Forensic Science International: Genetics 27 (2017) 132–141

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

A nearest neighbour approach by genetic distance to the assignment of individual trees to geographic origin

Bernd Degen[a,*], Céline Blanc-Jolivet[a], Katrin Stierand[a], Elizabeth Gillet[b]

[a] Thünen Institute of Forest Genetics, Sieker Landstrasse 2, Grosshansdorf, 22927, Germany
[b] Forest Genetics and Forest Tree Breeding, Faculty of Forest Sciences and Forest Ecology, Georg-August-University of Göttingen, Büsgenweg 2, 37077 Göttingen, Germany

ARTICLE INFO

Article history:
Received 15 March 2016
Received in revised form 21 December 2016
Accepted 27 December 2016
Available online 29 December 2016

Keywords:
Forensics
Genetic assignment
Genetic distance
GeoAssign
Geographic origin
Timber
Trees

ABSTRACT

During the past decade, the use of DNA for forensic applications has been extensively implemented for plant and animal species, as well as in humans. Tracing back the geographical origin of an individual usually requires genetic assignment analysis. These approaches are based on reference samples that are grouped into populations or other aggregates and intend to identify the most likely group of origin. Often this grouping does not have a biological but rather a historical or political justification, such as "country of origin".

In this paper, we present a new nearest neighbour approach to individual assignment or classification within a given but potentially imperfect grouping of reference samples. This method, which is based on the genetic distance between individuals, functions better in many cases than commonly used methods. We demonstrate the operation of our assignment method using two data sets. One set is simulated for a large number of trees distributed in a 120 km by 120 km landscape with individual genotypes at 150 SNPs, and the other set comprises experimental data of 1221 individuals of the African tropical tree species Entandrophragma cylindricum (Sapelli) genotyped at 61 SNPs. Judging by the level of correct self-assignment, our approach outperformed the commonly used frequency and Bayesian approaches by 15% for the simulated data set and by 5–7% for the Sapelli data set.

Our new approach is less sensitive to overlapping sources of genetic differentiation, such as genetic differences among closely-related species, phylogeographic lineages and isolation by distance, and thus operates better even for suboptimal grouping of individuals.

© 2016 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

In recent years, the application of forensic methods based on genetic markers to assign individual plants and animals to their geographic origin [1,2] has gained importance for the control of trade regulations and consumer protection. Whereas many animal and plant species are protected by the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), the protection level can depend on the country of origin (species listed in CITES annex 3). For the timber trade, regulations such as the US Lacey Act and the EU timber regulation require the trader to declare the geographic origin of the material. Therefore it is important to be able to determine the true origin of the individuals that are traded [2].

In different areas of the world, genetic reference data bases have been developed to aid in the identification of the true geographic origin of timber [3,4]. To this end, reference samples of individual trees are collected throughout the natural distribution area of the species or in specific target regions. For each individual, the data base records the geographic coordinates of the individual and its genotype at gene loci representing different types of gene marker, especially molecular markers of the types nSSR, cp-DNA, mt-DNA and more recently SNP [5–8]. Usually, more than one individual is sampled at each reference location, and reference locations and their sampled individuals are aggregated *a priori* into groups (or classes) by non-genetic criteria such as country of origin. The objective is to assign an individual of unknown origin, such as imported timber, to that group of reference individuals to which its multilocus genotype conforms best by some criterion. Examples

* *Corresponding author.
E-mail addresses: bernd.degen@thuenen.de (B. Degen), celine.blanc-jolivet@thuenen.de (C. Blanc-Jolivet), katrin.stierand@thuenen.de (K. Stierand), egillet@gwdg.de (E. Gillet).

http://dx.doi.org/10.1016/j.fsigen.2016.12.011
1872-4973/© 2016 Elsevier Ireland Ltd. All rights reserved.

for the assignment of individuals to place of origin have been published for timber tree species, elephants, and salmon [9–12].

Most of the previously applied approaches to the assignment of an individual of unknown origin to its putative group are using the frequency method [13] and/or the Bayesian method [14]. These approaches estimate or use the allele frequencies at all loci in every group, compute the probability that the multilocus genotype of the individual in question would have been generated within each of the groups assuming that mating was at random within a large parental population with the group-wide allele frequencies at each of the mutually unlinked loci, and assign the individual to that group for which this probability of generation is highest [1]. Application of these approaches is based on assumptions that (a) the structure of the groups fits to Wright's Island Model of drift [15], (b) the alleles at each locus are in Hardy-Weinberg-Proportions (HWP) (i.e., no stochastic association between alleles at the same locus), and (c) that the loci are in linkage equilibrium or are completely unlinked (i.e., no stochastic association between the genotypes between loci). Unfortunately, these assumptions are often violated in real populations. For example, molecular markers such as nSSRs often show deviation from HWP in at all spatial scales [16,17]. Also the use of large SNP marker sets derived from Next Generation Sequencing often leads to linkage disequilibrium (LD) in the data, which would require pruning process of loci in the data [18]. To overcome this problem and to avoid loss in the Gain of Informativeness for Assignment (GIA), other methods combine some loci into haplotypes [18]. Indeed, the use of haplotypes provides information on the ancestry and recombination events, which are useful when differentiation among groups is low. However, these methods are mostly interesting when dense marker sets are used [19], and it is not clear how these methods handle missing data, which is unfortunately the main issue with genotyping of timber or other material with degraded DNA.

The success of these assignment methods depends on the extent of genetic differentiation among the groups [1]. The sampling scheme and the variability of the spatial genetic structure have an additional impact on the success of the assignment methods, especially if the groups represent political units such as countries.

When pre-defined aggregation to groups does not reflect the genetic structure among the reference individuals, genetic assignment approaches based on allele frequencies may fail [20]. Grouping according to political (e.g. country) borders instead of genetic boundaries between populations as reproductive units is particularly critical and may lead to confusing genetic mixtures within groups [21]. In the case that reference groups contain individuals of more than one reproductively isolated deme, such as different regions or even cryptic sympatric species, the mean genetic difference among individuals of different demes should be larger than the mean genetic difference among individuals of the same deme [22]. The ability of markers to identify such demes differs, however. For instance, the incomplete lineage sorting or chloroplast capture that can be observed among closely-related species at chloroplast markers [23] could hinder attempts to recognize these species in the reference data. Also, the genetic differentiation among the groups of reference samples is biased by the proportion of species mixture or the mixing of individuals from different phylogeographic lineages (e.g. refugia) within the groups or by the assignment of individuals from the same spatial genetic unit (e.g. cross-border populations) to different groups.

The opposite approach to the *a priori* specification of groups is to attempt to partition individuals into reference groups that show HWP and linkage equilibrium. Bayesian clustering approaches, as implemented for example in the program STRUCTURE [24], have become common in recent years. These aggregations are, however, also based on estimates of allele frequencies and could be biased by unequal sample sizes among genetic groups or violation of the assumption that the populations fit Wright's Island Model with relatively small, clearly differentiated populations [20]. Another major drawback of the partitioning of reference individuals into genetic groups by Bayesian clustering methods is that the genetic groups they detect may be spread over more than one country. This collides with existing legislation that requires declaration of the country of origin, the political borders of which may cut through the middle of a genetic group, making it difficult to issue a statement from genetic testing on how likely this declaration is.

This paper describes a nearest neighbour classification approach that assigns unclassified individuals to predefined classes of reference individuals, such as by the country of origin that is of relevance for the timber trade or any CITES listed species with different country restrictions. The distance between individuals is measured by the genetic distance between their multilocus genotypes, defining the nearest neighbours of a specific individual as those individuals with the smallest genic distance to it. An unclassified individual is assigned to a particular class if this class has the highest representation among a limited set of nearest neighbours by a new index I_r and if this representation is statistically significant. Since assignment is not based on estimation of allele frequencies within entire reference classes, the approach avoids the problems of possible discrepancy between political and genetic boundaries described above. We demonstrate its application using two data sets, a large set of simulated data for a hypothetical tropical tree species and experimental data of an African tree species. When applied to test whether individuals of known origin are correctly assigned to their class, it turns out that the probability of correct self-assignment is better than for conventional methods based on allele frequencies.

2. Material and methods

2.1. The frequency method

The frequency method assigns an individual to the group (population) r in which the individual's genotype is most likely to occur [25]. The allele frequency of allele n at locus m in group r is p_{rmn}. Under the assumption of Hardy-Weinberg equilibrium, a genotype $A_nA_{n'}$ has the likelihood to occur in the group r at locus m of $p_{rmn}{}^2$ if $n = n'$ and otherwise of $2\ p_{rmn}\ p_{rmn'}$. For independent loci M the likelihood is computed as the product of the likelihoods of each locus.

2.2. The Bayesian method

The Bayesian method is looking for the derivation of the probability density of group allele frequencies from sample group frequencies [25]. Here assuming an equal *prior* probability density to the allele frequencies of each locus in each group we get the marginal probability of observing an individual with genotype A_nA_n' at locus m in group r as [14]:

$$\frac{\left(b_{rmn}+\frac{1}{B_m}+1\right)(b_{rmn}+1/N_m)}{(b_{rm}+2)(b_{rm}+1)}\ \textit{if}\quad n=n'$$

$$\frac{2\left(b_{rmn}+\frac{1}{B_m}\right)(b_{rmn'}+1/N_m)}{(b_{rm}+2)(b_{rm}+1)}\ \textit{if}\quad n\neq n'$$

with b_{rmn} is the number of alleles n sampled at locus m in population r, b_{rm} equals the number of gene copies sampled at locus m in population r, N_m is the total number of alleles observed in the whole set of groups at locus m.

Download English Version:

https://daneshyari.com/en/article/6462920

Download Persian Version:

https://daneshyari.com/article/6462920

Daneshyari.com