Forensic Science International: Genetics 13 (2014) 121–127

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Familias 3 – Extensions and new functionality

Daniel Kling [a,b,*], Andreas O. Tillmar [c,d], Thore Egeland [b,e]

[a] Department of Family Genetics, Norwegian Institute of Public Health, Oslo, Norway
[b] Department for Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Aas, Norway
[c] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden
[d] Department of Clinical and Experimental Medicine, Faculty of Health Sciences, Linköping University, Linköping, Sweden
[e] Department of Forensic Genetics, Norwegian Institute of Public Health, Oslo, Norway

ARTICLE INFO

Article history:
Received 14 March 2014
Received in revised form 23 June 2014
Accepted 1 July 2014

Keywords:
Familias
Paternity
Likelihood computations
Simulation
Mutation
Disaster victim identification

ABSTRACT

In relationship testing the aim is to determine the most probable pedigree structure given genetic marker data for a set of persons. *Disaster Victim Identification* (DVI) based on DNA data from presumed relatives of the missing persons can be considered to be a collection of relationship problems. Forensic calculations in investigative mode address questions like "How many markers and reference persons are needed?" Such questions can be answered by *simulations*. Mutations, deviations from Hardy–Weinberg Equilibrium (or more generally, accounting for population substructure) and silent alleles cannot be ignored when evaluating forensic evidence in case work. With the advent of new markers, so called microvariants have become more common. Previous mutation models are no longer appropriate and a new model is proposed. This paper describes methods designed to deal with *DVI* problems and a *new simulation model* to study distribution of likelihoods. There are softwares available, addressing similar problems. However, for some problems including DVI, we are not aware of freely available validated software. The Familias software has long been widely used by forensic laboratories worldwide to compute likelihoods in relationship scenarios, though previous versions have lacked desired functionality, such as the above mentioned. The extensions as well as some other novel features have been implemented in the new version, freely available at www.familias.no. The implementation and validation are briefly mentioned leaving complete details to Supplementary sections.

1. Introduction

There are several applications that require determination of genetic relatedness. The focus of this paper is to describe methods and implementations for complex relationships problems and disaster victim identification (DVI). While we have forensic applications in mind similar problems occur in a wide range of areas. The core computational problem is to calculate the likelihood of the data given competing hypotheses and from this to form the *likelihood ratio* (LR). We may further use a Bayesian approach with prior information to compute the posterior probabilities. In this paper we restrict attention to unlinked STR markers and then likelihoods are typically calculated using extensions of the Elston–Stewart (ES) algorithm [1] accommodating correction for population substructure (theta-correction), mutations and silent alleles [2]. The algorithm is in concept a peeling algorithm, where we consider subsets of a pedigree as conditionally independent given the connecting node. As a consequence, the algorithm may require long computation time, when marriage and inbreeding loops are present [3]. An implementation of the ES algorithm is provided in the software Familias [4]. The program is used by a large number of laboratories worldwide [5] when calculating likelihoods in relationship scenarios. Though previous versions of the software have included several important features, such as null/silent alleles, advanced mutation models and subpopulation correction, Familias has also lacked some desired functionality [6]. With the advent of new STR markers, micro-variant alleles (i.e., 9.3) have become more common necessitating an appropriate mutation model to handle transitions to and from such alleles. Whereas previous models are generally not designed to handle these transitions, this paper presents a new model, providing an extension of the stepwise mutation model [7,8], thereby accommodating for microvariants.

* Corresponding author at: Norwegian Institute of Public Health Department of Family Genetics, Gaustadalléen 30, NO-0027 Oslo, Norway. Tel.: +47 210 77663.
E-mail addresses: daniel.kling@fhi.no (D. Kling), andreas.tillmar@rmv.se (A.O. Tillmar), thore.egeland@nmbu.no (T. Egeland).

http://dx.doi.org/10.1016/j.fsigen.2014.07.004
1872-4973/

Monte Carlo simulation is a generic approach of relevance to virtually all areas of science. In our context, simulations can be used to get an idea of what evidential strength we will achieve for a given case. Based on simulations, one may for instance conclude that it is not worthwhile to proceed with a case unless more reference persons are genotyped or the number of genetic markers is increased. Simulation also extends the results from a point estimate of the LR to a complete description of its probability distribution. The model used for simulation is the same as the one used for likelihood and LR calculation. In other words, the simulations reflect the chosen mutation model, silent alleles and incorporate theta correction.

Disaster victim identification (DVI) applications can be considered as a potentially large collection of relationship estimation problems. Typically, LR ratios (sometimes converted to posterior probabilities) are reported and the aim is to compare large amounts of reference data, i.e., family members or personal belongings of missing persons, with unidentified remains. The underlying core computational model remains the same as in *standard* likelihood calculations. Since the early report on the successful use of DNA as a tool to identify victims of a mass disaster by Olaisen et al. [9], numerous papers have been published demonstrating its application and utility [10–15]. For the scope of this paper we consider smaller to medium sized DVI situation where the number of missing persons is typically limited to 1000.

As previously stated, the emphasis of this paper is on the new methods. Details on implementation and validation of the new software Familias 3 [hereafter called only Familias], which extends on Familias 2.0 [4], appear as supplementary material and in the manual. Some of the functionality of the new version or similar features can be found in other software [6,16,17]. However, (i) Familias is validated Drabek [6], (ii) widely used [5] (iii) freely available and (iv) the basic code is open (see http://familias.name/OpenFamilias). In addition, the implementation benefits from integrating similar problems (LR calculations, simulations and DVI feature) into one user friendly environment.

2. Methods

In relationship testing, mutually exclusive hypotheses are normally formulated. A hypothesis H corresponds to a *pedigree*, where the latter connects two or more individuals in a relationship tree. The core problem is to calculate the $P(data|H,\phi)$ where the *data* consists of alleles for different genetic markers and ϕ represents parameters needed to model e.g. mutations and subpopulation structure. The computation of the likelihood is in this paper based on the Elston–Stewart algorithm [1] and later extensions described in [18]. Briefly the algorithm peels the pedigree by calculating conditional probabilities for *cutsets*, where each *cutset* is conditionally independent given the rest of the pedigree, and can thus be effectively used on large pedigrees. The algorithm can also effectively accommodate many unlinked markers. Should we need to account for dependency between markers, other algorithms and implementations must be considered, e.g. FamLink[19] or Merlin [20]. For two different hypotheses H_1 and H_2, the likelihood ratio $\mathrm{LR} = P(data|H_1,\phi)/P(data|H_2,\phi)$ is typically calculated and reported.

The next section first describes the new mutation model, then the simulation approach and a framework to deal with DVI problems. Thereafter, some general principles related to validation are described. Finally, the implementation is briefly described deferring more complete descriptions to supplementary sections and the manual.

2.1. Mutation model

As mentioned, there is a need for a new mutation model capable of handling transitions to and from microvariants, e.g. between 9 and 9.3. Some current models treat such *microvariant mutations* (MVM) in the same way as *integer mutations* (IM) or neglect them as the mentioned transitions are considered improbable. This is biologically unreasonable and the problem has become more pronounced as MVM are more common in the latest STR kits. We provide a new stepwise mutation model accounting for MVM. The model is called the extended step wise model in the implementation.

We specify the model by letting M be the mutation matrix, with elements m_{ij}, where $i,j = 1, \ldots, N$ and where N is the number of alleles. Each element m_{ij} is the probability of a transition from allele A_i to allele A_j. The current model separates the overall mutation rate, denoted μ, into two parts, one corresponding to integer mutations, R, and one to the micro-variants α, i.e., $\mu = R + \alpha$. Biologically R is often explained by slippage error during DNA replication [8] while α is connected to insertions/deletions and point mutations. The last parameter, the mutation range r, is defined as for previous IM models; it is the value with which the probability decreases for each further step away from the original allele mutates.

Next the model is specified precisely by the transition probabilities m_{ij}. There are three different alternatives:

1. $m_{ij} = (1 - (R + \alpha))$ if $i = j$, i.e. the probability that an allele does not mutate.
2. $m_{ij} = k_i(1-\alpha)r^{|i-j|}$ for integer mutations.
3. $m_{ij} = k_i\alpha/N_i$ for micro variant mutations. N_i is the number of MVM-s from allele i. The rows must sum to unity and therefore the normalizing constants k_i are determined by the constraints $\sum_{j=1}^{N} m_{ij} = 1$.

Example 1. Consider a marker containing the alleles 9, 9.3, 10, 10.3 and 15. The transition matrix M is then given by:

$$M = \begin{bmatrix} 1-(R+\alpha) & k_1\alpha/2 & (1-\alpha)k_1r^1 & k_1\alpha/2 & (1-\alpha)k_1r^6 \\ k_2\alpha/3 & 1-(R+\alpha) & k_2\alpha/3 & (1-\alpha)k_2r^1 & k_2\alpha/3 \\ (1-\alpha)k_3r^1 & k_3\alpha/2 & 1-(R+\alpha) & k_3\alpha/2 & (1-\alpha)k_1r^5 \\ k_4\alpha/3 & (1-\alpha)k_4r^1 & k_4\alpha/3 & 1-(R+\alpha) & k_4\alpha/3 \\ (1-\alpha)k_5r^6 & k_5\alpha/2 & (1-\alpha)k_5r^5 & k_5\alpha/2 & 1-(R+\alpha) \end{bmatrix}$$

In this case, k_1 is found as follows $1 = 1-(R+\alpha) + k_1\alpha/2 + (1-\alpha)k_1r + k_1\alpha/2 + (1-\alpha)k_1r^6 \Leftrightarrow k_1 = (R+\alpha)/(\alpha + (1-\alpha)(r + r^6))$. Similar calculation can be shown for the other k_i. Note that, the matrix M is not necessarily symmetric, meaning that the probability of observing a mutation from 9 to 9.3 is not the same as observing a mutation from 9.3 to 9. This is a consequence of the definition of M. Further note that for transitions from allele 9 for example, $N_i = 2$ as there are two MVM:s given allele 9 as starting point.

2.2. Simulation

Simulations provide means to calculate prediction intervals and investigate specific likelihood ratio thresholds for a given case. The probability of falsely including/excluding a true hypothesis with a given LR threshold can be estimated. The interface may be utilized to examine the number of genetic markers we need to obtain a sufficiently good LR, prior to deciding to accept a case, as well as providing intervals. The simulation interface accounts for all parameters in the model.

Specifically, the simulation algorithm starts by detecting all founders for a given pedigree. Founder genotypes are sampled using defined allele frequencies in combination with possible subpopulation correction, modeled by the parameter θ (sometimes denoted Fst in the literature). Furthermore, transitions within the pedigree are sampled using a transition matrix, the latter depending on the selected mutation model. Interested users may use raw data from the simulations to study observed mutation rates or the occurrence of silent alleles. Moreover, in addition to providing prediction intervals, the interface provide relevant functionality to study thresholds and false positive/negative rates, i.e., given two mutually exclusive hypotheses, H_1 and H_2, the probability $P(\mathrm{LR} \geq x|H)$ is estimated for a given threshold x and an assumed hypothesis H. Simply put, it gives the probability of obtaining a LR at least as great as a given threshold.

2.3. DVI

Since the introduction of DNA, genetic data from relatives or personal belongings of missing persons have become one of the most important and reliable means of identification [10–12,14,21]. The disaster victim identification (DVI) module in Familias is provided to assist in any operation that requires an all-against-all search. To specify, we have K number of unidentified DNA profiles and M number of reference DNA profiles. The former data set may be reduced to K' as identical DNA profiles are found through blind matching while the latter is reduced to L' if some of the M reference profiles belong to the same cluster, i.e., in this setting meaning the same reference family. We have $K \geq K'$ and $L \geq L'$. For $k = 1, \ldots, K'$ we compare each unidentified DNA profile k with the $l = 1, \ldots, L'$ reference families. In Familias we specify two sets of data; PM (Post Mortem) data – obtained from unidentified remains, where several of the remains as mentioned may originate from the same individual and AM (Ante Mortem) data, where we define missing persons. In the AM data we define reference families for each missing person, where we may have genetic data from relatives of the missing person or direct matching samples such as personal belongings. The reference family can contain arbitrary pedigree structures., Complex pedigrees, mutation models (see below) and theta correction, will typically produce longer computation times. The module calculates likelihoods for each combination of PM data and AM data. Using a Bayesian approach the likelihoods are converted to posterior probabilities including prior probabilities set by the user. The choice of priors has been debated elsewhere [22] and can be influenced in Familias by changing the size of the DVI operation. As of now, meta data is not used to adjust priors or to exclude unidentified persons based on gender. This may in some situations be appropriate as the meta data may have been incorrectly specified.

In addition to the DVI module, there is a blind search interface, allowing the user to search a set of persons for unknown relations. The feature may be used on any data set, e.g. to search for relations in a set of individuals before creating a population frequency database or as in the DVI situation to find direct matches or relations between PM samples. The blind searching is restricted to pairwise searches for a number of predefined relationships and implements a fast algorithm based on the formulas presented in Table 4 in Hepler et al. [23]. The algorithm does not account for mutation unless the parent–child relation is chosen; while theta correction is applied in all scenarios should the value be nonzero. Briefly, the formulas implemented are based on identical by descent (IBD) sharing probabilities not accounting for inbreeding. The general formula is,

$$P(data|H) = P(IBD = 0|H)g_0 + P(IBD = 1|H)g_1 + P(IBD = 2|H)g_2 \quad (1)$$

where $P(IBD = 0|H) = k_0$, $P(IBD = 1|H) = k_1$ and $P(IBD = 2|H) = k_2$ are the probabilities that two individuals share 0, 1 respectively 2 alleles identical by descent.; g_0, g_1 and g_2 are functions of allele probabilities depending only on the genotype *data*. A more general formula, also accounting for inbreeding can be derived, though its utility in the current setting is limited.

For the new direct matching feature, Familias implements a general approach. To specify, consider two profiles G_1 and G_2. Further, consider the competing hypotheses:

H_1: The profiles belong to the same person
H_2: The profiles belong to two unrelated persons

The hypotheses, and the current setting, is distinct from the more common situation where we have some trace evidence from a crime scene and a reference profile to compare with. The former being uncertain while the latter is commonly considered to be accurate.

To compute the LR we require some more definitions. We consider a *latent* genotype G_{true}, consisting of all possible genotypes for the current marker. We can now specify the LR as

$$LR = \frac{P(G_1, G_2|H_1)}{P(G_1, G_2|H_2)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} P(G_{true,i,j})P(G_1|G_{true,i,j})P(G_2|G_{true,i,j})}{P(G_1)P(G_2)} \quad (2)$$

where N is the number of alleles at the current marker and $P(G_{true,i,j})$ is the genotype probability, $p_i{}^*p_j$, for the *latent* genotype with alleles i and j. $P(G_1|G_{true,i,j})$ and $P(G_2|G_{true,i,j})$ are the transition probabilities from the *latent* genotype to the observed genotypes. To calculate the transition probabilities in the direct matching we specify three parameters, d = allelic dropout probability, c = allelic dropin probability and e = typing error probability. Here, we specify dropout as the probability of one allele not being unobserved for a heterozygous genotype (allelic dropout), dropin as the probability of an extra allele being observed for a homozygous genotype (allelic dropin) and typing error as the probability of some other laboratory error leading to an incorrect genotype [24]. Dropouts, dropins and errors are assumed to occur independently. Note that these parameters only apply to direct matching function and are not used in the kinship calculations. See Table 1 below for a list of $\mathbf{P(G_1,G_2|H_1)}$ **and** $\mathbf{P(G_1,G_2|H_2)}$ for some combinations of genotypes G_1 and G_2. (The formulas are simplified to fit, removing terms negligible in the calculations assuming $d >> c > e$; the implementation is exact, see Supplementary data 1 for a more thorough walkthrough of Eq. (2), including an example where the simplifying assumptions are omitted)

We see that if $d = c = e = 0$, the LR $\mathbf{[P(G_1,G_2|H_1)/P(G_1,G_2|H_2)]}$ in the first and fifth line of Table 1 reduces to $1/P(A,A)$ and $1/P(A,B)$, while the remaining lines simplifies to zero. Further note that if $d >> c > e$ and d is comparatively small, say below 0.1, several latent genotypes are unlikely as the transition probabilities are very small. Moreover, if subpopulation correction is nonzero the allele probabilities are not independent. The user-friendlyness of handling three parameters (d, c and e) instead of one can be discussed. Similar to Merlin [20], one may instead use a general error variable, including all the effects possibly causing an erroneous genotype.

Table 1
LRs based on the direct matching feature of Familias.

G_1	G_2	$P(G_1,G_2\|H_1)$	$P(G_1,G_2\|H_2)$
A,A	A,A	$[(1-d^2)^2(1-e)^2(1-c)^2]P(A,A)$	P(A,A)*P(A,A)
A,A	A,B	$(1-e)^2[P(A,A)(1-d^2)^2cP(B)+P(A,B)d(1-d)(1-c)(1-d^2)]$	P(A,A)*P(A,B)
A,A	B,B	$(1-e)^2[(1-c)^2d^2(1-d)^2]P(A,B)$	P(A,A)*P(B,B)
A,A	B,C	$(1-e)^2[P(A,B)d(1-d)cP(C)+P(A,C)d(1-d)cP(B)]$	P(A,A)*P(B,C)
A,B	A,B	$(1-e)^2[(1-d)^2(1-c)^2]$ P(A,B)	P(A,B)*P(A,B)
A,B	B,C	$(1-e)^2[P(B,B)(1-d)^2c^2P(A)P(C)]$[a]	P(A,B)*P(B,C)
A,B	C,D	$(1-d^2)^2(1-c)^2[P(A,B)e+P(C,D)e]$[a]	P(A,B)*P(C,D)

[a] Note that neither of the observed genotypes, G_1 or G_2, is probable as the latent genotype.

2.4. Validation

Validation can mean several things, including validation of methods and validation of the implementation. Here we focus on approaches that may be of general interest and which can be used to validate also other programs. Specific validation examples showing correct numerical results, i.e. results that can be derived by other means, typically exact formulae, appear in Supplementary data 2. (Some useful validation files are available at the Familias homepage)

2.4.1. Some useful validation formulae in simulations

The expected value of the LR assuming the denominator hypothesis H_2 to be true is 1

$$E(\mathrm{LR}|H_2) = (1-p)0 + p\frac{1}{p} = 1 \tag{3}$$

where p is the random match probability and 0 and 1/p are the two possible values for the likelihood ratio. This follows directly from the definition of the likelihood ratio and expectation as pointed out by Thompson [18]. This is true also if mutations and population substructure are modeled. Slooten and Egeland [25] presents further theoretical properties of LR:s. For instance, the identity

$$SD(\mathrm{LR}|H_2) = \sqrt{E(\mathrm{LR}|H_1) - 1} \tag{4a}$$

relating the standard deviation (SD) under H_2 to the expected value under H_1. This last equation, however, is not valid when there are mutations or theta correction is made.

Eqs. (3) and (4) can be used to check simulations under the denominator hypothesis when p is not too small, typically for one marker. When p is small, say below 10^{-10} any reasonable number of simulations should lead to all LR-s being 0 as the probability of a random match is then negligible.

Turning to validation for simulations under the numerator hypothesis, the general formula for the expected value for all pairwise, non-inbred relationships presented in Slooten and Egeland [25] can be used

$$E(\mathrm{LR}|H_1) = \alpha L^2 + \beta L + (1-\alpha-\beta) \tag{4b}$$

$$\text{where } L = \text{alleles}, \alpha = \frac{k_2^2}{2}, \text{ and } \beta = \frac{k_1^2 + 4k_1k_2 + 2k_2^2}{4}$$

As an example, note for a parent–child relation $k_1 = 1$ and $k_2 = 0$ and the expected LR is therefore $(L+3)/4$ for one marker. This generalizes directly to n independent markers

$$E(\mathrm{LR}|H_1) = \prod_{i=1}^{n}\frac{L_i+3}{4}$$

where L_i is the number of alleles for marker i.

We have checked the code using the above formulae for one marker at the time. To get an indication of the simulation uncertainty, several simulations can be run with different seeds.

Exact calculations are hard for general mutation models. There is, however, one exemption as explained next. Consider the hypotheses H_1:*AF is the father CH* and H_2:*AF and CH are unrelated.* The genotypes of AF and CH are denoted a/b and c/d. For instance, if both individuals are homozygote 9,9 then $a = b = c = d = 9$. A case which would need a mutation to be consistent with paternity occurs for genotypes 9,9.3 and 10,10.3 corresponding to $a = 9$, $b = 9.3$, $c = 10$ and $d = 10.3$. The likelihood ratio may be written [26]

$$\mathrm{LR} = \frac{1}{4}\frac{(m_{ac}+m_{bc})\,p_d + (m_{ad}+m_{bd})\,p_c}{p_c\,p_d} \tag{5}$$

where p denotes allele frequency. Example 2 below relies heavily on the above equation.

2.5. Implementation

The software functionality described herein is implemented in a Windows friendly software, Familias version 3.1.4 at the time of writing. See Supplementary data 2 for some validation examples. The mayor changes since Familias 2.0 is the introduction of the new mutation model, the simulation interface as well as the new DVI module. We also introduce a new blind match searching function implementing some new functionality, primarily connected to the direct matching, see previous description. The latest version of Familias is freely available at www.familias.no. Moreover, several other new features will be presented in the next releases, e.g. the possibility to model profiles with dropouts [Manuscript submitted].

3. Results

3.1. New mutation model and simulation

Example 2. In this example both simulation and the new mutation model is illustrated. Consider one marker with the mutation model and alleles as described in Section 2 of this paper. The mutation parameters are specified as:

$$R = 0.005,\ r = 0.1 \text{ and } \alpha = 0.001.$$

The mutation matrix M becomes

Allele	9	9.3	10	10.3	15
9	9.940e−01	2.973e−05	5.945e−03	2.973e−05	5.945e−08
9.3	1.982e−05	9.940e−01	1.982e−05	5.945e−03	1.982e−05
10	5.939e−03	2.973e−05	9.940e−01	2.973e−05	5.939e−07
10.3	1.982e−05	5.946e−03	1.982e−05	9.940e−01	1.982e−05
15	5.929e−06	2.967e−03	5.929e−05	2.967e−03	9.940e−01

For the numerical examples below, the allele frequencies for the alleles (9, 9.3, 10, 10.3, 15) are (0.05, 0.05, 0.20, 0.30, 0.40). From Eq. (5) we find, when the alleged father is 9, 9.3 and the child 10, 10.3

$$\mathrm{LR} = \frac{1}{4}((5.94\mathrm{e}-03 + 1.98\mathrm{e}-05)*0.2 + (2.97\mathrm{e}-05 + 5.94\mathrm{e}-03)*0.3)/(0.2*0.3)$$
$$= 0.0124.$$

which is accurately reproduced by Familias 3. Similarly, simulations closely reproduce the theoretical values. For instance, the expected value of the *LR* assuming AF and CH to be unrelated is 1

according to Eq. (3) and the computer output based on 10,000 simulations gives a value close to the theoretical. Furthermore, the expected LR assuming AF to be the father, $(L + 3)/4 = (5 + 3)/5 = 2$, from Eq. (4b) is also consistent with simulations.

3.2. DVI module and blind search interface

To validate the DVI module simulated data was constructed for a number of relationships (Data available upon request). Specifically, 100 pairs of siblings, 100 pairs of grandparents/grandchildren and 100 pairs of parent/childs were generated using the simulation interface. For each pair one of the individuals was withdrawn and denoted as missing. All missing persons, in total 300, were collected into a data set of unidentified remains. The reference families were constructed according to the simulated relationship, i.e., 100 families where the reference data was from siblings, 100 families where the reference data was from grandparents and 100 families where the reference data was from a parent. An all-against-all search was performed in the DVI module, where LRs were calculated for all possible combinations of unidentified remain and reference family. In total $300 \times 300 = 90{,}000$ comparisons were done, producing a list of matches above a given threshold (in this case set as low as LR = 1). The match list indicated some false matches (i.e. false inclusions), which is most probably due to the low LR threshold. However, no false match obtained a LR higher than the *true* match. Some true matches for the missing persons obtained very low LR barely above 1.0, which was in some of the cases explained by simulated mutations (grandparents and parent) and in other cases by low number of shared alleles (sibling cases), (Data available upon request). See also Ge et al. for a discussion on choice of reference family relatives in DVI operations [27]. The point with this validation is not to investigate the match threshold but rather to demonstrate the accuracy in the calculations.

We further use constructed data to validate the blind searching function. Consider a system with alleles similar to the first example, i.e., the allele frequencies for the alleles [9, 9.3, 10, 10.3, 15] are [0.05, 0.05, 0.20, 0.30, 0.40]. For simplicity we let the mutation rate be zero, while we consider both $\theta = 0$ and $\theta = 0.01$. Consider two persons P1 and P2 with genotypes G_1 and G_2. We can now easily calculate the likelihood ratio for the predefined relationships in the blind search interface using Eq. (1). Note that the interface allows us to scale versus some other relationship rather than unrelated, but for the current calculation we use unrelated as the alternative hypothesis.

Let $G_1 = 9,9$ and $G_2 = 9,10$. For $\theta = 0.01$ we need to calculate the updated set of frequencies, $[p(9), p(9|9), p(9|9,9), p(10|9,9,9)] = [0.05, 0.0595, 0.0688, 0.194]$, using formulas in Balding et al. [28]. Note that this set of frequencies will change if two alleles are IBD, i.e. for IBD = 2 and IBD = 1 we need to update the frequencies as only two respectively three alleles are drawn from the population.

Table 2
LRs for some relationship hypotheses, calculated versus unrelated as alternative hypothesis, for a pair of individuals P1 and P2.

Relationship	LR ($\theta = 0$)	LR ($\theta = 0.01$)
Direct match	29.07	18.12
Siblings	5.25	3.919
Half siblings	5.5	4.169
Cousins	3.25	2.584
Parent–child	10	7.338
2nd cousins	1.5625	1.396

The likelihood for the different hypotheses of relatedness can now be calculated from Eq. (1) as, $L(Data|H) = k_0 2\,p(9)^2\,p(9)\,p(10) + k_1\,p(9)\,p(9)\,p(10)$, where k_0 and k_1 are replaced by the values according to the relationship H.

The direct match LR can be calculated according to Eq. (2), by summing over all possible genotypes for the *latent* genotype and compute the likelihood for each case according to: (We specify $d = 0.1$, $e = 0.001$ and $c = 0.001$, which are the default values in Familias)

$$\begin{aligned}
\mathrm{LR} &= \frac{P(G_1, G_2|H_0)}{P(G_1, G_2|H_1)} = \frac{\sum_{i=1}^{A}\sum_{j=i}^{A} P(G_{true,i,j})P(G_1|G_{true,i,j})P(G_2|G_{true,i,j})}{P(G_1)P(G_2)} \\
&= \langle \text{We simplify and remove terms which is negligible in the numerator} \rangle \\
&= \frac{p(9)\,p(9|9,\theta)P(9,9|9,9)P(9,10|9,9) + 2\,p(9)\,p(9|10,\theta)P(9,9|9,10)P(9,10|9,10)}{2\,p(9)\,p(9|9,\theta)\,p(9|9,9,\theta)\,p(10|9,9,9,\theta)} \\
&= \frac{(1-e)^2[\,p(9)\,p(9|9,\theta)(1-d^2)c\,p(10) + 2\,p(9)\,p(9|10,\theta)d(1-d)(1-d)^2(1-c)]}{2\,p(9)\,p(9|9,\theta)\,p(9|9,9,\theta)\,p(10|9,9,9,\theta)}
\end{aligned}$$

The theoretical values in coincide with the values calculated in Familias (Table 2). Also note that the Direct match obtain a high LR even though the profiles are not identical, this is due to the high values on the parameters d, c and e.

3.3. Simulations

To further corroborate output from the simulation interface we compared results on some standard forensic cases with simulations reported in Table 6 of Ge et al. [27], see Table 3. The investigated relationships are described elsewhere, op.cit., but are based on simulations on the standard 13 CODIS STR markers, in order to determine how many relatives are necessary in a given

Table 3
Distribution of log10 likelihood ratios for 10,000 simulations using three different methods.

Method	Pedigree	Mean	5percentile	1percentile
Familias3_no_mut	Both parents	10.25	8.1	7.4
Ge et al.	Both parents	10.26	8.07	7.34
Familias3_mut	Both parents	10.17	7.85	6.63
Familias3_no_mut	One parent/One child	4.08	2.47	1.90
Ge et al.	One parent/One child	4.09	2.48	1.92
Familias3_mut	One parent/One child	4.07	2.43	1.69
Familias3_no_mut	2 full sibs	5.88	2.64	1.25
Ge et al.	2 full sibs	5.88	2.65	1.34
Familias3_mut	2 full sibs	5.86	2.48	1.09
Familias3_no_mut	1 halfsib	0.92	−0.59	−1.16
Ge et al.	1 halfsib	0.91	−0.57	−1.16
Familias3_mut	1 halfsib	1.16	−0.70	−1.29
Familias3_no_mut	2 children (same parent 2)	6.97	4.31	3.43
Ge et al.	2 children (same parent 2)	6.98	4.33	3.35
Familias3_mut	2 children (same parent 2)	6.94	4.24	3.14

The methods are Familias 3 (with and without mutations considered) as well as results presented by Ge et al., the pedigrees are described elsewhere [27].

case to obtain sufficient LRs. The Familias simulation interface produces almost identical output as presented by Ge and colleagues. As a comparison we also included simulations using the extended stepwise mutation model and the results are still close to the simulations without mutations.

4. Discussion

Familias is a well-known software in the forensic community and used by a number of laboratories [5]. The software facilitates the interpretation of the evidence by computing likelihood ratios and posterior probabilities for a given set of relationship hypotheses and genetic marker data. This paper describes methods implemented in the new version (Familias 3), providing considerable extensions to previous versions [4].

A comprehensive simulation interface provides versatile functionality for studying distribution of likelihood ratios for a given case. Users may now investigate a case prior to accepting it by computing prediction intervals and decide whether decisive evidence is likely to be obtained. The authors are aware of the discussion in the forensic community on the use of case specific thresholds rather than using an general LR/Posterior probability threshold for all cases. We do not propagate for lowering the threshold only because for a given case the evidence will never reach the required value. The users should instead study the false positive/negative rates to find an appropriate limit. As presented in this paper, the algorithm can simulate arbitrary pedigree structures where the only limitation is set by the computation time.

To assist in mass disaster identifications, we have developed a DVI module, allowing users to handle small to medium scale identifications. There are several papers and online discussions following previous larger scale mass disaster incidents, e.g. the Tsunami disaster [12], the WTC terror attack [11,14] and the hurricane Katrina [29,30]. This paper includes some points on the implementation and interested users should follow the references given above for further mathematical discussions. Similar to the simulation interface, the DVI module adopts the full functionality of Familias, allowing for subpopulation frequency correction, silent alleles and mutations. The module further allows the definition of multiple alternative family hypotheses, within each family, thus permitting each reference family to have several missing persons and the user can weigh the evidence given a match based on the possibility that the unidentified person may fit in several locations in a family tree.

To further aid in the identification of unidentified remains, a blind search tool is included. As presented in this paper the tool can be used to rapidly scan data sets for unknown relations; unknown in the sense that we have no prior knowledge how the individuals in the data set are related. In addition to assist in DVI operations the search can also be performed to verify that data sets for the creation of population frequency databases do not contain related individuals. The blind search is restricted to pair wise comparisons on a number of predefined relationships implementing the formulas presented in Hepler et al. not accounting for inbreeding and mutations [23]. As the formulas are general in the sense that any non-inbreed pair wise relationship can be defined, the implementation in Familias opens up for future extensions where any non-inbred relationship between two individuals could be specified using the k_0, k_1 and k_2 parameters, see Eq. (1). Furthermore, the search also includes a newly developed direct matching function (also part of the DVI module), which incorporates dropout, dropin and typing error probabilities. The latter is probably hard to estimate but can in some situations not be neglected, and therefore equally important as the two first mentioned probabilities.

Further, to cope with the increasing polymorphism in the new STR markers, we have developed a new mutation model. The model builds on the stepwise model [7], but provides extensions for microvariants, e.g. 9.3. Microvariants are more and more common, for instance the STR marker SE33 (ACTPB2) includes several alleles with a non-integer repeat unit and even though mutation rates for transitions between non-integer alleles and integer alleles may sometimes be negligible we require an appropriate model to handle them. This transition model is not stationary. In other words, the distribution of allele frequencies will change slightly with each generation in the pedigree. A stationary version of the above model would be a welcomed extension. Such an extension should preserve the main features like the diagonal elements, i.e., the overall mutation probability. We have not yet been able to derive such a stationary model.

In summary, the software Familias has previously been proven to be a resourceful tool in calculations concerning genetic relatedness [4–6]. We believe the extensions provided in this paper will be important for many users where previous versions have lacked desired functionality. The latest version can be freely downloaded at http://www.familias.no.

Acknowledgements

The authors would like to thank the contributions from users testing and evaluating the new features of Familias. The work of the last author leading to these results was financially supported from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 285487 (EUROFORGEN-NoE).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2014.07.004.

References

[1] R.C. Elston, J. Stewart, A general model for the genetic analysis of pedigree data, Hum. Hered. 21 (1971) 523–542.
[2] J. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press Inc., Bosa Roca, USA, 2004.
[3] C. Cannings, E. Thompson, M. Skolnick, Probability functions on complex pedigrees [domesticated mammals, laboratory animals], Adv. Appl. Probab. 10 (1978) 26–61.
[4] T. Egeland, P.F. Mostad, B. Mevåg, et al., Beyond traditional paternity and identification cases. Selecting the most probable pedigree, Forensic Sci. Int. 110 (2000) 47–59.
[5] L. Poulsen, S.L. Friis, C. Hallenberg, et al., A report of the 2009–2011 paternity and relationship testing workshops of the English Speaking Working Group of the International Society For Forensic Genetics, Forensic Sci. Int. Genet. 9 (2013) e1–e2.
[6] J. Drabek, Validation of software for calculating the likelihood ratio for parentage and kinship, Forensic Sci. Int. Genet. 3 (2009) 112–118.
[7] T. Ota, M. Kimura, A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population, Genet. Res. 22 (1973) 201–204.
[8] H. Ellegren, Heterogeneous mutation processes in human microsatellite DNA sequences, Nat. Genet. 24 (2000) 400–402.
[9] B. Olaisen, M. Stenersen, B. Mevag, Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster, Nat. Genet. 15 (1997) 402–405.
[10] B. Leclair, C.J. Fregeau, K.L. Bowen, et al., Enhanced kinship analysis and STR-based DNA typing for human identification in mass fatality incidents: the Swissair flight 111 disaster, J. Forensic Sci. 49 (2004) 939–953.
[11] L.G. Biesecker, J.E. Bailey-Wilson, J. Ballantyne, et al., Epidemiology. DNA identifications after the 9/11 World Trade Center attack, Science 310 (2005) 1122–1123.
[12] C.H. Brenner, Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities, Forensic Sci. Int. 157 (2006) 172–180.
[13] M. Prinz, A. Carracedo, W.R. Mayr, 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 (2007) 3–12.

[14] B. Leclair, R. Shaler, G.R. Carmody, et al., Bioinformatics and human identification in mass fatality incidents: the world trade center disaster, J. Forensic Sci. 52 (2007) 806–819.
[15] L. Bradford, J. Heal, J. Anderson, et al., Disaster victim investigation recommendations from two simulated mass disaster scenarios utilized for user acceptance testing CODIS 6.0, Forensic Sci. Int. Genet. 5 (2011) 291–296.
[16] C.H. Brenner, Symbolic kinship program, Genetics 145 (1997) 535–542.
[17] K. Slooten, Validation of DNA-based identification software by computation of pedigree likelihood ratios, Forensic Sci. Int. Genet. 5 (2011) 308–315.
[18] E.A. Thompson, Statistical Inference from Genetic Data on Pedigrees, JSTOR, 2000.
[19] D. Kling, T. Egeland, A.O. Tillmar, FamLink – a user friendly software for linkage calculations in family genetics, Forensic Sci. Int. Genet. 6 (2012) 616–620.
[20] G.R. Abecasis, S.S. Cherny, W.O. Cookson, et al., Merlin – rapid analysis of dense genetic maps using sparse gene flow trees, Nat. Genet. 30 (2002) 97–101.
[21] C.H. Brenner, B.S. Weir, Issues and strategies in the DNA identification of World Trade Center victims, Theor. Popul. Biol. 63 (2003) 173–178.
[22] B. Budowle, J. Ge, R. Chakraborty, et al., Use of prior odds for missing persons identifications, Investig. Genet. 2 (2011) 15.
[23] B.S. Weir, A.D. Anderson, A.B. Hepler, Genetic relatedness analysis: modern data and new challenges, Nat. Rev. Genet. 7 (2006) 771–780.
[24] A. Kloosterman, M. Sjerps, A. Quak, Error rates in forensic DNA analysis: definition, numbers, impact and communication, Forensic Sci. Int. Genet. 12 (2014) 77–85.
[25] K.-J. Slooten, T. Egeland, Exclusion probabilities and likelihood ratios with applications to kinship problems, Int. J. Legal Med. (2013) 1–11.
[26] F. Ricciardi, K. Slooten, Mutation Models for DVI analysis, Forensic Sci. Int. Genet. 11 (2014) 85–95.
[27] J. Ge, B. Budowle, R. Chakraborty, Choosing relatives for DNA identification of missing persons, J. Forensic Sci. 56 (Suppl. 1) (2011) S23–S28.
[28] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[29] S. Donkervoort, S.M. Dolan, M. Beckwith, et al., Enhancing accurate data collection in mass fatality kinship identifications: lessons learned from Hurricane Katrina, Forensic Sci. Int. Genet. 2 (2008) 354–362.
[30] S.M. Dolan, D.S. Saraiya, S. Donkervoort, 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 (2009) 414–417.