Forensic Science International: Genetics Supplement Series 5 (2015) e559–e560

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

Comparison of two online algorithm methods for forensic ancestry inference

L.B. Yun[a,b], T.Z. Gao[b], K. Sun[b], Y. Gu[b], Y.P. Hou[b,*]

[a] Shanghai Key Laboratory of Forensic Medicine (Institute of Forensic Science, Ministry of Justice), Shanghai, PR China
[b] Institute of Forensic Medicine, West China School of Basic Science and Forensic Medicine, Sichuan University, Chengdu, PR China

ARTICLE INFO

Article history:
Received 19 August 2015
Accepted 23 September 2015
Available online 26 September 2015

Keywords:
Ancestry informative markers
Forensic ancestry

ABSTRACT

The forensic prediction of the biogeographic ancestry based on DNA typing has become more widespread. The search for optimized panels, consisting of a small but efficient and robust set of ancestry informative markers (AIMs), has received increased interest. Several panels published in recent literature would provide excellent information on ancestry as a useful forensic investigative tool. Relying on the genetic profiles from these panels, accurate and efficient estimation of an individual ancestry depends on excellent algorithmic methods. In the present work, a comparison of the FROG-kb (http://frog.med.yale.edu) and Bayesian classification approaches of Snipper (http://mathgene.usc.es/snipper/) was carried out to assess the ability of genetic ancestry inference for the Kidd Lab 55 AISNP panel (Kidd et al., FSI Genetics 2014 (10), 23–32). Preliminary reported data from the online SPSmart browser illustrated that both algorithmic inference methods were adaptable for forensic ancestry assignments of an individual. For a few individuals, especially originating from admixed populations, ancestral assignment was inaccurate or discordant according to the likelihood calculations in FROG-kb or the ratio from Snipper forensic ancestry analysis portal. Therefore, future improvement will require more populations adding into the reference populations for the likelihood function in FROG-kb, as well as appropriate training sets applying to the online Snipper analysis tool.

1. Introduction

There is a growing interest in developing panels of Ancestry Informative Markers (AIMs) aimed at forensic prediction of an individual biogeographic ancestry. With the arrival of new genotyping technologies and large-scale genomic projects, many published AIM panels have been designed by way of selecting ancestry informative SNPs (AISNPs), which exhibit large differences in allele frequencies between populations from different geographical or ethnic groups [1]. Several optimized panels consisting of a small but efficient and robust set of AISNPs would provide excellent information on ancestry as a useful forensic investigative tool. Seven of them including Kidd Lab 55 AISNP panel have been incorporated in the web site Forensic Resource/Reference on Genetics knowledge base, FROG-kb, (http://frog.med.yale.edu) as well as available for calculating likelihoods functionality [2]. Some studies have shown that accurate and efficient estimation of an individual ancestry also depends on excellent algorithmic methods. Over the last several years, methods or algorithms such as real-time tools in Snipper (http://mathgene.usc.es/snipper/) have been developed to improve forensic ancestry analysis based on an individual's profile [3]. It is beneficial to assess the ability of algorithmic methods used to forensic ancestry inference. In the present work, comparison of these two online algorithm methods based on Kidd Lab 55 AISNP panel was carried out.

2. Materials and methods

The data set explored in this study is obtained from 1000 Genomes Phase I May 2011, which is available in the online SPSmart browser [4]. The data set consisted of 55 AISNPs according the Kidd Lab 55 AISNP panel, which was downloaded as a list containing each rs number and an active link to the dbSNP record for that SNP. These 55 SNPs were extracted from 1093 samples representing 14 populations distributed in four continents: Africa (ASW, LWK,YRI), Europe (CEU, FIN, GBR, IBS,TSI), East Asia (CHB, CHS, JPT) and America (CLM, MXL, PUR). Then ten profiles for 55 AISNP randomly extracted from each population. So the data set

* Corresponding author at: Institute of Forensic Medicine, West China School of Basic Science and Forensic Medicine, Sichuan University, 610041 Chengdu, PR China. Fax: +86 28 85501550.
E-mail address: forensic@scu.edu.cn (Y.P. Hou).

http://dx.doi.org/10.1016/j.fsigss.2015.09.221
1875-1768/

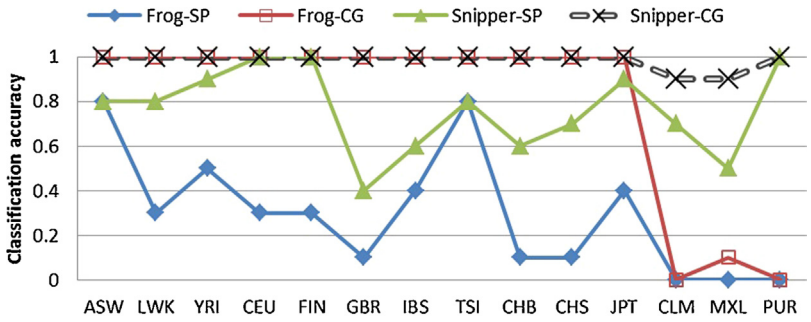

Fig. 1. Comparison of ancestry classification accuracy for two online algorithm methods in 14 populations. The blue line and red represent the accuracy of the test samples as the inferred ancestry in accordance with their specific population (Frog-SP) and original continental group (Frog-CG) using the calculations in FROG-kb. The green and black line represent the accuracy of the test samples as the inferred ancestry in accordance with their specific population (Snipper-SP) and original continental group (Snipper-CG) using the Snipper analysis portal. These two online algorithm methods were adaptable for forensic ancestry assignments of an individual, but their ability of for ancestry inference varied obviously.

using for calculation was composed of 140 test samples from 14 populations.

According to the pipeline of Kidd Lab 55 AISNP panel in FROG-kb, the 'file upload' option for users to enter genotype data was chosen to perform the analysis. After pasting the genotype information as the file format, the complete file was copied and pasted in the text area provided in the 'Input Genotype for a Panel' function. An individual's genotype was calculated follow the instructions and details especially given above.

Following the function of 'Binary AIM classification of individuals', classification with a custom Excel file of populations was downloaded. During creating the Excel file of populations in the required format, 1000 profiles from 1093 samples were considered to input the file due to the limitation. The other paramaters were chosen as classifier Naive Bayes (Hardy–Weinberg principle applies) after the excel file uploaded with the Data input function. The classification was carried out while 110 bases of 55 AISNP were typed into individual Data input function.

3. Results and discussion

Kidd Lab 55 AISNP panel with the higher population classification performance can be used to infer individual ancestry and compare to the expected ancestry. The distribution of classification accuracy for the two online algorithm methods is summarized in Fig. 1.

It can be clearly seen that most of test samples obtained correct ancestry predictions at continent level from the likelihood calculations in FROG-kb and the Bayes LRs calculated in Snipper. When the population with the highest likelihood was defined as the inference of ancestry for the test sample, some individual profiles received virtually 100% of the inferred ancestry in accordance with their geographic origin ancestry. In contrast, ancestral assignment of a few individuals, especially originating from CLM, MXL and PUR, were inaccurate according to the likelihood calculations in FROG-kb or the ratio from Snipper forensic ancestry analysis portal. To the best of our knowledge, admixture or divergent population origins can create higher levels of heterogeneity that may affect how well the ancestry panel and classification method differentiate all populations worldwide [5]. The results also indicated substantial divergence for these admixed populations from 1000 Genomes. Adding more multiple ancestral populations into the reference populations should contribute to estimate accurately admixed ancestry.

As a results, the ability of two online algorithm methods for forensic ancestry inference varied significantly from population to population. This variability is extremely large in the populations distributed in Africa and America continent. Although more individual profile were misclassified by the calculation function in FROG-kb than the Snipper portal, it should be noted that the number of reference population in FROG-kb was many times greater than the training set consisted of 14 populations in Snipper. It was noticeable that greater reference populations, especially considering very closely related groups would improve forensic ancestry inference to routinely the specific population. Since the difference between the performance of both ancestry estimating methods, it is worth to explore the reasonable size of reference population or training set in the future.

4. Conclusions

Our comparison study demonstrated both online algorithmic inference methods were adaptable for forensic ancestry assignments of an individual. Caution should be exercised when inferring ancestries for individuals from populations poorly represented biogeographic regions and origined with admixture components. Therefore, future improvement will require more populations adding into the reference populations for the likelihood function in FROG-kb, as well as appropriate training sets applying to the online Snipper analysis tool.

Role of funding

This study was supported by the open fund (KF1408) from Shanghai Key Laboratory of Forensic Medicine (Institute of Forensic Science, Ministry of Justice), China.

Conflict of interest

None.

References

[1] C. Santos, et al., Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: results of a collaborative EDNAP exercise, Forensic Sci. Int. Genet. 19 (2015) 56–67.
[2] K.K. Kidd, et al., Progress toward an efficient panel of SNPs for ancestry inference, Forensic Sci. Int. Genet. 10 (2014) 23–32.
[3] C. Phillips, et al., Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set, Forensic Sci. Int. Genet. 11 (2014) 13–25.
[4] J. Amigo, A. Salas, C. Phillips, A. Carracedo, SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access, BMC Bioinf. 9 (2008) 428.
[5] J. Pardo-Seco, F. Martinon-Torres, A. Salas, Evaluating the accuracy of AIM panels at quantifying genome ancestry, BMC Genomics 15 (2014) 543.