Documentation for FamLink version 1.1

1. Preface

This is the manual for the statistical tool FamLink that can be used for the statistical interpretation of pairs of linked marker in relationship testing.

If you need help getting started or have comments, please send us an e-mail:
daniel.kling@fhi.no
andreas.tillmar@rmv.se

To reference FamLink, please cite: Kling D, Egeland T, Tillmar AO, *FamLink – A user friendly software for linkage calculations in family genetics.* Forensic Science International Genetics (2012) DOI: 10.1016/j.fsigen.2012.01.012

1.1 New features in FamLink version 1.1

- A new simulation tool "Analyze Merlin files". Pedigree simulation with data from mutiple DNA markers (see 5.3 below).
- The map of the physical location of the markers available for the "Quick analyis" option has been updated.
- The report for "Quick analyis"-results has been updated.
- Some minor bugs have been corrected.

2. Introduction

The software FamLink provides functions for likelihood calculation for family relationships/pedigrees using pairwise linked DNA marker data. FamLink is a freely available software, accessible via http://www.FamLink.se. The software provides an easy-to-use graphical user interface which allows for linkage calculation based on the Lander-Green algorithm. FamLink uses the same implementation as in the Merlin engine (Abecasis et al., 2002).

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 Famlink. See Thompson (2000) for a general introduction to linkage and statstical genetics.

In criminal casework, linkage only has an impact on match probability calculations when the alternative hypothesis is a close relative (Buckleton et al, 2006). For kinship/relationship testing, however, linkage can be relevant in the transition probabilities for alleles passing from founder to a child for certain pedigrees, e.g. pedigrees with one individual that is present in two or more meioses.

Traditional software used in forensic genetics for pedigree likelihood calculations (e.g. Familias, DNA-view) normally assume unlinked markers and make use of the product rule when combining likelihoods from multiple markers.

Lately, the international forensic DNA community has focused attention on the use of two closely located STR loci (vWA and D12S391) that have been included in the European standard set of STR markers (ESS). Moreover, there is a possibility to have several closely located pairs of loci when combining information from multiple common PCR multiplexes (Philips et al., 2011)

FamLink has two main functions, (1) calculate case specific likelihood ratio (LR) for two (or more) hypotheses with observed DNA-data for two linked DNA markers and (2) perform simulations for two or more pedigrees (hypotheses) to study the impact of ignoring linkage for a specified pair of linked STR markers.

3. Installation

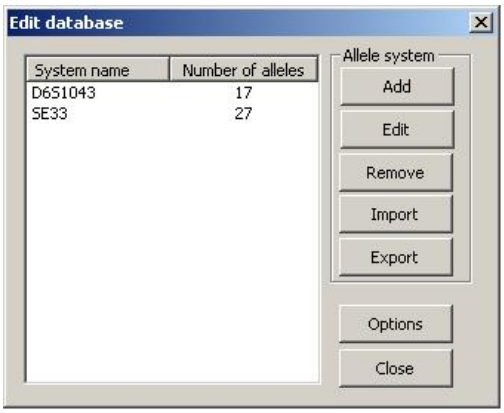
FamLink is compatible with all Windows version from XP and the installation files are provided in FamLinkSetup.exe

4. Getting started

There is a separate Getting started document which provides a simple example aimed at first time users. This document is available from within FamLink under ‘Help’.

4.1 Setting the Option parameters

This applies to functionality obtained by entering File->Frequency database



4.1.1 Allele frequency database

The allele frequency database contain information for two different loci and can be edited either (a) manually by typing the allele and its frequency or (b) via import of a pre-defined file. There is no general limit to the number of alleles allowed at each locus, but the simulations only allows for 64 alleles and therefore only the first 64 will be used.

4.1.1.a Manual editing of allele frequency database

Click on “Frequency database” in the File menu. Mark the first marker in the list and click edit. Here it is possible to edit the name of the locus and to designate name of alleles and their frequencies.

4.1.1.b Importing allele frequency database from file.

Click on “Frequency database” in the File menu. Click “import” and chose the text-file (*.txt) with the allele frequency data according the following format: (FamLink allows for both , and . as decimal separators)

SE33
22 0.04
33 0.96

D6S1043
11 0.12
12 0.87
13 0.01

Mark both markers and click on “import” and select the two systems you would like to use.

Note that the file may contain a list of more markers, though the software only allows the user to import two systems. If no systems are selected, FamLink will import the two first systems by default.

FamLink also has the ability to import a Familias database. (Egeland et al., 2000) The format is not described here but is simply a save file from the Familias software containing a frequency database.

4.1.2 Parameter options

This applies to functionality obtained by entering File->Frequency database->Options

4.1.1.a Theta

Here the recombination frequency is defined. Please give the recombination frequency, which is later on converted to a genetic distance (cM) using Haldane's map function (Thompson, 2000). (To convert a genetic distance to recombination frequency, see 4.1.2)

4.1.1.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.

4.1.1.b Error rate

FamLink cannot presently accommodate kinship (theta) corrections nor mutations. It is, however, possible to model genotyping error whereby the true allele is recorded as a randomly chosen allele with a specified probability.

4.1.2 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). 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 in three ways, either via pre-defined pedigrees, via importation of a ped-file or via a Familias-file (Quick analysis).

4.2.1 Wizard using pre-defined pedigrees

1. Select the "New wizard" option in the File-menu.
2. Select the main hypothesis, and then click the "next" button. (Alternatively double click the desired hypothesis).
3. Select the alternative hypothesis (if multiple alternatives, select all by simultaneously holding Ctrl down).
4. Add the DNA data for the typed individuals
    a. It is possible to give each individual a name.
    b. Chose alleles from the dropdown-list, or add a new allele by typing the allele name in the appropriate box.
    c. It is possible to import DNA data using a pre-defined file of the Familias-import DNA-data format.
5. Click next.
6. Press "Calculate" and the LR will be calculated.
7. It is possible to alter the prior for the hypotheses via the "Prior" tab, when the posterior probability is chosen as the outcome.
8. The result can be scaled via the "Scale" button.
9. The results can be saved by either
    a. Save LR
    b. Generate a report.
    c. Save a multiplication factor: MF=LR(linkage)/LR(no linkage) .This can be used if a LR for the case has been calculated elsewhere assuming unlinked markers, and an adjustment is desired in order to account for linkage. Example: Assume a LR=1000 is obtained using all markers and disregarding linkage between a pair of markers The correct LR_{corr} accounting for linkage is LR_{corr}=LR*MF=1000*MF.

4.2.2 Import a pre-defined ped-file

We refer to http://www.sph.umich.edu/csg/abecasis/Merlin/tour/input_files.html for a complete description of the ped-file. What is important to notice is that the ped-file must contain at least two hypotheses, and include all DNA-data. (The frequency database is still defined in the software)

Example

Consider a paternity duo, with the alleged father having alleles 12,12.3 and 21,20 for marker 1 and 2 respectively, and the child is 13,14 and 22,25 for the same pair of markers. The ped-file will have the following appearance for the two hypotheses “father-child” vs “unrelated”

1 1 0 0 1 12.3 12 21 20
1 2 0 0 2 x x x x
1 3 1 2 1 13 14 22 25
2 1 0 0 1 12 12 21 20
2 2 0 0 2 x x x x
2 3 1 2 1 x x x x
2 4 3 2 1 13 14 22 25

The columns are defined as follows: (1) pedigree ID (2) Individual ID in a specific pedigree (3) Ind ID of father in pedigree, 0 if founder (4) Ind ID of mother in pedigree, 0 if founder (5) sex, 1 if male, 2 if female (6) marker 1 allele 1 (7) marker 1 allele 2 etc…

Calculation with ped-file

1. Chose the “New wizard” option in the File-menu.
2. Click on the “import ped file” button
3. Select the pre defined ped file (An error message is received if the ped file is incorrect)
4. Click on next (The result dialog will appear since all hypotheses and DNA data should be in the imported ped file)
5. Same as points 6-9 in *4.1.2* above.

Please not that simulations is not possible for imported pedigrees in the current version of FamLink.

4.2.3 Import a Familias-file

See section 5.1 below.

4.3 Perform simulation to study the impact of linkage for a given pedigree

The simulation tool is helpful to examine the impact of taking linkage into account or not, for any pair of linked markers and any given pedigrees. The simulation is a modified simulation part of the existing simulation module in Merlin.

The pedigree data is simulated based on the chosen pedigrees and the allele frequency database and the parameter defined in the options dialog (Theta , error rate). The LR is computed twice, once with the specified theta and once with theta=0.5. In the output, the effect of not accounting and accounting for linkage is expressed as the ratio LR(theta=0.5)/LR(theta=r). In the simulation report, a summary of the results from the simulations can be found, such as medians and percentiles.

To perform the simulation, just follow the steps given in 3.2.1 and 3.2.2 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) The simulation report can be generated using the Save Results button.

Comment: Generally, when simulating assuming a hypothesis H2, it may happen that the simulated data is inconsistent under the alternative hypothesis H1. Currently FamLink, estimates the impact of linkage, i.e., LR(unlinked)/LR(linked) assuming data consistent for both hypotheses. In addition, FamLink reports the number of inconsistent simulations.

5.1 Quick analysis (Beta)

The feature allows the import of a complete save file from Familias, containing marker allele frequencies (see appendix for a compilation of the currently available marker systems), family DNA-profiles for the included individuals and pedigree information for the tested hypotheses. A likelihood calculation (accounting for linkage) is performed for all included markers and stored as a txt-file (input filename + _FamLinked.txt). Additionally, a simple report can be generated. STR markers not included in the appendix will be excluded from the calculations. Unless at least two different pedigrees are specified, the function will return an error message. This function is useful when the intention is to compare a complete case with several markers in linkage (e.g. using PP16/ESX17/HDplex in the same case) and also with complex relationships.

5.2 Match probability calculation for relatives

Although FamLink is developed for use in relationship testing, some of the features of FamLink can be used to estimate the numerical effect of linkage in match probability calculation when the alternative man is a close relative to the suspect. See Buckleton et al for a theoretical description of the issue. As an example, think of a case where you have DNA data from a stain and from the suspect for two closely located loci. The hypotheses are that

the DNA in the stain comes from the suspect vs the DNA in the stain comes from a full brother of the suspect. Assume that you have the same data as in the upper part of Table 1 in Buckleton et al, meaning DNA data P,Q for marker 1 and U,V for marker 2. We further assume that the recombination frequnecy is 0.316 and that all allele frequencies are 0.1.

Buckleton et al calculated the numerical effect of linkage to 1.18. In FamLink this figure can be achived in the following way. Edit the database and adjust the recombination frequency to 0.316. Select the hypothesis "full sibling" together with the "Unrelated" hypothesis and enter the same DNA data for individual 1 and indivudal 2 (heterozygous at both loci). Calculate and select the "Full sibling" hypothesis and then click on "Save results" and chose "Multiplication factor". In this case FamLink estimate the factor to 1.18305, which is the same as in Buckleton et al. This factor means the if linkage is accounted for, it will result in a LR that is 1.18 times higher then the LR assuming no linkage.

5.3 Analyze Merlin files (Beta)

This feature is analougus to the simulation tool described above (4.3) with the differences that there is no limit in the number of markers that can be included and likelihood ratios calculated from different map-files (physical location of the markers) can be compared. This option is intended for more advanced users. The tool can be useful for mainly two things 1) Estimate the information content (given as distribution of LRs) for a defined case scenario and a given set of markers. 2) Estimate the effect of linkage when using multiple DNA markers.

Five files are needed for the analysis (for a comprehensive description of the input-files we refer to http://www.sph.umich.edu/csg/abecasis/Merlin/index.html). In short, first you need a ped-file (.ped) describing the pedigrees (hypotheses) for the different case scenarios that will be tested. All persons which are genotyped should be uniquelly identified in all pedigrees in order for the simulations to be correct. This is achieved by assigning unique dummy genotypes for each typed person. Further, persons which are not genotyped should have zeros as genotypes. Then you need a data file (.dat) which is a list of the DNA-markers and a file with allele frequencies (.freq) for the markers defined in the data file. Finally, two map-files (.map) are needed which contain information of the physical location of the markers defined in the data file. If the purpose of the study is to estimate the information content for a given case scenario, the two map-files can be identical. If the purpose is to study the effect of accounting for linkage or not, the first map-file should describe the "true" location of the

markers, and the second map-file should be constructed to artifially separate the markers to illustrate no linkage (i.e put the markers on a virtual distance of 1000cM).

The simulation is performed using the first map-file for the data generation for each pedigree listed in the ped-file. The LRs are computed for both map-files.

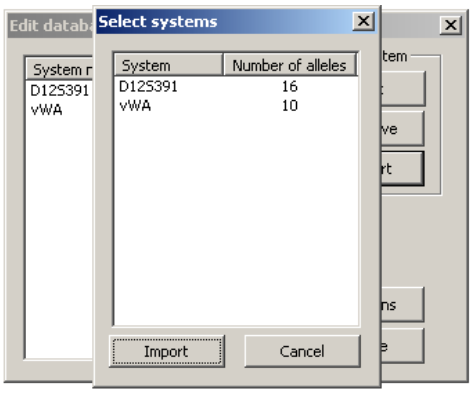
The output is a textfile that lists the LRs for all simulated cases (for all hypotheses) and also lists of the effect of linkage (LR(ignoring linkage)/LR(accounting for linkage)).

6. Examples

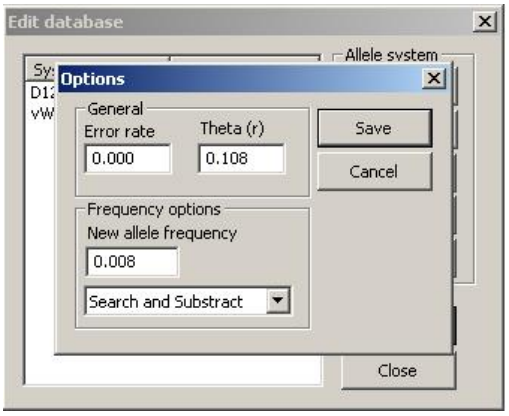
6.1 Example 1 – A standard case ("Uncle-child" versus "Unrelated")

In this example we want to calculate the likelihood ratio (LR) for the case "uncle-child" vs "unrelated", with data from a man (alleged uncle) and a child. The data consist of alleles for the marker vWA and D12S391, with a recombination frequency of 0.108. The alleged uncle is 12, 14 at vWA and 19, 29 at D12S391, and the child is 13, 13 at vWA and 19,21 at D12S391. We will use the U.S caucasian allele frequency database. (Found in Install folder\Examples, C:\Program files\FamLink\Examples if this it the FamLink install directory)

We start by importing the allele frequencies stored in the file Allele freq D12_vWA_US_cauc.txt

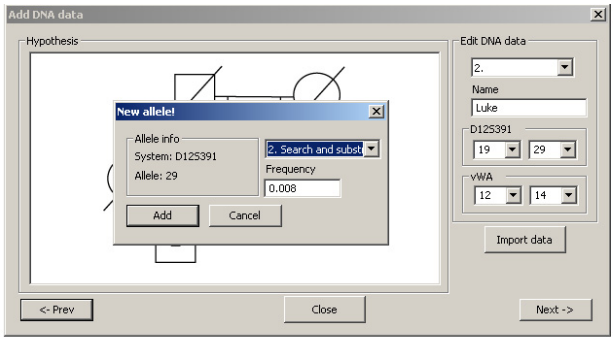


We continue to set the Theta (r)-value to 0.108, the error rate to 0 and the new frequency parameter to 0.008 with the "Search and Substract" option

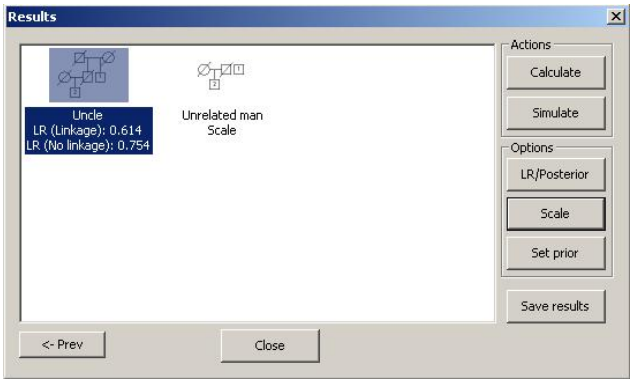


In the New wizard option, we select the “Uncle” hypothesis (Not “Uncle (data mother)”) among the predefined pedigrees and the click on next and then choose the “Unrelated man” hypothesis.

Moving on the Add DNA data tab, we select the first person, name him Luke and type his DNA data (19, 29 at D12S391 and 12, 14 at vWA). Note that since allele 29 is not in the database, we need to type this and choose appropriate allele frequency and method for addition in the existing database. (Default values are previously defined in the Options dialog)



We do the same for the child Sam (19,21 at D12S391 and 13,13 at vWA), and move on to the result tab where we perform the calculation and get the following output.



LR(Linkage)=0.614 and LR(No linkage)=0.754. We generate the report (via Save results).

6.2 Example 2 – Simulation

In this example we will study the impact of ignoring linkage between the STR loci SE33 and D6S1043, which are located only 4.4 cM apart (e.g. r=0.044), for the case incest vs no incest.

We import the appropriate allele frequencies. In this example, we will use published allele frequencies for the Chinese Han population (see example files). We set the Theta (r)-value to 0.044 and the error rate to 0. Via the “new wizard” option, we chose the “Incest” hypothesis first and then the “No incest” hypothesis. We skip the add DNA data window by pressing “next”, and then click on the simulate button in the results tab. We chose to perform the simulation with 10,000 runs. We also chose to save the LRs from the simulation, together with a summary of the outcome of the simulations. We finally generate a report.

7. References

Hill, C.R., Duewer, D.L., Kline, M.C., Sprecher, C.J., McLaren, R.S., Rabbach, D.R., Krenke, B.E., Ensenberger, M.G., Fulmer, P.M., Stort, D.R., Butler, J.M. (2011) Concordance and population studies along with stutter and peak height ratio analysis for the PowerPlex® ESX 17 and ESI 17 Systems. *Forensic Sci. Int. Genet.* 5(4): 269-275.

Huang S., Zhu Y., Shen X., Le X.,Yan H. (2010) Genetic variation analysis of 15 autosomal STR loci of AmpF'STR Sinofiler™ PCR Amplification Kit in Henan (central China) Han population. *Legal Medicine* 12: 160–161.

Liu C., Harashima N., Katsuyama Y., Ota M., Arakura A., Fukushima H. (1997) ACTBP2 gene frequency distribution and sequencing of the allelic ladder and variants in the Japanese and Chinese populations. *Int J Legal Med* 110 : 208–212.

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.

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.

C. Phillips, D. Ballard, P. Gill, D. S. Court, A. Carracedo and M.V. Lareu, The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data. *Forensic Sci. Int. Genet* (2011) doi:10.1016/j.fsigen.2011.07.012 (in press).

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.

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) 47-59.