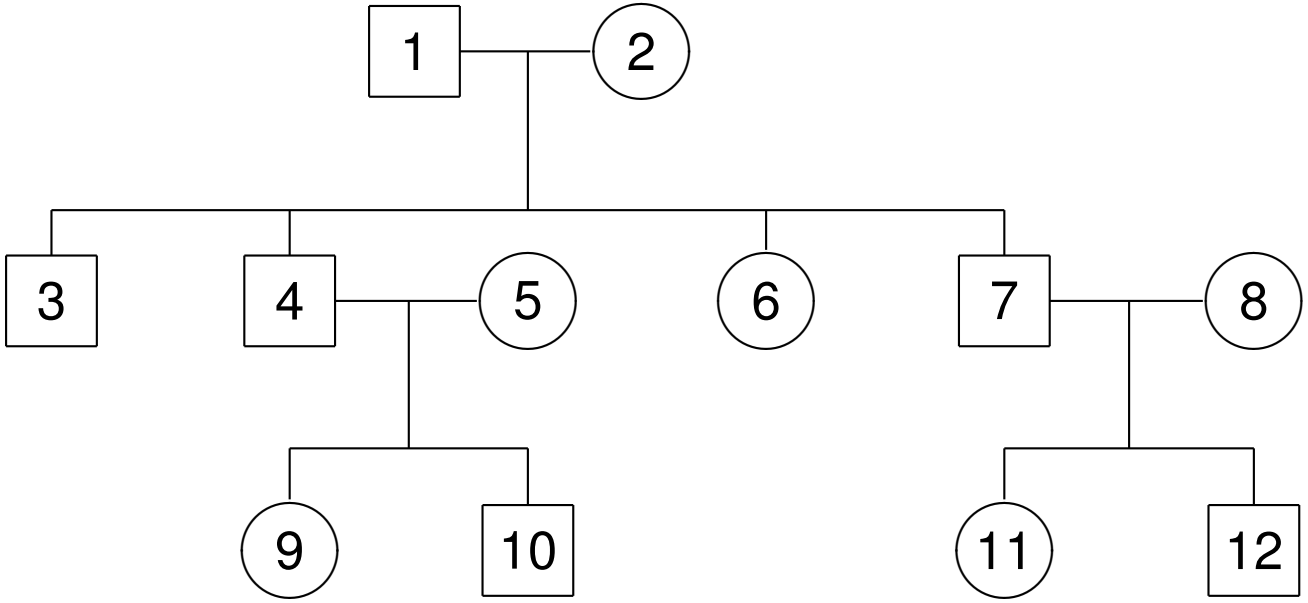
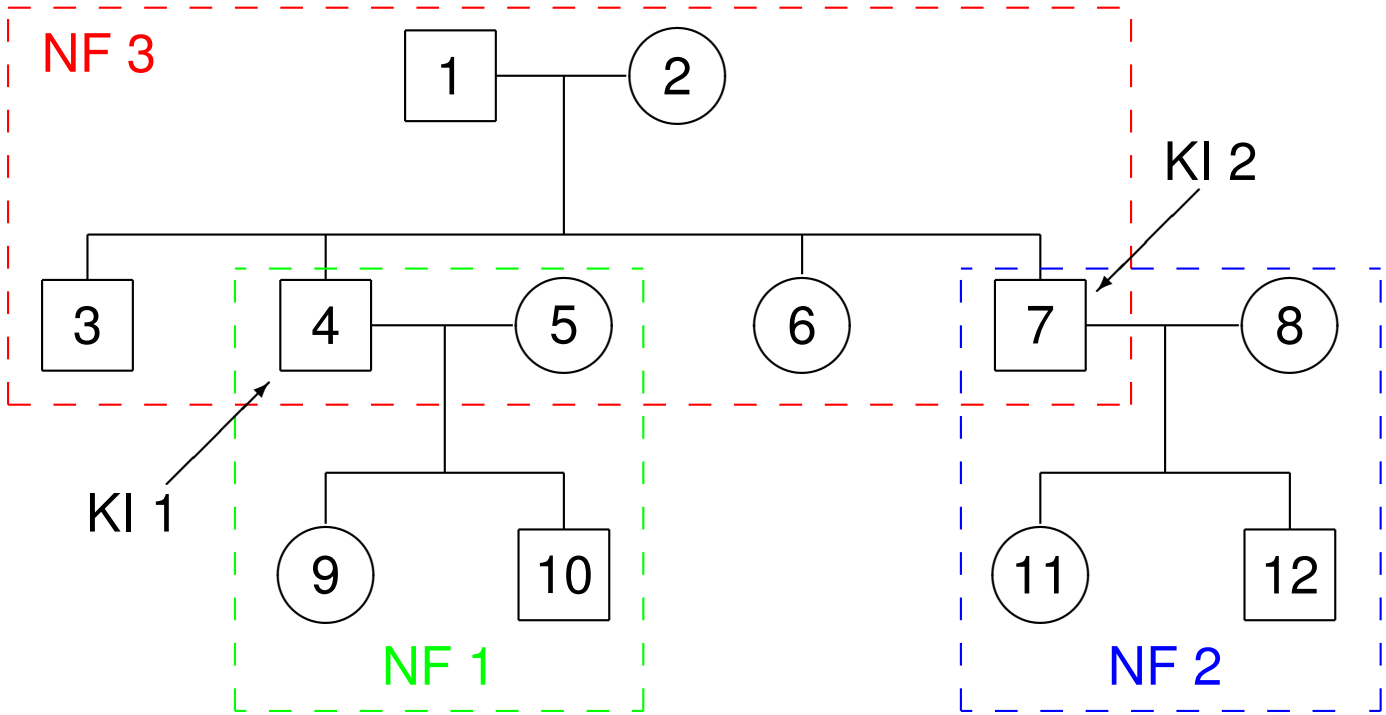
Calculation of the pedigree likelihood: Algorithms



Elston-Stewart-Algorithm

Idea:

- Break down the pedigree into nuclear families (NF)
- Peel the result from each nuclear-family calculation onto the key individual (KI) linking that nuclear family to the rest of the pedigree



Elston-Stewart-Algorithm

$$
\begin{aligned}
L(\theta \mid y) &= \sum_{g_1} \cdots \sum_{g_{12}} \left[\prod_{j=1}^{12} P(y_j \mid g_j) \right] \cdot \left[\prod_{i \in \{1,2,5,8\}} P(g_j) \right] \cdot \left[\prod_{i \notin \{1,2,5,8\}} P_\theta(g_j \mid g_{F_j}, g_{M_j}) \right] \\
&= \sum_{g_1} \sum_{g_2} \left[\prod_{j=1}^{2} P(y_j \mid g_j) \right] \cdot \left[\prod_{j=1}^{2} P(g_j) \right] \\
&\quad \cdot \left[\sum_{g_3} P(y_3 \mid g_3) \cdot P_\theta(g_3 \mid g_1, g_2) \right] \cdot \left[\sum_{g_6} P(y_6 \mid g_6) \cdot P_\theta(g_6 \mid g_1, g_2) \right] \\
&\quad \cdot \left[\sum_{g_4} P(y_4 \mid g_4) \cdot P_\theta(g_4 \mid g_1, g_2) \right. \\
&\qquad \left. \cdot \sum_{g_5} \sum_{g_9} \sum_{g_{10}} \left[\prod_{j \in \{5,9,10\}} P(y_j \mid g_j) \right] \cdot P(g_5) \cdot \left[\prod_{j \in \{9,10\}} P_\theta(g_j \mid g_4, g_5) \right] \right] \\
&\quad \cdot \left[\sum_{g_7} P(y_7 \mid g_7) \cdot P_\theta(g_7 \mid g_1, g_2) \right. \\
&\qquad \left. \cdot \sum_{g_8} \sum_{g_{11}} \sum_{g_{12}} \left[\prod_{j \in \{8,11,12\}} P(y_j \mid g_j) \right] \cdot P(g_8) \cdot \left[\prod_{j \in \{11,12\}} P_\theta(g_j \mid g_7, g_8) \right] \right]
\end{aligned}
$$

Complexity of Elston-Stewart-Algorithm

Problem:

Multi-point linkage analysis, i.e., l marker loci $M_1,\ldots,M_l$ with recombination fraction $\theta_{i,i+1}$ $(1 \leq i < l)$ between locus M_i and locus M_{i+1} ($\rightarrow$ map function)

$l = 6$ and five alleles at each marker locus, one diallelic disease locus

$\Rightarrow 2 \cdot 5^6 = 31,250$ haplotypes and

$2 \cdot 5^6 \cdot (2 \cdot 5^6 + 1)/2 = 488,296,875$ genotypes

Running time and memory requirements of Elston-Stewart-algorithm are

- linear in the number of individuals
- exponential in the number of loci

Limitations of algorithms for the calculation of pedigree likelihoods

Algorithm	**Program**[1]	**Likelihood**	**Limitations**
Elston-Stewart	Linkage FastLink Vitesse	exact	$<$ 6 marker loci
Lander-Green	Genehunter Allegro Merlin	exact	$<$ 20 individuals
Monte-Carlo	Loki SimWalk	approximate	$\sim$ 1000 individuals, $\sim$ 1000 marker loci

[1] "List of Genetic Analysis Software" at

`https://github.com/gaow/genetic-analysis-software`

Distribution of $Z(\hat{\theta})$ under the null hypothesis

Suppose that two-point linkage analysis results in a maximum lod score of $Z(\hat{\theta}) = 4.2$. What are the consequences of such a result? Does $Z(\hat{\theta}) = 4.2$ provide sufficient evidence for linkage between between marker and disease locus, i.e., can the null hypothesis $H_0 : \theta = 0.5$ be rejected?

To answer this question, it is required to obtain the distribution of

$$Z(\hat{\theta}) = \log_{10} \frac{\max_{\theta \in [0,1/2]} L(\theta \mid y)}{L(\theta = 1/2 \mid y)}$$

under the null hypothesis $H_0 : \theta = 0.5$.

Distribution of $Z(\hat{\theta})$ under the null hypothesis

It can be shown that the distribution of $2 \cdot \ln(10) \cdot Z(\hat{\theta})$ is approximated by a $1/2 : 1/2$ mixture of a point mass at zero and a χ^2_1-distribution (c.f. P/16).

Therefore,

$$\begin{aligned} P_{\theta=1/2}(Z(\hat{\theta}) \geq 4.2) &= P_{\theta=1/2}(2 \cdot \ln(10) \cdot Z(\hat{\theta}) \geq 2 \cdot \ln(10) \cdot 4.2) \\ &\approx \frac{1}{2} \cdot P(\chi^2_1 \geq 19.34) = 5 \cdot 10^{-6} \end{aligned}$$

Exercise:

Historically, human geneticists declared linkage in case that $Z(\hat{\theta}) \geq 3$. What is the P-value corresponding to $Z(\hat{\theta}) = 3$?

(Hint: $P(\chi^2_1 \geq 13.8155) = 0.0002$)

Effect of misspecification of the genetic model

Calculation of lod scores requires knowledge/specification of (c.f. PLI/3)

1. disease model (frequency of alleles at the disease locus, penetrances)
2. parameters related to the marker locus (marker allele frequencies)

Disease model specification

- may be correct for Mendelian diseases
- is almost surely incorrect for genetically complex diseases

In principle, model misspecification can

- increase the type I error rate (i.e., large values of $Z(\widehat{\theta})$ occur more frequently than predicted by the null distribution of $Z(\widehat{\theta})$ under the "correct" model)
- increase the type II error rate (decrease the power to detect linkage)

Effect of misspecification of the genetic model

In the following, it is assumed that the families in the sample were selected on the basis of the disease phenotypes.

Then, misspecification of the disease model

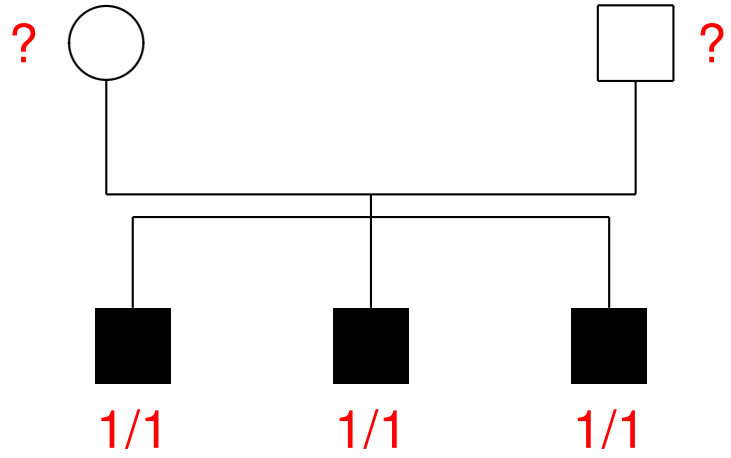
- will not increase the type I error rate
- but can reduce the power to detect linkage (and result in estimates of θ which are too large in tendency)

Misspecification of the parameters related to the marker locus

- can increase the type I error rate
- reduce the power to detect linkage

Effect of misspecification of the genetic model

Example: Increased type I error rate due to misspecification of marker allele frequencies



Disease model: $f_2 = 1, f_1 = f_0 = 0$ and

- $P(\text{allele } 1) = .99$: $Z(\widehat{\theta}) = 0.0065$
- $P(\text{allele } 1) = .01$: $Z(\widehat{\theta}) = 1.1784$

Effect of misspecification of the genetic model

How to guard against an increased type I error rate due to misspecification of marker allele frequencies?

1. Avoid samples with unavailable/untyped individuals.

   (often impracticable)
2. Estimate marker allele frequencies from the data.
3. Repeat the analysis for different assumed marker allele frequencies to assess the sensitivity of results in regard to marker allele frequencies.

Effect of misspecification of the genetic model

How to guard against decreased power due to misspecification of the disease model?

1. Repeat the analysis for a few disease models. Select the model which results in the largest maximum lod score.
2. Repeat the analysis for **all** disease models ($\rightarrow$ MOD score analysis).
3. Apply methods of nonparametric linkage analysis.