Final Report

Population Genetics of SNPs for Forensic Purposes

NIJ Grant# 2004-DN-BX-K-25

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

Portions of this report are taken verbatim from the four research publications supported by this grant or using material from the project:

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.

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

Pakstis, A. J., W. C. Speed, J. R. Kidd, and K. K. Kidd, 2007. SNPs for Individual Identification. Progress in Forensics Genetics 12 (in press)

Butler, J. M., B. Budowle, P. Gill, K. K. Kidd, C. Phillips, P. M. Schneider, P. M. Vallone, and N. Morling, 2007. Report on ISFG SNP Panel Discussion. Progress in Forensics Genetics 12 (in press)

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. Our specific goal was to identify panels of SNP markers (1) with globally low Fst and high average heterozygosity and (2) with globally high Fst 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 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 initially describe an efficient strategy for identifying and characterizing SNPs useful for individual identification. Then we present a panel of 40 best SNPs studied on 40 population samples from around the world that have both low Fst (<0.06) and high heterozygosity (>0.4). Collectively, these SNPs give average match probabilities of less than 10^{-16} in most of the 40 populations and less than 10^{-14} in all but one small isolated population; the range is 2.02×10^{-17} to 1.29×10^{-13}. From other resources we have accumulated a total of 109 SNPs meeting our criteria on a reduced set of 31 populations that are likely to be of greatest forensic relevance because we eliminated small, isolated populations. We expect that many genetically independent (unlinked) markers will be found suitable. We still advocate screening more SNPs and evaluating the better candidates on many additional populations so that reasonably small (e.g. $<10^{-12}$) genotype frequencies can be demonstrated to occur in a wider range of populations. 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. A 10-SNP set developed by others has already been shown to allow easy, though rough, resolution of the four major continental groups. However, their analyses on the HGDP-CEPH panel (and their 10 SNPs on our 40 populations) of those markers did not allow any further geographic subdivision of populations. Our developing AISNP panel currently consists of 249 candidate SNPs that, *in toto* and in some subsets, give greatly improved resolution of the four continental groupings of populations.

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, 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 di-allelic 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. Such markers represent a potentially optimal resource for individual identification. Our project undertook the task of determining how readily we could identify a sufficient quantity

of such markers. Our unique collection of cell lines on population samples from around the world offers a special advantage in accomplishing this task.

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 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 Fst and high average heterozygosity and (2) with globally high Fst 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 is to identify 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 optimizing an ancestry informative panel and, indeed, our progress in that area has necessarily focused on developing criteria.

2.3 Strategy and Methods for Individual Identification [Kidd et al., 2006]

We describe both an efficient strategy for identifying and characterizing such SNPs that would be valuable for individual identification, and then test that strategy on a broad

representation of world populations. 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.

Our preliminary efforts demonstrated the feasibility of identifying SNPs with the useful properties desired and resulted in a panel of 19 SNPs, from an initial selection of 195 candidate SNPs. This set of 19 markers gave an average match probability of less than 10^{-7} in most of the 40 populations studied and no greater than 10^{-6} in the most isolated, inbred populations.

2.4 Provisional panel of 40 best SNPs for individual identification [Pakstis et al., 2007]

Here we reported on our progress in identifying SNPs that show little allele frequency variation among a worldwide sample of 40 populations, i.e., have a low F_{st}, while remaining highly informative. Such markers have match probabilities that are nearly uniform irrespective of population and become candidates for a universally applicable 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. With the NIJ funding we screened a total of 432 SNPs that were likely *a priori* to have high heterozygosity and low allele frequency variation and from these have selected the markers with the lowest F_{st} in our set of 40 populations to produce a panel of 40 low F_{st}, high heterozygosity SNPs. Collectively these SNPs give average match probabilities of less than 10^{-16} in most of the

40 populations and less than 10^{-14} in all but one small isolated population; the range is 2.02×10^{-17} to 1.29×10^{-13}.

These 40 SNPs constitute excellent candidates for the global forensic community to consider for a universally applicable SNP panel for human identification. The best technology and multiplexing sets for routinely studying such markers still needs to be determined in the future. It would also be useful for additional population samples from around the world to be studied on our candidate panel of 40 best SNPs in order to extend the evidence that the candidate SNPs qualify as a universal panel for individual identification. Identifying additional candidate SNPs will be helpful to provide more options to consider when evaluating the best technology, marker combinations, and optimal characteristics for routine lab work. The relative ease with which our panel of 40 best markers could be identified also provides a cautionary lesson for investigations of possible balancing selection.

2.5 Expanding the panel of candidate SNPs for individual identification

[Presentations: NIJ grantees mtg. July 2007; ISFG mtg. Copenhagen Aug 2007]

While the provisional panel of 40 best SNP markers we identified give genotype probabilities of $<10^{-16}$ in almost all populations studied, some forensic scientists suggested that our criteria are too stringent in that we have included several small, isolated groups among the populations used to screen SNPs. We re-evaluated our data, as well as some comparable data we have generated for SNPs proposed by other groups, after excluding the most isolated populations from consideration, reducing the screening panel from 40 to 31 populations. This does result in a larger panel of candidate SNPs using an even more stringent level of interpopulation variation in allele frequencies--an Fst < 0.05 instead of our initial criterion of an Fst < 0.06--while

maintaining heterozygosity > 0.40. In addition to the previously published 40 SNPs we are able to include 23 from among the 36 previously excluded as well as 5 from among the markers proposed by the SNPforID consortium. From our other studies using the same population samples we have identified several additional SNPs that meet the original criteria applied to 31 populations. Many of these candidate SNPs (now >108 with Fst <0.06) are molecularly close and/or genetically linked making them unsuitable for studies involving relationships. However, since the ability of various SNPs to be robustly typed by various methodologies, ideally in multiplex reactions, needs to be evaluated before deciding on a final panel, it is appropriate to keep all these markers among the candidates until the laboratory aspects can be evaluated. We think it likely that many genetically independent (unlinked) markers will be found suitable. We advocate screening still more SNPs to assure identifying a sufficient number meeting broad forensic criteria. We also believe that all of the near-final candidates should be evaluated on many additional populations so that reasonably small (e.g. $<10^{-12}$) genotype frequencies can be demonstrated to occur broadly.

We continue to search for additional SNPs meeting the same criteria that we applied in identifying the best 40 SNPs. We have also been collaborating with the SNPforID consortium in evaluating some of their more promising markers to see which ones might be comparable to our best 40 SNPs. By October 2007 we have found that 3 out of 47 SNPforID markers meet the dual requirements of high heterozygosity (≥0.4) and Fst ≤0.060 when typed on the same panel of 40 population samples that we studied in finding the best 40 SNPs. When the stringency of the criteria are reduced somewhat by eliminating from the population panel various small, isolated groups and keeping the remaining 31 groups representative of many of the world's largest

populations, then we find a total of 9 SNPforID markers with minimum heterozygosities of 0.4 and Fst(31pops)≤0.060.

We have also sifted through SNPs that we have studied on other (non-NIJ) projects for additional markers that could be useful to consider for various identification panels. In this way we have identified a resource consisting of 31 additional SNPs that meet the combined high information content (heterozygosity ≥0.4) and low variability across ethnic groups (Fst (40pops) ≤0.060) but at least half of these markers are not sufficiently far away from the SNPs in the best 40 panel to be considered immediately as useful additions to the best 40 panel. Some of these markers may be useful alternate polymorphisms to consider when the candidates for a universal identification panel are being optimized for the typing method(s) that are yet to be evaluated. We also need to screen for more SNPs separated by larger chromosomal distances for other applications such as situations in which close biological relatives are routinely present. In our panel of 40-best SNPs only 25 of the 40 SNPs would meet the more stringent criteria needed for evaluating close biological relatives. In October of 2007 we placed on our website from the various sources discussed (NIJ-funded, other Kidd lab projects, and collaboration with SNPfor ID group) (http://info.med.yale.edu/genetics/kkidd/SNPdata2007.pdf) a list of the 108 candidate SNPs meeting the combined Fst (<0.06) and heterozygosity (>0.4) criteria for 31-populations. After that list was placed on the web we finished evaluating more of the SNPforID markers on the 31 and 40 population sets and identified one additional SNPforID marker meeting the criteria for 31 populations but not for the 40 population set. Thus, as of November 2007 we have a total of 109 candidate SNPs meeting the criteria for the 31-populations.

2.6 Progress on AISNPs (Ancestry informative SNPs)

We have made a strong start on developing a panel of high Fst 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. 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 our 40 populations) of those markers did not allow any further subdivision of populations even when regions were examined separately using the program STRUCTURE. We have sought appropriate markers for robustly resolving geographic and population structure with multiple screening procedures: (1) high Fst 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 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. The first two screening approaches are aimed at providing good assignment to continent (with North and South America combined). The first three approaches yielded 109 markers as an initial exploratory dataset. Using these resources one cannot know from the limited data available how informative any marker will be. Indeed, not all of these SNPs have high Fst values when typed on the 40 populations, though all but 18 are above the mean of the random distribution. Thus, we have continued to collect data on high Fst SNPs.

Our developing AISNP panel currently consists of 249 candidate SNPs. When four continental clusters are considered, the populations in Africa, Europe, east Asia, and the Americas our 249-SNP panel gives greater certainty of assignment of individuals using various

statistics and visually reflected in the greater homogeneity of the graphics produced by STRUCTURE relative to the Lao et al. (2006) panel.

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 assignment of individuals.

With such a large number of SNPs, we can extend our analyses to populations located between continents. However, we realize that 249 SNPs is not a reasonable size for an investigative AISNP panel, and we plan to continue exploring methods of decreasing the number of SNPs while retaining informativeness and precision. We also continue searching through a variety of data sets that other groups have already created looking for potential AISNP candidates that can be tested more thoroughly on our large collection of population samples from around the world.

3. Background and Rationale

Single Nucleotide Polymorphisms (SNPs) are being considered for a potentially useful role in forensic human identification [Amorim & Pereira, 2005; Sanchez et al., 2004; Sanchez et al., 2003.]. 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. 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. (4) Finally, SNP typing can be done very quickly for large numbers of SNPs on a chip.

There are two commonly recognized problems with SNPs replacing STRPs for individual identification in forensics. 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. 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. While those probabilities might differ by several orders of magnitude, the individual probabilities calculated for 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 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 in forensics, some of the criticisms of Lewontin and Hartl [Lewontin & Hartl, 1991] might have some validity.

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. The problem is how to identify appropriate individual identification SNPs (IISNPs) and demonstrate their low allele frequency variation sufficiently well for forensic purposes.

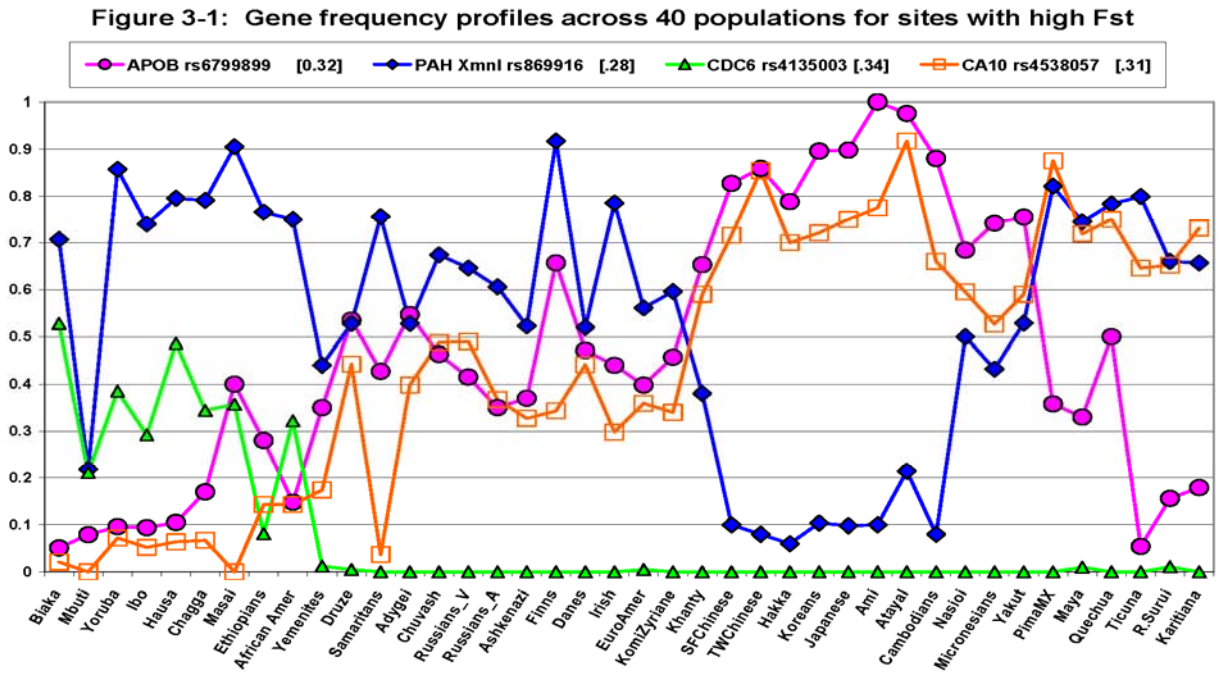


Figure 3-1. The frequencies of one allele at each of four SNPs with high variation in allele frequencies among populations. The SNPs are identified by their rs number in dbSNP and the symbol of the genetic locus in which each occurs; the data are in ALFRED. The populations are arranged by geographic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

4. Goals

With a unique collection of population samples (Table 4-1), a well-equipped molecular laboratory, 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 need for the population data for forensic SNPs was made 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 seen in various publications [e.g. Kidd et al. 2004; Tishkoff & Kidd 2004]

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 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 Fst and high average heterozygosity and (2) with globally high Fst 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 is to identify 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 optimizing an ancestry informative panel and indeed, our progress in that area has focused on developing criteria.

TABLE 4-1 The 40 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
	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
	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
	EuroAmericans▼	92	PO000020C	SA000020C
N.W. Asia	Komi Zyriane	40	PO000326L	SA000489V
	Khanty	50	PO000325K	SA000488U
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
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

▼ indicates the seven population samples included in the initial screening of polymorphisms.
* Samples (usually a subset) contributed to the HGDP-CEPH panel, Paris.
† Samples with many related individuals; most analyses include only unrelated individuals.
☼ Source: National Laboratory for the Genetics of Israeli Populations
╬ EuroAmericans are unrelated individuals married into large, multigenerational pedigrees that were collected for studies of genetic linkage and human variation..

Table 4-1. The 40 populations studied. The seven population samples included in the initial screen are indicated by ▼. The ALFRED UIDs can be used to retrieve the descriptions of the populations and of the specific samples of those populations.

5. Individual Identification Panel: Proof of Principle

5.1 Strategy

To obtain SNPs with high global heterozygosity and low inter-population variation, we pursued a strategy of four steps to successively enrich for appropriate SNPs. First, we identify likely candidate polymorphisms. We then screen these on a few populations. We then test the "best" of those markers on many populations. Finally, we retain 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 Fst [Wright 1951] as a standardized measure of the variance in allele frequencies among populations.

For our initial identification of likely candidates, we have 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 provides off-the-shelf assays that are guaranteed to work with no effort on our part to design and optimize an assay. From Applied Biosystems we obtained the frequencies for those TaqMan markers that had allele frequency data on four populations (African Americans, European Americans, Chinese, and Japanese). These markers were then rank ordered by both average heterozygosity and minimal difference in allele frequency among the four populations. We then choose markers with average heterozygosity >0.45 and Fst <0.01. Once a marker is selected for testing, no other markers are selected within 1Mb of that marker.

For the initial screen in our lab we selected a total of 371 individuals from seven populations in order to sample genetic variation from all major geographical regions: European Americans (92), Biaka (66), Hausa (39), Ibo (48), Cambodians (25), Taiwanese Chinese (49), and Maya (52). These and the other populations studied are listed in Table 4-1 along with the unique identifiers (UIDs) in ALFRED, the ALlele FREquency Database (http://alfred.med.yale.edu), for the descriptions of the populations and samples.

The second screening of the best of the markers from the initial screen consisted of samples from the additional 33 populations (Table 4-1). Thus, markers making it through the second screen have been typed on ~2,100 individuals from 40 populations. By geographic region the numbers of samples are: Africa (including African Americans) (459), Southwest Asia (211), Europe (558), Northwest Asia (90), East Asia (345), Northeast Asia/Siberia (51), Pacific Islands (60), North America (105), and South America (191).

5.2 Screening Criteria

To determine reasonable screening values we analyzed data we had collected on other projects. About 900 SNPs, more or less randomly selected with respect to Fst, had been typed on 38-42 populations including all or most of the 40 populations being used in this study. 277 of these SNPs had average heterozygosities ≥0.4 for the 7 populations. For each of these markers we plotted its Fst across all of the populations against its Fst calculated for the seven populations in the initial screen (Figure 5-1). There is a significant, but far from perfect correlation. We chose an initial cut-off value of 0.02 for the 7-population Fst as giving the largest proportion of markers with low Fst for all populations. Inspection of the scatterplot shows that we could

increase this value and still identify markers with low Fst on the larger population set and that option may be considered in the future if more markers are needed.

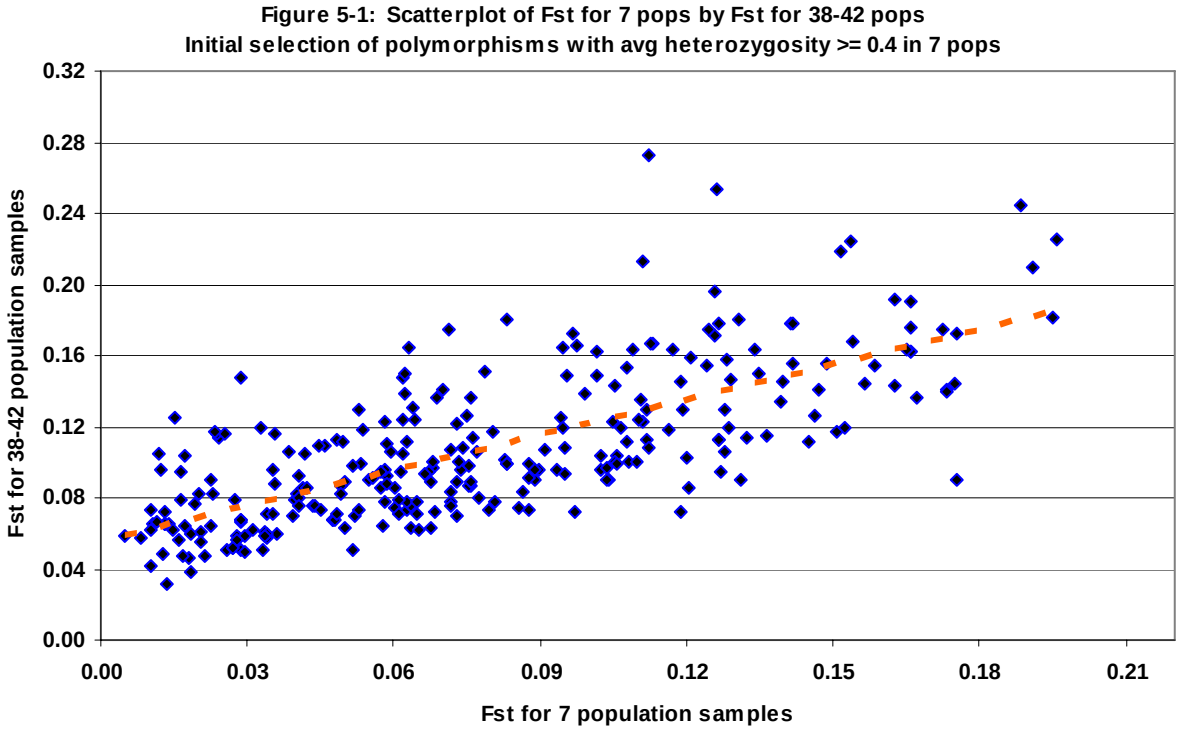


Figure 5-1. Scatterplot of Fst values for 277 SNPs (selected for high heterozygosity on 7 populations) calculated on 7 populations and for 38-42 populations that include the 7. The Pearson correlation coefficient is 0.72.

Finally, we are using an Fst 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. 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 630 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.

The statistical independence of the markers was assessed by calculating Δ^2 [Kidd et al., 2004] for all of the 171 unique, pairwise combinations of the final 19 markers within each of the 40 populations. The Δ^2 value, sometimes called r^2, is a measure of linkage disequilibrium (LD),

i.e., association of alleles at different loci. The LD values were then examined in various ways for evidence of meaningful associations among the markers.

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 19 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 19 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 Yield

After screening the Applied Biosystems Taqman Assays catalogue list of 90,483 SNPs, we identified 2,723 with Fst < 0.01 and average heterozygosity > 0.45 across all three of their populations (African American, European American and East Asians). We selected the best 195 markers separated by at least 1 Mb for testing on the seven populations listed earlier. Results for two markers were unacceptable. (one failed Hardy-Weinberg, the other did not allow clear allele calling) leaving data for 193 SNPs.

The Fst distribution of the 193 SNPs using data from seven populations is given in Fig. 5-2. This figure shows that Fst values for the seven populations can be considerably larger than the value of 0.01 for three populations that was the initial selection criterion. Yet, the distribution is shifted to lower values than that for the 38–42 populations. Given the correlation (Fig. 5-1) between the seven population and 38–42 population Fst values, we should be enriching for low Fst across all populations. Thirty-five SNPs had an Fst of 0.02 or less and these were then typed on all 40 populations. Fig. 5.3 compares the Fst values for these 35 markers on seven and 40 populations. Interestingly, in contrast to the positive correlation of the two Fst calculations seen in Fig. 5.1, at this low end of the distribution no significant correlation exists. The heterozygosities calculated for the initial three populations (>0.45) remain high for the 40 populations (>0.43 for 19 best SNPs; >0.37 for 35 SNPs). Finally, 19 SNPs met the criterion of Fst of 0.06 or less for all 40 populations (Fig. 5-3). These SNPs are listed in Table 5-1.

Figure 5-2. Comparison of Fst distributions

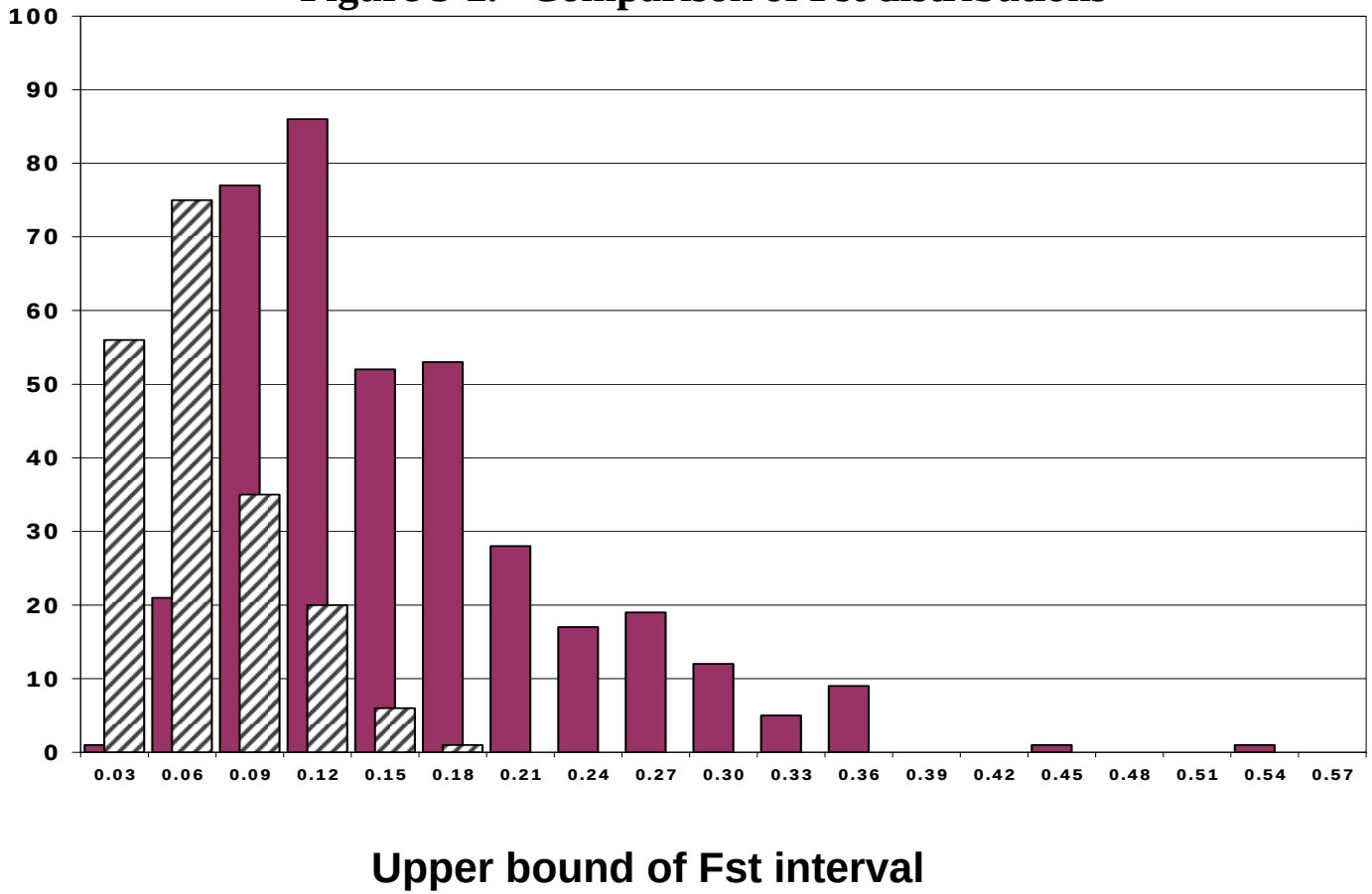


Figure 5-2. Comparison of Fst distributions. The solid bars represent the Fst for reference markers (not pre-selected for Fst) calculated for 38-42 populations. The cross-hatched bars represent the Fst for the 193 markers calculated for seven populations.

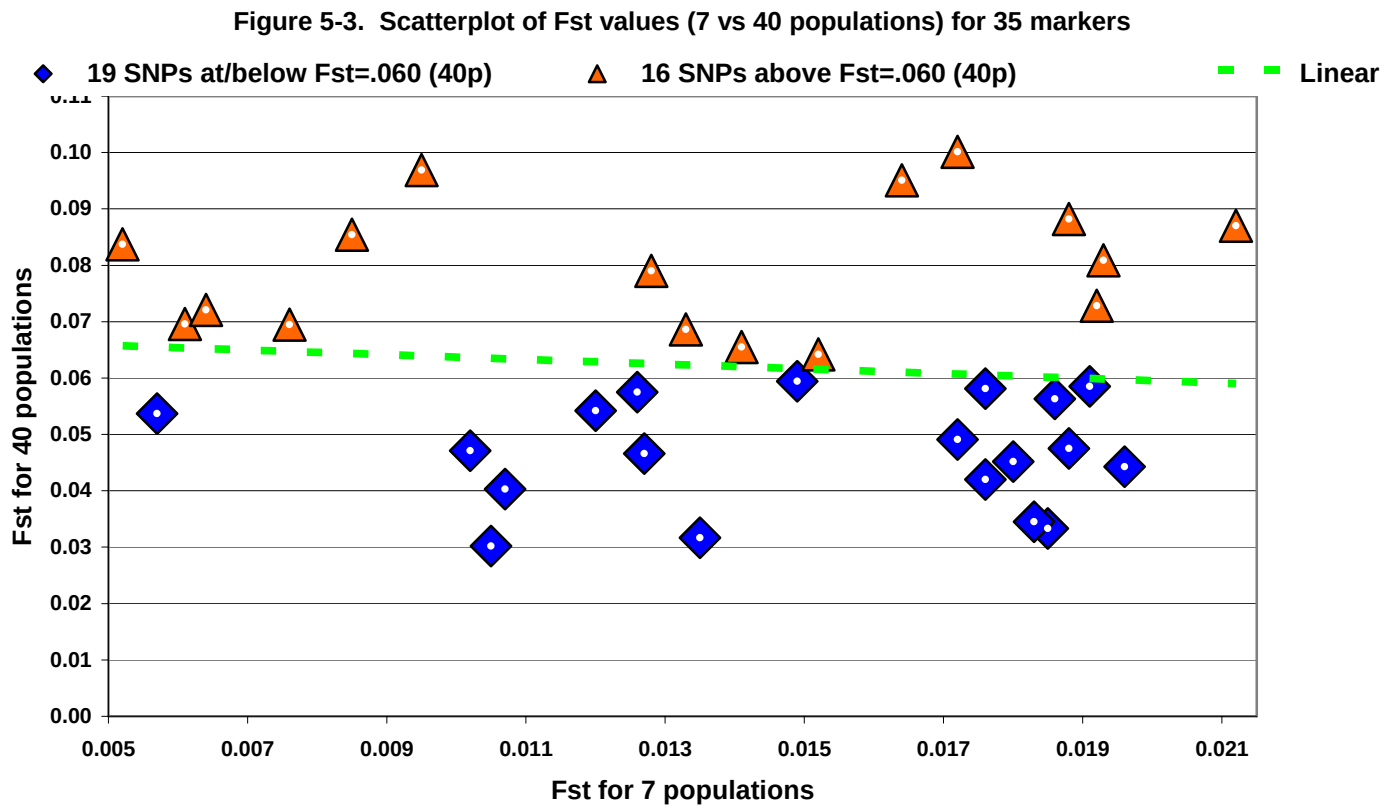


Figure 5-3. Scatterplot of the 35 markers tested on all 40 populations by the Fst values for the seven populations in the initial screen and for all 40 populations. The 19 SNPs included in the panel are plotted as diamonds; the 16 SNPs with final Fst above 0.06 are plotted as triangles. The regression is plotted as a dashed line; the Pearson correlation coefficient is –0.10. Note that five SNPs have 40-population Fst values between 0.06 and 0.07.

TABLE 5-1
The 19 best polymorphisms sorted by Fst value based on 40 population samples

Chr	Cytogenetic Band Position	Locus Symbol	ABI Catalog #	dbSNP rs#	Nt Position UCSC May 2004	ALFRED Site UID	Fst 40 p	Fst 7 p	Avg. Het. 40 p	Avg. Het. 7 p
4	p12	GABRA2	C___8263011_10	rs279844	46,170,583	SI001391O	0.0302	0.0105	0.485	0.495
13	q32.3	PHGDHL1	C___1619935_1_	rs1058083	98,836,234	SI001402H	0.0317	0.0135	0.464	0.484
5	q31	SPOCK	C___2556113_10	rs13182883	136,661,237	SI001390N	0.0333	0.0185	0.471	0.489
1	q21.3-q22	LY9	C___1006721_1_	rs560681	157,599,743	SI001392P	0.0345	0.0183	0.434	0.439
10	q26	HSPA12A	C___3254784_10	rs740598	118,496,889	SI001393Q	0.0403	0.0107	0.463	0.477
6	q22	TRDN	C___2140539_10	rs1358856	123,936,677	SI001407O	0.0400	0.0176	0.473	0.486
18	p11.3	RAB31	C___1371205_10	rs9951171	9,739,879	SI001395S	0.0443	0.0196	0.474	0.490
1	P36	PRDM2	C____340791_10	rs7520386	13,900,708	SI001394R	0.0452	0.0180	0.477	0.490
6	p24-p22.3	HIVEP1	C___9371416_10	rs13218440	12,167,940	SI001397U	0.0466	0.0127	0.457	0.479
6	q24.3	SASH1	C___1256256_1_	rs2272998	148,803,149	SI001398V	0.0471	0.0102	0.468	0.490
2	q31.3	CERKL	C___1276208_10	rs12997453	182,238,765	SI001396T	0.0475	0.0188	0.445	0.466
6	q25	SYNE1	C___2515223_10	rs214955	152,789,820	SI001403I	0.0491	0.0172	0.475	0.491
4	q21.1	RCHY1	C___1880371_10	rs13134862	76,783,075	SI001400F	0.0537	0.0057	0.456	0.467
10	q23.3-q24.1	SORBS1	C___7538108_10	rs1410059	97,162,585	SI001399W	0.0540	0.0120	0.471	0.482
5	qter	ADAMTS2	C___3153696_10	rs338882	178,623,331	SI001401G	0.0563	0.0186	0.467	0.490
6	q22-q23	THSD2	C____411273_10	rs2503107	127,505,069	SI001406N	0.0575	0.0126	0.454	0.463
5	q35	LCP2	C___3032822_1_	rs315791	169,668,498	SI001404J	0.0581	0.0176	0.471	0.485
11	q23	KBTBD3	C___1636106_10	rs6591147	105,418,194	SI001409O	0.0585	0.0191	0.449	0.481
18	q11.2	B4GALT6	C___7459903_10	rs985492	27,565,032	SI001413J	0.0594	0.0149	0.468	0.487

Notes:
The locus symbol is sometimes that for the closest named gene identifiable.
Avg. Het. is the average heterozygosity
Nt. Position. is the nucleotide position of the polymorphism along the chromosome using the May 2004 build information from the University of California Santa Cruz genome center (counting from pter as origin).

Table 5-1. The best 19 SNPs sorted by the final Fst. For each SNP the table gives the position, locus name, various identifiers in different databases, and various statistics.

5.6 Validation for Forensic Use: Independence of 19 Best SNPs

As shown in Table 5-1, the 19 best SNPs are distributed across nine different chromosomes with four chromosomes having more than one SNP. In order to assess the independence of variation for the 19 markers, all pairwise LD values (Δ^2) were computed in each of the 40 population samples. The pattern of results across the 171 x 40 = 6,840 LD values clearly supports the conclusion that each SNP contributes essentially independent variation for each of the 40 population samples tested (data not shown). For the 171 unique SNP pairings, the average Δ^2 (each based on 40 populations) ranges from 0.01 to 0.06. The vast majority of the Δ^2 values are close to zero (e.g., 82.9% are values $\leq$0.05 and 94.9% are $\leq$0.11) and these are certainly not statistically different from equilibrium given our sample sizes. There is a positive bias in LD estimates that increases as sample size decreases [Teare et al., 2002]. This bias is demonstrated in our results by a strong negative correlation of –0.689 between sample size and the proportion of Δ^2 values >0.10 among the 40 population samples (data not shown).

The largest LD values ranging from 0.25 to 0.47 were examined in detail to see if they might contain evidence of weak levels of association. There are only 34 LD values in this range, the most extreme ½ of 1% of the 6,840 calculated. Of these 34 largest LD values 31 involve SNPs paired across different chromosomes. Several populations had more than one of these large LD values: Masai (N=22) had four, Samaritans (N=41) had two, Archangel Russians (N=34) had two, Atayal (N=42) had three, Cambodians (N=25) had four, Nasioi (N=23) had five, Surui (N=47) had three, and Karitiana (N=57) had four. There are several reasons for believing these represent chance. We note that 171 comparisons were done for each population and that all but three of these large LD values involve different chromosomes. These larger LD values likely represent the chance occurrences that can arise when carrying out a large number of

calculations. This seems especially so in conjunction with the bias in LD values for small samples since half of these involve samples of less than 40 individuals and all involve samples with less than the average of ~52 individuals per sample. Because there is no plausible biological explanation to expect SNP alleles on different chromosomes or those far apart on the same chromosome to be associated except by chance, we provisionally conclude all of these large LD values are simply a chance deviation. Additional study will be necessary to confirm this.

The three large LD values that involve markers located on the same chromosome are also likely due to chance. One involves a pair of markers that are at opposite ends of chromosome 1. Two involve markers on chromosome 6 that are 3.57 Mbp and 21.30 Mbp apart in the Surui and Karitiana, respectively. These three are included in Table 5-2, which summarizes the LD results for all pairs of markers on the same chromosome. All of the pairs in Table 5-2 have median LD values of 0.02 or less and mean values of 0.04 or less. As is evident from these low mean and median values, the maximum values are global outliers in all cases and probably represent chance in light of the many comparisons. Moreover, most of the populations involved are the smaller ones and most of the distances involved are several times greater than reports of confirmed LD. We expect that independent samples of these populations would not show these associations and provisionally conclude that these 14 SNPs in Table 5-2 are statistically independent.

TABLE 5-2
Statistical summary of pairwise LD values (Δ^2) across 40 population samples for all of the SNP pairs located on the same chromosome and the physical distance separating those SNPs

Chr	SNP pair		Separation (M bp)	N Pops	Median	Avg.	Min.	Max.	Max LD pop.
1	LY9	PRDM2	143.70	40	.02	.04	.00	.25	Masai
4	GABRA2	RCHY1	30.61	40	.01	.03	.00	.23	Atayal
5	SPOCK	ADAMTS2	41.96	40	.01	.03	.00	.15	Mbuti
5	LCP2	SPOCK	33.01	40	.01	.03	.00	.20	Nasioi
5	LCP2	ADAMTS2	8.96	40	.02	.04	.00	.21	Russians,Arch.
6	TRDN	SYNE1	28.85	40	.01	.03	.00	.22	Nasioi
6	TRDN	HIVEP1	111.77	40	.01	.03	.00	.21	Karitiana
6	TRDN	SASH1	24.87	40	.02	.03	.00	.14	Quechua
6	SYNE1	HIVEP1	140.62	40	.01	.03	.00	.20	Karitiana
6	SYNE1	SASH1	3.99	40	.02	.04	.00	.22	Cambodians
6	HIVEP1	SASH1	136.64	40	.02	.03	.00	.18	Adygei
6	THSD2	TRDN	3.57	40	.02	.04	.00	.28	R. Surui
6	THSD2	SYNE1	25.29	40	.02	.03	.00	.10	Pima, Mexico
6	THSD2	HIVEP1	15.34	40	.01	.02	.00	.13	Yemenite Jews
6	THSD2	SASH1	21.30	40	.02	.04	.00	.26	Karitiana
18	B4GALT6	RAB31	17.83	40	.01	.03	.00	.16	Nasioi

Table 5-2. Pairwise LD comparisons (as Δ^2) across 40 populations for markers on the same chromosome. The marker pairs are identified by the names of the loci containing the SNPs as given in Table 5-1. Physical distance (in Megabases) and the population in which the maximum Δ^2 occurred are given.

5.7 Statistics for the Preliminary 19-SNP Panel

The frequency of the most common 19-locus genotype in each population is given in Figure 5-4. Most values are less than 2×10^{-6} and the largest values are between 6.0×10^{-6} and 1.6×10^{-5}. These larger values are in small isolated populations such as the Samaritans, Nasioi, and American Indian tribes. These values are relevant in that they provide an upper bound to the match probability in any population.

Figure 5-5 presents the average match probability by population for 19 preliminary best SNPs. This value is the weighted average of the match probabilities of the 319 possible genotypes, assuming exact H-W ratios within each population. The values range across approximately one order of magnitude, from greater than 10^{-7} to slightly greater than 10^{-8}. 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 probability of discrimination is greater than 0.999999.

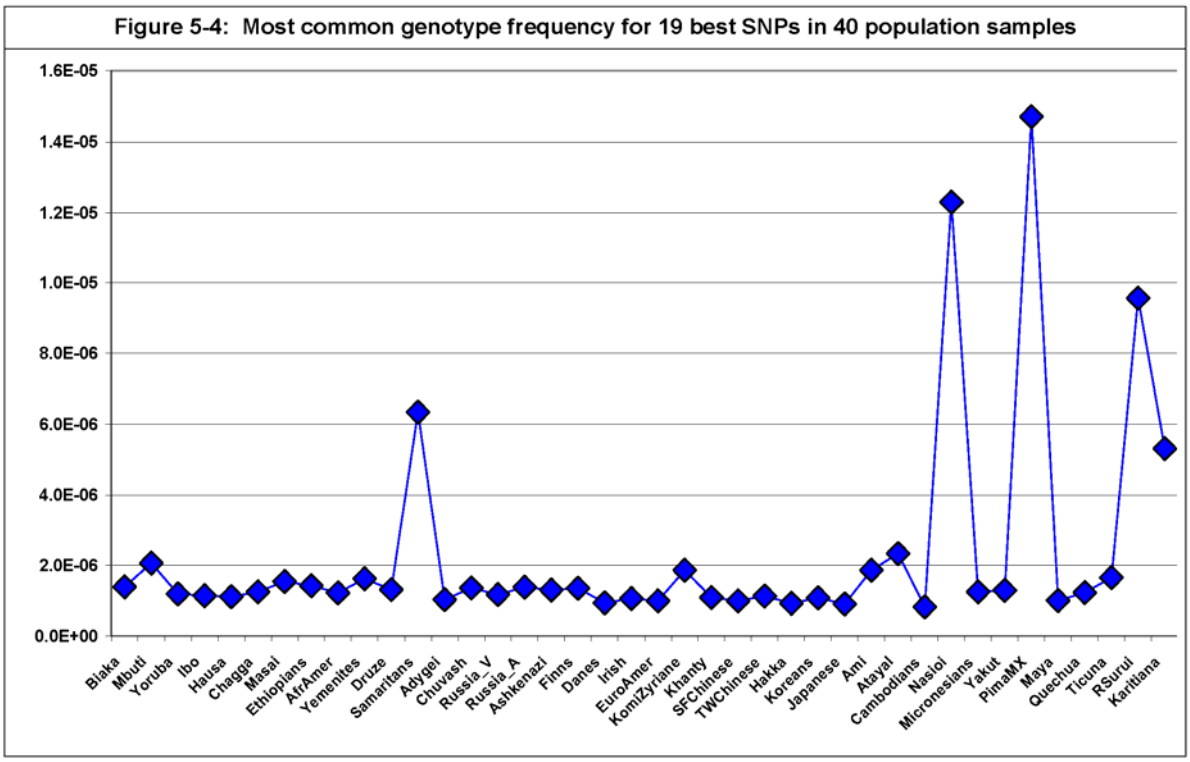


Figure 5-4. The frequency of the most frequent genotype for 19 SNPs in each population. Populations are ordered by geographic region from Africa on the left to South America on the right.

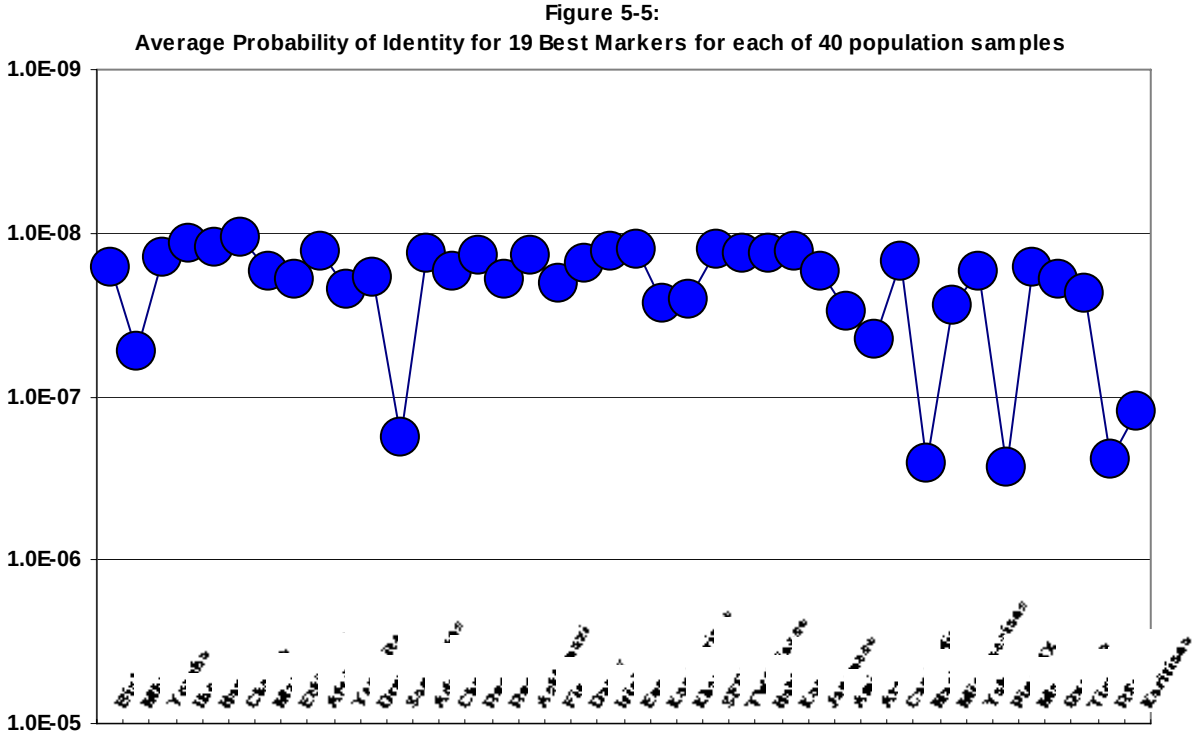


Figure 5-5. The average match probability for the best 19 markers for each of 40 population samples. Populations are ordered by geographic region as in Figure 5-4.

Vallone et al. (2005) tested 70 SNPs on three populations and found that 12 of them were sufficient to yield a unique genotype for each individual. While our panel of 19 markers did not result in unique genotypes for every individual, we tested over ten times as many individuals (~2,100 vs 189). The distribution (Table 5-3) of the number of loci matching for the more than 1.74 million pairwise comparisons of 1,895 individuals (with complete typings for the 19 best SNPs) shows that a very small percentage match at all the markers. We expect that doubling the number of markers will be more than sufficient to yield a unique genotype for each individual in our panel.

TABLE 5-3
All unique pairwise comparisons of individuals for 19 best SNPs; Overall results for 1,895 individuals in 40 population samples

Number of Genotype Differences	Within Groups	Across Groups	Total Comparisons
19	0	0	0
18	0	1	1
17	0	3	3
16	1	94	95
15	10	434	444
14	47	2,206	2,253
13	195	8,617	8,812
12	683	27,237	27,920
11	1,600	69,080	70,680
10	3,589	140,296	143,885
9	6,203	230,457	236,660
8	8,575	306,803	315,378
7	9,940	331,964	341,904
6	9,030	284,689	293,719
5	6,389	191,950	198,339
4	3,577	99,134	102,711
3	1,420	37,684	39,104
2	469	10,224	10,693
1	98	1,706	1,804
0	22	138	160
Total Pairings	51,848	1,742,717	1,794,565

Table 5-3. Unique pairwise comparisons of all individuals with complete typings for 19 best SNPs. The "within groups" column is the sum of all pairwise comparisons within each of the 40 populations. The "across groups" column summarizes all pairwise comparisons for which individuals are in different populations.

To explore the variation in match probabilities empirically we have calculated match probabilities for each individual in each of four populations: Yoruba, Adygei, Japanese, and Mexican Pima. Match probabilities were calculated using 10 sets of allele frequencies: one that varied by population--the empiric allele frequencies for the specific population--and nine

geographic region-specific frequencies that were used for all four populations. We then calculated the fold difference in match probabilities for each individual as the maximum/minimum of the ten match probabilities from the different allele frequency sets. Table 5-4 presents the mean, maximum, and minimum of those fold differences for the individuals in the population. These calculations were done for the 19 low-Fst marker panel in Table 5-1 and, as a "worst-case" example, for a panel of 19 high-Fst markers also tested on all 40 populations. The high-Fst markers included the APOB marker in Figure 3-1 and 18 others with similarly high Fst. As can be seen in Table 5-4, our proposed low-Fst panel had mean differences in match probabilities of 34- to 253-fold and maximum differences in match probabilities of essentially 1000-fold, depending on the frequency dataset used. In contrast, the high Fst panel had mean differences of 1.76×10^{9}- to 3.34×10^{14}-fold and could have had as much as a 10^{16}-fold difference, depending on frequency dataset used. For the low-Fst panel, the largest match probability for an individual was distributed quite randomly among the datasets, as expected for very similar frequency sets. For the high-Fst panel, the largest match probability tended to occur using the allele frequencies for the specific population.

TABLE 5-4
Empirical variation in match probabilities

Marker panel	Fold differences in match probabilities	Adygei	Japanese	Mexican Pima	Yoruba
19 low-Fst SNPs	Mean	1.02E+02	9.38E+01	1.31E+03	1.99E+02
	Maximum	6.67E+02	5.75E+02	3.01E+04	2.34E+03
	Minimum	7.82E+00	2.62E+00	1.31E+01	7.83E+00
19 high-Fst SNPs	Mean	5.96E+13	3.56E+13	5.11E+10	2.73E+16
	Maximum	2.43E+15	1.04E+15	7.99E+11	8.77E+17
	Minimum	3.38E+06	1.40E+05	1.37E+06	1.30E+09

Table 5-4. Empirical variation in match probabilities. Values given are for all individuals in the specific population samples. Calculations are based on 10 different sets of allele frequencies as described in the text.

6. A Provisional Panel of 40 IISNPs

6.1 Expanding the number of IISNPs

Our "final" provisional panel at the end of this funding period consists of the 40 best markers with a 40-population Fst below 0.06 and average heterozygosity > 0.4. Such markers correspond to 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 2.02×10^{-17} to 1.29×10^{-13}. These 40 SNPs therefore constitute 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. The strategy, methods and preliminary results for the first 19 of the 40 best markers identified were presented earlier in section 5. We have also published a paper (Pakstis et al., 2007) in the journal *Human Genetics* describing this panel of 40 best markers and some of the text and all the figures and tables in this section derive from that publication. We have deposited the gene frequency tables for all the markers we screened in ALFRED, the ALlele FREquency Database (http://alfred.med.yale.edu).

6.2 The Yield from Screening

We screened the 90,483 SNPs that have allele frequencies for four populations (European American, African American, Chinese, and Japanese) and identified 436 markers that we have typed on the seven-population screen described in section 5.1. Four failed to show acceptable clusters or failed Hardy-Weinberg ratios in multiple populations and were discarded as unacceptable/unreliable. In our initial study (Kidd et al. 2006) 193 of these were analyzed; those are included in these analyses. 73 SNPs or 17% of the total of 432 had an Fst of 0.02 or less on the seven populations and we typed these on all 40 populations.

The Fst distribution on the seven-population screen is shown in Figure 6-1. This is a very "wide" distribution considering that all of these markers had a three-population Fst of 0.01 or less. However, the majority of these Fst values (median=0.054) are below the mean and median of a distribution of markers unselected for Fst. Our published Fst distribution of 369 similarly unselected SNPs on 38 populations had a mean Fst of 0.138 and a standard deviation of 0.068 (Kidd et al., 2004); a recent update (unpublished) of this distribution has 813 SNPs on 40 populations with a mean of 0.139 and a standard deviation of 0.070. Clearly, for this range of populations and large number of SNPs the values are quite stable.

Figure 6-1. F_{st} distribution for 432 SNPs screened on 7 populations

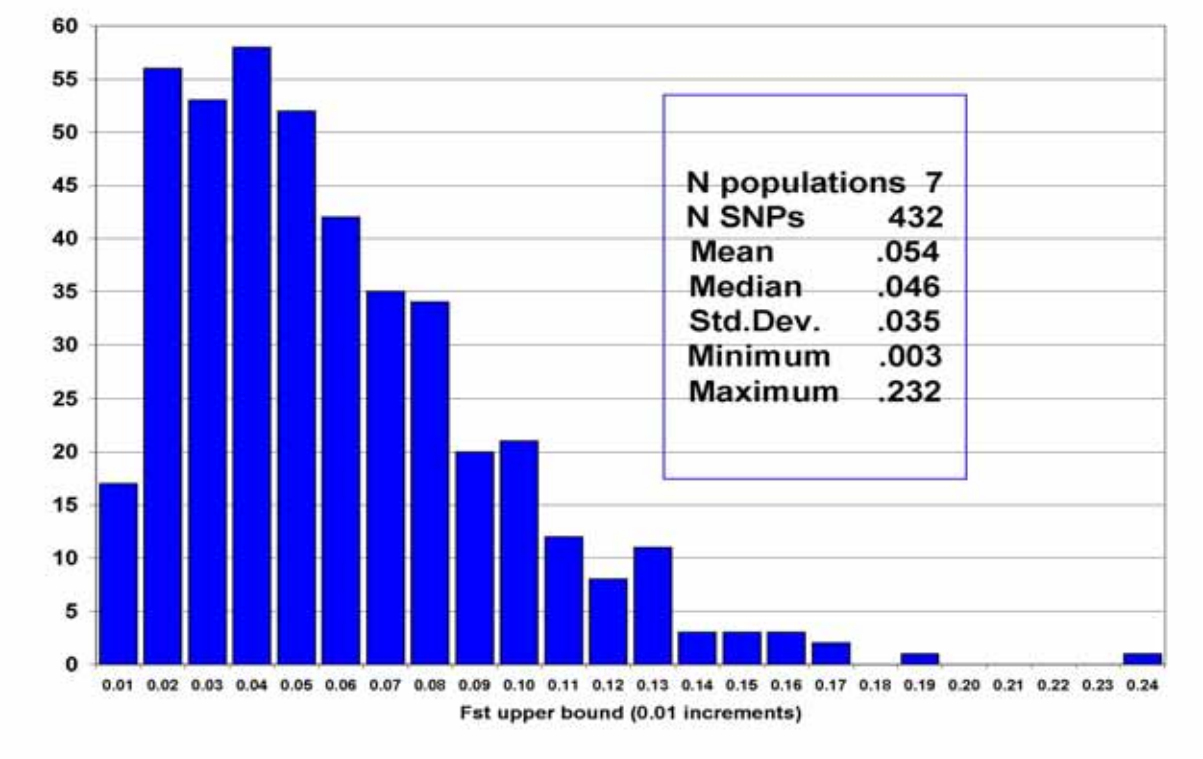


Figure 6-2. Scatterplot for 73 SNPs in follow-up

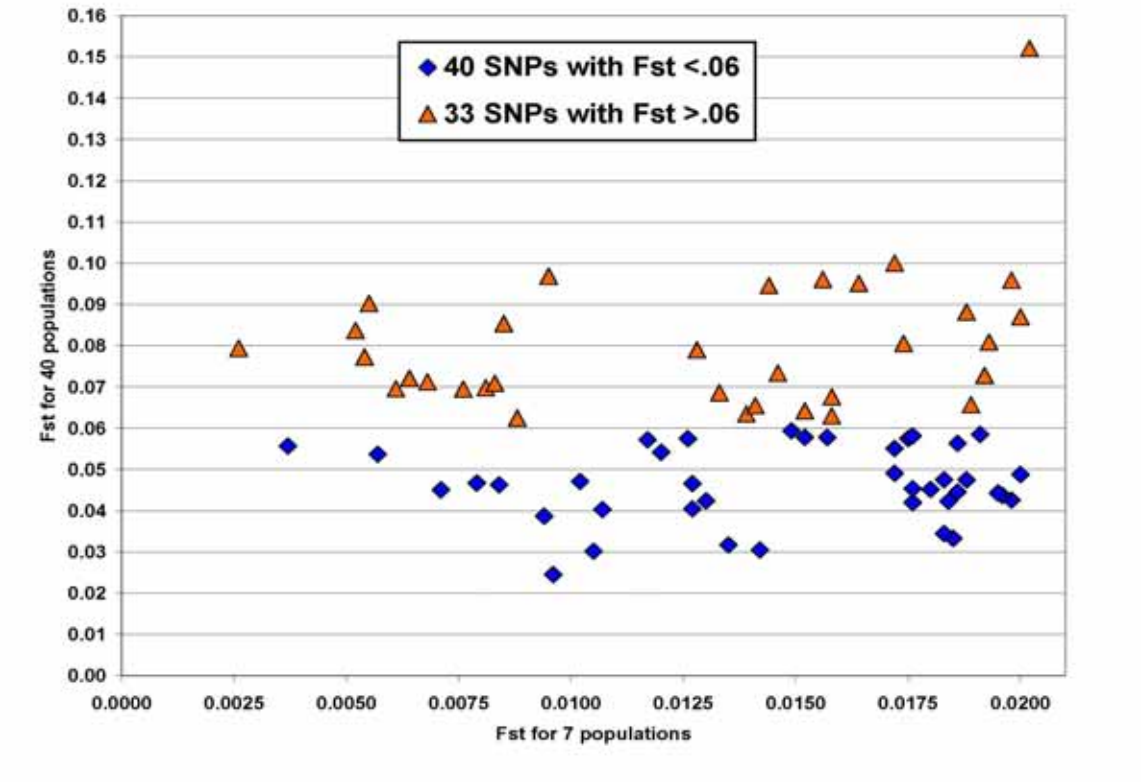


Figure 6-2 compares the Fst values for these 73 markers on 7 and 40 populations. Due to the contraction in the range of values studied at this low end of the global, multi-population Fst distribution no significant correlation exists. Having started our screening process with SNPs giving essentially identical allele frequencies in populations representing three regions of the world, we end with a relatively small fraction (~10%) of SNPs still showing little allele frequency variation when tested on a broader sample of populations from around the world. However, over 50% of those 73 SNPs with low Fst and high heterozygosity on our seven population screen still met our 40-population criteria. The heterozygosities calculated for the initial three populations (>0.45) remain high for the 40 populations (>0.43 for 40 best SNPs and

>0.37 for 73 SNPs). The 40 SNPs that met the criterion of an Fst of 0.06 or less for all 40 populations (Figure 6-2) are listed in Table 6-1.

TABLE 6-1. The 40 best polymorphisms sorted by Fst value based on 40 population samples

Chr	Cyto-genetic Band Position	†	Locus Symbol ‡	ABI Catalog #	dbSNP rs#	Nt Position UCSC May 2004	ALFRED Site UID	Fst 40 p	Fst 7 p	Avg. Het. 40 p	Avg. Het. 7 p
11	q23.2		IGSF4	C___2450075_10	rs10488710	114,712,386	SI001899B	0.025	0.010	0.441	0.460
4	p12	√	GABRA2	C___8263011_10	rs279844	46,170,583	SI0013910	0.030	0.011	0.485	0.495
4	q32.3	√	PALLD	C__11245682_10	rs6811238	170,038,345	SI001910L	0.031	0.014	0.485	0.492
13	q32.3	√	PHGDHL1	C___1619935_1_	rs1058083	98,836,234	SI001402H	0.032	0.014	0.464	0.484
5	q31	√	SPOCK	C___2556113_10	rs13182883	136,661,237	SI001390N	0.033	0.019	0.471	0.489
1	q23.3	√	LY9	C___1006721_1_	rs560681	157,599,743	SI001392P	0.035	0.018	0.434	0.439
8	p21	√	FZD3	C___2049946_10	rs10092491	28,466,991	SI001900K	0.039	0.009	0.456	0.458
10	q26	√	HSPA12A	C___3254784_10	rs740598	118,496,889	SI001393Q	0.040	0.011	0.463	0.477
20	p12.1	√	C20orf133	C___2997607_10	rs445251	15,072,933	SI001912N	0.041	0.013	0.463	0.473
6	q22		TRDN	C___2140539_10	rs1358856	123,936,677	SI001427O	0.042	0.018	0.473	0.486
15	q13	√	intergenic	C__11673733_10	rs1821380	37,100,694	SI001913O	0.042	0.018	0.464	0.474
20	q13.1	√	intergenic	C___2508482_10	rs1523537	50,729,569	SI001914P	0.042	0.013	0.472	0.476
18	q11.1		ZNF521	C____105475_10	rs7229946	20,992,999	SI001901L	0.043	0.020	0.464	0.456
20	p11.1		SSTR4	C___3206279_1_	rs2567608	22,965,082	SI001902M	0.044	0.020	0.475	0.490
18	p11.3	√	RAB31	C___1371205_10	rs9951171	9,739,879	SI001395S	0.044	0.020	0.474	0.490
3	q29	√	ATP13A4	C__25749280_10	rs6444724	194,690,082	SI001903N	0.045	0.019	0.468	0.489
6	q16.1	√	intergenic	C___1817429_10	rs1336071	94,593,976	SI001915Q	0.045	0.007	0.472	0.495
1	p36	√	PRDM2	C____342791_10	rs7520386	13,900,708	SI001394R	0.045	0.018	0.477	0.490
7	p22	√	intergenic	C___2572254_10	rs1019029	13,667,516	SI001916R	0.045	0.018	0.472	0.485
22	q11.2		loc388882	C__11522503_1_	rs2073383	22,126,725	SI001911M	0.046	0.008	0.452	0.474
6	p24.1	√	HIVEP1	C___9371416_10	rs13218440	12,167,940	SI001397U	0.047	0.013	0.457	0.479
6	q22.31		intergenic	C___1152009_10	rs1478829	120,602,393	SI001917S	0.047	0.008	0.474	0.491
6	q24.3		SASH1	C___1256256_1_	rs2272998	148,803,149	SI001398V	0.047	0.010	0.468	0.490
22	q12.3	√	loc650568	C__11887110_1_	rs987640	31,884,062	SI001918T	0.048	0.018	0.476	0.488
2	q31.3	√	CERKL	C___1276208_10	rs12997453	182,238,765	SI001396T	0.048	0.019	0.445	0.466
10	p15.1	√	DNMT2	C___2822618_10	rs3780962	17,233,352	SI0019040	0.049	0.020	0.475	0.490
6	q25	√	SYNE1	C___2515223_10	rs214955	152,789,820	SI001403I	0.049	0.017	0.475	0.491

4	q21.1		RCHY1	C___1880371_10	rs13134862	76,783,075	SI001400F	0.054	0.006	0.456	0.467
10	q24.3		SORBS1	C___7538108_10	rs1410059	97,162,585	SI001399W	0.054	0.012	0.471	0.482
16	p13.3	√	a2bp1	C__31419546_10	rs7205345	7,460,255	SI001905P	0.055	0.017	0.469	0.487
7	q33	√	PTN	C___3004178_10	rs321198	136,487,093	SI001906Q	0.056	0.004	0.457	0.489
5	qter	√	ADAMTS2	C___3153696_10	rs338882	178,623,331	SI001401G	0.056	0.019	0.467	0.490
4	q32.1		intergenic	C___7428940_10	rs1554472	157,847,511	SI001919U	0.057	0.012	0.471	0.494
2	p25.2	√	GRHL1	C___2073009_10	rs1109037	10,036,320	SI001909T	0.058	0.018	0.467	0.482
6	q22.3		RSPO3	C____411273_10	rs2503107	127,505,069	SI001426N	0.058	0.013	0.454	0.463
6	q24		EPM2A	C___2223883_10	rs447818	145,910,689	SI001907R	0.058	0.015	0.471	0.479
5	q33.3		TTC1	C___1995608_10	rs7704770	159,420,531	SI001908S	0.058	0.016	0.450	0.456
5	q35		LCP2	C___3032822_1_	rs315791	169,668,498	SI001404J	0.058	0.018	0.471	0.485
11	q23	√	KBTBD3	C___1636106_10	rs6591147	105,418,194	SI0014090	0.059	0.019	0.449	0.481
18	q11.2		B4GALT6	C___7459903_10	rs985492	27,565,032	SI001413J	0.059	0.015	0.468	0.487
							Averages:	0.047	0.015	0.465	0.480

Notes:

† Check (√) marks in this column identify the set of 25 polymorphisms that are "un-linked" (as well as being independent at the population level based on the LD tests) because they are more than 50 centi-Morgans (genetic map distance) from other markers on the same chromosome.

‡ The locus symbol is sometimes that for the closest named gene identifiable (e.g. LCP2 gene is ~11kb from rs315791). Gene symbols (e.g.a2bp1, loc650568) that are in lower case are un-official symbols in current use and may change in the future. Official gene symbols assigned by the Human Gene Nomenclature committee are typed in uppercase. "Intergenic" appears where no official or unofficial symbols are in use and the nearest known genes are very far away.

Avg. Het. is the average heterozygosity

Nt. Position. is the nucleotide position of the polymorphism along the chromosome using the May 2004 build information from the University of California Santa Cruz genome center (counting from pter as origin).

Some minor corrections and updates have been made here compared to overlapping entries in TABLE 2 of FSI (2005) preliminary report. Gene symbol RSPO3 replaced THSD2 as official gene symbol since publication of the preliminary report. The ALFRED Site UIDs have been corrected for rs1358856 and rs2503107 and the ABI Catalog # is corrected for rs7520386; in each case a single character has been changed.

The allele frequencies for the 73 SNPs in this study that were followed up on 40 population samples can be found in ALFRED. We are in the process of adding to ALFRED the allele frequencies for the additional 359 SNPs typed only in the 7 population sample screening step.

Missing typings were not concentrated in any population sample or SNP. For the 7 population screening (371 individuals) of 432 SNPs, 95.9% of the 160,272 typings succeeded and 4.1% failed. For the individual populations, missing/failed typings ranged from 1.6% in the Cambodians to 5.8% in the Maya. For 2,053 individuals in 40 population samples, 98.8% of the 82,120 possible typings for the 40 best SNPs succeeded and 1.2% failed. An average of 39.51 SNPs were typed per individual; 97.86% of the individuals had typings completed for 36 to 40 of the SNPs. For individual populations the rate of missing typings ranged from 0.3% (Chagga, Komi Zyrian) to 2.6% (Ethiopians, Nasioi) and had a simple average of 1.2% (1.1% median). For the 40 SNPs individually the rate of missing typings ranged from about 0.1% to 3.9%. So far as we can tell, it was the random occurrences of these few missing typings that resulted in the relatively low (~76%) frequency of individuals with complete typing results for all 40 SNPs (Table 6-3).

TABLE 6-3 All unique pairwise comparisons of individuals for 40 best SNPs Overall results for 1,568 individuals with complete typings in 40 population samples			
Number of Genotypes Matching	**Within Groups**	**Across Groups**	**Combined Comparisons**
0	**0**	**0**	**0**
1 or 2	**0**	**2**	**2**
3 or 4	**4**	**125**	**129**
5 or 6	**42**	**1992**	**2034**
7 or 8	**321**	**16101**	**16422**
9 or 10	**1527**	**67567**	**69094**
11 or 12	**4009**	**173446**	**177455**
13 or 14	**7178**	**279932**	**287110**
15 or 16	**8744**	**297429**	**306173**
17 or 18	**7090**	**211505**	**218595**
19 or 20	**4025**	**102294**	**106319**
21 or 22	**1613**	**33844**	**35457**
23 or 24	**515**	**7698**	**8213**
25 or 26	**174**	**1116**	**1290**
27 or 28	**70**	**127**	**197**
29 or 30	**21**	**8**	**29**
31 or 32	**7**	**1**	**8**
33 or 34	**1**	**0**	**1**
35 or 36	**0**	**0**	**0**
37 or 38	**0**	**0**	**0**
39 or 40	**0**	**0**	**0**
Totals	**35341**	**1193187**	**1228528**

6.3 Independence of the 40 Best SNPs

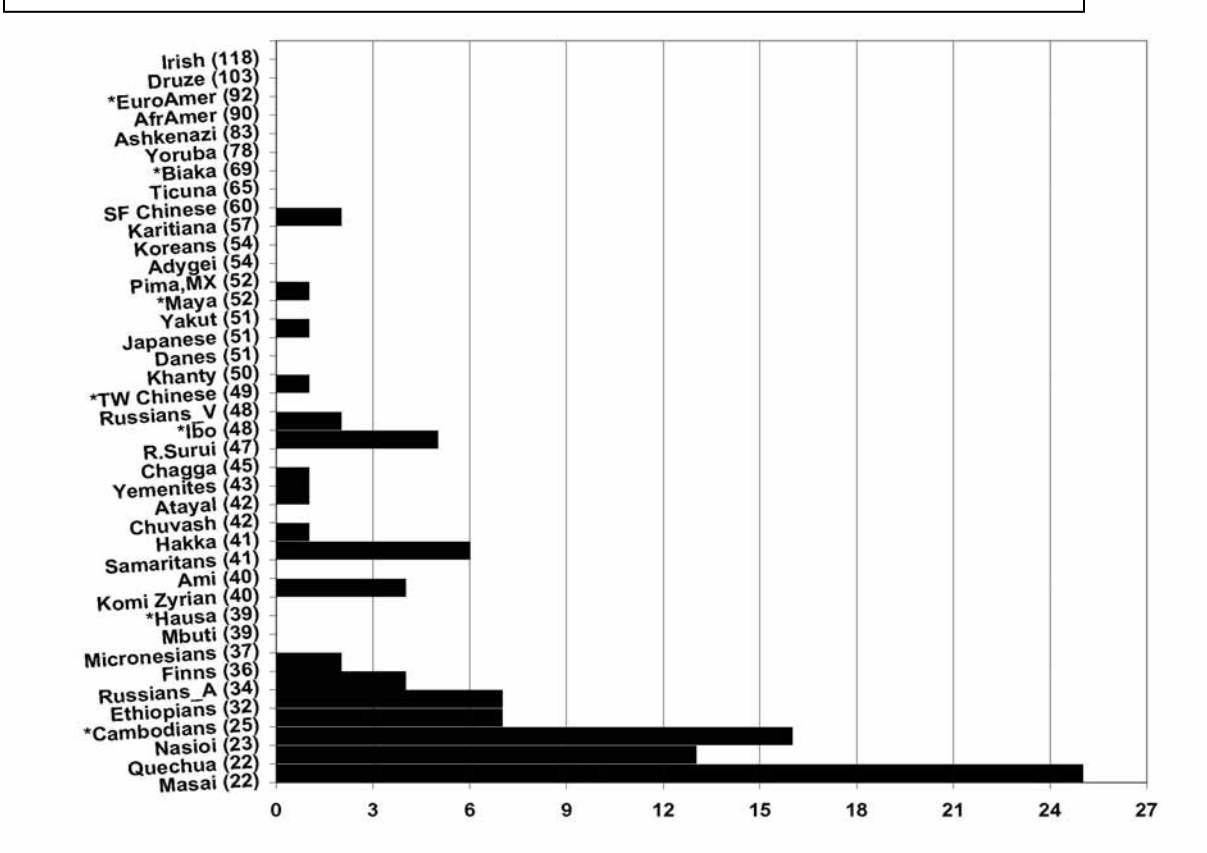
As shown in Table 6-1, because the ascertainment did not consider chromosomes per se, the 40 best SNPs are distributed across only 16 different autosomes with eleven chromosomes having more than one SNP. In order to assess the population independence of variation for the 40 markers, all pairwise LD values (r^2) were computed in each of the 40 population samples. The pattern of results across the 780 x 40 = 31,200 LD values clearly supports the conclusion that each SNP contributes essentially independent variation for each of the 40 population

samples tested. The vast majority of the r^2 values are close to zero (e.g., the median is 0.010 and the average is 0.029) and these are not statistically different from equilibrium given our sample sizes and the numbers of tests done. The distribution of nominal significance levels is approximately what can be expected by chance with an average across populations of 11.1% of the 780 comparisons in a population nominally significant at the 0.01 level, 3.7% nominally significant at the .001 level, and 1.3% nominally significant at the .0001 level. An ultraconservative Bonferroni correction assigns the equivalent 1% significance level to 0.0000128 (=0.01/780). 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 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 demonstrated in our results by our examination of the largest LD values ranging from 0.25 to 0.54 to see if they might contain evidence of weak levels of association. There are only 99 LD values in this range, the most extreme ⅓ of 1% of the 31,200 calculated. Of these 99 largest LD values 88 involve SNPs paired across different chromosomes. There are several reasons for believing these represent chance. We noted above that 780 comparisons were done for each population so that these large LD values that involve different chromosomes likely represent the chance occurrences that can arise when carrying out a large

number of comparisons. This seems especially so in conjunction with the bias in LD values for small samples since most of the 99 most extreme LD values involve samples of less than 40 individuals (Figure 6-3). Because there is no plausible biological explanation 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 except by chance, we provisionally conclude that all of these large LD values are chance deviations. Larger samples from these populations will be necessary to confirm this but they are not currently available.

Figure 6-3. Extreme LD (r^2) values by population



The 11 of the largest 99 LD values that involve markers located on the same chromosome are also likely due to chance. Table 6-2 summarizes the LD results for these SNP pairs on the same chromosome that have LD values >0.25. All of the marker pairs in Table 6-2 have median

LD values of 0.03 or less and mean values of 0.06 or less across the 40 populations. Most of these LD values for these pairs of markers are not significantly different from zero in the majority of population samples. These 11 SNP pairs involve distances of at least 2.8 megabases, most from 22 to 108 megabases. All of these distances are at least 10 times larger than the 200 or so kilobases that is the maximum extent of LD usually seen (Peltonen et al., 1999; Varilo et al., 2004). As is evident from these very low mean and median values, these maximum LD values are likely global outliers and probably represent chance in light of the many comparisons. Moreover, most of the populations involved are those with the smaller sample sizes and hence the values are biased upward. We expect that independent re-samplings of these populations would not show these associations and provisionally conclude that these 11 SNP pairs in Table 6-2 are statistically independent. 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) account for 3 of the 4 smallest intervals in Table 6-2.

TABLE 6-2

Statistical summary of pairwise LD (r^2) values across all populations and SNP pairs involving LD values >0.25 and that are located on the same chromosome plus the physical distance separating those SNPs

Chr	SNP pair ‡		Separation (M bp)	N Pops	Median	Avg.	Min.	Max.	Max LD pop.
4	2 GABRA2	28 RCHY1	30.612	40	0.01	0.03	0.00	0.30	Masai
5	5 SPOCK	37 TTC1	22.759	40	0.01	0.02	0.00	0.35	R. Surui
6	10 TRDN	17 intergenic	29.342	40	0.02	0.04	0.00	0.36	Masai
6	10 TRDN	22 intergenic	3.334	40	0.01	0.04	0.00	0.48	Quechua
6	10 TRDN	35 RSPO3	3.568	39	0.02	0.04	0.00	0.28	R. Surui
6	17 intergenic	27 SYNE1	58.195	40	0.01	0.03	0.00	0.31	Samaritans
6	21 HIVEP1	22 intergenic	108.434	40	0.01	0.04	0.00	0.29	Nasioi
6	22 intergenic	35 RSPO3	6.902	39	0.01	0.03	0.00	0.26	Karitiana
6	23 SASH1	36 EPM2A	2.892	40	0.03	0.06	0.00	0.53	R. Surui
20	9 C20orf133	14 SSTR4	7.891	40	0.02	0.05	0.00	0.44	Nasioi
20	12 intergenic	14 SSTR4	27.764	40	0.01	0.04	0.00	0.39	Nasioi

Notes:

‡ Under SNP pair column, the number in front of each marker symbol corresponds to the row in Table 6-1.

6.4 Statistics for the 40-SNP Panel

The frequencies of the most probable 40-locus genotype (assuming Hardy-Weinberg ratios) for each population are given in Figure 6-4 (by the line connecting the diamond shaped points). Most values are less than 10^{-12} and the largest value is less than 10^{-9}. 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.

Figure 6-4. Frequency of the most common genotype for 40 best SNPs (diamonds)
Match probabilities for 40 best SNPs (circles)

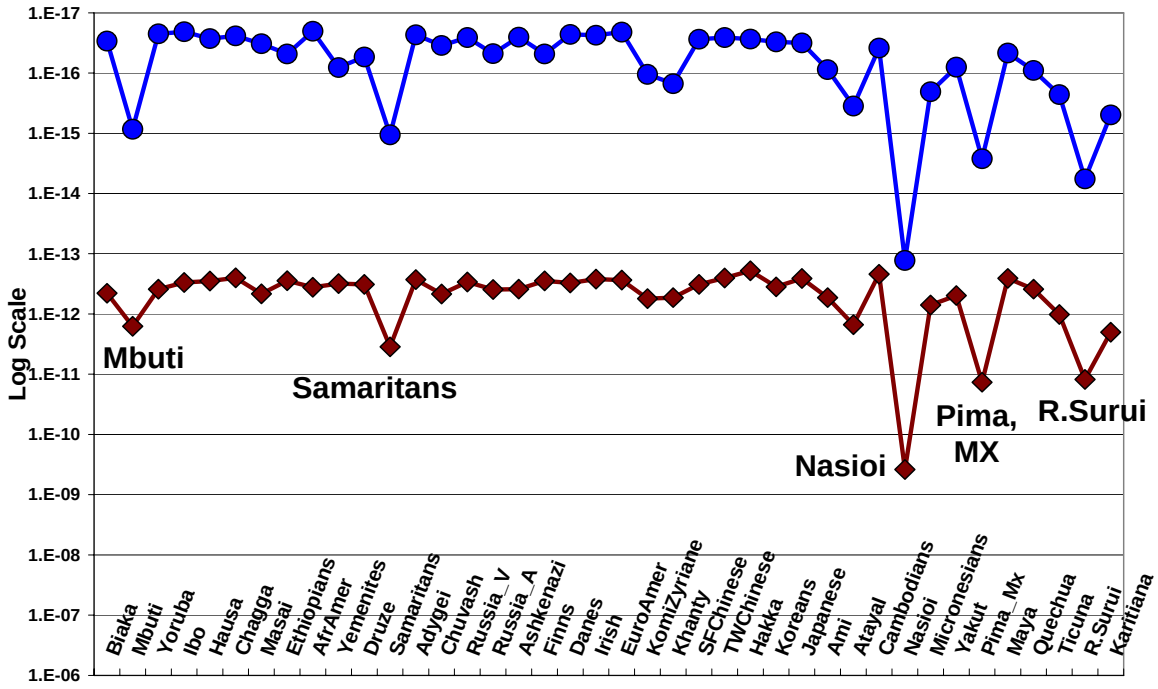


Figure 6-4 also presents the average match probability by population as shown by the values represented by filled circles. This value is the weighted average of the match probabilities of the 340 possible genotypes, assuming exact H-W ratios within each population. Most populations have values less than 10^{-16} but the values range across approximately four orders of magnitude, from less than 10^{-12} to less than 10^{-16}. We note only five populations have values about or larger than 10^{-15} and in none of those populations are there more than 10^{4} 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 probability of discrimination is greater than 0.999999999999.

6.5 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 Fst and high heterozygosity. The final panel of 40 SNPs has a narrow range for the average match probability across almost all populations. This validates the low Fst, high heterozygosity strategy for identifying SNPs that are appropriate for use in human identification. While Fst 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 with globally low Fst values. Also, our step-wise approach shows that the more different populations used to screen the more refined the result. A maximum global Fst of 0.06 functions well as a criterion even when small isolated populations are included. Similarly, because we also selected for high heterozygosity, the globally low Fst reflects not just similar allele frequency but also uniformly high heterozygosity. The actual cause of the low Fst in the SNPs we screen is most likely that they are drawn from the lower tail of the distribution of Fst for random neutral SNPs. The fact that 39 of the 40 best SNPs are located in intronic, intergenic, or untranslated regions reinforces this idea; one SNP is located in an exon of the SSTR4 gene and the polymorphism produces a nonsynonomous, missense change. We are not aware of any phenotypic consequences either of this polymorphism or of any polymorphism in linkage disequilibrium with any of the 40 SNPs. However, the possibility of such cannot be excluded.

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 Fst range for the 90,483 Applied BioSystems markers screened in the three populations we used for our original selection of candidate markers was $5.6x10^{-8}$ to 0.93, with mean = 0.087 and median = 0.063. Only 14,638 SNPs in this large pool had heterozygosities

>0.45 in all three populations and this marker subset had Fst values ranging from $5.6x10^{-8}$ to 0.1. We selected 436 SNPs to follow-up because they all had an Fst ≤0.01 on the initial three populations and the highest heterozygosities out of 2,723 SNPs with Fst ≤0.01. Nonetheless, on our seven-population screen we obtained a wide range of Fst values for the successful 432 SNPs extending from 0.003 to 0.232 (mean = 0.054, median = 0.046) (Figure 6-1). On the 813 essentially random markers we have tested on these seven populations the Fst range is even larger (range 0.020 to 0.534, mean = 0.139, sd = 0.070), but Fst for these potentially low Fst markers spans half of that range. The same imprecision in extrapolation occurs with our selection of markers with a seven-population Fst ≤ 0.02 for typing on all 40 populations, as can be seen in Figure 6-2. There is no correlation between the variation of Fst among SNPs in 40 populations and that in seven populations for this lower tail of the seven-population distribution. When markers are selected in a nearly random manner, there is a high correlation between the Fst seen on these seven populations and on all 40 populations (Figure 5-1) (Kidd et al., 2006), but the present results show the impossibility of accurately predicting or extrapolating to the relative Fst of a larger set of populations from values on a subset, even if that subset includes a set of populations from the four major continents.

We conclude that the 40 SNPs in our "final" panel are statistically independent at the population level. The median (0.01) and mean (0.03) LD values 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 the smallest populations. About 99.68% of all LD values are ≤0.25. The relatively small number of LD values greater than 0.25 (i.e., 99 values or <0.3%) occurred almost entirely between

unlinked markers (88 involve SNPs paired from different chromosomes and 2 are >50MB apart on the same chromosome) and predominantly involved different SNP pairs (89 of 99 SNP pairs).

6.6 Some general implications of this study

Two especially interesting aspects of our screening results are (1) the large variation among SNPs in Fst value when additional populations were tested (Figures 6-1 and 6-2) (2) yet the relatively high yield of markers having both low Fst 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 Fst and high heterozygosity. It is improbable that most would be maintained by balancing selection. In our screening study of 90,483 AB SNPs we found that 0.0442% or about 4.4 per 10,000 SNPs screened met our criteria for the combination of low Fst and high heterozygosity. Among our other research projects (enriched for SNPs and InDels varying around the world) 11 out of 887 markers screened (1.24%) could be identifed that met the same criteria for low Fst and high heterozygosity. Thus, it may be challenging to unequivocally demonstrate balancing selection in humans against a background of such SNPs.

6.7 Discrimination among individuals

Our panel of 40 SNPs resulted in unique genotypes for every one of the individuals with complete typings for all 40 SNPs. The distribution (Table 6-3) of the number of SNP genotypes matching for the more than 1.22 million pairwise comparisons of 1,568 individuals shows that no individuals match at all the markers. We obtained the nearly symmetric distribution around 15 (out of 40) matches expected by chance and no comparisons with more than 34 matches out of the 40. Thus, even with an occasional typing error generating an incorrect genotype and hence a false match or mismatch, the panel is robust. The expected number of real mismatches between unrelated samples is large enough to be certain of non-identity. A single mismatch between two 40-SNP profiles has a high probability of being an error and should be replicated. One would suspect biological relatedness or errors masking true identity if only a few mismatches occur. This also makes the marker set appropriate for tagging and tracking DNA samples in large biomedical, association, and epidemiological studies.

6.8 Toward a universal panel

This preliminary panel of 40 SNPs has excellent characteristics for individual identification, already yielding match probabilities that come close to the theoretical average match probability of just under 10^{-17} for 40 "perfect" SNPs, i.e., all with heterozygosity equal to 0.5. The yield of 40 acceptable SNPs from an initial set of 436 selected SNPs is encouraging. 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 40 SNPs may have much less uniformly high heterozygosities. Certainly, their rank is expected to change when any additional

populations are considered; some of the SNPs with Fst just larger than 0.06 may end up better than those with Fst just smaller than 0.06. Therefore, in order to obtain a universally applicable panel of SNPs it will be necessary to have an even larger panel of candidates from which to eventually select a final panel. That panel of candidates must also be sufficiently large that allowance is made for the inability of some markers to be included in multiplexed reactions. Other sources of potentially acceptable SNPs exist. Thousands of additional candidates for screening are available from the HapMap. Other researchers have identified SNPs with high heterozygosity in several diverse populations (e.g., Shriver et al., 2005; Sanchez et al., 2006) corresponding roughly to our seven-population screen. Our 40-population data from other projects can also yield suitable candidates. Thus, the forensic community should have no problem extending the panel of candidates to >>45 SNPs and even reducing the variation among populations provided many candidate markers can be tested on sufficiently large and diverse sets of populations. At the levels of heterozygosity we are achieving, a panel of 45 SNPs would give match probabilities less than 10^{-18} for most populations, easily in the range achieved with the CODIS markers. Were we to incorporate markers with $0.06 < Fst < 0.07$ into the preliminary panel, the variation in average match probability among populations we have studied would increase somewhat, but match probabilities would decrease for all populations.

Our panel should be considered in conjunction with markers in other panels to attempt to reach a consensus among the global research and forensic communities. Among SNP panels that have been proposed for use in individual identification (e.g., Inagaki et al., 2004; Lee et al., 2005; Sanchez et al., 2006), ours is the first to screen simultaneously for high heterozygosity and low Fst in a large global sample of populations. Others have tested only one or a few populations and/or have not imposed a specific criterion of low Fst to evaluate the uniformity of

the high heterozygosity. (Note, uniformly high heterozygosity means that the Fst will be low but a low Fst does not mean a high heterozygosity, just a relatively uniform heterozygosity.) When allele frequencies have been available for multiple populations, most previously published markers fail our criteria.

6.9 Independence in populations versus unlinked in families

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. While all 40 SNPs in our panel are statistically independent at the population level (the objective of our study), several of them are close enough molecularly to show linkage in families. If a universally applicable panel of SNPs is ever adopted by the international forensic community, it would be ideal for all markers in the panel to be both independent at the population level and unlinked.

The syntenic SNPs (those on the same chromosomes) among the best 40 in our study were examined to determine which would likely show genetic linkage among close biological relatives. The 25 syntenic SNP pairs are separated on average by 37.5 MB but cluster into two very distinct groups—6 pairs that are 75 to 172 MB apart and 19 pairs that are all <34 MB apart (median separation ~15MB). The 6 pairs >75 MB apart should be essentially unlinked. An estimate of the genetic map distance between each of the 19 SNP pairs that are <34 MB was obtained via the NCBI MapViewer (http://www.ncbi.nlm.nih.gov/mapview). The nucleotide positions for each pair were entered and the map distance was gauged by averaging the Genethon, deCode, and Marshfield estimates of map distance. A scatterplot (data not shown) of physical distance in MB by map distance in centi-Morgans (cM) for the 19 closest SNP pairs

displays a relationship not too different from the genome-wide expectation of roughly 1 cM per MB although most of the 19 points are above the 1 cM/MB line, such that the median ratio is 1.28 cM/MB and the range is 0.8 to 2.7 cM/MB. If we eliminate 15 SNPs because of linkage, retaining only the SNP with the best combination of low Fst and high heterozygosity from each set of linked SNPs, the 25 remaining SNPs are both unlinked and independent (Table 6-1, column 3). However, it is premature to discard any of these syntenic candidate SNPs for at least two reasons. The rank order of the 40 SNPs will likely change as additional populations are tested for these markers. Also, additional appropriate markers identified in the future may be unlinked to some of these syntenic loci but not others.

6.10 Some forensic considerations

The values in Figure 6-4 are calculated for ideal populations with no allowance for substructure. As noted by the NRC Committee (1996), the correction factor θ is equivalent to Fst for markers having Hardy-Weinberg 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 6-4 (filled-circles). We note that the relationships of measures of within population substructure to the global Fst 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 Fst we should also be reducing the likelihood of relevant substructure within each population. For these 40 loci the average "global" (40-population) Fst is 0.047. In an actual forensic application ignoring ethnicity one could use the global average

allele frequencies (appropriately weighted from population-specific data available for these 40 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.

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 for a potentially universally applicable panel many additional populations will need to 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 40 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). The 40 populations studied here cover most major regions of the world; the regions not covered are flanked by those that have been studied. However, as additional data accumulate on these markers and similar data become available for other markers, the rank order of markers for a universal panel may well change. Also, we would expect the Fst 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 ranges seen for the 40 populations that we have studied since we have deliberately included some isolated populations from various parts of the world as test of the robustness/generality of the results. Also important would be independent samples to show that the few large associations among markers 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.

We used TaqMan for the screening procedures because we were screening markers individually and did not have to develop or optimize the assays. While TaqMan low density arrays allow samples to be co-loaded, TaqMan is not capable of being multiplexed for the entire analysis through to the reading of the plate. It is not our intention to advocate any typing protocol nor, at this stage, to invest effort in developing multiplexing for these markers. Because dozens of SNPs can be routinely multiplexed, that is not an issue with modern "chip" methods such as those of Illumina or Affymetrix. Some typing methods might require a different multiplexing procedure and one would need to be developed. One important caveat is that any new typing method must be evaluated to demonstrate that there are not common nearby variants that would interfere with typing the target SNP (e.g., Osier et al., 2002). However, the SNPs we are identifying are in the public domain and any individual or corporation wishing to can work on developing methods for implementing this panel in a forensic or research setting. We do not advocate such effort for a forensic application of this panel. For a research application these SNPs are an efficient small panel but we do note that large numbers of "random" SNPs should also provide uniqueness irrespective of ethnicity. For a forensic application many more candidate SNPs need to be developed and all such need to be tested on more populations. In identifying those candidate SNPs we recommend researchers use screening criteria similar to those we have used because, though arbitrary, they have been demonstrated to yield SNPs with the desirable population genetic characteristics. When larger numbers of appropriate SNPs are available, the best set can be selected both in terms of their population genetics and the ability to develop an appropriate assay for forensic applications.

7. Expansion of the Set of Candidate IISNPs

7.1 Reducing the population panel

Our published population genetics criteria for SNPs for individual identification (IISNPs)--nearly maximum informativeness in populations from all parts of the world—seemed reasonably well accepted by the forensic community. However, our panel of 40 candidate SNPs meeting those criteria and giving 40-SNP genotype probabilities of $<10^{-16}$ in almost all populations 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. A much larger panel of 108 candidate SNPs meets our operationalized criteria of an Fst <0.06 and average heterozygosity >0.40. In addition to the previously published 40 SNPs we are now 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 is appropriate to keep all these markers among the candidates until they can be evaluated by laboratory and other criteria. We still advocate screening more SNPs to assure identifying a sufficient number meeting broad forensic criteria. We also believe that all of the near-final candidates should be evaluated on multiple, additional populations so that reasonably small (e.g.$<10^{-12}$) genotype frequencies can be demonstrated to occur even more broadly.

Our studies have led us to realize that different purposes require different panels of SNPs. We clarified our thinking in this regard in posters at the NIJ Forensics meeting in June, 2007 and in a poster at the ISFG meeting in Copenhagen in August, 2007. Our definitions of the four

types of panels was written up in Butler et al. (2007) and are given in Table 7.1. The remainder of Section 7 is taken from material on those two poster presentation.

Table 7.1. 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 di-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.

To date our studies have concentrated on the first two types of SNP panels with some preliminary investigation into the third. Most of our results are for IISNPs and we present here data on 108 SNPs that for a set of 31 populations (see Table 7-2) meet the criteria of high average informativeness (measured as heterozygosity) and low allele frequency variation among populations (measured as Fst) so that the panel is applicable anywhere in the world.

TABLE 7-2. Populations included in forensic studies						
Population samples at Kidd Lab	**Low Fst--High Het. 40 pop. samples**	**31 popula-tion samples**		**Population samples (continued)**	**Low Fst--High Het. 40 pop. samples**	**31 popu-lation 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				

7.2 Elaborating criteria for IISNPs in forensics

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 additional 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, these additional criteria include:

a. No medical or sensitive personal information is conveyed by the individual or combined data. Ideally the SNP is 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.

b. "Highly informative" is interpreted as high heterozygosity around the world and low allele frequency variation (measured as low Fst) so that the panel is informative irrespective of the ancestry of an individual. These criteria are important for use in modern multi-ethnic societies such as the USA. Choosing the "best" markers will be a function of the specific populations used to measure heterozygosity and Fst; thus, as more populations are studied for a set of markers, the rank order will change. Fortunately, the expectation is that once a large number of populations of diverse geographic origin is tested, the changes in rank order will be minimal.

c. Each of the SNPs should be statistically independent at the population level (no linkage disequilibrium with any other SNP in the panel) so that the product rule can be applied. This requires that some small, isolated populations be tested if markers are molecularly close since random genetic drift in such populations can generate LD over long molecular distances.

d. If the panel is also to be used in paternity testing, the markers should be unlinked as well. This requires knowing the correspondences between recombination distances and molecular distances for all markers within ~100 megabases of each other since there is considerable variation in the relationship around the genome. If markers meet the criterion of

being unlinked, they should also be statistically independent making it highly probable they will meet criterion *c* above.

e. Sufficient SNPs are needed to assure low probabilities of two randomly selected individuals having the same multi-site typing results. For SNPs with heterozygosities >0.4 and little allele frequency variation (low Fst), a panel of 40 to 45 SNPs gives probabilities $<10^{-15}$.

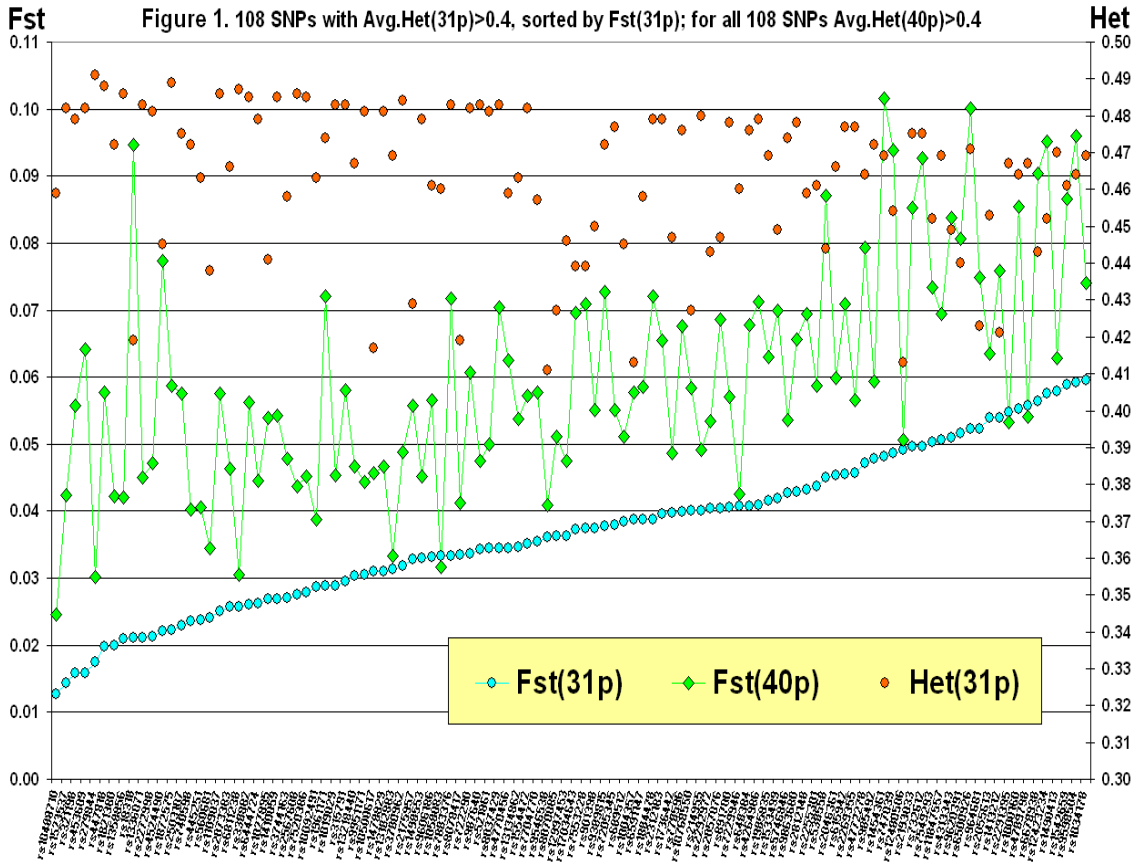
f. Documentation in the form of allele frequencies in a global set of populations must be in the public domain. The allele frequencies should be based on minimum samples of close to 50 individuals per population and/or close to 100 individuals from pooling closely related populations in a given region to allow moderate accuracy for each allele frequency estimate.

g. Laboratory criteria will need to be applied to any candidate SNPs for any of the types of panels. Depending on available equipment in forensics labs different typing techniques may be required and not all SNPs are amenable to all typing techniques. Some will require multiplexing in a way that may make some combinations unacceptable. One of the criticisms of our work has been that we have not developed a multiplex typing method, but that has never been our objective nor is it our expertise.

7.3 The expanded set of candidates

We have identified 108 candidate SNPs for an IISNP panel with Fst <0.06 and average heterozygosity ≥0.4. Their Fst values and heterozygosities based on 31 populations are given in Figure 7-1. These 31 are the larger populations more likely to be relevant in forensic settings, especially in the USA and Europe. Figure 7-1 shows the comparison of Fst values in the reduced set of 31 populations (blue circles) compared to the original set of 40 populations (green diamonds). The dbSNP rs numbers are given in the figure. This expanded set of 108 candidates

for an IISNP panel in Figure 7-1 meets criteria 1, 2, and 3 above and meets criteria *b*, *c*, *e*, and *f*. A large subset also meets criterion *d*. Criterion *a* is a particularly ambiguous one if one concentrates on "genes", as explained in the discussion following. Some sets of SNPs are genetically linked and we have not tested all pairwise combinations for absence of LD in all populations since other considerations will need to be considered in selecting which SNP to keep among the molecularly and genetically close SNPs.



7.4 When is multiplexing an issue?

One such consideration will be whether or not multiplexing is an issue. It is our assumption that the primary value of SNPs is the ability to quickly type a sample for large numbers of SNPs on a chip. With current techniques it is routine to be able to "multiplex" arbitrary sets of dozens to thousands of SNPs with no problems. With very small amounts of

DNA it should be possible to type several dozen arbitrarily selected SNPs simultaneously without multiplexing problems. However, if PCR product size is the assay, rather than assaying the "interior" of an amplicon, multiplexing does become an issue. Other considerations are uniqueness of the SNP and ease of typing using small amplicons. Since all of these 108 SNPs have been typed with TaqMan and have given high quality typing results, these criteria have been met for all.

7.5 How does one deal with SNPs in "genes"?

We believe the most controversial issue will be whether or not intronic SNPs must be excluded. Many of these 108 SNPs are in introns; some that are in intergenic regions (by current knowledge) show high sequence conservation in mammals. While we argue that intronic SNPs are acceptable as a rule, we will also argue that SNPs in highly conserved regions, intergenic or intronic, should be excluded. We are in the process of examining all 108 SNPs for these characteristics and will make the data available when complete. Some examples are presented in Table 7-3.

TABLE 7-3. Examples of genomic characteristics/locations of candidate IISNPs

Rank Fst 31p	dbSNP rs#	Het (31p)	Fst (31p)	Nucleotide position	Chr	Vertebrate Conserved (Y/N)	Known Gene (Y/N)	In Exon (Y/N/nr)	In Intron (Y/N/nr)	Distance Nearest Gene/Exon	Gene SYMBOL	Notes
10	rs1336071	0.472	0.0451	94,593,976	6	Y	N?	N	?	~5.5kb	spliced est	
16	rs445251	0.463	0.0237	15,072,933	20	N	?	N	Y	~50kb	C20orf133	1
20	rs6811238	0.487	0.0257	169,900,190	4	N	Y	N	Y	~30kb	PALLD	
26	rs2567608	0.486	0.0275	22,965,082	20	"N"	Y	Y	N	nr	SSTR4	2
27	rs7520386	0.485	0.0278	14,027,989	1	N	Y	N	N	~4kb	PRDM2	3
60	rs689512	0.445	0.0384	78,308,991	17	N	Y	N	Y	~1.6kb	TBCD	
71	rs891700	0.478	0.0405	237,948,549	1	N	Y	N	Y	~150bp	CHRM3	
75	rs1985835	0.469	0.0415	60,925,204	20	N	Y	N	Y	~800bp	COL9A3	
85	rs4772278	0.464	0.0472	99,732,276	13	N	Y	N	Y	~9kb	PCCA	
87	rs1454361	0.469	0.0481	24,920,672	14	N	N	N	N	>200kb	?	

Notes: (1) hypothetical protein LOC140733; (2) in non-conserved part of exon; (3) downstream of 3' UTR

What is the relevance of a gene to marker selection? What do the phrases "no medical or personal information" and "not in a gene" in criterion "a" (section 7.2) really mean as criteria for forensic SNPs? One can understand public apprehension over having medical information conveyed by the SNP alleles in a forensic database. That can easily be generalized to other sensitive, "personal" information. Indeed, ethical concerns over identifying high likelihood of an individual developing a cancer, Alzheimer disease, or Huntington disease does preclude using SNPs that would convey such information. However, from a scientific perspective that does not generalize to precluding all SNPs from even those genes, much less any gene, if the SNPs meet the population genetics criteria we have used for a panel for individual identification. The scientific logic is outlined in the following.

One of the criteria for a "universal" panel of IISNPs is that heterozygosity is high around the world. Thus, both alleles at the SNP are by definition normal, with nearly equal allele frequencies in all populations and cannot be deterministic for a Mendelian genetic disease. Similarly, the SNP cannot have a significant impact on risk for a common, complex disorder. This logic applies even if the SNP is in the coding sequence of a gene known to be involved in a Mendelian or complex genetic disorder, but there are very rare exceptions. Obviously there is no point in arguing for including SNPs in coding regions.

The more general question of linkage disequilibrium with a variant involved in a Mendelian or complex disorder is important. Since the Mendelian disorders are rare, the alleles of a SNP with high heterozygosity will not convey significant information about the mutations for a Mendelian disorder even if there is complete linkage disequilibrium. In the case of the disease-causing allele in complete LD with one of the SNP alleles, while the SNP genotype does alter the numeric probability of the mutation being present, it is not a very meaningful alteration

even in this extreme case of a relatively common disease-causing mutation. Extrapolated to complex disorders with no deterministic alleles and low risk conveyed by variants at any one locus, this logic indicates that genotypes for SNPs with globally high heterozygosity, e.g. ≥0.4, do not convey significant medical or other sensitive personal information.

While one can accept excluding SNPs in coding regions of a gene as a conservative measure, is there any reason to exclude SNPs from introns? Certainly, the Tyrosine Hydroxylase STR (TH01) currently used in CODIS is in an intron, intron 1. Even more significantly, the Von Willibrand Factor (vWF) STR in CODIS is in an intron (intron 40) of a gene with disease causing alleles. We would argue that there is no general scientific reason for excluding SNPs from introns of such genes if they meet our population genetics criteria of high heterozygosity and low Fst. There are two aspects to the argument. First, as noted above, the SNPs are clearly normal genetic variation and highly heterozygous around the world. Therefore, they cannot be medically important in themselves. Second, to argue that such SNPs might be in LD with functional variation does not hold up as a significant argument as also noted above and the LD argument has serious implications for any SNP. Those implications are twofold. First, scientists are increasingly identifying new genes in previously "empty" regions of the genome and identifying new functional elements that are not traditional protein-coding genes. Thus, any region in the genome might turn out to be of major functional importance at some time in the future. Second, an argument of LD cannot be universally applied since LD varies around the genome and among populations. Moreover, individual SNPs can show remote LD but not close LD. Thus, an argument that no SNP can be in a gene or in LD with a functional element will be impossible to prove for all populations and runs the serious risk of requiring revision of SNP panels as new information is learned about the genome.

7.6 Recently Completed Analyses

Following our awareness of the SNP for ID panel of 52 SNPs (Sanchez et al., 2003) proposed for individual identification, we applied our criteria to those data and tested their best markers. Most failed to meet our criteria on our 40-population panel and we did not pursue others of theirs. In order to determine whether any of their markers would meet our criteria on the 31-population panel (Table 7-2), we recently completed testing 47 of their 52 SNPs (TaqMan assays for the remaining 5 were not available). The actual allele frequencies for these are being entered into ALFRED. As shown by the highlighting in Table 7-4, nine of the markers meet our criteria for 31 populations but only three meet the criteria for 40 populations. As we test additional populations now available in our lab we will include the best 9 to 12 of these to evaluate their performance. It is interesting to note that for the 40-population analyses a dozen of these markers show inter-population variation above the average (Fst ~0.14) for unselected SNPs. We will be discussing a joint paper with the SNP for ID group in the coming months.

TABLE 7-4: Fst and Heterozygosity for 47 SNPforID markers

Markers are divided into subtables by AvgHet(31 pops); yellow highlights Avg.Het.>0.40;

then subtables are sorted by Fst(31pops); green highlights Fst<0.06 and AvgHet>0.4.

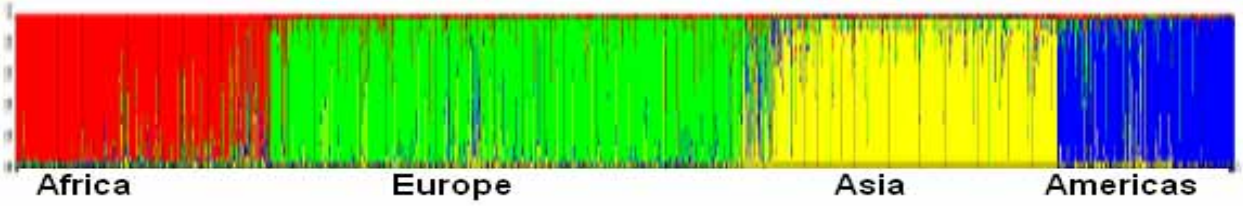
Chr	AB catalog #	dbSNP #	Avg.Het.	Fst	Avg.Het.	Fst
			31 pops	31 pops	40 pops	40 pops
15	C_29375514_10	rs8037429	0.483	0.0344	0.463	0.0705
11	C__7698393_10	rs901398	0.450	0.0375	0.440	0.0551
1	C__7539584_10	rs891700	0.478	0.0405	0.470	0.0571
4	C_11989432_10	rs2046361	0.466	0.0453	0.459	0.0598
14	C__2120263_10	rs1454361	0.469	0.0481	0.444	0.1016
10	C__8902740_10	rs964681	0.423	0.0523	0.411	0.0748
1	C__1732269_10	rs1413212	0.421	0.0540	0.425	0.0759
1	C__9630073_10	rs1490413	0.470	0.0579	0.467	0.0628
12	C__2881087_30	rs2111980	0.463	0.0598	0.451	0.0807
20	C__8953333_10	rs1031825	0.442	0.0709	0.436	0.0826
9	C__3175786_10	rs1463729	0.452	0.0735	0.440	0.1007
21	C__2528441_20	rs2831700	0.450	0.0736	0.442	0.1027
18	C__7485867_20	rs1493232	0.454	0.0830	0.459	0.0763
4	C_12098080_10	rs1979255	0.450	0.0836	0.448	0.0972
8	C__1083125_10	rs763869	0.436	0.0848	0.425	0.0921
5	C___574306_10	rs717302	0.404	0.1222	0.384	0.1543
6	C__2695128_10	rs727811	0.429	0.1291	0.424	0.1471
20	C__2203431_10	rs1005533	0.426	0.1313	0.419	0.1310
6	C__2513175_10	rs1029047	0.415	0.1335	0.395	0.1559
21	C__2688083_10	rs914165	0.432	0.1336	0.431	0.1371
13	C__3084646_10	rs354439	0.429	0.1375	0.405	0.1842
2	C__1553762_20	rs907100	0.424	0.1440	0.416	0.1650
10	C__7431207_20	rs735155	0.427	0.1451	0.413	0.1737
7	C__7608025_10	rs917118	0.417	0.1503	0.401	0.1784
22	C_11482429_10	rs2040411	0.412	0.1503	0.405	0.1691
17	C__7475537_10	rs938283	0.261	0.0421	0.243	0.0692
8	C___408450_10	rs2056277	0.243	0.0532	0.218	0.0693
3	C_11354314_10	rs1357617	0.320	0.0610	0.283	0.0772
9	C__1410631_20	rs1360288	0.381	0.0622	0.367	0.0829
14	C__1146837_10	rs873196	0.370	0.0627	0.353	0.0726
22	C_____27044_1	rs733164	0.390	0.0633	0.399	0.0852
19	C____10567_20	rs719366	0.380	0.0635	0.388	0.0897
12	C__2626420_10	rs2107612	0.343	0.0729	0.329	0.0960
5	C__3199379_20	rs251934	0.357	0.0883	0.336	0.0939
2	C__1611304_10	rs876724	0.376	0.0932	0.372	0.1143
16	C__1168681_20	rs729172	0.387	0.0954	0.361	0.1098
11	C_26325730_10	rs2076848	0.390	0.0982	0.371	0.1153
1	C_30511383_20	rs10495407	0.362	0.1030	0.352	0.1372
7	C__2604172_10	rs737681	0.388	0.1171	0.349	0.1504
16	C__1877107_10	rs1382387	0.398	0.1344	0.392	0.1411
13	C__1922667_10	rs1886510	0.358	0.1486	0.336	0.1701
9	C__1881082_10	rs1015250	0.392	0.1782	0.383	0.2003
3	C___233252_10	rs1355366	0.378	0.1809	0.348	0.2126

17	C__2653097_10	rs740910	0.256	0.2189	0.261	0.2579
21	C__2349786_10	rs722098	0.374	0.2511	0.388	0.2217
15	C___314944_10	rs1528460	0.370	0.2592	0.375	0.2503
13	C__7468761_10	rs1335873	0.350	0.2891	0.358	0.2702

8. Progress on Identifying Ancestry Informative SNPs (AISNPs)

We have made a strong start on developing a panel of high Fst 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. 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 our 40 populations, Figure 8-1) 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].

Figure 8-1: STRUCTURE solution at K=4 clusters for 40 populations with Lao et al. (2006) 10-SNP set. (Populations are in the same order as in Figure 6-4)

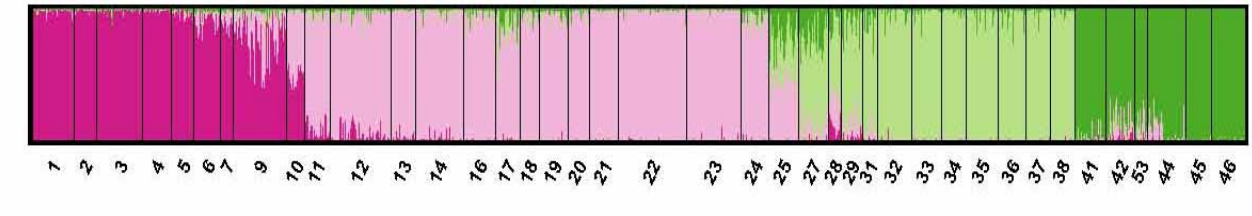


We have sought appropriate markers for robustly resolving geographic and population structure with multiple screening procedures: (1) high Fst markers identified in the Celera or HapMap databases, (2) the ten markers published by Lao et al. (2006), (3) the markers identified for the Kim et al. (2005) 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. The first two screening approaches are aimed at providing good assignment to continent (except the Americas). The first three approaches yielded 109 markers as an initial exploratory

dataset. Using these resources one cannot know from the limited data available how informative any marker will be. Indeed, not all of these SNPs have high Fst values when typed on the 40 populations, though all but 18 are above the mean of the random distribution. Thus, we have continued to collect data on high Fst SNPs.

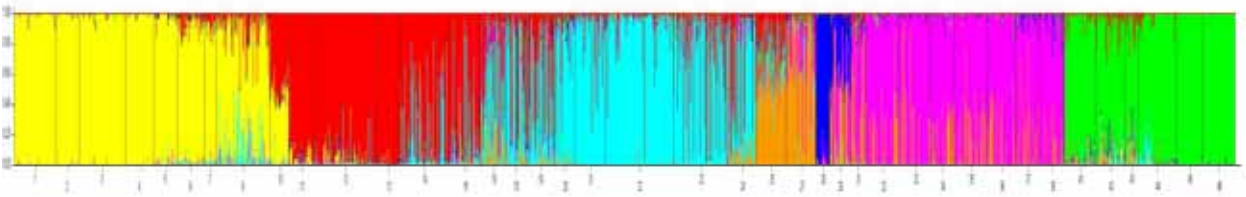
Our developing AISNP panel currently consists of 249 candidate SNPs. When four continental clusters are considered, the populations in Africa, Europe, east Asia, and the Americas our 249-SNP panel (Figure 8-2) gives greater certainty of assignment of individuals using various statistics (not shown) and visually reflected in the greater homogeneity of the colored bands relative to the Lao et al. (2006) panel (Figure 8-1). With such a large number of SNPs, we can extend our analyses to populations located between continents (Figure 8-3). However, we realize that 249 SNPs is not a reasonable size for an investigative AISNP panel, and we plan to continue exploring methods of decreasing the number of SNPs while retaining informativeness.

Figure 8-2. STRUCTURE solution at K=4 clusters for 40 populations with the 249 SNP data set. (Populations are in the same order as in Figure 6-4)



Africa **Europe** **Asia** **Americas**

Figure 8-3. STRUCTURE solution at K=7 clusters for 40 populations with the 249 SNP data set. (Populations are in the same order as in Figure 6-4)



Africa SWAsia Europe Siberia Asia Americas Oceania

Our ongoing studies suggest that a simple statistic can quantify the difference in "clarity" between Figures 8-1 and 8-2: the average frequency with which individuals are assigned to the populations that logic dictates should belong to the same cluster. Thus, populations 16-23 are all located in Europe proper (plus European Americans) and, for forensic purposes, should cluster unambiguously as "European". Clearly both visually and, we find, statistically that is not the case for the Lao 10-SNP set but is for the 249-SNP set. We are pursuing such approaches to provide hard statistical support for what to date has been largely visual.

As the first step in optimizing resolution among populations within regions we have analyzed ~1000 markers we have typed on these 40 populations and identified the 50 SNPs with the highest Fst within each of 7 regions (Africa, south-west Asia, Europe, Siberia, east Asia, Pacific, and the Americas). None was selected for global Fst and indeed the global resolution was not good. With some overlap of Fst between regions, the total number of SNPs identified in this way was 256. Surprisingly (at least to us), the regional resolution was not good either when all 256 markers were used, possibly because the random variation contributed by the markers that have high Fst in different regions obscured any structure indicated by the 50 SNPs that had high Fst within the region.

We have identified additional SNPs from publicly available data sets (e.g., Conrad et al., 2006; Shriver et al 2005; and new data on the HGDP panel) that show indication of being able to distinguish between populations. We plan to type a large number of these SNPs on our newly extended set of populations (now 44 in number, including 1 population sample from a geographically "intermediate" location).

9. Conclusion

9.1 IISNPs

The 40-SNP IISNP panel we developed meets our objective in the original application for such a panel of SNPs. However, we have learned as we conducted this research and do not advocate its adoption but do advocate its testing on additional populations and the testing of additional unlinked markers to make the panel valid for relationship inference without having to incorporate genetic linkage values into calculations.

Though we have analyzed SNPs on a reduced panel of populations, we do not ourselves advocate use of SNPs meeting only those 31-population criteria. Rather, we advocate even more strict criteria than our original 40-population values of heterozygosity >0.4 and Fst <0.06. Since we have demonstrated it is possible to find such markers, we see no reason not to attempt an extremely robust set of IISNPs. That extends to inclusion of more diverse populations through involvement of more laboratories testing the best of the proposed IISNPs.

9.2 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 good at distinguishing among a different set of populations, even if the same geographic regions are involved. Thus, we are initially focusing on a panel for robust assignment to four "continental" groups. Our progress in that area shows that a small number of AISNPs (~two dozen) can do very well for assigning individuals from the geographic regions of focus, but does not do well for individuals from intermediate geographic regions. Separate sets of AISNPs can be found for distinguishing among populations within a geographic region but a different set is needed for each 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 good set of AISNPs by statistical criteria that we are developing. Extensive analyses have begun but no answers are yet clear.

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.

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, Report on ISFG SNP Panel Discussion. Progress in Forensics Genetics 12 (2007) (in press)

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.

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.

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.

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.

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.

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.

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.

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, 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.

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.

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.

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.

Osier, M.V., A. J. Pakstis, D. Goldman, H. J. Edenberg, J. R. Kidd, and K. K. Kidd. A Proline-Theronine Substitution in Codon 351 of ADH1C is Common in Native Americans. Alcoholism: Clinical and Experimental Research Res 26 (2002) 1759-1763.

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

Pritchard, J.K., M Stephens, P Donnelly. Inference of population structure using multilocus genotype data. Genetics 155 (2000) 945-59.

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) 135-40.

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.

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.

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.