THE SEARCH FOR BETTER MARKERS FOR FORENSIC ANCESTRY INFERENCE

Kenneth K. Kidd*, William C. Speed, Andrew J. Pakstis, Judith R. Kidd
Department of Genetics, Yale University School of Medicine, PO Box 208005, New Haven, CT 06520-8005. *Corresponding author: 203-785-2654, kenneth.kidd@yale.edu

INTRODUCTION

In 2007 we presented definitions of four types of panels of autosomal SNPs that would be useful for different forensic problems/questions (1). These four types are

"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 diallelic 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."

In that same poster we presented our progress on a panel of IISNPs; our final panel was published two years later (2). We are now working on identifying panels of AISNPs, PISNPs, and LISNPs. (We note that an example of a LISNP locus, GRAMD1C with 3 SNPs haplotyped to form five common alleles (haplotypes), was included in that poster (1).) Very large numbers of SNPs, 10's to 100's of thousands, can do a very good job at major ancestral distinctions among ancestries from populations originating from around the world.(3, 4, 5). However, for routine forensic applications a smaller number of SNPs that can be tested quickly and cheaply is preferable.

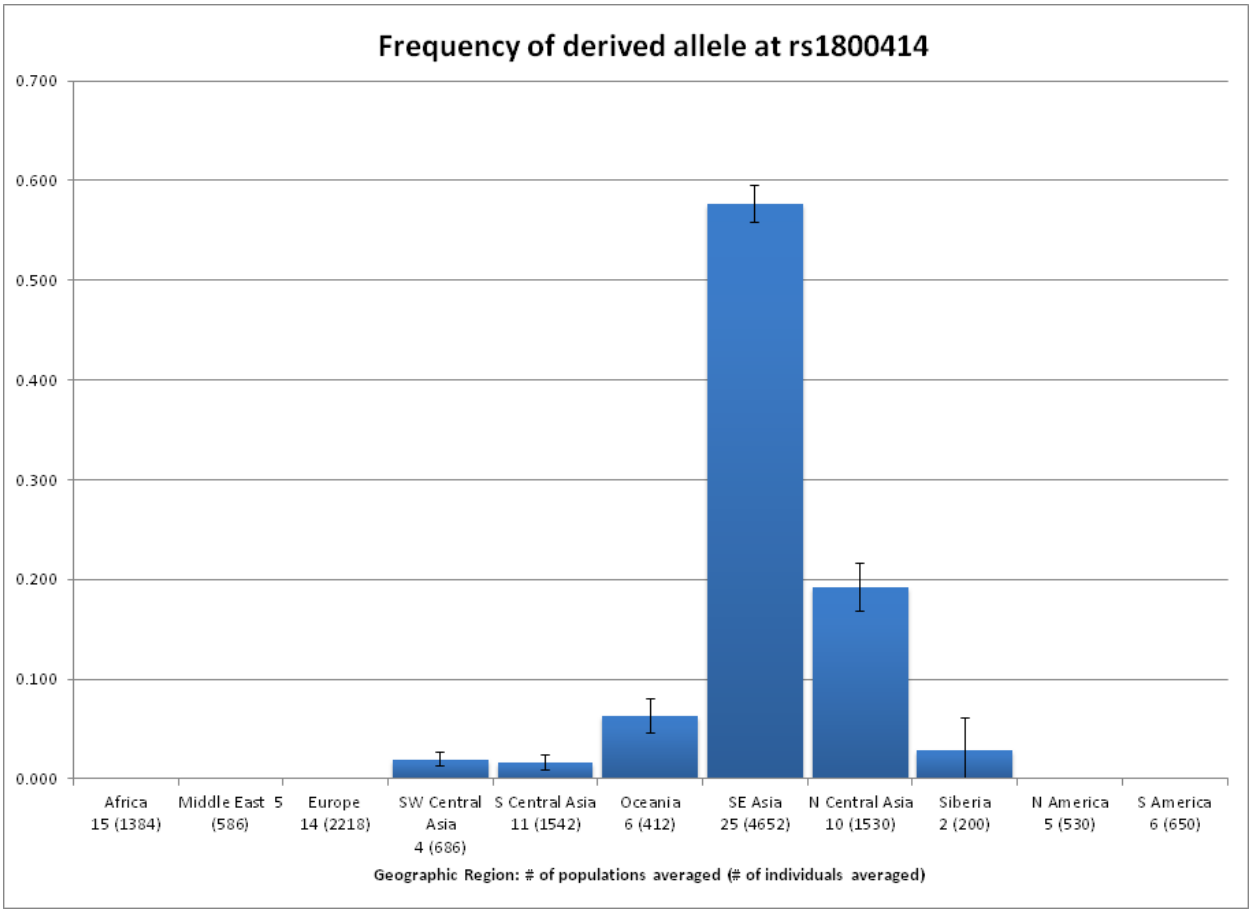
Many panels of AISNPs have been published over the past decade and all provide some resolution of ancestry to "continental" level (reviewed in 6). However, none of those panels is likely to be perfect for fine resolution of ancestry in all areas of the world because they differ in the populations used to identify the SNPs and, consequently, in the resulting SNPs. Since some studies have used a relatively small number of populations, the ability of those panels to extrapolate to broader inference involving other ancestries is questionable. For example, a small panel optimized for distinctions in Europe is likely to have little value in distinguishing various East Asian groups. Some ways of reporting the inference of ancestry also seem to be inappropriate. While triangle and/or tetrahedron plots with "racial" or "continental" vertices can be useful graphics in research, global variation in humans does not fit "racial' categories and models inherent in such plots. Ultimately, to be useful for forensics an AISNP panel will need greater specificity than is provided by "continental" assignment and very thorough documentation of the allele frequencies in populations globally. We are now working to identify such a robust panel of AISNPs.

PROGRESS

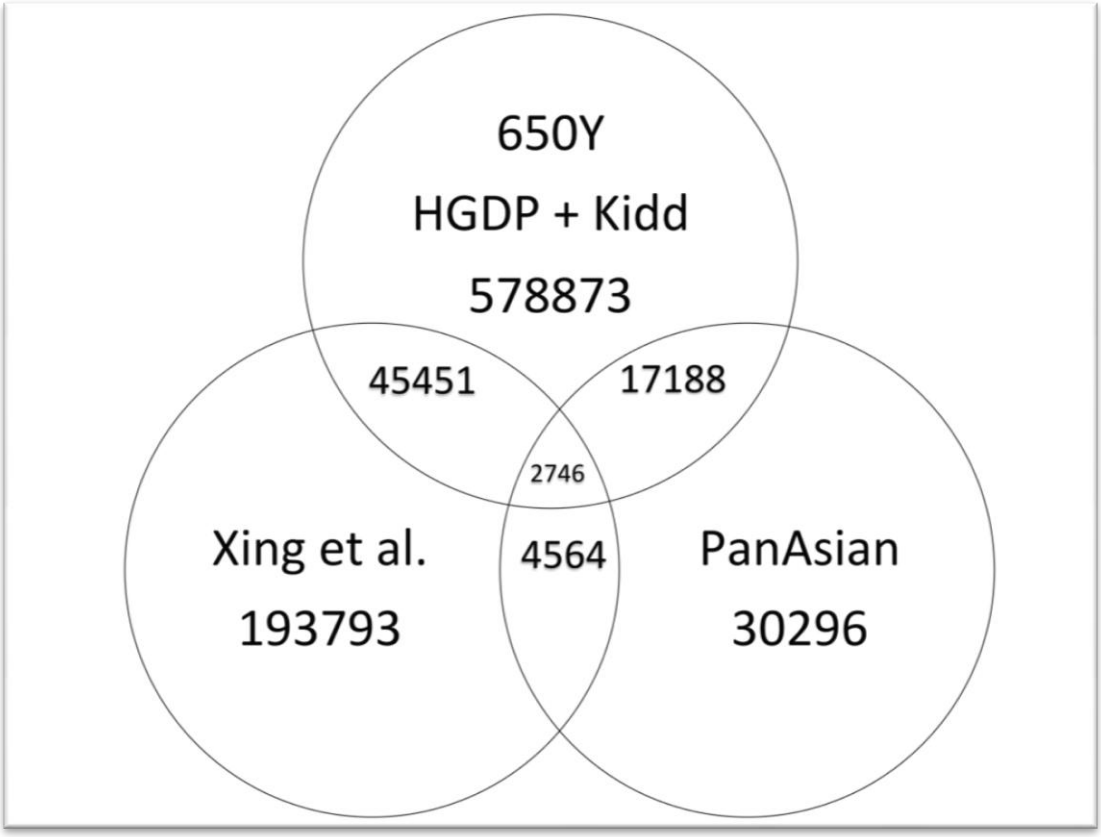
In our search for a better panel of AISNPs we have primarily pursued testing candidate AISNPs on our resource of >2500 individuals from >47 populations. Candidate SNPs that have shown large allele frequency differences among populations have been identified in data on the HapMap and HGDP populations and in other studies including our own. To date the largest test of ancestry informative markers was a study of 4,888 individuals from 119 different populations (7) using the 128 SNPs identified by the Seldin group (8) in the HGDP data (3,4). An analysis of all samples allows a clear grouping into 8

distinct ancestries. When 40 additional AISNP candidates identified by Caroline Nievergelt were added and analyses of Native American populations undertaken, most individual tribal groups could be clearly distinguished (9). We are continuing to work on improving these panels.

One way of improving the panel is to consider PISNPs that are also very informative in ancestry. Some of the SNPs involved in skin and eye color are excellent candidates, but to date most seem to distinguish Europeans from indigenous peoples in all other regions of the world and even subdivide Europe. Considerable work on some of these genes is ongoing (10) and we have extended the data on SNPs in the OCA2-HERC2 region to over 100 populations to determine their geographic distributions more definitively (11). In addition to the improved documentation for the SNPs most strongly associated with eye color in population frequencies around the world, that study also provided new information on a SNP associated with skin color in East Asia, rs1800414 (Figure 1).



One of the issues inhibiting identification of excellent AISNPs based on larger numbers of individuals and a set of populations with a more uniform geographic distribution is the lack of overlap in modern population studies. For example, we tested the Illumina 650Y panel (no longer manufactured) on 1300 individuals in our collection and have combined the data with the HGDP data for 578,873 autosomal SNPs. The Jorde lab (5) published a study of many populations for 193,793 autosomal SNPs and the Pan Asian consortium (12) published data on many Asian populations for 30,296 autosomal SNPs. Very few of the populations in these studies are the same, and even then the individuals are different. The combination of all these populations would provide an excellent global dataset. However, only 2,746 SNPs have been studied in common in all three studies (Figure 2), a paltry fraction of the total studied. While our ongoing studies may identify excellent AISNPs in this set, none of our existing very good AISNPs is included among these 2,746.



As noted earlier, for forensic panels of SNPs to be useful, the underlying data must be thoroughly documented in a public fashion. We are also working on those aspects through ALFRED, our ALlele FREquency Database http://alfred.med.yale.edu (13). Currently, ALFRED has over 35 million allele frequency tables that are linked to their molecular definitions and the specific population studied. ALFRED is supported by the NSF to make these data freely accessible to the scientific and educational communities.

CONCLUSION

We believe that no relatively small (<200) number of SNPs yet exists to constitute a panel of AISNPs with optimal characteristics. However, even small panels of ~40 AISNPs can easily identify ancestry from up to six or seven broad geographic regions. Based on our ongoing work we are convinced that one or more greatly improved panels of AISNPs will be available soon. However, ultimate ideal documentation will require cooperation among laboratories with different sets of population resources to have tested the same set of candidate SNPs.

In closing, it is important to note that the short tandem repeat polymorphisms in the CODIS and other forensic panels provide virtually no information on ancestry or phenotype, both very important in forensics. SNPs or other diallelic markers, such as indels, will be required. And, routine forensic use of well documented panels of AISNPs and PISNPs will require that forensic labs have the equipment, the protocols for SNPs, and trained technicians. Forensic laboratories need to be thinking now about the coming capabilities offered by current research and newer technologies.

ACKNOWLEDGEMENTS

This work was funded primarily by Grants 2007-DN-BX-K197 and 2010-DN-BX-K225 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. ALFRED is supported by the US National Science Foundation grant BCS0938633 to KKK as scientific infrastructure.

REFERENCES

1. Pakstis AJ, Speed WC, Kidd JR, Kidd KK, 2007. SNPs for Individual Identification, http://medicine.yale.edu/labs/kidd/www/ISFGposter2007.pdf. Definitions were then published in Butler JM, Budowle B, Gill P, Kidd KK, Phillips C, Schneider PM, Vallone PM, Morling N, 2008. Report on ISFG SNP Panel Discussion. *Forensic Science International: Genetics* Supplement Series 1:471-472.

2 Pakstis AJ, Speed WC, Fang R,. Hyland F.C.L,. Furtado M.R,. Kidd J.R,. Kidd K.K. **2010**. SNPs for a universal individual identification panel. *Human Genetics* **127**:315-324.

3. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100-1104.

4. Pemberton TJ, Jakobsson M, Conrad DF, Coop G, Wall JD, Pritchard JK, Patel PI, Rosenberg NA, 2008. Using population mixtures to optimize the utility of genomic databases: linkage disequilibrium and association study design in India. *Ann Hum Genet.* 72:535-46.

5. Xing J, Watkins WS, Shlien A, Walker E, Huff CD, Witherspoon DJ, Zhang Y, Simonson TS, Weiss RB, Schiffman JD, Malkin D, Woodward SR, Jorde LB, 2010. Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. *Genomics* 96:199-210.

6. Kayser M, de Knijff P, 2011. Improving human forensics through advances in genetics, genomics and molecular biology. *Nat Rev Genet* 12:179-92.

7. Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, de La Vega FM, Kidd KK. 2011. Analyses of a set of 128 ancestry informative SNPs (AISNPs) in a global set of 119 population samples. *Investigative Genetics* 2:1

8. Kosoy R, Nassir R, Tian C, White PA, Butler LLM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, and Seldin MF. 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Human Mutation* 30: 69-78.

9. Kidd JR, Friedlaender F, Pakstis AJ, Furtado M, Fang R, Wang X, Nievergelt CM, Kidd KK. 2011. Single nucleotide polymorphisms and haplotypes in Native American populations. *Am J Phys Anthropology*, in press for December, 2011.

10. Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M, Ballantyne KN, 2011. Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence. *Forensic Science International: Genetics* 5:464-471.

11. Donnelly MP, Barta C, Grigorenko E, Kim J-J, Li H, Lu R-B, Manolopolulos VG, New M, Paschou P, Siniscalco M, Zhukova OV, Speed WC, Pakstis AJ, Kidd JR, Kidd KK, 2011. A global view of the OCA2-HERC2 region and pigmentation. *Human Genetics .*

12. The HUGO Pan-Asian SNP Consortium, 2009. Mapping Human Genetics Diversity in Asia. *Science* 326:1541-1545.

13. Rajeevan H, Soundararajan U, Pakstis AJ, Kidd JR, Kidd KK, 2011. ALFRED: an allele frequency resource for research and teaching. *Nucleic Acids Research,* online October 28 pp1-6. Doi:10.1093/nar/gkr924.

FIGURE CAPTIONS

Figure 1. Allele frequencies of the derived OCA2 allele (615Arg) at the SNP rs1800414 for geographic regions using data available in ALFRED. The figure summarizes 103 populations and a total of over 14,000 individuals. The numbers of populations (individuals) typed are given for each region.

Figure 2. A Venn diagram showing the overlap in SNPs typed in the Kidd Lab and HDGP populations with two other studies (5,12).