

ESTIMATION OF LINKAGE

Multiple symbols have been used to denote recombination. In this class, we will use “ θ ”
 F_2 progeny from AaBb in repulsion (3:1)(3:1):

Gametes or backcross and F_2 phenotypes

	AB	Ab	aB	ab	Total
1. Frequency of gamete types	$\frac{\theta}{2}$	$\frac{1-\theta}{2}$	$\frac{1-\theta}{2}$	$\frac{\theta}{2}$	1
2. Exp. prop. of phenotypes in F_2	$\frac{2+\theta^2}{4}$	$\frac{1-\theta^2}{4}$	$\frac{1-\theta^2}{4}$	$\frac{\theta^2}{4}$	1
3. Calculated zygotic series in F_2	$\frac{n(2+\theta^2)}{4}$	$\frac{n(1-\theta^2)}{4}$	$\frac{n(1-\theta^2)}{4}$	$\frac{n(\theta^2)}{4}$	
4. Observed number in backcross or F_2	n_1	n_2	n_3	n_4	n

** “ θ ” is always AB + ab, so that in repulsion, θ is the recombination fraction, but in coupling, θ is the parental combination fraction, the recombination fraction being $1 - \theta$.

Some facts:

1. AB and ab gametes will not be observed in this cross unless there is recombination (i.e. $\theta > 0$).

2. We will only see “aabb” F_2 phenotypes if two “ab” gametes unite, frequency $\frac{\theta}{2} * \frac{\theta}{2} = \frac{\theta^2}{4}$

3. We will see 2/4 AB phenotypes, if there is no recombination between A&B (i.e. complete linkage). However, if there is recombination, we will see an increase in this class by the same amount as the number of “aabb” individuals we observe, i.e. $\theta^2/4$. Therefore, the number of AB phenotypes seen with a given level of recombination,

$$\theta = \frac{2}{4} + \frac{\theta^2}{4} = \frac{2+\theta^2}{4}$$

1. The remaining two classes, Ab and aB, represent parental types whose frequency will decline if there is any recombination. Their frequency of occurrence is equal, and their loss of numbers is also equal.

We would expect each to lose: $\theta^2/4$.

2. We can also calculate the expected proportions directly from the gametic array:
e.g. for AB:

$$\frac{(\theta)(\theta)}{4} + \frac{2(1-\theta)(1-\theta)}{4} + \frac{2(1-\theta)(\theta)}{4} + \frac{2(1-\theta)(\theta)}{4} + \frac{2(\theta)(\theta)}{4}$$

These terms relate to AB*AB + 2*Ab*aB + 2*AB*Ab + 2*AB*aB + 2*AB*ab

All these terms except the first are multiplied by 2, because each gamete could have come from either parent—i.e., there is no constraint on the gamete contributed by the first parent. However, the AB*AB class is constrained, because the first parent must produce an AB. In the other cases, the first parent could produce *either* gamete, thus multiply by two for the two alternatives; e.g., Ab*aB and aB*Ab.

Using similar reasoning, you can determine the expected proportions for any segregation type. You may need to do this if you have a strange segregation ratio for some traits you are studying. If you can determine these proportions, you can use maximum likelihood methods to establish a linkage value.

For now, we will assume that male and female recombination rates are the same. If not (and they have been shown to vary), different recombination rates need to be included in the above table.

In Row 2, by entering a given recombination value, you can calculate the expected proportions of phenotypes in the F₂. E.g., for 20% recombination in repulsion, enter 0.20 for θ , in coupling, enter 0.80 for θ .

The big problem, though, is that we do not know what value of θ to include—therefore we need to estimate its value using one of several methods. The most widely used methods of estimating linkage are (1) maximum likelihood and (2) the product method (Fisher and Balmukand, 1928), which provides good estimates, but only in certain situations.

Note:

- Backcross allows a direct estimation of recombination, whereas F₂ at best provides estimates of either only the square of the recombination fraction or of the square of one minus this fraction.
- When the recombination varies between male and female parents, the product of the two recombination fraction or one minus each fraction, is estimated.

I. MAXIMUM LIKELIHOOD ESTIMATION OF LINKAGE

The most precise estimate of recombination will be found using the method of maximum likelihood. Information on Maximum Likelihood Estimation can be found in Allard (1956), Liu (1998), Lynch and Walsh (1998), Mather (1957), and Weir (1996).

Advantages of MLE:

1. Always provides the lowest standard error of θ .
 2. Can be used when classes are missing or very small.
- However, it can give spurious results if there is segregation distortion (to be discussed later).
 - With maximum likelihood methods, one develops a likelihood equation that, when maximized, will give the most precise estimate of the recombination fraction that would result in the observed data. Maximizing an equation is most easily done by differentiation, as we know from basic calculus.
 - From our segregation data, we know that a certain number of individuals are observed in each of several phenotypic or genotypic classes, e.g. A-B-, A-bb, aaB-, and aabb. A multinomial model is statistically appropriate for this situation (with some assumptions), where we are randomly sampling individuals from a population, and those individuals can fall into several categories. Further information on multinomial models is available in Larson (1969) or other probability theory books.
 - We can write a probability for the data we observed by fitting a multinomial distribution in the following way:

Let,
 n_i = the number of individuals in the i th genotypic (phenotypic) class $\left(\sum_{i=1}^k n_i = n \right)$

k = the number of genotypic classes, each class indexed by a subscript i .

P_i = the probability of choosing a random individual of class i from the population $\left(\sum_{i=1}^k P_i = 1 \right)$

$$\Pr\{n_1, n_2, \dots, n_k\} = \frac{n!}{n_1! n_2! \dots n_k!} P_1^{n_1} P_2^{n_2} \dots P_k^{n_k}$$

Compare the case when we have a binomial random variable, e.g. heads and tails of a coin flip:

The possible outcomes of 5 flips are:

5 H, 0 T 4 H, 1 T 3 H, 2 T 2 H, 3 T 1 H, 4 T 0 H, 5 T

$$\Pr\{5 \text{ H, } 0 \text{ T}\} = \frac{5!}{5!0!} (0.5)^5 (0.5)^0 = (1)(0.03125)(1) = 0.03125$$

$$\Pr\{4 \text{ H, } 1 \text{ T}\} = \frac{5!}{4!1!} (0.5)^4 (0.5)^1 = (5)(0.0625)(0.5) = 0.15625$$

etc.

The factorial expression leading off the equation represents the number of different ways in which the result under consideration can arise: e.g. 5H, 0T can arise in only one way, *viz.* H-H-H-H-H, but 4H, 1T can arise in 5 ways, *viz.* T-H-H-H-H; H-T-H-H-H; H-H-T-H-H; H-H-H-T-H; H-H-H-H-T, and so on.

Of course, the sum of the probabilities for each of the six outcomes is one.

The only difference with our data is that we have a multinomial random variable—more classes are possible, and consequently, more combinations of data can be found.

Now, we know the probabilities (P_i) in terms of θ , and we have the n_i from our observed counts. What we really want to know is θ . Thus, we can write a *likelihood* equation that allows us to find the value of our unknown parameter, θ .

$$L(\theta) = \frac{n!}{n_1!n_2!\dots n_k!} P_1^{n_1} P_2^{n_2} \dots P_k^{n_k}$$

This expression is not easy to work with, so we usually work with its natural logarithm, $\ln L$, which is called the support function.

$$\ln L(\theta) = C + n_1 \ln P_1 + n_2 \ln P_2 + \dots + n_k \ln P_k$$

To find the maximum likelihood estimator (MLE) for this equation, we need to maximize it, which is easily done by differentiation and equating to zero. The derivatives are also called scores (S).

$$\frac{d \ln L}{d \theta} = S(\theta) = n_1 \frac{d(\ln P_1)}{d \theta} + n_2 \frac{d(\ln P_2)}{d \theta} + \dots + n_k \frac{d(\ln P_k)}{d \theta} = 0$$

Repulsion example:

Example: Vv (2 vs 6 row) and Ff (green vs chlorina) barley--Robertson, 1944

F₂ repulsion: a=753 b=292 c=351 d=19 n=1415
 F₂ coupling: a=1064 b=223 c=259 d=218 n=1764

	Gamete genotype or F ₂ phenotype				Total
	AB	Ab	aB	ab	
1. Frequency of gamete types	$\frac{\theta}{2}$	$\frac{1-\theta}{2}$	$\frac{1-\theta}{2}$	$\frac{\theta}{2}$	1
2. Exp. prop. of phenotypes in F ₂	$\frac{2+\theta^2}{4}$	$\frac{1-\theta^2}{4}$	$\frac{1-\theta^2}{4}$	$\frac{\theta^2}{4}$	1
3. Calculated zygotic series in F ₂	$\frac{n(2+\theta^2)}{4}$	$\frac{n(1-\theta^2)}{4}$	$\frac{n(1-\theta^2)}{4}$	$\frac{n(\theta^2)}{4}$	n
4. Observed numbers of progeny	753	292	351	19	1415

$$\ln L = C + 753 \ln \frac{2+\theta^2}{4} + (292 + 351) \ln \frac{1-\theta^2}{4} + 19 \ln \frac{\theta^2}{4}$$

differentiating:

$$\frac{d \ln L}{d \theta} = 753 \frac{2\theta}{2+\theta^2} - 643 \frac{2\theta}{1-\theta^2} + 19 \frac{2\theta}{\theta^2} = 0$$

This simplifies algebraically (you can do this in your spare time) to

$$(n_1 + n_2 + n_3 + n_4) \theta^4 - (n_1 - 2n_2 - 2n_3 - n_4) \theta^2 - 2n_4 = 0$$

$$1415 \theta^4 + 552 \theta^2 - 38 = 0$$

Recall basic algebra and the quadratic equation: $ax^2 + bx + c = 0$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

In this case, $x = \theta^2$ and:

$$\theta^2 = \frac{-552 \pm \sqrt{(552)^2 - 4(1415)(-38)}}{2(1415)} = 0.0597$$

$$\therefore \theta = \sqrt{0.0597} = 0.24434$$

i.e., there is 24.43% recombination between the two loci. Realize that you may get more than one solution—but only values in the interval $0 \leq \theta \leq 0.5$ make sense biologically!

Mathematics review:

1. $\frac{d(a * x^n)}{dx} = a * \frac{d(x^n)}{dx}$
2. $\frac{d(\ln(x + c))}{dx} = \frac{d(\ln u)}{du} * \frac{du}{dx}$, where $u = (x + c)$
3. $\frac{d(\ln x)}{dx} = \frac{1}{x}$
4. $\left(\frac{f}{g}\right)' = \frac{g * f' - f * g'}{g^2}$

Therefore, for the first term in our example above:

$$\begin{aligned} \frac{d753\ln\frac{2+\theta^2}{4}}{d\theta} &= 753 \frac{d\ln\left(\frac{2+\theta^2}{4}\right)}{du} \frac{d\left(\frac{2+\theta^2}{4}\right)}{d\theta} \\ &= 753 \left(\frac{4}{2+\theta^2}\right) \left(\frac{2\theta}{4}\right) = 753 \left(\frac{2\theta}{2+\theta^2}\right) \end{aligned}$$

Coupling example:

Substituting $(1-\theta)$ for (θ) in the likelihood expression before differentiating results in a first derivative that has $(1-\theta)$ substituted for (θ) and with each term multiplied by (-1) , to change the sign.

Now, the coupling data from before can also be entered directly into the simple formula above:

$$1764\theta^4 + 118\theta^2 - 2 * 218 = 0$$

To give $\theta^2 = 0.4648$ and $\theta = 0.6817$

However, since these data are in coupling and $\theta = AB + ab$, the recombination value is $1 - 0.6817$ or 0.3183 , i.e. 31.83%.

General form of the likelihood equation:

$$\ln L(\theta) = \sum_{i=1}^k n_i \ln P_i$$

That is, multiply the observed number of individuals in each class by the natural logarithm of the probability of observing a member of that class and sum over all the k classes of data.

To show that our estimate of θ is indeed the most likely estimate, we can enter the value back into our support equation:

$$\ln L = C + 753 \ln \frac{2 + \theta^2}{4} + (292 + 351) \ln \frac{1 - \theta^2}{4} + 19 \ln \frac{\theta^2}{4}$$

Substituting 0.2443 for θ from our F_2 repulsion example above:

$$\ln L(0.24) = -500 - 931 - 80 = -1511$$

Which appears to be rather unlikely. However, realize that *any* observed data will be unlikely due to the vast number of possible permutations of possible observations. The important point is how does this log likelihood compare to other possible values of θ ?

$$\ln L(0.20) = -507 - 918 - 87 = -1512$$

$$\ln L(0.30) = -489 - 952 - 72 = -1513$$

Our value of 0.24 is more likely than either 0.2 or 0.3, suggesting that it is indeed the MLE for θ .

Information and Variance of MLE:

The second derivative of the support equation--or the first derivative of the scores--multiplied by -1 is called the information. The inverse of the information value of a MLE is its variance:

$$I(\hat{\theta}) = -\left(\frac{d^2 \ln L(\theta)}{d\theta^2}\right) = \frac{1}{V(\hat{\theta})} \quad \text{Or} \quad I(\hat{\theta}) = -n \sum_{i=1}^k \left(n_i \frac{d^2 (\ln P_i)}{d\theta^2} \right)$$

$V(\hat{\theta})$ is the variance of the MLE of the recombination fraction,

$I(\hat{\theta})$ is the total amount of information present in the population,

n is the number of individuals in the population

n_i is the number of individuals in each segregation class,

P_i is the expected proportion in each segregation class.

The more information about the recombination fraction is available, the greater the precision, and the lower the variance of the estimate will be possible. Therefore, more individuals in the population will contribute to lower variance. In our example in repulsion (Allard 1956; Table 6 for 9:3:3:1):

$$\begin{aligned} \hat{\theta} &= 0.2443 \\ I(\hat{\theta}) &= n \frac{2(1+2\theta^2)}{(2+\theta^2)(1-\theta^2)} = n * i \\ V(\hat{\theta}) &= \frac{1}{I(\hat{\theta})} \\ SE(\hat{\theta}) &= \sqrt{V(\hat{\theta})} \end{aligned}$$

Where, i = the average amount of information per individual in the population

$$\begin{aligned} I(\hat{\theta}) &= 1415 \frac{2(1+2(0.2442)^2)}{(2+(0.2443)^2)(1-(0.2443)^2)} = 1635.74 \\ V(\hat{\theta}) &= \frac{1}{1635.74} = 6.114 \times 10^{-4} \\ SE(\hat{\theta}) &= \sqrt{6.114 \times 10^{-4}} = 0.0247 \text{ or } 2.47\% \end{aligned}$$

Thus, our estimate of recombination between these two loci (Cr and Ms) is $24.43 \pm 2.47\%$.

ALTERNATE APPROACHES FOR SOLVING MAXIMUM LIKELIHOOD EQUATIONS

The problem that is often faced when solving for a MLE is that the solution is not easily obtained—i.e., it doesn't factor, the quadratic equation cannot be used, etc. Therefore, a number of alternative methods can be used:

1. Grid search
 2. Newton-Raphson Iteration
 3. EM (Expectation-Maximization) Algorithm
- We'll discuss the first two in class. Weir (1996), Lynch and Walsh (1998) and Liu (1998) provide more discussion of these methods.

1. Grid search: Simplest to use, but could be time consuming
Enter numbers into the ML equation until you converge on the correct value (i.e. 0).
e.g. in our example:

$$1415\theta^4 + 552\theta^2 - 38 = 0$$

Try	$\theta = 0.25$	$5.53 + 34.50 - 38 = 2.03$
	$\theta = 0.20$	$2.26 + 22.08 - 38 = -13.66$
	$\theta = 0.24$	$5.53 + 31.79 - 38 = -0.675$
	$\theta = 0.2443$	$5.04 + 32.94 - 38 = -0.0152$
	$\theta = 0.245$	$5.10 + 33.13 - 38 = 0.232$
	etc.	

Should use a range of numbers and develop a plot to avoid a local maximum.

2. Newton-Raphson Iteration
 - a. The general idea is to make a choice for θ , modify that choice using the *score*, and then repeating the modifications until successive iterations differ by less than some specified threshold.
 - b. This method employs a Taylor series expansion of a function in a power series around a point:

$$f(x) \cong f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(x_0) + \dots$$

Only the first two components are of interest, since the others are quite small

- c. Thus, if $f(x) = 0$, we can solve the above equation for x

$$f(x) = 0 \cong f(x_0) + (x - x_0)f'(x_0)$$

$$0 = \frac{f(x_0)}{f'(x_0)} + (x - x_0)$$

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

- d. We can apply this concept for our score function $[S(\theta)] = 0$ at the MLE, and obtain an *improved estimate* of the MLE (k+1) by using an *initial estimate* (k), and then using the

improved estimate as the initial estimate in a second iteration to get a further improved estimate:

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + \frac{S(\hat{\theta})^{(k)}}{I(\hat{\theta})^{(k)}} = \hat{\theta}^{(k)} - \frac{\left[\frac{d \ln L(\hat{\theta})}{d(\hat{\theta})} \right]_{\theta=\hat{\theta}^{(k)}}}{\left[-\frac{d^2 \ln L(\hat{\theta})}{d\hat{\theta}^2} \right]_{\theta=\hat{\theta}^{(k)}}}$$

When $\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}$ becomes very small (below some threshold), we declare convergence and accept the value as the MLE for θ .

Quick summary:

Steps in linkage estimation, so far:

1. Identify the segregation classes possible and number of individuals in each.
2. Determine the expected proportions of individuals in each class in terms of θ .
3. Write a log likelihood equation (*support*) for the multinomial distribution of your data; each segregation class will represent a single term in the equation.
4. Differentiate each term in the support to give the *scores*, the sum of which is equated to zero.
5. Solve the equation for θ using quadratic equation, grid search, N-R Iteration, EM Algorithm, etc.

Linkage example:

Allard (1956) has developed formulae and tabulated solutions of various scores and amounts of information over a range of θ (he uses 'p') from 0.01 to 0.50. Use the tables as follows:

1. Tables 4-5. Identify the phenotypes/genotypes observed, and their expected frequencies to develop the MLE.
2. Table 6. Identify the type of data or cross with which you are working (note that he uses segregations *that would be observed under no linkage* for descriptive purposes) and the corresponding estimation equation, which represents the score (i.e., they are the first derivatives of the support) and the amount of information per individual.
3. Solve the estimation equation. Application of N-R iteration is actually quite easy:
 - a. Begin with some value of θ , and find that value along the top of Table 7.
 - b. For each term (score) in your equation, find the corresponding solution, which is then entered into the equation and multiplied by the observed number of individuals for that particular class. Then sum the individual terms to get an overall value for the equation.
 - c. Go to Table 8, find the appropriate type of data, and then get the *information* for the θ being evaluated.

- d. Divide “b” by “c”, and subtract the result from the chosen value of θ . This gives a revised value of θ , to start the whole procedure over again from “a”. Repeat until convergence is met.

Using his examples from Table 1:

D vs. d (determinate vs indeterminate growth); R vs. r (dark red vs red seed coat)
9 different estimates of “ θ ” (Table 2).

Take, for example, data set #4--the common repulsion case, 9:3:3:1 ratio:

$$\frac{d \ln L}{d \theta} = a \frac{2\theta}{2 + \theta^2} - (b + c) \frac{2\theta}{1 - \theta^2} + d \frac{2\theta}{\theta^2} = 0 \quad (\text{Table 6}).$$

We can substitute the numbers 293, 107, 119, and 35 for a, b, c, and d, respectively.

We can also find the scores at various levels of “ θ ” for each derivative in Table 7:

Steps in solving for θ : (based on Allard's paper)

Begin estimation at $\theta = 0.5$ (or any other number of your choice):

- 1) Insert the observed numbers and the scores into the maximum likelihood equation:

$$\begin{aligned} 293(0.444444444) + (107 + 119)(-1.33333333) + 35(4) &= 0 \\ 130.22 + (-301.33) + 140 &= 0 \\ -31.11 &\neq 0 \end{aligned}$$

Clearly, setting θ equal to 0.50 has not resulted in a maximization of the equation.

- 2) Calculate the amount of information in the data set using Table 8.
 $\theta = 0.5$, equation 6, repulsion

$$i = 1.778, n = 293 + 107 + 119 + 35 = 554$$

$$\text{Therefore, } I = 1.778 * 554 = 985.012$$

- 3) Apply Newton-Raphson Iteration:
The first derivative (i.e., the \sum of the scores) is divided by the second derivative (i.e., I) to estimate the correction needed in estimating θ .

$$\text{Thus, } \hat{\theta}^{(k+1)} = 0.50 + \frac{-3111}{985.012} = 0.47$$

Therefore, we next choose to evaluate this equation at $\theta = 0.47$

If we continued, we would find that the actual value of θ for this data set is 0.47 (Table 3).
In other cases, further refinement may be necessary.

4) We can calculate the standard error of this linkage value quite simply as:

$$s_{\theta} = \sqrt{\frac{1}{I(\theta)}} = \sqrt{\frac{1}{(1667)(554)}} = 0.033$$

5) For data set 4, we can say that the recombination value between R and D is $47.0 \pm 3.3\%$.

II PRODUCT METHOD OF ESTIMATING LINKAGE (Fisher and Balmukand, 1928)

Simplest, easy to use, and tables are available (Immer, 1930).

Basic premise: the ratio $\frac{ad}{bc}$ varies with the strength of linkage. From line 2:

$$\frac{ad}{bc} = \frac{\theta^2(2 + \theta^2)}{(1 - \theta^2)^2}$$

*For repulsion, ad/bc is calculated, but for coupling, $\frac{bc}{ad}$ is calculated

Standard error of the θ value determined by the product method:

$$SE(\theta) = \sqrt{\frac{(1 - \theta^2)(2 + \theta^2)}{2n(1 + 2\theta^2)}} = \frac{\sqrt{(1 - \theta^2)(2 + \theta^2)}}{\sqrt{n}} = \frac{f}{\sqrt{n}}, \text{ where } f \text{ has been tabulated by Immer (1930).}$$

Example: Vv (2 vs 6 row) and Ff (green vs chlorina) barley–Robertson, 1944

F2 repulsion:	a=753	b=292	c=351	d=19	n=1415
F2 coupling:	a=1064	b=223	c=259	d=218	n=1764

Ratio of products:

$$\text{Repulsion} = (753 \cdot 19) / (292 \cdot 351) = 0.1396$$

looking up tabled value of 0.1396 (and interpolating) gives us a $\theta = 24.473$

f value for 0.1396 based from the table by Immer (1960) is 0.6271

$$SE_{\theta} = \frac{0.6271}{\sqrt{1415}} = 0.01667 = 1.67\%$$

Therefore, $\theta = 24.47 \pm 1.67\%$

Coupling: ratio of products = 0.2490, $\theta = 31.776$.

f value for 0.2490 (coupling) based from the table by Immer (1960) is 0.3955

$$SE_{\theta} = \frac{0.3955}{\sqrt{1764}} = 0.0094 = 0.94\%$$

Therefore, $\theta = 31.78 \pm 0.94\%$.

Advantages of product method:

1. Easy with tables
2. If recessives have reduced viability, this method is disturbed the least (less than even maximum likelihood).
3. Mathematical efficiency equal to max likelihood.

Disadvantages of product method:

1. Missing classes result in one product = 0. With close linkage in repulsion, the “d” class may be missing or very low. If very low, one individual more or less makes a big difference in the ratio from which “ θ ” is determined. In these cases, use maximum likelihood for 3-class data.

IMPROVING ESTIMATION THROUGH MORE COMPLETE CLASSIFICATION

The maximum amount of information can only be extracted from a segregating population if all genotypes can be identified. With a 3:1/3:1 situation, only four of 10 possible genotypic classes are identified. Progeny tests are useful in completing the classification of genotypes. In other words, there is information on recombination *hidden* in any segregation class that is composed of a number of distinct genotypes, e.g. the AB phenotypic class is composed of the AABB, AaBB, AaBb, and AABb genotypes. By breaking each phenotype into its constituent genotypes, we can get more information on recombination and get a better estimate of linkage.

One way of getting all genotypes out is by growing F₃ lines to ascertain what the F₂ plant's genotype was. For example, if the F₂ plant has an AB phenotype, but its genotype is actually AaBb, we would expect to see some segregation at both loci in the F₃. If the F₂ were ABAB, we would not see any segregation in the F₃.

Thus, by growing progeny tests, we can write a complete maximum likelihood expression to include all, or at least more, than four classes. The expected proportions of all genotypes (repulsion):

	AA	Aa (2)	aa	Total
BB	e $\frac{\theta^2}{4}$	f $\frac{2\theta(1-\theta)}{4}$	l $\frac{(1-\theta)^2}{4}$	$\frac{1}{4}$
Bb (2)	g $\frac{2\theta(1-\theta)}{4}$	h and h' * $\frac{2\theta^2}{4}$ and $\frac{2(1-\theta)^2}{4}$	m $\frac{2\theta(1-\theta)}{4}$	$\frac{2}{4}$
bb	j $\frac{(1-\theta)^2}{4}$	k $\frac{2\theta(1-\theta)}{4}$	n $\frac{\theta^2}{4}$	$\frac{1}{4}$
Tot	1/4	1/2	1/4	1

*There are two types of heterozygotes: those derived from two nonrecombinant gametes and those from two recombinant gametes, i.e., $\frac{AB}{ab}$ and $\frac{Ab}{aB}$.

So, if we have 4 classes the ln-likelihood equation is:

$$\ln L = C + a \ln \frac{2 + \theta^2}{4} + (b + c) \ln \frac{1 - \theta^2}{4} + d \ln \frac{\theta^2}{4}$$

But for 9 classes:

$$\ln L = C + (e + n) \ln \frac{\theta^2}{4} + (f + g + k + m) \ln \frac{2\theta(1-\theta)}{4} + (j + l) \ln \frac{(1-\theta)^2}{4} + (h + h') \ln \frac{2\theta^2 + 2(1-\theta)^2}{4}$$

The second equation will provide a better estimate of the recombination value. Distinguishing between h and h' will provide an even better estimate of θ by giving complete classification of the progeny.

INFORMATION AND PLANNING EXPERIMENTS

The relative amount of information (i.e., i) available to extract from a given cross depends on two factors:

1. The number of classes recovered (i.e., the completeness of classification), and
2. The tightness of the linkage.

The key factor is recovering recombinant individuals.

For example, in an F_2 in repulsion, the most informative recoveries are the doubly recombinant AB/AB, AB/ab, and ab/ab individuals, because these represent recombinations from both male and female meioses. Note the first two, *viz.*, AB/AB and AB/ab, are obscured in a two dominant gene model within the A-B- class. The only unambiguously recombinant individuals are in the ab/ab class; however, we recover few of these individuals if the linkage is tight. Thus, a single individual contributes little information when two dominant loci are tightly linked, because we recover few ab/ab individuals unless the progeny size is large. Thus, estimates of linkage have large standard errors in this situation.

Information of various classifications: (see Allard, p. 240, 242)

With complete classification, the amount of information changes dramatically depending on θ , from >200 units per individual at $\theta = 0.01$ to about 8 at $\theta = 0.50$ (see Allard's Table 8).

Backcrosses with all four classes recognizable have $\frac{1}{2}$ the information as a completely classified F_2 , because only one gamete's meioses are observed (the other parent's being obscured because it is homozygous). However, backcrosses are useful in estimating male and female recombination independently.

If the number of classes recovered is less than complete classification, the amount of information per individual also declines.

F_2 populations with two dominant loci linked in repulsion:

As we know, in this situation, the likelihood equation reduces to a quadratic equation when solving for the MLE of θ . If the observed numbers in the doubly recessive class, aabb, is zero, then the solution of the maximum likelihood equation is either 0 or outside the parameter space (i.e., outside 0 to 0.50). So regardless of the numbers in the other classes, if the doubly recessive class is 0, the linkage value will be 0. However, as θ moves from 0 and if more progeny are evaluated, this problem dissipates. Nevertheless, the ability to estimate linkage values with low standard errors in these cases is limited. See Allard (1956) Table 1, Data Set 5 and compare the recombination estimate in Table 3.

Progeny sizes needed to estimate θ at a given level of accuracy:

If you have to work with a given progeny population, *what is the progeny size needed* to attain a certain degree of accuracy?

$$n = \frac{1}{iV(\hat{\theta})}$$

For example, if both loci have complete dominance and are in repulsion, θ is estimated to be 0.05, and a standard error not larger than 5% is desired, then

$n = [(1.006)(0.05)^2]^{-1} = 398$ progeny must be evaluated, because few recombinants (ab/ab) can be unambiguously seen.

Compare a coupling cross—only 21 individuals need to be evaluated to attain the same SE! In this case, recombinants (A-/bb and aa/B-) can be easily seen (none are present with complete linkage); each class can result from a single recombinant gamete or from two recombinant gametes—thus the chances of one occurring are high.

Compare a repulsion cross segregating 9:7--133,334 individuals need to be evaluated!

Progeny testing vs. growing a larger F₂

$\frac{i_{\theta} F_3}{i_{\theta} F_2} >$ the number of F₃ progeny needed to grow to ensure recovery of recessives

The bottom line with dominant markers:

F₃ are better than more F₂ if $\theta < 0.11$ and 16 plants are grown of each F₃
(99% sure of correct classification)

i.e., F₃ families of 16 plants provide more than 16 times the information of a single F₂ plant
or better than F₂ if $\theta < 0.08$ if 99.9% accuracy of determination is desired
(24 plants/F₃).

DETERMINING RECOMBINATION VALUES FROM PARTIAL RATIOS

In addition to complete sets of data, such as the previous examples, we can also calculate recombination values among segregate classes that become apparent after progeny testing, or in cases where one or more of the genotypes are lethal.

Linkage estimation with lethal alleles:

Assume an F₂ population with segregation of two genes, A and B: uppercase alleles inherited from one parent; lowercase from the other (i.e., coupling).

(i) Single gene defect

Suppose that the homozygous class “aa” die before genotyping can occur. Analyzing linkage in this situation requires recalculation of the expected progeny ratios and redesigning a likelihood equation.

- (1) The way to figure this out is to figure the expectations without any lethality.
- (2) Then adjust the expectations based on the total number of survivors you can measure or score.

In this case, $\frac{1}{4}$ of the plants will die, so all the expectations of the remaining classes need to be divided by $\frac{3}{4}$. That is, each of these classes will now be expected in some frequency out of the $\frac{3}{4}$ progeny that are viable.

For example: normally, we would expect AABB to be present in $\frac{(1-\theta)^2}{4}$, but since we only see

75% of the normally expected progeny, our new expectation of AABB is $\frac{\frac{(1-\theta)^2}{4}}{\frac{3}{4}} = \frac{(1-\theta)^2}{4} * \frac{4}{3} = \frac{(1-\theta)^2}{3}$

From here, it is a simple matter to develop the log likelihood equation:

$$\ln L(\theta) = n_{AABB} \ln\left(\frac{(1-\theta)^2}{3}\right) + n_{AaBB} \ln\left(\frac{2\theta(1-\theta)}{3}\right) + \text{etc.}$$

For a single gene defect, we can figure the proportion of individuals not surviving (or not able to be scored) according to our expectations for that locus alone—that is, linkage is not involved in our estimation of the proportion of surviving plants. For a single gene with “aa” lethal, $\frac{1}{4}$ of the plants will be missing.

(ii) Two gene defect

In the case of a two-gene defect, the number missing will be a proportion that depends on the recombination fraction between the loci. Therefore, we cannot make a simple statement like the single gene model.

So, if the aabb class dies, and we can classify 8 total classes in an F_2 , then:

Since the aabb class would be expected to be present (in coupling) at: $\frac{(1-\theta)^2}{4}$

We would expect all other classes to be present at: $1 - \frac{(1-\theta)^2}{4}$

Therefore, to get the expected frequency of all other classes, we need to divide the expectations derived in the normal manner (i.e. without any lethal classes) by the expectations of those we observe:

$$\text{e.g. for AABB: } \frac{\frac{(1-\theta)^2}{4}}{1 - \frac{(1-\theta)^2}{4}}$$

and so on.

The likelihood equation is then developed from these expectations.

SEGREGATION DISTORTION AND LINKAGE ESTIMATION

Segregation distortion at individual loci, due to gametic or zygotic selection, is prevalent in many organisms. In rice and some other crops, known genes (gametophyte–ga; sterility–S) cause distorted ratios; in other crops the cause is unknown, though may be related to inbreeding depression (e.g., in alfalfa). Interspecific hybridization often causes segregation distortion in the progeny.

Computing recombination fractions between loci with segregation distortion is more complicated than we have discussed. Generally, maximum likelihood will not give the best estimate when computed under the assumption of no segregations distortion, as we have been doing. If the correct expectations can be specified—that is, if we know the actual expectations given some disturbance—then ML provides an unbiased estimate.

The product formula can be used in many cases of disturbed segregation, if two dominant genes are being considered (Mather 1957: Page 94-95 Chapter VIII).

These papers provide general formulae to allow successful application of maximum likelihood to cases of segregation distortion:

Lorieux et al. (1995b) for F_2 populations

Lorieux et al. (1995a) for backcrosses

COMBINING ESTIMATES OF θ

All we do in determining a value of θ for a single experiment is sum the scores from each individual phenotypic (or genotypic) class to get an overall score for the likelihood equation.

Analogously, we can pool a number of estimates of θ by simply adding the likelihood equations and solving via N-R iteration or other method.

$$L(\theta)_{pooled} = L(\theta)_{pop 1} + L(\theta)_{pop 2} + \dots + L(\theta)_{pop n}$$

Allard presents a more complete analysis in Table 2. He has 9 different sets of data from which he wants to determine the overall value for “ θ ”. By pooling all his data, he will get an average estimate of the recombination value across genetic backgrounds and environments.

- 1) Calculate the score and amount of information for each data set at $\theta = 0.5$.
- 2) Sum the scores and the information across populations to produce an overall total.
- 3) Calculate the deviation from 0.50 which needs to be applied.
- 4) Repeat at the new θ value.

In this case, he found after (2): $\theta = 0.50 - 0.11 = 0.39$. Thus, he reevaluated at 0.39. This gave a deviation of +0.01, so he reevaluated at $\theta = 0.40$ and found a deviation of only +0.0025. Thus, the combined θ is 40.25%.

Heterogeneity of linkage estimates (Allard, 1956)

Fisher (1949) showed that the sum of the squares of the deviations from zero of the logarithm of the maximum likelihood expression for each body of data, divided by total amount of information provided by that body of data, is in a χ^2 distribution.

$$\chi^2 = \sum_1^N \frac{S^2}{I}$$

where S are the scores (D in Allard), I is the information, and N are the total number of data sets. S in this regard is the deviation from zero of the maximized log likelihood equation—, the squared deviations are used to determine the χ^2 .

In Allard’s example (Table 2 and p. 238):

The total χ^2 , with N = 9 df, is calculated as the total of the χ^2 values for each individual estimation equation; these deviations are due to heterogeneity as well as to experimental error.

The pooled χ^2 , with 1 df, can be estimated as the squared deviation for the overall analysis (i.e. the sum of the scores evaluated at the MLE). This value indicates the extent to which θ was not calculated perfectly.

The heterogeneity χ^2 , with N-1 df, can be calculated as the difference between the total χ^2 and the pooled χ^2 .

For this example, we reject H_0 that all nine data sets are homogeneous, i.e. that they all provided the same estimate of θ . Three sets (4, 5, and 6) appear to be the cause of the deviation. This could be due to environmental and/or genetic background effects. Therefore, we can say that $\theta = 0.40$ as an average, but that deviations from this are possible in some situations, depending on parents used in the cross or on the environment tested. Estimate 5, in particular, is highly skewed (0%)—this is due to “0” progeny in the n_4 (d) class (see Table 1).

Alternative approach:

Log likelihood ratio test statistic approach: (Liu, 1998). This method uses the log likelihood equation to calculate a statistic, G , that evaluates the goodness of fit of the estimated value of θ relative to the fit of $\theta = 0.50$. G is distributed as a χ^2 with 1 df (because each test is between an estimated value of θ and a fixed value of 0.50).

Assume values for θ have been determined in p populations, denoted 1, 2, ..., i :

Then,

$$G_i = 2 \ln \left[\frac{L(\hat{\theta}_i)}{L(0.50)} \right] = 2 [\ln L(\hat{\theta}_i) - \ln L(0.50)]$$

The total log likelihood ratio test is simply:

$$G_{Total} = \sum_{i=1}^p G_i$$

The pooled log likelihood ratio test:

$$G_{Pooled} = 2 [\ln L(\hat{\theta}_{pooled}) - \ln L(0.50)]$$

The heterogeneity log likelihood ratio test:

$$G_{Heterogeneity} = G_{Total} - G_{Pooled}$$

The Lod Score—Defining the strength of linkage

In the examples above, we tested the statistical significance of linkage between two loci using the G statistic or a χ^2 value. These give a probability value for θ . Another way to assess the likelihood of linkage between two loci is the “lod score.”

For a given set of observed data, we can construct a likelihood equation $L(\theta)$, from which the recombination value can be determined. The ratio of the likelihood that two loci are linked at some value θ to the likelihood that the loci are unlinked (i.e., $\theta = 0.50$) provides a measure of the strength of the evidence for linkage.

$$Z(\theta) = \log_{10} \left[\frac{L(\theta)}{L(0.50)} \right] = \log_{10} L(\theta) - \log_{10} L(0.50)$$

$Z(\theta)$ represents the lod (or lod score), defined as the logarithm (base 10) of the odds ratio (loose use of “odds”, because it usually refers to probabilities, not likelihoods).

If Z is positive, then we have tentative evidence for linkage.

If Z is negative, then we have no evidence for linkage.

$Z(\hat{\theta}) = Z_{\max} = \hat{Z}$ represents the maximum lod score; i.e., Z is maximized at the MLE of θ .

Significance of linkage estimate based on a lod score

Linkage is often determined to be “significant” if the lod exceeds 3. A lod of 3 means that the likelihood that the loci are 1000 times more likely to be linked than unlinked ($\log 1000 = 3$), given the available data.

Lod scores less than 3, but greater than 0, indicate that there is potentially linkage, but that the strength of the evidence is not strong enough to be considered “significant” in most cases. However, the threshold of 3 is somewhat arbitrary, so the researcher needs to take his requirements into account when specifying a threshold.

Support interval for Z_{\max}

A support interval for Z_{\max} can be constructed as $Z_{\max} - 1$, given that Z is significant (i.e. lod3). This gives the range of possible values of θ that could also possibly explain the data.

Relating Z to G

The lod score, Z , is related to the log likelihood test statistic, G :

$$Z = 0.2172 G$$

Z has been commonly used in genomics, but since G is distributed as a χ^2 it may be more easily understood by those familiar with biological χ^2 tests. However, interpretation of Z and G may differ in some instances. Note that a LOD of 3 corresponds to a χ^2 probability level of approximately 0.0002 (see Liu, 1998, Table 4.3, p. 100). The reason that such a stringent test is made in accepting a LOD of 3 is that genome wide, when making tests of linkage among many loci, we need to control the probability of type I errors through use of a more restrictive probability level. Thus, the lod value chosen should account for the number of markers being tested.