Forensic Science International: Genetics Supplement Series 7 (2019) 221–223

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

relMix: An open source software for DNA mixtures with related contributors

Elias Hernandis[a], Guro Dørum[b], Thore Egeland[c,*]

[a] Universidad Autónoma de Madrid, Spain
[b] Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland
[c] Norwegian University of Life Sciences, Norway

ARTICLE INFO

Keywords:
Forensic genetics
Mixtures
LR
Relatives
relMix

ABSTRACT

In both criminal cases and relationship inference there is an increasing demand for analysis of DNA mixtures where relatives are involved. The goal might be to identify the contributors to a mixture where the donors may or may not be related, or to determine the relationship between individuals based on a mixture. relMix is an open source software for analysing DNA mixtures involving relatives, available as a graphical user interface in R. We explain the model behind relMix and give an overview of the new features (including improved checking of input) in the latest version.

Introduction

In both criminal cases and relationship inference there is an increasing demand for analysis of DNA mixtures where relatives are involved. One example is prenatal paternity cases based on a mother-fetus mixture and reference samples from the mother, the alleged father, but obviously not the child. In crime cases one may encounter stains where two or more contributors are related. relMix is an open source software for analysing DNA mixtures involving relatives, available from https://CRAN.R-project.org/package=relMix as a graphical user interface in `R`. Compared to commonly used mixture software, relMix can account for arbitrary kinship between more than two contributors in addition to mutations and silent alleles.

Motivating example

Investigators want to determine the father of an unborn child where the candidates are brothers. Available evidence consists of DNA reference samples from the mother, brother 1 and brother 2. In addition, a sample from the mother contains a mixture between her DNA and the DNA of her unborn child. Based on this we formulated

H_1 : Brother1isthefather
H_2 : Brother2isthefather

as shown in Figure 1. For the discussion we also included

H_3: Anunrelatedmanisthefather.

For this case we consider an equal mutation model with mutation probabilities 0.001 and 0.003 for females and males, respectively. The dropout probabilities were 0.05 for the child and 0 for the mother.

The evidence will be summarised by the likelihood ratios

$$\mathrm{LR}_1 = \frac{P(\text{data}\,|H_1)}{P(\text{data}\,|H_2)} \qquad \mathrm{LR}_2 = \frac{P(\text{data}\,|H_1)}{P(\text{data}\,|H_3)}$$

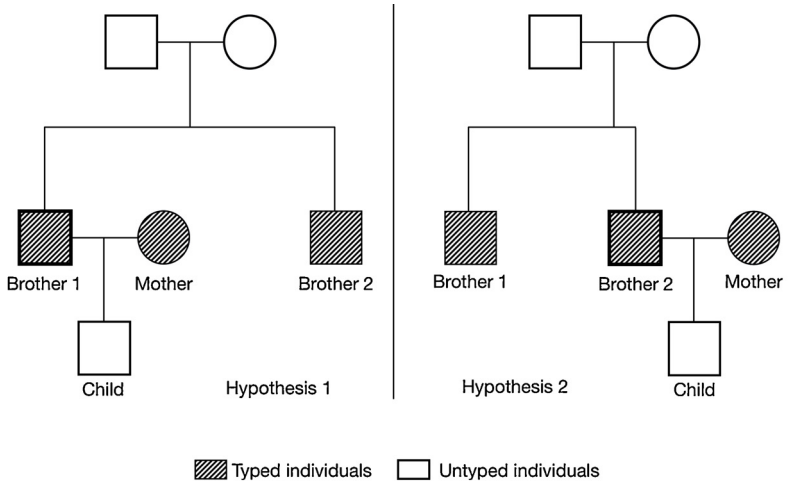
Consider the table in Figure 1. The first row is consistent with all hypotheses. For D19S433 a mutation or a dropout is needed for H_1 but not for H_2. D21S11 is consistent with H_1 but a mutation is needed for H_2. The final line shows clear evidence, but by most standards not conclusive, in favor of H_1. It is correct to report $LR_1 = 1380$ since we were asked to compare brother 1 to brother 2. If we inappropriately compared brother 1 to an unrelated man, we would get an *LR* that overestimates the evidence.

Program input

relMix works with tab-separated files to import DNA and allele frequency data. These can be exported from DNA profiling or spreadsheet software. Pedigrees for paternity cases are included with the program while other arbitrarily complex pedigrees can be loaded using the `Familias` (https://www.familias.name/openfamilias.html) format. Finally, parameters describing mutation, drop-in, drop-out, silent alleles, and population substructure (θ) are entered manually through a user friendly interface as shown in Figure 2.

* Corresponding author.
E-mail address: thore.egeland@nmbu.no (T. Egeland).

https://doi.org/10.1016/j.fsigss.2019.09.085
Received 12 September 2019; Accepted 25 September 2019
Available online 17 October 2019



Marker	Brother 1	Brother 2	Mother	Mixture	LR1	LR2
D10S1248	11/14	11/14	13/15	11/13/15	1.00E+00	2.54E+02
D19S433	13/13	13/14	12/14	12/14	9.62E-02	1.18E-01
D21S11	32/36	31/35	27/27	27/32	3.60E+03	4.03E+01
				Total LR (20 markers)	**1.38E+03**	4.83E+09

Fig. 1. Pedigrees and excerpt from the result table of the motivating example.

New in relMix version 1.3

relMix now checks for common mistakes such as marker name inconsistencies, duplicate markers, invalid file formats and more. In particular, manually typing reference data or manipulating frequency databases can lead to subtle errors that previously resulted in wrong calculations or programme termination (e.g., `TPOX vs TP0X`). In addition, reuse of frequency databases coming from other programmes can lead to problems if marker naming is not consistent. We introduce specific checks for these kinds of errors which are largely based on computing the Levenshtein distance [1] between identifiers to find those that are suspiciously similar. The Levenshtein distance counts the minimum number of edits (substitutions, insertions or deletions) required to go from one string of text to another. Detected errors are presented to the user with an explanation and automatically fixed if possible. We found that setting a threshold of 2 for automatic correction of inconsistencies was beneficial because it also allows for transpositions in addition to the previous edits which are a common typing error (e.g., `Plasma vs. Palsma`).

Development of this last version was done in GitHub, a platform that enables efficient collaboration between different authors in a project. In addition, all changes made to the codebase and the codebase itself are public, allowing for greater transparency and encouraging collaboration with other external developers. Towards end users, GitHub provides a mechanism for bug reporting and contacting the authors in which the questions and answers posted remain public and searchable for the benefit of the community. The adoption of this new workflow and development methodology is an important step for open/free software.

Discussion

The case presented demonstrates that relMix can deal with complex cases of practical significance. The importance of modelling relationships and mutations, is clearly demonstrated. LRmixStudio (https://lrmixstudio.org) is based on a model similar to the one we use. This software includes important functionality not available in relMix, but only simple pairwise relationships. Alternative software like EuroForMix (http://www.euroformix.com) is based on continuous models. Peak height information, which may or may not be important as discussed in [2], is therefore accounted for. Alternative models and implementations based on Bayesian networks are exemplified in [3].

The model

We adopt the mixture model described in [4] and [5]. The model accounts for dropout and drop-in, but not peak heights. For a given locus, the probability that allele a will not appear or will appear in the mixture $\mathcal{M}$, respectively, is found as

$$P(a \notin \mathcal{M}|\mathbf{g}, \mathbf{d}, c) = (1 - \mathrm{cp}_a) \prod_i d_i^{n_{i,b}},$$
$$P(a \in \mathcal{M}|\mathbf{g}, \mathbf{d}, c) = 1 - (1 - \mathrm{cp}_a) \prod_i d_i^{n_{i,a}},$$

where

$\mathbf{g}$ = genotypesofallcontributors
$\mathbf{d}$ = dropoutprobabilitiesforallcontributors
d_i = dropoutprobabilityforcontributor i
cp_a = probabilitythat a willdropin
$n_{i,a}$ = numberoftimes a isobservedincontributor i

The probability of observing a set M of mixture alleles is thus

$$P(\mathcal{M} = M|\mathbf{g}, \mathbf{d}, c) = \prod_{a \notin M} P(a \notin M|\mathbf{g}, \mathbf{d}, c) \cdot \prod_{a \in M} P(a \in M|\mathbf{g}, \mathbf{d}, c).$$

Finally, the probability of the evidence E conditioned on hypothesis H_j is found by combining the probability of the mixture with the probability of the kinship as

$$P(E|H_j) = \sum_{u \in U} P(\mathcal{M} = M|\mathbf{g}_\mathbf{K}, \mathbf{g}_\mathbf{U} = u, \mathbf{d}, c) \cdot P(\mathbf{g}_\mathbf{A}, \mathbf{g}_\mathbf{K}, \mathbf{g}_\mathbf{U} = u|H_j),$$

where

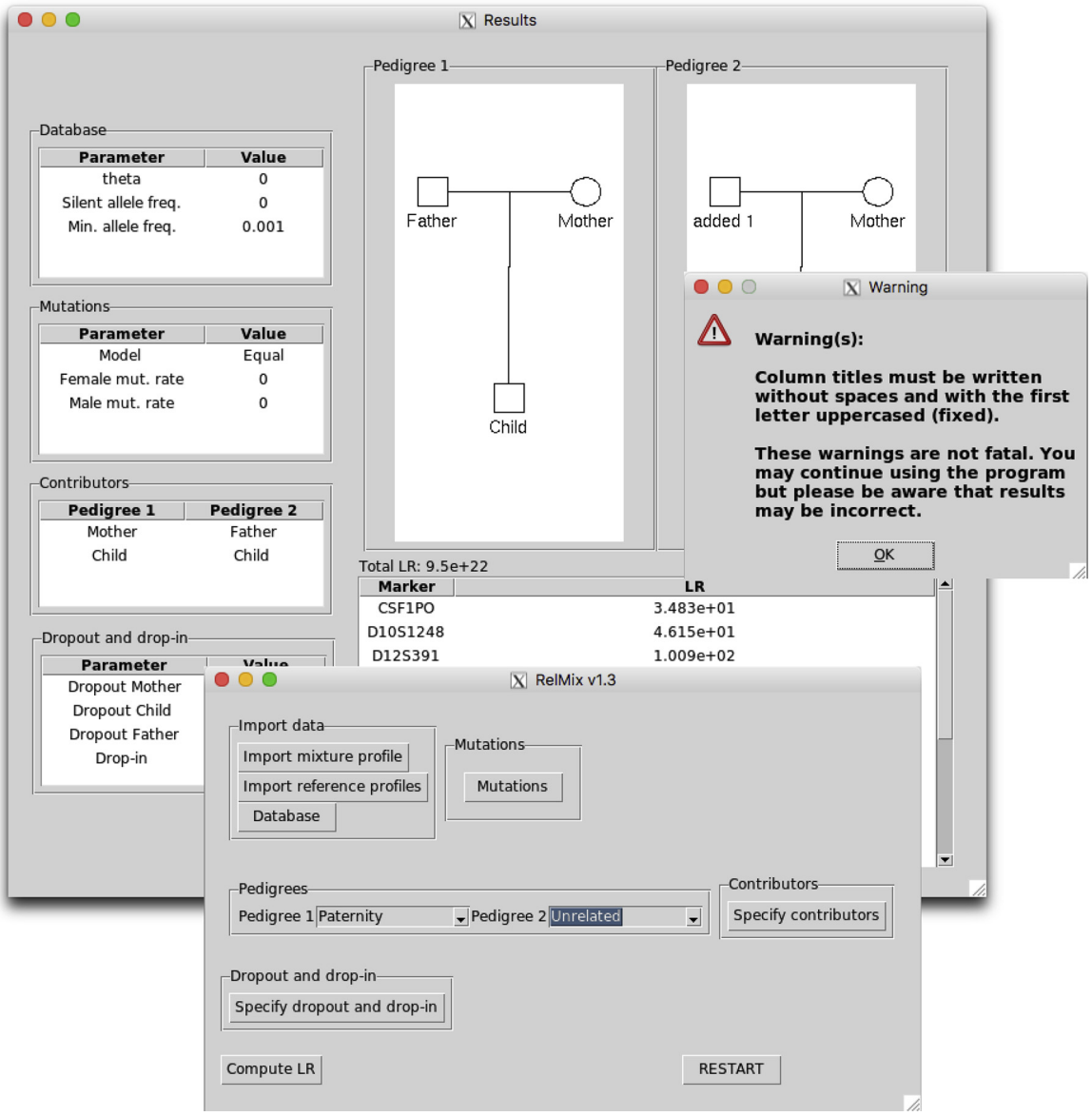


Fig. 2. Examples from the relMix user interface, including the new input validation system.

$\mathbf{g_K}$ = Genotypesofknowncontributors
$\mathbf{g_U}$ = Genotypesofunknowncontributors
$\mathbf{g_A}$ = Genotypesofadditionalgenotypedindividuals
U = Setofpossiblegenotypesfortheunknown contributor(s)

Calculations are based on the `R` version of `Familias`.

References

[1] Levenshtein distance. https://en.wikipedia.org/wiki/ Levenshtein_distance.
[2] K. Slooten, The information gain from peak height data in DNA mixtures, Forensic Sci. Int. Genet. 36 (2018) 119–123.
[3] P.J. Green, J. Mortera, Paternity testing and other inference about relationships from DNA mixtures, Forensic Sci. Int. Genet. 28 (2017) 128–137.
[4] G. Dørum, N. Kaur, M. Gysi, Pedigree-based relationship inference from complex DNA mixtures, Int. J. Legal Med. 131 (3) (2017) 629–641.
[5] H. Haned, K. Slooten, P. Gill, Exploratory data analysis for the interpretation of low template DNA mixtures, Forensic Sci. Int. Genet. 6 (6) (2012) 762–774.