Relationship Inference with Familias.
EUROFORGEN–NoE Webinar

Thore Egeland
Ana Mosquera

November 10, 2016

Contents

1 Background and preparations for the webinar

This document is prepared for the EUROFORGEN-NoE http://www.euroforgen.eu/ webinar on “Relationship Inference and Familias”, the first scheduled for Nov 9 2016. The participant we have in mind is a case worker or a scientist working in a forensic lab or an academic institution and may have attended some of the many EUROFORGEN NoE courses listed here http://familias.name/book.html. At any rate, we assume basic knowledge of the topic summarised in the title. We have included a tutorial in the appendix which may serve as reference. Chapters 2 and 3 of [2] presents the topics with all details and the exercises in this document are revised and updated versions of similarly numbered exercises in [2].

The purpose of the webinar is to provide a review and an update and also provide the opportunity to discuss with tutors and colleagues. In this way we hope to maintain and strengthen the network established in the EUROFORGEN NoE project.

Prior to the webinair participants are encouraged to download the last version of the Familias software, freely available from http://familias.no (Released 2016-09-19). Furthermore, please download the input files needed to follow the the exercises discussed in the webinair. These files are contained in the zipped folders http://familias.name/Ch2Input.zip and http://familias.name/Ch3Input.zip.

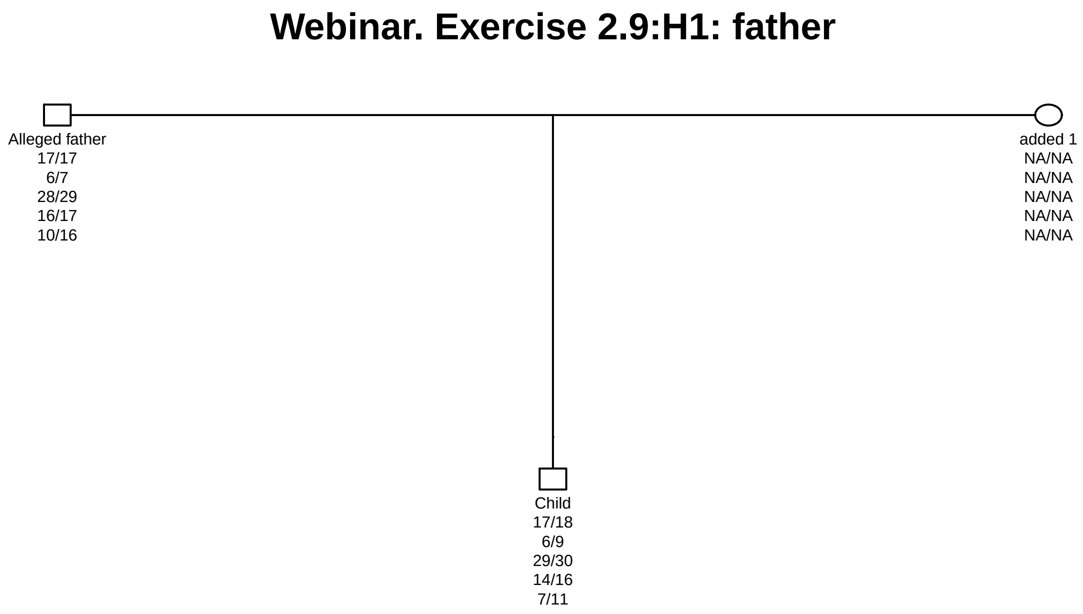


Figure 1: Pedigree for Exercise 2.9. The first five markers are shown.

2 Exercises 2.9 and 2.17: Relationship testing and Simulation

Exercise 2.9 (Paternity case with mutation)**.**

Load the file `Exercise2_9.fam`, please see Figure 16.

a) How many markers are there? What are the persons of the case? Where can you find the genotypes. Formulate the hypotheses. Verify that $LR = 0$.

b) There is one marker where the child and the alleged father do not share an allele. Find this marker.

c) Use the `Stepwise (Stationary)` model, for females and males with mutation rate 0.001 and mutation range 0.5 for all markers and calculate LR. Explain what is meant by a stationary model. Explain the idea behind the extended stepwise model. Try this model.

d) Assume you are asked to consider the hypotheses H_3: Brother of alleged father is father. Calculate LR (H_1/H_3).

e) Is there a best mutation model? Should a mutation model be used routinely for all markers?

Exercise 2.17 (Simulation).

Load the file `Exercise2_17.fam`. How many markers are there? What are the persons of the case? Formulate the hypotheses implied by the input file. The file contains no genotype information. Use the simulation in `Familias` to simulate genotypes for both individuals. Untick `Random seed` and set seed to 12345. What is effect of specifying a `Random seed`? Use 1000 simulations and find

a) The mean LR(H_1/H_2) when H_1 is true.

b) The mean LR(H_1/H_2) when H_2 is true.

c) The probability of observing a LR larger than 50 when H_1 is true.

3 Exercise 3.3. Disaster victim identification

Exercise 3.3 (DVI - An extended example).

Consider the crash of a small plane with 10 passengers, see Figure 16. We have obtained reference data from 5 different families. There are many steps and the exercise may take some time, but we encourage users to push through all steps as there is a lot to learn by doing this.

a) In `Familias`, open the **Exercise3_3.fam** file, which contains frequency data for 23 autosomal markers.

b) Enter the first step in the DVI module, `Add unidentified persons`. We may define individuals manually, similar to normal `Familias` procedure, though we prefer importing data from file to skip as much manual input as possible. Import the file **Exercise3_3_pm.txt**. Note: `Familias` can import different files formats, e.g. CODIS xml and tab-separated text files.

c) The file only contains 8 unidentified remains. Discuss why this may be a realistic scenario, especially in larger scale scenarios. How may this effect the calculations?

d) Deselect `Use list` and enter 10 in the `Size` box. This is used to define the priors. We will not dwell on the discussion of priors for now. Briefly we define the number of missing persons to 10.

e) Press `Next` to define reference families. We may now either define families manually or we may import them from file. We will here consider two different alternatives. Define the first family manually by selecting `Add`. Enter a name for the family, *Family 1*.

f) Import data for the persons in the family (a father). Import the file **Exercise3_3_am1.txt**. (Note: it is not necessary to first manually define the typed persons.)

   If relevant, now is the time to define other persons included in the family, in the current family none. Note, this may be untyped persons necessary to define the relations between the reference persons and the missing person(s). We will return to an example of this later.

g) We continue with defining the relation between the defined person(s) and the missing person. (Note: simply naming the person father/mother/brother etc. does not define the relationships). Select `Add` in the pedigree section to add a new pedigree. Name the pedigree appropriately, *Father*, and add necessary relation between the reference person(s) and the missing person. Press `Close` and then `Close` again to return to the list of reference families.

h) Define also a second family, where data is available for a brother of a missing person, by pressing `Add`. Enter a name, *Family 2*.

i) Import reference persons from file **Exercise3_3_am2.txt**

j) Add necessary additional persons, untyped mother and father, and then define the reference person as brother to the missing person. *Hint*: Add a pedigree as for Family 1 and specify that the brother and the missing persons share the same parents.

k) Add the rest of the reference families by selecting the import option `Simple` and select files **Exercise3_3_am3.txt**, **Exercise3_3_am4.txt**, and

Exercise3_3_am5.txt. Change the names of the families to *Family 3*, *Family 4* and *Family 5*. Also, check the persons and pedigrees in each imported family to make sure you know the relationships. Rename the pedigrees to reflect the defined relationships.

l) Press `Next` and `Search` to start the matching. Select the threshold for a match to be reported. Enter 1.0, as we would rather obtain more matches at this stage and later remove matches which may be spurious.

m) Interpret the results. Were all remains identified?

n) Select a match and press `View match` to investigate the individual LRs for each system.

o) Change the size of the accident in step c) to 100 and see how this affects the priors in the current case. How does this in turn affect the

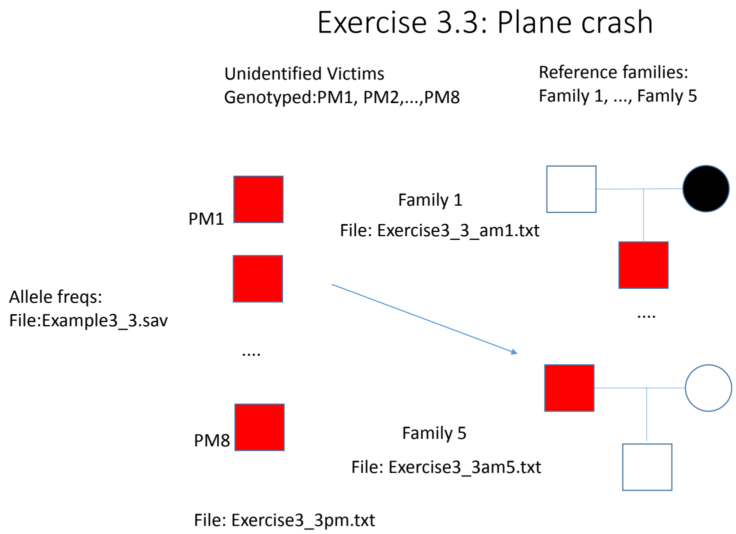
p) We suspect there might be relatives among the unidentified persons. Enter the first step, `Add unidentified persons` and select `Blind search`. Perform a blind search for siblings relations. (Use 10 as match threshold, leave all other options at default) How may the results be used in the DVI operation? posteriors? *Hint*: Perform a new search to see the effect.

q) * New information is added to the case. The first family, defined manually in d) also contains a second missing person. The brother of the reference father is also missing. Try finding out how this could be solved using the means available in the DVI module.

r) * Perform a new search, use the same match threshold as in e).

s) * Discuss the solution and other ways to improve the algorithm.

t) Save the project.



Exercise 3.3: Plane crash

Unidentified Victims
Genotyped:PM1, PM2,...,PM8

Reference families:
Family 1, ..., Famly 5

PM1

Family 1

File: Exercise3_3_am1.txt

Allele freqs:
File:Example3_3.sav

....

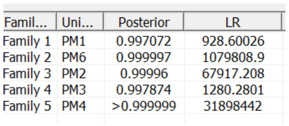
....

PM8

Family 5

File: Exercise3_3am5.txt

File: Exercise3_3pm.txt



Famil...	Uni...	Posterior	LR
Family 1	PM1	0.997072	928.60026
Family 2	PM6	0.999997	1079808.9
Family 3	PM2	0.99996	67917.208
Family 4	PM3	0.997874	1280.2801
Family 5	PM4	>0.999999	31898442

Figure 2: Pedigree for Exercise 3.3. The first five markers are shown.

4 Solutions Exercises 2.9 and 2.17

4.1 Exercise 2.9

Regarding b): The marker with 0 LR, Penta_E is most easily found using `View result`. Regarding c), $LR = 4421152$, d) $LR(H_1/H_3) = 1.39$ (answers differ if mutations are only modelled for Penta_E). There is also solution file, `Solutions_2_9.fam` available. Regarding the last question, there is no consensus. One can argue that a model should be formulated before calculations and then appropriate mutation models should be specified for all markers. On the other hand, introducing mutations complicates calculations and this is a problem if it is desired to verify by hand. This is discussed at greater length in the Section 2.4.4 "Dealing with mutations in practice".

4.2 Exercise 2.17

- Simulation: In `Pedigrees` click `Simulate`. Move both AF and Child to `Will be genotyped`. The simulation will produce slightly different results each time it is run unless a seed is set. If you untick `random seed` and set seed to 12345, you should get the same results as below. Click `Simulate`. The mean LR is shown for both H_1 true and H_2 true.

  a) The mean LR when H1 is true is 40.86.

  b) The mean LR when H2 is true is 0.8979.

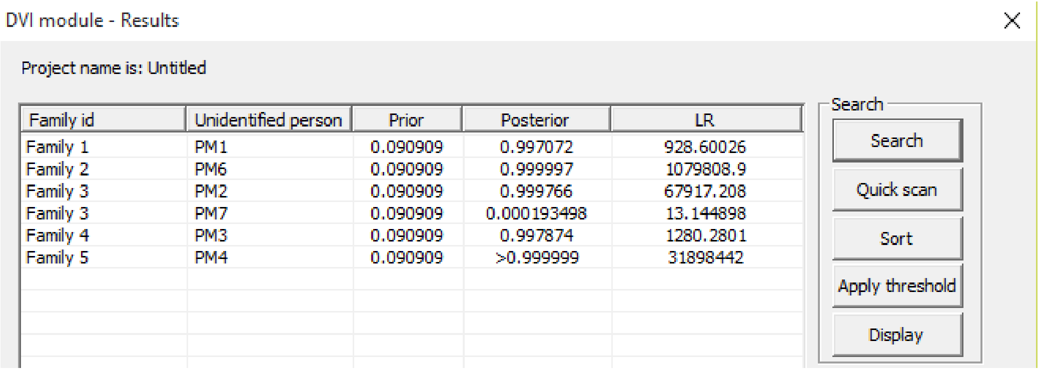
  c) Click `LR limit`, choose LR threshold 50 and click update. The probability of observing a LR larger than 50 is 0.09.

5 Solution Exercise 3.3

- c) This may be a realistic scenario for different reasons. A simple reason may be that not all missing persons have been found. Another may be that not all remains produce DNA profiles. The fact that only 8 profiles is in the set even though the total number of missing persons is greater is accounted for in the prior.

  d) This means we have some prior belief that the number of missing persons is 10.

  l) For results see Figure 3.



Family id	Unidentified person	Prior	Posterior	LR
Family 1	PM1	0.090909	0.997072	928.60026
Family 2	PM6	0.090909	0.999997	1079808.9
Family 3	PM2	0.090909	0.999766	67917.208
Family 3	PM7	0.090909	0.000193498	13.144898
Family 4	PM3	0.090909	0.997874	1280.2801
Family 5	PM4	0.090909	>0.999999	31898442

Figure 3: List of results from the DVI search for Exercise 3.3 l)

m) Not all remains where identified, this is expected as we only have reference data from 5 families. All posterior probabilities are above 99% (except for the match between Family 3 and PM7) though only three are greater than 99.99%.

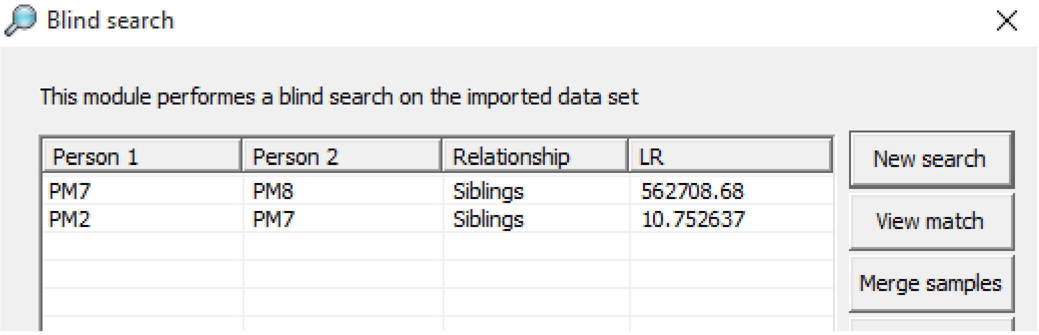
n) The user will find a possible mutation for the match between Family 4 and PM3 for the marker vWA.

p) The posterior becomes considerably lower as the priors are lowered.

o) For results see Figure 4. We see that PM7 and PM8 has a possible sibling relation. If either of the two persons match in a family we may use this information to match both into that family. We may also combine this information with meta data such as known relationship between missing persons.

q) * One way is to add another pedigree in the reference family. Another solution may be to add another reference family with the same reference person. The difference would be how the posteriors are calculated. We will use the first option, i.e., add another pedigree to Family 1, where we now need to define extra persons in order to define the brother relationship.

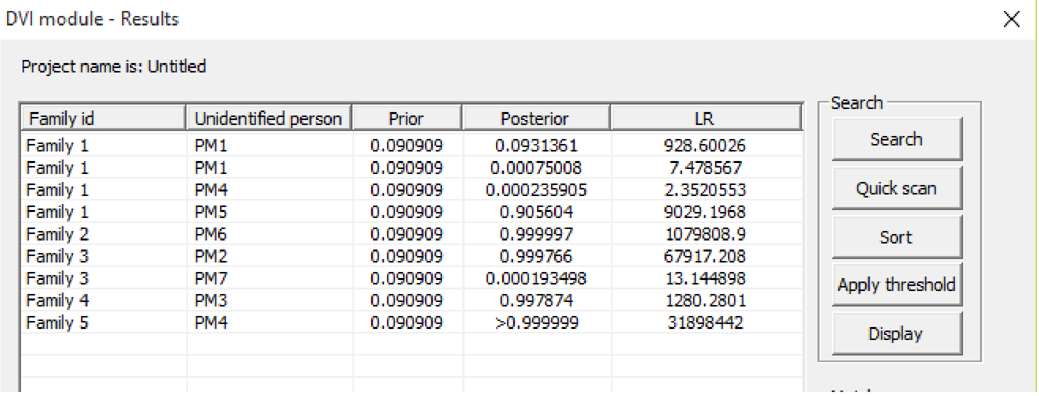
r) * For results see Figure 5. We see that we now have 4 possible matches for Family 1, where three is with the Brother pedigree. We see that PM5 has the highest LR in the Brother pedigree. We also see that the posterior probabilities are spread out between the



This module performes a blind search on the imported data set

Person 1	Person 2	Relationship	LR
PM7	PM8	Siblings	562708.68
PM2	PM7	Siblings	10.752637

Figure 4: Results for Exercise 3.3 o).



Project name is: Untitled

Family id	Unidentified person	Prior	Posterior	LR
Family 1	PM1	0.090909	0.0931361	928.60026
Family 1	PM1	0.090909	0.00075008	7.478567
Family 1	PM4	0.090909	0.000235905	2.3520553
Family 1	PM5	0.090909	0.905604	9029.1968
Family 2	PM6	0.090909	0.999997	1079808.9
Family 3	PM2	0.090909	0.999766	67917.208
Family 3	PM7	0.090909	0.000193498	13.144898
Family 4	PM3	0.090909	0.997874	1280.2801
Family 5	PM4	0.090909	>0.999999	31898442

Figure 5: Results for Exercise 3.3 r).

matches for Family 1, thus considerably lowering the probabilities for the match against PM1.

s) A better solution, but more complex, would be to allow the definition of several missing persons in the same pedigree. Familias would then either search for each missing persons individually, or try matching all unidentified persons with the missing persons at once. The complexity using the latter approach grows exponentially with the number of missing persons.

A Familias tutorial

This tutorial[1] supplements and refers to the Elsevier-book "Relationship Inference with Familias and R. Statistical Methods in Forensic Genetics", by Egeland, Kling and Mostad, and is based on the windows program `Familias` 3.1.9 or later versions; as of Oct 23, 2016, the version is 3.1.9.6, referred to as `Familias`. There are some references to the mentioned book in this tutorial. However, these references are not necessary for the understanding of the tutorial. The purpose to explain briefly the *basic functionality* of the program, a complete description is provided in the manual available from `http://familias.no`.

The tutorial starts by discussing a standard paternity case. Then, we address the most important complicating factors: mutation, theta-correction and silent alleles. Finally, we present a more complicated example where more than two alternatives are considered. Some relevant papers include [3][2], and [1][3].

Four basic steps

There are four basic steps involved in a typical application of the program as illustrated in Figure 6. These steps suffice to perform the calculations for standard paternity cases. Below these steps are detailed for the paternity case summarised in Figure 7.

1. `General DNA data` window, Figure 8. Click `Add` to enter a marker. In the new window, enter `Marker1` and the two alleles `A` and `B`, both with frequencies 0.05. Enter the `C` allele with frequency 0.9. Press `Save`.

2. `Persons` window, Figure 9. Enter the persons: `AF` (alleged father), `Mother`, and `CH` (child) and their gender. Close window (this should generally be done before continuing).

3. `Case DNA data` window, Figure 10. Double-click each person to enter his or her DNA data as given Figure 7. In the new window, enter the

[1] Available from `http://familias.name/book.html`

[2] `Familias 3` reference: Kling et al. "Familias 3–Extensions and new functionality". FSI: Genetics, 13:12-127, 2014

[3] Drábek. "Validation of software for calculating the likelihood ratio for parentage and kinship". FSI: Genetics, 3:112-118, 2009

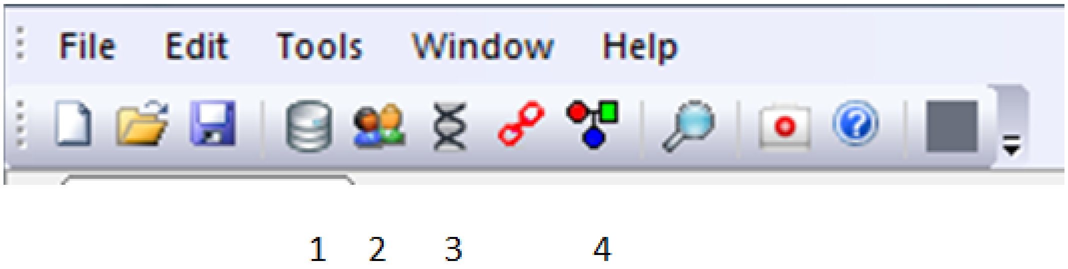


Figure 6: The basic windows of `Familias`.

appropriate allele system (use the pull-down menu) and the observed alleles for this person, then press `Add` and `OK`.

4. `Pedigrees` window, Figure 11. Click `Add` to enter the pedigree corresponding to hypothesis H1 (paternity). Enter `H1: AF father` as `Pedigree name`. Enter the `Mother` as the parent of `CH` in the pull-down menu and click `Add`. Similarly, enter `AF` as the parent of `CH`. Click `OK` to finish the definition of the pedigree corresponding to hypothesis H1. Click `Add` in the pedigree window once more to add the pedigree corresponding to hypothesis H2. Enter `H2: Unrelated` as `Pedigree name`. Press `Calculate`. Normally one would answer `Yes` when asked to save.

The output is shown in Figure 12, page 18. The $LR = 20$ as it should according to the equation $LR = 1/p_A = 1/0.05 = 20$. Furthermore, a prior probability of 0.5 for each alternative, gives the posterior as $LR/(LR+1) = 20/21 = 0.952381$, also indicated in the output window.

Advanced software is not required for the simple paternity case considered so far. However, mutation, theta corrections and silent alleles complicate matters as described below.

Specific mutation models

The default value for mutation rates is zero. However, if it is known or reasons to suspect that there is a non-zero mutation rate, it should be specified. A reasonable mutation rate could be around 0.005. The program offers the

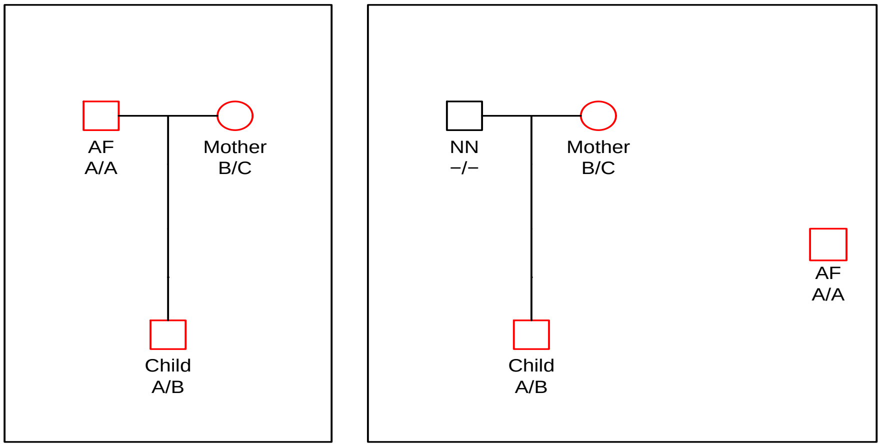
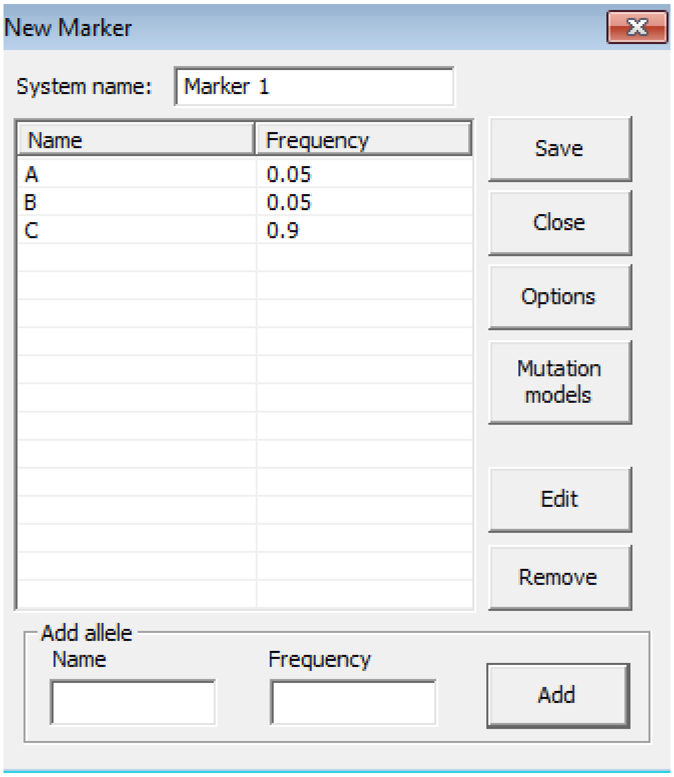


Figure 7: A standard paternity case. Left panel corresponds to hypothesis H_1, AF being the father while AF is unrelated, H_2, to the right.

possibility to distinguish between male and female mutation rates. The reason for this is that paternal alleles tend to mutate more often than maternal alleles. There are 5 different mutation models to choose from as shown in Figure 13:

1. `Equal probability (Simple)`
2. `Proportional to freq.`
3. `Stepwise (Unstationary)`
4. `Stepwise (Stationary)`
5. `Extended stepwise`

If a model is stationary this implies that adding irrelevant persons will not affect the result. Conversely, for unstable models adding irrelevant persons may lead to slightly different results. Stationarity is not a natural biological condition, as allele frequencies do change over time. However, non-stationarity has the somewhat unpleasant consequence that the exact LR will change by including extra irrelevant persons in the calculations. Furthermore, a person's allele frequencies will be different if they are derived directly from the data base compared to if they are derived, with mutations, from parents



New Marker

System name: Marker 1

Name	Frequency
A	0.05
B	0.05
C	0.9

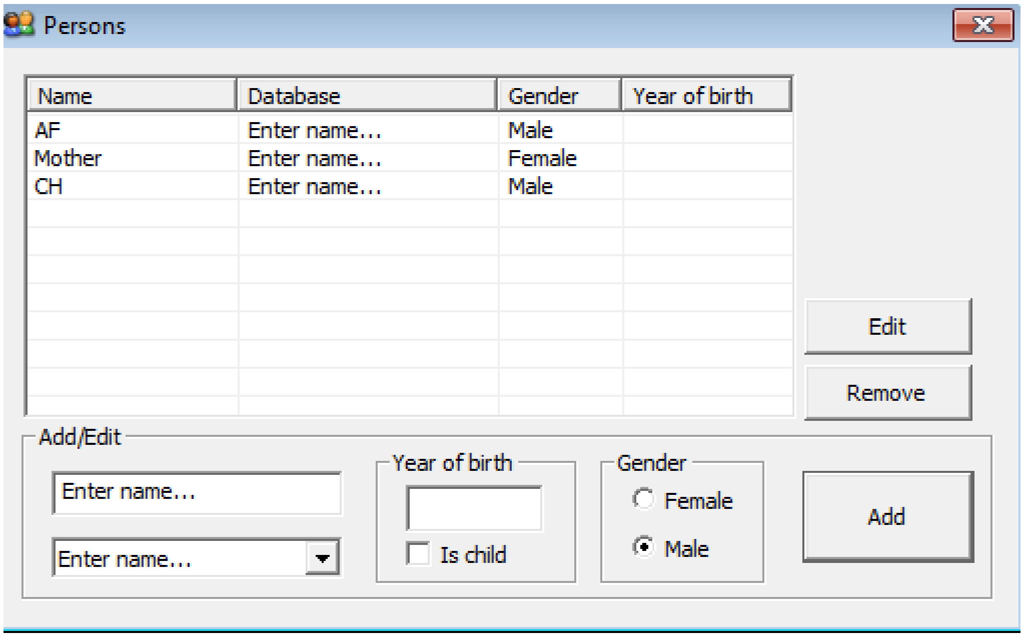
Figure 8: Defining a marker.

having the database allele frequencies. Models 2 and 4 above are stationary.

Exercise 2.7 and 2.8, available from `http://familias.name/book.html`, exemplify the above mutation models. The alleged father is 14/15 and the child 16/17. Without a model for mutations, the likelihood ratio would be 0. Using Model 1, `Equal probability (Simple)`, all mutations are equally likely. With reasonable parameter choices for Models 3, 4, and 5, the shortest mutation, the one from 15 to 16, is the more likely.

Theta correction

Deviation from Hardy Weinberg Equilibrium is the simplest case where the so called theta correction is needed. The input is illustrated in Figure 14.



Name	Database	Gender	Year of birth
AF	Enter name...	Male	
Mother	Enter name...	Female	
CH	Enter name...	Male	

Enter name...

Enter name...

Year of birth: □ Is child

Gender: ○ Female ● Male

Figure 9: Defining persons.

Silent alleles

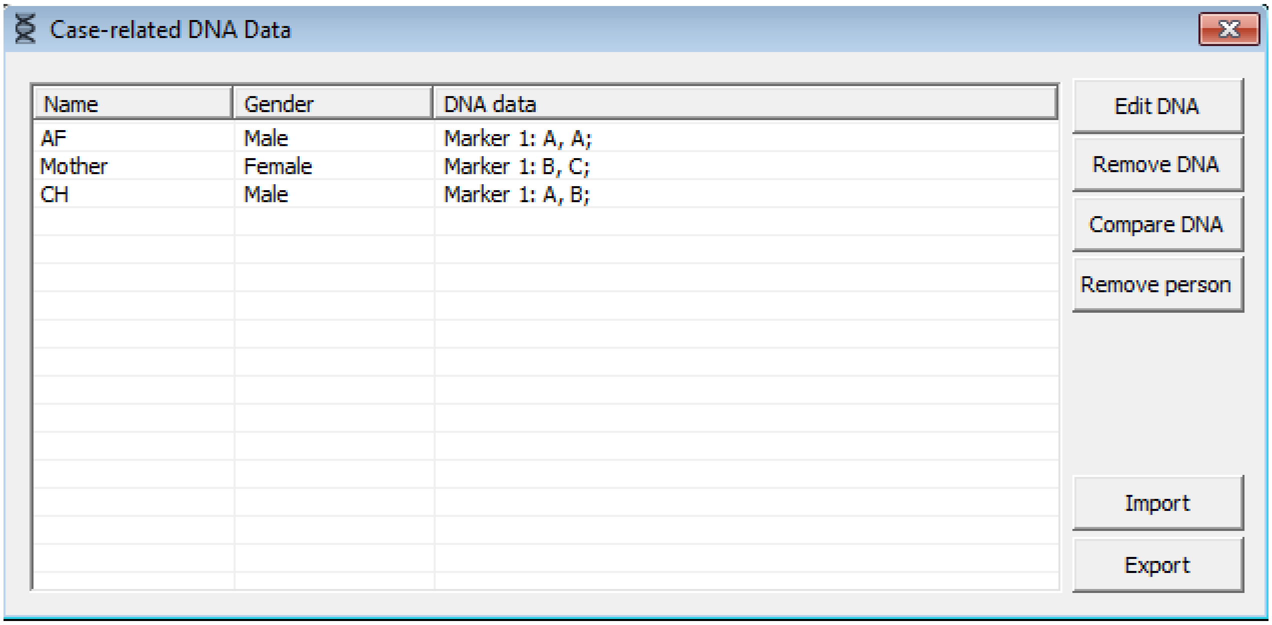
Silent alleles may be present when some homozygotes are observed. Figure 15 demonstrates the input. Note that both alleles need to be provided in the `Case DNA data` window.

Example. Brothers?

A woman M has 3 sons S1, S2, and S3, and the question is if a putative father PF is the father of all, some, or none of these sons. DNA data is available for S1, S2 and S3.[4] Data from 8 loci is given. In all loci, all alleles have frequency 0.05. The alleles are numbered 1, 2, 3, 4. With this notation, S1, S2, and S3 have the observations given in Table 1.

Note that `Familias` contains functions for automatic generation of sets of pedigrees. This may be useful in situations when a large number of pedigrees should be considered possible. In the example above, clicking the button

[4] The file `http://familias.name/TutorialBrothers.fam` is available for those who would like to skip manual input.



Case-related DNA Data

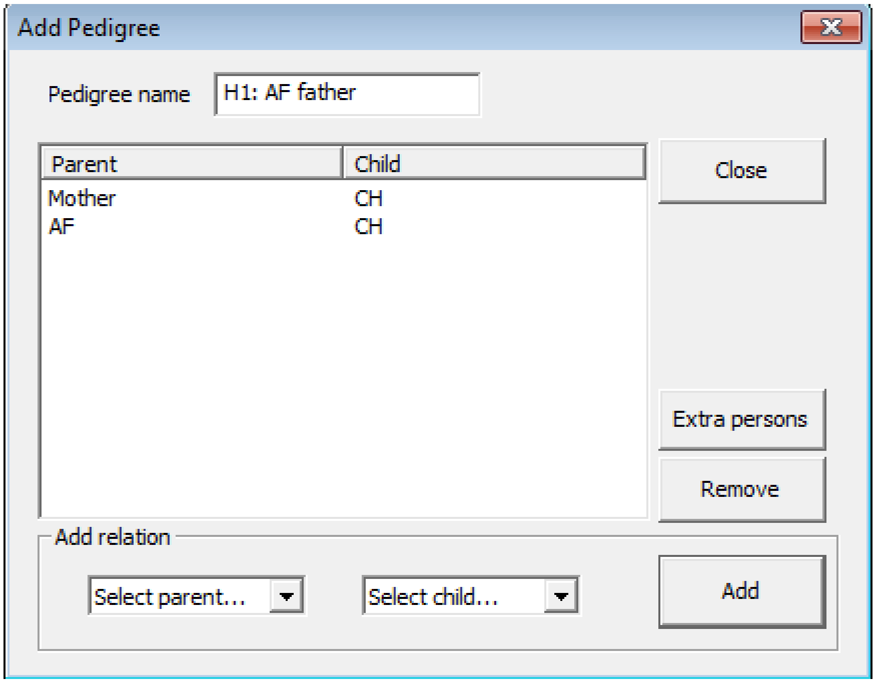
Name	Gender	DNA data
AF	Male	Marker 1: A, A;
Mother	Female	Marker 1: B, C;
CH	Male	Marker 1: A, B;

Figure 10: Genotype data.

Locus	sys1	sys2	sys3	sys4	sys5	sys6	sys7	sys8
S1	1/2	1/2	1/2	1/2	1/2	1/1	1/2	1/2
S2	3/4	3/3	3/4	3/4	3/4	1/2	1/2	2/3
S3	3/4	1/2	1/2	3/4	3/3	3/3	3/4	3/4

Table 1: Genotype data for brother example.

Generate (and keeping the default settings) will generate a total of 8 pedigrees provided S1, S2 and S3 are defined as children, M and PF are given the same birth data and `M` is fixed to be the mother of S1, S2 and S3. The results are given in Figure 17. We are using by default a flat prior of $1/8 = 0.125$. Observe that pedigree8, the full brother alternative gives a likelihood (and hence a posterior) of 0. To understand this consider sys5 in Table 1. S3 is homozygous 3/3. This implies that the two other can display at most two alleles different from 3. However, they have three alleles, 1,2 and 4. The alternative specifying S2 and S3 as full brothers and half brother of S1, is the most likely. This pedigree appears as `Ped4` in Figure 16, The posterior



Pedigree name: H1: AF father

Parent	Child
Mother	CH
AF	CH

Figure 11: Pedigree definition of trio.

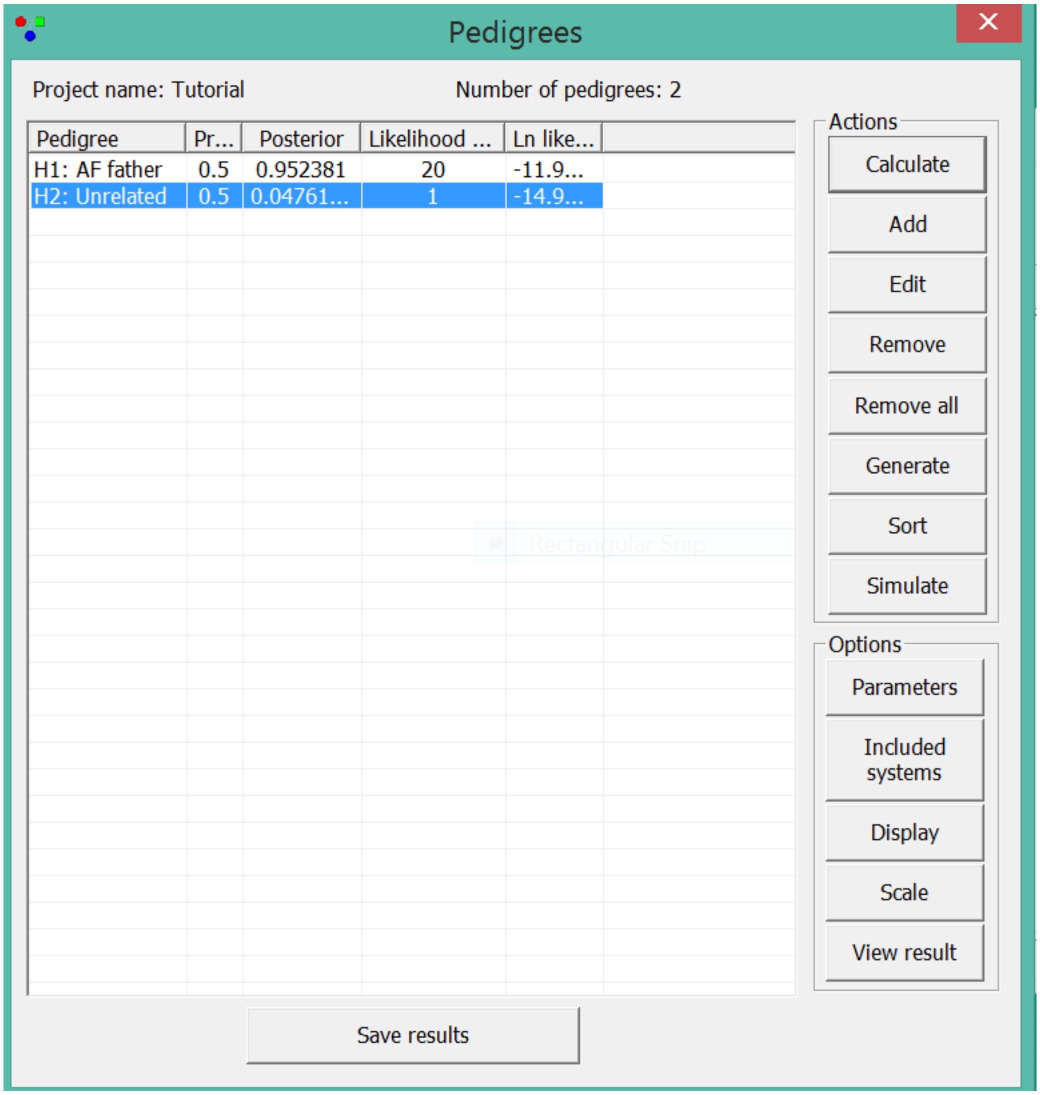
probability 0.6245 from `Familias` can be confirmed by

$$\frac{\exp(-120.2663)}{4\exp(-122.6019)+\exp(-120.2663)+\exp(-122.058)+\exp(-123.3108))}$$

where the numerator is the likelihood for `Ped 4` and the denominator the sum of the likelihoods. Rather, than reporting the posterior probability, we can obviously report more conventional LR-s. Then a choice of reference, a denominator, need to be decided on, and several values have to be reported and this may be inconvenient.

Brother example continued

A stepwise stationary mutation model with mutation rate 0.005 and range 0.1 is used for all markers. As can be seen from Figure 18, the results are now completely changed. The full brother alternative now comes out as the by far most likely alternative. This makes sense intuitively as there are several marker where pairs of individuals share both alleles.

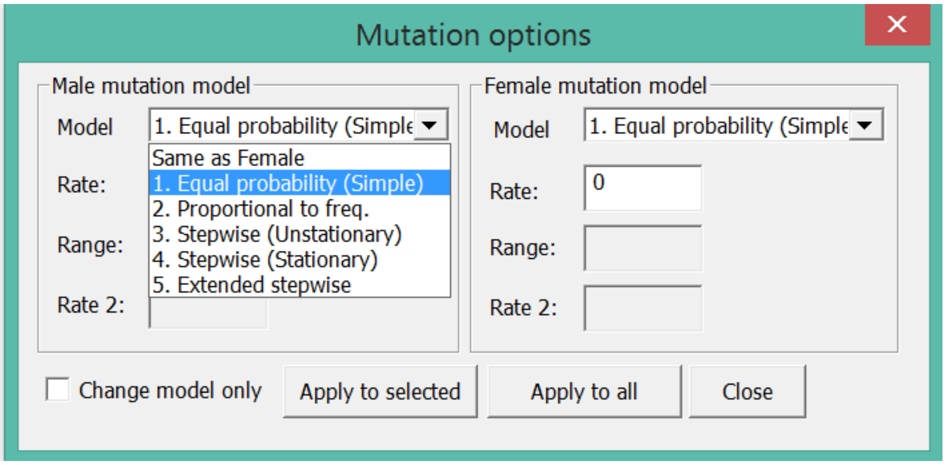


Project name: Tutorial

Number of pedigrees: 2

Pedigree	Pr...	Posterior	Likelihood ...	Ln like...
H1: AF father	0.5	0.952381	20	-11.9...
H2: Unrelated	0.5	0.04761...	1	-14.9...

Figure 12: `Familias` output.



Mutation options

Male mutation model

Model 1. Equal probability (Simple

Rate:

Range:

Rate 2:

Female mutation model

Model 1. Equal probability (Simple

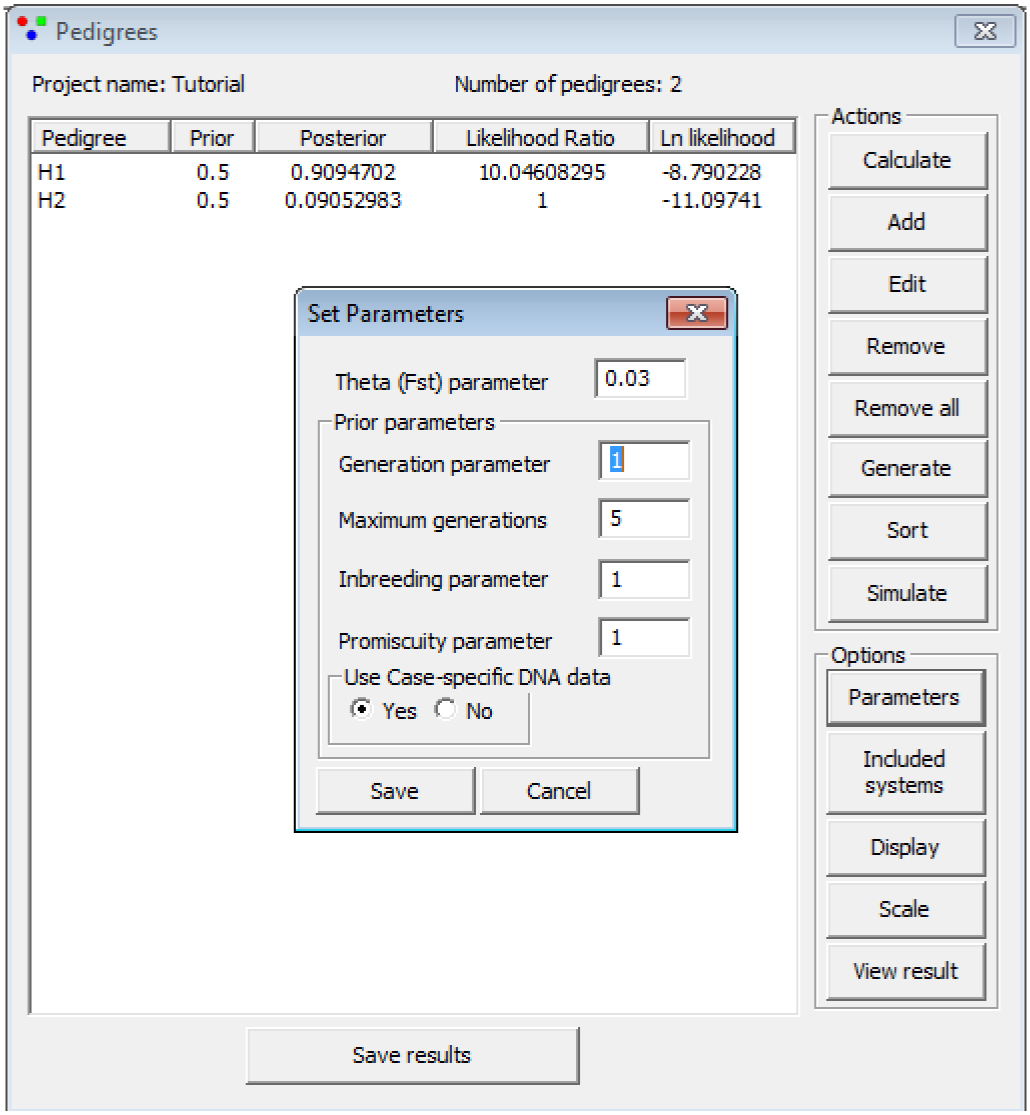
Rate: 0

Range:

Rate 2:

□ Change model only

Figure 13: Mutation models



Project name: Tutorial

Number of pedigrees: 2

Pedigree	Prior	Posterior	Likelihood Ratio	Ln likelihood
H1	0.5	0.9094702	10.04608295	-8.790228
H2	0.5	0.09052983	1	-11.09741

Theta (Fst) parameter: 0.03

Generation parameter: 1

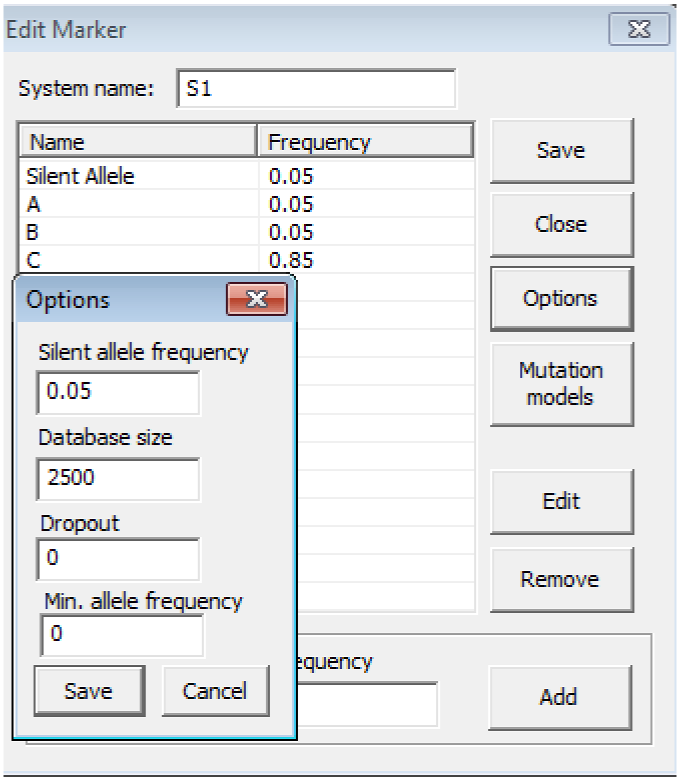
Maximum generations: 5

Inbreeding parameter: 1

Promiscuity parameter: 1

Use Case-specific DNA data: ● Yes ○ No

Figure 14: Press `Parameter` in the `Pedigrees` window to provide a θ value, 0.03 in the example. The LR is reduced from 20 to 10.05. The remaining parameters of this window are normally left unchanged. The default values correspond to a flat prior and changing them only affects the posterior probability, not LR.



Edit Marker

System name: S1

Name	Frequency
Silent Allele	0.05
A	0.05
B	0.05
C	0.85

Options

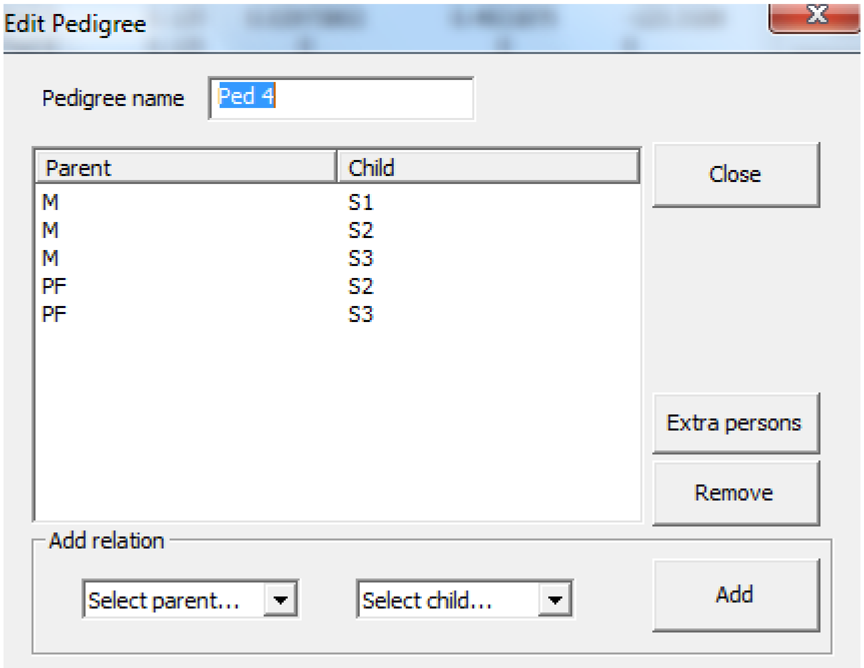
Silent allele frequency: 0.05

Database size: 2500

Dropout: 0

Min. allele frequency: 0

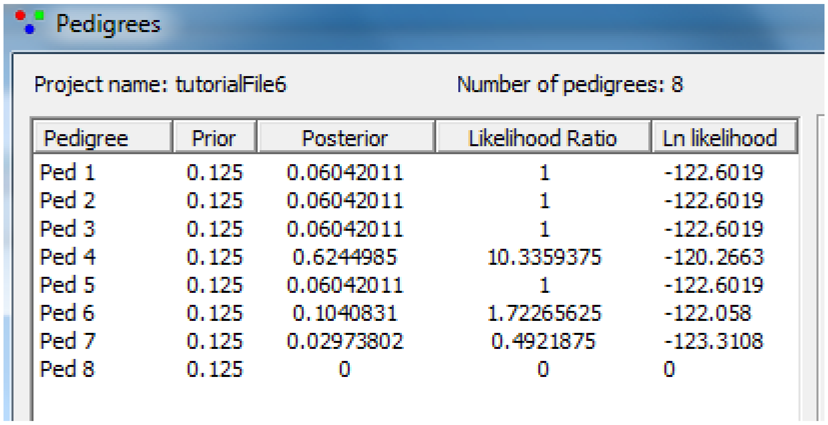
Figure 15: Press `Options` in the `Edit Marker` window to provide a silent allele frequency, 0.05, in the example. Note that allele frequencies including the slient allele must sum to 1. This can be achieved by reducing the frequency of the `C` allele to 0.85. The LR is reduced from 20 to 13.33.



Pedigree name: Ped 4

Parent	Child
M	S1
M	S2
M	S3
PF	S2
PF	S3

Figure 16: 8-pedigree example. The most likely pedigree.

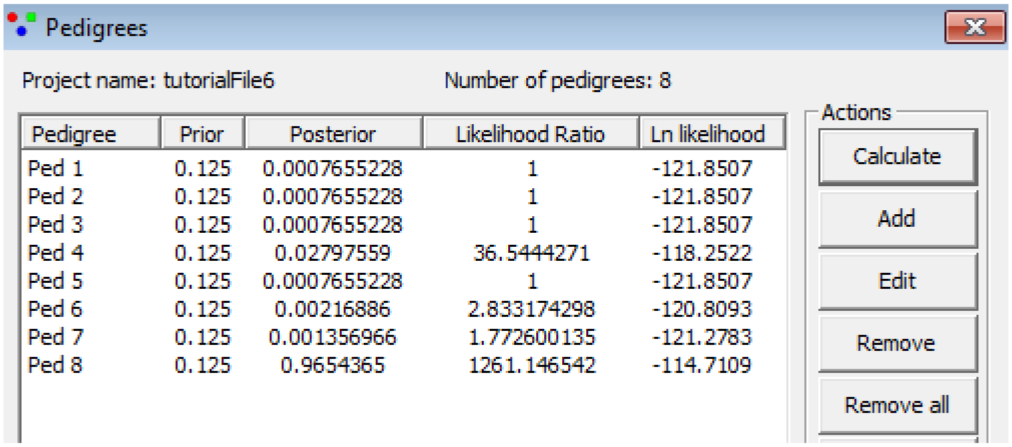


Project name: tutorialFile6

Number of pedigrees: 8

Pedigree	Prior	Posterior	Likelihood Ratio	Ln likelihood
Ped 1	0.125	0.06042011	1	-122.6019
Ped 2	0.125	0.06042011	1	-122.6019
Ped 3	0.125	0.06042011	1	-122.6019
Ped 4	0.125	0.6244985	10.3359375	-120.2663
Ped 5	0.125	0.06042011	1	-122.6019
Ped 6	0.125	0.1040831	1.72265625	-122.058
Ped 7	0.125	0.02973802	0.4921875	-123.3108
Ped 8	0.125	0	0	0

Figure 17: `Familias` output for 8-pedigree example.



Project name: tutorialFile6

Number of pedigrees: 8

Pedigree	Prior	Posterior	Likelihood Ratio	Ln likelihood
Ped 1	0.125	0.0007655228	1	-121.8507
Ped 2	0.125	0.0007655228	1	-121.8507
Ped 3	0.125	0.0007655228	1	-121.8507
Ped 4	0.125	0.02797559	36.5444271	-118.2522
Ped 5	0.125	0.0007655228	1	-121.8507
Ped 6	0.125	0.00216886	2.833174298	-120.8093
Ped 7	0.125	0.001356966	1.772600135	-121.2783
Ped 8	0.125	0.9654365	1261.146542	-114.7109

Figure 18: 8-pedigree example with mutation model.

References

[1] J. Drábek. Validation of software for calculating the likelihood ratio for parentage and kinship. *Forensic Science International: Genetics*, 3(2):112–118, 2009.

[2] T. Egeland, L. Kling, and P. Mostad. *Relationship Inference with Familias and R. Statistical Methods in Forensic Genetics.* Elsevier, 2015.

[3] D. Kling, A. O. Tillmar, and T. Egeland. Familias 3–extensions and new functionality. *Forensic Science International: Genetics*, 13:121–127, 2014.