Non-parametric methods for linkage analysis

To this point, we have discussed “model-based” linkage analyses. These require one to specify a genetic model. For example, we looked at a diallelic disease locus and parameterized the disease allele frequency and the three penetrances ($f_1, f_2, f_3$) associated with the three genotypes.

Clearly, such a model is not sufficient to accurately characterize the underlying etiology of complex traits. The true model for a complex trait is likely not caused by a single gene but rather depends on multiple genetic and environmental factors.

So what do we do? There have been two general approaches to this problem:

1. continue to use the “parametric” framework, acknowledging that it is an approximation to reality
2. abandon the “model-based” or “parametric” framework for linkage analysis and focus on the sharing of disease status and sharing of marker alleles by sets of relatives

We will discuss two “model-free” or “non-parametric” methods of linkage analysis.

Measures of allele-sharing by relatives

Two alleles share “identity by state” (IBS) if they are the same, regardless of whether they were inherited from the same source. Two alleles share “identity by descent” (IBD) if they are copies of the same ancestral allele.
Exercise: Suppose that marker locus A and B are tightly linked so that no recombination between them is expected. Consider the pair of siblings. For each locus, how many alleles are IBS? How many alleles are IBD?

A locus: 2 alleles IBS
B locus: 1 allele IBS

IBD at A locus: both must have received the $A_1$ allele from dad, so that is IBD. Since Sue received $A_2B_2$ from mom and Mary received $A_2B_1$, they don’t share $A_2$ IBD because of the tight linkage between the loci.

IBD at B locus: Sue got $B_2$ from mom, Mary got $B_1$, so not IBD. On the other hand, they got the same $A_1$ allele from dad, so they got the same linked B allele. 1 allele IBD.
The point of the previous example is that we could figure out more about IBD status by having information on two tightly linked markers than if we had had unlinked markers. IBD status does not vary randomly along the chromosome.

The basic idea behind non-parametric linkage analysis is to find genes or markers where IBD sharing correlates with phenotypic sharing. IBD values for loci that are unlinked to a disease locus will be independent of IBD values of a disease locus. Conversely, IBD values will be positively correlated for linked loci with the correlation increasing for loci increasingly close.

**Exercise:** For a marker locus that is unlinked to the disease locus, we would like to know the expected IBD distribution for a random marker. To do this, suppose the parents have genotypes AB and CD so that they will produce children with genotypes AC, AD, BC and BD each with a probability of ¼. The following table characterizes the IBD sharing for all possible sib-pairs with each cell occurring with probability 1/16. Given this, what is the probability that the sibs share 0, 1 and 2 alleles IBD? In addition, what is the expected number of alleles shared IBD?

<table>
<thead>
<tr>
<th>Sibling 1 Genotype</th>
<th>Sibling 2 Genotype</th>
<th>AC</th>
<th>AD</th>
<th>BC</th>
<th>BD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>BC</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>BD</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

\[
P(0 \text{ alleles IBD})=\frac{4}{16}=\frac{1}{4}\\P(1 \text{ alleles IBD})=\frac{8}{16}=\frac{1}{2}\\P(2 \text{ alleles IBD})=\frac{4}{16}=\frac{1}{4}\\
\]

Let \(X=\text{number of alleles IBD}\)

\[
E(X) = \sum_{i=0}^{2} iP(X = i) = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1
\]
This distribution corresponds to the expected allele sharing under the null hypothesis of no linkage. For marker loci linkage to the disease locus, how would we expect the IBD sharing to change?

**IBD sharing should be elevated for diseased siblings, e.g. \( P(\text{siblings share 2 alleles IBD}) > \frac{1}{4} \) for loci linked to the disease locus.**

**Affected Sib–Pair Methods**

We sample families with two affected siblings and genotype the siblings and their parents. At a genetic marker, count the number of alleles IBD at the marker. If the marker is linked to a disease locus, we expected a greater amount of allele sharing than the null distribution of \( \frac{1}{4} \), \( \frac{1}{2} \), and \( \frac{1}{4} \) for 0, 1, and 2 alleles IBD.

Two commonly used test statistics are:

- the Pearson chi-squared statistic for comparing the observed counts of sib pairs sharing 0, 1 and 2 alleles IBD with that expected under the null of no linkage, i.e. \( E_0(n_0, n_1, n_2) = (N/4, N/2, N/4) \) where \( N \) is the number of sib-pairs. This statistic is asymptotically chi-squared with 2 degrees of freedom under the null hypothesis:

\[
S_1 = \sum_{i=0}^{2} \frac{(n_i - e_i)^2}{e_i}
\]

- the mean test compares the proportion of alleles shared IBD with the expected sharing under the null of no linkage. In simulation studies, the mean test performs well across a wide-range of conditions (appropriate Type I error, powerful across most conditions). The statistic is asymptotically standard normal under the null hypothesis:

\[
S_2 = \frac{\left( \frac{n_1}{2} + \frac{n_2}{2} \right) - \frac{N}{2}}{\frac{n}{8}}
\]

Other allele-sharing statistics are based on the same idea: we compare the allele sharing pattern we would expect under the null hypothesis of no linkage with the observed allele sharing pattern. We expected that markers more closely linked to a disease locus will exhibit greater deviation from the null hypothesis proportions of \( \frac{1}{4} \), \( \frac{1}{2} \), \( \frac{1}{4} \) for 0, 1, and 2 alleles IBD.
Haseman-Elston regression

The same ideas can be applied to quantitative traits. If a marker locus is linked to a QTL, the difference in trait values between two relatives is expected to decrease as they share more marker alleles IBD. Another way to say this is that trait values are expected to be more similar in relatives who share more alleles IBD.

Basic idea behind Haseman-Elston regression:

- suppose that the marker under study is linked to a QTL
- pairs of relatives that share marker alleles IBD will share QTL alleles IBD if the marker is linked to the QTL
- the quantitative traits for these relatives should be more similar than those relatives that are not IBD

In Haseman-Elson regression, the “dependent variable” is the difference in trait values between two relatives, \( Y_j = (Z_{1j} - Z_{2j})^2 \). \( Y \) is expected to decrease as relatives share more alleles IBD. Specifically, the regression equation is

\[
E(Y_j \mid \pi_{jm}) = \alpha + \beta \pi_{jm},
\]

where \( \pi_{jm} \) is the proportion of alleles shared IBD for the \( j \)th pair of relatives at marker \( m \).

Note that this strategy can work for different types of relatives, not just sib-pairs. However:

- different types of relative pairs cannot be used in the same regression
- cannot use parent-offspring pairs...why?

**They always share 1 allele IBD**
It can be shown that

<table>
<thead>
<tr>
<th>Relative Type</th>
<th>slope $\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>siblings</td>
<td>$\beta = -2(1 - 2\theta)^2 \frac{\sigma_A^2}{\theta}$</td>
</tr>
<tr>
<td>grandparent-grandchild</td>
<td>$\beta = -2(1 - 2\theta)\sigma_A^2$</td>
</tr>
<tr>
<td>aunt/uncle-nephew/niece</td>
<td>$\beta = -2(1 - 2\theta)(1 - \theta)\sigma_A^2$</td>
</tr>
</tbody>
</table>

where $\theta$ is the recombination fraction between the marker and disease locus.

The null and alternative hypotheses of the test are:

$H_0: \beta = 0$

$H_A:$

Focusing on siblings, a true null hypothesis corresponds to:

**Comments**

- Residuals tend to be heteroscedastic because the variance of the trait usually varies with the number of alleles IBD. It is important to use regression methods that do not assume homoscedasticity.
- $\beta$ depends on $\theta$ and $\sigma_A^2$, but these cannot be separately estimated by estimating $\beta$. This is a testing procedure, not an estimation procedure.
- The power of the test varies with $\theta$ and $\sigma_A^2$.
- The method depends on determining IBD status. In the case of sibships, we want the parents to be segregating four different alleles at the markers.
Advantages of Non-Parametric Linkage Analysis over Parametric Linkage Analysis

1. Conceptually simple
2. Do not need to specify a genetic model
3. Lots of small pedigrees might be more easily obtained than a large pedigree

Disadvantages of Non-Parametric Linkage Analysis over Parametric Linkage Analysis

1. It is strictly a testing procedure, does not allow estimation of important parameters like $\theta$ and $\sigma^2_i$.
2. The distortion of the IBD-sharing distribution will vary with the true genetic model, and the optimal way to test for IBD-sharing depends on the true genetic model. So you really implicitly assume a genetic model when you decide what test to use.
Other issues in linkage analysis

We have discussed the basics of linkage analysis for both model based and non-model based methods. Here is a partial summary of some of the issues we have discussed, and some additional issues:

- **Pedigree structure** – While we don’t have complete control over the types of pedigrees available, we will generally have some flexibility with regard to pedigree structure. It is important to consider what relatives provide more information than others. Chapter 5 of Ott¹ provides guidance for simple family structures and model based methods. New methods in statistical genetics are generally published with information on power. This will provide you some data regarding power for your study; however, each study is certainly unique. Simulation studies provide a flexible framework for evaluating the power of various pedigree structures and methods. Section 9.7 of Ott is a good place to start if you are considering simulation studies.

- **Random versus non-random ascertainment of pedigrees** – For disease studies, sampling must be done through diseased individuals in order to obtain enough data to assess segregation of disease within families. This ascertainment must be accounted for when trying to apply estimates of parameters from the study sample to the overall population. (In handout 4b we discussed ascertainment bias with an example that made the bias clear.)

- **Phenotypic characterization** – The power of any segregation or linkage study is greatly affected by the homogeneity of the phenotype under study. Unfortunately, we often don’t have a good handle on how to best characterize complex traits so that additional phenotypic characterization could lead to better defined sub-groups of disease that yield greater power than analyzing the overall heterogeneous sample!

- **Marker informativeness and marker spacing** – The power for any method of linkage analysis will depend on the informativeness of the marker or marker set, i.e. in order to test for linkage you need to be able to see the segregation in the region under study. In addition, you need to have markers spaced such that you are close enough to detect a trait locus.