SNP panels for individual identification and for ancestry inference

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

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

ABSTRACT

We have completed the development of a panel of SNPs for individual identification (IISNP) and we are optimizing a SNP panel for ancestry informativeness (AISNP) employing our resource of 44 population samples representing the major continents of the world. After testing over 500 potential candidate SNPs we have completed a final, universal panel of 92 best IISNPs based on data on 44 population samples. The list of 92 IISNPs is available at the Kidd Laboratory website (http://info.med.yale.edu/genetics/kkidd/92snpJan2009.pdf); it includes dbSNP rs-numbers and chromosome locations. All 92 IISNPs have an average heterozygosity >0.4 and the Fst values are all <0.06 on our 44 populations making these a universally applicable panel irrespective of ethnicity or ancestry. A subset of 45 unlinked SNPs constitutes an excellent panel for individual identification providing match probabilities comparable to the 13 CODIS STR markers (match probabilities of less than 10^{-17} in most of the world); their unlinked status also makes them useful for situations involving close biological relationships. Chance level linkage disequilibrium (LD) values are observed for all unique pairings of 86 of the 92 IISNPs (median LD=0.01) in each of 44 populations. The remaining 6 IISNPs show strong LD in most of the 44 populations for a small subset (7) of the unique pairings in which they occur due to close linkage; these 6 SNPs can therefore only be alternative candidates for inclusion in an IISNP panel of 86 SNPs independent at the population level.

We continue to identify candidate AISNPs and are evaluating alternative methods for optimizing the SNP panels. We have studied well over 300 candidate SNPs on up to 73 population samples. Inferences of continental origin are relatively easy to assure with various subsets of the candidate AISNPs already identified. The clinal change in allele frequencies characteristic of adjacent populations makes within-continent inference more difficult and a single compact panel of SNPs might not be achievable for all population comparisons. Multiple panels optimized for subregional inference of ancestry likely will be the endpoint of our search with an initial panel to differentiate continental regions and at least one appropriate subpanel to optimize inference within the targeted geographical region. However, we are currently able to distinguish probabilistically six groups across Eurasia and expect to be able to improve the differentiation.

PUBLIC AVAILABILITY OF SNP FREQUENCIES

As publications are submitted summarizing various stages of our work, we continue to deposit the SNP gene frequencies for the population samples studied into ALFRED, the Allele Frequency Database (http://alfred.med.yale.edu). We contribute the SNP frequencies not only for the best SNPs found for different purposes, but also the frequencies for screened SNPs studied on the small, preliminary population panels that did not have characteristics that merited additional typings on the full population panels. In addition, ALFRED continues to incorporate from publications and other sources additional allele frequency data on more polymorphisms and on more populations. Thus it is a useful reference source of allele frequency data for many DNA polymorphisms that might be used in a particular forensic setting.

SCREENING AND FINAL RESULTS FOR IISNP CANDIDATES

Our previous IISNP panel consisted of 40 SNPs based on 40 populations. While there was no significant pairwise LD in any of the populations, some pairs were sufficiently close that linkage existed. This made those SNP pairs more difficult to use in studies involving biological relationships. Therefore, in our more recent search to develop a panel of IISNPs that were universally applicable and unlinked, we preferentially targeted regions of the genome in which we did not already have good IISNP candidates in order to enlarge the number of unlinked IISNPs. We also enlarged our set of populations by adding four populations for geographic regions poorly represented in the initial 40 populations: East Africa, East Europe, South Asia, and Southeast Asia. We gleaned candidates from a very large SNP dataset (Li et al., 2008) that became available online in 2008 for the populations studied on the Human Genome Diversity Panel (HGDP). We obtained other candidate markers that we identified from the Shriver et al. (2005) dataset which studied 14 populations from around the world. We typed all interesting SNPs on 44 population samples (Table 1). With better data for selecting candidates, we had a higher percentage meeting our acceptance criteria of average heterozygosity >0.4 and the Fst values <0.06 on our 44 populations. Figure 1 shows these values based on 44 populations for the final 92 candidate IISNPs with the SNPs rank-ordered (left to right) from lowest to highest Fst. No meaningful departures from Hardy-Weinberg ratios were seen for any of the 92 IISNPs in the populations studied. All 92 IISNPs have been reliably typed by TaqMan; how best to multiplex specific subsets to use for different identification tasks will likely depend on the application.

When pairwise LD does not exist, as among the 45 unlinked IISNPs, the SNPs are statistically independent at the population level and the "product rule" can be used to calculate match probabilities. Figure 2 displays match probabilities and most common genotype frequencies for each population for this set of 45 unlinked IISNPs. Most of the populations have match probabilities $<10^{-17}$ and many are $<10^{-18}$; even some of the smaller, more isolated populations have match probabilities $<10^{-15}$. Thus, this set of 45 unlinked SNPs is an excellent panel for individual identification with match probabilities comparable to the CODIS STR panel and these are not highly dependent on ethnicity. Thus, it is safe to say with considerable scientific justification that a maximum match probability of $<10^{-15}$ can be used for any forensic match between any crime scene and any defendant anywhere in the world. The unlinked status of these 45 SNPs also makes them useful for situations involving close biological relationships. If relationships are not involved, more of the 92 IISNPs can be added to the set to make the match probabilities even smaller. Computing match probabilities based on all 86 IISNPs that show no LD gives results in the range of 10^{-31} to 10^{-35} for the 44 populations. At this level, the actual probability has no realistic meaning other than uniqueness among all humans.

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

Figure 1. Fst & Avg.Heterozygosity for the Final 92-IISNP Panel

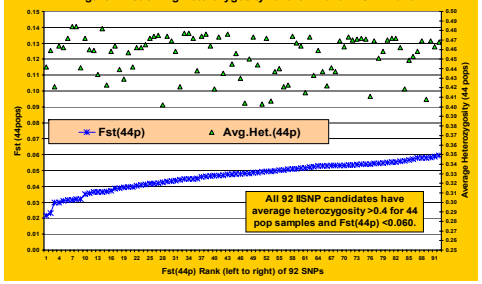


Figure 2. Match probabilities & most freq. genotype frequency--45 IISNPs

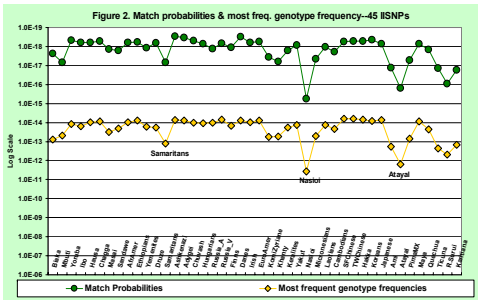


Table 1. 73 population samples studied

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

Legend:
X green highlighting of X indicates 4 newest population samples for IISNP project
* Samples (usually a subset) contributed to the HGDP-CEPH panel in Paris
† Samples with many related individuals; most analyses only include unrelated individuals

STATUS OF ANCESTRY INFERENCE STUDIES

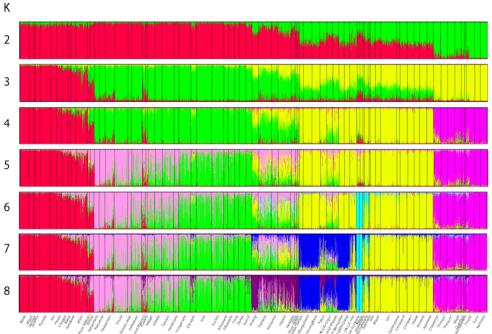
Our goal is to optimize a panel of high-Fst Ancestry Informative SNPs (AISNPs) that can provide useful information on the geographic-ethnic ancestry of the person whose DNA is being analyzed. Such a panel could help evaluating eye witness testimony and identifying ethnicity of skeletal remains or other forensic evidence. We have not attempted to identify phenotype other than whatever may be correlated with geographic ancestry. A panel of AISNPs should be highly differentiating of ancestral origins of an individual DNA sample with a reasonable number of SNPs (that is, ideally less than100) and have the population genetics support that would allow a high enough probability of correct ancestral assignment to make it a strong investigative tool. We are striving for greater specificity of ancestry than is generally provided by "continental" assignment. It is clear that differentiation, on average, among even closely clustering groups (e.g., European populations) is possible if enough markers are used (Li et al., 2008)—that is not the problem. The problem is identifying ancestry for a single individual with a reasonable number of SNPs. Ultimately, it may be that no single small panel will be optimal for all questions.

We started by genotyping individuals in our 40-44 population samples for a large number of potential AISNPs. Of these SNPs, many did not show large frequency differences and others failed our internal standards for data quality, leaving a current working set of 320 SNPs typed on 44 populations. These have been assembled from several sources as SNPs with high Fst among many populations or large allele frequency differences between populations from different geographic regions. Originally, we selected from data on only three geographic regions, Africa, Europe, and East Asia. More recently we have had larger datasets from which to select candidates and are finding a higher proportion to be useful. We have incorporated two published sets of AISNPs in their entirety: the 10 from Lao et al. (2006) and the 128 from Kosoy et al. (2009). The difficulty all along in achieving satisfactory results on our goal of an efficient yet robust AISNP panel has been determining which of many SNPs contribute to a clear distinction between population groups. We have employed the STRUCTURE program to evaluate the ability of the data on these markers to give a clear pattern of the four "continental" regions while clearly differentiating the several intermediate populations. Some small subsets of high Fst SNPs (~20 SNPs) are excellent for discriminating ancestry from the major continental regions of the world (K=4 in STRUCTURE analyses). The geographically intermediate populations, however, still show "mixed ancestry", an expected statistical artifact of a largely clinal distribution being forced into a small number of discrete clusters. In order to have a population set that will allow us to search for AISNPs giving finer geographic resolution, we have an expanded dataset 73 populations and a total of 3464 individuals. Some of the additional samples are newly arrived in our lab (e.g., Zaramo); most are small DNA samples sent to us by many different collaborators who will be coauthors of final results. These samples are sufficient to use for SNPs that have already been shown to be good at a global level for distinguishing between different continental level ancestries.

Figure 3 presents our preliminary results of an expanded pilot study on the 73 population samples using all 128 of the Kosoy et al (2009) SNPs. The genotyping is not yet complete on all samples for all SNPs but partially missing data should not greatly alter the results. These preliminary results provide a replication of the value of these SNPs that were originally studied on only 8 populations by the authors. The results are encouraging in that we find we can distinguish eight clusters rather than just the four continental groups and their value for detecting admixture is good. However, in terms of inference of ancestry the panel is not great for Eurasia. Individuals from Sub-Saharan Africa, Northwest Europe, far East Asia, and the Americas are very clearly placed into their respective clusters. The degree of certainty for the other groups, however, is not good in that considerable individual to individual variation exists at the higher numbers of clusters. For example the populations from the Middle East, Pakistan, and southern Europe have rather high probabilities of miss-assignment. Averaged by population, people from four Middle East populations have a 4-22% chance of being miss-assigned; three Pakistani groups have a 42-64% chance of being miss-assigned; and the southern Europeans have about an equal probability of being assigned to Europe or the middle East. Given the variation among individuals in these populations, the results are insufficient for a forensic application applied to a single individual. Similarly, South Asia and Central Asia are probabilistically distinguishable at K=8, but only in half the replicate runs of STRUCTURE (not shown). Our overall objective will be to identify markers that will as much as possible clarify those additional clusters. We are striving for a universal panel of AISNPs, but these clarifications are also specifically areas of forensic relevance within the United States given our increasingly heterogeneous population.

Figure 3. STRUCTURE results for 128 SNPs on 73 population samples; K=clusters

The STRUCTURE analyses summarized in Figure 3 represent the most frequent patterns seen at each K value among the 10 independent runs of the MCMC search in STRUCTURE. The number on the right for each K value represents the number of times out of the 10 runs that the program gave that pattern, the most frequent for that K value.



STATUS OF ANCESTRY INFERENCE STUDIES (continued)

There will continue to be individuals and populations that show significant non-zero probabilities of belonging to more than one cluster. That is not strictly evidence of admixture (though admixture could be a cause), but rather indicates that the SNPs being used have intermediate allele frequencies in those populations as expected for a clinal distribution. This is illustrated for African Americans in Figure 3. It is expected that individuals will vary in their level of admixture but it is highly unlikely that all of the partial cluster assignments seen for individuals actually represent those levels of admixture of those ancestries.

While these 128 SNPs are clearly useful for determining admixture, they are not necessarily good for identifying ethnicity for an unknown sample coming from an admixed population. The elaboration in Figure 4 of the African American sample from the K=8 STRUCTURE analysis (in Figure 3) illustrates this. All individuals are self-identified African Americans in the Coriell cell line repository. At the top the population averages are plotted for all 73 populations. On average, about 25% of the African American sample shows non-African signal (upper left enlargement). However, when individuals are considered (lower left and bottom enlargements) there is extensive variation. When sorted by probability of individual assignment to different "geographic-ethnic"clusters, the variation can be seen to be considerable. Several of the individuals are more likely to be considered non-African than African.

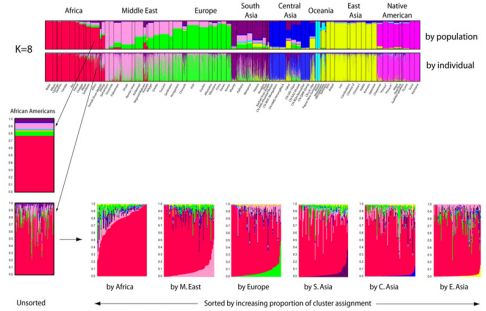


Figure 4. Preliminary STRUCTURE analyses for 73 populations using nearly complete data for the 128 SNPs of Kosoy et al. (2009). Eight clusters can be resolved with reasonable correspondence to geographic origins of the populations. At the top the proportional cluster assignments are shown averaged by populations and for each individual. At the side and across the bottom the cluster assignments for African Americans are shown in greater detail. Individuals are sorted by amount of assignment to each of the 6 major clusters showing partial assignments. No individuals have appreciable assignments to Oceania and Native American clusters.

PUBLICATIONS RELATED TO THIS NIJ FUNDED PROJECT

Butler et al. **2008**. *Prog in For Genet* Genetics Suppl Series 1:471-472.

Kidd et al. **2006**. *For Sci Intl* 164:20-32.

Pakstis et al. **2007**. *Human Genetics* 121:304-317.

Pakstis et al. **2008**. *Prog in For Genet* Genetics Suppl Series 1:479-481.

Note: PDF files for the above papers are downloadable (Pubs.#468, #449, #461, & #467 respectively) at: http://info.med.yale.edu/genetics/kkidd/pubs.html..

OTHER REFERENCES

Kosoy et al. **2009** *Human Mutation* 30:69-78

Lao et al. **2006**. *Am J Hum Genetics* 78:680-689.

Phillips et al. **2007**. *For Sci Intl:Genet* 1:273-280.

Pritchard et al. **2000**. *Genetics* 155:945-959.

Rosenberg **2005**. *J Computational Biol* 12:1183-1201.

Shriver et al. **2005**. *Human Genomics*. **2**:81-89.

DATABASES

ALFRED, The Allele Frequency Database; http://alfred.med.yale.edu

ACKNOWLEDGEMENTS

This work was funded primarily by NIJ Grants 2004-DN-BX-K025 and 2007-DN-BX-K197 to KKK awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. Points of view in this presentation 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. We also thank Applied Biosystems for supplying some of the TaqMan reagents used in these studies.