Parametric Linkage Analysis

Two-point linkage analysis:

Analysis of linkage between a disease gene and a marker gene

Goals:

1. Estimation of the recombination fraction θ between marker and disease locus
2. Testing the null hypothesis $H_0 : \theta = 1/2$ (i.e., marker and disease locus are unlinked) against the alternative hypothesis $H_1 : \theta < 1/2$ (i.e., marker and disease locus are linked)

Parametric Linkage Analysis

Method:

- Sample of families with the disease (i.e., one or more affected individuals in each family)
- Calculation of the likelihood $L_i(\theta \mid y_i)$ of pedigree i
- Likelihood of the whole sample: $\prod_{i=1}^{n} L_i(\theta \mid y_i)$
- Estimation of θ:

$$\hat{\theta} = \arg \max_{\theta \in [0,1/2]} \prod_{i=1}^{n} L_i(\theta \mid y_i) = \arg \max_{\theta \in [0,1/2]} \sum_{i=1}^{n} \ln L_i(\theta \mid y_i)$$

- Testing of H_0:

$$Z(\theta) = \log_{10} \frac{\prod_{i=1}^{n} L_i(\theta \mid y_i)}{\prod_{i=1}^{n} L_i(\theta = 1/2 \mid y_i)} = \sum_{i=1}^{n} \log_{10} \frac{L_i(\theta \mid y_i)}{L_i(\theta = 1/2 \mid y_i)}$$

($Z(\theta)$: *lod score*, $Z(\hat{\theta})$: *maximum lod score*)

Calculation of the pedigree likelihood

Requires knowledge of

1. disease model
   - number of alleles at the disease locus (usually: two),

     frequencies p_i of the alleles at the disease locus
   - penetrance parameters f_2, f_1, f_0

     (f_i is the conditional probability that an individual is affected given

     his/her genotype at the disease locus contains i disease alleles)
2. parameters related to the marker locus
   - number and frequencies q_j of alleles at the marker locus
   - relationship between marker genotype and marker phenotype

Calculation of the pedigree likelihood

$y = (y_1, \ldots, y_I)$:

y_j describes observed marker and disease phenotypes of individual j

$L(\theta \mid y)$:

probability of observing y, given θ and the pedigree structure (and assuming all model parameters f_i, p_i, q_i to be known)

$\Rightarrow$

$$L(\theta \mid y) = \sum_{g \in \mathcal{G}} P_\theta(y, g) = \sum_{g \in \mathcal{G}} P_\theta(y \mid g) \cdot P_\theta(g) \qquad (1)$$

with $\mathcal{G}$ denoting the set of all joint marker-disease genotypes (including phase)

Calculation of the pedigree likelihood

Founders are pedigree members without parents in the pedigree; $\mathcal{F}$ denotes the set of founders in the pedigree.

Non-founders are pedigree members with parents in the pedigree. If individual j is a non-founder, then let F_j and M_j denote the father and mother of individual j.

Assumption 1: Genotypes of founders are assumed to be independent

$$\Rightarrow \qquad P_\theta(g) = \prod_{j \in \mathcal{F}} P_\theta(g_j) \cdot \prod_{j \notin \mathcal{F}} P_\theta(g_j \mid g_{F_j}, g_{M_j}) \tag{2}$$

Assumption 2:

a) $y_1, \ldots, y_I$ are independent conditional on $g_1, \ldots, g_I$

b) y_j only depend on g_j, i.e., $P_\theta(y_j \mid g_1, \ldots, g_I) = P_\theta(y_j \mid g_j)$

$$\Rightarrow \qquad P_\theta(y \mid g) = \prod_{j=1}^{I} P_\theta(y_j \mid g_j) \tag{3}$$

Calculation of the pedigree likelihood

$\Rightarrow$

$$L(\theta \mid y) = \sum_{g \in \mathcal{G}} \left[\prod_{j=1}^{I} P(y_j \mid g_j) \right] \cdot \left[\prod_{j \in \mathcal{F}} P(g_j) \right] \cdot \left[\prod_{j \notin \mathcal{F}} P_\theta(g_j \mid g_{F_j}, g_{M_j}) \right]$$

(obtained from (1) by inserting (2) and (3) and by noting that $P(y_j \mid g_j)$ and, for $j \in \mathcal{F}$, $P(g_j)$ does not depend on θ)

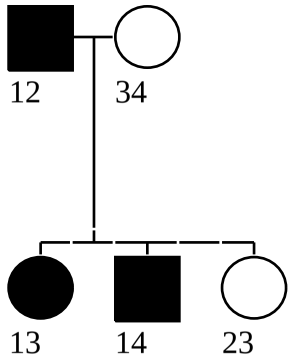
Calculation of the pedigree likelihood: genotype elimination

Let $\mathcal{G}^\star := \{g \in \mathcal{G} : \prod_{j=1}^{I} P(y_j \mid g_j) > 0\}$ denote the set of genotypes being compatible with the observed phenotype y

$\Rightarrow$

$$L(\theta \mid y) = \sum_{g \in \mathcal{G}^\star} \left[\prod_{j=1}^{I} P(y_j \mid g_j) \right] \cdot \left[\prod_{j \in \mathcal{F}} P(g_j) \right] \cdot \left[\prod_{j \notin \mathcal{F}} P_\theta(g_j \mid g_{F_j}, g_{M_j}) \right]$$

Calculation of the pedigree likelihood: example



□ : male, unaffected

○ : female, unaffected

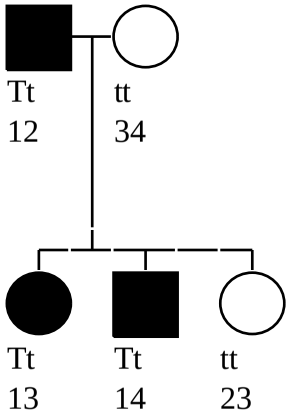
■ : male, affected

● : female, affected

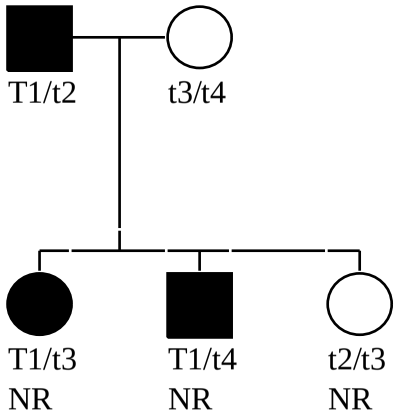
Assumptions:

- $f_2 = f_1 = 1$, $f_0 = 0$, i.e., the disease is dominantly inherited (because of $f_2 = f_1$), fully penetrant (because of $f_2 = f_1 = 1$), and without phenocopies (because of $f_0 = 0$)
- two alleles T and t with frequencies p_T and p_t at the disease locus
- four alleles 1, 2, 3, and 4 with frequencies q_i at the marker locus

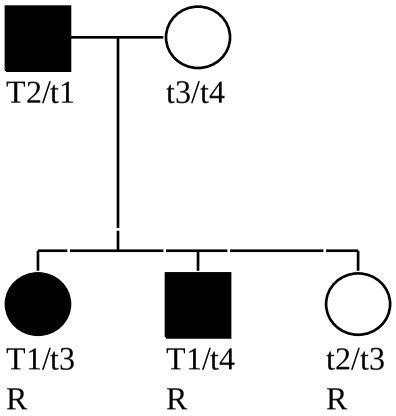
Calculation of the pedigree likelihood: example



Phase in the father?



$\tilde{g}$



$\tilde{\tilde{g}}$

$$\mathcal{G}^\star = \{\tilde{g}, \tilde{\tilde{g}}\}$$

Calculation of the pedigree likelihood: example

- $P(g_j)$ for $j \in \mathcal{F}$:

  Assumption 3: Marker and disease locus haplotypes are in Hardy-Weinberg equilibrium, i.e., with $g_j = (m_{j1}d_{j1}, m_{j2}d_{j2})$ denoting the two marker/disease haplotypes of individual j, it follows that

$$P(g_j) = \begin{cases} P^2(m_{j1}d_{j1}) & \text{if} \quad m_{j1}d_{j1} = m_{j2}d_{j2} \\ 2 \cdot P(m_{j1}d_{j1}) \cdot P(m_{j2}d_{j2}) & \text{if} \quad m_{j1}d_{j1} \neq m_{j2}d_{j2} \end{cases}$$

  Assumption 4: There exists linkage equilibrium between alleles at the marker and the disease locus, i.e.,

$$P(m_{j1}d_{j1}) = P(m_{j1}) \cdot P(d_{j1}) = q_{m_{j1}} \cdot p_{d_{j1}}$$

Application to the pedigree of the example:

$\tilde{g}$ and $\tilde{\tilde{g}}$: $\prod_{j \in \mathcal{F}} P(g_j) = 4 \cdot p_T \cdot p_t^3 \cdot q_1 \cdot q_2 \cdot q_3 \cdot q_4$

Calculation of the pedigree likelihood: example

- $P_\theta(g_j \mid g_{F_j}, g_{M_j})$ for $j \notin \mathcal{F}$:

  $\tilde{g}$: $P_\theta(g_j \mid g_1, g_2) = 0.25 \cdot (1-\theta)$ for $j = 3,4,5$

  $\tilde{\tilde{g}}$: $P_\theta(g_j \mid g_1, g_2) = 0.25 \cdot \theta$ for $j = 3,4,5$

- $P(y_j \mid g_j) = 1$ for $j = 1, \ldots, 5$

$\Rightarrow$

$$
\begin{aligned}
L(\theta \mid y) &= 4 \cdot p_T \cdot p_t^3 \cdot q_1 \cdot q_2 \cdot q_3 \cdot q_4 \cdot \left[\frac{1}{4} \cdot (1-\theta)\right]^3 \\
&\quad + 4 \cdot p_T \cdot p_t^3 \cdot q_1 \cdot q_2 \cdot q_3 \cdot q_4 \cdot \left[\frac{1}{4} \cdot \theta\right]^3 \\
&= \frac{1}{16} \cdot p_T \cdot p_t^3 \cdot q_1 \cdot q_2 \cdot q_3 \cdot q_4 \cdot \left[(1-\theta)^3 + \theta^3\right] \\
\widehat{\theta} &= 0 \\
Z(\theta) &= \log_{10}\left[4 \cdot (1-\theta)^3 + 4 \cdot \theta^3\right] \\
Z(\widehat{\theta}) &= \log_{10} 4 \sim .6021
\end{aligned}
$$

Calculation of the pedigree likelihood: example

Recall the assumptions for the calculation of the lod score $Z(\theta)$:

- dominant mode of inheritance
- disease allele frequencies p_T and p_t
- marker allele frequencies $q_1, \ldots, q_4$

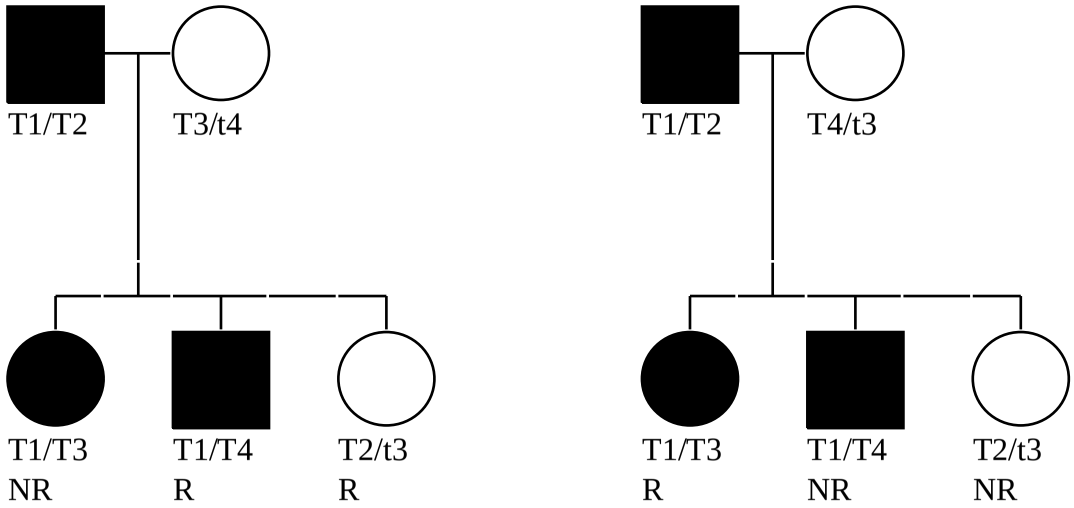
In this specific example, $Z(\theta)$ does not depend on

- the assumed marker allele frequencies, since all members of the pedigree are genotyped at the marker locus.
- the assumed disease allele frequencies, since the observed disease phenotypes in the pedigree imply a unique disease genotype in all members of the pedigree.

However, the assumed mode of inheritance is crucial for $Z(\theta)$!

Calculation of the pedigree likelihood: example

- $f_2 = 1, f_1 = f_0 = 0$, i.e., recessive mode of inheritance:



$\Rightarrow$

$$
\begin{aligned}
L(\theta \mid y) &= \frac{1}{16} \cdot p_T^3 \cdot p_t \cdot q_1 \cdot q_2 \cdot q_3 \cdot q_4 \cdot \theta \cdot (1-\theta) \\
Z(\theta) &= \log_{10}\left[4\theta(1-\theta)\right] \\
\hat{\theta} &= 1/2 \\
Z(\hat{\theta}) &= 0
\end{aligned}
$$