Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International 164 (2006) 20–32
Forensic Science International
www.elsevier.com/locate/forsciint

Developing a SNP panel for forensic identification of individuals

Kenneth K. Kidd [a,*], Andrew J. Pakstis [a], William C. Speed [a], Elena L. Grigorenko [b], Sylvester L.B. Kajuna [c], Nganyirwa J. Karoma [c], Selemani Kungulilo [d], Jong-Jin Kim [e], Ru-Band Lu [f], Adekunle Odunsi [g], Friday Okonofua [h], Josef Parnas [i], Leslie O. Schulz [j], Olga V. Zhukova [k], Judith R. Kidd [a]

[a] *Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA*
[b] *Child Study Center and Department of Psychology, Yale University, New Haven, CT, USA*
[c] *The Hubert Kairuki Memorial University, Dar-es-Salaam, Tanzania*
[d] *Muhimbili University College of Health Sciences, Dar-es-Salaam, Tanzania*
[e] *DNA Analysis Division, National Institute of Scientific Investigation, Seoul, Korea*
[f] *Department of Psychiatry, College of Medicine and Hospital, National Cheng-Kung University, Tainan, Taiwan, ROC*
[g] *Department of Gynecological Oncology, Roswell Park Cancer Institute, Buffalo, NY, USA*
[h] *Department of Obstetrics and Gynecology, Faculty of Medicine, University of Benin, Benin City, Nigeria*
[i] *The Danish National Research Foundation, Center for Subjectivity Research, University of Copenhagen, Købmagergade 46, DK-1150 Copenhagen N, Denmark*
[j] *College of Health Sciences, University of Texas at El Paso, El Paso, TX, USA*
[k] *N.I. Vavilov Institute of General Genetics RAS, Moscow, Russia*

Received 2 September 2005; received in revised form 3 November 2005; accepted 8 November 2005
Available online 19 December 2005

Abstract

Single nucleotide polymorphisms (SNPs) are likely in the near future to have a fundamental role in forensics in both human identification and description. However, 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. We describe here both an efficient strategy for identifying and characterizing such SNPs, and 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 are 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 are then screened on a total of 40 populations (~2100 individuals) and the most promising retained. The preliminary panel of 19 SNPs, from an initial selection of 195 SNPs, gives an average match probability of $<10^{-7}$ in most of 40 populations studied and no greater than 10^{-6} in the most isolated, inbred populations. Expansion of this panel to ~50 comparable SNPs should give match probabilities of about 10^{-15} with a small global range.

Keywords: Human identification; SNPs; Population genetics; Fst; Heterozygosity

* Corresponding author at: Department of Genetics, Yale University School of Medicine, 333 Cedar Street, P.O. Box 208005, New Haven, CT 06520, USA. Tel.: +1 203 785 2654; fax: +1 203 785 6568.
E-mail address: Kenneth.Kidd@yale.edu (K.K. Kidd).

0379-0738/$ – see front matter
doi:10.1016/j.forsciint.2005.11.017

1. Introduction

Single nucleotide polymorphisms (SNPs) are being considered for a potentially useful role in forensic human identification [1–3]. Among their advantages are: (1) SNPs have essentially zero rate of recurrent mutation. With mutation rates for SNPs estimated at 10^{-8} [4] compared with rates of 10^{-3} to 10^{-5} for STRPs [5,6], 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 [7–9] have demonstrated the value of reducing amplicon size from the 100 to 450 bp range of the COmbined DNA Index System (CODIS) 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 (merely 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 multiplexed and done very quickly.

There are two commonly recognized problems with SNPs replacing STRPs 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} [10]. 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. Fig. 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 [11] might have some validity.

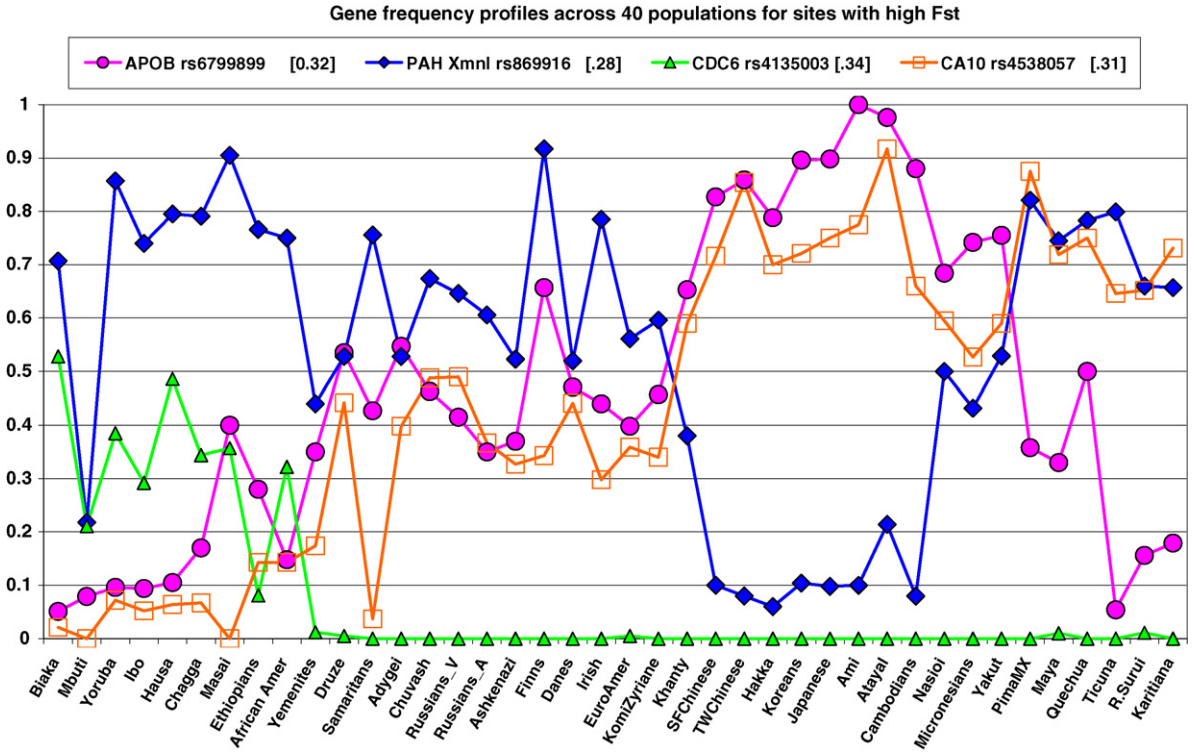

Fig. 1. The frequencies of one allele at each of four SNPs with high variation in allele frequencies among populations. The SNPs are identified by their rs number in dbSNP and the symbol of the genetic locus in which each occurs; the data are in ALFRED. The populations are arranged by geographic region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

A SNP 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. High heterozygosity maximizes the information at each SNP and low Fst minimizes the chance effects between populations. Thus, the combination of high heterozygosity and low Fst increases the efficiency of a forensic panel—that is, it will take fewer SNPs to produce lower match probabilities than if random SNPs are used. Fortunately, not all SNPs are as varied in allele frequency among

Table 1
The 40 population samples

Geographic region	Name	*N*	Population ALFRED UID	Sample ALFRED UID
Africa	Biaka[a,b]	70	PO000005F	SA000005F
	Mbuti[a]	39	PO000006G	SA000006G
	Yoruba[a]	78	PO000036J	SA000036J
	Ibo[b]	48	PO000096P	SA000096S
	Hausa[b]	39	PO000097Q	SA000100B
	Chagga	45	PO000324J	SA000487T
	Masai	22	PO000456P	SA000854R
	Ethiopian Jews[d]	32	PO000015G	SA000015G
	African Americans	90	PO000098R	SA000101C
S.W. Asia	Yemenite Jews	43	PO000085N	SA000016H
	Druze[a,d]	127[c]	PO000008I	SA0000846S
	Samaritans	41	PO000095O	SA000098R
Europe	Adygei[a]	54	PO000017I	SA000017I
	Chuvash	40	PO000032M	SA000491O
	Russians, Vologda[a]	48	PO000019K	SA000019K
	Russians, Archangelsk	34	PO000019K	SA001530J
	Ashkenazi Jews[d]	83	PO000038L	SA000490N
	Finns	36	PO000018J	SA000018J
	Danes	51	PO000007H	SA000007H
	Irish	118	PO00057M	SA000057M
	EuroAmericans[b]	92	PO000020C	SA000020C
N.W. Asia	Komi Zyriane	40	PO000326L	SA000489V
	Khanty	50	PO000325K	SA000488U
East Asia	SF Chinese[a]	60	PO000009J	SA000009J
	TW Chinese[b]	49	PO000009J	SA000001B
	Hakka	41	PO000003D	SA000003I
	Koreans	66	PO000030D	SA000936S
	Japanese[a]	51	PO000010B	SA000010B
	Ami	40	PO000002C	SA000002C
	Atayal	40	PO000021D	SA000021D
	Cambodians[a,b]	25	PO000022E	SA000022E
N.E. Asia	Yakut[a]	51	PO000011C	SA000011C
Pacific Islands	Nasioi[a]	23	PO000012D	SA000012D
	Micronesians	37	PO000063J	SA000063J
N. America	Pima, Mexico[a]	99[c]	PO000034H	SA000026I
	Maya[a,b]	52	PO000013E	SA000013E
S. America	Quechua	22	PO000069P	SA000069P
	Ticuna	65	PO000027J	SA000027J
	Rondonian Surui[a]	47	PO000014F	SA000014F
	Karitiana[a]	57	PO000028K	SA000028K

The ALFRED UIDs can be used to retrieve the descriptions of the populations and of the specific samples of those populations. EuroAmericans are unrelated individuals married into large, multigenerational pedigrees that were collected for studies of genetic linkage and human variation.

[a] Samples (usually a subset) contributed to the HGDP-CEPH panel, Paris.

[b] Indicates the seven population samples included in the initial screening of polymorphisms.

[c] Samples with many related individuals; most analyses include only unrelated individuals.

[d] Source: National Laboratory for the Genetics of Israeli Populations.

populations as those in Fig. 1. Some have remarkably little variation in allele frequency around the world. The problem is how to identify appropriate SNPs and demonstrate their low allele frequency variation sufficiently well for forensic purposes. We have developed an efficient screening procedure for finding such SNPs and report here the strategy and the initial results.

2. Methods

2.1. Strategy

Our strategy consists of five steps. 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 [12] 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. Our objective is to identify appropriate SNPs; subsequently others could determine the appropriate typing methods for forensic applications of the set of markers identified. 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 test markers with average heterozygosity >0.45 and Fst <0.01. Once a marker is selected for testing, no other markers are selected within 1 Mb of that marker.

For the initial screen we have 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 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 an additional 33 populations (Table 1). Thus, markers making it through the second screen will have been typed on ~2100 individuals from 40 populations. By geographic region the number 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).

2.2. Screening criteria

To determine reasonable screening values we analyzed data we have collected on other projects (in ALFRED and unpublished). About 900 SNPs, more or less randomly selected with respect to Fst, have been typed on 38–42 populations including all or most of the 40 populations being used in this study ([13] and unpublished data). Two hundred and seventy-seven of these SNPs have average heterozygosities ≥ 0.4 for the seven 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 (Fig. 2). There is a significant, but far from perfect correlation. We chose an initial cut-off value of 0.02 for the seven 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.

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.

2.3. 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 3 μl 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.

2.4. 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 χ^2-test comparing the expected and observed number of individuals occurring for each

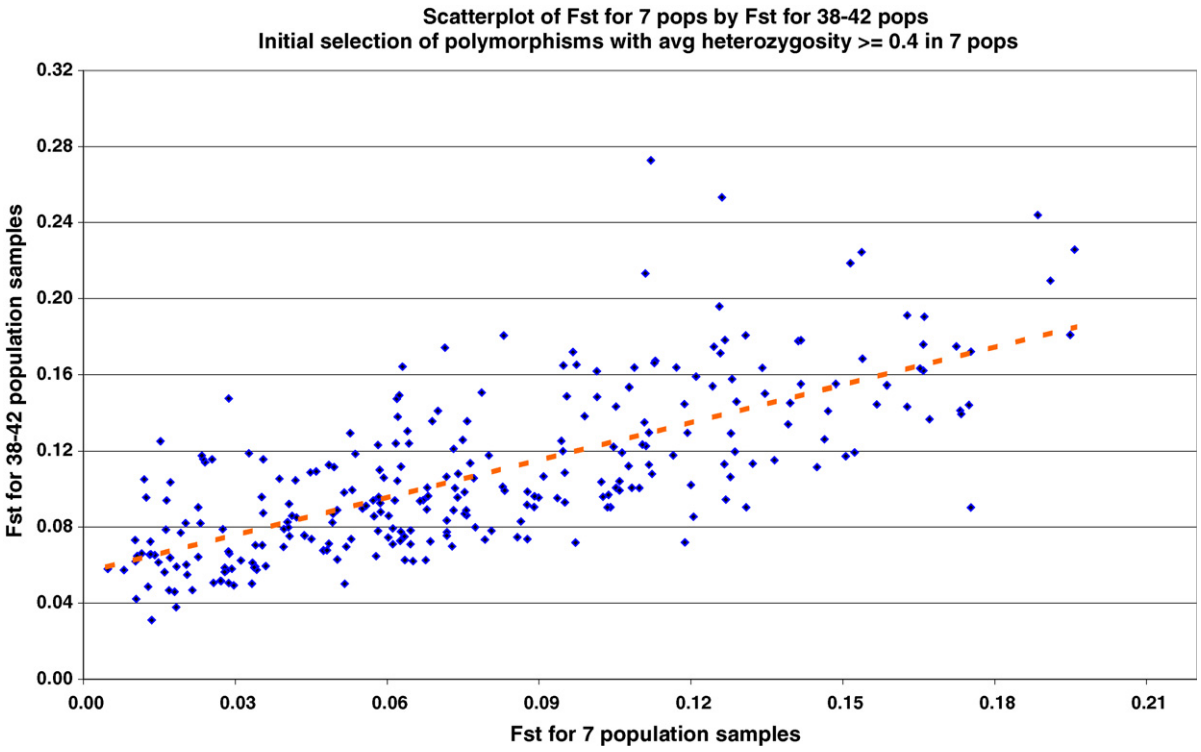

Fig. 2. Scatterplot of Fst values for 277 SNPs (selected for high heterozygosity on seven populations) calculated on seven populations and for 38–42 populations that include the seven. The Pearson correlation coefficient is 0.72.

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 [14] 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 19 locus-specific values were multiplied together within each population to give the most common genotype frequency.

3. Results

3.1. The yield from screening

The AB TaqMan Assays-on-Demand catalog lists 90,483 SNPs that have allele frequencies for four populations (European American, African American, Chinese, and Japanese). Because Japanese and Chinese allele frequencies are more similar to each other than either is to the other two (European American and African American) their allele frequencies and heterozygosity were averaged and this average was used along with African Americans and European Americans in screening the Applied Biosystems database. Of these, 14,638 have an average heterozygosity $\geq$0.45 across all three populations. The Fst for 2723 of those SNPs is $<$0.01 for the three populations. We started screening from these 2723 markers, initially selecting those that had the highest heterozygosity and lowest Fst but not

choosing any that were within 1.0 megabases of any other. Markers from the Applied Biosystems catalog located on the X or Y chromosome were excluded. To date we have screened 195 markers on the seven populations listed earlier. The original Fst calculated for this set of 195 markers ranges from 5.6×10^{-8} to 6.2×10^{-4}, using the European American, African American, and an average of the Chinese and Japanese frequency data provided by Applied Biosystems. One of the 195 markers showed deviations from Hardy–Weinberg ratios in some populations indicative of a silent allele. One marker did not yield clear genotype clusters. These markers were discarded. All of the rest ''behaved well'' and had no significant deviations from Hardy–Weinberg ratios. The Fst distribution of these 193 SNPs using data from seven populations is given in Fig. 3. 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. 2) 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 have been typed on all 40 populations. Fig. 4 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. 2, 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. 4). These SNPs are listed in Table 2.

3.2. Independence of 19 best SNPs

As shown in Table 2, 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 $171 \times 40 = 6840$ LD values clearly supports the conclusion that each SNP contributes essentially independent variation for each of the 40 population samples tested. 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 ≤ 0.05 and 94.9% are ≤ 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 [15]. 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

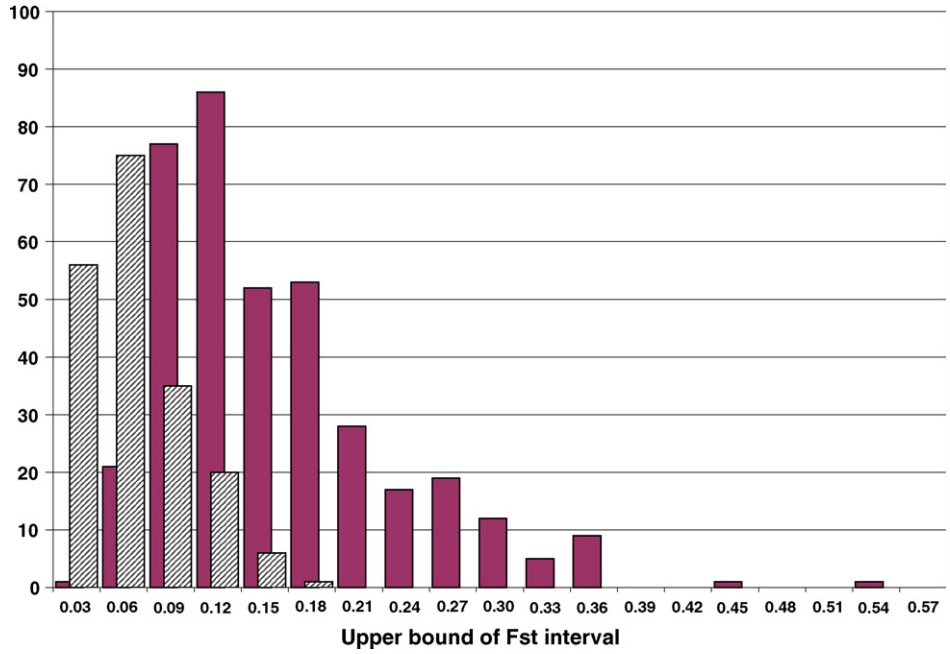

Fig. 3. Comparison of Fst distributions. The solid bars represent the Fst for reference markers (not pre-selected for Fst) calculated for 38–42 populations. The cross-hatched bars represent the Fst for the 193 markers calculated for seven populations.

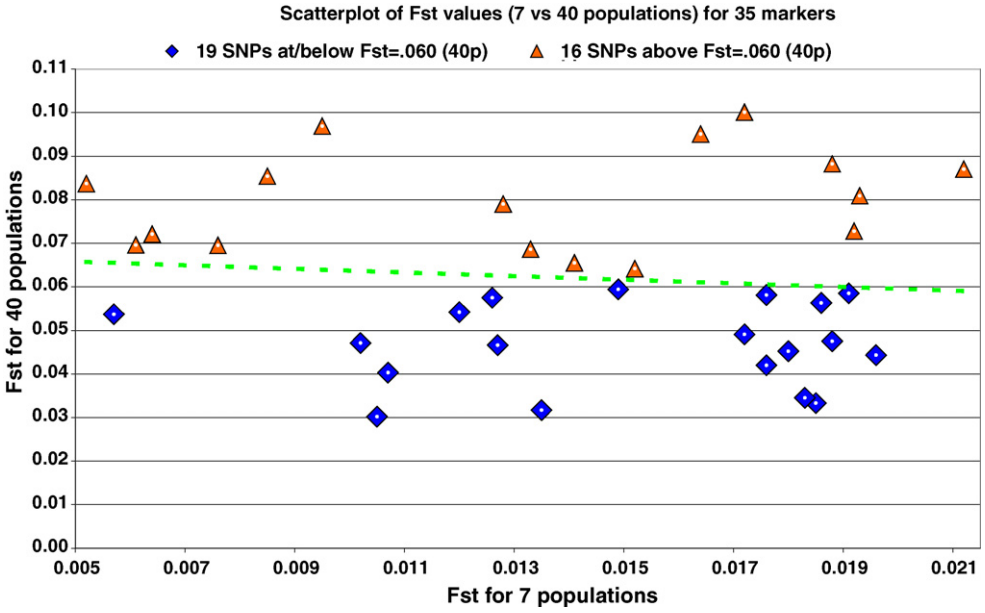

Fig. 4. Scatterplot of the 35 markers tested on all 40 populations by the Fst values for the seven populations in the initial screen and for all 40 populations. The 19 SNPs included in the panel are plotted as diamonds; the 16 SNPs with final Fst above 0.06 are plotted as triangles. The regression is 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.

weak levels of association. There are only 34 LD values in this range, the most extreme 1/2 of 1% of the 6840 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 <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 and 21.30 Mbp apart in the Surui and Karitiana, respectively. These three are included in Table 3, which summarizes the LD results for all pairs of markers on the same chromosome. All of the pairs in Table 3 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 3 are statistically independent.

3.3. Statistics for the preliminary 19-SNP panel

The frequency of the most common 19-locus genotype in each population is given in Fig. 5. 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.

Fig. 6 presents the average match probability by population. This value is the weighted average of the match probabilities of the 3^{19} possible genotypes, assuming exact H–W ratios within each population. The values range

Table 2
The 19 best polymorphisms sorted by Fst value based on 40 population samples

Chromosome	Cytogenetic band position	Locus symbol	ABI catalog #	dbSNP rs#	Nt. position UCSC May 2004	ALFRED site UID	Fst 40 population	Fst 7 population	Avg. Het. 40 population	Avg. Het. 7 population
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

For each SNP the table gives the position, locus name, various identifiers in different databases, and various statistics. *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 at Santa Clara genome center (counting from pter as origin).

Table 3
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

Chromosome	SNP pair		Separation (M bp)	*N* populations	Median	Average	Minimum	Maximum	Maximum LD population
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

The marker pairs are identified by the names of the loci containing the SNPs as given in Table 2. Physical distance (in Megabases) and the population in which the maximum Δ^2 occurred are given.

across approximately one order of magnitude, from $>10^{-7}$ to $>10^{-8}$. 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 9populations, the probability of discrimination is >0.9999999.

4. Discussion

The narrow range in the distribution of the average match probability across populations validates the low Fst strategy for identifying SNPs for use in forensic human identification. While Fst depends on the specific set of populations

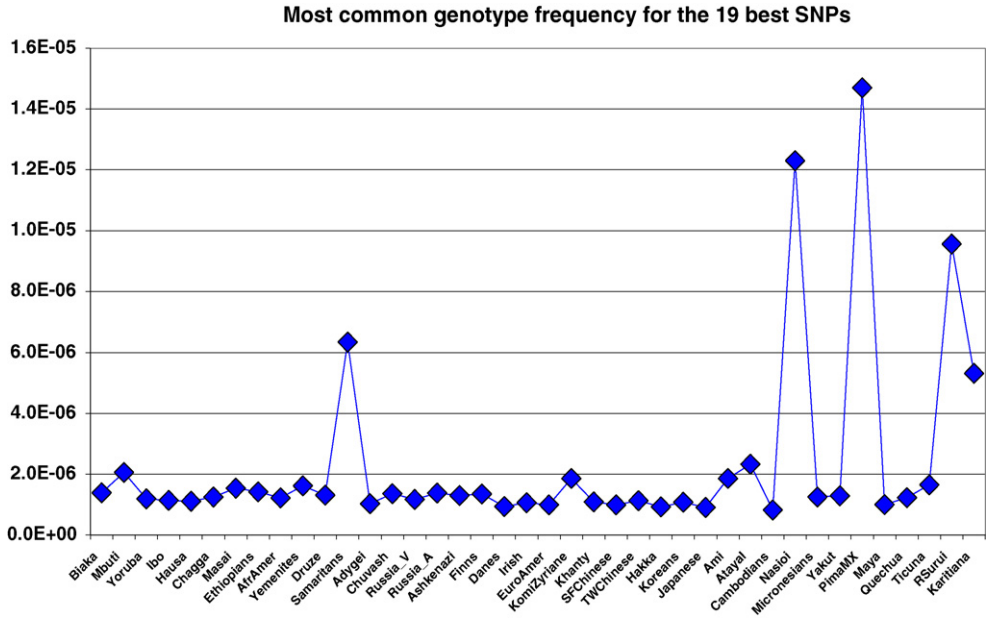

Fig. 5. The frequency of the most frequent genotype for 19 SNPs in each population. Populations are ordered by geographic region as in Fig. 1 and Table 1.

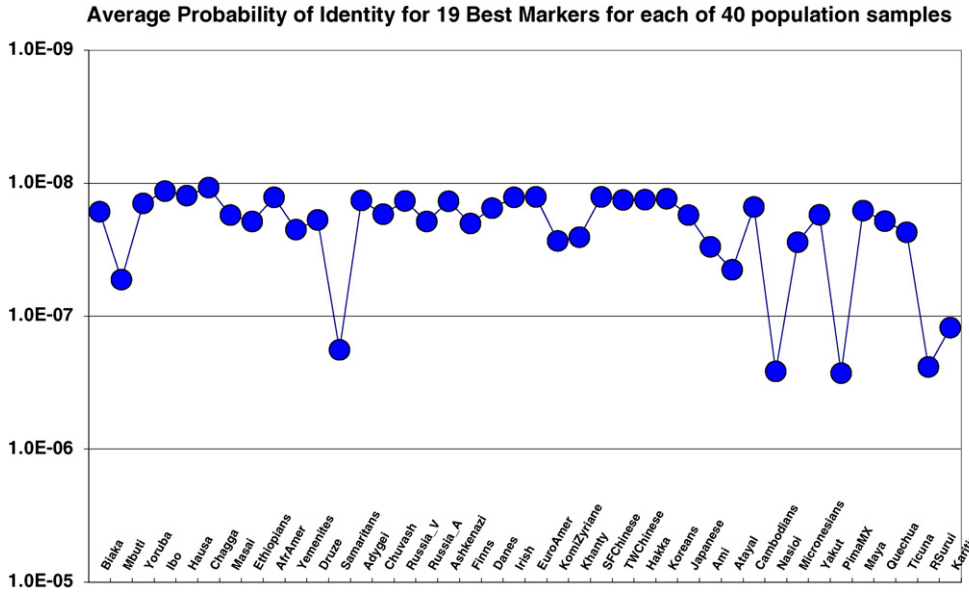

Fig. 6. The average match probability for the best 19 markers for each of 40 population samples. Populations are ordered by geographic region as in Fig. 1 and Table 1.

studied, it is clear that a global set of DNA samples can be used to screen for markers with globally uniform Fst values. Selecting markers based on low Fst has an additional benefit of minimizing any differential effect balancing selection in a particular population or geographical region may have. With low Fst SNPs, whatever balancing selection may exist at any SNP must exist in all populations. The actual cause of the low Fst in the SNPs we screen is most likely that they are drawn from the low Fst tail of the random distribution of neutral SNPs.

Gill et al. [2] had recommended that candidate SNPs would need to be tested by several laboratories before being considered for forensic applications. While that is still advisable for these markers, it seems unlikely that very different allele frequencies will result since we know from many years of data being accumulated on populations that allele frequencies tend to be similar in geographically close populations [16,17]. The 40 populations studied here cover most major regions of the world; the regions not covered are flanked by those that have been studied. More important will be independent samples to show that the few large associations among markers are indeed the chance events they seem to be.

It is also noteworthy that the populations with the largest average probability of identity (and the highest frequency for the most common genotype) tend to be the more genetically isolated populations (Mbuti, Samaritans, Komi Zyriane, Taiwanese aboriginals, Nasioi) and the Native American groups. This is not unexpected. The isolated populations have undergone more genetic drift and tend to have less variation than the less isolated and historically larger populations. The Native American groups are known to be diverse [16], possibly because of historical reproductive barriers between tribes. We would expect the Fst values to increase as more such populations are studied for these markers. However, the much more relevant factors – the frequency of the most common genotype and the average probability of identity – are not likely to greatly exceed the ranges seen for the 40 populations that we have studied.

The values in Figs. 5 and 6 are calculated for ideal populations. Actual calculations in a forensic setting could include θ, the parameter for within-population substructure, as recommended in the 1996 report of the NRC Committee on DNA Technology in Forensic Science [18]. That report recommended that a value of θ of 0.01–0.03 would be appropriate in most situations. However, if the subgroups have identical allele frequencies, correction for substructure is unnecessary. While absence of significant substructure cannot be known from our data, the similarity of allele frequencies globally greatly reduces the likelihood of substantial allele frequency differences among subgroups within a population. Thus, we assume that any correction factor will be small and not greatly alter the match probabilities from what we have calculated.

The preliminary panel of 19 SNPs works quite well, already yielding match probabilities that have probative value, even if not at the level of individualization achievable with CODIS markers. Given the yield of at least 19 from an initial 195 SNPs and more than 2000 more from which to select, we should have no problem extending the panel to >45 SNPs. At the levels of heterozygosity we are achieving, a panel of 45 SNPs would give match probabilities in the range of 10^{-15}, easily in the range achieved with the CODIS markers. Moreover, with the HapMap project [19] now

complete on the first four populations for $\sim 10^6$ markers, there are potentially thousands of additional markers that could pass the initial screening criteria. We could also immediately increase the size of our panel by accepting markers with Fst <0.07. Five markers had Fst values above our cut-off of 0.06 but <0.07. Were we to incorporate these markers into the preliminary panel, the variation among populations in average match probability would increase somewhat with a range from about 10^{-8} to 10^{-10} although 35 out of 40 populations would be in a narrower interval between 10^{-9} and 10^{-10}.

Vallone et al. [20] 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 10 times as many individuals (~2100 versus 189). The distribution (Table 4) of the number of loci matching for the more than 1.74 million pairwise comparisons of 1895 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.

Other SNP panels have been proposed [20–22]. While the SNPs have usually been selected for high heterozygosity in the target population, they have not generally been tested in multiple populations. Some have subsequently been tested as part of the HapMap project [19] or by Perlegen [23] or by Celera (unpublished) and show substantial variation in allele frequencies among the populations. For example, all 12 of the SNPs in the Syvanen et al. [21] panel fail our criteria. Of the 36 autosomal SNPs in the Petkovski et al. [22] panel, 12 fail on the heterozygosity >0.45 criterion and the rest fail on the allele frequency variation criterion (Fst) using data from dbSNP. Vallone et al. [20] did test their panel on samples of three ethnic groups and while still useful in all three, most appear to fail our criterion of low Fst.

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 10 match probabilities from the different allele frequency sets. Table 5 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 2 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

Table 4
All unique pairwise comparisons of individuals for 19 best SNPs; overall results for 1895 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	2206	2253
13	195	8617	8812
12	683	27237	27920
11	1600	69080	70680
10	3589	140296	143885
9	6203	230457	236660
8	8575	306803	315378
7	9940	331964	341904
6	9030	284689	293719
5	6389	191950	198339
4	3577	99134	102711
3	1420	37684	39104
2	469	10224	10693
1	98	1706	1804
0	22	138	160
Total pairings	51848	1742717	1794565

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.

Table 5
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

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.

APOB marker in Fig. 1 and 18 others with similarly high Fst. As can be seen in Table 5, 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.

We conclude that the 19 SNPs in our panel are statistically independent. The median (0.01) and mean (0.03) LD values are close to zero and the computed LD values that are significantly different from zero are approximately what would be expected by chance. About 99.5% of all LD values are <0.25. The small number of larger LD values (≥ 0.25) occurred between unlinked markers (31 involve SNPs paired from different chromosomes and three are far apart on the same chromosome) and almost all (31 of 34) of the larger LD values involved different SNP pairs. The three different marker pairs that repeat once (in different populations) involve SNPs on different chromosomes. Of course, epistatic effects on fitness are possible but in the absence of supportive patterns such as very strong associations and/or multiple populations displaying the same association, the likelihood of an epistatic effect distorting population frequencies in only one population is vanishingly small, especially when compared to chance levels of association. As we identify additional appropriate markers we will be able to discard any that might not be fully independent.

It is not our intention to advocate any typing protocol. We used TaqMan for the screening procedures because we were screening markers individually and did not have to develop or optimize the assays. Because TaqMan is not capable of being multiplexed, as will be essential in any actual forensic application, any forensic application will require a different typing method. However, the SNPs we are identifying are in the public domain and any individual or corporation wishing to can develop methods for implementing this panel in a forensic setting.

Acknowledgements

This work was funded primarily by NIJ Grant 2004-DN-BX-K025 to KKK. We thank Applied Biosystems for making their allele frequency database available to us. We would like to thank Jia-Nee Foo and Michael Donnelly for their excellent technical work on initial phases of this project. We also want to acknowledge and thank the following people who helped assemble the samples from the diverse populations: F.L. Black, B. Bonne-Tamir, L.L. Cavalli-Sforza, K.Dumars, J. Friedlaender, D. Goldman, L. Giuffra, K. Kendler, W. Knowler, F. Oronsaye, J. Parnas, L. Peltonen, and K. Weiss. In addition, some of the cell lines were obtained from the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University, Israel, and the African American samples were obtained from the Coriell Institute for Medical Research, Camden, NJ. Special thanks are due to the many hundreds of individuals who volunteered to give blood samples for studies of gene frequency variation.

References

[1] A. Amorim, L. Pereira, Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs, Foren. Sci. Int. 150 (2005) 17–21.
[2] P. Gill, 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), Sci. Justice 44 (2004) 51–53.
[3] J.J. Sanchez, C. Borsting, C. Hallenberg, A. Buchard, A. Hernandez, N. Morling, Multiplex PCR and minisequencing of SNPs—a model with 35 Y chromosome SNPs, Foren. Sci. Int. 137 (2003) 74–84.
[4] D.E. Reich, 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, Nat. Genet. 32 (2002) 135–140.
[5] Q.Y. Huang, 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, Am. J. Hum. Genet. 70 (2002) 625–634.
[6] B.M. Dupuy, M. Stenersen, T. Egeland, B. Olaisen, Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci, Hum. Mutat. 23 (2004) 117–124.
[7] M.D. Coble, J.M. Butler, Characterization of new miniSTR loci to aid analysis of degraded DNA, J. Foren. Sci. 50 (2005) 43–53.
[8] J.M. Butler, Y. Shen, B.R. McCord, The development of reduced size STR amplicons as tools for analysis of degraded DNA, J. Foren. Sci. 48 (2003) 1054–1064.
[9] M.M. Holland, 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, Croat. Med. J. 44 (2003) 264–272.
[10] R. Chakraborty, K.K. Kidd, (Perspective) The utility of DNA typing in forensic work, Science 254 (1991) 1735–1739.
[11] R.C. Lewontin, D.L. Hartl, Population genetics in forensic DNA typing, Science 254 (1991) 1745–1750.
[12] S. Wright, The genetical structure of populations, Ann. Eugenics 15 (1951) 323–354.
[13] K.K. Kidd, A.J. Pakstis, W.C. Speed, J.R. Kidd, Understanding human DNA sequence variation, J. Hered. 95 (2004) 406–420.
[14] B. Devlin, N. Risch, A comparison of linkage disequilibrium measures for fine-scale mapping, Genomics 29 (1995) 311–322.

[15] M.D. Teare, A.M. Dunning, F. Durocher, G. Rennart, D.F. Easton, Sampling distribution of summary linkage disequilibrium measures, Ann. Hum. Genet. 66 (2002) 223–233.
[16] L.L. Cavalli-Sforza, P. Menozzi, A. Piazza, The History and Geography of Human Genes, Princeton University Press, Princeton, 1994.
[17] S.A. Tishkoff, K.K. Kidd, Implications of biogeography of human populations for race and medicine, Nat. Genet. 36 (Suppl) (2004) s21–s27.
[18] National Research Council Committee on DNA Technology in Forensic Science, The evaluation of Forensic DNA Evidence/ Committee on DNA Forensic Science: An update., National Academy Press, Washington, DC, 1996.
[19] International HapMap Consortium, The International HapMap Project, Nature 406 (2003) 789–796.
[20] P.M. Vallone, A.E. Decker, J.M. Butler, Allele frequencies for 70 autosomal SNP loci with U.S. Caucasian, African American, and Hispanic samples, Foren. Sci. Int. 149 (2005) 279–286.
[21] A.C. Syvanen, A. Sajantila, M. Lukka, Identification of individuals by analysis of biallelic DNA markers, using PCR and solid-phase minisequencing, Am. J. Hum. Genet. 52 (1993) 46–59.
[22] E. Petkovski, C. Keyser-Tracqui, R. Hienne, B. Ludes, SNPs and MALDI-TOF MS: tools for DNA typing in forensic paternity testing and anthropology, J. Foren. Sci. 50 (2005) 535–541.
[23] D.A. Hinds, L.L. Stuve, G.B. Nilsen, E. Haperin, E. Eskin, D.G. Ballinger, K.A. Frazer, D.R. Cox, Whole-genome patterns of common DNA variation in three human populations, Science 307 (2005) 1052–1053.