PRELIMINARY VERSION, THE PRESENTATION WILL BE SUBJECT TO CHANGES

Linked markers (FamLink and FamLinkX)

Daniel Kling
Norwegian Institute of Public Health and Norwegian University of Life Sciences
daniel.kling@fhi.no

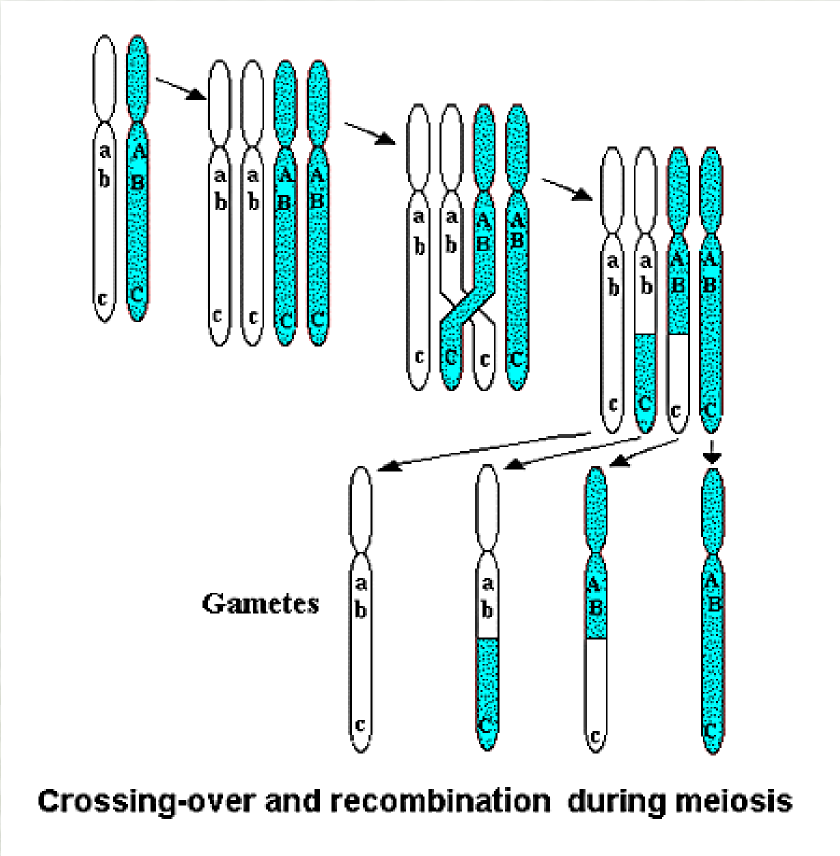
Copenhagen, May 20-23, 2014

Outline

- Linkage explained
- Linkage disequilibrium explained
- FamLinkX

Linkage

- What is linkage?

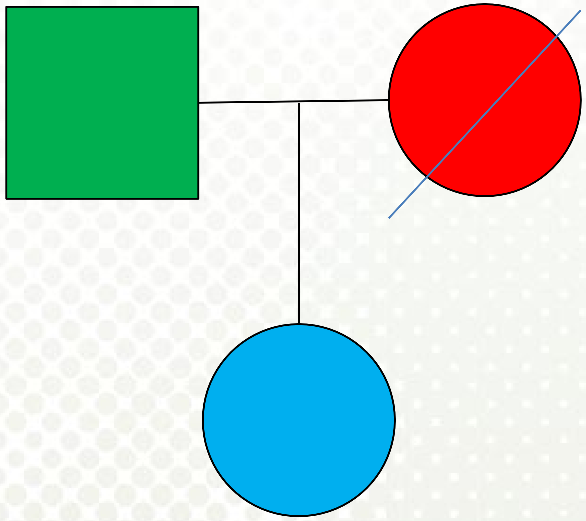


Crossing-over and recombination during meiosis

http://www.accessexcellence.org/RC/VL/GG/comeiosis.php

Linkage

- Worked example, paternity with two markers



a,b
c,d

a,e
c,f

$$LR = \frac{p(e,f)\left[p(a,c)p(b,d)(1-r)*0.5 + p(a,d)p(b,c)r*0.5\right]}{\left(p(a,c)p(e,f) + p(a,f)p(e,c)\right)\left(p(a,c)p(b,d) + p(a,d)p(b,c)\right)}$$

Linkage

- Worked example, paternity with two markers

$$LR=\frac{p(e,f)\left[p(a,c)p(b,d)(1-r)*0.5+p(a,d)p(b,c)r*0.5\right]}{\left(p(a,c)p(e,f)+p(a,f)p(e,c)\right)\left(p(a,c)p(b,d)+p(a,d)p(b,c)\right)}=$$

$$=\langle\text{Given LE}\rangle=\frac{0.5p(e)p(f)\left[p(a)p(c)p(b)p(d)(1-r)+p(a)p(d)p(b)p(c)r\right]}{\left(p(a,c)p(e,f)+p(a,f)p(e,c)\right)\left(p(a,c)p(b,d)+p(a,d)p(b,c)\right)}$$

$$=\frac{0.5p(e)p(f)p(a)p(c)p(b)p(d)\left[(1-r)+r\right]}{\left(p(a,c)p(e,f)+p(a,f)p(e,c)\right)\left(p(a,c)p(b,d)+p(a,d)p(b,c)\right)}=$$

$$=\frac{0.5p(e)p(f)p(a)p(c)p(b)p(d)\left[(1-r)+r\right]}{p(a)^2p(c)^2\left(2p(e,f)\right)\left(2p(b,d)\right)}=\frac{0.5}{4p(a)p(c)}$$

Linkage disequilibrium

- What is linkage disequilibrium
  - Allelic association
  - Two alleles (at two different markers) which is observed more often/less often than can be expected.
  - Effects the allele probabilities not the transmission probabilities.

Example

Marker1 (vWa): Alleles 13 and 14, frequencies 0.2 and 0.8

Marker2 (D12S391): Alleles 16 and 17, frequencies 0.4 and 0.6

*Expected frequency of [13, 16] is 0.2*0.4 = 0.08*

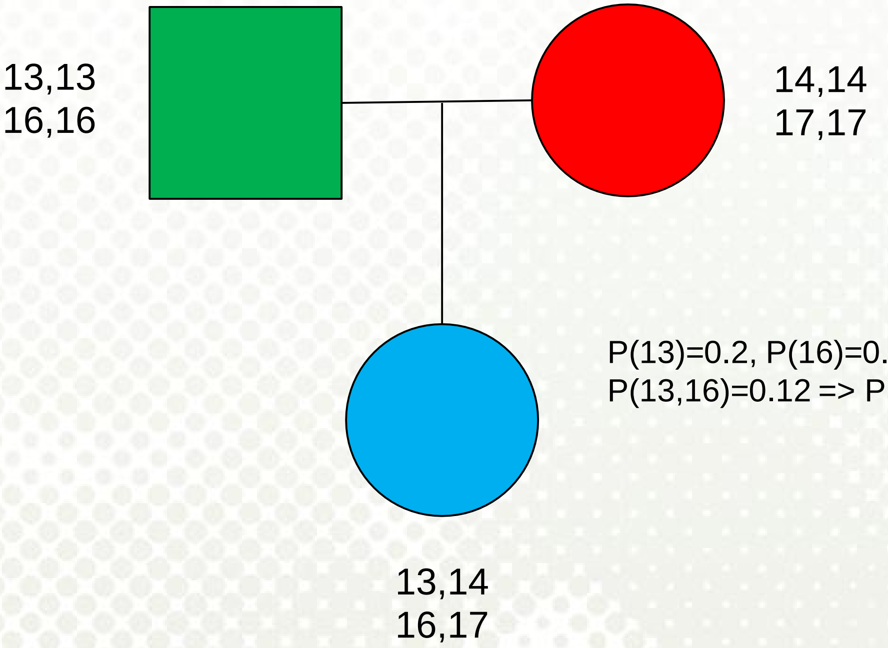
Observed frequency of [13, 16] is 0.12

$$r_{ij}^2 = \frac{\left[p_{ij} - p_i \times p_j\right]^2}{p_i(1-p_i)p_j(1-p_j)} \qquad \left\{\begin{array}{l} i = 13,14 \\ j = 16,17 \end{array}\right\}$$

$$p_{13,16} = 0.12,\ p_{13} = 0.2,\ p_{16} = 0.4 \Rightarrow r_{13,16}^2 = \frac{(0.12-0.08)^2}{0.2(1-0.2)0.4(1-0.4)} \simeq$$

Linkage disequilibrium

- Worked example, paternity with two markers



13,13
16,16

14,14
17,17

13,14
16,17

P(13)=0.2, P(16)=0.4 (Linkage equilibrium)
P(13,16)=0.12 => P(16|13)=0.6 (Linkage disequilibrium)

LR1 = 1/P(13)*1/P(16)=12.5 (Linkage equilibrium)

LR2 = 1/P(13)*1/P(16|13)=8.33 (Linkage disequilibrium)

Summary

Linkage	Linkage disequilibrium
Dependency between neighbouring markers	Dependency between alleles at different loci
Observed within a pedigree	Observed in a population
Extends long distances >10 cM	Usually extends short distances <1 cM
Do not affect random match probability (unless related)	Affect random match probabilities
Take into account for extended pedigrees	Always take into account for all pedigrees
Always take into account if also LD is present, for all pedigrees	Measured by the deviation from expectations, decays with recombinations
Measured by the recombination rate, constant	Used to find alleles associated with a disease, in the population
Used to find markers linked to a disease, in families	

X-chromosomal markers

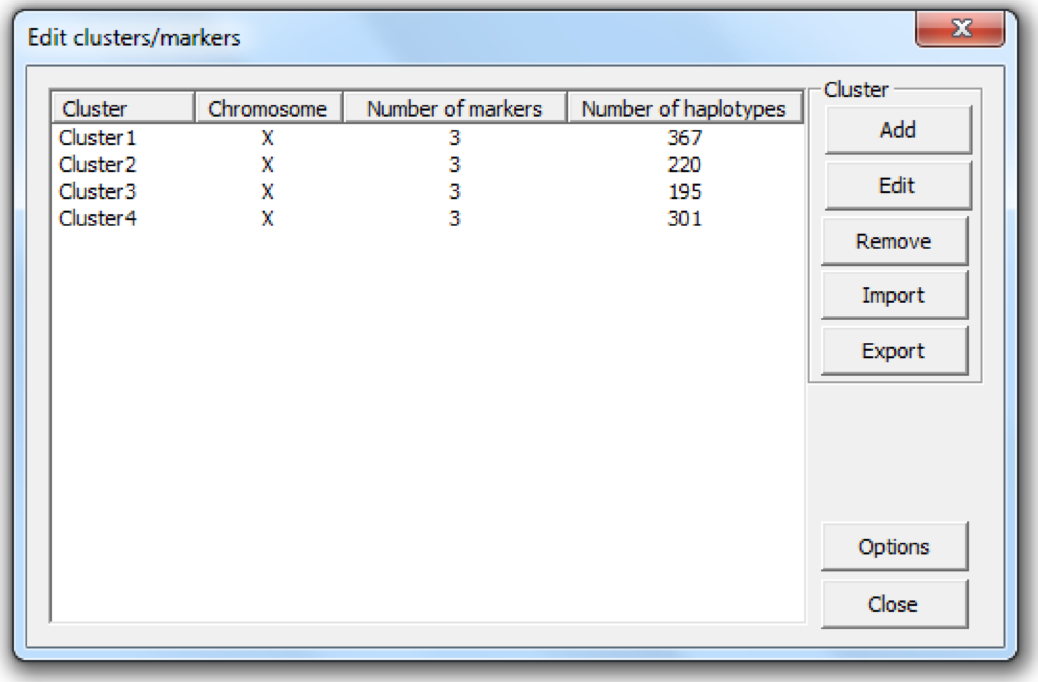
- Used where autosomal markers fail
- Argus X12
  - (4 clusters with three tightly linked markers)
- Linkage
- Linkage disequilibrium
- Mutations
- FamLinkX!
  - New joint probability model
  - Released autumn 2013

FamLinkX

- Markov chain to handle linkage
  - Similar to Lander-Green
- Multistep Markov chain to handle LD
- Uses a Dirichlet distribution to estimate haplotype frequencies

FamLinkX – At a glance

- Define clusters of markers



Cluster	Chromosome	Number of markers	Number of haplotypes
Cluster1	X	3	367
Cluster2	X	3	220
Cluster3	X	3	195
Cluster4	X	3	301

- Account for linkage between clusters
- Account for linkage and LD within each cluster

FamLinkX – Creating the database

- Size of the database?
  - Depend on the cluster size
- Include only males
  - Why?
- Input format for FamLinkX

ClusterID	Marker1	Marker2	…	HaploCounts
1	13	20		10
2	13	21		11
.				
.				
.				

FamLinkX – Creating the database

Example 1

Consider two SNP:s located in the same cluster/haploblock

S1: Alleles A, a [0.9 0.1]

S2: Alleles B, b [0.9 0.1]

How many possible haplotypes do we have? 4

What is the expected frequency of the most uncommon haplotype? 0.01

It may suffice to genotype 200 individuals

FamLinkX – Creating the database

Example 2

Consider two STR:s located in the same cluster/haploblock

S1: Alleles A1, A2....A20 [1/20 1/20 .... 1/20]

S2: Alleles B1, B2....B20 [1/20 1/20 .... 1/20]

How many possible haplotypes do we have? 20*20=400

What is the expected frequency of the most uncommon haplotype? 1/400=0.0025

Some haplotypes may be even more uncommon due to LD.

We may want to genotype >400 individuals

FamLinkX – Creating the database

Example 3

Consider three STR:s located in the same cluster/haploblock

S1: Alleles A1, A2....A20 [1/20 1/20 .... 1/20]

S2: Alleles B1, B2....B20 [1/20 1/20 .... 1/20]

S3: Alleles C1, C2....C20 [1/20 1/20 .... 1/20]

How many possible haplotypes do we have? 20*20*20=8000

What is the expected frequency of the most uncommon haplotype? 1/8000

Some haplotypes may be even more uncommon due to LD.

As many individuals as possible

FamLinkX – Creating the database

– Estimation of updated haplotype frequencies
– The model

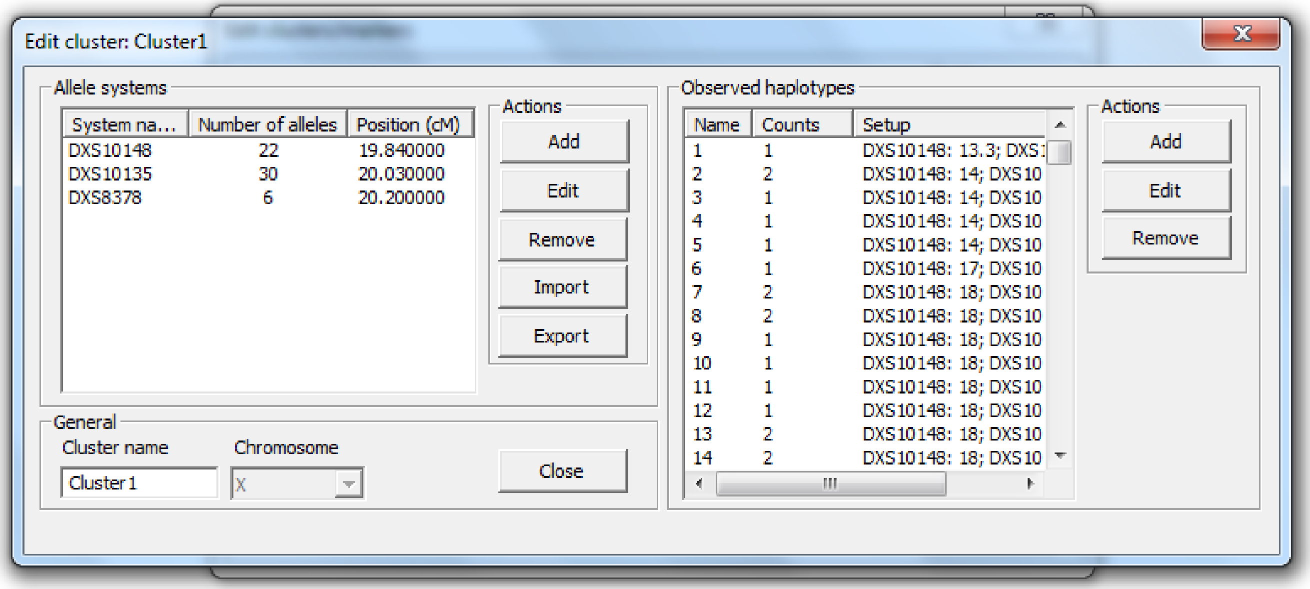
$$H_i = \frac{c_i + \lambda p_i}{C + \lambda}$$

Hi = Updated haplotype frequency
ci = Counts for haplotype *i*
C = Total number of haplotypes
pi = Expected haplotype frequency
λ = Prior weight given to expected frequency

– If λ=0, only observed haplotypes have a nonzero frequency.
– If λ=large, all haplotypes have a frequency.

FamLinkX – At a glance

- Add markers
  - Define genetic position
  - Mutation parameters



System na...	Number of alleles	Position (cM)
DXS10148	22	19.840000
DXS10135	30	20.030000
DXS8378	6	20.200000

Cluster name: Cluster1

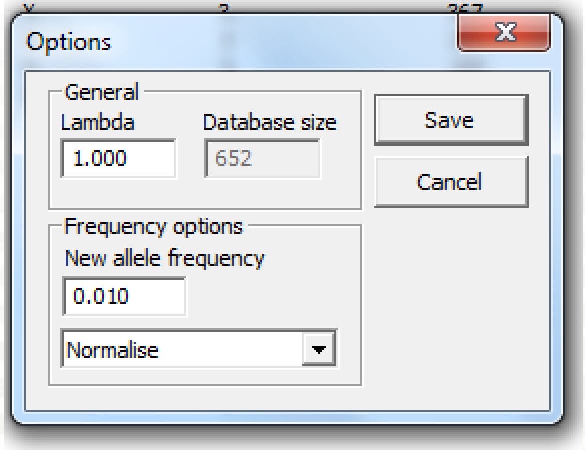
Chromosome: X

Name	Counts	Setup
1	1	DXS10148: 13.3; DXS
2	2	DXS10148: 14; DXS10
3	1	DXS10148: 14; DXS10
4	1	DXS10148: 14; DXS10
5	1	DXS10148: 14; DXS10
6	1	DXS10148: 17; DXS10
7	2	DXS10148: 18; DXS10
8	2	DXS10148: 18; DXS10
9	1	DXS10148: 18; DXS10
10	1	DXS10148: 18; DXS10
11	1	DXS10148: 18; DXS10
12	1	DXS10148: 18; DXS10
13	2	DXS10148: 18; DXS10
14	2	DXS10148: 18; DXS10

- Add haplotypes(!)
  - Define setup
  - Counts

FamLink – At a glance

- Selecting value for Lambda



Options

General

Lambda: 1.000

Database size: 652

Frequency options

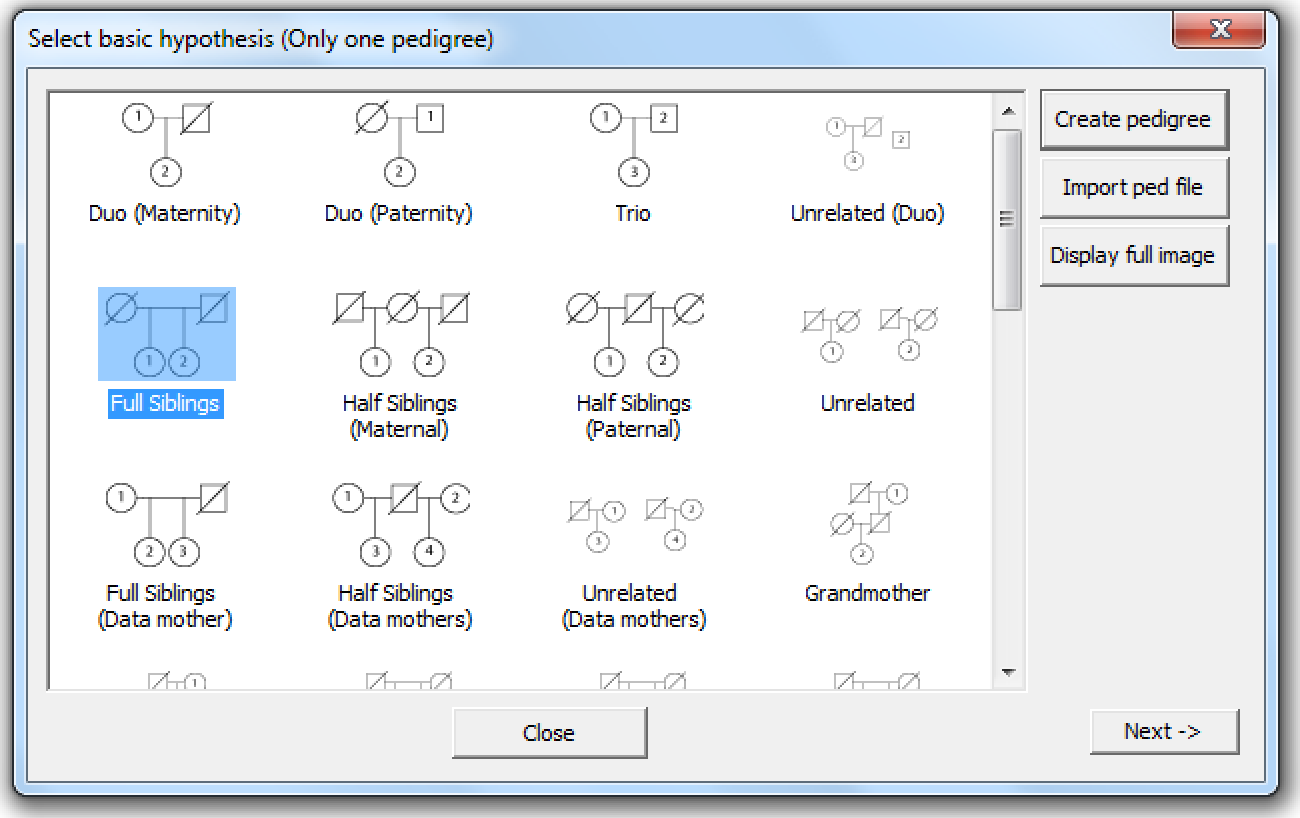
New allele frequency: 0.010

Normalise

- We display two methods

FamLinkX – At a glance

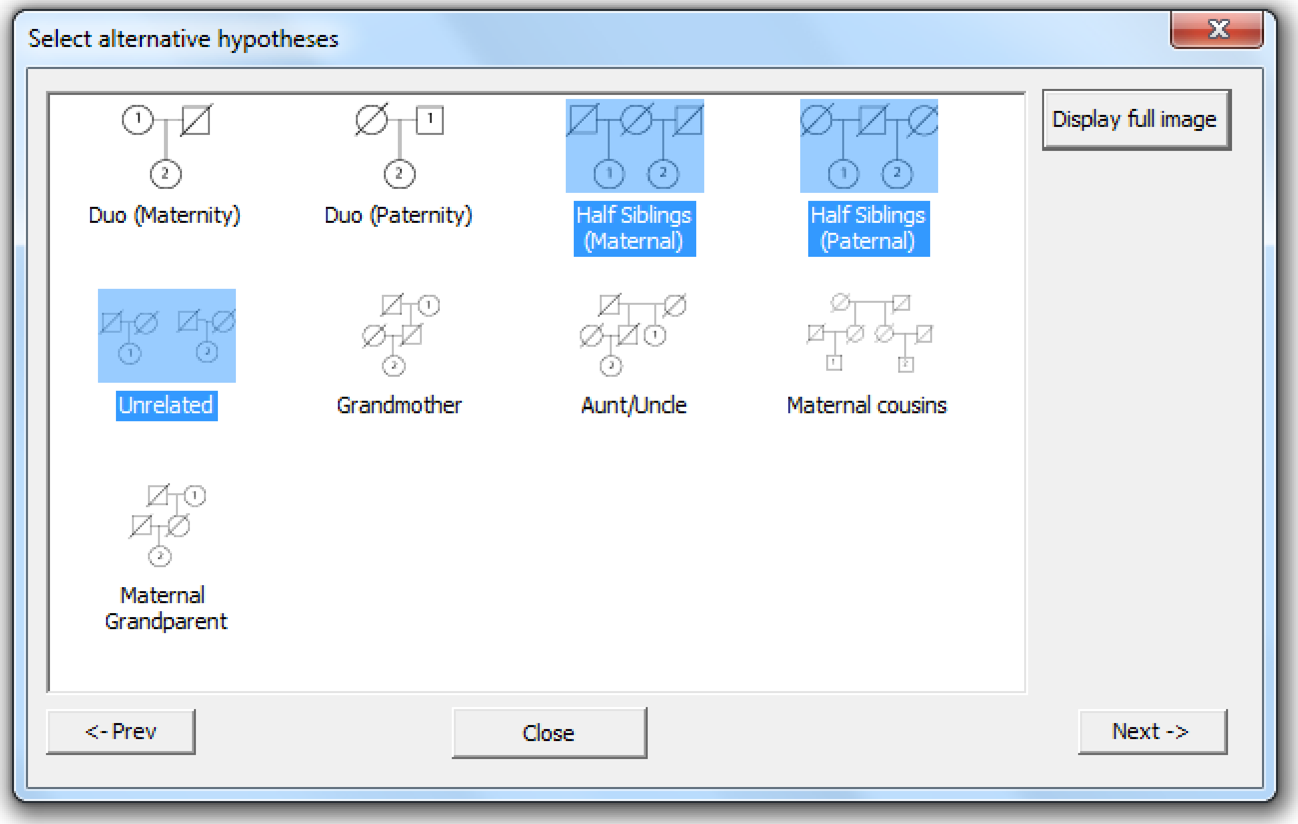
– Select main hypothesis



Select basic hypothesis (Only one pedigree)

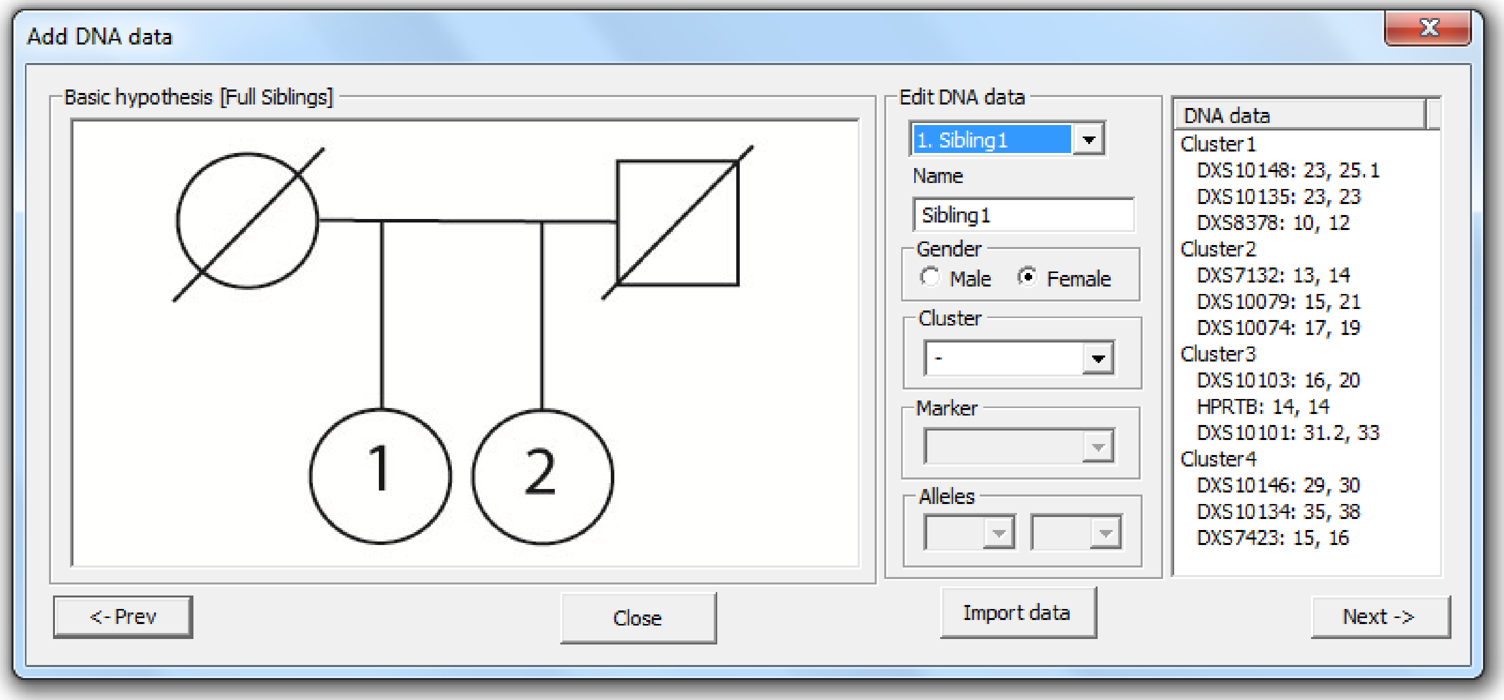
FamLinkX – At a glance

- Select alternative hypotheses



FamLink – At a glance

- Define DNA data



Add DNA data

Basic hypothesis [Full Siblings]

Edit DNA data

1. Sibling1

Name: Sibling1

Gender: ○ Male ● Female

Cluster: -

Marker

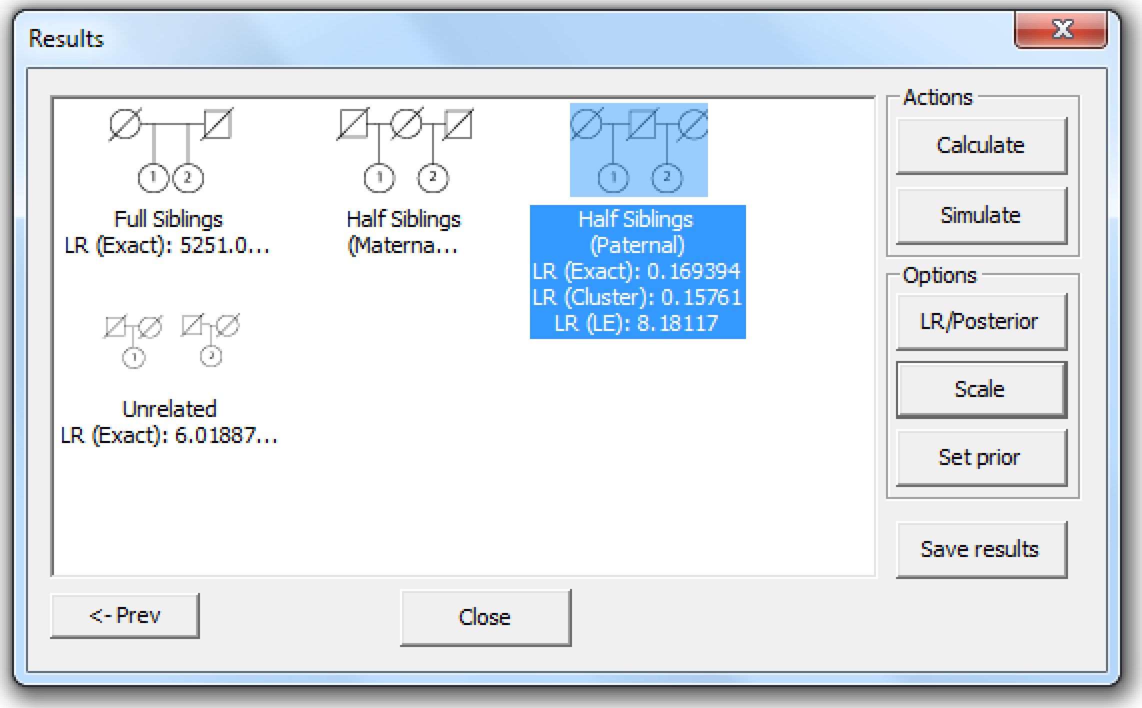
Alleles

DNA data

Cluster1
  DXS10148: 23, 25.1
  DXS10135: 23, 23
  DXS8378: 10, 12
Cluster2
  DXS7132: 13, 14
  DXS10079: 15, 21
  DXS10074: 17, 19
Cluster3
  DXS10103: 16, 20
  HPRTB: 14, 14
  DXS10101: 31.2, 33
Cluster4
  DXS10146: 29, 30
  DXS10134: 35, 38
  DXS7423: 15, 16

FamLinkX – At a glance

- Calculate likelihoods



Results

Full Siblings
LR (Exact): 5251.0...

Half Siblings
(Materna...

Half Siblings
(Paternal)
LR (Exact): 0.169394
LR (Cluster): 0.15761
LR (LE): 8.18117

Unrelated
LR (Exact): 6.01887...

- We display three computation methods

FamLinkX – Keep in mind

– Scaling against a pedigree where mutations are present(!)
– Depends on the value of lambda
    - Try different values, e.g. 1, Dbsize and 1000000
– Three markers or more in a cluster will require long computation times
– Speedup is being developed

FamLinkX – Questions?

EXERCISES