Better panels of SNPs for ancestry inference and 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

In our earlier presentations we noted that panels of SNPs for forensic applications fall into four categories: individual identification (iiSNPs), ancestry inference (aiSNPs), lineage (family) inference (liSNPs), and phenotype inference (piSNPs). In all cases the basic information required is knowledge of how the allele frequencies vary around the world. Our research has focused on developing optimized panels of iiSNPs and aiSNPs using our unique resources of 44 population samples (Table 1) representing populations from around the world. Here we present updates on our progress. With the objective of identifying a small number of iiSNPs (~45) with globally high heterozygosity combined with very low Fst as well as no linkage disequilibrium (LD) or close linkage, we are studying additional markers and additional populations. We have almost completed typing our expanded panel of 108 iiSNP candidates on four additional populations adding a new region (India) and strengthening coverage in other parts of the world (East Africa, Eastern Europe, and Southeast Asia). These updated results for 44 population samples alter somewhat the rank order of the existing markers. We have also identified new candidate SNPs that are on completely different chromosomes compared to the current best markers and we are typing them on all 44 populations. Other researchers are typing many of these markers on additional populations that are available to them. An improved panel of candidates will result from the combined results of these studies. Our initial work on aiSNPs has produced over 200 new candidate SNPs that are excellent for discriminating ancestry from the major continental regions (K=4 in STRUCTURE analyses) with geographically intermediate populations showing “mixed ancestry”, an expected statistical artifact of a clinal distribution being forced into discrete clusters. Most of the information resides in a small number (12 to 30) of those SNPs, but the small subsets are not capable of greater subdivision of the populations. We have also shown that the best of the SNPforID consortium aiSNPs continue to perform reasonably well at the four cluster step on the extended panel of population samples in our laboratory but they are also not sufficient for clean separation within the major continental groups. Additional candidate aiSNPs are being typed and analyses are ongoing to attempt to broaden the panel of aiSNPs and global samples in order to identify those most informative in identifying ancestry that is more refined than assuming (incorrectly) that human populations falls into just four clusters.

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 various population samples studied to 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.

SCREENING FOR ADDITIONAL SNP CANDIDATES

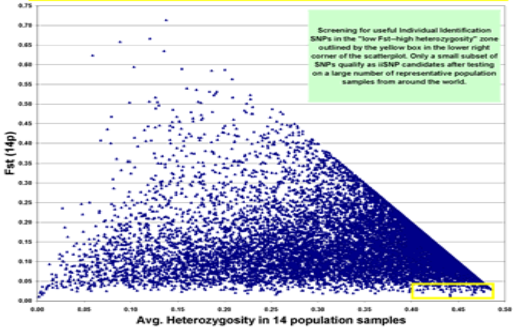
iiSNPs. We are targeting SNPs that not only meet our combined low Fst and high heterozygosity criteria but we are also preferentially targeting regions of the genome (unlinked as well as those not showing LD with existing candidates) in which we do not already have good iiSNP candidates. Currently, we are in the process of typing 14 new SNPs on our 44 population samples that were gleaned from analyses of a very large SNP dataset recently available online for the populations studied on the Human Genome Diversity Panel (HGDP). We have also generated lists of markers to investigate that we have identified from the large number of SNPs in the Shriver et al. (2005) dataset which studied 14 populations from around the world. Figure 1 displays the scatterplot of Fst by average heterozygosity values for the Shriver dataset and the yellow box in the lower right corner of the figure shows the zone of special interest for finding iiSNPs.

aiSNPs. Previously, we identified through various methods (including publications like Lao et al. 2006) a set of 249 SNP candidates for use in our ancestry inference studies. Recently, we have added 19 SNPs from a larger panel of 34 identified by the SNPforID consortium (Phillips et al. 2007) and another 128 SNPs from various admixture studies focused on studying African American and Hispanic groups in the United States. We have also screened the Shriver et al. (2005) dataset and have lists of other potential aiSNPs to evaluate as time and resources allow. Currently, we have almost 400 SNPs that we have typed on our panel of 40 populations (Table 1) and will soon finish typings on these SNPs for 5 new populations—the four new ones noted in Table 1 as well as a sample of 40 Zaramo individuals from Tanzania.

TABLE 1. Populations included in forensic studies; Note:4 new population samples have green highlighting

Population samples (n) at Kidd Lab	Low Fst--High Het. 44 pop. samples	31 population samples	Population samples (continued)	Low Fst--High Het. 44 pop. samples	31 population samples
Biaka (68)	X	X	Komi Zyrian (47)	X	X
Mbuti (39)	X		Khanty (50)	X	X
Yoruba (78)	X	X	Keralites (30)	X	
Ibo (48)	X	X	Yakut (51)	X	
Hausa (39)	X	X	Nasioi (23)	X	
Chagga (45)	X	X	Micronesians (37)	X	
Masai (22)	X	X	Cambodians (25)	X	X
Sandawe (40)	X		Laotians (118)	X	
African Americans (90)	X	X	Chinese, San Francisco (59)	X	X
Ethiopian Jews (32)	X	X	Chinese, Taiwan (49)	X	X
Yemenite Jews (43)	X	X	Hakka (41)	X	X
Druze (102)	X	X	Koreans (52)	X	X
Samaritans (41)	X		Japanese (50)	X	X
Ashkenazi (83)	X	X	Ami (40)	X	
Adygei (54)	X	X	Atayal (42)	X	
Chuvash (42)	X	X	Pima, Mexico (53)	X	X
Hungarians (89)	X		Maya (52)	X	X
Russians, Archangel (34)	X	X	Quechua (22)	X	X
Russians, Vologda (48)	X	X	Ticuna (65)	X	
Finns (36)	X	X	Rondonian Surui (47)	X	
Danes (51)	X	X	Karitiana (57)	X	
Irish (118)	X	X	Average(R.Surui,Karitiana)		X
European Americans (92)	X	X			

Figure 1: Fst by avg. heterozygosity scatterplot for 11,454 SNPs--Shriver dataset



Screening for useful Individual Identification SNPs in the "low Fst--high heterozygosity" zone outlined by the yellow box in the lower right corner of the scatterplot. Only a small subset of SNPs qualify as iiSNP candidates after testing on a large number of representative population samples from around the world.

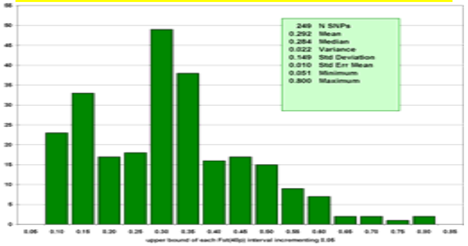
RESULTS for 108 iiSNP candidates on expanded population panel

The 4 new populations have been typed for the set of 108 iiSNP candidates defined last year. As can be seen in Figure 2, the average heterozygosities across the 44 population samples are all still above the minimum threshold of 0.4 that we adopted at an earlier stage of our studies. The Fst values (green triangles in Figure 2) for the expanded panel do change but they are strongly correlated with the Fst values obtained earlier on the panel of 40 populations (red circles in Figure 2). The Fst rankings of the 108 SNP candidates can and do change a great deal across the 3 population panels considered; figure 3 visualizes this by plotting the 108 iiSNP ranks for the 31 population panel on the x-axis relative to the y-axis which tracks the ranks for the Fst values for each of the three panels. Each of the 4 new population samples studied represents large populations of potential relevance for forensic applications. More such populations need to be studied to assure the general value of iiSNPs. The 31 population sample excludes (see Table 1) some of the small and inbred populations (intentionally included in the 40 population panel) that would not be quite so typical of groups studied in forensic applications in Europe and the U.S.A.

STATUS OF ANCESTRY INFERENCE STUDIES

The ultimate goal of this project is to identify a relatively small subset of SNPs that will do the best job of inferring the ethnic group membership of individuals. We have been exploring a variety of strategies for identifying an optimal set of markers for this purpose. The different statistical approaches do produce different sets of SNPs and they can be compared for how well the test sets work to infer ancestry. SNPs with a variety of strongly differentiating patterns among ethnic groups are required and, while some SNPs do give roughly equivalent contributions to the task, the essential problem is that there are virtually no SNPs that uniquely identify ethnic groups but rather even the best SNPs identified show clinal gradations across various subsets of the world’s ethnic groups. At the annual meeting of the American Society of Human Genetics this November, we (along with some colleagues in statistics) will be presenting the results of one of the more interesting strategies which uses a greedy algorithm to identify SNPs that minimize the error rate in predicting ethnic group membership. We are also pursuing additional statistical approaches for identifying optimal subsets of aiSNPs but as yet do not have detailed results to present.

Figure 4. Fst(40p) distribution for 249 aiSNP candidates



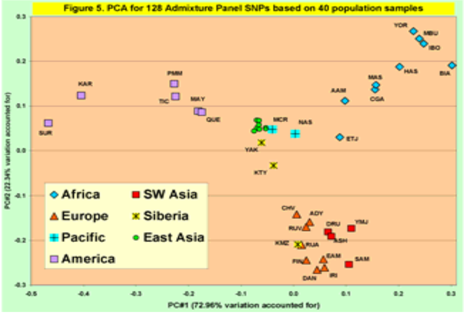


Figure 5. PCA for 128 Admixture Panel SNPs based on 40 population samples

STATUS OF ANCESTRY INFERENCE STUDIES (CONT’D)

We have also been accumulating additional candidate SNPs (as noted earlier ~400 candidate aiSNPs are being typed currently and more are being sought) as well as enlarging our panel of population samples on which to evaluate the merits of the SNPs. To date, the various candidate markers we have accumulated (and various, selected subsets thereof) have succeeded best only at differentiating individuals from four major geographical regions of the world—(1) sub-Saharan Africa, (2) Europe plus SW Asia, (3) East Asia, and (4) the Americas. The initial 249 of the almost 400 candidate aiSNPs that we are currently studying, for example, give an Fst distribution (based on the 40 population samples) in which one can see (Figure 4) that most of the SNPs have Fst values well above the average of 0.14 that one sees for a distribution of many hundreds of SNPs not selected for differentiating frequency variation. A similar pattern of strong population differentiation occurs for the 128 candidate aiSNPs ascertained from the various admixture studies. Principal components analysis (PCA) results for the 128 admixture aiSNP candidates presented in Figure 5 documents the overall informativeness of this set of candidate markers. The first 2 principal components of the PCA account for 95% of the variation and the overall pattern compares favorably to that seen for a PCA analysis based on over 2,500 “random” SNPs organized as multi-SNP haplotypes (not shown). As shown in Figure 6, the 249 marker set gives a clear pattern of four regions with several intermediate populations when analyzed with the STRUCTURE program of Pritchard et al. (2000). Figure 7 displays the results of STRUCTURE runs for 2 subsets of the 249 marker set. Figure 7a shows how the 2,044 individuals are assigned for the 20 highest SNPs by the informativeness measure estimated by INFOCALC (Rosenberg 2005) while Figure 7b gives the results for the 20 SNPs with the highest ranks based on Fst values. The diverse individuals studied are arrayed in the same order in Figures 6 and 7. As can be seen from the images, neither of the 2 subsets (nor others employed thus far) give very “clean” assignments to the 4 geographical groupings or clusters.

Figure 6. STRUCTURE results for 249 candidate aiSNPs with 50,000 burn-ins, 200,000 MCMC, N=2044 individuals, K=4

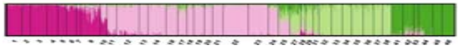
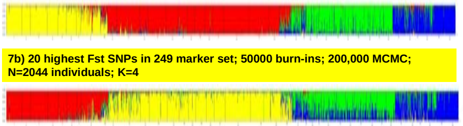


Figure 7. STRUCTURE analyses for 2 subsets of the 249 marker set.

7a) 20 highest INFOCALC-Informativeness SNPs, 50,000 burn-ins, 200,000 MCMC, N=2044 individuals, K=4



7b) 20 highest Fst SNPs in 249 marker set; 50000 burn-ins; 200,000 MCMC; N=2044 individuals; K=4

PUBLICATIONS RELATED TO THIS NIJ FUNDED PROJECT

Butler et al. **2008**. *For Sci Intl Genet (Suppl.)* In Press, e-pub online doi:10.1016/j.fsigss.2007.10.159.

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

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

Pakstis et al. **2008**. *For Sci Intl Genet (Suppl.)* In Press, e-pub online doi:10.1016/j.fsigss.2007.10.200.

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

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

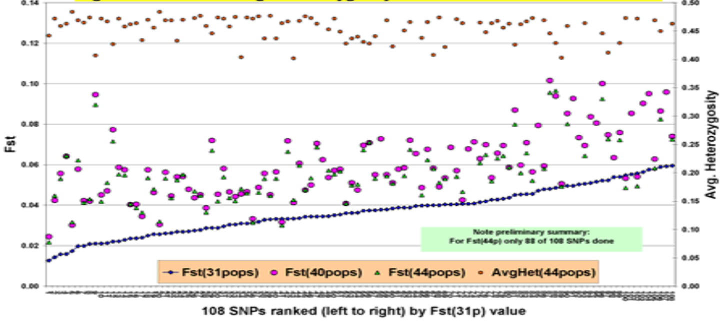
ALFRED, The Allele Frequency Database; http://alfred.med.yale.edu

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.

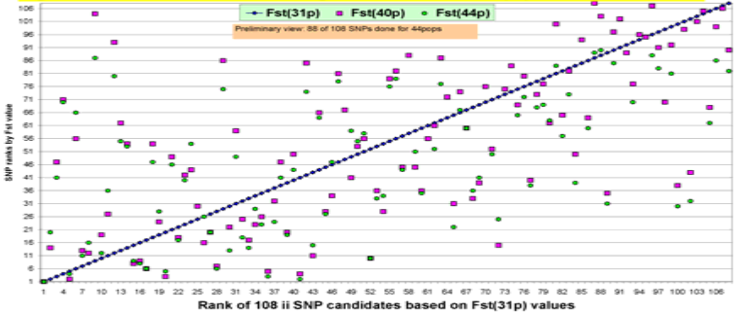
108 iiSNP candidates: The list of dbSNP rs-numbers and chromosomal locations is available online at http://info.med.yale.edu/genetics/kkidd/SNPdata2007.pdf

Figure 2: Fst and Average Heterozygosity for 108 iiSNP candidates



Note preliminary summary: For Fst(44p) only 88 of 108 SNPs done

Figure 3: 108 ii SNPs--how ranks change with Fst for different population sets



Preliminary view: 88 of 108 SNPs done for 44pops