Developing SNP panels for ancestry identification useful in forensic investigations

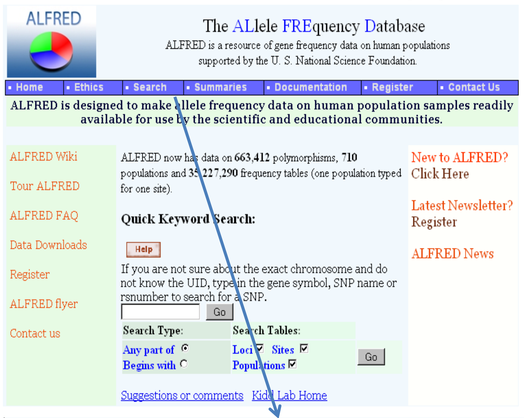
Kenneth K. Kidd, Judith R. Kidd, Andrew J. Pakstis, William C. Speed, Michael P. Donnelly

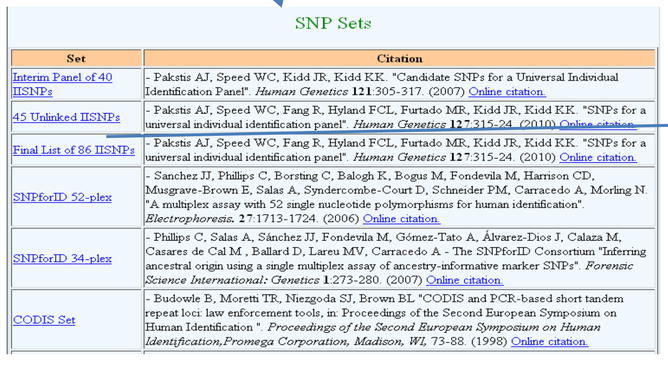
Department of Genetics, Yale University School of Medicine, New Haven CT 06520

ABSTRACT

Given a large enough number of SNPs and other markers it is certainly possible to distinguish on average the ancestry of closely related populations and individuals within the same population. However, routine forensic investigations have limited time and resources. Our specific research aim is to identify ancestry for single individuals with a limited number of SNPs. Virtually all of the ancestry informative SNPs (AISNPs) published in the scientific literature are designed for much more limited tasks because the data are based on a very limited number of populations studied and in general do not allow a distinction of ancestry beyond origin in major continental regions of the world. Our recent research has begun to identify SNP sets that can begin to identify ancestry within some continental regions. Whether or not we can achieve a large likelihood ratio for distinguishing ancestry within a geographic region like Europe or the native populations of North America with only a few hundred SNP markers remains to be demonstrated--this is a research project. We will present an overview of our recent work collecting existing and identifying new AISNP sets and validating their utility in an increasing number of population samples. We are also collaborating on the development of statistical methodologies to optimize the number of SNPs required for various comparisons. We have a growing set of prospective AISNPs (currently numbering around 1,177 SNPs) identified in a variety of ways. Subsets of this superset have already been studied and published on increasingly large numbers of population samples. For example, we have studied the 128 AISNPs identified by the Seldin group on 119 populations including 73 population samples that we have typed. We have also analyzed another set of 40 SNPs (developed by the Nievergelt group) on 57 populations. Our presentation will provide an overview of our project thus far, how we are already sharing intermediate results via the ALFRED database (http://ALFRED.med.yale.edu) and various publications, as well as what next steps will be undertaken in the near future.

"SNP sets" Available on "Search tab" of ALFRED homepage





Testing published AISNP panels on more populations in search of the best SNPs

Several SNP panels for ancestry identification have been published. Most have documented differential allele frequencies for a relatively small number of populations representing major continental geographical regions but cannot predict their ability to predict ancestry for untested populations. We have studied some of the better SNP panels on a much more extensive roster of population samples from around the world in search of the best subset of SNPs that can distinguish populations within and across the major geographical areas. (ALFRED has a convenient feature for identifying various published SNP lists.) An illustrative group of 39 Pilot AISNPs (Test Panel Table) with above average Fst values are highlighted here. In the analyses reported here for 39 Pilot AISNPs we have studied roughly comparable numbers of individuals in the five relevant geographic regions—Sub-Saharan Africa, SW Asia & Southern Europe, Northern & Western Europe, East Asia, and the Americas. Note in the STRUCTURE figure for 43 populations, the results for K=6 look virtually identical to the results for 128 markers in 119 populations except that South and Central Asia have fewer populations included for the 39 Pilot AISNPs and do not so obviously share a color with Pacific populations.

The 39 Pilot AISNPs were selected from a larger dataset derived from six main sources of AISNP candidates from the literature and unpublished sources:

(1)Seldin and colleagues have published a number of reports on AISNPs (Collins-Schramm et al., 2004; Yang et al., 2005; Tian et al., 2006; 2007; Kosoy et al., 2009; Nassir et al., 2009). We have also evaluated the 128 SNPs, which the Seldin group identified, on 119 populations so far (Kidd et al., 2011; in *Investigative Genetics*).

(2) A set of 40 AISNPs identified by Caroline Nievergelt at UCSD from HGDP data of Li et al. (2008). We have studied them on 47 of our population samples and presented an analysis focused on Native American populations (Kidd et al., 2011; *Am J Phys Anthro*).

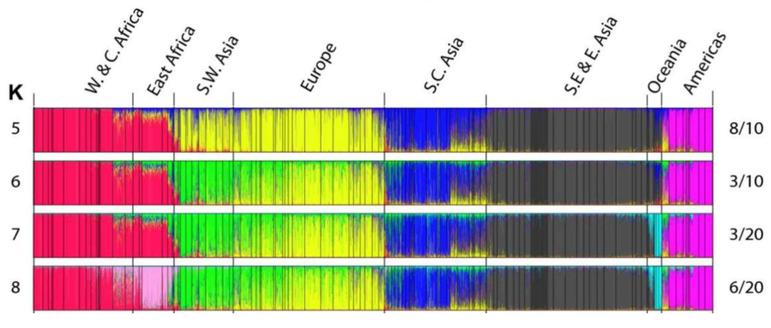
(3) The AISNPs of the Goldman group (Enoch et al., 2006) have been analyzed for their AI value on HGDP plus HAPMAP populations and a promising subset are being typed on 47+ populations at Kidd Lab.

(4) The SNPs reported by Lao et al. (2006, 2008) are being typed on Kidd Lab populations.

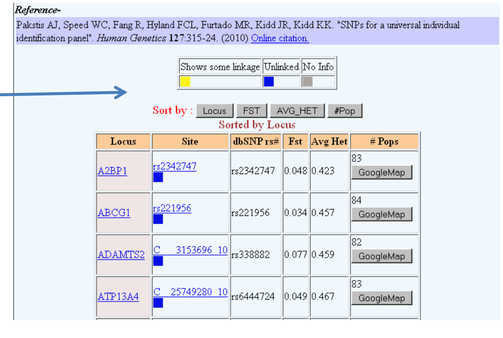
(5) The 34-plex AIMs developed by the SNPforID consortium (Phillips et al., 2007) have been checked for their variation patterns across HGDP & HAPMAP populations. Selected markers are being tested on Kidd Lab samples.

(6) SNPs originally selected for study on non-forensic projects at Kidd Lab have also been screened (global and continental Fst) for their utility as AISNPs. A number have been found with strong, distinctive geographic patterns (e.g., at ADH1B, DRD2).

STRUCTURE: 119 populations, 128 AISNPs (Kidd et al., 2011, Investigative Genetics)



Summary Information for sites in selected SNP Set



Announcing

Forensic Reference On Genetics knowledge base

http://frog.med.yale.edu

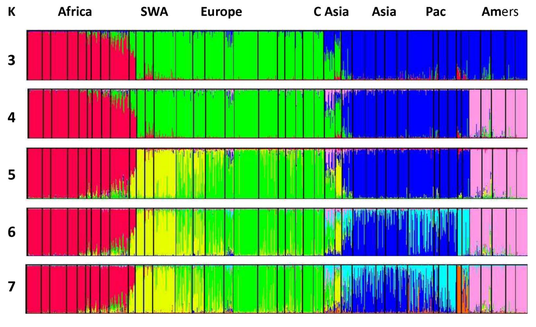
As of June 20, 2011, a pilot version of a new web site/database has gone online: Forensic Reference/Resource on Genetics knowledge base.

To date the FROG website has only a pilot (pre-beta) version of a function to calculate probabilities of genotypes by population for certain IISNP and AISNP panels. This is available to elicit feedback from the forensic community on this function while other aspects are being developed. In the future this database will be linked to the data in ALFRED, allowing access to all the data for diverse published sets of SNPs.

A provisional/pilot panel of AISNPs has been assembled from our set of over 4000 SNPs studied on 45 to 57 of our populations. These include the "better" AISNPs from several different published sources. It is based on very preliminary analyses specifically to provide data for this pilot version of FROGkb

Pilot functionality of FROG*kb* allows input of the genotypes for an unknown individual for all or a subset of either the 45 IISNP panel or the 39 Pilot AISNP panel. The output is a numeric and graphic depiction of the likelihoods of the input multisite genotype originating in each of the many populations already entered. For the IISNP panel, this is essentially the match probability. For the AISNP panel one obtains the likelihood of origin in each population. The likelihoods can then be evaluated using the relative values.

STRUCTURE analysis: 43 populations, 39 Pilot AISNPs, 2278 individuals



Test Panel of 39 Pilot AISNPs for Evaluating FROGkb

dbSNP rs#	LOCUS	CHR	HGDP	HAPMAP
rs2814778	DARC	1	no	no
rs3737576	S1PR1	1	yes	yes
rs7554936	SEMA6C	1	yes	yes
rs10497191	ACVR1	2	yes	no
rs6754311	DARS	2	no	yes
rs260690	EDAR	2	yes	yes
rs3827760	EDAR	2	no	yes
rs798443	LOC339788	2	yes	yes
rs1834619	SMC6	2	yes	yes
rs12498138	GOLGB1	3	yes	yes
rs1229984	ADH1B	4	no	yes
rs7657799	CXXC4	4	yes	yes
rs870347	PAPD7	5	yes	yes
rs16891982	SLC45A2	5	no	yes
rs192655	MDN1	6	yes	yes
rs917115	JAZF1	7	yes	yes
rs4918664	CYP26A1	10	yes	yes
rs1079597	DRD2	11	no	yes
rs174570	FADS2	11	yes	yes
rs2238151	ALDH2	12	no	yes
rs9522149	ARHGEF7	13	yes	yes
rs7997709	NBEA	13	yes	yes
rs1572018	KBTBD6	13	yes	yes
rs200354	RPL3P4	14	yes	yes
rs12439433	D15S118	15	yes	yes
rs12913832	HERC2	15	yes	yes
rs1426654	SLC24A5	15	no	yes
rs1834640	SLC24A5	15	yes	yes
rs735480	TRIM69	15	yes	yes
rs17642714	ABCC3	17	no	yes
rs11652805	AMZ2P1	17	yes	yes
rs4411548	ATP6V0A1	17	no	no
rs2593595	G6PC	17	yes	yes
rs4471745	TMEM100	17	yes	yes
rs7226659	RIT2	18	yes	yes
rs4891825	RTTN	18	yes	yes
rs7251928	ZBTB7A	19	yes	yes
rs310644	PTK6	20	yes	yes
rs2024566	ZC3H7B	22	yes	yes

NOTE: all SNP Frequencies are in ALFRED

ACKNOWLEDGEMENTS

This work was funded primarily by NIJ Grants 2010-DN-BX-K225, 2010-DN-BX-K226, 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 U.S. 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. ALFRED is supported by NSF grant BCS0938633.

REFERENCES

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 Oct 20, 2010.

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

Li, J.Z., D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, et al., Worldwide human relationships inferred from genome-wide patterns of variation, Science 319:1100-1104 (2008).

Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, et al., An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genetics Jul 24;10:39 (2009).

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