Document Title:	Population Genetics of SNPs for Forensic Purposes (Updated)
Author:	Kenneth K. Kidd
Document No.:	236433
Date Received:	November 2011
Award Number:	2007-DN-BX-K197

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

NIJ Final Report

September 1, 2007 to February 28, 2011

Population Genetics of SNPs for Forensic Purposes

NIJ Grant# 2007-DN-BX-K197, including supplement

Kenneth K. Kidd (PI), Yale University School of Medicine

Portions of this report are taken from ten research publications. two submitted manuscripts, and a number of poster presentations--all supported by this grant or the preceding grant (NIJ 2004-DN-BX-K025).

Kidd, J.R., F.R. Friedlaender, A.J. Pakstis, M.R. Furtado, R. Fang, X. Wang, C.R. Nievergelt, K.K. Kidd. SNPs and haplotypes in Native American populations. Submitted to American Journal of Physical Anthropology.

Kidd, K.K., J.R. Kidd, W.C. Speed, R. Fang, M.R. Furtado, F.C.L. Hyland, A.J. Pakstis Expanding data and resources for forensic use of SNPs in individual identification. Submitted to Forensic Science International: Genetics.

Sampson, J., K.K. Kidd, J.R. Kidd, and H. Zhao, Select SNPs to identify ancestry, Annals of Human Genetics, In press, early 2011.

Kidd, J.R., F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, K.K. Kidd, **2011**, Analyses of a set of 128 ancestry informative SNPs (AISNPs) in a global set of 119 population samples. Investigative Genetics 2:1 e-pub January 5, 2011; In press October 20, 2010.

Donnelly, M.P., P. Paschou, E. Grigorenko, D. Gurwitz, S.Q. Mehdi, S.L.B. Kajuna, C. Barta, S. Kungulilo, N.J. Karoma, R.-B. Lu, O.V. Zhukova, J.-J. Kim, D. Comas, M. Siniscalco, M. New, P. Li, H. Li, W.C. Speed, H. Rajeevan, A.J. Pakstis, J.R. Kidd, K.K. Kidd, 2010. The distribution and most recent common ancestor of the 17q21 inversion in humans. American Journal of Human Genetics 86:161-171.

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

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.

Li H., K. Cho, J.R. Kidd, K.K. Kidd, 2009. Genetic Landscape of Eurasia and ‘‘Admixture’’ in Uyghurs. *The American Journal of Human Genetics* 85(6):934-937 (Letter to Editor).

Pakstis, A.J., W. C. Speed, J. R. Kidd, and K. K. Kidd, 2008. SNPs for Individual Identification. Progress in Forensics Genetics Genetics Suppl Series 1:479-481.

Butler, J. M., B. Budowle, P. Gill, K. K. Kidd, C. Phillips, P. M. Schneider, P. M. Vallone, and N. Morling, 2008. Report on ISFG SNP Panel Discussion. Progress in Forensics Genetics Genetics Suppl Series 1:471-472.

Pakstis A.J., W. C. Speed, J. R. Kidd, and K. K. Kidd, 2007. Candidate SNPs for a universal individual identification panel. Human Genetics 121:305-317

Kidd K.K., A.J. Pakstis, W.C. Speed, E.L. Grigorenko, S.L.B. Kajuna, N.J. Karoma, S. Kungulilo, J-J. Kim, R-B. Lu, A. Odunsi, F. Okonofua, J. Parnas, L.O. Schulz, O.V. Zhukova, and J. Kidd, 2006. Developing a SNP panel for forensic identification of individuals. Forensic Science International 164 (1): 20-32.

Other noteworthy information

Here is a recent publication from Chinese researchers who developed a multiplexing assay for 44 of the IISNPs that we published in Human Genetics [Pakstis et al. 2010]:

Lou, C., B. Cong, S. Li, L. Fu, X. Zhang, T. Feng, S. Su, C. Ma, F. Yu, J. Ye, L. Pei, 2011. A SNaPshot assay for genotyping 44 individual identification single nucleotide polymorphisms, Electrophoresis, 32:1-11.

1. Abstract

Some SNPs show little allele frequency variation among populations while remaining highly informative. Such SNPs represent a potentially useful supplemental resource for individual identification in forensics especially when considered in light of several advantageous characteristics of SNPs generally compared to STRPs (Simple Tandem Repeat Polymorphisms). Our specific goals were to improve two preliminary panels of SNP markers: (1) SNPs with globally low Fst and high average heterozygosity for use in individual identification and (2) SNPs with globally high Fst and at least moderate average heterozygosity for use in ancestry inference. The first of those panels would provide exclusion probabilities (or match probabilities) for individual identification with especially low dependence on ancestry. The second panel would provide highly accurate specificity of biological ancestry for forensic investigation. Using our previously described efficient strategy for identifying and characterizing SNPs useful for individual identification, we have identified a sufficient number of SNPs for individual identification (IISNPs) using our unique collection of cell lines on population samples from around the world. We identified and published [Pakstis et al., 2010] a panel of 92 best SNPs studied on 44 population samples from around the world. These SNPs have both low Fst (<0.06) and high heterozygosity (>0.4). Of these, 45 SNPs have no genetic linkage and give average match probabilities of less than 10^{-17} in most of the 44 populations and less than 10^{-15} in all, including the several small isolated populations. Of the remaining SNPs most show no significant pairwise linkage disequilibrium. If only 6 SNPs are set aside as “alternatives”, the remaining set of 86 IISNPs are statistically independent at the population level and give match probabilities less than 10^{-31}

irrespective of population. We now consider our IISNP panel to be final and have made the list of IISNPs public (Pakstis et al., 2010) and are preparing a second manuscript for publication providing additional analyses and noting the additional populations that have been studied for many of the specific IISNPs.

We have made a strong start on developing a panel of ancestry informative SNPs (AISNPs) as an investigative tool. One initial focus has been on developing statistical criteria for evaluating the quality of a panel of AISNPs. We have used multiple approaches to the identification of SNPs potentially informative for biological-ethnic ancestry. Our developing AISNP panel currently consists of 430 candidate AISNPs that, *in toto* and in some subsets, give greatly improved resolution of the four continental groupings of populations. A subset of 128 of those SNPs has now been studied on 119 populations, including several samples of populations shared with us as DNA samples. A paper analyzing the data we have collected with data made public by others for a total of 119 population samples has been published [Kidd et al., 2011]. We are able with these data to distinguish, probabilistically, Southwest Asia from Europe, Siberia from East Asia, and other relevant Eurasian subregions. Additional SNPs are now being selected to refine and make more robust the finer-scale ethnic distinctions. Newly developed statistical methods are being used to select those additional SNPs from public databases and will be used on the full dataset we are developing to select the optimal subset for robust ancestry inference.

In a pilot effort we have designed a specific interface for ALFRED, the ALele FREquency Database, to make forensic SNP sets readily accessible. Allele frequencies from our IISNP sets, the SNPforID individual identification panel and the ancestry SNP sets have been entered into ALFRED.

Table of Contents

2. Executive Summary

2.1 Background and rationale

Single Nucleotide Polymorphisms (SNPs) are likely in the near future to have a fundamental role in forensics, both in human identification and description. Among their many advantages, several are especially relevant. **(1)** SNPs have an essentially zero rate of recurrent mutation. With mutation rates for SNPs estimated at 10^{-8} compared with rates of 10^{-3} to 10^{-5} for STRPs (Simple Tandem Repeat Polymorphisms), the likelihood of a mutation confounding typing is negligible and far less than other potential artifacts in typing. **(2)** SNPs have the potential for accurate automated typing and allele calling. The diallelic nature of SNPs means that allele calling is a qualitative issue not a quantitative issue, and thus more amenable to automation. **(3)** Small amplicon size is achievable with SNPs. Recent studies on mini-STRs have demonstrated the value of reducing amplicon size from the 100-450 bp range of the standard kits for CODIS (COmbined DNA Index System) loci to the 60-130 bp range especially in typing degraded forensic or archaeological samples. With a reliable multiplex procedure, many SNPs can potentially be typed using very short recognition sequences—in the range of 45-55 bp. Such short amplicons (barely exceeding the length of the two flanking PCR primers) will clearly be extremely valuable when DNA samples are severely degraded. **(4)** Finally, SNP typing can be done very quickly for large numbers of SNPs on a chip.

Considerable research is necessary to establish adequate scientific foundations for these applications. In the case of identification, because allele frequencies can vary greatly among populations, the population genetics of match probabilities is a critical issue. Some SNPs, however, show little allele frequency variation among populations while remaining highly informative, i.e., they have high heterozygosity in all populations. Such markers represent the optimal resource for individual identification. Our project determined that we could identify a sufficient quantity of such markers and we have now identified and characterized 92 such SNPs. In contrast to these SNPs, SNPs that show large allele frequency differences among populations can be very useful for inference of biological-ethnic ancestry of an individual from a DNA sample. We have made progress in identifying such SNPs. Our unique collection of cell lines on population samples from around the world was a special advantage in accomplishing these tasks.

2.2 Goals

The original purpose of the research undertaken under NIJ funding was to develop two forensic panels of SNPs that could be used, respectively, for individual and biological-ancestry identification. These panels needed sufficient research so that forensic applications would not be rejected by the courts because of inadequate scientific basis. The specific goal was to identify panels of SNP markers (1) with globally low allele frequency variation (measured as F_{st}) and high average heterozygosity and (2) with globally high F_{st} and at least moderate average heterozygosity. The first of those panels would provide exclusion probabilities (or match probabilities) for individual identification with especially low dependence on ancestry. The second panel would provide highly accurate specificity of biological ancestry for forensic investigation. Our objective has been to identify sufficient numbers of appropriate SNPs; subsequently others could determine the appropriate typing methods for forensic applications of the set of markers identified. The initial and primary emphasis was on an individual identification panel because the optimization criteria for such a panel were clear. Less clear were the procedures and criteria for optimizing an ancestry informative panel and, indeed, our progress in that area has necessarily focused on developing criteria. In addition, we have tested over 400 AISNP candidates on our population samples.

2.3 Strategy and methods for individual identification

We described both an efficient strategy for identifying and characterizing SNPs that would be valuable for individual identification (IISNPs), and then tested that strategy on a broad representation of world populations [Kidd et al., 2006]. Initially, markers with high heterozygosity and little frequency variation among African American, European American, and East Asian populations were selected for additional screening on seven populations that provide a sampling of genetic variation from the world's major geographical regions. Those with little allele frequency variation on the seven populations were then screened on a total of 40 population samples (~2,100 individuals) and the most promising retained. We not only demonstrated the feasibility of identifying SNPs with the useful properties desired but also developed in a panel of 40 statistically independent IISNPs by the time the present grant started [Pakstis et al., 2007]. While all 40 IISNPs showed no pairwise linkage disequilibrium and hence were statistically independent, the panel was not optimal. Problems with the 40-IISNP panel were the lack of alternatives and the genetic linkage among some of the SNPs making their use in situations involving relationships statistically complicated. The current project was designed to remedy those weaknesses. To address the issue of genetic linkage we preferentially selected candidate IISNPs from regions unlinked to any in the 40 IISNP panel. The new candidates were selected primarily from recently available public datasets that included data on many SNPs in many populations. We simultaneously added four populations (see table 4-1) to our panel of populations to increase the stringency of our selection process.

2.4 The best SNPs identified for individual identification

With NIJ funding we tested a total of several hundred SNPs that we identified in public databases as likely *a priori* to have high heterozygosity and low allele frequency variation globally. From these we have selected the markers with the lowest F_{st} in our expanded set of 44 populations. The result is a panel of 92 IISNPs. The dbSNP rs-numbers of the 92 IISNPs, their chromosome locations, nucleotide positions, genetic map positions, F_{st} and average heterozygosity on the 44 populations studied and other useful information can be found in a pdf file at our laboratory website: http://info.med.yale.edu/genetics/kkidd/92snpJan2009.pdf; a reformatted copy is included in this report in the Appendix. No meaningful departures from Hardy-Weinberg ratios were seen for any of the 92 IISNPs in the populations studied.

We think that with this final set of IISNPs we have finished the individual identification aspects of our IISNP project; what remains are final descriptive statistical analyses and publication of the panel and supporting statistics. This final set of SNPs is the result of searching for even better IISNPs by including additional populations and searching for unlinked SNPs. All 92 IISNPs have been reliably typed by TaqMan; how best to multiplex specific subsets to use for different identification tasks will likely depend on the application. Six of these 92 SNPs are very closely linked with others in the set and do show significant linkage disequilibrium; they should only be used as substitutes if necessary. While there is loose linkage among some pairs of the remaining 86 IISNPs they show no pairwise LD and collectively give results in the range of 10^{-31} to 10^{-35} for the 44 populations. At this level, the actual probability has no realistic meaning other than uniqueness among all humans.

The requirement that markers be unlinked led to a subset of 45 unlinked IISNPs that show little allele frequency variation among a worldwide sample of 44 populations, i.e., have a low F_{st}, while remaining highly informative. Collectively these SNPs give average match probabilities of less than 10^{-18} in most of the 44 populations and less than 10^{-15} in even the smallest most isolated population; the range of match probabilities is 2.90×10^{-19} to 5.71×10^{-16}. This 45-IISNP panel is the primary panel we advocate for ordinary forensic use with the remainder of the 86 for use if markers fail or very close relatives are involved.

These 45 SNPs are excellent for the global forensic community to consider as a universally valid individual identification panel applicable in forensics and paternity testing. They are also immediately useful for efficient sample identification/tagging in large biomedical, association, and epidemiologic studies. The best technology for multiplexing sets and for routinely using such markers still needs to be determined through empiric studies in forensic laboratories. The relative ease with which our panel of 45 best markers was identified also provides a cautionary lesson for investigations of possible balancing selection.

We also collaborated with the SNPforID consortium in evaluating some of their more promising markers [Sanchez et al., 2006] to determine which ones might be comparable to our earlier best 40 SNPs. We found that four out of 47 SNPforID markers meet the dual requirements of high heterozygosity (≥0.4) and Fst ≤0.060 when typed on our panel of 44 population samples; these four are among the 92 IISNPs based on those populations. However, none of the four made our panel of 45 unlinked IISNPs.

2.5 Progress on AISNPs (Ancestry Informative SNPs)

We have made a strong start on developing a panel of high F_{st} SNPs as an investigative tool, with an initial focus on resolution at the “continental” level but also on developing criteria for evaluating the quality of a panel of AISNPs. We initially sought appropriate markers for robustly resolving geographic and population structure with multiple screening procedures: (1) high F_{st} markers identified in the Celera or HapMap databases, (2) the ten markers published by Lao et al. [2006], (3) the markers identified in our previous study [Kim et al., 2005] as having a very large difference between Chinese and Japanese allele frequencies, (4) markers from our studies that have above average F_{st} within each region and (5) more recently, markers from other studies designed to identify SNPs to detect admixture in biomedical disease studies [e.g., Kosoy et al [2009]; Enoch et al [2006]). Our developing AISNP panel currently consists of over 400 candidate AISNPs tested on a minimum of the 44 populations used for the IISNP study.

Currently we are pursuing several approaches to improve the panel. First, existing markers that appear to be good based on the 44 population data are being typed on additional recently available population samples. For most of these markers we have extended the AISNP study to include smaller samples of a wider variety of populations. We have completed data for 128 SNPs on 68 populations from our lab and an additional 51 population samples from the literature. A manuscript has been submitted and recently published [Kidd et al., 2011]. Second, we have just finished typing 65 of our populations on 40 of the 41 AISNPs identified by Caroline Nievergelt at UCSD (unpublished) and have sent her the data for joint analyses with her existing data. Third, we are analyzing existing data to determine which geographic-ethnic distinctions are most poorly resolved. The region of Central and East Asia is one problem area; the region stretching from Europe through the Middle East to South Asia is another. We are using this new knowledge and a new greedy algorithm approach [Sampson et al., 2008, 2011] to identify in other public databases, primarily the HGDP (Human Genome Diversity Project) data, SNPs that should be especially informative in refining those distinctions. That Sampson algorithm will be used to refine the AISNP panel into the smallest number of SNPs still providing excellent distinction among the expanded panel of populations we are studying. We have also begun integrating three different genome-wide SNP studies based on the Illumina 650Y SNP array as a basis for identifying even better AISNPs. We are using two additional approaches to identify redundancy in a panel of AISNPs and identify SNPs that contribute little to the specific ancestry inference: PCA (Principal Components Analysis) and heatmap analyses.

3. Background and Rationale

Single Nucleotide Polymorphisms (SNPs) are being considered for a potentially useful role in forensic human identification [Gill et al., 2004; Amorim & Pereira, 2005;

Sanchez et al., 2003; Sanchez et al., 2006]. Among their advantages are: (1) SNPs have essentially zero rate of recurrent mutation. With mutation rates for SNPs estimated at 10^{-8} [Reich et al., 2002] compared with rates of 10^{-3} to 10^{-5} for STRPs [Huang et al., 2002; Dupuy et al., 2004], the likelihood of a mutation confounding typing is negligible and far less than other potential artifacts in typing. (2) SNPs have the potential for accurate automated typing and allele calling using chips (e.g., commercial products by Illumina and Affymetrix) or other multiple-SNP typing procedures (e.g., commercial products of Applied Biosystems). With these methods SNP typing can be done very quickly for large numbers of SNPs. The diallelic nature of SNPs means that allele calling is a qualitative issue not a quantitative issue, and thus more amenable to automation. (3) Small amplicon size is achievable with SNPs. Recent studies on miniSTRs [Coble & Butler, 2005; Butler et al., 2003; Holland et al., 2003] have demonstrated the value of reducing amplicon size from the 100-450 bp range of the standard kits for CODIS (COmbined DNA Index System) loci to the 60-130 bp range especially in typing degraded forensic or archaeological samples. With a reliable multiplex procedure, many SNPs can potentially be typed using very short recognition sequences—in the range of 45-55 bp. Such short amplicons (barely exceeding the length of the two flanking PCR primers) will clearly be extremely valuable when DNA samples are severely degraded.

Two problems with SNPs replacing STRPs for individual identification in forensics are commonly recognized. One is the “inability” to reliably detect mixtures, which are a significant occurrence in case work. The other is the inertia created by the large existing databases of CODIS markers. However, SNPs do not have to be all-purpose to have a useful role in forensics. SNPs can be very useful in missing persons cases in which the CODIS databases are not relevant. Also, many local cases involving a minor crime and suspect in hand could quickly use SNPs for confirmatory evidence. A much more significant problem is the population genetics of SNPs. With multiallelic markers, such as the standard CODIS STRPs, most of the alleles at most of the loci are low frequency in most populations. This means that match probabilities are low irrespective of population. Nonetheless, those probabilities might differ by several orders of magnitude. For example, the match probabilities for individuals that were calculated some years ago for some VNTRs lie in the realm of 10^{-10} to 10^{-13} [Chakraborty & Kidd, 1991]. Probabilities of 10^{-10} or less also occur for the CODIS markers (unpublished data). Probability differences of three orders of magnitude in such ranges are not relevant to decisions about the meaning of/cause of the match. The problem with SNPs is that the frequency of an allele can range from zero to one among different populations, causing a very large dependence of the match probability on the population frequencies used for the calculation. Figure 3-1 is an example of SNPs that have widely varying allele frequencies around the world. Were this level of variation true of SNPs used for calculating match probabilities in forensics, some of the criticisms of Lewontin and Hartl [Lewontin & Hartl, 1991] might have some validity. In contrast, that high level of variation in allele frequencies among populations can be valuable for other forensic purposes, such as an investigative tool for inference of ancestry, as discussed below.

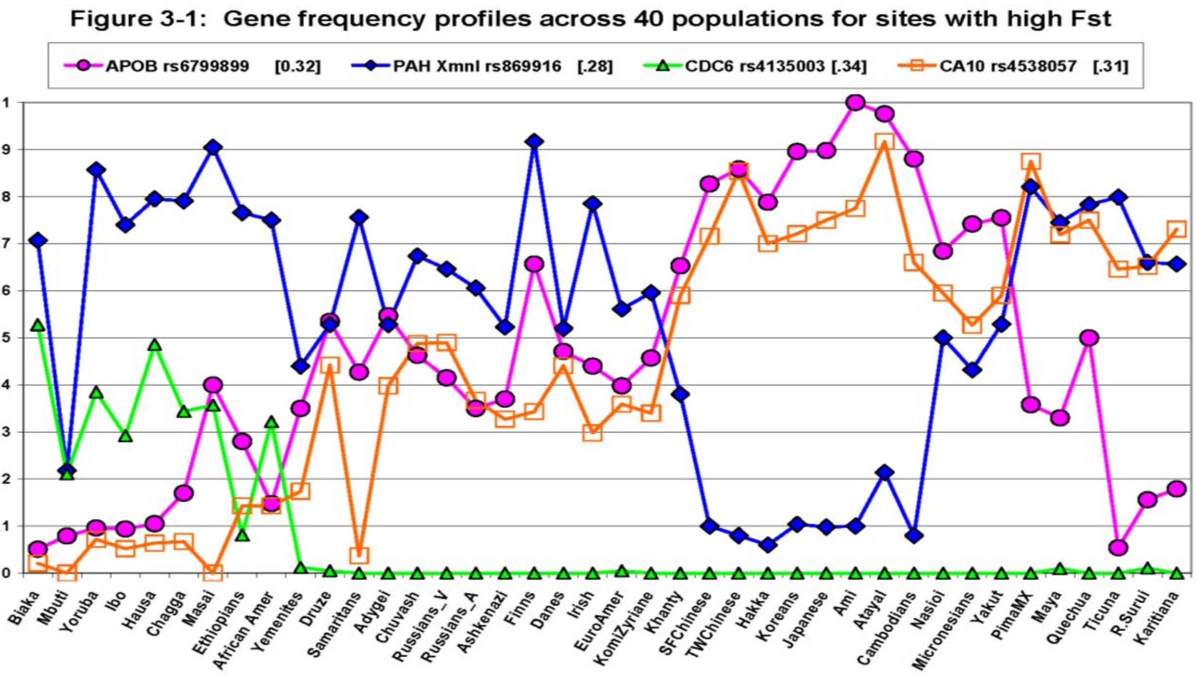


Figure 3-1. The frequencies of one allele at each of four SNPs with high variation in allele frequencies among populations. The SNPs are identified by their rs number in dbSNP and the symbol of the genetic locus in which each occurs; the data are in ALFRED. The populations are arranged by geographic region in rough order of distance from Africa but arbitrarily within each geographic region. See Table 4-1 for more detail on the populations.

Thus, we distinguish “individual identification” from “ancestry inference.” We have also noted that two other types of studies require SNPs with different characteristics: SNPs that are phenotype informative (in the sense of visually discernible features such as hair, eye, or skin color) and SNPs that are lineage-family informative (Table 3-1; cf. also Butler et al., 2008). Identifying gross phenotypes other than whatever may be significantly correlated with geographic ancestry is a highly problematic area [Kayser and Schneider, 2009; Royal et al., 2010]. Similarly, lineage informative markers will likely be multi-allelic markers such as haplotypes of SNPs. The considerable work needed for those types of SNPs has not been a major part of our forensic research, other than a minor demonstration in a poster for the NIJ Grantee meeting in 2007 (available on the Kidd Lab web site: <http://info.med.yale.edu/genetics/kkidd>). Recently we have begun pilot work for future forensic studies in this area [e.g., Donnelly et al., 2010]. However, in this report most of our references to pigmentation-related phenotypes (hair, eye, skin) are a convenient labeling of some known biological functions of genes. Underlying SNPs at those genes have been identified which show sufficiently different allele frequencies across ethnic groups to be useful contributors to panels of ancestry informative SNPs. They do not indicate any demonstrable method of identification via visually discernible features. Many different genetic loci and environmental factors typically contribute in complex ways to visually discernible phenotypes such as hair color, eye color, skin color, height, weight, etc.

Table 3.1. Types of Panels of SNPs for Forensic Applications

Individual Identification SNPs (IISNPs): SNPs that collectively give very low probabilities of two individuals having the same multisite genotype.

Ancestry Informative SNPs (AISNPs): SNPs that collectively give a high probability of an individual’s ancestry being from one part of the world or being derived from two or more areas of the world.

Lineage Informative SNPs (LISNPs): Sets of tightly linked SNPs that function as multiallelic markers that can serve to identify relatives with higher probabilities than simple bi-allelic SNPs.

Phenotype Informative SNPs (PISNPs): SNPs that provide high probability that the individual has particular phenotypes, such as a particular skin color, hair color, eye color, etc.

We can elaborate the criteria for IISNPs to be used in forensic applications to include:

1. An easily typed unique locus.

2. Highly informative for the stated purpose.

3. Well documented relevant characteristics.

Each of the types of panels requires a different set of elaborated criteria. For IISNPs our research has concentrated on these three characteristics as relevant to individual identification, but we recognize that other characteristics are important for SNPs that can be put into a database analogous to CODIS. For individual identification, comparable to the standard use of CODIS markers in forensics, a panel of SNPs all with high heterozygosity and essentially identical allele frequencies in all populations would be ideal because the match probability would be nearly constant irrespective of population. Fortunately, not all SNPs are as varied in allele frequency among populations as those in Figure 3-1. Some have remarkably little variation in allele frequency around the world as shown in Figure 3-2.

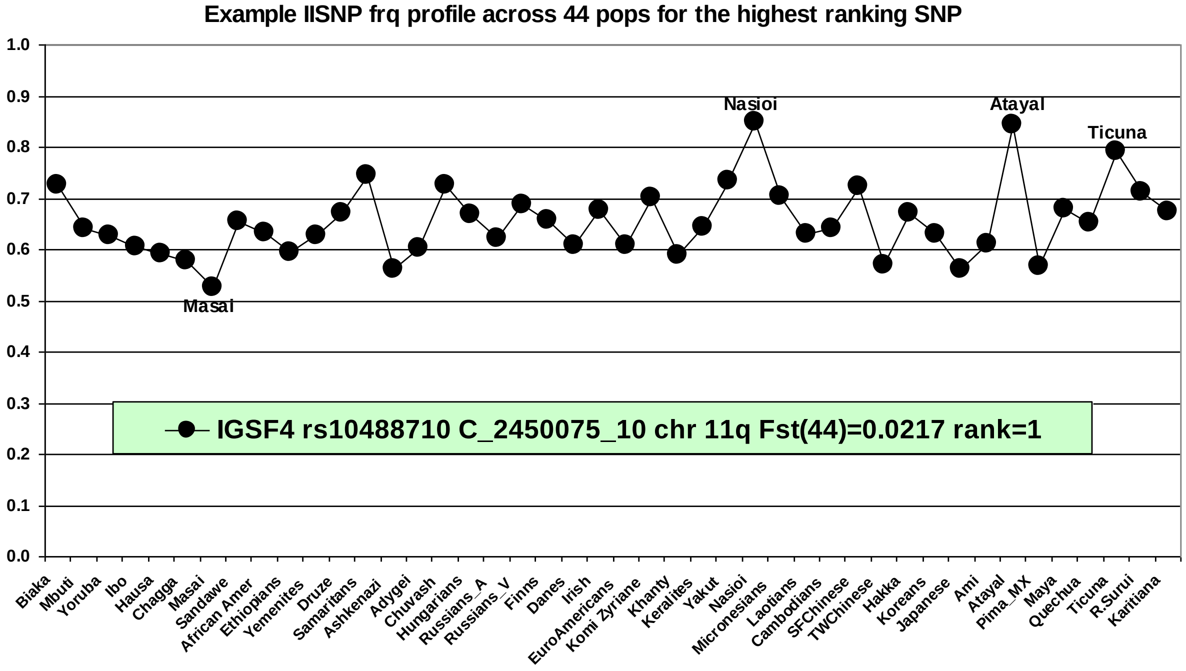


Figure 3-2. The frequencies of one allele of the best IISNP found to date in 44 populations. While the frequency varies around the world, it varies much much less than the average and is highly heterozygous in almost all populations. The populations are arranged as in Figure 3-1. See Table 4-1 for more detail on the populations.

The problem is to identify appropriate individual identification SNPs (IISNPs) and demonstrate their low allele frequency variation sufficiently well for forensic purposes. Most of our research to date has been directed to identification of SNPs that have high heterozygosity around the world with minimal allele frequency variation so that a single small match probability can be used universally between a crime scene sample and an individual's DNA profile because it will be sufficiently independent of ethnicity/ancestry.

In contrast to the use of SNPs for matching individual profiles it is also possible to use SNPs to infer ancestry of the individual from the DNA, as Figure 3-1 implies. A panel of AISNPs should be highly differentiating of ancestral origins of an individual DNA sample with a reasonable number of SNPs and have the population genetics support that would allow a high enough probability of correct ancestral assignment to make it a strong investigative tool. We are striving for greater specificity of ancestry than is generally provided by "continental" assignment. It is clear that differentiation, on average, among even closely clustering groups (e.g., European populations) is possible if enough markers are used [Li et al., 2008, Novembre et al., 2008]—that is not the problem. The problem is identifying ancestry for a single individual with a reasonable number of SNPs. Ultimately, it may be that no single small panel will be optimal for all questions

Comparing allele frequency differences for different SNPs studied on different sets of populations is not straightforward and involves inference based on other knowledge of the population relationships. We are also aware that a SNP with a very unusual frequency in one population may not be reliable because of errors in sampling or typing. Obviously, SNPs with larger allele frequency differences between populations will provide better differentiation. Proper comparison requires that we test all candidates on a single large collection of population samples.

SNPs have already been shown to allow the easy (though fairly rough) resolution of the four continental groups with as few as 10 SNPs [Lao et al., 2006]. However, their analyses on the HGDP-CEPH panel (and their 10 SNPs on 40 of our populations) of those markers did not allow any further subdivision of populations even when regions were examined separately using the program STRUCTURE.

We are in the process of identifying additional SNPs that are targeted at discriminating within specific geographic regions. We are using multiple sources. One of those is our own laboratory. With the already large number of SNPs now typed on our population samples as part of other studies, we expect to identify those SNPs that are most informative for the identification of populations within specific geographical regions. For example, we have shown that rs671 at ALDH2 varies greatly within East Asia [Li et al., 2009] and is fixed elsewhere. Such SNPs have not been incorporated into the AISNP analyses yet. Of the ~4000 SNPs we have typed on 37 to 44 of our populations as part of other research projects, most do not have sufficiently varying frequencies around the world to be useful for an AISNP panel, but we have undertaken additional analyses to screen for those that may be useful but not yet recognized as such.

The other resources are the published and available data from other studies, the HapMapIII data on 11 populations and the extensive data on the HGDP-CEPH populations. (We note that about one-third of the HGDP-CEPH samples are from our lab and already being studied by us but the additional SNPs and populations in that collection are a major resource.) Other studies have not used these populations, but have published the frequency data on their set of AISNPs. We can use those to identify particularly informative SNPs that complement or supplement the distinctions already made by our panel. As an example of one way to use those resources to identify SNPs to type on our populations we have preliminary heatmap analyses of the data from Hodgkinson et al. [2008] on their panel of 186 AISNPs typed on the HGDP-CEPH panel [Figure 3-3]. (Dr. David Goldman kindly sent us the raw data.)

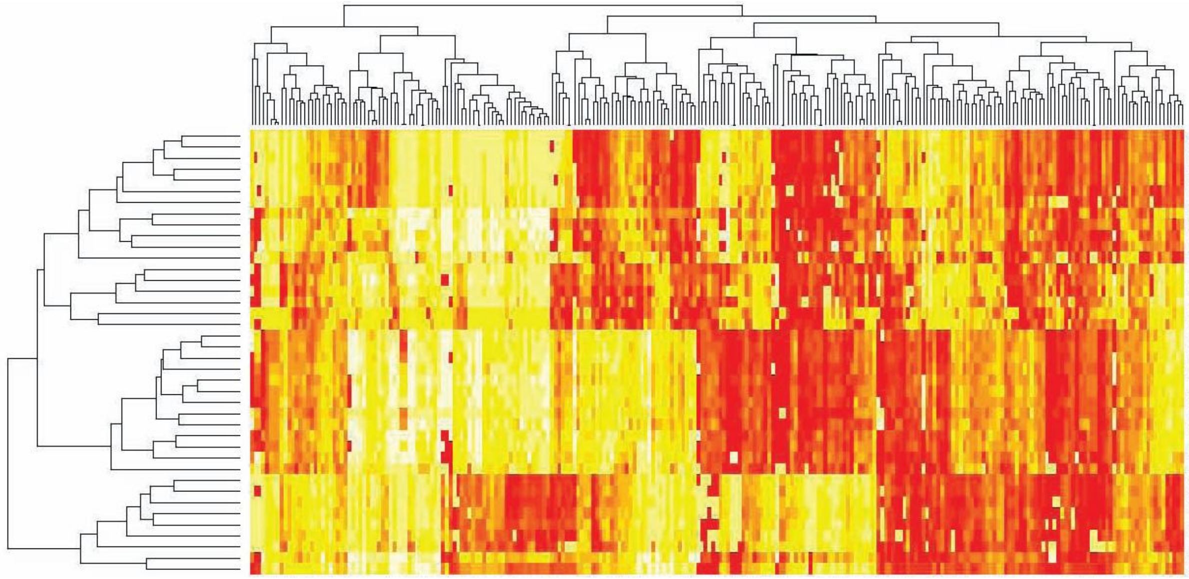


Figure 3-3. Heatmap image of 186 SNPs (X axis) typed on 40 populations (Y axis). This image shows which SNPs contribute most (red) and least (yellow) to the discriminations of the populations as they determine the phenogram of populations shown on the Y axis. (Data from Hodgkinson et al. (2008)

At the beginning of the grant now ending, we had published our preliminary work [Kidd et al., 2006; Pakstis et al., 2007] on a panel of IISNPs that would be universally applicable and had presented in posters some very preliminary work on ancestry inference with AISNPs. Neither was an ideal panel. The 40 IISNPs identified at that time contained several pairs that were linked, reducing the applicability in situations involving family relationships. The AISNPs could easily distinguish four continental regions (Northwest Europe, Africa, Far East Asia, and the Americas) but could not clearly allow any inference for individuals from geographically intermediate regions.

4. Goals

The original purpose of the research undertaken under NIJ funding now ending was to develop two forensic panels of SNPs that could be used, respectively, for individual identification and inference of biological ancestry. These panels needed sufficient research so that when attempting to introduce them for forensic applications they would not be rejected by the courts because of inadequate scientific basis. The specific goal was to identify panels of SNP markers (1) with globally low F_{st} and high average heterozygosity and (2) with globally high F_{st} and at least moderate average heterozygosity. The first of those panels would provide exclusion probabilities (or match probabilities) for individual identification with especially low dependence on ancestry. The second panel would provide highly accurate specificity of biological ancestry for forensic investigation.

We justified our goals of continuing to develop both IISNP and AISNP panels based on our unique collection of population samples (Table 4-1), our well-equipped molecular laboratory, our extensive experience in population genetics, and considerable experience testifying during the early use of DNA in forensics. We felt we knew what the Courts would require as scientific support for use of SNP panels and that we were in an ideal position to develop panels meeting those criteria. The necessity for population data for forensic SNPs was especially evident when the need for SNPs in identification of victims on the World Trade Center attacks could not find any with adequate scientific support for use in a multiethnic population.

Our collection of population samples also provides a unique resource for validating SNPs that can be used in investigations to identify the ethnic ancestry of the individual leaving a DNA sample at a crime scene. As seen in Figure 3-1, SNPs that vary considerably in frequency can carry information on ancestry. Our populations provide an excellent global overview of human variation as shown in various publications [e.g. Kidd et al. 2004; Tishkoff & Kidd 2004]

Our objective continued to be to identify appropriate SNPs; subsequently others could determine the appropriate typing methods for forensic applications of all or a subset of markers identified. The initial and primary emphasis was on an individual identification panel because the optimization criteria for such a panel were clear. Less clear were the procedures and criteria optimizing an ancestry informative panel and indeed, our progress in that area has focused on developing criteria for optimization and a dataset of candidates from which to select a robust set of AISNPs. We have now finished identification of SNPs for an IISNP panel and have published the panel and initial statistical analyses [Pakstis et al. 2010]. We are in the process of finishing a second manuscript for publication including some additional statistical support for the panel and an update with additional allele frequencies for many of the IISNPs culled from the literature. We have also made considerable progress in developing an AISNP panel with a preliminary set of good SNPs [Kidd et al., 2011], a clear procedure and resources for identifying additional excellent candidates to add to the developing panel, and methods for refining the developing pool of candidates based on the specific question being asked.

TABLE 4-1 The 44 population samples

Geographic Region	Name	N	Population ALFRED UID	Sample ALFRED UID
Africa	Biaka	70	PO000005F	SA000005F
	Mbuti	39	PO000006G	SA000006G
	Yoruba	78	PO000036J	SA000036J
	Ibo	48	PO000096P	SA000096S
	Hausa	39	PO000097Q	SA000100B
	Chagga	45	PO000324J	SA000487T
	Masai	22	PO000456P	SA000854R
	Sandawe	40	PO000661N	SA001773S
	Ethiopian Jews	32	PO000015G	SA000015G
	African Americans	90	PO000098R	SA000101C
S.W. Asia	Yemenite Jews	43	PO000085N	SA000016H
	Druze	† 127	PO000008I	SA0000846S
	Samaritans	41	PO000095O	SA000098R
Europe	Adygei	54	PO000017I	SA000017I
	Chuvash	40	PO00032M	SA000491O
	Hungarians	† 145	PO000453M	SA002023H
	Russians, Vologda	48	PO000019K	SA000019K
	Russians, Archangelsk	34	PO000019K	SA001530J
	Ashkenazi Jews	83	PO000038L	SA000490N
	Finns	36	PO000018J	SA000018J
	Danes	51	PO000007H	SA000007H
	Irish	118	PO00000M	SA000057M
	Euro Americans	92	PO000020C	SA000020C
N.W. Asia	Komi Zyriane	40	PO000326L	SA000489V
	Khanty	50	PO000325K	SA000488U
S.C. Asia	Keralites	30	PO000672P	SA001854S
East Asia	SF Chinese	60	PO000009J	SA000009J
	TW Chinese	49	PO000009J	SA000001B
	Hakka	41	PO000003D	SA000003I
	Koreans	66	PO000030D	SA000936S
	Japanese	51	PO000010B	SA000010B
	Ami	40	PO000002C	SA000002C
	Atayal	40	PO000021D	SA000021D
	Cambodians	25	PO000022E	SA000022E
	Laotians	119	PO000671O	SA001853R
N.E. Asia	Yakut	51	PO000011C	SA000011C
Pacific Islands	Nasioi	23	PO000012D	SA000012D
	Micronesians	37	PO000063J	SA000063J
N. America	Pima, Mexico	† 99	PO000034H	SA000026I
	Maya	52	PO000013E	SA000013E
S. America	Quechua	22	PO000069P	SA000069P
	Ticuna	65	PO000027J	SA000027J
	Rondonian Surui	47	PO000014F	SA000014F
	Karitiana	57	PO000028K	SA000028K

Notes:
† Samples with many related individuals; most analyses include only unrelated individuals.
The four samples added most recently--Sandawe, Hungarians, Keralites, Laotians—increasing our population panel from 40 to 44 groups.

5. Initial Development of an Individual Identification Panel

5.1 Strategy

To obtain SNPs with high global heterozygosity and low inter-population variation, we initially pursued a strategy of four steps to successively enrich for appropriate SNPs. First, we identified likely candidate polymorphisms. We then screened these on a few populations. We then tested the “best” of those markers on many populations. Finally, we retained the “best of the best” (i.e., those with highest average heterozygosity and lowest variation among populations, being the most likely to be useful for individual forensic identification). As our measure of variation among populations, we have used F_{st} [Wright 1951] as a standardized measure of the variance in allele frequencies among populations.

For our initial identification of likely candidates, we used the Applied Biosystems catalog database of SNPs for which there are pre-designed, synthesized, and pre-tested TaqMan assays. We chose this source because it provided off-the-shelf assays that are guaranteed to work with no effort on our part to design and optimize an assay. From Applied Biosystems (AB) we obtained the frequencies for those TaqMan markers that had allele frequency data on four populations (African Americans, European Americans, Chinese, and Japanese). We later expanded our selection from the AB data set to include the HapMap. These markers were then rank ordered by both average heterozygosity and minimal difference in allele frequency among the four populations. We then chose markers with average heterozygosity >0.45 and F_{st} <0.01. Once a marker was selected for testing, no other markers were selected within 1Mb of that marker.

For the initial screen in our lab we chose a total of 371 individuals from seven populations selected from all major geographical regions. Finally, the SNPs that continued to have low F_{st} and high heterozygosity were tested on the remaining populations. That procedure was initially a proof of principle [Kidd et al., 2006] and then expanded to a set of 40 IISNPs based on 40 population samples [Pakstis et al., 2007]. The current 44 population samples are listed in Table 4-1.

Our most recent efforts to identify additional IISNPs to produce a panel of *unlinked* SNPs involved screening recently available public databases with data on larger numbers of populations that are at least the equivalent of our seven-population screen. Given the greater public resources available, we have targeted regions of the genome unlinked to any of our initial set of 40 IISNPs. Once selected, a candidate was typed directly on an expanded set of 44 populations. The final criteria of $F_{st} < 0.06$ and average heterozygosity > 0.4 have been maintained.

5.2 Screening criteria

To determine reasonable screening values we analyzed data we had collected on other projects and decided empirically which values to use for four populations, seven populations, and the final set of 40 or 44 populations. Those criteria were more stringent on the smaller numbers of populations since increasing the number of populations from different geographic regions is likely to only increase the Fst value. We published the details in earlier papers [Kidd et al., 2006; Pakstis et al., 2007].

Finally, we used an F_{st} of 0.06 provisionally as the upper limit for selecting “good” SNPs at the end of the second screening. This is also an arbitrary limit based on examination of the initial results. A higher value would allow inclusion of more markers that are almost as good. A lower value would decrease the number of markers but they would be even more homogeneous in allele frequencies among populations.

5.3 Marker typing

Marker typing was done with TaqMan assays ordered from the Assays-on-Demand catalog of Applied Biosystems. The manufacturer's protocol was followed using 3µl reactions in 384-well plates. PCR was done on either an AB9600 or MJ tetrad. Reactions were read in an AB7900 and interpreted using Sequence Detection System (SDS) 2.1 software. All scans were manually checked for accurate genotype clustering by the software. Assays which failed to give distinct genotype clusters or failed the Hardy-Weinberg test were discarded. All individual DNA samples that failed to give a result on the first or second screen were repeated once only to provide the final data set.

5.4 Statistical-analytic methods

Allele frequencies for each marker were estimated by gene counting within each population sample assuming each marker is a two-allele, co-dominant system. Agreement with Hardy-Weinberg ratios was tested for each marker in each population using a simple Chi-square test comparing the expected and observed number of individuals occurring for each possible genotype. When small numbers of any genotype were present, as occasionally in the smallest samples, a simulation-based exact test was used [Cubells et al. 1997]. Tests with p-values falling below thresholds such as

0.05, 0.01, and especially 0.001 were then inspected for patterns worth investigating. However, among the tests carried out for the final set of markers the numbers of tests that failed at the 5% and 1% levels were close to the numbers expected by chance and did not appear to cluster preferentially in particular markers or populations.

Because the ascertainment did not consider chromosome locations per se, the 40 best SNPs are distributed across only 16 different autosomes with eleven chromosomes having more than one SNP. Thus, not all were necessarily statistically independent in calculations of a random match probability. Therefore, the statistical independence of the markers was assessed by calculating r^2 [Kidd et al., 2004] for all of the unique, pairwise combinations of the final markers within each of the 44 populations. The r^2 value is a measure of linkage disequilibrium (LD), i.e., association of alleles at different loci.

The match probability was calculated in two steps. First, the match probability for each marker within a population was computed by finding the squared frequency of each possible genotype; these were then added together to get the locus match probability. Then, assuming the essential independence of genetic variation across markers, the locus match probabilities for each of the best markers were multiplied together within each population separately to obtain the overall average match probability for the set of best SNPs.

The frequency of the most common extended genotype for the set of best markers was calculated assuming Hardy-Weinberg ratios and the independence of the best SNP loci. For each population the most common genotype at each locus was determined using the allele frequencies in that population and then identifying which genotype has the largest expected frequency. The locus-specific values were multiplied together within each population to give the most common genotype frequency.

5.5 The 40-SNP, 40-Population Panel

At the beginning of this grant, we had established a provisional panel of the 40 best markers with a 40-population Fst below 0.06 and average heterozygosity > 0.4. These SNPs included the best of those screened as described above and the least varying 1.24% of SNP markers studied in our lab for other purposes [Kidd et al. 2004 and unpublished data]. Collectively these SNPs give average match probabilities of less than 10^{-16} in most of the 40 populations we studied and less than 10^{-14} in all but one small isolated population; the range is 2.02 x 10^{-17} to 1.29 x 10^{-13}. These 40 SNPs therefore constituted excellent candidates for the global forensic community to consider for a universally applicable SNP panel for human identification. The relative ease with which these markers could be identified also provides a cautionary lesson for investigations of possible balancing selection. We described this panel of 40 best markers in Pakstis et al [2007], and have deposited the gene frequency tables for all the markers we screened into ALFRED, the ALlele FREquency Database (http://alfred.med.yale.edu).

5.6 The 31-Population Panel

However, our panel of 40 candidate SNPs was criticized by some as being too stringent because those studies included several small, isolated groups. Therefore, we re-evaluated our data, as well as other data, after excluding the most isolated populations from consideration, reducing the screening panel from 40 to 31 populations, those most likely to be forensically relevant [Table 5-1]. A much larger panel of 108 candidate SNPs met our operationalized criteria of an Fst <0.06 and average heterozygosity >0.40 when considered on these 31 populations. In addition to the previously published 40 SNPs we were able to include some of the markers proposed by the SNPforID consortium [Sanchez et al., 2006]. Some of these 108 candidate SNPs are molecularly close and/or genetically linked making them unsuitable for studies involving relationships. However, it seemed appropriate to make all these markers

TABLE 5-1. Relationships among population sets

Population samples at Kidd Lab	Low Fst-- High Het. 40 pop. samples	31 population samples		Population samples (continued)	Low Fst-- High Het. 40 pop. samples	31 population samples
Biaka	X	X		Komi Zyrian	X	X
Mbuti	X			Khanty	X	X
Yoruba	X	X		Yakut	X	
Ibo	X	X		Nasioi	X	
Hausa	X	X		Micronesians	X	
Chagga	X	X		Cambodians	X	X
Masai	X	X		Chinese, San Francisco	X	X
African Americans	X	X		Chinese, Taiwan	X	X
Ethiopian Jews	X	X		Hakka	X	X
Yemenite Jews	X	X		Koreans	X	X
Druze	X	X		Japanese	X	X
Samaritans	X			Ami	X	
Ashkenazi	X	X		Atayal	X	
Adygei	X	X		Pima, Mexico	X	X
Chuvash	X	X		Maya	X	X
Russians, Archangel	X	X		Quechua	X	X
Russians, Vologda	X	X		Ticuna	X	
Finns	X	X		Rondonian Surui	X	
Danes	X	X		Karitiana	X	
Irish	X	X		Average(R.Surui, Karitiana)		X
European Americans	X	X				

publically available so other researchers could evaluate them by laboratory and other criteria for possible forensic use. The data were presented in posters at the NIJ meeting and at the ISFG meeting in Copenhagen in 2007and exist on our web site, <http://info.med.yale.edu/genetic/kkidd>. We do not believe it is appropriate to publish this list given that for our own studies we have retained the small isolated populations because we believe they are important for demonstrating the universality of the low match probabilities.

6. A “Final”, Universal Panel of 45-92 IISNPs

6.1 Expanding the number of populations and candidate IISNPs

There was no significant pairwise LD among the 40 SNPs in any of the 40 populations, but some pairs were sufficiently close that linkage existed. This made those SNP pairs more difficult to use in studies involving biological relationships. Therefore, in our more recent search to develop a panel of IISNPs that were universally applicable and unlinked, we preferentially targeted regions of the genome in which we did not already have good IISNP candidates in order to enlarge the number of unlinked IISNPs. We also enlarged our set of populations by adding four populations for geographic regions poorly represented in the initial 40 populations: East Africa, East Europe, South Asia, and Southeast Asia. We gleaned candidates from a very large SNP dataset [Li et al., 2008] that became available online in 2008 for the populations studied on the Human Genome Diversity Panel (HGDP). Conrad et al [2006] also made their data on the HGDP available as part of Pemberton et al [2008]. We obtained other candidate markers that we identified from the large number of SNPs in the Shriver et al. [2005] dataset which studied 14 populations from around the world.

6.2 The Yield from Screening

After screening these new large public data bases (e.g., HGDP [Li et al., 2008]), we selected additional markers likely to be unlinked to any of the original 40 IISNPs for testing on the full set of 44 populations listed in Table 4-1. As we used larger data bases for the initial screening for low Fst markers, we skipped the intermediate stage of screening seven of our populations because those datasets already had more than seven populations globally distributed. Also, as we began to have available larger screening databases, our priority was to screen for SNPs in genomic regions unlinked to existing good IISNPs that had already been identified. As public databases became more comprehensive and included more populations than the original three, our yield percentage of usable SNPs increased dramatically from about 1 in 10 to about 1 in 2 from the largest of the databases of candidate SNPs we typed on our 44 population samples (Table 4-1). Our numeric acceptance criteria remained the same--an average heterozygosity >0.4 and the F_{st} values <0.06—using all 44 populations. The increase in population samples studied from 40 to 44 did increase the stringency of the evaluation.

Genetic linkage among the SNPs meeting those acceptance criteria was assessed by first comparing the molecular maps (on which the individual SNPs could be placed) with the three common genetic linkage maps (Genethon, Marshfield, and DeCode) which are based on STR markers spaced at distances of several megabases along each chromosome (Figure 6-1). As seen in the figure, there is not a simple relationship between physical nucleotide distances and the genetic map distances, but we are able to infer the centiMorgan distances between adjacent pairs of markers. Since a Kosambi correction appears to relate recombination frequency with genetic (linkage) map distances, we assumed any markers more than about 80 to 100cM apart are unlinked and markers closer but more than 50cM apart are loosely linked.

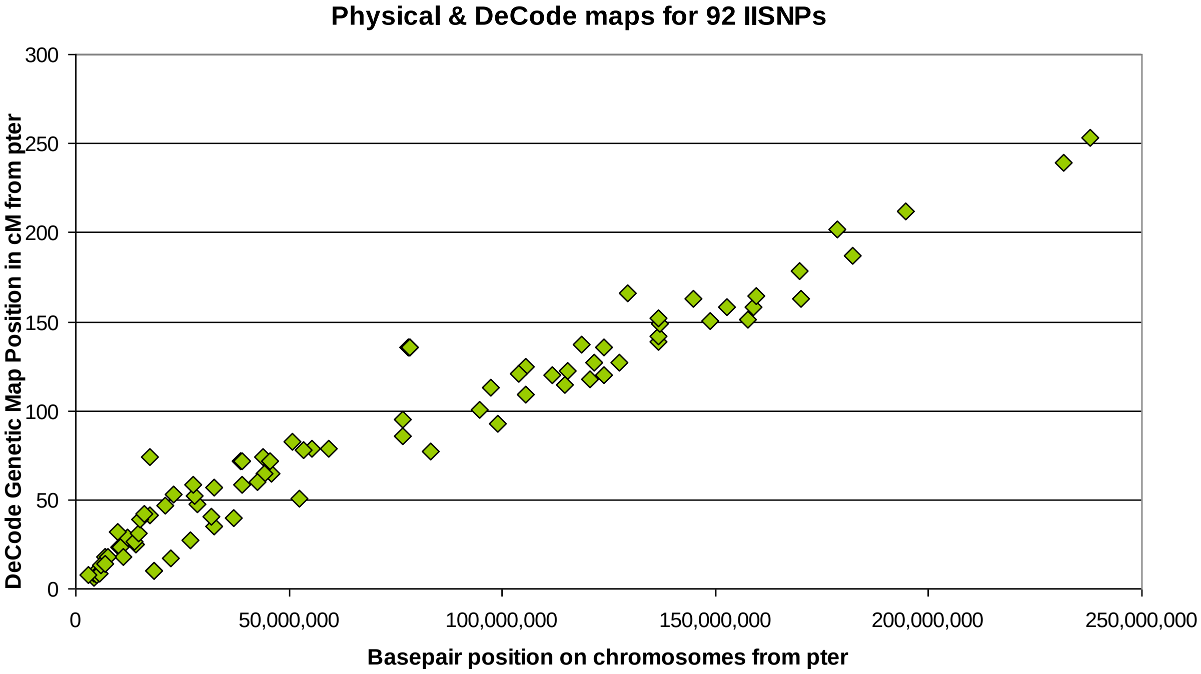


Figure 6-1. An example of the comparison of physical and recombination map locations of the 92 IISNPs. The pattern shown by the Genethon and Marshfield linkage maps is very similar.

Figure 6-2 shows the heterozygosity and F_{st} values based on 44 populations for the final 92 candidate IISNPs with the SNPs rank-ordered (left to right) from lowest to highest Fst based on the 44 populations. All 92 IISNPs have an average heterozygosity greater than or equal to 0.4. The detailed overview in Figure 6-2 shows the extraordinary level of informativeness of each SNP across the population samples studied. The median heterozygosity is 0.477 among the 4,048 values computed and

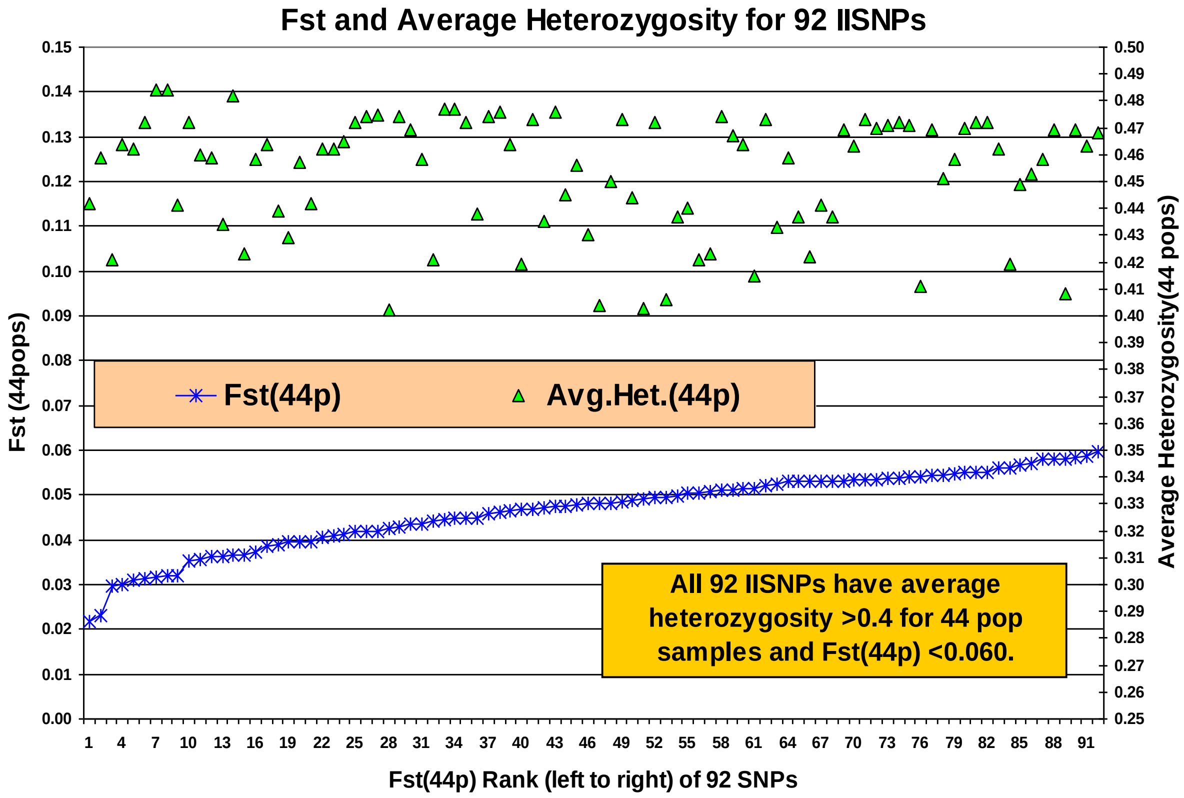


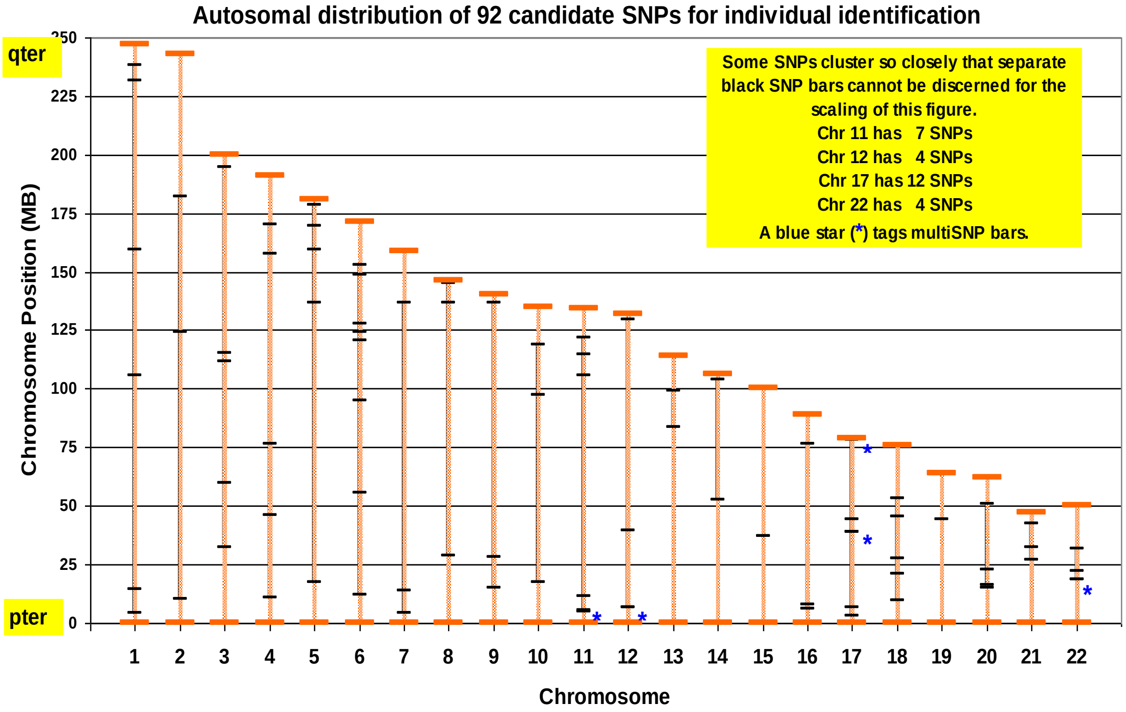
Figure 6-2. The acceptance criteria for these SNPs were a 44-population F_{st} less than 0.06 and average heterozygosity greater than 0.4.

86% of these heterozygosity values for individual populations are greater than or equal to 0.4. Less than 1% of the individual population heterozygosities are <0.2 in this highly selected group of SNPs; the small number of very low heterozygosities occur entirely in the samples from relatively small, isolated populations that often also have a high degree of inbreeding. To put in perspective how extraordinarily informative the 92 IISNPs are, we find in a set of 2,000 autosomal SNPs that we have studied on 47 population samples around the world that 26% of the heterozygosities are <0.2 compared to less than 1% of the heterozygosities for the 92 IISNPs studied on 44 population samples. The 2,000 autosomal SNPs were selected in part to be variable across the major regions of the world. A set of randomly selected SNPs would likely have a higher proportion of heterozygosities <0.2 when typed on our sampling of populations from around the world.

One of the previously reported best 40 SNPs [Pakstis et al., 2007] is not on the list of 92 IISNPs because the Fst value (0.0622) for that SNP exceeded the 0.06 threshold when tested on the 44 population samples. The remaining 39 of 40 SNPs have shifted somewhat in relative rank position due both to the expansion in the number of population samples tested which altered the average heterozygosities and F_{st} values computed as well as the fact that additional SNPs were identified that qualified for admission to the IISNP list.

No meaningful departures from Hardy-Weinberg ratios were seen for any of the 92 IISNPs in the 44 population samples studied. In addition to a standard Chi-square test, a Monte Carlo permutation test procedure was employed and 1,000 iterations were generated for each test; the proportion of probabilities obtained falling below the 5%, 1%, and 0.1% significance level thresholds were generally somewhat smaller than the values expected by chance perhaps due in part to the extensive selection procedure and very high heterozygosities of the IISNPs that made it onto the final list. All 92 IISNPs have been reliably typed by TaqMan; how best to multiplex specific subsets to use for different identification tasks will likely depend on the application. These 92 SNPs are distributed around the genome, as shown in Figure 6-3.

Figure 6-3

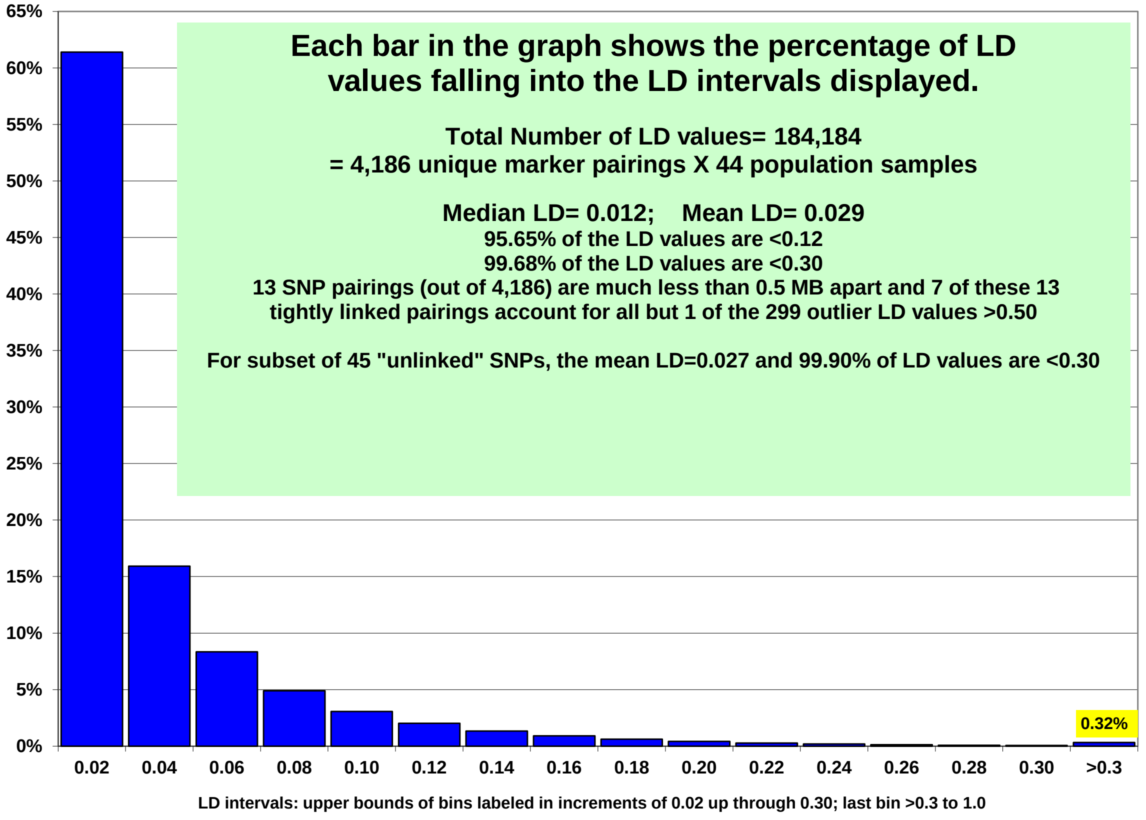


6.3. Evaluating Independence of the 92 Best SNPs

Other groups [e.g. Sanchez et al., 2006; Lee et al., 2005] have screened for unlinked SNPs so that the panel would also be appropriate for paternity testing and for forensic work that involved relatives. Obviously (Figure 6.3), some of the 92 SNPs are molecularly close (close physical linkage) and likely to show tight genetic linkage as well. Therefore we evaluated whether the closely linked markers were independent at the population level (the objective of our study) by evaluating pairwise LD among all 92 SNPs in all 44 populations. Figure 6-4 provides an overview of the results; for 86 of the 92 SNPs including the 45 unlinked SNPs (see Figure 6-5 and Appendix in section 9), the results are consistent with linkage equilibrium; the overwhelming majority of the LD values cluster close to 0.0 with the median LD value for the overall distribution equal to

0.012. Six of the 92 IISNPs show strong LD in most of the population samples in some SNP pairings due to very close physical linkage. Consequently, these six SNPs can only be considered as alternative candidates for inclusion in an IISNP panel independent at the population level; the footnote in the appendix table listing the 92 IISNPs identifies these six SNPs. When pairwise LD does not exist, as among 86 of the 92 including all of the 45 unlinked IISNPs, the SNPs are statistically independent at the population level and the "product rule" can be used to calculate match probabilities.

Figure 6-4. LD (r^2) distribution for 92 IISNPs in 44 population samples



Each bar in the graph shows the percentage of LD values falling into the LD intervals displayed.

Total Number of LD values= 184,184
= 4,186 unique marker pairings X 44 population samples

Median LD= 0.012; Mean LD= 0.029
95.65% of the LD values are <0.12
99.68% of the LD values are <0.30
13 SNP pairings (out of 4,186) are much less than 0.5 MB apart and 7 of these 13 tightly linked pairings account for all but 1 of the 299 outlier LD values >0.50

For subset of 45 "unlinked" SNPs, the mean LD=0.027 and 99.90% of LD values are <0.30

6.4 Identifying Unlinked SNPs: The 45-SNP Panel

Among the 92 IISNPs, we identified a subset of 45 unlinked SNPs that are distributed across all 22 autosomal chromosomes. Figure 6-5 displays a schematic representation of the 22 human autosomes and shows the relative positions of the 45 unlinked SNPs; locations of nine IISNPs that might be possible alternatives for one of the 45 unlinked SNPs are also shown in aqua-filled circles; these are possible alternatives for the nearest unlinked SNP (black-/gold-filled). In selecting the 45 unlinked SNPs we considered primarily the genetic map distances separating IISNPs that are located on the same chromosomes as well as the F_{st} ranks based on 44 population samples for the SNPs. Where there were alternative choices for SNPs that are genetically unlinked, SNPs were selected that have the smallest F_{st} values. If the SNPs considered happened to have identical Fst values, then the SNP with the higher average heterozygosity was selected. The 92 IISNPs are ranked in this fashion in the table available as a pdf file at our laboratory website. The genetic map distance evaluated for a particular SNP interval was typically the average of the Genethon, Marshfield, and DeCode map distances for that interval. Since the marker densities and beginning/end points of these 3 genetic maps vary on each chromosome, the genetic map distance obtained for a particular interval between syntenic SNPs is not necessarily as precise as the nucleotide distance based on the latest release of the human reference sequence for the same SNPs but the average genetic distance so obtained for the long physical intervals considered should be very satisfactory for the task of selecting unlinked SNPs. As shown in Figure 6-5, the 33 SNPs labeled by black-filled circles are very clearly unlinked as they are either on different chromosomes or else separated from other SNPs identified by black-circles by average genetic map distances of 95 centi-Morgans (cM) or more. The 12 SNPs labeled by gold-filled circles show some partial, weak linkage with adjacent SNPs labeled by black-circles but these are all still very large intervals with most much larger than 50 cM. The thirteen intervals where the genetic map is <95 centi-Morgans for these 12 SNPs (sorted by increasing interval size in cM) include: 41 [chr12], 43 [chr20], 48 [chr18], 53 [chr6], 54 [chr2], 55 [chr5], 68 [chr3], 70 [chr5], 78 [chr6], 79 [chr14], 80 [chr1], 84 [chr16], and 93 [chr3].

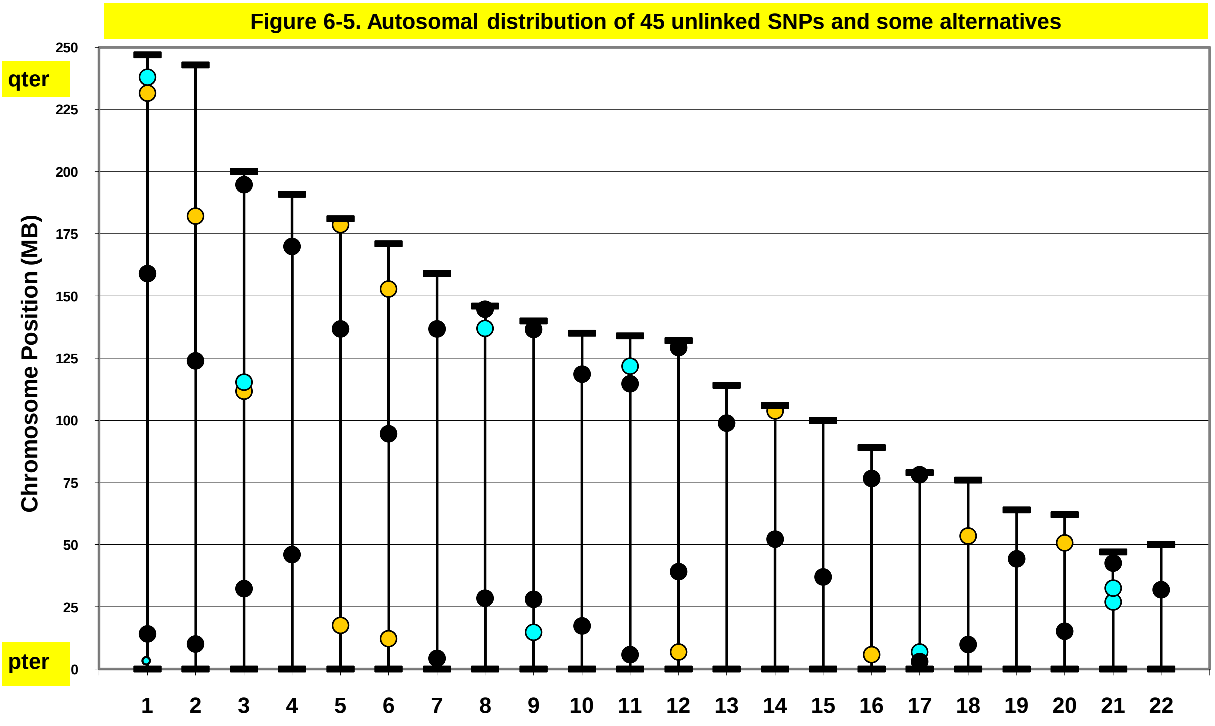


Figure 6-5. The autosomal distribution of 45 unlinked SNPs (black-, gold-filled circles) and some alternatives (aqua-filled). The 33 black-filled circles represent SNPs that are either on separate chromosomes or are separated from the nearest black-filled circle by 95 cM or more. See text discussion.

6.5 Statistics for the 45-SNP Panel

Figure 6-6 displays match probabilities and most common genotype frequencies for each population for this set of 45 unlinked IISNPs. Most of the populations have match probabilities $<10^{-17}$ and many are $<10^{-18}$; even some of the smaller, more isolated populations have match probabilities $<10^{-15}$. Thus, this set of 45 unlinked SNPs is an excellent panel for individual identification with match probabilities comparable to the CODIS STR panel and these are not highly dependent on ethnicity. Thus, it is safe to say with considerable scientific justification that a maximum match probability of $<10^{-15}$ can be used for any forensic match between any crime scene and any defendant anywhere in the world. The unlinked status of these 45 SNPs also makes them useful for situations involving close biological relationships. If relationships are not involved, more of the 92 IISNPs can be added to the set to make the match probabilities even smaller. Computing match probabilities based on all 86 IISNPs that show no LD gives results in the range of 10^{-31} to 10^{-35} for the 44 populations (Figure 6-9). At this level, the actual probability has no realistic meaning other than uniqueness among all humans.

The frequencies of the most probable 45-locus genotype (assuming Hardy-Weinberg ratios) for each population and the 45-SNP match probabilities are also quite small. Most are less than 10^{-13} and the largest is less than 10^{-11}. The larger values in the small isolated populations are relevant in that they should provide a reasonable upper bound to the match probability in any population. Of course, the caveat is that since these populations are all less than 10^{11} in size, the empiric smallest genotype frequency is 1/N in a population of size N.

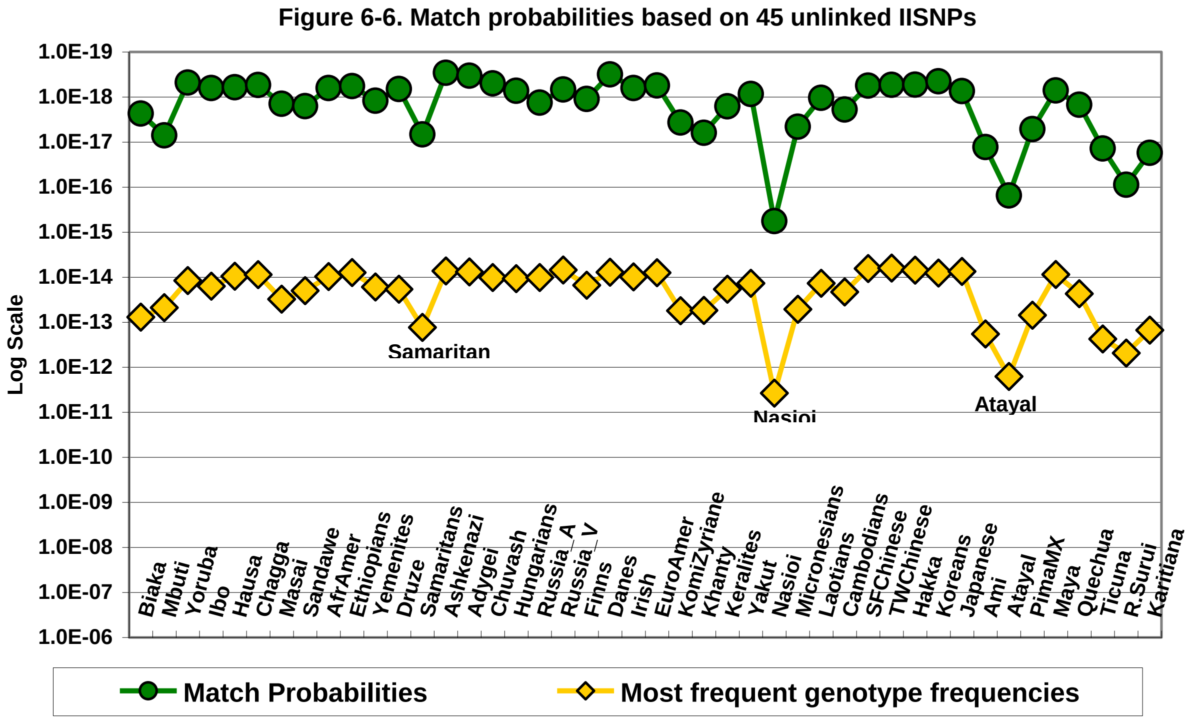


Figure 6-6. The average match probability by population as shown by the values represented by filled circles. This value assumes exact H-W ratios within each population and independence of the 45 SNPs. Most populations have values less than 10^{-17} but the values range across approximately three orders of magnitude, from less than 10^{-15} to less than 10^{-18}. We note only three populations have values about or larger than 10^{-16} and in none of those populations are there more than 10^{6} individuals. The probability of discrimination, i.e., the probability that two individuals are different, for each population is one minus the values shown in this figure. Thus, in all populations, the theoretical probability of discrimination is greater than 0.999999999999.

Empirical confirmation of the utility of the 92 IISNPs in additional populations is desirable, but we do not think it is cost effective to undertake additional specific typing at this point. We can be confident that the 45-marker panel will have essentially the same useful properties for individual identification in other large human populations. Given the global ubiquity and common frequency of both alleles at all 92 SNPs only extremely small and highly inbred populations are expected to have many of the 45 loci approach fixation of one allele. We have deliberately included several small isolated and inbred populations from different geographic regions in our studies: Mbuti from Africa, Samaritans from Southwest Asia, Khanty from West Siberia, Nasioi from Melanesia, Ami and Atayal from Taiwan, Surui and Karitiana from the Amazon. While these do show larger match probabilities (Figure 6-6) than the large populations, those probabilities are still $<10^{-15}$. Some of these smaller populations are among the smallest, most isolated in the world making it exceedingly improbable that another small population would be dramatically different. Should an individual match show few heterozygotes, that in itself is information. If necessary, additional SNPs from the remaining 47 IISNPs could be typed to yield a smaller statistical value. (However, any DNA match probability of even 10^{-2} can be meaningful in conjunction with other evidence.) Thus, while we have obtained additional population samples as this study was concluding, we have not invested the money and effort into testing additional populations for these markers. However, as noted below, additional data on many of these markers already exists in the public domain. We have begun assembling those data into ALFRED.

In all of these comparisons two populations are noticeable outliers: the Karitiana and Ticuna. Both are known to contain significant numbers of close relatives. While the exact relationships among these samples are not known, the entire Karitiana population

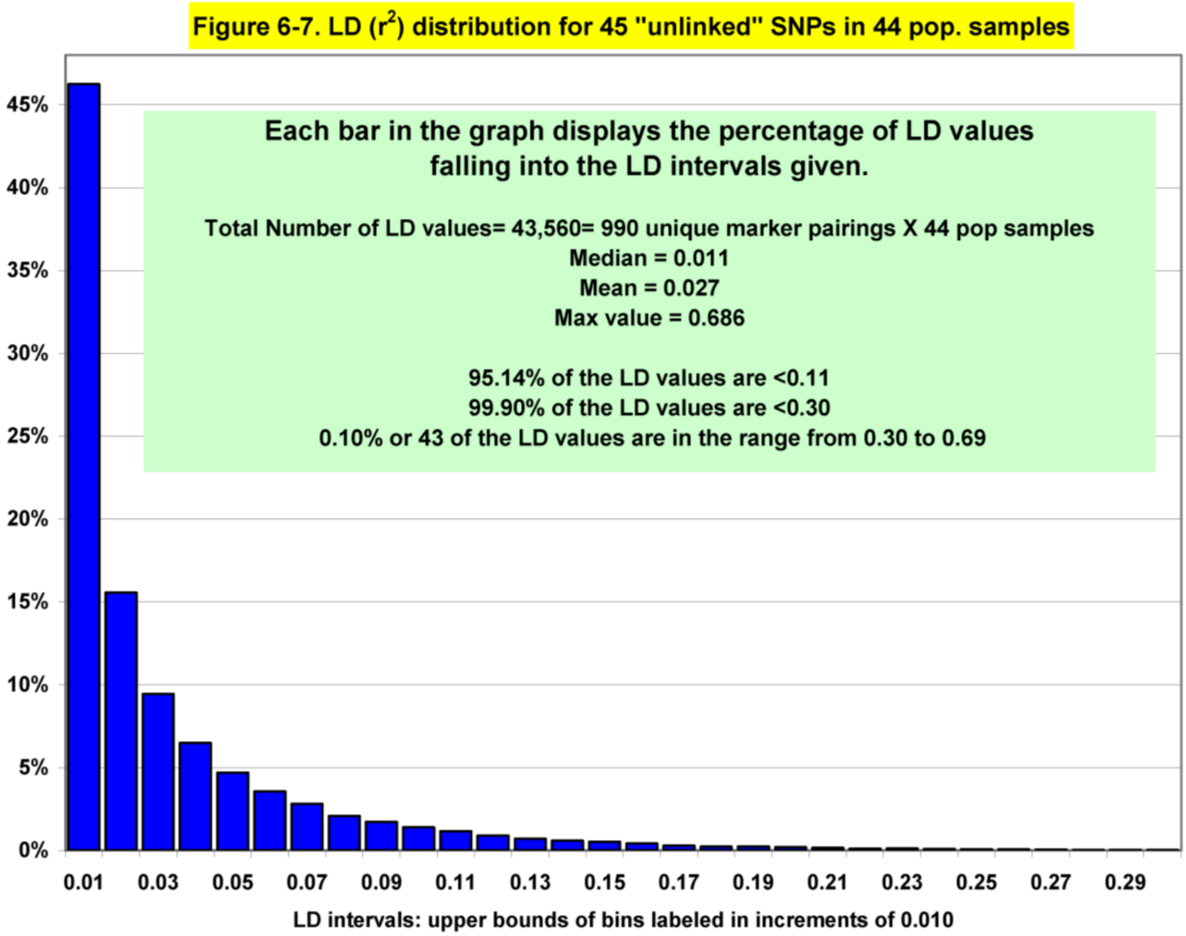


Figure 6-7. LD (r^2) distribution for 45 "unlinked" SNPs in 44 pop. samples

Each bar in the graph displays the percentage of LD values falling into the LD intervals given.

Total Number of LD values= 43,560= 990 unique marker pairings X 44 pop samples
Median = 0.011
Mean = 0.027
Max value = 0.686

95.14% of the LD values are <0.11
99.90% of the LD values are <0.30
0.10% or 43 of the LD values are in the range from 0.30 to 0.69

Figure 6.7 The pairwise LD values for the 45-IISNP panel. All of the values above the empiric 95% level occur sporadically between unlinked markers, often between markers on different chromosomes. They also occur primarily in the populations with the smaller sample sizes and we know there is a positive bias on LD in small samples. There is no consistent pattern and no biological explanation; we conclude these are chance events.

is equivalent to a single extended family so a sample of unrelated individuals is an impossibility [Kidd et al., 1993]. Inclusion of biological relatives in a sample does not bias gene frequency estimates [Cotterman, 1954] but does bias LD measures upward. Not surprisingly, other small populations such as the Rondonian Surui and Samaritans also consistently have among the highest percentages of nominally significant comparisons at all levels of significance.

There is also a positive bias in LD estimates that increases as sample size decreases [Teare et al., 2002]. This bias is observable in our results. Choosing various arbitrary thresholds as one moves farther out along the tail of the distribution of LD values, one finds higher proportions of the values involving the smaller sample sizes. For instance, among the LD values ≥0.30 we find that 39 of the 43 occurrences have sample sizes well below the median sample size of 96 chromosomes that characterizes the 43,560 LD values computed for the 45 unlinked SNPs. The Pearson correlation between LD values and sample size measured as the number of chromosomes is -0.23 for these 43,560 LD values. The medians and especially the means of the LD values summarized by population sample tend to be a little higher for the groups with the smaller sample sizes.

The median (0.011) and mean (0.027) LD values for the 45 unlinked SNPs (Figure 6-7) are close to zero and the computed LD values that are nominally significantly different from zero are approximately what would be expected by chance and primarily involve markers on different chromosomes and/or occur in the smallest populations in which the LD values are biased upward. In addition, small inbred populations necessarily contain related individuals and can be expected to show extended LD—the R.Surui [Calafell et al., 1999] and Karitiana [Kidd et al., 1993]. About 99.90% of all the LD values are <0.3 for the 45 unlinked SNPs. The relatively small number of LD values ≥0.3 (i.e., 43 values) occurred almost entirely between unlinked markers (42 of 43). The two largest LD values observed are 0.686 in the Masai and 0.475 in the Nasioi and these outliers occur for different SNP pairings. Both of these outlier LD values involve SNP pairings across chromosomes and the mean LD values for 44 populations are 0.04 and 0.03, respectively for these two SNP pairings where the 2 highest outlier values occurred. All SNP pairings for the 45 unlinked SNPs were inspected in which outlier values occurred that are ≥0.30. In all 43 instances, the outlier values appear to be isolated cases with no evidence to even suggest a supportive pattern of other moderate level LD values among the other populations studied.

The 29 unique SNP pairings that involve intra-chromosomal comparisons for the 45 unlinked SNPs were also examined more closely; a total of 1,276 LD values (=29 pairings times 44 pops) occur for these comparisons and the 29 physical intervals range from 32.3 MB to 217.5 MB. These distances are vastly larger than the 200 or so kilobases that is the maximum extent of LD usually seen in larger populations [Peltonen et al., 1999; Varilo et al., 2004]. Averaging across the 44 population samples, the 29 mean LD values range from 0.02 to 0.04. Only 1 of the 1,276 LD values involved is ≥0.3; it is an LD value of 0.45 occurring for a 161.2 MB interval on chromosome 5.

Because there is no plausible biological explanation except by chance for expecting SNP alleles on different chromosomes or those far apart on the same chromosome to be associated only in a few small samples but not in the majority of samples, we conclude that all of these large LD values between distant or unlinked SNPs are chance deviations. Larger, independent samples from these populations will be necessary to confirm this but they are not currently available.

The accumulated evidence leads us to conclude that the 45 unlinked SNPs in our “final” IISNP panel are statistically independent at the population level.

In the last few months (summer of 2010) we have undertaken additional analyses of our existing 45-IISNP panel to explore other aspects of these SNPs. We have determined the number of SNPs showing identical genotypes for all pairs of individuals using our 45-IISNP panel and compared with the numbers for the same individuals but using a set of "random" SNPs (used in Pakstis et al. 2010 for other comparisons). The data are displayed in Figure 6-8. We have calculated the number of genotypes out of 45 that match in each possible pairwise comparison of all of the individuals that are typed for all 45 SNPs. The numbers in section A of Figure 6-8 are based on the 45 unlinked IISNPs and show results of all unique pairwise comparisons of individuals who are completely typed for the 45 IISNPs; results are shown for comparisons within populations, between populations, and total. Comparable analyses for 45 "random" SNPs typed on nearly all of the same individuals in the 44 populations are also shown in section B. Two points are noteworthy. First, the IISNP distribution for the number of loci (genotypes) matching is shifted towards lower numbers than the "random" SNP distribution, resulting in a larger number of sites that do not match. Second, the within population comparisons have a higher proportion of the loci matching than the between population comparisons. As shown by Pakstis et al. (2007) for a subset of these data, the pairs of individuals involved are almost exclusively in the smaller, more tribal, populations in which the samples undoubtedly contain at least second-degree relatives and in which the allele frequencies tend to deviate somewhat more from the ideal 50% heterozygosity. The highest numbers of loci matching occur for the populations that have undergone the largest amount of drift and/or are most likely to have complex relationships among individuals in the study.

We have also calculated the probabilities that a random individual would have a full sibling with an identical genotype at all 45 IISNPs. We used the specific allele frequencies in each population for each SNP. Those sibling match numbers fall in the range of 10^{-9} to 10^{-11} (Figure 6-9). This quantification is encouraging because of the question sometimes raised by defense attorneys that a close relative of a defendant with an exact match might also have an exact match. The low probability is in the range often found (or used) for a random match and is also relevant to the issue of pursuing relatives of an individual with a "partial match" (variously defined). These analyses are being included in a manuscript nearly finished for submission. In that paper we also call attention to the new resource we have developed as part of ALFRED, described in the following paragraph, and its relevance to the IISNP panel.

However much we believe in the universality of the 45-IISNP panel, additional empirical evidence from more populations is always useful. Therefore, since all of our data are in ALFRED, we have made a specific effort to accumulate additional data from the literature and public data repositories, e.g., HapMap, PopRes, HGDP, etc., on these and other sets of SNPs and add those allele frequency data to ALFRED. We have prototyped a "set" interface allowing users to access all 45 IISNPs and see how many populations are now typed for each marker and the current average heterozygosity and F_{st}. Figure 6-10 shows the SNP sets currently prototyped and Figure 6-11 shows the top of the web page for the 45-IISNP panel. However, though we have found additional data for all 45 IISNPs, the data for different individual SNPs are frequently based on different populations. (As of mid-January 2011, 48 of the 86 IISNPs had population frequencies in the ALFRED database for 64 to 101 different total populations representing a very substantial increase beyond those populations we published in early 2010. Another 43 of the 86 IISNPs have frequency data on 45 to 50 populations.) Therefore, additional summary analyses are not possible at this time due to the rather different sets of populations with frequency data across the IISNPs, but the accumulating data continue to support the universality of the set of IISNPs with high average heterozygosity and low Fst, though, of course, with additional data the values for each SNP change somewhat.

Figure 6-8. The distributions of unique pairwise comparisons of individuals for the number of genotypes matching for two datasets. A: the 45 unlinked IISNP set fully typed on individuals in 44 population samples showing all pairwise comparisons of individuals within the same population, all of those involving individuals in different populations, and the total. B: the same calculations for 45 random SNPs (Set #1 in Pakstis et al. 2010) on the same individuals in the same 44 populations.

A			
Number Genotype Matches	Within	Between	Combined
0	0	0	0
1 or 2	0	0	0
3 or 4	0	18	18
5 or 6	7	348	355
7 or 8	82	4150	4232
9 or 10	514	25346	25860
11 or 12	1974	94245	96219
13 or 14	5040	225443	230483
15 or 16	8933	362366	371299
17 or 18	10873	398947	409820
19 or 20	9307	308707	318014
21 or 22	5770	168386	174156
23 or 24	2731	64779	67510
25 or 26	929	18030	18959
27 or 28	342	3484	3826
29 or 30	121	477	598
31 or 32	39	31	70
33 or 34	12	4	16
35 or 36	3	1	4
37 or 38	1	0	1
39 or 40	0	0	0
41 or 42	0	0	0
43 or 44	0	0	0
45	0	0	0
Totals	46678	1674762	1721440

B			
Within	Between	Combined	Number Genotype Matches
0	0	0	0
0	0	0	1 or 2
0	0	0	3 or 4
0	3	3	5 or 6
0	47	47	7 or 8
0	703	703	9 or 10
6	5867	5873	11 or 12
51	29089	29140	13 or 14
271	92440	92711	15 or 16
1100	195830	196930	17 or 18
3010	287372	290382	19 or 20
5799	306656	312455	21 or 22
8045	243195	251240	23 or 24
8253	150954	159207	25 or 26
6513	75483	81996	27 or 28
4200	31188	35388	29 or 30
2332	10170	12502	31 or 32
1038	2611	3649	33 or 34
361	406	767	35 or 36
114	48	162	37 or 38
26	1	27	39 or 40
4	0	4	41 or 42
0	0	0	43 or 44
0	0	0	45
41123	1432063	1473186	Totals

Figure 6-9. Match probabilities for 45 unlinked IISNPs and for 86 IISNPs across 44 population samples for random pairings of individuals. Also plotted are the population specific probabilities for a match of the full sibling of a random person for 45 unlinked IISNPs only. The populations are grouped by geographical regions with Africa leftmost followed by SW Asia, Europe, South Central Asia, Siberia, Central Asia, Pacific Islands, East Asia, and the Americas.

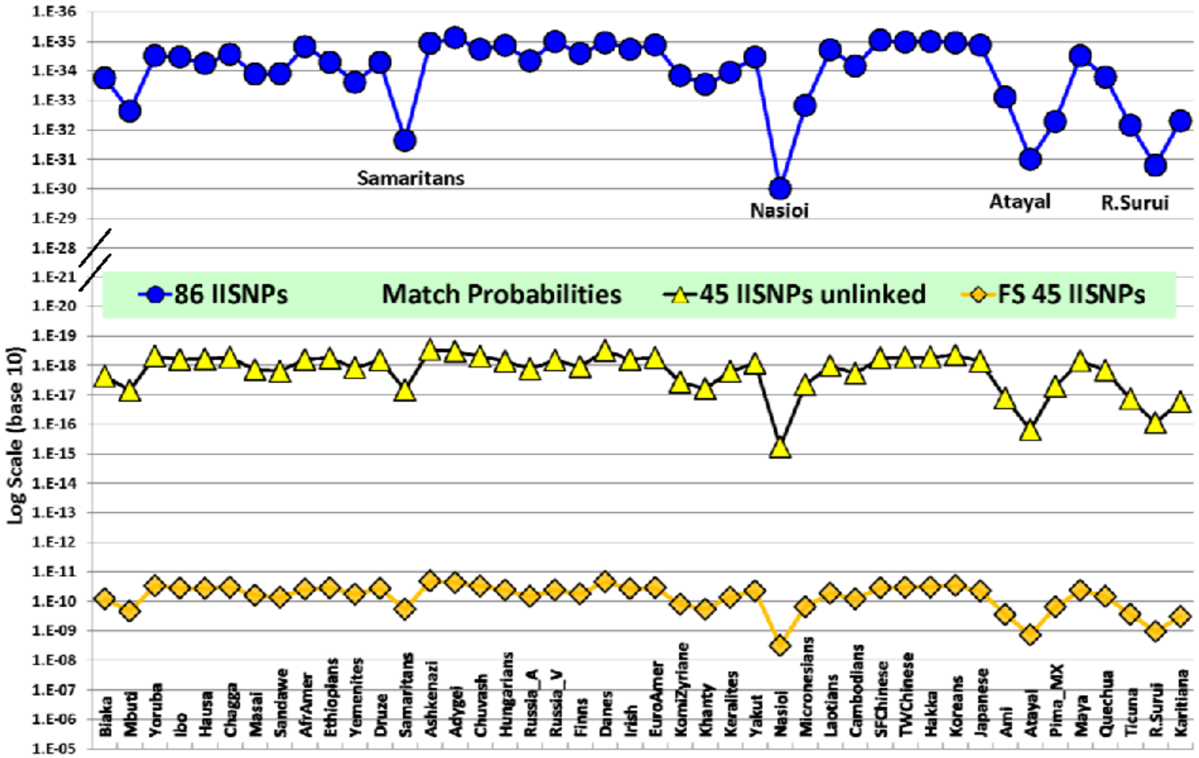
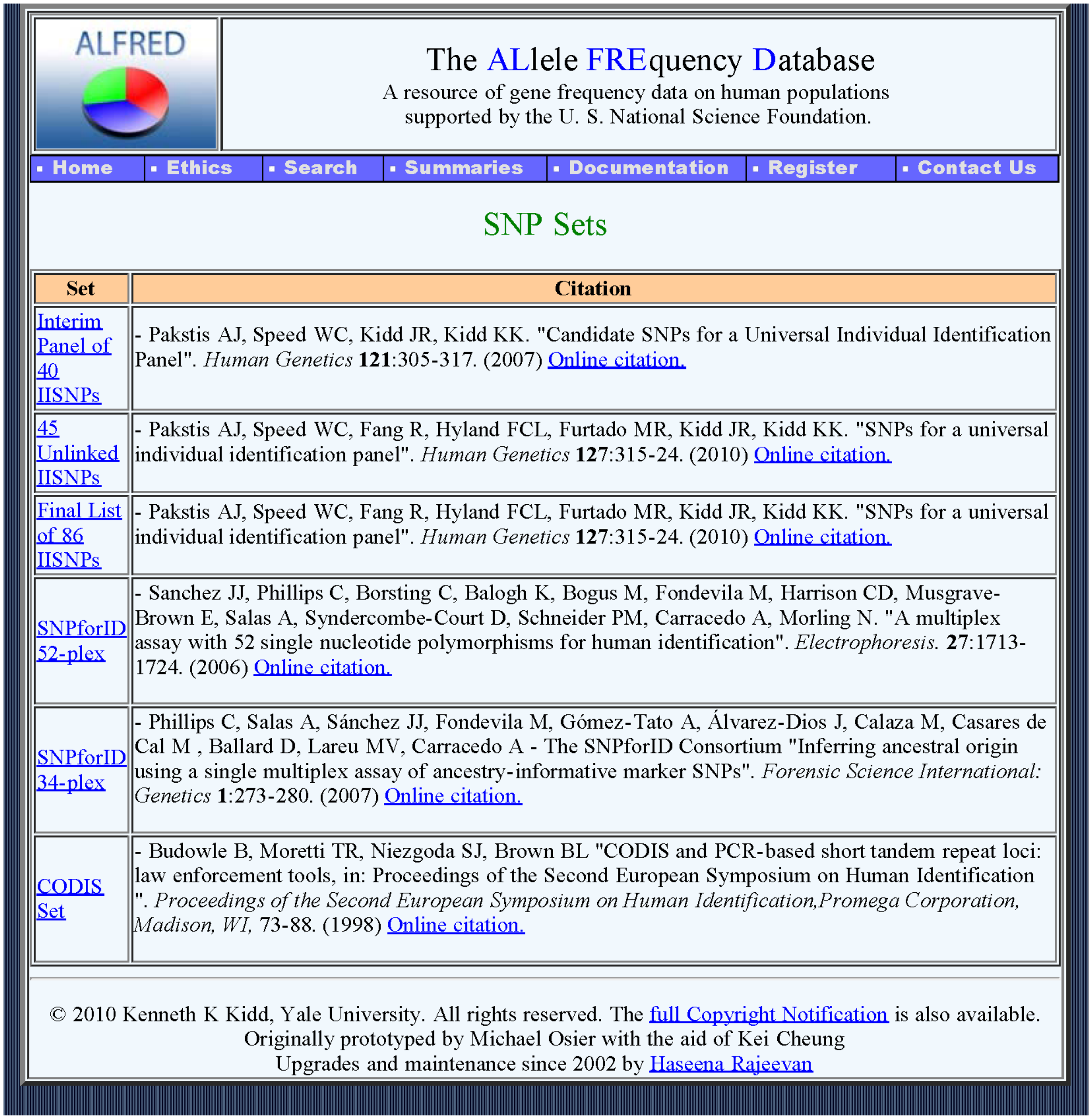


Figure 6-10. SNP-Set page in ALFRED (http://alfred.med.yale.edu/alfred/snpSets.asp). Clicking on the “set name” link under the “Set” column header brings up the full list of individual SNPs as shown in Figure 6-11.



The ALlele FREquency Database

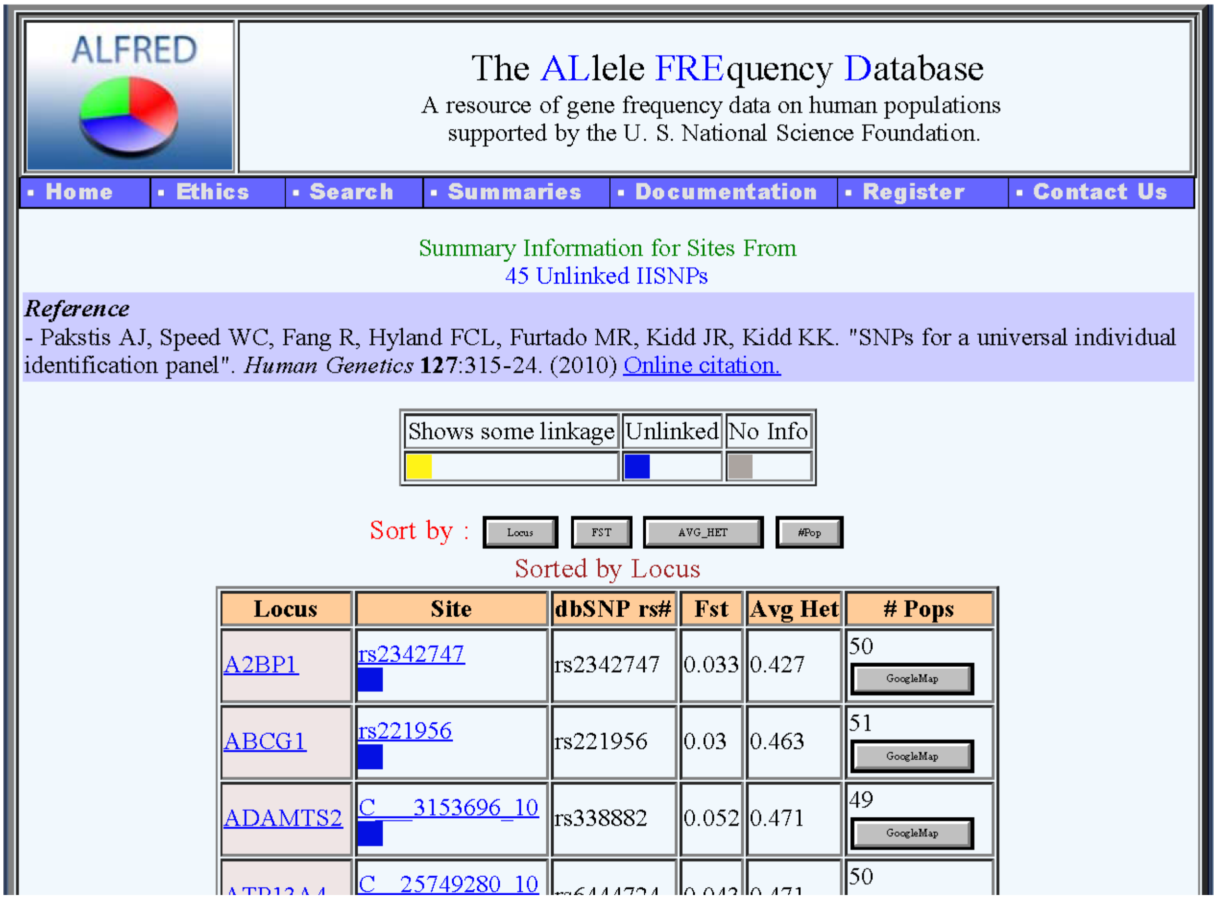
A resource of gene frequency data on human populations
supported by the U. S. National Science Foundation.

SNP Sets

Set	Citation
Interim Panel of 40 IISNPs	- Pakstis AJ, Speed WC, Kidd JR, Kidd KK. "Candidate SNPs for a Universal Individual Identification Panel". *Human Genetics* **121**:305-317. (2007) Online citation.
45 Unlinked IISNPs	- Pakstis AJ, Speed WC, Fang R, Hyland FCL, Furtado MR, Kidd JR, Kidd KK. "SNPs for a universal individual identification panel". *Human Genetics* **127**:315-24. (2010) Online citation.
Final List of 86 IISNPs	- Pakstis AJ, Speed WC, Fang R, Hyland FCL, Furtado MR, Kidd JR, Kidd KK. "SNPs for a universal individual identification panel". *Human Genetics* **127**:315-24. (2010) Online citation.
SNPforID 52-plex	- Sanchez JJ, Phillips C, Borsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N. "A multiplex assay with 52 single nucleotide polymorphisms for human identification". *Electrophoresis.* **27**:1713-1724. (2006) Online citation.
SNPforID 34-plex	- Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Álvarez-Dios J, Calaza M, Casares de Cal M , Ballard D, Lareu MV, Carracedo A - The SNPforID Consortium "Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs". *Forensic Science International: Genetics* **1**:273-280. (2007) Online citation.
CODIS Set	- Budowle B, Moretti TR, Niezgoda SJ, Brown BL "CODIS and PCR-based short tandem repeat loci: law enforcement tools, in: Proceedings of the Second European Symposium on Human Identification ". *Proceedings of the Second European Symposium on Human Identification,Promega Corporation, Madison, WI,* 73-88. (1998) Online citation.

© 2010 Kenneth K Kidd, Yale University. All rights reserved. The full Copyright Notification is also available.
Originally prototyped by Michael Osier with the aid of Kei Cheung
Upgrades and maintenance since 2002 by Haseena Rajeevan

Figure 6-11. The top of the list for the 45 IISNP panel is displayed as it appears in ALFRED along with current summary statistics. Clicking on the rs# link brings up the full information on the specific SNP. Clicking on the “Google Map” button under the “# Pops” column header brings up Google Map with all data plotted as pie charts in their ancestral geographic locations. As noted just above the table, this is the top of the 45 IISNPs when sorted by “Locus”; the other sorting options are noted.



The ALlele FREquency Database

A resource of gene frequency data on human populations
supported by the U. S. National Science Foundation.

Summary Information for Sites From
45 Unlinked IISNPs

Reference
- Pakstis AJ, Speed WC, Fang R, Hyland FCL, Furtado MR, Kidd JR, Kidd KK. "SNPs for a universal individual identification panel". *Human Genetics* **127**:315-24. (2010) Online citation.

Shows some linkage	Unlinked	No Info
🟨	🟦	▨

Sort by : Locus FST AVG_HET #Pop

Sorted by Locus

Locus	Site	dbSNP rs#	Fst	Avg Het	# Pops
A2BP1	rs2342747 🟦	rs2342747	0.033	0.427	50 GoogleMap
ABCG1	rs221956 🟦	rs221956	0.03	0.463	51 GoogleMap
ADAMTS2	C___3153696_10 🟦	rs338882	0.052	0.471	49 GoogleMap
ATP12A4	C___25749280_10	rs6444724	0.043	0.471	50

6.6 Assessment of what was accomplished to this point

In terms of the diversity of the populations on which data have been collected this study represents the largest single study to date to find SNPs with globally low F_{st} and high heterozygosity. The final panel of 45 SNPs has a narrow range for the average match probability across almost all populations. This validates the low F_{st}, high heterozygosity strategy for identifying SNPs that are appropriate for use in human identification. While F_{st} depends on the specific set of populations studied, it is clear that a global set of DNA samples needs to be used to screen for markers that have globally low F_{st} values. Also, our step-wise approach shows that the more different populations used to screen the more refined the result. A maximum global F_{st} of 0.06 functions well as a criterion even when small isolated populations are included. Similarly, because we also simultaneously selected for high heterozygosity, the globally low F_{st} reflects not just similar allele frequency but also uniformly high heterozygosity. The actual cause of the low F_{st} in the SNPs we screen is most likely that they are drawn from the lower tail of the distribution of F_{st} for random neutral SNPs. We are not aware of any phenotypic consequences either of any one of these 92 polymorphisms or of any polymorphism in linkage disequilibrium with any of the 92 SNPs. While this interesting possibility cannot be excluded, it is irrelevant to their use in individual identification.

The data from our step-wise screening also demonstrate an important fact relevant to extrapolating to a global level the allele frequency variation found in a smaller set of population samples. The F_{st} range for the 90,483 Applied BioSystems markers screened in the three populations we used for our original selection of candidate markers was 5.6×10^{-8} to 0.93, with mean = 0.087 and median = 0.063. As described in our initial publications [Kidd et al., 2006; Pakstis et al., 2007] we ended up with a final yield of less than 10% of the initial 436 SNPs chosen for the intermediate screen on seven populations. Following the 7-population screen we had roughly 50% yield of acceptable IISNPs when those 78 SNPs were tested on the full 40 populations. In this latest project we used much larger public databases to identify over 100 additional SNPs as candidates in chromosomal regions where we did not already have useful IISNPs. Screening these on our 44 populations had a yield of roughly 50%. The present results show the impossibility of accurately predicting or extrapolating the frequency distribution from one set of populations to another, even if both sets include populations from the four major continental regions. We believe our panel of populations is better than the others used because it is over twice as big as the next largest panel of populations, the HGDP-CEPH collection, and has comparable coverage of the world. We have also learned that, as high-throughput, genome-wide studies of SNPs are being done on diverse populations, multiple sets of SNPs meeting our IISNP criteria could be developed. However, it is unlikely that a significantly better set can be found. The extensive validation of the panel we have developed is more than sufficient for forensic use as it is and the effort at developing a new panel would not be cost effective or scientifically justifiable.

6.7 Some forensic considerations

The values in Figures 6-6 and 6-9 are calculated for ideal populations with no allowance for substructure. As noted by the NRC Committee (1996), the correction factor θ is equivalent to F_{st} for markers having Hardy-Weinberg genotype frequency ratios, as is the case for all our markers within each population. We assume that any correction factor for substructure within a large ethnically more homogeneous population will be small and not greatly alter the match probabilities for the large populations in Figure 6-6 (filled-circles). We note that the relationships of measures of within population substructure to the global F_{st} are not simple [Balding, 2003]. However, the similarity of allele frequencies globally greatly reduces the likelihood of substantial allele frequency differences among subgroups within an ethnically heterogeneous population. Moreover, by selecting for a globally low F_{st} we should also be reducing the likelihood of relevant substructure within each population. For these 45 unlinked loci the average "global" (44-population) F_{st} is 0.042**.** In an actual forensic application ignoring ethnicity one could use the global average allele frequencies (appropriately weighted from population-specific data available for these 45 SNPs in ALFRED) and the average global Fst as the value of θ used in standard forensic calculations (NRC Committee, 1996) to account for global substructure. Alternatively, and certainly valid, one could simply use the value of 10^{-15} as the largest value seen to date and that only in a very isolated population.

Candidate SNPs being considered for forensic applications need to be tested by several laboratories before being introduced into actual casework, both to demonstrate robustness of the methodology and to provide additional population data, especially on the samples commonly used as the basis for allele frequencies for the CODIS markers and hence already accepted in the courts. For a universally applicable panel to be accepted as such many additional populations should be tested and independent samples of those we have studied should be tested. Except for very small endogamous (tribal) populations it seems unlikely that very different allele frequencies will result for the 45 SNPs we have identified since we know from many years of data being accumulated on populations that allele frequencies tend to be similar in geographically close populations [Cavalli-Sforza et al. 1994; Rosenberg et al., 2002; Tishkoff & Kidd 2004; Kosoy et al., 2009]. The 44 populations studied here cover most major regions of the world; the regions not covered are flanked by those that have been studied. We would expect the F_{st} values to increase as more small, isolated populations are studied for these markers. Even so, the frequencies of the most common genotype and the average probabilities of identity are not likely to greatly exceed the largest seen for the 44 populations that we have studied since we have deliberately included some isolated populations from various parts of the world as a test of the robustness/generality of the results. Also important would be independent samples to show that the few large associations among markers observed are indeed the chance events they seem to be. That may be impossible for the very isolated populations such as the Nasioi because of the cost of a specific expedition as well as the problems of obtaining cooperation of a new group of individuals.

An important forensic aspect of any panel is the ability to apply the panel in actual forensic casework. We have collaborated with Drs. Manohar Furtado and Rixun Fang at Applied Biosystems [Fang et al., 2009]. They have developed two Gen-Plex panels that cover the 45 unlinked and the remaining 41 SNPs. They have also demonstrated the successful typing of all of these IISNPs on highly degraded DNA that will not allow results for the majority of the standard CODIS markers.

6.8 A universal panel

Our final set of 45 unlinked SNPs and the additional 47 SNPs, 41 of which have no significant LD among themselves or with any of the unlinked 45, has excellent characteristics that qualify it for being accepted as a universal panel for individual identification. The 45 unlinked IISNPs already yield match probabilities that come close to the theoretical average match probability of just under 10^{-19} for 45 "perfect" IISNPs, i.e., all with heterozygosity equal to 0.5. While our use of Fst <0.06 is arbitrary, it has proven to be very good at identifying markers with very similar allele frequencies in most populations. As more populations are typed, especially smaller and/or more isolated populations, some of these 45 SNPs may have less uniformly high heterozygosities. Certainly, their rank order is expected to change when any additional populations are considered; some of the SNPs with F_{st} just larger than 0.06 may end up better than those with F_{st} just smaller than 0.06. However, it is extremely unlikely that match probabilities for the 45 unlinked SNPs will exceed 10^{-12}, which is still a forensically very meaningful low value. Also, with 86 SNPs independent at the population level, some of which could be substituted for some of the 45 unlinked SNPs should technical (e.g., multiplexing) problems arise, we think that pursuit of additional IISNPs is not necessary.

Other SNP panels have been proposed for use in individual identification [e.g., Inagaki et al., 2004; Lee et al., 2005; Sanchez et al., 2006], but ours is the first to be based simultaneously on high heterozygosity and low F_{st} in a large global sample of populations. The SNPforID panel has been tested on the HGDP-CEPH panel and does give low, but highly varying, match probabilities because of considerable variation in heterozygosity among the populations. At least for European populations it is quite useful. However, we do not feel it qualifies as a "universal" panel--only four of those SNPs fall within our set of 92 and none within the 45 unlinked SNPs. (Note: uniformly high heterozygosity means that the F_{st} will be low but a low F_{st} does not mean a high heterozygosity, just a relatively uniform heterozygosity.)

We note that these 92 IISNPs meet one important criterion beyond the purely population genetic criteria: No medical or sensitive personal information is conveyed by the individual or combined data. To our knowledge these SNPs are not in a “gene” but what is a gene is an ongoing research issue as modern human molecular genetics continues to identify new types of functional elements in addition to conventional protein coding sequences. However, since these SNPs approach the ideal of 50% heterozygosity, close to the average of 37.5% of the global human population will share a randomly chosen genotype at any one of the loci. That minimizes the level of concern should some functional effect of one of these SNPs be determined in the future.

6.9 Some general implications of this study

Two especially interesting aspects of our screening results are (1) the large variation among SNPs in F_{st} value when additional populations were tested (Figures 6-1 and 6-2); (2) yet the relatively high yield of markers having both low F_{st} values and high heterozygosity when a large number of population samples was studied. Forensic researchers are reminded of the genetic diversity of the human species. The first point of interest also has implications beyond forensics for researchers interested in the search for balancing selection based solely on data for a small number of populations, such as is true for the HapMap data [The International HapMap Consortium, 2003, 2005]. The HapMap data are a very valuable resource but cannot be considered to represent the extent of global allele frequency variation very accurately. The second finding also has implications for the search for balancing selection in that there must be a very large number of such SNPs with low F_{st} and high heterozygosity. It is improbable that most would be maintained by balancing selection. Thus, it may be challenging to demonstrate unequivocally that balancing selection in humans against a background of such SNPs.

7. Progress on identifying Ancestry Informative SNPs (AISNPs)

We have been aiming to achieve greater specificity of ancestry identification than is generally provided by "continental" assignment or by forensic anthropology based on a skeleton and our desired level of specificity approaches the individual ethnic group or extended family (clan). Whether or not we can achieve a large likelihood ratio for distinction within a geographic region the size of Europe with only a few hundred SNP markers remains to be seen--this is a research project. However, our initial results show we can do much better than four "continental" populations and suggest we may be able to obtain a certainty of ancestry or clan membership strong enough to be useful to investigators. It will also be important to have associated probabilities of *incorrect* assignment, so investigators can readily understand the power of the panel.

Differentiation, on average, among even closely related groups or individuals within such groups (e.g., European populations) is possible if enough markers are used)—that is not the problem. The problem is identifying ancestry for a ***single individual*** with a ***reasonable*** number of SNPs. It is likely that the utilization of SNP haplotypes or SNPs combined with STRPs (SNPSTR; Mountain et al., 2002;

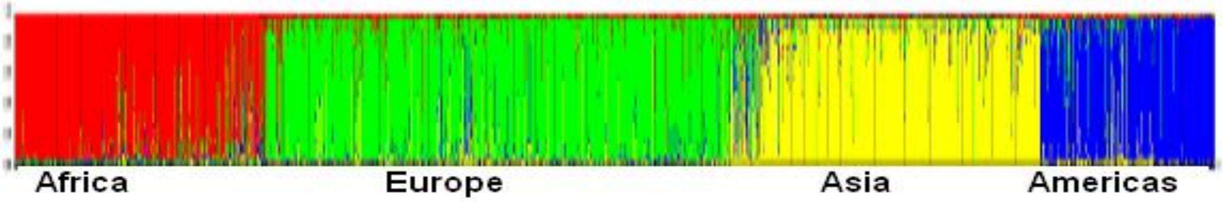
Ramakrishnan & Mountain, 2004) may help achieve the goal of maximizing accuracy of our prediction of ancestry.

7.1 Existing AISNP panels

We have identified several studies with a biomedical focus that present panels of AISNPs reportedly suitable for determining admixture levels in specific admixed populations (e.g., Jorde et al., 2000; Collins-Schramm et al., 2004; Shriver et al., 2005; Vallone et al., 2005; Yang et al., 2005; Enoch et al., 2006; Lao et al., 2006; Tian et al., 2006, 2007, 2008, 2009; 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; Nassir et al., 2009). Other studies have examined the HGDP-CEPH dataset on the Illumina Hap650Y bead array and have shown there is considerable information on population relationships in the ~650,000 SNPs but have not published lists of those SNPs providing the most information (Paschou et al., 2007; Biswas et al., 2009). These panels provided a resource for our development of an AISNP panel, and we have made a very strong start on developing a panel of high Fst SNPs as an investigative tool; we have progressed beyond simple resolution at the "continental" level and are developing criteria for evaluating the quality of a panel of AISNPs. SNPs have already been shown to allow the easy (though fairly rough) resolution of the four continental groups with as few as 10 SNPs [Lao et al., 2006]. This set of ten SNPs was one of our starting points for developing a much more comprehensive AISNP panel. We began by typing the 10 SNPs from Lao et al [2006] on our then 40 populations. Their analyses on the HGDP-CEPH panel (and their 10

SNPs on our 40 populations, Figure 7-1) of those markers did not allow any further subdivision of populations even when regions were examined separately using the program STRUCTURE [Pritchard et al. 2000; Falush et al., 2003]; we felt more could be accomplished.

Figure 7-1: STRUCTURE solution at K=4 clusters for 40 populations from our lab with Lao et al. (2006) 10-SNP set. This is one of the early efforts at an AISNP panel with a small number of SNPs.



7.2 Targeting the selection of SNPs for ancestry inference

We initially sought appropriate markers for robustly resolving geographic and population structure using four sources: (1) high Fst markers identified in the Celera or HapMap databases, (2) the ten markers published by Lao et al. [2006] summarized above, (3) the markers identified for the Kim et al. [2005] study as having a very large difference between Chinese and Japanese allele frequencies, and (4) markers from our studies that have above average Fst within each region. Kosoy et al. [2009] published an admixture panel consisting of 128 SNPs with reasonably good resolution. One of our first steps was to type these same SNPs on our populations. As the results were

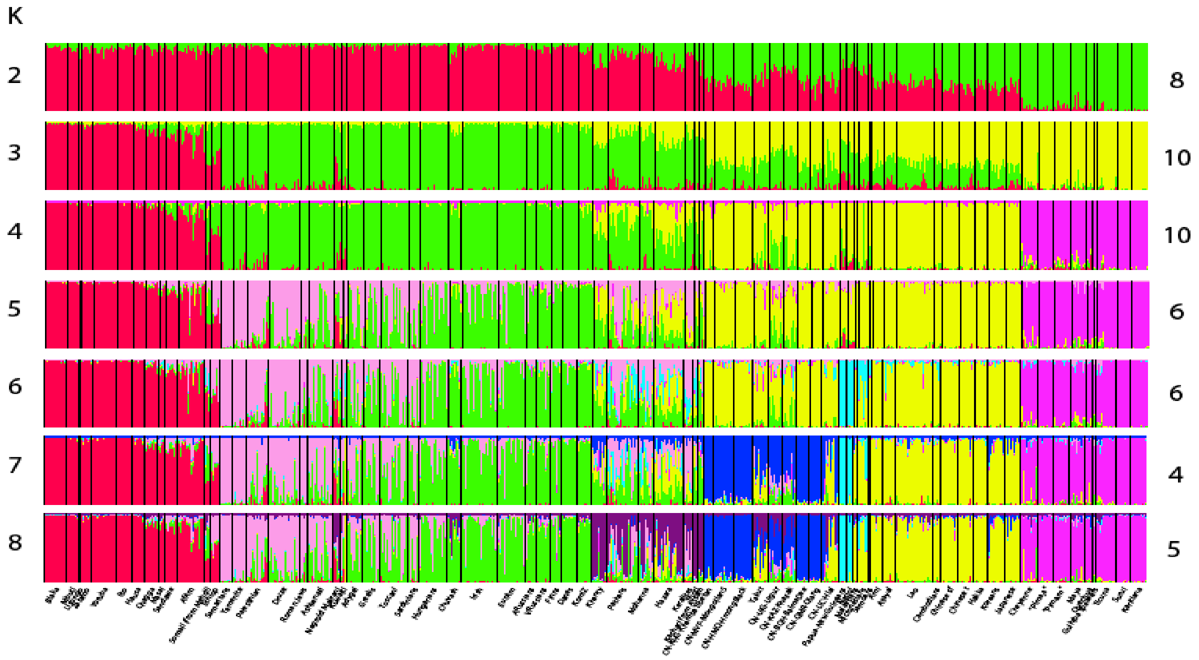


Figure 7-2. Preliminary STRUCTURE analyses for 73 populations using nearly complete data for the 128 SNPs of Kosoy et al [2009]. Eight clusters can be resolved with reasonable correspondence to geographic origins of the populations. Populations are in the order of Table 7-1. The numbers on the right indicate the number of times the most common pattern (shown) occurred in 10 independent replicates of that K value.

good, we extended the analysis to 73 populations. Figure 7.2 is a preliminary analysis on populations from our lab, including both those population samples with small amounts of DNA available and those from cell lines from our lab (Table 7.1). This figure gives the correct impression that at larger K values it becomes more difficult to cleanly assign individuals to a cluster, particularly for populations that are geographically intermediate (Southeast Europe between Western Europe and Southwest Asia) or admixed (African Americans).

We have used the results summarized in Figure 7-2 to identify regions that needed better discrimination, such as Europe and Southwest Asia, Southwest Asia and Southern Asia, Western China and more eastern East Asia. In order to explore those regions further, we have extended the number of populations from 73 to 119 to include more geographically intermediate populations and have typed the 128 SNPs of Kosoy et al. [2009] on all the samples. We obtained the data on the additional 46 populations from the literature (HapMap and J. Li et al., 2008). Table 7-2 lists all 119 population samples we analyzed and summarized in Figure 7-4 with the number of individual DNA samples in each population and the source.   Figure 7-4 shows STRUCTURE analyses up to K=8 using these populations and the 128 SNPs.

Table 7-1. 73 Population samples from Kidd lab used for the preliminary analysis of 128 admixture SNPs from Kosoy et al [2009] shown in Figure 7.2.

Geographic Region	Name	N	IISNP 44 pops	Geographic Region	Name	N	IISNP 44 pops
Africa	Biaka, C.A.R. *	70	X	S.C.Asia	Hazara, Pakistan	96	
	Mbuti, D.R.Congo *	39	X		Keralites, S.India	30	X
	Lisongo	8			Thoti, Andhra Pradesh	14	
	Zaramo, Tanzania	40			Kachari, Assam	18	
	Yoruba, Nigeria *	78	X	C.Asia	CN-KHG-KhambaTibetan	31	
	Ibo, Nigeria	48	X		CN-MVF-Mongolians	64	
	Hausa, Nigeria	39	X		CN-HMQ-HmongBlack	59	
	Chagga, Tanzania	45	X		Yakut *	51	X
	Masai, Tanzania	22	X		CN-UIG-Uigur	47	
	Sandawe, Tanzania	40	X		CN-KAZ-Khazak	48	
	AfrAmericans	90	X		CN-BQH-BaimaDee	42	
	Somali	22			CN-QMR-Qiang	40	
	Ethiopian Jews	32	X		CN-LIC-Hlai	59	
S.W.Asia	Samaritans	41	X	W.Pacific	Papua-New Guineans	22	
	Yemenite Jews	43	X		Nasioi, Melanesia *	23	X
	Palestinians	69			Malaysians	11	
	Druze *	† 127	X		Micronesians	37	X
	Kuwaiti	16			Samoans	8	
Europe	Roman Jews	27			Ami, Taiwan	40	X
	Ashkenazi	83	X		Atayal, Taiwan	42	X
	Adygei *	54	X	E.Asia	Laotians	119	X
	Greeks	56			Cambodians *	25	X
	Toscani, Italy	89			Chinese, SFB *	60	X
	Sardinians	35			Chinese, Taiwan	49	X
	Hungarians	† 145	X		Hakka, Taiwan	41	X
	Chuvash	42	X		Koreans	54	X
	Irish	118	X		Japanese *	51	X
	EuroAmericans	92	X	N.America	Cheyenne	56	
	Russians, Archangelsk	34	X		Pima, Arizona	51	
	Russians, Vologda *	48	X		Pima, Mexico *	† 99	X
	Finns	36	X		Maya, Yucatan *	52	X
	Danes	51	X	S.America	Quechua, Peru	22	X
N.W.Asia	Komi Zyriane	47	X		Guihiba speakers,	13	
	Khanty	50	X		Ticuna	65	X
S.C.Asia	Pathans, Pakistan	111			Rondonian Surui *	47	X
	Negroid Makrani	27			Karitiana *	57	X
	Mohanna, Pakistan	51					

Legend: * Samples (usually a subset) contributed to the HGDP-CEPH panel in Paris
† Samples with many related individuals; most analyses only include unrelated individuals

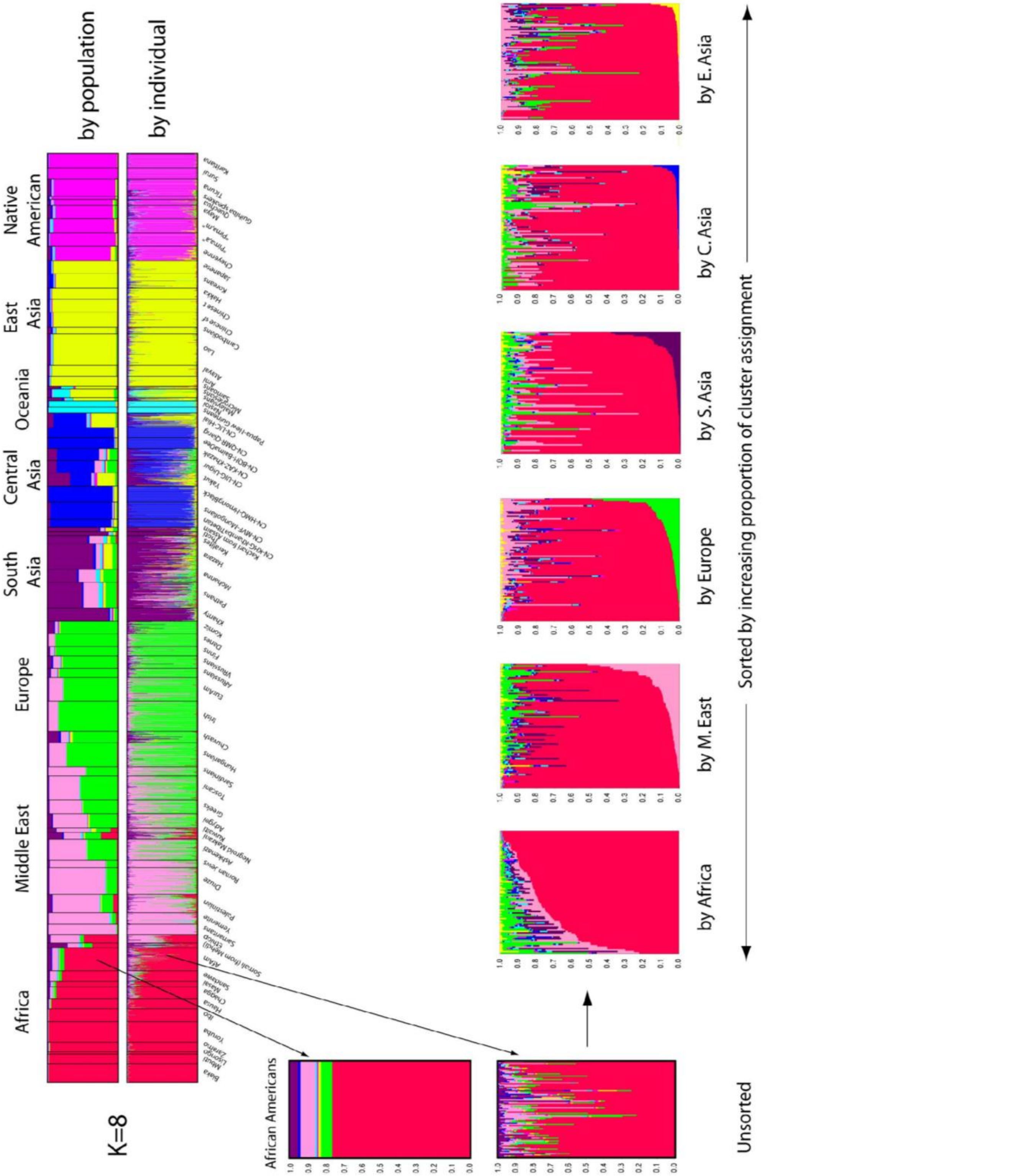

K=8
by population
by individual
Africa
Middle East
Europe
South Asia
Central Asia
Oceania
East Asia
Native American
African Americans
Unsorted
by Africa
by M.East
by Europe
by S.Asia
by C.Asia
by E.Asia
Sorted by increasing proportion of cluster assignment
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Figure 7-3. An elaboration of the K=8 solution in Figure 7-2. At the top, the proportional cluster assignments are shown both averaged by populations and for each individual. At the side and across the bottom the cluster assignments for African Americans are shown in greater detail. Individuals are sorted by amount of assignment to each of the six major clusters showing partial assignments. No African American individuals have appreciable assignments to Oceania and Native American Clusters. .

While these 128 SNPs are clearly useful for determining admixture, they are not necessarily good for identifying ethnicity for an unknown sample coming from an admixed population. The elaboration in Figure 7-3 of the African American sample from the K=8 STRUCTURE analysis (in Figure 7-2) illustrates this. The population averages are plotted across the top for all 73 populations; many populations show significant fractions of multiple clusters. This is evident for the African Americans, all of whom are self-identified African Americans in the Coriell cell line collection. On average, about 25% of the African American sample shows non-African signal (upper left enlargement). However, when individuals are considered (lower left and bottom enlargements) there is extensive variation. When individuals are sorted by probability of individual assignment to different “geographic-ethnic” clusters, the variation can be seen to be considerable. Several of the individuals are more likely to be considered non-African than African.

Table 7-2: 119 Population samples from Kidd et al. (2011) with number of DNA samples for each population and origin of data. These samples (from the literature and our lab) were typed on the Kosoy et al. [2009] set of 128 admixture SNPs. See Figure 7-4 for Structure results.

Population	Abbrev	N	Source of data
Biaka	BIA	67	Yale*
Mbuti	MBU	39	Yale*
Mandenka	MND	24	HGDP*
Lisongo	LSG	8	Yale
Yoruba	YOR	77	Yale
YorubaYRI	YRI	113	HapMap*
Ibo	IBO	48	Yale
Zaramo	ZRM	36	Yale
Hausa	HAS	39	Yale
Bantu_N.E.	BTN	12	HGDP*
Bantu_S	BTS	8	HGDP*
San	SAN	6	HGDP*
Luhya LWK	LWK	90	HapMap
African Amer 1	AAM	90	Yale
African Amer ASW	ASW	56	HapMap
Chagga	CGA	45	Yale
Maasai, T	MAS	20	Yale
Maasai MKK	MKK	144	HapMap
Sandawe	SND	40	Yale
Ethiopian Jews	ETH	32	Yale
Somali	SML	12	Yale
Mozabite	MOZ	30	HGDP*
Kuwaiti	KWT	16	Yale
Samaritans	SAM	40	Yale
Yemenite Jews	YMJ	42	Yale
Palestinian 1	PLA-1	49	Yale
Palestinian 2	PLA-2	51	HGDP*
Druze 1	DRU-1	75	Yale
Druze 2	DRU-2	47	HGDP*
Bedouin	BDN	48	HGDP*
Roman Jews	RMJ	26	Yale
Adygei	ADY	54	Yale*
Greeks	GRK	53	Yale
Ashkenazi Jews	ASH	79	Yale
Tuscan 1	Tus	8	HGDP
Tuscan TSI	TSI	88	Hapmap
Sardinian 1	SRD-1	34	Yale
Sardinian 2	SRD-2	28	HGDP
Orcadian	ORC	16	HGDP
North_Italian	ITN	13	HGDP
French_Basque	FRB	24	HGDP*
French	FRN	29	HGDP
Hungarians	HGR	89	Yale
Irish	IRI	114	Yale
European Amer 1	EAM	89	Yale
European Amer CEU	CEU	115	HapMap*
Russians 1	RUA	33	Yale
Russians 2	RUV	47	Yale*

Finns	FIN	34	Yale
Danes	DAN	51	Yale
Komi Zyriane	KMZ	47	Yale
Chuvash	CHV	42	Yale
Makrani 1	MKR-2	26	Yale
Makrani 2	MKR-1	25	HGDP
Kalash	KLS	25	HGDP*
Brahui	BRH	25	HGDP
Balochi	BCH	25	HGDP*
Sindhi	SDI	25	HGDP
Keralite	KER	30	Yale
Thoti	THT	14	Yale
Kachari	KCH	17	Yale
Gujarati GIH	GIH	88	HapMap
Pathan 1	PTH-1	75	Yale
Pathan 2	PTH-2	23	HGDP
Mohanna	MHN	48	HGDP
Burusho	BSH	25	HGDP*
Khanty	KTY	50	Yale
Hazara 1	HZR-1	87	Yale
Hazara 2	HZR-2	24	HGDP
Uygur 2	UYG	10	HGDP*
Uygur 1	UIG	45	Yale
Khazak	KAZ	44	Yale
Khamba Tibetan	KHG	27	Yale
Mongolians 1	MVF	62	Yale
Mongolians 2	MGL	10	HGDP*
HmongBlack	HMQ	46	Yale
BaimaDee	BQH	40	Yale
Qiang	QMR	38	Yale
Hlai	LIC	47	Yale
Yakut	YAK	51	Yale*
Dai	DAI	10	HGDP
Lahu	LHU	10	HGDP*
Miaozu	MIZ	10	HGDP
Naxi	NXI	9	HGDP
Oroqen	OQN	10	HGDP
She	SHE	10	HGDP
Tu	TU	10	HGDP
Tujia	TUJ	10	HGDP
Xibo	XBO	9	HGDP
Yizu	YIZ	10	HGDP
Daur	DUR	9	HGDP*
Hezhen	HEZ	9	HGDP
Han, S.F.	HAN	43	HGDP
Han CHD	CHD	85	HapMap
Han CHB	CHB	84	HapMap*
Han, Taiwan	CHT	50	YALE
Hakka	HKA	41	YALE
Koreans	KOR	54	YALE
Japanese	JPN	50	YALE
Japanese JPT	JPT	86	HapMap*
Laotians	LAO	118	YALE

Cambodians	CBD	24	YALE*
Ami	AMI	40	YALE
Atayal	ATL	42	YALE
Malaysians	MLY	11	YALE
Micronesians	MCR	34	YALE
Samoans	SMO	8	YALE
P-NG 1	PNG	13	YALE
P-NG 2	PNG	17	HGDP*
Nasioi	NAS	22	YALE
Mexican Amer MEX	MEX	49	HapMap*
Pima Mexico	PMM	53	YALE*
Maya	MAY	51	YALE*
Quechua	QUE	22	YALE
Colombians	COL-2	13	HGDP*
Guihiba	COL-1	11	YALE
Ticuna	TIC	65	YALE
Surui R	SUR	45	YALE
Karitiana	KAR	55	YALE

*These or subsets of these samples were included in Nassir et al (2008).

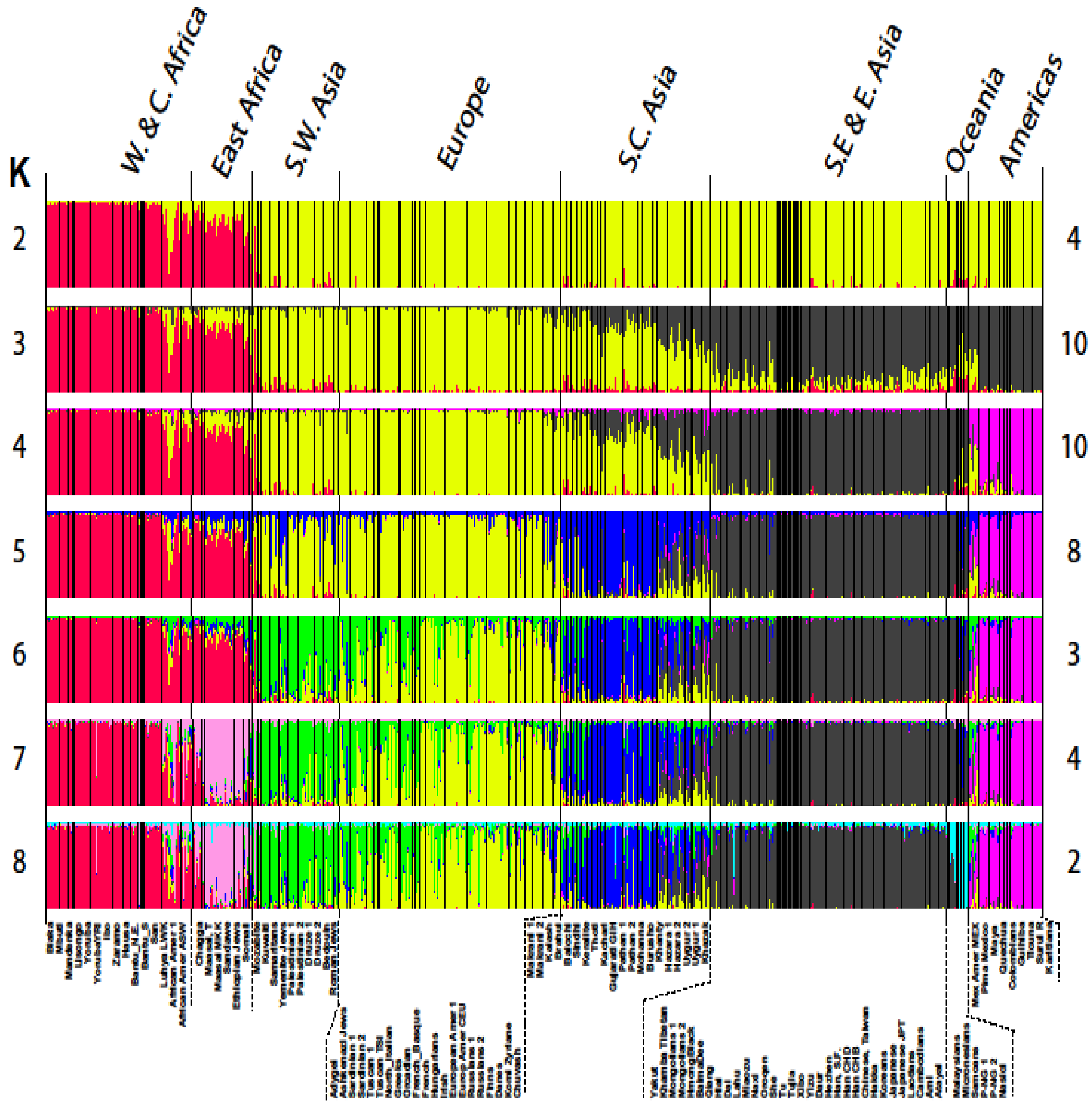


Figure 7-4. AISNP Structure analysis of 119 population samples (Kidd et al., 2011) with 128 SNPs from Kosoy et al. (2009) based on 10 independent Structure runs.

We have begun analysis of a final panel of 40 AISNPs developed by Carolyn Nievergelt at UCSD but as yet unpublished by her. We have typed 66 of our lab populations for these SNPs and have just begun analyses. Unfortunately we have begun to run out of DNA on some of the samples that we had available for the 73 SNP analysis accounting for the decrease in the number of population samples we are able to routinely analyze.

Finally, we have continued to search the many available databases with populations that are in two or more such regions to identify the SNPs that have the greatest allele frequency differences between the two. We believe that such SNPs will greatly improve these distinctions. In the latter quest we have also examined the 100 “best” SNPs identified in the HGDP-CEPH dataset using the Sampson et al. [2008, 2011] greedy algorithm on a subset of populations including the African and Native American populations. Not only can those regions be separated from all others in a global analysis but also when population subsets are analyzed individual populations in Africa and in the New World can be distinguished.

Cumulatively, we have assembled a larger set of about 430 high Fst SNPs, including those with very region-specific patterns of variation, typed on 55 populations, we believe we can use the Sampson et al. [2008, 2011] algorithm, the heatmap method illustrated in Figure 3-3, and PCA to eliminate duplication of information provided by different SNPs and thereby reduce the size of an AISNP panel while maintaining the key information on ancestry.

There will continue to be individuals and populations that show significant non-zero probabilities of belonging to more than one cluster. That is not strictly evidence of admixture (though admixture could be a cause), but rather indicates that the SNPs being used have intermediate allele frequencies in those populations as expected for a clinal distribution. It is expected that individuals will vary in their level of admixture but it is highly unlikely that all of the partial cluster assignments seen for individuals actually represent those levels of admixture of those ancestries. Kalinowski et al. [2010] discusses various limitations in the interpretation of the STRUCTURE program results with emphasis on finding the appropriate cluster number, the influence of sample size, and the artifactual nature of some partial cluster assignments.

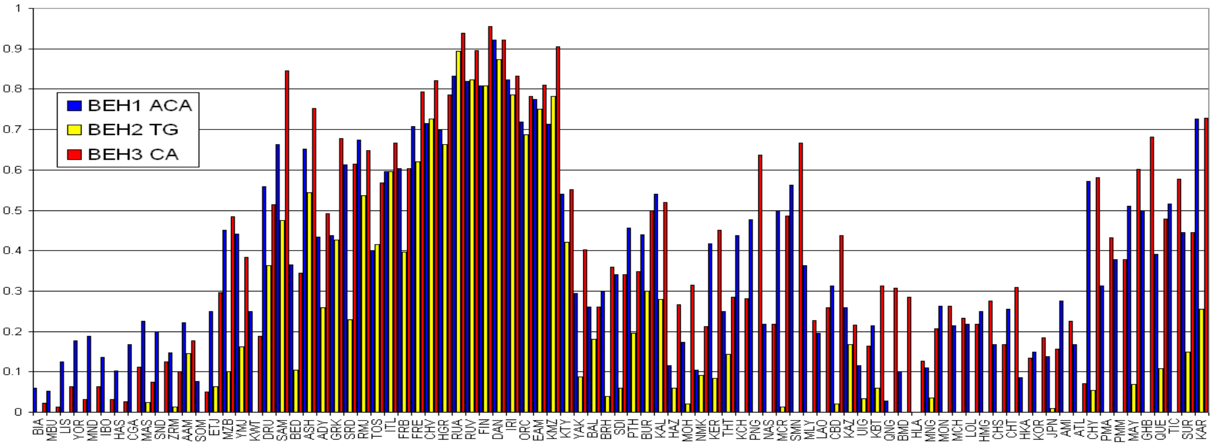
7.3 Optimizing SNP sets for intra-regional comparisons

An approach to the identification of appropriate SNPS for an AISNP panel is that of Sampson et al. [2008, 2011]. He has developed a statistical approach to identifying ancestry informative markers. Development of those statistics and their application to our data were not funded by our NIJ grants but clearly the data collection was and the results are relevant. The initial analyses have helped identify the most informative of the markers in our AISNP working data set. As part of his ongoing methodological research, Dr. Sampson has been analyzing the HGDP SNP data reported in Li et al. [2008]. As a result we are now able to identify some SNPs that are especially discriminating among groups in that panel. At the moment we are working to identify SNPs in those data that help resolve the clinal distribution that exists from Western Europe through the Middle East and Pakistan to India. Very preliminary results indicate that this method is the best we have examined so far for selecting a limited number of SNPs that will show the greatest ability to recognize *intra*-continental ancestry. Using the HGDP data published in Li et al. [2008], he has identified 100 SNPs from the nearly 650,000 SNPs reported.

We have found that previous studies have not evaluated statistically the precision with which individuals known to belong to a "cluster" are assigned to that cluster. This is clearly an important question to consider for the use of AISNPs as an investigative tool. Our statistical approaches to that question show that different sets of SNPs can vary greatly in that aspect and yet be quite robust in appropriate assignment of individuals. It is clear, however, that individuals from geographically intermediate populations often have partial assignments that are difficult to distinguish from admixture. Consequently, to be able to address this issue, we have increased our "standard" set of populations from 44 to 55—those populations for which we have cell lines in our lab. With reasonable clarity we have shown that we can distinguish eight clusters of populations (Figures 7-2 and 7-3). However, individuals from Southwest Asian and Central Asian populations are not cleanly identifiable as members of their "cluster" Consequently, we are examining SNPs typed on other sets of populations, primarily on the CEPH-HGDP panel, for SNPs that show large allele frequency variation across Eurasia.

Many other analyses are ongoing. We have shown that the HapMap Mexicans are highly admixed and not representative of other Native Americans as seen in the 119 population sample (Figure 7-4) at K=3 through K=8. We are extending that finding by a regional analysis of Native American populations using an enlarged set of SNPs. We also have a regional analysis underway for Europe using several of the "European-specific SNPs" (Figure 7-5).

Figure 7-5. The global frequencies of three different haplotypes at OCA2 that have been associated with blue eye color. All have high frequencies in northern Europe and show a north-south cline, but only the BEH2 haplotype corresponds to the global pattern of the blue-eyed phenotype.



Accomplishments

7.3.1. Data sets developed

We currently have three AISNP datasets: (1) a superset of 430 SNPs, selected because our studies indicated they were high Fst globally and each of them has been typed on at least the 45-55 core populations; (2) the Seldin 128 AISNPs (a subset of the 430 SNPs) with allele frequencies on 119 population samples including 73 populations typed in our lab; and (3) the Nievergelt unpublished set of 40 AISNPs (also a subset of the 430) typed in our lab on 63 populations. As noted above our total set of AISNPs includes both markers that have been published by other authors and those we have selected as candidate SNPs. Rarely do SNPs published in various papers overlap, but often markers we (or others) identified are very close together and have nearly complete LD giving virtually identical information on population relationships.

7.3.2. Analyses

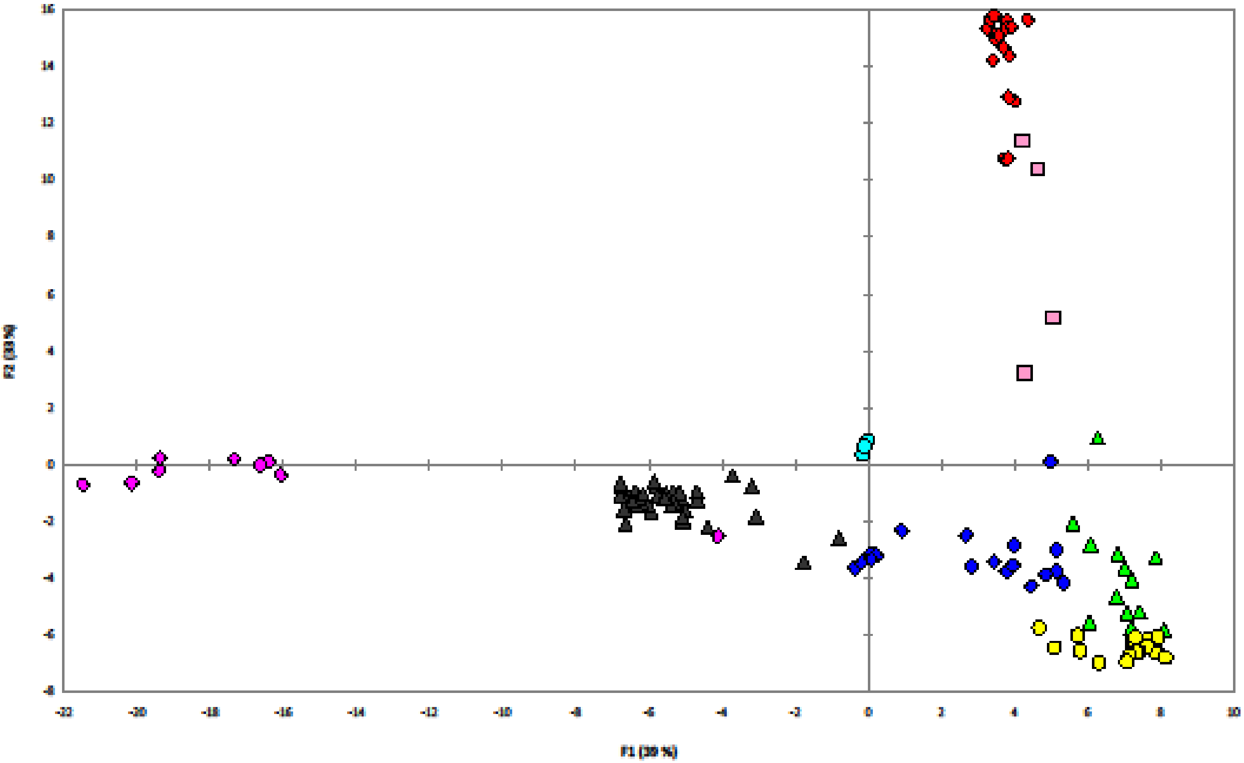
In the process of preparing to submit funding proposals during the period covered by this progress report we also carried out experiments and analyses of relevance to our ancestry informative marker project. Specifically, (1) we were able to demonstrate that the LD Block strategy for Lineage Informative SNPs advocated by Ge et al., (2009) has serious flaws. Along the way we identified an alternative method in a pilot study that will be more productive. (2) We have also carried out a pilot resequencing study to identify region/haplotype specific SNPs that increase the resolution of determining population membership within geographic regions. (3) We have additional analyses of the pigmentation phenotype marker OCA2 that identify the subset of the several markers in strong LD in Europe which could be the most useful for identifying ancestry as well as the blue-eye phenotype (Figure 7-5).

7.3.3. AISNP Progress

Our goal has been fine geographic resolution of ancestry. We have accumulated an AISNP dataset of over 400 SNPs that we have identified as globally informative (using F_{st}, Informativeness, and other statistics) and typed on a minimum of 45 of our previous populations. We have already tested the best SNPs from several of the multiple ancestry informative panels on the set of 55 populations on which we have unlimited amounts of DNA (from our cell lines) and we will be bringing more of these SNPs up to 57 populations. We are similarly expanding the number of populations typed for the best of the 3000+ we previously identified from our other research projects. We have expanded the populations being typed on the Seldin group's 128 marker set (Kosoy et al. (2009)) that we have been working on for some time. Sufficient analyses have been generated with these 128 SNPs that we have completed analyses and published a manuscript (Kidd et al., 2011) summarizing the most important findings thus far. Very recently we began typing our population samples on an additional AISNP panel consisting of 40 SNPs (Caroline Nievergelt's panel) which is largely independent of the data already collected. Those data are currently being analyzed in collaboration with Dr. Nievergelt. We plan to identify the best integrated subset of all markers and by the end of the current project we shall have completed analyses of a panel combining the 128 AISNPs now in the *Investigative Genetics* paper (Kidd et al., 2011) with the 40 highly informative AISNPs identified by C. Nievergelt along with some of our other AISNPs identified as informative in particular regions of the world.

Our recent analytic efforts have focused on extending one good set of AISNPs to a very large number of populations to serve as a base-line for improvement. Figure 7-6 presents our recent results for the first 2 factors of a Principal Components Analysis on 4,873 individuals in 119 population samples using the 128 admixture SNPs of the Seldin group (Kosoy et al., 2009) (Figure 7-6). In addition to the 73 populations

Figure 7-6: Principal Components Analysis (PCA) of data on 119 populations (N=4,873 individuals) including samples from the HGDP (Cann et al., 2002), the HapMap III, and our own lab. The sites used are the 128 sites published by Seldin's group (Kosoy et al., 2009; Nassir et al., 2009) as a set of AISNPs for the identification of admixture in the common populations in the U.S. Because of the way these sites were selected, it is clear that the first factor strongly distinguishes among Europeans and Native Americans, and the second factor distinguishes among Europeans and Africans. The colors are the same as those in the Structure Figure 7-4.



typed in our lab [see Table 7-2], we imported the data on these markers from the HGDP dataset (Li et al., 2008) and from the 11 HapMap III populations. Our results extend the Kosoy et al. findings and validate these SNPs for somewhat finer resolution (Table 7-2 and Figures 7-2, 3, & 4.

We are in the process of integrating all of our candidate AISNPs into a single data set so that we can identify the “best” subset, eliminating those of less value and those that are completely redundant. This integration will entail including the “best” of the 3000+ SNPs typed on our populations for other, non-forensic projects in our lab. For example, we have recently shown that rs671 at *ALDH2* varies greatly within East Asia (Li et al., 2009) and is fixed elsewhere. Similarly, we have also observed on our samples that rs12203592 at *IRF4* varies considerably across Europe and the Middle East and is essentially fixed elsewhere. We recently published the worldwide distribution of the 17q21 inversion haplotype [Donnelly et al., 2010]; this inversion can be identified reliably by typing a small number of SNPs (3) making it a useful ancestry informative marker. None of these SNPs has been incorporated into the AISNP analyses yet. Our overall objective will be to identify markers that will as much as possible clarify additional clusters. We are striving for a universal panel of AISNPs, but these clarifications are also specifically areas of forensic relevance within the United States given our increasingly heterogeneous population.

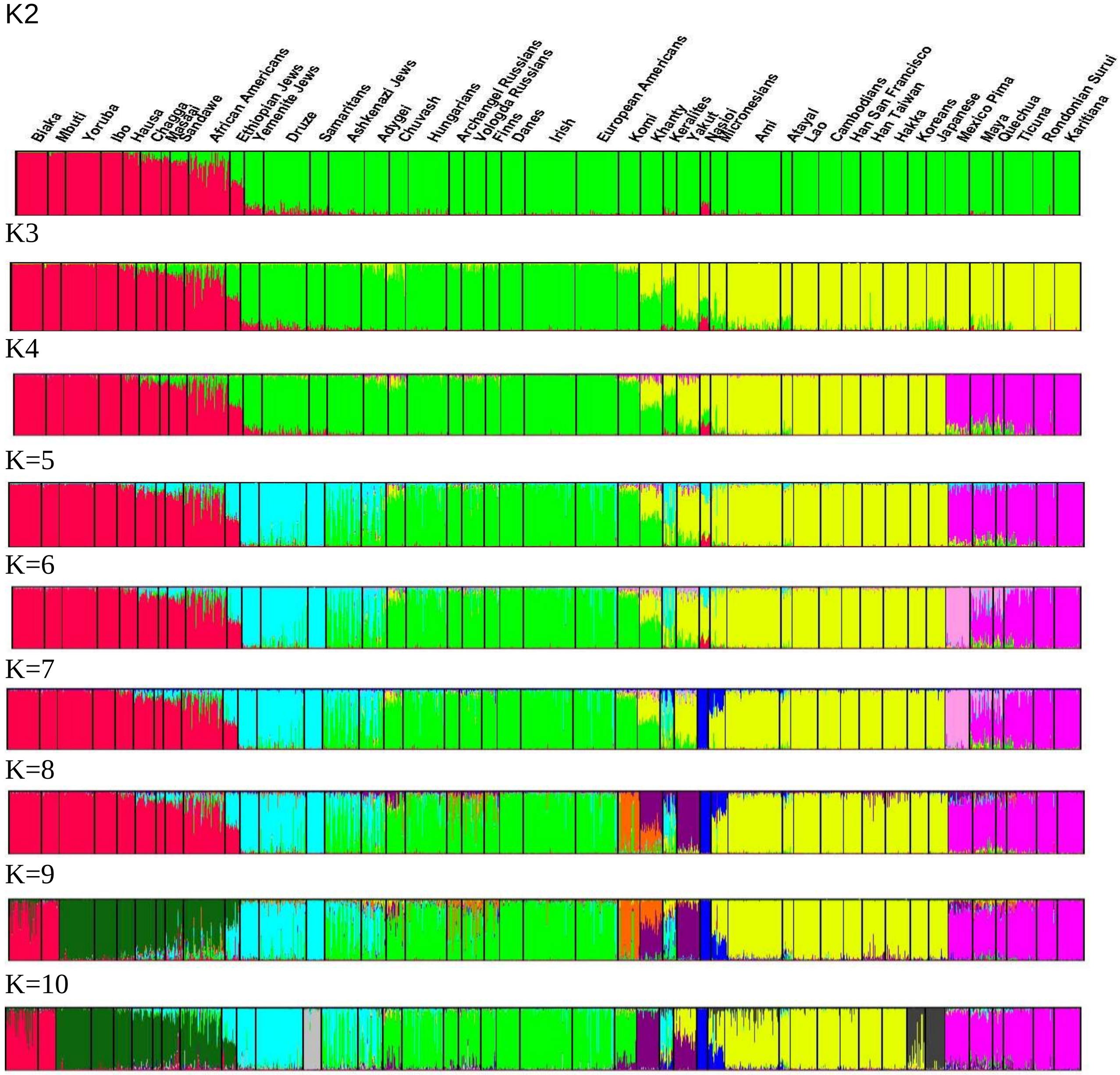
7.3.4. Ongoing enhancement of AISNP data

In reviewing our previous AISNP data, we have determined that it is useful to maximize the number of population samples. Thus, we now routinely genotype 55 population samples, and, where HGDP-CEPH and/or HapMap data are available, we will integrate our data with those data. Our results to date and the existence of multiple, non-overlapping sets of AISNPs in the literature suggest that different subsets will be nearly equal in some distinctions but differ in others. Dr. Christine Nievergelt at UCSD has made available an unpublished set of 40 AISNPs that includes several SNPs not previously identified in our lab or in the literature; these SNPs provide excellent separation of certain geographic regions and we will type those SNPs on our samples. Ultimately, it may be that no single small panel will be optimal for all questions. It now seems clear that we will be able to continue into the proposed project to make improvements in the fine distinctions among geographic regions as ancestry for an individual.

One approach to further differentiation of specific geographical regions will be through resequencing efforts as part of other projects in this lab in which we focused on finding region-specific variation. For example, we have identified a new SNP that has moderate heterozygosity only in South West Asia and North East Africa and is essentially monomorphic elsewhere in the world. As such SNPs are identified for diverse regions of the world, they can provide additional information to refine ancestry inference whenever an individual carries the variant allele.

Our STRUCTURE analyses of 320 of our AISNPs on 47 populations show much clearer distinction between Europe, Southwest Asia, and South Asia. Thus, it is evident that more informative SNPs do exist. We just need to identify the optimal combination. STRUCTURE and *frappe* will also provide help and demonstrate quality of the results. Heatmaps and PCA analyses will similarly assist. Our objective is a small but robust set

Figure 7-7: Preliminary results with 321 AISNP candidates and 44 populations.



of AISNPs, but that may not be possible if multiple geographic regions are to be reliably differentiated. The process will be iterative and if a small set is not capable of sufficient resolution we may identify different region-specific panels. We are exploring methods to identify SNPs that are informative for specific subregions, but have not been able to identify/invent a method that will be optimal.

We know of no algorithmic method to identify an optimized subset of the 500 or so AISNPs we will soon have data on; empiricism will be a factor. One approach to reducing the number of markers is to test for redundancy of information by simple correlation of the allele frequencies across populations for all pairs of SNPs. SNPs that are highly correlated because of strong LD are redundant; those which are highly correlated but not in LD provide independent information and are candidates for providing robustness to the panel and a basis for balancing the numbers of SNPs informative for different relationships. Another approach is the use of a heatmap to identify the clusters of SNPs and select the first marker of each cluster. A third is the method of Sampson et al. [2008, 2011] that selects SNPs sequentially based on maximizing assignment of individuals to the regions of known origin.

7.3.5. Preliminary studies of LISNPs and PISNPs

As noted above, we have identified several sets of very close and tightly linked SNPs that define haplotypes with exceedingly rare internal recombination. These will be excellent lineage informative SNPs. We have begun typing a few additional SNPs to explore whether or not we can find additional such mini-haplotypes. Full examination of the issue is part of a new grant application submitted to NIJ.

We have already typed several phenotype-informative SNPs as part of our AISNP project since some of the best known phenotype SNPs also show a restricted geographic distribution (e.g., Figure 7-5). In future work we plan to extend this to additional SNPs identified in the literature since our OCA2 studies have shown that

SNPs/haplotypes identified in Europe as highly correlated with the blue eye phenotype are NOT informative in other parts of the world because the LD in Europe breaks down elsewhere.

8. Conclusions

8.1 IISNPs

Our previous IISNP panel published in Pakstis et al. (2007) consisted of 40 best SNPs based on 40 populations and it accomplished the objective in our original application for such a panel of SNPs. While there was no significant pairwise LD in any of the populations, some pairs were sufficiently close that linkage existed. This made those SNP pairs more difficult to use in studies involving close biological relationships. Therefore, in our more recent search to develop a panel of IISNPs that were universally applicable and unlinked (Pakstis et al., 2010), we preferentially targeted regions of the genome in which we did not already have good IISNP candidates in order to enlarge the number of unlinked IISNPs. We also enlarged our set of populations by adding four populations for geographic regions poorly represented in the initial set of 40 populations: East Africa, East Europe, South Asia, and Southeast Asia. Expanding the set of populations studied made our evaluation of IISNP candidates more stringent and helps bolster support for the universal applicability of the 45-SNP panel of unlinked IISNPs that we have developed. The unlinked marker panel is valid for relationship inference without having to incorporate genetic linkage values into calculations. If relationships are not involved in the applied application, more of the 92 IISNPs can be added to the set to make the match probabilities even smaller. When pairwise LD does not exist, as among the 45 unlinked IISNPs we have developed, the SNPs are statistically independent at the population level and the “product rule” can be used to calculate match probabilities. Most of the populations we studied have match probabilities $<10^{-17}$ and many of the match probabilities are $<10^{-18}$; even some of the smaller, more isolated populations have match probabilities $<10^{-15}$. Thus, this set of 45 unlinked SNPs is an excellent panel for individual identification with match probabilities comparable to the CODIS STR panel and these are not highly dependent on ethnicity. Computing match probabilities based on all 86 IISNPs that show no LD gives results in the range of 10^{-31} to 10^{-35} for the 44 populations. At this level, the actual probability has no realistic meaning other than uniqueness among all humans.

Empirical confirmation of the utility of the 92 IISNPs in additional populations may be desirable, but we do not think it is cost effective at this point. We can be confident that the 45-marker panel will have essentially the same useful properties for individual identification in other large human populations. Given the global ubiquity and common frequency of both alleles at all 92 SNPs only extremely small and highly inbred populations are expected to have many of the 45 loci approach fixation of one allele. We have deliberately included several small isolated and inbred populations from different geographic regions in our studies: Mbuti from Africa, Samaritans from Southwest Asia, Khanty from West Siberia, Nasioi from Melanesia, Ami and Atayal from Taiwan, Surui and Karitiana from the Amazon. While these do show larger match probabilities than the large populations, those probabilities are still $<10^{-15}$. Some of these smaller populations are among the smallest, most isolated in the world making it exceedingly improbable that another small population would be dramatically different. Should an individual match show few heterozygotes, that in itself is informative. If necessary, additional SNPs from the remaining 47 IISNPs could be typed to yield a smaller statistical value. (However, any DNA match probability of even 10^{-2} can be meaningful in conjunction with other evidence.) Thus, while we have obtained additional population samples as this IISNP study was concluding, we have not invested money and effort into testing additional populations for these markers.

8.2 AISNPs

Our efforts to identify AISNPs has shown us that the problem is much more complex than usually discussed in the literature. Foremost is the fact that markers useful for distinguishing among one specific set of populations is likely to be much less useful at distinguishing among a different set of populations, even if the same geographic regions are involved. Our progress in that area shows that a small number of AISNPs (~two dozen) can do very well for distinguishing among individuals with ancestry exclusively or primarily from sub-Saharan Africa, the Americas, and opposite ends of Eurasia. However, there is a clear clinal distribution across Eurasia and probably from north-central Siberia through North and then South America. (Data across this distribution are extremely sparse.) Clearly different SNPs provide information about different parts of the global distribution. Therefore, separate sets of AISNPs may be required for distinguishing among populations within a geographic region.

In conclusion we have made progress but from a purely scientific perspective conclude that much more work is required to find robust sets of AISNPs for specific purposes. We have produced a large dataset of markers on multiple populations and find that no obvious algorithm or statistic appears to define a single, good set of AISNPs based on the statistical criteria that we are developing. Extensive analyses are underway but no answers are yet clear pending testing of additional candidate AISNPs.

8.3 Dissemination of results

This project has primarily involved searching for new appropriate SNPs with no meaningful intermediate results worthy of publication. Results of this project are now being made available through different dissemination strategies and publications. We have circulated our unpublished results by making available on our web site the lists of markers in our provisional panel prior to publication for the current 92-SNP IISNP data. A concomitant effort has involved making all of the raw allele frequency data publically available in ALFRED. Our policy is that the forensic data are available in ALFRED as soon as possible after we check the data for errors, etc. All of the IISNP data that we generated are already entered into ALFRED. The allele frequency data on most of the 320 AISNPs for 40 to 44 populations are also in ALFRED. The additional data on the 128-AISNP subset will be entered as the data sets are completed. We have individually notified the NIJ and some individual members of the forensic community of the material on our web site. Our new IISNP paper (submitted to Forensic Science International Genetics) will publicize the existence of a SNP Set function in ALFRED to allow access to accumulated data on SNPs in various published sets. That functionality is a prototype for the development of a more functional forensic interface to data that are accumulating in ALFRED; that project has just been funded by NIJ. 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 in 2007, 2008, and 2009 as well as the ISFG 2007 annual meeting in Copenhagen. Copies of those poster presentations are still accessible at our laboratory website as pdf files. We are also publishing papers to provide full background and analyses documenting the SNP panels we have developed in a peer reviewed setting that will make their strengths and limitations clear and help to make the panels acceptable in the courts. Now that the panel of IISNPs is final, a major paper has been published in Human Genetics in early 2010. The broader forensic community will become more aware of our panels by placing some of our papers in the forensic literature. Our paper on the 128 AISNPs on 119 populations has been published in the new journal *Investigative Genetics* [Oct 2010, in press; Jan 2011, online publication] and the new IISNP paper is under review at *Forensic Science International Genetics*. While we do not consider that set of AISNPs sufficient for accurate ancestry inference, the dataset does represent a larger and more comprehensive dataset than anything yet published and provides new cautionary insights into the general problem of ancestry inference.

A poster presentation [Cho et al.] was made at the American Society for Human Genetics's meeting in early November 2010 exploring multidimensional PCA and hierarchical approaches to identifying useful AISNPs.

An analysis of inference on Native American populations used the dataset assembled as part of our NIJ studies for an invited symposium paper at the annual American Association of Physical Anthropologists in April, 2010. A decision was subsequently made to publish all the Symposium papers in a special issue of the American Journal of Physical Anthropology. We have submitted the paper based on this work and it is being reviewed, as will all in the symposium, before being accepted. While the analyses were not part of or supported by the NIJ project, the results are clearly of relevance.

At the annual NIJ grantees meeting in June 2010 we also demonstrated a prototype interface to ALFRED to retrieve data for published SNP sets for individual identification and ancestry inference. The extra data on the IISNP panel can be found through that new SNP Sets interface in ALFRED. We have subsequently added several additional SNP Sets of forensic interest to the ALFRED interface. With the new funding of NIJ 2010-DN-BX-K226 we will develop a more detailed interface and tools useful for forensic investigators that build upon the infrastructure already being piloted and already available via the ALFRED allele frequency database.

Recently, we learned of a paper in the journal Electrophoresis [Lou et al., 2011] by a Chinese research group which has developed a multiplexed assay for individual identification based on 44 of the IISNPs that we reported in Pakstis et al. (2010).

9. APPENDIX

Kidd Laboratory (January 2009) list of candidate SNPs for individual identification (IISNPs)

92 SNPs with average heterozygosity ≥ 0.4 and Fst(44pops) <0.06

Including a suggested set of 45 IISNP markers that are also "unlinked"

The SNPs are sorted by the Fst value based on a total of 44 population samples (more than 2,200 individuals typed). Four new population samples have been tested on all the SNPs screened since the "Provisional List of Candidates, Summer 2007"†. The 4 new samples include the Sandawe (East Africa), Hungarians (Europe), Keralites (South Central Asia), and Laotians (Southeast Asia). One of the original 40 best SNPs was dropped from the list after the expanded population testing due to an Fst>0.060; for convenience that marker is identified at the bottom of the table. Note that under the column labeled *Fst(44p) ranks* the aqua-blue highlighting of ranks indicate markers that were in the original published list of 40 best SNPs (Pakstis et al., 2007). Markers studied by the SNPforID consortium (Sanchez et al., 2006) have a single asterisk tag after the Fst(44p) rank. Publications describing the identification of all but the most recently screened SNPs can be found in the appended citation list.

In column 1 of the table ("unlinked" IISNPs), the green-highlighted check marks (√) indicate 45 SNPs among the 92 candidate SNPs that appear to be the most useful for individual identification at this time; 33 of these 45 proposed IISNPs are more than 95 cM apart while the other 12 SNPs in the list of "unlinked" SNPs range from 41 to 94 cM apart. The 45 proposed IISNPs are spread across the 22 autosomes. The set of "unlinked" SNPs might still need adjustment depending on the typing procedures developed for the implementation of this recommended panel. For example, it may not be possible to include all 45 SNPs due to multiplexing problems. Substitute SNPs may be needed and the additional SNPs in the list below offer some alternate candidates on various chromosomes. All 92 SNPs meet the population genetics criteria (Fst<0.06 and average heterozygosity >0.4); however, genetic map distances for substitute SNPs on the IISNP list need to be considered carefully to avoid markers that are too closely linked and that thus may have a degree of linkage disequilibrium that renders the substitute marker too correlated with existing nearby IISNPs. In such a case the substitute SNP would not add a full marker's worth of independent information to the overall IISNP panel.

The table column labeled *Avg cM position* is a simple average of the centi-Morgan value of the polymorphism on the DeCode, Genethon, and Marshfield genetic maps (which were obtained from the NCBI Map Viewer). The reader is reminded that each of these extensive maps does not necessarily have the same starting point on each chromosome and that the density of markers will vary in different chromosome regions. The starting or zero positions are near the pter end of each chromosome.

Except for some of the most recently screened markers, the information here was included in figures and tables presented in posters at various scientific meetings. (See footnote †.) PDF files of the poster presentations and of the earlier preliminary candidate list as of summer 2007 can be found at the following "contents" web page under the Kidd Lab Library header (http://info.med.yale.edu/genetics/kkidd/contents.html).

Allele frequency tables for all 92 best candidate SNPs have been deposited into ALFRED, the Allele Frequency Database. ALFRED is freely accessible on the web at http://ALFRED.med.yale.edu. Allele frequency tables for several hundred SNPs that were screened for this low Fst—high heterozygosity project are in the process of being entered into ALFRED; many of these SNPs did not pass beyond the early screening stage in which they were typed for 7 population samples representing the major continental regions of the world.

√ marks 45 "unlinked" IISNPs; # indicates one of 40 best SNPs (Pakstis et al., 2007);
*** next to Fst rank tags SNPforID marker**

unlinked IISNPs	Fst 44p rank	TaqMan Catalog ID	dbSNP rs#	ALFRED UID	Avg.Het. (44p)	Fst (44p)	Chr	Chr arm	Nucleotide Position Map Build 36.2	Avg cM posi-tion
√	1	C___2450075_10	rs10488710	SI001899B	0.442	0.0217	11	q	114,712,386	111.6
√	2	C__16156638_10	rs2920816	SI015053O	0.459	0.0232	12	q	39,149,319	57.9
√	3	C__29220288_10	rs6955448	SI015041L	0.421	0.0298	7	p	4,276,891	7.6
√	4	C___1619935_1_	rs1058083	SI001402H	0.464	0.0300	13	q	98,836,234	84.6
√	5	C____824925_10	rs221956	SI015402M	0.462	0.0310	21	q	42,480,066	54.6
√	6	C___2556113_10	rs13182883	SI001390N	0.472	0.0314	5	q	136,661,237	140.6
√	7	C___8263011_10	rs279844	SI001391O	0.484	0.0316	4	p	46,024,412	61.8
√	8	C__11245682_10	rs6811238	SI001910L	0.484	0.0319	4	q	169,900,190	166.9
√	9	C___9603287_10	rs430046	SI015042M	0.441	0.0321	16	q	76,574,552	94.1
√	10	C____788229_10	rs576261	SI015043N	0.472	0.0352	19	q	44,251,647	63.6
	11	C__16071557_10	rs2833736	SI015401L	0.460	0.0356	21	q	32,504,593	32.2
√	12	C___2049946_10	rs10092491	SI001900K	0.459	0.0364	8	p	28,466,991	52.5
√	13	C___1006721_1_	rs560681	SI001392P	0.434	0.0364	1	q	159,053,294	167.3
	14	C___1056251_10	rs590162	SI015390S	0.482	0.0366	11	q	121,701,199	124.6
√	15	C___9084395_10	rs2342747	SI015395X	0.423	0.0367	16	p	5,808,701	10.1
√	16	C__26449463_10	rs4364205	SI015054P	0.458	0.0372	3	p	32,392,648	56.3
√	17	C___2997607_10	rs445251	SI001912N	0.464	0.0386	20	p	15,072,933	36.8
√	18	C__29060279_10	rs7041158	SI015389A	0.439	0.0389	9	p	27,975,938	51.3
	19	C___1797119_10	rs9546538	SI003897B	0.429	0.0395	13	q	83,354,736	69.6
√	20	C___1304451_10	rs1294331	SI015382T	0.457	0.0396	1	q	231,515,036	247.4
√	21	C___1454681_10	rs159606	SI015134O	0.442	0.0396	5	p	17,427,898	70.3
√	22	C___3254784_10	rs740598	SI001393Q	0.462	0.0406	10	q	118,496,889	139.1
	23	C___3031045_1_	rs464663	SI015400K	0.462	0.0410	21	q	26,945,241	25.7
√	24	C__11673733_10	rs1821380	SI001913O	0.465	0.0413	15	q	37,100,694	38.2
√	25	C___1817429_10	rs1336071	SI001915Q	0.472	0.0418	6	q	94,593,976	102.3
	26	C___2572254_10	rs1019029	SI001916R	0.474	0.0419	7	p	13,860,801	23.0
√	27	C___1371205_10	rs9951171	SI001395S	0.475	0.0420	18	p	9,739,879	31.4
√	28	C___7968314_10	rs8078417	SI015122L	0.402	0.0426	17	q	78,055,224	130.0
	29	C___2140539_10	rs1358856	SI001427O	0.474	0.0430	6	q	123,936,677	121.3
√	30	C__25749280_10	rs6444724	SI001903N	0.469	0.0435	3	q	194,690,074	217.4
√	31	C___9371416_10	rs13218440	SI001397U	0.458	0.0436	6	p	12,167,940	24.6
	32	C__15957782_10	rs2270529	SI015388Z	0.421	0.0443	9	p	14,737,133	28.9
√	33	C___1452175_	rs1498553	SI015123M	0.477	0.0446	11	p	5,665,604	11.4
√	34	C____342791_10	rs7520386	SI001394R	0.477	0.0447	1	p	14,027,989	29.7
√	35	C___2508482_10	rs1523537	SI001914P	0.472	0.0447	20	q	50,729,569	79.4
√	36	C___3285337_	rs1736442	SI015124N	0.438	0.0450	18	q	53,376,775	79.4
	37	C___1152009_10	rs1478829	SI001917S	0.474	0.0459	6	q	120,602,393	119.8
√	38	C___2822618_10	rs3780962	SI001904O	0.476	0.0462	10	p	17,233,352	42.7
	39	C____105475_10	rs7229946	SI001901L	0.464	0.0466	18	q	20,992,999	49.8
	40	C__30281961_10	rs9866013	SI015044O	0.419	0.0468	3	p	59,463,380	77.4
	41	C___3206279_1_	rs2567608	SI001902M	0.473	0.0469	20	p	22,965,082	49.8
√	42	C___1541359_10	rs2399332	SI015385W	0.435	0.0472	3	q	111,783,816	124.5
√	43	C__11887110_1_	rs987640	SI001918T	0.476	0.0476	22	q	31,889,508	34.9
	44	C____376875_10	rs4847034	SI015135P	0.445	0.0476	1	p	105,519,154	134.1
	45	C__11522503_1_	rs2073383	SI001911M	0.456	0.0479	22	q	22,132,171	15.8
	46	C__26227271_10	rs3744163	SI015125O	0.430	0.0480	17	q	78,333,148	130.0
	47	C___1605841_10	rs10500617	SI003936V	0.404	0.0481	11	p	5,055,969	9.0
√	48	C___9530932_10	rs993934	SI015136Q	0.450	0.0482	2	q	123,825,683	134.2
	49	C___7969752_	rs2291395	SI015126P	0.473	0.0486	17	q	78,119,428	130.0
√	50	C___2715242_10	rs10773760	SI015392U	0.444	0.0487	12	q	129,327,649	165.1
	51	C___1274218_	rs12480506	SI001169R	0.403	0.0492	20	p	16,189,416	39.1
	52	C__11258596_	rs4789798	SI015127Q	0.472	0.0494	17	q	78,124,932	130.0
√	53	C____187613_10	rs4530059	SI015393V	0.406	0.0495	14	q	103,840,194	126.5
	54	E_rs8070085_10	rs8070085	SI014994B	0.437	0.0498	17	q	38,595,510	66.4
√	55	C___1276208_10	rs12997453	SI001396T	0.440	0.0503	2	q	182,121,504	188.1
√	56	C__27999762_10	rs4606077	SI015387Y	0.421	0.0503	8	q	144,727,897	164.2

	57	C_____19853_	rs689512	SI001329P	0.423	0.0507	17	q	78,308,991	130.0
√	58	C___2515223_10	rs214955	SI001403I	0.474	0.0511	6	q	152,739,399	155.7
	59	C___1256256_1_	rs2272998	SI001398V	0.467	0.0511	6	q	148,803,149	148.6
	60	C___2539254_	rs5746846	SI003887A	0.464	0.0515	22	q	18,300,646	9.0
	61	C__26372385_10	rs4288409	SI015386X	0.415	0.0515	8	q	136,908,411	152.0
√	62	C___2184724_	rs2269355	SI015128R	0.473	0.0521	12	p	6,816,175	17.0
	63	C___1570295_10	rs1027895	SI000905O	0.433	0.0524	17	q	43,865,696	69.4
√	64	C___3004178_10	rs321198	SI001906Q	0.459	0.0530	7	q	136,680,378	143.5
	65	C__11631183_	rs2175957	SI015129S	0.437	0.0530	17	q	38,540,348	66.3
	66	C___3080506_1_	rs2292972	SI001330H	0.422	0.0530	17	q	78,359,077	130.0
	67 *	C___7698393_	rs901398	SI003975Y	0.441	0.0531	11	p	11,052,797	18.2
	68	C___2539253_	rs9606186	SI001586U	0.437	0.0531	22	q	18,300,359	9.0
√	69	C__3153696a_10	rs338882	SI001401G	0.469	0.0532	5	q	178,623,331	195.8
√	70	C___2002375_10	rs10776839	SI015046Q	0.463	0.0533	9	q	136,557,129	152.6
	71	C___2714437_	rs521861	SI001163L	0.473	0.0534	18	q	45,625,012	70.7
√	72	C___2073009_10	rs1109037	SI001909T	0.470	0.0534	2	p	10,003,173	21.5
	73	C__29487208_10	rs4796362	SI015397Z	0.471	0.0536	17	p	6,752,253	14.2
	74	C___3032822_1_	rs315791	SI001404J	0.472	0.0539	5	q	169,668,498	176.3
	75 *	C___7539584_	rs891700	SI003976Z	0.471	0.0541	1	q	237,948,549	261.3
	76	C___7477802_	rs1004357	SI015131L	0.411	0.0541	17	q	39,047,052	67.1
	77	E_rs7205345_10	rs7205345	SI001905P	0.469	0.0544	16	p	7,460,255	14.2
	78	C__1636106a_10	rs6591147	SI001409O	0.451	0.0545	11	q	105,418,194	106.3
	79	C____411273_10	rs2503107	SI001426N	0.458	0.0548	6	q	127,505,069	125.9
	80	C___7538108_10	rs1410059	SI001399W	0.470	0.0551	10	q	97,162,585	117.6
	81	C__11907549_1_	rs1872575	SI003924S	0.472	0.0552	3	q	115,287,669	128.2
	82	C___7428940_10	rs1554472	SI001919U	0.472	0.0552	4	q	157,709,356	155.7
	83 *	C__11989432_10	rs2046361	SI003977A	0.462	0.0559	4	p	10,578,157	23.1
√	84	C___7945874_10	rs9905977	SI015045P	0.419	0.0561	17	p	2,866,143	7.9
	85	C___1995608_10	rs7704770	SI001908S	0.449	0.0567	5	q	159,420,531	163.0
	86	C___1880371_10	rs13134862	SI001400F	0.453	0.0571	4	q	76,644,920	84.2
	87	C____282853_10	rs2811231	SI015137R	0.458	0.0579	6	p	55,263,663	78.9
	88	C___7459903_10	rs985492	SI001413J	0.469	0.0580	18	q	27,565,032	58.6
	89	C___1605842_	rs10768550	SI003937W	0.408	0.0580	11	p	5,055,290	9.0
	90 *	C___9630073_	rs1490413	SI003978B	0.469	0.0583	1	p	4,267,183	8.3
	91	C__11338582_	rs2255301	SI001069Q	0.463	0.0587	12	p	6,779,703	16.9
√	92	C____611046_10	rs722290	SI003542O	0.468	0.0596	14	q	52,286,473	47.6

SNP below dropped from IISNP list when Fst exceeded 0.06 after expanding to 44 population samples

XXX	C___2223883_10	rs447818	SI001907R	0.469	0.0622	6	q	145,910,689	145.1

Note:

Only chance level linkage disequilibrium (LD) values are observed for all unique pairings of 86 of the 92 IISNPs (median LD =0.011) in each of 44 population samples. However, six of the 92 SNPs show strong LD in most of the 44 populations for a small subset (7) of the unique pairings due to close linkage; these 6 SNPs can therefore only be alternative candidates for inclusion in an applied IISNP panel of 86 SNPs independent at the population level. These six SNPs showing some LD are those in the above table with Fst ranks numbered: 52, 57, 65, 66, 68, and 89.

Citations

Pakstis et al. 2007. *Human Genetics* 121:304-317. A PDF file of this paper (publication #461) can be downloaded at: http://info.med.yale.edu/genetics/kkidd/pubs.html. (See also publications 467 and 468.)

Sanchez et al. 2006. *Electrophoresis* 27:1713-1724.

† Scientific meetings where much of this information was presented:

Figures 1 and 2 of Poster presentation July 24, 2007 for the annual meeting of grantees of the U.S. National Institue of Justice, Washington, D.C.
Title: An expanded, nearly universal, panel of SNPs for individual identification.
Authors: Andrew J. Pakstis, William C. Speed, Judith R. Kidd, Kenneth K. Kidd
Affiliation: Dept of Genetics, Yale University School of Medicine, New Haven, CT

Figure 1 of Poster presentation August 22-25, 2007 for the meeting of the International Society of Forensic Geneticists (ISFG) in Copenhagen, Denmark.
Title: SNPs for individual identification
Authors: Andrew J. Pakstis, William C. Speed, Judith R. Kidd, Kenneth K. Kidd
Affiliation: Dept of Genetics, Yale University School of Medicine, New Haven, CT

Poster presentation July 21-23, 2008 for the annual meeting of grantees of the U.S. National Institute of Justice, Arlington, Virginia .
Title: Better panels of SNPs for ancestry inference and individual identification
Authors: Andrew J. Pakstis, William C. Speed, Judith R. Kidd, Kenneth K. Kidd
Affiliation: Dept of Genetics, Yale University School of Medicine, New Haven, CT

Poster presentation June 15-17, 2009 for the annual meeting of grantees of the U.S. National Institute of Justice, Arlington, Virginia.
Title: SNP panels for individual identification and for ancestry inference
Authors: Kenneth K. Kidd, Judith R. Kidd, William C. Speed, Andrew J. Pakstis
Affiliation: Dept of Genetics, Yale University School of Medicine, New Haven, CT

Acknowledgments

In addition to direct funding from grants (see below) we thank Applied Biosystems for providing some of the TaqMan assays used for screening SNPs on our population samples. We also thank the many hundreds of individuals who volunteered to provide samples for genetic marker testing.

Financial Support

This work was supported in part by multiple funding sources. The assembly of some of the population resources was funded in part by NIH grants (GM057672 and AA009379). The forensic aspects of the project were funded in part by the U.S. DOJ NIJ; specifically, this work was supported in part by awards 2004-DN-BX-K025 and 2007-DN-BX-K197 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the authors and do not necessarily reflect those of the Department of Justice.

END OF APPENDIX

10. Recent scientific meetings where AISNP and IISNP work was presented

Slide presentation October 23, 2009 at the annual meeting of the American Society of Human Genetics, Honolulu, Hawaii.
Title: A universal SNP panel for individual identification
Authors: Kenneth K. Kidd(1), Judith R. Kidd(1), Eva Straka(1), William C. Speed(1), Rixun Fang(2), Fiona Hyland(2), Manohar R. Furtado(2), Andrew J. Pakstis(1)
Affiliation: (1) Dept of Genetics, Yale University School of Medicine, New Haven, CT
And (2) Genetic Systems Division, Applied Biosystems, Foster City, CA

Slide presentation December 2009 at NIJ Technology Transition Workshop, Texas
Title: Genetics of SNP markers
Author: Kenneth K. Kidd

Presentation June 20-22, 2010 for the annual meeting of grantees of the U.S. National Institute of Justice, Arlington, Virginia.
Title: Demonstration of ALFRED, a reference database for forensic SNPs
Authors: Kenneth K. Kidd, Andrew J. Pakstis, Haseena Rajeevan, William C. Speed, Usha Soundararajan, Judith R. Kidd

Slide presentation April 14-17, 2010 for the annual meeting of the American Association of Physical Anthropologists held in Albuquerque, New Mexico.
Title: Ancestry informative SNPs and haplotypes in Native American populations.
Authors: Kenneth K. Kidd, Judith R. Kidd, Francoise Friedlaender, Andrew J. Pakstis.

Poster presentation Nov 2-6, 2010 for the annual meeting of the American Society of Human Genetics, Washington, D.C.
Title: Selecting genetic marker panels to detect population stratification using multidimensional principal components and hierarchical clustering approaches
Authors: Kelly Cho, Judith R. Kidd, Kenneth K. Kidd

References

Amorim, A., L. Pereira, Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. Forensic Science International 150 (2005) 17-21.

Balding, D.J. Likelihood-based inference for genetic correlation coefficients. Theoretical Population Biology 63 (2003) 221-230.

Bauchet M., B. McEvoy, L.N. Pearson, E.E. Quillen, T. Sarkisian, K. Hovhannesyan, R. Deka, D.G. Bradley, M.D. Shriver, Measuring European population stratification with microarray genotype data, American Journal of Human Genetics, 80 (2007) 948-954.

Biswas, S., L.B. Scheinfeldt, J.M. Akey, Genome-wide insights into the patterns and determinants of fine-scale population structure in humans, Am J Hum Genet 84 (2009) 641-650.

Butler, J.M., Y. Shen, B.R. McCord, The development of reduced size STR amplicons as tools for analysis of degraded DNA. Journal of Forensic Sciences 48 (2003) 1054-1064.

Butler, J. M., B. Budowle, P. Gill, K. K. Kidd, C. Phillips, P. M. Schneider, P. M. Vallone, and N. Morling, 2008. Report on ISFG SNP Panel Discussion. Progress in Forensics Genetics, Genetics Suppl Series 1:471-472.

Calafell, F., A. Shuster, W.C. Speed, J.R. Kidd, F.L. Black, and K.K. Kidd. Genealogy reconstruction from short tandem repeat genotypes in an Amazonian population. American Journal of Physical Anthropology 108 (1999) 137-146.

Cann, H.M., C. deToma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. chu, C. Carcassi, L. Contu, F. Du, L. Excoffier, G.B. Ferrara, J.S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R.J. Herrera, X. Huang, J. Kidd, K.K. Kidd, A. Langaney, A.A. Lin, S.Q. Mehdi, P. Parham, A. Piazza, M.P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J.L. Weber, H.T. Greely, M.W. Feldman, G. Thomas, J. Dausset, and L.L. Cavalli-Sforza. A human genome diversity cell line panel. Science 296:261-262 (2002).

Cavalli-Sforza, L.L., P. Menozzi, A. Piazza, *The History and Geography of Human Genes,* Princeton University Press, Princeton, 1994.

Chakraborty, R., K.K. Kidd, (Perspective) The utility of DNA typing in forensic work, Science 254 (1991) 1735-1739.

Coble, M.D., J.M. Butler, Characterization of new miniSTR loci to aid analysis of degraded DNA. Journal of Forensic Sciences 50 (2005) 43-53.

Collins-Schramm, H.E., B. Chima, T. Morii, K. Wah, Y. Figueroa, L.A. Criswell, R.L. Hanson, W.C. Knowler, G. Silva, J.W. Belmont, and M.F. Seldin, Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians, Human Genetics 114 (2004) 263-271.

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. Nature Genetics 38 (2006) 1251-1260

Cotterman, C.W. 1954. Estimation of gene frequencies in nonexperimental populations. Chapt. 35 p. 449-465 Eds: O. Kempthorne, T. A. Bancroft, J. W. Gowen and J. L. Lush. Statistics and mathematics in biology.

Cubells, J.F., K. Kobayashi, T. Nagatsu, K.K. Kidd, J.R. Kidd, F. Calafell, H. Kranzler, H. Ichinose, and J. Gelernter. Population genetics of a functional variant of the dopamine beta hydroylase gene (DBH). American Journal of Medical Genetics 74 (1997) 374-379.

Donnelly, M.P., P. Paschou, E. Grigorenko, D. Gurwitz, S.Q. Mehdi, S.L.B. Kajuna, C. Barta, S. Kungulilo, N.J. Karoma, R.-B. Lu, O.V. Zhukova, J.-J. Kim, D. Comas, M. Siniscalco, M. New, P. Li, H. Li, W.C. Speed, H. Rajeevan, A.J. Pakstis, J.R. Kidd, K.K. Kidd **2010**. The distribution and most recent common ancestor of the 17q21 inversion in humans. *American Journal of Human Genetics* 86:161-171.

Dupuy, B.M., M. Stenersen, T. Egeland, B. Olaisen, Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. Human Mutation 23 (2004) 117-124.

Enoch M-A, Shen P-H, Xu K, Hodgkinson C, Goldman D: Using ancestry-informative markers to define populations and detect population stratification. Journal of Psychopharmacology, 20 (2006) 199-226.

Falush, D., M. Stephens, J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164 (2003) 1567-87.

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

Ge, J., B. Budowle, J.V. Planz, and R. Chakraborty, Haplotype block: a new type of forensic DNA markers, International Journal of Legal Medicine, e-pub (2009).

Gill, P., D.J. Werrett, B. Budowle, R. Guerrieri, An assessment of whether SNPs will replace STRs in national DNA databases—joint considerations of the DNA working group of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDAM), Science & Justice 44 (2004) 51-53.

Halder I., M. Shriver, M. Thomas, J.R. Fernandez, T. Frudakis, A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications, Human Mutation, 29 (2008) 648-658.

Hodgkinson, C.A., Q. Yuan, K. Xu, P.H. Shen, E. Heinz, E.A. Lobos, E.B. Binder, J. Cubells, C.L. Ehlers, J. Gelernter, J. Mann, B. Riley, A. Roy, B. Tabakoff, R.D. Todd, Z. Zhou, and D. Goldman, Addictions biology: haplotype-based analysis for 130 candidate genes on a single array, Alcohol 43:505-515 (2008).

Holland, M.M., C.A. Cave, C.A. Holland, T.W. Bille, Development of a Quality, High Throughput DNA Analysis Procedure for Skeletal Samples to Assist with the Identification of Victims from the World Trade Center Attacks. Croatian Medical Journal 44 (2003) 264-272.

Huang, Q.Y., F.H. Xu, H. Shen, H.Y. Deng, Y.J. Liu, Y.Z. Liu, J.L. Li, R.R. Recker, H.W. Deng, Mutation patterns at dinucleotide microsatellite loci in humans. American Journal of Human Genetics 70 (2002) 625-634.

Inagaki, S., Y. Yamamoto, Y. Doi, T. Takata, T. Ishikawa, K. Imabayashi, K. Yoshitome, S. Miyaishi, H. Ishizu. A new 39-plex analysis method for SNPs including 15 blood group loci. Forensic Science International 144 (2004) 45-57.

International HapMap Consortium, The International HapMap Project, Nature 406 (2003) 789-796.

International HapMap Consortium. A haplotype map of the human genome. Nature 437 (2005) 1299-1320.

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 (2008) 998-1003.

Jorde, L B., P.A. Shortsleeve, J.W. Henry, R.T. Vanburen, L.E. Hutchinson, T.M. Rigley, Genetic analysis of the Utah population: a comparison of STR and VNTR loci, Human Biology 72 (2000) 927-936

Kalinowski, S.T. The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. (2010) *Heredity* In press, e-pub August 4.

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

Kidd, J.R., A.J. Pakstis, and K.K. Kidd, 1993. Global levels of DNA variation. Proceedings of the 4th International Symposium on Human Identification 1993 (Promega) pp 21-30.

Kidd, J.R., F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, K.K. Kidd, **2011**, Analyses of a set of 128 ancestry informative SNPs (AISNPs) in a global set of 119 population samples. Investigative Genetics 2:1 e-pub January 5, 2011; In press October 20, 2010

Kidd, J.R., F.R. Friedlaender, A.J. Pakstis, M.R. Furtado, R. Fang, X. Wang, C.R. Nievergelt, K.K. Kidd. SNPs and haplotypes in Native American populations. Submitted to American Journal of Physical Anthropology.

Kidd, K.K., A.J. Pakstis, W.C. Speed, and J.R. Kidd, Understanding human DNA sequence variation. Journal of Heredity 95 (2004) 406-420.

Kidd, K.K., A.J. Pakstis, W.C. Speed, E.L. Grigorenko, S.L.B. Kajuna, N.J. Karoma, S. Kungulilo, J-J. Kim, R-B. Lu, A. Odunsi, F. Okonofua, J. Parnas, L.O. Schulz, O.V. Zhukova, and J. Kidd. Developing a SNP panel for forensic identification of individuals. Forensic Science International 164 (2006) 20-32.

Kim, J.J., P. Verdu, A.J. Pakstis, W.C. Speed, J.R. Kidd, and K.K. Kidd. Use of autosomal loci for clustering individuals and populations of East Asian origin, Human Genetics 117 (2005) 511-519.

Kosoy, R., R. Nassir, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, P.K. Gregersen, J.W. Belmont, F.M. De La Vega, and M.F. Seldin, Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America, Hum Mutat 30:69-78 (2009)

Lao, O., K. van Duijn, P. Kersbergen, P. de Knijff, M. Kayser. Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. American Journal of Human Genetics 78 (2006) 680-90.

Lao, O., T.T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf, A. Caliebe, M. Balascakova, J. Bertranpetit, L.A. Bindoff, D. Comas, G. Holmlund, A. Kouvatsi, M. Macek, I. Mollet, W. Parson, J. Palo, R. Ploski, A. Sajantila, A. Tagliabracci, U. Gether, T. Werge, F.

Rivadeneira, A. Hofman, A.G. Uitterlinden, C. Gieger, H.E. Wichmann, A. Ruther, S. Schreiber, C. Becker, P. Nurnberg, M.R. Nelson, M. Krawczak, and M. Kayser, Correlation between genetic and geographic structure in Europe, Curr Biol 18:1241-1248 (2008).

Lee, H.Y., M. J. Park, J-E Yoo, U. Chung, G-R Han, K-J Shin. Selection of twenty-four highly informative SNP markers for human identification and paternity analysis in Koreans. Forensic Science International 148 (2005)107-112.

Lewontin, R.C., D.L. Hartl, Population genetics in forensic DNA typing, Science 254 (1991) 1745-1750.

Li, H., S. Borinskaya, K. Yoshimura, N. Kal'ina, A. Marusin, V.A. Stepanov, Z. Qin, S. Khaliq, M.-Y. Lee, Y. Yang, A. Mohyuddin, D. Gurwitz, S.Q. Mehdi, E. Rogaev, L. Jin, N. Yankovsky, J.R. Kidd, and K.K. Kidd. Refined Geographic Distribution of the Oriental ALDH2*504Lys (nee 487Lys) Variant, Annals of Human Genetics 73:335–345 (2009).

Li H., K. Cho, J.R. Kidd, K.K. Kidd. Genetic Landscape of Eurasia and ‘‘Admixture’’ in Uyghurs. The American Journal of Human Genetics 85(6):934-937 (2009) (Letter to Editor).

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).

Mountain, J.L., A. Knight, M. Jobin, C. Gignoux, A. Miller, A.A. Lin, and P.A. Underhill, SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes, Genome Res 12:1766-1772 (2002).

Nassir R., R. Kosoy, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, P.K. Gregersen, J.W. Belmont, F.M. De La Vega, M.F. Seldin, 2009. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BioMedCentral Genetics 10:39.

National Research Council Committee on DNA Technology in Forensic Science.
The evaluation of forensic DNA evidence/Committee on DNA Forensic Science: An update. Washington D.C., National Academy Press, 1996.

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).

Pakstis A.J., W.C. Speed, J.R. Kidd, K.K. Kidd, 2007. Candidate SNPs for a universal individual identification panel. Human Genetics 121:305-317

Pakstis, A. J., W. C. Speed, J. R. Kidd, and K. K. Kidd, 2008. SNPs for Individual Identification. Progress in Forensics Genetics Genetics Suppl Series 1:479-481.

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.

Paschou P., E. Ziv, E.G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M.W. Mahoney, P. Drineas, PCA-correlated SNPs for structure identification in worldwide human populations, *PLoS Genet*ics, 3 (2007) e160.

Peltonen, L., A. Jalanko, T. Varilo. Molecular genetics of the Finnish disease heritage. Human Molecular Genetics 8 (1999)1913-23.

Pemberton, T.J., M. Jakobsson, D.F. Conrad, G. Coop, J.D. Wall, J.K. Pritchard, P.I. Patel, and N.A. Rosenberg, Using population mixtures to optimize the utility of genomic databases: linkage disequilibrium and association study design in India, Ann Hum Genet 72:535-546 (2008).

Price, A.L., J. Butler, N. Patterson, C. Capelli, V.L. Pascali, F. Scarnicci, A. Ruiz-Linares, L. Groop, A.A. Saetta, P. Korkolopoulou, U. Seligsohn, A. Waliszewska, C. Schirmer, K. Ardlie, A. Ramos, J. Nemesh, L. Arbeitman, D.B. Goldstein, D. Reich, J.N. Hirschhorn, Discerning the ancestry of European Americans in genetic association studies, PLoS Genetics 4 (2008) e236.

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

Ramakrishnan, U., J.L. Mountain, Precision and accuracy of divergence time estimates from STR and SNPSTR variation, Molecular Biology Evolution 21 (2004)1960-1971.

Reich, D.E., S.F. Schaffner, M.J. Daly, G. McVean, J.C. Mullikin, J.M. Higgins, D.J. Richter, E.S. Lander, D. Altshuler, Human genome sequence variation and the influence of gene history, mutation and recombination. Nature Genetics 32 (2002) 135-140.

Rosenberg, N.A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, M. W. Feldman. Genetic Structure of Human Populations. Science 298 (2002) 2381-2385.

Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ, and Clark AG. Inferring genetic ancestry: Opportunities, challenges, and implications. American Journal of Human Genetics, 2010, 86: 661-673.

Sampson, J., K.K. Kidd, J.R. Kidd, and H. Zhao, Selecting SNPs to correctly predict ethnicity, in Annual meeting of the American Society of Human Genetics, Philadelphia, PA, 2008.

Sampson, J., K.K. Kidd, J.R. Kidd, and H. Zhao, Select SNPs to identify ancestry, Annals of Human Genetics, In press, early 2011.

Sanchez, J.J., C. Borsting, C. Hallenberg, A. Buchard, A. Hernandez., N. Morling, Multiplex PCR and minisequencing of SNPs—a model with 35 Y chromosome SNPs, Forensic Science International 137 (2003) 74-84.

Sanchez, J.J., C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C. D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P. M. Schneider, A. Carracedo, N. Morling. A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 1713-24.

Shriver, M.D., R. Mei, E. J. Parra, V. Sonpar, I. Halder, S. A. Tishkoff, T. G. Schurr, S. I. Zhadanov, L. P. Osipova, T. D. Brutsaert, J. Friedlaender, L. B. Jorde, W. S. Watkins, M. J. Bamshad, G. Gutierrez, H. Loi, H. Matsuzaki, R. A. Kittles, G. Argyropoulos, J. R. Fernandez, J. M. Akey, K. W. Jones. Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation, Human Genomics 2 (2005) 81-9.

Teare, M.D., A.M. Dunning, F. Durocher, G. Rennart, D.F. Easton, Sampling distribution of summary linkage disequilibrium measures. Annals of Human Genetics, 66 (2002) 223-233.

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, American Journal of Human Genetics 79 (2006) 640-649.

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, American Journal of Human Genetics 80 (2007) 1014-1023.

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 300K SNP information, PLoS Genetics 4 (2008) e4.

Tian C., R. Kosoy, R. Nassir, A. Lee, P. Villoslada, L. Klareskog, L. Hammarström, H.J. Garchon, A.E. Pulver, M. Ransom, P.K. Gregersen, M.F. Seldin, European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups, Molecular Medicine 15 (2009) 371-383.

Tishkoff, S.A., K.K. Kidd, Implications of biogeography of human populations for race" and medicine. Nature Genetics 36 (suppl) (2004) s21-s27.

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

Varilo, T., L. Peltonen. Isolates and their potential use in complex gene mapping efforts. Current Opinion in Genetics & Development 14 (2004) 316-23.

Wright, S. The genetical structure of populations. Annals of Eugenics 15 (1951) 323-354.

Yang, N., H. Li, L.A. Criswell, P.K. Gregersen, M.E. Alarcon-Riquelme, R. Kittles, R. Shigeta, G. Silva, P.I. Patel, J.W. Belmont, and M.F. Seldin, Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine, Human Genetics 118 (2005) 382-392.