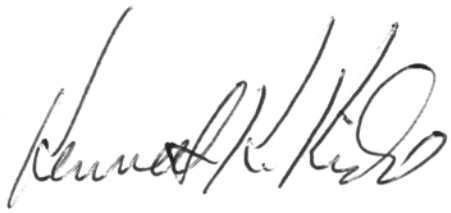
Final Summary Overview
National Institute of Justice
Grant # 2015-DN-BX-K023
Starting 01-01-2016 Ending 12-31-2018
Project Title:
"Powerful Forensic Markers Optimized for Massively Parallel Sequencing"

Kenneth K. Kidd (PI)
Professor Emeritus of Genetics
Senior Research Scientist in Genetics
Email: Kenneth.Kidd@yale.edu
Telephone: 203-785-2654
Department of Genetics
Yale University School of Medicine

Revision submitted by K.K. Kidd on May 6, 2019

Signature of submitting official



What are the major goals and objectives of this project?

As stated in the funding proposal: "Our overall purpose continues to be to develop better sets of SNP markers for forensics using our genomic analysis expertise and unique resources to document the validity of these panels for their specific purposes. Our goals for this application are: (1) to enhance our developing ancestry informative SNP (AISNP) panels to provide more robust differentiation among populations for an accurate estimate of an individual's ancestral origin at even finer geographic levels with an optimized multi-tier panel of SNPs, (2) to identify and characterize additional microhaplotypes that can be genotyped by massively parallel sequencing (MPS)."

What was accomplished under these goals?

Our accomplishments are best measured by the published papers and the growing list of citations of those papers. The publications are listed at the end of this Project Overview. The talks presented at international meetings are also relevant as many submitted abstracts have been accepted for oral presentation, including by the National Institute of Justice (NIJ) at the 2018 NIJ Conference associated with the AAFS meeting in Seattle but are not summarized since the material covered is included in the full publications.

Considering just the AISNP panels there are now 25 new reference populations with frequencies on all 55 of the Kidd panel AISNPs. These new reference populations and the SNP frequencies have been entered into ALFRED and FROG-kb during the July-December 2018 reporting period. There are now 164 total reference populations (up from the 139 at the beginning of the grant, Pakstis et al., 2017) for the 55 AISNP panel in FROG-kb. Studies of additional populations have been completed (many still unpublished) for a large number of microhaplotypes and comparisons of microhaps selected for different purposes have been completed.

Project Design and Methods

Briefly, the design has been to identify candidate marker loci using publicly available data and our own laboratory resources. Once identified, the better of the candidate loci have been genotyped on our resources of over 2500 individuals from 57 distinct population samples. We previously established cell lines on these individuals and therefore have large amounts of DNA. Based on evaluation in those populations the more promising markers have been tested on additional population samples. Analyses of these additional samples are based on small amounts of DNA collected by collaborators and shipped to us for analysis. Data analyses have involved multiple biostatistical methods to identify those markers that are best for different forensic applications.

Data Collection

Purified DNA is typed by TaqMan assays obtained from ThermoFisher. Robotic DNA handling allows moderate throughput of the specific SNPs identified as likely to be informative. For the tests of the DNA-only samples a multiplex pre-amplification protocol developed in the lab allows 100 SNPs to be typed on the amount of DNA needed to type a single SNP. Computer programs allow interpretation of the ABI SDS TaqMan reads into individual genotypes and uploading into the lab's genotype database.

Findings and Analyses in Progress

AISNPs

A database of several thousand SNPs typed on all 2500 individuals represented by cell-line DNA has been assembled; over 1000 of those have been added by this project. In addition, data on over 600,000 SNPs already exist for over half of those individuals. This large database allows new searches as new data are added and new statistics, such as random forest, are implemented to identify good markers for distinguishing among new groups of populations such as the Southwest Asia populations currently being analyzed.

Microhaplotypes

So far, 182 microhaplotypes (also referred to as microhaps) have been characterized on our 57 core lab populations, a significant increase from the 129 that were characterized when the current project started. More than 200 candidate microhaps have been considered along the way and more than 25 have been discarded based on preliminary data showing them to be much less informative than the top half of the existing microhaps. Data collection is ongoing for microhaplotype data on a dozen new populations from Southwest Asia, North Africa, and Central Asia on which we have DNA from collaborators. We have evaluated the existing 182 microhaps with complete data on 83 populations (our data from cell lines plus 1000 Genomes data) and only the "best" microhaps are being pursued. We are exploring the differences

between the best 50 microhaps selected using effective number of alleles, A_e, and the best 50 selected using informativeness, I_n. The union of those two sets (78 microhaps) is the focus of data collection on the new DNA-only populations. We have also identified several microhaps that can be elaborated by typing one or more additional SNPs within the current extent or very close to the existing microhap. This future work will increase the number of highly informative microhaps both for mixture deconvolution and ancestry inference.

Data Uploaded to ALFRED Database

We continue to contribute new allele frequency data from our lab to the ALFRED database. New frequency data is entered into ALFRED coordinated with the submission of new publications employing the new data. Allele and haplotype frequency data for all individual SNPs and for the haplotypes that we have been studying have already been made public in ALFRED. Searching ALFRED with the key word “microhap” retrieves a table with links for a total of 198 of the microhaps including 33 microhaps published in two papers by Chinese researchers in mid-2018.

Impact

The only highly differentiating sets of AISNPs that are both extensively validated on a large number of individuals and populations and are currently available in the public domain are the three developed or studied by us (Kidd et al., 2014), by the Seldin Lab (Kosoy et al., 2009; Kidd et al., 2011), and by Nievergelt et al. (2013). To date, only the Kidd 55 AISNP panel is commercially available from both Verogen and Thermo Fisher. The Seldin Lab panel is also part of the ThermoFisher kit. Given the desire of several U.S. Government agencies (personal communication), and many forensic labs in general for small (≤200 SNPs) panels of ancestry informative SNPs, results of our work to improve biogeographic resolution and robustness are likely to be made commercially available as kits. (Both Illumina/Verogen and ThermoFisher have recognized our 55 AISNP panel as one of the better, if not the best, ancestry panels.) Investigators can use commercial ancestry companies, but their markers and statistics are often

proprietary and the underlying science unavailable. Forensic laboratories may be reluctant to use such labs for those reasons. Currently, no microhaplotypes are commercially available for forensic use but Thermo Fisher is developing one (cf., Kidd et al., 2014, 2015) involving 74 of our better microhaps. Our extensively validated and documented data and our analyses of those data already exist in the public domain through FROG-kb and/or ALFRED. Because of the extensive public documentation, forensic labs will have greater reason to use these markers than proprietary ones. Our global data on Phenotype Informative SNPs will provide bases for preventing simplistic/erroneous interpretation (cf., Yun et al., 2014) until the biological basis for interpretation is clear. We have shown theoretically and practically that microhaplotypes allow excellent mixture identification and resolution, much better than STRPs. Similarly, familial relationships can be much more definitively demonstrated with the microhaps now being developed into commercial kits than with the existing STRP kits.

The high impact of this project over the years is documented by the high percentiles for citations of our papers (see Table 1). While the most recent papers are too new (less than 2 years since publication) to be evaluated (labeled as “not rated” in Table 1) we expect similar high percentiles based on the citations we already know of.

Our work on developing LISNPs in the form of microhaps, may be especially important in mass disaster situations in which ethnicity AND extended family matching may be extremely important. We are providing the basic population data to make such compound markers statistically tractable.

Practice. We expect the SNPs identified as part of this project (up to now and in ongoing work) to be useful for many investigative purposes. To the degree that SNPs identified from this study are brought before the courts, this work will provide a firm scientific basis for their acceptability. Data collected on the SNPs identified will provide a strong statistical foundation for conclusions when used as investigative tools for inference of ancestry or family/clan membership. As forensic labs begin sequencing using MiSeq or PGM, our results will be placed into practice

because both Illumina/Verogen and ThermoFisher kits have already incorporated our 55-AISNP and 45-IISNP panels into MPS kits. ThermoFisher is working with us to develop a commercial kit incorporating 74 of our better microhaps. Both companies have expressed interest in adding more AISNPs when we have verified them for better resolution of ancestry. We think that with the growing importance of mixture deconvolution and probabilistic genotyping using STRPs, the superior performance of MPS with microhaps will result in rapid incorporation of the new technology into practice. The Precision ID Ancestry Panel of ThermoFisher for use with MPS is already validated and implemented for casework in one forensic laboratory (Jin et al., 2018). It is worth noting that the majority of the reference population data in the ThermoFisher Torrent Suite software are allele frequencies generated in our laboratory that were taken from material we made public in ALFRED. This is another development attributable to our NIJ grants.

What is the impact on other disciplines?

The data being collected and made public are useful for research in many aspects of anthropology and recent human evolution. For example, several of our papers are written to emphasize the relationships among populations from an anthropologic perspective. However, the data were collected for and incorporated in research on the ancestry inference forensic AISNP panels.

How were the results disseminated to communities of interest?

The primary mode of dissemination is through publications and through the accessibility of the data in ALFRED and FROG-kb. In addition, Dr. Kidd has presented nine talks at scientific meetings and university departments and posters at two meetings.

Literature cited in this report

Butler, J. M., B. Budowle, P. Gill, K. K. Kidd, C. Phillips, P. M. Schneider, P. M. Vallone, and N. Morling, 2008. Report on ISFG SNP Panel Discussion. *Forensic Science International: Genetics Supplement Series* 1: 471–472.

Fang R., A.J. Pakstis, F. Hyland, D. Wang, J. Shewale, J.R. Kidd, K.K. Kidd, M.R. Furtado. 2009. Multiplexed SNP detection panels for human identification. *Forensic Science International: Genetics Supplement Series* 2:538-539.

Jin, S., M. Chase, M. Henry, G. Alderson, J.M. Morrow, S. Malik, D. Ballard, J. McGrory, N. Fernandopulle, J. Millman, J. Laird, 2018. Implementing a biogeographic ancestry inference service for forensic casework. *Electrophoresis* doi: 10.1002/elps.201800171; online Aug 20, 2018.

Kidd, J.R., F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, K.K. Kidd (2011). Analyses of a set of 128 ancestry informative SNPs (AISNPs) in a global set of 119 population samples. *Investigative Genetics* 2:1.

Kidd, K.K, A.J. Pakstis, W.C. Speed, R. Lagace, J. Chang, S. Wootton, N. Ihuegbu, 2013. Microhaplotype loci are a powerful new type of forensic marker. *Forensic Science International: Genetics Supplement Series* 4:e123-e124.

Kidd, K.K., W. C. Speed, A. J. Pakstis, M. R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, F. R. Friedlaender and J. R. Kidd (2014). Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Science International: Genetics* 10: 23-32.

Kidd,K.K., W.C. Speed, S. Wootton, R. Lagace, R. Langit, E. Haigh, J. Chang, A.J. Pakstis, 2015. Genetic markers for massively parallel sequencing in forensics. *Forensic Science International: Genetics Supplement Series* 5:e677-e679.

Kosoy R, Nassir R, Tian C, White PA, Butler LLM, Silva G, Kittles R, Alarcon- Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF (2009) Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Human Mutation* 30:69-78.

Nievergelt, C.M., A.X. Maihofer, T. Shekhtman, O. Libiger, X. Wang, K.K. Kidd, J.R. Kidd, 2013. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investigative Genetics* 4:13.

Pakstis, A. J., W. C. Speed, J. R. Kidd, and K. K. Kidd, 2008. SNPs for Individual Identification. Forensic Science International: Genetics: Genetics Supplement Series 1: 479–481.

Yun, L., Y. Gu, H. Rajeevan, K.K. Kidd (2014) Application of six IrisPlex SNPs and comparison of two eye color prediction systems in diverse Eurasia populations. *International Journal of Legal Medicine* 128:447-453.

Manuscripts submitted/under review

Andrew J. Pakstis, Cemal Gurkan, Mustafa Dogan, Hasan E Balkaya, Serkan Dogan, Pavlos Neophytou, Lotfi Cherni, Sami Boussetta, Houssein Khodjet-El-Khil, Amel Ben Ammar ElGaaied, Nina Mjolsnes, Kirstin Janssen, Gunn-Hege Olsen, Sibte Hadi, Eida Khalaf Almohammed, Vania Pereira, Ditte Mikkelsen Truelsen, Ozlem Bulbul, Usha Soundararajan, Haseena Rajeevan, Judith R. Kidd, Kenneth K. Kidd, 2018. Genetic relationships for European, Mediterranean and SW Asian Populations using a panel of 55 AISNPs**.** Submitted Oct 30, 2018; Ms revised after initial reviews; still in review process as of May 3, 2019.

Manuscripts accepted in 2019 but not yet published

Petros Drineas, Fotis Tsetsos, Anna Plantinga, Iosif Lazaridis, Evangelia Yannaki, Anna Razou, Katerina Kanaki, Manolis Michalodimitrakis, Francisco Perez-Jimenez, Giustina De Silvestro, Maria C Renda, John A Stamatoyannopoulos, Kenneth Kidd, Brian Browning, Peristera Paschou, George Stamatoyannopoulos, 2019. Genetic history of the population of Crete. *Annals of Human Genetics*. In Press as of May 1, 2019.

Manuscripts accepted/published--Oct 2018 to Mar 2019

Fabio Oldoni, Kenneth K. Kidd, Daniele Podini, 2019. Microhaplotypes in forensic genetics. *Forensic Science International: Genetics* 38:54-69, doi: 10.1016/j.fsigen.2018.09.009.

Lindsay Bennett, Fabio Oldoni, Kelly Long, Selena Cisana, Katrina Maddela, Sharon Wootton, Joseph Chang, Ryo Hasegawa; Robert Lagace, Kenneth K. Kidd, Daniele Podini, 2019. Mixture deconvolution by massively parallel sequencing of microhaplotypes. *International Journal of Legal Medicine* 133:719-729, doi: 10.1007/s00414-019-02010-7. In Press Jan. 22, 2019. Online Feb 13, 2019. Erratum published May 2019: “The original version of this article contained an author name error. In this article, Katrina Madella has been corrected to Katrina Maddela.”

Sami Boussetta, Lotfi Cherni, Andrew J. Pakstis, Nesrine Ben Salem, Sarra Elkamel, Kenneth K. Kidd, Houssein Khodjet-el-Khil, Amel Ben Ammar ElGaaied (2019) Usefulness of COMT gene polymorphism in North African populations. *Gene* 696:186-196, doi: 10.1016/j.gene.2019.02.021. In press Feb 1, 2019; online Feb 18, 2019; Feb 27, 2019 assigned to May issue.

Publications 2016 to Sept 2018

U. Soundararajan, L. Yun, M. Shi, K.K. Kidd, 2016. Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. *Forensic Science International: Genetics* 23:25-32.

Cai-Xia Li, Andrew J. Pakstis, Li Jiang, Yi-Liang Wei, Qi-Fan Sun, Hong Wu, Ozlem Bulbul, Ping Wang, Long-Li Kang, Judith R. Kidd, Kenneth. K. Kidd, 2016. A panel of 74 AISNPs: Improved Ancestry Inference within Eastern Asia. *Forensic Science International: Genetics*. 23:101-110.

Lotfi Cherni, Andrew J. Pakstis, Sami Boussetta, Sarra Elkamel, Sabeh Frigi, Houssein Khodjet-El-Khil, Alison Barton, Eva Haigh, William C. Speed, Amel Ben Ammar Elgaaied, Judith R. Kidd, Kenneth K. Kidd, 2016. Genetic variation in Tunisia in the context of human diversity worldwide. *American Journal of Physical Anthropology* 161:62-71.

Ozlem Bulbul, Lotfi Cherni, Houssein El Khil Khodjet, Haseena Rajeevana, Kenneth K. Kidd, 2016. Evaluating a subset of ancestry informative SNPs for discriminating among Southwest Asian and Circum-Mediterranean populations. *Forensic Science International: Genetics*. 23:153-158.

Kenneth K. Kidd, 2016. Proposed nomenclature for microhaplotypes. *Human Genomics* **10**:16; DOI**:**10.1186/s40246-016-0078-y

Andrew J Pakstis, Longli Kang, Lijun Liu, Zhiying Zhang, Tianbo Jin, Elena L. Grigorenko, Frank R.Wendt, Bruce Budowle, Sibte Hadi, Mariam Salam Al Qahtani, Niels Morling, Helle Smidt Mogensen, Goncalo E. Themudo, Usha Soundararajan, Haseena Rajeevan, Judith R. Kidd,

Kenneth K. Kidd, **2017**. Increasing the reference populations for the 55 AISNP panel: the need and benefits. *International Journal of Legal Medicin*e 131:913-917*;*

Kenneth K. Kidd, William C. Speed, Andrew J. Pakstis, Daniele S. Podini, Robert Lagace, Joseph Chang, Sharon Wootton, Eva Haigh, Usha Soundararajan, 2017. Evaluating 130 microhaplotypes across a global Set of 83 populations. *Forensic Science International: Genetics* 29:29-37. In Press March 12, 2017.

G. Stamatoyannopoulos, A. Bose, A. Teodosiadis, F. Tsetsos, A. Plantinga, N. Psatha, N. Zogas, E. Yannaki, P. Zalloua, K.K. Kidd, B.L. Browning, J. Stamatoyannopoulos, P. Paschou, P. Drineas, 2017. Genetics of the Peloponnesean populations and the theory of extinction of the medieval Peloponnesean Greeks. European Journal of Human Genetics 25:637-645.

Kenneth K. Kidd, Usha Soundararajan, Haseena Rajeevan, Andrew J. Pakstis, Katherine N. Moore, Jeri D. Ropero-Miller, **2018**. The redesigned Forensic Research/Reference on Genetics-knowledge base, FROG-kb. *Forensic Science International: Genetics* 33:33-37

Ozlem Bulbul, Andrew J. Pakstis, Usha Soundararajan, Cemal Gurkan, Jane E. Brissenden, Janet M. Roscoe, Baigalmaa Evsanaa, Ariunaa Togtokh, Peristera Paschou, Elena L. Grigorenko, David Gurwitz, Sharon Wootton, Robert Lagace, Joseph Chang, William C. Speed, Kenneth K. Kidd, 2017. Ancestry inference of 96 population samples using microhaplotypes. *International Journal of Legal Medicine* 132:703-711.

Ozlem Bulbul, William C Speed, Cemal Gurkan, Usha Soundararajan, Haseena Rajeevan, Andrew J Pakstis, Kenneth K Kidd, 2018. Improving Ancestry Distinctions among Southwest Asian Populations. *Forensic Science International Genetics* 35:14-20.

Kenneth K. Kidd, Andrew J. Pakstis, William C. Speed, Robert Lagace, Sharon Wootton, Joseph Chang, 2018. Selecting microhaplotypes optimized for different purposes. *Electrophoresis* doi.org/10.1002/elps.201800092; online June 21, 2018.

Sheng Gu, Hui Li, Andrew J. Pakstis, William C. Speed, David Gurwitz, Judith R. Kidd, Kenneth K. Kidd, 2018. Recent selection on a Class I ADH locus distinguishes Southwest Asian populations including Ashkenazi Jews. *Genes* 9, E452, 16pp; doi:10.3390/genes9090452

Table 1. Our forensic-related journal publications sorted by RCR value as ranked by NIH *iCite* website (https://icite.od.nih.gov/analysis) on Sep. 17, 2018. Papers from this project are in bold, full citations are above. This includes 38 of our 43 forensic publications. RCR values are not calculated until one year after the paper has been published. Five of our forensic papers [Butler et al., 2008; Pakstis et al., 2008; Fang et al., 2009; Kidd et al., 2013 and 2015] that appear in the Supplement Series of Forensic Science International: Genetics are not included in RCR calculations at the NIH website presumably because they do not get tracked in PubMed.

PubMed ID	Total Cites	Cites /Year	Relative Citation Ratio (RCR)	NIH Percentile	Pub Year	Citation First Author	Journal
30205534	0	0	Not rated	Not rated	**2018**	**S Gu**	**Genes**
29931757	0	0	Not rated	Not rated	**2018**	**KK Kidd**	**Electrophoresis**
29625264	0	0	Not rated	Not rated	**2018**	**O Bulbul**	**Forensic Sci Int Genet**
29248957	1	1.00	Not rated	Not rated	**2018**	**O Bulbul**	**Int J Legal Med**
29175726	0	0	Not rated	Not rated	**2018**	**KK Kidd**	**Forensic Sci Int Genet**
28272534	0	0	Not rated	Not rated	**2017**	**G Stamatoyannopoulos**	**Eur J Hum Genet**
24508742	44	8.80	3.45	88.2	2014	KK Kidd	Forensic Sci Int Genet
19937056	79	8.78	3.32	87.5	2010	AJ Pakstis	Hum Genet
19456322	79	7.90	3.20	86.8	2009	Hui Li	Ann Hum Genet
16360294	87	6.69	2.67	82.8	2006	KK Kidd	Forensic Sci Int
25038325	25	5.00	2.56	81.8	2014	KK Kidd	Forensic Sci Int Genet
23815888	37	6.17	2.48	80.9	2013	CM Nievergelt	Investig Genet
21208434	59	7.38	2.41	80.1	2011	JR Kidd	Investig Genet
27077960	9	3.00	2.10	76.3	**2016**	**C-X Li**	**Forensic Sci Int Genet**
26977931	9	3.00	2.01	74.9	**2016**	**U Soundararajan**	**Forensic Sci Int Genet**
28359046	6	3.00	1.98	74.4	**2017**	**KK Kidd**	**Forensic Sci Int Genet**
27316555	7	2.33	1.91	73.4	**2016**	**KK Kidd**	**Hum Genomics**
25750707	12	3.00	1.84	72.0	2015	KK Kidd	Investig Genet
17333283	47	3.92	1.67	68.9	2007	AJ Pakstis	Hum Genet
22039151	35	5.00	1.65	68.5	2012	H Rajeevan	Nucleic Acids Res
26355664	14	3.50	1.61	67.7	2015	AJ Pakstis	Forensic Sci Int Genet
27192181	8	2.67	1.59	67.1	**2016**	**Lotfi Cherni**	**Am J Phys Anthropol**
28070634	6	3.00	1.33	60.7	**2017**	**AJ Pakstis**	**Int J Legal Med**
12519999	44	2.75	1.18	56.2	2003	H Rajeevan	Nucleic Acids Res
22938150	18	2.57	1.11	53.9	2012	H Rajeevan	Investig Genet
10592274	40	2.11	1.07	52.4	2000	KH Cheung	Nucleic Acids Res
22445421	15	2.14	1.04	51.5	2012	KK Kidd	Forensic Sci Int Genet
27160361	4	1.33	0.96	48.5	**2016**	**O Bulbul**	**Forensic Sci Int Genet**
12209575	35	2.06	0.82	42.8	2002	MV Osier	Am J Phys Anthropol
22535184	15	2.14	0.82	42.7	2012	AJ Pakstis	Eur J Hum Genet
16028061	28	2.00	0.76	40.1	2005	J-J Kim	Hum Genet
24395150	8	1.60	0.52	28.2	2014	L Yun	Int J Legal Med
26829292	4	1.00	0.48	26.0	2015	JE Brissenden	Hum Biol
11125124	18	1.00	0.45	24.1	2001	MV Osier	Nucleic Acids Res
21913176	9	1.13	0.34	18.0	2011	JR Kidd	Am J Phys Anthropol
19325849	10	0.83	0.26	13.1	2007	H Rajeevan	Evol Bioinform Online
10902212	10	0.53	0.20	9.8	2000	KH Cheung	Pac Symp Biocomput
21668909	1	0.13	0.04	2.4	2011	JN Sampson	Ann Hum Genet

Description: RCR methodology (referenced in Table 1) copied from the NIH iCite website:

"The Relative Citation Ratio is a new metric developed within the Office of Portfolio Analysis (OPA) that represents a citation-based measure of scientific influence of one or more articles. It is calculated as cites/year of each paper, normalized to the citations per year received by NIH-funded papers in the same field and year. This benchmarking process, performed with quantile regression, ensures that a paper with an RCR of 1.0 has received the same number of cites/year as the median NIH-funded paper in its field, while a paper with an RCR of 2.0 has received twice as many cites/year as the median NIH-funded paper in its field."