Using all available evidence to solve kinship cases

Thore Egeland[1,2] and Franco Marsico[3]

[1]Faculty of Chemistry, Biotechnology and Food Science, NMBU, Oslo, Norway.

[2]Forensic Genetics Research Group, Dept of Forensic Sciences, OUS, Oslo, Norway.

[3]Faculty of Exact and Natural Sciences, UBA, Buenos Aires, Argentina.

Abstract

Kinship cases, ranging from standard paternity tests to complex Disaster Victim Identifications, are typically evaluated using likelihood ratios (LR) based on forensic genetic markers. However, genetic evidence alone often proves insufficient in certain scenarios, particularly when determining which individual is the parent and which is the child in a relationship pair, or when establishing distant familial connections. While forensic practitioners frequently incorporate supplementary evidence (SE), such as age, biological sex, or phenotypic traits, in these cases, this integration typically occurs informally, without rigorous probability estimation, compromising procedural transparency and reliability. Here, we present a comprehensive methodological framework that formally synthesizes forensic DNA evidence (FDE) with SE through innovative Markov chain models and customized transition matrices designed for various biological traits. This approach generates combined likelihood assessments expressed as LRs or posterior probabilities. Validation through both simulated and real-world case studies demonstrates that the systematic incorporation of SE substantially improves resolution accuracy in kinship determinations. To facilitate adoption, we have implemented this methodology in mispitools, an open-source R package available to the forensic community.

1 Introduction

The general topic of this paper is forensic kinship cases. Our main goal is to solve problems that cannot be solved using only Forensic DNA Evidence (FDE), i.e., forensic genetic markers. An example of these types of cases is the typical paternity case. The evidence in favour of a parent - child relationship may be overwhelming based on genetic marker data. However, we cannot determine who is the parent and who is the child considering only FDE. Another example involves individuals found to be siblings, but based on forensic DNA makers, we cannot distinguish the younger from the older.

Traditionally, the problem is solved using additional information, such as age, but usually in an informal way without reporting probabilities [1]. In addition, there may be cases, such as distant relationships or in which marker data are of poor quality, that lead to low statistical power when only relying FDE [2, 3, 4, 5]. In these cases, it is common to include Supplementary Evidence (SE) [6]. SE can be age information, biological sex, pigmentation traits, information about tattoos, among other features. Note that SE can be based on genetic markers, used, for example, to predict eye colour [7]. The FDEs are assumed to be conventional markers that do not predict known phenotypes.

Figures 1 and 2 show simple versions of the kinship cases addressed in this paper. The available FDE is similar in both cases, with the same members genotyped. However, this is not true for the SE. Figure 1 shows a parent-child relationship between two individuals, A and B. However, we do not know who is father and who is the son. We refer to this as a *pedigree* case. The example shown in Figure 2 corresponds to a *Disaster Victim Identification* (DVI) case [8]. V_1 and V_2 indicate victims or, more generally, Unidentified Persons (UPs). In a real case, we may have multiple samples from the same UP [9]. Here, we assume that these multiple samples have been merged prior to the analyses. The identities of V_1 and V_2 have not been established. The individual R_1 to the right is a genotyped reference individual in a family with two missing persons M_1 and M_2. The identities of M_1 and M_2 are known, but they have gone missing.

Usually, there is more available SE for the DVI case than for the pedigree case [10]. For example, a feature, such as a tattoo, an injury, or a pigmentation trait, observed

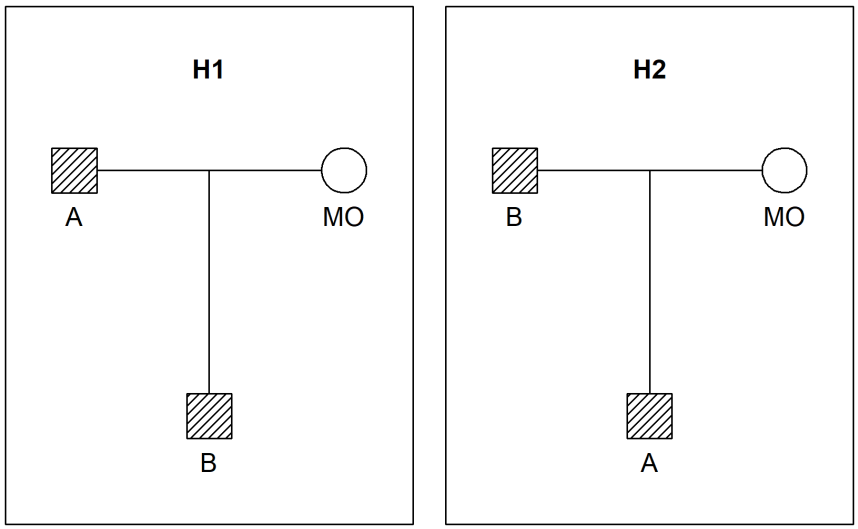

Figure 1: Pedigree scenario. Father - son hypotheses. Squares represent males and circles females. A and B are genotyped (dashed). H_1 and H_2 represent the two hypothesis which switch the position of A and B, therefore defining two pedigrees, $\mathcal{P}_1$ or $\mathcal{P}_2$.

both in an unidentified person and in a missing person, may be helpful in DVI cases [6, 11]. We refer to this as 'Comparison data' as it permits direct comparison between an UP and a missing person. Such data is typically not helpful in the pedigree scenario. For example, even if we find that A has blue eyes and B has brown eyes, this cannot help to distinguish between hypotheses H_1 and H_2 in Figure 1. This means, there is not relationship between the pigmentation trait and the role (parent or child) in the pedigree. However, information on age, such as knowing that A is older than B, would be relevant. Similarly, information on fertility status helps, that is, knowing that only one of them could be a parent. We refer to data of this kind as 'Directional'.

Some papers [1, 6], have studied the general aspects of the search process carried out with the goal of resolving missing person cases, using FDE and SE data. Others focused on the main aspects related to DVI problems, also including SE in some steps of the process [8, 12]. Nevertheless, none of them proposed a general framework for computing LRs and combining FDE-SE to solve kinship problems, ranging from basic cases, such as a paternity, to complex missing person cases hampered by low statistical power. Filling the gap in this area is the main goal of this work. We formulate and implement a model

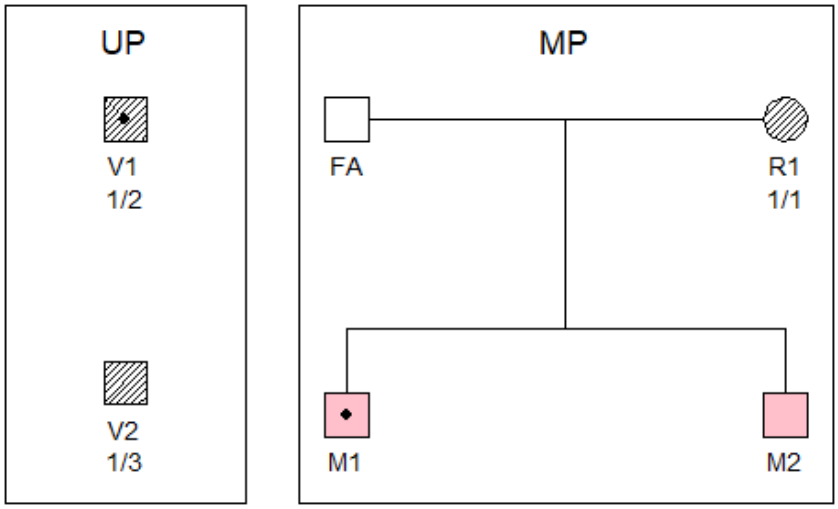

Figure 2: DVI case. Genotypes are available for the unidentified persons (samples) V_1, V_2 and the reference R_1. V_1 and M_1 share a feature, indicated by a dot, not seen in V_2 and M_2.

to combine all available data and exemplify it by resolving cases of practical importance.

The data and methods are presented in Section 2. The books by Egeland, Kling and Mostad [13] and Meester and Slooten [14] present the mathematical basis for kinship problems. We discuss models for directional and comparison data in the next section and show how these data sources are combined to give LR or posterior probabilities. Then examples based on real and simulated data are presented in Section 3. Finally, we discuss the relevance of the models, limitations, and future work in Section 4.

2 Data and methods

2.1 Hypotheses

Consider first pedigree cases; Figure 1 shows an example. Then, there are typically hypotheses H_1 and H_2 specifying that the individuals are related according to pedigrees $\mathcal{P}_1$ or $\mathcal{P}_2$, respectively. Commonly, there could be more hypotheses, say H_3 and H_4, specifying that, for example, A is the uncle of B or vice versa, respectively. However, in DVI applications, there are generally many more hypotheses. For example, those

corresponding to the case presented in Figure 2. The seven possible solutions to the identification problem are listed along with likelihoods and an LRs, explained in the following sections. An hypothesis, referred to as an assignment a, for the DVI problem we are addressing, is a one-to-one correspondence between a subset of $\mathcal{V} = \{V_1, ..., V_s\}$ and a subset of $\mathcal{M} = \{M_1, ..., M_m\}$, typically with the requirement that all identifications are sex consistent. For example, in Table 1, a consistent assignment is $\{V_1 = M_2, V_2 = M_1\}$. Alternatively, we may write this more compactly as a tuple (M_2, M_1).

a	V_1	V_2	L_{SE}	L_{FDE}	$\mathrm{LR}_{.:7}$
a_1	M_1	M_2	$(1-m)\cdot(1-m)$	$\frac{1}{2}p_2 \cdot p_3 \cdot p_1^2$	$\frac{(1-m)^2}{\alpha\beta}\frac{1}{8p_1^2}$
a_2	M_2	M_1	$m \cdot m$	$\frac{1}{2}p_2 \cdot p_3 \cdot p_1^2$	$\frac{m^2}{\alpha\beta}\frac{1}{8p_1^2}$
a_3	M_1	*	$(1-m)\cdot\beta$	$p_2 \cdot 2p_1p_3 \cdot p_1^2$	$\frac{1-m}{\alpha}\frac{1}{2p_1}$
a_4	M_2	*	$m \cdot \beta$	$p_2 \cdot 2p_1p_3 \cdot p_1^2$	$\frac{m}{\alpha}\frac{1}{2p_1}$
a_5	*	M_1	$\alpha \cdot m$	$2p_1p_2 \cdot p_3 \cdot p_1^2$	$\frac{m}{\beta}\frac{1}{2p_1}$
a_6	*	M_2	$\alpha\cdot(1-m)$	$2p_1p_2 \cdot p_3 \cdot p_1^2$	$\frac{1-m}{\beta}\frac{1}{2p_1}$
a_7	*	*	$\alpha \cdot \beta$	$2p_1p_2 \cdot 2p_1p_3 \cdot p_1^2$	1

Table 1: The seven possible solutions to the DVI problem in Figure 2. Each setting is represented by a different a value. The likelihoods based on SE (L_{SE}) and FDE (L_{FDE}) are followed by the LR, using the last assignment where no one are identified as reference. The calculations are explained in Example 3.4.

2.2 Priors

We will in some cases present Bayesian approaches. This amounts to specifying prior probabilities $\pi_1, \pi_2, \ldots$ for the hypotheses or assignments. Obviously, $\pi_i \geq 0$, and $\sum_i \pi_i = 1$. In many cases a flat prior is used, that is, $\pi_1 = \pi_2 = \cdots$.

In some applications, posterior probabilities obtained from previously analyzed data are used as priors in subsequent analyses [1]. Parametric models are also available for pedigree priors [15, 16] as we explain below in Example 3.2. Finally, *empirical Bayes procedures* can be used to incorporate information. As an example, consider the problem of finding the most likely pairwise relationship among alternatives $\mathcal{P}_1, \ldots, \mathcal{P}_n$. If we

have access to a database of pairwise relationships, we can set priors $\pi_1, \ldots, \pi_n$ based on the observed frequencies. This would, for instance, reduce the weight of rare inbred relationships compared to more frequently occurring alternatives and is an example of an empirical Bayes procedure [17].

2.3 Likelihood

2.3.1 Forensic DNA Evidence

The forensic genetic markers are assumed not to be informative of phenotypes beyond possibly being informative about sex. The likelihood calculation will be performed using standard methods and software [18]. We will use the likelihood calculations implemented in the pedsuite libraries and described in the book by Vigeland [19]. In a DVI case, the likelihood can be written

$$L_{FDE}(a) = P(d \mid a, \Phi) \tag{1}$$

where d denotes the genotype data from the forensic markers. We replace the assignment a with the hypothesis H in pedigree cases. We assume that the parameters specifying, e.g., allele frequencies and mutation models, denoted by Φ, are fixed. In the following, we omit Φ in the notation. Throughout, we make the standard simplifying assumptions: Hardy Weinberg Equilibrium, no artifacts (dropout, drop-in, silent alleles, errors) and assume that markers are independent (unlinked and in linkage disequilibrium). However, all these assumptions can be relaxed.

As mentioned previously, we distinguish between evidence that can be used for direct comparisons (comparison data) and information that can be used to order individuals in a pedigree based on age or fertility information (directional data). The latter type is first addressed.

2.3.2 Directional data

Consider Figure 1 and the hypotheses H_1: A is the father of B *versus* H_2: B is the father of A. We choose to use priors to incorporate information that could help distinguish between H_1 and H_2. As mentioned previously, these priors could be based on age

estimates of A or B or on fertility assumptions. It could, for instance, be that one of them is not likely to be a parent, and this may or may not be based on age information.

The mentioned example is only one of many. The same situation occurs whenever the likelihood remains unchanged if two individuals are swapped. This will typically be true if only two individuals are genotyped. There are, however, some mutation models that convey directional information, and the likelihoods may change, typically only marginally, if genotypes are swapped [20].

2.3.3 Comparison data

Consider a single feature x_i observed in an unidentified individual (V_i) and its counterpart y_j in a missing person (M_j). Features may assume multiple levels ($c > 1$), which can be ordered (e.g., height intervals) or unordered (e.g., eye colour). A floating bin approach [21] has been proposed to effectively handle continuous variables in forensic evaluations by conservatively assigning match windows that account for measurement uncertainty. Consequently, continuous variables can be recoded as binary match/nonmatch indicators. For example, assigning a match value 1 when the age window of an unidentified person contains the age of a missing person, and a non-match value 0 otherwise [1].

The conditional distribution of $(x_i \mid y_j, V_i = M_j)$ is modeled by a matrix $c \times c$ $M = (m_{st})$ where $m_{st} \geq 0$ and $\sum_{t=1}^{c} m_{st} = 1$. In other words, M is a transition matrix, similar to those used to model mutations in forensic genetics, as explained in Dawid et al [22]. In other words,

$$P(x_i = s \mid y_j = t, V_i = M_j) = m_{ts}, \tag{2}$$

$$P(x_i = s \mid y_j = t, V_i \text{ and } M_j \text{ unrelated}) = P(x_i = s) = \alpha_s. \tag{3}$$

Note that we model conditionally on the data in the missing person. In other words, if there is uncertainty in the value in the missing person, this is not accommodated by the model.

The transition matrix corresponding to Equation (2) may be written

$$M = \begin{pmatrix} m_{11} & \cdots & m_{1c} \\ \cdot & \cdots & \cdot \\ m_{c1} & \cdots & m_{cc} \end{pmatrix}. \tag{4}$$

Knowledge of the feature may be reflected in the transition matrix. If, we can assume that a possible error can only result in a classification to the closest value, we would use a band matrix like

$$\begin{pmatrix} m_{11} & 1-m_{11} & 0 & \cdot & \cdots & \cdot & 0 \\ m_{21} & m_{22} & m_{23} & 0 & \cdots & \cdot & 0 \\ & & & & \cdots & \cdot & \cdot \\ 0 & 0 & & 0 & \cdots & 1-m_{cc-1} & m_{cc} \end{pmatrix}.$$

The likelihood based on the SE can be written

$$L_{SE}(a) = P(x \mid y, a) = \Pi P(x_i \mid y_j, V_i = M_j) \times \Pi P(x_i \mid V_i \neq M_j) \tag{5}$$

where the first product is over all pairs with $V_i = M_j$, while the second product includes the remaining terms. The transition probabilities on the right hand side of (5) are given by Equations (2) and (3). Note that (5) assumes independence between the feature values for UP-s. In the next section we extend from one to several features.

2.3.4 Modelling conditional dependency between traits

This section describes how we can estimate the dependence between features directly from the data and use this to obtain more realistic values for likelihood and LR. An alternative *parametric* model is presented in Appendix A.

In practice, one feature will not suffice to resolve cases where some likelihoods based on FDE are identical. We therefore assume that k features have been observed for each UP and each M, and recorded, respectively, in $x_i = (x_{i1}, \ldots, x_{ik})$ and $y_j = (y_{j1}, \ldots, y_{jk})$. Consider the first product in Equation (5). We continue to assume that features between

individuals are independent. Therefore, we need only consider $P(x_i \mid y_j, V_i = M_j)$. We introduce variables to indicate if the feature values in UP and M coincide

$$\delta_r(i,j) = \begin{cases} 1 & \text{if } x_{ir} = y_{jr} \\ & \qquad\qquad r = 1, \ldots, k, \\ 0 & \text{otherwise} \end{cases}$$

and $P(\delta_r = 0) = \epsilon_r$. Note that we simplify to a binary error model rather than using more general matrices as in the previous section to limit the number of parameters. We assume that the features in an UP are independent given $V_i = M_j$ and so

$$P(x_i \mid y_j, V_i = M_j) = \Pi_{r=1}^{k} (1 - \epsilon_r)^{\delta_r(i,j)} \epsilon_r^{1-\delta_r(i,j)}. \tag{6}$$

Note that $P(x_i \mid y_j, V_i = M_j) = 1$ if errors do not occur (we define $0^0 = 1$), i.e., when all epsilons are 0.

Consider next the likelihood under the hypothesis that $V_i \neq M_j$. We assume that the feature vector x_j can be ordered to secure a markovian dependence structure, i.e.,

$$P(x_i \mid y_j, V_i \neq M_j) = P(x_{i1}) P(x_{i2} \mid x_{i1}) \cdots P(x_k \mid x_{k-1}).$$

We estimate $P(x_{is} = u \mid x_{i,s-1} = v)$ by $(N_{u,v} + 1)/(N_v + U)$, where $N_{u,v}$ is the observed frequency of the combination (u, v), $u = 1, \ldots, U$, $v = 1, \ldots, V$, and N_v is the total sample size when $x_{i,s-1} = v$.

2.3.5 Complete Likelihood

The complete likelihood based on the SE and FDE can be written

$$\begin{aligned} L(a) &= P(d, x \mid y, a) = P(d \mid x, y, a) P(x \mid y, a) \\ &= P(d \mid a) P(x \mid y, a) \\ &= L_{FDE}(a) \times L_{SE}(a) \end{aligned} \tag{7}$$

where the terms on the right hand side are given in Equations (1) and (5). We have assumed above that the FDE is independent of SE.

2.4 Likelihood Ratios

The LR comparing assignments a to a^* follows directly from Eq. (7):

$$\begin{aligned}\mathrm{LR}(a, a^*) &= \frac{L_{FDE}(a)}{L_{FDE}(a^*)} \times \frac{L_{SE}(a)}{L_{SE}(a^*)} \\ &= \mathrm{LR}_{FDE}(a, a^*) \times \mathrm{LR}_{SE}(a, a^*). \end{aligned} \tag{8}$$

2.5 Posterior probabilities

There is nothing new in terms of converting LRs to posteriors:

$$P(H_i \mid data) = \frac{\pi_i L_i}{\pi_1 L_1 + \cdots + \pi_N L_N} = \frac{\pi_i LR_{i:1}}{\pi_1 LR_{1:N} + \cdots + \pi_N LR_{N:N}} \tag{9}$$

where N is the number of assignments and $LR_{N:N} = 1$.

We next show that the following two models give the same posterior probabilities

1. Assume a flat prior, i.e., $\pi = 1/N$, and calculate the posterior from (9).
2. Assume the prior is defined by the SE data as

$$\gamma_i = \frac{L_{i,SE}}{\sum_{j=1}^{N} L_{j,SE}}$$

   and calculate the posterior from the FDE data as explained below.

Starting with item 1 above, we get

$$\begin{aligned}P(H_i \mid data) &= \frac{L_i}{L_1 + \cdots + L_N} \\ &= \frac{L_{i,SE} L_{i,FDE}}{L_{1,SE} L_{1,FDE} + \cdots + L_{N,SE} L_{N,FDE}} \\ &= \frac{\gamma_i L_{i,FDE}}{\gamma_1 L_{1,FDE} + \cdots + \gamma_N L_{N,FDE}}\end{aligned}$$

which completes the argument.

2.6 Statistical power calculation

To evaluate the statistical power of a given LR model, we define the true positive rate (tp) as $P(LR > T|H_1)$ and the false positive rate (fp) as $P(LR > T|H_2)$. Similarly, the true negative rate (tn) is $P(LR \leq T|H_2)$ and the false negative rate (fn) is $P(LR \leq T|H_1)$. These ratios are used to compute the Matthews-Correlation Coefficient.

To obtain error rates and evaluate performance metrics, a methodology using simulation has been suggested [2, 3, 23]. In this section, we describe this general approach, summarised in Algorithm 1 below, which is applicable to various types of evidence, both genetic and non-genetic.

Algorithm 1 Evidence simulations

```
for j in (MP_1, MP_2, ...) do
    Simulate UP-s
    for i in (UP_1, UP_2, ...) do
        Simulate data considering P(data|H)
    end for
end for
```

For each MP_j in the database, the UP-s are simulated. Then, for each UP, *data* (that could be FDE or SE) is simulated considering H_1 or H_2 to be true.

2.7 The Balsac data

We employed the BALSAC dataset, a comprehensive genealogical and demographic repository derived primarily from Quebec's civil marriage records [24], to evaluate the models based on real data. Established in 1972 at the Université du Québec à Chicoutimi, BALSAC spans nearly four centuries, providing a resource for reconstructing historical population structures. Balsac data can be accessed for research purposes only upon request, see https://balsac.uqac.ca. Further details on Balsac data are provided in Appendix B.

2.8 Implementation

This work uses R libraries available for both FDE and SE models. For FDE, the main computations of the LR distribution are done using `forrel` [23], which is part of the collection of packages `pedsuite` [19], and `Familias` [13, 15]. For SE models, specific functions have been developed and are available in the `mispitools` R package [25] (https://CRAN.R-project.org/package=mispitools). Moreover, we provide R code, available as supplementary material (`usingAll.R`), that allows us to run both SE and FDE based LR-computations, simulations, and evaluation of the performance. In addition, the R library `dvir` [8] can be used to calculate the likelihoods of all assignments.

3 Results

In the first examples, cases with directional data are presented. Various ways of modeling priors are discussed. Then the comparison SE data are introduced and exemplified in different settings. Finally, examples combining FDE and SE LRs are shown.

3.1 Directional data

In this subsection, we analyze some examples based on directional data, where SE can inform about possible kinship configurations.

Example 3.1. We return to the example in Figure 1. This is one example in a large class of similar cases. The defining feature is that the likelihood based on FDE remains unchanged for all permutations of the genotyped individuals. In the figure, there are only two permutations. In a case with three full siblings, there are six permutations giving the same likelihood. The posteriors in (9) simplify to the prior, that is, $P(H_i \mid data) = \pi_i$ by definition since the likelihoods are identical. It is therefore only possible to distinguish between the hypotheses if we have information that can be used to inform the prior, resulting in some posterior being sufficiently close to 1. As discussed in Section 2.2, the priors can be obtained by

- a direct specification,

- a parametric model,
- empirical Bayes.

The first alternative is perhaps what is most commonly done in practice. It may, for example, be known that only one of the pedigrees in Figure 1 is possible. The parametric approach would not be helpful here since it would give a flat prior in symmetric cases like this one.

In the next example, we demonstrate how the parametric prior can be helpful. The final option, empirical Bayes, requires information on the individuals and technically leads to the DVI approach discussed later.

Example 3.2. In this example we are only concerned with the type of relation between individuals A and B. Specifically, we consider the hypotheses H_1 : 'Grandparent-grandchild' and H_2 : 'half-sib'. It is well known that no amount of unlinked marker data can resolve the case. Prior information (or SE discussed later) can help. We next consider the previously mentioned parametric prior [15]. This prior is proportional to

$$M_I^{b_I} M_P^{b_P} M_G^{b_G}$$

where M_I, M_P, M_G, refer to inbreeding, promiscuity and generations and are non-negative parameters. A value of M_G below 1 decreases the prior probability for pedigrees with few generations and similarly for the two other parameters where also unity corresponds to a neutral position. The corresponding integer exponentials b_I, b_P, b_G, are calculated by the Familias program [13]. b_I is the number of children where both parents are present in the pedigree and where the parents have a common ancestor in the pedigree. In our example, there is no inbreeding and hence $b_I = 0$. For promiscuity, the number of pairs having precisely one parent in common is calculated and denoted b_P. Here, $b_P = 0$ for the grandparent - grandchild relation while it is 1 for half-sibs. Finally, turning to generations, the longest chain of generations starting with a named person (i.e., a person to be genotyped) and ending in an adult, is calculated and assigned the value b_G. Here,

$b_G = 2$ and 1, respectively. Based on this we find the following prior

$$\pi_1 = cM_G^2,\ \pi_2 = cM_G M_P,\ c = (M_G^2 + M_G M_P)^{-1}. \tag{10}$$

The individuals are believed to be of roughly the same age and this is judged to correspond to $M_G = 0.001$. We have no useful information to assign a value pertaining to promiscuity and we therefore set $M_P = 1$. The priors, and also the posterior in this case, are approximately 0.001 and 0.999 based on (10). In practice, one would perform sensitivity analyses to study the impact of different prior specifications.

Example 3.3. Assume we are asked whether individuals A and B are paternal half-siblings (H_1) or maternal half-siblings (H_2). With extensive linked data, one might distinguish between these hypotheses by exploiting the fact that recombination rates are higher in females [19]. However, when only unlinked markers are available, the FDE yields identical likelihoods for both hypotheses, i.e., $L_{FDE}(H_1) = L_{FDE}(H_2)$, resulting in a LR of 1.

Genealogical information, such as that provided by the BALSAC dataset (Appendix B and Figure S1), can be used to inform prior probabilities. Empirical evidence indicates that approximately 62.5% of half-sibling pairs are paternal. Hence, an empirical Bayes approach suggests setting $\pi(H_1) = 0.625$ and $\pi(H_2) = 0.375$. Since $L_{FDE}(H_1) = L_{FDE}(H_2)$, the posteriors coincide with the priors and so $P(H_1 \mid \text{data}) = 0.625$ and $P(H_2 \mid \text{data}) = 0.375$. Thus, although the FDE is uninformative in a directional sense, the incorporation of empirical directional data yields posterior probabilities that favor the paternal half-sibling hypothesis.

3.2 Example: Comparison data

In this subsection, we analyze a set of examples based on comparison data.

Example 3.4. This example is based on Figure 2. We have a binary feature with value 1 (indicated by a dot in the figure), and value 2 (no dot) with frequencies α and $\beta = 1 - \alpha$,

respectively. The probability of missclassification is m, i.e.,

$$P(x_i = 1 \mid y_j = 2, V_i = M_j) = P(x_i = 2 \mid y_j = 1, V_i = M_j) = m.$$

We see from the figure that the SE data is $x_1 = 1, x_2 = 2, y_1 = 1, y_2 = 2$ while the FDE is $g_{V_1} = 1/2, g_{V_2} = 1/3$ and $g_{R_1} = 1/1$. The possible assignments, the corresponding likelihoods and the LRs are given in Table 1. These calculations are based on the general likelihood given in (7) and the corresponding LR (8). We provide the details for the last line of the table. Regarding the SE, we find

$$L_{FDE}(a_7) = P(x_1 = 1, x_2 = 2 \mid y_1 = 1, y_2 = 2, V_1 = *, V_2 = *) = \alpha \cdot \beta.$$

The conditioning on the missing persons is in this case irrelevant since they are unrelated to the unidentified persons. Moreover, we have assumed that the feature occurs independently in the UP-s, explaining the above equality. Turning to the FDE, we multiply the genotype probabilities of the genotyped individuals since they are unrelated in this case to arrive at

$$L_{FDE}(a_7) = 2p_1p_2 \cdot 2p_1p_3 \cdot p_1^2.$$

We can divide the assignments of Table 1 into three groups:

$$\{a_1, a_2\},\ \{a_3, a_4, a_5, a_6\},\ \{a_7\}. \tag{11}$$

The assignments within the first two groups can not be distinguished based on autosomal genetic markers.

We next give some numerical examples. Obviously, estimates of the parameters are then needed. Allele frequencies are obviously treated similarly as in standard forensic applications. Regarding, the feature frequency, α and $\beta = 1 - \alpha$, reliable estimates may or may not be available. The hardest parameter to estimate is m, the probability of missclassification. If $m = 0$, corresponding to perfect feature classification, assignments a_2, a_4 and a_5 are excluded. At the other extreme, $m = 0.5$, there is no information to distinguish within the three groups $\{a_1, a_2\}$, $\{a_3, a_4\}$ and $\{a_5, a_6\}$. Table 2 provides

a numerical example based on Table 1. In most practical cases, there will be enough forensic markers to distinguish *between* the tree mentioned groups (11). The challenge is to distinguish within the groups and then we need more SE evidence as discussed in the next examples.

	V_1	V_2	combinedLR	LRSE	LRFDE	posterior
a1	M1	M2	63.2812	5.0625	12.5000	0.6570
a2	M2	M1	0.7813	0.0625	12.5000	0.0081
a3	M1	*	5.6250	1.1250	5.0000	0.0584
a4	M2	*	0.6250	0.1250	5.0000	0.0065
a5	*	M1	2.5000	0.5000	5.0000	0.0260
a6	*	M2	22.5000	4.5000	5.0000	0.2336
a7	*	*	1.0000	1.0000	1.0000	0.0104

Table 2: LRs and posteriors calculated based on Table 1. Here $p = (0.1, 0.2, 0.7)$, $\alpha = 0.8$ and $m = 0.1$. The posterior is calculated as explained in Section 2.5.

Example 3.5. This example is a simulated case based on a real pedigree from the Balsac database [24]. The purpose is to demonstrate how we can deal with several variables of different types (binary, factors with *unordered* and *ordered levels*). The variables are assumed to be independent. Dependence is addressed in the next example. Consider Figure 3. The father 4 has 11 children by 3 and 8 by 5. The following SE evidence is shown in Figure 3:

1. A binary feature with presence indicated by green (G). The prevalence of green in the population is p_G while a missclassification occurs with probability μ_G.
2. An *unordered* factor taking values 'B', 'D' or 'X', indicated above the individuals in the figure, showing blue, dark or other eye colour. The population frequencies are p_B, p_D, p_X, respectively, and sum to 1. The missclassification probabilities are given by the 'equal' mutation model.
3. An *unordered* factor taking values $1, \ldots, 5$ with probabilities $p_1, \ldots, p_5$, indicating

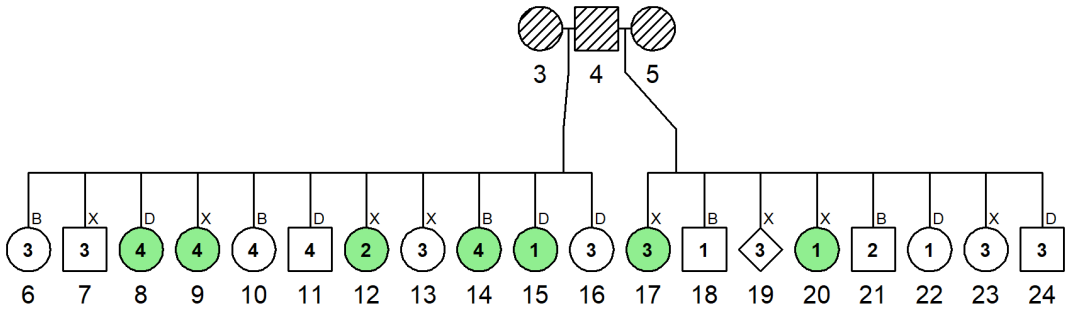

Figure 3: A Balsac pedigree based on genealogical registers from Quebec, Canada. Squares correspond to males, circles to females, and rhombus showing unknown biological sex

age category from youngest to oldest. We use a band matrix with overall mutation rate μ_A, as explained earlier to model the transitions.

An unidentified male (V_1) known to be one of the children or someone not related to the family is found. FDE excludes V_1 as been among the individuals 17 - 24. We therefore only consider candidates among individuals 6 - 16. V_1 is a male and hence there are three possible assignments a_1, a_2 and a_3 corresponding to V_1 being '7', '11', or '*' (an unrelated individual), respectively. The individual V_1 is not green, has other ('X') eye colour, and belongs to age group 3. The corresponding likelihoods are

$$
\begin{aligned}
L_{SE}(a_1) &= (1-\mu_G)(1-\mu_E/2)(1-\mu_A/2),\\
L_{SE}(a_2) &= (1-\mu_G)\,\mu_E/2\,\mu_A/2,\\
L_{SE}(a_3) &= p_G p_X p_3.
\end{aligned}
$$

This gives the following likelihood ratios

$$
\begin{aligned}
\mathrm{LR}_{SE}(a_1, a_3) &= \frac{(1-\mu_G)(1-\mu_E/2)(1-\mu_A/2)}{p_G p_X p_3},\\
\mathrm{LR}_{SE}(a_2, a_3) &= \frac{\mu_G\,\mu_E/2\,\mu_A/2}{p_G p_X p_3}.
\end{aligned}
$$

It could be more relevant to compare the alternatives that are identical wrt to the FDE

likelihood:

$$\mathrm{LR}_{SE}(a_1, a_2) = \frac{(1-\mu_E/2)(1-\mu_A/2)}{\mu_E/2\,(\mu_A/2)}.$$

If the probability of miss-classifying eye colour and age is 0.05, the above likelihood ratio is 1521.

Example 3.6. We illustrate how dependence between two phenotypic traits like hair and eye colour can affect the LR in a kinship context. Let $\mathbf{x} = (x_1, x_2)$ be the pigmentation traits, hair and eye colour respectively, of the unidentified person and $\mathbf{y} = (y_1, y_2)$ for the missing person. We apply the model presented in 2.3.4 section.

Joint frequency estimates for hair and eye colour can be obtained from large forensic DNA-phenotyping databases [7, 26]. Figure 4 depicts an example of the resulting probabilities for hair colour (1 = black, 2 = brown, 3 = red, 4 = blond) and eye colour (1 = brown, 2 = blue, 3 = green). Darker cells indicate more common trait combinations.

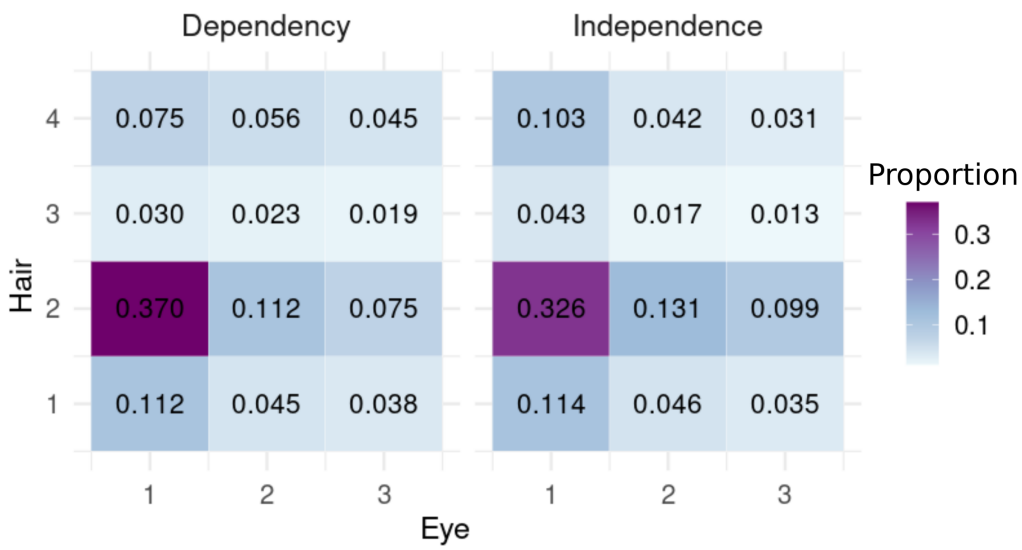

Figure 4: Joint frequencies $f_{(x_1,x_2)}$ under H_2, using (left) a conditional dependency model and (right) an independence model. Each cell corresponds to a combination (x_1, x_2), where $x_1 \in \{1 = \text{black}, 2 = \text{brown}, 3 = \text{red}, 4 = \text{blond}\}$ and $x_2 \in \{1 = \text{brown}, 2 = \text{blue}, 3 = \text{green}\}$. The darkest cells represent the most frequent combination, here brown hair with brown eyes.

Under H_1, each feature x_i is assumed to match its counterpart y_i with probability

$1 - \epsilon_i$ and to be misclassified with probability ϵ_i. The LR comparing H_1 to H_2 is:

$$\mathrm{LR}_{SE} = \frac{P(\mathbf{x} \mid \mathbf{y}, H_1)}{P(\mathbf{x} \mid H_2)}.$$

The denominator $P(\mathbf{x} \mid H_2)$ follows either the dependency or independence assumption.

As an example, we consider a case where two most common pigmentation characteristics match between M and UP. For the dependency model,

$$LR^{(\mathrm{dep})} = \frac{0.95^2}{0.370} = 2.44$$

while independency gives

$$LR^{(\mathrm{ind})} = \frac{0.95^2}{0.330} = 2.77.$$

This means, that not considering the dependency leads to a overrestimation of the LR; the ratio ‘independence‘/‘dependence‘ is $2.77/2.44 = 1.14$. In this case overestimation is reasonable since ‘brown hair’ and ‘brown eyes’ are positively correlated. The denominator of the LR is then smallest for independence. Therefore, if we assume independence, we overestimate the power of observing this joint feature. In the extreme case of correlation 1, the second feature would contain no additional information if the first is observed. The features ‘brown eyes’ and ‘blond hair’ are negatively correlated and in this case the ratio is $0.075/0.103 = 0.73$. In other words, ignoring dependence leads to under estimation. The effect of modelling dependence can be explored for different traits combinations, as shown in Table 3. It illustrates that for certain observed combinations, especially rare ones, the LRs can differ markedly between the two models.

Example 3.7. Power considerations have been useful for FDE [2, 3, 23]. In this example, we extend the discussion to SE. Given the likelihood functions for H_1 and H_2 , it is possible to evaluate $P(LR_{SE} \leq x|H_1)$ and $P(LR_{SE} \leq x|H_2)$. In particular, this approach allows researchers to:

i) determine the expected LR values under both H_1 and H_2 [27];

ii) assess the discrimination capacity of LR by analyzing the overlap between $P(LR_{SE}|H_1)$ and $P(LR_{SE}|H_2)$ [3];

x_1^*	x_2^*	δ_1	δ_2	$f^{(\text{dep})}$	$f^{(\text{ind})}$	$LR^{(\text{dep})}$	$LR^{(\text{ind})}$	Ratio
2	1	1	1	0.370	0.330	2.44	2.77	1.14
4	3	1	1	0.045	0.031	19.94	28.88	1.45
1	3	1	0	0.038	0.035	1.25	1.38	1.10
3	2	1	1	0.023	0.017	39.03	52.35	1.33

Table 3: LR values under the dependency vs. independence models for four selected trait combinations. The columns $f^{(\text{dep})}$ and $f^{(\text{ind})}$ show the smoothed population frequencies under H_2, with and without conditional dependence. The indicators δ_1, δ_2 specify which traits match under H_1. The Ratio is obtained by dividing $LR^{(\text{ind})}$ by $LR^{(\text{dep})}$.

iii) use simulations to predict which new evidence is most likely to enhance the method's discrimination power [23, 28].

We analyze these properties using a missing person case from the Balsac database (see Appendix B). The subject (M) is a female, with an age at death of 40 ∓ 5 years) and brown hair. Population frequencies for age and sex are taken from the Balsac dataset, while hair color frequencies are based on published data [29]. A conservative error rate of $\epsilon = 0.05$ is used for all models [6] and sensitivity analyses of all SE models are provided in Appendix C.

In total, 10,000 UPs were simulated under both H_1 and H_2. For example, under H_1, approximately 95% of the simulated UPs were female (consistent with M), while under H_2 the frequency was about 50%, matching the reference population. Similar simulations were carried out for hair color and age. For each simulated UP, an LR was computed using the described model and case, yielding LR distributions for each trait under both hypotheses. Also, combined LR distribution for all traits were obtained using direct product between LRs.

Figure 5 illustrates these LR distributions. In Panel A, for instance, the blue bar (corresponding to H_1) shows a higher frequency of cases with $\log_{10}(LR) = 0.28$, whereas the brown bar (for H_2) indicates that lower LR values are more common. Similar patterns are observed for hair color (Panel B), age (C) and combined (D).

The results reveal a clear trend in the ability of SE variables to differentiate between

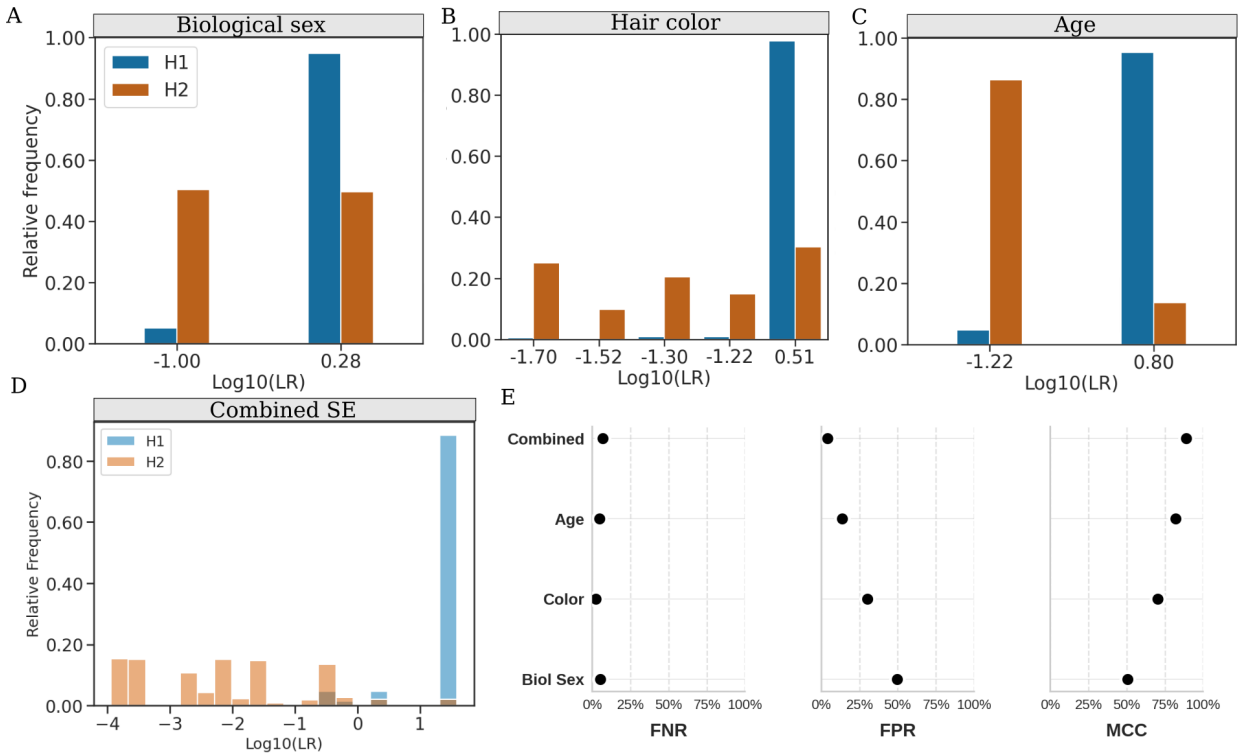

Figure 5: Panels A-C show the distributions of Log_{10} for three separate variables: biological sex, hair colour, and age, respectively, under the hypotheses H_1 : UP is M (blue bars) and H_2: UP is not M (brown bars). Panel D illustrates how combining these variables (under conditional independence) produces a sharper separation between LR distributions. Panel E shows performance metrics (FNR, FPR and Matthews Correlation Coefficient, MCC) compared for each approach, highlighting the improved discrimination achieved when multiple supplementary evidences are used. In all cases, the LR threshold considered is 1.

H_1 (UP is M) and H_2 (UP is not M). As shown in panels A–C, Biological Sex provides the weakest discriminatory signal, with an MCC = 0.5049, meaning that nearly half of non-matching individuals could be incorrectly classified.

This aligns with expectations, given the limited number of sex categories and their relatively balanced distribution in human populations. Hair colour improves performance (FPR = 0.3018, MCC = 0.7027), likely due to its increased variability and predictive power in some populations.

However, Age stands out as the most informative single variable, with a notably low FPR (0.1361) and the highest MCC (0.8203) among individual traits.

The greatest improvement occurs when all SE variables are combined (Figure 5D).

Although the FNR slightly increases (0.0692), the FPR drops dramatically to 0.0408, and the MCC reaches 0.8904, indicating a high classification performance.

3.3 Combining SE and FDE data

Finally, we investigated two pedigrees, each of which had low statistical power when relying solely on FDE (Figure 6). This implies that low LR values ($LR < 1$) can be obtained when H_1 is true (FNR), and high values when H_2 is true. The SE data are the same as those analyzed in the previous example. For FDE, 23 autosomal STRs makers were considered. In Pedigree 1, the maternal grandmother and great-grandmother were available for genotyping, in pedigree 2, only a paternal first cousin.

Figure 6 illustrates the distributions of $\log_{10}(\text{LR})$ under two scenarios: one using FDE alone, and one where FDE is combined with SE (Section 2.4, Equation 8). In both cases a classification threshold of $\text{LR} = 1$ was applied.

For Pedigree 1, the FDE only approach already showed relatively strong performance, reaching a FNR of 0.031 and a FPR of 0.039, corresponding to a MCC of 0.93. However, once SE was incorporated, the FPR dropped to 0.006, and the MCC rose to 0.96.

Even more importantly, improvements were observed for Pedigree 2, which initially proved difficult to resolve with FDE alone, with $FNR = 0.246$, $FPR = 0.22$ and $MCC = 0.53$. By integrating SE data, we substantially reduced both FNR and FPR, to 0.053 and 0.029, respectively, and increased the MCC to 0.92.

4 Discussion

Our study focuses on the integration of Supplementary Evidence (SE) into kinship evaluation frameworks to resolve cases where Forensic DNA Evidence (FDE) alone proves insufficient. We address some fundamental limitations, such as cases where (i) FDE likelihoods remain invariant under permutation of individuals, preventing discrimination of, for example, parent-child directionality; and cases where (ii) adding SE is crucial for practical reason, for example when FDE is insufficient due to degraded samples or missing informative relatives of a missing person.

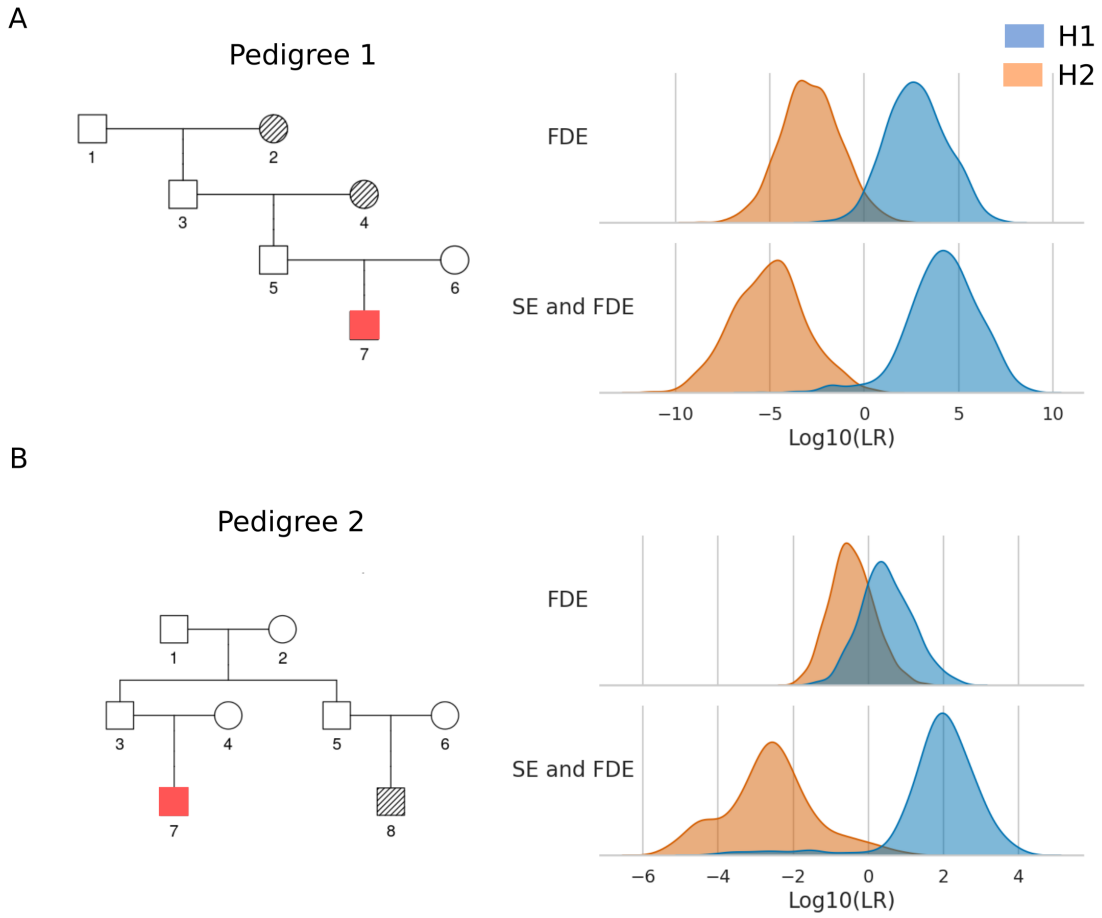

Figure 6: The distributions of $\log_{10}$(LR) for two different pedigrees (labeled as Pedigree 1 and Pedigree 2), under two scenarios: using only the FDE (*top row*) versus combining FDE with SE (*bottom row*).

For (i), we propose a solution by incorporating Bayesian priors derived from empirical age distributions and intergenerational intervals (see Examples 3.1 and 3.2). Our model generates informative posterior probabilities for alternative pedigree configurations.

For (ii), we introduce the *comparison data*, implementing transition matrices to model the correspondence between phenotypic traits (e.g., age, biological sex or hair colour) observed in UPs and Ms, as formalised in Equations (2) and (3). In DVI scenarios (Tables 1 and 2), the joint likelihood, calculated as the product of FDE and SE contributions, enables discrimination between assignment configurations that would remain indistinguishable using genetic data alone.

Below we address theoretical and practical challenges related to the use of SE data.

4.1 On the incorporation of SE

Despite the potential value of SE in kinship analysis, formal methods for its incorporation remain largely unexplored in the literature, with the exception of few works [1]. This lack of research can be attributed to several factors: (i) the perceived sufficiency of conventional genetic evidence without supplementary data; (ii) the preference for qualitative expert assessments over quantitative approaches when supplementary evidence is considered; and (iii) the inherent challenges in developing robust statistical models with accurately estimated parameters that can adequately accommodate the various forms of supplementary evidence.

Argument (i) is valid for numerous cases where FDE serves as the primary investigative tool, particularly in contexts with access to high-quality DNA processing capabilities. However, multiple studies emphasise that the identification phase in missing persons investigations must adopt an integrative approach that considers all available evidence [6]. Furthermore, previous research has demonstrated how even well-powered genetic analyses may yield erroneous conclusions due to the inappropriate application of non-genetic data, such as biological sex information, when SE characteristics are used as filtering criteria but are potentially inaccurate [1].

The argument (ii) is highly context dependent. In large scale databases containing thousands of unidentified persons, case-by-case expert review becomes logistically unfeasible. Furthermore, in scenarios where FDE is insufficient due to limited reference samples or distant genetic relationships between the missing individual and available relatives, comparisons frequently generate numerous false positives for specific pedigrees, making exhaustive manual verification impractical. Of particular concern are false negatives arising from underpowered cases, which represent silent errors, since low FDE-based LRs typically remain unexamined in extensive databases [3]. This systematic oversight precludes the opportunity to incorporate SE that might substantially alter identification decisions and ultimately lead to successful resolutions that would otherwise be missed.

Regarding (iii), specifying models and parameters for supplementary evidence presents

challenges compared to FDE-based approaches, which benefit from established biological foundations and extensive parameter estimation data. However, similar difficulties exist within FDE modeling itself, particularly for mutation and dropout parameters, necessitating sensitivity analyses to validate conclusions across plausible parameter ranges [13]. We applied this same approach to our SE models, as demonstrated in Figure C, revealing systematic patterns: increasing uncertainty (higher error rates) consistently decreased LRs under H_1 while increasing them under H_2, confirming that our model appropriately adjusts the evidential weight in response to uncertainty. Furthermore, population-rare traits generated higher LR, whether biological sex or specific hair pigmentation, reinforcing the fundamental forensic principle that trait rarity correlates with informational value [28].

4.2 Alternative approaches

In this subsection, we examine alternative methodological approaches to those presented in the manuscript. Our framework can be conceptualized as modular building blocks, where the integration of novel evidence types, alternative data sources, and additional variables presents distinct analytical challenges. Here, we introduce several identified limitations that represent promising avenues for future research and potential model refinements.

4.2.1 Directional data

In Section 2.3.2 we used a prior to include directional SE evidence. Alternative approaches that use data are possible as discussed next. Assume that the distributions of ages for A and B are normally distributed and independent, i.e., $x_A \sim N(\mu_A, \sigma_A^2)$ and $x_B \sim N(\mu_B, \sigma_B^2)$. Furthermore, we could say that the hypotheses H_1 and H_2 in Figure 1 are possible if $q_1 = P(x_A - x_B > g)$ and $q_2 = P(x_B - x_A > g)$, respectively, where g is a minimal difference in age between parent and offspring, say $g = 15$. Assume that point estimates for the parameters μ_1, μ_2, σ_1 and σ_2 have been provided so that q_1, q_2 and $q3 = 1 - q_1 - q_2$ can be calculated. We could use these q-estimates as priors to get

the posteriors

$$P(H_i) = \frac{q_i LR_{i3}}{q_1 LR_{13} + q_2 LR_{23} + q_3}$$

where H_3 specifies A and B to be unrelated. Assume the LRs comparing H_1 and H_2 to H_3, are $LR_{13} = LR_{23} = 100000$. If the ages of A and B are estimated to 30 and 10 with standard deviation estimates at 5 and 2, $q_1 = 0.82, q_2 \approx 0, q_3 = 0.18$. This gives $P(H_1) \approx 1$.

The integration of age-based directional evidence provides a statistical framework that can resolve relationship ambiguities in cases where traditional DNA evidence alone is insufficient.

4.2.2 Comparison data

Comparison data is crucial in the investigative process [6, 30]. Some variables are easier to model mathematically due to their categorical nature (e.g., biological sex). However, continuous variables, such as age, require either the formulation of continuous probability distributions or the establishment of fixed categories based on predetermined ranges. As previously proposed for the interpretation of FDE, the floating bin approach [21] can be used to handle continuous variables in forensic statistical evaluations. This method conservatively assigns matching windows that account for measurement uncertainty and remains easily interpretable.

Other challenge in these calculations is defining the likelihood model for H_2, particularly since UP is not M. In this context, population based statistics are used to estimate the probability of observing a given characteristic, under the assumption that UP represents a randomly selected individual from the reference population. For several SE variables, including age, biological sex, and others, readily accessible sources such as census records can be used. However, for traits like pigmentation, determining these likelihoods is considerably more challenging, and their definition has proven particularly problematic [31]. For the latter, various efforts are underway to build reliable databases [29, 26].

Moreover, by assuming the independence between the characteristics, LRs can be

combined through direct multiplication. In our case, we assumed that the FDE did not contain any markers predictive of phenotypes, making the independence assumption reasonable. While assuming independence is logical for variables such as biological sex and age in specific contexts, it is well known that this assumption does not hold for pigmentation traits [29].

Modeling the dependency between two variables can significantly enhance the identification process by capturing informative relationships that would otherwise be overlooked under the assumption of independence. By incorporating conditional dependency, this LR now reflects both the direct probabilities of trait concordance and the population structure of these traits.

However, such dependencies are often challenging to specify explicitly or detect without detailed computations. Moreover, incorporating SE models allows researchers to simulate expected statistical outcomes, providing clear insights into, for example, the range of LR values when H_1 or H_2 is true. This simulation approach has proven valuable in various applications of forensic DNA evidence [2, 23, 28].

5 Conclusion

Our framework integrates forensic DNA evidence with supplementary evidence through a statistical approach that addresses limitations in kinship analysis. By formalizing the use of directional data via Bayesian priors and implementing transition matrices for comparison data, we enable identification in scenarios where genetic markers alone are insufficient. The implementation in the open-source mispitools package provides forensic practitioners with quantitative tools that enhance statistical power in complex cases. This work demonstrates that properly modeled supplementary evidence contributes discriminatory information to kinship analyses, offering solutions for challenging scenarios ranging from standard paternity testing to large-scale disaster victim identification where successful resolutions might otherwise remain elusive.

References

[1] Franco Marsico and Inés Caridi. Incorporating non-genetic evidence in large scale missing person searches: A general approach beyond filtering. *Forensic Science International: Genetics*, page 102891, 2023.

[2] Daniel Kling, Thore Egeland, Mariana Herrera Piñero, and Magnus Dehli Vigeland. Evaluating the statistical power of DNA-based identification, exemplified by 'The missing grandchildren of Argentina'. *Forensic Science International: Genetics*, 31:57–66, 2017.

[3] Franco L Marsico, Magnus D Vigeland, Thore Egeland, and Mariana Herrera Piñero. Making decisions in missing person identification cases with low statistical power. *Forensic Science International: Genetics*, 54:102519, 2021.

[4] François-Xavier Laurent, Andrea Fischer, Robert F Oldt, Sree Kanthaswamy, John S Buckleton, and Susan Hitchin. Streamlining the decision-making process for international DNA kinship matching using Worldwide allele frequencies and tailored cutoff log10LR thresholds. *Forensic Science International: Genetics*, 57:102634, 2022.

[5] Maarten Kruijver, Ronald Meester, and Klaas Slooten. Optimal strategies for familial searching. *Forensic Science International: Genetics*, 13:90–103, 2014.

[6] Mercedes Salado Puerto, Denise Abboud, Jose Pablo Baraybar, Angel Carracedo, Stephen Fonseca, William Goodwin, Pierre Guyomarc'h, Alejandra Jimenez, Udo Krenzer, Maria Dolores Morcillo Mendez, et al. The search process: Integrating the investigation and identification of missing and unidentified persons. *Forensic Science International: Synergy*, 3:100154, 2021.

[7] Manfred Kayser, Wojciech Branicki, Walther Parson, and Christopher Phillips. Recent advances in Forensic DNA Phenotyping of appearance, ancestry and age. *Forensic Science International: Genetics*, page 102870, 2023.

[8] Magnus D Vigeland and Thore Egeland. Joint DNA-based disaster victim identification. *Scientific Reports*, 11(1):13661, 2021.

[9] Daniel Corach. Mass disaster victim identification assisted by dna typing. In *Molecular Diagnostics*, pages 407–415. Elsevier, 2010.

[10] Bruce Budowle, Jianye Ge, Ranajit Chakraborty, and Harrell Gill-King. Use of prior odds for missing persons identifications. *Investigative genetics*, 2:1–6, 2011.

[11] Camilla Tettamanti, Francesca Frigiolini, Lorenzo Franceschetti, Rosario Barranco, Sara Lo Pinto, Lucia Casarino, Simonetta Verdiani, Mattia Porcu, Cristina Cattaneo, Danilo De Angelis, et al. A forensic approach to complex identification cases: The collapse of an italian cemetery into the sea. *Genes*, 16(3):277, 2025.

[12] Inés Caridi, Enrique E Alvarez, Carlos Somigliana, and Mercedes Salado Puerto. Using already-solved cases of a mass disaster event for prioritizing the search among remaining victims: a Bayesian approach. *Scientific reports*, 10(1):1–11, 2020.

[13] Thore Egeland, Daniel Kling, and Petter Mostad. *Relationship Inference with Familias and R: Statistical methods in Forensic Genetics*. Academic Press, 2015.

[14] Ronald Meester and Klaas Slooten. *Probability and forensic evidence: Theory, philosophy, and applications*. Cambridge University Press, 2021.

[15] T Egeland, P F Mostad, B Mevåg, and M Stenersen. Beyond traditional paternity and identification cases. Selecting the most probable pedigree. *Forensic Science International*, 110:47–59, 2000.

[16] Sheehan NA and Egeland T. Structured incorporation of prior information in relationship identification problems. *Annals of Human Genetics*, 71:501–518, 2007.

[17] George Casella. An introduction to empirical bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.

[18] Daniel Kling, Andreas O Tillmar, and Thore Egeland. Familias 3–extensions and new functionality. *Forensic Science International: Genetics*, 13:121–127, 2014.

[19] Magnus Dehli Vigeland. *Pedigree analysis in R*. Academic Press, 2021.

[20] T Egeland and MD Vigeland. Properties of mutation models with applications in forensic genetics. Manuscript, 2023.

[21] James F Crow, MA Berger, SS Diamond, DH Kaye, HH Kazazian, AG Motulsky, TA Nagylaki, M Nei, GF Sensabaugh, DO Siegmund, et al. The evaluation of forensic DNA evidence. *National Re-899 search Council*, 900, 1996.

[22] A P Dawid, J Mortera, and V L Pascali. Non-fatherhood or mutation? A probabilistic approach to parental exclusion in paternity testing. *Forensic Science International*, 124:55–61, 2001.

[23] Magnus D Vigeland, Franco L Marsico, Mariana Herrera Pinero, and Thore Egeland. Prioritising family members for genotyping in missing person cases: a general approach combining the statistical power of exclusion and inclusion. *Forensic Science International: Genetics*, 49:102376, 2020.

[24] Hélène Vézina, Jean-Sébastien Bournival, et al. An overview of the BALSAC population database. Past developments, current state and future prospects. *Historical Life Course Studies*, 9:114–129, 2020.

[25] Franco L Marsico. Mispitools: An R package for comprehensive statistical methods in Kinship Inference. *bioRxiv*, pages 2024–08, 2024.

[26] Susan Walsh, Lakshmi Chaitanya, Krystal Breslin, Charanya Muralidharan, Agnieszka Bronikowska, Ewelina Pospiech, Julia Koller, Leda Kovatsi, Andreas Wollstein, Wojciech Branicki, et al. Global skin colour prediction from dna. *Human genetics*, 136:847–863, 2017.

[27] Thore Egeland, Nadia Pinto, and Magnus Dehli Vigeland. A general approach to power calculation for relationship testing. *Forensic Science Internaname.textional: Genetics*, 9:186–190, 2014.

[28] Franco Marsico, Gustavo Sibilla, Ma Soledad Escobar, and Ariel Chernomoretz. The missing person problem through the lens of information theory. *Forensic Science International: Genetics*, page 103025, 2024.

[29] Nina G Jablonski and George Chaplin. The colours of humanity: the evolution of pigmentation in the human lineage. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1724):20160349, 2017.

[30] Mercedes Salado Puerto and Hugh Tuller. Large-scale forensic investigations into the missing: Challenges and considerations. *Forensic science international*, 279:219–228, 2017.

[31] Peter M Schneider, Barbara Prainsack, and Manfred Kayser. The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry. *Deutsches Ärzteblatt International*, 116(51-52):873, 2019.

A Parametric dependence model

In general, there appears not to be any general parametric models for dependence suitable for our applications. However, in the simple case with two binary variables a model can be established and studied. Let $P(x_1 = 1) = 1 - P(x_1 = 0) = p$ and $P(x_2 = 1) = 1 - P(x_2 = 0) = q$ and consider the joint distribution given in Table 4. The

x_1/x_2	0	1	sum
0	$1 - p - q + pq(1 + \theta)$	$q(1 - p(1 + \theta))$	$1 - p$
1	$p(1 - q(1 + \theta))$	$pq(1 + \theta)$	p
sum	$1 - q$	q	1

Table 4: Joint bivariate distribution. Note that admissible values for θ are those leading to a proper probability distribution.

main advantages of this parametric distribution is that we can estimate θ from data and study the impact of dependence as a function of θ. Note that

$$\mathrm{corr}(x_1, x_2) = \frac{pq\theta}{\sqrt{p(1-p)}\sqrt{q(1-q)}}$$

If $\theta > 0$ there is positive dependence and then $L_{dep}(0,0) > L_{ind}(0,0)$ and $L_{dep}(1,1) > L_{ind}(1,1)$ whereas the opposite inequalities apply for (1,0) and (0,1). This allows us to conclude on the impact of dependence depending on whether $\theta > 0$ or not. For instance, the ratio of LR assuming independence to the one for dependence is $(1 + \theta)$ when $x_1 = x_2 = 1$.

B Balsac genealogies

Our analysis is based on a cohort of 2,077 individuals, with a nearly balanced sex distribution (50.5% male [n = 1,049] and 49.3% female [n = 1,023]). Analysis of mortality patterns yields a mean age at death of 47.2 years and a median of 55 years, with 25% of individuals succumbing before the age of 18. Despite evidence of elevated early-life mortality, the majority of deaths occur between 55 and 72 years, and the maximum

observed lifespan reaches 102 years (Figure S1A). The birth years ranged from 1665 to 1822 (Figure S1B). Examination of half-sibling relationships reveals a marked asymmetry: 62.5% (n = 346) of half-sibling pairs share a paternal lineage, while only 37. 5% (n = 208) are maternal half-siblings. For anonymity purposes, in the implemented models in `mispitools` we do not add the real datasets, but only summary metrics (frequency distribution of traits) obtained from it.

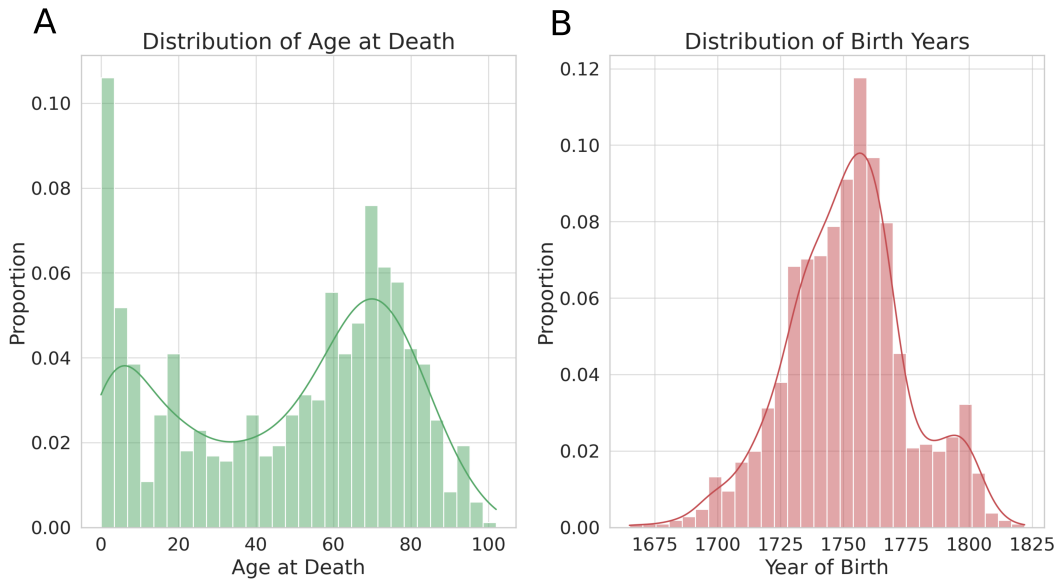

Figure S1: Demographic and genealogical distributions from the BALSAC database. The first panel (A) Distribution of Age at death (B) Year of birth distribution.

C Sensitivity analysis

To evaluate the robustness of the LR models, we performed a sensitivity analysis systematically varying key model parameters. In the sex-based analysis (Figure S2A), the error parameter (ε) was varied from 0.01 to 0.4 while the female population proportion was adjusted from 0.1 to 0.4. For the LR based on age (Figure S2B), with the missing person age (MPa) fixed at 40 and the simulation parameters kept constant (that is, $\gamma = 0.07$ and $\varepsilon_{\text{age}} = 0.05$), the age error range or also called the matching window varied from 1 to 20. In the hair colour analysis (Figure S2C), the baseline hair colour error

(ε) varied from 0.01 to 0.4 and the population proportion for hair colour 1 (p1) varied between 0.1 and 0.2 (with the remaining proportions scaled proportionally to sum to 1). In each scenario, 100,000 simulations were performed under both hypotheses: H_1 (the unidentified individual is the missing person) and H_2 (the unidentified individual is not the missing person), and the median LR was calculated.

The results reveal a clear trend: as the uncertainty increases (for example, with higher error rates), the LR values under H_1 decrease while those under H_2 increase. This behavior indicates that the model effectively captures the diminishing weight of the evidence as the uncertainty grows. In addition, a similar pattern is observed when the specific trait value (female in Figure S2A and the selected hair colour in Figure S2C) becomes more prevalent. This finding reinforces a well-known principle in forensic analysis: When the observed trait is less common, it is more informative and contributes greater weight to the evidence.

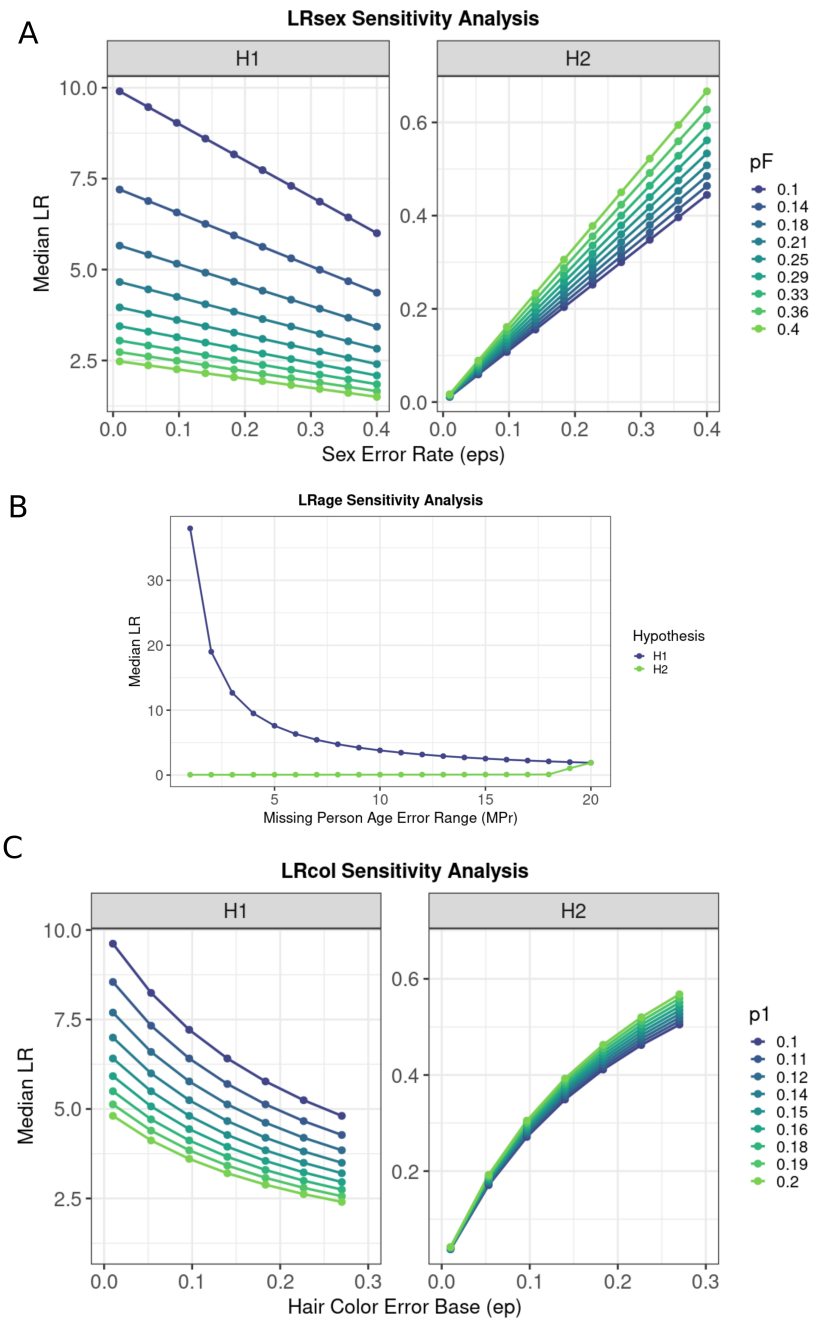

Figure S2: Sensitivity analysis of likelihood ratio estimates for (a) sex, (b) age, and (c) hair colour. For LRsex, the error rate (ε) and female population proportion (pF) were varied from 0.01–0.4 and 0.1–0.4, respectively. For LRage, with age fixed at 40, the age error range was varied from 1 to 20. For LRcol, the hair colour error (ε) ranged from 0.01 to 0.4 and the proportion for hair colour 1 (p1) from 0.1 to 0.2. Median LR values (computed from 500 simulations) are displayed for both hypotheses (H_1 and H_2), illustrating the impact of parameter uncertainty on LR estimates.