Sibling Pair Linkage Tests

Biostatistics 666
Today ...

- Introduction to linkage analysis of affected siblings

- A simple disease model
  - Probability of sampling affected relative pairs

- Linkage analysis with sibling pairs using Risch’s Maximum LOD Score (MLS)

- Distribution of IBD in affected sibling pairs and Holman’s “Possible Triangle Constraint”
Examplar Linkage Study


- Affected sibling pair study of type 1 diabetes
  - Common chronic disease of childhood
  - 292 affected sibpairs for initial screen
  - 467 affected sibpairs for follow-up

- Highest LOD score reaches 34.2 near HLA on chr. 6
  - At this locus, chromosomes carried by affected siblings are identical 73% of the time.
Examplar Linkage Study Results

Single Locus Disease Model

1. Allele frequencies
   • For normal and susceptibility alleles

2. Penetrances
   • Probability of disease for each genotype

   - Useful in exploring behavior of linkage and association tests

   - Simplification of reality, ignores other loci and the environment
Penetrance

- $f_{ij} = P(\text{Affected} \mid G = ij)$
- Probability someone with genotype $ij$ is affected
- Models the marginal effect of each locus
Using Penetrances

- Allele frequency $p$
- Genotype penetrances $f_{11}$, $f_{12}$, $f_{22}$

- Probability of genotype given disease
  - $P(G = ij \mid D) =$

- Prevalence
  - $K =$
A genetic model can predict probability of sampling different affected relative pairs.

We will consider some simple cases:
- Unrelated individuals
- Parent-offspring pairs
- Monozygotic twins

What do the pairs above have in common?
- HINT: Think about the amount of shared genetic material
What we might expect ...

- Related individuals have similar genotypes

- For a genetic disease...

- Probability that two relatives are both affected must be greater or equal to the probability that two randomly sampled unrelated individuals are affected
Relative Risk and Prevalence

- In relation to affected proband, define
  - $K_R$ prevalence in relatives of type R
  - $\lambda_R = K_R/K$ increase in risk for relatives of type R
- $\lambda_R$ is a measure of the overall effect of a locus
  - Useful for predicting power of linkage studies
Unrelated Individuals

- Probability of affected pair of unrelateds

\[ P(a \text{ and } b \text{ affected}) = P(a \text{ affected})P(b \text{ affected}) \]

\[ = P(\text{affected})^2 \]

\[ = \left[ p^2 f_{11} + 2p(1-p)f_{12} + (1-p)^2 f_{22} \right]^2 \]

\[ = K^2 \]

- For any two related individuals, probability that both are affected should be greater
Monozygotic Twins

- Probability of affected pair of identical twins

\[
P(MZ \text{ pair affected}) = \sum_{G} P(G)P(a \text{ affected} | G)P(b \text{ affected} | G)
\]

\[
= p^2 f_{11}^2 + 2p(1-p)f_{12}^2 + (1-p)^2 f_{22}^2
\]

\[
= K_{MZ}K
\]

\[
= \lambda_{MZ}KK
\]

- \(\lambda_{MZ}\) will be greater than for any other relationship
Parent Offspring Pairs

- Probability of affected parent-offspring pair

\[ P = P(\text{parent and child affected}) \]
\[ = \sum_{G_p} \sum_{G_o} P(G_p, G_o) f_{G_p} f_{G_o} \]
\[ = \sum_{i} \sum_{j} \sum_{k} P(i, j, k) f_{ij} f_{ik} \]
\[ = p^3 f_{11}^2 + (1 - p)^3 f_{22}^2 + p(1 - p) f_{12}^2 + 2 p^2 (1 - p) f_{11} f_{12} + 2 p(1 - p)^2 f_{22} f_{12} \]
\[ = KK_o \]
\[ = \lambda_o KK \]

- \( \lambda_o \) will be between 1.0 and \( \lambda_{MZ} \)
IBD – Identity by Descent

- Sharing of segregating stretch of chromosome within a family

- If a stretch of chromosome is shared IBD, all variants within the stretch will be shared

- At any locus siblings share 0, 1 or 2 alleles IBD
  - Baseline probabilities of IBD 0, 1 and 2 are $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$
For a single locus model...

\[
\begin{align*}
\lambda_{IBD=2} &= \lambda_{MZ} \\
\lambda_{IBD=1} &= \lambda_O \\
\lambda_{IBD=0} &= 1
\end{align*}
\]

- Model ignores contribution of other genes and environment

\[
\begin{align*}
K_{IBD=2} &= K_{MZ} \\
K_{IBD=1} &= K_O \\
K_{IBD=0} &= K
\end{align*}
\]

- Simple model that allows for useful predictions
  - Risk to half-siblings
  - Risk to cousins
  - Risk to siblings
Point of Situation

- Probabilities of affected pairs for
  - Unrelated Individuals
  - Monozygotic Twins
  - Parent-Offspring Pairs

- Each of these shares a fixed number of alleles IBD ...
Affected Half-Siblings

- IBD sharing
  - 0 alleles with probability 50%
  - 1 allele with probability 50%

- This gives ...

\[
\lambda_H = \frac{1}{2} \lambda_O + \frac{1}{2} = \frac{1}{2}(\lambda_O + 1)
\]

\[
K_H = \frac{1}{2} K_O + \frac{1}{2} K = \frac{1}{2}(K_O + K)
\]
Affected Sibpairs

- IBD sharing …
  - 0 alleles with probability 25%
  - 1 alleles with probability 50%
  - 2 alleles with probability 25%

- This gives …

\[
\lambda_S = \frac{1}{4} \lambda_{MZ} + \frac{1}{2} \lambda_O + \frac{1}{4} = \frac{1}{4} (\lambda_{MZ} + 2\lambda_O + 1)
\]

which implies

\[
\lambda_{MZ} = 4\lambda_S - 2\lambda_O - 1
\]
Important Notes...

- We can use allele frequencies and penetrances to estimate probability of affected relative pairs.

- Among sibling pairs, pairs with two alleles “identical-by-descent” have the highest probability of both being affected.
  - Most like “identical twins” for single locus models.
Affected Sibpair Linkage Analyses

- Consider affected sibling pairs
- Consider one genetic marker at a time
- Are paired genotypes more similar than expected?
- Only a subset of all genetic markers must be examined
Likelihood Based Linkage Test

- Depends on three parameters $z_0, z_1, z_2$
  - Probability of sharing 0, 1 and 2 alleles IBD

- Null likelihood uses $z_0=\frac{1}{4}, z_1=\frac{1}{2}, z_2=\frac{1}{4}$

- Alternative likelihood uses MLE for $z_0, z_1, z_2$

- Compare likelihoods with likelihood ratio test
Potential Sib-Pair Likelihood

Under the null hypothesis:

\[ L = \left( \frac{1}{4} \right)^{n_{IBD0}} \left( \frac{1}{2} \right)^{n_{IBD1}} \left( \frac{1}{4} \right)^{n_{IBD2}} \]

Under the alternative hypothesis

\[ L = \left( \hat{z}_0 \right)^{n_{IBD0}} \left( \hat{z}_1 \right)^{n_{IBD1}} \left( \hat{z}_2 \right)^{n_{IBD2}} \]
Likelihood Ratio Based Test Statistics

\[ LOD = \log_{10} \frac{L(\hat{z}_0, \hat{z}_1, \hat{z}_2)}{L(z_0 = \frac{1}{4}, z_1 = \frac{1}{2}, z_2 = \frac{1}{4})} \]

\[ \chi^2 = 2 \ln \frac{L(\hat{z}_0, \hat{z}_1, \hat{z}_2)}{L(z_0 = \frac{1}{4}, z_1 = \frac{1}{2}, z_2 = \frac{1}{4})} \]

\[ = 2 \ln L(\hat{z}_0, \hat{z}_1, \hat{z}_2) - 2 \ln L(z_0 = \frac{1}{4}, z_1 = \frac{1}{2}, z_2 = \frac{1}{4}) \]
In real life...

- Markers are only partially informative
- IBD sharing is equivocal
  - Uncertainty can only be partly reduced by examining relatives
- Need an alternative likelihood
  - Should allow for partially informative data
Desirable Properties

- Models IBD probabilities $z_0$, $z_1$, $z_2$
  - Probability of sharing 0, 1 and 2 alleles IBD

- Uses partial information on IBD sharing

- For unambiguous data, equivalent to previous likelihood
For A Single Family

\[ L_i = \sum_{j=0}^{2} P(\text{IBD} = j \mid \text{ASP}) P(\text{Genotypes}_i \mid IBD = j) = \sum_{j=0}^{2} z_j w_{ij} \]

Risch (1990) defines

\[ w_{ij} = P(\text{Genotypes}_i \mid IBD = j) \]

We only need proportionate \( w_{ij} \)
Likelihood and LOD Score

\[ L(z_0, z_1, z_2) = \prod_i \sum_j z_j w_{ij} \]

\[ LOD = \log_{10} \prod_i \frac{\hat{z}_0 w_{i0} + \hat{z}_1 w_{i1} + \hat{z}_2 w_{i2}}{\frac{1}{4} w_{i0} + \frac{1}{2} w_{i1} + \frac{1}{4} w_{i2}} \]

The MLS statistic is the LOD evaluated at the MLEs of \( z_0, z_1, z_2 \).
### P(Marker Genotype|IBD State)

<table>
<thead>
<tr>
<th>Relative</th>
<th>IBD</th>
<th>IBD</th>
<th>IBD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>(a,b) (c,d)</td>
<td>$4p_a p_b p_c p_d$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(a,a) (b,c)</td>
<td>$2p_a^2 p_b p_c$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(a,a) (b,b)</td>
<td>$p_a^2 p_b$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(a,b) (a,c)</td>
<td>$4p_a^2 p_b p_c$</td>
<td>$p_a p_b p_c$</td>
<td>0</td>
</tr>
<tr>
<td>(a,a) (a,b)</td>
<td>$2p_a^3 p_b$</td>
<td>$p_a^2 p_b$</td>
<td>0</td>
</tr>
<tr>
<td>(a,b) (a,b)</td>
<td>$4p_a^2 p_b^2$</td>
<td>$(p_a p_b^2 + p_a^2 p_b)$</td>
<td>$2p_a p_b$</td>
</tr>
<tr>
<td>(a,a) (a,a)</td>
<td>$p_a$</td>
<td>$p_a$</td>
<td>$p_a^2$</td>
</tr>
</tbody>
</table>

Prior Probability: $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$

These probabilities apply to pair of individuals, when no other genotypes in the family are known.
Example scoring for $w_{ij}$

In this case, relative weights depend on allele frequency.
In these cases, multiple weights are non-zero (but equal) for each family.
How to maximize likelihood?

- If all families are informative
  - Use sample proportions of IBD=0, 1, 2

- If some families are uninformative
  - Use an E-M algorithm
  - At each stage generate complete dataset with fractional counts
  - Iterate until estimates of LOD and z parameters are stable
Assigning Partial Counts in E-M

\[
P(\text{IBD} = j \mid \text{Genotypes}) = \]
\[
= \frac{P(\text{IBD} = j \mid \text{ASP})P(\text{Genotypes} \mid \text{IBD} = j)}{L_i}
\]
\[
= \frac{P(\text{IBD} = j \mid \text{ASP})P(\text{Genotypes} \mid \text{IBD} = j)}{\sum_{k=0}^{2} P(\text{IBD} = k \mid \text{ASP})P(\text{Genotypes} \mid \text{IBD} = k)}
\]
\[
= \frac{z_j w_{ij}}{\sum_{k=0}^{2} z_k w_{ik}}
\]
Example

Assume a bi-allelic marker where the two alleles have identical frequencies.
### Example of E-M Steps

<table>
<thead>
<tr>
<th>Parameters</th>
<th></th>
<th></th>
<th>IBD=0</th>
<th>IBD=1</th>
<th>IBD=2</th>
<th>IBD=2</th>
<th>LOD</th>
<th>LODi</th>
<th>LODu</th>
</tr>
</thead>
<tbody>
<tr>
<td>z0</td>
<td>0.250</td>
<td>0.500</td>
<td>0.250</td>
<td>0.56</td>
<td>2.22</td>
<td>2.22</td>
<td>5</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>z1</td>
<td>0.056</td>
<td>0.222</td>
<td>0.722</td>
<td>0.08</td>
<td>0.66</td>
<td>4.26</td>
<td>5</td>
<td>3.19</td>
<td>2.30</td>
</tr>
<tr>
<td>z2</td>
<td>0.008</td>
<td>0.066</td>
<td>0.926</td>
<td>0.01</td>
<td>0.17</td>
<td>4.82</td>
<td>5</td>
<td>4.01</td>
<td>2.84</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.017</td>
<td>0.982</td>
<td>0.00</td>
<td>0.04</td>
<td>4.96</td>
<td>5</td>
<td>4.20</td>
<td>2.97</td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.004</td>
<td>0.996</td>
<td>0.00</td>
<td>0.01</td>
<td>4.99</td>
<td>5</td>
<td>4.25</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.001</td>
<td>0.999</td>
<td>0.00</td>
<td>0.00</td>
<td>5.00</td>
<td>5</td>
<td>4.26</td>
<td>3.01</td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.00</td>
<td>0.00</td>
<td>5.00</td>
<td>5</td>
<td>4.26</td>
<td>3.01</td>
</tr>
</tbody>
</table>
Properties of Pair Analyses Explored by Risch

- Effect of marker informativeness
- Effect of adding relative genotypes
- Size of genetic effect
- Degree of relationship
Marker Informativeness

Proportion of LOD Retained

Proportion of Expected MLS

Marker Informativeness

- Sibs
- 2nd Degree
- 3rd Degree
Marker Informativeness
Gene of Modest Effect ($\lambda_0=3$)
Marker Informativeness
Gene of Larger Effect ($\lambda_0=10$)

Expected LOD Score

- Sibs
- 2nd Degree
- 3rd Degree
Genotypes of Other Family Members

- Genotyping only pair decreases LOD score by
  - Up to 33% if only sib-pairs are genotyped
  - Up to 60% for second degree relatives
  - Up to 70% for third degree relatives

- Genotyping effort decreases by
  - 50% if only sib-pairs are typed
  - 60% if only second degree relatives typed
  - 75% if only third degree relatives typed
Point of Situation ...

- Noted that affected siblings are more likely to share two alleles identical by descent.

- Derived a likelihood based linkage test that compares sharing probabilities to null defaults.

- Let’s examine these probabilities in more detail ...
Next ...

- Predicting distribution of IBD
  - Modeling marginal effect of a single locus
  - Relative risk ratio ($\lambda_R$)

- The Possible Triangle for Sibling Pairs
  - Plausible IBD values for affected siblings
  - Refinement of the model of Risch (1990)
Recurrence Risks vs. IBD

\[ \lambda_{IBD=2} = \lambda_{MZ} = \frac{P(\text{affected} \mid IBD = 2 \text{ with affected relative})}{P(\text{affected})} \]

\[ \lambda_{IBD=1} = \lambda_{O} = \frac{P(\text{affected} \mid IBD = 1 \text{ with affected relative})}{P(\text{affected})} \]

\[ \lambda_{IBD=0} = 1 = \frac{P(\text{affected} \mid IBD = 0 \text{ with affected relative})}{P(\text{affected})} \]
Bayes' Theorem: Predicting IBD Sharing

\[
P(\text{IBD} = i \mid \text{affected pair}) = \]

\[
= \frac{P(\text{IBD} = i)P(\text{affected pair} \mid \text{IBD} = i)}{\sum_j P(\text{IBD} = j)P(\text{affected pair} \mid \text{IBD} = j)}
\]

\[
= \frac{P(\text{IBD} = i)\lambda_{\text{IBD}=i}}{\sum_j P(\text{IBD} = j)\lambda_{\text{IBD}=j}}
\]
Sibpairs

Expected Values for $z_0$, $z_1$, $z_2$

\[ z_0 = 0.25 \frac{1}{\lambda_s} \]
\[ z_1 = 0.50 \frac{\lambda_o}{\lambda_s} \]
\[ z_2 = 0.25 \frac{\lambda_{MZ}}{\lambda_s} \]

\[ 1 \leq \lambda_o \leq \lambda_s \leq \lambda_{MZ} \text{ for any genetic model} \]
Possible Triangle

$z_0 = \frac{1}{4}, \; z_1 = \frac{1}{2}$

Area covering all possible values for sharing parameters
The yellow triangle indicates possible true values for the sharing parameters for any genetic model.

\[ H_0: \quad z_0 = \frac{1}{4}, \quad z_1 = \frac{1}{2} \]

\[ H_1 \]
Intuition

- Under the null
  - True parameter values are \(\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)\)
  - Estimates will wobble around this point

- Under the alternative
  - True parameter values are within triangle
  - Estimates will wobble around true point
Idea (Holmans, 1993)

- Testing for linkage
  - Do IBD patterns suggest a gene is present?

- Focus on situations where IBD patterns are compatible with a genetic model
  - Restrict maximization to possible triangle
The possible triangle method

1. Estimate $z_0$, $z_1$, $z_2$ without restrictions

2. If estimate of $z_1 > \frac{1}{2}$ then ...
   a) Repeat estimation with $z_1 = \frac{1}{2}$
   b) If this gives $z_0 > \frac{1}{4}$ then revert to null (MLS=0)

3. If estimates imply $2z_0 > z_1$ then ...
   a) Repeat estimation with $z_1 = 2z_0$
   b) If this gives $z_0 > \frac{1}{4}$ then revert to null (MLS=0)

4. Otherwise, leave estimates unchanged.
Possible Triangle

Holman's Example:

IBD Pairs
0  8
1  60
2  32

MLS = 4.22 (overall)
MLE = (0.08, 0.60, 0.32)

MLS = 3.35 (triangle)
MLE = (0.10, 0.50, 0.40)
MLS Combined With Possible Triangle

- Under null, true $z$ is a corner of the triangle
  - Estimates will often lie outside triangle
  - Restriction to the triangle decreases MLS
  - MLS threshold for fixed type I error decreases

- Under alternative, true $z$ is within triangle
  - Estimates will lie outside triangle less often
  - MLS decreases less
  - Overall, power should be increased
Example

- Type I error rate of 0.001
- LOD of 3.0 with unrestricted method
  - Risch (1990)
- LOD of 2.3 with possible triangle constraint
  - Holmans (1993)
  - For some cases, almost doubles power
Recommended Reading

- Holmans (1993)
  Asymptotic Properties of Affected-Sib-Pair Linkage Analysis
  *Am J Hum Genet* **52**:362-374

- Introduces possible triangle constraint
- Good review of MLS method
Recommended Reading

- Risch (1990)
  - Linkage Strategies for Genetically Complex Traits. III. The Effect of Marker Polymorphism on Analysis of Affected Relative Pairs

- Introduces MLS method for linkage analysis
  - Still, one of the best methods for analysis pair data

- Evaluates different sampling strategies
  - Results were later corrected by Risch (1992)