Theoretical Population Biology xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Weight of the evidence of genetic investigations of ancestry informative markers

Torben Tvedebrink [a,*], Poul Svante Eriksen [a], Helle Smidt Mogensen [b], Niels Morling [b]

[a] Department of Mathematical Sciences, Aalborg University, Denmark
[b] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 30 June 2017
Available online xxxx

Keywords:
Forensic genetics
Ancestry informative markers
Population genetics
Exact likelihood ratio test

ABSTRACT

Ancestry-informative markers (AIMs) are markers that give information about the ancestry of individuals. They are used in forensic genetics for predicting the geographic origin of the investigated individual in crime and identification cases. In the exploration of the genogeographic origin of an AIMs profile, the likelihoods of the AIMs profile in various populations may be calculated. However, there may not be an appropriate reference population in the database. The fact that the likelihood ratio (LR) of one population compared to that of another population is large does not imply that any of the populations is relevant.

To handle this phenomena, we derived a likelihood ratio test (LRT) that is a measure of absolute concordance between an AIMs profile and a population rather than a relative measure of the AIMs profile's likelihood in two populations. The LRT is similar to a Fisher's exact test. By aggregating over markers, the central limit theorem suggests that the resulting quantity is approximately normally distributed. If only a few markers are genotyped or if the majority of the markers are fixed in a given population, the approximation may fail. We overcome this using importance sampling and show how exponential tilting results in an efficient proposal distribution.

By simulations and published AIMs profiles, we demonstrate the applicability of the derived methodology. For the genotyped AIMs, the LRT approach achieves the nominal levels of rejection when tested on data from five major continental regions.

© 2017 Elsevier Inc. All rights reserved.

UNCORRECTED PROOF

1. Introduction

Ancestry-informative markers (AIMs) are genetic markers that give information about the geographic ancestry of individuals. Cavalli-Sforza et al. (1994) reviewed the markers that were available at the time containing information about ancestry. However, more recently research has focused on identifying carefully selected markers with higher information about ancestry (see e.g. Jobling et al., 2014). Ancestry, the geography of human populations, and genetic polymorphisms are closely associated with each other (Rosenberg et al., 2002, 2003; Serre and Pääbo, 2004; Manica et al., 2005; Wang et al., 2012). In forensic genetics, typing of Short Tandem Repeats (STRs), which are presently the standard tool in forensic genetic identification and relationship testing, gives information on *genogeographic* ancestry (Brinkmann et al., 1998), but this information is only used to some extent (Phillips, 2015). Throughout the text, we use the term *genogeographic* (Harrison, 1977) rather than *biogeographic* to emphasise that our analysis is solely based on genetic markers and not other available biomarkers. Hence, genogeographic markers may be considered a subset of biogeographic markers. Other genetic markers besides AIMs are available, including Y-chromosome markers (Jobling and Tyler-Smith, 2004) and mitochondrial (mtDNA) sequence variation (Egeland et al., 2005). These markers have benefits and limitations that relate to their paternal and maternal lineages. Autosomal markers – especially Single Nucleotide Polymorphisms (SNPs) – presently are the preferred ancestry markers in the forensic community due to the fact that many SNPs have very different allele frequencies in various populations (e.g. Kidd et al., 2014). Insertion–deletion polymorphisms (indels) may also be valuable AIMs (Yang et al., 2005). Analysis of indels has a number of advantages compared to those of SNPs, but the number of indels and their informative value are less than those of SNPs (Phillips, 2015).

Investigations of AIMs have been used in crime cases in order to predict the genogeographic ancestry of the donor of a biological trace at the scene of crime, and biostatistical predictive tools for forensic genetic use have been developed (e.g. Phillips et al., 2007). In forensic genetic investigations in crime cases, it is important to perform the investigations on very small amounts of DNA, i.e. less than one nanogram of DNA. Investigations of

* Correspondence to: Fredrik Bajers Vej 7G, DK-9220 Aalborg East, Denmark.
E-mail addresses: tvede@math.aau.dk (T. Tvedebrink), svante@math.aau.dk (P.S. Eriksen), helle.smidt@sund.ku.dk (H.S. Mogensen), niels.morling@sund.ku.dk (N. Morling).

https://doi.org/10.1016/j.tpb.2017.12.004
0040-5809/© 2017 Elsevier Inc. All rights reserved.

genogeographic ancestry are especially valuable if a relatively large number of markers (e.g. SNPs) can be investigated. Therefore, the introduction of investigations of large sets of SNPs with Massively Parallel Sequencing—MPS (Themudo et al., 2016; Pereira et al., 2017) was very welcome. This technology makes it possible to investigate a large number of SNPs simultaneously with less than 1 ng of DNA (Børsting et al., 2014). During the last decade, a number of investigations have supplemented our knowledge about the distribution of genetic markers in various human populations (e.g. Kidd et al., 2014), and AIMs panels and databases with information on genogeographical markers have been established (e.g. Phillips et al., 2009; SNIPPER: http://mathgene.usc.es/snipper, Cheung et al., 2000; Pakstis et al., 2017; FROG-kb: http://frog.med.yale.edu/FrogKB/).

The practical use of investigations of AIMs in crime and identification cases is aiming at predicting the ancestry or genogeographic origin of the investigated individual. This may substitute and/or support eyewitness testimony when descriptions are unavailable or uncertain. This is of special interest in cases, in which DNA from the perpetrator is available, but no suspect is identified and no match is found in crime DNA databases. Similarly, AIMs testing is of value in identifications of missing persons and disaster victims.

The strategy for interpretation of the results of AIMs investigations can be either explorative (hypothesis generating) or hypothesis testing. In the explorative situation, the likelihoods of the AIMs profile of the individual in various populations may be calculated, and the one with the highest likelihood may be considered the genogeographical population of origin. This simple approach has, however, some difficulties that will be discussed below. In the hypothesis testing situation, two highly relevant populations may be identified a priori. The likelihoods of observing the AIMs profile in the two populations may be calculated, and the ratio between the two likelihoods may be calculated (i.e. the likelihood ratio). This process may be repeated with other combinations of relevant populations. This strategy will offer the likelihood ratio as the weight of the evidence as recommended by the International Society for Forensic Genetics—ISFG (Morling et al., 2002; Gill et al., 2006).

However, there may not be an appropriate population in the database of reference populations leading to suspicious results of the ancestry prediction (Kidd et al., 2014; Themudo et al., 2016). The fact that the likelihood is substantially larger in one population than in another does not prove that any of the two populations are relevant to the AIMs profile at hand. This is due to the fact that even though the populations may be exclusive, they are not exhaustive in the sense that they cover all possible human populations.

To handle this phenomenon, we derived a likelihood ratio test, by which we can assess if there is at least one population in our database of reference populations that is "sufficiently close" to the "true" ancestry population of the AIMs profile at hand.

In case the null hypothesis is rejected for all population samples in the database of reference populations, we refrain from computing further quantities of forensic interest. However, in case we fail to reject all null hypotheses, we compute pairwise likelihood ratios, comparing the probability of the evidence under the competing hypotheses, i.e. different populations of origin. As the allele frequencies/counts are based on samples taken from the various reference populations, we use the methods of Chakraborty et al. (1993) to assess the variance of the genotype probabilities due to sampling effects.

The manuscript is organised as follows. In Section 2, we briefly describe the public available data used. Section 2 also contains the derivation of the likelihood ratio test and a brief discussion on computation of the evidential weight in the AIMs framework. Some results based on the derived test and evidential weight are presented in Section 3. Discussions and conclusions are given in Section 4.

2. Material and methods

Section 2.1 describes the data used in this study and methodology development. The majority of the data originate from public available repositories (ALFRED, https://alfred.med.yale.edu/).

In Section 2.2, we describe the likelihood ratio test that assesses whether a certain population can be accepted as a possible population of genogeographic origin of the AIMs profile. This test is an absolute measure of concordance between the AIMs profile and a population, rather than a relative measure of the AIMs profile's likelihood in two populations (the likelihood ratio). Methods for computing p-values using importance sampling are derived in Section 2.3, and weight of evidence computations is discussed in Section 2.4.

2.1. Materials

We used the 128 AIMS SNPs from Kosoy et al. (2009), and the genotype frequencies of 119 populations from Kidd et al. (2011). The populations are geographically scattered over most continents, e.g. Africa, Americas, Asia, and Europe (Kidd et al., 2011). The frequencies for a Greenlandish population (77 individuals) were from Themudo et al. (2016). Frequencies for a Danish population (142 individuals) and the data from a Somali population (98 individuals) were from Pereira et al. (2017). Frequencies for two populations (22 and 25 individuals, respectively) from Ecuador were from Santangelo et al. (2017). Typing of additional populations from Morocco (86 individuals, unpublished), Iran (87 individuals, unpublished), and Turkey (114 individuals, unpublished) was essentially as described by Pereira et al. (2017).

Briefly, DNA libraries were constructed using the Ion AmpliSeq™ Library Kit 2.0 (Thermo Fisher Scientific, Waltham, USA) and the Precision ID Ancestry Panel (Thermo Fisher Scientific, Waltham, USA). The DNA was amplified with 25 PCR cycles, and the amplificates were converted into libraries on a Biomek® 3000 Laboratory Automation Workstation (Beckman Coulter Inc., CA, USA) with an in-house customised script (available upon request). The resulting libraries were quantitated using the Qubit® 3.0 and 20–25 libraries, and were subsequently pooled in equimolar amounts using the Biomek® 3000. The pooled libraries were converted into sequencing templates by emulsion PCR and enrichment of templated Ion Sphere™ Particles using the Ion Chef™ instrument (Thermo Fisher Scientific, Waltham, USA). Sequencing templates were loaded onto 318 chips by the Ion Chef™, and sequenced on the Ion PGM™.

Samples from 89 individuals from Morocco were collected (Tomas et al., 2008). Samples from 87 individuals from Iran were collected (Farzad et al., 2013). Samples from 114 individuals with self-declared Turkish birthplace were from Danish paternity and immigration cases.

The work was approved by the Danish ethical committee (H-1-2011-081).

2.2. Testing for appropriate population in database

As mentioned in Section 1, the inexhaustibility of the reference populations in a database implies that the relative comparison of profile likelihoods may be irrelevant and potentially misleading. In order to measure the absolute concordance between an AIMs profile and a population, we constructed a hypothesis test that assesses whether the AIMs profile and the sample from the population are likely to originate from the same allelic distribution.

For a given panel of AIMs, we assume that L bi-allelic markers have been genotyped in J distinct populations, with n_j individuals being genotyped in population $j \in \{1, \ldots, J\}$. Since the AIMs investigated here are bi-allelic, only the allele frequency of one of

UNCORRECTED PROOF

Download English Version:

https://daneshyari.com/en/article/8877460

Download Persian Version:

https://daneshyari.com/article/8877460

Daneshyari.com