Final Summary Report
National Institute of Justice Grant #2013-DN-BX-K023
Covering January 1, 2014 to December 31, 2016
Project Title: "High resolution SNP panels for forensic Identification of ancestry, family, and phenotype"

Kenneth K. Kidd (PI), Professor Emeritus of Genetics
Email: Kenneth.Kidd@yale.edu Telephone: 203-785-2654
Department of Genetics, Yale University School of Medicine
Submitted by K.K. Kidd

Introduction

NIJ 2013-DN-BX-K023 was originally funded from January 1, 2014 through December 31, 2015. A GAN authorizing a 12-month unfunded extension of the grant until December 31, 2016 was approved in late October of 2015. Justification for the additional time was based on the existence of residual funds and additional work to be done on the project, as well as to ensure adequate funding until new award funds for 2015-DN-BX-K023 became available. The new grant, which commenced officially on January 1 of 2016, is essentially a continuation and extension of project 2013-DN-BX-K023.

During this project 19 publications (see lists at end of this report) have appeared and several more are in preparation. Most publications document different facets of our progress toward developing improved single nucleotide polymorphism (SNP) panels for ancestry and lineage inference. Much of the effort was focused on identifying multi-SNP haplotypes that we call microhaplotypes and then characterizing them in population samples from around the world. Through various research collaborations we have also extended the number of reference populations that have been genotyped for the panel of 55 ancestry informative SNPs (AISNPs) that we developed earlier (Kidd et al., 2014a) and made the frequencies and sample sizes available on the worldwide web via the allele frequency database ALFRED and the forensic resource knowledge-base FROGkb. Progress has also been made on developing second-tier AISNP panels for two areas of the world. Here we provide an overview of what has already been accomplished and indicate what results are currently in preparation for publication.

Major goals and specific aims

The overall purpose of this project as proposed and carried through after funding was awarded was to develop better sets of single nucleotide polymorphisms (SNP) markers for forensic applications. The two primary goals were (1) to enhance our developing ancestry informative (AISNP) panels to provide more robust differentiation at even finer geographic levels with an optimized panel of SNPs with large allele frequency differences among populations, (2) to identify and characterize additional multi-SNP haplotypes for inference of lineage (clan and extended family) relationships (LISNPS) focusing on molecularly smaller regions that still define multiple alleles (haplotypes) to take advantage of the newest genome sequencing technology to eliminate ambiguity in phase. These broad goals are interrelated. Many phenotype informative (PISNPs) markers for skin or eye color, hair type, etc. are excellent AISNPs. Thus, we proposed to characterize the population genetics of additional PISNPs as they are also candidate AISNPs. Multiallelic haplotypes also convey more information on identity and ancestry than the same SNPs considered as independent single SNPs.

Aims for AISNPs. First, we proposed to improve the current 55-AISNP panel as a global "first tier" panel by incorporating additional populations fully typed for all markers and adding additional markers that improve regional distinctions while removing the less discriminating markers to optimize the "smallest" panel that can reliably distinguish at least nine geographic regions on a minimum of 100 populations. Second, we aimed to develop "second tier" panels of markers that are highly discriminating within a geographic region but may be poor at a global level.

Aims for LISNPs. In our publications on minihaplotypes we demonstrated that multiallelic loci have greater statistical power to identify relatives than two-allele polymorphisms considered singly (such as SNPs). The decreasing cost and improved reliability of sequencing led us to move from emphasizing minihaplotypes (molecular regions extending up to 10 thousand basepairs) to focusing on microhaplotypes with molecular extents under 300 basepairs. At the time the proposal was submitted we had fully typed 10 microhaplotypes that we identified in our existing data on 45 populations. So the first specific aim for the LISNP panel aimed to collect the necessary data on at least 60 populations for those first 10 microhaplotypes AND on the more than 30 microhaps already identified from screening our existing data and published findings accumulating in public databases (such as HGDP, HapMap, 1000

Genomes) and other published results researchers made accessible on individual whole genome sequences for multiple individuals from different populations. The second specific aim for LISNPs was to identify the number of additional microhaps necessary for making forensic applications more powerful. Our objective for a final panel was for likelihood ratios considerably better than currently exist using the CODIS markers.

Aims for PISNPs. We indicated our efforts on PISNPs would be limited to those most relevant to ancestry inference. However, we would aim to accumulate results on different populations that provide a better global understanding of the variation for PISNPs previously identified in studies of only one region of the world. Such understanding is important since the same visible trait in one part of the world may have a different genetic cause than exists in other regions and conversely a variant strongly associated in one global region may be unimportant in another region because of background genotype or linkage disequilibrium differences with the true causal variant.

Accomplishments

The specific aims implementing the broader goals just described can be summarized as contributing either to identifying additional useful polymorphisms or else to expanding the number of population samples that have been genotyped so that we can better characterize the utility of the panels for different forensic purposes.

Identifying more useful SNPs

Most of the effort on studying new SNPs has been focused on finding new microhaplotypes. Because they are multiallelic, microhaplotypes can serve as LISNPs as originally defined by Kidd (cf. Butler et al., 2008). To date about 156 different potential microhaplotypes defined by 438 SNPs distributed across the 22 autosomal chromosomes have been identified from diverse sources. A total of 132 of these microhaplotypes (defined by 362 SNPs) have been well characterized thus far on 83 populations from around the world represented by over 5,000 individuals. The 83 populations include 57 groups that are routinely studied at Kidd lab using our “unlimited” DNA from cell lines (our unique resource) and another 26 groups for which comparative genotypes are extracted from the thousand genomes project browser (Auton et al. 2015). A manuscript reporting analyses of this dataset will soon be ready for submission to a journal. Publications since 2013 have been documenting our developing microhaplotype studies (see publication list). In 2015 we made public then-current results on 129 microhaplotypes

studied on 55 populations as a downloadable spreadsheet on our lab web site and publicized that in several talks and posters at meetings.

Because mixtures of biological material from more than one individual is such a major issue in modern forensics, we are emphasizing identification of microhaplotypes that can aid in identification and deconvolution of mixtures. The absence of stutter makes the microhaplotypes especially useful. Using the concept of effective number of alleles (A_e) (Kidd and Speed, 2014), the 28 microhaplotypes with global average A_e >3 give a probability of NOT detecting a mixture of DNA from two or more random individuals that is less than 10^{-7}. That is a major conclusion of the recent presentations and of the manuscript almost finished (awaiting approval from co-authors). A_e is also a statistic that can be used to estimate the ability of a microhaplotype locus to provide information on relationships beyond second degree. We are currently working to increase the number of loci with $A_e > 4$. While individually these microhaplotypes are not as informative as the forensic STRPs, larger numbers can be multiplexed and there is no degredation of information by the high mutation rates of the STRPs.

Though not the major emphasis on identifying microhaplotypes for ancestry, the microhaplotypes with higer A_e values can be selected to show large differences among populations. At the moment, our microhaplotypes characterized on 83 populations can distinguish 7 biogeographic regions. Though there is low overall correlation of the ancestry information and mixture information statistics, about 30% of the loci rank in the top half by both measures.

A proposed nomenclature for microhaplotypes was recently published (Kidd, 2016). A clear and simple way of labeling these multi-SNP loci is needed that allows microhaplotype loci identified by different labs to be distinguished. The proposal follows the HUGO Gene Nomenclature Committee guidelines with a unique prefix, “mh”, followed by chromosome number, lab symbol, and a number that with the chromosome number and lab symbol results in a unique combination.

In the course of studying our new population samples (Mongolians, North Africans, Saudi Arabians, etc.) for existing AISNP panels we are re-evaluating the existing markers and typing new candidate AISNPs on all the population samples. We demonstrated the large empty matrix problem in existing published panels of AISNPs (Soundararajan et al., 2016) and proposed a new 75-SNP panel that is more discriminatory among East Asian populations (Li et al.,2016). Analyses are ongoing to identify a more informative panel of fewer than 75 SNPs.

One other recent effort has involved identifying additional single AISNPs for developing second-tier panels for particular regions of the world. Our work in developing the Kidd 55 AISNP panel (Kidd et al., 2014a) can be considered an evolving first-tier panel that is helpful in identifying one of seven to ten or

more major world regions (depending on the set of populations analyzed) that an individual's ancestry may primarily derive from. A second-tier panel would consist of an additional set of SNPs that are especially good at differentiating among ethnic groups within a particular world region. Our collaborations with other researchers have resulted in preliminary studies exploring the creation of second-tier AISNP panels for Eastern Asia (Li et al., 2016) and for Southwest Asia-Mediterranean Europe (Bulbul et al., 2016). The work on Mongolian samples (Brissenden et al., 2015) and North African samples (Cherni et al., 2016) has demonstrated the value of the original 55 AISNPs and is providing data for better second teir panels.

Characterizing more populations for AISNPs In Soundararjan et al. (2016) we reviewed the AISNP literature to demonstrate the extent of the empty matrix problem that plagues the field. Among the 21 different published AISNP panels that were examined there was very little overlap of SNPs. Only 46 of the 1,397 different SNPs across the 21 panels are present in three or more of the panels, none in more than 6 of the 21 panels. In addition, most of the proposed AISNP panels have been typed on a relatively small number of ethnic groups and many the development datasets of these panels omitted important regions of the world. The inherent value of an AISNP panel for inferring ancestry depends on having comprehensive, comparable allele frequencies on the diverse range of ethnic groups and of potentially different regional populations within ethnic groups that span large geographical regions (cf, book chapter by Kidd, 2016). Thus, a set of reference population allele frequencies must exist for all the markers in an AISNP panel on a very large range of populations/ethnic groups from around the world in order to calculate the likelihood (match probabilities) and have a reasonable chance of successfully identifying an individual's ancestral origin. Pakstis et al. (2015) reported that at the time of publication 125 populations from around the world had reference allele frequencies on all of the 55 Kidd AISNPs. Among the published AISNP panels, the 55 Kidd AISNPs has the largest number of reference populations available. The frequency data for this set are accessible via the ALFRED database. The FROGkb knowledge-base has access to this data and facilitates calculation and inspection of the likelihoods of a match among the reference populations when online users submit the genotype profile for an individual online. Recently, additional reference populations have been added to ALFRED and are accessible for calculations via FROGkb. Currently there are 139 reference populations for the Kidd 55 AISNP panel. Analyses indicate there are ten geographical regions (inferred from STRUCTURE results at K=9) that the 55 AISNP panel differentiates. A short publication updating these additions is currently in press (Pakstis et al., 2017). Of course, this accumulation of 139 reference populations for the Kidd 55 AISNP panel is still only the beginning of the process of assembling a satisfactory set of reference populations that does an adequate job of covering human diversity at the global level. (The exact number needed is an empirical question

that is difficult to be exact about. The number of different human languages might be an approximate and rather imperfect proxy for the number of reference populations needed. According to the Ethnologue summary, there are over 7,000 living languages documented and over 1,300 of those languages have estimated population sizes of more than 100,000 people.) We are typing our populations for the better AISNPs of the published panels (such as the SNPforID and EUROFORGEN panels) to provide more reference data for more of the AISNPs published by others in the hope of reducing the empty matrix problem.

Training and professional development opportunities

This research project was not funded to provide training opportunities. One postdoc supported on fellowship funds participated in these studies and two visiting scientists supported by funds from their organizations also participated.

Impact of this project for policy and practice

Policy. The only highly differentiating sets of ancestry informative SNPs 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., 2008; Kidd et al., 2011), and by Nievergelt et al. (2013). To date, the Kidd 55 and the Seldin 128 AISNP panels (170 AISNPs in the combined set) are the only ones commercially available as a kit for investigators to use. 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 proposed work to improve biogeographic resolution and robustness are likely to be made commercially available as kits. Both Illumina and LifeTech have recognized our 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 Life Technologies is developing one based on the work from this project. Our extensively validated and documented data and our analyses of those data are being put in the public domain through the ALFRED and FROG-kb databases. Because of the extensive public documentation, forensic labs may 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 clarified by basic research.

Practice. The SNPs identified in this project should be useful for many investigative purposes. To the degree that SNPs identified from this study are brought before the courts, the work of our project provides a firm scientific basis for their acceptability. Population samples collected on the SNPs identified help to provide a strong statistical foundation for conclusions when used as investigative tools for inference of ancestry, phenotype, or family/clan membership. Microhaplotypes with 3 or more effective alleles are statistically powerful in deconvoluting samples that are mixtures from more than one individual. As forensic laboratories begin sequencing routinely, our results will be placed into practice because both Illumina and LifeTech kits have already incorporated our 55-AISNP and 45-IISNP panels. LifeTech is working to develop multiplex reactions for microhaplotypes for their kits based on the microhaplotypes we have identified. Both companies have expressed interest in adding more AISNPs when we have verified them for better resolution of ancestry.

Dissemination of results/products

Many of the results of this project have already been published. Lists at the end of this report show the ten papers published in peer-reviewed scientific journals, one book chapter in a handbook of forensic genetics, and a short, "extended abstract" paper based on a presentation at an international symposium on forensic genetics. Some additional manuscripts still in preparation are also noted at the end of this report.

As results have been published, SNP and haplotype allele frequencies and sample sizes for the populations studied have been made freely available on the world-wide web via ALFRED (Allele Frequency Database; http://alfred.med.yale.edu) and via the FROGkb (Forensic Resource on Genetics Knowledge Base; http://frog.med.yale.edu).

Various aspects of the work supported by this project dealing with the value of SNPs for forensic applications were also presented at numerous slide talks and poster sessions at international meetings and visits to University laboratories. These meetings of forensic researchers, anthropologists, and population geneticists were held at locations in the United States, Europe, and China. Examples of such meetings include but were not limited to: Bode Technology meetings, the International Symposium on Human Identification (ISHI), the International Society of Forensic Genetics (ISFG), Green Mountain DNA Conferences, Gordon Research Conferences, the American Academy of Forensic Sciences, and the International Conference on Genomics.

Literature Cited

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM et al. **2015**. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.

Butler JM, Budowle B, Gill P, Kidd KK, Phillips P, Schneider PM, Vallone PM, Morling N. **2008**. Report on ISFG Panel Discussion. *Forensic Science International: Genetics Supplement Series* 1: 471–472

Nievergelt, C. M., A. X. Maihofer, T. Shekhtman, O. Libiger, X. Wang, K. K. Kidd and J. R. Kidd, **2013** "Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel." *Investigative Genetics* 4(1):13.

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 andadmixture proportions in common populations in America. *Human Mutation* 30:69-78.

Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, de La Vega FM, Kidd KK. **2011**. Analyses of a set of 128 ancestry informative SNPs (AISNPs) in a global set of 119 population samples. *Investigative Genetics* 2:1

Publications 2013-2016 in peer-reviewed journals (* = peripheral to grant)

*Boyden SE, Desai A, Cruse G, Young ML, Bolan HC, Scott LM, Eisch AR, Long RD, Lee CCR, Satorius CL, Pakstis AJ, Olivera A, Mullikin JC, Chouery E, Megarbane A, Medlej-Hashim M, Kidd KK, Kastner DL, Metcalfe DD, Komarow HD, **2016**. Vibratory urticaria associated with a missense variant in ADGRE2. *New England Journal of Medicine* 374(7):656-663

Brissenden, J., J.R. Kidd, B. Evsanaa, A.Togtokh, A.J. Pakstis, F. Friedlaender, K.K. Kidd, J.Roscoe, **2015**. Mongolians in the genetic landscape of Central Asia: Exploring the genetic relations among Mongolians and other world populations. *Human Biology* 87(2):5-23.

Bulbul, O., L. Cherni, H. El-Khil-Khodjet, H. Rajeevana, K.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. DOI:10.1016/j.fsigen.2016.04.010.

Cherni, L., A.J. Pakstis, S. Boussetta, S. Elkamel, S. Frigi, H. Khodjet-El-Khil, A. Barton, E. Haigh, W.C. Speed, A. Ben Ammar Elgaaied, J.R. Kidd, K.K. Kidd, **2016**. Genetic variation in Tunisia in the

context of human diversity worldwide. *American Journal of Physical Anthropology* 161:62-71. DOI:10.1002/ajpa.23008. In press April 22, 2016; Online May 18, 2016.

*Keramati AR, Fathzadeh M, Go G-W, Singh R, Choi M, Faramarzia S, Mane S, Kasaei M, Sarajzadeh-Fard K, Hwa J, Kidd KK, Bigi MAB, Malekzadeh R, Hosseinian A, Babaie M, Lifton RP, Mani A, **2014**. A form of the metabolic syndrome associated with mutations in DYRK1B. *New England Journal of Medicine* 370(20):1909-1919.

Kidd, K.K., and W.C. Speed, **2014**. Criteria for selecting microhaplotypes: mixtures and deconvolution. *Investigative Genetics* 6:1.

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

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

Kidd, K.K., A.J. Pakstis, W.C. Speed, R. Lagace, J. Chang, S. Wootton, E. Haigh, J.R. Kidd, **2014b**. Current sequencing technology makes micro-haplotypes a powerful new type of genetic marker for forensics. *Forensic Science International: Genetics* 12:215-224. DOI:10.1016/j.fsigen.2014.06.014.

Kidd KK, Pakstis AJ, Yun L, **2014**. An historical perspective on "The world-wide distribution of allele frequencies at the human dopamine D4 receptor locus" *Human Genetics* 133:431-433.

*Kim M, Chen X, Chin LJ, Paranjape T, Speed WC, Kidd KK, Zhao H, Weidhaas JB, Slack FJ, **2014**. Extensive sequence variation in the 3' untranslated region of the KRAS gene in lung and ovarian cancer cases. *Cell Cycle* 13:1030-1040.

Li, C.-X., A.J. Pakstis, L. Jiang, Y.-L. Wei, Q.-F. Sun, H. Wu, O. Bulbul, P.Wang, L.-L. Kang, J.R. Kidd, K.K. Kidd, **2016**. A panel of 74 AISNPs: Improved Ancestry Inference within Eastern Asia. *Forensic Science International: Genetics* 23:101-110. DOI:10.1016/j.fsigen.2016.04.002.

Pakstis, A.J., E Haigh, L Cherni, A Ben Ammar ElGaaied, A Barton, B Evsanaa, A Togtokh, J Brissenden, J Roscoe, O Bulbul, G Filoglu, C Gurkan, KA.Meiklejohn, JM Robertson, C-X Li, Y-L Wei, H Li, U Soundararajan, H Rajeevan, JR Kidd, KK Kidd, **2015**. 52 additional reference population samples for the 55 AISNP panel. *Forensic Science International: Genetics* 19:269-271.

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. *Int J Legal Med in press*

Paschou, P., P. Drineas, E. Yannaki, A. Razou, K. Kanaki, F. Tsetsos, S.S. Padmanabhuni, M. Michalodimitrakis, M.C, Renda, S. Pavlovic, A. Anagnostopoulos, J.A. Stamatoyannopoulos, K.K. Kidd, G. Stamatoyannopoulos, **2014**. Maritime route of colonization of Europe. *Proc. Natl. Acad. Sci. USA* 111(25):9211-9216. doi:10.1073/pnas.1320811111.

Soundararajan, U., 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. DOI: 10.1016/j.fsigen.2016.01.013.

Xu, H., C.C. Wang, R. Shrestha, L.X. Wang, M. Zhang, Y He, J.R. Kidd, K.K. Kidd, L. Jin, H. Li, **2014**. Inferring population structure and demographic history using Y-STR data from worldwide populations. *Molecular Genetics and Genomics* 290:141-150.

Yun, L., Y. Gu, H. Rajeevan and K. K. Kidd (2014) "Application of six IrisPlex SNPs and comparison of two eye color prediction systems in diverse Eurasia populations." *Int J Legal Med* 128(3): 447-453.

Other Publications

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.

Kidd, K.K., **2016**. Chapter 7: "Thoughts on estimating ancestry" In: A. Amorim and B. Budowle (Eds), Handbook of Forensic Genetics--Biodiversity and heredity in civil and criminal investigation. London: Imperial College Press.

Manuscripts in preparation

K.K. Kidd, W.C. Speed, A.J. Pakstis, other co-authors including those from collaboration with ThermoFisher Scientific, order to be determined depending on participation in analyses and writing; 2016. Working title: “Evaluating 132 Microhaplotypes across a Global Set of 83 Populations”.

U. Soundararajan, H. Rajeevan, K.K. Kidd, 2016; New enhancements for FROG-kb.

A.J. Pakstis, L. Cherni, O. Bulbul. L. Kang, J.R. Kidd, W.C. Speed, K.K. Kidd. 2016. A global overview of functional variation at the OCA2 locus. The OCA2-HERC2 gene region has functional SNP variants relevant to skin/eye/hair pigmentation.

A.J. Pakstis, E. Haigh, J.R. Kidd, K. K. Kidd, 2016. Validation of minihaplotypes as valuable markers for familial identification and ancestry inference.