Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International: Genetics 2 (2008) 9–18
FSI GENETICS
www.elsevier.com/locate/fsig

Estimating mutation rates from paternity casework

P. Vicard [a,*], A.P. Dawid [b], J. Mortera [a], S.L. Lauritzen [c]

[a] *Dipartimento diEconomia, Università Roma Tre, Via Silvio D'Amico 77, Roma 00145, Italy*
[b] *University of Cambridge, United Kingdom*
[c] *University of Oxford, United Kingdom*

Received 5 July 2007; accepted 19 July 2007

Abstract

We present a statistical methodology for making inferences about mutation rates from paternity casework. This takes account of a number of sources of potential bias, including hidden mutation, incomplete family triplets, uncertain paternity status and differing maternal and paternal mutation rates, while allowing a wide variety of mutation models. An object-oriented Bayesian network is used to facilitate computation of the likelihood function for the mutation parameters. This can process either full or summary genotypic information, both from complete putative father–mother–child triplets and from defective cases where only the child and one of its parents are observed. We use a dataset from paternity casework to illustrate the effects on inferences about mutation parameters of various types of biases and the mutation model assumed. In particular, we show that there can be relevant information in cases of unconfirmed paternity, and that excluding these, as has generally been done, can lead to biased conclusions.

Keywords: DNA profile; Hidden mutation; Likelihood function; Mutation models; Object-oriented Bayesian network; Uncertain paternity

1. Introduction

DNA testing is often conducted to resolve a disputed attribution of paternity. However, when this results in an *incompatibility* – values of a child's genotype and those of its presumed parents at some forensic marker locus that appear inconsistent with simple genetic segregation – the obvious interpretation of this as due to *non-paternity* is clouded by an alternative possibility that it is in fact due to *mutation*. Dawid et al. [1] gave formulae for, and illustrated the effect of, allowing for mutation when conducting paternity analyses. Bayesian networks to conduct such analyses, for assumed values of the mutation rates, were presented in Refs. [2,3]. Since the conclusions are very sensitive to the values assumed, it is important to have good estimates of these rates.

Mutation rates for the STR markers used in forensic paternity tests range from around 5×10^{-4} to 7×10^{-3} per generation [4–6]. Such a rate is commonly estimated by the observed frequency, s/n, of inferred mutation at that marker in casework triplets, where n is the total number of meioses, and s the number of these deemed to be mutations. However, this naïve estimate can be very misleading. The aim of this paper is to show how one can estimate mutation rates while correctly accounting for uncertain paternity – an unavoidable feature of data collected at forensic laboratories – and other sources of bias [4,7,8]. Once we have good mutation rate estimates, these can be used for many purposes; in particular, they can be fed back into individual paternity case analyses.

We encapsulate our statistical inferences in the *likelihood function* for the unknown parameters based on the data analysed. The *maximum likelihood estimate* can then readily be calculated, while the spread of the likelihood around this point indicates the appropriate uncertainty to be attached to it. Alternatively, the likelihood function can be combined, by Bayes's theorem, with a prior distribution for the parameters based on external information, to yield the posterior distribution: a full probabilistic description of the remaining uncertainty after taking the new data into account. Our technical approach is two-pronged, based both on developing algebraic expressions, and on the construction and numerical analysis of an object-oriented Bayesian network (OOBN) model [3].

* Corresponding author. Tel.: +39 0657335684; fax: +39 0657335771.
E-mail addresses: vicard@uniroma3.it (P. Vicard), apd@statslab.cam.ac.uk (A.P. Dawid), mortera@uniroma3.it (J. Mortera), steffen@stats.ox.ac.uk (S.L. Lauritzen).
URL: http://tinyurl.com/d3vgz, http://tinyurl.com/2675a, http://tinyurl.com/cxhw8, http://tinyurl.com/yu8ns2

1872-4973/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2007.07.002

Throughout this paper, for simplicity, we assume both Hardy–Weinberg and linkage equilibrium, *i.e.* independence of an individual's genes both within and across markers, so that each DNA marker in the profile may be handled separately. We further restrict attention to the case that all unrelated founder individuals considered can be regarded as having DNA profiles drawn independently from a common randomly mating population with known allele frequencies.

The paper is organised as follows. In Section 2 we describe the paternity casework data used for mutation estimation, and the biases that need to be accounted for. In Section 3 we consider some models of inter-allelic mutation transitions. In Section 4 we study the algebraic structure of the likelihood function. In Section 5 we present a Bayesian network model to analyse single-marker data for any single case. For illustrative purposes and to highlight the principal issues affecting estimation of a mutation rate, in Section 6 we analyse a dataset from a paternity laboratory. We conclude with some cautionary comments on the construction and analysis of such datasets. In particular, cases with non-negligible uncertainty about paternity are often omitted from the dataset, but this can lead to biased estimates. Instead, such cases should be retained and subjected to appropriate analysis.

For fuller details and justification of the methodologies and analyses presented here, see Ref. [9].

2. Background

The DNA STR markers used for forensic purposes are particularly prone to mutation. Data on inter-allelic mutational transitions are very sparse since, in the datasets collected at forensic laboratories and used to assess mutation rates, the number of meioses where mutation appears plausible is typically very small. In order to have any chance of estimating allele-specific mutation rates from such data, we need to construct models expressing them in terms of a small number of parameters. The main thrust of this work is to estimate such mutational parameters, and thereby the overall mutation rate.

2.1. Paternity casework

Data collected by paternity testing laboratories mainly consist of DNA profiles for mother–putative father–child triplets, as well as some "defective" cases where one of the putative parents has not been profiled. For each putative family, the data at a given marker can be: *either*

Compatible: For example, mother (5, 5), putative father (5, 12), and child (5, 12). This can be explained by paternity and Mendelian segregation, with the mother handing down either of her 5's, and the father his 12.
or
Incompatible: For example, mother (5, 5), putative father (5, 12) and child (5, 8). This cannot be explained as above, since there would then be no source for the child's 8.

2.2. Biases

Among the features that should be properly accounted for when using paternity cases to estimate mutation rates are: hidden mutation [7]; differential mutation [4] and unknown paternity status. Previous work accounting for all these features [8] required various simplifying assumptions and approximations. In this paper we relax or remove these. In particular:

(1) We can analyse detailed genotype data, with case-specific prior probabilities.
(2) Defective cases, *i.e.* mother–child and putative father–child pairs, can be included.
(3) General mutation models can be used.

As will be seen in Section 6, a further source of bias not accounted for in Ref. [8], but having possibly dramatic effects on mutation rate estimates, is *preselection*: cases regarded as of insufficiently certain paternity may have been excluded from the database, leading to loss of information and biased conclusions. Wherever possible complete databases should be used, but subjected to an appropriate analysis, such as described here, that takes full account of the uncertainty over paternity.

3. Mutation models

For a given marker, let $q_{i\to j}$ $(j\neq i)$ denote the probability that allele i mutates to allele j in a parent–child transmission. The probability that allele i is transmitted unmutated is thus $q_{i\to i} := 1-\sum_{j\neq i} q_{i\to j}$, and the overall mutation rate is

$$\mu := \sum_{(i,j):i\neq j} p_i\, q_{i\to j} = 1-\sum_i p_i\, q_{i\to i},$$

where p_i denotes the population frequency of allele i.

Various models describing the allele-specific mutation rates $q_{i\to j}$ in terms of a small set of adjustable parameters have been proposed [1,2,10–13]. The estimate of the overall mutation rate μ can be sensitive to the particular mutation model assumed for the inter-allelic transitions, which could in turn affect individual case paternity analyses. It is therefore important that the statistical methodology used should be able to analyse a range of alternative plausible models.

3.1. Scalar models

A very flexible class of mutation models comprises the *scalar mutation models* [8], having the general form:

$$q_{i\to j} = \lambda\, s_{i\to j} \quad (j\neq i)$$

where $\lambda \geq 0$ is an unknown scale parameter to be estimated and $S := (s_{i\to j})$ is a specified transition matrix, *i.e.* $s_{i\to j}\geq 0$ and, for each i, $\sum_j s_{i\to j} = 1$. The overall mutation rate per transmission is then

$$\mu = \kappa\,\lambda \tag{1}$$

where

$$\kappa := 1 - \sum_i p_i s_{i\to i}. \tag{2}$$

3.2. Parametrisations

Assume a common scalar mutation model in both the maternal and paternal lines, but with possibly differing line-specific scale parameters, λ_M and λ_P, respectively. The line-specific mutation rates are thus $\mu_M = \kappa\lambda_M$, $\mu_P = \kappa\lambda_P$, with κ given by (2). We introduce $\tau := \mu_M + \mu_P$, the *total mutation rate*, and $\rho := \mu_P/\tau$, the *paternal fraction* of the total mutation rate. We thus have $\mu_P = \rho\tau$ and $\mu_M = (1-\rho)\tau$. We further introduce the *total mutation parameter* $\xi := \lambda_P + \lambda_M$. Then

$$\tau = \kappa\xi \tag{3}$$

and

$$\rho = \frac{\lambda_P}{\xi}. \tag{4}$$

In this paper we treat ρ as known, and regard the total mutation rate τ, or equivalently ξ, as the parameter of principal interest (note that $\lambda = \xi/2$ was used in Ref. [14]).

4. Likelihood function

We consider mother–putative father–child triplets, on each of which we have full or partial information (''findings'') about their DNA profiles. We suppose that we have a collection of such cases randomly drawn from the relevant population. For any case, full information at any given marker would comprise the genotypes of all three individuals for that marker. Partial information comprises both defective cases, in which the data on one of the ''parents'' are missing, and/or summary data, for example a simple report as to compatibility or otherwise.

4.1. General structure

We wish to use all the available data to make inferences about the mutation rates of the various markers, allowing in particular for the possibility of non-paternity. In Ref. [9, Appendix] it is shown that, under reasonable assumptions, estimation for each marker can be carried out separately. Therefore from now on we focus on a single marker.

Suppose we have a model – not necessarily scalar – describing the mutation process in terms of a parameter ξ. The overall likelihood function for ξ can be computed as the product of the individual case likelihood functions. Considering a single such case, with information f for that marker, it is shown in Ref. [9] that its contribution to the likelihood for ξ has the form:

$$\ell(\xi) \propto \pi^* \,\mathrm{pr}(f|\xi, P) + (1-\pi^*)\,\mathrm{pr}(f|\xi, \bar{P}), \tag{5}$$

where P [resp. $\bar{P}$] denotes paternity [resp. non-paternity]; $\mathrm{pr}(f|\xi, P)$ [resp. $\mathrm{pr}(f|\xi, \bar{P})$] is the probability, assuming paternity [resp. non-paternity], that for this case we would obtain the information f on that marker, when the value of the mutation parameter is ξ and π^* is the ''prior'' probability of paternity for this case, calculated taking into account the findings on all the *other* DNA markers in the profile assuming no further mutations (after incorporating any relevant external evidence into the initial probability of paternity before any DNA evidence): this may be obtained from standard formulae [15].

Finally we multiply all such likelihood contributions across the different families in the dataset to obtain the overall likelihood function for ξ.

4.2. Algebraic form

One way of proceeding is to develop algebraic formulae for the terms $\mathrm{pr}(f|\xi, P)$ and $\mathrm{pr}(f|\xi, \bar{P})$ in (5).

Suppose that we have full triplet genotype data: mother's genotype *AB*, putative father genotype *CD*, and child's genotype *EF* (here *A, B, C, D, E, F* are arbitrary and any of them could be identical). Let $q^M_{i\to j}$ denote the mutation transition rate from allele i to allele j for the maternal line, and $q^P_{i\to j}$ that for the paternal line. Then under paternity, P, we have

$$\begin{aligned}\mathrm{pr}(f|\xi, P) &\propto \mathrm{pr}(EF|AB, CD; P)\\ &\propto (q^M_{A\to E} + q^M_{B\to E})(q^P_{C\to F} + q^P_{D\to F})\\ &\quad + (q^M_{A\to F} + q^M_{B\to F})(q^P_{C\to E} + q^P_{D\to E}).\end{aligned} \tag{6}$$

Formula (6) has been used [16] to develop a likelihood analysis and comparison of various mutation models on the assumption of paternity in all cases.

Since $\sum_j q_{i\to j} = 1$, apart from highly exceptional cases where neither of the child's alleles agrees with any of those of its ''parents'', to a very good approximation we can treat (6) as a linear function of the small q's.

Under non-paternity, $\bar{P}$,

$$\begin{aligned}\mathrm{pr}(f|\xi, \bar{P}) &\propto \mathrm{pr}(EF|AB, CD; \bar{P})\\ &\propto 2\{(q^M_{A\to E} + q^M_{B\to E})\,p_F + (q^M_{A\to F} + q^M_{B\to F})\,p_E\}.\end{aligned} \tag{7}$$

Expression (7) is a linear function of $(q^M_{i\to j} : i \neq j)$, the constant term being non-zero except for a case of maternal incompatibility (when neither of the child's alleles is present in the mother).

The omitted constant of proportionality is the same in both (6) and (7), see [9]. The likelihood contribution $\ell(\xi)$ for this case can thus be calculated by substituting (6) and (7) into (5).

4.3. Scalar model

Henceforth we restrict attention to scalar mutation models. We have $q^M_{i\to j} = \lambda_M s_{i\to j}$ $(i \neq j)$, $q^M_{i\to i} = 1 - \lambda_M(1 - s_{i\to i})$, and similarly for the paternal line. Expressing $\lambda_P = \rho\xi$, $\lambda_M = (1-\rho)\xi$, it is readily seen that, as a function of ξ for fixed ρ, for any findings f on a case, $\ell(\xi)$ is exactly quadratic, and very close to linear, in ξ.[1]

[1] The quadratic term in $\ell(\xi)$ is entirely due to considering the possibility of two simultaneous mutations—a very rare event.

4.4. Incompatible cases

For an incompatible case $\ell(\xi)$ will typically be increasing in ξ. Using (3) and rescaling, we then have an expression for the likelihood contribution, in terms of the total mutation rate τ, of the form

$$\ell(\tau) \propto a + \tau \tag{8}$$

where $a > 0$ determines the behaviour of the likelihood contribution for this case.

4.4.1. Influence of the constant

Plausible values of τ typically range from 0.002 to 0.01. When the intercept $a > 0.02$, the likelihood contribution (8) is effectively constant in this region, and the case can thus be discarded as essentially uninformative about τ. On the other hand, when $a < 0.001$, $\ell(\tau)$ can be well approximated by τ, equivalent to taking paternity as confirmed. These are the only two options that have typically been considered in previous analyses. However, in the intermediate range, $0.001 < a < 0.02$, the function $\ell(\tau)$ is not well approximated either by a constant or by τ. Then, rather than either discard the case or assume paternity, we need to compute a and use the correct likelihood contribution $\ell(\tau) \approx a + \tau$.

4.5. Compatible cases

For any marker the vast majority of cases considered will be compatible (an attribute henceforth denoted by ''Comp''). Using the above algebraic approach we could in principle obtain, for each such case, the likelihood function based on its full genotype data. However, this detailed analysis may be impossible or inconvenient, whether because the full data are unavailable, or simply because of the large number of cases to be analysed. Instead we could use, as the finding on each such case, just the relevant summary ''compatibility'' property, as described in Ref. [9], and compute a likelihood contribution based on this summary finding alone.

There are three types of compatible case to consider.

4.5.1. Fully compatible triplets

A large number, n_{FC} say, of the compatible cases will be triplets that are *fully compatible*, *i.e.* compatible on all markers. For a general scalar model, the summary likelihood contribution $\ell(\xi) = \mathrm{pr}(\mathrm{Comp}|\xi)$ from such a case, being a sum of terms (one for each possible compatible configuration) each of which is very close to linear in ξ, will itself be very close to linear. Moreover, the appropriate ''prior probability'' π^* in (5), although varying from case to case in the light of the findings on the other markers (and other evidence), will always be near 1, and can to a good approximation be taken to be 1. Thus, to a very good approximation we have:

$$\ell(\xi) \propto \mathrm{pr}(\mathrm{Comp}|\xi, P) \propto 1 - \alpha\xi \tag{9}$$

for a suitable value of α—which is, however, difficult to obtain by algebraic methods.

Equivalently,

$$\ell(\tau) \propto 1 - \alpha'\tau \tag{10}$$

for $\alpha' := \alpha/\kappa$ with κ given by (2). We note ([8] Section 5) that $\mathrm{pr}(\mathrm{Comp}|\xi, P)$ does not depend on ρ, and hence neither does α'.

If there are n_{FC} fully compatible cases, the overall likelihood contribution from the summary data ''full compatibility'' on these cases is thus

$$L_{\mathrm{FC}}(\tau) \approx (1 - \alpha'\tau)^{n_{\mathrm{FC}}}. \tag{11}$$

4.5.2. Locally compatible triplets

A further (typically small) number n_{LC} of the cases compatible on this marker will be only *locally compatible*, *i.e.* incompatible on one or more of the other markers. Then π^* is essentially 0, and the summary likelihood contribution is, to a very good approximation,

$$\ell(\tau) \propto \mathrm{pr}(\mathrm{Comp}|\tau, \bar{P}). \tag{12}$$

Clearly $\mathrm{pr}(\mathrm{Comp}|\tau, \bar{P})$ only depends on the maternal mutation rate, $\mu_M = (1 - \rho)\tau$, and will be of the form $\mathrm{pr}(\mathrm{Comp}|\tau, \bar{P}) \propto 1 - \beta'(1 - \rho)\,\tau$. Again, β' is not readily determined by algebraic methods.[2]

The total likelihood contribution from the summary data on these locally compatible cases is

$$L_{\mathrm{LC}}(\tau) \approx \{1 - \beta'(1 - \rho)\,\tau\}^{n_{\mathrm{LC}}}. \tag{13}$$

There is thus some information about τ in locally compatible cases, although if, as is typical, $n_{\mathrm{LC}} \ll n_{\mathrm{FC}}$, this will be of a smaller order of magnitude than that in fully compatible cases.

4.5.3. Compatible pairs

We can conduct similar analyses for cases of mother–child or putative father–child compatible pairs.

For a mother–child case we can without loss of generality always assume non-paternity, generating the approximate likelihood contribution (accounting for possible *hidden mutation*, *i.e.* one not resulting in incompatibility):

$$\ell(\tau) \propto 1 - (1 - \rho)\,\gamma'\,\tau, \tag{14}$$

for a suitable value of γ', this linear approximation again being excellent.

For a putative father–child case, if there is also an incompatibility at some other marker we can assume non-paternity—but since that now leaves only the data on the child, which is clearly uninformative about mutation, such cases can be discarded. Otherwise, there being no other incompatibility, it will again usually be an acceptable approximation to assume paternity. This case is then formally identical to a mother–child pair except for the interchange of maternal and paternal

[2] Note that the algebraic analysis of Ref. [8] for this case was incorrect, and its subsequent results are thus potentially inaccurate, since it applied an approximation valid only for small values of the parameter λ to the case $\lambda = 1$.

mutation rates. It will therefore generate a likelihood contribution that can be excellently approximated by

$$\ell(\tau) \propto 1 - \rho\,\gamma'\,\tau, \tag{15}$$

with the same γ' as in (14).

5. Object-oriented Bayesian network for mutation rate estimation

The algebraic approach is not well suited to the analysis of incomplete information, as in Section 4.5; for example, simple expressions for the constants α', β', γ' are generally unavailable. To handle such cases we have explored an alternative route, involving the construction and application of a computer software system to calculate the appropriate likelihood expression numerically. This was originally built as a regular Bayesian network (BN) [9], and more recently reconfigured as an *object-oriented Bayesian network* (OOBN). The use of OOBN technology to simplify the requisite tasks of network specification and construction is explained in Ref. [14].

The same BN can be used to analyse any information, whether complete or incomplete. In addition, the OOBN interface makes it straightforward to extend such a system to allow for additional complexities, such as laboratory errors, silent alleles, *etc.* [3]. An OOBN extended to handle non-stationary mutation models and incomplete cases is described in Ref. [17].

Fig. 1 shows the top level network representing the mutation rate estimation problem for a single marker. For detailed descriptions of all our OOBN networks, see Refs. [3,14,17].

Here **m** (for the mother), **pf** (for the putative father) and **af** (for the alternative father) are founders. The true father is represented by means of a query node **tf**, choosing either **pf** or **af** according to the value of the Boolean hypothesis node **tf = pf**? Node **c** represents the child: its internal structure incorporates Mendel's law and the mutation model used. We have built and used network modules implementing the *mixed mutation model*, including its special case the *proportional mutation model* [8]. However, our methods can readily be modified to handle other (not necessarily scalar) mutation models as desired.

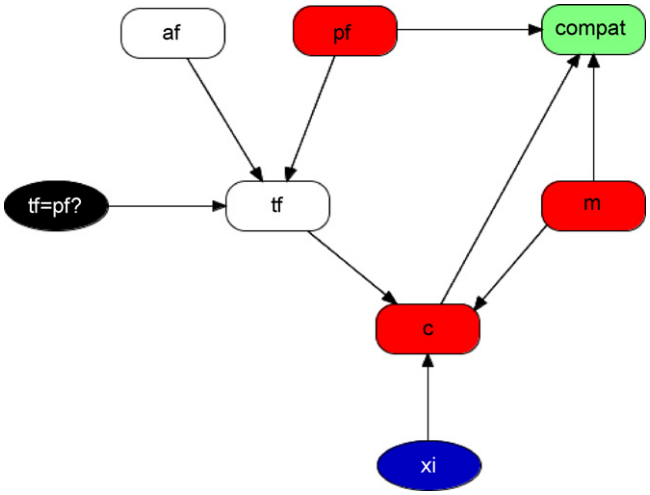

Fig. 1. Object-oriented Bayesian network to assess mutation rate: top level network.

Node **xi** is the total mutation parameter ξ; we gave it 43 possible states ranging from 0 to 0.0198, more concentrated around the mutation rate values reported in the literature. Node *compat* represents an additional module necessary for analysing summary data on compatibility of complete and defective (putative father–child and mother–child pairs) cases as described in Ref. [14].

5.1. Using the network

5.1.1. Computation of the likelihood

When implemented in appropriate software,[3] the OOBN enables numerical calculation of the contribution $\ell(\xi)$ to the likelihood function for the mutation parameter ξ, based on the (complete or incomplete) information on any single case.

We first set the parameters ρ, the paternal fraction of the mutation rate, and h, the mixing parameter for the mixed model,[4] and specify whether the case is a triplet or a pair. We next enter the findings (complete or incomplete) on the case. If we have data on the putative father, we also set the prior probability π of paternity. We now propagate, using the software, and interrogate the node **xi** to obtain the posterior probabilities for its various states. Since we have set ξ to be uniformly distributed *a priori*, these posterior probabilities will be proportional to the desired likelihood function $\ell(\xi)$, based on this case, evaluated at these points. The overall likelihood at these points is now obtained by multiplication across all cases. (This can also be interpreted as the likelihood for τ, on making the substitution $\xi = \tau/\kappa$.)

Unlike the analysis of Ref. [8], this analysis allows for simultaneous mutation in both the paternal and the maternal germline. However, it does not directly yield the continuous likelihood function, but only its values at the discrete set of values chosen as states for node **xi** in the network. We therefore fit a curve to these points to obtain, either exactly or approximately, the continuous likelihood contribution $\ell(\xi)$. This is particularly straightforward for a scalar model, since, according to Section 4.3, $\ell(\xi)$ is exactly quadratic, and very close to linear.

5.1.2. Incompatible cases

The number of incompatible cases will typically be very small, but influential on the analysis. It is important to use full genotype evidence for these. The associated likelihood contributions can be obtained using either the network, as described above, or the algebraic approach of Section 4.

6. Illustrative data analysis

We illustrate our methodology by applying it to a set of data collected at the Institut für Rechtsmedizin and provided to us by Professor Bernd Brinkmann. We use these data only to show the

[3] We used Hugin Version 6, available from http://www.hugin.com.

[4] ρ and h are contained in a subnetwork (not visible in Fig. 1) of **c**.

Table 1
Population gene frequencies for marker vWA

Allele	12	13	14	15	16	17	18	19	20	21	22
Frequency	0.0003	0.0018	0.1009	0.1004	0.1949	0.2834	0.2162	0.0866	0.0137	.0015	.0003

possible biasing effects of the standard procedures applied in the forensic laboratories: consequently our data analysis must be taken as illustrative only, and not used to support any specific numerical values for mutation rates.

We focus attention on estimation of the total mutation rate τ for the STR marker vWA. The allele frequencies for this marker in the reference population are given in Table 1. We use the mixed mutation model, considering values 0 (proportional model), 0.5, 0.9 and 1 (single-step mutation model, SMM) for the mixing parameter h.

6.1. Data and assumptions

The dataset comprises 2013 meioses. Information about these cases is largely taken from Ref. [4], with some variations and additional information for incompatible cases supplied by Brinkmann's laboratory.

The total number of meioses is odd because some of the compatible cases are deficient, and whereas each triplet contributes two meioses, a pair contributes only one. In the naïve analysis, each meiosis, be it associated with a triplet or a pair, is taken as contributing the same information to the estimation of the mutation rate. In fact different compositions of the meioses in terms of triplets and pairs will affect the mutation rate estimate differently. In the dataset analysed here 4% of meioses are from deficient cases. Therefore, we assume that we have 8 meioses from 4 triplets incompatible for vWA, 1924 meioses from 962 compatible triplets, and a further 81 meioses from 81 compatible pairs. Among the 962 triplets compatible for vWA, 943 are fully compatible, and 19 only locally compatible. According to Ref. [4], the overall ratio of maternal to paternal meioses is approximately 1: we therefore assume that there are 40 mother–child pairs and 41 putative father–child pairs, all these cases being fully compatible.

We have used full genotype data for the four incompatible cases. We did not have access to the detailed genotype data on the compatible cases. We have thus based our analysis of these on summary information only, as described in Section 4.5.

6.2. Preselection

Traditionally, mutation analysis of paternity casework data has been preceded by a preliminary *preselection* stage, where only those cases for which paternity is regarded as sufficiently firmly established are retained for further analysis. Then, at the analysis stage, each retained incompatible case is regarded as contributing one mutation event (possibly further classified as a paternal or a maternal mutation), while compatibility is taken as evidence of no mutation. If paternal and maternal mutation rates are assumed to be equal, a naïve estimate of their common value is given by the ratio of the number of incompatibilities to the total number of meioses.

The cases supplied to us had already been preselected by the laboratory as those for which paternity was regarded as established with very high probability. As we show below, such preselection of cases can lead to biases in estimation. To illustrate this, we henceforth proceed as if our data had not been preselected, but consider the effect of further selection of cases.

We compute, for each of the four incompatible cases, the probability π^* of paternity, after taking account of the data on the remaining markers: this can be calculated using the formula of Ref. [15]. However, in order to illustrate the possible biasing effects of preselection we have used only a subset of the other markers, leading to the values given in Table 2. Using 0.999 as a threshold for the probability of paternity, for the cases as reported in Table 2, Case 4 would have been discarded at the preselection stage.

Having discarded Case 4, three incompatibilities are left from a total of 2011 meioses. The naïve estimate of the combined mutation rate τ is thus $2 \times 3/2011 = 0.003$. We shall see below that taking proper account of the information in the discarded Case 4, *i.e.* not applying the preselection process, would lead to estimates closer to 0.004. If used in paternity analysis of a new case of *prima facie* exclusion, such a difference of around 30% in the estimated mutation rate would lead to a similar percentage difference in the calculated paternity index [1].

6.3. Compatible cases

Compatible cases carry information on hidden mutation. The approximate likelihood contribution from the summary information on compatible triplets is computed using formulae (11) for the fully compatible cases and (12) for the locally compatible cases, while that from the compatible pairs is obtained from (14) for mother–child cases and (15) for fully compatible putative father–child cases. The overall approximate likelihood term for all compatible cases is thus

$$(1-\alpha'\tau)^{943}\{1-\beta'(1-\rho)\tau\}^{19}\{1-\gamma'(1-\rho)\tau\}^{40}(1-\gamma'\rho\tau)^{41}. \tag{16}$$

The parameter α in (9) is evaluated by first computing $\mathrm{pr}(\mathrm{Comp}|\xi, P)$ from the network, simultaneously for each of the values used for ξ, and fitting a linear relationship to these values. Then α' is obtained as α/κ. The parameters β' and γ' are

Table 2
Prior probability π^* of paternity in the four incompatible cases

	Case 1	Case 2	Case 3	Case 4
π^*	0.99984	0.99959	0.99920	0.99612

Table 3
Coefficients α', β' and γ' for different values of h

	h			
	1	0.9	0.5	0
α'	0.790	0.788	0.780	0.768
β'	0.620	0.616	0.607	0.590
γ'	0.561	0.561	0.557	0.548

obtained similarly. Table 3 shows the coefficients α', β' and γ' as functions of the mixing parameter h.[5] We see that the dependence on the specific mutation model, determined by h, is essentially ignorable.

For our further analysis we set the parameter ρ at 0.5. The overall likelihood function from compatible cases then becomes

$$(1-\alpha'\tau)^{943}(1-0.5\,\beta'\tau)^{19}(1-0.5\,\gamma'\tau)^{81}.$$

6.4. Incompatible cases

The detailed findings on vWA for the four incompatible cases are given in Table 4. These four triplets are compatible on all other markers.

In contrast to the compatible cases, the value π assumed for the probability of paternity affects the information about τ contained in an incompatible case. As shown in [9, Appendix] and expressed in (5), it is appropriate to use $\pi = \pi^*$, the probability of paternity based on the findings on the other markers. These values (using non-committal prior probabilities) are given in Table 2.

6.4.1. Individual cases

We first investigate, for the incompatible cases, the sensitivity of the associated likelihood contribution $\ell(\tau)$ to the assumed paternity probability π. For each incompatible case we evaluated the constant a in (8) for various values of π and h.[6] Tables 5 and 6 report the results for Cases 1 and 4; Cases 2 and 3 were very similar to Case 1.

Case 1 is a case of paternal exclusion, where either a one-step or a six-step paternal mutation could explain the incompatibility. From Table 5 we see that, for the proportional model, we have $a \leq 0.001$, effectively confirming paternity, so long as $\pi \geq 0.9995$; however $a \leq 0.02$, and thus the case is already informative about τ, as soon as π exceeds about 0.995. For the SMM and mixed models, the case is informative even for π around 0.9, although paternity is confirmed only for $\pi \geq 0.9995$. Thus, Case 1 contains non-negligible information about τ for values of π that would normally lead to rejection of this case. Since the likelihood contribution $\ell(\tau) \approx a + \tau$ increases with τ, such a rejection would result in a downward bias in the mutation rate estimate.

Case 4 is a case of ambiguous exclusion, which can be explained under paternity either by a three-step maternal mutation or by a two-step paternal mutation. Under the SMM neither of these can happen and non-paternity becomes the only possible explanation. Otherwise (see Table 6), the case is already informative for $\pi \geq 0.995$ (for the proportional model, and mixed model with $h = 0.5$) or $\pi \geq 0.9995$ (for $h = 0.9$); while paternity can be taken as established for $\pi \geq 0.9995$ (proportional model) or $\pi \geq 0.9997$ (mixed model, $h = 0.5$). However, under the mixed model with $h = 0.9$ we cannot safely proceed as if paternity were established even when $\pi = 0.9997$. Intuitively this is because this model, which is close to the SMM, gives very low probability to the observed data under paternity, so favouring the alternative explanation of non-paternity—which is much less informative about mutation.

Table 4
Findings on incompatible cases for marker vWA

	Child	Mother	Putative father
Case 1	18 20	18 19	14 19
Case 2	16 18	16 17	17 19
Case 3	18 19	17 18	16 18
Case 4	15 18	18 18	17 18

Table 5
Intercept a of the likelihood contribution $\ell(\tau) \approx a + \tau$ from incompatible Case 1 (π^*)

π	h			
	1 (SMM)	0.9	0.5	0 (Prop. model)
0.9	0.01165	0.01263	0.02002	0.18207
0.95	0.00565	0.00612	0.00969	0.08556
0.995	0.00055	0.00060	0.00094	0.00812
0.9995	0.00005	0.00006	0.00009	0.00081
0.9997	0.00003	0.00004	0.00006	0.00048
0.99984	0.00002	0.00002	0.00003	0.00026
1	0	0	0	0

6.5. Overall likelihood

We here examine the sensitivity of the overall likelihood for τ (incorporating all cases, both compatible and incompatible) to

Table 6
Intercept a of the likelihood contribution $\ell(\tau) \approx a + \tau$ from incompatible Case 4 (π^*)

π	h		
	0.9	0.5	0 (Prop. model)
0.9	4.98779	0.30287	0.12480
0.95	1.03600	0.13419	0.05778
0.995	0.06788	0.01217	0.00542
0.99612	0.05228	0.00943	0.00420
0.9995	0.00656	0.00121	0.00054
0.9997	0.00393	0.00072	0.00032
1	0	0	0

[5] The naïve approach, which simply counts each compatible triplet as exhibiting no mutation, is equivalent to taking $\alpha' = \beta' = \gamma' = 1$; departure of these coefficients from 1 is due to taking proper account of hidden mutation. It is clear from Table 3 that this effect is important and would lead to the naïve estimates being strongly biased [7,8].

[6] We used both the algebraic approach of Section 4.3 and, as a check, numerical computation in the network.

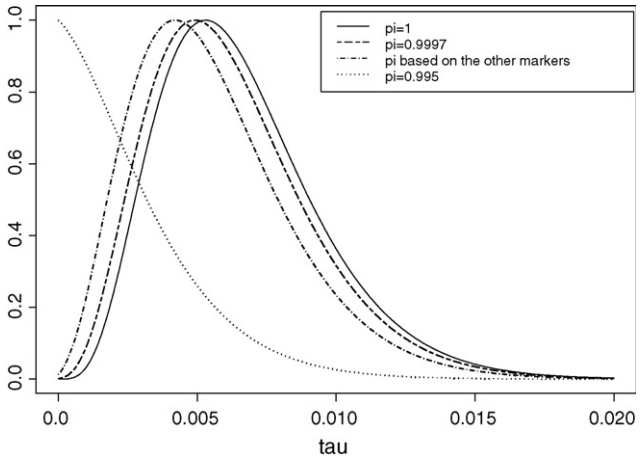

Fig. 2. Likelihood for τ, proportional model.

the mutation model and the probabilities of paternity used for the four incompatible cases. We consider the proportional ($h = 0$) and mixed ($h = 0.9$) models, together with the following choices for π:

(i) $\pi = 1$. This ignores any effect of non-paternity, although hidden mutation is correctly accounted for.
(ii) $\pi = 0.9997$. At this value paternity would typically be taken as established, further analysis proceeding as for (i) above.
(iii) $\pi = \pi^*$. This uses data on other markers to assess the relevant probability of paternity.
(iv) $\pi = 0.995$. Such cases would normally be discarded. We investigate how much information on the mutation rate they contain.

The corresponding overall likelihoods for τ are shown in Fig. 2 for the proportional model and in Fig. 3 for the mixed model. We see that, on introducing even a very small probability of non-paternity, smaller values of τ become more likely; the smaller π is, the more the curve shifts to the left. Intuitively this is because, as we decrease π, we can more easily explain away an incompatibility as due to non-paternity rather than mutation. The effect of this on the maximum likelihood estimate is shown in Table 7.

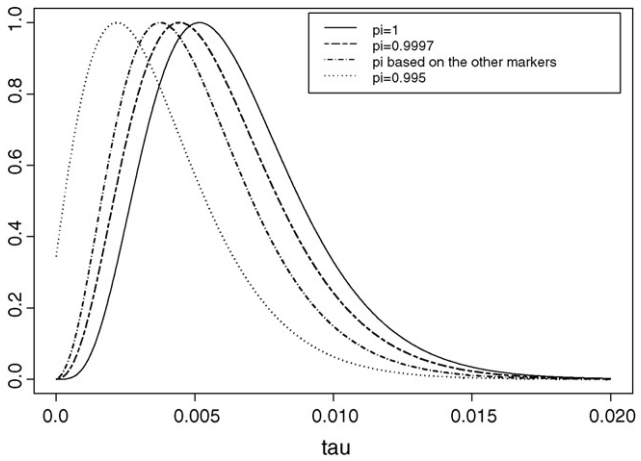

Fig. 3. Likelihood for τ, mixed ($h = 0.9$) model.

Table 7
Maximum likelihood estimates of τ for the proportional and mixed ($h = 0.9$) mutation models

π	Proportional model	Mixed model
1	0.0053	0.0052
0.9997	0.0049	0.0044
π^*	0.0042	0.0038
0.995	0	0.0022

The value of π affects the overall likelihood differently for the proportional model and the mixed model. This is essentially due to the differences between the columns for $h = 0$ and $h = 0.9$ in Table 5 for Case 1 (with similar behaviour for Cases 2 and 3) and in Table 6 for Case 4. Under either model, when $\pi = 0.9997$ the contribution from the first three cases was well approximated by τ, just as if paternity were established. However, (see Section 6.4.1), this approximation is poor for Case 4 under the mixed model with $h = 0.9$, on account of its intercept $a = 0.00393$. Allowing a very small chance that the observed incompatibility might be due to non-paternity has led to a flatter likelihood contribution. It is essentially just this effect for this one case that has caused the overall likelihood to move noticeably to the left under the mixed model.

Very similar considerations apply when we take $\pi = \pi^*$. For Case 4, the flattening effect of its low value $\pi^* = 0.9961$ is now seen even under the proportional model; while under the mixed model this case is effectively discarded as a non-paternity (paternity would require a mutation of at least two steps, which is very unlikely for this model).

When $\pi = 0.995$, under the proportional model all four incompatible cases have non-negligible intercept a. These flatter contributions result in a likelihood favouring very small values of τ. The behaviour is quite different under the mixed model. For Cases 1–3, the observed incompatibilities could be explained by one-step mutations, which are quite probable under this mutation model, to the extent that it becomes safe to assume paternity, together with mutation, for these cases; conversely Case 4, which cannot now be easily explained by mutation, can be discarded as a non-paternity.

7. Discussion, extensions, recommendations

We have described and illustrated how algebraic expressions and Bayesian networks can be used to make inferences about mutation rates from paternity casework, taking full account of a variety of logical subtleties and computational complexities, including in particular hidden mutation and uncertain paternity. We have not considered such complications as more complex pedigrees, allelic dropout, measurement error or uncertainty in the population gene frequencies. In principle these could be incorporated into a more elaborate Bayesian network description of the problem [3].

For simplicity, in our illustrative data analysis we fixed ρ, the paternal fraction of the total mutation rate, at 0.5 (so taking $\mu_M = \mu_P$). One could easily explore the sensitivity of

inferences to the value assumed for ρ.[7] Our basic framework can also be used to make inferences about ρ, treated as an unknown parameter along with τ [18]. Many more incompatible triplets are then needed to obtain a reliable estimate of ρ.[8] However, if we could assume a common value of ρ across all markers the combined information could become appreciable. Evidence could also be combined across different studies.

We have seen that mutation estimates can be quite sensitive to the mutation model assumed, expressed in our case by the value of the mixing parameter h. With a little further elaboration we could treat h as another unknown parameter, and estimate it, too, from the same data. In particular we could then compare the relative support for the SMM ($h = 1$) as against the proportional model ($h = 0$). We could similarly compare other, distinct, mutation models.

An issue deserving further investigation is how to quantify the loss of information from analysing only summary data, rather than full genotype information, for compatible cases. Although this does not lead to systematic biases, it can reduce the precision of parameter estimates, as well as affecting their values. When full data are available, the large number of cases could be analysed either numerically, using a suitable batch-processing interface to the Bayesian network, or by constructing a program to handle the algebra and related calculations as described in Section 4.

We cannot emphasise too strongly that our focus in this paper has been entirely methodological; in particular, our data analysis is intended purely to illustrate this methodology, and its results must not be taken as substantively meaningful. This is because the paternity dataset supplied to us, like most of those used for mutation estimation, was not randomly selected from casework, but instead comprised only preselected cases, where the assessed probability of paternity was very close to 1. As we have observed in our treatment of Case 4 in Section 6, there can be valuable information about mutation even in cases that would be excluded by this criterion, and ignoring this is likely to lead to biases in the estimates obtained. We therefore recommend retaining a much more complete collection of cases than usual for analysis. If it is not possible to analyse all cases (or a large random sample), then the choice of those to exclude should be based on an assessment, according to the criteria of Section 4.4.1, of the constant a in (8), for a variety of realistic assumptions about the probability of paternity and the mutation model. Taking proper account of cases where the value of the probability of paternity is below the threshold commonly considered as confirming paternity should lead to more precise estimates of both τ and ρ—which could differ notably from those based on the biased sets of cases usually analysed.

We hope that future work based on appropriate analysis of full paternity casework will provide better estimates of mutation rates on specific markers. Analysis of casework data from various reference populations could enable valuable assessment of the variation of mutation rates across populations. But, prior to any analysis, however simple or sophisticated, an appropriate data collection, selection and recording process is essential if we are to obtain credible estimates of mutation parameters.

Acknowledgments

This research was supported by a Research Interchange Grant from the Leverhulme Trust. We are indebted to our collaborators in this research programme for helpful discussions. We are also grateful to Bernd Brinkmann and colleagues for information about their data.

References

[1] A.P. Dawid, J. Mortera, V.L. Pascali, Non-fatherhood or mutation? A probabilistic approach to parental exclusion in paternity testing, Forensic Sci. Int. 124 (2001) 55–61.
[2] A.P. Dawid, J. Mortera, V.L. Pascali, D.W. van Boxel, Probabilistic expert systems for forensic inference from genetic markers, Scand. J. Stat. 29 (2002) 577–595.
[3] A.P. Dawid, J. Mortera, P. Vicard, Object-oriented Bayesian networks for complex forensic DNA profiling problems, Forensic Sci. Int. 169 (2007) 195–205.
[4] B. Brinkmann, M. Klintschar, F. Neuhuber, J. Hühne, B. Rolf, Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat, Am. J. Hum. Genet. 62 (1998) 1408–1415.
[5] L. Henke, J. Henke, Mutation rate in human microsatellites, Am. J. Hum. Genet. 64 (1999) 1473 (with reply by B. Rolf and B. Brinkmann, 1473–1474).
[6] A. Sajantila, N. Lukka, A.C. Syvanen, Experimentally observed germline mutations at human micro- and minisatellite loci, European Journal of Human Genetics 7 (1999) 263–266.
[7] R. Chakraborty, D.N. Stivers, Y. Zhong, Estimation of mutation rates from parentage exclusion data: applications to STR and VNTR loci, Mutat. Res. 354 (1996) 41–48.
[8] P. Vicard, A.P. Dawid, A statistical treatment of biases affecting the estimation of mutation rates, Mutat. Res. 547 (2004) 19–33.
[9] P. Vicard, A.P. Dawid, J. Mortera, S.L. Lauritzen, Estimation of mutation rates from paternity cases using a Bayesian network, Research Report 249, Department of Statistical Science, University College London, 2004 (http://tinyurl.com/2cjpq8).
[10] T. Ohta, M. Kimura, A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population, Genet. Res. 22 (1973) 201–204.
[11] A.M. Valdes, M. Slatkin, N.B. Freimer, Allele frequencies at microsatellite loci: the stepwise mutation model revisited, Genetics 133 (1993) 737–749.
[12] R. Durret, S. Kruglyak, A new stochastic model of microsatellite evolution, J. Appl. Probability 36 (1999) 621–631.
[13] T. Egeland, P.F. Mostad, Statistical genetics and genetical statistics: a forensic perspective, Scand. J. Stat. 29 (2002) 297–308.
[14] A.P. Dawid, An object-oriented Bayesian network for estimating mutation rates, in: C.M. Bishop, B.J. Frey (Eds.), Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key West, Florida, January 3–6, 2003 (http://tinyurl.com/39bmh).
[15] E. Essen-Möller, Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis, Theoretische Grundlagen, Mitteilungen der Anthropologischen Gesellschaft 68 (1938) 9–53.

[7] Brinkmann et al. [4] interpret their data as showing a higher mutation rate in the paternal than in the maternal germline in a ratio about 17:3, corresponding to $\rho = 0.85$.

[8] In Ref. [8] the information about ρ in the data was shown to be on a par with the information about the bias of a coin that would be obtained from tossing it k times, where k is the number of unambiguous incompatibilities observed in the data (for vWA in our dataset, $k = 2$).

[16] J.C. Whittaker, R.M. Harbord, N. Boxall, I. Mackay, G. Dawson, R.M. Sibly, Likelihood-based estimation of microsatellite mutation rates, Genetics 164 (2003) 781–787 (http://tinyurl.com/ywk7gg).
[17] P. Vicard, New statistical approaches for estimating mutation parameters, in: Atti della XLII Riunione Scientifica della Società Italiana di Statistica—volume delle sessioni plenarie e specializzate, CLEUP, pp. 417–428, 2004 (http://tinyurl.com/35nq27).
[18] S. Vorkas, Estimation of mutation rates using a Bayesian network, B.Sc. Dissertation, Department of Statistical Science, University College London, 2005.