An expanded, nearly universal, panel of SNPs for individual identification

Andrew J. Pakstis, William C. Speed, Judith R. Kidd, Kenneth K. Kidd

Department of Genetics, Yale University School of Medicine, New Haven CT 06520

ABSTRACT

We have previously published population genetics criteria for SNPs for individual identification (IISNPs)--nearly maximum informativeness in populations from all parts of the world--as well as a panel of 40 candidate SNPs meeting those criteria.* This panel gave 40-SNP genotype probabilities of $<10^{-16}$ in almost all populations. Some have suggested that our criteria are too stringent in that we have included several small, isolated groups among the populations used to screen SNPs. Therefore, we have 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 results in a much 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, more stringent criteria as well as the relaxed criteria. Many of these candidate SNPs (now >80) 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 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 broadly.

*Kidd et al. 2006. *Forensic Science International* 164:20-32; Pakstis et al. 2007. *Human Genetics* 121:304-317. PDF files of these papers can be downloaded at: http://info.med.yale.edu/genetics/kkidd/pubs.html. (Publ. #449 & #461 respectively)
**Sanchez et al. 2006. *Electrophoresis* 27:1713-1724.

TYPES OF PANELS OF SNPs FOR FORENSIC APPLICATIONS

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

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

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

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

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 1) meet the criterion of high average informativeness (measured as heterozygosity) and low allele frequency variation among populations (measured as Fst). Also presented are examples of AISNPs and LISNPs.

GENERAL CRITERIA FOR USE OF A SNP 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 that will 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? See panel).

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 multiethnic societies such as the USA.

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.

d. If the panel is also to be used in paternity testing, the markers should be unlinked as well.

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, 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 samples of close to 50 individuals per population and/or close to 100 individuals from closely related populations in a given region to allow moderate accuracy for each allele frequency estimate.

Using screening procedures described in our two publications (Kidd et al., 2006; Pakstis et al., 2007) we have identified a large number of candidates for an IISNP panel. These candidates meet criteria 1, 2, and 3 above and meet 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.

TABLE 1. Populations included in forensic studies

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

RESULTS AND DISCUSSION

We have identified 108 candidate SNPs for an IISNP panel with Fst <0.06 and average heterozygosity >0.4. Their Fst values and heterozygosities are given in Figures 1 and 2. Because our original set of 40 populations included several small isolated and/or inbred groups, we have reduced our set of populations to the larger populations more likely to be relevant in forensic settings, especially in the USA and Europe. Figure 1 shows the comparison of Fst values in the reduced set of 31 populations compared to the original set of 40 populations. The dbSNP rs numbers are given in the figures.

Many of the SNPs are closely 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 among the molecularly and genetically close SNPs.

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 large numbers 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. While we do not consider multiplexing a relevant issue, we recognize others might. Another consideration is uniqueness of the SNP and ease of typing using small amplicons. Since all of these 108 have been typed with TaqMan and have given high quality typing results, these criteria have been met for all.

The most controversial issue will be whether or not intronic SNPs must be excluded. In our initial search for appropriate IISNPs we did not consider whether or not the SNP was in a functional element or an intron, because of the logic described under "Relevance of a gene to marker selection". Many of these 108 SNPs are in introns and some are in intergenic regions (by current knowledge) but in regions with high conservation in mammals. While we argue that intronic SNPs are acceptable, we will also argue that intergenic SNPs in highly conserved regions 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 2.

TABLE 2. 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

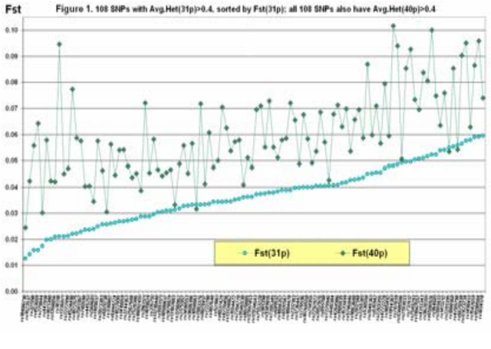
Figure 1. 108 SNPs with Avg.Het(31p)>0.4, sorted by Fst(31p); all 108 SNPs also have Avg.Het(40p)>0.4

RELEVANCE OF A GENE TO MARKER SELECTION

What do "no medical or personal information" and "not in a gene" 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 for a panel for individual identification. The scientific logic is outlined in the following.

Since one of the criteria for a "Universal" panel of IISNPs is that heterozygosity is high around the world, the SNP itself is by definition normal genetic variation with nearly equal allele frequencies in all populations and cannot be deterministic for a Mendelian genetic disease. Similarly, it cannot have a significant impact on 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 obviously there is no point in arguing for including such SNPs.

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. Consider this example. A SNP with alleles A(60%) and G(40%) has heterozygosity of 48%. Consider it is in complete LD with a mutation M(0.1%) and the normal allele N(99.9%), such that chromosomes in the population are AN(60%), AM(0%), GM(0.1%), GN(39.9%). If the marker result is AA, there is no risk of the mutation or the disorder. If the marker result is AG, the risk of the mutation being present is 0.25%. If the marker result is GG, the risk of the mutation being heterozygous is 0.5% and being homozygous is 0.001%. Thus, while the SNP genotype does alter the risk 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? We would argue that there is no general scientific reason for excluding such SNPs, especially if the intron is large and the SNP is far from an exon. 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 has serious implications for any SNP. The 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 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.

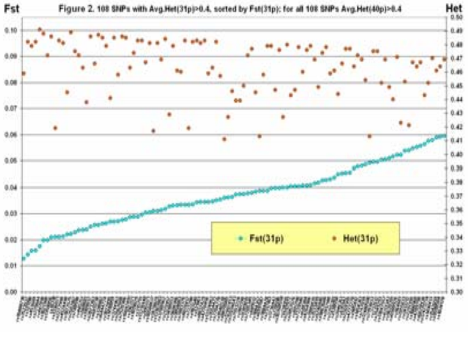
Figure 2. 108 SNPs with Avg.Het(31p)>0.4, sorted by Fst(31p); for all 108 SNPs Avg.Het(40p)>0.4

Figure 3 is an example of an AISNP. The global contour plot of the *ADH1B*47His* allele frequency is based on data from 168 populations. While this SNP does not uniquely identify a single geographic region it contributes significantly to defining ancestry geographically when employed in combination with other SNPs showing a variety of geographic patterns of allele frequencies. Taken from Li et al. "Geographically separate increases in the frequency of the derived *ADH1B*47His* allele in East and West Asia." *Am. J. Hum. Gen.* (2007) In Press.

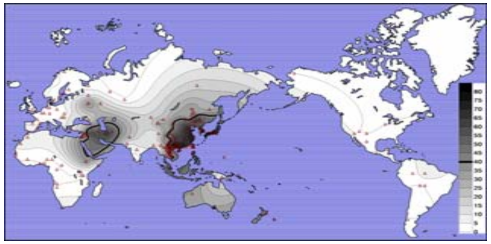

Figure 4 provides an example of a "locus" for a LISNP panel. The three SNPs are in introns of the GRAMD1C gene located in 3q13.3 and they define five haplotypes globally with at least four being common in most populations. The molecular span (~6.1kb) is so short that recombination among the SNPs will be so rare that the possibility can generally be ignored. As a 4- to 5-allele system such a locus will be more informative in determining relationships among individuals than a bi-alleleic polymorphism.

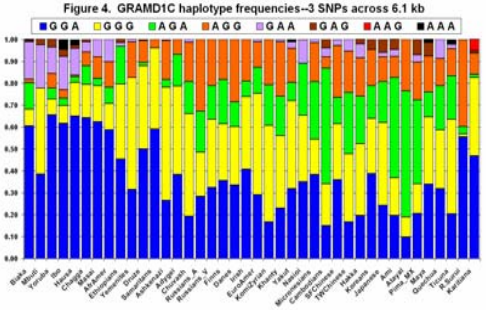
Figure 4. GRAMD1C haplotype frequencies--3 SNPs across 6.1 kb

TABLE 3. GRAMD1C gene 3-SNP haplotype example

In almost all the populations studied, 65% to 91% of phenotypes can be resolved unambiguously into genotypes by direct examination because no more than one SNP typing is heterozygous. The 7 theoretically possible ambiguous phenotypes (i.e. those with heterozygotes at 2 or more SNPs) are examined below.

Ambiguous Phenotype Frequencies for selected populations

Ambiguous	Yoruba(n=77)		AfrAm(n=84)		EuroAm(n=85)		Japanese(n=50)		Maya(n=46)	
Phenotype	Exp	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp	Obs
AG GG AG	0.043	0.078	0.090	0.143	0.171	0.176	0.176	0.160	0.176	0.217
GG AG AG	0.059	0.078	0.043	0.036	0.007	0.000	0.036	0.060	0.035	0.043
AG AG AG	0.016	0.013	0.020	0.012	0.004	0.000	0.031	0.000	0.029	0.022
AG AG AA	0.037	0.000	0.033	0.012	0.001	0.000	0.013	0.000	0.013	0.022
AG AG GG	0.002	0.000	0.003	0.000	0.003	0.000	0.019	0.000	0.017	0.022
AA AG AG	0.001	0.000	0.002	0.000	0.001	0.000	0.007	0.000	0.006	0.000
AG AA AG	0.001	0.000	0.001	0.000	<.001	0.000	0.001	0.000	0.001	0.000

Ambiguous Phenotype	Possible Genotypes Explaining Phenotypes			
AG GG AG	AGA/GGG	AGG/GGA		
GG AG AG	GAA/GGG	GAG/GGA		
AG AG AG	AAA/GGG	AAG/GGA	AGA/GAG	AGG/GAA
AG AG AA	AAA/GGA	AGA/GAA		
AG AG GG	AAG/GGG	AGG/GAG		
AA AG AG	AAA/AGG	AAG/AGA		
AG AA AG	AAA/GAG	GAA/AAG		

Probability of possible genotypes for common ambiguous phenotype AG GG AG

	Yoruba	AfrAm	EuroAm	Japanese	Maya
AGA/GGG	0.0036	0.0090	0.0550	0.0711	0.0350
AGG/GGA	0.0440	0.0687	0.0338	0.0271	0.0559
Relative Likelihood					
AGA/GGG	1.0	1.0	1.6	2.6	1.0
AGG/GGA	12.2	7.7	1.0	1.0	1.6

ACKNOWLEDGEMENTS

This work was funded primarily by NIJ Grant 2004-DN-BX-K025 to KKK awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice. Assembly of the population resource was funded by several NIH grants over many years. Recently the resource has been enlarged by funds from GM57672 and AA09379 to KKK. We thank the many collaborating researchers who helped assemble the samples from diverse populations. Special thanks are due to the many hundreds of individuals who volunteered to give blood samples for studies of gene frequency variation.