4 Solution for exercises in Chapter 4

Updated: April 27th 2021

Figures and text are reproduced in this file to make the text more comprehensive. Solutions are generally given in italics.

4.9 Solutions

4.9.1 Warm-up

We will first consider a small example to explore a DVI search. This exercise can be solved without access to a computer. To simplify, we can disregard any complicating factors (mutations, silent alleles, subpopulation structure, dropout/dropins etc).

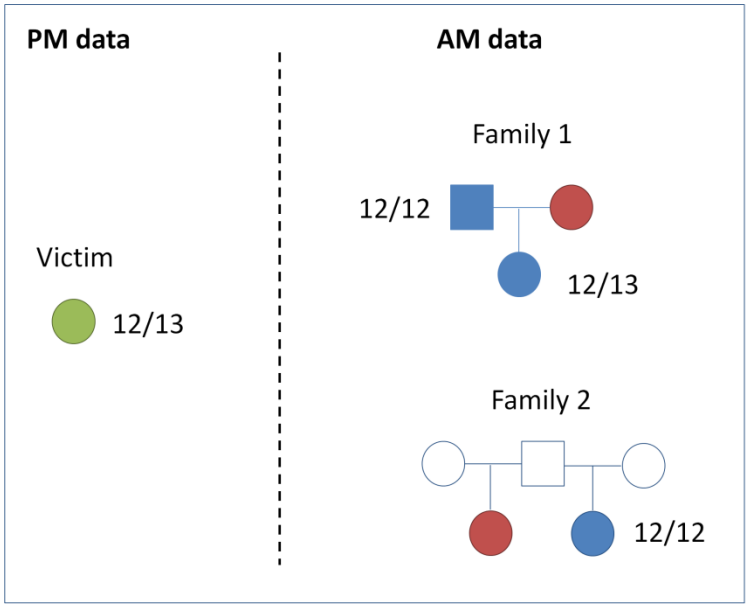
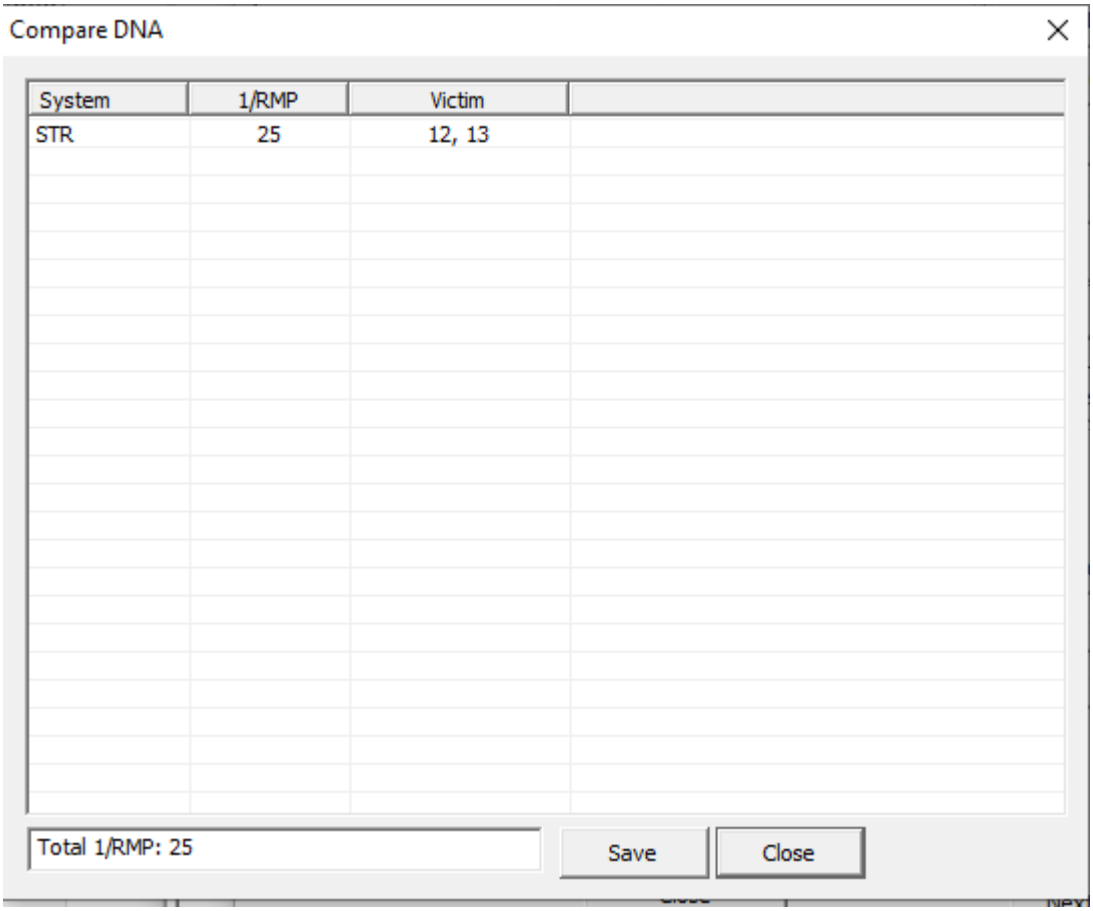


Figure 1. Illustration of the identification problem described in exercise 4.9.1. Data available for a single marker.

a) Consider data as given in **Figure 1**. The population frequency of allele 12 is 0.1 and allele 13 is 0.2. A personal belonging is found and genotyped as 12/13. Find the LR comparing the hypothesis H1: 'the belonging comes from the Victim' to
H2: 'the belonging comes from someone unrelated to the Victim'.
Hint: compute the 1/random match probability for the victim's genotypes.

*The solution is given by 1/(2*0.1*0.2)=25 and the interpretation is that if a personal belonging is available with the same genotype as the victim the likelihood ratio (LR) is 25 that the personal belonging and the victim is one and the same person. The drawback with using personal belongings is of course that an exclusion can be explained by a different owner of the personal belonging.*



b) Compute the likelihood ratio (LR) for the victim to belong to each of the two families respectively given the allele frequencies in a). Assume the alternative hypothesis is that the victim is unrelated to each family.

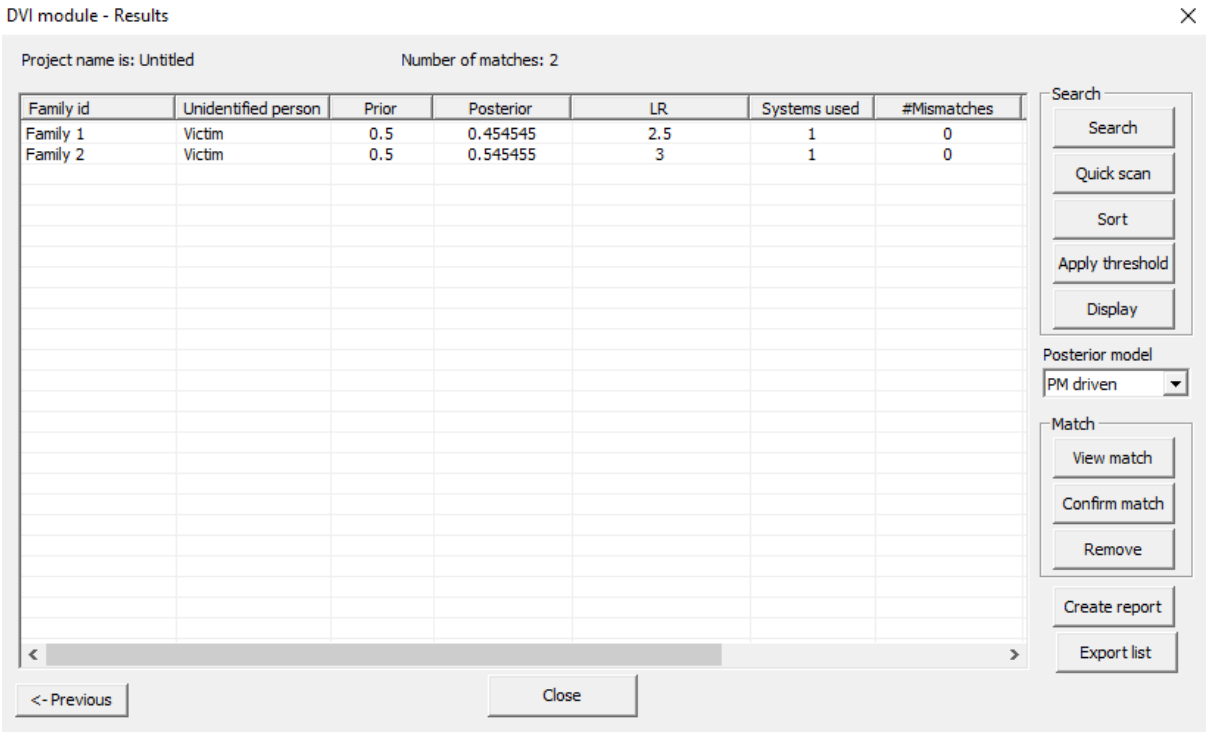
In the first family (Family 1) we have data for both the spouse and the child of the missing person. The spouse cannot be used to directly identify the missing person but can increase the power by giving clues to what the maternal/paternal alleles of the child are. The LR is computed as a regular trio case,

$$LR_{Victim,Family1} = \frac{\Pr(12/12)\Pr(12/13)\cdot 0.5}{\Pr(12/12)\Pr(12/13)p_{13}} = \frac{0.5}{p_{13}} = 2.5$$

In the second family we only have data for a paternal half sibling and the LR is computed as,

$$LR_{Victim,Family2} = \frac{0.5\Pr(12/12)p(13) + 0.5\Pr(12/12)\Pr(12/13)}{\Pr(12/12)\Pr(12/13)} = \frac{0.5 + p_{12}}{2p_{12}} = 3$$

Where we also note that X-chromosomal data is a strong candidate to increase the evidential weight.



c) Assume there is a prior probability of 0.3 that the victim belongs to Family 1 and 0.6 that the victim belongs to Family 2. Compute the posterior probability that the victims belongs to each of the two families respectively. Hint: Use the naïve method whereby each family is considered separately.

We use Bayes' formula and compute the posterior probability using the likelihood ratios from b) as,

$$\Pr(\text{Victim belongs to Family 1}) = \frac{0.3 \cdot LR_{Victim,Family1}}{0.3 \cdot LR_{Victim,Family1} + 0.6 \cdot LR_{Victim,Family2} + 0.1 \cdot 1} = \frac{0.75}{0.75 + 1.8 + 0.1} = 28\%$$

$$\Pr(\text{Victim belongs to Family 2}) = \frac{0.6 \cdot LR_{Victim,Family2}}{0.3 \cdot LR_{Victim,Family1} + 0.6 \cdot LR_{Victim,Family2} + 0.1 \cdot 1} = \frac{1.8}{0.75 + 1.8 + 0.1} = 68\%$$

$$\Pr(\text{Victim is unrelated to the families}) = \frac{0.1 \cdot 1}{0.3 \cdot LR_{Victim,Family1} + 0.6 \cdot LR_{Victim,Family2} + 0.1 \cdot 1} = 4\%$$

We have used what we in the book refer to as a PM-driven approach focusing on the victim since we had access to specific information about the prior location of the victim.

4.9.2 *Covering the basics of a DVI search*

Assume you are asked to do the identification following a helicopter accident. The example is illustrated in **Figure 4.2**. The exercise will cover the major steps in the DVI process with realistic forensic data. There are seven unidentified remains (denoted V1 through V7) and three reference families, in total four missing persons. All files are available at http://familias.name/BookKETP/Exercises/Ch4/Exercise_4_9_3.zip or following links from the main repository.

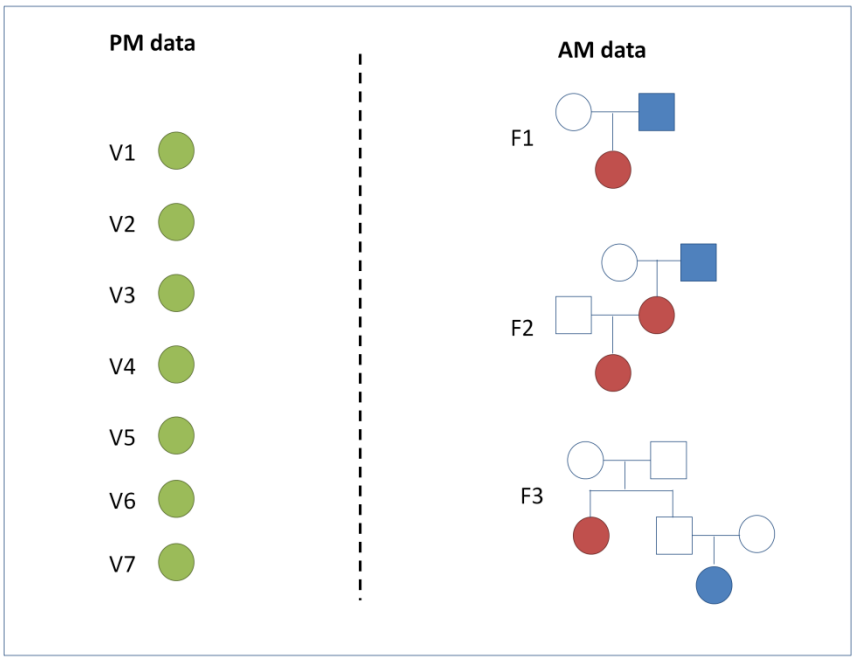


Figure 4.2. Illustration of pedigree data for exercise 0*. V1-V7 represent unidentified victims (PM data) while F1, F2, and F3 represent reference families where individuals highlighted in red are missing persons and individuals highlighted with blue are available reference persons.*

a) How many combinations of missing persons and unidentified remains can you enumerate? That is, what are exactly the number of possible combinations of victims and missing persons we need to investigate. V1=MP1 and V1=MP2 are two such combinations. Note, reference family F2 involves two missing persons.

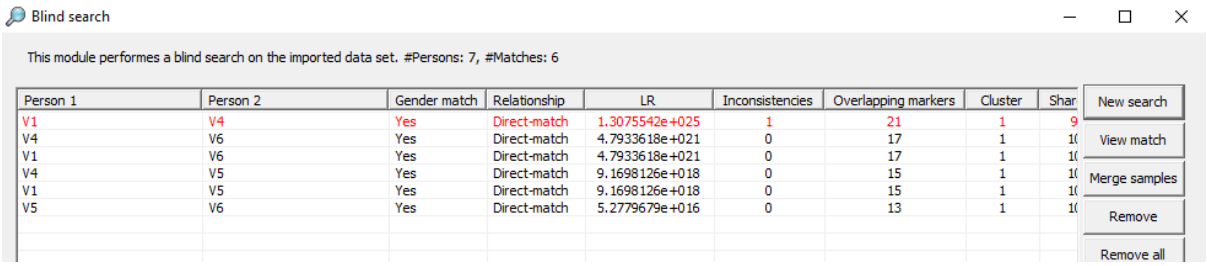
*In total seven victims and three families, one with two missing persons. In total this yields 7 * 4 = 28 combinations of victims and missing persons. If instead we simultanously fit victims with missing persons we get in total 209*

b) -

c) -

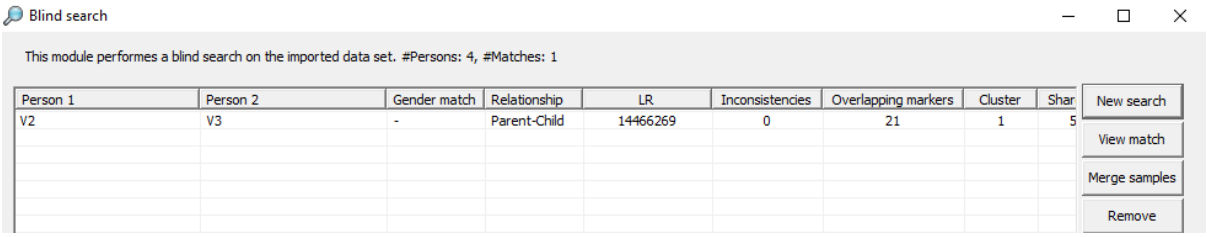
d) Conduct a blind search in the PM data set to reveal any identical remains. Merge the identical samples with a combined LR of more then 1,000,000. Use a dropout probability of 0.1. How many samples are merged?

Accepting all matches above 1,000,000 we can merge samples V1=V4 (with a single inconsistency), V4=V6 and V4=V5. Yielding in total four unique PM samples.



e) Next, conduct a new blind search in the PM data to identify parent/child relations. Use a LR threshold of 1000. How can the results be used?

A single match is obtained, indicating a parent/child relation between PM samples V2 and V3. Since we know that one of the reference families contain two missing persons, our finding could support the hypothesis that they do belong in this particular family.

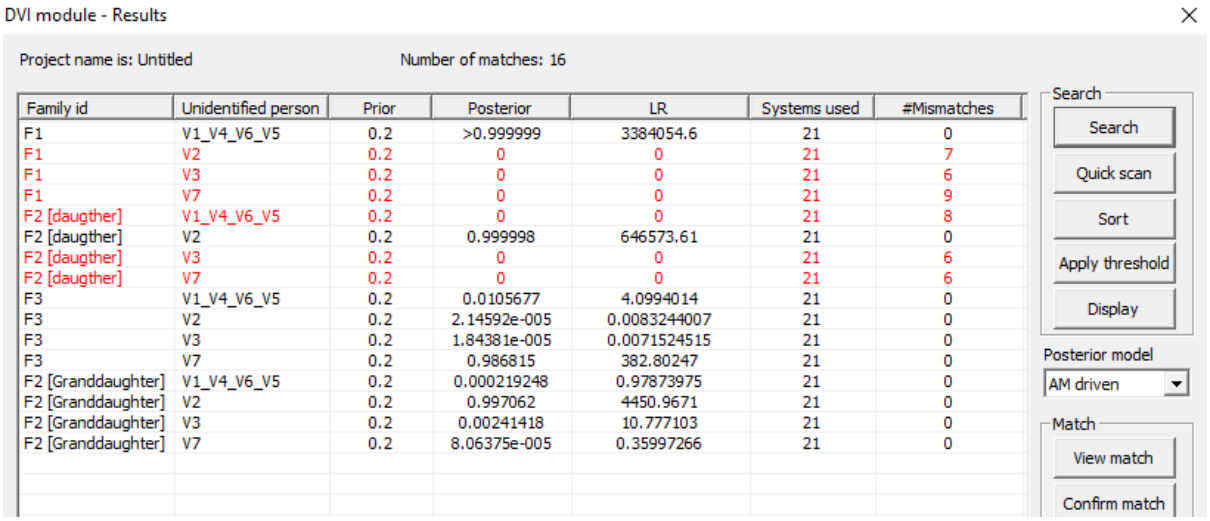


f) Import the AM data, given for three reference families illustrated in **Figure 4.2**.

The reference data is imported into Familias using the "Multiple families" options. Relationships should be defined automatically. Since the missing persons in Family 2 are to be treated separately (not jointly), we make a copy of Family 2 where one of the families indicate the connection between the Grandfather and his missing daughter and second between the Grandfather and his missing granddaughter.

g) Perform a search whereby each unidentified sample is compared to each of the reference families and a likelihood ratio for each comparison.

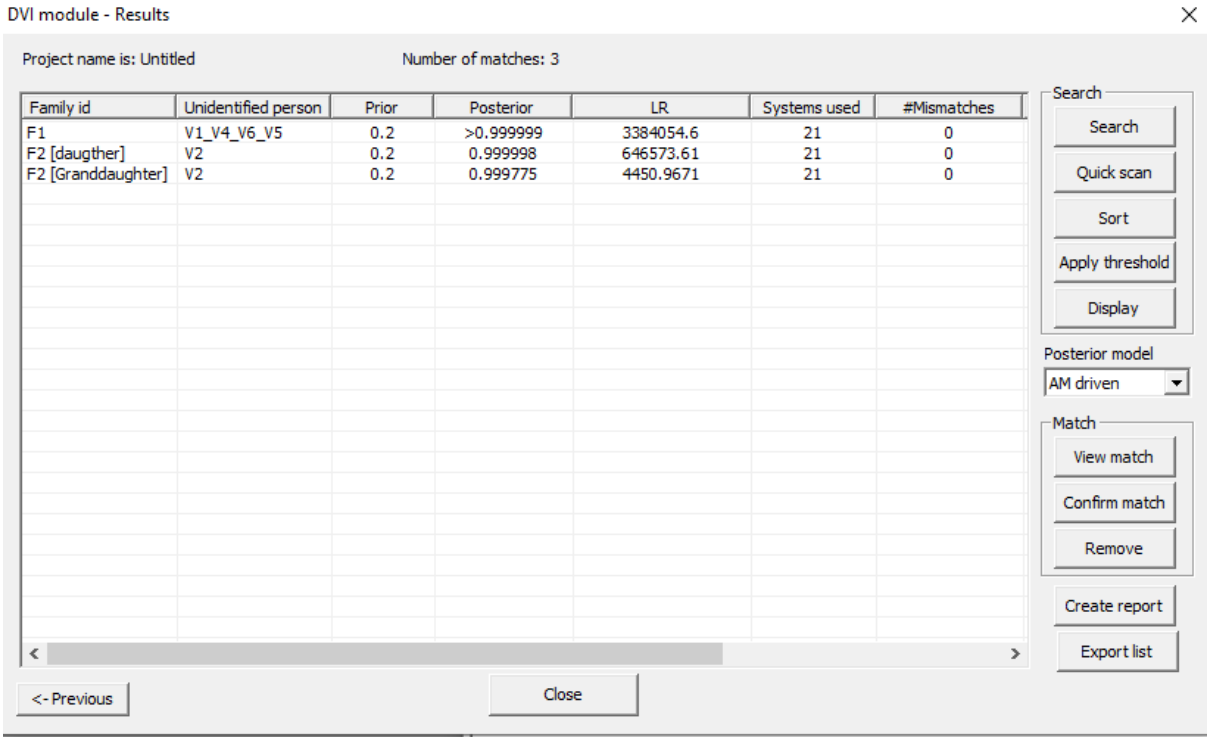
See illustration below



Family id	Unidentified person	Prior	Posterior	LR	Systems used	#Mismatches
F1	V1_V4_V6_V5	0.2	>0.999999	3384054.6	21	0
F1	V2	0.2	0	0	21	7
F1	V3	0.2	0	0	21	6
F1	V7	0.2	0	0	21	9
F2 [daugther]	V1_V4_V6_V5	0.2	0	0	21	8
F2 [daugther]	V2	0.2	0.999998	646573.61	21	0
F2 [daugther]	V3	0.2	0	0	21	6
F2 [daugther]	V7	0.2	0	0	21	6
F3	V1_V4_V6_V5	0.2	0.0105677	4.0994014	21	0
F3	V2	0.2	2.14592e-005	0.0083244007	21	0
F3	V3	0.2	1.84381e-005	0.0071524515	21	0
F3	V7	0.2	0.986815	382.80247	21	0
F2 [Granddaughter]	V1_V4_V6_V5	0.2	0.000219248	0.97873975	21	0
F2 [Granddaughter]	V2	0.2	0.997062	4450.9671	21	0
F2 [Granddaughter]	V3	0.2	0.00241418	10.777103	21	0
F2 [Granddaughter]	V7	0.2	8.06375e-005	0.35997266	21	0

h) Report identifications using an LR threshold of 1000.

See illustration below. Victim V1 (merged with V4, V5 and V6) belongs to F1, V2 to F2 either as the daughter or the granddaughter in F2.



Family id	Unidentified person	Prior	Posterior	LR	Systems used	#Mismatches
F1	V1_V4_V6_V5	0.2	>0.999999	3384054.6	21	0
F2 [daugther]	V2	0.2	0.999998	646573.61	21	0
F2 [Granddaughter]	V2	0.2	0.999775	4450.9671	21	0

i) Report potential further testing that could resolve the cases where the LR is below 1000.

We note that V2 and V3 have a high LR of being in a parent/child relation and V3 has a LR of 10 to belong in F2. This could be used to compute a new LR for the complete set of individuals. We will later return to what we call a global soluation including such problems.We further note that V7 has a LR of roughly 382 to belong to F3. Additional

autosomal markers as well as X-chromosomal markers could potentially increase the strength to provide a conclusive result.

4.9.3 On the use of thresholds

This exercise will deal with problems related to the choice of threshold in a mass identification. We will divide the exercise into likelihood ratio thresholds and posterior probability thresholds. Consider the pedigree in **Figure 4.3**.

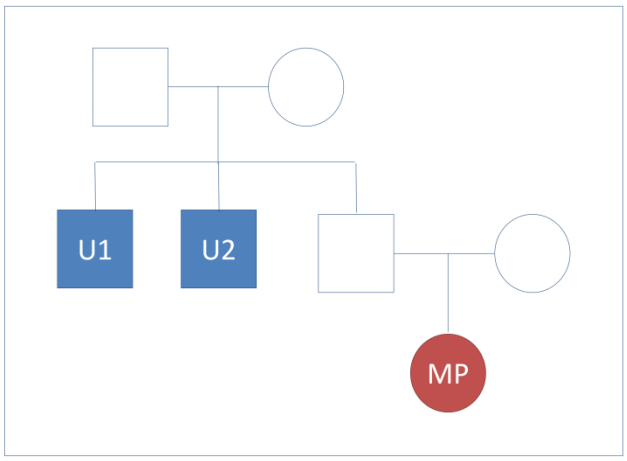


Figure 4.3. Illustration of pedigree with two uncles (U1 and U2) of a missing person (MP).

We will evaluate expectations and appropriate case thresholds using data for either a single uncle (U1) of the missing person (MP) or using data for two uncles (U1 and U2) of the missing person (MP). Data is simulated using a set of 16 STR markers. For each simulation we compute the likelihood ratio and summarize some of the results in **Figure 4.4** below.

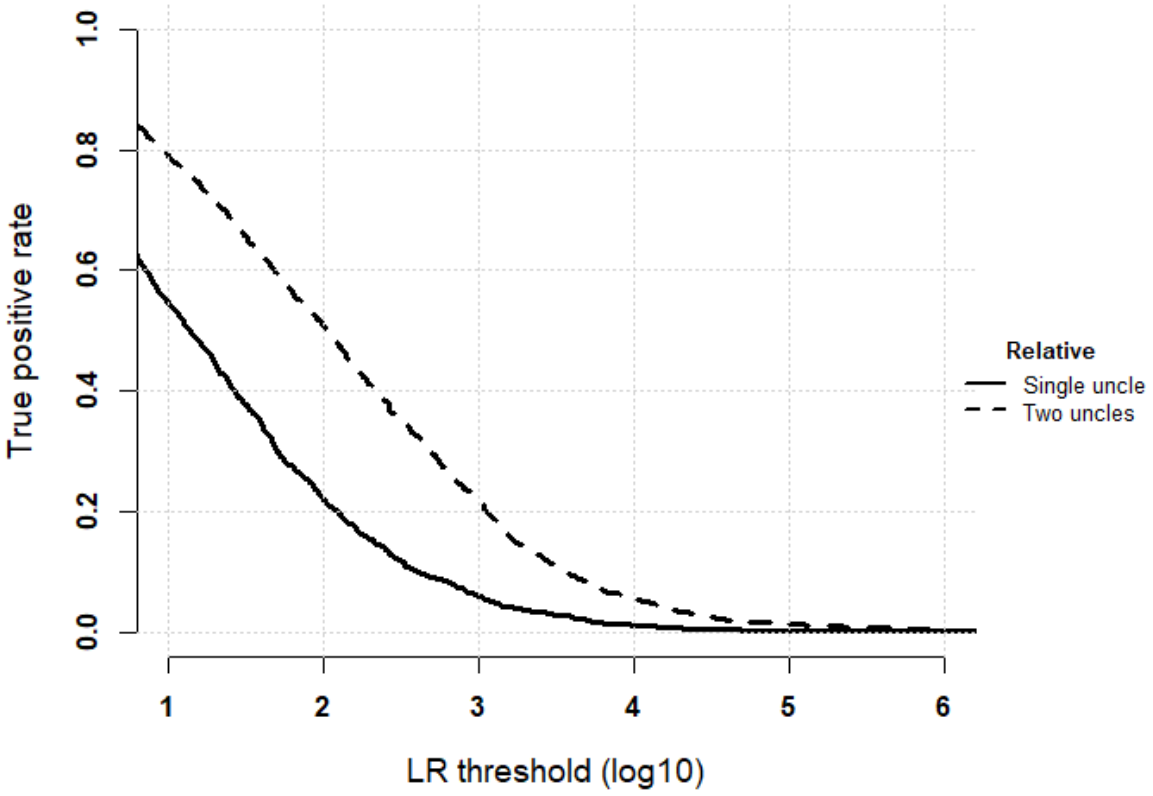


Figure 4.4. Illustration of true positive rates for data simulated using a single uncle and two uncles as references of a missing person.

a) What is the probability (approximately) that we will find the missing person given a LR threshold of 100 using a single uncle? That is, the probability that an uncle can be used to identify a niece/nephew.

The probability can be found where the curve entitled "Single uncle" intersects with 2 on the x-axis. This is approximately 0.22 (or 22%). The interpretation is that we will be able to successfully identify a victim using a single uncle in only 22% of the identifications (given our 16 STR markers and the LR threshold).

b) What is the probability that we will find the missing person given a LR threshold of 100 using a two uncles? That is, the probability that two uncle can be used to identify a niece/nephew.

The probability can be found where the curve entitled "Two uncles" intersects with 2 on the x-axis. This is approximately 0.5 (or 50%). The interpretation is that we will be able to successfully identify a victim using two uncles in 50% of the identifications (given our 16 STR markers and the LR threshold).

c) Use the information in a) to answer what posterior thresholds this translates to for priors of 1/10, 1/100 and 1/1000. For simplicity we can assume there are only two competing hypotheses that we need to consider, so a LR of 1000 and a prior of 1/10 equals a posterior probability of $(1000 \cdot 1/1000)/(1000 \cdot 1/1000 + 1 \cdot 999/1000) = 0.5$

For the single uncle case, this translates to a 22% probability to exceed 99%, 50% and 9% posterior probability respectively.
For the two uncles case, this translates to a 50% probability to exceed 99%, 50% and 9% posterior probability respectively.

Now consider the illustration in **Figure 4.5**.

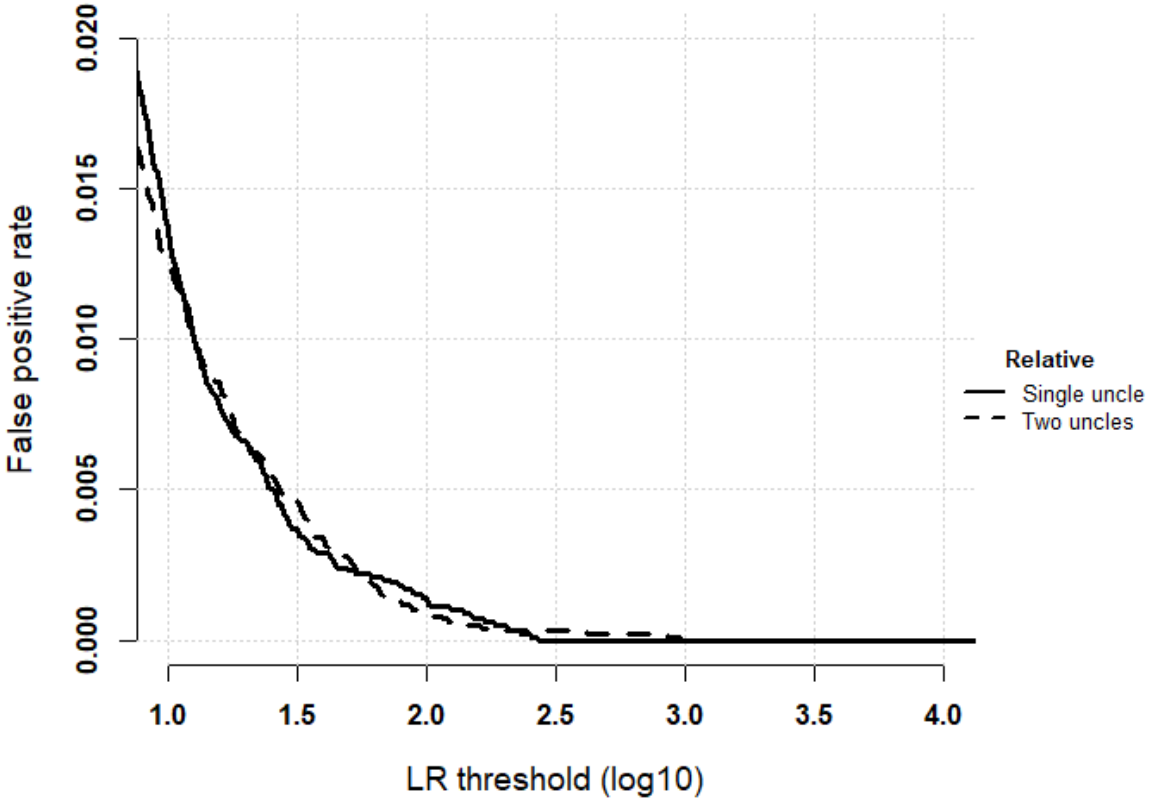


Figure 4.5. Illustration of false positive rates for data simulated using a single uncle and two uncles as references of a missing person.

d) What is the probability that we will falsely include an unrelated individual as the missing person given a LR threshold of 100 when using a single uncle?

Using a single uncle the value is found at the intersection between the curve entitled "Single uncle" and 2 on the x-axis. It is approximately 0.2%.

e) What is the probability that we will falsely include an unrelated individual as the missing person given a LR threshold of 100 when using two uncles?

Using a single uncle the value is found at the intersection between the curve entitled "Single uncle" and 2 on the x-axis. It is approximately 0.2%. Only slightly lower than for a single uncle.

4.9.4 Exploring the potential of screening

Consider data for eight victims and eight different reference families, data are given in files where all samples have been genotyped for 16 autosomal STR markers. The reference families are shown in **Figure 4.6**. All files are available at http://familias.name/BookKETP/Exercises/Ch4/Exercise_4_9_4.zip or following links from the main repository.

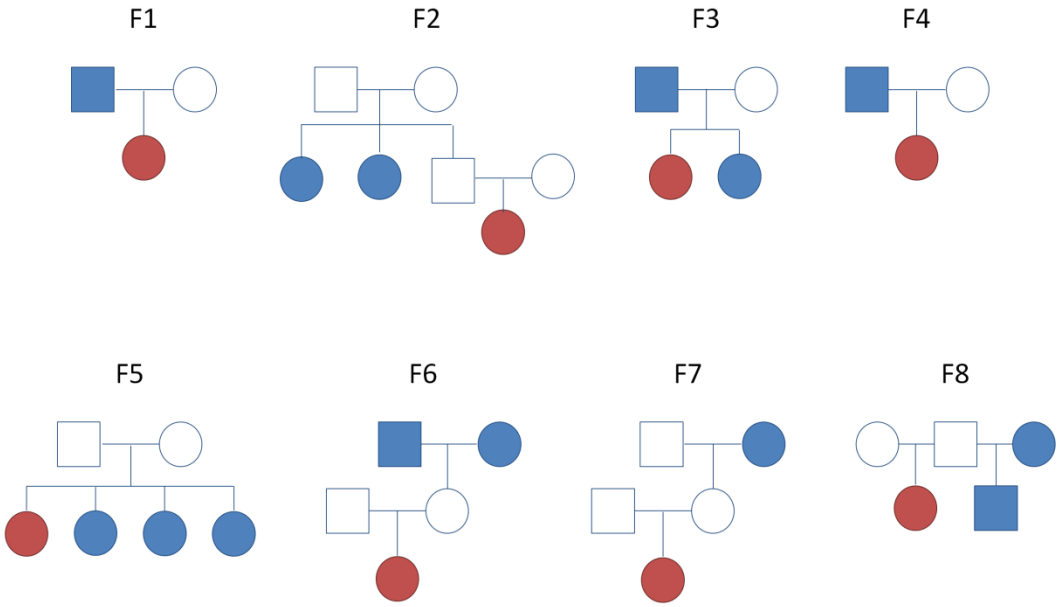


Figure 4.6. Illustration of pedigree data for exercise 4.9.4*. Individuals highlighted in red are missing persons and individuals highlighted with blue are available reference persons.*

a) Import the population frequency data file with data given for 16 autosomal STR markers. Mutation rates are 0.001 for all markers and we may use a simple mutation model (equal probability for all mutations). Other complicating factors can be disregarded.

b) Perform a blind search among the victims to find if there are related individuals in the PM data. Use the likelihood ratio (LR) as a measure of the weight of evidence and set the threshold at 100 to include a candidate as a potential relative. Search for parent/child and sibling relations.

   Results are visualized below and show two matches, both between V1 and V2. LR are both in favor of parent/child relation as well as siblings.

This module performes a blind search on the imported data set. #Persons: 8, #Matches: 2

Person 1	Person 2	Gender match	Relationship	LR	Inconsistencies	Overlapping markers
V1	V2	-	Parent-Child	477481.72	0	16
V1	V2	-	Siblings	21394.745	NA	16

The sharing pattern (see below) might suggest that parent/child is the true relation. We note the connection for subsequent searches and also note that none of the pedigrees suggest a link between two missing persons.

IBS=2	IBS=1	IBS=0
31.3%	68.8%	0.0%
31.3%	68.8%	0.0%

c) How could a set of relatives in the PM data affect the subsequent DVI search?

Knowing that there are relatives among the victims is essential to avoid misidentifications (one person is wrongly identified as his relative). Also, this information can help to increase the chances of identifications, if for instance a pair of relatives is taken into account at the same time.

d) * Evaluate the reference families

i. Find any inconsistencies in the data

ii. Perform simulations (unconditional) to evaluate the potential of the given relatives for each family. Report the probability that the LR will exceed 100 for each set of relatives. Use the frequency data supplied in the exercise.

iii. * Perform conditional simulations to evaluate the potential of each reference family, report the probability that the LR will exceed 100 for each family. Use the frequency data supplied in the exercise.

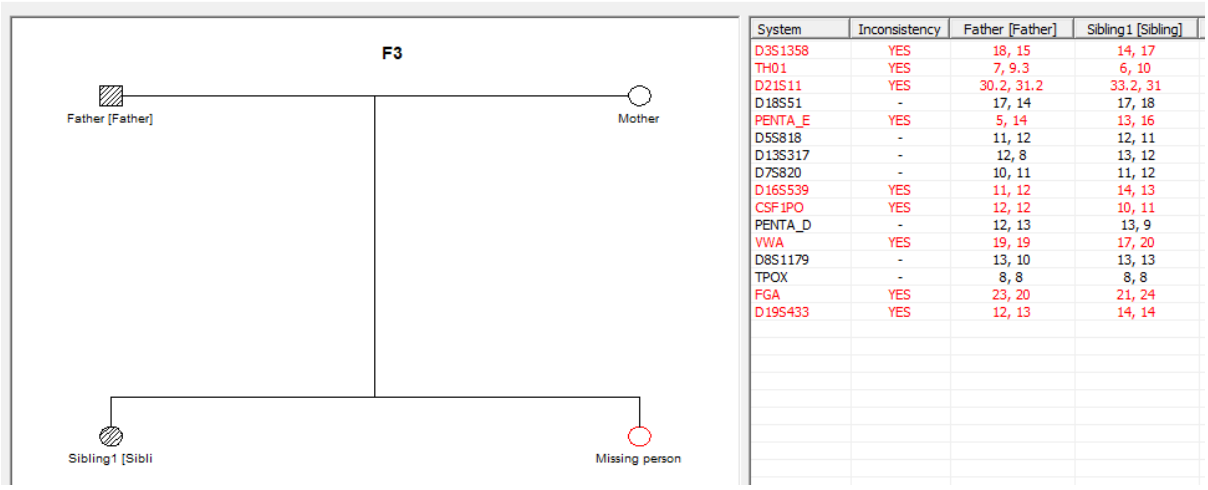
We first import the AM data using the "Multiple families" option in Familias. The software recognizes the relationship tags, but some manual re-arrangement is necessary. In F8, the relation between the Mother and the Missing person has to be removed.

Next we evaluate the reference data and find that F3 shows inconsistencies, see illustration below. Father cannot be the father of sibling 1. Therefore, there is uncertainty in this pedigree and we need to account for this in the subsequent search.

Options:

- *F3-1: Father is the real father of the MP, sibling 1 is unrelated to MP*
- *F3-2: Father is the real father of the MP, sibling1 is a maternal half-sib of the MP*
- *F3-3: Sibling 1 is the real sibling of the MP, father is not related to the MP*
- *F3-4: Sibling1 is a maternal half-sib of the MP, father is not related to the MP*

View pedigree



System	Inconsistency	Father [Father]	Sibling1 [Sibling]
D3S1358	YES	18, 15	14, 17
TH01	YES	7, 9.3	6, 10
D21S11	YES	30.2, 31.2	33.2, 31
D18S51	-	17, 14	17, 18
PENTA_E	YES	5, 14	13, 16
D5S818	-	11, 12	12, 11
D13S317	-	12, 8	13, 12
D7S820	-	10, 11	11, 12
D16S539	YES	11, 12	14, 13
CSF1PO	YES	12, 12	10, 11
PENTA_D	-	12, 13	13, 9
VWA	YES	19, 19	17, 20
D8S1179	-	13, 10	13, 13
TPOX	-	8, 8	8, 8
FGA	YES	23, 20	21, 24
D19S433	YES	12, 13	14, 14

We redefine F3 according to the illustration below yielding in total four new pedigrees.

c) Evaluate the reference families:

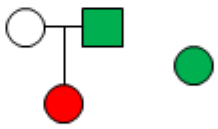
i. Find if there are any inconsistencies in the data

F3

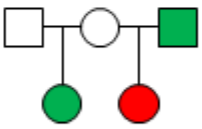
Reference family	#Typed persons	#Markers	Inconsistencies
F3-1	1	16	0
F3-2	2	16	0
F3-3	1	16	0
F3-4	1	16	0

- Redefine F3 taking into account all the possible options:
  - F3-1: Father is the real father of the MP, sibling 1 is unrelated to MP
  - F3-2: Father is the real father of the MP, sibling1 is a maternal half-sib of the MP
  - F3-3: Sibling 1 is the real sibling of the MP, father is not related to the MP
  - F3-4: Sibling1 is a maternal half-sib of the MP, father is not related to the MP

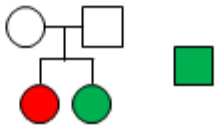
F3-1



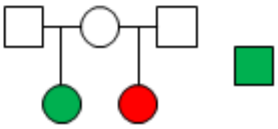
F3-2



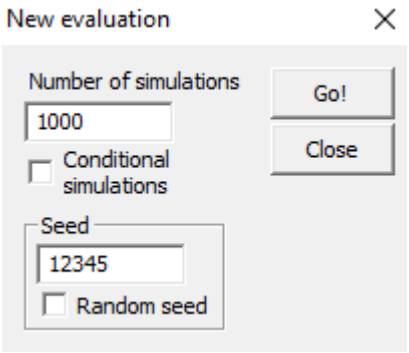
F3-3



F3-4



We next proceed to perform unconditional simulations for the reference families. The exact results depend on the seed (we use 12345) and the number of simulations (we use 1000).



The unconditional simulations will evaluate the potential of the specific set of relatives and the number of typed genetic markers. However, it will not take into account the available genotypes. The results are illustrated below and shows that F2 ,F6, F7, F8 as well as F3-4 (see pedigree above) have low probability to yield LR above 1000 (see column with P(LR>1000)). In addition, F7, F8 as well as F3-3 and F3-4 will never be able to exclude an unrelated individual (Exclusion probability=0).

Reference family evaluation tool

Reference family	#Typed per...	#Markers	Inconsistenc...	P(LR>100)	P(LR>1000)	P(LR>10000)	Exclusion probability
F1	1	16	0	0.995000	0.979000	0.888000	1.000000
F2	2	16	0	0.472000	0.204000	0.061000	0.313000
F3-1	2	16	0	0.996000	0.984000	0.895000	1.000000
F4	1	16	0	0.995000	0.979000	0.888000	1.000000
F5	3	16	0	1.000000	0.999000	0.998000	1.000000
F6	2	16	0	0.828000	0.530000	0.228000	0.988000
F8	2	16	0	0.457000	0.203000	0.076000	0.009000
F7	1	16	0	0.240000	0.070000	0.014000	0.000000
F3-2	2	16	0	0.999000	0.996000	0.968000	1.000000
F3-3	2	16	0	0.866000	0.762000	0.593000	0.000000
F3-4	2	16	0	0.240000	0.065000	0.016000	0.000000

We next proceed to perform conditional simulations for the reference families. The exact results depend on the seed (we use 12345) and the number of simulations (we use 1000).

The conditional simulations will evaluate the potential of the specific set of relatives and the number of typed genetic markers. In addition, they will account for the available genotypes. The results are illustrated below and shows that F2 ,F6, F7, F8 as well as F3-4 (see pedigree above) have low probability to yield LR above 1000. The results are similar to the unconditional simulations from a general perspective. However, closer investigation reveals that for instance the P(LR>1000) for F7 is greatly reduced (from 7% to 1.8%) and similarly increase for F1 (97.9% to 100%).

Reference family	#Typed per...	#Markers	Inconsistenc...	P(LR>100)	P(LR>1000)	P(LR>10000)	Exclusion probability
F1	1	16	0	1.000000	1.000000	0.996000	0.999993
F2	2	16	0	0.336000	0.081000	0.010000	0.087464
F3-1	2	16	0	1.000000	0.998000	0.924000	0.999960
F4	1	16	0	1.000000	0.976000	0.784000	0.999876
F5	3	16	0	1.000000	0.998000	0.997000	0.999873
F6	2	16	0	0.885000	0.614000	0.264000	0.991635
F8	2	16	0	0.617000	0.365000	0.171000	0.000000
F7	1	16	0	0.122000	0.018000	0.000000	0.000000
F3-2	2	16	0	0.998000	0.992000	0.962000	0.999960
F3-3	2	16	0	0.922000	0.815000	0.693000	0.000000
F3-4	2	16	0	0.339000	0.110000	0.018000	0.000000

e) Perform a screening (each PM sample versus each AM sample) whereby no relations within the reference families are considered. Screen for parent/child and sibling relations using a LR threshold of 100.

The results are visualized below. The screening is useful to detect non-paternity cases between PM and Pedigrees. We note that families F1-F6 have a match with V1-V6 (correctly).The results further suggest that in F3, the father might not be the father of V3 (correctly).

Family id	Unidentified person	LR	Systems used	#Mismatches	Pedigree
Father [Father] (F1)	V1 (Parent-Child)	3672946.7	16	0	Parent-Child
Father [Father] (F1)	V1 (Siblings)	115666.64	16	0	Siblings
Uncle2 [Uncle] (F2)	V2 (Parent-Child)	608.00518	16	0	Parent-Child
Sibling1 [Sibling] (F3-1)	V3 (Siblings)	150721.09	16	0	Siblings
Sibling1 [Sibling] (F3-2)	V3 (Siblings)	150721.09	16	0	Siblings
Sibling1 [Sibling] (F3-3)	V3 (Siblings)	150721.09	16	0	Siblings
Sibling1 [Sibling] (F3-4)	V3 (Siblings)	150721.09	16	0	Siblings
Mother [Mother] (F4)	V4 (Parent-Child)	420597.4	16	0	Parent-Child
Mother [Mother] (F4)	V4 (Siblings)	9556.7634	16	0	Siblings
Sibling1 [Sibling] (F5)	V5 (Parent-Child)	1412.9954	16	1	Parent-Child
Sibling1 [Sibling] (F5)	V5 (Siblings)	606763.72	16	0	Siblings
Sibling2 [Sibling] (F5)	V5 (Parent-Child)	5341617.7	16	0	Parent-Child
Sibling2 [Sibling] (F5)	V5 (Siblings)	7451825	16	0	Siblings
Sibling3 [Sibling] (F5)	V5 (Parent-Child)	7363614.1	16	0	Parent-Child
Sibling3 [Sibling] (F5)	V5 (Siblings)	71576448	16	0	Siblings
Grandmother [Grandmother] (F6)	V6 (Siblings)	1779.5662	16	0	Siblings

f) Perform a full search using a LR threshold of 100. Compare the results to d) and discuss the importance of the screening step.

The results are visualized below. In comparison to the results in e), we see that F8 matches (correctly) with V8, although the LR is comparatively low at 185. We note that the match between F2 and V2, appearing in the screening procedure, is missing suggesting some error in the specification of the relationship. We see that in F5 (three siblings available), there is a mismatch, but the final LR is still high suggesting a mutation as the explanation.

Project name is: Untitled Number of matches: 7

Family id	Unidentified person	LR	Systems used	#Mismatches	Pedigree
F1	V1	3672320.3	16	0	Missing person
F4	V4	426143.59	16	0	Missing person
F5	V5	6.1134078e...	16	1	Missing person
F6	V6	5555.0323	16	0	Missing person
F8	V8	185.76486	16	0	Missing person
F3-3	V3	150721.09	16	0	Missing person
F3-4	V3	3625.4199	16	0	Missing person

Running a search with LR=10 as threshold we note that also F7 matches (correctly) with V7, however the LR is only at 82, suggesting that additional relatives should be typed (or more

genetic markers). This is in line with our previous simulations. We also note that the match between F2 and V2 appears.

Project name is: Untitled Number of matches: 9

Family id	Unidentified person	LR	Systems used	#Mismatches	Pedigree
F1	V1	3672320.3	16	0	Missing person
F2	V2	66.077932	16	0	Missing person
F4	V4	426143.59	16	0	Missing person
F5	V5	6.1134078e...	16	1	Missing person
F6	V6	5555.0323	16	0	Missing person
F8	V8	185.76486	16	0	Missing person
F7	V7	82.574902	16	0	Missing person
F3-3	V3	150721.09	16	0	Missing person
F3-4	V3	3625.4199	16	0	Missing person

4.9.5 The impact of posterior models

In this exercise we will consider some different DVI scenarios and evaluate different models for posterior probabilities. For each scenario, compute the posterior probability given the different models (One-to-one, PM-driven, AM-driven) discussed in this chapter. We will later return to the global model. The likelihood ratios are given as tables in the online files. For computational reasons we assume all non-reported combinations of missing person and victim have a LR of zero. No information is given with regards to what relatives are genotyped or if there are any suspected relatives among the victims. We can assume each missing person is represented by exactly one reference family. All files are available at http://familias.name/BookKETP/Exercises/Ch4/Exercise_4_9_5.zip or following links from the main repository.

a) Consider a small plane crash with 10 unidentified remains and only two missing persons. 8 missing persons are unaccounted for.

In this first scenario, the number of PM samples exceeds the number if missing persons accounted for. That is, there are unknown missing persons where we do not have access to reference data. The prior is 1/11. We note similar results between the models except for the match between V1 and F1 where the posterior is high for the One-to-one and PM-driven model whereas the AM-driven model gives this match a lower posterior owing to a much higher LR between V1 and F1. Also V2 and V3 seems to be identical or at least the LR is identical where the AM driven model divides the posterior onto these two while the other models yields high posteriors for both identifications.

			Posterior		
PM sample	**AM sample**	**LR**	**One-to-one**	**AM driven**	**PM driven**
V1	F1	10000	0.9990	0.9894	0.9991
V5	F1	6	0.3750	0.0006	0.4000
V10	F1	100	0.9091	0.0099	0.9174
V2	F2	600	0.9836	0.4950	0.9852
V3	F2	600	0.9836	0.4950	0.9852
V8	F2	8	0.4444	0.0066	0.4706
V9	F2	3	0.2308	0.0025	0.2500

	Prior:	0.09	0.09	0.09

b) Consider a car crash with four unidentified victims and four missing persons. All missing persons are accounted for.

In this second scenario, the number of AM samples and PM samples are identical. The prior is 1/5. We note that the match between V1 and F3 (LR=100) yields a high posterior in the One-to-one as well as the AM-driven models whereas in the PM-driven model, V1 fits much better in F1. Similar argument for the match between V3 and F2 but where the AM-driven model finds a better match between V2 and F2.

			Posterior		
PM sample	**AM sample**	**LR**	**One-to-one**	**AM driven**	**PM driven**
V1	F1	100000	0.99996	0.99999	0.998991
V1	F3	100	0.961538	0.900901	0.000999
V2	F2	1000	0.996016	0.949668	0.999001
V3	F2	50	0.925926	0.047483	0.819672
V3	F3	10	0.714286	0.09009	0.163934
V4	F2	2	0.333333	0.001899	0.037736
V4	F4	50	0.925926	0.980392	0.943396
		Prior	0.20	0.20	0.20

c) Consider an avalanche accident where only two unidentified remains are recovered and 10 persons are reported as missing. All missing persons are accounted for.

In this third scenario, the number of AM samples exceeds the number if unidentified remains. The prior is 1/11. We see similar results but note that the AM driven model gives a fairly high posterior 69% to V2=F2 whereas it is more probable that V2=F3 (96% posterior).

			Posterior		
PM sample	**AM sample**	**LR**	**One-to-one**	**AM driven**	**PM driven**
V1	F1	1000	0.99009901	0.9910803	0.9861933
V2	F2	20	0.666666667	0.6896552	0.0623053
V1	F3	3	0.230769231	0.0096154	0.0029586
V2	F3	300	0.967741935	0.9615385	0.9345794
V1	F10	10	0.5	0.5263158	0.0098619
		Prior	0.09	0.09	0.09

4.9.6 The power of relatives

Consider a missing person database with 10,000 individuals and 100 reference families (missing persons). Two siblings of a missing person are available in all 100 reference families. Data has been generated using a standard set of 16 autosomal STR markers. All files are available at http://familias.name/BookKETP/Exercises/Ch4/Exercise_4_9_6.zip or following links from the main repository.

a) In a screening procedure (pairwise AM versus PM data), how many comparisons are conducted in total given that each reference family contains two reference relatives.

In total, 10,000 x 100 x 2 = 2,000,000 comparisons are performed.

b) In a complete search, how many comparisons are conducted in total? (Given an all-versus-all search).

In total, 10,000 x 100 = 1,000,000 comparisons are performed.

c) If we assume the false positive rate is 0.00001 and false negative rate is 0.01 for an inclusion threshold of LR = 1000, what is the expected number of false positive candidates in both a screening and a complete search? Note that we can generally expect a higher false positive/negative rate when the screening procedure is performed since less information (fewer relatives) is leveraged.

We expect in total 2,000,000 x 0.00001 = 20 and 1,000,000 x 0.00001 = 10 false positives given these conditions for the two searches respectively. We expect in total 200 x 0.01 = 2 and 100 x 0.01 = 1 false negatives given these conditions for the two searches respectively.

d) Frequency data are available for the 16 STR markers, further define the mutation rates as uniform with rates 0.0001 for all markers and across genders, etc. Use the simplest mutation model (equal probability for all mutations) for computational reasons. Open the frequency database, import the file containing the PM data as well as the file containing the AM data. Check that all reference families have been correctly defined. Each family is represented by two full siblings of a missing person

-

e) Conduct a screening procedure where each sibling is compared to each victim individually as well as a complete search where the power of both siblings in the 100 reference families is used. -Use LR = 1000 as inclusion threshold to retain a match.

Top results from screening procedure is visualized below.

Family id	Unidentified person	Prior	Posterior	LR
Sibling1 [Brother] (F1)	V1 (Siblings)	1.00E-04	0.999955	2.24E+08
Sibling2 [Brother] (F1)	V1 (Siblings)	1.00E-04	0.999303	1.43E+07
Sibling1 [Brother] (F2)	V2 (Siblings)	1.00E-04	0.999043	1.04E+07
Sibling1 [Brother] (F3)	V3 (Siblings)	1.00E-04	0.299189	4269.18
Sibling2 [Brother] (F3)	V3 (Siblings)	1.00E-04	0.964529	271924
Sibling2 [Brother] (F4)	V4 (Siblings)	1.00E-04	0.996411	2.78E+06
Sibling2 [Brother] (F5)	V5 (Siblings)	1.00E-04	0.982562	563470
Sibling2 [Brother] (F6)	V6 (Siblings)	1.00E-04	0.811702	43107.2
Sibling1 [Brother] (F7)	V7 (Siblings)	1.00E-04	0.984118	619640
Sibling2 [Brother] (F7)	V7 (Siblings)	1.00E-04	0.839418	52273.4
Sibling1 [Brother] (F8)	V8 (Siblings)	1.00E-04	0.832858	49829.5
Sibling2 [Brother] (F8)	V8 (Siblings)	1.00E-04	0.984041	616589
Sibling1 [Brother] (F10)	V10 (Siblings)	1.00E-04	0.998601	7.14E+06
Sibling2 [Brother] (F10)	V10 (Siblings)	1.00E-04	0.963912	267099
Sibling1 [Brother] (F11)	V11 (Siblings)	1.00E-04	0.999905	1.05E+08
Sibling2 [Brother] (F11)	V11 (Siblings)	1.00E-04	0.919465	114170

Top results from a complete search is visualized below.

Family id	Unidentified person	Prior	Posterior	LR
F1	V1	1.00E-04	1	9.04E+12
F2	V2	1.00E-04	1	1.70E+09
F3	V3	1.00E-04	1	3.08E+07
F4	V4	1.00E-04	1	2.76E+06
F5	V5	1.00E-04	1	1.29E+07
F6	V6	1.00E-04	1	2.37E+08
F7	V7	1.00E-04	1	4.06E+08
F8	V8	1.00E-04	1	5.31E+09
F9	V9	1.00E-04	1	2.98E+06
F10	V10	1.00E-04	1	2.00E+08
F11	V11	1.00E-04	1	1.05E+11
F12	V12	1.00E-04	1	2.47E+07
F13	V13	1.00E-04	1	4.92E+09
F14	V14	1.00E-04	1	4.61E+08
F15	V15	1.00E-04	1	1.49E+10
F16	V16	1.00E-04	1	3.34E+07
F17	V17	1.00E-04	1	1.95E+11
F18	V18	1.00E-04	0.999999	1.30E+11
F18	V711	1.00E-04	9.50E-07	123474
F19	V19	1.00E-04	1	6.01E+06
F20	V20	1.00E-04	1	2.92E+09

f) Report the number of matches exceeding LR=1000 in both searches. Assume that the correct solution is given by V1=MP1...V100=MP100. How does this number correspond with the expected number in (c)?

We can process the output in Excel or R. We find that in the screening procedure, where we expect 200 matches (e.g. Sibling 1 in F1 should match V1 as well as Sibling 2 in F1). We find 163 true matches indicating 37 missed matches (false negatives), compared to 2 expected and 124 false matches compared to 20 expected. In the complete search we find all true matches (100 in total) compared to one expected missed match (false negative) and five false matches, compared to10 expected. The results are in line with expectations since the screening procedure only leverage the information from one of the relatives and the true false positive/negative rates are most likely higher then stated in the exercise when a single sibling is used.

*4.9.7 * The global solution*

a) *The commands and output:*

```
> library(dvir)
> data(dataExercise497)
> pm = dataExercise497$pm
> am = dataExercise497$am
> missing = dataExercise497$missing
> res = jointDVI(pm, am, missing, verbose = F)
> res
   V1  V2  V3    loglik           LR   posterior
1 MP1 MP2 MP3 -502.9837 4.286272e+21 0.994957849
2 MP1 MP2   * -508.2686 2.172155e+19 0.005042151
```

The number of solutions is found through typing the commands

```
library(dvir)
ncomb(3,3,0,0)

[1] 34
```

A direct argument confirms this

$$\binom{3}{0}\binom{3}{0}0! + \binom{3}{1}\binom{3}{1}1! + \binom{3}{2}\binom{3}{2}2! + \binom{3}{3}\binom{3}{3}3! = 1 + 9 + 18 + 6 = 34$$

The numbers 1, 9, 27, 6 correspond to respectively 0, 1, 2, 3 persons being identified.

```
jointDVI(from, to, ids.to)   V1  V2  V3    loglik           LR   posterior

1 MP1 MP2 MP3 -502.9837 4.286272e+21 0.994957849

2 MP1 MP2   * -508.2686 2.172155e+19 0.005042151
```

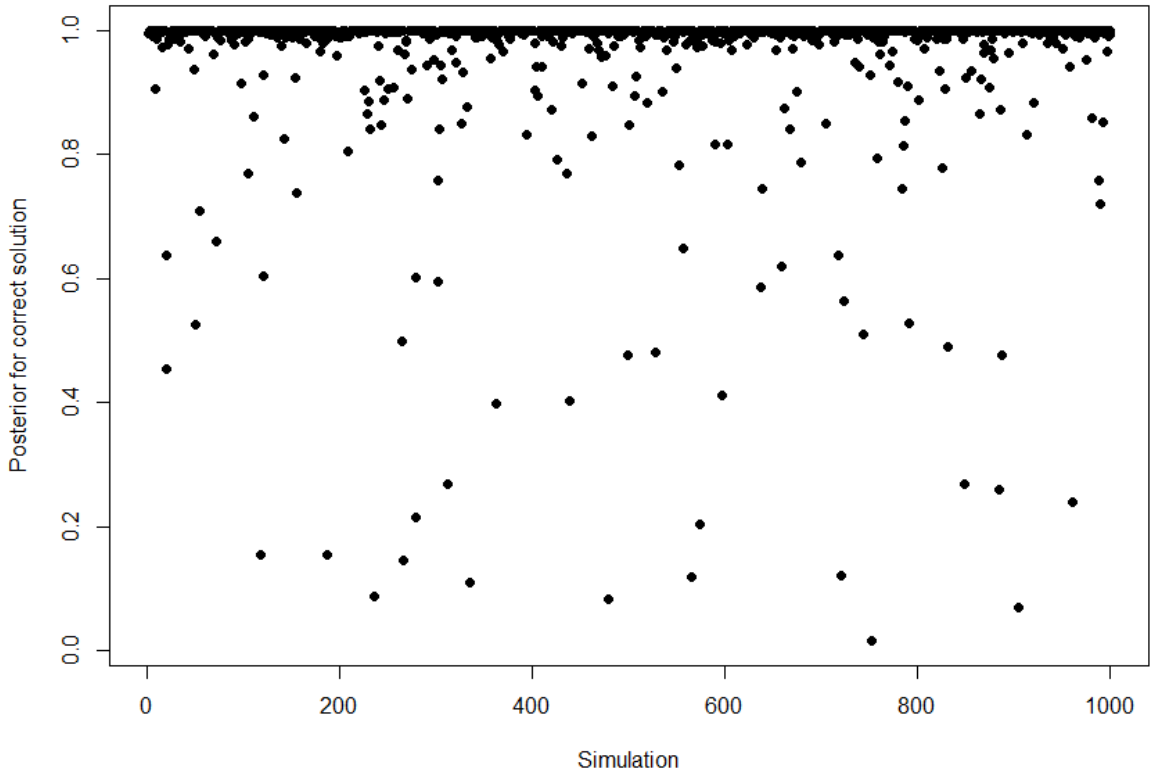
We see that the posterior for V1 = MP1, V2 = MP2 and V3 = MP3 is 0.995.

As an alternative we may use R Familias, the solution is provided in the ad-hoc script *http://familias.name/BookKETP/Exercises/Ch4/Exercise4_9_7a_solution.r* which will produce similar results as above.

b) The solution based on dvir is given in the function dvir:::exercise497.

.

An alternative solution is provided at (running time is fairly extensive)
http://familias.name/BookKETP/Exercises/Ch4/Exercise4_9_7b_solution.r
Generating data that can be plotted, see illustration below illustrating the posterior for the correct solution (i.e. V1=MP1, V2=MP2 and V3=MP3). Through the command length(which(results[,1]==apply(results,1,max))) in R, we get 976 out of total 1000 simulation where the correct solution has the maximum posterior.



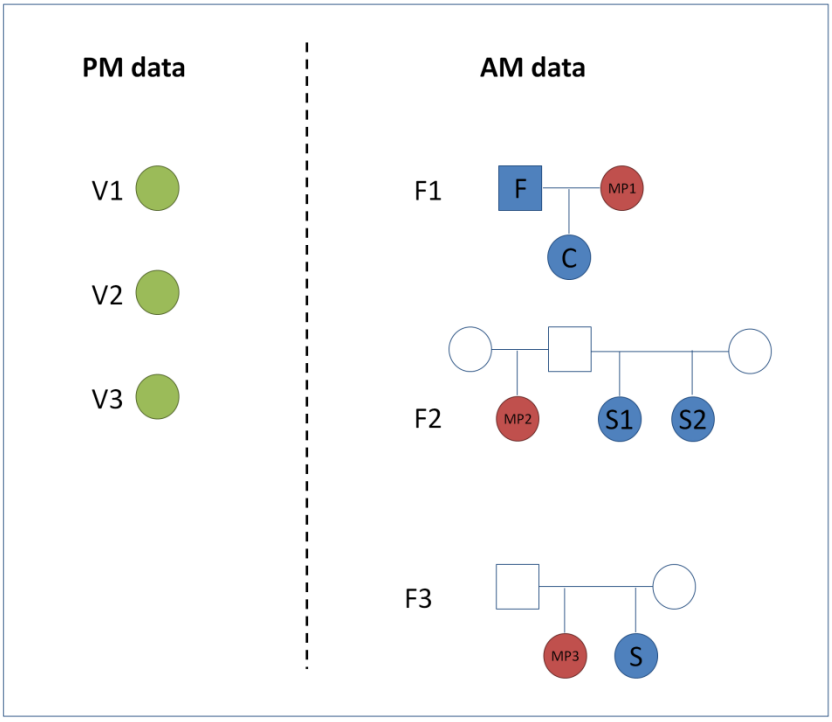


Figure 4.7. Illustration of pedigree data for Exercise 4.9.7*. Individuals highlighted in red are missing persons and individuals highlighted with blue are available reference persons.*

*4.9.8 * A missing family*

Following an accident where a car is crashed into a sea, you are tasked with the mission to identify the family of three persons, deceased in the accident. The pedigree is given in **Figure 4.8**. There are exactly three unique DNA profiles extracted from the scene of the accident. All files are available at http://familias.name/BookKETP/Exercises/Ch4/Exercise_4_9_8.zip or following links from the main repository. The exercise can be solved in standard forensic software (e.g. Familias) but we recommend R and the library https://github.com/thoree/dvir, (except for a) below) described in Section 4.8.2.

Below, from b), the R library `dvir` is used. The library can be obtained from https://github.com/thoree/dvir. See http://familias.name/BookKETP/Solutions/Solution-Family-grave-Part-II.pdf in case you run into problems.

```
library(dvir)
```

a) * Conduct a one-to-one calculation where the available relative (R1 in **Figure 4.8**) is used to identify the victims one by one. Report the LRs for each combination of victim missing person respectively. *Hint: Use a blind search function to compute the LR for parent/child as well as half sibling relations.*

Start with Familias and import PM and AM data to obtain http://familias.name/BookKETP/Solutions/Exercise_4_9_8_TE.fam. Then go `Tools > DVI module > Search > Quick scan` to obtain

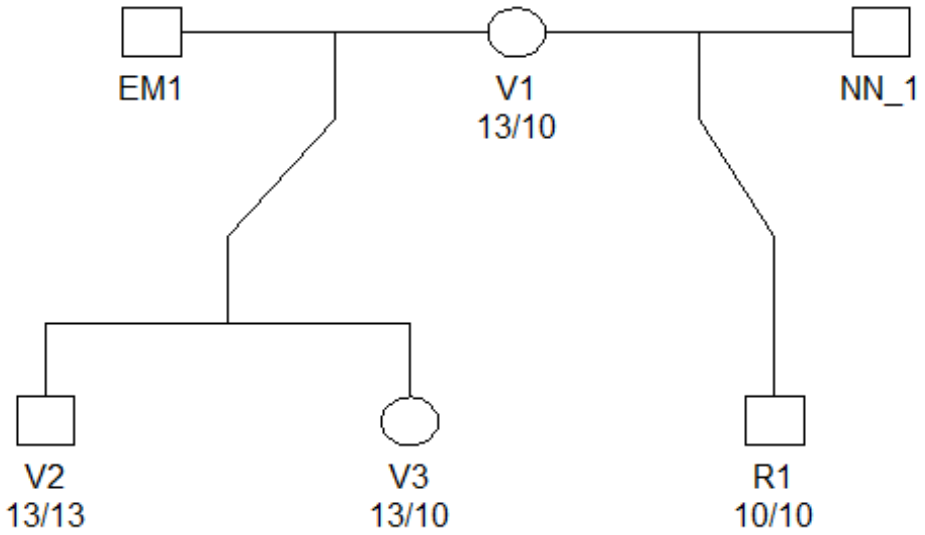
Project name is: Untitled Number of matches: 6

Family id	Unidentified person	Prior	Posterior	LR
R1 (Exercise_4_9...	V1 (Parent-Child)	0.25	>0.999999	32580331
R1 (Exercise_4_9...	V1 (Half-siblings)	0.25	0.999976	122682.05
R1 (Exercise_4_9...	V3 (Half-siblings)	0.25	0.999593	7368.3172
R1 (Exercise_4_9...	V2 (Half-siblings)	0.25	0.987755	241.99908
R1 (Exercise_4_9...	V3 (Parent-Child)	0.25	0.213915	0.81638127
R1 (Exercise_4_9...	V2 (Parent-Child)	0.25	6.72411e-008	2.0172343e-007

Blind search in a)

The findings are consistent with V1 = MP1, V2 = MP2 and V3 = MP3, however the evidence for V3=MP3 and MP2=V2 is below 10,000.

The below plot, with marker 4 shown, indicates that V1 is the father of R1 and not the other way around:



b) * We now instead turn to a global approach. How many combinations do we need to try to exhaust all solutions taking sex into account? That is, if we use the joint power of all relatives and try all combinations of victims and missing persons. [V1=MP1, V2=MP2, V3=MP3] is one such combination. *Hint: Use the ncomb function in the R library described in the introduction.*

We find the number of combinations of missing persons and victims through

```
ncomb(2,2,1,1)

[1] 14
```

c) The commands and output are

```
> data(dataExercise498)
> pm = dataExercise498$pm
> am = dataExercise498$am
> missing = dataExercise498$missing
> res = jointDVI(pm, am, missing, verbose = F)
> res
    V1  V2  V3    loglik           LR    posterior
1  MP1 MP2 MP3 -174.6096 4.899248e+27 1.000000e+00
2  MP1   * MP3 -201.5697 9.583306e+15 1.956077e-12
3  MP1 MP2   * -202.4876 3.826983e+15 7.811368e-13
4    * MP2 MP3 -205.9574 1.191034e+14 2.431054e-14
5  MP3 MP2   * -213.1740 8.745861e+10 1.785144e-17
6  MP1   *   * -221.0690 3.258772e+07 6.651576e-21
7  MP3   *   * -226.6511 1.226820e+05 2.504100e-23
8    *   * MP3 -229.4635 7.368317e+03 1.503969e-24
9    * MP2   * -232.8795 2.419991e+02 4.939515e-26
10   *   *   * -238.3685 1.000000e+00 2.041130e-28
> library(dvir)
```


```
5  MP3 MP2   * -213.1740 8.745861e+10 1.785144e-17

6  MP1   *   * -221.0690 3.258772e+07 6.651576e-21

7  MP3   *   * -226.6511 1.226820e+05 2.504100e-23

8    *   * MP3 -229.4635 7.368317e+03 1.503969e-24

9    * MP2   * -232.8795 2.419991e+02 4.939515e-26

10   *   *   * -238.3685 1.000000e+00 2.041130e-28
```


The solution is obtained with a posterior virtually equal 1. It is interesting to note that even though we only have a single relative R1 we can make identification with a very high posterior probability through the use of all missing persons. Also, in the one-to-one comparison a), none of the half-siblings reach LR>10,000 (if that is our identification threshold).

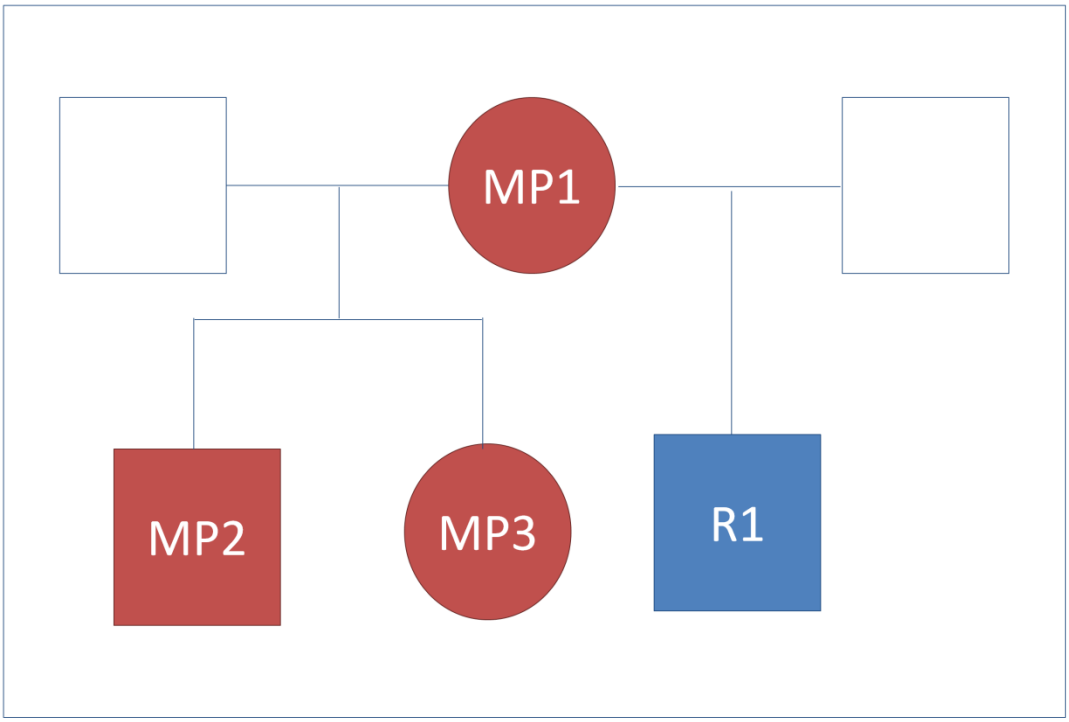


Figure 4.8. Illustration of pedigree data for exercise 4.9.8*. Individuals highlighted in red are missing persons and individuals highlighted in blue are available reference persons.*