

## QTL Mapping 2

CS741  
2009  
Jim Holland

## Linked Markers Share Information

- The combined effect of the significant markers on linkage group 2 is not the sum of the individual effects, it is much less.
- The linked markers represent much of the same information.
- Typically, we select the single most important marker in a region to represent that region's effect.

## Unlinked markers may share information!

- In a mapping population of typical size (less than 500 lines), the effects of markers even on different chromosomes are not completely independent, simply due to sampling effects.
- Therefore, you cannot accurately estimate the combined effects of QTLs by summing up the independent (one-at-a-time) estimates.
- The combined effects of multiple markers is typically less than the sum of the independently estimated effects.

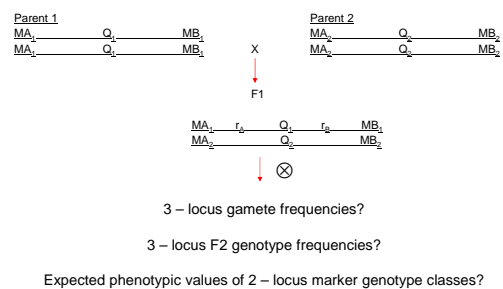
## Multiple Marker Models

- Different combinations of markers represent significant QTL regions can be tested.
- The combination in which all markers are significant that maximizes the model  $R^2$  value can be selected as the "best".
- If you find several distinct models with similar  $R^2$  values but with different subsets of loci, this suggests that your sample size is not sufficiently large to accurately estimate all of these locus effects simultaneously.

## Interval Mapping

- Single marker ANOVA underestimates the true effect of a QTL unless there is zero recombination between marker and QTL.
- Interval mapping provides a method to test the effects of positions in intervals between markers.
- If you can test a position very near the true QTL position, you will have higher power to detect the QTL.
- You do not know the genotypes at positions between markers, but based on the linkage map distances and the flanking marker genotypes, you can get the probability of genotypes at that position.

## Interval Mapping Model



### 3 – locus gamete frequencies

- $F(A_1Q_1B_1) = (1/2)(1-r_A)(1-r_B)$
- $F(A_1Q_2B_1) = (1/2)(r_Ar_B)$
- etc... (8 total gamete types)

### 3 – locus genotype frequencies

- $F(A_1A_1Q_1Q_1B_1B_1) = (1/4)(1-r_A)^2(1-r_B)^2$
- $F(A_1A_1Q_1Q_2B_1B_1) = (1/2)(1-r_A)(1-r_B)(r_Ar_B)$
- $F(A_1A_1Q_2Q_2B_1B_1) = (1/4)(r_Ar_B)^2$
- etc...(27 total genotype classes)

### Expected Value of 2-locus Marker Classes:

Genotype	Frequency	Value
$A_1A_1Q_1Q_1B_1B_1$	$(1/4)(1-r_A)^2(1-r_B)^2$	$m + a$
$A_1A_1Q_1Q_2B_1B_1$	$(1/4)(1-r_A)(1-r_B)(r_Ar_B)$	$m + d$
$A_1A_1Q_2Q_2B_1B_1$	$(1/4)(r_Ar_B)^2$	$m - a$

Expected value of  $A_1A_1B_1B_1$  is weighted mean:  
 $E(A_1A_1B_1B_1) = m + a[(1-r_A)^2(1-r_B)^2 - r_A^2r_B^2]/(1-r)^2 + d[2(1-r_A)(1-r_B)r_Ar_B]/(1-r)^2$

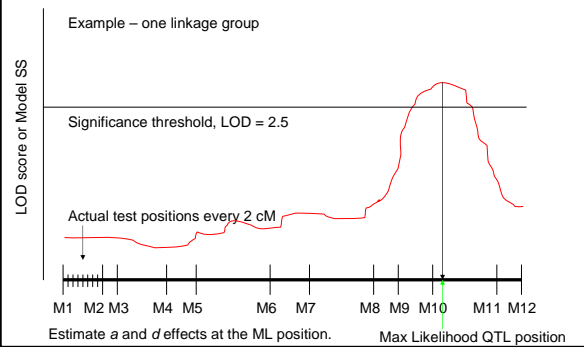
- Do same for 7 other genotypic classes →

### Expected Values of $F_2$ Marker Classes

Marker genotype	Coefficients of Expected Genotypic Value	
	$a$ (additive genetic effect)	$d$ (dominance genetic effect)
$A_1A_1B_1B_1$	$[(1-r_A)^2(1-r_B)^2 - r_A^2r_B^2](1-r)^2$	$[2r_A(1-r_A)r_B(1-r_B)](1-r)^2$
$A_1A_1B_1B_2$	$[(1-r_A)^2r_B(1-r_B) - r_A^2r_B(1-r_B)]r(1-r)$	$[r_A(1-r_A)(1-r_B)^2 + r_A(1-r_A)r_B^2]r(1-r)$
$A_1A_1B_2B_2$	$[(1-r_A)^2r_B^2 - r_A^2(1-r_B)^2]r^2$	$[2r_A(1-r_A)r_Br(1-r_B)]r^2$
$A_1A_2B_1B_1$	$[r_A(1-r_A)(1-r_B)^2 - r_A(1-r_A)r_B^2]r(1-r)$	$[(1-r_A)^2r_B(1-r_B) - r_A^2r_B(1-r_B)]r(1-r)$
$A_1A_2B_1B_2$	0	$[r_A^2r_B^2r(1-r_B)^2 + (1-r_A)^2r_B^2r(1-r_B)^2]r^2$
$A_1A_2B_2B_2$	$[r_A(1-r_A)r_B^2r(1-r_B) - r_A(1-r_A)r_B^2r(1-r)]r(1-r)$	$[(1-r_A)^2r_B(1-r_B) + r_A^2r_B(1-r_B)]r(1-r)$
$A_2A_1B_1B_1$	$[r_A^2(1-r_B)^2 - (1-r_A)^2r_B^2]r^2$	$[2r_A(1-r_A)r_Br(1-r_B)]r^2$
$A_2A_1B_1B_2$	$[r_A^2r_B(1-r_B) - (1-r_A)^2r_B(1-r_B)]r(1-r)$	$[r_A(1-r_A)(1-r_B)^2 + r_A(1-r_A)r_B^2]r(1-r)$
$A_2A_1B_2B_2$	$[r_A^2r_B^2 - (1-r_A)^2r_B^2]r^2$	$[2r_A(1-r_A)r_Br(1-r_B)]r^2$

To test the effects of a position within the interval, select  $r_A$ , then  $r_B = (r - r_A)/(1 - 2r_A)$ . This table becomes two columns of coefficients for  $a$  and  $d$ , and regression or max. likelihood analysis is used to estimate best fits for  $a$  and  $d$  to observed data for the marker class means.

### Interval Mapping Tests Positions Every 1 – 2 cM Through Genome

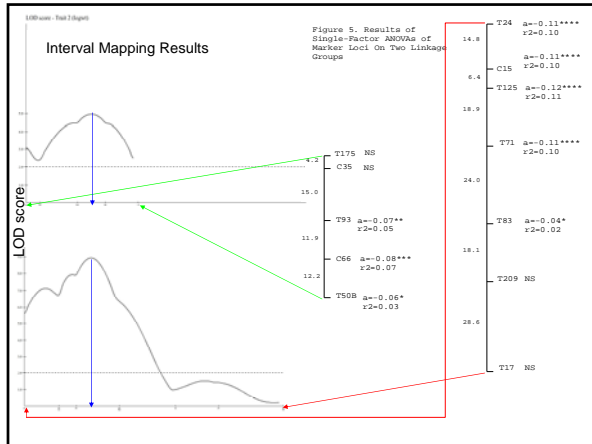


### Interval Mapping QTL Estimates

- By selecting the most likely position of the QTL, you can estimate the QTL effects directly at the position of the QTL.
- This eliminates the bias that occurs with single marker analysis due to recombination between marker and QTL positions.

### Interval Mapping Example

- Analyze the same data set of 333 tomato  $F_2$  plants using Mapmaker/QTL
- This requires the linkage map to be made first.
- Then linkage map is scanned every 2 cM for QTL.
- Results displayed as LOD scores for each position.
- $LOD = \log_{10}[\text{likelihood of model including QTL effect} / \text{likelihood of model with no QTL effect}]$



### Interval Mapping Results

POS	WEIGHT	DOM	%VAR	LOG-LIKE	Interval	Marker
0.0	-0.102	-0.007	9.0%	5.645	4-11 14.8 cM	T24
2.0	-0.110	-0.008	10.4%	6.159		
4.0	-0.116	-0.008	11.4%	6.584		
6.0	-0.119	-0.007	12.1%	6.897		
8.0	-0.120	-0.006	12.3%	7.083		
10.0	-0.120	-0.005	12.1%	7.135		
12.0	-0.117	-0.006	11.4%	7.054		
14.0	-0.111	-0.009	10.4%	6.853		
0.0	-0.109	-0.010	9.9%	6.752	11-8 6.4 cM	C15
2.0	-0.118	-0.012	11.4%	7.418		
4.0	-0.122	-0.014	12.0%	7.802		
6.0	-0.122	-0.016	11.8%	7.932		
0.0	-0.121	-0.016	11.7%	7.931	8-12 18.9 cM	T125
2.0	-0.130	-0.014	13.6%	8.409		
4.0	-0.136	-0.011	15.1%	8.753		
6.0	-0.140	-0.009	16.0%	8.926		
8.0	-0.140	-0.009	16.3%	8.914		
10.0	-0.138	-0.010	16.0%	8.723		
12.0	-0.134	-0.013	15.2%	8.369		
14.0	-0.128	-0.016	13.9%	7.880		
16.0	-0.119	-0.020	12.2%	7.292		
18.0	-0.109	-0.022	10.3%	6.647	12-9 24.0 cM	T71

Part of linkage group 2 results

Compare to single marker analysis at T125: a = -0.12, r<sup>2</sup> = 11%

Max. Likelihood Estimates

- ### Composite Interval Mapping
- First do single marker ANOVA, then build best fitting multiple marker model using model selection techniques.
  - Then scan the genome using interval mapping to identify QTL *after* accounting for marker effects that are unlinked to test position.
  - By fitting unlinked QTL in the model, the residual variation due to other QTL is reduced, increasing the power to detect QTL.

- ### Multiple Interval Mapping
- Build multiple QTL models, fitting all QTLs at their maximum likelihood positions.
  - Permits simultaneous estimation of QTL effects while also using the power/precision of interval mapping.
  - All the same problems of model selection occur with MIM. The best way to obtain a robust model is to use a large population size.

- ### QTL Mapping Results
- Positions of QTL are hard to estimate precisely. Confidence intervals often include 10 – 20 cM.
  - Traits with low heritability require large population sizes and extensive replication to obtain accurate QTL position/effect estimates.
  - Better statistical methods help but do not solve the problem.

- ### QTL Estimation Problems
- When mapping in small populations (less than ~500 lines), QTL with small effects are often missed.
  - Those QTL that are identified have overestimated effects (because they "absorb" some of the information from the undetected QTL).
  - Thus, QTL estimates from one population sample often poorly predict their effects in an independent sample of the same population.
  - Typical QTL mapping studies are probably robust only for QTL with effects of 10% or more.
  - QTLs for the same trait can vary dramatically across mapping populations!

## Association Analysis

- Instead of making new populations from crosses between divergent lines, can we identify QTL in already existing germplasm collections, breeding lines, or natural populations?
- Maybe, but first we need to account for population structure.
- Why?

## Gene – Phenotype Associations in General Populations

- Statistical association between a gene and phenotypic variation occurs if:
  - gene actually affects phenotype, or
  - tested gene is in gametic phase disequilibrium with the causal gene(s).

We might be happy to detect genes that are linked to QTL, but the problem is that gametic (“linkage”) disequilibrium in many populations does not imply linkage!

## Gametic Phase Disequilibrium (aka: “Linkage” Disequilibrium)

- Nonrandom association of alleles at different loci
- Measured as:  $D_{ab} = p_{ab} - p_a p_b$

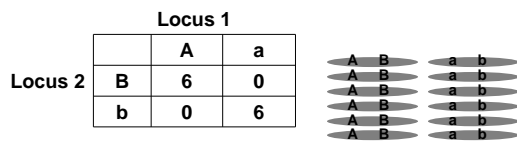
### INCREASED/MAINTAINED BY:

Population subdivision  
Recent population hybridization  
Mutation  
Physical linkage  
Selection on epistatically interacting loci

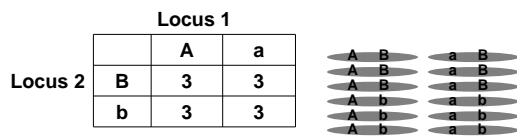
### DECREASED BY:

Recombination  
Independent assortment

So, linkage tends to maintain disequilibrium, but tightly linked genes can be in equilibrium, and conversely, unlinked genes can be in disequilibrium.



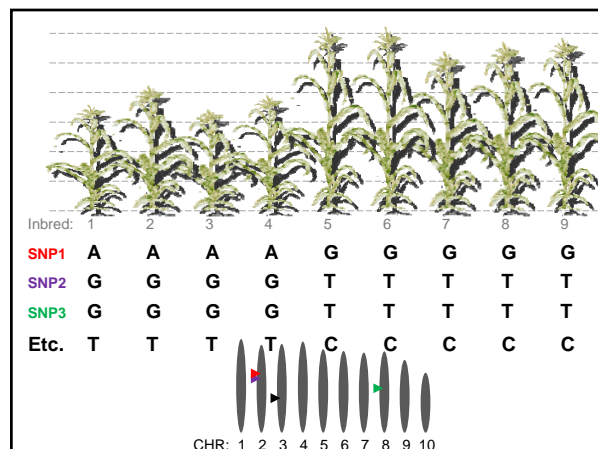
$$D_{ab} = 0.5 - 0.5 \cdot 0.5 = 0.25$$



$$D_{ab} = 0.25 - 0.50 \cdot 0.50 = 0$$

## Population Structure: Typical mapping populations vs. germplasm collections

- In QTL mapping populations, LD *only* occurs between physically linked genes...there is no population structure - why?
- In general populations, we need to first estimate population structure, then account for that in the association analysis.



## Controlling Population Structure in Association Analysis

- Use a set of random markers distributed across the genome to determine relationships between individuals/subpopulation structure.
- Ex: 260 maize inbred lines from around the world fingerprinted with 94 SSR markers revealed three major groups: Stiff Stalk, non-Stiff Stalk temperate, and Tropical. Correspond to heterotic groups recognized by maize breeders.
- Each line can be assigned a probability that it belongs to each of the 3 subpopulations: "Q matrix"
- Or you can estimate pairwise genetic similarities among lines.

## Association Mapping Model

$$y = X\beta + S\alpha + Qv + Zu + e$$

Trait values  $\uparrow$   $X\beta$   $\uparrow$  Environments, etc.  $\uparrow$  Candidate Gene effects  $\uparrow$  Subpopulation effects  $\uparrow$  Background Genetic effects,  $\text{Var}(u) = KVg$

## If candidate gene has a significant effect:

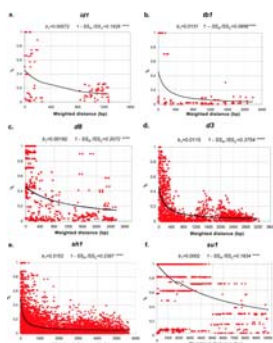
- Is it due to causality or due to linkage to causal genes?
- Each gene region/population/species combination must be studied carefully to determine the extent of linkage disequilibrium.
- In diverse maize populations, LD tends to be reduced over short distances (<1 kb), but in highly selected elite lines, it can extend ~100 kb.
- What do you expect for wheat?

## Extent of LD determines resolution of association analysis

- If LD is extensive, then the detected effect may be due to linkage with the causal gene. So, you are not sure if tested gene is causal gene.
- Higher LD causes lower resolution
- But it also means you can scan with random markers and localize QTL to regions.
- Lower LD increases resolution.
- But you may have to have causal gene to detect the effect. Without good candidate genes, association analysis with high LD may be hopeless.

## Association Mapping

Linkage disequilibrium between polymorphic sites reduces resolution because you cannot be certain which site is responsible for association with phenotype.



LD varies among genes, species, and populations within species!