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. Those studies included several small, isolated groups. Therefore, we have 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.** Some of these 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.

*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. (Publications #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 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 1) 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.

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 multi-ethnic 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.

The 108 candidates for an IISNP panel in Figure 1 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.

Though the SNPs in **Figure 1** are a provisional list based on studies in progress, interested individuals will find the full list of 108 SNPs by rs# along with the values plotted here on the Kidd Lab web site http://info.med.yale.edu/genetics/kkidd>.

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 based on 31 populations are given in Figure 1. These 31 are 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 (blue circles) compared to the original set of 40 populations (green diamonds). The dbSNP rs numbers are given in the figure.

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.

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. Other considerations are 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. 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 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	rs1338071	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	0.457	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	14,027,869	1	N	Y	N	N	~4kb	PRDM2	3
60	rs689512	0.445	0.0384	78,306,991	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

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.

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 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.

FAMILY LINEAGE INFORMATIVE SNPs: An Example

In an initial attempt to identify LISNPs we have analyzed three SNPs in introns of the GRAMD1C gene (in 3q13.3) that 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. In almost all of the 40 population samples studied, 65% to 91% of the 3-site phenotypes can be resolved unambiguously into haplotypic genotypes by direct examination because no more than one SNP is heterozygous. Figure 2 shows the haplotype frequencies as bar graphs in each of the population samples for this potential LISNP. As a 4- to 5-allele system such a locus will be more informative in determining relationships among individuals than a di-alleleic polymorphism would be. Table 3 gives examples of the probabilistic "resolution" of ambiguous phenotypes.

TABLE 3. GRAMD1C gene 3-SNP haplotype example

The 7 theoretically possible ambiguous phenotypes (i.e. those with two or more heterozygous SNPs) are examined below. Results from populations on different continents are used as illustrations.

Ambiguous Phenotype Frequencies for selected populations

Ambiguous Phenotype	Yoruba(n=77) Exp	Yoruba(n=77) Obs	Afr.Am(n=84) Exp	Afr.Am(n=84) Obs	EuroAm(n=95) Exp	EuroAm(n=95) Obs	Japanese(n=50) Exp	Japanese(n=50) Obs	Maya(n=46) Exp	Maya(n=46) Obs
AG GG AG	0.043	0.078	0.080	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				Probability of possible genotypes for common ambiguous phenotype AG GG AG				
						Yoruba	AfrAm	EuroAm	Japanese	Maya
AG GG AG	AGA/GGG	AGG/GGA								
GG AG AG	GAA/GGG	GAG/GGA			AGA/GGG	0.0036	0.0090	0.0550	0.0711	0.0350
AG AG AG	AAA/GGG	AAG/GAA	AGA/GAG	AGG/GAA	AGG/GGA	0.0440	0.0687	0.0338	0.0271	0.0559
AG AG AA	AAA/GGA	AGA/GAA								
AG AG GG	AAG/GGG	AGG/GAG								
AA AG AG	AAA/AGG	AAG/AGA				Relative Likelihood				
AG AA AG	AAA/GAG	GAA/AAG			AGA/GGG	1.0	1.0	1.6	2.6	1.0
					AGG/GGA	12.2	7.7	1.0	1.0	1.8