MLE of θ in the three major racial groups of man (from data of Mohrenweiser et al. 1987 and Neel et al. 1988).

Populations	n	k	$\hat{\theta}_{MLE}$	Gene-diversity	
				Obs.	Exp.
Japanese	29,272	5.531	0.440	0.087 ± 0.026	0.305
US White	4,435	2.608	0.185	0.050 ± 0.018	0.156
US Black	374	1.667	0.105	0.052 ± 0.015	0.095

Cause and implication of excess of rare alleles in cosmopolitan populations

Having shown that the excess of rare alleles is perhaps a general rule in many cosmopolitan populations, several causal mechanisms may be proposed which may explain this phenomenon, indicated in the introductory section of this presentation. We argue that hidden substructuring could be a major factor, although it cannot be clearly shown from these data alone. However, if we examine the above observations in the light of the natural experiment conducted in Chakraborty et al. (1988), we can easily argue that none of the populations examined here are probably single breeding

units within themselves, and hence the observed departure is an effect of hidden subdivision. The questions, therefore, are: How many subpopulations are necessary to explain the observed discrepancy, and what are the implications of such discrepancy. The first question cannot be unequivocally answered from the above data alone, because there are two factors of hidden subdivision which pertains to the above-noted discrepancy, shown in Chakraborty et al. (1988). To re-capitulate that analysis, this earlier study showed that within each of the individual South and Central American Indian tribes, the allele frequency profile follows the prediction of the neutral allele theory suggesting that the genetic variations of the isozymes used in the major racial group analysis here may indeed be selectively neutral. However, when the data from 12 tribes are pooled to compose a single hypothetical Amerindian population, the allele frequency profile in such a composite population departs from the expectations of the neutral model, all features of which parallel to the present observations. This is graphically shown in Fig. 2, where the expected allele frequency distributions (denoted by black bars) are derived based on the gene-diversity estimator $\hat{\theta}_{GS}$.

Chakraborty et al. (1988) further showed that the complexity of hidden subdivision is composed of two factors: 1) The number of subpopulations involved, and 2) the average degree of genetic differentiation among them. A larger number of subpopulations with small degree of genetic divergence among them may have effect similar to a small number of subpopulations with large genetic distances among them. This result implies that unless there is any other independent information regarding the extent of subdivision, the number of subpopulations cannot be esti-

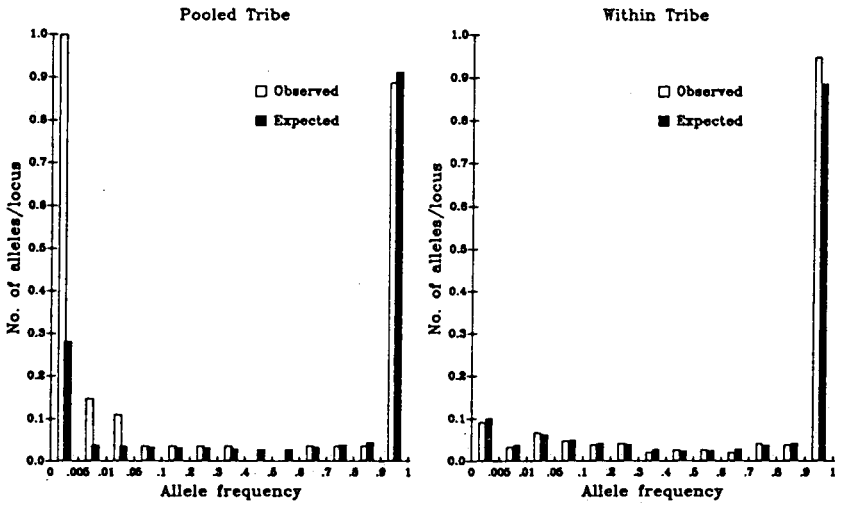


Fig. 2. Observed (blank bars) and expected (black bars) allele frequency distributions from isozyme data in 12 Amerindian tribes from South and Central America. The pooled distribution is from the allele frequency distribution in the total population (pooled over 12 tribes), and the within tribe distribution is the average of the individual tribe-specific distributions. The raw data are given in Chakraborty et al. (1988).

mated precisely. However, some knowledge of the history of a population may provide information regarding the time depth of isolation of subpopulations within it. If such data are available, one might suggest a reasonable value of effective genetic distance among the constituent subpopulations, from which an estimate of the necessary number of subpopulations may be made to explain the observed excess of allele numbers.

This brings us to the second question, namely the implication of the observed excess of allele numbers in human populations. Note that a translation of the time depth of isolation to a genetic distance (see Nei 1975, 1987) is possible only when the mutation rate estimates are available. The current indirect estimates of mutation rate from isozyme studies are all based on theories that depend upon the assertion that the observed number of rare (or total) alleles is in accordance with the neutral expectations. This is where the hidden subdivision effect is most prominent. Therefore, if hidden subdivision is prevalent in most cosmpolitan populations, the indirect estimates of mutation rate may be grossly overestimated. Note that in the most recent attempt to provide an indirect estimate of mutation rate from human isozyme data, Chakraborty & Neel (1989) used the Amerindian populations that within themselves follow the neutral allele frequency profile, and these estimates are much closer to the direct estimates (although still differ by a factor of two). Comparison of mutation rate estimates in different populations, therefore, should be made with caution, because the extent of hidden subdivisions may be substantially different among them.

Conclusion and discussion

The main conclusion of the above analyses of the mtDNA and isozyme data is that there is an apparent excess of the total number of alleles or mtDNA-morphs in all populations analyzed. This excess is probably due to hidden subdivision within each population, since this observation is parallel to what is seen in the Amerindian study (Chakraborty et al. 1988) where such a discrepancy could be created artificially by amalgamating tribes with different degrees of genetic differentiation among them.

Comparing the results of Tables 2 and 6, it is reassuring that the excess of total number of variants in the Japanese population is quite consistent between the mtDNA survey (2.89-fold increase in relation to the neutral expectation) and in the isozyme data (2.74-fold increase). This is somewhat larger than the excess seen in the Amerindian study where 12 tribes have been amalgamated (resulting in a 1.6-fold increase). Although these numbers should not be interpreted at their face value, since the extent of excess is a complex function of average degree of polymorphism within a population, number of constituent subpopulations, and their genetic isolation. However, considering the three populations of the present isozyme data analysis, we see that the Japanese population exhibit a higher extent of the excess of total number of alleles. Even though the Japanese isozyme survey (Neel et al. 1988) involves a smaller set of loci (32) than the US Whites and US Blacks (Mohrenweiser et al. 1987; where 51 loci were scored), and the average gene-diversity in the Japanese population is somewhat larger (8.7 %) in comparison with the US Whites or Blacks (approximately 5 %), the 2.7-fold excess in the Japanese population as compared to the 1.8-fold excess in the US Whites, or 1.3-fold excess in the US Blacks is reasonable,

since the US Blacks represent an admixed population formed by slave-trading from a smaller region of Africa (mostly West Africa; see Adams & Ward 1973) and has the history of only five centuries of gene-admixture with the Caucasians. The US white population probably consists of substructuring due to their different European ancestry, where the genetic differences can be substantially smaller than the probable constituent Japanese subpopulations.

These deductions are of course tentative, because a firm conclusion in this regard should require an adjustment for sample size differences between the surveys, eliminate any possible bias due to differences of the loci employed in the analysis, and finally must adjust for probable differences of genetic distances among the constituent subpopulations. This issue will be considered in a greater details elsewhere (Chakraborty et al., in prep.).

As mentioned earlier, one important implication of this demonstration is that the excess of allele numbers in comparison to the neutral expectation is dependent on the internal genetic structure of the population. Therefore, when an indirect estimate of mutation rate is based on the number of rare alleles, any inter-survey comparison of mutation rate estimates from different populations should be made with caution. While in the past there had been some suggestions for the possibility of primitive populations exhibiting larger mutation rates than the modern cosmopolitan populations, in view of the present observation this postulation cannot be persued any longer vigorously (see Chakraborty & Neel 1989 for a different reason for this).

Another implication of the present set of observations relates to the comparability of molecular variation at the DNA level, detected by restriction fragment length polymorphisms (RFLPs) with that at the isozyme loci detected by electrophoresis. It is true that electrophoresis detects only a fraction (between one-third to one-fourth) of molecular variation because of the fact that nucleotide changes that do not change the charge of a protein molecule is not detected by electrophoresis. In contrast, as long as the recognition sequence is altered, such changes will be detected by the RFLP-technique. However, at present population data on RFLPs exist for populations that are very loosely defined, leaving enough scope of internal hidden subdivisions within them. Therefore, from the greater number of variant RFLP alleles (or haplotypes) alone one cannot judge how enhanced is the degree of detection of genetic variation by this improved technique. The unit of population should be precise for both technologies, should a more precise estimate of detectability of molecular polymorphism be made, even when we adjust for hidden variation in both techniques.

The analyses presented here also reflect that even if the populations of slightly deleterious mutations and population bottleneck do not apply to a particular population, there may be a more realistic factor (namely, hidden subdivision) that could cause specific departures of the allele frequency profile from the prediction of the neutral mutation hypothesis. Hidden subdivision is a factor that can be more incisively examined. For example, enlarged sampling from different geographic regions within a population may reveal the extent of subdivision. Such geographic microdifferentiation has been demonstrated with allele frequency data at specific loci in Japan (e.g., Nei & Imaizumi 1966a, b), or by population structure analysis employing Wright's F-statistics (e.g., Neel & Ward 1972). Therefore, should hidden subdivision be the only cause of departure of the allele frequency profile from the neutral expectations, this can be easily tested. Such validations of the bottleneck hypothesis or slightly deleterious mutation model are not readily available.

Finally, it should be noted that while this presentation emphatically argues for the more thorough examination of the possibility that many cosmopolitan populations may indeed have internal hidden substructuring, causing its allele frequency profile to depart substantially from the neutral expectations, this does not negate the existence of other probable causes, which might simultaneously affect a population during the course of its evolution. Nevertheless, since it is a testable proposition, hidden subdivision may be detected and validated. Even after this is done, if it cannot account for all the departures, other hypotheses may be entertained for understanding the evolutionary mechanisms that may generate the observed allele frequency structure of the population.

Acknowledgements

This work was supported by US Public Health Service grant GM 41399 from the US National Institutes of Health. I thank Drs. J.V. Neel and P.E. Smouse for their collaborations for some of the findings cited here, and Drs. M. Nei and W.J. Schull for their suggestions and comments during the conduct of this work.

References

Adams, J. & Ward, R.H., 1973: Admixture studies and detection of selection. – Science 180, 1137–1143.

Ayala, F.J., 1976: Molecular Evolution. – Sinauer, Sunderland.

Blanc, H., Chen, K.H., D'Amore, M.A. & Wallace, D.C., 1983: Amino acid change associated with the major polymorphic *Hinc*-II site of Oriental and Caucasian mitochondrial DNAs. – Amer. J. Hum. Genet. 35, 167–176.

Brown, W.M., 1980: Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. – Proc. Natl. Acad. Sci. USA 77, 3605–3609.

Brown, W.M. & Goodman, H.M., 1979: Quantitation of intra-population variation by restriction endonuclease analysis of human mitochondrial DNA. – In: Cummings, D., Borst, P., Dawid, I., Weissman, S. & Fox, C.F. (eds.): Extrachromosomal DNA, ICN-UCLA Symposia, 485–500. – Academic Press, New York.

Bruce, E.J. & Ayala, F.J., 1979: Phylogenetic relationships between man and the apes: Electrophoretic evidence. – Evolution 33, 1040–1056.

Cann, R.L., Brown, W.M. & Wilson, A.C., 1982: Evolution of human mitochondrial DNA: A preliminary report. – In: Bonne-Tamir, B., Cohen, P. & Goodman, R.N. (eds.): Human Genetics – Part A: The Unfolding Genome, 157–165. – Alan R. Liss, New York.

Cann, R.L., Stoneking, M. & Wilson, A.C., 1987: Mitochondrial DNA and human evolution. – Nature 325, 31–36.

Cann, R.L. & Wilson, A.C., 1983: Length mutations in human mitochondrial DNA. – Genetics 104, 699–711.

Chakraborty, R., 1981: Expected number of rare alleles per locus in a sample and estimation of mutation rates. – Amer. J. Hum. Genet. 33, 481–483.

Chakraborty, R., Fuerst, P.A. & Nei, M., 1980: Statistical studies on protein polymorphism in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. – Genetics 94, 1039–1063.

Chakraborty, R. & Griffiths, R.C., 1982: Correlation of heterozygosity and number of alleles in different frequency classes. – Theor. Pop. Biol. 21, 205–218.

Chakraborty, R. & Neel, J.V., 1989: Description and validation of a method for simultaneous estimation of effective population size and mutation rate from human population data. - Proc. Natl. Acad. Sci. USA 86, 9407-9411.

Chakraborty, R. & Nei, M., 1977: Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. - Evolution 31, 347-356.

Chakraborty, R. & Schwartz, R.J., 1990: Selective neutrality of surname distribution in an immigrant Indian community of Houston, Texas. - Amer. J. Hum. Biol. 2, 1-15.

Chakraborty, R., Smouse, P.E. & Neel, J.V., 1988: Population amalgamation and genetic variation: Observations on artificially agglomerated tribal populations of Central and South America. - Amer. J. Hum. Genet. 43, 709-725.

Denaro, M., Blanc, H., Johnson, M.J., Chen, K.H., Wilmsen, E., Cavalli-Sforza, L.L. & Wallace, D.C., 1981: Ethnic variation in *Hpa*-I endonuclease cleavage patterns of human mitochondrial DNA. - Proc. Natl. Acad. Sci. USA 78, 5768-5772.

Ewens, W.J., 1972: The sampling theory of selectively neutral alleles. - Theor. Pop. Biol. 3, 87-112.

Fox, W.R. & Lasker, G.W., 1983: The distribution of surname frequencies. - Int. Stat. Rev. 51, 81-87.

Goodnight, C.J., 1987: On the effect of founder events on epistatic genetic variance. - Evolution 41, 80-91.

Harihara, S., Saitou, N., Hirai, M., Gojobori, T., Park, K.S., Misawa, S., Ellepola, S.B., Ishida, T. & Omoto, K., 1988: Mitochondrial DNA polymorphism among five Asian populations. - Amer. J. Hum. Genet. 43, 134-143.

Horai, S., Gojobori, T. & Matsunaga, E., 1986: Distinct clustering of mitochondrial DNA types among Japanese, Caucasians and Negroes. - Jap. J. Genet. 61, 271-275.

——— 1987: Evolutionary implications of mitochondrial DNA polymorphism in human populations. - Hum. Genet. 74, 177-181.

Horai, S. & Matsunaga, E., 1986: Mitochondrial DNA polymorphism in Japanese. II. Analysis with restriction enzymes of four or five base pair reconition. - Human Genet. 72, 105-117.

Johnson, M.J., Wallace, D.C., Ferris, S.D., Rattazzi, M.C. & Cavalli-Sforza, L.L., 1983: Radiation of human mitochondrial DNA types analyzed by restriction endonuclease cleavage patterns. - J. Mol. Evol. 19, 255-271.

Kimura, M., 1983: The Neutral Theory of Evolution. - Cambridge University Press, Cambridge.

King, M.C. & Wilson, A.C., 1975: Evolution at two levels: Molecular similarities and biological differences between humans and chimpanzees. - Science 188, 107-116.

Landsteiner, K. & Levine, P., 1928: On the inheritance of agglutinogens of human blood demonstrable by immune agglutinins. - J. Exp. Med. 48, 731-749. (Reprinted in: Schull, W.J. & Chakraborty, R. (eds.): Human Genetics - A Selection of Insights, 39-57. Dowden, Hutchinson & Ross, Stroudsburg, PA.)

Lewontin, R.C., 1974: The Genetic Basis of Evolutionary Change. - Columbia University Press, New York.

Lewontin, R.C. & Cockerham, C.C., 1959: The goodness-of-fit test for detecting natural selection in random mating populations. - Evolution 13, 561-564.

Li, W.-H., 1978: Maintenance of genetic variability under the joint effect of mutation, selection and random drift. - Genetics 90, 349-382.

Li, W.-H. & Nei, M., 1975: Drift variances of heterozygosity and genetic distance in transient states. - Genet. Res. 25, 229-248.

Maruyama, T. & Fuerst, P.A., 1984: Population bottlenecks and non-equilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. - Genetics 108, 745-763.

—— 1985: Population bottlenecks and non-equilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. - Genetics 111, 691-703.

Milkman, R., 1973: Electrophoretic variation in *Escherichia coli* from natural sources. - Science **182**, 1024-1026.

Mohrenweiser, H.W., Wurzinger, K.H. & Neel, J.V., 1987: Frequency and distribution of rare electrophoretic mobility variants in a population of newborns in Ann Arbor, Michigan. - Ann. Hum. Genet. **51**, 303-316.

Mourant, A.E., Kopec, A.C. & Domaniewska-Sobczak, K., 1976: The Distribution of the Human Blood Groups and Other Polymorphisms, 2nd ed. - Oxford University Press, New York.

Neel, J.V. & Rothman, E.D., 1978: Indirect estimates of mutation rates in tribal Amerindians. - Proc. Natl. Acad. Sci. USA **75**, 5585-5588.

Neel, J.V., Satoh, C., Smouse, P.E., Asakawa, J., Takahashi, N., Goriki, K., Fujita, M., Kageoka, T. & Hazama, R., 1988: Protein variants in Hiroshima and Nagasaki: Tales of two cities. - Amer. J. Hum. Genet. **43**, 870-893.

Neel, J.V. & Ward, R.H., 1972: The genetic structure of a tribal population, the Yanomama Indians. VI. Analysis by F-statistics (including a comparison with the Makiritare and Xavante). - Genetics **72**, 639-666.

Nei, M., 1975: Molecular Population Genetics and Evolution. - North-Holland - American Elsevier, New York.

— 1977: Estimation of mutation rate from rare protein variants. - Amer. J. Hum. Genet. **29**, 225-232.

— 1978: Estimation of average heterozygosity and genetic distance from a small number of individuals. - Genetics **89**, 583-590.

— 1987: Molecular Evolutionary Genetics. - Columbia University Press, New York.

Nei, M. & Graur, D., 1984: Extent of protein polymorphism and the neutral mutation theory. - Evol. Biol. **17**, 73-118.

Nei, M. & Imaizumi, Y., 1966a: Genetic structure of human populations. I. Local differentiation of blood group gene frequencies in Japan. - Heredity **21**, 9-25.

— — 1966b: Genetic structure of human populations. II. Differentiation of blood gruop gene frequencies among isolated populations. - Heredity **21**, 183-190.

Nei, M., Maruyama, T. & Chakraborty, R., 1975: The bottleneck effect and genetic variability in populations. - Evolution **29**, 1-10.

Ohta, T., 1973: Slightly deleterious mutant substitutions in evolution. - Nature **246**, 98-98.

— 1976: Role of very slightly deleterious mutations in molecular evolution and polymorphism. - Theor. Pop. Biol. **10**, 254-275.

Roychoudhury, A.K. & Nei, M., 1987: Human Polymorphic Genes: World Distribution. - Oxford University Press, New York.

Wallace, D.C., Garrison, K. & Knowler, W.C., 1985: Dramatic founder effects in Amerindian mitochondrial DNAs. - Amer. J. Phys. Anthrop. **68**, 149-155.

Watterson, G.A., 1984: Allele frequencies after a bottleneck. - Theor. Pop. Biol. **26**, 387-407.

Zei, G., Matessi, R.G., Siri, E., Moroni, A. & Cavalli-Sforza, L.L., 1983: Surnames in Sardinia. I. Fit of frequency distributions for neutral alleles and genetic population structure. - Ann. Hum. Genet. **47**, 329-352.

Received November 9, 1989

Address for correspondence:
Prof. Dr. R. Chakraborty, Center for Demographic and Population Genetics, The University of Texas, Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, Texas 77225, USA.

Anthrop. Anz.	Jg. 48	4	313 - 331	Stuttgart, Dezember 1990

Genetic profile of cosmopolitan populations: Effects of hidden subdivision

R. Chakraborty

Center for Demographic and Population Genetics, The University of Texas, Graduate School of Biomedical Sciences, Houston, TX, USA

With 2 figures and 10 tables in the text

Summary: Natural populations of many organisms exhibit excess of rare alleles in comparison with the predictions of the neutral mutation hypothesis. It has been shown before that either a population bottleneck or the presence of slightly deleterious mutations can explain this phenomenon. A third explanation is presented in this work, showing that hidden subdivision within a population can also lead to an excess of rare alleles in the total population when the expectations of the neutral model are based on the allele frequency profile of the entire population data.

With two examples (mitrochondrial DNA-morph distribution and isozyme allele frequency distributions), it is shown that most cosmopolitan human populations exhibit excess of rare as well as total allele counts, when these are compared with the expectations of the neutral mutation hypothesis. The mitochondrial data demonstrate that such excesses can be detected from genetic variation at a single locus as well, and this is not due to stochastic error of allele frequency distributions. Contrast of the present observations with the allele frequency profiles in agglomerated tribal populations from South and Central America shows that even when the neutral expectations hold for individual subpopulations, if all subpopulations are grouped into a single population, the pooled data exhibit an excess of total number of alleles that is mainly due to the excess of rare alleles. Therefore, a primary cause of the excess number of rare alleles could be the hidden subdivision, and the magnitude of the excess indicates the extent of substructuring. The two components of hidden subdivision are: 1) Number of subpopulations, and 2) the average genetic distance among them. The implications of this observation in estimating mutation rate are discussed indicating the difficulties of comparing mutation rates form different population surveys.

Zusammenfassung: Natürliche Populationen zahlreicher Organismen weisen einen Überschuß an seltenen Allelen auf, der nicht mit der Hypothese neutraler Mutationen in Übereinstimmung steht. Es wurde zur Erklärung dieses Phänomens bisher angenommen, daß entweder „Bottleneck-Effekte" oder schwach nachteilige Mutationen in diesem Zusammenhang eine Rolle spielen. In dieser Untersuchung wird eine dritte Erklärungsmöglichkeit aufgezeigt, indem dargestellt wird, daß eine nicht offen erkennbare Strukturierung einer Population zu einem Überschuß an seltenen Allelen in der Gesamtpopulation führen kann, und zwar dann, wenn die Erwartungen nach dem Modell neutraler Mutationen auf dem Allelenfrequenzprofil für die Gesamtpopulation basieren.

An zwei Beispielen (mitochondriale DNA-morph-Verteilung und Verteilung der Allelenfrequenzen von Isoenzymen) wird gezeigt, daß die meisten menschlichen Großbevölkerungen Überschüsse sowohl seltener Allele als auch der Allelzahlen insgesamt zeigen, wenn diese mit den Erwartungswerten nach der Hypothese neutraler Mutationen verglichen werden. So lassen

0003-5548/90/0048-0313 $ 4.75

die Daten für die mitochondriale DNA erkennen, daß ein solcher Überschuß ebenso anhand der genetischen Variabilität an einem einzigen Genlocus entdeckt werden kann, was nicht durch stochastische Fehler der Allelenfrequenzverteilungen bedingt ist. Beobachtungen an zusammengesetzten süd- und mittelamerikanischen Stammesbevölkerungen zeigen, daß die Hypothese neutraler Mutationen für die einzelnen Stammesbevölkerungen durchaus zutreffen kann. Wenn jedoch alle Subpopulationen zu einer Gesaamtpopulation zusammengefaßt werden, lassen die gepoolten Frequenzdaten einen Überschuß bezüglich der Gesamtallelenzahl erkennen, welcher hauptsächlich durch einen Überschuß an seltenen Allelen bedingt ist. Eine wesentliche Ursache hierfür ist offenbar in einer nicht offen erkennbaren Strukturierung der Gesamtbevölkerung zu sehen, und das Ausmaß des Überschusses reflektiert den Grad dieser Strukturierung. Die beiden Komponenten der verborgenen Bevölkerungsgliederung sind 1. die Zahl der Subpopulationen und 2. der durchschnittliche genetische Abstand zwischen ihnen. Die Bedeutung dieser Beobachtung für die Schätzung von Mutationsraten wird diskutiert, wobei auch auf die Schwierigkeiten hinsichtlich des Vergleichs von Mutationsraten eingegangen wird, die an verschiedenen Populationen ermittelt worden sind.

Introduction

In genetic analysis of population data, the genetic make-up of a population is generally studied in a variety of ways. The basic data for such analyses are frequencies of various alleles (or genotypes) at one or more loci, estimated from random samples drawn from a population. The population, in this context, is usually defined as a breeding unit, within which mating occurs with a well-specified pattern (generally assumed to be random).

Over the history of population genetics, the technique of detecting genetic variation has changed considerably. Initially, before the advent of serological techniques of blood grouping (Landsteiner & Levine 1928), morphological and physiological traits had been popular for studying genetic variations within and between populations. Soon after the discovery of blood group systems in humans, anthropological and human genetic studies started using these techniques extensively. As a result, morphological data had been now replaced by voluminous gene frequency surveys in various ethnic groups around the world (Mourant et al. 1976). The development of electrophoretic techniques introduced another set of traits which detect genetic variations based on charge and/or molecular size differences of protein-enzyme molecules. Since such changes are due to new mutations at the nucleotide level that are translated into mRNA during protein synthesis, the electrophoretically determined genetic variations were the first step of studying evolution at a molecular level. Genetic variation detected by electrophoresis, furthermore, does not depend upon the antigen-antibody specificity which is essential for the serological methods applied to detect the genetic variation at blood groups and immunological systems. Use of electrophoretic techniques, therefore, produced data not only in humans, but also on other organisms, ranging from *E. coli* (Milkman 1973) to primates (King & Wilson 1975, Bruce & Ayala 1979). It soon became apparent that genetic variation is widespread; its extent varies from organism to organism (Nei 1975), and for some organisms the extent of genetic variation vary widely over different geographic regions, depending upon other factors such as population size, reproductive isolation, and ecological conditions. These assertions have been even more firmly established by the recent molecular techniques of restriction fragment length polymorphisms (RFLPs), and nucleotide-sequencing technique, whereby the detection

of genetic variation is extended beyond the translated region of the DNA. Although these later techniques are far more powerful to study genetic variation at a molecular level, it should be noted that current knowledge of DNA polymorphism at a population level is still scanty compared to the electrophoretically determined polymorphisms (see Roychoudhury & Nei 1987 for a comparative compilation of the current data).

While such data convincingly established the ubiquity of genetic variation, there is still a question as to which evolutionary factors play a dominant role in maintaining such variations in natural populations. In other words, it is not certain if the main cause of extensive genetic polymorphism is natural selection, nor it is clear whether or not the genetic variation is being maintained by counteracting forces of mutation and random genetic drift (Lewontin 1974, Ayala 1976, Nei 1975, 1987, Kimura 1983). This controversy, called the selectionist-neutralist controversy, still remains unresolved for the reason that there are several features of data on genetic polymorphisms that cannot be rigorously explained by either of these hypotheses. For example, the observed average level of heterozygosity is generally too low in contrast with the predictions of the balancing selection hypothesis of genetic polymorphism (Nei & Graur 1984). At the same time there is a relative excess of the frequency of "rare" alleles in comparison with the expectation of the neutral mutation hypothesis, which is particularly noteworthy in many species, ranging from Drosophila to human (Ohta 1976, Chakraborty et al. 1980). This later observation led Ohta (1973) to propose that many of the new mutations that occur in nature are deleterious, but they are quickly eliminated from the population because of the negative selection against them. However, there are some "slightly" deleterious mutants which are not so quickly eliminated from the population because of the weakness of selection intensity against them. Negative selection against them, however, prevents their frequencies to attain intermediate or high level. As a result, these slightly deleterious mutations account for the relative excess of "rare" alleles in a population.

An alternative explanation of the excess of rare alleles is given by Nei et al. (1975), Chakraborty & Nei (1977), Maruyama & Fuerst (1984, 1985), and Watterson (1984) who proposed that in nature many populations (or species) are subject to drastic fluctuations of population sizes over time due to ecological and/or environmental changes. When a population goes through a sudden reduction of its size, the genetic variability in the gene pool is substantially reduced, and it takes a long time (of the order of the inverse of mutation rate, in units of generation length) to recover from the loss of genetic variation. On the other hand, the population size may return to the resource capacity of the population/species comparatively much earlier. This phenomenon, called "bottleneck", can easily explain the apparent excess of rare alleles under the premises of the neutral mutation hypothesis.

While both of these explanations are based on mathematical arguments that are difficult to refute, direct validations of "bottleneck" and/or "slightly" deleterious mutations are not available from data of natural populations. This is so because of the lack of past historical data on population sizes for many organisms, and the accurate measurement of selection coefficient for or against any specific allele is a difficult task (Lewontin & Cockerham 1959). In summary, it is not universally accepted that the population "bottleneck" is an wide-spread evolutionary phenomenon applicable for all species/populations in which an excess of rare alleles is observed (Goodnight 1987). In a similar vein, the hypothesis of slightly deleterious mutations has also its own caveat (Li 1978).

The purpose of this presentation is to show that there is another factor which can give rise to the same observation (excess of rare or total number of alleles compared to the expectations under the neutral mutation hypothesis). In an earlier publication (Cahkraborty et al. 1988) it is shown that when there is a hidden subdivision within a population, caused by microdifferentiation within a population, the allele frequency profile in the total population deviates from the neutral expectation in such a fashion that the total number of alleles exhibit an excess which mainly occurs through an excess of rare alleles. This presentation extends the above study demonstrating that hidden subdivision possibly are present in many national (cosmopolitan) human populations which can cause rare alleles to be observed in frequencies higher than the expectations of the mutation-drift model (neutral mutation hypothesis). This is shown first with the mitochondrial DNA (mtDNA) survey data from several oriental populations (Harihara et al. 1988), which implies that the effects of hidden subdivision may be detected even with data from a single locus, provided that it contains enough variability. Secondly it is shown that the summary observations from the three major cosmopolitan populations (Japanese, US Whites, and US Blacks), studied by Neel et al. (1988) and Mohrenweiser et al. (1987), also exhibit the same phenomenon when several isozyme loci are simultaneously considered. This indicates that the observations from the mtDNA data are not artifacts of stochastic errors of single-locus data. A recapitulation of Chakraborty et al.'s (1988) computations is presented to demonstrate that the relative excess of rare alleles present in a population, produced by the phenomenon of amalgamation, is determined by two factors: 1) the number of subpopulations hidden within the population studied; and 2) their genetic dissimilarities (average genetic distances among them). Therefore, the complexity of the population can be examined in terms of the observed level of excess number of alleles encountered in any given survey. Lastly, it is argued that since the excess mainly occurs through the excess of rare alleles, unless the above factor is critically taken into consideration, the indirect estimate of mutation rate per locus per generation may be overestimated from the frequencies of rare alleles, as proposed by Nei (1977), Neel & Rothman (1978), Chakraborty (1981), and others.

Genetic diversity at the mtDNA genome in some Oriental populations

The mitochondrial DNA (mtDNA) is particularly useful in evolutionary studies of ethnic origins of various human populations (e.g., Brown 1980, Denaro et al. 1981, Blanc et al. 1983, Johnson et al. 1983, Horai et al. 1987, Harihara et al. 1988) and in detecting DNA polymorphisms that existed before the geographic dispersal of the human species in the world (Cann et al. 1982, Cann & Wilson 1983, Cann et al. 1987). The mtDNA has a distinct advantage over nuclear DNA because the substantial sequence variation present in the mtDNA that can be detected by the restriction fragment length polymorphisms (RFLPs) is produced unequivocally by new mutations and no recombination is involved in the generation of the mtDNA-morphs that can be defined by using various restriction enzymes (Brown & Goodman 1979, Horai et al. 1986, Horai & Matsunaga 1986). The molecular sequence variation at the mtDNA genome has also provided evidence for founder effects in several human populations (Wallace et al. 1985).

The power of resolution of sequence variability at the mtDNA genome, however, varies substantially from laboratory to laboratory, depending upon the restriction enzymes used and detectability of fragment size differences. In comparing mtDNA-morph frequency differences among various populations, therefore, the uniformity of laboratory methods must be taken into account. For the present purpose, here we consider the data from a recent survey where mtDNA polymorphisms were detected using 13 restriction enzymes on the total DNA obtained from blood samples of five Asian populations; Japanese and Ainu of Northern Japan, Korean, Negrito (Aëta) of the Philippines, and Vedda of Sri Lanka (Harihara et al. 1988). In the total sample of 243 individuals 20 different mtDNA-morphs were detected from the combination of 28 different restriction enzyme morphs. Since the rate of nucleotide substitutions and the extent of nucleotide diversity at the mtDNA genome, by and large, follow the pattern of the predictions of the neutral mutation hypothesis (for a review see Nei 1987), it is interesting to ask whether the mtDNA-morph distributions observed in the survey of Harihara et al. (1988) are in accordance with the sampling theory of neutral mutations (Ewens 1972, Chakraborty & Griffiths 1982).

Estimation

In a survey of n genes from a steady-state population of effective size N_e, the expected number of alleles (morphs) that occur with r copies in the sample is given by

$$E(k_r) = \frac{\theta}{r} \cdot \frac{n!}{(n-r)!} \cdot \frac{\Gamma(n+\theta-r)}{\Gamma(n+\theta)} \tag{1}$$

for $r = 1, 2, \ldots$; where $\theta = 2N_{e(f)}v$, in which v is the mutation rate per generation (for mtDNA $v = m\mu$, where μ is the mutation rate per nucleotide site per generation, and m is the length of the mtDNA genome ≈ 16.5 kb in man) $N_{e(f)}$ is the effective female population size, and $\Gamma(.)$ is a Gamma function (Chakraborty & Griffiths 1982).

In addition, the expectation and variance of the total number of alleles (mtDNA-morphs) in a sample of size n are given by Ewens (1972)

$$E(k) = \theta \cdot \sum_{r=0}^{n-1} (\theta + r)^{-1} \tag{2}$$

and

$$V(k) = \theta \cdot \sum_{r=0}^{n-1} r / (\theta + r)^2 \tag{3}$$

in which θ is the same as defined in equation (1).

Obviously, evaluations of equations (1)–(3) require the knowledge of the composite parameter $\theta = 2N_{e(f)}v$, for which two alternatives are suggested in the literature. In the first, called the gene-diversity estimator, the observed mtDNA-morph frequency distribution $(k_r; r = 1, 2, \ldots)$ is used to generate an unbiased estimator of the function $\theta/(1+\theta)$, the expected gene-diversity in the population. Nei (1978) has shown that

$$\hat{H} = \frac{n}{n-1}\left[1 - \sum_{r=1}^{n} r^2 k_r / n^2\right] \tag{4}$$

is an unbiased estimator of $\theta/(1+\theta)$; i.e., $E(\hat{H}) = \theta/(1+\theta)$. Therefore, a candidate estimator of θ is given by the gene-diversity estimator

$$\hat{\theta}_{GS} = \hat{H}/(1-\hat{H}), \tag{5}$$

for which an approximate sampling variance formula is given by Chakraborty & Schwartz (1990) as

$$V(\hat{\theta}_{GS}) \approx \frac{2\theta(1+\theta)^2}{(2+\theta)(3+\theta)} - \frac{2(1+\theta)^3}{n} \tag{6}$$

Alternatively, θ can be estimated using the maximum likelihood (ML) method suggested by Ewens (1972), in which the ML-estimator of θ, denoted by $\hat{\theta}_{MLE}$, satisfies the equation

$$k = \hat{\theta}_{MLE} \cdot \sum_{r=0}^{n-1} (\hat{\theta}_{MLE} + r)^{-1} \tag{7}$$

whose sampling variance has the close form (see Chakraborty & Schwartz 1990 for a derivation)

$$V(\hat{\theta}_{MLE}) \approx \theta / [\sum_{r=0}^{n-1} r / (\theta + r)^2] \tag{8}$$

While the use of either of the above two alternative estimators of θ can be used to evaluate the expected number of alleles (mtDNA-morphs) with a specified number of copies in a sample (equation 1), from a pure statistical consideration one might be inclined to use the MLE, $\hat{\theta}_{MLE}$, because it is more efficient than $\hat{\theta}_{GS}$ (i.e., $\hat{\theta}_{MLE}$ has smaller sampling variance compared to $\hat{\theta}_{GS}$). However, as we shall see below, when the observed distribution of k_r is at discrepancy with the prediction of equation (1), $\hat{\theta}_{GS}$ is a more realistic estimator, because problems such as hidden subdivision affects $\hat{\theta}_{MLE}$ to deviate from the true value more substantially compared to $\hat{\theta}_{GS}$ (Chakraborty et al. 1988, Chakraborty & Schwartz 1990). It should also be stated that none of these estimators are unbiased estimator of θ, for which no formulation is available in the current literature.

For our purpose, we shall use both estimators ($\hat{\theta}_{GS}$ and $\hat{\theta}_{MLE}$) to demonstrate that certain features of the allele frequency distribution always detect hidden subdivision in a population irrespective of the choice of estimators.

Results

Table 1 shows the summary statistics of the mtDNA survey reported by Harihara et al. (1988) for each of the five Asian populations mentioned earlier, and for the pooled sample. In addition, this table also presents the two estimators of θ, ($\hat{\theta}_{MLE}$ from k through an iterative solution of equation 7, and $\hat{\theta}_{GS}$ from H using equation 5), along with their standard errors.

Two features of these estimates are notworthy. First, there is a direct positive association between the estimates of θ from k (i.e., $\hat{\theta}_{MLE}$) with the sample size, while this is not so for the estimate from H (i.e., $\hat{\theta}_{GS}$). This feature is parallel to the observations noted by Chakraborty & Schwartz (1990) in the context of analyzing the surname frequency distributions in England and Wales (Fox & Lasker 1983) and in Italy (Zei et al. 1983). This raises doubt as to whether the relative magnitude of the

Table 1. Summary statistics of mtDNA surveys from five Asian populations (adapted from Harihara et al. 1988).

Populations	n	k	$\hat{H}$	Estimate of $\theta = 2N_{e(f)}v$ from	
				$\hat{H}$ ($\hat{\theta}_{GS}$)	k ($\hat{\theta}_{MLE}$)
Japanese	74	11	0.400 ± 0.072	0.681 ± 0.202	3.341 ± 1.240
Ainu	48	6	0.231 ± 0.081	0.309 ± 0.136	1.588 ± 0.808
Korean	64	7	0.332 ± 0.075	0.509 ± 0.169	1.796 ± 0.832
Aëta	37	3	0.199 ± 0.086	0.257 ± 0.132	0.569 ± 0.434
Vedda	20	4	0.510 ± 0.104	1.159 ± 0.462	1.209 ± 0.808
Pooled	243	20	0.340 ± 0.040	0.517 ± 0.092	4.991 ± 1.307

n = Number of individuals sampled.
k = Observed number of different mtDNA-morphs in the sample.
$\hat{H}$ = Gene-diversity in the sample (computed by equation 4).

$\hat{\theta}_{MLE}$ values for these five populations truly reflect their differences of effective sizes, as it should, since the same mtDNA-genome has been examined in these surveys and hence the mutation rate component (v) should be the same, unless it varies across populations. **Second**, the estimate of θ from k for the pooled sample is much larger than those of the individual samples, while the pooled estimate of θ from $\hat{H}$ (pooled gene diversity in the entire sample) is within the range of the individual sample estimates. These two features suggest that the total number of mtDNA-morphs observed in the sample may not follow the prediction of equation (2), which is the premise of the maximum likelihood estimator of θ ($\hat{\theta}_{MLE}$, equation 5).

To check this assertion, we substituted the stimators of θ in equation (1) to get the expectation k_r for all values of r = 1, 2, ..., for each of the populations sampled by Harihara et al. (1988). These are shown in Tables 2 through 4 for the Japanese, the Ainu, and the Koreans, respectively; and in Table 5 for the pooled sample.

The standard errors of the estimates of k_r, shown in these tables are computed by the formula given in Chakraborty & Griffiths (1982). A comparison of the expected values of k_r with the observed frequencies show three features: 1) Within each population when θ is estimated from $\hat{H}$ (i.e., $\hat{\theta}_{GS}$), the observed mtDNA-morph distributions show excess of the total number of mtDNA-morphs compared to the neutral expectation, and this excess is mainly due to the excess of rare morphs (i.e., those occurring with few number of copies in the samples); 2) this phenomenon is more conspicuous in the pooled sample suggesting that perhaps the reason of the above observation is the fact that within each of the defined populations there is hidden subdivision; and 3) θ is estimated from the total number of observed mtDNA-morphs in the sample (i.e., $\hat{\theta}_{MLE}$), although the expectation and observed for the total number of morphs agree with each other (as it should, because of equation 7), there are excesses of the rare morphs, compensated by deficiencies in the number of common morphs in the sample, and this phenomenon is more prominent in the pooled sample. In other words, irrespective of the estimator used for θ, these data exhibit a departure from the predictions of the neutral allele theory, suggesting

Table 2. mtDNA-morph profile in the Japanese population (based on data of Harihara et al. 1988).

Number of copies	Frequencies		
		Expected based on θ from	
	Obs.	$\hat{H}$ ($\hat{\theta}_{GS}$)	k ($\hat{\theta}_{MLE}$)
1	6	0.684 ± 0.827	3.238 ± 1.798
2	2	0.343 ± 0.586	1.569 ± 1.251
3	1	0.230 ± 0.480	1.013 ± 1.004
4	1	0.173 ± 0.416	0.736 ± 0.855
57*	1	2.370 ± 1.556	4.444 ± 2.206
Total	11	3.800 ± 2.434	11.000 ± 2.693

* The expected for this category represents frequencies for 5 or more copies.

Table 3. mtDNA-morph profile in the Ainu population (based on data of Harihara et al. 1988).

Number of copies	Frequencies		
		Expected based on θ from	
	Obs.	$\hat{H}$ ($\hat{\theta}_{GS}$)	k ($\hat{\theta}_{MLE}$)
1	4	0.313 ± 0.560	1.569 ± 1.252
2	1	0.159 ± 0.399	0.775 ± 0.880
42*	1	1.770 ± 1.346	3.656 ± 2.013
Total	6	2.243 ± 1.067	6.000 ± 1.965

* The expected for this category represents frequencies for 3 or more copies.

Table 4. mtDNA-morph profile in the Korean population (based on data of Harihara et al. 1988).

Number of copies	Frequencies		
		Expected based on θ from	
	Obs.	$\hat{H}$ ($\hat{\theta}_{GS}$)	k ($\hat{\theta}_{MLE}$)
1	3	0.513 ± 0.716	1.774 ± 1.332
2	1	0.259 ± 0.508	0.876 ± 0.936
3	1	0.174 ± 0.416	0.576 ± 0.759
4	1	0.131 ± 0.362	0.427 ± 0.653
52*	1	2.018 ± 1.542	3.346 ± 1.937
Total	7	3.094 ± 1.363	7.000 ± 2.160

* The expected for this category represents frequencies for 5 or more copies.

Table 5. mtDNA-morph profile in the pooled Asian population (based on data of Harihara et al. 1988).

Number of copies	Frequencies		
		Expected based on θ from	
	Obs.	$\hat{H}$ ($\hat{\theta}_{GS}$)	k ($\hat{\theta}_{MLE}$)
1	8	0.519 ± 0.720	4.910 ± 2.216
2	6	0.260 ± 0.510	2.415 ± 1.554
3	1	0.174 ± 0.417	1.584 ± 1.258
4	1	0.130 ± 0.361	1.169 ± 1.080
5	2	0.105 ± 0.323	0.920 ± 0.958
9	1	0.059 ± 0.241	0.478 ± 0.690
197*	1	2.570 ± 1.806	8.525 ± 2.995
Total	20	3.815 ± 1.603	20.000 ± 3.818

* The expected for this category represents frequencies for 10 or more copies.

that the mtDNA-morph distributions in these Asian populations show excess of rare morphs, indicative of the presence of hidden subdivisions within each of the populations studied.

As a consequence of these results, if one uses $\hat{\theta}_{MLE}$ as a valid estimator for θ in such surveys, we further note that the expected gene diversity, $\hat{\theta}_{MLE}/(1 + \hat{\theta}_{MLE})$, becomes much larger than the observed. This is shown in Table 6 for each population sample, and for the pooled data; suggesting that if internal subdivision is the cause of the above-noted discrepancy, $\hat{\theta}_{MLE}$ may not be an appropriate estimator of θ. On the contrary, there are some circumstances under which even in the presence of hidden subdivision, the estimator $\hat{\theta}_{GS}$ (from observed gene diversity) may not be in too much error, as argued in Chakraborty et al. (1988) and Chakraborty & Schwartz (1990).

Table 6. Gene-diversity for the mtDNA genome for five Asian populations (from data of Harihara et al. 1988).

Population	n	Heterozygosity (H)	
		Obs. ± s.e.	Exp. (from $\hat{\theta}_{MLE}$)
Japanese	74	0.400 ± 0.072	0.770
Ainu	48	0.231 ± 0.081	0.609
Korean	64	0.332 ± 0.075	0.642
Aēta	37	0.199 ± 0.086	0.363
Vedda	20	0.510 ± 0.104	0.547
Pooled	243	0.340 ± 0.040	0.833

Allele frequency profile at isozyme loci in three major races of man

While the above section demonstrates that the presence of hidden subdivision may be detected by studying data from single locus, one criticism of such an analysis is that the theory used in this context is known to have a large sampling variance, due to stochastic error accumulated during evolution (Li & Nei 1975, Nei 1978). Therefore, data from a single locus may easily cause deviation due to this artifact. In this section, therefore, we show that the observations noted above from the analysis of mtDNA data are not due to stochastic errors alone, which can be substantially reduced if a large number of loci are used together to perform similar analysis. This is done here using the isozyme surveys reported by Neel et al. (1988) and Mohrenweiser et al. (1987) who used uniform laboratory methods to detect isozyme variations in three cosmopolitan populations: US Whites and US Blacks from a survey of cord blood samples from new borns in Ann Arbor, Michigan; and Japanese from Hiroshima and Nagasaki, studied to examine the effect of radiation exposure during the atom-bomb exposure. Table 7 presents a summary of their findings, pertinent details of which can be found in the original reports.

Table 7. Summary statistics from isozyme surveys in the three major racial groups of humans (adapted from Neel et al. 1988, Mohrenweiser et al. 1987).

Statistics	Japanese	US White	US Black
No. of loci surveyed	32	51	51
No. of gene sampled/locus	29,272	4,435	374
Av. heterozygosity/locus (in %)	8.699	5.011	5.230
Av. no. of alleles/locus	5.531	2.608	1.667
Av. no. of singleton alleles/locus	2.031	0.843	0.196
Prop. of variant loci	0.875	0.667	0.471

From these summary statistics, we can use equations (5) and (7) to obtain the two alternative estimators of θ, and derive expectations for the number of alleles per locus for each specific allele frequency classes, substituting such parameter estimates in equation (1). Figs. 1a, 1b, and 1c and Table 8 show the contrast of the observed allele frequency distributions for all loci pooled together in these three populations separately. For brevity, we show only the predictions based on the estimator $\hat{\theta}_{GS}$, since the qualitative results are similar for the other estimator ($\hat{\theta}_{MLE}$) as well. These figures and Table 8 clearly show that within each of the three major racial groups of man, there is a conspicuous excess of total number of alleles compared to the neutral expectation, and the excess is mainly due an increase in the number of rare alleles.

Fig. 1. Observed (blank bars) and expected (black bars) allele frequency distributions from isozyme data in the three major racial groups of man. Panel (a) = Japanese; Panel (b) = US White; and Panel (c) = US Black. The raw data are given in Mohrenweiser et al. (1987) and Neel et al. (1988). The expectations are based on the parameter estimate $\hat{\theta}_{GS}$, and employs the theory of Chakraborty & Griffiths (1982).

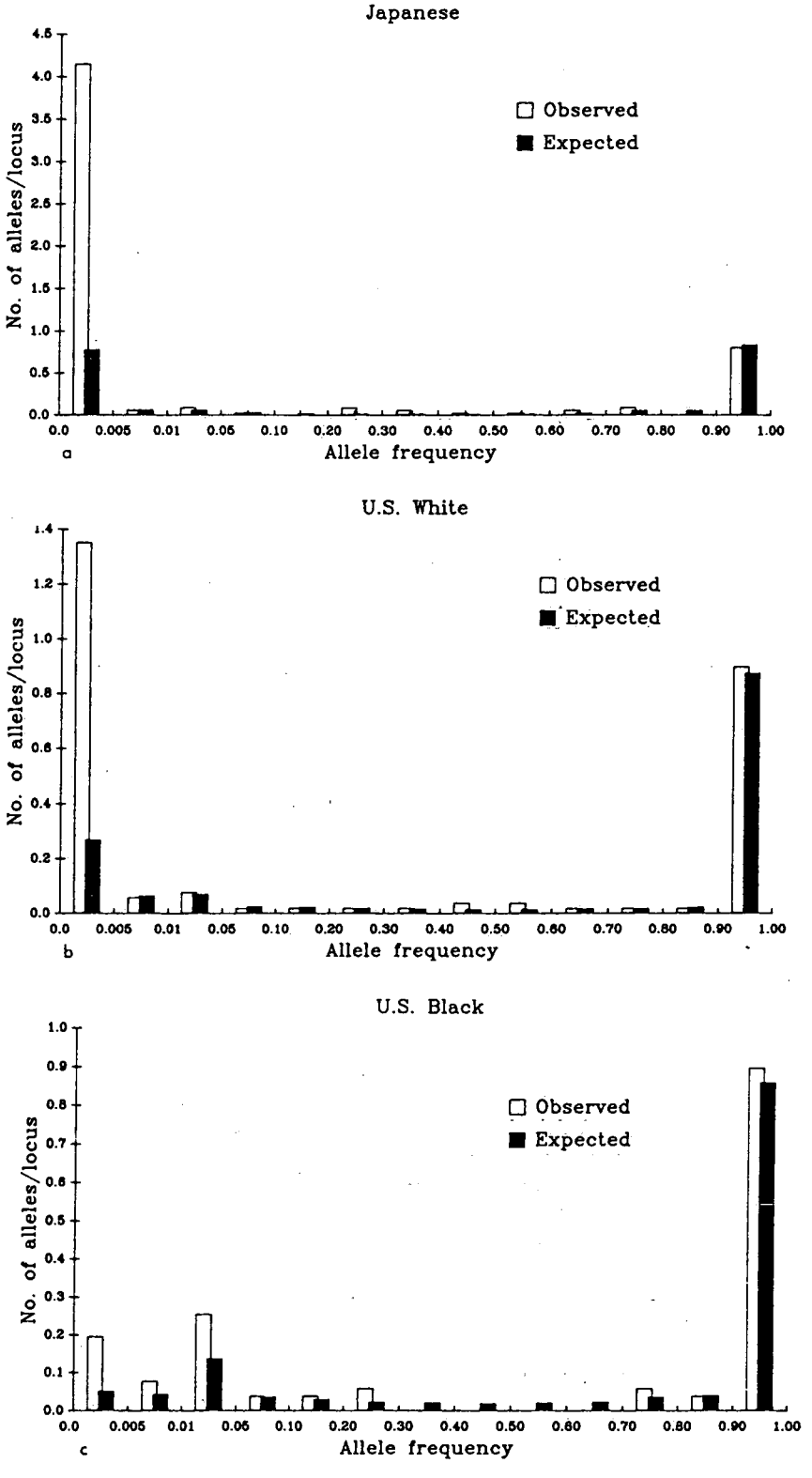

Japanese
Observed
Expected
No. of alleles/locus
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0 0.005 0.01 0.05 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
Allele frequency
a
U.S. White
Observed
Expected
No. of alleles/locus
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.005 0.01 0.05 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
Allele frequency
b
U.S. Black
Observed
Expected
No. of alleles/locus
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0 0.005 0.01 0.05 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
Allele frequency
c

Table 8. Allele frequency distributions per locus from isozyme data in the three major racial groups of man.

Allele frequency class	Japanese		US White		US Black	
	Obs.	Exp.*	Obs.	Exp.*	Obs.	Exp.*
0 -0.005	4.156	0.783	1.353	0.270	0.196	0.052
0.005-0.01	0.063	0.068	0.059	0.065	0.078	0.044
0.01 -0.05	0.094	0.063	0.078	0.072	0.255	0.138
0.05 -0.10	0.031	0.032	0.020	0.026	0.039	0.037
0.10 -0.20	0.0	0.019	0.020	0.024	0.039	0.030
0.20 -0.30	0.094	0.018	0.020	0.020	0.059	0.024
0.30 -0.40	0.063	0.018	0.020	0.018	0.0	0.022
0.40 -0.50	0.031	0.016	0.039	0.016	0.0	0.020
0.50 -0.60	0.031	0.016	0.039	0.016	0.0	0.022
0.60 -0.70	0.063	0.026	0.020	0.019	0.0	0.024
0.70 -0.80	0.094	0.058	0.020	0.020	0.059	0.037
0.80 -0.90	0.0	0.060	0.020	0.024	0.039	0.041
0.90 -1.00	0.813	0.843	0.902	0.879	0.902	0.864
Total	5.531	2.021	2.608	1.469	1.667	1.354

* The expected number of alleles in a gene frequency class is computed by the theory of Chakraborty & Griffiths (1982), using the estimate $\hat{\theta}_{GS}$ from the average gene-diversity. All expectations refer to the appropriate sample size of the relevant surveys (see Table 7).

The distributions shown in these three figures suggest that such excess can be detected even by a graphic display of the data. An objective test of significance of such departures (not attempted here) can be done using a result from the theory of Chakraborty & Griffiths (1982), who showed that the number of rare alleles (e.g., alleles with gene frequency in the range below 1 % or 5 %) approximately follows a Poisson distribution, and hence appropriate exact tests may be conducted using the expected frequency of such alleles as the estimated parameter of the Poisson distribution, with sample size being the number of loci involved in such calculations (e.g., the significance of the departure of the observed frequency of 4.156 alleles with gene frequency below 0.5 % per locus in the Japanese population compared to its expectation of 0.783 can be tested with a Poisson test assuming the parameter value of 0.783 and sample size 32, the number of loci).

In summary, these data show that the observed isozyme allele frequency profile in each of the three major races of man depart from the neutral predictions, exhibiting an excess of rare alleles, which is the main contributor of the apparent excess of total number of alleles per locus. In conjunction to this we also note that the proportion of variant loci (i.e., the loci with at least one variant allele observed) is also in excess of the expectation. These summary observations are shown in Table 9, particularly using the gene-diversity estimator, $\hat{\theta}_{GS}$.

As in the case of mtDNA data, we also note that when $\hat{\theta}_{MLE}$ is used as an estimator of θ, the predicted level of heterozygosity ($\hat{H}$) becomes too large compared to the observed in each of these three populations. This is shown in Table 10.

All of the findings of the single-locus analysis using mtDNA-morph distributions in the five Asian populations are comparable with the isozyme surveys in the major racial groups of man. In other words, the conclusions derived from the mtDNA survey are not due to an artifact of stochastic errors associated with single-locus data.

Table 9. Summary statistics of departures from the neutral model in the isozyme variation in three racial groups of man.

Statistics	Observed	Expectd ± s.e.
Japanese ($\hat{\theta}_{GS} = 0.095$; n = 29,272)		
Total no. of alleles/locus	5.531	2.021 ± 0.177
No. of singleton alleles/locus	2.031	0.095 ± 0.055
Proportion of variant loci	0.875	0.625 ± 0.086
US White ($\hat{\theta}_{GS} = 0.053$; n = 4,435)		
Total no. of alleles/locus	2.608	1.469 ± 0.095
No. of singleton alleles/locus	0.843	0.053 ± 0.032
Proportion of variant loci	0.667	0.358 ± 0.067
US Black ($\hat{\theta}_{GS} = 0.055$; n = 374)		
Total no. of alleles/locus	1.667	1.354 ± 0.083
No. of singleton alleles/locus	0.196	0.052 ± 0.032
Proportion of variant loci	0.471	0.279 ± 0.063

Table 10. Observed and expected gene-diversity from the MLE of θ in the three major racial groups of man (from data of Mohrenweiser et al. 1987 and Neel et al. 1988).

Populations	n	k	$\hat{\theta}_{MLE}$	Gene-diversity	
				Obs.	Exp.
Japanese	29,272	5.531	0.440	0.087 ± 0.026	0.305
US White	4,435	2.608	0.185	0.050 ± 0.018	0.156
US Black	374	1.667	0.105	0.052 ± 0.015	0.095

Cause and implication of excess of rare alleles in cosmopolitan populations

Having shown that the excess of rare alleles is perhaps a general rule in many cosmopolitan populations, several causal mechanisms may be proposed which may explain this phenomenon, indicated in the introductory section of this presentation. We argue that hidden substructuring could be a major factor, although it cannot be clearly shown from these data alone. However, if we examine the above observations in the light of the natural experiment conducted in Chakraborty et al. (1988), we can easily argue that none of the populations examined here are probably single breeding

units within themselves, and hence the observed departure is an effect of hidden subdivision. The questions, therefore, are: How many subpopulations are necessary to explain the observed discrepancy, and what are the implications of such discrepancy. The first question cannot be unequivocally answered from the above data alone, because there are two factors of hidden subdivision which pertains to the above-noted discrepancy, shown in Chakraborty et al. (1988). To re-capitulate that analysis, this earlier study showed that within each of the individual South and Central American Indian tribes, the allele frequency profile follows the prediction of the neutral allele theory suggesting that the genetic variations of the isozymes used in the major racial group analysis here may indeed be selectively neutral. However, when the data from 12 tribes are pooled to compose a single hypothetical Amerindian population, the allele frequency profile in such a composite population departs from the expectations of the neutral model, all features of which parallel to the present observations. This is graphically shown in Fig. 2, where the expected allele frequency distributions (denoted by black bars) are derived based on the gene-diversity estimator $\hat{\theta}_{GS}$.

Chakraborty et al. (1988) further showed that the complexity of hidden subdivision is composed of two factors: 1) The number of subpopulations involved, and 2) the average degree of genetic differentiation among them. A larger number of subpopulations with small degree of genetic divergence among them may have effect similar to a small number of subpopulations with large genetic distances among them. This result implies that unless there is any other independent information regarding the extent of subdivision, the number of subpopulations cannot be esti-

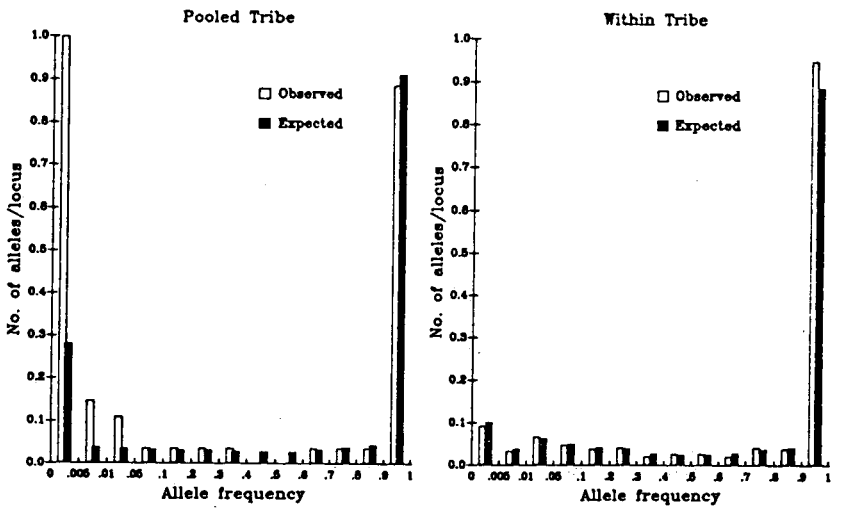


Fig. 2. Observed (blank bars) and expected (black bars) allele frequency distributions from isozyme data in 12 Amerindian tribes from South and Central America. The pooled distribution is from the allele frequency distribution in the total population (pooled over 12 tribes), and the within tribe distribution is the average of the individual tribe-specific distributions. The raw data are given in Chakraborty et al. (1988).

mated precisely. However, some knowledge of the history of a population may provide information regarding the time depth of isolation of subpopulations within it. If such data are available, one might suggest a reasonable value of effective genetic distance among the constituent subpopulations, from which an estimate of the necessary number of subpopulations may be made to explain the observed excess of allele numbers.

This brings us to the second question, namely the implication of the observed excess of allele numbers in human populations. Note that a translation of the time depth of isolation to a genetic distance (see Nei 1975, 1987) is possible only when the mutation rate estimates are available. The current indirect estimates of mutation rate from isozyme studies are all based on theories that depend upon the assertion that the observed number of rare (or total) alleles is in accordance with the neutral expectations. This is where the hidden subdivision effect is most prominent. Therefore, if hidden subdivision is prevalent in most cosmpolitan populations, the indirect estimates of mutation rate may be grossly overestimated. Note that in the most recent attempt to provide an indirect estimate of mutation rate from human isozyme data, Chakraborty & Neel (1989) used the Amerindian populations that within themselves follow the neutral allele frequency profile, and these estimates are much closer to the direct estimates (although still differ by a factor of two). Comparison of mutation rate estimates in different populations, therefore, should be made with caution, because the extent of hidden subdivisions may be substantially different among them.

Conclusion and discussion

The main conclusion of the above analyses of the mtDNA and isozyme data is that there is an apparent excess of the total number of alleles or mtDNA-morphs in all populations analyzed. This excess is probably due to hidden subdivision within each population, since this observation is parallel to what is seen in the Amerindian study (Chakraborty et al. 1988) where such a discrepancy could be created artificially by amalgamating tribes with different degrees of genetic differentiation among them.

Comparing the results of Tables 2 and 6, it is reassuring that the excess of total number of variants in the Japanese population is quite consistent between the mtDNA survey (2.89-fold increase in relation to the neutral expectation) and in the isozyme data (2.74-fold increase). This is somewhat larger than the excess seen in the Amerindian study where 12 tribes have been amalgamated (resulting in a 1.6-fold increase). Although these numbers should not be interpreted at their face value, since the extent of excess is a complex function of average degree of polymorphism within a population, number of constituent subpopulations, and their genetic isolation. However, considering the three populations of the present isozyme data analysis, we see that the Japanese population exhibit a higher extent of the excess of total number of alleles. Even though the Japanese isozyme survey (Neel et al. 1988) involves a smaller set of loci (32) than the US Whites and US Blacks (Mohrenweiser et al. 1987; where 51 loci were scored), and the average gene-diversity in the Japanese population is somewhat larger (8.7 %) in comparison with the US Whites or Blacks (approximately 5 %), the 2.7-fold excess in the Japanese population as compared to the 1.8-fold excess in the US Whites, or 1.3-fold excess in the US Blacks is reasonable,

since the US Blacks represent an admixed population formed by slave-trading from a smaller region of Africa (mostly West Africa; see Adams & Ward 1973) and has the history of only five centuries of gene-admixture with the Caucasians. The US white population probably consists of substructuring due to their different European ancestry, where the genetic differences can be substantially smaller than the probable constituent Japanese subpopulations.

These deductions are of course tentative, because a firm conclusion in this regard should require an adjustment for sample size differences between the surveys, eliminate any possible bias due to differences of the loci employed in the analysis, and finally must adjust for probable differences of genetic distances among the constituent subpopulations. This issue will be considered in a greater details elsewhere (Chakraborty et al., in prep.).

As mentioned earlier, one important implication of this demonstration is that the excess of allele numbers in comparison to the neutral expectation is dependent on the internal genetic structure of the population. Therefore, when an indirect estimate of mutation rate is based on the number of rare alleles, any inter-survey comparison of mutation rate estimates from different populations should be made with caution. While in the past there had been some suggestions for the possibility of primitive populations exhibiting larger mutation rates than the modern cosmopolitan populations, in view of the present observation this postulation cannot be persued any longer vigorously (see Chakraborty & Neel 1989 for a different reason for this).

Another implication of the present set of observations relates to the comparability of molecular variation at the DNA level, detected by restriction fragment length polymorphisms (RFLPs) with that at the isozyme loci detected by electrophoresis. It is true that electrophoresis detects only a fraction (between one-third to one-fourth) of molecular variation because of the fact that nucleotide changes that do not change the charge of a protein molecule is not detected by electrophoresis. In contrast, as long as the recognition sequence is altered, such changes will be detected by the RFLP-technique. However, at present population data on RFLPs exist for populations that are very loosely defined, leaving enough scope of internal hidden subdivisions within them. Therefore, from the greater number of variant RFLP alleles (or haplotypes) alone one cannot judge how enhanced is the degree of detection of genetic variation by this improved technique. The unit of population should be precise for both technologies, should a more precise estimate of detectability of molecular polymorphism be made, even when we adjust for hidden variation in both techniques.

The analyses presented here also reflect that even if the populations of slightly deleterious mutations and population bottleneck do not apply to a particular population, there may be a more realistic factor (namely, hidden subdivision) that could cause specific departures of the allele frequency profile from the prediction of the neutral mutation hypothesis. Hidden subdivision is a factor that can be more incisively examined. For example, enlarged sampling from different geographic regions within a population may reveal the extent of subdivision. Such geographic microdifferentiation has been demonstrated with allele frequency data at specific loci in Japan (e.g., Nei & Imaizumi 1966a, b), or by population structure analysis employing Wright's F-statistics (e.g., Neel & Ward 1972). Therefore, should hidden subdivision be the only cause of departure of the allele frequency profile from the neutral expectations, this can be easily tested. Such validations of the bottleneck hypothesis or slightly deleterious mutation model are not readily available.

Finally, it should be noted that while this presentation emphatically argues for the more thorough examination of the possibility that many cosmopolitan populations may indeed have internal hidden substructuring, causing its allele frequency profile to depart substantially from the neutral expectations, this does not negate the existence of other probable causes, which might simultaneously affect a population during the course of its evolution. Nevertheless, since it is a testable proposition, hidden subdivision may be detected and validated. Even after this is done, if it cannot account for all the departures, other hypotheses may be entertained for understanding the evolutionary mechanisms that may generate the observed allele frequency structure of the population.

Acknowledgements

This work was supported by US Public Health Service grant GM 41399 from the US National Institutes of Health. I thank Drs. J.V. Neel and P.E. Smouse for their collaborations for some of the findings cited here, and Drs. M. Nei and W.J. Schull for their suggestions and comments during the conduct of this work.

References

Adams, J. & Ward, R.H., 1973: Admixture studies and detection of selection. - Science 180, 1137–1143.

Ayala, F.J., 1976: Molecular Evolution. - Sinauer, Sunderland.

Blanc, H., Chen, K.H., D'Amore, M.A. & Wallace, D.C., 1983: Amino acid change associated with the major polymorphic *Hinc*-II site of Oriental and Caucasian mitochondrial DNAs. - Amer. J. Hum. Genet. 35, 167–176.

Brown, W.M., 1980: Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. - Proc. Natl. Acad. Sci. USA 77, 3605–3609.

Brown, W.M. & Goodman, H.M., 1979: Quantitation of intra-population variation by restriction endonuclease analysis of human mitochondrial DNA. - In: Cummings, D., Borst, P., Dawid, I., Weissman, S. & Fox, C.F. (eds.): Extrachromosomal DNA, ICN-UCLA Symposia, 485–500. - Academic Press, New York.

Bruce, E.J. & Ayala, F.J., 1979: Phylogenetic relationships between man and the apes: Electrophoretic evidence. - Evolution 33, 1040–1056.

Cann, R.L., Brown, W.M. & Wilson, A.C., 1982: Evolution of human mitochondrial DNA: A preliminary report. - In: Bonne-Tamir, B., Cohen, P. & Goodman, R.N. (eds.): Human Genetics - Part A: The Unfolding Genome, 157–165. - Alan R. Liss, New York.

Cann, R.L., Stoneking, M. & Wilson, A.C., 1987: Mitochondrial DNA and human evolution. - Nature 325, 31–36.

Cann, R.L. & Wilson, A.C., 1983: Length mutations in human mitochondrial DNA. - Genetics 104, 699–711.

Chakraborty, R., 1981: Expected number of rare alleles per locus in a sample and estimation of mutation rates. - Amer. J. Hum. Genet. 33, 481–483.

Chakraborty, R., Fuerst, P.A. & Nei, M., 1980: Statistical studies on protein polymorphism in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. - Genetics 94, 1039–1063.

Chakraborty, R. & Griffiths, R.C., 1982: Correlation of heterozygosity and number of alleles in different frequency classes. - Theor. Pop. Biol. 21, 205–218.

Chakraborty, R. & Neel, J.V., 1989: Description and validation of a method for simultaneous estimation of effective population size and mutation rate from human population data. - Proc. Natl. Acad. Sci. USA 86, 9407-9411.

Chakraborty, R. & Nei, M., 1977: Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. - Evolution 31, 347-356.

Chakraborty, R. & Schwartz, R.J., 1990: Selective neutrality of surname distribution in an immigrant Indian community of Houston, Texas. - Amer. J. Hum. Biol. 2, 1-15.

Chakraborty, R., Smouse, P.E. & Neel, J.V., 1988: Population amalgamation and genetic variation: Observations on artificially agglomerated tribal populations of Central and South America. - Amer. J. Hum. Genet. 43, 709-725.

Denaro, M., Blanc, H., Johnson, M.J., Chen, K.H., Wilmsen, E., Cavalli-Sforza, L.L. & Wallace, D.C., 1981: Ethnic variation in *Hpa*-I endonuclease cleavage patterns of human mitochondrial DNA. - Proc. Natl. Acad. Sci. USA 78, 5768-5772.

Ewens, W.J., 1972: The sampling theory of selectively neutral alleles. - Theor. Pop. Biol. 3, 87-112.

Fox, W.R. & Lasker, G.W., 1983: The distribution of surname frequencies. - Int. Stat. Rev. 51, 81-87.

Goodnight, C.J., 1987: On the effect of founder events on epistatic genetic variance. - Evolution 41, 80-91.

Harihara, S., Saitou, N., Hirai, M., Gojobori, T., Park, K.S., Misawa, S., Ellepola, S.B., Ishida, T. & Omoto, K., 1988: Mitochondrial DNA polymorphism among five Asian populations. - Amer. J. Hum. Genet. 43, 134-143.

Horai, S., Gojobori, T. & Matsunaga, E., 1986: Distinct clustering of mitochondrial DNA types among Japanese, Caucasians and Negroes. - Jap. J. Genet. 61, 271-275.

——— 1987: Evolutionary implications of mitochondrial DNA polymorphism in human populations. - Hum. Genet. 74, 177-181.

Horai, S. & Matsunaga, E., 1986: Mitochondrial DNA polymorphism in Japanese. II. Analysis with restriction enzymes of four or five base pair reconition. - Human Genet. 72, 105-117.

Johnson, M.J., Wallace, D.C., Ferris, S.D., Rattazzi, M.C. & Cavalli-Sforza, L.L., 1983: Radiation of human mitochondrial DNA types analyzed by restriction endonuclease cleavage patterns. - J. Mol. Evol. 19, 255-271.

Kimura, M., 1983: The Neutral Theory of Evolution. - Cambridge University Press, Cambridge.

King, M.C. & Wilson, A.C., 1975: Evolution at two levels: Molecular similarities and biological differences between humans and chimpanzees. - Science 188, 107-116.

Landsteiner, K. & Levine, P., 1928: On the inheritance of agglutinogens of human blood demonstrable by immune agglutinins. - J. Exp. Med. 48, 731-749. (Reprinted in: Schull, W.J. & Chakraborty, R. (eds.): Human Genetics - A Selection of Insights, 39-57. Dowden, Hutchinson & Ross, Stroudsburg, PA.)

Lewontin, R.C., 1974: The Genetic Basis of Evolutionary Change. - Columbia University Press, New York.

Lewontin, R.C. & Cockerham, C.C., 1959: The goodness-of-fit test for detecting natural selection in random mating populations. - Evolution 13, 561-564.

Li, W.-H., 1978: Maintenance of genetic variability under the joint effect of mutation, selection and random drift. - Genetics 90, 349-382.

Li, W.-H. & Nei, M., 1975: Drift variances of heterozygosity and genetic distance in transient states. - Genet. Res. 25, 229-248.

Maruyama, T. & Fuerst, P.A., 1984: Population bottlenecks and non-equilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. - Genetics 108, 745-763.

—— 1985: Population bottlenecks and non-equilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. - Genetics 111, 691-703.

Milkman, R., 1973: Electrophoretic variation in *Escherichia coli* from natural sources. - Science 182, 1024-1026.

Mohrenweiser, H.W., Wurzinger, K.H. & Neel, J.V., 1987: Frequency and distribution of rare electrophoretic mobility variants in a population of newborns in Ann Arbor, Michigan. - Ann. Hum. Genet. 51, 303-316.

Mourant, A.E., Kopec, A.C. & Domaniewska-Sobczak, K., 1976: The Distribution of the Human Blood Groups and Other Polymorphisms, 2nd ed. - Oxford University Press, New York.

Neel, J.V. & Rothman, E.D., 1978: Indirect estimates of mutation rates in tribal Amerindians. - Proc. Natl. Acad. Sci. USA 75, 5585-5588.

Neel, J.V., Satoh, C., Smouse, P.E., Asakawa, J., Takahashi, N., Goriki, K., Fujita, M., Kageoka, T. & Hazama, R., 1988: Protein variants in Hiroshima and Nagasaki: Tales of two cities. - Amer. J. Hum. Genet. 43, 870-893.

Neel, J.V. & Ward, R.H., 1972: The genetic structure of a tribal population, the Yanomama Indians. VI. Analysis by F-statistics (including a comparison with the Makiritare and Xavante). - Genetics 72, 639-666.

Nei, M., 1975: Molecular Population Genetics and Evolution. - North-Holland - American Elsevier, New York.

— 1977: Estimation of mutation rate from rare protein variants. - Amer. J. Hum. Genet. 29, 225-232.

— 1978: Estimation of average heterozygosity and genetic distance from a small number of individuals. - Genetics 89, 583-590.

— 1987: Molecular Evolutionary Genetics. - Columbia University Press, New York.

Nei, M. & Graur, D., 1984: Extent of protein polymorphism and the neutral mutation theory. - Evol. Biol. 17, 73-118.

Nei, M. & Imaizumi, Y., 1966a: Genetic structure of human populations. I. Local differentiation of blood group gene frequencies in Japan. - Heredity 21, 9-25.

— — 1966b: Genetic structure of human populations. II. Differentiation of blood gruop gene frequencies among isolated populations. - Heredity 21, 183-190.

Nei, M., Maruyama, T. & Chakraborty, R., 1975: The bottleneck effect and genetic variability in populations. - Evolution 29, 1-10.

Ohta, T., 1973: Slightly deleterious mutant substitutions in evolution. - Nature 246, 98-98.

— 1976: Role of very slightly deleterious mutations in molecular evolution and polymorphism. - Theor. Pop. Biol. 10, 254-275.

Roychoudhury, A.K. & Nei, M., 1987: Human Polymorphic Genes: World Distribution. - Oxford University Press, New York.

Wallace, D.C., Garrison, K. & Knowler, W.C., 1985: Dramatic founder effects in Amerindian mitochondrial DNAs. - Amer. J. Phys. Anthrop. 68, 149-155.

Watterson, G.A., 1984: Allele frequencies after a bottleneck. - Theor. Pop. Biol. 26, 387-407.

Zei, G., Matessi, R.G., Siri, E., Moroni, A. & Cavalli-Sforza, L.L., 1983: Surnames in Sardinia. I. Fit of frequency distributions for neutral alleles and genetic population structure. - Ann. Hum. Genet. 47, 329-352.

Received November 9, 1989

Address for correspondence:
Prof. Dr. R. Chakraborty, Center for Demographic and Population Genetics, The University of Texas, Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, Texas 77225, USA.

Am. J. Hum. Genet. 47:87–94, 1990

Mitochondrial DNA Polymorphism Reveals Hidden Heterogeneity within Some Asian Populations

Ranajit Chakraborty

Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston

Summary

Use of data on mtDNA morph distributions from six Asian populations has shown that the observed number of different mtDNA morphs is too large when compared with the number expected on the basis of the observed gene diversity in the mtDNA genome. This excess number of morphs mainly occurs through an excess of rare morphs, and this discrepancy is more pronounced in a pooled sample of five Asian populations. It is suggested that this discrepancy is probably due to internal heterogeneity in each of the anthropologically defined populations. This analysis demonstrates the utility that population data for a single locus, such as the mtDNA genome, have for detecting hidden heterogeneity in populations, provided that the locus has substantial genetic variability, so that many variant alleles can be detected.

Introduction

Mitochondrial DNA (mtDNA) is particularly useful in evolutionary studies of the ethnic origins of human populations (e.g., see Brown 1980; Denaro et al. 1981; Blanc et al. 1983; Johnson et al. 1983; Horai et al. 1987; Harihara et al. 1988) and in detecting DNA polymorphisms that existed before the geographic dispersal of the human species (Cann et al. 1982, 1987; Cann and Wilson 1983). mtDNA has a distinct advantage over nuclear DNA for population genetic studies because (1) the evolutionary rate of nucleotide substitutions appears to be larger in the mtDNA genome compared with the nuclear genes (e.g., see Nei 1987), (2) the determination of the various mtDNA morphs (haplotypes) is unequivocal from population data, since mtDNA is maternally inherited, and (3) the generation of different mtDNA morphs can only occur through new mutations, and no recombination has to be invoked in studying the maintenance of mtDNA polymorphisms.

In a recent study, Whittam et al. (1986) analyzed allelic variations in 145 human mtDNAs representing samples from five geographic regions. They concluded that while the allele frequency distributions at different loci in the mtDNA genome follow the general predictions of the equilibrium theory of a mutation-drift model of selectively neutral mutations, certain deviations (e.g., observed gene diversity lower than that expected and excesses in the frequencies of common alleles and in the number of singleton alleles) can be attributed to possible bottleneck effect during recent human evolution and to the action of purifying selection. Since such observations can also be explained by hidden substructuring of populations, as evidenced in the study of electrophoretic variations in South and Central American Indians (Chakraborty et al. 1988), the purpose of the present paper is to demonstrate that the substantial variation of the mtDNA genome can be used to reveal hidden heterogeneity within anthropologically defined populations. This is shown by examining the mtDNA-morph distributions in several Asian populations, studied by Brega et al. (1986) and Harihara et al. (1988), and by utilizing the sampling theory of selectively neutral alleles (Ewens 1972; Chakraborty and Griffiths 1982). It is suggested that the discrepancies between the observed and expected distributions of the mtDNA morphs in most Asian populations are probably due to their internal hidden heterogeneity, and this conclusion probably applies to the populations examined by Whittam et al. (1986) as well.

Received December 19, 1989; revision received February 22, 1990.

Address for correspondence and reprints: Dr. Ranajit Chakraborty, Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, TX 77225.

0002-9297/90/4701-0012$02.00

Material and Methods

Data

It is well known that the power of resolution of molecular variability in the mtDNA genome varies in RFLP studies, depending on the restriction enzymes used and the detectability of fragment size differences. Therefore, uniformity of laboratory methods must be established in comparing the mtDNA-morph distributions among different populations. Recently, Harihara et al. (1988) published mtDNA-morph distributions in five Asian populations—the Japanese and the Ainu from northern Japan, the Koreans, the Negrito (Aeta) from the Philippines, and the Veddas of Sri Lanka—by using 13 restriction enzymes and following uniform laboratory conditions. Brega et al. (1986) used six enzymes from the above set to survey the mtDNA-morph distribution in the Tharu population of Nepal. These data form the basis of the present analysis. In the study of five Asian populations Harihara et al. (1988) observed 20 different mtDNA morphs in a total sample of 243 individuals, whereas in 91 individuals from the Tharu population of Nepal Brega et al. (1986) observed 13 mtDNA morphs (haplotypes).

Theory

Since the rate of nucleotide substitutions and the extent of nucleotide diversity in the mtDNA genome roughly follow the predictions of the neutral mutation hypothesis (for a review, see Nei 1987), I ask whether the various aspects of the mtDNA distributions in these six populations are consistent with the expectations from the sampling theory of selectively neutral mutations (Ewens 1972; Chakraborty and Griffiths 1982), which are based on the assumption that the sampling has occurred from a single homogeneous population in each case. This is accomplished by examining the expectations of two summary statistics of the mtDNA-morph distributions—gene diversity (H) and the number of different mtDNA morphs observed (k) in the samples in terms of a common parameter, $\theta = 2N_{e(f)}\nu$, where $N_{e(f)}$ is the effective female population size and where ν is the mutation rate/generation/mtDNA genome. Suppose that an observed distribution is represented by $\{k_r; r = 1, 2, \ldots\}$, where k_r is the number of mtDNA morphs each of which occurs r times in a sample of size n. An unbiased estimate of the population H is given by

$$H = \frac{n}{n-1}\left(1 - \sum_{r=1}^{n} r^2 k_r / n^2\right), \qquad (1)$$

(Nei 1978) and k becomes

$$k = \sum_{r=1}^{n} k_r . \qquad (2)$$

It is well known that the expected values of these two sample statistics are given by

$$E(H) = \theta/(1+\theta), \qquad (3)$$

(Kimura and Crow 1964; Ewens 1972) and

$$E(k) = \theta \cdot \sum_{r=0}^{n-1} (\theta + r)^{-1} . \qquad (4)$$

Equations (3) and (4) provide two alternative estimators of the composite parameter θ, equating the observed values of H and k to their respective expectations (yielding estimators θ_H, the gene-diversity estimator of θ from H, and θ_k, which is also the maximum-likelihood estimator of θ from k). Chakraborty and Schwartz (1990) derived the approximate sampling variances of these two estimators, which can be used to judge whether these two estimators are in accordance with each other.

However, since the sampling distribution of k may not conform to a standard form (such as the normal distribution; Ewens 1972), an alternative, and probably more effective, way of judging the congruence of these two estimators is to check whether the observed value of k deviates substantially from its distribution, when θ_H is used to compute the expected distribution. This is done by computing the tail of the cumulative probability function, i.e., the probability of observing k or more morphs in a sample of size n, given $\theta = \theta_H$, which becomes

$$P(k) = 1 - \sum_{r=1}^{k-1} [\Gamma(\theta)\theta^r n! B(r,n)/\{\Gamma(\theta+n)r!\}] , \qquad (5)$$

where $\Gamma(.)$ is a gamma function, and

$$B(r,n) = \sum \left(\prod_{i=1}^{r} n_i\right)^{-1} , \qquad (6)$$

where $n_1, n_2, \ldots, n_r$ are partitions of the integer n into r classes such that each n_i is greater than zero and $n_1 + n_2 + \ldots + n_r = n$. The summation in expression (6) is over all permutations of $(n_1, n_2, \ldots, n_r)$.

This alternative form of Ewens's (1972) sampling distribution of k is given by F. M. Stewart (see the appendix of Fuerst et al. 1977). This test allows one to judge whether the observed value of k is too large for the given gene diversity. It should be noted that this test is in contrast with Watterson's (1978) test of selective neutrality, where the observed value of H (or its complement) is judged on the basis of its sampling property when θ is estimated from k (i.e., when the estimator θ_k is used to represent the true value of θ). For the present purpose, I prefer the above test procedure as opposed to Watterson's test, since, in the presence of hidden subdivision, θ_H is a better estimator of θ than is θ_k (Chakraborty et al. 1988).

Since this analysis reveals that the observed H in the mtDNA is inconsistent (too low) for the observed k, I address the question of whether this discrepancy is due to the apparent excess of some specific frequency classes of morphs or to uniform over all frequency classes. This is done by using the theory of Chakraborty and Griffiths (1982), where the expected k_r is given by

$$E(k_r) = \frac{\theta}{r} \cdot \frac{n!}{(n-r)!} \cdot \frac{\Gamma(n+\theta-r)}{\Gamma(n+\theta)}, \qquad (7)$$

which can be contrasted with the observed values of k_r for all r, to see whether the discrepancies of the observed k are due to some specific r values only. Note that, because of equation (2), if θ is estimated by θ_k, even though the expected value of k will agree with the observed k, there is no guarantee that, for each r, the observed k_r will agree with expected k_r, given by equation (7). Therefore, the agreement of the observed and expected morph distributions can be checked irrespective of the choice of parameters θ_H or θ_k.

Finally, these tests are performed on the total sample of five Asian populations surveyed by Harihara et al. (1988) to show that the most likely explanation of the observed discrepancies of the mtDNA distribution in some of the individual populations is heterogeneity in the populations, caused by hidden amalgamation of subpopulations.

Results

Estimators of θ for the mtDNA Genome in the Presence of Population Heterogeneity

Table 1 shows the summary statistics (H and k) and n of the mtDNA survey reported by Brega et al. (1986) and Harihara et al. (1988) for six Asian populations. In addition, this table also presents the two estimators of θ (θ_H based on H from eq. [3] and θ_k based on k computed by an iterative solution of eq. [4]), along with their standard errors based on the theory of Chakraborty and Schwartz (1990). Since the restriction enzymes used by Harihara et al. (1988) for the first five populations are more extensive than the ones employed by Brega et al. (1986) for the Tharu population of Nepal, the last row of this table represents the pooled sample excluding the Tharu population, to avoid the problems associated with nonuniformity of laboratory methods in interpreting the present results.

Two features of these estimates are noteworthy. First, the values of θ_k are always larger than θ_H, and this difference is more pronounced in the pooled sample. Since the mtDNA genome behaves like a single genetic locus, and because the standard errors of these estimates therefore are large because of stochastic factors (Nei and Roychoudhury 1974; Li and Nei 1975), these differences may not be statistically significant. However, the larger values of θ_k compared with θ_H, noted in this analysis, are contrary to the single-locus esti-

Table 1

Parameter Estimates from the mtDNA Morph Distributions in Six Asian Populations

Population (n)	k	H	Estimate of $\theta = 4N_e\nu$ from $H(\theta_H)$	$k(\theta_k)$
Japanese (74)	11	.40 ± .07	.68 ± .20	3.34 ± 1.24
Ainu (48)	6	.23 ± .08	.31 ± .14	1.59 ± .81
Korean (64)	7	.33 ± .08	.51 ± .17	1.80 ± .83
Aeta (37)	3	.20 ± .09	.26 ± .13	.57 ± .43
Vedda (20)	4	.51 ± .10	1.16 ± .46	1.21 ± .81
Tharu (91)	13	.65 ± .04	1.93 ± .37	3.92 ± 1.33
Pooled (243)[a]	20	.34 ± .04	.52 ± .09	4.99 ± 1.31

[a] Excluding the Tharu population (see text for details).

mates of θ, seen in the isozyme data analysis of Zouros (1979) and in the mtDNA data analysis of several *Drosophila* species (Nei 1987). Under the assumption that these samples are drawn from homogeneous equilibrium populations, θ_H is an overestimate of the true parameter θ, while θ_k should be closer to the true value of θ, because it is the maximum-likelihood estimator. Given the present observation, I may suspect that the above assumptions may not hold for these populations. Noting that the six populations are anthropologically distinct and that in the pooled sample the discrepancy between the two estimators is more pronounced, I could postulate that hidden heterogeneity in each population is probably one of the reasons for these observations. In view of this, the question is which estimator of θ should be regarded as more reliable when hidden amalgamation is present in a sample.

In order to address this question, I must consider the second noteworthy feature of these estimates—i.e., that there is a direct positive association between the estimates θ_k and n, while this is not so for the estimator θ_H. Even though it is known that the statistic k is critically dependent on n whereas H is comparatively more stable over differences of n (particularly when the unbiased estimator, eq. [1], is used), there is no apparent justification for the n dependency of the estimator θ_k, since its computation adjusts for the n employed in a survey (see eq. [4]). To check whether the observed differences of θ_H and θ_k can be explained by the differences in n of the different populations, I recomputed the θ_k estimators by using iterative solutions of equation (4), after obtaining an adjusted value for k (k_{adj}), reducing n from each population to the minimum value of 20 (equivalent to the n from the Veddas of Sri Lanka). k_{adj} is computed from the equation (Chakraborty et al. 1988).

$$k_{adj} = k - \sum_{i=1}^{k} e^{-np_i}, \qquad (8)$$

where the p_i's are the relative frequencies of the different morphs in the original samples. Table 2 presents these k_{adj} values for each of the six populations, as well as for the pooled sample (excluding the Tharu population). These k_{adj} values, along with an n of 20 are then substituted for iterative solutions of equation (4), to provide n-adjusted estimators θ_k(adj), which can be contrasted with the θ_H values of table 1, because the effect of variation in n is now completely removed from the θ_k values of table 1. The last column of table 2 shows the θ_k (adj) values. Although these values no longer show any correlation with the original n values, they still remain larger than the θ_H values both for each population and for the pooled sample. Therefore, I might conclude that the θ_k estimators are affected by the hidden amalgamation within each population sample and that this cannot be removed by simple adjustment of n alone. Even though amalgamation may also inflate the θ_H estimates from their true values, the effect of amalgamation on θ_H values is comparatively smaller, since H is less affected by amalgamation (Chakraborty et al. 1988). These observations together suggest that, of the two estimators of θ, θ_H is the one preferred for the present data.

Table 2

k_{adj} and θ_k(adj) in Six Asian Populations

Population	k	k_{adj}[a]	θ_k(adj)
Japanese	11	4.47	1.48 ± .94
Ainu	6	3.17	.80 ± .61
Korean	7	3.59	1.00 ± .71
Aeta	3	2.32	.44 ± .41
Vedda	4	4.00	1.21 ± .81
Tharu	13	5.43	2.09 ± 1.22
Pooled[b]	20	4.24	1.34 ± .87

[a] Based on n = 20, computed by eq. (8).
[b] Excluding the Tharu population.

Excess of Total k for the Observed Gene Diversity in the mtDNA Genome

Since the θ_H estimates are comparatively better than the θ_k values, table 3 presents the expectations and standard errors of the k values in these different samples, for their respective n values. In all cases, I see that the observed k is larger than the expected k. Simple t-tests should not be used to see whether there is an excess of k in each sample, because the sampling distribution of k is not normal. Using the cumulative distribution function, given in equation (5), the last column of table 3 shows the large deviation probabilities, $P(k)$, for each sample and for the pooled data (excluding the Tharu sample).

This analysis immediately reflects that there is an excess in the total k for the given H in these populations. The excess is statistically significant at the 5% level in all of the populations except the Aetas of the Philippines and the Veddas of Sri Lanka. The excess is conspicuous in the Japanese (a large population) and in the Ainu population (which probably received genes from outside recently). The nonsignificant excess is the Aeta and Vedda populations probably is due to the small

Table 3

Observed and Expected *k* Values in Six Asian Populations

Population (n)	k		$P(k)$
	Observed	Expected ± SE[a]	
Japanese (74)	11	3.80 ± 1.56	<.001
Ainu (48)	6	2.24 ± 1.07	.006
Korean (64)	7	3.09 ± 1.36	.014
Aeta (37)	3	1.98 ± .95	.256
Vedda (20)	4	3.91 ± 1.48	.584
Tharu (91)	13	8.02 ± 2.36	.035
Pooled (243)[b]	20	3.82 ± 1.60	$<10^{-8}$

[a] Computed on the basis of H shown in table 1.
[b] Excluding the Tharu population.

number of individuals sampled from these populations. Omoto et al. (1978) and Ellepola and Wikramanayake (1986), in their isozyme surveys based on electrophoretically detected genetic variation, suggested that these two populations are also probably affected by recent gene migration from outside. Although there are other alternative explanations for the excess k (see Discussion and Conclusion), the reason that heterogeneity is the most likely explanation is evident from the fact that the pooled sample shows a more conspicuous excess ($P < 10^{-8}$).

These results can also be alternatively stated as observed H being too low for the observed k seen in the samples from these populations. For showing this, I computed the expected H from equation (3), using the estimates of θ_k for each population. Figure 1*b* shows these results, and figure 1*a* shows the contrast of observed k and the expected k based on H. These two panels together suggest that the observed H and the total k are not consistent with each other, as would be predicted if these samples were drawn from homogeneous equilibrium populations.

The Excess in k Is Due to Excess of Rare Ones

Having shown that there are too many different mtDNA morphs in several of these populations as compared with the expected values based on H values, I asked whether such excess is uniformly distributed over all frequency classes of the mtDNA-morph distribution. This is addressed with the help of equation (7), which can be evaluated for any value of r, once the θ parameter is suitably estimated. Note that $r = 1$ refers to the singleton morphs, i.e., the morphs with only one copy in a sample, and that $r = 2$ refers to the doubletons, i.e., the morphs with two copies in a sample. On the basis of the n values in the present data, the rare morph class (i.e., the morphs with frequency less than 1% or 5% in the population) consists mainly of the morphs with one or two copies each per sample. Chakraborty and Griffiths (1982) also evaluated the sampling variances of k_r, which show that for small r (such as 1 or 2) the asymptotic distribution of k_r is Poisson, so that $V(k_r) \approx E(k_r)$. Note that if θ_k estimates are used in expression (7) to evaluate the expected k_r, their sum over all values of $r = 1, 2, \ldots, n$ should equal the observed value of k, whereas when θ_H is used the sum will be less than k. Therefore, the congruence of the observed and expected values of k_r can be tested irrespective of the choice of θ estimators. Table 4 shows these computations for each population separately, as well as for the pooled sample (excluding the Tharu sam-

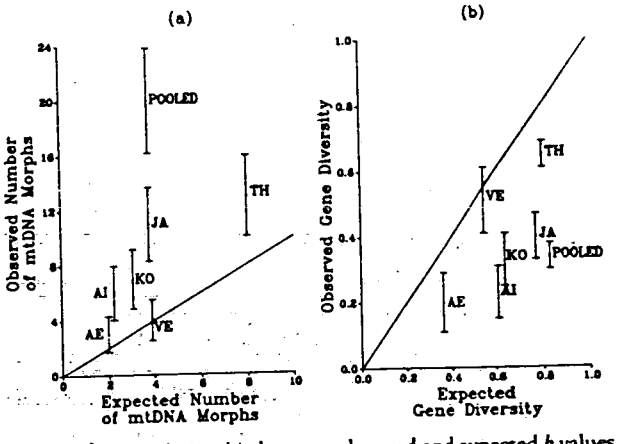


Figure 1 Relationship between observed and expected k values, on the basis of observed H in six Asian populations (*a*) and relationship between observed and expected H, on the basis of observed k (*b*). JA = Japanese; AI = Ainu; KO = Korean; AE = Aeta; VE = Vedda; TH = Tharu. The pooled data represent the total of the first five populations (excluding the Tharu sample). The vertical bars represent ± SE deviations.

Table 4

Observed and Expected k_r Values in Six Asian Populations

Population (n)	r	Observed Frequency	Expected Frequency of $k_r \pm$ SE, P	
			Based on θ_H	Based on θ_k
Japanese (74)	1	6	$.68 \pm .83, <10^{-4}$	$3.24 \pm 1.80, .110$
	2	2	$.34 \pm .59, .047$	$1.57 \pm 1.25, .465$
Ainu (48)	1	4	$.31 \pm .56, <10^{-3}$	$1.57 \pm 1.25, .075$
	2	1	$.16 \pm .40, .147$	$.78 \pm .88, .540$
Korean (64)	1	3	$.51 \pm .72, .015$	$1.77 \pm 1.33, .263$
	2	1	$.26 \pm .51, .228$	$.88 \pm .94, .584$
Aeta[a] (37)	2	2	$.13 \pm .37, .008$	$.29 \pm .54, .035$
Vedda[b] (20)	1	2	$1.15 \pm 1.07, .319$	$1.20 \pm 1.09, .336$
Tharu (91)	1	5	$1.91 \pm 1.38, .045$	$3.80 \pm 1.95, .332$
	2	5	$.95 \pm .97, .003$	$1.84 \pm 1.36, .039$
Pooled[c] (243)	1	8	$.52 \pm .72, <10^{-7}$	$4.91 \pm 2.22, .124$
	2	6	$.26 \pm .51, <10^{-6}$	$2.42 \pm 1.55, .037$

[a] No singleton morph observed.
[b] No doubleton morph observed.
[c] Excluding the Tharu population.

ple). The P shown in this table represents the probability of observing k_r or more morphs for $r = 1$ (or 2), given the respective expected value of k_r, in which a Poisson distribution of k_r is assumed (Chakraborty and Griffiths 1982).

It is clear from table 4 that the observed excess k is predominantly due to an excess in the observed number of rare morphs. When θ is estimated from the observed H, all populations except the Aetas of the Philippines show a significant ($P < .05$) excess of rare morphs, and this is more conspicuous in the pooled sample. When θ is estimated from k, the excess in the number of rare morphs does not reach significance in most cases, but the observed values of k_r are consistently larger than the expected values of k_r, even though for this estimator of θ the expected total k values equal the respective observed total k values. It can therefore be concluded that, irrespective of the choice of the parameter θ, there is an excess of rare mtDNA morphs in most of these samples, in contrast with the expectations based on the assumption that each sample is drawn from an equilibrium-homogenous population. Since this departure from the expectation is more profound in the pooled sample, it may be suspected that the most likely explanation is that the majority of the populations are heterogeneous in nature, so that they do not represent single homogeneous panmictic populations.

Discussion and Conclusion

The above findings indicate that mtDNA variability, as detected by the number of different RFLP morphs, is too high for the observed H in this genome in most of the Asian populations sampled. This excess is mainly due to the increased frequencies of rare morphs and is more conspicuous in the pooled sample than in the small populations (Aeta and Vedda). These observations are parallel to the findings reported for Amerindian populations of South and Central America (Chakraborty et al. 1988), findings based on isozyme polymorphisms. Also note that these observations are completely parallel to Whittam et al.'s (1986) findings based on mtDNA data. This parallelism indicates that the present results on the mtDNA genome are probably not an artifact of stochastic sampling errors due to data for a single locus. Single-locus data generally would have produced a deviation in the other direction, since Zouros (1979) has shown that the single-locus estimate of θ_H is generally larger than that of θ_k, just the opposite of the present findings.

There are several other possible explanations for the excess number of alleles produced through an apparent increase in rare ones. Ohta (1973, 1976) argued that this can occur in the presence of slightly deleterious mutations, while Nei et al. (1975), Maruyama and Fuerst (1984, 1985), and Watterson (1984) have shown that recent population bottlenecks can also cause a disproportionate increase in the number of variant alleles, in contrast with the expectations based on observed H. Whittam et al. (1986) ascribed the observed deviations of the mtDNA allele frequency distributions to such causes, apparently not realizing that most of their population data are not really from any anthropologically well-defined populations (e.g., at least two of their

five sampling units are quite heterogeneous: Asians [group II] and Europe, North Africa, and Middle East [group IV]). There are at least two lines of evidence suggesting that hidden heterogeneity is the most likely explanation in such contexts. First, the analysis performed in the present paper clearly demonstrates that the pooled sample (which is obviously an amalgamated sample consisting of five anthropologically distinct populations) shows a more profound deviation in the direction seen in each individual population. Whittam et al. (1986) also found this, but they failed to interpret their findings accordingly. Second, Chakraborty et al. (1988) have demonstrated that the isozyme loci reveal the effect of amalgamation when data from relatively homogeneous populations are sequentially agglomerated, producing an excess of allele numbers (total as well as rare), whose extent depends both on how many subpopulations are amalgamated into a sample and on how distinct the subpopulations are (measured by the average genetic distance between them).

In addition, for several of these populations there is independent evidence of internal substructuring. For example, Nei and Imaizumi (1966*a*, 1966*b*) showed that the coefficient of gene differentiation within the different subpopulations in Japan is not negligible. Since their analysis was based on blood-group genes, their results cannot be directly compared with the present findings. Analysis of data on 32 isozyme loci from a large sample of Japanese (Neel et al. 1988), however, corroborates the present finding, since in that survey 5.53 alleles/locus were observed whereas the observed average H (.087) at these loci predicts only 2.02 ± 0.18 alleles/locus (Chakraborty, in press). A phylogenetic analysis of mtDNA RFLP variation, conducted by Horai and Matsunaga (1986), indicates that at least two distinct lineages of mtDNA exist in the Japanese population. There is no such independent documentation of internal fragmentation of the other populations analyzed here. Nevertheless, the isozyme studies of Omoto (1972) in the Ainu population of Japan, of Omoto et al. (1978) in the Aeta population of the Philippines, and of Ellepola and Wikramanayake (1986) in the Vedda population of Sri Lanka suggest internal substructuring as well as recent gene migration from the outside, both of which should have a net effect similar to that of amalgamation. The typological classification of the Tharu mtDNA morphs also indicated some evidence of foreign-gene admixture (Brega et al. 1986), suggesting that the present evidence of heterogeneity in the Tharu population is not inconsistent with what is known about the genetic structure of this population.

Finally, methodological differences between the present study and that of Whittam et al. (1986) should be noted; the earlier study is based on the estimate of θ based on the total number of alleles (k), which is quite error prone in the presence of hidden substructuring within a population (Chakraborty et al. 1988; Chakraborty and Schwartz 1990). On the contrary, use of θ_H as an estimator of θ circumvents this problem. In addition, this approach avoids tedious simulations employed by Whittam et al. (1986), since I can use analytical expressions such as equations (5) and (7) to conduct the relevant statistical tests. In principle the present technique can be applied to other broad sets of data; however, at present, the lack of uniformity of laboratory protocols (as mentioned earlier) makes it difficult to use the reported data to establish allelic homologies as a means of limiting such analysis. Nevertheless, this analysis shows that the effect of hidden heterogeneity within a population can be detected with allele frequency data from a single locus, provided that substantial genetic variation exists at the locus. Therefore, the hypervariable minisatellite loci—i.e., the VNTR loci—should be useful in revealing the genetic structure of populations, because such loci generally exhibit substantially larger variability than that seen in either the RFLP or the isozyme loci.

Acknowledgments

This work was supported by U.S. Public Health Service research grant GM 41399 from the National Institutes of Health. I thank Drs. W. J. Schull and M. Nei for their helpful suggestions during the conduct of this study.

References

Blanc H, Chen K-H, D'Amore MA, Wallace DC (1983) Amino acid change associated with the major polymorphic Hinc II site of Oriental and Caucasian mitochondrial DNAs. Am J Hum Genet 35:167–176

Brega A, Gardella R, Semino O, Morpurgo G, Astaldi Ricotti GB, Wallace DC, Santachiara Benerecetti AS (1986) Genetic studies on the Tharu population of Nepal: restriction endonuclease polymorphisms of mitochondrial DNA. Am J Hum Genet 39:502–512

Brown WM (1980) Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc Natl Acad Sci USA 77:3605–3609

Cann RL, Brown WM, Wilson AC (1982) Evolution of human mitochondrial DNA: a preliminary report. In: Bonne-Tamir B, Cohen P, Goodman RN (eds). Human Genetics. Part A: The Unfolding Genome. Alan R. Liss, New York, pp 157–166

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cann RL, Wilson AC (1983) Length mutations in human mitochondrial DNA. Genetics 104:699–711

Chakraborty R. Genetic profile of cosmopolitan populations: effects of hidden subdivision. Anthropol Anz (in press)

Chakraborty R, Griffiths RC (1982) Correlation of heterozygosity and number of alleles in different frequency classes. Theor Popul Biol 21:205–218

Chakraborty R, Schwartz RJ (1990) Selective neutrality of surname distribution in an immigrant Indian community of Houston, Texas. Am J Hum Biol 2:1–15

Chakraborty R, Smouse PE, Neel JV (1988) Population amalgamation and genetic variation: observations on artifically agglomerated tribal populations of Central and South America. Am J Hum Genet 43:709–725

Denaro M, Blanc H, Johnson MJ, Chen KH, Wilmsen E, Cavalli-Sforza LL, Wallace DC (1981) Ethnic variation in *Hpa*I endonuclease cleavage patterns of human mitochondrial DNA. Proc Natl Acad Sci USA 78:5768–5772

Ellepola SB, Wikramanayake ER (1986) A genetic study of the Veddas and the Sinhalese. Ceylon J Med Sci 29:1–21

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87–112

Fuerst PA, Chakraborty R, Nei M (1977) Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics 86:455–483

Harihara S, Saitou N, Hirai M, Gojobori T, Park KS, Misawa S, Ellopola SB, et al (1988) Mitochondrial DNA polymorphism among five Asian populations. Am J Hum Genet 43:134–143

Horai S, Gojobori T, Matsunaga E (1987) Evolutionary implications of mitochondrial DNA polymorphism in human populations. Hum Genet 74:177–181

Horai S, Matsunaga E (1986) Mitochondrial DNA polymorphism in Japanese. II. Analysis with restriction enzymes of four or five base pair recognition. Hum Genet 72:105–117

Johnson MJ, Wallace DC, Ferris SD, Rattazzi MC, Cavalli-Sforza LL (1983) Radiation of human mitochondrial DNA types analyzed by restriction endonuclease cleavage patterns. J Mol Evol 19:255–271

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738

Li WH, Nei M (1975) Drift variances of heterozygosity and genetic distance in transient states. Genet Res 25:229–248

Maruyama T, Fuerst PA (1984) Population bottlenecks and non-equilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. Genetics 108:745–763

——— (1985) Population bottlenecks and non-equilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. Genetics 111:691–703

Neel JV, Satoh C, Smouse P, Asakawa J-I, Takahashi N, Goriki K, Fujita M, et al (1988) Protein variants in Hiroshima and Nagasaki: tales of two cities. Am J Hum Genet 43:870–893

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590

——— (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Imaizumi Y (1966*a*) Genetic structure of human populations. I. Local differentiation of blood group frequencies in Japan. Heredity 21:9–25

——— (1966*b*) Genetic structure of human populations. II. Differentiation of blood group gene frequencies among isolated populations. Heredity 21:183–190

Nei M, Maruyama T, Chakraborty R (1975) The bottleneck effect and genetic variability in populations. Evolution 29:1–10

Nei M, Roychoudhury AK (1974) Sampling variance of heterozygosity and genetic distance. Genetics 76:379–390

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246:96–98

——— (1976) Role of slightly deleterious mutations in molecular evolution and polymorphism. Theor Popul Biol 10:254–275

Omoto K (1972) Polymorphism and genetic affinities of the Ainu of Hokkaido. Hum Biol Oceania 1:278–288

Omoto K, Misawa S, Harada S, Sumpaico JS, Medado PM, Ogonuki H (1978) Population genetic studies of the Philippine Negritos. I. A pilot survey of red cell enzyme and serum protein groups. Am J Hum Genet 30:190–201

Watterson GA (1978) The homozygosity test of neutrality. Genetics 88:405–417

——— (1984) Allele frequencies after a bottleneck. Theor Popul Biol 26:387–407

Whittam TS, Clark AG, Stoneking M, Cann RL, Wilson AC (1986) Allelic variation in human mitochondrial genes based on patterns of restriction site polymorphism. Proc Natl Acad Sci USA 83:9611–9615

Zouros E (1979) Mutation rates, population sizes, and amounts of electrophoretic variation of enzyme loci in natural populations. Genetics 92:623–646

Am. J. Hum. Genet. 47:87–94, 1990

Mitochondrial DNA Polymorphism Reveals Hidden Heterogeneity within Some Asian Populations

Ranajit Chakraborty

Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston

Summary

Use of data on mtDNA morph distributions from six Asian populations has shown that the observed number of different mtDNA morphs is too large when compared with the number expected on the basis of the observed gene diversity in the mtDNA genome. This excess number of morphs mainly occurs through an excess of rare morphs, and this discrepancy is more pronounced in a pooled sample of five Asian populations. It is suggested that this discrepancy is probably due to internal heterogeneity in each of the anthropologically defined populations. This analysis demonstrates the utility that population data for a single locus, such as the mtDNA genome, have for detecting hidden heterogeneity in populations, provided that the locus has substantial genetic variability, so that many variant alleles can be detected.

Introduction

Mitochondrial DNA (mtDNA) is particularly useful in evolutionary studies of the ethnic origins of human populations (e.g., see Brown 1980; Denaro et al. 1981; Blanc et al. 1983; Johnson et al. 1983; Horai et al. 1987; Harihara et al. 1988) and in detecting DNA polymorphisms that existed before the geographic dispersal of the human species (Cann et al. 1982, 1987; Cann and Wilson 1983). mtDNA has a distinct advantage over nuclear DNA for population genetic studies because (1) the evolutionary rate of nucleotide substitutions appears to be larger in the mtDNA genome compared with the nuclear genes (e.g., see Nei 1987), (2) the determination of the various mtDNA morphs (haplotypes) is unequivocal from population data, since mtDNA is maternally inherited, and (3) the generation of different mtDNA morphs can only occur through new mutations, and no recombination has to be invoked in studying the maintenance of mtDNA polymorphisms.

In a recent study, Whittam et al. (1986) analyzed allelic variations in 145 human mtDNAs representing samples from five geographic regions. They concluded that while the allele frequency distributions at different loci in the mtDNA genome follow the general predictions of the equilibrium theory of a mutation-drift model of selectively neutral mutations, certain deviations (e.g., observed gene diversity lower than that expected and excesses in the frequencies of common alleles and in the number of singleton alleles) can be attributed to possible bottleneck effect during recent human evolution and to the action of purifying selection. Since such observations can also be explained by hidden substructuring of populations, as evidenced in the study of electrophoretic variations in South and Central American Indians (Chakraborty et al. 1988), the purpose of the present paper is to demonstrate that the substantial variation of the mtDNA genome can be used to reveal hidden heterogeneity within anthropologically defined populations. This is shown by examining the mtDNA-morph distributions in several Asian populations, studied by Brega et al. (1986) and Harihara et al. (1988), and by utilizing the sampling theory of selectively neutral alleles (Ewens 1972; Chakraborty and Griffiths 1982). It is suggested that the discrepancies between the observed and expected distributions of the mtDNA morphs in most Asian populations are probably due to their internal hidden heterogeneity, and this conclusion probably applies to the populations examined by Whittam et al. (1986) as well.

Received December 19, 1989; revision received February 22, 1990.

Address for correspondence and reprints: Dr. Ranajit Chakraborty, Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, TX 77225.

0002-9297/90/4701-0012$02.00

Material and Methods

Data

It is well known that the power of resolution of molecular variability in the mtDNA genome varies in RFLP studies, depending on the restriction enzymes used and the detectability of fragment size differences. Therefore, uniformity of laboratory methods must be established in comparing the mtDNA-morph distributions among different populations. Recently, Harihara et al. (1988) published mtDNA-morph distributions in five Asian populations—the Japanese and the Ainu from northern Japan, the Koreans, the Negrito (Aeta) from the Philippines, and the Veddas of Sri Lanka—by using 13 restriction enzymes and following uniform laboratory conditions. Brega et al. (1986) used six enzymes from the above set to survey the mtDNA-morph distribution in the Tharu population of Nepal. These data form the basis of the present analysis. In the study of five Asian populations Harihara et al. (1988) observed 20 different mtDNA morphs in a total sample of 243 individuals, whereas in 91 individuals from the Tharu population of Nepal Brega et al. (1986) observed 13 mtDNA morphs (haplotypes).

Theory

Since the rate of nucleotide substitutions and the extent of nucleotide diversity in the mtDNA genome roughly follow the predictions of the neutral mutation hypothesis (for a review, see Nei 1987), I ask whether the various aspects of the mtDNA distributions in these six populations are consistent with the expectations from the sampling theory of selectively neutral mutations (Ewens 1972; Chakraborty and Griffiths 1982), which are based on the assumption that the sampling has occurred from a single homogeneous population in each case. This is accomplished by examining the expectations of two summary statistics of the mtDNA-morph distributions—gene diversity (H) and the number of different mtDNA morphs observed (k) in the samples in terms of a common parameter, $\theta = 2N_{e(f)}\nu$, where $N_{e(f)}$ is the effective female population size and where ν is the mutation rate/generation/mtDNA genome. Suppose that an observed distribution is represented by $\{k_r; r = 1, 2, \ldots\}$, where k_r is the number of mtDNA morphs each of which occurs r times in a sample of size n. An unbiased estimate of the population H is given by

$$H = \frac{n}{n-1}\left(1 - \sum_{r=1}^{n} r^2 k_r / n^2\right), \qquad (1)$$

(Nei 1978) and k becomes

$$k = \sum_{r=1}^{n} k_r . \qquad (2)$$

It is well known that the expected values of these two sample statistics are given by

$$E(H) = \theta/(1+\theta), \qquad (3)$$

(Kimura and Crow 1964; Ewens 1972) and

$$E(k) = \theta \cdot \sum_{r=0}^{n-1} (\theta + r)^{-1} . \qquad (4)$$

Equations (3) and (4) provide two alternative estimators of the composite parameter θ, equating the observed values of H and k to their respective expectations (yielding estimators θ_H, the gene-diversity estimator of θ from H, and θ_k, which is also the maximum-likelihood estimator of θ from k). Chakraborty and Schwartz (1990) derived the approximate sampling variances of these two estimators, which can be used to judge whether these two estimators are in accordance with each other.

However, since the sampling distribution of k may not conform to a standard form (such as the normal distribution; Ewens 1972), an alternative, and probably more effective, way of judging the congruence of these two estimators is to check whether the observed value of k deviates substantially from its distribution, when θ_H is used to compute the expected distribution. This is done by computing the tail of the cumulative probability function, i.e., the probability of observing k or more morphs in a sample of size n, given $\theta = \theta_H$, which becomes

$$P(k) = 1 - \sum_{r=1}^{k-1} [\Gamma(\theta)\theta^r n! B(r,n)/\{\Gamma(\theta+n) r!\}] , \qquad (5)$$

where $\Gamma(.)$ is a gamma function, and

$$B(r,n) = \sum \left(\prod_{i=1}^{r} n_i\right)^{-1} , \qquad (6)$$

where $n_1, n_2, \ldots, n_r$ are partitions of the integer n into r classes such that each n_i is greater than zero and $n_1 + n_2 + \ldots + n_r = n$. The summation in expression (6) is over all permutations of $(n_1, n_2, \ldots, n_r)$.

This alternative form of Ewens's (1972) sampling distribution of k is given by F. M. Stewart (see the appendix of Fuerst et al. 1977). This test allows one to judge whether the observed value of k is too large for the given gene diversity. It should be noted that this test is in contrast with Watterson's (1978) test of selective neutrality, where the observed value of H (or its complement) is judged on the basis of its sampling property when θ is estimated from k (i.e., when the estimator θ_k is used to represent the true value of θ). For the present purpose, I prefer the above test procedure as opposed to Watterson's test, since, in the presence of hidden subdivision, θ_H is a better estimator of θ than is θ_k (Chakraborty et al. 1988).

Since this analysis reveals that the observed H in the mtDNA is inconsistent (too low) for the observed k, I address the question of whether this discrepancy is due to the apparent excess of some specific frequency classes of morphs or to uniform over all frequency classes. This is done by using the theory of Chakraborty and Griffiths (1982), where the expected k_r is given by

$$E(k_r) = \frac{\theta}{r} \cdot \frac{n!}{(n-r)!} \cdot \frac{\Gamma(n+\theta-r)}{\Gamma(n+\theta)}, \qquad (7)$$

which can be contrasted with the observed values of k_r for all r, to see whether the discrepancies of the observed k are due to some specific r values only. Note that, because of equation (2), if θ is estimated by θ_k, even though the expected value of k will agree with the observed k, there is no guarantee that, for each r, the observed k_r will agree with expected k_r, given by equation (7). Therefore, the agreement of the observed and expected morph distributions can be checked irrespective of the choice of parameters θ_H or θ_k.

Finally, these tests are performed on the total sample of five Asian populations surveyed by Harihara et al. (1988) to show that the most likely explanation of the observed discrepancies of the mtDNA distribution in some of the individual populations is heterogeneity in the populations, caused by hidden amalgamation of subpopulations.

Results

Estimators of θ for the mtDNA Genome in the Presence of Population Heterogeneity

Table 1 shows the summary statistics (H and k) and n of the mtDNA survey reported by Brega et al. (1986) and Harihara et al. (1988) for six Asian populations. In addition, this table also presents the two estimators of θ (θ_H based on H from eq. [3] and θ_k based on k computed by an iterative solution of eq. [4]), along with their standard errors based on the theory of Chakraborty and Schwartz (1990). Since the restriction enzymes used by Harihara et al. (1988) for the first five populations are more extensive than the ones employed by Brega et al. (1986) for the Tharu population of Nepal, the last row of this table represents the pooled sample excluding the Tharu population, to avoid the problems associated with nonuniformity of laboratory methods in interpreting the present results.

Two features of these estimates are noteworthy. First, the values of θ_k are always larger than θ_H, and this difference is more pronounced in the pooled sample. Since the mtDNA genome behaves like a single genetic locus, and because the standard errors of these estimates therefore are large because of stochastic factors (Nei and Roychoudhury 1974; Li and Nei 1975), these differences may not be statistically significant. However, the larger values of θ_k compared with θ_H, noted in this analysis, are contrary to the single-locus esti-

Table 1

Parameter Estimates from the mtDNA Morph Distributions in Six Asian Populations

Population (n)	k	H	Estimate of $\theta = 4N_e\nu$ from $H(\theta_H)$	$k(\theta_k)$
Japanese (74)	11	.40 ± .07	.68 ± .20	3.34 ± 1.24
Ainu (48)	6	.23 ± .08	.31 ± .14	1.59 ± .81
Korean (64)	7	.33 ± .08	.51 ± .17	1.80 ± .83
Aeta (37)	3	.20 ± .09	.26 ± .13	.57 ± .43
Vedda (20)	4	.51 ± .10	1.16 ± .46	1.21 ± .81
Tharu (91)	13	.65 ± .04	1.93 ± .37	3.92 ± 1.33
Pooled (243)[a]	20	.34 ± .04	.52 ± .09	4.99 ± 1.31

[a] Excluding the Tharu population (see text for details).

mates of θ, seen in the isozyme data analysis of Zouros (1979) and in the mtDNA data analysis of several *Drosophila* species (Nei 1987). Under the assumption that these samples are drawn from homogeneous equilibrium populations, θ_H is an overestimate of the true parameter θ, while θ_k should be closer to the true value of θ, because it is the maximum-likelihood estimator. Given the present observation, I may suspect that the above assumptions may not hold for these populations. Noting that the six populations are anthropologically distinct and that in the pooled sample the discrepancy between the two estimators is more pronounced, I could postulate that hidden heterogeneity in each population is probably one of the reasons for these observations. In view of this, the question is which estimator of θ should be regarded as more reliable when hidden amalgamation is present in a sample.

In order to address this question, I must consider the second noteworthy feature of these estimates—i.e., that there is a direct positive association between the estimates θ_k and n, while this is not so for the estimator θ_H. Even though it is known that the statistic k is critically dependent on n whereas H is comparatively more stable over differences of n (particularly when the unbiased estimator, eq. [1], is used), there is no apparent justification for the n dependency of the estimator θ_k, since its computation adjusts for the n employed in a survey (see eq. [4]). To check whether the observed differences of θ_H and θ_k can be explained by the differences in n of the different populations, I recomputed the θ_k estimators by using iterative solutions of equation (4), after obtaining an adjusted value for k (k_{adj}), reducing n from each population to the minimum value of 20 (equivalent to the n from the Veddas of Sri Lanka). k_{adj} is computed from the equation (Chakraborty et al. 1988).

$$k_{adj} = k - \sum_{i=1}^{k} e^{-np_i}, \qquad (8)$$

where the p_i's are the relative frequencies of the different morphs in the original samples. Table 2 presents these k_{adj} values for each of the six populations, as well as for the pooled sample (excluding the Tharu population). These k_{adj} values, along with an n of 20 are then substituted for iterative solutions of equation (4), to provide n-adjusted estimators θ_k(adj), which can be contrasted with the θ_H values of table 1, because the effect of variation in n is now completely removed from the θ_k values of table 1. The last column of table 2 shows the θ_k (adj) values. Although these values no longer show any correlation with the original n values, they still remain larger than the θ_H values both for each population and for the pooled sample. Therefore, I might conclude that the θ_k estimators are affected by the hidden amalgamation within each population sample and that this cannot be removed by simple adjustment of n alone. Even though amalgamation may also inflate the θ_H estimates from their true values, the effect of amalgamation on θ_H values is comparatively smaller, since H is less affected by amalgamation (Chakraborty et al. 1988). These observations together suggest that, of the two estimators of θ, θ_H is the one preferred for the present data.

Table 2

k_{adj} and θ_k(adj) in Six Asian Populations

Population	k	k_{adj}[a]	$\theta_k(adj)$
Japanese	11	4.47	1.48 ± .94
Ainu	6	3.17	.80 ± .61
Korean	7	3.59	1.00 ± .71
Aeta	3	2.32	.44 ± .41
Vedda	4	4.00	1.21 ± .81
Tharu	13	5.43	2.09 ± 1.22
Pooled[b]	20	4.24	1.34 ± .87

[a] Based on n = 20, computed by eq. (8).
[b] Excluding the Tharu population.

Excess of Total k for the Observed Gene Diversity in the mtDNA Genome

Since the θ_H estimates are comparatively better than the θ_k values, table 3 presents the expectations and standard errors of the k values in these different samples, for their respective n values. In all cases, I see that the observed k is larger than the expected k. Simple t-tests should not be used to see whether there is an excess of k in each sample, because the sampling distribution of k is not normal. Using the cumulative distribution function, given in equation (5), the last column of table 3 shows the large deviation probabilities, $P(k)$, for each sample and for the pooled data (excluding the Tharu sample).

This analysis immediately reflects that there is an excess in the total k for the given H in these populations. The excess is statistically significant at the 5% level in all of the populations except the Aetas of the Philippines and the Veddas of Sri Lanka. The excess is conspicuous in the Japanese (a large population) and in the Ainu population (which probably received genes from outside recently). The nonsignificant excess is the Aeta and Vedda populations probably is due to the small

Table 3

Observed and Expected *k* Values in Six Asian Populations

Population (n)	k		$P(k)$
	Observed	Expected ± SE[a]	
Japanese (74)	11	3.80 ± 1.56	<.001
Ainu (48)	6	2.24 ± 1.07	.006
Korean (64)	7	3.09 ± 1.36	.014
Aeta (37)	3	1.98 ± .95	.256
Vedda (20)	4	3.91 ± 1.48	.584
Tharu (91)	13	8.02 ± 2.36	.035
Pooled (243)[b]	20	3.82 ± 1.60	$<10^{-8}$

[a] Computed on the basis of H shown in table 1.
[b] Excluding the Tharu population.

number of individuals sampled from these populations. Omoto et al. (1978) and Ellepola and Wikramanayake (1986), in their isozyme surveys based on electrophoretically detected genetic variation, suggested that these two populations are also probably affected by recent gene migration from outside. Although there are other alternative explanations for the excess k (see Discussion and Conclusion), the reason that heterogeneity is the most likely explanation is evident from the fact that the pooled sample shows a more conspicuous excess ($P < 10^{-8}$).

These results can also be alternatively stated as observed H being too low for the observed k seen in the samples from these populations. For showing this, I computed the expected H from equation (3), using the estimates of θ_k for each population. Figure 1*b* shows these results, and figure 1*a* shows the contrast of observed k and the expected k based on H. These two panels together suggest that the observed H and the total k are not consistent with each other, as would be predicted if these samples were drawn from homogeneous equilibrium populations.

The Excess in k Is Due to Excess of Rare Ones

Having shown that there are too many different mtDNA morphs in several of these populations as compared with the expected values based on H values, I asked whether such excess is uniformly distributed over all frequency classes of the mtDNA-morph distribution. This is addressed with the help of equation (7), which can be evaluated for any value of r, once the θ parameter is suitably estimated. Note that $r = 1$ refers to the singleton morphs, i.e., the morphs with only one copy in a sample, and that $r = 2$ refers to the doubletons, i.e., the morphs with two copies in a sample. On the basis of the n values in the present data, the rare morph class (i.e., the morphs with frequency less than 1% or 5% in the population) consists mainly of the morphs with one or two copies each per sample. Chakraborty and Griffiths (1982) also evaluated the sampling variances of k_r, which show that for small r (such as 1 or 2) the asymptotic distribution of k_r is Poisson, so that $V(k_r) \approx E(k_r)$. Note that if θ_k estimates are used in expression (7) to evaluate the expected k_r, their sum over all values of $r = 1, 2, \ldots, n$ should equal the observed value of k, whereas when θ_H is used the sum will be less than k. Therefore, the congruence of the observed and expected values of k_r can be tested irrespective of the choice of θ estimators. Table 4 shows these computations for each population separately, as well as for the pooled sample (excluding the Tharu sam-

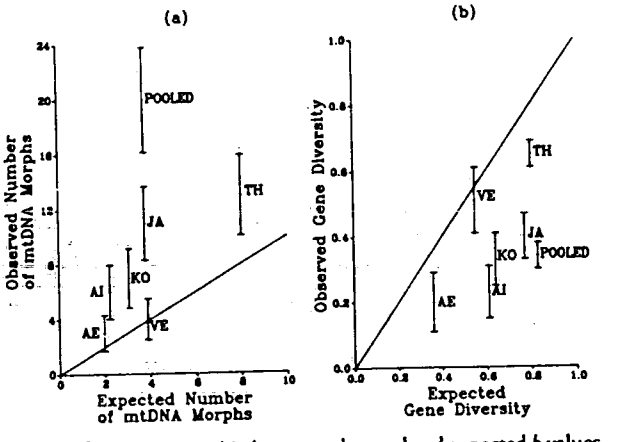


Figure 1 Relationship between observed and expected k values, on the basis of observed H in six Asian populations (*a*) and relationship between observed and expected H, on the basis of observed k (*b*). JA = Japanese; AI = Ainu; KO = Korean; AE = Aeta; VE = Vedda; TH = Tharu. The pooled data represent the total of the first five populations (excluding the Tharu sample). The vertical bars represent ± SE deviations.

Table 4

Observed and Expected k_r Values in Six Asian Populations

Population (n)	r	Observed Frequency	Expected Frequency of $k_r \pm$ SE, P	
			Based on θ_H	Based on θ_k
Japanese (74)	1	6	.68 ± .83, $<10^{-4}$	3.24 ± 1.80, .110
	2	2	.34 ± .59, .047	1.57 ± 1.25, .465
Ainu (48)	1	4	.31 ± .56, $<10^{-3}$	1.57 ± 1.25, .075
	2	1	.16 ± .40, .147	.78 ± .88, .540
Korean (64)	1	3	.51 ± .72, .015	1.77 ± 1.33, .263
	2	1	.26 ± .51, .228	.88 ± .94, .584
Aeta[a] (37)	2	2	.13 ± .37, .008	.29 ± .54, .035
Vedda[b] (20)	1	2	1.15 ± 1.07, .319	1.20 ± 1.09, .336
Tharu (91)	1	5	1.91 ± 1.38, .045	3.80 ± 1.95, .332
	2	5	.95 ± .97, .003	1.84 ± 1.36, .039
Pooled[c] (243)	1	8	.52 ± .72, $<10^{-7}$	4.91 ± 2.22, .124
	2	6	.26 ± .51, $<10^{-6}$	2.42 ± 1.55, .037

[a] No singleton morph observed.
[b] No doubleton morph observed.
[c] Excluding the Tharu population.

ple). The P shown in this table represents the probability of observing k_r or more morphs for $r = 1$ (or 2), given the respective expected value of k_r, in which a Poisson distribution of k_r is assumed (Chakraborty and Griffiths 1982).

It is clear from table 4 that the observed excess k is predominantly due to an excess in the observed number of rare morphs. When θ is estimated from the observed H, all populations except the Aetas of the Philippines show a significant ($P < .05$) excess of rare morphs, and this is more conspicuous in the pooled sample. When θ is estimated from k, the excess in the number of rare morphs does not reach significance in most cases, but the observed values of k_r are consistently larger than the expected values of k_r, even though for this estimator of θ the expected total k values equal the respective observed total k values. It can therefore be concluded that, irrespective of the choice of the parameter θ, there is an excess of rare mtDNA morphs in most of these samples, in contrast with the expectations based on the assumption that each sample is drawn from an equilibrium-homogenous population. Since this departure from the expectation is more profound in the pooled sample, it may be suspected that the most likely explanation is that the majority of the populations are heterogeneous in nature, so that they do not represent single homogeneous panmictic populations.

Discussion and Conclusion

The above findings indicate that mtDNA variability, as detected by the number of different RFLP morphs, is too high for the observed H in this genome in most of the Asian populations sampled. This excess is mainly due to the increased frequencies of rare morphs and is more conspicuous in the pooled sample than in the small populations (Aeta and Vedda). These observations are parallel to the findings reported for Amerindian populations of South and Central America (Chakraborty et al. 1988), findings based on isozyme polymorphisms. Also note that these observations are completely parallel to Whittam et al.'s (1986) findings based on mtDNA data. This parallelism indicates that the present results on the mtDNA genome are probably not an artifact of stochastic sampling errors due to data for a single locus. Single-locus data generally would have produced a deviation in the other direction, since Zouros (1979) has shown that the single-locus estimate of θ_H is generally larger than that of θ_k, just the opposite of the present findings.

There are several other possible explanations for the excess number of alleles produced through an apparent increase in rare ones. Ohta (1973, 1976) argued that this can occur in the presence of slightly deleterious mutations, while Nei et al. (1975), Maruyama and Fuerst (1984, 1985), and Watterson (1984) have shown that recent population bottlenecks can also cause a disproportionate increase in the number of variant alleles, in contrast with the expectations based on observed H. Whittam et al. (1986) ascribed the observed deviations of the mtDNA allele frequency distributions to such causes, apparently not realizing that most of their population data are not really from any anthropologically well-defined populations (e.g., at least two of their

five sampling units are quite heterogeneous: Asians [group II] and Europe, North Africa, and Middle East [group IV]). There are at least two lines of evidence suggesting that hidden heterogeneity is the most likely explanation in such contexts. First, the analysis performed in the present paper clearly demonstrates that the pooled sample (which is obviously an amalgamated sample consisting of five anthropologically distinct populations) shows a more profound deviation in the direction seen in each individual population. Whittam et al. (1986) also found this, but they failed to interpret their findings accordingly. Second, Chakraborty et al. (1988) have demonstrated that the isozyme loci reveal the effect of amalgamation when data from relatively homogeneous populations are sequentially agglomerated, producing an excess of allele numbers (total as well as rare), whose extent depends both on how many subpopulations are amalgamated into a sample and on how distinct the subpopulations are (measured by the average genetic distance between them).

In addition, for several of these populations there is independent evidence of internal substructuring. For example, Nei and Imaizumi (1966*a*, 1966*b*) showed that the coefficient of gene differentiation within the different subpopulations in Japan is not negligible. Since their analysis was based on blood-group genes, their results cannot be directly compared with the present findings. Analysis of data on 32 isozyme loci from a large sample of Japanese (Neel et al. 1988), however, corroborates the present finding, since in that survey 5.53 alleles/locus were observed whereas the observed average H (.087) at these loci predicts only 2.02 ± 0.18 alleles/locus (Chakraborty, in press). A phylogenetic analysis of mtDNA RFLP variation, conducted by Horai and Matsunaga (1986), indicates that at least two distinct lineages of mtDNA exist in the Japanese population. There is no such independent documentation of internal fragmentation of the other populations analyzed here. Nevertheless, the isozyme studies of Omoto (1972) in the Ainu population of Japan, of Omoto et al. (1978) in the Aeta population of the Philippines, and of Ellepola and Wikramanayake (1986) in the Vedda population of Sri Lanka suggest internal substructuring as well as recent gene migration from the outside, both of which should have a net effect similar to that of amalgamation. The typological classification of the Tharu mtDNA morphs also indicated some evidence of foreign-gene admixture (Brega et al. 1986), suggesting that the present evidence of heterogeneity in the Tharu population is not inconsistent with what is known about the genetic structure of this population.

Finally, methodological differences between the present study and that of Whittam et al. (1986) should be noted; the earlier study is based on the estimate of θ based on the total number of alleles (k), which is quite error prone in the presence of hidden substructuring within a population (Chakraborty et al. 1988; Chakraborty and Schwartz 1990). On the contrary, use of θ_H as an estimator of θ circumvents this problem. In addition, this approach avoids tedious simulations employed by Whittam et al. (1986), since I can use analytical expressions such as equations (5) and (7) to conduct the relevant statistical tests. In principle the present technique can be applied to other broad sets of data; however, at present, the lack of uniformity of laboratory protocols (as mentioned earlier) makes it difficult to use the reported data to establish allelic homologies as a means of limiting such analysis. Nevertheless, this analysis shows that the effect of hidden heterogeneity within a population can be detected with allele frequency data from a single locus, provided that substantial genetic variation exists at the locus. Therefore, the hypervariable minisatellite loci—i.e., the VNTR loci—should be useful in revealing the genetic structure of populations, because such loci generally exhibit substantially larger variability than that seen in either the RFLP or the isozyme loci.

Acknowledgments

This work was supported by U.S. Public Health Service research grant GM 41399 from the National Institutes of Health. I thank Drs. W. J. Schull and M. Nei for their helpful suggestions during the conduct of this study.

References

Blanc H, Chen K-H, D'Amore MA, Wallace DC (1983) Amino acid change associated with the major polymorphic Hinc II site of Oriental and Caucasian mitochondrial DNAs. Am J Hum Genet 35:167–176

Brega A, Gardella R, Semino O, Morpurgo G, Astaldi Ricotti GB, Wallace DC, Santachiara Benerecetti AS (1986) Genetic studies on the Tharu population of Nepal: restriction endonuclease polymorphisms of mitochondrial DNA. Am J Hum Genet 39:502–512

Brown WM (1980) Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc Natl Acad Sci USA 77:3605–3609

Cann RL, Brown WM, Wilson AC (1982) Evolution of human mitochondrial DNA: a preliminary report. In: Bonne-Tamir B, Cohen P, Goodman RN (eds). Human Genetics. Part A: The Unfolding Genome. Alan R. Liss, New York, pp 157–166

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cann RL, Wilson AC (1983) Length mutations in human mitochondrial DNA. Genetics 104:699–711

Chakraborty R. Genetic profile of cosmopolitan populations: effects of hidden subdivision. Anthropol Anz (in press)

Chakraborty R, Griffiths RC (1982) Correlation of heterozygosity and number of alleles in different frequency classes. Theor Popul Biol 21:205–218

Chakraborty R, Schwartz RJ (1990) Selective neutrality of surname distribution in an immigrant Indian community of Houston, Texas. Am J Hum Biol 2:1–15

Chakraborty R, Smouse PE, Neel JV (1988) Population amalgamation and genetic variation: observations on artifically agglomerated tribal populations of Central and South America. Am J Hum Genet 43:709–725

Denaro M, Blanc H, Johnson MJ, Chen KH, Wilmsen E, Cavalli-Sforza LL, Wallace DC (1981) Ethnic variation in *Hpa*I endonuclease cleavage patterns of human mitochondrial DNA. Proc Natl Acad Sci USA 78:5768–5772

Ellepola SB, Wikramanayake ER (1986) A genetic study of the Veddas and the Sinhalese. Ceylon J Med Sci 29:1–21

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87–112

Fuerst PA, Chakraborty R, Nei M (1977) Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics 86:455–483

Harihara S, Saitou N, Hirai M, Gojobori T, Park KS, Misawa S, Ellopola SB, et al (1988) Mitochondrial DNA polymorphism among five Asian populations. Am J Hum Genet 43:134–143

Horai S, Gojobori T, Matsunaga E (1987) Evolutionary implications of mitochondrial DNA polymorphism in human populations. Hum Genet 74:177–181

Horai S, Matsunaga E (1986) Mitochondrial DNA polymorphism in Japanese. II. Analysis with restriction enzymes of four or five base pair recognition. Hum Genet 72:105–117

Johnson MJ, Wallace DC, Ferris SD, Rattazzi MC, Cavalli-Sforza LL (1983) Radiation of human mitochondrial DNA types analyzed by restriction endonuclease cleavage patterns. J Mol Evol 19:255–271

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738

Li WH, Nei M (1975) Drift variances of heterozygosity and genetic distance in transient states. Genet Res 25:229–248

Maruyama T, Fuerst PA (1984) Population bottlenecks and non-equilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. Genetics 108:745–763

——— (1985) Population bottlenecks and non-equilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. Genetics 111:691–703

Neel JV, Satoh C, Smouse P, Asakawa J-I, Takahashi N, Goriki K, Fujita M, et al (1988) Protein variants in Hiroshima and Nagasaki: tales of two cities. Am J Hum Genet 43:870–893

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590

——— (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Imaizumi Y (1966*a*) Genetic structure of human populations. I. Local differentiation of blood group frequencies in Japan. Heredity 21:9–25

——— (1966*b*) Genetic structure of human populations. II. Differentiation of blood group gene frequencies among isolated populations. Heredity 21:183–190

Nei M, Maruyama T, Chakraborty R (1975) The bottleneck effect and genetic variability in populations. Evolution 29:1–10

Nei M, Roychoudhury AK (1974) Sampling variance of heterozygosity and genetic distance. Genetics 76:379–390

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246:96–98

——— (1976) Role of slightly deleterious mutations in molecular evolution and polymorphism. Theor Popul Biol 10:254–275

Omoto K (1972) Polymorphism and genetic affinities of the Ainu of Hokkaido. Hum Biol Oceania 1:278–288

Omoto K, Misawa S, Harada S, Sumpaico JS, Medado PM, Ogonuki H (1978) Population genetic studies of the Philippine Negritos. I. A pilot survey of red cell enzyme and serum protein groups. Am J Hum Genet 30:190–201

Watterson GA (1978) The homozygosity test of neutrality. Genetics 88:405–417

——— (1984) Allele frequencies after a bottleneck. Theor Popul Biol 26:387–407

Whittam TS, Clark AG, Stoneking M, Cann RL, Wilson AC (1986) Allelic variation in human mitochondrial genes based on patterns of restriction site polymorphism. Proc Natl Acad Sci USA 83:9611–9615

Zouros E (1979) Mutation rates, population sizes, and amounts of electrophoretic variation of enzyme loci in natural populations. Genetics 92:623–646

of the text, deal with karyotype interpretation, with a brief synopsis of material from the International System for Cytogenetic Nomenclature (ISCN 1985); with cell culture, maintenance, and storage; and with chromosomes in clinical medicine. The authors have reproduced (from ISCN 1985) mitotic and meiotic ideograms and have provided classification and nomenclature of chromosomes. They briefly touch on the maintenance and storage of cell cultures and provide a cursory look at the use of chromosomes for clinical medicine.

Overall, the information on tissue culture technique, chromosome preparation, banding techniques, and specialized techniques makes this a worthwhile book for any cytogenetics laboratory director, technologist, or student.

STUART SCHWARTZ

University of Maryland School of Medicine
Baltimore

0002-9297/91/4801-0024$02.00

Am. J. Hum. Genet. 48:173–174, 1991

DNA Technology and Forensic Science. Banbury report 32. Edited by J. Ballantyne, C. Sensabaugh, and J. Witkowski. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 1989. Pp. xiii + 368. $95.00.

The impact of academic research on society generally has a lag time, which is not often as short as the one being witnessed in the case of the development of DNA technology. It is the societal impact of this technology that constitutes the theme of this volume, the proceedings of a conference held at the Banbury Center, Cold Spring Harbor Laboratory. It represents the opinions and views of panelists as diverse as molecular biologists, legal authorities, forensic scientists, and policy analysts. The editors rightly surmised that the purpose of this conference was to address some of the key questions surrounding the legal application of DNA techniques, and hence the volume's focus is on the policy issues and not on the DNA technology per se.

Organized in five sections, this volume starts with five essays on legal and social issues arising in the use of genetic information in forensic applications. Section 2 addresses, through six presentations, the question of admissibility and interpretation of DNA data in a legal setting. The third section deals with the subject of the transfer of DNA technology to forensic laboratories and describes some general features of the implementation of the technology in a forensic setting. The five articles in the fourth section address the current and potential future approaches to the use of DNA techniques in identity determinations, and the last section discusses the relevant issues of establishment, construction, and management of a DNA data bank. Each section contains the open discussions by the panelists and by other participants who voiced their opinions on the main articles, as well as on relevant associated problems.

The style of the entire volume is truly "freewheeling, open, and informative" (p. xii). As a result, it contains a wealth of information regarding the technical aspects of the use of recombinant-DNA research, the opportunity it provides to the legal expert in handling criminal cases, and scenarios through which possible misuse of "pseudogenetic information" (to quote Motulsky [p. 3]) might occur in the unregulated application of this technology. The organizers and editors should be congratulated for providing this forum, from which future advancement of this important interface of science and society will surely evolve.

Without denigrating the importance and relevance of the main presentations, it should be stated that there are some aspects of the application of DNA technology in forensic science that did not receive the attention they deserve. For example, I do not believe that the mechanisms of generation and maintenance of new variation at the hypervariable loci is unrelated to the population issues raised by Lander. Although Jeffrey's article briefly addresses some of them, there is no explicit statement as to how departures of single-locus and multiple-locus genotype frequencies from their respective equilibrium values can be related to factors such as incomplete resolution of similar-size alleles and the undetectability of alleles of very small (or large) sizes. While it may be argued that such issues fall under the category of technology development and hence that discussion of them was not the focus of this presentation, some of the policy statements hinge critically on them. Therefore, some attempt to cover these aspects would have been at least academically profitable to the practitioners of this area. Furthermore, the legal community should be informed about the multidisciplinary nature of human genetic research, in view of which the "general acceptability" (p. 75) criterion of a scientific method (or concept) in a legal setting should be reassessed. It may not be prudent to state that time has come to reevaluate the Frye test, because that test does not seem to be tangible any longer.

Even with these limitations, this volume certainly will have its place in the history of forensic applications of new scientific developments, and it will definitely spawn other volumes on this subject. For the futurists, Judge Boggs's closing article carries several important suggestions that are worth noting. It is almost an axiom that DNA technology is potentially a very powerful tool in the context of forensic science. The high degree of sensitivity and specificity offered by this technic is unparalleled in comparison with the other available genetic tools. At the same time, the possibility of technical errors, as well as other concerns (such as marker independence and band-width reading), are substantive, and the witness stand is not an appropriate venue to resolve these concerns. Experiences gathered through extensive human population genetic studies can easily deal with such issues under

a scientific forum, and work is in progress in this direction in many institutions. Therefore, these concerns alone should not be used to label the application of DNA in forensic science "hasty" simply because of the current "expert" witnesses' inabilities to deal with them. Just as in the case of a laboratory that performs the DNA tests, an assurance of the quality of the expertise to examine these population genetic issues is urgently needed. Several times, the attorneys as well as the jury bench are mesmerized by numbers brought forward to them by "bootstrapping" and "jackknifing," and hence the *mojo* aspect of DNA is depicted. However, the same argument can be presented in simpler terms without distorting the scientific basis, and this is crucially needed in the dissemination of knowledge to the layman.

In summary, the organizers of this conference must be congratulated for making this happen, and the editors deserve credit for bringing these proceedings to the public. While there are distinct possibilities of citations of these writings by the pros and cons of DNA technology in a court setting, there is in this volume enough food for thought to satisfy the medicolegal geneticists who derive pleasure from doing science and properly advocating the utility of science in solving public problems.

RANAJIT CHAKRABORTY
Center for Demographic and Population Genetics
University of Texas Graduate School of Biomedical Sciences
Houston

0002-9297/91/4801-0025$02.00

Am. J. Hum. Genet. 48:1/4–175, 1990

Back Door to Eugenics. By Troy Duster. New York and London: Routledge, Chapman Hall, 1990. Pp. 201. $39.95 (cloth); $13.95 (paper).

The author is a sociologist at the University of California, Berkeley. He addresses various social issues raised by the application of genetic testing and screening. As the title of the book implies, Duster is much worried that our new technologic capabilities carried out in the name of health, prevention, and treatment of genetic disease may bring back the "old eugenics," with its emphasis on race, class, and "inferior" human beings. In a far-ranging discussion of many recent developments, he points out the important role that the social climate at a given historical period has in shaping scientific research and public policy. The author claims that costly high technology with emphasis on genetics is encouraged by "the medical establishment, research interests, the biotechnology lobby and insurance companies" and contrasts these trends with the need for less expensive programs, such as the provision of better and more universal prenatal care, which could have a marked impact on public health. While Duster is correct in deploring our society's unwillingness to use its resources for low-cost health care that would benefit millions of people, it is unlikely that the funds currently used for high-technology expenditures would be redirected.

The author cites several examples of his general thesis. Faculties involved in genetic counselor training (as at the University of California, Berkeley, program for which he was an adviser) stress high grades and technical preparation in applicants over humane and empathic skills. The author does not consider that high academic ability does not preclude humanistic qualities. The general success of Tay-Sachs screening programs as compared with the failure of sickle cell screening programs is explained by participation of the Jewish community at all levels, as contrasted with failure to involve the black community in the screening programs. The author documents that prenatal diagnosis is more widely used by women with higher incomes and adduces a variety of plausible reasons, including the role of the state health department bureaucracy which is not well attuned to poorer and minority populations.

The stress on genetic factors in multifactorial disorders is deplored, since emphasis on hereditary determinants will deflect attention from searching for and dealing with the many environmental factors in common diseases. Geneticists would counter that detection of those at high risk will identify those persons who will benefit maximally from treatments designed to manipulate the environmental causes of those conditions. Furthermore, the author (following Lewontin et al. in *Not in our Genes*) takes issue with various data that ascribe a genetic basis to IQ, schizophrenia, and propensity to crime. As an interesting example, the data showing a high frequency of *low* IQs among Jews in the early part of this century are contrasted with more recent findings which demonstrate a significantly *higher* mean IQ as compared with the Caucasian U.S. population and are cited as an example of the futility of this kind of investigation.

As might be expected from the author's orientation, he gives no attention to the internal logic of modern biomedical science as compared with the role of social forces in shaping research and practice. Once medical geneticists learned about DNA and restriction enzymes, the application of this molecular technology to monogenic hereditary diseases was inevitable and had little to do with social forces. Once it is realized that undefined genetic factors play a role in many common diseases, the obvious next step is the utilization of the new biochemical and molecular methodology in attempting to elucidate the role of *specific* genes that predispose to these diseases. Again, the development of science per se leads to such investigations. Similarly, various discoveries in molecular genetics and the neurosciences now allow us to search for specific genes that may affect normal and abnormal behavior.

The book is valuable in depicting the many forces that may shape research and practices in medical genetics. It reminds us that, with the best intentions, abuses in the name of genetics could enter by the back door. Many facts cited in this book point out the importance of ongoing dialogue regarding these

of the text, deal with karyotype interpretation, with a brief synopsis of material from the International System for Cytogenetic Nomenclature (ISCN 1985); with cell culture, maintenance, and storage; and with chromosomes in clinical medicine. The authors have reproduced (from ISCN 1985) mitotic and meiotic ideograms and have provided classification and nomenclature of chromosomes. They briefly touch on the maintenance and storage of cell cultures and provide a cursory look at the use of chromosomes for clinical medicine.

Overall, the information on tissue culture technique, chromosome preparation, banding techniques, and specialized techniques makes this a worthwhile book for any cytogenetics laboratory director, technologist, or student.

STUART SCHWARTZ

University of Maryland School of Medicine
Baltimore

0002-9297/91/4801-0024$02.00

Am. J. Hum. Genet. 48:173–174, 1991

DNA Technology and Forensic Science. Banbury report 32. Edited by J. Ballantyne, C. Sensabaugh, and J. Witkowski. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 1989. Pp. xiii + 368. $95.00.

The impact of academic research on society generally has a lag time, which is not often as short as the one being witnessed in the case of the development of DNA technology. It is the societal impact of this technology that constitutes the theme of this volume, the proceedings of a conference held at the Banbury Center, Cold Spring Harbor Laboratory. It represents the opinions and views of panelists as diverse as molecular biologists, legal authorities, forensic scientists, and policy analysts. The editors rightly surmised that the purpose of this conference was to address some of the key questions surrounding the legal application of DNA techniques, and hence the volume's focus is on the policy issues and not on the DNA technology per se.

Organized in five sections, this volume starts with five essays on legal and social issues arising in the use of genetic information in forensic applications. Section 2 addresses, through six presentations, the question of admissibility and interpretation of DNA data in a legal setting. The third section deals with the subject of the transfer of DNA technology to forensic laboratories and describes some general features of the implementation of the technology in a forensic setting. The five articles in the fourth section address the current and potential future approaches to the use of DNA techniques in identity determinations, and the last section discusses the relevant issues of establishment, construction, and management of a DNA data bank. Each section contains the open discussions by the panelists and by other participants who voiced their opinions on the main articles, as well as on relevant associated problems.

The style of the entire volume is truly "freewheeling, open, and informative" (p. xii). As a result, it contains a wealth of information regarding the technical aspects of the use of recombinant-DNA research, the opportunity it provides to the legal expert in handling criminal cases, and scenarios through which possible misuse of "pseudogenetic information" (to quote Motulsky [p. 3]) might occur in the unregulated application of this technology. The organizers and editors should be congratulated for providing this forum, from which future advancement of this important interface of science and society will surely evolve.

Without denigrating the importance and relevance of the main presentations, it should be stated that there are some aspects of the application of DNA technology in forensic science that did not receive the attention they deserve. For example, I do not believe that the mechanisms of generation and maintenance of new variation at the hypervariable loci is unrelated to the population issues raised by Lander. Although Jeffrey's article briefly addresses some of them, there is no explicit statement as to how departures of single-locus and multiple-locus genotype frequencies from their respective equilibrium values can be related to factors such as incomplete resolution of similar-size alleles and the undetectability of alleles of very small (or large) sizes. While it may be argued that such issues fall under the category of technology development and hence that discussion of them was not the focus of this presentation, some of the policy statements hinge critically on them. Therefore, some attempt to cover these aspects would have been at least academically profitable to the practitioners of this area. Furthermore, the legal community should be informed about the multidisciplinary nature of human genetic research, in view of which the "general acceptability" (p. 75) criterion of a scientific method (or concept) in a legal setting should be reassessed. It may not be prudent to state that time has come to reevaluate the Frye test, because that test does not seem to be tangible any longer.

Even with these limitations, this volume certainly will have its place in the history of forensic applications of new scientific developments, and it will definitely spawn other volumes on this subject. For the futurists, Judge Boggs's closing article carries several important suggestions that are worth noting. It is almost an axiom that DNA technology is potentially a very powerful tool in the context of forensic science. The high degree of sensitivity and specificity offered by this technic is unparalleled in comparison with the other available genetic tools. At the same time, the possibility of technical errors, as well as other concerns (such as marker independence and band-width reading), are substantive, and the witness stand is not an appropriate venue to resolve these concerns. Experiences gathered through extensive human population genetic studies can easily deal with such issues under

a scientific forum, and work is in progress in this direction in many institutions. Therefore, these concerns alone should not be used to label the application of DNA in forensic science "hasty" simply because of the current "expert" witnesses' inabilities to deal with them. Just as in the case of a laboratory that performs the DNA tests, an assurance of the quality of the expertise to examine these population genetic issues is urgently needed. Several times, the attorneys as well as the jury bench are mesmerized by numbers brought forward to them by "bootstrapping" and "jackknifing," and hence the *mojo* aspect of DNA is depicted. However, the same argument can be presented in simpler terms without distorting the scientific basis, and this is crucially needed in the dissemination of knowledge to the layman.

In summary, the organizers of this conference must be congratulated for making this happen, and the editors deserve credit for bringing these proceedings to the public. While there are distinct possibilities of citations of these writings by the pros and cons of DNA technology in a court setting, there is in this volume enough food for thought to satisfy the medicolegal geneticists who derive pleasure from doing science and properly advocating the utility of science in solving public problems.

RANAJIT CHAKRABORTY
Center for Demographic and Population Genetics
University of Texas Graduate School of Biomedical Sciences
Houston

0002-9297/91/4801-0025$02.00

Am. J. Hum. Genet. 48:1/4–175, 1990

Back Door to Eugenics. By Troy Duster. New York and London: Routledge, Chapman Hall, 1990. Pp. 201. $39.95 (cloth); $13.95 (paper).

The author is a sociologist at the University of California, Berkeley. He addresses various social issues raised by the application of genetic testing and screening. As the title of the book implies, Duster is much worried that our new technologic capabilities carried out in the name of health, prevention, and treatment of genetic disease may bring back the "old eugenics," with its emphasis on race, class, and "inferior" human beings. In a far-ranging discussion of many recent developments, he points out the important role that the social climate at a given historical period has in shaping scientific research and public policy. The author claims that costly high technology with emphasis on genetics is encouraged by "the medical establishment, research interests, the biotechnology lobby and insurance companies" and contrasts these trends with the need for less expensive programs, such as the provision of better and more universal prenatal care, which could have a marked impact on public health. While Duster is correct in deploring our society's unwillingness to use its resources for low-cost health care that would benefit millions of people, it is unlikely that the funds currently used for high-technology expenditures would be redirected.

The author cites several examples of his general thesis. Faculties involved in genetic counselor training (as at the University of California, Berkeley, program for which he was an adviser) stress high grades and technical preparation in applicants over humane and empathic skills. The author does not consider that high academic ability does not preclude humanistic qualities. The general success of Tay-Sachs screening programs as compared with the failure of sickle cell screening programs is explained by participation of the Jewish community at all levels, as contrasted with failure to involve the black community in the screening programs. The author documents that prenatal diagnosis is more widely used by women with higher incomes and adduces a variety of plausible reasons, including the role of the state health department bureaucracy which is not well attuned to poorer and minority populations.

The stress on genetic factors in multifactorial disorders is deplored, since emphasis on hereditary determinants will deflect attention from searching for and dealing with the many environmental factors in common diseases. Geneticists would counter that detection of those at high risk will identify those persons who will benefit maximally from treatments designed to manipulate the environmental causes of those conditions. Furthermore, the author (following Lewontin et al. in *Not in our Genes*) takes issue with various data that ascribe a genetic basis to IQ, schizophrenia, and propensity to crime. As an interesting example, the data showing a high frequency of *low* IQs among Jews in the early part of this century are contrasted with more recent findings which demonstrate a significantly *higher* mean IQ as compared with the Caucasian U.S. population and are cited as an example of the futility of this kind of investigation.

As might be expected from the author's orientation, he gives no attention to the internal logic of modern biomedical science as compared with the role of social forces in shaping research and practice. Once medical geneticists learned about DNA and restriction enzymes, the application of this molecular technology to monogenic hereditary diseases was inevitable and had little to do with social forces. Once it is realized that undefined genetic factors play a role in many common diseases, the obvious next step is the utilization of the new biochemical and molecular methodology in attempting to elucidate the role of *specific* genes that predispose to these diseases. Again, the development of science per se leads to such investigations. Similarly, various discoveries in molecular genetics and the neurosciences now allow us to search for specific genes that may affect normal and abnormal behavior.

The book is valuable in depicting the many forces that may shape research and practices in medical genetics. It reminds us that, with the best intentions, abuses in the name of genetics could enter by the back door. Many facts cited in this book point out the importance of ongoing dialogue regarding these

Genetic Data Analysis. By Bruce S. Weir. Sunderland, Mass.: Sinauer Associates, 1990. Pp. 377 + xii. $48.00 (cloth); $27.00 (paper).

One major impact that recent developments in molecular biology have had on population genetics is that the evolutionary processes can now be studied at the molecular level. The purpose of this book is to illustrate that classic statistical methodologies developed for population genetic studies with serological and isozyme markers can also be applied to molecular data, with suitable modifications. In eight chapters, accompanied by tables of standard statistical distributions, a few computer programs, and notes on data randomization, the author covers topics of interest that justify the title of the volume, particularly considering its subtitle—*Methods for Discrete Population Genetic Data.*

I undertake this reviewing task with trepidation. Being a statistician by training, I obviously support the author's view that rigorous statistics must accompany population genetic data reporting, even when the methods of gathering such data are as detailed as the molecular techniques currently in use. Therefore, the need of a book on this topic is well justified, to update the R. C. Elandt-Johnson and O. Kempthorne texts published more than 2 decades back. On the other hand, being a data analyst for >20 years, I am firmly convinced that the role of statistics both in applied science in general and in biology, in particular should be to provide objective appraisals of certain hypotheses and to check whether the observed data are in conformity with such hypotheses. Mathematical models of data analysis and statistical methods of estimation and hypothesis testing must, therefore, strictly adhere to assumptions that are intrinsic to the biological processes that led to the observed data. Mathematical elegance and statistical sophistication is not a compromise of these premises, but mathematical rigor without valid assumptions can only introduce obscurity and intimidate practical scientists. An elegant advocate of statistics, Professor P. C. Mahalanobis, has argued that a statistician must approach an applied problem through four phases: (1) formulation of the problem, (2) data-collection design appropriate for answering the question, (3) methods for data analysis, and (4) inference. Statistical methods should be data driven, and avoidance of any of the above four phases can only make them infelicitous.

Genetic Data Analysis, I suspect, has failed to guide the readers through the above four phases. It describes a wide variety of methods covering gene frequency estimation, computations of disequilibria indices and diversity measures, analysis of population structure and data gathered across generations, analysis of molecular data on restriction sites and sequences, and, finally, phylogeny reconstruction. Although the introductory chapter briefly presents the nature and sampling procedures attendant to genetic data, clear statements regarding either the questions that can be answered by using the methods outlined or the biological interpretations of the numbers generated by these methods are critically lacking in the presentation. Furthermore, the readers are often kept at bay with regard to (*a*) the assumptions underlying the methods and/or (*b*) their adequacy in the context of data from natural populations. In my opinion, this is a serious flaw, and readers with less statistical expertise may have difficulty in understanding the biological meaning of the statistics proposed by the author. Some may also wonder about the legitimacy of assumptions, such as "replicate subpopulations," in the context of data on the evolution of a set of specific natural populations.

The book is also marred by a score of mistakes, some gross and annoying, others subtle but critical for young readers who accept formulae as ex-cathedra statements. The errors and inconsistencies are too numerous to list exhaustively. However, some are noteworthy. For example, figure 1.1 (p. 3) mislabels the heterozygotes and one of the homozygotes for the first six loci. Other errors include eight (not nine) possible

sample values of a binomial distribution with eight trials (p. 29), wrong formula (e.g., p. 81—last equation), wrong source codes in the program (p. 309), missing references (p. 272), and wrong internal chapter citations (e.g., p. 256). Probably most errors reflect a hastiness in the preparation of this book.

This is also a somewhat opinionated presentation. For example, an 11-page discussion on the possibility of Mendel's data pruning is unwarranted. A 54-page-long description of the variance-component approach to diversity and population structure analysis, without any mention of the adequacy of the replicate-subpopulation assumption for nonexperimental natural populations, is unfair to students who would like to learn from this book the biology of population structure. The depth of treatment is also quite uneven from chapter to chapter. It does not take long to realize that disequilibria statistics are the forte of this author, while the topics covered in the last three chapters are quite superficial and depend often on secondary sources of citation. As a result, readers may be misguided by the general conclusions of these chapters. For example, the discussion on the comparative performances of different methods of phylogeny reconstruction (chap. 8, p. 272) erroneously concludes that the UPGMA and Fitch-Margoliash methods do not differ much in the presence of a molecular clock, which is demonstrated to be wrong in some work published in *MBE*.

In summary, it is difficult to recommend this book in a self-study course on genetic data analysis. This edition imposes an onerous task on tutors who would like to equip their students with state-of-the-art methods of handling genetic data, because of both its lack of fair coverage of these methods and numerous errors of presentations. The small typeface of the prints also makes the reading difficult.

RANAJIT CHAKRABORTY
Center for Demographic and Population Genetics
University of Texas Graduate School of Biomedical Sciences

Genetic Data Analysis. By Bruce S. Weir. Sunderland, Mass.: Sinauer Associates, 1990. Pp. 377 + xii. $48.00 (cloth); $27.00 (paper).

One major impact that recent developments in molecular biology have had on population genetics is that the evolutionary processes can now be studied at the molecular level. The purpose of this book is to illustrate that classic statistical methodologies developed for population genetic studies with serological and isozyme markers can also be applied to molecular data, with suitable modifications. In eight chapters, accompanied by tables of standard statistical distributions, a few computer programs, and notes on data randomization, the author covers topics of interest that justify the title of the volume, particularly considering its subtitle—*Methods for Discrete Population Genetic Data.*

I undertake this reviewing task with trepidation. Being a statistician by training, I obviously support the author's view that rigorous statistics must accompany population genetic data reporting, even when the methods of gathering such data are as detailed as the molecular techniques currently in use. Therefore, the need of a book on this topic is well justified, to update the R. C. Elandt-Johnson and O. Kempthorne texts published more than 2 decades back. On the other hand, being a data analyst for >20 years, I am firmly convinced that the role of statistics both in applied science in general and in biology, in particular should be to provide objective appraisals of certain hypotheses and to check whether the observed data are in conformity with such hypotheses. Mathematical models of data analysis and statistical methods of estimation and hypothesis testing must, therefore, strictly adhere to assumptions that are intrinsic to the biological processes that led to the observed data. Mathematical elegance and statistical sophistication is not a compromise of these premises, but mathematical rigor without valid assumptions can only introduce obscurity and intimidate practical scientists. An elegant advocate of statistics, Professor P. C. Mahalanobis, has argued that a statistician must approach an applied problem through four phases: (1) formulation of the problem, (2) data-collection design appropriate for answering the question, (3) methods for data analysis, and (4) inference. Statistical methods should be data driven, and avoidance of any of the above four phases can only make them infelicitous.

Genetic Data Analysis, I suspect, has failed to guide the readers through the above four phases. It describes a wide variety of methods covering gene frequency estimation, computations of disequilibria indices and diversity measures, analysis of population structure and data gathered across generations, analysis of molecular data on restriction sites and sequences, and, finally, phylogeny reconstruction. Although the introductory chapter briefly presents the nature and sampling procedures attendant to genetic data, clear statements regarding either the questions that can be answered by using the methods outlined or the biological interpretations of the numbers generated by these methods are critically lacking in the presentation. Furthermore, the readers are often kept at bay with regard to (*a*) the assumptions underlying the methods and/or (*b*) their adequacy in the context of data from natural populations. In my opinion, this is a serious flaw, and readers with less statistical expertise may have difficulty in understanding the biological meaning of the statistics proposed by the author. Some may also wonder about the legitimacy of assumptions, such as "replicate subpopulations," in the context of data on the evolution of a set of specific natural populations.

The book is also marred by a score of mistakes, some gross and annoying, others subtle but critical for young readers who accept formulae as ex-cathedra statements. The errors and inconsistencies are too numerous to list exhaustively. However, some are noteworthy. For example, figure 1.1 (p. 3) mislabels the heterozygotes and one of the homozygotes for the first six loci. Other errors include eight (not nine) possible

sample values of a binomial distribution with eight trials (p. 29), wrong formula (e.g., p. 81—last equation), wrong source codes in the program (p. 309), missing references (p. 272), and wrong internal chapter citations (e.g., p. 256). Probably most errors reflect a hastiness in the preparation of this book.

This is also a somewhat opinionated presentation. For example, an 11-page discussion on the possibility of Mendel's data pruning is unwarranted. A 54-page-long description of the variance-component approach to diversity and population structure analysis, without any mention of the adequacy of the replicate-subpopulation assumption for nonexperimental natural populations, is unfair to students who would like to learn from this book the biology of population structure. The depth of treatment is also quite uneven from chapter to chapter. It does not take long to realize that disequilibria statistics are the forte of this author, while the topics covered in the last three chapters are quite superficial and depend often on secondary sources of citation. As a result, readers may be misguided by the general conclusions of these chapters. For example, the discussion on the comparative performances of different methods of phylogeny reconstruction (chap. 8, p. 272) erroneously concludes that the UPGMA and Fitch-Margoliash methods do not differ much in the presence of a molecular clock, which is demonstrated to be wrong in some work published in *MBE*.

In summary, it is difficult to recommend this book in a self-study course on genetic data analysis. This edition imposes an onerous task on tutors who would like to equip their students with state-of-the-art methods of handling genetic data, because of both its lack of fair coverage of these methods and numerous errors of presentations. The small typeface of the prints also makes the reading difficult.

RANAJIT CHAKRABORTY
Center for Demographic and Population Genetics
University of Texas Graduate School of Biomedical Sciences

Professional Ethics Report

Newsletter of the American Association for the Advancement of Science
Committee on Scientific Freedom & Responsibility
Professional Society Ethics Group

VOLUME V ✦ NUMBER 2 ✦ SPRING 1992

IN THE NEWS

This past February voters in Switzerland rejected a proposed referendum banning scientific experiments using animals. The ban proposed by the Swiss Animal Protection League would have tightened existing restrictions for obtaining a license to experiment on animals. The Swiss government, medical groups, and Swiss-based pharmaceutical companies all opposed the initiative, claiming that it could lead to the relocation of research facilities outside Switzerland. Another major concern was the ban's implications for commercial confidentiality. Allowing animal rights groups to challenge in court individual research projects presumably would require that companies release their detailed research plans. Another aspect of the proposal that worried researchers was a clause requiring the Swiss government to enact an animal experimentation law within five years. Failure to do so could ban animal research completely.

On April 22, the National Academy of Sciences issued its report on *Responsible Science: Ensuring the Integrity of the Research Process* (see In Print) by a special Panel on Scientific Responsibility and the Conduct of Research. The Panel made twelve recommendations, including the creation of an independent Scientific Integrity Advisory Board, which would serve as a clearinghouse for the exchange of information and experiences related to scientific misconduct and efforts to promote responsible research conduct, and one that called on scientific societies and journals to "provide and expand resources and forums to foster responsible research practices and to address misconduct in science and questionable research practices." The Panel also urged the government to adopt a common definition of misconduct in science and common policies and procedures for handling allegations of scientific misconduct. The report also draws a distinction between misconduct in science, which includes "fabrication, falsification or plagiarism in proposing, performing, or reporting research," and "questionable research practices." The latter includes "actions that violate traditional values of the research enterprise and that may be detrimental to the research process," but for which there is currently "neither broad agreement as to the seriousness of these actions nor any consensus on standards for behavior in such matters."

IN THE SOCIETIES

The revision of the American Psychological Association's "Ethical Principles of Psychologists" is an action item on its Council of Representatives' agenda for the APA's August 1992 convention meeting. The revision involves major changes from the current (1989) version. In particular, it involves a clear identification of the aspirational versus enforceable sections. Contact Stanley E. Jones, Director, Office of Ethics, APA, 750 First Street, NE, Washington, DC 20002-4242; (202) 336-5500.

At its May meeting, the Council of Scientific Society Presidents adopted several resolutions related to professional ethics. Among them was one on "The Role of Professional Societies in Setting Ethical Standards in Science," which urges scientific societies to "develop mechanisms to educate members regarding standards of research practice, the ethical conduct and reporting of science, and the traditions, values, and paradigms of the discipline." Two other resolutions endorsed recommendations contained in the National Academy of Sciences' report *Responsible Science: Ensuring the Integrity of the Research Process* — one calling for uniform federal policies and procedures for handling allegations of misconduct in science, and the other urging institutions to "assure both accusers and accused the fundamental elements of due process...." For more details, contact CSSP at (202) 872-4452.

CASES AND COMMENTARIES

In past issues, PER has used this section to print invited commentaries on hypothetical cases. In this issue, we present a case that is currently unfolding in federal court. In recent years, scientists and attorneys have engaged in heated arguments—inside and outside the courtroom—over the reliability of DNA fingerprinting in criminal cases. The debate has encompassed a number of issues: the techniques used to determine whether samples match, the statistical methods used to interpret a match, and the standards and practices of quality control of the laboratories that perform the analyses.

Connected to this more public debate is a series of less visible, yet volatile skirmishes that highlight tensions

between law and science and raise issues of professional ethics for both the legal and scientific communities. In order to focus more attention on these matters and their implications for the relationship between law and science, PER has prepared this essay and invited all of the parties involved in the different incidents it describes to share their positions on the controversies with our readers. Those contributions received follow this essay. Readers of PER are invited to send us their reactions to the essay as well as to the responses that follow it for publication in our next issue. These should be received by August 15.—Editor

In a motion for a new trial in the 1989 murder case, *U.S. v. Yee, et al.*, two New York defense attorneys–Barry Scheck and Peter Neufeld–have filed an affidavit with a U.S. District Court in Ohio alleging misconduct by law enforcement officials and scientists in disputes over the reliability of DNA fingerprinting in criminal cases. The two attorneys accuse federal and state law enforcement of efforts "to intimidate, harass, and deter expert witnesses from testifying against the FBI and other forensic DNA laboratories and publishing their views." They go on to describe "the most troubling aspect of the government's campaign [as] its direct interference in the publication and peer review process...to prevent and undermine the publication of scientific opinions law enforcement does not like." They also accuse several scientists of conflicts of interest in providing testimony on the reliability of DNA fingerprinting and in participating in the peer review process of scientific journals. In both published reports and additional motions and affidavits submitted to the Ohio court, law enforcement officials and scientists named in the defense motion have refuted the allegations.

The defense attorneys refer to a manuscript accepted in September 1991 for publication in *Science* in which two population geneticists–Richard Lewontin and Daniel Hartl–claimed that proponents of DNA fingerprinting have made unwarranted assumptions regarding the rarity of genetic patterns in populations. They concluded that DNA fingerprinting as currently practiced should not be admissible as evidence. Scheck and Neufeld report that James Wooley, an Assistant U.S. Attorney, telephoned Hartl, pressuring him to withdraw the paper. The co-authors refused to do so, and Lewontin wrote Wooley, accusing him of a "very serious breach of ethics" and "intimidation."[1] According to the affidavit, federal law enforcement officials, in association with scientists who dispute the claims made in the manuscript (the affidavit mentions C. Thomas Caskey and Ranajit Chakraborty), "succeeded in delaying publication..., altering the content of the article, and [getting] an unprecedented simultaneous publication of a rebuttal article which did not go through the same peer review process."

In a counter affidavit submitted to the court, Wooley accuses the defense affidavit of being a "vehicle through which to launch a vicious, mean-spirited and baseless attack on the character, ethics and actions of numerous FBI agents, prosecutors, and nationally prominent scientists,... who support the admissibility of DNA testing...." He confirms speaking with Hartl, but denies intimidating or threatening him. He states that after the conversation with Hartl, he "took no action relative to the publication of the paper."

The paper by Lewontin and Hartl as well as the rebuttal article were published in *Science*.[2] In that same issue,[3] the editor of *Science*, Daniel Koshland, is reported to have been "disturbed that the data did not support the paper's conclusions" and asked the coauthors to make revisions. Koshland denied having received any communications on the matter from government officials, although he did hear complaints from other scientists. He defended his decision to solicit a rebuttal article, which, contrary to the claims of Scheck and Neufeld, was also peer-reviewed, "to give a more balanced view of the subject," although the normal procedure followed by *Science* is to publish rebuttals in a subsequent issue and to give the authors of the original article an opportunity to respond.

Another example cited in the affidavit by Scheck and Neufeld is a paper criticizing the statistical methods used by forensic laboratories submitted to the *American Journal of Human Genetics* (*AJHG*) in November 1991 by Seymour Geisser, a University of Minnesota statistician. The paper had come to the attention of Stephen L. Redding, a prosecutor in the Office of the Hennipen County Attorney in Minnesota, where the paper's author was scheduled to testify as a witness for the defense at a DNA admissibility hearing in January 1992. Geisser received a fax from Redding demanding that he produce in court any manuscript he authored, whether accepted or under review, related to DNA fingerprinting. According to the defense affidavit, fifteen minutes after receiving that fax, the author received a fax from Charles Epstein, editor of *AJHG*, along with comments on his manuscript from three anonymous reviewers, one of whom strongly recommended against publication. This latter reviewer was later identified as Ranajit Chakraborty, a scientist who has testified for the prosecution in criminal cases and who coauthored the rebuttal article in *Science* referred to earlier.

The affidavit quotes from Epstein's letter that "since this work will certainly be used in court cases, the writing needs to be more careful, and the work must apply to the data in the way they are actually used." Scheck and Neufeld conclude that this series of events "provides compelling circumstantial evidence that the rules of journal and reviewer confidentiality were breached and that prosecutors meddled in the process." They continue, "journals are

Editor: Mark S. Frankel
Associate and Managing Editor: Amy Crumpton
Contributing Editor: Alexander Fowler
Editorial Board: William Anderson, Brian Boom, John Gardenier, Jonathan Knight, William Middleton
The *Professional Ethics Report* is published quarterly under the auspices of the Committee on Scientific Freedom and Responsibility and the Professional Society Ethics Group, American Association for the Advancement of Science, 1333 H Street, NW, Washington, DC 20005, (202)326-6798.

supposed to evaluate a...manuscript on its scientific merits, not based on the uses courts may make of a scientist's opinion." Epstein denies being influenced by anyone, and claims that he was unaware that Chakraborty was co-investigator on a Justice Department grant intended to establish scientific support for the FBI's statistical methods used in evaluating DNA evidence when he asked him to review the Geisser manuscript.[4]

These incidents raise potentially serious conflict of interest problems, as scientists with strong positions on the reliability of DNA fingerprinting and/or financial interests in companies associated with the technique are called on to testify in court or to review related journal manuscripts or grant applications. Indeed, Scheck and Neufeld claim that two scientists who gave expert testimony in a hearing related to the *Yee* case failed to disclose relevant financial ties. They argue that one of the scientists, Stephen P. Daiger of the Graduate School of Biomedical Sciences at the University of Texas at Houston, had a grant application on DNA typing for forensic applications with the National Institute of Justice that constitutes a "substantial financial bias and conflict of interest that should have been revealed to the court and counsel when he testified." They argue further that "testimony favorable to the FBI and the Justice Department would no doubt preserve or enhance the prospect of receiving [the] $300,000 NIJ grant."

In an affidavit submitted to the Ohio court, Daiger expresses irritation at "the criticisms and insults contained" in the defense document. He notes that he did not stand to profit personally since grant funds go directly to his institution for disbursement. He adds that "At no time during [the review] process was it stated or implied to me that my testimony in any legal case, or in any other forum, would have an effect on the review process."

The defense attorneys also accuse C. Thomas Caskey of Baylor College of Medicine of similar conflicts of interest. Like Daiger, Caskey had a grant application pending at the NIJ. The proposed research was intended to develop "a low cost, rapid analysis PCR based DNA profiling system...[that would] be a highly marketable product." Caskey also has an agreement to license his diagnostic techniques to Cellmark Diagnostics, Inc., a major DNA fingerprinting company. The defense affidavit argues that "the failure of the government and Dr. Caskey to disclose his $200,000 Justice Department grant deprived the defense counsel of an opportunity to explore fully and fairly the bias, conflict of interest, and financial motives of the government's most important witness." [Caskey's ties to Cellmark Diagnostics led to his resignation from a committee of the National Academy of Sciences established to report on the use of DNA technology in forensic science.]

In his own affidavit to the court, Caskey notes that he has "not personally profited from the NIJ grant or from the licensing of the newly developed personal identification technology to Cellmark...." He states further that he has "freely discussed the NIJ award and the Cellmark license with anyone who has asked about these matters and I would have responded truthfully...." if questioned about them at the hearing. Commenting more generally on potential conflicts of interest in court proceedings related to DNA fingerprinting, Caskey added that he did "not feel that any scientist associated with an institution which is a recipient of competitive federal (or state) financial support should automatically be disqualified from serving as a witness for the prosecution or defense in any case. To automatically exclude all...grantees from serving as expert witnesses will deprive all interested parties of the services of highly qualified scientists."

PER has received a letter from the Medical Faculty of Rijeka in Croatia requesting assistance in "resuscitating ethics" in their curriculum. Because of the war and poor economic conditions, the faculty is badly in need of literature. Below are edited excerpts from the letter:

> We are applying to you to help us with at least a small contribution. You could send us a book, magazines, or reference manual on medical ethics (or medical sociology) from your personal library, with your dedication or a message of solidarity on it.
>
> Your institution could send us some books on medical ethics, or subscribe to a magazine for us dealing with medical ethics (or forward us funds so that we can purchase materials).
>
> We ask you to support the idea of organizing a symposium on the ethical aspects of this war in Dubrovavik, once this unhappy war is over.
>
> We kindly thank you in advance for your understanding and support.
>
> *Prof. dr. Ivan Segota*
> *Head, Department of Social Sciences*
> *Medical Faculty of Rijeka*
> *O. Ban 22, 51000 Rijeka Croatia*

Contact Prof. Segota directly at the above address.

1. Christopher Anderson, "DNA fingerprinting discord," *Nature*, vol. 54, December 1991, p. 500.

2. R.C. Lewontin and Daniel L. Hartl, "Population Genetics in Forensic DNA Typing," *Science*, vol. 254, December 20, 1991, pp. 1745-50, and R. Chakraborty and Kenneth K. Kidd, "The Utility of DNA Typing in Forensic Work," *Ibid*, pp. 1735-39.

3. L. Roberts, "Was *Science* Fair to its Authors?" *Ibid*, p. 1722.

4. Christopher Anderson, "Conflict concerns disrupt panels, cloud testimony," *Nature*, vol. 355, February 27, 1992, p. 753-54.

Commentaries

The reported allegations[1] in the affidavit of Scheck and Neufeld requesting a new trial are nothing but unsubstantiated innuendos. I was never involved in the *Yee* case as an expert. Nevertheless, since my research deals with the subject of the effects of population substructure on genetic variation, I have read both Hartl's and Lewontin's reports on this subject, which are public records. I did not find their arguments convincing, nor did I find any support for their criticisms in our analysis of DNA typing data on populations around the world.[2] I had a draft version of a rebuttal to their criticisms prepared even before they submitted their paper to *Science*. I first learned of their *Sci-*

ence article in a court case around the third week of September 1991 where it was submitted as a defense exhibit. In the first week of October 1991, at the 8th International Congress of Human Genetics in Washington, DC, I found that a large number of human geneticists knew of their forthcoming article, and I expressed my intent to publish a rebuttal, should their article appear in *Science*. Upon return from the Congress (second week of October 1991), I received a telephone request from the editorial office of *Science* to submit my rebuttal. At that stage, I asked Dr. Kenneth Kidd to co-author the rebuttal,.... At no stage in the preparation of our paper, did I encourage *Science* to "delay" the publication of Lewontin and Hartl's paper, nor did I attempt to "alter the content" of their article....our rebuttal was reviewed by at least two reviewers. Scientists working in a forefront area should have no reason to believe that this procedure was "unprecedented" since simultaneous publications of controversial opinions are practiced by several leading journals.[3]

The account[1] of the review of Dr. Geisser's manuscript submitted to the *AJHG* is also a misrepresentation of actual events. It is true that at the request of the AJHG editor I acted as a reviewer of that article. I sent my comments to the journal office by the end of December 1991. According to the regular practice of reviewers, I have never discussed this review nor the paper with anyone. I was critical of the manuscript, because I believed that it was unprofessionally written, it contained several fatal errors, and it only reported parts of unpublished data from other laboratories without appropriate credit or consent of the data gatherers. My co-investigatorship in a NIJ grant had no connection with my reviewing this manuscript, and my review was to the point of evaluating a "scientific manuscript on its scientific merit." In the last week of January 1992, in a court case in Seattle, a defense exhibit produced all reviews and Dr. Geisser's correspondence with the journal editor, through which I knew that his manuscript was being used in the courts. Since under a court order I revealed that I reviewed this manuscript for *AJHG*, I gave this information to the journal editor, Dr. Epstein, on February 4, 1992....

In summary, such allegations are deliberate attempts to divert the scientific community's attention from the basic issues of validity and reliability of DNA typing results for forensic applications. If active researchers who are engaged in DNA research are barred from reviewing work in this area, or are excluded from expressing scientific opinion based on their research because of possible "conflict of interest," it would constitute not only a disservice to the courts and society, but would also grossly compromise the quality of scientific progress in this area of research.

Ranajit Chakraborty, Ph.D.
Center for Demographic and Population Genetics
Univ. of Texas Graduate School of Biomedical Sciences

1. Christopher Anderson, "Conflict concerns disrupt panels, cloud testimony," *Nature*, vol. 355, February 27, 1992, pp. 753-754, and C. Anderson, "Coincidence or conspiracy?" *Ibid.*, p. 753.

2. Ranajit Chakraborty and Stephen P. Daiger, "Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah," *Human Biology*, vol. 63, October 1991, pp. 571-58[illegible] Ranjan Deka, Ranajit Chakraborty and Robert E. Ferrell, "A pop[illegible]ulation genetic study of six VNTR loci in three ethnically defined populations," *Genomics*, vol. 11, 1991, pp. 83-92; and Ranajit Chakraborty, "Sample size requirements for addressing the population genetic issues of forensic use of DNA typing," *Human Biology*, vol. 64, April 1992, pp. 141-159.

3. *The American Journal of Human Genetics*, vol. 35, July 1983, delayed publication of criticisms on path analysis by Samuel Karlin et al. (pp. 695-732) to accommodate simultaneous printing of a rebuttal from C.R. Cloninger et al. (pp. 733-756) and a commentary by Sewall Wright (pp. 757-768). The same journal simultaneously published an invited editorial (pp. 819-823) of E. Lander commenting on articles by A.J. Jeffreys et al. (pp. 823-824) and B. Budowle et al. (pp. 824-840) in vol. 48, May 1991.

◇◇◇◇◇◇◇◇◇◇◇◇

In *U.S. vs. Yee*, the evidence included the fact that blood found in the vehicle identified with the crime matched to one of the defendants who managed to wound himself during the murder....

- Following a lengthy, thorough, and costly...Frye hearing, Magistrate Judge James Carr ruled to admit the DNA evidence.
- District Court Judge John Potter affirmed the ruling to admit the DNA evidence.
- The jury found the defendants guilty of all charges.

The defendants and defense lawyers have had their days in court and lost. They lost because the evidence, including DNA, was persuasive to a jury of peers.

Neufeld and Scheck, having lost their case based on their earlier defense positions, are attempting to discredit expert witnesses, myself included. This attempt to personally discredit expert witnesses is deplorable and reflects their desperate position in this case....

It is significant that both the Office of Technology Assessment and the National Academy of Sciences reports have supported the application of DNA technology to forensic science. Furthermore, in the case of *U.S. vs. Randolph Jakobetz*, a case of the 2nd Circuit Court involving DNA evidence, the Federal Court of Appeals...affirmed the lower court ruling on admitting DNA evidence and the conviction of Jakobetz. It is important to keep focus on the major substantive issues and not be diverted or manipulated by these defense lawyers, whose objective and responsibility is to free their clients. No lofty objectives by Neufeld and Scheck on the DNA issues are credible given their client interest.

C. Thomas Caskey, M.D.
Director, Institute for Molecular Genetics
Baylor College of Medicine

◇◇◇◇◇◇◇◇◇◇◇◇

What may be "compelling evidence" to Scheck and Neufeld is nothing but sheer coincidence as far as I am concerned. Their allegations concerning the handling of Dr. Geisser's manuscript, which is still under consideration

by the *AJHG*..., are wholly without merit. As Editor of the *Journal*, my staff and I have bent over backwards to get a fair hearing for Dr. Geisser's manuscript, as we do for any manuscript submitted to us....

During the period that I have been Editor, the *Journal* has served as an open forum on the forensic uses of DNA technology. We have published highly "partisan" but nevertheless carefully reviewed papers on all sides of the issue. The irony of all this, in the present context, is that the last time I was personally called to task and I felt it necessary to explain my position publicly is when I invited and published an editorial sympathetic to Scheck and Neufeld's cause. I take the past and current events together as constituting more than circumstantial evidence that the *Journal* can not really be perceived as exercising bias in one direction or another.

For all of the reasons cited in your article, it is becoming more and more difficult to secure unbiased reviews of papers dealing with the contentious issues surrounding the forensic uses of DNA technology. The fault is not with...my journal, or, I suspect,...other journals which are already or will soon find themselves in a similar position. It lies, instead, in the obscenely polarized atmosphere which puts scientists, law enforcement officials, attorneys, and laboratory directors into two camps which are at loggerhead with one another—those *for* and those *against*. And while there may cerainly be ethical issues, including conflicts of interest, on both sides of the argument, I am even more concerned with the ethics of how the two sides are dealing with one another and with their vehicles for communication. What should be handled as matters of rational scientific debate have taken on the tone of negative political campaigning.

The motivations of the various members of legal, law enforcement, and DNA laboratory communities are reasonably discernible and even understandable. To me, at least, they appear to be related to issues of who will gain and who will lose—in general, and in specific instances—if evidence based on DNA anaylsis is used for forensic purposes. I am, however, somewhat at a loss to explain the behavior of my genetic colleagues on either side of the argument. What should be issues of science have become, in large measure, matters of philosophy and social policy. If this is not the case, the only conclusion that I can come to in the face of the shrill rhetoric emanating from distinguished geneticists in both camps is that the logical basis of population and statistical genetics is more flawed than any of us would like to believe. Let us hope that this is not truly the case.

Charles J. Epstein, Editor
American Journal of Human Genetics

◇◇◇◇◇◇◇◇◇◇◇◇

The article referred to in your essay that Wesley Johnson and I submitted in November 1991 presented some methods for testing statistical independence within a locus. The methods presented have applications to DNA profile data when allelic resolution is subject to measurement error. The product rule, used by the various DNA laboratories for estimating certain relative frequencies, is a result of the assumption of statistical independence.

You have mentioned Ranajit Chakraborty as a referee on our paper. The second referee was Bruce Weir. Both have frequently submitted reports and testified for the prosecution when FBI DNA profiles were at issue. I have testified for the defense in some of these cases. They have collaborated with FBI forensic workers, gained access to their data, and have published it. Certainly they should have recused themselves from serving as referees, or at the very least informed the editor of their situation. Chakraborty did neither, but continually attempted to conceal the facts so that he could deliver a blatantly derogatory referee's report, accusing me of ignorance and error, among other things. His intent was to have me confronted with his "anonymous" referee's report in court by the prosecution. However, his remarks were so transparently identifiable that he finally admitted that he refereed the paper. In what appeared to me to be a misguided effort to counterbalance the blatant attempt to discredit me, Bruce Weir recently sent me a paper by Chakraborty to referee for the journal *Genetics*. I declined to return the favor.

Although the submitted article made no mention of the FBI, the referees' reports indicated that our paper was critical of the FBI. I believe that this resulted from the fact that I was an author of the paper and the two referees worked so closely with the FBI that they did not perceive the larger issues that were addressed.

As an illustration of the methods, a statistical summary of a selected subset of one of the FBI databases that were provided by a local Public Defender's office was used....without attribution to the FBI because the intent was only to display the arithmetic of the methods proposed and not to reach a conclusion. It is our position that the methods, when applied to appropriate databases, can be used to guide conclusions concerning the use of the product rule by any laboratory that generates DNA profiles subject to measurement error.

The paper was returned to us by the editor stating that we should seek permission for using the subset of data from its generator. Further, the editor in telephone conversations stated that we should also provide an analysis of all FBI databases for this paper and resubmit such a revision. I told him that I strongly doubted that they would give us permission to make an independent analysis for publication. They had continually balked at providing the data in a utilizable form that allowed testing of the critical assumptions that they make. When they finally acceded to court orders to do so, they arranged to have the data sealed under a protective order. The editor suggested that I should prepare an analysis anyway. I asked him whether such a protective order could be legally contravened. An opinion was solicited by the editor from his lawyer as to his potential liability as well as ours (authors) if such data were used without the generator's permission. The opinion by his lawyer clearly indicated no liability for the editor, but uncertain liability for the authors of such a publication.

In the light of this fact, the editor instructed me to seek permission from the FBI to make such an analysis of their databases. With some reluctance, because of my past

experience with the FBI and prosecutors working closely with them, I finally requested permission of Bruce Budowle.

A response came about a month later from James Kearney, head of the section on Forensic Science Research. After first expressing concern regarding the use of FBI databases by me, he also questioned my intent.... He then criticized me for not seeking such permission earlier (as if it would have been granted). He went on to indicate that the FBI had already provided the data to Chakraborty, Devlin, Risch and Weir. Finally, he wrote "we are willing to approve your use of FBI population data with certain provisions. You must be sensitive to the fact that previous commitments have been made with other researchers, and the particular study you are doing must not conflict with these projects. The FBI data may be used only in a joint collaboration with Dr. Budowle.... The use of the data is restricted to this one paper. All parties (i.e., authors) must agree to the entire contents of a final manuscript prior to submission to a journal. Any changes whatsoever in the manuscript must be agreed upon by all collaborating parties."

Obviously an independent study under such provisions would be totally compromised, if not impossible. It completely violates the NAS report on DNA Technology in Forensic Science (page 3-23, section 3.73): "If scientific evidence is not yet ready for scientific scrutiny and *public re-evaluation by others*, it is not yet ready for court." By the way, Chakraborty, Devlin, Risch and Weir have all published articles based on the FBI databases without Budowle as a co-author. My analysis of the FBI data obtained from court cases indicates that the assumption of statistical independence should be rejected for many of the probes (loci) for the three major databases they use—Caucasian, Black and Hispanic.

Recently, I analyzed Cellmark databases for a court case in Ann Arbor, Michigan. At the insistence of Cellmark, the prosecutor requested that the judge rule that I not be allowed to submit my analysis of their data for publication. So much for open science!

Seymour Geisser, Ph.D.
Director, School of Statistics
University of Minnesota, Twin City Campus

◇◇◇◇◇◇◇◇◇◇◇◇

In July 1990, I testified as an expert witness on DNA typing in a pretrial admissibility hearing in the case of *U.S. v. Yee, et al.* The DNA testing had been done by the FBI. One of the central areas of contention in the admission of DNA evidence is the validity of the calculation used to predict the chance that a randomly selected individual would match the DNA pattern of the forensic sample, that is, the "significance" of a match.... I expressed the view, which I still hold, that no matter how this calculation is made (within reason), the chance of a match by coincidence alone is extremely small; thus the DNA evidence is probative and significant.

Several months prior to my testimony in *Yee*, two colleagues and I submitted a grant proposal for research on DNA fingerprinting to the National Institutes of Justice (NIJ). The grant application requested approximately $200,000 total direct costs to support two years of proposed research. [It was approved] in August of 1990.... The grant supports the research of three faculty-level investigators, myself included, but provides no salary or other personal benefits for us....the grant funds are administered by the University of Texas, not the investigators.

Our proposed NIJ grant was not discussed in my testimony in the *Yee* case. However, the appeals brief filed by Scheck and Neufeld alleges two types of conflict arising from the grant. First, the brief states that I should have spontaneously reported that a grant had been submitted, even though no directly relevant questions were asked of me. I see this as a largely procedural issue, since the adversarial nature of testimony in criminal cases makes it extremely difficult to "volunteer" information, especially when its relevance to the particular case is unclear.

The second alleged conflict, though, is...likely to be troubling to the scientific community. The appeals brief suggests that since our proposed research project was to be funded by a federal agency, and since that agency oversees an interested party in this case (the FBI), there was, *a priori*, a conflict of interest in my testimony. What makes this doctrine troubling is that by this standard any research scientist receiving support from a governmental agency might be precluded from providing expert testimony in a case to which the agency is party, however indirectly. If nothing else, this would severely limit the availability of expert advice to the courts and the government.

As your essay notes, there are still areas of contention on the admission of DNA typing results in criminal cases....to limit the supply of expert testimony by excluding individuals with research support in this area would be a disservice to the courts and to society.

Stephen P. Daiger, Ph.D.
Professor, Medical Genetics Center
The University of Texas Health Science Center

◇◇◇◇◇◇◇◇◇◇◇◇

[F]orensic DNA evidence...is now being used by law enforcement agencies throughout the world. However,...a few vocal critics persist in their attempts to discredit the validity of DNA technology as a reliable and powerful forensic tool. As these challenges have failed on legal and technical grounds, some critics have unfortunately resorted to vengeful personal attacks and political debate to distract and confuse the public and the courts. For a number of these critics who frequently appear as defense experts, their livelihood depends to some degree on keeping a controversy alive.

The primary area still being exploited by critics and defense experts is the interpretation of the population statistics used to estimate the significance of a DNA match. However, the National Academy of Sciences' study endorsed the current statistical methods used by forensic laboratories with recommended modifications in the procedures until the availability of additional world-wide population data....

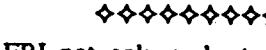
There are, of course, a number of other specific and general recommendations with respect to the forensic use of DNA technology. Two of those which are being studied carefully deal with developing appropriate models to control and assure the reliability of DNA technology and alternative approaches to statistical interpretation of DNA results. These issues are being studied in consideration of federal and state legislative initiatives and ongoing efforts of other government agencies and professional groups....

DNA analysis has been conducted in over 14,700 criminal cases and admitted in over 612 criminal trials. DNA evidence has been rejected in only five reported cases and seven unreported cases. A few courts have allowed the evidence of a match to be introduced, but reduced or excluded altogether the statistics associated with a match.

There have been 53 appellate decisions in the United States directly addressing the admissibility of forensic DNA test results. Every appellate court addressing the general acceptance or relevancy/reliability of the Restriction Fragment Length Polymorphism (RFLP) technology has ruled that the technology met the applicable standard. Two appellate decisions have remanded the cases back to the trial court for failure of the prosecution to lay an adequate foundation for the admission of the population statistics. Only one appellate court has excluded DNA evidence. In that Minnesota case, the court cited the laboratory's failure to follow certain minimum guidelines it deemed necessary for the admission of DNA test results, as well as previous case precedent which restricted the use of population frequencies in criminal cases.

The forensic use of DNA technology will enhance the effectiveness of law enforcement throughout the United States.... For the full potential of the technology to be realized, a coordinated national effort is essential to establish testing standards which meet stringent criteria for compatibility and reliability. This has been the cornerstone of the FBI's program for the nationwide implementation of DNA technology for use by law enforcement. In working with the foresnic community and other scientists, the FBI has facilitated a consensus in defining such standards....

John W. Hicks
Assistant Director in Charge, Laboratory Division
Federal Bureau of Investigation



The FBI not only seeks to have its forensic DNA methods approved by courts as "generally accepted" among population geneticists and molecular geneticists, but, more importantly, it asks Congress to pass legislation giving it ultimate authority to set scientific and quality assurance standards for all forensic DNA laboratories that will contribute and receive genetic information from the Bureau's national DNA databank. This is the critical policy issue of the day. It must inform any assessment of the controversy arising from the government's intrusion into the publication and peer review processes of scientific journals, the intimidation of scientists who have criticized the government's forensic DNA methods, and the failure of prosecutors and scientist witnesses to make appropriate disclosures to courts about substantial government grants being awarded to those witnesses.

In this short space it is impossible to summarize all the relevant facts and documents contained in the *U.S. v. Yee, et al.* court papers. We will gladly make them available upon request, and particularly suggest attention be paid to a lengthy affidavit submitted by Dr. Sheldon Krimsky, a leading authority on social and ethical questions in science and technology....

One example we cited of interference in the publication and peer review process was the treatment accorded Dr. Seymour Geisser when he submitted a manuscript criticizing statistical methods of forensic DNA laboratories to the *AJHG*. Three clear problems emerged: 1) There was strong evidence the confidentiality of the peer review process had been breached through disclosures to prosecutors; 2) Dr. Charles Epstein, *AJHG*'s editor, did not know that Dr. Ranjit Chakraborty, the peer reviewer who vituperatively opposed publication, was a recipient of a Justice Department grant intended to establish scientific support for the statistical methods; and 3) Dr. Epstein's own evaluation was admittedly based on how the paper "would be used in court cases," as opposed to a strict consideration of the scientific merits.

The saga of Dr. Geisser's papers has now, however, taken an even more troubling turn. Dr. Geisser was asked by *AJHG* to include in his paper an analysis of FBI population data that had been disclosed in numerous court cases. The FBI objected, taking the position that unpublished FBI population data could not be used by Dr. Geisser unless he made Dr. Budowle of the FBI a co-author, and that Dr. Budowle would then have the right to approve or disapprove every word in the Geisser paper....

The FBI's position reflects a mean spirited form of censorship. Most importantly, this position has been severely criticized by the National Academy of Sciences (NAS) in its recent report, *DNA Technology in Forensic Sciences*:

> Any population databank used in support of forensic DNA typing should be openly available for reasonable scientific inspection. Presenting scientific conclusions in a criminal court is at least as serious as presenting scientific conclusions in an academic paper. According to long-standing and wise scientific tradition, the data underlying an important scientific conclusion must be freely available, so that others can evaluate the results and *publish* their own findings, whether in support or disagreement. There is no excuse for secrecy in raw data....If scientific evidence is not yet ready for both scientific scrutiny and public re-evaluation by others, it is not yet ready for court. (NAS Rpt, at 3-23.)

Another example of interference with the publication and peer review process was the FBI's efforts to influence improperly the National Academy's DNA report. It is now undisputed that John Hicks, Director of the FBI Laboratory, improperly obtained from two members of the NAS Committee a confidential draft of the report concerning statistical methods that was very critical of the FBI. Hicks subsequently wrote an unsolicited reply that NAS staff say they did not distribute to the Committee. In an FBI crime

Professional Ethics Report
Directorate for Science and Policy Programs
American Association for the Advancement of Science
1333 H Street, NW, Washington, DC 20005

Non-Profit Org.
U.S. Postage Paid
Washington, D.C
Permit no. 7304

laboratory publication, Hicks writes that an "internal investigation" was conducted of our allegations, including presumably our allegations against him, and that no evidence of wrongdoing was found. We wonder who conducted that investigation and how it could reach such conclusions given the clear evidence that Hicks himself inexcusably violated the peer review and publication processes of the NAS.

Finally, as far as the Justice Department grants to Drs. Caskey and Daiger are concerned, a few points need to be made: 1) FBI agent Kearney sat on the panels that awarded these grants, so that any claim they were awarded in a fashion analogous to NIH grants is simply absurd; 2) That Drs. Daiger and Caskey did not feel compromised or even financially rewarded by these grants is ethically irrelevant to the obligation to disclose; and 3) There is no legal justification whatsoever for failing to disclose these grants to the court.

The NAS pointedly recommends the Department of Health and Human Services, and not the Department of Justice (DOJ), be given responsibility for accrediting forensic DNA laboratories, and that this be done through contracting with private professional organizations. After concluding that the DOJ (i.e., the FBI) "lacks expertise in quality assurance and quality control in molecular genetics," the NAS observed "the DOJ may be perceived as an advocate for application of the technology," and "[o]versight by DOJ may not be perceived as providing adequate assurance to the public or to a defendant facing prosecution by DOJ." (NAS Rpt., at 4-12.)

The recommendations of the NAS should be strongly supported in Congress by all in the legal and scientific communities who care about the responsible, reliable, and ethical use of forensic DNA technology.

Barry Scheck, Esq.
Director, Clinical Education
Benjamin N. Cardozo School of Law

RESOURCES

Responsible Science: Ensuring the Integrity of the Research Process (Washington, DC: National Academy of Sciences Press, 1992); $24.95 plus $3.00 shipping. To order, call the Press at (800) 624-6242.

The AAAS has published, *The Genome, Ethics and the Law: Issues in Genetic Testing*. The book includes an overview paper of the discussion that took place at a conference in Berkeley Springs, West Virginia, June 14-16, 1991, and three background papers addressing the scientific basis of genetic testing, the ethical implications of recent and anticipated advances, and the legal issues those advances raise. Single copies are free. Contact Alexander Fowler, AAAS Science and Policy Programs, 1333 H Street NW, Washington, DC, 20005; (202) 326-6600.

Support from the following professional societies is gratefully acknowledged:

American Academy of Environmental Engineers
American Academy of Otolaryngology—Head and Neck Surgery
American Anthropological Association
American Association of Engineering Societies
American Association of University Professors
American Chemical Society
American Institute of Aeronautics and Astronautics
American Institute of Chemists
American Medical Association
American Occupational Therapy Association
American Pharmaceutical Association
American Philosophical Association
American Political Science Association
American Psychological Association
American Society for Engineering Education
American Society for Microbiology
American Society for Photogrammetry and Remote Sensing
American Society for Public Administration
American Society of Mechanical Engineers
American Sociological Association
American Speech-Language-Hearing Association
American Statistical Association
Association of American Medical Colleges
Botanical Society of America
Council of Biology Editors
Council of Scientific Society Presidents
Ecological Society of America
Federation of Behavioral, Psychological, & Cognitive Sciences
Institute of Electrical and Electronics Engineers
National Association of Social Workers
National Society of Professional Engineers
Society for Computer Simulation
Society for Epidemiologic Research
Society for the Scientific Study of Sex
Society of Professional Archeologists
University Center for Human Values, Princeton University
Vaughen Enterprises Ethics Consulting Services

Professional Ethics Report

Newsletter of the American Association for the Advancement of Science
Committee on Scientific Freedom & Responsibility
Professional Society Ethics Group

VOLUME V ✦ NUMBER 2 ✦ SPRING 1992

IN THE NEWS

This past February voters in Switzerland rejected a proposed referendum banning scientific experiments using animals. The ban proposed by the Swiss Animal Protection League would have tightened existing restrictions for obtaining a license to experiment on animals. The Swiss government, medical groups, and Swiss-based pharmaceutical companies all opposed the initiative, claiming that it could lead to the relocation of research facilities outside Switzerland. Another major concern was the ban's implications for commercial confidentiality. Allowing animal rights groups to challenge in court individual research projects presumably would require that companies release their detailed research plans. Another aspect of the proposal that worried researchers was a clause requiring the Swiss government to enact an animal experimentation law within five years. Failure to do so could ban animal research completely.

On April 22, the National Academy of Sciences issued its report on *Responsible Science: Ensuring the Integrity of the Research Process* (see In Print) by a special Panel on Scientific Responsibility and the Conduct of Research. The Panel made twelve recommendations, including the creation of an independent Scientific Integrity Advisory Board, which would serve as a clearinghouse for the exchange of information and experiences related to scientific misconduct and efforts to promote responsible research conduct, and one that called on scientific societies and journals to "provide and expand resources and forums to foster responsible research practices and to address misconduct in science and questionable research practices." The Panel also urged the government to adopt a common definition of misconduct in science and common policies and procedures for handling allegations of scientific misconduct. The report also draws a distinction between misconduct in science, which includes "fabrication, falsification or plagiarism in proposing, performing, or reporting research," and "questionable research practices." The latter includes "actions that violate traditional values of the research enterprise and that may be detrimental to the research process," but for which there is currently "neither broad agreement as to the seriousness of these actions nor any consensus on standards for behavior in such matters."

IN THE SOCIETIES

The revision of the American Psychological Association's "Ethical Principles of Psychologists" is an action item on its Council of Representatives' agenda for the APA's August 1992 convention meeting. The revision involves major changes from the current (1989) version. In particular, it involves a clear identification of the aspirational versus enforceable sections. Contact Stanley E. Jones, Director, Office of Ethics, APA, 750 First Street, NE, Washington, DC 20002-4242; (202) 336-5500.

At its May meeting, the Council of Scientific Society Presidents adopted several resolutions related to professional ethics. Among them was one on "The Role of Professional Societies in Setting Ethical Standards in Science," which urges scientific societies to "develop mechanisms to educate members regarding standards of research practice, the ethical conduct and reporting of science, and the traditions, values, and paradigms of the discipline." Two other resolutions endorsed recommendations contained in the National Academy of Sciences' report *Responsible Science: Ensuring the Integrity of the Research Process* — one calling for uniform federal policies and procedures for handling allegations of misconduct in science, and the other urging institutions to "assure both accusers and accused the fundamental elements of due process...." For more details, contact CSSP at (202) 872-4452.

CASES AND COMMENTARIES

In past issues, PER has used this section to print invited commentaries on hypothetical cases. In this issue, we present a case that is currently unfolding in federal court. In recent years, scientists and attorneys have engaged in heated arguments—inside and outside the courtroom—over the reliability of DNA fingerprinting in criminal cases. The debate has encompassed a number of issues: the techniques used to determine whether samples match, the statistical methods used to interpret a match, and the standards and practices of quality control of the laboratories that perform the analyses.

Connected to this more public debate is a series of less visible, yet volatile skirmishes that highlight tensions

between law and science and raise issues of professional ethics for both the legal and scientific communities. In order to focus more attention on these matters and their implications for the relationship between law and science, PER has prepared this essay and invited all of the parties involved in the different incidents it describes to share their positions on the controversies with our readers. Those contributions received follow this essay. Readers of PER are invited to send us their reactions to the essay as well as to the responses that follow it for publication in our next issue. These should be received by August 15.—Editor

In a motion for a new trial in the 1989 murder case, *U.S. v. Yee, et al.*, two New York defense attorneys–Barry Scheck and Peter Neufeld–have filed an affidavit with a U.S. District Court in Ohio alleging misconduct by law enforcement officials and scientists in disputes over the reliability of DNA fingerprinting in criminal cases. The two attorneys accuse federal and state law enforcement of efforts "to intimidate, harass, and deter expert witnesses from testifying against the FBI and other forensic DNA laboratories and publishing their views." They go on to describe "the most troubling aspect of the government's campaign [as] its direct interference in the publication and peer review process...to prevent and undermine the publication of scientific opinions law enforcement does not like." They also accuse several scientists of conflicts of interest in providing testimony on the reliability of DNA fingerprinting and in participating in the peer review process of scientific journals. In both published reports and additional motions and affidavits submitted to the Ohio court, law enforcement officials and scientists named in the defense motion have refuted the allegations.

The defense attorneys refer to a manuscript accepted in September 1991 for publication in *Science* in which two population geneticists–Richard Lewontin and Daniel Hartl–claimed that proponents of DNA fingerprinting have made unwarranted assumptions regarding the rarity of genetic patterns in populations. They concluded that DNA fingerprinting as currently practiced should not be admissible as evidence. Scheck and Neufeld report that James Wooley, an Assistant U.S. Attorney, telephoned Hartl, pressuring him to withdraw the paper. The co-authors refused to do so, and Lewontin wrote Wooley, accusing him of a "very serious breach of ethics" and "intimidation."[1] According to the affidavit, federal law enforcement officials, in association with scientists who dispute the claims made in the manuscript (the affidavit mentions C. Thomas Caskey and Ranajit Chakraborty), "succeeded in delaying publication..., altering the content of the article, and getting an unprecedented simultaneous publication of a rebuttal article which did not go through the same peer review process."

In a counter affidavit submitted to the court, Wooley accuses the defense affidavit of being a "vehicle through which to launch a vicious, mean-spirited and baseless attack on the character, ethics and actions of numerous FBI agents, prosecutors, and nationally prominent scientists,... who support the admissibility of DNA testing...." He confirms speaking with Hartl, but denies intimidating or threatening him. He states that after the conversation with Hartl, he "took no action relative to the publication of the paper."

The paper by Lewontin and Hartl as well as the rebuttal article were published in *Science*.[2] In that same issue,[3] the editor of *Science*, Daniel Koshland, is reported to have been "disturbed that the data did not support the paper's conclusions" and asked the coauthors to make revisions. Koshland denied having received any communications on the matter from government officials, although he did hear complaints from other scientists. He defended his decision to solicit a rebuttal article, which, contrary to the claims of Scheck and Neufeld, was also peer-reviewed, "to give a more balanced view of the subject," although the normal procedure followed by *Science* is to publish rebuttals in a subsequent issue and to give the authors of the original article an opportunity to respond.

Another example cited in the affidavit by Scheck and Neufeld is a paper criticizing the statistical methods used by forensic laboratories submitted to the *American Journal of Human Genetics* (*AJHG*) in November 1991 by Seymour Geisser, a University of Minnesota statistician. The paper had come to the attention of Stephen L. Redding, a prosecutor in the Office of the Hennipen County Attorney in Minnesota, where the paper's author was scheduled to testify as a witness for the defense at a DNA admissibility hearing in January 1992. Geisser received a fax from Redding demanding that he produce in court any manuscript he authored, whether accepted or under review, related to DNA fingerprinting. According to the defense affidavit, fifteen minutes after receiving that fax, the author received a fax from Charles Epstein, editor of *AJHG*, along with comments on his manuscript from three anonymous reviewers, one of whom strongly recommended against publication. This latter reviewer was later identified as Ranajit Chakraborty, a scientist who has testified for the prosecution in criminal cases and who coauthored the rebuttal article in *Science* referred to earlier.

The affidavit quotes from Epstein's letter that "since this work will certainly be used in court cases, the writing needs to be more careful, and the work must apply to the data in the way they are actually used." Scheck and Neufeld conclude that this series of events "provides compelling circumstantial evidence that the rules of journal and reviewer confidentiality were breached and that prosecutors meddled in the process." They continue, "journals are

Editor: Mark S. Frankel
Associate and Managing Editor: Amy Crumpton
Contributing Editor: Alexander Fowler
Editorial Board: William Anderson, Brian Boom, John Gardenier, Jonathan Knight, William Middleton
The *Professional Ethics Report* is published quarterly under the auspices of the Committee on Scientific Freedom and Responsibility and the Professional Society Ethics Group, American Association for the Advancement of Science, 1333 H Street, NW, Washington, DC 20005, (202)326-6798.

supposed to evaluate a...manuscript on its scientific merits, not based on the uses courts may make of a scientist's opinion." Epstein denies being influenced by anyone, and claims that he was unaware that Chakraborty was co-investigator on a Justice Department grant intended to establish scientific support for the FBI's statistical methods used in evaluating DNA evidence when he asked him to review the Geisser manuscript.[4]

These incidents raise potentially serious conflict of interest problems, as scientists with strong positions on the reliability of DNA fingerprinting and/or financial interests in companies associated with the technique are called on to testify in court or to review related journal manuscripts or grant applications. Indeed, Scheck and Neufeld claim that two scientists who gave expert testimony in a hearing related to the *Yee* case failed to disclose relevant financial ties. They argue that one of the scientists, Stephen P. Daiger of the Graduate School of Biomedical Sciences at the University of Texas at Houston, had a grant application on DNA typing for forensic applications with the National Institute of Justice that constitutes a "substantial financial bias and conflict of interest that should have been revealed to the court and counsel when he testified." They argue further that "testimony favorable to the FBI and the Justice Department would no doubt preserve or enhance the prospect of receiving [the] $300,000 NIJ grant."

In an affidavit submitted to the Ohio court, Daiger expresses irritation at "the criticisms and insults contained" in the defense document. He notes that he did not stand to profit personally since grant funds go directly to his institution for disbursement. He adds that "At no time during [the review] process was it stated or implied to me that my testimony in any legal case, or in any other forum, would have an effect on the review process."

The defense attorneys also accuse C. Thomas Caskey of Baylor College of Medicine of similar conflicts of interest. Like Daiger, Caskey had a grant application pending at the NIJ. The proposed research was intended to develop "a low cost, rapid analysis PCR based DNA profiling system...[that would] be a highly marketable product." Caskey also has an agreement to license his diagnostic techniques to Cellmark Diagnostics, Inc., a major DNA fingerprinting company. The defense affidavit argues that "the failure of the government and Dr. Caskey to disclose his $200,000 Justice Department grant deprived the defense counsel of an opportunity to explore fully and fairly the bias, conflict of interest, and financial motives of the government's most important witness." [Caskey's ties to Cellmark Diagnostics led to his resignation from a committee of the National Academy of Sciences established to report on the use of DNA technology in forensic science.]

In his own affidavit to the court, Caskey notes that he has "not personally profited from the NIJ grant or from the licensing of the newly developed personal identification technology to Cellmark...." He states further that he has "freely discussed the NIJ award and the Cellmark license with anyone who has asked about these matters and I would have responded truthfully...." if questioned about them at the hearing. Commenting more generally on potential conflicts of interest in court proceedings related to DNA fingerprinting, Caskey added that he did "not feel that any scientist associated with an institution which is a recipient of competitive federal (or state) financial support should automatically be disqualified from serving as a witness for the prosecution or defense in any case. To automatically exclude all...grantees from serving as expert witnesses will deprive all interested parties of the services of highly qualified scientists."

PER has received a letter from the Medical Faculty of Rijeka in Croatia requesting assistance in "resuscitating ethics" in their curriculum. Because of the war and poor economic conditions, the faculty is badly in need of literature. Below are edited excerpts from the letter:

> We are applying to you to help us with at least a small contribution. You could send us a book, magazines, or reference manual on medical ethics (or medical sociology) from your personal library, with your dedication or a message of solidarity on it.
>
> Your institution could send us some books on medical ethics, or subscribe to a magazine for us dealing with medical ethics (or forward us funds so that we can purchase materials).
>
> We ask you to support the idea of organizing a symposium on the ethical aspects of this war in Dubrovavik, once this unhappy war is over.
>
> We kindly thank you in advance for your understanding and support.
>
> *Prof. dr. Ivan Segota*
> *Head, Department of Social Sciences*
> *Medical Faculty of Rijeka*
> *O. Ban 22, 51000 Rijeka Croatia*

Contact Prof. Segota directly at the above address.

1. Christopher Anderson, "DNA fingerprinting discord," *Nature*, vol. 54, December 1991, p. 500.

2. R.C. Lewontin and Daniel L. Hartl, "Population Genetics in Forensic DNA Typing," *Science*, vol. 254, December 20, 1991, pp. 1745-50, and R. Chakraborty and Kenneth K. Kidd, "The Utility of DNA Typing in Forensic Work," *Ibid*, pp. 1735-39.

3. L. Roberts, "Was *Science* Fair to its Authors?" *Ibid*, p. 1722.

4. Christopher Anderson, "Conflict concerns disrupt panels, cloud testimony," *Nature*, vol. 355, February 27, 1992, p. 753-54.

Commentaries

The reported allegations[1] in the affidavit of Scheck and Neufeld requesting a new trial are nothing but unsubstantiated innuendos. I was never involved in the *Yee* case as an expert. Nevertheless, since my research deals with the subject of the effects of population substructure on genetic variation, I have read both Hartl's and Lewontin's reports on this subject, which are public records. I did not find their arguments convincing, nor did I find any support for their criticisms in our analysis of DNA typing data on populations around the world.[2] I had a draft version of a rebuttal to their criticisms prepared even before they submitted their paper to *Science*. I first learned of their *Sci-*

ence article in a court case around the third week of September 1991 where it was submitted as a defense exhibit. In the first week of October 1991, at the 8th International Congress of Human Genetics in Washington, DC, I found that a large number of human geneticists knew of their forthcoming article, and I expressed my intent to publish a rebuttal, should their article appear in *Science*. Upon return from the Congress (second week of October 1991), I received a telephone request from the editorial office of *Science* to submit my rebuttal. At that stage, I asked Dr. Kenneth Kidd to co-author the rebuttal,.... At no stage in the preparation of our paper, did I encourage *Science* to "delay" the publication of Lewontin and Hartl's paper, nor did I attempt to "alter the content" of their article....our rebuttal was reviewed by at least two reviewers. Scientists working in a forefront area should have no reason to believe that this procedure was "unprecedented" since simultaneous publications of controversial opinions are practiced by several leading journals.[3]

The account[1] of the review of Dr. Geisser's manuscript submitted to the *AJHG* is also a misrepresentation of actual events. It is true that at the request of the AJHG editor I acted as a reviewer of that article. I sent my comments to the journal office by the end of December 1991. According to the regular practice of reviewers, I have never discussed this review nor the paper with anyone. I was critical of the manuscript, because I believed that it was unprofessionally written, it contained several fatal errors, and it only reported parts of unpublished data from other laboratories without appropriate credit or consent of the data gatherers. My co-investigatorship in a NIJ grant had no connection with my reviewing this manuscript, and my review was to the point of evaluating a "scientific manuscript on its scientific merit." In the last week of January 1992, in a court case in Seattle, a defense exhibit produced all reviews and Dr. Geisser's correspondence with the journal editor, through which I knew that his manuscript was being used in the courts. Since under a court order I revealed that I reviewed this manuscript for *AJHG*, I gave this information to the journal editor, Dr. Epstein, on February 4, 1992....

In summary, such allegations are deliberate attempts to divert the scientific community's attention from the basic issues of validity and reliability of DNA typing results for forensic applications. If active researchers who are engaged in DNA research are barred from reviewing work in this area, or are excluded from expressing scientific opinion based on their research because of possible "conflict of interest," it would constitute not only a disservice to the courts and society, but would also grossly compromise the quality of scientific progress in this area of research.

Ranajit Chakraborty, Ph.D.
Center for Demographic and Population Genetics
Univ. of Texas Graduate School of Biomedical Sciences

1. Christopher Anderson, "Conflict concerns disrupt panels, cloud testimony," *Nature*, vol. 355, February 27, 1992, pp. 753-754, and C. Anderson, "Coincidence or conspiracy?" *Ibid.*, p. 753.

2. Ranajit Chakraborty and Stephen P. Daiger, "Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah," *Human Biology*, vol. 63, October 1991, pp. 571-58[illegible] Ranjan Deka, Ranajit Chakraborty and Robert E. Ferrell, "A population genetic study of six VNTR loci in three ethnically defined populations," *Genomics*, vol. 11, 1991, pp. 83-92; and Ranajit Chakraborty, "Sample size requirements for addressing the population genetic issues of forensic use of DNA typing," *Human Biology*, vol. 64, April 1992, pp. 141-159.

3. *The American Journal of Human Genetics*, vol. 35, July 1983, delayed publication of criticisms on path analysis by Samuel Karlin et al. (pp. 695-732) to accommodate simultaneous printing of a rebuttal from C.R. Cloninger et al. (pp. 733-756) and a commentary by Sewall Wright (pp. 757-768). The same journal simultaneously published an invited editorial (pp. 819-823) of E. Lander commenting on articles by A.J. Jeffreys et al. (pp. 823-824) and B. Budowle et al. (pp. 824-840) in vol. 48, May 1991.

◇◇◇◇◇◇◇◇◇◇◇◇

In *U.S. vs. Yee*, the evidence included the fact that blood found in the vehicle identified with the crime matched to one of the defendants who managed to wound himself during the murder....

- Following a lengthy, thorough, and costly...Frye hearing, Magistrate Judge James Carr ruled to admit the DNA evidence.
- District Court Judge John Potter affirmed the ruling to admit the DNA evidence.
- The jury found the defendants guilty of all charges.

The defendants and defense lawyers have had their days in court and lost. They lost because the evidence, including DNA, was persuasive to a jury of peers.

Neufeld and Scheck, having lost their case based on their earlier defense positions, are attempting to discredit expert witnesses, myself included. This attempt to personally discredit expert witnesses is deplorable and reflects their desperate position in this case....

It is significant that both the Office of Technology Assessment and the National Academy of Sciences reports have supported the application of DNA technology to forensic science. Furthermore, in the case of *U.S. vs. Randolph Jakobetz*, a case of the 2nd Circuit Court involving DNA evidence, the Federal Court of Appeals...affirmed the lower court ruling on admitting DNA evidence and the conviction of Jakobetz. It is important to keep focus on the major substantive issues and not be diverted or manipulated by these defense lawyers, whose objective and responsibility is to free their clients. No lofty objectives by Neufeld and Scheck on the DNA issues are credible given their client interest.

C. Thomas Caskey, M.D.
Director, Institute for Molecular Genetics
Baylor College of Medicine

◇◇◇◇◇◇◇◇◇◇◇◇

What may be "compelling evidence" to Scheck and Neufeld is nothing but sheer coincidence as far as I am concerned. Their allegations concerning the handling of Dr. Geisser's manuscript, which is still under consideration

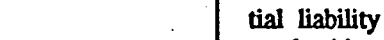
by the *AJHG*..., are wholly without merit. As Editor of the *Journal*, my staff and I have bent over backwards to get a fair hearing for Dr. Geisser's manuscript, as we do for any manuscript submitted to us....

During the period that I have been Editor, the *Journal* has served as an open forum on the forensic uses of DNA technology. We have published highly "partisan" but nevertheless carefully reviewed papers on all sides of the issue. The irony of all this, in the present context, is that the last time I was personally called to task and I felt it necessary to explain my position publicly is when I invited and published an editorial sympathetic to Scheck and Neufeld's cause. I take the past and current events together as constituting more than circumstantial evidence that the *Journal* can not really be perceived as exercising bias in one direction or another.

For all of the reasons cited in your article, it is becoming more and more difficult to secure unbiased reviews of papers dealing with the contentious issues surrounding the forensic uses of DNA technology. The fault is not with...my journal, or, I suspect,...other journals which are already or will soon find themselves in a similar position. It lies, instead, in the obscenely polarized atmosphere which puts scientists, law enforcement officials, attorneys, and laboratory directors into two camps which are at loggerhead with one another—those *for* and those *against*. And while there may cerainly be ethical issues, including conflicts of interest, on both sides of the argument, I am even more concerned with the ethics of how the two sides are dealing with one another and with their vehicles for communication. What should be handled as matters of rational scientific debate have taken on the tone of negative political campaigning.

The motivations of the various members of legal, law enforcement, and DNA laboratory communities are reasonably discernible and even understandable. To me, at least, they appear to be related to issues of who will gain and who will lose—in general, and in specific instances—if evidence based on DNA anaylsis is used for forensic purposes. I am, however, somewhat at a loss to explain the behavior of my genetic colleagues on either side of the argument. What should be issues of science have become, in large measure, matters of philosophy and social policy. If this is not the case, the only conclusion that I can come to in the face of the shrill rhetoric emanating from distinguished geneticists in both camps is that the logical basis of population and statistical genetics is more flawed than any of us would like to believe. Let us hope that this is not truly the case.

Charles J. Epstein, Editor
American Journal of Human Genetics



The article referred to in your essay that Wesley Johnson and I submitted in November 1991 presented some methods for testing statistical independence within a locus. The methods presented have applications to DNA profile data when allelic resolution is subject to measurement error. The product rule, used by the various DNA laboratories for estimating certain relative frequencies, is a result of the assumption of statistical independence.

You have mentioned Ranajit Chakraborty as a referee on our paper. The second referee was Bruce Weir. Both have frequently submitted reports and testified for the prosecution when FBI DNA profiles were at issue. I have testified for the defense in some of these cases. They have collaborated with FBI forensic workers, gained access to their data, and have published it. Certainly they should have recused themselves from serving as referees, or at the very least informed the editor of their situation. Chakraborty did neither, but continually attempted to conceal the facts so that he could deliver a blatantly derogatory referee's report, accusing me of ignorance and error, among other things. His intent was to have me confronted with his "anonymous" referee's report in court by the prosecution. However, his remarks were so transparently identifiable that he finally admitted that he refereed the paper. In what appeared to me to be a misguided effort to counterbalance the blatant attempt to discredit me, Bruce Weir recently sent me a paper by Chakraborty to referee for the journal *Genetics*. I declined to return the favor.

Although the submitted article made no mention of the FBI, the referees' reports indicated that our paper was critical of the FBI. I believe that this resulted from the fact that I was an author of the paper and the two referees worked so closely with the FBI that they did not perceive the larger issues that were addressed.

As an illustration of the methods, a statistical summary of a selected subset of one of the FBI databases that were provided by a local Public Defender's office was used....without attribution to the FBI because the intent was only to display the arithmetic of the methods proposed and not to reach a conclusion. It is our position that the methods, when applied to appropriate databases, can be used to guide conclusions concerning the use of the product rule by any laboratory that generates DNA profiles subject to measurement error.

The paper was returned to us by the editor stating that we should seek permission for using the subset of data from its generator. Further, the editor in telephone conversations stated that we should also provide an analysis of all FBI databases for this paper and resubmit such a revision. I told him that I strongly doubted that they would give us permission to make an independent analysis for publication. They had continually balked at providing the data in a utilizable form that allowed testing of the critical assumptions that they make. When they finally acceded to court orders to do so, they arranged to have the data sealed under a protective order. The editor suggested that I should prepare an analysis anyway. I asked him whether such a protective order could be legally contravened. An opinion was solicited by the editor from his lawyer as to his potential liability as well as ours (authors) if such data were used without the generator's permission. The opinion by his lawyer clearly indicated no liability for the editor, but uncertain liability for the authors of such a publication.

In the light of this fact, the editor instructed me to seek permission from the FBI to make such an analysis of their databases. With some reluctance, because of my past

experience with the FBI and prosecutors working closely with them, I finally requested permission of Bruce Budowle.

A response came about a month later from James Kearney, head of the section on Forensic Science Research. After first expressing concern regarding the use of FBI databases by me, he also questioned my intent.... He then criticized me for not seeking such permission earlier (as if it would have been granted). He went on to indicate that the FBI had already provided the data to Chakraborty, Devlin, Risch and Weir. Finally, he wrote "we are willing to approve your use of FBI population data with certain provisions. You must be sensitive to the fact that previous commitments have been made with other researchers, and the particular study you are doing must not conflict with these projects. The FBI data may be used only in a joint collaboration with Dr. Budowle.... The use of the data is restricted to this one paper. All parties (i.e., authors) must agree to the entire contents of a final manuscript prior to submission to a journal. Any changes whatsoever in the manuscript must be agreed upon by all collaborating parties."

Obviously an independent study under such provisions would be totally compromised, if not impossible. It completely violates the NAS report on DNA Technology in Forensic Science (page 3-23, section 3.73): "If scientific evidence is not yet ready for scientific scrutiny and *public re-evaluation by others*, it is not yet ready for court." By the way, Chakraborty, Devlin, Risch and Weir have all published articles based on the FBI databases without Budowle as a co-author. My analysis of the FBI data obtained from court cases indicates that the assumption of statistical independence should be rejected for many of the probes (loci) for the three major databases they use—Caucasian, Black and Hispanic.

Recently, I analyzed Cellmark databases for a court case in Ann Arbor, Michigan. At the insistence of Cellmark, the prosecutor requested that the judge rule that I not be allowed to submit my analysis of their data for publication. So much for open science!

Seymour Geisser, Ph.D.
Director, School of Statistics
University of Minnesota, Twin City Campus

◇◇◇◇◇◇◇◇◇◇◇◇

In July 1990, I testified as an expert witness on DNA typing in a pretrial admissibility hearing in the case of *U.S. v. Yee, et al.* The DNA testing had been done by the FBI. One of the central areas of contention in the admission of DNA evidence is the validity of the calculation used to predict the chance that a randomly selected individual would match the DNA pattern of the forensic sample, that is, the "significance" of a match.... I expressed the view, which I still hold, that no matter how this calculation is made (within reason), the chance of a match by coincidence alone is extremely small; thus the DNA evidence is probative and significant.

Several months prior to my testimony in *Yee*, two colleagues and I submitted a grant proposal for research on DNA fingerprinting to the National Institutes of Justice (NIJ). The grant application requested approximately $200,000 total direct costs to support two years of proposed research. [It was approved] in August of 1990.... The grant supports the research of three faculty-level investigators, myself included, but provides no salary or other personal benefits for us....the grant funds are administered by the University of Texas, not the investigators.

Our proposed NIJ grant was not discussed in my testimony in the *Yee* case. However, the appeals brief filed by Scheck and Neufeld alleges two types of conflict arising from the grant. First, the brief states that I should have spontaneously reported that a grant had been submitted, even though no directly relevant questions were asked of me. I see this as a largely procedural issue, since the adversarial nature of testimony in criminal cases makes it extremely difficult to "volunteer" information, especially when its relevance to the particular case is unclear.

The second alleged conflict, though, is...likely to be troubling to the scientific community. The appeals brief suggests that since our proposed research project was to be funded by a federal agency, and since that agency oversees an interested party in this case (the FBI), there was, *a priori*, a conflict of interest in my testimony. What makes this doctrine troubling is that by this standard any research scientist receiving support from a governmental agency might be precluded from providing expert testimony in a case to which the agency is party, however indirectly. If nothing else, this would severely limit the availability of expert advice to the courts and the government.

As your essay notes, there are still areas of contention on the admission of DNA typing results in criminal cases....to limit the supply of expert testimony by excluding individuals with research support in this area would be a disservice to the courts and to society.

Stephen P. Daiger, Ph.D.
Professor; Medical Genetics Center
The University of Texas Health Science Center

◇◇◇◇◇◇◇◇◇◇◇◇

[F]orensic DNA evidence...is now being used by law enforcement agencies throughout the world. However,...a few vocal critics persist in their attempts to discredit the validity of DNA technology as a reliable and powerful forensic tool. As these challenges have failed on legal and technical grounds, some critics have unfortunately resorted to vengeful personal attacks and political debate to distract and confuse the public and the courts. For a number of these critics who frequently appear as defense experts, their livelihood depends to some degree on keeping a controversy alive.

The primary area still being exploited by critics and defense experts is the interpretation of the population statistics used to estimate the significance of a DNA match. However, the National Academy of Sciences' study endorsed the current statistical methods used by forensic laboratories with recommended modifications in the procedures until the availability of additional world-wide population data....

There are, of course, a number of other specific and general recommendations with respect to the forensic use of DNA technology. Two of those which are being studied carefully deal with developing appropriate models to control and assure the reliability of DNA technology and alternative approaches to statistical interpretation of DNA results. These issues are being studied in consideration of federal and state legislative initiatives and ongoing efforts of other government agencies and professional groups....

DNA analysis has been conducted in over 14,700 criminal cases and admitted in over 612 criminal trials. DNA evidence has been rejected in only five reported cases and seven unreported cases. A few courts have allowed the evidence of a match to be introduced, but reduced or excluded altogether the statistics associated with a match.

There have been 53 appellate decisions in the United States directly addressing the admissibility of forensic DNA test results. Every appellate court addressing the general acceptance or relevancy/reliability of the Restriction Fragment Length Polymorphism (RFLP) technology has ruled that the technology met the applicable standard. Two appellate decisions have remanded the cases back to the trial court for failure of the prosecution to lay an adequate foundation for the admission of the population statistics. Only one appellate court has excluded DNA evidence. In that Minnesota case, the court cited the laboratory's failure to follow certain minimum guidelines it deemed necessary for the admission of DNA test results, as well as previous case precedent which restricted the use of population frequencies in criminal cases.

The forensic use of DNA technology will enhance the effectiveness of law enforcement throughout the United States.... For the full potential of the technology to be realized, a coordinated national effort is essential to establish testing standards which meet stringent criteria for compatibility and reliability. This has been the cornerstone of the FBI's program for the nationwide implementation of DNA technology for use by law enforcement. In working with the foresnic community and other scientists, the FBI has facilitated a consensus in defining such standards....

John W. Hicks
Assistant Director in Charge, Laboratory Division
Federal Bureau of Investigation

◇◇◇◇◇◇◇◇◇◇◇◇

The FBI not only seeks to have its forensic DNA methods approved by courts as "generally accepted" among population geneticists and molecular geneticists, but, more importantly, it asks Congress to pass legislation giving it ultimate authority to set scientific and quality assurance standards for all forensic DNA laboratories that will contribute and receive genetic information from the Bureau's national DNA databank. This is the critical policy issue of the day. It must inform any assessment of the controversy arising from the government's intrusion into the publication and peer review processes of scientific journals, the intimidation of scientists who have criticized the government's forensic DNA methods, and the failure of prosecutors and scientist witnesses to make appropriate disclosures to courts about substantial government grants being awarded to those witnesses.

In this short space it is impossible to summarize all the relevant facts and documents contained in the *U.S. v. Yee, et al.* court papers. We will gladly make them available upon request, and particularly suggest attention be paid to a lengthy affidavit submitted by Dr. Sheldon Krimsky, a leading authority on social and ethical questions in science and technology....

One example we cited of interference in the publication and peer review process was the treatment accorded Dr. Seymour Geisser when he submitted a manuscript criticizing statistical methods of forensic DNA laboratories to the *AJHG*. Three clear problems emerged: 1) There was strong evidence the confidentiality of the peer review process had been breached through disclosures to prosecutors; 2) Dr. Charles Epstein, *AJHG*'s editor, did not know that Dr. Ranjit Chakraborty, the peer reviewer who vituperatively opposed publication, was a recipient of a Justice Department grant intended to establish scientific support for the statistical methods; and 3) Dr. Epstein's own evaluation was admittedly based on how the paper "would be used in court cases," as opposed to a strict consideration of the scientific merits.

The saga of Dr. Geisser's papers has now, however, taken an even more troubling turn. Dr. Geisser was asked by *AJHG* to include in his paper an analysis of FBI population data that had been disclosed in numerous court cases. The FBI objected, taking the position that unpublished FBI population data could not be used by Dr. Geisser unless he made Dr. Budowle of the FBI a co-author, and that Dr. Budowle would then have the right to approve or disapprove every word in the Geisser paper....

The FBI's position reflects a mean spirited form of censorship. Most importantly, this position has been severely criticized by the National Academy of Sciences (NAS) in its recent report, *DNA Technology in Forensic Sciences*:

> Any population databank used in support of forensic DNA typing should be openly available for reasonable scientific inspection. Presenting scientific conclusions in a criminal court is at least as serious as presenting scientific conclusions in an academic paper. According to long-standing and wise scientific tradition, the data underlying an important scientific conclusion must be freely available, so that others can evaluate the results and *publish* their own findings, whether in support or disagreement. There is no excuse for secrecy in raw data....If scientific evidence is not yet ready for both scientific scrutiny and public re-evaluation by others, it is not yet ready for court. (NAS Rpt, at 3-23.)

Another example of interference with the publication and peer review process was the FBI's efforts to influence improperly the National Academy's DNA report. It is now undisputed that John Hicks, Director of the FBI Laboratory, improperly obtained from two members of the NAS Committee a confidential draft of the report concerning statistical methods that was very critical of the FBI. Hicks subsequently wrote an unsolicited reply that NAS staff say they did not distribute to the Committee. In an FBI crime

Professional Ethics Report
Directorate for Science and Policy Programs
American Association for the Advancement of Science
1333 H Street, NW, Washington, DC 20005

Non-Profit Org.
U.S. Postage Paid
Washington, D.C
Permit no. 7304

laboratory publication, Hicks writes that an "internal investigation" was conducted of our allegations, including presumably our allegations against him, and that no evidence of wrongdoing was found. We wonder who conducted that investigation and how it could reach such conclusions given the clear evidence that Hicks himself inexcusably violated the peer review and publication processes of the NAS.

Finally, as far as the Justice Department grants to Drs. Caskey and Daiger are concerned, a few points need to be made: 1) FBI agent Kearney sat on the panels that awarded these grants, so that any claim they were awarded in a fashion analogous to NIH grants is simply absurd; 2) That Drs. Daiger and Caskey did not feel compromised or even financially rewarded by these grants is ethically irrelevant to the obligation to disclose; and 3) There is no legal justification whatsoever for failing to disclose these grants to the court.

The NAS pointedly recommends the Department of Health and Human Services, and not the Department of Justice (DOJ), be given responsibility for accrediting forensic DNA laboratories, and that this be done through contracting with private professional organizations. After concluding that the DOJ (i.e., the FBI) "lacks expertise in quality assurance and quality control in molecular genetics," the NAS observed "the DOJ may be perceived as an advocate for application of the technology," and "[o]versight by DOJ may not be perceived as providing adequate assurance to the public or to a defendant facing prosecution by DOJ." (NAS Rpt., at 4-12.)

The recommendations of the NAS should be strongly supported in Congress by all in the legal and scientific communities who care about the responsible, reliable, and ethical use of forensic DNA technology.

Barry Scheck, Esq.
Director, Clinical Education
Benjamin N. Cardozo School of Law

RESOURCES

Responsible Science: Ensuring the Integrity of the Research Process (Washington, DC: National Academy of Sciences Press, 1992); $24.95 plus $3.00 shipping. To order, call the Press at (800) 624-6242.

The AAAS has published, *The Genome, Ethics and the Law: Issues in Genetic Testing*. The book includes an overview paper of the discussion that took place at a conference in Berkeley Springs, West Virginia, June 14-16, 1991, and three background papers addressing the scientific basis of genetic testing, the ethical implications of recent and anticipated advances, and the legal issues those advances raise. Single copies are free. Contact Alexander Fowler, AAAS Science and Policy Programs, 1333 H Street NW, Washington, DC, 20005; (202) 326-6600.

Support from the following professional societies is gratefully acknowledged:

American Academy of Environmental Engineers
American Academy of Otolaryngology—Head and Neck Surgery
American Anthropological Association
American Association of Engineering Societies
American Association of University Professors
American Chemical Society
American Institute of Aeronautics and Astronautics
American Institute of Chemists
American Medical Association
American Occupational Therapy Association
American Pharmaceutical Association
American Philosophical Association
American Political Science Association
American Psychological Association
American Society for Engineering Education
American Society for Microbiology
American Society for Photogrammetry and Remote Sensing
American Society for Public Administration
American Society of Mechanical Engineers
American Sociological Association
American Speech-Language-Hearing Association
American Statistical Association
Association of American Medical Colleges
Botanical Society of America
Council of Biology Editors
Council of Scientific Society Presidents
Ecological Society of America
Federation of Behavioral, Psychological, & Cognitive Sciences
Institute of Electrical and Electronics Engineers
National Association of Social Workers
National Society of Professional Engineers
Society for Computer Simulation
Society for Epidemiologic Research
Society for the Scientific Study of Sex
Society of Professional Archeologists
University Center for Human Values, Princeton University
Vaughen Enterprises Ethics Consulting Services

)N V BILITY

tudy: effect of
Hypertension 2

e of metabolic
Phys Anthropol

pidemiology of
J Hum Genet

12

Generalized Occupancy Problem and Its Applications in Population Genetics

RANAJIT CHAKRABORTY

Analysis of categorical observations constitutes one of the principal ways of handling genetic data, evidenced in the contributions of the honoree of this symposium in relation to his studies on inbreeding effects (Schull and Neel, 1965), genetic effects of radiation exposure (Schull et al., 1981) and adaptation effects of high altitude hypoxia (Schull and Rothhammer, 1990). In the context of hypothesis testing based on specific genetic models in such data analysis, often we need to contrast frequencies of different endpoints in samples of unequal sizes or predict expected frequencies in populations from observations in samples of small sampling proportions. For example, Neel (Chapter 4, this volume) illustrates one such problem where the frequencies of different genetic variants are contrasted in two Japanese cities, Hiroshima and Nagasaki, where the sample sizes are widely different. In a previous study (Chakraborty et al., 1988) the issue of sample size adjustment was addressed by deriving the expectation and variance of the observed number of alleles in samples of fixed sizes.

In this chapter, a more complete solution will be given from a combinatorial approach, whereby the sampling distribution can be completely specified for an arbitrary sample size. Although the problem is formulated and solved in general terms of a multinomial distribution with known number of classes, it will be illustrated that the general theory can be used in solving several types of genetic problems, such as, non-randomness of mutagen-induced mutations across loci; tests of Hardy-Weinberg equilibrium in the presence of a large number of alleles in samples of comparatively moderate sizes; and global tests of gametic phase disequilibria within a defined DNA segment. Because the combinatorial approach used in this context emerges from the combinatorics of the classical occupancy problem, I call this class of problems *Generalized Occupancy Problems* in population genetics.

FORMULATION OF THE GENERALIZED OCCUPANCY PROBLEM

Consider a multinomial distribution with K classes (K can be arbitrarily large), with class probabilities represented by the vector $\pi' = (\pi_1, \pi_2, \ldots, \pi_K)$.

Obviously, the elements of π form a simplex on a K-dimensional space satisfying

$$0 < \pi_i < 1, \quad \text{and} \sum_{i=1}^{K} \pi_i = 1. \qquad [12.1]$$

When a random sample (with replacement) of size n is drawn from such a distribution, several characteristics of the sample are of interest; e.g., the number of classes represented in the sample, or the number of classes are represented with a specified number of sampling units in each of these classes; and the sampling distributions of such variables. For brevity, I shall only discuss the sampling distribution of the observed number of classes, although the combinatorial solution can be easily extended to many other variables. Denote the observed number of classes that are represented in the sample by X. Obviously, X is a random variable that can take values between 1 and min (K,n), since in the event $K > n$, we cannot observe more than n classes in any sample of size n. Before attempting to solve the problem, note several interesting properties of the random variable X.

First, when all π_i's are equal ($\pi_i = 1/K$ for all i), a situation more commonly known as the classical occupancy problem, Arnold and Beaver (1988) showed that X is a sufficient statistic for the parameter K, and its sampling distribution is given by

$$P_{[m]} = \text{Prob.}\,(X = m) = \binom{K}{m} \frac{m!\, S_n^{(m)}}{K^n}, \qquad [12.2]$$

for $m = 1,2,\ldots,\min(K,n)$, where $S_n^{(m)}$ is a Stirling number of the second kind (Abramowitz and Stegun, 1965; p 824). Also note that in the case of the classical occupancy problem, a value of X together with the sample size n, and number of possible classes K, uniquely determines the number (R) of sampling units that are "repeated" within the observed classes in a sample. In this context, a sampling unit is called a "repeat" if the class in which this sampling unit belongs already contains another sampling unit in a given sample, Therefore, the sampling distribution of R can be uniquely specified from that of X, and going through the computation of equation [12.2] is preferable, since X is a sufficient statistic for the only parameter (K) of the multinomial distribution with equi-probable classes.

These abstract definitions may be better understood through the following example. Consider an experiment where 20 ($= n$) mutants are observed which could belong to 30 ($= K$) loci numbered from 1 through 30. Suppose that there are 3 loci each containing 4 mutants (i.e., $n_1 = n_2 = n_3 = 4$), two loci containing 2 mutants each ($n_4 = n_5 = 2$), and four loci containing one mutant each ($n_6 = n_7 = n_8 = n_9 = 1$). The remaining loci (10 through 30) do not contain any mutants (i.e., $n_{10} = \cdots = n_{30} = 0$). In this example, the observed number of loci containing at least one mutant (X) is 9, and the number of "repeat" mutants is 11. Noting that $11 = 20 - 9$, (i.e., $R = n - X$, more generally), it is clear that the probability distribution of R is completely specified by that of X for any fixed sample size (n).

Second, in the general case of unequal π_i's, the expectation and variance

of X have been derived using the indicator variable approach (Chakraborty et al., 1988), which are simple functions of n, K, and the π_i's. I shall show in the following that both of these properties are compatible with the general probability distribution function derived by the combinatorial approach used here.

COMBINATORIAL SOLUTION OF THE PROBABILITY FUNCTION OF X

Instead of working with the variable X, the observed number of non-empty classes, it is easier to work with its complement, $Y = K - X$, which is equivalent to the number of classes not represented in a sample of size n. Define events $A_1, A_2, \ldots, A_K$, such that A_i is the collection of all partitions of n into K segments, such that $n_i = 0$ (but $\sum n_i = n$). In other words, A_i represents the collection of sample points where the ith class ($i = 1, 2, \ldots, K$) remains empty in a specific sample. Note that A_i's are not exclusive of each other; i.e., there can be sample configurations where more than one n_i can be simultaneously zero.

For sampling with replacement, the following equations hold:

$$P_i = \text{Prob}(A_i) = (1 - \pi_i)^n, \qquad [12.3a]$$

$$P_{ij} = \text{Prob}(A_i A_j) = (1 - \pi_i - \pi_j)^n, \qquad [12.3b]$$

$$P_{ijk} = \text{Prob}(A_i A_j A_k) = (1 - \pi_i - \pi_j - \pi_k)^n, \qquad [12.3c]$$

etc., for all $i \neq j \neq k = 1, 2, \ldots, K$.

Following Feller (1968, p 99), define a sequence of summations $\{T_1, T_2, \ldots, T_K\}$ where

$$T_1 = \sum_i p_i, \; T_2 = \sum_i \sum_j p_{ij}, \; T_3 = \sum_i \sum_j \sum_k p_{ijk}, \text{ etc.}$$

where the summations are taken such that $i < j < k < \cdots \leqslant K$, so that each combination appears once and only once; hence, the summation $T_r (1 \leqslant r \leqslant K)$ contains ${}^K C_r$ terms. The last term T_K reduces to only one term,

$$T_K = \text{Prob}(A_1 A_2 \cdots A_K) = P_{123\cdots K},$$

which is the probability of simultaneous occurrences of all K events A_1 through A_K. Invoking condition [12.1] on equations [12.3a–c], we note that $T_K = 0$, and furthermore,

$$T_{K-1} = \sum_{i=1}^{K} \pi_i^n, \qquad [12.4a]$$

$$T_{K-2} = \sum_{i>j=1}^{K} \sum^{K} (\pi_i + \pi_j)^n, \qquad [12.4b]$$

$$T_{K-3} = \sum_{i>j>k=1}^{K} \sum^{K} \sum^{K} (\pi_i + \pi_j + \pi_k)^n, \qquad [12.4c]$$

etc.

Applying Feller's theorem (Feller, 1968, p 106), we obtain

$$P_{[K-m]} = \text{Prob}(X = K - m) = \text{Prob}(Y = m)$$
$$= \sum_{i=m}^{K} (-1)^{i-m} \binom{i}{m} T_i, \qquad [12.5]$$

for $K - \min(K, n) \leqslant m \leqslant K - 1$, giving the sampling distribution of X, the number of non-empty classes in a sample of size n. Note that in [12.5], T_0 is conventionally defined as unity (see also Feller, 1968).

EXPECTATION AND VARIANCE OF THE NUMBER OF NON-EMPTY CLASSES IN A SAMPLE

As mentioned before, the expectation and variance of X can be derived by an indicator variable approach. For this we define K indicator variables, $Y_1, Y_2, \ldots, Y_K$, so that

$$Y_i = \begin{cases} 1, \text{ if the } i\text{th class is unobserved in the sample,} \\ 0, \text{ otherwise.} \end{cases}$$

Note that $X = K - Y$, and $Y = \sum_{i=1}^{K} Y_i$.

Therefore, the expectation of X is given by

$$E(X) = K - E(Y) = K - \sum_{i=1}^{K} E(Y_i)$$
$$= K - \sum_{i=1}^{K} \text{Prob}(Y_i = 1)$$
$$= K - \sum_{i=1}^{K} (1 - \pi_i)^n = K - T_1, \qquad [12.6]$$

derived by Emigh (1983) and Chakraborty et al. (1988).

Furthermore, the variance of X is given by

$$V(X) = V(Y) = \sum_{i=1}^{K} V(Y_i) + \sum_{i \neq j=i}^{K} \sum^{K} \text{Cov}(Y_i, Y_j). \qquad [12.7]$$

Since Y_i's are Bernoulli variables,

$$V(Y_i) = E(Y_i^2) - [E(Y_i)]^2 = (1 - \pi_i)^n [1 - (1 - \pi_i)^n], \qquad [12.8]$$

and

$$\text{Cov}(Y_i, Y_j) = E(Y_i Y_j) - E(Y_i) \cdot E(Y_j)$$
$$= (1 - \pi_i - \pi_j)^n - (1 - \pi_i)^n (1 - \pi_j)^n. \qquad [12.9]$$

Substituting equations [12.8] and [12.9] into equation [12.7], we obtain

$$V(X) = \sum_{i=1}^{K} (1-\pi_i)^n [1-(1-\pi_i)^n] + \sum_{i\neq j=1}^{K}\sum^{K} [(1-\pi_i-\pi_j)^n - (1-\pi_i)^n(1-\pi_j)^n]$$

$$= \sum_{i=1}^{K} (1-\pi_i)^n \cdot [1-\sum(1-\pi_i)^n] + \sum_{i\neq j=1}^{K}\sum^{K} [(1-\pi_i-\pi_j)^n \qquad [12.10]$$

$$= T_1(1-T_1) + T_2.$$

Note that when n is large, and each π_i small, we may approximate each term of the summations T_r's [12.3a–c] by $(1-\pi_i)^n \approx e^{-n\pi_i}$, $(1-\pi_i-\pi_j)^n \approx e^{-n(\pi_i+\pi_j)}$, etc., so that the variance of X can be approximated by

$$V(X) \approx \sum_{i=1}^{K} e^{-n\pi_i}(1-e^{-n\pi_i}), \qquad [12.11]$$

as shown in Chakraborty et al. (1988). Equation [12.10] is, however, exact, and not difficult to compute numerically even if the number of classes is large.

Let us now establish that equations [12.6] and [12.10] are compatible with the probability function [12.5]. To do this, first note that invoking equation [12.5] we have

$$\sum_m \text{Prob}(Y=m) = \sum_m \sum_{i=m}^{K} (-1)^{i-m}\binom{i}{m} T_i$$

$$= \sum_{i=0}^{K}\left[\sum_{m=0}^{i}(-1)^{i-m}\binom{i}{m}\right] T_i$$

$$= 1 + \sum_{i=1}^{K}\left[\sum_{m=0}^{i}(-1)^{i-m}\binom{i}{m}\right] T_i,$$

since $T_0 = 1$. Furthermore, the summation within the parenthesis is the binomial expansion of $(1-1)^i$, for $i = 1, 2, \ldots, K$. Therefore, we establish that [12.5] is a proper probability function since the entire probability mass equals unity.

The expectation of Y, from [12.5], then becomes

$$E(Y) = \sum_m m \cdot \text{Prob}(Y=m)$$

$$= \sum_m \sum_{i=m}^{K} (-1)^{i-m} m \binom{i}{m} T_i$$

$$= \sum_m \left[\sum_{i=m}^{K} (-1)^{i-m}\binom{i-1}{m-1} i T_i\right]$$

$$= T_1 + \sum_{i=2}^{K}\left[\sum_{m-1=0}^{i-1}(-1)^{i-m}\binom{i-1}{m-1}\right] i T_i \qquad [12.12]$$

$$= T_1,$$

since the summation within the parenthesis vanishes for all $i = 2, 3, \ldots, K$. Similar algebraic manipulations show that the second moment of Y can be written as

$$E(Y^2) = E(Y) + \sum_m m(m-1)\cdot \text{Prob}(Y = m)$$
$$= T_1 + 2T_2. \qquad [12.13]$$

Since $X = K - Y$, I therefore complete the proof of equations [12.6] and [12.10] starting from [12.5]. Furthermore, this computational logic also yields the rth factorial moment of Y, $\mu_{[r]}(Y)$, given by

$$\mu_{[r]}(Y) = E[Y(Y-1)\cdots(Y-r+1)] = r!\cdot T_r, \qquad [12.14]$$

for any $r \geqslant 1$, giving the complete characterization of the probability function [12.5] through its moments.

When all π_i's are equal (i.e., $\pi_i = 1/K$ for all i), note that

$$T_1 = K[(K-1)/K]^n, \text{ and } T_2 = K(K-1)\cdot[(K-2)/K]^n/2,$$

and hence,

$$E(X) = K[1 - \{(K-1)/K\}^n], \qquad [12.15]$$

and

$$V(X) = K\cdot\{(K-1)/K\}^n\cdot[1 - K\cdot\{(K-1)/K\}^n]$$
$$+ K(K-1)\cdot\{(K-2)/K\}^n, \qquad [12.16]$$

which are derived in Arnold and Beaver (1988) while studying the sampling properties of the observed number of classes in the context of the classical occupancy problem. When $\pi_i = 1/K$, note also that the summations $\{T_i\}$ take the form

$$T_i = \binom{K}{i}\{(K-i)/K\}^n,$$

so that the probability function [12.5] reduces to

$$P_{[K-m]} = \sum_{i=m}^{K} (-1)^{i-m}\binom{i}{m}\binom{K}{i}(K-i)^n/K^n$$
$$= \sum_{i=m}^{K} (-1)^{i-m}\binom{K}{K-m}\binom{K-m}{i-m}(K-i)^n/K^n$$
$$= \binom{K}{K-m}\cdot\sum_{i=m}^{K} (-1)^{i-m}\binom{K-m}{i-m}(K-i)^n/K^n$$
$$= \binom{K}{K-m}\cdot\frac{m!\,S_n^{(K-m)}}{K^n},$$

invoking the definition of a Stirling number of the second kind (Abramowitz and Stegun, 1965; p 824).

The above derivations, therefore, show that the sampling distribution of the number of observed classes in a finite sample can be analytically specified for any arbitrary multinomial distribution. This generalizes the special case solution of the problem discussed in Arnold and Beaver (1988) in the context of the classical occupancy problem. The algebraic solutions of other relevant random variables (e.g., the number of classes containing a specified number of sampling units within each of them) are also similar, although more cumbersome to compute numerically.

APPLICATIONS

I mention here three applications of this generalized occupancy problem, each of which has considerable genetic implications.

Are mutagen-induced mutations equally likely to occur at all loci?

Hanash et al. (1988) recently demonstrated that somatic cell gene mutations altering protein structure do not occur with equal probability at all loci when cultured human lymphoblastoid cell lines are treated with mutagens like ethylnitrosourea. To show this, they used the technique of two-dimensional polyacrylamide gel electrophoresis, and found 65 mutants occurring at 49 of the 263 loci scored in their experiments. The locus-specific distributions of the mutation frequencies in their work were: three mutants observed at each of five loci ($n_1 = n_2 = n_3 = n_4 = n_5 = 3$), two mutants at each of six loci ($n_6 = \cdots = n_{11} = 2$), and one mutant at each of 38 loci ($n_{12} = \cdots = n_{49} = 1$). No mutation was detected at each of the remaining 214 loci ($n_{50} = \cdots = n_{263} = 0$). The total number of mutations ($n = 65$) was, thus, distributed in $K = 263$ classes. The null hypothesis to be tested in H_0: $\{\pi_i =$ the probability of mutation occurring at the ith locus $= 1/K = 1/263$, for all $i\}$. In their work, the authors defined the concept of "repeat" mutations (R), noting that 16 mutations occurred at loci each of which contained already one mutation (i.e., $R = n - X$, where X is the number of loci containing at least one mutant). Under the null hypothesis of equiprobable mutation frequencies across loci, the number of "repeat" mutations should be small, since $K = 263$ is much larger than the sample size $n = 65$. Through a simulation experiment of the occupancy problem, they determined that the probability of 16 or more "repeat" mutants is below 0.0005, and hence, they conclude that mutagen-induced mutations are not equally likely to occur at all loci.

The theory described above provides a complete analytical solution, avoiding any simulation. Note that the observed value of X in the above experiment is $m = 49$, and hence the observed number of empty classes (number of loci having no occurrences of mutations, Y) is 214 ($= 263 - 49$). With $n = 65$, and $K = 263$, the range of possible X values in this experiment

is 1 through 65, and consequently the possible values of Y (number of empty classes) are from 198 (= 263 − 65) through 262. Figure 12.1 shows the exact probability distribution (shown by the histogram) of non-empty classes (X, the number of loci at which one or more mutants can occur) under the null hypothesis of equiprobable mutations across the 263 loci. The observed value of X ($m = 49$) is marked with an arrow, the area below which is the total probability ($P = 0.0005$) of all other sample configurations which represent deviations from the hypothesis more extreme under the null hypothesis. Note that Hanash et al.'s simulation also resulted in a P-value consistent with the present result, suggesting that the qualitative conclusion of their analysis is the same as the one obtained by the present analytical solution.

Since under the null hypothesis ($\pi_i = 1/K = 1/263$ for all i) the observed number of non-empty classes (X) is the sufficient statistic for K, I further checked to see if the exact distribution of X can be approximated by any standard distribution. Employing equations [12.5] and [12.6], the mean and variance of X for this experiment are given by $E(X) = 57.687$ and $V(X) = 5.289$. The smooth curve of Figure 12.1 represents the density function

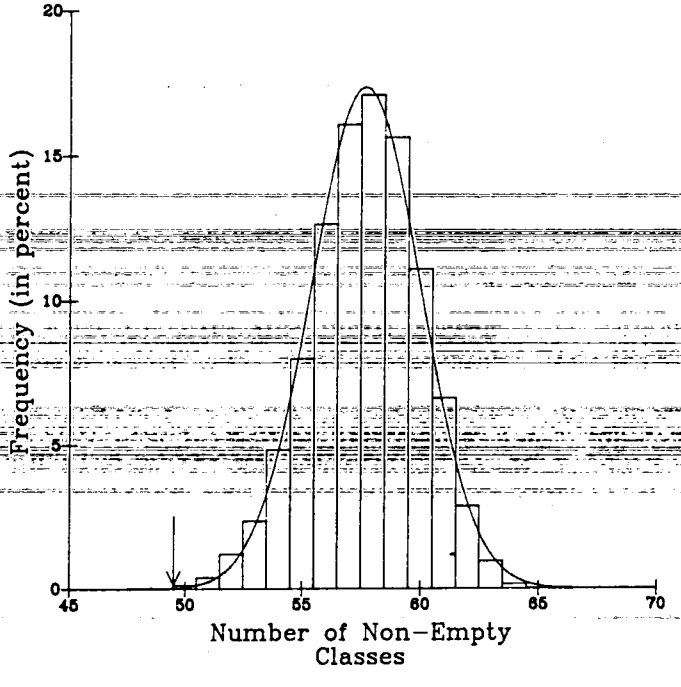


Figure 12.1. The sampling distribution of the number of non-empty classes in a sample of 65 observations drawn from a multinomial distribution with 263 equi-frequent classes. The histogram shows the exact analytical distribution evaluated from equation [12.5] and the smooth curve is the normal approximation using the expectation and variance given in equations [12.6] and [12.10]. The arrow indicates the observed value $m = 49$ (See text for details).

of a normal distribution with these mean and variance values. Note that the normal deviate for $m = 49$, then becomes $z = -3.78$, giving a corresponding P value of 0.00002, which is somewhat smaller than the p value obtained from the exact distribution. Nevertheless, the normal approximation is quite satisfactory, when compared with the histogram shown in Figure 12.1.

Test of Hardy-Weinberg expectation based on the observed numbers of distinct genotyeps in a finite sample

The generalized occupancy problem can also be used to examine whether or not the genotype distribution of a given number of alleles follows the Hardy-Weinberg expectation (HWE). Generally, this is done by either a likelihood ratio test or a goodness of fit chi-square test, contrasting the observed and expected frequencies of all possible genotypes. However, there are occasions when the number of alleles are so large that many of the genotypes are either not observed in a sample, or the observed frequencies of several genotypes are so small that the large sample approximation of these test statistics in unwarranted. The recently discovered VNTR polymorphisms provide examples of this nature, where the number of possible alleles is often so large that no reasonably sized survey can encompass all possible genotypes in any given sample. Assuming that there are K segregating alleles at a locus, there are K possible homozygote genotypes and $K(K-1)/2$ possible distinct heterozygote genotypes that can be encountered. One might ask, what would be the distribution of the numbers of distinct genotypes (of homozygote and heterozygote types, separately) observed in a sample of n individuals. Under the Hardy-Weinberg expectation of genotypic probabilities given by p_i^2 for homozygotes and $2p_ip_j$ for heterozygotes, where p_i represents the allele frequencies in the population, we can use the above analytical formulation to compute the exact distributions of the distinct numbers of homozygote and heterozygote genotypes seen in a sample.

Figure 12.2 shows a numerical example of such computations. Deka et al. (1991) recently surveyed the New Guinea population for VNTR polymorphisms at six loci. At the D1S76 locus, they discovered 6 alleles in a sample of 35 individuals. Gene counting showed that in the sample of 70 genes at this locus, the allele counts of these 6 alleles are 1, 3, 7, 9, 25, and 25. In total they observed 20 heterozygous individuals (consisting of 7 distinct genotypes). However, under the HWE assumption, the expected frequency of heterozygotes from the above allele counts is 25.4, showing a significant deficiency of heterozygotes ($P < 0.05$). Since the observed numbers of distinct homozygote and heterozygote genotypes in their sample were 4 and 7, respectively, we can ask if these observations deviate from their respective expectations under the HWE assumption. Figure 12.2a shows the exact distribution of the observed number of distinct homozygote genotypes (drawn as histogram) and Figure 12.2b gives the same for the observed number of distinct heterozygote genotypes, under the HWE assumption for the given allele frequencies. The arrows represent the observed statistics. Clearly, the

observed number of distinct homozygote genotypes ($m = 4$) is not at variance with the HWE, since the probability of observing four or more distinct homozygote genotypes is 0.411. Under HWE the probability of observing seven or less distinct heterozygote genotypes is 0.017, suggesting that a significant deficiency is observed in the total number of heterozygotes as well as in the number of distinct heterozygote genotypes. Of course, as in the case of traditional likelihood ratio or chi-square tests, this test cannot ascertain the real cause of such heterozygote deficiency.

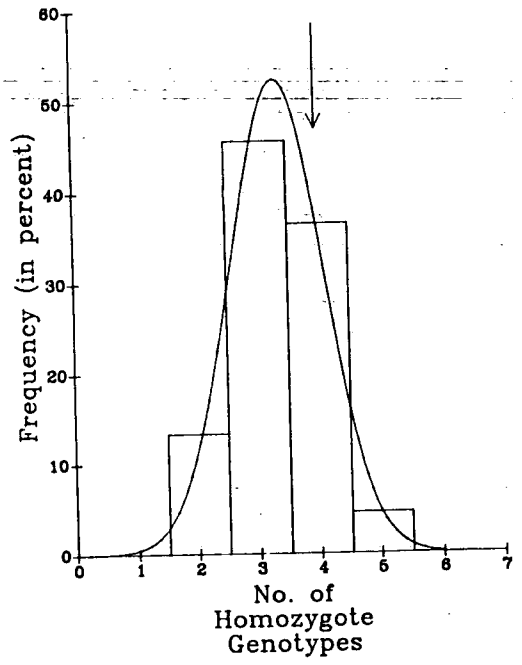
From equations [12.6] and [12.10], the mean and variance of the number of distinct genotypes were computed as 3.329 and 0.578 for the homozygotes, and 10.106 and 1.652 for the heterozygote genotypes, respectively. The expected distributions under the normality approximation are also shown by the smooth curves in both panels of Figure 12.2. As in the earlier case, it shows that the normal approximation is fairly adequate for the distribution of distinct heterozygote genotypes. This is not so for the homozygotes because of the narrow range of variation in the number of distinct homozygote genotypes. Under the normality approximation, the normal deviate corresponding to observing seven or less distinct heterozygote genotypes is $z = -2.41$, with a P-value of 0.008, which is again smaller than the exact P-value shown above.

Global test of disequilibrium based on multiple-locus haplotype data

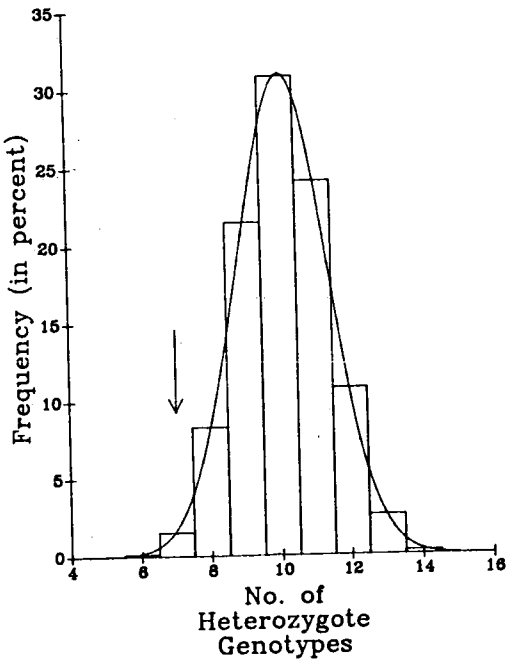
As a third application, consider the haplotype frequency data surveyed by Wainscoat et al. (1986) at the β-globin gene cluster detected by five polymorphic restriction sites, at each of which there are two segregating alleles. This results in $2^5 = 32$ possible haplotypes at this gene region, but in a sample of 55 chromosomes sampled from a Polynesian population, these authors found only 5 observed haplotypes (see Table 1 of Wainscoat et al., 1986). One might ask, what is the expected distribution of the number of haplotypes given that these five sites are independently segregating. Figure 12.3 shows the exact distribution (represented by histogram), following the general analytical formulation (equation 12.5), where the expected haplotype frequencies are assumed to follow the independent segregation rule. Clearly, almost the entire distribution is to the right of the observed number ($m = 5$) of haplotypes, giving a rare probability of observing five or less haplotypes ($P < 10^{-5}$), suggesting that the observed number of haplotypes is incompatible with the assumption of independent segregation. Under independent

Figure 12.2. The sampling distributions of the number of distinct homozygote (A) and heterozygote (B) genotypes at the D1S76 VNTR locus in a sample of 35 individuals from Papua New Guinea (Deka et al., 1990) under the assumption of Hardy-Weinberg equilibrium frequencies of genotypic proportions. The histograms are exact computations (equation 12.5) and the smooth curves are the normal approximations based on mean and variance, given in equations [12.6] and [12.10]. The arrows indicate the observed numbers of distinct homozygote (4) and heterozygote genotypes (7) found in the sample.

A


Frequency (in percent)
60
50
40
30
20
10
0
0 1 2 3 4 5 6 7
No. of
Homozygote
Genotypes

B


Frequency (in percent)
35
30
25
20
15
10
5
0
4 6 8 10 12 14 16
No. of
Heterozygote
Genotypes

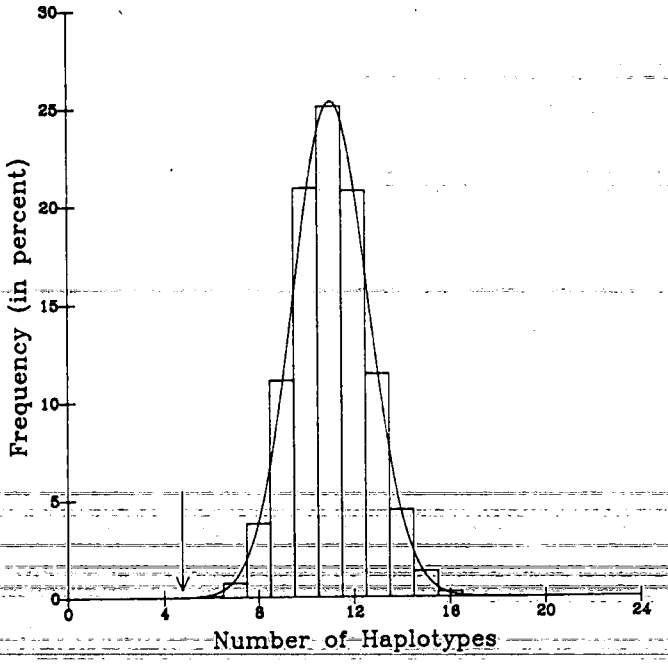


Figure 12.3. The sampling distribution of the number of DNA haplotypes at the β-globin gene cluster, defined by 5 restriction site polymorphisms (Wainscoat et al., 1986), in a sample of 55 chromosomes from a Polynesian population under the assumption of complete linkage equilibrium. The histogram is the exact distribution based on equation [12.5] and the smooth curve is its normal approximation based on mean and variance given by equations [12.6] and [12.10]. The arrow indicates the observed number of 5 different haplotypes found in the sample.

segregation, the expected mean and variance of the observed number of haplotypes in a sample of 55 chromosomes are 11.059 and 2.477, respectively. The normal approximation of the sampling distribution is again shown by the smooth curve of Figure 12.3. While the normal approximation appears satisfactory, the normal deviate corresponding to the observed number 5 is $z = -3.85$, giving a P-value of 0.00006, which is larger than the exact P-value. Note that Blanton and Chakravarti (1987) suggested this global test for disequilibrium, although unlike here their sampling distribution was obtained by simulation.

DISCUSSION

The analytical theory presented here along with the specific applications indicate that the generalized occupancy problem has a number of interesting applications in population genetics. This is particularly true in the context of sparse data, where by the very nature of the problem, the exact sampling

distribution must be evaluated and no adequate large sample approximation is available. This theory enables comparison of occurrences of several biological endpoints in cross-survey comparisons, adjusting for sample size differences, as shown in Chakraborty et al. (1988), and in this chapter three other applications are mentioned. In the first application, no loss of information is attendent to the consideration of the sample statistic X, the observed number of non-empty classes, since under the equiprobable mutation rate (across loci), X is a sufficient statistic of the only parameter of the underlying distribution. In the other two cases, the consideration of observed number of classes raise the possibility of some loss of information, since the frequencies of the different observed categories do not enter into the present analysis. However, when the number of categories is large compared to the sample size, most of the observed categories are likely to have one or a few sample points in them and such loss of information is not critical. As shown through the applications here, the exact distribution evaluation does indeed detect deviations from the null hypothesis even when the sample size is larger than the total number of possible classes. A comprehensive power analysis of this approach to deal with such specific genetic applications will be attempted in the future.

In closing I should note the close resemblance of the methodology of this presentation with a percentage testing problem that Schull and I resolved several years ago (Chakraborty and Schull, 1976), where we evaluated the sampling distribution of the numbr of loci with reference to which a randomly accused man could be excluded if this man is not the father of a child born to a specific mother. Although all of these problems can be resolved by simulation, the real advantage of the present theory is that such problems can be addressed analytically, avoiding the natural bias and tediousness of computer simulations.

ACKNOWLEDGMENTS

This chapter is dedicated to Professor William J. Schull, whose encouragement and motivation primarily led to this work. This research is partially supported by Grants GM 41399 and 90-IJ-CX-0038 from the National Institutes of Health and National Institute of Justices, respectivcly.

REFERENCES

Abramowitz M, Stegun IA (1965) *Handbook of Mathematical Functions.* New York, Dover.

Arnold BC, Beaver RJ (1988) Estimation of the number of classes in a population. *Biometrical Journal* 30:413–424.

Blanton SH, Chakravarti A (1987) A global test of linkage disequilibrium. *Amer J Hum Genet* 41:A250.

Chakraborty R, Schull WJ (1976) A note on the distribution of the number of exclusions to be expected in paternity testing. *Amer J Hum Genet* 28:615–618.

Chakraborty R, Smouse PE, Neel JV (1988) Population amalgamation and genetic variation: observations on artificially agglomerated tribal populations of Central and South America. *Amer J Hum Genet* 43:709–725.

Deka R, Chakraborty R, Ferrell RE (1991) A population genetic study of six VNTR loci in three ethnically defined populations. *Genomics* 11:83–92.

Emigh TH (1983) On the number of observed classes from a multinomial distribution. *Biometrics* 39:485–491.

Feller W (1968) *An Introduction to Probability Theory and its Applications.* New York, Wiley.

Hanash SM, Boehnke M, Chu EHY, Neel JV, Kuick RD (1988) Nonrandom distribution of structural mutants in ethylnitrosourea treatment of cultured human lymphoblastoid cells. *Proc Natl Acad Sci USA* 85:165–169.

Schull WJ, Neel JV (1965) *The Effects of Inbreeding on Japanese Children.* New York, Harper and Row.

Schull WJ, Otake M, Neel JV (1981) Genetic effects of the atomic bombs: a reappraisal. *Science* 213:1220–1227.

Schull WJ, Rothhammer F (1990) *The Aymara: Strategies in Human Adaptation to a Rigorous Environment.* Amsterdam, Kluwer Academic Publishers.

Wainscoat JS, Hill AVS, Boyce AL, Flint J, Hernandez M, Thein SL, Old JM, Lynch JR, Falusi AG, Weatherall DJ, Clegg JB (1986) Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. *Nautre* 319:491–493.

ON ABILITY

study: effect of
Hypertension 2

ıe of metabolic
Phys Anthropol

pidemiology of
J Hum Genet

12

Generalized Occupancy Problem and Its Applications in Population Genetics

RANAJIT CHAKRABORTY

Analysis of categorical observations constitutes one of the principal ways of handling genetic data, evidenced in the contributions of the honoree of this symposium in relation to his studies on inbreeding effects (Schull and Neel, 1965), genetic effects of radiation exposure (Schull et al., 1981) and adaptation effects of high altitude hypoxia (Schull and Rothhammer, 1990). In the context of hypothesis testing based on specific genetic models in such data analysis, often we need to contrast frequencies of different endpoints in samples of unequal sizes or predict expected frequencies in populations from observations in samples of small sampling proportions. For example, Neel (Chapter 4, this volume) illustrates one such problem where the frequencies of different genetic variants are contrasted in two Japanese cities, Hiroshima and Nagasaki, where the sample sizes are widely different. In a previous study (Chakraborty et al., 1988) the issue of sample size adjustment was addressed by deriving the expectation and variance of the observed number of alleles in samples of fixed sizes.

In this chapter, a more complete solution will be given from a combinatorial approach, whereby the sampling distribution can be completely specified for an arbitrary sample size. Although the problem is formulated and solved in general terms of a multinomial distribution with known number of classes, it will be illustrated that the general theory can be used in solving several types of genetic problems, such as, non-randomness of mutagen-induced mutations across loci; tests of Hardy-Weinberg equilibrium in the presence of a large number of alleles in samples of comparatively moderate sizes; and global tests of gametic phase disequilibria within a defined DNA segment. Because the combinatorial approach used in this context emerges from the combinatorics of the classical occupancy problem, I call this class of problems *Generalized Occupancy Problems* in population genetics.

FORMULATION OF THE GENERALIZED OCCUPANCY PROBLEM

Consider a multinomial distribution with K classes (K can be arbitrarily large), with class probabilities represented by the vector $\pi' = (\pi_1, \pi_2, \ldots, \pi_K)$.

Obviously, the elements of π form a simplex on a K-dimensional space satisfying

$$0 < \pi_i < 1, \quad \text{and} \quad \sum_{i=1}^{K} \pi_i = 1. \qquad [12.1]$$

When a random sample (with replacement) of size n is drawn from such a distribution, several characteristics of the sample are of interest; e.g., the number of classes represented in the sample, or the number of classes are represented with a specified number of sampling units in each of these classes; and the sampling distributions of such variables. For brevity, I shall only discuss the sampling distribution of the observed number of classes, although the combinatorial solution can be easily extended to many other variables. Denote the observed number of classes that are represented in the sample by X. Obviously, X is a random variable that can take values between 1 and $\min(K,n)$, since in the event $K > n$, we cannot observe more than n classes in any sample of size n. Before attempting to solve the problem, note several interesting properties of the random variable X.

First, when all π_i's are equal ($\pi_i = 1/K$ for all i), a situation more commonly known as the classical occupancy problem, Arnold and Beaver (1988) showed that X is a sufficient statistic for the parameter K, and its sampling distribution is given by

$$P_{[m]} = \text{Prob.}(X = m) = \binom{K}{m} \frac{m! S_n^{(m)}}{K^n}, \qquad [12.2]$$

for $m = 1, 2, \ldots, \min(K, n)$, where $S_n^{(m)}$ is a Stirling number of the second kind (Abramowitz and Stegun, 1965; p. 824). Also note that in the case of the classical occupancy problem, a value of X together with the sample size n, and number of possible classes K, uniquely determines the number (R) of sampling units that are "repeated" within the observed classes in a sample. In this context, a sampling unit is called a "repeat" if the class in which this sampling unit belongs already contains another sampling unit in a given sample, Therefore, the sampling distribution of R can be uniquely specified from that of X, and going through the computation of equation [12.2] is preferable, since X is a sufficient statistic for the only parameter (K) of the multinomial distribution with equi-probable classes.

These abstract definitions may be better understood through the following example. Consider an experiment where 20 ($= n$) mutants are observed which could belong to 30 ($= K$) loci numbered from 1 through 30. Suppose that there are 3 loci each containing 4 mutants (i.e., $n_1 = n_2 = n_3 = 4$), two loci containing 2 mutants each ($n_4 = n_5 = 2$), and four loci containing one mutant each ($n_6 = n_7 = n_8 = n_9 = 1$). The remaining loci (10 through 30) do not contain any mutants (i.e., $n_{10} = \cdots = n_{30} = 0$). In this example, the observed number of loci containing at least one mutant (X) is 9, and the number of "repeat" mutants is 11. Noting that $11 = 20 - 9$, (i.e., $R = n - X$, more generally), it is clear that the probability distribution of R is completely specified by that of X for any fixed sample size (n).

Second, in the general case of unequal π_i's, the expectation and variance

of X have been derived using the indicator variable approach (Chakraborty et al., 1988), which are simple functions of n, K, and the π_i's. I shall show in the following that both of these properties are compatible with the general probability distribution function derived by the combinatorial approach used here.

COMBINATORIAL SOLUTION OF THE PROBABILITY FUNCTION OF X

Instead of working with the variable X, the observed number of non-empty classes, it is easier to work with its complement, $Y = K - X$, which is equivalent to the number of classes not represented in a sample of size n. Define events $A_1, A_2, \ldots, A_K$, such that A_i is the collection of all partitions of n into K segments, such that $n_i = 0$ (but $\sum n_i = n$). In other words, A_i represents the collection of sample points where the ith class ($i = 1, 2, \ldots, K$) remains empty in a specific sample. Note that A_i's are not exclusive of each other; i.e., there can be sample configurations where more than one n_i can be simultaneously zero.

For sampling with replacement, the following equations hold:

$$P_i = \text{Prob}(A_i) = (1 - \pi_i)^n, \qquad [12.3a]$$

$$P_{ij} = \text{Prob}(A_i A_j) = (1 - \pi_i - \pi_j)^n, \qquad [12.3b]$$

$$P_{ijk} = \text{Prob}(A_i A_j A_k) = (1 - \pi_i - \pi_j - \pi_k)^n, \qquad [12.3c]$$

etc., for all $i \neq j \neq k = 1, 2, \ldots, K$.

Following Feller (1968, p 99), define a sequence of summations $\{T_1, T_2, \ldots, T_K\}$ where

$$T_1 = \sum_i p_i,\ T_2 = \sum_i \sum_j p_{ij},\ T_3 = \sum_i \sum_j \sum_k p_{ijk}, \text{ etc.}$$

where the summations are taken such that $i < j < k < \cdots \leqslant K$, so that each combination appears once and only once; hence, the summation $T_r (1 \leqslant r \leqslant K)$ contains ${}^K C_r$ terms. The last term T_K reduces to only one term,

$$T_K = \text{Prob}(A_1 A_2 \cdots A_K) = P_{123\cdots K},$$

which is the probability of simultaneous occurrences of all K events A_1 through A_K. Invoking condition [12.1] on equations [12.3a–c], we note that $T_K = 0$, and furthermore,

$$T_{K-1} = \sum_{i=1}^{K} \pi_i^n, \qquad [12.4a]$$

$$T_{K-2} = \sum_{i>j=1}^{K} \sum^{K} (\pi_i + \pi_j)^n, \qquad [12.4b]$$

$$T_{K-3} = \sum_{i>j>k=1}^{K} \sum^{K} \sum^{K} (\pi_i + \pi_j + \pi_k)^n, \qquad [12.4c]$$

etc.

Applying Feller's theorem (Feller, 1968, p 106), we obtain

$$P_{[K-m]} = \text{Prob}(X = K - m) = \text{Prob}(Y = m)$$
$$= \sum_{i=m}^{K} (-1)^{i-m} \binom{i}{m} T_i, \qquad [12.5]$$

for $K - \min(K, n) \leqslant m \leqslant K - 1$, giving the sampling distribution of X, the number of non-empty classes in a sample of size n. Note that in [12.5], T_0 is conventionally defined as unity (see also Feller, 1968).

EXPECTATION AND VARIANCE OF THE NUMBER OF NON-EMPTY CLASSES IN A SAMPLE

As mentioned before, the expectation and variance of X can be derived by an indicator variable approach. For this we define K indicator variables, $Y_1, Y_2, \ldots, Y_K$, so that

$$Y_i = \begin{cases} 1, \text{if the } i\text{th class is unobserved in the sample,} \\ 0, \text{otherwise.} \end{cases}$$

Note that $X = K - Y$, and $Y = \sum_{i=1}^{K} Y_i$.

Therefore, the expectation of X is given by

$$E(X) = K - E(Y) = K - \sum_{i=1}^{K} E(Y_i)$$
$$= K - \sum_{i=1}^{K} \text{Prob}(Y_i = 1)$$
$$= K - \sum_{i=1}^{K} (1 - \pi_i)^n = K - T_1, \qquad [12.6]$$

derived by Emigh (1983) and Chakraborty et al. (1988).

Furthermore, the variance of X is given by

$$V(X) = V(Y) = \sum_{i=1}^{K} V(Y_i) + \sum_{i \neq j=i}^{K} \sum^{K} \text{Cov}(Y_i, Y_j). \qquad [12.7]$$

Since Y_i's are Bernoulli variables,

$$V(Y_i) = E(Y_i^2) - [E(Y_i)]^2 = (1 - \pi_i)^n [1 - (1 - \pi_i)^n], \qquad [12.8]$$

and

$$\text{Cov}(Y_i, Y_j) = E(Y_i Y_j) - E(Y_i) \cdot E(Y_j)$$
$$= (1 - \pi_i - \pi_j)^n - (1 - \pi_i)^n (1 - \pi_j)^n. \qquad [12.9]$$

Substituting equations [12.8] and [12.9] into equation [12.7], we obtain

$$V(X) = \sum_{i=1}^{K} (1-\pi_i)^n [1-(1-\pi_i)^n] + \sum_{i\neq j=1}^{K}\sum^{K} [(1-\pi_i-\pi_j)^n - (1-\pi_i)^n(1-\pi_j)^n]$$

$$= \sum_{i=1}^{K} (1-\pi_i)^n \cdot [1-\sum(1-\pi_i)^n] + \sum_{i\neq j=1}^{K}\sum^{K} [(1-\pi_i-\pi_j)^n \qquad [12.10]$$

$$= T_1(1-T_1) + T_2.$$

Note that when n is large, and each π_i small, we may approximate each term of the summations T_r's [12.3a–c] by $(1-\pi_i)^n \approx e^{-n\pi_i}$, $(1-\pi_i-\pi_j)^n \approx e^{-n(\pi_i+\pi_j)}$, etc., so that the variance of X can be approximated by

$$V(X) \approx \sum_{i=1}^{K} e^{-n\pi_i}(1-e^{-n\pi_i}), \qquad [12.11]$$

as shown in Chakraborty et al. (1988). Equation [12.10] is, however, exact, and not difficult to compute numerically even if the number of classes is large.

Let us now establish that equations [12.6] and [12.10] are compatible with the probability function [12.5]. To do this, first note that invoking equation [12.5] we have

$$\sum_m \text{Prob}(Y=m) = \sum_m \sum_{i=m}^{K} (-1)^{i-m}\binom{i}{m} T_i$$

$$= \sum_{i=0}^{K}\left[\sum_{m=0}^{i}(-1)^{i-m}\binom{i}{m}\right] T_i$$

$$= 1 + \sum_{i=1}^{K}\left[\sum_{m=0}^{i}(-1)^{i-m}\binom{i}{m}\right] T_i,$$

since $T_0 = 1$. Furthermore, the summation within the parenthesis is the binomial expansion of $(1-1)^i$, for $i = 1, 2, \ldots, K$. Therefore, we establish that [12.5] is a proper probability function since the entire probability mass equals unity.

The expectation of Y, from [12.5], then becomes

$$E(Y) = \sum_m m \cdot \text{Prob}(Y=m)$$

$$= \sum_m \sum_{i=m}^{K} (-1)^{i-m} m \binom{i}{m} T_i$$

$$= \sum_m \left[\sum_{i=m}^{K} (-1)^{i-m}\binom{i-1}{m-1} i T_i\right]$$

$$= T_1 + \sum_{i=2}^{K}\left[\sum_{m-1=0}^{i-1}(-1)^{i-m}\binom{i-1}{m-1}\right] i T_i \qquad [12.12]$$

$$= T_1,$$

since the summation within the parenthesis vanishes for all $i = 2, 3, \ldots, K$. Similar algebraic manipulations show that the second moment of Y can be written as

$$E(Y^2) = E(Y) + \sum_m m(m-1) \cdot \text{Prob}(Y = m)$$

$$= T_1 + 2T_2. \qquad [12.13]$$

Since $X = K - Y$, I therefore complete the proof of equations [12.6] and [12.10] starting from [12.5]. Furthermore, this computational logic also yields the rth factorial moment of Y, $\mu_{[r]}(Y)$, given by

$$\mu_{[r]}(Y) = E[Y(Y-1)\cdots(Y-r+1)] = r!\, T_r, \qquad [12.14]$$

for any $r \geq 1$, giving the complete characterization of the probability function [12.5] through its moments.

When all π_i's are equal (i.e., $\pi_i = 1/K$ for all i), note that

$$T_1 = K[(K-1)/K]^n, \text{ and } T_2 = K(K-1)\cdot[(K-2)/K]^n/2,$$

and hence,

$$E(X) = K[1 - \{(K-1)/K\}^n], \qquad [12.15]$$

and

$$V(X) = K\cdot\{(K-1)/K\}^n\cdot[1 - K\cdot\{(K-1)/K\}^n] + K(K-1)\cdot\{(K-2)/K\}^n, \qquad [12.16]$$

which are derived in Arnold and Beaver (1988) while studying the sampling properties of the observed number of classes in the context of the classical occupancy problem. When $\pi_i = 1/K$, note also that the summations $\{T_i\}$ take the form

$$T_i = \binom{K}{i}\{(K-i)/K\}^n,$$

so that the probability function [12.5] reduces to

$$P_{[K-m]} = \sum_{i=m}^{K} (-1)^{i-m}\binom{i}{m}\binom{K}{i}(K-i)^n/K^n$$

$$= \sum_{i=m}^{K} (-1)^{i-m}\binom{K}{K-m}\binom{K-m}{i-m}(K-i)^n/K^n$$

$$= \binom{K}{K-m}\cdot\sum_{i=m}^{K} (-1)^{i-m}\binom{K-m}{i-m}(K-i)^n/K^n$$

$$= \binom{K}{K-m}\cdot\frac{m!\, S_n^{(K-m)}}{K^n},$$

invoking the definition of a Stirling number of the second kind (Abramowitz and Stegun, 1965; p 824).

The above derivations, therefore, show that the sampling distribution of the number of observed classes in a finite sample can be analytically specified for any arbitrary multinomial distribution. This generalizes the special case solution of the problem discussed in Arnold and Beaver (1988) in the context of the classical occupancy problem. The algebraic solutions of other relevant random variables (e.g., the number of classes containing a specified number of sampling units within each of them) are also similar, although more cumbersome to compute numerically.

APPLICATIONS

I mention here three applications of this generalized occupancy problem, each of which has considerable genetic implications.

Are mutagen-induced mutations equally likely to occur at all loci?

Hanash et al. (1988) recently demonstrated that somatic cell gene mutations altering protein structure do not occur with equal probability at all loci when cultured human lymphoblastoid cell lines are treated with mutagens like ethylnitrosourea. To show this, they used the technique of two-dimensional polyacrylamide gel electrophoresis, and found 65 mutants occurring at 49 of the 263 loci scored in their experiments. The locus-specific distributions of the mutation frequencies in their work were: three mutants observed at each of five loci ($n_1 = n_2 = n_3 = n_4 = n_5 = 3$), two mutants at each of six loci ($n_6 = \cdots = n_{11} = 2$), and one mutant at each of 38 loci ($n_{12} = \cdots = n_{49} = 1$). No mutation was detected at each of the remaining 214 loci ($n_{50} = \cdots = n_{263} = 0$). The total number of mutations ($n = 65$) was, thus, distributed in $K = 263$ classes. The null hypothesis to be tested in H_0: $\{\pi_i =$ the probability of mutation occurring at the ith locus $= 1/K = 1/263$, for all $i\}$. In their work, the authors defined the concept of "repeat" mutations (R), noting that 16 mutations occurred at loci each of which contained already one mutation (i.e., $R = n - X$, where X is the number of loci containing at least one mutant). Under the null hypothesis of equiprobable mutation frequencies across loci, the number of "repeat" mutations should be small, since $K = 263$ is much larger than the sample size $n = 65$. Through a simulation experiment of the occupancy problem, they determined that the probability of 16 or more "repeat" mutants is below 0.0005, and hence, they conclude that mutagen-induced mutations are not equally likely to occur at all loci.

The theory described above provides a complete analytical solution, avoiding any simulation. Note that the observed value of X in the above experiment is $m = 49$, and hence the observed number of empty classes (number of loci having no occurrences of mutations, Y) is 214 ($= 263 - 49$). With $n = 65$, and $K = 263$, the range of possible X values in this experiment

is 1 through 65, and consequently the possible values of Y (number of empty classes) are from 198 ($= 263 - 65$) through 262. Figure 12.1 shows the exact probability distribution (shown by the histogram) of non-empty classes (X, the number of loci at which one or more mutants can occur) under the null hypothesis of equiprobable mutations across the 263 loci. The observed value of X ($m = 49$) is marked with an arrow, the area below which is the total probability ($P = 0.0005$) of all other sample configurations which represent deviations from the hypothesis more extreme under the null hypothesis. Note that Hanash et al.'s simulation also resulted in a P-value consistent with the present result, suggesting that the qualitative conclusion of their analysis is the same as the one obtained by the present analytical solution.

Since under the null hypothesis ($\pi_i = 1/K = 1/263$ for all i) the observed number of non-empty classes (X) is the sufficient statistic for K, I further checked to see if the exact distribution of X can be approximated by any standard distribution. Employing equations [12.15] and [12.16], the mean and variance of X for this experiment are given by $E(X) = 57.687$ and $V(X) = 5.289$. The smooth curve of Figure 12.1 represents the density function

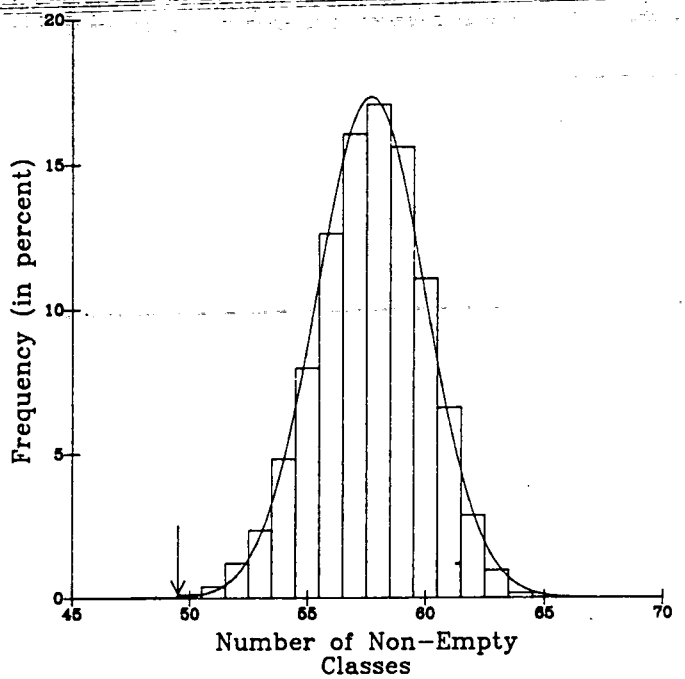


Figure 12.1. The sampling distribution of the number of non-empty classes in a sample of 65 observations drawn from a multinomial distribution with 263 equi-frequent classes. The histogram shows the exact analytical distribution evaluated from equation [12.5] and the smooth curve is the normal approximation using the expectation and variance given in equations [12.6] and [12.10]. The arrow indicates the observed value $m = 49$ (See text for details).

of a normal distribution with these mean and variance values. Note that the normal deviate for $m = 49$, then becomes $z = -3.78$, giving a corresponding P value of 0.00002, which is somewhat smaller than the p value obtained from the exact distribution. Nevertheless, the normal approximation is quite satisfactory, when compared with the histogram shown in Figure 12.1.

Test of Hardy-Weinberg expectation based on the observed numbers of distinct genotyeps in a finite sample

The generalized occupancy problem can also be used to examine whether or not the genotype distribution of a given number of alleles follows the Hardy-Weinberg expectation (HWE). Generally, this is done by either a likelihood ratio test or a goodness of fit chi-square test, contrasting the observed and expected frequencies of all possible genotypes. However, there are occasions when the number of alleles are so large that many of the genotypes are either not observed in a sample, or the observed frequencies of several genotypes are so small that the large sample approximation of these test statistics in unwarranted. The recently discovered VNTR polymorphisms provide examples of this nature, where the number of possible alleles is often so large that no reasonably sized survey can encompass all possible genotypes in any given sample. Assuming that there are K segregating alleles at a locus, there are K possible homozygote genotypes and $K(K-1)/2$ possible distinct heterozygote genotypes that can be encountered. One might ask, what would be the distribution of the numbers of distinct genotypes (of homozygote and heterozygote types, separately) observed in a sample of n individuals. Under the Hardy-Weinberg expectation of genotypic probabilities given by p_i^2 for homozygotes and $2p_ip_j$ for heterozygotes, where p_i represents the allele frequencies in the population, we can use the above analytical formulation to compute the exact distributions of the distinct numbers of homozygote and heterozygote genotypes seen in a sample.

Figure 12.2 shows a numerical example of such computations. Deka et al. (1991) recently surveyed the New Guinea population for VNTR polymorphisms at six loci. At the D1S76 locus, they discovered 6 alleles in a sample of 35 individuals. Gene counting showed that in the sample of 70 genes at this locus, the allele counts of these 6 alleles are 1, 3, 7, 9, 25, and 25. In total they observed 20 heterozygous individuals (consisting of 7 distinct genotypes). However, under the HWE assumption, the expected frequency of heterozygotes from the above allele counts is 25.4, showing a significant deficiency of heterozygotes ($P < 0.05$). Since the observed numbers of distinct homozygote and heterozygote genotypes in their sample were 4 and 7, respectively, we can ask if these observations deviate from their respective expectations under the HWE assumption. Figure 12.2a shows the exact distribution of the observed number of distinct homozygote genotypes (drawn as histogram) and Figure 12.2b gives the same for the observed number of distinct heterozygote genotypes, under the HWE assumption for the given allele frequencies. The arrows represent the observed statistics. Clearly, the

observed number of distinct homozygote genotypes ($m = 4$) is not at variance with the HWE, since the probability of observing four or more distinct homozygote genotypes is 0.411. Under HWE the probability of observing seven or less distinct heterozygote genotypes is 0.017, suggesting that a significant deficiency is observed in the total number of heterozygotes as well as in the number of distinct heterozygote genotypes. Of course, as in the case of traditional likelihood ratio or chi-square tests, this test cannot ascertain the real cause of such heterozygote deficiency.

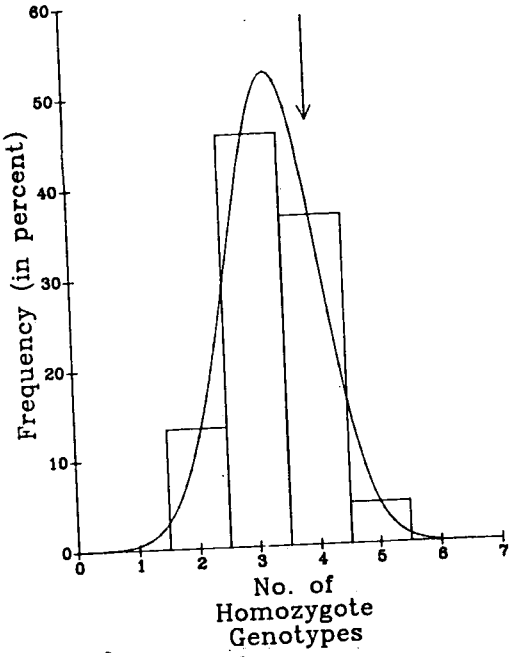
From equations [12.6] and [12.10], the mean and variance of the number of distinct genotypes were computed as 3.329 and 0.578 for the homozygotes, and 10.106 and 1.652 for the heterozygote genotypes, respectively. The expected distributions under the normality approximation are also shown by the smooth curves in both panels of Figure 12.2. As in the earlier case, it shows that the normal approximation is fairly adequate for the distribution of distinct heterozygote genotypes. This is not so for the homozygotes because of the narrow range of variation in the number of distinct homozygote genotypes. Under the normality approximation, the normal deviate corresponding to observing seven or less distinct heterozygote genotypes is $z = -2.41$, with a P-value of 0.008, which is again smaller than the exact P-value shown above.

Global test of disequilibrium based on multiple-locus haplotype data

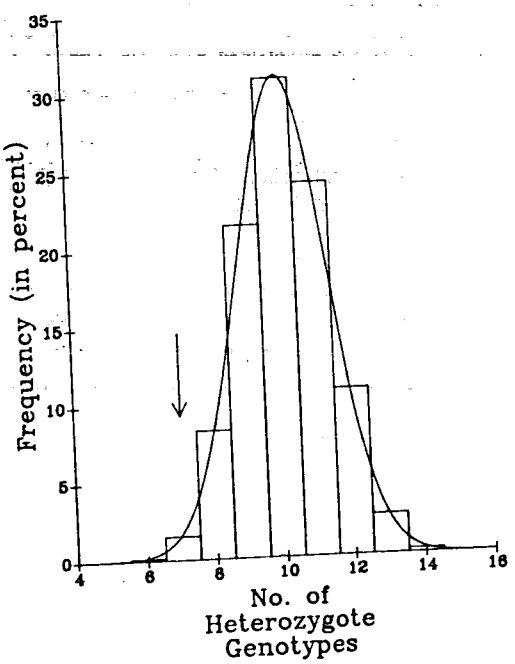
As a third application, consider the haplotype frequency data surveyed by Wainscoat et al. (1986) at the β-globin gene cluster detected by five polymorphic restriction sites, at each of which there are two segregating alleles. This results in $2^5 = 32$ possible haplotypes at this gene region, but in a sample of 55 chromosomes sampled from a Polynesian population, these authors found only 5 observed haplotypes (see Table 1 of Wainscoat et al., 1986). One might ask, what is the expected distribution of the number of haplotypes given that these five sites are independently segregating. Figure 12.3 shows the exact distribution (represented by histogram), following the general analytical formulation (equation 12.5), where the expected haplotype frequencies are assumed to follow the independent segregation rule. Clearly, almost the entire distribution is to the right of the observed number ($m = 5$) of haplotypes, giving a rare probability of observing five or less haplotypes ($P < 10^{-5}$), suggesting that the observed number of haplotypes is incompatible with the assumption of independent segregation. Under independent

Figure 12.2. The sampling distributions of the number of distinct homozygote (A) and heterozygote (B) genotypes at the D1S76 VNTR locus in a sample of 35 individuals from Papua New Guinea (Deka et al., 1990) under the assumption of Hardy-Weinberg equilibrium frequencies of genotypic proportions. The histograms are exact computations (equation 12.5) and the smooth curves are the normal approximations based on mean and variance, given in equations [12.6] and [12.10]. The arrows indicate the observed numbers of distinct homozygote (4) and heterozygote genotypes (7) found in the sample.

A


Frequency (in percent)
60
50
40
30
20
10
0
0
1
2
3
4
5
6
7
No. of
Homozygote
Genotypes

B


Frequency (in percent)
35
30
25
20
15
10
5
0
4
6
8
10
12
14
16
No. of
Heterozygote
Genotypes

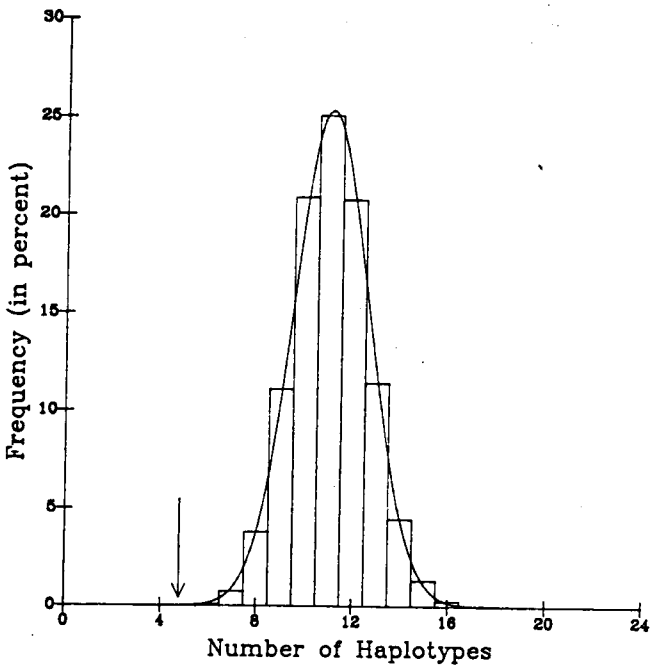


Figure 12.3. The sampling distribution of the number of DNA haplotypes at the β-globin gene cluster, defined by 5 restriction site polymorphisms (Wainscoat et al., 1986), in a sample of 55 chromosomes from a Polynesian population under the assumption of complete linkage equilibrium. The histogram is the exact distribution based on equation [12.5] and the smooth curve is its normal approximation based on mean and variance given by equations [12.6] and [12.10]. The arrow indicates the observed number of 5 different haplotypes found in the sample.

segregation, the expected mean and variance of the observed number of haplotypes in a sample of 55 chromosomes are 11.059 and 2.477, respectively. The normal approximation of the sampling distribution is again shown by the smooth curve of Figure 12.3. While the normal approximation appears satisfactory, the normal deviate corresponding to the observed number 5 is $z = -3.85$, giving a P-value of 0.00006, which is larger than the exact P-value. Note that Blanton and Chakravarti (1987) suggested this global test for disequilibrium, although unlike here their sampling distribution was obtained by simulation.

DISCUSSION

The analytical theory presented here along with the specific applications indicate that the generalized occupancy problem has a number of interesting applications in population genetics. This is particularly true in the context of sparse data, where by the very nature of the problem, the exact sampling

distribution must be evaluated and no adequate large sample approximation is available. This theory enables comparison of occurrences of several biological endpoints in cross-survey comparisons, adjusting for sample size differences, as shown in Chakraborty et al. (1988), and in this chapter three other applications are mentioned. In the first application, no loss of information is attendent to the consideration of the sample statistic X, the observed number of non-empty classes, since under the equiprobable mutation rate (across loci), X is a sufficient statistic of the only parameter of the underlying distribution. In the other two cases, the consideration of observed number of classes raise the possibility of some loss of information, since the frequencies of the different observed categories do not enter into the present analysis. However, when the number of categories is large compared to the sample size, most of the observed categories are likely to have one or a few sample points in them and such loss of information is not critical. As shown through the applications here, the exact distribution evaluation does indeed detect deviations from the null hypothesis even when the sample size is larger than the total number of possible classes. A comprehensive power analysis of this approach to deal with such specific genetic applications will be attempted in the future.

In closing I should note the close resemblance of the methodology of this presentation with a percentage testing problem that Schull and I resolved several years ago (Chakraborty and Schull, 1976), where we evaluated the sampling distribution of the numbr of loci with reference to which a randomly accused man could be excluded if this man is not the father of a child born to a specific mother. Although all of these problems can be resolved by simulation, the real advantage of the present theory is that such problems can be addressed analytically, avoiding the natural bias and tediousness of computer simulations.

ACKNOWLEDGMENTS

This chapter is dedicated to Professor William J. Schull, whose encouragement and motivation primarily led to this work. This research is partially supported by Grants GM 41399 and 90-IJ-CX-0038 from the National Institutes of Health and National Institute of Justices, respectively.

REFERENCES

Abramowitz M, Stegun IA (1965) *Handbook of Mathematical Functions.* New York, Dover.

Arnold BC, Beaver RJ (1988) Estimation of the number of classes in a population. *Biometrical Journal* 30:413–424.

Blanton SH, Chakravarti A (1987) A global test of linkage disequilibrium. *Amer J Hum Genet* 41:A250.

Chakraborty R, Schull WJ (1976) A note on the distribution of the number of exclusions to be expected in paternity testing. *Amer J Hum Genet* 28:615–618.

Chakraborty R, Smouse PE, Neel JV (1988) Population amalgamation and genetic variation: observations on artificially agglomerated tribal populations of Central and South America. *Amer J Hum Genet* 43:709–725.

Deka R, Chakraborty R, Ferrell RE (1991) A population genetic study of six VNTR loci in three ethnically defined populations. *Genomics* 11:83–92.

Emigh TH (1983) On the number of observed classes from a multinomial distribution. *Biometrics* 39:485–491.

Feller W (1968) *An Introduction to Probability Theory and its Applications* New York, Wiley.

Hanash SM, Boehnke M, Chu EHY, Neel JV, Kuick RD (1988) Nonrandom distribution of structural mutants in ethylnitrosourea treatment of cultured human lymphoblastoid cells. *Proc Natl Acad Sci USA* 85:165–169.

Schull WJ, Neel JV (1965) *The Effects of Inbreeding on Japanese Children*. New York, Harper and Row.

Schull WJ, Otake M, Neel JV (1981) Genetic effects of the atomic bombs: a reappraisal. *Science* 213:1220–1227.

Schull WJ, Rothhammer F (1990) *The Aymara: Strategies in Human Adaptation to a Rigorous Environment*. Amsterdam, Kluwer Academic Publishers.

Wainscoat JS, Hill AVS, Boyce AL, Flint J, Hernandez M, Thein SL, Old JM, Lynch JR, Falusi AG, Weatherall DJ, Clegg JB (1986) Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. *Nautre* 319:491–493.

Simposios

ORIGEN DEL HUMANO

Coordinador: Dr. Cristián Orrego

GEOGRAFIA GENICA DE SUDAMERICA : CONTRASTANDO MODELOS DE DESPLAZAMIENTO POBLACIONAL PRECOLOMBINOS. (Gene geography of South America : testing models of pre-Columbian population displacements). **F. Rothhammer, C. Silva y E.Llop. Departamento Biología Celular y Genética, Facultad de Medicina, Universidad de Chile y Departamento de Matemática y Computación, Facultad de Ciencias, Universidad de Santiago.**

Aprovechando avances recientes de la graficación computarizada se han presentado los resultados del tratamiento estadistico multivariado de frecuencias génicas para determinadas áreas geográficas en forma de mapas sintéticos de variación genética. Estos mapas han sido utilizados para poner a prueba hipótesis sobre la difusión de la agricultura en Europa como también sobre la acción de factores evolutivos, tanto a nivel mundial como también para Norte, Centro y Sudamerica. Los resultados obtenidos para America en general son poco concluyentes debido al pequeño número de sistemas genéticos incluidos en el análisis. En esta ocasión presentaremos mapas sinteticos de frecuencias génicas para Sudamerica utilizando el mayor número posible de sistemas. Posteriormente utilizaremos estas representaciones gráficas para poner a prueba modelos de desplazamiento poblacional precolombinos.

GRANT N° 91-1110 FONDECYT.

HOMINIZACION UNA PERSPECTIVA BIOANTROPOLOGICA Eugenio Aspillaga, Departamento de Antropología y Departamento de Anatomía Normal, Universidad de Chile. Para una mayor comprensión del proceso de Hominización es necesario remontarse al origen de los primates como grupo, hace como unos 70 millones de años; y discutir en torno a las adaptaciones y caracteristicas de dicho grupo, que constribuyeron a la génesis de la propiedades biologicas más notorio del hombre y que en su origen constituyeron un conjunto de propiedades necesarias para la aparición de éste y del fenómeno adaptativo extra somático que llamamos cultura. Es posible que las restrictivas condiciones del medio arboreo, propiciará la seleccion de variabilidad biologica consistente con una vida exitosa en dicho medio; es asi como aspectos tan caracteristicos del Hombre como son su capacidad manipulacion, la vision estereoscópica combinada con una desarrollada capacidad de enfoque y visión de colores, el desarrollo de áreas asociativas en el cerebro, asi como un cerebelo más complejo y otras caracteristicas relacionadas, pueden comprenderse mejor si se analiza los presuntos requerimientos vitales de los primates primitivos en su relación con la vida arbórea.

Alistralophedus afarensis, probablemente fue el primer nominidos donde todas las propiedades biológicas necesarias para la aparición del Hombre y la cultura estaban ya esbozadas, hace unos 3,5 a 4 millones de años. Dichas propiedades se asentaran una vez "gatillado" el fenómeno de la cultura, probablemente por el **homo habilis**, hace unos 2,8 millones de años. La cultura acelerará el proceso enormemente, contribuyendo a hacer exponencial el incremento de la capacidad craneana hasta alcanzar en un plazo muy breve, en términos macroevolutivos, los limites de nuestra propia especie. Esta última en su variedad **Homo sapiens sapiens** se encargará de ocupar casi todos los espacios de nuestro planeta, incluido nuestro continente, América, el cual comenzó a conquistar hace unos 40.000 años y donde han tenido lugar interesantes fenómenos de tipo microevolutivo.

IMPACT OF MOLECULAR GENETICS IN STUDYING ORIGIN OF HUMAN POPULATIONS.

Ranajit Chakraborty. Center for Demographic and Population Genetics. The University of Texas Graduate School of Biomedical Sciences, Houston, Texas.

The origin of specific human populations has always been an intriguing question to biological anthopologists. While early works on this subject reliad on typological patterns of human variation, recent advances in molecular genetics make such studies far more incisive; the level at which genetic variation can now be studied is much closer to the underlying molecular typing can now be done on people who existed several thousand years back, dating the origins of populations can now be done with a precision that was not feasible from the traditional biological traits. In this presentation, presentation, preliminary data on molecular variation in humans will be used to show that population genetic theories are available to utilize biochemical variation detected at a molecular level, from which precise genetic profiles of populations can be determined. Anthropologists, human biologists, as well as forensic scientist can profitably use the concept of genetic variations among individuals, allowing reconstruction of past events of colonization and expansion of populations, from which origins of specific populations can be predicted. (Research supported by NIH grant GM-41399

R98

ISSN 0004-0533

ARCHIVOS DE BIOLOGIA Y MEDICINA EXPERIMENTALES

VOL. 24 NOVIEMBRE 1991 Nº 2

34a. REUNION ANUAL

SOCIEDAD DE BIOLOGIA DE CHILE
Y SOCIEDADES AFILIADAS

Sociedad Chilena de Biología de la Reproducción y Desarrollo
Sociedad de Biología Celular de Chile
Sociedad de Bioquímica de Chile
Sociedad de Botánica de Chile
Sociedad Chilena de Ciencias Fisiológicas
Sociedad de Farmacología de Chile
Sociedad de Genética de Chile

RESUMENES DE
CONFERENCIAS, SIMPOSIOS Y COMUNICACIONES

27 - 30 de noviembre de 1991
Puyehue, Chile

Sociedad de Biología de Chile

Simposios

ORIGEN DEL HUMANO

Coordinador: Dr. Cristián Orrego

GEOGRAFIA GENICA DE SUDAMERICA : CONTRASTANDO MODELOS DE DESPLAZAMIENTO POBLACIONAL PRECOLOMBINOS. (Gene geography of South America : testing models of pre-Columbian population displacements). **F. Rothhammer, C. Silva y E.Llop. Departamento Biología Celular y Genética, Facultad de Medicina, Universidad de Chile y Departamento de Matemática y Computación, Facultad de Ciencias, Universidad de Santiago.**

Aprovechando avances recientes de la graficación computarizada se han presentado los resultados del tratamiento estadistico multivariado de frecuencias génicas para determinadas áreas geográficas en forma de mapas sintéticos de variación genética. Estos mapas han sido utilizados para poner a prueba hipótesis sobre la difusión de la agricultura en Europa como también sobre la acción de factores evolutivos, tanto a nivel mundial como también para Norte, Centro y Sudamerica. Los resultados obtenidos para America en general son poco concluyentes debido al pequeño número de sistemas genéticos incluidos en el análisis. En esta ocasión presentaremos mapas sinteticos de frecuencias génicas para Sudamerica utilizando el mayor número posible de sistemas. Posteriormente utilizaremos estas representaciones gráficas para poner a prueba modelos de desplazamiento poblacional precolombinos.

GRANT N° 91-1110 FONDECYT.

HOMINIZACION UNA PERSPECTIVA BIOANTROPOLOGICA Eugenio Aspillaga, Departamento de Antropologia y Departamento de Anatomia Normal, Universidad de Chile. Para una mayor comprensión del proceso de Hominización es necesario remontarse al origen de los primates como grupo, hace como unos 70 millones de años; y discutir en torno a las adaptaciones y caracteristicas de dicho grupo, que constribuyeron a la génesis de la propiedades biológicas más notorio del hombre y que en su origen constituyeron un conjunto de propiedades necesarias para la aparición de este y del fenomeno adaptativo extra somático que llamamos cultura. Es posible que las restrictivas condiciones del medio arboreo, propiciará la seleccion de variabilidad biologica consistente con una vida exitosa en dicho medio; es asi como aspectos tan caracteristicos del Hombre como son su capacidad manipulacion, la visión estereoscópica combinada con una desarrollada capacidad de enfoque y visión de colores, el desarrollo de áreas asociativas en el cerebro, asi como un cerebelo más complejo y otras caracteristicas relacionadas, pueden comprenderse mejor si se analiza los presuntos requerimientos vitales de los primates primitivos en su relación con la vida arbórea.

Alistralophedus afarensis, probablemente fue el primer nominidos donde todas las propiedades biológicas necesarias para la aparición del Hombre y la cultura estaban ya esbozadas, hace unos 3,5 a 4 millones de años. Dichas propiedades se asentaran una vez "gatillado" el fenómeno de la cultura, probablemente por el homo habilis, hace unos 2,8 millones de años. La cultura acelerará el proceso enormemente, contribuyendo a hacer exponencial el incremento de la capacidad craneana hasta alcanzar en un plazo muy breve, en términos macroevolutivos, los limites de nuestra propia especie. Esta última en su variedad Homo sapiens sapiens se encargará de ocupar casi todos los espacios de nuestro planeta, incluido nuestro continente, América, el cual comenzó a conquistar hace unos 40.000 años y donde han tenido lugar interesantes fenómenos de tipo microevolutivo.

IMPACT OF MOLECULAR GENETICS IN STUDYING ORIGIN OF HUMAN POPULATIONS.

Ranajit Chakraborty. Center for Demographic and Population Genetics. The University of Texas Graduate School of Biomedical Sciences, Houston, Texas.

The origin of specific human populations has always been an intriguing question to biological anthopologists. While early works on this subject reliad on typological patterns of human variation, recent advances in molecular genetics make such studies far more incisive; the level at which genetic variation can now be studied is much closer to the underlying molecular typing can now be done on people who existed several thousand years back, dating the origins of populations can now be done with a precision that was not feasible from the traditional biological traits. In this presentation, presentation, preliminary data on molecular variation in humans will be used to show that population genetic theories are available to utilize biochemical variation detected at a molecular level, from which precise genetic profiles of populations can be determined. Anthropologists, human biologists, as well as forensic scientist can profitably use the concept of genetic variations among individuals, allowing reconstruction of past events of colonization and expansion of populations, from which origins of specific populations can be predicted. (Research supported by NIH grant GM-41399

R98

ISSN 0004-0533

ARCHIVOS DE BIOLOGIA Y MEDICINA EXPERIMENTALES

VOL. 24 NOVIEMBRE 1991 Nº 2

34a. REUNION ANUAL

SOCIEDAD DE BIOLOGIA DE CHILE Y SOCIEDADES AFILIADAS

Sociedad Chilena de Biología de la Reproducción y Desarrollo
Sociedad de Biología Celular de Chile
Sociedad de Bioquímica de Chile
Sociedad de Botánica de Chile
Sociedad Chilena de Ciencias Fisiológicas
Sociedad de Farmacología de Chile
Sociedad de Genética de Chile

RESUMENES DE CONFERENCIAS, SIMPOSIOS Y COMUNICACIONES

27 - 30 de noviembre de 1991
Puyehue, Chile

Sociedad de Biología de Chile

le Fonds de la Recherche en Santé du Québec, Medical Research Council of Canada, and the Network of Centres of Excellent (Genetics).

References

Beaudet AL (1990) Invited editorial: carrier screening for cystic fibrosis. Am J Hum Genet 47:603–605

Boat TJ, Welsh M, Beaudet AL (1989) Cystic fibrosis. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic basis of inherited disease, 6th ed. McGraw-Hill, New York, pp 2649–2682

Caskey CT, Kaback MM, Beaudet AL (1990) The American Society of Human Genetics statement on cystic fibrosis screening. Am J Hum Genet 46:393

Clow CL, Scriver CR (1977) Knowledge about and attitudes toward genetic screening among high-school students: the Tay-Sachs experience. Pediatrics 59:86–91

Cystic Fibrosis Genetic Analysis Consortium (1990) Worldwide survey of the ΔF508 mutation—report from the Cystic Fibrosis Genetic Analysis Consortium. Am J Hum Genet 47:354–359

European Working Group on CF Genetics (1990) Gradient of distribution in Europe of the major CF mutation and its associated haplotype. Human Genetics 85:436–441

Kerem B-S, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, et al (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Scriver CR, Bardanis M, Cartier L, Clow CL, Lancaster GA, Ostrowsky JT (1984) β-Thalassemia disease prevention: genetic medicine applied. Am J Hum Genet 36:1024–1038

Triggs-Raine BL, Gravel RA (1990) Diagnostic heteroduplexes: simple detection of carriers of a 4-bp insertion mutation in Tay-Sachs disease. Am J Hum Genet 46:183–184

Wilfond BS, Fost N (1990) The cystic fibrosis gene: medical and social implications for heterozygote detection. JAMA 263:2777–2783

0002-9297/91/4901-0027$02.00

Am. J. Hum. Genet. 49:242–243, 1991

Inclusion of Data on Relatives for Estimation of Allele Frequencies

To the Editor:

In a recent issue of the *Journal,* Boehnke (1991) suggests a general method for estimating allele (or haplotype) frequencies from data on relatives. He provides the maximum likelihood estimates of allele frequencies for any arbitrary structure of pedigree relationships among relatives. It is a clever application of the maximum likelihood method originally designed for pedigree analysis. Nevertheless, in support of this method a few remarks may be added that might be particularly helpful to the users who are unfamiliar with the literature in this area of research.

First, it may be noted that the inclusion of relatives for estimating allele frequencies at a locus has a comparatively long history in human genetics. Fisher (1940) examined the effect that inclusion of relatives had on the estimate of the proportion of recessives in the population; for estimating allele frequencies and their precision Cotterman (1947) developed a weighting system from family data by using the maximum likelihood principle. Finney (1948*a*, 1948*b*) and Smith (1957) suggested alternative methods to address the same problem. Finally, Chakraborty (1978), in an appendix to the work of Ferrell et al. (1978), suggested a combinatorial approach, showing that, in addition to the estimation of allele frequencies, one can estimate the number of independent genes sampled in a survey that includes data on relatives. In principle, while Boehnke's (1991) method is based on similar logic, none of the above developments of this subject is referred to in his work.

Second, the versatility of the computer algorithms, such as MENDEL (Lange et al. 1988), yields allele frequency estimates at any locus even if the latter has a complex mode of inheritance, while all previous attempts deal with simple Mendelian transmission rules and specific family structures. However, the estimate of the equivalent number of alleles, given by Boehnke (1991), needs an extra cautionary remark. It should not be equated to the number of independent genes sampled, derived by Chakraborty (1978). In a set of family data the number of independent genes sampled is truly a random variable (let us denote it by N), whose expectation and variance can be analytically obtained from the relationships among individuals included in the analysis. Chakraborty (1978) showed that the distribution of N in family data is dependent on the family structure and size. For example, in a nuclear family with genotype information available on both parents and s children, N becomes 4, irrespective of s, and hence the offspring genotypes do not give any extra information when all alleles are codominant. In fact, for a codominant locus, the inclusion of offspring genotypes when both parental genotypes are known introduces errors of random fluctuation of Mendelian segregation ratios.

When genotypic data are available only on s (≥ 1)

sibs, N can take values 2, 3, and 4, with mean and variance given by

$$E(N) = 4(2^s - 1)/2^s, \quad (1)$$

and

$$V(N) = 16 - (28/2^s) + (8/4^s) - E^2(N), \quad (2)$$

so that the ratio of the expected number of independent genes sampled to the total number of alleles assayed, $E(N)/2s$, can be quite small, because $E(N) \to 4$ as $s \to \infty$. Chakraborty (1978) also considered more complex situations, such as the inclusion of individuals with one parent, $s - 1$ of his/her sibs, and k offspring tested. On the basis of such evaluations, it was shown that it might be necessary to attach different weights to genotype data on individuals belonging to different generations, to arrive at a statistically consistent estimator of allele frequencies. Although Boehnke's (1991) likelihood function (represented by his eq. [1]) accomplishes that, his n^* (equivalent number of alleles sampled) does not truly represent N. This is so because n^* depends on the estimated allele frequency as well as on the family structure and size, while N is independent of the allele frequency estimates. Therefore, for a given data structure, N will remain the same for all alleles, while n^* can vary substantially over alleles.

Furthermore, a more intricate problem relates to the reference population for which allele frequency estimates are being sought. Is it the entire collection of individuals in a given space at a given point of time, or does it relate to individuals of different generations recorded at a particular point of time? Population biologists interested in allele frequency estimates may approach the task of estimating allele frequencies by depending on the definition of the reference population used. Some might prefer to ignore the relationship structure altogether, and others may prefer to follow either the weighting schemes suggested earlier or Boehnke's suggested algorithm.

Boehnke's (1991) work is reassuring in the sense that, like his predecessors, he also concludes that ignoring the familial relationship does not introduce any systematic bias into allele frequency estimates; it only makes the allele frequency estimates appear more precise than they actually are. Therefore, inclusion of relatedness might be extremely important when one compares allele frequencies between samples that have large differences in their family structures, particularly when the alleles are rare, such as is the case with recently arisen mutations (e.g., see Neel et al. 1988).

RANAJIT CHAKRABORTY
Center for Demographic and Population Genetics
University of Texas Graduate School of Biomedical Sciences
Houston

References

Boehnke M (1991) Allele frequency estimation from data on relatives. Am J Hum Genet 48:22–25

Chakraborty R (1978) Number of independent genes examined in family surveys and its effect on gene frequency estimation. Am J Hum Genet 30:550–552

Cotterman CW (1947) A weighting system for the estimation of gene frequencies from family records. In: Contributions to the Laboratory of Vertebrate Biology, no. 33. Ann Arbor, University of Michigan, pp 1–21

Ferrell RE, Bertin T, Young R, Barton SA, Murillo F, Schull WJ (1978) The Aymara of western Bolivia. IV. Gene frequencies for eight blood groups and 19 protein and erythrocyte enzyme systems. Am J Hum Genet 30:539–549

Fisher RA (1940) The estimation of the proportion of recessives from tests carried on a sample not wholly unrelated. Ann Engenics 10:160–170

Finney DJ (1948*a*) The estimation of gene frequencies from family records. I. Factors without dominance. Heredity 2:199–218

——— (1948*b*) The estimation of gene frequencies from family records. II. Factors with dominance. Heredity 2: 369–390

Lange K, Weeks D, Boehnke M (1988) Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol 5:471–472

Neel JV, Satoh C, Smouse P, Asakawa J-I, Takahashi N, Goriki K, Fujita M, et al (1988) Protein variants in Hiroshima and Nagasaki: tales of two cities. Am J Hum Genet 43:870–893

Smith CAB (1957) Counting methods in genetical statistics. Ann Hum Genet 21:254–276

0002-9297/91/4901-0028$02.00

Am. J. Hum. Genet. 49:243–244, 1991

Reply to Chakraborty

To the Editor:

I thank Dr. Chakraborty for pointing out several references that describe other methods for allele frequency

le Fonds de la Recherche en Santé du Québec, Medical Research Council of Canada, and the Network of Centres of Excellent (Genetics).

References

Beaudet AL (1990) Invited editorial: carrier screening for cystic fibrosis. Am J Hum Genet 47:603–605

Boat TJ, Welsh M, Beaudet AL (1989) Cystic fibrosis. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic basis of inherited disease, 6th ed. McGraw-Hill, New York, pp 2649–2682

Caskey CT, Kaback MM, Beaudet AL (1990) The American Society of Human Genetics statement on cystic fibrosis screening. Am J Hum Genet 46:393

Clow CL, Scriver CR (1977) Knowledge about and attitudes toward genetic screening among high-school students: the Tay-Sachs experience. Pediatrics 59:86–91

Cystic Fibrosis Genetic Analysis Consortium (1990) Worldwide survey of the ΔF508 mutation—report from the Cystic Fibrosis Genetic Analysis Consortium. Am J Hum Genet 47:354–359

European Working Group on CF Genetics (1990) Gradient of distribution in Europe of the major CF mutation and its associated haplotype. Human Genetics 85:436–441

Kerem B-S, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, et al (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Scriver CR, Bardanis M, Cartier L, Clow CL, Lancaster GA, Ostrowsky JT (1984) β-Thalassemia disease prevention: genetic medicine applied. Am J Hum Genet 36:1024–1038

Triggs-Raine BL, Gravel RA (1990) Diagnostic heteroduplexes: simple detection of carriers of a 4-bp insertion mutation in Tay-Sachs disease. Am J Hum Genet 46:183–184

Wilfond BS, Fost N (1990) The cystic fibrosis gene: medical and social implications for heterozygote detection. JAMA 263:2777–2783

0002-9297/91/4901-0027$02.00

Am. J. Hum. Genet. 49:242–243, 1991

Inclusion of Data on Relatives for Estimation of Allele Frequencies

To the Editor:

In a recent issue of the *Journal,* Boehnke (1991) suggests a general method for estimating allele (or haplotype) frequencies from data on relatives. He provides the maximum likelihood estimates of allele frequencies for any arbitrary structure of pedigree relationships among relatives. It is a clever application of the maximum likelihood method originally designed for pedigree analysis. Nevertheless, in support of this method a few remarks may be added that might be particularly helpful to the users who are unfamiliar with the literature in this area of research.

First, it may be noted that the inclusion of relatives for estimating allele frequencies at a locus has a comparatively long history in human genetics. Fisher (1940) examined the effect that inclusion of relatives had on the estimate of the proportion of recessives in the population; for estimating allele frequencies and their precision Cotterman (1947) developed a weighting system from family data by using the maximum likelihood principle. Finney (1948*a*, 1948*b*) and Smith (1957) suggested alternative methods to address the same problem. Finally, Chakraborty (1978), in an appendix to the work of Ferrell et al. (1978), suggested a combinatorial approach, showing that, in addition to the estimation of allele frequencies, one can estimate the number of independent genes sampled in a survey that includes data on relatives. In principle, while Boehnke's (1991) method is based on similar logic, none of the above developments of this subject is referred to in his work.

Second, the versatility of the computer algorithms, such as MENDEL (Lange et al. 1988), yields allele frequency estimates at any locus even if the latter has a complex mode of inheritance, while all previous attempts deal with simple Mendelian transmission rules and specific family structures. However, the estimate of the equivalent number of alleles, given by Boehnke (1991), needs an extra cautionary remark. It should not be equated to the number of independent genes sampled, derived by Chakraborty (1978). In a set of family data the number of independent genes sampled is truly a random variable (let us denote it by N), whose expectation and variance can be analytically obtained from the relationships among individuals included in the analysis. Chakraborty (1978) showed that the distribution of N in family data is dependent on the family structure and size. For example, in a nuclear family with genotype information available on both parents and s children, N becomes 4, irrespective of s, and hence the offspring genotypes do not give any extra information when all alleles are codominant. In fact, for a codominant locus, the inclusion of offspring genotypes when both parental genotypes are known introduces errors of random fluctuation of Mendelian segregation ratios.

When genotypic data are available only on s (≥ 1)

sibs, N can take values 2, 3, and 4, with mean and variance given by

$$E(N) = 4(2^s - 1)/2^s, \quad (1)$$

and

$$V(N) = 16 - (28/2^s) + (8/4^s) - E^2(N), \quad (2)$$

so that the ratio of the expected number of independent genes sampled to the total number of alleles assayed, $E(N)/2s$, can be quite small, because $E(N) \to 4$ as $s \to \infty$. Chakraborty (1978) also considered more complex situations, such as the inclusion of individuals with one parent, $s - 1$ of his/her sibs, and k offspring tested. On the basis of such evaluations, it was shown that it might be necessary to attach different weights to genotype data on individuals belonging to different generations, to arrive at a statistically consistent estimator of allele frequencies. Although Boehnke's (1991) likelihood function (represented by his eq. [1]) accomplishes that, his n^* (equivalent number of alleles sampled) does not truly represent N. This is so because n^* depends on the estimated allele frequency as well as on the family structure and size, while N is independent of the allele frequency estimates. Therefore, for a given data structure, N will remain the same for all alleles, while n^* can vary substantially over alleles.

Furthermore, a more intricate problem relates to the reference population for which allele frequency estimates are being sought. Is it the entire collection of individuals in a given space at a given point of time, or does it relate to individuals of different generations recorded at a particular point of time? Population biologists interested in allele frequency estimates may approach the task of estimating allele frequencies by depending on the definition of the reference population used. Some might prefer to ignore the relationship structure altogether, and others may prefer to follow either the weighting schemes suggested earlier or Boehnke's suggested algorithm.

Boehnke's (1991) work is reassuring in the sense that, like his predecessors, he also concludes that ignoring the familial relationship does not introduce any systematic bias into allele frequency estimates; it only makes the allele frequency estimates appear more precise than they actually are. Therefore, inclusion of relatedness might be extremely important when one compares allele frequencies between samples that have large differences in their family structures, particularly when the alleles are rare, such as is the case with recently arisen mutations (e.g., see Neel et al. 1988).

RANAJIT CHAKRABORTY
Center for Demographic and Population Genetics
University of Texas Graduate School of Biomedical Sciences
Houston

References

Boehnke M (1991) Allele frequency estimation from data on relatives. Am J Hum Genet 48:22–25

Chakraborty R (1978) Number of independent genes examined in family surveys and its effect on gene frequency estimation. Am J Hum Genet 30:550–552

Cotterman CW (1947) A weighting system for the estimation of gene frequencies from family records. In: Contributions to the Laboratory of Vertebrate Biology, no. 33. Ann Arbor, University of Michigan, pp 1–21

Ferrell RE, Bertin T, Young R, Barton SA, Murillo F, Schull WJ (1978) The Aymara of western Bolivia. IV. Gene frequencies for eight blood groups and 19 protein and erythrocyte enzyme systems. Am J Hum Genet 30:539–549

Fisher RA (1940) The estimation of the proportion of recessives from tests carried on a sample not wholly unrelated. Ann Engenics 10:160–170

Finney DJ (1948*a*) The estimation of gene frequencies from family records. I. Factors without dominance. Heredity 2:199–218

——— (1948*b*) The estimation of gene frequencies from family records. II. Factors with dominance. Heredity 2: 369–390

Lange K, Weeks D, Boehnke M (1988) Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol 5:471–472

Neel JV, Satoh C, Smouse P, Asakawa J-I, Takahashi N, Goriki K, Fujita M, et al (1988) Protein variants in Hiroshima and Nagasaki: tales of two cities. Am J Hum Genet 43:870–893

Smith CAB (1957) Counting methods in genetical statistics. Ann Hum Genet 21:254–276

0002-9297/91/4901-0028$02.00

Am. J. Hum. Genet. 49:243–244, 1991

Reply to Chakraborty

To the Editor:

I thank Dr. Chakraborty for pointing out several references that describe other methods for allele frequency

Am. J. Hum. Genet. 49:895–897, 1991

Statistical Interpertation of DNA Typing Data

To The Editor:

Both the invited editorial by Lander (1991) and similar earlier commentaries on the subject of courtroom applications of DNA typing data have led to numerous arguments that simply defy well-known human population-genetic principles. In such criticisms, the authors employ a logic that may be called "reverse logic," whose mathematical validity is highly questionable. It is true that population substructure leads to genotypic proportions that deviate from Hardy-Weinberg expectations (HWE). Population substructure also produces gametic (as well as nongametic) disequilibria. These are well-known population-genetic principles. But Lander (1989*a*, 1989*b*, 1991) and others (e.g., see Cohen 1990) fail to recognize that there are other factors, particularly relevant to the RFLP analysis of DNA typing, which may produce these end results. Therefore, from the observed deviation from HWE and from an observed linkage disequilibrium, one cannot necessarily infer population substructure. It is unfortunate that in the peer-reviewed journals the above-mentioned authors have been allowed to make this inference without validating whether other associated features of DNA typing data conform to the substructuring hypothesis.

First, one might note that deviations from HWE, in the direction of deficiency of overall proportions of heterozygotes, have been noted in the DNA typing data in binned classification of alleles (Budowle et al. 1991). In contrast, it is demonstrated that, when we consider both the incomplete resolution of similar-sized alleles and measuremental errors of allele sizing, no deviation from HWE is detected (Devlin et al. 1990). One could argue that such tests do not have sufficient statistical power for detection of deviation from HWE. To ameliorate this problem, population data from several law-enforcement agencies have been subjected to nonparametric correlation analyses to check whether alleles of different sizes aggregate in any nonrandom fashion to form DNA types of individuals. Such tests, when properly applied (considering that the paternal and materal alleles cannot be distinguished in individuals in a population data base), result in no deviation from HWE. A correlation measure, originally devised for any general continuous trait with unknown (and possibly complex-shape) distribution (Karlin 1981), has substantially more power for detection of deviation from HWE. It can also be shown that Karlin's (1981) nonparametric correlation measure applies for quasi-continuous traits such as allele sizes at VNTR loci; it is distribution free, and its expectation can be derived even if nonrandom aggregation of alleles within individauls occurs because of population substructuring. These results indicate that, even if populations such as U.S. Caucasians, U.S. blacks, or Hispanics are truly substructured, their consequence on deviations from HWE is only trivial and cannot produce effects as gross as the ones indicated in the fictitious examples given (e.g., see Cohen 1990). Furthermore, even though it is well known that in RFLP analysis by Southern blot protocol the possibility exists that certain alleles of extreme sizes may remain undetected, Lander and others pay no attention to this in explaining the observed heterozygote deficiency. There is a voluminous literature (e.g., see Skibinski et al. 1983; Gart and Namm 1984; and cited references) that deals with such issues. It can be shown that even an extent of 6%–10% overall heterozygote deficiency can be explained if the frequency of such "nondetectable" alleles is 3%–6%. Samples of quite large sizes (e.g., more than 1,500–5,000 individuals/population) would be required for one to observe any single homozygote individual both of whose alleles are nondetectable. Even if this is found, there is no way to distinguish this type from those due to other vagaries of DNA typing (such as DNA degradation, insufficient DNA, etc.). Therefore, covert nondetectability of extreme-size alleles is a much *simpler* explanation of heterozygote deficiency of binned allele data.

Second, if the observed deviations from HWE were truly due to substructuring, what nature of substructuring (in terms of both the number of subpopulations and their evolutionary time of divergence) can produce such an extent of deviation can be shown. Such analyses reveal that, if we were to generate a 10% proportional heterozygote deficiency at a VNTR locus that has 90% heterozygosity, we have to invoke more than 20–30 subpopulations each of which should not have exchanged any gene among them for more than 40,000 years since their divergence from a common ancestry. This is clearly contrary to the origin and demography of the U.S. populations, where even among the orthodox religious populations the gene migration has been rather substantial (at least of the order of 10%/generation during the last century; e.g., see Kennedy 1944). Lander also fails to note that Lewontin's (1972) observation—i.e., that mean genetic difference between populations is far too small compared with genetic variation within populations—has been validated with a concept of populations that is much narrower than that of racial classification (e.g., see Chakraborty and Leimer 1986; Nei 1987).

Third, although Lander and others (e.g., see Cohen 1990) claim that there are substantial linkage disequilibria among VNTR alleles of unlinked loci, no specific data has been shown to this effect. Because of the presence of multiple alleles at such loci, this is particularly important, since the methods of estimation and detection of such multiallelic disequilibria are on a relatively softer ground. Applications of a recently proposed method (Hernández and Weir 1989) in specific case studies (for courtroom applications) reveal no disequilibria. Therefore, in the absence of any solid data on the extent of disequilibria, the claim that nonrandom aggregation of alleles of unlinked loci exists in individuals should not have been published in any peer-reviewed journal. In analogy with the departure from HWE, one can easily show that, for subpopulations that diverged from their common ancestry during the past 10,000–15,000 years (which would be the extreme for U.S. Caucasians, U.S. blacks, and U.S. Hispanics) and that exchanged genes among themselves, linkage disequilibria cannot attain any significant value at all. Cohen's (1990) numerical illustration requires linkage disequilibria that can be produced only if the subpopulations represent different species. I do not think that any human genetics will support such a statistical view!

Finally, if there are technical limitations to the RFLP analysis of DNA typing that generate deviation from the square law (HWE) or multiplication rule (linkage equilibria), the question is, Can we devise any modification to guard against biased probability calculation? The answer is yes! This is so, because, first, if there are covert nondetectable alleles, the gene-count method of estimating allele frequencies already gives overestimates of allele frequencies. Second, binning provides further cushions for allele-frequency estimates, cushions that are much larger than the expected deviations—and, when binning is used in conjunction with the use of 2p for the frequency of homozygotes (single-band patterns), this cushion is even greater. All these lead to estimations of chance occurrence of specific DNA types that can be biased only in the upward direction, establishing an objectivity in statistical interpretation of DNA typing data—an objectivity not portrayed in Lander's editorial.

RANAJIT CHAKRABORTY

Center for Demographic and Population Genetics
University of Texas Graduate School
of Biomedical Sciences
Houston

References

Budowle B, Giusti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, Presley LA, et al (1991) Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. Am J Hum Genet 48:841–855

Chakraborty R, Leimer O (1986) Genetic variation within subdivided population. In: Ryman N, Utter F (eds) Population genetics and fishery management. Washington University Press, Seattle, pp 89–120

Cohen JE (1990) DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number variability. Am J Hum Genet 46:358–368

Devlin B, Risch N, Roeder K (1990) No excess of homozygosity at loci used for DNA fingerprinting. Science 249: 1416–1420

Hernández JL, Weir BS (1989) A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics 45:53–70

Gart JJ, Namm J (1984) A score test for the possible presence of recessive alleles in generalized ABO-like genetic systems. Biometrics 40:887–894

Karlin S (1981) Sibling and parent-offspring correlation estimation with variable family size. Proc Natl Acad Sci USA 78:2664–2668

Kennedy RJR (1944) Single or triple melting pot? intermarriage trends in New Haven, 1870–1940. Am J Sociol 49: 331–339

Lander ES (1989*a*) DNA fingerprinting on trial. Nature 339: 501–505

——— (1989*b*) Population genetic considerations in the forensic use of DNA typing. In: DNA technology and forensic science. Banbury rep 32. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp 143–156

———. (1991) Research on DNA typing catching up with courtroom application. Am J Hum Genet 48:819–823

Lewontin RC (1972) The apportionment of human diversity. Evol Biol 6:381–398

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Skibinski DOF, Beardmore JA, Cross TF (1983) Aspects of the population genetics of *Mytilus* (Mylilidae; Mollusca) in the British isles. Biol J Linnaean Soc 19:137–183

0002-9297/91/4904-0026$02.00

Am. J. Hum. Genet. 49:895–897, 1991

Statistical Interpertation of DNA Typing Data

To The Editor:

Both the invited editorial by Lander (1991) and similar earlier commentaries on the subject of courtroom applications of DNA typing data have led to numerous arguments that simply defy well-known human population-genetic principles. In such criticisms, the authors employ a logic that may be called "reverse logic," whose mathematical validity is highly questionable. It is true that population substructure leads to genotypic proportions that deviate from Hardy-Weinberg expectations (HWE). Population substructure also produces gametic (as well as nongametic) disequilibria. These are well-known population-genetic principles. But Lander (1989*a*, 1989*b*, 1991) and others (e.g., see Cohen 1990) fail to recognize that there are other factors, particularly relevant to the RFLP analysis of DNA typing, which may produce these end results. Therefore, from the observed deviation from HWE and from an observed linkage disequilibrium, one cannot necessarily infer population substructure. It is unfortunate that in the peer-reviewed journals the above-mentioned authors have been allowed to make this inference without validating whether other associated features of DNA typing data conform to the substructuring hypothesis.

First, one might note that deviations from HWE, in the direction of deficiency of overall proportions of heterozygotes, have been noted in the DNA typing data in binned classification of alleles (Budowle et al. 1991). In contrast, it is demonstrated that, when we consider both the incomplete resolution of similar-sized alleles and measuremental errors of allele sizing, no deviation from HWE is detected (Devlin et al. 1990). One could argue that such tests do not have sufficient statistical power for detection of deviation from HWE. To ameliorate this problem, population data from several law-enforcement agencies have been subjected to nonparametric correlation analyses to check whether alleles of different sizes aggregate in any nonrandom fashion to form DNA types of individuals. Such tests, when properly applied (considering that the paternal and materal alleles cannot be distinguished in individuals in a population data base), result in no deviation from HWE. A correlation measure, originally devised for any general continuous trait with unknown (and possibly complex-shape) distribution (Karlin 1981), has substantially more power for detection of deviation from HWE. It can also be shown that Karlin's (1981) nonparametric correlation measure applies for quasi-continuous traits such as allele sizes at VNTR loci; it is distribution free, and its expectation can be derived even if nonrandom aggregation of alleles within individauls occurs because of population substructuring. These results indicate that, even if populations such as U.S. Caucasians, U.S. blacks, or Hispanics are truly substructured, their consequence on deviations from HWE is only trivial and cannot produce effects as gross as the ones indicated in the fictitious examples given (e.g., see Cohen 1990). Furthermore, even though it is well known that in RFLP analysis by Southern blot protocol the possibility exists that certain alleles of extreme sizes may remain undetected, Lander and others pay no attention to this in explaining the observed heterozygote deficiency. There is a voluminous literature (e.g., see Skibinski et al. 1983; Gart and Namm 1984; and cited references) that deals with such issues. It can be shown that even an extent of 6%–10% overall heterozygote deficiency can be explained if the frequency of such "nondetectable" alleles is 3%–6%. Samples of quite large sizes (e.g., more than 1,500–5,000 individuals/population) would be required for one to observe any single homozygote individual both of whose alleles are nondetectable. Even if this is found, there is no way to distinguish this type from those due to other vagaries of DNA typing (such as DNA degradation, insufficient DNA, etc.). Therefore, covert nondetectability of extreme-size alleles is a much *simpler* explanation of heterozygote deficiency of binned allele data.

Second, if the observed deviations from HWE were truly due to substructuring, what nature of substructuring (in terms of both the number of subpopulations and their evolutionary time of divergence) can produce such an extent of deviation can be shown. Such analyses reveal that, if we were to generate a 10% proportional heterozygote deficiency at a VNTR locus that has 90% heterozygosity, we have to invoke more than 20–30 subpopulations each of which should not have exchanged any gene among them for more than 40,000 years since their divergence from a common ancestry. This is clearly contrary to the origin and demography of the U.S. populations, where even among the orthodox religious populations the gene migration has been rather substantial (at least of the order of 10% / generation during the last century; e.g., see Kennedy 1944). Lander also fails to note that Lewontin's (1972) observation—i.e., that mean genetic difference between populations is far too small compared with genetic variation within populations—has been validated with a concept of populations that is much narrower than that of racial classification (e.g., see Chakraborty and Leimer 1986; Nei 1987).

Third, although Lander and others (e.g., see Cohen 1990) claim that there are substantial linkage disequilibria among VNTR alleles of unlinked loci, no specific data has been shown to this effect. Because of the presence of multiple alleles at such loci, this is particularly important, since the methods of estimation and detection of such multiallelic disequilibria are on a relatively softer ground. Applications of a recently proposed method (Hernández and Weir 1989) in specific case studies (for courtroom applications) reveal no disequilibria. Therefore, in the absence of any solid data on the extent of disequilibria, the claim that nonrandom aggregation of alleles of unlinked loci exists in individuals should not have been published in any peer-reviewed journal. In analogy with the departure from HWE, one can easily show that, for subpopulations that diverged from their common ancestry during the past 10,000–15,000 years (which would be the extreme for U.S. Caucasians, U.S. blacks, and U.S. Hispanics) and that exchanged genes among themselves, linkage disequilibria cannot attain any significant value at all. Cohen's (1990) numerical illustration requires linkage disequilibria that can be produced only if the subpopulations represent different species. I do not think that any human genetics will support such a statistical view!

Finally, if there are technical limitations to the RFLP analysis of DNA typing that generate deviation from the square law (HWE) or multiplication rule (linkage equilibria), the question is, Can we devise any modification to guard against biased probability calculation? The answer is yes! This is so, because, first, if there are covert nondetectable alleles, the gene-count method of estimating allele frequencies already gives overestimates of allele frequencies. Second, binning provides further cushions for allele-frequency estimates, cushions that are much larger than the expected deviations—and, when binning is used in conjunction with the use of 2p for the frequency of homozygotes (single-band patterns), this cushion is even greater. All these lead to estimations of chance occurrence of specific DNA types that can be biased only in the upward direction, establishing an objectivity in statistical interpretation of DNA typing data—an objectivity not portrayed in Lander's editorial.

RANAJIT CHAKRABORTY

Center for Demographic and Population Genetics
University of Texas Graduate School
of Biomedical Sciences
Houston

References

Budowle B, Giusti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, Presley LA, et al (1991) Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. Am J Hum Genet 48:841–855

Chakraborty R, Leimer O (1986) Genetic variation within subdivided population. In: Ryman N, Utter F (eds) Population genetics and fishery management. Washington University Press, Seattle, pp 89–120

Cohen JE (1990) DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number variability. Am J Hum Genet 46:358–368

Devlin B, Risch N, Roeder K (1990) No excess of homozygosity at loci used for DNA fingerprinting. Science 249: 1416–1420

Hernández JL, Weir BS (1989) A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics 45:53–70

Gart JJ, Namm J (1984) A score test for the possible presence of recessive alleles in generalized ABO-like genetic systems. Biometrics 40:887–894

Karlin S (1981) Sibling and parent-offspring correlation estimation with variable family size. Proc Natl Acad Sci USA 78:2664–2668

Kennedy RJR (1944) Single or triple melting pot? intermarriage trends in New Haven, 1870–1940. Am J Sociol 49: 331–339

Lander ES (1989*a*) DNA fingerprinting on trial. Nature 339: 501–505

——— (1989*b*) Population genetic considerations in the forensic use of DNA typing. In: DNA technology and forensic science. Banbury rep 32. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp 143–156

——— (1991) Research on DNA typing catching up with courtroom application. Am J Hum Genet 48:819–823

Lewontin RC (1972) The apportionment of human diversity. Evol Biol 6:381–398

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Skibinski DOF, Beardmore JA, Cross TF (1983) Aspects of the population genetics of *Mytilus* (Mylilidae; Mollusca) in the British isles. Biol J Linnaean Soc 19:137–183

0002-9297/91/4904-0026$02.00

Session 48: Genetic Variability and Population Differentiation

252

Population genetics of hypervariable loci. R.Chakraborty*. Genetics Centers. Graduate School of Biomedical Sciences. Univ. of Texas Health Science Center. Houston, Texas, USA.

The human genome contains a large number of loci at which the level of polymorphism is extremely high (often exceeding 95%). This hypervariability is produced by copy number variation of tandem repeats of core sequences that vary from locus to locus, and are detectable either by Southern gel electrophoresis or by polymerase chain reaction (PCR)-based methods. PCR-based methods make allelic variations easier to detect. The mechanisms of production of new alleles at such hypervariable loci differ substantially from those at traditional protein-coding loci. Nevertheless, empirical observations on genetic variation within and between human groups as well as interspecific comparisons between man and chimpanzee suggest that the evolutionary models used for protein data provide important insight about population structure and genetic differentiation from data on hypervariable loci. In comparison with the protein loci, the hypervariable loci exhibit a greater extent of inter-locus variation caused by their high mutation rates, about 10- to 50-fold larger than that at the protein-coding loci. The PCR-based protocols for assaying genetic variation at such loci are better suited for evolutionary studies, although Southern gel data also can be used provided precautionary measures are adopted to take account of imperfect resolution of alleles in such data. (Research supported by grant GM 41399 from NIH and 90-IJ-CX-0038 from NIJ).

253

Mutations in the G6PD molecule and their population distribution. R. Lisker. Instituto Nacional de la Nutrición Salvador Zubirán, México City, México.

Around 400 supposedely unique variants of G6PD have been biochemically characterized. Most of them were identified by a standardized methodology recommended in 1967 by the W.H.O. In spite of the above there is some unavoidable degree of interlaboratory variability such as differences in commercial reagents or in the degree of enzyme denaturation during storage and purification, and the question of how many of these variants are truly unique is a real one. The cloning and sequencing of G-6-PD has made possible and entirely new approach for it's study and it is already clear that variants that were supposedely different are really identical and some thought to be the same are heterogeneous.

G-6-PD variants are distributed world wide and the most comon one, A+, present in 20% of black African males has an A to G transition at nucleotide 376. Ninety % of the A- variants have in addition a G to A transition at nucleotide 202 and in a few patients the second mutation occurs at nucleotides 680 or 968. It has now been proven that G-6-PD's Betica (Spain), Matera (Italy), Alabama (U.S.A.), Tepic, Castilla and Distrito Federal (México) are identical to the A-, indicating the widespread distribution of this African gene. Similar findings are being encountered in G-6-PD Mediterranean, also widely distributed.

254

History and geography of human genes. A. Piazza* (1), P. Menozzi (2), and L.L. Cavalli-Sforza (3). (1) Dpt. of Genetics, Torino University, Torino, Italy, (2) Inst. of Ecology, Parma University, Parma, Italy and (3) Dpt. of Genetics, Stanford University, Stanford, California, USA.

The geographical distribution of gene frequencies can provide insight into the evolution of the various human populations. A major problem of interpretation is that genetic similarity may point to a common historical origin, but it may also be due to experience of similar geographical environments. Nor should we forget that similar cultural ways of life may favour the survival of the same genes.

Geography of human genes - as more thoroughly discussed in a forthcoming book by Princeton University Press - seem to show a remarkable correlation with archaeological and linguistic data. This means that a great deal of human diversity achieves a very simple explanation in history itself. As Gould (1989) pointed out "the primary signature of time and history is not effaced by immediate adaptation to prevailing circumstances or by recent episodes of conquest and amalgamation: we remain the children of our past". In fact an interesting result of our analyses is that this past has been mostly frozen in the memory of all our genes.

255

Genetics and the peopling of the Americas. E. J. E. Szathmary. University of Western Ontario, London, Canada.

Disputed questions regarding the peopling of North America concern the timing and location of human entry to the continent. A single place of entry (Beringia) is likely, but controversy over the timing of entry remains. One view favours early migration between 25-20,000 years ago. Another posits three different times of entry.

The hypothesis of three distinct migrations is said to be supported by biologic and linguistic data. In fact, evidence based on nuclear genes at many loci can support two distinct entry models. One posits a single occupying population migrating more than 20,000 years ago, the descendants of which became separated by coalescing glaciers during the last glacial maximum. Genetic differentiation between and within the separated groups then occurred. This model accounts for the observed genetic closeness of Athapaskan-speaking Indians to Eskimos, in comparison to that of other Indians. However, the model of three distinct migrations is equally supported by the multilocus nuclear data if one assumes three different times of migration.

Attempts to resolve the issue include analysis of patterns of affinity revealed by single "loci", e.g., Immunoglobulin Gm, and mtDNA. Gm data show closeness of Athapaskans to Eskimos, and mtDNA also reveal the Athapaskan-Eskimo link (Shields, 1991). The affinities revealed by mtDNA and Gm data make sense if the effective size of the ancestral population included all the peoples of Beringia, the group remaining northwest of the glaciers. Time depths of mtDNA divergences within other Native Americans support the notion of larger effective population sizes.

256

Gm haplotype distribution in Amerindians. F. M. Salzano*(1) and S. M. Callegari-Jacques(1)(2). (1) Genetics Dept., Fed. Univ. Rio Grande do Sul, Porto Alegre, RS, Brazil; (2) Statistics Dept., Fed. Univ.Rio Grande do Sul,Porto Alegre RS, Brazil.

A review was made of all available data on the Gm system in Amerindians. The most informative set included 60 Indian groups with less than 5% of non-Indian admixture, tested for at least 6 antigens. Two haplotypes Gm*1;21 and Gm*1,2;21 constitute in average 95% of this gene pool. But Gm*3;5, Gm*1;11,15,16 and Gm*1;5 also contribute to population discrimination. The amount of genetic diversity between groups, as evaluated by F, is almost identical in North and South American Indians (0.11 and 0.10, respectively). There is a significant association between the Gm distribution and the languages spoken by these individuals. Axis 1 of a correspondence analysis, strongly influenced (52%) by Gm*1,2;21 shows a clear geographical gradient. Starts with very low values in St. Lawrence island, Alaska, increases to a maximum in Panamá, and afterwards decreases to medium numbers in southwestern South America. A factor associated in some way with climate may be responsible for this trend.

257

HLA class II genes and haplotypes: a molecular analysis of the colonization of the Pacific. S.W. Serjeantson* and X. Gao.
Human Genetics Group, John Curtin School of Medical Research, Australian National University, Canberra, Australia.

The relative distributions of DR,DQ haplotypes have been determined in Australian Aborigines, Papua New Guinea highlanders, coastal Melanesians, Micronesians, Polynesians, Javanese and Southern and Northern Chinese. Using sequence-specific oligonucleotides (SSOs) for hybridization of PCR products from DRB1, DRB3, DRB5, DQA1 and DQB1 genes, more than 2,200 haplotypes have been examined. Many haplotypes were unique to Oceania. For instance, the predominant DR2 haplotype in Oceania involves a novel combination of DRB1*1502, DRB5*0101 alleles; this haplotype occurs sporadically in Java, but not in China. In Southern China, the most frequent DR2 haplotype has the unusual arrangement DRB1*1602, DRB5*0101. Novel reaction patterns with SSOs have led to the identification of several new DR4 and DRw6 DRB1 alleles, as confirmed by DNA sequencing, and the distributions of these provide new insights into population affinities in Asia-Oceania.

Session 48: Genetic Variability and Population Differentiation

252

Population genetics of hypervariable loci. R.Chakraborty*. Genetics Centers, Graduate School of Biomedical Sciences, Univ. of Texas Health Science Center, Houston, Texas, USA.

The human genome contains a large number of loci at which the level of polymorphism is extremely high (often exceeding 95%). This hypervariability is produced by copy number variation of tandem repeats of core sequences that vary from locus to locus, and are detectable either by Southern gel electrophoresis or by polymerase chain reaction (PCR)-based methods. PCR-based methods make allelic variations easier to detect. The mechanisms of production of new alleles at such hypervariable loci differ substantially from those at traditional protein-coding loci. Nevertheless, empirical observations on genetic variation within and between human groups as well as interspecific comparisons between man and chimpanzee suggest that the evolutionary models used for protein data provide important insight about population structure and genetic differentiation from data on hypervariable loci. In comparison with the protein loci, the hypervariable loci exhibit a greater extent of inter-locus variation caused by their high mutation rates, about 10- to 50-fold larger than that at the protein-coding loci. The PCR-based protocols for assaying genetic variation at such loci are better suited for evolutionary studies, although Southern gel data also can be used provided precautionary measures are adopted to take account of imperfect resolution of alleles in such data. (Research supported by grant GM 41399 from NIH and 90-IJ-CX-0038 from NIJ).

253

Mutations in the G6PD molecule and their population distribution. R. Lisker. Instituto Nacional de la Nutrición Salvador Zubirán, México City, México.

Around 400 supposedely unique variants of G6PD have been biochemically characterized. Most of them were identified by a standardized methodology recommended in 1967 by the W.H.O. In spite of the above there is some unavoidable degree of interlaboratory variability such as differences in commercial reagents or in the degree of enzyme denaturation during storage and purification, and the question of how many of these variants are truly unique is a real one. The cloning and sequencing of G-6-PD has made possible and entirely new approach for it's study and it is already clear that variants that were supposedely different are really identical and some thought to be the same are heterogeneous.

G-6-PD variants are distributed world wide and the most comon one, A+, present in 20% of black African males has an A to G transition at nucleotide 376. Ninety % of the A- variants have in addition a G to A transition at nucleotide 202 and in a few patients the second mutation occurs at nucleotides 680 or 968. It has now been proven that G-6-PD's Betica (Spain), Matera (Italy), Alabama (U.S.A.), Tepic, Castilla and Distrito Federal (México) are identical to the A-, indicating the widespread distribution of this African gene. Similar findings are being encountered in G-6-PD Mediterranean, also widely distributed.

254

History and geography of human genes. A. Piazza* (1), P. Menozzi (2), and L.L. Cavalli-Sforza (3). (1) Dpt. of Genetics, Torino University, Torino, Italy, (2) Inst. of Ecology, Parma University, Parma, Italy and (3) Dpt. of Genetics, Stanford University, Stanford, California, USA.

The geographical distribution of gene frequencies can provide insight into the evolution of the various human populations. A major problem of interpretation is that genetic similarity may point to a common historical origin, but it may also be due to experience of similar geographical environments. Nor should we forget that similar cultural ways of life may favour the survival of the same genes.

Geography of human genes - as more thoroughly discussed in a forthcoming book by Princeton University Press - seem to show a remarkable correlation with archaeological and linguistic data. This means that a great deal of human diversity achieves a very simple explanation in history itself. As Gould (1989) pointed out "the primary signature of time and history is not effaced by immediate adaptation to prevailing circumstances or by recent episodes of conquest and amalgamation: we remain the children of our past". In fact an interesting result of our analyses is that this past has been mostly frozen in the memory of all our genes.

255

Genetics and the peopling of the Americas. E. J. E. Szathmary. University of Western Ontario, London, Canada.

Disputed questions regarding the peopling of North America concern the timing and location of human entry to the continent. A single place of entry (Beringia) is likely, but controversy over the timing of entry remains. One view favours early migration between 25-20,000 years ago. Another posits three different times of entry.

The hypothesis of three distinct migrations is said to be supported by biologic and linguistic data. In fact, evidence based on nuclear genes at many loci can support two distinct entry models. One posits a single occupying population migrating more than 20,000 years ago, the descendants of which became separated by coalescing glaciers during the last glacial maximum. Genetic differentiation between and within the separated groups then occurred. This model accounts for the observed genetic closeness of Athapaskan-speaking Indians to Eskimos, in comparison to that of other Indians. However, the model of three distinct migrations is equally supported by the multilocus nuclear data if one assumes three different times of migration.

Attempts to resolve the issue include analysis of patterns of affinity revealed by single "loci", e.g., Immunoglobulin Gm, and mtDNA. Gm data show closeness of Athapaskans to Eskimos, and mtDNA also reveal the Athapaskan-Eskimo link (Shields, 1991). The affinities revealed by mtDNA and Gm data make sense if the effective size of the ancestral population included all the peoples of Beringia, the group remaining northwest of the glaciers. Time depths of mtDNA divergences within other Native Americans support the notion of larger effective population sizes.

256

Gm haplotype distribution in Amerindians. F. M. Salzano*(1) and S. M. Callegari-Jacques(1)(2). (1) Genetics Dept., Fed. Univ. Rio Grande do Sul, Porto Alegre, RS, Brazil; (2) Statistics Dept., Fed. Univ.Rio Grande do Sul.Porto Alegre RS, Brazil.

A review was made of all available data on the Gm system in Amerindians. The most informative set included 60 Indian groups with less than 5% of non-Indian admixture, tested for at least 6 antigens. Two haplotypes Gm*1;21 and Gm*1,2;21 constitute in average 95% of this gene pool. But Gm*3;5, Gm*1;11,15,16 and Gm*1;5 also contribute to population discrimination. The amount of genetic diversity between groups, as evaluated by F, is almost identical in North and South American Indians (0.11 and 0.10, respectively). There is a significant association between the Gm distribution and the languages spoken by these individuals. Axis 1 of a correspondence analysis, strongly influenced (52%) by Gm*1,2;21 shows a clear geographical gradient. Starts with very low values in St. Lawrence island, Alaska, increases to a maximum in Panamá, and afterwards decreases to medium numbers in southwestern South America. A factor associated in some way with climate may be responsible for this trend.

257

HLA class II genes and haplotypes: a molecular analysis of the colonization of the Pacific. S.W. Serjeantson* and X. Gao.
Human Genetics Group, John Curtin School of Medical Research, Australian National University, Canberra, Australia.

The relative distributions of DR,DQ haplotypes have been determined in Australian Aborigines, Papua New Guinea highlanders, coastal Melanesians, Micronesians, Polynesians, Javanese and Southern and Northern Chinese. Using sequence-specific oligonucleotides (SSOs) for hybridization of PCR products from DRB1, DRB3, DRB5, DQA1 and DQB1 genes, more than 2,200 haplotypes have been examined. Many haplotypes were unique to Oceania. For instance, the predominant DR2 haplotype in Oceania involves a novel combination of DRB1*1502, DRB5*0101 alleles; this haplotype occurs sporadically in Java, but not in China. In Southern China, the most frequent DR2 haplotype has the unusual arrangement DRB1*1602, DRB5*0101. Novel reaction patterns with SSOs have led to the identification of several new DR4 and DRw6 DRB1 alleles, as confirmed by DNA sequencing, and the distributions of these provide new insights into population affinities in Asia-Oceania.

Book Reviews

Biomechanics and Exercise Physiology. By Arthur P. Johnson, xv + 493 pp. New York: John Wiley & Sons. 1991. $85.00 (cloth).

This book is a refreshing change from the usual texts on the biomechanics of sport focusing primarily on the physics of human movement. Arthur Johnson's *Biomechanics and Exercise Physiology* focuses on the mechanics of the cardiovascular, respiratory, and thermoregulatory systems; relatively little attention is given to the physics of movement. The text neatly integrates concepts of applied physiology with those of engineering and emphasizes the mathematical quantification of cardiorespiratory physiology under the stressful conditions of exercise.

The author has taken a quantitative approach to describing physiological processes and provides many mathematical models that can be used for predicting physiological responses to a variety of exercise conditions. Units of measurement for physiological processes are given considerable attention. Because units of measurement vary considerably among the subdivisions of the disciplines of physiology and engineering, the author has standardized the models presented to metric units.

The book begins with a chapter on exercise limitations that includes a discussion of the role of models in describing how physiological systems work. Models for exercise intensity and duration, muscle metabolism during exercise, recovery from exercise, lactate threshold, and oxygen uptake kinetics are included. The traditional biomechanics of exercise (e.g., physics of human movement) are presented on a limited basis in Chapter 2. The remainder of the book includes chapters on cardiovascular responses, respiratory responses, and thermal response to exercise. Detailed attention is given to the control and regulation of these systems, and these three chapters make up the bulk of the book. Skeletal muscle physiology and the mechanism of muscular contraction are not considered.

This book is definitely written from the viewpoint of an engineer. It is not suitable as an introductory text for either exercise physiology or biomechanics students. Other sources must be relied on to obtain basic information about these disciplines. This book could serve as a sole text in a graduate-level course in biomechanics and exercise physiology or as a supplementary text in courses in exercise physiology, biomechanics, biophysics, or bioengineering. The advanced nature of the book makes it best suited for graduate-level course work. A background in human physiology, physics, and mathematics is required to grasp fully many of the concepts presented.

Parts of the book may serve as a valuable reference source for applied physiologists and biomechanists. However, due to the broad scope of the material, it is doubtful that the researcher would find the entire text useful as a reference source. Because of the unique content of this book, I recommend it highly as a reference source for applied physiologists, biomechanists, and bioengineers.

James E. Graves
Center for Exercise Science
University of Florida
Gainesville, Florida

Convergent Issues in Genetics and Demography. Edited by J.A. Adams, A. Hermalin, D. Lam, and P.E. Smouse. xii + 361 pp. New York: Oxford University Press. 1990. $49.95 (cloth).

The Malthusian theory of population growth is a common denominator in many research themes of social as well as population biological sciences. This alone justifies Malthus' place as a common ancestor of the disciplines of population biology and demography. However, it is also true that demographers and population biologists took rather diverse paths for almost two centuries in spite of sharing some common goals. Even though this was caused mainly by different departmental affiliations of the demographers and population biologists, it would be wrong to assert that the convergent issues of demography and population biology remained unnoticed during the past two centuries. The institutional affiliation of this reveiwer, the Center for Demographic and population Genetics, suggests that an integrative approach of studying the demo-

graphic and genetic aspects of population biology might bring a synergistic development of these two disciplines. The academic success of our twenty year-old center, however, does not imply that the congruent issues of (population) genetics and demography have been completely resolved. This volume represents statements on some key congruent issues, how they are addressed by demographers and geneticists, where their apparently divergent paths intersect, and what can be learned from the experiences gathered to solve the current problems.

Convergent Issues in Genetics and Demography is an outgrowth of a series of papers presented at an international conference held at the University of Michigan, Ann Arbor, October 7–8, 1988. Unlike many symposia volumes, this is much more than a simple collection of loosely connected chapters. Organized in four sections, the theme always remains the major points of congruence between the areas of interest to demographers and geneticists. This is clearly stated in the prefatory notes of the editors, but I would go further to state that there is sufficient food for thought in this volume for economists, health practioners, and epidemiologists when demographic and population genetic principles can be effectively utilized.

This book begins with five contributions in the first section in which the utility of historical information in demographic and genetic investigations is explored. Do historical records explain the population genetic and demographic patterns of extant populations, or with the help of historical data can we reconstruct the past genetic or demographic patterns from those observed at present? These are some of the questions addressed by contributors in this section.

Variation among individuals is a cornerstone observation in population biology, which would be called *heterogeneity* in demography. The second section, consisting of five contributions, deals with this issue. In my opinion, this subject is of great importance because it serves as the stepping stone for an area that needs further exploration. Without a clear understanding of the variation of individual contributions to the next generation or chance of survival until reproduction, the role of the nature-nurture interaction in determining the quality of life or disease burden in a population cannot be fully depicted. While this section provides several lucid discussions on the frailty models applicable to genetics and demography, a clearer statement of the need for correlated frailty models would have been helpful. This is so because, for many chronic disease studies, when the involvement of genes is of interest to geneticists, the definition of the penetrance function (frailty) should incorporate familial correlations with regard to common family background in addition to their dependence on genotypes.

The third section consists of four contributions exploring the interface of demography and genetics in epidemiological research. Human populations exhibit substantial differences with regard to health detriments, no matter what metric is used to measure such detriments. As Schull (in Chapter 14) points out, there are countless questions, genetic and nongenetic, that arise in the search for an understanding of the origin of these differences. While the basic scientists are trying to unravel the biological basis of the repair mechanisms of exposures to environmental insults, and thereby characterize the basis of individual variation of the capacities to repair damages, the applicability of such findings in predicting risk and in policy making for the betterment of the quality of life has been rather limited. The interdisciplinary subject of genetic epidemiology, therefore, stands at the crossroads of genetics and demography, having a tremendous prognostic value for future work in this direction (see Chapter 14).

The final section consists of four papers in which some "persistent" issues of genetics and demography are discussed. Although this is not explicitly stated, these papers are closely related to the subject of conservation biology, in which both genetics and demography play a pivotal role in examining what features would make a population viable or stable. Even though Ewens (Chapter 20) discusses the concept of minimum viable population (MVP) size, except for Christiansen's discussion (Chapter 19) on the natural populations of a marine fish species (*Zoarces viviparus*), no other nonhuman organism is discussed at length in this volume.

In the above count, I intentionally did not enumerate the opening chapters of each section, where the editors provide additional notes on the chapters. I view this as a strong point of this volume, because the introductory comments of the editors establish a cohesiveness for the subsequent presenta-

tions. These comments also place the contributions in the framework of the current problems in these areas of research. The cross referencing of the individual contributions, similarly, should also be useful to readers to place each in appropriate perspective.

To the human biologists, I recommend this volume as valuable material for a seminar course for advanced graduate students, since it contains not only a current statement of the problem areas of the interface of genetics and demography but also provides a good listing of future areas for exploration. Obviously, as indicated above, the topics covered in the volume are not exhaustive, nor was the volume intended to be so. Notwithstanding this minor defect, my enthusiasm about recommending this collection of papers is high, because the editors have done their homework well. The quality of the production is good, and there are ample references to other areas of research not discussed at length in this work. Uniform typesetting, clear and concise presentation of theory, and discussion of relevant data provide a good balance between analytical and empirical findings. With the hope that future attempts at continuing such a fruitful dialogue among the experts in the divergent fields of demography and genetics, this volume could become a strong foundation from which one can build a successful synthesis.

RANAJIT CHAKRABORTY
Center for Demographic and Population Genetics
University of Texas Graduate School of Biomedical Sciences
Houston, Texas

Santé Communautaire et Soins de Santé Primaires/Community Health and Primary Health Care. xii + 82 pp. Les Bulletins du Centre International de l'Enfance No. 1. Paris: Centre International de l'Enfance. 1991. $40.00 for four issues (paper).

This small volume published by the International Children's Centre (ICC) in Paris is the first of its series. Its aim is to provide a bibliographic bulletin with brief summaries of international research published on the topics of health care. The contents reflect the journals received by the ICC in 1990, and thus, as noted in the volume itself, it is somewhat selective and nonexhaustive. Brief synopses are provided for the selected publications, bilingually in French and English. The publishers offer the service of sending xerox copies of articles reviewed in the bulletin at the rate of three dollars per ten pages.

The 66 reviews in this issue are of variable quality and range in content from descriptions of professional, traditional and family health care providers; to reports of the social, economic, and political contexts in which health care providers function; to the anthropological, economic, and social relationships between populations and health care providers. An index of research by country and author is provided, and the publications are classified by four main themes: 1) professional health services, 2) other health resources; 3) social processes (including macroeconomic aspects, sociocultural aspects, and epidemiological aspects), and 4) the interfaces between health services and society (which includes health-seeking behavior, community participation, linkages between cultural models, financing of health services, health economic models, and drugs).

While much of the contents can be accessed by *Medline*, one of the advantages of such a bulletin for the U.S. community is the coverage of research reports from French journals, many of which present interesting data collected from Francophone Africa. I found the bulletin to be a particularly useful reference in identifying research reports of interest in medical anthropology. One of the weaker aspects of the present bulletin is that a number of the publications included were published from 1986 to 1988. However, in that this is the first volume in the series, it is understandable that there is a need to be comprehensive in terms of inclusive reporting from the journals at hand. I would recommend the publication as a library resource rather than for a personal subscription.

It should be noted that the International Children's Centre publishes a number of bibliographical bulletins of interest to medical anthropologists and public health specialists and has set up a bibliographical database on maternal and child health that can be accessed by correspondence or in person and is cited as being available on laser disc. In addition, the ICC has a library for researchers, which includes documentation from international agencies working in developing countries and documents from un-

Book Reviews

Biomechanics and Exercise Physiology. By Arthur P. Johnson, xv + 493 pp. New York: John Wiley & Sons. 1991. $85.00 (cloth).

This book is a refreshing change from the usual texts on the biomechanics of sport focusing primarily on the physics of human movement. Arthur Johnson's *Biomechanics and Exercise Physiology* focuses on the mechanics of the cardiovascular, respiratory, and thermoregulatory systems; relatively little attention is given to the physics of movement. The text neatly integrates concepts of applied physiology with those of engineering and emphasizes the mathematical quantification of cardiorespiratory physiology under the stressful conditions of exercise.

The author has taken a quantitative approach to describing physiological processes and provides many mathematical models that can be used for predicting physiological responses to a variety of exercise conditions. Units of measurement for physiological processes are given considerable attention. Because units of measurement vary considerably among the subdivisions of the disciplines of physiology and engineering, the author has standardized the models presented to metric units.

The book begins with a chapter on exercise limitations that includes a discussion of the role of models in describing how physiological systems work. Models for exercise intensity and duration, muscle metabolism during exercise, recovery from exercise, lactate threshold, and oxygen uptake kinetics are included. The traditional biomechanics of exercise (e.g., physics of human movement) are presented on a limited basis in Chapter 2. The remainder of the book includes chapters on cardiovascular responses, respiratory responses, and thermal response to exercise. Detailed attention is given to the control and regulation of these systems, and these three chapters make up the bulk of the book. Skeletal muscle physiology and the mechanism of muscular contraction are not considered.

This book is definitely written from the viewpoint of an engineer. It is not suitable as an introductory text for either exercise physiology or biomechanics students. Other sources must be relied on to obtain basic information about these disciplines. This book could serve as a sole text in a graduate-level course in biomechanics and exercise physiology or as a supplementary text in courses in exercise physiology, biomechanics, biophysics, or bioengineering. The advanced nature of the book makes it best suited for graduate-level course work. A background in human physiology, physics, and mathematics is required to grasp fully many of the concepts presented.

Parts of the book may serve as a valuable reference source for applied physiologists and biomechanists. However, due to the broad scope of the material, it is doubtful that the researcher would find the entire text useful as a reference source. Because of the unique content of this book, I recommend it highly as a reference source for applied physiologists, biomechanists, and bioengineers.

JAMES E. GRAVES
Center for Exercise Science
University of Florida
Gainesville, Florida

Convergent Issues in Genetics and Demography. Edited by J.A. Adams, A. Hermalin, D. Lam, and P.E. Smouse. xii + 361 pp. New York: Oxford University Press. 1990. $49.95 (cloth).

The Malthusian theory of population growth is a common denominator in many research themes of social as well as population biological sciences. This alone justifies Malthus' place as a common ancestor of the disciplines of population biology and demography. However, it is also true that demographers and population biologists took rather diverse paths for almost two centuries in spite of sharing some common goals. Even though this was caused mainly by different departmental affiliations of the demographers and population biologists, it would be wrong to assert that the convergent issues of demography and population biology remained unnoticed during the past two centuries. The institutional affiliation of this reveiwer, the Center for Demographic and population Genetics, suggests that an integrative approach of studying the demo-

graphic and genetic aspects of population biology might bring a synergistic development of these two disciplines. The academic success of our twenty year-old center, however, does not imply that the congruent issues of (population) genetics and demography have been completely resolved. This volume represents statements on some key congruent issues, how they are addressed by demographers and geneticists, where their apparently divergent paths intersect, and what can be learned from the experiences gathered to solve the current problems.

Convergent Issues in Genetics and Demography is an outgrowth of a series of papers presented at an international conference held at the University of Michigan, Ann Arbor, October 7–8, 1988. Unlike many symposia volumes, this is much more than a simple collection of loosely connected chapters. Organized in four sections, the theme always remains the major points of congruence between the areas of interest to demographers and geneticists. This is clearly stated in the prefatory notes of the editors, but I would go further to state that there is sufficient food for thought in this volume for economists, health practioners, and epidemiologists when demographic and population genetic principles can be effectively utilized.

This book begins with five contributions in the first section in which the utility of historical information in demographic and genetic investigations is explored. Do historical records explain the population genetic and demographic patterns of extant populations, or with the help of historical data can we reconstruct the past genetic or demographic patterns from those observed at present? These are some of the questions addressed by contributors in this section.

Variation among individuals is a cornerstone observation in population biology, which would be called *heterogeneity* in demography. The second section, consisting of five contributions, deals with this issue. In my opinion, this subject is of great importance because it serves as the stepping stone for an area that needs further exploration. Without a clear understanding of the variation of individual contributions to the next generation or chance of survival until reproduction, the role of the nature-nurture interaction in determining the quality of life or disease burden in a population cannot be fully depicted. While this section provides several lucid discussions on the frailty models applicable to genetics and demography, a clearer statement of the need for correlated frailty models would have been helpful. This is so because, for many chronic disease studies, when the involvement of genes is of interest to geneticists, the definition of the penetrance function (frailty) should incorporate familial correlations with regard to common family background in addition to their dependence on genotypes.

The third section consists of four contributions exploring the interface of demography and genetics in epidemiological research. Human populations exhibit substantial differences with regard to health detriments, no matter what metric is used to measure such detriments. As Schull (in Chapter 14) points out, there are countless questions, genetic and nongenetic, that arise in the search for an understanding of the origin of these differences. While the basic scientists are trying to unravel the biological basis of the repair mechanisms of exposures to environmental insults, and thereby characterize the basis of individual variation of the capacities to repair damages, the applicability of such findings in predicting risk and in policy making for the betterment of the quality of life has been rather limited. The interdisciplinary subject of genetic epidemiology, therefore, stands at the crossroads of genetics and demography, having a tremendous prognostic value for future work in this direction (see Chapter 14).

The final section consists of four papers in which some "persistent" issues of genetics and demography are discussed. Although this is not explicitly stated, these papers are closely related to the subject of conservation biology, in which both genetics and demography play a pivotal role in examining what features would make a population viable or stable. Even though Ewens (Chapter 20) discusses the concept of minimum viable population (MVP) size, except for Christiansen's discussion (Chapter 19) on the natural populations of a marine fish species (*Zoarces viviparus*), no other nonhuman organism is discussed at length in this volume.

In the above count, I intentionally did not enumerate the opening chapters of each section, where the editors provide additional notes on the chapters. I view this as a strong point of this volume, because the introductory comments of the editors establish a cohesiveness for the subsequent presenta-

tions. These comments also place the contributions in the framework of the current problems in these areas of research. The cross referencing of the individual contributions, similarly, should also be useful to readers to place each in appropriate perspective.

To the human biologists, I recommend this volume as valuable material for a seminar course for advanced graduate students, since it contains not only a current statement of the problem areas of the interface of genetics and demography but also provides a good listing of future areas for exploration. Obviously, as indicated above, the topics covered in the volume are not exhaustive, nor was the volume intended to be so. Notwithstanding this minor defect, my enthusiasm about recommending this collection of papers is high, because the editors have done their homework well. The quality of the production is good, and there are ample references to other areas of research not discussed at length in this work. Uniform typesetting, clear and concise presentation of theory, and discussion of relevant data provide a good balance between analytical and empirical findings. With the hope that future attempts at continuing such a fruitful dialogue among the experts in the divergent fields of demography and genetics, this volume could become a strong foundation from which one can build a successful synthesis.

RANAJIT CHAKRABORTY
Center for Demographic and Population Genetics
University of Texas Graduate School of Biomedical Sciences
Houston, Texas

Santé Communautaire et Soins de Santé Primaires/Community Health and Primary Health Care. xii + 82 pp. Les Bulletins du Centre International de l'Enfance No. 1. Paris: Centre International de l'Enfance. 1991. $40.00 for four issues (paper).

This small volume published by the International Children's Centre (ICC) in Paris is the first of its series. Its aim is to provide a bibliographic bulletin with brief summaries of international research published on the topics of health care. The contents reflect the journals received by the ICC in 1990, and thus, as noted in the volume itself, it is somewhat selective and nonexhaustive. Brief synopses are provided for the selected publications, bilingually in French and English. The publishers offer the service of sending xerox copies of articles reviewed in the bulletin at the rate of three dollars per ten pages.

The 66 reviews in this issue are of variable quality and range in content from descriptions of professional, traditional and family health care providers; to reports of the social, economic, and political contexts in which health care providers function; to the anthropological, economic, and social relationships between populations and health care providers. An index of research by country and author is provided, and the publications are classified by four main themes: 1) professional health services, 2) other health resources; 3) social processes (including macroeconomic aspects, sociocultural aspects, and epidemiological aspects), and 4) the interfaces between health services and society (which includes health-seeking behavior, community participation, linkages between cultural models, financing of health services, health economic models, and drugs).

While much of the contents can be accessed by *Medline*, one of the advantages of such a bulletin for the U.S. community is the coverage of research reports from French journals, many of which present interesting data collected from Francophone Africa. I found the bulletin to be a particularly useful reference in identifying research reports of interest in medical anthropology. One of the weaker aspects of the present bulletin is that a number of the publications included were published from 1986 to 1988. However, in that this is the first volume in the series, it is understandable that there is a need to be comprehensive in terms of inclusive reporting from the journals at hand. I would recommend the publication as a library resource rather than for a personal subscription.

It should be noted that the International Children's Centre publishes a number of bibliographical bulletins of interest to medical anthropologists and public health specialists and has set up a bibliographical database on maternal and child health that can be accessed by correspondence or in person and is cited as being available on laser disc. In addition, the ICC has a library for researchers, which includes documentation from international agencies working in developing countries and documents from un-

Letters to the Editor

Multiple Alleles and Estimation of Genetic Parameters: Computational Equations Showing Involvement of All Alleles

Genetic loci that exhibit multiple (more than two) segregating alleles are generally more useful than bi-allelic ones for population genetic studies simply because they offer greater potential for variation in observed number of alleles as well as allele frequency differences across populations. Since allele frequencies at a locus in a population are structurally constrained (they always add to one), a matrix treatment of allele frequency data at a multi-allelic locus requires deleting one allele from the analysis. Hence the resultant estimator may be construed as dependent on which allele is being eliminated in the process of estimation (BALAKRISHNAN 1973). Such situations have been faced by BALAKRISHNAN and SANGHVI (1968) and SMOUSE and SPIELMAN (1977) when they attempted to estimate genetic distances between populations by statistics parallel to Mahalanobis-D^2 (MAHALANOBIS 1936) for multivariate data. ROBERTS and HIORNS (1962) also suggested a method of estimating genetic admixture in a hybrid population using allele frequency data that requires elimination of one allele of a multiallelic locus. Recently, this issue has resurfaced in the least-square estimation of admixture components in a hybrid population (LONG 1991). Since these investigators generally presented their estimating equations in terms of "shortened" vectors of allele frequencies (by deleting one allele from each locus) and the variance-covariance matrix of such "shortened" vectors of sampled allele frequencies, in general it is not obvious whether or not the resultant estimators depend upon the allele that is eliminated from the analysis. As a result, such methods are criticized on the ground of the subjectivity involved in selecting the allele to be eliminated (BALAKRISHNAN 1973) although in some applications algebraic verifications are given to show that any allele can be dropped without affecting the estimate (LONG 1991). The purpose of this communication is to show that by exploiting a well-known property of the variance-covariance matrix of the cell frequencies of a multinomial distribution (KURCZYNSKI 1970) a simple translation of the matrix estimators can be obtained, which indicates that even though the formal representation requires deleting one allele, the computational equation truly needs the frequencies of all alleles. Therefore, such estimators are functions of the full array of allele frequencies.

This simple exercise has at least three implications. First, it shows that the resultant estimators can be computed by algebraic operations involving all allele frequencies (which consequently results in numerically more accurate estimates, because matrix inversions generally introduce round off errors, which can be substantial particularly when the array size is large). Second, analytical relationships between different estimators of genetic parameters (e.g., distance, fixation indices, or admixture components) can be studied with greater ease with such representations (see *e.g.*, CHAKRABORTY and RAO 1991). Finally, genetic polymorphisms detected by DNA markers such as the variable number of tandem repeat (VNTR) loci often involve allele numbers (per locus) exceeding several dozen, and treating them with matrix operations requires a large array size, and even with that numerical inaccuracies cannot be avoided. On the contrary, algebraic expressions such as the ones presented here should make the analysis of such allele frequency data easier and certainly numerically more accurate.

Although the technique suggested here has wider applications, I consider only two specific estimation problems.

Genetic distance with multiple alleles: Denoting p_{ijk} as the frequency of the kth allele ($k = 1, 2, \ldots, s_j + 1$) of the jth locus ($j = 1, 2, \ldots, r$) in the ith population ($i = 1, 2$), estimated from $n_{ij}/2$ individuals sampled from the ith population for the jth locus, BALAKRISHNAN and SANGHVI (1968) suggested an estimator of the genetic distance between the two populations, given by

$$G_c^2 = \sum_{j=1}^{r} \mathbf{d}_j' \mathbf{C}_j^{-1} \mathbf{d}_j, \tag{1}$$

where $\mathbf{d}_j$ is a column vector of dimension s_j (one less than the number of segregating alleles at the jth locus, $s_j + 1$) whose kth element is $d_{jk} = p_{1jk} - p_{2jk}$, and $\mathbf{C}_j$ is a square matrix of size $s_j \times s_j$ whose elements are

$$C_{jkl} = \begin{cases} p_{jk}(1 - p_{jk}), & \text{for } k = l, \\ -p_{jk}p_{jl}, & \text{for } k \neq l \end{cases} \tag{2}$$

for $k, l = 1, 2, \ldots, s_j$; in which p_{jk} is the average of the kth allele frequency at the jth locus across populations; *i.e.*,

$$p_{jk} = \sum_i n_{ij}p_{ijk} / \sum_i n_{ij}. \tag{3}$$

Obviously, the quadratic form of equation (1) is the analog of Mahalanobis-D^2 (MAHALANOBIS 1936) since $\mathbf{C}_j$, given by (2), is the common dispersion matrix of the "shortened" vector of allele frequencies, estimated from the average allele frequencies across populations. Equation 1 may be written in the algebraic form

$$G_c^2 = \sum_{j=1}^{r} \sum_{k=1}^{s_j} \sum_{l=1}^{s_j} (p_{1jk} - p_{2jk}) C_j^{kl} (p_{1jl} - p_{2jl}), \quad (4)$$

where C_j^{kl} is the (k,l)th element of the $\mathbf{C}_j^{-1}$ matrix.

In order to show that G_c^2 is dependent on all allele frequencies, KURCZYNSKI (1970) noted that the inverse of the matrix $\mathbf{C}_j$ (of Equation 2) has the form

$$C_j^{kl} = \begin{cases} p_{jk}^{-1} + p_{j,s_j+1}^{-1}, & \text{for } k = l, \\ p_{j,s_j+1}^{-1}, & \text{for } k \neq l, \end{cases} \quad (5)$$

for $k, l = 1, 2, \ldots, s_j$.

Inserting (5) in (4) and noting that

$$\sum_{k=1}^{s_j} (p_{1jk} - p_{2jk})^2 + \sum_{k \neq j=1}^{s_j} \sum (p_{1jk} - p_{2jk})(p_{1jl} - p_{2jl})$$

$$= \left[\sum_{k=1}^{s_j} (p_{1jk} - p_{2jk}) \right]^2 = (p_{1j,s_j+1} - p_{2j,s_j+1})^2,$$

we obtain

$$G_c^2 = \sum_{j=1}^{r} \sum_{k=1}^{s_j - 1} (p_{1jk} - p_{2jk})^2 / p_{jk}, \quad (6)$$

which depends on frequencies of every segregating allele, irrespective of which alleles are being dropped in the definition of the $\mathbf{d}_j$-vectors or $\mathbf{C}_j$ matrices. Equation 6 not only shows the involvement of all allele frequencies in the estimation, but also it is numerically simpler to compute than Equation 4. Note that the above proof also applies to SMOUSE and WILLIAM's (1982) measure of disease-gene association, where such equivalence is stated without a formal derivation. Furthermore, it demonstrates that BALAKRISHNAN and SANGHVI's (1968) measure is equivalent to the original estimator of SANGHVI (1953), except a multiplication factor. In addition, the above derivation shows that the alternative two estimators (G_c^2 and G_s^2) proposed by BALAKRISHNAN and SANGHVI (1968) are mathematically identical.

Another advantage of the representation of Equation 6 is that it clearly shows how SANGHVI's estimator of genetic distance is related to others. For example, considering the allele frequencies at a single locus (say, the jth locus), BHATTACHARYYA (1946) defined a distance statistic, θ^2, between populations, which satisfies the relationship

$$\text{Cos } \theta = \sum_{k=1}^{s_{j+1}} \{p_{1jk} p_{2jk}\}^{1/2}, \quad (7)$$

which can be written as

$$\begin{aligned} \text{Cos } \theta &= \frac{1}{2} \cdot \sum_{k=1}^{s_{j+1}} [(p_{1jk} + p_{2jk})^2 - (p_{1jk} - p_{2jk})^2]^{1/2} \\ &= \frac{1}{2} \cdot \sum_{k=1}^{s_{j+1}} (p_{1jk} + p_{2jk}) \left[1 - \frac{(p_{1jk} - p_{2jk})^2}{(p_{1jk} + p_{2jk})^2} \right]^{1/2} \quad (8) \\ &= 1 - \frac{1}{4} \cdot \sum_{k=1}^{s_{j+1}} \frac{(p_{1jk} - p_{2jk})^2}{p_{1jk} + p_{2jk}}. \end{aligned}$$

However, since $\text{Cos } \theta \approx 1 - \theta^2/2$, for small θ, Equation 8 approximates to

$$\theta^2 \approx \frac{1}{2} \cdot \sum_{k=1}^{s_{j+1}} (p_{1jk} - p_{2jk})^2 / (p_{1jk} + p_{2jk}), \quad (9)$$

showing that for genetically close populations (*i.e.*, for small θ), SANGHVI's (1953) and BHATTACHARYYA's (1946) distance estimators are equivalent, barring a multiplication factor. Equivalence of Equations 9 and 6 with 4 further shows that they are analogs of Mahalanobis-D^2 for categorial data. Several other such equivalence relationships between various distance functions are discussed in CHAKRABORTY and RAO (1991) who utilize representations such as Equation 6.

The same logic provides a formal proof of the assertion that in the absence of disequilibria (WEIR 1979), SMOUSE and SPIELMAN's (1977) multivariate distance function based on multiple-allele genotype score vectors reduces to the form of Equation 6.

Weighted least square estimate of admixture proportions: For a dihybrid population whose gene pool consists of a fraction M of genes from a parental population 1 and a fraction $(1 - M)$ from parental population 2, LONG (1991) recently suggested a weighted least square estimator of M, which in matrix notation takes the form

$$m_j = (\mathbf{x}_j' \mathbf{V}_j^{-1} \mathbf{x}_j)^{-1} \mathbf{x}_j' \mathbf{V}_j^{-1} \mathbf{y}_j, \quad (10)$$

where $\mathbf{x}_j$ and $\mathbf{y}_j$ are column vectors of dimension s_j (one less than the number of segregating alleles, s_{j+1}, at the jth locus), with their kth elements defined by $x_{jk} = p_{1jk} - p_{2jk}$ and $y_{jk} = p_{hjk} - p_{2jk}$, for $k = 1, 2, \ldots, s_j$, and $\mathbf{V}_j$ is a $s_j \times s_j$ matrix with elements

$$V_{jkl} = \begin{cases} E(p_{hjk}) \cdot [1 - E(p_{hjk})], & \text{for } k = l \\ -E(p_{hjk}) \cdot E(p_{hjl}), & \text{for } k \neq l \end{cases} \quad (11)$$

in which p_{ijk} is the frequency of the kth allele ($k = 1, 2, \ldots, s_{j+1}$) at the jth locus in the ith population ($i = 1$ or 2 for the parental populations) and $E(p_{hjk})$ is the expected allele frequency in the admixed population under the admixture model.

The estimator m_j (Equation 10) is based on the "shortened" vectors of allele frequency differences (dropping the (s_{j+1})th allele). However, noting that the elements of the $\mathbf{V}_j^{-1}$ matrix are given by

$$V_j^{kl} = \begin{cases} 1/E(p_{hjk}) + 1/E(p_{hj,s_j+1}), & \text{for } k = l, \\ 1/E(p_{hj,s_j+1}), & \text{for } k \neq l, \end{cases} \quad (12)$$

for $k, l = 1, 2, \ldots, s_j$, LONG (1991) verified that the estimator m_j of Equation 10 is invariant of the allele dropped from the analysis. To show explicitly that Equation 10 does not depict that it depends on all allele frequencies, first note that

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{x}_j = \sum_{k=1}^{s_j}\sum_{l=1}^{s_j}(p_{1jk} - p_{2jk})V_j^{kl}(p_{1jk} - p_{2jk}), \quad (13a)$$

and

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{y}_j = \sum_{k=1}^{s_j}\sum_{l=1}^{s_j}(p_{1jk} - p_{2jk})V_j^{kl}(p_{hjk} - p_{2jk}). \quad (13b)$$

Invoking (12) in (13a) and (13b), and noting that

$$p_{1j,s_j+1} - p_{2j,s_j+1} = -\sum_{i=1}^{s_j}(p_{ijk} - p_{2jk}), \quad (14a)$$

and

$$p_{hj,s_j+1} - p_{2j,s_j+1} = -\sum_{i=1}^{s_j}(p_{hjk} - p_{2jk}), \quad (14b)$$

we can rewrite (13a) and (13b) as

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{x}_j = \sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk}), \quad (15a)$$

and

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{Y}_j = \sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk}), \quad (15b)$$

so that Equation 10 becomes

$$m_j = \frac{\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk})}{\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk})}, \quad (16)$$

which is an equation of scalers. Expressed in this fashion, m_j involves each of the s_{j+j} segregating allele frequencies of both parental populations and the admixed one.

This representation (Equation 16) of the weighted least squares (WLS) estimator of LONG (1991) further shows that m_j (the WLS estimator) is identical to the classical BERNSTEIN (1931) estimator of admixture proportion for a bi-allelic locus, noted in LONG and SMOUSE (1983). With the notation p_{ij} and $q_{ij} = (1 - p_{ij})$ of the two allele frequencies at a locus the numerator and denominator of Equation 16 become

$$\frac{(p_{1j} - p_{2j})(p_{hj} - p_{2j})}{E(p_{hj})} + \frac{(q_{1j} - q_{2j})(q_{hj} - q_{2j})}{E(q_{hj})} = \frac{(p_{1j} - p_{2j})(p_{hj} - p_{2j})}{E(p_{hj})\cdot E(q_{hj})},$$

and

$$\frac{(p_{1j} - p_{2j})^2}{E(p_{hj})} + \frac{(q_{1j} - q_{2j})^2}{E(q_{hj})} = \frac{(p_{1j} - p_{2j})^2}{E(p_{hj})\cdot E(q_{hj})},$$

so that the cancellation of their common denominators results in

$$m_j = (p_{hj} - p_{2j})/(p_{1j} - p_{2j}) = (q_{hj} - q_{2j})/(q_{1j} - q_{2j}),$$

establishing the identity of the WLS and Bernstein's estimators for bi-allelic loci.

The combined estimator for allele frequency data on r loci, based on LONG's (1991) method, becomes

$$m = \frac{\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk})}{\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk})} \quad (17)$$

which avoids matrix manipulations of even bigger dimension.

The sampling error of m also has a corresponding scalar form. In LONG's notation, the sampling variance is

$$V(m) = \text{MSE}\cdot(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1}, \quad (18)$$

where the mean square error (MSE) of the admixture model is

$$\text{MSE} = (\mathbf{y} - m\mathbf{x})'\mathbf{V}^{-1}(\mathbf{y} - m\mathbf{x})/\sum_{j=1}^{r} s_j. \quad (19)$$

Invoking (12) in these quadratic forms, and using the identities (14a) and (14b), we can rewrite (19) as

$$\text{MSE} = \frac{\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}[(p_{hjk} - p_{2jk}) - m(p_{1jk} - p_{2jk})]^2/E(p_{hjk})}{\sum_{j=1}^{r} s_j} \quad (20)$$

which yields the sampling variance of m,

$$V(m) = \frac{\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}[(p_{hjk} - p_{2jk}) - m(p_{1jk} - p_{2jk})]^2/E(p_{hjk})}{\left[\sum_{j=1}^{r} s_j\right]\cdot\left[\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk})\right]}. \quad (21)$$

The variance of the admixture estimate based on the jth locus data is

$$V(m_j) = \text{MSE}\cdot\left[\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk})\right]^{-1}, \quad (22)$$

in which the expression (20) is used for evaluating the MSE.

In addition to the demonstration that Equations 16 and 22, or 17 and 21 involve all allele frequencies

from each population, their computational simplicity remain unaltered even if the sample sizes for different loci are different. Since the V matrix refers to the expected allele frequencies in the admixed population, all terms of the summation over j will have to be weighted by n_{hj}, the number of genes sampled for the jth locus from the admixed population. For example, the combined estimator becomes

$$m = \frac{\sum_{j=1}^{r} n_{hj} \sum_{k=1}^{s_j+1} (p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk})}{\sum_{j=1}^{r} n_{hj} \sum_{k=1}^{s_j+1} (p_{1jk} - p_{2jk})^2/E(p_{hjk})}. \quad (23)$$

The corresponding changes in its sampling variance are also similar.

Other population genetic applications of algebraic representations of quadratic forms involving inverses of multinomial variance-covariance matrices include the estimation of Wright's fixation indices in the context of analysis of population structure. Using approaches similar to the above, LONG's (1986) multiallelic generalizations of COCKERHAM's (1969, 1973) variance-covariance estimators of the fixation indices can also be reduced to algebraic forms, which indicate their relationship with some existing estimators suggested earlier (see *e.g.*, LI and HORVITZ 1953; CURIE-COHEN 1982; ROBERTSON and HILL 1984; WEIR and COCKERHAM 1984).

To close this commentary, I must mention that the algebraic reductions of the matrix estimators such as the ones mentioned above are not meant to denigrate the utility of matrix notations in population genetics. Matrix representations of functions of allele frequencies at multiallelic loci have their importance and place that cannot be denied. They serve the purpose of establishing the basis of the method of estimation that is not always obvious in the closed form algebraic expression. In some instances matrix estimators are unavoidable. For example, the estimator of admixture contributions from multiple (more than two) ancestral populations is straightforward in matrix notation (ELSTON 1971; CHAKRABORTY 1986) and the incorporation of all orders of disequilibria (WEIR 1979) in estimating parameters of population structure and genetic distance analyses requires matrix notations, although nearly equivalent algebraic forms are also available (see *e.g.*, WEIR and COCKERHAM 1984). Nevertheless, the primary intent of this note has been to demonstrate that the principle that these are independent of which allele is dropped from the analysis.

This work was supported by U.S. Public Health Service research grants GM 41399 from the National Institutes of Health and 90-IJ-CX-0038 from the National Institute of Justice. I thank P. E. SMOUSE, B. S. WEIR and an anonymous reviewer for their comments and suggestions on the work.

RANAJIT CHAKRABORTY
Center for Demographic and Population Genetics
University of Texas Graduate School of Biomedical Sciences
P. O. Box 20334
Houston, Texas 77225

LITERATURE CITED

BALAKRISHNAN, V., 1973 Use of distance in hybrid analysis, pp. 268–273 in *Genetic Structure of Populations*, edited by N. E. MORTON. University of Hawaii Press, Honolulu.

BALAKRISHNAN, V., and L. D. SANGHVI, 1968 Distances between populations on the basis of attribute data. Biometrics **24:** 859–865.

BERNSTEIN, F., 1931 Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung. Comitato Italiano per lo Studio dei Problemi della Popolazione. Instituto Poligrafico dello Stato. Rome.

BHATTACHARYYA, A., 1946 On a measure of divergence between two multinomial populations. Sankhya **7:** 401–406.

CHAKRABORTY, R., 1986 Gene admixture in human populations: models and predictions. Yearb. Phys. Anthropol. **29:** 1–43.

CHAKRABORTY, R., and C. R. RAO, 1991 Measurement of genetic variation for evolutionary studies, in *Handbook of Statistics*, Vol. 8, edited by C. R. RAO and R. CHAKRABORTY. Elsevier, Amsterdam (in press).

COCKERHAM, C. C., 1969 Variance of gene frequencies. Evolution **23:** 72–84.

COCKERHAM, C. C., 1973 Analyses of gene frequencies. Genetics **74:** 679–700.

CURIE-COHEN, M., 1982 Estimates of inbreeding in a natural population: a comparison of sampling properties. Genetics **100:** 339–358.

ELSTON, R. C., 1971 The estimation of admixture in racial hybrids. Ann. Hum. Genet. **35:** 9–17.

KURCZYNSKI, T. W., 1970 Generalized distance and discrete variables. Biometrics **26:** 525–534.

LI, C. C., and D. G. HORVITZ, 1953 Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet. **5:** 107–117.

LONG, J. C., 1986 The allelic correlation structure of Gainj- and Kalam-speaking people. I. the estimation and interpretation of Wright's *F*-statistics. Genetics **112:** 629–647.

LONG, J. C., 1991 The genetic structure of admixed populations. Genetics **127:** 417–428.

LONG, J. C., and P. E. SMOUSE, 1983 Intertribal geneflow between the Ye'cuana and Yanamamö: genetic analysis of an admixed village. Am. J. Phys. Anthropol. **61:** 411–422.

MAHALANOBIS, P. C., 1936 On the generalized distance in statistics. Proc. Natl. Inst. Sci. India **12:** 49–55.

ROBERTS, D. F., and R. W. HIORNS, 1962 The dynamics of racial admixture. Am. J. Hum. Genet. **14:** 261–277.

ROBERTSON, A., and W. G. HILL, 1984 Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics **107:** 703–718.

SANGHVI, L. D., 1953 Comparison of genetical and morphological methods for a study of biological differences. Am. J. Phys. Anthropol. **11:** 385–404.

SMOUSE, P. E., and R. S. SPIELMAN, 1977 How allocation of individuals depends on genetic differences among populations. Excerpta Med. Congr. Ser. No. **411:** 255–260.

SMOUSE, P. E., and R. C. WILLIAMS, 1982 Multivariate analysis of HLA-disease associations. Biometrics **38:** 757–768.

WEIR, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **35:** 235–254.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

Letters to the Editor

Multiple Alleles and Estimation of Genetic Parameters: Computational Equations Showing Involvement of All Alleles

Genetic loci that exhibit multiple (more than two) segregating alleles are generally more useful than bi-allelic ones for population genetic studies simply because they offer greater potential for variation in observed number of alleles as well as allele frequency differences across populations. Since allele frequencies at a locus in a population are structurally constrained (they always add to one), a matrix treatment of allele frequency data at a multi-allelic locus requires deleting one allele from the analysis. Hence the resultant estimator may be construed as dependent on which allele is being eliminated in the process of estimation (BALAKRISHNAN 1973). Such situations have been faced by BALAKRISHNAN and SANGHVI (1968) and SMOUSE and SPIELMAN (1977) when they attempted to estimate genetic distances between populations by statistics parallel to Mahalanobis-D^2 (MAHALANOBIS 1936) for multivariate data. ROBERTS and HIORNS (1962) also suggested a method of estimating genetic admixture in a hybrid population using allele frequency data that requires elimination of one allele of a multiallelic locus. Recently, this issue has resurfaced in the least-square estimation of admixture components in a hybrid population (LONG 1991). Since these investigators generally presented their estimating equations in terms of "shortened" vectors of allele frequencies (by deleting one allele from each locus) and the variance-covariance matrix of such "shortened" vectors of sampled allele frequencies, in general it is not obvious whether or not the resultant estimators depend upon the allele that is eliminated from the analysis. As a result, such methods are criticized on the ground of the subjectivity involved in selecting the allele to be eliminated (BALAKRISHNAN 1973) although in some applications algebraic verifications are given to show that any allele can be dropped without affecting the estimate (LONG 1991). The purpose of this communication is to show that by exploiting a well-known property of the variance-covariance matrix of the cell frequencies of a multinomial distribution (KURCZYNSKI 1970) a simple translation of the matrix estimators can be obtained, which indicates that even though the formal representation requires deleting one allele, the computational equation truly needs the frequencies of all alleles. Therefore, such estimators are functions of the full array of allele frequencies.

This simple exercise has at least three implications. First, it shows that the resultant estimators can be computed by algebraic operations involving all allele frequencies (which consequently results in numerically more accurate estimates, because matrix inversions generally introduce round off errors, which can be substantial particularly when the array size is large). Second, analytical relationships between different estimators of genetic parameters (e.g., distance, fixation indices, or admixture components) can be studied with greater ease with such representations (see *e.g.*, CHAKRABORTY and RAO 1991). Finally, genetic polymorphisms detected by DNA markers such as the variable number of tandem repeat (VNTR) loci often involve allele numbers (per locus) exceeding several dozen, and treating them with matrix operations requires a large array size, and even with that numerical inaccuracies cannot be avoided. On the contrary, algebraic expressions such as the ones presented here should make the analysis of such allele frequency data easier and certainly numerically more accurate.

Although the technique suggested here has wider applications, I consider only two specific estimation problems.

Genetic distance with multiple alleles: Denoting p_{ijk} as the frequency of the kth allele ($k = 1, 2, \ldots, s_j + 1$) of the jth locus ($j = 1, 2, \ldots, r$) in the ith population ($i = 1, 2$); estimated from $n_{ij}/2$ individuals sampled from the ith population for the jth locus, BALAKRISHNAN and SANGHVI (1968) suggested an estimator of the genetic distance between the two populations, given by

$$G_c^2 = \sum_{j=1}^{r} \mathbf{d}_j' \mathbf{C}_j^{-1} \mathbf{d}_j, \tag{1}$$

where $\mathbf{d}_j$ is a column vector of dimension s_j (one less than the number of segregating alleles at the jth locus, $s_j + 1$) whose kth element is $d_{jk} = p_{1jk} - p_{2jk}$, and $\mathbf{C}_j$ is a square matrix of size $s_j \times s_j$ whose elements are

$$C_{jkl} = \begin{cases} p_{jk}(1 - p_{jk}), & \text{for } k = l, \\ -p_{jk}p_{jl}, & \text{for } k \neq l \end{cases} \tag{2}$$

for $k, l = 1, 2, \ldots, s_j$; in which p_{jk} is the average of the kth allele frequency at the jth locus across populations; *i.e.*,

$$p_{jk} = \sum_i n_{ij}p_{ijk} / \sum_i n_{ij}. \tag{3}$$

Obviously, the quadratic form of equation (1) is the analog of Mahalanobis-D^2 (MAHALANOBIS 1936) since $\mathbf{C}_j$, given by (2), is the common dispersion matrix of the "shortened" vector of allele frequencies, estimated from the average allele frequencies across populations. Equation 1 may be written in the algebraic form

$$G_c^2 = \sum_{j=1}^{r} \sum_{k=1}^{s_j} \sum_{l=1}^{s_j} (p_{1jk} - p_{2jk}) C_j^{kl} (p_{1jl} - p_{2jl}), \quad (4)$$

where C_j^{kl} is the (k,l)th element of the $\mathbf{C}_j^{-1}$ matrix.

In order to show that G_c^2 is dependent on all allele frequencies, KURCZYNSKI (1970) noted that the inverse of the matrix $\mathbf{C}_j$ (of Equation 2) has the form

$$C_j^{kl} = \begin{cases} p_{jk}^{-1} + p_{j,s_j+1}^{-1}, & \text{for } k = l, \\ p_{j,s_j+1}^{-1}, & \text{for } k \neq l, \end{cases} \quad (5)$$

for $k, l = 1, 2, \ldots, s_j$.

Inserting (5) in (4) and noting that

$$\sum_{k=1}^{s_j} (p_{1jk} - p_{2jk})^2 + \sum_{k \neq j=1}^{s_j} \sum (p_{1jk} - p_{2jk})(p_{1jl} - p_{2jl})$$

$$= \left[\sum_{k=1}^{s_j} (p_{1jk} - p_{2jk}) \right]^2 = (p_{1j,s_j+1} - p_{2j,s_j+1})^2,$$

we obtain

$$G_c^2 = \sum_{j=1}^{r} \sum_{k=1}^{s_j+1} (p_{1jk} - p_{2jk})^2 / p_{jk}, \quad (6)$$

which depends on frequencies of every segregating allele, irrespective of which alleles are being dropped in the definition of the $\mathbf{d}_j$-vectors or $\mathbf{C}_j$ matrices. Equation 6 not only shows the involvement of all allele frequencies in the estimation, but also it is numerically simpler to compute than Equation 4. Note that the above proof also applies to SMOUSE and WILLIAM's (1982) measure of disease-gene association, where such equivalence is stated without a formal derivation. Furthermore, it demonstrates that BALAKRISHNAN and SANGHVI's (1968) measure is equivalent to the original estimator of SANGHVI (1953), except a multiplication factor. In addition, the above derivation shows that the alternative two estimators (G_c^2 and G_r^2) proposed by BALAKRISHNAN and SANGHVI (1968) are mathematically identical.

Another advantage of the representation of Equation 6 is that it clearly shows how SANGHVI's estimator of genetic distance is related to others. For example, considering the allele frequencies at a single locus (say, the jth locus), BHATTACHARYYA (1946) defined a distance statistic, θ^2, between populations, which satisfies the relationship

$$\text{Cos } \theta = \sum_{k=1}^{s_j+1} \{p_{1jk} p_{2jk}\}^{1/2}, \quad (7)$$

which can be written as

$$\text{Cos } \theta = \frac{1}{2} \cdot \sum_{k=1}^{s_j+1} [(p_{1jk} + p_{2jk})^2 - (p_{1jk} - p_{2jk})^2]^{1/2}$$

$$= \frac{1}{2} \cdot \sum_{k=1}^{s_j+1} (p_{1jk} + p_{2jk}) \left[1 - \frac{(p_{1jk} - p_{2jk})^2}{(p_{1jk} + p_{2jk})^2} \right]^{1/2} \quad (8)$$

$$= 1 - \frac{1}{4} \cdot \sum_{k=1}^{s_j+1} \frac{(p_{1jk} - p_{2jk})^2}{p_{1jk} + p_{2jk}}.$$

However, since $\text{Cos } \theta \approx 1 - \theta^2/2$, for small θ, Equation 8 approximates to

$$\theta^2 \approx \frac{1}{2} \cdot \sum_{k=1}^{s_j+1} (p_{1jk} - p_{2jk})^2 / (p_{1jk} + p_{2jk}), \quad (9)$$

showing that for genetically close populations (*i.e.*, for small θ), SANGHVI's (1953) and BHATTACHARYYA's (1946) distance estimators are equivalent, barring a multiplication factor. Equivalence of Equations 9 and 6 with 4 further shows that they are analogs of Mahalanobis-D^2 for categorial data. Several other such equivalence relationships between various distance functions are discussed in CHAKRABORTY and RAO (1991) who utilize representations such as Equation 6.

The same logic provides a formal proof of the assertion that in the absence of disequilibria (WEI 1979), SMOUSE and SPIELMAN's (1977) multivariate distance function based on multiple-allele genotype score vectors reduces to the form of Equation 6.

Weighted least square estimate of admixture proportions: For a dihybrid population whose gene pool consists of a fraction M of genes from a parental population 1 and a fraction $(1 - M)$ from parental population 2, LONG (1991) recently suggested a weighted least square estimator of M, which in matrix notation takes the form

$$m_j = (\mathbf{x}_j' \mathbf{V}_j^{-1} \mathbf{x}_j)^{-1} \mathbf{x}_j' \mathbf{V}_j^{-1} \mathbf{y}_j, \quad (10)$$

where $\mathbf{x}_j$ and $\mathbf{y}_j$ are column vectors of dimension s_j (one less than the number of segregating alleles, s_{j+1}, at the jth locus), with their kth elements defined by $x_{jk} = p_{1jk} - p_{2jk}$ and $y_{jk} = p_{hjk} - p_{2jk}$, for $k = 1, 2, \ldots, s_j$, and $\mathbf{V}_j$ is a $s_j \times s_j$ matrix with elements

$$V_{jkl} = \begin{cases} E(p_{hjk}) \cdot [1 - E(p_{hjk})], & \text{for } k = l \\ -E(p_{hjk}) \cdot E(p_{hjl}), & \text{for } k \neq l \end{cases} \quad (11)$$

in which p_{ijk} is the frequency of the kth allele ($k = 1, 2, \ldots, s_{j+1}$) at the jth locus in the ith population ($i =$ 1 or 2 for the parental populations) and $E(p_{hjk})$ is the expected allele frequency in the admixed population under the admixture model.

The estimator m_j (Equation 10) is based on the "shortened" vectors of allele frequency differences (dropping the (s_{j+1})th allele). However, noting that the elements of the $\mathbf{V}_j^{-1}$ matrix are given by

$$V_j^{kl} = \begin{cases} 1/E(p_{hjk}) + 1/E(p_{hj,s_j+1}), & \text{for } k = l, \\ 1/E(p_{hj,s_j+1}), & \text{for } k \neq l, \end{cases} \quad (12)$$

for $k, l = 1, 2, \ldots, s_j$, LONG (1991) verified that the estimator m_j of Equation 10 is invariant of the allele dropped from the analysis. To show explicitly that Equation 10 does not depict that it depends on all allele frequencies, first note that

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{x}_j = \sum_{k=1}^{s_j}\sum_{l=1}^{s_j}(p_{1jk} - p_{2jk})V_j^{kl}(p_{1jk} - p_{2jk}), \quad (13a)$$

and

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{y}_j = \sum_{k=1}^{s_j}\sum_{l=1}^{s_j}(p_{1jk} - p_{2jk})V_j^{kl}(p_{hjk} - p_{2jk}). \quad (13b)$$

Invoking (12) in (13a) and (13b), and noting that

$$p_{1j,s_j+1} - p_{2j,s_j+1} = -\sum_{i=1}^{s_j}(p_{ijk} - p_{2jk}), \quad (14a)$$

and

$$p_{hj,s_j+1} - p_{2j,s_j+1} = -\sum_{i=1}^{s_j}(p_{hjk} - p_{2jk}), \quad (14b)$$

we can rewrite (13a) and (13b) as

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{x}_j = \sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk}), \quad (15a)$$

and

$$\mathbf{x}_j'\mathbf{V}_j^{-1}\mathbf{Y}_j = \sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk}), \quad (15b)$$

so that Equation 10 becomes

$$m_j = \frac{\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk})}{\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk})}, \quad (16)$$

which is an equation of scalers. Expressed in this fashion, m_j involves each of the s_{j+j} segregating allele frequencies of both parental populations and the admixed one.

This representation (Equation 16) of the weighted least squares (WLS) estimator of LONG (1991) further shows that m_j (the WLS estimator) is identical to the classical BERNSTEIN (1931) estimator of admixture proportion for a bi-allelic locus, noted in LONG and SMOUSE (1983). With the notation p_{ij} and $q_{ij} = (1 - p_{ij})$ of the two allele frequencies at a locus the numerator and denominator of Equation 16 become

$$\frac{(p_{1j} - p_{2j})(p_{hj} - p_{2j})}{E(p_{hj})} + \frac{(q_{1j} - q_{2j})(q_{hj} - q_{2j})}{E(q_{hj})} = \frac{(p_{1j} - p_{2j})(p_{hj} - p_{2j})}{E(p_{hj})\cdot E(q_{hj})},$$

and

$$\frac{(p_{1j} - p_{2j})^2}{E(p_{hj})} + \frac{(q_{1j} - q_{2j})^2}{E(q_{hj})} = \frac{(p_{1j} - p_{2j})^2}{E(p_{hj})\cdot E(q_{hj})},$$

so that the cancellation of their common denominators results in

$$m_j = (p_{hj} - p_{2j})/(p_{1j} - p_{2j}) = (q_{hj} - q_{2j})/(q_{1j} - q_{2j}),$$

establishing the identity of the WLS and Bernstein's estimators for bi-allelic loci.

The combined estimator for allele frequency data on r loci, based on LONG's (1991) method, becomes

$$m = \frac{\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk})}{\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk})} \quad (17)$$

which avoids matrix manipulations of even bigger dimension.

The sampling error of m also has a corresponding scalar form. In LONG's notation, the sampling variance is

$$V(m) = \text{MSE}\cdot(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1}, \quad (18)$$

where the mean square error (MSE) of the admixture model is

$$\text{MSE} = (\mathbf{y} - m\mathbf{x})'\mathbf{V}^{-1}(\mathbf{y} - m\mathbf{x})/\sum_{j=1}^{r} s_j. \quad (19)$$

Invoking (12) in these quadratic forms, and using the identities (14a) and (14b), we can rewrite (19) as

$$\text{MSE} = \frac{\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}[(p_{hjk} - p_{2jk}) - m(p_{1jk} - p_{2jk})]^2/E(p_{hjk})}{\sum_{j=1}^{r} s_j} \quad (20)$$

which yields the sampling variance of m,

$$V(m) = \frac{\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}[(p_{hjk} - p_{2jk}) - m(p_{1jk} - p_{2jk})]^2/E(p_{hjk})}{\left[\sum_{j=1}^{r} s_j\right]\cdot\left[\sum_{j=1}^{r}\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk})\right]}. \quad (21)$$

The variance of the admixture estimate based on the jth locus data is

$$V(m_j) = \text{MSE}\cdot\left[\sum_{k=1}^{s_j+1}(p_{1jk} - p_{2jk})^2/E(p_{hjk})\right]^{-1}, \quad (22)$$

in which the expression (20) is used for evaluating the MSE.

In addition to the demonstration that Equations 16 and 22, or 17 and 21 involve all allele frequencies

from each population, their computational simplicity remain unaltered even if the sample sizes for different loci are different. Since the V matrix refers to the expected allele frequencies in the admixed population, all terms of the summation over j will have to be weighted by n_{hj}, the number of genes sampled for the jth locus from the admixed population. For example, the combined estimator becomes

$$m = \frac{\sum_{j=1}^{r} n_{hj} \sum_{k=1}^{s_j+1} (p_{1jk} - p_{2jk})(p_{hjk} - p_{2jk})/E(p_{hjk})}{\sum_{j=1}^{r} n_{hj} \sum_{k=1}^{s_j+1} (p_{1jk} - p_{2jk})^2/E(p_{hjk})}. \quad (23)$$

The corresponding changes in its sampling variance are also similar.

Other population genetic applications of algebraic representations of quadratic forms involving inverses of multinomial variance-covariance matrices include the estimation of Wright's fixation indices in the context of analysis of population structure. Using approaches similar to the above, LONG's (1986) multiallelic generalizations of COCKERHAM's (1969, 1973) variance-covariance estimators of the fixation indices can also be reduced to algebraic forms, which indicate their relationship with some existing estimators suggested earlier (see *e.g.*, LI and HORVITZ 1953; CURIE-COHEN 1982; ROBERTSON and HILL 1984; WEIR and COCKERHAM 1984).

To close this commentary, I must mention that the algebraic reductions of the matrix estimators such as the ones mentioned above are not meant to denigrate the utility of matrix notations in population genetics. Matrix representations of functions of allele frequencies at multiallelic loci have their importance and place that cannot be denied. They serve the purpose of establishing the basis of the method of estimation that is not always obvious in the closed form algebraic expression. In some instances matrix estimators are unavoidable. For example, the estimator of admixture contributions from multiple (more than two) ancestral populations is straightforward in matrix notation (ELSTON 1971; CHAKRABORTY 1986) and the incorporation of all orders of disequilibria (WEIR 1979) in estimating parameters of population structure and genetic distance analyses requires matrix notations, although nearly equivalent algebraic forms are also available (see *e.g.*, WEIR and COCKERHAM 1984). Nevertheless, the primary intent of this note has been to demonstrate that the principle that these are independent of which allele is dropped from the analysis.

This work was supported by U.S. Public Health Service research grants GM 41399 from the National Institutes of Health and 90-IJ-CX-0038 from the National Institute of Justice. I thank P. E. SMOUSE, B. S. WEIR and an anonymous reviewer for their comments and suggestions on the work.

RANAJIT CHAKRABORTY
Center for Demographic and
Population Genetics
University of Texas Graduate School
of Biomedical Sciences
P. O. Box 20334
Houston, Texas 77225

LITERATURE CITED

BALAKRISHNAN, V., 1973 Use of distance in hybrid analysis, pp. 268–273 in *Genetic Structure of Populations*, edited by N. E. MORTON, University of Hawaii Press, Honolulu.

BALAKRISHNAN, V., and L. D. SANGHVI, 1968 Distances between populations on the basis of attribute data. Biometrics **24:** 859–865.

BERNSTEIN, F., 1931 Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung. Comitato Italiano per lo Studio dei Problemi della Popülazione. Instituto Poligrafico dello Stato. Rome.

BHATTACHARYYA, A., 1946 On a measure of divergence between two multinomial populations. Sankhya **7:** 401–406.

CHAKRABORTY, R., 1986 Gene admixture in human populations: models and predictions. Yearb. Phys. Anthropol. **29:** 1–43.

CHAKRABORTY, R., and C. R. RAO, 1991 Measurement of genetic variation for evolutionary studies, in *Handbook of Statistics*, Vol. 8, edited by C. R. RAO and R. CHAKRABORTY. Elsevier, Amsterdam (in press).

COCKERHAM, C. C., 1969 Variance of gene frequencies. Evolution **23:** 72–84.

COCKERHAM, C. C., 1973 Analyses of gene frequencies. Genetics **74:** 679–700.

CURIE-COHEN, M., 1982 Estimates of inbreeding in a natural population: a comparison of sampling properties. Genetics **100:** 339–358.

ELSTON, R. C., 1971 The estimation of admixture in racial hybrids. Ann. Hum. Genet. **35:** 9–17.

KURCZYNSKI, T. W., 1970 Generalized distance and discrete variables. Biometrics **26:** 525–534.

LI, C. C., and D. G. HORVITZ, 1953 Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet. **5:** 107–117.

LONG, J. C., 1986 The allelic correlation structure of Gainj- and Kalam-speaking people. I. the estimation and interpretation of Wright's *F*-statistics. Genetics **112:** 629–647.

LONG, J. C., 1991 The genetic structure of admixed populations. Genetics **127:** 417–428.

LONG, J. C., and P. E. SMOUSE, 1983 Intertribal geneflow between the Ye'cuana and Yanamamö: genetic analysis of an admixed village. Am. J. Phys. Anthropol. **61:** 411–422.

MAHALANOBIS, P. C., 1936 On the generalized distance in statistics. Proc. Natl. Inst. Sci. India **12:** 49–55.

ROBERTS, D. F., and R. W. HIORNS, 1962 The dynamics of racial admixture. Am. J. Hum. Genet. **14:** 261–277.

ROBERTSON, A., and W. G. HILL, 1984 Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics **107:** 703–718.

SANGHVI, L. D., 1953 Comparison of genetical and morphological methods for a study of biological differences. Am. J. Phys. Anthropol. **11:** 385–404.

SMOUSE, P. E., and R. S. SPIELMAN, 1977 How allocation of individuals depends on genetic differences among populations. Excerpta Med. Congr. Ser. No. **411:** 255–260.

SMOUSE, P. E., and R. C. WILLIAMS, 1982 Multivariate analysis of HLA-disease associations. Biometrics **38:** 757–768.

WEIR, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **35:** 235–254.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

Sample Size Requirements for Addressing the Population Genetic Issues of Forensic Use of DNA Typing

RANAJIT CHAKRABORTY[1]

Abstract DNA typing offers a unique opportunity to identify individuals for medical and forensic purposes. Probabilistic inference regarding the chance occurrence of a match between the DNA type of an evidentiary sample and that of an accused suspect, however, requires reliable estimation of genotype and allele frequencies in the population. Although population-based data on DNA typing at several hypervariable loci are being accumulated at various laboratories, a rigorous treatment of the sample size needed for such purposes has not been made from population genetic considerations. It is shown here that the loci that are potentially most useful for forensic identification of individuals have the intrinsic property that they involve a large number of segregating alleles, and a great majority of these alleles are rare. As a consequence, because of the large number of possible genotypes at the hypervariable loci that offer the maximum potential for individualization, the sample size needed to observe all possible genotypes in a sample is large. In fact, the size is so large that even if such a huge number of individuals could be sampled, it could not be guaranteed that such a sample was drawn from a single homogeneous population. Therefore adequate estimation of genotypic probabilities must be based on allele frequencies, and the sample size needed to represent all possible alleles is far more reasonable. Further economization of sample size is possible if one wants to have representation of only the frequent alleles in the sample, so that the rare allele frequencies can be approximated by an upper bound for forensic applications.

It is now well known that interspersed in the human genome are numerous DNA regions that have genetic variation whose magnitude is much larger than that coded by the traditional serological loci (Jeffreys et al. 1985a,b; Nakamura et al. 1987). Such genetic hypervariability can

[1]Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, PO Box 20334, Houston, TX 77225.

Human Biology, April 1992, Vol. 64, No. 2, pp. 141–159.

KEY WORDS: VNTR POLYMORPHISM, SAMPLE SIZE, INFINITE ALLELE MODEL, DNA TYPING, FORENSIC GENETICS

be used profitably to identify specific individuals by DNA typing at such regions of the genome. In recent years, therefore, civil and criminal courts in the United States and Europe have been admitting DNA typing evidence in resolving legal controversies involved in both paternity disputes and criminal cases [see, for example, Ballantyne et al. (1989)]. Because DNA typing technology is relatively new and because population-based surveys on variability at such genetic loci are less abundant compared to the classical serological markers, concerns have been raised about their general applicability for legal purposes (Lander 1989a,b, 1991; Cohen 1990).

The criticisms with regard to admitting DNA typing evidence in legal cases include the inappropriateness of using population data from small samples (Lander 1989b). It is therefore necessary to determine how large a sample must be for the statistical analysis of DNA typing data to be regarded as appropriate for legal applications. Note that for forensic use DNA typing data can be represented in terms of multinomial distributions with a large number of possible classes [see, for example, Budowle, Giusti et al. (1991)] whose theory is completely characterized (Rao 1957, 1958; Johnson and Kotz 1969). However, there is no systematic discussion on this subject in the context of allele and genotype frequencies generally observed at the hypervariable loci that currently are being used for forensic cases.

My purpose here is to address this issue by asking what sample sizes are adequate for conservative evaluation of genotype or allele frequencies. Based on the population genetic characteristics of the hypervariable loci, I show that the large heterozygosities at such loci necessarily imply that the expected number of alleles at each of these loci is generally quite large (often larger than 50) and that there is a predominance of rare alleles (i.e., alleles that occur in frequencies as small as 0.01) at such loci. Furthermore, the total number of alleles and the number of rare alleles are increasing functions of sample size. Consequently, if we want to determine a minimum sample size based on the criterion that all possible genotypes must be represented in the sample, the needed sample size would be so large that it would be impractical to test that many individuals from a single homogeneous population. It therefore becomes necessary to evaluate genotype frequencies from allele frequency data, and the sample size requirements can be derived by estimating allele frequencies from the sample. Based on representation of all alleles in the sample, feasible sample sizes can be obtained from well-defined homogeneous populations. However, because many of the alleles are characteristically of rare frequency (as defined before), sample size requirements still can be severe. It might be more economical to ask for a sample size whereby the frequencies of more frequent alleles are estimated with reliable precision and to determine a threshold for the

frequencies of rare alleles for which an upper bound can be prescribed based on such threshold values. This procedure should yield a conservative estimate (biased in the upward direction) of probabilities of obtaining a match for cases involving rare genotypes, even when the exact evaluation of chance occurrence of a match becomes impossible based on the sampled allele frequencies.

Throughout this paper I call the class of loci that are shown to be useful for forensic purposes VNTR (variable number of tandem repeat) loci, following Nakamura et al. (1987), although various other names for such loci have been proposed depending on the core motifs of their nucleotide sequences [see, for example, Edwards et al. (1991)]. I also assume that the allelic distinctions at such loci are made without ambiguity, so that the different alleles are discrete and no measurement error is involved in the size classification of alleles. When allele sizes are quasi-continuous in a population, the present theory can be applied with suitable binning of alleles [e.g., Budowle, Giusti et al. (1991)] without any major changes in the qualitative conclusions of the present results. The theory also assumes that VNTR loci have an autosomal codominant mode of transmission and that the genotype frequencies satisfy Hardy-Weinberg equilibrium (HWE) expectations. Note that, in the absence of measurement error, data on such loci collected from well-defined populations generally show that this assumption is appropriate [see, for example, Boerwinkle et al. (1989), Ludwig et al. (1989), Budowle, Chakraborty et al. (1991), Chakraborty, Fornage et al. (1991), Deka et al. (1991), and Edwards et al. (1991)]. Polymorphisms at VNTR loci with discretized allelic distinctions also have been shown to follow the predictions of the neutral mutation model of the infinite allele model [e.g., Jeffreys et al. (1988), Budowle, Chakraborty et al. (1991), Chakraborty, Fornage et al. (1991), Chakraborty and Daiger (1991), Deka et al. (1991), and Edwards et al. (1991)], which is assumed to generate the expected number of alleles at VNTR loci.

Throughout this article the sample size is defined as the number of individuals sampled from the population (designated by n) so that, whenever sample size is considered with regard to the number of alleles sampled, it is equated to $2n$ (because of autosomal inheritance of the loci and diploidy of the human genome).

Theory and Results

Expected Total Number of Alleles and Number of Rare Alleles in a Sample from an Equilibrium Population. Although the high degree of polymorphism at VNTR loci has been noted in almost all human populations—and this is characterized in terms of high heterozygosity at

these loci and by a large number of alleles—it is not possible to determine the exact number of alleles that can occur at VNTR loci in any population. This is because the allelic possibilities are truly infinite for such loci, and hence predictions can be made only with regard to the sampling distribution of the number of alleles found in any sample of alleles drawn from a population. By assuming selective neutrality of VNTR alleles, one can employ the sampling theory of selectively neutral alleles (Ewens 1972; Chakraborty and Griffiths 1982) to determine the expected total number of alleles and the expected number of alleles for any specific allele frequency class. In a sample of $2n$ alleles (n individuals) randomly drawn from a population that is at mutation-drift equilibrium, Ewens (1972) showed that the expected total number of alleles can be expressed in terms of an unknown parameter (θ) and the sample size (n) by the equation

$$E(k) = \theta \sum_{i=0}^{2n-1} (\theta + i)^{-1}, \tag{1}$$

where $\theta = 4Nv$, in which N is the effective size of the population and v is the mutation rate at the locus per generation. Chakraborty (1981) and Chakraborty and Griffiths (1982) further showed that the expected number of alleles whose frequency lies in the range p_1 to p_2 ($0 \leq p_1 < p_2 \leq 1$) in such a sample is given by

$$E[k(p_1, p_2)] = \sum_{i=[2np_1]+1}^{[2np_2]} \frac{\theta}{i} \frac{n!}{(n-i)!} \frac{\Gamma(n+\theta-i)}{\Gamma(n+\theta)}, \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function and $[n]$ is the largest integer contained in n. Equation (2) immediately proscribes the expected number of rare alleles in a sample by substituting $p_1 = 0$ and $p_2 = p$ for any small p (say, $p = 0.001$, 0.01, or 0.05). Therefore the expected number of rare alleles in a sample of n individuals becomes

$$E[k(p)] = \sum_{i=1}^{[2np]} \frac{\theta}{i} \frac{n!}{(n-i)!} \frac{\Gamma(n+\theta-i)}{\Gamma(n+\theta)}, \tag{3}$$

which can be approximated with good precision (Chakraborty 1981) by

$$E[k(p)] \approx A\theta - B\theta^2, \tag{4}$$

where

$$A = \sum_{i=1}^{[2np]} (1/i) \quad \text{and} \quad B = \sum_{i=1}^{[2np]} (2n-i)^{-1} \tag{5}$$

are constants depending on the sample size n and the definition of rare alleles ($p = 0.001$, 0.01, 0.05, or any arbitrary small number). It should

Table 1. Expected Total Number of Alleles in a Sample of n Individuals for a Given Level of Heterozygosity

	Heterozygosity				
n	*0.20*	*0.50*	*0.75*	*0.90*	*0.95*
50	2.2	5.2	11.1	22.9	35.3
100	2.4	5.9	13.2	28.8	46.9
500	2.8	7.5	18.0	43.0	76.2
1,000	3.0	8.2	20.0	49.2	89.2
10,000	3.5	10.5	26.9	69.9	132.7

be noted that both k (the total number of alleles) and $k(p)$ (the number of rare alleles) in a sample are random variables and hence do not have fixed values, but their sampling distributions are known. Ewens (1972) gave the exact sampling distribution for k [which is quite complex but can be numerically evaluated; see Stewart's algorithm in the appendix of Fuerst et al. (1977)], and Chakraborty and Griffiths (1982) showed that for small values of p (such as the ones noted here), the variable $k(p)$ follows a Poisson distribution so that expression (3) or (4) also gives the variance of the number of rare alleles in a sample.

Equations (1) and (4) can be used to evaluate the expected numbers of rare alleles for any selectively neutral locus, for which we must know the composite parameter θ. One way of estimating this parameter is from the level of heterozygosity at the locus, which is relatively less sensitive to sample size. Denote the sample value of heterozygosity by H. It is known that under the assumption of equilibrium the expectation of H is $\theta/(1 + \theta)$ so that the moment estimator of θ is given by $H/(1 - H)$. The justifications for this moment estimator of θ in the present context are given by Chakraborty (1990a,b) and Chakraborty and Schwartz (1990). Therefore Eqs. (1) and (4) can be used to predict roughly how many alleles are expected at any selectively neutral locus if the sample is drawn from an equilibrium population and how many rare alleles there may be in such a sample.

Table 1 presents the expected total number of alleles for some representative heterozygosity values that are generally seen at VNTR loci for several sample sizes (number of individuals). It is clear from this table that the observed number of alleles is an increasing function of heterozygosity and of sample size, and, in particular, when the heterozygosity is high (say, 95%), the number of alleles can easily exceed 40 when more than 100 individuals are sampled from the population. Therefore a large number of alleles at any VNTR locus is an intrinsic property of such polymorphisms.

Table 2. Expected Number of Rare Alleles in a Sample of n Individuals for Selected Levels of Heterozygosity at a Locus

		n				
H	P	*50*	*100*	*500*	*1,000*	*10,000*
0.20	0.001	–[a]	–	0.3	0.4	0.9
	0.01	0.3	0.4	0.7	0.9	1.5
	0.05	0.6	0.7	1.1	1.3	1.9
0.50	0.001	–	–	1.0	1.5	3.6
	0.01	1.0	1.5	2.9	3.6	5.9
	0.05	2.2	2.9	4.5	5.1	7.4
0.75	0.001	–	–	3.0	4.5	10.8
	0.01	2.9	4.4	8.7	10.7	17.5
	0.05	6.4	8.3	13.0	15.1	22.0
0.90	0.001	–	–	8.9	13.4	32.3
	0.01	8.2	12.7	25.6	31.6	52.1
	0.05	16.4	22.2	36.3	42.5	63.2
0.95	0.001	–	–	18.6	28.1	68.0
	0.01	15.4	24.9	52.0	64.7	108.1
	0.05	24.8	37.1	67.0	80.0	123.7

a. Dash denotes cases where this definition of rare alleles is inappropriate.

Such a large number of alleles is also observed in many empirical surveys in accordance with this theory. For example, Boerwinkle et al. (1989) observed 12 alleles at the ApoB VNTR locus in a French population of 125 unrelated individuals, and Ludwig et al. (1989) found 14 alleles at the same locus in a survey of 318 US whites. Because their observed heterozygosity values were 75% and 78%, respectively, by using Eq. (1), we get expectations for the total number of alleles for these sample sizes of 13.8 and 18.9, respectively, which are in fair agreement with the observations [particularly because k, the total number of alleles, also has a large variance; see Ewens (1972)]. An even higher number of alleles was observed in the Utah white population at several VNTR loci (Odelberg et al. 1989). This can be explained by the high heterozygosity (75–95%) at the loci surveyed. The largest number of alleles observed by Odelberg et al. (1989) was 67 at the D2S44 locus, which was reported as having a heterozygosity of 95%. All these results indicate that high heterozygosity at VNTR loci necessarily leads to a large number of segregating alleles, and this is consistent with the expectations of the pattern of polymorphism at these loci.

Table 2 shows another important feature of such polymorphisms. Using Eq. (4), one can compute the expected number of rare alleles for any given level of heterozygosity in a sample of size n. In the computations presented in Table 2, I used three definitions of rare alleles (p

= 0.001, 0.01, and 0.05), noting that, when less than 50 individuals were surveyed, rare alleles could not be defined with the criterion $p < 0.01$. The general implications of the results shown in this table are: (1) Many of the segregating alleles at VNTR loci are rare; (2) like the total number of alleles, the number of rare alleles is also an increasing function of heterozygosity and sample size; and (3) the proportion of rare alleles increases with increasing sample size and heterozygosity. For example, when $H = 90\%$, in a sample of 50 individuals 16.4 of the 23 alleles will have a frequency below 5%, whereas in a sample of 10,000 individuals 63 of the 70 alleles at the locus are expected to have an allele frequency below 5%. For higher heterozygosities the rare alleles will constitute an even larger fraction of the total number of alleles. Analysis of the empirical survey data [e.g., Odelberg et al. (1989)] indicates that such theoretical expectations are congruent with the observations. For example, at the D2S44 locus Odelberg et al. (1989) observed 37 of the 67 alleles with a frequency less than 1% ($p = 0.01$), although 42.5 ± 6.5 are expected to be in this class using Eq. (4) and assuming a Poisson distribution of rare alleles.

The features of polymorphism seen in the computations of Tables 1 and 2 have important implications for sample size requirements in VNTR surveys. These results indicate that, given large heterozygosities, the number of possible genotypes at VNTR loci often can be large. For example, when $H = 0.95$, we can easily expect as many as 90 alleles (for $n = 1000$), and the number of possible genotypes at a VNTR locus will be 4095. Furthermore, because at a VNTR locus there can be 80 alleles with frequency 5% or less, 3240 of these genotypes should have a frequency less than 0.5% should the population be in Hardy-Weinberg equilibrium with respect to the genotype frequencies. Obviously, we would like to know what sample size will be required to represent all these genotypes in the sample if we are to use genotype data to estimate all genotype frequencies directly.

Probability of Observing All Possible Genotypes in a Sample of Fixed Size and Minimum Sample Size Requirement for Representation of All Genotypes in a Sample. The minimum sample size requirement can be addressed in two alternative ways. First, we can evaluate the probability that all possible genotypes are represented in a sample of a given size. Elsewhere I have shown that this probability can be evaluated only when the number of alleles and their frequencies in the population are known (Chakraborty 1991). Consider a locus for which there are k segregating alleles, where $p_1, p_2, \ldots, p_k$ (the p-vector) represent the true allele frequencies. If we assume that in the population the genotype frequencies follow Hardy-Weinberg expectations, the vector of genotype

frequencies is of dimension $K = k(k + 1)/2$, and it can be represented by $q_1, q_2, \ldots, q_K$, where the q_i are either of the form p_i^2 or $2p_ip_j$ depending on the specific genotypes. In a sample of n individuals scored for such a locus, the probability that all genotypes are represented is given by [see Chakraborty (1991) for derivation]

$$P = 1 - \sum_{r=1}^{K} (-1)^{r-1} S_r, \tag{6}$$

where S_r is the probability that at least r of the K genotypes are not observed in the sample; that is,

$$S_1 = \sum_{i=1}^{K} (1 - q_i)^n, \quad S_2 = \sum_{i>j=1}^{K}\sum (1 - q_i - q_j)^n, \text{ etc.} \tag{7}$$

Although Eq. (6) can be numerically evaluated for any given values of k and p-vector, it involves tedious enumerations of a large number of summations, particularly for k as large as the ones noted earlier. However, Chakraborty (1991) derived the exact sampling distribution of the random variable X, the number of distinct genotypes observed in a sample for known values of k and p-vector, and this distribution also can be approximated by a normal distribution with mean μ and variance σ^2, where

$$\mu = \frac{k(k + 1)}{2} - S_1, \tag{8a}$$

$$\sigma^2 = S_1(1 - S_1) + 2S_2, \tag{8b}$$

and S_1 and S_2 are as defined in Eqs. (7).

Therefore the probability P of Eq. (6) can be approximated by

$$P \approx 1 - \Phi(S_1/\sigma), \tag{9}$$

where $\Phi(x)$ is the cumulative probability of a standard normal variate up to x.

Table 3 shows the result of such a computation for some short tandem repeat (STR) loci, where the number of alleles and their frequencies are taken from the survey of 40 or more unrelated white individuals reported by Edwards et al. (1991). It is clear from this table that, at these STR loci (where the number of alleles generally ranges from 6 to 17) and because many of the alleles are rare, all possible genotypes may not generally be observed even in an extremely large sample. For example, for the $(AGAT)_n$ (HUMHPRTB) locus, even in a sample of 1 million individuals, the probability of observing all possible genotypes is only 0.5. Therefore in any sample of a fixed size n the chance of observing all possible genotypes in general is very small.

Table 3. Probability of Observing All Possible Genotypes in Samples of Fixed Size for Some Representative Short Tandem Repeat Loci

n^a	*Loci*[b]				
	$(AGAT)_n$	$(AATG)_n$	$(ACAG)_n$	$(AAT)_n$	$(AGC)_n$
50	2.06×10^{-37}	4.83×10^{-7}	0.025	2.72×10^{-14}	(c)
100	2.51×10^{-32}	7.05×10^{-5}	0.106	3.37×10^{-12}	2.97×10^{-14}
200	1.03×10^{-26}	2.13×10^{-3}	0.243	1.97×10^{-9}	8.45×10^{-9}
500	1.26×10^{-19}	0.019	0.414	1.12×10^{-5}	1.03×10^{-4}
1,000	1.67×10^{-15}	0.029	0.482	3.09×10^{-4}	3.96×10^{-3}
10,000	1.04×10^{-5}	0.317	1.0	0.014	0.364
100,000	0.174	0.500	1.0	0.204	1.0
1,000,000	0.500	1.0	1.0	0.496	1.0

a. Number of individuals typed.
b. Data on allele frequencies and number of alleles on these short tandem repeat loci are taken from the white sample examined by Edwards et al. (1991).
c. Sample size is smaller than the number of possible genotypes giving a zero probability for observing all possible genotypes in the sample.

These computations, however, do not prescribe any well-defined minimum sample size requirement for observing all possible genotypes; nor can the minimum sample size be evaluated analytically by any direct method. A crude conservative sample size estimate can be obtained by the following alternative method. By using Eq. (6), the probability P of observing all possible genotypes in a sample of n individuals satisfies the inequality

$$P \geq 1 - \sum_{i=1}^{k} (1 - p_i^2)^n - \sum_{i>j=1}^{k}\sum^{k} (1 - 2p_ip_j)^n, \tag{10}$$

in which the right-hand side is at a maximum when all allele frequencies are equal, that is, when $p_i = 1/k$ for all i. Therefore a conservative estimate of the minimum sample size requirement for ensuring that all genotypes are represented in the sample with confidence $(1 - \alpha)$ is given by the inequality

$$1 - k(1 - k^{-2})^n - \tfrac{1}{2}k(k-1)(1 - 2k^{-2})^n \geq 1 - \alpha, \tag{11}$$

in which the substitutions $(1 - k^{-2})^n \approx e^{-n/k}$ and $(1 - 2k^{-2})^n \approx e^{-2n/k}$ yield

$$n \geq -k^2 \ln\left[\frac{\sqrt{k^2 + 2\alpha k(k-1)} - k}{k(k-1)}\right]. \tag{12}$$

Table 4. Conservative Estimates of the Number of Individuals Needed to Represent All Possible Genotypes in a Sample for a k-Allelic Codominant Locus

	Minimum Sample Size[a] Needed for:				
k	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.25$
5	213	156	116	99	77
10	921	691	532	465	379
15	2,164	1,647	1,288	1,137	944
20	3,962	3,043	2,406	2,137	1,794
25	6,330	4,893	3,899	3,479	2,943
30	9,279	7,210	5,778	5,174	4,402
40	16,955	13,278	10,733	9,659	8,287
50	27,051	21,305	17,329	15,651	13,507
100	115,134	92,153	76,248	69,539	60,970

a. Sample size in these computations refers to the number of individuals to be typed for each k-allelic codominant locus. The values of $\alpha = 0.001$, 0.01, 0.05, 0.10, and 0.25 represent 99.9%, 99%, 95%, 90%, and 75% confidence, respectively, of being assured that all possible genotypes are represented in the sample.

When the allele frequencies are not equal (as is the case for all VNTR loci), sample size requirements for representation of all possible genotypes can far exceed the bound prescribed by expression (12), and numerical evaluation of this expression is instructive enough to show that it is not feasible to collect samples large enough to encompass all possible genotypes for any VNTR locus in any population. Table 4 presents numerical evaluations of expression (12) for some representative values of k, the number of alleles that are in the general range seen in Table 1. It is clear that, even with this conservative minimum sample size estimate, a sample of 15,651 individuals is required to encompass all possible genotypes with 90% confidence if there are 50 alleles segregating at a VNTR locus. This is generally too much to ask in a survey study, and even if such a large sample could be collected, there is no guarantee that the individuals truly came from a single homogeneous population.

The analysis clearly establishes that, if we are to use the observed relative frequencies of all genotypes as the estimates of genotypic probabilities in the population, a sample of adequate size cannot be collected because from any reasonable homogeneous population this large a sample cannot be gathered. An appropriate alternative way to estimate the genotype frequencies is therefore to use the estimate of allele frequencies and to invoke assumptions through which genotype probabilities can be derived based on allele frequency estimates (such as the Hardy-Weinberg equilibrium assumption).

Sample Size Requirement Based on Allele Frequencies. Having shown that the only practical and reliable way to estimate genotypic

Table 5. Minimum Number of Individuals Needed to Represent All Alleles in a Sample for a k-Allelic Codominant Locus

k	*Minimum Sample Size[a] Needed for:*				
	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.25$
5	22	16	12	10	8
10	46	35	27	23	19
15	72	55	43	38	31
20	99	76	60	53	44
25	127	98	78	69	58
30	155	120	96	86	72
40	212	166	134	120	102
50	271	213	171	156	133
100	576	461	380	346	300

a. Number of individuals to be typed.

probabilities at VNTR loci is from the allele frequencies, I can now turn to the evaluation of the minimum sample size requirement based on allele frequencies. Again the logic of deriving expression (12) can be used to determine a crude conservative estimate of minimum sample size. For a locus with k segregating alleles whose frequencies in a population are p_1, p_2, etc., the probability that all alleles are represented in a sample of n individuals should exceed the quantity

$$1 - \sum_{i=1}^{k} (1 - p_i)^{2n}. \tag{13}$$

In order for expression (13) to exceed the level of confidence $(1 - \alpha)$, we must ensure that

$$1 - k(1 - k^{-1})^{2n} \geq 1 - \alpha, \tag{14}$$

or

$$n \geq \tfrac{1}{2} \ln(\alpha/k)/\ln(1 - k^{-1}). \tag{15}$$

Admittedly, this bound of minimum sample size is too crude because, when the allele frequencies are not equal, far larger sample sizes are needed for all alleles to be represented in a sample. Nevertheless, Table 5 shows that use of expression (15) leads to sample size estimates that are feasible to collect from any well-defined population. For example, to ensure that all 50 equifrequent alleles are represented in a sample with 95% confidence ($\alpha = 0.05$), we need to type 171 individuals from the population. In practice, however, the required sample size may be larger, because these computations do not represent the reality of the situation,

Table 6. Minimum Number of Individuals Needed to Have r Alleles with Frequency p or Above Represented in the Sample

		Minimum Sample Size Needed for:				
r	p	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.25$
1	0.001	3452	2302	1498	1151	693
	0.01	344	230	150	115	69
	0.05	68	45	30	23	14
2	0.001	3798	2647	1838	1485	1005
	0.01	379	264	183	148	100
	0.05	74	52	36	29	20
5	0.001	4257	3104	2292	1935	1442
	0.01	424	309	229	193	144
	0.05	84	61	45	38	29
10	0.001	4603	3450	2637	2278	1781
	0.01	459	344	263	227	178
	0.05	90	68	52	45	35

namely, the allele frequencies are not equal, and thus such a direct evaluation of minimum sample size is not possible.

Alternatively, we might ask what sample size would be required if we want to ensure that all alleles with frequencies above a certain small value will be represented in the sample with a proscribed level of confidence. Because the probability that an allele with frequency p remains unobserved in a sample of n individuals is given by $(1 - p)^{2n}$, if there are r alleles at a locus that have frequencies p or above in the population [reasonable values of r can be obtained from $k - k(p)$, from the first section of the previous analysis], in order for all these common alleles to be represented in the sample, we must have

$$[1 - (1 - p)^{2n}]^r \geq 1 - \alpha, \tag{16}$$

or

$$n \geq \ln[1 - (1 - \alpha)^{1/r}]/2 \ln(1 - p). \tag{17}$$

Table 6 presents sample size estimates based on this inequality. As seen in Tables 1 and 2, for most VNTR loci, even when the total number of alleles is large, the expected number of alleles having frequency p or above is generally below 10 for $p = 0.001$, 0.01, or 0.05. Therefore in Table 6 the minimum sample size requirement is presented for values of $r \leq 10$. It is clear from this table that, if we sacrifice the alleles of frequency below 0.01, a sample of 300 individuals will ensure that all common alleles (alleles with frequency greater than 1%) will be represented in a sample with at least 95% confidence. This is a much more feasible sampling strategy and should guarantee reliable estimation of

Table 7. Frequency of Alleles That Will Be Represented in a Sample of n Individuals with a Given Level of Confidence

	Allele Frequencies for:					
	$r = 1$		$r = 5$		$r = 10$	
n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
50	0.0450	0.0295	0.0602	0.0448	0.0667	0.0514
100	0.0228	0.0149	0.0306	0.0227	0.0339	0.0260
200	0.0114	0.0075	0.0154	0.0114	0.0171	0.0131
500	0.0046	0.0030	0.0062	0.0046	0.0069	0.0053
1000	0.0023	0.0015	0.0003	0.0002	0.0003	0.0003

the frequencies of common alleles in a population. Note that the sample size estimates of Table 6 are even more economical if we sacrifice all alleles having frequency 0.05 or less, in which case 50 individuals may be sufficient to ensure the presence of all common alleles in the sample with 95% confidence.

The inequality (17) can also be written in the form

$$p \geq 1 - [1 - (1 - \alpha)^{1/r}]^{1/2n}, \tag{18}$$

which can be used to examine which allele frequencies are reliably estimated in a survey of n individuals. This bound also proscribes a threshold value for the rare allele frequencies that would yield a conservative probability of a match in forensic cases involving previously unseen DNA types. Table 7 presents some representative values of such minimum bounds of allele frequencies. It shows that in the VNTR surveys involving 200 or more individuals the alleles with frequency above 1% are generally represented, and even when the sample size is 50, alleles with frequency above 5% should be encompassed in the sample.

Discussion and Conclusion

The final step in using DNA typing data in forensic applications consists in using estimates of specific genotype frequencies to determine how often by chance alone two biological specimens from two different individuals have identical DNA type results. Obviously, reliable conservative estimates of genotype probabilities are required for such a purpose, and population-based data must provide such estimates. Activities at various laboratories are currently geared toward providing such data. It is intuitively clear that a hypervariable locus that provides greater heterozygosity is also more efficient for resolving a forensic case, because the chance of a match by chance alone decreases as heterozygosity in-

creases. Therefore from a strategic point of view hypervariable loci that contain larger heterozygosities should be considered first for gathering population data. Of course, the cost efficiency and technical reproducibility of typing results also must be considered in selecting the loci that serve the purpose better.

Here, I first show that one of the intrinsic population genetic characteristics of VNTR polymorphisms is that VNTR loci generally contain a large number of segregating alleles whose exact number in any population is a random variable and hence is strictly unknown. The expected number of alleles, however, can be derived by assuming that the pattern of VNTR polymorphism follows the predictions of the infinite allele model of selectively neutral alleles. Validation of this assumption is provided by several recent articles [see, for example, Jeffreys et al. (1988), Budowle, Chakraborty, de Andrade et al. (1991), Chakraborty, Fornage et al. (1991), Edwards et al. (1991), and Deka et al. (1991)], particularly when the DNA typing protocol can discretize the allelic distinctions by techniques such as high-resolution Southern gel electrophoresis following polymerase chain reaction (PCR) techniques.

In view of the recent article by Jeffreys et al. (1990) that VNTR alleles of identical size may not always be iso-allelic at a molecular level and that generation of new alleles at VNTR loci may not exactly correspond to the infinite allele model, one might question the applicability of Ewens's sampling theory invoked in the present analysis. To this effect several comments are noteworthy. First, earlier studies in relation to protein variation have shown that in the presence of hidden variation (within allelic classes) the proportion of rare alleles in any given sample is even more elevated compared to the prediction of Ewens's sampling theory (Chakraborty et al. 1980). Therefore the minimum sample size requirements established here should serve as adequate guidelines even if the size classification of alleles by agarose gel electrophoresis involves undetected hidden variation. Second, Jeffreys et al.'s (1990) study also indicates that the rate of mutation (and therefore θ) may not be constant for all same-size alleles. Nei et al. (1976) entertained such a model, called the variable mutation rate model, the consequences of which are again seen in the preponderance of rare alleles, more than that predicted by Ewens's sampling theory. Therefore variability of mutation rate also does not preclude use of the theory discussed here. Moreover, Clark's (1987) and Flint et al.'s (1989) empirical studies of allele frequency distributions with quasi-continuous size classification of VNTR alleles justify the adequacy of Ewens's sampling theory in the present context.

These comments together with the observation of the preponderance of rare alleles noted in surveys such as that of Odelberg et al. (1989), Boerwinkle et al. (1989), and Ludwig et al. (1989) imply that the number of possible genotypes at VNTR loci is generally quite large (easily of

the order of thousands) and that many of these genotypes should occur in a population with minute probabilities. In fact, some of the genotypes may not even exist in a population at any specific time (generation).

Remember that, if we want to determine a minimum sample size so that the direct estimation of all possible genotype frequencies is possible from their observed relative frequencies in a sample, we must ensure that all possible genotypes are represented in the sample. But the noted characteristics dictate that this is not feasible because the sample size needed to encompass all possible genotypes in the sample is quite large. Sometimes it can be so large that, even if that many individuals could be tested, there is no guarantee that all of them would belong to a single homogeneous population. Because of this, it can be concluded that sample size determination should not be decided using criteria based on direct estimation of genotypic probabilities. In fact, the nature of VNTR polymorphisms necessarily dictates that genotypic probabilities be evaluated from the frequencies of segregating alleles.

This is possible by assuming the HWE law, which should first be validated. Tests for HWE can be varied in nature, although it is claimed that none of them generally have adequate statistical power [see, for example, Ward and Sing (1970) and Emigh (1980)]. Although there are some concerns that VNTR genotype data sometimes fail to conform to HWE predictions (Lander 1989a,b; Cohen 1990), elsewhere it has been shown that, when the allelic distinctions are of a discrete nature, in general, assumption of HWE is reasonable (Boerwinkle et al. 1989; Ludwig et al. 1989; Odelberg et al. 1989; Budowle, Chakraborty et al. 1991; Chakraborty, de Andrade et al. 1991; Edwards et al. 1991; Deka et al. 1991), particularly when the samples are drawn from well-defined populations.

Earlier claims of deviation of genotype frequencies from HWE predictions in DNA typing data in the presence of quasi-continuous allele frequencies have been recently disproved by the development of an appropriate test criterion (Devlin et al. 1990). Furthermore, Chakraborty, de Andrade et al. (1991) have shown that other factors, such as the possibility of incomplete resolution of nearly similar size alleles, the presence of short alleles that are not detectable by RFLP (restriction fragment length polymorphism) analysis, and measurement error, must be taken into account before ascribing the presence of population structure and inbreeding to the source of the apparent heterozygote deficiency seen in RFLP typing of VNTR data. Chakraborty and Jin (1991) have also demonstrated that the apparent heterozygote deficiencies observed in some VNTR surveys employing RFLP analysis of allele size determinations [e.g., Budowle, Giusti et al. (1991)] cannot be due to the presence of population substructure within the populations sampled. Failure to detect deviations from HWE in such tests does not arise from the low statistical

power of the tests employed in these studies, and the fact that such tests do indeed detect deviations from HWE resulting from population heterogeneity is empirically shown by Chakraborty et al. (1988) and Chakraborty (1990b).

The results here show that the sample size requirements for reliable estimation of allele frequencies are relatively more modest. In particular, because most segregating alleles at any VNTR locus are likely to be rare, a sample that encompasses all common alleles with a certain level of confidence and that substitutes an upper bound for the rare allele frequencies is required. This economizes the sample size requirement further. For example, because the number of alleles whose frequency exceeds 1% in a population is generally 10 or less, we can ask for a sample size that will ensure that all these alleles are represented in the sample. Numerical evaluations presented in Table 6 indicate that samples of 300 individuals may be adequate for such purposes.

An issue that needs special attention is that appropriate populations must be studied so that for any given forensic case the relevant population-based data are used. This brings up the issue of allele frequency differences across populations at VNTR loci. One might note that since the discovery of protein-enzyme variations three decades ago, there are still several gaps of population-based allele frequency studies at such loci (Roychoudhury and Nei 1988), and it will take a great deal of effort to come to this stage of allele frequency surveys for VNTR polymorphisms. Therefore, at present, one may have to substitute the most genetically similar population for any specific case study. The question, therefore, is what sample size is adequate for studying interpopulational distances with respect to VNTR polymorphisms. The present results are instructive in this respect as well. Because the rare alleles contribute little to the heterozygosity or genetic distance, one might conclude that, if we sacrifice the rare alleles and concentrate only on reliable estimation of common allele frequencies, the sample size needed would not be very large. Even 50 individuals per population might be enough if we are ready to substitute the frequencies of all rare alleles with an upper bound, such as 0.05. Note that these results are in direct agreement with Nei's (1978) suggestions, which established general guidelines of sample size evaluations in the context of electrophoretic surveys for evolutionary studies.

In summary, population-based VNTR surveys will serve forensics and evolutionary studies of population genetics better if we concentrate on developing more VNTR loci that can be reliably typed. Although the present theory ideally requires an accurate estimation of heterozygosity, the discussion indicates that some underestimation of heterozygosity is of no concern, and hence sacrificing some rare alleles can be tolerated at the expense of cost reduction of sampling. Therefore I conclude that for conservative estimates of allele and genotype frequencies at VNTR

loci, 100–150 individuals per population may be adequate for such surveys. Allele frequency data generated in this process can be used to provide statistical evaluations of false matching for most forensic cases, particularly because all rare events will be cushioned with a higher probability, reducing the chance of biased accusation.

Acknowledgments I thank S.P. Daiger, B. Budowle, E. Boerwinkle, C.T. Caskey, and A. Edwards for their comments and suggestions during the preparation of this paper. I am also thankful to two anonymous reviewers for their efforts to edit an earlier version of this paper. This research is supported by the National Institutes of Health under grant GM 41399 and by the National Institute of Justice, Office of Justice programs, US Department of Justice, under grant 90-IJ-CX-0038. Of course, the opinions expressed in this paper are my own and not the endorsements of the granting agencies or the views of my colleagues who reviewed this paper.

Received 1 July 1991; revision received 28 August 1991.

Literature Cited

Ballantyne, J., G. Sensabaugh, and J. Witkowski. 1989. *DNA Technology and Forensic Science*. Banbury Report 32. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Boerwinkle, E., W. Xiong, E. Fourest, and L. Chan. 1989. Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: Application to the apolipoprotein B 3′ hypervariable region. *Proc. Natl. Acad. Sci. USA* 86:212–216.

Budowle, B., R. Chakraborty, A.M. Giusti, A.E. Eisenberg, and R.C. Allen. 1991. Analysis of the VNTR locus D1S80 by PCR followed by high resolution PAGE. *Am. J. Hum. Genet.* 48:137–144.

Budowle, B., A.M. Giusti, J.S. Waye, F.S. Baechtel, R.M. Fourney, D.E. Adams, L.A. Presley, H.A. Deadman, and K.L. Monson. 1991. Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic purposes. *Am. J. Hum. Genet.* 48:841–855.

Chakraborty, R. 1981. Expected number of rare alleles per locus in a sample and estimation of mutation rates. *Am. J. Hum. Genet.* 33:481–484.

Chakraborty, R. 1990a. Genetic profile of cosmopolitan populations: Effects of hidden subdivision. *Anthropol. Anz.* 48:313–331.

Chakraborty, R. 1990b. Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. *Am. J. Hum. Genet.* 47:87–94.

Chakraborty, R. 1991. Generalized occupancy problem and its applications in population genetics. In *Impact of Genetic Variation on Populations, Families, and Individuals*, C.F. Sing and C.L. Hanis, eds. New York: Oxford University Press (in press).

Chakraborty, R., and S.P. Daiger. 1991. Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah. *Hum. Biol.* 63:571–587.

Chakraborty, R., and R.C. Griffiths. 1982. Correlation of heterozygosity and the number of alleles in different frequency classes. *Theor. Popul. Biol.* 21:205–218.

Chakraborty, R., and L. Jin. 1991. Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. *Hum. Genet.* (in press).

Chakraborty, R., and R.J. Schwartz. 1990. Selective neutrality of surname distribution in an immigrant Indian community of Houston, Texas. *Am. J. Hum. Biol.* 2:1–15.

Chakraborty, R., P.A. Fuerst, and M. Nei. 1980. Statistical studies of protein polymorphism in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. *Genetics* 94:1039–1063.

Chakraborty, R., P.E. Smouse, and J.V. Neel. 1988. Population amalgamation and genetic variation: Observations on artificially agglomerated tribal populations of Central and South America. *Am. J. Hum. Genet.* 43:709–725.

Chakraborty, R., M. de Andrade, S.P. Daiger, and B. Budowle. 1991. Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. *Ann. Hum. Genet.* (in press).

Chakraborty, R., M. Fornage, R. Gueguen, and E. Boerwinkle. 1991. Population genetics of hypervariable loci: Analysis of PCR based VNTR polymorphism within a population. In *DNA Fingerprinting: Approaches and Applications*, T. Burke, G. Dolf, A. Jeffreys, and R. Wolff, eds. Basel: Birkhäuser, 127–143.

Clark, A.G. 1987. Neutrality tests of highly polymorphic restriction-fragment-length polymorphisms. *Am. J. Hum. Genet.* 41:948–956

Cohen, J.E. 1990. DNA fingerprinting for forensic identification: Potential effects of data interpretation of subpopulation heterogeneity and band number variability. *Am. J. Hum. Genet.* 46:358–368.

Deka, R., R. Chakraborty, and R.E. Ferrell. 1991. Population genetics of hypervariable loci in three ethnic groups. *Genomics* 11:83–92.

Devlin, B., N. Risch, and K. Roeder. 1990. No excess of homozygosity at loci used for DNA fingerprinting. *Science* 249:1416–1420.

Edwards, A., A. Civitello, H.A. Hammond, and C.T. Caskey. 1991. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am. J. Hum. Genet.* 49:746–756.

Emigh, T.H. 1980. A comparison of tests for Hardy-Weinberg equilibrium. *Biometrics* 36:627–642.

Ewens, W.J. 1972. The sampling theory of selectivity neutral alleles. *Theor. Popul. Biol.* 3:87–112.

Flint, J., A.J. Boyce, J.J. Martinson, and J.B. Clegg. 1989. Population bottlenecks in Polynesia revealed by minisatellites. *Hum. Genet.* 83:257–263.

Fuerst, P.A., R. Chakraborty, and M. Nei. 1977. Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. *Genetics* 86:455–483.

Jeffreys, A.J., R. Neumann, and V. Wilson. 1990. Repeat unit sequence variation in minisatellites: A novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* 60:473–485.

Jeffreys, A.J., V. Wilson, and S.L. Thein. 1985a. Hypervariable "minisatellite" regions of human DNA. *Nature* 314:67–73.

Jeffreys, A.J., V. Wilson, and S.L. Thein. 1985b. Individual-specific "fingerprints" of human DNA. *Nature* 316:76–79.

Jeffreys, A.J., N.J. Royle, V. Wilson, and Z. Wong. 1988. Spontaneous mutation rates to new alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* 332:278–281.

Johnson, N.L., and S. Kotz. 1969. *Distributions in Statistics: Discrete Distributions*. Boston, MA: Houghton Mifflin.

Lander, E. 1989a. DNA fingerprinting on trial. *Nature* 339:501–505.

Lander, E. 1989b. Population genetic considerations in the forensic use of DNA typing. In *DNA Technology and Forensic Science*, J. Ballantyne, G. Sensabaugh, and J. Witkowski, eds. Banbury Report 32. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 143–156.

Lander, E. 1991. Invited editorial: Research on DNA typing catching up with courtroom applications. *Am. J. Hum. Genet.* 48:819–823.

Ludwig, E.H., W. Friedl, and B.J. McCarthy. 1989. High-resolution analysis of a hypervariable region in the human apolipoprotein B gene. *Am. J. Hum. Genet.* 45:458–464.

Nakamura, Y., M. Leppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin, and R.L. White. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* 235:1616–1622.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583–590.

Nei, M., R. Chakraborty, and P.A. Fuerst. 1976. Infinite allele model with varying mutation rate. *Proc. Natl. Acad. Sci. USA* 73:4164–4168.

Odelberg, S.J., R. Platke, J.R. Eldridge, L. Ballard, P. O'Connell, Y. Nakamura, M. Leppert, J.M. Lalouel, and R. White. 1989. Characterization of eight VNTR loci by agarose gel electrophoresis. *Genomics* 5:915–924.

Rao, C.R. 1957. Maximum likelihood estimation for the multinomial distributions. *Sankhyā* 18:139–148.

Rao, C.R. 1958. Maximum likelihood estimation for the multinomial distribution with infinite number of classes. *Sankhyā* 20:211–218.

Roychoudhury, A.K., and M. Nei. 1988. *Human Polymorphic Genes: World Distribution*. New York: Oxford University Press.

Ward, R.H., and C.F. Sing. 1970. A consideration of the power of the χ^2 test to detect inbreeding effects in natural populations. *Am. Natur.* 104:355–363.

Sample Size Requirements for Addressing the Population Genetic Issues of Forensic Use of DNA Typing

RANAJIT CHAKRABORTY[1]

Abstract DNA typing offers a unique opportunity to identify individuals for medical and forensic purposes. Probabilistic inference regarding the chance occurrence of a match between the DNA type of an evidentiary sample and that of an accused suspect, however, requires reliable estimation of genotype and allele frequencies in the population. Although population-based data on DNA typing at several hypervariable loci are being accumulated at various laboratories, a rigorous treatment of the sample size needed for such purposes has not been made from population genetic considerations. It is shown here that the loci that are potentially most useful for forensic identification of individuals have the intrinsic property that they involve a large number of segregating alleles, and a great majority of these alleles are rare. As a consequence, because of the large number of possible genotypes at the hypervariable loci that offer the maximum potential for individualization, the sample size needed to observe all possible genotypes in a sample is large. In fact, the size is so large that even if such a huge number of individuals could be sampled, it could not be guaranteed that such a sample was drawn from a single homogeneous population. Therefore adequate estimation of genotypic probabilities must be based on allele frequencies, and the sample size needed to represent all possible alleles is far more reasonable. Further economization of sample size is possible if one wants to have representation of only the frequent alleles in the sample, so that the rare allele frequencies can be approximated by an upper bound for forensic applications.

It is now well known that interspersed in the human genome are numerous DNA regions that have genetic variation whose magnitude is much larger than that coded by the traditional serological loci (Jeffreys et al. 1985a,b; Nakamura et al. 1987). Such genetic hypervariability can

[1]Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, PO Box 20334, Houston, TX 77225.

Human Biology, April 1992, Vol. 64, No. 2, pp. 141–159.

KEY WORDS: VNTR POLYMORPHISM, SAMPLE SIZE, INFINITE ALLELE MODEL, DNA TYPING, FORENSIC GENETICS

be used profitably to identify specific individuals by DNA typing at such regions of the genome. In recent years, therefore, civil and criminal courts in the United States and Europe have been admitting DNA typing evidence in resolving legal controversies involved in both paternity disputes and criminal cases [see, for example, Ballantyne et al. (1989)]. Because DNA typing technology is relatively new and because population-based surveys on variability at such genetic loci are less abundant compared to the classical serological markers, concerns have been raised about their general applicability for legal purposes (Lander 1989a,b, 1991; Cohen 1990).

The criticisms with regard to admitting DNA typing evidence in legal cases include the inappropriateness of using population data from small samples (Lander 1989b). It is therefore necessary to determine how large a sample must be for the statistical analysis of DNA typing data to be regarded as appropriate for legal applications. Note that for forensic use DNA typing data can be represented in terms of multinomial distributions with a large number of possible classes [see, for example, Budowle, Giusti et al. (1991)] whose theory is completely characterized (Rao 1957, 1958; Johnson and Kotz 1969). However, there is no systematic discussion on this subject in the context of allele and genotype frequencies generally observed at the hypervariable loci that currently are being used for forensic cases.

My purpose here is to address this issue by asking what sample sizes are adequate for conservative evaluation of genotype or allele frequencies. Based on the population genetic characteristics of the hypervariable loci, I show that the large heterozygosities at such loci necessarily imply that the expected number of alleles at each of these loci is generally quite large (often larger than 50) and that there is a predominance of rare alleles (i.e., alleles that occur in frequencies as small as 0.01) at such loci. Furthermore, the total number of alleles and the number of rare alleles are increasing functions of sample size. Consequently, if we want to determine a minimum sample size based on the criterion that all possible genotypes must be represented in the sample, the needed sample size would be so large that it would be impractical to test that many individuals from a single homogeneous population. It therefore becomes necessary to evaluate genotype frequencies from allele frequency data, and the sample size requirements can be derived by estimating allele frequencies from the sample. Based on representation of all alleles in the sample, feasible sample sizes can be obtained from well-defined homogeneous populations. However, because many of the alleles are characteristically of rare frequency (as defined before), sample size requirements still can be severe. It might be more economical to ask for a sample size whereby the frequencies of more frequent alleles are estimated with reliable precision and to determine a threshold for the

frequencies of rare alleles for which an upper bound can be prescribed based on such threshold values. This procedure should yield a conservative estimate (biased in the upward direction) of probabilities of obtaining a match for cases involving rare genotypes, even when the exact evaluation of chance occurrence of a match becomes impossible based on the sampled allele frequencies.

Throughout this paper I call the class of loci that are shown to be useful for forensic purposes VNTR (variable number of tandem repeat) loci, following Nakamura et al. (1987), although various other names for such loci have been proposed depending on the core motifs of their nucleotide sequences [see, for example, Edwards et al. (1991)]. I also assume that the allelic distinctions at such loci are made without ambiguity, so that the different alleles are discrete and no measurement error is involved in the size classification of alleles. When allele sizes are quasi-continuous in a population, the present theory can be applied with suitable binning of alleles [e.g., Budowle, Giusti et al. (1991)] without any major changes in the qualitative conclusions of the present results. The theory also assumes that VNTR loci have an autosomal codominant mode of transmission and that the genotype frequencies satisfy Hardy-Weinberg equilibrium (HWE) expectations. Note that, in the absence of measurement error, data on such loci collected from well-defined populations generally show that this assumption is appropriate [see, for example, Boerwinkle et al. (1989), Ludwig et al. (1989), Budowle, Chakraborty et al. (1991), Chakraborty, Fornage et al. (1991), Deka et al. (1991), and Edwards et al. (1991)]. Polymorphisms at VNTR loci with discretized allelic distinctions also have been shown to follow the predictions of the neutral mutation model of the infinite allele model [e.g., Jeffreys et al. (1988), Budowle, Chakraborty et al. (1991), Chakraborty, Fornage et al. (1991), Chakraborty and Daiger (1991), Deka et al. (1991), and Edwards et al. (1991)], which is assumed to generate the expected number of alleles at VNTR loci.

Throughout this article the sample size is defined as the number of individuals sampled from the population (designated by n) so that, whenever sample size is considered with regard to the number of alleles sampled, it is equated to $2n$ (because of autosomal inheritance of the loci and diploidy of the human genome).

Theory and Results

Expected Total Number of Alleles and Number of Rare Alleles in a Sample from an Equilibrium Population. Although the high degree of polymorphism at VNTR loci has been noted in almost all human populations—and this is characterized in terms of high heterozygosity at

these loci and by a large number of alleles—it is not possible to determine the exact number of alleles that can occur at VNTR loci in any population. This is because the allelic possibilities are truly infinite for such loci, and hence predictions can be made only with regard to the sampling distribution of the number of alleles found in any sample of alleles drawn from a population. By assuming selective neutrality of VNTR alleles, one can employ the sampling theory of selectively neutral alleles (Ewens 1972; Chakraborty and Griffiths 1982) to determine the expected total number of alleles and the expected number of alleles for any specific allele frequency class. In a sample of $2n$ alleles (n individuals) randomly drawn from a population that is at mutation-drift equilibrium, Ewens (1972) showed that the expected total number of alleles can be expressed in terms of an unknown parameter (θ) and the sample size (n) by the equation

$$E(k) = \theta \sum_{i=0}^{2n-1} (\theta + i)^{-1}, \tag{1}$$

where $\theta = 4Nv$, in which N is the effective size of the population and v is the mutation rate at the locus per generation. Chakraborty (1981) and Chakraborty and Griffiths (1982) further showed that the expected number of alleles whose frequency lies in the range p_1 to p_2 ($0 \leq p_1 < p_2 \leq 1$) in such a sample is given by

$$E[k(p_1, p_2)] = \sum_{i=[2np_1]+1}^{[2np_2]} \frac{\theta}{i} \frac{n!}{(n-i)!} \frac{\Gamma(n+\theta-i)}{\Gamma(n+\theta)}, \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function and $[n]$ is the largest integer contained in n. Equation (2) immediately proscribes the expected number of rare alleles in a sample by substituting $p_1 = 0$ and $p_2 = p$ for any small p (say, $p = 0.001$, 0.01, or 0.05). Therefore the expected number of rare alleles in a sample of n individuals becomes

$$E[k(p)] = \sum_{i=1}^{[2np]} \frac{\theta}{i} \frac{n!}{(n-i)!} \frac{\Gamma(n+\theta-i)}{\Gamma(n+\theta)}, \tag{3}$$

which can be approximated with good precision (Chakraborty 1981) by

$$E[k(p)] \approx A\theta - B\theta^2, \tag{4}$$

where

$$A = \sum_{i=1}^{[2np]} (1/i) \quad \text{and} \quad B = \sum_{i=1}^{[2np]} (2n-i)^{-1} \tag{5}$$

are constants depending on the sample size n and the definition of rare alleles ($p = 0.001$, 0.01, 0.05, or any arbitrary small number). It should

Table 1. Expected Total Number of Alleles in a Sample of n Individuals for a Given Level of Heterozygosity

	Heterozygosity				
n	*0.20*	*0.50*	*0.75*	*0.90*	*0.95*
50	2.2	5.2	11.1	22.9	35.3
100	2.4	5.9	13.2	28.8	46.9
500	2.8	7.5	18.0	43.0	76.2
1,000	3.0	8.2	20.0	49.2	89.2
10,000	3.5	10.5	26.9	69.9	132.7

be noted that both k (the total number of alleles) and $k(p)$ (the number of rare alleles) in a sample are random variables and hence do not have fixed values, but their sampling distributions are known. Ewens (1972) gave the exact sampling distribution for k [which is quite complex but can be numerically evaluated; see Stewart's algorithm in the appendix of Fuerst et al. (1977)], and Chakraborty and Griffiths (1982) showed that for small values of p (such as the ones noted here), the variable $k(p)$ follows a Poisson distribution so that expression (3) or (4) also gives the variance of the number of rare alleles in a sample.

Equations (1) and (4) can be used to evaluate the expected numbers of rare alleles for any selectively neutral locus, for which we must know the composite parameter θ. One way of estimating this parameter is from the level of heterozygosity at the locus, which is relatively less sensitive to sample size. Denote the sample value of heterozygosity by H. It is known that under the assumption of equilibrium the expectation of H is $\theta/(1 + \theta)$ so that the moment estimator of θ is given by $H/(1 - H)$. The justifications for this moment estimator of θ in the present context are given by Chakraborty (1990a,b) and Chakraborty and Schwartz (1990). Therefore Eqs. (1) and (4) can be used to predict roughly how many alleles are expected at any selectively neutral locus if the sample is drawn from an equilibrium population and how many rare alleles there may be in such a sample.

Table 1 presents the expected total number of alleles for some representative heterozygosity values that are generally seen at VNTR loci for several sample sizes (number of individuals). It is clear from this table that the observed number of alleles is an increasing function of heterozygosity and of sample size, and, in particular, when the heterozygosity is high (say, 95%), the number of alleles can easily exceed 40 when more than 100 individuals are sampled from the population. Therefore a large number of alleles at any VNTR locus is an intrinsic property of such polymorphisms.

Table 2. Expected Number of Rare Alleles in a Sample of n Individuals for Selected Levels of Heterozygosity at a Locus

		n				
H	P	50	100	500	1,000	10,000
0.20	0.001	–[a]	–	0.3	0.4	0.9
	0.01	0.3	0.4	0.7	0.9	1.5
	0.05	0.6	0.7	1.1	1.3	1.9
0.50	0.001	–	–	1.0	1.5	3.6
	0.01	1.0	1.5	2.9	3.6	5.9
	0.05	2.2	2.9	4.5	5.1	7.4
0.75	0.001	–	–	3.0	4.5	10.8
	0.01	2.9	4.4	8.7	10.7	17.5
	0.05	6.4	8.3	13.0	15.1	22.0
0.90	0.001	–	–	8.9	13.4	32.3
	0.01	8.2	12.7	25.6	31.6	52.1
	0.05	16.4	22.2	36.3	42.5	63.2
0.95	0.001	–	–	18.6	28.1	68.0
	0.01	15.4	24.9	52.0	64.7	108.1
	0.05	24.8	37.1	67.0	80.0	123.7

a. Dash denotes cases where this definition of rare alleles is inappropriate.

Such a large number of alleles is also observed in many empirical surveys in accordance with this theory. For example, Boerwinkle et al. (1989) observed 12 alleles at the ApoB VNTR locus in a French population of 125 unrelated individuals, and Ludwig et al. (1989) found 14 alleles at the same locus in a survey of 318 US whites. Because their observed heterozygosity values were 75% and 78%, respectively, by using Eq. (1), we get expectations for the total number of alleles for these sample sizes of 13.8 and 18.9, respectively, which are in fair agreement with the observations [particularly because k, the total number of alleles, also has a large variance; see Ewens (1972)]. An even higher number of alleles was observed in the Utah white population at several VNTR loci (Odelberg et al. 1989). This can be explained by the high heterozygosity (75–95%) at the loci surveyed. The largest number of alleles observed by Odelberg et al. (1989) was 67 at the D2S44 locus, which was reported as having a heterozygosity of 95%. All these results indicate that high heterozygosity at VNTR loci necessarily leads to a large number of segregating alleles, and this is consistent with the expectations of the pattern of polymorphism at these loci.

Table 2 shows another important feature of such polymorphisms. Using Eq. (4), one can compute the expected number of rare alleles for any given level of heterozygosity in a sample of size n. In the computations presented in Table 2, I used three definitions of rare alleles (p

= 0.001, 0.01, and 0.05), noting that, when less than 50 individuals were surveyed, rare alleles could not be defined with the criterion $p < 0.01$. The general implications of the results shown in this table are: (1) Many of the segregating alleles at VNTR loci are rare; (2) like the total number of alleles, the number of rare alleles is also an increasing function of heterozygosity and sample size; and (3) the proportion of rare alleles increases with increasing sample size and heterozygosity. For example, when $H = 90\%$, in a sample of 50 individuals 16.4 of the 23 alleles will have a frequency below 5%, whereas in a sample of 10,000 individuals 63 of the 70 alleles at the locus are expected to have an allele frequency below 5%. For higher heterozygosities the rare alleles will constitute an even larger fraction of the total number of alleles. Analysis of the empirical survey data [e.g., Odelberg et al. (1989)] indicates that such theoretical expectations are congruent with the observations. For example, at the D2S44 locus Odelberg et al. (1989) observed 37 of the 67 alleles with a frequency less than 1% ($p = 0.01$), although 42.5 ± 6.5 are expected to be in this class using Eq. (4) and assuming a Poisson distribution of rare alleles.

The features of polymorphism seen in the computations of Tables 1 and 2 have important implications for sample size requirements in VNTR surveys. These results indicate that, given large heterozygosities, the number of possible genotypes at VNTR loci often can be large. For example, when $H = 0.95$, we can easily expect as many as 90 alleles (for $n = 1000$), and the number of possible genotypes at a VNTR locus will be 4095. Furthermore, because at a VNTR locus there can be 80 alleles with frequency 5% or less, 3240 of these genotypes should have a frequency less than 0.5% should the population be in Hardy-Weinberg equilibrium with respect to the genotype frequencies. Obviously, we would like to know what sample size will be required to represent all these genotypes in the sample if we are to use genotype data to estimate all genotype frequencies directly.

Probability of Observing All Possible Genotypes in a Sample of Fixed Size and Minimum Sample Size Requirement for Representation of All Genotypes in a Sample. The minimum sample size requirement can be addressed in two alternative ways. First, we can evaluate the probability that all possible genotypes are represented in a sample of a given size. Elsewhere I have shown that this probability can be evaluated only when the number of alleles and their frequencies in the population are known (Chakraborty 1991). Consider a locus for which there are k segregating alleles, where $p_1, p_2, \ldots, p_k$ (the p-vector) represent the true allele frequencies. If we assume that in the population the genotype frequencies follow Hardy-Weinberg expectations, the vector of genotype

frequencies is of dimension $K = k(k + 1)/2$, and it can be represented by $q_1, q_2, \ldots, q_K$, where the q_i are either of the form p_i^2 or $2p_ip_j$ depending on the specific genotypes. In a sample of n individuals scored for such a locus, the probability that all genotypes are represented is given by [see Chakraborty (1991) for derivation]

$$P = 1 - \sum_{r=1}^{K} (-1)^{r-1} S_r, \tag{6}$$

where S_r is the probability that at least r of the K genotypes are not observed in the sample; that is,

$$S_1 = \sum_{i=1}^{K} (1 - q_i)^n, \quad S_2 = \sum_{i>j=1}^{K}\sum (1 - q_i - q_j)^n, \text{ etc.} \tag{7}$$

Although Eq. (6) can be numerically evaluated for any given values of k and p-vector, it involves tedious enumerations of a large number of summations, particularly for k as large as the ones noted earlier. However, Chakraborty (1991) derived the exact sampling distribution of the random variable X, the number of distinct genotypes observed in a sample for known values of k and p-vector, and this distribution also can be approximated by a normal distribution with mean μ and variance σ^2, where

$$\mu = \frac{k(k + 1)}{2} - S_1, \tag{8a}$$

$$\sigma^2 = S_1(1 - S_1) + 2S_2, \tag{8b}$$

and S_1 and S_2 are as defined in Eqs. (7).

Therefore the probability P of Eq. (6) can be approximated by

$$P \approx 1 - \Phi(S_1/\sigma), \tag{9}$$

where $\Phi(x)$ is the cumulative probability of a standard normal variate up to x.

Table 3 shows the result of such a computation for some short tandem repeat (STR) loci, where the number of alleles and their frequencies are taken from the survey of 40 or more unrelated white individuals reported by Edwards et al. (1991). It is clear from this table that, at these STR loci (where the number of alleles generally ranges from 6 to 17) and because many of the alleles are rare, all possible genotypes may not generally be observed even in an extremely large sample. For example, for the $(AGAT)_n$ (HUMHPRTB) locus, even in a sample of 1 million individuals, the probability of observing all possible genotypes is only 0.5. Therefore in any sample of a fixed size n the chance of observing all possible genotypes in general is very small.

Table 3. Probability of Observing All Possible Genotypes in Samples of Fixed Size for Some Representative Short Tandem Repeat Loci

n^a	*Loci*[b]				
	$(AGAT)_n$	$(AATG)_n$	$(ACAG)_n$	$(AAT)_n$	$(AGC)_n$
50	2.06×10^{-37}	4.83×10^{-7}	0.025	2.72×10^{-14}	(c)
100	2.51×10^{-32}	7.05×10^{-5}	0.106	3.37×10^{-12}	2.97×10^{-14}
200	1.03×10^{-26}	2.13×10^{-3}	0.243	1.97×10^{-9}	8.45×10^{-9}
500	1.26×10^{-19}	0.019	0.414	1.12×10^{-5}	1.03×10^{-4}
1,000	1.67×10^{-15}	0.029	0.482	3.09×10^{-4}	3.96×10^{-3}
10,000	1.04×10^{-5}	0.317	1.0	0.014	0.364
100,000	0.174	0.500	1.0	0.204	1.0
1,000,000	0.500	1.0	1.0	0.496	1.0

a. Number of individuals typed.
b. Data on allele frequencies and number of alleles on these short tandem repeat loci are taken from the white sample examined by Edwards et al. (1991).
c. Sample size is smaller than the number of possible genotypes giving a zero probability for observing all possible genotypes in the sample.

These computations, however, do not prescribe any well-defined minimum sample size requirement for observing all possible genotypes; nor can the minimum sample size be evaluated analytically by any direct method. A crude conservative sample size estimate can be obtained by the following alternative method. By using Eq. (6), the probability P of observing all possible genotypes in a sample of n individuals satisfies the inequality

$$P \geq 1 - \sum_{i=1}^{k} (1 - p_i^2)^n - \sum_{i>j=1}^{k}\sum^{k} (1 - 2p_ip_j)^n, \tag{10}$$

in which the right-hand side is at a maximum when all allele frequencies are equal, that is, when $p_i = 1/k$ for all i. Therefore a conservative estimate of the minimum sample size requirement for ensuring that all genotypes are represented in the sample with confidence $(1 - \alpha)$ is given by the inequality

$$1 - k(1 - k^{-2})^n - \tfrac{1}{2}k(k - 1)(1 - 2k^{-2})^n \geq 1 - \alpha, \tag{11}$$

in which the substitutions $(1 - k^{-2})^n \approx e^{-n/k}$ and $(1 - 2k^{-2})^n \approx e^{-2n/k}$ yield

$$n \geq -k^2 \ln\left[\frac{\sqrt{k^2 + 2\alpha k(k - 1)} - k}{k(k - 1)}\right]. \tag{12}$$

Table 4. Conservative Estimates of the Number of Individuals Needed to Represent All Possible Genotypes in a Sample for a k-Allelic Codominant Locus

	Minimum Sample Size[a] Needed for:				
k	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.25$
5	213	156	116	99	77
10	921	691	532	465	379
15	2,164	1,647	1,288	1,137	944
20	3,962	3,043	2,406	2,137	1,794
25	6,330	4,893	3,899	3,479	2,943
30	9,279	7,210	5,778	5,174	4,402
40	16,955	13,278	10,733	9,659	8,287
50	27,051	21,305	17,329	15,651	13,507
100	115,134	92,153	76,248	69,539	60,970

a. Sample size in these computations refers to the number of individuals to be typed for each k-allelic codominant locus. The values of $\alpha = 0.001$, 0.01, 0.05, 0.10, and 0.25 represent 99.9%, 99%, 95%, 90%, and 75% confidence, respectively, of being assured that all possible genotypes are represented in the sample.

When the allele frequencies are not equal (as is the case for all VNTR loci), sample size requirements for representation of all possible genotypes can far exceed the bound prescribed by expression (12), and numerical evaluation of this expression is instructive enough to show that it is not feasible to collect samples large enough to encompass all possible genotypes for any VNTR locus in any population. Table 4 presents numerical evaluations of expression (12) for some representative values of k, the number of alleles that are in the general range seen in Table 1. It is clear that, even with this conservative minimum sample size estimate, a sample of 15,651 individuals is required to encompass all possible genotypes with 90% confidence if there are 50 alleles segregating at a VNTR locus. This is generally too much to ask in a survey study, and even if such a large sample could be collected, there is no guarantee that the individuals truly came from a single homogeneous population.

The analysis clearly establishes that, if we are to use the observed relative frequencies of all genotypes as the estimates of genotypic probabilities in the population, a sample of adequate size cannot be collected because from any reasonable homogeneous population this large a sample cannot be gathered. An appropriate alternative way to estimate the genotype frequencies is therefore to use the estimate of allele frequencies and to invoke assumptions through which genotype probabilities can be derived based on allele frequency estimates (such as the Hardy-Weinberg equilibrium assumption).

Sample Size Requirement Based on Allele Frequencies. Having shown that the only practical and reliable way to estimate genotypic

Table 5. Minimum Number of Individuals Needed to Represent All Alleles in a Sample for a k-Allelic Codominant Locus

k	*Minimum Sample Size[a] Needed for:*				
	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.25$
5	22	16	12	10	8
10	46	35	27	23	19
15	72	55	43	38	31
20	99	76	60	53	44
25	127	98	78	69	58
30	155	120	96	86	72
40	212	166	134	120	102
50	271	213	171	156	133
100	576	461	380	346	300

a. Number of individuals to be typed.

probabilities at VNTR loci is from the allele frequencies, I can now turn to the evaluation of the minimum sample size requirement based on allele frequencies. Again the logic of deriving expression (12) can be used to determine a crude conservative estimate of minimum sample size. For a locus with k segregating alleles whose frequencies in a population are p_1, p_2, etc., the probability that all alleles are represented in a sample of n individuals should exceed the quantity

$$1 - \sum_{i=1}^{k} (1 - p_i)^{2n}. \tag{13}$$

In order for expression (13) to exceed the level of confidence $(1 - \alpha)$, we must ensure that

$$1 - k(1 - k^{-1})^{2n} \geq 1 - \alpha, \tag{14}$$

or

$$n \geq \tfrac{1}{2} \ln(\alpha/k)/\ln(1 - k^{-1}). \tag{15}$$

Admittedly, this bound of minimum sample size is too crude because, when the allele frequencies are not equal, far larger sample sizes are needed for all alleles to be represented in a sample. Nevertheless, Table 5 shows that use of expression (15) leads to sample size estimates that are feasible to collect from any well-defined population. For example, to ensure that all 50 equifrequent alleles are represented in a sample with 95% confidence ($\alpha = 0.05$), we need to type 171 individuals from the population. In practice, however, the required sample size may be larger, because these computations do not represent the reality of the situation,

Table 6. Minimum Number of Individuals Needed to Have r Alleles with Frequency p or Above Represented in the Sample

		Minimum Sample Size Needed for:				
r	p	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.25$
1	0.001	3452	2302	1498	1151	693
	0.01	344	230	150	115	69
	0.05	68	45	30	23	14
2	0.001	3798	2647	1838	1485	1005
	0.01	379	264	183	148	100
	0.05	74	52	36	29	20
5	0.001	4257	3104	2292	1935	1442
	0.01	424	309	229	193	144
	0.05	84	61	45	38	29
10	0.001	4603	3450	2637	2278	1781
	0.01	459	344	263	227	178
	0.05	90	68	52	45	35

namely, the allele frequencies are not equal, and thus such a direct evaluation of minimum sample size is not possible.

Alternatively, we might ask what sample size would be required if we want to ensure that all alleles with frequencies above a certain small value will be represented in the sample with a proscribed level of confidence. Because the probability that an allele with frequency p remains unobserved in a sample of n individuals is given by $(1 - p)^{2n}$, if there are r alleles at a locus that have frequencies p or above in the population [reasonable values of r can be obtained from $k - k(p)$, from the first section of the previous analysis], in order for all these common alleles to be represented in the sample, we must have

$$[1 - (1 - p)^{2n}]^r \geq 1 - \alpha, \tag{16}$$

or

$$n \geq \ln[1 - (1 - \alpha)^{1/r}]/2 \ln(1 - p). \tag{17}$$

Table 6 presents sample size estimates based on this inequality. As seen in Tables 1 and 2, for most VNTR loci, even when the total number of alleles is large, the expected number of alleles having frequency p or above is generally below 10 for $p = 0.001$, 0.01, or 0.05. Therefore in Table 6 the minimum sample size requirement is presented for values of $r \leq 10$. It is clear from this table that, if we sacrifice the alleles of frequency below 0.01, a sample of 300 individuals will ensure that all common alleles (alleles with frequency greater than 1%) will be represented in a sample with at least 95% confidence. This is a much more feasible sampling strategy and should guarantee reliable estimation of

Table 7. Frequency of Alleles That Will Be Represented in a Sample of n Individuals with a Given Level of Confidence

	Allele Frequencies for:					
	$r = 1$		$r = 5$		$r = 10$	
n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
50	0.0450	0.0295	0.0602	0.0448	0.0667	0.0514
100	0.0228	0.0149	0.0306	0.0227	0.0339	0.0260
200	0.0114	0.0075	0.0154	0.0114	0.0171	0.0131
500	0.0046	0.0030	0.0062	0.0046	0.0069	0.0053
1000	0.0023	0.0015	0.0003	0.0002	0.0003	0.0003

the frequencies of common alleles in a population. Note that the sample size estimates of Table 6 are even more economical if we sacrifice all alleles having frequency 0.05 or less, in which case 50 individuals may be sufficient to ensure the presence of all common alleles in the sample with 95% confidence.

The inequality (17) can also be written in the form

$$p \geq 1 - [1 - (1 - \alpha)^{1/r}]^{1/2n}, \tag{18}$$

which can be used to examine which allele frequencies are reliably estimated in a survey of n individuals. This bound also proscribes a threshold value for the rare allele frequencies that would yield a conservative probability of a match in forensic cases involving previously unseen DNA types. Table 7 presents some representative values of such minimum bounds of allele frequencies. It shows that in the VNTR surveys involving 200 or more individuals the alleles with frequency above 1% are generally represented, and even when the sample size is 50, alleles with frequency above 5% should be encompassed in the sample.

Discussion and Conclusion

The final step in using DNA typing data in forensic applications consists in using estimates of specific genotype frequencies to determine how often by chance alone two biological specimens from two different individuals have identical DNA type results. Obviously, reliable conservative estimates of genotype probabilities are required for such a purpose, and population-based data must provide such estimates. Activities at various laboratories are currently geared toward providing such data. It is intuitively clear that a hypervariable locus that provides greater heterozygosity is also more efficient for resolving a forensic case, because the chance of a match by chance alone decreases as heterozygosity in-

creases. Therefore from a strategic point of view hypervariable loci that contain larger heterozygosities should be considered first for gathering population data. Of course, the cost efficiency and technical reproducibility of typing results also must be considered in selecting the loci that serve the purpose better.

Here, I first show that one of the intrinsic population genetic characteristics of VNTR polymorphisms is that VNTR loci generally contain a large number of segregating alleles whose exact number in any population is a random variable and hence is strictly unknown. The expected number of alleles, however, can be derived by assuming that the pattern of VNTR polymorphism follows the predictions of the infinite allele model of selectively neutral alleles. Validation of this assumption is provided by several recent articles [see, for example, Jeffreys et al. (1988), Budowle, Chakraborty, de Andrade et al. (1991), Chakraborty, Fornage et al. (1991), Edwards et al. (1991), and Deka et al. (1991)], particularly when the DNA typing protocol can discretize the allelic distinctions by techniques such as high-resolution Southern gel electrophoresis following polymerase chain reaction (PCR) techniques.

In view of the recent article by Jeffreys et al. (1990) that VNTR alleles of identical size may not always be iso-allelic at a molecular level and that generation of new alleles at VNTR loci may not exactly correspond to the infinite allele model, one might question the applicability of Ewens's sampling theory invoked in the present analysis. To this effect several comments are noteworthy. First, earlier studies in relation to protein variation have shown that in the presence of hidden variation (within allelic classes) the proportion of rare alleles in any given sample is even more elevated compared to the prediction of Ewens's sampling theory (Chakraborty et al. 1980). Therefore the minimum sample size requirements established here should serve as adequate guidelines even if the size classification of alleles by agarose gel electrophoresis involves undetected hidden variation. Second, Jeffreys et al.'s (1990) study also indicates that the rate of mutation (and therefore θ) may not be constant for all same-size alleles. Nei et al. (1976) entertained such a model, called the variable mutation rate model, the consequences of which are again seen in the preponderance of rare alleles, more than that predicted by Ewens's sampling theory. Therefore variability of mutation rate also does not preclude use of the theory discussed here. Moreover, Clark's (1987) and Flint et al.'s (1989) empirical studies of allele frequency distributions with quasi-continuous size classification of VNTR alleles justify the adequacy of Ewens's sampling theory in the present context.

These comments together with the observation of the preponderance of rare alleles noted in surveys such as that of Odelberg et al. (1989), Boerwinkle et al. (1989), and Ludwig et al. (1989) imply that the number of possible genotypes at VNTR loci is generally quite large (easily of

the order of thousands) and that many of these genotypes should occur in a population with minute probabilities. In fact, some of the genotypes may not even exist in a population at any specific time (generation).

Remember that, if we want to determine a minimum sample size so that the direct estimation of all possible genotype frequencies is possible from their observed relative frequencies in a sample, we must ensure that all possible genotypes are represented in the sample. But the noted characteristics dictate that this is not feasible because the sample size needed to encompass all possible genotypes in the sample is quite large. Sometimes it can be so large that, even if that many individuals could be tested, there is no guarantee that all of them would belong to a single homogeneous population. Because of this, it can be concluded that sample size determination should not be decided using criteria based on direct estimation of genotypic probabilities. In fact, the nature of VNTR polymorphisms necessarily dictates that genotypic probabilities be evaluated from the frequencies of segregating alleles.

This is possible by assuming the HWE law, which should first be validated. Tests for HWE can be varied in nature, although it is claimed that none of them generally have adequate statistical power [see, for example, Ward and Sing (1970) and Emigh (1980)]. Although there are some concerns that VNTR genotype data sometimes fail to conform to HWE predictions (Lander 1989a,b; Cohen 1990), elsewhere it has been shown that, when the allelic distinctions are of a discrete nature, in general, assumption of HWE is reasonable (Boerwinkle et al. 1989; Ludwig et al. 1989; Odelberg et al. 1989; Budowle, Chakraborty et al. 1991; Chakraborty, de Andrade et al. 1991; Edwards et al. 1991; Deka et al. 1991), particularly when the samples are drawn from well-defined populations.

Earlier claims of deviation of genotype frequencies from HWE predictions in DNA typing data in the presence of quasi-continuous allele frequencies have been recently disproved by the development of an appropriate test criterion (Devlin et al. 1990). Furthermore, Chakraborty, de Andrade et al. (1991) have shown that other factors, such as the possibility of incomplete resolution of nearly similar size alleles, the presence of short alleles that are not detectable by RFLP (restriction fragment length polymorphism) analysis, and measurement error, must be taken into account before ascribing the presence of population structure and inbreeding to the source of the apparent heterozygote deficiency seen in RFLP typing of VNTR data. Chakraborty and Jin (1991) have also demonstrated that the apparent heterozygote deficiencies observed in some VNTR surveys employing RFLP analysis of allele size determinations [e.g., Budowle, Giusti et al. (1991)] cannot be due to the presence of population substructure within the populations sampled. Failure to detect deviations from HWE in such tests does not arise from the low statistical

power of the tests employed in these studies, and the fact that such tests do indeed detect deviations from HWE resulting from population heterogeneity is empirically shown by Chakraborty et al. (1988) and Chakraborty (1990b).

The results here show that the sample size requirements for reliable estimation of allele frequencies are relatively more modest. In particular, because most segregating alleles at any VNTR locus are likely to be rare, a sample that encompasses all common alleles with a certain level of confidence and that substitutes an upper bound for the rare allele frequencies is required. This economizes the sample size requirement further. For example, because the number of alleles whose frequency exceeds 1% in a population is generally 10 or less, we can ask for a sample size that will ensure that all these alleles are represented in the sample. Numerical evaluations presented in Table 6 indicate that samples of 300 individuals may be adequate for such purposes.

An issue that needs special attention is that appropriate populations must be studied so that for any given forensic case the relevant population-based data are used. This brings up the issue of allele frequency differences across populations at VNTR loci. One might note that since the discovery of protein-enzyme variations three decades ago, there are still several gaps of population-based allele frequency studies at such loci (Roychoudhury and Nei 1988), and it will take a great deal of effort to come to this stage of allele frequency surveys for VNTR polymorphisms. Therefore, at present, one may have to substitute the most genetically similar population for any specific case study. The question, therefore, is what sample size is adequate for studying interpopulational distances with respect to VNTR polymorphisms. The present results are instructive in this respect as well. Because the rare alleles contribute little to the heterozygosity or genetic distance, one might conclude that, if we sacrifice the rare alleles and concentrate only on reliable estimation of common allele frequencies, the sample size needed would not be very large. Even 50 individuals per population might be enough if we are ready to substitute the frequencies of all rare alleles with an upper bound, such as 0.05. Note that these results are in direct agreement with Nei's (1978) suggestions, which established general guidelines of sample size evaluations in the context of electrophoretic surveys for evolutionary studies.

In summary, population-based VNTR surveys will serve forensics and evolutionary studies of population genetics better if we concentrate on developing more VNTR loci that can be reliably typed. Although the present theory ideally requires an accurate estimation of heterozygosity, the discussion indicates that some underestimation of heterozygosity is of no concern, and hence sacrificing some rare alleles can be tolerated at the expense of cost reduction of sampling. Therefore I conclude that for conservative estimates of allele and genotype frequencies at VNTR

loci, 100–150 individuals per population may be adequate for such surveys. Allele frequency data generated in this process can be used to provide statistical evaluations of false matching for most forensic cases, particularly because all rare events will be cushioned with a higher probability, reducing the chance of biased accusation.

Acknowledgments I thank S.P. Daiger, B. Budowle, E. Boerwinkle, C.T. Caskey, and A. Edwards for their comments and suggestions during the preparation of this paper. I am also thankful to two anonymous reviewers for their efforts to edit an earlier version of this paper. This research is supported by the National Institutes of Health under grant GM 41399 and by the National Institute of Justice, Office of Justice programs, US Department of Justice, under grant 90-IJ-CX-0038. Of course, the opinions expressed in this paper are my own and not the endorsements of the granting agencies or the views of my colleagues who reviewed this paper.

Received 1 July 1991; revision received 28 August 1991.

Literature Cited

Ballantyne, J., G. Sensabaugh, and J. Witkowski. 1989. *DNA Technology and Forensic Science*. Banbury Report 32. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Boerwinkle, E., W. Xiong, E. Fourest, and L. Chan. 1989. Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: Application to the apolipoprotein B 3′ hypervariable region. *Proc. Natl. Acad. Sci. USA* 86:212–216.

Budowle, B., R. Chakraborty, A.M. Giusti, A.E. Eisenberg, and R.C. Allen. 1991. Analysis of the VNTR locus D1S80 by PCR followed by high resolution PAGE. *Am. J. Hum. Genet.* 48:137–144.

Budowle, B., A.M. Giusti, J.S. Waye, F.S. Baechtel, R.M. Fourney, D.E. Adams, L.A. Presley, H.A. Deadman, and K.L. Monson. 1991. Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic purposes. *Am. J. Hum. Genet.* 48:841–855.

Chakraborty, R. 1981. Expected number of rare alleles per locus in a sample and estimation of mutation rates. *Am. J. Hum. Genet.* 33:481–484.

Chakraborty, R. 1990a. Genetic profile of cosmopolitan populations: Effects of hidden subdivision. *Anthropol. Anz.* 48:313–331.

Chakraborty, R. 1990b. Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. *Am. J. Hum. Genet.* 47:87–94.

Chakraborty, R. 1991. Generalized occupancy problem and its applications in population genetics. In *Impact of Genetic Variation on Populations, Families, and Individuals*, C.F. Sing and C.L. Hanis, eds. New York: Oxford University Press (in press).

Chakraborty, R., and S.P. Daiger. 1991. Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah. *Hum. Biol.* 63:571–587.

Chakraborty, R., and R.C. Griffiths. 1982. Correlation of heterozygosity and the number of alleles in different frequency classes. *Theor. Popul. Biol.* 21:205–218.

Chakraborty, R., and L. Jin. 1991. Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. *Hum. Genet.* (in press).

Chakraborty, R., and R.J. Schwartz. 1990. Selective neutrality of surname distribution in an immigrant Indian community of Houston, Texas. *Am. J. Hum. Biol.* 2:1–15.

Chakraborty, R., P.A. Fuerst, and M. Nei. 1980. Statistical studies of protein polymorphism in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. *Genetics* 94:1039–1063.

Chakraborty, R., P.E. Smouse, and J.V. Neel. 1988. Population amalgamation and genetic variation: Observations on artificially agglomerated tribal populations of Central and South America. *Am. J. Hum. Genet.* 43:709–725.

Chakraborty, R., M. de Andrade, S.P. Daiger, and B. Budowle. 1991. Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. *Ann. Hum. Genet.* (in press).

Chakraborty, R., M. Fornage, R. Gueguen, and E. Boerwinkle. 1991. Population genetics of hypervariable loci: Analysis of PCR based VNTR polymorphism within a population. In *DNA Fingerprinting: Approaches and Applications*, T. Burke, G. Dolf, A. Jeffreys, and R. Wolff, eds. Basel: Birkhäuser, 127–143.

Clark, A.G. 1987. Neutrality tests of highly polymorphic restriction-fragment-length polymorphisms. *Am. J. Hum. Genet.* 41:948–956

Cohen, J.E. 1990. DNA fingerprinting for forensic identification: Potential effects of data interpretation of subpopulation heterogeneity and band number variability. *Am. J. Hum. Genet.* 46:358–368.

Deka, R., R. Chakraborty, and R.E. Ferrell. 1991. Population genetics of hypervariable loci in three ethnic groups. *Genomics* 11:83–92.

Devlin, B., N. Risch, and K. Roeder. 1990. No excess of homozygosity at loci used for DNA fingerprinting. *Science* 249:1416–1420.

Edwards, A., A. Civitello, H.A. Hammond, and C.T. Caskey. 1991. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am. J. Hum. Genet.* 49:746–756.

Emigh, T.H. 1980. A comparison of tests for Hardy-Weinberg equilibrium. *Biometrics* 36:627–642.

Ewens, W.J. 1972. The sampling theory of selectivity neutral alleles. *Theor. Popul. Biol.* 3:87–112.

Flint, J., A.J. Boyce, J.J. Martinson, and J.B. Clegg. 1989. Population bottlenecks in Polynesia revealed by minisatellites. *Hum. Genet.* 83:257–263.

Fuerst, P.A., R. Chakraborty, and M. Nei. 1977. Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. *Genetics* 86:455–483.

Jeffreys, A.J., R. Neumann, and V. Wilson. 1990. Repeat unit sequence variation in minisatellites: A novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* 60:473–485.

Jeffreys, A.J., V. Wilson, and S.L. Thein. 1985a. Hypervariable "minisatellite" regions of human DNA. *Nature* 314:67–73.

Jeffreys, A.J., V. Wilson, and S.L. Thein. 1985b. Individual-specific "fingerprints" of human DNA. *Nature* 316:76–79.

Jeffreys, A.J., N.J. Royle, V. Wilson, and Z. Wong. 1988. Spontaneous mutation rates to new alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* 332:278–281.

Johnson, N.L., and S. Kotz. 1969. *Distributions in Statistics: Discrete Distributions*. Boston, MA: Houghton Mifflin.

Lander, E. 1989a. DNA fingerprinting on trial. *Nature* 339:501–505.

Lander, E. 1989b. Population genetic considerations in the forensic use of DNA typing. In *DNA Technology and Forensic Science*, J. Ballantyne, G. Sensabaugh, and J. Witkowski, eds. Banbury Report 32. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 143–156.

Lander, E. 1991. Invited editorial: Research on DNA typing catching up with courtroom applications. *Am. J. Hum. Genet.* 48:819–823.

Ludwig, E.H., W. Friedl, and B.J. McCarthy. 1989. High-resolution analysis of a hypervariable region in the human apolipoprotein B gene. *Am. J. Hum. Genet.* 45:458–464.

Nakamura, Y., M. Leppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin, and R.L. White. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* 235:1616–1622.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583–590.

Nei, M., R. Chakraborty, and P.A. Fuerst. 1976. Infinite allele model with varying mutation rate. *Proc. Natl. Acad. Sci. USA* 73:4164–4168.

Odelberg, S.J., R. Platke, J.R. Eldridge, L. Ballard, P. O'Connell, Y. Nakamura, M. Leppert, J.-M. Lalouel, and R. White. 1989. Characterization of eight VNTR loci by agarose gel electrophoresis. *Genomics* 5:915–924.

Rao, C.R. 1957. Maximum likelihood estimation for the multinomial distributions. *Sankhyā* 18:139–148.

Rao, C.R. 1958. Maximum likelihood estimation for the multinomial distribution with infinite number of classes. *Sankhyā* 20:211–218.

Roychoudhury, A.K., and M. Nei. 1988. *Human Polymorphic Genes: World Distribution*. New York: Oxford University Press.

Ward, R.H., and C.F. Sing. 1970. A consideration of the power of the χ^2 test to detect inbreeding effects in natural populations. *Am. Natur.* 104:355–363.

American Journal of Human Genetics
Volume 47, No. 3, September 1990
(0504) 11.4

Population genetics of VNTR polymorphism in humans. R. Chakraborty and E. Boerwinkle. Genetics Centers, University of Texas Graduate School of Biomedical Sciences, Houston, Texas, USA.

Several regions of the human genome have been identified that exhibit a high degree of polymorphism due to a variable number of tandemly repeated (VNTR) DNA sequences. While the utility of these VNTR loci have been well publicized, their population genetic characteristics are poorly understood because of: (1) the large number of rare alleles; (2) presumptive high rate of "mutation"; and (3) possibility of incomplete resolution of similar size alleles. Understanding the population genetic characteristics is necessary for optimal utilization of these highly informative loci for gene mapping and genetic identification purposes.

We are studying the population genetic characteristics of several VNTR and microsatellite loci in a sample of 600 individuals, in addition to the data available in the published literature. Because of the large number of alleles and a relatively moderate sample size, standard tests of Hardy-Weinberg equilibrium (HWE) and gametic disequilibrium are inadequate, and alternative tests are proposed. These conservative tests are based on the exact sampling distributions of numbers of observed homozygotes and heterozygotes in a finite sample. When the typing method is such that all alleles are distinguishable (e.g., PCR typing of the 3' apolipoprotein-B VNTR), the genotype distribution fits the predictions of HWE, and no disequilibria are observed among unlinked VNTR loci. In the presence of incomplete resolution of alleles (e.g., D2S44 VNTR) significant departures from equilibrium expectations are observed. In addition, when complete resolution of alleles and genotypes is achieved, the classic mutation-drift (infinite allele) model accounts for the large amount of allelic diversity. The lack of complete resolution causes conspicuous discrepencies between the observed and expected allele frequency profiles. We propose a new model of forward-backward "mutational" changes that represents the population dynamics of VNTR allelic diversity more adequately. Our results indicate that the laboratory techniques applied for typing VNTR alleles plays a large role in dictating the population genetic features of VNTR loci. Ignoring this aspect may result in a wrong inference about population structure, consequently handicapping the optimal utility of these loci. (Research supported by grants GM-41399 and HL-40613 from the US National Institutes of Health).

with the current indication that there
is. Supported by Cystic Fibrosis

ie Country: European
on. I Casals, V.
C Lázaro, C Vázquez
Genetics Department,
Pau 08025 Barcelona
l cystic fibrosis (CF)
ons shows frequencies
British and North
Italian, and 30% in
have analysed the
r DNA polymorphism
508 mutation. The
grandparents were
were classified by
es of Basque origin
nd only 51% of non
he mutation. The
lation in respect to
ted at other genetic
istic differences of
the Indo-European
Basque population was
fore the arrival, in
. The high frequency
the Basque Country
hat the mutated CFTR
nts from the Middle
urope, and suggests
Ir[illegible] present within
nti[illegible]t, before this
ago. If this is the
gration diluted the
n Δ F508 by bringing
e.
de Investigaciones
Social" 90E1254 and

American Journal of Human Genetics
Volume 47, No. 3, September 1990
(0504) 11.4

Population genetics of VNTR polymorphism in humans. R. Chakraborty and E. Boerwinkle. Genetics Centers, University of Texas Graduate School of Biomedical Sciences, Houston, Texas, USA.

Several regions of the human genome have been identified that exhibit a high degree of polymorphism due to a variable number of tandemly repeated (VNTR) DNA sequences. While the utility of these VNTR loci have been well publicized, their population genetic characteristics are poorly understood because of: (1) the large number of rare alleles; (2) presumptive high rate of "mutation"; and (3) possibility of incomplete resolution of similar size alleles. Understanding the population genetic characteristics is necessary for optimal utilization of these highly informative loci for gene mapping and genetic identification purposes.

We are studying the population genetic characteristics of several VNTR and microsatellite loci in a sample of 600 individuals, in addition to the data available in the published literature. Because of the large number of alleles and a relatively moderate sample size, standard tests of Hardy-Weinberg equilibrium (HWE) and gametic disequilibrium are inadequate, and alternative tests are proposed. These conservative tests are based on the exact sampling distributions of numbers of observed homozygotes and heterozygotes in a finite sample. When the typing method is such that all alleles are distinguishable (e.g., PCR typing of the 3' apolipoprotein-B VNTR), the genotype distribution fits the predictions of HWE, and no disequilibria are observed among unlinked VNTR loci. In the presence of incomplete resolution of alleles (e.g., D2S44 VNTR) significant departures from equilibrium expectations are observed. In addition, when complete resolution of alleles and genotypes is achieved, the classic mutation-drift (infinite allele) model accounts for the large amount of allelic diversity. The lack of complete resolution causes conspicuous discrepencies between the observed and expected allele frequency profiles. We propose a new model of forward-backward "mutational" changes that represents the population dynamics of VNTR allelic diversity more adequately. Our results indicate that the laboratory techniques applied for typing VNTR alleles plays a large role in dictating the population genetic features of VNTR loci. Ignoring this aspect may result in a wrong inference about population structure, consequently handicapping the optimal utility of these loci. (Research supported by grants GM-41399 and HL-40613 from the US National Institutes of Health).

Polymorphisms at VNTR Loci Suggest Homogeneity of the White Population of Utah

RANAJIT CHAKRABORTY[1] AND STEPHEN P. DAIGER[2]

Abstract Apparent departure from equilibrium of genetic parameters measured for multiallelic single-locus markers such as VNTR (variable number of tandem repeat) loci has been suggested as evidence of underlying heterogeneity of the tested population. Using allele frequency distributions at eight VNTR loci from the white population of Utah, we show that the observed number of alleles and the gene diversity at each locus are congruent according to expectations of the neutral mutation model. This demonstrates the genetic homogeneity of the white population of Utah with reference to the allele (total and rare) frequency distribution at eight VNTR loci. The importance of such procedures is discussed in the context of using VNTR polymorphism data for forensic and medicolegal applications. Recommendations for reporting population data for hypervariable loci are also made to aid potential users in conducting similar analyses.

The discovery of hypervariable loci was one of the significant achievements of human genetics in the 1980s because of the impact of such polymorphisms on gene mapping, parentage ascertainment, and forensic identity determination. Although numerous hypervariable loci have been described [e.g., Wyman and White (1980), Bell et al. (1982), Proudfoot et al. (1982), Jeffreys et al. (1985, 1988), Nakamura et al. (1987), Wong et al. (1987), and Chimera et al. (1989)], large-scale population data on such loci are relatively scarce or not reported in a useful fashion. The alleles found at such loci vary in the number of tandem repeats of a specific DNA sequence. The characteristics of such variable number of tandem repeat (VNTR) loci include (1) a near continuum of allelic diversity according to molecular size in all populations tested, (2) low frequency of each individual allele, (3) high heterozygosity, and (4) mutation rates several-fold higher than those of traditional protein-coding loci.

[1]Center for Demographic and Population Genetics, Graduate School of Biomedical Sciences, University of Texas Health Science Center, PO Box 20334, Houston, TX 77225.

[2]Medical Genetics Center, Graduate School of Biomedical Sciences, University of Texas Health Science Center, PO Box 20334, Houston, TX 77225.

Human Biology, October 1991, Vol. 63, No. 5, pp. 571–587.

Although the practical utility of VNTR loci is enhanced enormously by these four characteristics, there are attendant limitations as well. For example, the nearly continuous allelic variation in molecular weight implies that the assessment of different alleles requires high-resolution laboratory protocols to resolve alleles that differ by only a single repeating unit. Because the length of the repeating unit may be small in relation to allele length [e.g., 30 bp, as in the case of the *D2S44* locus on human chromosome 2q; see Odelberg et al. (1989)], assessment can be a technically difficult task for large-scale population surveys. The current method of describing alleles by their length is also compromised by this limitation. Furthermore, because each individual allele occurs with low frequency in any population, reporting allele frequencies is cumbersome. Also, without a large sample size the estimated frequencies of such alleles are not generally reliable. With a few exceptions published data on polymorphisms at such loci generally are not suitable for conventional methods of genetic analyses, which can take into account the great extent of genetic variability displayed by such loci.

Recently, Lander (1989) and Cohen (1990) raised some of these issues and asked for establishment of standards for proper utilization of such data. They also called for appropriate statistical tests for validating several population genetic assumptions inherent in forensic and medicolegal applications of VNTR polymorphisms. Our purpose here is to demonstrate that the necessary theoretical tools for such analyses exist and that they can be routinely practiced, provided that laboratory protocols are adequately described and that the data are appropriate for such analyses. We use the allele frequency data reported by Odelberg et al. (1989) to show that the sampling theory of selectively neutral alleles (Ewens 1972; Chakraborty and Griffiths 1982) is applicable to VNTR loci. We find that the white population in Utah is genetically homogeneous according to these tests. Because the allelic diversity at VNTR loci is much greater than that at the traditional protein-coding loci, guidelines for minimum sample size for such loci are suggested. Some general recommendations regarding data requirements are also made for future applications of the present theory.

Materials and Methods

Odelberg et al. (1989) characterized eight VNTR loci (*D17S5, D2S44, D9S7, D14S13, D19S20, D16S83, D1S74*, and *D3S42*). Each locus exhibits a high degree of polymorphism in human populations. Allelic variability at these loci was detected by agarose gel electrophoresis. Alleles differing by a single repeating unit can be detected at the *D17S5, D2S44*, and *D9S7* loci. For the remaining five loci the resolving power

is lower [see Odelberg et al. (1989) for details]. By examining 78–151 unrelated individuals from the white population of Utah, Odelberg and co-workers detected allelic variation ranging from 13 to 67 distinct alleles per locus and locus heterozygosity ranging from 75.5% to 97.8%. The individuals included in Odelberg's study consisted of the unrelated parents or grandparents of 46 large, three-generation Utah Mormon pedigrees (part of the Utah panel of human linkage studies), and therefore the sample can be regarded as representative of the white population of Utah (Odelberg et al. 1989; White et al. 1985).

Results: Theory

Odelberg and co-workers tested the concordance of the observed genotypic proportions with expectations based on Hardy-Weinberg equilibrium (HWE) by combining the frequencies of all heterozygotes and homozygotes. Although demonstration of departures from HWE predictions are one approach for detecting population substructure, in the theory described in what follows we show an alternative approach in which only allelic counts are used to examine the genetic homogeneity of the population from such data. This suggested test avoids the problem of combining data over all alleles, necessitated by their small counts, and circumvents the problem of resolution of heterozygosity and homozygosity of nearly equal size alleles (Devlin et al. 1990).

Test Based on Total Number of Alleles. Under the premises of the neutral mutation hypothesis, when each mutation yields a new allele (infinite allele model; Wright 1949), the expected gene diversity, defined by H (Kimura and Crow 1964), in a population is given by

$$H = \theta/(1 + \theta), \tag{1}$$

where $\theta = 4N_e v$, in which N_e is the effective population size and v is the rate of mutation per locus per generation. In a sample of n genes drawn from a population, the expectation of the total number of alleles (k) is given by (Ewens 1972)

$$E(k) = \theta \sum_{i=0}^{n-1} (\theta + i)^{-1}. \tag{2}$$

Equations (1) and (2) are characteristics of a single random mating population that reached a steady state (equilibrium) under mutation-drift balance. Because the gene diversity (or heterozygosity in a random mating population) in Eq. (1) is equivalent to the complement of the

sum of the squares of allele frequencies and because the variable k can be observed directly, it is possible to examine whether the observed gene diversity and the number of alleles are congruent, satisfying Eq. (1) and (2). This can be done in two ways.

First, an estimate of θ is obtained from the gene diversity H calculated from observed allele frequencies. If $x_1, x_2, \ldots, x_k$ denote the observed frequencies of k different alleles in a sample of n genes drawn at random from a population, an unbiased estimate of gene diversity is given by (Nei 1978)

$$\hat{H} = n\left(1 - \sum x_i^2\right)/(n-1). \tag{3}$$

Although in previous works θ has been estimated from gene diversity as $t = \hat{H}/(1-\hat{H})$ [see, e.g., Fuerst et al. (1977), Chakraborty et al. (1988), and Chakraborty (1990a,b)], Zouros (1979) has shown that this estimator is biased in the upward direction, because

$$E[\hat{H}/(1-\hat{H})] \approx \theta\left[1 + \frac{2(1+\theta)}{(2+\theta)(3+\theta)}\right] \tag{4}$$

is larger than θ. Therefore a more reasonable estimator of θ from gene diversity can be obtained by equating the observed value of $t = \hat{H}/(1-\hat{H})$ to its expectation given by the right-hand side of Eq. (4). This is equivalent to solving the cubic equation

$$\theta^3 + (7-t)\theta^2 + (8-5t)\theta - 6t = 0 \tag{5}$$

for any observed value of $t = \hat{H}/(1-\hat{H})$. This equation can be solved by iteration, and it always provides one real root greater than 0. We denote the solution of this equation by $\hat{\theta}_H$, the gene diversity estimator of θ. A Taylor series approximation of Eq. (4) provides an approximate standard error of this estimate:

$$\text{s.e.}(\hat{\theta}_H) \approx \frac{(2+\theta)^2(3+\theta)^2 s(\hat{H})}{(1-\hat{H})^2(1+\theta)[(2+\theta)(3+\theta)(4+\theta) + 10(2+\theta) + 4]}, \tag{6}$$

in which $s(\hat{H})$ denotes the standard error of the estimate $\hat{H}$ [see Nei (1978) for the computational formula of $s(\hat{H})$]. When this estimated value of θ is substituted in Eq. (2) to compute the expected number of alleles, hidden subdivision within a population results in an *excess* of the observed number of alleles; that is, $k > E(k)$. The amount of excess depends on the number of subpopulations within the population and the degree of genetic divergence among them [see Chakraborty et al. (1988,

Figure 4)]. A formal test of the discrepancy between k and $E(k)$ can be obtained from the distribution of k.

Following Ewens (1972), the probability of observing k or more alleles in a sample of n genes can be written

$$P(k) = 1 - \sum_{r=1}^{k-1} \Gamma(\theta)\theta^r n! B(r, n)/[r!\Gamma(n+\theta)], \tag{7}$$

where $\Gamma(\cdot)$ is a gamma function and

$$B(r, n) = \sum \left(\prod_{i=1}^{r} n_i \right)^{-1}, \tag{8}$$

in which $n_1, n_2, \ldots, n_r$ are partitions of the integer n into r classes such that each n_i is greater than 0 and $n_1 + n_2 + \ldots + n_r = n$. The summation in this expression is over all permutations of the n_i [see the appendix by Stewart in Fuerst et al. (1977) and Chakraborty (1990b)]. When the estimate $\hat{\theta}_H$ is substituted for θ in Eq. (7), it allows a test of whether or not the observed value of k is too large for the given gene diversity.

Because the gene diversity estimator of θ may not be the most efficient one (Ewens 1972), alternatively one might ask whether the observed gene diversity is in congruence with its expectation when the estimate of θ is obtained from other features of the allele frequency distribution. Ewens (1972) showed that the right-hand side of Eq. (2), when equated to the observed number of alleles in a sample, provides the maximum likelihood estimate of θ (denoted by $\hat{\theta}_k$). Although a closed-form expression of this estimate does not exist, Chakraborty and Schwartz (1990) showed that this estimate can be obtained iteratively and that its approximate standard error also can be obtained from given values of k and n. Chakraborty (1990a,b) showed that, if the sample is drawn from a genetically heterogeneous population, the observed gene diversity $\hat{H}$ is generally smaller than its expectation based on the estimator $\hat{\theta}_k$ [i.e., when $E(\hat{H})$ is computed by substituting the estimator $\hat{\theta}_k$ for θ in Eq. (1)]. However, a formal test of the discrepancy between the observed and the expected H is tedious because an analytical sampling distribution of H is not available (Watterson 1978).

Test Based on Rare Alleles. A test for substructuring in a population can also include examination of the numbers of alleles in different gene frequency classes. Chakraborty et al. (1988) showed that in the presence of heterogeneity there is an excess of the total number of alleles, which is largely the result of an excess of rare alleles. More recently, Chakraborty (1990a,b) demonstrated that hidden heterogeneity can be

revealed through an excess of rare alleles irrespective of which estimator ($\hat{\theta}_H$ or $\hat{\theta}_k$) is chosen for computing the expected number of rare alleles. To conduct this test, with any defined criteria of rare alleles (such as alleles that occur with frequency q or less, with q generally taken as 0.01 or 0.05), one computes the observed number of rare alleles by summing (over r) the number of alleles (k_r), each of which occurs with r copies in a sample. Chakraborty and Griffiths (1982) showed that the expectation of k_r is

$$E(k_r) = \frac{\theta}{r} \frac{n!}{(n-r)!} \frac{\Gamma(n+\theta-r)}{\Gamma(n+\theta)}, \tag{9}$$

for $r = 1, 2, \ldots, n$, where θ is defined as in Eqs. (1) and (2). Furthermore, for rare alleles (i.e., when r is much smaller than n), the distribution of k_r is a Poisson distribution (Chakraborty and Griffiths 1982). Therefore the deviation of observed k_r from $E(k_r)$ can be tested by computing the cumulative Poisson probabilities simply from the knowledge of $E(k_r)$. This can be computed by substituting either of the two alternative estimators of θ mentioned earlier. The estimator $\hat{\theta}_k$ is preferred for this purpose because it is generally larger than the estimator $\hat{\theta}_H$, and our intent is to look for deviation in the direction $k_r > E(k_r)$, which is expected in the presence of hidden heterogeneity.

Results: Data Analysis

Table 1 shows the allele frequency distributions at the eight VNTR loci in the white population of Utah, as surveyed by Odelberg et al. (1989). Note that this table is a convenient form for presenting the basic data on allele frequency distribution even when the observed number of alleles is large. Of course, the specific allele designations cannot be represented in such a table. Nevertheless, such summary information is enough to compute any statistics of allele frequency distribution (e.g., number of alleles in each gene frequency class and total and expected homozygosity or heterozygosity). With this notation the observed estimate of gene diversity based on the allele frequency distribution ($\hat{H}$) becomes

$$\hat{H} = n\left(1 - \sum_{r=1}^{n} r^2 k_r / n^2\right) / (n-1), \tag{10}$$

where k_r is the number of alleles with r copies in a sample of n genes drawn from a population. The estimates of H obtained from Eq. (10) are slightly different from the ones reported by Odelberg et al. (1989, Table

Table 1. Allele Frequency Distribution at Eight VNTR Loci in the White Population of Utah

Number of Copies (r)	*Observed Number of Alleles* (k_r)							
	D17S5	*D2S44*	*D9S7*	*D14S13*	*D19S20*	*D16S83*	*D1S74*	*D3S42*
1	1	17	1	3	2	3	4	7
2	2	11	1	3	1	2	3	1
3	–	9	4	5	–	–	1	1
4	1	4	1	5	–	1	1	–
5	–	4	1	2	1	1	–	1
6	–	5	–	4	1	1	1	–
7	–	2	–	1	3	–	2	–
8	1	5	–	4	2	1	2	1
9	–	–	–	–	–	1	–	–
10	1	2	–	–	–	–	2	–
11	–	6	1	–	–	–	1	–
12	–	–	–	1	–	–	1	–
13	–	–	–	–	–	–	1	1
15	1	2	–	–	–	1	1	–
16	1	–	–	–	–	–	2	–
17	1	–	1	1	1	–	–	–
18	–	–	–	1	–	1	–	–
>18	5[a]	–	5[b]	–	2[c]	3[d]	–	3[e]
Total	14	67	16	30	13	15	22	15

a. Includes 3 alleles that have 25, 51, and 87 copies and 2 alleles with 32 copies each.
b. Includes 5 alleles that have 25, 33, 38, 42, and 71 copies.
c. Includes 2 alleles that have 38 and 61 copies.
d. Includes 1 allele with 30 copies, and 2 alleles with 27 copies each.
e. Includes 3 alleles that have 24, 36, and 70 copies.

3) because Odelberg et al. ignored the bias correction factor $n/(n-1)$ in their computations.

Table 2 provides the estimates of θ based on $\hat{H}$ and k and their standard errors. Note that the gene diversity estimators of θ are generally larger than the maximum likelihood estimators ($\hat{\theta}_k$) based on k; the only exception is the locus *D3S42*. This is so despite our use of a new bias-correcting algorithm to avoid the upward bias of the traditional gene diversity estimator t of θ. Approximate heterogeneity tests of the difference of the two estimators (data not shown here) suggest that at three loci, *D2S44*, *D14S13*, and *D1S74*, the estimate $\hat{\theta}_H$ is significantly larger than $\hat{\theta}_k$, whereas for the remaining five loci their difference is not significant. On the contrary, had this sample been drawn from a genetically heterogeneous population, we would have found the opposite, namely, the maximum likelihood estimators of θ ($\hat{\theta}_k$) larger than their respective gene diversity estimators ($\hat{\theta}_H$). This effect had been noticed by Chakraborty et al. (1988), Chakraborty and Schwartz (1990), and Chakraborty (1990a,b). When data from the 8 loci are combined, the estimators of θ from the average gene diversity per locus and the average

Table 2. Parameter Estimates from Eight VNTR Loci in the White Population of Utah

Locus	*Sample Size*[a] (n)	*Number of Alleles* (k)	*Gene Diversity*[b] (H)	*Estimates of* θ ($\pm$ *1 s.e.*)	
				θ_H	θ_k
D17S5	302	14	0.851 ± 0.011	4.70 ± 0.46	2.89 ± 0.89
D2S44	302	67	0.978 ± 0.002	43.13 ± 3.59	26.39 ± 4.06
D9S7	272	16	0.870 ± 0.010	5.07 ± 0.47	3.56 ± 1.03
D14S13	164	30	0.954 ± 0.005	19.05 ± 2.49	10.51 ± 2.37
D19S20	168	13	0.799 ± 0.021	3.16 ± 0.46	3.12 ± 1.02
D16S83	156	15	0.877 ± 0.010	5.94 ± 0.61	3.90 ± 1.20
D1S74	154	22	0.937 ± 0.005	13.40 ± 1.23	6.76 ± 1.75
D3S42	168	15	0.755 ± 0.024	2.39 ± 0.33	3.80 ± 1.16
Average	211	24	0.876 ± 0.010	5.93 ± 0.62	6.77 ± 1.61

a. Sample size refers to the number of genes sampled.
b. Calculated from observed allele frequencies using Eq. (10).

number of alleles per locus are fairly close (5.93 versus 6.77; heterogeneity $\chi^2 = 0.24$ with 1 d.f.; $p > 0.58$), suggesting that the assumption of genetic homogeneity of the sample is quite reasonable based on the pooled data on these 8 loci.

Table 3 compares the observed total number of alleles with the expected number based on the estimator $\hat{\theta}_H$ for each locus and for the average of the eight loci. For each locus, except *D3S42*, the observed number of alleles is smaller than its expectation, and hence no excess in the total number of alleles is demonstrated in this analysis. The average number of alleles per locus observed at these 8 loci (24) is in close agreement with its expectation [21.86; computed by using Eq. (2), where θ is estimated from the average heterozygosity, $\bar{H} = 0.876$, of the 8 loci, substituted into Eq. (5) and by using $n = 211$, the average number of genes per locus; see Table 2], suggesting that the sample is probably drawn from a homogeneous population. This is so because, in the presence of genetic heterogeneity, we would expect an excess number of alleles, and therefore the observed values of k would be larger than their expectations $E(k)$, with a probability [given by Eq. (7)] smaller than usual levels of statistical significance (0.05 or 0.01).

Table 4 shows the observed and expected (based on the estimator $\hat{\theta}_k$) gene diversity values. Also shown in this table are the proportions of the actual number of heterozygotes reported by Odelberg et al. (1989). Unlike the test of the number of alleles, no formal tests of significance can be done for these statistics because the sampling distribution of gene diversity is not known. Nevertheless, because hidden heterogeneity in a population results in an observed $\hat{H}$ smaller than $E(\hat{H})$ based on the estimator $\hat{\theta}_k$ (Chakraborty 1990a,b) and because this is not generally

Table 3. Observed and Expected Total Number of Alleles at Eight VNTR Loci in the White Population of Utah

	Number of Alleles		
Locus	*Observed*	*Expected*[a]	*Probability*[b]
D17S5	14	20.13 ± 3.87	0.963
D2S44	67	90.13 ± 7.20	>0.999
D9S7	16	20.81 ± 3.91	0.917
D14S13	30	43.56 ± 5.10	0.998
D19S20	13	13.13 ± 3.08	0.567
D16S83	15	20.13 ± 3.73	0.940
D1S74	22	34.46 ± 4.65	0.998
D3S42	15	10.73 ± 2.79	0.093
Average	24	21.86 ± 3.94	0.330

a. Based on the gene diversity estimator of θ ($\hat{\theta}_H$).
b. Probability that the number of alleles in a sample is equal to or less than the one observed, computed by substituting $\theta = \hat{\theta}_H$ in Eq. (7).

Table 4. Observed and Expected Gene Diversity at Eight VNTR Loci in the White Population of Utah

	Gene Diversity		*Observed Proportion of Heterozygotes*[c]
Locus	*Calculated*[a]	*Expected*[b]	
D17S5	0.851 ± 0.011	0.743 ± 0.059	0.861 ± 0.028
D2S44	0.978 ± 0.002	0.964 ± 0.005	0.947 ± 0.018
D9S7	0.870 ± 0.010	0.781 ± 0.050	0.824 ± 0.033
D14S13	0.954 ± 0.005	0.913 ± 0.018	0.854 ± 0.039
D19S20	0.799 ± 0.021	0.757 ± 0.060	0.810 ± 0.043
D16S83	0.877 ± 0.010	0.796 ± 0.050	0.897 ± 0.034
D1S74	0.937 ± 0.005	0.871 ± 0.029	0.872 ± 0.038
D3S42	0.755 ± 0.024	0.792 ± 0.050	0.786 ± 0.045
Average	0.876 ± 0.010	0.871 ± 0.024	0.856 ± 0.034

a. Calculated from the observed allele frequencies, using Eq. (10).
b. Based on the maximum likelihood estimator of θ ($\hat{\theta}_k$).
c. Obtained from the actual number of heterozygotes observed, reported in Table 3 of Odelberg et al. (1989).

seen in these computations, we surmise that the gene diversity test also suggests genetic homogeneity of the sampled population.

Table 5 presents a test of genetic equilibrium based on the frequency of rare alleles. Because the sample size (number of genes sampled) per locus varies between 156 and 302 in this survey, we used the criteria of 1% and 5% for defining rare alleles. For example, with $n = 302$, rare alleles with a 1% criterion represent those alleles that have counts of

Table 5. Observed and Expected Number of Rare Alleles at Eight VNTR Loci in the White Population of Utah

	Number of Alleles with Frequency ≤0.01			*Number of Alleles with Frequency ≤0.05*		
Locus	Observed	Expected[a]	*Probability*[b]	*Observed*	*Expected*[a]	*Probability*[b]
D17S5	3	5.25 ± 2.29	0.232	7	9.33 ± 3.06	0.286
D2S44	37	42.47 ± 6.52	0.226	67	63.46 ± 7.97	0.345
D9S7	2	5.27 ± 2.30	0.103	10	10.90 ± 3.30	0.472
D14S13	3	9.93 ± 3.15	0.011	27	24.32 ± 4.93	0.320
D19S20	2	3.08 ± 1.75	0.406	10	8.16 ± 2.86	0.304
D16S83	3	3.83 ± 1.96	0.467	8	9.63 ± 3.10	0.376
D1S74	4	6.52 ± 2.55	0.221	12	15.92 ± 3.99	0.199
D3S42	7	3.74 ± 1.93	0.085	11	9.84 ± 3.14	0.397
Average	7.75	9.80 ± 3.13	0.239	18.38	18.12 ± 4.26	0.551

a. Computed from Eq. (8) using the maximum likelihood estimator of θ ($\hat{\theta}_k$). Observed and expected numbers for the average reflect per locus estimates.

b. Probability of deviation from expectation; based on Poisson distribution. That is, these are probabilities of a value less than or equal to the observed value when the actual observed value is less than the expected, or of a value greater than or equal to the observed value when the actual observed value is greater than the expected.

3 or less in the sample. Although the observed numbers of such rare alleles can be obtained directly from the data in Table 1, the expected numbers are based on Eq. (9), summing over relevant r values (3 or less, for the given example), in which the estimate $\hat{\theta}_k$ is substituted for θ. Because the number of rare alleles follows a Poisson distribution, the congruence of the expected and observed numbers in this table is tested by computing the tail probability of a Poisson distribution. The probability column shows the exact significance values reached in each case. With the exception of the *D14S13* locus, the observed number of rare alleles is in statistical agreement with the expectations. For *D14S13*, with the 1% criterion of rare alleles, we find a deficiency of rare alleles. Therefore this test also suggests that there is no hidden substructuring in the population. As in the case of total number of alleles, hidden heterogeneity would have produced excess rare alleles.

Discussion and Conclusion

The analyses indicate that, even though the genetic variation revealed by the eight VNTR loci is extensive, there is no general indication of hidden subdivision within the white population of Utah. Jorde (1982) came to a similar conclusion by studying migration patterns of the founders of this population. Although our present study does not

provide a new anthropologic conclusion, several features of the analyses are of general significance in understanding the population genetic characteristics of hypervariable loci. First, unlike protein-coding loci, data from even a single VNTR locus can be subjected to this type of analysis because of the extensive number of alleles found at such loci. Second, although the mechanisms producing new variants in VNTR loci [e.g., nonhomologous sister chromatid exchange, unequal crossover, gene conversion, replication slippage; see Jeffreys et al. (1988)] are different from those producing variation in the protein-coding loci (mainly point mutation or small deletion), the infinite allele model of selectively neutral alleles applies equally well to population data for both types of polymorphic loci. This observation is also consistent with the pattern of allele frequency distributions at other VNTR and short tandem repeat (STR) loci found in recent population surveys (Deka et al. 1991; Edwards et al. 1991). Third, although the pooled data on the eight loci satisfy the predictions from the hypothesis of a single homogeneous population rather strikingly, we observe some deviations for the individual loci, but these deviations are in the direction *opposite* to the ones that can be caused by genetic heterogeneity within a population.

Our results apparently contradict Odelberg et al.'s (1989) analysis of deviations from HWE based on the comparison of observed and expected homozygosity and heterozygosity at these loci. Odelberg and co-workers found excess homozygosity at three loci (*D2S44*, *D14S13*, and *D1S74*), which might be construed as evidence of heterogeneity. A likely explanation for this apparent excess homozygosity is the technical difficulty of distinguishing closely spaced alleles on Southern gels (Devlin et al. 1990). We note that the same three loci exhibit significant differences in the two estimators of θ ($\hat{\theta}_H$ and $\hat{\theta}_k$). Hidden subdivision is not the cause of these deviations because, as noted earlier, the direction of deviation is opposite to what would be expected under heterogeneity.

To examine a possible cause of these discrepancies, we note some sample size considerations. Recall that at each VNTR locus a substantial number of alleles are detected, and almost all alleles occur at low frequencies in the population. Of the 192 alleles detected in this survey, there are only 3 alleles at these 8 loci that have frequencies exceeding 25%. Given this extensive allelic diversity, one might ask whether the available sample sizes are enough to capture all possible genotypes in these samples. With k alleles at a locus, there are $k(k+1)/2$ possible different genotypes, k of which are homozygotes, and $k(k-1)/2$ heterozygotes. Hence, if the sample size (number of individuals surveyed) is less than $k(k+1)/2$, several of these different genotypes will not be recorded in the sample. Exactly how many distinct genotypes were encountered in the survey was not reported by Odelberg et al. (1989). Nevertheless, under the assumption of HWE we can compute the minimum sample size required

to have all genotypes detected in the sample based on the observed allele frequencies. For example, if p_i is the true frequency of the ith allele at a locus, the probability that each of the $K = k(k+1)/2$ possible genotypes will be found in a sample of n individuals is

$$P = \sum \frac{n!}{\prod_{i=1}^{k} n_{ii}! \prod_{i>j=1}^{k}\prod^{k} n_{ij}!} \prod_{i=1}^{k}(p_i^2)^{n_{ii}} \prod_{i>j=1}^{k}\prod^{k}(2p_ip_j)^{n_{ij}}, \tag{11}$$

where the summation is over all n_{ii} and n_{ij} values such that none is 0 and such that they add to the total sample size (n). Although expression (11) is tedious to compute numerically when k and n are both large, it is easy to show that

$$P \geq 1 - \sum_{i=1}^{k}(1 - p_i^2)^n - \sum_{i>j=1}^{k}\sum^{k}(1 - 2p_ip_j)^n. \tag{12}$$

The right-hand side of expression (12) is at a maximum when all allele frequencies are equal, that is, when $p_i = 1/k$ for all i. Therefore a conservative estimate of the minimum sample size requirement for ensuring that all genotypes are represented in the sample with confidence $(1 - \alpha)$ is given by the inequality

$$1 - k(1 - k^{-2})^n - \frac{1}{2}k(k-1)(1 - 2k^{-2})^n \geq 1 - \alpha, \tag{13}$$

which reduces to

$$n \geq -k^2 \log_e \left[\frac{\sqrt{k^2 + 2\alpha k(k-1)} - k}{k(k-1)}\right]. \tag{14}$$

Table 6 shows the values of n for $\alpha = 0.10$, 0.05, and 0.01 that represent the minimum sample size required for the specific 8 loci in the present data. Note that, because the observed allele frequencies are not all equal at these loci, the actual sample size requirement may be even more stringent. Nevertheless, these computations indicate that with the available sample sizes it is unlikely that all possible genotypes are included in the data collected in this specific survey.

In terms of these minimum sample size requirements, it is clear that the smallest sample sizes are for the three loci *D2S44*, *D14S13*, and *D1S74*, which showed significant excess homozygosity in the analysis of Odelberg et al. (1989) and which exhibited significant differences in the two estimators of θ. Because the allele frequency distributions at

Table 6. Minimum Number of Individuals Needed to Detect all VNTR Genotypes in a Population Sample Given the Observed Allele Frequencies

Locus	*Observed Number of Alleles*	*Actual Sample Size*	*Minimum Sample Size Needed*[a]		
			$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
D17S5	14	151	977	1,107	1,421
D2S44	67	151	29,418	32,430	39,570
D9S7	16	136	1,311	1,483	1,890
D14S13	30	82	5,174	5,778	7,210
D19S20	13	84	830	944	1,213
D16S83	15	78	1,137	1,289	1,647
D1S74	22	78	2,632	2,957	3,727
D3S42	15	84	1,137	1,289	1,647
Average	24	106	3,183	3,570	4,486

a. Sample sizes in these computations refer to the number of individuals to be typed for each locus. The values of $\alpha = 0.10$, 0.05, and 0.01 represent 90%, 95%, and 99% confidence of being assured that all possible genotypes are represented in the sample.

these loci are in a direction opposite to that expected in the presence of hidden heterogeneity, we suspect that the observed excess homozygosity at these loci is an artifact of inadequate sample size alone and cannot be attributed to a Wahlund effect or population substructure.

We also note that, even though a formal power analysis of the tests proposed here is not yet available, previous applications of this method have succeeded in detecting hidden heterogeneity, evidenced by analysis of distributions of protein polymorphism (Chakraborty et al. 1988) and mitochondrial morphs (Chakraborty 1990a,b).

We should note that the present methodology rests on the assumption that the allelic diversity revealed by VNTR polymorphism obeys the infinite allele model. This approximation has been justified by Ohta (1986) and is borne out in the present analysis. Other population surveys on VNTR and STR loci also demonstrate that this model is appropriate for predicting allele frequency distributions at such loci (Budowle et al. 1991; Deka et al. 1991; Edwards et al. 1991).

Another characteristic feature of VNTR loci is also revealed through our present calculations. Because the parameter θ equals $4N_e\nu$, the relative values of the estimates of θ at the 8 VNTR loci when compared with protein-coding loci provide an indirect estimate of mutation rates for the VNTR loci. Mohrenweiser et al. (1987) estimated gene diversity at general protein-coding loci of 0.08 for the white population. By using the present gene diversity estimator of θ [solution of Eq. (5)], this leads to $\hat{\theta}_H = 0.0651$. A recent estimate of mutation rate at protein-coding loci is 1.1×10^{-5} per locus per generation (Chakraborty and Neel 1989). By

using the gene diversity estimates of θ from Table 2 and calibrating them against the estimate from protein-coding loci, we obtain mutation rates at the 8 VNTR loci ranging from 4.0×10^{-4} (for the locus *D3S42*) to 7.3×10^{-3} (for the locus *D2S44*) per locus per generation. The average mutation rate over the 8 loci is approximately 1.0×10^{-3} per locus per generation. In other words, these 8 VNTR loci are subject to mutational alterations at a rate 35–660 times (average 90 times) higher than protein-coding loci. These indirect estimates of the mutation rate at VNTR loci are almost an order of magnitude smaller than Jeffreys et al.'s (1988) estimate (average 0.012 per locus per generation for the 5 loci *D5S43, D12S11, D7S22, D7S21*, and *D1S7*) but are close to the estimate reported by Wolff et al. (1988). These observations are also in accordance with the indirect mutation rate estimates derived for 5 STR loci (average 6.1×10^{-5} per locus per generation; Edwards et al. 1991).

Although high mutation rates at VNTR loci are not relevant to the use of these loci in forensic identification problems, a frequency of 1 mutant per 1000 cases can have a substantial impact on parentage testing. Statistical methods to circumvent this problem for conventional genetic markers have been discussed elsewhere [e.g., Chakraborty and Schull (1976) and Chakraborty and Ryman (1981)]; these methods can be reformulated easily for VNTR loci.

Finally, we must close with a note of caution. Even though there are several claims of population subheterogeneity based on a simple demonstration of heterozygote deficiency at VNTR loci (Lander 1989; Cohen 1990), the results described here indicate that additional genetic parameters must be examined before it is valid to conclude that subheterogeneity exists. This is necessary because heterozygote deficiency alone can result from causes other than subheterogeneity, such as incomplete resolution of similar size alleles (Devlin et al. 1990) and loss of allelic bands resulting from very small (or very large) alleles (Skibinski et al. 1983). Lack of congruence between gene diversity and number of alleles can also be produced by evolutionary events [such as the bottleneck effect and the fluctuation of population size (Nei et al. 1975; Maruyama and Fuerst 1984, 1985; Watterson 1984)], which can be checked by examining the relationship of H and $E(k)$ with N_e [Eqs. (1) and (2)]. Ewens (1972) provided extensive tables of relationship between $E(k)$ and N_e (or θ) for an equilibrium population, and Nei et al. (1975), Maruyama and Fuerst (1984, 1985), and Watterson (1984) showed that, when a population goes through a severe bottleneck, the expected relationship between H and k is disturbed (in the direction of excess allele numbers compared with expectation) and the effect persists for a long time (of the order of the inverse of mutation rate).

We suggest that each specific survey should be subjected to tests similar to the ones suggested here. If the observed heterozygote deficiency

is due to factors other than population subheterogeneity, there should be a deficiency of allele numbers (total and rare) in contrast to the excess expected in the presence of population heterogeneity. Data requirements for such tests might include (1) a statement regarding the resolving power of the gel system used, in particular, whether even a single repeating unit can be detected; (2) tabulation of the observed allele frequency distribution in a form similar to our Table 1; (3) documentation of the exact number of distinct genotypes (both homozygotes and heterozygotes) observed in the sample; (4) an attempt to type a sufficient number of individuals so that a substantial fraction of all possible genotypes are detected; and (5) use of a well-defined (geographically or culturally homogeneous) sample population. Admittedly, several of these desiderata are compromised in the data analyzed here, and some are difficult to achieve in practice. For example, it is possible that allelic diversity will increase with an increase in sample size so that a substantial fraction of all genotypes can never be seen in any sample of feasible size. So long as only rare alleles (say, with a frequency less than 0.5%) are missed in a sample, the gene diversity estimator should be relatively insensitive to sample size. Our analysis of the Odelberg et al. (1989) data demonstrates that such analysis is feasible even when the requirements are met only approximately. Further empirical studies in appropriate populations are needed to capitalize fully on the forensic utility of VNTR popymorphisms.

Acknowledgments We thank A. Edwards for his comments on the manuscript. This work was supported by the National Institutes of Health under grant GM 41399 and by the National Institutes of Justice, Office of Justice Programs, US Department of Justice, under grant 90–IJ–CX–0038. Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice.

Received 17 December 1990; revision received 1 March 1991.

Literature Cited

Bell, G.I., M.J. Selby, and W.J. Rutter. 1982. The highly polymorphic region near the insulin gene is composed of simple tandemly repeating sequences. *Nature* 295:31–35.

Budowle, B., R. Chakraborty, A.M. Giusti, A.E. Eisenberg, and R.C. Allen. 1991. Analysis of the variable number of tandem repeats locus *D1S80* by the polymerase chain reaction followed by high resolution polyacrylamide gel electrophoresis. *Am. J. Hum. Genet.* 48:137–144.

Chakraborty, R. 1990a. Genetic profile of cosmopolitan populations: Effects of hidden subdivision. *Anthropol. Anz.* 48:313–331.

Chakraborty, R. 1990b. Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. *Am. J. Hum. Genet.* 47:87–94.

Chakraborty, R., and R.C. Griffiths. 1982. Correlation of heterozygosity and number of alleles in different frequency classes. *Theor. Popul. Biol.* 21:205–218.

Chakraborty, R., and J.V. Neel. 1989. Description and validation of a method for simultaneous estimation of effective population size and mutation rate from human population data. *Proc. Natl. Acad. Sci. USA* 86:9407–9411.

Chakraborty, R., and N. Ryman. 1981. Use of odds of paternity computations in determining the reliability of single exclusions in paternity testing. *Hum. Hered.* 31:363–369.

Chakraborty, R., and W.J. Schull. 1976. A note on the distribution of the number of exclusions to be expected in paternity testing. *Am. J. Hum. Genet.* 28:615–618.

Chakraborty, R., and R.J. Schwartz. 1990. Selective neutrality of surname distribution in an immigrant Indian community of Houston, Texas. *Am. J. Hum. Biol.* 2:1–15.

Chakraborty, R., P.E. Smouse, and J.V. Neel. 1988. Population amalgamation and genetic variation: Observations on artificially agglomerated tribal populations of Central and South America. *Am. J. Hum. Genet.* 43:709–725.

Chimera, J.A., C.R. Harris, and M. Litt. 1989. Population genetics of the highly polymorphic locus *D16S7* and its use in paternity evaluation. *Am. J. Hum. Genet.* 45:926–931.

Cohen, J.E. 1990. DNA fingerprinting for forensic identification: Potential effects of data interpretation of subpopulation heterogeneity and band number variability. *Am. J. Hum. Genet.* 46:358–368.

Deka, R., R. Chakraborty, and R.E. Ferrell. 1991. Population genetics of hypervariable loci in three ethnic groups. *Genomics* (in press).

Devlin, B., N. Risch, and K. Roeder. 1990. No excess of homozygosity at loci used for DNA fingerprinting. *Science* 249:1416–1420.

Edwards, A., H.A. Hammond, C.T. Caskey, L. Jin, and R. Chakraborty. 1991. Population genetics of trimeric and tetrameric tandem repeats in four human ethnic groups. *Genomics* (in press).

Ewens, W.J. 1972. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3:87–112.

Fuerst, P.A., R. Chakraborty, and M. Nei. 1977. Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. *Genetics* 86:455–483.

Jeffreys, A.J., V. Wilson, and S.L. Thein. 1985. Hypervariable "minisatellite" regions of human DNA. *Nature* 314:67–73.

Jeffreys, A.J., N.J. Royle, V. Wilson, and Z. Wong. 1988. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* 332:278–281.

Jorde, L.B. 1982. The genetic structure of the Utah Mormons: Migration analysis. *Hum. Biol.* 54:583–597.

Kimura, M., and J.F. Crow. 1964. The number of alleles that can be maintained in a finite population. *Genetics* 49:725–738.

Lander, E.S. 1989. DNA fingerprinting on trial. *Nature* 339:501–505.

Maruyama, T., and P.A. Fuerst. 1984. Population bottlenecks and nonequilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. *Genetics* 108:745–763.

Maruyama, T., and P.A. Fuerst. 1985. Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in small population that was formed by a recent bottleneck. *Genetics* 111:691–703.

Mohrenweiser, H.W., K.H. Wurzinger, and J.V. Neel. 1987. Frequency and distribution of rare electrophoretic mobility variants in a population of newborns in Ann Arbor, Michigan. *Ann Hum. Genet.* 51:303–316.

Nakamura, Y., M. Leppert, P. O'Connell et al. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* 235:1616–1622.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583–590.

Nei, M., T. Maruyama, and R. Chakraborty. 1975. The bottleneck effect and genetic variability in populations. *Evolution* 29:1–10.

Odelberg, S.J., R. Platke, J.R. Eldridge, L. Ballard, P. O'Connell, Y. Nakamura, M. Leppert, J.-M. Lalouel, and R. White. 1989. Characterization of eight VNTR loci by agarose gel electrophoresis. *Genomics* 5:915–924.

Ohta, T. 1986. Actual number of alleles contained in a multigene family. *Genet. Res.* 48:119–123.

Proudfoot, N.J., A. Gil, and T. Maniatis. 1982. The structure of the human zeta-globin gene and a closely linked, nearly identical pseudogene. *Cell* 31:553–563.

Skibinski, D.O.F., J.A. Beardmore, and T.F. Cross. 1983. Aspects of the population genetics of *Mytilus* (Mytilidae; Mollusca) in the British Isles. *Biol. J. Linn. Soc.* 19:137–183.

Watterson, G.A. 1978. The homozygosity test of neutrality. *Genetics* 88:405–417.

Watterson, G.A. 1984. Allele frequencies after a bottleneck. *Theor. Popul. Biol.* 26:387–407.

White, R., M. Leppert, D.T. Bishop, D. Barker, J. Berkowitz, C. Brown, P. Callahan, T. Holm, and L. Jerominski. 1985. Construction of linkage maps with DNA markers for human chromosomes. *Nature* 313:101–105.

Wolff, R.K., Y. Nakamura, and R. White. 1988. Molecular characterization of a spontaneously generated new allele at a VNTR locus: No exchange of flanking DNA sequences. *Genomics* 3:347–351.

Wong, Z., V. Wilson, I. Patel, S. Povey, and A.J. Jeffreys. 1987. Characterization of a panel of highly variable minisatellites cloned from human DNA. *Ann. Hum. Genet.* 51:269–288.

Wright, S. 1949. Genetics of populations. In *Encyclopaedia Britannica*, 14th ed., v. 10, 111–112.

Wyman, A.R., and R. White. 1980. A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* 77:6754–6758.

Zouros, E. 1979. Mutation rates, population sizes, and amounts of electrophoretic variation of enzyme loci in natural populations. *Genetics* 92:623–646.

Polymorphisms at VNTR Loci Suggest Homogeneity of the White Population of Utah

RANAJIT CHAKRABORTY[1] AND STEPHEN P. DAIGER[2]

Abstract Apparent departure from equilibrium of genetic parameters measured for multiallelic single-locus markers such as VNTR (variable number of tandem repeat) loci has been suggested as evidence of underlying heterogeneity of the tested population. Using allele frequency distributions at eight VNTR loci from the white population of Utah, we show that the observed number of alleles and the gene diversity at each locus are congruent according to expectations of the neutral mutation model. This demonstrates the genetic homogeneity of the white population of Utah with reference to the allele (total and rare) frequency distribution at eight VNTR loci. The importance of such procedures is discussed in the context of using VNTR polymorphism data for forensic and medicolegal applications. Recommendations for reporting population data for hypervariable loci are also made to aid potential users in conducting similar analyses.

The discovery of hypervariable loci was one of the significant achievements of human genetics in the 1980s because of the impact of such polymorphisms on gene mapping, parentage ascertainment, and forensic identity determination. Although numerous hypervariable loci have been described [e.g., Wyman and White (1980), Bell et al. (1982), Proudfoot et al. (1982), Jeffreys et al. (1985, 1988), Nakamura et al. (1987), Wong et al. (1987), and Chimera et al. (1989)], large-scale population data on such loci are relatively scarce or not reported in a useful fashion. The alleles found at such loci vary in the number of tandem repeats of a specific DNA sequence. The characteristics of such variable number of tandem repeat (VNTR) loci include (1) a near continuum of allelic diversity according to molecular size in all populations tested, (2) low frequency of each individual allele, (3) high heterozygosity, and (4) mutation rates several-fold higher than those of traditional protein-coding loci.

[1]Center for Demographic and Population Genetics, Graduate School of Biomedical Sciences, University of Texas Health Science Center, PO Box 20334, Houston, TX 77225.

[2]Medical Genetics Center, Graduate School of Biomedical Sciences, University of Texas Health Science Center, PO Box 20334, Houston, TX 77225.

Human Biology, October 1991, Vol. 63, No. 5, pp. 571–587.

Although the practical utility of VNTR loci is enhanced enormously by these four characteristics, there are attendant limitations as well. For example, the nearly continuous allelic variation in molecular weight implies that the assessment of different alleles requires high-resolution laboratory protocols to resolve alleles that differ by only a single repeating unit. Because the length of the repeating unit may be small in relation to allele length [e.g., 30 bp, as in the case of the *D2S44* locus on human chromosome 2q; see Odelberg et al. (1989)], assessment can be a technically difficult task for large-scale population surveys. The current method of describing alleles by their length is also compromised by this limitation. Furthermore, because each individual allele occurs with low frequency in any population, reporting allele frequencies is cumbersome. Also, without a large sample size the estimated frequencies of such alleles are not generally reliable. With a few exceptions published data on polymorphisms at such loci generally are not suitable for conventional methods of genetic analyses, which can take into account the great extent of genetic variability displayed by such loci.

Recently, Lander (1989) and Cohen (1990) raised some of these issues and asked for establishment of standards for proper utilization of such data. They also called for appropriate statistical tests for validating several population genetic assumptions inherent in forensic and medicolegal applications of VNTR polymorphisms. Our purpose here is to demonstrate that the necessary theoretical tools for such analyses exist and that they can be routinely practiced, provided that laboratory protocols are adequately described and that the data are appropriate for such analyses. We use the allele frequency data reported by Odelberg et al. (1989) to show that the sampling theory of selectively neutral alleles (Ewens 1972; Chakraborty and Griffiths 1982) is applicable to VNTR loci. We find that the white population in Utah is genetically homogeneous according to these tests. Because the allelic diversity at VNTR loci is much greater than that at the traditional protein-coding loci, guidelines for minimum sample size for such loci are suggested. Some general recommendations regarding data requirements are also made for future applications of the present theory.

Materials and Methods

Odelberg et al. (1989) characterized eight VNTR loci (*D17S5, D2S44, D9S7, D14S13, D19S20, D16S83, D1S74*, and *D3S42*). Each locus exhibits a high degree of polymorphism in human populations. Allelic variability at these loci was detected by agarose gel electrophoresis. Alleles differing by a single repeating unit can be detected at the *D17S5, D2S44*, and *D9S7* loci. For the remaining five loci the resolving power

is lower [see Odelberg et al. (1989) for details]. By examining 78–151 unrelated individuals from the white population of Utah, Odelberg and co-workers detected allelic variation ranging from 13 to 67 distinct alleles per locus and locus heterozygosity ranging from 75.5% to 97.8%. The individuals included in Odelberg's study consisted of the unrelated parents or grandparents of 46 large, three-generation Utah Mormon pedigrees (part of the Utah panel of human linkage studies), and therefore the sample can be regarded as representative of the white population of Utah (Odelberg et al. 1989; White et al. 1985).

Results: Theory

Odelberg and co-workers tested the concordance of the observed genotypic proportions with expectations based on Hardy-Weinberg equilibrium (HWE) by combining the frequencies of all heterozygotes and homozygotes. Although demonstration of departures from HWE predictions are one approach for detecting population substructure, in the theory described in what follows we show an alternative approach in which only allelic counts are used to examine the genetic homogeneity of the population from such data. This suggested test avoids the problem of combining data over all alleles, necessitated by their small counts, and circumvents the problem of resolution of heterozygosity and homozygosity of nearly equal size alleles (Devlin et al. 1990).

Test Based on Total Number of Alleles. Under the premises of the neutral mutation hypothesis, when each mutation yields a new allele (infinite allele model; Wright 1949), the expected gene diversity, defined by H (Kimura and Crow 1964), in a population is given by

$$H = \theta/(1 + \theta), \tag{1}$$

where $\theta = 4N_e v$, in which N_e is the effective population size and v is the rate of mutation per locus per generation. In a sample of n genes drawn from a population, the expectation of the total number of alleles (k) is given by (Ewens 1972)

$$E(k) = \theta \sum_{i=0}^{n-1} (\theta + i)^{-1}. \tag{2}$$

Equations (1) and (2) are characteristics of a single random mating population that reached a steady state (equilibrium) under mutation-drift balance. Because the gene diversity (or heterozygosity in a random mating population) in Eq. (1) is equivalent to the complement of the

sum of the squares of allele frequencies and because the variable k can be observed directly, it is possible to examine whether the observed gene diversity and the number of alleles are congruent, satisfying Eq. (1) and (2). This can be done in two ways.

First, an estimate of θ is obtained from the gene diversity H calculated from observed allele frequencies. If $x_1, x_2, \ldots, x_k$ denote the observed frequencies of k different alleles in a sample of n genes drawn at random from a population, an unbiased estimate of gene diversity is given by (Nei 1978)

$$\hat{H} = n\left(1 - \sum x_i^2\right)/(n-1). \tag{3}$$

Although in previous works θ has been estimated from gene diversity as $t = \hat{H}/(1-\hat{H})$ [see, e.g., Fuerst et al. (1977), Chakraborty et al. (1988), and Chakraborty (1990a,b)], Zouros (1979) has shown that this estimator is biased in the upward direction, because

$$E[\hat{H}/(1-\hat{H})] \approx \theta\left[1 + \frac{2(1+\theta)}{(2+\theta)(3+\theta)}\right] \tag{4}$$

is larger than θ. Therefore a more reasonable estimator of θ from gene diversity can be obtained by equating the observed value of $t = \hat{H}/(1-\hat{H})$ to its expectation given by the right-hand side of Eq. (4). This is equivalent to solving the cubic equation

$$\theta^3 + (7-t)\theta^2 + (8-5t)\theta - 6t = 0 \tag{5}$$

for any observed value of $t = \hat{H}/(1-\hat{H})$. This equation can be solved by iteration, and it always provides one real root greater than 0. We denote the solution of this equation by $\hat{\theta}_H$, the gene diversity estimator of θ. A Taylor series approximation of Eq. (4) provides an approximate standard error of this estimate:

$$\text{s.e.}(\hat{\theta}_H) \approx \frac{(2+\theta)^2(3+\theta)^2 s(\hat{H})}{(1-\hat{H})^2(1+\theta)[(2+\theta)(3+\theta)(4+\theta) + 10(2+\theta) + 4]}, \tag{6}$$

in which $s(\hat{H})$ denotes the standard error of the estimate $\hat{H}$ [see Nei (1978) for the computational formula of $s(\hat{H})$]. When this estimated value of θ is substituted in Eq. (2) to compute the expected number of alleles, hidden subdivision within a population results in an *excess* of the observed number of alleles; that is, $k > E(k)$. The amount of excess depends on the number of subpopulations within the population and the degree of genetic divergence among them [see Chakraborty et al. (1988,

Figure 4)]. A formal test of the discrepancy between k and $E(k)$ can be obtained from the distribution of k.

Following Ewens (1972), the probability of observing k or more alleles in a sample of n genes can be written

$$P(k) = 1 - \sum_{r=1}^{k-1} \Gamma(\theta)\theta^r n! B(r, n)/[r!\Gamma(n+\theta)], \tag{7}$$

where $\Gamma(\cdot)$ is a gamma function and

$$B(r, n) = \sum \left(\prod_{i=1}^{r} n_i \right)^{-1}, \tag{8}$$

in which $n_1, n_2, \ldots, n_r$ are partitions of the integer n into r classes such that each n_i is greater than 0 and $n_1 + n_2 + \ldots + n_r = n$. The summation in this expression is over all permutations of the n_i [see the appendix by Stewart in Fuerst et al. (1977) and Chakraborty (1990b)]. When the estimate $\hat{\theta}_H$ is substituted for θ in Eq. (7), it allows a test of whether or not the observed value of k is too large for the given gene diversity.

Because the gene diversity estimator of θ may not be the most efficient one (Ewens 1972), alternatively one might ask whether the observed gene diversity is in congruence with its expectation when the estimate of θ is obtained from other features of the allele frequency distribution. Ewens (1972) showed that the right-hand side of Eq. (2), when equated to the observed number of alleles in a sample, provides the maximum likelihood estimate of θ (denoted by $\hat{\theta}_k$). Although a closed-form expression of this estimate does not exist, Chakraborty and Schwartz (1990) showed that this estimate can be obtained iteratively and that its approximate standard error also can be obtained from given values of k and n. Chakraborty (1990a,b) showed that, if the sample is drawn from a genetically heterogeneous population, the observed gene diversity $\hat{H}$ is generally smaller than its expectation based on the estimator $\hat{\theta}_k$ [i.e., when $E(\hat{H})$ is computed by substituting the estimator $\hat{\theta}_k$ for θ in Eq. (1)]. However, a formal test of the discrepancy between the observed and the expected H is tedious because an analytical sampling distribution of H is not available (Watterson 1978).

Test Based on Rare Alleles. A test for substructuring in a population can also include examination of the numbers of alleles in different gene frequency classes. Chakraborty et al. (1988) showed that in the presence of heterogeneity there is an excess of the total number of alleles, which is largely the result of an excess of rare alleles. More recently, Chakraborty (1990a,b) demonstrated that hidden heterogeneity can be

revealed through an excess of rare alleles irrespective of which estimator ($\hat{\theta}_H$ or $\hat{\theta}_k$) is chosen for computing the expected number of rare alleles. To conduct this test, with any defined criteria of rare alleles (such as alleles that occur with frequency q or less, with q generally taken as 0.01 or 0.05), one computes the observed number of rare alleles by summing (over r) the number of alleles (k_r), each of which occurs with r copies in a sample. Chakraborty and Griffiths (1982) showed that the expectation of k_r is

$$E(k_r) = \frac{\theta}{r} \frac{n!}{(n-r)!} \frac{\Gamma(n+\theta-r)}{\Gamma(n+\theta)}, \tag{9}$$

for $r = 1, 2, \ldots, n$, where θ is defined as in Eqs. (1) and (2). Furthermore, for rare alleles (i.e., when r is much smaller than n), the distribution of k_r is a Poisson distribution (Chakraborty and Griffiths 1982). Therefore the deviation of observed k_r from $E(k_r)$ can be tested by computing the cumulative Poisson probabilities simply from the knowledge of $E(k_r)$. This can be computed by substituting either of the two alternative estimators of θ mentioned earlier. The estimator $\hat{\theta}_k$ is preferred for this purpose because it is generally larger than the estimator $\hat{\theta}_H$, and our intent is to look for deviation in the direction $k_r > E(k_r)$, which is expected in the presence of hidden heterogeneity.

Results: Data Analysis

Table 1 shows the allele frequency distributions at the eight VNTR loci in the white population of Utah, as surveyed by Odelberg et al. (1989). Note that this table is a convenient form for presenting the basic data on allele frequency distribution even when the observed number of alleles is large. Of course, the specific allele designations cannot be represented in such a table. Nevertheless, such summary information is enough to compute any statistics of allele frequency distribution (e.g., number of alleles in each gene frequency class and total and expected homozygosity or heterozygosity). With this notation the observed estimate of gene diversity based on the allele frequency distribution ($\hat{H}$) becomes

$$\hat{H} = n\left(1 - \sum_{r=1}^{n} r^2 k_r / n^2\right) / (n-1), \tag{10}$$

where k_r is the number of alleles with r copies in a sample of n genes drawn from a population. The estimates of H obtained from Eq. (10) are slightly different from the ones reported by Odelberg et al. (1989, Table

Table 1. Allele Frequency Distribution at Eight VNTR Loci in the White Population of Utah

Number of Copies (r)	*Observed Number of Alleles* (k_r)							
	D17S5	*D2S44*	*D9S7*	*D14S13*	*D19S20*	*D16S83*	*D1S74*	*D3S42*
1	1	17	1	3	2	3	4	7
2	2	11	1	3	1	2	3	1
3	–	9	4	5	–	–	1	1
4	1	4	1	5	–	1	1	–
5	–	4	1	2	1	1	–	1
6	–	5	–	4	1	1	1	–
7	–	2	–	1	3	–	2	–
8	1	5	–	4	2	1	2	1
9	–	–	–	–	–	1	–	–
10	1	2	–	–	–	–	2	–
11	–	6	1	–	–	–	1	–
12	–	–	–	1	–	–	1	–
13	–	–	–	–	–	–	1	1
15	1	2	–	–	–	1	1	–
16	1	–	–	–	–	–	2	–
17	1	–	1	1	1	–	–	–
18	–	–	–	1	–	1	–	–
>18	5[a]	–	5[b]	–	2[c]	3[d]	–	3[e]
Total	14	67	16	30	13	15	22	15

a. Includes 3 alleles that have 25, 51, and 87 copies and 2 alleles with 32 copies each.
b. Includes 5 alleles that have 25, 33, 38, 42, and 71 copies.
c. Includes 2 alleles that have 38 and 61 copies.
d. Includes 1 allele with 30 copies, and 2 alleles with 27 copies each.
e. Includes 3 alleles that have 24, 36, and 70 copies.

3) because Odelberg et al. ignored the bias correction factor $n/(n-1)$ in their computations.

Table 2 provides the estimates of θ based on $\hat{H}$ and k and their standard errors. Note that the gene diversity estimators of θ are generally larger than the maximum likelihood estimators ($\hat{\theta}_k$) based on k; the only exception is the locus *D3S42*. This is so despite our use of a new bias-correcting algorithm to avoid the upward bias of the traditional gene diversity estimator t of θ. Approximate heterogeneity tests of the difference of the two estimators (data not shown here) suggest that at three loci, *D2S44*, *D14S13*, and *D1S74*, the estimate $\hat{\theta}_H$ is significantly larger than $\hat{\theta}_k$, whereas for the remaining five loci their difference is not significant. On the contrary, had this sample been drawn from a genetically heterogeneous population, we would have found the opposite, namely, the maximum likelihood estimators of θ ($\hat{\theta}_k$) larger than their respective gene diversity estimators ($\hat{\theta}_H$). This effect had been noticed by Chakraborty et al. (1988), Chakraborty and Schwartz (1990), and Chakraborty (1990a,b). When data from the 8 loci are combined, the estimators of θ from the average gene diversity per locus and the average

Table 2. Parameter Estimates from Eight VNTR Loci in the White Population of Utah

Locus	*Sample Size*[a] (n)	*Number of Alleles* (k)	*Gene Diversity*[b] (H)	*Estimates of* θ ($\pm$ *1 s.e.*)	
				θ_H	θ_k
D17S5	302	14	0.851 ± 0.011	4.70 ± 0.46	2.89 ± 0.89
D2S44	302	67	0.978 ± 0.002	43.13 ± 3.59	26.39 ± 4.06
D9S7	272	16	0.870 ± 0.010	5.07 ± 0.47	3.56 ± 1.03
D14S13	164	30	0.954 ± 0.005	19.05 ± 2.49	10.51 ± 2.37
D19S20	168	13	0.799 ± 0.021	3.16 ± 0.46	3.12 ± 1.02
D16S83	156	15	0.877 ± 0.010	5.94 ± 0.61	3.90 ± 1.20
D1S74	154	22	0.937 ± 0.005	13.40 ± 1.23	6.76 ± 1.75
D3S42	168	15	0.755 ± 0.024	2.39 ± 0.33	3.80 ± 1.16
Average	211	24	0.876 ± 0.010	5.93 ± 0.62	6.77 ± 1.61

a. Sample size refers to the number of genes sampled.
b. Calculated from observed allele frequencies using Eq. (10).

number of alleles per locus are fairly close (5.93 versus 6.77; heterogeneity $\chi^2 = 0.24$ with 1 d.f.; $p > 0.58$), suggesting that the assumption of genetic homogeneity of the sample is quite reasonable based on the pooled data on these 8 loci.

Table 3 compares the observed total number of alleles with the expected number based on the estimator $\hat{\theta}_H$ for each locus and for the average of the eight loci. For each locus, except *D3S42*, the observed number of alleles is smaller than its expectation, and hence no excess in the total number of alleles is demonstrated in this analysis. The average number of alleles per locus observed at these 8 loci (24) is in close agreement with its expectation [21.86; computed by using Eq. (2), where θ is estimated from the average heterozygosity, $\bar{H} = 0.876$, of the 8 loci, substituted into Eq. (5) and by using $n = 211$, the average number of genes per locus; see Table 2], suggesting that the sample is probably drawn from a homogeneous population. This is so because, in the presence of genetic heterogeneity, we would expect an excess number of alleles, and therefore the observed values of k would be larger than their expectations $E(k)$, with a probability [given by Eq. (7)] smaller than usual levels of statistical significance (0.05 or 0.01).

Table 4 shows the observed and expected (based on the estimator $\hat{\theta}_k$) gene diversity values. Also shown in this table are the proportions of the actual number of heterozygotes reported by Odelberg et al. (1989). Unlike the test of the number of alleles, no formal tests of significance can be done for these statistics because the sampling distribution of gene diversity is not known. Nevertheless, because hidden heterogeneity in a population results in an observed $\hat{H}$ smaller than $E(\hat{H})$ based on the estimator $\hat{\theta}_k$ (Chakraborty 1990a,b) and because this is not generally

Table 3. Observed and Expected Total Number of Alleles at Eight VNTR Loci in the White Population of Utah

	Number of Alleles		
Locus	*Observed*	*Expected*[a]	*Probability*[b]
D17S5	14	20.13 ± 3.87	0.963
D2S44	67	90.13 ± 7.20	>0.999
D9S7	16	20.81 ± 3.91	0.917
D14S13	30	43.56 ± 5.10	0.998
D19S20	13	13.13 ± 3.08	0.567
D16S83	15	20.13 ± 3.73	0.940
D1S74	22	34.46 ± 4.65	0.998
D3S42	15	10.73 ± 2.79	0.093
Average	24	21.86 ± 3.94	0.330

a. Based on the gene diversity estimator of θ ($\hat{\theta}_H$).
b. Probability that the number of alleles in a sample is equal to or less than the one observed, computed by substituting $\theta = \hat{\theta}_H$ in Eq. (7).

Table 4. Observed and Expected Gene Diversity at Eight VNTR Loci in the White Population of Utah

	Gene Diversity		*Observed Proportion of Heterozygotes*[c]
Locus	*Calculated*[a]	*Expected*[b]	
D17S5	0.851 ± 0.011	0.743 ± 0.059	0.861 ± 0.028
D2S44	0.978 ± 0.002	0.964 ± 0.005	0.947 ± 0.018
D9S7	0.870 ± 0.010	0.781 ± 0.050	0.824 ± 0.033
D14S13	0.954 ± 0.005	0.913 ± 0.018	0.854 ± 0.039
D19S20	0.799 ± 0.021	0.757 ± 0.060	0.810 ± 0.043
D16S83	0.877 ± 0.010	0.796 ± 0.050	0.897 ± 0.034
D1S74	0.937 ± 0.005	0.871 ± 0.029	0.872 ± 0.038
D3S42	0.755 ± 0.024	0.792 ± 0.050	0.786 ± 0.045
Average	0.876 ± 0.010	0.871 ± 0.024	0.856 ± 0.034

a. Calculated from the observed allele frequencies, using Eq. (10).
b. Based on the maximum likelihood estimator of θ ($\hat{\theta}_k$).
c. Obtained from the actual number of heterozygotes observed, reported in Table 3 of Odelberg et al. (1989).

seen in these computations, we surmise that the gene diversity test also suggests genetic homogeneity of the sampled population.

Table 5 presents a test of genetic equilibrium based on the frequency of rare alleles. Because the sample size (number of genes sampled) per locus varies between 156 and 302 in this survey, we used the criteria of 1% and 5% for defining rare alleles. For example, with $n = 302$, rare alleles with a 1% criterion represent those alleles that have counts of

Table 5. Observed and Expected Number of Rare Alleles at Eight VNTR Loci in the White Population of Utah

	Number of Alleles with Frequency ≤0.01			*Number of Alleles with Frequency ≤0.05*		
Locus	Observed	Expected[a]	*Probability*[b]	*Observed*	*Expected*[a]	*Probability*[b]
D17S5	3	5.25 ± 2.29	0.232	7	9.33 ± 3.06	0.286
D2S44	37	42.47 ± 6.52	0.226	67	63.46 ± 7.97	0.345
D9S7	2	5.27 ± 2.30	0.103	10	10.90 ± 3.30	0.472
D14S13	3	9.93 ± 3.15	0.011	27	24.32 ± 4.93	0.320
D19S20	2	3.08 ± 1.75	0.406	10	8.16 ± 2.86	0.304
D16S83	3	3.83 ± 1.96	0.467	8	9.63 ± 3.10	0.376
D1S74	4	6.52 ± 2.55	0.221	12	15.92 ± 3.99	0.199
D3S42	7	3.74 ± 1.93	0.085	11	9.84 ± 3.14	0.397
Average	7.75	9.80 ± 3.13	0.239	18.38	18.12 ± 4.26	0.551

a. Computed from Eq. (8) using the maximum likelihood estimator of θ ($\hat{\theta}_k$). Observed and expected numbers for the average reflect per locus estimates.

b. Probability of deviation from expectation; based on Poisson distribution. That is, these are probabilities of a value less than or equal to the observed value when the actual observed value is less than the expected, or of a value greater than or equal to the observed value when the actual observed value is greater than the expected.

3 or less in the sample. Although the observed numbers of such rare alleles can be obtained directly from the data in Table 1, the expected numbers are based on Eq. (9), summing over relevant r values (3 or less, for the given example), in which the estimate $\hat{\theta}_k$ is substituted for θ. Because the number of rare alleles follows a Poisson distribution, the congruence of the expected and observed numbers in this table is tested by computing the tail probability of a Poisson distribution. The probability column shows the exact significance values reached in each case. With the exception of the *D14S13* locus, the observed number of rare alleles is in statistical agreement with the expectations. For *D14S13*, with the 1% criterion of rare alleles, we find a deficiency of rare alleles. Therefore this test also suggests that there is no hidden substructuring in the population. As in the case of total number of alleles, hidden heterogeneity would have produced excess rare alleles.

Discussion and Conclusion

The analyses indicate that, even though the genetic variation revealed by the eight VNTR loci is extensive, there is no general indication of hidden subdivision within the white population of Utah. Jorde (1982) came to a similar conclusion by studying migration patterns of the founders of this population. Although our present study does not

provide a new anthropologic conclusion, several features of the analyses are of general significance in understanding the population genetic characteristics of hypervariable loci. First, unlike protein-coding loci, data from even a single VNTR locus can be subjected to this type of analysis because of the extensive number of alleles found at such loci. Second, although the mechanisms producing new variants in VNTR loci [e.g., nonhomologous sister chromatid exchange, unequal crossover, gene conversion, replication slippage; see Jeffreys et al. (1988)] are different from those producing variation in the protein-coding loci (mainly point mutation or small deletion), the infinite allele model of selectively neutral alleles applies equally well to population data for both types of polymorphic loci. This observation is also consistent with the pattern of allele frequency distributions at other VNTR and short tandem repeat (STR) loci found in recent population surveys (Deka et al. 1991; Edwards et al. 1991). Third, although the pooled data on the eight loci satisfy the predictions from the hypothesis of a single homogeneous population rather strikingly, we observe some deviations for the individual loci, but these deviations are in the direction *opposite* to the ones that can be caused by genetic heterogeneity within a population.

Our results apparently contradict Odelberg et al.'s (1989) analysis of deviations from HWE based on the comparison of observed and expected homozygosity and heterozygosity at these loci. Odelberg and co-workers found excess homozygosity at three loci (*D2S44, D14S13,* and *D1S74*), which might be construed as evidence of heterogeneity. A likely explanation for this apparent excess homozygosity is the technical difficulty of distinguishing closely spaced alleles on Southern gels (Devlin et al. 1990). We note that the same three loci exhibit significant differences in the two estimators of θ ($\hat{\theta}_H$ and $\hat{\theta}_k$). Hidden subdivision is not the cause of these deviations because, as noted earlier, the direction of deviation is opposite to what would be expected under heterogeneity.

To examine a possible cause of these discrepancies, we note some sample size considerations. Recall that at each VNTR locus a substantial number of alleles are detected, and almost all alleles occur at low frequencies in the population. Of the 192 alleles detected in this survey, there are only 3 alleles at these 8 loci that have frequencies exceeding 25%. Given this extensive allelic diversity, one might ask whether the available sample sizes are enough to capture all possible genotypes in these samples. With k alleles at a locus, there are $k(k+1)/2$ possible different genotypes, k of which are homozygotes, and $k(k-1)/2$ heterozygotes. Hence, if the sample size (number of individuals surveyed) is less than $k(k+1)/2$, several of these different genotypes will not be recorded in the sample. Exactly how many distinct genotypes were encountered in the survey was not reported by Odelberg et al. (1989). Nevertheless, under the assumption of HWE we can compute the minimum sample size required

to have all genotypes detected in the sample based on the observed allele frequencies. For example, if p_i is the true frequency of the ith allele at a locus, the probability that each of the $K = k(k+1)/2$ possible genotypes will be found in a sample of n individuals is

$$P = \sum \frac{n!}{\prod_{i=1}^{k} n_{ii}! \prod_{i>j=1}^{k}\prod^{k} n_{ij}!} \prod_{i=1}^{k}(p_i^2)^{n_{ii}} \prod_{i>j=1}^{k}\prod^{k}(2p_ip_j)^{n_{ij}}, \tag{11}$$

where the summation is over all n_{ii} and n_{ij} values such that none is 0 and such that they add to the total sample size (n). Although expression (11) is tedious to compute numerically when k and n are both large, it is easy to show that

$$P \geq 1 - \sum_{i=1}^{k}(1 - p_i^2)^n - \sum_{i>j=1}^{k}\sum^{k}(1 - 2p_ip_j)^n. \tag{12}$$

The right-hand side of expression (12) is at a maximum when all allele frequencies are equal, that is, when $p_i = 1/k$ for all i. Therefore a conservative estimate of the minimum sample size requirement for ensuring that all genotypes are represented in the sample with confidence $(1 - \alpha)$ is given by the inequality

$$1 - k(1 - k^{-2})^n - \frac{1}{2}k(k-1)(1 - 2k^{-2})^n \geq 1 - \alpha, \tag{13}$$

which reduces to

$$n \geq -k^2 \log_e \left[\frac{\sqrt{k^2 + 2\alpha k(k-1)} - k}{k(k-1)}\right]. \tag{14}$$

Table 6 shows the values of n for $\alpha = 0.10$, 0.05, and 0.01 that represent the minimum sample size required for the specific 8 loci in the present data. Note that, because the observed allele frequencies are not all equal at these loci, the actual sample size requirement may be even more stringent. Nevertheless, these computations indicate that with the available sample sizes it is unlikely that all possible genotypes are included in the data collected in this specific survey.

In terms of these minimum sample size requirements, it is clear that the smallest sample sizes are for the three loci *D2S44*, *D14S13*, and *D1S74*, which showed significant excess homozygosity in the analysis of Odelberg et al. (1989) and which exhibited significant differences in the two estimators of θ. Because the allele frequency distributions at

Table 6. Minimum Number of Individuals Needed to Detect all VNTR Genotypes in a Population Sample Given the Observed Allele Frequencies

Locus	*Observed Number of Alleles*	*Actual Sample Size*	*Minimum Sample Size Needed*[a]		
			$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
D17S5	14	151	977	1,107	1,421
D2S44	67	151	29,418	32,430	39,570
D9S7	16	136	1,311	1,483	1,890
D14S13	30	82	5,174	5,778	7,210
D19S20	13	84	830	944	1,213
D16S83	15	78	1,137	1,289	1,647
D1S74	22	78	2,632	2,957	3,727
D3S42	15	84	1,137	1,289	1,647
Average	24	106	3,183	3,570	4,486

a. Sample sizes in these computations refer to the number of individuals to be typed for each locus. The values of $\alpha = 0.10$, 0.05, and 0.01 represent 90%, 95%, and 99% confidence of being assured that all possible genotypes are represented in the sample.

these loci are in a direction opposite to that expected in the presence of hidden heterogeneity, we suspect that the observed excess homozygosity at these loci is an artifact of inadequate sample size alone and cannot be attributed to a Wahlund effect or population substructure.

We also note that, even though a formal power analysis of the tests proposed here is not yet available, previous applications of this method have succeeded in detecting hidden heterogeneity, evidenced by analysis of distributions of protein polymorphism (Chakraborty et al. 1988) and mitochondrial morphs (Chakraborty 1990a,b).

We should note that the present methodology rests on the assumption that the allelic diversity revealed by VNTR polymorphism obeys the infinite allele model. This approximation has been justified by Ohta (1986) and is borne out in the present analysis. Other population surveys on VNTR and STR loci also demonstrate that this model is appropriate for predicting allele frequency distributions at such loci (Budowle et al. 1991; Deka et al. 1991; Edwards et al. 1991).

Another characteristic feature of VNTR loci is also revealed through our present calculations. Because the parameter θ equals $4N_e\nu$, the relative values of the estimates of θ at the 8 VNTR loci when compared with protein-coding loci provide an indirect estimate of mutation rates for the VNTR loci. Mohrenweiser et al. (1987) estimated gene diversity at general protein-coding loci of 0.08 for the white population. By using the present gene diversity estimator of θ [solution of Eq. (5)], this leads to $\hat{\theta}_H = 0.0651$. A recent estimate of mutation rate at protein-coding loci is 1.1×10^{-5} per locus per generation (Chakraborty and Neel 1989). By

using the gene diversity estimates of θ from Table 2 and calibrating them against the estimate from protein-coding loci, we obtain mutation rates at the 8 VNTR loci ranging from 4.0×10^{-4} (for the locus *D3S42*) to 7.3×10^{-3} (for the locus *D2S44*) per locus per generation. The average mutation rate over the 8 loci is approximately 1.0×10^{-3} per locus per generation. In other words, these 8 VNTR loci are subject to mutational alterations at a rate 35–660 times (average 90 times) higher than protein-coding loci. These indirect estimates of the mutation rate at VNTR loci are almost an order of magnitude smaller than Jeffreys et al.'s (1988) estimate (average 0.012 per locus per generation for the 5 loci *D5S43, D12S11, D7S22, D7S21*, and *D1S7*) but are close to the estimate reported by Wolff et al. (1988). These observations are also in accordance with the indirect mutation rate estimates derived for 5 STR loci (average 6.1×10^{-5} per locus per generation; Edwards et al. 1991).

Although high mutation rates at VNTR loci are not relevant to the use of these loci in forensic identification problems, a frequency of 1 mutant per 1000 cases can have a substantial impact on parentage testing. Statistical methods to circumvent this problem for conventional genetic markers have been discussed elsewhere [e.g., Chakraborty and Schull (1976) and Chakraborty and Ryman (1981)]; these methods can be reformulated easily for VNTR loci.

Finally, we must close with a note of caution. Even though there are several claims of population subheterogeneity based on a simple demonstration of heterozygote deficiency at VNTR loci (Lander 1989; Cohen 1990), the results described here indicate that additional genetic parameters must be examined before it is valid to conclude that subheterogeneity exists. This is necessary because heterozygote deficiency alone can result from causes other than subheterogeneity, such as incomplete resolution of similar size alleles (Devlin et al. 1990) and loss of allelic bands resulting from very small (or very large) alleles (Skibinski et al. 1983). Lack of congruence between gene diversity and number of alleles can also be produced by evolutionary events [such as the bottleneck effect and the fluctuation of population size (Nei et al. 1975; Maruyama and Fuerst 1984, 1985; Watterson 1984)], which can be checked by examining the relationship of H and $E(k)$ with N_e [Eqs. (1) and (2)]. Ewens (1972) provided extensive tables of relationship between $E(k)$ and N_e (or θ) for an equilibrium population, and Nei et al. (1975), Maruyama and Fuerst (1984, 1985), and Watterson (1984) showed that, when a population goes through a severe bottleneck, the expected relationship between H and k is disturbed (in the direction of excess allele numbers compared with expectation) and the effect persists for a long time (of the order of the inverse of mutation rate).

We suggest that each specific survey should be subjected to tests similar to the ones suggested here. If the observed heterozygote deficiency

is due to factors other than population subheterogeneity, there should be a deficiency of allele numbers (total and rare) in contrast to the excess expected in the presence of population heterogeneity. Data requirements for such tests might include (1) a statement regarding the resolving power of the gel system used, in particular, whether even a single repeating unit can be detected; (2) tabulation of the observed allele frequency distribution in a form similar to our Table 1; (3) documentation of the exact number of distinct genotypes (both homozygotes and heterozygotes) observed in the sample; (4) an attempt to type a sufficient number of individuals so that a substantial fraction of all possible genotypes are detected; and (5) use of a well-defined (geographically or culturally homogeneous) sample population. Admittedly, several of these desiderata are compromised in the data analyzed here, and some are difficult to achieve in practice. For example, it is possible that allelic diversity will increase with an increase in sample size so that a substantial fraction of all genotypes can never be seen in any sample of feasible size. So long as only rare alleles (say, with a frequency less than 0.5%) are missed in a sample, the gene diversity estimator should be relatively insensitive to sample size. Our analysis of the Odelberg et al. (1989) data demonstrates that such analysis is feasible even when the requirements are met only approximately. Further empirical studies in appropriate populations are needed to capitalize fully on the forensic utility of VNTR popymorphisms.

Acknowledgments We thank A. Edwards for his comments on the manuscript. This work was supported by the National Institutes of Health under grant GM 41399 and by the National Institutes of Justice, Office of Justice Programs, US Department of Justice, under grant 90–IJ–CX–0038. Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice.

Received 17 December 1990; revision received 1 March 1991.

Literature Cited

Bell, G.I., M.J. Selby, and W.J. Rutter. 1982. The highly polymorphic region near the insulin gene is composed of simple tandemly repeating sequences. *Nature* 295:31–35.

Budowle, B., R. Chakraborty, A.M. Giusti, A.E. Eisenberg, and R.C. Allen. 1991. Analysis of the variable number of tandem repeats locus *D1S80* by the polymerase chain reaction followed by high resolution polyacrylamide gel electrophoresis. *Am. J. Hum. Genet.* 48:137–144.

Chakraborty, R. 1990a. Genetic profile of cosmopolitan populations: Effects of hidden subdivision. *Anthropol. Anz.* 48:313–331.

Chakraborty, R. 1990b. Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. *Am. J. Hum. Genet.* 47:87–94.

Chakraborty, R., and R.C. Griffiths. 1982. Correlation of heterozygosity and number of alleles in different frequency classes. *Theor. Popul. Biol.* 21:205–218.

Chakraborty, R., and J.V. Neel. 1989. Description and validation of a method for simultaneous estimation of effective population size and mutation rate from human population data. *Proc. Natl. Acad. Sci. USA* 86:9407–9411.

Chakraborty, R., and N. Ryman. 1981. Use of odds of paternity computations in determining the reliability of single exclusions in paternity testing. *Hum. Hered.* 31:363–369.

Chakraborty, R., and W.J. Schull. 1976. A note on the distribution of the number of exclusions to be expected in paternity testing. *Am. J. Hum. Genet.* 28:615–618.

Chakraborty, R., and R.J. Schwartz. 1990. Selective neutrality of surname distribution in an immigrant Indian community of Houston, Texas. *Am. J. Hum. Biol.* 2:1–15.

Chakraborty, R., P.E. Smouse, and J.V. Neel. 1988. Population amalgamation and genetic variation: Observations on artificially agglomerated tribal populations of Central and South America. *Am. J. Hum. Genet.* 43:709–725.

Chimera, J.A., C.R. Harris, and M. Litt. 1989. Population genetics of the highly polymorphic locus *D16S7* and its use in paternity evaluation. *Am. J. Hum. Genet.* 45:926–931.

Cohen, J.E. 1990. DNA fingerprinting for forensic identification: Potential effects of data interpretation of subpopulation heterogeneity and band number variability. *Am. J. Hum. Genet.* 46:358–368.

Deka, R., R. Chakraborty, and R.E. Ferrell. 1991. Population genetics of hypervariable loci in three ethnic groups. *Genomics* (in press).

Devlin, B., N. Risch, and K. Roeder. 1990. No excess of homozygosity at loci used for DNA fingerprinting. *Science* 249:1416–1420.

Edwards, A., H.A. Hammond, C.T. Caskey, L. Jin, and R. Chakraborty. 1991. Population genetics of trimeric and tetrameric tandem repeats in four human ethnic groups. *Genomics* (in press).

Ewens, W.J. 1972. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3:87–112.

Fuerst, P.A., R. Chakraborty, and M. Nei. 1977. Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. *Genetics* 86:455–483.

Jeffreys, A.J., V. Wilson, and S.L. Thein. 1985. Hypervariable "minisatellite" regions of human DNA. *Nature* 314:67–73.

Jeffreys, A.J., N.J. Royle, V. Wilson, and Z. Wong. 1988. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* 332:278–281.

Jorde, L.B. 1982. The genetic structure of the Utah Mormons: Migration analysis. *Hum. Biol.* 54:583–597.

Kimura, M., and J.F. Crow. 1964. The number of alleles that can be maintained in a finite population. *Genetics* 49:725–738.

Lander, E.S. 1989. DNA fingerprinting on trial. *Nature* 339:501–505.

Maruyama, T., and P.A. Fuerst. 1984. Population bottlenecks and nonequilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. *Genetics* 108:745–763.

Maruyama, T., and P.A. Fuerst. 1985. Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in small population that was formed by a recent bottleneck. *Genetics* 111:691–703.

Mohrenweiser, H.W., K.H. Wurzinger, and J.V. Neel. 1987. Frequency and distribution of rare electrophoretic mobility variants in a population of newborns in Ann Arbor, Michigan. *Ann Hum. Genet.* 51:303–316.

Nakamura, Y., M. Leppert, P. O'Connell et al. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* 235:1616–1622.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583–590.

Nei, M., T. Maruyama, and R. Chakraborty. 1975. The bottleneck effect and genetic variability in populations. *Evolution* 29:1–10.

Odelberg, S.J., R. Platke, J.R. Eldridge, L. Ballard, P. O'Connell, Y. Nakamura, M. Leppert, J.-M. Lalouel, and R. White. 1989. Characterization of eight VNTR loci by agarose gel electrophoresis. *Genomics* 5:915–924.

Ohta, T. 1986. Actual number of alleles contained in a multigene family. *Genet. Res.* 48:119–123.

Proudfoot, N.J., A. Gil, and T. Maniatis. 1982. The structure of the human zeta-globin gene and a closely linked, nearly identical pseudogene. *Cell* 31:553–563.

Skibinski, D.O.F., J.A. Beardmore, and T.F. Cross. 1983. Aspects of the population genetics of *Mytilus* (Mytilidae; Mollusca) in the British Isles. *Biol. J. Linn. Soc.* 19:137–183.

Watterson, G.A. 1978. The homozygosity test of neutrality. *Genetics* 88:405–417.

Watterson, G.A. 1984. Allele frequencies after a bottleneck. *Theor. Popul. Biol.* 26:387–407.

White, R., M. Leppert, D.T. Bishop, D. Barker, J. Berkowitz, C. Brown, P. Callahan, T. Holm, and L. Jerominski. 1985. Construction of linkage maps with DNA markers for human chromosomes. *Nature* 313:101–105.

Wolff, R.K., Y. Nakamura, and R. White. 1988. Molecular characterization of a spontaneously generated new allele at a VNTR locus: No exchange of flanking DNA sequences. *Genomics* 3:347–351.

Wong, Z., V. Wilson, I. Patel, S. Povey, and A.J. Jeffreys. 1987. Characterization of a panel of highly variable minisatellites cloned from human DNA. *Ann. Hum. Genet.* 51:269–288.

Wright, S. 1949. Genetics of populations. In *Encyclopaedia Britannica*, 14th ed., v. 10, 111–112.

Wyman, A.R., and R. White. 1980. A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* 77:6754–6758.

Zouros, E. 1979. Mutation rates, population sizes, and amounts of electrophoretic variation of enzyme loci in natural populations. *Genetics* 92:623–646.

Patterns of Genetic Variation Within and Between Populations Detected by PCR-Based VNTR Polymorphisms

Ranajit Chakraborty, Stephen P. Dalger and Eric Boerwinkle
University of Texas Health Science Center
Houston, Texas

It is now well established that dispersed in the human genome there are many hypervariable loci where the polymorphism is due to the copy number variation of tandemly repeated conserved nucleotide sequences. Depending upon the length of the core sequences (from two bases to several kilobases), such loci are called microsatellites, or short tandem repeat (STR) loci and minisatellites, or variable number of tandem repeat (VNTR) loci. They have the following characteristics: 1) quasi-continuous distributions of allele sizes (determined by the total length of the alleles), 2) large number of alleles per locus (sometimes exceeding 50), and 3) high heterozygosity (often in excess of 95%). Due to these characteristics, such hypervariable loci have been proven to be quite effective for gene mapping, genetic counseling when they are closely located to a disease susceptibility locus, forensic identification of individuals, determination of relatedness among individuals, and strategic policy making in conversation biology.

For efficient use of STR and VNTR polymorphisms it is; however, imperative to know the following:

1) the measured variation at these loci are inherited in a Mendelian fashion;
2) how many different alleles are segregating within a population;
3) their allele frequency distributions in populations of diverse origin;
4) the effective methods of prediction of genotype frequencies from their respective allele frequencies;
5) features of randomness of allele frequency distributions at different loci within populations;
6) stability of the alleles over generations, so as to assert biological relatedness between individuals from their allele sharing according to the identity by descent rule;
7) how similar (or dissimilar) different populations are with regard to the allele frequencies so that appropriate reference population database can be used for specific applications; and
8) what is the pattern of evolutionary dynamics of such polymorphisms.

The purpose of this paper is to discuss several issues that are important for the analysis of the pattern of within and between population variation at hypervariable loci, each of which in turn relates to the previously mentioned features of genetic variation. The ensuing discussions are specifically made in the context of data generated by the polymerase chain reaction (PCR)-based protocols. In principle, however, they should also apply to data gathered by means of the traditional Southern blot restriction fragment length polymorphism (RFLP) analysis.

SOUTHERN BLOT GENERATED DATA VERSUS PCR-BASED DATA

Before discussing the methods of data analysis, it is worthwhile to consider some distinctive features of data that are generated by the Southern blot data and PCR-based data. Since most population genetic methods of analyzing genotype frequency data require complete resolution of all alleles, data generated by a Southern blot RFLP analysis may be considered sometimes inappropriate. Therefore, it is not always possible by an RFLP analysis to distinguish between copy number variation of tandem repeats of short core sequences when the repeat units vary by one or a few copies. This is known to introduce pseudohomozygosity (i.e., heterozygote

individuals carrying two nearly similar size alleles may be classified as homozygotes). Furthermore, when RFLP analysis is specifically designed for high resolution of alleles, some alleles of extreme sizes (high as well as low size alleles) may not be detected on a Southern gel. In addition, the determination of allele sizes is not perfectly accurate even with ladder lanes on the gel. Comparison of allelic homology by size determination from different gels is even more problematic.

Since PCR-based protocols are designed for discretization of alleles, through PCR studies it is possible to examine the sequence homology of identical size alleles as well as sequence identity of different copies of core sequences for specific individuals. This in turn results in determination of allelic identity by descent between relatives with greater accuracy. The PCR-based studies on several STR loci in populations of diverse origin indicate that allelic distinctions with a single repeat variation when the core motif is only two bases long are also fairly reliable (Chakraborty *et al.* 1991; Edwards *et al.* 1991a). In short, even with the current knowledge of PCR technology, the limitations of Southern blot RFLP analysis for the detection of hypervariable polymorphisms can be greatly circumvented, making such data amenable for population genetic analysis.

ASPECTS OF WITHIN POPULATION VARIATION

Once the allelic distinctions are made and the allele frequencies are determined from a random sample of individuals drawn from a population, the next important question regarding the features of within population variation is whether or not the genotype frequencies can be reliably predicted from allele frequencies assuming the traditional Hardy-Weinberg (or the Square) rule. This is needed particularly with regard to hypervariable loci. Unlike the traditional blood group and protein polymorphisms, direct estimation of genotype probabilities from relative counts of different genotypes at a hypervariable locus is not recommended, since too many possible genotypes exist at almost all such loci. Further, many of them will remain unobserved even when the sample size is respectable. For example, with 50 alleles segregating at a locus, there are 1,275 possible distinct genotypes. The prospect of finding all of them in a sample is quite low even if the sample size is increased to 5,000 individuals sampled from the population (Chakraborty and Daiger 1991). Therefore, application of the Hardy-Weinberg rule is also essential for genotypic probability calculations at hypervariable loci. The tests of checking the appropriateness of such calculations has been recently questioned (Lander 1989; 1991). Such criticisms, however, seem to disregard the fact that the context of applications of the traditional statistical tests for Hardy-Weinberg rule to hypervariable loci must take into account the sparse nature of data. This is important since alleles numbers are large, and many alleles occur in low frequencies, resulting in too many missing genotypes existing in a sample even when the sample size is respectable and the fluctuations of observed genotype frequencies from their Hardy-Weinberg expectations become asymmetric.

The validity of Hardy-Weinberg expectations of genotype distributions can be checked from data on unrelated individuals as well as from data on relatives. The available test procedures applicable to the genotype frequency data on unrelated individuals can be broadly classified as: 1) observed versus expected frequencies of all heterozygote (or homozygote) individuals, 2) observed versus expected numbers of distinct genotypes seen in a sample (heterozygote as well as homozygote genotypes examined separately), 3) congruence of observed number of alleles with its prediction based on heterozygosity at the locus, 4) likelihood ratio test based on contrasts of observed versus expected frequencies of each possible genotype, 5) intraclass correlation of allele sizes (or copy numbers) within individuals (Karlin *et al.* 1981), and 6) association array analysis (Karlin 1983). Based on genotype data on nuclear families (both parents and one or more children), tests of Hardy-Weinberg equilibrium (HWE) may be constructed by: 1) tests based on genotypic association of spouse dichotomized by heterozygosity and homozygosity and numbers of shared alleles, 2) tests based on Mendelian segregation of alleles in children, subdividing the data by parental mating classes based on heterozygosity status and shared alleles, 3) interclass correlation of allele sizes (or copy numbers) between spouses, and 4) association array analysis of allele sizes observed in spouses. This multiplicity of tests is particularly meaningful in the

context of studies on hypervariable loci, since each of these tests individually does not generally have an appreciable power of detecting deviation from the Hardy-Weinberg rule. Together, they emphasize different aspects of the genotype frequency distribution in a population. In particular, one should recognize that because of the sparse nature of the data instead of the traditional large sample distribution of the test statistics (such as the Chi-square or the normal distribution), one must determine the level of significance of the test statistics by empirical means of suffling the observed alleles across individuals to from new genotypes (permutation methods).

Applications of such permutation methods demonstrate that the nonparametric approaches of intraclass and interclass correlations and association array analysis (Karlin *et al.* 1981; Karlin 1983) are particularly suited for VNTR data. Accordingly, the allele size distributions are often complex (Shriver *et al.* 1991), and they generally provide a greater statistical power should there exist any deviation from the Hardy-Weinberg rule because of the presence of population substructuring and/or the presence of pseudohomozygosity due to incomplete resolution of alleles.

These test procedures, when applied to population data on the apolipoprotein B (Apo B) and D1S80 VNTR loci, and on 5 STR loci and a (CA)-repeat locus at the apolipoprotein C II (Apo C II) gene, all scored by PCR-based studies, suggest no deviation from the HWE (Budowle *et al.* 1991; Chakraborty *et al.* 1991; Edwards *et al.* 1991b). Hence, we conclude that the estimation of genotype frequencies from allele frequencies by the use of HWE assumption is quite reliable.

Another aspect of within population variation at hypervariable loci relates to the estimation of multilocus genotype frequencies from the single locus genotype frequencies. This is generally done by multiplying the respective single locus genotype frequencies across all loci, on the grounds that the loci under investigation are independently segregating. Hence, there is no cosegregation of alleles leading to gametic phase disequilibrium. The criticism against the application of this product rule is that in the presence of population substructuring, there will be a gametic phase disequilibria even among alleles of unlinked genes. Tests for appropriateness of the product rule may be based on: 1) the distribution of a number of heterozygous loci among individuals each scored for multiple loci, 2) computation of linkage disequilibria for every pair of locu-allele combinations, 3) interclass correlation of allele sizes (or copy numbers) between pairs of loci, and 4) association array analysis. As in the case of HWE tests, simulation analysis suggests that the interclass correlation approach or an association array analysis has a comparatively greater power of detecting deviations from the multiplication rule. This is particularly true when a large number (greater than 10) alleles are segregating at each locus. Of the previously mentioned alternative test procedures, the distribution of the number of heterozygous loci has the flexibility that it can handle any number of loci simultaneously even though it condenses the genotype data into two classes (heterozygote and homozygote). While in all other test procedures the appropriateness of the multiplication rule is checked by testing the pairwise independence of loci, it should be noted that the interdependence of genotype frequencies among loci caused by population substructure is generally such that most of the dependence can be explained only by (locus) pairwise dependence parameters. Hence, the test for higher order gametic phase disequilibria produced by mixtures of populations is not critical for judging the appropriateness of the multiplication rule.

Applications of these test procedures to PCR-based studies on two VNTR loci (Apo B and Apo CII) and 5 STR loci suggest that in populations defined by major racial classifications (such as American Caucasians, French Caucasians, Texas Blacks, Texas Mexican-Americans or Asians of Texas), the multiplication rule is quite adequate for estimating multilocus genotype frequencies (Chakraborty *et al.* 1991; Edwards *et al.* 1991b).

ASPECTS OF BETWEEN POPULATION VARIATION

Genetic variation among populations of different origin is a geological truism and should exist for hypervariable loci. Therefore, when we study the allele frequency differences across populations and examine their significance, two questions become important: 1) Is the difference statistically significant? and 2) Is the pattern of allele frequency differences consistent with

the ethnohistory of populations? Studies done on anthropologically defined populations reveal that significant allele frequency differences among diverse populations exist. However, their pattern is quite congruent with the ones found at the traditional blood groups and protein loci and these are consistent with the ethnohistory of populations (Deka *et al.* 1991). Even more broadly defined populations satisfy these characteristics, observed in a study of 5 STR loci in four major racial groups (Edwards *et al.* 1991b).

In this context, it is worth noting that summary measures of allele frequency differences such as coefficient of gene differentiation (Nei 1973), genetic distance (Nei 1978), or proportion of shared alleles across populations in relation to the average frequencies of such alleles are also meaningful concepts in the context of hypervariable loci. However, the length polymorphism of repeat core sequences does not reveal the total nucleotide diversity seen at these loci (Chakraborty *et al.* 1991; Jin *et al.* 1991).

It should be noted, however, that hypervariability at VNTR and STR loci generally leads to larger interpopulational differences compared to that at the traditional serological and biochemical loci. This is due to larger rates of mutations at VNTR and STR loci, whose relative magnitude in relation to the protein mutation rate can be indirectly estimated. This has been attempted for several VNTR and STR loci (Chakraborty and Daiger 1991; Deka *et al.* 1991; Edwards *et al.* 1991b), which indicates that the rate of mutation at VNTR and STR loci are generally higher (by 6- to 50-fold, sometimes up to 500-fold) than that at the protein loci.

IMPLICATIONS FOR FORENSIC IDENTIFICATION

Such issues of intra- and interpopulational variation at VNTR and STR loci, when analyzed appropriately, suggest that the population genetic assumptions of HWE and product rule are quite adequately validated. When departures are noted, they could be explained by the technical limitations (such as nondetectability of alleles of extreme size in Southern blot analysis or pseudohomozygosity due to incomplete resolution of alleles). Several of these technical limitations can be circumvented by PCR-based studies, and even though data on PCR-based allele frequencies are not as widely available at present (in comparison to protein variation), a general picture of interpopulational differences is quite evident even from the limited studies. In several cases of forensic applications when match probability was estimated for a specific multilocus genotype profile, using data of allele frequencies from a variety of populations, it is seen that the estimates are quite consistent; their differences may proportionally differ by several fold, but in no case does a genotype profile that was rare for a population data become common for another population.

From these considerations, we conclude that hypervariable polymorphism studied by PCR-based protocol meets the standard of population genetic analysis. Further, their use in the applications mentioned previously is justifiable on the basis of biological as well as statistical characteristics of these polymorphic systems.

ACKNOWLEDGEMENT

This work was supported in part by grant 90-IJ-CX-0038 from the National Institute of Justice Programs, United States Department of Justice. Points of view or opinions expressed in this paper are those of the authors and do not necessarily represent the official opinion of the United States Department of Justice.

REFERENCES

Budowle, B., Chakraborty, R., Giusti, A., Eisenberg, A. and Allen, R. (1991). Analysis of the variable number of tandem repeats locus D1S80 by the polymerase chain reaction followed by high resolution polyacrylamide gel electrophoresis, Am. J. Hum. Genet. 48:137-144.

Chakraborty, R. and Daiger, S. (1991). Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah, Hum. Biol. 63:571-587.

Chakraborty, R., Fornage, M., Gueguen, R. and Boerwinkle, E. (1991). Population Genetics of Hypervariable Loci: Analysis of PCR-Based VNTR Polymorphism Within a Population. In: DNA Fingerprinting: Approaches and Applications. T. Burke, G. Dolf, A. J. Jeffreys and R. Wolff (eds.), Birkhauser Verlag, Basel, pp. 127-143.

Deka, R., Chakraborty, R. and Ferrell, R. (1991). A population genetic study of six VNTR loci in three ethnically defined populations, Genomics 11:83-92.

Edwards, A., Civitello, A., Hammond, H. and Caskey, C. (1991a). DNA typing and genetic mapping with trimeric and tetrameric tandem repeats, Am. J. Hum. Genet. 49:746-756.

Edwards, A., Hammond, H., Jin, L., Caskey, C. and Chakraborty R. (1991b). Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups, Genomics, in press.

Jin, L., Chakraborty, R., Hammond, H. and Caskey C. (1991). Polymorphisms at short tandem repeat (STR) loci within and between four ethnic populations of Texas, Am. J. Hum. Genet., in press.

Karlin, S. (1983). Association arrays in assessing forms of dependencies between bivariate random variables, Proc. Natl. Acad. Sci. USA 80:647-651.

Karlin, S., Cameron, E. and Williams, P. (1981). Sibling and parent-offspring correlation with variable family size, Proc. Natl. Acad. Sci. USA 78:2664-2668.

Lander, E. (1989). DNA fingerprinting on trail, Nature 339:501-505.

Lander, E. (1991). Invited editorial: Research on DNA typing catching up with courtroom application, Am. J. Hum. Genet. 48:819-823.

Nei, M. (1973). Analysis of gene diversity in subdivided populations, Proc. Natl. Acad. Sci. USA 70:3321-3323.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals, Genetics 89:583-590.

Shriver, M., Daiger, S. P., Chakraborty, R. and Boerwinkle, E. (1991). Multimodal distribution of length variation in VNTR loci detected using PCR, Crime Lab. Digest 18:144-147.

Patterns of Genetic Variation Within and Between Populations Detected by PCR-Based VNTR Polymorphisms

Ranajit Chakraborty, Stephen P. Dalger and Eric Boerwinkle
University of Texas Health Science Center
Houston, Texas

It is now well established that dispersed in the human genome there are many hypervariable loci where the polymorphism is due to the copy number variation of tandemly repeated conserved nucleotide sequences. Depending upon the length of the core sequences (from two bases to several kilobases), such loci are called microsatellites, or short tandem repeat (STR) loci and minisatellites, or variable number of tandem repeat (VNTR) loci. They have the following characteristics: 1) quasi-continuous distributions of allele sizes (determined by the total length of the alleles), 2) large number of alleles per locus (sometimes exceeding 50), and 3) high heterozygosity (often in excess of 95%). Due to these characteristics, such hypervariable loci have been proven to be quite effective for gene mapping, genetic counseling when they are closely located to a disease susceptibility locus, forensic identification of individuals, determination of relatedness among individuals, and strategic policy making in conversation biology.

For efficient use of STR and VNTR polymorphisms it is; however, imperative to know the following:

1) the measured variation at these loci are inherited in a Mendelian fashion;
2) how many different alleles are segregating within a population;
3) their allele frequency distributions in populations of diverse origin;
4) the effective methods of prediction of genotype frequencies from their respective allele frequencies;
5) features of randomness of allele frequency distributions at different loci within populations;
6) stability of the alleles over generations, so as to assert biological relatedness between individuals from their allele sharing according to the identity by descent rule;
7) how similar (or dissimilar) different populations are with regard to the allele frequencies so that appropriate reference population database can be used for specific applications; and
8) what is the pattern of evolutionary dynamics of such polymorphisms.

The purpose of this paper is to discuss several issues that are important for the analysis of the pattern of within and between population variation at hypervariable loci, each of which in turn relates to the previously mentioned features of genetic variation. The ensuing discussions are specifically made in the context of data generated by the polymerase chain reaction (PCR)-based protocols. In principle, however, they should also apply to data gathered by means of the traditional Southern blot restriction fragment length polymorphism (RFLP) analysis.

SOUTHERN BLOT GENERATED DATA VERSUS PCR-BASED DATA

Before discussing the methods of data analysis, it is worthwhile to consider some distinctive features of data that are generated by the Southern blot data and PCR-based data. Since most population genetic methods of analyzing genotype frequency data require complete resolution of all alleles, data generated by a Southern blot RFLP analysis may be considered sometimes inappropriate. Therefore, it is not always possible by an RFLP analysis to distinguish between copy number variation of tandem repeats of short core sequences when the repeat units vary by one or a few copies. This is known to introduce pseudohomozygosity (i.e., heterozygote

individuals carrying two nearly similar size alleles may be classified as homozygotes). Furthermore, when RFLP analysis is specifically designed for high resolution of alleles, some alleles of extreme sizes (high as well as low size alleles) may not be detected on a Southern gel. In addition, the determination of allele sizes is not perfectly accurate even with ladder lanes on the gel. Comparison of allelic homology by size determination from different gels is even more problematic.

Since PCR-based protocols are designed for discretization of alleles, through PCR studies it is possible to examine the sequence homology of identical size alleles as well as sequence identity of different copies of core sequences for specific individuals. This in turn results in determination of allelic identity by descent between relatives with greater accuracy. The PCR-based studies on several STR loci in populations of diverse origin indicate that allelic distinctions with a single repeat variation when the core motif is only two bases long are also fairly reliable (Chakraborty *et al.* 1991; Edwards *et al.* 1991a). In short, even with the current knowledge of PCR technology, the limitations of Southern blot RFLP analysis for the detection of hypervariable polymorphisms can be greatly circumvented, making such data amenable for population genetic analysis.

ASPECTS OF WITHIN POPULATION VARIATION

Once the allelic distinctions are made and the allele frequencies are determined from a random sample of individuals drawn from a population, the next important question regarding the features of within population variation is whether or not the genotype frequencies can be reliably predicted from allele frequencies assuming the traditional Hardy-Weinberg (or the Square) rule. This is needed particularly with regard to hypervariable loci. Unlike the traditional blood group and protein polymorphisms, direct estimation of genotype probabilities from relative counts of different genotypes at a hypervariable locus is not recommended, since too many possible genotypes exist at almost all such loci. Further, many of them will remain unobserved even when the sample size is respectable. For example, with 50 alleles segregating at a locus, there are 1,275 possible distinct genotypes. The prospect of finding all of them in a sample is quite low even if the sample size is increased to 5,000 individuals sampled from the population (Chakraborty and Daiger 1991). Therefore, application of the Hardy-Weinberg rule is also essential for genotypic probability calculations at hypervariable loci. The tests of checking the appropriateness of such calculations has been recently questioned (Lander 1989; 1991). Such criticisms, however, seem to disregard the fact that the context of applications of the traditional statistical tests for Hardy-Weinberg rule to hypervariable loci must take into account the sparse nature of data. This is important since alleles numbers are large, and many alleles occur in low frequencies, resulting in too many missing genotypes existing in a sample even when the sample size is respectable and the fluctuations of observed genotype frequencies from their Hardy-Weinberg expectations become asymmetric.

The validity of Hardy-Weinberg expectations of genotype distributions can be checked from data on unrelated individuals as well as from data on relatives. The available test procedures applicable to the genotype frequency data on unrelated individuals can be broadly classified as: 1) observed versus expected frequencies of all heterozygote (or homozygote) individuals, 2) observed versus expected numbers of distinct genotypes seen in a sample (heterozygote as well as homozygote genotypes examined separately), 3) congruence of observed number of alleles with its prediction based on heterozygosity at the locus, 4) likelihood ratio test based on contrasts of observed versus expected frequencies of each possible genotype, 5) intraclass correlation of allele sizes (or copy numbers) within individuals (Karlin *et al.* 1981), and 6) association array analysis (Karlin 1983). Based on genotype data on nuclear families (both parents and one or more children), tests of Hardy-Weinberg equilibrium (HWE) may be constructed by: 1) tests based on genotypic association of spouse dichotomized by heterozygosity and homozygosity and numbers of shared alleles, 2) tests based on Mendelian segregation of alleles in children, subdividing the data by parental mating classes based on heterozygosity status and shared alleles, 3) interclass correlation of allele sizes (or copy numbers) between spouses, and 4) association array analysis of allele sizes observed in spouses. This multiplicity of tests is particularly meaningful in the

context of studies on hypervariable loci, since each of these tests individually does not generally have an appreciable power of detecting deviation from the Hardy-Weinberg rule. Together, they emphasize different aspects of the genotype frequency distribution in a population. In particular, one should recognize that because of the sparse nature of the data instead of the traditional large sample distribution of the test statistics (such as the Chi-square or the normal distribution), one must determine the level of significance of the test statistics by empirical means of suffling the observed alleles across individuals to from new genotypes (permutation methods).

Applications of such permutation methods demonstrate that the nonparametric approaches of intraclass and interclass correlations and association array analysis (Karlin *et al.* 1981; Karlin 1983) are particularly suited for VNTR data. Accordingly, the allele size distributions are often complex (Shriver *et al.* 1991), and they generally provide a greater statistical power should there exist any deviation from the Hardy-Weinberg rule because of the presence of population substructuring and/or the presence of pseudohomozygosity due to incomplete resolution of alleles.

These test procedures, when applied to population data on the apolipoprotein B (Apo B) and D1S80 VNTR loci, and on 5 STR loci and a (CA)-repeat locus at the apolipoprotein C-II (Apo C II) gene, all scored by PCR-based studies, suggest no deviation from the HWE (Budowle *et al.* 1991; Chakraborty *et al.* 1991; Edwards *et al.* 1991b). Hence, we conclude that the estimation of genotype frequencies from allele frequencies by the use of HWE assumption is quite reliable.

Another aspect of within population variation at hypervariable loci relates to the estimation of multilocus genotype frequencies from the single locus genotype frequencies. This is generally done by multiplying the respective single locus genotype frequencies across all loci, on the grounds that the loci under investigation are independently segregating. Hence, there is no cosegregation of alleles leading to gametic phase disequilibrium. The criticism against the application of this product rule is that in the presence of population substructuring, there will be a gametic phase disequilibria even among alleles of unlinked genes. Tests for appropriateness of the product rule may be based on: 1) the distribution of a number of heterozygous loci among individuals each scored for multiple loci, 2) computation of linkage disequilibria for every pair of locu-allele combinations, 3) interclass correlation of allele sizes (or copy numbers) between pairs of loci, and 4) association array analysis. As in the case of HWE tests, simulation analysis suggests that the interclass correlation approach or an association array analysis has a comparatively greater power of detecting deviations from the multiplication rule. This is particularly true when a large number (greater than 10) alleles are segregating at each locus. Of the previously mentioned alternative test procedures, the distribution of the number of heterozygous loci has the flexibility that it can handle any number of loci simultaneously even though it condenses the genotype data into two classes (heterozygote and homozygote). While in all other test procedures the appropriateness of the multiplication rule is checked by testing the pairwise independence of loci, it should be noted that the interdependence of genotype frequencies among loci caused by population substructure is generally such that most of the dependence can be explained only by (locus) pairwise dependence parameters. Hence, the test for higher order gametic phase disequilibria produced by mixtures of populations is not critical for judging the appropriateness of the multiplication rule.

Applications of these test procedures to PCR-based studies on two VNTR loci (Apo B and Apo CII) and 5 STR loci suggest that in populations defined by major racial classifications (such as American Caucasians, French Caucasians, Texas Blacks, Texas Mexican-Americans or Asians of Texas), the multiplication rule is quite adequate for estimating multilocus genotype frequencies (Chakraborty *et al.* 1991; Edwards *et al.* 1991b).

ASPECTS OF BETWEEN POPULATION VARIATION

Genetic variation among populations of different origin is a geological truism and should exist for hypervariable loci. Therefore, when we study the allele frequency differences across populations and examine their significance, two questions become important: 1) Is the difference statistically significant? and 2) Is the pattern of allele frequency differences consistent with

the ethnohistory of populations? Studies done on anthropologically defined populations reveal that significant allele frequency differences among diverse populations exist. However, their pattern is quite congruent with the ones found at the traditional blood groups and protein loci and these are consistent with the ethnohistory of populations (Deka *et al.* 1991). Even more broadly defined populations satisfy these characteristics, observed in a study of 5 STR loci in four major racial groups (Edwards *et al.* 1991b).

In this context, it is worth noting that summary measures of allele frequency differences such as coefficient of gene differentiation (Nei 1973), genetic distance (Nei 1978), or proportion of shared alleles across populations in relation to the average frequencies of such alleles are also meaningful concepts in the context of hypervariable loci. However, the length polymorphism of repeat core sequences does not reveal the total nucleotide diversity seen at these loci (Chakraborty *et al.* 1991; Jin *et al.* 1991).

It should be noted, however, that hypervariability at VNTR and STR loci generally leads to larger interpopulational differences compared to that at the traditional serological and biochemical loci. This is due to larger rates of mutations at VNTR and STR loci, whose relative magnitude in relation to the protein mutation rate can be indirectly estimated. This has been attempted for several VNTR and STR loci (Chakraborty and Daiger 1991; Deka *et al.* 1991; Edwards *et al.* 1991b), which indicates that the rate of mutation at VNTR and STR loci are generally higher (by 6- to 50-fold, sometimes up to 500-fold) than that at the protein loci.

IMPLICATIONS FOR FORENSIC IDENTIFICATION

Such issues of intra- and interpopulational variation at VNTR and STR loci, when analyzed appropriately, suggest that the population genetic assumptions of HWE and product rule are quite adequately validated. When departures are noted, they could be explained by the technical limitations (such as nondetectability of alleles of extreme size in Southern blot analysis or pseudohomozygosity due to incomplete resolution of alleles). Several of these technical limitations can be circumvented by PCR-based studies, and even though data on PCR-based allele frequencies are not as widely available at present (in comparison to protein variation), a general picture of interpopulational differences is quite evident even from the limited studies. In several cases of forensic applications when match probability was estimated for a specific multilocus genotype profile, using data of allele frequencies from a variety of populations, it is seen that the estimates are quite consistent; their differences may proportionally differ by several fold, but in no case does a genotype profile that was rare for a population data become common for another population.

From these considerations, we conclude that hypervariable polymorphism studied by PCR-based protocol meets the standard of population genetic analysis. Further, their use in the applications mentioned previously is justifiable on the basis of biological as well as statistical characteristics of these polymorphic systems.

ACKNOWLEDGEMENT

This work was supported in part by grant 90-IJ-CX-0038 from the National Institute of Justice Programs, United States Department of Justice. Points of view or opinions expressed in this paper are those of the authors and do not necessarily represent the official opinion of the United States Department of Justice.

REFERENCES

Budowle, B., Chakraborty, R., Giusti, A., Eisenberg, A. and Allen, R. (1991). Analysis of the variable number of tandem repeats locus D1S80 by the polymerase chain reaction followed by high resolution polyacrylamide gel electrophoresis, Am. J. Hum. Genet. 48:137-144.

Chakraborty, R. and Daiger, S. (1991). Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah, Hum. Biol. 63:571-587.

Chakraborty, R., Fornage, M., Gueguen, R. and Boerwinkle, E. (1991). Population Genetics of Hypervariable Loci: Analysis of PCR-Based VNTR Polymorphism Within a Population. In: DNA Fingerprinting: Approaches and Applications. T. Burke, G. Dolf, A. J. Jeffreys and R. Wolff (eds.), Birkhauser Verlag, Basel, pp. 127-143.

Deka, R., Chakraborty, R. and Ferrell, R. (1991). A population genetic study of six VNTR loci in three ethnically defined populations, Genomics 11:83-92.

Edwards, A., Civitello, A., Hammond, H. and Caskey, C. (1991a). DNA typing and genetic mapping with trimeric and tetrameric tandem repeats, Am. J. Hum. Genet. 49:746-756.

Edwards, A., Hammond, H., Jin, L., Caskey, C. and Chakraborty R. (1991b). Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups, Genomics, in press.

Jin, L., Chakraborty, R., Hammond, H. and Caskey C. (1991). Polymorphisms at short tandem repeat (STR) loci within and between four ethnic populations of Texas, Am. J. Hum. Genet., in press.

Karlin, S. (1983). Association arrays in assessing forms of dependencies between bivariate random variables, Proc. Natl. Acad. Sci. USA 80:647-651.

Karlin, S., Cameron, E. and Williams, P. (1981). Sibling and parent-offspring correlation with variable family size, Proc. Natl. Acad. Sci. USA 78:2664-2668.

Lander, E. (1989). DNA fingerprinting on trail, Nature 339:501-505.

Lander, E. (1991). Invited editorial: Research on DNA typing catching up with courtroom application, Am. J. Hum. Genet. 48:819-823.

Nei, M. (1973). Analysis of gene diversity in subdivided populations, Proc. Natl. Acad. Sci. USA 70:3321-3323.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals, Genetics 89:583-590.

Shriver, M., Daiger, S. P., Chakraborty, R. and Boerwinkle, E. (1991). Multimodal distribution of length variation in VNTR loci detected using PCR, Crime Lab. Digest 18:144-147.

C. R. Rao and R. Chakraborty, eds., *Handbook of Statistics*, *Vol.* 8

7

Analysis of Population Structure: A Comparative Study of Different Estimators of Wright's Fixation Indices

Ranajit Chakraborty and Heidi Danker-Hopfe

1. Introduction

Computations of Wright's fixation indices (F_{IT}, F_{ST}, and F_{IS}) are pivotal for studying the genetic differentiation of populations. It is well known that these indices can be conceptually defined in terms of correlations between uniting gametes (Wright, 1943, 1951); as functions of heterozygosities and their Hardy–Weinberg expectations (Nei, 1973, 1977), or as functions of variance components from a nested analysis of variance (Cockerham, 1969, 1973; Weir and Cockerham, 1984; Long, 1986). Nei (1977) and Nei and Chesser (1983) considered the question of estimating the fixation indices through a decomposition of gene diversity in the total population, while Cockerham (1969, 1973) and Weir and Cockerham (1984) provided estimation procedures by a variance component analysis. Long (1986) extended the variance component approach of estimation to the case of multiple (greater than two) allelic loci, which gives numerically different results from the Weir–Cockerham estimates (see equation (10) of Weir and Cockerham, 1984 vs. equations (9a), (10a) and (11a) of Long, 1986). Although there are several studies drawing comparisons of these different estimates in simulated data (Van Den Bussche et al., 1986; Chakraborty and Leimar, 1987; Slatkin and Barton, 1989), it is not generally known how these different estimates differ in real data in practice. Furthermore, there is no comprehensive computer algorithm which computes all of these estimates simultaneously.

The purpose of this chapter is twofold:

(1) to review the different conceptualizations of Wright's fixation indices using an uniform set of notations, and to examine the question of estimation of parameters and hypothesis testing in the context of an analysis of categorical data; and

(2) to document a computational algorithm for deriving the different estimators (with their standard errors, and test criteria) developed here (called WRIGHT, with three components: NEI, CLARK, and LONG) that can be used for any given data for population structure analysis.

In doing so, we provide the description of the parameters and express the estimators as functions of the observed data statistics, since there is a misconception that some of the formulations are in terms of the data statistics, and not the underlying parameters. The estimation equations are given encompassing the situations where the genotype or the allele frequencies are available. Note that when allele frequencies are used as observed data characteristics (which is usually the case for loci involving dominance relationships among alleles at a locus or in analysis of data collected from the literature), because of the lack of information on observed heterozygosities within populations, the two fixation indices F_{IT} and F_{IS} cannot be estimated, hence the only parameter that needs estimation is F_{ST}.

Empirical comparisons of these different estimators are provided with a gene diversity analysis of the populations of Sikkim, India published by Bhasin et al. (1986). Finally, we discuss the relative merits of these different estimators in terms of their complexity of computation, and generality in various practical situations. While there are several recent reviews of the difficulties of the estimators of F_{IS}, F_{ST}, and F_{IT} in the literature (see, e.g., Curie-Cohen, 1982; Robertson and Hill, 1984; Weir and Cockerham, 1984), they do not encompass all of the different estimators as fully as presented here. Consequently, these reviews do not explicitly demonstrate why the different methods of estimation produce numerically different results, or how different they can be in practice. Therefore, this review, together with the documentation of a single computer program (available from the authors upon request) should serve as an up-to-date description of the applicability of the estimators of Wright's fixation indices to the analysis of any combination of immunological (blood groups, immunoglobulin-Gm, HLA), biochemical (red-cell isozymes and serum proteins), and DNA polymorphism (Restriction Fragment Length Polymorphism, RFLP's) data in the study of the genetic structure of a subdivided population.

2. Parameters of population structure

2.1. Wright's fixation indices and Nei's gene diversity

When F_{IT} and F_{IS} are defined as correlations between two uniting gametes to produce the individuals relative to the total population and relative to the subpopulations, respectively, the correlation between two gametes drawn at random for each subpopulation (F_{ST}) is known to satisfy the identity (Wright, 1943, 1951)

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST}). \tag{2.1}$$

Consider a population which is subdivided into s subpopulations in each of which Hardy–Weinberg equilibrium (HWE) does not necessarily hold (i.e., $F_{IS} \neq 0$). For a locus with r alleles (denoted as $A_1, A_2, \ldots, A_r$), deviation from HWE can be fully specified by $\frac{1}{2}r(r-1)$ F_{IS} parameters (Rao et al. 1973).

However, if only the homozygotes are considered, r F_{IS} parameters are enough to specify deviations from HWE.

In the latter event, the frequency of the homozygotes for the k-th allele (A_kA_k) in the i-th subpopulation may be written as

$$P_{ik} = p_{ik}^2 + F_{ISik}p_{ik}(1 - p_{ik}), \tag{2.2}$$

where p_{ik} is the frequency of the A_k allele in the i-th subpopulation for $i = 1, 2, \ldots, s$; $k = 1, 2, \ldots, r$. Therefore the allele-specific F_{IS} in the i-th subpopulation can be written as

$$F_{ISik} = (P_{ik} - p_{ik}^2)/[p_{ik}(1 - p_{ik})]. \tag{2.3}$$

The deviation from HWE in the total population, with reference to the same homozygote frequency, can be parameterized in the same fashion, giving

$$P_{\cdot k} = \bar{p}_{\cdot k}^2 + F_{ITk}\bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k}), \tag{2.4}$$

where

$P_{\cdot k} = \sum_{i=1}^{s} w_i P_{ik}$ is the proportion of A_kA_k genotypes in the total population,
$\bar{p}_{\cdot k} = \sum_{i=1}^{s} w_i p_{ik}$ is the frequency of the A_k allele in the total population, and
w_i = weight of the i-th subpopulation relative to the total population size, which yields

$$F_{ITk} = (P_{\cdot k} - \bar{p}_{\cdot k}^2)/[\bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k})]. \tag{2.5}$$

With these notations the average F_{IS} (within population deviation from HWE) over all subpopulations for the k-th allele, takes the form

$$\begin{aligned} F_{ISk} &= \sum_{i=1}^{s} w_i(P_{ik} - p_{ik}^2) \Big/ \left[\sum_{i=1}^{s} w_i p_{ik}(1 - p_{ik})\right] \\ &= (P_{\cdot k} - \overline{p_{\cdot k}^2})/(\bar{p}_{\cdot k} - \overline{p_{\cdot k}^2}), \end{aligned} \tag{2.6}$$

where

$$\overline{p_{\cdot k}^2} = \sum_{i=1}^{s} w_i p_{ik}^2.$$

From equation (2.1), we therefore have

$$F_{STk} = (\overline{p_{\cdot k}^2} - \bar{p}_{\cdot k}^2)/(\bar{p}_{\cdot k} - \bar{p}_{\cdot k}^2). \tag{2.7}$$

Note that, in this formulation, the definitions of allele-specific F_{IS}, F_{IT}, and F_{ST} values are indeed parameters defined in terms of allele frequencies in the population.

Furthermore, to obtain the locus specific values of these fixation indices, we can

sum the numerators and denominators over all alleles at a locus ($k = 1, 2, \ldots, r$) to get the following formulae. From equation (2.6), we have

$$\begin{aligned} F_{IS} &= \left[\sum_{k=1}^{r}\sum_{i=1}^{s} w_i(P_{ik} - p_{ik}^2)\right] \Big/ \left[\sum_{k=1}^{r}\sum_{i=1}^{s} w_i p_{ik}(1 - p_{ik})\right] \\ &= \left[\sum_{i=1}^{s} w_i \sum_{k=1}^{r} p_{ik}(1 - p_{ik}) - \sum_{i=1}^{s} w_i\left(1 - \sum_{k=1}^{r} P_{ik}\right)\right] \\ &\quad \times \left[\sum_{i=1}^{s} w_i \sum_{k=1}^{r} p_{ik}(1 - p_{ik})\right]^{-1} \\ &= (H_S - H_0)/H_S\,, \end{aligned} \tag{2.8}$$

where

$$H_S = \sum_{i=1}^{s} w_i \sum_{k=1}^{r} p_{ik}(1 - p_{ik}) = \sum_{i=1}^{s} w_i H_{Si}$$

is the average within population heterozygosity expected under HWE, and

$$H_0 = \sum_{i=1}^{s} w_i\left(1 - \sum_{k=1}^{r} P_{ik}\right) = 1 - \sum_{i=1}^{s}\sum_{k=1}^{r} w_i P_{ik}$$

is the actual proportion of heterozygotes in the total population. Similarly, from equation (2.5),

$$\begin{aligned} F_{IT} &= \left[\sum_{k=1}^{r} (P_{\cdot k} - \bar{p}_{\cdot k}^2)\right] \Big/ \left[\sum_{k=1}^{r} \bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k})\right] \\ &= \left[\sum_{k=1}^{r} \bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k}) - \left(1 - \sum_{k=1}^{r} P_{\cdot k}\right)\right] \Big/ \left[\sum_{k=1}^{r} \bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k})\right] \\ &= (H_T - H_0)/H_T\,, \end{aligned} \tag{2.9}$$

where

$$H_T = \sum_{k=1}^{r} \bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k}) = 1 - \sum_{k=1}^{r} \bar{p}_{\cdot k}^2$$

is the heterozygosity in the total population (expected under HWE).

Lastly, from equation (2.7), we have

$$F_{ST} = (H_T - H_S)/H_T\,. \tag{2.10}$$

Therefore, estimation of the fixation indices are equivalent to estimation of the parameters H_T, H_S, and H_0 for a locus (Nei, 1973, 1977). Note that these parametric relationships also hold when the fixation indices are defined by pooling over several loci. In this case, H_S, H_T, and H_0 are the respective heterozygosities

averaged over all loci. The criticism that the relationship between fixation indices with heterozygosities is true for data statistics (and not for parameters) is not valid. Weir and Cockerham's (1984) comments are perhaps due to the misconception that under the mutation-drift model, the expectation of F_{ST} in a population with a finite number of subpopulations (expectation under the evolutionary process) is a function of the number of subpopulations as well. Two comments are worth noting at this point.

First, the above parameterization does not depend upon the evolutionary model of genetic differentiation among subpopulations, and hence the relationships (2.8)–(2.10) hold for any general mating system, irrespective of the selective differentials that may exist among the alleles. Second, even though Nei (1977) defined the pooled F_{IS} in terms of a weighted average of the subpopulation-specific F_{IS_i} values (equation (4) of Nei, 1977), with Wright's (1965) and Kirby's (1975) weight functions, such weights are not needed if we first define F_{IS}, F_{IT}, and F_{ST} as allele-specific parameters and obtain the locus-specific parameters by summing numerators and denominators over all alleles at a locus. It is also clear from equations (2.8)–(2.10) that while estimation of F_{IS} and F_{IT} would require sampling of genotypes from all subpopulations (as they are functions of the actual proportion of heterozygotes in the subpopulations, H_0), F_{ST} can be estimated with allele frequency data alone without making any assumption regarding F_{IS}. Furthermore, estimates of F_{ST} from genotype or allele frequency data should be identical, as long as the allele frequencies are obtained by the gene counting method. These issues will be detailed in the estimation section to follow.

2.2. Fixation indices and Cockerham's variance component representation

Cockerham (1969, 1973) redefined the fixation indices in terms of intra-class correlation derived from an analysis of variance of allele frequencies. In this formulation, indicator variables are defined for both alleles of a random individual sampled, which are in turn expressed as a linear model of additive effects of between-subpopulations (a), between-individuals within a subpopulation (b), and within-individual (c or w) variations. Following the classical analysis of variance model of random effects, where the subpopulations are treated as replicates of each other (Weir and Cockerham, 1984), Cockerham (1969, 1973) showed that the component of variance ascribed to the above factors (a, b and c) yield a parametric relationship with the fixation indices. In particular, Cockerham's results for a specific allele can be written in terms of our notation as

$$F_{ITk} = (a_k + b_k)/(a_k + b_k + c_k), \tag{2.11}$$

$$F_{ISk} = b_k/(b_k + c_k), \tag{2.12}$$

$$F_{STk} = a_k/(a_k + b_k + c_k), \tag{2.13}$$

where a_k, b_k, and c_k are the variance components associated with the above factors, in which the genotype frequencies in the population are tabulated with

regard to a specific allele, A_k (and thus only the frequencies of the three genotypes A_kA_k, $A_k\overline{A}_k$, and $\overline{A}_k\overline{A}_k$ enter into the analysis, $\overline{A}_k$ being a combination of all alleles of type other than A_k).

Before reviewing the estimation equations for these variance components, it is worthwhile to examine how these variance component parameters translate into the gene frequency parameters in a subdivided population.

It is easy to note that

$$a_k + b_k + c_k = \bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k}), \tag{2.14}$$

where $\bar{p}_{\cdot k} = \sum_{i=1}^{s} w_i p_{ik}$, as defined in equation (2.4). Invoking equation (2.13) into equation (2.2), we also have

$$a_k = \overline{p^2_{\cdot k}} - \bar{p}^2_{\cdot k} = \sum_{i=1}^{s} w_i (p_{ik} - \bar{p}_{\cdot k})^2, \tag{2.15}$$

and similarly, from equations (2.4) and (2.12), we have

$$\begin{aligned} b_k &= P_{\cdot k} - \overline{p^2_{\cdot k}} = \sum_{i=1}^{s} w_i (P_{ik} - p_{ik}^2) \\ &= \sum_{i=1}^{s} w_i F_{\mathrm{IS}ik} p_{ik}(1 - p_{ik}), \end{aligned} \tag{2.16}$$

since $P_{ik} = p_{ik}^2 + F_{\mathrm{IS}ik} p_{ik}(1 - p_{ik})$, according to our equation (2.2).

Putting equations (2.15) and (2.16) in equation (2.14), we get

$$c_k = \bar{p}_{\cdot k} - P_{\cdot k} = \sum_{i=1}^{s} w_i (p_{ik} - P_{ik}). \tag{2.17}$$

Since $-p_{ik}/(1 - p_{ik}) \leqslant F_{\mathrm{IS}ik} \leqslant 1$ for all $i = 1, 2, \ldots, s$ and all k, it is easy to see that $c_k \geqslant 0$. Because of equation (2.15), it is also ensured that $a_k \geqslant 0$. However, there is no guarantee that b_k is non-negative. It is, therefore, peculiar that even a parametric value of the variance component due to between individual variation can assume negative values in this formulation. Cockerham (1969) acknowledged this feature, and ascribed this to either a mating system where mates are less related than the average within a subpopulation, or to certain types of selection (Cockerham, 1969, p. 74). Since this arises for $F_{\mathrm{ST}} > F_{\mathrm{IT}}$ ($\theta > F$, in Cockerham's 1969 notation), this occurs whenever F_{IS} takes negative values (see equation (2.1)).

The above translation of parameters reveals that the negative value of b may not necessarily arise only in estimation; it is an inherent feature of the proposed linear model itself (Cockerham, 1969, 1973). It is particularly uncomfortable, since the linear model is not supposed to produce negative variance components.

In Cockerham's formulation, the locus-specific parameters are defined by

summing a_k, b_k, and c_k values over all alleles, and expressing the fixation indices as respective ratios of sums, analogous to equations (2.11)–(2.13). The same pooling algorithm is suggested for definition of parameters pooled over all loci studied (see Weir and Cockerham, 1984, equation (10)).

2.3. Long's extension of Cockerham's model

Long (1986) and Smouse and Long (1988) provided a multivariate extension of the Cockerham model, where a pair of $(r-1)$-dimensional indicator vectors is defined for a r-allelic genotypic system. This yields a multivariate decomposition of the total dispersion matrix; Σ_a, Σ_b, Σ_c in the analogy of a, b, and c of a bi-allelic locus. With $\Sigma = \Sigma_a + \Sigma_b + \Sigma_c$, the locus-specific fixation indices take the form of

$$F_{IT} = (r-1)^{-1}\,\mathrm{tr}[\Sigma^{-1/2}(\Sigma_a + \Sigma_b)\Sigma^{-1/2}]\,, \tag{2.18}$$

$$F_{ST} = (r-1)^{-1}\,\mathrm{tr}[\Sigma^{-1/2}\Sigma_a\Sigma^{-1/2}]\,, \tag{2.19}$$

$$F_{IS} = (r-1)^{-1}\,\mathrm{tr}[(\Sigma_b + \Sigma_c)^{-1/2}\Sigma_b(\Sigma_b + \Sigma_c)^{-1/2}]\,, \tag{2.20}$$

where tr denotes the trace of a matrix. In this formulation, again, while Σ_a and Σ_c are positive semi-definite matrices, the parametric form of Σ_b can be negative-definite, introducing peculiarities in the interpreting of the decomposition of dispersion matrices.

Note that for $r = 2$, equations (2.18)–(2.20) are mathematically identical to Cockerham's definition of parameters; but for $r > 2$, since equations (2.18)–(2.20) involve covariances of allele or genotype frequencies within and between subpopulations (off-diagonal elements of the Σ-matrices), the locus-specific fixation indices, according to Long's approach, are parametrically different from Weir–Cockerham's parameters. A multi-locus extension of Long's formulation is also available, where the respective Σ matrices for a group of loci are written as block-diagonal locus-specific Σ matrices (see Long, 1986, equation (8)).

In summary, the above parameterization of the genetic structure of a subdivided population indicates that Wright's fixation indices can be expressed in terms of the actual proportion of heterozygotes (H_0) and its expectation (under HWE) in the total population (H_T) and within subpopulations (H_S), without invoking any specific model of the mating system or gene differentiation between or within subpopulations. This mathematical equivalence is shown in the form of parameters, and they are consistent with Wright's identity (equation (2.1)), while the variance–component parameterization is more complex in nature, and could yield possible inconsistencies (e.g., $b < 0$, whenever F_{IS} is negative) for certain evolutionary factors (selection) or social structure of subdivision. Having defined the parameters, let us now turn to estimation and hypothesis testing issues.

3. Estimation of fixation indices

The above discussion indicates that while the heterozygosities or variance components are quadratic functions of allele and/or genotype frequencies within each subpopulation, and their weighted (by relative subpopulation size) averages, the fixation indices are ratios of functions of parameters. While estimation of a ratio of parametric functions is an unpleasant statistical problem for categorical data, to the extent that we may approximate the expectation of a ratio by the ratio of expectations, reasonable estimators of fixation indices may be obtained. Weir and Cockerham (1984) called such estimators (ratio of unbiased estimators of a numerator and a denominator) 'unbiased', while in the strict statistical sense, such estimators are at best consistent (i.e., approach the true value in terms of probability in large samples). A further problem arises, because of the categorical nature of the observations (allele frequencies, or genotype frequencies). The properties of ratio estimators are generally studied in the statistical literature for continuous traits which have Gaussian probability distributions. Even those who are concerned with distinctions between parameters and statistics have been rather cavalier about this aspect of the problem.

In this section we consider some estimators and present estimating equations in terms of the observed frequencies, which in turn indicate how much bias might arise in using these estimating equations. We might also mention that the definition of sample size has been quite elusive in the literature; because it is not always explicit whether it refers to the number of genes sampled, or that of individuals (see, e.g., Weir and Cockerham, 1984).

3.1. Estimation from genotype data

Let us first consider the case where all genotypes are recognizable, so that unequivocally all different alleles can be counted in a sample. As noted before, the F_{IT}, F_{IS}, and F_{ST} parameters depend on the sizes of subpopulations, relative to the total population size. In practice these are unknown, and furthermore, subpopulation sizes generally fluctuate over an evolutionary time period. The temporal change in population sizes has a substantial effect on the coefficients of gene-differentiation as well as heterozygosity (see, e.g., Nei et al., 1975; Chakraborty and Nei, 1977). Therefore, we shall assume all subpopulations to have equal size. This assumption is also explicitly made in Weir and Cockerham (1984, p. 1359). This, however, does not imply that the numbers of individuals sampled from the subpopulations are all equal.

3.2. Estimators of fixation indices by Nei's approach

As before consider a r-allelic locus, and define N_{ikl} to be the number of individuals of genotype A_kA_l in the i-th subpopulation ($k = 1, 2, \ldots, r$; $l = k, \ldots, r$; $i = 1, 2, \ldots, s$). Let N_i be the total number of individuals (sample size) sampled from the i-th subpopulation. The total sample size (number of individuals)

sampled from the entire subdivided population is

$$N = \sum_{i=1}^{s} N_i, \quad \text{where } N_i = \sum_{k \geq l=1}^{r} \sum N_{ikl}.$$

When $(N_{ikl};\ k = 1, 2, \ldots, r;\ l = k, \ldots, r)$ is a genotype-specific categorized subdivision of a random sample of N_i individuals from the i-th subpopulation, it is easy to note that

$$X_{ik} = N_{ikk}/N_i \tag{3.1}$$

and

$$x_{ik} = \left(2N_{ikk} + \sum_{l>k=1}^{r} N_{ikl}\right) \Big/ 2N_i, \tag{3.2}$$

are unbiased estimates of P_{ik} and p_{ik}, the proportion of A_kA_k homozygotes, and the allele frequency of A_k in the i-th subpopulation.

An unbiased estimator for p_{ik}^2 can be obtained as

$$\hat{p_{ik}^2} = x_{ik}(2N_i x_{ik} - 1)/(2N_i - 1), \tag{3.3}$$

which in turn, provides an unbiased estimator of $p_{ik}(1 - p_{ik})$, namely,

$$\frac{2N_i}{2N_i - 1}\, x_{ik}(1 - x_{ik}). \tag{3.4}$$

Note that if $x_{ik}(1 - x_{ik})$ is used as an estimator for $p_{ik}(1 - p_{ik})$, the extent of bias is

$$\begin{aligned} b &= [(2N_i - 1)/(2N_i) - 1]p_{ik}(1 - p_{ik}) \\ &= -p_{ik}(1 - p_{ik})/(2N_i), \end{aligned} \tag{3.4a}$$

i.e., $x_{ik}(1 - x_{ik})$ is an under-estimator of $p_{ik}(1 - p_{ik})$, with proportional bias being $1/2N_i$.

With equations (3.2)–(3.4), the estimator of $F_{\mathrm{IS}ik}$ (given by equation (2.3) is

$$\hat{F}_{\mathrm{IS}ik} = \frac{X_{ik} - x_{ik}(2N_i x_{ik} - 1)/(2N_i - 1)}{2N_i x_{ik}(1 - x_{ik})/(2N_i - 1)}, \tag{3.5}$$

which is a consistent estimator to the extent that the numerator and denominator of the ratio are estimated by their respective unbiased estimators.

Note that if x_{ik}^2 is used as an estimator for p_{ik}^2 (with a negative bias of the order $1/2N_i$), the estimator for $F_{\mathrm{IS}ik}$ becomes

$$\hat{F}'_{\mathrm{IS}ik} = (X_{ik} - x_{ik}^2)/[x_{ik}(1 - x_{ik})], \tag{3.5a}$$

which is identical to Curie-Cohen's (1982) estimator $\hat{f}_1 = 1 - (y/x)$, where y is the observed heterozygosity for the A_k allele in the sample, and $x = 2N_i x_{ik}(1 - x_{ik})$, an estimator of its expectation under HWE.

It might be further noted that Nei's unbiased estimator (equation (3.5)) takes the form

$$\hat{F}_{\mathrm{IS}ik} = 1 - [1 - 1/(2N_i)]\, y/x\,, \tag{3.5b}$$

which will be useful in deriving its standard error, shown in the next section.

Curie-Cohen (1982) showed that these equations have a natural multiple-allele extension, when y and x are interpreted as the total observed (H_0) and expected (H_E, under HWE) heterozygosity for all alleles at a locus. He, however, did not note the equivalence of his f_1 estimator with Nei's estimate, written in terms of H_0 and H_S, at a locus (Nei, 1977).

Let us now consider the joint analysis of data from several subpopulations. Since $\bar{p}_{\cdot k} = \sum_{i=1}^{s} w_i p_{ik}$, we have

$$\bar{p}^2_{\cdot k} = \sum_{i=1}^{s} w_i^2 p_{ik}^2 + \sum_{i \neq i'=1}^{s}\sum w_i w_{i'} p_{ik} p_{i'k}\,,$$

and hence, when the samples from the subpopulations are drawn independently of each other (as usually is the case), we obtain an unbiased estimator of $\bar{p}^2_{\cdot k}$, given by

$$\widehat{\bar{p}^2_{\cdot k}} = \sum_{i=1}^{s} w_i^2 \frac{x_{ik}(2N_i x_{ik} - 1)}{2N_i - 1} + \sum_{i \neq i'=1}^{s}\sum w_i w_{i'} x_{ik} x_{i'k}$$

$$= \left(\sum_{i=1}^{s} w_i x_{ik}\right)^2 - \sum_{i=1}^{s} w_i^2 \frac{x_{ik}(1 - x_{ik})}{2N_i - 1}\,.$$

Therefore, estimating the numerators and denominators in an unbiased fashion, we obtain the following estimators of the allele-specific fixation indices at a particular locus:

$$\hat{F}_{\mathrm{IS}k} = \frac{\sum_{i=1}^{s} w_i(X_{ik} - x_{ik}^2) + \sum_{i=1}^{s} w_i x_{ik}(1 - x_{ik})/(2N_i - 1)}{\sum_{i=1}^{s} w_i 2N_i x_{ik}(1 - x_{ik})/(2N_i - 1)}\,, \tag{3.6}$$

$$\hat{F}_{\mathrm{IT}k} = \frac{\sum_{i=1}^{s} w_i X_{ik} - (\sum_{i=1}^{s} w_i x_{ik})^2 + \sum_{i=1}^{s} w_i^2 x_{ik}(1 - x_{ik})/(2N_i - 1)}{\sum_{i=1}^{s} w_i x_{ik} - (\sum_{i=1}^{s} w_i x_{ik})^2 + \sum_{i=1}^{s} w_i^2 x_{ik}(1 - x_{ik})/(2N_i - 1)}\,, \tag{3.7}$$

$$\hat{F}_{\mathrm{ST}k} = \frac{\sum_{i=1}^{s} w_i x_{ik}^2 - (\sum_{i=1}^{s} w_i x_{ik})^2 - \sum_{i=1}^{s} w_i(1 - w_i) x_{ik}(1 - x_{ik})/(2N_i - 1)}{\sum_{i=1}^{s} w_i x_{ik} - (\sum_{i=1}^{s} w_i x_{ik})^2 + \sum_{i=1}^{s} w_i^2 x_{ik}(1 - x_{ik})/(2N_i - 1)}\,. \tag{3.8}$$

Note that while all of these estimators are consistent, to the extent that the numerators and denominators of the parameters defined in equations (2.4)–(2.6) are estimated with their respective unbiased estimators, in applying these equations we need the relative sizes (w_i's) for all subpopulations. These are, however not known in practice; nor can they always be reliably substituted by relative sample sizes. Nei (1977) and Nei and Chesser (1983), therefore assumed that the w_i's are all equal, $w_i = 1/s$ for all i. In that event, equations (3.6)–(3.8) take the form

$$\hat{F}'_{\mathrm{IS}k} = \frac{\sum_{i=1}^{s}[(X_{ik} - x_{ik}^2) + x_{ik}(1 - x_{ik})/(2N_i - 1)]}{\sum_{i=1}^{s} 2N_i x_{ik}(1 - x_{ik})/(2N_i - 1)}, \tag{3.6a}$$

$$\hat{F}'_{\mathrm{IT}k} = \frac{\sum_{i=1}^{s} X_{ik} - (1/s)(\sum_{i=1}^{s} x_{ik})^2 + (1/s)\sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N_i - 1)}{\sum_{i=1}^{s} x_{ik} - (1/s)(\sum_{i=1}^{s} x_{ik})^2 + (1/s)\sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N_i - 1)}, \tag{3.7a}$$

and

$$\hat{F}'_{\mathrm{ST}k} = \frac{\sum_{i=1}^{s} x_{ik}^2 - (1/s)(\sum_{i=1}^{s} x_{ik})^2 - (1 - 1/s)\sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N_i - 1)}{\sum_{i=1}^{s} x_{ik} - (1/s)(\sum_{i=1}^{s} x_{ik})^2 + (1/s)\sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N_i - 1)}. \tag{3.8a}$$

When the sample sizes are large enough, so that $2N_i \approx 2N_i - 1$ and $\sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N_i - 1)$ is negligible, these equations take a much simpler form:

$$\hat{F}'_{\mathrm{IS}k} \approx \sum_{i=1}^{s} (X_{ik} - x_{ik})^2 \Big/ \sum_{i=1}^{s} x_{ik}(1 - x_{ik}), \tag{3.6b}$$

$$\hat{F}'_{\mathrm{IT}k} \approx \left[\sum_{i=1}^{s} X_{ik} - \frac{1}{s}\left(\sum_{i=1}^{s} x_{ik}\right)^2\right] \Big/ \left[\sum_{i=1}^{s} x_{ik} - \frac{1}{s}\left(\sum_{i=1}^{s} x_{ik}\right)^2\right], \tag{3.7b}$$

$$\hat{F}'_{\mathrm{ST}k} \approx \left[\sum_{i=1}^{s} x_{ik}^2 - \frac{1}{s}\left(\sum_{i=1}^{s} x_{ik}\right)^2\right] \Big/ \left[\sum_{i=1}^{s} x_{ik} - \frac{1}{s}\left(\sum_{i=1}^{s} x_{ik}\right)^2\right]. \tag{3.8b}$$

Note that equation (3.8b) takes the well-known form

$$\hat{F}'_{\mathrm{ST}k} \approx s_k^2/x_{\cdot k}(1 - x_{\cdot k}), \tag{3.8c}$$

where s_k^2 is the variance of the A_k-allele frequency over all subpopulations, $s_k^2 = \sum_{i=1}^{s}(x_{ik} - \bar{x}_{\cdot k})^2/s$, with $\bar{x}_{\cdot k}$ representing the average frequency of the A_k-allele over all subpopulations, $\bar{x}_{\cdot k} = \sum_{i=1}^{s} x_{ik}/s$.

When the investigators have sufficient reason to believe that the sampling from each subpopulation has been conducted in such a manner that the relative sample sizes (N_i/N) reflect their respective relative sizes (population values), one might replace the w_i's in equations (3.6)–(3.8) by their respective sample size weights, $\hat{w}_i = N_i/N$, and obtain the allele-specific estimates of the fixation indices. However, note that while this weighting may serve the purpose of taking into account the

relative contribution of each subpopulation in the total population in the current generation, they are not evolutionary stable, as N_i's can fluctuate drastically over time.

Pooling over all alleles at a locus, the locus-specific estimates of F_{IS}, F_{IT} and F_{ST} values can be obtained easily, since the respective parameter values have been defined by summing the numerators and denominators over all alleles at a locus (see equations (2.8)–(2.10). Since these equations are represented in terms of heterozygosities in the population, it may be worthwhile to express the unbiased estimators of H_S, H_0, and H_T explicitly. Nei and Chesser (1983) obtained such estimators, with the assumption that all w_i's are equal (= $1/s$).

In our terminology, with any general weight (w_i's unequal), we may use the above mentioned unbiased estimators of p_{ik}, P_{ik}^2, and $\overline{p^2}_k$ to obtain

$$\hat{H}_0 = 1 - \sum_{i=1}^{s} \sum_{k=1}^{r} w_i \hat{P}_{ik} = 1 - \sum_{i=1}^{s} \sum_{k=1}^{r} w_i X_{ik}\,, \tag{3.9}$$

$$\hat{H}_S = 1 - \sum_{i=1}^{s} \sum_{k=1}^{r} w_i \widehat{p_{ik}^2}$$

$$= 1 - \sum_{i=1}^{s} \sum_{k=1}^{r} \frac{w_i}{2N_i - 1}\left[2N_i \sum_{k=1}^{r} x_{ik}^2 - 1\right], \tag{3.10}$$

and

$$\hat{H}_T = \sum_{k=1}^{r}\left[\sum_{i=1}^{s} w_i x_{ik}\left(1 - \sum_{i=1}^{s} w_i x_{ik}\right) + \sum_{i=1}^{s} w_i^2 x_{ik}(1 - x_{ik})\Big/(2N_i - 1)\right] \tag{3.11}$$

as respective unbiased estimators of H_0, H_S, and H_T. Substitution of these estimators in equations (2.8)–(2.10) provide consistent estimators of the locus specific fixation indices.

When the w_i's are all equal, equations (3.9)–(3.11) reduce to

$$\hat{H}_0' = 1 - \frac{1}{s} \sum_{i=1}^{s} \sum_{k=1}^{r} X_{ik}\,, \tag{3.9a}$$

$$\hat{H}_S' = \frac{1}{s} \sum_{i=1}^{s} 2N_i \hat{H}_{Si}\Big/(2N_i - 1)\,, \tag{3.10a}$$

$$\hat{H}_T' = \sum_{k=1}^{r}\left[\bar{x}_{\cdot k}(1 - \bar{x}_{\cdot k}) + \frac{1}{s^2} \sum_{i=1}^{s} x_{ik}(1 - x_{ik})\Big/(2N_i - 1)\right], \tag{3.11a}$$

where

$$\hat{H}_{Si} = 1 - \sum_{k=1}^{r} x_{ik}^2 \quad \text{and} \quad \bar{x}_{\cdot k} = \sum_{i=1}^{s} x_{ik}/s\,.$$

Note that these estimators are exact unbiased estimators of H_0, H_S, and H_T when all subpopulations are of equal size (but N_i's need not be equal), whereas

the estimators given by Nei and Chesser (1983) involve some approximations (see their equation (8) in particular). As before, when the N_i's are large, we may equate $2N_i/(2N_i - 1)$ to unity, and neglect the last term of $\hat{H}'_T$, to get

$$\hat{H}'_S \approx \frac{1}{s} \sum_{i=1}^{s} \hat{H}_{Si} \tag{3.10b}$$

and

$$\hat{H}'_T \approx 1 - \sum_{k=1}^{r} \bar{x}^2_{\cdot k} . \tag{3.11b}$$

Thus, when all genotypes are recognizable, unbiased estimators of H_0, H_S, and H_T can be obtained simply by enumerating all allele frequencies in each subpopulation (by gene counting) and evaluating the sum total of all homozygotes (X_{ik}'s). The resulting estimators

$$\hat{F}_{IS} = 1 - \hat{H}_0/\hat{H}_S , \tag{3.12}$$

$$\hat{F}_{IT} = 1 - \hat{H}_0/\hat{H}_T , \tag{3.13}$$

and

$$\hat{F}_{ST} = 1 - \hat{H}_S/\hat{H}_T , \tag{3.14}$$

are again consistent, to the extent that in these the numerators and denominators are estimated by their respective unbiased statistics.

Estimation of parameters pooled over several loci can be achieved exactly in the same manner, by defining the heterozygosities (H_0, H_S, H_T) as averages over all loci.

3.3. Estimators by Cockerham's approach

As shown in equations (2.11)–(2.13), Cockerham (1973) derived the allele-specific fixation indices in terms of components of variance in a nested analysis of variance. In this approach, the estimation of fixation indices reduces to the problem of estimating the components a_k, b_k, and c_k. Weir and Cockerham (1984) gave the explicit forms of these estimators, they are

$$\hat{a}_k = \frac{\bar{N}}{N_c} s^2_k(\hat{w}) - \frac{1}{\bar{N} - 1}\left[\bar{x}_{\cdot k}(\hat{w})(1 - \bar{x}_{\cdot k}(\hat{w})) - \frac{s-1}{s} s^2_k(\hat{w}) - \tfrac{1}{4}\bar{h}(\hat{w})\right], \tag{3.15}$$

$$\hat{b}_k = \frac{\bar{N}}{\bar{N} - 1} \bar{x}_{\cdot k}(\hat{w})[1 - \bar{x}_{\cdot k}(\hat{w})] - \frac{s-1}{s} s^2_k(\hat{w}) - \frac{2\bar{N} - 1}{4\bar{N}} \bar{h}(\hat{w}) , \tag{3.16}$$

and

$$\hat{c}_k = \tfrac{1}{2}\bar{h}(\hat{w}) , \tag{3.17}$$

where

$\bar{N} = \sum_{i=1}^{s} N_i/s$ is the average number of individuals sampled per subpopulation,
$N_c = [s\bar{N} - \sum_{i=1}^{s} N_i^2/s\bar{N}]/(s-1) = \bar{N}(1 - C^2/s)$, where C is the coefficient of variation of sample sizes (N_i's),
$\bar{x}_{.k}(\hat{w}) = \sum_{i=1}^{s} N_i x_{ik}/s\bar{N}$, the weighted average allele frequency of A_k per subpopulation,
$s_k^2(\hat{w}) = \sum_{i=1}^{s} N_i(x_{ik} - \bar{x}_{.k}(\hat{w}))^2/(s-1)\bar{N}$, is the variance of A_k allele frequencies over subpopulations,
$\bar{h}(\hat{w}) = \sum_{i=1}^{s} N_i h_i(\hat{w})/s\bar{N}$, the average observed heterozygote frequency for allele A_k.

In parallel to equations (2.11)–(2.13), the estimators $F_{\mathrm{IT}k}$, $F_{\mathrm{IS}k}$, and $F_{\mathrm{ST}k}$ become

$$\hat{F}_{\mathrm{IT}k} = (\hat{a}_k + \hat{b}_k)/(\hat{a}_k + \hat{b}_k + \hat{c}_k)\,, \tag{3.15a}$$

and

$$\hat{F}_{\mathrm{IS}k} = \hat{b}_k/(\hat{b}_k + \hat{c}_k)\,, \tag{3.16a}$$

$$\hat{F}_{\mathrm{ST}k} = \hat{a}_k/(\hat{a}_k + \hat{b}_k + \hat{c}_k)\,. \tag{3.17a}$$

Note that these expressions are defined in terms of weighted variance components, where sample sizes from the subpopulations are taken as weights, irrespective of their true relative population sizes (i.e., $\hat{w}_i = N_i/N$). These explicit forms are obtained by algebraic manipulations of the estimated mean square errors in Table 3 (Cockerham, 1973). It should be noted that Cockerham's Table 3 (Cockerham, 1973, p. 688) has an inadvertant error, where the expressions S_a and S_a' should have an additional coefficient 2, which is missing.

Cockerham (1973) also gave an explicit estimator for $F_{\mathrm{IS}ik}$, the F_{IS} estimator for a specific allele (A_k) in the i-th subpopulation, which has the form

$$\hat{F}_{\mathrm{IS}ik} = 1 - \frac{4(N_i - 1)\,[N_i x_{ik} - N_{ikk}]}{4N_i^2 x_{ik}(1 - x_{ik}) - 2(N_i x_{ik} - N_{ikk})}\,, \tag{3.18}$$

that can be computed from the respective subpopulation-specific genotype data. A pooled estimator of F_{IS}, pooled over all alleles at a locus, can be obtained by summing the numerator and the denominator of equation (3.18), as done in the other cases.

In particular, when N_i is large, the pooled estimator over all alleles at a locus takes the form

$$\hat{F}_{\mathrm{IS}i} = 1 - \frac{1}{2N_i}\left[\sum_{i=1}^{r} h_{ik}\right] \Big/ \left[1 - \sum_{i=1}^{r} x_{ik}^2\right], \tag{3.18a}$$

where h_{ik} is the observed number of heterozygotes carrying the A_k allele in the i-th subpopulation.

While equation (3.18) can be derived even without invoking the variance com-

ponents (see Cockerham, 1969, pp. 689–690), this is different from Nei's estimator (our equation (3.5)), which estimates $F_{IS\,ik}$ as a ratio estimator, based on equation (2.3). Both estimators are asymptotically unbiased (since each of them estimates the numerator and denominator by their respective unbiased statistics).

Setting up the equivalence of Cockerham's (1973) and Curie-Cohen's (1982) notations, it may be shown that the above estimator takes the form

$$\hat{F}_{IS\,ik} = [2N_i(x - y) + y]/(2N_i x - y)\,, \tag{3.18b}$$

where $x = 2N_i x_{ik}(1 - x_{ik})$, and $y\ (= \sum'_{l>k} N_{ikl})$ is the observed heterozygosity for the A_k-allele in a particular subpopulation. This equivalence will also be useful in deriving the standard error of this estimator (discussed in the next section).

At this stage, since we have three alternative estimators of $F_{IS\,ik}$: Nei's unbiased (equation (3.5)), biased (equation (3.5a)), and Cockerham's (equation (3.18)), it might be worthwhile to study how they behave for a given sample.

It can be shown that

$$\hat{F}_{IS\,ik} - \hat{F}'_{IS\,ik} = (x_{ik} - X_{ik})/2N_i x_{ik}(1 - x_{ik})\,, \tag{3.19}$$

where $\hat{F}_{IS\,ik}$ is from equation (3.5) and $\hat{F}'_{IS\,ik}$ is from equation (3.5a).

Since $x_{ik} \geqslant X_{ik}$ in any given sample, we have the inequality

$$\text{Nei's unbiased estimator} \geqslant \text{Nei's biased estimator}\,, \tag{3.20}$$

over the entire sample space.

Furthermore, the expected difference of these two estimators,

$$\mathrm{E}[\hat{F}_{IS\,ik} - \hat{F}'_{IS\,ik}] \approx (1 - F_{IS\,ik})/(2N_i - 1)\,, \tag{3.21}$$

which is usually very small, of the order $(2N_i - 1)^{-1}$.

Similarly, we can show that

$$\text{Cockerham's estimator (equation (3.18))} > \text{Nei's biased estimator (equation (3.5a))}, \tag{3.22}$$

over the entire sample space.

The relationship between Nei's unbiased and Cockerham's estimator is a little bit more involved. For simplicity, using Curie-Cohen's notation [y = observed number of heterozygotes and x = an estimator of the expected number of heterozygotes, for a specific allele = $2N_i x_{ik}(1 - x_{ik})$], we get

$$\text{Cockerham's estimator (equation (3.18)} - \text{Nei's unbiased estimator (equation 3.5)} = y/(2N_i x) \cdot \text{Cockerham's estimator (equation (3.18))}. \tag{3.23}$$

Hence, when Cockerham's estimator is negative we have the string of inequalities

$$\text{equation (3.5)} \geqslant \text{equation (3.18)} \geqslant \text{equation (3.5a)}, \tag{3.24}$$

i.e., Cockerham's estimator is bounded by Nei's biased and unbiased estimators.

However, when Cockerham's estimator is positive, from equation (3.23) we have

$$\text{equation (3.18)} \geqslant \text{equation (3.5)} \geqslant \text{equation (3.5a)}, \tag{3.25}$$

i.e., Nei's unbiased estimator is bounded by his biased estimator and that of Cockerham.

These inequalities also hold for locus-specific estimators, irrespective of the number of alleles and allele frequencies. To our knowledge, this mathematical relationship among these three estimators has not been demonstrated before. Since the expected differences are of the order of inverse of the number of genes sampled ($2N_i$) in a subpopulation, they are generally much smaller than their standard errors, which will be shown later.

It is worthwhile to note that while Nei's (1977) or Nei and Chesser's (1983) estimate of F_{STk} (see equation (3.8) or (3.8a)) is only a function of allele frequencies in all subpopulations, Weir and Cockerham's (1984) estimator of F_{STk} also depends on the frequencies of observed heterozygosity for the A_k allele in the sample.

Weir and Cockerham (1984) also gave explicit expressions for approximations for these general estimators under several special cases. In particular, they note that when the N_i's are large, the above estimators take the form

$$\hat{F}'_{ITk} = 1 - \frac{[1 - C^2/s]\bar{h}(\hat{w})}{2[1 - C^2/s]\bar{x}_{\cdot k}(\hat{w})\{1 - \bar{x}_{\cdot k}(\hat{w})\} + 2[1 + (s-1)/s \cdot C^2]s_k^2(\hat{w})/s}, \tag{3.15b}$$

$$\hat{F}'_{ISk} = 1 - \frac{\bar{h}(\hat{w})}{2\bar{x}_{\cdot k}(\hat{w})\{1 - \bar{x}_{\cdot k}(\hat{w})\} - 2(s-1)s_k^2(\hat{w})/s}, \tag{3.16b}$$

and

$$\hat{F}'_{STk} = \frac{s_k^2(\hat{w})}{[1 - C^2/s]\bar{x}_{\cdot k}(\hat{w})\{1 - \bar{x}_{\cdot k}(\hat{w})\} + [1 + (s-1)/s \cdot C^2]s_k^2(\hat{w})/s}, \tag{3.17b}$$

in which $\hat{F}'_{STk}$ can be calculated only from allele frequency data. In addition to the N_i's being large, if s (the number of subpopulations) is also large, Weir–Cockerham's estimate of F_{STk} takes the well known form of

$$\hat{F}''_{STk} = s_k^2/\bar{x}_{\cdot k}(1 - \bar{x}_{\cdot k}).$$

Note that, while the general estimator of F_{STk} in Cockerham's approach depends upon the genotype frequencies (equation (3.17a)), its large sample approximation (equation (3.17b)) is only dependent on allele frequencies.

Weir and Cockerham (1984) suggested that locus-specific estimators for F_{IS}, F_{IT}, and F_{ST} can be derived by summing $\hat{a}_k$, $\hat{b}_k$, and $\hat{c}_k$ over all alleles, so that

$$\hat{F}_{IS} = \sum_{k=1}^{r} \hat{b}_k \Big/ \sum_{k=1}^{r} (\hat{b}_k + \hat{c}_k), \tag{3.26}$$

$$\hat{F}_{IT} = \sum_{k=1}^{r} (\hat{a}_k + \hat{b}_k) \Big/ \sum_{k=1}^{r} (\hat{a}_k + \hat{b}_k + \hat{c}_k), \tag{3.27}$$

and

$$\hat{F}_{ST} = \sum_{k=1}^{r} \hat{a}_k \Big/ \sum_{k=1}^{r} (\hat{a}_k + \hat{b}_k + \hat{c}_k). \tag{3.28}$$

Although other methods of pooling data of multiple alleles exist (e.g., Wright, 1965; Kirby, 1975; Robertson and Hill, 1984), Weir and Cockerham (1984) advocate that the method presented above (equations (3.26)–(3.28)) is more appropriate for ratio estimators (see also Reynolds et al., 1983).

Note that since the parametric value of b_k can be negative (see equation (2.16)), it is quite possible that in this approach $\hat{F}_{ST}$ can often exceed $\hat{F}_{IT}$. Van Den Bussche et al. (1986) also noted that negative estimates of F_{STk} (or F_{ST}) can arise in Weir–Cockerham's approach when the following inequality holds:

$$s_k^2(\hat{w}) < \frac{1}{N-1}\left[\bar{x}_{\cdot k}(\hat{w})\{1 - \bar{x}_{\cdot k}(\hat{w})\} - \frac{s-1}{s} s_k^2(\hat{w}) - \tfrac{1}{4}\bar{h}(\hat{w})\right]. \tag{3.29}$$

While it is possible that Nei and Chesser's (1983) estimator of F_{ST} can also be negative (where $\hat{H}_S > \hat{H}_T$ occur), several simulation studies show that the negative estimates of F_{ST} are more common in the variance component approach (Chakraborty and Leimar, 1987; Van Den Bussche et al., 1986; Slatkin and Barton, 1989).

Finally, equations (3.26)–(3.28) can be extended to obtain pooled estimators of all indices, summing the numerators and denominators over all alleles over several loci (see equation (10) of Weir and Cockerham, 1984, p. 1364).

3.4. Long's estimators for multiple alleles and multiple loci

Long (1986) provided an interesting extension of Cockerham's approach for multiple alleles. He noted that when multiple alleles ($r > 2$) are involved at a locus, summation of a_k, b_k, and c_k over alleles (as suggested by Weir and Cockerham, 1984) ignores the correlation of allele and genotype frequencies (that is inherent in a multinomial sampling of genotypes) within subpopulations. Although this idea is imbedded in Weir–Cockerham's work (see their Appendix, termed as matrix estimation method), the formulation is explicitly stated in Long (1986) in terms

of the decomposition of multivariate dispersion matrices. The parameters, as defined by equations (2.18)–(2.20), can be estimated substituting the estimators for the Σ_a, Σ_b, and Σ_c matrices. Long (1986) provided computational formulae for such estimators (see Appendix of Long, 1986) which involve the genotype and allele counts within each subpopulation and their totals over all subpopulations.

Since there are several misprints in the formulae in Long's (1986) paper (see pp. 646–647), we present the general estimation procedure for a r-allelic codominant locus. This has two purposes: first, this exposition clearly indicates how Weir and Cockerham's (1984) expressions have their natural multivariate extensions and, second, it will indicate why Long's algorithm gives numerical results different from those of Weir and Cockerham for a multiallelic locus ($r > 2$). Furthermore, we derive here the explicit closed expressions for the Σ_a, Σ_b, and Σ_c matrices, that are not available in Long (1986). For a single subpopulation, closed expressions for $\Sigma_b + \Sigma_c$ matrix are also shown through this exposition.

For a specific subpopulation, when an estimator for F_{IS} is sought (in parallel to $F_{IS_{ik}}$ estimator, as done for Nei's and Cockerham's method earlier—only difference being in Long's procedure we need a different pooling algorithm over all alleles), a multivariate variance–covariance decomposition can be done in analogy of Table 3 of Cockerham (1973). The within-individual mean-square cross-product matrix (MSCP) S_c (equivalent to S_{wk} of Cockerham) for the i-th subpopulation takes the form, whose k-th diagonal element,

$$h_{ik}/2N_i, \quad \text{where } h_{ik} = \sum_{l>k=1}^{r} N_{ikl},$$

is the observed number of heterozygotes with reference to the A_k-allele in the i-th subpopulation, and the (k, l)-th off-diagonal element of the S_c matrix is $-h_{ikl}/2N_i$, where $h_{ikl} = N_{ikl}$, the observed number of A_kA_l heterozygotes in the i-th subpopulation.

Algebraic manipulation of the MSCP matrix for between individual source of variation, S_b matrix has:

$$k\text{-th diagonal element} = \frac{4N_i x_{ik}(1 - x_{ik}) - h_{ik}}{2(N_i - 1)} \tag{3.30a}$$

and

$$(k, l)\text{-th off-diagonal element} = \frac{h_{ikl} - 4N_i x_{ik} x_{il}}{2(N_i - 1)}, \tag{3.30b}$$

where the x_{ik}'s are as defined in equation (3.2).

These matrices are square matrices of dimension $r - 1$, since the linear constraint of allele frequencies (summation of all allele frequencies at a particular locus being one) has to be used in order to make such matrices non-singular (a requirement needed for the computations done in the sequel).

Estimator of Σ_b matrix (variance–covariance component due to between-

individual source of variation) is obtained as

$$\hat{\Sigma}_b = \tfrac{1}{2}[\mathrm{MSCP}(b) - \mathrm{MSCP}(c)] = \tfrac{1}{2}[S_b - S_c]\,, \tag{3.31}$$

since

$$\mathrm{E}[\mathrm{MSCP}(b)] = \Sigma_c + 2\Sigma_b \quad \text{and} \quad \mathrm{E}[\mathrm{MSCP}(c)] = \Sigma_c$$

(see Cockerham (1973; p. 688).

Therefore, $\hat{\Sigma}_b$ matrix has the form, whose

$$k\text{-th diagonal element} = \frac{4N_i^2 x_{ik}(1 - x_{ik}) - (2N_i - 1)h_{ik}}{4N_i(N_i - 1)} \tag{3.32a}$$

and

$$(k, l)\text{-th element} = \frac{h_{ikl}(2N_i - 1) - 4N_i^2 x_{ik} x_{il}}{4N_i(N_i - 1)}\,, \tag{3.32b}$$

for $k, l = 1, 2, \ldots, r - 1$.

In order to estimate $F_{\mathrm{IS}i}$, we need the matrix $\hat{\Sigma}_b + \hat{\Sigma}_c$, whose

$$k\text{-th diagonal element} = \frac{4N_i^2 x_{ik}(1 - x_{ik}) - h_{ik}}{4N_i(N_i - 1)} \tag{3.33a}$$

and

$$(k, l)\text{-th element} = \frac{h_{ikl} - 4N_i^2 x_{ik} x_{il}}{4N_i(N_i - 1)}\,, \tag{3.33b}$$

for $k, l = 1, 2, \ldots, r - 1$.

With these computations, the estimator for $F_{\mathrm{IS}i}$ is

$$\hat{F}_{\mathrm{IS}i} = \frac{1}{r - 1}\,\mathrm{tr}[(\hat{\Sigma}_b + \hat{\Sigma}_c)^{-1/2}\,\hat{\Sigma}_b(\hat{\Sigma}_b + \hat{\Sigma}_c)^{-1/2}]\,. \tag{3.34}$$

Although no closed explicit expression for $\hat{F}_{\mathrm{IS}i}$ can be given in general (for $r > 2$), the explicit expressions for the elements of $\hat{\Sigma}_b + \hat{\Sigma}_c$ and $\hat{\Sigma}_c$ matrices are instructive to understand why the numerical values of Long's estimators are different from Weir–Cockerham's estimators. For example, even if all off-diagonal elements are neglected, equation (3.34) would yield

$$\hat{F}'_{\mathrm{IS}i} = \frac{1}{r - 1}\sum_{k=1}^{r-1}\left[\frac{4N_i^2 x_{ik}(1 - x_{ik}) - (2N_i - 1)h_{ik}}{4N_i^2 x_{ik}(1 - x_{ik}) - h_{ik}}\right]$$

$$= 1 - \frac{1}{r - 1}\sum_{k=1}^{r-1}\left[\frac{2(N_i - 1)h_{ik}}{4N_i^2 x_{ik}(1 - x_{ik}) - h_{ik}}\right], \tag{3.34a}$$

whereas Weir and Cockerham's (1984) algorithm would suggest the computation of

$$\hat{F}''_{IS_i} = 1 - \left[\sum_{k=1}^{r} 2(N_i - 1)h_{ik}\right] \Big/ \left[\sum_{k=1}^{r} [4N_i^2 x_{ik}(1 - x_{ik}) - h_{ik}]\right]. \tag{3.34b}$$

While for a bi-allelic locus ($r = 2$), equations (3.34), (3.34a), and (3.34b) are identical, there are a number of practical limitations of equation (3.34) which are worth noting. For instance, suppose that there are multiple alleles ($r > 2$) in the total population, but in each subpopulation one or several are not present (either in the sample, or in the subpopulation as a whole), and the missing alleles vary across subpopulations. In such an event, for each subpopulation the S_b and S_c matrices will be of different dimension, and would refer to different sets of alleles. Therefore, in the strict sense F_{IS_i} values computed from equation (3.34) cannot be contrasted across subpopulations, since they are based on different sets of alleles even when they belong to the same locus.

Nevertheless, the large sample estimator for F_{IS_i}, following the matrix method has a closed form, not noted by Weir and Cockerham (1984) or Long (1986). Note that when the N_i's are large, ignoring terms of the order N_i^{-2}, we have

$$(\hat{\Sigma}_b + \hat{\Sigma}_c)_{kl} = \begin{cases} x_{ik}(1 - x_{ik}) & \text{for } k = l, \quad (3.33c) \\ -x_{ik}x_{il} & \text{for } k \neq l, \quad (3.33d) \end{cases}$$

for $k, l = 1, 2, \ldots, r - 1$ at a locus.

The (k, l)-th element of the $(\hat{\Sigma}_b + \hat{\Sigma}_c)^{-1}$ matrix has the form

$$(\hat{\Sigma}_b + \hat{\Sigma}_c)^{-1}_{kl} = \begin{cases} 1/x_{ik} + 1/x_{ir} & \text{for } k = l, \\ 1/x_{ir} & \text{for } k \neq l, \end{cases}$$

for $k, l = 1, 2, \ldots, r - 1$.

Therefore, if we estimate F_{IS_i} by

$$\hat{F}'_{IS_i} = (r - 1)^{-1} \operatorname{tr}[(\hat{\Sigma}_b + \hat{\Sigma}_c)^{-1} \hat{\Sigma}_b],$$

it has a closed form

$$\begin{aligned} \hat{F}'_{IS_i} &= 1 - [2N_i(r - 1)]^{-1} \sum_{k=1}^{r} h_{ik}/x_{ik} \\ &= (r - 1)^{-1} \left[\sum_{k=1}^{r} (x_{ik}/p_{ik}) - 1\right], \end{aligned} \tag{3.35}$$

while Cockerham's estimator, pooled over alleles has a large sample form given in equation (3.18a).

Note that equation (3.35) is identical to the estimator $\hat{f}_2$ used by Curie-Cohen (1982), although he arrived at this estimator by a different logic.

When several subpopulations are analysed together, nested multivariate variance–covariance analysis was performed by Long (1986), to obtain the estimators for three variance–covariance component matrices (VCCM's) as

$$S_c = \text{MSCP}(c), \tag{3.36}$$

$$S_b = \tfrac{1}{2}[\text{MSCP}(b) - \text{MSCP}(c)], \tag{3.37}$$

$$S_a = (1/2N_c)[\text{MSCP}(a) - \text{MSCP}(b)], \tag{3.38}$$

where N_c is as defined in equations (3.15)–(3.17). Here again, each of these matrices are square matrices of dimension $(r-1)$. As in the univariate case (equations (3.15)–(3.17)), explicit closed forms of these three matrices can be written which are not given in Long (1986). Long's equation for the MSCP(c) matrix (called MSCP(W) in Long, 1986) for a three allelic locus has a misprint (see his equation on top of p. 647) which fails to show how such a matrix can be computed for a multi-allelic locus. If we write the (k, l)-th element of S_a, S_b, and S_c as a_{kl}, b_{kl}, and c_{kl}, respectively, algebraic manipulation yields

$$\hat{a}_{kk} = \frac{\bar{N}}{\bar{N}_C}\left\{s_k^2(\hat{w}) - \frac{1}{\bar{N}-1}\left[\bar{x}_{\cdot k}(\hat{w})\{1-\bar{x}_{\cdot k}(\hat{w})\} - [(s-1)/s]s_k^2(\hat{w}) - \tfrac{1}{4}\bar{h}_k(\hat{w})\right]\right\}, \tag{3.37a}$$

$$\hat{a}_{kl} = \frac{\bar{N}}{\bar{N}_C}\left\{s_{kl}(\hat{w}) - \frac{1}{\bar{N}-1}\left[\bar{x}_{\cdot k}(\hat{w})\bar{x}_{\cdot l}(\hat{w}) + [(s-1)/s]s_{kl}(\hat{w}) - \tfrac{1}{4}\bar{h}_{kl}(\hat{w})\right]\right\}, \tag{3.37b}$$

$$\hat{b}_{kk} = \frac{\bar{N}}{\bar{N}-1}\left\{\bar{x}_{\cdot k}(\hat{w})\{1-\bar{x}_{\cdot k}(\hat{w})\} - [(s-1)/s]s_k^2(\hat{w}) - [(2\bar{N}-1)/4\bar{N}]\bar{h}_k(\hat{w})\right\}, \tag{3.38a}$$

$$\hat{b}_{kl} = \frac{\bar{N}}{\bar{N}-1}\left\{[(2\bar{N}-1)/4\bar{N}]\bar{h}_{kl}(\hat{w}) - \bar{x}_{\cdot k}(\hat{w})\bar{x}_{\cdot l}(\hat{w}) - [(s-1)/s]s_{kl}(\hat{w})\right\}, \tag{3.38b}$$

$$\hat{c}_{kk} = \tfrac{1}{2}\bar{h}_k(\hat{w}), \tag{3.39a}$$

$$\hat{c}_{kl} = \tfrac{1}{2}\bar{h}_{kl}(\hat{w}), \tag{3.39b}$$

for $k, l = 1, 2, \ldots, r-1$, where $\bar{x}_{\cdot k}(\hat{w})$ and $s_k^2(\hat{w})$ are as defined in the context of equations (3.15)–(3.17), and

$$s_{kl}(\hat{w}) = \left[\sum_{i=1}^{s} N_i x_{ik} x_{il} - s\bar{N}\bar{x}_{\cdot k}(\hat{w})\bar{x}_{\cdot l}(\hat{w})\right] \Big/ \bar{N}(s-1)$$

is the covariance of the allele frequencies of A_k and A_l over all subpopulations; $\bar{h}_k(\hat{w})$, the observed heterozygote frequency of the A_k allele over subpopulations ($= s\bar{N}\sum_{i=1}^{s} h_{ik}$), and $\bar{h}_{kl}(\hat{w}) = \sum_{i=1}^{s} h_{ikl}/s\bar{N}$ is the average observed frequency of a specific heterozygote A_kA_l over all subpopulations.

Note that equations (3.37a), (3.38a), and (3.39a) are identical to the A_k-allele specific variance components described by Weir and Cockerham (1984), while equations (3.37b), (3.38b), and (3.39b) are direct extensions of these with multinomial sampling of genotypes.

With these explicit general closed form expressions of the elements of S_a, S_b, and S_c matrices one can compute the F_{IS}, F_{IT}, and F_{ST} estimators:

$$\hat{F}_{IS} = \frac{1}{r-1}\,\mathrm{tr}[(S_b + S_c)^{-1/2} S_b (S_b + S_c)^{-1/2}], \tag{3.40}$$

$$\hat{F}_{ST} = \frac{1}{r-1}\,\mathrm{tr}[(S_a + S_b + S_c)^{-1/2} S_a (S_a + S_b + S_c)^{-1/2}], \tag{3.41}$$

and

$$\hat{F}_{IT} = \frac{1}{r-1}\,\mathrm{tr}[(S_a + S_b + S_c)^{-1/2}(S_a + S_b) \times (S_a + S_b + S_c)^{-1/2}], \tag{3.42}$$

with far more ease than following Long's (1986) suggestion. Note that like the one-subpopulation situation, even if the off-diagonal elements ($\hat{a}_{kl}$, $\hat{b}_{kl}$, $\hat{c}_{kl}$) are neglected, instead of Weir–Cockerham's estimates (equations (3.26)–(3.28)), equations (3.40)–(3.42) take the respective forms

$$\hat{F}'_{IS} = \frac{1}{r-1}\sum_{k=1}^{r-1} \frac{\hat{b}_{kk}}{(\hat{b}_{kk} + \hat{c}_{kk})}, \tag{3.40a}$$

$$\hat{F}'_{ST} = \frac{1}{r-1}\sum_{k=1}^{r-1} \frac{\hat{a}_{kk}}{(\hat{a}_{kk} + \hat{b}_{kk} + \hat{c}_{kk})}, \tag{3.41a}$$

$$\hat{F}'_{IT} = \frac{1}{r-1}\sum_{k=1}^{r-1} \frac{\hat{a}_{kk} + \hat{b}_{kk}}{(\hat{a}_{kk} + \hat{b}_{kk} + \hat{c}_{kk})}, \tag{3.42a}$$

which perform worse than the estimators (3.26)–(3.28) in Weir and Cockerham's (1984) simulation experiments. Furthermore, the F_{IS} estimator obtained from equation (3.40) is not a weighted average of the subpopulation-specific $\hat{F}_{IS_i}$ values obtained from equation (3.34) since for each specific subpopulation the matrices can have different dimensions for reasons stated earlier.

At this point it is worthwhile to mention that this multivariate extension has not been presented explicitly before. Although Weir and Cockerham (1984) found that the estimators by such matrix method have the smallest standard errors in comparison with various other estimators they examined, their computations of the

matrix estimators are somewhat different from those of Long (1986). Instead of $\Sigma^{-1/2}\Sigma_a\Sigma^{-1/2}$, Weir and Cockerham used $\Sigma^{-1}\Sigma_a$. Since the Σ matrices, as well as their estimators, are always symmetric square matrices, it is not clear why Long's procedure of pre- and post-multiplication with $-\frac{1}{2}$ power of the $S_a + S_b + S_c$ or $S_b + S_c$ matrices is needed. In fact, since such estimators can be computed only for non-singular $S_a + S_b + S_c$ and $S_b + S_c$ matrices, if we define the fixation indices by

$$F_{IT} = (r-1)^{-1}\,\mathrm{tr}[\Sigma^{-1}(\Sigma_a + \Sigma_b)]\,, \tag{2.18a}$$

$$F_{ST} = (r-1)^{-1}\,\mathrm{tr}[\Sigma^{-1}\Sigma_a]\,, \tag{2.19a}$$

$$F_{IS} = (r-1)^{-1}\,\mathrm{tr}[(\Sigma_b + \Sigma_c)^{-1}\Sigma_b]\,, \tag{2.20a}$$

instead of equations (2.18)–(2.20), only matrix-inversion routines are needed as opposed to the evaluation of eigen values and eigen vectors and inverse computations of the eigen vector matrices that are required in Long's algorithm.

Like the Weir and Cockerham estimator of F_{ST} (equation (3.28)), the estimator given by equation (3.41) also depends on the observed frequency of heterozygotes (see equations (3.37a) and (3.37b)) in addition to allele frequency data, which makes these estimators qualitatively different from that in Nei's approach (equation (3.8a)). Since in most practical situations the off-diagonal elements (a_{kl}, b_{kl}, c_{kl} for $k \neq l$) are small, because the subpopulations are sampled independently; the complexity of computations can be greatly reduced when Weir–Cockerham estimators are computed (according to equations (3.26)–(3.28)) for multi-allelic loci in the variance–component approach to estimation.

3.5. Estimation where genotype data are not available

Sometimes population structure analyses may have to be done in the absence of genotype data. Such is the case where the population structure is to be inferred from the allele frequency data reported in the literature, or the allele frequencies are estimated from phenotypic data at loci where complex dominance relationships exist among various alleles or haplotypes (e.g., ABO, Rh, and HLA system in man). Obviously, since such data do not provide any direct information regarding the observed number (or proportion) of homozygotes or heterozygotes, a somewhat different estimation procedure must be adopted.

In this case, Nei's approach can be easily adopted for estimating F_{ST}, since H_S and H_T parameters can be obtained simply from the estimated allele frequencies (with the assumption that the x_{ik}'s are multinomial proportions from a sample of $2N_i$ genes sampled from the i-th subpopulation). Equation (3.8) or its variant, equation (3.8a) with $w_i = 1/s$, is the estimator of preference here. Since F_{IS} is defined in terms of the deviation of genotype frequencies from their HWE expectations, no direct estimation of this quantity is possible. However, some approximate theory of estimation may be suggested.

Note that in the case of genotype data, the goodness-of-fit χ^2 statistic (of

testing for deviation from HWE expectations) for a r-allelic locus is $\chi^2 = N_i(r-1)F_{IS_i}$ (Li, 1955), and hence an approximate absolute value of F_{IS} can be obtained from $\sqrt{\chi^2/N_i(r-1)}$ where N_i is the number of individuals sampled from the i-th subpopulation. However, this suggested estimator is quite approximate, since for the loci in a dominance system, the goodness-of-fit statistic has a more complex parametric form (see Rao and Chakraborty, 1974). Furthermore, the sign of F_{IS} cannot be directly inferred from the χ^2 statistics. We advocate that for such data, only F_{ST} estimation is legitimate.

If one prefers the analysis of variance approach even the exact estimation of F_{ST} is not possible, unless large sample approximations are made. This is so because Weir and Cockerham's (1984) estimator of F_{ST} requires estimation of the observed heterozygosity for each allele (see equation (3.15) and so is the case with Long's approach (see equations (3.37a), (3.37b) and (3.41)). Under the assumption $F_{IS} = 0$ (random union of gametes within subpopulations), since $F_{IT} = F_{ST}$, Weir and Cockerham (1984, 1963) obtained the estimator

$$\hat{F}_{STk} = \frac{s_k^2(\hat{w}) - \left[\bar{x}_{\cdot k}(\hat{w})(1-\bar{x}_{\cdot k}(\hat{w})) - \frac{s-1}{s}s_k^2(\hat{w})\right] \Big/ [2\bar{N}-1]}{\left\{1 - \frac{2\bar{N}C^2}{(2\bar{N}-1)s}\right\}\bar{x}_{\cdot k}(\hat{w})\{1-\bar{x}_{\cdot k}(\hat{w})\} + \left\{1 + \frac{2\bar{N}(s-1)C^2}{(2\bar{N}-1)s}\right\}\frac{s_k^2(\hat{w})}{s}}, \tag{3.43}$$

where $\bar{x}_{\cdot k}(\hat{w})$ and $s_k^2(\hat{w})$ are the weighted mean and variance of the A_k-allele frequency over all subpopulations (defined in equations (3.15)–(3.17)), and C^2 is the coefficient of variation of N_i's over all subpopulations (note that $1 - C^2/s = N_c$, where N_c is as defined in equations (3.15)–(3.17)). Clearly this estimator depends only on allele frequency data. Therefore, the analysis of variance approach, when applied to allele frequency data, also yields a consistent estimator for F_{ST} under the assumption that $F_{IS} = 0$. For large sample sizes, this assumption is, however, not needed (see equation (3.17b).

When all subpopulations have the same sample size (i.e., $N_i = N$), equation (3.43) takes the form

$$\hat{F}_{STk} = \frac{s_k^2 - \{\bar{x}_{\cdot k}(1-\bar{x}_{\cdot k}) - [(s-1)/s]s_k^2\}/(2\bar{N}-1)}{\bar{x}_{\cdot k}(1-\bar{x}_{\cdot k}) + s_k^2/s}, \tag{3.43a}$$

which reduces to the well known formula $s_k^2/\bar{x}_{\cdot k}(1-\bar{x}_{\cdot k})$ when $\bar{N}$ and s are large.

4. Standard errors and hypothesis testing

The discussions in the earlier sections clearly indicate that the problem of estimation of the fixation indices arises because these are defined as ratios of functions of allele and genotype frequencies in the subpopulations, and hence, strictly speaking none of the estimators suggested above can be claimed most efficient. We arrived at consistent estimators by estimating the numerators and denominators by their respective unbiased statistics. Although several expressions for the standard errors of these estimators are suggested, and the question of hypothesis testing has been addressed in a variety of ways, we agree with Cockerham (1973) that such procedures are on much less sound grounds than estimation. Nevertheless, since all estimators derived above are of the general form $\hat{\theta} = t_1/t_2$, where t_1 and t_2 are the estimators of the numerators and the denominators of the respective fixation index parameters, using Taylor's expansion (Kendall and Stuart, 1977, p. 247), an approximate formula for the variance of $\hat{\theta}$ can be written as

$$V(\hat{\theta}) \approx \left[\frac{\mathrm{E}(t_1)}{\mathrm{E}(t_2)}\right]^2 \left[\frac{V(t_1)}{\mathrm{E}^2(t_1)} + \frac{V(t_2)}{\mathrm{E}^2(t_2)} - \frac{2\,\mathrm{Cov}(t_1, t_2)}{\mathrm{E}(t_1)\cdot \mathrm{E}(t_2)}\right], \tag{4.1}$$

where $\mathrm{E}(\cdot)$, $V(\cdot)$, and $\mathrm{Cov}(\cdot,\cdot)$ represent the expectation, variance, and covariance of the respective statistics.

For the analysis of data from a single subpopulation, where only F_{IS} is to be estimated, Curie-Cohen (1982) derived the sampling variance of such estimators. As shown earlier, the estimators by Nei's and Cockerham's approach (equations (3.5) and (3.18)) are related to Curie-Cohen's (1982) estimator $\hat{f}_1 = 1 - (y/x)$, for which he derived a general expression for $\mathrm{Var}(\hat{f}_1)$ at a multi-alleleic codominant locus. His expression (equation (5); Curie-Cohen, 1982, p. 345) can be further reduced to

$$V(\hat{f}_1) = \frac{(1 - F_{IS})\,[(1 - \mu_2) + (1 - F_{IS})\,(1 - \mu_2)^2 - (1 - F_{IS})^2\,(\mu_3 - \mu_2^2)]}{n(1 - \mu_2)^2}, \tag{4.2}$$

where $\mu_2 = \sum_{k=1}^{r} p_{ik}^2$ and $\mu_3 = \sum_{k=1}^{r} p_{ik}^3$, are parameters that depend upon the true allele frequencies at a locus. In practice the estimates of F_{IS}, μ_2, and μ_3 based on sample statistics can be used to estimate $V(\hat{f}_1)$. Using our equation (3.5b), we may immediately note that Nei's unbiased estimator of $F_{IS_{ik}}$ has a sampling variance

$$[1 - 1/(2N_i)]^2\, V(\hat{f}_1) \tag{4.2a}$$

while, Cockerham's estimator (equation (3.18)) has the variance

$$\frac{(1 - F_{IS})\,[(\mu_2 - 2\mu_3 + \mu_2^2) + F_{IS}(1 - 2\mu_2 + 4\mu_3 - 3\mu_2^2) - 2F_{IS}^2(\mu_3 - \mu_2^2)]}{N_i(1 - \mu_2)^2}, \tag{4.2b}$$

in which terms of the order $(2N_i)^{-2}$ or less are neglected.

As mentioned earlier, for large samples F_{IS_i} at a locus, estimated by Long's procedure, is identical to the estimator $\hat{f}_2$, used by Curie-Cohen (1982). Since he derived its sampling variance (equation (7); Curie-Cohen, 1982, p. 346), in our notation for Long's estimator we get

$$V(\hat{F}_{IS}) = \frac{1 - F_{IS}}{2N_i(r-1)^2}\left[2(r-1) - 2(2r-1)F_{IS} + r^2F_{IS}^2 + F_{IS}(2 - F_{IS})\sum_{k=1}^{r} 1/p_{ik}\right]. \quad (4.2c)$$

Equations (4.2), (4.2a), (4.2b), and (4.2c), therefore provide the approximate sampling variance of Nei's biased, Nei's unbiased, Cockerham's and Long's estimator for F_{IS} for a specific subpopulation, for any general multi-allelic codominant locus. When estimators of a specific allele are sought, the equations (4.2), (4.2a), and (4.2b) can be used taking $r = 2$, as shown for a specific case by Curie-Cohen (1982).

Although for a given sample, these sampling variances are to be evaluated with sample estimates of $F_{IS_{ik}}$, μ_2, and μ_3; it is possible to compare the relative efficiencies of Nei's unbiased (equation (3.5)), Nei's biased (equation (3.5a)), Cockerham's estimates (equation (3.18)), and its multivariate extension (equation (3.35)) by contrasting their sampling variances for known parametric values of F_{IS}, μ_2, and μ_3.

Equation (4.2a) suggests that when these parameters are fixed, Nei's unbiased estimator has a smaller sampling variance than the biased estimator. Of course, in reality, when estimates are used in variance evaluation this might not occur in a given set of data (since F_{IS} estimates would differ for these two estimators).

Note that for a bi-allelic locus (with allele frequencies p and q), equations (4.2), (4.2b), and (4.2c) all take the common form

$$N_i V(\hat{F}_{IS}) \approx \frac{1 - F_{IS}}{2pq}\left[2pq + 2(1 - 3pq)F_{IS} - (p-q)^2F_{IS}^2\right], \quad (4.3)$$

suggesting that the large-sample standard errors of Nei's biased estimator Cockerham's estimator, and Long's estimator are all identical to the extent that the terms of the order $(1/2N_i)^{-1}$ or less are neglected. Equation (4.3) is also identical to equation (3) of Curie-Cohen (1982).

To our knowledge, this equivalence of the standard errors of the different F_{IS} estimators has not been demonstrated before. In view of this mathematical equivalence, one might wonder why the empirical values of the standard errors of the different estimators vary in the simulation experiments of Weir and Cockerham (1984), Van Den Busche et al. (1986), and Chakraborty and Leimar (1987). Note that the standard error of F_{IS} is dependent on the true value of F_{IS} and the allele

frequencies at a locus (equation (4.3)). Hence, in the computation of the empirical values of the standard errors it is customary to replace the true values of the parameters by their respective estimates (i.e., $\hat{F}_{IS}$ is substituted for F_{IS}). Since we have shown earlier that the estimates differ depending upon the method of estimation satisfying the inequalities (3.24) and (3.25), it is obvious that the same analytical formula for variance (evaluated by equation (4.3)) will give different values when F_{IS} is replaced by its different estimates.

In order to study the empirical differences in the standard errors, it is therefore important to see how expression (4.3) varies as a function of F_{IS}. Curie-Cohen (1982) examined this in his Figure 1 (for a two allelic locus) and Figures 5 and 7 (for two different three allelic loci). His Figure 1 is somewhat confusing, since expression (4.3) does not decrease to zero as F_{IS} approaches its lower limit ($-p/q$ for $q > p$). Substituting $F_{IS} = -p/q$, it reduces to $p(q-p)/2q^4$, which is zero only if $p = q$. In Figure 1, we therefore plotted $\{N_i V(\hat{F}_{IS})\}^{1/2}$ as a function of F_{IS} for four values of p ($p = 0.01$, 0.1, 0.25, and 0.5). It is clear that for $F_{IS} = 0$, $V(\hat{F}_{IS}) = 1/N_i$, irrespective of the allele frequencies at a bi-allelic locus. In general, $V(\hat{F}_{IS})$ is a cubic function of F_{IS}, which attains its maximum at a value of F_{IS} depending upon the allele frequencies at the locus. When the allele frequencies are very skewed (p close to zero or one), the curve rises very fast for negative values of F_{IS}, and similarly drops fast when F_{IS} approaches one. Since Cockerham's estimator (equation (3.18)) is always larger than Nei's biased estimator (equation (3.5a)), unless the true value of F_{IS} is large, substitution of the respective estimates will yield smaller standard error for Nei's biased estimator as compared to that of Cockerham's estimator. The nature of the curves in Figure 1 indicate that such is the case for negative values of F_{IS}, irrespective of the allele frequencies at the locus. In theory, the situation can be reversed for large positive F_{IS}. But, since large positive estimates of F_{IS} are rare in natural populations (unless the organism is highly inbred), this theoretical possibility is not commonly seen. For skewed allele frequencies, the difference in the empirical values of the standard errors can be substantial, because of the sharp rise of the curve. We therefore claim that the observed discrepancies in the standard errors of the various estimators of F_{IS} are the artifacts of substituting the estimates in the variance formula (equation (4.3)). Indeed, there is no inherent difference in the standard errors, as seen in the analytical formulae established here.

Another comment regarding the standard error evaluation of Long's large-sample estimator of F_{IS} (or f_2 of Curie-Cohen, 1982) is worth mentioning at this point. Note that for a multi-allelic locus, this estimator is defined by contrasting the observed proportion of the homozygosity of each allele with the respective allele frequency (equation (3.35)). However, when equation (4.2c) is used to evaluate its standard error $\{V(\hat{F}_{IS})\}^{1/2}$, substituting $\hat{p}_{ik}$ for p_{ik} and $\hat{F}_{IS}$ for F_{IS}, one might encounter negative variance estimators, particularly when one (or more) allele is rare in the population, and $\hat{F}_{IS}$ is negative. In the application section to follow, we have several situations when it occurred. There does not appear to be any simple solution to circumvent this problem of a negative variance estimate. We simply note that the substitution of estimates for parameters (e.g., $\hat{F}_{IS}$ for F_{IS}

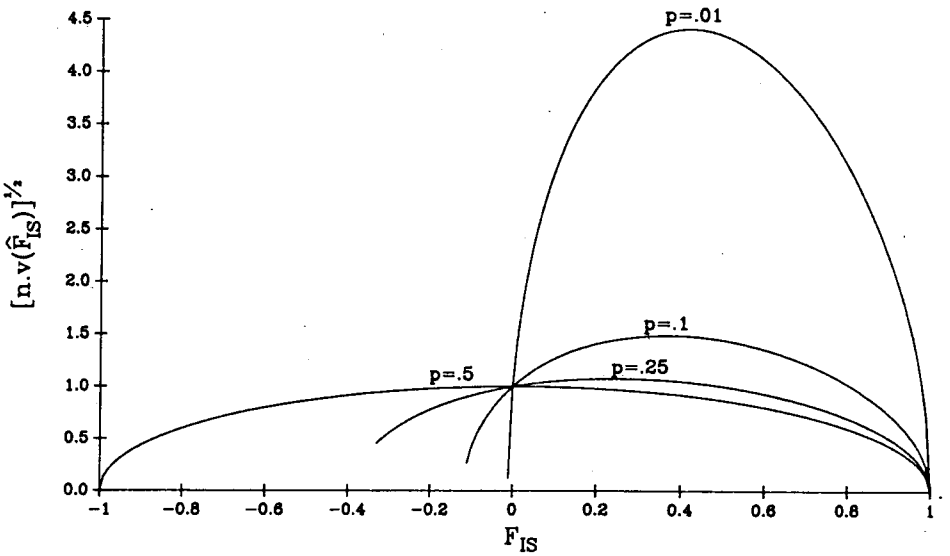


Fig. 1. Relationship between the sampling error of the large-sample estimate of F_{IS} and the true value of parameter (F_{IS}), as studied by plotting $\{n \text{ Var}(\hat{F}_{IS})\}^{1/2}$ versus F_{IS} for a bi-allelic codominant locus with allele frequencies p and q ($= 1 - p$).

and $\hat{p}_{ik}$ for p_{ik}) in equation (4.2c) yields a poor estimate of $V(\hat{F}_{IS})$ because of the inverse function of $\hat{p}_{ik}$'s (last term of equation (4.2c)).

One solution to this problem, admittedly an ad-hoc one, is to note that when some alleles are rare, since they generally appear in a sample only as heterozygotes, they do not contribute to the estimate of F_{IS} (equation (3.35)). They can be deleted in the variance computation, which is equivalent to computing the $\Sigma\,(1/\hat{p}_{ik})$ term only for alleles that contribute to the estimate of F_{IS}. This avoids the occurrence of a negative variance estimate, as seen in our empirical study. Obviously, more work is needed to provide a justifiable estimator for the standard error of $\hat{F}_{IS}$ in such situations.

The mathematical equivalence of the standard errors shown here apply only for bi-allelic loci. For a general multi-allelic locus such comparisons are more difficult, since the variances also depend on the sum of squares and cubes of allele frequencies (see equations (4.2), (4.2b), and (4.2c)). Nevertheless, for a r-allelic locus with equal gene frequencies (i.e., $p_{ik} = 1/r$ for all k), we have

$$n\,V(\hat{F}_{IS}) = \frac{(1 - F_{IS})\,[1 + (r - 1)F_{IS}]}{r - 1} \tag{4.4}$$

which holds for all of these estimators.

When data from several subpopulations are jointly used for parameter estimation, equation (4.1) can again be used to obtain approximate variances of these

estimators, in which case the variances and covariances reflect the inter-locus variation and covariation of the observed statistics. Chakraborty (1974) was the first to use this idea to evaluate the sampling variance of $\hat{F}_{ST}$, which he represented by

$$V(\hat{F}_{ST}) \approx F_{ST}^2 \left[\frac{V(\hat{H}_S)}{H_S^2} + \frac{V(\hat{H}_T)}{H_T^2} - \frac{2\,\mathrm{Cov}(\hat{H}_S, \hat{H}_T)}{H_S H_T} \right], \tag{4.5}$$

where the variances and covariances of $\hat{H}_S$ and $\hat{H}_T$ [$V(\hat{H}_S)$, $V(\hat{H}_T)$, and $\mathrm{Cov}(\hat{H}_S, \hat{H}_T)$] are obtained from inter-locus variations of these statistics. While Nei and Chakravarti (1977) demonstrated that the equation (4.5) is approximately adequate, Weir and Cockerham (1984) advocated a jackknife procedure in this context (Miller, 1974; Efron, 1982). In principle, if $\hat{\theta}$ represents an estimator of a parameter θ (not to be confused with Cockerham's notation), based on n observations, then the jackknife variance of $\hat{\theta}$ can be expressed as

$$V(\hat{\theta}) \approx \frac{n-1}{n} \sum_{i=1}^{n} \left[\hat{\theta}(i) - \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}(i) \right]^2, \tag{4.6}$$

where $\hat{\theta}(i)$ is the estimator based on $(n-1)$ observations, omitting the i-th observation. If $\hat{\theta}$ involves some bias in estimating θ (as is the case with ratio estimators), a less biased estimator of θ is

$$\hat{\theta}^* = n\hat{\theta} - [(n-1)/n] \sum_{i=1}^{n} \hat{\theta}(i). \tag{4.7}$$

This technique is adopted in estimating the standard errors of the F_{IS}, F_{IT}, F_{ST} estimators of the variance–component approach by Weir and Cockerham (1984), where jackknifing was done over loci (i.e., estimators of a, b, and c components were computed omitting one locus at a time). In particular, when the L-th locus data is omitted, the respective estimators for F_{IS}, F_{IT}, and F_{ST} used are

$$\hat{F}_{IT}(L) = \left[\sum_{l \neq L} \sum_{k} (\hat{a}_{lk} + \hat{b}_{lk}) \right] \Big/ \left[\sum_{l \neq L} \sum_{k} (\hat{a}_{lk} + \hat{b}_{lk} + \hat{c}_{lk}) \right], \tag{4.8}$$

$$\hat{F}_{IS}(L) = \left[\sum_{l \neq L} \sum_{k} \hat{b}_{lk} \right] \Big/ \left[\sum_{l \neq L} \sum_{k} (\hat{b}_{lk} + \hat{c}_{lk}) \right], \tag{4.9}$$

and

$$\hat{F}_{ST}(L) = \left[\sum_{l \neq L} \sum_{k} \hat{a}_{lk} \right] \Big/ \left[\sum_{l \neq L} \sum_{k} (\hat{a}_{lk} + \hat{b}_{lk} + \hat{c}_{lk}) \right]. \tag{4.10}$$

Note that the same approach can be adopted for Nei's estimation procedure as well, where $\hat{H}_S(L)$, $\hat{H}_T(L)$, and $\hat{H}_0(L)$ values are to be evaluated omitting the L-th locus data.

While the jackknifing over loci may provide standard errors of the estimator,

pooled over loci, there has been no explicit formulation for evaluating the sampling errors of individual allele-specific estimators. There is no simple formula for the standard errors of the variance–component estimators for a particular allele, although the sampling theory of categorical analysis of variance (CATANOVA) developed by Light and Margolin (1971), or analysis of diversity (ANODIV) of Rao (1982), indicated in Nayak (1983) may be adopted in this context. Further work is needed to provide computational formulae in this regard.

In principle, under multinomial sampling of genotypes, sampling variances of estimators $\hat{F}_{\mathrm{IT}k}$, $\hat{F}_{\mathrm{IS}k}$, and $\hat{F}_{\mathrm{ST}k}$ (equations (3.6)–(3.8)) can be derived, following Nei and Roychoudhury (1974) and Nei (1978) which refer to the sampling variance computations of heterozygosities and genetic distance. No explicit form of the intra-locus standard errors of the fixation indices are yet available.

Although the utility of the estimators is greatly increased when such standard error evaluation is available, this does not immediately resolve hypothesis testing problems, because with categorical data such ratio estimators do not have simple sampling distributions. Nayak (1983) showed that while exact sampling distributions of the variance components (or sum of squares) are not available, in large samples (N_i's large), the mean square error terms can be represented by linear combinations of χ^2 variables. However, the coefficients of such linear combinations are again not estimable, and hence such theory is difficult to apply in practice.

Cockerham (1973) suggested some heuristic test criteria for specific hypotheses. His test criteria require notations somewhat different from the rest of this paper. In order to avoid confusion, let us introduce for each allele (A_k) at a locus, three genotypes A_kA_k, $A_k\bar{A}_k$, and $\bar{A}_k\bar{A}_k$, where $\bar{A}_k$ is the combination of alleles except the A_k allele. Let M_{ik2}, M_{ik1}, and M_{ik0} be the observed frequencies of these three genotypes in a sample of N_i individuals from the i-th subpopulation. Note that M_{ikl} represents the number of individuals with l copies ($l = 0, 1, 2$) of the A_k allele in the i-th subpopulation, and $M_{ik0} + M_{ik1} + M_{ik2} = N_i$ for $i = 1, 2, \ldots, s$. As before let $N = N_1 + N_2 + \cdots + N_s$, the total number of individuals in the entire survey.

Furthermore, let x_{ik} represent the estimated allele frequency of A_k in the i-th subpopulation, given by our equation (3.2), which is equivalent to

$$x_{ik} = (2M_{ik2} + M_{ik1})/2N_i\,, \tag{3.1a}$$

Under the hypothesis that $F_{\mathrm{ST}} = 0$ and $F_{\mathrm{IS}ik} = 0$ for all i and k, the expectations of M_{ikl}'s are given by Cockerham (1973) as:

$$\bar{\eta}_{ik1} = \mathrm{E}(M_{ik1}) = 2N_i[2N/(2N-1)]\bar{x}_{\cdot k}(1-\bar{x}_{\cdot k})\,, \tag{4.11}$$

where $\bar{x}_{\cdot k} = \sum_{i=1}^{s} N_i x_{ik}/N$, the weighted average frequency of the A_k allele over all subpopulations,

$$\bar{\eta}_{ik2} = \mathrm{E}(M_{ik2}) = N\bar{x}_{\cdot k} - \tfrac{1}{2}\bar{\eta}_{ik1} \tag{4.12}$$

and

$$\bar{\eta}_{ik0} = \mathrm{E}(M_{ik0}) = N(1 - \bar{x}_{\cdot k}) - \tfrac{1}{2}\bar{\eta}_{ik1}\,, \tag{4.13}$$

so that the deviation from $F_{ST} = 0$ and $F_{IS\,ik} = 0$ can be measured by the goodness-of-fit statistic

$$\chi_1^2 = \sum_{i=1}^{s} \sum_{l=0}^{2} (M_{ikl} - \bar{\eta}_{ikl})^2 / \bar{\eta}_{ikl}\,, \tag{4.14}$$

which has a χ^2 distribution with d.f. $2s - 1$ ($2s$ independent genotypes, and one parameter, $\bar{p}_{\cdot k}$ being estimated).

The test-statistic for $F_{IS\,ik} = 0$ for all i and k, given by Cockerham (1973) is the sum-total of $s\chi^2$ values measuring deviations from HWE within individual subpopulations. However, since the unbiased estimator of the $A_k\bar{A}_k$ heterozygote proportions in the i-th subpopulation, is $2N_i x_{ik}(1 - x_{ik})/(2N_i - 1)$, under this hypothesis the expectations of M_{ikl}'s are given by

$$\hat{\eta}_{ik1} = \mathrm{E}(M_{ik1}) = 4N_i^2 x_{ik}(1 - x_{ik})/(2N_i - 1)\,, \tag{4.15}$$

$$\hat{\eta}_{ik2} = \mathrm{E}(M_{ik2}) = N_i x_{ik} - \tfrac{1}{2}\hat{\eta}_{ik1}\,, \tag{4.16}$$

$$\hat{\eta}_{ik0} = \mathrm{E}(M_{ik0}) = N_i(1 - x_{ik}) - \tfrac{1}{2}\hat{\eta}_{ik1}\,. \tag{4.17}$$

Departure from this hypothesis can be tested by the χ^2 statistic

$$\chi_2^2 = \sum_{i=1}^{s} \sum_{l=0}^{2} (M_{ikl} - \hat{\eta}_{ikl})^2 \Big/ \hat{\eta}_{ikl}\,, \tag{4.18}$$

with s d.f.

Cockerham (1973) suggested $\chi_1^2 - \chi_2^2$ as the test criterion with d.f. $s - 1$ for testing the hypothesis $F_{ST} = 0$. While this may approximately hold for large samples, when the A_k-allele is rare in one or more subpopulations, because of small values of M_{ikl}, or $\hat{\eta}_{ikl}$, or $\hat{\eta}_{ikl}$ this approximation may not be accurate. Workman and Niswander (1970) suggested a more direct test for $F_{ST} = 0$ by the usual χ^2 test of heterogeneity (Rao, 1965, p. 323) which is commonly employed in most anthropogenetic studies (see, e.g., Chakraborty et al., 1977).

When $F_{IS\,ik}$ is assumed to be equal in all subpopulations, the test for $F_{IS} = 0$ (common value over all subpopulations) can also be tested with a χ^2 statistic. In this case, the expectations of M_{ikl}'s are computed as

$$\hat{\eta}_1 = \mathrm{E}\left(\sum_{i=1}^{s} M_{ik1}\right) = 4N \sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N - s)\,, \tag{4.19}$$

$$\hat{\eta}_2 = \mathrm{E}\left(\sum_{i=1}^{s} M_{ik2}\right) = N\bar{x}_{\cdot k} - \tfrac{1}{2}\hat{\eta}_1\,, \tag{4.20}$$

$$\hat{\eta}_0 = \mathrm{E}\left(\sum_{i=1}^{s} M_{ik0}\right) = N(1 - \bar{x}_{\cdot k}) - \tfrac{1}{2}\hat{\eta}_1\,, \tag{4.21}$$

which yields

$$\chi_5^2 = \sum_{l=0}^{2} \left[\sum_{i=1}^{s} M_{ikl} - \hat{\eta}_l \right]^2 \Big/ \hat{\eta}_l , \tag{4.22}$$

which also has a χ^2 distribution with one d.f.

In a similar vein, Cockerham (1973) suggested a test criterion for $F_{STk} = 0$ from allele frequency data, which takes the form

$$\chi_k^2 = \left[\sum_{i=1}^{s} 2N_i [x_{ik} - \bar{x}_{\cdot k}(\hat{w})]^2 \right] \Big/ \{ \bar{x}_{\cdot k}(\hat{w}) [1 - \bar{x}_{\cdot k}(\hat{w})] \} , \tag{4.23}$$

with $(s - 1)$ d.f., for each specific allele A_k. Although for genotypic data, several alternative test criteria for F_{ST} exist, there is no definitive theory that suggests which should be the preferred one. We might note that expression (4.23) is the most commonly employed test criterion for F_{STk} in empirical studies of population structure (see also Workman and Niswander, 1970).

Although the test criteria (4.13), (4.18), (4.22) and (4.23) have their own intuitive appeal, Rao (1982) and Nayak (1983) showed that when the N_i's are not equal, these χ^2 statistics do not quite reflect an orthogonal decomposition of the total sum of squares in terms of a categorical analysis of variance. Further investigation is needed to address the question of most powerful test criteria in the analysis of such data. Furthermore, since these statistics refer to a single allele (A_k), a combined analysis for multiple allelic loci is not provided by these test criteria.

Long (1986) approached this problem while providing locus-specific estimates of the fixation indices. As shown earlier, Long's (1986) estimators are derived in terms of the three MSCP matrices: MSCP(a), MSCP(b), and MSCP(c), respectively (in Long's notation MSCP(c) = MSCP(W)). He suggested that the significance of F_{IS}, F_{IT} can be tested by

$$\Lambda_1^* = \det[\text{MSCP}(c)]/\det[\text{MSCP}(b)] \approx \Lambda(G, N - s, N) , \tag{4.24}$$

$$\Lambda_2^* = \det[\text{MSCP}(b)]/\det[\text{MSCP}(a)] \approx \Lambda(G, s - 1, N - s) , \tag{4.25}$$

and

$$\Lambda_3^* = \det[(N - 2)\,\text{MSCP}(b) + 2\text{MSCP}(a)]/\det[N\,\text{MSCP}(c)] \approx \Lambda(G, N - 1, N) , \tag{4.26}$$

respectively, where $\det(Z)$ is the determinant of a matrix Z, G is the dimension of S-matrices (number of independent alleles); and $\Lambda(\text{df}_1, \text{df}_2, \text{df}_3)$ is a Wilk's Λ variate with d.f. df_1, df_2, and df_3 (Anderson, 1984, p. 299).

Although the rationale of these test criteria results from the convergence of the multinomial to the multivariate normal distribution for fairly large sample sizes, there are several problems with these test statistics. First, for unequal sample sizes S_a, S_b, and S_c are not independently distributed (and so too are their respective

MSCP matrices). Nayak (1983) showed that their correlations can be quite substantial, and hence, the criteria Λ_1^*, Λ_2^*, and Λ_3^* do not satisfy the conditions under which Wilk's Λ distribution is valid (see equation (3) of Anderson, 1984, p. 299). Second, the assumption that a MSCP matrix follows a Wishart distribution is true for multivariate normal variates. A multinomial sampling of genotypes where one or more alleles are rare in the population and consequently may be absent in one or more subpopulations, will not approach multivariate normality unless the sample sizes are very large. Third, Wilk's Λ distribution approximation will also require a large number of subpopulations in addition to large N_i values. Since, in the earlier sections we showed that a great deal of work is needed to reduce bias due to small N_i and s values in estimating the fixation indices, the attempt to sweep out all these troubles by using such approximations cannot be generally advocated. Fourth and lastly, as indicated earlier, the variance–component approach may yield a negative-definite MSCP(B) matrix (see Cockerham, 1969, p. 74 for the univariate result), which also makes the Λ-distribution approximation invalid.

In summary, we argue that a rigorous test procedure for studying the significance of the fixation indices is not yet available. All suggested test criteria are only approximate, and caution must be exercised in interpreting their results.

5. An application

Bhasin et al. (1986) studied the genetic structure of the people of Sikkim of North India in order to determine the extent of genetic differentiation among the various subdivisions of their social units. They recognized 13 social groups in this population: North Sikkim, Sherpas, Tamangs, Gurungs, Rais, Limboos, Pradhans, Brahmins, Chhetris, and Scheduled Castes who are ethnohistorically as well as socially isolated to a certain extent. They studied 17 polymorphic blood groups and protein loci in each of these subpopulations. Of these loci, 11 are codominant. Haptoglobin (Hp), Group-Specific Component (Gc), Transferrin (Tf), Acid phosphatase (aP), Phosphoglucomutase-1 (PGM_1), 6-phosphogluconate dehydrogenase (6-PGD), Esterase D (EsD), Adenylate kinase (Ak), Hemoglobin (Hb), Duffy (Fy), and Kidd (Ik), at which the number of detected alleles vary from 2 to 5 (Gc and aP have 3 alleles each, Tf has 5 alleles, all of the remaining having 2 alleles each). The remaining six loci—AB0, MNSs, Rh, Kell and Immunoglobulin Gm and Km—have variable degrees of complex dominant relationships among their alleles/haplotypes, so that at each of these loci not all genotypes are distinguishable. The genotype/phenotype data at these loci for each subpopulation are presented in Bhasin et al. (1986).

We consider this survey for illustrating empirically the differences in the various estimators for several reasons. First, as part of a larger study of the extent of genetic differentiation among the populations of Sikkim, the estimates of the fixation indices provide the basis of our further studies, and hence it is important to determine their stability over different estimation methods employed. Second,

as the sample sizes of this study drastically differ over subpopulations as well as over loci, this example should also provide insight regarding the stability of parameter estimates with or without invoking the large sample approximations discussed in our theoretical exposition. Finally, the availability of loci with and without dominance relationships in this study will help us to examine some features of the statistical properties of the parameter estimates based on genotype vs. allele frequency data, not commonly found in all applications of this nature. Notwithstanding these issues, we should note that since this review deals with a comparative study of the various estimation procedures, and *not* the population structure of the Sikkimese people, only the results pertaining to the comparative analyses are reported here.

5.1. *Comparison of the estimates of F_{IS} in a single subpopulation*

We have seen earlier that estimation of F_{IS} is possible from genotype data by several methods. Since only codominant loci can be used for this purpose, Table 1 contrasts the estimates of F_{IS} in the Lepchas of North Sikkim for 11 loci, as computed using equations (3.5) (Nei's unbiased estimator), (3.5a) (Nei's biased estimator) and (3.18) (Cockerham's estimator). Although several other estimators of F_{IS} are available in such situations (Li and Horvitz, 1953; Curie-Cohen, 1982; Robertson and Hill, 1984), we contend that the data presented in Table 1 are sufficient to contrast most of the theoretically justifiable estimators. Note that Curie-Cohen's (1982) estimator $\hat{f}_1$ is identical to Nei's biased estimator, and his $\hat{f}_2$ is exactly the same as the multivariate large sample estimate at a locus, while his $\hat{f}_3$ estimator can be computed from the χ^2 values presented in this table. We computed two different χ^2 values, one corresponding to Nei's unbiased estimator (evaluated by equation (4.18) with $s = 1$), and the other corresponding to the biased estimator (where the expected genotype frequencies are computed by $N_i x_{ik}^2$ for the genotype $A_k A_k$ and $2N_i x_{ik} x_{il}$ for the genotype $A_k A_l$). The χ^2 values for Cockerham's estimator are identical to those of Nei's unbiased estimator, and hence they are not repeated in the table. The allele-specific χ^2's have a single d.f. in every case, while the locus-specific χ^2's have d.f. $r(r-1)/2$, where r is the number of segregating alleles at a locus in the specific subpopulation. Since for the bi-allelic loci the estimators and χ^2 values are identical for both alleles, and hence their locus-specific values are exactly the same as those based on any specific allele, only allele-specific estimates are given for such loci (Hp, PGM_1, PGD, EsD, Ak, Hb, Duffy, and Kidd in our example). Further note that although the transferrin locus has 5 segregating alleles in the total Sikkim population, in Lepchas of North Sikkim only 3 segregating alleles were found (Bhasin et al., 1986), and hence this was treated as a 3-allelic locus for this computation.

We chose not to present the estimator of F_{IS} based on χ^2 values for two reasons. First, since χ^2 values represent deviation from HWE for the two-sided alternative $F_{IS} \neq 0$, the sign of F_{IS} cannot be inferred from the value of χ^2 (Li and Horvitz, 1953; Curie-Cohen, 1982), and second, the χ^2 values can be greatly

Table 1
Allele-specific estimates of $F_{IS/k}$ for the Lepchas sampled in North Sikkim

Locus	Allele	N	Frequency	Nei's unbiased estimate			Nei's biased estimate			Cockerham's estimate	Long's estimate	
				$F_{IS/k} \pm$ s.e.	χ^2	d.f.	$F_{IS/k} \pm$ s.e.	χ^2	d.f.	$F_{IS/k} \pm$ s.e.[a]	Weighted[b]	Large sample[c]
Hp	Hp^1	65	0.1154	0.177 ± 0.162	2.18	1	0.171 ± 0.163	1.90	1	0.179 ± 0.164	0.179 ± 0.127	0.171 ± 0.163
Gc	Gc^{1F}		0.6129	0.393 ± 0.119	9.75[f]	1	0.388 ± 0.120	9.34[f]	1	0.395 ± 0.120		
	Gc^{1S}		0.2581	0.415 ± 0.130	10.97[g]	1	0.410 ± 0.131	10.44[g]	1	0.417 ± 0.131		
	Gc^2		0.1290	0.004 ± 0.127	0.00	1	−0.005 ± 0.126	0.00	1	0.004 ± 0.128		
	Pooled	33		0.320 ± 0.106	13.70[f]	3	0.314 ± 0.107	13.44[f]	3	0.322 ± 0.107	0.233	0.225 ± 0.108
Tf	Tf^{C1}		0.7097	−0.087 ± 0.120	0.48	1	−0.096 ± 0.120	0.57	1	−0.088 ± 0.121		
	Tf^{C2}		0.2823	−0.066 ± 0.121	0.28	1	−0.075 ± 0.122	0.35	1	−0.067 ± 0.122		
	Tf^D		0.0081	0.000 ± 0.126	0.00	1	−0.008 ± 0.008	0.00	1	0.000 ± 0.127		
	Pooled	62		−0.75 ± 0.117	0.74	3	−0.084 ± 0.117	0.82	3	−0.076 ± 0.118	−0.037	−0.047 ± 0.124[d]
AP	P^a		0.1638	−0.061 ± 0.115	0.22	1	−0.070 ± 0.113	0.28	1	−0.061 ± 0.116		
	P^b		0.8190	0.020 ± 0.134	0.02	1	0.012 ± 0.133	0.01	1	0.020 ± 0.135		
	P^c		0.0172	−0.009 ± 0.093	0.01	1	−0.018 ± 0.012	0.02	1	−0.009 ± 0.094		
	Pooled	58		−0.018 ± 0.116	1.79	3	−0.027 ± 0.115	1.86	3	−0.018 ± 0.117	−0.028	−0.037 ± 0.122[d]
PGM_1	PGM_1^1	50	0.6600	−0.324 ± 0.115	5.37[e]	1	−0.337 ± 0.114	5.68[e]	1	−0.328 ± 0.115	−0.328 ± 0.115	−0.337 ± 0.114
PGD	PGD^A	53	0.8679	0.022 ± 0.143	0.03	1	0.012 ± 0.141	0.01	1	0.022 ± 0.144	0.022 ± 0.144	0.012 ± 0.141
EsD	EsD^1	50	0.7400	0.485 ± 0.139	12.17[g]	1	0.480 ± 0.141	11.53[g]	1	0.488 ± 0.141	0.488 ± 0.141	0.480 ± 0.141
AK	AK^1	58	0.9914	0.000 ± 0.130	0.00	1	−0.009 ± 0.009	0.00	1	0.000 ± 0.131	0.000 ± 0.131	−0.009 ± 0.009
Hb	Hb^A	61	0.9754	−0.017 ± 0.075	0.03	1	−0.025 ± 0.015	0.04	1	−0.017 ± 0.075	−0.017 ± 0.075	−0.025 ± 0.015
Duffy	Fy^a	66	0.8485	0.181 ± 0.149	2.27	1	0.175 ± 0.149	2.02	1	0.182 ± 0.150	0.182 ± 0.150	0.175 ± 0.149
Kidd	Ik^a	47	0.4681	−0.268 ± 0.139	3.45	1	−0.282 ± 0.139	3.73	1	−0.272 ± 0.140	−0.272 ± 0.140	−0.282 ± 0.139

[a] The χ^2 values for Cockerham's estimate of $F_{IS/k}$ are exactly the same as that for Nei's unbiased estimates.
[b] Long's weighted estimates are locus-specific estimates, which are identical to Cockerham's estimator for two-allelic loci.
[c] Long's large sample estimator is identical to that of Curie-Cohen's (1982) estimator f_2, and hence their standard errors are also the same (see text for details).
[d] These s.e. values are computed deleting the alleles that do not contribute to the estimate (see text for details).
[e] $p < 0.05$.
[f] $p < 0.01$.
[g] $p < 0.001$.

affected by rare genotypes and their expected values, giving unstable estimates in specific situations, an example of which will be discussed later in this section.

The standard errors of the three estimators are evaluated by equation (4.2) (for Nei's biased estimator), (4.2a) (for Nei's unbiased estimator) and (4.2b) (for Cockerham's estimator), where $r = 2$ is used for allele specific values, and the entire locus data used for locus-specific standard errors (represented by s.e. in the table).

In terms of the values of the estimates, it is clear that Nei's unbiased, biased, and Cockerham's estimates of F_{IS} are quite close to each other, with biased estimates always being the smallest. The differences of these estimates (the allele-specific ones as well as their pooled values over all alleles at a locus) are always encompassed by their respective standard errors (see Table 1).

The standard errors of Nei's unbiased and Cockerham's estimates are also very similar, while for negative $\hat{F}_{IS_{ik}}$ (or $\hat{F}_{IS_i}$) values Cockerham's estimators have slightly larger standard errors, the situation reverses when the estimates are positive. The differences in the standard errors are however very small, and in no case change the qualitative results of hypothesis testing ($F_{IS} = 0$) either by the χ^2 value shown in the table, or by a crude test of the normal deviate $[\hat{F}_{IS_{ik}}/\text{s.e.}(\hat{F}_{IS_{ik}})]$–the latter test not explictly shown in this table. As noted earlier, for the bi-allelic loci the differences in the standard errors are produced only because of substituting the respective estimates of F_{IS} in equation (4.3). Curie-Cohen (1982) also showed that in multiallelic loci the standard errors of the various estimators are only slightly different (see Figure 5 of Curie–Cohen, 1982, p. 352).

A comment regarding the standard errors of Nei's biased estimators is worth noting. While these are quite close to those of Nei's unbiased and Cockerham's estimators, where the allele frequency is close to 0 or 1 (e.g., Tf^D, p^C, Ak^1, and Hb^A) the s.e.'s of the biased estimators are substantially smaller than those of Nei's unbiased and Cockerham's estimators. This feature may not be intuitively clear. Nevertheless, Figure 1 indicates that for skewed allele frequencies (p, small), the s.e. of Nei's biased estimate sharply rises from a very small value in the range of negative F_{IS} values. Since we evaluated the s.e. of each estimate by substituting the obtained F_{IS} estimates of the same method, these computations are indeed a comparison of different trajectories. For example, in the case of the Tf^D allele at the Transferrin locus, the standard error of Nei's unbiased estimator is evaluated at $F_{IS} = -0.008$, while that of Nei's unbiased and Cockerham's estimates is evaluated for $F_{IS} = 0.0$. The frequency of this allele in the Lepcha subpopulation is 0.008. For this allele frequency, even for the biased estimator of Nei, the s.e. rises from 0.008 to 0.127 as F_{IS} is changed from -0.008 to 0.0. Therefore, the differences in the standard errors noticed in Table 1 are largely due to the fact that the estimates are somewhat different in these three methods. When allele frequencies are at an intermediate range, small differences between parameter estimates do not substantially change the standard errors, but for skewed allele frequencies even minute changes in the estimates can induce a large difference in standard errors, particularly when the F_{IS} estimate is negative. In spite of such

differences, there is no change in the conclusion regarding hypothesis testing either from the χ^2 values or from normal deviates. Even though we do not present similar analyses for the other 12 subpopulations studied by Bhasin et al. (1986), this statement is valid in general.

Table 1 further shows that of the 17 allele-specific tests performed, significant deviation from $F_{IS} = 0$ is found in 4 cases, due to 3 loci (Gc, PGM_1, and EsD). One of these significant deviations is due to a negative F_{IS} (at the PGM_1 locus). This finding is consistent for all three methods employed in this analysis. We also have evidence of negative significant F_{IS} values for the Tf^{C1} allele in Tamangs and Scheduled Castes, Tf^{C2} allele in Rais, Gurungs and Scheduled Castes and for the Kidd locus (either allele) in Rais and Gurungs.

Table 2 presents a summary of the significant (at 5% level) positive and negative $\hat{F}_{IS/k}$ (or $\hat{F}_{IS/}$) values in 158 independent allele-specific and 127 locus-specific tests in the total data on 11 loci in the 13 subpopulations mentioned earlier. For allele-specific tests five test procedures are considered in this table: 2 χ^2 tests (one based on biased estimates of genotype frequencies, and the other based on unbiased estimates), and 3 normal deviates (based on Nei's biased and unbiased estimators, and that of Cockerham). For locus-specific tests, in addition to the above five test procedures, normal deviates based on Long's large-sample estimates of $F_{IS/}$ (which is identical to the estimator $\hat{f}_2$ of Curie-Cohen, 1982) are also used, since the standard error of such estimators is known (see equation (4.2c)).

The total number of significant deviations from $F_{IS} = 0$ is almost the same for each χ^2 statistic. The normal deviates based on Nei's unbiased and Cockerham's estimators also behave identically, as do the normal deviates based on Nei's biased and Long's large-sample estimates in the case of locus-specific tests. Furthermore, the numbers of positive and negative significant F_{IS} values according to the χ^2 statistics are not equal; there are far more positive significant values than negative ones.

Table 2
Number of significant ($p < 0.05$) F_{IS} values in the Sikkim survey as detected by various estimators

Test criterion	Allele-specific tests with $F_{IS/k}$ value[a]		Locus-specific tests with $F_{IS/}$ value[b]	
	Positive	Negative	Poitive	Negative
χ^2: Unbiased	22	7	12	5
Biased	20	10	10	7
Normal Deviate based on				
Nei's unbiased estimate	16	17	8	14
Nei's biased estimate	17	29	18	27
Cockerham's estimate	16	18	9	13
Long's large sample est.	–	–	18	25

[a] Total number of independent allele-specific tests = 158.
[b] Total number of independent locus-specific tests = 127.

These features are not unique to this data alone, and can be explained on the basis of the theory we presented before. First, note that χ^2 statistics only detect deviation in either direction, and since the range of negative F_{IS} is narrower ($F_{ISik} \geq -p_{ik}/(1-p_{ik})$ for every allele A_k) than the range of positive F_{IS} ($F_{IS} \leq 1$), it is expected that more significant positive F_{IS} values will be encountered based on χ^2 goodness-of-fit. Second, since Nei's biased estimator has empirically smaller sampling variance than that of Cockerham's estimator for negative F_{IS} (Figure 1) for all allele frequencies (unless the alleles are equi-frequent), it is expected that this will pick up more significant negative $\hat{F}_{IS}$ values than the normal deviates based on Cockerham's estimator. This is also predicted from Figure 1, which shows that the sampling variance sharply drops off even if the $\hat{F}_{ISik}$ values are slightly decreased, particularly when $\hat{F}_{ISik}$ is negative. Since in all cases Nei's biased estimate is smaller than all other estimates, a normal deviate based on this estimator would necessarily pick up more significant negative F_{IS} values as compared to any other test criteria.

The estimate of F_{IS} for a single subpopulation, combining all alleles at a locus shows exactly the same picture. We have not explicitly shown the behavior of the test criteria based on Long's weighted estimator for the reason that its sampling variance is not yet available. However, its large-sample variance can be computed based on (4.2c), which is used in the computations shown in this table.

5.2. *Comparison of F_{IT}, F_{IS}, and F_{ST} estimates over all subpopulations*

In Tables 3, 4, and 5 we provide a comparative study of the estimators of the three fixation indices over all 13 subpopulations of Sikkim. Nei's weighted, unweighted, and large sample estimates are computed by equations (3.6)–(3.8), (3.6a)–(3.8a), and (3.6b)–(3.8b), respectively. While the standard errors of these estimators for allele- and locus-specific cases cannot be evaluated, approximate tests for allele-specific F_{IS} and F_{ST} values may be conducted by χ^2 statistics, according to equations (4.22), and (4.23), respectively. These results are shown in Table 3. In Table 4 computations of Cockerham's estimators are shown for the same data. In addition to the weighted (equations (3.15a), (3.16a), and (3.17a)) and large sample estimates (equations (3.15b), (3.16b), and (3.17b)), Cockerham's estimators of F_{ST} are also obtained under the assumption of $F_{IS} = 0$ (equation (3.43)) whose values and χ^2 test criteria are shown in this table. It should be noted that the χ^2 test criteria for F_{IS} and F_{ST} for Cockerham's general estimates are exactly the same as those of Nei's weighted estimators (shown in Table 3), and hence they are not repeated in Table 4. Multivariate estimators, according to the generalized formulation of variance component analysis (equations (2.18a)–(2.20a)), are presented in Table 5. Only locus-specific estimates are needed here, and for two-allelic loci these estimates are identical to those of Cockerham's weighted analysis, as shown earlier.

Several interesting findings emerge from these computations. First, the weighted estimators of each fixation index are nearly the same for Cockerham's and Nei's method. Second, while the F_{ST} estimates of Nei's weighted and unweighted

Table 3
Nei's allele- and locus-specific F_{IT}, F_{IS}, and F_{ST} estimates

Locus	Allele	Weighted							Unweighted			Large N		
		F_{IT}	F_{IS}	χ^2	d.f.	F_{ST}	χ^2	d.f.	F_{IT}	F_{IS}	F_{ST}	F_{IT}	F_{IS}	F_{ST}
Hp	Hp^1	−0.007	−0.021	0.25	1	0.013	23.96[a]	12	−0.013	−0.029	0.016	−0.014	−0.042	0.027
Gc	Gc^{1F}	0.278	0.214	23.48[c]	1	0.081	133.39[c]	12	0.261	0.190	0.087	0.260	0.181	0.097
	Gc^{1S}	0.238	0.192	19.76[c]	1	0.057	92.08[c]	12	0.207	0.161	0.055	0.206	0.151	0.066
	Gc^2	0.188	0.162	14.08[c]	1	0.031	43.86[c]	12	0.193	0.166	0.032	0.192	0.156	0.043
	Pooled	0.240	0.192			0.059			0.224	0.173	0.061	0.223	0.163	0.072
Tf	Tf^{C1}	−0.151	−0.163	14.62[c]	1	0.010	26.20[a]	12	−0.162	−0.177	0.013	−0.163	−0.193	0.025
	Tf^{C2}	−0.166	−0.176	17.20[c]	1	0.009	22.72[a]	12	−0.179	−0.192	0.011	−0.180	−0.209	0.024
	Tf^{C3}	−0.009	−0.017	0.11	1	0.008	24.78[a]	12	−0.011	−0.018	0.006	−0.013	−0.034	0.021
	Tf^{C12}	−0.003	−0.009	0.03	1	0.006	23.39[a]	12	−0.004	−0.010	0.006	−0.005	−0.027	0.022
	Tf^{D}	−0.003	−0.003	0.00	1	−0.000	7.66	12	−0.002	−0.002	−0.000	−0.002	−0.011	0.008
	Pooled	−0.152	−0.163			0.009			−0.163	−0.177	0.012	−0.164	−0.193	0.024
aP	P^A	0.080	0.067	2.28	1	0.013	29.55[b]	12	0.055	0.039	0.016	0.054	0.026	0.029
	P^B	0.088	0.077	2.98	1	0.012	28.81[b]	12	0.065	0.051	0.014	0.064	0.038	0.027
	P^C	0.193	0.187	9.40[b]	1	0.008	458.95[c]	12	0.253	0.246	0.010	0.252	0.234	0.024
	Pooled	0.087	0.075			0.013			0.066	0.052	0.015	0.065	0.039	0.028
PGM_1	PGM_1^1	0.025	0.025	0.27	1	−0.000	11.98	12	0.070	0.074	−0.005	0.068	0.057	0.013
PGD	PGD^A	0.150	0.127	6.09[a]	1	0.26	30.95[b]	12	0.176	0.147	0.034	0.175	0.133	0.048
EsD	EsD^1	0.145	0.143	11.19[c]	1	0.003	13.78	12	0.134	0.132	0.003	0.134	0.121	0.015
Ak	Ak^1	0.097	0.067	1.24	1	0.032	188.82[c]	12	0.126	0.100	0.029	0.125	0.088	0.041
Hb	Hb^A	−0.007	−0.016	0.08	1	0.008	22.21[a]	12	−0.007	−0.019	0.011	−0.008	−0.031	0.022
Duffy	Fy^a	0.152	0.122	3.52	1	0.034	46.81[c]	12	0.180	0.142	0.045	0.179	0.114	0.072
Kidd	Ik^a	−0.204	−0.210	11.30[c]	1	0.005	12.67	12	−0.210	−0.218	0.007	−0.213	−0.259	0.037

[a] $p < 0.05$.
[b] $p < 0.01$.
[c] $p < 0.001$.

Table 4
Cockerham's allele- and locus-specific F_{IT}, F_{IS}, and F_{ST} estimates

Locus	Allele	Weighted[a]			Under $F_{IS} = 0$			Large N		
		F_{IT}	F_{IS}	F_{ST}	F_{ST}	χ^2	d.f.	F_{IT}	F_{IS}	F_{ST}
Hp	Hp^1	−0.006	−0.021	0.015	0.014	27.72[c]	12	−0.006	−0.032	0.025
Gc	Gc^{1F}	0.284	0.216	0.086	0.088	108.84[d]	12	0.284	0.206	0.098
	Gc^{1S}	0.242	0.193	0.061	0.063	80.96[d]	12	0.242	0.183	0.073
	Gc^2	0.190	0.163	0.032	0.034	49.15[d]	12	0.191	0.153	0.044
	Pooled	0.245	0.194	0.063	0.065			0.245	0.184	0.075
Tf	Tf^{C1}	−0.150	−0.165	0.014	0.012	23.71[b]	12	−0.149	−0.177	0.023
	Tf^{C2}	−0.165	−0.179	0.012	0.010	22.18[b]	12	−0.164	−0.190	0.022
	Tf^{C3}	−0.008	−0.021	0.012	0.012	24.37[b]	12	−0.008	−0.032	0.024
	Tf^{C12}	−0.002	−0.013	0.011	0.011	23.26[b]	12	−0.001	−0.025	0.023
	Tf^{D}	−0.003	0.001	−0.004	−0.004	7.62	12	−0.003	−0.010	0.007
	Pooled	−0.151	−0.166	0.013	0.011			−0.151	−0.177	0.022
aP	P^A	0.081	0.068	0.014	0.015	26.23[c]	12	0.081	0.056	0.027
	p^B	0.089	0.077	0.013	0.014	25.08[b]	12	0.089	0.065	0.026
	p^C	0.194	0.187	0.009	0.011	22.41[b]	12	0.194	0.175	0.023
	Pooled	0.088	0.076	0.013	0.014			0.088	0.064	0.026
PGM_1	PGM_1^1	0.025	0.025	−0.000	0.000	12.30	12	0.025	0.010	0.015
PGD	PGD^A	0.152	0.131	0.025	0.027	32.19[c]	12	0.152	0.115	0.042
EsD	EsD^1	0.145	0.144	0.002	0.003	15.38	12	0.146	0.132	0.015
Ak	Ak^1	0.100	0.067	0.036	0.036	51.88[d]	12	0.101	0.056	0.047
Hb	Hb^A	−0.007	−0.016	0.009	0.009	21.91[b]	12	−0.006	−0.027	0.020
Duffy	Fy^a	0.156	0.126	0.035	0.038	29.76[c]	12	0.157	0.101	0.062
Kidd	Ik^a	−0.203	−0.216	0.011	0.005	14.41	12	−0.202	−0.241	0.031

[a] The χ^2 for Cockerham's weighted estimates of F_{IS} and F_{ST} are exactly the same as those for Nei's weighted estimates (see Table 2).
[b] $p < 0.05$.
[c] $p < 0.01$.
[d] $p < 0.01$.

analyses are almost identical, there are some differences in the corresponding F_{IT} and F_{IS} estimates. Third, the large sample F_{ST} estimates of Nei and Cockerham are almost identical, even though these two methods yield somewhat different F_{IT} and F_{IS} values in large samples. Fourth, even when the estimate of allele- and locus-specific F_{IS} is significantly different from zero (tested by the χ^2 values), Cockerham's special case estimate of F_{ST} (equation (3.43) under $F_{ST} = 0$) is almost identical to that of his weighted analysis (Table 4). Fifth, while the large sample values of F_{IT} are very similar to those based on weighted analysis (in Nei's as well as Cockerham's approaches), the F_{ST} values do not behave similarly. Indeed, F_{ST} values are generally larger when large sample approximations are

Table 5
Locus-specific estimates of F_{IT}, F_{IS}, and F_{ST} by the multivariate technique of Long (1986)

Locus	Weighted estimators			Large sample estimators		
	F_{IT}	F_{IS}	F_{ST}	F_{IT}	F_{IS}	F_{ST}
Hp	−0.006	−0.021	0.015	−0.006	0.032	0.025
Gc	0.238	0.185	0.065	0.233	0.164	0.083
Tf	−0.156	−0.182	0.022	−0.167	−0.204	0.031
AP	0.081	0.063	0.019	0.073	0.045	0.029
PGM_1	0.025	0.025	0.000	0.025	0.010	0.015
PGD	0.152	0.131	0.025	0.152	0.115	0.042
EsD	0.145	0.144	0.002	0.146	0.132	0.015
AK	0.100	0.067	0.036	0.101	0.056	0.047
Hb	−0.007	−0.016	0.009	−0.006	−0.027	0.020
Duffy	0.156	0.126	0.035	0.157	0.101	0.062
Kidd	−0.203	−0.216	0.011	−0.202	−0.241	0.031

made. Because of equation (2.1), the F_{IS} should be under-estimated in large sample approximations (since F_{IT} does not change substantially). This is the case for every comparison of Cockerham's estimators, while there are some minor discrepancies in Nei's approach. These differences are due to changes in F_{IT} values in large sample vs. weighted analysis. Sixth, the multivariate estimators are the most deviant ones. There is no general trend of these estimators as compared to Nei's and Cockerham's estimators. This is also theoretically justifiable, since the wieghting scheme in the multivariate approach is quite different (equations (2.18a)–(2.20a)).

5.3. Comparison of the estimates pooled over loci

Table 6 presents the estimates and their standard errors pooled over all codominant loci. As mentioned before, pooling over loci can be done in two ways for every method of estimation:

(1) by taking the ratio of sums of locus-specific estimates, and

(2) by the technique of jackknifing.

Since each fixation index is described as a function of ratios of parameters (population allele frequencies and their inter-locus variances across subpopulations), Weir and Cockerham (1984) advocated the jackknifing procedure suggesting that this might reduce the bias of estimation and in turn make the standard errors more reliable (Miller, 1974; Efron, 1982). We, however, do not see any substantial change in the estimates as well as in their standard errors through jackknifing. In fact, there is a tendency for the jackknife estimators to have somewhat larger s.e.'s for each fixation index. This table also shows that while Cockerham's and Nei's estimators are virtually identical (weighted as well as large sample), the large sample approximations involve over-estimation of F_{ST} and under-estimation of F_{IS}, F_{IT} remaining very similar. The small difference of

Table 6
Estimates of F_{IT}, F_{IS}, and F_{ST} pooled over loci and their standard errors by the three different methods

	Ratio of sums			Jackknife		
	F_{IT}	F_{IS}	F_{ST}	F_{IT}	F_{IS}	F_{ST}
Nei's estimates						
Weighted	0.045 ± 0.060	0.025 ± 0.055	0.020 ± 0.009	0.045 ± 0.064	0.025 ± 0.058	0.020 ± 0.009
Unweighted	0.044 ± 0.060	0.023 ± 0.055	0.022 ± 0.009	0.044 ± 0.063	0.023 ± 0.058	0.022 ± 0.010
Large N	0.043 ± 0.060	0.005 ± 0.058	0.038 ± 0.008	0.043 ± 0.063	0.005 ± 0.061	0.038 ± 0.009
Cockerham's estimates						
Weighted	0.047 ± 0.060	0.025 ± 0.056	0.022 ± 0.009	0.047 ± 0.064	0.025 ± 0.059	0.022 ± 0.010
Under $F_{IS} = 0$			0.022 ± 0.009			0.022 ± 0.010
Large N	0.047 ± 0.060	0.011 ± 0.057	0.037 ± 0.009	0.048 ± 0.064	0.011 ± 0.060	0.037 ± 0.009
Long's estimates						
Weighted	0.048 ± 0.063	0.028 ± 0.059	0.021 ± 0.011	0.048 ± 0.064	0.029 ± 0.060	0.023 ± 0.011
Large N	0.046 ± 0.062	0.011 ± 0.058	0.036 ± 0.010	0.047 ± 0.063	0.010 ± 0.059	0.038 ± 0.009

Long's estimators as compared with others is mainly produced by the difference of the pooling algorithm in his procedure, as noted earlier. However, unless a survey has a large number of multi-allelic loci, this method is likely to produce an almost identical qualitative conclusion about the genetic structure of the population, as seen in this example.

5.4. Comparison of the estimates of F_{ST} from allele frequency data

As mentioned earlier, analysis of population structure is sometimes necessary from allele frequency data alone. This occurs when either the loci involves dominance relationships among their alleles, or the allele frequency data are collected from the literature for comparative studies. In such cases, the only estimable parameter is F_{ST}. It is shown earlier, that in Nei's gene diversity approach, the estimators (weighted or unweighted) remain the same even if allele frequencies are used in estimation instead of genotype data (see equations (3.8) and (3.8a)). Cockerham's estimator of F_{ST} takes the form of equation (3.43), whose multivariate extension (Long's approach) is obvious from equation (3.41). The variance–component approach (univariate or multivariate) of estimation of F_{ST} from allele frequency data is therefore mathematically equivalent to the estimation of the same parameter from genotype data with the additional assumption that $F_{IS} = 0$. In order to compare the empirical values of these estimators from allele frequency data, we computed Nei's weighted unbiased, Cockerham's, and Long's estimates of F_{ST} for all 17 loci studied by Bhasin et al. (1986). The allele frequencies used in these

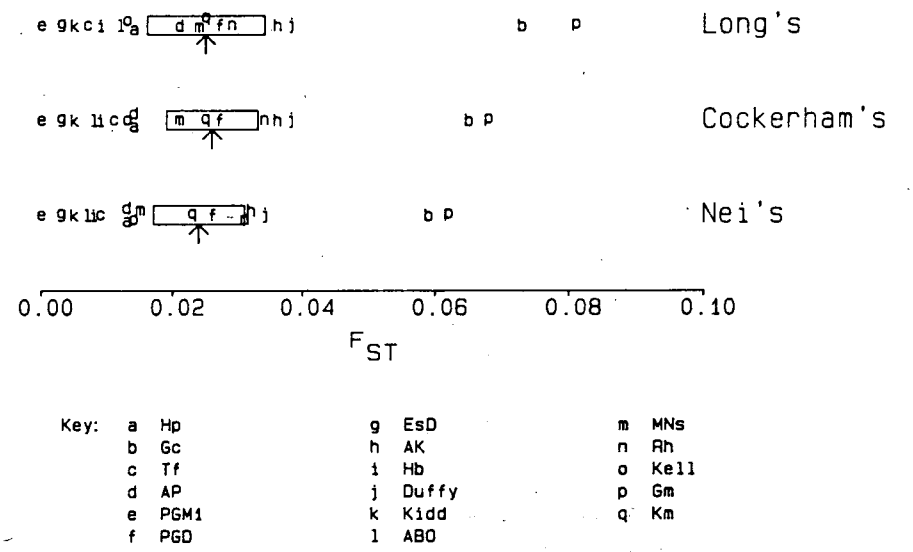


Fig. 2. A comparison of three locus-specific estimators of F_{ST} from allele frequency data on 17 loci in 13 subpopulations of Sikkim, India (Bhasin et al., 1986). The loci are indexed alphabatically (see Key). The averages over loci are indicated by arrow, and the boxes around these means represent ± s.e. range of the estimates (see text for the explanation of the estimators).

computations are the same as the ones reported in Bhasin et al. (1986). Figure 2 shows a diagramatical comparison of these locus-specific estimates, where the loci are indexed as a to q (see Key of Figure 2). The pooled estimates of F_{ST} over loci are indicated by an arrow, the box around which indicates the range with $\pm$ s.e.

It is clear that the estimates are again empirically very similar. Long's estimates are identical to the Cockerham estimates for all bi-allelic loci, although for multiallelic loci (Gc, Tf, AP, AB0, MNSs, Rh, and Gm) some discrepancies are noticable due to the different pooling (over alleles) algorithm employed in this method, as noted earlier. Nevertheless, the pooled estimates over loci are strikingly similar. Finally, we note that while the computation of the standard errors shown in this figure are based on equation (4.5), the jackknife estimates (equation (4.6)) of these standard errors are almost identical to the ones shown here. Hence, as in the case of genotype data, estimates of F_{ST} from allele frequency data also have similar empirical properties.

6. Discussion

As mentioned in Section 1, the purpose of this paper is to make a comprehensive comparative analysis of estimation of fixation indices by Nei's gene diversity approach with that of the variance component approach developed by Cockerham, or its multivariate extension. Keeping a distinction of parameters and sample statistics, throughout our presentation we have shown that these methods yield empirically very similar results. Even though these approaches have been described in a number of publications (see, e.g., Cockerham, 1969, 1973; Cockerham and Weir, 1986, 1988; Weir and Cockerham, 1984; Nei, 1973, 1977; Nei and Chesser, 1983; Chakraborty, 1974; Chakraborty and Leimar, 187; Long, 1986; Smouse and Long, 1988) and several other related statistics have been developed by others (Haldane, 1954; Li and Horvitz, 1953; Smith, 1970, 1977; Curie-Cohen, 1982; Robertson and Hill, 1984), to our knowledge, the analytical relationships between the two major approaches have not been studied explicitly before. In this discussion, first we re-iterate the new results presented here; and then we provide some arguments regarding the method we would suggest to practioners. Nevertheless, since during the conduct of this study, we developed a comprehensive computer-program for analyzing data on population structure, every estimator discussed in this paper can be computed by our computer algorithm. Interested readers can obtain a copy of the FORTRAN source codes of these programs by writing to the authors (compatible for IBM-AT type computers with a numerical co-procesor).

Our new results are as follows. First, the string of inequalities for the F_{IS} estimators in a single subpopulation shows that the expected differences among the estimators are of the order $1/2N$, N being the number of individuals sampled. While Nei's biased estimator of F_{IS} is always the smallest for any allele, Cockerham's variance–component estimator can be larger (when positive) or

smaller (when negative) than Nei's unbiased estimator. Second, even though Long (1986) and Smouse and Long (1988) generalized Cockerham's approach for a multivariate case (three or more alleles at a locus), they failed to note that their method yields F_{IS} estimators mathematically identical to Curie-Cohen's $\hat{f}_2$ (based on the ratio of observed and expected homozygote frequencies), in large samples. Third, for a single subpopulation, Nei's unbiased, biased (identical to $\hat{f}_1$ of Curie-Cohen, 1982–although not stated by him) and Cockerham's estimators of F_{IS} have closed form expressions of standard errors, for specific alleles as well as for the locus as a whole, which are also documented here for the first time. Much of the ground work for these derivations was, however, done by Curie-Cohen (1982) and Robertson and Hill (1984).

These new findings allow more rigorous comparative analyses of the different estimators, than the ones done before. Our empirical data analysis shows the closeness of the different estimators, which are based on somewhat different premises. There have been a number of misconceptions about the gene diversity approach, which should be clarified in this context. Note that the gene diversity approach does not need the correlation interpretation of the fixation indices. The total heterozygosity in subdivided populations is decomposed here on the basis of the number of extant subpopulations. No assumption of the replicative nature of subpopulations is needed. While Cockerham's linear model (of random effects) makes the assumption that the subpopulations studied are replicates from the universe of all subpopulations that exist within the total population, a situation that might apply to experimental populations, in the context of evolutionary significance, it is not clear if this assumption is realistic. In the specific example considered here, Sikkimese people are indeed subdivided into the present 13 subpopulations which during their history have assembled in this geographic region by following different migration routes (Bhasin et al., 1986). They are not replicates of each other, and indeed there may not be any further subpopulation among the people of Sikkim. If a statistical framework forms the basis of the variance–component analysis, the question is: should we treat the underlying linear model (Cockerham, 1969, 1973) as a random effects model in such a situation? Our answer to this question would be no as this subdivided structure represents a fixed-effect model. On the contrary, our exposition clearly indicates that Nei's gene diversity approach has a formal statistical basis, since all components of the decomposition of heterozygosity can be represented in terms of the underlying parameters, and they can be related with Wright's fixation indices without invoking their interpretation through correlations.

At this point it is worthwhile to note that for a single subpopulation the probabilistic interpretation of F_{IS} has been used by Haldane and Moshinsky (1939), Cotterman (1940), and Malécot (1948), where F_{IS} is interpreted as the probability that the two genes at a locus in an individual are identical. This probabilistic interpretation implicitly implies that the F_{IS} can take only non-negative values in the unit interval. Similar probabilistic interpretations of F_{IT} and F_{ST} are also used by Crow and Kimura (1970, pp. 105–106) to prove the Wright's identity (equation (2.1)). They, however, note that since F_{IT} and F_{IS} can be

negative, correlational interpretations of these fixation indices also yield the Wright's identity (Crow and Kimura, 1970, pp. 107–108). It is apparently implicit in their derivation that the subpopulations do not exchange migrants during the process of gene differentiation, so that the allele frequency variations across subpopulations do not depend upon the F_{IS} values within the subpopulations. In contrast, in Nei's formulation of gene diversity analysis the Wright's identity is established simply by the notion that F_{IS} and F_{IT} represent summary measures of deviations from the Hardy–Weinberg expectations in the subpopulations and in the total population, respectively, and F_{ST} represents the extent of genetic differentiation (standardized variance of allele frequencies across subpopulations). No assumption regarding migration and selection is needed in such derivation (Nei, 1973, 1977). The Wright's identity (equation (2.1)) simply becomes a mathematical consequence of the parametric definitions of F_{IT} (equation (2.5)), F_{IS} (equation (2.6)), and F_{ST} (equation (2.7)).

When the parameters are so defined, our equations (2.8), (2.9), and (2.10) suggest that all fixation indices have their natural bounds, namely F_{ST} lies between 0 and 1, while F_{IS} and F_{IT} can take positive as well as negative values, depending on H_S being smaller or larger than H_0 for F_{IS} (equation (2.8)) and H_T being smaller or larger than H_0 for F_{IT} (equation (2.9)). In such formulations no assumption is needed regarding the evolutionary mechanism that determines the process of genetic differentiation within and between subpopulations.

Since the variance–component approach can yield a negative value for the variance component b (equation (2.16)), in order to interprete the linear model (equation (16) of Cockerham, 1969) one must assume that $\sum_{k=1}^{r} w_i F_{IS_{ik}} p_{ik}(1 - p_{ik})$ must be positive. Cockerham (1969, 1973) recognized this feature of his model, and yet justified it on the ground that evolutionary factors that generally produce negative F_{IS} values are not usually strong enough to produce large negative F_{IS} (or F_{IT}) values. Our data analysis provides evidence contrary to this argument. We indeed found several negative estimates of F_{IS} (Tables 1 and 2). Even if their significance is discounted, because the normal deviates or the χ^2 statistics may not attain their large sample distribution in samples of the size analyzed here, it is unpleasant to deal with a linear model with negative variance components (not only the estimates, but also in parametric form).

Nei (1986) addressed some of these issues along with other evidences where the implicit assumptions of the variance component formulations are unrealistic for natural populations. He also noted that his original definition of F_{ST} (= D_{ST}/H_T, called G_{ST} by Nei, 1973) has one deficiency, since it is dependent on the number of subpopulations (s). He suggested one modification, defining $D'_{ST} = sD_{ST}/(s - 1)$, to take into account this deficiency (Nei, 1986). According to this suggestion, H'_T, the gene diversity in the total population is defined as $D'_{ST} + H_S$, yielding the three fixation indices $F_{IS} = H_0/H_S$ (unchanged from the previous definition), $F'_{IT} = H_0/H'_T$, and $F'_{ST} = D'_{ST}/H'_T$, for which the estimation technique presented here works with only minor modifications (see also Nei and Chesser, 1983). When s is large (say, 10 or more), these re-defined fixation indices change only slightly, and hence they are not computed in our application (since

for the present example $s = 13$). However, when s is small, it is preferable to calculate these modified indices with the above modifications. Also note that the re-defined F'_{ST} is identical to the parameter β defined by Cockerham and Weir (1988), not recognized by these authors. Therefore an estimator of β can also be obtained by estimating F'_{ST} in the gene diversity approach without the intraclass correlation interpretation. Nevertheless, we must reiterate the point that adjustment for the number of subpopulations does not necessarily help in comparing the coefficient of gene differentiation estimates in different data sets from different natural populations. An extrapolation of such estimates from one set of populations to another can be misleading, since their evolutionary histories are usually different. Cockerham's approach is more ideally suited for experimental populations, where the number of subpopulations represent the replicate of populations designed with a given experimental situation, and hence extrapolation from one experiment to another must need adjustments for variations in number of replicate subpopulations within each experiment.

Notwithstanding these philosophical differences, given the empirical similarity of the various estimators, a recommendation regarding the choice of estimators should be of interest to investigators who deal with real data. On the basis of statistical principles, unfortunately, there is no general recommendation. We claim this for several reasons. First, in every formulation, we have shown that consistent estimators can be derived. The study of large sample variances either by theoretical variances evaluated with intra-locus data, or by empirical evaluation of inter-locus variation shows that all estimators are subjected to similar sampling fluctuations. Second, even though with the aid of computer-algorithms the numerical task of computation can be left to computers, the choice is simply a matter of taste.

Since the gene diversity approach relates F_{ST} to the average genetic distances among subpopulations (Nei's minimum distance; Nei, 1972) a genetic distance interpretation of the coefficient of gene differentiation is also possible. Note that this interpretation does not assume, again, any evolutionary mechanism, and hence this interpretation should hold with or without mutation and selection. While Cockerham's F_{ST} parameters, and its multivariate extension have been shown to have a genetic distance interpretation as well (Reynolds et al., 1983; Long et al., 1987) in order for the measures of co-ancestry to be interpreted as genetic distances one must assume that genetic differentiation occurs without the aid of mutation and selection (Reynolds et al., 1983; Weir and Cockerham, 1984). Furthermore, in this latter paper they also assume that the same population size is maintained for all subpopulations and for all generations. While these assumptions are not needed in formulating Nei's genetic distances, thus far the evolutionary expectation and drift variances of genetic distances have been worked out under the neutral model of evolution without constant population size (Li and Nei, 1976; Nei and Chakravarti, 1977).

We advocate the use of the gene diversity approach for its simplicity and generality for natural populations. No loss of statistical rigor is attendant to this recommendation, as explicitly shown here—because we did not make use of any

evolutionary model in this presentation, and as a method of estimation, what we used can be called the method of moments in the terminology of statistical inference. This is the only appropriate estimation technique that yields analytically closed form estimators. We might add here that Curie-Cohen (1982) and Robertson and Hill (1984) investigated the properties of the maximum likelihood estimators of F_{IS} based on multinomial sampling of genotypes, which behave worse than Nei's biased estimator in most practical situations (Curie-Cohen, 1982).

Although we presented analytically closed form expressions of intra-locus variances of F_{IS} estimators, these are applicable only for single-locus data. Generally, large sample sizes are needed to apply these formulae, since the estimators are rather unstable (the drift variance is quite large; as shown by Li and Nei, 1976—for heterozygosities, Nei and Chakravarti, 1978—for $G_{ST} \approx F_{ST}$), and the power of detection of significant deviations of these indices is generally low (Brown, 1970; Ward and Sing, 1970; Chakraborty and Rao, 1972; Haber, 1980; Emigh, 1980). Evolutionary interpretation of the coefficients of gene differentiation or deviations from $F_{IS} = 0$, $F_{IT} = 0$ should be based on data on multiple loci. We have shown that multi-allelic and/or multiple-loci can be analyzed easily without the aid of Long's (1986) multivariate extensions. Indeed Nei's formulation of the decomposition of gene diversity is philosophically based on samples of genomes drawn from the population. He defined gene diversity as the complement of the probability that the two genomes are identical at each locus. Therefore, he computed gene diversity based on a sample of loci (polymorphic and monomorphic, see Nei, 1975, 1987). Even though the parameter F_{ST} (in Nei's terminology, G_{ST}) or its estimate does not change even if the monomorphic loci are excluded, the absolute value of H_T, H_S, and H_0 (averaged over all loci in a genome) changes. Even with a limited number of loci, we have shown that the variances of these quantities can be examined by studying their inter-locus variation, which yields inter-locus variances of the fixation indices as well. Since the inter-locus variance is the major component of the total sampling variance (Nei and Roychoudhury, 1974; Nei, 1978), jackknifing helps a little to provide a more reliable estimate of the extent of sampling variance. This finding is in disagreement with Mueller's study (Mueller, 1979) on genetic distance, but it is consistent with our own simulation results published before (Chakraborty, 1985). Weir and Cockerham (1984) claimed that jackknifing worked 'satisfactorily' in the two-population situation in their simulations (Reynolds et al., 1983), while we find that jackknifing does not add any particular advantage in terms of parameter estimates or their standard errors (Table 6).

Finally, we should return to the issue of hypothesis testing in the context of population structure data analysis. Considerable labor is needed to provide estimators adjusting for the effect of limited sample sizes. It is seen that when sample sizes are small (of the order of 100 or less individuals per locus per subpopulation, in the specific example given here), the use of large sample approximations yield over-estimates of F_{ST} and under-estimates of F_{IS} (F_{IT} remaining almost identical), irrespective of the method of analysis (Nei's vs. Cockerham's). Since such esti-

mates are invoked in evaluating standard errors or for computing test criteria, the question is: are these test criteria reliable, and can we justify the large sample properties of these test criteria? Our answer, although we cannot prove it analytically, is a probable no. We say so, for the reason that if the normal deviates are to be regarded as reliable, we must evaluate the standard errors accurately. We have seen that in some region of the parametric space, the standard errors can be drastically affected, even by a minute change in the parameter estimates. For the χ^2 tests, on the other hand, we must regard the variance components as independently distributed. This assumption, we might note, is also needed in Long's (1986) Wilk's Λ-test criteria. Nayak (1983) has shown that when the genotype data from several subpopulations are represented in the form of an analysis of variance of categorical data (Light and Margolin, 1971), the mean square errors of the different sources of variation are not independently distributed. The correlations between them can often be substantial. Furthermore, for every source of variation, the large sample distribution of the mean square errors is of the form of composite χ^2's, where the coefficients are also functions of unknown parameters. They cannot be simply equated to a χ^2 statistics as is done commonly invoking large sample theory of continuously distributed random variables. Therefore, we argue that the test statistics generally suggested for population structure analysis have much poorer statistical justifiability than the parameter estimates. Cockerham (1973) arrived at this general conclusion, although the sampling theory of weighted categorical data analysis was not available at that time.

7. Summary

A comprehensive comparative study of the various estimators of the fixation indices (F_{IT}, F_{IS}, and F_{ST}) shows that the properties of the estimators based on Nei's gene diversity and Cockerham's variance component analysis are very similar, in spite of their philosophical differences. In the analysis of genotypic data from a single population, a string of inequalities of the different estimators of F_{IS} is mathematically established, with regard to which the discrepancies in the sampling precision of these estimators can be reconciled. The analytical expression for the large sample variance of these estimators suggests that the parametric value of their sampling variance is identical. Empirical evaluation of the bias and standard errors of the three fixation indices from a genetic survey of 17 loci from 13 subpopulations of Sikkim, India suggests that for these ratio estimators the Jackknife method and Taylor's series approximation yield almost identical bias and standard error. These conclusions also hold for the estimation of F_{ST} from allele frequency data alone. A comprehensive computer program for obtaining all estimators has been developed, and is available from the authors upon request.

Acknowledgements

This paper is dedicated to the memory of Sewall Wright, the pioneer of population structure analysis, who passed away during the progress of this work. Dr Masatoshi Nei and Dr William J. Schull are to be acknowledged for their help in critically reviewing earlier versions of this review. Thanks are also due to Mr R. Schwartz for his help in computation and graphic works. This work was supported by grants from the National Institutes of Health and National Science Foundation. HDH was supported by the German research fellowship program of DAAD during the conduct of this work.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd edition). John Wiley, New York.

Bhasin, M. K., Walter, H., Chahal, S. M. S., Bhardwaj, V., Sudhakar, K., Danker-Hopfe, H., Dannewitz, A., Singh, I. P., Bhasin, V., Shil, A. P., Sharma, M. B. and Wadhavan, D. (1986). Biology of the people of Sikkim, India. 1. Studies on the variability of genetic markers. *Z. Morph. Anthrop.* **77**, 49–86.

Brown, A. H. D. (1970). The estimation of Wright's fixation index from genotype frequencies. *Genetica* **41**, 399–406.

Chakraborty, R. (1974). A note on Nei's measure of gene diversity in a substructured population. *Humangenetik* **21**, 85–88.

Chakraborty, R. (1985). Genetic distance and gene diversity: Some statistical considerations. In: *Multivariate Analysis – VI*, P. R. Krishnaiah (ed.). Elsevier, Amsterdam, 77–96.

Chakraborty, R., Chakravarti, A. and Malhotra, K. C. (1977). Variation of allele frequencies among caste groups of the Dhangars of Maharashtra, India: An analysis with Wright's *F*-statistics. *Ann. Hum. Biol.* **4**, 275–280.

Chakraborty, R. and Nei, M. (1977). Bottleneck effect with stepwise mutation model of electrophoretically detectable alleles. *Evolution* **31**, 347–356.

Chakraborty, R. and Leimar, O. (1987). Genetic variation within a subdivided population. In: *Population Genetics and Fishery Management*, N. Ryman and F. Utter (eds.). Sea Grant Program, University of Washington Press, Seattle, WA, 89–120.

Chakraborty, R. and Rao, D. C. (1972). On the detection of *F* from ABO blood group data. *Am. J. Hum. Genet.* **24**, 352–353.

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution* **23**, 72–84.

Cockerham, C. C. (1973). Variance of gene frequencies. *Evolution* **27**, 679–700.

Cockerham, C. C. and Weir, B. S. (1986). Estimation of inbreeding parameters in stratified populations. *Ann. Hum. Genet.* **50**, 271–281.

Cockerham, C. C. and Weir, B. S. (1987). Correlations, descent measures: Drift with migration and mutation. *Proc. Natl. Acad. Sci. USA* **84**, 8512–8514.

Cotterman, C. W. (1940). A calculas for statistic-genetics. Ph.D. dissertation, Ohio University, Columbus, OH.

Crow, J. F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row, New York.

Curie-Cohen, M. (1982). Estimates of inbreeding in a natural population: A comparison of sampling properties. *Genetics* **100**, 339–358.

Efron, B. (1982). *The Jackknife, the Bootstrap and other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia.

Haber, M. (1980). Detection of inbreeding effects by the chisquare test on genotypic and phenotypic frequencies. *Am. J. Hum. Genet.* **32**, 754–760.

Haldane, J. B. S. (1954). An exact test for randomness of mating. *J. Genet.* **52**, 631–635.
Haldane, J. B. S. and Moshinsky, P. (1939). Inbreeding in Mendelian populations with special reference to human cousin marriage. *Ann. Eugen.* **9**, 321–340.
Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistica. Vol. 1* (4th edition). MacMillan, New York.
Kirby, G. C. (1975). Heterozygote frequencies in small subpopulations. *Theoretical Population Biology* **8**, 31–48.
Li, C. C. (1955). *Population Genetics*. University of Chicago Press, Chicago, IL.
Li, C. C. and Horvitz, D. G. (1953). Some methods of estimating the inbreeding coefficient. *Am. J. Hum. Genet.* **5**, 107–117.
Li, W.-H. and Nei, M. (1975). Drift variances of heterozygosity and genetic distance in transient states. *Genet. Res.* **25**, 229–248.
Light, R. J. and Margolin, B. H. (1971). An analysis of variance for categorical data. *J. Am. Stat. Assoc.* **66**, 534–544.
Long, J. C. (1986). The allelic correlation structure of Gainj- and Kaam-speaking people. I. The estimation and interpretation of Wright's *F*-statistics. *Genetics* **112**, 629–647.
Malécot, G. (1948). *Les Mathématiques de l'Hérédité*. Masson et Cie, Paris.
Miller, R. G. (1974). The jackknife – A review. *Biometrika* **61**, 1–15.
Mueller, L. D. (1979). A comparison of two methods for making statistical inferences on Nei's measure of genetic distance. *Biometrics* **35**, 757–763.
Nayak, T. K. (1983). Applications of entropy functions in measurement and analysis of diversity. Ph.D. Dissertation, University of Pittsburgh, Pittsburgh, PA.
Nei, M. (1972). Genetic distance between populations. *Am. Nat.* **106**, 283–292.
Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc.* Natl. Acad. Sci. USA **70**, 3321–3323.
Nei, M. (1975). *Molecular Population Genetics and Evolution*: North-Holland, Amsterdam.
Nei, M. (1977). *F*-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.* **41**, 225–233.
Nei, M. (1978). Estimation of average hterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583–590.
Nei, M. (1986). Definition and estimation of fixation indices. *Evolution* **40**, 643–645.
Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.
Nei, M. and Chakravarti, A. (1977). Drift variances of F_{ST} and G_{ST} statistics obtained from a finite number of isolated populations. *Theor. Pop. Biol.* **11**, 307–325.
Nei, M. and Chesser, R. K. (1983). Estimation of fixation indices and gene diversities. *Ann. Hum. Genet.* **47**, 253–259.
Nei, M., Maruyama, T. and Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution* **29**, 1–10.
Nei, M. and Roychoudhury, A. K. (1974). Sampling variance of hterozygosity and genetic distance. *Genetics* **76**, 379–390.
Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. Wiley, New York.
Rao, C. R. (1982). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhya A* **44**, 1–21.
Rao, C. R., Rao, D. C. and Chakraborty, R. (1973). The generalized Wright's model. In: *Genetic Structure of Populations*, N. E. Morton, (ed.). University of Hawaii Press, Honolulu, 55–59.
Rao, D. C. and Chakraborty, R. (1974). The generalized Wright's model and population structure. *Am. J. Hum. Genet.* **26**, 444–453.
Reynolds, J., Weir, B. S. and Cockerham, C. C. (1983). Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* **105**, 767–779.
Robertson, A. and Hill, W. G. (1984). Deviation from Hardy–Weinberg proportions: Sampling variances and use in estimation of inbreeding coefficients. *Genetics* **107**, 703–718.
Slatkin, M. and Barton, N. H. (1989). A comparison of three methods for estimating average level of gene flow. *Evolution* **43**, 1349–1368.
Smith, C. A. B. (1970). A note on testing the Hardy–Weinberg law. *Ann. Hum. Genet.* **33**, 377–383.

mates are invoked in evaluating standard errors or for computing test criteria, the question is: are these test criteria reliable, and can we justify the large sample properties of these test criteria? Our answer, although we cannot prove it analytically, is a probable no. We say so, for the reason that if the normal deviates are to be regarded as reliable, we must evaluate the standard errors accurately. We have seen that in some region of the parametric space, the standard errors can be drastically affected, even by a minute change in the parameter estimates. For the χ^2 tests, on the other hand, we must regard the variance components as independently distributed. This assumption, we might note, is also needed in Long's (1986) Wilk's Λ-test criteria. Nayak (1983) has shown that when the genotype data from several subpopulations are represented in the form of an analysis of variance of categorical data (Light and Margolin, 1971), the mean square errors of the different sources of variation are not independently distributed. The correlations between them can often be substantial. Furthermore, for every source of variation, the large sample distribution of the mean square errors is of the form of composite χ^2's, where the coefficients are also functions of unknown parameters. They cannot be simply equated to a χ^2 statistics as is done commonly invoking large sample theory of continuously distributed random variables. Therefore, we argue that the test statistics generally suggested for population structure analysis have much poorer statistical justifiability than the parameter estimates. Cockerham (1973) arrived at this general conclusion, although the sampling theory of weighted categorical data analysis was not available at that time.

7. Summary

A comprehensive comparative study of the various estimators of the fixation indices (F_{IT}, F_{IS}, and F_{ST}) shows that the properties of the estimators based on Nei's gene diversity and Cockerham's variance component analysis are very similar, in spite of their philosophical differences. In the analysis of genotypic data from a single population, a string of inequalities of the different estimators of F_{IS} is mathematically established, with regard to which the discrepancies in the sampling precision of these estimators can be reconciled. The analytical expression for the large sample variance of these estimators suggests that the parametric value of their sampling variance is identical. Empirical evaluation of the bias and standard errors of the three fixation indices from a genetic survey of 17 loci from 13 subpopulations of Sikkim, India suggests that for these ratio estimators the Jackknife method and Taylor's series approximation yield almost identical bias and standard error. These conclusions also hold for the estimation of F_{ST} from allele frequency data alone. A comprehensive computer program for obtaining all estimators has been developed, and is available from the authors upon request.

Smith, C. A. B. (1977). A note on genetic distance. *Ann. Hum. Genet.* **40**, 463–479.

Smouse, P. E. and Long, J. C. (1988). A comparative *F*-statistics analysis of the genetic structure of human populations from the Lowland South America and Highland New Guinea. In: *Quantitative Genetics*, B. S. Weir, E. J. Eison, M. M. Goodman and G. Namkoong (eds.). Sinaur Association Inc., Sunderland, 32–46.

Van Den Bussche, R. A., Hamilton, M. J. and Chesser, R. K. (1986). Problems of estimating gene diversity among populations. *The Texas Journal of Science* **38**, 281–287.

Weir B. S. and Cockerham, C. C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.

Workman, P. L. and Niswander, J. D. (1970). Population studies on Southwestern Indian tribes. II. Local genetic differentiation in the Papago. *Am. J. Hum. Genet.* **22**, 24–49.

Wright, S. (1943). Isolation by distance. *Genetics* **28**, 114–138.

Wright, S. (1951). The genetical structure of populations. *Ann. Eugenics* **15**, 323–354.

Wright, S. (1965). The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution* **19**, 395–420.

C. R. Rao and R. Chakraborty, eds., *Handbook of Statistics*, *Vol.* 8

7

Analysis of Population Structure: A Comparative Study of Different Estimators of Wright's Fixation Indices

Ranajit Chakraborty and Heidi Danker-Hopfe

1. Introduction

Computations of Wright's fixation indices (F_{IT}, F_{ST}, and F_{IS}) are pivotal for studying the genetic differentiation of populations. It is well known that these indices can be conceptually defined in terms of correlations between uniting gametes (Wright, 1943, 1951); as functions of heterozygosities and their Hardy–Weinberg expectations (Nei, 1973, 1977), or as functions of variance components from a nested analysis of variance (Cockerham, 1969, 1973; Weir and Cockerham, 1984; Long, 1986). Nei (1977) and Nei and Chesser (1983) considered the question of estimating the fixation indices through a decomposition of gene diversity in the total population, while Cockerham (1969, 1973) and Weir and Cockerham (1984) provided estimation procedures by a variance component analysis. Long (1986) extended the variance component approach of estimation to the case of multiple (greater than two) allelic loci, which gives numerically different results from the Weir–Cockerham estimates (see equation (10) of Weir and Cockerham, 1984 vs. equations (9a), (10a) and (11a) of Long, 1986). Although there are several studies drawing comparisons of these different estimates in simulated data (Van Den Bussche et al., 1986; Chakraborty and Leimar, 1987; Slatkin and Barton, 1989), it is not generally known how these different estimates differ in real data in practice. Furthermore, there is no comprehensive computer algorithm which computes all of these estimates simultaneously.

The purpose of this chapter is twofold:

(1) to review the different conceptualizations of Wright's fixation indices using an uniform set of notations, and to examine the question of estimation of parameters and hypothesis testing in the context of an analysis of categorical data; and

(2) to document a computational algorithm for deriving the different estimators (with their standard errors, and test criteria) developed here (called WRIGHT, with three components: NEI, CLARK, and LONG) that can be used for any given data for population structure analysis.

In doing so, we provide the description of the parameters and express the estimators as functions of the observed data statistics, since there is a misconception that some of the formulations are in terms of the data statistics, and not the underlying parameters. The estimation equations are given encompassing the situations where the genotype or the allele frequencies are available. Note that when allele frequencies are used as observed data characteristics (which is usually the case for loci involving dominance relationships among alleles at a locus or in analysis of data collected from the literature), because of the lack of information on observed heterozygosities within populations, the two fixation indices F_{IT} and F_{IS} cannot be estimated, hence the only parameter that needs estimation is F_{ST}.

Empirical comparisons of these different estimators are provided with a gene diversity analysis of the populations of Sikkim, India published by Bhasin et al. (1986). Finally, we discuss the relative merits of these different estimators in terms of their complexity of computation, and generality in various practical situations. While there are several recent reviews of the difficulties of the estimators of F_{IS}, F_{ST}, and F_{IT} in the literature (see, e.g., Curie-Cohen, 1982; Robertson and Hill, 1984; Weir and Cockerham, 1984), they do not encompass all of the different estimators as fully as presented here. Consequently, these reviews do not explicitly demonstrate why the different methods of estimation produce numerically different results, or how different they can be in practice. Therefore, this review, together with the documentation of a single computer program (available from the authors upon request) should serve as an up-to-date description of the applicability of the estimators of Wright's fixation indices to the analysis of any combination of immunological (blood groups, immunoglobulin-Gm, HLA), biochemical (red-cell isozymes and serum proteins), and DNA polymorphism (Restriction Fragment Length Polymorphism, RFLP's) data in the study of the genetic structure of a subdivided population.

2. Parameters of population structure

2.1. Wright's fixation indices and Nei's gene diversity

When F_{IT} and F_{IS} are defined as correlations between two uniting gametes to produce the individuals relative to the total population and relative to the subpopulations, respectively, the correlation between two gametes drawn at random for each subpopulation (F_{ST}) is known to satisfy the identity (Wright, 1943, 1951)

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST}). \tag{2.1}$$

Consider a population which is subdivided into s subpopulations in each of which Hardy–Weinberg equilibrium (HWE) does not necessarily hold (i.e., $F_{IS} \neq 0$). For a locus with r alleles (denoted as $A_1, A_2, \ldots, A_r$), deviation from HWE can be fully specified by $\frac{1}{2}r(r-1)$ F_{IS} parameters (Rao et al. 1973).

However, if only the homozygotes are considered, r F_{IS} parameters are enough to specify deviations from HWE.

In the latter event, the frequency of the homozygotes for the k-th allele (A_kA_k) in the i-th subpopulation may be written as

$$P_{ik} = p_{ik}^2 + F_{ISik}p_{ik}(1 - p_{ik}), \tag{2.2}$$

where p_{ik} is the frequency of the A_k allele in the i-th subpopulation for $i = 1, 2, \ldots, s$; $k = 1, 2, \ldots, r$. Therefore the allele-specific F_{IS} in the i-th subpopulation can be written as

$$F_{ISik} = (P_{ik} - p_{ik}^2)/[p_{ik}(1 - p_{ik})]. \tag{2.3}$$

The deviation from HWE in the total population, with reference to the same homozygote frequency, can be parameterized in the same fashion, giving

$$P_{\cdot k} = \bar{p}_{\cdot k}^2 + F_{ITk}\bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k}), \tag{2.4}$$

where

$P_{\cdot k} = \sum_{i=1}^{s} w_i P_{ik}$ is the proportion of A_kA_k genotypes in the total population,
$\bar{p}_{\cdot k} = \sum_{i=1}^{s} w_i p_{ik}$ is the frequency of the A_k allele in the total population, and
w_i = weight of the i-th subpopulation relative to the total population size, which yields

$$F_{ITk} = (P_{\cdot k} - \bar{p}_{\cdot k}^2)/[\bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k})]. \tag{2.5}$$

With these notations the average F_{IS} (within population deviation from HWE) over all subpopulations for the k-th allele, takes the form

$$\begin{aligned} F_{ISk} &= \sum_{i=1}^{s} w_i(P_{ik} - p_{ik}^2) \Big/ \left[\sum_{i=1}^{s} w_i p_{ik}(1 - p_{ik})\right] \\ &= (P_{\cdot k} - \overline{p_{\cdot k}^2})/(\bar{p}_{\cdot k} - \overline{p_{\cdot k}^2}), \end{aligned} \tag{2.6}$$

where

$$\overline{p_{\cdot k}^2} = \sum_{i=1}^{s} w_i p_{ik}^2.$$

From equation (2.1), we therefore have

$$F_{STk} = (\overline{p_{\cdot k}^2} - \bar{p}_{\cdot k}^2)/(\bar{p}_{\cdot k} - \bar{p}_{\cdot k}^2). \tag{2.7}$$

Note that, in this formulation, the definitions of allele-specific F_{IS}, F_{IT}, and F_{ST} values are indeed parameters defined in terms of allele frequencies in the population.

Furthermore, to obtain the locus specific values of these fixation indices, we can

sum the numerators and denominators over all alleles at a locus ($k = 1, 2, \ldots, r$) to get the following formulae. From equation (2.6), we have

$$\begin{aligned} F_{IS} &= \left[\sum_{k=1}^{r}\sum_{i=1}^{s} w_i(P_{ik} - p_{ik}^2)\right] \Big/ \left[\sum_{k=1}^{r}\sum_{i=1}^{s} w_i p_{ik}(1 - p_{ik})\right] \\ &= \left[\sum_{i=1}^{s} w_i \sum_{k=1}^{r} p_{ik}(1 - p_{ik}) - \sum_{i=1}^{s} w_i\left(1 - \sum_{k=1}^{r} P_{ik}\right)\right] \\ &\quad \times \left[\sum_{i=1}^{s} w_i \sum_{k=1}^{r} p_{ik}(1 - p_{ik})\right]^{-1} \\ &= (H_S - H_0)/H_S\,, \end{aligned} \tag{2.8}$$

where

$$H_S = \sum_{i=1}^{s} w_i \sum_{k=1}^{r} p_{ik}(1 - p_{ik}) = \sum_{i=1}^{s} w_i H_{Si}$$

is the average within population heterozygosity expected under HWE, and

$$H_0 = \sum_{i=1}^{s} w_i\left(1 - \sum_{k=1}^{r} P_{ik}\right) = 1 - \sum_{i=1}^{s}\sum_{k=1}^{r} w_i P_{ik}$$

is the actual proportion of heterozygotes in the total population. Similarly, from equation (2.5),

$$\begin{aligned} F_{IT} &= \left[\sum_{k=1}^{r} (P_{\cdot k} - \bar{p}^2_{\cdot k})\right] \Big/ \left[\sum_{k=1}^{r} \bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k})\right] \\ &= \left[\sum_{k=1}^{r} \bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k}) - \left(1 - \sum_{k=1}^{r} P_{\cdot k}\right)\right] \Big/ \left[\sum_{k=1}^{r} \bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k})\right] \\ &= (H_T - H_0)/H_T\,, \end{aligned} \tag{2.9}$$

where

$$H_T = \sum_{k=1}^{r} \bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k}) = 1 - \sum_{k=1}^{r} \bar{p}^2_{\cdot k}$$

is the heterozygosity in the total population (expected under HWE).

Lastly, from equation (2.7), we have

$$F_{ST} = (H_T - H_S)/H_T\,. \tag{2.10}$$

Therefore, estimation of the fixation indices are equivalent to estimation of the parameters H_T, H_S, and H_0 for a locus (Nei, 1973, 1977). Note that these parametric relationships also hold when the fixation indices are defined by pooling over several loci. In this case, H_S, H_T, and H_0 are the respective heterozygosities

averaged over all loci. The criticism that the relationship between fixation indices with heterozygosities is true for data statistics (and not for parameters) is not valid. Weir and Cockerham's (1984) comments are perhaps due to the misconception that under the mutation-drift model, the expectation of F_{ST} in a population with a finite number of subpopulations (expectation under the evolutionary process) is a function of the number of subpopulations as well. Two comments are worth noting at this point.

First, the above parameterization does not depend upon the evolutionary model of genetic differentiation among subpopulations, and hence the relationships (2.8)–(2.10) hold for any general mating system, irrespective of the selective differentials that may exist among the alleles. Second, even though Nei (1977) defined the pooled F_{IS} in terms of a weighted average of the subpopulation-specific F_{IS_i} values (equation (4) of Nei, 1977), with Wright's (1965) and Kirby's (1975) weight functions, such weights are not needed if we first define F_{IS}, F_{IT}, and F_{ST} as allele-specific parameters and obtain the locus-specific parameters by summing numerators and denominators over all alleles at a locus. It is also clear from equations (2.8)–(2.10) that while estimation of F_{IS} and F_{IT} would require sampling of genotypes from all subpopulations (as they are functions of the actual proportion of heterozygotes in the subpopulations, H_0), F_{ST} can be estimated with allele frequency data alone without making any assumption regarding F_{IS}. Furthermore, estimates of F_{ST} from genotype or allele frequency data should be identical, as long as the allele frequencies are obtained by the gene counting method. These issues will be detailed in the estimation section to follow.

2.2. Fixation indices and Cockerham's variance component representation

Cockerham (1969, 1973) redefined the fixation indices in terms of intra-class correlation derived from an analysis of variance of allele frequencies. In this formulation, indicator variables are defined for both alleles of a random individual sampled, which are in turn expressed as a linear model of additive effects of between-subpopulations (a), between-individuals within a subpopulation (b), and within-individual (c or w) variations. Following the classical analysis of variance model of random effects, where the subpopulations are treated as replicates of each other (Weir and Cockerham, 1984), Cockerham (1969, 1973) showed that the component of variance ascribed to the above factors (a, b and c) yield a parametric relationship with the fixation indices. In particular, Cockerham's results for a specific allele can be written in terms of our notation as

$$F_{ITk} = (a_k + b_k)/(a_k + b_k + c_k)\,, \tag{2.11}$$

$$F_{ISk} = b_k/(b_k + c_k)\,, \tag{2.12}$$

$$F_{STk} = a_k/(a_k + b_k + c_k)\,, \tag{2.13}$$

where a_k, b_k, and c_k are the variance components associated with the above factors, in which the genotype frequencies in the population are tabulated with

regard to a specific allele, A_k (and thus only the frequencies of the three genotypes A_kA_k, $A_k\bar{A}_k$, and $\bar{A}_k\bar{A}_k$ enter into the analysis, $\bar{A}_k$ being a combination of all alleles of type other than A_k).

Before reviewing the estimation equations for these variance components, it is worthwhile to examine how these variance component parameters translate into the gene frequency parameters in a subdivided population.

It is easy to note that

$$a_k + b_k + c_k = \bar{p}_{\cdot k}(1 - \bar{p}_{\cdot k}), \tag{2.14}$$

where $\bar{p}_{\cdot k} = \sum_{i=1}^{s} w_i p_{ik}$, as defined in equation (2.4). Invoking equation (2.13) into equation (2.2), we also have

$$a_k = \overline{p_{\cdot k}^2} - \bar{p}_{\cdot k}^2 = \sum_{i=1}^{s} w_i (p_{ik} - \bar{p}_{\cdot k})^2, \tag{2.15}$$

and similarly, from equations (2.4) and (2.12), we have

$$\begin{aligned} b_k &= P_{\cdot k} - \overline{p_{\cdot k}^2} = \sum_{i=1}^{s} w_i (P_{ik} - p_{ik}^2) \\ &= \sum_{i=1}^{s} w_i F_{\mathrm{IS}ik} p_{ik}(1 - p_{ik}), \end{aligned} \tag{2.16}$$

since $P_{ik} = p_{ik}^2 + F_{\mathrm{IS}ik} p_{ik}(1 - p_{ik})$, according to our equation (2.2).

Putting equations (2.15) and (2.16) in equation (2.14), we get

$$c_k = \bar{p}_{\cdot k} - P_{\cdot k} = \sum_{i=1}^{s} w_i (p_{ik} - P_{ik}). \tag{2.17}$$

Since $-p_{ik}/(1 - p_{ik}) \leqslant F_{\mathrm{IS}ik} \leqslant 1$ for all $i = 1, 2, \ldots, s$ and all k, it is easy to see that $c_k \geqslant 0$. Because of equation (2.15), it is also ensured that $a_k \geqslant 0$. However, there is no guarantee that b_k is non-negative. It is, therefore, peculiar that even a parametric value of the variance component due to between individual variation can assume negative values in this formulation. Cockerham (1969) acknowledged this feature, and ascribed this to either a mating system where mates are less related than the average within a subpopulation, or to certain types of selection (Cockerham, 1969, p. 74). Since this arises for $F_{\mathrm{ST}} > F_{\mathrm{IT}}$ ($\theta > F$, in Cockerham's 1969 notation), this occurs whenever F_{IS} takes negative values (see equation (2.1)).

The above translation of parameters reveals that the negative value of b may not necessarily arise only in estimation; it is an inherent feature of the proposed linear model itself (Cockerham, 1969, 1973). It is particularly uncomfortable, since the linear model is not supposed to produce negative variance components.

In Cockerham's formulation, the locus-specific parameters are defined by

summing a_k, b_k, and c_k values over all alleles, and expressing the fixation indices as respective ratios of sums, analogous to equations (2.11)–(2.13). The same pooling algorithm is suggested for definition of parameters pooled over all loci studied (see Weir and Cockerham, 1984, equation (10)).

2.3. Long's extension of Cockerham's model

Long (1986) and Smouse and Long (1988) provided a multivariate extension of the Cockerham model, where a pair of $(r-1)$-dimensional indicator vectors is defined for a r-allelic genotypic system. This yields a multivariate decomposition of the total dispersion matrix; Σ_a, Σ_b, Σ_c in the analogy of a, b, and c of a bi-allelic locus. With $\Sigma = \Sigma_a + \Sigma_b + \Sigma_c$, the locus-specific fixation indices take the form of

$$F_{IT} = (r-1)^{-1} \operatorname{tr}[\Sigma^{-1/2}(\Sigma_a + \Sigma_b)\Sigma^{-1/2}], \tag{2.18}$$

$$F_{ST} = (r-1)^{-1} \operatorname{tr}[\Sigma^{-1/2}\Sigma_a\Sigma^{-1/2}], \tag{2.19}$$

$$F_{IS} = (r-1)^{-1} \operatorname{tr}[(\Sigma_b + \Sigma_c)^{-1/2}\Sigma_b(\Sigma_b + \Sigma_c)^{-1/2}], \tag{2.20}$$

where tr denotes the trace of a matrix. In this formulation, again, while Σ_a and Σ_c are positive semi-definite matrices, the parametric form of Σ_b can be negative-definite, introducing peculiarities in the interpreting of the decomposition of dispersion matrices.

Note that for $r = 2$, equations (2.18)–(2.20) are mathematically identical to Cockerham's definition of parameters; but for $r > 2$, since equations (2.18)–(2.20) involve covariances of allele or genotype frequencies within and between subpopulations (off-diagonal elements of the Σ-matrices), the locus-specific fixation indices, according to Long's approach, are parametrically different from Weir–Cockerham's parameters. A multi-locus extension of Long's formulation is also available, where the respective Σ matrices for a group of loci are written as block-diagonal locus-specific Σ matrices (see Long, 1986, equation (8)).

In summary, the above parameterization of the genetic structure of a subdivided population indicates that Wright's fixation indices can be expressed in terms of the actual proportion of heterozygotes (H_0) and its expectation (under HWE) in the total population (H_T) and within subpopulations (H_S), without invoking any specific model of the mating system or gene differentiation between or within subpopulations. This mathematical equivalence is shown in the form of parameters, and they are consistent with Wright's identity (equation (2.1)), while the variance–component parameterization is more complex in nature, and could yield possible inconsistencies (e.g., $b < 0$, whenever F_{IS} is negative) for certain evolutionary factors (selection) or social structure of subdivision. Having defined the parameters, let us now turn to estimation and hypothesis testing issues.

3. Estimation of fixation indices

The above discussion indicates that while the heterozygosities or variance components are quadratic functions of allele and/or genotype frequencies within each subpopulation, and their weighted (by relative subpopulation size) averages, the fixation indices are ratios of functions of parameters. While estimation of a ratio of parametric functions is an unpleasant statistical problem for categorical data, to the extent that we may approximate the expectation of a ratio by the ratio of expectations, reasonable estimators of fixation indices may be obtained. Weir and Cockerham (1984) called such estimators (ratio of unbiased estimators of a numerator and a denominator) 'unbiased', while in the strict statistical sense, such estimators are at best consistent (i.e., approach the true value in terms of probability in large samples). A further problem arises, because of the categorical nature of the observations (allele frequencies, or genotype frequencies). The properties of ratio estimators are generally studied in the statistical literature for continuous traits which have Gaussian probability distributions. Even those who are concerned with distinctions between parameters and statistics have been rather cavalier about this aspect of the problem.

In this section we consider some estimators and present estimating equations in terms of the observed frequencies, which in turn indicate how much bias might arise in using these estimating equations. We might also mention that the definition of sample size has been quite elusive in the literature; because it is not always explicit whether it refers to the number of genes sampled, or that of individuals (see, e.g., Weir and Cockerham, 1984).

3.1. Estimation from genotype data

Let us first consider the case where all genotypes are recognizable, so that unequivocally all different alleles can be counted in a sample. As noted before, the F_{IT}, F_{IS}, and F_{ST} parameters depend on the sizes of subpopulations, relative to the total population size. In practice these are unknown, and furthermore, subpopulation sizes generally fluctuate over an evolutionary time period. The temporal change in population sizes has a substantial effect on the coefficients of gene-differentiation as well as heterozygosity (see, e.g., Nei et al., 1975; Chakraborty and Nei, 1977). Therefore, we shall assume all subpopulations to have equal size. This assumption is also explicitly made in Weir and Cockerham (1984, p. 1359). This, however, does not imply that the numbers of individuals sampled from the subpopulations are all equal.

3.2. Estimators of fixation indices by Nei's approach

As before consider a r-allelic locus, and define N_{ikl} to be the number of individuals of genotype A_kA_l in the i-th subpopulation ($k = 1, 2, \ldots, r$; $l = k, \ldots, r$; $i = 1, 2, \ldots, s$). Let N_i be the total number of individuals (sample size) sampled from the i-th subpopulation. The total sample size (number of individuals)

sampled from the entire subdivided population is

$$N = \sum_{i=1}^{s} N_i, \quad \text{where } N_i = \sum_{k \geqslant l=1}^{r} \sum N_{ikl}.$$

When $(N_{ikl};\ k = 1, 2, \ldots, r;\ l = k, \ldots, r)$ is a genotype-specific categorized subdivision of a random sample of N_i individuals from the i-th subpopulation, it is easy to note that

$$X_{ik} = N_{ikk}/N_i \tag{3.1}$$

and

$$x_{ik} = \left(2N_{ikk} + \sum_{l>k=1}^{r} N_{ikl}\right) \Big/ 2N_i, \tag{3.2}$$

are unbiased estimates of P_{ik} and p_{ik}, the proportion of A_kA_k homozygotes, and the allele frequency of A_k in the i-th subpopulation.

An unbiased estimator for p_{ik}^2 can be obtained as

$$\hat{p_{ik}^2} = x_{ik}(2N_i x_{ik} - 1)/(2N_i - 1), \tag{3.3}$$

which in turn, provides an unbiased estimator of $p_{ik}(1 - p_{ik})$, namely,

$$\frac{2N_i}{2N_i - 1}\, x_{ik}(1 - x_{ik}). \tag{3.4}$$

Note that if $x_{ik}(1 - x_{ik})$ is used as an estimator for $p_{ik}(1 - p_{ik})$, the extent of bias is

$$\begin{aligned} b &= [(2N_i - 1)/(2N_i) - 1]\, p_{ik}(1 - p_{ik}) \\ &= -p_{ik}(1 - p_{ik})/(2N_i), \end{aligned} \tag{3.4a}$$

i.e., $x_{ik}(1 - x_{ik})$ is an under-estimator of $p_{ik}(1 - p_{ik})$, with proportional bias being $1/2N_i$.

With equations (3.2)–(3.4), the estimator of $F_{\mathrm{IS}ik}$ (given by equation (2.3) is

$$\hat{F}_{\mathrm{IS}ik} = \frac{X_{ik} - x_{ik}(2N_i x_{ik} - 1)/(2N_i - 1)}{2N_i x_{ik}(1 - x_{ik})/(2N_i - 1)}, \tag{3.5}$$

which is a consistent estimator to the extent that the numerator and denominator of the ratio are estimated by their respective unbiased estimators.

Note that if x_{ik}^2 is used as an estimator for p_{ik}^2 (with a negative bias of the order $1/2N_i$), the estimator for $F_{\mathrm{IS}ik}$ becomes

$$\hat{F}'_{\mathrm{IS}ik'} = (X_{ik} - x_{ik}^2)/[x_{ik}(1 - x_{ik})], \tag{3.5a}$$

which is identical to Curie-Cohen's (1982) estimator $\hat{f}_1 = 1 - (y/x)$, where y is the observed heterozygosity for the A_k allele in the sample, and $x = 2N_i x_{ik}(1 - x_{ik})$, an estimator of its expectation under HWE.

It might be further noted that Nei's unbiased estimator (equation (3.5)) takes the form

$$\hat{F}_{IS\,ik} = 1 - [1 - 1/(2N_i)]\, y/x\,, \tag{3.5b}$$

which will be useful in deriving its standard error, shown in the next section.

Curie-Cohen (1982) showed that these equations have a natural multiple-allele extension, when y and x are interpreted as the total observed (H_0) and expected (H_E, under HWE) heterozygosity for all alleles at a locus. He, however, did not note the equivalence of his f_1 estimator with Nei's estimate, written in terms of H_0 and H_S, at a locus (Nei, 1977).

Let us now consider the joint analysis of data from several subpopulations. Since $\bar{p}_{\cdot k} = \sum_{i=1}^{s} w_i p_{ik}$, we have

$$\bar{p}^2_{\cdot k} = \sum_{i=1}^{s} w_i^2 p_{ik}^2 + \sum_{i \neq i'=1}^{s}\sum w_i w_{i'} p_{ik} p_{i'k}\,,$$

and hence, when the samples from the subpopulations are drawn independently of each other (as usually is the case), we obtain an unbiased estimator of $\bar{p}^2_{\cdot k}$, given by

$$\widehat{\bar{p}^2_{\cdot k}} = \sum_{i=1}^{s} w_i^2 \frac{x_{ik}(2N_i x_{ik} - 1)}{2N_i - 1} + \sum_{i \neq i'=1}^{s}\sum w_i w_{i'} x_{ik} x_{i'k}$$

$$= \left(\sum_{i=1}^{s} w_i x_{ik}\right)^2 - \sum_{i=1}^{s} w_i^2 \frac{x_{ik}(1 - x_{ik})}{2N_i - 1}\,.$$

Therefore, estimating the numerators and denominators in an unbiased fashion, we obtain the following estimators of the allele-specific fixation indices at a particular locus:

$$\hat{F}_{IS\,k} = \frac{\sum_{i=1}^{s} w_i(X_{ik} - x_{ik}^2) + \sum_{i=1}^{s} w_i x_{ik}(1 - x_{ik})/(2N_i - 1)}{\sum_{i=1}^{s} w_i 2N_i x_{ik}(1 - x_{ik})/(2N_i - 1)}\,, \tag{3.6}$$

$$\hat{F}_{IT\,k} = \frac{\sum_{i=1}^{s} w_i X_{ik} - (\sum_{i=1}^{s} w_i x_{ik})^2 + \sum_{i=1}^{s} w_i^2 x_{ik}(1 - x_{ik})/(2N_i - 1)}{\sum_{i=1}^{s} w_i x_{ik} - (\sum_{i=1}^{s} w_i x_{ik})^2 + \sum_{i=1}^{s} w_i^2 x_{ik}(1 - x_{ik})/(2N_i - 1)}\,, \tag{3.7}$$

$$\hat{F}_{ST\,k} = \frac{\sum_{i=1}^{s} w_i x_{ik}^2 - (\sum_{i=1}^{s} w_i x_{ik})^2 - \sum_{i=1}^{s} w_i(1 - w_i) x_{ik}(1 - x_{ik})/(2N_i - 1)}{\sum_{i=1}^{s} w_i x_{ik} - (\sum_{i=1}^{s} w_i x_{ik})^2 + \sum_{i=1}^{s} w_i^2 x_{ik}(1 - x_{ik})/(2N_i - 1)}\,. \tag{3.8}$$

Note that while all of these estimators are consistent, to the extent that the numerators and denominators of the parameters defined in equations (2.4)–(2.6) are estimated with their respective unbiased estimators, in applying these equations we need the relative sizes (w_i's) for all subpopulations. These are, however not known in practice; nor can they always be reliably substituted by relative sample sizes. Nei (1977) and Nei and Chesser (1983), therefore assumed that the w_i's are all equal, $w_i = 1/s$ for all i. In that event, equations (3.6)–(3.8) take the form

$$\hat{F}'_{\mathrm{IS}k} = \frac{\sum_{i=1}^{s} [(X_{ik} - x_{ik}^2) + x_{ik}(1 - x_{ik})/(2N_i - 1)]}{\sum_{i=1}^{s} 2N_i x_{ik}(1 - x_{ik})/(2N_i - 1)}, \tag{3.6a}$$

$$\hat{F}'_{\mathrm{IT}k} = \frac{\sum_{i=1}^{s} X_{ik} - (1/s)(\sum_{i=1}^{s} x_{ik})^2 + (1/s)\sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N_i - 1)}{\sum_{i=1}^{s} x_{ik} - (1/s)(\sum_{i=1}^{s} x_{ik})^2 + (1/s)\sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N_i - 1)}, \tag{3.7a}$$

and

$$\hat{F}'_{\mathrm{ST}k} = \frac{\sum_{i=1}^{s} x_{ik}^2 - (1/s)(\sum_{i=1}^{s} x_{ik})^2 - (1 - 1/s)\sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N_i - 1)}{\sum_{i=1}^{s} x_{ik} - (1/s)(\sum_{i=1}^{s} x_{ik})^2 + (1/s)\sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N_i - 1)}. \tag{3.8a}$$

When the sample sizes are large enough, so that $2N_i \approx 2N_i - 1$ and $\sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N_i - 1)$ is negligible, these equations take a much simpler form:

$$\hat{F}'_{\mathrm{IS}k} \approx \sum_{i=1}^{s} (X_{ik} - x_{ik})^2 \Big/ \sum_{i=1}^{s} x_{ik}(1 - x_{ik}), \tag{3.6b}$$

$$\hat{F}'_{\mathrm{IT}k} \approx \left[\sum_{i=1}^{s} X_{ik} - \frac{1}{s}\left(\sum_{i=1}^{s} x_{ik}\right)^2\right] \Big/ \left[\sum_{i=1}^{s} x_{ik} - \frac{1}{s}\left(\sum_{i=1}^{s} x_{ik}\right)^2\right], \tag{3.7b}$$

$$\hat{F}'_{\mathrm{ST}k} \approx \left[\sum_{i=1}^{s} x_{ik}^2 - \frac{1}{s}\left(\sum_{i=1}^{s} x_{ik}\right)^2\right] \Big/ \left[\sum_{i=1}^{s} x_{ik} - \frac{1}{s}\left(\sum_{i=1}^{s} x_{ik}\right)^2\right]. \tag{3.8b}$$

Note that equation (3.8b) takes the well-known form

$$\hat{F}'_{\mathrm{ST}k} \approx s_k^2/x_{\cdot k}(1 - x_{\cdot k}), \tag{3.8c}$$

where s_k^2 is the variance of the A_k-allele frequency over all subpopulations, $s_k^2 = \sum_{i=1}^{s} (x_{ik} - \bar{x}_{\cdot k})^2/s$, with $\bar{x}_{\cdot k}$ representing the average frequency of the A_k-allele over all subpopulations, $\bar{x}_{\cdot k} = \sum_{i=1}^{s} x_{ik}/s$.

When the investigators have sufficient reason to believe that the sampling from each subpopulation has been conducted in such a manner that the relative sample sizes (N_i/N) reflect their respective relative sizes (population values), one might replace the w_i's in equations (3.6)–(3.8) by their respective sample size weights, $\hat{w}_i = N_i/N$, and obtain the allele-specific estimates of the fixation indices. However, note that while this weighting may serve the purpose of taking into account the

relative contribution of each subpopulation in the total population in the current generation, they are not evolutionary stable, as N_i's can fluctuate drastically over time.

Pooling over all alleles at a locus, the locus-specific estimates of F_{IS}, F_{IT} and F_{ST} values can be obtained easily, since the respective parameter values have been defined by summing the numerators and denominators over all alleles at a locus (see equations (2.8)–(2.10). Since these equations are represented in terms of heterozygosities in the population, it may be worthwhile to express the unbiased estimators of H_S, H_0, and H_T explicitly. Nei and Chesser (1983) obtained such estimators, with the assumption that all w_i's are equal ($= 1/s$).

In our terminology, with any general weight (w_i's unequal), we may use the above mentioned unbiased estimators of p_{ik}, P_{ik}^2, and $p_{\cdot k}^2$ to obtain

$$\hat{H}_0 = 1 - \sum_{i=1}^{s} \sum_{k=1}^{r} w_i \hat{P}_{ik} = 1 - \sum_{i=1}^{s} \sum_{k=1}^{r} w_i X_{ik}\,, \tag{3.9}$$

$$\hat{H}_S = 1 - \sum_{i=1}^{s} \sum_{k=1}^{r} w_i \widehat{p_{ik}^2}$$

$$= 1 - \sum_{i=1}^{s} \sum_{k=1}^{r} \frac{w_i}{2N_i - 1}\left[2N_i \sum_{k=1}^{r} x_{ik}^2 - 1\right], \tag{3.10}$$

and

$$\hat{H}_T = \sum_{k=1}^{r}\left[\sum_{i=1}^{s} w_i x_{ik}\left(1 - \sum_{i=1}^{s} w_i x_{ik}\right) + \sum_{i=1}^{s} w_i^2 x_{ik}(1 - x_{ik})\Big/(2N_i - 1)\right] \tag{3.11}$$

as respective unbiased estimators of H_0, H_S, and H_T. Substitution of these estimators in equations (2.8)–(2.10) provide consistent estimators of the locus specific fixation indices.

When the w_i's are all equal, equations (3.9)–(3.11) reduce to

$$\hat{H}_0' = 1 - \frac{1}{s} \sum_{i=1}^{s} \sum_{k=1}^{r} X_{ik}\,, \tag{3.9a}$$

$$\hat{H}_S' = \frac{1}{s} \sum_{i=1}^{s} 2N_i \hat{H}_{Si}\Big/(2N_i - 1)\,, \tag{3.10a}$$

$$\hat{H}_T' = \sum_{k=1}^{r}\left[\bar{x}_{\cdot k}(1 - \bar{x}_{\cdot k}) + \frac{1}{s^2} \sum_{i=1}^{s} x_{ik}(1 - x_{ik})\Big/(2N_i - 1)\right], \tag{3.11a}$$

where

$$\hat{H}_{Si} = 1 - \sum_{k=1}^{r} x_{ik}^2 \quad \text{and} \quad \bar{x}_{\cdot k} = \sum_{i=1}^{s} x_{ik}/s\,.$$

Note that these estimators are exact unbiased estimators of H_0, H_S, and H_T when all subpopulations are of equal size (but N_i's need not be equal), whereas

the estimators given by Nei and Chesser (1983) involve some approximations (see their equation (8) in particular). As before, when the N_i's are large, we may equate $2N_i/(2N_i - 1)$ to unity, and neglect the last term of $\hat{H}'_T$, to get

$$\hat{H}'_S \approx \frac{1}{s} \sum_{i=1}^{s} \hat{H}_{Si} \tag{3.10b}$$

and

$$\hat{H}'_T \approx 1 - \sum_{k=1}^{r} \bar{x}^2_{\cdot k} . \tag{3.11b}$$

Thus, when all genotypes are recognizable, unbiased estimators of H_0, H_S, and H_T can be obtained simply by enumerating all allele frequencies in each subpopulation (by gene counting) and evaluating the sum total of all homozygotes (X_{ik}'s). The resulting estimators

$$\hat{F}_{IS} = 1 - \hat{H}_0/\hat{H}_S , \tag{3.12}$$

$$\hat{F}_{IT} = 1 - \hat{H}_0/\hat{H}_T , \tag{3.13}$$

and

$$\hat{F}_{ST} = 1 - \hat{H}_S/\hat{H}_T , \tag{3.14}$$

are again consistent, to the extent that in these the numerators and denominators are estimated by their respective unbiased statistics.

Estimation of parameters pooled over several loci can be achieved exactly in the same manner, by defining the heterozygosities (H_0, H_S, H_T) as averages over all loci.

3.3. Estimators by Cockerham's approach

As shown in equations (2.11)–(2.13), Cockerham (1973) derived the allele-specific fixation indices in terms of components of variance in a nested analysis of variance. In this approach, the estimation of fixation indices reduces to the problem of estimating the components a_k, b_k, and c_k. Weir and Cockerham (1984) gave the explicit forms of these estimators, they are

$$\hat{a}_k = \frac{\bar{N}}{N_c} s^2_k(\hat{w}) - \frac{1}{\bar{N} - 1}\left[\bar{x}_{\cdot k}(\hat{w})(1 - \bar{x}_{\cdot k}(\hat{w})) - \frac{s-1}{s} s^2_k(\hat{w}) - \tfrac{1}{4}\bar{h}(\hat{w})\right], \tag{3.15}$$

$$\hat{b}_k = \frac{\bar{N}}{\bar{N} - 1} \bar{x}_{\cdot k}(\hat{w})[1 - \bar{x}_{\cdot k}(\hat{w})] - \frac{s-1}{s} s^2_k(\hat{w}) - \frac{2\bar{N} - 1}{4\bar{N}} \bar{h}(\hat{w}) , \tag{3.16}$$

and

$$\hat{c}_k = \tfrac{1}{2}\bar{h}(\hat{w}) , \tag{3.17}$$

where

$\bar{N} = \sum_{i=1}^{s} N_i/s$ is the average number of individuals sampled per subpopulation,
$N_c = [s\bar{N} - \sum_{i=1}^{s} N_i^2/s\bar{N}]/(s-1) = \bar{N}(1 - C^2/s)$, where C is the coefficient of variation of sample sizes (N_i's),
$\bar{x}_{.k}(\hat{w}) = \sum_{i=1}^{s} N_i x_{ik}/s\bar{N}$, the weighted average allele frequency of A_k per subpopulation,
$s_k^2(\hat{w}) = \sum_{i=1}^{s} N_i(x_{ik} - \bar{x}_{.k}(\hat{w}))^2/(s-1)\bar{N}$, is the variance of A_k allele frequencies over subpopulations,
$\bar{h}(\hat{w}) = \sum_{i=1}^{s} N_i h_i(\hat{w})/s\bar{N}$, the average observed heterozygote frequency for allele A_k.

In parallel to equations (2.11)–(2.13), the estimators F_{ITk}, F_{ISk}, and F_{STk} become

$$\hat{F}_{ITk} = (\hat{a}_k + \hat{b}_k)/(\hat{a}_k + \hat{b}_k + \hat{c}_k), \tag{3.15a}$$

and

$$\hat{F}_{ISk} = \hat{b}_k/(\hat{b}_k + \hat{c}_k), \tag{3.16a}$$

$$\hat{F}_{STk} = \hat{a}_k/(\hat{a}_k + \hat{b}_k + \hat{c}_k). \tag{3.17a}$$

Note that these expressions are defined in terms of weighted variance components, where sample sizes from the subpopulations are taken as weights, irrespective of their true relative population sizes (i.e., $\hat{w}_i = N_i/N$). These explicit forms are obtained by algebraic manipulations of the estimated mean square errors in Table 3 (Cockerham, 1973). It should be noted that Cockerham's Table 3 (Cockerham, 1973, p. 688) has an inadvertant error, where the expressions S_a and S_a' should have an additional coefficient 2, which is missing.

Cockerham (1973) also gave an explicit estimator for $F_{IS\,ik}$, the F_{IS} estimator for a specific allele (A_k) in the i-th subpopulation, which has the form

$$\hat{F}_{IS\,ik} = 1 - \frac{4(N_i - 1)[N_i x_{ik} - N_{ikk}]}{4N_i^2 x_{ik}(1 - x_{ik}) - 2(N_i x_{ik} - N_{ikk})}, \tag{3.18}$$

that can be computed from the respective subpopulation-specific genotype data. A pooled estimator of F_{IS}, pooled over all alleles at a locus, can be obtained by summing the numerator and the denominator of equation (3.18), as done in the other cases.

In particular, when N_i is large, the pooled estimator over all alleles at a locus takes the form

$$\hat{F}_{IS\,i} = 1 - \frac{1}{2N_i}\left[\sum_{i=1}^{r} h_{ik}\right] \Big/ \left[1 - \sum_{i=1}^{r} x_{ik}^2\right], \tag{3.18a}$$

where h_{ik} is the observed number of heterozygotes carrying the A_k allele in the i-th subpopulation.

While equation (3.18) can be derived even without invoking the variance com-

ponents (see Cockerham, 1969, pp. 689–690), this is different from Nei's estimator (our equation (3.5)), which estimates $F_{IS\,ik}$ as a ratio estimator, based on equation (2.3). Both estimators are asymptotically unbiased (since each of them estimates the numerator and denominator by their respective unbiased statistics).

Setting up the equivalence of Cockerham's (1973) and Curie-Cohen's (1982) notations, it may be shown that the above estimator takes the form

$$\hat{F}_{IS\,ik} = [2N_i(x - y) + y]/(2N_i x - y)\,, \tag{3.18b}$$

where $x = 2N_i x_{ik}(1 - x_{ik})$, and $y\ (= \sum_{l>k}^{r} N_{ikl})$ is the observed heterozygosity for the A_k-allele in a particular subpopulation. This equivalence will also be useful in deriving the standard error of this estimator (discussed in the next section).

At this stage, since we have three alternative estimators of $F_{IS\,ik}$: Nei's unbiased (equation (3.5)), biased (equation (3.5a)), and Cockerham's (equation (3.18)), it might be worthwhile to study how they behave for a given sample.

It can be shown that

$$\hat{F}_{IS\,ik} - \hat{F}'_{IS\,ik} = (x_{ik} - X_{ik})/2N_i x_{ik}(1 - x_{ik})\,, \tag{3.19}$$

where $\hat{F}_{IS\,ik}$ is from equation (3.5) and $\hat{F}'_{IS\,ik}$ is from equation (3.5a).

Since $x_{ik} \geqslant X_{ik}$ in any given sample, we have the inequality

$$\text{Nei's unbiased estimator} \geqslant \text{Nei's biased estimator}\,, \tag{3.20}$$

over the entire sample space.

Furthermore, the expected difference of these two estimators,

$$\mathrm{E}[\hat{F}_{IS\,ik} - \hat{F}'_{IS\,ik}] \approx (1 - F_{IS\,ik})/(2N_i - 1)\,, \tag{3.21}$$

which is usually very small, of the order $(2N_i - 1)^{-1}$.

Similarly, we can show that

$$\text{Cockerham's estimator (equation (3.18))} > \text{Nei's biased estimator (equation (3.5a))}, \tag{3.22}$$

over the entire sample space.

The relationship between Nei's unbiased and Cockerham's estimator is a little bit more involved. For simplicity, using Curie-Cohen's notation [y = observed number of heterozygotes and x = an estimator of the expected number of heterozygotes, for a specific allele = $2N_i x_{ik}(1 - x_{ik})$], we get

$$\text{Cockerham's estimator (equation (3.18)} - \text{Nei's unbiased estimator (equation 3.5)} = y/(2N_i x) \cdot \text{Cockerham's estimator (equation (3.18))}. \tag{3.23}$$

Hence, when Cockerham's estimator is negative we have the string of inequalities

$$\text{equation (3.5)} \geqslant \text{equation (3.18)} \geqslant \text{equation (3.5a)}, \tag{3.24}$$

i.e., Cockerham's estimator is bounded by Nei's biased and unbiased estimators.

However, when Cockerham's estimator is positive, from equation (3.23) we have

$$\text{equation (3.18)} \geqslant \text{equation (3.5)} \geqslant \text{equation (3.5a)}, \tag{3.25}$$

i.e., Nei's unbiased estimator is bounded by his biased estimator and that of Cockerham.

These inequalities also hold for locus-specific estimators, irrespective of the number of alleles and allele frequencies. To our knowledge, this mathematical relationship among these three estimators has not been demonstrated before. Since the expected differences are of the order of inverse of the number of genes sampled ($2N_i$) in a subpopulation, they are generally much smaller than their standard errors, which will be shown later.

It is worthwhile to note that while Nei's (1977) or Nei and Chesser's (1983) estimate of F_{ST_k} (see equation (3.8) or (3.8a)) is only a function of allele frequencies in all subpopulations, Weir and Cockerham's (1984) estimator of F_{ST_k} also depends on the frequencies of observed heterozygosity for the A_k allele in the sample.

Weir and Cockerham (1984) also gave explicit expressions for approximations for these general estimators under several special cases. In particular, they note that when the N_i's are large, the above estimators take the form

$$\hat{F}'_{IT_k} = 1 - \frac{[1 - C^2/s]\bar{h}(\hat{w})}{2[1 - C^2/s]\bar{x}_{\cdot k}(\hat{w})\{1 - \bar{x}_{\cdot k}(\hat{w})\} + 2[1 + (s-1)/s \cdot C^2]s_k^2(\hat{w})/s}, \tag{3.15b}$$

$$\hat{F}'_{IS_k} = 1 - \frac{\bar{h}(\hat{w})}{2\bar{x}_{\cdot k}(\hat{w})\{1 - \bar{x}_{\cdot k}(\hat{w})\} - 2(s-1)s_k^2(\hat{w})/s}, \tag{3.16b}$$

and

$$\hat{F}'_{ST_k} = \frac{s_k^2(\hat{w})}{[1 - C^2/s]\bar{x}_{\cdot k}(\hat{w})\{1 - \bar{x}_{\cdot k}(\hat{w})\} + [1 + (s-1)/s \cdot C^2]s_k^2(\hat{w})/s}, \tag{3.17b}$$

in which $\hat{F}'_{ST_k}$ can be calculated only from allele frequency data. In addition to the N_i's being large, if s (the number of subpopulations) is also large, Weir–Cockerham's estimate of F_{ST_k} takes the well known form of

$$\hat{F}''_{ST_k} = s_k^2/\bar{x}_{\cdot k}(1 - \bar{x}_{\cdot k}).$$

Note that, while the general estimator of F_{STk} in Cockerham's approach depends upon the genotype frequencies (equation (3.17a)), its large sample approximation (equation (3.17b)) is only dependent on allele frequencies.

Weir and Cockerham (1984) suggested that locus-specific estimators for F_{IS}, F_{IT}, and F_{ST} can be derived by summing $\hat{a}_k$, $\hat{b}_k$, and $\hat{c}_k$ over all alleles, so that

$$\hat{F}_{IS} = \sum_{k=1}^{r} \hat{b}_k \Big/ \sum_{k=1}^{r} (\hat{b}_k + \hat{c}_k), \tag{3.26}$$

$$\hat{F}_{IT} = \sum_{k=1}^{r} (\hat{a}_k + \hat{b}_k) \Big/ \sum_{k=1}^{r} (\hat{a}_k + \hat{b}_k + \hat{c}_k), \tag{3.27}$$

and

$$\hat{F}_{ST} = \sum_{k=1}^{r} \hat{a}_k \Big/ \sum_{k=1}^{r} (\hat{a}_k + \hat{b}_k + \hat{c}_k). \tag{3.28}$$

Although other methods of pooling data of multiple alleles exist (e.g., Wright, 1965; Kirby, 1975; Robertson and Hill, 1984), Weir and Cockerham (1984) advocate that the method presented above (equations (3.26)–(3.28)) is more appropriate for ratio estimators (see also Reynolds et al., 1983).

Note that since the parametric value of b_k can be negative (see equation (2.16)), it is quite possible that in this approach $\hat{F}_{ST}$ can often exceed $\hat{F}_{IT}$. Van Den Bussche et al. (1986) also noted that negative estimates of F_{STk} (or F_{ST}) can arise in Weir–Cockerham's approach when the following inequality holds:

$$s_k^2(\hat{w}) < \frac{1}{N-1}\left[\bar{x}_{\cdot k}(\hat{w})\{1 - \bar{x}_{\cdot k}(\hat{w})\} - \frac{s-1}{s}\, s_k^2(\hat{w}) - \tfrac{1}{4}\bar{h}(\hat{w})\right]. \tag{3.29}$$

While it is possible that Nei and Chesser's (1983) estimator of F_{ST} can also be negative (where $\hat{H}_S > \hat{H}_T$ occur), several simulation studies show that the negative estimates of F_{ST} are more common in the variance component approach (Chakraborty and Leimar, 1987; Van Den Bussche et al., 1986; Slatkin and Barton, 1989).

Finally, equations (3.26)–(3.28) can be extended to obtain pooled estimators of all indices, summing the numerators and denominators over all alleles over several loci (see equation (10) of Weir and Cockerham, 1984, p. 1364).

3.4. Long's estimators for multiple alleles and multiple loci

Long (1986) provided an interesting extension of Cockerham's approach for multiple alleles. He noted that when multiple alleles ($r > 2$) are involved at a locus, summation of a_k, b_k, and c_k over alleles (as suggested by Weir and Cockerham, 1984) ignores the correlation of allele and genotype frequencies (that is inherent in a multinomial sampling of genotypes) within subpopulations. Although this idea is imbedded in Weir–Cockerham's work (see their Appendix, termed as matrix estimation method), the formulation is explicitly stated in Long (1986) in terms

of the decomposition of multivariate dispersion matrices. The parameters, as defined by equations (2.18)–(2.20), can be estimated substituting the estimators for the Σ_a, Σ_b, and Σ_c matrices. Long (1986) provided computational formulae for such estimators (see Appendix of Long, 1986) which involve the genotype and allele counts within each subpopulation and their totals over all subpopulations.

Since there are several misprints in the formulae in Long's (1986) paper (see pp. 646–647), we present the general estimation procedure for a r-allelic codominant locus. This has two purposes: first, this exposition clearly indicates how Weir and Cockerham's (1984) expressions have their natural multivariate extensions and, second, it will indicate why Long's algorithm gives numerical results different from those of Weir and Cockerham for a multiallelic locus ($r > 2$). Furthermore, we derive here the explicit closed expressions for the Σ_a, Σ_b, and Σ_c matrices, that are not available in Long (1986). For a single subpopulation, closed expressions for $\Sigma_b + \Sigma_c$ matrix are also shown through this exposition.

For a specific subpopulation, when an estimator for F_{IS} is sought (in parallel to $F_{IS\,ik}$ estimator, as done for Nei's and Cockerham's method earlier—only difference being in Long's procedure we need a different pooling algorithm over all alleles), a multivariate variance–covariance decomposition can be done in analogy of Table 3 of Cockerham (1973). The within-individual mean-square cross-product matrix (MSCP) S_c (equivalent to S_{wk} of Cockerham) for the i-th subpopulation takes the form, whose k-th diagonal element,

$$h_{ik}/2N_i, \quad \text{where } h_{ik} = \sum_{l>k=1}^{r} N_{ikl},$$

is the observed number of heterozygotes with reference to the A_k-allele in the i-th subpopulation, and the (k, l)-th off-diagonal element of the S_c matrix is $-h_{ikl}/2N_i$, where $h_{ikl} = N_{ikl}$, the observed number of A_kA_l heterozygotes in the i-th subpopulation.

Algebraic manipulation of the MSCP matrix for between individual source of variation, S_b matrix has:

$$k\text{-th diagonal element} = \frac{4N_i x_{ik}(1 - x_{ik}) - h_{ik}}{2(N_i - 1)} \tag{3.30a}$$

and

$$(k, l)\text{-th off-diagonal element} = \frac{h_{ikl} - 4N_i x_{ik} x_{il}}{2(N_i - 1)}, \tag{3.30b}$$

where the x_{ik}'s are as defined in equation (3.2).

These matrices are square matrices of dimension $r - 1$, since the linear constraint of allele frequencies (summation of all allele frequencies at a particular locus being one) has to be used in order to make such matrices non-singular (a requirement needed for the computations done in the sequel).

Estimator of Σ_b matrix (variance–covariance component due to between-

individual source of variation) is obtained as

$$\hat{\Sigma}_b = \tfrac{1}{2}[\mathrm{MSCP}(b) - \mathrm{MSCP}(c)] = \tfrac{1}{2}[S_b - S_c]\,, \tag{3.31}$$

since

$$\mathrm{E}[\mathrm{MSCP}(b)] = \Sigma_c + 2\Sigma_b \quad \text{and} \quad \mathrm{E}[\mathrm{MSCP}(c)] = \Sigma_c$$

(see Cockerham (1973, p. 688).

Therefore, $\hat{\Sigma}_b$ matrix has the form, whose

$$k\text{-th diagonal element} = \frac{4N_i^2 x_{ik}(1 - x_{ik}) - (2N_i - 1)h_{ik}}{4N_i(N_i - 1)} \tag{3.32a}$$

and

$$(k, l)\text{-th element} = \frac{h_{ikl}(2N_i - 1) - 4N_i^2 x_{ik} x_{il}}{4N_i(N_i - 1)}\,, \tag{3.32b}$$

for $k, l = 1, 2, \ldots, r - 1$.

In order to estimate $F_{\mathrm{IS}i}$, we need the matrix $\hat{\Sigma}_b + \hat{\Sigma}_c$, whose

$$k\text{-th diagonal element} = \frac{4N_i^2 x_{ik}(1 - x_{ik}) - h_{ik}}{4N_i(N_i - 1)} \tag{3.33a}$$

and

$$(k, l)\text{-th element} = \frac{h_{ikl} - 4N_i^2 x_{ik} x_{il}}{4N_i(N_i - 1)}\,, \tag{3.33b}$$

for $k, l = 1, 2, \ldots, r - 1$.

With these computations, the estimator for $F_{\mathrm{IS}i}$ is

$$\hat{F}_{\mathrm{IS}i} = \frac{1}{r - 1}\,\mathrm{tr}[(\hat{\Sigma}_b + \hat{\Sigma}_c)^{-1/2}\,\hat{\Sigma}_b(\hat{\Sigma}_b + \hat{\Sigma}_c)^{-1/2}]\,. \tag{3.34}$$

Although no closed explicit expression for $\hat{F}_{\mathrm{IS}i}$ can be given in general (for $r > 2$), the explicit expressions for the elements of $\hat{\Sigma}_b + \hat{\Sigma}_c$ and $\hat{\Sigma}_c$ matrices are instructive to understand why the numerical values of Long's estimators are different from Weir–Cockerham's estimators. For example, even if all off-diagonal elements are neglected, equation (3.34) would yield

$$\hat{F}'_{\mathrm{IS}i} = \frac{1}{r - 1}\sum_{k=1}^{r-1}\left[\frac{4N_i^2 x_{ik}(1 - x_{ik}) - (2N_i - 1)h_{ik}}{4N_i^2 x_{ik}(1 - x_{ik}) - h_{ik}}\right]$$

$$= 1 - \frac{1}{r - 1}\sum_{k=1}^{r-1}\left[\frac{2(N_i - 1)h_{ik}}{4N_i^2 x_{ik}(1 - x_{ik}) - h_{ik}}\right], \tag{3.34a}$$

whereas Weir and Cockerham's (1984) algorithm would suggest the computation of

$$\hat{F}''_{IS_i} = 1 - \left[\sum_{k=1}^{r} 2(N_i - 1)h_{ik}\right] \Big/ \left[\sum_{k=1}^{r} [4N_i^2 x_{ik}(1 - x_{ik}) - h_{ik}]\right]. \tag{3.34b}$$

While for a bi-allelic locus ($r = 2$), equations (3.34), (3.34a), and (3.34b) are identical, there are a number of practical limitations of equation (3.34) which are worth noting. For instance, suppose that there are multiple alleles ($r > 2$) in the total population, but in each subpopulation one or several are not present (either in the sample, or in the subpopulation as a whole), and the missing alleles vary across subpopulations. In such an event, for each subpopulation the S_b and S_c matrices will be of different dimension, and would refer to different sets of alleles. Therefore, in the strict sense F_{IS_i} values computed from equation (3.34) cannot be contrasted across subpopulations, since they are based on different sets of alleles even when they belong to the same locus.

Nevertheless, the large sample estimator for F_{IS_i}, following the matrix method has a closed form, not noted by Weir and Cockerham (1984) or Long (1986). Note that when the N_i's are large, ignoring terms of the order N_i^{-2}, we have

$$(\hat{\Sigma}_b + \hat{\Sigma}_c)_{kl} = \begin{cases} x_{ik}(1 - x_{ik}) & \text{for } k = l, \quad (3.33c) \\ -x_{ik}x_{il} & \text{for } k \neq l, \quad (3.33d) \end{cases}$$

for $k, l = 1, 2, \ldots, r - 1$ at a locus.

The (k, l)-th element of the $(\hat{\Sigma}_b + \hat{\Sigma}_c)^{-1}$ matrix has the form

$$(\hat{\Sigma}_b + \hat{\Sigma}_c)^{-1}_{kl} = \begin{cases} 1/x_{ik} + 1/x_{ir} & \text{for } k = l, \\ 1/x_{ir} & \text{for } k \neq l, \end{cases}$$

for $k, l = 1, 2, \ldots, r - 1$.

Therefore, if we estimate F_{IS_i} by

$$\hat{F}'_{IS_i} = (r - 1)^{-1} \operatorname{tr}[(\hat{\Sigma}_b + \hat{\Sigma}_c)^{-1} \hat{\Sigma}_b],$$

it has a closed form

$$\begin{aligned} \hat{F}'_{IS_i} &= 1 - [2N_i(r - 1)]^{-1} \sum_{k=1}^{r} h_{ik}/x_{ik} \\ &= (r - 1)^{-1}\left[\sum_{k=1}^{r} (x_{ik}/p_{ik}) - 1\right], \end{aligned} \tag{3.35}$$

while Cockerham's estimator, pooled over alleles has a large sample form given in equation (3.18a).

Note that equation (3.35) is identical to the estimator $\hat{f}_2$ used by Curie-Cohen (1982), although he arrived at this estimator by a different logic.

When several subpopulations are analysed together, nested multivariate variance–covariance analysis was performed by Long (1986), to obtain the estimators for three variance–covariance component matrices (VCCM's) as

$$S_c = \text{MSCP}(c), \tag{3.36}$$

$$S_b = \tfrac{1}{2}[\text{MSCP}(b) - \text{MSCP}(c)], \tag{3.37}$$

$$S_a = (1/2N_c)[\text{MSCP}(a) - \text{MSCP}(b)], \tag{3.38}$$

where N_c is as defined in equations (3.15)–(3.17). Here again, each of these matrices are square matrices of dimension $(r-1)$. As in the univariate case (equations (3.15)–(3.17)), explicit closed forms of these three matrices can be written which are not given in Long (1986). Long's equation for the MSCP(c) matrix (called MSCP(W) in Long, 1986) for a three allelic locus has a misprint (see his equation on top of p. 647) which fails to show how such a matrix can be computed for a multi-allelic locus. If we write the (k, l)-th element of S_a, S_b, and S_c as a_{kl}, b_{kl}, and c_{kl}, respectively, algebraic manipulation yields

$$\hat{a}_{kk} = \frac{\overline{N}}{N_C}\left\{s_k^2(\hat{w}) - \frac{1}{\overline{N}-1}\left[\bar{x}_{\cdot k}(\hat{w})\{1 - \bar{x}_{\cdot k}(\hat{w})\} - [(s-1)/s]s_k^2(\hat{w}) - \tfrac{1}{4}\bar{h}_k(\hat{w})\right]\right\}, \tag{3.37a}$$

$$\hat{a}_{kl} = \frac{\overline{N}}{N_C}\left\{s_{kl}(\hat{w}) - \frac{1}{\overline{N}-1}\left[\bar{x}_{\cdot k}(\hat{w})\bar{x}_{\cdot l}(\hat{w}) + [(s-1)/s]s_{kl}(\hat{w}) - \tfrac{1}{4}\bar{h}_{kl}(\hat{w})\right]\right\}, \tag{3.37b}$$

$$\hat{b}_{kk} = \frac{\overline{N}}{\overline{N}-1}\left\{\bar{x}_{\cdot k}(\hat{w})\{1 - \bar{x}_{\cdot k}(\hat{w})\} - [(s-1)/s]s_k^2(\hat{w}) - [(2\overline{N}-1)/4\overline{N}]\bar{h}_k(\hat{w})\right\}, \tag{3.38a}$$

$$\hat{b}_{kl} = \frac{\overline{N}}{\overline{N}-1}\left\{[(2\overline{N}-1)/4\overline{N}]\bar{h}_{kl}(\hat{w}) - \bar{x}_{\cdot k}(\hat{w})\bar{x}_{\cdot l}(\hat{w}) - [(s-1)/s]s_{kl}(\hat{w})\right\}, \tag{3.38b}$$

$$\hat{c}_{kk} = \tfrac{1}{2}\bar{h}_k(\hat{w}), \tag{3.39a}$$

$$\hat{c}_{kl} = \tfrac{1}{2}\bar{h}_{kl}(\hat{w}), \tag{3.39b}$$

for $k, l = 1, 2, \ldots, r-1$, where $\bar{x}_{\cdot k}(\hat{w})$ and $s_k^2(\hat{w})$ are as defined in the context of equations (3.15)–(3.17), and

$$s_{kl}(\hat{w}) = \left[\sum_{i=1}^{s} N_i x_{ik} x_{il} - s\overline{N}\bar{x}_{\cdot k}(\hat{w})\bar{x}_{\cdot l}(\hat{w})\right] \Big/ \overline{N}(s-1)$$

is the covariance of the allele frequencies of A_k and A_l over all subpopulations; $\bar{h}_k(\hat{w})$, the observed heterozygote frequency of the A_k allele over subpopulations ($= s\bar{N}\sum_{i=1}^{s} h_{ik}$), and $\bar{h}_{kl}(\hat{w}) = \sum_{i=1}^{s} h_{ikl}/s\bar{N}$ is the average observed frequency of a specific heterozygote A_kA_l over all subpopulations.

Note that equations (3.37a), (3.38a), and (3.39a) are identical to the A_k-allele specific variance components described by Weir and Cockerham (1984), while equations (3.37b), (3.38b), and (3.39b) are direct extensions of these with multinomial sampling of genotypes.

With these explicit general closed form expressions of the elements of S_a, S_b, and S_c matrices one can compute the F_{IS}, F_{IT}, and F_{ST} estimators:

$$\hat{F}_{IS} = \frac{1}{r-1}\,\mathrm{tr}[(S_b + S_c)^{-1/2} S_b (S_b + S_c)^{-1/2}], \tag{3.40}$$

$$\hat{F}_{ST} = \frac{1}{r-1}\,\mathrm{tr}[(S_a + S_b + S_c)^{-1/2} S_a (S_a + S_b + S_c)^{-1/2}], \tag{3.41}$$

and

$$\hat{F}_{IT} = \frac{1}{r-1}\,\mathrm{tr}[(S_a + S_b + S_c)^{-1/2}(S_a + S_b) \times (S_a + S_b + S_c)^{-1/2}], \tag{3.42}$$

with far more ease than following Long's (1986) suggestion. Note that like the one-subpopulation situation, even if the off-diagonal elements ($\hat{a}_{kl}$, $\hat{b}_{kl}$, $\hat{c}_{kl}$) are neglected, instead of Weir–Cockerham's estimates (equations (3.26)–(3.28)), equations (3.40)–(3.42) take the respective forms

$$\hat{F}'_{IS} = \frac{1}{r-1}\sum_{k=1}^{r-1}\frac{\hat{b}_{kk}}{(\hat{b}_{kk} + \hat{c}_{kk})}, \tag{3.40a}$$

$$\hat{F}'_{ST} = \frac{1}{r-1}\sum_{k=1}^{r-1}\frac{\hat{a}_{kk}}{(\hat{a}_{kk} + \hat{b}_{kk} + \hat{c}_{kk})}, \tag{3.41a}$$

$$\hat{F}'_{IT} = \frac{1}{r-1}\sum_{k=1}^{r-1}\frac{\hat{a}_{kk} + \hat{b}_{kk}}{(\hat{a}_{kk} + \hat{b}_{kk} + \hat{c}_{kk})}, \tag{3.42a}$$

which perform worse than the estimators (3.26)–(3.28) in Weir and Cockerham's (1984) simulation experiments. Furthermore, the F_{IS} estimator obtained from equation (3.40) is not a weighted average of the subpopulation-specific $\hat{F}_{IS,i}$ values obtained from equation (3.34) since for each specific subpopulation the matrices can have different dimensions for reasons stated earlier.

At this point it is worthwhile to mention that this multivariate extension has not been presented explicitly before. Although Weir and Cockerham (1984) found that the estimators by such matrix method have the smallest standard errors in comparison with various other estimators they examined, their computations of the

matrix estimators are somewhat different from those of Long (1986). Instead of $\Sigma^{-1/2}\Sigma_a\Sigma^{-1/2}$, Weir and Cockerham used $\Sigma^{-1}\Sigma_a$. Since the Σ matrices, as well as their estimators, are always symmetric square matrices, it is not clear why Long's procedure of pre- and post-multiplication with $-\frac{1}{2}$ power of the $S_a + S_b + S_c$ or $S_b + S_c$ matrices is needed. In fact, since such estimators can be computed only for non-singular $S_a + S_b + S_c$ and $S_b + S_c$ matrices, if we define the fixation indices by

$$F_{IT} = (r-1)^{-1}\,\mathrm{tr}[\Sigma^{-1}(\Sigma_a + \Sigma_b)], \tag{2.18a}$$

$$F_{ST} = (r-1)^{-1}\,\mathrm{tr}[\Sigma^{-1}\Sigma_a], \tag{2.19a}$$

$$F_{IS} = (r-1)^{-1}\,\mathrm{tr}[(\Sigma_b + \Sigma_c)^{-1}\Sigma_b], \tag{2.20a}$$

instead of equations (2.18)–(2.20), only matrix-inversion routines are needed as opposed to the evaluation of eigen values and eigen vectors and inverse computations of the eigen vector matrices that are required in Long's algorithm.

Like the Weir and Cockerham estimator of F_{ST} (equation (3.28)), the estimator given by equation (3.41) also depends on the observed frequency of heterozygotes (see equations (3.37a) and (3.37b)) in addition to allele frequency data, which makes these estimators qualitatively different from that in Nei's approach (equation (3.8a)). Since in most practical situations the off-diagonal elements (a_{kl}, b_{kl}, c_{kl} for $k \neq l$) are small, because the subpopulations are sampled independently; the complexity of computations can be greatly reduced when Weir–Cockerham estimators are computed (according to equations (3.26)–(3.28)) for multi-allelic loci in the variance–component approach to estimation.

3.5. Estimation where genotype data are not available

Sometimes population structure analyses may have to be done in the absence of genotype data. Such is the case where the population structure is to be inferred from the allele frequency data reported in the literature, or the allele frequencies are estimated from phenotypic data at loci where complex dominance relationships exist among various alleles or haplotypes (e.g., ABO, Rh, and HLA system in man). Obviously, since such data do not provide any direct information regarding the observed number (or proportion) of homozygotes or heterozygotes, a somewhat different estimation procedure must be adopted.

In this case, Nei's approach can be easily adopted for estimating F_{ST}, since H_S and H_T parameters can be obtained simply from the estimated allele frequencies (with the assumption that the x_{ik}'s are multinomial proportions from a sample of $2N_i$ genes sampled from the i-th subpopulation). Equation (3.8) or its variant, equation (3.8a) with $w_i = 1/s$, is the estimator of preference here. Since F_{IS} is defined in terms of the deviation of genotype frequencies from their HWE expectations, no direct estimation of this quantity is possible. However, some approximate theory of estimation may be suggested.

Note that in the case of genotype data, the goodness-of-fit χ^2 statistic (of

testing for deviation from HWE expectations) for a r-allelic locus is $\chi^2 = N_i(r-1)F_{IS_i}$ (Li, 1955), and hence an approximate absolute value of F_{IS} can be obtained from $\sqrt{\chi^2/N_i(r-1)}$ where N_i is the number of individuals sampled from the i-th subpopulation. However, this suggested estimator is quite approximate, since for the loci in a dominance system, the goodness-of-fit statistic has a more complex parametric form (see Rao and Chakraborty, 1974). Furthermore, the sign of F_{IS} cannot be directly inferred from the χ^2 statistics. We advocate that for such data, only F_{ST} estimation is legitimate.

If one prefers the analysis of variance approach even the exact estimation of F_{ST} is not possible, unless large sample approximations are made. This is so because Weir and Cockerham's (1984) estimator of F_{ST} requires estimation of the observed heterozygosity for each allele (see equation (3.15) and so is the case with Long's approach (see equations (3.37a), (3.37b) and (3.41)). Under the assumption $F_{IS} = 0$ (random union of gametes within subpopulations), since $F_{IT} = F_{ST}$, Weir and Cockerham (1984, 1963) obtained the estimator

$$\hat{F}_{STk} = \frac{s_k^2(\hat{w}) - \left[\bar{x}_{.k}(\hat{w})(1-\bar{x}_{.k}(\hat{w})) - \frac{s-1}{s}s_k^2(\hat{w})\right] \Big/ [2\bar{N}-1]}{\left\{1 - \frac{2\bar{N}C^2}{(2\bar{N}-1)s}\right\}\bar{x}_{.k}(\hat{w})\{1-\bar{x}_{.k}(\hat{w})\} + \left\{1 + \frac{2\bar{N}(s-1)C^2}{(2\bar{N}-1)s}\right\}\frac{s_k^2(\hat{w})}{s}}, \tag{3.43}$$

where $\bar{x}_{.k}(\hat{w})$ and $s_k^2(\hat{w})$ are the weighted mean and variance of the A_k-allele frequency over all subpopulations (defined in equations (3.15)–(3.17)), and C^2 is the coefficient of variation of N_i's over all subpopulations (note that $1 - C^2/s = N_c$, where N_c is as defined in equations (3.15)–(3.17)). Clearly this estimator depends only on allele frequency data. Therefore, the analysis of variance approach, when applied to allele frequency data, also yields a consistent estimator for F_{ST} under the assumption that $F_{IS} = 0$. For large sample sizes, this assumption is, however, not needed (see equation (3.17b).

When all subpopulations have the same sample size (i.e., $N_i = N$), equation (3.43) takes the form

$$\hat{F}_{STk} = \frac{s_k^2 - \{\bar{x}_{.k}(1-\bar{x}_{.k}) - [(s-1)/s]s_k^2\}/(2\bar{N}-1)}{\bar{x}_{.k}(1-\bar{x}_{.k}) + s_k^2/s}, \tag{3.43a}$$

which reduces to the well known formula $s_k^2/\bar{x}_{.k}(1-\bar{x}_{.k})$ when $\bar{N}$ and s are large.

4. Standard errors and hypothesis testing

The discussions in the earlier sections clearly indicate that the problem of estimation of the fixation indices arises because these are defined as ratios of functions of allele and genotype frequencies in the subpopulations, and hence, strictly speaking none of the estimators suggested above can be claimed most efficient. We arrived at consistent estimators by estimating the numerators and denominators by their respective unbiased statistics. Although several expressions for the standard errors of these estimators are suggested, and the question of hypothesis testing has been addressed in a variety of ways, we agree with Cockerham (1973) that such procedures are on much less sound grounds than estimation. Nevertheless, since all estimators derived above are of the general form $\hat{\theta} = t_1/t_2$, where t_1 and t_2 are the estimators of the numerators and the denominators of the respective fixation index parameters, using Taylor's expansion (Kendall and Stuart, 1977, p. 247), an approximate formula for the variance of $\hat{\theta}$ can be written as

$$V(\hat{\theta}) \approx \left[\frac{\mathrm{E}(t_1)}{\mathrm{E}(t_2)}\right]^2 \left[\frac{V(t_1)}{\mathrm{E}^2(t_1)} + \frac{V(t_2)}{\mathrm{E}^2(t_2)} - \frac{2\,\mathrm{Cov}(t_1, t_2)}{\mathrm{E}(t_1)\cdot \mathrm{E}(t_2)}\right], \tag{4.1}$$

where $\mathrm{E}(\cdot)$, $V(\cdot)$, and $\mathrm{Cov}(\cdot,\cdot)$ represent the expectation, variance, and covariance of the respective statistics.

For the analysis of data from a single subpopulation, where only F_{IS} is to be estimated, Curie-Cohen (1982) derived the sampling variance of such estimators. As shown earlier, the estimators by Nei's and Cockerham's approach (equations (3.5) and (3.18)) are related to Curie-Cohen's (1982) estimator $\hat{f}_1 = 1 - (y/x)$, for which he derived a general expression for $\mathrm{Var}(\hat{f}_1)$ at a multi-alleleic codominant locus. His expression (equation (5); Curie-Cohen, 1982, p. 345) can be further reduced to

$$V(\hat{f}_1) = \frac{(1 - F_{IS})\,[(1 - \mu_2) + (1 - F_{IS})\,(1 - \mu_2)^2 - (1 - F_{IS})^2\,(\mu_3 - \mu_2^2)]}{n(1 - \mu_2)^2}, \tag{4.2}$$

where $\mu_2 = \sum_{k=1}^{r} p_{ik}^2$ and $\mu_3 = \sum_{k=1}^{r} p_{ik}^3$, are parameters that depend upon the true allele frequencies at a locus. In practice the estimates of F_{IS}, μ_2, and μ_3 based on sample statistics can be used to estimate $V(\hat{f}_1)$. Using our equation (3.5b), we may immediately note that Nei's unbiased estimator of $F_{IS_{ik}}$ has a sampling variance

$$[1 - 1/(2N_i)]^2\, V(\hat{f}_1) \tag{4.2a}$$

while, Cockerham's estimator (equation (3.18)) has the variance

$$\frac{(1 - F_{IS})\,[(\mu_2 - 2\mu_3 + \mu_2^2) + F_{IS}(1 - 2\mu_2 + 4\mu_3 - 3\mu_2^2) - 2F_{IS}^2(\mu_3 - \mu_2^2)]}{N_i(1 - \mu_2)^2}, \tag{4.2b}$$

in which terms of the order $(2N_i)^{-2}$ or less are neglected.

As mentioned earlier, for large samples F_{IS_i} at a locus, estimated by Long's procedure, is identical to the estimator $\hat{f}_2$, used by Curie-Cohen (1982). Since he derived its sampling variance (equation (7); Curie-Cohen, 1982; p. 346), in our notation for Long's estimator we get

$$V(\hat{F}_{IS}) = \frac{1 - F_{IS}}{2N_i(r-1)^2}\left[2(r-1) - 2(2r-1)F_{IS} + r^2F_{IS}^2 + F_{IS}(2 - F_{IS})\sum_{k=1}^{r} 1/p_{ik}\right]. \quad (4.2c)$$

Equations (4.2), (4.2a), (4.2b), and (4.2c), therefore provide the approximate sampling variance of Nei's biased, Nei's unbiased, Cockerham's and Long's estimator for F_{IS} for a specific subpopulation, for any general multi-allelic codominant locus. When estimators of a specific allele are sought, the equations (4.2), (4.2a), and (4.2b) can be used taking $r = 2$, as shown for a specific case by Curie-Cohen (1982).

Although for a given sample, these sampling variances are to be evaluated with sample estimates of $F_{IS_{ik}}$, μ_2, and μ_3; it is possible to compare the relative efficiencies of Nei's unbiased (equation (3.5)), Nei's biased (equation (3.5a)), Cockerham's estimates (equation (3.18)), and its multivariate extension (equation (3.35)) by contrasting their sampling variances for known parametric values of F_{IS}, μ_2, and μ_3.

Equation (4.2a) suggests that when these parameters are fixed, Nei's unbiased estimator has a smaller sampling variance than the biased estimator. Of course, in reality, when estimates are used in variance evaluation this might not occur in a given set of data (since F_{IS} estimates would differ for these two estimators).

Note that for a bi-allelic locus (with allele frequencies p and q), equations (4.2), (4.2b), and (4.2c) all take the common form

$$N_i V(\hat{F}_{IS}) \approx \frac{1 - F_{IS}}{2pq}[2pq + 2(1 - 3pq)F_{IS} - (p-q)^2F_{IS}^2], \quad (4.3)$$

suggesting that the large-sample standard errors of Nei's biased estimator Cockerham's estimator, and Long's estimator are all identical to the extent that the terms of the order $(1/2N_i)^{-1}$ or less are neglected. Equation (4.3) is also identical to equation (3) of Curie-Cohen (1982).

To our knowledge, this equivalence of the standard errors of the different F_{IS} estimators has not been demonstrated before. In view of this mathematical equivalence, one might wonder why the empirical values of the standard errors of the different estimators vary in the simulation experiments of Weir and Cockerham (1984), Van Den Busche et al. (1986), and Chakraborty and Leimar (1987). Note that the standard error of F_{IS} is dependent on the true value of F_{IS} and the allele

frequencies at a locus (equation (4.3)). Hence, in the computation of the empirical values of the standard errors it is customary to replace the true values of the parameters by their respective estimates (i.e., $\hat{F}_{IS}$ is substituted for F_{IS}). Since we have shown earlier that the estimates differ depending upon the method of estimation satisfying the inequalities (3.24) and (3.25), it is obvious that the same analytical formula for variance (evaluated by equation (4.3)) will give different values when F_{IS} is replaced by its different estimates.

In order to study the empirical differences in the standard errors, it is therefore important to see how expression (4.3) varies as a function of F_{IS}. Curie-Cohen (1982) examined this in his Figure 1 (for a two allelic locus) and Figures 5 and 7 (for two different three allelic loci). His Figure 1 is somewhat confusing, since expression (4.3) does not decrease to zero as F_{IS} approaches its lower limit ($-p/q$ for $q > p$). Substituting $F_{IS} = -p/q$, it reduces to $p(q-p)/2q^4$, which is zero only if $p = q$. In Figure 1, we therefore plotted $\{N_i V(\hat{F}_{IS})\}^{1/2}$ as a function of F_{IS} for four values of p ($p = 0.01$, 0.1, 0.25, and 0.5). It is clear that for $F_{IS} = 0$, $V(\hat{F}_{IS}) = 1/N_i$, irrespective of the allele frequencies at a bi-allelic locus. In general, $V(\hat{F}_{IS})$ is a cubic function of F_{IS}, which attains its maximum at a value of F_{IS} depending upon the allele frequencies at the locus. When the allele frequencies are very skewed (p close to zero or one), the curve rises very fast for negative values of F_{IS}, and similarly drops fast when F_{IS} approaches one. Since Cockerham's estimator (equation (3.18)) is always larger than Nei's biased estimator (equation (3.5a)), unless the true value of F_{IS} is large, substitution of the respective estimates will yield smaller standard error for Nei's biased estimator as compared to that of Cockerham's estimator. The nature of the curves in Figure 1 indicate that such is the case for negative values of F_{IS}, irrespective of the allele frequencies at the locus. In theory, the situation can be reversed for large positive F_{IS}. But, since large positive estimates of F_{IS} are rare in natural populations (unless the organism is highly inbred), this theoretical possibility is not commonly seen. For skewed allele frequencies, the difference in the empirical values of the standard errors can be substantial, because of the sharp rise of the curve. We therefore claim that the observed discrepancies in the standard errors of the various estimators of F_{IS} are the artifacts of substituting the estimates in the variance formula (equation (4.3)). Indeed, there is no inherent difference in the standard errors, as seen in the analytical formulae established here.

Another comment regarding the standard error evaluation of Long's large-sample estimator of F_{IS} (or f_2 of Curie-Cohen, 1982) is worth mentioning at this point. Note that for a multi-allelic locus, this estimator is defined by contrasting the observed proportion of the homozygosity of each allele with the respective allele frequency (equation (3.35)). However, when equation (4.2c) is used to evaluate its standard error $\{V(\hat{F}_{IS})\}^{1/2}$, substituting $\hat{p}_{ik}$ for p_{ik} and $\hat{F}_{IS}$ for F_{IS}, one might encounter negative variance estimators, particularly when one (or more) allele is rare in the population, and $\hat{F}_{IS}$ is negative. In the application section to follow, we have several situations when it occurred. There does not appear to be any simple solution to circumvent this problem of a negative variance estimate. We simply note that the substitution of estimates for parameters (e.g., $\hat{F}_{IS}$ for F_{IS}

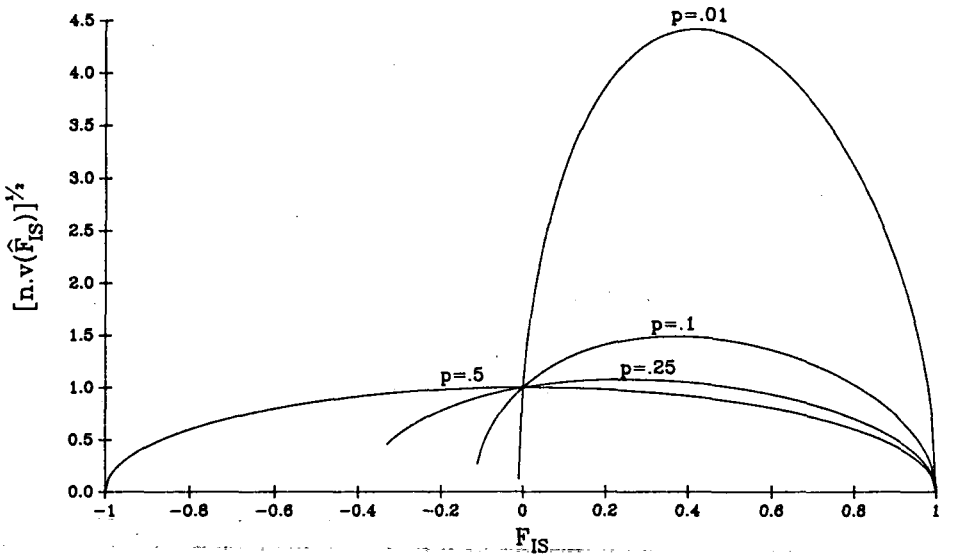


Fig. 1. Relationship between the sampling error of the large-sample estimate of F_{IS} and the true value of parameter (F_{IS}), as studied by plotting $\{n \text{ Var}(\hat{F}_{IS})\}^{1/2}$ versus F_{IS} for a bi-allelic codominant locus with allele frequencies p and q (= 1 − p).

and $\hat{p}_{ik}$ for p_{ik}) in equation (4.2c) yields a poor estimate of $V(\hat{F}_{IS})$ because of the inverse function of $\hat{p}_{ik}$'s (last term of equation (4.2c)).

One solution to this problem, admittedly an ad-hoc one, is to note that when some alleles are rare, since they generally appear in a sample only as heterozygotes, they do not contribute to the estimate of F_{IS} (equation (3.35)). They can be deleted in the variance computation, which is equivalent to computing the $\Sigma\,(1/\hat{p}_{ik})$ term only for alleles that contribute to the estimate of F_{IS}. This avoids the occurrence of a negative variance estimate, as seen in our empirical study. Obviously, more work is needed to provide a justifiable estimator for the standard error of $\hat{F}_{IS}$ in such situations.

The mathematical equivalence of the standard errors shown here apply only for bi-allelic loci. For a general multi-allelic locus such comparisons are more difficult, since the variances also depend on the sum of squares and cubes of allele frequencies (see equations (4.2), (4.2b), and (4.2c)). Nevertheless, for a r-allelic locus with equal gene frequencies (i.e., $p_{ik} = 1/r$ for all k), we have

$$n\,V(\hat{F}_{IS}) = \frac{(1 - F_{IS})\,[1 + (r - 1)F_{IS}]}{r - 1} \tag{4.4}$$

which holds for all of these estimators.

When data from several subpopulations are jointly used for parameter estimation, equation (4.1) can again be used to obtain approximate variances of these

estimators, in which case the variances and covariances reflect the inter-locus variation and covariation of the observed statistics. Chakraborty (1974) was the first to use this idea to evaluate the sampling variance of $\hat{F}_{ST}$, which he represented by

$$V(\hat{F}_{ST}) \approx F_{ST}^2 \left[\frac{V(\hat{H}_S)}{H_S^2} + \frac{V(\hat{H}_T)}{H_T^2} - \frac{2\,\mathrm{Cov}(\hat{H}_S, \hat{H}_T)}{H_S H_T} \right], \tag{4.5}$$

where the variances and covariances of $\hat{H}_S$ and $\hat{H}_T$ [$V(\hat{H}_S)$, $V(\hat{H}_T)$, and $\mathrm{Cov}(\hat{H}_S, \hat{H}_T)$] are obtained from inter-locus variations of these statistics. While Nei and Chakravarti (1977) demonstrated that the equation (4.5) is approximately adequate, Weir and Cockerham (1984) advocated a jackknife procedure in this context (Miller, 1974; Efron, 1982). In principle, if $\hat{\theta}$ represents an estimator of a parameter θ (not to be confused with Cockerham's notation), based on n observations, then the jackknife variance of $\hat{\theta}$ can be expressed as

$$V(\hat{\theta}) \approx \frac{n-1}{n} \sum_{i=1}^{n} \left[\hat{\theta}(i) - \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}(i) \right]^2, \tag{4.6}$$

where $\hat{\theta}(i)$ is the estimator based on $(n-1)$ observations, omitting the i-th observation. If $\hat{\theta}$ involves some bias in estimating θ (as is the case with ratio estimators), a less biased estimator of θ is

$$\hat{\theta}^* = n\hat{\theta} - [(n-1)/n] \sum_{i=1}^{n} \hat{\theta}(i). \tag{4.7}$$

This technique is adopted in estimating the standard errors of the F_{IS}, F_{IT}, F_{ST} estimators of the variance–component approach by Weir and Cockerham (1984), where jackknifing was done over loci (i.e., estimators of a, b, and c components were computed omitting one locus at a time). In particular, when the L-th locus data is omitted, the respective estimators for F_{IS}, F_{IT}, and F_{ST} used are

$$\hat{F}_{IT}(L) = \left[\sum_{l \neq L} \sum_{k} (\hat{a}_{lk} + \hat{b}_{lk}) \right] \Big/ \left[\sum_{l \neq L} \sum_{k} (\hat{a}_{lk} + \hat{b}_{lk} + \hat{c}_{lk}) \right], \tag{4.8}$$

$$\hat{F}_{IS}(L) = \left[\sum_{l \neq L} \sum_{k} \hat{b}_{lk} \right] \Big/ \left[\sum_{l \neq L} \sum_{k} (\hat{b}_{lk} + \hat{c}_{lk}) \right], \tag{4.9}$$

and

$$\hat{F}_{ST}(L) = \left[\sum_{l \neq L} \sum_{k} \hat{a}_{lk} \right] \Big/ \left[\sum_{l \neq L} \sum_{k} (\hat{a}_{lk} + \hat{b}_{lk} + \hat{c}_{lk}) \right]. \tag{4.10}$$

Note that the same approach can be adopted for Nei's estimation procedure as well, where $\hat{H}_S(L)$, $\hat{H}_T(L)$, and $\hat{H}_0(L)$ values are to be evaluated omitting the L-th locus data.

While the jackknifing over loci may provide standard errors of the estimator,

pooled over loci, there has been no explicit formulation for evaluating the sampling errors of individual allele-specific estimators. There is no simple formula for the standard errors of the variance–component estimators for a particular allele, although the sampling theory of categorical analysis of variance (CATANOVA) developed by Light and Margolin (1971), or analysis of diversity (ANODIV) of Rao (1982), indicated in Nayak (1983) may be adopted in this context. Further work is needed to provide computational formulae in this regard.

In principle, under multinomial sampling of genotypes, sampling variances of estimators $\hat{F}_{\mathrm{IT}k}$, $\hat{F}_{\mathrm{IS}k}$, and $\hat{F}_{\mathrm{ST}k}$ (equations (3.6)–(3.8)) can be derived, following Nei and Roychoudhury (1974) and Nei (1978) which refer to the sampling variance computations of heterozygosities and genetic distance. No explicit form of the intra-locus standard errors of the fixation indices are yet available.

Although the utility of the estimators is greatly increased when such standard error evaluation is available, this does not immediately resolve hypothesis testing problems, because with categorical data such ratio estimators do not have simple sampling distributions. Nayak (1983) showed that while exact sampling distributions of the variance components (or sum of squares) are not available, in large samples (N_i's large), the mean square error terms can be represented by linear combinations of χ^2 variables. However, the coefficients of such linear combinations are again not estimable, and hence such theory is difficult to apply in practice.

Cockerham (1973) suggested some heuristic test criteria for specific hypotheses. His test criteria require notations somewhat different from the rest of this paper. In order to avoid confusion, let us introduce for each allele (A_k) at a locus, three genotypes A_kA_k, $A_k\bar{A}_k$, and $\bar{A}_k\bar{A}_k$, where $\bar{A}_k$ is the combination of alleles except the A_k allele. Let M_{ik2}, M_{ik1}, and M_{ik0} be the observed frequencies of these three genotypes in a sample of N_i individuals from the i-th subpopulation. Note that M_{ikl} represents the number of individuals with l copies ($l = 0, 1, 2$) of the A_k allele in the i-th subpopulation, and $M_{ik0} + M_{ik1} + M_{ik2} = N_i$ for $i = 1, 2, \ldots, s$. As before let $N = N_1 + N_2 + \cdots + N_s$, the total number of individuals in the entire survey.

Furthermore, let x_{ik} represent the estimated allele frequency of A_k in the i-th subpopulation, given by our equation (3.2), which is equivalent to

$$x_{ik} = (2M_{ik2} + M_{ik1})/2N_i\,, \tag{3.1a}$$

Under the hypothesis that $F_{\mathrm{ST}} = 0$ and $F_{\mathrm{IS}ik} = 0$ for all i and k, the expectations of M_{ikl}'s are given by Cockerham (1973) as:

$$\bar{\eta}_{ik1} = \mathrm{E}(M_{ik1}) = 2N_i[2N/(2N-1)]\bar{x}_{\cdot k}(1-\bar{x}_{\cdot k})\,, \tag{4.11}$$

where $\bar{x}_{\cdot k} = \sum_{i=1}^{s} N_i x_{ik}/N$, the weighted average frequency of the A_k allele over all subpopulations,

$$\bar{\eta}_{ik2} = \mathrm{E}(M_{ik2}) = N\bar{x}_{\cdot k} - \tfrac{1}{2}\bar{\eta}_{ik1} \tag{4.12}$$

and

$$\bar{\eta}_{ik0} = \mathrm{E}(M_{ik0}) = N(1 - \bar{x}_{\cdot k}) - \tfrac{1}{2}\bar{\eta}_{ik1}\,, \tag{4.13}$$

so that the deviation from $F_{ST} = 0$ and $F_{IS\,ik} = 0$ can be measured by the goodness-of-fit statistic

$$\chi_1^2 = \sum_{i=1}^{s} \sum_{l=0}^{2} (M_{ikl} - \bar{\eta}_{ikl})^2 / \bar{\eta}_{ikl}\,, \tag{4.14}$$

which has a χ^2 distribution with d.f. $2s - 1$ ($2s$ independent genotypes, and one parameter, $\bar{p}_{\cdot k}$ being estimated).

The test-statistic for $F_{IS\,ik} = 0$ for all i and k, given by Cockerham (1973) is the sum-total of $s\chi^2$ values measuring deviations from HWE within individual subpopulations. However, since the unbiased estimator of the $A_k\bar{A}_k$ heterozygote proportions in the i-th subpopulation, is $2N_i x_{ik}(1 - x_{ik})/(2N_i - 1)$, under this hypothesis the expectations of M_{ikl}'s are given by

$$\hat{\eta}_{ik1} = \mathrm{E}(M_{ik1}) = 4N_i^2 x_{ik}(1 - x_{ik})/(2N_i - 1)\,, \tag{4.15}$$

$$\hat{\eta}_{ik2} = \mathrm{E}(M_{ik2}) = N_i x_{ik} - \tfrac{1}{2}\hat{\eta}_{ik1}\,, \tag{4.16}$$

$$\hat{\eta}_{ik0} = \mathrm{E}(M_{ik0}) = N_i(1 - x_{ik}) - \tfrac{1}{2}\hat{\eta}_{ik1}\,. \tag{4.17}$$

Departure from this hypothesis can be tested by the χ^2 statistic

$$\chi_2^2 = \sum_{i=1}^{s} \sum_{l=0}^{2} (M_{ikl} - \hat{\eta}_{ikl})^2 \Big/ \hat{\eta}_{ikl}\,, \tag{4.18}$$

with s d.f.

Cockerham (1973) suggested $\chi_1^2 - \chi_2^2$ as the test criterion with d.f. $s - 1$ for testing the hypothesis $F_{ST} = 0$. While this may approximately hold for large samples, when the A_k-allele is rare in one or more subpopulations, because of small values of M_{ikl}, or $\hat{\eta}_{ikl}$, or $\hat{\eta}_{ikl}$ this approximation may not be accurate. Workman and Niswander (1970) suggested a more direct test for $F_{ST} = 0$ by the usual χ^2 test of heterogeneity (Rao, 1965, p. 323) which is commonly employed in most anthropogenetic studies (see, e.g., Chakraborty et al., 1977).

When $F_{IS\,ik}$ is assumed to be equal in all subpopulations, the test for $F_{IS} = 0$ (common value over all subpopulations) can also be tested with a χ^2 statistic. In this case, the expectations of M_{ikl}'s are computed as

$$\hat{\eta}_1 = \mathrm{E}\left(\sum_{i=1}^{s} M_{ik1}\right) = 4N \sum_{i=1}^{s} x_{ik}(1 - x_{ik})/(2N - s)\,, \tag{4.19}$$

$$\hat{\eta}_2 = \mathrm{E}\left(\sum_{i=1}^{s} M_{ik2}\right) = N\bar{x}_{\cdot k} - \tfrac{1}{2}\hat{\eta}_1\,, \tag{4.20}$$

$$\hat{\eta}_0 = \mathrm{E}\left(\sum_{i=1}^{s} M_{ik0}\right) = N(1 - \bar{x}_{\cdot k}) - \tfrac{1}{2}\hat{\eta}_1\,, \tag{4.21}$$

which yields

$$\chi_5^2 = \sum_{l=0}^{2} \left[\sum_{i=1}^{s} M_{ikl} - \hat{\eta}_l \right]^2 \Big/ \hat{\eta}_l , \tag{4.22}$$

which also has a χ^2 distribution with one d.f.

In a similar vein, Cockerham (1973) suggested a test criterion for $F_{STk} = 0$ from allele frequency data, which takes the form

$$\chi_k^2 = \left[\sum_{i=1}^{s} 2N_i [x_{ik} - \bar{x}_{\cdot k}(\hat{w})]^2 \right] \Big/ \{ \bar{x}_{\cdot k}(\hat{w}) [1 - \bar{x}_{\cdot k}(\hat{w})] \} , \tag{4.23}$$

with $(s - 1)$ d.f., for each specific allele A_k. Although for genotypic data, several alternative test criteria for F_{ST} exist, there is no definitive theory that suggests which should be the preferred one. We might note that expression (4.23) is the most commonly employed test criterion for F_{STk} in empirical studies of population structure (see also Workman and Niswander, 1970).

Although the test criteria (4.13), (4.18), (4.22) and (4.23) have their own intuitive appeal, Rao (1982) and Nayak (1983) showed that when the N_i's are not equal, these χ^2 statistics do not quite reflect an orthogonal decomposition of the total sum of squares in terms of a categorical analysis of variance. Further investigation is needed to address the question of most powerful test criteria in the analysis of such data. Furthermore, since these statistics refer to a single allele (A_k), a combined analysis for multiple allelic loci is not provided by these test criteria.

Long (1986) approached this problem while providing locus-specific estimates of the fixation indices. As shown earlier, Long's (1986) estimators are derived in terms of the three MSCP matrices: MSCP(a), MSCP(b), and MSCP(c), respectively (in Long's notation MSCP(c) = MSCP(W)). He suggested that the significance of F_{IS}, F_{IT} can be tested by

$$\Lambda_1^* = \det[\text{MSCP}(c)]/\det[\text{MSCP}(b)] \approx \Lambda(G, N - s, N) , \tag{4.24}$$

$$\Lambda_2^* = \det[\text{MSCP}(b)]/\det[\text{MSCP}(a)] \approx \Lambda(G, s - 1, N - s) , \tag{4.25}$$

and

$$\Lambda_3^* = \det[(N - 2)\,\text{MSCP}(b) + 2\text{MSCP}(a)]/\det[N\,\text{MSCP}(c)] \approx \Lambda(G, N - 1, N) , \tag{4.26}$$

respectively, where $\det(Z)$ is the determinant of a matrix Z, G is the dimension of S-matrices (number of independent alleles); and $\Lambda(\text{df}_1, \text{df}_2, \text{df}_3)$ is a Wilk's Λ variate with d.f. df_1, df_2, and df_3 (Anderson, 1984, p. 299).

Although the rationale of these test criteria results from the convergence of the multinomial to the multivariate normal distribution for fairly large sample sizes, there are several problems with these test statistics. First, for unequal sample sizes S_a, S_b, and S_c are not independently distributed (and so too are their respective

MSCP matrices). Nayak (1983) showed that their correlations can be quite substantial, and hence, the criteria Λ_1^*, Λ_2^*, and Λ_3^* do not satisfy the conditions under which Wilk's Λ distribution is valid (see equation (3) of Anderson, 1984, p. 299). Second, the assumption that a MSCP matrix follows a Wishart distribution is true for multivariate normal variates. A multinomial sampling of genotypes where one or more alleles are rare in the population and consequently may be absent in one or more subpopulations, will not approach multivariate normality unless the sample sizes are very large. Third, Wilk's Λ distribution approximation will also require a large number of subpopulations in addition to large N_i values. Since, in the earlier sections we showed that a great deal of work is needed to reduce bias due to small N_i and s values in estimating the fixation indices, the attempt to sweep out all these troubles by using such approximations cannot be generally advocated. Fourth and lastly, as indicated earlier, the variance–component approach may yield a negative-definite MSCP(B) matrix (see Cockerham, 1969, p. 74 for the univariate result), which also makes the Λ-distribution approximation invalid.

In summary, we argue that a rigorous test procedure for studying the significance of the fixation indices is not yet available. All suggested test criteria are only approximate, and caution must be exercised in interpreting their results.

5. An application

Bhasin et al. (1986) studied the genetic structure of the people of Sikkim of North India in order to determine the extent of genetic differentiation among the various subdivisions of their social units. They recognized 13 social groups in this population: North Sikkim, Sherpas, Tamangs, Gurungs, Rais, Limboos, Pradhans, Brahmins, Chhetris, and Scheduled Castes who are ethnohistorically as well as socially isolated to a certain extent. They studied 17 polymorphic blood groups and protein loci in each of these subpopulations. Of these loci, 11 are codominant. Haptoglobin (Hp), Group-Specific Component (Gc), Transferrin (Tf), Acid phosphatase (aP), Phosphoglucomutase-1 (PGM_1), 6-phosphogluconate dehydrogenase (6-PGD), Esterase D (EsD), Adenylate kinase (Ak), Hemoglobin (Hb), Duffy (Fy), and Kidd (Ik), at which the number of detected alleles vary from 2 to 5 (Gc and aP have 3 alleles each, Tf has 5 alleles, all of the remaining having 2 alleles each). The remaining six loci—AB0, MNSs, Rh, Kell and Immunoglobulin Gm and Km—have variable degrees of complex dominant relationships among their alleles/haplotypes, so that at each of these loci not all genotypes are distinguishable. The genotype/phenotype data at these loci for each subpopulation are presented in Bhasin et al. (1986).

We consider this survey for illustrating empirically the differences in the various estimators for several reasons. First, as part of a larger study of the extent of genetic differentiation among the populations of Sikkim, the estimates of the fixation indices provide the basis of our further studies, and hence it is important to determine their stability over different estimation methods employed. Second,

as the sample sizes of this study drastically differ over subpopulations as well as over loci, this example should also provide insight regarding the stability of parameter estimates with or without invoking the large sample approximations discussed in our theoretical exposition. Finally, the availability of loci with and without dominance relationships in this study will help us to examine some features of the statistical properties of the parameter estimates based on genotype vs. allele frequency data, not commonly found in all applications of this nature. Notwithstanding these issues, we should note that since this review deals with a comparative study of the various estimation procedures, and *not* the population structure of the Sikkimese people, only the results pertaining to the comparative analyses are reported here.

5.1. *Comparison of the estimates of F_{IS} in a single subpopulation*

We have seen earlier that estimation of F_{IS} is possible from genotype data by several methods. Since only codominant loci can be used for this purpose, Table 1 contrasts the estimates of F_{IS} in the Lepchas of North Sikkim for 11 loci, as computed using equations (3.5) (Nei's unbiased estimator), (3.5a) (Nei's biased estimator) and (3.18) (Cockerham's estimator). Although several other estimators of F_{IS} are available in such situations (Li and Horvitz, 1953; Curie-Cohen, 1982; Robertson and Hill, 1984), we contend that the data presented in Table 1 are sufficient to contrast most of the theoretically justifiable estimators. Note that Curie-Cohen's (1982) estimator $\hat{f}_1$ is identical to Nei's biased estimator, and his $\hat{f}_2$ is exactly the same as the multivariate large sample estimate at a locus, while his $\hat{f}_3$ estimator can be computed from the χ^2 values presented in this table. We computed two different χ^2 values, one corresponding to Nei's unbiased estimator (evaluated by equation (4.18) with $s = 1$), and the other corresponding to the biased estimator (where the expected genotype frequencies are computed by $N_i x_{ik}^2$ for the genotype $A_k A_k$ and $2N_i x_{ik} x_{il}$ for the genotype $A_k A_l$). The χ^2 values for Cockerham's estimator are identical to those of Nei's unbiased estimator, and hence they are not repeated in the table. The allele-specific χ^2's have a single d.f. in every case, while the locus-specific χ^2's have d.f. $r(r-1)/2$, where r is the number of segregating alleles at a locus in the specific subpopulation. Since for the bi-allelic loci the estimators and χ^2 values are identical for both alleles, and hence their locus-specific values are exactly the same as those based on any specific allele, only allele-specific estimates are given for such loci (Hp, PGM_1, PGD, EsD, Ak, Hb, Duffy, and Kidd in our example). Further note that although the transferrin locus has 5 segregating alleles in the total Sikkim population, in Lepchas of North Sikkim only 3 segregating alleles were found (Bhasin et al., 1986), and hence this was treated as a 3-allelic locus for this computation.

We chose not to present the estimator of F_{IS} based on χ^2 values for two reasons. First, since χ^2 values represent deviation from HWE for the two-sided alternative $F_{IS} \neq 0$, the sign of F_{IS} cannot be inferred from the value of χ^2 (Li and Horvitz, 1953; Curie-Cohen, 1982), and second, the χ^2 values can be greatly

Table 1
Allele-specific estimates of $F_{IS/k}$ for the Lepchas sampled in North Sikkim

Locus	Allele	N	Frequency	Nei's unbiased estimate			Nei's biased estimate			Cockerham's estimate	Long's estimate	
				$F_{IS/k} \pm$ s.e.	χ^2	d.f.	$F_{IS/k} \pm$ s.e.	χ^2	d.f.	$F_{IS/k} \pm$ s.e.[a]	Weighted[b]	Large sample[c]
Hp	Hp^1	65	0.1154	0.177 ± 0.162	2.18	1	0.171 ± 0.163	1.90	1	0.179 ± 0.164	0.179 ± 0.127	0.171 ± 0.163
Gc	Gc^{1F}		0.6129	0.393 ± 0.119	9.75[f]	1	0.388 ± 0.120	9.34[f]	1	0.395 ± 0.120		
	Gc^{1S}		0.2581	0.415 ± 0.130	10.97[g]	1	0.410 ± 0.131	10.44[g]	1	0.417 ± 0.131		
	Gc^2		0.1290	0.004 ± 0.127	0.00	1	−0.005 ± 0.126	0.00	1	0.004 ± 0.128		
	Pooled	33		0.320 ± 0.106	13.70[f]	3	0.314 ± 0.107	13.44[f]	3	0.322 ± 0.107	0.233	0.225 ± 0.108
Tf	Tf^{C1}		0.7097	−0.087 ± 0.120	0.48	1	−0.096 ± 0.120	0.57	1	−0.088 ± 0.121		
	Tf^{C2}		0.2823	−0.066 ± 0.121	0.28	1	−0.075 ± 0.122	0.35	1	−0.067 ± 0.122		
	Tf^{D}		0.0081	0.000 ± 0.126	0.00	1	−0.008 ± 0.008	0.00	1	0.000 ± 0.127		
	Pooled	62		−0.75 ± 0.117	0.74	3	−0.084 ± 0.117	0.82	3	−0.076 ± 0.118	−0.037	−0.047 ± 0.124[d]
AP	P^a		0.1638	−0.061 ± 0.115	0.22	1	−0.070 ± 0.113	0.28	1	−0.061 ± 0.116		
	P^b		0.8190	0.020 ± 0.134	0.02	1	0.012 ± 0.133	0.01	1	0.020 ± 0.135		
	P^c		0.0172	−0.009 ± 0.093	0.01	1	−0.018 ± 0.012	0.02	1	−0.009 ± 0.094		
	Pooled	58		−0.018 ± 0.116	1.79	3	−0.027 ± 0.115	1.86	3	−0.018 ± 0.117	−0.028	−0.037 ± 0.122[d]
PGM_1	PGM_1^1	50	0.6600	−0.324 ± 0.115	5.37[e]	1	−0.337 ± 0.114	5.68[e]	1	−0.328 ± 0.115	−0.328 ± 0.115	−0.337 ± 0.114
PGD	PGD^A	53	0.8679	0.022 ± 0.143	0.03	1	0.012 ± 0.141	0.01	1	0.022 ± 0.144	0.022 ± 0.144	0.012 ± 0.141
EsD	EsD^1	50	0.7400	0.485 ± 0.139	12.17[g]	1	0.480 ± 0.141	11.53[g]	1	0.488 ± 0.141	0.488 ± 0.141	0.480 ± 0.141
AK	AK^1	58	0.9914	0.000 ± 0.130	0.00	1	−0.009 ± 0.009	0.00	1	0.000 ± 0.131	0.000 ± 0.131	−0.009 ± 0.009
Hb	Hb^A	61	0.9754	−0.017 ± 0.075	0.03	1	−0.025 ± 0.015	0.04	1	−0.017 ± 0.075	−0.017 ± 0.075	−0.025 ± 0.015
Duffy	Fy^a	66	0.8485	0.181 ± 0.149	2.27	1	0.175 ± 0.149	2.02	1	0.182 ± 0.150	0.182 ± 0.150	0.175 ± 0.149
Kidd	Ik^a	47	0.4681	−0.268 ± 0.139	3.45	1	−0.282 ± 0.139	3.73	1	−0.272 ± 0.140	−0.272 ± 0.140	−0.282 ± 0.139

[a] The χ^2 values for Cockerham's estimate of $F_{IS/k}$ are exactly the same as that for Nei's unbiased estimates.
[b] Long's weighted estimates are locus-specific estimates, which are identical to Cockerham's estimator for two-allelic loci.
[c] Long's large sample estimator is identical to that of Curie-Cohen's (1982) estimator f_2, and hence their standard errors are also the same (see text for details).
[d] These s.e. values are computed deleting the alleles that do not contribute to the estimate (see text for details).
[e] $p < 0.05$.
[f] $p < 0.01$.
[g] $p < 0.001$.

affected by rare genotypes and their expected values, giving unstable estimates in specific situations, an example of which will be discussed later in this section.

The standard errors of the three estimators are evaluated by equation (4.2) (for Nei's biased estimator), (4.2a) (for Nei's unbiased estimator) and (4.2b) (for Cockerham's estimator), where $r = 2$ is used for allele specific values, and the entire locus data used for locus-specific standard errors (represented by s.e. in the table).

In terms of the values of the estimates, it is clear that Nei's unbiased, biased, and Cockerham's estimates of F_{IS} are quite close to each other, with biased estimates always being the smallest. The differences of these estimates (the allele-specific ones as well as their pooled values over all alleles at a locus) are always encompassed by their respective standard errors (see Table 1).

The standard errors of Nei's unbiased and Cockerham's estimates are also very similar, while for negative $\hat{F}_{IS_{ik}}$ (or $\hat{F}_{IS_i}$) values Cockerham's estimators have slightly larger standard errors, the situation reverses when the estimates are positive. The differences in the standard errors are however very small, and in no case change the qualitative results of hypothesis testing ($F_{IS} = 0$) either by the χ^2 value shown in the table, or by a crude test of the normal deviate $[\hat{F}_{IS_{ik}}/\text{s.e.}(\hat{F}_{IS_{ik}})]$–the latter test not explictly shown in this table. As noted earlier, for the bi-allelic loci the differences in the standard errors are produced only because of substituting the respective estimates of F_{IS} in equation (4.3). Curie-Cohen (1982) also showed that in multiallelic loci the standard errors of the various estimators are only slightly different (see Figure 5 of Curie–Cohen, 1982, p. 352).

A comment regarding the standard errors of Nei's biased estimators is worth noting. While these are quite close to those of Nei's unbiased and Cockerham's estimators, where the allele frequency is close to 0 or 1 (e.g., Tf^D, p^C, Ak^1, and Hb^A) the s.e.'s of the biased estimators are substantially smaller than those of Nei's unbiased and Cockerham's estimators. This feature may not be intuitively clear. Nevertheless, Figure 1 indicates that for skewed allele frequencies (p, small), the s.e. of Nei's biased estimate sharply rises from a very small value in the range of negative F_{IS} values. Since we evaluated the s.e. of each estimate by substituting the obtained F_{IS} estimates of the same method, these computations are indeed a comparison of different trajectories. For example, in the case of the Tf^D allele at the Transferrin locus, the standard error of Nei's unbiased estimator is evaluated at $F_{IS} = -0.008$, while that of Nei's unbiased and Cockerham's estimates is evaluated for $F_{IS} = 0.0$. The frequency of this allele in the Lepcha subpopulation is 0.008. For this allele frequency, even for the biased estimator of Nei, the s.e. rises from 0.008 to 0.127 as F_{IS} is changed from -0.008 to 0.0. Therefore, the differences in the standard errors noticed in Table 1 are largely due to the fact that the estimates are somewhat different in these three methods. When allele frequencies are at an intermediate range, small differences between parameter estimates do not substantially change the standard errors, but for skewed allele frequencies even minute changes in the estimates can induce a large difference in standard errors, particularly when the F_{IS} estimate is negative. In spite of such

differences, there is no change in the conclusion regarding hypothesis testing either from the χ^2 values or from normal deviates. Even though we do not present similar analyses for the other 12 subpopulations studied by Bhasin et al. (1986), this statement is valid in general.

Table 1 further shows that of the 17 allele-specific tests performed, significant deviation from $F_{IS} = 0$ is found in 4 cases, due to 3 loci (Gc, PGM_1, and EsD). One of these significant deviations is due to a negative F_{IS} (at the PGM_1 locus). This finding is consistent for all three methods employed in this analysis. We also have evidence of negative significant F_{IS} values for the Tf^{C1} allele in Tamangs and Scheduled Castes, Tf^{C2} allele in Rais, Gurungs and Scheduled Castes and for the Kidd locus (either allele) in Rais and Gurungs.

Table 2 presents a summary of the significant (at 5% level) positive and negative $\hat{F}_{IS/k}$ (or $\hat{F}_{IS/}$) values in 158 independent allele-specific and 127 locus-specific tests in the total data on 11 loci in the 13 subpopulations mentioned earlier. For allele-specific tests five test procedures are considered in this table: 2 χ^2 tests (one based on biased estimates of genotype frequencies, and the other based on unbiased estimates), and 3 normal deviates (based on Nei's biased and unbiased estimators, and that of Cockerham). For locus-specific tests, in addition to the above five test procedures, normal deviates based on Long's large-sample estimates of $F_{IS/}$ (which is identical to the estimator $\hat{f}_2$ of Curie-Cohen, 1982) are also used, since the standard error of such estimators is known (see equation (4.2c)).

The total number of significant deviations from $F_{IS} = 0$ is almost the same for each χ^2 statistic. The normal deviates based on Nei's unbiased and Cockerham's estimators also behave identically, as do the normal deviates based on Nei's biased and Long's large-sample estimates in the case of locus-specific tests. Furthermore, the numbers of positive and negative significant F_{IS} values according to the χ^2 statistics are not equal; there are far more positive significant values than negative ones.

Table 2
Number of significant ($p < 0.05$) F_{IS} values in the Sikkim survey as detected by various estimators

Test criterion	Allele-specific tests with $F_{IS/k}$ value[a]		Locus-specific tests with $F_{IS/}$ value[b]	
	Positive	Negative	Poitive	Negative
χ^2: Unbiased	22	7	12	5
Biased	20	10	10	7
Normal Deviate based on				
Nei's unbiased estimate	16	17	8	14
Nei's biased estimate	17	29	18	27
Cockerham's estimate	16	18	9	13
Long's large sample est.	–	–	18	25

[a] Total number of independent allele-specific tests = 158.
[b] Total number of independent locus-specific tests = 127.

These features are not unique to this data alone, and can be explained on the basis of the theory we presented before. First, note that χ^2 statistics only detect deviation in either direction, and since the range of negative F_{IS} is narrower ($F_{IS_{ik}} \geq -p_{ik}/(1-p_{ik})$ for every allele A_k) than the range of positive F_{IS} ($F_{IS} \leq 1$), it is expected that more significant positive F_{IS} values will be encountered based on χ^2 goodness-of-fit. Second, since Nei's biased estimator has empirically smaller sampling variance than that of Cockerham's estimator for negative F_{IS} (Figure 1) for all allele frequencies (unless the alleles are equi-frequent), it is expected that this will pick up more significant negative $\hat{F}_{IS}$ values than the normal deviates based on Cockerham's estimator. This is also predicted from Figure 1, which shows that the sampling variance sharply drops off even if the $\hat{F}_{IS_{ik}}$ values are slightly decreased, particularly when $\hat{F}_{IS_{ik}}$ is negative. Since in all cases Nei's biased estimate is smaller than all other estimates, a normal deviate based on this estimator would necessarily pick up more significant negative F_{IS} values as compared to any other test criteria.

The estimate of F_{IS} for a single subpopulation, combining all alleles at a locus shows exactly the same picture. We have not explicitly shown the behavior of the test criteria based on Long's weighted estimator for the reason that its sampling variance is not yet available. However, its large-sample variance can be computed based on (4.2c), which is used in the computations shown in this table.

5.2. *Comparison of F_{IT}, F_{IS}, and F_{ST} estimates over all subpopulations*

In Tables 3, 4, and 5 we provide a comparative study of the estimators of the three fixation indices over all 13 subpopulations of Sikkim. Nei's weighted, unweighted, and large sample estimates are computed by equations (3.6)–(3.8), (3.6a)–(3.8a), and (3.6b)–(3.8b), respectively. While the standard errors of these estimators for allele- and locus-specific cases cannot be evaluated, approximate tests for allele-specific F_{IS} and F_{ST} values may be conducted by χ^2 statistics, according to equations (4.22), and (4.23), respectively. These results are shown in Table 3. In Table 4 computations of Cockerham's estimators are shown for the same data. In addition to the weighted (equations (3.15a), (3.16a), and (3.17a)) and large sample estimates (equations (3.15b), (3.16b), and (3.17b)), Cockerham's estimators of F_{ST} are also obtained under the assumption of $F_{IS} = 0$ (equation (3.43)) whose values and χ^2 test criteria are shown in this table. It should be noted that the χ^2 test criteria for F_{IS} and F_{ST} for Cockerham's general estimates are exactly the same as those of Nei's weighted estimators (shown in Table 3), and hence they are not repeated in Table 4. Multivariate estimators, according to the generalized formulation of variance component analysis (equations (2.18a)–(2.20a)), are presented in Table 5. Only locus-specific estimates are needed here, and for two-allelic loci these estimates are identical to those of Cockerham's weighted analysis, as shown earlier.

Several interesting findings emerge from these computations. First, the weighted estimators of each fixation index are nearly the same for Cockerham's and Nei's method. Second, while the F_{ST} estimates of Nei's weighted and unweighted

Table 3
Nei's allele- and locus-specific F_{IT}, F_{IS}, and F_{ST} estimates

Locus	Allele	Weighted							Unweighted			Large N		
		F_{IT}	F_{IS}	χ^2	d.f.	F_{ST}	χ^2	d.f.	F_{IT}	F_{IS}	F_{ST}	F_{IT}	F_{IS}	F_{ST}
Hp	Hp^1	−0.007	−0.021	0.25	1	0.013	23.96[a]	12	−0.013	−0.029	0.016	−0.014	−0.042	0.027
Gc	Gc^{1F}	0.278	0.214	23.48[c]	1	0.081	133.39[c]	12	0.261	0.190	0.087	0.260	0.181	0.097
	Gc^{1S}	0.238	0.192	19.76[c]	1	0.057	92.08[c]	12	0.207	0.161	0.055	0.206	0.151	0.066
	Gc^2	0.188	0.162	14.08[c]	1	0.031	43.86[c]	12	0.193	0.166	0.032	0.192	0.156	0.043
	Pooled	0.240	0.192			0.059			0.224	0.173	0.061	0.223	0.163	0.072
Tf	Tf^{C1}	−0.151	−0.163	14.62[c]	1	0.010	26.20[a]	12	−0.162	−0.177	0.013	−0.163	−0.193	0.025
	Tf^{C2}	−0.166	−0.176	17.20[c]	1	0.009	22.72[a]	12	−0.179	−0.192	0.011	−0.180	−0.209	0.024
	Tf^{C3}	−0.009	−0.017	0.11	1	0.008	24.78[a]	12	−0.011	−0.018	0.006	−0.013	−0.034	0.021
	Tf^{C12}	−0.003	−0.009	0.03	1	0.006	23.39[a]	12	−0.004	−0.010	0.006	−0.005	−0.027	0.022
	Tf^{D}	−0.003	−0.003	0.00	1	−0.000	7.66	12	−0.002	−0.002	−0.000	−0.002	−0.011	0.008
	Pooled	−0.152	−0.163			0.009			−0.163	−0.177	0.012	−0.164	−0.193	0.024
aP	P^A	0.080	0.067	2.28	1	0.013	29.55[b]	12	0.055	0.039	0.016	0.054	0.026	0.029
	P^B	0.088	0.077	2.98	1	0.012	28.81[b]	12	0.065	0.051	0.014	0.064	0.038	0.027
	P^C	0.193	0.187	9.40[b]	1	0.008	458.95[c]	12	0.253	0.246	0.010	0.252	0.234	0.024
	Pooled	0.087	0.075			0.013			0.066	0.052	0.015	0.065	0.039	0.028
PGM_1	PGM_1^1	0.025	0.025	0.27	1	−0.000	11.98	12	0.070	0.074	−0.005	0.068	0.057	0.013
PGD	PGD^A	0.150	0.127	6.09[a]	1	0.26	30.95[b]	12	0.176	0.147	0.034	0.175	0.133	0.048
EsD	EsD^1	0.145	0.143	11.19[c]	1	0.003	13.78	12	0.134	0.132	0.003	0.134	0.121	0.015
Ak	Ak^1	0.097	0.067	1.24	1	0.032	188.82[c]	12	0.126	0.100	0.029	0.125	0.088	0.041
Hb	Hb^A	−0.007	−0.016	0.08	1	0.008	22.21[a]	12	−0.007	−0.019	0.011	−0.008	−0.031	0.022
Duffy	Fy^a	0.152	0.122	3.52	1	0.034	46.81[c]	12	0.180	0.142	0.045	0.179	0.114	0.072
Kidd	Ik^a	−0.204	−0.210	11.30[c]	1	0.005	12.67	12	−0.210	−0.218	0.007	−0.213	−0.259	0.037

[a] $p < 0.05$.
[b] $p < 0.01$.
[c] $p < 0.001$.

Table 4
Cockerham's allele- and locus-specific F_{IT}, F_{IS}, and F_{ST} estimates

Locus	Allele	Weighted[a]			Under $F_{IS} = 0$			Large N		
		F_{IT}	F_{IS}	F_{ST}	F_{ST}	χ^2	d.f.	F_{IT}	F_{IS}	F_{ST}
Hp	Hp^1	−0.006	−0.021	0.015	0.014	27.72[c]	12	−0.006	−0.032	0.025
Gc	Gc^{1F}	0.284	0.216	0.086	0.088	108.84[d]	12	0.284	0.206	0.098
	Gc^{1S}	0.242	0.193	0.061	0.063	80.96[d]	12	0.242	0.183	0.073
	Gc^2	0.190	0.163	0.032	0.034	49.15[d]	12	0.191	0.153	0.044
	Pooled	0.245	0.194	0.063	0.065			0.245	0.184	0.075
Tf	Tf^{C1}	−0.150	−0.165	0.014	0.012	23.71[b]	12	−0.149	−0.177	0.023
	Tf^{C2}	−0.165	−0.179	0.012	0.010	22.18[b]	12	−0.164	−0.190	0.022
	Tf^{C3}	−0.008	−0.021	0.012	0.012	24.37[b]	12	−0.008	−0.032	0.024
	Tf^{C12}	−0.002	−0.013	0.011	0.011	23.26[b]	12	−0.001	−0.025	0.023
	Tf^{D}	−0.003	0.001	−0.004	−0.004	7.62	12	−0.003	−0.010	0.007
	Pooled	−0.151	−0.166	0.013	0.011			−0.151	−0.177	0.022
aP	P^A	0.081	0.068	0.014	0.015	26.23[c]	12	0.081	0.056	0.027
	p^B	0.089	0.077	0.013	0.014	25.08[b]	12	0.089	0.065	0.026
	p^C	0.194	0.187	0.009	0.011	22.41[b]	12	0.194	0.175	0.023
	Pooled	0.088	0.076	0.013	0.014			0.088	0.064	0.026
PGM_1	PGM_1^1	0.025	0.025	−0.000	0.000	12.30	12	0.025	0.010	0.015
PGD	PGD^A	0.152	0.131	0.025	0.027	32.19[c]	12	0.152	0.115	0.042
EsD	EsD^1	0.145	0.144	0.002	0.003	15.38	12	0.146	0.132	0.015
Ak	Ak^1	0.100	0.067	0.036	0.036	51.88[d]	12	0.101	0.056	0.047
Hb	Hb^A	−0.007	−0.016	0.009	0.009	21.91[b]	12	−0.006	−0.027	0.020
Duffy	Fy^a	0.156	0.126	0.035	0.038	29.76[c]	12	0.157	0.101	0.062
Kidd	Ik^a	−0.203	−0.216	0.011	0.005	14.41	12	−0.202	−0.241	0.031

[a] The χ^2 for Cockerham's weighted estimates of F_{IS} and F_{ST} are exactly the same as those for Nei's weighted estimates (see Table 2).
[b] $p < 0.05$.
[c] $p < 0.01$.
[d] $p < 0.01$.

analyses are almost identical, there are some differences in the corresponding F_{IT} and F_{IS} estimates. Third, the large sample F_{ST} estimates of Nei and Cockerham are almost identical, even though these two methods yield somewhat different F_{IT} and F_{IS} values in large samples. Fourth, even when the estimate of allele- and locus-specific F_{IS} is significantly different from zero (tested by the χ^2 values), Cockerham's special case estimate of F_{ST} (equation (3.43) under $F_{ST} = 0$) is almost identical to that of his weighted analysis (Table 4). Fifth, while the large sample values of F_{IT} are very similar to those based on weighted analysis (in Nei's as well as Cockerham's approaches), the F_{ST} values do not behave similarly. Indeed, F_{ST} values are generally larger when large sample approximations are

Table 5
Locus-specific estimates of F_{IT}, F_{IS}, and F_{ST} by the multivariate technique of Long (1986)

Locus	Weighted estimators			Large sample estimators		
	F_{IT}	F_{IS}	F_{ST}	F_{IT}	F_{IS}	F_{ST}
Hp	−0.006	−0.021	0.015	−0.006	0.032	0.025
Gc	0.238	0.185	0.065	0.233	0.164	0.083
Tf	−0.156	−0.182	0.022	−0.167	−0.204	0.031
AP	0.081	0.063	0.019	0.073	0.045	0.029
PGM_1	0.025	0.025	0.000	0.025	0.010	0.015
PGD	0.152	0.131	0.025	0.152	0.115	0.042
EsD	0.145	0.144	0.002	0.146	0.132	0.015
AK	0.100	0.067	0.036	0.101	0.056	0.047
Hb	−0.007	−0.016	0.009	−0.006	−0.027	0.020
Duffy	0.156	0.126	0.035	0.157	0.101	0.062
Kidd	−0.203	−0.216	0.011	−0.202	−0.241	0.031

made. Because of equation (2.1), the F_{IS} should be under-estimated in large sample approximations (since F_{IT} does not change substantially). This is the case for every comparison of Cockerham's estimators, while there are some minor discrepancies in Nei's approach. These differences are due to changes in F_{IT} values in large sample vs. weighted analysis. Sixth, the multivariate estimators are the most deviant ones. There is no general trend of these estimators as compared to Nei's and Cockerham's estimators. This is also theoretically justifiable, since the wieghting scheme in the multivariate approach is quite different (equations (2.18a)–(2.20a)).

5.3. *Comparison of the estimates pooled over loci*

Table 6 presents the estimates and their standard errors pooled over all co-dominant loci. As mentioned before, pooling over loci can be done in two ways for every method of estimation:

(1) by taking the ratio of sums of locus-specific estimates, and

(2) by the technique of jackknifing.

Since each fixation index is described as a function of ratios of parameters (population allele frequencies and their inter-locus variances across subpopulations), Weir and Cockerham (1984) advocated the jackknifing procedure suggesting that this might reduce the bias of estimation and in turn make the standard errors more reliable (Miller, 1974; Efron, 1982). We, however, do not see any substantial change in the estimates as well as in their standard errors through jackknifing. In fact, there is a tendency for the jackknife estimators to have somewhat larger s.e.'s for each fixation index. This table also shows that while Cockerham's and Nei's estimators are virtually identical (weighted as well as large sample), the large sample approximations involve over-estimation of F_{ST} and under-estimation of F_{IS}, F_{IT} remaining very similar. The small difference of

Table 6
Estimates of F_{IT}, F_{IS}, and F_{ST} pooled over loci and their standard errors by the three different methods

	Ratio of sums			Jackknife		
	F_{IT}	F_{IS}	F_{ST}	F_{IT}	F_{IS}	F_{ST}
Nei's estimates						
Weighted	0.045 ± 0.060	0.025 ± 0.055	0.020 ± 0.009	0.045 ± 0.064	0.025 ± 0.058	0.020 ± 0.009
Unweighted	0.044 ± 0.060	0.023 ± 0.055	0.022 ± 0.009	0.044 ± 0.063	0.023 ± 0.058	0.022 ± 0.010
Large N	0.043 ± 0.060	0.005 ± 0.058	0.038 ± 0.008	0.043 ± 0.063	0.005 ± 0.061	0.038 ± 0.009
Cockerham's estimates						
Weighted	0.047 ± 0.060	0.025 ± 0.056	0.022 ± 0.009	0.047 ± 0.064	0.025 ± 0.059	0.022 ± 0.010
Under $F_{IS} = 0$			0.022 ± 0.009			0.022 ± 0.010
Large N	0.047 ± 0.060	0.011 ± 0.057	0.037 ± 0.009	0.048 ± 0.064	0.011 ± 0.060	0.037 ± 0.009
Long's estimates						
Weighted	0.048 ± 0.063	0.028 ± 0.059	0.021 ± 0.011	0.048 ± 0.064	0.029 ± 0.060	0.023 ± 0.011
Large N	0.046 ± 0.062	0.011 ± 0.058	0.036 ± 0.010	0.047 ± 0.063	0.010 ± 0.059	0.038 ± 0.009

Long's estimators as compared with others is mainly produced by the difference of the pooling algorithm in his procedure, as noted earlier. However, unless a survey has a large number of multi-allelic loci, this method is likely to produce an almost identical qualitative conclusion about the genetic structure of the population, as seen in this example.

5.4. Comparison of the estimates of F_{ST} from allele frequency data

As mentioned earlier, analysis of population structure is sometimes necessary from allele frequency data alone. This occurs when either the loci involves dominance relationships among their alleles, or the allele frequency data are collected from the literature for comparative studies. In such cases, the only estimable parameter is F_{ST}. It is shown earlier, that in Nei's gene diversity approach, the estimators (weighted or unweighted) remain the same even if allele frequencies are used in estimation instead of genotype data (see equations (3.8) and (3.8a)). Cockerham's estimator of F_{ST} takes the form of equation (3.43), whose multivariate extension (Long's approach) is obvious from equation (3.41). The variance–component approach (univariate or multivariate) of estimation of F_{ST} from allele frequency data is therefore mathematically equivalent to the estimation of the same parameter from genotype data with the additional assumption that $F_{IS} = 0$. In order to compare the empirical values of these estimators from allele frequency data, we computed Nei's weighted unbiased, Cockerham's, and Long's estimates of F_{ST} for all 17 loci studied by Bhasin et al. (1986). The allele frequencies used in these

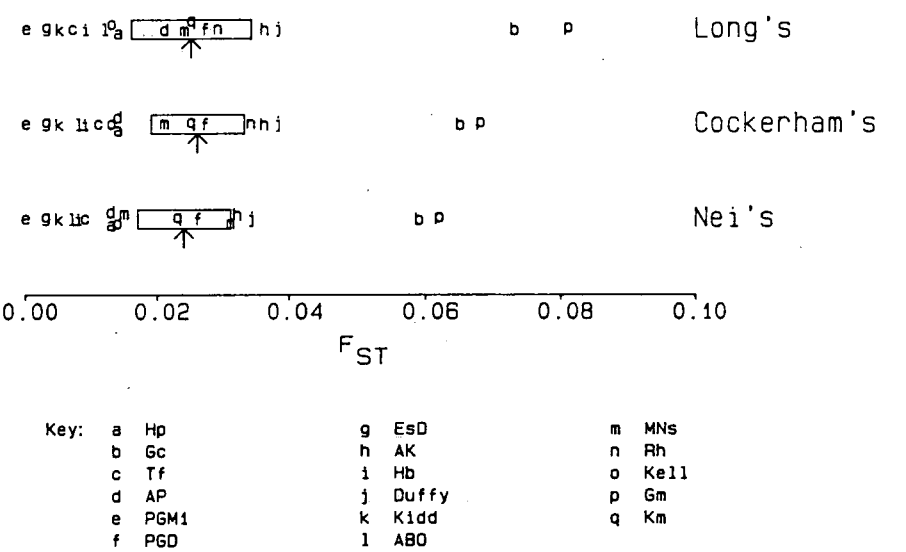


Fig. 2. A comparison of three locus-specific estimators of F_{ST} from allele frequency data on 17 loci in 13 subpopulations of Sikkim, India (Bhasin et al., 1986). The loci are indexed alphabatically (see Key). The averages over loci are indicated by arrow, and the boxes around these means represent $\pm$ s.e. range of the estimates (see text for the explanation of the estimators).

computations are the same as the ones reported in Bhasin et al. (1986). Figure 2 shows a diagramatical comparison of these locus-specific estimates, where the loci are indexed as a to q (see Key of Figure 2). The pooled estimates of F_{ST} over loci are indicated by an arrow, the box around which indicates the range with $\pm$ s.e.

It is clear that the estimates are again empirically very similar. Long's estimates are identical to the Cockerham estimates for all bi-allelic loci, although for multiallelic loci (Gc, Tf, AP, AB0, MNSs, Rh, and Gm) some discrepancies are noticable due to the different pooling (over alleles) algorithm employed in this method, as noted earlier. Nevertheless, the pooled estimates over loci are strikingly similar. Finally, we note that while the computation of the standard errors shown in this figure are based on equation (4.5), the jackknife estimates (equation (4.6)) of these standard errors are almost identical to the ones shown here. Hence, as in the case of genotype data, estimates of F_{ST} from allele frequency data also have similar empirical properties.

6. Discussion

As mentioned in Section 1, the purpose of this paper is to make a comprehensive comparative analysis of estimation of fixation indices by Nei's gene diversity approach with that of the variance component approach developed by Cockerham, or its multivariate extension. Keeping a distinction of parameters and sample statistics, throughout our presentation we have shown that these methods yield empirically very similar results. Even though these approaches have been described in a number of publications (see, e.g., Cockerham, 1969, 1973; Cockerham and Weir, 1986, 1988; Weir and Cockerham, 1984; Nei, 1973, 1977; Nei and Chesser, 1983; Chakraborty, 1974; Chakraborty and Leimar, 187; Long, 1986; Smouse and Long, 1988) and several other related statistics have been developed by others (Haldane, 1954; Li and Horvitz, 1953; Smith, 1970, 1977; Curie-Cohen, 1982; Robertson and Hill, 1984), to our knowledge, the analytical relationships between the two major approaches have not been studied explicitly before. In this discussion, first we re-iterate the new results presented here; and then we provide some arguments regarding the method we would suggest to practioners. Nevertheless, since during the conduct of this study, we developed a comprehensive computer-program for analyzing data on population structure, every estimator discussed in this paper can be computed by our computer algorithm. Interested readers can obtain a copy of the FORTRAN source codes of these programs by writing to the authors (compatible for IBM-AT type computers with a numerical co-procesor).

Our new results are as follows. First, the string of inequalities for the F_{IS} estimators in a single subpopulation shows that the expected differences among the estimators are of the order $1/2N$, N being the number of individuals sampled. While Nei's biased estimator of F_{IS} is always the smallest for any allele, Cockerham's variance-component estimator can be larger (when positive) or

smaller (when negative) than Nei's unbiased estimator. Second, even though Long (1986) and Smouse and Long (1988) generalized Cockerham's approach for a multivariate case (three or more alleles at a locus), they failed to note that their method yields F_{IS} estimators mathematically identical to Curie-Cohen's $\hat{f}_2$ (based on the ratio of observed and expected homozygote frequencies), in large samples. Third, for a single subpopulation, Nei's unbiased, biased (identical to $\hat{f}_1$ of Curie-Cohen, 1982—although not stated by him) and Cockerham's estimators of F_{IS} have closed form expressions of standard errors, for specific alleles as well as for the locus as a whole, which are also documented here for the first time. Much of the ground work for these derivations was, however, done by Curie-Cohen (1982) and Robertson and Hill (1984).

These new findings allow more rigorous comparative analyses of the different estimators, than the ones done before. Our empirical data analysis shows the closeness of the different estimators, which are based on somewhat different premises. There have been a number of misconceptions about the gene diversity approach, which should be clarified in this context. Note that the gene diversity approach does not need the correlation interpretation of the fixation indices. The total heterozygosity in subdivided populations is decomposed here on the basis of the number of extant subpopulations. No assumption of the replicative nature of subpopulations is needed. While Cockerham's linear model (of random effects) makes the assumption that the subpopulations studied are replicates from the universe of all subpopulations that exist within the total population, a situation that might apply to experimental populations, in the context of evolutionary significance, it is not clear if this assumption is realistic. In the specific example considered here, Sikkimese people are indeed subdivided into the present 13 subpopulations which during their history have assembled in this geographic region by following different migration routes (Bhasin et al., 1986). They are not replicates of each other, and indeed there may not be any further subpopulation among the people of Sikkim. If a statistical framework forms the basis of the variance–component analysis, the question is: should we treat the underlying linear model (Cockerham, 1969, 1973) as a random effects model in such a situation? Our answer to this question would be no as this subdivided structure represents a fixed-effect model. On the contrary, our exposition clearly indicates that Nei's gene diversity approach has a formal statistical basis, since all components of the decomposition of heterozygosity can be represented in terms of the underlying parameters, and they can be related with Wright's fixation indices without invoking their interpretation through correlations.

At this point it is worthwhile to note that for a single subpopulation the probabilistic interpretation of F_{IS} has been used by Haldane and Moshinsky (1939), Cotterman (1940), and Malécot (1948), where F_{IS} is interpreted as the probability that the two genes at a locus in an individual are identical. This probabilistic interpretation implicitly implies that the F_{IS} can take only non-negative values in the unit interval. Similar probabilistic interpretations of F_{IT} and F_{ST} are also used by Crow and Kimura (1970, pp. 105–106) to prove the Wright's identity (equation (2.1)). They, however, note that since F_{IT} and F_{IS} can be

negative, correlational interpretations of these fixation indices also yield the Wright's identity (Crow and Kimura, 1970, pp. 107–108). It is apparently implicit in their derivation that the subpopulations do not exchange migrants during the process of gene differentiation, so that the allele frequency variations across subpopulations do not depend upon the F_{IS} values within the subpopulations. In contrast, in Nei's formulation of gene diversity analysis the Wright's identity is established simply by the notion that F_{IS} and F_{IT} represent summary measures of deviations from the Hardy–Weinberg expectations in the subpopulations and in the total population, respectively, and F_{ST} represents the extent of genetic differentiation (standardized variance of allele frequencies across subpopulations). No assumption regarding migration and selection is needed in such derivation (Nei, 1973, 1977). The Wright's identity (equation (2.1)) simply becomes a mathematical consequence of the parametric definitions of F_{IT} (equation (2.5)), F_{IS} (equation (2.6)), and F_{ST} (equation (2.7)).

When the parameters are so defined, our equations (2.8), (2.9), and (2.10) suggest that all fixation indices have their natural bounds, namely F_{ST} lies between 0 and 1, while F_{IS} and F_{IT} can take positive as well as negative values, depending on H_S being smaller or larger than H_0 for F_{IS} (equation (2.8)) and H_T being smaller or larger than H_0 for F_{IT} (equation (2.9)). In such formulations no assumption is needed regarding the evolutionary mechanism that determines the process of genetic differentiation within and between subpopulations.

Since the variance–component approach can yield a negative value for the variance component b (equation (2.16)), in order to interprete the linear model (equation (16) of Cockerham, 1969) one must assume that $\sum_{k=1}^{r} w_i F_{IS_{ik}} p_{ik}(1 - p_{ik})$ must be positive. Cockerham (1969, 1973) recognized this feature of his model, and yet justified it on the ground that evolutionary factors that generally produce negative F_{IS} values are not usually strong enough to produce large negative F_{IS} (or F_{IT}) values. Our data analysis provides evidence contrary to this argument. We indeed found several negative estimates of F_{IS} (Tables 1 and 2). Even if their significance is discounted, because the normal deviates or the χ^2 statistics may not attain their large sample distribution in samples of the size analyzed here, it is unpleasant to deal with a linear model with negative variance components (not only the estimates, but also in parametric form).

Nei (1986) addressed some of these issues along with other evidences where the implicit assumptions of the variance component formulations are unrealistic for natural populations. He also noted that his original definition of F_{ST} ($= D_{ST}/H_T$, called G_{ST} by Nei, 1973) has one deficiency, since it is dependent on the number of subpopulations (s). He suggested one modification, defining $D'_{ST} = sD_{ST}/(s - 1)$, to take into account this deficiency (Nei, 1986). According to this suggestion, H'_T, the gene diversity in the total population is defined as $D'_{ST} + H_S$, yielding the three fixation indices $F_{IS} = H_0/H_S$ (unchanged from the previous definition), $F'_{IT} = H_0/H'_T$, and $F'_{ST} = D'_{ST}/H'_T$, for which the estimation technique presented here works with only minor modifications (see also Nei and Chesser, 1983). When s is large (say, 10 or more), these re-defined fixation indices change only slightly, and hence they are not computed in our application (since

for the present example $s = 13$). However, when s is small, it is preferable to calculate these modified indices with the above modifications. Also note that the re-defined F'_{ST} is identical to the parameter β defined by Cockerham and Weir (1988), not recognized by these authors. Therefore an estimator of β can also be obtained by estimating F'_{ST} in the gene diversity approach without the intraclass correlation interpretation. Nevertheless, we must reiterate the point that adjustment for the number of subpopulations does not necessarily help in comparing the coefficient of gene differentiation estimates in different data sets from different natural populations. An extrapolation of such estimates from one set of populations to another can be misleading, since their evolutionary histories are usually different. Cockerham's approach is more ideally suited for experimental populations, where the number of subpopulations represent the replicate of populations designed with a given experimental situation, and hence extrapolation from one experiment to another must need adjustments for variations in number of replicate subpopulations within each experiment.

Notwithstanding these philosophical differences, given the empirical similarity of the various estimators, a recommendation regarding the choice of estimators should be of interest to investigators who deal with real data. On the basis of statistical principles, unfortunately, there is no general recommendation. We claim this for several reasons. First, in every formulation, we have shown that consistent estimators can be derived. The study of large sample variances either by theoretical variances evaluated with intra-locus data, or by empirical evaluation of inter-locus variation shows that all estimators are subjected to similar sampling fluctuations. Second, even though with the aid of computer-algorithms the numerical task of computation can be left to computers, the choice is simply a matter of taste.

Since the gene diversity approach relates F_{ST} to the average genetic distances among subpopulations (Nei's minimum distance; Nei, 1972) a genetic distance interpretation of the coefficient of gene differentiation is also possible. Note that this interpretation does not assume, again, any evolutionary mechanism, and hence this interpretation should hold with or without mutation and selection. While Cockerham's F_{ST} parameters, and its multivariate extension have been shown to have a genetic distance interpretation as well (Reynolds et al., 1983; Long et al., 1987) in order for the measures of co-ancestry to be interpreted as genetic distances one must assume that genetic differentiation occurs without the aid of mutation and selection (Reynolds et al., 1983; Weir and Cockerham, 1984). Furthermore, in this latter paper they also assume that the same population size is maintained for all subpopulations and for all generations. While these assumptions are not needed in formulating Nei's genetic distances, thus far the evolutionary expectation and drift variances of genetic distances have been worked out under the neutral model of evolution without constant population size (Li and Nei, 1976; Nei and Chakravarti, 1977).

We advocate the use of the gene diversity approach for its simplicity and generality for natural populations. No loss of statistical rigor is attendant to this recommendation, as explicitly shown here—because we did not make use of any

evolutionary model in this presentation, and as a method of estimation, what we used can be called the method of moments in the terminology of statistical inference. This is the only appropriate estimation technique that yields analytically closed form estimators. We might add here that Curie-Cohen (1982) and Robertson and Hill (1984) investigated the properties of the maximum likelihood estimators of F_{IS} based on multinomial sampling of genotypes, which behave worse than Nei's biased estimator in most practical situations (Curie-Cohen, 1982).

Although we presented analytically closed form expressions of intra-locus variances of F_{IS} estimators, these are applicable only for single-locus data. Generally, large sample sizes are needed to apply these formulae, since the estimators are rather unstable (the drift variance is quite large; as shown by Li and Nei, 1976—for heterozygosities, Nei and Chakravarti, 1978—for $G_{ST} \approx F_{ST}$), and the power of detection of significant deviations of these indices is generally low (Brown, 1970; Ward and Sing, 1970; Chakraborty and Rao, 1972; Haber, 1980; Emigh, 1980). Evolutionary interpretation of the coefficients of gene differentiation or deviations from $F_{IS} = 0$, $F_{IT} = 0$ should be based on data on multiple loci. We have shown that multi-allelic and/or multiple-loci can be analyzed easily without the aid of Long's (1986) multivariate extensions. Indeed Nei's formulation of the decomposition of gene diversity is philosophically based on samples of genomes drawn from the population. He defined gene diversity as the complement of the probability that the two genomes are identical at each locus. Therefore, he computed gene diversity based on a sample of loci (polymorphic and monomorphic, see Nei, 1975, 1987). Even though the parameter F_{ST} (in Nei's terminology, G_{ST}) or its estimate does not change even if the monomorphic loci are excluded, the absolute value of H_T, H_S, and H_0 (averaged over all loci in a genome) changes. Even with a limited number of loci, we have shown that the variances of these quantities can be examined by studying their inter-locus variation, which yields inter-locus variances of the fixation indices as well. Since the inter-locus variance is the major component of the total sampling variance (Nei and Roychoudhury, 1974; Nei, 1978), jackknifing helps a little to provide a more reliable estimate of the extent of sampling variance. This finding is in disagreement with Mueller's study (Mueller, 1979) on genetic distance, but it is consistent with our own simulation results published before (Chakraborty, 1985). Weir and Cockerham (1984) claimed that jackknifing worked 'satisfactorily' in the two-population situation in their simulations (Reynolds et al., 1983), while we find that jackknifing does not add any particular advantage in terms of parameter estimates or their standard errors (Table 6).

Finally, we should return to the issue of hypothesis testing in the context of population structure data analysis. Considerable labor is needed to provide estimators adjusting for the effect of limited sample sizes. It is seen that when sample sizes are small (of the order of 100 or less individuals per locus per subpopulation, in the specific example given here), the use of large sample approximations yield over-estimates of F_{ST} and under-estimates of F_{IS} (F_{IT} remaining almost identical), irrespective of the method of analysis (Nei's vs. Cockerham's). Since such esti-

mates are invoked in evaluating standard errors or for computing test criteria, the question is: are these test criteria reliable, and can we justify the large sample properties of these test criteria? Our answer, although we cannot prove it analytically, is a probable no. We say so, for the reason that if the normal deviates are to be regarded as reliable, we must evaluate the standard errors accurately. We have seen that in some region of the parametric space, the standard errors can be drastically affected, even by a minute change in the parameter estimates. For the χ^2 tests, on the other hand, we must regard the variance components as independently distributed. This assumption, we might note, is also needed in Long's (1986) Wilk's Λ-test criteria. Nayak (1983) has shown that when the genotype data from several subpopulations are represented in the form of an analysis of variance of categorical data (Light and Margolin, 1971), the mean square errors of the different sources of variation are not independently distributed. The correlations between them can often be substantial. Furthermore, for every source of variation, the large sample distribution of the mean square errors is of the form of composite χ^2's, where the coefficients are also functions of unknown parameters. They cannot be simply equated to a χ^2 statistics as is done commonly invoking large sample theory of continuously distributed random variables. Therefore, we argue that the test statistics generally suggested for population structure analysis have much poorer statistical justifiability than the parameter estimates. Cockerham (1973) arrived at this general conclusion, although the sampling theory of weighted categorical data analysis was not available at that time.

7. Summary

A comprehensive comparative study of the various estimators of the fixation indices (F_{IT}, F_{IS}, and F_{ST}) shows that the properties of the estimators based on Nei's gene diversity and Cockerham's variance component analysis are very similar, in spite of their philosophical differences. In the analysis of genotypic data from a single population, a string of inequalities of the different estimators of F_{IS} is mathematically established, with regard to which the discrepancies in the sampling precision of these estimators can be reconciled. The analytical expression for the large sample variance of these estimators suggests that the parametric value of their sampling variance is identical. Empirical evaluation of the bias and standard errors of the three fixation indices from a genetic survey of 17 loci from 13 subpopulations of Sikkim, India suggests that for these ratio estimators the Jackknife method and Taylor's series approximation yield almost identical bias and standard error. These conclusions also hold for the estimation of F_{ST} from allele frequency data alone. A comprehensive computer program for obtaining all estimators has been developed, and is available from the authors upon request.

Acknowledgements

This paper is dedicated to the memory of Sewall Wright, the pioneer of population structure analysis, who passed away during the progress of this work. Dr Masatoshi Nei and Dr William J. Schull are to be acknowledged for their help in critically reviewing earlier versions of this review. Thanks are also due to Mr R. Schwartz for his help in computation and graphic works. This work was supported by grants from the National Institutes of Health and National Science Foundation. HDH was supported by the German research fellowship program of DAAD during the conduct of this work.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd edition). John Wiley, New York.

Bhasin, M. K., Walter, H., Chahal, S. M. S., Bhardwaj, V., Sudhakar, K., Danker-Hopfe, H., Dannewitz, A., Singh, I. P., Bhasin, V., Shil, A. P., Sharma, M. B. and Wadhavan, D. (1986). Biology of the people of Sikkim, India. I. Studies on the variability of genetic markers. *Z. Morph. Anthrop.* **77**, 49–86.

Brown, A. H. D. (1970). The estimation of Wright's fixation index from genotype frequencies. *Genetica* **41**, 399–406.

Chakraborty, R. (1974). A note on Nei's measure of gene diversity in a substructured population. *Humangenetik* **21**, 85–88.

Chakraborty, R. (1985). Genetic distance and gene diversity: Some statistical considerations. In: *Multivariate Analysis – VI*, P. R. Krishnaiah (ed.). Elsevier, Amsterdam, 77–96.

Chakraborty, R., Chakravarti, A. and Malhotra, K. C. (1977). Variation of allele frequencies among caste groups of the Dhangars of Maharashtra, India: An analysis with Wright's *F*-statistics. *Ann. Hum. Biol.* **4**, 275–280.

Chakraborty, R. and Nei, M. (1977). Bottleneck effect with stepwise mutation model of electrophoretically detectable alleles. *Evolution* **31**, 347–356.

Chakraborty, R. and Leimar, O. (1987). Genetic variation within a subdivided population. In: *Population Genetics and Fishery Management*, N. Ryman and F. Utter (eds.). Sea Grant Program, University of Washington Press, Seattle, WA, 89–120.

Chakraborty, R. and Rao, D. C. (1972). On the detection of *F* from ABO blood group data. *Am. J. Hum. Genet.* **24**, 352–353.

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution* **23**, 72–84.

Cockerham, C. C. (1973). Variance of gene frequencies. *Evolution* **27**, 679–700.

Cockerham, C. C. and Weir, B. S. (1986). Estimation of inbreeding parameters in stratified populations. *Ann. Hum. Genet.* **50**, 271–281.

Cockerham, C. C. and Weir, B. S. (1987). Correlations, descent measures: Drift with migration and mutation. *Proc. Natl. Acad. Sci. USA* **84**, 8512–8514.

Cotterman, C. W. (1940). A calculas for statistic-genetics. Ph.D. dissertation, Ohio University, Columbus, OH.

Crow, J. F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row, New York.

Curie-Cohen, M. (1982). Estimates of inbreeding in a natural population: A comparison of sampling properties. *Genetics* **100**, 339–358.

Efron, B. (1982). *The Jackknife, the Bootstrap and other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia.

Haber, M. (1980). Detection of inbreeding effects by the chisquare test on genotypic and phenotypic frequencies. *Am. J. Hum. Genet.* **32**, 754–760.

Haldane, J. B. S. (1954). An exact test for randomness of mating. *J. Genet.* **52**, 631–635.
Haldane, J. B. S. and Moshinsky, P. (1939). Inbreeding in Mendelian populations with special reference to human cousin marriage. *Ann. Eugen.* **9**, 321–340.
Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistica. Vol. 1* (4th edition). MacMillan, New York.
Kirby, G. C. (1975). Heterozygote frequencies in small subpopulations. *Theoretical Population Biology* **8**, 31–48.
Li, C. C. (1955). *Population Genetics*. University of Chicago Press, Chicago, IL.
Li, C. C. and Horvitz, D. G. (1953). Some methods of estimating the inbreeding coefficient. *Am. J. Hum. Genet.* **5**, 107–117.
Li, W.-H. and Nei, M. (1975). Drift variances of heterozygosity and genetic distance in transient states. *Genet. Res.* **25**, 229–248.
Light, R. J. and Margolin, B. H. (1971). An analysis of variance for categorical data. *J. Am. Stat. Assoc.* **66**, 534–544.
Long, J. C. (1986). The allelic correlation structure of Gainj- and Kaam-speaking people. I. The estimation and interpretation of Wright's *F*-statistics. *Genetics* **112**, 629–647.
Malécot, G. (1948). *Les Mathématiques de l'Hérédité*. Masson et Cie, Paris.
Miller, R. G. (1974). The jackknife – A review. *Biometrika* **61**, 1–15.
Mueller, L. D. (1979). A comparison of two methods for making statistical inferences on Nei's measure of genetic distance. *Biometrics* **35**, 757–763.
Nayak, T. K. (1983). Applications of entropy functions in measurement and analysis of diversity. Ph.D. Dissertation, University of Pittsburgh, Pittsburgh, PA.
Nei, M. (1972). Genetic distance between populations. *Am. Nat.* **106**, 283–292.
Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc.* Natl. Acad. Sci. USA **70**, 3321–3323.
Nei, M. (1975). *Molecular Population Genetics and Evolution*. North-Holland, Amsterdam.
Nei, M. (1977). *F*-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.* **41**, 225–233.
Nei, M. (1978). Estimation of average hterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583–590.
Nei, M. (1986). Definition and estimation of fixation indices. *Evolution* **40**, 643–645.
Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.
Nei, M. and Chakravarti, A. (1977). Drift variances of F_{ST} and G_{ST} statistics obtained from a finite number of isolated populations. *Theor. Pop. Biol.* **11**, 307–325.
Nei, M. and Chesser, R. K. (1983). Estimation of fixation indices and gene diversities. *Ann. Hum. Genet.* **47**, 253–259.
Nei, M., Maruyama, T. and Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution* **29**, 1–10.
Nei, M. and Roychoudhury, A. K. (1974). Sampling variance of hterozygosity and genetic distance. *Genetics* **76**, 379–390.
Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. Wiley, New York.
Rao, C. R. (1982). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhya A* **44**, 1–21.
Rao, C. R., Rao, D. C. and Chakraborty, R. (1973). The generalized Wright's model. In: *Genetic Structure of Populations*, N. E. Morton, (ed.). University of Hawaii Press, Honolulu, 55–59.
Rao, D. C. and Chakraborty, R. (1974). The generalized Wright's model and population structure. *Am. J. Hum. Genet.* **26**, 444–453.
Reynolds, J., Weir, B. S. and Cockerham, C. C. (1983). Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* **105**, 767–779.
Robertson, A. and Hill, W. G. (1984). Deviation from Hardy–Weinberg proportions: Sampling variances and use in estimation of inbreeding coefficients. *Genetics* **107**, 703–718.
Slatkin, M. and Barton, N. H. (1989). A comparison of three methods for estimating average level of gene flow. *Evolution* **43**, 1349–1368.
Smith, C. A. B. (1970). A note on testing the Hardy–Weinberg law. *Ann. Hum. Genet.* **33**, 377–383.

mates are invoked in evaluating standard errors or for computing test criteria, the question is: are these test criteria reliable, and can we justify the large sample properties of these test criteria? Our answer, although we cannot prove it analytically, is a probable no. We say so, for the reason that if the normal deviates are to be regarded as reliable, we must evaluate the standard errors accurately. We have seen that in some region of the parametric space, the standard errors can be drastically affected, even by a minute change in the parameter estimates. For the χ^2 tests, on the other hand, we must regard the variance components as independently distributed. This assumption, we might note, is also needed in Long's (1986) Wilk's Λ-test criteria. Nayak (1983) has shown that when the genotype data from several subpopulations are represented in the form of an analysis of variance of categorical data (Light and Margolin, 1971), the mean square errors of the different sources of variation are not independently distributed. The correlations between them can often be substantial. Furthermore, for every source of variation, the large sample distribution of the mean square errors is of the form of composite χ^2's, where the coefficients are also functions of unknown parameters. They cannot be simply equated to a χ^2 statistics as is done commonly invoking large sample theory of continuously distributed random variables. Therefore, we argue that the test statistics generally suggested for population structure analysis have much poorer statistical justifiability than the parameter estimates. Cockerham (1973) arrived at this general conclusion, although the sampling theory of weighted categorical data analysis was not available at that time.

7. Summary

A comprehensive comparative study of the various estimators of the fixation indices (F_{IT}, F_{IS}, and F_{ST}) shows that the properties of the estimators based on Nei's gene diversity and Cockerham's variance component analysis are very similar, in spite of their philosophical differences. In the analysis of genotypic data from a single population, a string of inequalities of the different estimators of F_{IS} is mathematically established, with regard to which the discrepancies in the sampling precision of these estimators can be reconciled. The analytical expression for the large sample variance of these estimators suggests that the parametric value of their sampling variance is identical. Empirical evaluation of the bias and standard errors of the three fixation indices from a genetic survey of 17 loci from 13 subpopulations of Sikkim, India suggests that for these ratio estimators the Jackknife method and Taylor's series approximation yield almost identical bias and standard error. These conclusions also hold for the estimation of F_{ST} from allele frequency data alone. A comprehensive computer program for obtaining all estimators has been developed, and is available from the authors upon request.

Smith, C. A. B. (1977). A note on genetic distance. *Ann. Hum. Genet.* **40**, 463–479.
Smouse, P. E. and Long, J. C. (1988). A comparative *F*-statistics analysis of the genetic structure of human populations from the Lowland South America and Highland New Guinea. In: *Quantitative Genetics*, B. S. Weir, E. J. Eison, M. M. Goodman and G. Namkoong (eds.). Sinaur Association Inc., Sunderland, 32–46.
Van Den Bussche, R. A., Hamilton, M. J. and Chesser, R. K. (1986). Problems of estimating gene diversity among populations. *The Texas Journal of Science* **38**, 281–287.
Weir B. S. and Cockerham, C. C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.
Workman, P. L. and Niswander, J. D. (1970). Population studies on Southwestern Indian tribes. II. Local genetic differentiation in the Papago. *Am. J. Hum. Genet.* **22**, 24–49.
Wright, S. (1943). Isolation by distance. *Genetics* **28**, 114–138.
Wright, S. (1951). The genetical structure of populations. *Ann. Eugenics* **15**, 323–354.
Wright, S. (1965). The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution* **19**, 395–420.

Ann. Hum. Genet. (1992), **56**, 45–57
Printed in Great Britain

Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications

R. CHAKRABORTY[1]*, M. DE ANDRADE[1], S. P. DAIGER[2] AND B. BUDOWLE[3]

[1] *Center for Demographic and Population Genetics and*
[2] *Medical Genetics Center, University of Texas Graduate School of Biomedical Sciences, Houston, Texas*
[3] *Forensic Science Research and Training Center, Laboratory Division, FBI Academy, Quantico, Virginia*

SUMMARY

Restriction fragment length polymorphisms (RFLP) analysis using the Southern blot technique can be used to recognize copy number variation of variable number of tandem repeats (VNTR) of conserved core sequences at several regions of the human genome. This new class of polymorphisms reveals a high degree of genetic variation, useful for individual identification purposes. Criticisms against forensic applications of such DNA typing data include the limitation of employing Hardy–Weinberg expectation of genotype frequencies, since several surveys indicate apparent deficiency of heterozygosity (or excess homozygosity) in comparison with Hardy–Weinberg expectations. This research postulates an alternative explanation of deficiency of apparent heterozygosity which is caused by the inability to detect extremely small-sized alleles (called 'non-detectable' alleles) due to the sensitivity of Southern gel electrophoresis. We show that the presence of 'non-detectable' alleles can produce pseudo-homozygosity and their frequencies can be predicted from the observed proportional heterozygote deficiency. Furthermore, in the covert presence of such 'non-detectable' alleles, we show that the gene-count method provides over-estimates of allele frequencies in the sample population, and hence the Hardy–Weinberg predictions of genotype frequencies avoid wrongful bias against suspects in forensic applications of DNA typing data. Applications of this theory to population data on six VNTR loci in US Caucasians and US Blacks suggest that the presence of 'non-detectable' alleles could be the major cause of apparent heterozygote deficiency, and the current approaches of predicting the population frequency of specific DNA phenotypes are practically free of the possible wrongful bias in courtroom applications of DNA typing data.

INTRODUCTION

Scientific as well as social implications of the discovery of hypervariable VNTR loci in the human genome are by now well-recognized (Wyman & White, 1980; Jeffreys *et al.* 1985; Nakamura *et al.* 1987; Ballantyne *et al.* 1989). While the general criteria of the presence of large numbers of alleles and high heterozygosities at most VNTR loci make them ideal candidates for genetic 'fingerprinting' of individuals, ensuing controversies (Lander, 1989, 1991; Thompson & Ford, 1989; Cohen, 1990) with regard to their forensic applications prompted careful attention

* Correspondence to: Ranajit Chakraborty, Ph.D., Center for Demographic and Population Genetics, University of Texas, Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, TX 77225, U.S.A.

to various statistical as well as population genetic characteristics of such hypervariable loci (Devlin *et al.* 1990; Chakraborty & Daiger, 1991; Chakraborty *et al.* 1991). In a realistic as well as rigorous study, Devlin *et al.* (1990) demonstrated that the quasi-continuous variations of allele sizes at many VNTR loci could be produced at least in part by measuremental errors of allele-size determination. Individuals who exhibit both alleles of similar (but different) sizes may be wrongly typed as homozygotes, and therefore the population data may indicate evidence of heterozygote deficiency (or equivalently, excess of homozygosity) in comparison with the predictions of Hardy–Weinberg equilibrium (HWE) law. Since most genotypes at VNTR are rare, it has been demonstrated that by necessity genotype frequencies at VNTR loci best be predicted from their allele frequencies (Chakraborty *et al.* 1991). As a matter of fact, for forensic applications it may be enough to prescribe an upper bound for every genotype probability, since this will avoid wrongful bias in criminal offense cases (Ballantyne *et al.* 1989; Budowle *et al.* 1991*a*). Since under the assumption of HWE the probability of observing a pair of alleles constituting an individual's genotype is given by the product of the respective alleles in the population or multiplying this product by a factor of two if the alleles are dissimilar, it may also be called a 'product rule' of genotype probability determination. The most common cause of deviation from the above product rule (in the direction of excess homozygozity) observed in the content of the VNTR polymorphism is claimed to be population subdivision (Lander, 1989; Cohen, 1990). The purpose of this communication is to examine the plausibility of another important 'technical' causal factor, the non-detectability of a class of small-sized alleles. This is analytically equivalent to the problem of 'null' alleles that exist in human and other organisms at protein-coding loci (Martin, 1983; Foltz, 1986*a*, *b*). The objectives of this research are: (1) to show that the presence of such 'non-detectable' alleles produces pseudo-homozygosity, and their frequencies in a population can be predicted from the proportional heterozygote deficiency, although other rigorous methods of estimation of such allele frequencies are available in the literature (Gart & Nam, 1984*a*); (2) to demonstrate that their presence may remain covert, although there might be considerable numbers of heterozygote individuals in the sample, where such 'non-detectable' alleles are found in combination with other alleles so that such individuals are observed as apparent homozygotes (for the detected alleles) because of their single-band Southern-gel profiles; and finally, (3) to show that in the presence of such alleles, the gene-count method of estimation of allele frequencies over-estimates the actual frequencies of all detectable alleles, and hence the product-rule gives enough cushion in the prediction of genotype probabilities, avoiding bias against suspects in forensic applications of DNA typing data.

Analytical explorations of this technical problem is followed by an analysis of population data on six VNTR loci (D2S44, D14S13, D4S139, D17S79, D1S7 and D16S85) in US Caucasians and US Blacks, classified by the fixed-bin approach suggested earlier (Budowle *et al.* 1991*a*). It has been shown earlier that the fixed-bin approach of allele classification greatly circumvents the problem of measuremental error of allele size determination, and hence analytical strategies of detecting pseudo-homozygosity by the approach of Devlin *et al.* (1990) do not have to be involved in this study (Budowle *et al.* 1991*a*).

FREQUENCY CONSEQUENCES OF 'NON-DETECTABLE' ALLELES

Consider the case of k detectable alleles ($A_1, A_2, \ldots, A_k$) and a class of 'non-detectable' alleles (i.e. alleles of too small sizes) that may truly contain alleles of dissimilar sizes, each of which remains undetected on a Southern gel. We designate this class of alleles by A_0. This scenario is reminiscent of the HLA system in humans, where A_0 is the blank allele and A_is are the ones detected by the specific HLA-allelic antisera. We can therefore designate the different genotype and phenotype frequencies and their respective probabilities under HWE as shown in Table 1, where p_i is the frequency of the allele A_i ($i = 1, 2, \ldots, k$) and r is the frequency of 'non-detectable' class of alleles (A_0).

The HWE probabilities of the observed phenotypes, conditioned on the observed absence of homozygotes for the 'non-detectable' alleles (A_0, A_0), then, become

$$Pr(A_i-) = \frac{(p_i^2+2p_i r)}{(1-r^2)}, \quad \text{for } i = 1,2,\ldots,k. \tag{1}$$

and

$$Pr(A_iA_j) = \frac{2p_i p_j}{(1-r^2)}, \quad \text{for } j > i = 1,2,\ldots,k. \tag{2}$$

In contrast, if we ignore the existence of 'non-detectable' alleles altogether, we would regard each A_i- phenotype as the true homozygote A_iA_i (for $i = 1, 2, \ldots, k$) and A_iA_j phenotype as heterozygote, so that the gene-count estimators (which are also the maximum likelihood estimators) of allele frequencies are

$$\tilde{p}_i = \frac{2n_{i-}+\Sigma_{i\neq j} n_{ij}}{2n}, \quad \text{for } i = 1,2,\ldots,k. \tag{3}$$

where $n = \Sigma_{i=1}^{k} n_{i-} + \Sigma_{j>i=1}^{k} n_{ij}$.

The expectations of $\tilde{p}_i$ in the conditional data set, in the presence of A_0 allele(s) are

$$\begin{aligned} E(\tilde{p}_i) &= \frac{2n[p_i^2+2p_i r]+2np_i\Sigma_{i\neq j}p_j}{2n(1-r^2)} \\ &= \frac{p_i^2+2p_i r+p_i(1-p_i-r)}{1-r^2} \\ &= \frac{p_i(1+r)}{1-r^2} \\ &= \frac{p_i}{1-r} \end{aligned} \tag{4}$$

for $i = 1, 2, \ldots, k$.

For $r > 0$, obviously $\tilde{p}_i$ is an over-estimate of the true allele frequency, p_i, in the sampled population.

Ignoring the non-detectable alleles (A_0), one would assert that under HWE the expected heterozygosity, H_E, is

$$H_E = \sum_{i\neq j} \tilde{p}_i\tilde{p}_j = 1-\sum_{i=1}^{k} \tilde{p}_i^2, \tag{5}$$

Table 1. *Genotypes, phenotypes and frequencies of a VNTR locus in the presence of 'non-detectable' alleles*

Genotype	Phenotype	Observed frequency	Probability under HWE
A_iA_i	A_i-	n_i	$p_i^2+2p_ir$ for $i=1,2,\ldots,k$
A_iA_j	A_iA_j	n_{ij}	$2p_ip_j$ for $i<j=1,2,\ldots,k$
A_0A_0	Blank	n_{00}	r^2

Note: A_i- phenotype appears as a single-band-lane on a Southern gel (for $i=1,2,\ldots,k$), the size of which can be detected on a control ladder-lane.

A_iA_j phenotype (genotype) is the unequivocal two-banded lane, sizes of both alleles being detectable.

Blank phenotypes are rare, showing no band in the lane; often ignored because they can also be caused by insufficient DNA or other factors (such as DNA degradation, etc.).

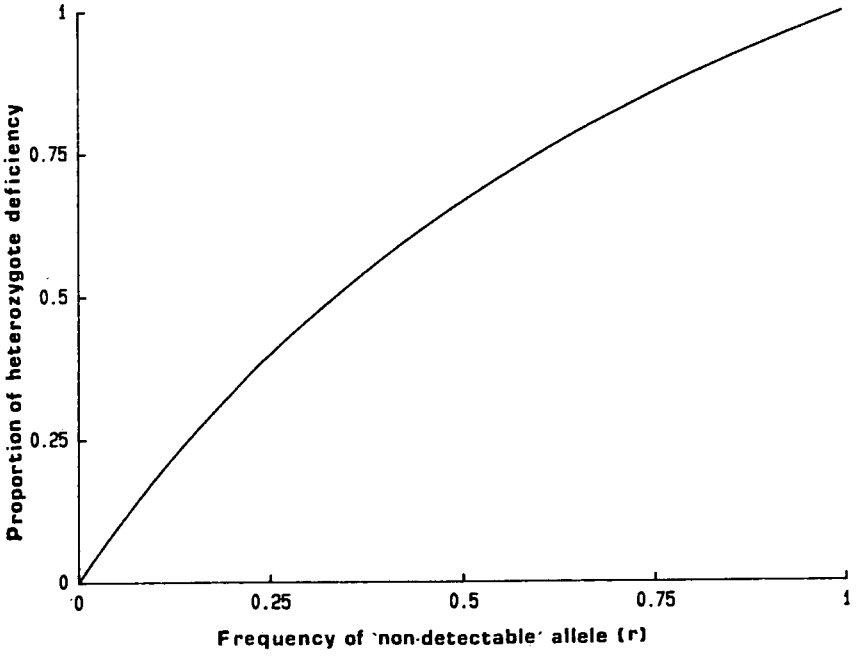


Fig. 1. Proportional heterozygote deficiency as a function of the frequency of 'non-detectable' alleles.

which would be contrasted with the observed heterozygosity, H_O, given by

$$H_O = \frac{\Sigma_{i\neq j} p_i p_j}{1-r^2}. \tag{6}$$

Note that $1-r^2=(1-r)(1+r)>(1-r)^2$, hence from equation (6) we have

$$H_O < \sum_{i\neq j}\left(\frac{p_i}{1-r}\right)\left(\frac{p_j}{1-r}\right) = \sum_{i\neq j}\tilde{p}_i\tilde{p}_j$$

$$= 1-\sum_{i=1}^{k}\tilde{p}_i^2 = H_E, \tag{7}$$

showing that in conditional data (with homozygosity for non-detectable alleles absent), non-detectable alleles can produce an observed apparent heterozygote deficiency (or equivalently, excess homozygosity) in comparison with the HWE prediction.

Furthermore, invoking (4) in (5) and (6) we get the expected heterozygote deficiency proportional to expected heterozygosity

$$D = \frac{H_E - H_O}{H_E} = \frac{2r}{1+r}, \tag{8}$$

since

$$\begin{aligned} H_O - H_E &= \frac{\Sigma_{i \neq j} p_i p_j}{1-r^2} - \sum_{i \neq j} \tilde{p}_i \tilde{p}_j \\ &= \sum_{i \neq j} p_i p_j \left[\frac{1}{1-r^2} - \frac{1}{(1-r)^2} \right] \\ &= \sum_{i \neq j} p_i p_j \left[\frac{(1+r)-(1-r)}{(1+r)(1-r)^2} \right] \\ &= -\frac{2r}{1+r} \sum_{i \neq j} \tilde{p}_i \tilde{p}_j. \end{aligned} \tag{9}$$

Therefore, the expected frequency of the non-detectable alleles (r) that can produce an observed level of proportional heterozygote deficiency (D) can be evaluated as

$$r_* = \frac{D}{(2-D)}. \tag{10}$$

Figure 1 shows the value of D as function of r. The concavity of the curve indicates that the proportional heterozygote deficiency (D) is at least as large as r (since $D \geqslant r$, the equality holding only in the extreme cases when $r = 0$ or 1). In other words, even a rare occurrence of non-detectable allele would produce a noticeable amount of heterozygote deficiency, particularly when the expected heterozygosity at the locus is 70% or greater (which is usually the case with most VNTR loci).

COVERTNESS OF NON-DETECTABLE ALLELES

The algebra of the preceding section demonstrates that the existence of rare 'non-detectable' alleles could produce an appreciable amount of observed heterozygote deficiency (or, excess homozygosity); the frequency of such alleles can be predicted from the observed proportional heterozygote deficiency (D). One might then ask, if this were the case why do such alleles remain covert; i.e. not seen in homozygote form? There is the possibility that a blank lane on a Southern gel might often be attributed to other technical problems, such as insufficient DNA and/or DNA degradation. Therefore, even if 'non-detectable' alleles appear in homozygous form they would not necessarily be scored. Further reasoning on statistical grounds can be given from the expected frequency of homozygosity for the non-detectable alleles (nr^2) and the probability of observing at least one homozygote of the type for a given sample size.

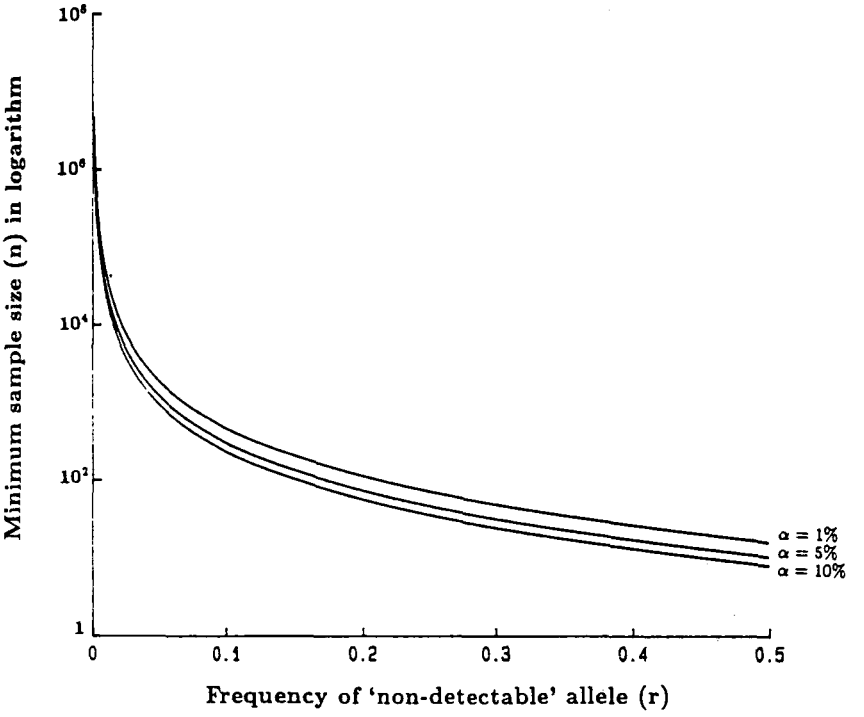


Fig. 2. Minimum sample size (n) needed to observe homozygotes for 'non-detectable' alleles as functions of frequency of such alleles (r) for different levels of significance (α).

The expected number of homozygotes for the non-detectable alleles is

$$F = nr_*^2 = \frac{nD^2}{(2-D)^2}, \tag{11}$$

which can remain very low even if D is large. For example, for $D = 0{\cdot}10$ F becomes 1 in a sample of 100 individuals, 5 in 500 individuals. Therefore, there is a high chance that such individuals might be missed in a survey. The probability that no such homozygotes would be seen in a sample of n individuals is given by

$$P_0(n) = (1-r_*^2)^n, \tag{12}$$

from a simple binomial distribution treatment.

Alternatively, the sample size needed (n) to observe at least one such homozygote in a survey with a confidence of $100(1-\alpha)\%$ must satisfy the inequality

$$1-(1-r^2)^n \geqslant 1-\alpha,$$

or
$$(1-r^2)^n \leqslant \alpha,$$

or
$$n \log(1-r^2) \leqslant \log(\alpha),$$

or
$$n \geqslant \frac{\log(\alpha)}{\log(1-r^2)}. \tag{13}$$

Table 2. *Observed heterozygote deficiencies at six VNTR loci in US Caucasians and Blacks and the frequencies of 'non-detectable' alleles that explain these deficiencies*

Locus	Sample size (n)	Obs. freq.		Exp. freq.		Prop. hetero. def. (D) in %	Pred. value of r in %
		Homo.	Hetero.	Homo.	Hetero.		
				US Caucasians			
D2S44	218	19	199	17·49	200·51	0·753	0·378
D14S13	218	22	196	19·18	198·82	1·418	0·714
D4S139	144	18	126	14·26	129·74	2·883	1·462
D17S79	209	61	148**	43·17	165·83	10·752	5·681
D1S7	210	19	191*	12·07	197·93	3·501	0·416
D16S85	210	21	189	19·42	190·58	0·829	0·416
				US Blacks			
D2S44	295	35	260**	19·41	275·59	5·657	2·911
D14S13	258	24	234*	15·46	242·54	3·521	1·792
D4S139	304	28	276	22·46	281·54	1·968	0·994
D17S79	281	54	227*	39·76	241·24	5·903	3·041
D1S7	268	23	245*	15·24	252·76	3·070	1·559
D16S85	212	47	165*	34·53	177·47	7·027	3·641

*$P < 0·05$; **$P < 0·01$ by a χ^2 test with 1 D.F.

Figure 2 shows plots of such minimum sample size requirement $n = \log(\alpha)/\log(1-r^2)$, when n is plotted in logarithmic scale as function of r for $\alpha = 0·01$, 0·05 and 0·10. Clearly, the sample size requirement is rather stringent. For example, for $r = 0·10$ [which would produce almost 18% proportional heterozygote deficiency, see equation (13)] we would need at least 1000 individuals to be screened before encountering at least one 'non-detectable' homozygote genotype with a confidence of 95%. When $r = 0·01$, the 95% confidence minimum sample size for observing at least one such individual is 29995. It is no surprise, therefore, that 'non-detectable' alleles remain covert in samples observed.

ANALYSIS OF SIX VNTR LOCI IN US CAUCASIANS AND US BLACKS

Budowle *et al.* (1991*a*) recently presented the analysis of phenotypes at six VNTR loci (D1S7, D2S44, D4S139, D14S13, D16S85 and D17S79) for the US Caucasians and US Blacks that exist in the FBI data base. Fixed bin approach (Budowle *et al.* 1991*a*) gives allele counts of binned alleles from which concordance of the observed numbers of heterozygotes (or homozygotes) with their HWE predictions can be examined (see Table 9 or Budowle *et al.* 1991*a*). Table 2 presents the summary of their computations. At each locus heterozygote deficiency is noticed in both populations, and in several cases (e.g. D1S7 and D17S79 in US Caucasians, and at all loci except the D4S139 locus in US Blacks) the observed heterozygote deficiencies are significant (at least 5% level). Ordinarily, one would have asserted the presence of substantial population substructuring in both populations from such data. The computations of Table 2 indicate that the heterozygote deficiency observed at each locus in this data can be easily explained by the nominal presence of 'non-detectable' alleles. Even the most deviant locus (D17S79 in the US Caucasians) exhibiting more than 10% proportional heterozygote deficiency could be regarded

Table 3. *Sample size needed to observe at least one 'non-detectable' homozygote*

Locus	Sample size (n)	Prop. hetero. def. (D) in %	Exp. freq. of homo. non-detec. alleles (nr^2)	Sample size required to observe at least one non-detec. homo. for $\alpha = 0{\cdot}01$	$\alpha = 0{\cdot}05$	$\alpha = 0{\cdot}10$
			US Caucasians			
D2S44	218	0·753	0·003	322365	209703	161183
D14S13	218	1·418	0·011	90268	58721	45134
D4S139	144	2·883	0·031	21531	14006	10765
D17S79	209	10·752	0·675	1425	927	713
D1S7	210	3·501	0·067	14503	9435	7252
D16S85	210	0·829	0·004	265787	172899	132894
			US Blacks			
D2S44	295	5·657	0·250	5433	3535	2717
D14S13	258	3·521	0·083	14337	9327	7169
D4S139	304	1·968	0·030	46640	30340	23320
D17S79	281	5·903	0·260	4977	3238	2489
D1S7	268	3·070	0·065	18946	12325	9475
D16S85	212	7·027	0·281	3472	2259	1726

as satisfying the HWE predictions with 'non-detectable' alleles that have a combined frequency below 6%. Frequencies of such alleles required to explain the other observed heterozygote deficiencies are even smaller (see Table 2, last column).

The supposition that these frequencies (r) of 'non-detectable' alleles are reasonable can be judged from the computation presented in Table 3. We note that with these predicted r values, the expected frequency of homozygotes for 'non-detectable' alleles in this database does not exceed one in any case (the maximum value of nr^2 is 0·675 for the D17S79 locus for the US Caucasian sample), and hence it is not a surprise that no such homozygote was detected. The sample size required to observe at least one such homozygote is also very large, as shown in the last three columns of Table 3. For example, even in the worst case (D17S79 in Caucasians) we would have needed 713 individuals to be 90% sure of getting at least one 'blank' phenotype. We may therefore conclude that even the case of a substantial heterozygote deficiency can be explained easily by covertness of 'non-detectable' alleles due to Southern gel electrophoresis.

IMPLICATIONS FOR FORENSIC APPLICATIONS OF VNTR TYPING DATA

As mentioned before, VNTR polymorphisms are useful for identification of individuals because the number of possible genotypes at VNTR loci is large, and most of them are rare enough in a population, so that the probability that two individuals will have identical genotypes (particularly for multiple numbers of such loci) by chance alone is very low. The rareness of the genotypes, however, poses a limitation, in the sense that the chance occurrence of genotypes must be evaluated from a population data base using population genetics principles such as the HWE (or the product rule). The theory and data analysis presented above suggest that in the presence of 'non-detectable' alleles such a rule of estimating genotype frequencies from allele-frequency data should not apply, because the combined frequency of

heterozygotes will fall below the predicted one. We argue that this should not be of any concern in the use of such allele frequency data for forensic application, because all that is necessary to establish that the chance occurrence of a specific genotype is rare is to prescribe an upper bound for the frequency of every conceivable genotype. If 'non-detectable' alleles are the predominant source of observed heterozygote deficiency, then gene-count estimates of all detectable alleles provide enough cushion to prescribe such upper bounds.

For example, if the Southern gel profile for a particular VNTR locus depicts genotype A_iA_j (bands of size i and j are detected), the true probability of this genotype in the population (under HWE) is $2p_ip_j$, but based on the allele-count data on conditional data (no 'blank' homozygote found) is $2\tilde{p}_i\tilde{p}_j$ ($\tilde{p}_i$ is estimated from equation (3)).

Note that

$$E(2\tilde{p}_i\tilde{p}_j) = \frac{2p_ip_j}{(1-r)^2}\left[1-\frac{1}{2n}\right]$$

and hence

$$E\left(\frac{2(2n-1)\tilde{p}_i\tilde{p}_j}{2n}\right) = \frac{2p_ip_j}{(1-r)^2} \geqslant 2p_ip_j$$

so that $2(2n-1)\tilde{p}_i\tilde{p}_j/(2n)$ is an over-estimate of the true frequency of the genotype A_iA_j in the population. When n is larger than 100, $2\tilde{p}_i\tilde{p}_j$ should reasonably serve as an adequate upper bound for the true heterozygote frequency in the population.

For the apparent homozygotes (single-band pattern of type A_i) the situation is somewhat more involved. Since the actual probability of the A_i- phenotype in the population is $p_i^2+2p_ir$ (under HWE in the presence of 'non-detectable' alleles), we cannot guarantee that $\tilde{p}_i^2$ will always be an over-estimate of $p_i^2+2p_ir$. However, Budowle *et al.* (1991*a*) suggested $2\tilde{p}_i$ as an estimator of the probability of A_i- phenotype in the population. Now note that $E(\tilde{p}_i) = p_i/(1-r)$, and $p_i^2+2p_ir = p_i(p_i+2r) < 2p_i(p_i+r) \leqslant 2p_i$, since $p_i+r \leqslant 1$. Therefore,

$$E(2\tilde{p}_i) \geqslant p_i^2+2p_ir,$$

establishing that Budowle *et al*'s (1991*a*) estimator always over-estimates the chance occurrence of A_i- phenotypes in the population. Therefore, as long as 'non-detectable' alleles are the predominant source of causing deviation of the observed phenotype frequencies from the HWE predictions, use of $2\tilde{p}_i$ for the probability of A_i- phenotype, and $2\tilde{p}_i\tilde{p}_j$ for the probability of heterozygote A_iA_j should cause no concern for the forensic applications of VNTR typing data.

Even though our discussion thus far focused on clearly distinguished alleles, on a practical level, with binning of quasi-continuous allele size data (Budowle *et al.* 1991*a*) an additional cushion is placed on the bin frequency estimates. Some of the single-band patterns may not be true homozygotes; however, double allelic counts are placed in bins that contain single-band patterns. Those bin frequencies are likely to be over-estimated. In contrast, the bin that contains the 'non-detectable' alleles will have its frequency under-estimated. That is generally the bin which contains the smallest-sized alleles (Budowle *et al.* 1991*a*). Since probability calculations for such bins are generally not done, the under-estimation is of no concern. It might also be noted that Budowle *et al*'s fixed bin method is deliberately an excessively conservative approach unless the heterozygosity approaches 1.

Multiplication of such genotype (phenotype) frequencies over independent segregating loci is also justifiable when the VNTR alleles are non-syntenic, or far apart on a chromosome. Since 'blank' alleles at such alleles should not be co-segregating in a population, this also should be of no concern.

CONCLUSION

The theory discussed here assumes clearly distinguished alleles, although VNTR allelic designations by size-classification of Southern gel banding patterns do not exactly produce discrete alleles. We resorted to an analysis of binned allele data (Budowle *et al.* 1991*a*) to circumvent this problem. These suggest that the deficiency of combined heterozygotes observed in the VNTR polymorphism surveys conducted by Southern gel electrophoresis can be explained simply by the presence of 'non-detectable' alleles. The combined frequencies of such alleles in a population do not have to be large to produce substantial apparent deficiency of heterozygosity. This situation is equivalent to the occurrence of null alleles at protein coding loci (Martin, 1983; Foltz, 1986*a*, *b*; Skibinski *et al.* 1983; Milkman & Beatty, 1970) and their implications in causing deviations from HWE expectations of genotype frequencies have been studied extensively (Gart & Nam, 1984*a*, *b*; Nam & Gart, 1985, 1987). We provide an estimate of the combined frequencies of such 'non-detectable' alleles (r) from the observed proportional heterozygote deficiency (D); it is possible to obtain more refined estimators from the full array of data on all specific phenotypes (Gart & Nam, 1984*b*).

The demonstration that the usual gene-count estimators of allele frequencies ($\tilde{p}_i$) over-estimate the true allele frequencies in the presence of 'non-detectable' alleles is helpful in the forensic context, because liberal over-estimates of actual genotype probabilities can be obtained from them (by the $2\tilde{p}_i$ rule prescribed by Budowle *et al.* 1991*a*), without knowing the true frequencies of 'non-detectable' alleles.

The main emphasis of this work is to demonstrate the possibility of 'non-detectable' alleles as the principal cause of an apparent heterozygote deficiency. A rigorous study has not yet been done on the population databases to attempt to determine whether or not single-band patterns are operationally true homozygotes or pseudo-homozygotes (by using a restriction endonuclease that yields larger DNA fragments than *Hae*III-digested DNA and/or increasing the quantity of DNA analysed). However, there is intuitive and empirical evidence that supports the existence of covert alleles. First, the size of a VNTR fragment generally is dictated by the number of repeat units it contains. Since the probes (used to detect genetic variation at the loci described in this paper) hybridize to the repeat regions, the larger fragment of a heterozygote profile usually is more intense or more readily detectable than the smaller fragment. Due to the quantity of DNA subjected to RFLP analysis, hybridization efficiency, and/or autoradiographic exposure times, it can be anticipated that some alleles will go undetected. In fact, multiple analyses of the same samples have shown that several heterozygote individuals have appeared as single-band homozygotes, the smaller, weaker band being the difference (data not shown). Second, Budowle *et al.* (1991*a*) demonstrated for the D16S85 locus that some *Hae*III-digested DNA showed that single-banded profiles were heterozygotes when digested with the restriction enzyme *Pvu*II. Third, Eisenberg (Texas College of Osteopathic Medicine, personal communication) observed for D2S44 in the Texas Black population that there was a class of small-sized alleles (approximately 300 bp in length) that were difficult to detect by

hybridization since there were very few repeat units within the fragments. Fourth, Jeffreys *et al.* (1991) present other data and cite examples of true 'non-detectable' alleles, although from such initial studies their population frequencies are not precisely known. Fornage *et al.* (1991) also present direct evidence of small-sized alleles at the Apo-CII VNTR locus which are detectable by PCR but would have remained undetected by a traditional RFLP analysis.

It is true that in principle the possibility of heterogeneity within a population (population substructuring) cannot be distinguished from the scenario presented here. However, it can be argued that the presence of 'non-detectable' alleles by size-classifications of Southern gel banding patterns is more plausible. First, if population substructuring is responsible for causing the observed heterozygote deficiency, we should have seen that for other loci as well. Traditional blood groups and protein polymorphisms do not generally reveal such a high degree of heterozygote deficiency. 'Null' alleles are rare for such loci, and therefore it is expected. On the contrary, even if we were to assume that deviations from HWE expectations are difficult to detect (Ward & Sing, 1970; Chakraborty & Rao, 1972) with loci where variability is limited (as in the case of blood groups and protein polymorphisms), hypervariable VNTR loci scored by polymerase chain reaction (PCR)-based protocols do not reveal heterozygote deficiency of the amount shown in Table 2. PCR protocols do not provide any scope of 'non-detectability' and all alleles should be clearly defined by this method. Published population data on PCR-based studies, such as the D1S80 locus studied for Caucasians (Budowle *et al.* 1991*b*) or the Apo-B and Apo-CII VNTR loci studied for US Caucasians and Europeans (Boerwinkle *et al.* 1989; Ludwig *et al.* 1989; Chakraborty *et al.* 1991) exhibit no deviation from HWE frequencies. If population structure was to be an issue, we would have expected heterozygote deficiency even in the case of PCR-based studies. Second, under the hypothesis that population structure is the cause of the observed heterozygote deficiency at VNTR loci (such as the ones shown in Tables 2 and 3), Chakraborty & Jin (1992) have shown that the observed proportional heterozygote deficiency is equivalent to the coefficient of gene differentiation, G_{ST} (Nei, 1973), among subpopulations of a substructured population. This expectation of this coefficient is a composite function of s, the number of subpopulations, and their evolutionary time of divergence measured in units of $2N$ generations ($T = t/2N$, N is the effective size, assumed constant for all subpopulations over their evolution), and the amount of heterozygosity within subpopulations, H (see Nei, 1975). Isolines of G_{ST} for different combinations of s and T computed for different H (Chakraborty & Jin, 1992) suggest that when within-population heterozygosity is of the level $> 70\%$ (as in the case of VNTR loci), G_{ST} of the range of 1–10% can be generated only when $T = t/2N$ is large when s is small, or $t/2N$ is small when s is large. Furthermore, this assumes no gene-flow between populations. Only a small amount of gene-flow substantially retards the accumulation of G_{ST} (Nei & Feldman, 1972; Chakraborty & Nei, 1974). Since the history of US Caucasian and US Black populations shows ample evidence of gene-flow, even in the religiously orthodox communities (Kennedy, 1944; Spuhler & Clark, 1961), we argue that a level of 10% proportional heterozygote deficiency is virtually inconsistent with the hypothesis of population substructure being the cause of the deficiency. On the other hand, 6% frequency of 'non-detectable' alleles quite reasonably explains 10% proportional heterozygote deficiency.

In principle, measuremental errors of band sizes may still affect accurate classification of alleles by a fixed bin approach (Budowle *et al.* 1991*a*). Quality control experiments (data not shown) suggest that this is not a critical concern, because sizing errors are small for small size

alleles where bin widths are narrow, and wide width of the large size bins can easily encompass up to 5% sizing error. In such cases, the $\tilde{p}_i$ value for the adjacent bins that have higher allele frequency should be used. Since the $2\tilde{p}_i$ already gives enough cushion in overestimating the actual genotype frequency for single band profiles, Budowle *et al*'s (1991*a*) liberal suggestion is more than sufficient to encompass sizing errors so that the detailed treatment of measuremental errors with additional assumptions, such as the ones suggested by Devlin *et al.* (1990), are not critically needed.

This work was supported by a grant 90-IJ-CX-0038 from the National Institute of Justice. The conclusions reached in this work, however, are solely the opinion of the authors and are not endorsements of the granting agencies supporting this research.

REFERENCES

BALLANTINE, J., SENSABAUGH, G. & WITKOWSKI, J. (1989). DNA technology and forensic science. Banbury Report 32. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

BOERWINKLE, E., XIONG, W., FOUREST, E. & CHAN, L. (1989). Rapid typing of tandemly repeated hypervariable loci by polymerase chain reaction: application to the apolipoprotein B3' hypervariable region. *Proc. Natl Acad. Sci. USA* **86**, 212–216.

BUDOWLE, B., GIUSTI, A. M., WAYE, J. S., BAECHTEL, F. S., FOURNEY, R. M., ADAMS, D. E., PRESLEY, L. A., DEADMAN, H. A. & MONSON, K. L. (1991*a*). Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic comparisons. *Am. J. Hum. Genet.* **48**, 841–855.

BUDOWLE, B., CHAKRABORTY, R., GIUSTI, A. M., EISENBERG, A. E. & ALLEN, R. C. (1991*b*). Analysis of the VNTR locus D1S80 by PCR followed by high-resolution PAGE. *Am. J. Hum. Genet.* **48**, 137–144.

CHAKRABORTY, R. & RAO, D. C. (1972). On the detection of F from ABO blood group data. *Am. J. Hum. Genet.* **24**, 352–353.

CHAKRABORTY, R. & NEI, M. (1974). Dynamics of gene differentiation between incompletely isolated populations of unequal sizes. *Theor. Pop. Biol.* **5**, 460–469.

CHAKRABORTY, R. & DAIGER, S. P. (1991). Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah. *Hum. Biol.* **63**, 571–587.

CHAKRABORTY, R., FORNAGE, M., GUEGUEN, R. & BOERWINKLE, E. (1991). Population genetics of hypervariable loci: analysis of PCR based VNTR polymorphism within a population. In *DNA Fingerprinting: Approaches and Applications* (ed. T. Burke, G. Dolf, A. J. Jeffreys and R. Wolff), pp. 127–143. Berne: Birkhäuer Verlag.

CHAKRABORTY, R. & JIN, L. (1992). Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. *Hum. Genet.* (in press).

COHEN, J. E. (1990). DNA fingerprinting for forensic identification: potential effects of data interpretation of subpopulation heterogeneity and band number variability. *Am. J. Hum. Genet.* **46**, 358–368.

DEVLIN, B., RISCH, N. & ROEDER, K. (1990). No excess of homozygosity at loci used for DNA fingerprinting. *Science* **249**, 1416–1420.

FOLTZ, D. W. (1986*a*). Segregation and linkage studies of allozyme loci in pair crosses of the oyster *Crassostrea virginica*. *Biochem. Genet.* **24**, 941–956.

FOLTZ, D. W. (1986*b*). Null alleles as a possible cause of heterozygote deficiencies in the oyster *Crassostrea virginica* and other bivalves. *Evolution* **40**, 869–870.

FORNAGE, M., CHAN, L., SIEST, G. & BOERWINKLE, E. (1991). Frequency distribution of $[TG]_n$ $[AG]_m$ minisatellite in the apolipoprotein c-II gene. *Genomics* (in the Press).

GART, J. J. & NAM, J. (1984*a*). A score test for the possible presence of recessive alleles in generalized ABO-like genetic systems. *Biometrics* **40**, 887–894.

GART, J. J. & NAM, J. (1984*b*). Statistical methods for genetic studies of HLA and cancer. In *Statistical Methods for Cancer Studies* (ed. R. G. Cornell), pp. 229–266. New York: Marcel Decker.

JEFFREYS, A. J., WILSON, V. & THEIR, S. L. (1985). Individual specific 'fingerprints' of human DNA. *Nature* **316**, 76–79.

JEFFREYS, A. J., ROYLE, N. J., PATEL, I., ARMOUR, J. A. L., MACLEOD, A., COLLICK, A., GRAY, I. C., NEUMANN, R., GIBBS, M., CROSIER, M., HILL, M., SIGNER, E. & MONCKTON, D. (1991). Principles and recent advances in human DNA fingerprinting. In *DNA Fingerprinting: Approaches and Applications* (ed. T. Burke, G. Dolf, A. J. Jeffreys and R. Wolff), pp. 1–19. Berlin: Birkhäuer Verlag.

KENNEDY, R. J. R. (1944). Single or triple melting pot? Intermarriage trends in New Haven, 1870–1940. *Am. J. Sociology* **49**, 331–339.

LANDER, E. (1989). DNA fingerprinting on trial. *Nature* **339**, 501–505.

LANDER, E. (1991). Invited editorial: research on DNA typing catching up with courtroom application. *Am. J. Hum. Genet.* **48**, 819–823.

LUDWIG, E. H., FRIEDL, W. & MCCARTHY, B. J. (1989). High-resolution analysis of a hypervariable region in the human apolipoprotein B gene. *Am. J. Hum. Genet.* **45**, 458–464.

MARTIN, W. (1983). Consideration of 'silent genes' in the statistical evaluation of blood group findings in paternity testing. In *Inclusion Probabilities in Parentage Testing* (ed. R. H. Walker), pp. 245–247. Arlington, VA: American Association of Blood Banks.

MILKMAN, R. & BEATTY, L. D. (1970). Large-scale electrophoretic studies of allelic variation in *Mytilus edulis*. *Biol. Bull.* **139**, 430.

NAKAMURA, Y., LEPPERT, M., O'CONNELL, P., WOLFF, R., HOLM, T., CULVER, M., MARTIN, C., FUJIMOTO, E., HOFF, M., KUMLIN, E. & WHITE, R. L. (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**, 1616–1622.

NAM, J. & GART, J. J. (1985). The ML estimation and testing of generalized ABO-like data with no observed double recessives. *Biometrics* **41**, 455–466.

NAM, J. & GART, J. J. (1987). On two tests of fit for HLA data with no double blanks. *Am. J. Hum. Genet.* **41**, 70–76.

NEI, M. & FELDMAN, M. W. (1972). Identity of genes by descent within and between populations under mutation and migration pressures. *Theor. Pop. Biol.* **3**, 460–465.

NEI, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl Acad. Sci. USA* **70**, 3321–3323.

NEI, M. (1975). *Molecular Population Genetics and Evolution*. Amsterdam: North-Holland.

SKIBINSKI, D. O. F., BEARDMORE, J. A. & CROSS, T. F. (1983). Aspects of the population genetics of *Mytilus* (Mytilidae; Mollusca) in the British Isles. *Biol. J. Linn. Soc.* **10**, 137–183.

SPUHLER, J. N. & CLARK, P. J. (1961). Migration into the human breeding population of Ann Harbor, Michigan, 1900–1950. *Hum. Biol.* **33**, 223–231.

THOMPSON, W. C. & FORD, S. (1989). DNA typing: acceptance and weight of the new genetic identification tests. *Virginia Law Review* **75**, 45–108.

WARD, R. H. & SING, C. F. (1970). A consideration of the power of the chi^2-test to detect inbreeding effects in natural populations. *Amer. Nat.* **104**, 355–363.

WYMAN, A. R. & WHITE, R. (1980). A highly polymorphic locus in human DNA. *Proc. Natl Acad. Sci. USA* **77**, 6754–6758.

Ann. Hum. Genet. (1992), **56**, 45–57
Printed in Great Britain

Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications

R. CHAKRABORTY[1]*, M. DE ANDRADE[1], S. P. DAIGER[2] AND B. BUDOWLE[3]

[1] *Center for Demographic and Population Genetics and*
[2] *Medical Genetics Center, University of Texas Graduate School of Biomedical Sciences, Houston, Texas*
[3] *Forensic Science Research and Training Center, Laboratory Division, FBI Academy, Quantico, Virginia*

SUMMARY

Restriction fragment length polymorphisms (RFLP) analysis using the Southern blot technique can be used to recognize copy number variation of variable number of tandem repeats (VNTR) of conserved core sequences at several regions of the human genome. This new class of polymorphisms reveals a high degree of genetic variation, useful for individual identification purposes. Criticisms against forensic applications of such DNA typing data include the limitation of employing Hardy–Weinberg expectation of genotype frequencies, since several surveys indicate apparent deficiency of heterozygosity (or excess homozygosity) in comparison with Hardy–Weinberg expectations. This research postulates an alternative explanation of deficiency of apparent heterozygosity which is caused by the inability to detect extremely small-sized alleles (called 'non-detectable' alleles) due to the sensitivity of Southern gel electrophoresis. We show that the presence of 'non-detectable' alleles can produce pseudo-homozygosity and their frequencies can be predicted from the observed proportional heterozygote deficiency. Furthermore, in the covert presence of such 'non-detectable' alleles, we show that the gene-count method provides over-estimates of allele frequencies in the sample population, and hence the Hardy–Weinberg predictions of genotype frequencies avoid wrongful bias against suspects in forensic applications of DNA typing data. Applications of this theory to population data on six VNTR loci in US Caucasians and US Blacks suggest that the presence of 'non-detectable' alleles could be the major cause of apparent heterozygote deficiency, and the current approaches of predicting the population frequency of specific DNA phenotypes are practically free of the possible wrongful bias in courtroom applications of DNA typing data.

INTRODUCTION

Scientific as well as social implications of the discovery of hypervariable VNTR loci in the human genome are by now well-recognized (Wyman & White, 1980; Jeffreys *et al.* 1985; Nakamura *et al.* 1987; Ballantyne *et al.* 1989). While the general criteria of the presence of large numbers of alleles and high heterozygosities at most VNTR loci make them ideal candidates for genetic 'fingerprinting' of individuals, ensuing controversies (Lander, 1989, 1991; Thompson & Ford, 1989; Cohen, 1990) with regard to their forensic applications prompted careful attention

* Correspondence to: Ranajit Chakraborty, Ph.D., Center for Demographic and Population Genetics, University of Texas, Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, TX 77225, U.S.A.

to various statistical as well as population genetic characteristics of such hypervariable loci (Devlin *et al.* 1990; Chakraborty & Daiger, 1991; Chakraborty *et al.* 1991). In a realistic as well as rigorous study, Devlin *et al.* (1990) demonstrated that the quasi-continuous variations of allele sizes at many VNTR loci could be produced at least in part by measuremental errors of allele-size determination. Individuals who exhibit both alleles of similar (but different) sizes may be wrongly typed as homozygotes, and therefore the population data may indicate evidence of heterozygote deficiency (or equivalently, excess of homozygosity) in comparison with the predictions of Hardy–Weinberg equilibrium (HWE) law. Since most genotypes at VNTR are rare, it has been demonstrated that by necessity genotype frequencies at VNTR loci best be predicted from their allele frequencies (Chakraborty *et al.* 1991). As a matter of fact, for forensic applications it may be enough to prescribe an upper bound for every genotype probability, since this will avoid wrongful bias in criminal offense cases (Ballantyne *et al.* 1989; Budowle *et al.* 1991*a*). Since under the assumption of HWE the probability of observing a pair of alleles constituting an individual's genotype is given by the product of the respective alleles in the population or multiplying this product by a factor of two if the alleles are dissimilar, it may also be called a 'product rule' of genotype probability determination. The most common cause of deviation from the above product rule (in the direction of excess homozygozity) observed in the content of the VNTR polymorphism is claimed to be population subdivision (Lander, 1989; Cohen, 1990). The purpose of this communication is to examine the plausibility of another important 'technical' causal factor, the non-detectability of a class of small-sized alleles. This is analytically equivalent to the problem of 'null' alleles that exist in human and other organisms at protein-coding loci (Martin, 1983; Foltz, 1986*a*, *b*). The objectives of this research are: (1) to show that the presence of such 'non-detectable' alleles produces pseudo-homozygosity, and their frequencies in a population can be predicted from the proportional heterozygote deficiency, although other rigorous methods of estimation of such allele frequencies are available in the literature (Gart & Nam, 1984*a*); (2) to demonstrate that their presence may remain covert, although there might be considerable numbers of heterozygote individuals in the sample, where such 'non-detectable' alleles are found in combination with other alleles so that such individuals are observed as apparent homozygotes (for the detected alleles) because of their single-band Southern-gel profiles; and finally, (3) to show that in the presence of such alleles, the gene-count method of estimation of allele frequencies over-estimates the actual frequencies of all detectable alleles, and hence the product-rule gives enough cushion in the prediction of genotype probabilities, avoiding bias against suspects in forensic applications of DNA typing data.

Analytical explorations of this technical problem is followed by an analysis of population data on six VNTR loci (D2S44, D14S13, D4S139, D17S79, D1S7 and D16S85) in US Caucasians and US Blacks, classified by the fixed-bin approach suggested earlier (Budowle *et al.* 1991*a*). It has been shown earlier that the fixed-bin approach of allele classification greatly circumvents the problem of measuremental error of allele size determination, and hence analytical strategies of detecting pseudo-homozygosity by the approach of Devlin *et al.* (1990) do not have to be involved in this study (Budowle *et al.* 1991*a*).

FREQUENCY CONSEQUENCES OF 'NON-DETECTABLE' ALLELES

Consider the case of k detectable alleles ($A_1, A_2, \ldots, A_k$) and a class of 'non-detectable' alleles (i.e. alleles of too small sizes) that may truly contain alleles of dissimilar sizes, each of which remains undetected on a Southern gel. We designate this class of alleles by A_0. This scenario is reminiscent of the HLA system in humans, where A_0 is the blank allele and A_is are the ones detected by the specific HLA-allelic antisera. We can therefore designate the different genotype and phenotype frequencies and their respective probabilities under HWE as shown in Table 1, where p_i is the frequency of the allele A_i ($i = 1, 2, \ldots, k$) and r is the frequency of 'non-detectable' class of alleles (A_0).

The HWE probabilities of the observed phenotypes, conditioned on the observed absence of homozygotes for the 'non-detectable' alleles (A_0, A_0), then, become

$$Pr(A_i-) = \frac{(p_i^2+2p_i r)}{(1-r^2)}, \quad \text{for } i = 1,2,\ldots,k. \tag{1}$$

and

$$Pr(A_iA_j) = \frac{2p_i p_j}{(1-r^2)}, \quad \text{for } j > i = 1,2,\ldots,k. \tag{2}$$

In contrast, if we ignore the existence of 'non-detectable' alleles altogether, we would regard each A_i- phenotype as the true homozygote A_iA_i (for $i = 1, 2, \ldots, k$) and A_iA_j phenotype as heterozygote, so that the gene-count estimators (which are also the maximum likelihood estimators) of allele frequencies are

$$\tilde{p}_i = \frac{2n_{i-}+\Sigma_{i\neq j} n_{ij}}{2n}, \quad \text{for } i = 1,2,\ldots,k. \tag{3}$$

where $n = \Sigma_{i=1}^{k} n_{i-} + \Sigma_{j>i=1}^{k} n_{ij}$.

The expectations of $\tilde{p}_i$ in the conditional data set, in the presence of A_0 allele(s) are

$$\begin{aligned} E(\tilde{p}_i) &= \frac{2n[p_i^2+2p_i r]+2np_i\Sigma_{i\neq j}p_j}{2n(1-r^2)} \\ &= \frac{p_i^2+2p_i r+p_i(1-p_i-r)}{1-r^2} \\ &= \frac{p_i(1+r)}{1-r^2} \\ &= \frac{p_i}{1-r} \end{aligned} \tag{4}$$

for $i = 1, 2, \ldots, k$.

For $r > 0$, obviously $\tilde{p}_i$ is an over-estimate of the true allele frequency, p_i, in the sampled population.

Ignoring the non-detectable alleles (A_0), one would assert that under HWE the expected heterozygosity, H_E, is

$$H_E = \sum_{i\neq j} \tilde{p}_i\tilde{p}_j = 1-\sum_{i=1}^{k} \tilde{p}_i^2, \tag{5}$$

Table 1. *Genotypes, phenotypes and frequencies of a VNTR locus in the presence of 'non-detectable' alleles*

Genotype	Phenotype	Observed frequency	Probability under HWE
A_iA_i	A_i-	n_i	$p_i^2+2p_ir$ for $i = 1, 2, \ldots, k$
A_iA_j	A_iA_j	n_{ij}	$2p_ip_j$ for $i<j = 1, 2, \ldots, k$
A_0A_0	Blank	n_{00}	r^2

Note: A_i- phenotype appears as a single-band lane on a Southern gel (for $i = 1, 2, \ldots, k$), the size of which can be detected on a control ladder-lane.

A_iA_j phenotype (genotype) is the unequivocal two-banded lane, sizes of both alleles being detectable.

Blank phenotypes are rare, showing no band in the lane; often ignored because they can also be caused by insufficient DNA or other factors (such as DNA degradation, etc.).

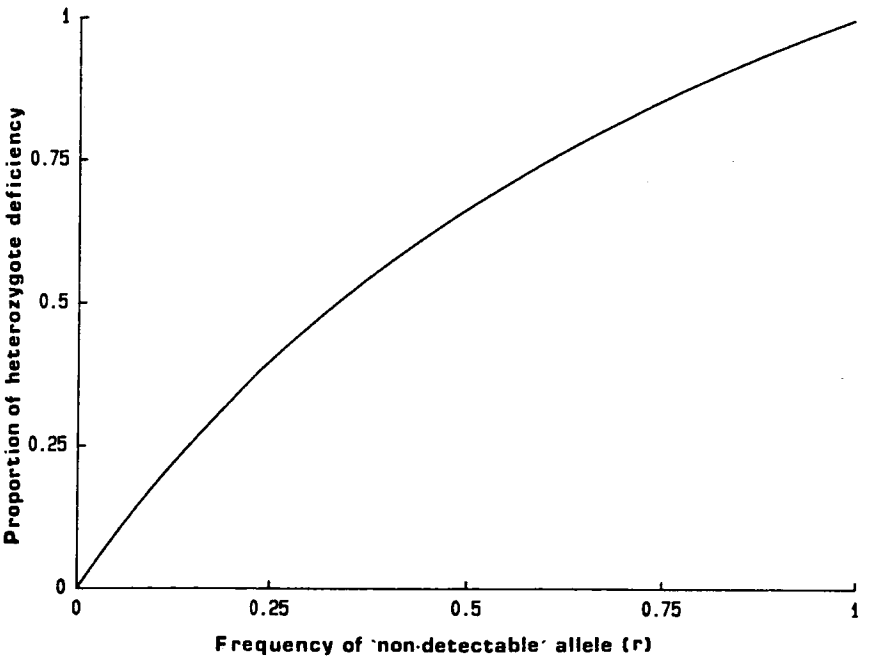


Fig. 1. Proportional heterozygote deficiency as a function of the frequency of 'non-detectable' alleles.

which would be contrasted with the observed heterozygosity, H_O, given by

$$H_O = \frac{\Sigma_{i\neq j} p_i p_j}{1-r^2}. \tag{6}$$

Note that $1-r^2 = (1-r)(1+r) > (1-r)^2$, hence from equation (6) we have

$$H_O < \sum_{i\neq j}\left(\frac{p_i}{1-r}\right)\left(\frac{p_j}{1-r}\right) = \sum_{i\neq j}\tilde{p}_i\tilde{p}_j$$

$$= 1-\sum_{i=1}^{k}\tilde{p}_i^2 = H_E, \tag{7}$$

showing that in conditional data (with homozygosity for non-detectable alleles absent), non-detectable alleles can produce an observed apparent heterozygote deficiency (or equivalently, excess homozygosity) in comparison with the HWE prediction.

Furthermore, invoking (4) in (5) and (6) we get the expected heterozygote deficiency proportional to expected heterozygosity

$$D = \frac{H_E - H_O}{H_E} = \frac{2r}{1+r}, \tag{8}$$

since

$$\begin{aligned} H_O - H_E &= \frac{\Sigma_{i \neq j} p_i p_j}{1-r^2} - \sum_{i \neq j} \tilde{p}_i \tilde{p}_j \\ &= \sum_{i \neq j} p_i p_j \left[\frac{1}{1-r^2} - \frac{1}{(1-r)^2} \right] \\ &= \sum_{i \neq j} p_i p_j \left[\frac{(1+r)-(1-r)}{(1+r)(1-r)^2} \right] \\ &= -\frac{2r}{1+r} \sum_{i \neq j} \tilde{p}_i \tilde{p}_j. \end{aligned} \tag{9}$$

Therefore, the expected frequency of the non-detectable alleles (r) that can produce an observed level of proportional heterozygote deficiency (D) can be evaluated as

$$r_* = \frac{D}{(2-D)}. \tag{10}$$

Figure 1 shows the value of D as function of r. The concavity of the curve indicates that the proportional heterozygote deficiency (D) is at least as large as r (since $D \geqslant r$, the equality holding only in the extreme cases when $r = 0$ or 1). In other words, even a rare occurrence of non-detectable allele would produce a noticeable amount of heterozygote deficiency, particularly when the expected heterozygosity at the locus is 70% or greater (which is usually the case with most VNTR loci).

COVERTNESS OF NON-DETECTABLE ALLELES

The algebra of the preceding section demonstrates that the existence of rare 'non-detectable' alleles could produce an appreciable amount of observed heterozygote deficiency (or, excess homozygosity); the frequency of such alleles can be predicted from the observed proportional heterozygote deficiency (D). One might then ask, if this were the case why do such alleles remain covert; i.e. not seen in homozygote form? There is the possibility that a blank lane on a Southern gel might often be attributed to other technical problems, such as insufficient DNA and/or DNA degradation. Therefore, even if 'non-detectable' alleles appear in homozygous form they would not necessarily be scored. Further reasoning on statistical grounds can be given from the expected frequency of homozygosity for the non-detectable alleles (nr^2) and the probability of observing at least one homozygote of the type for a given sample size.

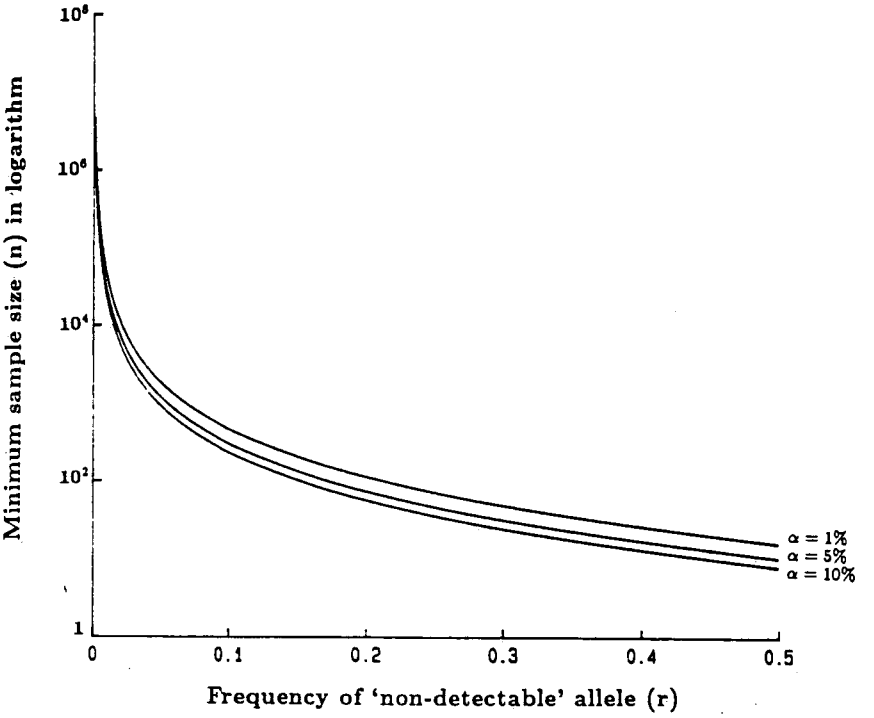


Fig. 2. Minimum sample size (n) needed to observe homozygotes for 'non-detectable' alleles as functions of frequency of such alleles (r) for different levels of significance (α).

The expected number of homozygotes for the non-detectable alleles is

$$F = nr_*^2 = \frac{nD^2}{(2-D)^2}, \tag{11}$$

which can remain very low even if D is large. For example, for $D = 0{\cdot}10$ F becomes 1 in a sample of 100 individuals. 5 in 500 individuals. Therefore, there is a high chance that such individuals might be missed in a survey. The probability that no such homozygotes would be seen in a sample of n individuals is given by

$$P_0(n) = (1-r_*^2)^n, \tag{12}$$

from a simple binomial distribution treatment.

Alternatively, the sample size needed (n) to observe at least one such homozygote in a survey with a confidence of $100(1-\alpha)\%$ must satisfy the inequality

$$1-(1-r^2)^n \geqslant 1-\alpha,$$

or
$$(1-r^2)^n \leqslant \alpha,$$

or
$$n \log(1-r^2) \leqslant \log(\alpha),$$

or
$$n \geqslant \frac{\log(\alpha)}{\log(1-r^2)}. \tag{13}$$

Table 2. *Observed heterozygote deficiencies at six VNTR loci in US Caucasians and Blacks and the frequencies of 'non-detectable' alleles that explain these deficiencies*

Locus	Sample size (n)	Obs. freq. Homo.	Obs. freq. Hetero.	Exp. freq. Homo.	Exp. freq. Hetero.	Prop. hetero. def. (D) in %	Pred. value of r in %
			US Caucasians				
D2S44	218	19	199	17·49	200·51	0·753	0·378
D14S13	218	22	196	19·18	198·82	1·418	0·714
D4S139	144	18	126	14·26	129·74	2·883	1·462
D17S79	209	61	148**	43·17	165·83	10·752	5·681
D1S7	210	19	191*	12·07	197·93	3·501	0·416
D16S85	210	21	189	19·42	190·58	0·829	0·416
			US Blacks				
D2S44	295	35	260**	19·41	275·59	5·657	2·911
D14S13	258	24	234*	15·46	242·54	3·521	1·792
D4S139	304	28	276	22·46	281·54	1·968	0·994
D17S79	281	54	227*	39·76	241·24	5·903	3·041
D1S7	268	23	245*	15·24	252·76	3·070	1·559
D16S85	212	47	165*	34·53	177·47	7·027	3·641

*$P < 0·05$; **$P < 0·01$ by a χ^2 test with 1 D.F.

Figure 2 shows plots of such minimum sample size requirement $n = \log(\alpha)/\log(1-r^2)$, when n is plotted in logarithmic scale as function of r for $\alpha = 0·01$, 0·05 and 0·10. Clearly, the sample size requirement is rather stringent. For example, for $r = 0·10$ [which would produce almost 18% proportional heterozygote deficiency, see equation (13)] we would need at least 1000 individuals to be screened before encountering at least one 'non-detectable' homozygote genotype with a confidence of 95%. When $r = 0·01$, the 95% confidence minimum sample size for observing at least one such individual is 29995. It is no surprise, therefore, that 'non-detectable' alleles remain covert in samples observed.

ANALYSIS OF SIX VNTR LOCI IN US CAUCASIANS AND US BLACKS

Budowle *et al.* (1991*a*) recently presented the analysis of phenotypes at six VNTR loci (D1S7, D2S44, D4S139, D14S13, D16S85 and D17S79) for the US Caucasians and US Blacks that exist in the FBI data base. Fixed bin approach (Budowle *et al.* 1991*a*) gives allele counts of binned alleles from which concordance of the observed numbers of heterozygotes (or homozygotes) with their HWE predictions can be examined (see Table 9 or Budowle *et al.* 1991*a*). Table 2 presents the summary of their computations. At each locus heterozygote deficiency is noticed in both populations, and in several cases (e.g. D1S7 and D17S79 in US Caucasians, and at all loci except the D4S139 locus in US Blacks) the observed heterozygote deficiencies are significant (at least 5% level). Ordinarily, one would have asserted the presence of substantial population substructuring in both populations from such data. The computations of Table 2 indicate that the heterozygote deficiency observed at each locus in this data can be easily explained by the nominal presence of 'non-detectable' alleles. Even the most deviant locus (D17S79 in the US Caucasians) exhibiting more than 10% proportional heterozygote deficiency could be regarded

Table 3. *Sample size needed to observe at least one 'non-detectable' homozygote*

Locus	Sample size (n)	Prop. hetero. def. (D) in %	Exp. freq. of homo. non-detec. alleles (nr^2)	Sample size required to observe at least one non-detec. homo. for $\alpha = 0{\cdot}01$	$\alpha = 0{\cdot}05$	$\alpha = 0{\cdot}10$
			US Caucasians			
D2S44	218	0·753	0·003	322365	209703	161183
D14S13	218	1·418	0·011	90268	58721	45134
D4S139	144	2·883	0·031	21531	14006	10765
D17S79	209	10·752	0·675	1425	927	713
D1S7	210	3·501	0·067	14503	9435	7252
D16S85	210	0·829	0·004	265787	172899	132894
			US Blacks			
D2S44	295	5·657	0·250	5433	3535	2717
D14S13	258	3·521	0·083	14337	9327	7169
D4S139	304	1·968	0·030	46640	30340	23320
D17S79	281	5·903	0·260	4977	3238	2489
D1S7	268	3·070	0·065	18946	12325	9475
D16S85	212	7·027	0·281	3472	2259	1726

as satisfying the HWE predictions with 'non-detectable' alleles that have a combined frequency below 6%. Frequencies of such alleles required to explain the other observed heterozygote deficiencies are even smaller (see Table 2, last column).

The supposition that these frequencies (r) of 'non-detectable' alleles are reasonable can be judged from the computation presented in Table 3. We note that with these predicted r values, the expected frequency of homozygotes for 'non-detectable' alleles in this database does not exceed one in any case (the maximum value of nr^2 is 0·675 for the D17S79 locus for the US Caucasian sample), and hence it is not a surprise that no such homozygote was detected. The sample size required to observe at least one such homozygote is also very large, as shown in the last three columns of Table 3. For example, even in the worst case (D17S79 in Caucasians) we would have needed 713 individuals to be 90% sure of getting at least one 'blank' phenotype. We may therefore conclude that even the case of a substantial heterozygote deficiency can be explained easily by covertness of 'non-detectable' alleles due to Southern gel electrophoresis.

IMPLICATIONS FOR FORENSIC APPLICATIONS OF VNTR TYPING DATA

As mentioned before, VNTR polymorphisms are useful for identification of individuals because the number of possible genotypes at VNTR loci is large, and most of them are rare enough in a population, so that the probability that two individuals will have identical genotypes (particularly for multiple numbers of such loci) by chance alone is very low. The rareness of the genotypes, however, poses a limitation, in the sense that the chance occurrence of genotypes must be evaluated from a population data base using population genetics principles such as the HWE (or the product rule). The theory and data analysis presented above suggest that in the presence of 'non-detectable' alleles such a rule of estimating genotype frequencies from allele-frequency data should not apply, because the combined frequency of

heterozygotes will fall below the predicted one. We argue that this should not be of any concern in the use of such allele frequency data for forensic application, because all that is necessary to establish that the chance occurrence of a specific genotype is rare is to prescribe an upper bound for the frequency of every conceivable genotype. If 'non-detectable' alleles are the predominant source of observed heterozygote deficiency, then gene-count estimates of all detectable alleles provide enough cushion to prescribe such upper bounds.

For example, if the Southern gel profile for a particular VNTR locus depicts genotype A_iA_j (bands of size i and j are detected), the true probability of this genotype in the population (under HWE) is $2p_ip_j$, but based on the allele-count data on conditional data (no 'blank' homozygote found) is $2\tilde{p}_i\tilde{p}_j$ ($\tilde{p}_i$ is estimated from equation (3)).

Note that

$$E(2\tilde{p}_i\tilde{p}_j) = \frac{2p_ip_j}{(1-r)^2}\left[1-\frac{1}{2n}\right]$$

and hence

$$E\left(\frac{2(2n-1)\,\tilde{p}_i\tilde{p}_j}{2n}\right) = \frac{2p_ip_j}{(1-r)^2} \geqslant 2p_ip_j$$

so that $2(2n-1)\,\tilde{p}_i\tilde{p}_j/(2n)$ is an over-estimate of the true frequency of the genotype A_iA_j in the population. When n is larger than 100, $2\tilde{p}_i\tilde{p}_j$ should reasonably serve as an adequate upper bound for the true heterozygote frequency in the population.

For the apparent homozygotes (single-band pattern of type A_i) the situation is somewhat more involved. Since the actual probability of the A_i- phenotype in the population is $p_i^2+2p_ir$ (under HWE in the presence of 'non-detectable' alleles), we cannot guarantee that $\tilde{p}_i^2$ will always be an over-estimate of $p_i^2+2p_ir$. However, Budowle *et al.* (1991*a*) suggested $2\tilde{p}_i$ as an estimator of the probability of A_i- phenotype in the population. Now note that $E(\tilde{p}_i) = p_i/(1-r)$, and $p_i^2+2p_ir = p_i(p_i+2r) < 2p_i(p_i+r) \leqslant 2p_i$, since $p_i+r \leqslant 1$. Therefore,

$$E(2\tilde{p}_i) \geqslant p_i^2+2p_ir,$$

establishing that Budowle *et al*'s (1991*a*) estimator always over-estimates the chance occurrence of A_i- phenotypes in the population. Therefore, as long as 'non-detectable' alleles are the predominant source of causing deviation of the observed phenotype frequencies from the HWE predictions, use of $2\tilde{p}_i$ for the probability of A_i- phenotype, and $2\tilde{p}_i\tilde{p}_j$ for the probability of heterozygote A_iA_j should cause no concern for the forensic applications of VNTR typing data.

Even though our discussion thus far focused on clearly distinguished alleles, on a practical level, with binning of quasi-continuous allele size data (Budowle *et al.* 1991*a*) an additional cushion is placed on the bin frequency estimates. Some of the single-band patterns may not be true homozygotes; however, double allelic counts are placed in bins that contain single-band patterns. Those bin frequencies are likely to be over-estimated. In contrast, the bin that contains the 'non-detectable' alleles will have its frequency under-estimated. That is generally the bin which contains the smallest-sized alleles (Budowle *et al.* 1991*a*). Since probability calculations for such bins are generally not done, the under-estimation is of no concern. It might also be noted that Budowle *et al*'s fixed bin method is deliberately an excessively conservative approach unless the heterozygosity approaches 1.

Multiplication of such genotype (phenotype) frequencies over independent segregating loci is also justifiable when the VNTR alleles are non-syntenic, or far apart on a chromosome. Since 'blank' alleles at such alleles should not be co-segregating in a population, this also should be of no concern.

CONCLUSION

The theory discussed here assumes clearly distinguished alleles, although VNTR allelic designations by size-classification of Southern gel banding patterns do not exactly produce discrete alleles. We resorted to an analysis of binned allele data (Budowle *et al.* 1991*a*) to circumvent this problem. These suggest that the deficiency of combined heterozygotes observed in the VNTR polymorphism surveys conducted by Southern gel electrophoresis can be explained simply by the presence of 'non-detectable' alleles. The combined frequencies of such alleles in a population do not have to be large to produce substantial apparent deficiency of heterozygosity. This situation is equivalent to the occurrence of null alleles at protein coding loci (Martin, 1983; Foltz, 1986*a*, *b*; Skibinski *et al.* 1983; Milkman & Beatty, 1970) and their implications in causing deviations from HWE expectations of genotype frequencies have been studied extensively (Gart & Nam, 1984*a*, *b*; Nam & Gart, 1985, 1987). We provide an estimate of the combined frequencies of such 'non-detectable' alleles (r) from the observed proportional heterozygote deficiency (D): it is possible to obtain more refined estimators from the full array of data on all specific phenotypes (Gart & Nam, 1984*b*).

The demonstration that the usual gene-count estimators of allele frequencies ($\tilde{p}_i$) over-estimate the true allele frequencies in the presence of 'non-detectable' alleles is helpful in the forensic context, because liberal over-estimates of actual genotype probabilities can be obtained from them (by the $2\tilde{p}_i$ rule prescribed by Budowle *et al.* 1991*a*), without knowing the true frequencies of 'non-detectable' alleles.

The main emphasis of this work is to demonstrate the possibility of 'non-detectable' alleles as the principal cause of an apparent heterozygote deficiency. A rigorous study has not yet been done on the population databases to attempt to determine whether or not single-band patterns are operationally true homozygotes or pseudo-homozygotes (by using a restriction endonuclease that yields larger DNA fragments than *Hae*III-digested DNA and/or increasing the quantity of DNA analysed). However, there is intuitive and empirical evidence that supports the existence of covert alleles. First, the size of a VNTR fragment generally is dictated by the number of repeat units it contains. Since the probes (used to detect genetic variation at the loci described in this paper) hybridize to the repeat regions, the larger fragment of a heterozygote profile usually is more intense or more readily detectable than the smaller fragment. Due to the quantity of DNA subjected to RFLP analysis, hybridization efficiency, and/or autoradiographic exposure times, it can be anticipated that some alleles will go undetected. In fact, multiple analyses of the same samples have shown that several heterozygote individuals have appeared as single-band homozygotes, the smaller, weaker band being the difference (data not shown). Second, Budowle *et al.* (1991*a*) demonstrated for the D16S85 locus that some *Hae*III-digested DNA showed that single-banded profiles were heterozygotes when digested with the restriction enzyme *Pvu*II. Third, Eisenberg (Texas College of Osteopathic Medicine, personal communication) observed for D2S44 in the Texas Black population that there was a class of small-sized alleles (approximately 300 bp in length) that were difficult to detect by

hybridization since there were very few repeat units within the fragments. Fourth, Jeffreys *et al.* (1991) present other data and cite examples of true 'non-detectable' alleles, although from such initial studies their population frequencies are not precisely known. Fornage *et al.* (1991) also present direct evidence of small-sized alleles at the Apo-CII VNTR locus which are detectable by PCR but would have remained undetected by a traditional RFLP analysis.

It is true that in principle the possibility of heterogeneity within a population (population substructuring) cannot be distinguished from the scenario presented here. However, it can be argued that the presence of 'non-detectable' alleles by size-classifications of Southern gel banding patterns is more plausible. First, if population substructuring is responsible for causing the observed heterozygote deficiency, we should have seen that for other loci as well. Traditional blood groups and protein polymorphisms do not generally reveal such a high degree of heterozygote deficiency. 'Null' alleles are rare for such loci, and therefore it is expected. On the contrary, even if we were to assume that deviations from HWE expectations are difficult to detect (Ward & Sing, 1970; Chakraborty & Rao, 1972) with loci where variability is limited (as in the case of blood groups and protein polymorphisms), hypervariable VNTR loci scored by polymerase chain reaction (PCR)-based protocols do not reveal heterozygote deficiency of the amount shown in Table 2. PCR protocols do not provide any scope of 'non-detectability' and all alleles should be clearly defined by this method. Published population data on PCR-based studies, such as the D1S80 locus studied for Caucasians (Budowle *et al.* 1991*b*) or the Apo-B and Apo-CII VNTR loci studied for US Caucasians and Europeans (Boerwinkle *et al.* 1989; Ludwig *et al.* 1989; Chakraborty *et al.* 1991) exhibit no deviation from HWE frequencies. If population structure was to be an issue, we would have expected heterozygote deficiency even in the case of PCR-based studies. Second, under the hypothesis that population structure is the cause of the observed heterozygote deficiency at VNTR loci (such as the ones shown in Tables 2 and 3), Chakraborty & Jin (1992) have shown that the observed proportional heterozygote deficiency is equivalent to the coefficient of gene differentiation, G_{ST} (Nei, 1973). among subpopulations of a substructured population. This expectation of this coefficient is a composite function of s, the number of subpopulations, and their evolutionary time of divergence measured in units of $2N$ generations ($T = t/2N$, N is the effective size, assumed constant for all subpopulations over their evolution), and the amount of heterozygosity within subpopulations, H (see Nei, 1975). Isolines of G_{ST} for different combinations of s and T computed for different H (Chakraborty & Jin, 1992) suggest that when within-population heterozygosity is of the level $> 70\%$ (as in the case of VNTR loci), G_{ST} of the range of 1–10% can be generated only when $T = t/2N$ is large when s is small, or $t/2N$ is small when s is large. Furthermore, this assumes no gene-flow between populations. Only a small amount of gene-flow substantially retards the accumulation of G_{ST} (Nei & Feldman, 1972; Chakraborty & Nei, 1974). Since the history of US Caucasian and US Black populations shows ample evidence of gene-flow, even in the religiously orthodox communities (Kennedy, 1944; Spuhler & Clark, 1961), we argue that a level of 10% proportional heterozygote deficiency is virtually inconsistent with the hypothesis of population substructure being the cause of the deficiency. On the other hand, 6% frequency of 'non-detectable' alleles quite reasonably explains 10% proportional heterozygote deficiency.

In principle, measuremental errors of band sizes may still affect accurate classification of alleles by a fixed bin approach (Budowle *et al.* 1991*a*). Quality control experiments (data not shown) suggest that this is not a critical concern, because sizing errors are small for small size

alleles where bin widths are narrow, and wide width of the large size bins can easily encompass up to 5% sizing error. In such cases, the $\tilde{p}_i$ value for the adjacent bins that have higher allele frequency should be used. Since the $2\tilde{p}_i$ already gives enough cushion in overestimating the actual genotype frequency for single band profiles, Budowle *et al*'s (1991*a*) liberal suggestion is more than sufficient to encompass sizing errors so that the detailed treatment of measuremental errors with additional assumptions, such as the ones suggested by Devlin *et al.* (1990), are not critically needed.

This work was supported by a grant 90-IJ-CX-0038 from the National Institute of Justice. The conclusions reached in this work, however, are solely the opinion of the authors and are not endorsements of the granting agencies supporting this research.

REFERENCES

BALLANTINE, J., SENSABAUGH, G. & WITKOWSKI, J. (1989). DNA technology and forensic science. Banbury Report 32. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

BOERWINKLE, E., XIONG, W., FOUREST, E. & CHAN, L. (1989). Rapid typing of tandemly repeated hypervariable loci by polymerase chain reaction: application to the apolipoprotein B3' hypervariable region. *Proc. Natl Acad. Sci. USA* **86**, 212–216.

BUDOWLE, B., GIUSTI, A. M., WAYE, J. S., BAECHTEL, F. S., FOURNEY, R. M., ADAMS, D. E., PRESLEY, L. A., DEADMAN, H. A. & MONSON, K. L. (1991*a*). Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic comparisons. *Am. J. Hum. Genet.* **48**, 841–855.

BUDOWLE, B., CHAKRABORTY, R., GIUSTI, A. M., EISENBERG, A. E. & ALLEN, R. C. (1991*b*). Analysis of the VNTR locus D1S80 by PCR followed by high-resolution PAGE. *Am. J. Hum. Genet.* **48**, 137–144.

CHAKRABORTY, R. & RAO, D. C. (1972). On the detection of F from ABO blood group data. *Am. J. Hum. Genet.* **24**, 352–353.

CHAKRABORTY, R. & NEI, M. (1974). Dynamics of gene differentiation between incompletely isolated populations of unequal sizes. *Theor. Pop. Biol.* **5**, 460–469.

CHAKRABORTY, R. & DAIGER, S. P. (1991). Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah. *Hum. Biol.* **63**, 571–587.

CHAKRABORTY, R., FORNAGE, M., GUEGUEN, R. & BOERWINKLE, E. (1991). Population genetics of hypervariable loci: analysis of PCR based VNTR polymorphism within a population. In *DNA Fingerprinting: Approaches and Applications* (ed. T. Burke, G. Dolf, A. J. Jeffreys and R. Wolff), pp. 127–143. Berne: Birkhäuer Verlag.

CHAKRABORTY, R. & JIN, L. (1992). Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. *Hum. Genet.* (in press).

COHEN, J. E. (1990). DNA fingerprinting for forensic identification: potential effects of data interpretation of subpopulation heterogeneity and band number variability. *Am. J. Hum. Genet.* **46**, 358–368.

DEVLIN, B., RISCH, N. & ROEDER, K. (1990). No excess of homozygosity at loci used for DNA fingerprinting. *Science* **249**, 1416–1420.

FOLTZ, D. W. (1986*a*). Segregation and linkage studies of allozyme loci in pair crosses of the oyster *Crassostrea virginica*. *Biochem. Genet.* **24**, 941–956.

FOLTZ, D. W. (1986*b*). Null alleles as a possible cause of heterozygote deficiencies in the oyster *Crassostrea virginica* and other bivalves. *Evolution* **40**, 869–870.

FORNAGE, M., CHAN, L., SIEST, G. & BOERWINKLE, E. (1991). Frequency distribution of $[TG]_n$ $[AG]_m$ minisatellite in the apolipoprotein c-II gene. *Genomics* (in the Press).

GART, J. J. & NAM, J. (1984*a*). A score test for the possible presence of recessive alleles in generalized ABO-like genetic systems. *Biometrics* **40**, 887–894.

GART, J. J. & NAM, J. (1984*b*). Statistical methods for genetic studies of HLA and cancer. In *Statistical Methods for Cancer Studies* (ed. R. G. Cornell), pp. 229–266. New York: Marcel Decker.

JEFFREYS, A. J., WILSON, V. & THEIR, S. L. (1985). Individual specific 'fingerprints' of human DNA. *Nature* **316**, 76–79.

JEFFREYS, A. J., ROYLE, N. J., PATEL, I., ARMOUR, J. A. L., MACLEOD, A., COLLICK, A., GRAY, I. C., NEUMANN, R., GIBBS, M., CROSIER, M., HILL, M., SIGNER, E. & MONCKTON, D. (1991). Principles and recent advances in human DNA fingerprinting. In *DNA Fingerprinting: Approaches and Applications* (ed. T. Burke, G. Dolf, A. J. Jeffreys and R. Wolff), pp. 1–19. Berlin: Birkhäuer Verlag.

KENNEDY, R. J. R. (1944). Single or triple melting pot? Intermarriage trends in New Haven, 1870–1940. *Am. J. Sociology* **49**, 331–339.

LANDER, E. (1989). DNA fingerprinting on trial. *Nature* **339**, 501–505.

LANDER, E. (1991). Invited editorial: research on DNA typing catching up with courtroom application. *Am. J. Hum. Genet.* **48**, 819–823.

LUDWIG, E. H., FRIEDL, W. & MCCARTHY, B. J. (1989). High-resolution analysis of a hypervariable region in the human apolipoprotein B gene. *Am. J. Hum. Genet.* **45**, 458–464.

MARTIN, W. (1983). Consideration of 'silent genes' in the statistical evaluation of blood group findings in paternity testing. In *Inclusion Probabilities in Parentage Testing* (ed. R. H. Walker), pp. 245–247. Arlington, VA: American Association of Blood Banks.

MILKMAN, R. & BEATTY, L. D. (1970). Large-scale electrophoretic studies of allelic variation in *Mytilus edulis*. *Biol. Bull.* **139**, 430.

NAKAMURA, Y., LEPPERT, M., O'CONNELL, P., WOLFF, R., HOLM, T., CULVER, M., MARTIN, C., FUJIMOTO, E., HOFF, M., KUMLIN, E. & WHITE, R. L. (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**, 1616–1622.

NAM, J. & GART, J. J. (1985). The ML estimation and testing of generalized ABO-like data with no observed double recessives. *Biometrics* **41**, 455–466.

NAM, J. & GART, J. J. (1987). On two tests of fit for HLA data with no double blanks. *Am. J. Hum. Genet.* **41**, 70–76.

NEI, M. & FELDMAN, M. W. (1972). Identity of genes by descent within and between populations under mutation and migration pressures. *Theor. Pop. Biol.* **3**, 460–465.

NEI, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl Acad. Sci. USA* **70**, 3321–3323.

NEI, M. (1975). *Molecular Population Genetics and Evolution*. Amsterdam: North-Holland.

SKIBINSKI, D. O. F., BEARDMORE, J. A. & CROSS, T. F. (1983). Aspects of the population genetics of *Mytilus* (Mytilidae; Mollusca) in the British Isles. *Biol. J. Linn. Soc.* **10**, 137–183.

SPUHLER, J. N. & CLARK, P. J. (1961). Migration into the human breeding population of Ann Harbor, Michigan, 1900–1950. *Hum. Biol.* **33**, 223–231.

THOMPSON, W. C. & FORD, S. (1989). DNA typing: acceptance and weight of the new genetic identification tests. *Virginia Law Review* **75**, 45–108.

WARD, R. H. & SING, C. F. (1970). A consideration of the power of the chi^2-test to detect inbreeding effects in natural populations. *Amer. Nat.* **104**, 355–363.

WYMAN, A. R. & WHITE, R. (1980). A highly polymorphic locus in human DNA. *Proc. Natl Acad. Sci. USA* **77**, 6754–6758.

AMERICAN JOURNAL OF HUMAN BIOLOGY 4:387–397 (1992)

Allele Sharing at Six VNTR Loci and Genetic Distances Among Three Ethnically Defined Human Populations

RANAJIT CHAKRABORTY[1], RANJAN DEKA[2], LI JIN[1], AND ROBERT E. FERRELL[2]
[1]*Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston, Texas 77225;* [2]*Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania 15261*

ABSTRACT Because of their high degree of polymorphisms, the variable number of tandem repeat (VNTR) loci have become extremely useful in studies involving gene mapping, determination of identity and relatedness of individuals, and evolutionary relationships among populations. However, there are some concerns regarding whether or not the patterns of such genetic variation can be studied by the classical population models that are developed for studying genetic variation at blood groups and protein loci, since VNTR alleles detected by molecular size may not always be identical by descent. Although theoretical and empirical studies demonstrate that this concern is overstated, this study provides further support of the application of the traditional mutation-drift models to predict the pattern of intra- and inter-populational variation at VNTR loci. By comparing genetic variation at six VNTR loci with that at 16 blood groups and protein loci in three ethnically defined populations, we show that the patterns of variability at these two sets of loci are in general parallel to each other. Shared VNTR alleles among populations are generally more frequent than the ones which are not present in every population; the proportion of shared alleles among populations increases with increasing genetic similarity of populations; and the number of VNTR alleles is positively correlated with gene diversity at these loci. All of these observations are in agreement with the prediction of the mutation-drift models, particularly when the possibility of forward-backward mutations are taken into account. This parallelism of genetic variation at VNTR loci and blood groups/protein loci further asserts the potential of using such hypervariable loci for microevolutionary studies, where closely related populations may exhibit considerably less allele frequency differences at the classical blood group and protein loci.

The discovery of the presence of tandemly repeated DNA sequences at several regions of the human genome has led to the characterization of a new class of polymorphic loci, known as variable number of tandem repeat (VNTR) loci. These loci are becoming increasingly popular in studies involving gene mapping, identification of individuals or familial relatedness, and evolutionary relationships between populations (see Burke et al., 1991). Several single locus VNTR probes have been developed (Wyman and White, 1980; Jeffreys et al., 1985; Nakamura et al., 1987) and their characterizations with regard to intra- and inter-populational genetic variation are also now available (Jeffreys et al., 1988; Clark, 1987; Flint et al., 1989; Odelberg et al., 1989; Deka et al., 1991). However, since this class of polymorphism reflects copy number variation of short conserved core sequences (ranging in length from one to hundreds of nucleotides) and the molecular mechanism of production and maintenance of VNTR alleles is still obscure, there are some concerns as to whether or not the patterns of VNTR polymorphisms

Received September 24, 1991; accepted November 12, 1991.

can be studied by the classical population genetic models developed for studying the blood group and protein variation (Lander, 1989, 1991; Kidd et al., 1991). Empirical data on intra-population variation at several VNTR loci has demonstrated that the traditional population genetic models are applicable to such loci (see Clark, 1987; Jeffreys et al., 1988; Flint et al., 1989; Chakraborty and Daiger, 1991; Chakraborty et al., 1991). Our earlier work on six VNTR loci studied in three ethnically defined populations exhibited the utility of VNTR polymorphisms for inter-populational genetic differentiation as well.

One characteristic feature of VNTR polymorphism is that these loci are generally more variable than the traditional blood groups and protein loci, reflected in higher levels of heterozygosity per locus as well as a larger number of alleles. Jeffreys et al. (1988) showed that these are caused by a higher rate of mutation at the VNTR loci in comparison to traditional loci. Such hypervariability, of course, necessitates certain modifications of statistical analyses of VNTR polymorphism data (see Chakraborty et al., 1991; Deka et al., 1991), but these are not sufficient reason to invalidate the application of the classic mutation-drift models of population genetics to analyze such data.

The purpose of this paper is to provide further evidence of comparability of intra- and inter-populational variation at the VNTR loci with those at traditional blood group and protein loci. Using a previous survey (Deka et al., 1991) of genetic variation at six VNTR loci (D1S57, D1S61, D1S76, D1S77, RB1, and α-globin 5′ HVR) in three ethnically defined populations (Kachari of Northeastern India, New Guinea Highlanders of Papua New Guinea, and Dogrib Indians of Canada), we show that among these populations the pattern of allele sharing is identical with that at 16 blood groups and protein loci studied previously in these populations. This pattern is also in accord with the prediction of a mutation-drift model of gene differentiation. Furthermore, we show that the proportion of shared alleles between populations is also parallel to their genetic similarities. Finally, the empirical relationship between the number of alleles and gene diversity at the six VNTR loci is also in agreement with the prediction of the mutation-drift model. These observations on the parallelism between the pattern of genetic variation at VNTR loci and blood groups and enzyme loci, along with the fact that the genetic distances among these populations as detected by the VNTR loci are in accordance with their historical records, suggest that such hypervariable loci are extremely useful for studying genetic differentiation between human populations. Since the DNA materials for the Kachari population used for the VNTR study (Deka et al., 1991) were all from individuals belonging to the Sonowal subgroup of Kacharis, the Kachari allele frequencies at the blood group and protein loci used in the present comparative analysis (reproduced in Appendix A), are the ones represented as that of the Sonowals (Sandwals in Walter et al., 1986) in the original papers (see Appendix A for the data source).

MATERIALS AND METHODS

Populations

For the study of comparability of patterns of polymorphisms at VNTR loci and blood groups and enzyme loci, three ethnically defined diverse human populations were selected: the Kachari of Northeastern India, Kalam and Gainj speaking New Guinea Highlanders of Papua New Guinea, and Dogrib Indians of Canada, who have been previously examined for genetic variation at several blood groups and protein loci. The Kachari is a distinct Mongoloid tribal group who live in the plains of the northeast Indian state of Assam. They belong to the Bodo subgroup of the Tibetoburman language family, and are known to have a population size exceeding 50,000 during the present century (B.M. Das, personal communication). Kalam and Gainj speaking New Guinea Highlanders represent a culturally inter-related interbreeding group of inhabitants of the northern fringe of Papua New Guinea's central highlands. They are one of the least acculturated, genetically unadmixed populations of the South Pacific (Wood et al., 1982). Although the precise estimate of the size of this population is unknown, in 1978 the de facto population size was a little over 1,100 (Long et al., 1986) and their demographic profile appears quite stable (Wood and Smouse, 1982). The Dogrib Indians are one of the 16 Athapaskan-speaking Amerindian tribes, and they reside in the Northwest Territories of Canada. Genetically, linguistically, and anthropologically, this is a well-characterized population

with low levels of non-Amerindian admixture (Szathmary et al., 1983) and from a demographic perspective they resemble to some extent the New Guinea Highlanders insofar as their current population size is concerned (Szathmary, 1983). From the recent history of migration, there are suggestions that the Kacharis might have in their gene pool, genes of Polynesian origin, suggesting genetic proximity with the New Guinea Highlanders (Walter et al., 1986).

Data

To generate data on VNTR polymorphims, high-molecular-weight DNA was isolated from buffy coats of 45 Kachari, 46 New Guinea Highlanders, and 30 Dogrib Indians by phenol-chloroform extraction (Aldridge et al., 1984). Further details of restriction endonuclease digestion, Southern blot technique of restriction fragment length polymorphism (RFLP) analysis to detect the different VNTR alleles, and the degree of resolution of allele size determinations, along with the chromosomal localization of the six VNTR loci (D1S57, D1S61, D1S76, D1S77, RB1, and α-globin 5′ HVR) are given in Deka et al. (1991). With the exception that the two loci D1S76 and D1S77 are closely linked ($\theta = 0.043$; O'Connell et al., 1989), the six loci represent a set of independently segregating loci at each of which genetic variation is governed by copy number variation of tandemly repeated core sequences. Deka et al. (1991) present the allele frequency distributions of each of these loci in the three populations examined here.

To contrast the pattern of VNTR polymorphisms at these six VNTR loci with that at the blood groups and protein loci, allele frequency data at 16 blood groups and protein loci which were previously published were extracted. While the New Guinea Highlanders and the Dogrib Indians were examined for a comparatively larger set of blood groups and protein loci (see Long et al., 1986; Szathmary, 1983; Szathmary et al., 1983), data on the traditional loci in the Kachari population are more limited. This led to the consideration of data at 16 blood groups and protein loci (ABO, Rh, MNSs, Diego, and Duffy blood groups, and AK, ADA, TF, ESD, GC, HP, ACP, LDH-A, HB-β, Gm, and Km protein loci) for the present analysis. Appendix A gives a compilation of allele frequencies at these loci for the three populations used in this analysis, along with their respective sample sizes (number of individuals sampled) and source of data.

Statistical methods

Patterns of polymorphism studied in this paper use several summary measures of genetic variation: (1) number of alleles, (2) gene diversity (Nei, 1973), and (3) genetic distance (bias-corrected estimate of Nei's standard genetic distance; Nei, 1978). These are directly computed from the allele frequency data (see Deka et al., 1991, and Appendix A). In order to analyze the pattern of allele sharing, the mean frequencies (and standard errors) were computed for the alleles that are present in all three, in two of the three, and in one of the three populations for both sets of loci (VNTR versus the traditional ones). The expectation is that the alleles that are present in all three populations should be more frequent than the ones which are not found in all of them. This is expected under a mutation-drift model, since Watterson and Guess (1977) have shown that the oldest allele is also likely to be the most frequent in a population, and hence, an allele which is present in all three populations is likely to have existed before the split of these three diverse populations, and thereby might also be more frequent than the newer ones which might not be shared by all. Second, for the three pairwise contrasts of populations, we computed the proportion of shared alleles, using the formula

$$P_{XY} = 2k_{XY}/(k_X + k_Y) \qquad (1)$$

where k_X and k_Y are the number of alleles in populations X and Y, and k_{XY} is the number of alleles present in both populations X and Y. The values of p_{XY} were computed for the six VNTR loci jointly and contrasted with that of the 16 blood groups and protein loci. For each choice of X and Y, the p_{XY} values were contrasted with Nei's gene identity for the corresponding loci (Nei, 1978), because under the mutation-drift model of gene differentiation, the proportion of shared alleles is expected to be positively correlated with gene identity (a measure of genetic similarity) between populations. Finally, the relationship between the number of alleles and gene diversity (heterozygosity) is studied for the six VNTR loci by plotting a scatter diagram of these two variables to examine

TABLE 1. Frequencies of shared alleles at VNTR and blood group and protein loci

Alleles present in—	For six VNTR loci		For 16 blood group/protein loci	
	Number	Frequency Mean ± S.E.	Number	Frequency Mean ± S.E.
3 populations	15	0.276 ± 0.052	30	0.488 ± 0.062
2 populations	15	0.166 ± 0.048	7	0.247 ± 0.088
1 population	17	0.041 ± 0.013	7	0.101 ± 0.063

TABLE 2. Relationship between the proportion of shared alleles (p_{XY}) and Nei's Gene Identity (I_{XY}) at the VNTR and blood groups/protein loci

Populations	VNTR loci		Blood groups/protein loci	
	I_{XY}	p_{XY}	I_{XY}	p_{XY}
Kachari vs. New Guinea	0.846	0.712	0.891	0.883
Kachari vs. Dogrib	0.654	0.581	0.870	0.892
Dogrib vs. New Guinea	0.760	0.653	0.855	0.873

whether or not these two measures of genetic variation are positively correlated. Under the mutation-drift model, these two quantities are positively correlated whose magnitude can be predicted from the observed average heterozygosity at the loci scored (Chakraborty and Griffiths, 1982).

RESULTS

Frequencies of shared and non-shared alleles

Table 1 shows the numbers and mean frequencies (± S.E.) of all alleles observed at the six VNTR loci and 16 blood groups and protein loci, categorized by their presence in all three, two of the three, and only in one of the three populations. It is clear that for both sets of loci, the pattern is identical; the alleles that are present in all three populations are more frequent than the ones which are present in two of the three populations, which are in turn more frequent than the alleles that are present in only one of the three populations. The differences of average frequencies of these three classes of alleles appear to be greater for the blood groups and protein loci, compared with the VNTR loci, which is expected, because the larger extent of polymorphism at the VNTR loci makes the individual allele frequencies generally smaller for each VNTR allele, in comparison to the blood group and protein alleles.

Relationship between proportion of shared alleles and gene identity

Table 2 shows the relationship between the proportion of shared alleles and gene identity for the two groups of loci. For the VNTR loci the correspondence between gene identity and proportion of shared alleles is perfect, reflecting that genetically similar populations are also sharing a proportionally larger number of VNTR alleles. On the contrary, for the blood groups and protein loci, such correspondence is equivocal. For example, even though the Kacharis are genetically closer to the New Guinea Highlanders, the proportion of shared blood groups/protein alleles between them is comparatively smaller than what they share with the Dogribs. There are two possible explanations for this discordance. First, since a great majority of the blood groups/protein loci are 2-allele systems (9 out of 16), the variation of the proportion of shared alleles is rather limited even among populations that are very diverse genetically. Second, the differences of the gene identity values among these populations detected at the blood groups and protein loci do not appear to be significant, as can be seen from the standard errors of the corresponding genetic distances (discussed later). In spite of these, the pattern of allele sharing at the VNTR loci indicates that even though the designation of allelic distinctions by copy number variation at these loci raises the possibility that two similar size alleles may not be of identical by descent, the present data shows that such allelic distinctions are quite adequate for the purpose of genetic comparisons between populations.

Relationship between gene diversity and number of alleles

Deka et al. (1991) demonstrated that the genotype distributions at these six VNTR loci are in accordance with the Hardy-Wein-

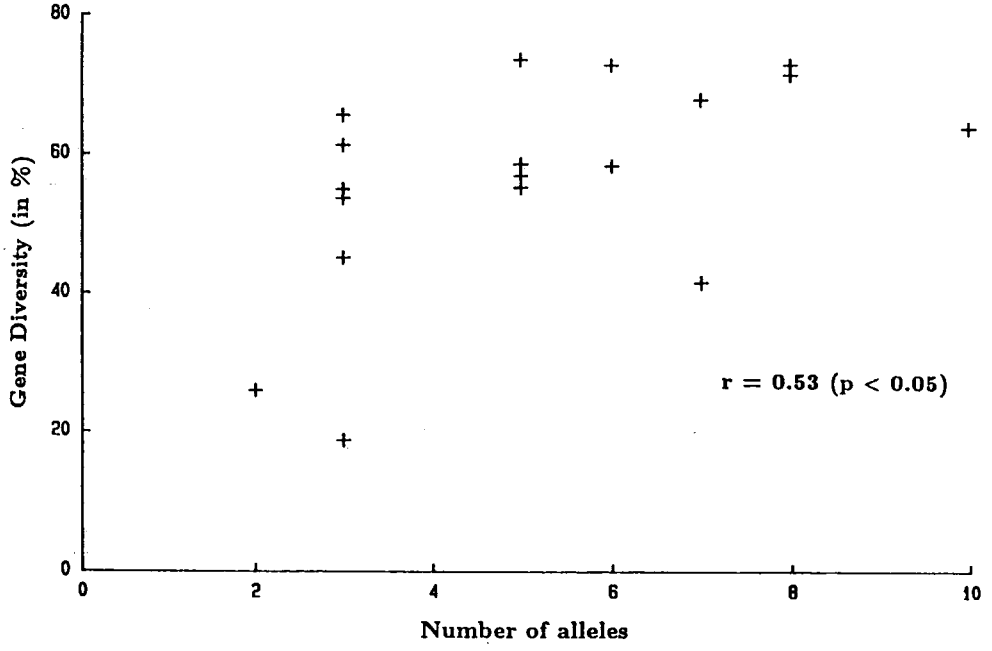


Fig. 1. Relationship between gene diversity (H) and number of alleles at six VNTR loci in three ethnically defined populations (Kachari, Dogrib, and New Guinea Highlander). The observed correlation, r = 0.526, is statistically significant ($P < 0.05$) and is in approximate agreement with the prediction of the classical infinite allele model (expected r = 0.562).

berg proportions for each of these three populations. Therefore, Nei's measure of gene diversity (estimated by the bias-correction method suggested in Nei, 1978) provides an adequate estimate of the heterozygosity at these loci. Figure 1 shows the scatter diagram of the number of alleles (k) and gene diversity (H) for the 18 population-loci combinations for the VNTR loci. It is clear that there is a positive trend in the scatter plot. Chakraborty and Griffiths (1982) have shown that, while gene diversity (H) and number of alleles (k) should be positively correlated under a mutation-drift model, their correlation is not perfect and is critically dependent on sample size. The observed correlation for the data shown in Figure 1 is 0.53, which is statistically significant ($P < 0.05$) even though the number (18) of data points is limited. Using the average gene diversity at these six loci as the base line of estimation, the expected correlation under the infinite allele model (Chakraborty and Griffiths, 1982) is 0.562, which is not statistically different from the one observed. This also reflects that designation of VNTR allelic distinctions by their lengths (or copy numbers) does not compromise the utility of such polymorphisms for evolutionary studies.

Genetic distances among populations at VNTR loci

Table 3 shows the comparison of genetic distances among these populations detected at the VNTR loci and the blood groups/protein loci. Both sets of loci demonstrate that the Kacharis and the New Guinea Highlanders are the closest of three contrasts that can be made from this data. From an anthropological view point, this is re-assuring, since the Dogrib Indians are the descendants of Mongoloid tribes that entered the New World through the Bering land-bridge around 12,000 to 15,000 years ago (at the latest), while the migration to New Guinea was much more recent and its contact with the mainland Asia was hardly uninterrupted during the history of civilization (Kirk and Szathmary, 1985). There are

TABLE 3. Genetic distances among the three populations for the six VNTR and 16 blood groups/protein Loci

Populations	Nei's standard distance (±S.E.) for VNTR Loci	Nei's standard distance (±S.E.) for blood groups/protein loci
Kachari vs. New Guinea	0.167 ± 0.067	0.115 ± 0.050
Kachari vs. Dogrib	0.424 ± 0.290	0.140 ± 0.062
Dogrib vs. New Guinea	0.275 ± 0.137	0.157 ± 0.078

some apparent discordances in the other estimates shown in Table 3. For example, the VNTR loci show that the Dogribs are the farthest from the Kacharis, while the blood group/protein loci predicts that the New Guinea and Dogrib distance is the largest. Upon a closer examination, these discrepancies can be ascribed to small number of loci used in both analyses. For both sets of loci, the genetic distances of the Dogribs from the Kacharis and New Guineans are statistically similar, because of their large standard errors. These standard errors, it should be noted, are more critically dependent on the number of loci used in the analysis, rather than the number of individuals surveyed (Nei, 1978). With this in mind, we may conclude that the genetic distance analysis of the VNTR allele frequency data is in agreement with that of the blood groups/protein loci data.

DISCUSSION

The results indicate that even when the VNTR allele designations are made from the RFLP analysis of allele sizing, the pattern of genetic variation observed at these hypervariable loci are almost parallel to the ones found at blood groups and protein loci. Nonidentity by descent of identical size VNTR alleles, which is a possibility, does not therefore compromise the utility of VNTR polymorphisms for evolutionary studies, although Kidd et al. (1991) mentioned this as a limitation of VNTR polymorphism detected through the RFLP analysis. It is true that designation of alleles by only copy number variation does not detect the allelic distinctions at the molecular level. Indeed, Deka et al. (1991) showed that the six loci studied here conform to a mutation-drift model, somewhere in between the classic infinite allele model (Kimura and Crow, 1964) and one-step forward-backward stepwise mutation model (Kimura and Ohta, 1978). The second model takes into account the hidden variation within alleles that have the same copy number of tandemly repeated core sequences. For this reason, a strict adherence to the inifinite allele model for every VNTR locus is not recommendable. Until the molecular mechanism of production of new VNTR alleles is precisely known, the exact calibration of the pattern of VNTR polymorphism data cannot be made. However, the above two models prescribe two limits of calibration, e.g., with regard to evolutionary time of divergence, rate of mutation, etc. As shown by Jeffreys et al. (1988), the correspondence between heterozygosity (determined by the proportion of two-banded genotype profiles detected by the RFLP analysis of single-locus VNTR probes) and mutation rate at several VNTR loci assures that such population genetic models are appropriate for using VNTR polymorphism data in evolutionary studies.

Even though the present study is the first direct demonstration of parallelism between the patterns of genetic variation at VNTR loci and blood groups/protein loci in ethnically defined populations, the results should be treated with some caution because of their preliminary nature. First, the sample sizes (number of individuals) for the VNTR assays are rather limited (30 to 46), which compromises the precision of the allele frequencies estimates. However, theoretical and empirical studies of sample size requirements for VNTR polymorphism indicate that the sample size limitations affect the precision of frequencies of only rare alleles, because when a limited number of individuals are sampled, rare alleles may not be observed (Evett and Gill, 1991; Chakraborty, 1992). Since the rare alleles do not contribute much to the estimation of summary measures such as gene diversity, gene identity, or genetic distance (Nei, 1978, 1987), small sample size does not constitute a major drawback of the qualitative conclusions reached in the present analysis. Second, there are also concerns that the RFLP analysis of allele detection results in incomplete resolution of true alleles. Although some critics advocate that a discrete allele theory is not applicable to such quasi-continuous variation of allele sizes, we argue that we followed a uniform protocol of allelic resolution for all sampled individuals, enabling us to detect minute differences between even

closely spaced fragments, which would have been missed by simply comparing different autoradiograms (see Deka et al., 1991 for details). Thus, the data can be treated fairly accurately with a discrete allele model, completely parallel to the treatment of the blood groups and protein alleles.

We postulated that the alleles that are present in all three populations examined here have supposedly existed before the split of these populations. This conclusion should also be treated with some caution, since there are suggestions that should the molecular mechanism of production of copy number variation of VNTR alleles behave in some form of forward-backward events, the hidden variability is expected to be more pronounced in the frequent alleles (Nei and Chakraborty, 1976; Chakraborty and Nei, 1976). Since the shared alleles are seen more frequently in our analysis (see Table 1), further experimental studies are needed to establish their molecular identity. To this end, we might note that Boerwinkle et al. (1989) and Jeffreys et al. (1990) have noted molecular heterogeneity of identical size VNTR alleles. Our speculation is that such molecular heterogeneity should exist more among these ancestral frequent alleles, for which internal mapping of shared VNTR alleles are currently being attempted with polymerase-chain-reaction (PCR) based sequencing studies. If this speculation is correct, the utility of VNTR polymorphism will be even greater for studying micro-evolutionary divergence between genetically closely related populations.

The present study also indicates that in spite of the caveats of allelic designations achieved by RFLP analysis of VNTR alleles, the degree of genetic variation detected by VNTR loci is larger than that at the blood groups/protein loci. This is reflected in higher genetic distance for these loci as well (Table 3). As a comparison, the average gene diversity for the 16 blood groups/protein loci for the three populations are 0.338 ± 0.050 (Kachari), 0.213 ± 0.051 (Dogrib), and 0.194 ± 0.056 (New Guinea Highlander). These are considerably lower than the levels of gene diversity at the six VNTR loci (0.670 ± 0.024, 0.432 ± 0.071, and 0.576 ± 0.043, respectively). Although these values should not be interpreted in absolute terms, since the loci are heavily biased towards being polymorphic, they provide a number of interesting implications with regard to the evolutionary history of the populations examined and the biology of the loci studied.

First, the Kacharis appear to be the most variable of the three populations studied at both sets of loci. Since the above contrasts are based on a set of common loci examined for each populations, they reflect that perhaps the Kacharis have a larger effective population size compared with the others. The history of these populations support this view, since at least during the past century the Kacharis maintained a considerably larger census size than the other two populations. The larger effective size in the Kacharis may also have been caused by a substantial amount of gene admixture in this population. Being situated in the northeastern corridor of the Indian subcontinent, this population received genes of Caucasian as well as Mongoloid ancestry (Walter et al., 1986, 1987), while the other two populations are relatively less admixed. Szathmary (1983) estimated that the maximum amount of non-Amerindian admixture in the Dogrib population could be 8.7%, while Long et al. (1986) asserted that the New Guinea Highlanders are perhaps the most unaccultrated. Although the average gene diversity levels at the six VNTR loci in the Dogrib and New Guinea Highlander populations are not significantly different (at 5% level), Szathmary et al.'s (1983) estimate of average gene diversity ($12.8 \pm 3.0\%$) in the Dogribs for a larger set of blood groups/protein loci (36 loci) is almost 2.5-times larger than that for the New Guinea Highlanders ($5.3 \pm 1.4\%$) estimated by Long et al. (1986) at an even larger set of 39 loci. This indicates that perhaps some VNTR variants may have been either missed in our survey of 30 Dogrib individuals analyzed here, or because of small size of this population, there had been a true loss of genetic variation in this tribal population. Second, under some restrictive assumptions these comparative data provide indirect estimate of the mutation rate at the VNTR loci as well. For example, under the two extreme mutation models, the expected gene diversities in an equilibrium population are given by

$$H = \begin{cases} 4N_e v/(1 + 4N_e v), & \text{for the infinite allele model,} \quad (2.1) \\ [(1 + 8N_e v)^{1/2} - 1]/(1 + 8N_e v)^{1/2}, & \text{for the stepwise mutation model.} \quad (2.2) \end{cases}$$

TABLE 4. Estimates of relative mutation rate at VNTR loci compared with blood groups and protein (BG/P) loci

Populations	Average gene diversity		Estimates of relative mutation rate	
	VNTR loci	BG/P Loci[1]	Using Equation 2.1	Using Equation 2.2
New Guinea	0.576	0.053	24.3	39.7
Kachari	0.670	0.139	12.6	23.5
Dogrib	0.432	0.118	5.7	7.3

[1]These estimates are adjusted assuming that the additional loci surveyed in Long et al. (1986) would be monomorphic in the Dogrib and Kachari populations.

N_e being the effective size and v, the mutation rate per locus per generation (Kimura and Crow, 1964; Ohta and Kimura, 1973), from which the relative mutation rate at these two sets of loci can be crudely estimated following Zouros (1979). Table 4 presents the result of such computations, where for uniformity all estimates of gene diversity are adjusted for the largest set of 39 loci (examined in Long et al., 1986), assuming that these additional loci would be monomorphic in the Kachari and Dogrib populations as well.

Although these estimates are quite crude, they reflect that the rate of mutation at these six VNTR loci is between 6- and 40-fold of that at the blood groups and protein loci. Assuming that the traditional loci mutate at a rate of 1.1×10^{-5}/locus/generation (Chakraborty and Neel, 1989), the rate of VNTR mutability would become somewhere between 6.6×10^{-5} to 4.4×10^{-4} per locus per generation. While these estimates are considerably lower than the rate of spontaneous mutations at the VNTR loci, as reported by Jeffreys et al. (1988), these are in the range seen in the studies of Wolff et al. (1988), Chakraborty and Daiger (1991) and Edwards et al. (1992).

Using the same equations (2.1 and 2.2), we can also estimate the relative effective sizes of the Dogrib and New Guinea populations in comparison to that of the Kacharis. Such calculations suggest that when the VNTR data are used, the effective size of the New Guinea Highlanders appear to be about 56 to 67% of that of the Kacharis, while the Dogrib size is about 26 to 37% of that of the Kacharis. The estimates for the blood groups/protein data are about 33 to 35% and 82 to 84%, respectively. Although these estimates are rather variable (and perhaps, quite imprecise), qualitatively we may conclude that the Kacharis have a comparatively larger effective population size, achieved due to the fact that they have intermixed with larger Caucasian as well as Mongoloid gene pools. For the additional analyses we assumed that the three populations are, within themselves, homogeneous. Although, Szathmary (1983), Szathmary et al. (1983) and Long et al. (1986) examined the extent of substructuring within the Dogribs and New Guineans, the coefficient of gene diversity (either by Nei's measure, G_{ST}; Nei, 1973; or by Wright's F_{ST}; Weir and Cockerham, 1984; Long, 1986) are not large enough to substantially change the above qualitative conclusions with regard to either the relative mutation rate, or the relative effective size. In fact, the allele frequency spectra in the total population data satisfies the premises of homogeneity within each of these three populations (details of such data not shown, but can be inferred from the allele frequency data of Deka et al., 1991 and Appendix A), for both sets of loci.

In conclusion, this preliminary comparative study of genetic variation at the six VNTR loci and 16 blood groups/protein loci exhibits that the pattern of genetic variation at the VNTR loci, with allelic designations determined by RFLP analysis, is parallel to that of blood groups and protein variation. In addition to this empirical support of the application of a mutation-drift model to examine the features of VNTR polymorphism data, we might note that Ohta (1986) provides a theoretical treatment, suggesting that the infinite allele model may also be appropriate to study genetic variation detected by copy number variation of repeated short DNA sequences. While the theoretical predictions of the summary measures of genetic variation are dependent on the assumption that the populations are at equilibrium due to mutation-drift balance, previous analysis of allele frequency distributions at the loci studied here suggests that this assumption approximately holds for the VNTR polymorphisms in these populations (Deka et al., 1991). The caveats of the

RFLP typing of VNTR alleles, mentioned here, can be circumvented by more refined methods (such as PCR-based techniques, and internal mapping of specific VNTR alleles), and in fact, the above discussions indicate that such future studies should strengthen the notion that the hypervariability present at the VNTR loci will be extremely useful for microevolutionary studies, where gene frequency variation at traditional blood groups and protein loci may give equivocal results.

ACKNOWLEDGMENTS

This work was supported in part by NIH grant GM 41399 from the National Institutes of Health and grant 90-IJ-CX-0038 from the U.S. National Institute of Justice (both to R.C.), and the Central Research Development Fund Award 5-33344 from the University of Pittsburgh (to R.D.). The comments of an anonymous reviewer were helpful in simplifying the presentation. Points of view or opinions are, however, the sole responsibility of the authors.

LITERATURE CITED

Aldridge J, Kunkel L, Bruns G, Tantaravahi U, Lalande M, Brewster T, Moreau E, Wilson M, Bromley W, Roderick T, and Latt SA (1984) A strategy to reveal high-frequency RFLPs along the human X chromosome. Am. J. Hum. Genet. *36:*546–564.

Boerwinkle E, Xiong W, Fourest E, and Chan L (1989) Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: Application to the apolipoprotein B 3′ hypervariable region. Proc. Natl. Acad. Sci. U.S.A. *86:*212–216.

Burke T, Dolf G, Jeffreys AJ, and Wolff R (eds.) (1991) DNA Fingerprinting: Approaches and Applications. Basel: Birkhäuser.

Chakraborty R (1992) Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Hum. Biol. *64:*141–159.

Chakraborty R, and Daiger SP (1991) Polymorphisms at VNTR loci suggest homogeneity of the White population of Utah. Hum. Biol. *63:*571–587.

Chakraborty R, Fornage M, Gueguen R, and Boerwinkle E (1991) Population genetics of hypervariable loci: Analysis of PCR based VNTR polymorphism within a population. In T Burke, G Dolf, AJ Jeffreys, and R Wolff (eds.): DNA Fingerprinting: Approaches and Applications. Basel: Birkhäuser, pp. 127–143.

Chakraborty R, and Griffiths RC (1982) Correlation of heterozygosity and number of alleles in different frequency classes. Theor. Pop. Biol. *21:*205–218.

Chakraborty R, and Neel JV (1989) Description and validation of a method for simultaneous estimation of effective population size and mutation rate from human population data. Proc. Natl. Acad. Sci. U.S.A. *86:*9407–9411.

Chakraborty R, and Nei M (1976) Hidden genetic variability in electromorphs in finite populations. Genetics *84:*385–393.

Clark AG (1987) Neutrality tests of highly polymorphic restriction fragment length polymorphisms. Am. J. Hum. Genet. *41:*948–956.

Das BM, Walter H, Gilbert K, Lindenberg P, Malhotra KC, Mukherjee BN, Deka R, and Chakraborty R (1987) Genetic variation of five blood group polymorphisms in ten populations of Assam, India. Int. J. Anthropol. *2:*325–340.

Deka R, Chakraborty R, and Ferrell RE (1991) A population genetic study of six VNTR loci in three ethnically defined populations. Genomics *11:*83–92.

Deka R, Gogoi BC, Hundrieser J, and Flatz G (1987) Hemoglobinopathies in Northeast India. Hemoglobin *11:*531–538.

Edwards A, Hammond HA, Caskey CT, Jin L, and Chakraborty R (1992) Population genetics of trimeric and tetrameric tandem repeats in four human ethnic groups. Genomics (in Press).

Evett IW, and Gill P (1991) A discussion of the robustness of methods assessing the evidential value of DNA single locus profiles in crime investigations. Electrophoresis *12:*226–230.

Flint J, Boyce AJ, Martinson JJ, and Clegg JB (1989) Population bottlenecks in Polynesia revealed in minisatellites. Hum. Genet. *83:*257–263.

Jeffreys AJ, Neumann R, and Wilson V (1990) Repeat unit sequence variation in minisatellites: A novel source of DNA polymorphism for studying variation and mutation by single nucleotide analysis. Cell *60:*473–485.

Jeffreys AJ, Royle NJ, Wilson V, and Wong Z (1988) Spontaneous mutation rates to new length alleles at tandem-repeatitive loci in human DNA. Nature *332:*278–281.

Jeffreys AJ, Wilson V, and Thein SL (1985) Hypervariable "minisatellite" regions of human DNA. Nature *314:*67–73.

Kidd JR, Black FL, Weiss KM, Balazs I, and Kidd KK (1991) Studies of three Amerindian populations using nuclear DNA polymorphisms. Hum. Biol. *63:*775–794.

Kimura M, and Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics *49:*725–738.

Kimura M, and Ohta T (1978) Stepwise mutation model and distribution of allele frequencies in a finite population. Proc. Natl. Acad. Sci. U.S.A. *75:*2868–2872.

Kirk RL, and Szathmary E (1985) Out of Asia. Canberra: Australian National University Press.

Lander ES (1989) DNA fingerprinting on trial. Nature *339:*501–505.

Lander ES (1991) Invited editorial: Research on DNA typing catching up with courtroom applications. Am. J. Hum. Gnet. *48:*819–823.

Long JC (1986) The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's F-statistics. Genetics *112:*629–647.

Long JC, Naidu JM, Mohrenweiser HW, Gershowitz H, Johnson PL, Wood JW, and Smouse PE (1986) Genetic characterization of Gainj- and Kalam-speaking peoples of Papua New Guinea. Am. J. Phys. Anthropol. *70:*75–96.

Mukherjee BN, Malhotra KC, Roy M. Banerjee S. Walter H, Chakraborty R (1989) Genetic heterogeneity and population structure in eastern India: Red cell enzyme variability in ten Assamese populations. Z. Morph. Anthropol. 77:287–296.

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T,

Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, and White R (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science *235*:1616–1622.

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. U.S.A. *70*:3321–3323.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics *89*:583–590.

Nei M (1987) Molecular Evolutionary Genetics. New York: Columbia University Press.

Nei M, and Chakraborty R (1976) Electrophoretically silent alleles in a finite population. J. Mol. Evol. *8*: 381–385.

O'Connell P, Lathrop GM, Nakamura Y, Leppert ML, Ardinger RH, Murray JL, Lalouel JM, and White R (1989) Twenty-eight loci form a continuous linkage map of markers for human chromosome 1. Genomics *4*:12–20.

Odelberg SJ, Paltke R, Eldridge JR, Ballard L, O'Connell P, Nakamura Y, Leppert M, Lalouel JM, and White R (1989) Characterization of eight VNTR loci by agarose gel electrophoresis. Genomics *5*:915–924.

Ohta T (1986) Actual number of alleles contained in a multigene family. Genet. Res. *48*:119–123.

Ohta T, and Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. *22*:201–204.

Szathmary EJE (1983) Dogrib Indians of the Northwest Territories, Canada: Genetic diversity and genetic relationship among subarctic Indians. Ann. Hum. Biol. *10*:147–162.

Szathmary EJE, Ferrell RE, and Gershowitz H (1983) Genetic differentiation in Dogrib Indians: Serum protein and eyrthrocyte enzyme variation. Am. J. Phys. Anthropol. *62*:249–254.

Walter H, Matsumoto H, Miyasaki T, Mukherjee BN, Malhotra KC, Das BM, Gilbert K, and Lindenberg P (1987) Distribution of Gm and Km allotypes among ten populations of Assam, India. Am. J. Phys. Anthropol. *73*:439–445.

Walter H, Mukherjee BN, Gilbert K, Lindenberg P, Dannewitz A, Malhotra KC, Das BM, and Deka R (1986) Investigations on the variability of haptoglobin, transferin and Gc polymorphisms in Assam, India. Hum. Hered. *36*:388–396.

Watterson GA, and Guess HA (1977) Is the most frequent allele the oldest? Theor. Pop. Biol. *11*:141–160.

Weir BS, and Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution *38*:1358–1370.

Wolff RK, Nakamura Y, and White R (1988) Molecular characterization of a spontaneously generated new allele at a VNTR locus: No exchange of flanking DNA sequence. Genomics *3*:347–351.

Wood JW, Johnson PL, Kirk RL, McLoughlin K, Blake NM, and Matheson FA (1982) The genetic demography of the Gainj of Papua New Guinea. I. Local differentiation of blood group, red cell enzyme, and serum protein allele frequencies. Am. J. Phys. Anthropol. *57*:15–25.

Wood JW, and Smouse PE (1982) A method of analyzing density-dependent vital rates with an application to the Gainj population of Papua New Guinea. Am. J. Phys. Anthropol. *58*:304–411.

Wyman AR, and White R (1980) A highly polymorphic locus in human DNA. Proc. Natl. Acad. Sci. U.S.A. *77*:6754–6758.

Zouros E (1979) Mutation rates, population sizes, and amounts of electrophoretic variation of enzyme loci in natural populations. Genetics *92*:623–646.

APPENDIX A. Allele frequency data at 16 blood groups and protein loci in the Kacharis, New Guinea Highlanders, and Dogrib Indians[1]

Locus	Allele	Dogrib	New Guinea	Kachari
ABO	A_1	0.177	0.347	0.171
	A_2	—	—	0.041
	B	—	0.126	0.186
	O	0.823	0.527	0.602
	(n)	158	415	107
Rh	CDe	0.221	0.933	0.812
	cDE	0.695	0.021	0.064
	CDE	0.014	0.010	0.047
	cDe	0.023	0.036	0.040
	cde	0.047	—	0.037
	(n)	158	412	107
MNSs	MS	0.095	0.015	0.071
	Ms	0.820	—	0.672
	NS	0.022	0.111	0.023
	Ns	0.063	0.874	0.234
	(n)	158	229	107
Diego	Di^a	—	—	0.028
	Di^b	1.0	1.0	0.972
	(n)	158	54	107
Duffy	Fy^a	0.953	0.994	0.744
	Fy^b	0.047	0.006	0.256
	(n)	158	390	107
AK	AK_1	1.0	1.0	0.948
	AK_2	—	—	0.052
	(n)	158	277	106
TF	Tf^C	1.0	0.946	1.0
	Tf^D	—	0.054	—
	(n)	158	575	64
ESD	ESD_1	0.826	0.934	0.617
	ESD_2	0.174	0.066	0.383
	(n)	158	570	107
GC	Gc^1	0.930	0.825	0.769
	Gc^2	0.070	0.152	0.231
	Gc^{Ab}	—	0.023	—
	(n)	158	522	104
HP	Hp_1	0.361	0.739	0.240
	Hp_2	0.639	0.261	0.760
	(n)	158	470	104
ACP	p^A	0.462	0.270	0.262
	p^B	0.538	0.730	0.738
	(n)	158	570	107
ADA	ADA_1	0.997	0.973	0.723
	ADA_2	0.003	0.026	0.277
	ADA_6	—	0.001	—
	(n)	158	570	83
HB-β	Hb_A	1.0	1.0	0.493
	Hb_E	—	—	0.507
	(n)	158	588	1082
LDH-A	Normal	1.0	1.0	1.0
	(n)	158	589	107
Gm	a;g	0.789	0.398	0.021
	a,x:g	0.066	0.045	0.117
	f;b0, 1, 3	0.010	—	0.201
	a;b	0.135	0.126	—
	f,a;b	—	0.311	0.523
	Others	—	0.120	0.138
	(n)	156	513	76
Km	Km_1	0.596	0.040	0.186
	Km_3	0.404	0.960	0.814
	(n)	156	512	77

[1]The sources of allele frequency data are as follows: Dogribs—Szathmary (1983) and Szathmary et al. (1983); New Guinea—Long et al. (1986); Kachari—Walter et al. (1986, 1987), Das et al. (1987), Deka et al. (1987), and Mukherjee et al. (1989). The chromosome frequencies for the MNSs locus in the New Guinea Highlanders were recomputed from the total phenotype data of Long et al.'s (1986) Table 3 considering the MNSs typing only. (n) refers to the No. of individuals sampled.

AMERICAN JOURNAL OF HUMAN BIOLOGY 4:387–397 (1992)

Allele Sharing at Six VNTR Loci and Genetic Distances Among Three Ethnically Defined Human Populations

RANAJIT CHAKRABORTY[1], RANJAN DEKA[2], LI JIN[1], AND ROBERT E. FERRELL[2]

[1]Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston, Texas 77225; [2]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania 15261

ABSTRACT Because of their high degree of polymorphisms, the variable number of tandem repeat (VNTR) loci have become extremely useful in studies involving gene mapping, determination of identity and relatedness of individuals, and evolutionary relationships among populations. However, there are some concerns regarding whether or not the patterns of such genetic variation can be studied by the classical population models that are developed for studying genetic variation at blood groups and protein loci, since VNTR alleles detected by molecular size may not always be identical by descent. Although theoretical and empirical studies demonstrate that this concern is overstated, this study provides further support of the application of the traditional mutation-drift models to predict the pattern of intra- and inter-populational variation at VNTR loci. By comparing genetic variation at six VNTR loci with that at 16 blood groups and protein loci in three ethnically defined populations, we show that the patterns of variability at these two sets of loci are in general parallel to each other. Shared VNTR alleles among populations are generally more frequent than the ones which are not present in every population; the proportion of shared alleles among populations increases with increasing genetic similarity of populations; and the number of VNTR alleles is positively correlated with gene diversity at these loci. All of these observations are in agreement with the prediction of the mutation-drift models, particularly when the possibility of forward-backward mutations are taken into account. This parallelism of genetic variation at VNTR loci and blood groups/protein loci further asserts the potential of using such hypervariable loci for microevolutionary studies, where closely related populations may exhibit considerably less allele frequency differences at the classical blood group and protein loci.

The discovery of the presence of tandemly repeated DNA sequences at several regions of the human genome has led to the characterization of a new class of polymorphic loci, known as variable number of tandem repeat (VNTR) loci. These loci are becoming increasingly popular in studies involving gene mapping, identification of individuals or familial relatedness, and evolutionary relationships between populations (see Burke et al., 1991). Several single locus VNTR probes have been developed (Wyman and White, 1980; Jeffreys et al., 1985; Nakamura et al., 1987) and their characterizations with regard to intra- and inter-populational genetic variation are also now available (Jeffreys et al., 1988; Clark, 1987; Flint et al., 1989; Odelberg et al., 1989; Deka et al., 1991). However, since this class of polymorphism reflects copy number variation of short conserved core sequences (ranging in length from one to hundreds of nucleotides) and the molecular mechanism of production and maintenance of VNTR alleles is still obscure, there are some concerns as to whether or not the patterns of VNTR polymorphisms

Received September 24, 1991; accepted November 12, 1991.

can be studied by the classical population genetic models developed for studying the blood group and protein variation (Lander, 1989, 1991; Kidd et al., 1991). Empirical data on intra-population variation at several VNTR loci has demonstrated that the traditional population genetic models are applicable to such loci (see Clark, 1987; Jeffreys et al., 1988; Flint et al., 1989; Chakraborty and Daiger, 1991; Chakraborty et al., 1991). Our earlier work on six VNTR loci studied in three ethnically defined populations exhibited the utility of VNTR polymorphisms for inter-populational genetic differentiation as well.

One characteristic feature of VNTR polymorphism is that these loci are generally more variable than the traditional blood groups and protein loci, reflected in higher levels of heterozygosity per locus as well as a larger number of alleles. Jeffreys et al. (1988) showed that these are caused by a higher rate of mutation at the VNTR loci in comparison to traditional loci. Such hypervariability, of course, necessitates certain modifications of statistical analyses of VNTR polymorphism data (see Chakraborty et al., 1991; Deka et al., 1991), but these are not sufficient reason to invalidate the application of the classic mutation-drift models of population genetics to analyze such data.

The purpose of this paper is to provide further evidence of comparability of intra- and inter-populational variation at the VNTR loci with those at traditional blood group and protein loci. Using a previous survey (Deka et al., 1991) of genetic variation at six VNTR loci (D1S57, D1S61, D1S76, D1S77, RB1, and α-globin 5′ HVR) in three ethnically defined populations (Kachari of Northeastern India, New Guinea Highlanders of Papua New Guinea, and Dogrib Indians of Canada), we show that among these populations the pattern of allele sharing is identical with that at 16 blood groups and protein loci studied previously in these populations. This pattern is also in accord with the prediction of a mutation-drift model of gene differentiation. Furthermore, we show that the proportion of shared alleles between populations is also parallel to their genetic similarities. Finally, the empirical relationship between the number of alleles and gene diversity at the six VNTR loci is also in agreement with the prediction of the mutation-drift model. These observations on the parallelism between the pattern of genetic variation at VNTR loci and blood groups and enzyme loci, along with the fact that the genetic distances among these populations as detected by the VNTR loci are in accordance with their historical records, suggest that such hypervariable loci are extremely useful for studying genetic differentiation between human populations. Since the DNA materials for the Kachari population used for the VNTR study (Deka et al., 1991) were all from individuals belonging to the Sonowal subgroup of Kacharis, the Kachari allele frequencies at the blood group and protein loci used in the present comparative analysis (reproduced in Appendix A), are the ones represented as that of the Sonowals (Sandwals in Walter et al., 1986) in the original papers (see Appendix A for the data source).

MATERIALS AND METHODS

Populations

For the study of comparability of patterns of polymorphisms at VNTR loci and blood groups and enzyme loci, three ethnically defined diverse human populations were selected: the Kachari of Northeastern India, Kalam and Gainj speaking New Guinea Highlanders of Papua New Guinea, and Dogrib Indians of Canada, who have been previously examined for genetic variation at several blood groups and protein loci. The Kachari is a distinct Mongoloid tribal group who live in the plains of the northeast Indian state of Assam. They belong to the Bodo subgroup of the Tibetoburman language family, and are known to have a population size exceeding 50,000 during the present century (B.M. Das, personal communication). Kalam and Gainj speaking New Guinea Highlanders represent a culturally inter-related interbreeding group of inhabitants of the northern fringe of Papua New Guinea's central highlands. They are one of the least acculturated, genetically unadmixed populations of the South Pacific (Wood et al., 1982). Although the precise estimate of the size of this population is unknown, in 1978 the de facto population size was a little over 1,100 (Long et al., 1986) and their demographic profile appears quite stable (Wood and Smouse, 1982). The Dogrib Indians are one of the 16 Athapaskan-speaking Amerindian tribes, and they reside in the Northwest Territories of Canada. Genetically, linguistically, and anthropologically, this is a well-characterized population

th low levels of non-Amerindian admixre (Szathmary et al., 1983) and from a deographic perspective they resemble to me extent the New Guinea Highlanders sofar as their current population size is ncerned (Szathmary, 1983). From the reent history of migration, there are suggesons that the Kacharis might have in their ene pool, genes of Polynesian origin, sugesting genetic proximity with the New uinea Highlanders (Walter et al., 1986).

Data

To generate data on VNTR polymorphims, igh-molecular-weight DNA was isolated om buffy coats of 45 Kachari, 46 New uinea Highlanders, and 30 Dogrib Indians y phenol-chloroform extraction (Aldridge t al., 1984). Further details of restriction ndonuclease digestion, Southern blot techique of restriction fragment length polymorphism (RFLP) analysis to detect the different VNTR alleles, and the degree of esolution of allele size determinations, long with the chromosomal localization of he six VNTR loci (D1S57, D1S61, D1S76, D1S77, RB1, and α-globin 5′ HVR) are given n Deka et al. (1991). With the exception hat the two loci D1S76 and D1S77 are closely linked ($\theta = 0.043$; O'Connell et al., 1989), the six loci represent a set of independently segregating loci at each of which genetic variation is governed by copy number variation of tandemly repeated core sequences. Deka et al. (1991) present the allele frequency distributions of each of these loci in the three populations examined here.

To contrast the pattern of VNTR polymorphisms at these six VNTR loci with that at the blood groups and protein loci, allele frequency data at 16 blood groups and protein loci which were previously published were extracted. While the New Guinea Highlanders and the Dogrib Indians were examined for a comparatively larger set of blood groups and protein loci (see Long et al., 1986; Szathmary, 1983; Szathmary et al., 1983), data on the traditional loci in the Kachari population are more limited. This led to the consideration of data at 16 blood groups and protein loci (ABO, Rh, MNSs, Diego, and Duffy blood groups, and AK, ADA, TF, ESD, GC, HP, ACP, LDH-A, HB-β, Gm, and Km protein loci) for the present analysis. Appendix A gives a compilation of allele frequencies at these loci for the three populations used in this analysis, along with their respective sample sizes (number of individuals sampled) and source of data.

Statistical methods

Patterns of polymorphism studied in this paper use several summary measures of genetic variation: (1) number of alleles, (2) gene diversity (Nei, 1973), and (3) genetic distance (bias-corrected estimate of Nei's standard genetic distance; Nei, 1978). These are directly computed from the allele frequency data (see Deka et al., 1991, and Appendix A). In order to analyze the pattern of allele sharing, the mean frequencies (and standard errors) were computed for the alleles that are present in all three, in two of the three, and in one of the three populations for both sets of loci (VNTR versus the traditional ones). The expectation is that the alleles that are present in all three populations should be more frequent than the ones which are not found in all of them. This is expected under a mutation-drift model, since Watterson and Guess (1977) have shown that the oldest allele is also likely to be the most frequent in a population, and hence, an allele which is present in all three populations is likely to have existed before the split of these three diverse populations, and thereby might also be more frequent than the newer ones which might not be shared by all. Second, for the three pairwise contrasts of populations, we computed the proportion of shared alleles, using the formula

$$P_{XY} = 2k_{XY}/(k_X + k_Y) \qquad (1)$$

where k_X and k_Y are the number of alleles in populations X and Y, and k_{XY} is the number of alleles present in both populations X and Y. The values of p_{XY} were computed for the six VNTR loci jointly and contrasted with that of the 16 blood groups and protein loci. For each choice of X and Y, the p_{XY} values were contrasted with Nei's gene identity for the corresponding loci (Nei, 1978), because under the mutation-drift model of gene differentiation, the proportion of shared alleles is expected to be positively correlated with gene identity (a measure of genetic similarity) between populations. Finally, the relationship between the number of alleles and gene diversity (heterozygosity) is studied for the six VNTR loci by plotting a scatter diagram of these two variables to examine

TABLE 1. Frequencies of shared alleles at VNTR and blood group and protein loci

	For six VNTR loci		For 16 blood group/protein loci	
Alleles present in—	Number	Frequency Mean ± S.E.	Number	Frequency Mean ± S.E.
3 populations	15	0.276 ± 0.052	30	0.488 ± 0.062
2 populations	15	0.166 ± 0.048	7	0.247 ± 0.088
1 population	17	0.041 ± 0.013	7	0.101 ± 0.063

TABLE 2. Relationship between the proportion of shared alleles (p_{XY}) and Nei's Gene Identity (I_{XY}) at the VNTR and blood groups/protein loci

	VNTR loci		Blood groups/protein loci	
Populations	I_{XY}	p_{XY}	I_{XY}	p_{XY}
Kachari vs. New Guinea	0.846	0.712	0.891	0.883
Kachari vs. Dogrib	0.654	0.581	0.870	0.892
Dogrib vs. New Guinea	0.760	0.653	0.855	0.873

whether or not these two measures of genetic variation are positively correlated. Under the mutation-drift model, these two quantities are positively correlated whose magnitude can be predicted from the observed average heterozygosity at the loci scored (Chakraborty and Griffiths, 1982).

RESULTS

Frequencies of shared and non-shared alleles

Table 1 shows the numbers and mean frequencies (± S.E.) of all alleles observed at the six VNTR loci and 16 blood groups and protein loci, categorized by their presence in all three, two of the three, and only in one of the three populations. It is clear that for both sets of loci, the pattern is identical; the alleles that are present in all three populations are more frequent than the ones which are present in two of the three populations, which are in turn more frequent than the alleles that are present in only one of the three populations. The differences of average frequencies of these three classes of alleles appear to be greater for the blood groups and protein loci, compared with the VNTR loci, which is expected, because the larger extent of polymorphism at the VNTR loci makes the individual allele frequencies generally smaller for each VNTR allele, in comparison to the blood group and protein alleles.

Relationship between proportion of shared alleles and gene identity

Table 2 shows the relationship between the proportion of shared alleles and gene identity for the two groups of loci. For the VNTR loci the correspondence between gene identity and proportion of shared alleles is perfect, reflecting that genetically similar populations are also sharing a proportionally larger number of VNTR alleles. On the contrary, for the blood groups and protein loci, such correspondence is equivocal. For example, even though the Kacharis are genetically closer to the New Guinea Highlanders, the proportion of shared blood groups/protein alleles between them is comparatively smaller than what they share with the Dogribs. There are two possible explanations for this discordance. First, since a great majority of the blood groups/protein loci are 2-allele systems (9 out of 16), the variation of the proportion of shared alleles is rather limited even among populations that are very diverse genetically. Second, the differences of the gene identity values among these populations detected at the blood groups and protein loci do not appear to be significant, as can be seen from the standard errors of the corresponding genetic distances (discussed later). In spite of these, the pattern of allele sharing at the VNTR loci indicates that even though the designation of allelic distinctions by copy number variation at these loci raises the possibility that two similar size alleles may not be of identical by descent, the present data shows that such allelic distinctions are quite adequate for the purpose of genetic comparisons between populations.

Relationship between gene diversity and number of alleles

Deka et al. (1991) demonstrated that the genotype distributions at these six VNTR loci are in accordance with the Hardy-Wein-

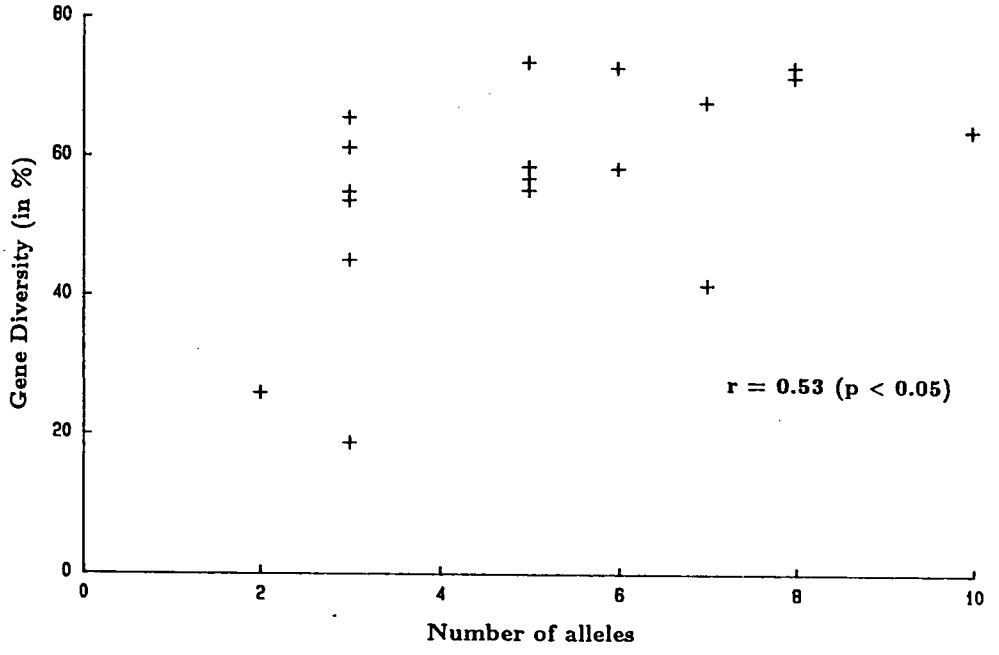


Fig. 1. Relationship between gene diversity (H) and number of alleles at six VNTR loci in three ethnically defined populations (Kachari, Dogrib, and New Guinea Highlander). The observed correlation, r = 0.526, is statistically significant ($P < 0.05$) and is in approximate agreement with the prediction of the classical infinite allele model (expected r = 0.562).

berg proportions for each of these three populations. Therefore, Nei's measure of gene diversity (estimated by the bias-correction method suggested in Nei, 1978) provides an adequate estimate of the heterozygosity at these loci. Figure 1 shows the scatter diagram of the number of alleles (k) and gene diversity (H) for the 18 population-loci combinations for the VNTR loci. It is clear that there is a positive trend in the scatter plot. Chakraborty and Griffiths (1982) have shown that, while gene diversity (H) and number of alleles (k) should be positively correlated under a mutation-drift model, their correlation is not perfect and is critically dependent on sample size. The observed correlation for the data shown in Figure 1 is 0.53, which is statistically significant ($P < 0.05$) even though the number (18) of data points is limited. Using the average gene diversity at these six loci as the base line of estimation, the expected correlation under the infinite allele model (Chakraborty and Griffiths, 1982) is 0.562, which is not statistically different from the one observed. This also reflects that designation of VNTR allelic distinctions by their lengths (or copy numbers) does not compromise the utility of such polymorphisms for evolutionary studies.

Genetic distances among populations at VNTR loci

Table 3 shows the comparison of genetic distances among these populations detected at the VNTR loci and the blood groups/protein loci. Both sets of loci demonstrate that the Kacharis and the New Guinea Highlanders are the closest of three contrasts that can be made from this data. From an anthropological view point, this is re-assuring, since the Dogrib Indians are the descendants of Mongoloid tribes that entered the New World through the Bering land-bridge around 12,000 to 15,000 years ago (at the latest), while the migration to New Guinea was much more recent and its contact with the mainland Asia was hardly uninterrupted during the history of civilization (Kirk and Szathmary, 1985). There are

TABLE 3. Genetic distances among the three populations for the six VNTR and 16 blood groups/protein Loci

Populations	Nei's standard distance (±S.E.) for VNTR Loci	Nei's standard distance (±S.E.) for blood groups/protein loci
Kachari vs. New Guinea	0.167 ± 0.067	0.115 ± 0.050
Kachari vs. Dogrib	0.424 ± 0.290	0.140 ± 0.062
Dogrib vs. New Guinea	0.275 ± 0.137	0.157 ± 0.078

some apparent discordances in the other estimates shown in Table 3. For example, the VNTR loci show that the Dogribs are the farthest from the Kacharis, while the blood group/protein loci predicts that the New Guinea and Dogrib distance is the largest. Upon a closer examination, these discrepancies can be ascribed to small number of loci used in both analyses. For both sets of loci, the genetic distances of the Dogribs from the Kacharis and New Guineans are statistically similar, because of their large standard errors. These standard errors, it should be noted, are more critically dependent on the number of loci used in the analysis, rather than the number of individuals surveyed (Nei, 1978). With this in mind, we may conclude that the genetic distance analysis of the VNTR allele frequency data is in agreement with that of the blood groups/protein loci data.

DISCUSSION

The results indicate that even when the VNTR allele designations are made from the RFLP analysis of allele sizing, the pattern of genetic variation observed at these hypervariable loci are almost parallel to the ones found at blood groups and protein loci. Nonidentity by descent of identical size VNTR alleles, which is a possibility, does not therefore compromise the utility of VNTR polymorphisms for evolutionary studies, although Kidd et al. (1991) mentioned this as a limitation of VNTR polymorphism detected through the RFLP analysis. It is true that designation of alleles by only copy number variation does not detect the allelic distinctions at the molecular level. Indeed, Deka et al. (1991) showed that the six loci studied here conform to a mutation-drift model, somewhere in between the classic infinite allele model (Kimura and Crow, 1964) and one-step forward-backward stepwise mutation model (Kimura and Ohta, 1978). The second model takes into account the hidden variation within alleles that have the same copy number of tandemly repeated core sequences. For this reason, a strict adherence to the inifinite allele model for every VNTR locus is not recommendable. Until the molecular mechanism of production of new VNTR alleles is precisely known, the exact calibration of the pattern of VNTR polymorphism data cannot be made. However, the above two models prescribe two limits of calibration, e.g., with regard to evolutionary time of divergence, rate of mutation, etc. As shown by Jeffreys et al. (1988), the correspondence between heterozygosity (determined by the proportion of two-banded genotype profiles detected by the RFLP analysis of single-locus VNTR probes) and mutation rate at several VNTR loci assures that such population genetic models are appropriate for using VNTR polymorphism data in evolutionary studies.

Even though the present study is the first direct demonstration of parallelism between the patterns of genetic variation at VNTR loci and blood groups/protein loci in ethnically defined populations, the results should be treated with some caution because of their preliminary nature. First, the sample sizes (number of individuals) for the VNTR assays are rather limited (30 to 46), which compromises the precision of the allele frequencies estimates. However, theoretical and empirical studies of sample size requirements for VNTR polymorphism indicate that the sample size limitations affect the precision of frequencies of only rare alleles, because when a limited number of individuals are sampled, rare alleles may not be observed (Evett and Gill, 1991; Chakraborty, 1992). Since the rare alleles do not contribute much to the estimation of summary measures such as gene diversity, gene identity, or genetic distance (Nei, 1978, 1987), small sample size does not constitute a major drawback of the qualitative conclusions reached in the present analysis. Second, there are also concerns that the RFLP analysis of allele detection results in incomplete resolution of true alleles. Although some critics advocate that a discrete allele theory is not applicable to such quasi-continuous variation of allele sizes, we argue that we followed a uniform protocol of allelic resolution for all sampled individuals, enabling us to detect minute differences between even

closely spaced fragments, which would have been missed by simply comparing different autoradiograms (see Deka et al., 1991 for details). Thus, the data can be treated fairly accurately with a discrete allele model, completely parallel to the treatment of the blood groups and protein alleles.

We postulated that the alleles that are present in all three populations examined here have supposedly existed before the split of these populations. This conclusion should also be treated with some caution, since there are suggestions that should the molecular mechanism of production of copy number variation of VNTR alleles behave in some form of forward-backward events, the hidden variability is expected to be more pronounced in the frequent alleles (Nei and Chakraborty, 1976; Chakraborty and Nei, 1976). Since the shared alleles are seen more frequently in our analysis (see Table 1), further experimental studies are needed to establish their molecular identity. To this end, we might note that Boerwinkle et al. (1989) and Jeffreys et al. (1990) have noted molecular heterogeneity of identical size VNTR alleles. Our speculation is that such molecular heterogeneity should exist more among these ancestral frequent alleles, for which internal mapping of shared VNTR alleles are currently being attempted with polymerase-chain-reaction (PCR) based sequencing studies. If this speculation is correct, the utility of VNTR polymorphism will be even greater for studying micro-evolutionary divergence between genetically closely related populations.

The present study also indicates that in spite of the caveats of allelic designations achieved by RFLP analysis of VNTR alleles, the degree of genetic variation detected by VNTR loci is larger than that at the blood groups/protein loci. This is reflected in higher genetic distance for these loci as well (Table 3). As a comparison, the average gene diversity for the 16 blood groups/protein loci for the three populations are 0.338 ± 0.050 (Kachari), 0.213 ± 0.051 (Dogrib), and 0.194 ± 0.056 (New Guinea Highlander). These are considerably lower than the levels of gene diversity at the six VNTR loci (0.670 ± 0.024, 0.432 ± 0.071, and 0.576 ± 0.043, respectively). Although these values should not be interpreted in absolute terms, since the loci are heavily biased towards being polymorphic, they provide a number of interesting implications with regard to the evolutionary history of the populations examined and the biology of the loci studied.

First, the Kacharis appear to be the most variable of the three populations studied at both sets of loci. Since the above contrasts are based on a set of common loci examined for each populations, they reflect that perhaps the Kacharis have a larger effective population size compared with the others. The history of these populations support this view, since at least during the past century the Kacharis maintained a considerably larger census size than the other two populations. The larger effective size in the Kacharis may also have been caused by a substantial amount of gene admixture in this population. Being situated in the northeastern corridor of the Indian subcontinent, this population received genes of Caucasian as well as Mongoloid ancestry (Walter et al., 1986, 1987), while the other two populations are relatively less admixed. Szathmary (1983) estimated that the maximum amount of non-Amerindian admixture in the Dogrib population could be 8.7%, while Long et al. (1986) asserted that the New Guinea Highlanders are perhaps the most unaccultrated. Although the average gene diversity levels at the six VNTR loci in the Dogrib and New Guinea Highlander populations are not significantly different (at 5% level), Szathmary et al.'s (1983) estimate of average gene diversity ($12.8 \pm 3.0\%$) in the Dogribs for a larger set of blood groups/protein loci (36 loci) is almost 2.5-times larger than that for the New Guinea Highlanders ($5.3 \pm 1.4\%$) estimated by Long et al. (1986) at an even larger set of 39 loci. This indicates that perhaps some VNTR variants may have been either missed in our survey of 30 Dogrib individuals analyzed here, or because of small size of this population, there had been a true loss of genetic variation in this tribal population. Second, under some restrictive assumptions these comparative data provide indirect estimate of the mutation rate at the VNTR loci as well. For example, under the two extreme mutation models, the expected gene diversities in an equilibrium population are given by

$$H = \begin{cases} 4N_e v/(1 + 4N_e v), & \text{for the infinite allele model,} \quad (2.1) \\ [(1 + 8N_e v)^{1/2} - 1]/(1 + 8N_e v)^{1/2}, & \text{for the stepwise mutation model.} \quad (2.2) \end{cases}$$

TABLE 4. Estimates of relative mutation rate at VNTR loci compared with blood groups and protein (BG/P) loci

Populations	Average gene diversity		Estimates of relative mutation rate	
	VNTR loci	BG/P Loci[1]	Using Equation 2.1	Using Equation 2.2
New Guinea	0.576	0.053	24.3	39.7
Kachari	0.670	0.139	12.6	23.5
Dogrib	0.432	0.118	5.7	7.3

[1]These estimates are adjusted assuming that the additional loci surveyed in Long et al. (1986) would be monomorphic in the Dogrib and Kachari populations.

N_e being the effective size and v, the mutation rate per locus per generation (Kimura and Crow, 1964; Ohta and Kimura, 1973), from which the relative mutation rate at these two sets of loci can be crudely estimated following Zouros (1979). Table 4 presents the result of such computations, where for uniformity all estimates of gene diversity are adjusted for the largest set of 39 loci (examined in Long et al., 1986), assuming that these additional loci would be monomorphic in the Kachari and Dogrib populations as well.

Although these estimates are quite crude, they reflect that the rate of mutation at these six VNTR loci is between 6- and 40-fold of that at the blood groups and protein loci. Assuming that the traditional loci mutate at a rate of 1.1×10^{-5}/locus/generation (Chakraborty and Neel, 1989), the rate of VNTR mutability would become somewhere between 6.6×10^{-5} to 4.4×10^{-4} per locus per generation. While these estimates are considerably lower than the rate of spontaneous mutations at the VNTR loci, as reported by Jeffreys et al. (1988), these are in the range seen in the studies of Wolff et al. (1988), Chakraborty and Daiger (1991) and Edwards et al. (1992).

Using the same equations (2.1 and 2.2), we can also estimate the relative effective sizes of the Dogrib and New Guinea populations in comparison to that of the Kacharis. Such calculations suggest that when the VNTR data are used, the effective size of the New Guinea Highlanders appear to be about 56 to 67% of that of the Kacharis, while the Dogrib size is about 26 to 37% of that of the Kacharis. The estimates for the blood groups/protein data are about 33 to 35% and 82 to 84%, respectively. Although these estimates are rather variable (and perhaps, quite imprecise), qualitatively we may conclude that the Kacharis have a comparatively larger effective population size, achieved due to the fact that they have intermixed with larger Caucasian as well as Mongoloid gene pools. For the additional analyses we assumed that the three populations are, within themselves, homogeneous. Although, Szathmary (1983), Szathmary et al. (1983) and Long et al. (1986) examined the extent of substructuring within the Dogribs and New Guineans, the coefficient of gene diversity (either by Nei's measure, G_{ST}; Nei, 1973; or by Wright's F_{ST}; Weir and Cockerham, 1984; Long, 1986) are not large enough to substantially change the above qualitative conclusions with regard to either the relative mutation rate, or the relative effective size. In fact, the allele frequency spectra in the total population data satisfies the premises of homogeneity within each of these three populations (details of such data not shown, but can be inferred from the allele frequency data of Deka et al., 1991 and Appendix A), for both sets of loci.

In conclusion, this preliminary comparative study of genetic variation at the six VNTR loci and 16 blood groups/protein loci exhibits that the pattern of genetic variation at the VNTR loci, with allelic designations determined by RFLP analysis, is parallel to that of blood groups and protein variation. In addition to this empirical support of the application of a mutation-drift model to examine the features of VNTR polymorphism data, we might note that Ohta (1986) provides a theoretical treatment, suggesting that the infinite allele model may also be appropriate to study genetic variation detected by copy number variation of repeated short DNA sequences. While the theoretical predictions of the summary measures of genetic variation are dependent on the assumption that the populations are at equilibrium due to mutation-drift balance, previous analysis of allele frequency distributions at the loci studied here suggests that this assumption approximately holds for the VNTR polymorphisms in these populations (Deka et al., 1991). The caveats of the

RFLP typing of VNTR alleles, mentioned here, can be circumvented by more refined methods (such as PCR-based techniques, and internal mapping of specific VNTR alleles), and in fact, the above discussions indicate that such future studies should strengthen the notion that the hypervariability present at the VNTR loci will be extremely useful for microevolutionary studies, where gene frequency variation at traditional blood groups and protein loci may give equivocal results.

ACKNOWLEDGMENTS

This work was supported in part by NIH grant GM 41399 from the National Institutes of Health and grant 90-IJ-CX-0038 from the U.S. National Institute of Justice (both to R.C.), and the Central Research Development Fund Award 5-33344 from the University of Pittsburgh (to R.D.). The comments of an anonymous reviewer were helpful in simplifying the presentation. Points of view or opinions are, however, the sole responsibility of the authors.

LITERATURE CITED

Aldridge J, Kunkel L, Bruns G, Tantaravahi U, Lalande M, Brewster T, Moreau E, Wilson M, Bromley W, Roderick T, and Latt SA (1984) A strategy to reveal high-frequency RFLPs along the human X chromosome. Am. J. Hum. Genet. *36*:546–564.

Boerwinkle E, Xiong W, Fourest E, and Chan L (1989) Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: Application to the apolipoprotein B 3′ hypervariable region. Proc. Natl. Acad. Sci. U.S.A. *86*:212–216.

Burke T, Dolf G, Jeffreys AJ, and Wolff R (eds.) (1991) DNA Fingerprinting: Approaches and Applications. Basel: Birkhäuser.

Chakraborty R (1992) Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Hum. Biol. *64*:141–159.

Chakraborty R, and Daiger SP (1991) Polymorphisms at VNTR loci suggest homogeneity of the White population of Utah. Hum. Biol. *63*:571–587.

Chakraborty R, Fornage M, Gueguen R, and Boerwinkle E (1991) Population genetics of hypervariable loci: Analysis of PCR based VNTR polymorphism within a population. In T Burke, G Dolf, AJ Jeffreys, and R Wolff (eds.): DNA Fingerprinting: Approaches and Applications. Basel: Birkhäuser, pp. 127–143.

Chakraborty R, and Griffiths RC (1982) Correlation of heterozygosity and number of alleles in different frequency classes. Theor. Pop. Biol. *21*:205–218.

Chakraborty R, and Neel JV (1989) Description and validation of a method for simultaneous estimation of effective population size and mutation rate from human population data. Proc. Natl. Acad. Sci. U.S.A. *86*:9407–9411.

Chakraborty R, and Nei M (1976) Hidden genetic variability in electromorphs in finite populations. Genetics *84*:385–393.

Clark AG (1987) Neutrality tests of highly polymorphic restriction fragment length polymorphisms. Am. J. Hum. Genet. *41*:948–956.

Das BM, Walter H, Gilbert K, Lindenberg P, Malhotra KC, Mukherjee BN, Deka R, and Chakraborty R (1987) Genetic variation of five blood group polymorphisms in ten populations of Assam, India. Int. J. Anthropol. *2*:325–340.

Deka R, Chakraborty R, and Ferrell RE (1991) A population genetic study of six VNTR loci in three ethnically defined populations. Genomics *11*:83–92.

Deka R, Gogoi BC, Hundrieser J, and Flatz G (1987) Hemoglobinopathies in Northeast India. Hemoglobin *11*:531–538.

Edwards A, Hammond HA, Caskey CT, Jin L, and Chakraborty R (1992) Population genetics of trimeric and tetrameric tandem repeats in four human ethnic groups. Genomics (in Press).

Evett IW, and Gill P (1991) A discussion of the robustness of methods assessing the evidential value of DNA single locus profiles in crime investigations. Electrophoresis *12*:226–230.

Flint J, Boyce AJ, Martinson JJ, and Clegg JB (1989) Population bottlenecks in Polynesia revealed in minisatellites. Hum. Genet. *83*:257–263.

Jeffreys AJ, Neumann R, and Wilson V (1990) Repeat unit sequence variation in *minisatellites*: A novel source of DNA polymorphism for studying variation and mutation by single nucleotide analysis. Cell *60*:473–485.

Jeffreys AJ, Royle NJ, Wilson V, and Wong Z (1988) Spontaneous mutation rates to new length alleles at randem-repeatitive loci in human DNA. Nature *332*:278–281.

Jeffreys AJ, Wilson V, and Thein SL (1985) Hypervariable "minisatellite" regions of human DNA. Nature *314*:67–73.

Kidd JR, Black FL, Weiss KM, Balazs I, and Kidd KK (1991) Studies of three Amerindian populations using nuclear DNA polymorphisms. Hum. Biol. *63*:775–794.

Kimura M, and Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics *49*:725–738.

Kimura M, and Ohta T (1978) Stepwise mutation model and distribution of allele frequencies in a finite population. Proc. Natl. Acad. Sci. U.S.A. *75*:2868–2872.

Kirk RL, and Szathmary E (1985) Out of Asia. Canberra: Australian National University Press.

Lander ES (1989) DNA fingerprinting on trial. Nature *339*:501–505.

Lander ES (1991) Invited editorial: Research on DNA typing catching up with courtroom applications. Am. J. Hum. Gnet. *48*:819–823.

Long JC (1986) The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's F-statistics. Genetics *112*:629–647.

Long JC, Naidu JM, Mohrenweiser HW, Gershowitz H, Johnson PL, Wood JW, and Smouse PE (1986) Genetic characterization of Gainj- and Kalam-speaking peoples of Papua New Guinea. Am. J. Phys. Anthropol. *70*:75–96.

Mukherjee BN, Malhotra KC, Roy M. Banerjee S. Walter H, Chakraborty R (1989) Genetic heterogeneity and population structure in eastern India: Red cell enzyme variability in ten Assamese populations. Z. Morph. Anthropol. 77:287–296.

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T,

Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, and White R (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science *235*:1616–1622.

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. U.S.A. *70*:3321–3323.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics *89*:583–590.

Nei M (1987) Molecular Evolutionary Genetics. New York: Columbia University Press.

Nei M, and Chakraborty R (1976) Electrophoretically silent alleles in a finite population. J. Mol. Evol. *8*: 381–385.

O'Connell P, Lathrop GM, Nakamura Y, Leppert ML, Ardinger RH, Murray JL, Lalouel JM, and White R (1989) Twenty-eight loci form a continuous linkage map of markers for human chromosome 1. Genomics *4*:12–20.

Odelberg SJ, Paltke R, Eldridge JR, Ballard L, O'Connell P, Nakamura Y, Leppert M, Lalouel JM, and White R (1989) Characterization of eight VNTR loci by agarose gel electrophoresis. Genomics *5*:915–924.

Ohta T (1986) Actual number of alleles contained in a multigene family. Genet. Res. *48*:119–123.

Ohta T, and Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. *22*:201–204.

Szathmary EJE (1983) Dogrib Indians of the Northwest Territories, Canada: Genetic diversity and genetic relationship among subarctic Indians. Ann. Hum. Biol. *10*:147–162.

Szathmary EJE, Ferrell RE, and Gershowitz H (1983) Genetic differentiation in Dogrib Indians: Serum protein and eyrthrocyte enzyme variation. Am. J. Phys. Anthropol. *62*:249–254.

Walter H, Matsumoto H, Miyasaki T, Mukherjee BN, Malhotra KC, Das BM, Gilbert K, and Lindenberg P (1987) Distribution of Gm and Km allotypes among ten populations of Assam, India. Am. J. Phys. Anthropol. *73*:439–445.

Walter H, Mukherjee BN, Gilbert K, Lindenberg P, Dannewitz A, Malhotra KC, Das BM, and Deka R (1986) Investigations on the variability of haptoglobin, transferin and Gc polymorphisms in Assam, India. Hum. Hered. *36*:388–396.

Watterson GA, and Guess HA (1977) Is the most frequent allele the oldest? Theor. Pop. Biol. *11*:141–160.

Weir BS, and Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution *38*:1358–1370.

Wolff RK, Nakamura Y, and White R (1988) Molecular characterization of a spontaneously generated new allele at a VNTR locus: No exchange of flanking DNA sequence. Genomics *3*:347–351.

Wood JW, Johnson PL, Kirk RL, McLoughlin K, Blake NM, and Matheson FA (1982) The genetic demography of the Gainj of Papua New Guinea. I. Local differentiation of blood group, red cell enzyme, and serum protein allele frequencies. Am. J. Phys. Anthropol. *57*:15–25.

Wood JW, and Smouse PE (1982) A method of analyzing density-dependent vital rates with an application to the Gainj population of Papua New Guinea. Am. J. Phys. Anthropol. *58*:304–411.

Wyman AR, and White R (1980) A highly polymorphic locus in human DNA. Proc. Natl. Acad. Sci. U.S.A. *77*:6754–6758.

Zouros E (1979) Mutation rates, population sizes, and amounts of electrophoretic variation of enzyme loci in natural populations. Genetics *92*:623–646.

APPENDIX A. Allele frequency data at 16 blood groups and protein loci in the Kacharis, New Guinea Highlanders, and Dogrib Indians[1]

Locus	Allele	Dogrib	New Guinea	Kachari
ABO	A_1	0.177	0.347	0.171
	A_2	—	—	0.041
	B	—	0.126	0.186
	O	0.823	0.527	0.602
	(n)	158	415	107
Rh	CDe	0.221	0.933	0.812
	cDE	0.695	0.021	0.064
	CDE	0.014	0.010	0.047
	cDe	0.023	0.036	0.040
	cde	0.047	—	0.037
	(n)	158	412	107
MNSs	MS	0.095	0.015	0.071
	Ms	0.820	—	0.672
	NS	0.022	0.111	0.023
	Ns	0.063	0.874	0.234
	(n)	158	229	107
Diego	Di^a	—	—	0.028
	Di^b	1.0	1.0	0.972
	(n)	158	54	107
Duffy	Fy^a	0.953	0.994	0.744
	Fy^b	0.047	0.006	0.256
	(n)	158	390	107
AK	AK_1	1.0	1.0	0.948
	AK_2	—	—	0.052
	(n)	158	277	106
TF	Tf^C	1.0	0.946	1.0
	Tf^D	—	0.054	—
	(n)	158	575	64
ESD	ESD_1	0.826	0.934	0.617
	ESD_2	0.174	0.066	0.383
	(n)	158	570	107
GC	Gc^1	0.930	0.825	0.769
	Gc^2	0.070	0.152	0.231
	Gc^{Ab}	—	0.023	—
	(n)	158	522	104
HP	Hp_1	0.361	0.739	0.240
	Hp_2	0.639	0.261	0.760
	(n)	158	470	104
ACP	p^A	0.462	0.270	0.262
	p^B	0.538	0.730	0.738
	(n)	158	570	107
ADA	ADA_1	0.997	0.973	0.723
	ADA_2	0.003	0.026	0.277
	ADA_6	—	0.001	—
	(n)	158	570	83
HB-β	Hb_A	1.0	1.0	0.493
	Hb_E	—	—	0.507
	(n)	158	588	1082
LDH-A	Normal	1.0	1.0	1.0
	(n)	158	589	107
Gm	a;g	0.789	0.398	0.021
	a,x:g	0.066	0.045	0.117
	f;b0, 1, 3	0.010	—	0.201
	a;b	0.135	0.126	—
	f,a;b	—	0.311	0.523
	Others	—	0.120	0.138
	(n)	156	513	76
Km	Km_1	0.596	0.040	0.186
	Km_3	0.404	0.960	0.814
	(n)	156	512	77

[1]The sources of allele frequency data are as follows: Dogribs—Szathmary (1983) and Szathmary et al. (1983); New Guinea—Long et al. (1986); Kachari—Walter et al. (1986, 1987), Das et al. (1987), Deka et al. (1987), and Mukherjee et al. (1989). The chromosome frequencies for the MNSs locus in the New Guinea Highlanders were recomputed from the total phenotype data of Long et al.'s (1986) Table 3 considering the MNSs typing only. (n) refers to the No. of individuals sampled.

DNA Fingerprinting: Approaches and Applications
ed. by T. Burke, G. Dolf, A. J. Jeffreys & R. Wolff

Population Genetics of Hypervariable Loci: Analysis of PCR Based VNTR Polymorphism Within a Population

R. Chakraborty[a], M. Fornage[a,b], R. Gueguen[b], and E. Boerwinkle[a]

[a]*Genetics Centers, University of Texas Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, Texas 77225, USA;* [b]*Center for Preventive Medicine, Nancy, France*

Summary
Using a polymerase chain reaction (PCR) based method, genotypes at two hypervariable loci (3′ to the Apo-B-structural gene and at the ApoC-II gene) were determined by size classification of alleles. Genotype data at the Apo-B locus (Apo-B VNTR) were obtained on 240 French Caucasians; the sample size for the ApoC-II VNTR was 162. For 160 individuals two-locus genotype data were available. Applications of some recently developed statistical methods to these data indicate that both of these loci are at Hardy-Weinberg equilibrium (HWE) and there is no indication of allelic associations between these two unlinked loci. In addition, the observed numbers of alleles (12 for the Apo-B and 11 for the ApoC-II VNTR loci) are also consistent with their respective expectations based on the observed heterozygosities (76.9% for the Apo-B and 85.9% for the ApoC-II loci) suggesting genetic homogeneity of this population-based sample. The multimodal distribution of allele sizes observed for both loci indicate that the production of new alleles at such VNTR loci may be caused by more than one molecular mechanism. The utility of such highly polymorphic loci for human genetic research and forensic applications are discussed in the context of these findings.

Introduction

A large number of DNA segments in the human genome contain a variable number of short tandemly repeated sequences. To varying degree, the core repeat unit is conserved from one repeat to another. The core sequence may also vary among such loci from 2 bases to several kilobases. The copy number of such core sequences reveals genetic variation several orders larger than that detected by classical serologic and biochemical genetic markers. Since the first demonstration of such a highly polymorphic, locus-specific sequence in the human genome by Wyman and White (1980), numerous hypervariable regions have been described that flank structural loci in the human genome (e.g., Bell *et al.*, 1982; Capon *et al.*, 1983; Goodbourn *et al.*, 1983; Jeffreys *et al.*, 1985; Boerwinkle *et al.*, 1989; Ludwig *et al.*, 1989; reviewed also by Jeffreys and Wolff, this volume). It is now well-recognized that the human genome contains a large number of these polymorphic segments, numbering possibly in the thousands. Several acronyms of such polymorphic systems are proposed. Jeffreys *et al.*

(1985) suggested locus-specific 'minisatellites'; Nakamura *et al.* (1987) coined the term 'VNTR' (Variable Number of Tandem Repeats), while several others (e.g., Balazs *et al.*, 1989; Ludwig *et al.*, 1989) used the terminology 'HVR' (Hypervariable Region) to describe this type of genetic variation. When the core unit is small (di-, tri, or tetra-nucleotide), Edwards *et al.* (1991) called them 'STR'-loci (Short Tandemly Repeated loci).

Since the allelic designation at these VNTR loci can be conveniently defined by the number of repeat units of the core sequence and each locus conforms to simple codominant Mendelian mode of inheritance, such loci are extremely useful for human genetic research. Recent tabulation of genetic markers used for human gene mapping indicates that collectively nearly 50 per cent of all genetic markers belong to this category (Kidd *et al.*, 1989). The increased efficiency of VNTR loci, compared to classical biochemical and RFLP markers, arises from the fact that the number of different alleles found at any VNTR locus is generally much larger. In addition, these VNTR loci have a high heterozygosity (sometimes approaching levels as high as 95–99 percent). As a consequence, detection of recombination between a VNTR locus and a disease gene or other genetic markers is simple because both parents can be heterozygous and provide four distinct alleles at such loci far more commonly than other classical markers.

The presence of large numbers of alleles at VNTR loci also makes them useful in the context of paternity testing and forensic medicine. A growing body of literature suggests that their utility is not only limited to academic circles and biomedical research; criminal justice and social welfare agencies can also benefit immensely from the application of such loci (Craig *et al.* 1988).

These advantages notwithstanding, concerns have been raised with regard to the population genetic characteristics of such polymorphisms (Lander, 1989; Cohen, 1990). In part this is caused by limited population data of VNTR polymorphisms from genetically well-characterized populations. Five studies in this regard are noteworthy. Baird *et al.* (1986) have shown extensive variation at two VNTR loci, HRAS-1 and D14S1, in three major ethnic groups. Using such data, Clark (1987) postulated that allelic variation at these loci follow the mutation-drift model of genetic variation (Kimura and Crow, 1964). Jeffreys *et al.* (1988) entertained other models of allele frequency distribution such as a finite allele mutation model or a stepwise mutation model to explain the relationship between heterozygosity and mutation rate at such loci. Flint *et al.* (1989) showed that the differences of heterozygosity levels of specific VNTR loci across populations can be used to postulate whether or not events such as a population bottleneck could have occurred during the geographic dispersal of humans. Chakraborty (1990a, b)

argued that even a single VNTR locus can provide information concerning substructuring within a population with a statistical power far greater than several classical genetic markers studied simultaneously.

The well known advantages of VNTR loci such as their high heterozygosity and a large number of alleles also poses problems in the statistical interpretation of population data. For example, the presence of a large number of alleles increases sample size requirements for population survey studies, since even with respectable sample sizes all possible genotypes can not be generally detected. As a result, precisions of allele frequency estimates are compromised which subsequently hinders statistical calculations based on classical text book methods of analysis. Since all genotypes are not observed, and even the ones observed occur in low frequencies, the application of the large sample theory of chi-square goodness-of-fit test is not valid for checking whether or not the observed genotypic distribution conforms to their Hardy-Weinberg expectations. Similarly, the association among alleles at multiple loci cannot be adequately determined from their genotypic distributions by standard summary measures such as linkage disequilibrium. In addition, a single VNTR locus may exhibit more than one allele of similar size and incomplete resolution of distinct alleles by Southern gel electrophoresis is not uncommon. Such inescapable laboratory phenomena can not be overlooked in the statistical interpretation of VNTR population data (Devlin *et al.*, 1990). There are also claims that the hypervariability at these loci is caused by 'mutational' changes whose rate is high in comparison with other classical markers. Hence, when used singly, a VNTR locus may lead to a wrong conclusion regarding biological relationships between individuals (Odelberg *et al.*, 1989). Furthermore, very little is known about the molecular mechanism of production of new alleles at VNTR loci, although it has been suggested that replication slippage, sister chromatid exchange, and/or unequal recombination may be involved in this process. Virtually nothing is known regarding the functional requirements of such loci in general, even though DNA-binding proteins which appear to bind specifically to VNTR loci are known to exist (Collick and Jeffreys, 1990). In view of these uncertainties, it is difficult to determine the mode and rate of evolution of genetic variation at VNTR loci.

In spite of these difficulties, we believe that the statistical interpretation of VNTR polymorphism data is not an insurmountable problem. Certain modifications of classic population genetic methods of data analysis can be introduced which lead to rigorous and legitimate estimation and hypothesis-testing principles for analyzing such data. Such methods also may suggest molecular mechanisms that generate and maintain genetic variation at these loci. We have initiated a series of studies on the population genetics of VNTR polymorphism at our Center, and this presentation is a preliminary summary of several results from

these studies. Two VNTR loci have been typed by PCR-based experimental protocols which can be employed for population surveys at such loci and which minimize the problem of incomplete resolution of alleles (Boerwinkle *et al.*, 1989). Here we describe the genotype and allele frequency distributions at two VNTR loci; one locus is 3′ to the apolipoprotein-B (ApoB) gene and the other is within the ApoC-II gene, which have been scored on individuals belonging to 121 nuclear families. Several alternative methods are suggested for examining the conformity of the genotypic distribution of VNTR data with HWE and for testing independent segregation of alleles at unlinked loci in the presence of large number of alleles. Futhermore, the allele frequency distributions of these loci are used to predict possible molecular mechanisms that generate and maintain such genetic variation. Characterization of such population genetic features implies that VNTR polymorphisms detected through PCR-based studies conform to classic population genetic principles, and hence are useful in human genetic research and forensic applications.

Materials and Methods

The genotype data analyzed here were obtained from a random sample of 121 nuclear families taking part in routine health examinations at the Center for Preventive Medicine in Nancy, France. Genomic DNA was isolated by phenol/chloroform extraction of proteinase K treated crude buffy coat preparations. Two VNTR loci with different core sequences were selected for the present analyses. The first is an AT-rich VNTR 3′ to the human apolipoprotein B gene on chromosome 2 (Huang and Breslow, 1987). Detailed PCR-based methods for typing this VNTR have been previously presented (Boerwinkle *et al.*, 1989). Our previous results indicate that this locus differs in the number of copies (from 29 to 51) of a conserved core sequence 14 or 15 base pair long (Boerwinkle *et al.*, 1989). The second VNTR locus is a microsatellite located in the first intron of the human apolipoprotein C-II gene on chromosome 19 (Fojo *et al.*, 1987). This ApoC-II VNTR consists of a $(TG)_n(AG)_m$ core motif repeated from 16 to 34 times. The oligonucleotides used for priming the PCR bind immediately adjacent to the $(TG)_n(AG)_m$ block. One member of the pair of primers was 5′ end-labeled using bacteriophage T4 kinase. The PCR was carried out in a 50 μl volume containing approximately 0.5 μg of genomic DNA and samples were processed through 30 temperature cycles consisting of 1 minute at 92°C (denaturation), 1 minute at 55°C (annealing), and 1.5 minutes at 72°C (elongation). After 10 cycles with only cold oligonucleotide the reaction mixture was spiked with the end-labeled oligonucleotide for the remaining 20 cycles. The amplified DNA was analyzed after being electrophoresed on 8% denaturing polyacrylamide gels and

exposed to X-ray film for 20 hours. Size standards were created by dideoxy sequencing using M13 mp18 as a template. The size of the PCR products were directly determined from the size of the co-migrating M13 fragment in the ladder (data not shown).

The family data were used to verify Mendelian segregation of the identified alleles at the ApoB and ApoC-II VNTR loci. The parents of these families represent a sample of unrelated individuals and were used for allele frequency estimation and other calculations. Therefore, these data can be regarded as from a random sample of Caucasian individuals of French ancestry. Genotype data on 240 individuals for the ApoB locus and 162 individuals for the ApoC-II locus were used for the analyses presented here. Two locus genotype data were available for 160 individuals.

Because one purpose of this paper is to describe the analytical tools for the analysis of allele and genotype frequency data at VNTR loci, the statistical methods are not presented in this section but rather are given along with a corresponding question and resulting inference in the next section. Statistical analyses consist of: (1) analysis of genotype and allele frequency distribution for each locus individually to test whether or not HWE predictions hold for these two loci, (2) joint analysis of two-locus genotype data to determine independent segregation of alleles at these two unlinked loci; (3) examination of the relationship between heterozygosity and the number of alleles at each locus to determine the underlying mechanism of production of new alleles at these loci, and finally (4) to postulate possible reasons for the shape of the allele size distributions at these loci.

Results

Single-Locus Genotypic Distributions

Typically VNTR locus variation is codominant and multi-allelic. Letting k be the number of segregating alleles, there are $k(k+1)/2$ possible genotypes at a locus, k of which are homozygous A_iA_i ($i = 1, 2, \ldots, k$) and $k(k-1)/2$ are heterozygous A_iA_j ($i < j = 2, 3, \ldots, k$). It should be possible to observe all possible $k(k+1)/2$ genotypes in any given sample. However, the sample size needed to observe all possible genotypes is generally large when k is large.

We observed 12 different alleles at the ApoB VNTR locus and 11 at the ApoC-II VNTR locus in a sample of 240 and 162 individuals, respectively. The number of possible genotypes ($k(k+1)/2$), therefore, are 78 and 66, respectively. Table 1 shows the observed genotype and allele frequency distributions at the ApoB locus, demonstrating that even though the sample size ($n = 240$) is much larger than the number

of possible genotypes, we observed only 42 distinct genotypes. Of the possible 12 homozygote genotypes only 8 are observed and of the possible 66 heterozygote genotypes only 34 are observed. A large fraction of the observed genotypes have frequencies below 5 individuals per genotype.

Obviously, such data cannot be subjected to a classical goodness-of-fit test examining whether or not the genotypic distribution conforms to HWE predictions. This is so because a chi-square approximation is inaccurate when the observed frequencies are small, and also we cannot assign any well-defined degrees of freedom because of the absence of a large fraction of genotypes in the sample. Three alternative test statistics are used for this purpose. First, grouping the entire data of the genotype table (Tab. 1) into two classes, heterozygotes and homozygotes (irrespective of the allele types, observed), we found that there are 59 individuals that are homozygous at this locus and 181 heterozygous. The expected numbers of individuals in these two categories, based on the observed allele frequencies (shown on the last column of Tab. 1), are 55.38 and 184.62, respectively, and their difference from the observed ones is not significant by a chi-square test with 1 d.f. ($P > 0.55$). Second, using the theory of Chakraborty *et al.* (1988) and Chakraborty (1991), we also determined the expectations (and their standard errors) of the number of distinct heterozygous and homozygous genotypes that can be observed in a sample of 240 individuals at the ApoB VNTR locus when the alleles unite at random to form the genotypes of these individuals. The sampling distributions of these statistics, under the HWE assumption, can be evaluated exactly (Chakraborty, 1991). In particular, denoting the expected genotype frequencies (under HWE) by $Q_1, Q_1, \ldots, Q_K$, the mean and variance of the number of distinct

Table 1. Genotype and allele frequency distributions at the ApoB VNTR Locus in a random sample of 240 individuals from Nancy, France

	Alleles												
	29	31	33	35	37	39	41	43	45	47	49	51	Freq.
29	—	1	1	—	—	—	—	—	—	—	—	—	2
31		3	2	7	6	1	2	1	—	2	2	1	31
33			2	2	10	2	—	—	—	1	3	1	26
35				12	50	1	2	—	—	10	6	2	104
37					36	11	4	—	2	16	17	6	194
39						2	1	—	—	—	3	2	25
41							1	—	—	—	—	—	11
43								—	—	—	—	—	1
45									—	—	—	—	2
47										1	2	1	34
49											2	—	37
51												—	13

genotypes observed in a sample of n individuals are given by

$$\mu = K - T_1, \quad \text{and} \quad \sigma^2 = T_1(1 - T_1) + 2T_2, \tag{1a}$$

respectively, where

$$T_1 = \sum_{i=1}^{K} (1 - Q_i)^n, \quad \text{and} \quad T_2 = \sum\sum_{i>j=1}^{K} (1 - Q_i - Q_j)^n. \tag{1b}$$

Note that in equations (1) and (2) $K = k$ for the homozygous and $Q_i = p_i^2$, the square of the i-th allele frequency, and $K = k(k-1)/2$ for the heterozygous genotypes, with Q_i's being an array of $k(k-1)/2$ probabilities representing each heterozygote genotypes ($2p_ip_j$). Application of this method to the present data showed that the observed numbers of distinct homozygous and heterozygous genotypes for the ApoB locus, 8 and 34, are not significantly different from their expectations, 6.37 and 32.98. Therefore, we conclude that the observed numbers of distinct genotypes are also in concordance with HWE predictions.

Since all of these statistics disregard the specific allele types observed in the sample, and hence these tests do not detect deviations of each specific genotype frequency from its HWE prediction, we used a third test criterion which does not have this limitation. This is the G-statistic (Sokal and Rohlf, 1969), a likelihood ratio, which should not be significant if the HWE prediction is correct. Unfortunately, although the G-statistic is a contrast of every observed genotype frequencies with their respective expectations, no standard statistical distribution can be applied to determine the significance level of the G-statistic because of the absence (or small frequencies) of several genotypes. We employed a shuffling algorithm to determine the empirical distribution of the G-statistic, by randomly permuting the 480 allele labels (12 of them, since there are 12 different observed alleles) and reconstructing genotypes by pairing the shuffled alleles at random. The observed value of G in the given data (of Tab. 1) was 60.18, and its empirical probability level was 0.62, suggesting that by chance 62% times we could have observed G-values larger than the one observed under the assumption of HWE. Therefore, none of the three alternative tests offered any suggestion of non-random association of alleles at the ApoB VNTR locus. In conclusion, the genotype distribution at this locus among the French Caucasians can be assumed to satisfy the HWE predictions. The observed heterozygosity at this locus is 0.745 ± 0.028, and its expectation based on the estimated allele frequencies is 0.769 ± 0.014.

Table 2 shows the observed genotypic and allele frequency distributions at the ApoC-II VNTR locus. Eleven segregating alleles are found at this locus, but of the possible 66 genotypes only 42 are observed. As in the case of the ApoB VNTR, each of the three alternative test criteria shows that the observed genotypic distribution is in accordance with the

Table 2. Genotype and allele frequency distributions at the ApoC-II VNTR Locus in a random sample of 162 individuals from Nancy, France

	Alleles											
	16	20	24	25	26	27	28	29	30	31	34	Freq.
16	3	3	3	1	1	6	6	5	7	1	1	40
20		2	1	—	—	6	7	4	8	—	1	34
24			—	—	—	1	1	5	—	1	—	12
25				2	—	—	—	1	1	—	—	7
26					—	4	1	2	—	—	—	8
27						13	14	10	4	1	—	72
28							5	3	10	1	2	55
29								4	7	1	—	46
30									2	—	—	41
31										—	—	5
34											—	4

HWE predictions, based on the observed allele frequencies. The overall heterozygosity at this locus is 0.809 ± 0.031, while its expectation based on the allele frequencies is 0.859 ± 0.007.

Table 3 summarizes each of the three test statistics mentioned above. Even though the statistics based on the total number of heterozygotes or homozygotes or the number of distinct genotypes of these two categories disregard each specific genotypic combinations, their use in testing departures from HWE predictions is not invalid when the number of alleles is large and the sample size is inadequate to observe

Table 3. Tests for Hardy-Weinberg Equilibrium (HWE) of genotype frequencies at the ApoB and ApoC-II VNTR Loci

Locus and statistics	Observed value	Expected ± s.e. (under HWE)	*P*
ApoB VNTR:			
Number of			
heterozygotes	181	184.62 ± 6.53	>0.55
homozygotes	59	55.38 ± 6.53	>0.55
Number of distinct			
heterozygote genotypes	34	32.98 ± 2.56	0.689
homozygote genotypes	8	6.37 ± 1.17	0.162
Likelihood ratio	60.18	—	0.621
ApoC-II-VNTR:			
Number of			
heterozygotes	131	139.19 ± 4.43	>0.55
homozygotes	31	22.81 ± 4.43	>0.55
Number of distinct			
heterozygote genotypes	35	34.89 ± 2.61	0.968
homozygote genotypes	7	6.06 ± 0.83	0.258
Likelihood ratio	65.38	—	0.234

all possible genotypes in a given sample. The inference regarding the fit of the data to HWE prediction is identical when these simple test statistics are contrasted with the more complex likelihood ratio test (G-statistics). The latter is presumably the most powerful statistical test because no data summarization is involved in evaluating this statistic nor in computing its empirical significance level. The first two summary statistics have the advantage in the sense that standard large sample theory can be invoked in judging whether or not they reflect a departure from HWE, whereas tedious permutation tests are needed to determine the significance level of the G-statistic.

Two-Locus Genotypic Distribution

As mentioned earlier, information on the joint distribution of genotypes at the ApoB and ApoC-II VNTR loci is available for 160 unrelated individuals in the present sample. A complete tabulation of this joint distribution was made to determine whether or not there is any evidence of non-random association of alleles at these loci. We expect no association, since these two loci are not syntenic, and hence they should segregate independently of each other. Evidence of non-random association, on the contrary, would signify that the population from which this sample is derived is heterogeneous, since it is known that a pseudo-linkage disequilibrium can be generated due to mixture of two or more populations (Nei and Li, 1973).

In principle, the two-locus genotype data is a contingency table of categorical data. But, due to the sparse nature of data, the traditional large sample contingency chi-square test cannot be applied since many of the classes are not represented in the sample. Among the 160 individuals for which two-locus genotypic information is available, there are 31 different genotypes at the ApoB VNTR locus and 41 different genotypes at the ApoC-II VNTR locus, giving a total of 1271 possible gentoypes that could have been observed. We observed only 132 different two-locus genotypes among the 160 individuals. The likelihood ratio test statistic, G, for the observed genotypic combinations is 227.86, which is not significant ($P = 0.428$), after 1000 random permutations. Two alternative statistics are also computed to illustrate that some particular summary of such data can be used to check independence of their segregations. Individuals may be classified into heterozygous or homozygous types at each locus to form a standard 2×2 contigency table (Tab. 4). The expected frequencies of each of the four classes can be obtained from the heterozygosity values of each locus, under the assumption of independent segregation. These are shown in the third column of Tab. 4. Clearly, the observed frequencies are in agreement with the expected ones (χ^2 with 1 d.f. is 0.34, $P > 0.55$),

Table 4. Tests for independence of genotypic frequencies at the ApoB and ApoC-II VNTR Loci

(A) Test based on two-locus homozygosity/heterozygosity:

		ApoC-II Locus	
ApoB Locus		Homozygous	Heterozygous
Homozygous	obs	9	32
	exp	7.52	31.81
Heterozygous	obs	22	97
	exp	23.09	97.58

χ^2 with 1 d.f. = 0.34 ($P > 0.55$)

(B) Test based on variance of number of heterozygous loci:

	Observed	Expected
Mean:	1.55 ± 0.05	1.56 ± 0.05
Variance:	0.36	0.34[+]

[+] 95% Confidence interval for variance is (0.27–0.41)

(C) Test based on likelihood ratio test criterion:

$-2 \ln(L_0/L_1) = 227.86$ (Empirical probability = 0.428)

are in agreement with the expected ones (χ^2 with 1 d.f. is 0.34, $P > 0.55$), suggesting again that there is no evidence of allelic association between these two unlinked loci.

Data from Tab. 4 can also be subjected to an alternative test, originally suggested by Brown *et al.* (1980) and examined in further detail by Chakraborty (1984). In this test, a variable representing the number of loci for which the individual is heterozygous is defined from the two-locus genotype data for each individual. In this specific case, this variable can take values 0, 1, or 2, corresponding to homozygosity at both loci, homozygosity for one locus and heterozygosity at the other, or heterozygosity at both loci, respectively. From the observed distribution of this statistic, we can determine the mean and variance of number of heterozygous loci. Brown *et al.* (1980) suggested that the variance of this statistic can be used to test whether or not the two loci are in linkage equilibrium. In the middle panel (B) of Tab. 4 we present the observed mean and variance of this statistic, and their expectations based on the hypothesis of linkage equilibrium. Also shown is the 95% confidence limit of the variance. Clearly, there is no departure from the expectation under the null hypothesis of linkage equilibrium. Therefore, we conclude that the ApoB and ApoC-II VNTR loci are at linkage equilibrium in this French Caucasian population. Intuitively, this result is expected because the two loci are unlinked. Indirectly, this also establishes that the population is homogeneous and there is no evidence

of internal substructuring large enough to produce departure from HWE or linkage equilibrium.

Allele Size Distributions and Relationship Between Heterozygosity and Number of Alleles

Having shown that the genotype distributions at the ApoB and ApoC-II VNTR loci conform to their Hardy-Weinberg equilibrium predictions and these two loci are at linkage equilibrium, some additional information regarding the production of new alleles may be extracted from the allele size distributions. Figures 1 and 2 show the size distributions of alleles for the ApoB and ApoC-II VNTR, respectively. Alleles are designated by the copy numbers of their respective core-sequences of length 14 or 15 bp for the ApoB locus and 2 for the ApoC-II VNTR locus. Both distributions show multiple modes; the ApoB distribution is bimodal, while there are apparently three modal classes for the Apoc-II locus.

Size distributions of alleles at these and other VNTR loci show multiple modes. In the literature there is no clear indication as to how such multiple modes can be generated. Boerwinkle *et al.* (1989) argued that the presence of multiple modes cannot be readily explained by 'mutational events' such as unequal recombination resulting from mismatching of repeat units or from replication slippage. Two possible alternative explanations may be offered. First, presence of multiple modes may indicate some form of genetic heterogeneity within the

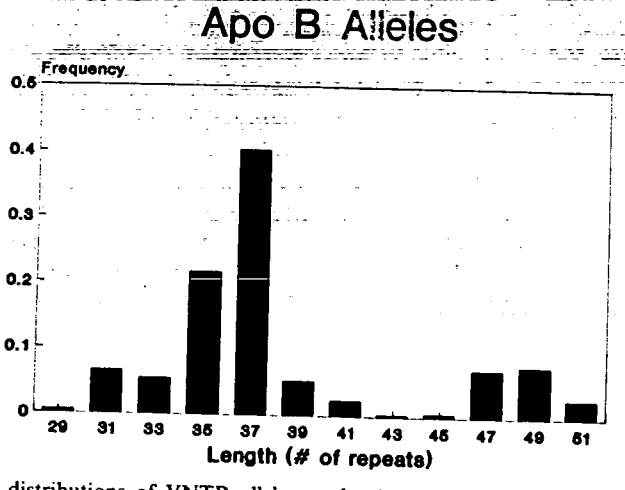


Figure 1. Size distributions of VNTR alleles at the ApoB Locus in the French Caucasian population (sizes are equivalent to the number of copies of a core sequence of length 14/15 bp)

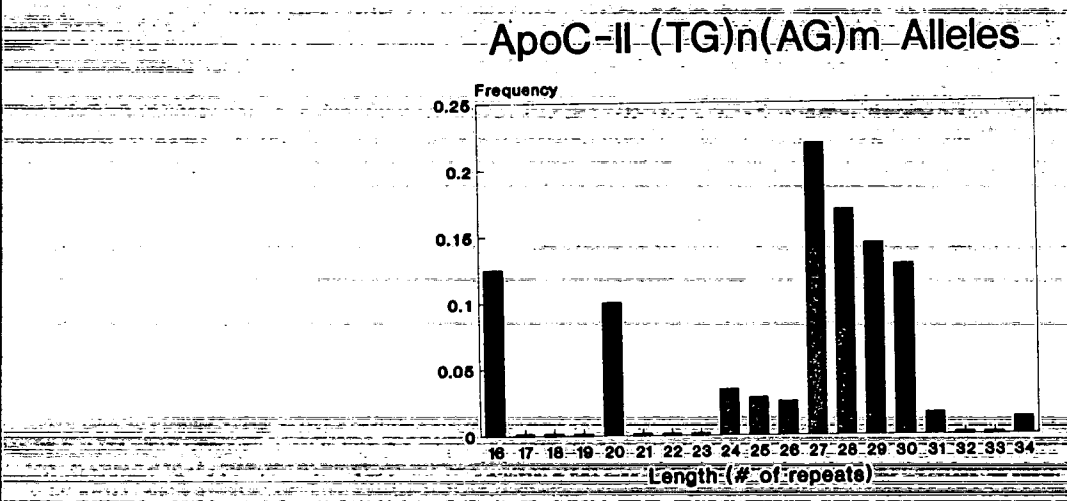


Figure 2. Size distributions of VNTR alleles at the ApoC-II Locus in the French Caucasian population (sizes are equivalent to the number of copies of a dinucleotide core sequence)

population. However, such heterogeneity would have produced significant departure from HWE predictions in the genotpye data analysis, and would have shown significant associations among alleles at the two loci. Since no such departure was found, we do not believe that population substructure is the cause of the observed multimodality. The second explanation is that the current distributions of alleles reflect their evolutionary antiquity, and therefore, it could be assumed that the modal classes reflect alleles that are older than the others. The suggestion of this possibility comes from the theory that under the infinite allele model, the age of an allele can be predicted from its frequency, in the sense that the most common allele is likely to be the oldest has a probability that equals its frequency (Watterson and Guess, 1977).

Such allele frequency profiles can be used to examine the relationship between heterozygosity and the observed number of alleles. In the context of electrophoretic loci, two mutation models have been proposed that can maintain genetic variation in a population. In one model, called the Infinite Allele Model (IAM), every mutational event is assumed to produce a new allele. When a population is at steady-state under the forces of such mutational events and random genetic drift, there is an expected relationship between heterozygosity and the observed number of alleles at a locus (Ewens, 1972; Chakraborty *et al.*, 1978; Chakraborty and Griffiths, 1982). Chakraborty and Weiss (1991) also showed that the sampling distribution of the observed number of alleles can be analytically evaluated. Table 5 shows the summary results of such computations for both loci.

The ApoB VNTR locus has an expected heterozygosity of 76.9%. Given this heterozygosity, we expect to find 14.13 ± 2.05 alleles in a

Table 5. Relationship between heterozygosity and number of alleles at the ApoB and ApoC-II VNTR Loci

		ApoB Locus	ApoC-II Locus
Heterozygosity	obs[a]	0.754 ± 0.028	0.819 ± 0.030
	exp[b]	0.769 ± 0.014	0.859 ± 0.007
Sample size (n)[c]		480	324
Number of alleles	obs	12	11
	exp[d] (IAM)	14.13 ± 2.05	21.53 ± 1.79
	exp[e] (SMM)	5.88	8.71

[a]The observed (obs) heterozygosity is from the actual genotype counts;
[b]The expected (exp) heterozygosity is based on the estimated allele frequencies;
[c]The sample size (n) refers to the number of genes sampled;
[d,e]The expected (exp) number of alleles under the Infinite Allele Model (IAM) and Stepwise Mutation Model (SMM) are based on the expected heterozygosity and sample size, n.

sample of 480 genes (240 individuals), whereas the observed number of alleles at this locus is 12. The prediction of the Infinite Allele Model (IAM) is in statistical agreement with the observation; the probability of observing 12 or less alleles is 0.783 and the probability of observing 12 or more alleles is 0.321.

The expected heterozygosity at the ApoC-II VNTR locus is 85.9%. Given this level of heterozygosity, we would have expected 24.86 ± 1.73 alleles to be observed in the sample of 324 genes (162 individuals). We actually observed only 11 alleles. Since the probability of observing 11 or less alleles in such a sample under the Infinite Allele Model is 0.003, we infer that there are too few alleles observed at this locus for the given heterozygosity. Two possible reasons could explain this discrepancy. First, since this VNTR locus has a dinucleotide core repeat unit, similar sized alleles migrate close to one another on a gel. When copy numbers are large, some rare alleles appearing in heterozygous state in combination with more common and similar size alleles may have been erroneously neglected. Such individuals may easily be scored homozygous for the common type allele. This can account for the observed deficiency in the number of alleles, without markedly reducing the heterozygosity level of the locus, since such unscored alleles are rare in the population. This possibility should have resulted in a heterozygote deficiency of our HWE test procedure as well. Although we did not detect any significant departure of genotype frequencies at this locus from the HWE predictions, there is a slight indication that the observed number of heterozygotes is somewhat lower (131 versus 139.19) than its expectation. The second reason could be that the Infinite Allele Model may not apply to such VNTR loci. When the core sequence is small, every 'mutational' event may not necessarily yield a new allele. A form of forward-backward mutation, called Step-Wise Mutation Model, may be more relevant in such a case. In the context of

electrophoretic studies, such a model has been proposed, where it is assumed that through a mutation the allelic state can either change by a single step in the forward or backward direction, or can keep the allelic state unaltered. Under such a model, Kimura and Ohta (1978) derived the relationship between number of alleles and heterozygosity. Applying their theory to the data on the ApoC-II VNTR locus, we found that for the given heterozygosity of 85.9%, we expect 8.71 alleles in a sample of 324 genes (162 individuals). The observed number of alleles, 11, is in between the expectations of the step-wise mutation model and the infinite allele model.

In summary, the relationship between heterozygosity and number of alleles at these two loci indicates that the genetic variation at such VNTR loci is maintained by joint effects of mutation and genetic drift, and the present population may be considered to be at a steady-state under these two counteracting forces.

Discussion and Conclusion

The above analyses of data on two VNTR loci performed on the same set of individuals from a genetically well-defined population showed that classic population genetic principles are applicable for understanding genetic characteristics of VNTR polymorphisms. The problems introduced by the large number of alleles can be circumvented by defining appropriate summary measures, such as the total number of heterozygotes, or the number of distinct genotypes observed in a sample. The sampling distributions of these summary measures are tractible and appropriate for hypothesis testing purposes. Alternatively, if one wishes to conduct genotype specific hypothesis testing, permutation tests can be performed on statistics relating expectations and observations of each specific genotype. Such permutation tests avoid problems inherent in sparse data (Efron, 1982). These alternative methods were shown here to result in identical conclusions.

Furthermore, our analyses also show that an apparent deficiency of observed heterozygosity should not be readily taken as evidence of substructuring within a population. This is so, because in the presence of substructuring we would have expected larger than expected number of alleles for the given value of heterozygosity (Chakraborty *et al.*, 1988; Chakraborty, 1990a, b). On the contrary, if incomplete resolution of alleles is responsible for an observed deficiency of heterozygosity, then it is generally accompanied with a smaller observed numbers of alleles.

Lastly, we note an important difference between the allele frequency distributions at the ApoB and ApoC-II VNTR loci. For the ApoB locus, the allele frequency distribution is in agreement with the predic-

tions of the infinite allele model, while this model does not apparently hold for the ApoC-II locus. The core sequence for the ApoB locus is substantially longer (14 or 15 bp) than that at the ApoC-II locus. When the core sequence is long, it may be true that replication slippage is relatively uncommon, while some form of unequal recombination or sister chromatid exchange may be the underlying mechanism of production of new alleles. In either of these two cases, recurrent mutations may not exactly revert allele sizes, because a fine tuning of such crossing-over events will be needed for generating an exact step-wise forward-backward form of mutation. Therefore, for VNTR loci characterized by relatively large core sequences, the infinite allele model may provide reasonable mathematical predictions of the allele frequency distribution, as in the case of the ApoB locus. On the other hand, when the core sequence is small, replication slippage can generate forward-backward mutations yielding several alleles of nearly similar sizes which can change from one to another. This process can occur in both a forward and backward fashion through recurrent mutational events. The large differences in some allele sizes at the ApoC-II locus may be produced by other mechanisms occurring at the same time. The observation that the observed number of alleles at the ApoC-II locus lies between the predictions of the Infinite Allele Model and the Step-wise Mutation Model indicates that at VNTR loci with a small repeat sequence, genetic variation may be generated by a mixture of two or more distinct molecular mechanisms. The first mechanism leads to new alleles not previously seen in a population and represents large differences of allele sizes, and the second produces small shifts of allele sizes in a forward-backward fashion. We speculate that the rate of occurrence of the first type of mutational changes is less than the second type. As a consequence of this, we observe a larger heterozygosity than expected at loci where step-wise changes are more common (reflected in larger heterozygosity at the ApoC-II locus compared to the ApoB locus).

A detailed mathematical study of such a mixed model of mutational changes is needed for a full understanding of the population dynamics of VNTR polymorphisms. Some initial attempt has been made in this direction. Li (1976) proposed a mixture model of mutation which incorporates the two types of mutations mentioned above. In his model, however, the step-wise changes were assumed to involve only one-step movements (forward or backward) in terms of allele states. Chakraborty and Nei (1982) proposed a step-mutation model where multiple step changes (in either direction) was introduced. Such models can be easily rationalized in the context of the molecular mechanisms of unequal recombination and replication slippage, and these should be examined in greater detail to study the evolutionary dynamics of VNTR polymorphism.

Acknowledgements
This work was supported by the grant GM-41399 from US National Institutes of Health and 90-IJ-CX-0038 from the National Institute of Justice. We thank Prof. A. J. Jeffreys for his constructive comments on the work and we are grateful to individuals from Nancy, France for their co-operation in our study.

References

Baird, M., Balazs, I., Giusti, A., Miyazaki, L., Wexler, K., Kanter, E., Glassberg, J., Rubinstin, P., and Sussman, L. (1986) Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity. *Am. J. Hum. Genet.* 39: 489–501.

Balazs, I., Baird, M., Clyne, M., and Meade, E. (1989) Human population genetic studies of five hypervariable loci. *Am. J. Hum. Genet.* 44: 182–190.

Bell, G. I., Selby, M. J., and Rutter, W. J. (1982) The highly polymorphic region near the insulin gene is composed of simple tandemly repeating sequences. *Nature* 295: 31–35.

Boerwinkle, E., Xiong, W., Fourest, E., and Chan, L. (1989) Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: Application to the apolipoprotein B 3′ hypervariable region. Proc. Natl. Acad. Sci. USA 86: 212–216.

Brown, A. H. D, Feldman, M. W., and Nevo, E. (1980) Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* 96: 523–536.

Capon, D. J., Chen, E. Y., Levinson, A. D., Seeburg, P. H., and Goeddel, D. V. (1983) Complete nucleotide sequence of the T24 human bladder carcinoma oncogene and its normal homologue. *Nature* 302: 33–37.

Chakraborty, R. (1984) Detection of nonrandom association of alleles from the distribution of the number of heterozygous loci in a sample. *Genetics* 108: 719–731.

Chakraborty, R. (1990a) Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. *Am. J. Hum. Genet.* 47: 87–94.

Chakraborty, R. (1990b) Genetic profile of cosmopolitan populations: Effects of hidden subdivision. *Anthrop. Anz.* 48: 313–331.

Chakraborty, R. (1991) Generalized occupancy problem and its application in population genetics. In: Sing, C. F., and Hanism, C. L. (eds), Impact of Genetic Variation on Individuals, Families and Populations. Oxford University Press, New York (in press).

Chakraborty, R., Fuerst, P. A., and Nei, M. (1978) Statistical studies on protein polymorphism in natural populations. II. Gene differentiation between populations. *Genetics* 88: 367–390.

Chakraborty, R., and Griffiths, R. C. (1982) Correlation of heterozygosity and number of alleles in different frequency classes. *Theor. Pop. Biol.* 21: 205–218.

Chakraborty, R., and Nei, M. (1982) Genetic differentiation of quantitative traits between populations or species. I. Mutation and random genetic drift. *Genet. Res.* 39: 303–314.

Chakraborty, R., Smouse, P. E., and Neel, J. V. (1988) Population amalgamation and genetic variation: Observations on artificially agglomerated tribal populations of Central and South America. *Am. J. Hum. Genet.* 43: 709–725.

Chakraborty, R., and Weiss, K. M. (1991) Genetic variation of the mitochondrial DNA genome in American Indians is at mutation-drift equilibrium. *Am. J. Phys. Anthrop.* (in press).

Cohen, J. E. (1990) DNA fingerprinting for forensic identification: Potential effects on data interpretation of subpopulation heterogeneity and band number variability. *Am. J. Hum. Genet.* 46: 358–368.

Collick, A., and Jeffreys, A.J. (1990) Detection of a novel minisatellite-specific DNA-binding protein. *Nucleic Acid Res.* 18: 625–629.

Clark, A. G. (1987) Neutrality tests of highly polymorphic restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 41: 948–956.

Craig, J., Fowler, S., Burgoyne, L. A., Scott, A. C., and Harding, H. W. J. (1988) Repetitive deoxyribonucleic acid (DNA) and human genome variation: A concise review relevant to forensic biology. *J. Forensic Sci.* 33: 1111–1126.

Devlin, B., Risch, N., and Roeder, K. (1990) No excess homozygosity at loci used for DNA fingerprinting, *Science* 249: 1416–1420.

Edwards, A., Hammond, H. A., Caskey, C. T., and Chakraborty, R. (1991) Population genetics of trimeric and tetrameric tandem repeats in four human ethnic groups. *Genomics* (in press).

Ewens, W. J. (1972) The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* 3: 87–112.

Efron, B. (1982) The Jackknife, the Bootstrap and Other Resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38. SIAM, Philadelphia.

Flint, J., Boyce, A.J., Martinson, J.J., and Clegg, J.B. (1989) Population bottlenecks in Polynesia revealed by minisatellites. *Hum. Genet.* 83: 257–263.

Fojo, S., Law, S., and Brewer, H. B. (1987) The human preapolipoprotein C-II gene complete nucleic acid sequence and genomic organization. *FEBS Letters* 213: 221–226.

Goodbourn, S. E. Y., Higgs, D. R., Clegg, J. B., and Weatherall, D. J. (1983) Molecular basis of length polymorphism in the human zeta-globin complex. Proc. Natl. Acad. Sci. USA 80: 5022–5026.

Huang, L. S., and Breslow, J. L. (1987) A unique AT-rich hypervariable minisatellite 3′ to the ApoB gene defines a high information restriction length polymorphism. *J. Biol. Chem.* 262: 8952–8955.

Jeffreys, A. J., Royle, V., Wilson, V., and Wong, Z. (1988) Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* 332: 278–281.

Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985) Hypervariable 'minisatellite' regions in human DNA. *Nature* 314: 67–73.

Kidd, K. K., Bowcock, A. M., Schmidtke, J., Track, R. K., Ricciuti, F., Hutchings, G., Bale, A., Perason, P., and Willard, H. F. (1989) Report of the DNA committee and catalogs of cloned and mapped genes and DNA polymorphisms. *Cytogenet. Cell Genet.* 51: 622–947.

Kimura, M., and Crow, J. F. (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49: 725–738.

Kimura, M., and Ohta, T. (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc. Natl. Acad. Sci. USA 75: 2868–2872.

Lander, E. S. (1989) DNA fingerprinting on trial. *Nature* 339: 501–505.

Li, W. H. (1976) A mixed model of mutation for electrophoretic identity of proteins within and between populations. *Genetics* 83: 423–432.

Ludwig, E. H., Friedl, W., and McCarthy, B. J. (1989) High-resolution analysis of a hypervariable region in the human apolipoprotein B gene. *Am. J. Hum. Genet.* 45: 458–464.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., and White, R. (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* 235: 1616–1622.

Nei, M., and Li, W. H. (1973) Linkage disequilibrium in subdivided populations. *Genetics* 75: 213–219.

Odelberg, S. J., Platke, R., Eldridge, J. R., Ballard, L., O'Connell, P., Nakamura, Y., Leppert, M., Lalouel, J. M., and White, R. (1989) Characterization of eight VNTR loci by agarose gel electrophoresis. *Genomics* 5: 915–924.

Sokal, R. R., and Rohlf, J. F. (1969) Biometry, 2nd edition. Freeman, New York.

Watterson, G.A., and Guess, H.A. (1977) Is the most frequent allele the oldest? *Theor. Pop. Biol.* 11: 141–160.

Wyman, A. R., and White, R. (1980) A highly polymorphic locus in human DNA. Proc. Natl Acad. Sci. USA 77: 6754–6758.

DNA Fingerprinting: Approaches and Applications
ed. by T. Burke, G. Dolf, A. J. Jeffreys & R. Wolff

Population Genetics of Hypervariable Loci: Analysis of PCR Based VNTR Polymorphism Within a Population

R. Chakraborty[a], M. Fornage[a,b], R. Gueguen[b], and E. Boerwinkle[a]

[a]*Genetics Centers, University of Texas Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, Texas 77225, USA;* [b]*Center for Preventive Medicine, Nancy, France*

Summary
Using a polymerase chain reaction (PCR) based method, genotypes at two hypervariable loci (3′ to the Apo-B-structural gene and at the ApoC-II gene) were determined by size classification of alleles. Genotype data at the Apo-B locus (Apo-B VNTR) were obtained on 240 French Caucasians; the sample size for the ApoC-II VNTR was 162. For 160 individuals two-locus genotype data were available. Applications of some recently developed statistical methods to these data indicate that both of these loci are at Hardy-Weinberg equilibrium (HWE) and there is no indication of allelic associations between these two unlinked loci. In addition, the observed numbers of alleles (12 for the Apo-B and 11 for the ApoC-II VNTR loci) are also consistent with their respective expectations based on the observed heterozygosities (76.9% for the Apo-B and 85.9% for the ApoC-II loci) suggesting genetic homogeneity of this population-based sample. The multimodal distribution of allele sizes observed for both loci indicate that the production of new alleles at such VNTR loci may be caused by more than one molecular mechanism. The utility of such highly polymorphic loci for human genetic research and forensic applications are discussed in the context of these findings.

Introduction

A large number of DNA segments in the human genome contain a variable number of short tandemly repeated sequences. To varying degree, the core repeat unit is conserved from one repeat to another. The core sequence may also vary among such loci from 2 bases to several kilobases. The copy number of such core sequences reveals genetic variation several orders larger than that detected by classical serologic and biochemical genetic markers. Since the first demonstration of such a highly polymorphic, locus-specific sequence in the human genome by Wyman and White (1980), numerous hypervariable regions have been described that flank structural loci in the human genome (e.g., Bell *et al.*, 1982; Capon *et al.*, 1983; Goodbourn *et al.*, 1983; Jeffreys *et al.*, 1985; Boerwinkle *et al.*, 1989; Ludwig *et al.*, 1989; reviewed also by Jeffreys and Wolff, this volume). It is now well-recognized that the human genome contains a large number of these polymorphic segments, numbering possibly in the thousands. Several acronyms of such polymorphic systems are proposed. Jeffreys *et al.*

(1985) suggested locus-specific 'minisatellites'; Nakamura *et al.* (1987) coined the term 'VNTR' (Variable Number of Tandem Repeats), while several others (e.g., Balazs *et al.*, 1989; Ludwig *et al.*, 1989) used the terminology 'HVR' (Hypervariable Region) to describe this type of genetic variation. When the core unit is small (di-, tri, or tetra-nucleotide), Edwards *et al.* (1991) called them 'STR' loci (Short Tandemly Repeated loci).

Since the allelic designation at these VNTR loci can be conveniently defined by the number of repeat units of the core sequence and each locus conforms to simple codominant Mendelian mode of inheritance, such loci are extremely useful for human genetic research. Recent tabulation of genetic markers used for human gene mapping indicates that collectively nearly 50 per cent of all genetic markers belong to this category (Kidd *et al.*, 1989). The increased efficiency of VNTR loci, compared to classical biochemical and RFLP markers, arises from the fact that the number of different alleles found at any VNTR locus is generally much larger. In addition, these VNTR loci have a high heterozygosity (sometimes approaching levels as high as 95–99 percent). As a consequence, detection of recombination between a VNTR locus and a disease gene or other genetic markers is simple because both parents can be heterozygous and provide four distinct alleles at such loci far more commonly than other classical markers.

The presence of large numbers of alleles at VNTR loci also makes them useful in the context of paternity testing and forensic medicine. A growing body of literature suggests that their utility is not only limited to academic circles and biomedical research; criminal justice and social welfare agencies can also benefit immensely from the application of such loci (Craig *et al.* 1988).

These advantages notwithstanding, concerns have been raised with regard to the population genetic characteristics of such polymorphisms (Lander, 1989; Cohen, 1990). In part this is caused by limited population data of VNTR polymorphisms from genetically well-characterized populations. Five studies in this regard are noteworthy. Baird *et al.* (1986) have shown extensive variation at two VNTR loci, HRAS-1 and D14S1, in three major ethnic groups. Using such data, Clark (1987) postulated that allelic variation at these loci follow the mutation-drift model of genetic variation (Kimura and Crow, 1964). Jeffreys *et al.* (1988) entertained other models of allele frequency distribution such as a finite allele mutation model or a stepwise mutation model to explain the relationship between heterozygosity and mutation rate at such loci. Flint *et al.* (1989) showed that the differences of heterozygosity levels of specific VNTR loci across populations can be used to postulate whether or not events such as a population bottleneck could have occurred during the geographic dispersal of humans. Chakraborty (1990a, b)

argued that even a single VNTR locus can provide information concerning substructuring within a population with a statistical power far greater than several classical genetic markers studied simultaneously.

The well known advantages of VNTR loci such as their high heterozygosity and a large number of alleles also poses problems in the statistical interpretation of population data. For example, the presence of a large number of alleles increases sample size requirements for population survey studies, since even with respectable sample sizes all possible genotypes can not be generally detected. As a result, precisions of allele frequency estimates are compromised which subsequently hinders statistical calculations based on classical text book methods of analysis. Since all genotypes are not observed, and even the ones observed occur in low frequencies, the application of the large sample theory of chi-square goodness-of-fit test is not valid for checking whether or not the observed genotypic distribution conforms to their Hardy-Weinberg expectations. Similarly, the association among alleles at multiple loci cannot be adequately determined from their genotypic distributions by standard summary measures such as linkage disequilibrium. In addition, a single VNTR locus may exhibit more than one allele of similar size and incomplete resolution of distinct alleles by Southern gel electrophoresis is not uncommon. Such inescapable laboratory phenomena can not be overlooked in the statistical interpretation of VNTR population data (Devlin *et al.*, 1990). There are also claims that the hypervariability at these loci is caused by 'mutational' changes whose rate is high in comparison with other classical markers. Hence, when used singly, a VNTR locus may lead to a wrong conclusion regarding biological relationships between individuals (Odelberg *et al.*, 1989). Furthermore, very little is known about the molecular mechanism of production of new alleles at VNTR loci, although it has been suggested that replication slippage, sister chromatid exchange, and/or unequal recombination may be involved in this process. Virtually nothing is known regarding the functional requirements of such loci in general, even though DNA-binding proteins which appear to bind specifically to VNTR loci are known to exist (Collick and Jeffreys, 1990). In view of these uncertainties, it is difficult to determine the mode and rate of evolution of genetic variation at VNTR loci.

In spite of these difficulties, we believe that the statistical interpretation of VNTR polymorphism data is not an insurmountable problem. Certain modifications of classic population genetic methods of data analysis can be introduced which lead to rigorous and legitimate estimation and hypothesis-testing principles for analyzing such data. Such methods also may suggest molecular mechanisms that generate and maintain genetic variation at these loci. We have initiated a series of studies on the population genetics of VNTR polymorphism at our Center, and this presentation is a preliminary summary of several results from

these studies. Two VNTR loci have been typed by PCR-based experimental protocols which can be employed for population surveys at such loci and which minimize the problem of incomplete resolution of alleles (Boerwinkle *et al.*, 1989). Here we describe the genotype and allele frequency distributions at two VNTR loci; one locus is 3′ to the apolipoprotein-B (ApoB) gene and the other is within the ApoC-II gene, which have been scored on individuals belonging to 121 nuclear families. Several alternative methods are suggested for examining the conformity of the genotypic distribution of VNTR data with HWE and for testing independent segregation of alleles at unlinked loci in the presence of large number of alleles. Futhermore, the allele frequency distributions of these loci are used to predict possible molecular mechanisms that generate and maintain such genetic variation. Characterization of such population genetic features implies that VNTR polymorphisms detected through PCR-based studies conform to classic population genetic principles, and hence are useful in human genetic research and forensic applications.

Materials and Methods

The genotype data analyzed here were obtained from a random sample of 121 nuclear families taking part in routine health examinations at the Center for Preventive Medicine in Nancy, France. Genomic DNA was isolated by phenol/chloroform extraction of proteinase K treated crude buffy coat preparations. Two VNTR loci with different core sequences were selected for the present analyses. The first is an AT-rich VNTR 3′ to the human apolipoprotein B gene on chromosome 2 (Huang and Breslow, 1987). Detailed PCR-based methods for typing this VNTR have been previously presented (Boerwinkle *et al.*, 1989). Our previous results indicate that this locus differs in the number of copies (from 29 to 51) of a conserved core sequence 14 or 15 base pair long (Boerwinkle *et al.*, 1989). The second VNTR locus is a microsatellite located in the first intron of the human apolipoprotein C-II gene on chromosome 19 (Fojo *et al.*, 1987). This ApoC-II VNTR consists of a $(TG)_n(AG)_m$ core motif repeated from 16 to 34 times. The oligonucleotides used for priming the PCR bind immediately adjacent to the $(TG)_n(AG)_m$ block. One member of the pair of primers was 5′ end-labeled using bacteriophage T4 kinase. The PCR was carried out in a 50 μl volume containing approximately 0.5 μg of genomic DNA and samples were processed through 30 temperature cycles consisting of 1 minute at 92°C (denaturation), 1 minute at 55°C (annealing), and 1.5 minutes at 72°C (elongation). After 10 cycles with only cold oligonucleotide the reaction mixture was spiked with the end-labeled oligonucleotide for the remaining 20 cycles. The amplified DNA was analyzed after being electrophoresed on 8% denaturing polyacrylamide gels and

exposed to X-ray film for 20 hours. Size standards were created by dideoxy sequencing using M13 mp18 as a template. The size of the PCR products were directly determined from the size of the co-migrating M13 fragment in the ladder (data not shown).

The family data were used to verify Mendelian segregation of the identified alleles at the ApoB and ApoC-II VNTR loci. The parents of these families represent a sample of unrelated individuals and were used for allele frequency estimation and other calculations. Therefore, these data can be regarded as from a random sample of Caucasian individuals of French ancestry. Genotype data on 240 individuals for the ApoB locus and 162 individuals for the ApoC-II locus were used for the analyses presented here. Two locus genotype data were available for 160 individuals.

Because one purpose of this paper is to describe the analytical tools for the analysis of allele and genotype frequency data at VNTR loci, the statistical methods are not presented in this section but rather are given along with a corresponding question and resulting inference in the next section. Statistical analyses consist of: (1) analysis of genotype and allele frequency distribution for each locus individually to test whether or not HWE predictions hold for these two loci, (2) joint analysis of two-locus genotype data to determine independent segregation of alleles at these two unlinked loci; (3) examination of the relationship between heterozygosity and the number of alleles at each locus to determine the underlying mechanism of production of new alleles at these loci, and finally (4) to postulate possible reasons for the shape of the allele size distributions at these loci.

Results

Single-Locus Genotypic Distributions

Typically VNTR locus variation is codominant and multi-allelic. Letting k be the number of segregating alleles, there are $k(k+1)/2$ possible genotypes at a locus, k of which are homozygous A_iA_i $(i = 1, 2, \ldots, k)$ and $k(k-1)/2$ are heterozygous A_iA_j $(i < j = 2, 3, \ldots, k)$. It should be possible to observe all possible $k(k+1)/2$ genotypes in any given sample. However, the sample size needed to observe all possible genotypes is generally large when k is large.

We observed 12 different alleles at the ApoB VNTR locus and 11 at the ApoC-II VNTR locus in a sample of 240 and 162 individuals, respectively. The number of possible genotypes $(k(k+1)/2)$, therefore, are 78 and 66, respectively. Table 1 shows the observed genotype and allele frequency distributions at the ApoB locus, demonstrating that even though the sample size $(n = 240)$ is much larger than the number

of possible genotypes, we observed only 42 distinct genotypes. Of the possible 12 homozygote genotypes only 8 are observed and of the possible 66 heterozygote genotypes only 34 are observed. A large fraction of the observed genotypes have frequencies below 5 individuals per genotype.

Obviously, such data cannot be subjected to a classical goodness-of-fit test examining whether or not the genotypic distribution conforms to HWE predictions. This is so because a chi-square approximation is inaccurate when the observed frequencies are small, and also we cannot assign any well-defined degrees of freedom because of the absence of a large fraction of genotypes in the sample. Three alternative test statistics are used for this purpose. First, grouping the entire data of the genotype table (Tab. 1) into two classes, heterozygotes and homozygotes (irrespective of the allele types observed), we found that there are 59 individuals that are homozygous at this locus and 181 heterozygous. The expected numbers of individuals in these two categories, based on the observed allele frequencies (shown on the last column of Tab. 1), are 55.38 and 184.62, respectively, and their difference from the observed ones is not significant by a chi-square test with 1 d.f. ($P > 0.55$). Second, using the theory of Chakraborty *et al.* (1988) and Chakraborty (1991), we also determined the expectations (and their standard errors) of the number of distinct heterozygous and homozygous genotypes that can be observed in a sample of 240 individuals at the ApoB VNTR locus when the alleles unite at random to form the genotypes of these individuals. The sampling distributions of these statistics, under the HWE assumption, can be evaluated exactly (Chakraborty, 1991). In particular, denoting the expected genotype frequencies (under HWE) by $Q_1, Q_1, \ldots, Q_K$, the mean and variance of the number of distinct

Table 1. Genotype and allele frequency distributions at the ApoB VNTR Locus in a random sample of 240 individuals from Nancy, France

	Alleles												
	29	31	33	35	37	39	41	43	45	47	49	51	Freq.
29	—	1	1	—	—	—	—	—	—	—	—	—	2
31		3	2	7	6	1	2	1	—	2	2	1	31
33			2	2	10	2	—	—	—	1	3	1	26
35				12	50	1	2	—	—	10	6	2	104
37					36	11	4	—	2	16	17	6	194
39						2	1	—	—	—	3	2	25
41							1	—	—	—	—	—	11
43								—	—	—	—	—	1
45									—	—	—	—	2
47										1	2	1	34
49											2	—	37
51												—	13

genotypes observed in a sample of n individuals are given by

$$\mu = K - T_1, \quad \text{and} \quad \sigma^2 = T_1(1 - T_1) + 2T_2, \tag{1a}$$

respectively, where

$$T_1 = \sum_{i=1}^{K} (1 - Q_i)^n, \quad \text{and} \quad T_2 = \sum_{i>j=1}^{K} (1 - Q_i - Q_j)^n. \tag{1b}$$

Note that in equations (1) and (2) $K = k$ for the homozygous and $Q_i = p_i^2$, the square of the i-th allele frequency, and $K = k(k-1)/2$ for the heterozygous genotypes, with Q_i's being an array of $k(k-1)/2$ probabilities representing each heterozygote genotypes ($2p_ip_j$). Application of this method to the present data showed that the observed numbers of distinct homozygous and heterozygous genotypes for the ApoB locus, 8 and 34, are not significantly different from their expectations, 6.37 and 32.98. Therefore, we conclude that the observed numbers of distinct genotypes are also in concordance with HWE predictions.

Since all of these statistics disregard the specific allele types observed in the sample, and hence these tests do not detect deviations of each specific genotype frequency from its HWE prediction, we used a third test criterion which does not have this limitation. This is the G-statistic (Sokal and Rohlf, 1969), a likelihood ratio, which should not be significant if the HWE prediction is correct. Unfortunately, although the G-statistic is a contrast of every observed genotype frequencies with their respective expectations, no standard statistical distribution can be applied to determine the significance level of the G-statistic because of the absence (or small frequencies) of several genotypes. We employed a shuffling algorithm to determine the empirical distribution of the G-statistic, by randomly permuting the 480 allele labels (12 of them, since there are 12 different observed alleles) and reconstructing genotypes by pairing the shuffled alleles at random. The observed value of G in the given data (of Tab. 1) was 60.18, and its empirical probability level was 0.62, suggesting that by chance 62% times we could have observed G-values larger than the one observed under the assumption of HWE. Therefore, none of the three alternative tests offered any suggestion of non-random association of alleles at the ApoB VNTR locus. In conclusion, the genotype distribution at this locus among the French Caucasians can be assumed to satisfy the HWE predictions. The observed heterozygosity at this locus is 0.745 ± 0.028, and its expectation based on the estimated allele frequencies is 0.769 ± 0.014.

Table 2 shows the observed genotypic and allele frequency distributions at the ApoC-II VNTR locus. Eleven segregating alleles are found at this locus, but of the possible 66 genotypes only 42 are observed. As in the case of the ApoB VNTR, each of the three alternative test criteria shows that the observed genotypic distribution is in accordance with the

Table 2. Genotype and allele frequency distributions at the ApoC-II VNTR Locus in a random sample of 162 individuals from Nancy, France

	Alleles											
	16	20	24	25	26	27	28	29	30	31	34	Freq.
16	3	3	3	1	1	6	6	5	7	1	1	40
20		2	1	—	—	6	7	4	8	—	1	34
24			—	—	—	1	1	5	—	1	—	12
25				2	—	—	—	1	1	—	—	7
26					—	4	1	2	—	—	—	8
27						13	14	10	4	1	—	72
28							5	3	10	1	2	55
29								4	7	1	—	46
30									2	—	—	41
31										—	—	5
34											—	4

HWE predictions, based on the observed allele frequencies. The overall heterozygosity at this locus is 0.809 ± 0.031, while its expectation based on the allele frequencies is 0.859 ± 0.007.

Table 3 summarizes each of the three test statistics mentioned above. Even though the statistics based on the total number of heterozygotes or homozygotes or the number of distinct genotypes of these two categories disregard each specific genotypic combinations, their use in testing departures from HWE predictions is not invalid when the number of alleles is large and the sample size is inadequate to observe

Table 3. Tests for Hardy-Weinberg Equilibrium (HWE) of genotype frequencies at the ApoB and ApoC-II VNTR Loci

Locus and statistics	Observed value	Expected $\pm$s.e. (under HWE)	P
ApoB VNTR:			
Number of heterozygotes	181	184.62 ± 6.53	>0.55
homozygotes	59	55.38 ± 6.53	>0.55
Number of distinct heterozygote genotypes	34	32.98 ± 2.56	0.689
homozygote genotypes	8	6.37 ± 1.17	0.162
Likelihood ratio	60.18	—	0.621
ApoC-II VNTR:			
Number of heterozygotes	131	139.19 ± 4.43	>0.55
homozygotes	31	22.81 ± 4.43	>0.55
Number of distinct heterozygote genotypes	35	34.89 ± 2.61	0.968
homozygote genotypes	7	6.06 ± 0.83	0.258
Likelihood ratio	65.38	—	0.234

all possible genotypes in a given sample. The inference regarding the fit of the data to HWE prediction is identical when these simple test statistics are contrasted with the more complex likelihood ratio test (G-statistics). The latter is presumably the most powerful statistical test because no data summarization is involved in evaluating this statistic nor in computing its empirical significance level. The first two summary statistics have the advantage in the sense that standard large sample theory can be invoked in judging whether or not they reflect a departure from HWE, whereas tedious permutation tests are needed to determine the significance level of the G-statistic.

Two-Locus Genotypic Distribution

As mentioned earlier, information on the joint distribution of genotypes at the ApoB and ApoC-II VNTR loci is available for 160 unrelated individuals in the present sample. A complete tabulation of this joint distribution was made to determine whether or not there is any evidence of non-random association of alleles at these loci. We expect no association, since these two loci are not syntenic, and hence they should segregate independently of each other. Evidence of non-random association, on the contrary, would signify that the population from which this sample is derived is heterogeneous, since it is known that a pseudo-linkage disequilibrium can be generated due to mixture of two or more populations (Nei and Li, 1973).

In principle, the two-locus genotype data is a contingency table of categorical data. But, due to the sparse nature of data, the traditional large sample contingency chi-square test cannot be applied since many of the classes are not represented in the sample. Among the 160 individuals for which two-locus genotypic information is available, there are 31 different genotypes at the ApoB VNTR locus and 41 different genotypes at the ApoC-II VNTR locus, giving a total of 1271 possible gentoypes that could have been observed. We observed only 132 different two-locus genotypes among the 160 individuals. The likelihood ratio test statistic, G, for the observed genotypic combinations is 227.86, which is not significant ($P = 0.428$), after 1000 random permutations. Two alternative statistics are also computed to illustrate that some particular summary of such data can be used to check independence of their segregations. Individuals may be classified into heterozygous or homozygous types at each locus to form a standard 2×2 contigency table (Tab. 4). The expected frequencies of each of the four classes can be obtained from the heterozygosity values of each locus, under the assumption of independent segregation. These are shown in the third column of Tab. 4. Clearly, the observed frequencies are in agreement with the expected ones (χ^2 with 1 d.f. is 0.34; $P > 0.55$),

Table 4. Tests for independence of genotypic frequencies at the ApoB and ApoC-II VNTR Loci

(A) Test based on two-locus homozygosity/heterozygosity:

		ApoC-II Locus	
ApoB Locus		Homozygous	Heterozygous
Homozygous	obs	9	32
	exp	7.52	31.81
Heterozygous	obs	22	97
	exp	23.09	97.58

χ^2 with 1 d.f. = 0.34 ($P > 0.55$)

(B) Test based on variance of number of heterozygous loci:

	Observed	Expected
Mean:	1.55 ± 0.05	1.56 ± 0.05
Variance:	0.36	0.34[+]

[+] 95% Confidence interval for variance is (0.27–0.41)

(C) Test based on likelihood ratio test criterion:

$-2 \ln(L_0/L_1) = 227.86$ (Empirical probability = 0.428)

are in agreement with the expected ones (χ^2 with 1 d.f. is 0.34; $P > 0.55$), suggesting again that there is no evidence of allelic association between these two unlinked loci.

Data from Tab. 4 can also be subjected to an alternative test, originally suggested by Brown *et al.* (1980) and examined in further detail by Chakraborty (1984). In this test, a variable representing the number of loci for which the individual is heterozygous is defined from the two-locus genotype data for each individual. In this specific case, this variable can take values 0, 1, or 2, corresponding to homozygosity at both loci, homozygosity for one locus and heterozygosity at the other, or heterozygosity at both loci, respectively. From the observed distribution of this statistic, we can determine the mean and variance of number of heterozygous loci. Brown *et al.* (1980) suggested that the variance of this statistic can be used to test whether or not the two loci are in linkage equilibrium. In the middle panel (B) of Tab. 4 we present the observed mean and variance of this statistic, and their expectations based on the hypothesis of linkage equilibrium. Also shown is the 95% confidence limit of the variance. Clearly, there is no departure from the expectation under the null hypothesis of linkage equilibrium. Therefore, we conclude that the ApoB and ApoC-II VNTR loci are at linkage equilibrium in this French Caucasian population. Intuitively, this result is expected because the two loci are unlinked. Indirectly, this also establishes that the population is homogeneous and there is no evidence

of internal substructuring large enough to produce departure from HWE or linkage equilibrium.

Allele Size Distributions and Relationship Between Heterozygosity and Number of Alleles

Having shown that the genotype distributions at the ApoB and ApoC-II VNTR loci conform to their Hardy-Weinberg equilibrium predictions and these two loci are at linkage equilibrium, some additional information regarding the production of new alleles may be extracted from the allele size distributions. Figures 1 and 2 show the size distributions of alleles for the ApoB and ApoC-II VNTR, respectively. Alleles are designated by the copy numbers of their respective core sequences of length 14 or 15 bp for the ApoB locus and 2 for the ApoC-II VNTR locus. Both distributions show multiple modes; the ApoB distribution is bimodal, while there are apparently three modal classes for the Apoc-II locus.

Size distributions of alleles at these and other VNTR loci show multiple modes. In the literature there is no clear indication as to how such multiple modes can be generated. Boerwinkle *et al.* (1989) argued that the presence of multiple modes cannot be readily explained by 'mutational events' such as unequal recombination resulting from mismatching of repeat units or from replication slippage. Two possible alternative explanations may be offered. First, presence of multiple modes may indicate some form of genetic heterogeneity within the

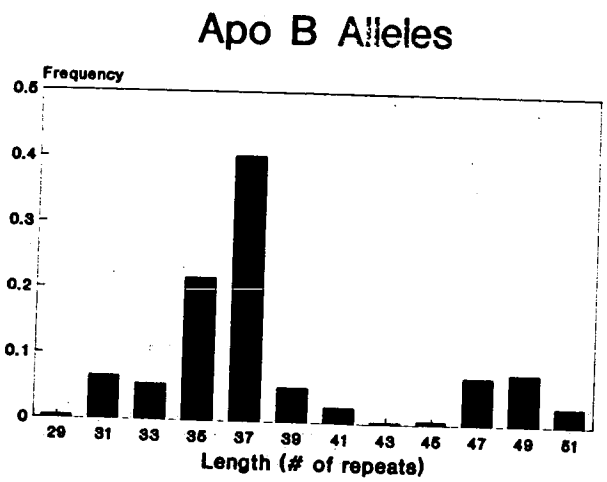


Figure 1. Size distributions of VNTR alleles at the ApoB Locus in the French Caucasian population (sizes are equivalent to the number of copies of a core sequence of length 14/15 bp)

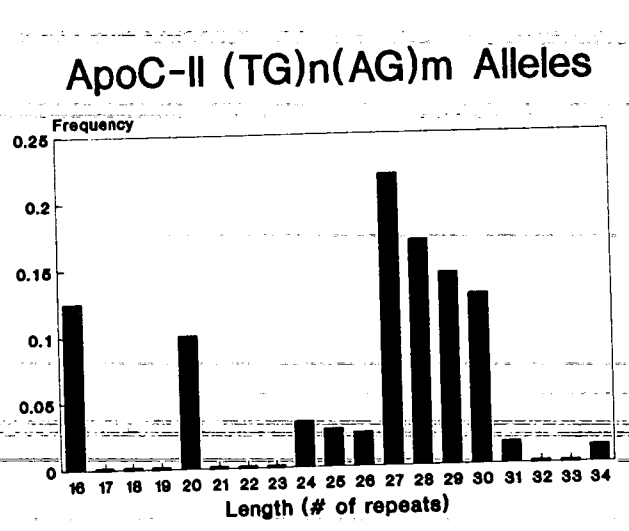


Figure 2. Size distributions of VNTR alleles at the ApoC-II Locus in the French Caucasian population (sizes are equivalent to the number of copies of a dinucleotide core sequence)

population. However, such heterogeneity would have produced significant departure from HWE predictions in the genotpye data analysis, and would have shown significant associations among alleles at the two loci. Since no such departure was found, we do not believe that population substructure is the cause of the observed multimodality. The second explanation is that the current distributions of alleles reflect their evolutionary antiquity, and therefore, it could be assumed that the modal classes reflect alleles that are older than the others. The suggestion of this possibility comes from the theory that under the infinite allele model, the age of an allele can be predicted from its frequency, in the sense that the most common allele is likely to be the oldest has a probability that equals its frequency (Watterson and Guess, 1977).

Such allele frequency profiles can be used to examine the relationship between heterozygosity and the observed number of alleles. In the context of electrophoretic loci, two mutation models have been proposed that can maintain genetic variation in a population. In one model, called the Infinite Allele Model (IAM), every mutational event is assumed to produce a new allele. When a population is at steady-state under the forces of such mutational events and random genetic drift, there is an expected relationship between heterozygosity and the observed number of alleles at a locus (Ewens, 1972; Chakraborty *et al.*, 1978; Chakraborty and Griffiths, 1982). Chakraborty and Weiss (1991) also showed that the sampling distribution of the observed number of alleles can be analytically evaluated. Table 5 shows the summary results of such computations for both loci.

The ApoB VNTR locus has an expected heterozygosity of 76.9%. Given this heterozygosity, we expect to find 14.13 ± 2.05 alleles in a

Table 5. Relationship between heterozygosity and number of alleles at the ApoB and ApoC-II VNTR Loci

		ApoB Locus	ApoC-II Locus
Heterozygosity	obs[a]	0.754 ± 0.028	0.819 ± 0.030
	exp[b]	0.769 ± 0.014	0.859 ± 0.007
Sample size (n)[c]		480	324
Number of alleles	obs	12	11
	exp[d] (IAM)	14.13 ± 2.05	21.53 ± 1.79
	exp[e] (SMM)	5.88	8.71

[a]The observed (obs) heterozygosity is from the actual genotype counts;
[b]The expected (exp) heterozygosity is based on the estimated allele frequencies;
[c]The sample size (n) refers to the number of genes sampled;
[d,e]The expected (exp) number of alleles under the Infinite Allele Model (IAM) and Stepwise Mutation Model (SMM) are based on the expected heterozygosity and sample size, n.

sample of 480 genes (240 individuals), whereas the observed number of alleles at this locus is 12. The prediction of the Infinite Allele Model (IAM) is in statistical agreement with the observation; the probability of observing 12 or less alleles is 0.783 and the probability of observing 12 or more alleles is 0.321.

The expected heterozygosity at the ApoC-II VNTR locus is 85.9%. Given this level of heterozygosity, we would have expected 24.86 ± 1.73 alleles to be observed in the sample of 324 genes (162 individuals). We actually observed only 11 alleles. Since the probability of observing 11 or less alleles in such a sample under the Infinite Allele Model is 0.003, we infer that there are too few alleles observed at this locus for the given heterozygosity. Two possible reasons could explain this discrepancy. First, since this VNTR locus has a dinucleotide core repeat unit, similar sized alleles migrate close to one another on a gel. When copy numbers are large, some rare alleles appearing in heterozygous state in combination with more common and similar size alleles may have been erroneously neglected. Such individuals may easily be scored homozygous for the common type allele. This can account for the observed deficiency in the number of alleles, without markedly reducing the heterozygosity level of the locus, since such unscored alleles are rare in the population. This possibility should have resulted in a heterozygote deficiency of our HWE test procedure as well. Although we did not detect any significant departure of genotype frequencies at this locus from the HWE predictions, there is a slight indication that the observed number of heterozygotes is somewhat lower (131 versus 139.19) than its expectation. The second reason could be that the Infinite Allele Model may not apply to such VNTR loci. When the core sequence is small, every 'mutational' event may not necessarily yield a new allele. A form of forward-backward mutation, called Step-Wise Mutation Model, may be more relevant in such a case. In the context of

electrophoretic studies, such a model has been proposed, where it is assumed that through a mutation the allelic state can either change by a single step in the forward or backward direction, or can keep the allelic state unaltered. Under such a model, Kimura and Ohta (1978) derived the relationship between number of alleles and heterozygosity. Applying their theory to the data on the ApoC-II VNTR locus, we found that for the given heterozygosity of 85.9%, we expect 8.71 alleles in a sample of 324 genes (162 individuals). The observed number of alleles, 11, is in between the expectations of the step-wise mutation model and the infinite allele model.

In summary, the relationship between heterozygosity and number of alleles at these two loci indicates that the genetic variation at such VNTR loci is maintained by joint effects of mutation and genetic drift, and the present population may be considered to be at a steady-state under these two counteracting forces.

Discussion and Conclusion

The above analyses of data on two VNTR loci performed on the same set of individuals from a genetically well-defined population showed that classic population genetic principles are applicable for understanding genetic characteristics of VNTR polymorphisms. The problems introduced by the large number of alleles can be circumvented by defining appropriate summary measures, such as the total number of heterozygotes, or the number of distinct genotypes observed in a sample. The sampling distributions of these summary measures are tractible and appropriate for hypothesis testing purposes. Alternatively, if one wishes to conduct genotype specific hypothesis testing, permutation tests can be performed on statistics relating expectations and observations of each specific genotype. Such permutation tests avoid problems inherent in sparse data (Efron, 1982). These alternative methods were shown here to result in identical conclusions.

Furthermore, our analyses also show that an apparent deficiency of observed heterozygosity should not be readily taken as evidence of substructuring within a population. This is so, because in the presence of substructuring we would have expected larger than expected number of alleles for the given value of heterozygosity (Chakraborty *et al.*, 1988; Chakraborty, 1990a, b). On the contrary, if incomplete resolution of alleles is responsible for an observed deficiency of heterozygosity, then it is generally accompanied with a smaller observed numbers of alleles.

Lastly, we note an important difference between the allele frequency distributions at the ApoB and ApoC-II VNTR loci. For the ApoB locus, the allele frequency distribution is in agreement with the predic-

tions of the infinite allele model, while this model does not apparently hold for the ApoC-II locus. The core sequence for the ApoB locus is substantially longer (14 or 15 bp) than that at the ApoC-II locus. When the core sequence is long, it may be true that replication slippage is relatively uncommon, while some form of unequal recombination or sister chromatid exchange may be the underlying mechanism of production of new alleles. In either of these two cases, recurrent mutations may not exactly revert allele sizes, because a fine tuning of such crossing-over events will be needed for generating an exact step-wise forward-backward form of mutation. Therefore, for VNTR loci characterized by relatively large core sequences, the infinite allele model may provide reasonable mathematical predictions of the allele frequency distribution, as in the case of the ApoB locus. On the other hand, when the core sequence is small, replication slippage can generate forward-backward mutations yielding several alleles of nearly similar sizes which can change from one to another. This process can occur in both a forward and backward fashion through recurrent mutational events. The large differences in some allele sizes at the ApoC-II locus may be produced by other mechanisms occurring at the same time. The observation that the observed number of alleles at the ApoC-II locus lies between the predictions of the Infinite Allele Model and the Step-wise Mutation Model indicates that at VNTR loci with a small repeat sequence, genetic variation may be generated by a mixture of two or more distinct molecular mechanisms. The first mechanism leads to new alleles not previously seen in a population and represents large differences of allele sizes, and the second produces small shifts of allele sizes in a forward-backward fashion. We speculate that the rate of occurrence of the first type of mutational changes is less than the second type. As a consequence of this, we observe a larger heterozygosity than expected at loci where step-wise changes are more common (reflected in larger heterozygosity at the ApoC-II locus compared to the ApoB locus).

A detailed mathematical study of such a mixed model of mutational changes is needed for a full understanding of the population dynamics of VNTR polymorphisms. Some initial attempt has been made in this direction. Li (1976) proposed a mixture model of mutation which incorporates the two types of mutations mentioned above. In his model, however, the step-wise changes were assumed to involve only one-step movements (forward or backward) in terms of allele states. Chakraborty and Nei (1982) proposed a step-mutation model where multiple step changes (in either direction) was introduced. Such models can be easily rationalized in the context of the molecular mechanisms of unequal recombination and replication slippage, and these should be examined in greater detail to study the evolutionary dynamics of VNTR polymorphism.

Acknowledgements
This work was supported by the grant GM-41399 from US National Institutes of Health and 90-IJ-CX-0038 from the National Institute of Justice. We thank Prof. A. J. Jeffreys for his constructive comments on the work and we are grateful to individuals from Nancy, France for their co-operation in our study.

References

Baird, M., Balazs, I., Giusti, A., Miyazaki, L., Wexler, K., Kanter, E., Glassberg, J., Rubinstin, P., and Sussman, L. (1986) Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity. *Am. J. Hum. Genet.* 39: 489–501.

Balazs, I., Baird, M., Clyne, M., and Meade, E. (1989) Human population genetic studies of five hypervariable loci. *Am. J. Hum. Genet.* 44: 182–190.

Bell, G. I., Selby, M. J., and Rutter, W. J. (1982) The highly polymorphic region near the insulin gene is composed of simple tandemly repeating sequences. *Nature* 295: 31–35.

Boerwinkle, E., Xiong, W., Fourest, E., and Chan, L. (1989) Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: Application to the apolipoprotein B 3′ hypervariable region. Proc. Natl. Acad. Sci. USA 86: 212–216.

Brown, A. H. D, Feldman, M. W., and Nevo, E. (1980) Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* 96: 523–536.

Capon, D. J., Chen, E. Y., Levinson, A. D., Seeburg, P. H., and Goeddel, D. V. (1983) Complete nucleotide sequence of the T24 human bladder carcinoma oncogene and its normal homologue. *Nature* 302: 33–37.

Chakraborty, R. (1984) Detection of nonrandom association of alleles from the distribution of the number of heterozygous loci in a sample. *Genetics* 108: 719–731.

Chakraborty, R. (1990a) Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. *Am. J. Hum. Genet.* 47: 87–94.

Chakraborty, R. (1990b) Genetic profile of cosmopolitan populations: Effects of hidden subdivision. *Anthrop. Anz.* 48: 313–331.

Chakraborty, R. (1991) Generalized occupancy problem and its application in population genetics. In: Sing, C. F., and Hanism, C. L. (eds), Impact of Genetic Variation on Individuals, Families and Populations. Oxford University Press, New York (in press).

Chakraborty, R., Fuerst, P. A., and Nei, M. (1978) Statistical studies on protein polymorphism in natural populations. II. Gene differentiation between populations. *Genetics* 88: 367–390.

Chakraborty, R., and Griffiths, R. C. (1982) Correlation of heterozygosity and number of alleles in different frequency classes. *Theor. Pop. Biol.* 21: 205–218.

Chakraborty, R., and Nei, M. (1982) Genetic differentiation of quantitative traits between populations or species. I. Mutation and random genetic drift. *Genet. Res.* 39: 303–314.

Chakraborty, R., Smouse, P. E., and Neel, J. V. (1988) Population amalgamation and genetic variation: Observations on artificially agglomerated tribal populations of Central and South America. *Am. J. Hum. Genet.* 43: 709–725.

Chakraborty, R., and Weiss, K. M. (1991) Genetic variation of the mitochondrial DNA genome in American Indians is at mutation-drift equilibrium. *Am. J. Phys. Anthrop.* (in press).

Cohen, J. E. (1990) DNA fingerprinting for forensic identification: Potential effects on data interpretation of subpopulation heterogeneity and band number variability. *Am. J. Hum. Genet.* 46: 358–368.

Collick, A., and Jeffreys, A.J. (1990) Detection of a novel minisatellite-specific DNA-binding protein. *Nucleic Acid Res.* 18: 625–629.

Clark, A. G. (1987) Neutrality tests of highly polymorphic restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 41: 948–956.

Craig, J., Fowler, S., Burgoyne, L. A., Scott, A. C., and Harding, H. W. J. (1988) Repetitive deoxyribonucleic acid (DNA) and human genome variation: A concise review relevant to forensic biology. *J. Forensic Sci.* 33: 1111–1126.

Devlin, B., Risch, N., and Roeder, K. (1990) No excess homozygosity at loci used for DNA fingerprinting, *Science* 249: 1416–1420.

Edwards, A., Hammond, H. A., Caskey, C. T., and Chakraborty, R. (1991) Population genetics of trimeric and tetrameric tandem repeats in four human ethnic groups. *Genomics* (in press).

Ewens, W. J. (1972) The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* 3: 87–112.

Efron, B. (1982) The Jackknife, the Bootstrap and Other Resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38. SIAM, Philadelphia.

Flint, J., Boyce, A.J., Martinson, J.J., and Clegg, J.B. (1989) Population bottlenecks in Polynesia revealed by minisatellites. *Hum. Genet.* 83: 257–263.

Fojo, S., Law, S., and Brewer, H. B. (1987) The human preapolipoprotein C-II gene complete nucleic acid sequence and genomic organization. *FEBS Letters* 213: 221–226.

Goodbourn, S. E. Y., Higgs, D. R., Clegg, J. B., and Weatherall, D. J. (1983) Molecular basis of length polymorphism in the human zeta-globin complex. Proc. Natl. Acad. Sci. USA 80: 5022–5026.

Huang, L. S., and Breslow, J. L. (1987) A unique AT-rich hypervariable minisatellite 3′ to the ApoB gene defines a high information restriction length polymorphism. *J. Biol. Chem.* 262: 8952–8955.

Jeffreys, A. J., Royle, V., Wilson, V., and Wong, Z. (1988) Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* 332: 278–281.

Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985) Hypervariable 'minisatellite' regions in human DNA. *Nature* 314: 67–73.

Kidd, K. K., Bowcock, A. M., Schmidtke, J., Track, R. K., Ricciuti, F., Hutchings, G., Bale, A., Perason, P., and Willard, H. F. (1989) Report of the DNA committee and catalogs of cloned and mapped genes and DNA polymorphisms. *Cytogenet. Cell Genet.* 51: 622–947.

Kimura, M., and Crow, J. F. (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49: 725–738.

Kimura, M., and Ohta, T. (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc. Natl. Acad. Sci. USA 75: 2868–2872.

Lander, E. S. (1989) DNA fingerprinting on trial. *Nature* 339: 501–505.

Li, W. H. (1976) A mixed model of mutation for electrophoretic identity of proteins within and between populations. *Genetics* 83: 423–432.

Ludwig, E. H., Friedl, W., and McCarthy, B. J. (1989) High-resolution analysis of a hypervariable region in the human apolipoprotein B gene. *Am. J. Hum. Genet.* 45: 458–464.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., and White, R. (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* 235: 1616–1622.

Nei, M., and Li, W. H. (1973) Linkage disequilibrium in subdivided populations. *Genetics* 75: 213–219.

Odelberg, S. J., Platke, R., Eldridge, J. R., Ballard, L., O'Connell, P., Nakamura, Y., Leppert, M., Lalouel, J. M., and White, R. (1989) Characterization of eight VNTR loci by agarose gel electrophoresis. *Genomics* 5: 915–924.

Sokal, R. R., and Rohlf, J. F. (1969) Biometry, 2nd edition. Freeman, New York.

Watterson, G.A., and Guess, H.A. (1977) Is the most frequent allele the oldest? *Theor. Pop. Biol.* 11: 141–160.

Wyman, A. R., and White, R. (1980) A highly polymorphic locus in human DNA. Proc. Natl Acad. Sci. USA 77: 6754–6758.

ordered notation permits investigation of the power of linkage studies by LOD score analysis from a new perspective. The theory developed can facilitate the mapping and characterization of complex human genetic traits. Additional heterogeneity in the HLA component of insulin dependent diabetes mellitus (IDDM) has been identified using this approach.

American Journal of Human Genetics
51(4) October 1992

46

Formal statistics of DNA Fingerprinting data and relatedness between individuals. R. Chakraborty and L. Jin. Genetics Centers, Grad. Sch. of Biomed. Sci., Univ. of Texas, Houston, Texas, USA.

In general genotypic similarities with traditional marker loci can not unequivocally specify biological relationships between individuals. DNA fingerprinting patterns, revealed by a single multilocus probe (MLP), or combination of patterns from several single locus probes (SLPs), provide opportunities for circumventing this problem, because of their extreme variability. In this research we derived analytical distributions of the observed number of distinct bands (n_X, n_Y) and the number of shared bands (n_{XY}) between two individuals (X and Y) of a specified kinship. These distributions, in turn, allow the formulation of a likelihood ratio approach to determine relatedness between individuals. The present theory avoids approximations that are currently required for determining relatedness between individuals using DNA fingerprinting data. Also, this unified theory is applicable to fingerprint data generated from a single MLP or from combinations of several SPLs. Applications of this theory with data on several variable number of tandem repeats (VNTR), and short tandem repeat (STR) loci, and their comparisons with several polymorphic protein loci used currently by the paternity testing laboratories suggest that the reliability of prediction of biological relatedness using VNTR and STR loci is far greater. These hypervariable loci are also able to provide statistical discrimination between more distant relatedness coefficients. Finally, we estimated the number of such loci that would be needed to discriminate different degrees of biological relatedness for a specified level of precision of discrimination. (Research supported by grants NIJ-90-IJ-CX-0038 and NIH-GM41399).

ordered notation permits investigation of the power of linkage studies by LOD score analysis from a new perspective. The theory developed can facilitate the mapping and characterization of complex human genetic traits. Additional heterogeneity in the HLA component of insulin dependent diabetes mellitus (IDDM) has been identified using this approach.

American Journal of Human Genetics
51(4) October 1992

46

Formal statistics of DNA Fingerprinting data and relatedness between individuals. R. Chakraborty and L. Jin. Genetics Centers, Grad. Sch. of Biomed. Sci., Univ. of Texas, Houston, Texas, USA.

In general genotypic similarities with traditional marker loci can not unequivocally specify biological relationships between individuals. DNA fingerprinting patterns, revealed by a single multilocus probe (MLP), or combination of patterns from several single locus probes (SLPs), provide opportunities for circumventing this problem, because of their extreme variability. In this research we derived analytical distributions of the observed number of distinct bands (n_X, n_Y) and the number of shared bands (n_{XY}) between two individuals (X and Y) of a specified kinship. These distributions, in turn, allow the formulation of a likelihood ratio approach to determine relatedness between individuals. The present theory avoids approximations that are currently required for determining relatedness between individuals using DNA fingerprinting data. Also, this unified theory is applicable to fingerprint data generated from a single MLP or from combinations of several SPLs. Applications of this theory with data on several variable number of tandem repeats (VNTR), and short tandem repeat (STR) loci, and their comparisons with several polymorphic protein loci used currently by the paternity testing laboratories suggest that the reliability of prediction of biological relatedness using VNTR and STR loci is far greater. These hypervariable loci are also able to provide statistical discrimination between more distant relatedness coefficients. Finally, we estimated the number of such loci that would be needed to discriminate different degrees of biological relatedness for a specified level of precision of discrimination. (Research supported by grants NIJ-90-IJ-CX-0038 and NIH-GM41399).