Forensic Science International: Genetics Supplement Series 5 (2015) e34–e36

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

# Assessing the suitability of different sets of InDels in ancestry estimation

Juliana G. Aquino[a], Juliana Jannuzzi[a], Elizeu F. Carvalho[a], Leonor Gusmão[a,b,c,*]

[a] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil
[b] IPATIMUP (Institute of Pathology and Molecular Immunology from the University of Porto), Porto, Portugal
[c] Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal

ARTICLE INFO

*Article history:*
Received 21 August 2015
Accepted 7 September 2015
Available online 10 September 2015

*Keywords:*
Admixture
InDel
Rio de Janeiro
Brazil

ABSTRACT

Ancestry informative markers (AIMs) are useful to estimate individual and population ancestries, providing important information to forensic investigations. Several AIM sets were described and evaluated by comparison with data from GWAS. Taking into account that an efficient set of AIMs shall provide identical results between full brothers and GWAS are not easily performed, we aimed to see if the accuracy of the ancestry estimates are correlated to differences obtained in siblings. Pairs of siblings from Brazil were genotyped for 83 InDels; and values of African, European and Native American contributions were compared using diverse sets of markers. The comparison of the ancestry in siblings was only meaningful for markers with high inter-populations variation. The lowest average differences between brothers were obtained for the complete set of 83 InDels, even including markers with low inter-populations variation.



## 1. Introduction

Ancestry informative markers – AIMs present significant differences in their allelic frequencies in different ancestral or geographically distant populations. They can be successfully used to estimate ancestry, at both individual and population levels, providing important information to forensic investigations [1]. There are several sets of markers described as being useful to determine ancestry, and their efficiency to estimate accurate ancestry proportions is generally evaluated by comparison with data generated by GWAS – Genome Wide Association Studies [2,3]. The determination of ancestry in siblings may be a good strategy to evaluate markers' performance, taking into account that an efficient set of markers shall provide identical results of ancestry between them.

The aim of this study was to compare ancestry values among siblings for different groups of InDel markers. More precisely, how the inter-population diversity as well as the number of markers would affect the accuracy and differences between siblings' ancestry estimates.

## 2. Materials and methods

A total of 26 pairs of siblings were selected from kinship cases investigated in the DNA Diagnostic Laboratory of the State University of Rio de Janeiro, Brazil. Written informed consent was obtained from all participants for cooperation in this study under strictly confidential conditions. DNA was extracted with Chelex [4]. Samples were genotyped for 83 InDels with different degrees of diversity and inter-population variation, using two PCR multiplex protocols previously described [1,5]. Capillary electrophoresis and detection were performed on a 3500 Genetic Analyser using POP-7™ polymer (Applied Biosystems). The genotypes were assigned using the software GeneMapper ID v4.1 (Applied Biosystems).

The apportionment of genetic ancestral contributions was estimated in all samples using the STRUCTURE v2.3.3 software [6]. A supervised analysis was performed using prior information on the geographic origin of the reference samples, assuming an essentially tri-hybrid contribution from Native Americans, Europeans and Africans (i.e., $K=3$). STRUCTURE runs consisted of 100,000 burnin steps followed by 100,000 Markov Chain Monte Carlo (MCMC) iterations. The option "*Use population Information to test for migrants*" was used with the Admixture model. Allele frequencies were correlated and updated using only individuals with POPFLAG = 1 (in this case, the HGDP-CEPH samples used as reference).

* Corresponding author. Fax: +55 21 23340594.
*E-mail address:* lgusmao@ipatimup.pt (L. Gusmão).

http://dx.doi.org/10.1016/j.fsigss.2015.09.014
1875-1768/

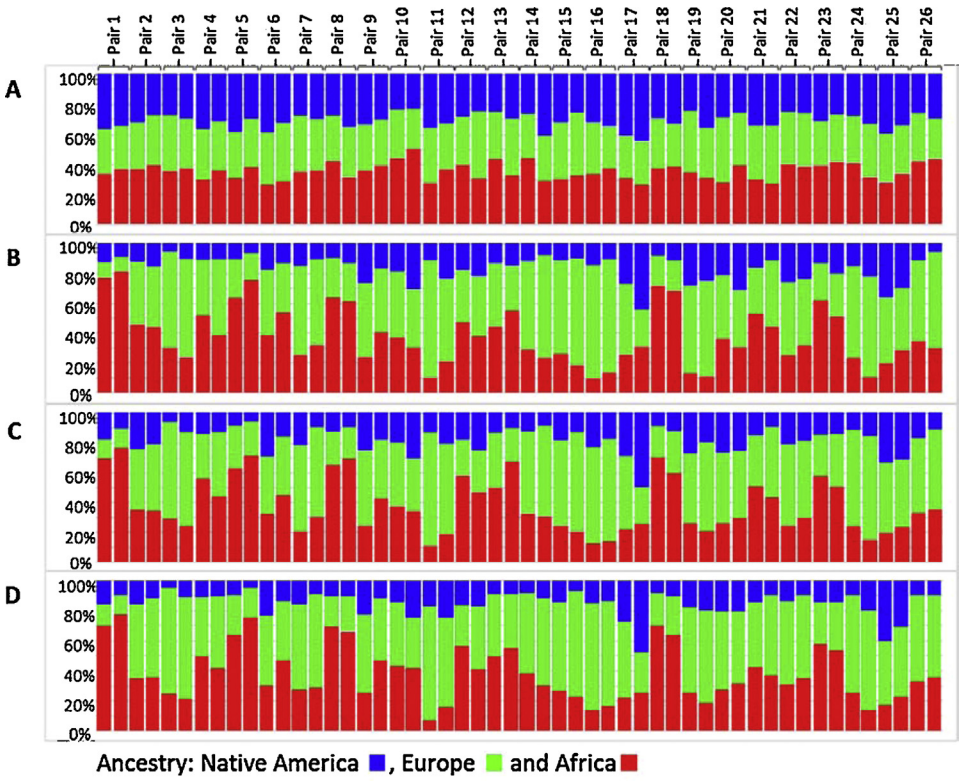


**Fig. 1.** Ancestry estimates for markers included in Set 1 (A), Set 2 (B), 46 AIMs (C) and for the 83 InDels (D).

## 3. Results and discussion

In order to test the effect of using markers with low vs. high levels of population differentiation, values of African, European and Native American contributions were calculated for 26 pairs of siblings from the admixed population of Rio de Janeiro, using the 30 markers with lowest (Set 1) and the highest (Set 2) inter-population variation (Fig. 1A and B). Ancestry estimates from the three contributing populations (both within and between the 26 pairs) were very similar for Set 1; contrasting with the higher variation presented by the Set 2. Despite the low efficiency of the first set of markers to estimate ancestry, the differences between siblings were lower than those obtained with the second set (Set 1 and Set 2 in Table 1). Such fact is due to the observation that markers with low levels of population differentiation tend to produced similar errors. Indeed, values of ancestry below 0.33 for set 2 were always overestimated by set 1, and higher values were underestimated. The non-random deviation of estimates for set 1 precludes the usefulness of a comparative analysis in siblings..

Results were also compared for set 2, 46 AIMs and the 83 full set (Fig. 1B–D). The differences observed between pairs of sibling were apparently random, which makes the comparison of siblings meaningful. Although the ancestry proportions were not significantly different for the three sets (Table 1), the highest differences between brothers were found for the 30 markers' set, followed by the 46 AIMs and were lower for the full set of 83 markers. These results, support a better performance of the complete set in ancestry estimation, although including markers with low inter-populations variation.

**Table 1**
Values of African (AFR), European (EUR) and Native American (NAM) ancestry estimated in the whole data set using different groups of markers, together with the sum and the average differences observed between siblings.

| | AFR | EUR | NAM |
|---|---|---|---|
| Set 1 | | | |
| Ancestry proportion | 0.356 | 0.330 | 0.315 |
| Total differences | 1.359 | 0.940 | 1.140 |
| Average differences | 0.054 | 0.038 | 0.046 |
| Set 2 | | | |
| Ancestry proportion | 0.361 | 0.466 | 0.173 |
| Total differences | 1.999 | 1.845 | 1.428 |
| Average differences | 0.080 | 0.074 | 0.057 |
| 46 Plex | | | |
| Ancestry proportion | 0.363 | 0.455 | 0.182 |
| Total differences | 1.746 | 1.811 | 1.480 |
| Average differences | 0.070 | 0.072 | 0.059 |
| 83 Plex | | | |
| Ancestry proportion | 0.389 | 0.482 | 0.156 |
| Total differences | 1.668 | 1.732 | 1.279 |
| Average differences | 0.067 | 0.069 | 0.051 |

## Conclusion

The approach followed in this study to evaluate the performance of groups of genetic marker using pairs of siblings, proved not to be adequate when markers with very different inter-populations variation are compared. Despite the low efficiency of some markers to produce accurate ancestry estimates, they produce the same type of errors, reducing, therefore, the differences observed among siblings.

The deviations of the estimates obtained for groups of markers with high inter-populations variation were apparently random, making the comparison of the ancestry in siblings relevant. Using this strategy, we observed that the average differences between

brothers decrease with the addition of more markers, supporting a better performance of large sets of markers, independently of their individual performance.

## Financial support

Financial support was granted by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and DNA Program – State University and Justice Court of Rio de Janeiro, Brazil. IPATIMUP integrates the i3S Research Unit, which is partially supported by FCT, the Portuguese Foundation for Science and Technology.

## Conflict of interest

None.

## References

[1] R. Pereira, C. Phillips, N. Pinto, et al., Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing, PLoS One 2 (1) (2012) e29684.
[2] M.G. Joshua, C.F.L. Juan, R.G. Christopher, et al., Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas, PLoS Genet. 8 (2012) e1002554.
[3] D. Michelle, M. Lize Van, G. Ushma, et al., A panel of ancestry informative markers for the complex five-way admixed South African coloured population, PLoS One 8 (2013) e82224.
[4] M.V. Lareu, C.P. Phillips, A. Carracedo, et al., Investigation of the STR locus HUMTH01 using PCR and two electrophoresis formats: UK and Galician Caucasian population surveys and usefulness in paternity investigations, Forensic Sci. Int. 66 (1994) 41–52.
[5] R. Pereira, C. Phillips, C. Alves, et al., A new multiplex for human identification using insertion/deletion polymorphisms, Electrophoresis 30 (21) (2009) 3682–3690.
[6] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.