

# Linkage Mapping

CS741  
2009  
Jim Holland

# Morgan's fly experiment

$Pr^+ Pr^+ Vg^+ Vg^+$  (red eyes, normal wings)  
×  $prprvgv$  (purple eyes, vestigial wings)  
↓  
 $Pr^+ prVg^+ vg$  (F1) ×  $prprvgv$

1339	$Pr^+ prVg^+ vg$
151	$Pr^+ prv$
151	$prprVg^+ vg$
1195	$prprv$

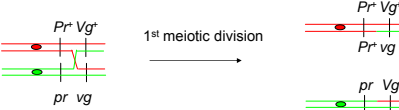
A heterozygous individual has two different homologous chromosomes:



DNA replication results in two sister chromatids per chromosome:



Homologous chromosomes pair at meiosis and can exchange chromatid strand pieces during crossing-over:



Following a cross-over between two genes, the resulting gametes are an equal mixture of parental and recombinant types:



# Crossing Over and Recombination

- Crossovers at meiosis are the CAUSE
- Recombinations observed between loci on the same chromosome are the RESULT
- Crossing-over and recombination are not related one to one.

# Genetic Recombination

- Recombination frequency = % of recombinant gametes produced by double heterozygotes:  
 $r = (151 + 154) / (1339 + 151 + 154 + 1195) = 0.107$
- Genes that are closer together on a chromosome *tend* to have lower recombination frequency, but there is not a direct relationship between PHYSICAL and GENETIC distances
- Why not?

# Genetic recombination

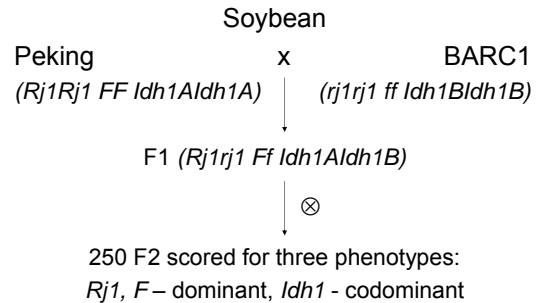
- What is the maximum possible value for recombination frequency?
- What is the recombination frequency between genes on different chromosomes?
- What proportion of recombinant gametes occur as the product of two crossovers between the same genes in the same meiosis? (more on this later...)

## Linkage Mapping

1. Detect linkage (for pairs) – identify linkage groups
2. Estimate recombination frequency (for pairs)
3. Order loci within linkage groups
4. Re-estimate pairwise linkage distances using “multi-point” (multi-gene) techniques

## Linkage Mapping Example

(Hedges et al., 1990)



## Detecting Linkage

- How to distinguish linkage from sampling error due to random chance?
- Chi-square test ( $\chi^2$ ) tests if **observed** deviations from **expected** segregation for unlinked genes would occur by chance less than 5% of time if genes are not linked.
- **Null hypothesis: genes are not linked.** We determine **expected** segregation ratios based on null hypothesis. We reject null hypothesis if **observed** ratios would occur less than 5% of the time if genes were unlinked.

## Detecting Linkage Example

Observed phenotypic segregation in an F2 population developed from the soybean cross Peking (*FFRj1Rj1*)  $\times$  BARC-1 (*ffrj1rj1*).

Phenotypic class	Number Observed
<i>F_Rj1_</i>	144
<i>F_rj1rj1</i>	44
<i>ffRj1-</i>	39
<i>ffrj1rj1</i>	23
Total	250

## First check that each locus segregates Mendelianly

Chi-square test for single-gene segregation of *Rj1* phenotype

Phenotypic class	Number observed	Number expected	(E-O) <sup>2</sup> /E
<i>Rj1_</i>	183	187.5	0.108
<i>rj1rj1</i>	67	62.5	0.324
Total	250	250.0	0.432

Degrees of freedom? Critical value of test at 5% probability level is 3.84.

We do NOT reject null hypothesis, assume that phenotype segregates as single gene.

## Test for linkage between *F* and *Rj1*

Chi-square test for linkage of *F* and *Rj1* genes.

Phenotypic class	Number observed	Number expected	(E-O) <sup>2</sup> /E
<i>F_Rj1_</i>	144	140.625	0.081
<i>F_rj1rj1</i>	44	46.875	1.458
<i>ffRj1-</i>	39	46.875	1.323
<i>ffrj1rj1</i>	23	15.625	3.481
Total	250	250.0	6.343

How many degrees of freedom?

Note that this test will be affected by any segregation distortion at the two genes AND by linkage.

Get the statistic for testing only linkage by subtracting the two single gene segregation chi-square values:  
 6.343 - 0.432 = 5.906

Get the df by total df minus two df for the single gene tests = 1 df.

We reject the null hypothesis and conclude genes are linked.



## Method of Maximum Likelihood

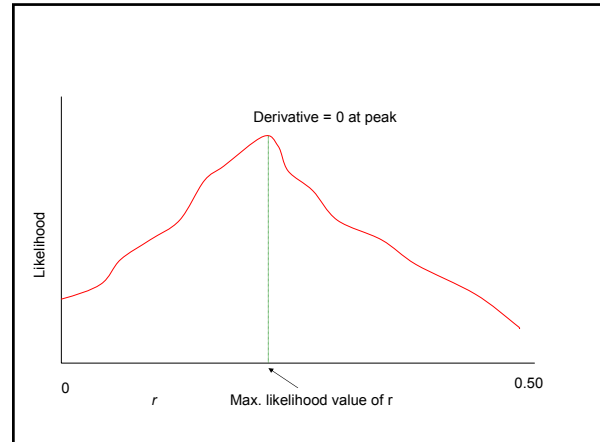
- Given the likelihood function:

$$L = \frac{n!}{a!b!c!d!} \left(\frac{3}{4} - \frac{r}{2} + \frac{r^2}{4}\right)^a \left(\frac{r}{2} - \frac{r^2}{4}\right)^b \left(\frac{r}{2} - \frac{r^2}{4}\right)^c \left(\frac{1}{4} - \frac{r}{2} + \frac{r^2}{4}\right)^d$$

The maximum likelihood estimate of  $r$  is that value of  $r$  that maximizes the likelihood of the observed data. Notice that  $a$ ,  $b$ ,  $c$ , and  $d$  above are all observed data, and  $r$  is the only unknown in the equation.

We could just plug in many different values for  $r$  and choose the one that gives maximum  $L$  (a numerical solution).

Or we could rely on a trick from calculus: the derivative of a function is equal to its slope at a specific position on the curve. Where the derivative (slope) is zero, the function must be at a maximum or minimum point.



## Solving for the maximum likelihood estimate

- General idea: the likelihood function is a curve relating the likelihood of observed data to the recombination frequency.
- The derivative of the likelihood function is its slope.
- The value of  $r$  at which the slope is zero must be a maximum or minimum likelihood point (in fact, it will be maximum if  $0 \leq r \leq 0.5$ )
- So, take the derivative of the likelihood function with respect to  $r$  and set it equal to zero, solve for  $r$ .

## Solving for max. likelihood estimate of $r$ :

- Additional tricks used to make math simpler:
- Replace  $(1 - r)^2$  with  $P$  and solve for max. likelihood estimate of  $P$ .
- Take derivative of natural log of likelihood function (calculus tells us that max. point of a function will also be the same as the max. point of the natural log of that function)

$$L = \frac{n!}{a!b!c!d!} \left(\frac{3}{4} - \frac{r}{2} + \frac{r^2}{4}\right)^a \left(\frac{r}{2} - \frac{r^2}{4}\right)^b \left(\frac{r}{2} - \frac{r^2}{4}\right)^c \left(\frac{1}{4} - \frac{r}{2} + \frac{r^2}{4}\right)^d$$

$$\ln(L) = \ln\left(\frac{n!}{a!b!c!d!}\right) + a \ln\left(\frac{1}{4}(2+P)\right) + b \ln\left(\frac{1}{4}(1-P)\right) + c \ln\left(\frac{1}{4}(1-P)\right) + d \ln\left(\frac{1}{4}P\right)$$

$$\frac{\partial \ln(L)}{\partial P} = \frac{a}{2+P} + \frac{b}{1-P} + \frac{c}{1-P} + \frac{d}{1-P} = 0$$

$$a(1-P)P - (2+P)P(b+c) + (2+P)(1-P)d = 0$$

$$a(P-P^2) - (2P+P^2)(b+c) + (2-P-P^2)d = 0$$

$$2d + (a-2b-2c-d)P - (a+b+c+d)P^2 = 0$$

Now, we fill in the numbers for  $a$ ,  $b$ ,  $c$ , and  $d$  from Hedges et al. (1990):  
 $46 + P(-45) + P^2(-250) = 0$

This has the form of a quadratic equation,  $ax^2 + bx + c = 0$ , which has roots:

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

Using the quadratic equation, we solve for  $P = -0.52, 0.348$ .  
 $P = (1-r)^2$ , so  $(1-r) = \text{square root of } P$ , which means that only 0.348 is a real solution for  $P$ .  
 $(1-r) = 0.590$   
 $r = 0.410$

### What if you can't solve the equation?

- Actually, this is common. So, we resort to a numerical solution by plugging in values for  $r$  between 0 and 0.5 and choosing the value of  $r$  that makes derivative of log of likelihood closest to zero.
- Why bother with this derivative business?
  - Often you cannot numerically solve original equation because powers are too huge
  - You can also get variance of estimate using the 2<sup>nd</sup> derivative of the log likelihood function (see Mather, 1951).