Linkage of Qualitative and Quantitative traits:

New tools have changed the way plant breeders go about the business of their science. This has primarily been accomplished by merging new biological techniques such as molecular markers with advances in computer hardware and software. Advances have made possible the generation of linkage maps with a resolution that approaches 1 centi-Morgan. In addition, techniques for mapping and characterizing quantitative trait loci have improved to the point where QTL have been cloned based on positional information in both mouse and tomato.

During the next several classes we will explore linkage analysis and genetic mapping. The implications of applying “genome sciences” to plant breeding are several:

“Map-based” breeding, an approach that incorporates positional information into breeding strategies, can improve the efficiency of crop improvement.

- How easy will it be to pyramid genes (e.g. tendency of resistance genes to cluster)?
- Do we need to select for coupling phase linkage or break unfavorable linkage?

We can more completely characterize quantitative traits (a process sometimes called “mendelizing QTL”).

- Identify the “gene action” and estimate the magnitude (\( V_{QTL/V_P} \) or \( V_{QTL/V_G} \)) of individual loci.
- Select for genotype when phenotypic selection is inefficient.
We can synthesize information from numerous crosses to understand patterns in genome organization.

- Clusters of disease resistance
- Candidate genes for QTL

The power of modern genetic mapping techniques to aid breeding will be realized only if we use the techniques with applied goals clearly in mind.

- Combine classical and new techniques.
- Place “economic” traits onto maps.
- Work in backgrounds that are appropriate for crop improvement.

**Building Genetic Maps.**

Constructing a genetic map requires the ordering of loci and the measurement of distance between them. There are several methods that can be used to determine linkage. These methods vary in efficiency and accuracy. Genetic analysis with the goal of building genetic maps requires the following general steps regardless of statistical methodology:

- Single-locus analysis
- Two-locus analysis
- Measuring linkage distance
- Linkage grouping
- Gene ordering
- Multi-point analysis

Today, we will discuss the linkage of qualitative traits. We will concentrate on the first steps, establishing linkage and measuring linkage distance. One way to view the early stages of linkage analysis is to use the context of Mendel’s laws of segregation and independent assortment. We ask the question “does a single locus segregate as expected?” and “do pairs of loci assort independently?” If the answer to the second
question is “no” we have evidence supporting genetic linkage and can proceed to estimate the distance between the loci in question.

**Definition for Genetic linkage**

- The association of genes located on the same chromosome
- The association or non-independence among alleles at more than one locus (statistical definition).

**I. Linkage between Qualitative traits**

The estimation of linkage between two qualitative traits requires systematically answering the following questions:

**Does each locus segregate as expected?** (Single locus analysis)

Statistic: $X^2$ to test Mendelian ratios appropriate for cross (BC, 1:1; F2, 3:1, 1:2:1, etc…)

**Do the loci assort independently?** (Two-locus analysis)

Statistic: $X^2$ to test independent assortment.

Log Likelihood Ratio

**What is the distance between loci?**

Estimation of linkage has been performed by multiple techniques with different strengths and weaknesses. A review, from a historical perspective, follows.

**Estimation of linkage**

If the two loci do not assort independently, we may proceed with the estimation of linkage (or recombination fraction estimation). Let’s walk through the process of linkage analysis for some simple cases. By crossing the double heterozygote back to the double recessive (AaBb x aabb) yields a direct measure of recombinant individuals. Selfing a heterozygote (AaBb with respect to two loci, A and B) will produce gametes of four types AB, Ab, aB, & ab (this is the case of an F2 population). If the loci are unlinked, equal numbers of the four types will be produced. If linked, the classes AB and ab will appear with a frequency different from the Ab and aB classes. For progeny of a selfed heterozygote, unlinked genes will give 9:3:3:1 ratio (assuming dominance) and linkage will be detected by an excess of the first and last classes in coupling phase and the second and third classes in repulsion phase. Recombination may be estimated from the observed numbers in the expected classes using a number of approaches. A brief overview follows.

First, “counting recombinants” offers an intuitive approach but may not always be practical. In our F2 example, it is not possible to determine whether the double
heterozygote (AaBb) resulted from non-recombinant or double recombinant gametes. Also, if the markers are dominant, not all recombinant progeny will be counted.

The first step in estimating linkage is to express the expected proportions as a probability. Note that probability expressions will be specific to the “experimental design” or cross. For an F2 derived from selfing the heterozygote we expect gametes in the pollen (male) and ovule (female) to have proportions as follows:

\[
\begin{array}{cccc}
\text{AB} & \text{Ab} & \text{aB} & \text{ab} \\
\text{Ovule} & 1/2p & 1/2q & 1/2q & 1/2p \\
\text{Pollen} & 1/2p' & 1/2q' & 1/2q' & 1/2p'
\end{array}
\]

Where \( p = (1-q) \)

Where \( p' = (1-q') \)

The mathematical description acknowledges that recombination may be different in male and female parents. From these proportions we can estimate the fraction of offspring in the double recessive class as \( 1/4 pp' \). Recombination can be expressed as a probability based in terms of the single quantity \( pp' \). If recombination is the same in both sexes, we can designate \( pp' \) as \( \theta \). Recombination is \( (\theta)^{1/2} \) in coupling phase and \( 1-(\theta)^{1/2} \) in repulsion phase.

Different authors have used different notations to describe the expected frequencies. Some of these follow in the table:

<table>
<thead>
<tr>
<th>Gametes</th>
<th>Class</th>
<th>Unlinked</th>
<th>Linked</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_B_</td>
<td>a</td>
<td>9n/16</td>
<td>( 1/4(2 + pp') )</td>
</tr>
<tr>
<td>A_bb</td>
<td>b</td>
<td>3n/16</td>
<td>( 1/4(1-pp') )</td>
</tr>
<tr>
<td>AaB_</td>
<td>c</td>
<td>3n/16</td>
<td>( 1/4(1-pp') )</td>
</tr>
<tr>
<td>aabb</td>
<td>d</td>
<td>n/16</td>
<td>( 1/4 pp' )</td>
</tr>
</tbody>
</table>

When reading about the estimation of linkage, pay close attention to notation. Fisher uses \( x \) to refer to \( pp' \), Wier uses \( \theta \) to refer to the same and \( c \) to refer to recombination frequency. Allard uses \( p \) to refer to recombination fraction (in other words \( p^2 = \theta \)). Liu uses \( \theta \) to refer to recombination frequency. Lynch and Walsh use \( c \) to refer to recombination frequency and \( m \) to refer to map distance (recombination frequency corrected by a mapping function for double recombination).

The estimation of linkage

- Additive method
- Weighted mean method
- Product method
- Maximum likelihood
- Log of the odds
The estimation of linkage

For a cross between a red fruited/yellow leafed inbred tomato and a pink fruited green leafed inbred line a red fruited green leafed hybrid was obtained. Selfing the hybrid lead to the following progeny classes:

<table>
<thead>
<tr>
<th></th>
<th>Red fruit</th>
<th>Pink fruit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green leaves</td>
<td>1997</td>
<td>904</td>
</tr>
<tr>
<td>Yellow leaves</td>
<td>906</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Total: 3839</td>
<td></td>
</tr>
<tr>
<td>Red fruit: pink fruit</td>
<td>2903:936</td>
<td></td>
</tr>
<tr>
<td>Green leaves: yellow leaves</td>
<td>2901:938</td>
<td></td>
</tr>
</tbody>
</table>

From these ratios we can see that Red fruit is dominant to pink fruit and green leaves is dominant to yellow leaves. The expected ratios for unlinked genes would be 9:3:3:1, or expected classes of 2159.4: 719.8: 719.8: 240. Using a Chi-square test we must reject the hypothesis that the genes controlling the two traits are unlinked. The linkage is in repulsion.

Additive method (Emerson, <1928).

Recombination is calculated from the sum of the first and fourth classes (AB and ab). Emerson’s method equates a + d (the observed frequencies of the double recessive and the double dominant phenotypes) to the expected frequency $n/4(2 + 2\theta) \equiv (n/4(2 + \theta) + n/4 \theta)$. The equation $a+d = n/4(2+2\theta)$ is equivalent to $N(\theta) = a-b-c+d$. After solving for $\theta$ and the solution is converted to recombination using the equation for repulsion $\theta^{1/2}$ or for coupling $1 - (\theta)^{1/2}$. A further conversion to % recombination (or centi-Morgans) is required.

Emerson’s method uses only some of the available information and the variance is therefore inflated. The variance of $\theta$ is $(1-(\theta)^2)/n$ which must also be converted to a variance for recombination (or to variance in centi-Morgans).

\[
\frac{1 - (\theta)^{1/2} \times 100}{\text{cM for repulsion phase}}
\]

\[
(\text{SE}(\theta) = ((1-(\theta)^2)/n)^{1/2} \quad \text{SE cM} = ((\text{cM} x ((1-(\theta)^2)/n))/\theta)^{1/2}
\]

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>SE $\theta$</th>
<th>cM</th>
<th>SE cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.057046</td>
<td>0.16113</td>
<td>23.8843</td>
<td>3.2971</td>
</tr>
</tbody>
</table>
**The weighted mean method**

The value of $\theta$ is estimated from an alternative linear function of the frequencies. The alternative linear function is developed using the expected frequency for unlinked loci to weight the observed frequencies (in our example 9:3:3:1). The expression $a - 3b - 3c + 9d$ is equated to $n(4\theta - 1)$ to derive the equation:

$$4n(\theta) = 2a - 2b - 2c + 10d$$

The sampling variance for $\theta$ is $1 + 6\theta - 4(\theta)^2 / 4n$

As in Emerson’s method, $\theta$ and its variance need to be converted to recombination and its variance.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>SE $\theta$</th>
<th>cM</th>
<th>SE cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.045194</td>
<td>0.00907</td>
<td>21.26</td>
<td>2.133</td>
</tr>
</tbody>
</table>

**Product method (Method of Bridges)**

For the product method (also referred to as Bridges’ method from C. B. Bridges, 1914, *The chromosome hypothesis of linkage. American Naturalist XLVIII 524*), the value $\theta$ is estimated from the products of the first and fourth classes and the second and third classes (for $F_2$ in repulsion) as follows:

$$\frac{a \times d}{b \times c} = \frac{\theta (2 + \theta)}{(1 - \theta)^2}$$

$$\frac{1997 \times 32}{906 \times 904} = \frac{\theta (2 + \theta)}{(1 - \theta)^2}$$

The positive solution to the quadratic equation is $\theta = 0.035645$ or 18.8799 cM

The random sampling variance for $\theta$ using the product method is:

$$2\theta (1 - \theta)(2 + \theta) / n(1 + 2\theta)$$

The SE of $\theta$ is 0.00584 and the SE in cM is 1.545

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>SE $\theta$</th>
<th>cM</th>
<th>SE cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.04356454</td>
<td>0.00584</td>
<td>18.8799</td>
<td>1.545</td>
</tr>
</tbody>
</table>
### Comparison of methods

<table>
<thead>
<tr>
<th></th>
<th>cM</th>
<th>SE</th>
<th>comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting recombinants</td>
<td>18.26</td>
<td>nc</td>
<td>no statistical test</td>
</tr>
<tr>
<td>Emerson’s method</td>
<td>23.88</td>
<td>3.373</td>
<td>provides most information when $\theta = 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(complete linkage measured in coupling)</td>
</tr>
<tr>
<td>Weighted mean method</td>
<td>21.26</td>
<td>2.133</td>
<td>provides most information when $(\theta)^{1/2} = 0.5$ (unlinked)</td>
</tr>
<tr>
<td>Product method</td>
<td>18.88</td>
<td>1.545</td>
<td>efficient at all values of $\theta$</td>
</tr>
<tr>
<td>Maximum Likelihood</td>
<td>see below</td>
<td></td>
<td>efficient at all values of $\theta$</td>
</tr>
</tbody>
</table>

Methods for estimating recombination may use only partial information and may not have a high level of accuracy. Emerson’s method and the Weighted mean method are imperfect estimators of linkage, but they provide useful tests of significance for apparent linkage or non-linkage (respectively). For the weighted mean method, the square of $a - 3b - 3c + 9d$ should exceed $36n$, when there is no linkage.

The product method will yield the same result as maximum likelihood for both $p$ and variance when population sizes are large. The product method is facilitated by the use of tables published by Bridges to help choose the proper products for multiplication and therefore is not easily generalized for computer analysis.

For mapping purposes we want to use methods that are accurate, can be generalized, and have a small sampling variance. It can be proven mathematically that the maximum likelihood approach yields the smallest possible sampling variance. Thus the product method and maximum likelihood are “efficient statistics” for two-point linkage analysis.

### Maximum likelihood estimation for the calculation of recombination values.

**The concept:**
- Derive the probability equation for expected recombination values.
- Differentiate and set equal to 0 (i.e. find the minimum/maximum). The solution is an estimate of recombination.

**The practice:**
- Probability equations are sometimes not easy to differentiate.
- The log (e) of the probability equations and the probability equations will show maxima at the same values.
Differentiation is easier for the log(e) of the probability equations.

The Generalized log-likelihood equation:

\[ \frac{dL}{dp} = a_1(d \ln m_1/dp) + a_2(d \ln m_2/dp) + a_3(d \ln m_3/dp) + \ldots a_j(d \ln m_j/dp) \]

\( p \) represents the recombination fraction \( (\theta)^{1/2} \).

\( a \) represents the observed frequency

\( m \) is the expected frequency

The derivative of the log likelihood equation will result in a quadratic equation. The positive solution to the quadratic equation will be an estimation of recombination. For our example (dominant genes in repulsion phase), the log (e) of the probability equations and would look something like this:

\[ = a\ln(2+\theta) + b\ln(1-\theta) + c\ln(1-\theta) + d\ln(\theta) \]  

\( \text{Note } \theta = pp' = p^{2} \)

Taking the derivative leads to the quadratic equation:

\[ 3839\theta^2 + 1655\theta - 64 = 0 \]

the positive solution for \( \theta \) is 0.035712 (18.898 cM  SE 1.545)

There is an alternative to taking the derivative and solving the quadratic equation. The first derivative of the log-likelihood equation for \( F_2 \) (9:3:3:1) in repulsion phase is equation 6 from Allard’s (1956) paper. Iteration of this Maximum-likelihood estimation equation can be used to arrive at an answer. Iterative processes are ideally suited for computers. Iteration works as follows:

Substitute trial values, and solve
+ sign estimate was too low
- sign estimate was too high
revise the estimate, and solve
revise the estimate, and solve (or interpolate)

\[ = 1997*(2*(p)/(2+p*p)))+((906+904)*(-2*(p))/(1-(p*p)))+(32*(2/p)) \]

Note that this equation could be written into EXCEL to favor iteration…

<table>
<thead>
<tr>
<th>( P )</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>-1397.78</td>
</tr>
<tr>
<td>0.1</td>
<td>473.05</td>
</tr>
<tr>
<td>0.16</td>
<td>121.06</td>
</tr>
<tr>
<td>0.19</td>
<td>-4.01</td>
</tr>
<tr>
<td>0.18</td>
<td>35.87</td>
</tr>
<tr>
<td>0.18898</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Reference:

**Estimation of linkage using the Log Likelihood approach.**

A test Statistics for testing the probability of linkage is the Log likelihood approach. The advantage of the generalized Log likelihood approach is that it facilitates combining data from multiple crosses (especially in situations where it is not possible to clearly partition all allelic classes).

Definition: The logarithm to base 10 of the ratio (likelihood of linkage/likelihood of unlinked loci) is the LOD SCORE (sometimes called Z). Note that the likelihood of unlinked loci is 0.5.

\[
Z = \log_{10}\left(\frac{\text{observed}}{\text{Expected}}\right)
= \log_{10}\left(\frac{L_p}{L_{0.5}}\right) = \log_{10}(L_p) - \log_{10}(L_{0.5})
\]

where \(L\) refers to the log likelihood function

Note that the LOD SCORE reflects the probability for linkage and still requires an estimate of \(p\). In practice this is almost always derived from a Maximum Likelihood estimation using some form of iteration.

The likelihood function takes the form of

\[
L = \sum \text{observed counts non-recombinant individuals} \times \log (\text{expected frequencies})
+ \sum \text{observed counts recombinant individuals} \times \log (\text{expected frequencies})
\]

\[
L = \sum \text{observed counts} \times \log (1-p) + \sum \text{observed counts} \times \log (p)
\]

Note that \(p\) can be estimated by maximizing the likelihood function as with the maximum likelihood equations from Allard’s paper. Maximum likelihood is the principle method for obtaining estimates of \(p\) and various techniques for iteration are used in practice.

The key feature is that information can be combined from independent families. The joint likelihood calculated from independent families is the product of the separate likelihoods and the overall lod score is the sum of the separate scores.
When the MLE recombination < 0.5 is more likely than = 0.5, the lod score is positive. It is conventional to conclude that lod 3 (1,000 to 1) is indicative of linkage and lod −2 is evidence of no linkage. Use of lod scores still require an estimation of recombination. Keep in mind that different significance criteria will be applicable to data from different experiments.

A note on the Allard paper. The description is a bit awkward “All equations in table 6 and elsewhere are given for repulsion phase, and the substitution of 1-p for p (accompanied by the sign change) must be made for all coupling data …”

This means: substitute 1-p for p, then multiply by −1.

Rep    Coupling
1/p    1/(1-p)*-1 = 1/(-1+p)

Rep    Coupling
1/(p-1) 1/((1-p)-1)*-1 = 1/(-p)*-1 = 1/p

etc…

Backcross data:

Repulsion Phase (Allard’s equation 2)

\[(a+d)\times\frac{1}{p} + (b+c)\times\frac{1}{(p-1)} = 0\]

Coupling Phase

\[(a+d)\times\frac{1}{(p-1)} + (b+c)\times\frac{1}{p} = 0\]

F₂ data (Allard’s equation 6)

\[a\left[\frac{2p}{(2+p^2)}\right] + (b+c)\left[-\frac{2p}{(1-p^2)}\right] + d\left(\frac{2}{p}\right) = 0\]

\[a\left[\frac{2(p-1)}{(3-2p+p^2)}\right] + (b+c)\left[\frac{2(1-p)}{p(2-p)}\right] + d\left(\frac{2}{(p-1)}\right)\]

Exercise:
Using the formulas from the Allard paper, write an equation in EXCEL (a “macro”) to estimate the linkage of two markers, TG23 and PTO, in the backcross data provided in the file “lnkspt.xls”. You can check your answer by counting recombinants. From the equations in Allard’s Hilgardia paper, you can also calculate a variance for your estimate. The file “lnkspt.xls” contains data for the backcross: (H7998 X 88119) X 88119. Data are a visual score (yes or no) for disease, the log of bacterial populations per gm of tissue, and marker data for TG23 and PTO.