Relationship Inference with Familias and R. Statistical Methods in Forensic Genetics

Exercise Chapter 2,3,4

Thore Egeland
Daniel Kling
Petter Mostad

April 28, 2016

2.13 Exercises

Solutions and input files can be found at `http://familias.name`.

Exercise 2.1 (Simple paternity case. Video)**.**

The purpose of this exercise is to illustrate the basic concepts based on a simple paternity case. Figure 2.6 shows a mother (undisputed), an alleged father (AF) and a child.

***** Insert Figure 2.6 around here *****

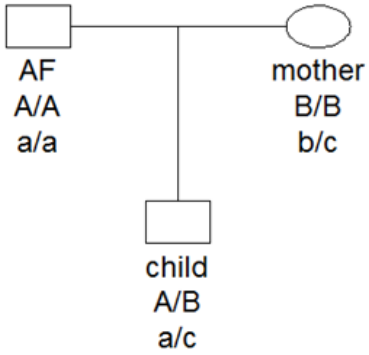

Figure 2.6: A standard paternity case with genotypes for two markers, see Exercise 2.1.

We consider

H_1**:** The alleged father (AF) is the real father.

H_2**:** The alleged father and the child are unrelated.

The mother is undisputed.

a) Consider first only one autosomal locus, called S1, with alleles A, B and C, as displayed in Figure 2.6. The allele frequencies are $p_A = p_B = 0.05$ and $p_C = 0.9$. Explain why the likelihood ratio is $LR = 1/p_A$. How do you interpret the LR?

b) Calculate the LR using `Familias`.

c) There is a second autosomal locus, called S2, with alleles a, b, c and d with allele frequencies 0.1, 0.1, 0.1 and 0.7, respectively. Calculate the likelihood ratio for this marker and also for the two first markers combined using `Familias`.

d) It can be shown that the likelihood ratio for two first markers is $(1/p_A) \times (1/p_B)$. Use this to verify the `Familias` answer calculated above.

e) Generate a report clicking `Save results` with options `Only report, Rtf report, Complete`. The report includes all input and all output. Check that the report file is correct and contains sufficient information to reproduce the calculations. In particular check that the LR for markers S1 and S2 and the combined likelihood.

f) Save the `Familias` file (we suggest the file extension fam). Exit `Familias`.

g) Start `Familias` and read the previously saved file. RMP abbreviates the random match probability. Calculate 1/RMP for AF by hand and in `Familias`. *Hint*: Mark AF in the `Case-Related DNA data` window and press `Compare`.

h) We next consider theta (θ) correction. For simplicity we will only use the first marker, S1. The θ parameter is called kinship parameter in `Familias` and is set using the `Options` in the `Pedigrees` window. Set the kinship parameter to 0.02. Calculate the LR for the first marker S1. To get calculations for selected markers only, in this case S1, use the `Included systems` button. Check that your answer coincide with the following theoretical result

$$LR = \frac{1 + 3\theta}{2\theta + (1 - \theta)p_A}. \tag{2.17}$$

i) Discuss the assumptions underlying the calculations of this exercise.

Exercise 2.2 (Simple paternity case with mutation)**.**

We consider a motherless paternity case, see Figure 2.7, with one marker, VWA.

***** Insert Figure 2.7 around here *****

The allele frequencies are given in Table 2.8. The alleged father is 14/15, the child 16/17 while the hypotheses are as in Exercise 2.1.

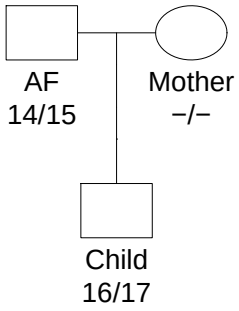

Figure 2.7: A paternity case with possible mutation.

Allele	Frequency
14	0.072
15	0.082
16	0.212
17	0.292
18	0.222
19	0.097
20	0.02
21	0.003

Table 2.8: Allele frequencies for Exercise 2.2

a) Explain why $LR = 0$. Confirm this answer using `Familias`.

b) Use the mutation model "Equal probability" with mutation rate $R = 0.007$ for both males females and calculate LR.

c) It can be shown (see Exercise 2.8) that

$$LR = \frac{m(p_{16} + p_{17})}{2p_{16}p_{17}}. \tag{2.18}$$

Use this formula to confirm the `Familias` calculation. Obtain m, the probability of mutating to one specific allele, using `File > Advanced options`. Explain the difference between R and m.

Exercise 2.3 (Missing sister)**.**

The College of American Pathologists (CAP) has several proficiency testing programs targeted to laboratories that perform DNA typing of STR loci. The below is a test from 2011: Hikers come across human skeletal remains in a forest. Evidence around the site provides a clue as to the identity of the individual. You are asked to test a bone to determine if the individual (bone) is related to an alleged mother (AM) and the mother's other daughter, the alleged full sister (AS,) see Figure 2.8.

***** Insert Figure 2.8 around here *****

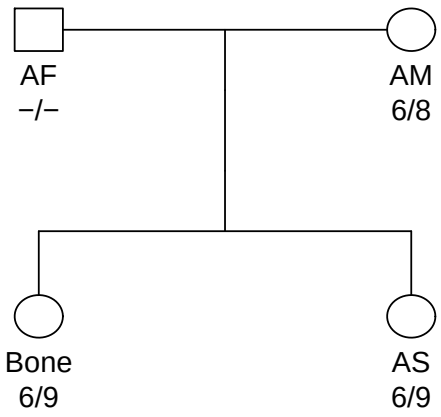

Figure 2.8: The case of the missing sister. The alleles (allele frequencies) are 6 (0.086), 8 (0.152) and 9 (0.328) for this marker (F13B).

The hypotheses are

H_1 : The bone belonged to the daughter of AM and sister of AS.

H_2 : The bone belonged to someone unrelated to AM and AS.

a) Enter the data manually and calculate the LR. The genotypes and allele frequencies for this marker (F13B) are given in Figure 2.8.

b) Read input from the file `Exercise2_3.fam`. Calculate the LR based on all markers.

c) Find the LR–s of the individual markers. Check that the answer for F13B corresponds to the one you found in problem a) above.

d) One of the markers, D7S820, gives a very large LR, namely 11189. What is the reason for this large LR? What is the combined LR if marker D7S820 is removed?

Exercise 2.4 (Grandfather - grandchild)**.**

Two individuals, GF and GS, are submitted to the laboratory for testing. The alternatives are

H_1 : GF is the grandfather of GS.

H_2 : The individuals GF and GS are unrelated.

Figure 2.9 shows the pedigree corresponding to H_1 for the first marker D3S1358.

***** Insert Figure 2.9 around here *****

a) Enter the data given in Figure 2.9 for D3S1358 manually into `Familias` and calculate the LR for the first marker shown in Figure 2.9.

b) Calculate the LR based on all markers. Read input from the file `Exercise2_4fam`.

c) Formulate a conclusion. In the CAP exercise it was stated that GF and GS share the same Y-haplotype and that the frequency of this haplotype is 0.0025. Can this information be used?

Exercise 2.5 (Simple paternity case. Probability of paternity)**.**

We revisit Exercise 2.1. Rather than calculating the LR we will now calculate the Essen-Möller index W defined as the probability of H_1 conditional on the genotypic data. Assume a priori that the hypotheses H_1 and H_2 are equally likely. Then, it can be shown that

$$W = Pr(H_1|\text{data}) = \frac{LR}{LR+1}. \tag{2.19}$$

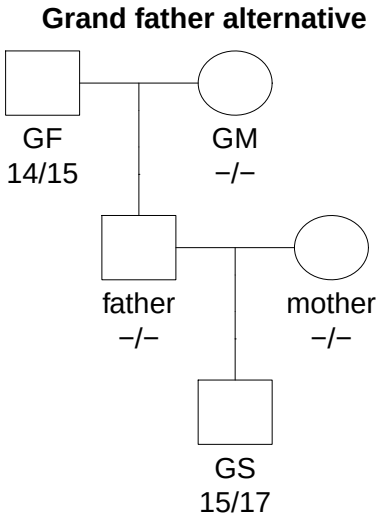

Figure 2.9: Grandfather-grandchild. The alleles (frequencies) for D3S1358 are: 14 (0.122), 15 (0.258) and 17 (0.197).

a) Recall that $LR = 20$ for the first marker. Calculate W.

b) Recall that $LR = 200$ for two markers. Calculate W.

c) Use `Familias` with `Exercise2_1.fam` to calculate W for the two above cases.

d) Do you prefer LR or W?

Exercise 2.6 (Inbreeding. LR. Several alternatives)**.**

Consider the following hypotheses

H_1 : AF, the father of mother (undisputed), is the father also of her child.

H_2 : An unrelated man is the father of the child.

Figure 2.10 shows the pedigree corresponding to hypothesis H_1. The allele frequencies are $p_1 = p_2 = p_3 = 0.05$.

***** Insert Figure 2.10 around here *****

***** Insert Figure 2.11 around here *****

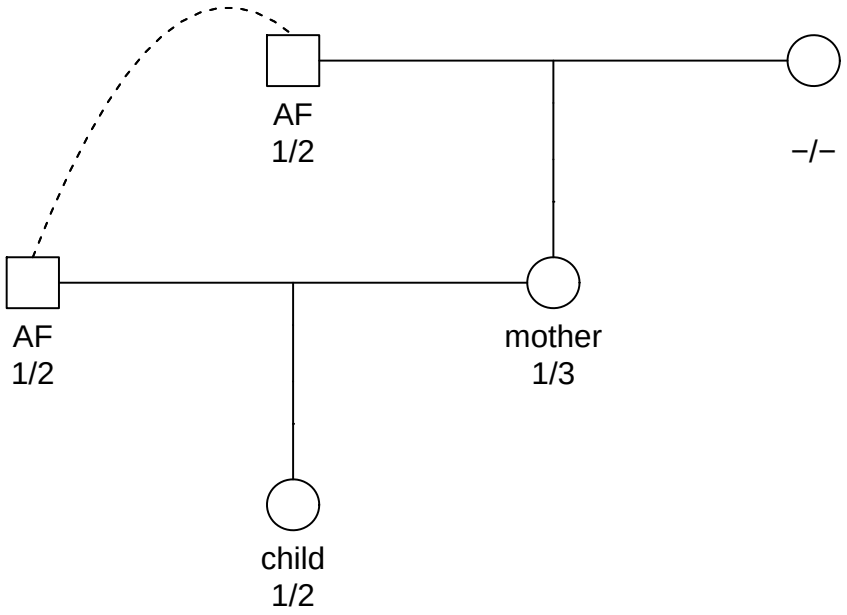

Figure 2.10: Incest by father. The stapled line indicates that AF appear in two roles. See Exercise 2.6.

a) Use `Familias` to calculate the LR based on the genotypes given in Figure 2.10. (You are encouraged to enter the data manually, but there is an input file `Exercise2_6.fam` available.) Does the incest influence the resulting LR in this particular case?

   The defense claims that one should rather consider the hypothesis

   H_3 : The brother of mother is the father of the child. See Figure 2.11.

   The LR can be calculated in several ways depending on the choice of the reference. Calculate $LR(H_1/H_2)$ and $LR(H_1/H_3)$.

b) When there are more than two hypotheses, as above, some prefer to rather calculate posterior probabilities for the hypotheses as there is then no need to define a reference (or denominator) pedigree. Assume that each

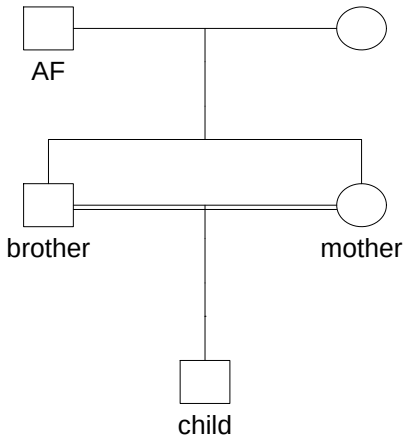

Figure 2.11: Incest by brother. A standard way of displaying incest is used in the figure. See Exercise 2.6.

of the three hypotheses are equally likely a priori. Calculate the posterior probabilities.

Exercise 2.7 (Several mutation models)**.**

In this exercise, which extends on Exercise 2.2, we will try the different mutation models. The models are described in detail in Section A.1.4 of the manual available from `http://familias.no`. Throughout we consider data from the system VWA given in Exercise 2.2. The alleged father is 14/15 and the child 16/17 while the hypotheses are

H_1 : The alleged father is the true father.

H_2 : The alleged father and child unrelated.

There are five different mutation models in `Familias`. In this exercise all will be tried. We will use overall mutation rate $R = 0.005$ and the same model for females and males. The answer for this exercise will be obtained using `Familias`; Exercise 2.8, on the other hand, is based on theoretical calculations.

a) Load the data `Exercise2_7.fam`. Use the `Equal probability` model. Calculate the LR. (Answer: $LR = 2.9e-03 = 0.0029$. *comment*: For this model mutations to all other alleles are equally likely).

b) Use the `Proportional to freq.` model. Calculate the LR. (Answer: $LR = 6.3e-03 = 0.0063$. *Comment*: For this model it is more likely to mutate to a common allele compared to a rare).

c) Use the `Stepwise (Unstationary)` model. For this model there are two parameters. The first is as before and should be set to $R = 0.005$. Set the second parameter, `Range` to $r = 0.5$. Calculate the LR. (Answer: $LR = 4.7e-03 = 0.0047$. *Comment*: For this model, mutation depends on the size of the mutation. With $r = 0.5$, a two step mutation occurs with half the probability of a one-step mutation. A three step mutation occurs with half the probability of a two step mutation and so on. The value $r = 0.1$ may be more realistic.)

d) Use the `Stepwise (Unstationary)` model with parameters as above. (Answer: $LR = 6.4e-03 = 0.0064$.)

e) Use `Extended stepwise` model with last parameter 0.1. In this case, with no microvariants, the results is the same as for `Stepwise(Unstationary)` regardless of values of last parameter.

Comment: The models `Proportional to freq.` and `Stepwise (Stationary)` are *stationary*, the others are not. If a model is stationary, introducing a new untyped person, say the father of the alleged father, does not change the LR. This is a reasonable property of a model as introducing irrelevant information should not change the result. For an unstationary model, however, the LR will change slightly as allele frequencies may then differ from one generation to the next. While this may appear mathematically inconsistent, it could be argued that allele frequencies change, i.e., a stationary distribution has not been reached.

f) Verify by means of an example that the LR does not change if a father of the alleged father is introduced for the stationary models, while slight changes occur for the two other models.

g) *Comment*: There is an other subtle point of all mutation models: In this case only five alleles (14, 15, 16, 17 and "Rest allele") are needed rather

than the 8 alleles defined. A five allele model will lead to slightly changing LR-s. Verify the above and comment.

Exercise 2.8 (*Mutation models, theoretical)**.**

In this exercise, which serves to confirm the answers obtained in Exercise 2.7, we will fill in some mathematical details related to the mutation models.

a) Show that

$$LR = \frac{p_{16}(m_{14,17} + m_{15,17}) + p_{17}(m_{14,16} + m_{15,16})}{4p_{16}p_{17}} \tag{2.20}$$

where p_{16} is the allele frequency for allele 16 and $m_{14,17}$ the probability of a mutation from 14 to 17 and so on; the formula is a special case of (2.12).

b) For the "Equal probability" model $m_{ij} = m = R/(n-1)$, where n is the number of alleles. Explain why Equation (2.18) in Exercise 2.2 follows from (2.20). Show that when $n = 8$ and $R = 0.005$, $LR = 0.0029$, as in Exercise 2.7b).

c) Consider next the proportional model. By definition,

$$m_{ij} = kp_j \text{ for } i \neq j, \text{ and } m_{ii} = 1 - k(1 - p_i).$$

Show that

$$LR = k = \frac{R}{\sum_{i=1}^{n} p_i(1 - p_i)}.$$

and from this verify the answer of the previous exercise.

Comment: Exact calculations for remaining mutation models are more technical. Section A.1.4 of the manual for `Familias 3` `http://familias.no/` contains some further examples.

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

Load the file `Exercise2_9.fam`. Consider the hypotheses of Exercise 2.1.

a) Verify that the $LR = 0$.

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.

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 2.10 (Sisters or half sisters?)**.**

We would like to determine whether two girls (called sister 1 and sister 2 in the left part of Figure 2.12) are sisters (corresponding to hypothesis H_1) or if they are half sisters (corresponding to hypothesis H_2 shown on the right hand side).

***** Insert Figure 2.12 around here *****

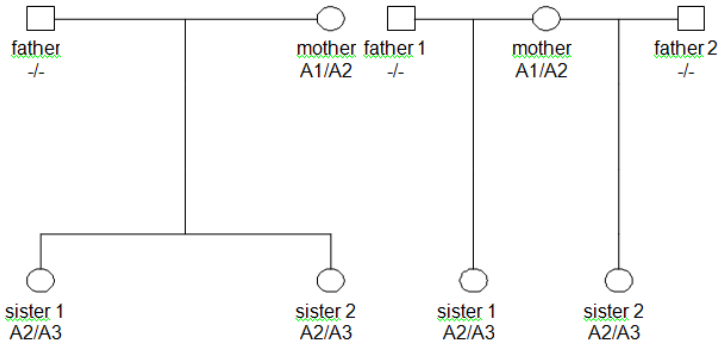

Figure 2.12: Pedigrees (hypotheses) in Exercise 2.10.

The genotypes are given in Table 2.9. with allele frequencies 0.1 for systems S1 and S2, and 0.05 for systems S3–S5.

a) What is the LR comparing the full sister alternative to the half sister alternative?

b) The LR in this case does not give rise to a clear conclusion. How would you determine the required number of further markers needed for a reliable conclusion? What markers would you use? You are not asked to do specific calculations.

Person	S1	S2	S3	S4	S5
Mother	A1/A2	A1/A2	A2/A3	A2/A4	A2/A3
Sister 1	A2/A3	A2/A3	A3/A4	A3/A4	A3/A4
Sister 2	A2/A3	A2/A3	A1/A3	A1/A3	A1/A3

Table 2.9: Marker data for Exercise 2.10

Exercise 2.11 (Silent allele)**.**

a) See Figure 2.13. This is a paternity case where there is suspicion of a silent allele. Include a silent allele frequency of $p_S = 0.05$, and calculate the LR(father/not father) using `Familias`. The allele frequencies for A and B are $p_A = p_B = 0.1$.

b) Confirm the above result using[3]

$$LR = \frac{p_s(p_A + p_s)}{(p_A + p_s)^2(p_B + 2p_s) + p_s p_A(p_B + 2p_s)}. \tag{2.21}$$

Exercise 2.12 (*Theta correction, derivation of formula)**.**

Verify theoretically the formula in Exercise 2.1h). *Hint*: Use the sampling formula described in Section 2.5.1.

Exercise 2.13 (Theta correction, `Familias`)**.**

This exercise expands on the previous Exercise 2.4 by introducing theta correction (kinship-parameter) of 0.02. Calculate the LR based on all markers. To save time you can read input from the file `Exercise2_4.fam`.

Exercise 2.14 (Input and output)**.**

A practical way to start work with `Familias` is to begin by reading a file containing the relevant database. Sometimes it is, however, of interest to read and write databases and case data and this will be the topic below.

a) Read input from the file `Exercise2_3`.

b) Export the data base from the `General DNA data` window. Name the output file `database.txt`.

[3]See http://dna-view.com/patform.htm

c) Export the case data from the `Case DNA Data`. Name the output file `casedata.txt`.

d) Open a new project.

e) Import `database.txt` from the `General DNA data` window.

f) Import `casedata.txt` from the `General DNA data` window.

g) Define the pedigrees, see Exercise 2.3, and calculate the LR for all markers.

Exercise 2.15 (*Paternity case with dropout)**.**

a) Load the file `Exercise2_15.fam`. Consider the hypotheses of Exercise 2.1. Confirm that that $LR = 0$ and find the one marker where the child and the alleged father do not share an allele.

b) Set the dropout probability to 0.1 for this marker, choose `Consider dropout` in the `Case DNA data` window for the child and recalculate the LR.

c) The data could also be explained by a mutation. Compare the above result with the LR you get with the mutation model (for this marker) `Stepwise (stationary)` for females and males with mutation rate 0.001 and mutation range 0.5. Remove dropout. *Comment*: One could consider both dropout and mutation.

d) Discuss: Should dropout be used for all homozygous markers?

Exercise 2.16 (Silent allele and dropout)**.**

Consider the paternity case in Figure 2.13. We analyze two scenarios, first that there is a silent allele passed on from the alleged father to the child, and second that there is a dropout in both AF and the child. Let the allele frequencies of A and B both be 0.2.

***** Insert Figure 2.13 around here *****

a) Include a silent allele frequency of 0.05 and calculate the LR using `Familias`.

b) Remove the silent allele frequency and instead include a dropout probability of 0.05 for both the alleged father and child. Calculate LR.

Exercise 2.17 (Simulation)**.**

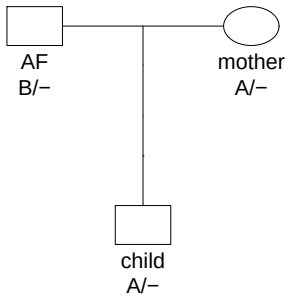

Figure 2.13: Pedigree for Exercise 2.11 and 2.16.

Load the file `Exercise2_17.fam`. The hypotheses considered are as in Exercise 2.1. 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. 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.

Exercise 2.18 (A fictitious paternity case)**.** A child was conceived as a result of a rape. The DNA-profiles of the defendant, mother, and child are available, see Table 2.10[4] below. The Likelihood Ratio $LR_{1,2}$ of this genetic evidence for the hypotheses

H_1 : The defendant is the father of the child,

H_2 : The defendant is unrelated to the father of the child,

is very high, thus providing strong evidence for paternity of the defendant.

Now suppose that the defendant claims that he is innocent, but that he believes his brother is the actual father of the child. We formulate a third hypothesis

[4]Thanks to Klaas Slooten for this exercise. This is fictitious case. Similar cases are sometimes encountered in practice.

	Locus	Mother	Child	Defendant	$LR_{1,2}$
1	CSF1PO	10/14	10/15	14/15	4.56
2	D2S1338	17/17	17/24	17/24	4.26
3	D3S1358	14/16	14/17	17/18	2.36
4	D5S818	11/13	12/13	11/12	2.83
5	D7S820	11/12	11/12	11/12	2.92
6	D8S1179	10/14	10/15	14/15	4.56
7	D13S317	8/13	12/13	12/12	3.24
8	D16S539	9/10	9/9	9/12	4.81
9	D18S51	13/14	14/18	13/18	5.45
10	D19S433	14/14	14/14	14/14	2.93
11	D21S11	29/29	29/30	30/33.2	2.15
12	FGA	22/24	24/24	22/24	3.63
13	TH01	9.3/9.3	9.3/9.3	7/9.3	1.64
14	TPOX	8/8	8/8	8/8	1.84
15	vWA	15/18	15/16	16/16	4.96
16	All				50218439.00

Table 2.10: Data for Exercise 2.18.

H_3 : The defendant's brother is the father of the child,

see Figure 2.14.

***** Insert Figure 2.14 around here *****

a) Give the algebraic formula for the Likelihood Ratio $LR_{1,2}$ for loci CSF1PO, D7S820, and D19S433.

b) Give the algebraic formula for the Likelihood Ratio $LR_{3,2}$ for the same loci.

c) Can you compute $LR_{3,2}$ numerically with the information above, or do you need a table of allele frequencies? Can you explain why?

d) What is the Likelihood Ratio for H_1 versus H_3 based on these three loci?

e) In the algebraic formula for $LR_{1,3}$ for locus CSF1PO, calculate its limits for $p_{15} \to 1$ and $p_{15} \to 0$, and explain the outcome.

f) Do the same for $LR_{1,3}$ on locus D19S433 when $p_{14} \to 1$ and $p_{14} \to 0$.

g) Discuss in the same way locus D7S820.

h) It can be shown that $LR_{3,2} = 500$. Can you calculate the probability that each hypothesis is true?

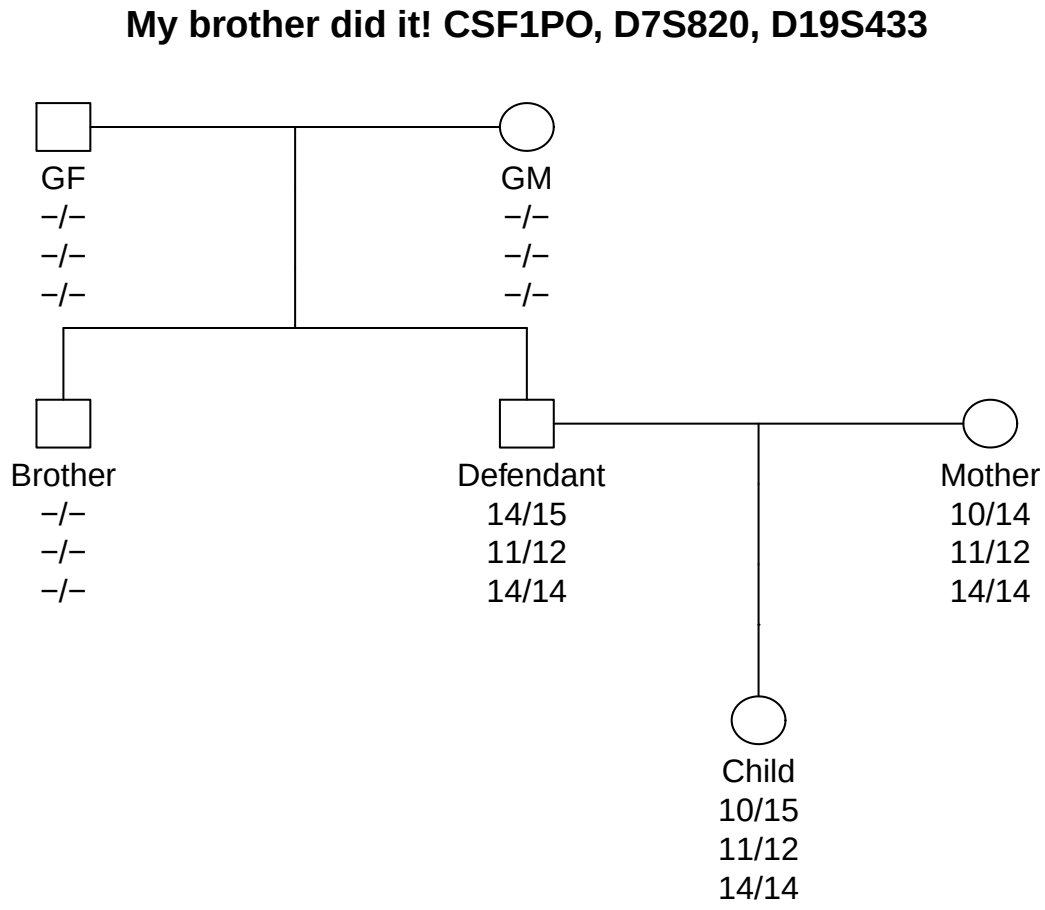

Figure 2.14: Hypothesis 2 in Exercise 2.18 with genotypes for CSF1PO, D7S820, D19S433.

3.5 Exercises

Solutions and input files can be found at `http://familias.name`.

Exercise 3.1 (DVI - A warm up example)**.** We consider a small DVI case with necessary data given in Figure 1.3, page 7 (also available as `http://familias.name/Fig_3_2.PNG`). There are three victims, called V1, V2 and V3. Reference data have been obtained from two families, F1 and F2. There is one marker (M1) with equifrequent alleles 1, 2, 3 and 4. (We assume mutation rates equal to zero.) In the DVI module of `Familias`, the reference families are treated one-by-one sequentially. The hypotheses for family F1 are:

- H_1: V1 belongs to F1.
- H_2: V2 belongs to F1.
- H_3: V3 belongs to F1.
- H_4: Some unknown person, NN, belongs to F1.

In other words, we consider four different possible hypotheses for the first family (F1). We start with some theoretical calculations before confirming the manual derivations in `Familias`.

a) Find the likelihood ratios $LR_i = L_i/L_4$.

b) We assume a flat prior, i.e., $\Pr(H_i) = 0.25$. Calculate the posterior probabilities $\Pr(H_i \mid \text{data})$.

c) The hypotheses for family F2 are:

   H_1 : V1 belongs to F2.

   H_2 : V2 belongs to F2.

   H_3 : V3 belongs to F2.

   H_4 : Some unknown person, V4, belongs to F2.

d) Repeat a) and b) for the above hypotheses.

e) Confirm the manual calculations in `Familias`. First, define the frequency database according to the specifications above.

f) Open the DVI interface.
Hint: `Tools -> DVI module -> Add Unidentified Persons`. Add persons and corresponding genotypes as specified for V1, V2 and V3 in Figure 1.3, page 7.

g) Continue by specifying the reference families *Hint*: `Next` and then `Add`. The families are specified by first specifying the persons, i.e., for F1 we define the typed father and mother while for F2 we define the typed father. Add the genotypes as specified in Figure 1.3, page 7. Next add the pedigree specifying the necessary relationships between the typed persons and the missing person. *Hint*: `Add` in the `Pedigrees` section. Never mind the pedigree named `Reference pedigree` for now, we will return to the importance of this later.

h) Perform a search using a `Threshold/Limit` of 0. *Hint*: `Next` and then `Search`.

i) Confirm the manual calculations above.

Exercise 3.2 (Blind search - A warm up example)**.** The blind search interface is, as the name suggests, a tool to blindly search for predefined pairwise relationships in a data set. The interface currently allows the user to search for Parent-Child, Siblings, Half-siblings, Cousins, Second cousins as well as Direct matches. For the direct matching, `Familias` implements a special algorithm, which will be explored in more detail later. A blind search may be performed in connection with the DVI module, e.g. to search a set of unidentified remains for direct matches or relationships within the data set. Another application may be to investigate a data set for unspecified relations before conducting a medical study or prior to creating a frequency database. The computations are swift and may be used to search large data set for relations. We will first test the module on a smaller data set.

a) Create one allele system with four alleles, 12, 13, 14 and 15 and allele frequencies uniformly distributed as 0.25.

b) Define four males P1, P2, P3 and P4. Enter DNA data as P1=12/12; P2=12/12; P3=13/14; P4=14/15. *Hint*: You should define individuals in `Tools->Persons`.

c) Enter the `Blind search` dialog *Hint*: `Tools->Blind search`.

d) Press `New search` and select `Direct-match` and `Siblings` as relationships. *Hint*: Hold down `Ctrl` to select multiple items in the list. Leave the `Match threshold` and θ at their default values. Set the remaining parameters (`Typing error`, `Dropout prob.` and `Dropin`) to 0.

e) Interpret the results.

f) Confirm the LR-values by manual calculations.

g) Try different values for `Typing error`, `Dropout prob` and `Dropin`. Specifically, change only one value to 0.1, i.e., leaving the other two at 0. Compare results to what you expect (exact calculations are not expected here).

Exercise 3.3 (DVI - An extended example)**.**

Consider the crash of a small plane with 10 passengers. 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) We suspect there might be relatives among the unidentified persons. Enter the first step, `Add unidentified persons` and select `Blind search`. Use your knowledge from Exercise 3.2 to 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?

p) 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 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 3.4 (DVI - Quick searching)**.**

Generally, pedigree structures may be complex and thus calculations with a large number of unidentified individuals will be computer intensive. `Familias` implements functionality to speed up calculation in the DVI interface. This exercise will illustrate the features and how they are used.

a) Open the file **Exercise3_4.fam,** which contains the final project from the previous exercise. Skip to the `Search` dialog in the DVI interface and press `Quick scan`. This will open a version of the `Blind search` dialog. Select `Parent-Child` and `Siblings` and match limit 10. Hit `Search`.

b) Compare the current results to the results obtained in the previous exercise.

c) What are the obvious benefits of using the `Quick scan` function?

d) In addition to the `Quick scan` function, `Familias` implements an algorithm to quickly compare reference families with PM samples using a zero mutation rate model. Enter the advanced settings, `File->Advanced` and select/deselect `Quick search` to activate/deactivate this feature.

e) Make sure the `Quick search` is selected and enter 1 as number of allowed mismatches. Return to the `Search` dialog in the DVI interface and perform a search with a match threshold of 0.0001.

f) Save the results using the `Export list` function.

g) Return to the advanced settings and change the number of allow mismatches to 2.

h) Perform a new search and compare the results to the ones obtained in f).

i) In light of the results, discuss "good" values on the number of mismatches and the benefits of using the quick search features.

Exercise 3.5 (DVI - A simulated example)**.**

This exercise is divided into two different parts,

1. Simulate data in `Familias` and prepare it for input to the DVI module. (Requires `Excel` or similar software)

2. Using the DVI module on the simulated data.

The exercise deals with a larger dataset of simulated pairs of sisters using real frequency data for 23 autosomal STR markers. As described, the first part of the exercise deals with how to simulate relationships in `Familias` and use the genotype data from the simulations. Skip to h) for DVI exercises only. Parts b) through g) are particularly useful for readers interested in simulating data for validation purposes and to understand some of the import formatting recognized by `Familias`.

a) Open the database from the file **Exercise3_5.fam,** containing frequency data for 23 autosomal STR markers. Briefly explore the database before continuing to make sure you are familliar with the markers and parameters.

b) Enter the `Advanced` dialog (*Hint*: `File->Advanced`) and make sure `Save genotype data` is selected and `Save complete data` is deselected. Press `Save`.

c) Define the persons necessary to simulate a pair of full siblings, i.e., Mother, Father, Sister1 and [Sister]. Note, it is important that you name the second sibling as indicated, i.e., [Sister]. In later steps `Familias` will recognize this as a relationship indicator and construct pedigrees automatically.

d) In the `Pedigree` dialog, create two pedigrees, the first specifying Sister1 and [Sister] as full siblings. The second one without any relations.

e) Press `Simulate`. Specify that we will have genotype data for Sister1 and [Sister]. Enter 100 simulations and set the seed to 12345. Also, select to save the raw data.

f) Press `Simulate` and save the genotype data to a text file. (Leave `Familias` open)

g) The following steps prepare the output for import into the DVI module of `Familias`:

1. Open the simulated genotype data in `Excel`, or similar software. (If the alleles appears as dates, please consult `http://familias.name`.)
2. Save the file as a new file using the xls or xlsx format for better compatibility.
3. Select the first row and select `Filter`. *Hint*: Usually found in the `Data` tab.
4. In the first column (`True ped`) select to only view `Ped 1` (or the name you gave to the full siblings pedigree) and in the second column select to only view `Sister1`.
5. Now, copy all the filtered data into a new file. *Hint*: `Ctrl+a` followed by `Ctrl+c` followed by `Ctrl+n` followed by `Ctrl+v`. Remove the first column, i.e., information about the pedigree.
6. In the first column select the first occurrence of Sister1. Fill down, such that the numbering is Sister1, Sister2,...,Sister100.

7. Save the new formatted file as **Exercise3_5_pm.txt** and make sure the format is tab-separated text file. (This file now contains data for 100 unrelated individuals serving as the PM data later)
8. Return to the original `Excel` document and remove the filtering. In the first column select again to only view `Ped 1` and in the second column select to only view `[Sister]`. Using the same procedure as in 5., copy the filtered data to a new document. In the first column in the second row, rename from `Ped 1` to `Family1`. Fill down such that the numbering becomes sequential, i.e., Family1, Family2,...,Family100.
9. Save as **Exercise3_5_am.txt**. This file now contains data for 100 reference sisters serving as the AM data later

h) Open the DVI interface and import the file**Exercise3_5_pm.txt** containing the data for 100 unrelated individuals serving as the set of unidentified remains. *Hint*: if you didn't perform the simulation steps, see a) before continuing.

i) Run a blind search and search for siblings using a match threshold of 10. (Leave all other parameters at their default values.)

j) Comment on the results

k) Continue to specify reference families. We will use the `Data only` import option, allowing a simple import. Select and import file **Exercise3_5_am.txt**. This will import 100 reference families with a brother as reference person. Explore the families and see how the import has automatically detected the relationships.

l) Perform a search using a match threshold of 10. Comment on false positives/negatives and investigate spurious matches using the `View match` button. For comparison, SisterX goes with Family X, where X is an integer 1-100.

Exercise 3.6 (Familial searching - Warm up example)**.** This exercise introduces the *Familial searching* module in a simple case. First some calculations are done by hand which are later checked using `Familias`. Consider a marker L1 with alleles 12, 13, 14 and 15, all with frequency 0.25, and a database of convicted offenders consisting of four individuals with genotypes P1=12/12,

P2=12/13, P3=13/14, and P4=14/15. For simplicity, we disregard complicating factors like mutation, theta-correction, dropin, dropout and typing error.

a) There is a stain S1=12/13, assumed to be from one individual. Consider

H_i: $S_1 = P_i$

H_0: S_1 is unrelated to P_i.

and calculate $LR_i = \Pr(\text{data} \mid H_i)/\Pr(\text{data} \mid H_0)$. (Answer: $LR_2 = 1/(2 \times 0.25 \times 0.25) = 8$ and $LR_i = 0$ for $i \neq 2$.

b) Repeat above calculations when $H_i : S_1$ is child of P_i. (Answer: $LR_1 = 2, LR_2 = 2, LR_3 = 1, LR_4 = 0$.)

c) There is a stain $S_2 = 12/13/14/15$ assumed to be from two contributors. Consider

H_i: Stain comes from P_i and an unrelated individual not in the database.

H_0: The stain comes from two unrelated individuals not in the database.

Calculate $LR_i = \Pr(\text{data} \mid H_i)/\Pr(\text{data} \mid H_0)$. (Answer: We first find the likelihoods

$$\Pr(\text{data} \mid H_0) = 24 \times p_{12}p_{13}p_{14}p_{15} = \frac{3}{32}$$
$$\Pr(\text{data} \mid H_1) = 0,$$
$$\Pr(\text{data} \mid H_i) = \frac{1}{8},\ i = 2, 3, 4.$$

Therefore $LR_1 = 0,\ LR_2 = LR_3 = LR_4 = \frac{4}{3}$.

d) Consider

H_i: Stain comes from son of P_1 and an unrelated individual (NN) not in the database.

H_0: The stain comes from two unrelated individuals not in the database.

Calculate $LR_1 = \Pr(\text{data} \mid H_1)/\Pr(\text{data} \mid H_0)$. Answer: The son must be 12/13, 12/14 or 12/15 and so

$$\begin{aligned}\Pr(\text{data} \mid H_1) &= \Pr(\text{son} = 12/13, NN = 14/15 \mid \text{father} = 12/12) \\ &+ \Pr(\text{son} = 12/14, NN = 13/15 \mid \text{father} = 12/12) \\ &+ \Pr(\text{son} = 12/15, NN = 13/14 \mid \text{father} = 12/12) \\ &= 6p_{13}p_{14}p_{15}.\end{aligned}$$

Therefore $LR_1 = \frac{1}{4p_{12}} = 1$.

e) Consider the above problem with P2=12/13 replacing P1 and calculate $LR_2 = \Pr(\text{data} \mid H_2)/\Pr(\text{data} \mid H_0)$. Answer: The son must be 12/13, 12/14, 12/15, 13/14 or 13/15 and so

$$\begin{aligned}\Pr(\text{data} \mid H_2) &= \frac{1}{2}(p_{12} + p_{13})2p_{14}p_{15} \\ &+ 2p_{13}p_{14}p_{15} + 2p_{12}p_{14}p_{15} \\ &= 3p_{12}p_{14}p_{15} + 3p_{13}p_{14}p_{15}\end{aligned}$$

Therefore $LR_2 = \frac{p_{12}+p_{13}}{8p_{12}p_{13}} = 1$.

f) Next, the above calculations are confirmed using `Familias`.

1. Start `Familias` and define the marker L1.
2. Enter `Tools->Familial searching`: Enter name P1 and `Add`. Add genotype data by selecting system L1. The manual input deviates from other parts of the program to allow for mixtures. Select `Allele 1` and 12. Select `Allele 2` and 12. Press `Add` to add the observations.
3. Define P2, P3 and P4 similarly.
4. Enter `Next` and define the stain $S1$.
5. Enter 0 for all search options, i.e., `LR threshold` and the other parameters should be set to 0.
6. Select `Parent-child` and `Direct match` in the upper right corner.
7. Enter `Next` and press `Search`. Confirm the answers in a) and b).
8. Return to the previous dialog, where we defined the stain. Import a mixture from the file **Exercise3_6.xml**. The sample is the previous mixture of two contributors: 12/13/14/15.

9. Confirm the previous calculations.

Exercise 3.7 (Looking for the relative of a stain)**.** The following exercise will explore the Familial searching module more thoroughly.

a) Start by importing the frequency database from the file **Exercise3_7.fam** and explore the contents.

b) The database containing convicted offenders and traces from previous crimes is contained in the file **Exercise3_7db.txt**. Import the file into the familial searching interface.
Hint: `Tools->Familial searching`. In total there should be 1000 elements, genotyped for different sets of markers. The latter may be common as different marker kits may have been used throughout the history of the database.

c) Perform a blind search on the database searching, using the option *Direct match*. Use default parameters values. Be patient, the search may take a while. How many comparisons are performed?

d) Comment on the results and whether the values on the dropout, dropin and typing error parameters (default values) are appropriate.

e) Can a blind search be performed on a database of say 5,000,000 elements? Compute the number of comparisons necessary for such an operation.

f) Press `Next` and continue with importing a batch of traces from some crime scenes. Import data from the file **Exercise3_7tr.txt**.

g) Specify that you wish to search for `Direct match` first. Let *Dropin*=0.01, *Dropout*=0.05 and *Typing error*=0.001. Set the match threshold to 1.

h) Perform a search and comment on the results.

i) Return to the previous dialog to specify a new search. Now select the *Parent-Child* and *Sibling* relations. Set the match threshold to 10 and leave the other settings at their default values.

j) Perform a new search and sort the matches using the `Sort` button. Select a subset of the matches to explore further. Use the `Subset` button and select the *top-k* method. Enter 10 as the number of matches to select. Review the results.

k) Press `Search` to redo the search. Select the *LR threshold* method in the subset feature. Enter 100 and apply. Review the results.

l) What are the benefits and downsides of using the *top-k* method compared to the *LR threshold* method?

m) `Familias` furthermore implements another subsetting method, *Profile centered*. The algorithm will perform conditional simulations based on the candidate (matching) profile and find a (LR) threshold based on the results. Perform the search again and apply the *Profile centered* method with the `alpha` parameters set to 0.9. Review the results and compare them to the ones obtained using the other two subset methods.

n) * Try explaining the benefits of using the *Profile centered* method and the meaning of the `alpha` parameter. Without changing the parameter, what will happen for low values on `alpha`? Confirm your answer in `Familias`.

o) Now, return to the search options and specify that you wish to scale against `Cousins`. In addition, specify the value of F_{st} to 0.02. When may these specifications be relevant? (Leave the other parameters at their default values).

p) Perform a search using the new settings and comment on the results. Compare with the results obtained in j).

Exercise 3.8 (* Mixtures and relatives)**.**

This exercise will demonstrate the *Familial searching* interface when the evidence is a mixture. The ideas were briefly touched upon in Exercise 3.6 and we will here consider some more theoretical points as well as confirm these, when possible, in `Familias`.

Beginning with the basics, for mixtures we may consider the hypotheses

H_1**:** The stain S_1 is a mixture of $N-1$ unrelated individuals and the unrelated profile of interest, P_1.

H_2**:** The stain S_1 is a mixture of N unrelated individuals. P_1 is not in the stain and is unrelated to all N individuals.

In the current exercise we consider $N = 2$.

a) We define one allele system with alleles and frequencies as indicated in Table 3.4 (mutation rates are set to zero).

Alleles	Frequencies
12	0.1
13	0.2
14	0.3
15	0.2
16	0.1
17	0.1

Table 3.4: Allele frequencies for a genetic marker.

b) * The stain S_1 is genotyped as 12/13/14, while the genotype of P_1 is typed as 12/13. Compute the LR comparing H_1 with H_2 by hand.

c) Prepare the data in `Familias` and compute the LR for the *Direct-match* between S_1 and P_1 in the `Familial searching` module of `Familias`. (Set all parameters to zero.)

We can extend the above mentioned hypotheses to include a number of alternative hypotheses about relatedness, relevant for the concept of *familial searching*. We may specify

H_3**:** The stain S_1 is a mixture of $N-1$ unrelated individuals and the relative R_1 of P_1.

d) * Using the same specifications as in a), compute the LR by hand comparing H_3 and H_2 and that the child of P_1 may be in S_1.

e) * Confirm the calculations in `Familias`.

f) ** Compute by hand the LR comparing H_3 and H_2, but assume the untyped mother of child is in S_1. In other words

H_1**:** The stain S_1 is a mixture of a mother and her child, P_1 is the father of the child.

H_2**:** The stain S_1 is a mixture of a mother and a her child. P_1 is unrelated to the child.

Note: The LR cannot be confirmed in `Familias`.

g) * Consider instead that the relative, R_1, is the brother of P_1. Compute the LR by hand comparing H_3 and H_2.

h) * Confirm the calculations in `Familias`.

i) * Add three more stains (S_2, S_3, S_4) with genotypes 12/13, 12/12 and 12/14 respectively. Given that these are stains with a single contributor and that we still search for siblings of P_1 in the stains, try predicting how the LR will be for these different genotype constellations.

j) * Confirm your ideas in `Familias` and compare the results to the one you obtained in h).

Exercise 3.9 (Further use of the blind search function)**.**

In this exercise we will take a closer look on how to use the blind search function in a different setting. We will consider the scenario where we wish to create a new database for a specific population.

a) In `Familias`, open the `Create database` function.
Hint: `File->Create database`.

b) Import the file **Exercise3_9.txt** containing output for 210, supposedly unrelated individuals, sampled from a small subpopulation. *Hint*: use the `Import` button.

The data is prepared using the standard `Familias` format (i.e., tab-separated text file), though direct output from e.g. Genemapper is accepted as well.

c) The first step before creating the database is to ensure that the individuals are truly unrelated. Open the `Blind search` function from within the `Create database` dialog using the `Check data` button.

The blind search will create a temporary database, based on the 210 individuals. The statistical calculations may be biased, but will provide a general idea of any relationships in the data set.

d) Perform a search for direct matches and parent-child relations. Use a match limit of 1000 and set the F_{st} correction to 0.05 (Leave all other parameters at default values).

e) Comment on the results. Export the list using the `Export list` function.

f) Return to the previous dialog and remove all samples with a LR above 100,000. *Note:* It is sufficient to remove one of the samples in a match.

   In a real situation it may be that we desire a more specific investigation to why the samples match, but for now remove the sample with the lowest number.

g) Create a summary of some statistical parameters. *Hint*: Use the `Statistics` button to export relevant information to a file. Review the file in a spreadsheet software such as `Excel`. (For problems with `Excel`, see `http://familias.name`).

h) Create the database and save the file. Explore the created database.

Exercise 3.10 (* Direct matching)**.**

In this exercise will take a closer look at the direct searching algorithm. We will consider data for one autosomal marker, vWA. We wish to compare the hypotheses

H_1**:** Two profiles, G_1 and G_2, belong to the same individual.

H_2**:** G_1 and G_2 belong to two unrelated individuals.

where the LR is computed according to

$$\begin{aligned} LR &= \frac{\Pr(G_1, G_2 \mid H_1)}{\Pr(G_1, G_2 \mid H_2)} \\ &= \frac{\sum_{i=1}^{N} \Pr(G_{true,i}) \Pr(G_1 \mid G_{true,i}) \Pr(G_2 \mid G_{true,i})}{\Pr(G_1) \Pr(G_2)}, \end{aligned} \tag{3.7}$$

see also Section 3.3.2 for details.

The direct matching feature allows any two profiles to be matched against each other and we may calculate a probability that they originate from the same latent profile ($G_{true,i}$) where the latent profile is a priori unknown and we have to sum over all possibilities. We compute transition probabilities, $\Pr(G_2 \mid G_{true,i})$ and $\Pr(G_2 \mid G_{true,i})$. For the current exercise we can use the probabilities listed in Table 3.5, where d is the dropout probability, c is the dropin parameter and e is the typing error probability. *Note:* the table displays a simplified version of probabilities, neglecting events such as two dropouts and one dropin occurring at the same time.

G_j	$G_{true,i}$	$\Pr(G_j \mid G_{true,i})$
a/a	a/a	$(1-d^2)(1-e)(1-c)$
a/a	a/b	$e+(1-d)d(1-c)(1-e)$
a/a	b/b	$e(1-d)^2(1-c)$
a/a	b/c	$e(1-d)^2(1-c)$
a/b	a/b	$(1-e)(1-c)(1-d)^2$
a/b	a/a	$e+(1-e)(1-d)^2cp_b$
a/b	a/c	$e(1-c)(1-d)^2+d(1-d)(1-e)cp_b$
a/b	c/c	$e(1-c)(1-d^2)$
a/b	c/d	$e(1-c)(1-d)^2$

Table 3.5: Transition probabilities between genotypes, $j = 1, 2$.

a) Open the file **Exercise3_10.fam** containing a frequency database for the marker vWA.

b) Import data for 10 individuals from the file **Exercise3_10.txt**.
Hint: Use `Tools->Case-related DNA data`. The file contains no information on gender and `Familias` will display a warning, which may be ignored.

c) In the `Blind search` dialog, press `New search`. Select *Direct-match* and specify the `Typing error`, `Dropin prob.` and `Dropout prob.` to zero. Use a match threshold of 1.

d) View the match between P9 and P10. Compute the LR by hand.

e) Perform a new search, now setting the `Dropout prob.` to 0.1. (Leave the other parameters as previously stated)

f) Explore the results, press `View match` to investigate the profiles and matches closer.

g) * Calculate the match between the individuals named P1 and P2 by hand.
Hint: Use Equation 3.7.

h) * In words, explain why the LR becomes so high even though the profiles are different. Explain also why the match between P9 and P10 gets a lower LR than in d)

i) Repeat e) but change the parameters to `Typing error`=0, `Dropin prob.`=0.1 and `Dropout prob.`=0.

j) * Calculate the match between the individuals named P1 and P2 by hand.

k) We may actually compare H_1 to some other hypothesis about relationship, say for example

H_2**:** G_1 and G_2 belong to two brothers.

This may for example be relevant in a criminal investigation, where the defendant states that *It was my brother who did it!*.

Perform a new search, using the same parameters as in e), but specifying that we wish to scale versus siblings.

l) Comment on the new results.

m) * Using the results in g), calculate the LR by hand for the match between P1 and P2. *Note:* relationship calculations do not consider dropout, dropin and typing errors.

4.5 Exercises

Solutions and input files can be found at `http://familias.name`.

4.5.1 Autosomal markers and FamLink

`Familias`, used in previous chapters, does not consider linkage, instead we introduce `FamLink` for pairs of linked autosomal markers. Some basic functionality will be explored as well as more advanced functions and theoretical points.

Note that genders in the predefined figures may deviate from the ones in the exercise. This can however be overwritten in a subsequent step.

Exercise 4.1 (Simple paternity case. Video)**.**

We will first consider a simple exercise where the purpose is to get familiar with the user interface and create your frequency database. Normally a saved frequency database will be loaded. Consider a paternity case (Duo) (for illustration see Figure 2.2) with hypotheses,

H_1**:** The alleged father (AF) is the real father.

H_2**:** The alleged father and the child are unrelated.

a) Specify two allele systems, L1 and L2 with alleles 12, 13 and 14 for both systems.
Hint: `File->Frequency database`. Let $p_{12} = 0.1$, $p_{13} = 0.2$ and $p_{14} = 0.7$ for both systems. The loci are closely linked, specify the recombination rate (θ) to 0.01. *Hint*: `File->Frequency database->Options`.

b) In several situations we only have the genetic distance between markers, measured in centiMorgans (cM). How can we convert between genetic distance and recombination rate?

c) Select the appropriate pedigrees using `File->New wizard` and specify the data for the father as homozygous 12/12 for both loci and the child heterozygous 12/13 for both loci. Calculate the LR which should coincide with the theoretical value of 25. Also compute the posterior probability using equal priors *Hint*: press the LR/Posterior button to change the displayed results.

d) Try changing the recombination rate to 0.5 and calculate the LR again. What happens? Explain why!

e) * Use pen and paper to show that LR=25. Use the same notation as in Equation 4.3, page 113.

f) The alleged father states that it could be his brother who is the true father of the child. Change H_2 to uncle by returning to the `Select alternative hypothesis` window. Calculate the LR with recombination rate specified to 0.01.

g) Save the `FamLink` file (we suggest the file extension sav). Exit `FamLink`.

Exercise 4.2 (A case of disputed sibship)**.**

The exercise involves two persons P1 and P2 interested to find out whether they share the same mother and/or father. We consider multiple hypotheses,

H_1**:** P1 and P2 are full siblings,

H_2**:** P1 and P2 are half siblings,

H_3**:** P1 and P2 are unrelated.

a) Reuse the frequency database specified in the previous exercise. *Hint*: `File->Open->Project`. Select the appropriate pedigrees, corresponding to the hypotheses above, and specify the data for P1 as homozygous 12/12 for both loci and P2 as homozygous 12/12 for both loci. *Hint*: Select multiple alternative hypotheses by holding `Ctrl` while selecting.

b) Calculate the LR. Scale versus the Unrelated pedigree and discuss the LR:s for full siblings and half siblings. Compute also the posterior probabilities using flat priors.

c) * Compute by hand the LR comparing H_1 versus H_2 from previous output.

d) Try changing the recombination rate to 0.5 and calculate the LR again. What happens? Explain why!

e) * Can linked autosomal markers be used to distinguish maternal from paternal half siblings?

Alleles	vWA (freqs)	D12S391 (freqs)
14	0.30	
15	0.10	
16	0.05	
17	0.20	0.20
18	0.25	0.10
19		0.30
20		0.20
21		0.20

Table 4.13: Allele frequencies for Exercise 4.3.

Exercise 4.3 (Immigration case)**.**

In cases of immigration, several alternative hypotheses may be relevant and linked markers may prove useful in the determination of the most probable one. In the current exercise we will explore a case where autosomal markers are unable to distinguish between the alternatives. Consider marker data for two markers vWA and D12S391, with frequency data as indicated in Table 4.13. We furthermore specify hypotheses,

H_1**:** P1 is the uncle of P2.

H_2**:** P1 is the grandfather of P2.

H_3**:** P1 and P2 are unrelated.

a) Specify the recombination rate to 0.1. Specify the genotypes for P1 to 15/15 and 20/20 for vWA and D12S391 respectively, while for P2 has identical genotypes as P1, for both loci. Calculate the LR and scale versus Unrelated. Discuss the results

b) Without changing the recombination rate to 0.5, discuss what will happen if you were to do this.

c) Change the recombination rate to 0.5 and compare to your result in b).

d) Calculate the LR also for $r = 0.25$ and compare the results from a), b) and c)

e) Save the project using the .sav file extension.

Exercise 4.4 (It was my brother who did it!)**.**

`FamLink` may be used in criminal cases where the defendant claims: "It was my brother who did it!". Suppose we have some stain, assumed to be from one person, from a crime scene where we wish to provide some probability as to whether a suspect could be the source of the stain. The random match probability is low, suggesting that the defendant may be the source of the stain. The defendant claims he is innocent and suggests that a brother of him did it. We consider hypotheses,

H_{D1} : The brother of the defendant is the source of profile in the stain.

H_{D2} : A random man, unrelated to the defendant, is the source of the profile in the stain.

H_P : The defendant is the source of the profile in the stain.

a) Open the saved project from Exercise 4.3, containing the frequency database. Specify the recombination rate to 0.09.

b) Select the pedigree `Full siblings` (corresponding to H_{D1}) as the main hypothesis and select `Unrelated` as the alternative hypothesis, H_{D2}.

c) For the two individuals enter the same data for both individuals, 14/14 for vWA and 21/21 for D12S931.

d) Calculate the LR. Discuss the result.

e) * Compute the likelihood for H_P and H_{D2} by hand. Calculate also the LR comparing H_P versus H_{D2}

f) * Calculate the remaining LR:s comparing H_P with H_{D1}. *Hint*: Use the LR calculated by hand with the one obtained in e).

g) Export the multiplication factor and discuss its meaning. *Hint*: Use `Save results`.

h) Given the results in e), the defendant is interested if some other relative could be blamed. As a scientist you are curious to find out and try "Uncle" and "Grandfather" as alternative hypotheses. Calculate the LR and discuss the results.

Exercise 4.5 (On the impact of linkage)**.**

The purpose of this exercise is to illustrate how `FamLink` can be used to study the general impact of accounting or not accounting of linkage for an incest case, see Exercise 2.6, page 51. To study the impact on a specific case we may use simulations, a resourceful tool included in the software. Simulations are performed assuming each of the specified hypotheses to be true and calculate the LR for each case. A summary statistic is then reported.

We specify the hypotheses as

H_1**:** The father of the mother is also the alleged father (AF) of the child.

H_2**:** Another man, unrelated to the alleged father is the father of the child. The alleged father is still the father of the mother.

a) We use real frequency data from the STR markers vWA and D12S391. Import the file **Exercise4_5.txt**, containing the frequency database for the two STR markers. *Hint*: `File->Frequency database` and `Import`.

b) Set the recombination rate to 0.089. (This corresponds to estimates in the literature). Select appropriate pedigrees but do not enter DNA data.

c) Perform simulations using a seed value of 12345. *Hint*: Instead of `Calculate`, hit `Simulate`. Do 1000 simulations which is generally a reasonable number to obtain an idea of the distribution of the LR. Select to `Save raw data` and place the file somewhere you can find it.

d) Once we have performed simulations we may store a summary. Save a simulation report and explore the contents:
`Save results` and `Simulation report`.

e) What is the median effect of linkage on the LR in the current case given H_1 is true?

f) * Open the raw data txt-file in `Excel` (or similar software), and estimate the average, given H_1 is true, from the ratio

$$LR(\text{Not accounting for linkage})/LR(\text{Accounting for linkage}).$$

g) * The LR assuming H_2 is true will often be zero in the current case. Explain why. Note that zeros will be disregarded when calculating summary statistics.

h) ** The expected LR under H_2 is 1 as shown in (8.3). Is this true in the current case? What can be done to improve the fit to the expectation?

Exercise 4.6 (A real example)**.**

Another pair of closely located markers is SE33 and D6S1043 with a genetic distance of only 4.4 cM. The markers are both included in some commercial STR multiplexes. We are interested to find the multiplication factor for the hypotheses

H_1 : Two females, Sister1 and Sister2, are full siblings. (Data available for the common mother.)

H_2 : Maternal half siblings. (Data available for the common mother.)

a) Compute the recombination rate in `FamLink` using Haldanes mapping function. *Hint*: `Tools->Conversion`)

b) We use real allele frequency data from a Chinese Han population. Import the file **Exercise4_6.txt**, containing the frequency database for the two STR markers. Specify the recombination rate calculated in a).

c) Specify the data for Sister1 as 14/14 for D6S1043 and 21/23.2 for SE33, for Sister2 as 14/14 for D6S1043 and 21/24 for SE33. Furthermore, specify the data for the mother as 14/19.3 and 17/21.

d) Calculate the LR in `FamLink` and compute the multiplication factor by hand using the results.

e) The multiplication factor may be combined with the total LR obtained in some other software, where linkage is not accounted for. Discuss how this is possible.

f) * Perform 1000 simulations (using `seed=12345`). Find the distribution for the multiplication factor in the simulation report for each simulated hypothesis. What are the values for the 5 and 95 percentiles? *Hint*: Use the fact that the multiplication factor is equal to $LR(\text{no linkage})/LR(\text{linkage})$.

Exercise 4.7 (* Defining pedigrees)**.**

FamLink includes a number of predefined pedigrees where the user only needs to select the required pictures indicating the family structure. We may also create our own pedigrees using the Merlin Input file notation (see http://www.sph.umich.edu/csg/abecasis/merlin/tour/input_files.html). Consider a case of three persons interested to know whether they are all full siblings or unrelated:

H_1: Three persons (P1, P2 and P3) are full siblings.

H_2: P1, P2 and P3 are unrelated.

a) We will use the same frequency database as in Exercise 4.2, but specify the recombination rate to 0.1.

b) Specify the needed relationships in a text file and rename it to **Exercise4_7.ped**. Specify the data for all persons to 12/12 at the first locus in the same file. For the second locus we will add a previously unseen allele, denoted 11; specify the data for all persons as 11/12.

c) Import the pedigree into FamLink.
(This is done by File>New wizard>Import ped file). When the new allele is imported we require a frequency. Specify 0.05 for allele 11 and select "Search and subtract" as method. (See manual at http://famlink.se for details on the "Search and subtract" method)

d) ** Discuss the Search and subtract method in the current context. Would it be applicable in e.g. Familias?

e) Calculate the LR and discuss the results.

f) Discuss the hypotheses and whether you would consider alternatives.

Exercise 4.8 (A case of identification)**.**

The following serves to illustrate the importance of linkage in a real case. Identification of unidentified remains is a recurring task at forensic laboratories. A skeleton was found and successfully genotyped for two overlapping commercial STR kits, yielding in total data for 23 autosomal STR markers. A putative sister of a missing person (MP) was genotyped for the same 23 markers. The hypotheses are defined as

H_1 : The sister is the true sister of MP.

H_2 : The sister is unrelated to MP.

The combined results gave $LR = 500$ in favor of H_1. We will use `FamLink` to examine the results for the markers vWA and D12S391 and how the linkage between the two markers influence the results.

a) Open the file **Exercise4_8.sav** in `FamLink`. The file contains a real frequency database for the indicated STR markers from a Norwegian population sample. Specify the recombination rate to 0.089.

b) Select the hypotheses indicated above and enter the data for the remains as 14/14 for vWA and 21/21 for D12S391. Similarly for the sister enter 14/14 for vWA and 21/21 for D12S391.

c) Calculate the LR and discuss the results. Would accounting for linkage change you conclusion?

d) Try changing the genotype data for the persons to see how this affects the outcome.

e) * In light of the results, try to discuss when we can expect an increase/decrease in the LR when accounting for linkage in the current case.

Exercise 4.9 (* Theoretical exercise)**.**

The following exercise is meant for mathematically oriented readers. We will consider the well discussed example of uncle versus half siblings, mentioned in the text. Consider two STR markers L1 and L2, where the genetic distance and hence the recombination is unspecified but can be denoted with ρ. We further consider hypotheses according to,

H_1 P1 and P2 are related as half siblings

H_2 Two persons, P1 and P2, are related as uncle/nephew.

a) * The genotype data is indicated in Table 4.14. Use your knowledge about IBD patterns to derive a theoretical formula for the LR for the two markers L1 and L2. *Hint*: Use methods described in Example 4.2 on page 115.

	P1	P2
L1	9/12	12/15
L2	19/21	21/25

Table 4.14: Genotype data for Exercise 4.10 .

b) * Confirm your formula with `FamLink`, using p_{12}=0.1, p_{21}=0.05 and $r = 0.1$. *Hint*: The frequencies of the other alleles are irrelevant, use any value.

c) * Plot the LR as a function of the recombination rate using the same allele frequencies as in b).

Exercise 4.10 (Extended example)**.**

In addition to creating our own pedigrees we may also analyze previous `Familias` projects (v 1.81 or above), to obtain an LR where linkage between virtually any number of markers are considered. All commonly used STR markers as well as a number of less common markers are predefined with their genetic distances specified, see file **markerInfo.ini** provided in the `FamLink` install directory. As indicated we may now consider more markers and more complicated pedigree structures. The following is extracted from a real case, see Figure 4.19, consider

H_1: The **Alleged Father** is the true father of the **Child** (see Figure 4.19).

H_2: The **Alternative father** is instead the father of the **Child**.

***** Insert Figure 4.19 around here *****

a) Open the file **Exercise4_10.fam** in `Familias` and explore the project. (*Note:* the person marked as Cousin in Figure 4.19 is not included since `Familias` versions below 3.0 could not handle the complexity of the project). The project contains genetic marker data for 35 autosomal STR markers.

b) Calculate the LR in `Familias`, be patient, the computations may take > 20 minutes depending on the performance of your computer. You may also choose to skip this.

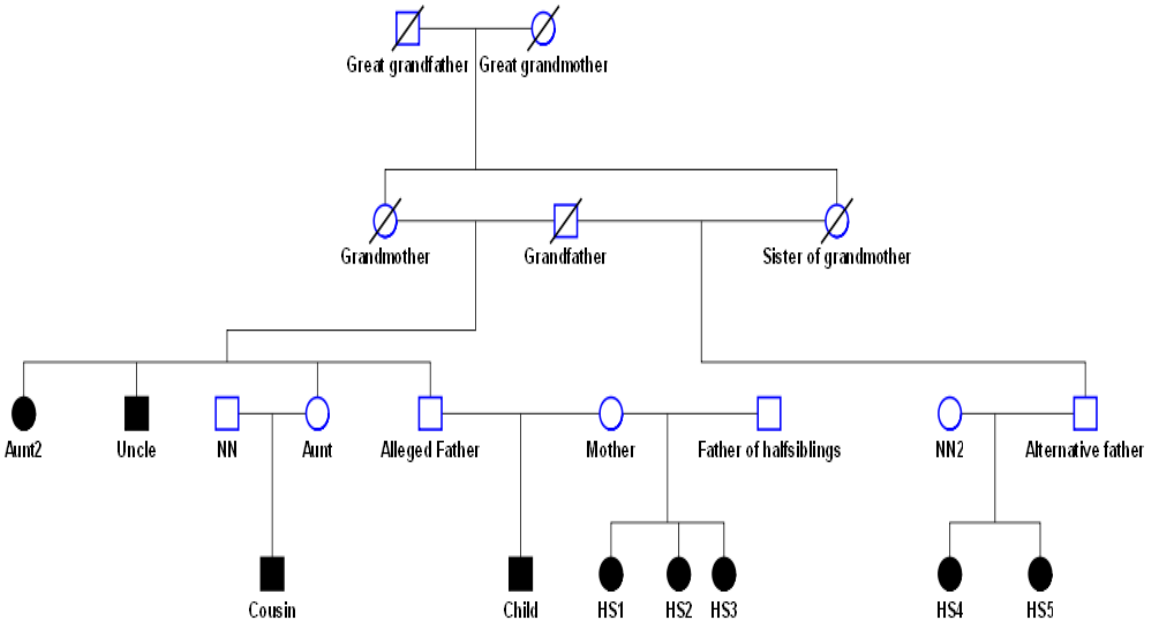

Figure 4.19: Complex kinship case.

c) Open the `Quick analysis` interface in `FamLink`. Mark generate report and browse to the file **Exercise4_10.fam**.
Hint: `Tools->Quick analysis`.

d) Perform an analysis of the `Familias` project by pressing `analyze`.

e) Open the generated report (found in the same directory as the exercise file) and browse the contents. Find the LR and compare it to the one obtained in `Familias`.

f) The analysis includes markers from the commercial kits HDplex (QIAGEN), PP16 (Promega) and ESX17 (Promega). What can be said about the number of linked markers on different chromosomes combining all the kits? *Hint*: See FamLink report)

g) Try using the function on your own `Familias` projects!

Exercise 4.11 (** Combining linkage and subpopulation effects)**.**

This exercise is interesting for the more mathematical oriented users. To allow for subpopulation structures we may adjust the allele frequencies using θ correction, see Section 2.5. We may actually combine the effect of linkage with correction for subpopulation structure in a model quite easily. Consider the hypotheses

H_1 Two persons, P1 and P2, are related as Half siblings.

H_2 P1 and P2 are Unrelated.

a) * P1 and P2 have genotype data as indicated in Table 4.14. Using the sampling formula, see p. 32, derive the theoretical expression for the LR. *Hint*: You may reuse formulas derived in Exercise 4.9)

b) * Using that $p_{12} = 0.2$, $p_{21} = 0.1$ and $r = 0.01$, compute the LR.

c) ** Using the sampling formula provided in chapter 2 and that $\theta = 0.02$, compute the updated set of allele frequencies. *Hint*: Given zero alleles are IBD we have four different observations, whereas given one allele IBD we have three different observations)

d) ** Compute the LR using that $\theta = 0.02$. *Hint*: Use the results in b) and c))

e) ** Plot the LR versus the value of θ using the specifications in b).

4.5.2 X-chromosomal markers and FamLinkX

`FamLinkX` implements an algorithm for linked markers on the X-chromosome. In addition to linkage the software accounts for linkage disequilibrium (allelic association) and mutations. The software is intended to be user-friendly but may provide obstacles for the inexperienced user. `FamLinkX` provides the likelihood ratios using three different methods, M1: Exact model, considering linkage, linkage disequilibrium and mutations; M2: Cluster approach, see manual for Merlin, linkage and linkage disequilibrium is considered but not recombinations within clusters and not mutations; M3: Only linkage is considered between markers. In the following exercises we are interested in M1

	12	13
16	59	1
17	1	39

Table 4.15: Haplotype observations for Exercise 4.12 .

as this is the preferred model, specially for STR markers, but comparisons to the other models will be made. For all calculations, unless stated otherwise, we consider X-chromosomal marker data and corresponding inheritance patterns.

Exercise 4.12 (A paternity case revisited. Video)**.**

We will first revisit the paternity case (Duo) (for illustration see Figure 2.6) with hypotheses

H_1**:** The alleged father (AF) is the true father of the female child.

H_2**:** The alleged father and the child are unrelated.

a) Open `FamLinkX` and specify the frequency database. *Hint*: `File->Frequency database`. Create a new cluster and specify two diallelic systems, L1 and L2, with alleles 12, 13 and 16, 17 respectively. Let $p_{12} = 0.6$, $p_{13} = 0.4$ for L1 and $p_{16} = 0.6$, $p_{17} = 0.4$ for L2. Select the `Simple mutation model` with the mutation rate set to 0 for both systems. Set the genetic position to 10 cM for L1 and 10.1 cM for L2. Furthermore, specify haplotype observations according to Table 4.15. Why is it important that we explicitly specify the gender of all persons in calculations for X-chromosomal marker data?

b) What is the estimated recombination rate between the two loci? Use Haldane's mapping function.

c) Why do we specify the number of observations for each haplotype?

d) Use the equation below to calculate a measure of the association between the alleles

$$r^2 = \frac{(p_{12}p_{16} - p_{12,16})^2}{p_{12}p_{13}p_{16}p_{17}}. \tag{4.12}$$

e) Specify λ to 0.0001 in `Options`. We will discuss the importance of λ in Exercise 4.14 and will not dwell further on it now. In brief, setting a low λ gives large weight to the observed haplotypes in Table 4.15. Select the appropriate pedigrees using the Wizard. The alternative hypothesis depicts two unrelated girls, but the first hypothesis will override the genders. *Hint*: `File->New wizard` and specify the data for the father as 12 for L1 and 17 for L2 and the child as 12/12 for L1 and 17/17 for L2. Calculate the LR which should coincide (approximately) with the theoretical value of 100. Choose to save the file when asked. There may be a small deviation from the theoretical value, which we will return to in later exercises.

f) Change the genotypes for the child to 12/13 for L1 and 16/17 for L2. Calculate the LR.

g) Change the number of observations for each haplotype. What happens? Explain why!

h) * Discuss the high degree of LD and if this is a likely situation to occur in reality.

Exercise 4.13 (A case of sibship revisited)**.**

In the second exercise we revisit the example in Exercise 4.2 concerning disputed sibship. Two females, F1 and F2, are interested to find out whether they are siblings in some way. We specify hypotheses

H_1**:** F1 and F2 are full siblings

H_2**:** F1 and F2 are maternal half siblings

H_3**:** F1 and F2 are paternal half siblings

H_4**:** F1 and F2 are unrelated

a) Explain why we may distinguish H_2 from H_3 with X-chromosomal markers but not with autosomal markers.

b) Use the same frequency data, and haplotype observations, as in Exercise 4.12, alternatively open the file **Exercise4_13.sav**.

c) Specify $\lambda = 0.0001$ and select the relevant hypotheses. Note: if you also stored the case-related DNA data in the previous exercise, you may be asked if you wish to erase all DNA data, answer yes.

d) Enter data for both F1 and F2 as 12/12 for L1 and 17/17 for L2. Calculate the LR Scale against H_4. Comment on the importance of accounting for LD and linkage in the current case.

Exercise 4.14 (On the importance of λ)**.**

This exercise is intended to provide some insight into how the haplotype frequencies are estimated and the importance of the parameter λ. Our model for haplotype frequency estimation is described by

$$F_i = \frac{c_i + p_i\lambda}{C + \lambda} \tag{4.13}$$

where F_i is the haplotype frequency for haplotype i, c_i is the number of observations for the haplotype, p_i is the expected haplotype frequency (assuming linkage equilibrium) calculated using the unconditional allele frequencies, C is the total number of observations for all haplotypes and λ is a parameter giving weight to the expected haplotype frequencies. This model allows for unobserved haplotypes to be accounted for, in contrast to models which estimate the haplotype frequency solely based on the counts. The difficulty lies in the choice of a good λ. Our recommendation is to compute the LR for a number of different values and select the least extreme value, i.e., the value closest to 1.

To start, we specify a case with an aunt of a female child

H_1 : The female is the aunt of the child.

H_2 : The two females are unrelated.

a) Again use the same frequency data as in Exercise 4.12. Select the relevant hypotheses. *Hint*: Select the *Aunt/Uncle* as the main hypothesis.

b) Enter data for the child as 12/12 for L1 and 16/17 for L2 and for the aunt as 12/13 for L1 and 17/17 for L2.

c) Calculate the LR for d $\lambda = 0.0001$, 0.01, 1, 100 and 10000.

d) What happens for large and small values of λ? *Hint*: use the equation for haplotype frequency estimation above.

e) Change the data for the child to 13/13 for L1. Repeat c) and discuss the results.

Exercise 4.15 (Extended example. Combining `Familias`, `FamLinkX`)**.**

This exercise provides a challenge where the user needs to combine the results from `Familias` and `FamLinkX` to obtain a final result. The data is extracted from a real case (anonymized) where three females provided DNA samples. The hypotheses are

H_1 : The three females are all full siblings.

H_2 : Any other pedigree constellations.

Obviously H_2 cannot be used in the current setting, in a simple way, and we need to refine possibly alternative hypotheses. a), b) and c) involve the use of `Familias`, however you may also skip to d) for `FamLinkX`.

a) * Open the file **Exercise4_15.fam** in `Familias 3`. We may assume that all females are children with no children of their own. Specify that the three females are children. Also define two parents, a mother and a father and specify that they are both born 1970.

b) * Use `Familias` (Generate pedigrees) to find the pedigrees with a posterior probability above 0.001. Which are the most probable relationships according to the results? Interpret the results.

c) * Discuss the constraints specified in a) and their impact on the results in b)

d) Open the database **Exercise4_15.sav** in `FamLinkX`, which contains frequency and haplotype data for the Argus X12 kit from QIAGEN based on a Swedish population sample. Explore the haplotype frequency database.

e) Based on the results in b), we specify the hypotheses

H_1 : The three females are all full siblings

H_2 : Two females are full siblings and the third (named F3) is a paternal half sibling

H_3 : Two females are full siblings and the third (named F3) is a maternal half sibling

f) Import the DNA data, available in **Exerecise4_15.txt**. Make sure to import the data in the file to the correct corresponding persons, the person denoted F3 should be imported to 3.

g) Calculate the LR and interpret the results. Be patient, the computation may require some time >20 min. *Hint*: To speed the computations up, go into `File->Advanced`, select and edit the hypothesis we have selected to investigate. For each pedigree set the `Threshold` value to 0.001 and the `Steps` value to 0.

h) Discuss if the LR in g) may be combined with the results in b)? What is your final conclusion on the case?

Exercise 4.16 (Further discussion of λ)**.**

We will provide an example of how the value of λ may crucial to the conclusion in a case. We use anonymized data from a real case with two typed females (F1 and F2) and consider hypotheses

H_1 : F1 and F2 are paternal half siblings, with different mothers.

H_2 : The two females are unrelated.

a) Open the file **Exercise4_16.sav**, containing the frequency database.

b) Select appropriate pedigrees and import genotype data from the file **Exercise4_16.txt**.

c) Compute the LR for a number of different values on λ, e.g., 0.001, 1, 100 and 1000.

d) Discuss the results in c). What conclusion can be drawn?

e) * Explore the genotype data and see if you can find an answer to the results. Use your knowledge about haplotype phases under the different hypotheses. *Hint*: Use the frequency estimation tool in the `Edit cluster` dialog.

f) Discuss what value on λ should be chosen. What is most conservative?

Exercise 4.17 (ESWG 2013 paper challenge)**.**

This exercise covers the calculation of the LR for the X-chromosomal data included in the ESWG paper challenge 2013 [51]. The question involved a paternity duo with the following hypotheses,

H_1 : AF is the biological father of a female child.

H_2 : AF and the child are unrelated.

a) Open the file **Exercise4_17.sav** containing the frequency database. Explore the specification of the haplotypes. Make sure λ is set to 1.

b) Select appropriate pedigrees and import genotype data from the file **Exercise4_17.txt**.

c) Calculate the LR.

d) Change the value of λ to 212 and see how this affects the results.

e) Compare the results calculated with the equilibrium (LE) model (M1) with the Exact (M3).

f) ** Try deriving the algebraic formula for the two markers in the first cluster. For simplicity assume mutations can be disregarded.

Exercise 4.18 (* Creating pedigrees)**.**

`FamLinkX` includes some predefined pedigrees where the calculations are performed using method M3. In addition, you may wish to create new pedigrees. The implementation currently does not allow method M3 to be used on user-defined pedigrees, therefore only calculation with methods M1 and M2 will be done. Two females F1 and F2 are related as maternal cousins. In addition, we are asked to find out whether they also share the same father,

H_1 : F1 and F2 are maternal cousins as well as paternal half siblings.

H_2 : F1 and F2 are maternal cousins with different fathers.

a) Open the file **Exercise4_18.sav** containing the frequency database. We will consider different λ, start by setting the parameter to 1.

b) Continue by creating the pedigree for H_1. You need do define several extra persons to specify the necessary relations.
Hint: `Tools->Select pedigree->Create/Edit pedigree`.

1. Rename the pedigree to H1
2. Add extra persons *Hint*: `Persons` button within the edit pedigree dialog. Add persons named F1, F2, grandmother, grandfather, mother1, mother2 and father. Make sure the genders are correct for all individuals.
3. Specify the relations in H1.
4. Close the edit pedigree dialog. The created pedigree will appear last in the list of pedigrees. Select the created pedigree and `Next`.
5. Create the alternative hypothesis, H_2, using the same procedure as for H1. Note: you need not define any new persons, use the same as create for H1.

c) Import genotype observations from the file **Exercise4_18.txt**. Make sure to import the person named *cousin1* to F1 and *cousin2* to F2.

d) Calculate the LR.

e) Repeat the calculations for different values of λ. What is your conclusion regarding the relationship between the two women.

Exercise 4.19 (* Theoretical considerations)**.**

For verification purposes, and to validate calculations/implementations, it may be interesting to derive theoretical formulas. Generally this is infeasible, but for some cases, using simplifications, algebraic formulas can be derived. We will start by consider one diallelic SNP marker, S1, with alleles 1 and 2. Frequencies are $p_1 = 0.4$ and $p_2 = 0.6$ and mutation rates are zero. Furthermore, we specify,

H_1 : A women (AM), is the alleged mother of the girl.

H_2 : The alleged mother and the girl are unrelated.

a) The AM has genotype 1/1 while the child has genotype 1/2. Compute the LR by hand and verify in `FamLinkX`.

The AM is discarded as the true mother for reasons other than the DNA. Instead an alleged father is presented. The hypotheses become,

H_1 : A man (AF), is the alleged father of a female child.

	1	2
3	1	59
4	39	1

Table 4.16: Haplotype observations for Exercise 4.19.

H_2 : The alleged father and the child are unrelated.

b) The AF has genotype 1 while the child still has genotype 1/2. Compute the LR by hand and verify in `FamLinkX` (Note: the alternative hypothesis displays two unrelated females, while as pointed out earlier, the genders in the main hypothesis overrides this information).

c) A second SNP, S2, is introduced with alleles 3 and 4. Frequencies are $p_3 = 0.6$ and $p_4 = 0.4$. The AF has genotype 3 while the child has genotype 3/3. Compute the LR by hand and verify in `FamLinkX`. You may assume that the two SNP:s are located far from each other and thus the calculations independent. This is in practice impossible in FamLinkX, since we only have one chromosome, a sufficient approximation is to define the genetic position well apart, e.g. 500 cM.

d) S1 and S2 are in fact closely located, separated by only 0.1 cM. Specify this in `FamLinkX`. Calculate the LR again.

e) There are further haplotype observations according to Table 4.16. Specify this in `FamLinkX`.

f) Use three different values for λ, 0.01, 1, 100, and compute the LR in `FamLinkX` for each.

g) *Derive the theoretical formula to confirm the calculations in g) for λ=100.

h) ** Plot the LR as a function of λ. *Hint*: Use the formula in Exercise 4.14)

i) ** Try deriving the theoretical formula for the paternity case in Exercise 4.12 using the specifications in e) and f) in the exercise.

Exercise 4.20 (** Exploring the algorithm)**.**

The exercise will explore the algorithm implemented in `FamLinkX`. The paper by Kling et al [68] provides further details that may be needed to solve some of the questions. We first introduce one genetic marker with alleles 12,13,14,15 and 16. Consider hypotheses

H_1 : Two girls, F1 and F2, are maternal half siblings.

H_2 : F1 and F2 are unrelated

a) * Given H_1, specify the founder alleles patterns, i.e., possible set of alleles for the founders, if the genotypes are 12/13 for F1 and 12/15 for F2. Assume mutation rates are zero. *Hint*: Plot the pedigree on a paper and specify the founders.

b) * Given H_1, specify the founder alleles patterns if the genotypes are 12/13 for F1 and 14/15 for F2. Now assume one-step mutations are possible.

c) * Set up the inheritance patterns, i.e. which are the meioses required to be accounted for?

d) * Using the results in a) and c), compute the number of different combinations we have to consider. In other words, the number of founder allele patterns times the number of inheritance patterns.

e) ** Now consider a second marker with the same data, i.e., the results in d) apply also for this marker. The two markers are linked such that for each combination for the second marker we also have to consider the different inheritance patterns in c). Using this information, compute the updated number of combination we have to consider for the second marker.

f) ** We also need to account for linkage disequilibrium using founder allele patterns, i.e., for each combination at the second marker we have to consider all possible states at the first marker. Using this information, compute the updated number of combination we have to consider for the second marker.

g) ** Finally, we consider a third marker, tightly linked with the second marker. Given that we have to consider linkage disequilibrium across all three markers what are the number of combinations we have to consider for the third marker? *Hint*: use that linkage only stretches one marker, i.e., given a specific inheritance state for the second marker, the third marker is independent of the different marker states at the first marker)

h) ** Consider instead H_2 and compute the number of combinations for the third marker. What effect does linkage have on the number of combinations we have to consider?

Exercise 4.21 (A case with mutation)**.**

We will consider a case where we introduce a possible mutation. Consider two females, F1 and F2, with hypotheses about relationship given as

H_1 : Maternal half siblings.

H_2 : Paternal half siblings.

which is not distinguishable using autosomal markers.

a) Import frequency data from the file **Exercise4_21.sav**.

b) Select maternal half siblings as the main hypothesis and select paternal half siblings as the alternative.

c) Import genotype data from the file **Exercise4_21.txt**

d) Compute the LR. Use the default $\lambda = 1$.

e) Change the data for the first individual (F1) to 19/25.1 for the marker DXS10148.

f) Compute the LR. Use default values for mutation parameters, i.e the simple model where each mutation is equally likely.

g) Change the data for the first individual (F1) to 19/19 and compute the LR.

h) Change the mutation model to `Extended model` with the `Range` parameter equal to 0.1 and the `Rate 2` parameter equal to 0.000001. ***** Thore: mention that this is only for the marker DXS10148, keep mutation rate. Just to check: there is apply mutation models to all. Why different mutation models for Familias and FamLinkX? *****

i) Compute the LR for the data as indicated in e) and g).

j) * Discuss the results and the importance of good models for mutations. Why is this decisive in the current case?

Exercise 4.22 (* Understanding complex parameters)**.** We will now have a closer look at some parameters that may further affect the results in `FamLinkX`, the *Threshold* and *Step*. The former, here denoted t, is a floating value, with the constraints ***** Changed: ***** $0 \leq t \leq 1$ and specifies the minimum value required for the overall pedigree likelihood to be included in the calculations. In other words, this limit apply to the $\Pr(D_i \mid V_i, F_i)$ in Equation 4.9.

The latter parameter, here denoted s, is a positive integer, with the constraints ***** Changed: ***** $s \geq 0$ and specifies the number of steps we consider for founder alleles. For instance, in a case of paternal half sisters, we need to sum over the possible alleles for the common father, given by the shared alleles of the sisters. If we consider mutations, all alleles are possible, while we can use the *Step* parameter do define the number of steps away from the sisters' alleles that we wish to include in the summation. In reference to Equation 4.9, this limits the founder alleles space F_i.

We will use an example to illustrate the effect of tuning the different parameters, consider

H_1 : Two females, F1 and F2, are paternal half sisters.

H_2 : The two females are unrelated.

a) Import frequency database from the file **Exercise4_22.sav**. The file contains a database for a Somali population [67]. The extended step wise mutation model is specified for all markers.

b) Specify the hypotheses according to H_1 and H_2. Import case data from the file **Exercise4_22.txt**.

c) Calculate the LR. Make sure λ is specified to 1.

d) Explore the genotype data to find an explanation for the results.

e) In the Advanced settings (`File->Advanced`), change the *Threshold* parameter for H_1 to 0.00001 leaving ***** other parameters unchanged? ***** the *Step* parameter unchanged for both H_1 and H_2. Save.

f) Compute the LR

g) Change the *Threshold* parameter for H_1 to 0.0001. Compute the LR.

h) * What can be said about the effect of the threshold parameter?

i) Change the *Threshold* parameter to 0.0001 and the *Step* parameter to 1 for H_1. Compute the LR.

j) Change the *Step* parameter to 2 for H_1 and compute the LR.

k) Change the genotypes for F1 for the locus DXS10101

l) * What can be said about the effect of the *Step* parameter?

We recommend keeping the value of *Threshold* high (e.g. 0.01-0.0001) and the value of *Step* low (e.g. 0 or 1) as this reduces the computation effort considerably compared to when we have lower values on the *Threshold* parameter and higher values on the *Step* parameter.