Forensic Science International: Genetics 28 (2017) 128–137

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Paternity testing and other inference about relationships from DNA mixtures

Peter J. Green [a,b,*], Julia Mortera [c]

[a] UTS, Sydney, Australia
[b] University of Bristol, UK
[c] Università Roma Tre, Italy

ARTICLE INFO

Article history:
Received 7 October 2016
Received in revised form 11 January 2017
Accepted 2 February 2017
Available online 16 February 2017

Keywords:
Bayesian networks
Coancestry
Deconvolution
Disputed paternity
Identity by descent
Kinship
Likelihood ratio

ABSTRACT

We present methods for inference about relationships between contributors to a DNA mixture and other individuals of known genotype: a basic example would be testing whether a contributor to a mixture is the father of a child of known genotype. The evidence for such a relationship is evaluated as the likelihood ratio for the specified relationship versus the alternative that there is no relationship. We analyse real casework examples from a criminal case and a disputed paternity case; in both examples part of the evidence was from a DNA mixture. DNA samples are of varying quality and therefore present challenging problems in interpretation. Our methods are based on a recent statistical model for DNA mixtures, in which a Bayesian network (BN) is used as a computational device; the present work builds on that approach, but makes more explicit use of the BN in the modelling. The R code for the analyses presented is freely available as supplementary material.

We show how additional information of specific genotypes relevant to the relationship under analysis greatly strengthens the resulting inference. We find that taking full account of the uncertainty inherent in a DNA mixture can yield likelihood ratios very close to what one would obtain if we had a single source DNA profile. Furthermore, the methods can be readily extended to analyse different scenarios as our methods are not limited to the particular genotyping kits used in the examples, to the allele frequency databases used, to the numbers of contributors assumed, to the number of traces analysed simultaneously, nor to the specific hypotheses tested.

1. Introduction

This paper presents methods for inference about the relationships between contributors to a DNA mixture with unknown genotype and other individuals of known genotype: a basic example would be testing whether a contributor to a mixture is the father of a child of known genotype (or indeed the similar question with the roles of parent and child reversed). Following commonly accepted practice, the evidence for such a relationship is presented as the likelihood ratio for the specified relationship versus the baseline, null hypothesis, that there is no relationship at all, so the father is taken to be a random member of the population. Our methods are based on the statistical model for DNA mixtures of [3], in which a Bayesian network (BN) is used as a computational device for efficiently computing likelihoods; the present work builds on that approach, but makes more explicit use of the BN in the modelling.

Other questions that can be answered by a similar approach include

- is a contributor to a mixture the brother of an individual of known genotype?
- is a contributor to a mixture the niece of an individual of known genotype *and* the great-aunt of another individual of known genotype?
- is a contributor to one mixture also a contributor to another mixture?
- is a contributor to one mixture a brother of a contributor to another mixture?
- is an individual of known genotype a family relative of two contributors to a mixture who are mother and child?

* Corresponding author at: School of Mathematics, University of Bristol, Bristol BS8 1TW, UK.
E-mail addresses: P.J.Green@bristol.ac.uk (P.J. Green), julia.mortera@uniroma3.it (J. Mortera).

http://dx.doi.org/10.1016/j.fsigen.2017.02.001
1872-4973/

A standard DNA paternity test compares the DNA profile of a putative father to that of his alleged child; the DNA profile of the mother might or might not be available. The case we report here (see Section 2) is one of disputed inheritance. The putative father died over 20 years ago and his corpse was exhumed in order to extract his DNA profile. The DNA extracted from the exhumed body sample was contaminated and appeared to be a mixture of at least two individuals. Furthermore, the DNA of the child's mother was not available. A preliminary analysis of this case was given in [15]. In that paper an approximate method based only the most probable genotype of a mixture contributor was used to specify the questioned relationship. Here we take all uncertainty about the mixture contributors into account.

Throughout the paper, our emphasis is on methodology. Real casework examples are presented, for illustration, but our methods are not limited to particular details of the genotyping kits, allele frequencies, number of contributors, or hypotheses in these examples.

The outline of the paper is as follows. After a brief description of the DNA mixture model and its modification for establishing potential relationships, we introduce the motivating example on paternity testing in Section 2. Four general methods for inference about relationships from DNA mixtures are illustrated in Section 3. Results for a real case where we assess if an alleged father of a typed actor is in the mixture are given in Section 4; results for a case where we try to identify an unknown contributor to a mixture through his potential mother's genotype are shown in Section 5. In Section 6 we illustrate a proposal for computing likelihood ratios for unions of hypotheses. Indications of the available open-source software are presented in Section 7. A general discussion and some concluding remarks are given in Section 8.

1.1. A model for DNA mixtures

We base the analysis of the DNA mixture on the model described in [3]. This model takes fully into account the peak heights and the possible artefacts, like stutter and dropout, that might occur in the DNA amplification process. We give a brief summary of the main features of the model, for further details we refer to [3]. The model is an extension of the gamma model developed in [4,5], and used in [6].

In summary, for a specific marker m and allele a, ignoring artefacts, the contribution H_{ia} from an individual i to the peak height at allele a has a gamma distribution, $H_{ia} \sim \Gamma(\rho\phi_i n_{ia},\ \eta)$, where ρ is proportional to the total amount of DNA in the mixture prior to amplification; ϕ_i denotes the *fraction* of DNA originating from individual i prior to PCR amplification, n_{ia} is the number of type a alleles for individual i; and η determines the scale. For an amplification without artefacts of one heterozygous contributor, *i.e.* ϕ_1=1 and $n_{1,a}$ = 1, $\mu = \rho\eta$ is the mean peak height and $\sigma = 1/\sqrt{\rho}$ is the coefficient of variation. In the following we use this reparametrization. The model is extended to take into account artefacts: stutter, whereby a proportion of a peak belonging to allele a appears as a peak at allele $a-1$; and dropout, when alleles are not observed because the peak height is below a detection threshold C. The parameter ξ denotes the mean stutter proportion.

The evidence E consists of the peak heights $\mathbf{z}$ as observed in the electropherograms, as well as any potential genotypes of known individuals. For given genotypes of the contributors, expressed as allele counts $\mathbf{n} = (n_{ia},\ i = 1, \ldots I;\ a = 1, \ldots, A)$, given proportions ϕ, and given values of the parameters $(\rho,\ \xi,\ \eta)$, all observed peak heights are independent and for a given hypothesis $\mathcal{H}$, the full likelihood is obtained by summing over all possible combinations of genotypes $\mathbf{n}$ with probabilities $P(\mathbf{n} \mid \mathcal{H})$ associated with $\mathcal{H}$:

$$L(\mathcal{H}) = Pr(E \mid \mathcal{H}) = \sum_{\mathbf{n}} L(\rho, \xi, \phi, \eta \mid \mathbf{z}, \mathbf{n}) P(\mathbf{n} \mid \mathcal{H}),$$

where

$$L(\rho, \xi, \phi, \eta \mid \mathbf{z}, \mathbf{n}) = \prod_m \prod_a L_{ma}(z_{ma})$$

and

$$L_{ma}(z_{ma}) = \begin{cases} g\{z_{ma};\ \rho D_a(\phi, \xi, \mathbf{n}), \eta\} & \text{if } z_{ma} \geq C \\ G\{C;\ \rho D_a(\phi, \xi, \mathbf{n}), \eta\} & \text{otherwise,} \end{cases} \quad (1)$$

with g and G denoting the gamma density and cumulative distribution function respectively, and D_a the effective allele counts after stutter. See [3] for full details: we use their notation above.

The number of terms in this sum is huge for a hypothesis which involves several unknown contributors to the mixture, but can be calculated efficiently by Bayesian network techniques that represent the genotypes using a Markovian structure, the allele counts for each individual being modelled sequentially over the alleles. The maximum likelihood estimate (MLE) parameters are obtained using the `R` package `DNAmixtures` [9] which interfaces to the HUGIN API (Hugin Expert A/S, 2012) through the `R` package `RHugin` [13].

In this paper we follow [3] in estimating parameters by maximum likelihood. In all computations of likelihood ratios, parameters in both numerator and denominator are fixed at the MLEs under the null hypothesis. Other choices are possible, depending on the demands of the legal environment, for example the likelihoods in the numerator and denominator could be separately maximised over values of the parameters; this would entail some additional computation.

1.2. Relationship inference with DNA mixtures

In this work we wish to establish whether one (or more) contributors to the DNA mixture has a potential relationship with one or more individuals whose genotypes are known and who have a known relationship to each other. To do this, we make more explicit use of the BN used as a computational device in [3].

This network represents the probabilistic dependence of the peak heights $\mathbf{z}$ on the allele counts $\mathbf{n}$ for the unknown contributors to the mixture, and the parameters (ϕ, ρ, ξ, η) of the gamma model. This dependence is represented in the right hand part of the directed acyclic graph in Fig. 1.

Our general strategy is to modify the Bayes net formulation of the model of [3], in ways described in the following sections, and then, as in that earlier paper, perform the necessary computations to deliver the required likelihood ratios, as laid out by [10], appropriately generalised. More details on this are given in the Appendix.

2. Motivating example: paternity testing

2.1. A case study

We now illustrate a real case from the Forensic Institute, Sapienza Università Roma, which provides the motivating example for this paper.

A man B, met a young lady C and began a secret relationship. One of C's sons A, learns as an adult that he is not the son of C's husband but probably B's son. Some years after B's death, A claims his share of B's substantial inheritance. After his mother's death and over 20 years after B's death, B's body is exhumed and DNA is extracted from a bone. This is to be used to establish whether A could be the son of B.