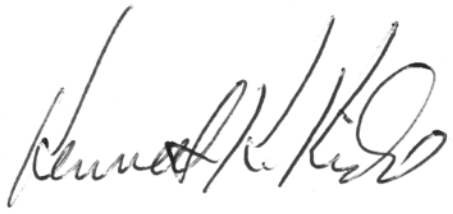
Final Summary Overview
National Institute of Justice
Grant # 2018-75-CX-0041
Starting 01-01-2019 Ending 06-30-2020
Project Title:
"Better Forensic Markers: Microhaplotypes and Ancestry SNPs"

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
Grantee DUNS#; EIN#
Submitted by K.K. Kidd on March 30, 2020
Signature of submitting official



What are the major goals and objectives of this project?

The overall purpose of this project was to develop better sets of SNP markers for forensics using our genomic analysis expertise and unique population resources to document the validity of these panels for their specific purposes. Two kinds of studies have been pursued: those involving panels of single SNPs and those involving microhaplotypes of multiple SNPs. Over the course of this and previous NIJ awards we have accumulated data on large numbers of SNPs on an extensive number of populations involving thousands of individuals. In our project microhaplotypes (microhaps) are DNA segments less than 250-300 basepairs long (hence *micro*) with 2 or more SNPs producing at least three common haplotypes in some populations. With the availability of Massively Parallel Sequencing (MPS) microhaplotype studies are coming to the fore because multiallelic microhaplotypes can convey more information on identity, ancestry, and mixture deconvolution than the same number of individual SNPs.

Our forensic research has been directed both toward identifying SNPs and haplotypes that are especially useful for several types of SNP panels--individual identification SNPs (IISNPs),

ancestry informative SNPs (AISNPs)-- and microhaplotypes informative for identifying all the above plus mixtures. We also have a small component of highly informative very small (<75 bp) microhaplotypes that are especially useful on degraded DNA and on exogenous cell-free DNA in blood. We have not pursued panels of SNPs for phenotype inference although we have data and analyses for SNPs and haplotypes at OCA2 (Kidd et al., in preparation).

What was accomplished under these goals?

Our accomplishments are best measured by the papers published during this project period and the growing list of citations of papers published under previous NIJ awards. The publications during this project period are listed at the end of this Project Overview. Talks were presented at the 2019 Green Mountain DNA Meeting and at the National Institute of Justice (NIJ) at the 2020 NIJ Conference associated with the AAFS meeting in Anaheim California. These are not summarized since the material covered is largely included in the full publications. New reference populations plus SNP and microhaplotype frequencies have been entered into ALFRED and FROG-kb during early 2019. At the end of March, 2019 ALFRED and FROG-kb became static when funding for them ended but they are still accessible on the worldwide web thanks to maintenance support from training grants at the Center for Medical Informatics at Yale University. We have been including detailed frequency tables for new results in the supplementary material of our recent publications (e.g., Pakstis et al., 2019b).

Project Design and Methods

Briefly, the design has been to identify candidate marker loci using publicly available data and our own extensive data resources on SNPs and populations collected over the past decades; that resource was built in part using previous NIJ grants. Once identified, the best 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. For example, the study of Middle East and nearby populations (Pakstis et al., 2019a) relied for the dense detail on DNA samples sent us by many collaborators. There was enough DNA for several of those populations that data on all of the Seldin markers could also be collected, resulting in another paper (Pakstis et al., 2019b). Data analyses have involved multiple biostatistical methods to identify those markers that are best for different forensic applications.

As massively parallel sequencing has become important for microhap studies we have worked to develop a highly informative panel of 90 microhaplotyes (mMHseq) and to add more populations to the mMHseq study (Gandotra et al., 2020).

Data Collection

Purified DNA typed by TaqMan assays obtained from ThermoFisher has been the major source of data for several years. Robotic DNA handling allowed 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 allowed 100 SNPs to be typed on the amount of DNA needed to type a single SNP. Computer programs provided interpretation of the ABI SDS TaqMan reads into individual genotypes and uploading into the lab's genotype database.

The microhaplotypes have been phased using the PHASE program (Stephens et al., 2001) on typing results for the individual SNPs. More recently we have been using the mMHseq panel of microhaplotypes to collect data on populations by MPS yielding phase directly.

Findings

AISNPs

A database of several thousand SNPs typed on all 2500 individuals represented by cell-line DNA has been assembled; over 1000 of those SNPs have been added by this project. In addition, data on over 660,000 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 recently analyzed for our 55-SNP panel for global ancestry inference (Pakstis et al., 2019a).

We have typed all of the Seldin panel of 128 AISNPs on all of our core populations and also on many of the DNA-only samples obtained from collaborators. That resulted in a dataset of the union of the two panels, 170 SNPs, with data complete for 81 populations. The analyses showed that these 81 populations could define 10 to 12 distinct clusters of similar populations (Pakstis et al., 2019b). This is a reference dataset for the commercial MPS kit available from ThermoFisher. We have collaborated on SNP analyses on selected regions of the world (Bousetta et al., 2019; Drineas et al., 2019).

Microhaplotypes

We collaborated on a major review paper on microhaplotypes (Oldoni, Kidd, and Podini, 2019). We participated in studies of mixture deconvolution based on microhaplotypes (Bennett et al. 2019; Oldoni et al., 2019). We collaborated with Runa Daniel and colleagues in helping develop microhaplotypes for the MAPlex assay designed for distinguishing Asian and Pacific populations (Phillips et al., 2019).

So far, 200+ microhaplotypes (microhaps) have been characterized on our 57 core lab populations, a significant increase from the 130 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 has been undertaken for microhaplotypes 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 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 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 work has provided the resource of ongoing analyses that will increase the number of highly informative microhaps both for mixture deconvolution and ancestry inference.

In the design of the mMHseq panel (Gandotra et al., 2020) we selected the 46 best microhaps based on the effective allele number (Ae) of the microhaps we had previously studied and a new search of the 1000 Genomes database for 44 loci with multiple SNPs and very high Ae. A total of 90 microhaps passed quality control in a multiplex typing and with a few additional population samples added to the 1000 Genomes database have been shown to be highly informative for ancestry. The average amplicon size is about 300 bp including the core microhap and the flanking sequence at each locus between the multiplexing PCR primers. That is sufficiently small that any sample that can yield a full result for the CODIS markers will yield a full sequence for these mMHseq loci. Data are available for 155 individual samples from our resources on the web site that was written as part of the project: (https://mmhseq.shinyapps.io/mMHseq/). Since the publication the panel has now been run on

additional 48 individuals for a total of 200+ of our population samples and these new data will be available on the web site soon.

Data Uploaded to ALFRED Database

We contributed new allele frequency data from our lab to the ALFRED database until it became static when funding ended at the end of March, 2019. New frequency data were 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 had 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 Precision ID Ancestry Panel 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, of the relatively small ancestry panels with reference data available on a large number of populations (Pakstis et al., 2019a).) 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. 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 OCA2 SNPs provide some basis for preventing simplistic/erroneous interpretation (cf., Yun et al., 2014) until the biological basis for interpretation of phenotype from genotype 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 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 lineage informative markers 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 and microhaplotypes identified as part of this project to be useful for many investigative purposes. To the degree that SNPs and microhaplotypes 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 worked with us to develop a kit incorporating 74 of our better microhaps but subsequently decided not to make it a commercial product. 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 relatively 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 dissemination method is through publications and through the accessibility of the data in ALFRED, FROG-kb and the mMHseq database https://mmhseq.shinyapps.io/mMHseq/). In addition, Dr. Kidd has presented talks at two scientific meetings during this project period; two additional talks this Spring have been cancelled because of the Covid-19 pandemic.

Literature and previous publications cited in this report

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.

Stephens M, N.J. Smith, P. Donnelly, 2001. A new statistical method for haplotype reconstruction from population data, Am. J. Hum. Genet. 68 (4) 978–989,

Publications during the current project period

Manuscripts in preparation

Pakstis AJ, WC Speed, M Murthy, KK Kidd. Global relationships of Central Asian populations based on 61 microhaplotypes.

Kidd KK, Andrew J. Pakstis, Michael P. Donnelly, O Bulbul, L Cherni, C Gurkan, L Kang, H Li, L Yun, P Paschou, KA Meiklejohn, E Haigh, WC Speed. The distinctive geographic patterns of common pigmentation variants at the OCA2 gene.

Manuscripts accepted in 2020

Gandotra N, WC Speed, Y Tang, AJ Pakstis, KK Kidd, C Scharfe, 2020. Validation of novel forensic DNA markers using multiplex microhaplotype sequencing. *Forensic Science International: Genetics.* In Press March 9, 2020. Online March 18 in pre-proof form. https://doi.org/10.1016/j.fsigen.2020.102275.

Rajeevan H, U Soundararajan, AJ Pakstis, KK Kidd, 2020. FrogAncestryCalc: A standalone batch likelihood computation tool for ancestry inference panels catalogued in FROG-kb. *Forensic Science International: Genetics* 46: In Press--assigned to May issue. https://doi.org/10.1016/j.fsigen.2020.102237.

Publications January 2019 to June 2020

Oldoni F, KK Kidd, D Podini, 2019. Microhaplotypes in forensic genetics, *Forensic Science International: Genetics* 38:54-69, doi: 10.1016/j.fsigen.2018.09.009.

Bennett L, F Oldoni, K Long, S Cisana, K Madella; S Wootton, J Chang, R Hasegawa, R Lagace, KK Kidd, D 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.

Boussetta S, L Cherni, AJ Pakstis, N Ben Salem, S Elkamel, KK Kidd, H Khodjet-el-Khil, A 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.

Drineas P, F Tsetsos, A Plantinga, I Lazaridis, E Yannaki, A Razou, K Kanaki, M Michalodimitrakis, F Perez-Jimenez, G De Silvestro, MC Renda, J Stamatoyannopoulos, K Kidd, B Browning, P Paschou, G Stamatoyannopoulos, 2019. Genetic history of the population of Crete. *Annals of Human Genetics* 83:373-388. doi: 10.1111/ahg.12328.

Bulbul O, KK Kidd; book chapter: Forensic Ancestry Inference. 2019. Eds. Elena Pilli and Andrea Berti; “Present and Future in Forensic Genetics”. Apple Academic Press., in Press.

Phillips C, D McNevin, KK Kidd, R Lagace, S Wootton, M de la Puente, A Freire-Aradas, A Mosquera-Miguel, M Eduardoff, TE. Gross, L. Dagostino, D Power, S Olsen, D. Hashiyada, C Oz, W Parson, PM. Schneider, MV Lareu, R Daniel, 2019. MAPlex - a massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations. *Forensic Science International: Genetics* 42:213-226.

Pakstis AJ, C Gurkan, M Dogan, HE Balkaya, S Dogan, P Neophytou, L Cherni, S Boussetta, H Khodjet-El-Khil, ABA ElGaaied, N Mjolsnes, K Janssen, G-H Olsen, S Hadi, EK Almohammed, V Pereira, D Mikkelsen Truelsen, O Bulbul, U Soundararajan, H Rajeevan, J R Kidd, KK Kidd, 2019a. Genetic relationships for European, Mediterranean and SW Asian Populations using a panel of 55 AISNPs**.** *European Journal of Human Genetics* 27:1885-1893. doi: 10.1038/s41431-019-0466-6.

Oldoni F, D Bader, C Fantinato, SC Wootton, R Lagacé, R Hasegawa, J Chang, K Kidd, D Podini, 2019. A massively parallel sequencing assay of microhaplotypes for mixture deconvolution. *Forensic Science International Genetics Supplement Series.* 7:522-524.

Pakstis AJ, WC. Speed, U Soundararajan, H Rajeevan, JR Kidd, H Li, KK Kidd, 2019b. Population relationships based on 170 ancestry SNPs from the combined Kidd and Seldin panels. *Scientific Reports* 9:18874; https://doi.org/10.1038/s41598-019-55175-x.

Description: RCR methodology (referenced in Table 1 caption on following page) 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."

Table 1. Our forensic-related journal publications sorted by RCR value as ranked by NIH *iCite* website (https://icite.od.nih.gov/analysis) on Jan 31, 2020. The list includes 46 of our 52 forensic publications. The website does not calculate RCR values until about 18 months after the paper has been published. Six of our forensic papers [Butler et al., 2008; Pakstis et al., 2008; Fang et al., 2009; Kidd et al., 2013 and 2015; Oldoni et al. 2019] that appear in the Supplement Series of Forensic Sci. Intl. 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
pending			Not rated	Not rated	2020	N Gandotra	Forensic Sci Int Genet
31991337			Not rated	Not rated	2020	H Rajeevan	Forensic Sci Int Genet
31827153			Not rated	Not rated	2019	AJ Pakstis	Scientific Reports
31377479			Not rated	Not rated	2019	C Phillips	Forensic Sci Int Genet
31285530	1	1.00	Not rated	Not rated	2019	AJ Pakstis	Eur J Hum Genet
31192450	1	1.00	Not rated	Not rated	2019	P Drineas	Ann Hum Genet
30790653			Not rated	Not rated	2019	S Boussetta	Gene
30758713	1	1.00	Not rated	Not rated	2019	L Bennett	Intl J Legal Med
30347322	3	3.00	Not rated	Not rated	2019	F Oldoni	Forensic Sci Int Genet
19456322	110	10.00	3.89	90.6	2009	Hui Li	Ann Hum Genet
24508742	62	10.33	3.83	90.3	2014	KK Kidd	Forensic Sci Int Genet
19937056	90	9.00	3.41	88.4	2010	AJ Pakstis	Hum Genet
16360294	94	6.71	2.71	83.6	2006	KK Kidd	Forensic Sci Int
21208434	70	7.78	2.52	81.8	2011	JR Kidd	Investig Genet
23815888	48	6.86	2.52	81.8	2013	CM Nievergelt	Investig Genet
27077960	17	4.25	2.33	79.6	2016	C-X Li	Forensic Sci Int Genet
25038325	29	4.83	2.28	79.1	2014	KK Kidd	Forensic Sci Int Genet
29625264	7	3.50	2.01	75.4	2018	O Bulbul	Forensic Sci Int Genet
28359046	10	3.33	1.98	74.9	2017	KK Kidd	Forensic Sci Int Genet
25750707	18	3.60	1.88	73.3	2015	KK Kidd	Investig Genet
28070634	13	4.33	1.79	71.6	2017	AJ Pakstis	Int J Legal Med
26355664	20	4.00	1.73	70.5	2015	AJ Pakstis	Forensic Sci Int Genet
26977931	17	4.25	1.76	71.1	2016	U Soundararajan	Forensic Sci Int Genet
22039151	41	5.13	1.70	69.9	2012	H Rajeevan	Nucleic Acids Res
17333283	49	3.77	1.59	67.7	2007	AJ Pakstis	Hum Genet
27316555	10	2.50	1.56	67.1	2016	KK Kidd	Hum Genomics
29931757	3	1.50	1.20	57.3	2018	KK Kidd	Electrophoresis
12519999	50	2.94	1.20	57.2	2003	H Rajeevan	Nucleic Acids Res
10592274	45	2.25	1.09	53.6	2000	KH Cheung	Nucleic Acids Res
22938150	20	2.50	1.05	52.0	2012	H Rajeevan	Investig Genet
27192181	9	2.25	1.03	51.4	2016	Lotfi Cherni	Am J Phys Anthropol
22445421	16	2.00	0.95	48.4	2012	KK Kidd	Forensic Sci Int Genet
27160361	6	1.50	0.86	44.5	2016	O Bulbul	Forensic Sci Int Genet
29175726	2	1.00	0.80	42.0	2018	KK Kidd	Forensic Sci Int Genet
12209575	37	2.06	0.80	42.0	2002	MV Osier	Am J Phys Anthropol
16028061	30	2.00	0.75	39.9	2005	J-J Kim	Hum Genet
22535184	15	1.88	0.71	37.6	2012	AJ Pakstis	Eur J Hum Genet
29248957	2	1.00	0.64	34.5	2018	O Bulbul	Int J Legal Med
24395150	11	1.83	0.58	31.6	2014	L Yun	Int J Legal Med
11125124	21	1.11	0.49	26.2	2001	MV Osier	Nucleic Acids Res
30205534	2	1.00	0.47	25.6	2018	S Gu	Gene
26829292	5	1.00	0.43	23.2	2015	JE Brissenden	Hum Biol
21913176	10	1.11	0.34	18.2	2011	JR Kidd	Am J Phys Anthropol
19325849	10	0.77	0.24	12.0	2007	H Rajeevan	Evol Bioinform Online
10902212	13	0.65	0.23	12.0	2000	KH Cheung	Pac Symp Biocomput
21668909	3	0.33	0.09	4.5	2011	JN Sampson	Ann Hum Genet