X-linked markers: Repetition, haplotype frequencies, FamLinkX introduction

Thore Egeland[(1),(2)]

(1) Norwegian University of Life Sciences, (2) NIPH

http://familias.name/ghep2016
thore.egeland@gmail.com

Contents

- Repetition
  - Inheritance
  - Linkage. Recombination.
  - Linkage disequilibrium (LD) (gametic association). Haplotypes.
  - LR
    - GHEP-ISFG Decaplex [3], ARGUS X-12 [4],....
- Estimating haplotype frequencies: λ model.
- FamLinkX [5]. Demo . Exercises

FamLinkX Exercises. http://familias.name/ghep2016

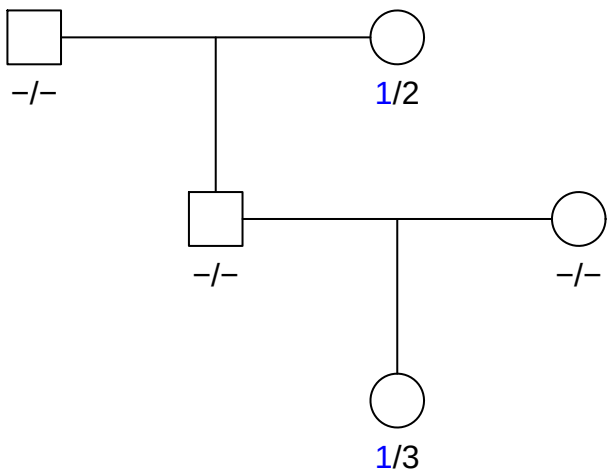
FamLinkX, developed by Daniel Kling. The participants should bring a laptop with FamLinkX, preferably version 2.5 (available from Aug 22 2016) downloaded (http://famlink.se/fx_download.html).

- Assistance for installation will be provided at the workshop, if needed. Note that installation is typically only possible if you are administrator on your laptop.

Time schedule

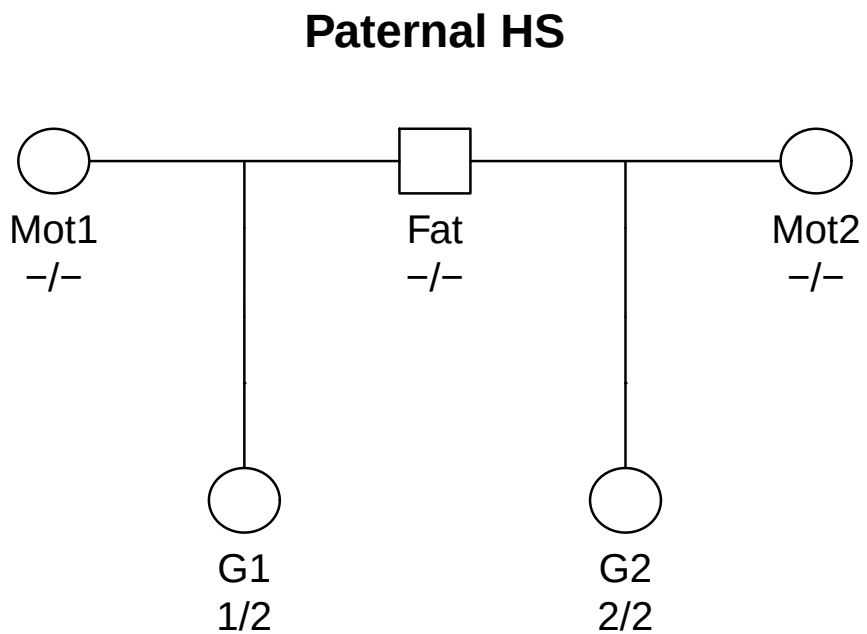
- 13.30 – 15.45 X-linked markers : Repetition, haplotype frequencies. FamLinkX demo video . More videos (Chapter 4). Thore Egeland
- 15.45 – 16.15 Coffee break
- 16.15 – 18.00 Exercises: The FamLinkX exercises 4.12-4.14, 4.16-4.18 are from the book "Relationship Inference with Familias and R" by Egeland, Kling, and Mostad available from Elsevier. The required material is freely available: exercises , solutions , zipped input files . Thore Egeland

X-chromosomal inheritance. Paternal Grand daughter



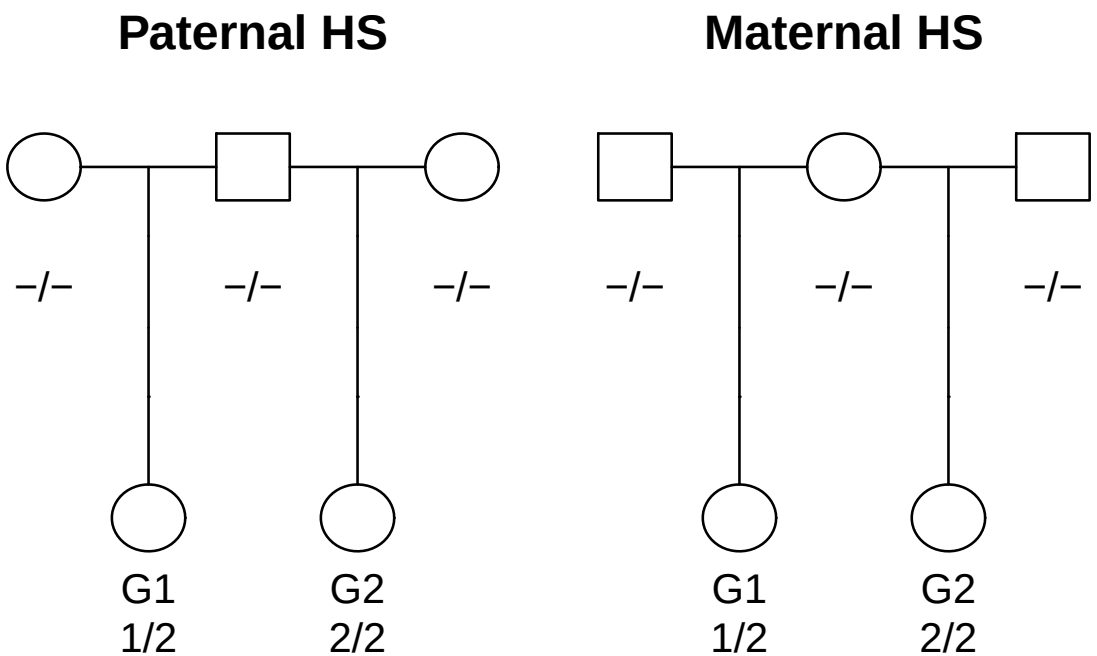
- Paternal grand daughter should share one X-allele IBD [6] with grand mother.
- One allele shared IBD with prob. 0.5 for *autosomal* marker.
- One X-allele shared IBD with prob. 0.5 for **maternal** grand daughter.

X-chromosomal inheritance. Half sisters



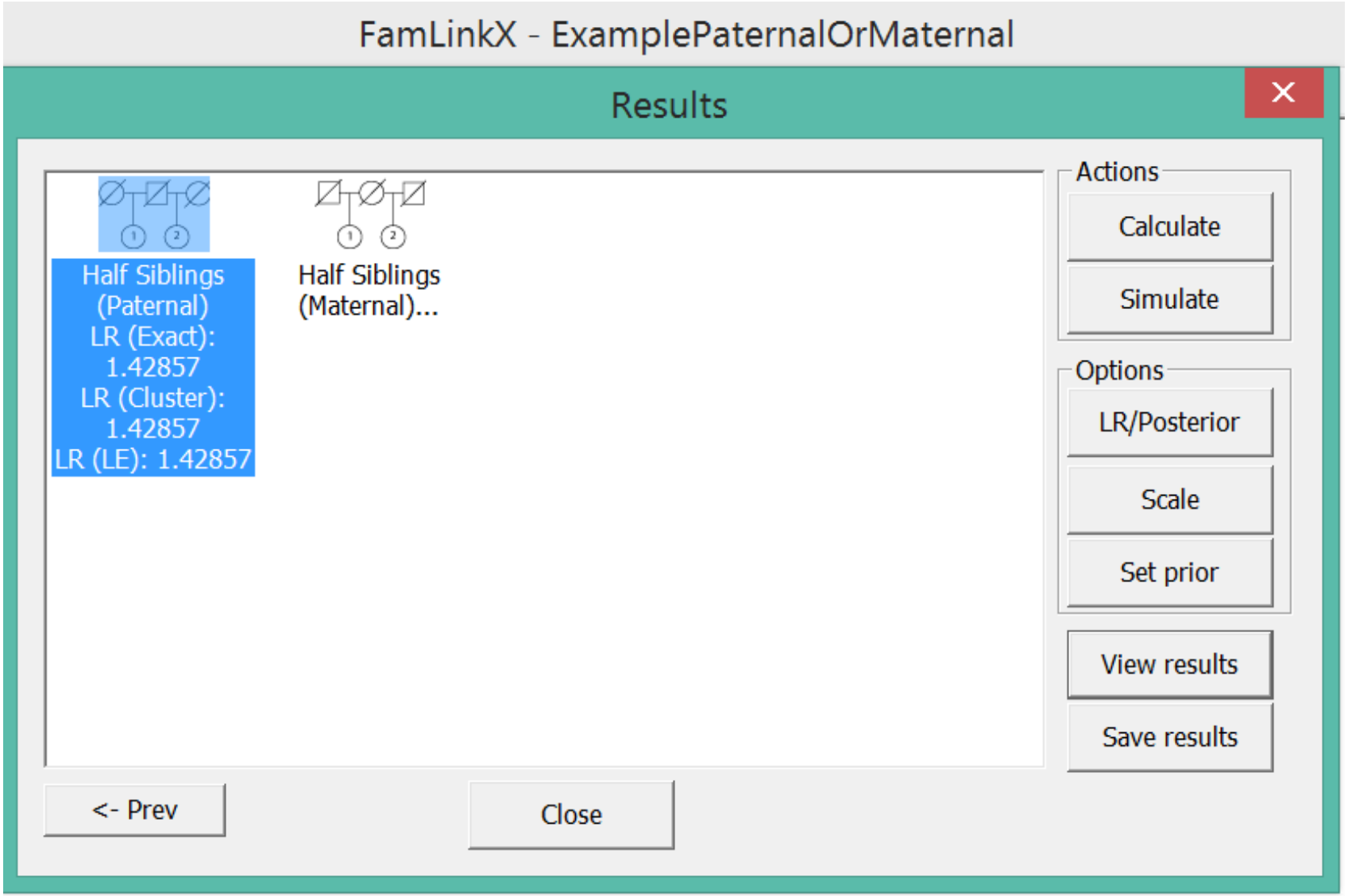
- Paternal half sisters should share one X-allele IBD.
- One allele shared IBD with prob. 0.5 for *autosomal* marker.
- One X-allele shared IBD with prob. 0.5 for **maternal** half sisters.

Paternal or maternal half sisters?

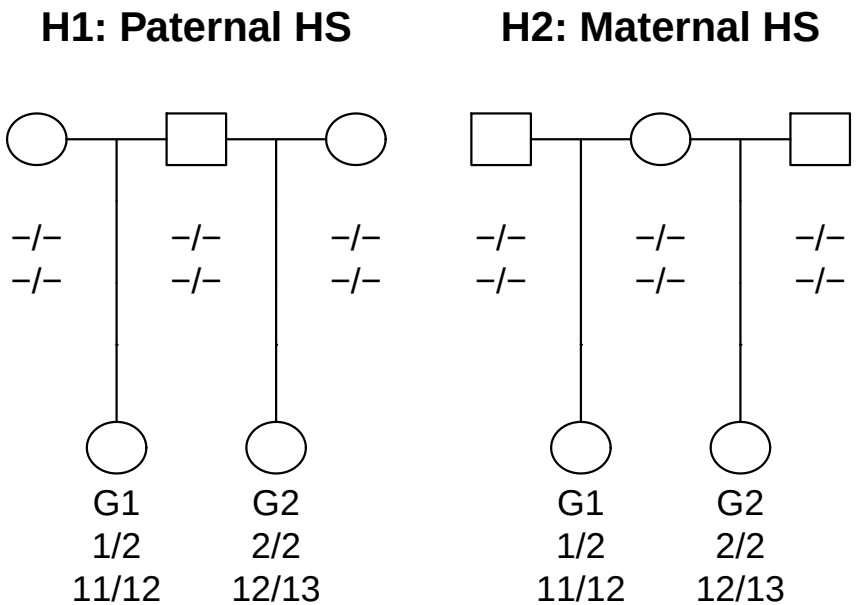


$$LR_1 = \frac{p_1 p_2^2}{\frac{1}{2} 2 p_1 p_2^3 + \frac{1}{2} p_1 p_2^2} = \frac{1}{p_2 + \frac{1}{2}} > 1 \text{ if } p_2 < \frac{1}{2}. \; LR_1 \overset{p_2=0.2}{=} 1.42857$$

FamLinkX: http://familias.name/ExamplePaternalOrMaternal.sav



Two markers: Multiply?: $LR =^{?} LR_1 * LR_2$

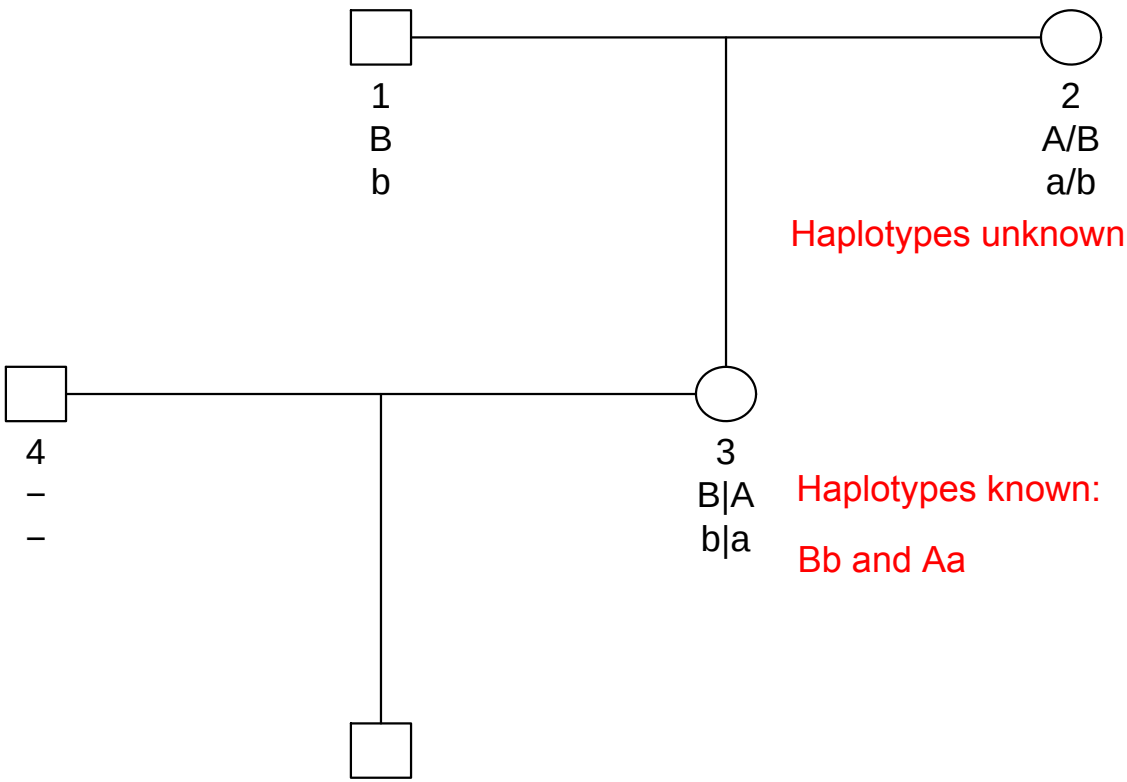


$$LR_2 = \frac{p_{11}p_{12}p_{13}}{\frac{1}{2}2p_{11}p_{12}2p_{12}p_{13} + \frac{1}{2}p_{11}p_{12}p_{13}} = \frac{1}{2p_{12} + \frac{1}{2}} > 1 \text{ if } p_2 < \frac{1}{4}.$$

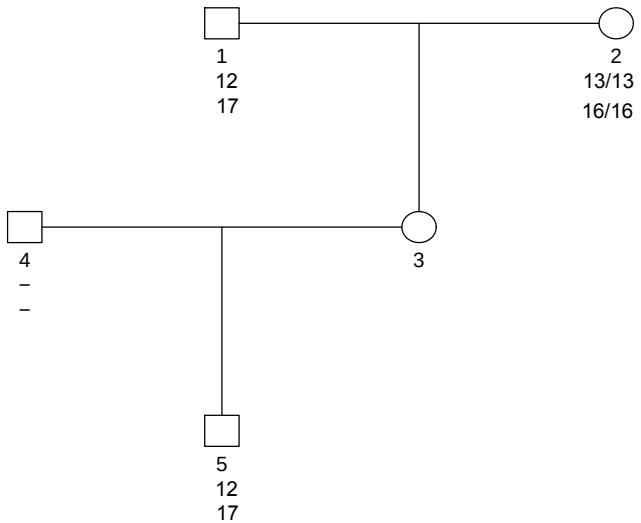
Assumptions

- We can only multiply LR-s when markers are independent, i.e., $LR \neq LR_1 * LR_2$.
- Most markers on the X–chromosome are **dependent**.
- Calculations need to account for dependence and software is needed.
- Dependence arises because of linkage and linkage disequilibrium (LD), Linkage Equilibrium (LE) (no gametic association), briefly reviewed next.

Linkage. Haplotypes. Recombination

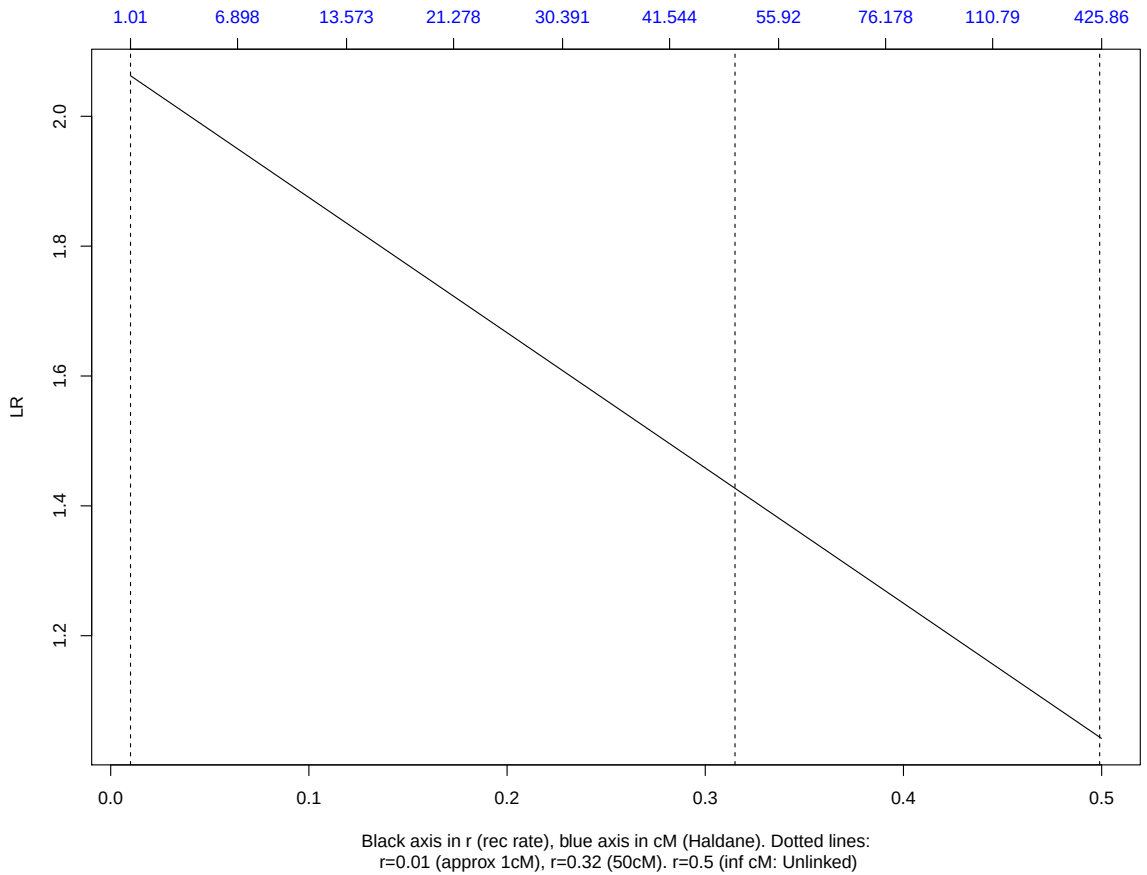


Maternal grandson or unrelated? Mother missing.



$$LR = \frac{\frac{1}{2}(1-r)}{p_{12,17}} \overset{LE}{=} \frac{\frac{1}{2}(1-r)}{p_{12}p_{17}} \overset{r=0.01}{=} \frac{\frac{1}{2}(1-0.01)}{0.6*0.4} = 2.0625$$

$$LR \overset{r=0.5}{=} \frac{1}{2p_{12}}\frac{1}{2p_{17}} = 1.04.$$

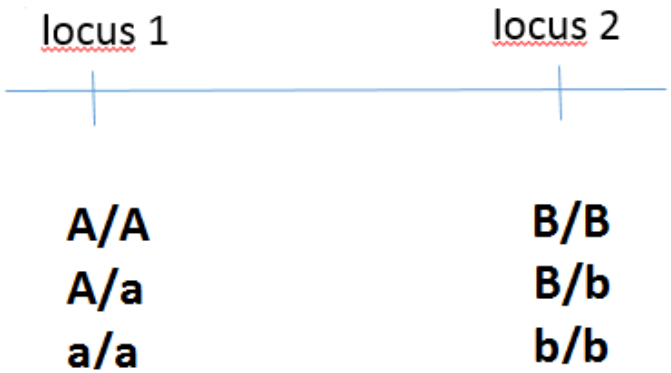


Black axis in r (rec rate), blue axis in cM (Haldane). Dotted lines:
r=0.01 (approx 1cM), r=0.32 (50cM). r=0.5 (inf cM: Unlinked)

- Haldane's map function:

$$r = \frac{1 - \exp(-2 * x/100)}{2} \underset{=}{x=50cM} \frac{1 - \exp(-2 * 50/100)}{2} = 0.316$$

LE: multiply



loc1	loc2	freq1	freq2	$P(hap \mid LE)$
A	B	0.2	0.3	0.2*0.3=0.06
A	b	0.2	0.7	0.2*0.7=0.14
a	B	0.8	0.3	0.8*0.3=0.24
a	b	0.8	0.7	0.8*0.7=0.56
tot				1.00

Table 1: LE based haplotype frequency estimates

Haplotype frequencies II. Data: count

loc1	loc2	freq1	freq2	$P(hap \mid LE)$	Count	$P(hap \mid Count)$
A	B	0.2	0.3	0.2*0.3=0.06	10	10/100=0.10
A	b	0.2	0.7	0.2*0.7=0.14	15	15/100=0.15
a	B	0.8	0.3	0.8*0.3=0.24	25	25/100=0.25
a	b	0.8	0.7	0.8*0.7=0.56	50	50/100=0.50
tot				1.00	100	1.00

Table 2: LE and count based haplotype frequency estimates

Problems with both methods (LE and count method)

- LE often far from valid.
- We typically don't observe all haplotypes:
  - Two markers, 10 alleles each: $10 \times 10 = 100$ haplotypes.
  - Three markers, 10 alleles each: $10 \times 10 \times 10 = 1000$ haplotypes.

Need to estimate frequency greater than 0 in a **non ad-hoc way**

How big databases are needed?

- Typically, larger number of haplotypes are needed compared with allele databases.
- Power calculations can be done in each case.
- Basic idea simplified: Assume true haplotype frequency is 0.1; specify accepted length of CI (confidence interval):
  - 95% CI from 0.09 to 0.11: Large sample needed.
  - 95% CI from 0.01 to 0.19: Small sample needed.
- Alternative approach: "Estimating Haplotype Frequency and Coverage of Databases" [2].

Estimating haplotype frequencies

$$F_i = \frac{c_i + \lambda p_i}{C + \lambda}. \tag{1}$$

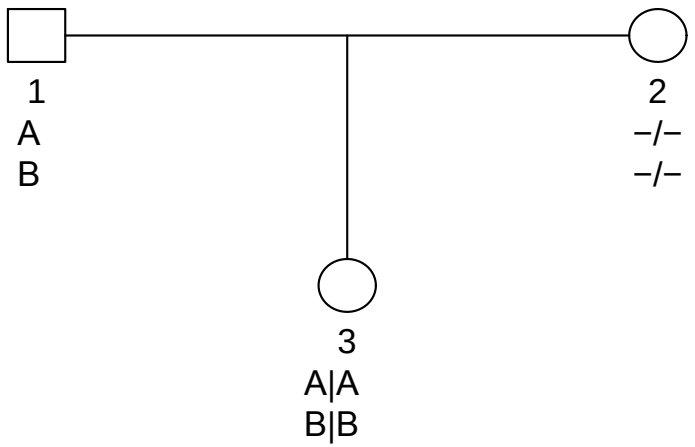
- F_i: Updated haplotype frequency,
- c_i: Count of haplotype i,
- C: Total number of haplotypes,
- p_i: Expected haplotype frequency,
- $\lambda > 0$: prior weight.

Example I

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

haplo	$P(hap \mid LE)$	Count	$\lambda = 100$ method
AB	$p_1 = 0.06$	$c_1 = 10$	$\frac{10+100*0.06}{100+100} = 0.080$
Ab	$p_2 = 0.14$	$c_2 = 15$	$\frac{15+100*0.14}{100+100} = 0.145$
aB	$p_3 = 0.24$	$c_3 = 25$	$\frac{25+100*0.24}{100+100} = 0.245$
ab	$p_4 = 0.56$	$c_4 = 50$	$\frac{50+100*0.56}{100+100} = 0.530$
tot	1.00	$C = 100$	1.000

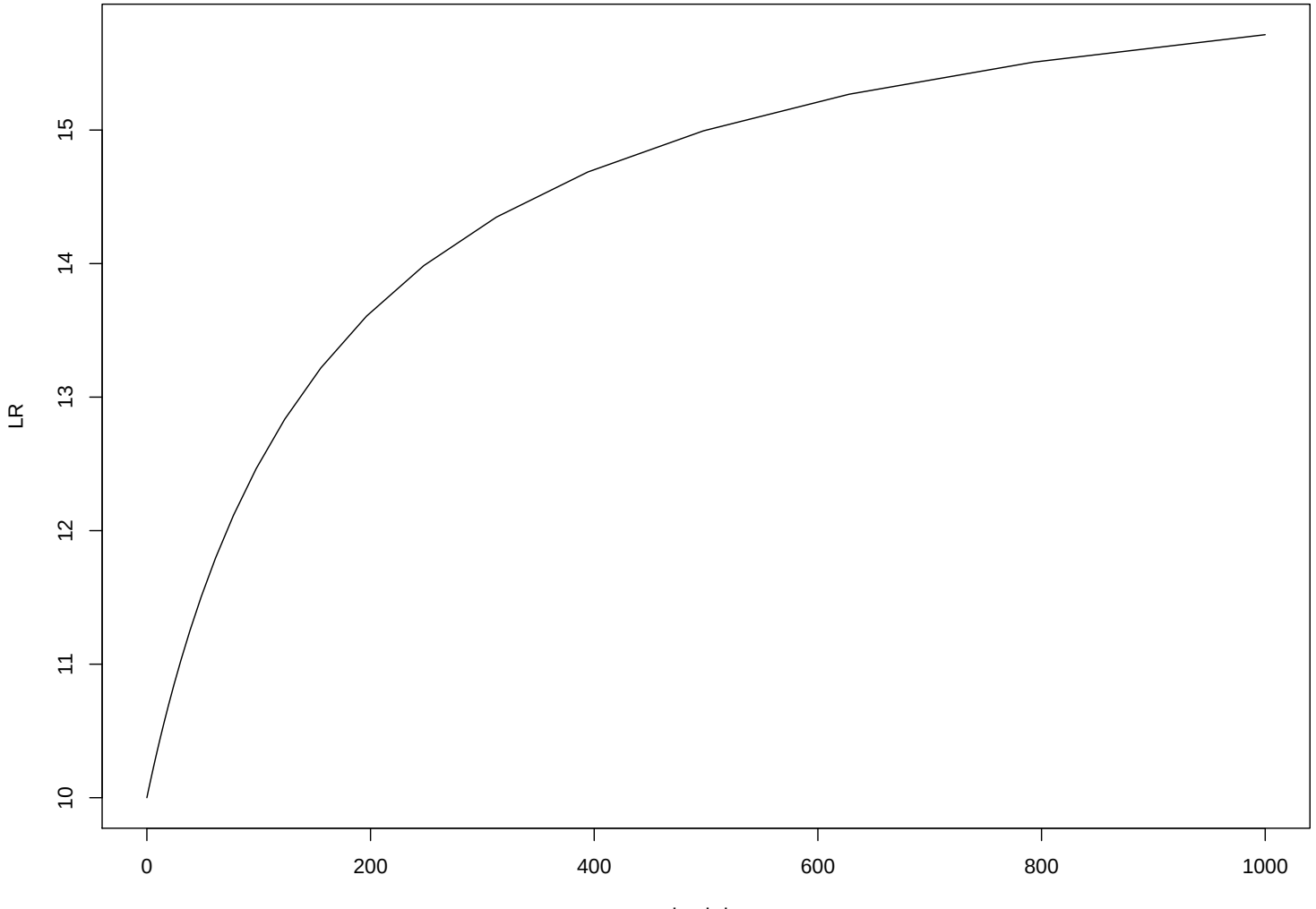
Example I contd: LR



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

$$LR = \frac{1}{\text{freq haplotype AB}} = \frac{C + \lambda}{c_i + \lambda p_i} = \frac{100 + \lambda}{10 + \lambda * 0.06}.$$

Example I contd: $LR = \frac{100+\lambda}{10+\lambda*0.06}$



Example II

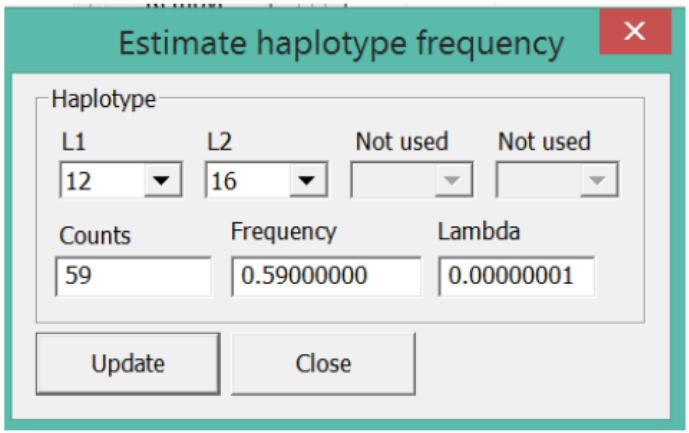
haplo	$P(hap \mid LE)$	Count	$\lambda = 0$	$\lambda = 100$	$\lambda = 10000$
AB	$p_1 = 0.06$	$c_1 = 10$	0.10	0.080	0.0604
Ab	$p_2 = 0.14$	$c_2 = 15$	0.15	0.145	0.1401
aB	$p_3 = 0.24$	$c_3 = 25$	0.25	0.245	0.2401
ab	$p_4 = 0.56$	$c_4 = 50$	0.50	0.530	0.5594
tot	1.00	$C = 100$	1.000	1.000	1.0000

Table 3: λ effect illustrated

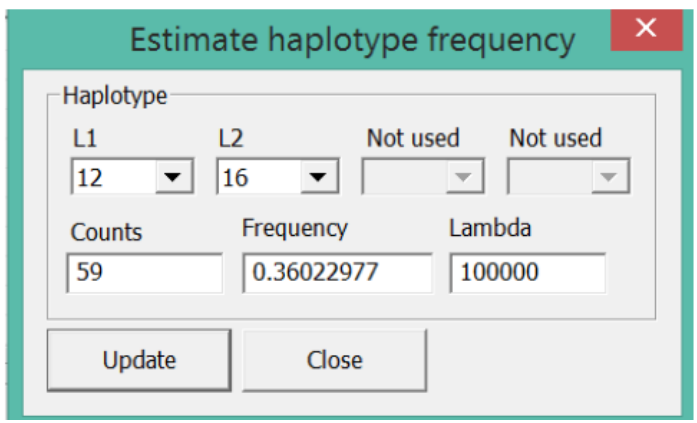
- If $\lambda = 0$: count estimate,
- If $\lambda = \infty$: LE estimate,
- otherwise weighted average.

Example III

Example. Data Exercise 4.12



$$\frac{59 + 0.00000001 \times 0.6 \times 0.6}{100 + 0.00000001} = 0.59$$



$$\frac{59 + 100000 \times 0.6 \times 0.6}{100 + 100000} = 0.36$$

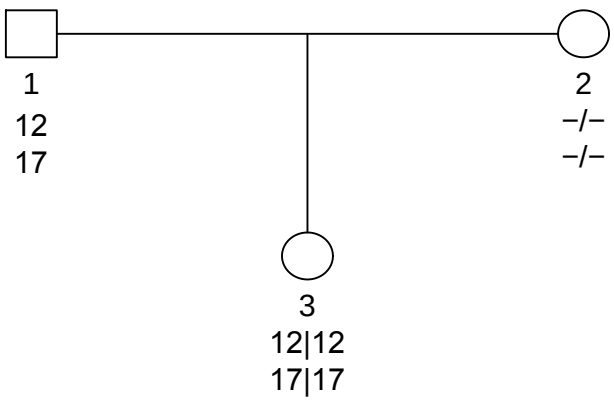
λ: Practical suggestion.

- λ estimates: Egeland, Kling, Mostad [1].
- FamLinkX 2.5 implementation: Generates R code. See video Exercise 4.14: `http://familias.name/VideosBook.pdf`
- Practical solution:
  calculate LRs with a selection of different values, say

$$\lambda = 0.01, 1, 100, 10000$$

and report the least extreme LR.

FamLinkX. Demo: Exercise 4.12. Video: http://familias.name/VideosBook.pdf

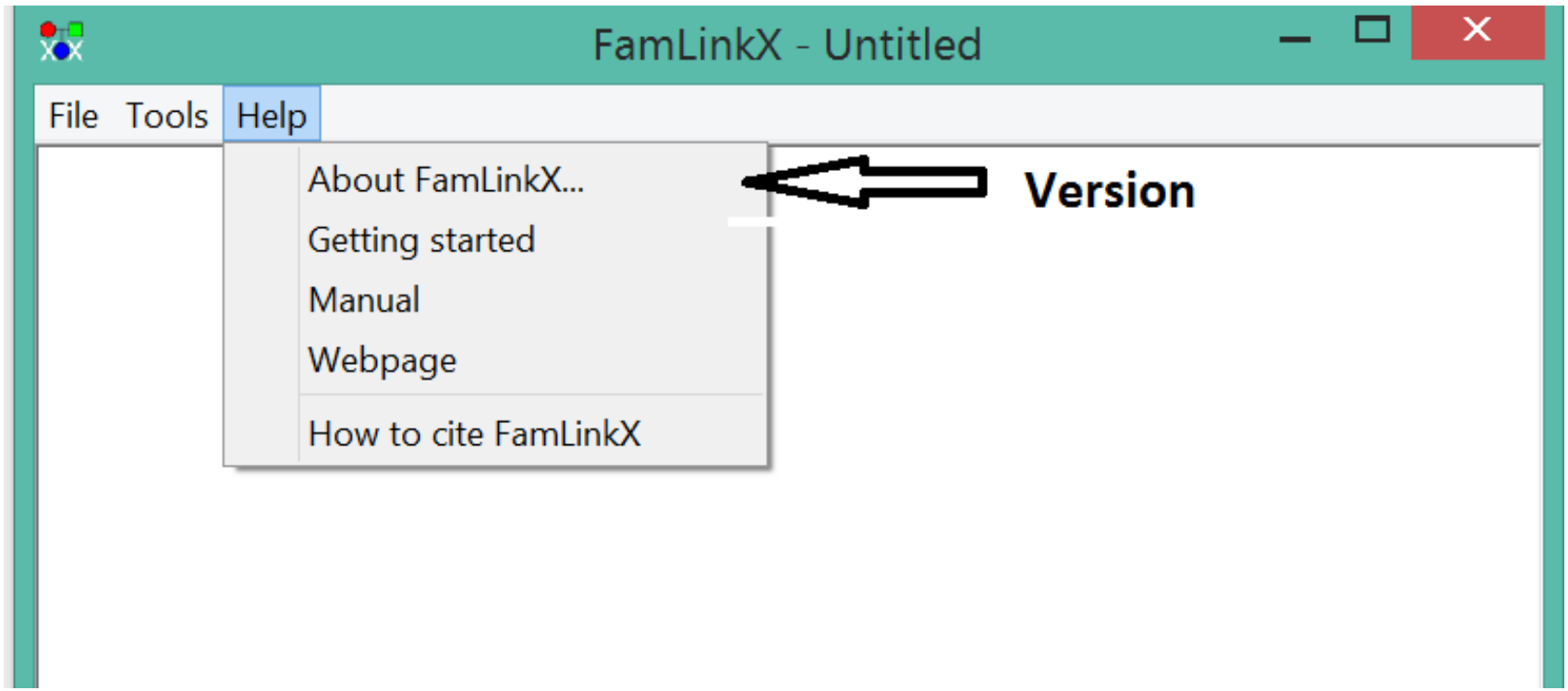


Haplotype ID	Number	L1	L2
1	59	12	16
2	1	12	17
3	1	13	16
4	39	13	17

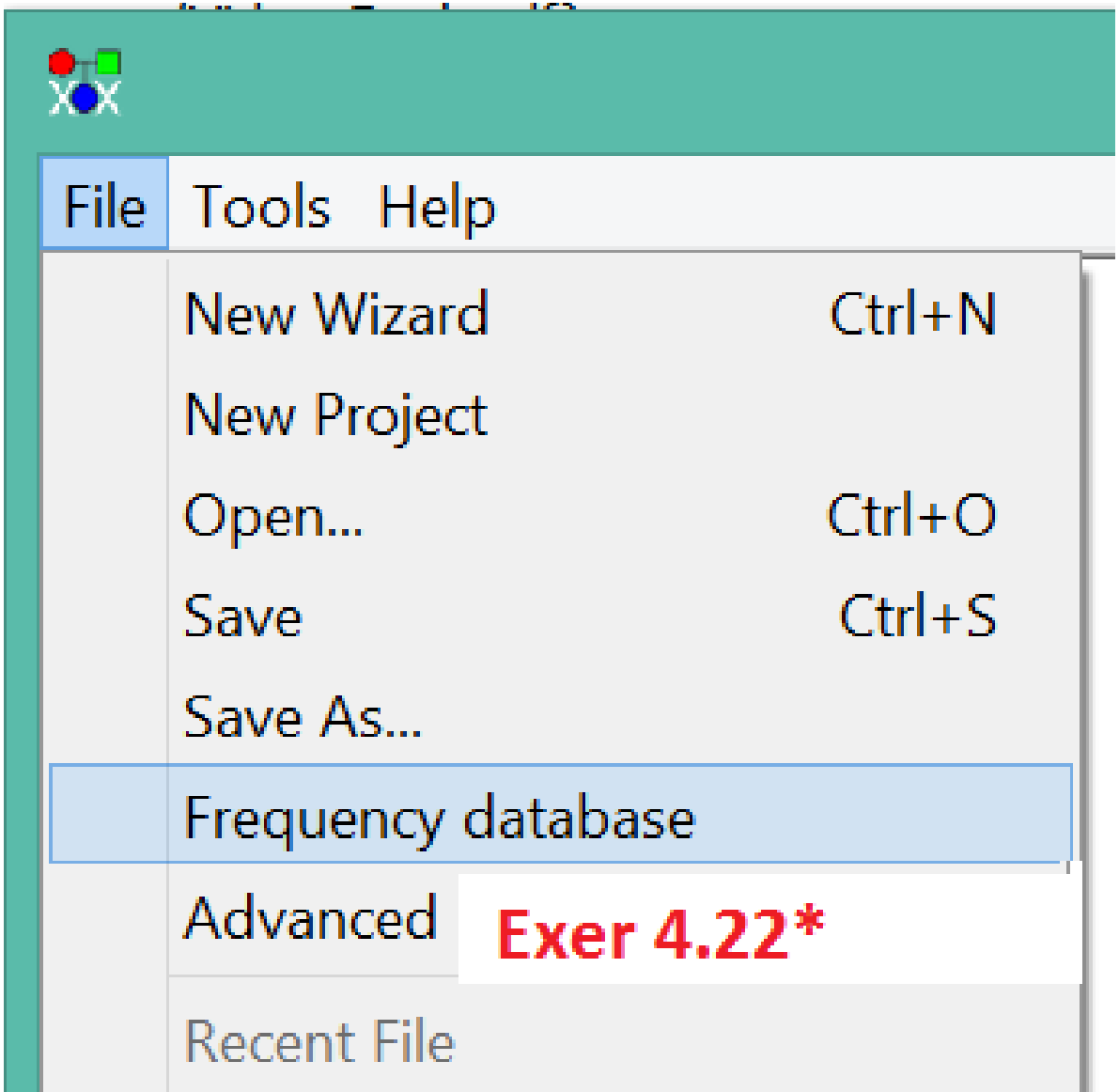
$H_{12,17}$ = estimated haplotype frequency

$$LR_{COUNT} = \frac{1}{H_{12,17}} = \frac{1}{1/100} = 100, LR_{LE} = \frac{1}{p_{12}p_{17}} = \frac{1}{0.6 \times 0.4} = 4.17.$$

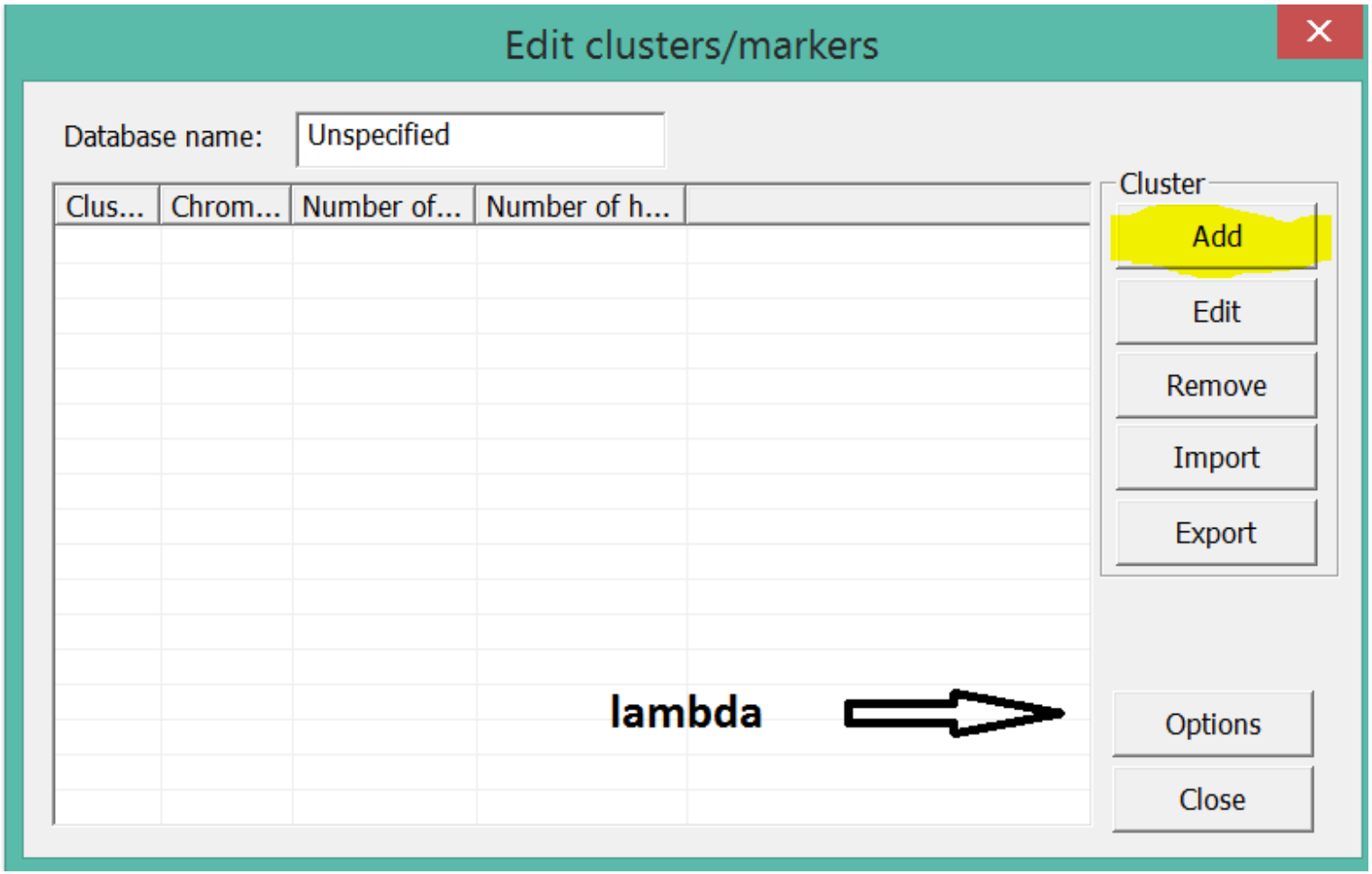
FamLinkX. Screenshots Exercise 4.12.



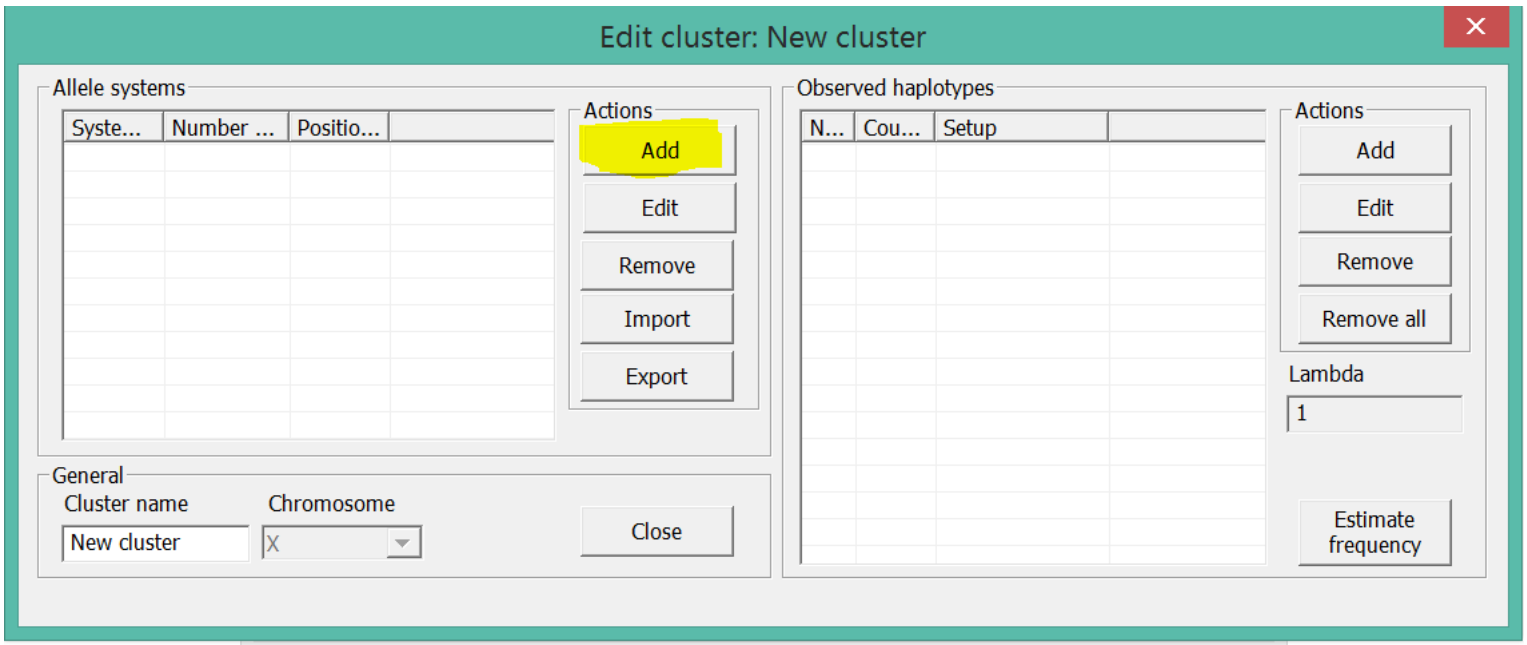
Screenshots Exercise 4.12. Frequency database



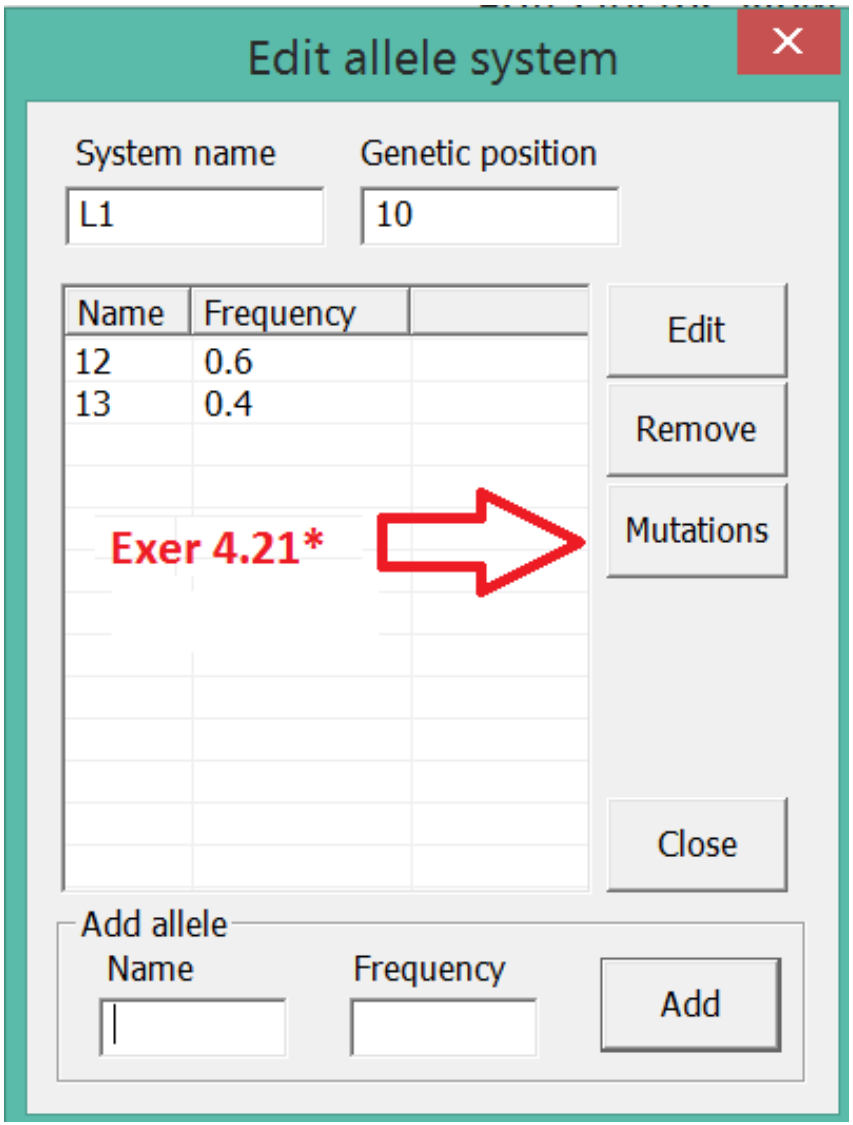
Editing clusters/markers I



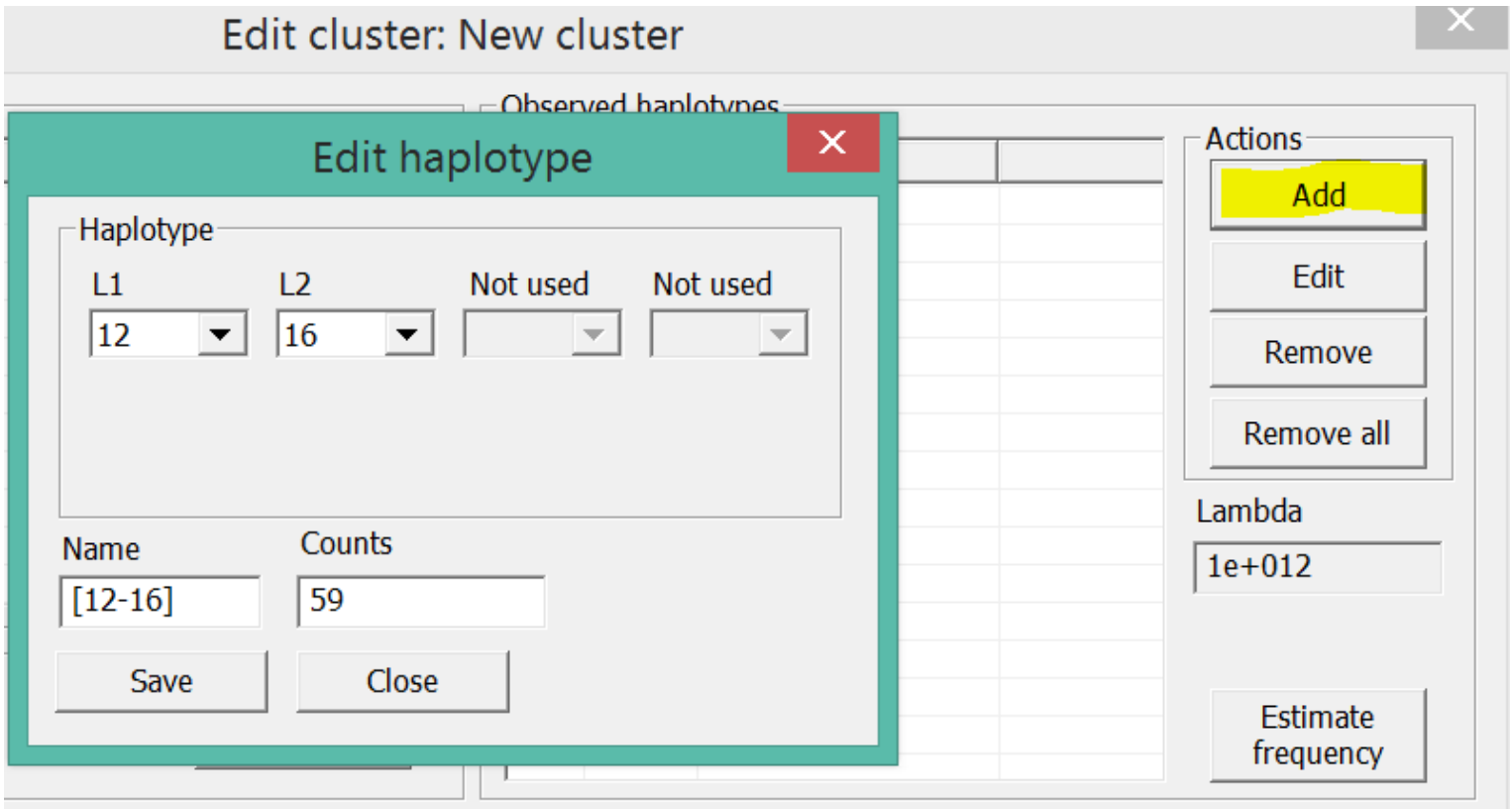
Editing clusters/markers II



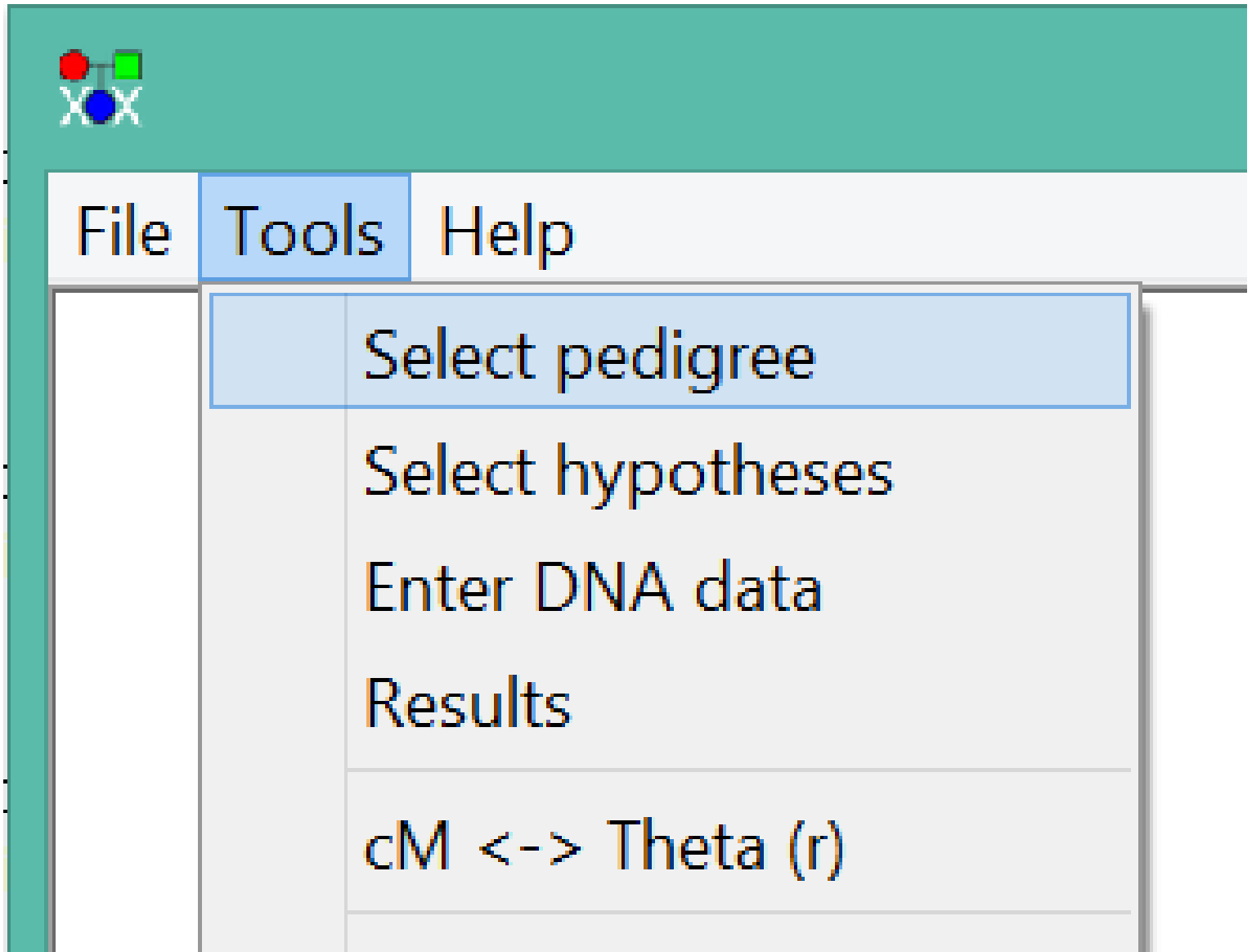
Editing clusters/markers III



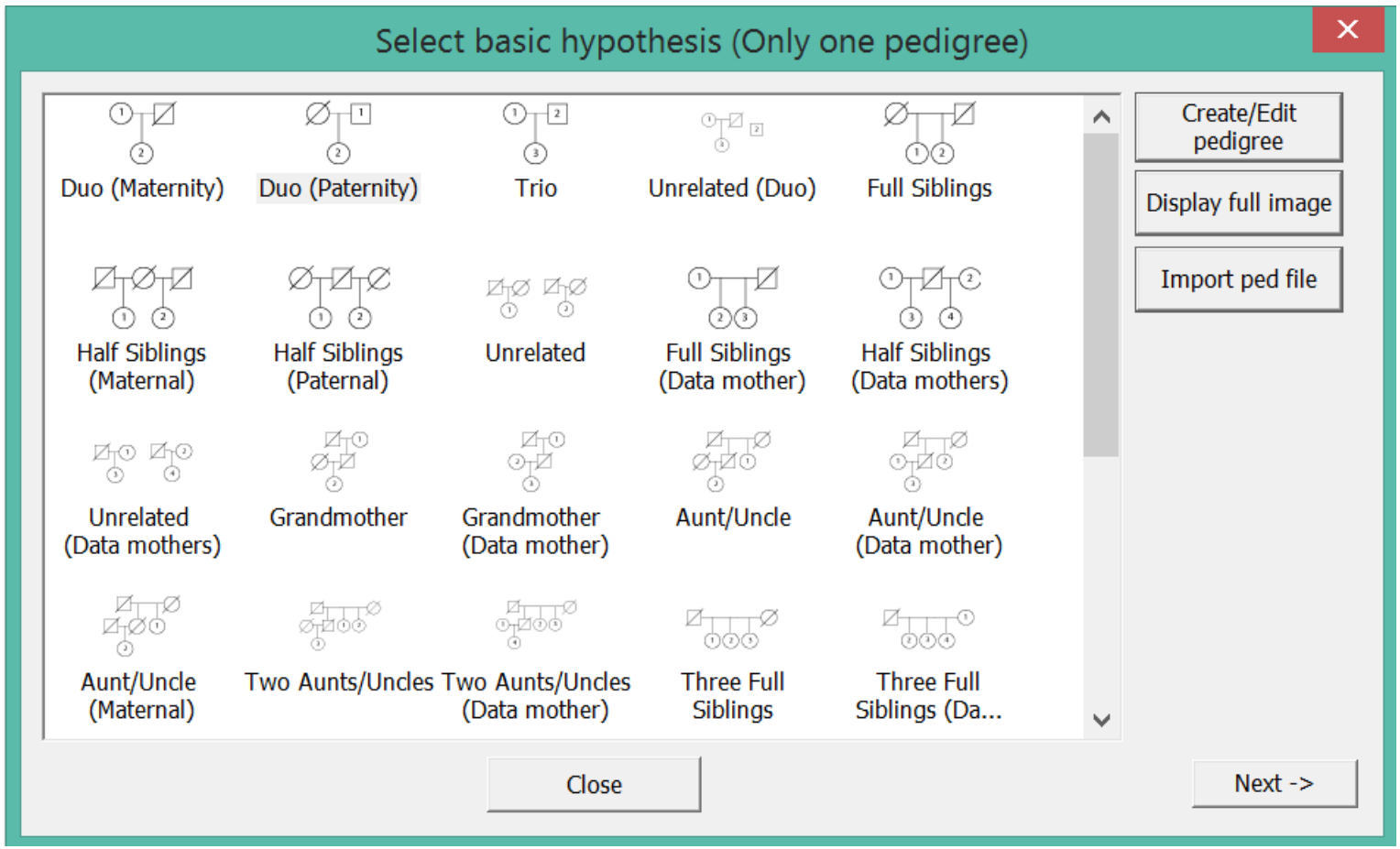
Editing clusters/markers IV



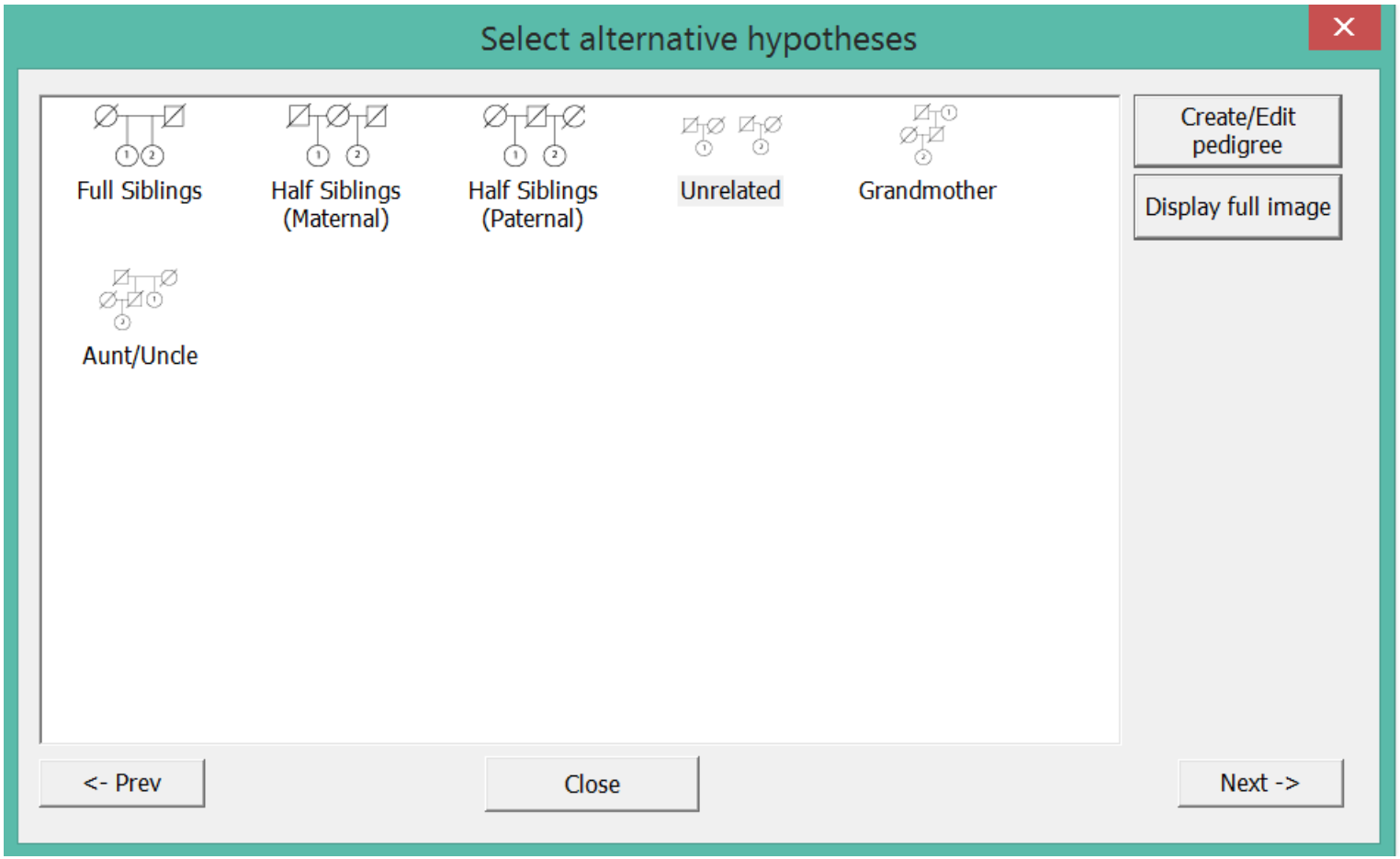
Tools I


File
Tools
Help
Select pedigree
Select hypotheses
Enter DNA data
Results
cM <-> Theta (r)

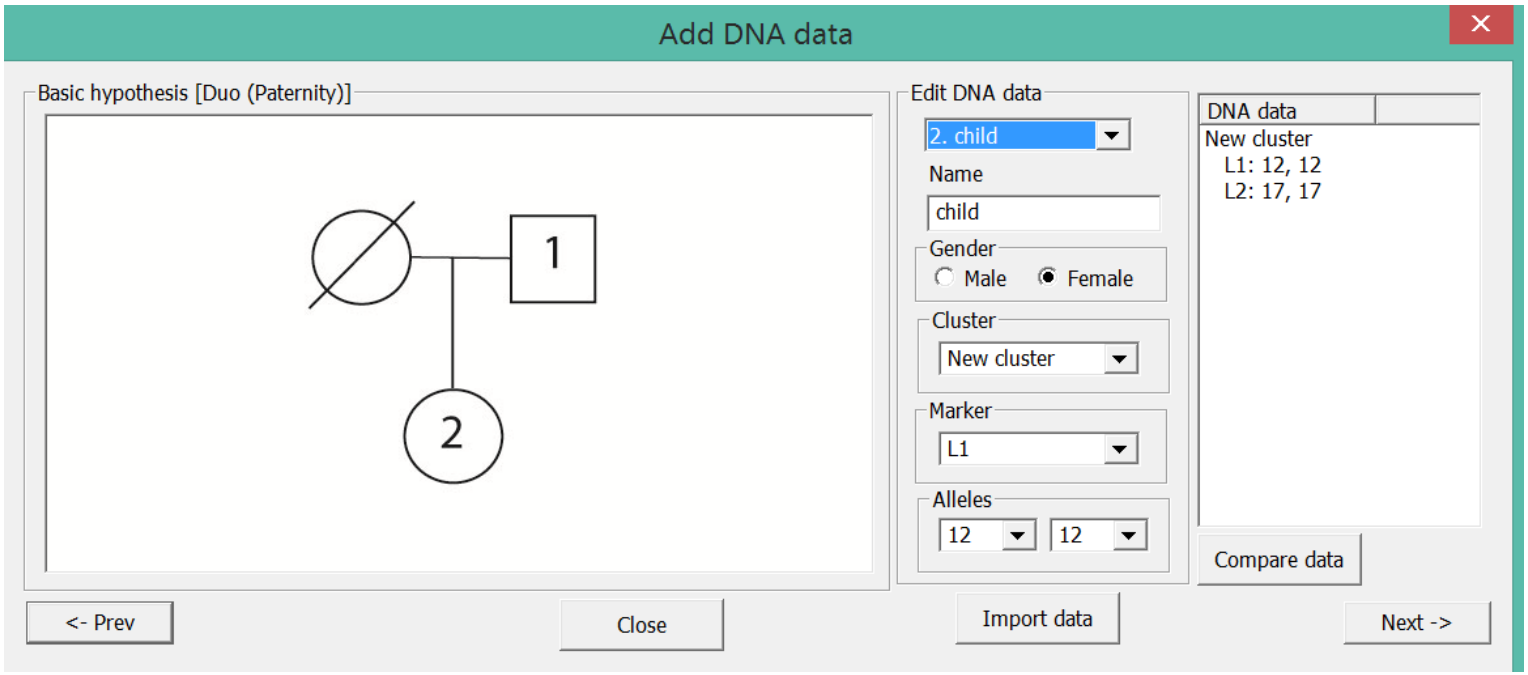
Tools II


Select basic hypothesis (Only one pedigree)
Duo (Maternity)
Duo (Paternity)
Trio
Unrelated (Duo)
Full Siblings
Half Siblings (Maternal)
Half Siblings (Paternal)
Unrelated
Full Siblings (Data mother)
Half Siblings (Data mothers)
Unrelated (Data mothers)
Grandmother
Grandmother (Data mother)
Aunt/Uncle
Aunt/Uncle (Data mother)
Aunt/Uncle (Maternal)
Two Aunts/Uncles
Two Aunts/Uncles (Data mother)
Three Full Siblings
Three Full Siblings (Da...
Create/Edit pedigree
Display full image
Import ped file
Close
Next ->

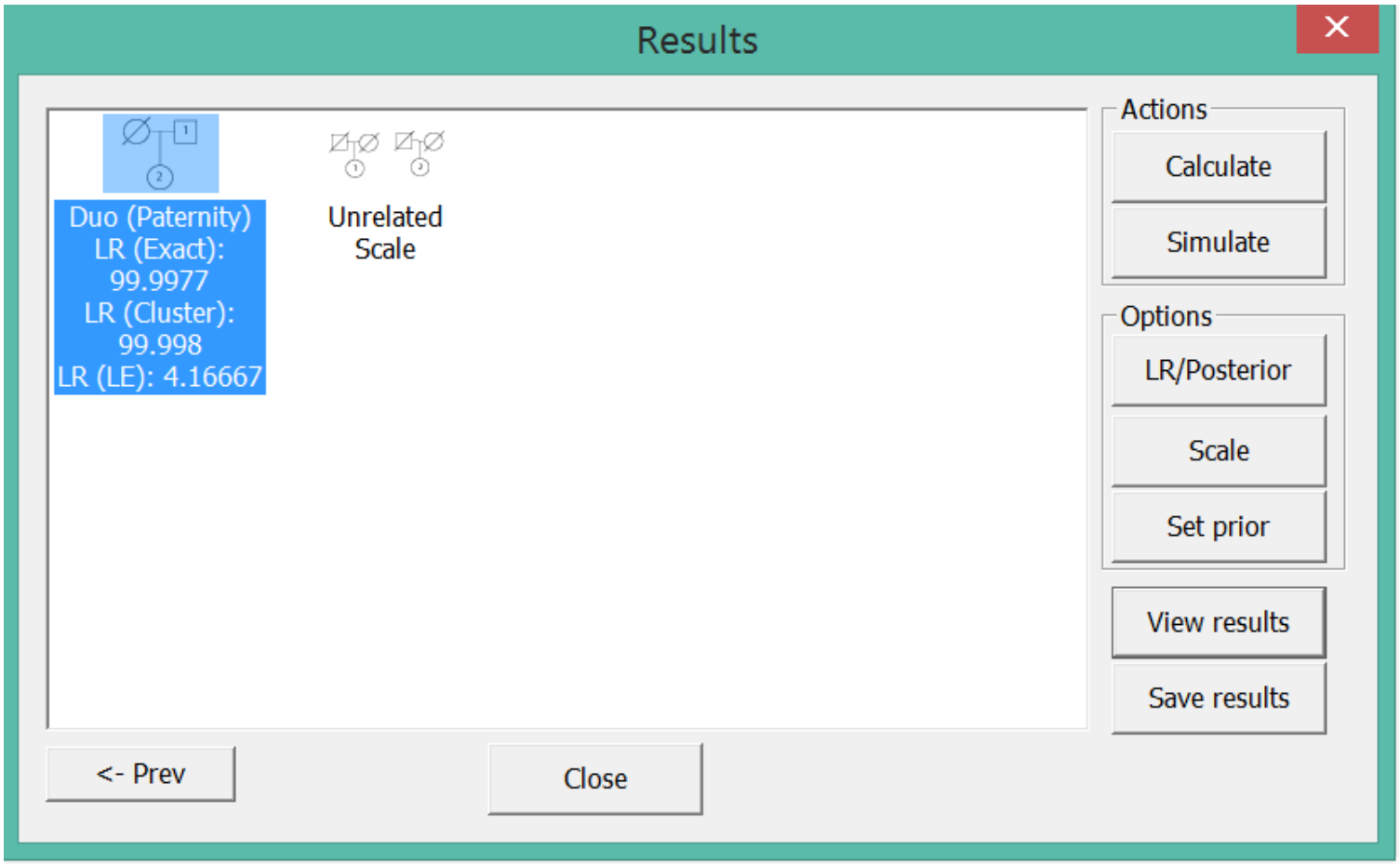
Tools III


Select alternative hypotheses
Full Siblings
Half Siblings (Maternal)
Half Siblings (Paternal)
Unrelated
Grandmother
Aunt/Uncle
Create/Edit pedigree
Display full image
<- Prev
Close
Next ->

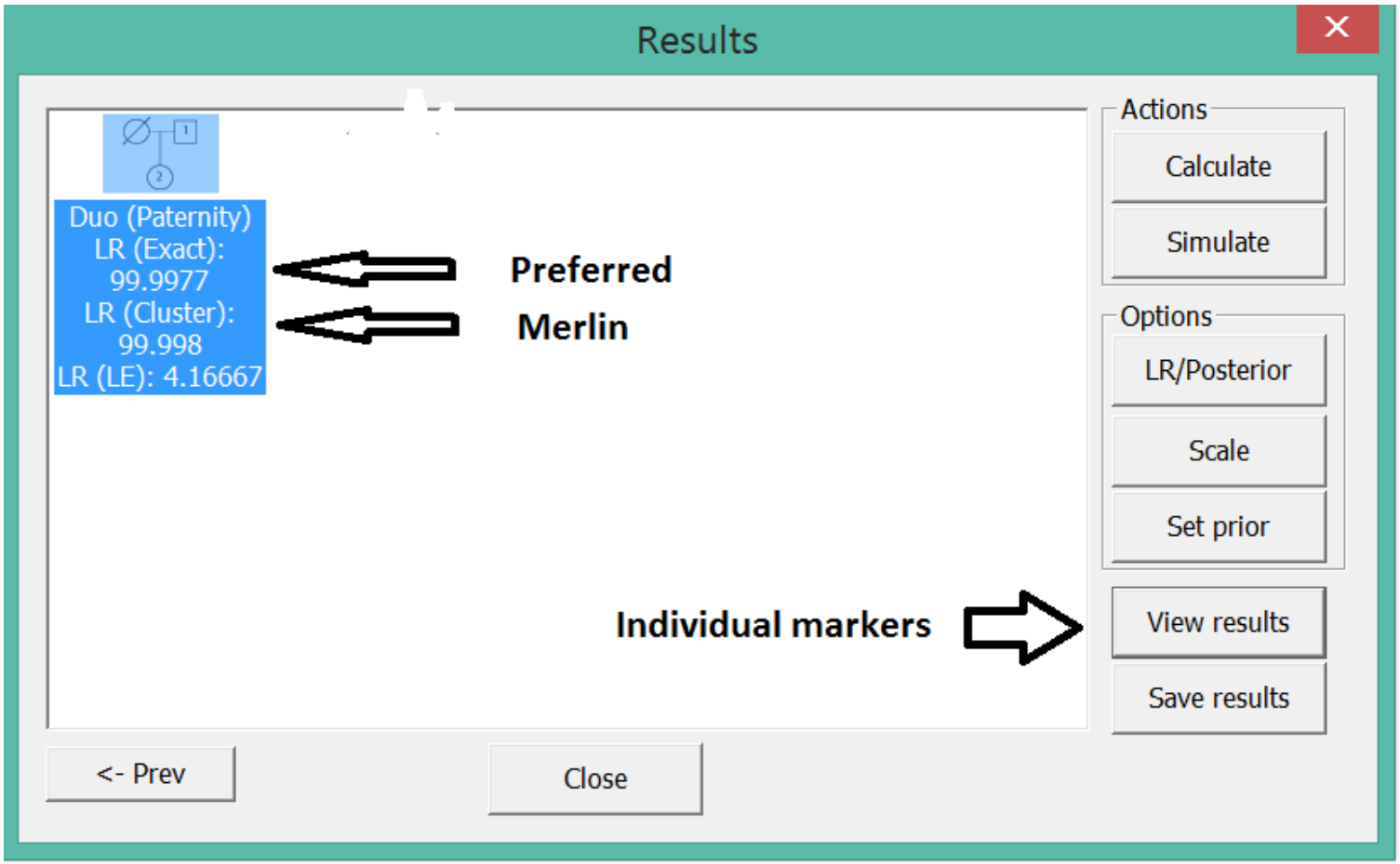
Tools IV


Add DNA data
Basic hypothesis [Duo (Paternity)]
1
2
Edit DNA data
2. child
Name
child
Gender
Male
Female
Cluster
New cluster
Marker
L1
Alleles
12
12
DNA data
New cluster
L1: 12, 12
L2: 17, 17
Compare data
<- Prev
Close
Import data
Next ->

Results


Results
Duo (Paternity)
LR (Exact):
99.9977
LR (Cluster):
99.998
LR (LE): 4.16667
Unrelated
Scale
Actions
Calculate
Simulate
Options
LR/Posterior
Scale
Set prior
View results
Save results
<- Prev
Close

Results


Results
Duo (Paternity)
LR (Exact):
99.9977
LR (Cluster):
99.998
LR (LE): 4.16667
Preferred
Merlin
Individual markers
Actions
Calculate
Simulate
Options
LR/Posterior
Scale
Set prior
View results
Save results
<- Prev
Close

Three computational models in FamLinkX

M1 Exact Linkage, LD (within clusters) and mutations accommodated.
Preferred model, but not implemented for *user defined* pedigrees.

M2 Cluster Not recombinations in clusters, LD (within clusters), not mutations.

M3 Only linkage.

Thore Egeland, Daniel Kling, and Petter Mostad.

Relationship Inference with Familias and R. Statistical Methods in Forensic Genetics.
Elsevier, 2015.

Thore Egeland and Antonio Salas.

Estimating haplotype frequency and coverage of databases.
PloS one, 3(12):e3988, 2008.

Leonor Gusmão, Paula Sánchez-Diz, Cíntia Alves, Iva Gomes, María Teresa Zarrabeitia, Mariel Abovich, Ivannia Atmetlla, Cecilia Bobillo, Luisa Bravo, Juan Builes, et al.
A GEP-ISFG collaborative study on the optimization of an X-STR decaplex: data on 15 Iberian and Latin American populations.
International journal of legal medicine, 123(3):227–234, 2009.

Daniel Kling, Barbara Dell'Amico, and Andreas O Tillmar.

FamLinkX–Implementation of a general model for likelihood computations for X-chromosomal marker data.
Forensic Science International: Genetics, 17:1–7, 2015.

Daniel Kling, Andreas Tillmar, Thore Egeland, and Petter Mostad.

A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations.
International journal of legal medicine, pages 1–12, 2014.

Nádia Pinto, Leonor Gusmão, and António Amorim.

X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial.
Forensic Science International: Genetics, 5(1):27–32, 2011.

FamLinkX Exercises. http://familias.name/ghep2016

FamLinkX, developed by Daniel Kling. The participants should bring a laptop with FamLinkX, preferably version 2.5 (available from Aug 22 2016) downloaded (http://famlink.se/fx_download.html).

- Assistance for installation will be provided at the workshop, if needed. Note that installation is typically only possible if you are administrator on your laptop.

Time schedule

- 13.30 – 15.45 X-linked markers : Repetition, haplotype frequencies. FamLinkX demo video . More videos (Chapter 4). Thore Egeland
- 15.45 – 16.15 Coffee break
- 16.15 – 18.00 Exercises: The FamLinkX exercises 4.12-4.14, 4.16-4.18 are from the book "Relationship Inference with Familias and R" by Egeland, Kling, and Mostad available from Elsevier. The required material is freely available: exercises , solutions , zipped input files . Thore Egeland