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
Repetition

- Inheritance
- Linkage. Recombination.
- LR
  - GHEP-ISFG Decaplex [3], ARGUS X-12 [4],....

Estimating haplotype frequencies: $\lambda$ model.

FamLinkX [5]. Demo . Exercises
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**

- 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
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?

\[ LR_1 = \frac{p_1 p_2^2}{\frac{1}{2} 2p_1 p_2^3 + \frac{1}{2} p_1 p_2^2} = \frac{1}{p_2 + \frac{1}{2}} > 1 \text{ if } p_2 < \frac{1}{2}. \]

\[ LR_1 \overset{p_2=0.2}{=} 1.42857 \]
FamLinkX: http://familias.name/ExamplePaternalOrMaternal.sav
Two markers: Multiply?: \( LR = ? \ LR_1 \times LR_2 \)

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

- We can only multiply $LR$-s when markers are independent, i.e., $LR \neq LR_1 \times LR_2$.
- 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

Haplotypes known:
- Bb and Aa

Haplotypes unknown

Recombination: depends on recombination probability r
Maternal grandson or unrelated? Mother missing.

\[
LR = \frac{1}{2} \frac{(1 - r)}{p_{12,17}} \quad LE = \frac{1}{2} \frac{(1 - r)}{p_{12}p_{17}} \\
\quad \quad \quad \quad \quad \quad r = 0.01 \quad \frac{1}{2} \frac{(1 - 0.01)}{0.6 \times 0.4} = 2.0625
\]

\[
LR \quad r = 0.5 \quad \frac{1}{2p_{12}} \frac{1}{2p_{17}} = 1.04
\]
Haldane’s map function:

\[
r = \frac{1 - \exp\left(-2 \times \frac{x}{100}\right)}{2}
\]

at 50 cM:

\[
x = 50 cM \quad 1 - \exp\left(-2 \times \frac{50}{100}\right) = 0.316
\]
**LE: multiply**

<table>
<thead>
<tr>
<th>loc1</th>
<th>loc2</th>
<th>freq1</th>
<th>freq2</th>
<th>( P(hap \mid LE) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2*0.3=0.06</td>
</tr>
<tr>
<td>A</td>
<td>b</td>
<td>0.2</td>
<td>0.7</td>
<td>0.2*0.7=0.14</td>
</tr>
<tr>
<td>a</td>
<td>B</td>
<td>0.8</td>
<td>0.3</td>
<td>0.8*0.3=0.24</td>
</tr>
<tr>
<td>a</td>
<td>b</td>
<td>0.8</td>
<td>0.7</td>
<td>0.8*0.7=0.56</td>
</tr>
</tbody>
</table>

| 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.
Estimating haplotype frequencies

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

- \( F_i \): Updated haplotype frequency,
- \( c_i \): Count of haplotype \( i \),
- \( C \): Total number of haplotypes,
- \( p_i \): Expected haplotype frequency,
- \( \lambda > 0 \): prior weight.
Repetition: Inheritance, linkage and LD

Haplotype frequencies

FamLinkX

### Example I

\[ F_i = \frac{c_i + \lambda p_i}{C + \lambda} \]

<table>
<thead>
<tr>
<th>haplo</th>
<th>( P(hap \mid LE) )</th>
<th>Count</th>
<th>( \lambda = 100 ) method</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>( p_1 = 0.06 )</td>
<td>( c_1 = 10 )</td>
<td>( \frac{10+100 \times 0.06}{100+100} = 0.080 )</td>
</tr>
<tr>
<td>Ab</td>
<td>( p_2 = 0.14 )</td>
<td>( c_2 = 15 )</td>
<td>( \frac{15+100 \times 0.14}{100+100} = 0.145 )</td>
</tr>
<tr>
<td>aB</td>
<td>( p_3 = 0.24 )</td>
<td>( c_3 = 25 )</td>
<td>( \frac{25+100 \times 0.24}{100+100} = 0.245 )</td>
</tr>
<tr>
<td>ab</td>
<td>( p_4 = 0.56 )</td>
<td>( c_4 = 50 )</td>
<td>( \frac{50+100 \times 0.56}{100+100} = 0.530 )</td>
</tr>
<tr>
<td>tot</td>
<td></td>
<td>1.00</td>
<td>1.000</td>
</tr>
</tbody>
</table>

\( \text{tot} \) = 1.00, \( C = 100 \)
Example I contd: LR

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

\[ LR = \frac{1}{\text{freq haplotype } AB} = \frac{C + \lambda}{c_i + \lambda p_i} = \frac{100 + \lambda}{10 + \lambda \times 0.06}. \]
Example I contd: \( LR = \frac{100 + \lambda}{10 + \lambda * 0.06} \)
## Example II

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

**Table 3: $\lambda$ effect illustrated**

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

Example. Data Exercise 4.12

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

\[
\frac{59 + 100000 \times 0.6 \times 0.6}{100 + 100000} = 0.36
\]
\(\lambda\): Practical suggestion.

- \(\lambda\) estimates: Egeland, Kling, Mostad [1].
- FamLinkX 2.5 implementation: Generates R code. See video Exercise 4.14: [http://familias.name/VideosBook.pdf](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

\[ H_{12,17} = \text{estimated haplotype frequency} \]

\[ LR_{\text{COUNT}} = \frac{1}{H_{12,17}} = \frac{1}{1/100} = 100, \quad LR_{\text{LE}} = \frac{1}{p_{12}p_{17}} = \frac{1}{0.6 \times 0.4} = 4.17. \]
FamLinkX. Screenshots Exercise 4.12.
Screenshots Exercise 4.12. Frequency database

![Screenshot of FamLinkX](attachment:image.png)
Editing clusters/markers I

![Edit clusters/markers window](Image)

- Database name: Unspecified
- Cluster options: Add, Edit, Remove, Import, Export
- Lambda arrow pointing to the cluster area

Outline
Repetition: Inheritance, linkage and LD
Haplotype frequencies
FamLinkX
Editing clusters/markers II
Editing clusters/markers III

Exer 4.21*
Editing clusters/markers IV
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)
- Grandmother
- Grandmother (Data mother)
- Aunt/Uncle
- Aunt/Uncle (Data mother)
- Aunt/Uncle (Maternal)
- Two Aunts/Undes
- Two Aunts/Undes (Data mother)
- Three Full Siblings
Select alternative hypotheses

- Full Siblings
- Half Siblings (Maternal)
- Half Siblings (Paternal)
- Unrelated
- Grandmother
- Aunt/Uncle

Create/Edit pedigree
Display full image
Results

Duo (Paternity)
LR (Exact): 99.9977
LR (Cluster): 99.998
LR (LE): 4.16667

Unrelated Scale
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.

Thore Egeland and Antonio Salas.
Estimating haplotype frequency and coverage of databases.

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.

Daniel Kling, Barbara Dell'Amico, and Andreas O Tillmar.
FamLinkX–Implementation of a general model for likelihood computations for X-chromosomal marker data.

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.

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.

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