Forensic Science International: Genetics 19 (2015) 269–271

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Short communication

52 additional reference population samples for the 55 AISNP panel

Andrew J. Pakstis [a,*], Eva Haigh [a], Lotfi Cherni [b,c], Amel Ben Ammar ElGaaied [b], Alison Barton [a], Baigalmaa Evsanaa [d,e], Ariunaa Togtokh [e], Jane Brissenden [d], Janet Roscoe [d,f], Ozlem Bulbul [a,g], Gonul Filoglu [g], Cemal Gurkan [h], Kelly A. Meiklejohn [i], James M. Robertson [j], Cai-Xia Li [k], Yi-Liang Wei [k,l], Hui Li [m], Usha Soundararajan [a], Haseena Rajeevan [a], Judith R. Kidd [a], Kenneth K. Kidd [a]

[a] Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA
[b] Laboratory of Molecular Genetics, Immunology and Human Pathology, Science Faculty of Tunis, University El Manar, 2092 Tunis, Tunisia
[c] High Institute of Biotechnology, University of Monastir, 5000 Monastir, Tunisia
[d] Department of Medicine, University of Toronto, Toronto, ON M5S, Canada
[e] Department of Nephrology, Health Sciences University of Mongolia, Khoroo 1, Ulaanbataar, Mongolia
[f] Department of Medicine, The Scarborough Hospital, Toronto, ON M1P 2V5, Canada
[g] Institute of Forensic Science, Istanbul University, Istanbul, Turkey
[h] Turkish Cypriot DNA Laboratory, Committee on Missing Persons in Cyprus Turkish Cypriot Member Office, Nicosia, North Cyprus, Cyprus
[i] Counterterrorism and Forensic Science Research Unit, Visiting Scientist Program, Federal Bureau of Investigation Laboratory Division, 2501 Investigation Parkway, Quantico, VA 22135, USA
[j] Counterterrorism and Forensic Science Research Unit, Federal Bureau of Investigation Laboratory Division, 2501 Investigation Parkway, Quantico, VA 22135, USA
[k] Research Centre of Applied Forensic Science Technologies, Institute of Forensic Science, Ministry of Public Security, Beijing 100038, PR China
[l] Department of Immunology, Biochemistry and Molecular Biology, 2011 Collaborative Innovation Center of Tianjin for Medical Epigenetics, Tianjin Medical University, Tianjin 300070, PR China
[m] MOE State Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, PR China

ARTICLE INFO

Article history:
Received 5 June 2015
Received in revised form 16 July 2015
Accepted 7 August 2015
Available online 14 August 2015

Keywords:
SNP
Ancestry
Reference database
FROG-kb
ALFRED

ABSTRACT

Ancestry inference for a person using a panel of SNPs depends on the variation of frequencies of those SNPs around the world *and* the amount of reference data available for calculation/comparison. The Kidd Lab panel of 55 AISNPs has been incorporated in commercial kits by both Life Technologies and Illumina for massively parallel sequencing. Therefore, a larger set of reference populations will be useful for researchers using those kits. We have added reference population allele frequencies for 52 population samples to the 73 previously entered so that there are now allele frequencies publicly available in ALFRED and FROG-kb for a total of 125 population samples.

1. Introduction

In 2014 we published a panel of 55 ancestry informative single nucleotide polymorphisms (AISNPs) and showed that seven to eight biogeographic regions could be distinguished using these markers on 3884 individuals from 73 populations [1]. The data on those populations for these 55 AISNPs are available in the ALlele FREquency Database ALFRED <http://alfred.med.yale.edu> [2] and for estimating ancestry using the Forensic Reference Resource on Genetics Knowledge Base (FROG-kb) <http://frog.med.yale.edu> [3]. How these SNPs help reveal ancestry was demonstrated [1] in principal components analysis (PCA) and STRUCTURE [4] analyses. We note that this panel of 55 AISNPs is now implemented for massively parallel sequencing (MPS) in the sequencing products offered by Illumina and by Life Technologies. Since there are now commercial kits using these 55 SNPs for ancestry inference, we have now added the allele frequencies for 52 more population samples for these 55 SNPs to ALFRED and FROG-kb,

* Corresponding author at: Department of Genetics, Yale University School of Medicine, 333 Cedar Street, SHM I-348, P.O. Box 208005, New Haven, CT 06520-8005, USA.
E-mail address: Andrew.Pakstis@yale.edu (A.J. Pakstis).

http://dx.doi.org/10.1016/j.fsigen.2015.08.003

making a more comprehensive reference database available for forensic inferences. The total dataset now includes data for 11 more of the 1000 Genomes populations (Phase 3) for a total of 22.

2. Materials and methods

For most new population samples studied collaborators sent DNA samples to Yale and the genotyping was done at Yale using the standard TaqMan assay system used for the original study [1]. In two cases the SNP genotypings were carried out in labs in China. Dr. Cai-Xia Li's group in China employed the custom Golden Gate genotyping assay procedure from Illumina, Inc.; Dr. Hui Li used the same TaqMan assays and protocols used at Yale. Supplementary Table S1 lists the 125 different population samples (6853 individuals) representing the diverse ethnic groups and biogeographic regions that have now been analyzed for the 55 AISNPs. The populations in the table are organized by geographic region. The table also includes the number of individuals from each group, the ALFRED unique identifier (UID) for looking up the description of each sample, and the names of the collaborating co-authors who provided SNP genotypes and/or collected the samples.

Supplementary Table S1 related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2015.08.003.

Table 1
Top 30 likelihoods calculated by FROG-kb [3] for 55 AISNP set from the 125 current reference populations for a Lybian individual.

Populations[a]	Probability of genotype in each population	Likelihood ratio
▶Palestinian Arabs	1.20E−13	1.0
Sousse, Tunisia	5.80E−14	2.1
Turkish Cypriots	5.70E−14	2.2
Mehdia, Tunisia	5.10E−14	2.4
*Lybians, Libya	5.00E−14	2.4
Nebeur, Tunisia	3.90E−14	3.1
Kairoun, Tunisia	2.30E−14	5.5
▶Kuwaiti	1.90E−14	6.5
Smar, Tunisia	1.10E−14	11
Kerkennah, Tunisia	7.70E−15	16
▶Druze	6.00E−15	21
▶Sardinians	4.90E−15	25
Turkish	3.80E−15	33
Kesra, Tunisia	3.70E−15	33
Tajiks	2.70E−15	45
▶Adygei	2.30E−15	53
Iranians	2.30E−15	53
▶Greeks	1.90E−15	66
▶Iberian (IBS)	1.30E−15	94
▶Ashkenazi	1.10E−15	110
▶Negroid Makrani	1.10E−15	110
▶Mohanna	4.80E−16	260
▶Roman Jews	3.80E−16	330
▶Hungarians	2.20E−16	560
▶Chuvash	1.80E−16	690
▶Russians, Vologda	1.20E−16	1.10E+03
▶Pathans	9.60E−17	1.30E+03
▶Toscani (TSI)	5.00E−17	2.50E+03
▶Yemenite Jews	2.20E−17	5.60E+03
Gujarati (GIH)	2.00E−17	6.10E+03

Results from FROG-kb [3] for a Libyan individual's 55 AISNP genotypes. The population sample from which the individual was taken is indicated by the asterisk; population samples previously used for FROG-kb calculations indicated with ▶ as in table. The likelihood ratio is calculated as the probability of the best population divided by the probability of the specified population. Only those populations with a likelihood ratio greater than 100 can be significantly eliminated as a population of origin but the new populations clearly give results favoring North Africa as opposed to Southwest Asia (Palestinian Arabs). Ethiopians, Somali, African Americans, and all Sub-Saharan populations tested had likelihood ratios from 10^{6} to 10^{55}, clearly excluding all other populations from the continent of Africa.

[a] Where symbol ▶ precedes a population name, the population was one of the 73 included in a previous publication—Kidd et al. [1]. The * precedes the population name of the group to which the Lybian individual belongs whose genotypes were employed in the FROG-kb calculation.

Genotypes were examined to ensure that the alleles called are on the positive strand and corrected if they were not. Allele frequencies were examined for possible misidentification of a locus and every locus was tested for Hardy-Weinberg ratios in every population sample with no significant deviations.

3. Results

Data from other laboratories are consistent with the TaqMan genotyping done at Yale in that alleles and frequencies agree with what could be expected for the geographic region and other population data in the general region. There are no significant deviations from Hardy-Weinberg ratios.

Allele frequencies for all 55 SNPs in all 125 population samples are accessible in ALFRED. In many cases additional populations have been studied for some of the SNPs and thus those SNPs have data on more than 125 population samples in ALFRED. In FROG-kb the "Kidd Lab – Set of 55 AISNPs" has complete data for all 125 reference population samples. The completeness of the data allows likelihoods and likelihood ratios to be calculated for all of these population samples for any input DNA profile for the 55 AISNPs (or a subset).

As an example of the added information provided by the new populations, we have summarized the results from FROG-kb analyses of an individual from the new population sample of Libyans (Table 1). We have listed the top 30 populations by the probability calculated by FROG-kb [3] of this individual originating from each of the population samples listed; the other 95 population samples representing other parts of the world had smaller probabilities. By the rules of likelihood, the population with the greatest probability of this genotype becomes the most likely population of origin. Likelihood ratios indicate how much more likely the best population is compared to others. By convention, a population with a ratio of 100 or more is significantly less likely to be the origin of the sample. Nineteen of the populations in Table S1 have ratios less than 100 and cannot be eliminated as the origin of this individual.

4. Discussion and conclusion

The additional population samples raise the 55 AISNP panel to having the largest number of reference population samples (125) and individuals (nearly 7000) of any public forensic ancestry panel. The absence of other population samples with data for all of these AISNPs illustrates the huge empty matrix problem with forensic panels of SNPs: different populations in the published literature have been typed for different sets of SNPs making comparison and integration impossible.

The value of more population samples is indicated by the results in Table 1. Without the North African samples from Tunisia and Libya, this Libyan individual's more likely populations of origin would be the Palestinian Arab sample, Kuwaiti, and Southwest Asian samples. The most likely broad region would have been suggested by those results, but the additional reference populations shift the interpretation toward North Africa. There will almost always be several populations of possible origin that are not significantly excluded and the denser the biogeographic coverage of a region, the more one expects to see several populations with low likelihood ratios. Thoughtful interpretation of the results will always be necessary for any forensic ancestry panel, especially since the population of origin may not be among the reference population samples. With more highly informative markers also tested on all 125 population samples (and more), it may be possible to narrow the range of possible ancestral populations. It is our opinion that this set of 55 AISNP is not the ultimate final panel nor are all of these 55 likely to be included in an improved panel.

The ideal forensic ancestry inference resource will consist of a large number of highly informative AISNPs with full data on a large number of population samples representing all regions of the world. In this context, the next best panel we are aware of for global ancestry is the one from the Seldin Lab [5,6]; we are currently working in our lab and with our collaborators to include most of those SNPs on these population samples. We are also adding at least the most informative SNPs from several of the other published ancestry panels, such as [7–9]. We encourage other researchers to consider adding their unique populations to this growing dataset of population samples which are all tested for the same set of ancestry informative SNPs.

Conflicts of interest

None.

Acknowledgments

This work was funded primarily by NIJ Grants 2013-DN-BX-K023 and 2014-DN-BX-K030 to KKK awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice and Grant BCS-1444279 from the US National Science Foundation. 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 or the Federal Bureau of Investigation. This work was partially funded by the National Natural Science Foundation of China (N.O.81471828) to CXL. OB was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under 2219-Grant Program. This research was supported in part by an appointment to the Visiting Scientist Program at the Federal Bureau of Investigation (FBI) Laboratory Division, administered by the Oak Ridge Institute of Science and Education, through an interagency agreement between the US Department of Energy and the FBI. This is publication number 15-16 of the Laboratory Division of the Federal Bureau of Investigation. Special thanks are due to the many hundreds of individuals who volunteered to give blood or saliva samples for studies of gene frequency variation and to the many colleagues who helped us collect the samples. In addition, some of the cell lines were obtained from the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University, and the African American samples were obtained from the Coriell Institute for Medical Research, Camden, New Jersey.

References

[1] K.K. Kidd, W.C. Speed, A.J. Pakstis, M.R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, F.R. Friedlaender, J.R. Kidd, Progress toward an efficient panel of SNPs for ancestry inference, Forensic Sci. Int.: Genet. 10 (2014) 23–32.
[2] H. Rajeevan, U. Soundararajan, J.R. Kidd, A.J. Pakstis, ALFRED: an allele frequency resource for research and teaching, Nucleic Acids Res. 40 (D1) (2012) D1010–D1015.
[3] H. Rajeevan, U. Soundararajan, A.J. Pakstis, K.K. Kidd, Introducing the forensic research/reference on genetics knowledge base, FROG-kb, Investig. Genet. 3 (2012) 18.
[4] D. Falush, M. Stephens, J.K. Pritchard, Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, Genetics 164 (2003) 1567–1587.
[5] R. Kosoy, R. Nassir, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, P.K. Gregersen, J.W. Belmont, et al., Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America, Hum. Mutat. 30 (2009) 69–78.
[6] J.R. Kidd, F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, K.K. Kidd, Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples, Investig. Genet. 2 (2011) 1.
[7] O. Lao, P.M. Vallone, M.D. Coble, T.M. Diegoli, M. van Oven, K.J. van der Gaag, J. Pijpe, P. de Knijff, M. Kayser, Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA, Hum. Mutat. 31 (2010) E1875–E1893.
[8] C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza, M. Casares de Cal, D. Ballard, M.V. Lareu, A. Carracedo, The SNPforID Consortium, Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci. Int.: Genet. 1 (2007) 273–280.
[9] C. Phillips, A.F. Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F.S. Rech, M.D. Carceles, A. Carracedo, P.M. Schneider, M.V. Lareu, Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci. Int.: Genet. 7 (2013) 359–366.

Electronic resources cited

ALFRED: http://alfred.med.yale.edu. FROG-kb: http://frog.med.yale.edu. 1000 Genomes project: http://www.1000genomes.org.