Final Technical Report

Title: **Development of improved Insertion-Deletion assays for Human and Ancestral Identifications from Degraded Samples**

Award: 2013-DN-BX-K036

Author: Bobby L LaRue

Abstract

Insertions-Deletions (INDELs) are a type of polymorphism where small sequences of DNA have been inserted or deleted in relation to a known consensus reference sequence. The differences between the alleles are based on amplicon size, rather than detecting a nucleotide substitution as used for Single Nucleotide Polymorphism (SNP) typing. These size differences can be easily resolved using capillary electrophoresis (CE) with traditional chemistries for assaying fragment length. Thus, no new instrumentation is required for INDEL analyses in standard forensic laboratories. Analytically, INDELs perform similar to that of STRs and can be multiplexed together to achieve a high power of discrimination, as well as be multiplexed with STRs to facilitate analyses of challenged samples. The amplicon sizes of INDELs can be designed to be short, which are comparable to those of SNPs and are optimal for highly degraded samples. Furthermore, unlike STRs, INDELs do not yield stutter peaks due to slippage during PCR. Thus, interpretation complexity of SNPs can be reduced compare with STRs, especially for Low Copy Number DNA profiles. The mutation rates of INDELs are about 10 times lower than those of the SNPs, which make INDELs desirable for DNA-based kinship analysis.

Development of INDEL panels will provide the forensic community, especially practitioners, new tools to enhance their abilities in analyzing low quantity or highly degraded samples with current standard technology. No new instrumentation is needed to implement these new assays in a standard forensic DNA laboratory. Higher success rates of genotyping can be obtained because of relatively small amplicon sizes and allele length differences based on allele detection. Less complicated interpretation protocols will be needed, since there is no stutter with INDELs. The final product eventually provides an alternative approach for cases with low copy number DNA evidence. In addition, the ancestry identification assay can be used in generating investigative leads and helping solve more cases.

In the following report the verification for the validity of our markers selected in our preliminary data are a robust method for individual identification, and we work with corporate partners to develop a retrotransposable element based strategy for genotyping individuals for HID purposes. We demonstrate how thes marker systems excel at genotyping both degraded and low quantity

DNA and surpass standard STR marker based systems with these type of specialized samples. We explore the use of massively paraleel sequencing to provide additional discriminatory power to INDEL systems and provide a mechanism to determine if there is a mixture present and potentially how to resolve a mixture typed with INDELs.

Additionally we describe the development of a two separate optimized INDEL genotyping panels with different purposes one Is suited for Human Identification (HID) purposes, and the other is for determining the biogeographic ancestry of an individual, or an ancestral informative marker (AIM).

Table of Contents

Executive Summary

Introduction

STATEMENT OF THE PROBLEM

Currently forensic genotyping relies on a type of genetic polymorphism referred to a s a short tandem repeat or STR. STR's can reliable type DNA samples with an astronomically high power of discrimination if provided with a small quantity of high quality DNA. STR's don't perform well with degraded DNA, and the standard approach with highly degraded DNA (environmental DNA, hair shafts, etc…). Additionally the shearing forces encountered during an explosion can mechanically degrade the DNA hindering efforts to associate biological material deposited by a potential bomb maker on an IED with the individual that deposited it,

Current Approaches to Problems

Standard approaches to Typing degraded DNA rely on technologies that are expensive, necessitate instrumentation that is not widely available to the forensic community at large, have greatly reduced discriminatory power, or all three combined. An approach that is routinely used is mitochondrial DNA typing. Mitochondrial DNA typing works for this type of sample because there are many 100 to 100's of copies of the mitochondrial Genome present in each cell, and it is relatively resistant to degradation. However Mitochondrial Genomes are small in physical size and have low discriminatory power as a result. Additionally, mitochondrial genotyping is very expensive and routinely is only done by a handful of practicing labs in the United States. Massively Paralell Sequencing of Single Nucleotide Polymorphisms (SNP)could be done in such a way to solve this issue, but it is still cost prohibitive for most laboratories to implement.

Current forensic DNA laboratory workflow is based on the separation of amplified fluorescently labeled DNA fragments by capillary electrophoresis. Substituting one set of primers for another and separating based on fragment length is both cost effective and requires little additional training on the part of existing laboratory staff. One method of doing this is to design primers around small insertions or deletions (INDEL) in the genome with the idea that the primers would be designed to set just outside of the INDEL. Depending on the distribution of the insertion or deletion alleles in populations, an INDLE allele could be selected to tell an individual from another regardless of population affinity, or whether or not an individual was from a certain ancestral population. The latter application could be quite valuable in cases where suspects don't exist, and the investigators are in search of "leads" as to the identity of potential suspects.

PURPOSE

The purpose of this project is to demonstate the utilility of INDEL and other similar bi-allelic markers that may be separated based on size. The following AIMS were proposed to help serve as guidelines to achieve these goals:

1. Select two panels of INDEL markers for human identification and ancestry identification, respectively, from the 1000 Genomes Project data with certain criteria one panel for HID and the other for AIM purposes
2. Design primers for the developed panels in Aim 1, including redesigning the primers for long INDELs to meet forensic needs and arranging the markers into 4 dye channel systems;
3. Develop multiplex assays based on the outcomes from Aim 1 & 2 and validate the assays following SWGDAM validation guidelines.

To date, Aims 1 and 2 are complete,and Aim 3 is only awaiting a release of the final project funds to complete the project as proposed.

Additionally a criticism arose during the project that INDELs are poor with mixture sample. As a bi-allelic marker (like SNPs), this is true, because they lack a true highly polymorphic nature like STR markers. To address this issue we utilized MPS to sequene the flanking region around our HID INDEL panel to identify additional polymorphisms that could be utilized to increase discriminatory power and to help resolve potential mixtures

We also genotyped multiple types of degraded samples to demonstrate that INDELs perform better with highly degraded DNA samples (Enbalmed remains, ancient DNA, rootless hairshafts, and explosive fragments).

RESEARCH DESIGN

Research design and methods

The ultimate goal of this study is to develop and validate two INDEL panels for human identification and ancestry identification, respectively. Three step-by-step aims are proposed to accomplish the goal as illustrated in Fig. 2.

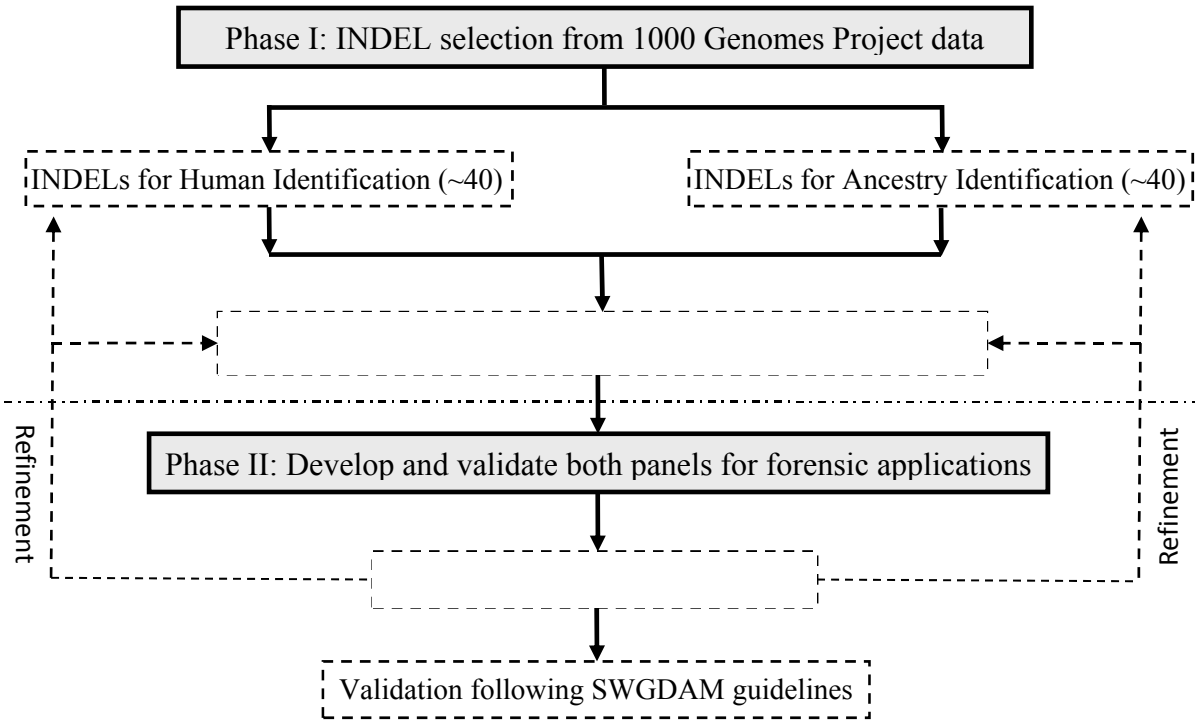


Figure 2. Schema chart of the research design.

The first aim is to select two panels of INDEL markers for human identification and ancestry identification, respectively, from the 1000 Genomes Project data. In the three pilot projects of the 1000 Genomes Projects, the low-coverage sequencing project meets the requirements for forensic markers development, because it covered the whole genome (including non-coding regions), and sequenced major populations with relatively large numbers of unrelated individuals. 500 samples were sequenced for each of the five major populations (i.e., African, Caucasian, East Asian, South Asian, and Native American). There are more than 1 million short INDELs and about 20,000 structure variants available for marker selection. Two panels of INDELs, for either human identification or ancestry identification, will be selected based on markers residing on the 22 autosomal chromosomes. Each panel will contain about 40 markers. Instead of selecting any INDEL at X or Y chromosomes, Amelogenin will be included for sex determination. Mitochondrial DNA will be ignored because of its relatively high mutation rate and heteroplasmy. The selection criteria of markers for **human identification** in the specified order are as follows:

1) The minimum allele size difference is at least 4, so that potential anomalies (e.g., +A issue during PCR) can be eliminated and the alleles are relatively easy to be differentiated by CE.
2) The Minimum Allele Frequency (MAF) is at least 0.3 for all three major populations (average heterozygosity ≥ 0.42 and match probability ≤ 0.4246).
3) Population substructure measure (F_{ST}) should be ≤ 0.06, which is a common threshold in SNP panel development for human identification.

4) No significant deviation from HWE should be detected beyond expectations observed by chance.
5) The marker is located in non-coding regions. The software "Variation Pattern Finder" provided by 1000 Genomes Project can be used to search associated coding regions [48]. dbSNP [31] at NCBI also provides the same function.
6) The sequences of the markers and their flanking regions will be obtained from the 1000 Genomes Projects by "BAMtools" [49]. These sequences will then be searched against NCBI to remove potential cross-reactivity with other species.
7) The distance between the markers on the same chromosome should be at least 50 Mb to avoid genetic linkage (but may be relaxed depending on specifications). LD tests will be performed to exclude markers with significant LD with other markers. Greedy algorithm will be used in this selection, in which the markers associated with the most LD pairs will be excluded first. But these excluded markers will be kept as suboptimal markers for multiplex assay refinement.

For the **ancestry identification** panel, the step in marker selection of human identification panel is still required, but the steps 2 and 4 are not necessary, because ancestry identification requires large differences of allele frequencies among populations and substructure within the major populations does not affect the accuracy of ancestry identification; indeed it can be indicative of candidate bioancestry markers. The step 3 in the human identification INDEL selection criteria will be modified. Only markers with high Fst (e.g., ≥ 0.2) will be selected. Steps 5 and 6 are still required. In step 7, because kinship analysis does not apply to the ancestry identification panel, the marker distance threshold can be reduced to 1 Mb, since markers with more than 1 Mb are generally not in LD at the population level [50, 51]. LD tests and succeeding marker selection are still necessary.

With ~40 INDELs meeting the selection criteria, the cumulative match probability can reach at least 1.3×10^{-15}, similar to that of 13 CODIS core STRs for human identity testing. The target accuracy of ancestry identification is >99%. Suboptimal markers (i.e., the markers excluded due to significant LD) will also be considered as backup selections for future panel refinement.

FINDINGS

We were able to verify the validity of our markers selected in our preliminary data are a robust method for individual identification, and we work with corporate partners to develop a retrotransposable element based strategy for genotyping individuals for HID purposes. We demonstrate how thes marker systems excel at genotyping both degraded and low quantity DNA and surpass standard STR marker based systems with these type of specialized samples. We explore the use of massively paraleel sequencing to provide additional discriminatory power to INDEL systems and provide a mechanism to determine if there is a mixture present and potentially how to resolve a mixture typed with INDELs.

Additionally we describe the development of a two separate optimized INDEL genotyping panels with different purposes one Is suited for Human Identification (HID) purposes, and the other is for determining the biogeographic ancestry of an individual, or an ancestral informative marker (AIM).

CONCLUSIONS

We were able to demonstrate the utility of INDEL and INNUL based markers for a variety of use as cost effective and reliable adjunct markers to standard STR genotyping. The INDEL panels we have developed are optimized for low quantities of of low quality DNA, and our primers and protocols are freely available to practicoioners in the field. These systems are ideal for poor quality exhumed remains, fragments from explosive devices, and with some refinement of extraction protocols, they have shown promise for genotyping rootless hair shafts.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

FINAL TECHNICAL REPORT

Introduction

Statement of the problem

The primary genetic markers for current human identification and forensic investigations are the Short Tandem Repeat (STR) loci. These markers are highly polymorphic and are amenable to semi-automated analyses. One area that could be improved upon however is the typing of degraded samples. Degraded DNA samples, which can be caused by environmental exposure, as a natural process of necrosis, or through exposure to mechanical shearing forces (such as in explosions associated with disasters) tend to fragment DNA into pieces of approximately 200 base pairs [1-3]. Single Nucleotide Polymorphisms (SNPs) have been suggested as adjunct tools to STRs for this particular niche, owing to their abundance in the human genome, utility for genotyping degraded DNA samples with their relatively short amplicon sizes, low mutation rate, and potential to be analyzed in an automated fashion [4-8]. However, SNP-based detection systems typically require that the chemistry and instrumentation reliably detect a single base substitution. To date, complex analytical approaches have been sought which are unwieldy and often not quantitative [9-14]. Another type of bi-allelic marker is Insertions-Deletions (INDELs), which can provide an alternative system for forensic purposes.

The polymorphism of INDELs, which is based on the presence or absence of an insertion or deletion, can be exploited more readily with a simplified analytical process [8, 14-16]. The difference between the alleles is based on size rather than detecting a nucleotide substitution and these differences can be resolved using capillary electrophoresis (CE). Thus, the instrumentation for INDEL analyses is commonly found in forensic laboratories. Essentially, INDELs perform analytically similar to that of STRs and in theory can be multiplexed together to achieve a high power of discrimination as well as be multiplexed with STRs to facilitate analyses of challenged samples. The amplicon sizes of INDELs are usually short (≤200 bp), which are comparable to those of SNPs and are optimal for highly degraded samples which typically. Furthermore, unlike STRs, INDELs do not yield stutter peaks due to slippage during PCR. Thus, interpretation complexity of SNPs can be reduced when compared with STRs, especially for degraded Low Copy Number DNA profiles.

Many INDELs have been found in the human genome and are contained within databases. Weber et al. [19] identified about 2,000 human diallelic short INDELs by comparing overlapping genomic or cDNA sequences and tested them with African, Japanese, European, and Native American samples. An INDEL database was initiated with Weber et al.'s study, the Marshfield Diallelic Insertion/Deletion Polymorphisms database, based on INDEL panels developed for human and ancestry identifications. In 2006, Mills et al. [30] built an initial map of human INDEL variation. This map contains 415,436 INDELs ranging from 1 bp to 9989 bp in length. These INDELs have been uploaded to dbSNP [28]. Both databases are publicly accessible.

Pereira et al. [14] in 2009 described the first INDEL multiplex assay for human identification. This assay contains 38 autosomal INDELs and was selected from ~4,000 markers which have been confirmed to be present in major populations (i.e., African, Caucasian, and Asian) in the Marshfield INDEL database. The selection criteria included: non-coding region markers, Minimum Allele Frequency ≥ 0.25 in major populations, average heterozygosity ≥ 0.4, and allele length differences around 2-5bp. Markers with known polymorphisms or mononucleotide repeats (≥7 bp) in their flanking sequences were excluded. Primers were designed using Primer3 [32]. The amplicon sizes were designed to be less than 160 bp. Optimum T_M was set at 60° C with a

minimum of 58° C and optimum GC content was set at 50% with a minimum of 45%. Non-specific hybridization was checked with BLAST at NCBI [33]. AutoDimer [34] was used to check for hairpin and primer–dimer secondary structures. To avoid linkage disequilibrium, the markers residing a short distance from each other on the same chromosome were excluded, so for single source profile comparisons these markers can be treated independently. However, the distances between several pairs of markers on the same chromosome were less than 50 Mb apart. For example, the distance between rs3080855 and rs34511541 on chromosome 18 is only about 13 Mb. This distance suggests that these markers can be genetically linked and independence between these markers may not be assumed. Thus, kinship analysis with this 38 INDEL panel needs more complicated interpretation, either incorporating recombination fractions between linked markers or removing the less informative marker of two linked markers from kinship analysis [35].

More recently, LaRue et al (manuscript in review) has looked at the distribution of 114 INDELs in 3 major North American populations. Of the 114 INDELs a primary panel of 38 candidate markers was selected that met the criteria of 1) a minimum allele frequency of greater than 0.20 across the populations studied; 2) general concordance with Hardy-Weinberg equilibrium (HWE) expectations; 3) relatively low FST based on the major populations; 4) physical distance between markers greater than 40 Mbp; and 5) a lack of linkage disequilibria between syntenic markers. Additionally, another 11 supplemental markers were selected for an expanded panel of 49 markers which met the above criteria, with the exception that they are separated at least by 20 Mbp. The resulting panels had Random Match Probabilities that were at least 10-16 and 10-19, respectively, and combined FST values of approximately 0.02. Given these findings, these INDELs should be useful for HID.

An INDEL kit for human identification is commercially available, Investigator DIPplex® kit, (Qiagen) [33]. It is a multiplex five-dye single-tube reaction assay for 30 bi-allelic INDEL markers and Amelogenin. This kit has been validated and population studies were also performed with four populations (African American, Asian, Caucasian, and Hispanic) [37]. Other validations or population studies were reported by Alvarez et al. [38] with the same four populations as LaRue et al. [37], Neuvonen et al. [39] with the Finnish and Somali populations, and Friis et al. [40] with the Danish population. According to LaRue et al. [37], this assay was able to type DNA from a number of forensically-relevant sample types and obtain full profiles with 62 pg of template DNA and partial profiles with as little as 16 pg of template DNA. The assay is reproducible, precise, and non-overlapping alleles from minor contributors were detectable in mixture analysis ranging from 6:1 to 19:1 mixtures. There were no significant departures from Hardy–Weinberg Equilibrium (HWE) or significant Linkage Disequilibrium (LD) between the markers (after correction for sampling). However, these markers were initially selected based on the Caucasian population. MAFs were at least 0.36 for Caucasian at all markers, but were lower than 0.2 for African American or Asian populations at several loci. “Off-ladder” and peak height microvariant alleles were observed at multiple loci in the population study, suggesting that these particular indel markers may not be optimal as forensic markers for all populations. In addition, the minimum distance between two of the markers was only 1.7 Mb (rs1636 and rs6481 at chromosome 22). Similar to the panel in Pereira et al. [14], independence between markers cannot be assumed in kinship analysis because of close genetic linkage [35]. Li et al. [16] also developed a 29-INDEL panel with MAF ≥ 0.2 for East Asian population based on dbSNP [31]. However, similar to [14] and [36], several pairs of markers were physically close (~3M bp) on the same chromosome.

Since the allele frequencies of many INDELs vary notably in different populations [23,30], INDELs also can be used to identify bioancestry of individuals. These markers are called Ancestry Informative Markers (AIM). Several AIM INDEL panels have been developed based on the Marshfield INDEL database [19]. Bastos-Rodrigues et al. [41] selected 40 slow-evolving short INDELs to analyze population genetic structure and tested them with seven populations. This panel also was evaluated for paternity testing [42]. Santos et al. [21] selected 48 AIM INDELs to measure the proportions of three different ancestries (sub-Saharan African, European, and Native American). This study did not include East Asian. Pereira et al. [20] selected 46 AIM INDELs to measure admixture proportions of four populations (African, European, East Asian and Native American). In these AIM selections and/or validations, either allele frequency variation measure (δ) or population substructure measure (Fst) was used to quantify the informativeness of these AIMs. Generally, more than 90% accuracies of identifying the major populations can be achieved with these AIMs.

X chromosome linked INDEL panels were also developed for both human identification and ancestry identification. Ribeiro-Rodrigues et al. [43] first analyzed 13 X-INDELs for a population admixture study in Brazil. This panel was extended further to 33 X-INDELs [44]. Pereira et al. [45] developed a 32 X-INDEL panel with African, European, and Asian populations but with MAF ≥ 0.1 for Asians. Its design was more focused on European and African populations.

Long INDELs or Retrotransposable Elements (REs) (e.g., ALUs) also have been used in human identification and bioancestry identification [24-28]. Novick et al. [24] first tested five human specific ALU insertions for forensic analysis. Allele frequencies of other ALU based panels also were reported, such as those by Dinç et al. [26] for the Anatolian population with ten ALUs and Selvaggi et al. [27] for the Piedmont (Northern Italy) population with eleven ALUs. Ray et al. [25] reported a large panel with 100 ALUs for ancestry identification, but the selection criteria were not explained. Recently, Mamedov et al. [28] selected 31 autosomal ALUs with at least a distance of 50 Mb between markers from a Russian database (http://labcfg.ibch.ru/Home.html) and one ALU on the X chromosome for sex identification. The MAF was at least 0.25. However, the amplicon size ranged from 200 – 600 bp, which is too large for forensic purposes, especially for degraded samples.

Most previous and current studies were based on relatively small databases, such as the Marshfield INDEL database. Because of a limited number of markers in this database, many selected INDELs had MAF ≤ 0.2, and thus, a relatively low power of discrimination for certain population(s). Recently, the 1000 Genomes Project aimed to explore human genome sequence variations by whole-genome sequencing hundreds of individuals from various populations [20]. In the pilot project, three projects were carried out: low-coverage sequencing (~3.6×) of 179 individuals from African, Caucasian, and East Asian; high-coverage sequencing (~42×) of two mother–father–child trios; and exon-targeted sequencing (~56×) of 697 individuals from seven populations. More than 1 million short INDELs (1-50bp) and 20,000 structural variants, including a good proportion of long INDELs (e.g., ALU and LINE), have been identified. Most of the INDELs were novel (i.e., not present in dbSNP previously). In the main project, more samples (~2,500) from more populations were sequenced, including 26 small populations from 5 major groups (i.e., African, Caucasian, East Asian, South Asian, and Native American). Although the data analysis is still underway, more INDELs are expected to be found. With this database, more INDELs with a high power of discrimination or high allele frequency difference among the population could be found. This database is potentially a great asset for developing INDEL multiplex assays for forensic purposes.

Preliminary study

REs range in size from hundreds (SINEs) to thousands (LINEs) of bp in length. Because the allele forms in these long INDELs are not the result of a deletion, they are actually insertion or null alleles. Earlier attempts to use Alu sequences for identity testing capitalized on the size difference between insertion and null alleles by amplifying the entire region with the same forward and reverse primers [28]. The insertion allele would be 200-600 bp larger than the null allele, and could be detected electrophoretically based on size differences. While useful for paternity testing and some population studies where DNA quality is not compromised, the large amplicon size difference between the alleles impacted amplification efficiency during PCR and is a limitation for forensic samples. The smaller amplicon (i.e., the null allele) is favored during amplification but the insertion larger amplicon with insertion element may drop out. The differential amplification or allele drop out can be exacerbated if the sample is highly degraded. Thus, the use of REs has not been embraced for the analysis of forensic samples [46].

Recently, LaRue et al (47) described a novel primer design, which can reduce the amplicon size and allele state differences of SINES and LINES such that these markers can be used effectively on forensic type samples. Thus, these markers now are amenable to analyzing degraded samples, as the amplicon size can be reduced from thousands and hundreds of bases to less than 100 bases in length and differential amplification of the allelic states can be dramatically reduced if not eliminated as the size of the insertion and null allele states can differ by only one to a few bases. Figure 1 illustrates this novel primer design. A common forward primer (FP) is used for both the insertion and null alleles. A "Null-Specific" reverse primer that straddles the insertion site is able to anneal with the null allele sequence (Fig. 1.a). "Insertion-Specific" reverse primers are designed to anneal with the insertion region, either straddling the insertion site (i.e., same as Null-Specific reverse primer) or just inside of the insertion allele (Fig. 1.b).

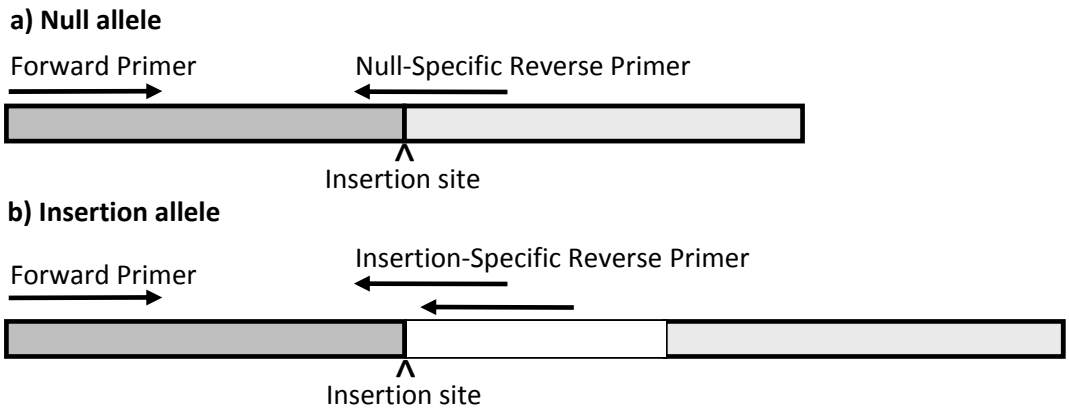


Figure 1. Novel primer design schema, including three primers: forward primer shared by both insertion and null alleles, Null-Specific reverse primer only for the null allele, and Insertion-Specific reverse primer only for the insertion allele.

To test this novel primer design, primers for nine REs, including both SINEs and LINEs (i.e., CH14-50-6236, CH4-12-7012, CH6-28-9163, LC3-2601, TARBP1, Yac52265, Ya5NBC51, Yb7AD155, and Yb8NBC106) were constructed using an *in silico* approach. A sensitivity study showed that full profiles could be obtained with as little as 125 pg of DNA. Sensitivity may be further increased by optimizing the primers and PCR conditions. Population studies were performed for these nine markers with 279 individuals from African-Americans, Caucasians, and Southwestern Hispanics (93 individuals for each population). No significant deviation from HWE and LD was detected after Bonferroni corrections. Both markers with high Fst (e.g., 0.26 at CH6-28-9163), which may be good for ancestry identification, and low Fst (e.g., ~0 at Yb7AD155),

which is applicable to for human identification, were found. With these nine markers, the cumulative match probability can reach at least 1.2×10^{-3}. The results demonstrated that REs have the potential for serving as a set of markers that can be used for forensic applications, especially for limited quantities of DNA. The "Supporting data" section in Appendices describes more details of this preliminary study.

Current Approaches to Problems

There have been several short INDEL panels developed for human identification [14-16]. They were designed either for the X-chromosome only [15], for certain populations [16], or based on relatively small data sets (e.g., Marshfield Diallelic Insertion/Deletion Polymorphisms database [13, 19]. INDEL panels for ancestry identification also have been reported [20-22] based on the same database [19].

Rationale for the research

The pilot project of the 1000 Genomes Project [23] has shown by low-coverage sequencing that there are more than 1 million short INDELs and 20,000 structural variants, including long INDELs (e.g., ALUs), in three major populations (African, Caucasian, and East Asian). Most of the INDELS are novel markers. Hence, a large pool of potential INDEL markers is available in this database to assess for forensic purposes.

The Marshfield Diallelic Insertion/Deletion Polymorphisms database [19] primarily contains short INDELs (allele length difference ≤50). However, there have been studies showing that long INDELs or Retrotransposable Elements (REs), such as ALUs, can be used in human identification and bioancestry identification [23-28]. These larger insert polymorphisms are expected to have very low mutation rates, less than short insertion elements. However, large amplicon size differences of long INDELs limit their use for analyzing degraded samples, because the amplicon sizes range from hundreds (i.e., Short Interspersed Nuclear Elements; SINE) to several thousand or more (i.e., Long Interspersed Nuclear Elements; LINE) base pairs in length. Recently, LaRue et al [47] demonstrated that alternative primer designs can reduce the amplicon size and allele state differences of SINES and LINES such that these markers can be used effectively on forensic type samples. Thus, these markers now are amenable to analyzing degraded samples, as the amplicon size can be reduced from thousands and hundreds of bases to less than 100 bases in length and differential amplification of the allelic states can be dramatically reduced if not eliminated as the size of the insertion and null allele states can differ by only one to a few bases. Better primer sets design is needed to reduce the amplicon size and difference between the two allele states of these large insert polymorphisms to exploit their power for human identity testing

II. **METHODS**

Samples and DNA Extraction

DNA was extracted from whole blood samples obtained from 190 unrelated individuals following the University of North Texas Health Science Center Institutional Review Board Approval. The sample set represented unrelated individuals of four major U.S. population groups with 49 Caucasians (CAU), 49 African Americans (AFA), 49 Hispanics (HIS), and 43 Asians (ASA). DNA extraction was performed using the Qiagen® QIAamp™ DNA Blood Mini Kit (Qiagen, Hilden, Germany), according to the manufacturer's protocol.

INDEL analysis of Enbalmed samples

Amplification of INDEL markers was performed on all samples using a prototypical 42-loci INDEL multiplex. DNA (0.8 ng in 3 µL) was added to 15 µL of AccuPrime ™ Taq DNA Polymerase System (ThermoFisher Scientific) PCR Master Mix, 4 µL of Primer Mix, and 3 µL of sterile, deionized water for a total reaction volume of 25 µL. Amplification was performed on a GeneAmp® PCR System 9700 (ThermoFisher Scientific) using the following parameters: an initial denaturation at 95 °C for 11 min, followed by 30 cycles of 94 °C for 20 s and 59 °C for 3 min, with a final extension of 60 °C for 30 min and a hold at 4 °C. Separation and detection of the amplified product was performed on a 3500XL Series Genetic Analyzer (ThermoFisher Scientific), using a 36 cm capillary array and POP-4 (ThermoFisher Scientific). An aliquot of PCR product (1 µL) was added to 11 µL of master mix (10.5 µL Hi-Di™ Formamide and 0.5 µL of Genescan™600 LIZ™ Size Standard v2.0 (ThermoFisher Scientific).

Results were analyzed using GeneMapper® ID-X Software v1.4 (ThermoFisher Scientific). The analysis method utilized a previously established analytical threshold of 100 RFUs as a minimal cutoff value for peak detection. For heterozygous loci with peak height ratios between 10-50 %, the data for the entire locus was discarded for that sample. Any minor heterozygote peak with a height below 10 % of the major peak was considered elevated baseline and disregarded. Random match probabilities (RMP) were then calculated in a population specific manner based on data presented in LaRue et al. for all samples using interpretable loci. Previous studies have demonstrated that three of the loci (2032678, 28362545, and Amelogenin) in the prototype multiplex violated assumptions necessary for reliably calculating RMP. As such these loci were excluded from calculations.

INNUL Analysis

Neat DNA (16 µL) was amplified in a 25 µL reaction volume using an early access version of the InnoTyper™ 21 Kit (InnoGenomics Technologies, LLC, New Orleans, LA) with the recommended cycling conditions. Amplified DNA products were separated on the 3500 XL Genetic Analyzer (ThermoFisher Scientific) using a 36 cm capillary array with POP-4 polymer. Data was analyzed with GeneMarker® HID software (SoftGenetics, State College, PA, USA) with an analytical threshold of 50 RFUs being applied. RMPs were calculated using allele frequencies from the Caucasian population

Primer Design

INDEL markers forward and reverse primers were designed to amplify each marker. Publically available software tools through the websites dbSNP, the UCSC Genome Browser and Primer-BLAST, were used to assist in designing the primers. Each of the chosen primer pairs were checked for potential dimerization with the other primer pairs using the publically available software tool, PriDimerCheck. After determining that no major issues should occur, unlabeled primers were ordered from Invitrogen™.

Unlabeled Primers

The unlabeled primers were run individually to ensure that each primer pair was performing as intended and would successfully amplify the DNA. For amplification, each sample contained the PCR mix made up of 5.5 µL of water, 2.5 µL of 10x buffer, 2.5 µL of BSA (10 mg/ mL), 2 µL of 50 mM MgCl2, 1 µL of 10 mM dNTPs, 0.5 µL Taq polymerase (5 U/ µL), 0.5 µL of the forward primer at a concentration of 10 µM, 0.5 µL of the reverse primer at a concentration of 10 µM, and 10 µL of DNA (1 ng/ µL). The samples were amplified on the Applied Biosystems® GeneAmp® PCR System 9700 thermocycler under the parameters of 95ºC for 11 minutes, 36

cycles of 95ºC for 10 seconds, 61ºC for 30 seconds, 72ºC for 30 seconds, and a final extension of 70ºC for 10 minutes. Once amplified, each marker was assessed using the Agilent© 2200 Tapestation using 2 µL of the TapeStation buffer with 2 µL of sample, following laboratory protocol (28). All individually run primer sets were evaluated to ensure each worked and produced a product around its estimated base pair range. After the primer pairs were run individually, they were arranged in groups of 5 primer pairs for an initial multiplex design. The amplification of the multiplex used the Qiagen® Multiplex PCR kit where each sample tube contained 5 µL of 10X primer mix, 10 µL of DNA (1 ng/ µL), 10 µL of water, and 25 µL of 2X multiplex PCR master mix, as instructed by the protocol (29). The 10X primer mix was made by 10 µL of the forward primer (10 µM) and 10 µL of the reverse primer (10 µM) added to 400 µL of water. Each sample was amplified on the thermocycler under the parameters of 95ºC for 5 minutes, 28 cycles of 95ºC for 30 seconds, 60ºC for 90 seconds, and 72ºC for 90 seconds, then 68ºC for 10 minutes. The products of this initial multiplex were run on the TapeStation to assess if the primers could be amplified together and still produce intended results.

Fluorescently Labeled Primers

After the initial multiplex was evaluated to determine if it would perform as intended, fluorescently labeled primers were ordered from Applied Biosystems®. The fluorescent labeled primer pairs were run first individually to ensure they were functioning properly. For amplification of the individual primer pairs, each sample contained the PCR mix made up of 5.5 µL of water, 2.5 µL of 10x buffer, 2.5 µL of BSA (10 mg/ mL), 2 µL of 50 mM MgCl2, 1 µL of 10 mM dNTPs, 0.5 µL Taq polymerase (5 U/ µL), 0.5 µL of the forward primer (10 µM), 0.5 µL of the reverse primer (10 µM), and 10 µL of DNA (0.5 ng/ µL). After amplification, the samples were analyzed using the Applied Biosystems® 3500 Genetic Analyzer. Each sample well for the CE contained 9.6 µL of HiDi formamide, 0.4 µL of Liz 600, and 1 µL of the amplified sample. The results from the CE were analyzed using GeneMapper ID-X (version 1.2). The next step was to create an initial multiplex of amplicons separated into 5 groups based on their fluorophore color. Using the Qiagen® Multiplex PCR kit where each sample tube contained 5 µL of 10X primer mix, 10 µL of DNA (0.05 ng/ µL), 10 µL of water, and 25 µL of 2X multiplex PCR master mix. The 10X primer mix was prepared by adding 10 µL of the forward fluorescent primer (20 µM) and 2 µL of the reverse unlabeled primer (100 µM) and diluting to 100 µL. Each sample was amplified on the thermocycler under the parameters of 95ºC for 5 minutes, 28 cycles of 95ºC for 30 seconds, 60ºC for 90 seconds, and 72ºC for 90 seconds, then 68ºC for 10 minutes. The amplified products were analyzed on the CE with each well containing 9.6 µL of HiDi formamide, 0.4 µL of Liz 600, and 1 µL of the amplified sample. After evaluating the profiles using GeneMapper ID-X, 1:10, 1:20, 1:50 and 1:100 dilutions of the amplified product were analyzed to reduce the amount of fluorescent dye overlap among amplicons. A new multiplex of the fluorophores was created with the amounts of each primer pair in the dye channel based on generating interlocus peak height balance. The new multiplex was amplified and run on the CE under the same parameters as described previously but with the 10X primer mix (20 µM) described depending on each primer pair and using a 1:100 dilution of the 0.5 ng/ µL DNA.

Library Preparation and Massively Parallel Sequencing of INDEL markers

Libraries were generated using a custom designed Nextera™ Rapid Capture Enrichment panel (Illumina, Inc., San Diego, CA) using the Illumina Design Studio as described by Zeng, et al.. The HID INDELs for this study were selected based on the results described by LaRue, et al. and Pereira, et al.. INDEL rs number, location, flanking region, and probe design are listed in Supplemental Table 1. Each sample library was diluted to 2 nM and paired-end sequencing

was performed on the Illumina MiSeq™ according to the manufacturer's recommended protocol with a read length of 250 bases.

STRait Razor Design

A configuration file was created for use with STRait Razor v2.5. To create the file, locus coordinates for each INDEL were located on the hg19 human reference genome using the Integrative Genomics Viewer (IGV). STRait Razor flanking regions up and downstream of the INDEL motif, and the complementary sequences, were recorded. The average size of the STRait Razor flank, used to mine sequence data for regions of interest, was 24 bases ± 0.10. The bases between STRait Razor flanks contained the INDEL motif and approximately 50 bases on either end. The STRait Razor flanks were designed to capture sequence variation in the flanking regions adjacent to the target INDEL (Supplemental Tables 3 and 4) while keeping total target size relatively small. The average length of this region (target INDEL plus approximately 50 bases on either end) was 99 bases ± 4 and 102 bases ± 4 for the deletion and insertion alleles, respectively. Lastly, a relatively short sequence between the flanking regions, but unique relative to the INDEL motif, was recorded; the average length of these sequences was 25 bases ± 1. Analysis of the resulting data was performed using the STRait Razor Sequence Analysis toolkit to assign genotypes to each sample and compile depth of coverage (DoC) and allele coverage ratio (ACR) data.

Analysis Concordance

A subset of 69 samples was analyzed manually to confirm STRait Razor allele calls. Fastq files were aligned using Burrows-Wheeler Aligner (BWA) and Sequence Alignment/Map Tools (SAMtools). The resulting binary alignment/map (.bam) files were used as input for the Genome Analysis Toolkit (GATK). The resulting variant call format (.vcf) files were analyzed using an in-house Excel-based workbook. The workbook assigned genotypes and compiled DoC and ACR data for each sample.

Population Statistical Analyses

Length-based and sequence-based allele frequencies, observed and expected heterozygosities, and testing for departures from Hardy-Weinberg Equilibrium (HWE) and linkage disequilibrium (LD) assessments were performed using Genetic Data Analysis (GDA). An in-house Excel-based workbook was used to generate power of discrimination values and single-locus and combined RMPs.

In silico Samples

Variant call files for chromosomes 1 through 22 were downloaded from the 1000 Genomes Project website (http://www.1000genomes.org/data). These files contain the autosomal genome data for 3500 individuals. When possible, the populations were binned into their associated major global population group (Table 1). Of the 3500 individuals, 550 individuals (Caucasian, N=244; African, N=156; and East Asian, N=150) were chosen comprising the training set used for marker selection for differentiating Caucasian, African, and East Asian ancestry.

Table M1. Populations from 1000 Genomes Project binned into three major global population.

African	Caucasian	East Asian
Yoruba in Ibadah, Nigeria	British in England and Scottland	Southern Han Chinese, China

African ancestry in southwest U.S.	Finnish in Finland	Han Chinese in Bejing, China
Luhya in Webuye, Kenya	Utah resident with Western Europe ancestry	Japanese in Tokyo, Japan
	Toscani in Italy	

Marker selection

Using the Linux-based software program, VCFtools, chromosomal data was filtered to include only INDEL markers. Pairwise population substructure, F_{ST}, was calculated for each INDEL. The output file was sorted in Excel® and a new text file containing only the positions with F_{ST} greater than 0.5 in at least one pairwise comparison was created. From this new file, population-specific allele frequencies were calculated. The markers were then manually sorted by length and markers of length 3-6 base pairs were considered for selection. This is so that they are large enough to resolve the alleles by CE, but small enough to conserve real estate in the CE system. Next, to verify that these markers are true INDELs, and part of a repeat region, the UCSC Genome Browser was used to eliminate markers that showed a presence of repeat or proto-repeat sequences near the INDEL by enabling the Repeat Masker algorithm. From this reduced list of markers, 20 per population group were selected based on high allele frequency divergence, or delta value, and genetic distance. Markers on the same chromosome were selected to be more than 1 Mb from its nearest neighbor in order to minimize potential for selection linked loci.

Statistical Analysis

Using VCFtools, genotype data for the 550 training set individuals were pulled out of the variant call files. The panel of 60 AIMs was evaluated by Principal Component Analysis using the software program Past3 to determine if the populations would cluster separately. Additionally, known populations were added to the analysis to determine if they would cluster with the expected population group. Next, the 60 AIMs were evaluated for Hardy-Weinberg Equilibrium (HWE) and linkage disequilibrium (LD). Using the software package, Genetic Data Analysis (GDA), exact tests for HWE and LD were performed. The AIMs panel was then evaluated for ancestry admixture using the program STRUCTURE v.2.3.4.

Table M2. Linux commands used for VCFtools.

	Command	Input File Format	Output Format
Pairwise F_{ST} calculations	vcftools --gzvcf ~/*chromosome1*.vcf.gz --keep-only-indels --weir-fst-pop ~/*pop1*.txt --weir-fst-pop ~/*pop2*.txt --out	*chromosome1.*vcf.gz- variant call file downloaded from 1000 Genomes Project website *pop1*.txt/ *pop2*.txt- text file containing 2 columns: (1)	*outfile*.weir.fst

	~/*outfile*	sample name (2) population	
Allele frequency calculations	vcftools --gzvcf ~/*chromosome1*.vcf.gz --positions ~/*fst*.txt --keep ~/*pop1*.txt --freq --out ~/*outfile*	*fst.*txt- text file containing the positions of INDEL markers with Fst >0.5 *pop1*.txt- text file containing a single column with one sample per line	*outfile*.frq
Genotype Data	vcftools --gzvcf ~/*chromosome1*.vcf.gz --keep ~/pop1.txt --positions ~/*markers1* --recode --recode-INFO-all --out ~/*outfile*	*pop1*.txt- text file containing a single column with one sample name per line *markers1*- gedit file containing 2 columns: (1) chromosome number (2) position	*outfile*.recode.vcf

Italicized words indicate file names

Additional Populations

Southwest Hispanic (SWH; N=243) and Southwest Asian (SWA; N=489) samples were compiled from the 1000 Genomes Project data. The genotype data for the 59 AIMs for these individuals were retrieved using VCFtools and added to the PCA of the original training set. Using Past3, these two population groups were also compared pairwise in PCA.

Table M3. Populations in 1000 Genomes Project binned into additional population groups.

Southwest Hispanic	Southwest Asian
Colombian in Medellin, Colombia	Gujarati Indian in Houtson, TX
Peruvian in Lima, Peru	Bengali in Bangladesh
Mexican Ancestry in Los Angeles, California	Sri Lankan Tamil in the UK
	Indian Telugu in the UK

III. RESULTS

HID Bi-Allelic Markers

Verification of the suitability of the HID markers in more diverse populations

This portion of the findings was published in *The International Journal of Legal Medecine* [REF]

The 49 INDEL markers described in the preliminary data of our proposal had yet to be assessed in admixed populations or isolated outside of North America. These markers with high discrimination power for the major populations of North America, but further populations studies were deemed necessary before the investment of resources was committed to the design of

primers that might not perform well in diverse global populations. Upon learning of our award for the project, we set out to rectify this problem.

To allay these concerns, the efficiency of the 49 INDEL markers was investigated in two urban admixed population samples Rio de Janeiro, Brazil, and Tripoli, Libya and one isolated Native Brazilian community. To do this we utilized our initial trial multiplexes of 114 INDELs and only interpreted the 49 loci selected by our preliminary study for analysis.

A panel of 49 INDEL loci was typed in three population samples:Native Brazilians (n=62) from the Amazon Basin in Brazil and two additional population sets from Rio de Janeiro (n=93)and Tripoli, Libya {n=77). The results indicated that all markers are in Hardy-Weinberg equilibrium (HWE) except for one marker in Native Brazilians {marker 36, rs28923216); three markers for the Rio de Janeiro population (markers 20, rs2308189; 27, rs3045264; and 28, rs3047269). and one marker for the Libyans (marker 40, rs34510056) (Table **1**).This number of departures is no more than would be expected to occur by chance. In addition, there were no significant departures from HWE after the Bonferroni correction.(«1 0.05/49,p<0.001).

To measure population differentiation due to substructure, the three groups were analyzed for Wright's F_{ST} estimates. Even though these three populations are not expected to mix due to their distance apart and characteristics, when combined for analysis purposes only, the combined F_{ST} value of the populations is relatively low, F_{ST} =0.05512. However, this value was higher than that reported for the North American sample populations. Overall, the results reveal that the markers constitute a suitable system for HID purposes to be used with the two urban groups. However, the panel was less efficient for the isolated community of native Brazillians.

The power of an INDEL panel is related to the number of markers with a Random Match Probability (RMP) near or below 0.4 (considering the ideal value of p=q=0.5, RMP is 0.375). For all markers, the RMP varied from 0.3 to 0.73 in the populations tested (Table 1). Assuming loci independence and no substructure effect, cumulative RMPs were 2.7x10-18, 1.5x10-20 and 4.5x10-20 for Native Brazilian, Rio de Janeiro, and Tripoli populations, respectively. The number of INDEL loci above and below, an RMP of 0.4, was varied by population (Fig. 1).There were 23 markers above this threshold in Native Brazilians, and only 12 and 16 markers in the Rio de Janeiro and Tripoli samples, respectively.

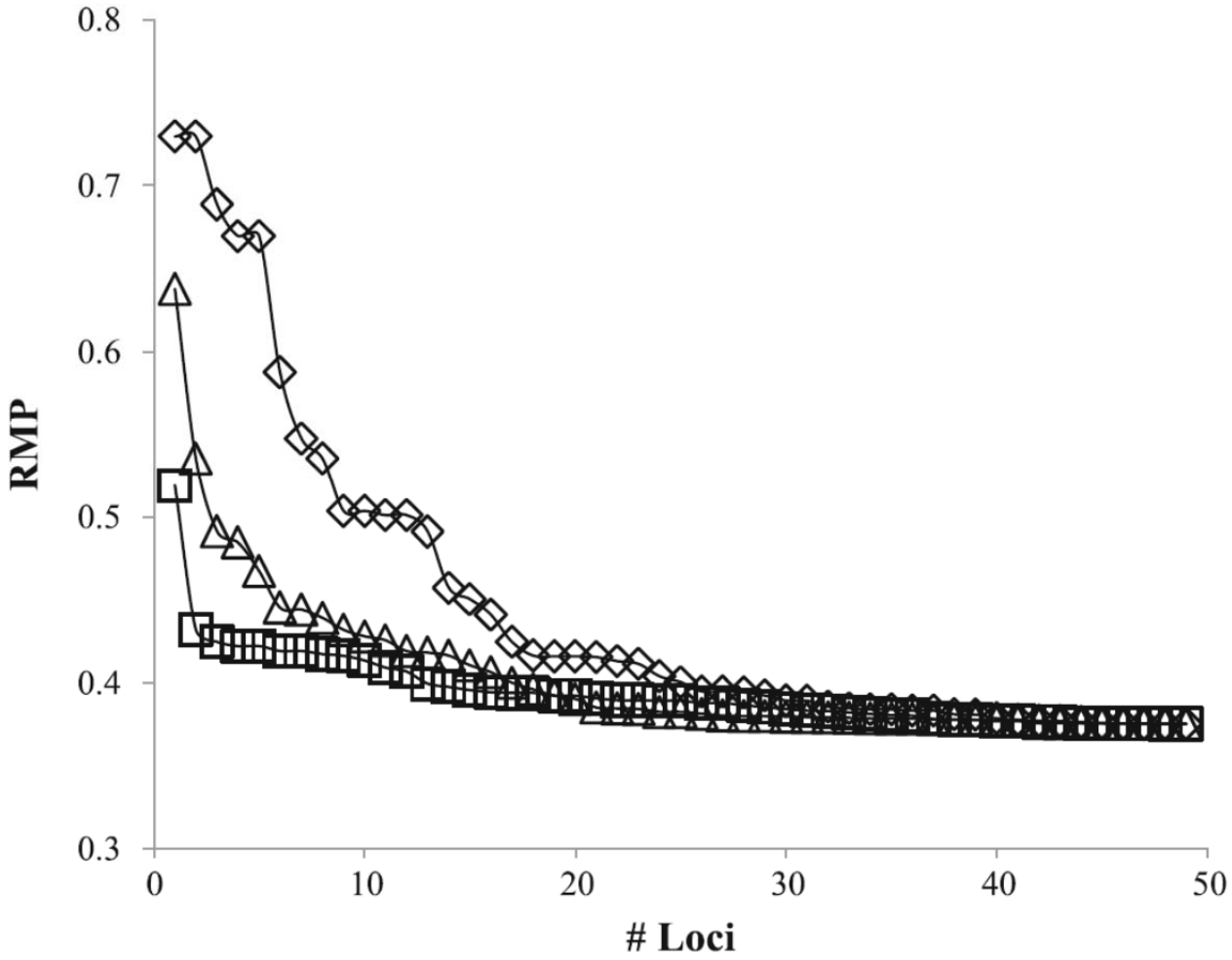


Figure 1. Random Match Probebili1y (RMP) plot against the Number of Loci (#Loci), in three population samples: Native Brazilian (white diamond); Rio de Janeiro (white square);and Tripoli (white triangle)

Linkage disequilibrium (LD) was determined using Fisher's exact test, with 10,000 shuffling. For the 49 INDEL markers, there were 1176 possible pairwise comparisons per population sample. A total of four pairs had a detectable LD at the 0.05 level, two from Rio de Janeiro and two from Native Brazilians. This proportion of detectable LD was lower than expected by chance. In addition, there were no significant departures from independence after the Bonferroni correction. The lack of detectable LD supports that the product rule can be applied.

Cumulative RMP values were 1.5xt0-20 and 4.Sx10-20 for the Rio de Janeiro and Tripoli populations, respectively. These values are one order of magnitude higher than those reported for the North American ethnic groups, and two orders of magnitude higher than the value for isolated Native Brazilians. This finding suggests that higher degrees of isolation leading to genetic bias and lower diversity may reduce the efficiency of the 49 INDEL panel. Similar reductions in efficiency of other INDEL panels have been reported elsewhere.
Population substructure was analyzed using the Structure software. The highest likelihood was achieved for K=3 (three parental populations, data not shown). The percentages of admixture within each of the populations are presented in Fig. 2. The Rio de Janeiro and Tripoli sample populations showed similar proportions of admixture levels for each of their respective parental clusters (30 %). However, for Native Brazilians, one of the parental clusters (CI) represented over 60 % ofthe total parental population.

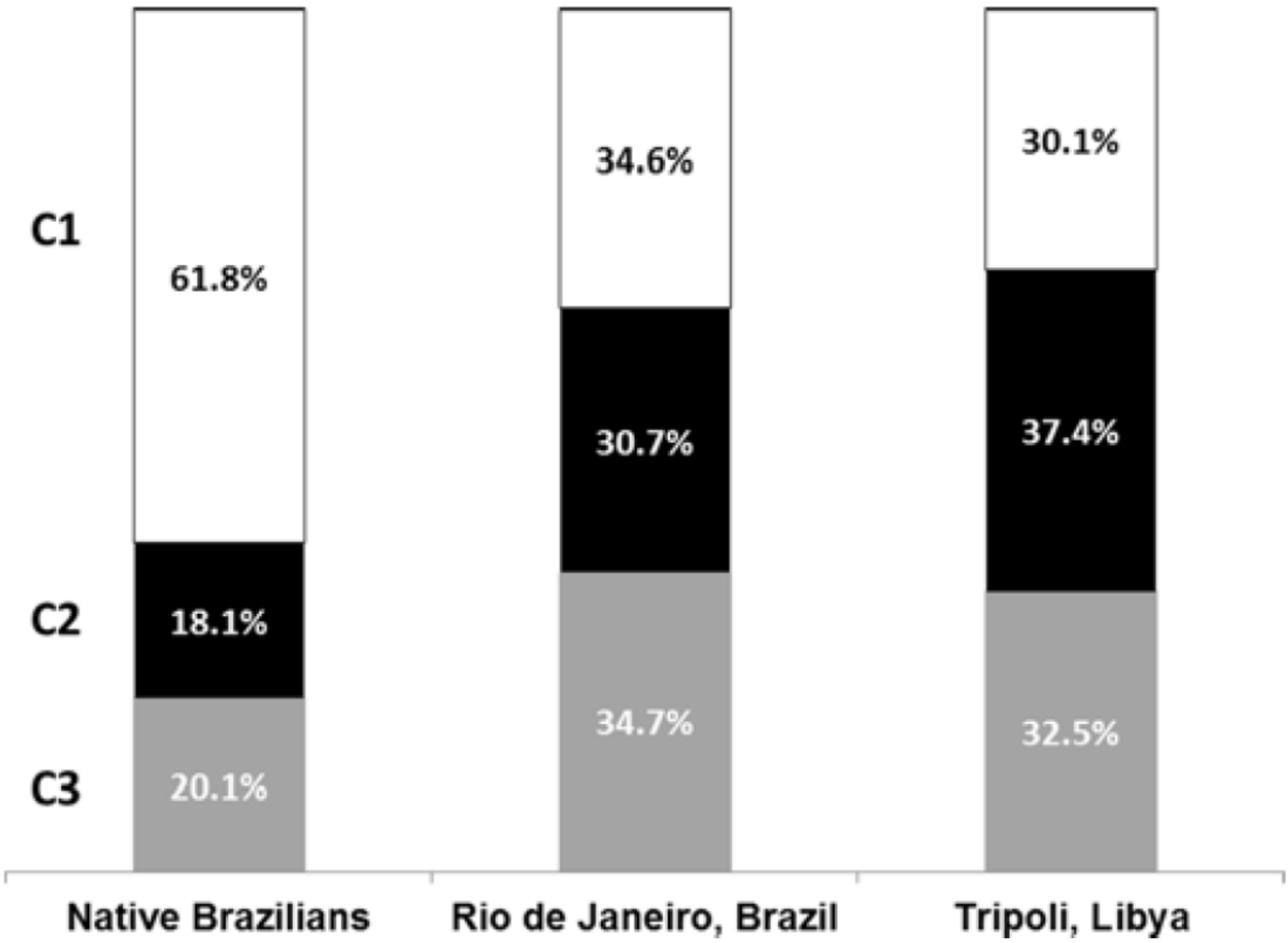


Figure 2. Schematic representation of the average substructure population. Using Structure v2.3, assuming K=3 and admixture model, three clusters (Cl, C2,. and C3) are represented for each population. The percentage of admixture is denoted inside the corresponding bars

This substructure cannot be explained by differences in sample size, because the other population samples (Rio de Janeiro, n=93 and Libya, n=77) had equivalent sizes and showed no bias towards any cluster. Substructure analysis reveals that admixed urban samples had equal contributions from three parental clusters. Native Brazilians had a major contributing cluster (over 60 %),indicating that some of these loci could behave as ancestry Amerindian markers.

In conclusion, the 49 INDEL marker panels could be used for HID and genetic studies in general, but caution should be exercised in the case of isolated, potentially substructured populations in which the RMP will not be as informative as in admixed populations.

Table 1. Description, location, and distribution of 49 INDEL markers in three populations

Marker number	RS number	Native Brazilians, Brazil (n=62)				Rio de Janeiro, Brazil (n=93)				Tripoli, Libya (n=77)				F_{ST}
		Freq (D)	H_O	HWE	RMP	Freq (D)	HO	HWE	RMP	Freq (D)	HO	HWE	RMP	
1	4187	0.3852	0.4754	1.0000	0.3892	0.5000	0.5435	0.5316	0.3750	0.4804	0.5098	1.0000	0.3754	0.0077
2	16,402	0.4262	0.3934	0.1235	0.3806	0.6141	0.4674	1.0000	0.3890	0.2192	0.3562	0.7368	0.4912	0.1586
3	16,458	0.5917	0.4500	0.5998	0.3838	0.4946	0.5109	1.0000	0.3750	0.5577	0.5385	0.5903	0.3784	0.0036
4	140,809	0.5574	0.4918	1.0000	0.3784	0.5924	0.4239	*02806*	0.3840	0.2945	0.3151	0.0489	0.4279	0.0986
5	1,160,886	0.3250	0.4500	1.0000	0.4113	0.3043	0.4565	0.6213	0.4221	0.4071	0.4714	0.8091	0.3841	0.0067
6	1,610,871	0.1750	0.2167	0.0678	0.5476	0.3352	0.4505	1.0000	0.4066	0.4933	0.6133	0.0669	0.3750	0.0919
7	2,067,140	0.4561	0.4211	0.2773	0.3769	0.3681	0.3846	0.1122	0.3942	0.4231	0.5000	1.0000	0.3811	0.0015
8	2,067,191	0.4083	0.4500	0.5964	0.3838	0.5163	0.5978	0.0932	0.3753	0.5704	0.4930	1.0000	0.3801	0.0170
9	2,307,507	0.3214	0.4643	0.7633	0.4130	0.5163	0.4674	0.5433	0.3753	0.3516	0.3906	0.2742	0.3999	0.0427
10	2,307,526	0.7000	0.4333	1.0000	0.4246	0.4022	0.5000	0.8322	0.3851	0.3630	0.4247	0.4468	0.3959	0.1097
11	2,307,579	0.3667	0.4333	0.5868	0.3947	0.3152	0.5217	0.0579	0.4161	0.4054	0.5135	0.6392	0.3844	0.0032
12	2,307,603	0.1083	0.2167	1.0000	0.6696	0.5435	0.5000	1.0000	0.3769	0.5486	0.4861	1.0000	0.3774	02108
13	2,307,656	0.5094	0.5660	0.4111	0.3751	0.4780	0.5165	0.8371	0.3755	0.5224	0.5075	1.0000	0.3755	--0.0049
Marker number	RS number	Native Brazilians, Brazil (n=62)				Rio de Janeiro, Brazil (n=93)				Tripoli, Libya (n=77)				F_{ST}

		Freq (D)	H_O	HWE	RMP	Freq (D)	HO	HWE	RMP	Freq (D)	HO	HWE	RMP	
14	2,307,696	0.5083	0.4167	0.2069	0.3751	0.5761	0.4565	0.5272	0.3810	0.3750	0.4722	1.0000	0.3921	0.0377
15	2,307,700	0.5351	0.5439	0.6023	0.3762	0.3913	0.4348	0.3938	0.3877	0.4342	0.5263	0.6359	0.3794	0.0134
16	2,307,710	0.2083	0.3500	1.0000	0.5035	0.3913	0.5652	0.1248	0.3877	0.3267	0.3867	0.2946	0.4105	0.0305
17	2,307,839	0.0833	0.1333	0.3270	0.7295	0.3696	0.4348	0.5077	0.3938	0.1233	0.1918	0.2837	0.6377	0.1417
18	2,307,850	0.1833	0.3333	0.6767	0.5356	0.1957	0.2826	0.3214	0.5191	0.2254	0.3099	0.3261	0.4846	--0.0042
19	2,308,112	0.3509	0.4912	0.7678	0.4002	0.3098	0.4891	02254	0.4190	0.5417	0.4444	0.3542	0.3768	0.0593
20	2,308,189	0.6140	0.4561	0.7855	0.3890	0.6467	0.3587	0.0385	0.3993	0.4167	0.4722	0.8140	0.3822	0.0577
21	2,308,196	0.5948	0.3276	0.0153	0.3845	0.3804	0.5000	0.6555	0.3905	0.7115	0.4231	1.0000	0.4318	0.1118
22	2,308,232	0.2105	0.3158	0.6931	0.5009	0.3098	0.3587	0.1451	0.4190	0.3359	0.5156	0.2734	0.4063	0.0113
23	2,308,276	0.4917	0.4167	0.1997	0.3751	0.4565	0.5217	0.6722	0.3769	0.4851	0.5224	0.8102	0.3752	--0.0055
24	2,308,292	0.4583	0.5500	0.6010	0.3768	02880	0.3804	0.4604	0.4320	0.1833	0.3333	0.6677	0.5356	0.0703
25	3,038,530	0.5877	0.5789	0.1839	0.3830	0.4130	0.4130	0.1895	0.3829	0.4437	0.6056	0.0903	0.3782	0.0246
26	3,042,783	0.6316	0.4912	0.7820	0.3941	0.5879	0.5385	0.3865	0.3831	0.7571	0.3143	0.2015	0.4674	0.0290
27	3,045,264	0.2193	0.3333	1.0000	0.4910	0.3641	0.3587	0.0421	0.3955	0.3116	0.3913	0.5801	0.4181	0.0172

Marker number	RS number	Native Brazilians, Brazil (n=62)	Rio de Janeiro, Brazil (n=93)	Tripoli, Libya (n=77)	F_{ST}

		Freq (D)	H_O	HWE	RMP	Freq (D)	HO	HWE	RMP	Freq (D)	HO	HWE	RMP	
28	3,047,269	0.4250	0.4500	o.6008	0.3808	0.4620	0.3587	0.0106	0.3765	0.5608	0.4730	0.8107	0.3788	0.0118
29	3,838,581	0.9000	0.2000	1.0000	0.6886	0.4457	0.5217	0.6725	0.3780	0.4375	0.4375	0.4554	0.3790	02351
30	3,841,948	0.1083	0.1833	0.5228	0.6696	0.3750	0.4457	0.6542	0.3921	0.4470	0.4697	0.8040	0.3779	0.1173
31	4,646,006	0.6250	0.4167	0.4147	0.3921	0.4185	0.5761	0.1309	0.3819	0.4769	0.4308	0.3155	0.3755	0.0369
32	10,623,496	0.3158	0.4561	0.7663	0.4158	0.3000	0.4222	1.0000	0.4246	0.4028	0.5556	0.2306	0.3850	0.0080
33	10,688,868	0.4500	0.4667	0.7935	0.3775	0.3132	0.4725	0.4745	0.4172	0.3151	0.4384	1.0000	0.4162	0.0174
34	13,447,508	0.3667	0.4000	0.2799	0.3947	0.3043	0.4348	1.0000	0.4221	0.2985	0.3284	0.0810	0.4255	--0.0011
35	17,859,968	0.5250	0.5167	1.0000	0.3756	0.3859	0.4457	0.6582	0.3890	0.4318	0.4697	0.8028	0.3798	0.0131
36	28,923,216	0.6583	0.3167	0.0228	0.4038	0.5489	0.4239	02108	0.3774	0.5833	0.4091	0.2109	0.3822	0.0057
37	33,951,431	0.8500	0.2000	0.1099	0.5875	0.6793	0.5109	0.1567	0.4134	0.5149	0.4627	0.6199	0.3752	0.1027
38	34,051,577	0.4000	0.5600	0.3682	0.3856	0.6703	0.3736	0.1502	0.4091	0.5571	0.3714	0.0516	0.3783	0.0622
39	34,495,360	0.3167	0.4333	1.0000	0.4154	0.5707	0.4891	1.0000	0.3801	0.7308	0.3846	1.0000	0.4453	0.1406
40	34,510,056	0.7917	0.3167	0.7031	0.5035	0.4837	0.4891	0.8420	0.3753	0.4318	0.6212	0.0449	0.3798	0.1246
41	34,511,541	0.7373	0.3220	0.1842	0.4503	0.3750	0.5109	0.5029	0.3921	0.4467	0.3867	0.0603	0.3779	0.1221
Marker number	RS number	Native Brazilians, Brazil (n=62)				Rio de Janeiro, Brazil (n=93)				Tripoli, Libya (n=77)				F_{ST}

		Freq (D)	H_O	HWE	RMP	Freq (D)	HO	HWE	RMP	Freq (D)	HO	HWE	RMP	
42	34,528,025	0.3158	0.3509	0.2212	0.4158	0.3846	0.4176	02715	0.3894	0.3099	0.3662	0.2666	0.4190	0.0016
43	34,535,242	0.7456	0.3684	0.7455	0.4572	0.6033	0.5543	0.1948	0.3863	0.7222	0.3611	0.3903	0.4390	0.0220
44	34,795,726	0.7895	0.3158	0.7021	0.5009	0.5380	0.5326	0.5425	0.3765	0.5643	0.4429	0.4778	0.3792	0.0640
45	34,811,743	0.9167	0.1333	0.3433	0.7295	0.6304	0.4565	0.8217	0.3938	0.7292	0.3750	0.7645	0.4441	0.0954
46	35,605,984	0.6852	0.3704	0.3438	0.4163	0.4286	0.5055	0.8295	0.3803	0.5423	0.4930	1.0000	0.3768	0.0560
47	35,716,687	0.7250	0.4167	1.0000	0.4410	0.5489	0.5761	0.1510	0.3774	0.5224	0.4776	0.8017	0.3755	0.0374
48	36,062,169	0.5167	0.5000	1.0000	0.3753	0.6413	0.5000	0.5012	0.3974	0.5000	0.5758	0.3235	0.3750	0.0201
49	60,901,515	0.5877	0.5789	0.1821	0.3830	0.6154	0.5714	0.0772	0.3894	0.6181	0.3750	0.0851	0.3901	--0.0060

Further development of Retrotransposable element markers

Portions of this manuscript are either submitted or in the process of being submitted to two separate articles in *FSI: Genetics*.

A sub-aim of our proposal was to work with corporate partners, Innogenomics, to further develop the Innotyper assay from a nine marker single-loci system to a multiplex system with high discriminatory power. Through pilot testing of various prototype versions of the assay we were able to obtain results from 19th century bones Figure 3, and 3 of 10 rootless hair shafts processed with the improved Wilson protocol (as described in NIJ final report from 2014) Figure 4. These samples prove challenging for STR based systems. This is likely due to the much smaller amplicon size of the Innotyper assay compared to STR genotyping amplicons which will not be as adversely affected by degradation of DNA.

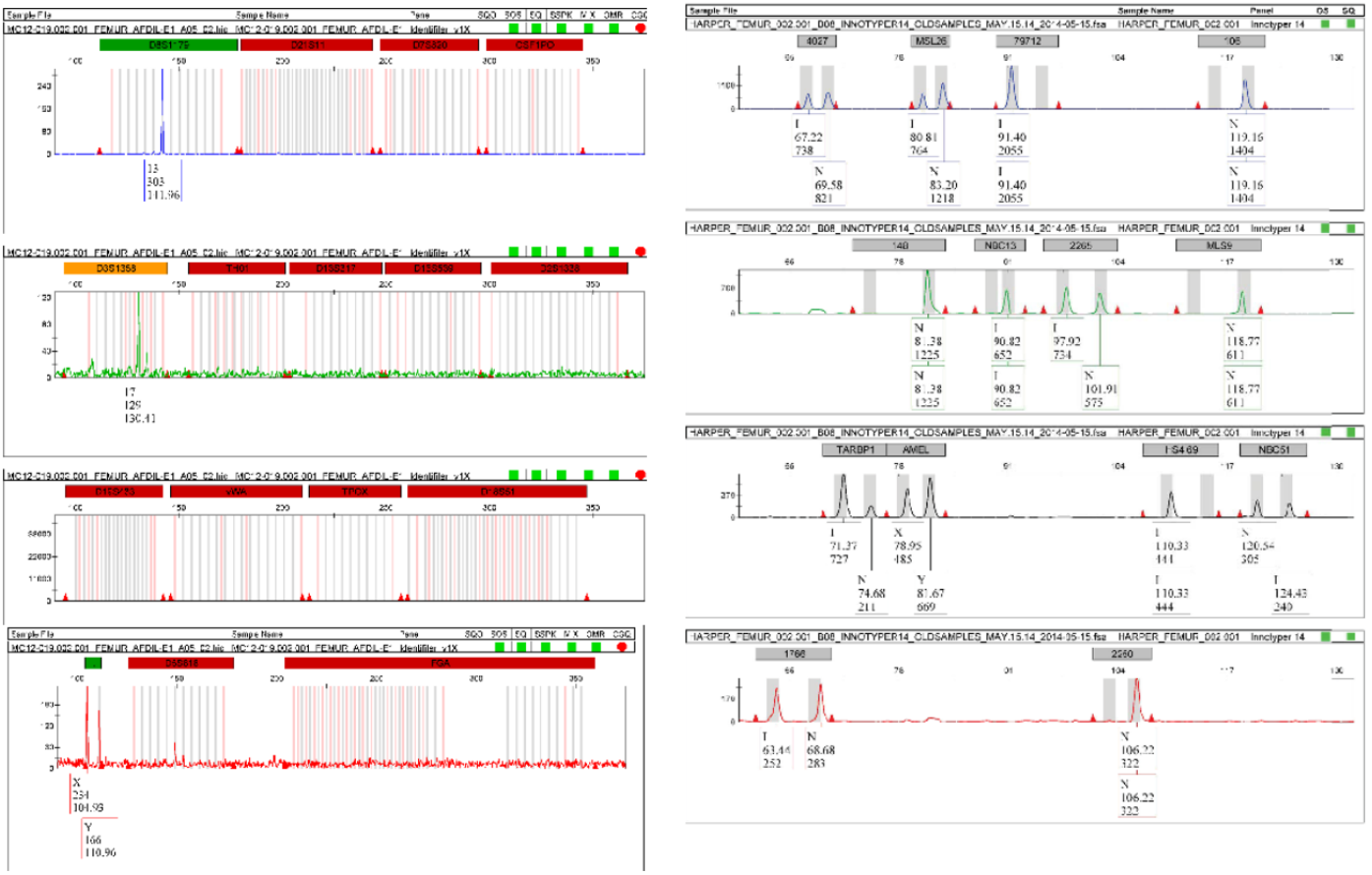


Figure 3. Identifiler profile of extract from 19th century bones and profile from same extract using Innotyper 14 prototype panel.

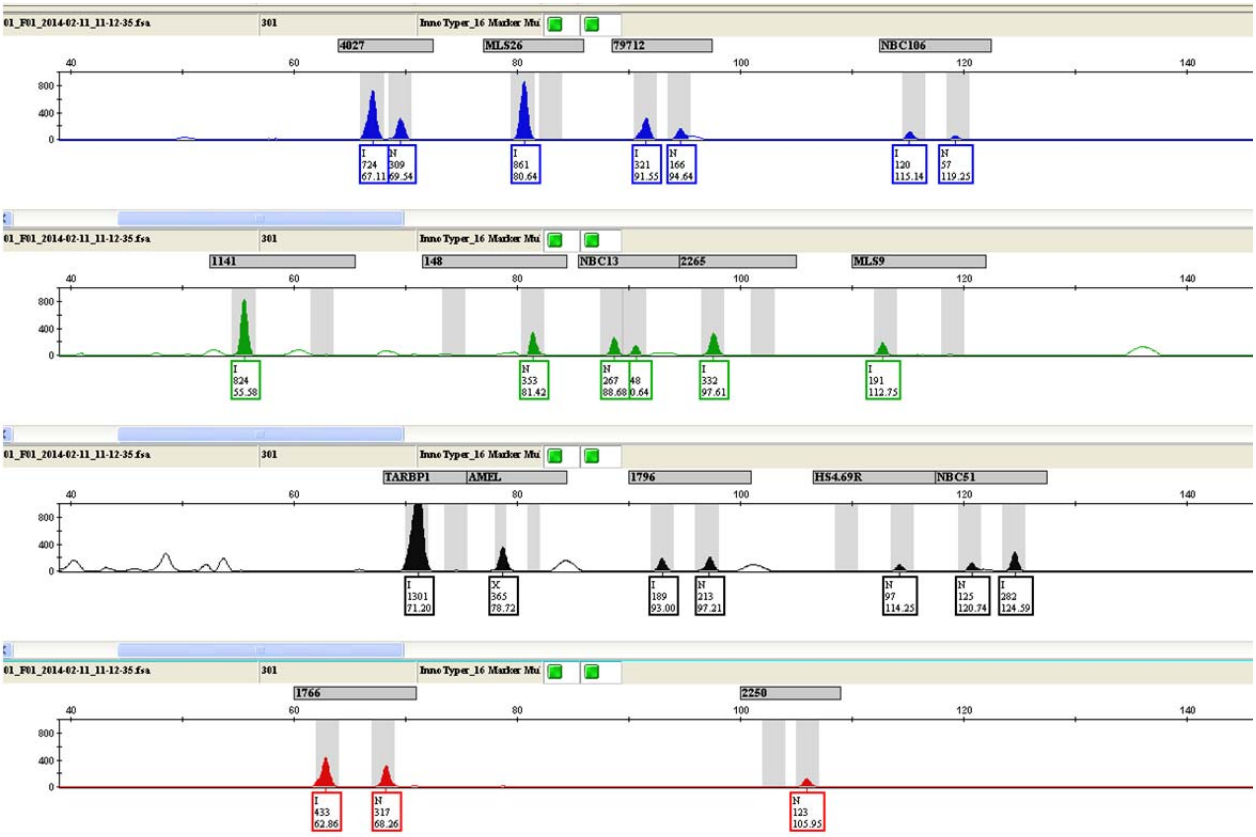


Figure 4. Representative electropherogram from pilot study of rootless 4cm hairshafts extracted utilizing the improved Wilson method and genotyped with Innotyper 16 prototype panel.

Through various iterative processes, Innogenomics has consulted with our research group for population testing, and chemistry "troubleshooting" first with a 14 marker system, then with a 16 marker system, and finally to Innotyper 21 (a 20 INNUL marker system with an amelogenin marker).

The distribution of alleles in four major populations, African American (*N=207*), Southwest Hispanic (*N=40*), Caucasian (*N=301*), East Asian (US) (*N=44*) were analyzed for Hardy Weinberg Equilibrium (HWE), Random Match Probability (RMP), Power of Exclusion (PE), and evidence of Substructure. In the four populations tested, there was no deviatin from HWE after a Bonferroni correction for multiple comparisons was used to correctly adjust the critical value for an alpha level of 0.05, and the combined random match probabilities were $2.47x\ 10^{-08}$, $3.52x10^{-08}$, $1.37x10^{-08}$, $2.64x10^{-07}$ for African American, Southwest Hispanic, Caucasian, and East Asian populations respectively. These results are summarized in Table 2. Additionally, no marker pairs demonstrated a departure from linkage disequilibrium after a Bonferroni correction for multiple comparisons appropriately lowered the critical value to an alpha level of 0.05 in any of the populations.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Table 2

Population specific Innotyper 21® allele frequencies and population parameters across 4 major global populations.

	African American *(N=207)*					Southwest Hispanic *(N=40)*					Caucasian *(N=301)*					East Asian (US) *(N=44)*					
Allele	**Freq I**	**Ho**	**HWE (p-value)**	**RMP**	**PE**	**Freq I**	**Ho**	**HWE (p-value)**	**RMP**	**PE**	**Freq I**	**Ho**	**HWE (p-value)**	**RMP**	**PE**	**Freq I**	**Ho**	**HWE (p-value)**	**RMP**	**PE**	**F_{ST}**
AC4027	0.53865	0.53623	0.26906	0.37650	0.18675	0.65000	0.35000	0.16031	0.40050	0.17574	0.42857	0.49834	0.81469	0.38030	0.18492	0.47727	0.50000	1.00000	0.37603	0.18724	0.01439
NBC216	0.59662	0.50725	0.17844	0.38490	0.18275	0.53750	0.42500	1.00000	0.37640	0.18679	0.73090	0.35216	0.20531	0.44540	0.15800	0.27027	0.43243	0.68410	0.45362	0.15833	0.09887
MLS26	0.14976	0.23188	0.86844	0.58790	0.11114	0.51250	0.52500	0.53656	0.37520	0.18742	0.34718	0.48837	0.56000	0.40160	0.17528	0.40698	0.58140	0.34150	0.44294	0.18310	0.11799
ALU79712	0.30918	0.43478	0.47125	0.41940	0.16797	0.50000	0.45000	0.36031	0.37500	0.18750	0.48007	0.48173	0.06813	0.37540	0.18730	0.03660	0.07317	1.00000	0.86437	0.03402	0.02807
NBC106	0.57488	0.43478	0.11906	0.38080	0.18466	0.37500	0.65000	0.02531	0.39210	0.17944	0.44020	0.52159	0.34375	0.37870	0.18570	0.51136	0.56818	0.54270	0.41632	0.18744	0.08016
RG148	0.53623	0.50242	1.00000	0.37630	0.18684	0.36250	0.47500	1.00000	0.39610	0.17769	0.30233	0.43854	0.57563	0.42320	0.16644	0.70455	0.31818	0.14520	0.41736	0.16483	0.03994
NBC13	0.21981	0.33333	0.67656	0.49050	0.14208	0.36250	0.47500	1.00000	0.39610	0.17769	0.35714	0.47508	0.59531	0.39790	0.17688	0.14290	0.28571	0.58280	0.59184	0.10748	-0.00198
AC2265	0.39130	0.46377	0.77188	0.38770	0.18145	0.76250	0.27500	0.17750	0.47240	0.14830	0.73754	0.35880	0.22750	0.45050	0.15610	0.83720	0.32558	0.57530	0.56084	0.11772	0.11967
MLS09	0.23671	0.36715	1.00000	0.47320	0.14803	0.38750	0.47500	1.00000	0.38860	0.18101	0.40532	0.47176	0.72469	0.38440	0.18294	0.82550	0.25581	0.58310	0.55435	0.12330	0.03192
AC1141	0.22947	0.34300	0.68969	0.48030	0.14556	0.71250	0.37500	0.68813	0.43240	0.16288	0.60797	0.52492	0.09125	0.38750	0.18153	0.60465	0.46512	1.00000	0.38129	0.18190	0.16113
TARBP	0.28502	0.37681	0.30781	0.43400	0.16225	0.36250	0.42500	0.73250	0.39610	0.17769	0.57475	0.50498	0.62719	0.38080	0.18467	0.35227	0.47727	1.00000	0.40806	0.17611	0.01745
AC2305	0.30676	0.42995	1.00000	0.42070	0.16744	0.66250	0.37500	0.32063	0.40560	0.17360	0.57807	0.47176	0.56688	0.38130	0.18442	0.85710	0.28571	0.57250	0.59184	0.10748	0.09161
HS4.69	0.31884	0.41546	0.52938	0.41430	0.17001	0.20000	0.30000	0.64031	0.51360	0.13440	0.38704	0.43522	0.14438	0.38870	0.18096	0.68605	0.48837	0.49290	0.43862	0.16900	0.05282
NBC51	0.59420	0.45411	0.38625	0.38430	0.18298	0.53750	0.57500	0.52438	0.37640	0.18679	0.52658	0.48837	0.73500	0.37570	0.18715	0.63953	0.48837	1.00000	0.40833	0.17739	0.08278
NBC102	0.39614	0.45411	0.73250	0.38650	0.18199	0.58750	0.42500	1.00000	0.38300	0.18361	0.40698	0.50166	0.62781	0.38410	0.18310	0.20450	0.31818	1.00000	0.50826	0.13622	0.01929
NBC120	0.59662	0.46860	0.76969	0.38490	0.18275	0.53750	0.37500	0.12281	0.37640	0.18679	0.41030	0.44186	0.15844	0.38340	0.18341	0.39286	0.45238	0.75490	0.37755	0.18163	0.00037
NBC10	0.65942	0.44928	1.00000	0.40430	0.17415	0.48750	0.52500	1.00000	0.37520	0.18742	0.43189	0.51163	0.48531	0.37980	0.18516	0.25581	0.37209	1.00000	0.45484	0.15413	0.03307
ACA1766	0.72222	0.39130	0.47844	0.43900	0.16038	0.80000	0.35000	0.51344	0.51360	0.13440	0.62791	0.45183	0.55438	0.39300	0.17905	0.78210	0.33333	1.00000	0.49244	0.14138	0.03241
SB19.12	0.39614	0.48309	1.00000	0.38650	0.18199	0.17500	0.30000	1.00000	0.54760	0.12353	0.30399	0.40199	0.42969	0.42230	0.16681	0.41463	0.43902	0.53320	0.36466	0.18380	0.09632
NBC148	0.54348	0.51691	0.57969	0.37690	0.18655	0.91250	0.12500	0.25438	0.71890	0.07347	0.87542	0.21595	0.79813	0.63510	0.09717	0.22090	0.34884	1.00000	0.48945	0.14248	0.20125
Overall				**2.4743E-08**	**0.9757**				**3.5180E-08**	**0.9740**				**1.3660E-08**	**0.9784**				**2.6417E-07**	**0.9626**	**0.0669**

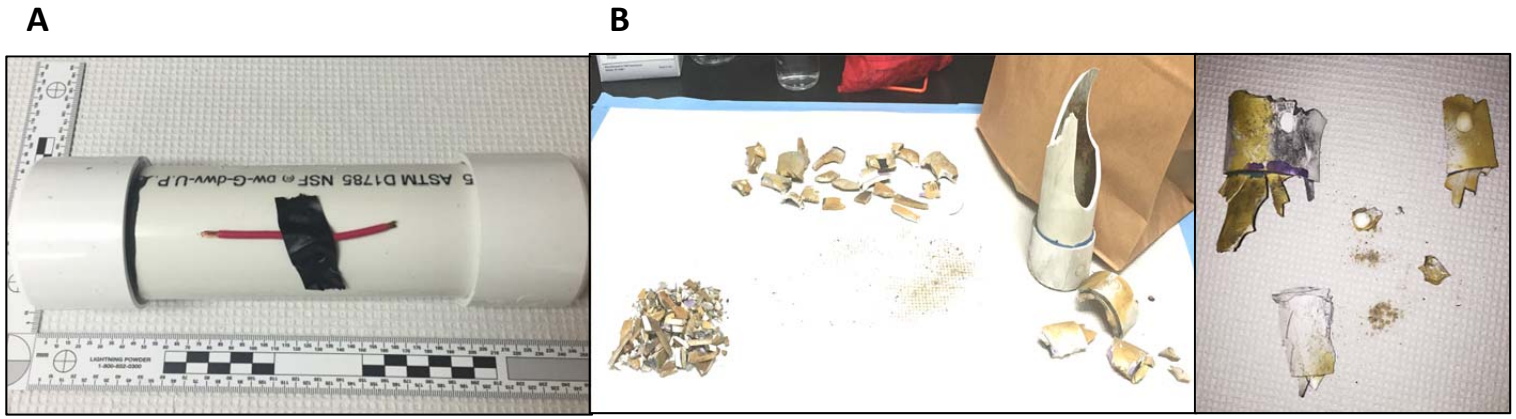


Figure 5A). Constructed pipe-bomb with wire attached. B) Post blast fragmentation of PVC device.

One of the topics mentioned in our proposal involved the use of bi-allelic markers to genotype degraded DNA from the post-balst fragments of Improvised Explosive devices . We collaborated with colleagues from Sam Houston State University, and the Montgomery County Fire Marshall's office to type IED post-blast fragments (Figure 5) with INNUL markers.

As mentioned earlier, INNULs have been proposed as an alternate marker system that has the potential to recover additional genetic information from challenging samples when traditional STR analysis fails as amplicons are smaller than STRs (up to 450 bp compared to 125 bp). Twenty-five DNA extracts which resulted in varying STR success (ranging from 2% to 87% reported alleles) were chosen for INNUL analysis.

Genotyping using INNUL markers resulted in a high degree of success with nine of the 25 samples resulting in complete genetic profiles, and eight samples with between 80% and 99% of alleles being reported. Only six samples resulted in < 70% of reported alleles being reported, and one sample completely failed INNUL analysis (Fig. 6).

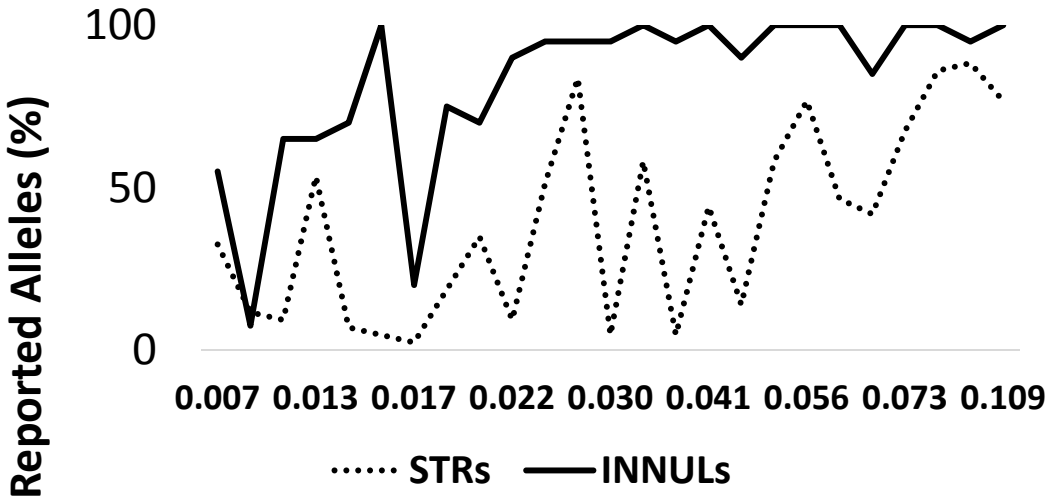


Figure 6 . Number of alleles reported (as a percentage) from 25 samples genotyped using STRs and INNULs ranked by increasing DNA quantity in PCR (ng). 15 µL and 16 µL of neat DNA was amplified in for STR and INNUL analysis, respectively.

DNA samples recovered from post-blast pipe bomb fragments showed significantly more complete genetic profiles when amplified with the INNUL kit than with the STR kit ($p < 0.001$) (Fig. 6). Six samples that had previously performed poorly or failed STR analysis (< 70% of STR alleles reported) resulted in complete INNUL profiles. Compared to STRs, INNULs resulted in more complete profiles for all but one sample which failed INNUL analysis (Fig. 6), suggesting that in addition to being able to amplify shorter DNA targets,

INNULs may also be more sensitive than STR panels and achieve greater genotyping success with lower quantities of DNA. In this study, a complete profile was produced from as little as 39 pg when STR typing of the same sample produced 7% alleles.

Genotyping Damaged samples with prototype INDEL panel

Portions of this work are being submitted to a manuscript in FSI:Genetics

The STR typing success of five embalmed tissue samples using the GlobalFiler® Amplification Kit was evaluated using the number of correct alleles detected (concordant to reference samples) and RMP values calculated from each STR profile. This was part of a larger study to observe various methods to improve STR genotype generation from these samples and the samples were geneotyped utilizing a prototype version of our HID INDEL panel in addition to a SNP panel via massively parallel sequencing (MPS).

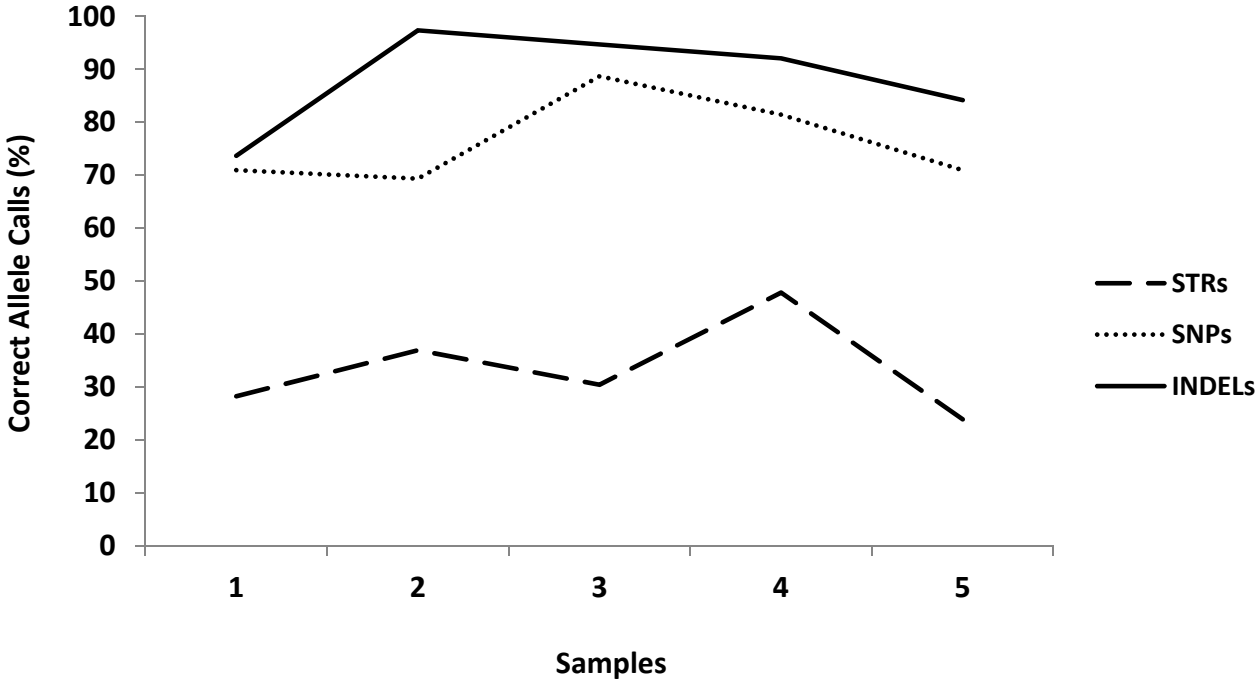


Figure 7. Genotyping Success for the five FD samples used in this study when non-treated samples were amplified using the GlobalFiler® STR Kit, the INDEL multiplex, and the HID-Ion AmpliSeq™ Identity Panel. Sample 1= Jejunum, 2 = Jejunum, 3 = Kidney, 4 = Stomach, 5 = Spleen

The number of alleles recovered from the untreated samples ranged from 23.9 % to 47.8 % (average of 33.5 % alleles) resulting in an average RMP value for the untreated samples of 2.03x10-7. All untreated DNA samples were also amplified using a multiplex INDEL system. The INDEL multiplex also contains markers substantially smaller (<200 bp) than the larger STR markers. Therefore, it was also expected that the INDEL multiplex would generate a greater number of correct allele calls than the GlobalFiler® STR kit.

Three loci (2032678, 28362545, and Amelogenin) contained with the INDEL multiplex found to violate equilibrium assumptions were excluded from the calculations (resulting in 39 loci used in this study). As this was a prototype multiplex it was assumed that the departure from HWE was associated with an unresolved chemistry issue.

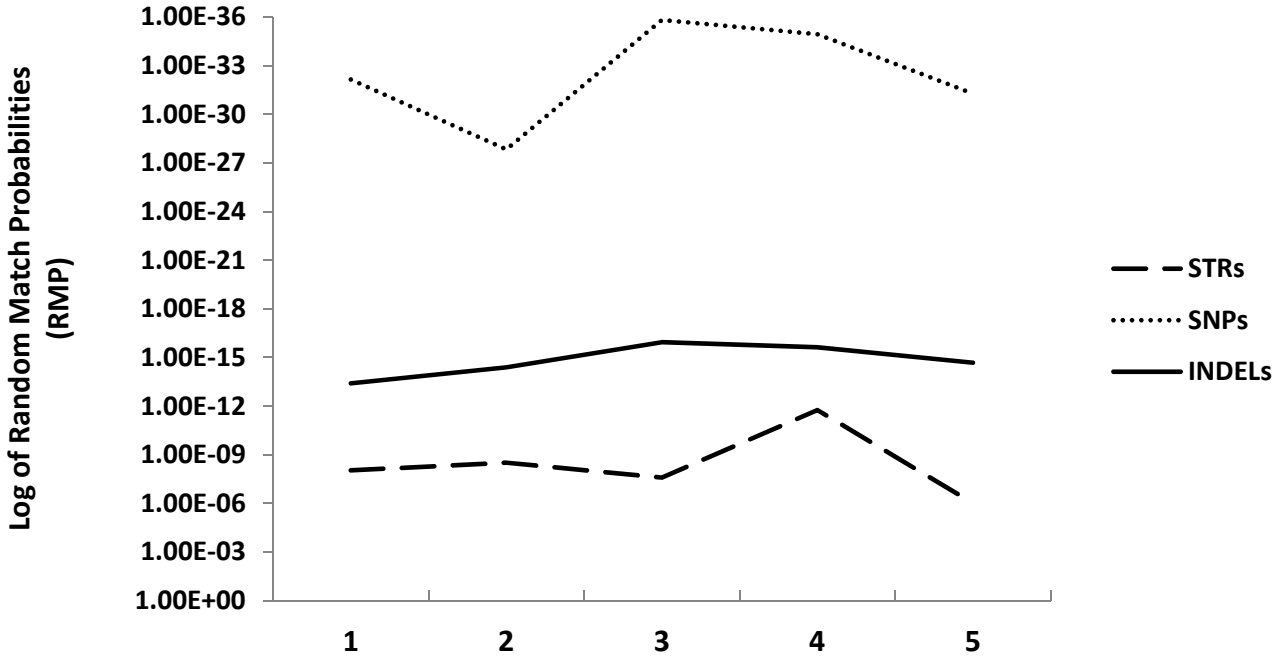


Figure 8. RMP values generated for the five FD samples used in this study when non-treated samples were amplified using the GlobalFiler® STR Kit (no treatment), the INDEL multiplex and the HID-Ion AmpliSeq™ Identity Panel. Sample 1= Jejunum, 2 = Jejunum, 3 = Kidney, 4 = Stomach, 5 = Spleen

The remaining INDEL panel resulted in a greater percentage of correct alleles being reported compared to the STRs and SNP panels for all five samples (Fig. 7).
Population specific allele frequencies were used to calculate RMP values for each sample genotyped using all three marker types. The RMPs generated using the INDEL system were notably lower than those calculated from the STR profiles generated using the GlobalFiler® STR kit for all five embalmed samples in this study (Fig. 8). Although a larger number of INDEL markers are needed to match the RMP values of current STR kits, they are more likely to be amplified in a highly degraded sample due to their smaller amplicon size. Therefore, it was not surprising that INDELs had a significantly higher percentage of allele calls for all five embalmed samples compared to the original samples amplified with GlobalFiler® (Fig. 7) ($p<0.001$).

Obviously when compared to the SNP panel that was analyzed via MPS, the INDEL panel has RMPs that are many orders of magnitude less, but the cost of a generating genotypes via MPS greatly outstrips the capillary electrophoresis based INDEL genotyping.

Additionally, the RMP of the INDEL panel approaches the RMP of STR marker systems prior to the expansion of the CODIS loci.

Finalizing the HID INDEL multiplex

Portions of this section will be submitted to the journal Electrophoresis

Aim two of the proposal describes the design of a singletube multiplex PCR reaction that is capable of genotyping DNA samples utilizing a refined INDEL panel. In order to do so, primers needed to be designed in silico, then tested for any unforeseen incompatabilities in the multiplex. Due to the expensive nature of fluorescently labeled primers, it was decided that putative multiplexes would be checked for compatability via unlabeled primers via microfluidic electrophoresis prior to ordering the labeled primers.

Primer Design

The FASTA sequence for each marker was found by using DbSNP, a free online program. For some of the markers, the DbSNP website either did not provide a FASTA sequence or the sequence available was too small for forward and reverse primers to be designed. When this limitation was the case, the UCSC genome browser was used to obtain the FASTA sequence. The FASTA sequence then could be entered into Primer Blast to design the primer pairs for each marker *in silica*. Primers were designed to amplify the INDEL markers by using the Primer-BLAST software. The primers were designed to generate PCR product sizes between 60 and 200 bp. For one marker, RS2307579, the PCR product size was 219 bp; all others were in the intended range. The results of the primer design from Primer-BLAST are listed in Table 3, with the primer pairs numbered in numerical order, with primer pair 1 corresponding to RS 4646006. After primers were selected for each marker, they were tested for possible dimerization using the PriDimerCheck option from the MPprimer website. This software compares all the primers against each other and provides the alignment, the matches, the 3ˈ-3ˈ dimerization, and the ΔG.

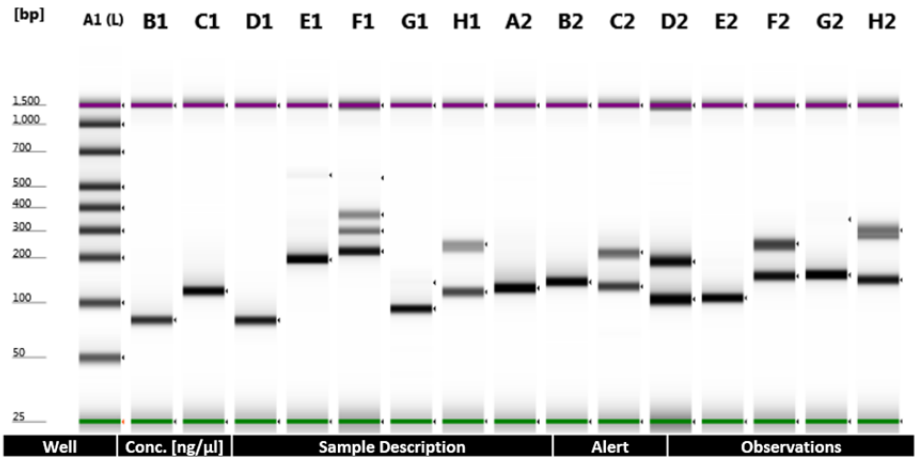
Unlabeled Primers

Since no major problems were observed when checking for dimerization, unlabeled primers were ordered. After amplification, the amplicons of the primers were run on the Agilent TapeStation. The TapeStation is a simple to use, automated electrophoresis platform that enables analysis of DNA fragments between 35 and 1000 bp in length. The TapeStation utilizes a ScreenTape gel matrix, similar to that of an agarose gel, to separate samples by molecular weight.

Table 3. Using Primer-BLAST, primers were designed for all 49 INDELs which are represented by their RS number. The estimated product length for each primer pair is also given along with the melting temperatures for the forward and reverse primers. The primer pairs are numbered in numerical order, with primer pair 1 being RS4646006.

RS Number	Forwar Primer (5'-3')	Reverse Primer (5'-3')	Product Length	TM Forward	TM Reverse
4646006	GCTGGGAAATGGGGAGACAA	GCCCGCTGTTTGGAAAGAAA	83	59.96	59.61
13447508	ATGGTTCAGTGGAAATAGCATGA	CATGTGGTCCAATCCCCTCA	118	58.14	59.37
3047269	TCATTCCCATGCTGGGTGAG	TGCACTGTACTTGCATGCTG	83	59.74	59.12
2307507	TGAAGGTGGGGCTATTGAGAAA	TTTCTCTTAGTTTGCATAAAACCCT	181	59.35	57.17
2307579	TTGTGACTGTGTCTCAGCAGTTAT	CACTGACTTGACTGAACCTTTCAAC	219	60.2	60.45
3838581	AGCATATGGAGAATGATTACTGGTG	TGCTCAAGATTTGTATGAGGAAGT	96	58.76	58.19
2308276	CTGAGAGACAATGGGATTTGCC	TTGCATGGAATTTCTCCATTTGA	120	59.31	57.26
3042783	TTCCCTGAGCTTACCGGAGTT	GATATTGACCTGAAGGCACACTG	132	60.83	59.38
3841948	AAACTACATGGCCCACAAGT	ATCCCATGGCACATTCCAGT	138	57.32	59.37
35716687	GGATGCAGTAGAGGCAGGTT	CATGCCATCATTAGGGGACT	127	59.46	57.03
2307603	AGTGTGCCTACAGATACCACTT	ACAGTCTTCATAGAACTATCTCACA	110	58.83	57.05
60901515	TGTGGATACCAAGCACTCCTG	TGCTGGTTCCAACCGGAAG	113	59.72	60.23
2308292	ACTCTGTCTCCACTGGGAATGT	CTATCTGTTAGGCGCACTGTGTCA	152	60.76	62.69
2307526	TGTTGGAGCCACATCAATGAC	GAGAAAGATCAAATTAATGCCAGGA	159	58.84	57.66
2307656	TCTGTGGGCAGAAGGCAAG	ACCAGGTTTGAAAAATGACATGCTA	146	59.93	59.7
2308196	AGCCTGTAAAAATTCCCCTCTTGT	AACGAACATCTTTTTCCACCACA	151	60.45	59.3
2067140	ACCCACCAAATGTCCCTGAC	TAGCTCACCTTGCAGTGCTC	67	59.89	60.04
2067191	ATTTCAGGGTAATCGGATTCTGTA	TGGCCTGTTTATCTTCTAAAGGG	141	57.51	58.14
1610871	TCACCTTCCTCCCAGTAACCA	TCCATTTCCCCTGCTACTCC	62	60.13	58.79
2307710	GCCCATACCTACTGTGACCA	AGGCTTGTCTACAAAATGAATGAA	79	58.8	57.1
2307839	TGCATGTAGGACAAGAGGTAGTT	GGTCTTGCAAAATTAATCACACTC	149	59.16	57.07
34510056	TAGATCCCGGCCCAAAGTCA	CGGTGGAATGCAAAACGACT	114	60.62	59.41
16458	AGCTCCCCAAAGACATGGTT	TGTAAGACTCAGAAGTTATAGGGCA	144	59.22	59.04
34535242	CTACAGACAGGTTTAAAATGAGCAA	ATTTACATAAGCCTCCTTCTGTGG	133	57.8	58.56
10623496	TCAGAGCAGGCTTATCTTAAACA	CTTGCTAAGACAGAAAGAAGAAACA	97	57.58	57.75
33951431	ACAAAGCCTCGGCGATAGAC	ACTCACAGCATGTGGGAGAAC	79	60.18	60.27
16402	ATGCGCCTTTTTGGTTTTGGT	GCATCAGGACTGTATGGGGC	106	60.13	60.54
2308112	CAGAAGAGGCGGTGCTGATG	TCTGGAGGGACCCAAGGTAT	72	61.09	59.28
2307850	TCACCGTTTCCTCCGCACT	GCCCAACCTGCGTGGAAAG	60	61.5	61.92
140809	AGGCTTTCAGATGTTCTTAGCC	CTCCTGAGTGACCACAGCG	87	58.38	60.08
1160886	TTCCCATTGTGCTTAAAACTCCT	CCAGTCTACCCAAATGTATTCCA	74	58.52	57.89
34051577	GTCATCCAGATTATCGAGTGAGA	GCTGCACTTTAGTCTTCCTGA	137	57.3	57.95
10688868	TTCCATCCCTCCTCTTGCCT	CGCTCTGCACATGCGTAAAA	80	60.25	59.83
34811743	ACACTTCGTACCCAGGATGC	GCCTCTCCTTTTTGTTCAACCC	69	59.75	59.96
2307696	CACTGACAGCAATCAGAACAC	CTGAGCCCATCTGACTGCTC	116	57.47	60.18
34528025	GTCTTGGAGAGGAGTCAAATCAGA	CTGGAACTCAAAGCAAAACGAG	75	59.78	58.38
3045264	CTTACCTACGTGGTTGGTGAC	GTACACGAGTAGCCGATGGA	94	58.32	58.98
2308232	GATTGATGCAATCTCACACTACC	CTTCCTATTCTCCTTGCTTCGT	95	57.58	57.87
4187	TCAAATAAGAGTTGTCATATCCTGC	TGGCAGTGAAGAGAACAGGTC	119	57.21	59.93
3038530	GGCAATGAATTCCTCCATATCAAAA	TCTGCAGAAATCGCTTTGTAAAT	80	58.18	57.37
2308189	ACAAAGGAACGACAAGAACAAAA	TGGAACCTGATTCATGCTGCT	78	57.62	59.99
34795726	GGAGAAAACATGGATAGGTAGCAA	TCTCTTCACTAACAGGATGAAGTAT	114	58.56	57.04
17859968	GCTGTCTTAAAAGATTGTGGGG	GGGTCCTACTAAATGCCATGTG	68	57.56	58.79
28923216	GTGAATTGATCACTTTGTTTCTTGC	GCCATTAGCTCAGATTCTCAGGA	128	58.13	59.93
36062169	ACTATTCTACTGCCATTTACCACA	AGAGGATATCTCAGAAGGATGGACT	74	57.69	59.92
34511541	TGGAGGACTTTAGTAGAAGAGGA	ACCCTTCTTAGGTTCAAAGACACT	70	57.47	59.59
34495360	TGTGGTTTGGTCTCAGTACTTGTTC	TTTCTAAGCTGAGTGGCAAGATG	74	61.14	58.99
35605984	ATAGTTTTCCTGCATTATCCCCAT	GCACAAAGAAGCTTATGTCATAGTA	146	58.03	57.46
2307700	CTGGCAGGGCCAGAGC	TCCTTCCTCGGAATCCCCAT	76	59.71	60.03

This analysis was performed to assess whether all the primers worked as intended. Figure 9 shows the results of the first 15 primer pairs and an allelic ladder along with the resulting bp sizes. The allelic ladder is a defined set of fragments of known sizes that is used to help determine the size of the unknown samples. All 49 primer pair amplicons were able to be amplified and produced results similar to those seen in Figure 5.

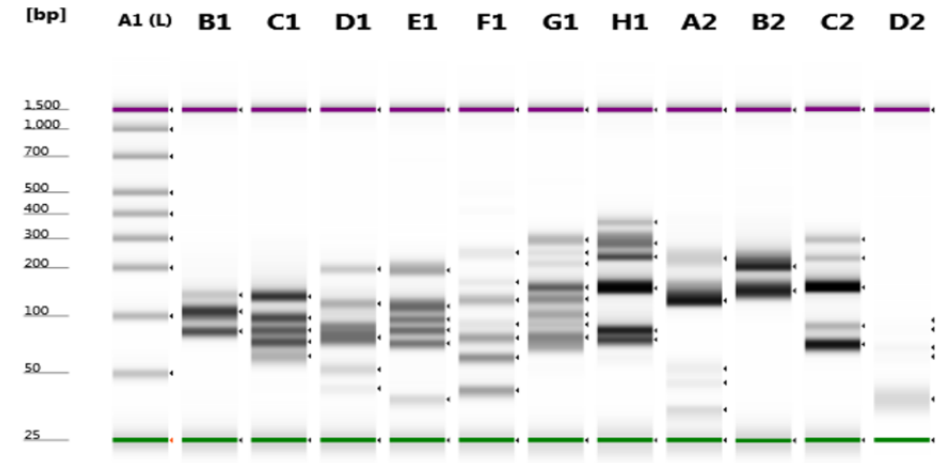


Well	Conc. [ng/µl]	Sample Description	Alert	Observations
A1	71.4	Ladder		Ladder
B1	11.4	1		Size bp: 84
C1	18.1	2		Size bp: 126
D1	13.6	3		Size bp: 84
E1	24.2	4		Size bp: 196
F1	19.9	5		Size bp: 224
G1	17.1	6		Size bp: 95
H1	16.3	7		Size bp: 123
A2	26.8	8		Size bp: 133
B2	22.1	9		Size bp: 146
C2	18.6	10		Size bp: 136
D2	31.1	11		Size bp: 108
E2	18.8	12		Size bp: 111
F2	28.9	13		Size bp: 159
G2	22.7	14		Size bp: 162
H2	20.5	15		Size bp: 150

Figure 9. The TapeStation gel image and summary table from the first 15 primer pairs and an allelic ladder. The table shows the base pair sizes for the respective primer pairs.

By running each of the primer pairs individually, the primers could be examined to ensure they were able to amplify the DNA and produce results. Since the TapeStation provided the approximate size for the amplicons, the primer pairs also were evaluated to determine if their bp size was approximately what was estimated during the design process. A review of the results given from the TapeStation showed that the bp sizes of the amplicons were as designed. After the primer pairs were run individually, they were arranged in groups of 5 primer pairs for a general multiplex design. This multiplex then was run on the TapeStation to assess if the primers could be amplified together and still produce intended results. Figure 10shows the results produced by the TapeStation of the multiplex. Although not every multiplex group produced 5 distinct bands, most banding that was observed was within the expected bp range for the primer pairs. Some of the primer pairs

in the multiplex groups may have been too close in bp length to produce a distinct separation by the TapeStation.



Well	Conc. [ng/µl]	Sample Description	Alert	Observations
A1	34.4	Ladder		Ladder
B1	47.5	1		Size bp: 87, 109, 144
C1	76.6	2		Size bp: 65, 78, 88, 98, 140
D1	44.0	3		Size bp: 44, 53, 81, 126, 197
E1	62.8	4		Size bp: 76, 88, 97, 120, 195
F1	34.6	5		Size bp: 64, 81, 93, 133, 170
G1	65.9	6		Size bp: 81, 93, 103, 135, 159
H1	117	7		Size bp: 79, 87, 158, 237
A2	56.5	8		Size bp: 132
B2	68.5	9		Size bp: 153, 205
C2	98.0	10		Size bp: 75, 91, 159
D2	8.25	-		Water

Figure 10. The TapeStation results from the general multiplex groupings. Group 1 contained primer pairs 25, 27, 49, 40, and 1; group 2 = 3, 6, 8, 19, and 28; group 3 = 26, 33, 30, 35, and 47; group 4 = 11, 22, 20, 34, and 37; group 5 = 29, 7, 14, 12, and 31; group 6 = 38, 36, 39, 43, and 15; group 7 = 18, 41, 46, 16, and 5; group 8 = 42, 13, 2, 9, and 44; group 9 = 10, 24, 4, and 32; group 10 = 48, 17, 21, 23, and 45.

With multiple bands being observed within the expected bp ranges for the primer pairs in the multiplex groups, these results showed that the primer pairs were working, were able to be multiplexed together and still produce results. Since no major issues were found, fluorescently labeled forward primers were ordered for each of the primer pairs.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Table 4. Fluorophores by which each amplicon was labeled. The highlighted primer pair amplicons were ordered as alternates with those respective fluorophore labels because they were too close in base pair length to the other amplicons.

Dye Channel	Primer Pair	Base pair length	INDEL length
Blue= 6-FAM	29	60/64	4
	46	70/75	5
	40	76/80	4
	37	90/94	4
	27	102/106	4
	2	118/124	6
	10	123/127	4
	32	132/137	5
	48	146/151	5
	14	155/159	4
	4	176/181	5
	5	216/219	3
	20	116/120	4
	7	74/79	5
Green= VIC	19	58/62	4
	28	72/77	5
	33	78/80	2
	38	95/101	6
	11	110/115	5
	39	119/125	6
	9	133/138	5
	21	147/149	2
	31	71/74	3
Yellow= NED	17	63/67	4
	47	70/74	4
	1	79/83	4
	6	92/96	4
	12	109/113	4
	44	128/133	5
	18	137/141	4
	16	147/151	4
	45	74/80	6
Red= TAZ	43	64/68	4
	41	73/78	5
	3	79/83	4
	25	93/97	4
	22	114/119	5
	8	127/132	5
	23	140/144	4
	13	147/152	5
	36	71/75	4
Purple= SID	34	67/69	2
	26	75/79	4
	30	84/87	3
	42	110/114	4
	35	112/116	4
	24	129/133	4
	15	146/151	5
	49	72/76	4

Fluorescently Labeled Primers

The 49 primer pairs were arranged into 5 different groups, each labeled with a different fluorophore: blue, green, yellow, red, and purple. The primer pairs were separated based on their resultant PCR product bp size. Amplicons labeled with the same fluorophore need to be different sizes to separate sufficiently and to have distinct products observed during analysis. Table 4 shows the arrangement of each of the amplicons by their fluorophore. The highlighted primer pair amplicons were alternates with those respective fluorophore labels because they were too close in base pair length to the other

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

amplicons, leaving only 43 primer pairs being used. Once the primer pair amplicons were separated by their fluorophores, they were again run individually to make sure they still worked properly with the new forward primer. By running the amplicons individually, it also provided a reference of where the peaks should be seen when multiplexed. After being run on the CE and analyzed using GeneMapper ID-X, all 43 primer pairs were found to work. Figure 11 shows an example of one of the amplified primer pairs for each of the dye channels.

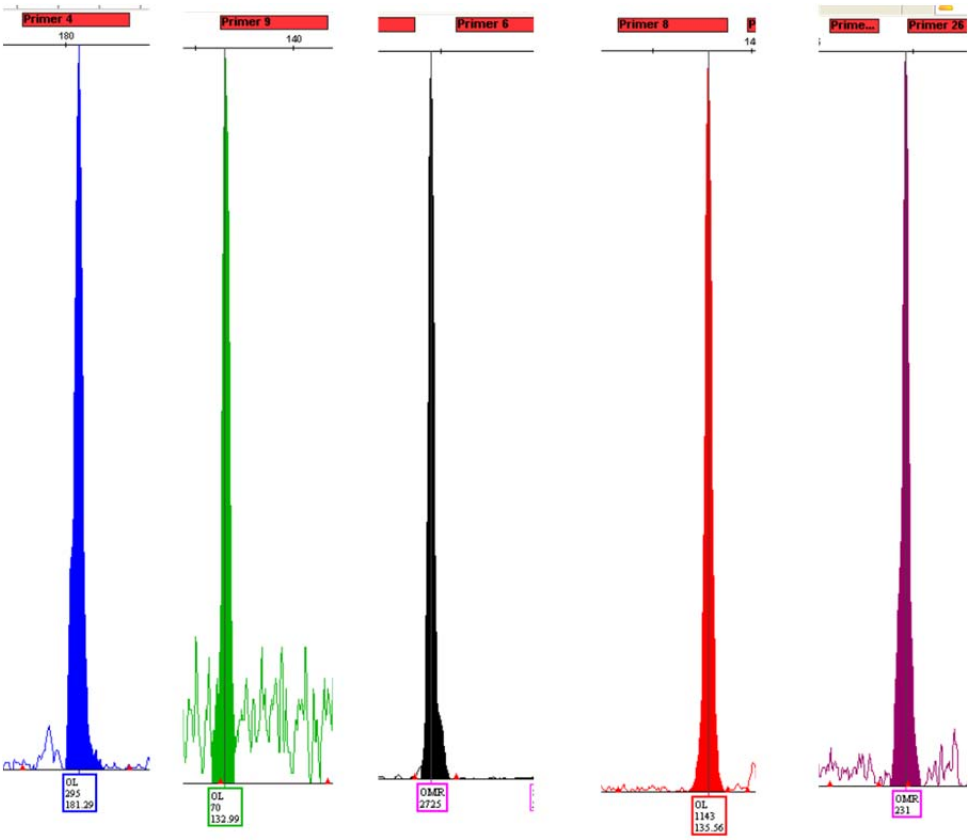


Figure 11. Electropherograms of amplicons 4 (RS2307507), 9 (RS3841948), 6 (RS3838581), 8 (RS3042783), and 34 (RS34811743) (left to right) are shown above. Each of the examples shows one peak indicating a homozygous insertion or deletion at those markers.

When creating a multiplex, it is expected that the size products seen with the individual amplicons would be consistent with the products observed in the multiplex. With the single amplicons known to be working, the amplicons were multiplexed into 5 groups based on the fluorophore label. The fluorophore multiplexes were designed to ensure no substantial spectral overlap from other amplicons. This assessment confirmed the fluorophore multiplexes could be amplified and multiplexed together and produce the desired results of product sizes remaining consistent. After running dilutions of the fluorophore multiplex to improve the primer overlap between amplicons, it was determined that the 1:100 dilution of the amplified product provided the best results. High signal produced by the

amplicons can cause saturation and spectral overlap (pull up) between the amplicons, diluting the sample helps to reduce this oversaturation. Figure 12 shows the electropherograms of the fluorophore multiplexes using the 1:100 dilution of PCR products (0.5 ng/ µL of input DNA).

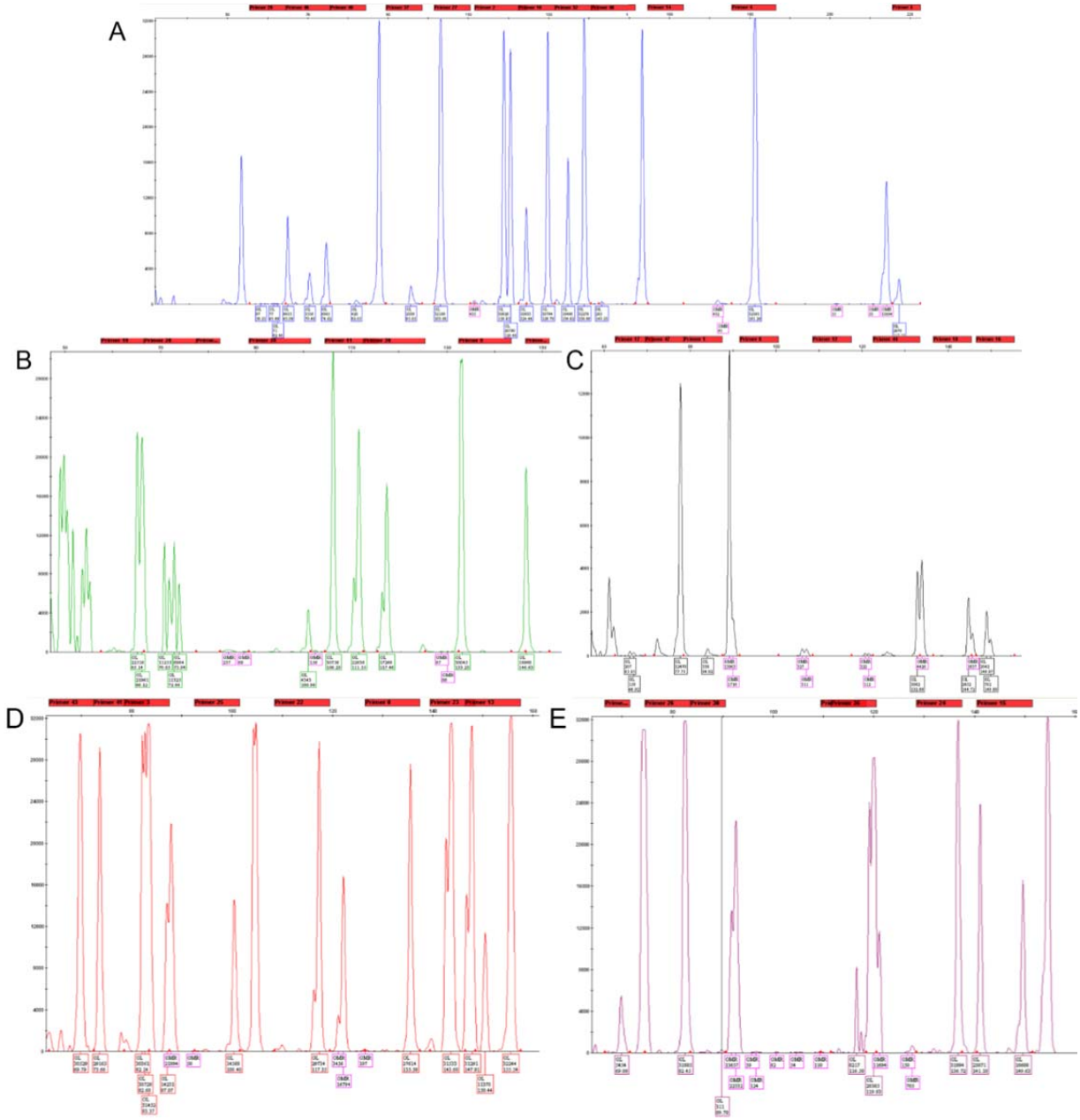


Figure 12. The electropherograms for each of the fluorophore multiplexes, with the x-axis representing the size in bp and the y-axis representing the reflective fluorescent units (RFUs). A) The blue fluorophore multiplex contains 12 amplicons. B) The green fluorophore multiplex contains 8 amplicons. C) The yellow fluorophore multiplex contains 8 amplicons. D) The red fluorophore multiplex contains 8 amplicons. E) The purple fluorophore multiplex contains 7 amplicons.

The fluorophore multiplexes show that most of the amplicons were able to amplify. The amplicon for primer pair 38 in the green fluorophore group did not generate a product in this multiplex (Figure 12). After adding 10 µL more of the forward (20 µM) and reverse (20 µM) primers, a peak could be detected indicating that it is possible to amplify this locus in the multiplex. Figure 13 shows a 1:1000 dilution of the 0.5 ng/ µL of DNA, where more of the primer pair 38 was added.

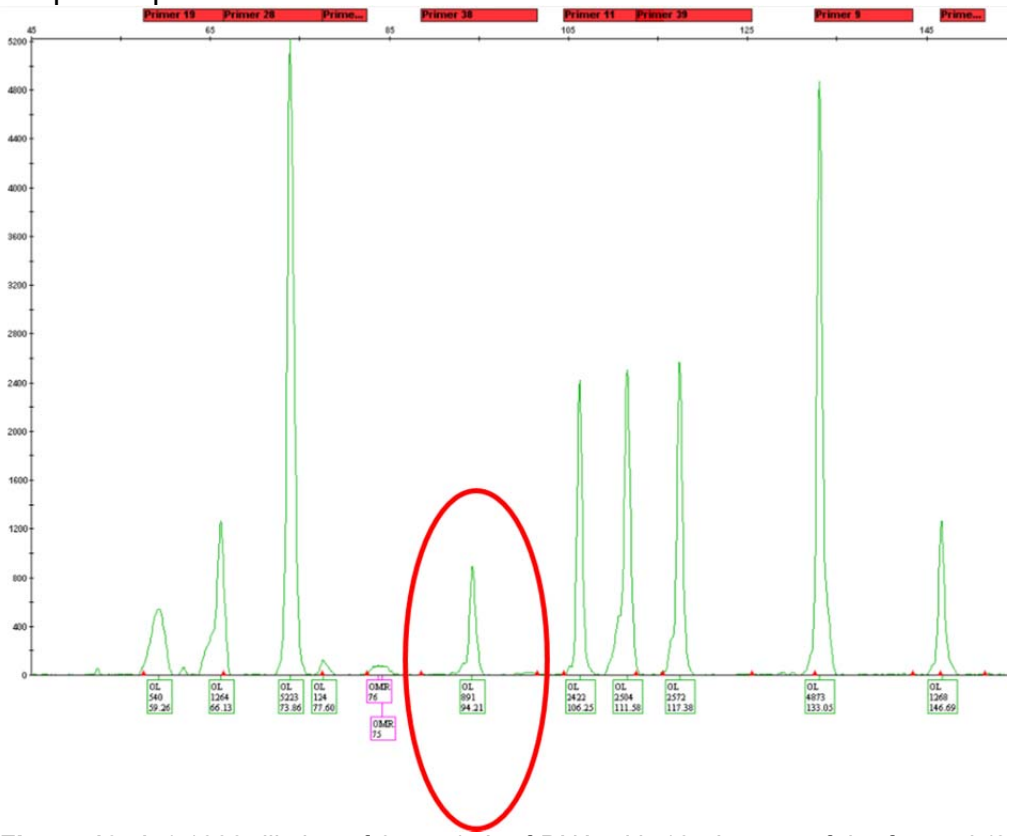


Figure 13. A 1:1000 dilution of 0.5 ng/ µL of DNA with 10 µL more of the forward (20 µM) and reverse (20 µM) primers of primer pair 38 added, with the x-axis representing the size in bp and the y-axis representing RFUs. A peak, circled in red, was seen, indicating this locus could be amplified in the multiplex. Amplicons included in this fluorophore multiplex are 19 (RS1610871), 28 (RS2308112), 33 (RS10688868), 38 (RS2308232), 11 (RS2307603), 39 (RS4187), 9 (RS3841948), 21 (RS2307839) (left to right).

In the yellow fluorophore group, two of the primer pair (12 and 44) amplicons did not produce peaks. An additional 10 µL of the forward (20 µM) and reverse (20 µM) primers for both pairs were added to the primer mix, but no improvement was seen. As a way to check if the amplicons were producing products of the same size in the fluorophore multiplex, the single amplicon images were evaluated next to the multiplex results. Figure 14 shows a representative comparison of some of the amplicons for each fluorophore group.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

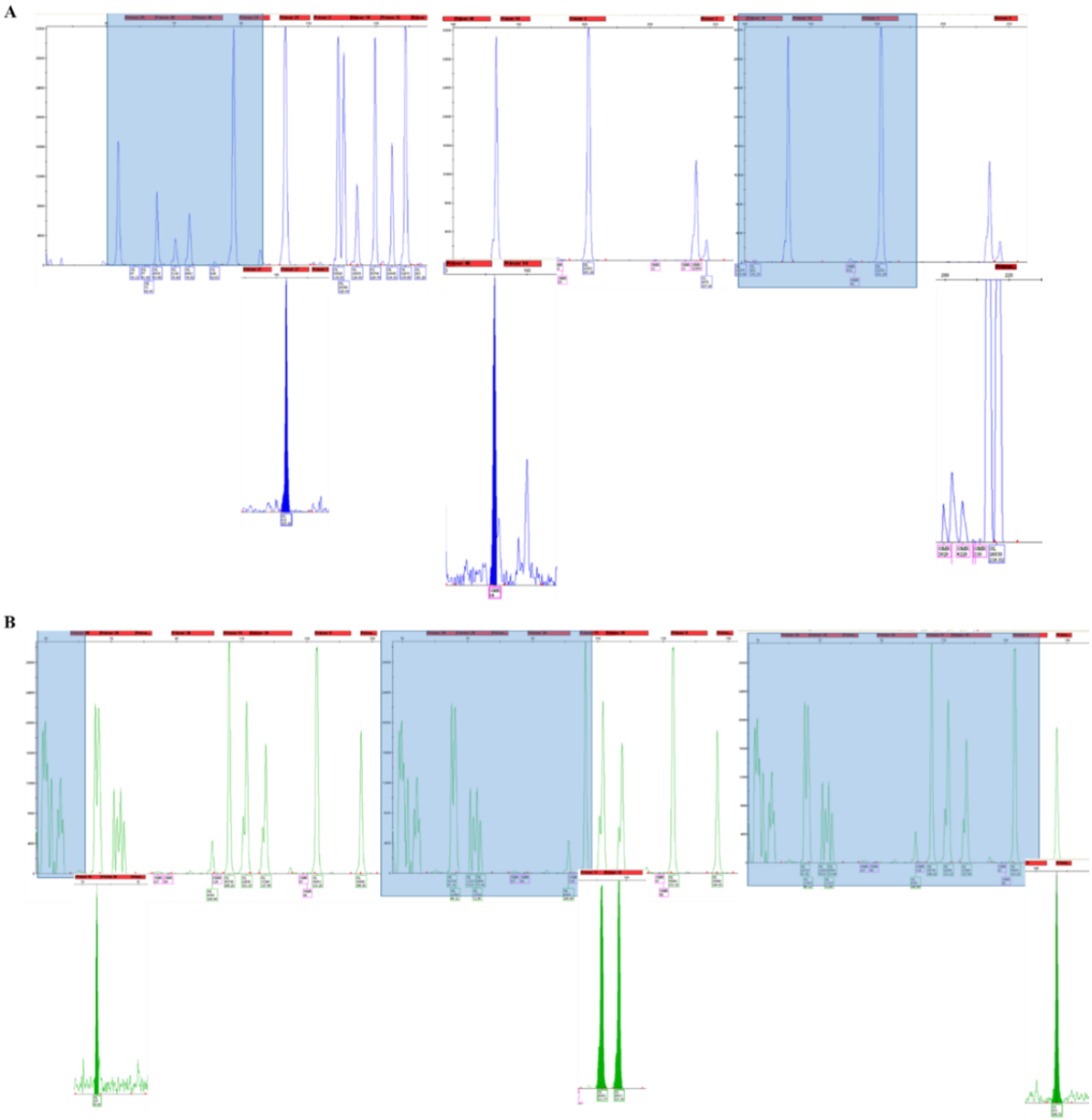


Figure 14. Fluor multiplex and single amplicon electropherograms comparing product size results, with the x-axis representing the size in bp and the y-axis representing RFUs. **A)** Blue fluorophore amplicon results. The INDELs represented are primer pairs 27 (RS16402), 14 (RS2307526), 4 (RS2307507). **B)** Green fluorophore amplicon results. The INDELs represented are primer pairs 28 (RS2308112), 39 (RS4187), 21 (RS2307839).

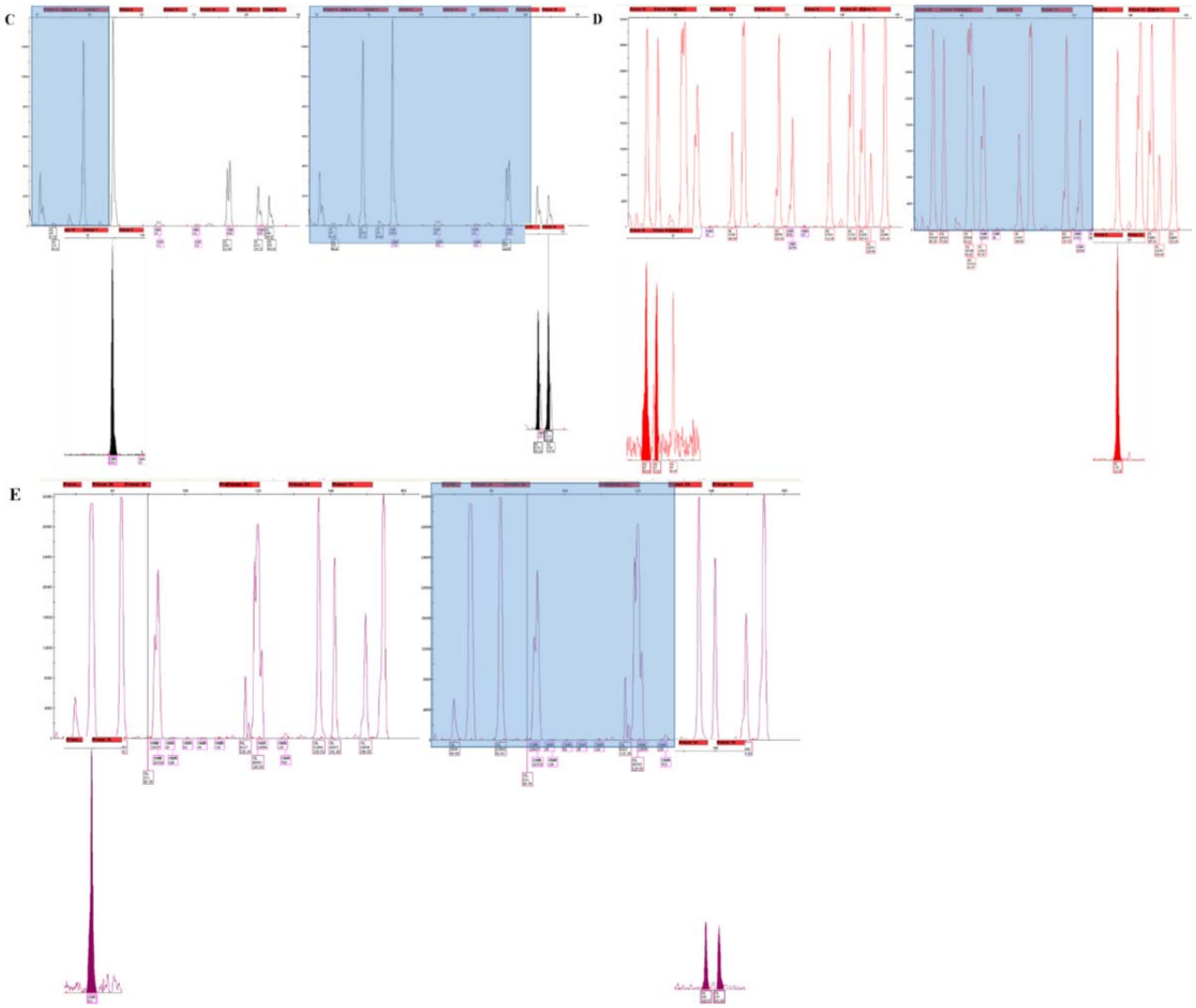
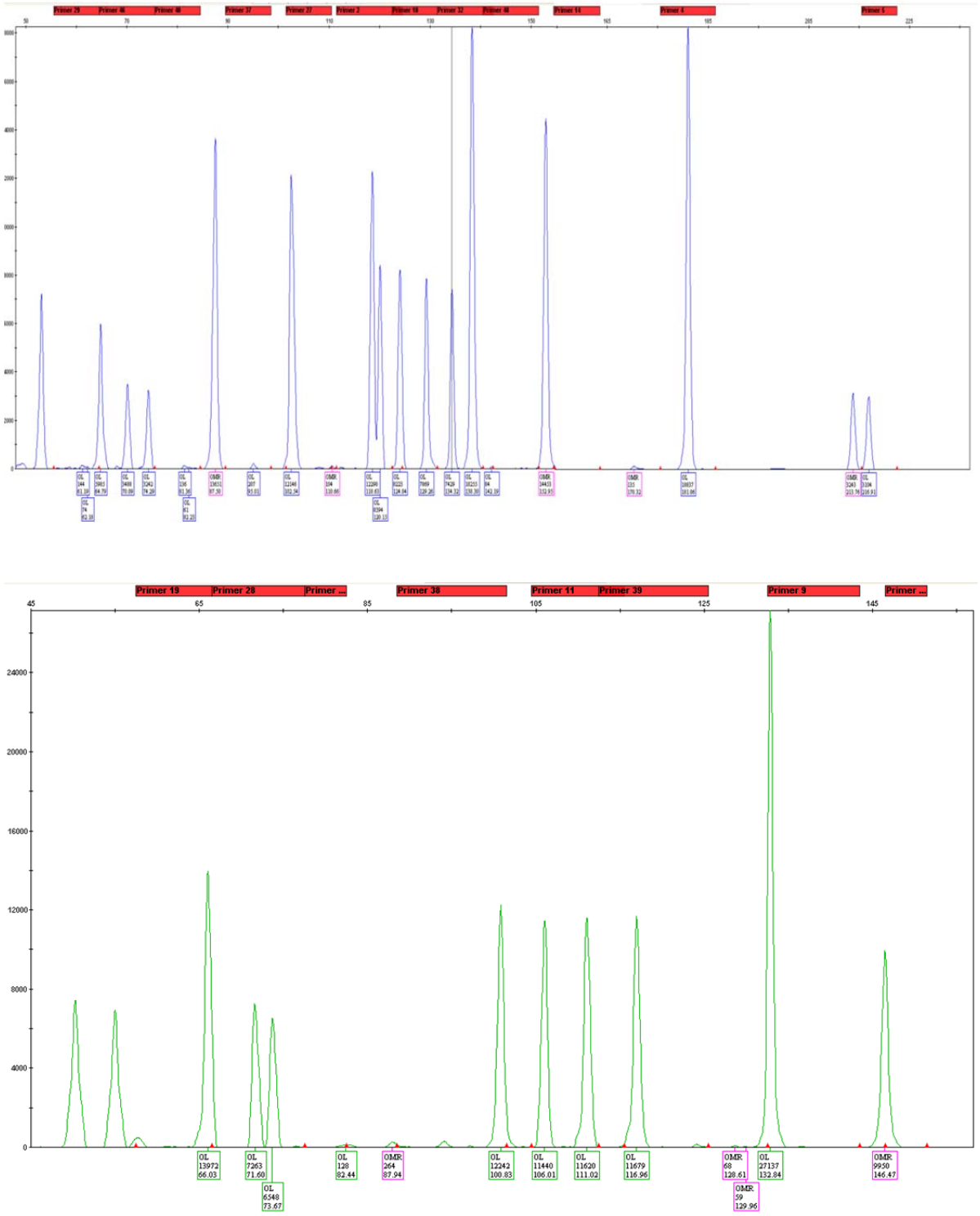
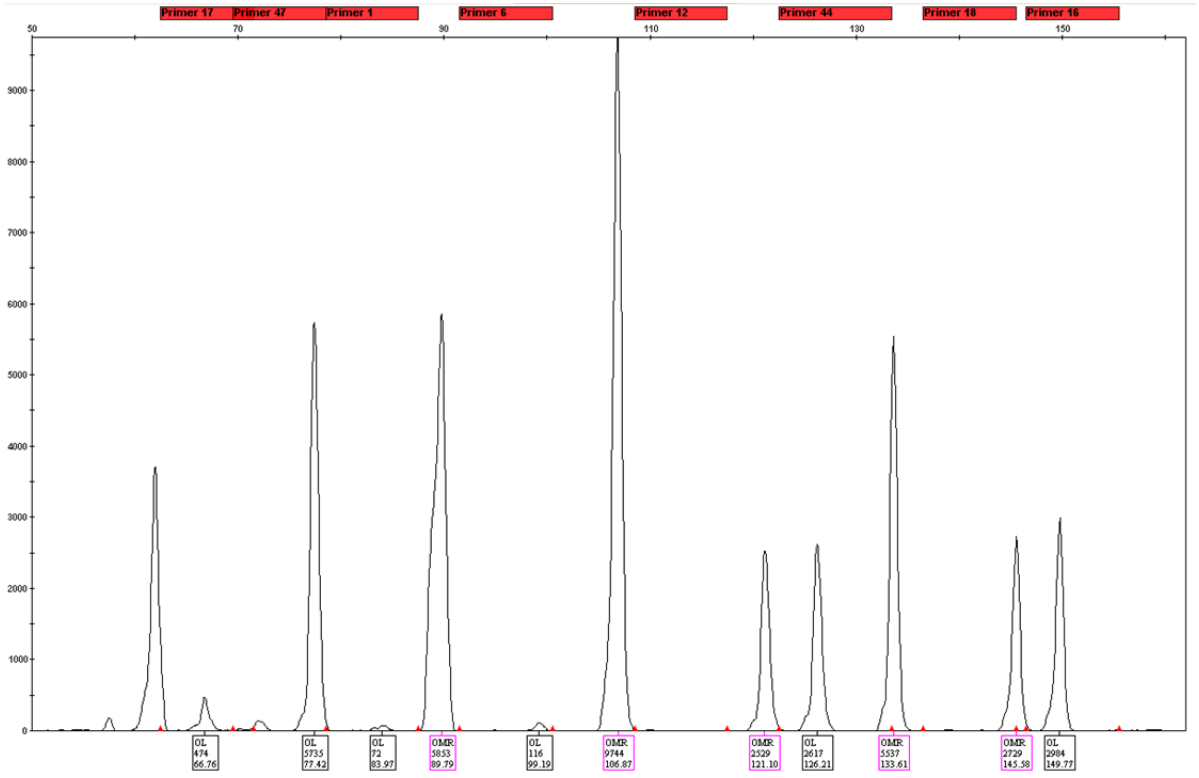


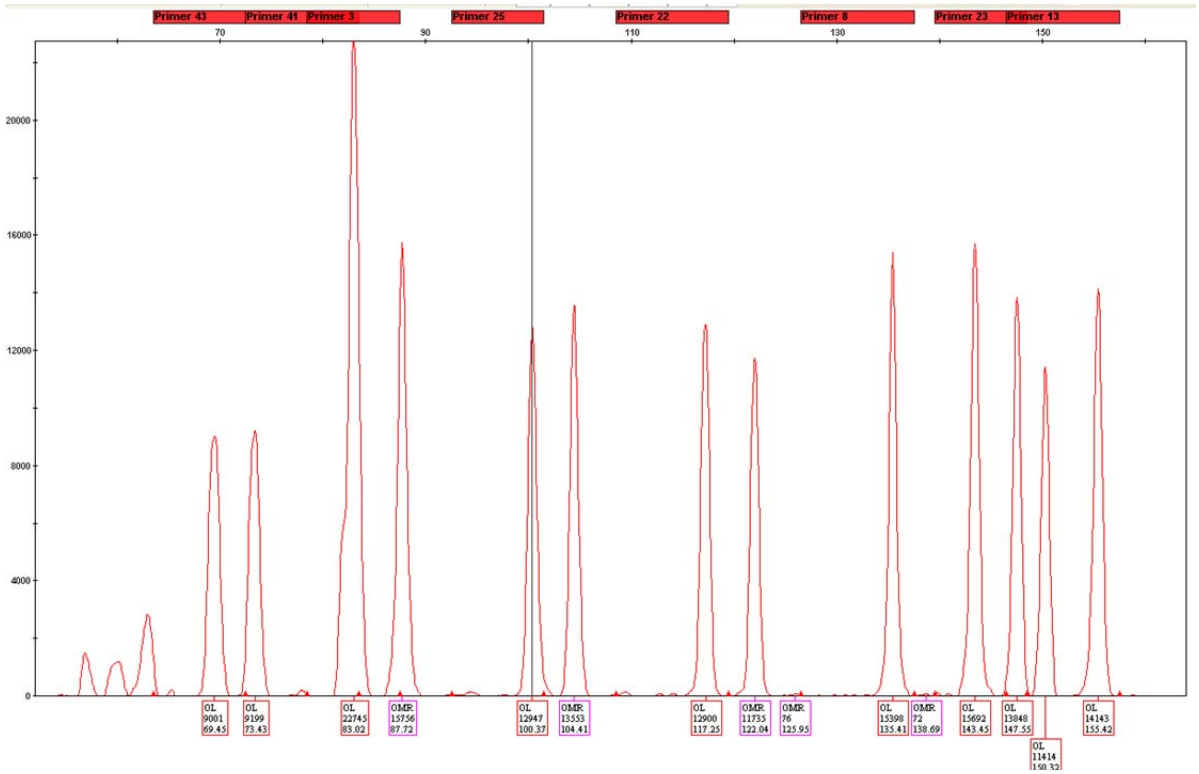
Figure 14. Fluor multiplex and single amplicon electropherograms comparing product size results, with the x-axis representing the size in bp and the y-axis representing RFUs. **C)** Yellow fluorophore amplicon results. The INDELs represented are primer pairs 6 (RS3838581) and 18 (RS2067191) **D)** Red fluorophore amplicon results. The INDELs represented are primer pairs 43 (RS17859968) and 8 (RS3042783). **E)** Purple fluorophore amplicon results. The INDELs represented are primer pairs 34 (RS34811743) and 24 (RS34535242).

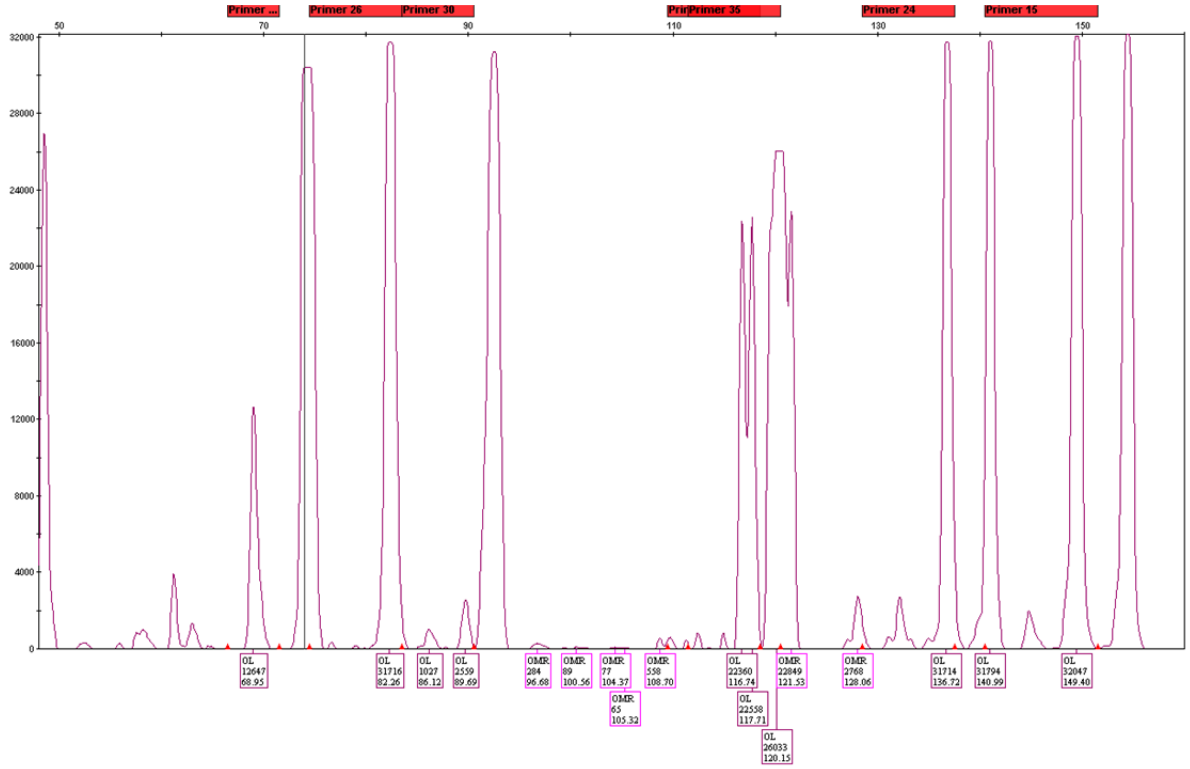
Further refinement of the multiplex by adjusting primers and PCR parameters has now allowed us to combine all of the markers into a single tube. One legitimate criticism of bi-allelic systems is that it is difficult to discern the presence of mixtures in such genotyping systems. We acknowledge this limitation and agree that with CE approaches, bi-allelic systems should only be used for single-source samples. To address this issue in the final multiplex Intra-loci (heterozygote) balance was given precedence over inter-loci balance as it would be more helpful in indicating mixtures. Where possible, primer concentrations were adjusted to give inter-loci balance with more of an emphasis on loci in the same dye channel with intra-dye channel balance when possible. Electorpherograms from this final multiplex are shown in figure 15.

Figure 15. Final INDEL multiplex PCR reaction with electropherograms in dye channels for Fam, Joe, Ned, Taz, and Sid respectively. ILS (not shown was LIZ 600).









Aim 3 of our proposal calls for a developmental validation of our final multiplex according to SWGDAM guidelines. We have made some progress in this area, but the completion of this last section of the proposal is dependent on the release of final funds for the project by NIJ ,and will be included in our revised report at the end of our extension period. Additionally, the markers in the SID dye channel while functional might be re-configured (Primer pair 35 removed; slightly less of all SID labeled primers added to the next lot of multiplex) slightly to improve the quality of data in that dye channel.

MPS assessment of our HID INDEL markers
Portions of this section have been published in FSI:Genetics

As mentioned in the previous section one criticism of bi-allelic markers is that they are a poor marker choice compared to more polymorphic STR markers for mixtures. While this is true when assaying bi-allelic markers via capillary electrophoresis, massively parallel sequencing allows for the detection of SNPs in flanking regions and other polymorphic elements. To address this issue, we utilized a custom enriched sequencing approach for 68 INDELs (our primary candidates plus additional markers described in the literature).

A total of 190 samples were sequenced. One run, containing 12 African American samples, performed poorly with insufficient sequencing Q scores (between 10 and 20) for all of read 2 and part of read 1. This run was removed from analysis due to poor

sequence quality. Ultimately, 178 samples were analyzed, consisting of 49 Caucasians, 37 African Americans, 49 Hispanics, and 43 Asians.

Locus Performance

Analysis of the resulting data was performed using operationally selected DoC and ACR thresholds of 10x and 0.20, respectively. Mean profile completion was 96.3% ± 0.108, ranging from 44.1% to 100% for the 178 samples. Full HID INDEL profiles were obtained with 70 samples. The average DoC and ACR for 68 HID INDELs was 96.9x ± 69.9 and 0.727 ± 0.182, respectively (Figures 16 and 17). One locus, rs33917182, fell below two standard deviations from the overall DoC mean with an average DoC of 26.5x ± 15.7. No loci fell below two standard deviations from the overall ACR mean. While DoC and ACR values were sufficient for analysis, the rs33917182 locus was typed successfully in only 41.6% of the samples (74/178) after application of the DoC and ACR thresholds. Removal of this locus due to poor success resulted in full HID INDEL profiles for 155 samples, increasing the overall mean profile completion to 97.1% ± 0.110.

Sequence Variation

Using STRait Razor, sequence data were obtained for the INDEL motif and approximately 50 bases on either side of the motif (Table 6). Based on 1000 Genomes data, 100 known polymorphisms (94 SNPs and 6 non-HID INDELs) exist within 50 bases of the target HID INDELs. Twenty-five and seventy-five of these polymorphisms have global allele frequencies (GAFs) ≥ 0.02 and < 0.02, respectively. The average distance of these polymorphisms from the target INDEL was 27 bases ± 13. All 25 flanking region polymorphisms with GAFs ≥ 0.02 were observed in the population data for four major US populations. Only 18/75 polymorphisms with GAFs < 0.02 were observed as would be expected due to sampling or being private variants.

In all four populations, 19 INDEL motifs had different sequences than previously reported. All but one of these motif sequence differences could be explained by differences in alignment by manual analysis in IGV. The rs35716687 locus has been reported previously as a TTAA deletion but the marker was identified as a TACT deletion. Fifteen markers were associated with a repeat motif; the initial INDEL selection criteria had sought to avoid such structures by excluding loci with three or more repeats. Four of the 15 markers contained three copies or repeats. The remaining 11 loci contained two copies. These motifs range in size from di- to penta-nucleotides (Table 5). While the number of repeats is limited, STR motifs may become problematic if stutter-type artifacts can be generated. Thus, special attention during validation studies should be paid for potential stutter product generation. Though possible, STRs with only a few repeat motifs are less subject to such PCR artifacts relative to STRs with several to many repeats.

Sequence variation was observed in the region adjacent to the INDEL motif at 42 loci, producing 65 novel microhaplotypes (Table 5) [31-33]. Forty-one HID INDEL loci are part of a microhaplotype containing one or two SNPs. One INDEL, rs34528025, is part of a microhaplotype containing the target INDEL, an adjacent SNP (rs202051643), and an

adjacent flanking-region INDEL (rs34247791). Twenty-two loci had sequence variants that account for ≥ 2% of total alleles in two or more populations. For these 22 microhaplotypes, the presence of additional, sequence-based alleles increased the average number of alleles per marker from 2 to 3.82 ± 1.14 with a range from 3 to 7 alleles (rs1408093). The observed heterozygosity for these 22 loci increased by an average of 0.154 ± 0.0895 for AFA, 0.108 ± 0.0824 for ASA, 0.182 ± 0.104 for CAU, and 0.123 ± 0.0959 for HIS (Table 3). All 68 loci were ranked based on length- and sequence-based observed heterozygosity (Table 7). By length, microhaplotypes containing a SNP and rs10688868, rs2308189, rs2308276, and rs2308292 ranked 33rd, 8th, 19th, and 32nd in the HIS, AFA, ASA, and CAU populations, respectively. However, when ranked by sequence-based observed heterozygosity, microhaplotypes containing these four INDELS displayed the highest heterozygoisties in the HIS, AFA, ASA, and CAU populations, respectively. The second highest heterozygosity microhaplotypes in the AFA, HIS, ASA, and CAU populations are rs10688868, rs2307526, rs2308189, and rs5895446, respectively. These four loci increased from their length-based ranks of 51st, 32nd, 3rd, and 17th, in AFA, HIS, ASA, and CAU, respectively. Microhaplotypes containing the rs10688868 and rs2308189 INDELs are ranked most, or second highest, heterozygosity in the AFA, ASA, and HIS populations, making them far more informative than even the top ranked length-based marker. Single-locus RMPs were decreased by an average of 0.166 ± 0.0816 for AFA, 0.130 ± 0.0661 for ASA, 0.176 ± 0.0837 for CAU, and 0.134 ± 0.0773 for HIS.

The remaining 20 loci with detectable adjacent sequence variants did not display substantial sequence variation (average frequency of 0.0234 ± 0.0250 across all four populations). It should be noted that the frequencies of microhaplotypes containing INDELs rs34528025, rs36062169, and rs3841948 were relatively high: 0.10 for HIS, 0.07 for ASA, and 0.08 for AFA, respectively. However, they were either observed once or not at all in the other population groups. While microhaplotypes containing these three INDELs did not substantially increase the discrimination power across the populations, these alleles may hold value for ancestry apportionment. The low allele frequency of these sequence variants, or lack of sufficient frequency in multiple populations, suggests that these 20 microhaplotypes do not have increased discrimination power over that of the current length-based allele polymorphism (Table 8) for HID applications.

Length-based allele frequencies and observed and expected heterozygosities were similar to those previously reported by our group and Pereira, et al. Prior to Bonferroni correction, three AFA, three ASA, no CAU, and three HIS length-based loci and four AFA, five ASA, three CAU, and two HIS sequence-based loci deviated significantly from HWE ($p < 0.05$). After Bonferroni correction, there were no significant departures from HWE for length- or sequence-based loci ($p = 0.00074$). Prior to Bonferroni correction, 185 AFA, 140 ASA, 197 CAU, and 216 HIS length-based and 205 AFA, 186 ASA, 124 CAU, and 186 HIS sequence-based pairwise LDs were observed ($p = 0.05$). Five (AFA), four (ASA), seven (CAU), and five (HIS) length-based and seven (AFA), eight (ASA), nine (CAU), and six (HIS) sequence-based significant pairwise LDs were observed for markers on the same chromosome but not on the same chromosomal arm. After Bonferroni correction, at most two pairwise locus comparisons showed significant LD for length- and sequence-

based alleles per population (rs2308112 and rs34795726 in AFA and rs34541393 and rs34811743 in ASA, $p < 0.0000219$). The observed significant pairwise LDs are less than that due to chance alone (~114). Assuming independence, the combined length-based RMPs were 1.36 x 10-26 for AFA, 5.42 x 10-27 for ASA, 2.94 x 10-27 for CAU, 1.33 x 10-27 for HIS and the combined sequence-based RMPs were 3.29 x 10-32 for AFA, 5.92 x 10-31 for ASA, 6.69 x 10-32 for CAU, and 5.67 x 10-32 for HIS for 68 HID INDEL-containing microhaplotypes.

The combined RMPs, under the assumption of independence, for 22 microhaplotypes were 3.84 x 10-14 for AFA, 3.87 x 10-13 for ASA, 7.76 x 10-14 for CAU, and 1.60 x 10-13 for HIS. These values are comparable to those obtained with larger INDEL panels described by Pereira, et al. and our group, and commercially available short tandem repeat (STR) kit

Figure 16. Depth of coverage (DoC) values for 68 human identity INDELs using the Nextera™ Rapid Capture Enrichment kit and the Illumina MiSeq™. Each box plot represents a single locus; the center horizontal line represents the median, the lower and upper boundaries of each box represent the first and third quartiles, respectively, the vertical dashed lines indicate plus/minus three times the interquartile range, and closed circles indicate outliers. The center horizontal line indicates the mean across 68 loci and the top and bottom horizontal lines indicate plus and minus two standard deviations, respectively for all loci combined.

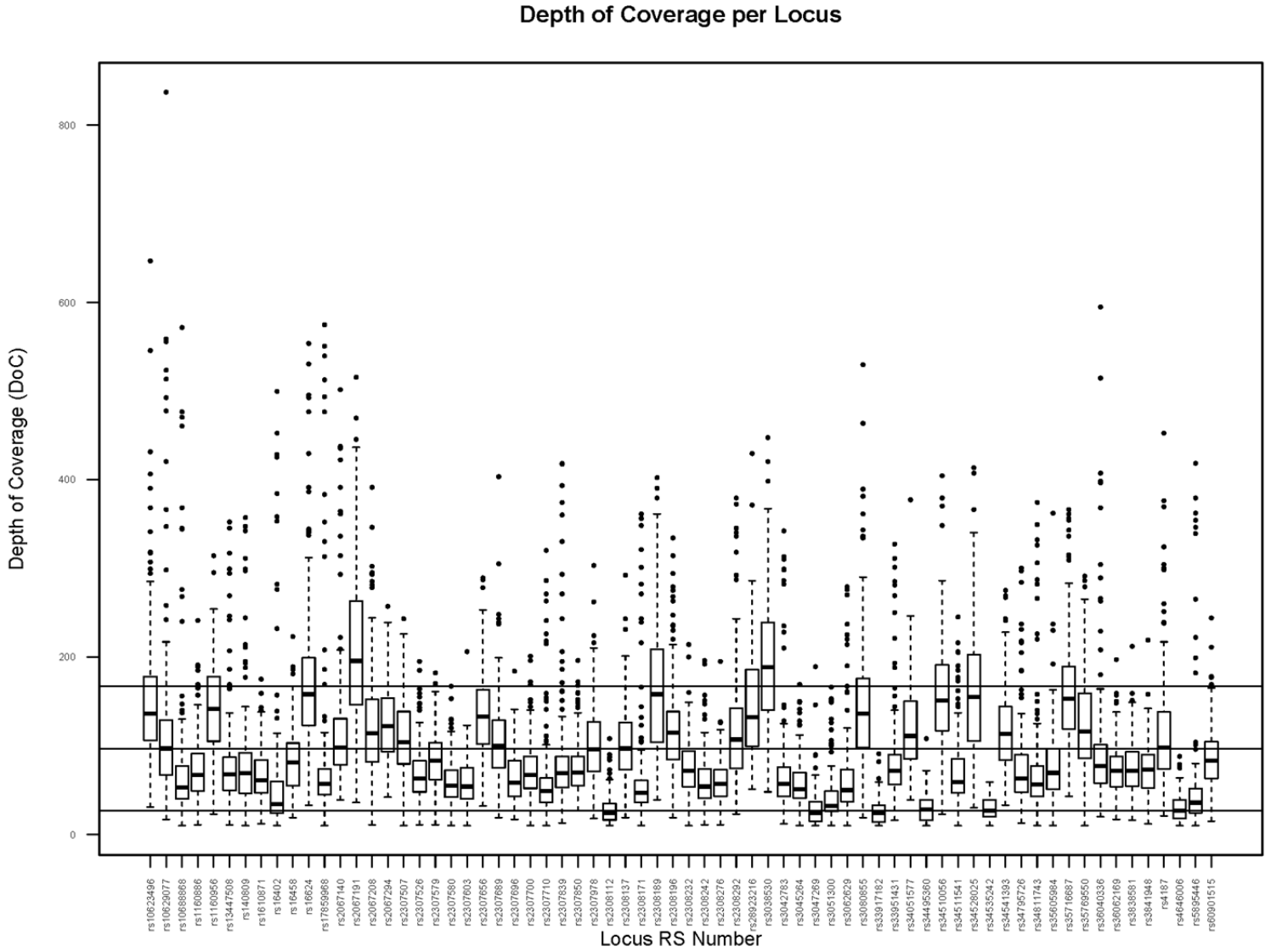


Figure 17. Allele coverage ratio (ACR) values for 68 human identity INDELs using the Nextera™ Rapid Capture Enrichment kit and the Illumina MiSeq™. Each box plot represents a single locus; the center horizontal line represents the median, the lower and upper boundaries of each box represent the first and third quartiles, respectively, the vertical dashed lines indicate plus/minus three times the interquartile range, and closed circles indicate outliers. The center horizontal line indicates the mean across 68 loci and the top and bottom horizontal lines indicate plus and minus two standard deviations, respectively for all loci combined.

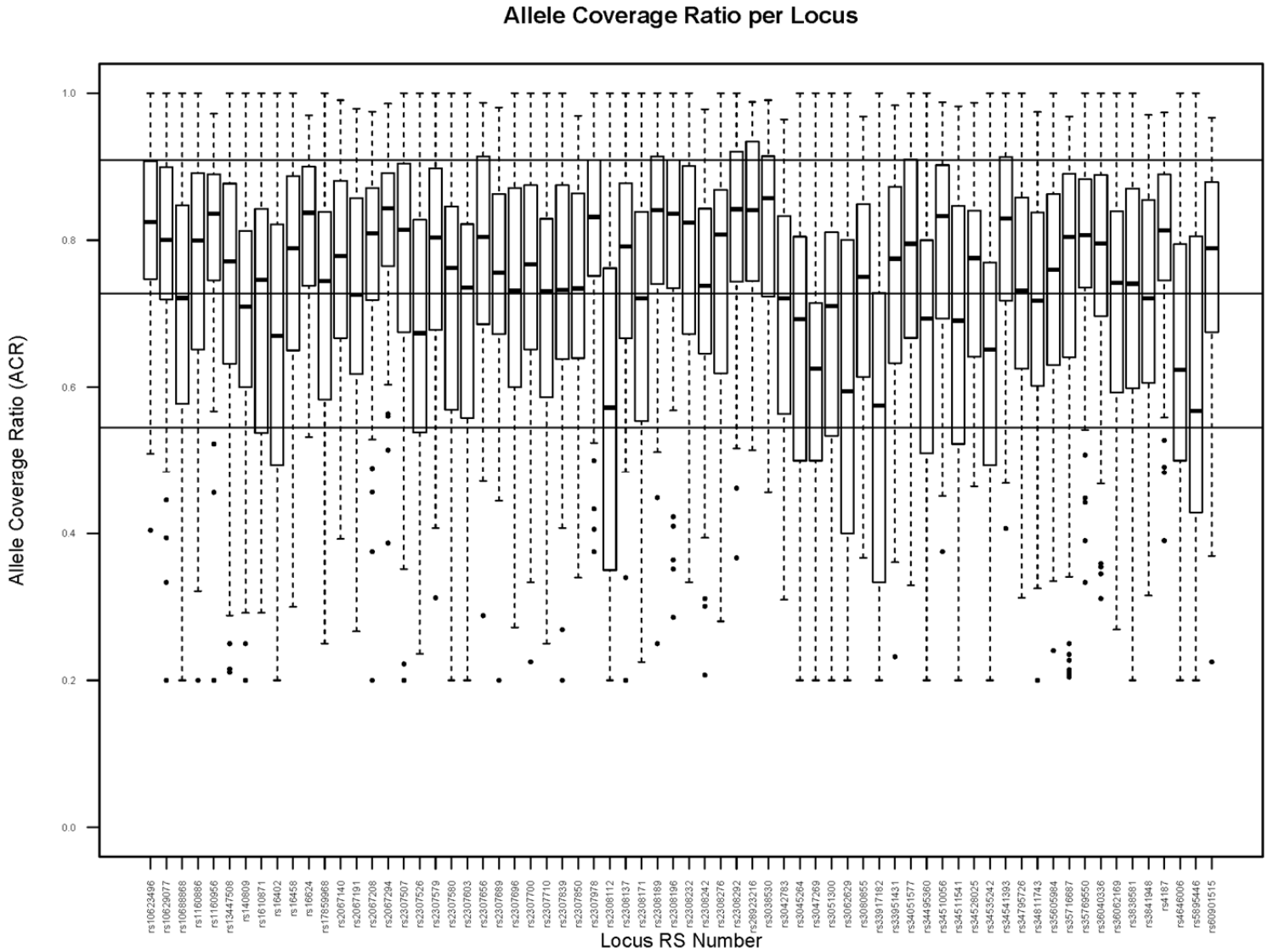


Table 5. Length-based (LBAF) and sequence-based (SBAF) allele frequencies by population for 68 insertion/deletion (INDEL) markers. Target INDEL motifs are underlined and flanking region variants bolded.

INDEL RS Number	Flanking RS Number(s) and hg19 Reference Allele	Length (bp)	Sequence	AFA		ASA		CAU		HIS	
				LBAF	SBAF	LBAF	SBAF	LBAF	SBAF	LBAF	SBAF
rs10623496	chr8:123945676 T[2]	100	TTAAACATTTGATAGTGCCTTATTTATTTATGTATGTGACACATAAAACCATGATTGTTTCTTCTTTCTGTCTTAGCAAGA**T**TTTTTTTTCTGCTTTCAG	0.3784	0.3784	0.3372	0.3256	0.3469	0.3469	0.3061	0.3061
		100	TTAAACATTTGATAGTGCCTTATTTATTTATGTATGTGACACATAAAACCATGATTGTTTCTTCTTTCTGTCTTAGCAAGA**C**TTTTTTTTCTGCTTTCAG		0		0.0116		0		0
		104	TTAAACATTTGATAGTGCCTTATTTATTTATGTATGTGACACATAAAACCAGAATTGATTGTTTCTTCTTTCTGTCTTAGCAAGA**T**TTTTTTTTTCTGCTTTCAG	0.6216	0.6216	0.6628	0.6628	0.6531	0.6531	0.6939	0.6939
rs10629077	rs537464320 C rs201421087 C	100	TTGGTTATCTCTGTCTAATCTA**C**TCTCTTGTGCCCACTTTTACTTACTCA**C**GTCTTTTCCCAACAGCATTCTGTCACACCTCCTAATAGTTTTGCTCATC	0.2568	0.2432	0.2791	0.2791	0.1735	0.1633	0.2755	0.2755
		100	TTGGTTATCTCTGTCTAATCTA**C**TCTCTTGTGCCCACTTTTACTTACTCA**T**GTCTTTTCCCAACAGCATTCTGTCACACCTCCTAATAGTTTTGCTCATC		0.0135		0		0		0
		100	TTGGTTATCTCTGTCTAATCTA**T**TCTCTTGTGCCCACTTTTACTTACTCA**C**GTCTTTTCCCAACAGCATTCTGTCACACCTCCTAATAGTTTTGCTCATC		0		0		0.0102		0
		102	TTGGTTATCTCTGTCTAATCTA**C**TCTCTTGTGCCCACTTTTACTTACTCA**C**ATGTCTTTTCCCAACAGCATTCTGTCACACCTCCTAATAGTTTTGCTCATC	0.7432	0.7432	0.7209	0.7209	0.8265	0.8265	0.7245	0.7245
rs10688868[3]	rs142634555 C rs56780729 C rs10902117 C	100	CCTGTTCCTCGCTCTAGTTCCCCACTTCCATCCCTCCTCTTG**C**CTCAGCCCTT**C**TCATCTCACACGCCACATGGGATCCACCCTTTTTA**T**GCATGTGCAG	0.2027	0.0676	0.5698	0.4767	0.3163	0.0714	0.3673	0.1531
		100	CCTGTTCCTCGCTCTAGTTCCCCACTTCCATCCCTCCTCTTG**C**CTCAGCCCTT**C**TCATCTCACACGCCACATGGGATCCACCCTTTTTA**C**GCATGTGCAG		0.1351		0.093		0.2449		0.2143
		102	CCTGTTCCTCGCTCTAGTTCCCCACTTCCATCCCTCCTCTTG**C**CTCAGCCC**CT**TT**C**TCATCTCACACGCCACATGGGATCCACCCTTTTTA**T**GCATGTGCAG		0.4459		0.4186		0.4898		0.4082
		102	CCTGTTCCTCGCTCTAGTTCCCCACTTCCATCCCTCCTCTTG**C**CTCAGCCC**CT**TT**T**TCATCTCACACGCCACATGGGATCCACCCTTTTTA**C**GCATGTGCAG	0.7973	0.3378	0.4302	0.0116	0.6837	0.1837	0.6327	0.2245
		102	CCTGTTCCTCGCTCTAGTTCCCCACTTCCATCCCTCCTCTTG**T**CTCAGCCC**CT**TT**T**TCATCTCACACGCCACATGGGATCCACCCTTTTA**C**GCATGTGCAG		0		0		0.0102		0
		102	CCTGTTCCTCGCTCTAGTTCCCCACTTCCATCCCTCCTCTTG**C**CTCAGCCC**CT**TT**C**TCATCTCACACGCCACATGGGATCCACCCTTTTTA**C**GCATGTGCAG		0.0135		0		0		0
rs1160886[1]		97	CAACTCTATTCCTTTTCCCATTGTGCTTAAAACTCCTTTGGAATAGTACTGTTTTCTCTATACTTGGAATACATTTGGGTAGACTGGACTATTTCTC	0.3919		0.4535		0.4082		0.3372	
		100	CAACTCTATTCCTTTTCCCATTGTGCTTAAAACTCCTTTGGAATAGTACTACTGTTTTCTCTATACTTGGAATACATTTGGGTAGACTGGACTATTTCTC	0.6081		0.5465		0.5918		0.6628	
rs1160956 [-]	rs138536239 T	101	CAAATTGTGTTCCTCCAAGGTATAGCTTTAGAAGGATCTTCTT**T**CAGTTGCTCTTTCAAACTTTTCCTTGTTAGAAAAGAAAAACTAATACAAATTGGTTA	0.5405	0.5405	0.6047	0.6047	0.8163	0.8163	0.6125	0.6125
		104	CAAATTGTGTTCCTCCAAGGTATAGCTTTAGAAGGATCTTCTT**T**CAGTTGCTCTTCTTTCAAACTTTTCCTTGTTAGAAAAGAAAAACTAATACAAATTGGTTA	0.4595	0.4459	0.3953	0.3953	0.1837	0.1837	0.3875	0.3875
		104	CAAATTGTGTTCCTCCAAGGTATAGCTTTAGAAGGATCTTCTT**G**CAGTTGCTCTTCTTTCAAACTTTTCCTTGTTAGAAAAGAAAAACTAATACAAATTGGTTA		0.0135		0		0		0
rs13447508	rs13447507 A	94	AATGTACATTATTAGATGTACT**A**TGGTTCAGTGGAAATAGCATGAACTAACAGAGTCTAATAATTTTTGACCTTAGACATGTCTCTTATCTCTG	0.3649	0.3649	0.5116	0.5116	0.2857	0.2857	0.3854	0.3854
		100	AATGTACATTATTAGATGTACT**A**TGGTTCAGTGGAAATAGCATGAACTAACAGACTTAGAGTCTAATAATTTTTGACCTTAGACATGTCTCTTATCTCTG	0.6351	0.6081	0.4884	0.4884	0.7143	0.7143	0.6146	0.6146
		100	AATGTACATTATTAGATGTACT**G**TGGTTCAGTGGAAATAGCATGAACTAACAGACTTAGAGTCTAATAATTTTTGACCTTAGACATGTCTCTTATCTCTG		0.0270		0		0		0
rs140809[3]	rs10905513 A rs687805 C	97	ATGTTCTTAGCCATGGA**A**TTCTTTAGGCTTAATTTTACTTCCAGTTAAATGCAGC**T**GCTGTGGTCACTCAGGAGGGGGATGGGCACCCAGAGTTCCT	0.4189	0.3514	0.3605	0.186	0.4688	0.3958	0.1633	0.0816
		97	ATGTTCTTAGCCATGGA**A**TTCTTTAGGCTTAATTTTACTTCCAGTTAAATGCAGC**C**GCTGTGGTCACTCAGGAGGGGGATGGGCACCCAGAGTTCCT		0.0541		0.1628		0.0729		0.0612
		97	ATGTTCTTAGCCATGGA**G**TTCTTTAGGCTTAATTTTACTTCCAGTTAAATGCAGC**T**GCTGTGGTCACTCAGGAGGGGGATGGGCACCCAGAGTTCCT		0.0135		0.0116		0		0.0204
		100	ATGTTCTTAGCCATGGA**A**TTCTTTAGGCTTAATTTTACTTCCAGTTAACAAATGCAGC**C**GCTGTGGTCACTCAGGAGGGGGATGGGCACCCAGAGTTCCT	0.5811	0.4595	0.6395	0.4651	0.5313	0.4479	0.8367	0.7347
		100	ATGTTCTTAGCCATGGA**G**TTCTTTAGGCTTAATTTTACTTCCAGTTAACAAATGCAGC**C**GCTGTGGTCACTCAGGAGGGGGATGGGCACCCAGAGTTCCT		0.0270		0.1628		0.0625		0.0918

		100	ATGTTCTTAGCCATGGA**G**TTCTTTAGGCTTAATTTTACTTCCAGTTAACAATGCAGC**T**GCTGTGGTCACTCAGGAGGGGGATGGGCACCCAGAGTTCCT		0.0811		0		0.0208		0.0102
		100	ATGTTCTTAGCCATGGA**A**TTCTTTAGGCTTAATTTTACTTCCAGTTAACAATGCAGC**T**GCTGTGGTCACTCAGGAGGGGGATGGGCACCCAGAGTTCCT		0.0135		0.0116		0		0
rs1610871[3]	rs111817892 G rs75866020 C chr5:171088015 T[2]	96	TGCAGATAGCCTCACCTTCCTCCCAGTAACCATCAA**G**CCCCATGAAGAAGGAGTAGCAGGGGAAATGGAGTCCACTAAAAGGC**C**GAAGCCC**T**CGGC	0.5000	0.5000	0.3721	0.3721	0.5918	0.5918	0.4744	0.4744
		100	TGCAGATAGCCTCACCTTCCTCCCAGTAACCATCAA**G**CCCCATGAAGTAGGAAGGAGTAGCAGGGGAAATGGAGTCCACTAAAAGGC**C**GAAGCCC**T**CGGC		0.3243		0.593		0.4082		0.5128
		100	TGCAGATAGCCTCACCTTCCTCCCAGTAACCATCAA**G**CCCCATGAAGTAGGAAGGAGTAGCAGGGGAAATGGAGTCCACTAAAAGGC**G**GAAGCCC**T**CGGC		0.1081		0.0349		0		0.0128
		100	TGCAGATAGCCTCACCTTCCTCCCAGTAACCATCAA**A**CCCCATGAAGTAGGAAGGAGTAGCAGGGGAAATGGAGTCCACTAAAAGGC**C**GAAGCCC**T**CGGC	0.5000	0.0541	0.6279	0	0.4082	0	0.5256	0
		100	TGCAGATAGCCTCACCTTCCTCCCAGTAACCATCAA**G**CCCCATGAAGTAGGAAGGAGTAGCAGGGGAAATGGAGTCCACTAAAAGGC**C**GAAGCCC**C**CGGC		0.0135		0		0		0
rs16402	-	100	GTTAATGCGCCTTTTTGGTTTTGGTGAAAATTTCAGACAGTGTGACATTAATTATTTGTCTTTGCACAGAATGATGATGCTAAAGCCAGCCCCATACAGT		0.3108		0.2326		0.2551		0.2396
		104	GTTAATGCGCCTTTTTGGTTTTGGTGAAAATTTCAGACAGTGTGACATTAATTATTTGTCTTTGCACAGAATGATGATGCTAAAGCCAGCCCCATACAGT		0.6892		0.7674		0.7449		0.7604
rs16458	-	100	GGTTAATCTTCCCCCAAAAACCTACCAATTAACAGTCTCAAAGTTTTACAATTCCTTCATTTCCAAACTCTCTTATGAGTAATAATCAAAGTATACATAT		0.5135		0.6047		0.6939		0.5000
		104	GGTTAATCTTCCCCCAAAAACCTACCAATTAACAGTCTCAAAGTTTTACAATTCCTTCCTTCATTTCCAAACTCTCTTATGAGTAATAATCAAAGTATACATAT		0.4865		0.3953		0.3061		0.5000
rs16624	rs146701576 C rs140263477 A rs250921 T	100	TCCCCTTTCCAC**C**ACAGTTACCTTTCAGGCTCTTGGGTTCTGGAAAGATGTTTTTGACATTGCAACTTGGAAGTCACA**A**TG**T**ATTTTTCACAAAACAGTG		0.2568		0.4302		0.7551		0.4286
		100	TCCCCTTTCCAC**T**ACAGTTACCTTTCAGGCTCTTGGGTTCTGGAAAGATGTTTTTGACATTGCAACTTGGAAGTCACA**A**TG**T**ATTTTTCACAAAACAGTG	0.2568	0	0.4419	0	0.7653	0.0102	0.4744	0.0102
		100	TCCCCTTTCCAC**C**ACAGTTACCTTTCAGGCTCTTGGGTTCTGGAAAGATGTTTTTGACATTGCAACTTGGAAGTCACA**A**TG**C**ATTTTTCACAAAACAGTG		0		0.0116		0		0
		102	TCCCCTTTCCAC**C**ACAGTTACCTTTCAGGCTCTTGGGTTCTGGAAAGATGTGTTTTTGACATTGCAACTTGGAAGTCACA**A**TG**C**ATTTTTCACAAAACAGTG		0.7162		0.5581		0.2347		0.5612
		102	TCCCCTTTCCAC**C**ACAGTTACCTTTCAGGCTCTTGGGTTCTGGAAAGATGTGTTTTTGACATTGCAACTTGGAAGTCACA**G**TG**C**ATTTTTCACAAAACAGTG	0.7432	0.027	0.5581	0	0.2347	0	0.5256	0
rs17859968[1,3]	rs9923304 C rs16955268 A	103	ATT**T**CAGAGTGCATGCTGTCTT**A**AAAGATTGTGGGGATTAAATTACAAGGCACATAAAGCACATGGCATTTAGTAGGACCCTCAATAAATGATAACTCTTACTC	0.3378	0.2297	0.3488	0.2674	0.4388	0.4286	0.4184	0.3878
		103	ATC**C**AGAGTGCATGCTGTCTT**A**AAAGATTGTGGGGATTAAATTACAAGGCACATAAAGCACATGGCATTTAGTAGGACCCTCAATAAATGATAACTCTTACTC		0.1081		0.0814		0.0102		0.0306
		107	ATC**C**AGAGTGCATGCTGTCTT**A**AAAGATTGTGGGGATTAAATTACAAGGCACATAATAAATAAAGCACATGGCATTTAGTAGGACCCTCAATAAATGATAACTCTTACTC	0.6622	0.6351	0.6512	0.6512	0.5612	0.5612	0.5816	0.5816
		107	ATC**C**AGAGTGCATGCTGTCTT**G**AAAGATTGTGGGGATTAAATTACAAGGCACATAATAAATAAAGCACATGGCATTTAGTAGGACCCTCAATAAATGATAACTCTTACTC		0.027		0		0		0
rs2067140[3]	rs192851878 T rs254233 C	96	AAAGGAAAATACAGGCA**T**GCCAATCACTACCCACCAAATGTC**A**CTGACACCAAAGAATCACTGATTAACCTGGAGAGCACTGCAAGGTGAGCTATA	0.2162	0.2162	0.6512	0.6512	0.6122	0.6122	0.6224	0.6224
		100	AAAGGAAAATACAGGCA**T**GCCAATCACTACCCACCAAATGTC**C**CTGACACCAGTCAAAGAATCACTGATTAACCTGGAGAGCACTGCAAGGTGAGCTATA		0.4730		0.1395		0.2245		0.1939
		100	AAAGGAAAATACAGGCA**T**GCCAATCACTACCCACCAAATGTC**A**CTGACACCAGTCAAAGAATCACTGATTAACCTGGAGAGCACTGCAAGGTGAGCTATA	0.7838	0.3108	0.3488	0.1977	0.3878	0.1633	0.3776	0.1837
		100	AAAGGAAAATACAGGCA**A**GCCAATCACTACCCACCAAATGTC**A**CTGACACCAGTCAAAGAATCACTGATTAACCTGGAGAGCACTGCAAGGTGAGCTATA		0		0.0116		0		0
rs2067191	rs115923419 C	91	GGATTAGAGTAATGTAAGTAATCTGATGAAATTTACCACTTCTAGTTATTTCCTTGTTTATGAACTACAGAATC**C**GATTACCCTGAAATTC	0.4189	0.4189	0.4302	0.4302	0.4898	0.4898	0.3523	0.3523
		95	GGATTAGAGTAATGTAAGTAATCTGATGAAATTTACCACTTCTAGATAGTTATTTCCTTGTTTATGAACTACAGAATC**C**GATTACCCTGAAATTC		0.5676		0.5698		0.5102		0.6477
		95	GGATTAGAGTAATGTAAGTAATCTGATGAAATTTACCACTTCTAGATAGTTATTTCCTTGTTTATGAACTACAGAATC**T**GATTACCCTGAAATTC	0.5811	0.0135	0.5698	0	0.5102	0	0.6477	0
rs2067208[1,3]	rs74531425 G	95	GGGGCTCAGGCAGCTGAAGAGAATGTTCTAGAATCCACAAAGAGCCTGGCAGGAGCCTGGGGACCTGGA**G**ATGCAGCCGGGTGCATCTATTCAGG	0.1892	0.1892	0.1744	0.1744	0.3163	0.3163	0.3061	0.3061
		100	GGGGCTCAGGCAGCTGAAGAGAATGTTCTAGAATCCACAAAGAGCCTGGCCTGGCAGGAGCCTGGGGACCTGGA**G**ATGCAGCCGGGTGCATCTATTCAGG		0.8108		0.7907		0.5408		0.6429
		100	GGGGCTCAGGCAGCTGAAGAGAATGTTCTAGAATCCACAAAGAGCCTGGCCTGGCAGGAGCCTGGGGACCTGGA**C**ATGCAGCCGGGTGCATCTATTCAGG	0.8108	0	0.8256	0.0349	0.6837	0.1429	0.6939	0.0510
rs2067294	-	100	GAAATGCAGCTTGATTCTCTATGCTATCCCGGCAATTCTCTATTTTTTGTCCTTTCTCTCCTCCATTCTGATATTCTCTTGCTTCTTGTTCACTCCTCGA		0.1351		0.2326		0.3265		0.5000
		103	GAAATGCAGCTTGATTCTCTATGCTATCCCGGCAATTCTCTATTTTTTGTCCTTCTTCTCTCCTCCATTCTGATATTCTCTTGCTTCTTGTTCACTCCTCGA		0.8649		0.7674		0.6735		0.5000

rs2307507	rs540604306 G	95	TTTTAAAATATGTTATATTTTAGGGCTTTGAAATTACT**G**AAAATAATTTAATGTTACATAGTTTTAAATCACTTTCATAGCTTAATAGGGTTTTA	0.1892	0.1757	0.3293	0.3293	0.4286	0.4286	0.4512	0.4512
		95	TTTTAAAATATGTTATATTTTAGGGCTTTGAAATTACT**A**AAAATAATTTAATGTTACATAGTTTTAAATCACTTTCATAGCTTAATAGGGTTTTA		0.0135	0		0		0	
		100	TTTTAAAATATGTTATATTTTAGGGCTTTGAAATTACT**G**AAAATAATTTAATTTTATGTTACATAGTTTTAAATCACTTTCATAGCTTAATAGGGTTTTA	0.8108	0.8108	0.6707	0.6707	0.5714	0.5714	0.5488	0.5488
rs2307526[3]	rs814781 C	100	TAATGCCAGGAGATAATTTATATACAAAGCAAAGGATGCTCAACAACATACGTAATGGGCTTCCGAATTACAAAGAACAGAATCAGGACT**C**AAGCAGATG	0.3649	0.3649	0.6341	0.6341	0.4694	0.4694	0.2955	0.2955
		104	TAATGCCAGGAGATAATTTATATACAAAGCAAAGGATGCTCAACAACATACGTGTGTAATGGGCTTCCGAATTACAAAGAACAGAATCAGGACT**G**AAGCAGATG	0.6351	0.3649	0.3659	0.3171	0.5306	0.4082	0.7045	0.4886
		104	TAATGCCAGGAGATAATTTATATACAAAGCAAAGGATGCTCAACAACATACGTGTGTAATGGGCTTCCGAATTACAAAGAACAGAATCAGGACT**C**AAGCAGATG		0.2703		0.0488		0.1224		0.2159
rs2307579	rs77911955 G	100	TAAAATTGTCACTGTACTAAAAATACCATTAATAATAAAGTGTGGAAAGACATATTCAGCTCTCTTCAAAAGTAGCTATTT**G**GTTATTTAAAGTTAGCTC	0.5811	0.5811	0.1744	0.1744	0.4184	0.4184	0.3500	0.3500
		103	TAAAATTGTCACTGTACTAAAAATACCATTAATAATAAAGTGTGGAAAGACATGATATTCAGCTCTCTTCAAAAGTAGCTATTT**G**GTTATTTAAAGTTAGCTC	0.4189	0.4054	0.8256	0.8256	0.5816	0.5816	0.6500	0.6500
		103	TAAAATTGTCACTGTACTAAAAATACCATTAATAATAAAGTGTGGAAAGACATGATATTCAGCTCTCTTCAAAAGTAGCTATTT**C**GTTATTTAAAGTTAGCTC		0.0135		0		0		0
rs2307580	-	100	TGGTTCTTAACCTCTGAAATTAATTTTTATATTTCTAAATTGCTTAATTGAATAGTTATGCTATGGTCCTGTTCCATATCTAGATACCTCCCAAGTATTC	0.2432		0.4286		0.4490		0.5513	
		104	TGGTTCTTAACCTCTGAAATTAATTTTTATATTTCTAAATTGCTTAATTAATTGAATAGTTATGCTATGGTCCTGTTCCATATCTAGATACCTCCCAAGTATTC	0.7568		0.5714		0.5510		0.4487	
rs2307603[1]		95	GAAAAAAGTCAGTGTGCCTACAGATACCACTTTATTTTGTCCACTTAGAACTAGGTTGCTATCTTTAGCTTAATTCTAGCATATGAGAAATGTGA	0.3919		0.4419		0.6122		0.3846	
		100	GAAAAAAGTCAGTGTGCCTACAGATACCACTTTATTTTGTCCACTTAGATCAGAACTAGGTTGCTATCTTTAGCTTAATTCTAGCATATGAGAAATGTGA	0.6081		0.5581		0.3878		0.6154	
rs2307656[3] -	rs13158027 T	95	AAATGTTGCTTCCC**T**ATCTACACAACCTGGTTTGCCCACGTGCCTCTTGATAATTAGAATAGAACTCTAGAAGGATATTAGCATGTCATTTTTCA	0.6216	0.6216	0.5581	0.5581	0.5000	0.5000	0.6429	0.6429
		100	AAATGTTGCTTCCC**C**ATCTACACAACCTGGTTTGCCCACGTGCCTCTTGATAAGTTAATTAGAATAGAACTCTAGAAGGATATTAGCATGTCATTTTTCA		0.2162		0.1163		0.3571		0.2262
		100	AAATGTTGCTTCCC**T**ATCTACACAACCTGGTTTGCCCACGTGCCTCTTGATAAGTTAATTAGAATAGAACTCTAGAAGGATATTAGCATGTCATTTTTCA	0.3784	0.1622	0.4419	0.3256	0.5000	0.1429	0.3571	0.1310
rs2307689[3]	rs36120065 A	86	CTTCCCCAAGGCCAAACCTCCTCCCTCAGAAGAGGGCTGTTCTTCCTCAAGTTACTAAGCCTCCTTCTTCAAACGAGGGG**A**TACCC	0.5676	0.5676	0.2674	0.2674	0.2143	0.2143	0.3605	0.3605
		89	CTTCCCCAAGGCCAAACCTCCTCCCTCAGAAGAGGGCTGTTCTTCCTCAAGTTACTAAGCCTCCTTCTTCAAACGAGGGG**A**TACCC		0.2703		0.407		0.5102		0.3488
		89	CTTCCCCAAGGCCAAACCTCCTCCCTCAGAAGAGGGCTGTTCTTCCTCAAGTTACTAAGCCTCCTTCTTCAAACGAGGGG**G**TACCC	0.4324	0.1622	0.7326	0.3256	0.7857	0.2755	0.6395	0.2907
rs2307696	-	96	ACACTACACAACAGAAACAATAACGAATCCCCCCCATCACACACACACAGCGACCTGAGGGTTCTGAAGAGTAAATAAGAGCAGTCAGATGGGCTC	0.5405		0.3023		0.4082		0.4459	
		100	ACACTACACAACAGAAACAATAACGAATCCCCCCCATCACACACACACAGCGACCGACCTGAGGGTTCTGAAGAGTAAATAAGAGCAGTCAGATGGGCTC	0.4595		0.6977		0.5918		0.5541	
rs2307700[3]	rs4239922 G	106	AACCTGGGACAGCTGGCAGGGCCAGAGCTCAGCT**G**ATCATTCATGGGCTTGTCAGTCAGTAAAAGAATGGGGATTCCGAGGAAGGACAAACAAAGGAAAGAAAAAA	0.3378	0.3378	0.2674	0.2674	0.5408	0.5408	0.3125	0.3125
		110	AACCTGGGACAGCTGGCAGGGCCAGAGCTCAGCT**C**ATCATTCATGGGCTTGTCACTCAGTCAGTAAAAGAATGGGGATTCCGAGGAAGGACAAACAAAGGAAAGAAAAAA	0.6622	0.3784	0.7326	0.6744	0.4592	0.2449	0.6875	0.5521
		110	AACCTGGGACAGCTGGCAGGGCCAGAGCTCAGCT**G**ATCATTCATGGGCTTGTCACTCAGTCAGTAAAAGAATGGGGATTCCGAGGAAGGACAAACAAAGGAAAGAAAAAA		0.2838		0.0581		0.2143		0.1354
rs2307710	-	100	CATTTGCTAAATTTGCAGAGCCCATACCTACTGTGACCAGAGAAGGAAGGAGCCTGAGATGTCATTAATTGAATTCATTCATTTTGTAGACAAGCCTAAC	0.4459		0.2674		0.3776		0.2041	
		104	CATTTGCTAAATTTGCAGAGCCCATACCTACTGTGACCAGAGAAGGAAGGAAGGAGCCTGAGATGTCATTAATTGAATTCATTCATTTTGTAGACAAGCCTAAC	0.5541		0.7326		0.6224		0.7959	
rs2307839	-	100	CTAAGAAATGTTTTAAGAAAAATTGTCAGCTAATAATGGTTATTCTTATTTGGAAATATATATACATAGATAAATACGTAAAGAAATGCACAGAGAAAGACAC	0.1944		0.4302		0.2347		0.2755	
		102	CTAAGAAATGTTTTAAGAAAAATTGTCAGCTAATAATGGTTATTCTTATTTCTGGAAATATATATACATAGATAAATACGTAAAGAAATGCACAGAGAAAGACAC	0.8056		0.5698		0.7653		0.7245	
rs2307850[1]		96	TCTGGAAAAGCTTTTACTGAGATGCCCAACCTGCGTGGAAAGCCTCCACCTTGCAGCATAAGTGCGGAGGAAACGGTGACAAAATCATTGCCCCGC	0.4324		0.2907		0.3367		0.3889	
		100	TCTGGAAAAGCTTTTACTGAGATGCCCAACCTGCGTGGAAAGCCTCCACCCACCTTGCAGCATAAGTGCGGAGGAAACGGTGACAAAATCATTGCCCCGC	0.5676		0.7093		0.6633		0.6111	
rs2307978[3] -	rs188547 G	105	CTAAGAAATGTTTTAAGAAAAATTGTCAGCTAATAATGGTTATTCTTATTTGGAAATATATATACATAGATAAATACGTAAAGAAATGCACAGAGAAAGACA**C**	0.3919	0.3919	0.4070	0.407	0.1735	0.1735	0.3553	0.3553
		107	CTAAGAAATGTTTTAAGAAAAATTGTCAGCTAATAATGGTTATTCTTATTTCTGGAAATATATATACATAGATAAATACGTAAAGAAATGCACAGAGAAAGACA**C**	0.6081	0.3514	0.5930	0.3256	0.8265	0.4694	0.6447	0.4079

		107	CTAAGAAATGTTTTAAGAAAAATTGTCAGCTAATAATGGTTATTCTTATTTCTGGAAATATATATACATAGATAAATACGTAAAGAAATGCACAGAGAAAGACA**G**		0.2568		0.2674		0.3571		0.2368
rs2308112	-	95	ATCCACAGAAGAGGCGGTGCTGATGGACTGGACCCCAGAGAGTCCTGCTTGGTGCCATACCTTGGGTCCCTCCAGATCAGGATGCTCCCAGCAAC		0.5405		0.5000		0.6020		0.5256
		100	ATCCACAGAAGAGGCGGTGCTGATGGACTGGACCCCAGAGAGTCCACACCTGCTTGGTGCCATACCTTGGGTCCCTCCAGATCAGGATGCTCCCAGCAAC		0.4595		0.5000		0.3980		0.4744
rs2308137	-	98	GGCAAGTTGGGAGTTTCCCGTTGAGCTTTGCTTCCCGTGAAAACAACCATGAGAACTCCAGAAATGGCCTTCTCAGGCCAGGCTTCGTGACTTTCAGG		0.7027		0.4535		0.3061		0.2692
		100	GGCAAGTTGGGAGTTTCCCGTTGAGCTTTGCTTCCCGTGAAAACAACCATGAGAGAACTCCAGAAATGGCCTTCTCAGGCCAGGCTTCGTGACTTTCAGG		0.2973		0.5465		0.6939		0.7308
rs2308171[1]		100	GGTAAGTGTTGCAACATCTGCCACAATTTAAAACACTTACTCAGGTGCCTGCTCAGTGCCTATGCACTAAGTCAAAGAGGCACAGGCCCAGCTGTCCAGAA		0.4730		0.0814		0.2292		0.1771
		104	GGTAAGTGTTGCAACATCTGCCACAATTTAAAACACTTACTCAGGTGCCTGTCTGCTCAGTGCCTATGCACTAAGTCAAAGAGGCACAGGCCAGCTGTCCAGAA		0.5270		0.9186		0.7708		0.8229
-		99	**T**ACATCCTGCTCCTTTCACCTGGAACCTGATTCA**C**GCTGCTGGAATCATTAGTTGTTTCGAATACTGTTTTCAATTTTGTTCTTGTCGTTCCTTTGTTT		0.1757		0.4643		0.3438		0.5405
		99	**C**ACATCCTGCTCCTTTCACCTGGAACCTGATTCA**C**GCTGCTGGAATCATTAGTTGTTTCGAATACTGTTTTCAATTTTGTTCTTGTCGTTCCTTTGTTT	0.5405	0.3108	0.4643	0	0.3542	0.0104	0.5541	0.0135
rs2308189[1,3]	rs176295 C rs176294 T	99	**T**ACATCCTGCTCCTTTCACCTGGAACCTGATTCA**T**GCTGCTGGAATCATTAGTTGTTTCGAATACTGTTTTCAATTTTGTTCTTGTCGTTCCTTTGTTT		0.0541		0		0		0
		104	**C**ACATCCTGCTCCTTTCACCTGGAACCTGATTCA**C**GCTGCTGGAATCATTAGTTAGTTGTTTCGAATACTGTTTTCAATTTTGTTCTTGTCGTTCCTTTGTTT		0.1486		0.3333		0.2813		0.2973
		104	**C**ACATCCTGCTCCTTTCACCTGGAACCTGATTCA**T**GCTGCTGGAATCATTAGTTAGTTGTTTCGAATACTGTTTTCAATTTTGTTCTTGTCGTTCCTTTGTTT	0.4595	0.3108	0.5357	0.2024	0.6458	0.3646	0.4459	0.1486
rs2308196	-	100	TTGAGAGATAAGTGTCTTGTAAAAATGATTTTGTGTTTCTGCAAACTTGTTACTGGAGACCATGGAATTTCCTTTTTCTTTTCTTCACTGTGGTGGAAAA		0.6757		0.6512		0.5714		0.6531
		104	TTGAGAGATAAGTGTCTTGTAAAAATGATTTTGTGTTTCTGCAAACTTGATTGTTACTGGAGACCATGGAATTTCCTTTTTCTTTTCTTCACTGTGGTGGAAAA		0.3243		0.3488		0.4286		0.3469
rs2308232[3]	rs1093240 C	100	TTAATCTAAATGGTGAATGATTGATGCAATCTCACACTAC**C**TTTGAAAAGCTATTCGTCTAGCTCTTGATCAAATTTAAGATCATATAATACGAAGCAAG		0.2432		0.0976		0.1979		0.0921
		100	TTAATCTAAATGGTGAATGATTGATGCAATCTCACACTAC**T**TTTGAAAAGCTATTCGTCTAGCTCTTGATCAAATTTAAGATCATATAATACGAAGCAAG	0.2568	0.0135	0.2683	0.1707	0.2917	0.0938	0.2895	0.1974
		106	TTAATCTAAATGGTGAATGATTGATGCAATCTCACACTAC**T**TTTGAAAAGCTAAACTTATTCGTCTAGCTCTTGATCAAATTTAAGATCATATAATACGAAGCAAG	0.7432	0.7432	0.7317	0.7317	0.7083	0.7083	0.7105	0.7105
rs2308242	chr3:8616681 T[2]	100	AGGAGAGCTCTGCGGAGTTCAT**T**TTCTGCTGATAAGGGGACAGGGAGAAAGCAGACGGGCAGTCATCCTTGGCCCACACAAGGGGACCATCCAGTGCCAA	0.3378	0.3378	0.2674	0.2674	0.2083	0.2083	0.2250	0.2250
		102	AGGAGAGCTCTGCGGAGTTCAT**T**TTCTGCTGATAAGGGGACAGGGAGAAAGAGCAGACGGGCAGTCATCCTTGGCCCACACAAGGGGACCATCCAGTGCCAA		0.6622		0.7326		0.7813		0.7750
		102	AGGAGAGCTCTGCGGAGTTCAT**C**TTCTGCTGATAAGGGGACAGGGAGAAAGAGCAGACGGGCAGTCATCCTTGGCCCACACAAGGGGACCATCCAGTGCCAA	0.6622	0	0.7326	0	0.7917	0.0104	0.7750	0
rs2308276[1,3]	rs10209911 T	100	CAAGTATATACCATTTCATTGTGACTCAATACTTTATAAAGATGAATTTAAATTTTGAAGTAAACTT**T**TTTTGTCTCAAATGGAGAAATTCCATGCAATA	0.5676	0.5676	0.4070	0.407	0.4694	0.4694	0.4487	0.4487
		105	CAAGTATATACCATTTCATTGTGACTCAATACTTTATAAAGATGAATTTAATTTAAATTTTGAAGTAAACTT**T**TTTTGTCTCAAATGGAGAAATTCCATGCAATA		0.3919		0.1977		0.5306		0.5128
		105	CAAGTATATACCATTTCATTGTGACTCAATACTTTATAAAGATGAATTTAATTTAAATTTTGAAGTAAACTT**C**TTTTGTCTCAAATGGAGAAATTCCATGCAATA	0.4324	0.0405	0.5930	0.3953	0.5306	0	0.5513	0.0385
rs2308292[3]	rs147933644 A rs140159023 A rs396196 C	95	GACCATGCTTTATATATTCTTA**A**AATT**A**TTGCAAACATTATATTTACTTTAAAGTTTCTGTGACACAGTG**T**GCCTAACAGATAGTGGGAATTTTA	0.5479	0.5405	0.4419	0.4419	0.2959	0.2857	0.3163	0.3163
		95	GACCATGCTTTATATATTCTTA**A**AATT**A**TTGCAAACATTATATTTACTTTAAAGTTTCTGTGACACAGTG**C**GCCTAACAGATAGTGGGAATTTTA		0		0		0.0102		0
		100	GACCATGCTTTATATATTCTTA**A**AATT**A**TTGCAAACATTATATTTACTTTTAAGTAAAGTTTCTGTGACACAGTG**C**GCCTAACAGATAGTGGGAATTTTA		0.3378		0.1977		0.398		0.2653
		100	GACCATGCTTTATATATTCTTA**A**AATT**A**TTGCAAACATTATATTTACTTTTAAGTAAAGTTTCTGTGACACAGTG**T**GCCTAACAGATAGTGGGAATTTTA		0.1081		0.3605		0.2551		0.3980
		100	GACCATGCTTTATATATTCTTA**A**AATT**T**TTGCAAACATTATATTTACTTTTAAGTAAAGTTTCTGTGACACAGTG**C**GCCTAACAGATAGTGGGAATTTTA	0.4521	0	0.5581	0	0.7041	0.0408	0.6837	0.0204
		100	GACCATGCTTTATATATTCTTA**C**AATT**A**TTGCAAACATTATATTTACTTTTAAGTAAAGTTTCTGTGACACAGTG**T**GCCTAACAGATAGTGGGAATTTTA		0.0135		0		0.0102		0
rs28923216	rs184394929 T	100	GTTTATAGTTTTGAAAGTGAATTGATCACTTTGTTTCTTGCTCTGAATCTTAATTTTTTTTCCTAAAGGAAAAAGATAATTTAC**T**TTTTATAGAGCAAAA	0.6757	0.6757	0.4405	0.4405	0.5408	0.5408	0.5417	0.5417
		105	GTTTATAGTTTTGAAAGTGAATTGATCACTTTGTTTCTTGCTCTGAATCTTTTGTAAATTTTTTTTCCTAAAGGAAAAAGATAATTTAC**T**TTTTATAGAGCAAAA		0.2973		0.5595		0.4592		0.4583
		105	GTTTATAGTTTTGAAAGTGAATTGATCACTTTGTTTCTTGCTCTGAATCTTTTGTAAATTTTTTTTCCTAAAGGAAAAAGATAATTTAC**G**TTTTATAGAGCAAAA	0.3243	0.0270	0.5595	0	0.4592	0	0.4583	0

rs3038530[3]	rs1923740 C rs115288378 C	104	**A**CT**C**TGGCAATTACTTAAGGCTATCATTCATTCTGCAGAAATCGCTTTGTAAATCAGCTTATTTGGCTCAGTTTATATTTGAAAATTTTGATATGGAGGAATTC	0.3514	0.3514	0.3605	0.3605	0.3367	0.3367	0.4615	0.4615
		108	**C**CT**C**TGGCAATTACTTAAGGCTATCATTCATTCTGCAGAAATCGCTTTGTAAATCAGCTTGATTATTTGGCTCAGTTTATATTTGAAAATTTTGATATGGAGGAATTC		0.5270		0.4186		0.2959		0.2308
		108	**A**CT**C**TGGCAATTACTTAAGGCTATCATTCATTCTGCAGAAATCGCTTTGTAAATCAGCTTGATTATTTGGCTCAGTTTATATTTGAAAATTTTGATATGGAGGAATTC	0.6486	0.0946	0.6395	0.2209	0.6633	0.3673	0.5385	0.3077
		108	**C**CT**T**TGGCAATTACTTAAGGCTATCATTCATTCTGCAGAAATCGCTTTGTAAATCAGCTTGATTATTTGGCTCAGTTTATATTTGAAAATTTTGATATGGAGGAATTC		0.0270		0		0		0
rs3042783[3]	chr2:222160737 T[2] rs3943815 C	100	TTCTGTTGTGGTTAGGAGGGATATTGACC**T**GAAGGCA**T**ACTGTTCCCTCTTAAAGAAGCAAGATCAACATTCAGAATCGCTCAGTGACAAATCTCAATTG	0.8169	0.1081	0.6163	0.5698	0.7245	0.602	0.4898	0.3878
		100	TTCTGTTGTGGTTAGGAGGGATATTGACC**T**GAAGGCA**C**ACTGTTCCCTCTTAAAGAAGCAAGATCAACATTCAGAATCGCTCAGTGACAAATCTCAATTG		0.6757		0.0465		0.1224		0.102
		105	TTCTGTTGTGGTTAGGAGGGATATTGACC**T**GAAGGCA**C**ACTGTTCCCTCTTAACTCAAAGAAGCAAGATCAACATTCAGAATCGCTCAGTGACAAATCTCAATTG	0.1831	0.2162	0.3837	0.3837	0.2755	0.2653	0.5102	0.5102
		105	TTCTGTTGTGGTTAGGAGGGATATTGACC**C**GAAGGCA**C**ACTGTTCCCTCTTAACTCAAAGAAGCAAGATCAACATTCAGAATCGCTCAGTGACAAATCTCAATTG		0		0		0.0102		0
rs3045264	rs183114846 G	100	AACAAAGTTTCCAAGGGCTCTTACCTACGTGGTTGGTGACCTTTTGGGGCTAATAATTAAGAAGGATGGGTATATATGAAATTATTTAAATCTCCATC**G**G	0.2838	0.2838	0.3571	0.3571	0.3367	0.3367	0.3293	0.3293
		104	AACAAAGTTTCCAAGGGCTCTTACCTACGTGGTTGGTGACCTTTTGGGGCTGTCTAATAATTAAGAAGGATGGGTATATATGAAATTATTTAAATCTCCATC**G**G	0.7162	0.7162	0.6429	0.6429	0.6633	0.6531	0.6707	0.6707
		104	AACAAAGTTTCCAAGGGCTCTTACCTACGTGGTTGGTGACCTTTTGGGGCTGTCTAATAATTAAGAAGGATGGGTATATATGAAATTATTTAAATCTCCATC**A**G		0		0		0.0102		0
rs3047269[1]		100	CACAGAGAGGCGGTGGGAGACAGGCACACCAGCATGCAAGTACAGTGCACTGGGATCTCTCTCACTAGATACAAGGAAGGTTTGGCAAATTAGTTTGTTT	0.6250		0.6429		0.4479		0.5500	
		104	CACAGAGAGGCGGTGGGAGACAGGCACACCAGCATGCAAGTACAGTGCACTGACTGGGATCTCTCTCACTAGATACAAGGAAGGTTTGGCAAATTAGTTTGTTT	0.3750		0.3571		0.5521		0.4500	
rs3051300[1] -	rs186936660 A	100	TTCTATTTAGAATTGGAAGATCTGGGTAGAAGTCCTATCTAGTCCATGTATACATTGTGTGAGACTGGGCAATATTGAATCTCCTAAATGTGGACAC**A**AA	0.1892	0.1892	0.3837	0.3837	0.5208	0.5208	0.3617	0.3617
		104	TTCTATTTAGAATTGGAAGATCTGGGTAGAAGTCCTATCTAGTCCATGTATGTATACATTGTGTGAGACTGGGCAATATTGAATCTCCTAAATGTGGACAC**A**AA		0.8108		0.6163		0.4792		0.6277
		104	TTCTATTTAGAATTGGAAGATCTGGGTAGAAGTCCTATCTAGTCCATGTATGTATACATTGTGTGAGACTGGGCAATATTGAATCTCCTAAATGTGGACAC**G**AA	0.8108	0	0.6163	0	0.4792	0	0.6383	0.0106
rs3062629	-	100	AATTCACCAGAATTTAAAATACTGCTGGGTATGATGTCACATGTCTGTAGTCCCTAGCTACTTGAGAGGCTGAGATGGGAGGATTCCTTGAGTCTAGGAA	0.3108		0.6977		0.3958		0.6531	
		105	AATTCACCAGAATTTAAAATACTGCTGGGTATGATGTCACATGTCTGTACTGTAGTCCCTAGCTACTTGAGAGGCTGAGATGGGAGGATTCCTTGAGTCTAGGAA	0.6892		0.3023		0.6042		0.3469	
rs3080855	-	100	GTATATTTATAAATTTCATGTAAGTAGATATTCTAAATACTAGTAAAGCTTGTTTTCATTTTGTGTTTTAAAAAAGAATTATGATATTTTTCTCCATGCC	0.2568		0.4651		0.3673		0.3776	
		104	GTATATTTATAAATTTCATGTAAGTAGATATTCTAAATACTAGTAAAGCTTAATTGTTTTCATTTTGTGTTTTAAAAAAGAATTATGATATTTTTCTCCATGCC	0.7432		0.5349		0.6327		0.6224	
rs33917182	-	100	GCTGTGCAGAGAGAGAGTAGGGGGAGGAGGTGGAGAACCCGGAAGAGCAGCTCTGAGCAGATTTCACACGATTAATAGGGGGTTTGTTGGCTGGTGGACT	0.5405		0.5465		0.5938		0.7105	
		102	GCTGTGCAGAGAGAGAGTAGGGGGAGGAGGTGGAGAACCCGGAAGAGCAGCCATCTGAGCAGATTTCACACGATTAATAGGGGGTTTGTTGGCTGGTGGACT	0.4595		0.4535		0.4063		0.2895	
rs33951431[3]	rs4741748 G	96	GGGACATAACAAAGCCTCGGCGATAGACAGCATTTGTGAGTTACATCACATGTTAGAACTTAATAGTTCTCCCACATGCTGTGA**A**TAAACAGGGTT	0.7042	0.3514	0.6163	0.5698	0.5816	0.5612	0.6224	0.5816
		96	GGGACATAACAAAGCCTCGGCGATAGACAGCATTTGTGAGTTACATCACATGTTAGAACTTAATAGTTCTCCCACATGCTGTGA**G**TAAACAGGGTT		0.3243		0.0465		0.0204		0.0408
		100	GGGACATAACAAAGCCTCGGCGATAGACAGCATTTGTGAGTTACATCACAAGTTTGTTAGAACTTAATAGTTCTCCCACATGCTGTGA**A**TAAACAGGGTT	0.2958	0.3243	0.3837	0.3605	0.4184	0.4184	0.3776	0.3776
		100	GGGACATAACAAAGCCTCGGCGATAGACAGCATTTGTGAGTTACATCACAAGTTTGTTAGAACTTAATAGTTCTCCCACATGCTGTGA**G**TAAACAGGGTT		0		0.0233		0		0
rs34051577[1]		100	CTGTCATAACAATAATGAGTCATCCAGATTATCGAGTGAGATACATATTTAAGAATTATCTTTAAAAATTTCAAAAATTTTAATTTTACTGTTGTGTTTT	0.4189		0.3837		0.6837		0.5000	
		105	CTGTCATAACAATAATGAGTCATCCAGATTATCGAGTGAGATACATATTTATCTTAAGAATTATCTTTAAAAATTTCAAAAATTTTAATTTTACTGTTGTGTTTT	0.5811		0.6163		0.3163		0.5000	
rs34495360 -	-	100	CCAGTTCTGTGGTTTGGTCTCAGTACTTGTTCTGTGCAGTAGTTAAGTCCTTCCAGCCATCTTGCCACTCAGCTTAGAAAAATACTCTCCAAAAACATGT	0.6944		0.5349		0.6224		0.4474	
		105	CCAGTTCTGTGGTTTGGTCTCAGTACTTGTTCTGTGCAGTAGTTAAGTCAAGTCCTTCCAGCCATCTTGCCACTCAGCTTAGAAAAATACTCTCCAAAAACATGT	0.3056		0.4651		0.3776		0.5526	
rs34510056[1]		95	ACAAAATATTCTGAATAGATCCCGGCCCAAAGTCATTTGATTTGGGAATAGTCTTAAAAACAGGCAGGCATACTGTTATTAACATTGTCATTATC	0.4189		0.8023		0.4796		0.4796	
		100	ACAAAATATTCTGAATAGATCCCGGCCCAAAGTCATTTGATTTGGGAATACTTTAGTCTTAAAAACAGGCAGGCATACTGTTATTAACATTGTCATTATC	0.5811		0.1977		0.5204		0.5204	

-

rs34511541[3]	rs57941925 T	95	AAGCTTATGAGATTTGGAGGACTTTAGTAGAAGAGGAAAATACCACATTTATTTATGAGTGTCT**T**TGAACCTAAGAAGGGTCTCATTTGTATACA	0.4054	0.4054	0.5000	0.5000	0.3265	0.3265	0.5102	0.5102
		100	AAGCTTATGAGATTTGGAGGACTTTAGTAGAAGAGGAAAATACCACATTTCTCTTATTTATGAGTGTCT**T**TGAACCTAAGAAGGGTCTCATTTGTATACA	0.5946	0.473	0.5000	0.3953	0.6735	0.4286	0.4898	0.3673
		100	AAGCTTATGAGATTTGGAGGACTTTAGTAGAAGAGGAAAATACCACATTTCTCTTATTTATGAGTGTCT**C**TGAACCTAAGAAGGGTCTCATTTGTATACA		0.1216		0.1047		0.2449		0.1224
rs34528025[1]	rs34247791 DEL rs202051643 G	100	GTCTCTAGCGTAGAAAGAGGAAATTTGACCCATGTCTTGGAGAGGAGTCAAATCAGAAACTCCTCCTATTAC**G**TCTTTTTCCTTTGCTCGTTTTGCTTTG	0.4324	0.4324	0.5581	0.5581	0.3469	0.3469	0.3163	0.3163
		106	GTCTCT**CT**AGCGTAGAAAGAGGAAATTTGACCCATGTCTTGGAGAGGAGTCAAATCAGAAACTCCTCCTGAGTATTAC**G**TCTTTTTCCTTTGCTCGTTTTGCTTTG	0.5676	0.5676	0.4419	0.4419	0.6531	0.6531	0.6837	0.5816
		106	GTCTCT**CT**AGCGTAGAAAGAGGAAATTTGACCCATGTCTTGGAGAGGAGTCAAATCAGAAACTCCTCCTGAGTATTAC**A**TCTTTTTCCTTTGCTCGTTTTGCTTTG		0		0		0		0.102
rs34535242[1]		100	AGGGGGTACTACAGACAGGTTTAAAATGAGCAAACCTAGCTGGTAGGTAGTGTCCTTAGAAGAGTTTTAAGTGAAAAAGGACATGATAAAATATGGCTTT	0.5811		0.4651		0.6122		0.5395	
		104	AGGGGGTACTACAGACAGGTTTAAAATGAGCAAACCTAGCTGGTAGGTAGGTAGTGTCCTTAGAAGAGTTTTAAGTGAAAAAGGACATGATAAAATATGGCTTT	0.4189		0.5349		0.3878		0.4605	
rs34541393 -	-	96	TACATTTCTAGATGTGTCAGGAGTCTAGAAACTTCAGTTTGGAGAATAACTACTTCCCTCACATCATTGTTCATACTGTTTTGGTTTTTATTATAA	0.4459		0.7442		0.4082		0.5395	
		100	TACATTTCTAGATGTGTCAGGAGTCTAGAAACTTCAGTTTGGAGAATAAAACTCTACTTCCCTCACATCATTGTTCATACTGTTTTGGTTTTTATTATAA	0.5541		0.2558		0.5918		0.4605	
rs34795726	rs189603436 G rs4646566 G	100	ATAATGTAGAGTTATTCAAAAAAAAG**G**TCTTTTAGAAATTCTTTTTAAATTATTTGCTACCTATCCATGTTTTCTCCAAATCTATCAGCAGCACAGAGT**G**	0.5946	0.5811	0.6977	0.6977	0.4592	0.4490	0.6042	0.6042
		100	ATAATGTAGAGTTATTCAAAAAAAAG**T**TCTTTTAGAAATTCTTTTTAAATTATTTGCTACCTATCCATGTTTTCTCCAAATCTATCAGCAGCACAGAGT**G**		0		0		0.0102		0
		100	ATAATGTAGAGTTATTCAAAAAAAAG**G**TCTTTTAGAAATTCTTTTTAAATTATTTGCTACCTATCCATGTTTTCTCCAAATCTATCAGCAGCACAGAGT**A**		0.0135		0		0		0
		104	ATAATGTAGAGTTATTCAAAAAAAAG**G**TCTTTTAGAAATTCTTTTTAAATTATTTGCTACCTATCCATGTTTTCTCCAAATCTATCAGCAGCACAGAGT**A**	0.4054	0.4054	0.3023	0.3023	0.5408	0.5408	0.3958	0.3958
rs34811743[3]	rs532272 C	100	TATGTCTCTACATCCCACCCCAA**C**TACAACACTTCGTACCCAGGATGCAACAGATCAAAGTAGTTGCTTACTATGGGTTGAACAAAAAGGAGAGGCACAC	0.5946	0.527	0.8095	0.8095	0.6633	0.6633	0.7604	0.7292
		100	TATGTCTCTACATCCCACCCCAA**A**TACAACACTTCGTACCCAGGATGCAACAGATCAAAGTAGTTGCTTACTATGGGTTGAACAAAAAGGAGAGGCACAC		0.0676		0		0		0.0313
		102	TATGTCTCTACATCCCACCCCAA**C**TACAACACTTCGTACCCAGGATGCAACTGAGATCAAAGTAGTTGCTTACTATGGGTTGAACAAAAAGGAGAGGCACAC	0.4054	0.4054	0.1905	0.1905	0.3367	0.3367	0.2396	0.2396
rs35605984	rs150571926 C	100	TATGTCATAGTAAAAA**C**TTGGAAATAATAAGATGTTGAATAATTGACATTATTAAATTATGCTACATTAGCATAATAAAATATTAGGTAGTTATTTTTAA	0.5541	0.5541	0.4167	0.4167	0.5204	0.5204	0.5816	0.5816
		105	TATGTCATAGTAAAAA**C**TTGGAAATAATAAGATGTTGAATAATTGACATTACTTTATTAAATTATGCTACATTAGCATAATAAAATATTAGGTAGTTATTTTTAA	0.4459	0.4459	0.5833	0.5833	0.4796	0.4694	0.4184	0.4184
		105	TATGTCATAGTAAAAA**A**TTGGAAATAATAAGATGTTGAATAATTGACATTACTTTATTAAATTATGCTACATTAGCATAATAAAATATTAGGTAGTTATTTTTAA		0		0		0.0102		0
rs35716687[1]		97	GTCATGCCATCATTAGGGGACTAAATGTGTTAATATCCTGAAAATTATAAGTAATCAATAATTTCTCTTTCGTGATACACCTTGTTTTGAAATATTT	0.6351		0.6395		0.6020		0.6020	
		101	GTCATGCCATCATTAGGGGACTAAATGTGTTAATATCCTGAAAATTATACTTAAGTAATCAATAATTTCTCTTTCGTGATACACCTTGTTTTTGAAATATTT	0.3649		0.3605		0.3980		0.3980	
rs35769550[1] -	rs1449554 A	100	**A**CTGCGTTTCTGTAGAGGAGTAAATGTACTAAGACTATTAAATAACTTACACCTTAACTAAAACTTTTAGGTGGAAACAAAAGACTGGTTAGAAAAAATG	0.0946	0.0946	0.4767	0.4767	0.4694	0.4694	0.6410	0.6410
		104	**G**CTGCGTTTCTGTAGAGGAGTAAATGTACTAAGACTATTAAATAACTTACTGACACCTTAACTAAAACTTTTAGGTGGAAACAAAAGACTGGTTAGAAAAAATG	0.9054	0.8919	0.5233	0.5233	0.5306	0.5306	0.3590	0.3590
		104	**A**CTGCGTTTCTGTAGAGGAGTAAATGTACTAAGACTATTAAATAACTTACTGACACCTTAACTAAAACTTTTAGGTGGAAACAAAAGACTGGTTAGAAAAAATG		0.0135		0		0		0
rs36040336	-	88	CACGGGTTAACAGATGCAGTTATTATGCCCATTTAACACGAGGGAAACTGAGGCCCAGAGAGGTTGAGGTTCACAGGTTGCAGCAGGG	0.4054		0.8023		0.7959		0.6735	
		90	CACGGGTTAACAGATGCAGTTATTATGCCCATTTAACACATGAGGGAAACTGAGGCCCAGAGAGGTTGAGGTTCACAGGTTGCAGCAGGG	0.5946		0.1977		0.2041		0.3265	
rs36062169[1]	rs114264449 C	94	TTAGGGTTTCCTGTCAACTATTCTACTGCCATTTACCACAGGGTCACCACATTCTAATAAGTCCATCCTT**C**TGAGATATCCTCTTCCTAACATG	0.5270	0.527	0.3023	0.3023	0.5714	0.5714	0.5513	0.5513
		100	TTAGGGTTTCCTGTCAACTATTCTACTGCCATTTACCACAGGGTCACCACCAGTACATTCTAATAAGTCCATCCTT**C**TGAGATATCCTCTTCCTAACATG	0.4730	0.4054	0.6977	0.686	0.4286	0.4286	0.4487	0.4487
		100	TTAGGGTTTCCTGTCAACTATTCTACTGCCATTTACCACAGGGTCACCACCAGTACATTCTAATAAGTCCATCCTT**A**TGAGATATCCTCTTCCTAACATG		0.0676		0.0116		0		0
rs3838581[1]	rs371883530 C	96	GATTA**C**TGGTGTTTACTTTTAATTCCAATAAATTAAAAGTTCTACTGTTTTTCTACTTCCTCATACAAATCTTGAGCAAGACAAACTTTAACATTC	0.3108	0.2973	0.3372	0.3372	0.3163	0.3163	0.5000	0.5000
		96	GATTA**T**TGGTGTTTACTTTTAATTCCAATAAATTAAAAGTTCTACTGTTTTTCTACTTCCTCATACAAATCTTGAGCAAGACAAACTTTAACATTC		0.0135		0		0		0

		100	GATTA**C**TGGTGTTTACTTTTAATTCCAATAAATTAAAAGTTCTACTGTTTGTTATTCTACTTCCTCATACAAATCTTGAGCAAGACAAACTTTAACATTC	0.6892	0.6892	0.6628	0.6628	0.6837	0.6837	0.5000	0.5000
rs3841948[1]	rs76509761 G	95	AAGTGATCCAGATTTGGTCTTTTACTGTGAAAATGCTTTTATACAATTTAGTAGAGATGTTATGCAATT**G**TACTATATCCTTTGCACACTGGAAT	0.3784	0.3784	0.5116	0.5116	0.3878	0.3878	0.2949	0.2949
		100	AAGTGATCCAGATTTGGTCTTTTACTGTGAAAATGCTTTTATACAATTTAATTTAGTAGAGATGTTATGCAATT**G**TACTATATCCTTTGCACACTGGAAT	0.6216	0.5405	0.4884	0.4884	0.6122	0.6122	0.7051	0.7051
		100	AAGTGATCCAGATTTGGTCTTTTACTGTGAAAATGCTTTTATACAATTTAATTTAGTAGAGATGTTATGCAATT**T**TACTATATCCTTTGCACACTGGAAT		0.0811		0		0		0
rs4187	-	100	ATGATTAACAAAAAAACAAGTAGAAAAATAAGAGAGTGTATTTAAAAAAAATAATCAAATGCTTTTTGAAAGACCTGTTCTCTTCACTGCCACACATATT	0.6892		0.5233		0.5204		0.5208	
		106	ATGATTAACAAAAAAACAAGTAGAAAAATAAGAGAGTGTATTTAAAAAAAATAAAGATAATCAAATGCTTTTTGAAAGACCTGTTCTCTTCACTGCCACACATATT	0.3108		0.4767		0.4796		0.4792	
rs4646006	rs562172870 G	100	TGTAAGTCTAAACAATCAGGCACGTGGGCAGCAATGGAGCTGCAGGT**G**CACTGTGTGCCATTTACCAGCCTTTGCTGATCTGTTCATTATTTTGCAGGGC	0.1622	0.1622	0.4070	0.3953	0.4271	0.4271	0.5366	0.5366
		100	TGTAAGTCTAAACAATCAGGCACGTGGGCAGCAATGGAGCTGCAGGT**A**CACTGTGTGCCATTTACCAGCCTTTGCTGATCTGTTCATTATTTTGCAGGGC		0		0.0116		0		0
		104	TGTAAGTCTAAACAATCAGGCACGTGGGCAGCAATGGAGCTGCAGGT**G**CACTGAGTGTGTGCCATTTACCAGCCTTTGCTGATCTGTTCATTATTTTGCAGGGC	0.8378	0.8378	0.5930	0.593	0.5729	0.5729	0.4634	0.4634
rs5895446[3]	rs2960102 G	108	**G**GGAGAGATATAGAGTTACTTTGTATCCTGCCACTATCACTGGGGAGATATGTTGGACAGAGTTCTATCGTGCAAAGTTAAGTGAAAGAGGTTCTAAGGAGATTGTTC	0.3243	0.3243	0.3256	0.3256	0.3673	0.3673	0.3265	0.3265
		110	**G**GGAGAGATATAGAGTTACTTTGTATCCTGCCACTATCACTGGGGAGATATGTTGGACACAGAGTTCTATCGTGCAAAGTTAAGTGAAAGAGGTTCTAAGGAGATTGTTC	0.6757	0.5811	0.6744	0.2326	0.6327	0.2551	0.6735	0.3469
		110	**A**GGAGAGATATAGAGTTACTTTGTATCCTGCCACTATCACTGGGGAGATATGTTGGACACAGAGTTCTATCGTGCAAAGTTAAGTGAAAGAGGTTCTAAGGAGATTGTTC		0.0946		0.4419		0.3776		0.3265
rs60901515[3]	rs9790699 C	100	TGCCTTATGCAATTTAAGCAACAATAGAAGACAAGTCAGGAACTGAGACTTATCTATTGAAACT**C**AGGAGTGCTTGGTATCCACAGTGGCAGATAAATTC	0.6389	0.5972	0.6860	0.6628	0.6735	0.6735	0.5976	0.5854
		100	TGCCTTATGCAATTTAAGCAACAATAGAAGACAAGTCAGGAACTGAGACTTATCTATTGAAACT**T**AGGAGTGCTTGGTATCCACAGTGGCAGATAAATTC		0.0417		0.0233		0		0.0122
		104	TGCCTTATGCAATTTAAGCAACAATAGAAGACAAGTCAGGAACTGAGACTTACTTATCTATTGAAACT**C**AGGAGTGCTTGGTATCCACAGTGGCAGATAAATTC	0.3611	0.3611	0.3140	0.3140	0.3265	0.3265	0.4024	0.4024

[1]Motif different from LaRue, et al., Pereira, et al., and dbSNP. Sequences confirmed with IGV .
[2]Due to lack of RS number for the observed SNP, the hg19 locus coordinates are provided.
[3]One of twenty-two INDELs, with substantial sequence variation, that are recommended for future HID INDEL panels.

Table 6. Insertion/deletion loci that are part of a short tandem repeat (STR) motif. The repeat motif for each locus is underlined and italicized letters indicate sequence that is not captured as part of the STRait Razor flank but was identified manually using the Integrative Genomics Viewer (IGV).

Locus rs#	STRait Razor Sequence for Insertions	Number of Repeat Motifs	
rs1160886	TAGTACTAC	2	
rs1160956	AAAGAAGAGCAAC	2	
rs16402	ATTAATTATTTATT	2	
rs16458	TTTTACAATTCCTTCCTTC	2	
rs17859968	GGCACATAAATAAA	2	
rs2067208	AAAGAGCCTGGCCTG	2	
rs2307580	TAATTAATTGAATA	2	
rs2307689	GGCTGTTCTTCTTC	3	
rs2307710	CCAGAGAAGGAAGGAAGGA	3	
rs2307839	TGAGAGAACAAC	3	
rs2307850	AGCCTCCACCCACC	2	
rs2308276	GATGAATTTAATTTAAA	2	
rs3051300	AGTCCATGTATGTA	2	
rs34535242	TAGCTGGTAGGTAGGTAG	3	
rs3841948	TATACAATTTAATTT	2	

Table 7. Length-based (LB) and sequence-based (SB) observed (H_o) and expected (H_e) heterozygosities in four major US population groups for 42 INDEL loci that exhibited sequence

variation. The change in H_o and H_e as a result of utilizing SB alleles is indicated by ΔH_o and ΔH_e, respectively.

Locus	AFA						ASA					
	LB H_e	SB H_e	ΔH_e	LB H_o	SB H_o	ΔH_o	LB H_e	SB H_e	ΔH_e	LB H_o	SB H_o	ΔH_o
rs10623496	0.48	0.48	0.00	0.59	0.59	0.00	0.45	0.46	0.01	0.40	0.40	0.00
rs10629077	0.39	0.39	0.01	0.41	0.41	0.00	0.41	0.41	0.00	0.33	0.33	0.00
rs10688868	0.33	0.67	0.35	0.35	0.73	0.38	0.50	0.60	0.10	0.53	0.65	0.12
rs1160956	0.50	0.52	0.01	0.49	0.51	0.03	0.48	0.48	0.00	0.42	0.42	0.00
rs13447508	0.47	0.50	0.03	0.51	0.57	0.05	0.51	0.51	0.00	0.51	0.51	0.00
rs140809	0.49	0.66	0.17	0.35	0.46	0.11	0.47	0.70	0.24	0.40	0.51	0.12
rs1610871	0.51	0.64	0.13	0.51	0.65	0.14	0.47	0.51	0.04	0.37	0.42	0.05
rs16624	0.39	0.43	0.04	0.30	0.35	0.05	0.50	0.51	0.01	0.42	0.42	0.00
rs17859968	0.45	0.54	0.09	0.46	0.57	0.11	0.46	0.50	0.04	0.51	0.53	0.02
rs2067140	0.34	0.64	0.30	0.38	0.65	0.27	0.46	0.52	0.06	0.42	0.49	0.07
rs2067191	0.49	0.51	0.02	0.51	0.54	0.03	0.50	0.50	0.00	0.58	0.58	0.00
rs2067208	0.31	0.31	0.00	0.32	0.32	0.00	0.29	0.35	0.06	0.26	0.28	0.02
rs2307507	0.31	0.32	0.00	0.27	0.27	0.00	0.45	0.45	0.00	0.41	0.41	0.00
rs2307526	0.47	0.67	0.20	0.51	0.70	0.19	0.47	0.50	0.03	0.54	0.54	0.00
rs2307579	0.49	0.50	0.01	0.51	0.51	0.00	0.29	0.29	0.00	0.30	0.30	0.00
rs2307656	0.48	0.55	0.07	0.43	0.49	0.05	0.50	0.58	0.08	0.47	0.56	0.09
rs2307689	0.50	0.59	0.09	0.59	0.62	0.03	0.40	0.66	0.27	0.40	0.72	0.33
rs2307700	0.45	0.67	0.22	0.41	0.59	0.19	0.40	0.48	0.08	0.35	0.42	0.07
rs2307978	0.48	0.67	0.18	0.46	0.68	0.22	0.49	0.66	0.18	0.44	0.60	0.16
rs2308189	0.50	0.76	0.26	0.54	0.81	0.27	0.50	0.64	0.14	0.60	0.74	0.14
rs2308232	0.39	0.39	0.01	0.41	0.41	0.00	0.40	0.43	0.03	0.39	0.44	0.05
rs2308242	0.45	0.45	0.00	0.30	0.30	0.00	0.40	0.40	0.00	0.49	0.49	0.00
rs2308276	0.50	0.53	0.03	0.54	0.57	0.03	0.49	0.65	0.16	0.49	0.77	0.28
rs2308292	0.50	0.59	0.0	0.49	0.59	0.1	0.50	0.64	0.1	0.47	0.60	0.1

			9			1			4			4
rs2892321 6	0.44	0.46	0.0 2	0.38	0.41	0.0 3	0.50	0.50	0.0 0	0.50	0.50	0.0 0
rs3038530	0.46	0.60	0.1 4	0.49	0.65	0.1 6	0.47	0.65	0.1 9	0.58	0.67	0.0 9
rs3042783	0.34	0.49	0.1 5	0.32	0.49	0.1 6	0.48	0.53	0.0 5	0.53	0.56	0.0 2
rs3045264	0.41	0.41	0.0 0	0.51	0.51	0.0 0	0.46	0.46	0.0 0	0.43	0.43	0.0 0
rs3051300	0.31	0.31	0.0 0	0.27	0.27	0.0 0	0.48	0.48	0.0 0	0.44	0.44	0.0 0
rs3395143 1	0.44	0.68	0.2 3	0.38	0.62	0.2 4	0.48	0.55	0.0 7	0.49	0.58	0.0 9
rs3451154 1	0.49	0.61	0.1 2	0.32	0.41	0.0 8	0.51	0.59	0.0 8	0.58	0.65	0.0 7
rs3452802 5	0.50	0.50	0.0 0	0.43	0.43	0.0 0	0.50	0.50	0.0 0	0.51	0.51	0.0 0
rs3479572 6	0.49	0.50	0.0 2	0.27	0.27	0.0 0	0.43	0.43	0.0 0	0.37	0.37	0.0 0
rs3481174 3	0.49	0.56	0.0 7	0.54	0.62	0.0 8	0.31	0.31	0.0 0	0.14	0.14	0.0 0
rs3560598 4	0.50	0.50	0.0 0	0.62	0.62	0.0 0	0.49	0.49	0.0 0	0.45	0.45	0.0 0
rs3576955 0	0.17	0.20	0.0 2	0.19	0.22	0.0 3	0.50	0.50	0.0 0	0.49	0.49	0.0 0
rs3606216 9	0.51	0.56	0.0 6	0.68	0.70	0.0 3	0.43	0.44	0.0 2	0.47	0.49	0.0 2
rs3838581	0.43	0.44	0.0 1	0.41	0.41	0.0 0	0.45	0.45	0.0 0	0.49	0.49	0.0 0
rs3841948	0.48	0.57	0.0 9	0.32	0.46	0.1 4	0.51	0.51	0.0 0	0.60	0.60	0.0 0
rs4646006	0.28	0.28	0.0 0	0.27	0.27	0.0 0	0.49	0.50	0.0 1	0.44	0.47	0.0 2
rs5895446	0.44	0.56	0.1 1	0.49	0.59	0.1 1	0.44	0.65	0.2 1	0.42	0.60	0.1 9
rs6090151 5	0.47	0.52	0.0 5	0.39	0.44	0.0 6	0.44	0.47	0.0 3	0.44	0.47	0.0 2

Locus	CAU						HIS					
	LB H_e	SB H_e	ΔH_e	LB H_o	SB H_o	ΔH_o	LB H_e	SB H_e	ΔH_e	LB H_o	SB H_o	ΔH_o
rs1062349 6	0.46	0.46	0.0 0	0.33	0.33	0.0 0	0.43	0.43	0.0 0	0.41	0.41	0.0 0
rs1062907 7	0.29	0.29	0.0 0	0.31	0.31	0.0 0	0.40	0.40	0.0 0	0.39	0.39	0.0 0
rs1068886 8	0.44	0.67	0.2 3	0.43	0.63	0.2 0	0.47	0.72	0.2 5	0.45	0.76	0.3 1
rs1160956	0.30	0.30	0.0 0	0.33	0.33	0.0 0	0.48	0.48	0.0 0	0.48	0.48	0.0 0
rs1344750 8	0.41	0.41	0.0 0	0.33	0.33	0.0 0	0.48	0.48	0.0 0	0.48	0.48	0.0 0

rs140809	0.50	0.64	0.14	0.44	0.54	0.10	0.28	0.45	0.17	0.24	0.41	0.16
rs1610871	0.49	0.49	0.00	0.53	0.53	0.00	0.51	0.52	0.01	0.38	0.41	0.03
rs16624	0.36	0.38	0.02	0.27	0.29	0.02	0.50	0.51	0.01	0.59	0.59	0.00
rs17859968	0.50	0.51	0.01	0.55	0.55	0.00	0.49	0.52	0.02	0.51	0.55	0.04
rs2067140	0.48	0.55	0.07	0.53	0.57	0.04	0.47	0.55	0.07	0.47	0.53	0.06
rs2067191	0.50	0.50	0.00	0.41	0.41	0.00	0.46	0.46	0.00	0.43	0.43	0.00
rs2067208	0.44	0.59	0.16	0.39	0.55	0.16	0.43	0.50	0.07	0.45	0.51	0.06
rs2307507	0.49	0.49	0.00	0.49	0.49	0.00	0.50	0.50	0.00	0.37	0.37	0.00
rs2307526	0.50	0.60	0.10	0.53	0.63	0.10	0.42	0.63	0.21	0.45	0.73	0.27
rs2307579	0.49	0.49	0.00	0.43	0.43	0.00	0.46	0.46	0.00	0.35	0.35	0.00
rs2307656	0.51	0.61	0.10	0.63	0.71	0.08	0.46	0.52	0.06	0.48	0.50	0.02
rs2307689	0.34	0.62	0.28	0.39	0.67	0.29	0.47	0.67	0.21	0.44	0.56	0.12
rs2307700	0.50	0.61	0.11	0.51	0.63	0.12	0.43	0.59	0.15	0.38	0.52	0.15
rs2307978	0.29	0.63	0.34	0.27	0.67	0.41	0.46	0.66	0.20	0.29	0.53	0.24
rs2308189	0.46	0.68	0.21	0.42	0.71	0.29	0.50	0.61	0.10	0.57	0.62	0.05
rs2308232	0.42	0.46	0.04	0.50	0.52	0.02	0.42	0.45	0.04	0.37	0.42	0.05
rs2308242	0.33	0.35	0.02	0.42	0.42	0.00	0.35	0.35	0.00	0.35	0.35	0.00
rs2308276	0.50	0.50	0.00	0.53	0.53	0.00	0.50	0.54	0.04	0.54	0.59	0.05
rs2308292	0.42	0.70	0.28	0.47	0.76	0.29	0.44	0.68	0.24	0.47	0.69	0.22
rs28923216	0.50	0.50	0.00	0.47	0.47	0.00	0.50	0.50	0.00	0.53	0.53	0.00
rs3038530	0.45	0.67	0.22	0.35	0.47	0.12	0.50	0.65	0.14	0.51	0.64	0.13
rs3042783	0.40	0.56	0.15	0.43	0.61	0.18	0.50	0.58	0.08	0.57	0.63	0.06
rs3045264	0.45	0.46	0.01	0.47	0.47	0.00	0.45	0.45	0.00	0.56	0.56	0.00
rs3051300	0.50	0.50	0.00	0.50	0.50	0.00	0.47	0.48	0.01	0.43	0.45	0.02
rs33951431	0.49	0.51	0.02	0.51	0.51	0.00	0.47	0.52	0.05	0.47	0.51	0.04
rs34511541	0.44	0.66	0.21	0.37	0.55	0.18	0.50	0.60	0.09	0.57	0.61	0.04

rs34528025	0.46	0.46	0.00	0.45	0.45	0.00	0.44	0.56	0.12	0.39	0.49	0.10
rs34795726	0.50	0.51	0.01	0.51	0.53	0.02	0.48	0.48	0.00	0.46	0.46	0.00
rs34811743	0.45	0.45	0.00	0.35	0.35	0.00	0.37	0.41	0.05	0.40	0.46	0.06
rs35605984	0.50	0.51	0.01	0.47	0.49	0.02	0.49	0.49	0.00	0.51	0.51	0.00
rs35769550	0.50	0.50	0.00	0.61	0.61	0.00	0.47	0.47	0.00	0.41	0.41	0.00
rs36062169	0.49	0.49	0.00	0.49	0.49	0.00	0.50	0.50	0.00	0.59	0.59	0.00
rs3838581	0.44	0.44	0.00	0.43	0.43	0.00	0.51	0.51	0.00	0.47	0.47	0.00
rs3841948	0.48	0.48	0.00	0.53	0.53	0.00	0.42	0.42	0.00	0.38	0.38	0.00
rs4646006	0.49	0.49	0.00	0.56	0.56	0.00	0.50	0.50	0.00	0.39	0.39	0.00
rs5895446	0.47	0.66	0.19	0.53	0.73	0.20	0.44	0.67	0.23	0.24	0.55	0.31
rs60901515	0.44	0.44	0.00	0.57	0.57	0.00	0.49	0.50	0.01	0.61	0.61	0.00

Table 8. Length-based (LB) and sequence-based (SB) observed heterozygosity rank (1= highest) in four major US population groups for 68 INDEL loci.

Locus	AFA		ASA		CAU		HIS	
	LB Rank	SB Rank	LB Rank	SB Rank	LB Rank	SB Rank	LB Rank	SB Rank
rs10623496[1]	5	15	50	55	62	63	44	51
rs10629077[1]	40	48	59	59	65	66	47	54
rs10688868[1,2]	51	2	9	6	42	7	33	1
rs1160886	11	24	10	18	25	34	37	44
rs1160956[1]	21	25	42	47	63	64	23	34
rs13447508[1]	12	20	12	20	64	65	21	33
rs140809[1,2]	52	37	51	21	41	22	67	52
rs1610871[1,2]	13	6	54	48	10	23	50	49
rs16402	14	26	65	65	59	61	58	61
rs16458	22	31	43	49	37	46	24	35
rs16624[1]	58	56	44	50	66	67	5	12
rs17859968[1,2]	29	21	13	19	8	18	16	19
rs2067140[1,2]	46	7	45	25	11	15	26	21
rs2067191[1]	15	23	4	12	50	55	40	46
rs2067208[1,2]	54	58	66	66	53	19	34	25
rs2067294	61	61	46	51	38	47	20	32
rs2307507[1]	62	62	49	54	26	35	57	60
rs2307526[1,2]	16	3	8	17	12	8	32	2
rs2307579[1]	17	27	62	62	43	50	59	62
rs2307580	63	63	23	31	27	36	39	45
rs2307603	30	38	60	60	3	11	10	15
rs2307656[1,2]	37	32	24	14	1	3	22	28
rs2307689[1,2]	6	9	52	3	54	5	38	17
rs2307696	38	46	25	32	57	60	2	7
rs2307700[1,2]	41	16	57	52	18	9	54	24
rs2307710	2	10	30	37	5	14	56	59
rs2307839	60	60	17	26	67	68	61	64
rs2307850	23	33	31	38	44	51	53	58
rs2307978[1,2]	31	5	32	8	68	6	64	23
rs2308112	64	64	1	4	19	30	51	56
rs2308137	47	54	33	39	28	37	62	65
rs2308171	53	57	67	67	56	59	65	67
rs2308189[1,2]	8	1	3	2	48	4	9	8
rs2308196	24	34	47	53	39	48	35	41
rs2308232[1,2]	42	49	53	44	22	29	55	48
rs2308242[1]	59	59	18	27	49	54	60	63
rs2308276[1,2]	9	22	19	1	13	24	13	13
rs2308292[1,2]	25	17	26	9	32	1	27	3

rs28923216[1]	48	50	16	24	33	42	14	22
rs3038530[1,2]	26	8	5	5	60	43	15	4
rs3042783[1,2]	55	35	11	15	45	12	7	5
rs3045264[1]	18	28	40	45	34	44	11	16
rs3047269	7	19	41	46	31	41	4	11
rs3051300[1]	65	65	34	40	23	32	42	43
rs3062629	32	39	27	33	24	33	18	29
rs3080855	33	40	14	22	14	25	28	37
rs33917182	49	55	58	58	52	57	66	68
rs33951431[1,2]	50	11	20	13	20	31	29	26
rs34051577	3	12	35	41	46	52	30	38
rs34495360	44	53	7	16	55	58	63	66
rs34510056	19	29	63	63	2	10	48	55
rs34511541[1,2]	56	51	6	7	58	20	8	9
rs34528025[1]	39	47	15	23	40	49	49	30
rs34535242	34	41	55	56	29	38	12	18
rs34541393	35	42	61	61	15	26	36	42
rs34795726[1]	66	66	56	57	21	27	31	39
rs34811743[1,2]	10	13	68	68	61	62	45	40
rs35605984[1]	4	14	29	36	35	39	17	27
rs35716687	20	30	36	42	9	21	41	47
rs35769550[1]	68	68	21	28	4	13	43	50
rs36040336	27	36	64	64	51	56	19	31
rs36062169[1]	1	4	28	29	30	40	6	14
rs3838581[1]	43	52	22	30	47	53	25	36
rs3841948[1]	57	43	2	10	16	28	52	57
rs4187	36	44	37	43	36	45	1	6
rs4646006[1]	67	67	38	34	7	17	46	53
rs5895446[1,2]	28	18	48	11	17	2	68	20
rs60901515[1,2]	45	45	39	35	6	16	3	10

[1]Marker is part of a microhaplotype observed in these population data
[2]Marker is one of the 22 INDELs/microhaplotypes with increased heterozygosity

AIM INDEL Markers

Portions of this section will be submitted to *The International Journal of Legal Medicine and Electrophoresis*

All INDEL markers identified in VCFtools were filtered based on pairwise F_{ST} comparisons of the three major global population groups, Caucasian, African, and East Asian. Those with at least one pairwise F_{ST} value greater than 0.5 were included. These markers were subsequently filtered to include only INDELs of length three to six base pairs. Next, the markers with allele frequency divergence in one of the three population groups were selected. A summary of the number of INDELs that meet these criteria can be seen in Table 10.

From the remaining INDELs, 60 markers, 20 for each population group, were selected as potential AIMs. These were chosen based on physical distance and allele frequency divergence. The allele frequency difference, or delta (δ) value was calculated between each population group. Markers with high delta value (> 0.5) were included in the panel as potential AIMs. Additionally, all syntenic markers, markers on the same chromosome, were selected to have a physical distance of at least 1Mb from its nearest neighbor.

Table 10. Summary of AIM-INDELs identified using VCFtools.

Chromosome	Caucasian	East Asian	African	Total
1	7	35	58	**100**
2	11	40	107	**158**
3	7	14	72	**93**
4	5	32	117	**154**
5	7	46	56	**109**
6	8	22	38	**68**
7	7	23	38	**68**
8	6	5	21	**32**
9	7	11	43	**61**
10	9	19	45	**73**
11	3	12	44	**59**
12	0	8	15	**23**
13	0	11	7	**18**
14	1	2	4	**7**
15	13	15	24	**52**
16	1	6	8	**15**
17	2	5	18	**25**
18	1	6	10	**17**
19	0	6	10	**16**
20	1	5	8	**14**
21	0	3	3	**6**
22	1	3	10	**14**
Total	**97**	**329**	**756**	**1182**

Statistical Analysis

To test whether these marker would cluster the population groups correctly, Principal Component Analysis (PCA) was performed using the software program Past3 (Figure 18A). Samples from each population group were labelled with a different color to show the distinct clusters among the training set. The first two principal components (PC1 and PC2) explained 40.3% of the variation. To further test the markers capacity to separate the major global population groups, populations from the 1000 Genomes Project that were not in the original training set were added to the PCA (Figure 18B-D). A population from Gambia in the Western Division, an Iberian population from Spain, and a

Kinh population in Ho Chi Minh City, Vietnam were selected to represent the African, Caucasian, and East Asian population groups, respectively.

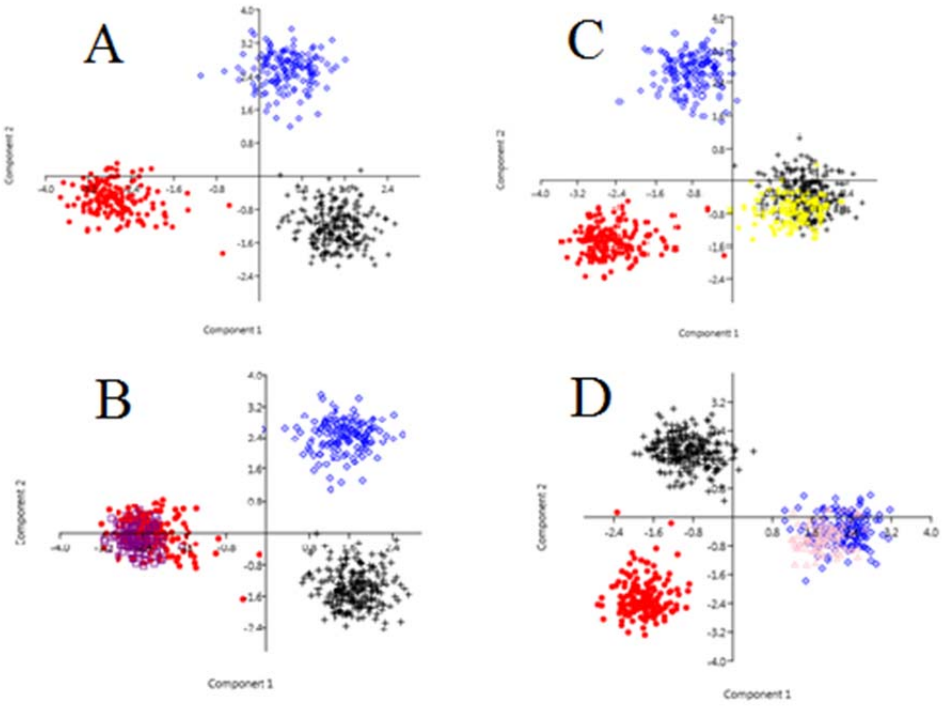


Figure 18. Principal Component Analysis (PCA) of 60 Ancestry Informative Markers (AIMs) using the software package, Past3. A) Original training set of 550 Individuals; Caucasian (Black Plus), East Asian (Blue Diamond) and African (Red Dot). B) Original training set with additional samples from Gambia in the Western Division (Purple Square). C) Original training set with additional Iberian samples from Spain (Yellow Star). D) Original training set with additional samples from Kinh in Ho Chi Minh City, Vietnam (Pink Triangle).

Additional statistical analysis was performed on the 60 markers using the software package Genetic Data Analysis (GDA). Exact tests for Hardy-Weinberg Equilibrium (HWE) were performed on the set of 60 AIMs. Of the 60 markers, five in the African population group, seven in the Caucasian population group, and three in the East Asian population group showed departure from HWE at a significance level of 0.05. After Bonferroni correction for multiple comparisons (α=0.05/60), only one marker showed

significant departure from HWE in all three population groups. INDEL rs78981054 showed a p-value of less than $8.33x10^{-4}$ in all three population groups.

The remaining 59 markers were evaluated for Linkage Disequilibrium (LD) to determine if there was an observable pattern of inheritance between any of the marker combinations. Of the 1710 combinations per population group, 222 in the African population group, 145 in the Caucasian population group, and 130 combinations in the East Asian population group showed LD. After Bonferroni correction for multiple comparisons (α= 0.05/1710), only three in the African population group, two in the Caucasian population group, and one combination in the East Asian population group showed significant LD. The marker combinations that showed significant LD in the African population group included, rs59009450/ rs67344973, rs67344973/rs10651200, and rs113043680/rs10528149. In the Caucasian population group, they were rs59009450/rs113501732 and rs59009450/rs35779249. Finally, in the East Asian population group, the markers that showed significant LD were rs141933116/ rs74515961. The 59 AIMs are described in Table 10.

The panel of 59 AIMs was analyzed for ancestry admixture in the software program STRUCTURE v.2.3.4 (31-34). After 20 simulations for 10 values of K, the *ad hoc* statistic, ΔK, was calculated (35). Figure 19A describes the distribution of ΔK for K values 1 through 10. ΔK is maximized when K=3. Figure xdescribes the STRUCTURE output for each individual (19C) and the population groups as a whole (19B).

Table 10. Descriptive Statistics of the 59 Ancestry-Informative Markers. (He and Ho refer to expected and observed heterozygosity, respectively).

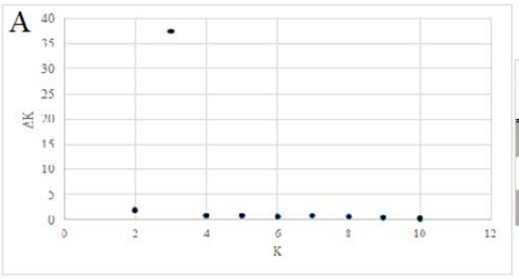
CAUCASIAN												
				Frequency of Insertion			Delta		Pairwise Fst		Heterozygosity	
rs#	Chrom.	Position	Sequence	African	Caucasian	East Asian	v.AFR	v.EAS	v.AFR	v.EAS	He	Ho
rs139570718	1	214397853	-/CCCAG	0.0352564	0.727459	0.223333	0.6922026	0.504126	0.640101	0.400736	0.477808	0.296364
rs3831920	1	1227664	-/TGAG	0.375	0.913934	0.293333	0.538934	0.620601	0.508888	0.600295	0.483578	0.289091
rs70958016	2	13725708	-/AGTTT	0.865385	0.278689	0.65	0.586696	0.371311	0.504749	0.244201	0.496152	0.372727
rs67934853	2	74943887	-/TAAC	0.923077	0.258197	0.81	0.66488	0.551803	0.603647	0.460751	0.481514	0.3
rs139220746	2	200205694	-/TATC	0.826923	0.227459	0.673333	0.599464	0.445874	0.52312	0.338786	0.499725	0.365455
rs140498743	3	139232513	-/TGTC	0.842949	0.360656	0.95	0.482293	0.589344	0.37517	0.517759	0.450366	0.287273
rs5864438	4	178146869	-/CTAT	0.839744	0.192623	0.803333	0.647121	0.61071	0.585954	0.542572	0.4968	0.303636
rs149676649	5	28495386	-/GATT	0.349359	0.79918	0.106667	0.449821	0.692513	0.350045	0.637831	0.499858	0.343636
rs57237250	6	110263002	-/GAGT	0.826923	0.260246	0.903333	0.566677	0.643087	0.479728	0.574514	0.481866	0.312727
rs1160871	7	28168745	-/TCTT	0.217949	0.788934	0.0233333	0.570985	0.7656007	0.491182	0.72318	0.487054	0.272727
rs55855642	8	122272251	-/ATAGAG	0.855769	0.381148	0.996667	0.474621	0.615519	0.368124	0.561435	0.432949	0.287273
rs67538813	9	30471814	-/CAGA	0.958333	0.383197	0.696667	0.575136	0.31347	0.507485	0.17589	0.465671	0.354545
rs10651200	10	69800907	-/TAACAA	0.939103	0.334016	0.83	0.605087	0.495984	0.525682	0.389713	0.460708	0.336364
rs71991275	10	28470438	-/AATA	0.74359	0.348361	0.996667	0.395229	0.648306	0.266669	0.596017	0.462733	0.329091
rs11576045	12	111799524	-/TGT	0.762821	0.235656	0.936667	0.527165	0.701011	0.433617	0.646023	0.488782	0.298182
rs35779249	13	43964476	-/TAA	0.961538	0.297131	0.82	0.664407	0.522869	0.607878	0.422731	0.467564	0.289091
rs370096890	14	65368820	-/CTTGA	0.910256	0.209016	0.63	0.70124	0.420984	0.648534	0.314874	0.499421	0.307273
rs138439822	15	35537968	-/TAACTC	0.858974	0.270492	0.713333	0.588482	0.442841	0.506957	0.327046	0.493679	0.345455
rs10528149	16	69989686	-/TGAT	0.0769231	0.721311	0.36	0.6443879	0.361311	0.578944	0.233118	0.493248	0.323636
rs55885844	17	79605107	-/ATTAA	0.304487	0.657787	0.00333333	0.3533	0.6544537	0.219042	0.602563	0.47119	0.321818

EAST ASIAN

				Frequency of Insertion			Delta		Pairwise Fst		Heterozygosity	
rs#	Chrom.	Position	Sequence	African	Caucasian	East Asian	v.AFR	v.CAU	v.AFR	v.CAU	He	Ho
rs141933116	1	8189066	-/AAGT	0.701923	0.956967	0.39	0.311923	0.566967	0.176461	0.579729	0.394559	0.310909
rs5839799	2	241417278	-/GTCT	0.88141	0.694672	0.286667	0.594743	0.408005	0.533799	0.282733	0.463231	0.352727
rs72375069	3	27427821	-/AATT	0.980769	0.657787	0.256667	0.724102	0.40112	0.7167	0.273236	0.461219	0.330909
rs33915414	4	21762063	-/CATGTT	0.0801282	0.385246	0.803333	0.7232048	0.418087	0.693429	0.295226	0.485208	0.334545
rs1610951	5	108999835	-/TTGG	0.971154	0.868852	0.336667	0.634487	0.532185	0.61792	0.475083	0.372597	0.232727
rs367799178	6	21621169	-/TTAA	0.285256	0.284836	0.89	0.604744	0.605164	0.544839	0.527296	0.49545	0.38
rs151280400	7	125249166	-/AATC	0.910256	0.659836	0.35	0.560256	0.309836	0.504593	0.17317	0.457571	0.358182
rs10581451	8	73854660	-/TGAG	0.894231	0.965164	0.18	0.714231	0.785164	0.677952	0.799592	0.39372	0.163636
rs150560593	9	95478810	-/TGCA	0.865385	0.739754	0.283333	0.582052	0.456421	0.514276	0.344585	0.454866	0.334545
rs67205569	10	94941566	-/TTGAC	0.971154	0.885246	0.1333333	0.8378207	0.7519127	0.831329	0.724881	0.416701	0.165455
rs143873637	11	97893598	-/TTGA	0.823718	0.866803	0.243333	0.580385	0.62347	0.504557	0.57738	0.432279	0.274545
rs66693708	12	77398405	-/TAAG	0.974359	0.805328	0.326667	0.647692	0.478661	0.633852	0.386204	0.40115	0.270909
rs10587399	13	37776954	-/TACT	0.887821	0.717213	0.243333	0.644488	0.47388	0.594295	0.362933	0.463231	0.341818
rs141122561	14	49242955	-/TTAGT	0.996795	0.963115	0.37	0.626795	0.593115	0.627839	0.612742	0.30695	0.174545
rs200047010	15	102264144	-/GCAGG	0.714744	0.702869	0.13	0.584744	0.572869	0.515882	0.486211	0.49545	0.336364
rs10549914	17	5328978	-/TTTA	0.852564	0.719262	0.18	0.672564	0.539262	0.622449	0.443694	0.476233	0.332727
rs74515961	18	52716306	-/ATGTC	0.983974	0.786885	0.376667	0.607307	0.410218	0.598112	0.301386	0.39372	0.269091
rs10668859	19	266759	-/GAAAG	0.86859	0.637295	0.14	0.72859	0.497295	0.692713	0.393972	0.491395	0.341818
rs11474791	20	19234875	-/GGACT	0.221154	0.108607	0.79	0.568846	0.681393	0.487299	0.652616	0.440101	0.278182
rs3074939	21	43422429	-/CAGT	0.205128	0.364754	0.836667	0.631539	0.471913	0.569109	0.361085	0.49508	0.358182

AFRICAN

				Frequency of Insertion			Delta		Pairwise Fst		Heterozygosity	
rs#	Chrom.	Position	Sequence	African	Caucasian	East Asian	v.EAS	v.CAU	v.EAS	v.CAU	He	Ho
rs59385244	1	16367160	-/AAGG	0.314103	0.821721	0.99	0.675897	0.507618	0.66481	0.424857	0.400337	0.261818
rs59009450	1	248818535	-/AAGAT	0.689103	0.0881148	0.246667	0.442436	0.6009882	0.325636	0.575336	0.421831	0.24
rs11277277	2	11273217	-/CACAG	0.339744	0.987705	0.936667	0.596923	0.647961	0.552569	0.687956	0.332102	0.176364
rs67344973	2	178513061	-/GTTT	0.875	0.256148	0.263333	0.611667	0.618852	0.551826	0.545406	0.491639	0.325455
rs148921522	3	85588405	-/TAAC	0.160256	0.625	0.86	0.699744	0.464744	0.656259	0.354217	0.493889	0.354545
rs112191273	3	7351968	-/GCTT	0.657051	0.0266393	0.0433333	0.6137177	0.6304117	0.580653	0.656176	0.332102	0.165455
rs70941213	4	106669965	-/AGTT	0.916667	0.243852	0.12	0.796667	0.672815	0.776951	0.613038	0.480799	0.256364
rs72255563	5	176226827	-/ACTT	0.772436	0.114754	0.136667	0.635769	0.657682	0.576689	0.622541	0.4261	0.229091
rs60234845	6	155859718	-/CCAA	0.75	0.239754	0.156667	0.593333	0.510246	0.521703	0.412464	0.462232	0.349091
rs35625334	7	79883089	-/AGAT	0.894231	0.354508	0.106667	0.787564	0.539723	0.764883	0.450782	0.493248	0.312727
rs56767439	8	12977501	-/TTAC	0.810897	0.204918	0.156667	0.65423	0.605979	0.59818	0.534393	0.463231	0.287273
rs113043680	9	126640635	-/TAAG	0.708333	0.139344	0.0966667	0.6116663	0.568989	0.556619	0.511686	0.411409	0.265455
rs113501732	10	128948642	-/CCTGT	0.272436	0.911885	0.763333	0.490897	0.639449	0.386335	0.616993	0.428189	0.249091
rs74499778	11	129941381	-/AGCT	0.375	0.952869	0.62	0.245	0.577869	0.110407	0.583277	0.421831	0.309091
rs2307553	14	80121686	-/TGAC	0.884615	0.252049	0.38	0.504615	0.632566	0.430009	0.562808	0.49819	0.383636
rs138123572	15	72786235	-/TGAC	0.185897	0.959016	0.946667	0.76077	0.773119	0.738709	0.782133	0.388618	0.149091
rs66913380	17	42191379	-/GCCA	0.195513	0.786885	0.85	0.654487	0.591372	0.598891	0.515387	0.463231	0.312727
rs10540310	20	59105205	-/CTTC	0.272436	0.75	0.87	0.597564	0.477564	0.531245	0.371308	0.457037	0.327273
rs10560659	21	17025686	-/CAAT	0.778846	0.17418	0.12	0.658846	0.604666	0.607315	0.54009	0.443219	0.272727



B

Population of Origin	Inferred Population Assignment (%)			# of Individuals
	AFR	EAS	CAU	
African	100	0	0	156
East Asian	0	100	0	150
Caucasian	0	0	100	244

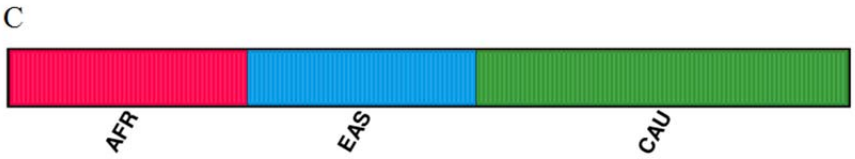
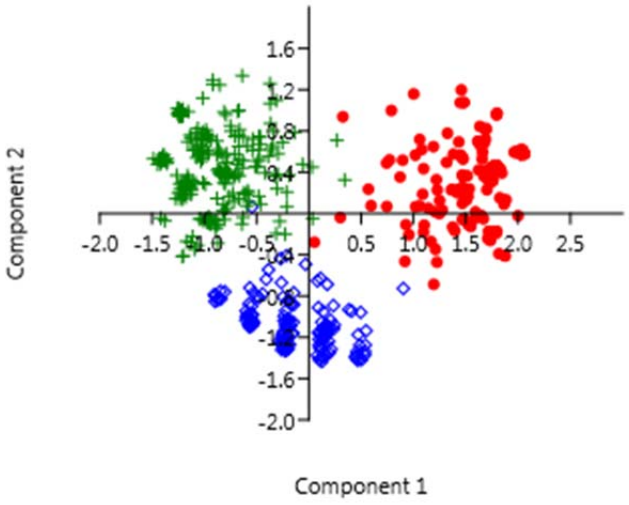


Figure 19. STRUCTURE v.2.3.4 Analysis of 59 AIMs. A) Graphical representation of the *ad hoc* statistic, ΔK. B) Table describing the overall population assignment of the training set samples for the 20 simulations at K=3. C) STRUCTURE plot for African (AFR), East Asian (EAS) and Caucasian (CAU) population groups compiled in CLUMPP v1.1.2 (36) and graphically displayed in *distruct v1.1* (37).

Upon further analysis it was determined that 59 INDELs were more than adequate to separate major global popualtions. To determine the minimum number of markers required to separate 3 major global populations, We used an iterative process where we performed PCA and structure on a decreasing number of our best markers (sorted by FST and then Delta value). The results were that we could separate 3 major global populations with 12 INDELs if care was taken to choose the appropriate markers (figure 20)



Population of Origin	Inferred Population Assignment (%)			# of Individuals
	AFR	EAS	CAU	
African	99.43	0.37	0.2	156
East Asian	0	99.1	0.9	150
Caucasian	0	0.2	99.8	244

Figure 20. (Counter-clockwise from upper left) PCA, and structure barplot depicting the separation of samples based on 12 best AIM INDEL markers. In the upper right, table displaying confidence when 3 populations are assumed during structure analysis.

Based on this analysis it was decided to reduce the number of markers to a number that would be more manageable in a multiplex PCR reaction. The list selected are described in table 11 and represent a redundant set of AIM markers to be multiplexed in the event of allele drop out. A set of 30 best markers were chosen to be multiplexed and in a similar manner to the way we checked primer compatibility for the HID primers, unlabeled primers were ordered and assessed for suitability on a microfluidic electrophoresis platform.

Primer pairs were chosen via Primer-BLAST produced desired primer sequences, lengths, and the associated melting temperatures, and G-C content percentage Once primer pairs were chosen, they were checked for potential dimerization using MPprimer. Each forward primer is compared to each reverse primer, and the output gives matches, an alignment score, 3'-3' dimer check, and ΔG (kcal/mol). Any primers with alignment scores of 5 or greater, or ΔG of -7 or less were discarded, and different primer pairs for those markers were chosen. The final 30 primer pairs are shown in Table11.

Table 11: Top 30 markers and selected forward and reverse primers.

Order #	rs#	Chrom	Position	Forward Primer	Reverse Primer	Seq Length	INDEL
2	rs139570718	1	214397853	CACTTCTAGGGATTTGTGGGGT	AGTTGAGACTTGGCTGACGG	147	CCCAG (Ins)
4	rs370096890	14	65368820	ACCAAATGCTTGGAAGTCTTGA	AACTGGGGCCAGGTGTTAAT	59	CTTGA (Del)
5							
6	rs67205569	10	94941566	CCAGGGTCTAAACAGAGGCA	TGACCCAGAATCCTGTGACTT	64	TTGAC (Del)
8	rs10668859	19	266759	CAGGAGTAGCCCATCATGAACA	CCCTAAGCTGGACTGTCTCC	128	GAAAG (Ins)
10	rs149676649	5	28495386	TTGTTTGTCCCTGTATTTAACAGAA	ATTGCATTGTGCATTTTTGTCATGT	171	GATT (Ins)
12	rs11474791	20	19234875	TCCCACAGAGTGACATTGCC	GAACCCCTGGACCATGTGAG	92	GGACT (Ins)
14	rs72375069	3	27427821	TAAATCCCTTGCACTACGCA	AGGTACTCTAATGTATTGCTGAAGA	140	AATT (Del)
15							
16	rs66913380	17	42191379	CAGCATGGCCTGGGAGC	GAGAGGGTTCAGCCAACACC	61	GCCA (Del)
18	rs72255563	5	176226827	ACACGCACACTCAGCACAC	GGAGACACACGTCTCCATGC	65	ACTT (Del)
20	rs3831920	1	1227664	TGAGCCGGGTAGCACTCA	GGGCATCAGGACCCAGATTT	94	TGAG (Del)
22	rs59385244	1	16367160	AAATCACCACCCTGCCTGAG	AAGTGCAGCAGGAAAAGCTC	73	AAGG (Ins)
24	rs5864438	4	178146869	CTGAACCTGGACGTGGTCAT	CCAGAGTGGATGCACCATAGAC	59	CTAT (Del)
25	rs57237250	6	110263002	TGCTGTTCTCATTCCACGTAT	AGTTAGCCATGGGAAGCACA	69	GAGT (Ins)
26	rs1610951	5	108999835	ATGTCAAGCACCGTGCCA	CTGTGTGACCTCTCTGAGC	83	TTGG (Del)
27	rs367799178	6	21621169	TTGCATTATGGCCAAAAATCATGT	CAGTTCCAACACAAAGGTAGCA	136	TTAA (Del)
28	rs10549914	17	5328978	AGCAATCAGTTCTCTTTGTCAAC	ACAGATACAGAATGTCAGGGTC	60	TTTA (Del)
29	rs112191273	3	7351968	TGGTGATGATTTTCAAATGGGACT	ACATTGCAGATTTAACTCATGAACC	61	GCTT (Ins)
30	rs56767439	8	12977501	ATGCCATAGTGAGAGAAGGAACA	ACCTGTCTTGCAGGAAGAACC	59	TTAC (Del)

The TapeStation, which was used to analyze the amplicon results, is an automated electrophoresis system that uses a ScreenTape matrix similar to agarose gel. The samples absorb an intercalating dye, are separated by size, and then fluorescence is

captured by a camera (Figure 20). Using a ladder, a reference of bp length, the approximate size of the amplified product can be determined. An electropherogram is then produced by the program to give a graphic representation of the sequence lengths (Figure 21). The observed amplicon size was close to that predicted for all primer pairs. Primer pair 18 produced no product after multiple attempts, and was therefore removed from further testing. The TapeStation results from the initial multiplex trial are show in Figure 22A-H.

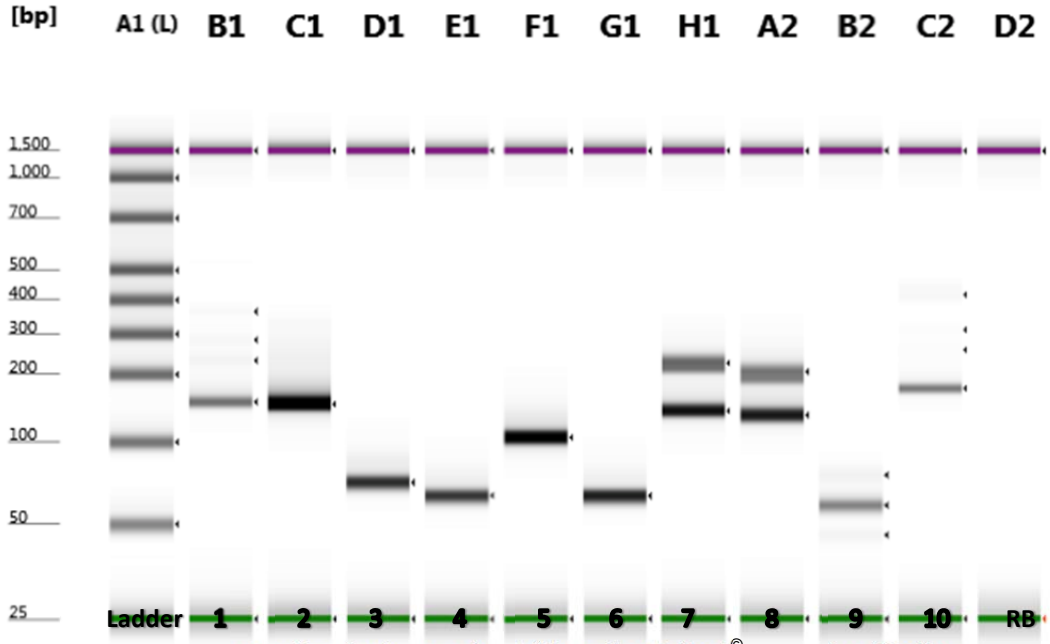


Figure 20: Electrophoresis results of primer pairs 1-10 on the Agilent© 2200 TapeStation.

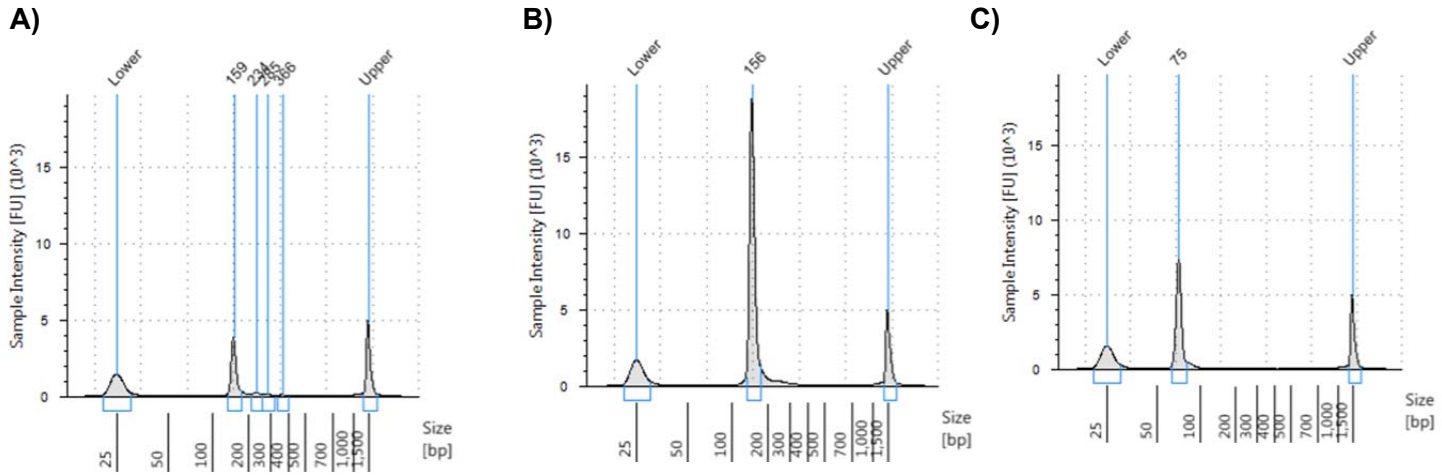


Figure 21: Electropherograms produced by electrophoresis of primer pairs 1 **(A)**, 2 **(B)**, and 3 **(C)** on the Agilent© 2200 TapeStation.

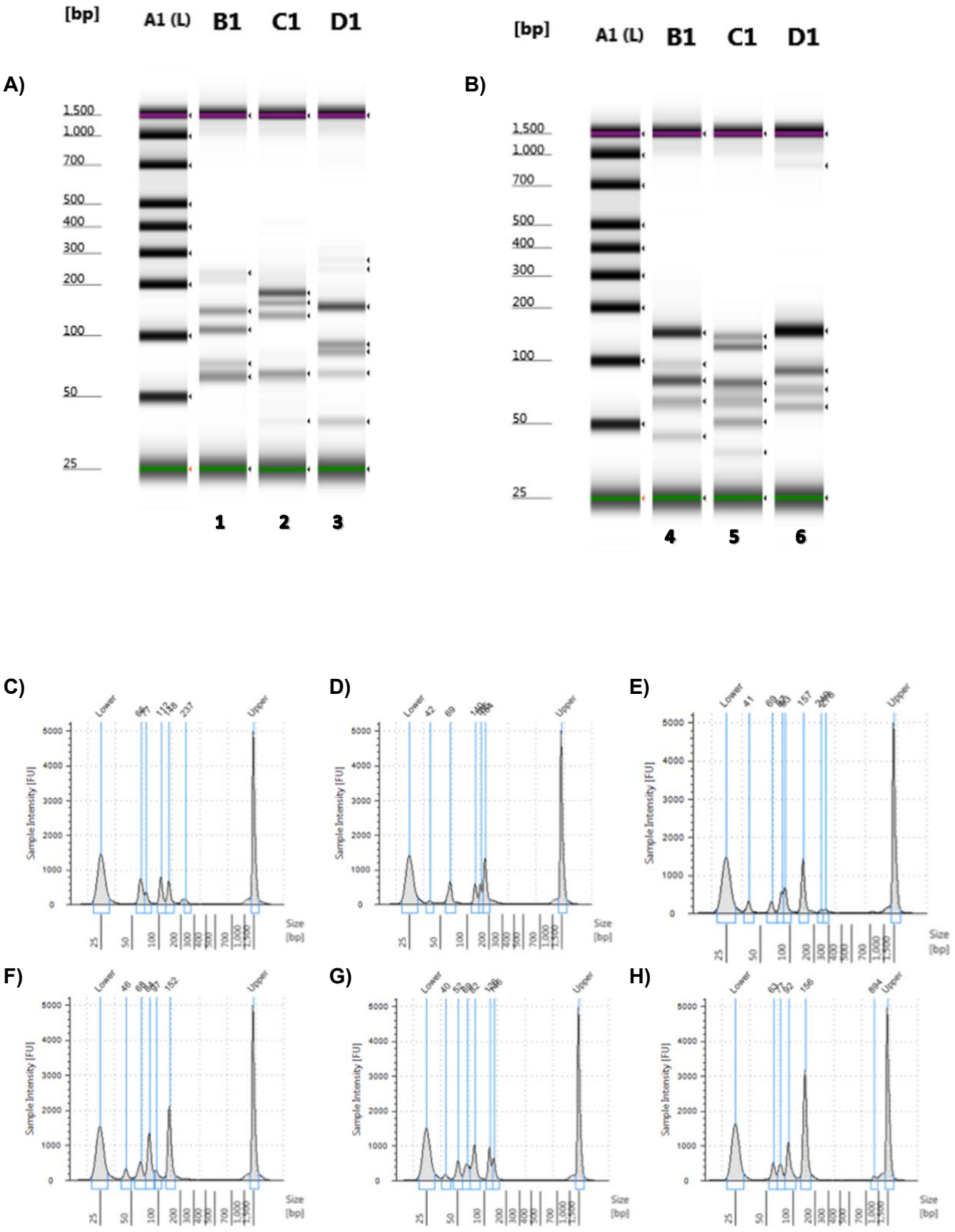


Figure 22: Electrophoresis results of multiplex trials 1-3 **(A)** and 4-6 **(B)** and electropherogram results **(C-H)** on the Agilent© 2200 TapeStation.

Table 12: Primer pairs arranged into dye channels and the expected sequence lengths of each (bps).

Dye Channel	Primer Pair	Alleles (Expected bp sequence lengths)
Blue (6-FAM)	4	59, 64
	28	60, 64
	25	69, 73
	13	83, 86
	23	114, 118
	15	134, 139
	1	151, 155
	10	171, 175
Green (VIC)	16	61, 65
	29	61, 65
	19	78, 82
	12	92, 97
	27	132, 136
	2	142, 147
Yellow (NED)	9	60, 64
	30	60, 64
	22	74, 78
	5	107, 112
	17	137, 143
	11	151, 155
Red (TAZ)	6	59, 64
	3	72, 76
	26	83, 87
	8	128, 133
	14	140, 144
Purple (SID)	24	59, 63
	21	65, 70
	20	94, 98
	7	136, 140

Fluorescently labeled primers were ordered according to table 12 and used to amplify DNA in singlet as well as multiplex by dye channel, and both resulted in oversaturation. When a large amount of amplified DNA is present, it may overwhelm the instrument's ability to measure the results; this is known as oversaturation. A 1:100 dilution of the DNA sample was made, and then the primer sets were used to amplify the DNA in singlet (Figure 23) as well as multiplexes of the same fluorophore (Figure 24A-E). Each single amplicon peak matched its respective location within the multiplex by a difference of no more than 1 base pair.

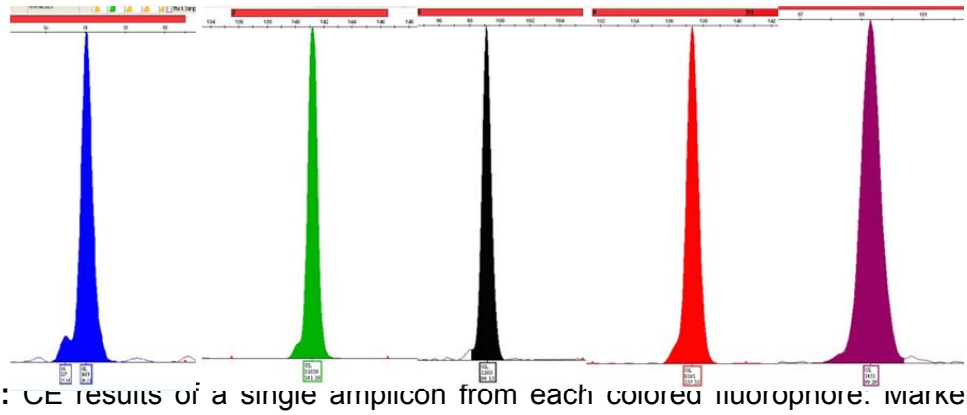
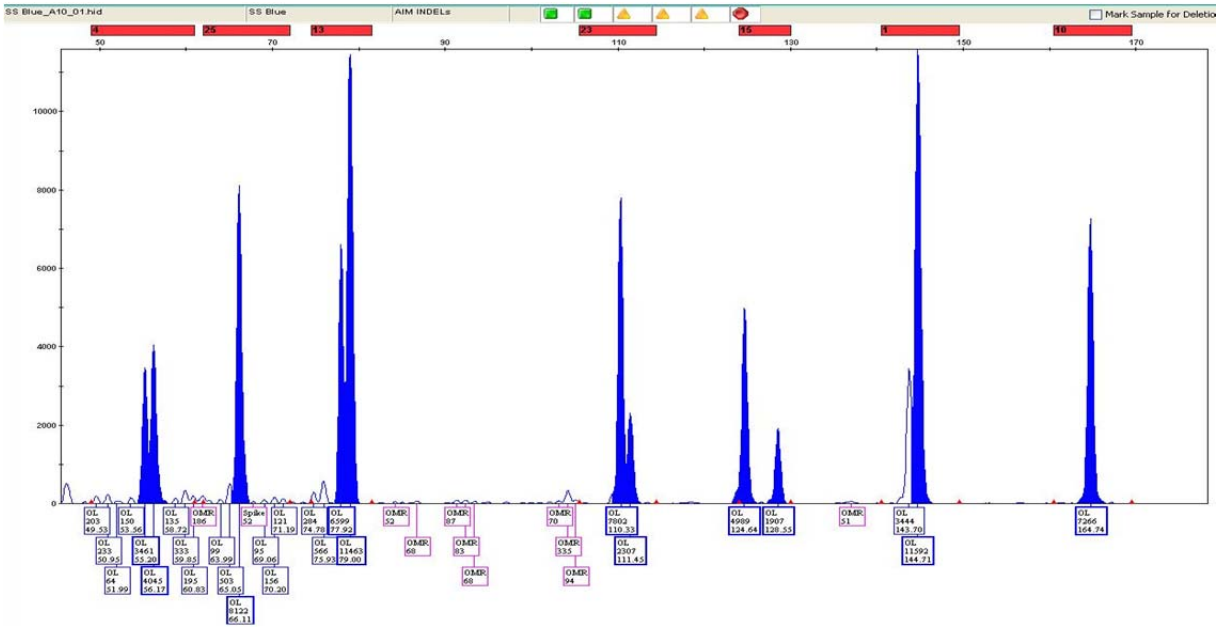


Figure 23: CE results of a single amplicon from each colored fluorophore. Markers 4 (blue), 2 (green), 5 (yellow), 8 (red), and 20 (purple).

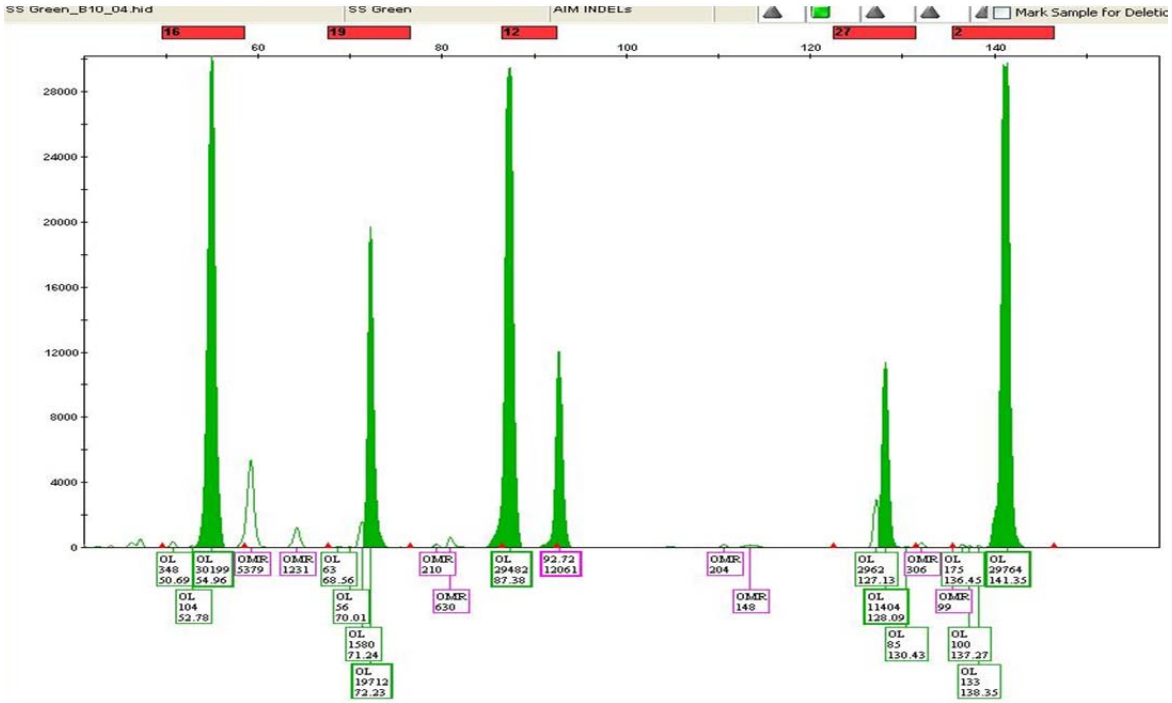
(Following pages)

Figure 24: Multiplex CE results of markers 4, 25, 13, 23, 15, 1, and 10 in the blue channel **(A)**, markers 16, 19, 12, 27, and 2 in the green channel **(B)**, markers 9, 22, 5, 17, and 11 in the yellow channel **(C)**, markers 6, 3, 26, 8, and 14 in the red channel **(D)**, and markers 24, 21, 20, and 7 in the purple channel **(E)** with diluted DNA sample.

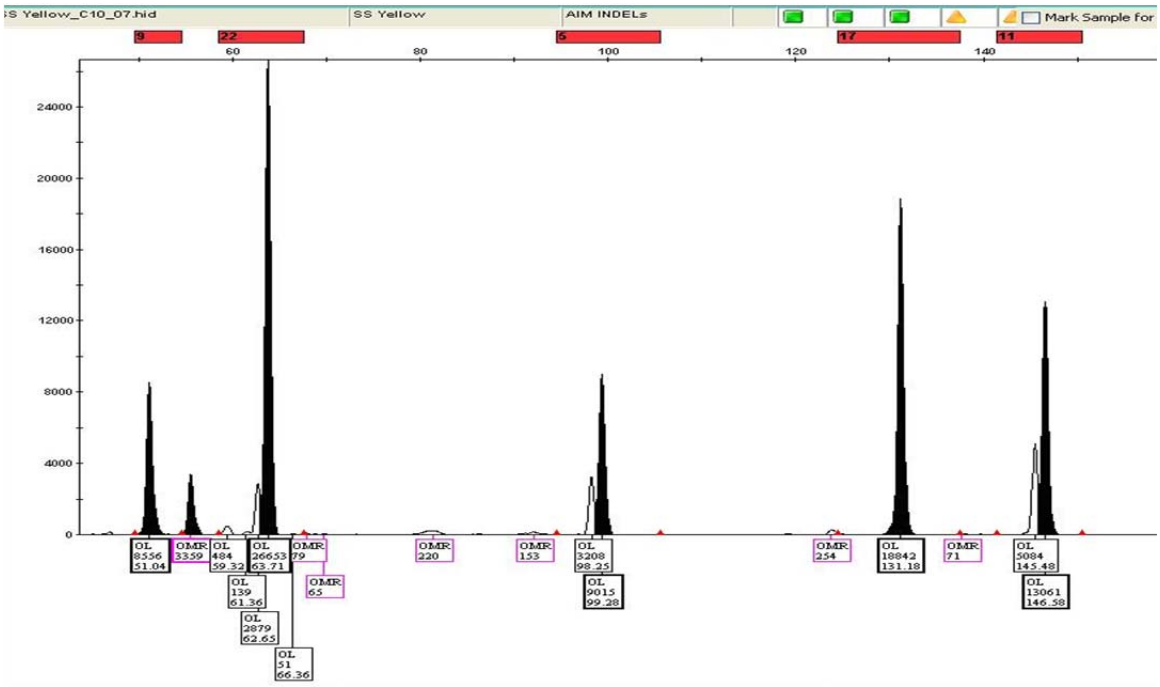
A)



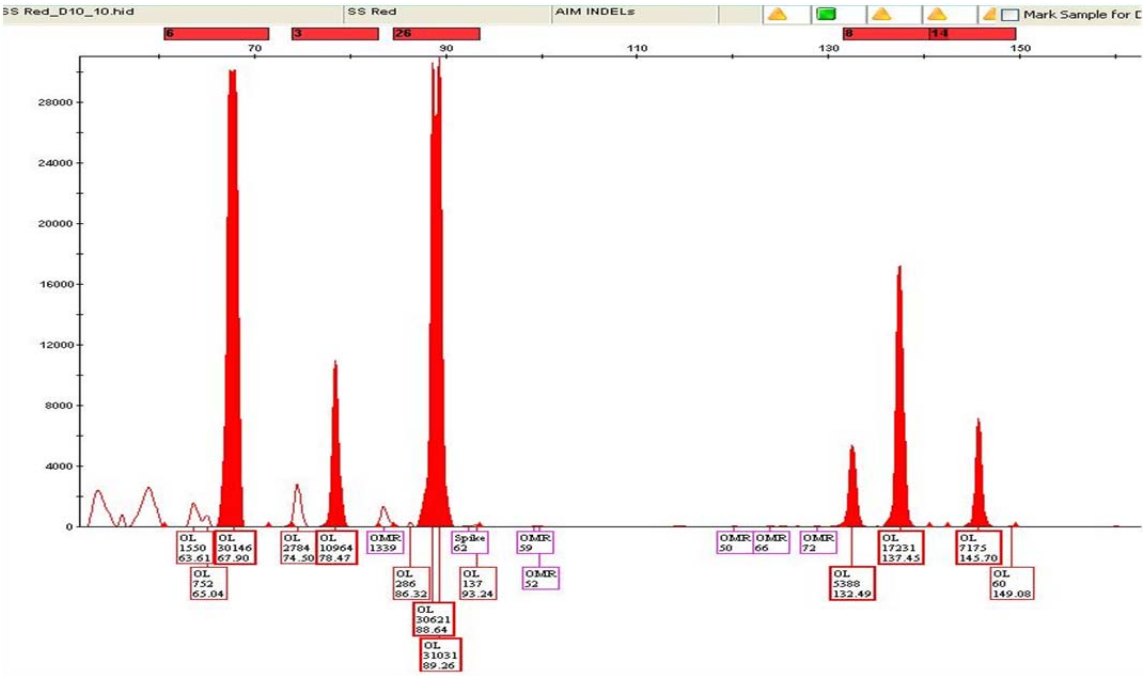
B)



C)



D)



E)

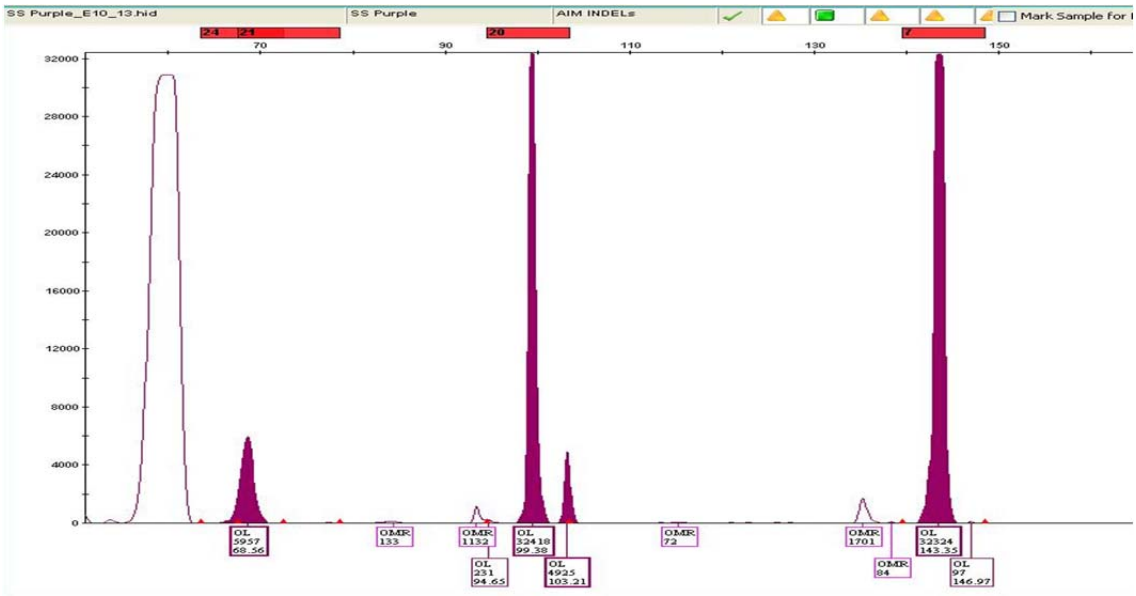


Figure 24: Multiplex CE results of markers 4, 25, 13, 23, 15, 1, and 10 in the blue channel **(A)**, markers 16, 19, 12, 27, and 2 in the green channel **(B)**, markers 9, 22, 5, 17, and 11 in the yellow channel **(C)**, markers 6, 3, 26, 8, and 14 in the red channel **(D)**, and markers 24, 21, 20, and 7 in the purple channel **(E)** with diluted DNA sample.

Once again, The SID channel has proved the most diifcult to balance, but usable data can be generated with the markers in this channel. Aim 3 of the grant calls for this multiplex to be developmentally validated according to SWGDAM guidelines as well, but we are dependent on the release of the remaining project funds to finish this portion of the project. Once the remaining funds are received, we will finish this portion of the proposal and update our revised final report at the end of our extension period. Once again, an additional adjustment of the SID labeled primers in the next lot of multiplex primer mix should correct the problems we have seen in our initial multiplex.

Additional Populations

When the Southwest Hispanic and Southwest Asian samples were added to the PCA, an additional cluster appeared in the plot (Figure 25 A-B). Both population groups tended to cluster between the African and Caucasian population groups with complete separation from the East Asian population group. Both SWH and SWA samples cluster more closely with the Caucasian population group than the other two. The first three principal components for the SWH and SWA PCA explained 38.2% and 33.8% variance, respectively. To determine if the SWH and SWA population groups could be separated from each other using these 59 AIMs, PCA was performed on these two population groups (Figure 25C). The PCA showed significant overlap between to two population groups with PC1 and PC2 explaining only 8.5% of the variance.

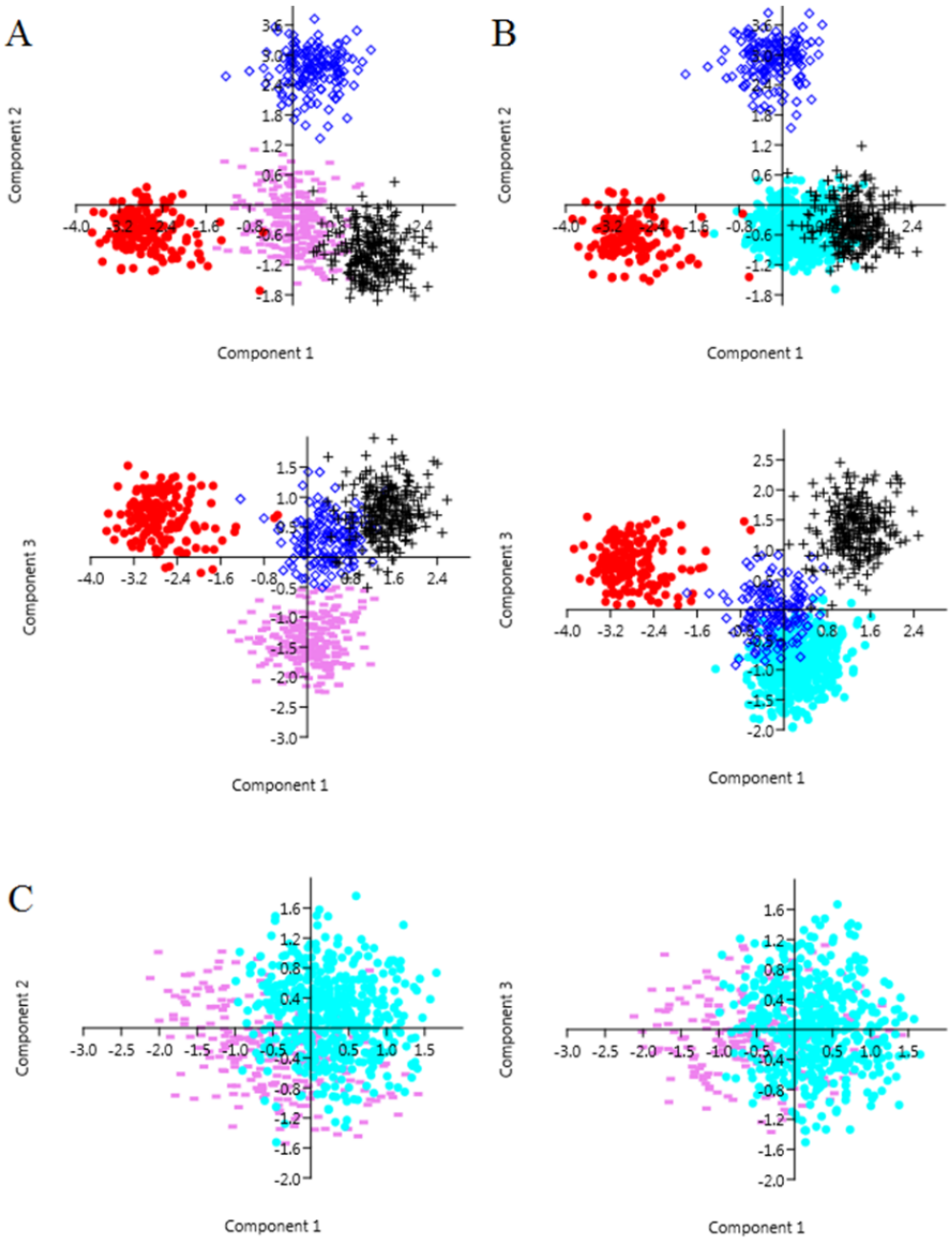


Figure 25. Principal Component Analysis (PCA) with Additional Population Groups Using Past3. A) Original training set of 550 Individuals; Caucasian (Black Plus), East Asian (Blue Diamond) and African (Red Dot) with Southwest Hispanic individuals (Pink Bar). B) Original training set individuals with Southwest Asian individuals (Turquoise Dot). C) PCA of SWH and SWA individuals.

To separate these admixed populations, another approach would be to develop a secondary panel of AIM markers to associate samples to various admixed populations when they have failed to be associated to major population groups. This was beyond the scope of our proposal, but in the interest of increasing the body of knowledge, we designed a set in silico which could discern between Southwest Asian and Southwest Hispanic populations samples utilizing the methods described for the major population groups (figure 26). We then designed 17 small amplicon primer sets that could be ordered and tested by individuals if they desired and they are listed in table 13 This will not be pursued further during this project by our group.

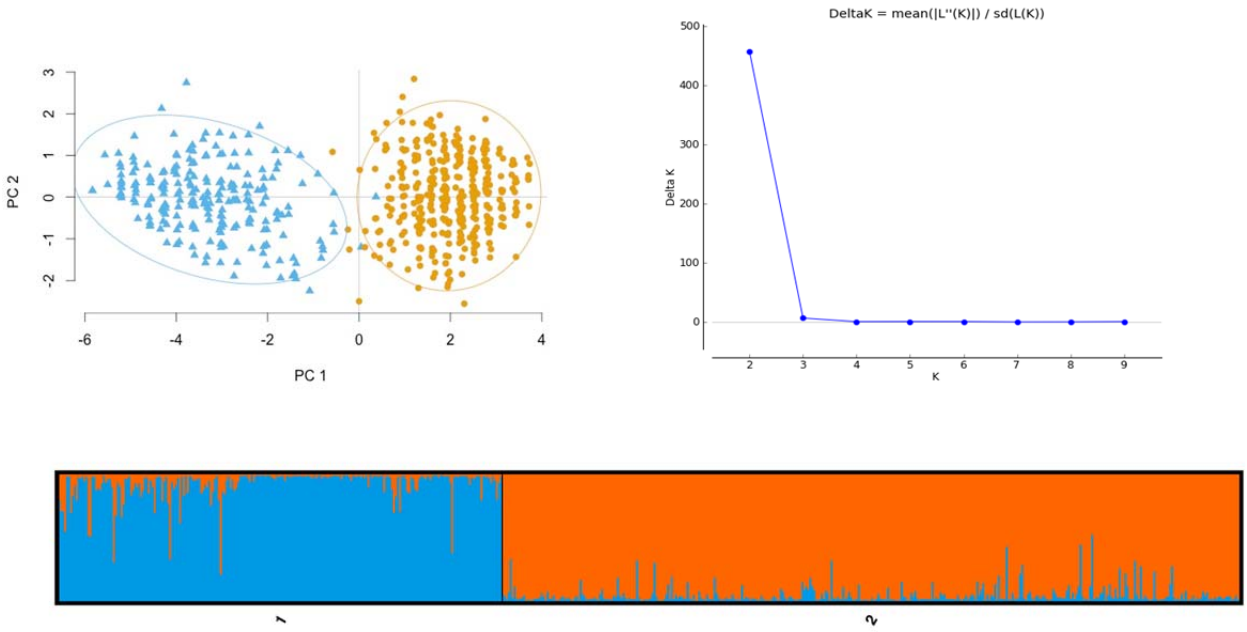


Figure 26 Clockwise from top left principal component analysis of separation of Southwest Hispanic and Southwest Asian populations at %95 confidence interval, plot of Delta K from Structure, and the resultant Structure bar plot with k=2 when observing allele distributions of 17 INDELs

Table 13 Potential supplemental panel of small amplicon INDELs for separating/assigning admixed samples to either Southwest Hispanic or Southwest Asian populations with %95 Confidence Interval

Gene	Chr	Position	Forward (+ strand)	Reverse (- Strand)	Amplicon	Tm (F;R)	%GC (F;R)	Column6	Dye Chane
rs5710711	16	30884552	CCAAAACAATGCTGTTCAGTTC	GGGCTTTATGTGGCCTTTTT	122	59.65	40.91	TTGA(Del)	
rs5775512	1	119716625	GCAGGGGTTGTTTCTGCT	ATGCCTGAGTGTTTCCCATC	145	58.8	55.56	AAACA(Ins)	
rs3836202	2	175665469	AGAGAACAGCAGCGACAGTAAT	AGGATTGCTTCCTGTTGGGATG	167	59.77	45.45	ATGGCT(Ins)	
rs1437357	5	146664264	CACCAACAGCAACCAAGTAGAG	TGATTCCTTCCAGCAGTGTTT	188	59.84	50	AAC(Del)	
rs1130171	11	68751537	CCTCAGCCTCCCAGATTACA	GGGCAACCAAGATTAACAGG	130	60.21	55	CCAG(Ins)	
rs1062460	20	33158217	GCCAGCCCCTAGTAAGCTCT	gaattgctcgaaccctgaag	150	60	60	TTTATT(Ins)	
rs5850158	3	74175219	CTCCAGCCTGGGTTACAGAG	CATGAAAGGTCTGGTCTGAGC	169	59.86	60	AAAA(Del)	
rs1064883	15	84001991	AGCCTGCAGCTAAGAAAGGA	AGATTGAGCCATTGCACTCC	189	59.36	50	CTTT(Ins)	
rs5668554	8	61829343	GGTTTTAAATGCTGTCGCCCA	TTTCCCTGCCTTGGTCTTGC	133	59.73	47.62	GTGT(Ins)	
rs1491263	5	107352467	CCCAAGGACACACTTTCCTG	GCTTGAACCTGAGAGGCAGA	152	60.54	55	TTTC(Del)	
rs7111244	14	46964326	GGCAAATTTATGGCATCTATGC	ACTCAAGACTCTGATACCAGTAATTGG	170	60.65	40.91	TTTCT(Ins)	
rs7744190	14	100844509	GGATTGTGCTTGGTGCTA	GTTGCAGTGAGCCAAGAT	190	59.86	50	TTTT(Del)	
rs2012552	3	11180390	AACAAGACCATCCCCTGGTA	ATCATGCCACTGCACTCAAA	118	59.26	50	TTT(Del)	
rs3753676	4	189102749	GGGAAGTTTCAGGGTTATTGGC	GTGTCTCTTCTTTCACGATCCCT	144	59.5	50	AAAT(Del)	
rs5568428	12	52191193	CGCCAAGCTCACTCTCTGTT	GAAAGGTCTGGGGAGAAACC	157	60.73	55	AAG(Ins)	
rs1406420	13	100095681	CTAGGTTCCTGTCGAGCCCA	GCTCAGCAAGTCAGTTCTTTGG	184	60.97	60	AAAAAG(Ins)	
rs1125972	21	47485737	TCCAGCAGACACCTAGGGC	AGAAAGGTCTGGGGAGAAACC	197	60.99	63.16	AGA(Ins)	

IV. . Conclusions

A. Discussion of findings:

In the results section of the report we have demonstrated that our initial selection of HID INDELs in our preliminary data are quite capable of robustly identifying individuals in admixed populations outside of North America (South America and North Africa). However some caution should be used when conferring identity to individuals from isolated populations. STR methods face similar issues, but the bi-allelic nature of INDELs and relative genetic stability make this expected in such niche populations (Brazilian Natives for example).

We also demonstrated that INDEL and INNUL markers by virtue of short amplicon design are better able to generate full profiles than standard STR panels with both DNA degraded by embalming and DNA sheared by the concussive blast of a simulated IED. The lowered RMP and inability to adequately resolve mixtures makes its routine use less useful, but when DNA degradation causes significant allele drop out, the utility of these highly sensitive markers becomes apparent. They are able to type samples that would be routinely relegated to mitochondrial DNA typing, but because the markers are not linked, that they have a much greater discriminatory power when multiplexed.

To address the problem with mixtures, we sequenced small amplicon enriched libraries surrounding INDELs. The sequencing was able to detect other variants in the flanking regionsthat could be combined into a microhaplotype which could give greater polymorphism to the INDEL increasing its discriminatory power and increasing the ability to resolve mixtures in samples typed with INDELs

We developed a single tube multiplex assay with 39 of the most highly performing INDEL markers and optimized for the follow on study to be performed once our final project funds are released.

Additionally we utilized in silico resources to identify INDELs that could be utilized to identify the biogeographic ancestry of an individual in relation to global major populations. A singletube multiplex PCR reaction for CE detection was developed for this panel of INDELs as well, and optimized it to reliably type degraded DNA samples for this purpose. This multiplex will also be validated upon the release of final project funds from NIJ and our revised report will reflect the results of both the HID and AIM INDEL developmental validations along with guidelines for use.

Finally, we identified a secondary AIM panel complete with putative primers that would be able to distinguish the biogeographic ancestry of individuals from Southwest Asian and Southwest Hispanic populations once the individual was determined to be from an admixed population by our primary AIM panel.

B. Implications for policy and practice

Our study has demonstrated the feasabiltiy of INDELs as supplemental markers with degraded samples. The discriminatory power of STR markers is well documented and the technology is much better for the majority of forensic cases. However, With highly degraded samples, (bomb fragments, degraded remains, chemically degraded remains, ancient DNA, and in some cases rootless hair shafts), INDELs are a fragment length based (use same instrumentation and analysis methods as standard STR's) with a very high discriminatory power without incurring the cost of sanger sequencing or MPS. In any instance where low quality sample would merit mitochondrial genotyping it is more than likely that INDELs would perform as well if not better. Additionally, since this technology is very similar to STR 's, then these markers are amenable to Rapid DNA genotyping in field forward environments with a much greater tolerance for degraded DNA.

C. Implications for further research

In the future INDEL multiplexes should utilize some of the increased/enhanced CE dye channels to include a few highly polymorphic STR's designed with mini-primers. This would allow an investigator to reliably detect the presence of a mixture in a sample. Additionally as evidenced by our MPS study, unlabeled INDEL primers could be spiked into standard CE based STR reactions to serve as an additional set of markers for subsequent analysis via MPS in the event that the CE based STR genotyping fails due to environmental degradation of the sample prior to collection (for example with touch DNA).

V. Selected References

1. Wilson, M.R., Polanskey, D., Butler, J., DiZinno, J.A., Replogle, J., and Budowle, B.: Extraction, PCR amplification, and sequencing of mitochondrial DNA from human hair shafts. BioTechniques 18:662-669, 1995.
2. Wilson, M.R., DiZinno, J.A., Polanskey, D., Replogle, J., and Budowle, B.: Validation of mitochondrial DNA sequencing for forensic casework analysis. Int. Journal Leg. Med. 108:68-74, 1995.
3. Budowle, B., Allard, M., Fisher, C.L., Isenberg, A.R., Monson, K.L., Stewart, J.E.B., Wilson, M.R., and Miller, K.W.P.: HVI and HVII Mitochondrial DNA population data in Apaches and Navajos. Int. J. Leg. Med. 116(4):212-215.
4. Budowle, B., Polanskey, D., Allard, M.W., and Chakraborty, R.: Addressing the use of phylogenetics for identification of sequences in error in the SWGDAM mitochondrial DNA database. J. Forens. Sci. 49(6):1256-1261, 2004
5. Budowle, B., Planz, J., Campbell, R., and Eisenberg, A.: SNPs and microarray technology in forensic genetics: development and application to mitochondrial DNA. Forens. Sci. Rev. 16:22-36, 2004.
6. Budowle, B.: SNP typing strategies. Forensic Sci. Int. 146 Suppl: S139-S142, 2004.
7. Budowle, B. and van Daal, A.: Forensically relevant SNP classes. Biotechniques 44(5):603-610, 2008.
8. Andreasson, H., Nilsson, M., Budowle, B., Lundberg, H., and Allen, M.: Nuclear and mitochondrial DNA quantification of various forensic materials. Forens. Sci. Int. 164:56-64, 2006.
9. Budowle, B., Ge, J., Aranda, X.G., Planz, J.V., Eisenberg, A.J., and Chakraborty, R.: Texas population substructure and estimating the rarity of Y STR haplotypes in forensic analyses. J. Forens. Sci. 54(5):1016-1021, 2009.
10. Budowle, B. and van Daal, A.: Extracting evidence from forensic DNA analyses: future molecular biology directions. Biotechniques 46(5):339-350, 2009.
11. Hall, T.A., Sannes-Lowery, K.A., McCurdy, L.D., Fisher, C., Anderson, T., Henthorne, A., Budowle, B., and Hofstadler, S.A.: Base composition profiling of human Mitochondrial DNA using PCR and direct automated electrospray ionization mass spectrometry. Analytical Chem. 81(18):7515-7526, 2009.
12. Budowle, B., Polanskey, D., Fisher, C.L., Den Hartog, B.K., Kepler, R.B., and Elling, J.W.: Automated alignment and nomenclature for consistent treatment of polymorphisms in the human mitochondrial DNA control region. J. Forens. Sci. 55(5):1190-1195, 2010.
13. Ge, J., Budowle, B., Planz, J.V., and Chakraborty, R.: Haplotype block: a new type of forensic DNA marker. Int. J. Leg. Med. 124(5):353-361, 2010.
14. Cummings, C.A., Bormann-Chung, C.A., Fang, R., Barker, M., Brzoska, P., Williamson, P.C., Beaudry, J., Matthews, M., Schupp, J., Wagner, D.M., Birdsell, D., Vogler, A.J., Furtado, M.R., Keim P., and Budowle, B.: Accurate, rapid, and high-throughput detection of strain-specific polymorphisms in Bacillus anthracis and Yersinia pestis by next-generation sequencing. BMC Investigative Genetics 1:5, 2010.
15. Budowle, B.: Familial searching: extending the investigative lead potential of DNA typing. Profiles in DNA 13(2), 2010, Available at: www.promega.com/profiles/1302/1302_07.html.
16. Ge, J., Budowle, B., Planz, J., Eisenberg, a., Ballantyne, J., and Chakraborty, R.: U.S. forensic Y chromosome short tandem repeats database. Leg. Med. 12(6):289-295, 2010.
17. Budowle, B., Ge, J., Chakraborty, R., Eisenberg, A.J., Green, R., Mulero, J., Lagace, R., and Hennessy, L.: Population Genetic Analyses of the NGM STR Loci. Int. J. Leg. Med. 125:101-109, 2011.
18. Ge, J., Chakraborty, R., Eisenberg, A. and Budowle, B.: Comparisons of the familial DNA database searching policies. J. Forens. Sci. 56(6):1448-1456, 2011.
19. Kavlick, M.F., Lawrence, H.S., Merritt, R.T., Fisher, C., Isenberg, A., Robertson, J.M., and Budowle, B.: Quantification of human mitochondrial DNA using synthesized DNA standards. J. Forens. Sci. 56(6):1457-1463, 2011
20. Frumkin, D., Wasserstrom, A., Budowle, B., and Davidson, A.: DNA methylation-based forensic tissue identification. Forens. Sci. Int. Genet. 5(5):517-524, 2011.
21. Davis, C., Ge, J., Chidambaram, A., King, J., Turnbough, M., Collins, M., Dym, O., Chakraborty, R., Eisenberg, A.J., and Budowle, B.: Y-STR loci diversity in native Alaskan populations. Int. J. Leg. Med. 125:559–563, 2011.
22. Budowle, B., Ge, J., Chakraborty, R., and Gill-King, H.: Use of prior odds for missing persons identifications. BMC Investigative Genetics 2:15, 2011.
23. Ge, J., Eisenberg, A., and Budowle, B.: Developing criteria and data to determine best options for expanding the core CODIS loci. BMC Investigative Genetics 3:1, 2012.

24. LaRue, B.L., Ge, J., King, J.L., and Budowle, B.: A validation study of the Qiagen Investigator DIPplex® Kit; an INDEL based assay for human identification. Int. J. Leg. Med. (126: 533-540, 2012).
25. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: A simulation study. Mol Ecol 2005;14(8):2611-20.
26. Jakobsson M, Rosenberg NA. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 2007 Jul;23(14):1801-6.
27. Rosenberg NA. DISTRUCT: A program for the graphical display of population structure. Mol Ecol Notes 2004;4:137-8.
28. Gettings KB, Kiesler KM, Vallone PM. Performance of a next generation sequencing SNP assay on degraded DNA. Forensic Sci Int Genet. 2015 Nov;19:1-9. doi: 10.1016/j.fsigen.2015.04.010. Epub 2015 May 27. PubMed PMID: 26036183.
29. Wendt FR, Churchill JD, Novroski NMM, King JL, Hg J, Oldt RF, McCulloh KL, Weise JA, Smith DG, Kanthaswamy S, Budowle B. Genetic analysis of the Yavapai Native Americans from West-Central Arizona. Forensic Sci Int Genet. http://dx.doi.org/10.1016/j.fsigen.2016.05.008
30. Quail MA, Smith M, Coupland P, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012 Jul 24;13:341. doi: 10.1186/1471-2164-13-341. PubMed PMID: 22827831; PubMed Central PMCID: PMC3431227.
31. Zeng X, King JL, Stoljarova M, et al. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Sci Int Genet. 2015 May;16:38-47. doi: 10.1016/j.fsigen.2014.11.022. Epub 2014 Dec 3. PubMed PMID: 25528025. 32. King JL, LaRue BL, Novroski NM, et al. High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet. 2014 Sep;12:128-35. doi: 10.1016/j.fsigen.2014.06.001. Epub 2014 Jun 7. PubMed PMID: 24973578.
33. Fordyce SL, Mogensen HS, Børsting C, et al. Second-generation sequencing of forensic STRs using the Ion Torrent™ HID STR 10-plex and the Ion PGM™. Forensic Sci Int Genet. 2015 J 438 an;14:132-40. doi: 10.1016/j.fsigen.2014.09.020. Epub 2014 Oct 5. PubMed PMID: 25450784.
34. Churchill JD, Chang J, Ge J, et al. Blind study evaluation illustrates utility of the Ion PGM™ system for use in human identity DNA typing. Croat Med J. 2015 Jun 19;56(3):218-29. PubMed PMID: 26088846.
35. Warshauer DH, King JL, Budowle B. STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. Forensic Sci Int Genet. 2015 Jan;14:182-6. doi: 10.1016/j.fsigen.2014.10.011. Epub 2014 Oct 22. PubMed PMID: 25450790.
36.QIAamp® DNA Mini and Blood Mini Handbook, 3rd Edition, June 2012. https://www.qiagen.com/us/resources/resourcedetail?id=67893a91-946f-49b5-8033-394fa5d752ea&lang=en.
37. Zeng X, Warshauer DH, King JL, Churchill JD, Chakraborty R, Budowle B. Empirical testing of a 23-AIMs panel of SNPs for ancestry evaluations in four major US populations. Int J Legal Med. 2016 Feb 25. [Epub ahead of print] PubMed PMID: 26914801.
38.MiSeq System User Guide.https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-system-guide-15027617-o.pdf.
39.Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high performance genomics data visualization and exploration. Brief Bioinform. 2013 Mar;14(2):178-92. doi: 10.1093/bib/bbs017. Epub 2012 Apr 19. PubMed PMID: 22517427; PubMed Central PMCID: PMC3603213
40.Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24-6. doi: 10.1038/nbt.1754. PubMed PMID: 21221095; PubMed Central PMCID: PMC3346182.
41.Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PubMed PMID: 19505943; PubMed Central PMCID: PMC2723002.
42.Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93. doi: 10.1093/bioinformatics/btr509. Epub 2011 Sep 8. PubMed PMID: 21903627; PubMed Central PMCID: PMC3198575.

43.Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18. PubMed PMID: 19451168; PubMed Central PMCID: PMC2705234.
44.McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19. PubMed PMID: 20644199; PubMed Central PMCID: PMC2928508.
45.Genetic Data Analysis Software. Lewis and Zaykin. 1999.
46.Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11.
47.Leclercq S, Rivals E, Jarne P. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. Genome Biol Evol. 2010 Jul 12;2:325-35. doi: 10.1093/gbe/evq023. PubMed PMID: 20624737; PubMed Central PMCID: PMC2997547.
48.Fan H, Chu JY. A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics. 2007 Feb;5(1):7-14. Review. PubMed PMID: 17572359.
49.Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci U S A. 1997 Feb 4;94(3):1041-6. PubMed PMID: 9023379; PubMed Central PMCID: PMC19636
50.Kidd KK, Pakstis AJ, Speed WC, Lagacé R, Chang J, Wootton S, Haigh E, Kidd JR. Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. Forensic Sci Int Genet. 2014 Sep;12:215-24. doi: 10.1016/j.fsigen.2014.06.014. Epub 2014 Jul 1. PubMed PMID: 25038325.
51.Kidd KK, Speed WC. Criteria for selecting microhaplotypes: mixture detection and deconvolution. Investig Genet. 2015 Jan 28;6:1. doi: 10.1186/s13323-014-0018-3. eCollection 2015. PubMed PMID: 25750707; PubMed Central PMCID: PMC4351693.
52.Ge J, Budowle B, Planz JV, Chakraborty R. Haplotype block: a new type of forensic DNA markers. Int J Legal Med. 2010 Sep;124(5):353-61. doi: 10.1007/s00414-009-0400-5. Epub 2009 Dec 22. PubMed PMID: 20033199.
53.Laurin N, Milot E. Population genetic data of the AmpFℓSTR® Identifiler® Plus and PowerPlex® 16 HS STR loci in four Canadian populations. Int J Legal Med. 2014 Mar;128(2):277-8. doi: 10.1007/s00414-013-0899-3. Epub 2013 Aug 13. PubMed PMID: 23942885.
56.Tomas C, Mogensen HS, Friis SL, Hallenberg C, Stene MC, Morling N. Concordance study and population frequencies for 16 autosomal STRs analyzed with PowerPlex® ESI 17 and AmpFℓSTR® NGM SElect™ in Somalis, Danes and Greenlanders.Forensic Sci Int Genet. 2014 Jul;11:e18-21. doi: 10.1016/j.fsigen.2014.04.004. Epub 2014 Apr 18. PubMed PMID: 24810256.
57.Turrina S, Ferrian M, Caratti S, De Leo D. Evaluation of genetic parameters of 22 autosomal STR loci (PowerPlex® Fusion System) in a population sample from Northern Italy. Int J Legal Med. 2014 Mar;128(2):281-3. doi: 10.1007/s00414-013-0934-4. Epub 2013 Nov 2. PubMed PMID: 24185983.

VI. Dissemination of Research Findings

Publications

Evaluation of a 49 InDel Marker HID panel in two specific populations of South America and one population of Northern Africa RS Moura-Neto, R Silva, IC Mello, T Nogueira, AA Al-Deib, B LaRue, J King, B Budowle International Journal of Legal Medicine 129 (2), 245-249

Massively parallel sequencing of 68 insertion/deletion markers identifies novel microhaplotypes for utility in human identity testing Frank R Wendt, David H Warshauer, Xiangpei Zeng, Jennifer D Churchill, Nicole MM Novroski, Bing Song, Jonathan L King, Bobby L LaRue, Bruce Budowle Forensic Science International: Genetics 25, 198-209

Development and Validation of a Novel Multiplexed DNA Analysis System, InnoTyper® 21

Hiromi Brown, Robyn Thompson, Gina Pineda Murphy, Dixie Peters, Bobby La Rue, Jonathan King, Anne H Montgomery, Marion Carroll, James Baus, Sid Sinha, Frank Wendt, Bing Song, Ranajit Chakraborty, Bruce Budowle and Sudhir K. Sinha Submitted FSI: Genetics

The Effectiveness of Various Strategies to Improve DNA Analysis of Formaldehyde-Damaged Tissues for Human Identification Purposes Natalia Czado, MS[1], Bobby LaRue, PhD[2], Amanda Wheeler, MS[1], Rachel Houston BS[1], Amy Sorensen MS[1], Kelly Grisedale, PhD[3], David Gangitano, PhD[1], Sheree Hughes-Stamm, PhD[1]. Pre-submission; FSI: Genetics

Use of alternative genotyping strategies to genotyped degraded DNA samples from improvised explosive devices Tasker E, Houston R, Sorensen A, LaRue B, Gangitano D, Hughes-Stamm Pre-submission; International Journal of Legal Medicine

Selection of a panel of 60 Ancestral Insertion and Deletion (INDEL) Markers that will distinguish between four major global populations Thompson L, Zeng X, Sage K, Sturm S, King JL, Budowle B, LaRue BL. . Human Genetics; 2016 manuscript in preparation

Developmental validation of a novel panel of 49 insertion/deletion markers (INDELs) for human identification for capillary electrophoresis. Sage K, Sturm S, Thompson L, King JL, Budowle B, LaRue BLInternational Journal of Legal Medicine; 2016 manuscript in preparation

Developmental validation of a novel panel of insertion/deletion markers (INDELs) for ancestral informativeness for capillary electrophoresis. Sturm S, Sage K, Song B, Zeng X, Thompson L, King JL, Budowle B, LaRue BL. Forensic Science International: Genetics; 2016 manuscript in preparation

Presentations

LaRue BL, Sinha SK, Montgomery AH, Pineda G, Thompson RA, King JL, Ge J, Chakraborty R, Budowle B *Innotyper- A novel approach to genotyping degraded DNA samples utilizing LINEs and SINEs* **AFDAA Summer meeting 2014**

Thompson L, King JL, Budowle B, and LaRue BL *Development of Insertion-Deletion (INDEL) panels for ancestral and individual identity genotyping.* **25th International Symposium for Identification 2014**

Kelly A. Sage, Sarah Strum, Bing Song, Jonathan L. King, Bruce Budowle, Bobby L. LaRue. *Two Novel Multiplex INDEL Assays for Determining Ancestral and Individual Identity for Use with Degraded DNA samples.* **26th Annual International Symposium on Human Identification. 2015**

Frank R Wendta, B.S., David H Warshauerb, Ph.D., Xiangpei Zenga, Ph.D., Jennifer D Churchilla, Ph.D., Nicole MM Novroskia, M.S., Bing Songa, B.S., Jonathan L Kinga, M.S., Bobby L LaRuea, Ph.D., Bruce Budowlea,c, Ph.D. *Sequencing of 68 Insertion/Deletion Markers: Motif and Microhaplotypes*
27th Annual International Symposium on Human Identification. 2016

Bing Song, Lindsey M. Tompson, Xiangpei Zeng, Sarah Sturm, Kelly Sage, Frank R.Wendt, Jonathan King, Ranajit Chakraborty, Bruce Budowle, Bobby LaRue, *Selection of a Supplementary Ancestry-Informative Marker (AIM) Panel of INDELs for Distinguishing Southwest Hispanics and Southwest Asians* **27th Annual International Symposium on Human Identification. 2016**