Joint DNA-based Disaster Victim Identification

Magnus Dehli Vigeland (magnusdv@gmail.com)
University of Oslo
Thore Egeland
Norwegian University of Life Sciences

Research Article

Keywords: computational and statistical aspects, DNA-based identification , victims , disasters

Posted Date: March 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-296414/v1

Joint DNA-based disaster victim identification

Magnus D. Vigeland[1,*] **and Thore Egeland**[2]

[1]Department of Medical Genetics, University of Oslo, Pb 4956 Nydalen, Oslo, Norway
[2]Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1433 Aas, Norway
[*]magnusdv@gmail.com

ABSTRACT

We address computational and statistical aspects of DNA-based identification of victims in the aftermath of disasters. Current methods and software for such identification typically consider each victim individually, leading to suboptimal power of identification and potential inconsistencies in the statistical summary of the evidence. We resolve these problems by performing joint identification of all victims, using the complete genetic data set. Individual identification probabilities, conditional on all available information, are derived from the joint solution in the form of *posterior pairing probabilities*.
A closed formula is obtained for the *a priori* number of possible joint solutions to a given DVI problem. This number increases quickly with the number of victims and missing persons, posing computational challenges for brute force approaches. We address this complexity with a preparatory sequential step aiming to reduce the search space. The examples show that realistic cases are handled efficiently. User-friendly implementations of all methods are provided in the R package **dvir**, freely available on all platforms.

Introduction

DNA-based *disaster victim identification* (DVI) is a rapidly developing field in forensic genetics, with important applications all around the world. Recent, high-profile cases include the after-math of the 1990-s Balkan conflicts[1], drowned migrants in Italy[2], the World Trade Center attack, USA[3], Thailand tsunami 2004[4], and the search for missing grandchildren in Argentina[5].

In a broader context, DVI involves a variety of data sources and experts from several branches of forensic science, including anthropology, odontology, pathology as well as genetics. The genetic data typically consists of *post mortem* (PM) DNA from victim samples and *ante mortem* (AM) DNA from relatives of the missing persons. Additional data like the sex and age is used if available. Extensive background and general guidelines for handling DVI problems are given in papers[6–8]. In this paper we restrict our attention to computational and statistical aspects of identification cases involving multiple victims, often called *mass identifications* in the literature. A simple example of such a case is shown in Figure 1.

Current approaches to mass identification typically employ either a (i) *one-to-one*, (ii) *PM-driven*, or (iii) *AM-driven* search strategy[9]. The one-to-one approach simply amounts to comparing each PM profile to each AM reference, looking for evidence of a close relationship. This method is widely used, at least for an initial screening, since easy cases, like direct matches and parent-child, often can be reliably resolved in this way[2,9]. In more complex cases, however, the one-to-one strategy is not sufficient. For a trivial example, observe that this method cannot identify the missing person M_1 in Figure 1, who is not genetically related to any of the reference individuals.

The PM-driven and AM-driven approaches proceed sequentially, considering one victim (PM-driven) or one family (AM-driven) at the time. We concentrate on the PM-driven in the following. Briefly, the idea is to calculate the likelihood ratios (LR) comparing a victim V to each of the missing persons. The largest LR points to the most likely match for V, and a successful identification is declared if this largest LR exceeds a prescribed threshold. Also, if *priors* are specified the LRs can be converted to *posterior probabilities*[10].

The above description glosses over several important points with substantial impact on the output solutions. For future reference we include below detailed descriptions of two possible implementations. Nevertheless, our main point is that while sequential methods may be useful in certain scenarios, they are generally not optimal and may produce inconsistent solutions.

The three approaches (i) – (iii) are all *restricted*, in the sense that they utilise only parts of the data in each step. A simple example of how this may lead to missed identifications is given in Figure 2, and the accompanying analysis. The take-home message is that the best match for one individual may obstruct the most likely overall solution. In the case of Figure 2, the missing persons (M_1) cannot be identified unless the data are considered jointly.

Another problem with restricted methods is that they may produce inconsistent results. For example, since conclusions are reached independently for each victim (or family), it may happen that a victim is classified as being the most likely member of two different families.

Our goal has been to present methods and implementations that provide consistent solutions to DVI problems, by considering

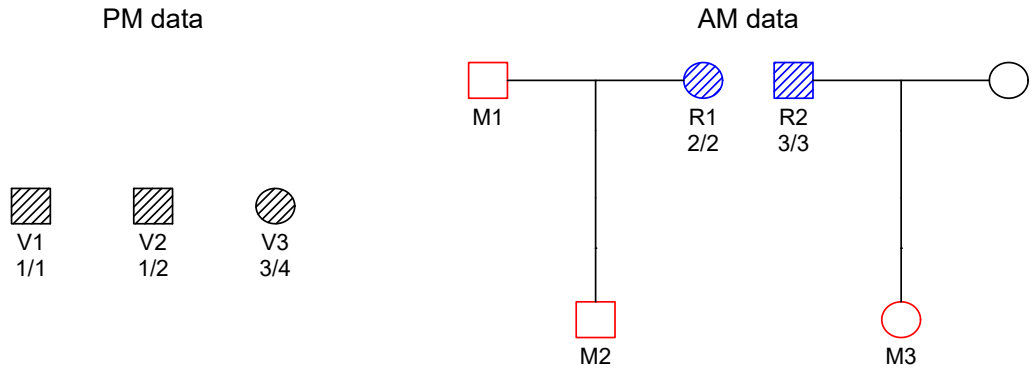

Figure 1. A toy DVI problem. The PM data consists of 3 victim samples to be matched against 3 missing persons (red) belonging to two different families. The AM data contains profiles from the reference individuals R_1 and R_2 (blue), one from each family. The hatched individuals are typed with a single marker.

all available data simultaneously. This is achieved by Algorithm 3 in the Methods section, which finds the most likely solution among all possible, while keeping the need for brute force calculations to a minimum.

Decision makers and the legal system often require independent conclusions for each missing person. In response to this, we provide formulas for *posterior pairing probabilities* for each victim - missing person pair.

To the best of our knowledge, no freely available software offer joint DVI computations. The restricted strategies mentioned above are implemented in Familias[11], but currently allow only one missing person in each family. Commercial software like Bonaparte[12,13] and DNA View[3] provide similar functionality, but precise details on the implementation are not publically available.

To rectify this we have developed the R package **dvir**, based on the **ped suite** ecosystem for pedigree analysis in R[14]. The data sets analysed in this paper are included as part of **dvir**, and further examples are given in the documentation. The source code is freely available from `https://github.com/thoree/dvir`.

Methods

The starting point of our investigations is a DVI situation involving s victim samples, hypothesised to belong to some or all of m missing persons (MPs). Identification is done by genetic matching against relatives of the missing persons, using a battery of forensic markers.

In our examples we consider independent autosomal markers in Hardy Weinberg Equilibrium (HWE). However, it should be noted that the overall approach applies very generally; in fact the only requirement is that likelihoods can be calculated. Some extensions are conceptually simple, like mutation modelling or X-chromosomal markers, while others are more challenging, as accommodating linkage disequilibrium.

We proceed to describe the input data in a bit more detail, and introduce some important notation.

PM data The *post mortem* (PM) samples are denoted $V_1, \ldots, V_s$. We assume throughout that these belong to different individuals. In practice, samples gathered from disaster victims often contain duplicates, requiring a preprocessing step in order to identify and merge these[15]. In our examples we also assume that the sex of each V_i is known, with s_F females and s_M males so that $s_F + s_M = s$. We note that this assumption is not vital to our methods, but helps to narrow the search space.

AM data The *ante mortem* (AM) data consist of one or more reference families, each containing at least one missing person (denoted $M_1, M_2, \ldots, M_m$) and at least one genotyped reference member (denoted $R_1, R_2, \ldots$). Again we assume known sex of all family members. In particular, let m_F and m_M be the number of female and male missing persons, respectively, with $m_F + m_M = m$.

A possible solution, referred to as an *assignment*, to the DVI problem we are addressing, is a one-to-one correspondence between a subset of $\mathscr{V} = \{V_1, \ldots, V_s\}$ and a subset of $\mathscr{M} = \{M_1, \ldots, M_m\}$, with the requirement that all identifications are sex consistent. For example, in Figure 1, a consistent assignment is $\{V_1 = M_2, V_3 = M_3\}$. Alternatively, we may write this more compactly as a tuple $(M_2, *, M_3)$, whose i'th element is the match for V_i, or '$*$' if the assignment does not include a match for

V_i. In the case of Figure 1 there are in total 14 assignments, as listed in the first three columns of Table 2. Note that the empty assignment $(*,*,*)$ is a valid solution, referred to as the *null model* below.

The *likelihood* $L(a)$ of an assignment a is defined as the probability

$$L(a) = P(\text{PM and AM data} \mid a, \Phi),$$

where the fixed parameters Φ include the reference pedigrees, marker allele frequencies and mutation models. To simplify the notation we write L_0 for the likelihood of the empty assignment, i.e., corresponding to the hypothesis that all victims are unrelated to all the missing persons. Moreover, we define $\mathrm{LR}_{i,j} = L(\mathrm{V}_i = \mathrm{M}_j)/L_0$ to be the likelihood ratio of the assignment $\{\mathrm{V}_i = \mathrm{M}_j\}$, giving rise to the *pairwise LR matrix*,

$$B = \begin{array}{c} \\ \mathrm{V}_1 \\ \vdots \\ \mathrm{V}_s \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{M}_1 & \dots & \mathrm{M}_m \end{array} \\ \begin{bmatrix} \mathrm{LR}_{1,1} & \dots & \mathrm{LR}_{1,m} \\ \vdots & \ddots & \vdots \\ \mathrm{LR}_{s,1} & \dots & \mathrm{LR}_{s,m} \end{bmatrix} \end{array}. \tag{1}$$

It should be noted that the likelihoods appearing in the definition of $\mathrm{LR}_{i,j}$ involve the complete PM and AM datasets. However, simpler calculations are obtained by considering the reduced DVI problem (PM_i, AM_j), where PM_i is just V_i, and AM_j consists of data from the relatives of M_j. Then it is straightforward to show that

$$\mathrm{LR}_{i,j} = \frac{P(\mathrm{PM}_i, \mathrm{AM}_j \mid \mathrm{V}_i = \mathrm{M}_j)}{P(\mathrm{PM}_i, \mathrm{AM}_j \mid \mathrm{V}_i \text{ unrelated to } \mathrm{M}_j)}.$$

In the simple case shown in Figure 1, the matrix B can be computed by hand. Let us assume that the marker has 10 alleles 1,2, ..., 10, with equal frequencies $p_1 = \cdots = p_{10} = 1/10$. We then have

$$B = \begin{array}{c} \\ \mathrm{V}_1 \\ \mathrm{V}_2 \\ \mathrm{V}_3 \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{M}_1 & \mathrm{M}_2 & \mathrm{M}_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix} \end{array}. \tag{2}$$

For example, the element $\mathrm{LR}_{2,2}$ is the LR when $\mathrm{V}_2 = \mathrm{M}_2$ is tested against the hypothesis that V_2 and M_2 are unrelated. This gives $\mathrm{LR} = p_1/(2p_1p_2) = 5$. Obviously, convincing LRs cannot be expected in this case, with only a single marker.

The zero elements of B correspond to sex-inconsistent pairings or exclusions. Furthermore, we see that the DNA data is uninformative for some of the pairings. The entries $\mathrm{LR}_{1,1} = \mathrm{LR}_{2,1} = 1$ result from the fact that M_1 is not related to either of the reference individuals, and imply that he can never be identified unless M_2 or M_3 is identified first.

The number of assignments

Let $\mathscr{A}$ be the set of all sex-consistent assignments for a given DVI problem. The total number of elements, $n = |\mathscr{A}|$, is a good measure of the problem's size, and may indicate whether a brute force approach is feasible. Consider first the situation where sex is not known neither for victims nor MPs. The total number of assignments is then

$$\sum_{k=0}^{\min(s,m)} \binom{s}{k}\binom{m}{k} k!. \tag{3}$$

The reasoning is as follows: For each k, there are $\binom{s}{k}$ different subsets of k victims. Each of these can be assigned to $\binom{m}{k}$ different subsets of the m missing persons. Finally, each assignment can be shuffled in $k!$ ways.

When the sexes are known, formula (3) applies to females and males independently, and the total number becomes

$$n = n(s_F, s_M, m_F, m_M) = \left[\sum_{k=0}^{\min(s_F,m_F)} \binom{s_F}{k}\binom{m_F}{k} k!\right]\left[\sum_{k=0}^{\min(s_M,m_M)} \binom{s_M}{k}\binom{m_M}{k} k!\right]. \tag{4}$$

Unsurprisingly, n increases rapidly with the number of victims and missing persons, but depends strongly on the distribution of sexes. To illustrate, Table 1 tabulates the number of assignments with 8 victims and 5 MPs, for all combinations of males/females. The total of 19081 assignments when all victims and MPs have the same sex, is considerably higher than in all other cases.

	s_F					
m_F	0	1	2	3	4	5
0	19081	3393	529	73	9	1
1	9276	3922	1074	228	40	6
2	4051	3135	1603	559	147	31
3	1546	2004	1768	1054	438	136
4	501	1045	1533	1533	1045	501
5	136	438	1054	1768	2004	1546
6	31	147	559	1603	3135	4051
7	6	40	228	1074	3922	9276
8	1	9	73	529	3393	19081

Table 1. The number of sex-consistent assignments in a DVI case with 5 victims and 8 MPs. The variables s_F and m_F denote the number of female victims and MPs, respectively.

Sequential approaches

Here we describe two natural implementations of the PM-driven search strategy. As alluded to in the introduction this sequential approach is suboptimal in several ways, but it may be the best option in very large-scale applications. The motivation for including these algorithms here is to expose and clarify implementational details, and to serve as reference for the novel methods described later.

Algorithm 1: Sequential (without updates)

Input: A DVI problem; a threshold $T > 1$.

Output: A proposed solution to the DVI case in the form of an assignment a. (In case of ties, more than one assignment may result.)

Procedure:

(i) Compute the pairwise LR matrix B.

(ii) If all elements of B are below T, then stop. Otherwise, let $\mathrm{LR}_{i,j}$ be the maximal element of B and store the identification $\mathrm{V}_i = \mathrm{M}_j$. If there are multiple maximal elements, branch off and proceed with one at a time.

(iii) Update B by deleting the row and column corresponding to $\mathrm{LR}_{i,j}$.

(iv) Repeat steps (ii) - (iii) until the procedure stops.

To illustrate Algorithm 1, consider our running example from Figure 1, for which the pairwise LR matrix was given in (2). If $T > 5$, no identifications are made. For any $T \leq 5$, the above algorithm identifies $\mathrm{V}_2 = \mathrm{M}_2$ and $\mathrm{V}_3 = \mathrm{M}_3$ (both with $\mathrm{LR} = 5$), after which the procedure stops. Hence the reported solution is $(*, \mathrm{M}_2, \mathrm{M}_3)$. As remarked earlier, M_1 cannot be identified with this approach.

The next algorithm is a refinement of Algorithm 1, with the crucial difference that the LR matrix is now recomputed in each step.

Algorithm 2: Sequential (with updates)

Input: A DVI problem; a threshold $T > 1$.

Output: A proposed solution to the DVI problem in the form of an assignment a. (In case of ties, more than one assignment may result.)

Procedure: As Algorithm 1, but where step (iii) is replaced with the following:

(iii) Update B by deleting the row and column corresponding to $\mathrm{LR}_{i,j}$, and recomputing the remaining LR values conditional on all previous identifications.

	V_1	V_2	V_3	loglik	LR	posterior
1	M_1	M_2	M_3	-16.12	250.00	0.72
2	M_1	M_2	$*$	-17.73	50.00	0.14
3	$*$	M_2	M_3	-18.42	25.00	0.07
4	M_1	$*$	M_3	-20.03	5.00	0.01
5	$*$	M_1	M_3	-20.03	5.00	0.01
6	$*$	M_2	$*$	-20.03	5.00	0.01
7	$*$	$*$	M_3	-20.03	5.00	0.01
8	M_1	$*$	$*$	-21.64	1.00	0.00
9	$*$	M_1	$*$	-21.64	1.00	0.00
10	$*$	$*$	$*$	-21.64	1.00	0.00
11	M_2	M_1	M_3	-Inf	0.00	0.00
12	M_2	M_1	$*$	-Inf	0.00	0.00
13	M_2	$*$	M_3	-Inf	0.00	0.00
14	M_2	$*$	$*$	-Inf	0.00	0.00

Table 2. The 14 possible assignments for the DVI problem in Figure 1, ranked according to LR.

When this strategy is applied to the example in Figure 1, the sequence of updated LR matrices becomes as follows:

$$\begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} M_1 & M_2 & M_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & \mathbf{5} & 0 \\ 0 & 0 & \mathbf{5} \end{bmatrix} \end{array} \quad \begin{array}{l} \nearrow \\ \\ \searrow \end{array} \quad \begin{array}{c} \\ V_1 \\ V_3 \end{array}\!\!\begin{array}{c} \begin{array}{cc} M_1 & M_3 \end{array} \\ \begin{bmatrix} \mathbf{10} & 0 \\ 0 & 5 \end{bmatrix} \end{array} \longrightarrow \begin{array}{c} \\ V_3 \end{array}\!\!\begin{array}{c} \begin{array}{c} M_3 \end{array} \\ \begin{bmatrix} \mathbf{5} \end{bmatrix} \end{array} \quad \begin{array}{c} \\ V_1 \\ V_2 \end{array}\!\!\begin{array}{c} \begin{array}{cc} M_1 & M_2 \end{array} \\ \begin{bmatrix} 1 & 0 \\ 1 & \mathbf{5} \end{bmatrix} \end{array} \longrightarrow \begin{array}{c} \\ V_1 \end{array}\!\!\begin{array}{c} \begin{array}{c} M_1 \end{array} \\ \begin{bmatrix} \mathbf{10} \end{bmatrix} \end{array} \tag{5}$$

In both cases, the identified solution is (M_1, M_2, M_3).

The joint approach

We now consider the possibility (and feasibility) of *joint* identification of the victims. Among the list $\mathscr{A}$ of all *a priori* possible assignments, we seek the one that maximises the overall likelihood: An assignment a^* is an optimal solution if $L(a^*) \geq L(a)$ for all $a \in \mathscr{A}$. In smaller cases where $|\mathscr{A}|$ (as given by formula (4)) is manageable , this may be solved by brute force, i.e., by calculating the likelihood of each assignment, and sorting them in descending order.

Applying this to our running example in Figure 1, Table 2 lists the likelihoods of all 14 possibilities. It shows that assignment (M_1, M_2, M_3) is a clear winner, five times more likely than the runner-up. In this case all calculations may be done manually. For example, under the null model $(*, *, *)$ all five genotyped individuals are unrelated, giving the likelihood $L_0 = 4 \cdot (0.1)^{10}$ and (natural) log-likelihood $\log(L_0) = -21.64$ as shown in line 10 of Table 2.

Combined approach

In larger cases the number of possible assignments may be prohibitive for brute force calculations. In this case, we propose the combination approach described below. In brief, the idea is to first use a modified version of Algorithm 2 in order to find *undisputed* pairings, and then use brute force on the remaining problem. A pairing (V_i, M_j) is said to be undisputed if its pairwise-search LR reaches the given threshold T, while all other LR values involving V_i or M_j are small.

Algorithm 3: Undisputed + joint

Input: A DVI problem; a threshold $T > 1$.
Output: A list of assignments, ranked by likelihood.
Procedure:

Step 1: Sequential.

(i) Compute the pairwise LR matrix B.

(ii) Identify all undisputed pairings $\mathrm{V}_i = \mathrm{M}_j$, characterised by $\mathrm{LR}_{i,j} \geq T$ while all other entries in the same row and column are ≤ 1. If no such elements are found, the procedure stops.

(iii) Update B by deleting rows and columns corresponding to undisputed pairings and recomputing the remaining LRs conditional on the same.

(iv) Repeat steps (ii) - (iii) until the procedure stops.

Step 2: Joint.

(i) Create a list $\mathscr{A}$ of sex-consistent assignments involving the remaining individuals.

(ii) Remove from $\mathscr{A}$ all impossible assignments, i.e., corresponding to zeroes in the updated LR matrix B.

(iii) Compute the likelihood of the remaining assignments in $\mathscr{A}$, conditional on the undisputed findings, and rank the output.

The combined approach may still fail if there are too many assignments to consider in Step 2. A practical solution is then to reduce the threshold T so that the number of undisputed matches is increased, leaving fewer remaining individuals for the joint analysis.

Posterior pairing probabilities

In this section we derive the *posterior pairing probabilities* $q_{i,j} = P(\mathrm{V}_i = \mathrm{M}_j \mid D)$ for $i = 1\ldots,s$ and $j = 1,\ldots,m$, where D denotes the PM and AM data. For each victim V_i we also compute the *posterior non-pairing probability*, $q_{i,*} = P(\mathrm{V}_i = * \mid D)$, i.e., the probability that V_i does not match any of the missing persons.

These probabilities are relevant since decisions often need to be made for each individual independently. Importantly, this approach opens for incorporating non-DNA information via a prior distribution. For any assignment $a \in \mathscr{A}$, let $\pi(a)$ denote the prior probability of a.

For a given pair $(\mathrm{V}_i, \mathrm{M}_j)$, let $\mathscr{A}_{i,j}$ denote the subset of $\mathscr{A}$ consisting of all assignments containing the pairing $\mathrm{V}_i = \mathrm{M}_j$. Bayes' theorem then gives

$$q_{i,j} = P(\mathrm{V}_i = \mathrm{M}_j \mid D) = \frac{\sum_{a\in\mathscr{A}_{i,j}} L(a)\pi(a)}{\sum_{a\in\mathscr{A}} L(a)\pi(a)}, \tag{6}$$

where, as before, $L(a)$ is the likelihood of a. Often a flat prior $\pi(a) = 1/|\mathscr{A}|$ is used, in which case (6) can be written in terms of likelihood ratios:

$$q_{i,j} = \frac{\sum_{a\in\mathscr{A}_{i,j}} \mathrm{LR}_a}{\sum_{a\in\mathscr{A}} \mathrm{LR}_a}. \tag{7}$$

Here, LR_a denotes the likelihood ratio comparing a to the empty (null) assignment.

The posterior non-pairing probabilities are computed similarly: If $\mathscr{A}_{i,*}$ denotes the set of assignments with no match for V_i, we have

$$q_{i,*} = P(\mathrm{V}_i = * \mid D) = \frac{\sum_{a\in\mathscr{A}_{i,*}} L(a)\pi(a)}{\sum_{a\in\mathscr{A}} L(a)\pi(a)} = \frac{\sum_{a\in\mathscr{A}_{i,*}} \mathrm{LR}_a}{\sum_{a\in\mathscr{A}} \mathrm{LR}_a}, \tag{8}$$

where the latter equality assumes a flat prior.

For our running example in Figure 1, the posterior probabilities with a flat prior are given in Table 3. Note that these probabilities are directly calculable from the LR column of Table 2, which provides the likelihood ratios required by formulas (7) and (8). For example, the top left entry is

$$q_{1,1} = \frac{250+50+5+1}{250+50+25+5+5+5+5+1+1+1} \approx 0.88.$$

It is reasonable to conclude that $\mathrm{V}_i = \mathrm{M}_j$ if $q_{i,j} > \alpha$ for some α close to 1, say $\alpha = 0.99$. A less stringent threshold $\alpha = 0.5$ could be used if the objective is only to find the most likely match. Importantly, as long as $\alpha > 0.5$ any pairings obtained in this way are *consistent*, in the sense that two victims cannot be paired with the same missing person. To show this, let V_i and $\mathrm{V}_{i'}$ denote two different victims. Then for any j the sets $\mathscr{A}_{i,j}$ and $\mathscr{A}_{i',j}$ are disjoint, so that

$$q_{i,j} + q_{i',j} = P(\mathrm{V}_i = \mathrm{M}_j \mid D) + P(\mathrm{V}_{i'} = \mathrm{M}_j \mid D) = \sum_{a\in\mathscr{A}_{i,j}} P(a \mid D) + \sum_{a\in\mathscr{A}_{i',j}} P(a \mid D) \leq \sum_{a\in\mathscr{A}} P(a \mid D) = 1.$$

This implies that $q_{i,j}$ and $q_{i',j}$ cannot both exceed 0.5.

	M_1	M_2	M_3	$*$
V_1	0.88	0.00	0.00	0.12
V_2	0.02	0.95	0.00	0.03
V_3	0.00	0.00	0.83	0.17

Table 3. Posterior pairing probabilities for the toy example in Figure 1.

Results

A comparison of methods

The purpose of the example is to illustrate that the joint approach may succeed in cases where the sequential methods fail.

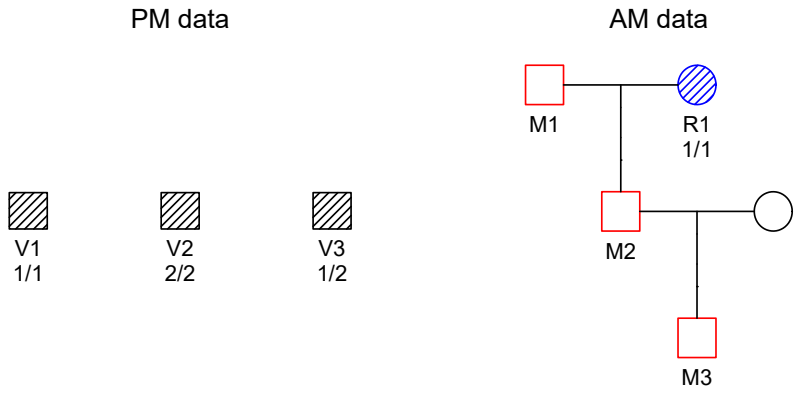

Figure 2. A simple case where sequential approaches fail.

Consider the DVI problem shown in Figure 2, where the genotypes correspond to a marker with alleles 1, 2, and 3, with frequencies 0.05, 0.05 and 0.9 respectively. (The precise values of these frequencies are not important.) Given this information, the pairwise LR matrix is found to be

$$B = \begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \end{array} \begin{array}{c} \begin{array}{ccc} M_1 & M_2 & M_3 \end{array} \\ \begin{bmatrix} 1 & 20 & 10.5 \\ 1 & 0 & 0.5 \\ 1 & 10 & 5.5 \end{bmatrix} \end{array}. \tag{9}$$

As a first observation we note that the column of 1's corresponding to M_1 implies that M_1 cannot identified by any method which uses data from only one victim at a time, such as Algorithm 1.

Next we consider the more reasonable Algorithm 2, which updates B after each new pairing. Clearly, since $LR_{1,2} = 20$ is the highest entry, the procedure starts by identifying $V_1 = M_2$. But this means that M_2 has genotype 1/1, which effectively blocks V_2 (who is 2/2) from being identified as M_1 or M_3. In full detail, the sequence of updated LR matrices becomes as follows:

$$\begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \end{array} \begin{array}{c} \begin{array}{ccc} M_1 & M_2 & M_3 \end{array} \\ \begin{bmatrix} 1 & \mathbf{20} & 10.5 \\ 1 & 0 & 0.5 \\ 1 & 10 & 5.5 \end{bmatrix} \end{array} \longrightarrow \begin{array}{c} \\ V_2 \\ V_3 \end{array} \begin{array}{c} \begin{array}{cc} M_1 & M_3 \end{array} \\ \begin{bmatrix} 0 & 0 \\ \mathbf{10} & \mathbf{10} \end{bmatrix} \end{array} \longrightarrow \begin{array}{c} \\ V_2 \end{array} \begin{array}{c} M_1 \\ \begin{bmatrix} 0 \end{bmatrix} \end{array} \text{ or } \begin{array}{c} \\ V_2 \end{array} \begin{array}{c} M_3 \\ \begin{bmatrix} 0 \end{bmatrix} \end{array}. \tag{10}$$

We conclude that Algorithm 2 produces two equally likely assignments, $(M_2, *, M_1)$ and $(M_2, *, M_3)$.

By contrast, Table 4 shows that the optimal solution, when all the data is considered jointly, is the assignment (M_3, M_1, M_2). In fact, this is $2,000/200 = 10$ times more likely than either of the solutions found by the sequential method above. Table 5 lists the posterior pairing probabilities under a flat prior.

In order to investigate the practical relevance of this effect, we conducted a series of simulation experiments based on standard forensic markers. After all, the genotypes in Figure 2 were particularly chosen so as to illustrate the effect, and with

	V_1	V_2	V_3	loglik	LR	posterior
1	M_3	M_1	M_2	-15.67	2,000.00	0.69
2	M_2	$*$	M_1	-17.97	200.00	0.07
3	M_2	$*$	M_3	-17.97	200.00	0.07
4	$*$	M_1	M_2	-17.97	200.00	0.07
5	M_3	$*$	M_2	-18.67	100.00	0.03
6	$*$	M_3	M_2	-18.67	100.00	0.03

Table 4. The most likely assignments in Figure 2.

	M_1	M_2	M_3	$*$
V_1	0.004	0.145	**0.736**	0.115
V_2	**0.766**	0.000	0.036	0.198
V_3	0.076	**0.831**	0.078	0.015

Table 5. Posterior pairing probabilities for the case in Figure 2. Values exceeding 0.5 are shown in bold.

multiple markers one might expect such anomalies to be drowned. Unfortunately, this is not the case. Figure 3 compares how the true positive rates (TPR) of Algorithms 1–3 vary with the number of markers, depending on the true solution. In each case, 500 sets of DNA profiles were simulated, using the database `NorwegianFrequencies` of 35 autosomal markers, available through the R package **forrel**. The LR threshold $T = 10,000$ was used for all three algorithms (for Algorithm 3, the threshold applied to the highest joint LR compared with the null). In addition, we included the TPR of the *most likely* solution reported by Algorithm 3, whether or not its LR exceeded T. For example, in the first panel the genotypes were simulated under the assumption that $V_1 = M_1$, $V_2 = M_2$ and $V_3 = M_3$. We see that the joint method (Algorithm 3) has a TPR near 1 already with 5 markers, while the best sequential (Algorithm 2) needs 20 markers to reach the same. Overall, Algorithm 3 clearly outperforms the others in all cases shown in the top row of Figure 3. Moreover, it is the only method to reliably reach a conclusion when the true assignment is $(M_1, *, M_3)$.

Example 1: Plane crash

Figure 4 shows a DVI problem based on a simulated plane crash. DNA profiles using 15 markers are available from 8 victims and 5 reference families. From the pairwise LR matrix in Table 6 we conclude that the identifications $V_2 = M_3$, $V_4 = M_5$ and $V_6 = M_2$ are undisputed, as defined in Algorithm 3, when applying the threshold $T = 10,000$. In addition, $V_1 = M_1$ also has a high LR, but does not reach the threshold. Based on these observations we anticipate the solution $(M_1, M_3, *, M_5, *, M_2, *, *)$. Indeed, this assignment is uniquely optimal, as shown in Table 7. It is noteworthy that that the many impossible or undisputed pairings in the pairwise LR matrix (Table 6) leave only two assignments for consideration in the joint step.

Next we introduce a mutational model, motivated by the possibility that a mutation may explain the lack of identification for M_4. In fact, an examination of the data reveals that V_3 and R_4 share alleles at all but one marker, suggesting that these may have a parent-child relationship.

We use a proportional model with mutation rate 0.001[16], mainly since it is stationary and hence calculations can be checked

	M_1	M_2	M_3	M_4	M_5
V_1	9.29e+02	9.03e-04			2.77e-01
V_2		6.75e-02	**6.79e+04**		6.66e-02
V_3		1.03e-04			3.82e-03
V_4		3.78e-05			**3.19e+07**
V_5		9.62e-04			3.92e-03
V_6		**1.08e+06**			1.29e-05
V_7		5.90e-04			1.90e-01
V_8		1.91e-04			2.72e-01

Table 6. Pairwise LRs for the plane crash example. Only nonzero elements are shown; entries reaching the threshold $T = 10,000$ are highlighted.

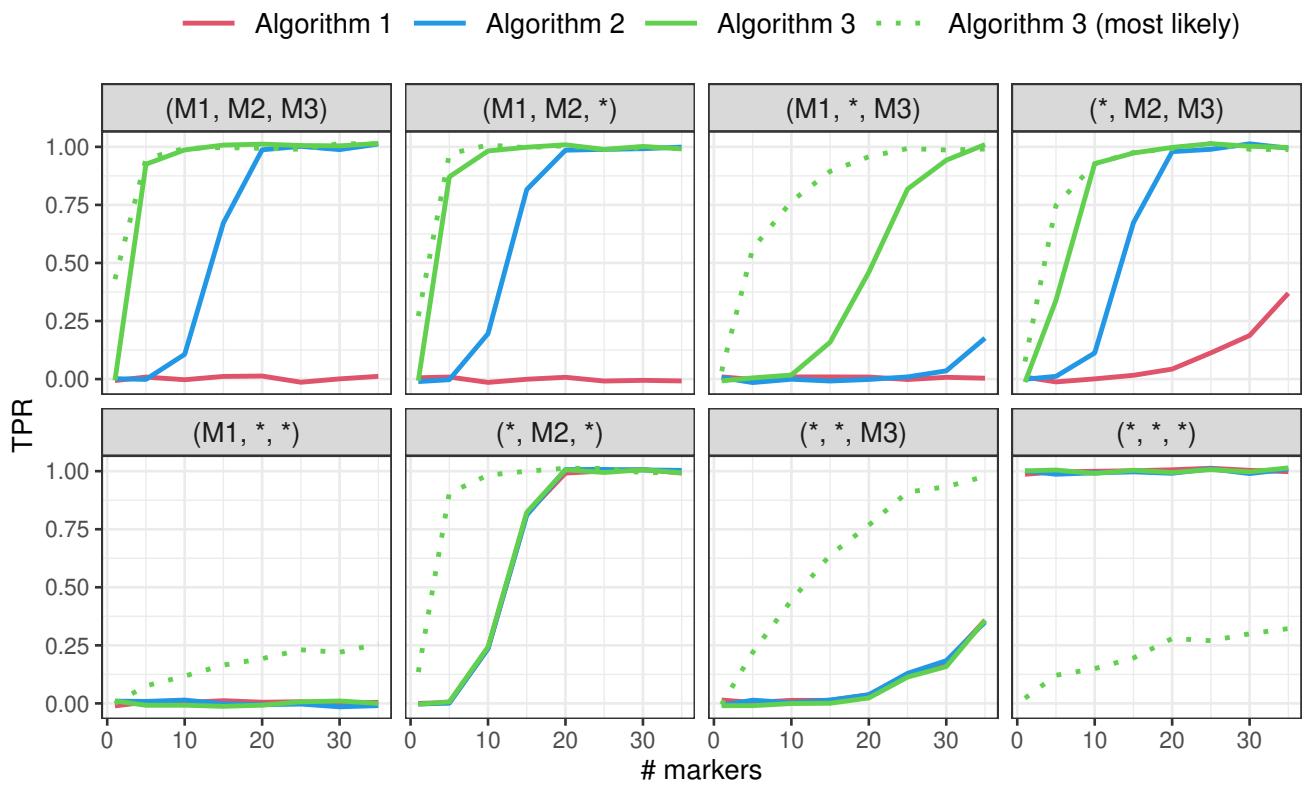

Figure 3. A comparison of the true positive rates (TPR) of different DVI algorithms. Each point is the result of 500 simulations of the AM and PM data conditional on the assignment indicated in the panel title. The threshold $T = 10,000$ was used throughout. A slight vertical jitter is applied to the points in order to increase visibility.

against other implementations. Under this model, the pairwise LR for $V_3 = M_4$ changes from 0 to 249. A joint analysis produces the top list shown in Table 8. We see that the identification $V_3 = M_4$ is now convincingly included in the most likely assignment. This observation is reinforced by the posterior pairing probabilities given in Table 9, calculated with a flat prior of $\pi(a) = 1/19081$.

Example 2: A large reference family

The pedigree in Figure 5 is based on an example from a workshop organized by the International Society for Forensic Genetics (ISFG)[17], featuring 5 victim samples to be matched against a large reference family with 12 missing individuals. For notational consistency we have renamed the individuals. Genetic marker data was simulated using 13 CODIS markers, assuming that the true solution is the assignment $(M_6, M_{10}, M_{12}, M_8, M_1)$.

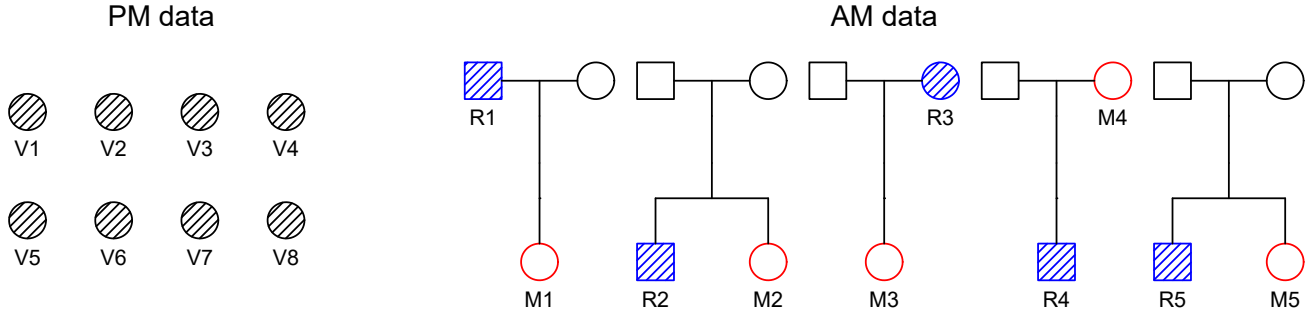

Figure 4. The plane crash example. Eight victims are to be matched against five reference families.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	loglik	LR	posterior
1	M_1	M_3	$*$	M_5	$*$	M_2	$*$	$*$	-562.80	2.17e+21	0.999
2	$*$	M_3	$*$	M_5	$*$	M_2	$*$	$*$	-569.64	2.34e+18	0.001

Table 7. Results of joint analysis of the plane crash example, without mutation modelling.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	loglik	LR	posterior
1	M_1	M_3	M_4	M_5	$*$	M_2	$*$	$*$	-557.31	5.27e+23	0.995
2	M_1	M_3	$*$	M_5	$*$	M_2	$*$	$*$	-562.83	2.11e+21	0.004
3	$*$	M_3	M_4	M_5	$*$	M_2	$*$	$*$	-564.14	5.69e+20	0.001
4	$*$	M_3	M_4	M_5	M_1	M_2	$*$	$*$	-566.54	5.18e+19	0.000
5	M_1	$*$	M_4	M_5	$*$	M_2	M_3	$*$	-566.82	3.89e+19	0.000

Table 8. The most likely assignments in the plane crash example, when mutations are modeled.

The pairwise LR matrix in Table 10 shows that there are no undisputed pairings (with $T = 10,000$). Admittedly, the pairing $V_4 = M_8$ has a high LR at 2.85e6, but it is disputed (in the sense of Step 1 - (ii) of Algorithm 3) by $LR_{4,10}$ and $LR_{4,11}$ which both exceed 1. (Relaxations of this step are considered in the Discussion.) Nevertheless, the many zeroes in Table 10 lead to a substantial reduction in the space of assignments, manifested in Step 2 - (ii) of Algorithm 3. More precisely, the *a priori* 9847 assignments given by equation (4) (with $s_F = 3, s_M = 2, m_F = m_M = 6$) is reduced to 1898 after removal of impossible pairings. Joint analysis of these assignments took $\sim$15 seconds on a standard laptop, and resulted in the top list presented in Table 11.

Note that two assignments tie for the best solutions, differing in their identification of victim V_2. This reflects the fact, deducible from Figure 5, that the pairings $\{V_2 = M_{10}\}$ and $\{V_2 = M_{11}\}$ cannot be distinguished based on DNA data. The posterior pairing probabilities with a flat prior are given in Table 12.

Discussion

The main contribution of this paper is to show that joint identification generally outperforms sequential DVI methods, and that careful implementation makes the joint approach computationally feasible even in fairly large cases.

From a computational point of view, DVI applications typically consist of a large number of kinship tests. For general issues concerning kinship testing, like the assumptions of Hardy-Weinberg equilibrium and independence of markers, we therefore refer to the rich literature on this subject[18]. Problems related to poor quality of DNA are also widely discussed in the forensic literature[19]. In the following we restrict the discussion to aspects that are particular to the methods of this paper.

For simplicity we used autosomal markers in our examples, but there are no methodological obstructions to including mtDNA, X or Y markers in joint DVI computations. In fact, our implementation in the **dvir** package already supports X-chromosomal markers. As previous authors have noted, it is not obvious how evidence from different types of markers should be reported, and opinions differ[20].

A well-known challenge in forensic genetics is that LR calculations are sensitive to misspecified allele frequencies. In some DVI cases it may therefore be difficult to decide on an appropriate frequency database, particularly if the individuals

	M_1	M_2	M_3	M_4	M_5	$*$
V_1	0.999					0.001
V_2			1.000			
V_3				0.996		0.004
V_4					1.000	
V_5						1.000
V_6		1.000				
V_7						1.000
V_8						1.000

Table 9. Posterior pairing probabilities in the plane crash example calculated using a flat prior and a proportional mutational model with rate 0.001.

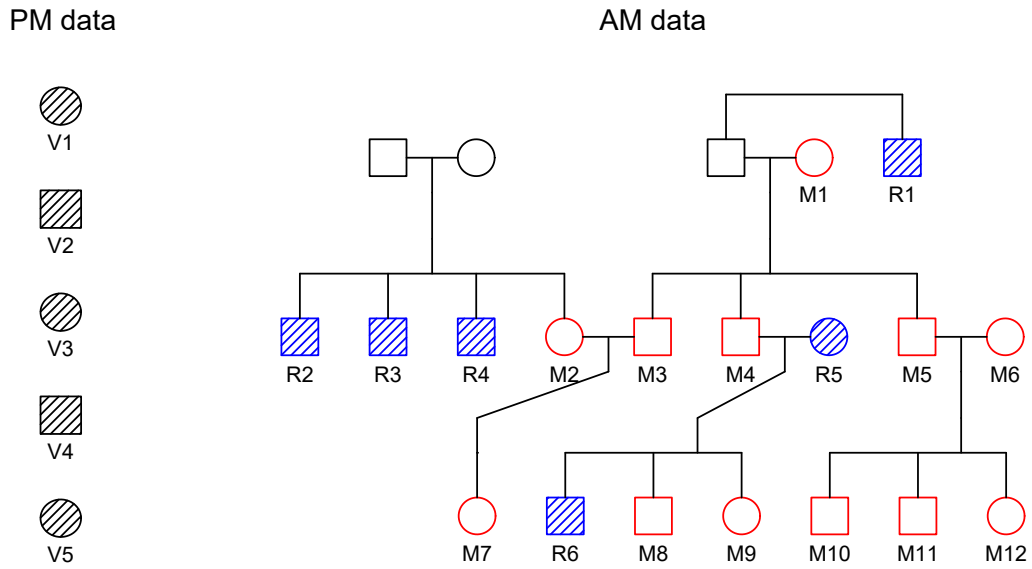

Figure 5. A large reference family with 12 missing individuals.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}
V_1	0.0647					1						0.164
V_2			0.152		0.152					1.25	1.25	
V_3	0.124					1	0.0156					2.27
V_4			0.792		0.792			2.85e+06		3.08	3.08	
V_5	17.5					1	0.00666					2.83

Table 10. Pairwise LRs in Example 2. Values below 0.001 are not shown.

originate from different populations. This problem has previously been addressed in the context of familial searching[21]. A practical approach is to do *ad hoc* sensitivity calculations. If the overall conclusions remain unchanged with different databases, this strengthens the confidence in the results. As a general comment we note that most autosomal forensic markers have been specifically selected for their relatively stable allele frequencies across populations, while this to a lesser extent holds for mtDNA, X or Y markers.

The algorithms we have presented can be modified or tuned in various ways. As in many similar applications, the most important parameter is arguably the LR threshold T. For a general discussion we refer to the established literature[22], also in connection with familial searching[23]. In the context of DVI we expect the value of T to depend both on the particular protocol and external factors. Simulation experiments like the one summarised in Figure 3 may provide guidance when deciding the threshold.

We mention one potential modification, which may have a significant impact on the run-time of Algorithm 3. Recall that Step 1- (ii) of this algorithm used the pairwise LR matrix to identify undisputed pairings $V_i = M_j$, characterised by

$LR_{i,j} \geq T$ while all other entries in the same row and column are ≤ 1.

	V_1	V_2	V_3	V_4	V_5	loglik	LR	posterior
1	M_6	M_{10}	M_{12}	M_8	M_1	-312.98	1.14E+24	0.50
2	M_6	M_{11}	M_{12}	M_8	M_1	-312.98	1.14E+24	0.50
3	M_6	M_{10}	M_{12}	M_8	M_7	-327.16	7.86E+17	0.00
4	M_6	M_{11}	M_{12}	M_8	M_7	-327.16	7.86E+17	0.00
5	M_6	*	M_{12}	M_8	M_1	-327.74	4.40E+17	0.00

Table 11. The five most likely assignments for the case in Figure 5.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}	*
V_1						1.000							
V_2										0.500	0.500		
V_3												1.000	
V_4								1.000					
V_5	1.000												

Table 12. Posterior pairing probabilities in the ICMP example. Numbers less than 0.001 are not shown.

The last part of this criterion may be relaxed, for instance by increasing the final limit 1 to $LR_{i,j}/T$. The effect of this change is easily seen in Example 2, specifically Table 10, where the pairing $V_4 = M_6$ would now be classified as undisputed.

The question of how the statistical evidence should be reported in identification cases is difficult and lacks general consensus. Although there is a tradition of specifying priors and reporting posterior probabilities in addition to likelihood ratios[10], our view is that specifying priors should be left to the decision makers. This is supported by ISFG recommendations[6], whose point 11 includes:

> In DVI work, DNA statistics are best represented as likelihood ratios that permit DNA results to be combined among multiple genetic systems or with other non-DNA evidence.

Nevertheless, we have included in several tables the posteriors with a flat prior, for reference. In real cases, information beyond the DNA data can be reflected by the prior.

Another point related to reporting is the choice of *reference*, i.e., the hypothesis in the denominator of the likelihood ratio. In our examples we have chosen to compare with the null, i.e., no relations between the victims and the missing persons, but it is not obvious that this is always the best choice. For instance, in order to communicate the uniqueness of the solution, a viable alternative is to compare the best solution to the second best. We suggest reporting the identification defined by the optimal assignment if its LR compared to closest contender exceeds a threshold, say 10,000. Clearly, this ensures that the LR against the null also exceeds the same threshold.

Conclusion

This paper presents and discusses methods for DNA-based identification. Restricted approaches, in which the victims are considered separately or sequentially, may give inconsistent, ambiguous results. We therefore generally recommend the combined approach summarised by Algorithm 3. The idea is simple: first take care of the virtually obvious pairings, and then do a complete search to resolve the remaining. The resulting joint solution should be supplemented by posterior pairing and non-pairing probabilities, which summarise the evidence for each individual identification.

All methods described in this paper are implemented in the R package **dvir**, which is freely available from the official R repository (CRAN) and runs on all platforms. The documentation of the package provides further details and instructive examples.

Author contributions statement

T.E. conceived the project. Both authors wrote and reviewed the manuscript. M.D.V prepared the figures.

Additional information

Competing interests statement

The authors declare no competing interests.

References

1. Parsons, T. J., Huel, R. M., Bajunović, Z. & Rizvić, A. Large scale DNA identification: The ICMP experience. *Forensic Sci. Int. Genet.* **38**, 236–244 (2019).
2. Bertoglio, B. *et al.* Disaster victim identification by kinship analysis: the Lampedusa October 3rd, 2013 shipwreck. *Forensic Sci. Int. Genet.* **44**, 102156 (2020).
3. Brenner, C. H. & Weir, B. S. Issues and strategies in the DNA identification of World Trade Center victims. *Theor. population biology* **63**, 173–178 (2003).

4. Brenner, C. H. Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities. *Forensic Sci. Int.* **157**, 172–180 (2006).
5. Vigeland, M. D., Marsico, F. L., Piñero, M. H. & Egeland, T. Prioritising family members for genotyping in missing person cases: A general approach combining the statistical power of exclusion and inclusion. *Forensic Sci. Int. Genet.* **49**, 102376 (2020).
6. Prinz, M. *et al.* DNA Commission of the International Society for Forensic Genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI). *Forensic Sci. Int. Genet.* **1**, 3–12 (2007).
7. OC, D. S. INTERPOL DVI best-practice standards—An overview. *Forensic Sci. Int.* **201**, 18–21 (2010).
8. Parsons, T. J. & Huel, R. L. DNA and missing persons identification: practice, progress and perspectives. In *Handbook of Forensic Genetics. Biodiversity and Heredity in Civil and Criminal Investigation* (World Scientific, New Jersey, 2016).
9. Kling, D., Egeland, T., Tillmar, A. & Prieto, L. *Mass identifications. Statistical methods in forensic genetics* (Elsevier. Academic press, 2021).
10. Vullo, C. M. *et al.* GHEP-ISFG collaborative simulated exercise for DVI/MPI: Lessons learned about large-scale profile database comparisons. *Forensic Sci. Int. Genet.* **21**, 45–53 (2016).
11. Kling, D., Tillmar, A. O. & Egeland, T. Familias 3–extensions and new functionality. *Forensic Sci. Int. Genet.* **13**, 121–127 (2014).
12. van Dongen, C., Slooten, K., Slagter, M., Burgers, W. & Wiegerinck, W. Bonaparte: Application of new software for missing persons program. *Forensic Sci. Int. Genet. Suppl. Ser.* **3**, e119–e120 (2011).
13. Slooten, K. Validation of DNA-based identification software by computation of pedigree likelihood ratios. *Forensic Sci. Int. Genet.* **5**, 308–315 (2011).
14. Vigeland, M. D. *Pedigree analysis in R* (Academic Press, 2021).
15. Egeland, T., Kling, D. & Mostad, P. *Relationship inference with familias and R: Statistical methods in forensic genetics* (Academic Press, 2015).
16. Egeland, T., Mostad, P. F., Mevåg, B. & Stenersen, M. Beyond traditional paternity and identification cases. Selecting the most probable pedigree. *Forensic Sci. Int.* **110**, 47–59 (2000).
17. ISFG Workshop, Seoul, South Korea, 2017. http://www.few.vu.nl/~ksn560/Block-III-PartI-KS-ISFG2017.pdf. Accessed: 2021-02-10.
18. Balding, D. J. & Steele, C. D. *Weight-of-evidence for Forensic DNA Profiles* (John Wiley & Sons, 2015).
19. Tvedebrink, T., Eriksen, P. S., Mogensen, H. S. & Morling, N. Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out. *Forensic Sci. Int. Genet.* **6**, 97–101 (2012).
20. Amorim, A. A cautionary note on the evaluation of genetic evidence from uniparentally transmitted markers. *Forensic Sci. Int. Genet.* **2**, 376–378 (2008).
21. Fortier, A. L., Kim, J. & Rosenberg, N. A. Human-Genetic Ancestry Inference and False Positives in Forensic Familial Searching. *G3: Genes, Genomes, Genet.* **10**, 2893–2902 (2020).
22. J, B., CM, T. & SJ, W. (eds.) *Forensic DNA Evidence Interpretation* (CRC Press, Florida, USA, 2005).
23. Slooten, K. & Meester, R. Probabilistic strategies for familial DNA searching. *J. Royal Stat. Soc. Ser. C: Appl. Stat.* 361–384 (2014).

Figures

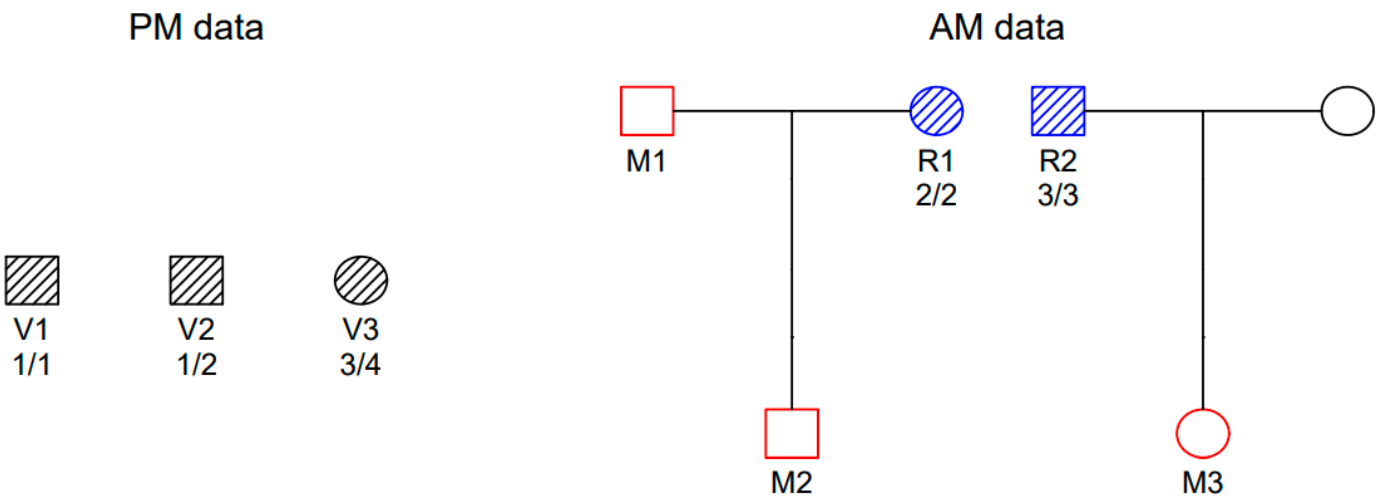

Figure 1

A toy DVI problem. The PM data consists of 3 victim samples to be matched against 3 missing persons (red) belonging to two different families. The AM data contains profiles from the reference individuals R1 and R2 (blue), one from each family. The hatched individuals are typed with a single marker.

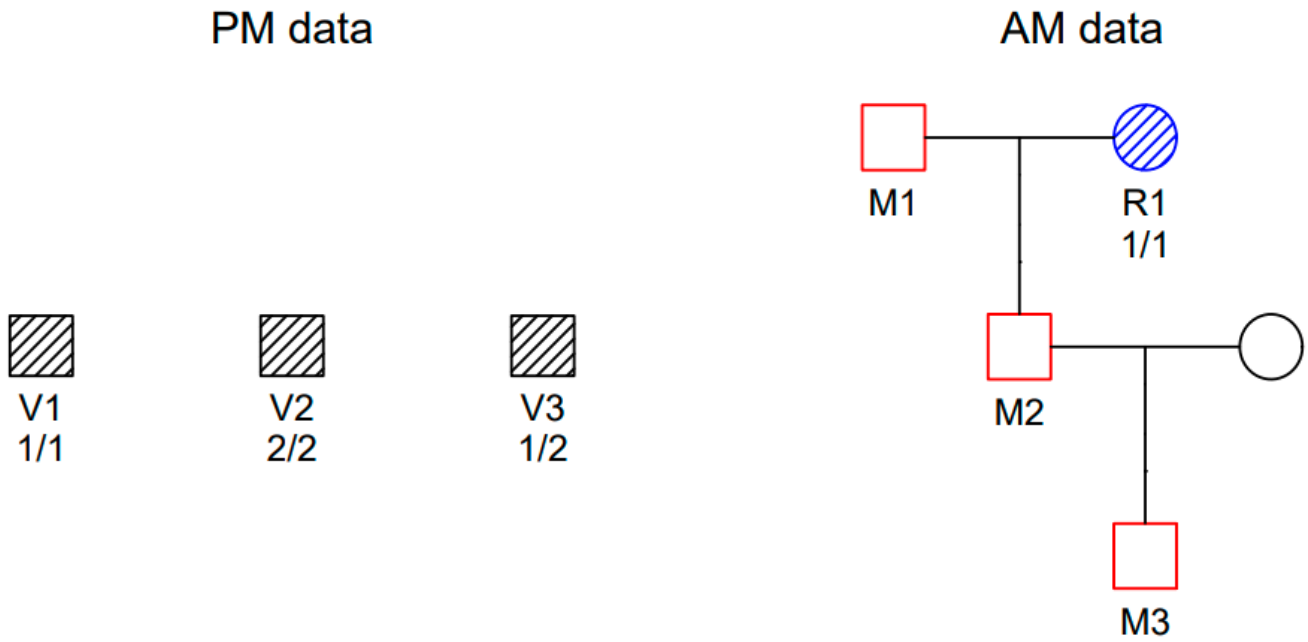

Figure 2

A simple case where sequential approaches fail.

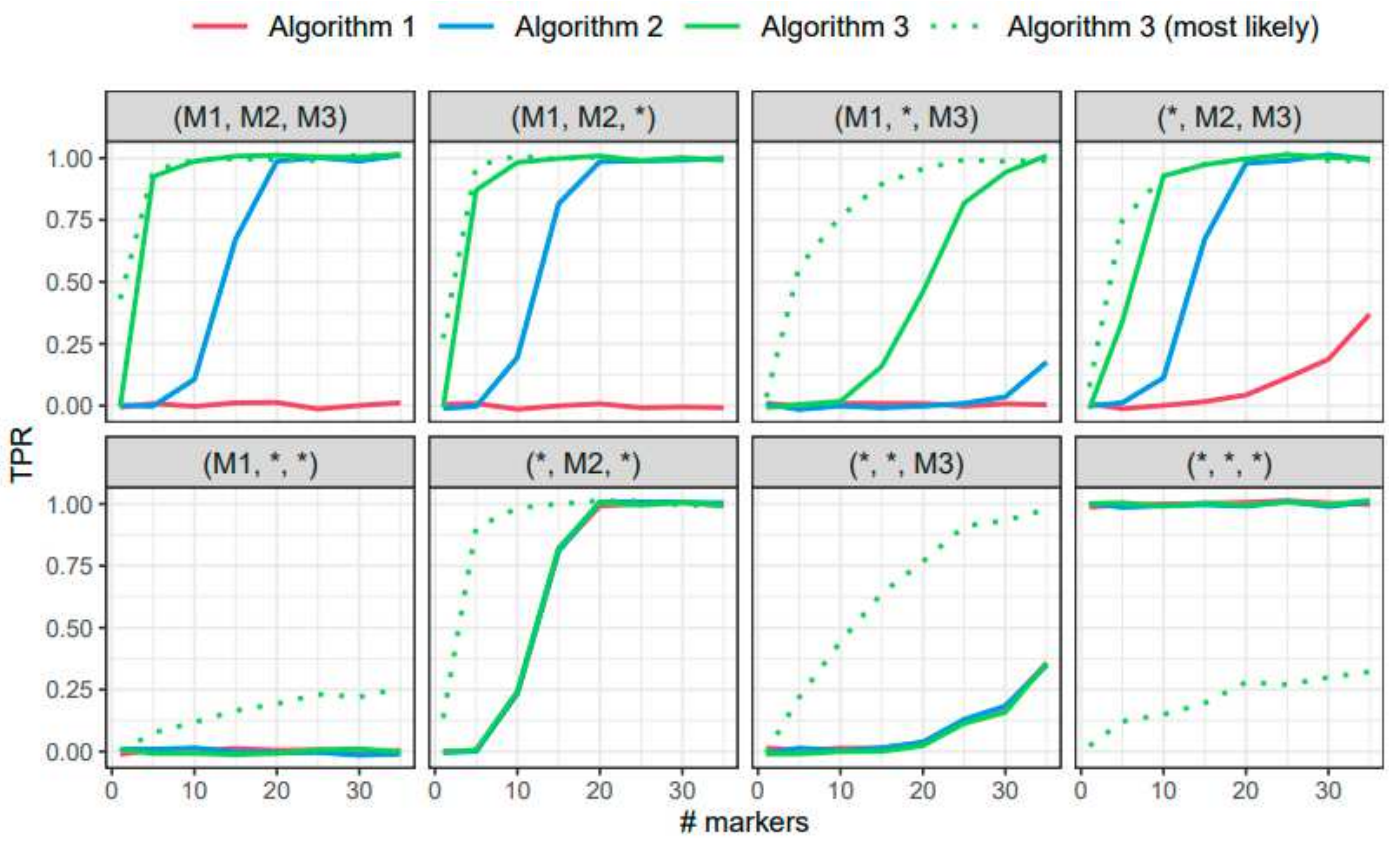

Figure 3

A comparison of the true positive rates (TPR) of different DVI algorithms. Each point is the result of 500 simulations of the AM and PM data conditional on the assignment indicated in the panel title. The threshold T = 10,000 was used throughout. A slight vertical jitter is applied to the points in order to increase visibility.

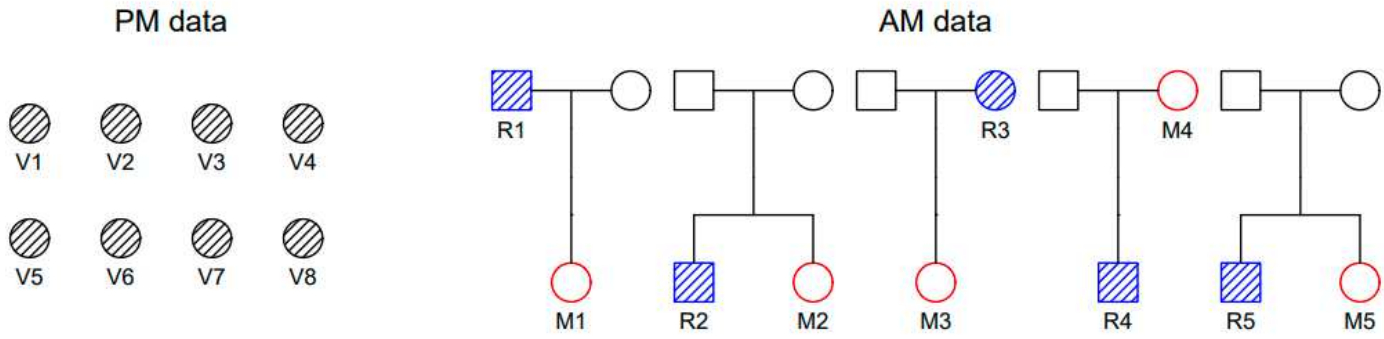

Figure 4

The plane crash example. Eight victims are to be matched against five reference families.

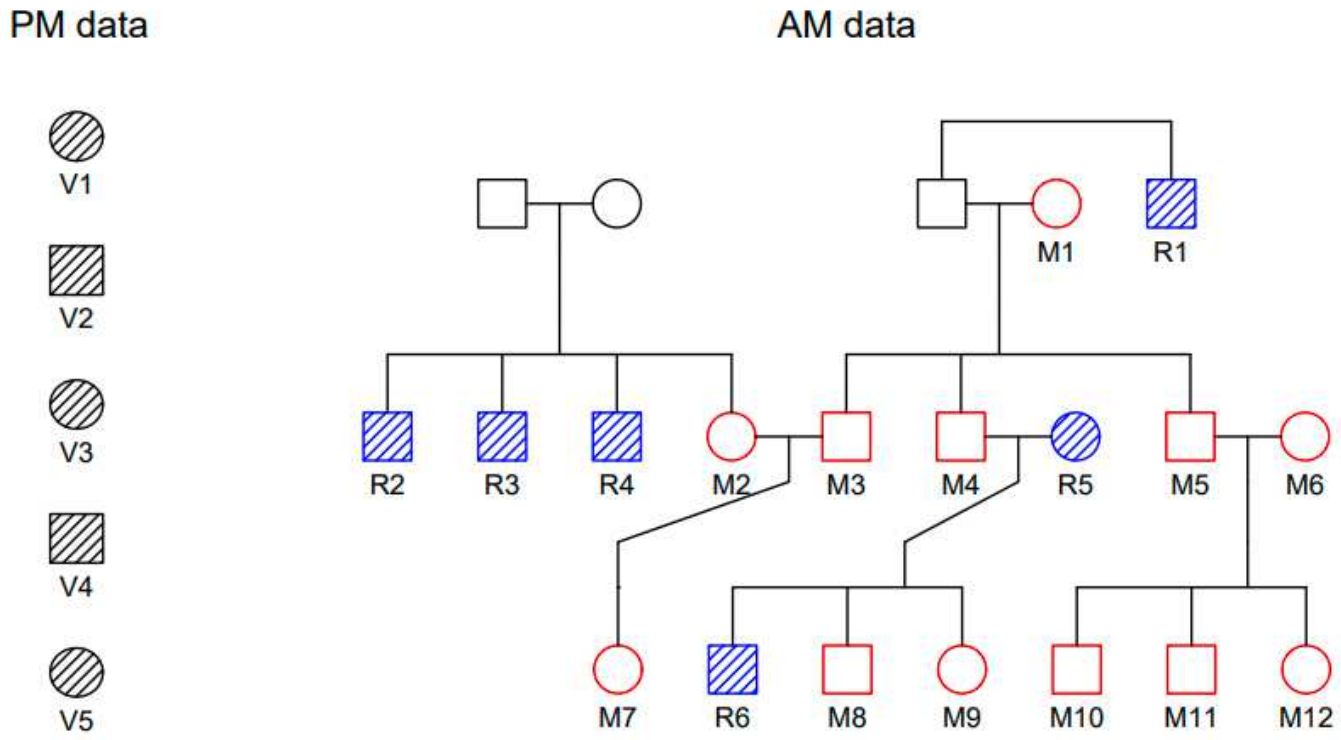

Figure 5

A large reference family with 12 missing individuals.