NIJ Final Technical Report

December 1, 2010 to April 30, 2014

Further Development of SNP Panels for Forensics

NIJ Grant# 2010-DN-BX-K225

Submitted by Kenneth K. Kidd (PI)

Professor of Genetics, Yale University School of Medicine

Email: Kenneth.Kidd@yale.edu

Telephone: 203-785-2654

NOTE: Portions of this report are taken from 10 published papers, 3 submitted manuscripts, a few manuscripts in preparation, along with various poster presentations and slide talks--all supported by this grant--that are listed in the sections of this report describing “Dissemination of Results” and “Publications & Poster Presentations”. Funding for this project ended on November 30, 2013; no cost extensions allowed time for the external reviews to be completed and final revisions of this report.

ABSTRACT

Single nucleotide polymorphisms (SNPs) are a largely untapped DNA resource for forensic applications. Because of the extensive databases and experience that have been established with the CODIS markers, SNPs are unlikely to replace STRPs for applications involving individual identification. However, SNPs offer many advantages over STRPs for inferring the ancestry of a DNA sample, for identifying close biological relatives, and eventually for inferring phenotypes like eye color. In the current project we built on our previous work and have provided improved SNP panels for forensic applications by (1) enhancing our developing ancestry informative (AISNP) panel, (2) identifying a panel of mini-haplotypes and a preliminary set of micro-haplotypes for inference of lineage (clan and extended family) relationships (LISNPS), and (3) providing population genetic background evidence for many Phenotype Informative markers (PISNPS).

We identified potential candidate SNPs useful for ancestry inference and lineage identification both by screening our accumulated datasets and large web-accessible SNP datasets (e.g. 1000 Genomes, CEPH Human Genome Diversity Panel) as well as the published literature. We employed a number of different statistical methods iteratively to extract the "best" subset of ancestry inference SNPs (AISNPs)--heatmaps, pairwise Fst, PCA analyses, and STRUCTURE--based on the SNPs that differentiated most between the populations. The best current set of 55 AISNPs identified during this project was made public on the FROG knowledge base (http://frog.med.yale.edu) in January, 2013; it is also described at length in a manuscript that has been submitted for publication.

LISNPs or Lineage Informative SNPs are haplotyped loci with two or more SNPs defining multiple alleles (haplotypes). Haplotypes can provide the multiple low frequency alleles that are optimal for familial and lineage assignment of a DNA sample. We began this project identifying small haplotypes that are less than 10,000 basepairs in extent and that have multiple haplotype alleles at common frequencies. Our published work on the concept of mini-haplotypes (Pakstis et al., 2012) provides proof of principle. We also developed a panel of 25 essentially unlinked mini-haplotypes for which a manuscript is in preparation.

New sequencing methodology that gives high throughput sequencing of individual DNA molecules of 200 plus basepairs led us to shift our priorities during the last year of the project. We emphasized finding micro-haplotypes with two or more SNPs within a segment of no more than 200 basepairs which define at least 3 haplotypes and also display high heterozygosity in most populations. By the end of our project we had identified a set of 28 microhaps, each defined by 2 to 3 SNPs, which we have studied on 54 of our populations. Because of the high information content of these markers, the random match probabilities based on them range from 2.93×10^{-13} to 1.03×10^{-18} in the populations studied, values comparable to the IISNP panel we published. Microhaps typed by sequencing have another desirable capability: identification of mixtures with the potential to quantify the components, i.e., to disentangle mixtures in a quantitative way. The presence of three or more different sequences becomes clear evidence of DNA from more than one person contributing to the sample. Thus, microhaps can become powerful markers to identify and quantify components of mixtures.

Carefully validated IISNP, AISNP, LISNP, and eventually PISNP panels offer great utility for forensic applications. Our work has contributed to all of these areas and documents the value of SNPs in forensics. This project underscores that continued work on SNP panels offers much more than a theoretical possibility of better SNPs for better forensics. By making the data public through the ALFRED and FROG databases, we also hope others will be able to build on what we have done.

Table of Contents

Executive Summary

Background, Rationale, Goals

The need for panels of SNPs (single nucleotide polymorphisms) for forensic applications is documented by a large number of publications in the literature. When DNA is badly damaged, when speed is essential, and when ease of typing and interpretation are critical, SNPs can perform better than STRPs, though SNPs are unlikely to replace STRPs entirely for individual identification. However, for inference of ancestry of a DNA sample, for inference of phenotype, and for identification of close relatives SNPs can provide much more accuracy than the CODIS STR markers. NIJ has supported our SNP research with three grants (2004-DN-BX-K025, 2007-DN-BX-K197, 2010-DN-BX-K225). The first two projects focused on developing two optimized panels of SNPs. The first panel was a Low Fst / High heterozygosity panel of SNPs for Individual Identification (IISNPs) that should be acceptable in the courts as reliable and highly probative. The second panel was a High Fst panel of SNPs that are Ancestry Informative (AISNPs) and provide a robust investigative tool. In the third grant, the subject of the current technical report, we built on our previous work and have provided improved SNP panels for forensic applications by (1) enhancing our developing AISNP panel, (2) identifying a panel of mini-haplotypes and a preliminary set of micro-haplotypes for inference of lineage (clan and extended family) relationships (LISNPS), and (3) providing population genetic background for many Phenotype Informative markers (PISNPS). In all of these studies, the goal has been to identify those SNPs likely to be good for the specific objective and to evaluate them on a large number of population samples. Our rationale for studying a global sample of populations is the highly diverse ancestries represented in the U.S. population.

Strategy and methods

At the beginning of this funding period we had an accumulated dataset of ~3000 SNPs with data on most of 54 populations in our population resource. During this project we have added several hundred new SNPs to the dataset and extended the nearly complete dataset to ~4000 markers on 58 populations. In addition we have incorporated additional populations into the dataset for the more informative of the SNPs for ancestry and phenotype. In those extensions and in the filling-in of markers not previously typed on some populations we have not included the panel of IISNPs developed on previous funding (Pakstis et al. 2010) although we did conduct additional analyses and write a paper on those SNPs (Kidd et al., 2012).

Our strategy for this funding period evolved as the science evolved. We initially worked (1) to identify the best markers for ancestry inference and (2) to characterize the population genetics of some of the SNPs that were implicated, usually only by association studies, with phenotypes. This involved analyzing existing data in successive iterations as the data increased, screening large public datasets for candidate SNPs, and incorporating into our data the SNPs being published in the literature for ancestry and phenotype. Our laboratory methods were the same for all of the different analyses: type by TaqMan the SNPs of interest on all of the relevant population samples. The minimum set of relevant populations were those for which we had unlimited DNA because over the previous two and a half decades we had established transformed cell lines (Table 1). Our lab procedures are organized in such a way that we are extremely efficient at typing small numbers of specific SNPs (that is fewer than 250) on very large numbers of samples (that is, more than 2,500). With this arrangement, it is more efficient and cost effective to type samples in-house rather than design a custom Illumina (for example) array. DNA samples are at standardized concentrations and all pipetting is robotic and performed in a very stereotypic format so that every different 384-well plate, labeled WPI to WPVIII (World Plate I, etc.), always contains DNA of the same individual in the same position matching our standard sample sheets and corresponding electronic files. Allele calling is automatic with Applied Biosystem's SDS software followed by visual inspection of the results. Data management software is designed to facilitate high throughput and includes many error-checking components. Functionality of our lab's database remains largely as we originally published (Cheung et al., 1996) to manage population genetic data and the system has been converted to a Web-accessible system with multiple security features to prevent unauthorized access. The database is firewall and password protected and backed up daily. In any case, no personally identifying information is included in the database. Indeed, all of our forensic studies are done on samples that are completely anonymous.

The analytic methods varied somewhat depending on the specific objective but always included testing for Hardy-Weinberg ratios in the data. Similarly, F_{st} is a useful approach for quantifying the relative amounts of variation shown by different SNPs across different populations and is automatically calculated for each SNP as simply basic information on the SNP. It is implemented in several programs we have written over the years. Specifics of other methods are described as part of the subprojects.

Ancestry Inference SNPs (AISNPs)

To ascertain the ethnicity of individuals from SNPs in forensic investigations a panel of ancestry informative markers (AIMs) is required; each of those needs to demonstrate

allele frequency differences among populations and have also been studied on a very broad set of the world's populations. AIMs cannot be expected to distinguish among populations automatically if the populations have not been studied previously. A panel of AISNPs is not necessarily useful for estimating an individual's ancestry if the ancestral population(s) relevant for that individual have not been included among the ethnic groups already studied for those AISNPs.

Most useful for forensic work will be a small but efficient and robust set of markers that can provide enough information so that ancestry can be assigned at a high enough probability to be helpful. In the forensic context a small number of SNPs can mean lower costs and possibly faster turnaround. Our panel of populations (Table 1, up to a total of 73) provides the basis for documenting the value of AISNPs for forensics. The problem is to identify the SNPs to test on our panel. Our screening of other datasets has evolved over the years as more SNPs have been typed on more populations. Recently we have used the HGDP-CEPH panel of over 1000 individuals from 51 populations as another source that others have also used (Kosoy et al., 2009; Nievergelt et al., 2013); we typed those published panels on our samples. We augmented the HGDP with the same SNPs tested on 1300 additional individuals not present in the HGDP; these included additional individuals from the populations we contributed to the HGDP as well as additional populations. Most recently we contributed a small but global sample to the Visigen/Identitas study (Keating et al., 2012) and have screened those results as well as the 1000 Genomes data.

The global coverage of our large number of candidate AISNPs consisted of 63 populations with a total of 3063 individuals (see Table 2). . We employed a number of methods to extract from this dataset the "best" subset of AISNPs balancing the ability to discriminate populations with the desire to have a relatively small number of SNPs. It is also important to balance the information from different SNPs to assure that different regions of the world are distinguishable in a robust way. The use of heatmaps for the candidates helped by graphically portraying redundancy in the SNP information. Pairwise Fst calculations for each SNP across populations from different regions helped identify those SNPs best at certain distinctions, such as Europe vs. East Asia, so the "best" SNPs were used in the balancing (Kersbergen et al., 2009). We also employed STRUCTURE (Pritchard et al., 2000) as one first-pass method of identifying the SNPs that differentiated most between the clusters identified. Once we had identified our set of 55 AISNPs on our 63 populations, we extracted the data for 10 populations from the 1000 Genomes data. The resulting data set included 73 populations and 3,884 individuals. Table 3 lists the final set of 55 AISNPs that was made public on the FROG-kb web site (http://frog.med.yale.edu) in January, 2013, and is described at length in a submitted manuscript (Kidd et al., 2013; see list of submitted papers at end of report). Figures 1 through 5 depict the results from several different methods applied to this set

of AISNPs. In the final months of the project, some additional candidate AISNPs have been typed and many of these 55 SNPs have been typed on other populations (mostly by other laboratories). When those data are assembled, it is likely that an improved set of AISNPs can be identified. We note specifically that we have typed these 55 on the two Mongolian population samples we have DNA on.

To move forward with ancestry inference the key problem is finding SNPs that provide a clearer differentiation of certain populations or groups of populations without detracting from differentiation among some other populations. As noted in Kersbergen et al. (2009) some SNPs can simply add noise. We note that several of the SNPs that help differentiate European individuals from the rest of the world are not fixed for the Europe-specific allele. With genotype differences among individuals we surmise that STRUCTURE analyses tend to use this Mendelian segregation to classify individuals in all European populations "randomly" into two or three different clusters. The SNP with the lowest Fst in these 73 populations, rs4411548 (Figure 5), has significant frequency variation between East Asians and Native Americans, but the frequencies in Europe and Southwest and South Asia are all intermediate in no strong pattern. Thus, that SNP is just adding noise to the differentiation of those populations.

Because it will be difficult to avoid this problem, we have begun identifying markers that will be especially useful in discriminating among populations within a region that is not cleanly subdivided by the 55 AISNPs. These will constitute a second-tier of markers that will be used for analyses only within a specific region once that region is identified by the first tier markers. Our data are not yet sufficient, but such second-tier markers will be a focus of future studies.

Lineage Informative SNPs (LISNPs)

We originally defined LISNPs as haplotyped loci with two or more SNPs defining multiple alleles (haplotypes) to be analogous to the STRPs used in forensics (Pakstis et al., 2007; Butler et al., 2008). To be useful in forensics, the SNPs must comprise multiple alleles (haplotypes) with sufficiently small rates of intra-locus recombination and mutation so that identity by state allows an assumption of identity by descent within a family. The haplotypes must also have known population frequencies and ideally at least moderate heterozygosity in most populations of relevance in forensics. Based on our research over many years on multi-SNP haplotypes and their global frequency patterns, we know that haplotypes in small molecular regions can provide the multiple low frequency alleles that are optimal for familial and lineage assignment of a DNA sample. Starting with the funding of this project, we began working to identify small haplotypes (<10kb in extent) that had multiple alleles. Our published work on the

concept of mini-haplotypes (Pakstis et al., 2012) provides proof of principle. We currently have a set of 25 minihaplotypes with a manuscript in preparation. Most have been entered into ALFRED; the last few are being entered into the database as the limited resources available makes this possible. The keyword *minhap* can be used to retrieve a list of the systems entered.

Tables 4 through 7 list the minihaplotypes we have identified to date and the results of several of the analyses involved in identifying these loci are presented in Figures 6 through 10 as applied to the 25 "best" of the minihaplotypes. All of these data have been obtained by TaqMan typing of the individual SNPs and phasing statistically. These analyses show that these 25 loci are very good for individual identification and have some useful ancestry information. With the high heterozygosity and multiple alleles, they are also clearly providing information on familial relationships.

While the statistical estimation of the haplotype phases in an individual is highly accurate even for individuals multiply heterozygous, new sequencing methodology that gives high throughput sequencing of individual DNA molecules of 200+ base pairs has caused us to shift our priorities in the past year. The emphasis is now on haplotype loci with two or more SNPs within a segment of no more than 200bp with at least 3 alleles (haplotypes) and high heterozygosity in most populations. Theoretically these micro-haplotype loci will have all of the characteristics of the mini-haplotypes and an even lower probability of de novo occurrence of a recombination event plus the ability to generate the data without statistically determining the phase of the individual SNP alleles. The objectives we have pursued are to identify such loci and type them on our populations to first determine their relative worth for forensics and, for the better ones, to provide the database needed to interpret the data when they are used in a forensic context.

Using database and public data screening procedures very analogous to the searches for AISNPs and mini-haplotype loci, we have also identified a set of 28 microhaps (Table 8), each defined by 2 to 3 SNPs, and have studied them on 54 of our population samples. Figure 11 presents a detailed view of one good microhaplotype, a 3-SNP microhap spanning 124 bp at the EDAR gene on chromosome 1. For the 28 microhaps on our 54 populations (2612 individuals) the global average heterozygosity is 0.543, with 21 of the loci above the single-SNP maximum of 0.5 (Figure 12). Because of this high heterozygosity (a.k.a. high information content), the random match probabilities range from 2.93×10^{-13} to 1.03×10^{-18} (Figure 13). These results are very comparable to our published panel of 45 unlinked individual identification SNPs (Kidd et al., 2012). These loci are also good for ancestry inference as shown in Figures 14 and 15. Additional documentation of the potential of these markers is shown by STRUCTURE analysis of the 48 microhaps that we have identified in the HGDP data

using criteria of high global heterozygosity and low frequency correlation across the globe (Figure 16).

The results of our studies strongly support the potential utility of microhaps as a general forensic tool for both individual identification and ancestry inference as well as a way to help in identifying family or clan relationships. Microhaps typed by sequencing have an additional capability: identification of mixtures with the potential to quantify the components, i.e., to disentangle mixtures in a quantitative way. The presence of three or more different sequences, each with a sufficient numbers of reads, becomes clear evidence of DNA from more than one person contributing to the sample. The relative number of reads for the different sequences becomes a basis for quantification of the input genotypes. With many loci multiplexed and with more loci with four or more haplotypes, the microhaps become powerful markers to identify and quantify components of mixtures. Microhaps with 3 or more common alleles (haplotypes) for one DNA segment of <200bp will allow computer software to accurately predict the likelihood and levels of mixture based on observing more than two sequence types at a locus and the numbers of occurrences of each type.

Phenotype Informative SNPs (PISNPs)

While we do not have phenotype information (or any information other than population origin) on our samples, our collection of population samples (cell lines and DNA-only) listed in Table 1 is valuable for gaining a better understanding of how SNPs that are reported to be associated with phenotypes that are observable, such as eye and skin color, actually vary among human populations. To that extent we have studied several of the primary phenotype-associated SNPs. Our main publication to date is on OCA2-region SNPs and it showed that most of the associated SNPs were identifying the same association because the alleles associated with light eye color were all in nearly complete linkage disequilibrium (LD) in Europe. One haplotype extends across a long region in the 5' half of the OCA2 gene up to the enhancer SNP subsequently shown by Visser et al. (2012) to be the primary "cause" of blue eyes. We also studied four of the OCA2 SNPs that have clear functional possibilities and that have an appreciable heterozygosity. Figure 17 shows the haplotype frequencies based upon those four SNPs. The haplotype defined by the enhancer variant (green) is clearly present only in populations in or near Europe with the possibility of European admixture in some other population samples. The haplotype defined by the functional SNP rs1800414 (His615Arg) (light blue in the figure) is restricted to East Asia with some evidence of gene flow into Southeastern Europe and is associated with lighter skin pigmentation among East Asians (Edwards et al.,2010).

We also carried out a study of population genetic variation for two skin color loci: SLC24A5 (chromosome 15q21.1) and SLC45A2 (chromosome 5p13.3) focused on two SNPs, rs1426654 (Ala111Thr) in SLC24A5 and rs16891982 (Leu374Phe) in SLC45A2. Figure 18 gives an overview of the global distributions of the “light skin” alleles in a minimum of 110 populations incorporating published data along with our own data.

Conclusion

Carefully crafted IISNP, AISNP, LISNP, and eventually PISNP panels offer great untapped advantages for forensic applications. Our work to date has contributed to all of these areas and documents the value of SNPs in forensics. This project underscores that continued work on SNP panels offers much more than a theoretical possibility of even better SNPs for better forensics. By making the data public through the ALFRED and FROG-kb databases and respective web-accessible interfaces, we make it possible for others to be able to build on what we have accomplished to date.

Introduction

Literature Background and Rationale

The need for panels of SNPs for forensic applications is documented by a large number of publications in the literature (e.g., Vallone et al., 2005; Giardina et al., 2007a 2007b; Sanchez et al., 2006; Dixon et al., 2005; Phillips et al., 2007a, 2007b; Kosoy et al., 2009; Pomeroy et al., 2009; Børsting et al., 2008, 2009; Pereira et al., 2008; Dario et al., 2009; Kayser and Schneider, 2009; Ge et al., 2010; Kim et al., 2010; and many more). When DNA is badly damaged, when speed is essential, and when ease of typing and interpretation are critical, SNPs can perform better than STRPs, though SNPs are unlikely to replace STRPs entirely for individual identification. However, for inference of ancestry of a DNA sample, for inference of phenotype, and for identification of close relatives SNPs can provide much more accuracy than the CODIS STR markers. The CODIS STRPs were chosen because they are highly polymorphic, consequently decreasing their differences around the world. The limited number of CODIS markers and the concern over STRP mutation rates weaken their utility within families. While increasing the number of STR markers can be difficult--witness the current efforts to increase the “standard” set of STRPs--large numbers of SNPs can readily be multiplexed and tested. NIJ has supported our SNP research with three grants (2004-DN-BX-K025, 2007-DN-BX-K197, 2010-DN-BX-K225). The first two projects focused on developing two optimized panels of SNPs. The first panel was a Low Fst / High heterozygosity panel of SNPs for Individual Identification (IISNPs) that should be acceptable in the courts as reliable and highly probative. The second panel was a High Fst panel of SNPs that are Ancestry Informative (AISNPs) and provide a robust investigative tool. In the third grant, the subject of the current technical report, we built on our previous work and provide improved SNP panels for forensic applications by (1) enhancing our developing AISNP panel, (2) identifying a panel of mini-haplotypes and a preliminary set of micro-haplotypes for inference of lineage (clan and extended family) relationships (LISNPS) , and (3) investigating what could be achieved at present with Phenotype Informative markers (PISNPS). We worked on all of these goals together because we have come to recognize in the course of our work that these are

not independent objectives, e.g., many excellent AISNPs are also potential PISNPs for skin color, eye color, hair type, etc. Also, mini-haplotypes (minihaps: very close SNPs under 10KB that define multiallelic haplotype loci) and micro-haplotypes (microhaps: two or more SNPs less than 200 basepairs apart that define multiallelic haplotype loci) convey more information on identity as well as ancestry than do single SNPs.

When SNPs associated with skin pigmentation, eye color, and hair texture are used as AISNPs, it also must be clear about the extent to which an AISNP panel will likely yield only highly problematic inference of gross physical appearance (phenotype) (Kayser & Schneider, 2009) even though optimized for reliable inference of ancestry (Cho and Sankar, 2004; Shriver et al., 2005; Cho and Sankar, 2005). For strong inference of phenotype per se, such as skin pigmentation, it will be necessary to have genotypes simultaneously on all of the relevant loci because of the likely existence of epistatic interactions in the normal range of variation as are already demonstrated with abnormal phenotypes, e.g., albinism.

With limited exceptions AISNP studies in the past have focused on tools for admixture quantification in disease gene searches. Very frequently those studies have focused on African Americans and “Hispanics”, essentially attempting to quantify “European” admixture with Africans and Native Americans, respectively. Studies specifically in forensics have been limited (Vallone et al., 2005; Phillips et al., 2007; Lao et al., 2006; Giardina et al., 2007a, b). Only Phillips et al. (2007) produced a set of AISNPs with extensive population support to qualify as a robust investigative tool. However, non-forensic studies to identify admixture markers can be a valuable resource for aiding our goal of identifying additional good candidate AISNPs. We identified many published studies with a biomedical focus that present panels of AISNPs reportedly suitable for determining admixture levels in specific admixed populations (e.g., Jorde & Wooding, 2004; Collins-Schramm et al., 2004; Shriver et al., 2005; Yang et al., 2005; Enoch et al., 2006; Lao et al., 2006; Tian et al., 2006, 2007, 2008; Bauchet et al., 2007; Halder et al., 2008; Hodgkinson et al., 2008, Jakobsson et al., 2008; Kosoy et al., 2009; Lao et al., 2008; Pemberton et al., 2008; Price et al., 2008). Other studies have found considerable ancestry information in the HGDP-CEPH 650K SNP dataset (Li et al., 2008; Paschou et al., 2007; Biswas et al., 2009). In examining several of these studies

of AISNPs we found little overlap between sets of SNPs published by different groups. Thus, we began working even before the current project was funded on those that appear to show the strongest distinctions among populations based on the published data, with an attempt to emphasize distinctions between populations not well distinguished by most AISNPs.

Lao et al. (2008) used large numbers of SNPs on DNA samples of 2,475 individuals from 23 European population samples and showed a significant correlation of the clustering by the first two principal components with European geography. (Unfortunately, the data are not publically available (Manfred Kayser, personal communication). Several other studies have shown differences among various European groups, usually on a North-South or East-West axis, but never with large numbers of populations (e.g., Li et al., 2008; Novembre et al., 2008; McEvoy et al., 2009a; Tian et al., 2009) and requiring large numbers of SNPs.

With the exception of Y-chromosome and mtDNA (haplotypes by definition), there are few publications using haplotype data for forensics or, more generally, ancestry identification. Jakobsson et al (2008) and Conrad et al (2006) have both used the data collected on the HGDP panel to examine genome and population structure with autosomal haplotypes. Odriozola et al. (2009) have combined the relatively highly conserved SNP, rs59186128, and the multi-allelic CODIS STR, D7S820, into a haplotype that subdivides the CODIS D7S820 alleles that are identical simply by number of repeat elements and does so differently in different populations. More recently Ge et al (2010) have proposed haplotypes based on haplotype blocks in the human genome as a tool for increasing the power of discrimination of SNPs in forensic applications. Schlebusch and Soodyall (2012) successfully identified a panel of 44 short 5-SNP haplotypes to study population structure and relationships among the San, Khoe, and groups of mixed ancestry in southern Africa. We have published more than a dozen papers showing global haplotype variation at multiple loci, but never in a forensic context. We know that haplotypes defined by a small number of SNPs covering short molecular distances of 10KB or less can be found that are highly informative markers. The very low mutation rate of the SNPs (compared to STRPs) and the low levels of recombination expected for haplotypes covering short molecular distances combine to

offer additional advantages for forensic applications that have been largely untapped up until now. In our original proposal for this project we emphasized the potential of mini-haplotypes and initially pursued the development of such a panel (Pakstis et al., 2012). During this project it became increasingly apparent that improvements in sequencing technology were making it practical to consider focusing on the development of micro-sized haplotypes that are less than 200 base pairs in extent which could reap even more advantages for forensic applications than a panel of mini-haplotypes.

Differentiation, on average, among even closely related groups or individuals within groups such as European populations is clearly possible if a large enough marker set is employed—that is not the problem. The key problem for routine forensic applications is identifying ancestry for a single individual with a reasonable number of SNPs. It is likely that the utilization of SNP haplotypes or possibly SNPs combined with STRPs (SNPSTR; Mountain et al., 2002; Ramakrishnan & Mountain, 2004) may help achieve the goal of maximizing accuracy of our prediction of ancestry.

As more genes influencing human phenotypes have become known, a growing literature has developed on their forensic uses (Branicki et al., 2008, 2007; Kayser & Schneider, 2009; Masui et al, 2009; Pulker et al., 2007; Soejima & Koda, 2007; Tully, 2007; Fujimoto et al., 2008a, 2008b; Mou et al., 2008; Valenzuela et al., 2010; Visser et al. 2012; Ruiz et al., 2013). MC1R, ASIP, EDAR, MATP (SLC45A2), SLC24A5, TYR, TYRP1, and OCA2 are among the genes with alleles associated with variation in skin color, hair texture, and eye color. The scientific literature provides extensive evidence that these particular genes and others that have been identified are not the whole story for these phenotypic traits of interest in forensics. For example, in a search for genes influencing the highly heritable trait height, McEvoy & Visscher (2009b) and Lettre (2009) report that about 50 genes, each with small effect, have been associated with height. Some of these genes may only be significant in disorders of height, but there may still be many, many genes responsible for the normal range of stature. The same can probably be said of nearly all of the traits of most interest in forensics. Until we have an understanding of the underlying biology of these traits, identifying their genetic markers will remain problematic. Absent that clear understanding, we can still begin to use what is known and develop markers or marker combinations that are informative for

ancestry and lineage identification, and perhaps, with many caveats, also for phenotypic traits.

Goals

The SNPs screened on this project fall into three categories: (1) potentially informative AISNPs (ancestry inference SNPs), (2) potentially helpful SNPs for forming good minihaps and microhaps useful as LISNPs, and (3) potentially useful PISNPs (Phenotype Informative SNPs) or SNPs which are in strong linkage disequilibrium with validated PISNPs. These are not mutually exclusive categories since several PISNPs (e.g., skin color SNPs) have characteristics that make them excellent AISNPs. Moreover, either AISNPs or PISNPs can serve as the basis for a minihap or a microhap. A minihap comprised of SNPs that are not individually good AISNPs can sometimes provide excellent discrimination among populations.

New collaborations established after the start of this project gave us access to DNA samples of Mongolian and Tsaatan populations through a collaboration with Canadian and Mongolian researchers. Through a collaboration with Dr. Lotfi Cherni we also obtained DNA samples for a number of populations from within Tunisia as well as a sample of Libyans which we have only begun to type systematically near the end of the current project's funding period. The choice of the SNPs typed on these new "DNA-only" populations has been motivated by our forensic projects on AISNPs, minihaps or microhaps, and PISNPs. For markers that are especially promising, we have typed some of our other additional samples (Table 1b) for which we have DNA only (e.g., Greeks, Italians, Catalans, Mohanna, Pathans, Hazara, etc.) in order to give us better global coverage. Near the end of this project we also gained access to DNA for a sample of Iranian individuals but data collection is on hold until additional funds are available so no analyses are available for this report.

AISNP and LISNP research is recursive in many senses. Based on existing data a SNP is evaluated for its potential usefulness. If the candidate is potentially useful as part of a minihap or for ancestry inference, we typed all of our large core populations for which we have cell lines established from which we could harvest DNA and then later

evaluated the usefulness of interesting candidate AISNPs or LISNPs on a larger dataset that included population samples for which we only have DNA from collaborations.

AISNP Studies. As sets of very good AISNP panels for continental level discrimination were available (ours and those of other research groups), we are evaluated them to identify the finer-scale population distinctions that needed better resolution. Other SNPs were then sought for allele frequency patterns that strengthened the distinctions needed. At that point, many candidates were evaluated on our core populations and the cycle was repeated.

LISNP Studies. Multiple LD analyses have been carried out on sliding window combinations of potential minihaps using the initial criterion of 3 to 5 SNP markers within a span of about 10kb or less. As might be imagined, many different combinations are possible in regions in which we have dense SNP coverage as a result of previous grants from NIH. By systematically searching through various resources (especially our collection of high density SNP typings on world populations as well as the Human Genome Diversity Panel SNPs), we identified 61 candidate minihaps including the 8 illustrative minihaps that we published in the *European Journal of Human Genetics* (Pakstis et al., 2012). After obtaining comparable typings on for the candidates on a core of 54 to 57 of our population samples we identified a reasonable panel of about 25 minihaps that are either unlinked or effectively independent on samples of unrelated individuals because of the large physical distances separating those minihaps that are on the same chromosome.

PISNP Studies. We do not have detailed phenotypes (such as hair and skin pigmentation) on our population samples from around the world but we studied the phenotype informative SNPs for their population generalization. Already published is our work on *OCA2* (Donnelly et al., 2011).

Methods

Laboratory Methods

For every SNP identified as potentially useful in other datasets, we generally typed it on all individuals of our current basic set of 57 populations being used for forsensic studies (Table 1a). Each SNP was genotyped by TaqMan using assays obtained from Applied Biosystems. Our procedures are organized in such a way that we are extremely efficient at typing small numbers of SNPs (that is fewer than 250) on very large numbers of samples (that is, more than 2,500). With this arrangement, it is more efficient and cost effective to type samples in-house rather than design a custom Illumina (for example) array. This is especially true when the decisions on which SNPs to type next often depend on what we find from SNPs recently typed. We have organized our DNA samples and typing procedures in such a way as to maximize robotic handling. The working stocks (0.1 mg/ml) of all DNA samples of individuals are aliquoted to 384-well plates in advance of need for them, utilizing an automated pipettor, and the plates are stored dried at room temperature until used. TaqMan assay is aliquoted into the plates by the same robot, the plates are sealed, and PCR'd immediately on stand-alone thermal cyclers. Thus, all pipetting is robotic and performed in a very stereotypic format so that every different 384-well plate, labeled WPI to WPVIII (World Plate I, etc.), always contains DNA of the same individual in the same position matching our standard sample sheets and corresponding electronic files. Electronic data recording is automatic with only the WP # and genetic system specified for the plate reader, an ABI 7900HT SDS. Allele calling is automatic with AB's SDS software with visual inspection of the results. We always carried out a second round of testing on the samples giving unclear or no results to get results for as many of the first-round failures as possible. Aliquoting of the failed DNA samples for the second round of typing is done by a Cherry Picker robot, reading a file generated by the SDS software and custom modules, thus automating the selection and aliquoting of DNA from the standard working stock boxes, but only of the “failures”. We developed an automated software system to convert the SDS interpretations into genotypes and to transfer the data into our local data management system, PhenoDB The software is designed to

facilitate high throughput and includes many error-checking components. For SNPs identified from scanning the HGDP Illumina 650Y data, we employed special DNA typing plates of those individuals that we did not include when we typed the Illumina 650Y SNPs on ~1300 of our samples.

Table 1a. 57 Population samples routinely studied at Kidd Lab from cell lines

Region	Population	N		Region	Population	N
AFRICA (12)	Biaka	68		Siberia (3)	Komi Zyriane	47
	Mbuti	39			Khanty	50
	Yoruba	77			Yakut	51
	Ibo	48		S.C. ASIA (3)	Keralites	30
	Hausa	39			Kachari	17
	Lisongo	8			Thoti	14
	Masai	20		E. ASIA (7)	Han, SF	60
	Chagga	45			Han, Taiwan	49
	Sandawe	40			Hakka	41
	Zaramo	40			Korean	54
	Ethiopian Jews	32			Japanese	49
	Afr. Americans	90			Ami	40
SW ASIA (5)	Kuwaiti	16			Atayal	42
	Yemenite Jews	43		S.E. ASIA (3)	Cambodians	24
	Druze	126			Laotians	119
	Samaritans	39			Malaysians	11
	Ashkenazi Jews	78		PACIFIC IS (4)	Nasioi	23
EUROPE (11)	Adygei	54			Micronesians	37
	Chuvash	42			Papuans/New Guinea	22
	Roman Jews	27			Samoans	8
	Sardinians	38		N. AMERICA (4)	Cheyenne	56
	Russian, Vologda	48			Pima, Arizona	51
	Russians, Archangel	33			Pima, Mexico	99
	Hungarians	144			Maya, Yucatan	52
	Finns	34		S. AMERICA (5)	Guihiba, Colombia	13
	Danes	51			Quechua, Peru	22
	Irish	114			Ticuna	65
	Euro. Americans	92			Surui, Rondonia	47
					Karitiana	57

Table 1b. Additional 25 population samples studied from limited amounts of DNA.

Region	Population	N		Region	Population	N
N. AFRICA	Libyans	71		E.&C. ASIA	Iranians	44
	Smar, S. Tunisia	65			Khazak	48
	Sousse, C.Tunisia	49			Uigur	48
E. AFRICA	Negroid Makrani	28			Mongols	74
	Somali	22			Baima Dee	42
	Ethiopian Jews*	21			Qiang	40
SW ASIA	Ashkenazi Jews*	100			Khamba, Tibet	36
	Palestinians	49			Hmong	30
	Catalans	60			Hlai	55
EUROPE	Greeks	56			Tsaatan, Mongolia	51
	Toscani	89			Outer Mongolians	71
S.CEN.ASIA	Hazara	87				
	Mohanna	53				
	Pathans	75				

Analytic and Statistical Methodology

Here we provide an overview of various statistical methods and software implementations of those methods that we have employed in work on AISNPs, LISNPs and PISNPs.

STRUCTURE (Pritchard et al., 2000; Falush et al., 2003) is an MCMC Bayesian approach for inferring, assuming a specific number of ancestral/underlying populations, the proportional ancestry of each individual based on multi-locus genotype data. The method uses existing genotypes of individuals to determine the allele frequencies for the assumed ancestral/underlying populations (clusters) and uses those inferred frequencies to simultaneously estimate the proportional ancestry of each individual. No prior knowledge of the ancestry of each individual is used. In practice, one does not know, a priori, the optimal number of underlying clusters to assume and must try several and consider the likelihoods of the different hypotheses. Also, since the method uses an MCMC procedure, multiple replicates for the same number of clusters must be run and the consistency of patterns evaluated. The output statistics allow ready inference of relative contribution of different loci to each cluster and probability of miss-assignment of an individual to a cluster as well as the probability of the mis-assignment of a population to a given cluster. The output graphics allow easy visualization of these

statistics. Figures 3, 8, and 15 in this report illustrate some of the ways STRUCTURE output can be visualized. We have extensive experience with STRUCTURE (e.g., Kim et al., 2005). A different program for inferring individual ancestry proportions is *frappe* (Tang et al., 2005) which was used by Li et al. (2008) for their very large dataset. This program is based on maximum likelihood and implements an EM algorithm rather than Bayesian estimation. It is especially designed to handle very high density SNP data.

A heatmap is a graphical representation of two-dimensional data in a high-resolution color graphic. It is well-known in the natural sciences and is one of the most widely used graphs in the biological sciences. For example in molecular biology, heatmaps are often used to represent the level of gene expression obtained from DNA microarrays. A heatmap displays and simultaneously reveals row and column hierarchical cluster structure in a data matrix (e.g., Figure 2). The basic idea implemented in a heatmap is the ensemble of cluster trees to the rows and columns of the data matrix; thus it is useful to visually summarize data by placing similar values near each other according to the clustering. The high-resolution color graphic of a heatmap is displayed as a rectangular tiling of a data matrix with cluster trees appended to its margins. Among many computer programs that have been developed to produce heatmaps to date, we have used The Heatmap.Plus Package version 1.3 implemented in R in our analysis.

The method developed by Sampson et al. (2008), differs from others in that it assumes our knowledge about overall population relationships and uses a greedy algorithm to identify the few SNPs, from a larger set, that are the most informative for predicting ancestry of many individuals from many populations. The method has been used to provide an initial set of candidate AISNPs to enrich our working panel. The method should also help us reduce our combined panel of ~800 SNPs to those most important for our large set of populations and for defined subsets of populations.

Multiple methods can be used for dimensionality reduction and characterization of the structure of the data, including widely used principal components analysis (PCA), singular value decomposition (SVD), independent component analysis, non-negative matrix factorization, eigen decomposition, random projection and factor analysis. Many variations of principal components may be used (e.g. Horne and Camp, 2004; Lin and

Altman, 2004). More recently in the context of characterization of population structure, Biswas et al. (2009) presents a PC-based approach by performing a SVD analysis of the covariance matrix between individuals using the Li et al. (2008) 650,000-SNP HGDP-CEPH data for 52 populations. SVD analysis using the covariance matrix after normalizing the data matrix is equivalent to performing PCA. One advantage of using the SVD analysis is that it can detect and extract small signals from noisy data. Thus rather than focusing solely on the top axes of variation as in a typical PCA, the SVD analysis allows a more detailed investigation of patterns of intracontinental structure in the lower-ranked significant PCs, where substantial information about population structure resides. In general, PC-based approaches are very useful techniques in data analysis and visualization.

F_{st}, is a common, useful approach for quantifying the relative amounts of variation shown by different SNPs across different populations. It is implemented in several programs we have written over the years. Informativeness (Rosenberg et al., 2003; Rosenberg, 2005) is a different statistic devised for the purpose of identifying genetic polymorphisms that are informative for population ancestry. We have also implemented the Infocalc program provided by Rosenberg in our laboratory.

Data Management

The structure and functionality of our lab's database, PhenoDB, remain largely as we originally published (Cheung et al., 1996), and the system has been converted to a Web-accessible system with multiple security features to prevent unauthorized access. The database is firewall and password protected and backed up daily. (In any case, no personally identifying information is included in the database.) Finally, the database allows analysts to retrieve data by individual or population for one or multiple polymorphisms and export those data in different formats ready for input to various analysis programs. Once data on a polymorphism are considered final, the data in PhenoDB can be semi-automatically transferred into ALFRED using software we have developed to use tables of metadata and intermediate XML representation of the data.

Implications for policy and practice

Policy. No highly differentiating set of AISNPs that is both extensively validated and replicated is currently available in the public domain. Investigators can use commercial ancestry companies, but their markers and statistics are often proprietary and the underlying science unavailable. Some past uses of such a commercial company have been controversial. Forensic laboratories may be reluctant to undertake such studies for all those reasons. Currently, no panels of mini-haplotypes or micro-haplotypes for forensic use are available and only one forensic SNPSTR has been published. To the extent our research demonstrates the utility of these enhanced markers, they can be useful in specific types of forensic cases. Our extensively validated and documented data and their analyses are either published or are being prepared for publication to place them in the public domain. Forensic labs will have greater reason to use these markers than proprietary ones. Our data on Phenotype Informative SNPs provides a basis for qualifying interpretations of DNA typings and preventing simplistic/erroneous interpretation until the biological basis for interpretation is clear.

Other forensic issues. This project has not addressed a large number of forensic issues. Chain of custody is not directly addressed. However, because SNPs have the potential for multiplexing and robotic handling, there can be many fewer transfer steps involving human handling once the DNA is isolated, somewhat simplifying chain of custody issues in the lab. The use of SNPs addresses the issue of degraded samples since small amplicons can be used. SNPs are superior to forensic STRPs for detection of ancestry because greater allele frequency variation can exist. Allele dropout is a potential problem with SNPs just as with STRPs; only extensive studies of the specific method used with diverse samples, including degraded samples, can determine whether it is a problem with any specific marker. Our studies have detected frequent systematic genetic problems with allele dropout, such as variants under a PCR primer or additional variants under the TaqMan probe, but they do not generalize to other typing methods for the same marker. Our work on developing LISNPs and PISNPs, especially in the forms of minihaps and microphaps, 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.

TaqMan is not especially useful in a forensic setting because it cannot be multiplexed at the detection stage. However, it is possible to multiplex at the initial PCR amplification stage. While we have used this approach to genotype many markers on a very small amount of DNA, we recognize that our research methods do not extend to actual casework. By making all our data public we make it possible for any lab or company to develop alternative assays for most or all of the SNPs. They need only validate the assay for the SNPs we identify, and do not need to validate the population genetics underlying the selection.

Practice. To the degree that SNPs identified from our studies are brought before the courts, this work that we have published and the data we have deposited into ALFRED provide a firm scientific basis for their acceptability.

Results

Progress on identifying Ancestry Inference SNPS (AISNPs)

Overview. To ascertain the ethnicity of individuals from SNPs in forensic investigations an optimized panel of ancestry informative markers (AIMs) is required; each of those needs to demonstrate large allele frequency differences among populations and have also been studied on a very broad set of the world's populations. AIMs cannot be expected to distinguish among populations automatically if the populations have not been studied previously. A panel of AISNPs is not necessarily useful for estimating an individual's ancestry if the ancestral population(s) relevant for that individual have not been included among the ethnic groups already studied for those AISNPs.

As referenced above, very large numbers of SNPs can be assembled that can provide accurate discrimination for at least five to seven geographic groupings of the world's human populations. However, it would be most useful for forensic work to identify a small but efficient and robust set of markers that can provide enough information so that ancestry can be assigned at a high enough probability to be useful for a forensic investigation. In the forensic context a small number of SNPs can mean lower costs and possibly faster turnaround. A small number of highly selected SNPs can be sufficient for accurate estimation of ancestry (Ding et al., 2011). The challenges for a productive search are to select population samples that are representative of diverse geographical regions, to use large enough sample sizes so that sampling errors are minimized, and finally to identify those polymorphisms most able to distinguish among those populations. We have met the first two of these challenges by having sample sizes averaging 50 individuals in the populations studied and the second by employing enough different population samples (up to a total of 73) that we have several samples from each major geographic region we are investigating. We have selected candidate SNPs using a wide variety of methods. As of the end of the reporting period for our project, we have identified a best set of 55 AISNPs that

constitute an efficient enough panel to assign an individual's ancestry to one of seven to eight biogeographic world regions.

***Strategy*.** The first two challenges above, number of population samples and sample sizes, are addressed by our use of many sources of data to identify potential AISNPs. We initially used the Applied Biosystems database of allele frequencies of four populations (Japanese, Chinese, Europeans, African Americans) for the TaqMan probes they sell. The approximately 650,000 SNPs tested on the HGDP-CEPH panel of over 1000 individuals from 51 populations was another source that others have also used (Kosoy et al., 2009; Nievergelt et al., 2013). We augmented the HGDP with the same SNPs tested on 1300 additional individuals not present in the HGDP; these included additional individuals from the populations we contributed to the HGDP and additional populations. We used our own laboratory database of about 4000 polymorphic markers typed on from 44 to 56 populations consisting of a total of nearly 3000 individuals. These populations and SNP data on them were the accumulation of many different studies of allele frequency variation done for a variety of reasons, e.g., pharmacogenetics (Speed et al., 2009). As they became available we screened other large datasets for promising candidate AISNPs.

We explored several approaches to select candidate SNPs, compare them, and balance the information a selection provided. Ultimately, the combination of approaches would have to be considered empiric. We calculated Fst across different sets of populations with data available in our lab and/or various public datasets. By "balancing" the selection of SNPs such that each geographical region was represented in the final selection, we sought to minimize the large excess of SNPs that have frequencies distinguishing African populations from populations in the rest of the world, a dichotomy that often outweighs most other distinctions among populations. After a considerable amount of testing alternative sets of SNPs and switching individual SNPs in and out, we arrived at a more efficient provisional panel of 55 AIMs.

Many candidate SNPs initially had data on a small number of populations; we evaluated those potential sites that had the largest absolute frequency differences or the largest Fst values for broader global variation on our initial set of 44-populations. We also typed two other published panels of AISNPs. In Kidd et al. (2011), the 128

SNPs identified by Seldin's group (Kosoy et al., 2009) were analyzed on 119 population samples that included samples in our laboratory, the HGDP-CEPH panel, and the HapMap. We collaborated on the Nievergelt et al. (2013) study to type 40 of the 41 SNPs in that panel on most of our populations. Those two studies have no SNPs in common even though both used the HGDP data (Li et al., 2008) to identify the AIMs selected. In both cases some SNPs had already been identified by us as good candidates; both studies also included other SNPs we had not previously identified as excellent candidates. All of the markers from those two studies were included in the set of several hundred candidate AISNPs that was typed on the remaining samples in our lab to complete a comprehensive dataset with no missing population-SNP data points. The global coverage of our large number of candidate AISNPs consisted of 63 populations with a total of 3063 individuals (see Table 2).

Table 2. The 73 populations studied for the best set of 55 AISNPs. The original populations were later supplemented with data on 1000 Genomes populations.

Region	Population	Sample Size	Source of data
Africa	Biaka	69	Kidd Lab
Africa	Mbuti	38	Kidd Lab
Africa	Lisongo	8	Kidd Lab
Africa	YRI	88	1000 Genomes
Africa	Yoruba	77	Kidd Lab
Africa	Ibo	48	Kidd Lab
Africa	Zaramo	40	Kidd Lab
Africa	Hausa	39	Kidd Lab
Africa	Masai	20	Kidd Lab
Africa	LWK	97	1000 Genomes
Africa	Chagga	45	Kidd Lab
Africa	Sandawe	40	Kidd Lab
Africa	ASW	61	1000 Genomes
Africa	Afr Americans	89	Kidd Lab
Africa	Somali	17	Kidd Lab
Africa	Ethiopians	37	Kidd Lab
S Asia	NegroidMakrani	26	Kidd Lab
SW Asia	Kuwaiti	14	Kidd Lab
SW Asia	Samaritans	39	Kidd Lab
SW Asia	Yemenites	41	Kidd Lab
SW Asia	Palestinians	68	Kidd Lab
SW Asia	Druze	102	Kidd Lab
Europe	Roman Jews	27	Kidd Lab

Europe	**Ashkenazi**	**116**	**Kidd Lab**
Europe	**TSI**	**98**	**1000 Genomes**
Europe	**IBS**	**14**	**1000 Genomes**
Europe	**Sardinians**	**35**	**Kidd Lab**
Europe	**Adygei**	**54**	**Kidd Lab**
Europe	**Greeks**	**53**	**Kidd Lab**
Europe	**Hungarians**	**89**	**Kidd Lab**
Europe	**Chuvash**	**42**	**Kidd Lab**
Europe	**Irish**	**114**	**Kidd Lab**
Europe	**CEU**	**85**	**1000 Genomes**
Europe	**EuroAmericans**	**89**	**Kidd Lab**
Europe	**Russians, Archangelsk**	**33**	**Kidd Lab**
Europe	**Russians, Vologda**	**47**	**Kidd Lab**
Europe	**FIN**	**93**	**1000 Genomes**
Europe	**Finns**	**34**	**Kidd Lab**
Europe	**GBR**	**89**	**1000 Genomes**
Europe	**Danes**	**51**	**Kidd Lab**
W Siberia	**Komi Zyriane**	**47**	**Kidd Lab**

It is important to balance the information from different SNPs to assure that different regions of the world are distinguishable in a robust way. We employed a number of methods. The use of heatmaps for the candidates helped by graphically portraying redundancy in the SNP information. Pairwise Fst calculations for each SNP across populations from different regions helped identify those SNPs best at certain distinctions, such as Europe vs. East Asia, so the “best” SNPs were used in the balancing (Kersbergen et al., 2009). We also employed STRUCTURE (Pritchard et al., 2000) as one first-pass method of identifying the SNPs that differentiated most between the clusters identified. Once we had identified our set of 55 AISNPs on our 63 populations, we extracted the data for 10 populations from the 1000 Genomes data. The resulting data set included 73 populations and 3,884 individuals.

Laboratory Methods. The population samples from our lab were typed for all SNPs by TaqMan SNP Genotyping Assays® (Applied Biosystems, Foster City, California, USA) in three microliter reactions following the manufacturer’s instructions. More details were described earlier. The genotypes of the samples in the 1000

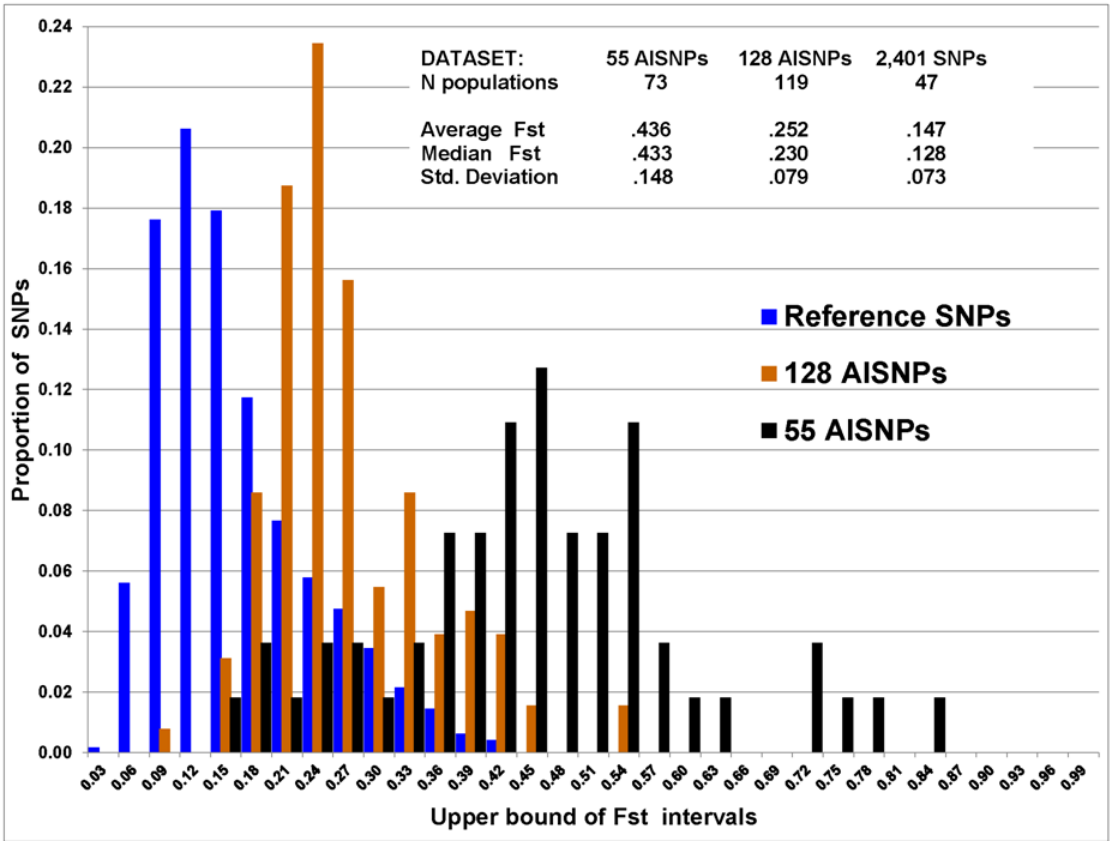
Genomes Project were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/. Overall, missing genotypes account for rather low percentages of the total (typically on the order of 1 to 1.5%).

Statistics. Fst was calculated for the allele frequencies using the formula of Wright with no modification for sample size variation among the population samples. We could have but did not also use I_n (Rosenberg et al., 2003) because it was shown to be highly correlated with Fst (Ding et al., 2011). We used both overall Fst in selecting candidate SNPs and pairwise Fst in balancing the panel to include SNPs informative for different distinctions among populations. Heatmaps were calculated using the public program in R. Principal component analysis (PCA) of population sample allele frequencies used XLSTAT (version 2009.4.07; Addinsoft SARL, http://www.xlstat.com/en/company/)). STRUCTURE (version 2.3.4; software freely available at http://pritch.bsd.uchicago.edu/structure.html) (Pritchard et al., 2000; Falush et al., 2003) was also used to evaluate and visualize the degree to which sets of sites distinguish among these populations. The various analyses typically used a burn-in of 20,000 followed by 10,000 iterations with a model of correlated allele frequencies specified. Specific solutions were then plotted using DISTRUCT 1.1 (this software is freely available at http://rosenberglab.bioinformatics.med.umich.edu/distruct.html) (Rosenberg, 2004). For the final set of 55 AISNPS, for example, ten replicates at each of the "K" cluster levels of 2 through 6 were used and 20 replicates were obtained at K = 7 and K= 8 and evaluated using CLUMPP. This software is freely available at http://rosenberglab.bioinformatics.med.umich.edu/clumpp.html (Jacobsson et al., 2007). The matrix of pairwise similarities among replicate runs was used to identify different overall patterns based on high G values among runs with the "same" pattern and lower values for runs with different patterns.

Results for best 55 AISNP panel. The final list of 55 AISNPs is given in Table 3. The allele frequencies are available in ALFRED for these 73 populations and any other populations that have available data. The data can be retrieved under the individual rs numbers or through the "SNP Sets" menu as "KiddLab Set of 55 AISNPs". There were

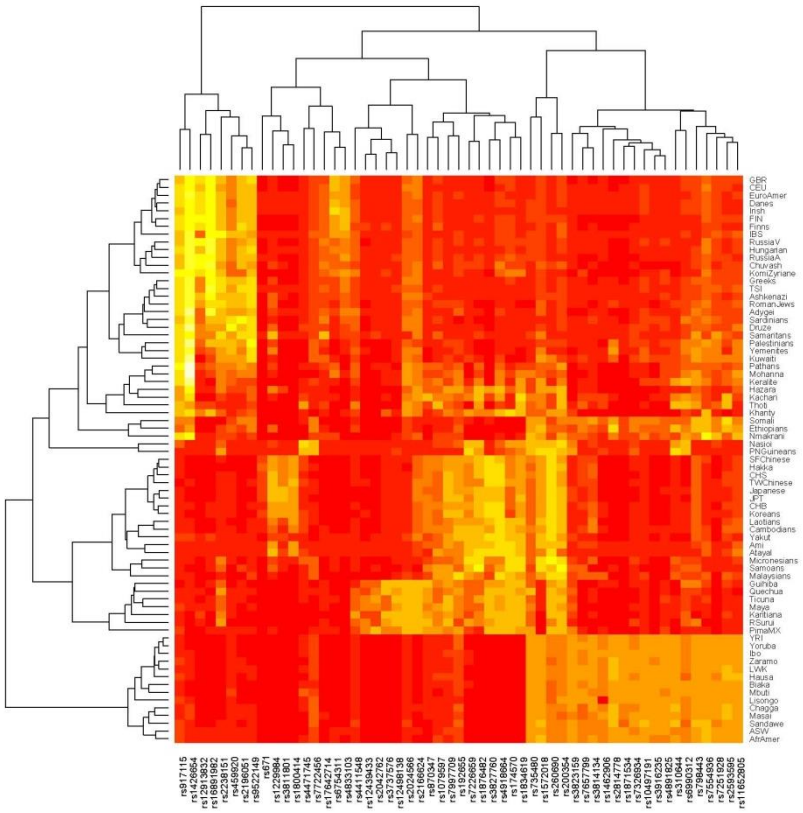
no significant deviations beyond chance levels from Hardy-Weinberg ratios given 55 x 73 = 4015 tests. Figure 1 compares the distribution of Fst for these 55 AISNPs with two other sets of markers: an essentially random set of SNPs and the 128 AISNPs (Kidd et al., 2011a). Though different numbers of populations are involved, there is considerable overlap in the population samples and the biogeographic ranges of populations are the same. Clearly, on average we are dealing with a set of SNPs with greater global variation than the 128 AISNPs. The Nievergelt et al. (2013) AISNPs, based on available population data in ALFRED, have a mean and median Fst of 0.36, intermediate between the 128- and 55-AISNP panels.

Figure 1. Comparison of Fst distributions. Two previously published distributions (Kidd et al., 2011a) are compared to the distribution for the set of 55 AISNPs. The two previous distributions are based on a reference set of SNPs typed on the Kidd Lab populations and on the Seldin group's set of 128 Ancestry Informative SNPs typed on a larger set of populations including the Kidd Lab populations. Because all three sets include the basic 47 Kidd Lab populations, the additional and different populations in the two larger studies are not sufficient to invalidate the marked differences in the distributions.



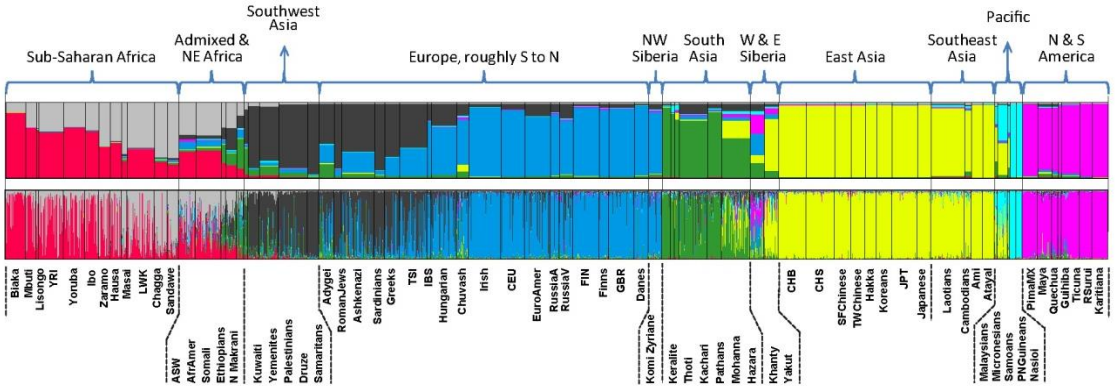
The heatmap in Figure 2 is based on the population allele frequencies for the 55 AISNPs. It allows a very quick visualization of (1) the relationship of each SNP in a data set to the others, and (2) of how each SNP contributes to distinguishing each population from the others. The heatmaps show the relationships of the SNPs graphically in the marginal dendrograms for both populations and markers. Of particular interest is the ability to examine the nature of the information provided by these individual markers for differentiation of the specific populations analyzed. Although STRUCTURE allows evaluation of potential AISNPs, it is cumbersome to use and not particularly useful in our effort to identify as small a set of SNPs as possible that distinguishes multiple regions of origin of individuals. The empiric approach using multiple methods as described above produced surprisingly good results.

Figure 2. The heatmap clustering of 73 populations and 55 AISNPs. The upper left block represents Europe through South Central Asia. The large middle block represents East Asia and below that the Native Americans. The bottom right block represents Africa. Different SNPs clearly contribute differently to population distinctions; The lengths of the branches in the dendograms give one perspective on this.



In the most likely STRUCTURE run at K=8 the individuals are assigned to seven distinct clusters in which most individuals in most populations fall into a single one of those clusters (Figure 3). At K=8 the results for most individuals in most populations are essentially unaltered from the pattern at K=7 (not shown) but a west to east cline is newly indicated for sub-Saharan African populations.

Figure 3. The most likely of the 10 Structure analyses at K=8 for the full dataset. The results are plotted as the average assignments for each population and as the individual assignments. Two clines are clearly indicated: one from West to East in Africa and one from the Middle East and Southeast Europe to North Europe.

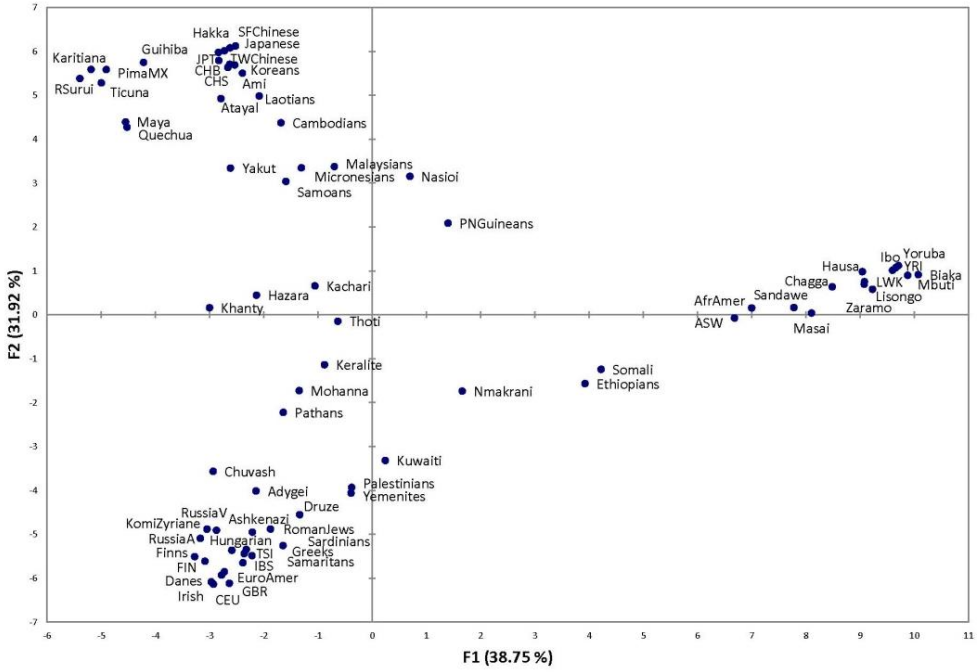


The STRUCTURE analysis is based on individual genotypes and shows how individuals can be clustered into a single population that comes close to meeting Hardy-Weinberg genotype expectations. Taken together the heatmap and STRUCTURE analyses show that clusters exist in which several populations are essentially indistinguishable. Thus, using a likelihood function such as implemented for this panel in FROG-kb cannot be expected to identify routinely the specific population from which an individual originates. Rather, the best resolution one can be reasonably confident of is that the cluster an individual belongs to will be identified.

Using Principal Components Analysis on the allele frequencies of the populations shows four distinct groupings of populations based on the first 3 components (Figure 4): a highly distributed African group, a more tightly clustered East Asian group, a modestly clustered Native American group, and a European-Southwest Asian group. This pattern reflects the geographic clustering of populations: geographically intermediate populations tend to be placed in more intermediate positions. African populations show a West-East cline toward non-Africans echoing the cline in the STRUCTURE analysis.

Figure 4. Principal Component Analysis of 73 populations using 55 AISNPs. The first PC accounts for 38.75% of the variation and primarily separates African populations from the rest of the world. The second PC accounts for 31.9% of the variation and primarily separates Europe from East Asia and the Americas. The two components account for 70.7% of the variance. The third PC accounts for 12.5% of the variance and completely separates American Indians from East Asians.

x-axis=PC#1, y-axis=PC#2



x-axis=PC#1, y-axis=PC#3

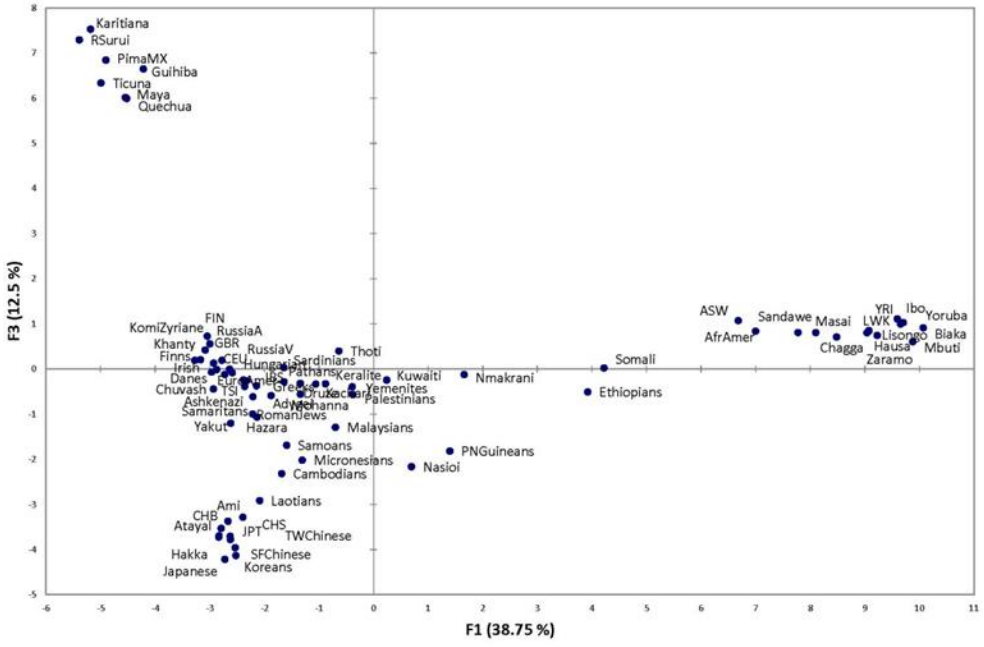


Table 3.The list of 55 best AISNPs

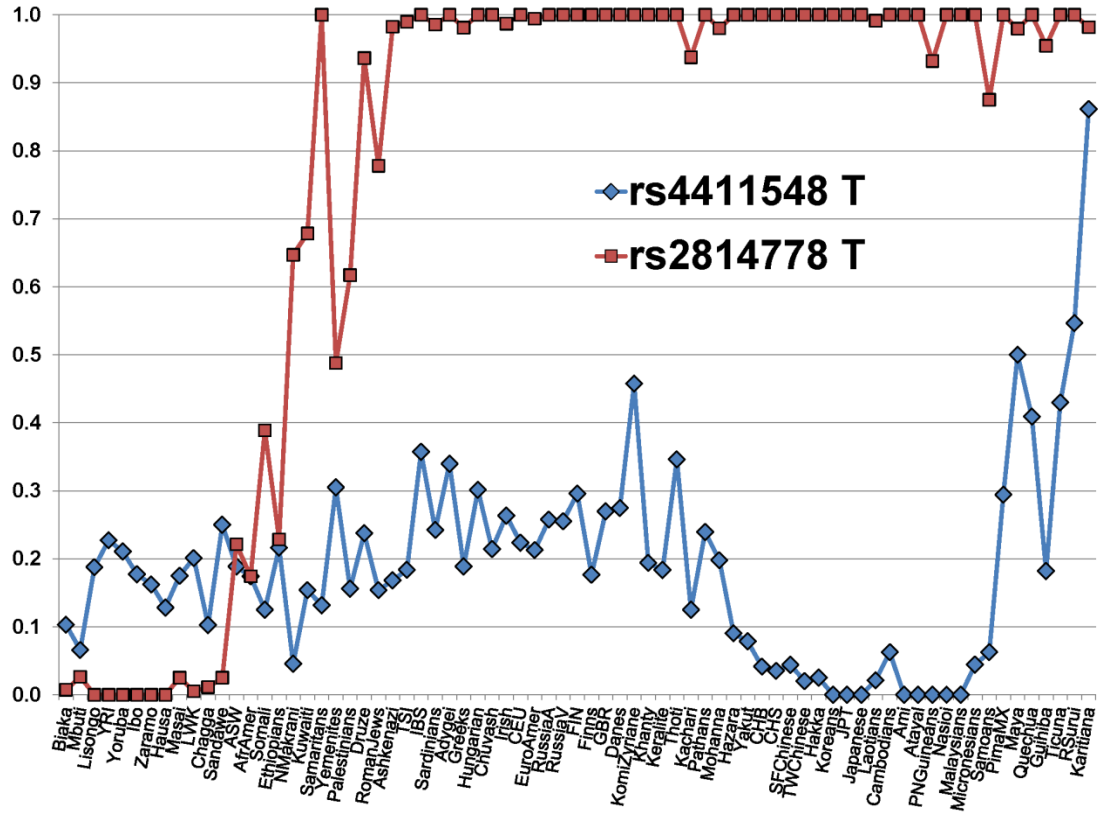
dbSNP rs#	Chr	Build 37 nt position	73-population Fst
rs3737576	1	101,709,563	0.44
rs7554936	1	151,122,489	0.39
rs2814778	1	159,174,683	0.82
rs798443	2	7,968,275	0.34
rs1876482	2	17,362,568	0.75
rs1834619	2	17,901,485	0.50
rs3827760	2	109,513,601	0.71
rs260690	2	109,579,738	0.49
rs6754311	2	136,707,982	0.41
rs10497191	2	158,667,217	0.54
rs12498138	3	121,459,589	0.48
rs4833103	4	38,815,502	0.37
rs1229984	4	100,239,319	0.43
rs3811801	4	100,244,319	0.45
rs7657799	4	105,375,423	0.44
rs16891982	5	33,951,693	0.69
rs7722456	5	170,202,984	0.20
rs870347	6	6,845,035	0.35
rs3823159	6	136,482,727	0.50
rs192655	7	90,518,278	0.21
rs917115	8	28,172,586	0.35
rs1462906	8	31,896,592	0.54
rs6990312	8	110,602,317	0.34
rs2196051	8	122,124,302	0.43
rs1871534	8	145,639,681	0.48
rs3814134	9	127,267,689	0.47
rs4918664	10	94,921,065	0.53
rs174570	11	61,597,212	0.51
rs1079597	11	113,296,286	0.16
rs2238151	12	112,211,833	0.36
rs671	12	112,241,766	0.22
rs7997709	13	34,847,737	0.37
rs1572018	13	41,715,282	0.41
rs2166624	13	42,579,985	0.30
rs7326934	13	49,070,512	0.54
rs9522149	13	111,827,167	0.44
rs200354	14	99,375,321	0.32
rs1800414	15	28,197,037	0.57
rs12913832	15	28,365,618	0.52
rs12439433	15	36,220,035	0.39
rs735480	15	45,152,371	0.39
rs1426654	15	48,426,484	0.73
rs459920	16	89,730,827	0.24
rs4411548	17	40,658,533	0.14
rs2593595	17	41,056,245	0.47
rs17642714	17	48,726,132	0.18
rs4471745	17	53,568,884	0.27
rs11652805	17	62,987,151	0.39
rs2042762	18	35,277,622	0.43
rs7226659	18	40,488,279	0.40
rs3916235	18	67,578,931	0.63
rs4891825	18	67,867,663	0.53
rs7251928	19	4,077,096	0.47
rs310644	20	62,159,504	0.58
rs2024566	22	41,697,338	0.31

Implications for Future Research. Clearly, this best panel of 55 AISNPs can be improved. The problem is finding SNPs that provide a clearer differentiation of certain populations or groups of populations without detracting from differentiation among some other populations. As noted in Kersbergen et al. (2009) some SNPs can simply add noise. We note that several of the SNPs that help differentiate European individuals from the rest of the world are not fixed for the Europe-specific allele. With genotype differences among individuals there is a distribution and some individuals will tend to have more of the non-European alleles than other individuals. At higher K values the Structure analyses tend to use this Mendelian segregation to classify individuals in all European populations "randomly" into two or three different clusters. In general, even if a SNP has extreme frequency variation between, say, East Asians and Native Americans, but the frequencies in Europe and Southwest and South Asia are all intermediate in no strong pattern, that SNP is adding noise to the differentiation of those populations. The SNP with the lowest Fst in these 73 populations, rs4411548, illustrates exactly that situation (Figure 5).

The frequency of one allele is near zero in East Asian and Pacific populations and ranges from 19 to 86 % in Native Americans. In contrast, that allele ranges from 2% to 45%, with most populations between 12 and 30%, in Africans, Europeans, and Southwest and South Central Asians. Thus, it will be a difficult task to find additional SNPs to differentiate both globally and within regions while minimizing the total number of SNPs. An alternative approach that we can pursue is to consider a second tier of SNPs that are good within a region but not necessarily good, or as good as existing AISNPs, for global differentiation. We are currently working on one such second tier of AISNPs for the eastern half of Asia. Phillips et al. (2013) have proposed such a regional panel focused on distinguishing European from South Asian populations.

While improvements will likely be possible for this panel, the analyses done show it is a very good first tier panel for identifying major biogeographic regions for the ancestry of an individual. Future tests of its robustness will require that additional populations be tested for these SNPs to determine how well these identify ancestries for individuals from populations that are intermediate to the existing 73 population samples.

Figure 5. The population frequency profiles for two of the 55 AISNPs. The SNP with the highest Fst (rs2814778) is in the upstream region of DARC, the classic Duffy blood group locus. The SNP with the lowest Fst (rs4411548) is intronic in ATP6V0A1 on chromosome 17.



Progress on identifying Lineage Informative SNPS (LISNPs)

Overview. Haplotype systems based on multiple SNPs that are closely linked have been advocated in recent years (Pakstis et al., 2007; Butler et al., 2008; Ge et al., 2010; Pakstis et al., 2012) as one type of forensically useful DNA marker (LISNPs, lineage informative SNPs). The multiple alleles, analogous to STRPs, of these haplotypes are more informative than simple two-allele SNPs for identifying biological relatives and have the potential to be more useful as well in inferring the ethnicity of an individual's ancestors. The value of a locus for identifying familial relationships, i.e., lineage informativeness, is inversely related to the heterozygosities in the relevant

population. Thus, the more polymorphic a locus, the greater is the chance that the relevant alleles are uncommon in general but more likely to be found more commonly among relatives. For large datasets containing many hundreds of SNPs quite sophisticated methods of inferring familial relationships have been developed (for example, Manichaikul et al., 2010). However, smaller numbers of loci can be used in a likelihood context if the loci are sufficiently polymorphic. SNPs can generate a multi-allelic locus by generating haplotypes of molecularly close SNPs (Ge et al., 2010; Pakstis et al., 2012). To be useful in forensics, the SNPs must comprise multiple alleles (haplotypes) with sufficiently small rates of intra-locus recombination and mutation so that identity by state allows an assumption of identity by descent within a family. The haplotypes must also have known population frequencies and ideally at least moderate heterozygosity in most populations of relevance in forensics.

CODIS STRPs are certainly appropriate markers for identifying familial relationships but at present there are too few for highly reliable inference. Our objective has been to generate a larger number of SNP-based, multi-allelic haplotype loci. Based on our research over many years on multi-SNP haplotypes and their global frequency patterns (e.g., recently: Donnelly et al., 2012; Murdoch et al., 2013; and others from earlier years listed at http://medicine.yale.edu/labs/kidd/www/pubs.html), we know that haplotypes in small molecular regions can provide the multiple low frequency alleles that are optimal for familial and lineage assignment of a DNA sample.

Our published work on the concept of mini-haplotypes (Pakstis et al., 2012) and the results we report here on both minihaplotypes and microhaplotypes document the process we have used to find, select and validate haplotyped SNPs for forensic work. In the original aims for this project and for the first full year or so after the start of the reporting period for this project we focused primarily on generating a useful set of minihaplotypes. The panel of 25 minihaplotypes that we identified do provide a useful tool for forensic investigations. However, it became evident from the progress being made in sequencing technology that developing a panel of microhaplotypes (≤200 base pairs) would be feasible and could be more useful for forensic applications since the act of sequencing the DNA segment encompassing the microhaplotypes could also provide definitive phasing of the multi-SNP haplotype alleles. The smaller physical interval of the

microhaps compared to the minihaps also reduces in magnitude the already tiny likelihood that recombination need be a concern in the use of such LISNPs. A large fraction of the minihaplotype alleles (roughly 60 to 80% depending on the particular minihaplotype system) presented here are automatically phase known because of the pattern of alleles present in the defining SNPs. An additional, large fraction of the individual minihap typings that are ambiguous for phase can be either inferred with certainty after a likelihood calculation or else can be inferred with strong odds ratios of 10:1 or better. (The likelihood calculation requires a knowledge of the appropriate haplotype frequencies for the ethnic group from which the individuals derive.) A small fraction, ~1 to 2%, of the individuals typed for the minihaps do remain ambiguous in the populations studied. Consequently, it is easy to see that sequenced microhaplotypes would require less work and more information for forensic applications. However, if the microhaps are not sequenced but simply studied as a composite of the individual SNP typings, as is done with the minihaps, then much of the potential advantage of pursuing microhaps is not gained since haplotype inference and calculation must still be applied in that case.

Because the last year of the project period was unfunded and because our recognition of the potential value of microhaplotypes did not arise until well after the grant period had begun, the panel of 25 minihaps presented here reached a much more satistfactory state as a potential tool for forensic work than the panel of 28 microhaps presented.

Minihaplotypes

Molecularly close (<10kb) SNPs with a very low recombination rate can define haplotypes at what becomes a multi-allelic locus if the linkage disequilibrium is not complete. Such loci have been suggested as potentially important for forensics in defining ancestry and in identifying familial relationships. We identified a set of 25 such mini-haplotype loci (minihaps) and present analyses validating their usefulness for ancestry inference, lineage-clan-family inference, and even individual identification. We show that the potential problem of establishing phase when individual SNPs are

genotyped independently is minimal. The 25 minihaps are highly polymorphic (analogous to STRPs). In the 54 populations studied, 69% of the haplotypes on average are determined with certainty by simple inspection of the SNP typings for those individuals with no more than one heterozgyous SNP per minihap. Another 23% of the multi-SNP haplotypes can be inferred with certainty based on likelihood calculations incorporating population-specific haplotype frequencies that are now available for reference. The 25 minihaps give match probabilities $<10^{-15}$ for most of the populations studied; exceptions include smaller/inbred groups not often encountered in forensic work. STRUCTURE, principal components, and population tree analyses illustrate the panel's utility for ancestry inference.

Methods, Material. Most of the methods and populations studied in characterizing the panel of 25 minihaps matches the information presented in Pakstis et al. (2012). We did expand the number of population samples systematically studied on all of the candidate minihaps from 45 to 54. Table 1a lists 57 populations for which we accumulated results; however, comparable data for analyses was not available for three of the groups (Malaysians, Cheyenne, and Arizona Pima) listed. As we searched for candidate minihaps consisting of ≤5 SNPs we gradually lowered the acceptable molecular extent from no more than 10 KB to no more than 5 KB since many such candidates were encountered. The 10 KB limit was initially selected to provide an upper limit of about 10^{-4} for the recombination rate assuming a genome-wide average of ~1% per megabase and no recombination hot spots within the locus. Because we wanted to obtain minihaps that were not just useful for distinguishing lineages/families, we also increasingly focused our candidate searches in chromosomal regions where useful ancestry markers had already been identified. The primary characteristics sought (high heterozygosity comparable to CODIS STRPs and the ability to resolve unambiguously each individual's SNP typings into the underlying genotypes at a high rate) for the minihaps did not change. When alternative candidates were available with similar satisfactory overall heterozygosity and resolvability of genotypes, candidates were preferred that achieved higher levels in more or all of the 8 major geographical regions into which the populations cluster.

For the extended minihap study software was written and validated to calculate and tally not only those genotypes that can be known by simple inspection (e.g. if no more than one heterozygote occurs among the SNP typings at a minihap for an individual then the two haplotypes are known unambiguously and with certainty) but also to compute relative likelihood of the alternative possibilities when two or more of the SNPs are heterozygous. The population-specific haplotype frequencies from the HAPLO analyses were used as input for the calculation of the population-specific genotype likelihoods. Since the population of origin of each individual was known, the relative likelihoods of the alternative genotypes when an individual had more than one heterozygous SNP at a minihap employed the appropriate frequencies. Because of the moderately strong linkage disequilibrium present and the small molecular extents of the minihaps, a substantial number of additional unambiguous genotypes (with a single nonzero likelihood) were identified from these likelihood calculations. In the remaining cases more than one genotype remained possible. Often the likelihood ratio of the most likely genotype to the alternative(s) exceeded 10, giving statistical confidence in the inference. We primarily summarize for this report those results where the resolvability was achieved unambiguously either by simple inspection or after calculation of the likelihood knowing each individual's population of origin. Analyses requiring the genotypes be known (such as STRUCTURE or the pairwise match comparisons reported in Table 5) included only the genotypes known either with certainty or inferred based on a likelihood ratio of 10:1 or more; otherwise the genotypes were left blank.

Minihap results. We studied 44 minihap candidates systematically on 54 populations and selected the best 25 of these that are unlinked (on separate chromosomes or far apart on the same chromosome) as a minihap panel to provide the proof in principle that such a resource can be of value in lineage/familial identification and ancestry inference. Seven of the 8 minihaps that we presented previously in Pakstis et al. (2012) are among the 44 minihap candidates studied on 54 populations. The PAH SNP from the earlier work was excluded because it was not studied in time on all of the 54 populations. Tables 4 and 6 along with Figure 6 cumulatively provide more detail on all 44 of the minihap candidates considered on the way to the set of 25 minihaps.

Table 4. The 25 “unlinked” mini-haplotypes—defining SNPs, molecular extent, and chromosomal location. When minihaps are syntenic, the distance shown for adjacent minihaps is between the first SNP in each minihap

Chr #	SNP s1 position GrCh37.p5 build 37.3 nt position	bp to next minihap SNP on same chr (s1 to s1 pos)	Closest Gene SYMBOL	Molecular Extent bp	Ordered SNPs by dbSNP rs-number s1	s2	s3	s4
1	76,311,291	82,864,063	MSH4	6,169	rs1498313	rs2047435	rs11161731	n/a
1	159,175,354	71,668,279	DARC	8,360	rs12075	rs3027041	rs3027048	n/a
1	230,843,633		AGT	5,464	rs3789669	rs699	rs1078499	n/a
2	136,552,517	102,627,397	LCT	9,219	rs3213892	rs1807356	rs2304370	n/a
2	239,179,914		PER2	6,676	rs2304676	rs2304674	rs2304672	n/a
3	113,646,657	UNLINKED	GRAMD1C	6,118	rs4422272	rs7612534	rs9865782	n/a
4	38,798,935	61,515,061	TLR1	3,710	rs5743614	rs5743604	rs5743595	n/a
4	100,313,996		ADH7	543	rs2851017	rs2032350	rs1442487	n/a
5	9,619,936	UNLINKED	TAS2R1	9,878	rs41462	rs2234233	rs41469	n/a
6	15,651,132	139,981,630	DTNBP1	9,740	rs760761	rs2619522	rs909706	n/a
6	155,632,762		TFB1M	3,283	rs56237	rs162984	rs9480107	rs1325045
7	27,107,433	95,517,337	HOXA1	9,316	rs2428424	rs2428427	rs6953314	n/a
7	122,624,770		TAS2R16	4,729	rs1204018	rs11980542	rs7785882	n/a
9	12,672,097	123,840,178	TYRP1	224	rs1408799	rs1408800	rs1408801	n/a
9	136,512,275		DBH	5,823	rs2519154	rs739398	rs77905	n/a
10	43,744,583	UNLINKED	RASGEF1A	8,348	rs10899786	rs4987092	rs4987093	n/a
11	5,094,638	UNLINKED	OR52E1	4,756	rs1378745	rs10768550	rs10500617	n/a
12	25,358,828	UNLINKED	KRAS	3,725	rs12587	rs7973450	rs9266	rs712
13	34,841,852	UNLINKED	GAMTP2	5,886	rs7993387	rs9540340	rs7997709	n/a
14	99,373,866	UNLINKED	RPL3P4	3,219	rs2038501	rs200354	rs200352	n/a
15	28,335,820	UNLINKED	OCA2	8,419	rs4778138	rs4778241	rs7495174	n/a
16	53,816,275	UNLINKED	FTO	9,214	rs8050136	rs12597786	rs7201850	rs9941349
17	38,178,149	UNLINKED	MED24	4,081	rs2302776	rs2302778	rs9916158	n/a
19	5,225,031	UNLINKED	PTPRS	1,381	rs12973477	rs2239363	rs4807699	n/a
21	33,034,892	UNLINKED	SOD1	7,112	rs4998557	rs1041740	rs17880135	n/a
			Average:	5,816				
			Median:	5,886				
			Minimum	224				
			Maximum	9,878				

Table 4 lists the 78 SNPs, chromosome locations, and molecular extents for the 25 unlinked minihaps selected for the panel. Figure 6 also illustrates a number of key characteristics (average heterozygosity, genotype resolvability, and Fst values) of these minihaps ranked by genotype resolvability.

The panel consists of twenty-two 3-SNP and three 4-SNP minihaps spread across 18 human autosomes. The average as well as the median molecular extent is about 5.8 KB (range is 224 bp to 9,878 bp). The overall levels of polymorphism and genotype resolvability are very good. The median heterozygosity is 0.62 for the 54

populations studied and ranges from 0.50 to 0.72. Just over two-thirds (68%) of the minihap genotypes are resolvable on average with certainty by simple inspection and, when the likelihood ratio calculation is applied to those cases with more than one heterozygous site, the frequency of unambiguously resolvable genotypes increases to 91% on average. Polymorphism levels and genotype resolvability are also very good when examined for the eight major geographical regions into which the populations are grouped. The native populations of the Pacific Island (4 populations) and the Americas (7 populations) have the lowest (but still very good) median heterozygosities of 0.53 and 0.54, respectively.

While most of the 25 minihaps selected for the panel are on separate chromosomes or separated by essentially unlinked distances (>95MB) when syntenic, there were two inter-minihap distances that were as close as 61.6 MB and 71.7 MB (see Table 4). However, these are essentially independent as well based on the non-significant pairwise linkage disequilibrium (LD) values for all unique pairings of the 78 SNPs. The overwhelming majority of the computed LD values (for SNPs paired from different minihaps) cluster near zero including the two minihaps closer than 95MB but >61MB. No meaningful, non-chance patterns were found for the small percentage of large LD values (r^2>0.6) observed beyond the known (published) bias that is introduced when sample sizes are small (especially when fewer than 25 individuals are sampled for some groups). Supplemental file 3 contains text, tables and figures summarizing the LD analyses. LD values were also computed and summarized for the unique SNP pairings within each minihap providing documentation of the moderate levels of LD embedded in each minihap that help to make them highly polymorphic at levels similar to the CODIS STRPs.

Table 5. Distribution of unique pairwise genotype comparisons for all individuals across the 25 minihaps using haplotype-allele genotypes.

Number of genotype matches	Number of pairwise comparisons of individual genotypes		
	Within populations	Between populations	Combined
0	4	6229	6233
1	41	21725	21766
2	159	38088	38247
3	384	45355	45739
4	621	42777	43398
5	829	33657	34486
6	922	23743	24665
7	890	14690	15580
8	791	8288	9079
9	663	4460	5123
10	504	2161	2665
11	415	898	1313
12	294	366	660
13	185	113	298
14	104	42	146
15	84	14	98
16	32	2	34
17	21	1	22
18	10	0	10
19	6	0	6
20	2	0	2
21	1	0	1
22	0	0	0
23	0	0	0
24	0	0	0
25	0	0	0

A number of analyses help demonstrate the utility of the 25 unlinked minihaps for distinguishing lineages or families in forensic work as well as their potential for ancestry inference. Table 5 shows the distribution of genotype matches (within and between populations separately) for all unique pairings of individuals. The possible genotype match scores range from zero to 25 and this panel of 25 unlinked minihaps provides ample scope to observe the wide range of overall relatedness found in the 54 populations studied. Individuals were included only if the minihap genotype was known

with certainty (on average 91% of individuals per population) or had a likelihood ratio of 10:1 or better against alternatives. For the within population comparisons the peak frequency occurs at 6 genotype matches (out of 25 possible). For the between population comparisons where genetic resemblance should be lower, the peak frequency occurs at 3 genotype matches. The tails of these genotype match distributions are quite long and none of the pairings that result are observed for 22 or more genotype matches. There are nine paired individuals with genotype match scores of 19 to 21 for the within group comparison. Eight of these nine pairs are from small and/or relatively inbred groups where one might predict the highest genotype match scores to occur.

Figure 7 plots the match probabilities and most common genotype frequencies for the panel of 25 unlinked minihaps in each of the 54 populations studied in a format similar to that in our earlier papers (Pakstis et al., 2010; Kidd et al., 2012) for an IISNP panel. The 44 population samples in that study are a subset of the 54 populations in the current study. For the IISNP panel all the populations had match probabilities $<10^{-15}$. By comparison this panel of 25 unlinked minihaps has match probabilities $<10^{-15}$ for only 44 of the 54 populations in the current study. However, all 10 of the populations with match probabilities between 10^{-10} and 10^{-15} are relatively small and/or inbred populations that are not commonly encountered in forensic work in Europe and North America. (The ten populations with the higher match probabilities include the Papuans from New Guinea, Nasioi from the Solomon Islands, the Ami and Atayal from Taiwan, Mexican Pima, Guihiba from Colombia, the Karitiana, Ticuna, and Rondonian Surui from the Amazon basin, and Samaritans from Israel).

STRUCTURE analyses (Pritchard et al., 2000) using version 2.3.4 were also carried out with genotypes for the 25 unlinked minihaps testing a range of clusters (K=2 through 10) with 20 iterations each. Figure 3 shows the results at K=7 for the result with the highest likelihood. The STRUCTURE analyses seem to distinguish clearly individuals from most of the major geographical regions, especially from the populations in Africa, Southwest Asia, East Asia, the Pacific Islands, and the Americas. The populations of Europe, South Central Asia and Siberia are somewhat less distinct blends of inferred ancestral clusters. Results of PCA and population tree analyses also

demonstrate the value of the panel of 25 unlinked minihaps for determining ancestry, at least at the level of major geographical regions (See Figures 9 and 10).

Figure 6. The 25 unlinked minihaps studied on 54 populations ranked by overall resolvability of haplotype-alleles with certainty. The Avg % resolvability by inspection (green triangles) is based on simple inspection of SNP typings. The Avg % ResolvInspection + Calc (green squares) combines simple inspection with calculation of odds ratio employing maximum likelihood haplotype frequency estimates for appropriate population group.

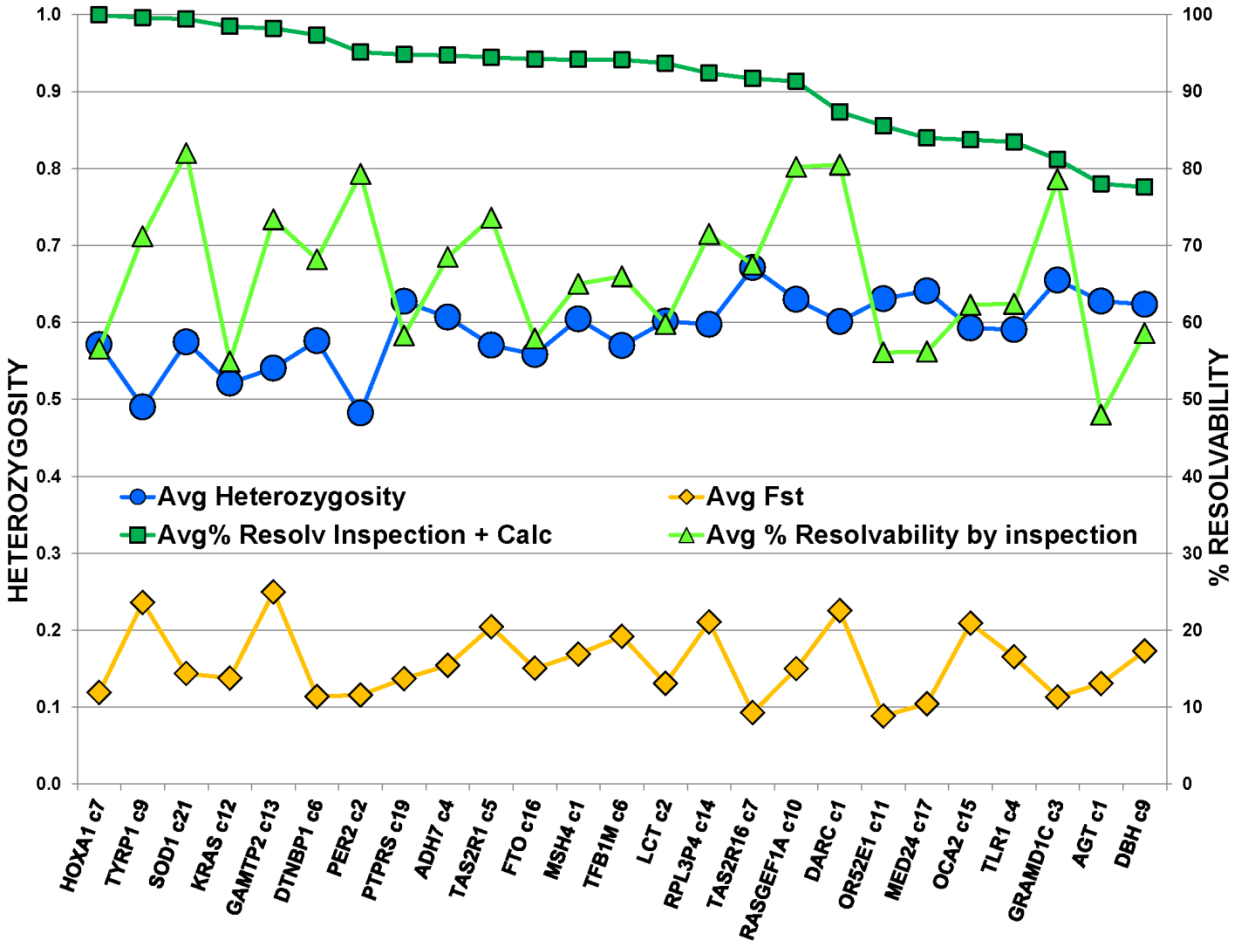


Figure 7. Match probabilities and most common genotype frequencies for 54 population samples utilizing the panel of 25 “unlinked” minihaps. Population abbreviations are explained in Table 7.

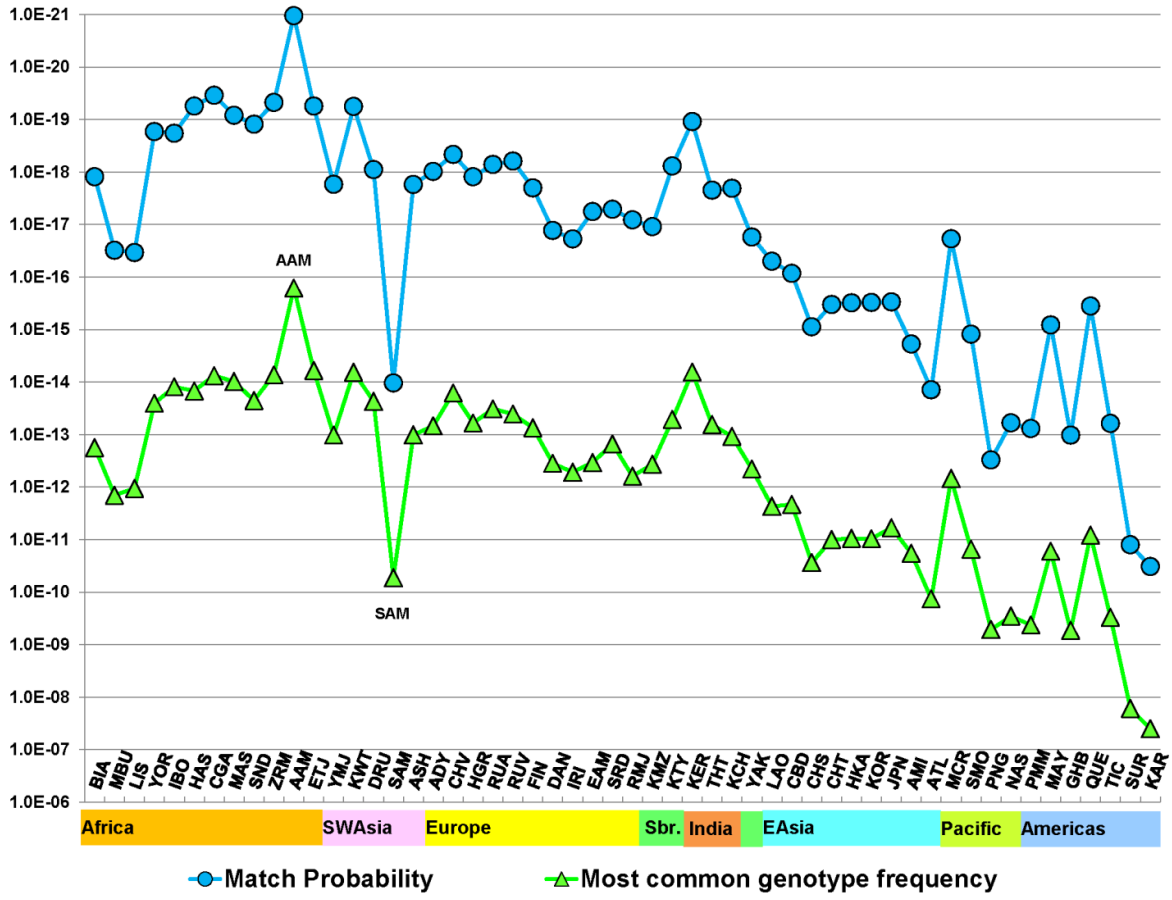


Figure 8. STRUCTURE 2.3.4 results based on 25 unlinked minihaps and 54 populations for K=7 clusters (20 iterations each).

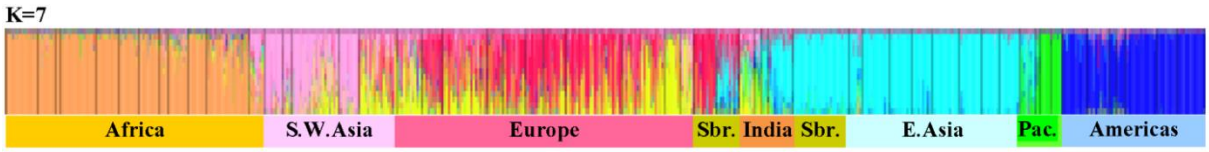
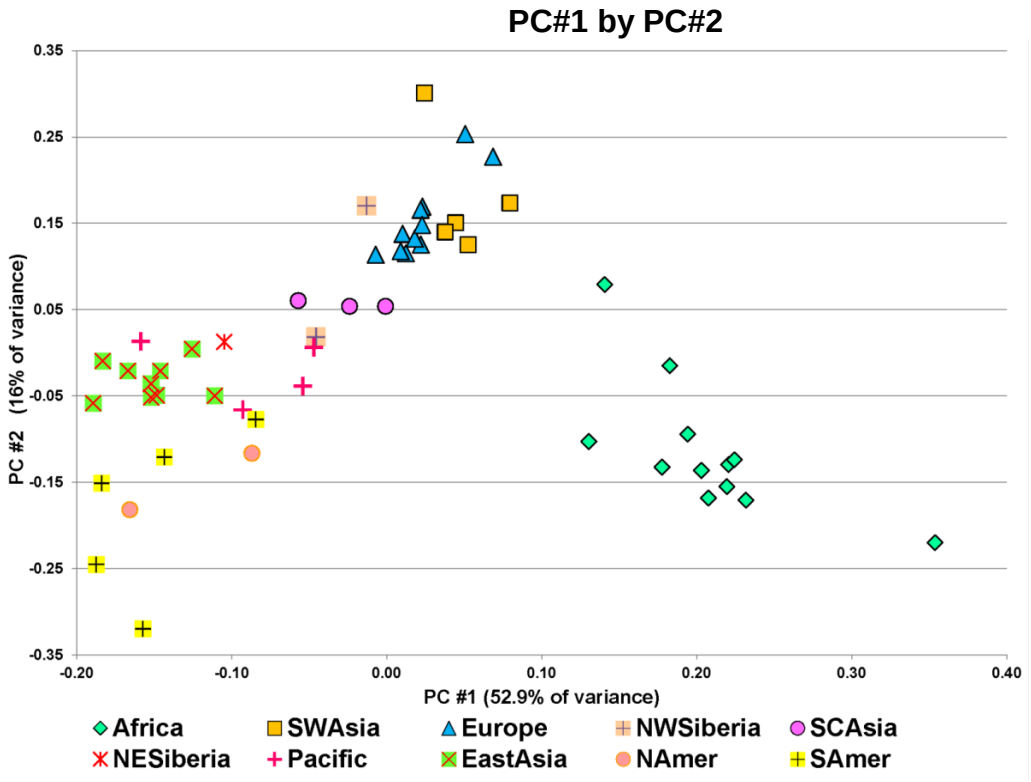


Figure 9. Principal components scatterplots. Analyses are based on a TAU genetic distance matrix for 25 "unlinked" minihaps studied on 54 population samples.



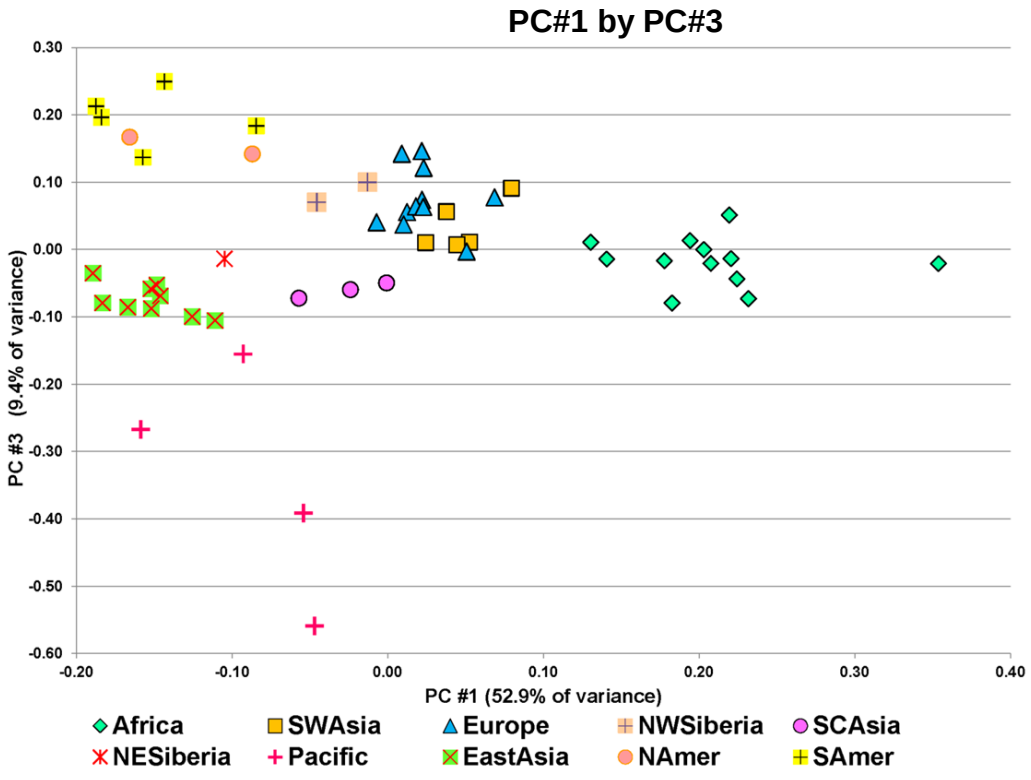
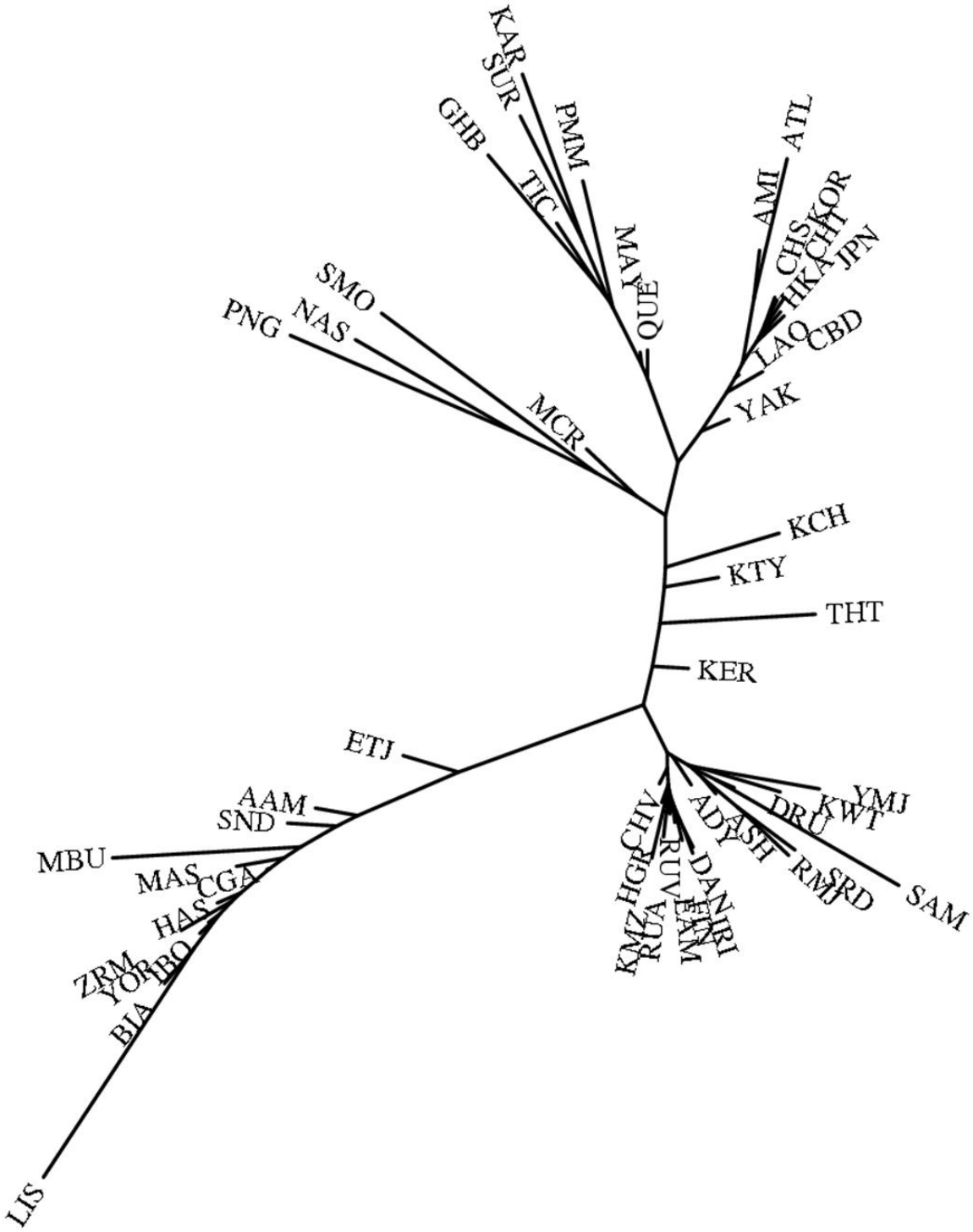


Figure 10. Best least squares tree for 25 “unlinked” minihaps studied on 54 population samples. Population abbreviations are explained in Table 7.



Implications for policy, practice, future research. This panel of 25 unlinked minihaps has valuable features that are useful for lineage identification and commend it as a documented research tool at least for the populations studied. The most notable features include high levels of polymorphism (like STRPs but based on SNPs with their low mutation rates) and high levels of resolvability for minihap genotypes that make haplotypes the equivalent of alleles identical by descent. These 25 unlinked minihaps already provide a substantial range of sensitivity for genetic relatedness for comparisons of individuals within and between a worldwide sampling of populations including some populations with relatively high levels of inbreeding (such as the three Amerindian groups from the Amazon basin). Certainly empirical studies demonstrating the ability of the panel to distinguish various degrees of close biological relationship (e.g. full and half siblings, cousins, etc.) should be counted among the additional evidence needed to substantiate the value of a minihap panel. The current panel of 25 unlinked minihaps might also find some immediate limited applications in actual forensic work especially when degradation of biological samples or other conditions do not allow the use of standard STRPs.

Ideally, before achieving status as a "final" minihap panel ready for routine applications it would be desirable to have other research groups replicate the results reported here on additional large samples from the sample populations as well as validate results on new populations. Since the 54 populations studied already cover much of the world and many of the as yet unstudied populations share similar demographic histories, it is reasonable to expect that most new populations studied will also be found to have excellent heterozygosities and genotype resolvabilities. In order to make the panel more generally useful it might also be desirable to find some additional unlinked minihaps that might have enhanced heterozygosities for Native American and Pacific Island populations. Fine-tuning the panel might also be desirable by replacing some of the loci in the current panel with loci that are found to have better average heterozygosities worldwide and also in particular geographical regions.

Carefully selected and documented SNP panels such as those described in this project and our earlier publications (Pakstis et al., 2010; 2012) have the potential to become major forensic tools because of their statistical power and low cost. In the

future the availability of inexpensive next generation sequencing (NGS) and complementary metal oxide semiconductor (CMOS) based methods (see reviews of Jones et al., 2011; Nielsen et al., 2011) for detecting SNPs will make carefully selected SNP-based panels an increasingly attractive alternative to STRPs in forensic applications such as individual identification, lineage inference, ancestry ascertainment, and phenotype inference. SNP panels can provide greater informativeness and accuracy than the current CODIS panels for all forensic applications. Incorporating well characterized SNP panels into national databases would help foster the acceptance of SNP-based tools in the courts.

A major aim of this project was to accumulate sufficient evidence to validate the feasibility and utility of minihaps for forensic work especially for distinguishing familial lineages. The 25 unlinked minihaps have high levels of heterozygosity in the 54 population samples from around the world that we have studied. The markers also have high levels of genotype resolvability. These minihaps have the evolutionary stability that allows haplotypes to be equated with alleles basically identical by descent in broader studies. Together, these aspects of the panel provide substantial support for the validity of this approach. The match probabilities achieved by the panel of 25 unlinked minihaps are already comparable to or better than the best CODIS STRPs and they compare favorably to the panel of 45 unlinked IISNPs that we reported in an earlier study at least for all the large major populations studied, including those routinely encountered in forensic labs in the U.S. and Europe. A bonus feature of the panel is that it also demonstrates distinct patterns of minihap frequencies for populations deriving from the major geographical regions of the world thereby helping when forensic applications deal with ancestry inference.

Table 6. The 19 additional minihaplotypes studied which did not qualify for inclusion in the panel of 25 "unlinked" minihaps noted in Table 4.

Chr #	SNP s1 position GrCh37.p5 build 37.3 nt position	bp to next minihap SNP on same chr (s1 to s1 pos)	Closest Gene SYMBOL	Molecular Extent bp	# SNPs	Ordered SNPs by dbSNP rs-number s1	s2	s3	s4
1	24,408,414	51,659,865	MYOM3	5,118	2	rs4313343	rs6678938	n/a	n/a
1	76,068,279	138,211	SLC44A5	8,112	3	rs6662665	rs211695	rs6703265	n/a
1	76,206,490	92,905,379	ACADM	6,535	3	rs12744608	rs17647178	rs1146581	n/a
1	169,111,869		NME7	2,599	4	rs17349222	rs12728466	rs12084964	rs1320964
2	101,587,455	34,895,900	NPAS2	6,737	3	rs1562313	rs2305160	rs11541353	n/a
2	136,483,355		R3HDM	5,624	3	rs12475553	rs724326	rs6739713	n/a
3	113,842,013	UNLINKED	DRD3	5,096	3	rs2087017	rs2399496	rs9817063	n/a
4	56,319,244	43,923,765	CLOCK	3,544	3	rs6855837	rs11240	rs1464490	n/a
4	100,243,009		ADH1B	3,929	4	rs1159918	rs1229982	rs3811801	rs9307239
7	96,744,453	UNLINKED	ACN9	4,560	3	rs10269566	rs6973504	rs7794886	n/a
9	88,084,730	UNLINKED	STK33P	1,956	3	rs11140984	rs17426617	rs6559867	n/a
10	96,609,568	10,240,212	CYP2C19	9,458	3	rs4917623	rs12268020	rs3862009	n/a
10	106,849,780		SORCS3	10,558	3	rs703472	rs1452269	rs791105	n/a
11	5,244,404	56,352,808	HBB-HBD	7,446	3	rs10837628	rs1609812	rs7944544	n/a
11	61,597,212		FADS2 prox	8,004	3	rs174570	rs1535	rs2072114	n/a
16	89,982,272	UNLINKED	MC1R	3,883	3	rs3212345	rs3212363	rs885479	n/a
17	28,549,898	10,296,956	SLC6A4	2,876	3	rs6354	rs2066713	rs8071667	n/a
17	38,846,854		KRT24	5,116	3	rs2109223	rs4890120	rs2429548	n/a
19	46,327,933	UNLINKED	SYMPK	7,023	3	rs10500292	rs8102876	rs4803866	n/a

Table 7. Abbreviations (3-characters) employed in various figures to indicate the 54 populations. The ethnic groups are organized by geographical region.

Region	ABR	Population		Region	ABR	Population
Africa	BIA	Biaka		Siberia	KMZ	Komi Zyrian
	MBU	Mbuti			KTY	Khanty
	LIS	Lisongo			YAK	Yakut
	YOR	Yoruba		SC Asia	KER	Keralites
	IBO	Ibo			THT	Thoti
	HAS	Hausa			KCH	Kachari
	CGA	Chagga		E Asia	CHS	Chinese, San Francisco
	MAS	Masai			CHT	Chinese, Taiwan
	SND	Sandawe			HKA	Hakka
	ZRM	Zaramo			KOR	Koreans
	AAM	Afr Americans			JPN	Japanese
	ETJ	Ethiopian Jews			LAO	Laotians
SW Asia	YMJ	Yemenite Jews			CBD	Cambodians
	KWT	Kuwaiti			AMI	Ami
	DRU	Druze			ATL	Atayal
	SAM	Samaritans		Pacific Is.	SMO	Samoans
	ASH	Ashkenazi			MCR	Micronesians
Europe	ADY	Adygei			PNG	Papuans, New Guinea
	CHV	Chuvash			NAS	Nasioi
	HGR	Hungarians		Americas	PMM	Pima, Mexico
	RUA	Russians, Archangelsk			MAY	Maya, Yucatan
	RUV	Russians, Vologda			GHB	Guihiba, Colombia
	FIN	Finns			QUE	Quechua, Peru
	DAN	Danes			TIC	Ticuna
	IRI	Irish			SUR	Rondonian Surui
	EAM	Euro Americans			KAR	Karitiana
	SRD	Sardinians				
	RMJ	Roman Jews				

Microhaplotypes

Key Idea. Modern sequencing technology makes it possible to genotype polymorphisms with high throughput and high multiplexing. We searched for and identified many loci with 2 or more SNPs within the limited expanse of a 200 base pair single sequence run and demonstrated that when linkage disequilibrium is moderate to strong but not complete, multi-SNP haplotype loci can be identified that have multiple haplotype alleles at common frequencies which are detected as phase-known haplotypes. These micro-haplotype loci (microhaps) are a powerful tool for individual identification, ancestry inference, and determining family/clan relationships.

Background and rationale. Multiallelic markers such as the short tandem repeat polymorphisms (STRPs) in CODIS have much greater information content than a similar number of diallelic markers such as SNPs. Yet SNPs offer a much lower mutation rate than STRPs and some methods available for genotyping are much cheaper, faster, and have lower error rates. SNP genotyping is also achievable with much shorter DNA fragments, comparable to those that occur in degraded forensic samples.

In previous work (Pakstis et al., 2012), we provided evidence for the utility of selected minihaplotypes, consisting of two or more SNPs within a molecular distance of less than 10 kb, in providing quite accurate ancestry and lineage information. At 10 kb distance, the recombination rate is on the order of the mutation rate for SNPs and it is considerably lower than the mutation rates for the standard forensic STRPs.
The new sequencing technologies are now making a new type of forensic marker possible: microhaplotypes based on two or more SNPs within a molecular distance less than the read length of the desktop machines like the Personal Genome Machine (PGM) from Life Technologies. A panel of carefully selected and characterized microhaplotype loci (microhaps) with multiple SNPs (1) in close proximity (<200 bp), (2) with high global heterozygosity, and (3) with high regional or global Fst should be able to provide information on global ancestry and on lineage-familial relationships.

Note that while it is not a large problem statistically, if the SNPs defining a microhap are assayed separately rather than sequenced so that all the results are phase known, then the phase on each chromosome must be estimated statistically for some individuals. A large number of individuals sampled from any given population will

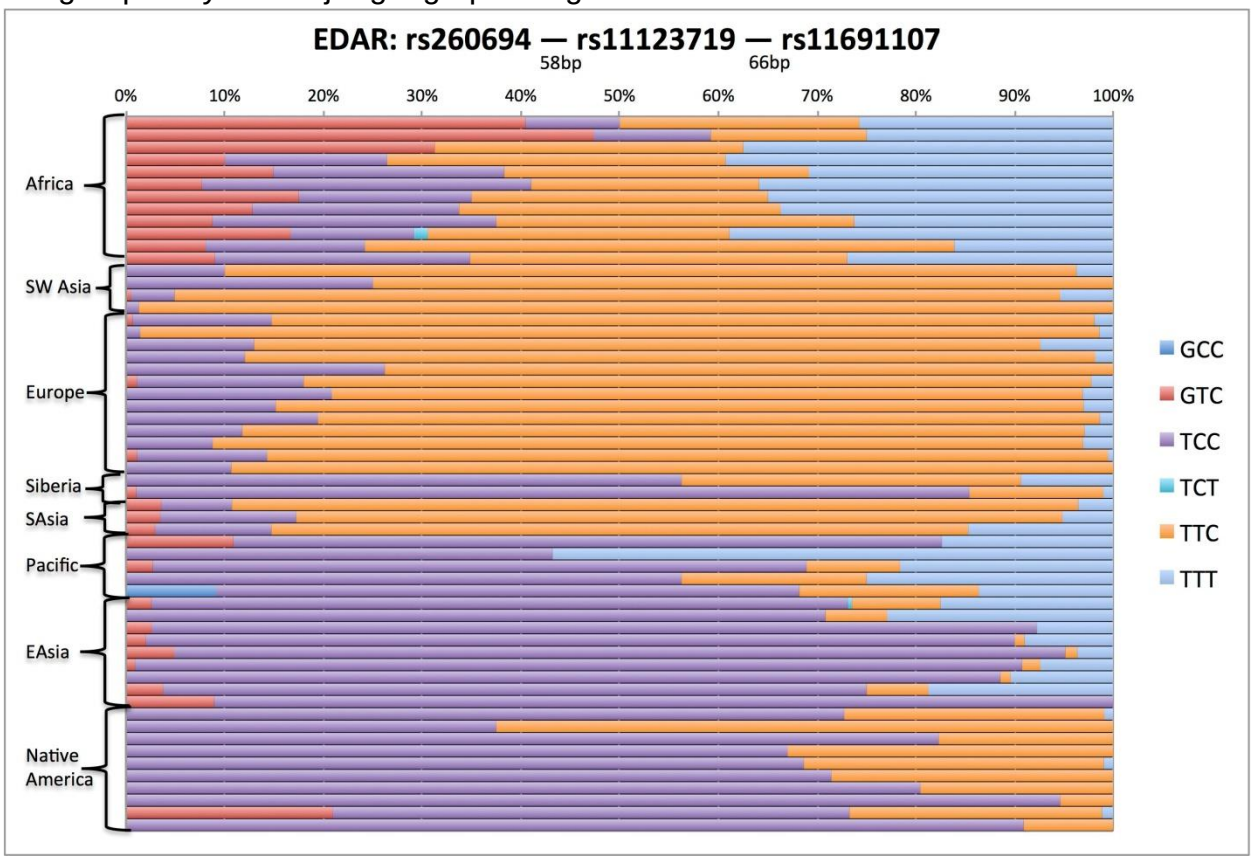
have unambiguous phase because the allelic arrangement on each chromosome is clearly known. This automatic knowledge of phase arises when all SNPs defining the microhap are either homozygous or else no more than one SNP is heterozygous. These phase known individuals also help make the estimation of the two chromosomes in the remaining individuals very accurate when haplotype inference is necessary.

Searching for candidate microhaps. We used pilot data from our existing datasets to choose the best criteria for identifying novel microhaplotype loci. We then used existing large SNP genotype datasets comprising many diverse populations that are accessible on the worldwide web (primarily the Human Genome Diversity Panel (HGDP) and the 1000 Genomes project) to identify 52 candidate SNP sets, and PHASEd these genotypes to determine global microhap frequencies at each locus. The distribution of the frequencies for these haplotypes globally as well as the microhap's heterozygosity within each population were used to determine the utility of each microhap locus for a forensic panel. Microhaps were sought that have high Fst values either globally or else between particular geographical regions of the world. The screening process also aimed for relatively low correlations of SNP allele frequencies within a microhap to avoid situations with strong linkage disequilibrium. Within the constraints allowed by maximizing Fst values we also aimed to maximize the average heterozygosity of the microhaps across the populations studied. For those candidate loci that looked most promising, we employed TaqMan assays to genotype the individuals in our 54 populations. We then determined phase for each individuals using software such as PHASE.

Candidate microhaps. As of the end of the current project we have a set of 28 microhaps (Table 8), each defined by 2 to 3 SNPs, which have been studied comparably on 54 of our population samples. The definitions, sample sizes, and frequencies for the first 20 of the 28 microhaps completed on our 54 populations have been entered into ALFRED (http://alfred.med.yale.edu) and can be retrieved using the keyword *microhap*. Figure 11 presents a detailed view of a 3-SNP microhap spanning 124 bp that has been identified at the EDAR gene on chromosome 1. The figure clearly shows multiple haplotypes at common frequencies across the populations studied. The

very large frequency differences between haplotypes make this microhap very informative for ancestry.

Figure 11. Haplotype frequencies for a microhap defined by 3-SNPs at the EDAR gene. In this stacked-bar graph each different haplotype allele is a different color and the extent of the color bar is proportional to the frequency of the haplotype. Populations are grouped by the major geographic regions of the world.



For the 28 microhaps fully genotyped and PHASEd on our 54 populations (2612 individuals) the global average heterozygosity is 0.543, with 21 of the loci above the single-SNP maximum of 0.5 (Figure 12). Because of this high heterozygosity (also known as informativeness), the random match probabilities range from 2.93×10^{-13} to 1.03×10^{-18}; a visual illustration of the match probabilities can be seen in Figure 13. These results are very comparable to our published panel of 45 unlinked individual identification SNPs (Kidd et al., 2012). Figure 14 displays the best population tree generated from this 28 microhap dataset employing the least squares estimation

procedure. The major continental regions clearly differentiate in the population tree. The principal components analysis results in Figure 15 are another way of visualizing this.

These results strongly support the potential utility of microhaps as a general forensic tool for both individual identification and ancestry inference. However, the original motivation was to develop multi-allelic SNP-based markers to help in identifying family or clan relationships. With the increased heterozygosity and multiple alleles (haplotypes) these loci have, they can approach the information content of the standard forensic short tandem repeat polymorphisms (STRPs). The ability to highly multiplex multiple such microhaps argues that these have the potential to surpass the "familial relationship" function of the STRPs.

Microhaps have an additional capability of interest for forensic work in the identification of mixtures and with the additional potential to quantify the components, i.e., to disentangle mixtures in a quantitative way. The presence of three or more different sequences at sufficient numbers of reads becomes clear evidence of DNA from more than one person contributing to the sample. With many loci multiplexed and with more loci with four or more haplotypes, the microhaps become powerful markers to identify and quantify components of mixtures. Microhaps with 3 or more common alleles (haplotypes) for one sequence of <200bp will allow computer software to accurately predict the likelihood and levels of mixture based on observing more than two sequence types at a locus and the numbers of occurrences of each type.

We will deposit in ALFRED the remaining 8 microhap definitions and frequency data on the 54 populations by the end of 2013. Work is underway to identify primers that will optimally amplify the entire region of each micro-haplotype locus for sequencing on the PGM. The work and expense involved in genotyping a micro-haplotype locus by sequencing is no greater than genotyping a single SNP by sequencing, but the yield in information is considerably greater.

***CONCLUSION*.** In this project we have generated evidence supporting the value of micro-haplotype loci as a powerful new type of forensic marker made possible by modern sequencing technology. Because our emphasis in this project switched to microhaps only after the project had been underway for some time we did not have the

funding or the time to assemble as good a panel of microhaps as we believe are possible. We did the best we could with resources available especially by focusing on assembling microhap candidates where we already had the component SNPs typed already on many of our core populations. (We think that the panel of minihaps that we did assemble before focusing on microhaps is an indicator of what could be accomplished with even better results for microhaps.) As additional microhaps with even more attractive characteristics are identified than the preliminary set of 28 that we have presented here and the necessary population databases are eventually assembled, it should be possible to identify a fine-tuned panel of microhaps that will outperform the existing SNP panels for individual identification and ancestry inference while also providing family/clan/lineage information and the ability to detect and quantify mixtures. The microhaps identified in the screen of the HGDP data clearly show that good ancestry inference will be possible (Figure 16). Save for the offender databases, the potential for completely supplanting STRPs clearly exists.

Table 8. List of 28 microhaplotypes.

Micro Hap Count	SNP Count		dbSNP rs-number	Nearby gene or label	Chr	Nt position Build 37	MicroHap extent in bp
1	1		rs260694	EDAR	2	109,586,313	125
	2		rs11123719	EDAR	2	109,586,371	
	3		rs11691107	EDAR	2	109,586,437	
2	4		rs2170607	LYPD6B	2	149,954,097	157
	5		rs10497052	LYPD6B	2	149,954,253	
3	6		rs4513489	CCR2	3	46,352,355	30
	7		rs6441961	CCR2	3	46,352,384	
4	8		rs3732783	DRD3	3	113,890,789	27
	9		rs6280	DRD3	3	113,890,815	
5	10		rs6808142	LRRC2	3	46,556,835	114
	11		rs17030627	LRRC2	3	46,556,948	
6	12		rs2584457	ADH7	4	100,304,166	83
	13		rs12648443	ADH7	4	100,304,248	
7	14		rs2851017	ADH7	4	100,313,996	67
	15		rs2032350	ADH7	4	100,314,062	
8	16		rs4699748	ADH7	4	100,321,443	153
	17		rs2584461	ADH7	4	100,321,573	
	18		rs1442492	ADH7	4	100,321,595	
9	19		rs1280100	FAT1	4	187,538,133	198
	20		rs1280099	FAT1	4	187,538,330	
10	21		rs870347	PAPD7	5	6,845,017	19
	22		rs870348	PAPD7	5	6,845,035	
11	23		rs41461	TAS2R1	5	9,619,905	32
	24		rs41462	TAS2R1	5	9,619,936	
12	25		rs17168174	ACN9	7	96,733,972	86
	26		rs10246622	ACN9	7	96,734,057	
13	27		rs1390950	GATA4	8	11,595,829	141
	28		rs2898295	GATA4	8	11,595,969	
14	29		rs1408800	TYRP1	9	12,672,275	46
	30		rs1408801	TYRP1	9	12,672,320	
15	31		rs3118582	RXRA	9	137,417,115	194
	32		rs10776839	RXRA	9	137,417,308	
16	33		rs10500616	OR52S1P1	11	5,109,946	123
	34		rs2499936	OR52S1P1	11	5,110,068	
17	35		rs2303377	NCAM1	11	113,111,501	165
	36		rs2303378	NCAM1	11	113,111,665	
18	37		rs4938013	ANKK1	11	113,264,470	42
	38		rs11214596	ANKK1	11	113,264,511	
19	39		rs6277	DRD2	11	113,283,459	19
	40		rs6275	DRD2	11	113,283,477	
20	41		rs1079727	DRD2	11	113,289,182	117
	42		rs2002453	DRD2	11	113,289,298	
21	43		rs2133298	PAH	12	103,260,634	186
	44		rs3817446	PAH	12	103,260,819	
22	45		rs1503767	SUDS3	12	118,889,488	72
	46		rs11068953	SUDS3	12	118,889,559	
23	47		rs12717560	C14ORF43	14	74,250,557	159
	48		rs12878166	C14ORF43	14	74,250,715	
24	49		rs1059504	ARHGAP27	17	43,472,321	187
	50		rs8327	ARHGAP27	17	43,472,507	
25	51		rs2233362	ABI3	17	47,287,067	43
	52		rs634370	ABI3	17	47,287,109	
26	53		rs11868709	TMEM235	17	73,740,166	60
	54		rs9907137	TMEM235	17	73,740,225	
27	55		rs1055919	PLIN3	19	4,852,137	201
	56		rs2271057	PLIN3	19	4,852,337	
28	57		rs10854214	intergenic	20	59,703,918	97
	58		rs10854215	intergenic	20	59,704,014	

Figure 12. **Average heterozygosity across 54 populations and Fst values for the set of 28 microhaps**

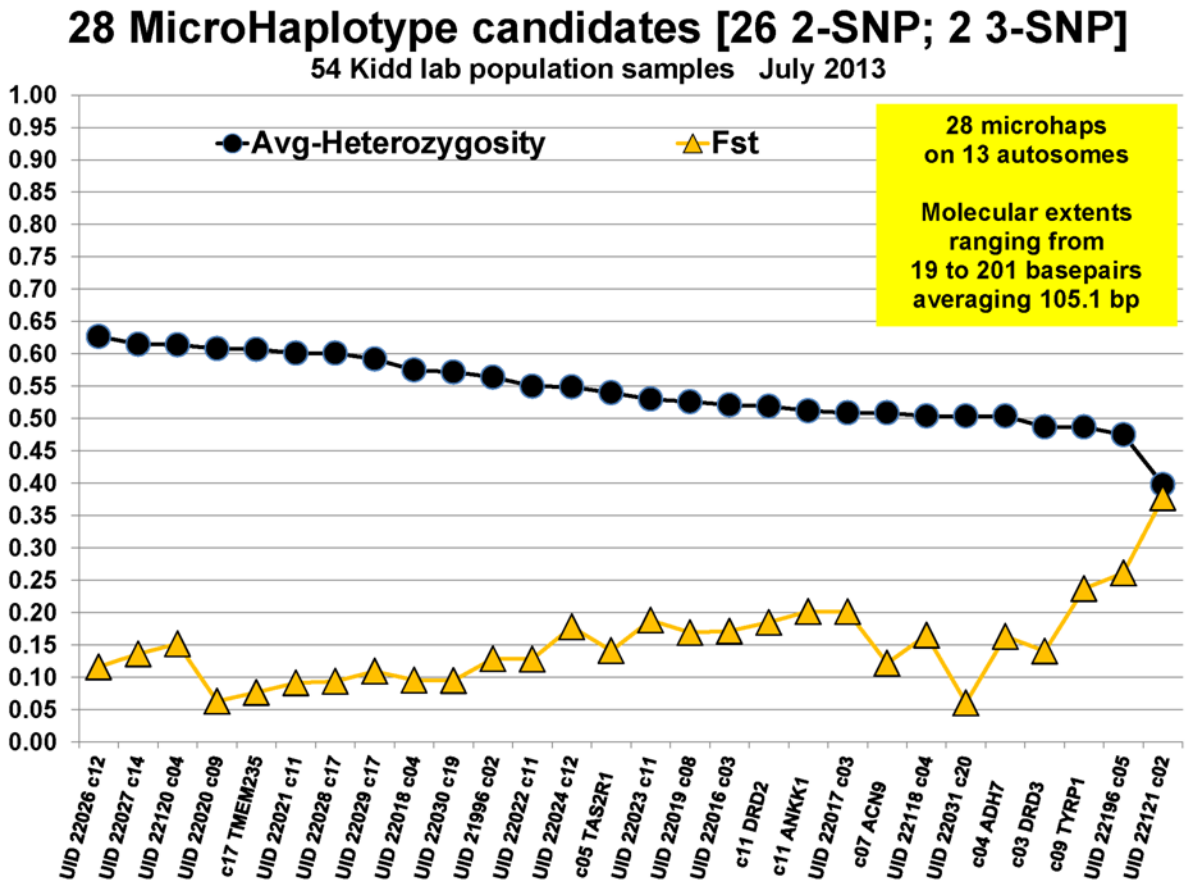


Figure 13. **Match probabilities for 54 populations employing 28 microhaps.**

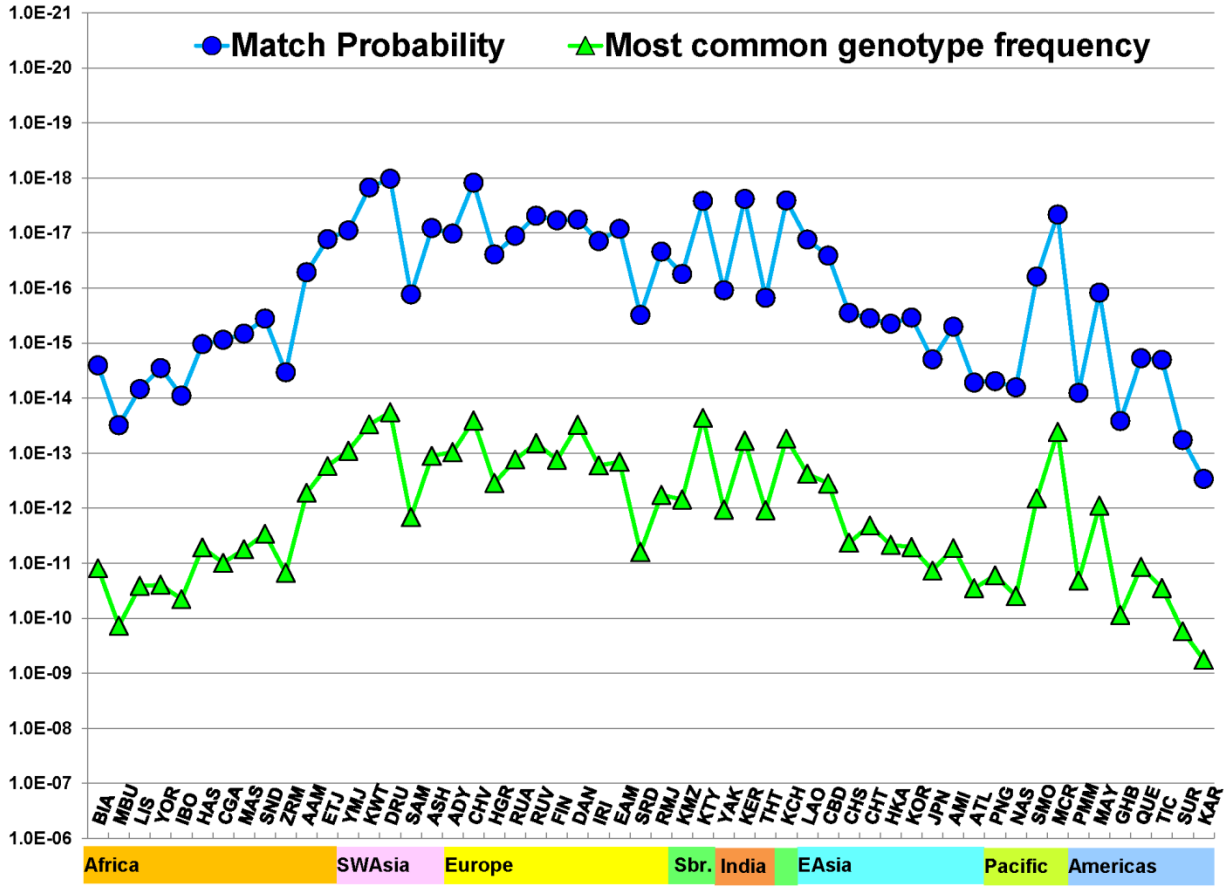


Figure 14. Best least squares population tree for 54 populations based on 28 microhaps. See Table 7 for the key to the population abbreviations.

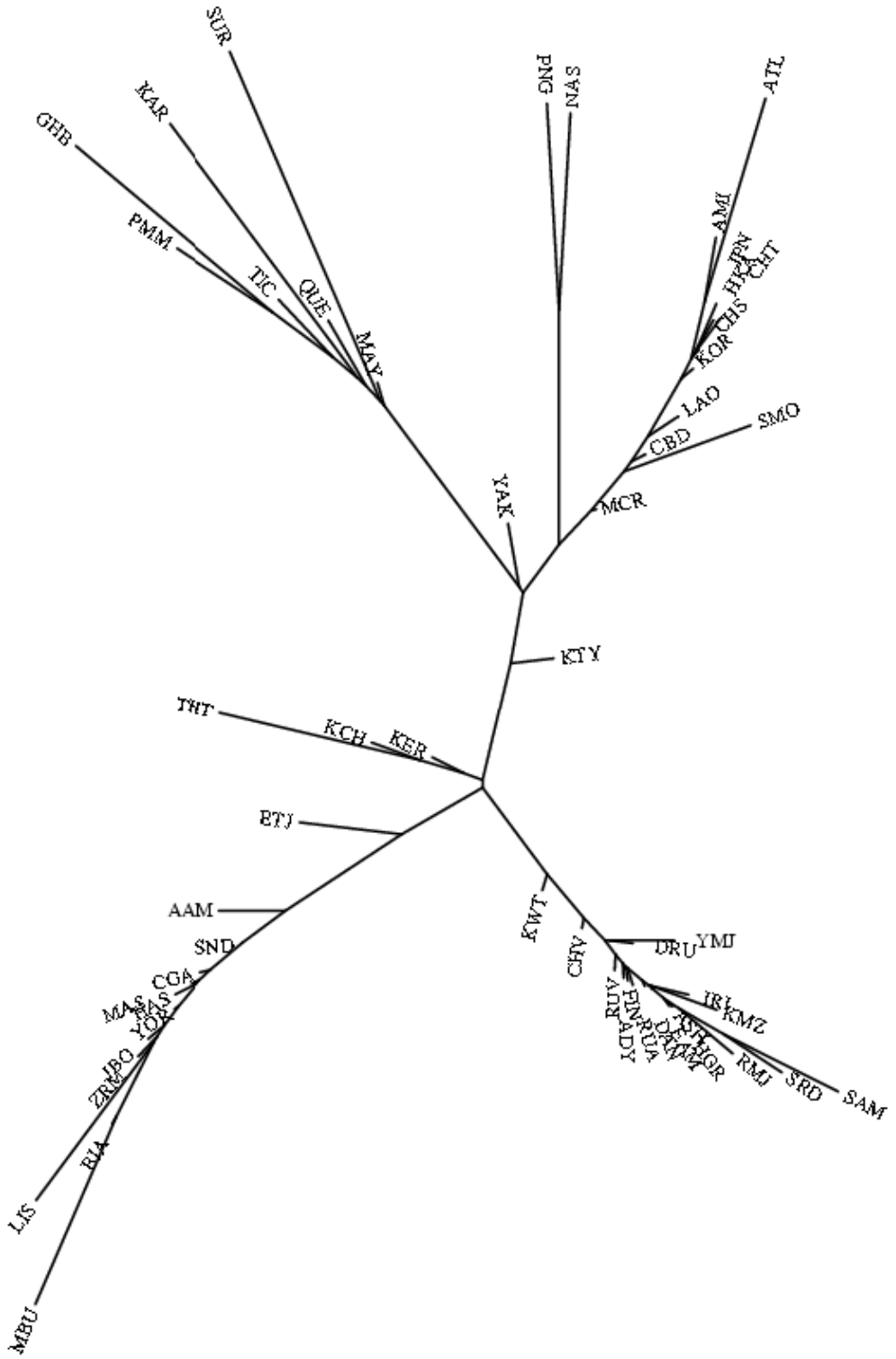


Figure 15. Principal components analysis based on TAU genetic distances for 54 populations studied on 28 microhaps. Table 7 explains population abbreviations.

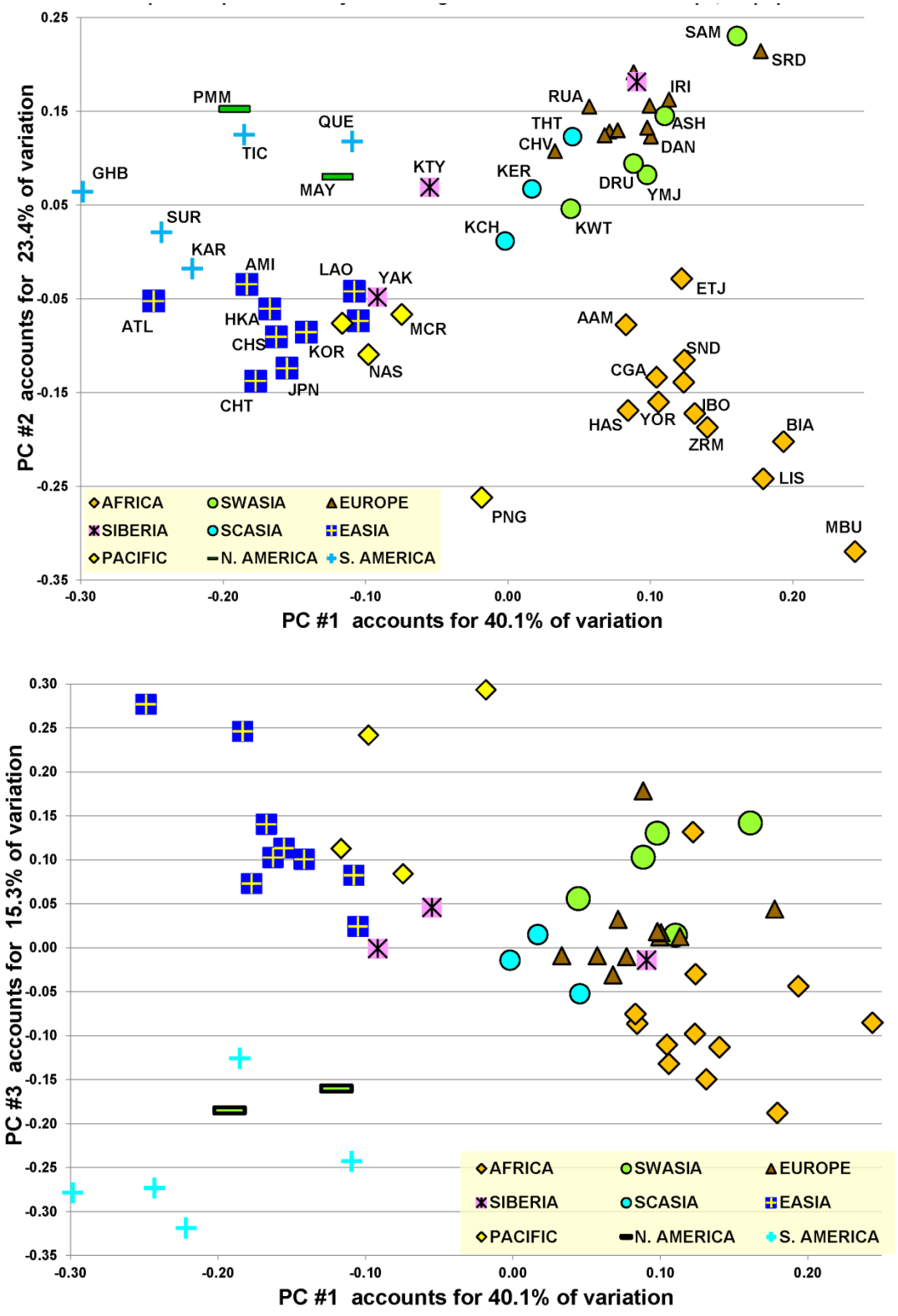
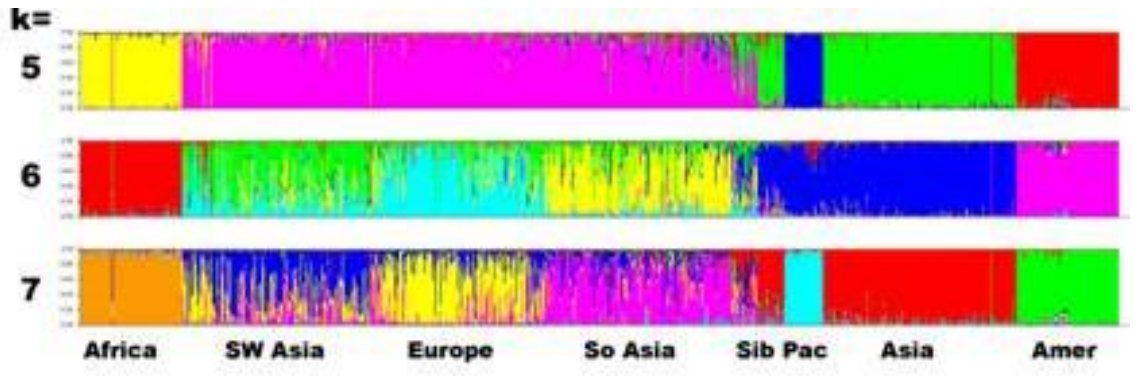


Figure 16. STRUCTURE analysis of 48 microhaps for the HGDP SNP data (Li et al., 2008). These loci are selected for high heterozygosity but not for high Fst; nonetheless good differentiation of the populations is seen across most regions of the world.

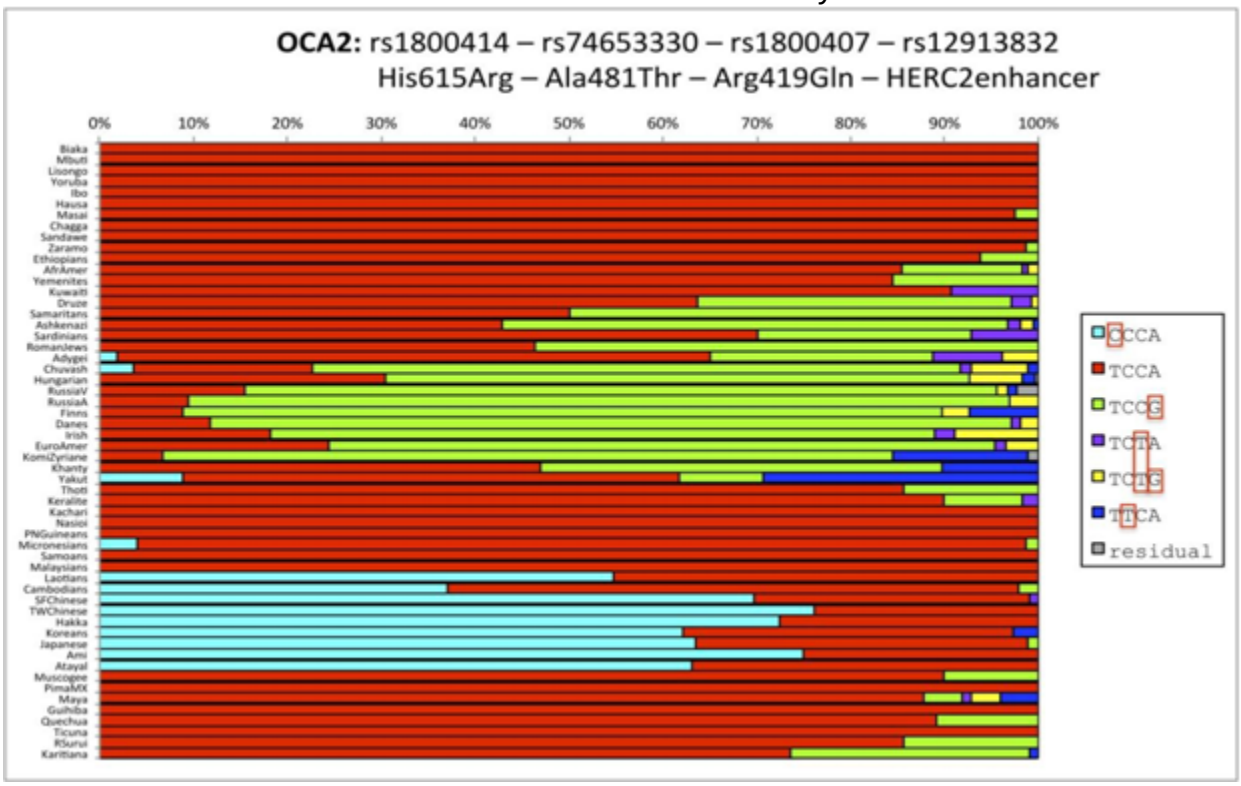


Progress on identifying Phenotype Informative SNPS (PISNPs)

Our population resources (Tables 1a, 1b) are valuable for gaining a better understanding of how SNPs that are reported to be associated with phenotypes actually vary among human populations. In Donnelly et al. (2012) we were able to show that many of the SNPs in the OCA2 gene (chr 15q12) that had been associated with eye color in Western Europe were in fact simply in linkage disequilibrium with the SNP in the upstream enhancer that Visser et al. (2012) showed to be the functionally relevant variation. For several of those non-coding SNPs in OCA2 the "blue eye" allele also occurred in other parts of the world at frequencies that would have predicted a noticeable frequency of blue eyes--if the allele caused blue eyes--when the population in fact has essentially no blue-eyed individuals.

We studied four of the OCA2 SNPs that have clear functional possibilities and that have an appreciable heterozygosity. Figure 17 shows the haplotype frequencies based upon those four SNPs. The haplotype defined by the enhancer variant (green) is clearly present only in populations in or near Europe with the possibility of European admixture in some other population samples. This is the haplotype actually functionally relevant to the "blue eye color" phenotype. The haplotype defined by the functional SNP rs1800414 (His615Arg) (light blue in the figure) is restricted to East Asia with some evidence of gene flow into Southeastern Europe. This SNP and haplotype are associated with lighter skin pigmentation among East Asians (Edwards et al.,2010). What our analyses show is that another non-synonymous substitution, rs74653330 (Ala481Thr) defines another haplotype reaching 10% to 30% in Siberia and the Finns; as yet there is no phenotype association known to be associated with this haplotype. Finally, we note that rs1800407 (Arg419Gln) occurs on haplotypes with the enhancer variant AND occurs on chromosomes with the ancestral allele at the enhancer site. Though uncommon, the cis-trans possibility may raise complexities in eye color prediction that is not currently considered in the formulae used to predict eye color for the IrisPlex SNP set.

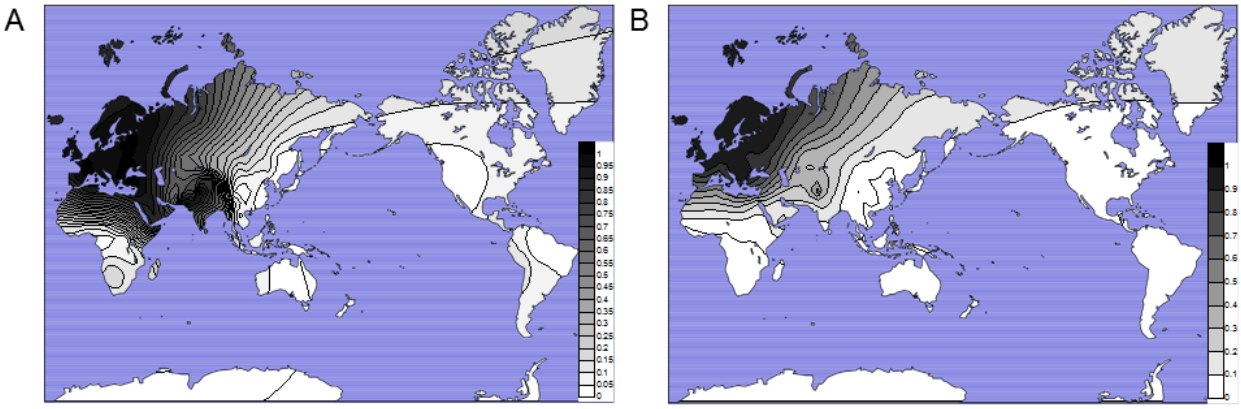
Figure 17. Bar graph of OCA2 haplotype frequencies based on four functional SNPs. rs12913832 is the enhancer SNP most relevant for blue eye color.



We also carried out a study of population genetic variation for two skin color loci: SLC24A5 (chr 15q21.1) and SLC45A2 (chr 5p13.3) in which we also examined the evidence for selection where there was sufficient variation to carry out the tests. A manuscript describing the details of these studies is in the process of being revised in light of the suggestions of the reviewers. We focused on two SNPs, rs1426654 (Ala111Thr) in SLC24A5 and rs16891982 (Leu374Phe) in SLC45A2, previously reported in the literature to be important contributors to variation in skin color. We typed a total of 47 SNPs (27 SNPs for SLC24A5; 20 SNPs for SLC45A2) in over 3600 individuals from 74 of our populations where we have either cell lines or DNA from various collaborating labs. We also incorporated into some analyses some overlapping published data on an additional 33 populations. The two SNPs previously reported to have important roles in skin color variation have rather different distributions in our sampling of populations from around the world. Figure 18 from the manuscript shows the global distribution of the allele frequencies for lighter skin pigmentation at these SNPs. Our results support the role for selection in different world regions at rs1426654

but not for rs16891982; our evidence points toward a different SNP in the SLC45A2 region. The combination of the genetic variation patterns we report and the differences we found in evidence for selection in different world regions suggest that in humans skin pigmentation evolved separately among the various populations but used the same genes to do so.

Figure 18. Worldwide allele frequency distributions for the lighter skin pigment alleles at SLC24A5 (rs1426654 “A”) and SLC45A2 (rs16891982 “G”).



Conclusions

Carefully crafted IISNP, AISNP, LISNP, and eventually PISNP panels offer great untapped advantages for forensic applications. Our work on this project has resulted in a panel of 55 Ancestry Informative SNPs, a panel of 25 “unlinked” minihaps, and a preliminary set of 28 microhaps based on SNPs which provide the evidence supporting this view. We have also improved global coverage of a previously developed panel of individual identification SNPs. We have provided global population data on several alleles that are strongly associated with various phenotypes, eye, skin, and hair color. Much more work remains to be done in developing the best SNP-based panels, especially those tailored for use in ancestry inference as well as lineage studies and eventually for phenotype studies as the more complete underlying genetic factors are identified and validated. This project underscores that continued work on SNP panels offers much more than a theoretical possibility of better SNPs for better forensics..

Dissemination of Research Findings

Results of this project have been and are currently being made available in several ways. First, results have been published in and are being submitted to the forensic/scientific literature. See the detailed list at the end of this section for ten published papers, two manuscripts under review, three manuscripts currently in preparation, and nine other publications from lab collaborations benefiting from this project.

We also have contributed allele and haplotype frequencies on the population samples studied to the freely accessible ALFRED database (http://alfred.med.yale.edu) on the worldwide web. Typically, our policy is that raw data for allele/haplotype frequencies and population sample sizes have been deposited in and made public through ALFRED as soon as possible after the data has been checked for correctness and the results of analyses are being assembled in manuscripts for publication in peer reviewed journals. In addition, special SNP marker sets of interest to forensic investigators have been made accessible in both the ALFRED and FROGkb (http://frog.med.yale.edu) databases.

Slide and poster presentations on the forensic applications of SNPs and various interim results from this project have been made at scientific meetings and during visits with forensic researchers at a number of universities. Copies of posters containing various results from this project have also been placed online (http://medicine.yale.edu/labs/kidd/www/contents.html) at the Kidd lab website as pdf files. We have also circulated our unpublished results by making available on our web site the lists of markers in various provisional SNP panels prior to publication. We have notified the NIJ and some individual members of the forensic community concerning the availability of the materials related to this project on our laboratory web site. Some members of the forensic science and related research communities have become aware of our IISNP and AISNP work based on the poster presentations we have made at the NIJ annual meetings as well as at the ISFG annual meetings. A slide presentation was made by K.K. Kidd on September 4th 2013 at the International Society of Forensic Genetics (ISFG) annual meeting held in Melbourne, Australia. K.K. Kidd has also made

Published papers supported by and citing NIJ 2010-DN-BX-K225

Joshua N. Sampson, Kenneth K. Kidd, Judith R. Kidd, Hongyu Zhao, **2011**. Selecting SNPs to identify ancestry. *Annals of Human Genetics* 75:539-553.

Kenneth K. Kidd, William C. Speed, Andrew J. Pakstis, Judith R. Kidd, **2011**. The search for better markers for forensic ancestry inference. *22nd International Symposium on Human Identification.* Washington D.C. October 3-6, 2011. Sponsored by Promega Corporation. Published at the sponsor's website, February, 2012.

Kenneth K. Kidd, Judith R. Kidd, William C. Speed, Rixun Fang, Manohar R. Furtado, Fiona C. Hyland, Andrew J. Pakstis, **2012**. Expanding data and resources for forensic use of SNPs in individual identification. *Forensic Science International: Genetics* 6:646-652.

Andrew J. Pakstis, Rixun Fang, Manohar R. Furtado, Judith R. Kidd, Kenneth K. Kidd, **2012**. Mini-haplotypes as lineage informative SNPs (LISNPs) and ancestry inference SNPs (AISNPs). *European Journal of Human Genetics* 20:1148-1154

Michael P. Donnelly, Peristera Paschou, Elena Grigorenko, David Gurwitz, Csaba Barta, Ru-Band Lu, Olga V. Zhukova, Jong-Jin Kim, Marcello Siniscalco, Maria New, Hui Li, Sylvester L. Kajuna, Vangelis G. Manolopolulos, William C. Speed, Andrew J. Pakstis, Judith R. Kidd, Kenneth K. Kidd, **2012**. A global view of the OCA2-HERC2 region and pigmentation. *Human Genetics* 131:683-696.

Caroline M. Nievergelt, Adam X. Maihofer, Tatyana Shekhtman, Ondrej Libiger, Xudong Wang, Kenneth K. Kidd, Judith R. Kidd, **2013**. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investigative Genetics* 4:13.

Christopher Heffelfinger, Andrew J. Pakstis, William C. Speed, Allison P. Clark, Eva Haigh, Rixun Fang, Manohar R. Furtado, Kenneth K. Kidd, Michael P. Snyder, **2014**. Positive selection and haplotype structure at TLR1. *European Journal of Human Genetics* 22:551-557. In Press July 24, 2013; e-pub Sept 4, 2013.

K.K. Kidd, A.J. Pakstis, W.C. Speed, R. Lagace, J. Chang, S. Wootton, N. Ihuegbu, **2013**. Microhaplotypes are a powerful new type of forensic marker. *Forensic Science International: Genetics Supplement Series* 4:e123-e124. In Press Oct 2, 2013; E-pub October 28, 2013. A short manuscript based on an invited talk presented on September 4, 2013 at the ISFG 2013 meeting in Melbourne Australia; it is being published "Proceedings of the ISFG 2013" issue. A copy of the submitted/"In press" version of the manuscript is in the appendix of this project report.

Libing Yun, Yan Gu, Haseena Rajeevan, Kenneth K. Kidd, **2013**. Application of six IrisPlex SNPs and comparison of two eye color prediction systems in diverse Eurasia populations. *International Journal of Legal Medicine,* In Press Nov 27 2013; E-pub Jan 5 2014.

Kenneth K. Kidd, William C. Speed, Andrew J. Pakstis, Manohar R. Furtado, Rixun Fang, Abeer Madbouly, Martin Maiers, Mridu Middha, Francoise R. Friedaender, Judith R. Kidd, **2014**. Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Science International: Genetics* 10:23-32; In Press Jan 7, 2014; e-pub Feb 7, 2014.

Project papers submitted to journals and under peer review:

Kenneth K. Kidd. Andrew J. Pakstis, William C. Speed, Robert Lagace, Joseph Chang, Sharon Wootton, Eva Haigh, Judith R. Kidd, **2014**. Current sequencing technology makes micro-haplotypes a powerful new type of genetic marker for forensics.
This manuscript summarizes the characteristics and utility of a preliminary panel of 28 micro-haplotypes studied on 54 human populations representing major continental regions of the world.

Libing Yun, Haseena Rajeevan, Usha Soundararajan, Kenneth K. Kidd, **2013**. An overview of global ancestry informative SNP panels in FROGkb. Letter submitted to *Investigative Genetics*, August 2013.

Michael P. Donnelly, Peristera Paschou, Elena L. Grigorenko, David Gurwitz, Csaba Barta, Ru-Band Lu, Olga V. Zhukova, Jong-Jin Kim, Marcello Siniscalco, Maria New, Hui Li, Sylvester L.B. Kajuna, Vangelis G. Manolopoulos, David Comas, William C. Speed, Andrew J. Pakstis, Judith R. Kidd, Kenneth K. Kidd, **2013**. Global distributions and selection in SLC24A5 and SLC45A2. Submitted to *Annals of Human Genetics*
Reviews received in June with provisional acceptance pending revisions that are now underway.

Project papers in preparation for submission to peer-reviewed journals:

Jane Brissenden, Judith R. Kidd, Baigalmaa Evsanaa, Ariunaa Togtokh, Andrew J. Pakstis , Françoise Friedlaender, Kenneth K. Kidd, Janet Roscoe, **2014**. Mongolians in the genetic landscape of central Asia: exploring the genetic relations among Mongolians and other world populations.

Andrew J. Pakstis, Rixun Fang, Manohar R. Furtado, Eva Haigh, Judith R. Kidd, Kenneth K. Kidd. Validation of mini-haplotypes as valuable markers for familial identification and ancestry inference, **2014**.

Other publications benefiting from this project:

Hui Li, Sheng Gu, Yi Han, Zhi Xu, Andrew J. Pakstis, Li Jin, Judith R. Kidd, Kenneth K. Kidd, **2011**. Diversification of the ADH1B gene during expansion of modern humans. *Annals of Human Genetics* 75:497-507.

Brenna M. Henn, Christopher R. Gignoux, Matthew Jobin, Julie M. Granka, J. Michael MacPherson, Jeffrey M. Kidd, Laura Rodriguez-Botigue, Sohini Ramachandran, Lawrence Hon, Abra Brisbin, Alice A. Lin, Peter Underhill, David Comas, Kenneth K. KIdd, Paul J. Norman, Peter Parham, Carlos D. Bustamante, Joanna L. Mountain, Marcus W. Feldman, **2011**. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences* 108:5154-5162.

J.R. Kidd, F. Friedlaender, A.J. Pakstis, M. Furtado, R. Fang, X. Wang, C.M. Nievergelt, K.K. Kidd, **2011**. SNPs and haplotypes in Native American populations. *American Journal of Physical Anthropology* 146:495-502.

Yan Lu, Longli Kang, Kang Hu, Chuanchao Wang, Xiaoji Sun, Feng Chen, Judith R. Kidd, Kenneth K. Kidd, Hui Li, **2012**. High diversity and no significant selection signal of human ADH1B gene in Tibet. *Investigative Genetics* 3:23; E-pub November 23, 2012.

Brendan Keating, Aruna T. Bansal, Susan Walsh, Jonathan Millman, Jonathan Newman, Kenneth Kidd, Bruce Budowle, Arthur Eisenberg, Joseph Donfack, Paolo Gasparini, Zoran Budimlija, Anjali K. Henders, Hareesh Chandrupatla, David L. Duffy, Scott D. Gordon, Pirro Hysi, Fan Liu, Sarah E. Medland, Laurence Rubin, Nicholas G. Martin, Timothy D. Spector, Manfred Kayser on behalf of the International Visible Trait Genetics (VisiGen) Consortium, 2013. First All-in-One diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness and sex with the Identitas v1 Forensic Chip. *International Journal of Legal Medicine* 127:559-572. E-pub November 13, 2012.

V. Stathias, G.R. Sotiris, I. Karagiannidis, G. Bourikas, G. Martinis, D. Papazoglou, A. Tavridou, N. Papanas, E. Maltezos, M. Theodoridis, V. Vargemezis, V.G. Manolopoulos, W.C. Speed, J.R. Kidd, K.K. Kidd, P. Drineas, P. Paschou, **2012**. Exploring genomic structure differences and similarities between the Greek and European HapMap populations: implications for association studies. *Annals of Human Genetics* 76:472-483.

Karyn M. Steinberg, Francesca Antonacci, Peter H. Sudmant, Jeffrey M. Kidd, Catarina D. Campbell, Laura Vives, Maika Malig, Laura Scheinfeldt, William Beggs, Muntaser Ibrahim, Godfrey Lema, Thomas B. Nyambo, Sabah A. Omar, Jean-Marie Bodo, Alain Froment, Michael P. Donnelly, Kenneth K. Kidd, Sarah A. Tishkoff, Evan E. Eichler, **2012**. Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature Genetics* 44:872-880; E-pub July 1, 2012.

D. Reich, N. Patterson, D. Campbell, A.Tandon, S.Mazieres, N. Ray, M.V. Parra, W. Rojas, C. Duque, N. Mesa, L.F. Garcia, O. Triana, S. Blair, A. Maestre, J.C. Dib, C.M. Bravi, G. Bailliet, D. Corach, T. Hunemeier, M.-C. Bortolini, F.M. Salzano, M.L. Petzl-Erler, V. Acuna-Alonzo, C. Aguilar-Salinas, S. Canizales-Quinteros, T. Tusie-Luna, L. Riba, M. Rodriguez-Cruz, M. Lopez-Alarcon, R. Coral-Vazquez, T. Canto-Cetina, I. Silva-Zolezzi, J.C. Fernandez-Lopez, A.V. Contreras, G. Jimenez-Sanchez, M.J. Gomez-Vazquez, J.Molina, A. Carracedo, A. Salas, C. Gallo, G. Poletti, D.B. Witonsky, G. Alkorta-Aranburu, R.I. Sukernik, L. Osipova, S. Fedorova, R, Vasquez, M. Villena, C. Moreau, R. Barrantes, D. Pauls, L. Excoffier, G. Bedoya, F. Rothhammer, J.M. Dugoujon, G. Larrouy, W. Klitz, D. Labuda, J. Kidd, K. Kidd, A. Di Rienzo, N.B. Freimer, A.L. Price, A. Ruiz-Linares, **2012**. Reconstructing Native American population history. *Nature* 488:370-374.

John D. Murdoch, William C. Speed, Andrew J. Pakstis, Christopher E. Heffelfinger, Kenneth K. Kidd, **2013**. Worldwide population variation and haplotype analysis at the serotonin transporter gene SLC6A4 and implications for association studies. *Biological Psychiatry* 74:879-889. In press: February 13, 2013; E-pub March 16, 2013.

Poster presentations reporting results of this project:

Poster presentation “Developing SNP panels for ancestry identification useful in forensic investigations” at the June **2011** annual meeting of the National Institute of Justice, Arlington, VA. Authors: Kenneth K. Kidd, Judith R. Kidd, Andrew J. Pakstis, William C. Speed, Michael P. Donnelly.

Poster presentation “Better SNPs for Better Forensics: Ancestry, Phenotype, and Family Identification” at the June **2012** annual meeting of the National Institute of Justice, Arlington, VA. Authors: Kenneth K. Kidd, Judith R. Kidd, Andrew J. Pakstis, William C. Speed

Poster presentation “Mongolians in the genetic landscape of Central Asia” ***HGM 2012***, March 11-14, Sydney, Australia. Authors: Jane Brissenden, Judith R. Kidd, Baigal Evsanaa, Ariunaa Togtokh, Kenneth K. Kidd, Janet Roscoe.

Poster presentation “Different haplotypes in East Asia and Europe both show positive selection in 1q24” presented at the annual meeting of the American Society of Human Genetics, November 6-10, 2012 in San Francisco, CA. Authors: C. Heffelfinger, A.J. Pakstis, W.C. Speed, M.P. Snyder, K.K. Kidd.

References

Beleza S., Johnson N.A., Candille S.I., Absher D.M., Coram M.A., Lopes J., Campos J., Araujo I.I., Anderson T.M., Vilhjalmsson B.J., Nordborg M., Correia E, Silva A., Shriver M.D., Rocha J., Barsh G.S., Tang H., Genetic architecture of skin and eye color in an African European admixed population, PLoS Genetics 9(3):e1003372, 2013 Mar.

Børsting C., Sanchez, J.J., Hansen, H.E., Hansen, A.M., Bruun, H.Q., Morling, N., Performance of the SNPforID 52 SNP-plex assay in paternity testing, Forensic Sci Int Genet. 2:2929-300 (2008).

Børsting C, Rockenbauer E, Morling N., Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard, Forensic Sci Int Genet. 4: 34-42, (2009).

Conrad, D.F., M. Jakobsson, G. Coop, X. Wen, J.D. Wall, N.A. Rosenberg, and J.K. Pritchard, A worldwide survey of haplotype variation and linkage disequilibrium in the human genome, Nat Genet 38:1251-1260 (2006).

Ding, L., Wiener, H., Abebe, T., Altaye, M., Go, R.C., Kercsmar, C., Grabowski, G., Martin, L.J., Khurana Hershey, G.K., Chakorborty, R., et al., 2011. Comparison of measures of marker informativeness for ancestry and admixture mapping. BMC Genomics 12, 622.

Dixon, L.A., C.M. Murray, E.J. Archer, A.E. Dobbins, P. Koumi, and P. Gill, Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes, Forensic Sci Int 154:62-77 (2005).

Donnelly M.P., P. Paschou, E. Grigorenko, D. Gurwitz, C. Barta, R.-B. Lu, O.V. Zhukova, J.-J. Kim, M. Siniscalco, M. New, H. Li, S.L. Kajuna, V.G. Manolopolulos, W.C. Speed, A.J. Pakstis, J.R. Kidd, K.K. Kidd, 2012. A global view of the OCA2- HERC2 region and pigmentation. *Human Genetics* 131:683-696.

Edwards M., A. Bigham, J. Tan, S. Li, A. Gozdzik, K. Ross, Li Jin, E.J. Parra, 2010. Association of the OCA2 polymorphism His615Arg with melanin content in East Asian populations: Further evidence of convergent evolution of skin pigmentation. *PLoS Genetics* 6:e1000867.

Enoch MA, Shen PH, Xu K, Hodgkinson C, Goldman D., 2006. Using ancestry informative markers to define populations and detect population stratification. *Journal of Psychopharmacology* 20:19-26.

Falush, D., M. Stephens, and J.K. Pritchard, Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, Genetics

164:1567-1587 (2003).

Ge, J., B. Budowle, J.V. Planz, and R. Chakraborty, Haplotype block: a new type of forensic DNA markers, . International Journal of Legal Medicine 124:353–361.

Giardina, E., I. Pietrangeli, C. Martone, P. Asili, I. Predazzi, P. Marsala, L. Gabriele, C. Pipolo, O. Ricci, G. Solla, L. Sineo, A. Spinella, and G. Novelli, In silico and in vitro comparative analysis to select, validate and test SNPs for human identification, BMC Genomics 8:457 (2007a).

Giardina, E., I. Predazzi, I. Pietrangeli, P. Asili, P. Marsala, L. Gabriele, C. Pipolo, O. Ricci, C. Martone, A. Spinella, and G. Novelli, Frequency assessment of SNPs for forensic identification in different populations, Forensic Sci Int Genet 1:e1-3 (2007b).

The HUGO Pan-Asian SNP Consortium et al., 2009. Mapping Human Genetic Diversity in Asia. *Science* 326:1541-1545.

Jakobsson, M., and Rosenberg, N.A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23, 1801-1806.

Jakobsson, M., S.W. Scholz, P. Scheet, J.R. Gibbs, J.M. VanLiere, H.C. Fung, Z.A. Szpiech, J.H. Degnan, K. Wang, R. Guerreiro, J.M. Bras, J.C. Schymick, D.G. Hernandez, B.J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H.M. Cann, J.A. Hardy, N.A. Rosenberg, and A.B. Singleton, Genotype, haplotype and copy-number variation in worldwide human populations, Nature 451:998-1003 (2008).

Jones IL et al. (2011) The potential of microelectrode arrays and microelectronics for biomedical research and diagnostics [review] Analytical & Bioanalytical Chemistry 399:2313-2329.

Kidd, J.R., F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, K.K. Kidd, **2011a**. Analyses of a set of 128 ancestry informative SNPs (AISNPs) in a global set of 119 population samples. *Investigative Genetics* 2:1 (epub January 5, 2011).

Kidd J.R., F.R.Friedlaender, A.J. Pakstis, M.R. Furtado, R. Fang, X. Wang, C.M. Nievergelt, K.K. Kidd, **2011b**. SNPs and Haplotypes in Native American Populations. *American Journal of Physical Anthropology* 146:495-502. (E-pub 13September2011).

Kayser, M., and P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, Forensic Sci Int Genet 3:154-161 (2009).

Kidd KK, Kidd JR, Speed WC, Fang R, Furtado MR, Hyland FC, Pakstis AJ, 2012.

Expanding data and resources for forensic use of SNPs in individual identification. Forensic Science International: Genetics 6:646-652.

Kersbergen, P., van Duijn, K., Kloosterman, A.D., den Dunnen, J.T., Kayser, M., and de Knijff, P., 2009. Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. BMC Genetics 10:69.

Kim, J.J., B.G. Han, H.I. Lee, H.W. Yoo, and J.K. Lee, Development of SNP-based human identification system, Int J Legal Med 124:125-131 (2010).

Kosoy, R., R. Nassir, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, et al., **2009**. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Human Mutation* 30:69-78.

Lettre, G., Genetic regulation of adult stature, Curr Opin Pediatr 21:515-522 (2009).

Li, J.Z., D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, and R.M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation, Science 319:1100-1104 (2008).

Lao O., Vallone P.M., Coble M.D., Diegoli T.M., van Oven M., van der Gaag K.J., Pijpe J., de Knijff P., Kayser M., 2010. Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. *Human Mutation* 31:E1875-93.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26:2867-2873.

McEvoy, B.P., G.W. Montgomery, A.F. McRae, S. Ripatti, M. Perola, T.D. Spector, L. Cherkas, K.R. Ahmadi, D. Boomsma, G. Willemsen, J.J. Hottenga, N.L. Pedersen, P.K. Magnusson, K.O. Kyvik, K. Christensen, J. Kaprio, K. Heikkila, A. Palotie, E. Widen, J. Muilu, A.C. Syvanen, U. Liljedahl, O. Hardiman, S. Cronin, L. Peltonen, N.G. Martin, and P.M. Visscher, Geographical structure and differential natural selection among North European populations, Genome Res 19:804-814 (2009a).

McEvoy, B.P., and P.M. Visscher, Genetics of human height, Econ Hum Biol 7:294-306 (2009b).

Murdoch JD, Speed WC, Pakstis AJ, Heffelfinger C, Kidd KK (2013) Worldwide population variation and haplotype analysis at the serotonin transporter gene SLC6A4 and implications for association studies. Biological Psychiatry In Press.

Nielsen R et al. (2011) Genotype and SNP calling from next-generation sequencing data. [Review] Nature Reviews Genetics 12:443-451.

Ngamphiw C, Assawamakin A, Xu S, Shaw PJ, Yang JO, Ghang H, Bhak J, Liu E, Tongsima S and the HUGO Pan-Asian SNP Consortium, 2011. PanSNPdb: The Pan- Asian SNP Genotyping Database. *PLoS One*. 6:e21451.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.R. Nelson, M. Stephens, and C.D. Bustamante, Genes mirror geography within Europe, Nature 456:98-101 (2008).

Odriozola, A., J.M. Aznar, L. Valverde, S. Cardoso, M.L. Bravo, J.J. Builes, B. Martinez, D. Sanchez, F. Gonzalez-Andrade, E. Sarasola, M.C. Gonzalez-Fernandez, B. Martinez Jarreta, and M.M. De Pancorbo, SNPSTR rs59186128_D7S820 polymorphism distribution in European Caucasoid, Hispanic, and Afro-American populations, Int J Legal Med 123:527-533 (2009).

Pakstis AJ, Speed WC, Kidd JR, Kidd KK, (2007) An expanded, nearly universal, panel of SNPs for individual identification. Poster presented at annual meeting of the National Institute of Justice 2007, available online at the Kidd lab website: http://medicine.yale.edu/labs/kidd/www/NIJposter2007.pdf

Pakstis A.J., W.C. Speed, R. Fang, F.C.L. Hyland, M.R. Furtado, J.R. Kidd, K.K. Kidd, 2010. SNPs for a universal individual identification panel. *Human Genetics* 127:315-324.

Pakstis A.J., R. Fang, M.R. Furtado, J.R. Kidd, K.K. Kidd, **2012**. Mini-haplotypes as lineage informative SNPs (LISNPs) and ancestry inference SNPs (AISNPs). *European Journal of Human Genetics.* 20:1148-1154.

Pereira R, Fondevila M, Phillips C, Amorim A, Carracedo A, Gusmao L, Genetic characterization of 52 autosomal SNPs in the Portuguese population, Forensic Sci. International Genetics 1: 358-360 (2008).

Phillips, C., R. Fang, D. Ballard, M. Fondevila, C. Harrison, F. Hyland, E. Musgrave-Brown, C. Proff, E. Ramos-Luis, B. Sobrino, A. Carracedo, M.R. Furtado, D. Syndercombe Court, and P.M. Schneider, Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel, Forensic Sci Int Genet 1:180-185 (2007a).

Phillips, C., A. Salas, J.J. Sanchez, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, M. Calaza, M.C. de Cal, D. Ballard, M.V. Lareu, and A. Carracedo, Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci Int Genet 1:273-280 (2007b).

Phillips, C., Freire Aradas, A., Kriegel, A.K., Fondevila, M., Bulbul, O., Santos, C., Serrulla Rech, F., Perez Carceles, M.D., Carracedo, Á., Schneider, P.M., et al., 2013. Eurasia-

plex: a forensic SNP assay for differentiating European and South Asian ancestries. Forensic Sci Int Genet 7, 359-366.

Pomeroy R, Duncan G, Sunar-Reeder B, Ortenberg E, Ketchum M, Wasiluk H, Reeder D., A low-cost, high-throughput, automated single nucleotide polymorphism assay for forensic human DNA applications, Anal Biochem. 395:61-7 (2009).

Pritchard, J.K., M. Stephens, P. Donnelly, 2000. Inference of population structure using multilocus genotype data, Genetics 155:945-959.

Rosenberg, N.A., Li, L.M., Ward, R., and Pritchard, J.K., 2003. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402-1422.

Ruiz Y, Phillips C, Gomez-Tato A, Alvarez-Dios J, Casares de Cal M, Cruz R, Maroñas O, Söchtig J, Fondevila M, Rodriguez-Cid MJ, Carracedo A, Lareu MV, **2013**. Further development of forensic eye color predictive tests. *Forensic Sci Int Genet* 7:28-40; Epub June 17, 2012.

Sanchez, J.J., C. Phillips, C. Borsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P.M. Schneider, A. Carracedo, and N. Morling, A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27:1713-1724 (2006).

Shriver, M., T. Frudakis, and B. Budowle, Getting the science and the ethics right in forensic genetics, Nat Genet 37:449-450; author reply 450-441 (2005).

Schlebusch C.M., H. Soodyall, 2012. Extensive population structure in San, Khoe, and mixed ancestry populations from Southern Africa revealed by 44 short 5-SNP haplotypes. Human Biology 54:695-724.

Speed, W.C., Kang, S.P., Tuck, D.P., Harris, L.N., and Kidd, K.K. (2009). Global variation in CYP2C8-CYP2C9 functional haplotypes. Pharmacogenomics J 9, 283-290.

Tian, C., D.A. Hinds, R. Shigeta, S.G. Adler, A. Lee, M.V. Pahl, G. Silva, J.W. Belmont, R.L. Hanson, W.C. Knowler, P.K. Gregersen, D.G. Ballinger, and M.F. Seldin, A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping, Am J Hum Genet 80:1014-1023 (2007).

Tian, C., R.M. Plenge, M. Ransom, A. Lee, P. Villoslada, C. Selmi, L. Klareskog, A.E. Pulver, L. Qi, P.K. Gregersen, and M.F. Seldin, Analysis and application of European genetic substructure using 300 K SNP information, PLoS Genet 4:e4 (2008).

Tian, C, Hinds DA, Shigeta R, Kittles R, Ballinger DG, and Seldin MF, A Genomewide Single-Nucleotide–Polymorphism Panel with High Ancestry Information for African American Admixture Mapping, Amer J Hum Gen 79: 640-649 (2006).

Tian C, Kosoy R, Nassir R, Lee A, Villoslada P, Klareskog L, Hammarström L, Garchon HJ, Pulver AE, Ransom M, Gregersen PK, Seldin MF, European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups, Mol Med. 15: 371-83 (2009).

Tully, G., Genotype versus phenotype: human pigmentation, Forensic Sci Int Genet 1:105-110 (2007).

Valenzuela, R.K., M.S. Henderson, M.H. Walsh, N.A. Garrison, J.T. Kelch, O. Cohen-Barak, D.T. Erickson, F. John Meaney, J. Bruce Walsh, K.C. Cheng, S. Ito, K. Wakamatsu, T. Frudakis, M. Thomas, and M.H. Brilliant, Predicting Phenotype from genotype: Normal Pigmentation, J Forensic Sci 55: 315-322 (2009).

Vallone, P.M., A.E. Decker, and J.M. Butler, Allele frequencies for 70 autosomal SNP loci with U.S. Caucasian, African-American, and Hispanic samples, Forensic Sci Int 149:279-286 (2005).

Yaeger R., A. Avila-Bront, K. Abdul, P.C. Nolan, V.R. Grann, M.G. Birchette, S. Choudhry, E.G. Burchard, K.B. Beckman, P. Gorroochurn, E. Ziv, N.S. Consedine, A.K. Joe, 2008. Comparing Genetic Ancestry and Self-Described Race in African Americans Born in the United States and in Africa. *Cancer Epidemiol Biomarkers Prevention* 17:1329-1338.

Visser M., M. Kayser, R-J. Palstra, 2012. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Research* 22:446-455.

APPENDIX

Attached here is a copy of the Kidd et al. paper “Microhaplotypes are a powerful new type of forensic marker”. Which was published online at the journal website: *Forensic Science International: Genetics Supplement Series* in December 2013. This short paper is based on a slide presentation made by K.K. Kidd on September 4, 2013 at the International Society of Forensic Genetics (ISFG) annual meeting that was held in Melbourne, Australia.