Forensic Science International: Genetics 31 (2017) 57–66

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Evaluating the statistical power of DNA-based identification, exemplified by 'The missing grandchildren of Argentina'

Daniel Kling[a,*], Thore Egeland[b], Mariana Herrera Piñero[c], Magnus Dehli Vigeland[d]

[a] Department of Forensic Services, Oslo University Hospital, Oslo, Norway
[b] IKBM, Norwegian University of Life Sciences, Ås, Norway
[c] Banco Nacional de Datos Genéticos, Ciudad de Buenos Aires, Argentina
[d] Department of Medical Genetics, Oslo University Hospital, Oslo, Norway

ARTICLE INFO

Article history:
Received 23 May 2017
Received in revised form 1 August 2017
Accepted 4 August 2017
Available online 12 August 2017

Keywords:
Missing persons
Familias
Paramlink
Simulations
DVI
Family reunification
DNA-based identification
Missing grandchildren

ABSTRACT

Methods and implementations of DNA-based identification are well established in several forensic contexts. However, assessing the statistical power of these methods has been largely overlooked, except in the simplest cases. In this paper we outline general methods for such power evaluation, and apply them to a large set of family reunification cases, where the objective is to decide whether a person of interest (POI) is identical to the missing person (MP) in a family, based on the DNA profile of the POI and available family members. As such, this application closely resembles database searching and disaster victim identification (DVI).

If parents or children of the MP are available, they will typically provide sufficient statistical evidence to settle the case. However, if one must resort to more distant relatives, it is not *a priori* obvious that a reliable conclusion is likely to be reached. In these cases power evaluation can be highly valuable, for instance in the recruitment of additional family members.

To assess the power in an identification case, we advocate the combined use of two statistics: the *Probability of Exclusion*, and the *Probability of Exceedance*. The former is the probability that the genotypes of a random, unrelated person are incompatible with the available family data. If this is close to 1, it is likely that a conclusion will be achieved regarding general relatedness, but not necessarily the specific relationship. To evaluate the ability to recognize a true match, we use simulations to estimate exceedance probabilities, i.e. the probability that the likelihood ratio will exceed a given threshold, assuming that the POI is indeed the MP. All simulations are done conditionally on available family data. Such conditional simulations have a long history in medical linkage analysis, but to our knowledge this is the first systematic forensic genetics application. Also, for forensic markers mutations cannot be ignored and therefore current models and implementations must be extended. All the tools are freely available in Familias (http://www.familias.no) empowered by the R library *paramlink*. The above approach is applied to a large and important data set: 'The missing grandchildren of Argentina'. We evaluate the power of 196 families from the DNA reference databank (Banco Nacional de Datos Genéticos, http://www.bndg.gob.ar. As a result we show that 58 of the families have poor statistical power and require additional genetic data to enable a positive identification.

1. Introduction

The missing grandchildren of Argentina are a well-known collection of missing person cases [1]. From 1976 to 1983, Argentina suffered a military civic dictatorship. It is estimated that 30,000 people were kidnapped, sent to clandestine centers, tortured and murdered. Most of them are still missing. Many women were pregnant at the time of abduction. Rapes were also common, frequently resulting in pregnancies. Children kidnapped with their parents or born in captivity were killed or delivered to families related to or from the military forces as "war booty", and their identities were forged. In most cases their biological parents were murdered and their bodies still remain missing.

[*] Corresponding author at: Oslo University Hospital, Department of Forensic Services, P.O. Box 4950, Nydalen, NO-0424, Norway.

E-mail addresses: daniel.l.kling@gmail.com (D. Kling), thore.egeland@nmbu.no (T. Egeland), mherrera@mincyt.gob.ar (M.H. Piñero), magnusdv@medisin.uio.no (M.D. Vigeland).

http://dx.doi.org/10.1016/j.fsigen.2017.08.006
1872-4973/© 2017 Elsevier B.V. All rights reserved.

As the regime eventually ended by 1983, grandmothers of the abducted children (organized in the association *Abuelas de Plaza de Mayo*) started to enquire the scientific community and the new democratic government to aid them in the search for their missing grandchildren. In 1987 the *Banco Nacional de Datos Genéticos* (BNDG) was created through the passing of specific laws (http://www.bndg.gob.ar). Since then, the BNDG has collected DNA from relatives (mainly grandparents, uncles/aunts and siblings) of appropriated children, and performed genetic analyses on thousands of children suspected to be one of the missing grandchildren. As of March 2017 such DNA analyses have contributed to the identification of 75 children. The BNDG has also assisted children born after abduction, through the aid of DNA studies, to register them with the surname of their true family.

The use of DNA in missing person identifications has gained increasing focus during the last two decades [2–10]. The success rate is generally high given that good quality genetic profiles can be generated. Since the early reports on the use of DNA in mass identifications by Olaisen et al. [11], several papers have increased our understanding of the pitfalls and precautions that need to be taken when evaluating the weight of evidence [5–7,12,13]. There are several complications that the forensic scientists are faced with, for instance degraded DNA with dropouts, complex family structures, large-scale comparisons and inconsistencies due to mutations. A particularly important problem is choosing relatives for identification; see for instance Ge et al. who performed a large simulation study to find the optimal family members [12]. Parents (or children) are the most informative, but these may not be available or may even be among the missing persons. In general it is a non-trivial problem to evaluate when enough genetic data has been collected to enable a robust conclusion when testing for a match. By ignoring known genotypes, computationally simple *unconditional* simulations can be performed, giving an idea of the average statistical power of a certain pedigree, for a given subset of

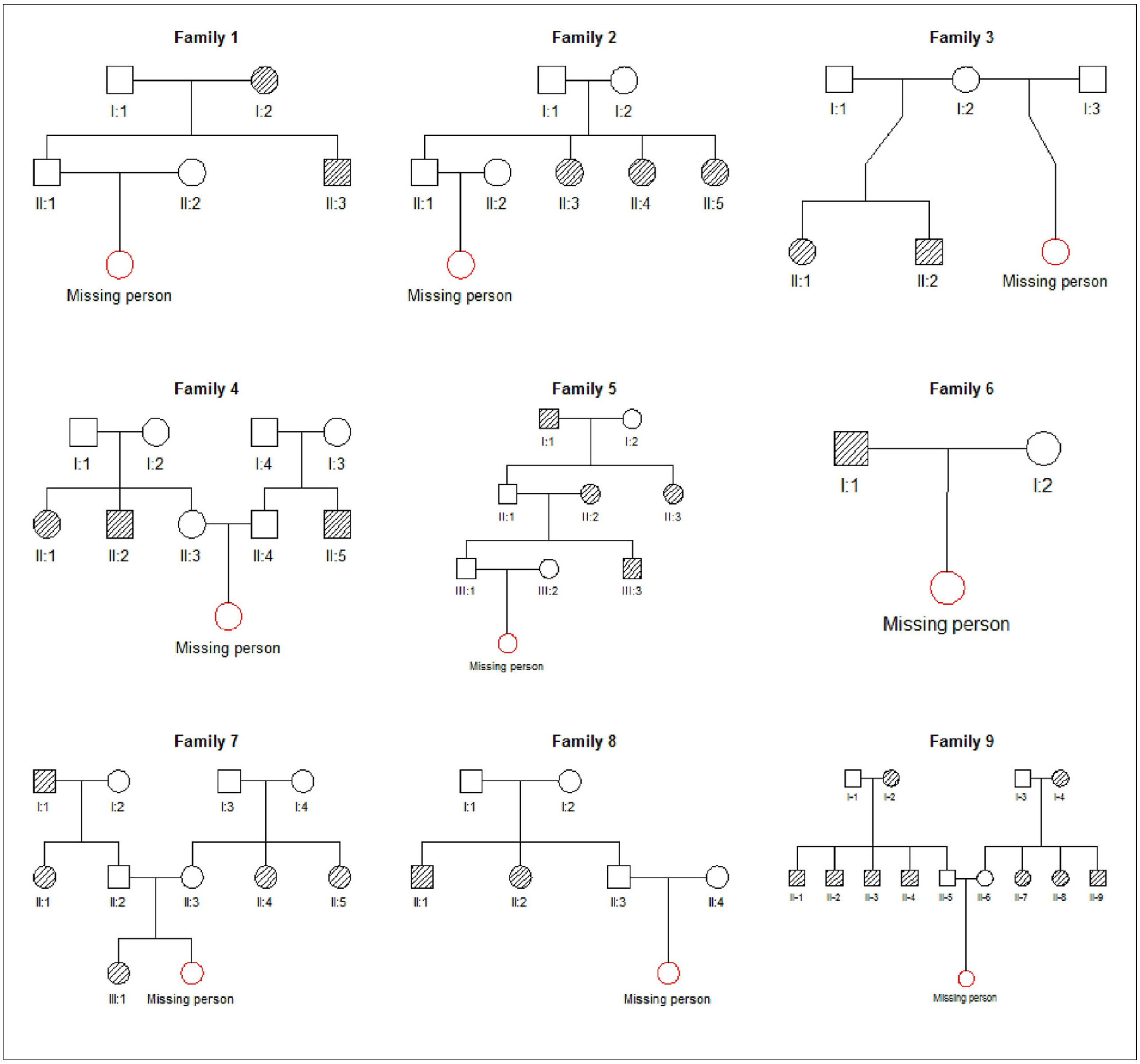

Fig. 1. Examples of reference families from the BNDG database, illustrating the variety of pedigree structures and available data. Hatched individuals are genotyped.

available individuals and a given marker kit. However, to obtain accurate estimates in a specific case, all the available information should be included, requiring *conditional* simulations. In this paper we describe and exemplify conditional simulation in models accommodating for mutations, and use this to evaluate a selection of the BNDG families. To our knowledge this is the first systematic application of such simulations. Some of the reference families in BNDG display prior inconsistencies that are most reasonably explained by mutations. It is therefore crucial that models and implementations allow for mutations. Our approach applies to any other DNA based identification process. We provide an implementation in the latest version of the freely available software Familias [14,15] available at http://www.familias.no, making use of the R library *paramlink* [16].

2. Material

The BNDG database currently contains more than 300 families searching for a missing family member. This study covers 196 of those reference families including 25 with successful reunification (the families not included in this study are still subject to thorough examination and verification of family bonds, genetic profiles etc). The task is to determine whether a given person of interest (POI) is identical to the missing person (MP) of a given family. In the present study only families with a single missing member are considered, but the methods and implementation can account for extended scenarios involving several missing persons in the same family. Fig. 1 shows a selection of reference families in the BNDG database. Summary statistics of all 196 reference families included in this study are given in Tables 1–2, providing an overview of the number of typed relatives and their relationship to the missing person. For some families, biological bonds are uncertain. For instance, paternity may be disputed and therefore, as a countermeasure, families are sometimes divided into different parts analyzed individually. Over the years several marker kits have been used for genotyping. Historically, each individual was typed with an average of 15 autosomal STR markers, while efforts are now being made to extend the number of markers to 24.

Conveniently for this study, a majority of reference families in the databank are stored in Familias files. Some of the previously identified individuals and their families had to be reconstructed in Familias for the purpose of this study.

Table 1
Summary statistics of the 196 reference families included in this study. The first column indicates the number of typed individuals, and the remaining columns show the number of families with the given properties. Parent(s): available data from one or both parent; 1st degree: available data from at least one parent or full sibling; 2nd degree: closest available relative is of 2nd degree (includes aunts, uncles, nieces, nephews, grandparents and halfsiblings). Exclusion possible: families where exclusion (of an unrelated POI) is theoretically possible using autosomal markers.

Typed	Families	Parent(s)	1st degree	2nd degree	Exclusion possible
1	11	5	6	5	5
2	20	5	5	15	14
3	29	7	8	21	22
4	41	6	8	33	33
5	29	7	15	14	26
6	26	5	12	14	21
7	14	0	6	8	13
8	14	0	5	9	9
9	5	0	2	3	2
10+	7	0	2	5	6
Total	**196**	35	69	127	151

Table 2
Summary statistics of the missing persons in the selected reference families. mtDNA = mitochondrial DNA available from female relatives; Y = Y-chromosomal data available from male relatives. The columns reflect families where these data may be used to confirm identification.

MP	Total	mtDNA	Y
Male	27	27	25
Female	17	17	–
Unknown	152	142	108
Total	**196**	**186**	**133**

3. Methods

The identification cases considered in this paper involve families with a single missing person. For a given person of interest we consider the following hypotheses:

H_1: POI is MP

H_2: POI is unrelated to MP

After genotyping the POI together with one or more family members, a statistical comparison of the above hypotheses is typically done using the likelihood ratio (LR):

$$LR = \frac{P(data|H_1, \varphi)}{P(data|H_2, \varphi)}.$$

Here *data* means the genotype data, and φ is a set of model parameters (such as allele frequencies and mutation rates) which will be assumed to be the same for all hypotheses. The probability $P(data \mid H, \varphi)$ is called the likelihood of H given the data, and sometimes written as $L_\varphi(H \mid data)$. When φ is clear from the context, we will omit it from the notation. In our context each hypothesis H is equivalent to specifying a (not necessarily connected) pedigree, and we may regard $L(H \mid data)$ as a *pedigree likelihood*. Several well-known methods for computing pedigree likelihoods exist of which the Elston-Stewart algorithm [17] is most suitable for forensic applications, which often involve markers with many alleles, mutations and complex pedigrees. Versions of the Elston-Stewart algorithm form the core of the software Familias [14] and the R library *paramlink* [16], which are used for the computations in this paper.

3.1. Probability of exclusion

We say that a hypothesis H can be excluded if the available genotype data is incompatible with the pedigree specified by H, or equivalently, if $L_\varphi(H \mid data) = 0$. Note that exclusion is possible only in certain cases (e.g. H_2 above can never be excluded) and only for certain model parameters φ (e.g. mutations must be assumed absent). Standard applications of exclusion include paternity testing (an alleged father is excluded if he does not share at least one allele with the child at each marker) and sibship testing (if in a group of individuals more than 4 alleles are observed at a single locus, the individuals cannot all have the same parents). In the context of comparing a POI with relatives of MP, we define the *probability of exclusion (PE)* as

$$PE = P(L(H_1|data) = 0|H_2).$$

Note that PE is computed on the basis of the typed family members, but before data from POI is available. If PE is close to 1, an unrelated POI can most likely be confidently excluded, while the opposite is true if PE is close to zero. In such families more members should be recruited if possible.

As shown in [16,18] *PE* can be computed exactly even in complex cases, by the formula

$$PE = \sum_g P(g|H_2)I_1(g).$$

The sum is over all conceivable genotypes g for POI at a single marker. $I_1(g)$ is the indicator of incompatibility, defined to be 1 if $P(g|H_1) = 0$, and 0 otherwise. All genotype probabilities involved are conditional on the available data from typed relatives of MP. The *PE* values presented in this paper were computed with the *paramlink* function *exclusionPower()*, which implements the above formula.

3.2. Exceedance probabilities

The exclusion probability discussed above gives a measure of the true negative rate of the testing procedure. Of equal importance is the true positive rate, i.e. the probability of a correct conclusion if the POI is in fact the missing person. The situation is somewhat different in nature than for exclusion, since a genetic test can never provide 100% certainty of a true match. In our cases we have used LR > 10,000 as the threshold for declaring a positive match.

For a given family the likelihood ratio LR can be viewed as a stochastic variable dependent on the marker genotypes of POI. We define the exceedance probability E_t for any $t > 0$ to be the probability that LR exceeds the threshold t, given that the POI is indeed the missing person:

$$E_t = P(LR > t|POI = MP).$$

The distribution of LR can be computed exactly in certain simple cases of pairwise relationships, as discussed by Kruijver et al. [19]. In the general case, however, simulation is the natural approach. Specifically in our context, we simulate the genotype of MP conditional on the known genotypes in the family, as we describe next.

3.3. Conditional simulation

Conditional simulation of marker genotypes has existed for several decades [20,21] but seems not to be widely known in the forensic community. Furthermore it is not offered by standard forensic software, possibly because of its computational complexity. The interaction between Familias and the *paramlink* package aims to rectify this, as the *markerSim()* function of *paramlink*

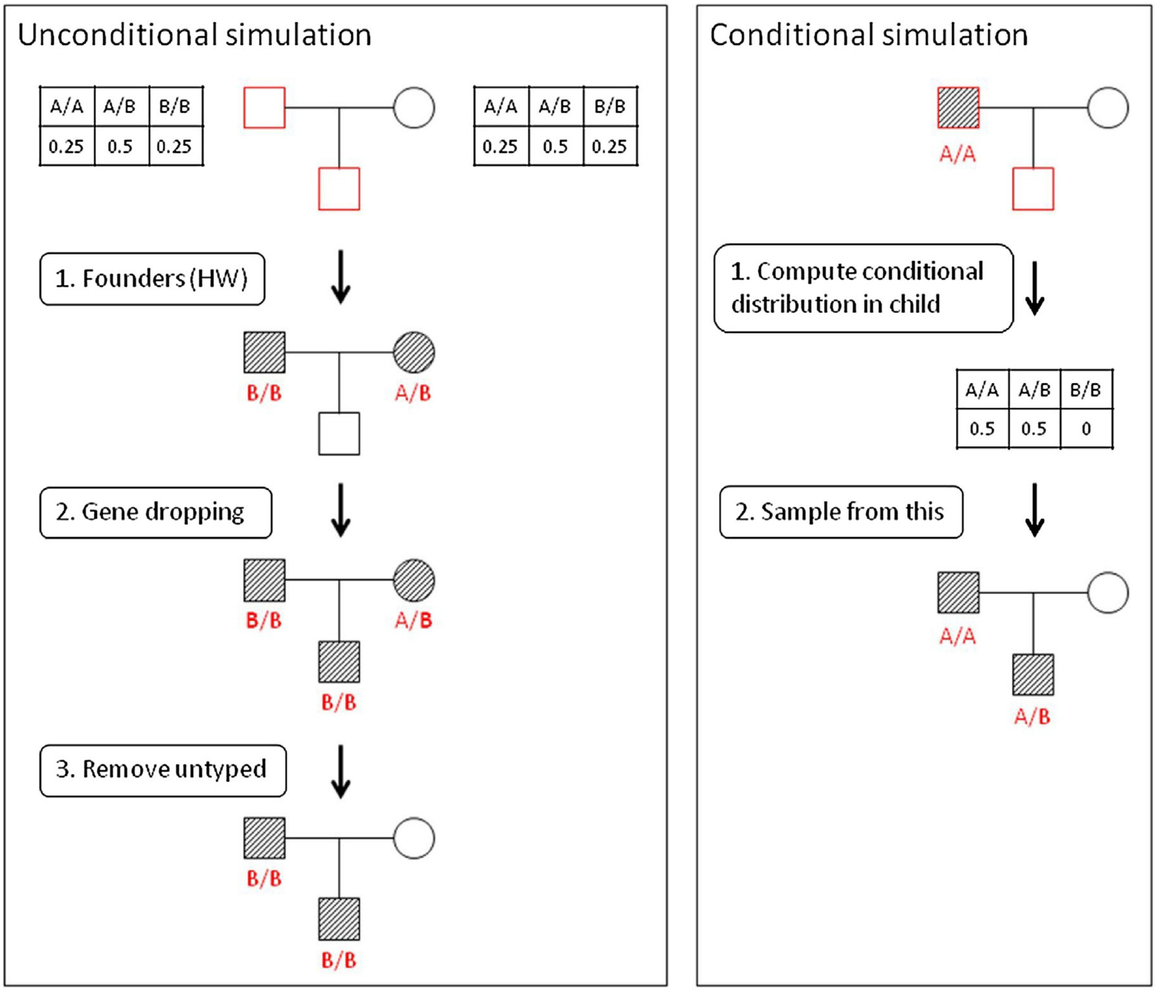

Fig. 2. Comparison of two approaches to simulate genetic data. For a SNP with equifrequent alleles A and B, the father is known to have genotype A/A, while the mother is unavailable for genotyping. To the left *Unconditional simulation* where founders are sampled from the Hardy-Weinberg (HW) distribution of genotypes followed by *gene dropping*, where a random allele is chosen from each parent to form the genotype of the child. To the right *Conditional simulation,* where the genotype distribution of the child is computed conditionally on the father's genotype. We note that given the genotype A/A of the father, the child cannot have B/B (as produced in the unconditional simulation) unless mutations are considered.

performs conditional simulations in virtually any pedigree. Furthermore, *markerSim()* allows modeling of mutations, which is essential in cases where the conditions (i.e. the known genotypes) contain Mendelian inconsistencies.

A basic algorithm [21] for conditional simulation of a single marker can be outlined as follows. Let the variable G_i be the genotype of individual i, and let K denote the set of currently known genotypes in the pedigree. Let T denote the target individuals, i.e. the subset of pedigree members we want to simulate.

1. Choose any untyped individual j in T

2. Compute the conditional distribution of G_j using Bayes formula:

$$P(G_j = g|K; pedigree) = \frac{P(G_j = g, K|pedigree)}{P(K|pedigree)}.$$

In the expression on the right hand side, both numerator and denominator are standard pedigree likelihoods and thus computable as discussed previously.

3. Sample G_j from the distribution computed above and include this in the set K of known genotypes. Repeat from point 1 until all members of T are simulated.

See Fig. 2 for a simple illustration of how unconditional simulations compare with the conditional simulations described above. The above conditional procedure is highly computer intensive due to the many likelihood computations necessary. To overcome this, the *paramlink* version of the algorithm implements several speedups, combining different simulation strategies in order to minimize the number of likelihood computations. These strategies include computing joint genotype distributions for selected individuals and gene dropping [22] where this is possible.

An additional complexity is given when the simulation model accounts for mutations. Mutations are accounted for by using proper transition models like the stepwise model reviewed in [23]. We refer to Ellegren et al. [24,25] for a general introduction. In our applications we have, for computational reasons, used a simple mutation model whereby each mutation is assigned the same probability. The mutation rates for all genetic markers were obtained from STRBase (http://strbase.nist.gov).

4. Results

We begin with a selection of examples illustrating how conditioning on the available genotypes affects the results. Finally, we present an extraction of the results from the evaluations of the reference families in the BNDG, outlined in the Material section.

4.1. Illustrative examples

Example 1. A standard duo case is studied using data for an autosomal marker with two alleles, A and B, whose population frequencies are given by p and q respectively. The genotype of the father (F) is known to be A/A, while the mother is unavailable. Fig. 2 explains how simulations could be performed in this case in order to investigate the distribution of the LR. In fact, this scenario is simple enough to allow exact calculations. Below are the unconditional (see [18] for details) and conditional expectations and variances for the LR assuming the hypothesis H_1 to be true, i.e. that the paternity is correct. In these expressions we use the notation E(X) and var(X) for the expectation and variance of a stochastic variable X.

$$\mathrm{E}(LR(H_1)) = \frac{5}{4} = 1.25,$$

$$var(LR(H_1)) = \frac{1}{16}\left(\frac{1}{pq} - 1\right),$$

$$E(LR(H_1)|F = A/A) = 1 + \frac{1}{2}\frac{q}{p},$$

$$var(LR(H_1)|F = A/A) = \left(\frac{1}{p} + \frac{1}{4}\frac{q}{p^2}\right) - \left(1 + \frac{1q}{2p}\right)^2.$$

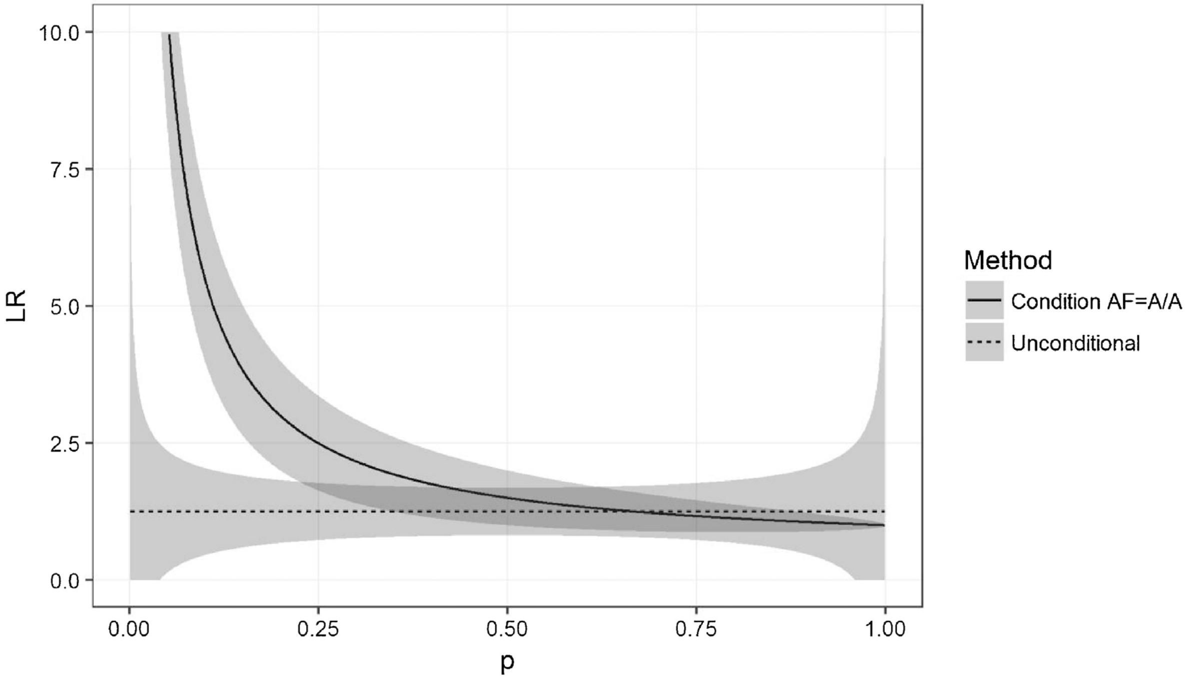

Fig. 3. The conditional and unconditional expectations of LR in Example 1. The bands correspond to one standard deviation. For small p, the unconditional estimate fails badly, leading to a large mean square error.

See Fig. 3 for a graphical representation of these expressions. The expected value 1.25 for unconditional simulation above can be interpreted as follows: Assume father-son pairs are simulated unconditionally. For each simulation an LR is calculated. The average value of these simulated LRs will be close to 1.25 if a sufficiently large number of simulations are performed. Note that, remarkably, this result is independent of allele frequencies. The unconditional variance above includes the term $1/(pq)$ and is therefore large if p (or q) is close to 0 and an extremely large number of simulations may then be needed for the average to be close to 1.25.

Given that the father is A/A, it is intuitively clear that LR(H_1) is large (in favor of paternity) if p is small, since father and son share a rare allele. For example, if p = 0.05, the above formula implies that the conditional expectation of LR is 1 + 0.5·(0.95/0.05) = 10.5.

A comparison of unconditional and conditional simulation should also account for the variance of the estimate. This can be done by the *mean squared error (MSE)*: Let $\hat{\theta}$ be an estimator of some stochastic variable θ. Then

$$MSE(\hat{\theta}) = var(\hat{\theta}) + (E\hat{\theta} - \theta)^2.$$

The second term above is the squared bias. The example above showed that this term may be large for unconditional simulation leading to a large *MSE* value even if the variance is small. If a large number of simulations are performed, *MSE* will be close to 0 for conditional simulations. However, for unconditional simulation $MSE \approx (0.25 - 0.50q/p)^2$. This grows without bound when p approaches zero, illustrating clearly the danger of ignoring available genotype information. See also Fig. 3.

Example 2. The pedigrees corresponding to the hypotheses are shown in Fig. 4. Genetic data is available from the four paternal relatives of the missing person with genetic data for 15 STR markers.

In this case 10,000 conditional simulations were performed for POI for both hypotheses. Fig. 5A illustrates the LR distributions and B the exceedance probability as a function of the logarithm of the threshold t. We have also included a dotted curve in Fig. 5B showing the probability that the LR exceeds t given that POI and MP are unrelated. Fig. 5 demonstrates that there is little risk of a false positive match in this family, but also that there is most likely not sufficient available data to achieve LR > 10,000 even if the true person is matched. In other words, more markers and/or additional relatives are needed in this case.

Example 3. The first part of this example addresses exclusion probabilities. Consider the hypotheses in Fig. 6. Genotypes for the first five markers are shown. Initially we study marker one and assume that mutations are not possible for *this* marker. Note that there are some *forced genotypes*: Both individuals “6” and ‘8’ must be 17/18. Therefore, a random person can be excluded as the MP if she does not have allele 17 or 18 (since MP must share an allele with person ‘8’). Given the allele frequencies $p_{17} = 0.161$ and $p_{18} = 0.117$, we find the probability of exclusion for this marker to be $(1 - p_{17} – p_{18})^2 = (1 - 0.161 – 0.117)^2 = 0.52$.

Consider next the second marker. As we do not question family relationships, and disregard artefacts like drop-out and genotyping error, there must be a mutation in the transition from ‘4’ to ‘5’, most likely 7 to 8, but possibly 9.3 to 8. Our implementation includes all mutation models of Familias. We will use the *proportional mutation* model in this example for markers displaying inconsistencies. The probability of mutating to an allele z is then proportional to the population frequency of z. In addition to being easy to work with mathematically and computationally, it satisfies desirable properties such as *detailed balance*, which implies *stationarity*, see Egeland et al. [26] for a more complete discussion. Obviously, the probability of exclusion is 0 when mutations are modeled in the model. Based on all 24 markers, the exclusion probability is 0.991 or 99.1%.

Next we simulate the likelihood ratio $LR = P(data|H_1)/P(data|H_2)$, using all 24 markers. When we simulate assuming H_2 as true, we expect 99.1% of the values to be 0 based on the exclusion probability calculated above; for our simulations 98.7% of values

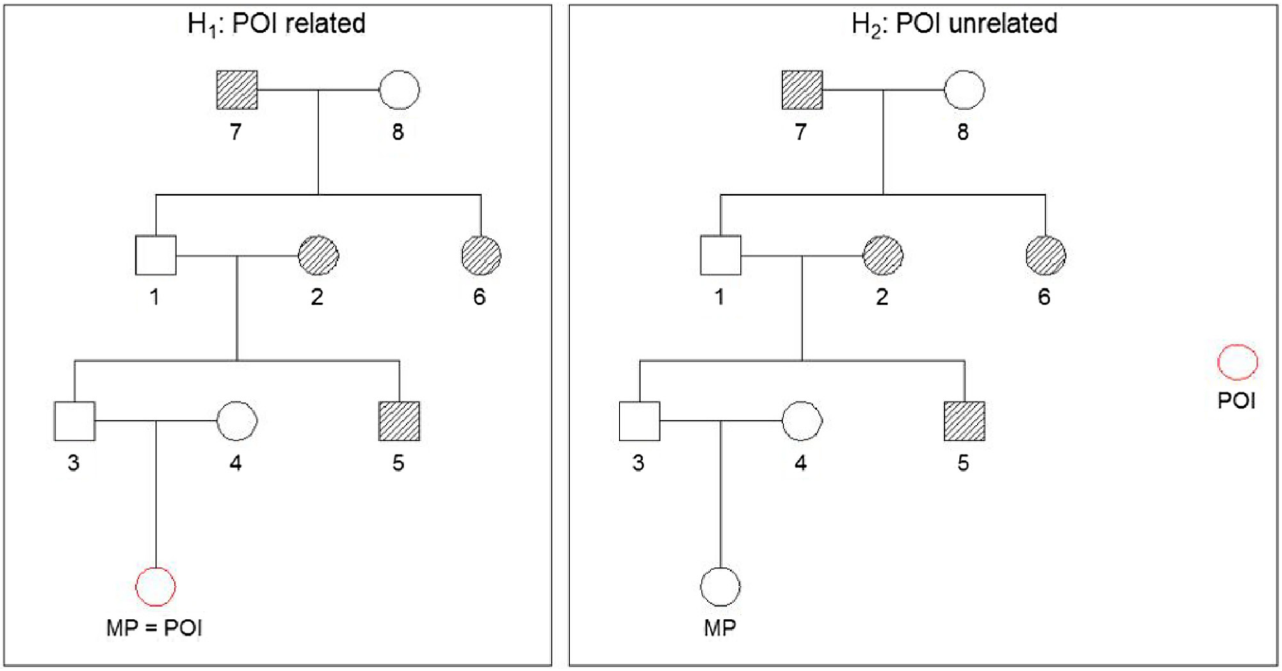

Fig. 4. Hypotheses and genotyped individuals for Example 2.

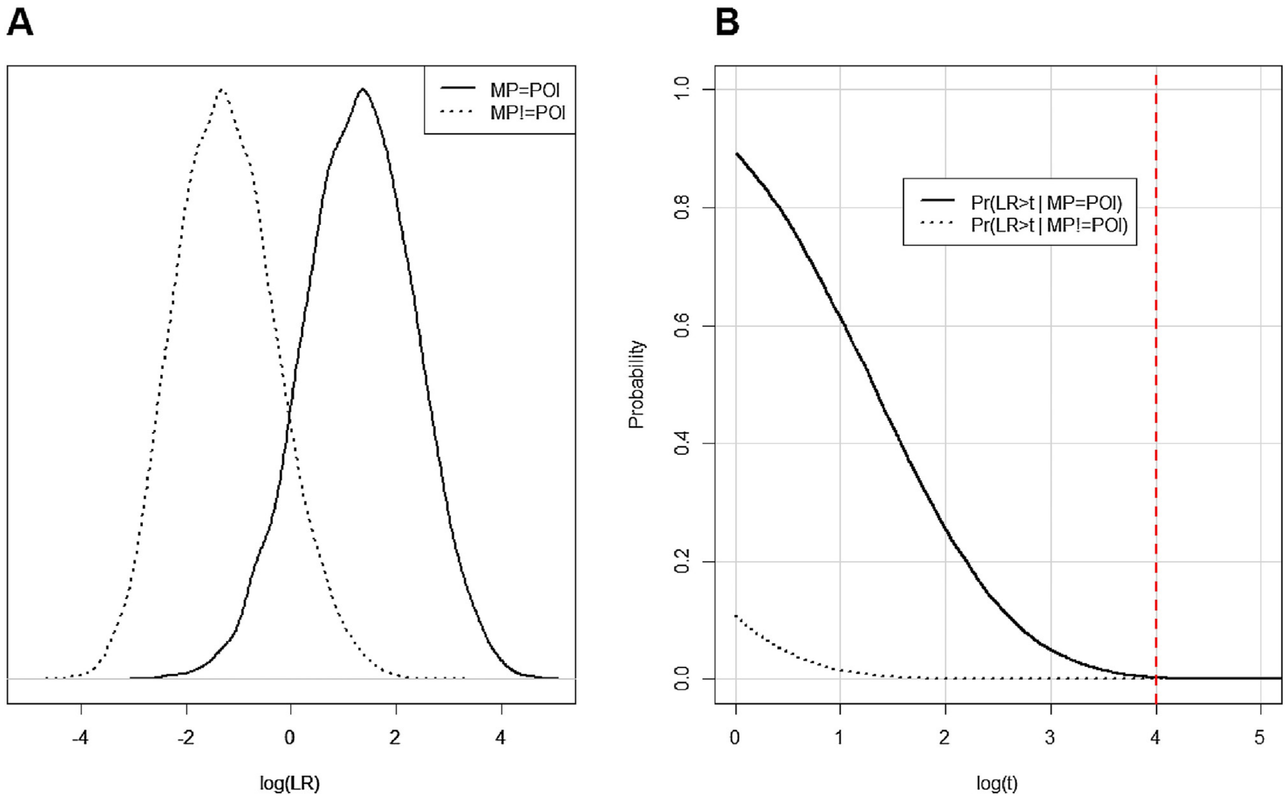

Fig. 5. LR distributions for Example 2 computed by conditional simulation, assuming either H_1 (solid curve) or H_2 (dotted curve). A – Density distributions of log(LR). B – Exceedance probabilities for LR. The red dashed line indicates the LR threshold t = 10,000.

equaled 0 which agrees well with the exact value. Fig. 7 shows the distribution of $\log_{10}(\text{LR}(H_1))$. The probability that $\log_{10}(\text{LR}(H_1))$ exceeds 4, i.e. LR > 10,000, is estimated at 0.992. In conclusion, these calculations indicate that there is sufficient power to solve this case.

4.2. Results from the reference families

We performed conditional simulations on all unsolved cases – in total 171 – in our selection of 196 reference families from the BNDG database. Fig. 8 shows the exclusion probability PE plotted against the estimated exceedance probability $E_{10000} = P(\text{LR} > 10{,}000 \mid H_1)$ for each family. As expected, cases with a high exclusion probability also tend to have a high exceedance probability as both statistics increase with the amount of genetic data available. One notable exception from this is the family with $(E_{10000}, \text{PE}) \approx (0.06, 0.85)$, where 8 members have been typed. A closer look at the pedigree (Fig. 9) reveals that the typed relatives are all on the maternal side, and most of them distant to the missing person, explaining the low power in this case. In total 58 families, one third of those investigated, have $E_{10000} < 0.8$, indicating low power and the necessity for additional genetic data. In contrast, 68 families have both E_{10000} and PE greater than 0.99 indicating excellent statistical power both for positive matching and exclusion. Included among these are all cases where parental data is available, except one case where few markers were typed. The remaining 45 families have decent power for positive identification, with $0.8 < E_{10000} < 0.99$. Most of these also have high PE, with a few exceptions where the particular configuration of typed relatives make exclusion theoretically impossible.

Finally, we demonstrate how the LR, obtained in 25 of the successfully reunited grandchildren, compares to the median and the 95% credibility intervals obtained with the conditional/regular simulations. This serves as validation of our implementation and as a comparison of conditional versus unconditional simulations. As shown in Fig. 10 all the actual LRs, except one, fall within the credibility intervals. This is exactly as expected. We further note that for some families the difference between conditional and unconditional simulations is big. Exploring the reasons reveal rare or common alleles for the typed family members either deflating or inflating the expected LR and distributions. Supplementary Table S1 summarizes the reference families with pedigrees.

5. Discussion

This paper presents the efforts conducted in a collection of missing person cases, the 'Missing grandchildren of Argentina'. In Argentina these cases are considered *Crimes Against Humanity* and the recovered grandchildren are fundamental proofs in trials against military officers and members of civil society who participated in these crimes. As far as we are aware, the BNDG database of missing grandchildren in Argentina is one of the first of its kind.[1]

A majority of the BNDG reference families lack first-degree relatives of the missing person. For example, parental data was available in only 35 of the 196 families included in this study. This makes accurate power assessment an essential tool when

[1] http://www.schofieldoration.org.au/wp-content/uploads/2014/03/MTBs-Oration.pdf.

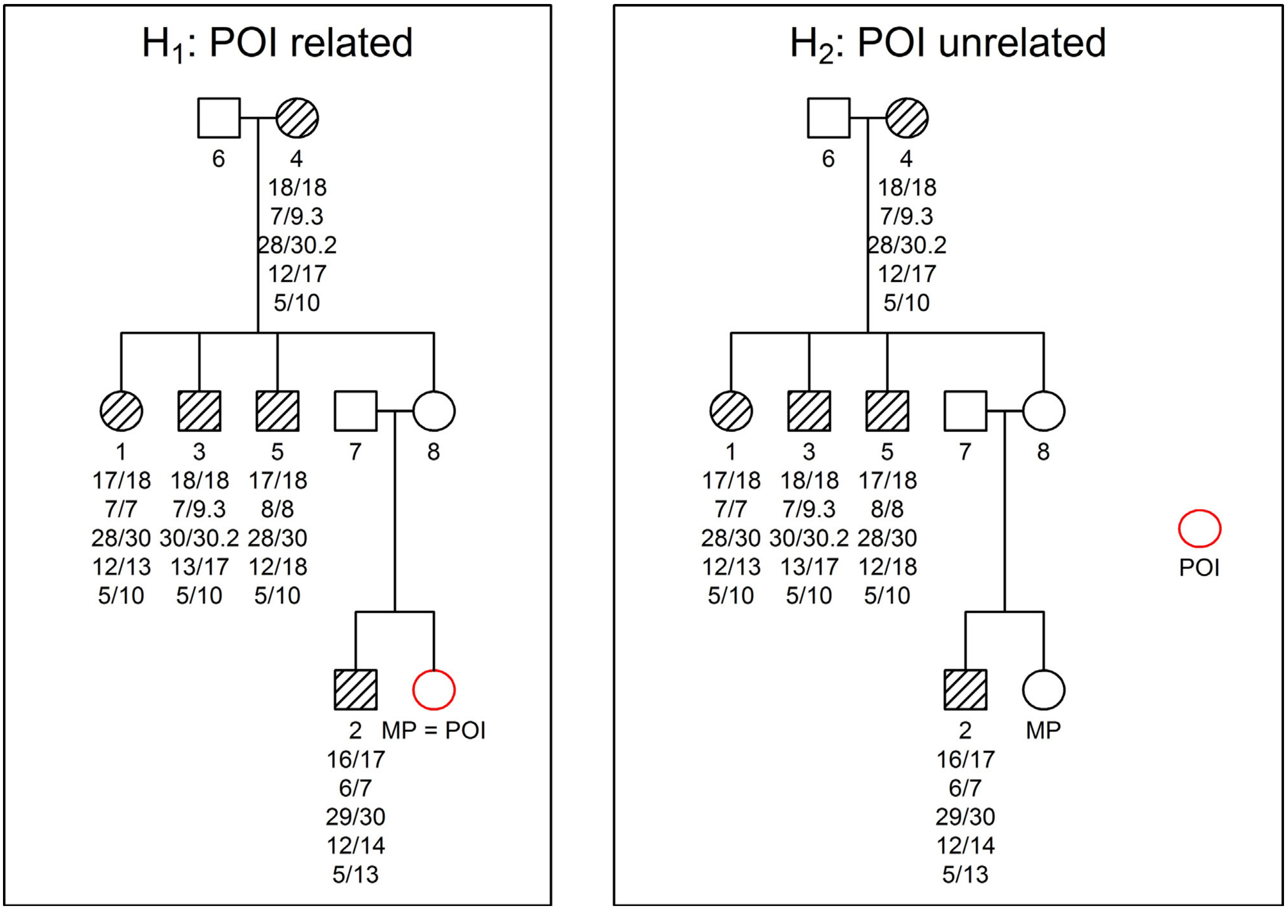

Fig. 6. Illustration for Example 3. The first five of 24 genetic markers are shown.

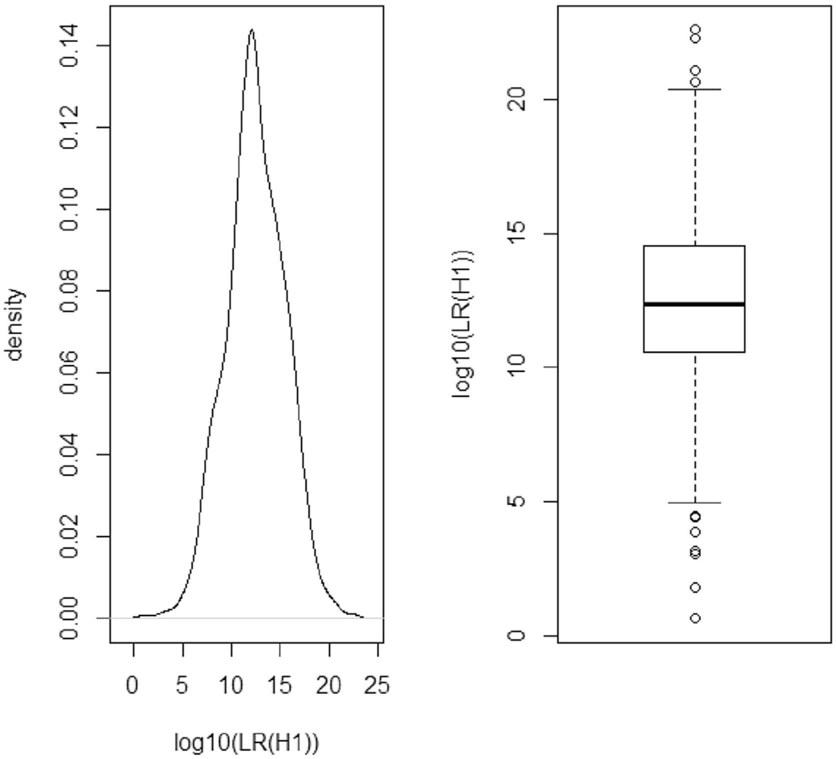

Fig. 7. Distribution of the log 10 likelihood ratio under H_1 in Example 3 shown as density and box plot.

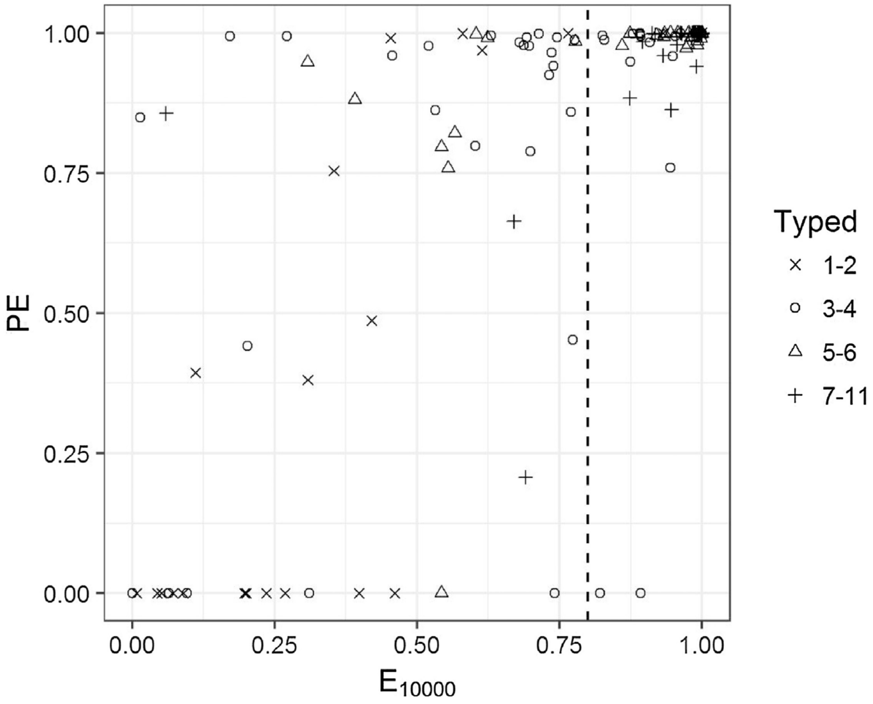

Fig. 8. The results of power evaluation of 171 reference families in the BNDG database. Each point represents a family, with the symbol indicating the number of typed individuals. The x-axis shows the exceedance probability $E_{10000} = P(LR > 10{,}000 \mid H_1)$ while the y-axis shows the probability PE of excluding an unrelated POI. $E_{10000} = 0.8$ is indicated with a dashed line.

recruiting family members. We approach the challenge by simulations and describe and exemplify two different simulation methods, the unconditional and the conditional approach. Although algorithms for conditional simulation are not as such new, implementations suitable for forensic data have not been available. By joining forces of the Familias software [14,15] with the R library *paramlink* [16], the functionality is now provided. In the conditional simulations we use all available information. The results will therefore be specific for the actual data from the reference family and is the approach we recommend when genotype data is available. The unconditional approach, on the other hand, provides a general idea of the identification power when genotype data is not available, typically prior to any recruitment of family members. Both simulation methods are implemented in the latest version of Familias (http://www.familias.no) and can be used on any number of families simultaneously.

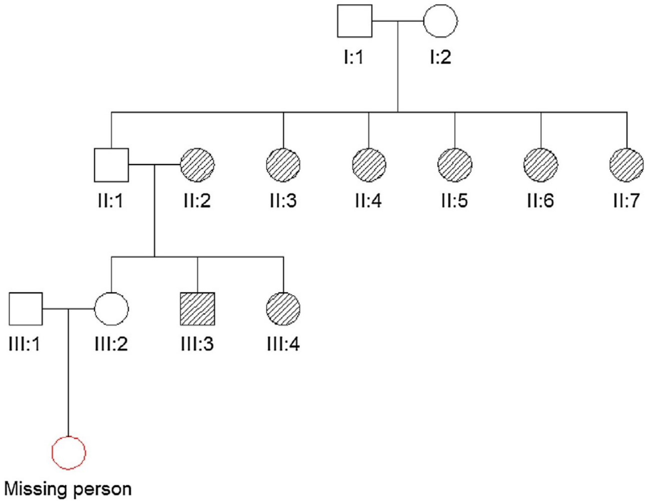

Fig. 9. Reference family from BNDG for which the probability of reunification is low despite many typed members. The typed members are all on the maternal side, and mostly distant to the missing person.

We advocate the use of two statistics in order to assess the statistical power of a given reference family. The exceedance probability E_t (that the LR exceeds a chosen threshold t) and the exclusion probability (of correctly excluding an unrelated POI). The former can be estimated by simulation, while the latter allows exact computation. In combination, these two statistics provide a good understanding of whether the available data is sufficient for successful identification.

Choosing the LR threshold for declaring a positive match is a difficult problem in many forensic applications. We do not pursue this question in this work, but conditional simulation along the lines of our methods is certainly a useful tool in this regard. We briefly mention the necessity of setting a threshold prior to conducting a search, here defined as the LR value a specific candidate (POI) needs to exceed in order to be further evaluated. This paper focuses on the LR and its properties. However, many forensic scientists would rather use the posterior probability, obtained via Bayes' theorem, when evaluating potential candidates. In fact, this probability is obtained in Familias so the end user can decide which approach to adopt.

We have not discussed the importance of lineage markers, e.g. Y and mitochondrial data in the context of missing person databases. The reference families in the databank are indeed subject to such extended genotyping and in case of a potential match, these data may be used to exclude (or include) candidates.

Finally, in light of the evaluations conducted in this study (see Fig. 8 for a summary) the BNDG has commenced efforts to genotype further autosomal markers (from 15 to 24). Furthermore an anthropology unit is created, with plans to exhume and

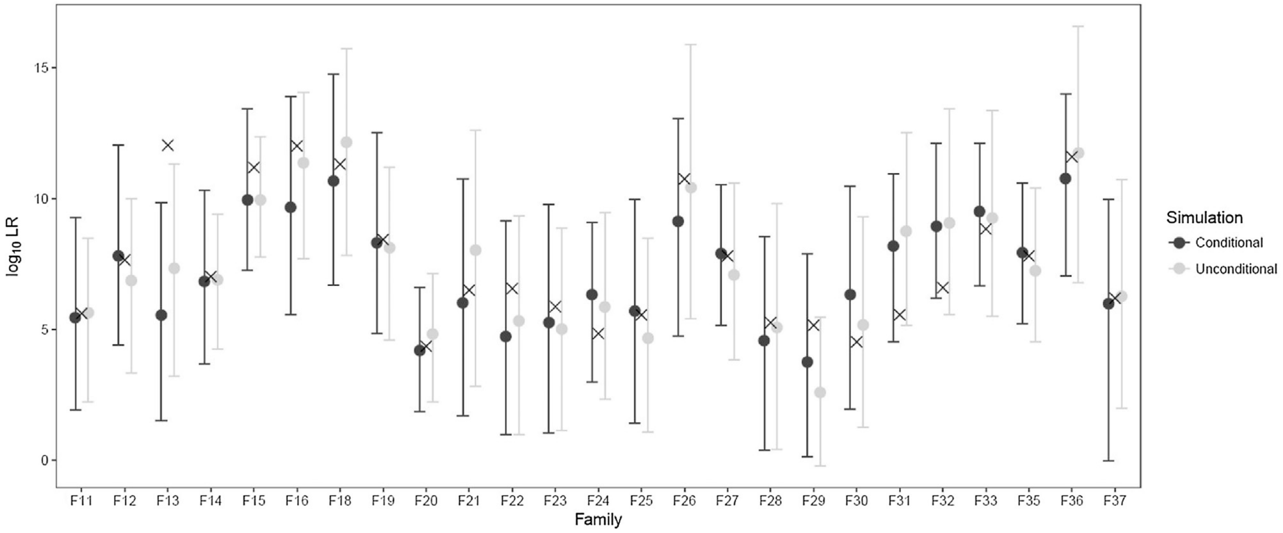

Fig. 10. 95% credibility intervals obtained by conditional/unconditional simulation for 25 reference families with successful identifications. The crosses mark the LR obtained in the identification while the dots show the median from the simulations.

genotype more than 80 deceased relatives in families with low statistical power.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2017.08.006.

References

[1] V.B. Penchaszadeh, Use of DNA Identification in Human Rights Work to Reunite Families in Latin America, in eLS, John Wiley & Sons, Ltd., 2001.

[2] S. Donkervoort, et al., Enhancing accurate data collection in mass fatality kinship identifications: lessons learned from Hurricane Katrina, Forensic Sci. Int. Genet. 2 (4) (2008) 354–362.

[3] S.M. Dolan, et al., The emerging role of genetics professionals in forensic kinship DNA identification after a mass fatality: lessons learned from Hurricane Katrina volunteers, Genet. Med. 11 (6) (2009) 414–417.

[4] L.G. Biesecker, et al., Epidemiology: DNA identifications after the 9/11 World Trade Center attack, Science 310 (5751) (2005) 1122–1123.

[5] C.H. Brenner, Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities, Forensic Sci. Int. 157 (2–3) (2006) 172–180.

[6] C.H. Brenner, B. Weir, Issues and strategies in the DNA identification of World Trade Center victims, Theor. Popul. Biol. 63 (3) (2003) 173–178.

[7] B. Budowle, et al., Use of prior odds for missing persons identifications, Invest. Genet. 2 (1) (2011) 15.

[8] M. Baeta, et al., Digging up the recent Spanish memory: genetic identification of human remains from mass graves of the Spanish Civil War and posterior dictatorship, Forensic Sci. Int. Genet. 19 (2015) 272–279.

[9] C. Phillips, et al., Ancestry analysis in the 11-M Madrid bomb attack investigation, PLoS One 4 (8) (2009) e6583.

[10] B. Rezic, et al., Identification of war victims from mass graves in Croatia: Bosnia, and Herzegovina by the use of standard forensic methods and DNA typing, J. Forensic Sci. 41 (5) (1996) 891–894.

[11] B. Olaisen, M. Stenersen, B. Mevåg, Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster, Nat. Genet. 15 (4) (1997) 402–405.

[12] J. Ge, B. Budowle, R. Chakraborty, Choosing relatives for DNA identification of missing persons, J. Forensic Sci. 56 (Suppl. 1) (2011) S23–8.

[13] C.M. Vullo, et al., GHEP-ISFG collaborative simulated exercise for DVI/MPI: Lessons learned about large-scale profile database comparisons, Forensic Sci. Int. Genet. 21 (2016) 45–53.

[14] T. Egeland, et al., Beyond traditional paternity and identification cases: selecting the most probable pedigree, Forensic Sci. Int. 110 (1) (2000) 47–59.

[15] D. Kling, A.O. Tillmar, T. Egeland, Familias 3-extensions and new functionality, Forensic Sci. Int. Genet. 13 (2014) 121–127.

[16] T. Egeland, N. Pinto, M.D. Vigeland, A general approach to power calculation for relationship testing, Forensic Sci. Int. Genet. 9 (2014) 186–190.

[17] R.C. Elston, J. Stewart, A general model for the genetic analysis of pedigree data, Hum. Hered. 21 (6) (1971) 523–542.

[18] K.-J. Slooten, T. Egeland, Exclusion probabilities and likelihood ratios with applications to kinship problems, Int. J. Legal Med. 128 (3) (2014) 415–425.

[19] M. Kruijver, R. Meester, K. Slooten, Optimal strategies for familial searching, Forensic Sci. Int. Genet. 13 (2014) 90–103.

[20] J. Ott, Computer-simulation methods in human linkage analysis, Proc. Natl. Acad. Sci. 86 (11) (1989) 4175–4178.

[21] J. Ott, G. Lathrop, SLINK: a general simulation program for linkage analysis, Am. J. Hum. Genet. 47 (1990) A204.

[22] J.W. MacCluer, et al., Pedigree analysis by computer simulation, Zoo Biol. 5 (2) (1986) 147–160.

[23] T. Egeland, D. Kling, P. Mostad, Relationship Inference with Familias and R: Statistical Methods in Forensic Genetics, Academic Press, 2015.

[24] H. Ellegren, Microsatellites: simple sequences with complex evolution, Nat. Rev. Genet. 5 (6) (2004) 435–445.

[25] H. Ellegren, Heterogeneous mutation processes in human microsatellite DNA sequences, Nat. Genet. 24 (4) (2000) 400–402.

[26] T. Egeland, N. Pinto, A. Amorim, Exact likelihood ratio calculations for pairwise cases, Forensic Sci. Int. Genet. (2017), doi:http://dx.doi.org/10.1016/j.fsigen.2017.04.018.