Documentation for FamLinkX

Andreas Tillmar
Thore Egeland
Daniel Kling

1. Preface

This is the manual for the statistical tool FamLinkX that can be used for the statistical interpretation of clusters of linked markers located on the X chromosome in relationship testing. The manual is valid for all versions above 2.2.

If you need help getting started or have comments, please send us an e-mail:
daniel.l.kling@gmail.com
andreas.tillmar@rmv.se

To reference FamLinkX, please cite:
Kling D, et al. (2014), *A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium and mutations* International Journal of Legal Medicine, (2015) 129: 943.

And/or

Kling D, et al. (2015), *FamLinkX - implementation of a general model for likelihood computations for X-chromosomal marker data*, Forensic Science International: Genetics, (2015) 17, 1-7.

The book Egeland, Kling, Mostad (2016) presents FamLinkX and the underlying theory.

2. Introduction

The software FamLinkX provides functionality for likelihood calculation on family relationships/pedigrees using linked DNA marker data located on the X-chromosome. FamLinkX is a freely available software, accessible via http://www.FamLink.se. The software provides an easy-to-use graphical user interface (GUI) which allows for calculations accounting for linkage, linkage

disequilibrium (LD) as well as mutations given some genetic data for a set of persons and some relationship hypotheses. Traditional software used in forensic genetics for pedigree likelihood calculations (e.g. Familias, DNA-view etc) normally assume unlinked markers and make use of the product rule when combining likelihoods from multiple markers. Merlin, and similar software, provides extensions for linked markers, but lacks some of the user-friendliness provided by FamLinkX as well as the possibility to model mutations and recombinations within a cluster.

The main function of FamLinkX is to calculate case specific likelihood ratios (LR) for two (or more) hypotheses with observed DNA-data for X-chromosomal DNA markers. In addition, it is possible to do simulations for two or more pedigrees (hypotheses) to study the impact of ignoring linkage and/or linkage disequilibrium for X-chromosomal STR markers.

Linkage, can be described as the co-segregation of closely located loci within a family or pedigree. The genetic distance between two loci is normally expressed in centiMorgan (cM), where 1 cM is defined as a distance that corresponds to 1 % chance of a recombination occurring between the loci during meiosis. Linkage can also be measured and discussed in terms of recombination frequency (*r*). The relationship between distance in centiMorgan and the recombination frequency can be studied via mapping functions, for example Haldane which is used in FamLinkX. See Thompson (2000) for a general introduction to linkage and statistical genetics.

Linkage disequilibrium (LD), can be defined as the non-random association of alleles at different loci. That is two alleles at different loci occur more/less often together than can be expected based on their allele frequencies. LD can stretch across chromosomes but in our application it is modeled for closely linked markers on the same chromosome.

We use a *cluster approach*, where tightly linked markers belong to the same cluster. We model linkage and LD within a cluster, while between cluster only linkage is considered. Our implementation allows for a two-step consideration of LD, meaning that we may produce inaccurate results given >3 markers in each cluster. (The inaccuracy depends of course on the degree of LD within the cluster and two steps may often be sufficient)

FamLinkX requires *haplotype frequencies*, or rather observed haplotypes counts, to properly account for LD. For X-chromosomal markers we can obtain phased data from males, i.e. samples where the gametic phase is known. (For female data we need more elaborate algorithms to estimate the frequencies) To account for unseen haplotypes we use a Dirichlet distribution where the updated haplotype frequencies are obtained through (1). The model depends on the choice of the parameter lambda, which gives prior weight to unobserved haplotypes.

$$f_{Dir,i} = \frac{n_i + \lambda f_{\exp,i}}{\sum_{i=1}^{I} n_i + \lambda \sum_{i=1}^{I} f_{\exp,i}} \quad (1)$$

It is possibly to estimate λ in the Edit/Clusters markers (Options) window: R code is generated which produces estimates for each cluster and an average estimate. This method will be further tested and a manuscript will be submitted.

In our model, *Mutations* are defined as the possibility of observing a transition from one allele to another within a pedigree. We specify a matrix M, containing the transition probabilities. We specify a model for transitions where each mutation depends on the allele we mutate from. M will have A_i x A_i dimensions, where A_i is the number of alleles at locus i and where each row in M must sum to 1.0, such that the row indicates the allele we start from and the column indicates the allele we have a transition to. Accounting for mutations is mostly relevant when dealing with STR marker data, where genetic inconsistencies are frequently observed. In such data, contrary to SNP marker data, alleles are determined as a distinct number of tandem repeats. For the stepwise model, denoted 'N step model' which we implement a version of, the probability of observing a transition between two alleles decreases with the difference in the number of repeats and is determined by the 'Range' parameter, such that a single step mutation can be relatively common while a four step mutation is very improbable. We recommend the 'Extended model' introduced in Kling et al. (2014) and explained further for the Familias (familias.no) software. See also Figure 4 for typical parameter values.

3. Installation

FamLinkX is compatible with all Windows version from XP and the installation files are provided at http://www.famlink.se/fx_download.html. Apple users may emulate a Windows environment to install the software.

4. Getting started

There is a separate Getting started document which provides two simple examples aimed at first time users. This document is available from within FamLinkX under 'Help'.

4.1 Editing marker information and setting the Option parameters

This applies to functionality obtained by entering File > Frequency database

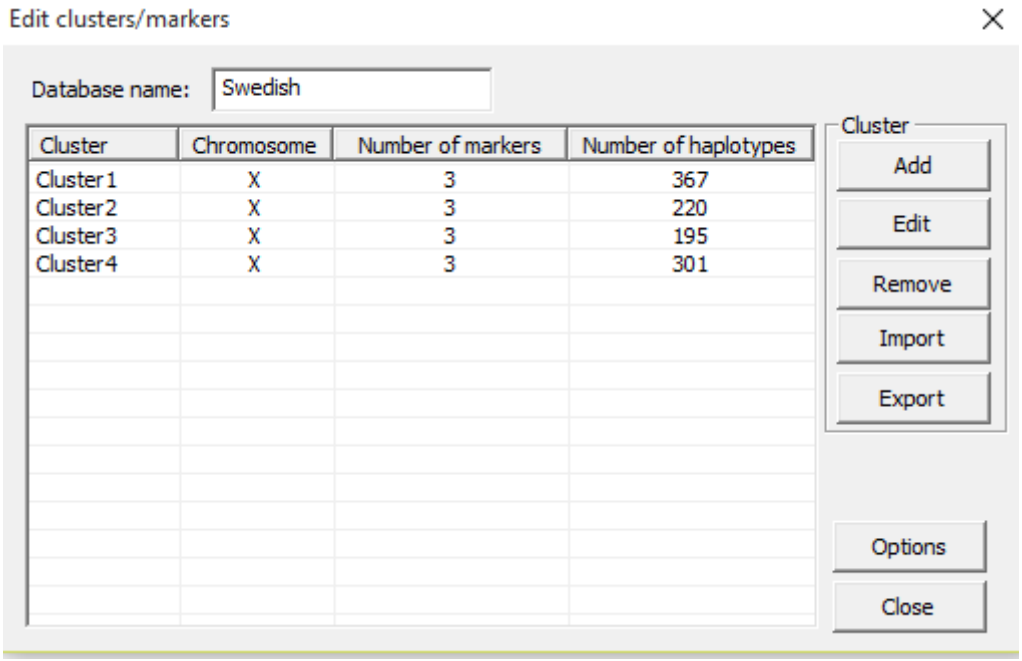


Figure 1. The "Edit clusters/markers" dialog.

4.1.1 Marker database

The marker database (see Figure 1) contains information about the Clusters and the specific genetic Markers. The information includes:

1. Cluster info
    - Information about markers in the same LD-group (i.e., markers for which haplotype-frequencies should be usedas determined by : values close to 0 implies that haplotype frequencies dominate whereas haplotype frequencies are weighted down for large values) and their observed haplotype data.
2. Marker info
    - Information about the name of the marker, its chromosomal location (in centiMorgans, cM), allele data and mutation models and mutation rates. (Note that the implementation is currently limited to X-chromosomal data and that alleles names must be numerical.)

4.1.1.1 Add a new cluster.

Click on “Frequency database” in the File menu. Click on the Add-button. The dialog displayed in Figure 2 should appear. The dialog is divided into two parts, to the left specification of genetic markers in the cluster and to the right specification of haplotype observations.

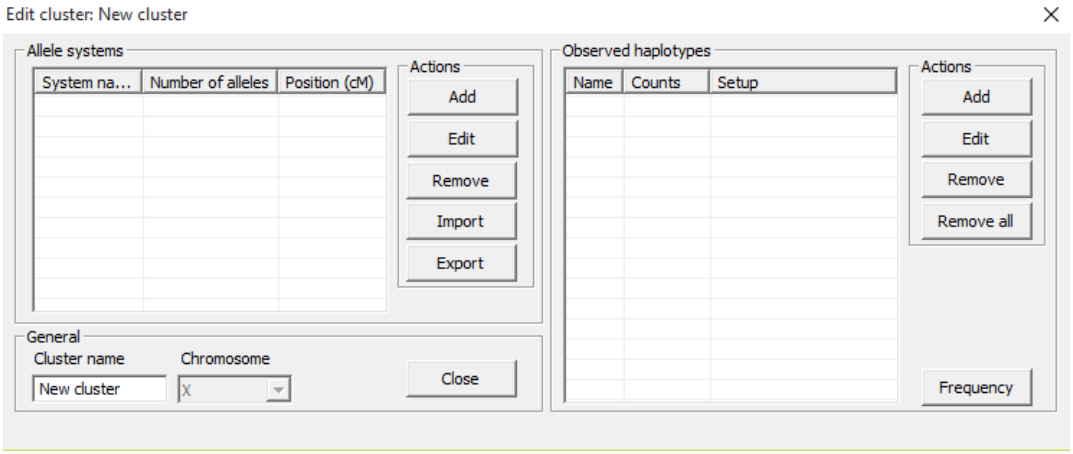


Figure 2. The "Edit cluster" dialog

- Add new markers, within the cluster, by clicking “Add” (left side)
- Edit marker information by marking a specific DNA-marker and then clicking "Edit"
- Add new haplotype data by clicking “Add” (right side).
- Edit haplotype information by marking a specific haplotype and then clicking "Edit"

4.1.1.2 Import haplotype database from file.

Observed haplotype data can be imported via a file-import option. This is done per cluster as follows: Mark one of the pre defined clusters (in the “Edit cluster/markers”-window) and click on the “import” button and chose the tab delimited text-file (*.txt) with the haplotype data according the following format:

Haplotype ID	Number		DXS10148	DXS10135	DXS8378
1	1	13.3	33.2	12	
2	2	14	21	10	
3	1	14	22	10	
4	1	14	26	12	
5	1	14	32	12	
6	1	17	27	10	
7	2	18	18	10	
8	2	18	18	10	

...

4.1.1.3 Manual editing of allele frequency database

Click on “Frequency database” in the File menu. Mark the cluster in which the marker is located, and click edit. Mark the marker and click edit. The dialog in Figure 3 should appear.

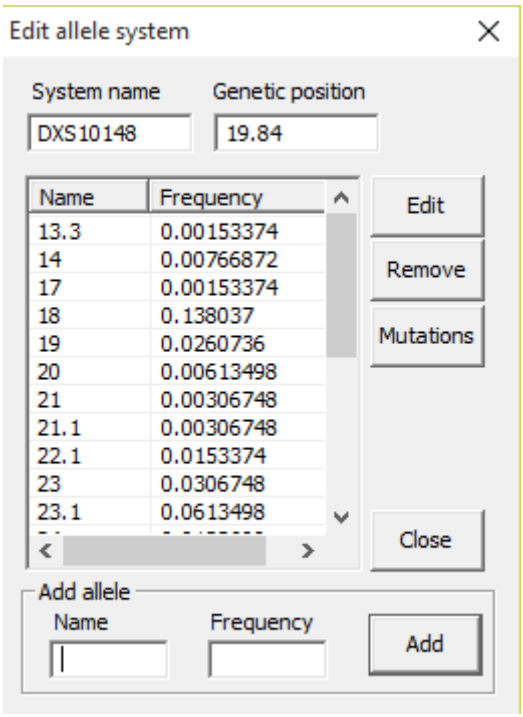


Figure 3. The "Edit allele system" dialog

Here it is possible to edit the name of the locus and to designate name of alleles and their frequencies. Please note that the name of an allele needs to be numeric. Iit is also possible to set the gentic location of the marker and, by clicking on the mutations-button, edit the mutation model and mutation rate (see 4.1.1.4)

4.1.1.4 Mutation model

To change mutation parameters, click on “Frequency database” in the File menu. Mark the cluster, in which the specific marker is located, and click edit. Mark the marker and click edit. Here it is possible to edit the name of the locus and to designate name of alleles and their frequencies. Click Mutations to edit the mutation model and mutation rate, see Figure 4.

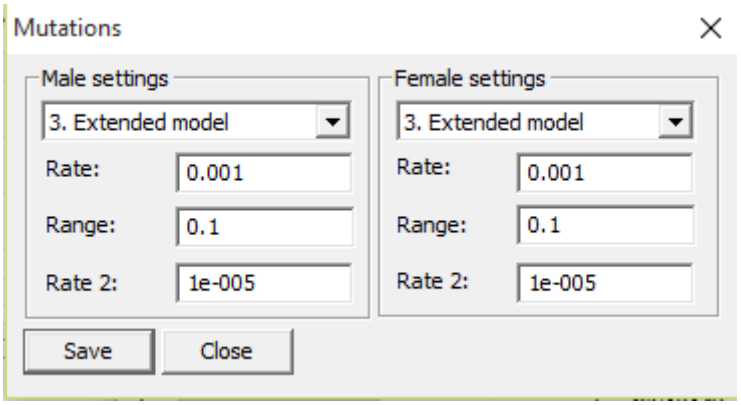


Figure 4. The "Mutations" dialog

We can have different mutation parameters for male and female transmission. We can select three different mutation models:

1. *Simple model (Equal).* The probability of observing a transition from any allele to any other allele is equal for all transitions. E.g. given a mutation rate of μ, the probability of observing a transition from allele A_i to allele A_j is μ/(n-1) for all j, where n is the number of alleles at the current locus.

2. *N step model.* We only consider mutations *N* step away. This model is only applicable to STR markers where we can observe distinct tandem repeat alleles. E.g. given a mutation rate of μ, the probability of observing a transition from allele A_i to allele A_j is $k\mu R^{|i-j|}$ if i≠j, where *R* is the mutation range, determining how probable a one step mutation is compared to a two step etc and *k* is a normalizing constant (not specified by the user). Any mutations requiring more than *N* steps will have zero probability.

3. *Extended model.* We consider a complete stepwise model. This model, as 2., is only applicable to STR markers where we can observe distinct tandem repeat alleles. To further complicate, we may have microvariants, not equal to a distinct number of repeats; this model accounts for those ”sidesteps” as well. We have three different parameters in this model, (Primary) Rate, μ, Range, R and (Secondary rate) Rate 2, α. The probability of observing a transition from allele A_i is 1-(μ+α), the probability of observing a transition to any non-distinct number of repeat alleles, i.e. ”sidesteps”, is α/(m-1), where m is the number of ”sidesteps” transitions for the current allele. The probability of observing a transition to an allele with a distinct number of tandem repeats is equal to $k\mu R^{|a-b|}$, where *k* is calculated such that each row in the mutation matrix sum to 1 and where a is the number of repeat units for allele A_i and b is the number of repeats units for allele A_j.

4.1.1.5 Import markers from file.

Marker data and allele frequencies can be imported via a file-import option. This is done per cluster as follows: Create an empty cluster (in the “Edit cluster/markers”-window). In the "Edit cluster" dialog that appears, click the "Import" button and select a tab-separated file with format as follows.

MarkerName
Allele tab Frequency

For instance,
DXS10148
13.3 0.1
14.3 0.9
...

4.1.1.5 Viewing haplotype frequencies

It is possible to compute the haplotype frequencies for investigative purposes. In the "Edit cluster" dialog, click the Frequency button (bottom right). The dialog displayed in Figure 5 should appear. Select the marker setup and select the number of observations (Counts) and the value of Lambda and click "Update". An estimated frequency will appear.

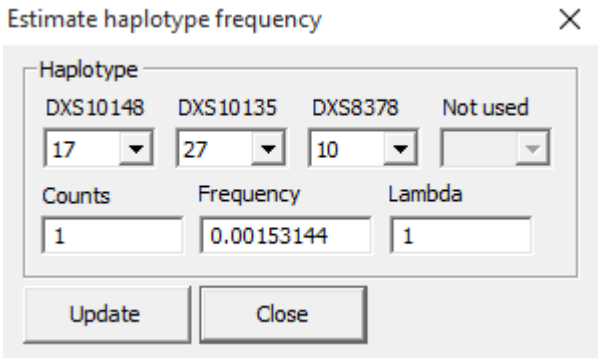


Figure 5. The "Estimate haplotype frequency" dialog

4.1.2 Parameter options

This applies to functionality obtained by entering File > Frequency database > Options, see Figure 6.

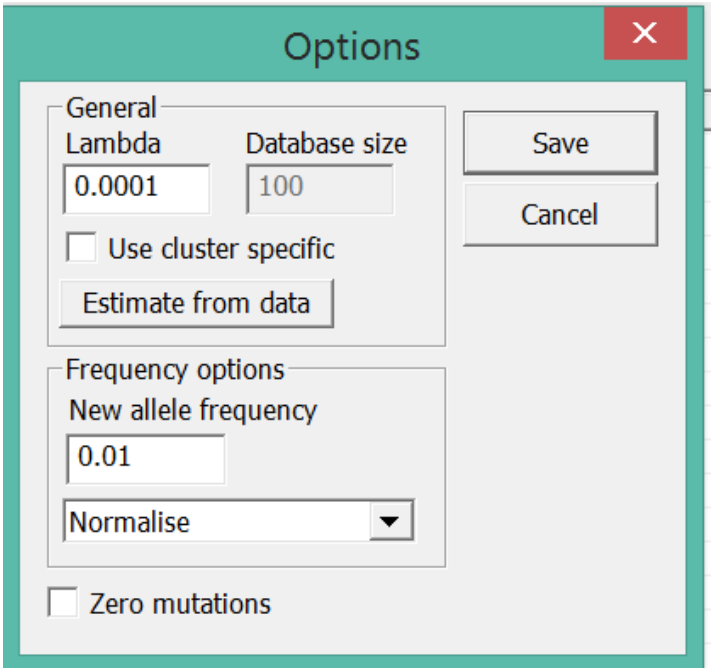


Figure 6. The "Options" dialog.

4.1.2.a Lambda

Here the value of Lambda is set. Lambda is a parameter that is used for the haplotype frequency estimation, see Introduction. A more detailed description of the lambda model can be found in Tillmar et al., 2008. Briefly small values of lambda results in a high weight of the observed data for the haplotype frequency estimation, while high values results in greater weight to the expected haplotype frequencies. If a sufficiently large Lambda is chosen, the calculated frequencies will be completely based on the expected haplotype frequencies. Observe that Lambda can be estimated as explained previously.

4.1.2.b New allele frequency and scaling

This frequency is used when a case involves an earlier not seen allele, i.e., an allele not present in the frequency database. Two options are possible for how the database should be updated, either via (1) normalisation (“Normalize”) of all allele frequencies in order to sum all the frequencies to 1, or (2) subtract the frequency of the new allele (“Search and Subtract”), from alleles not included in the current case. The update of the database is performed upon likelihood calculation and the frequencies, as imported/manually entered, will remain intact for the next case/calculation. This options also applies when the total sum of the allele frequencies are above or below 1.

4.1.2.c Zero mutations

If ticked, a zero mutation model will be used for all computations. This is an option in available in order to save the time from changing all the mutation rates to zero for all systems. In addition, the rates and models are still stored should the user decide to untick this option.

4.1.3 Advanced settings

This applies to functionality obtained by entering *File* > *Advanced*, see Figure 7.

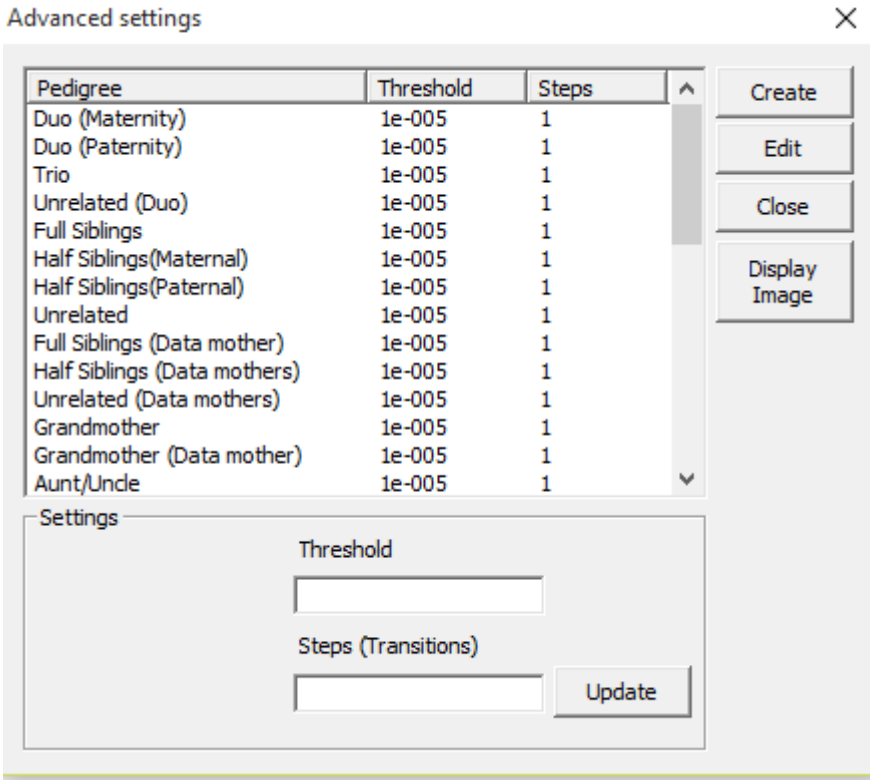


Figure 7. The "Advanced settings" dialog.

This part is mostly aimed at users interested in modeling mutations for STR markers. For each hardcoded relationship two different parameters can be specified; Steps, which is the number of transition steps allowed for, i.e. if the number of steps=2 we only allow for a two-step mutation; Threshold, which is a more arbitrarily defined value and specifies the overall pedigree likelihood threshold, i.e. given a specific combination of genetic data and a pedigree, what is the minimum threshold to include this combination in the subsequent calculations. We recommend keeping the Steps parameter low for each relationship (0-1) while keeping a high value on the Threshold parameter (0.01-0.0001). Should the LR become zero, you may wish to vary these parameters, primarily by decreasing the Threshold. If computations appear slow then the first move is to set the Steps parameters to 0.

4.1.4 Conversion Theta <-> cM

Under the tools menu, the “conversion” tab can be used to convert recombination frequencies to genetic distance (cM) via the three different mapping functions; Kosambi, Haldane and Morgan (Thompson, 2000), see Figure 8.

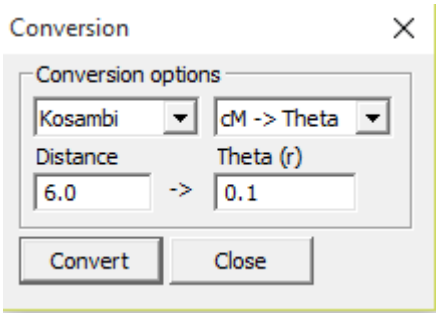


Figure 8. The "Conversion" dialog.

This function does not change anything in the current project, but can be used as calculator to convert between genetic distance and recombination distance.

4.2 Calculation of case-specific LR

The calculation of case specific LR can be done as follows (assuming that marker database has been specified in the ”Frequency database”):

4.2.1 Wizard using the pre-defined pedigrees

1. Select the “New wizard” option in the File-menu. (Alternatively select Tools > Select pedigree). The window in Figure 9 should appear.

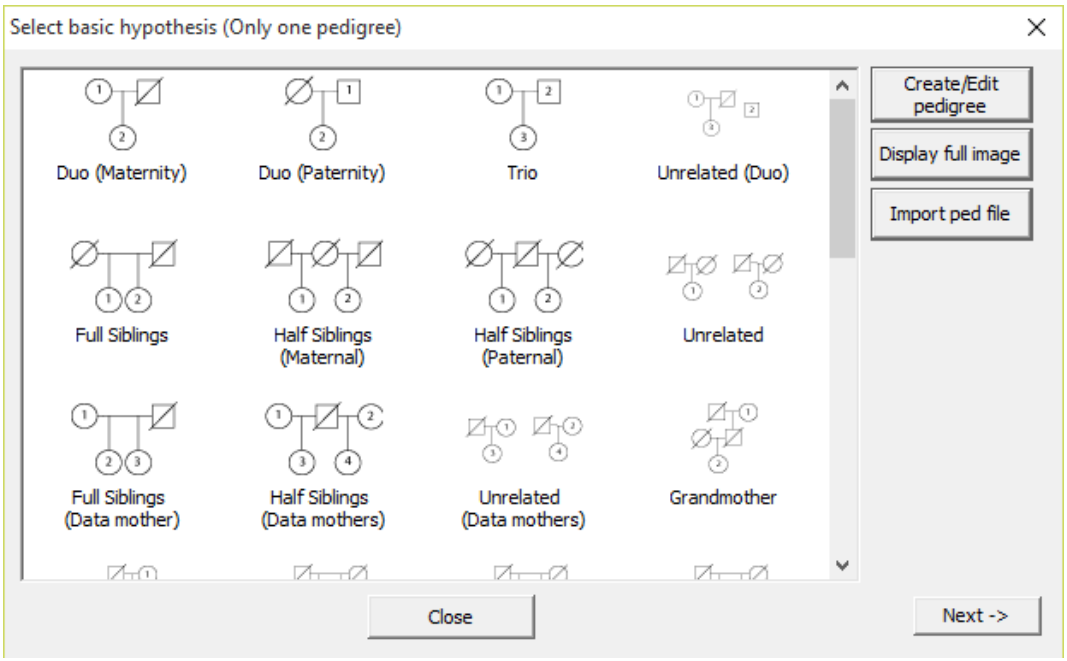


Figure 9. The "Select basic hypothesis" dialog.

2. Select the main hypothesis, and then click the “Next” button. (Alternatively double click the desired hypothesis). It is possible to create you own pedigrees by pressing "Create/Edit pedigree". Select the alternative hypothesis (if multiple alternatives, select all by

simultaneously holding Ctrl down). Click "Next" when finished. (Alternatively, if selecting only one alternative hypothesis, double click the desired hypothesis).

3. Add the DNA data for the typed individuals. The appearance should be as in Figure 10.

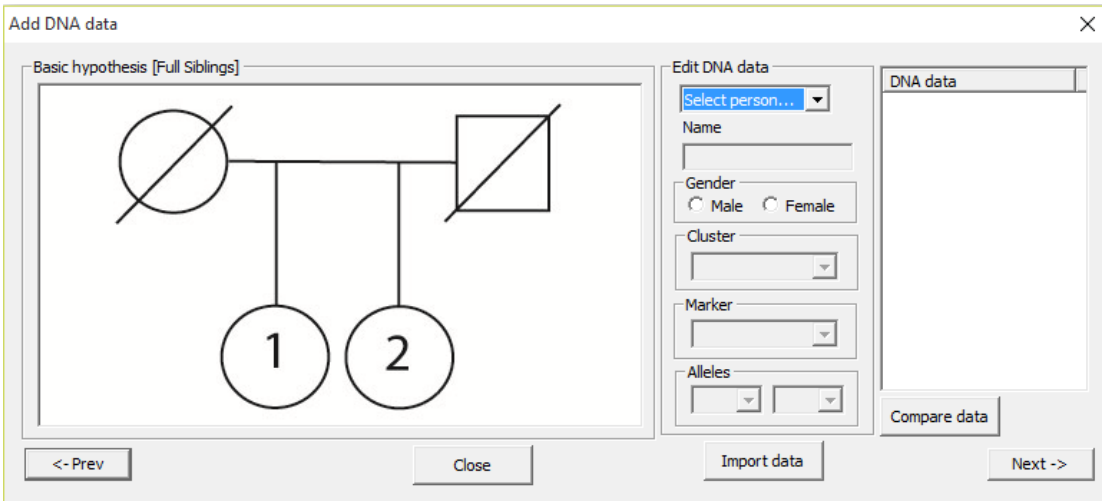


Figure 10. The "Add DNA data" dialog.

a. It is possible to give each individual a name and to set the gender.

b. Chose alleles from the dropdown-list, or add a new allele by typing the allele name in the appropriate box. (Alternatively double click a specific marker in the list to the right in order to automatically selected this marker for edit).

c. It is possible to import DNA data using a pre-defined file, either on the "Familias format" or in CODIS xml format, see Appendix 1 for specifications.

(Appendix 1).

d. To compare the specified DNA data for all individuals, click the "Compare data" button.

4. Click "Next" to get to the final step and to perform calculations.

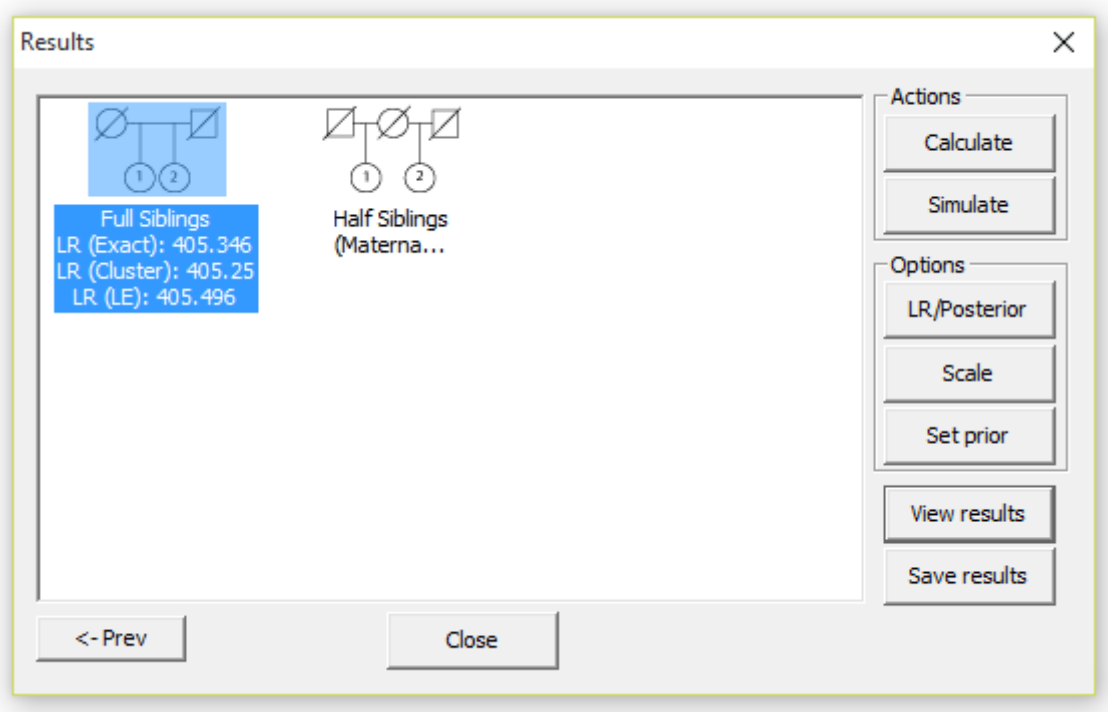


Figure 11. The "Results" dialog.

5. Press “Calculate” and the LR will be calculated. The results should look as in Figure 11. We present the LR using three different computation methods,
    I. LR (Exact) - the methods presented in Kling et al. (2015) are used. This is the most exact and recommended method to report. We account for mutations, recombinations within and between clusters as well as LD structure (haplotypes) within clusters.
    II. LR (Cluster) - similar to the previous method but we do not account for mutations or recombinations within a cluster.
    III. LR (LE) - the most naive method where we model the markers as single markers. No LD structure is accounted for, in other words, haplotype observations are disregarded.
6. It is possible to alter the prior for the hypotheses via the “Prior” tab, when the posterior probability is chosen as the outcome.
7. The result can be scaled via the “Scale” button. In practice this selected which pedigree likelihood should be in the denominator of the LR.
8. View the results and the marginal contribution from each single marker by pressing the "View results" button, see Figure 12 for an illustration.

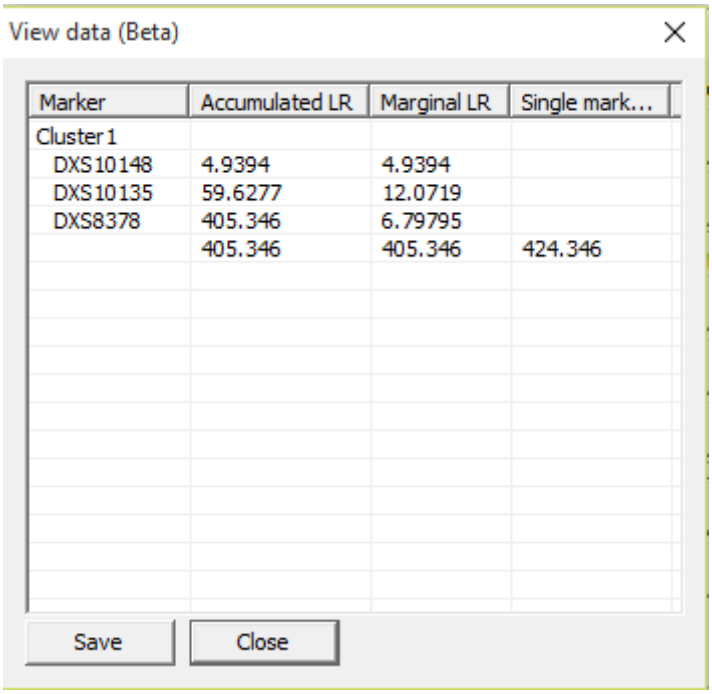


Figure 12. The "View data" dialog.

9. The results can be saved by clicking the "Save results" button.
    a. Save LR
    b. Case report
    c. Simulation report
    d. Txt file with the markers and the marginal LR:s for each system as well as total LR.

4.3 Perform simulation to study the impact of linkage and LD for a given pedigree and set of clusters/markers

The simulation tool is helpful to examine the impact of taking linkage and LD into account or not, for any set of clustered/linked markers and any given pedigrees. Note that a warning message is issued when running: “Simulation have not yet been tested and verified. Use with caution. The simulation algorithm simulates data accounting for linkage, LD and mutations. The algorithm starts by defining all founder alleles. Founder alleles are sampled using haplotype frequencies where LD structure is accounted for. The algorithm continues by simulating transitions from the founder alleles to all non-founders using a mutation matrix where each transition is assigned a specific probability. In addition, recombination is considered when two neighboring markers are simulated, i.e. the algorithm keeps track of the maternal/paternal chromosomes. Calculations are then performed using only genetic data from the typed persons.

To perform the simulation, just follow the steps given in 4.2.1 above but ignore the step with adding DNA data. On the result tab, select simulation and type the number of simulations. There is an option to save the “raw data”, to perform your own calculations. In addition a non random seed can be specified. (Not recommended unless to reproduce previous results or examine the generated data more closely). The simulation report can be generated using the "Save Results" button.

6. Examples

6.1 Example 1 – A standard case (“Paternal half-siblings” versus “Unrelated”)

In this example we want to calculate the likelihood ratio (LR) for the case where two females are tested wether they are paternal half-siblings or unrelated. The data consist of analysis of 12 X-chromosomal STR markers (Argus-12), located in four different clusters with three markers in each. Observed haplotype data is available for the different clusters. We will use a Swedish haplotype frequency database (Tillmar et al 2012, Input files can be found in Install folder\Examples, C:\Program files\FamLinkX\Examples if this is the FamLinkX install directory). In this example we will use information about marker location and mutation rates from Nothnagel et al., 2012.

We start a new project by selecting “New project” in the File menu, and then select the “Frequency database” in the File menu. We define all clusters and markers, including their genetic positions, mutation model etc (see 6.1.1). We import the frequency data (see 6.1.2). and then we use the “wizard” in order to specify pedigree-hypotheses (see 6.1.3) and to import case data (see 6.1.4), and finally we count the likelihood ratios and create a report (see 6.1.5).

6.1.1 Settting up the cluster/marker data set

Add a new cluster by clicking the ”Add” tab. A new cluster ”New cluster” appears in the cluster list, see Figure 13. (Newer versions of FamLinkX will automatically open the newly created cluster for editing).

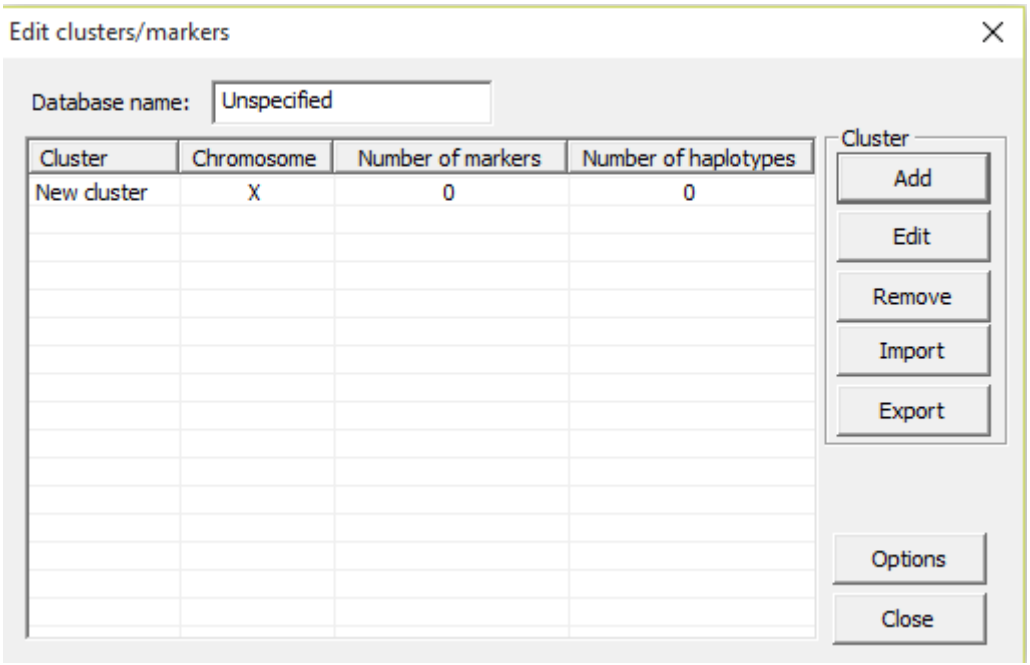


Figure 13. Adding an empty cluster.

To specify the marker information for the new cluster, mark ”New cluster” and click ”Edit”. A new window appears:

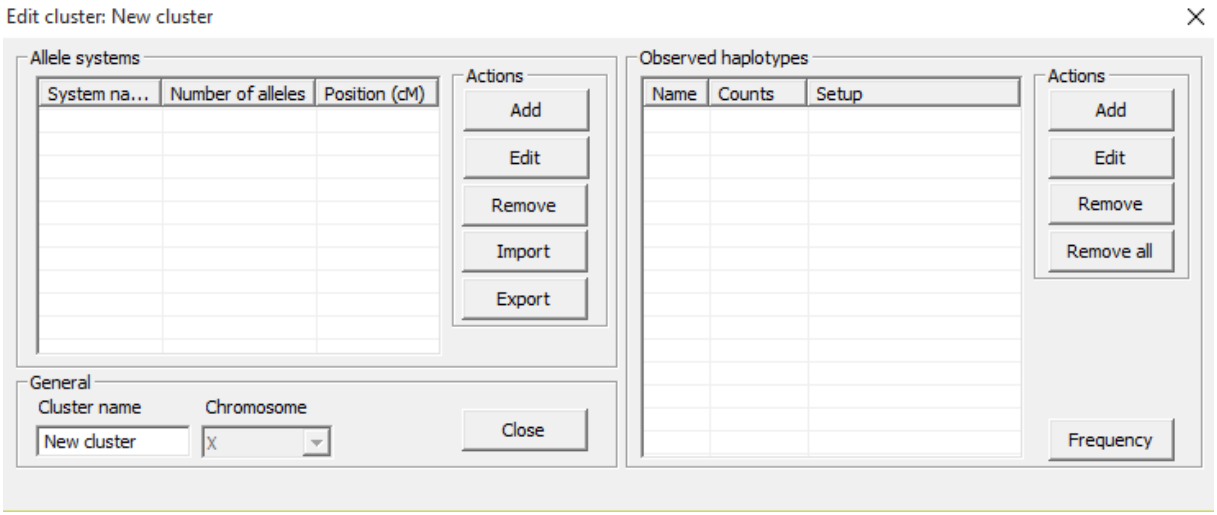


Figure 14. Editing the empty cluster.

First we change the name of the cluster (lower left). In this case we call the cluster ‟Cluster 1”. Since this cluster contains three STR markers, we have to ‟Add” three new markers, which is done by clicking ‟Add”-button three times.

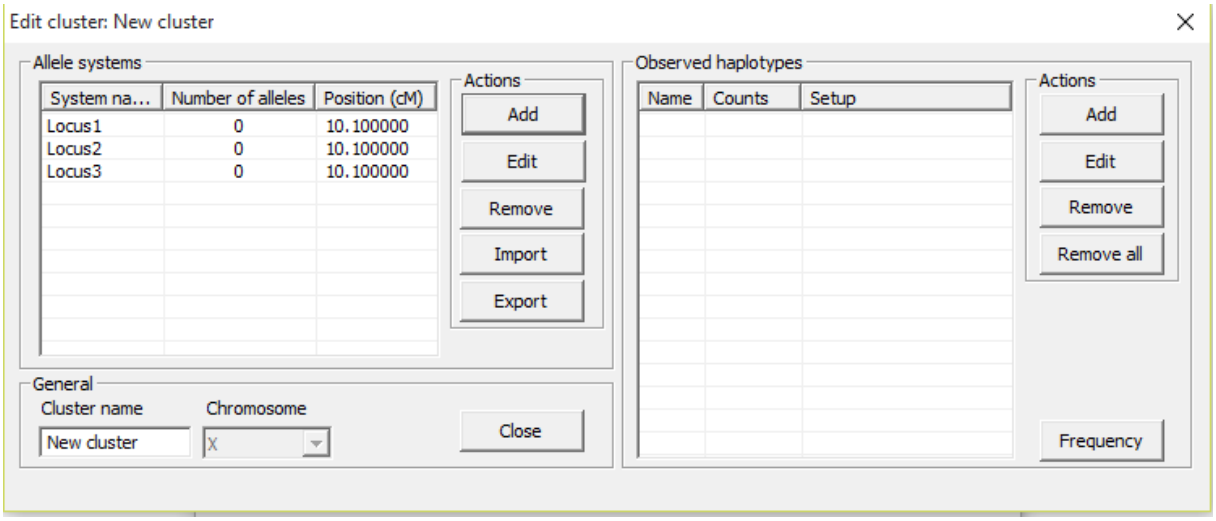


Figure 15. Inserting three empty markers.

To enter marker specific information, select "Locus1" and then click on "Edit". A new window appears:

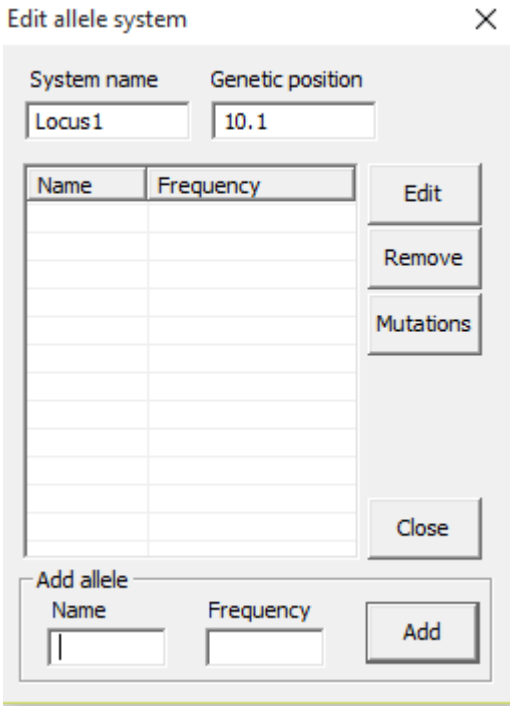


Figure 16. Editing Locus1.

Now enter the name of the locus in the field "System name" (e.g. "DXS10148") and its genetic position. To specify the mutation model and rates click on the "Mutatio"-button and edit the rates if necessary. In this case, for the DXS10148 marker, we select the "Extended model" and set the "Rate" to 0.0031, "Range" to "0.1" and " "Rate 2" to 1e-005 (for a more information regarding the mutation models, please see previous definitions).

Repeat the steps, above, for the remaining two markers within "cluster 1" and then add more clusters (four in total) and finalize the marker information for the remaining three clusters, all in all 12 markers.

6.1.2 Import haplotype data

The observed haplotype data is located in four different .txt files (one file for each cluster). The input files can be found in Install folder\Examples, C:\Program files\FamLinkX\Examples (if this is the FamLink install directory). Go to the " "Edit cluster/markers" – window via "Frequency database" in the File menu. Select the first cluster ("cluster 1") and click the "Import"-button. Select the .txt-file with the observed haplotype data for cluster 1 (see 4.1.1.2 for the design of the input-file). In this example the file is called "SweX12_hap_clu1.txt". When the data is imported the "number of haplotypes" should updated to reflect the changes. Repeat these steps for all four clusters. Now, your project is ready for case-specific computations

6.1.3 Specify pedigree-hypotheses

We use the "New wizard"-approach (available in the File-menu).

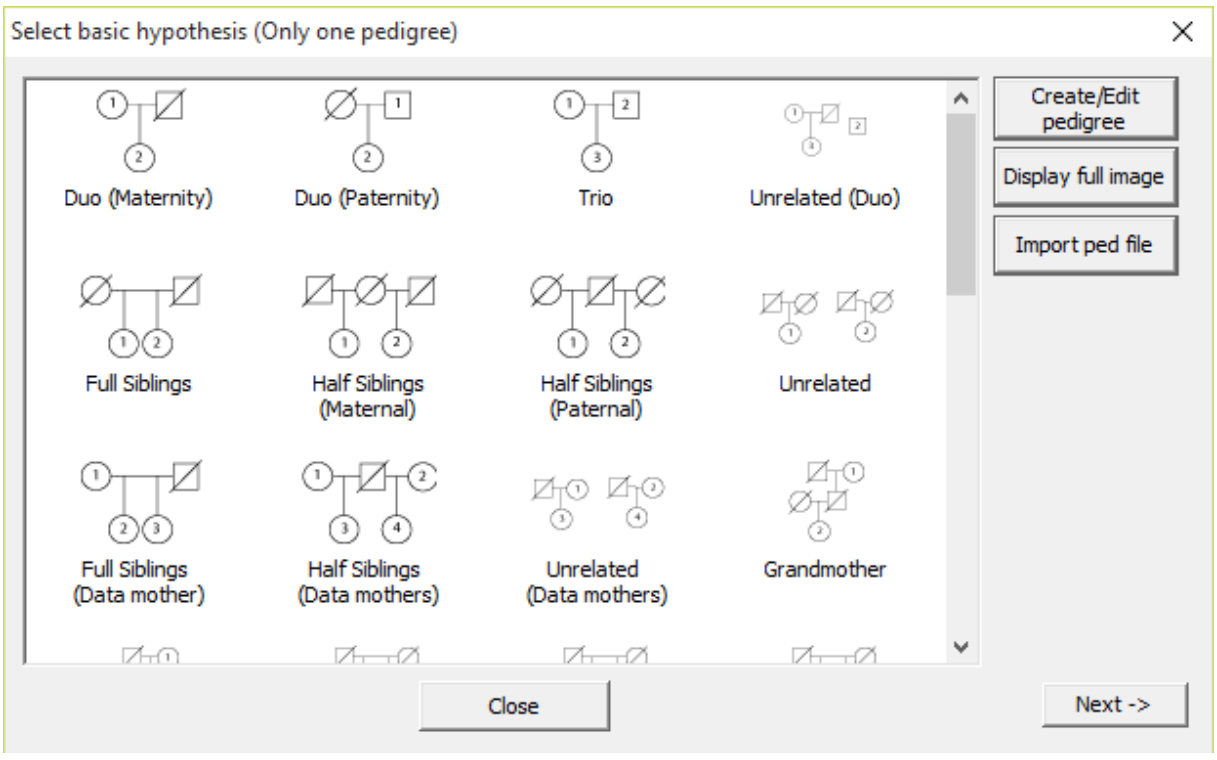
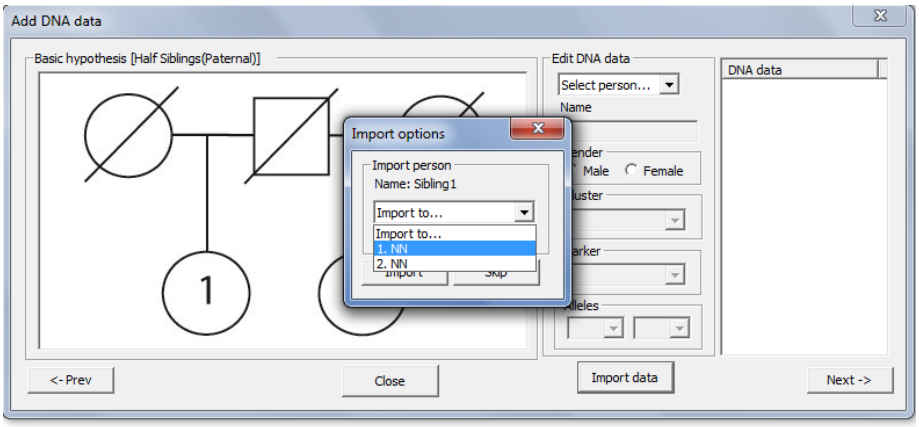


Figure 17. Selecting the main (first) hypothesis.

We select the pedigree “Half Siblings (Paternal)” as the basic hypothesis and the click ”Next ->”. In the next window we select the “Unrelated” hypothesis, and click “Next ->”.

6.1.4 Add case DNA data

In order to add the DNA profiles we use the “Import data” option in the “Add DNA data”-window. We click on “Import data” and navigate to the .txt-file with the STR data for the two individuals in the case (see Appendix 1 for info how to create the DNA data input-file).



Via “Import option” we select for which individual in the pedigree (1 or 2), “Sibling1” should be, and also tick the box “import name” followed by click on “import”. Then we do the same thing for “sibling 2”. We check the the correct DNA data has been imported by selecting each indiviual in the pull down menu below “Edit DNA data”. The we click on “Next ->” and the “Results”-window appears.

We start the calculation by clicking “Calculate”. The LR is the computed and presented under the “basic hypothesis” (i.e., scale vs “Unrelated” in this case). We create a report via the “Save results”-button and select “Case report”.

7. References

G.R. Abecasis, S.S. Cherny, W.O. Cookson and L.R. Cardon, Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* 30 (2002) 97-101.

J. Buckleton and C. Triggs, The effect of linkage on the calculation of DNA match probabilities for siblings and half siblings. *Forensic Sci. Int.* 160 (2006) 193-199.

T. Egeland, P.F. Mostad, B. Mevag and M. Stenersen, Beyond traditional paternity and identification cases. Selecting the most probable pedigree. Forensic Sci. Int. 110 (2000) 47-5

Egeland, Thore, Daniel Kling, and Petter Mostad. Relationship Inference with Familias and R: Statistical Methods in Forensic Genetics. Academic Press, 2015. .

Kling, Daniel; Tillmar, Andreas; Egeland, Thore. Familias 3 - Extensions and new functionality. Forensic Science International: Genetics 2014 ;Volum 13. s. 121-127

Kling, D., Tillmar, A., Egeland, T., & Mostad, P. (2014). A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations. International journal of legal medicine: Vol 129, Issue 5, pp 943-954.

Nothnagel, M., Szibor, R., Vollrath, O., Augustin, C., Edelmann, J., Geppert, M., Alves, C., Gusmao, L., Vennemann, M., Hou, Y., Immel, U. D., Inturri, S., Luo, H., Lutz-Bonengel, S., Robino, C., Roewer, L., Rolf, B., Sanft, J., Shin, K. J., Sim, J. E., Wiegand, P., Winkler, C., Krawczak, M., and Hering, S. (2012). Collaborative genetic mapping of 12 forensic short tandem repeat (str) loci on the human x chromosome. Forensic Sci Int Genet, 6(6), 778–84.

E.A. Thompson, Statistical inference from genetic data on pedigrees. NSF-CBMS Regional Conference Series in Probability and Statistics (2000) Volume 6. IMS, Beachwood, OH.

A.O. Tillmar, Population genetic analysis of 12 X-STRs in Swedish population. *Forensic Sci Int Genet*, 6 (2012), e80–81.

Appendix 1. Format of DNA data input-file

The format of the input file of DNA data follows the standard format of Familias DNA input-file. The file is most easily created in Excel-spread sheet and then saved as a .txt-file.

The first row is a header containing three columns for sample name and gender information, with the headers (“Sample name:”, “Amelogenin 1”, “Amelogenin 2”). The following columns represent the DNA markers used, two columns for each marker with the following headers for a two marker example (“mark1 1”, “mark1 2”, “mark2 1”, “mark2 2”). Note that male X-data is included as a homozygous.

Sample name:	Amelogenin 1	Amelogenin 2	mark1 1	mark1 2	mark2 1	mark2 2
Sibling1	X	X	24	24	12	19
Sibling2	X	X	24	27.1	13	19

Note, FamLinkX accepts CODIS xml files. This format is defined elsewhere.