Forensic Science International: Genetics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Forensic genetic analysis of bio-geographical ancestry

Chris Phillips *

Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Galicia, Spain

ARTICLE INFO

Article history:
Received 23 October 2014
Received in revised form 2 May 2015
Accepted 14 May 2015

Keywords:
Bio-geographical ancestry
Ancestry informative markers (AIMs)
Population genetics
SNP
PCA
STRUCTURE

ABSTRACT

With the great strides made in the last ten years in the understanding of human population variation and the detailed characterization of the genome, it is now possible to identify sets of ancestry informative markers suitable for relatively small-scale PCR-based assays and use them to analyze the ancestry of an individual from forensic DNA. This review outlines some of the current understanding of past human population structure and how it may have influenced the complex distribution of contemporary human diversity. A simplified description of human diversity can provide a suitable basis for choosing the best ancestry-informative markers, which is important given the constraints of multiplex sizes in forensic DNA tests. It is also important to decide the level of geographic resolution that is realistic to ensure the balance between informativeness and an over-simplification of complex human diversity patterns. A detailed comparison is made of the most informative ancestry markers suitable for forensic use and assessments are made of the data analysis regimes that can provide statistical inferences of a DNA donor's bio-geographical ancestry.

1. Introduction

1.1. The forensic context of bio-geographical ancestry analysis

In London seventeen years ago, seeing somebody acting suspiciously outside my neighbor's house, I contacted the police. They asked a simple question that often frames the eyewitness prompts made by UK police officers: "Was he White, Black or Asian". I said the person appeared White (resisting the need to correct these descriptions to the more neutral terminology of European, African, South Asian). As it was dark and I had brief glimpses, it was impossible to provide a concrete description. Eyewitness is notoriously unreliable and can be shaped by preconceptions or the circumstances of a crime [1]. Therefore, the inference of bio-geographical ancestry using markers with population-differentiated variation provides opportunities to strengthen eyewitness accounts or in their absence, gain information about a suspect.

This review explores the current viability of forensic DNA tests estimating ancestry that can provide investigative leads when eyewitness testimony or a database hit are not available. In simple lay terms, ancestry can be described as the genetic inheritance each individual carries from their ancestors, in the immediate past from their kinship, over longer periods from population members that have occupied the same place of origin. Bio-geographical ancestry analysis concentrates on the population variation found in an individual that can signal their origin from a particular geographic region. Forensic bio-geographical ancestry testing exploits much of the recent advances in the understanding of human genomic variation, with the key factor that tests must be sensitive enough to successfully genotype contact trace DNA or they will lack utility. Inference of ancestry in forensic analysis gives possibilities to substitute eyewitness testimony as described above—when descriptions are uncertain, unavailable or may misdirect investigators. Yet in forensic analysis, ancestry inference offers many other applications, including: (i) aiding cold case reviews with additional data on linked profiles; (ii) achieving more complete identifications of missing persons or disaster victims; (iii) confirming donor's self-declared ancestry and therefore maintaining the accuracy of databases for STRs, Y-markers and mitochondrial variation (mtDNA); (iv) refining familial search strategies highly dependent on STR allele frequency assumptions made prior to searching [2]; (v) assessing atypical combinations of physical characteristics in individuals with admixed parentage, e.g., using IrisPlex [3–5]); (vi) enhancing genetic studies where forensic sensitivity is necessary, e.g., testing medical archive material or archaeological DNA [6].

This review centers on autosomal markers, despite Y and mtDNA uniparental variation being highly differentiated geographically and therefore often forming the first and only step in forensic ancestry inference. Y and mtDNA variation is undisrupted by recombination, so is preserved in both lineages and correlates

* Tel.: +34 981 582 327; fax: +34 981 580 336.
E-mail address: c.phillips@mac.com (C. Phillips).

http://dx.doi.org/10.1016/j.fsigen.2015.05.012
1872-4973/

strongly with continental regions. However, Y and mtDNA variants collectively form single markers that can misrepresent an individual's overall ancestry when distant male/female lineages are inherited that have atypical ancestry. A notable example of this risk of misinterpretation was detection of African Y-chromosomes in a North Yorkshire kinship group [7]. As co-ancestry in an individual indicates population admixture, increasingly common in modern urban demographics, the probability of detecting atypical lineages and misinterpreting an individual's overall ancestry rises markedly. Another advantage of recombining autosomal loci compared to Y and mtDNA is the relative ease with which population data is obtained, with as few as 30–40 samples providing adequate population allele frequency estimates. In the 11-M Madrid bomb investigation [8], discrepancies between ancestry inferences from autosomal markers and both Y and mtDNA were seen. These stemmed from limited database coverage of North African populations, hampering interpretation of Y and mtDNA data based on very limited surveys of this region. The need for much larger databases to measure haplotype variation impacts reliable interpretation of uniparental variation in many less well-studied regions and has prompted the YHRD/EMPOP forensic-community databases [9,10].

Lastly, it is important to remember forensic estimation of bio-geographical ancestry is not confined to genetic analysis, nor is it unique to the DNA profiling age. Analysis of skeletal biometrics is used to estimate ancestry with statistical classification approaches (e.g., canonical plots) similar to principal component analysis applied to genetic data. Early forensic ancestry tests used the Duffy marker (rs2814778) 20 years before DNA profiling and it remains the most differentiated locus (for a brief survey of forensic ancestry analysis with classical markers, see [11]).

2. Patterns of human population structure

Any concise overview of human population structure, as it is currently understood, will be an oversimplification. However, before ancestry can be inferred from small sets of forensically viable markers it is necessary to attempt a definition of population groups based on the most strongly differentiated patterns of genetic structure. The worldwide human population is clearly not a single entity, nor is it always appropriate to define small populations confined to narrow regions. The constraints of forensic multiplex sizes and collection of sufficient reference data means a simplified description of complex human population structure is a necessary compromise.

Human populations are not fully interbreeding, since geographic distance by itself creates a strong constraint on random mating. Additionally, geophysical barriers such as oceans and mountains have restricted free movement of people away from regions defined by such barriers. Therefore, population structure in early human groups became established as they continued to mate with immediate neighbors that shared their ancestry. This means forensic tests estimating ancestry might expect some success, depending on the distribution of human population structure remaining intact today. Pre-genomics studies of population variation, starting with Lewontin [12], attempted to measure what structure existed in modern human population groups using limited numbers of polymorphic markers. Despite variation in loci and populations analyzed, later studies with the same approach

	SNP	Chr:position	Gene	RA	RA AFR (85 YRI)	RA EUR (85 CEU)	RA E ASN (85 CHB)	I_n3	I_n AFR	I_n EUR	I_n E ASN	AFR	EUR	E ASN
1	**rs4988235**	2:136608646	MCM6	C	1	0.276	1	*0.335*	*0.131*	*0.349*	*0.131*			
2	**rs12075**	1:159175354	DARC	G	0	0.435	0.947	*0.379*	*0.318*	*0.001*	*0.313*			
3	**rs2814778**	1:159174683	DARC	A	1	0	0	*0.599*	*0.662*	*0.200*	*0.200*			
4	**rs1426654**	15:48426484	SLC24A5	A	0.018	1	0.029	*0.557*	*0.188*	*0.617*	*0.169*			
5	**rs3827760**	2:109513601	EDAR	T	1	1	0.053	*0.531*	*0.186*	*0.186*	*0.576*			
								Cumulative I_n POP :	***1.48***	***1.35***	***1.39***			

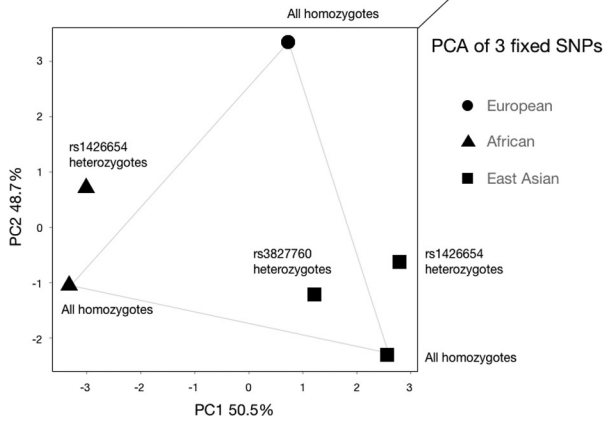

Fig. 1. Five examples of AIM-SNPs. SNP 1 shows a population group-specific allele, SNP 2 has near-fixed variation between Africans and East Asians. SNPs 3–5 are the most informative (reflected in the I_n values listed) with fixed alleles in each group. Combined I_nPOP divergences reach a reasonably comparable level of balance as SNP properties compensate for the distribution of variation amongst the groups analyzed. The PCA plot shows analysis of genotypes for SNPs 3–5, where a perfect triangle indicates genetic data was almost completely transformed to two PC axes. Genotypes are from sample size-adjusted 1000 Genomes Phase III data. RA, reference allele; Chr, chromosome. The promotor SNP for LCT, rs4988235 is sited in MCM6.

Please cite this article in press as: C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. (2015), http://dx.doi.org/10.1016/j.fsigen.2015.05.012