Better SNPs for Better Forensics: Ancestry, Phenotype, and Family Identification

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

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

2010-DN-BX-K225

2007-DN-BX-K197

Ancestry Informative SNP (AISNP) Studies

We have used our population resources to study many candidate SNPs for use in ancestry inference from a DNA profile (e.g., Kidd et al., 2011a, b; Donnelly et al., 2012). We originally presented a small panel of 39 AISNPs as a prototype AISNP panel for our FROG-kb web site <frog.med.yale.edu>. That panel, called FROG39, was better than anticipated when tested with STRUCTURE on 44 populations. For ranking of most likely population of origin for many profiles of known individuals the likelihood ranges were often >30 orders of magnitude with the actual origin at or near the top of the list. The most likely population was usually not significantly better than the true population or represented the same general geographic region. We have now identified additional good AISNPs (Table 1) and expanded analyses to our collection of samples from 55 populations (Table 2). STRUCTURE runs for the original FROG39, a new overlapping FROG45 AISNP set, and the combined set of all 56 AISNPs are presented in Figure 1.

The new FROG45 AISNP set was developed by first evaluating the FROG39 AISNPs for their contribution to the STRUCTURE results and kept the "best" 28, those with the highest Fst values that were collectively a balanced set of patterns of regional variation. The pattern balancing was done by selecting the markers with the highest values from each set of regional pairwise Fst values. To this set we added 17 additional AISNPs selected in the same manner but from a separate much larger set of candidate AISNPs, some that we identified as candidates and some that we typed but were originally suggested by other studies. The objective was to develop a "First-Tier" panel of AISNPs that would identify several geographic regions that might be more finely resolved with additional AISNPs selected to maximize variation within a region. This we now refer to as "FROG45" because it will shortly be accessible on FROG-kb while a paper giving more details is being prepared. The FROG45 STRUCTURE runs (Figure 1) are "better" in that South Asia is now clearly distinct at K=8 and K=9, but Africa shows no major variation. Also shown in Figure 1 are STRUCTURE runs for the combined set of SNPs representing the union of both panels. We have not been able to identify any subset of the 56 AISNPs that is as good as the FROG39 set. Our experience is that once we have fewer than about 40 AISNPs we lose some of the geographic distinctions we see in these analyses. Of course, many much smaller subsets will give good resolution of widely differing population groups such as West Africa, Northwest Europe, far East Asia, and Native Americans. We consider this too gross an approximation to the ancestral origins in the United States to be an acceptable indicator of likely ancestry for a forensic unknown.

Our current efforts are now focused on identifying AISNPs that are highly informative within a geographic region. We think such markers would constitute panels of second-tier AISNPs to be analyzed only for ancestry refinement within a region. Some of the phenotype informative SNPs will fall within this category as well as some of the lineage informative mini-haplotypes.

Table 1: 56 Ancestry Informative SNPs
Sorted by rs-number

dbSNP rs#	Panels	Chr	nt position Build 37	HGDP Stanford	# Hap Map pops	SOURCE
rs10497191	39s	2	158,667,21	Y	11	3
rs1079597	39s	11	113,296,28	N	11	1
rs11652805	39s & 45s	17	62,987,151	Y	11	2
rs1229984	39s & 45s	4	100,239,31	N	4	1
rs12439433	39s & 45s	15	36,220,035	Y	9	2
rs12498138	39s & 45s	3	121,459,58	Y	10	3
rs12913832	39s & 45s	15	28,365,618	Y	9	1,5,7
rs1426654	39s & 45s	15	48,426,484	N	11	1,7
rs1462906	45s	8	31,896,592	N	4	1
rs1572018	39s & 45s	13	41,715,282	Y	11	3
rs16891982	39s & 45s	5	33,951,693	N	4	1,4,7
rs174570	39s	11	61,597,212	Y	11	1
rs17642714	39s	17	48,726,132	N	4	1
rs1800414	45s	15	28,197,037	N	6	1
rs1834619	39s & 45s	2	17,901,485	Y	9	3
rs1834640	39s & 45s	15	48,392,165	Y	10	3
rs1871534	45s	8	145,639,68	N	4	1
rs1919550	45s	3	121,364,17	N	4	1,5
rs192655	39s	7	90,518,278	N	11	2
rs200354	39s	14	99,375,321	Y	11	2
rs2024566	39s	22	41,697,338	Y	11	3
rs2042762	45s	18	35,277,622	N	4	1,5
rs2166624	45s	13	42,579,985	Y	11	3
rs2196051	45s	8	122,124,30	Y	11	3
rs2238151	39s & 45s	12	112,211,83	N	10	1
rs2593595	39s & 45s	17	41,056,245	Y	11	3
rs2269961	39s & 45s	2	109,579,73	Y	11	2
rs2814778	39s & 45s	1	159,174,68	N	5	1,7
rs3811801	45s	4	100,244,31	N	none	1
rs3814134	45s	9	127,267,68	N	11	1,6
rs3823159	45s	6	136,482,72	Y	10	3
rs3827760	39s & 45s	2	109,513,60	N	7	1
rs3916235	45s	18	67,578,931	N	4	1
rs4411548	39s	17	40,658,533	N	4	1
rs4471745	39s & 45s	17	53,568,884	Y	11	3
rs459920	45s	16	88,730,827	N	11	1
rs4833103	45s	4	38,815,502	Y	9	3
rs4891825	39s & 45s	18	67,867,663	Y	10	2
rs4918664	39s & 45s	10	94,921,065	Y	11	3
rs671	45s	12	112,241,76	N	6	1
rs6754311	39s & 45s	2	136,707,98	N	9	1
rs6990312	45s	8	110,602,31	Y	11	3
rs7226659	39s & 45s	18	40,488,279	Y	11	3
rs7251928	39s	19	4,077,096	Y	11	3
rs7326934	45s	13	49,070,512	N	4	1
rs735480	39s	15	45,152,371	Y	11	3
rs7554936	39s	1	151,122,48	Y	11	2
rs7657799	39s & 45s	4	105,375,42	Y	11	2
rs7722456	45s	5	170,202,98	Y	10	3
rs798443	39s & 45s	2	7,968,275	Y	11	2
rs7997709	39s & 45s	13	34,847,737	Y	11	2
rs870347	39s & 45s	6	6,845,035	N	11	2
rs917115	39s & 45s	8	28,172,586	N	11	1
rs9522149	39s & 45s	13	111,827,16	Y	11	2

Notes

Source 1: Kidd lab SNPs	
Source 2: Seldin 128 SNPs	**Kosoy et al. (2009)**
Source 3: Nievergelt 41 SNPs	**unpublished**
Source 4: Kayser 24 SNPs	**Lao et al. (2010)**
Source 5: Yaeger 107 SNPs	**Yaeger et al. (2008)**
Source 6: Goldman 186 SNPs	**Enoch et al. (2006)**
Source 7: SNPforID 34plex	**Phillips et al (2007)**

TABLE 2.
Populations studied; ordered as in Figure 1.

Region and Population	N	Region and Population	N
West & Central Africa		**West Siberia**	
Biaka	68	Khanty	50
Mbuti	38	**India**	
Lisongo	8	Keralites	30
Yoruba	77	Thoti	14
Ibo	48	Kachari	16
Zaramo	38	**East Siberia**: Yakut	51
Hausa	38	**East Asia**	
East Africa		Chinese (San Francisco)	58
Maasai	20	Chinese (Taiwan)	50
Chagga	45	Hakka	40
Sandawe	40	Koreans	54
African Americans	89	Japanese	45
Ethiopian Jews	32	**Southeast Asia**	
Southwest Asia		Lao	118
Kuwaiti	14	Cambodians	23
Samaritans	39	Ami	40
Yemenites	41	Atayal	41
Druze	99	**South Pacific**	
Europe		Malaysians	11
Roman Jews	26	Samoans	9
Ashkenazi	78	Micronesians	34
Sardinians	33	Papua-New Guinea	22
Adygei	54	Nasioi	23
Chuvash	42	**North America**	
Hungarians	89	Pima (Mexico)	53
Irish	111	Maya	48
European Americans	89	**South America**	
Russians (Archangel)	33	Quechua	22
Russians (Vologda)	47	Guihiba	11
Finns	34	Ticuna	65
Danes	56	Surui (Rondonia)	86
Komi Zyriane	46	Karitiana	54

Figure 1. STRUCTURE analyses of two overlapping AISNP panels and their union

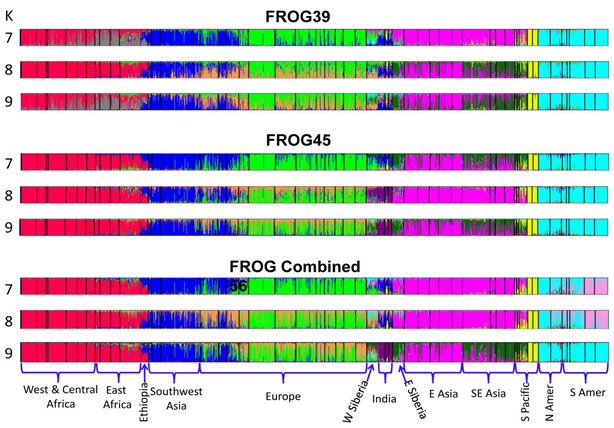


Figure 1. Two overlapping sets of AISNPs and their union run on STRUCTURE. The SNPs are listed in Table 1 and the 55 populations are listed in Table 2. Likelihoods in all cases were beginning to plateau at about K=8 but analyses were conducted up to K=12 with no additional regional clarification past K=9, just increased individual differences within populations. Results plotted are the most likely of 10 independent runs of the program.

Figure 2.

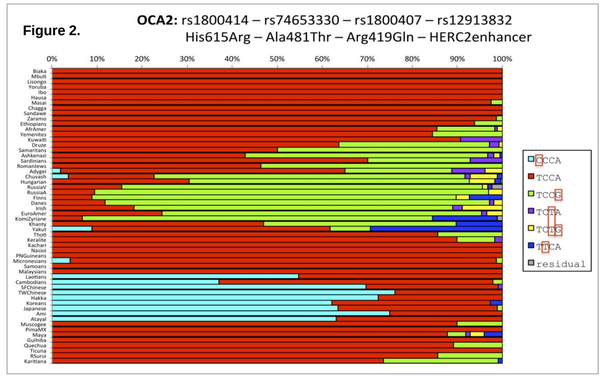


Figure 2. Bar graph of OCA2 haplotype frequencies based on four functional SNPs. rs12913832 is the enhancer SNP most relevant for blue eye color.

Lineage Informative SNP (LISNP) Studies

Markers that are especially good at helping to identify the relatives of an unknown DNA profile are generally those that are highly polymorphic while having a low enough mutation rate that identity by state (IBS) generally implies identity by descent (IBD). The standard forensic short tandem repeat polymorphisms (STRPs) meet the first of those requirements but do not always meet the second. While single nucleotide polymorphisms (SNPs) have very low mutation rates, they are di-allelic and hence heterozygosity is a maximum of 0.5. We have been working to identify haplotypes of 3 or 4 SNPs that involve only modest linkage disequilibrium and span a small molecular extent, less than 10kb. We have proposed these mini-haplotypes ("minihaps") as the LISNP class of SNP-based markers (Pakstis et al., 2012). Because minihaps have very low mutation and recombination rates, observing IBS gives a high probability of IBD, essential for identifying lineages/clans/families. We present an overview of 35 candidate minihaps, distributed across 17 autosomes, that we have studied on 54 population samples. In Figure 3 the Fst, average heterozygosity, and the percentage resolvability are plotted for each of the minihaps. Resolvability is the ability to unambiguously determine haplotypes because an individual is either homozygous for all the SNPs or else has at most one heterozygous SNP. The additive phylogenetic tree in Figure 4 is based on pairwise genetic distances for 34 of these minihaps and demonstrates their potential utility for ancestry inference. The bar graphs illustrate the allele frequencies (Figures 5) of the four minihap systems. NPAS2 and DRD3 are examples from the extremes of the percent resolvability distribution; while TAS2R16 has the highest average heterozygosity on 54 populations of the 35 minihaps studied so far and FADS2 has the lowest average heterozygosity.

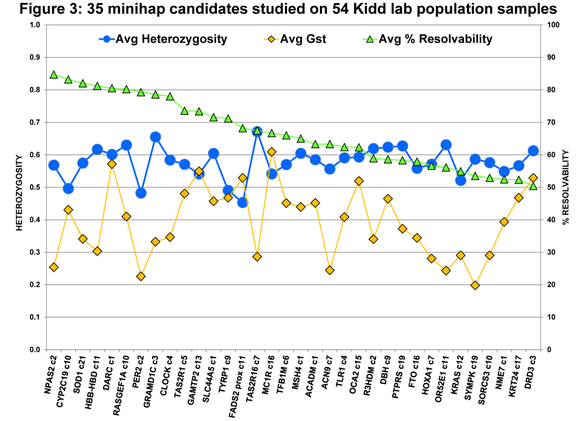
Phenotype Informative SNP (PISNP) Studies

Our population resources (Table 2) are valuable for gaining a better understanding of how SNPs associated with phenotypes vary among populations. In a recent study (Donnelly et al., 2012) we were able to show that many of the SNPs in OCA2 that had been associated with eye color in Western Europe were in fact simply in linkage disequilibrium with the SNP in the upstream enhancer that Visser et al. (2012) showed to be the functionally relevant variation. For several of those non-coding SNPs in OCA2 the "blue eye" allele also occurred in other parts of the world at frequencies that would have predicted a noticeable frequency of blue eyes if the allele caused blue eyes when the population has essentially no blue-eyed individuals.

We have now studied four of the OCA2 SNPs with clear functional possibilities that have an appreciable heterozygosity in any existing data. Figure 2 shows the haplotype frequencies of those four SNPs. The haplotype defined by the enhancer variant is clearly present only in populations in or near Europe with the possibility of European admixture in some other population samples. This is the haplotype actually functionally relevant to the "blue eye color" phenotype. The haplotype defined by the SNP rs1800414 (His615Arg) is restricted to East Asia with some evidence of gene flow into Southeastern Europe. This SNP and haplotype are associated with lighter pigmentation among East Asians (Edwards et al., 2010). What our new analyses show is that another non-synonymous substitution, rs74653330 (Ala481Thr) defines another haplotype reaching 10% to 30% in Siberia and the Finns; as yet there is no phenotype association known to be associated with this haplotype. Finally, we note that rs1800407 (Arg419Gln) occurs on haplotypes with the enhancer variant AND occurs on chromosomes with the ancestral allele at the enhancer site. Though uncommon, the cis-trans possibility may raise complexities in eye color prediction that is not currently considered in the formulae used to predict eye color for the IrisPlex SNP set.

Multi-population, multi-SNP (haplotype) studies are ongoing at several other loci known to be involved in phenotype. Primarily, these are skin color loci: SLC24A5, SLC45A2, MC1R, etc. As we obtain reasonable results, we will be making the data available in ALFRED—ALlele FRequency Database.

Figure 3: 35 minihap candidates studied on 54 Kidd lab population samples



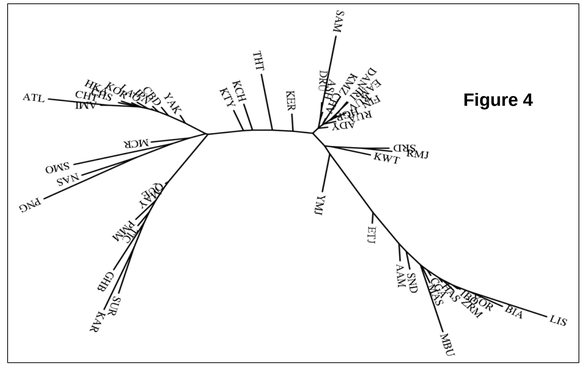
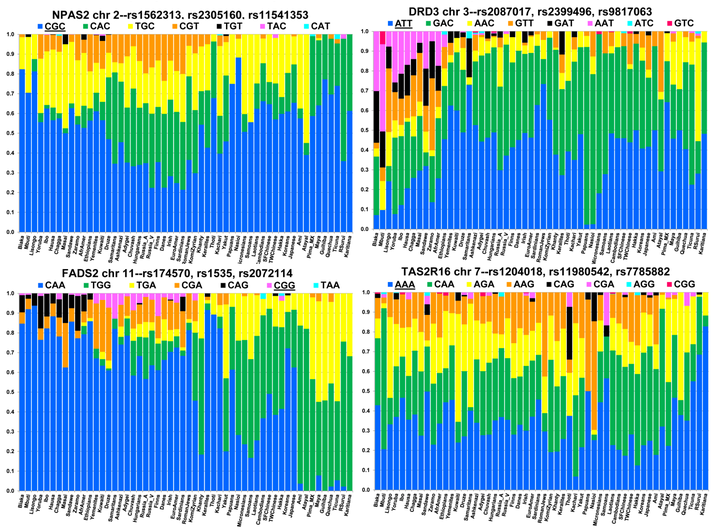


Figure 4

Figure 4. Best additive phylogenetic tree based on 34 minihaps and 54 of the 55 populations listed in Table 2. Malaysians were not included in tree analysis.

Figure 5. Examples of haplotype frequencies for four minihaps. Colored bars are proportional to frequencies. Underscored haplotypes are ancestral.



ACKNOWLEDGMENTS

This work was funded primarily by NIJ Grants 2010-DN-BX-K225 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 document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice. We thank Applied Biosystems for making their allele frequency database available to us and for supplying some of the TaqMan reagents that were employed in these studies. 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. ALFRED is supported by NSF grant BCS0938633.

REFERENCES

Donnelly M.P., P. Paschou, E. Grigorenko, D. Gurwitz, C. Barta, R.-B. Lu, O.V. Zhukova, J.-J. Kim, M. Siniscalco, M. New, H. Li, S.L. Kajuna, V.G. Manolopolulos, W.C. Speed, A.J. Pakstis, J.R. Kidd, K.K. Kidd, 2012. A global view of the OCA2-HERC2 region and pigmentation. *Human Genetics* 131:683-696.

Edwards M., A. Bigham, J. Tan, S. Li, A. Gozdzik, K. Ross, Li Jin, E.J. Parra, 2010. Association of the OCA2 polymorphism His615Arg with melanin content in East Asian populations: Further evidence of convergent evolution of skin pigmentation. *PLoS Genetics* 6:e1000867.

Enoch MA, Shen PH, Xu K, Hodgkinson C, Goldman D., 2006. Using ancestry informative markers to define populations and detect population stratification. *Journal of Psychopharmacology* 20:19-26.

Kidd, J.R., F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, K.K. Kidd, 2011. Analyses of a set of 128 ancestry informative SNPs (AISNPs) in a global set of 119 population samples. *Investigative Genetics* 2:1 (epub January 5, 2011) In press Oct2010.

Kosoy, R., R. Nassir, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, et al., 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America, Human Mutation 30:69-78.

Lao O., Vallone P.M., Coble M.D., Diegoli T.M., van Oven M., van der Gaag K.J., Pijpe J., de Knijff P., Kayser M., 2010. Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA. *Human Mutation* 31:E1875-93.

Pakstis A.J., R. Fang, M.R. Furtado, J.R. Kidd, K.K. Kidd, 2012. Mini-haplotypes as lineage informative SNPs (LISNPs) and ancestry inference SNPs (AISNPs). *European Journal of Human Genetics*. In Press.

Phillips C., Salas A., Sánchez J.J., Fondevila M., Gómez-Tato A., Álvarez-Dios J., Calaza M., Casares de Cal M. , Ballard D., Lareu M.V., Carracedo A., 2007. The SNPforID Consortium. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International: Genetics* 1:273-280.

Yaeger R., A. Avila-Bront, K. Abdul, P.C. Nolan, V.R. Grann, M.G. Birchette, S. Choudhry, E.G. Burchard, K.B. Beckman, P. Gorroochurn, E. Ziv, N.S. Consedine, A.K. Joe, 2008. Comparing Genetic Ancestry and Self-Described Race in African Americans Born in the United States and in Africa. *Cancer Epidemiol Biomarkers Prevention* 17:1329-1338.

Visser M., M. Kayser, R.-J. Palstra, 2012. HERC2 rs12913832 modulates human pigmentation by attentuating chromain-loop formation between a long-range enhancer and the OCA2 promoter. Genome Research 22:446-455.