# HOS 6236 Molecular Marker Assisted Plant Breeding

## Fall 2017

*Last Class:*

*Linkage phase, QTL analysis basic, project questions*

*Todays Class:*

*QTL analysis*

# Steps for a QTL analysis

1. Create or find a suitable population
2. Genotype with molecular markers
3. Use markers to build a linkage map
4. Phenotype for trait of interest (and more)
5. Use the linkage map with the phenotypic data to determine whether markers are correlated to traits; Number of QTLs, the amount of variation and position on genome

# Steps for a QTL analysis

1. Create or find a suitable population
2. Genotype with molecular markers
3. Use markers to build a linkage map
4. Phenotype for trait of interest (and more)
5. **Use the linkage map with the phenotypic data to determine whether markers are correlated to traits**; Number of QTLs, the amount of variation and position on genome

# Why QTL analysis?

- If traits were controlled by a single gene then Mendelian analysis should be enough to detect them

- Quantitative traits are controlled by many genes, so we are interested in:
  - number,
  - positions,
  - amount of variation they control

# Why QTL analysis?

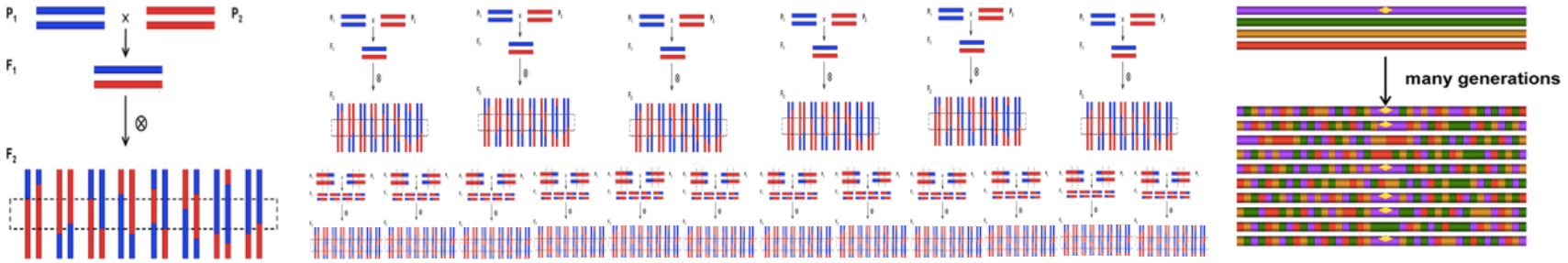
- QTL analysis depends on Linkage disequilibrium (LD) between the marker and gene controlling the trait

# Methods based on Linkage disequilibrium

I - QTL analysis

III – Genome-wide selection

II – Genetic association




Resolution

Low — Medium — High

Linkage Blocks

Large — Medium — Small

(Meuwissen et al., 2001)

# How to create LD

- F2, BC1, DHs, RILs are experimental designs to create this disequilibrium

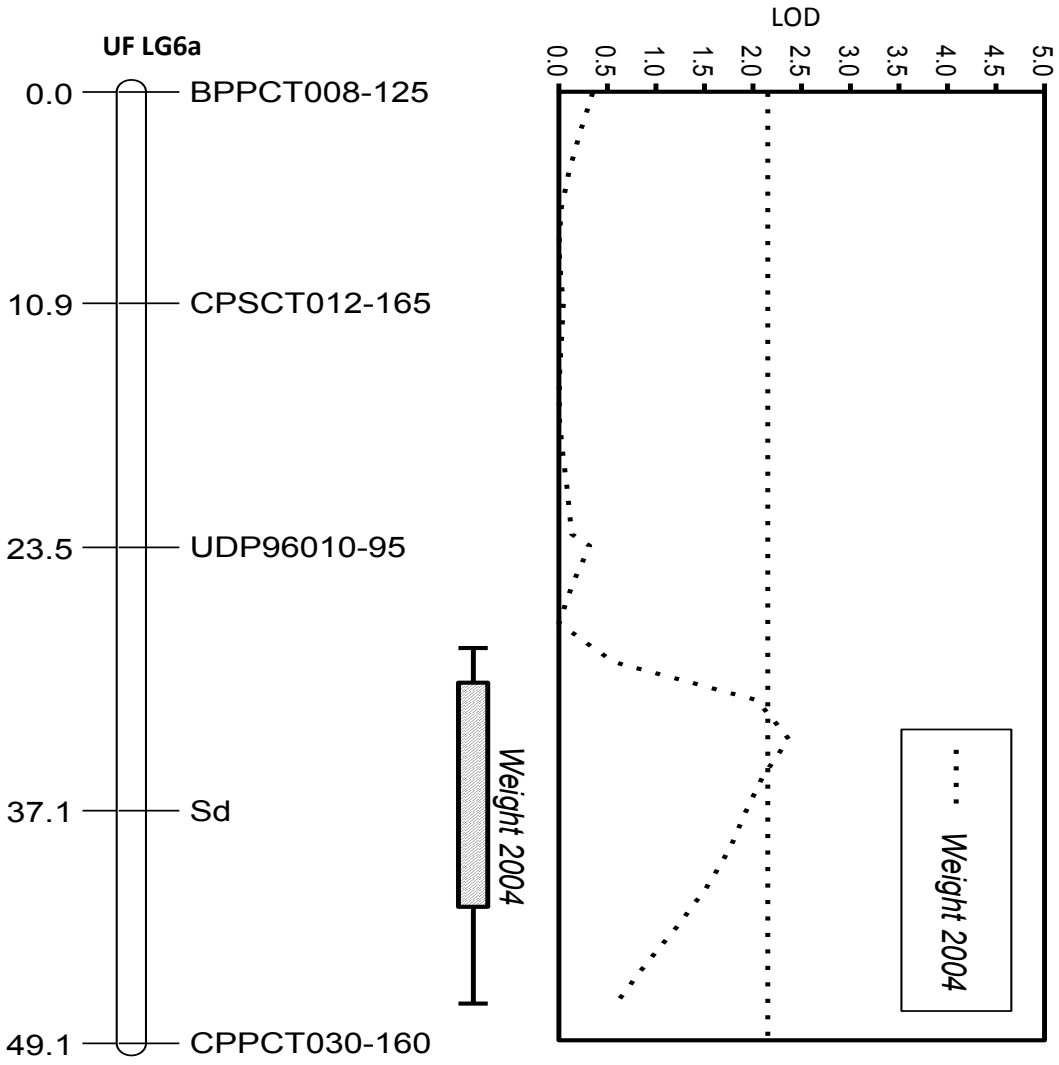
# Finding QTLs: Single Marker Regression (SMR)

- Simplest method
- The association of each marker is tested independently of other markers
- Uses ANOVA method
- Good method to only detect associations
- Problems:
  - Only larger markers indicate the location of QTL, but no indication of the distance
  - Recombination between marker and QTL underestimate the QTL effect

# Finding QTLs: Interval Mapping (IM)

- Analysis uses the intervals between adjacent markers instead of single markers
- Thus, uses marker position on the map
- And eliminates the problem of recombination between the marker and QTL
- Uses most powerful statistics methods Maximum likelihood approaches instead of ANOVA
- Better estimation of effect and position
- Problems:
  - If more than one QTL is linked to the interval being analyzed then the estimation of the effect size is bias

# Interval Mapping



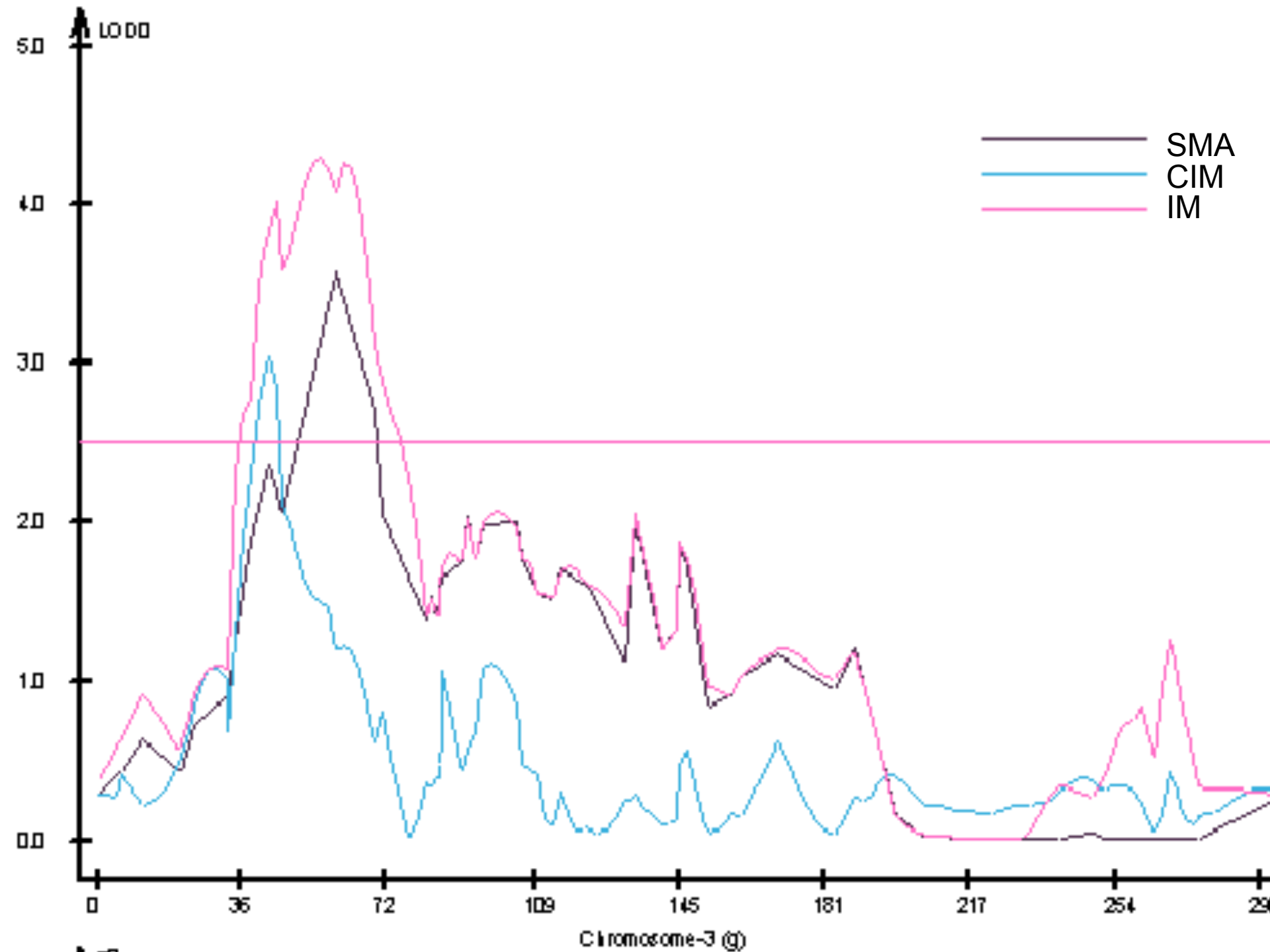

- $R^2$ = 26.4%
- Additive value: - 1.55g

# Finding QTLs: Composite Interval Mapping (CIM)

- Uses intervals between adjacent markers (uses the linkage map)
- No problem of recombination between the marker and QTL
- **Uses other markers as co-factors in the model to reduce the bias caused by multiple QTLs linked to the interval being analyzed**
- Uses Maximum likelihood method
- Best estimation of effect and position
- **Gold Standard**
- Problems:
  - What markers to chose as co-factors

# Composite Interval Mapping

- Focus on the interval and eliminates confounding effects from other QTLs in the genome
- Resolution of the QTL location is increased – the most likely position is more likely to be identified


LOD
5.0
4.0
3.0
2.0
1.0
0.0
SMA
CIM
IM
0
36
72
109
145
181
217
254
Chromosome-3 (g)

# Hypothesis Testing

- *True Positive:* QTL is correctly declared present
- *False Positive:* QTL is incorrectly declared present – Type I error
- *True Negative:* QTL is correctly declared absent
- *False Negative:* QTL is incorrectly declared absent – Type II error

# Determining QTL Significance

- LOD scores commonly used to determine statistical significance of genetic linkages and QTL

$$\text{LOD} = \log\left(\frac{\text{prob. of obtaining the observed data under linkage}}{\text{prob. of obtaining the observed data under random assortment}}\right)$$

$p < 0.01 \sim$ LOD 2.0

$p < 0.001 \sim$ LOD 3.0

UF UNIVERSITY of FLORIDA
The Foundation for The Gator Nation

# Determining QTL Significance

- Conventional threshold for declaring the presence of a QTL is LOD 3.0
  - False positive 1 in 1000 times
- Threshold commonly determined by permutation testing

# Permutation Testing

- LOD threshold for QTL is commonly determined by permutation testing
  - Shuffle phenotypic data
  - Keep genotypic data constant
  - Repeat 1000 or more times
  - Determine a significance level ($\alpha$)

# Permutation Testing

- If $\alpha = 0.05$, then with 1000 permutation tests, we would falsely declare a QTL-marker association 50 times
  - Genome-wide LOD score is calculated for each random iteration of phenotypic data
  - For 1000 permutation, the $950^{th}$ largest LOD score becomes the LOD threshold value

# QTL Map

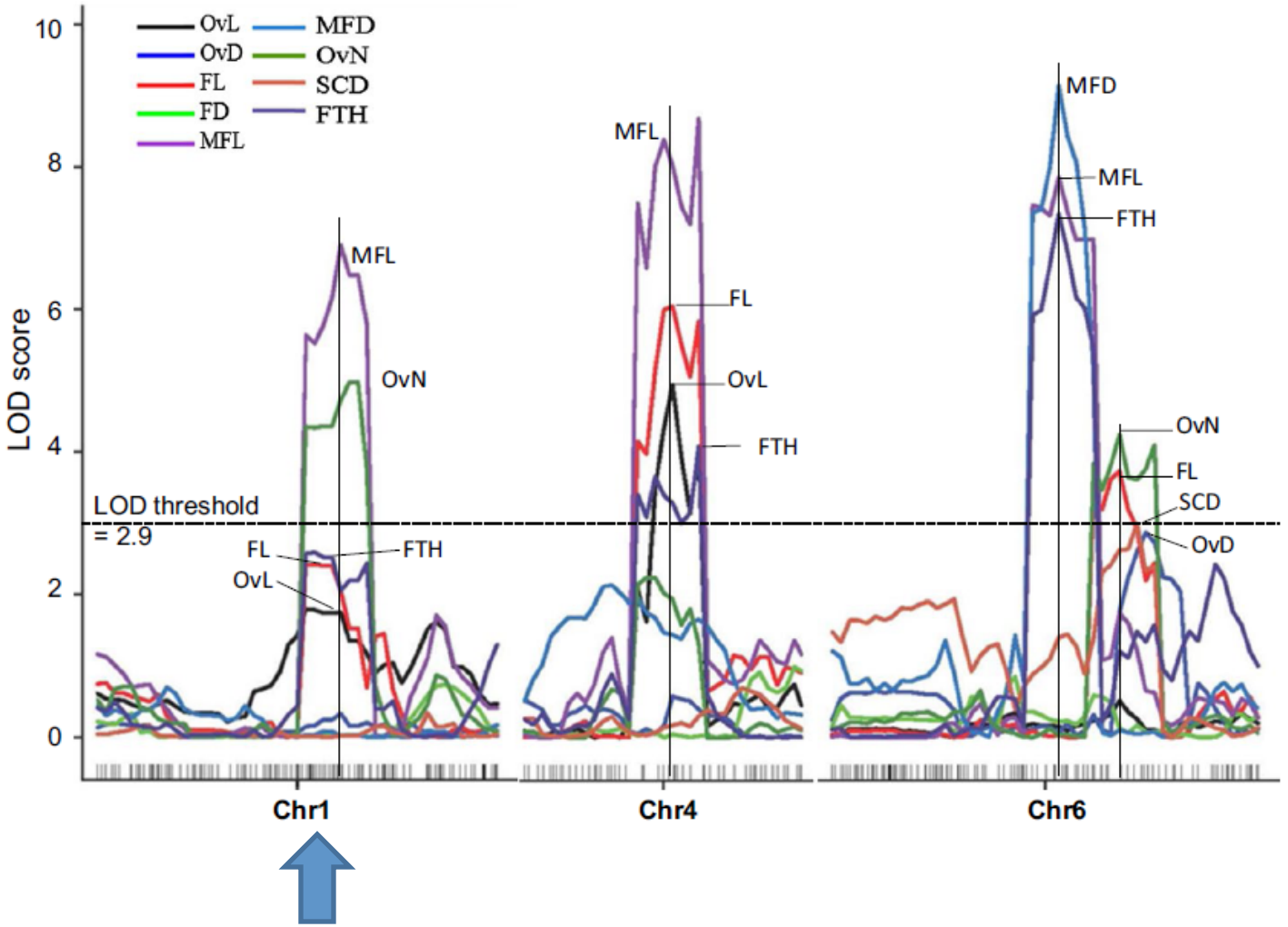



**Fig. 4** LOD profiles of fruit size-related QTLs detected with MQM model in the RIL population and high-density SNP maps in cucumber chromosomes 1 (*left*), 4 (*middle*), and 6 (*right*). The *dashed horizontal line* is LOD threshold for all QTLs (LOD = 2.9). *OvL* ovary length, *OvD* ovary diameter, *FL* immature fruit length, *FD* immature fruit diameter, *MFL* mature fruit length, *MFD* mature fruit diameter. *FTH* flesh thickness, *SCD* seed cavity size and *OvN* ovule number in mature fruit

Weng et al. (2015) Theor. Appl Genet. 128:1747

28.5 cM interval: The width of the interval will be determined by the marker density

# Estimating the QTL effect

Sample size plays a significant role in the QTL effect.

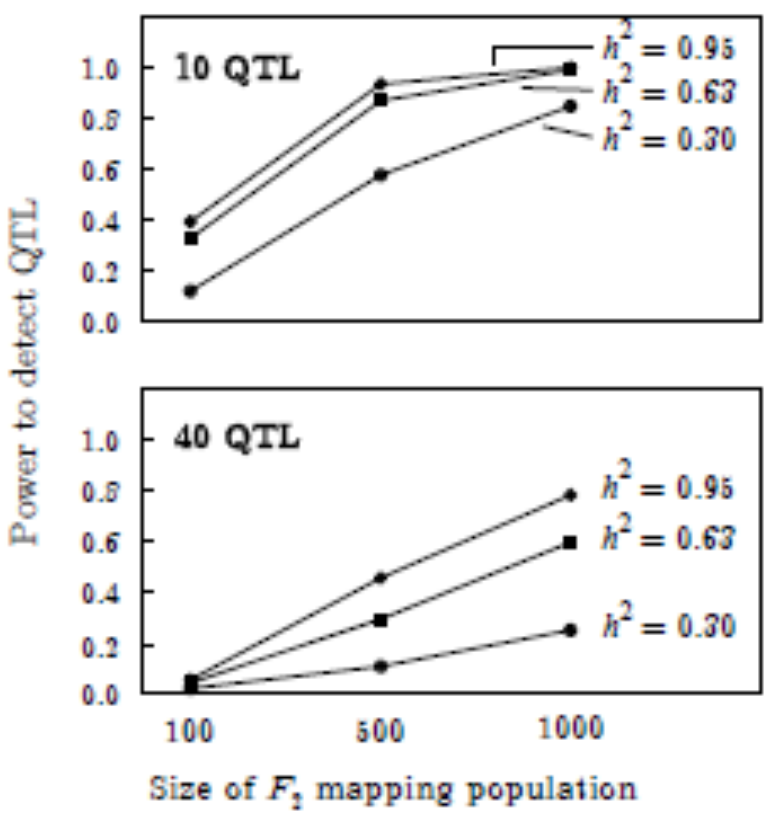
For example a small sample size may miss QTLs with small effect thus will overestimate the QTL effect of the ones that is able to identify, the “Beavis effect” (Beavis 1994, 1997)

Also, a small detected QTL effect could be either:

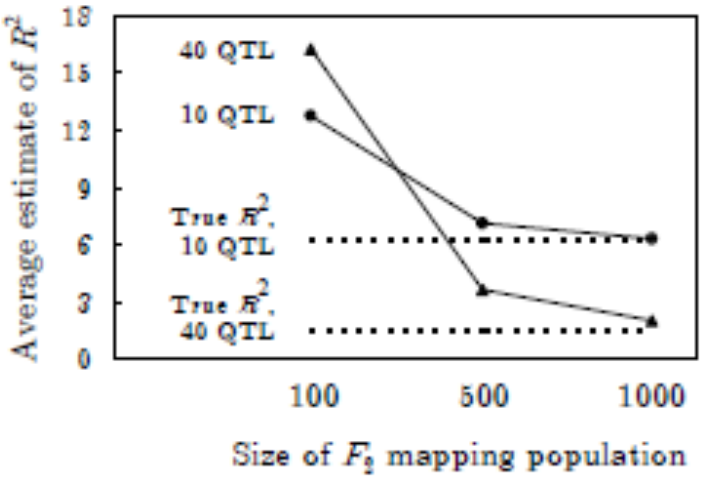
A tightly-linked QTL with a small effect

A loose linkage QTL with large effect

# The Beavis Effect




- Number of QTL are underestimated with small population sizes




- Phenotypic variance associated with QTL are overestimated when small population sizes are used

# Likelihood methods and LOD

Likelihood methods use the entire distribution of the data and not only the mean of a specific genotype

More powerful than ANOVA

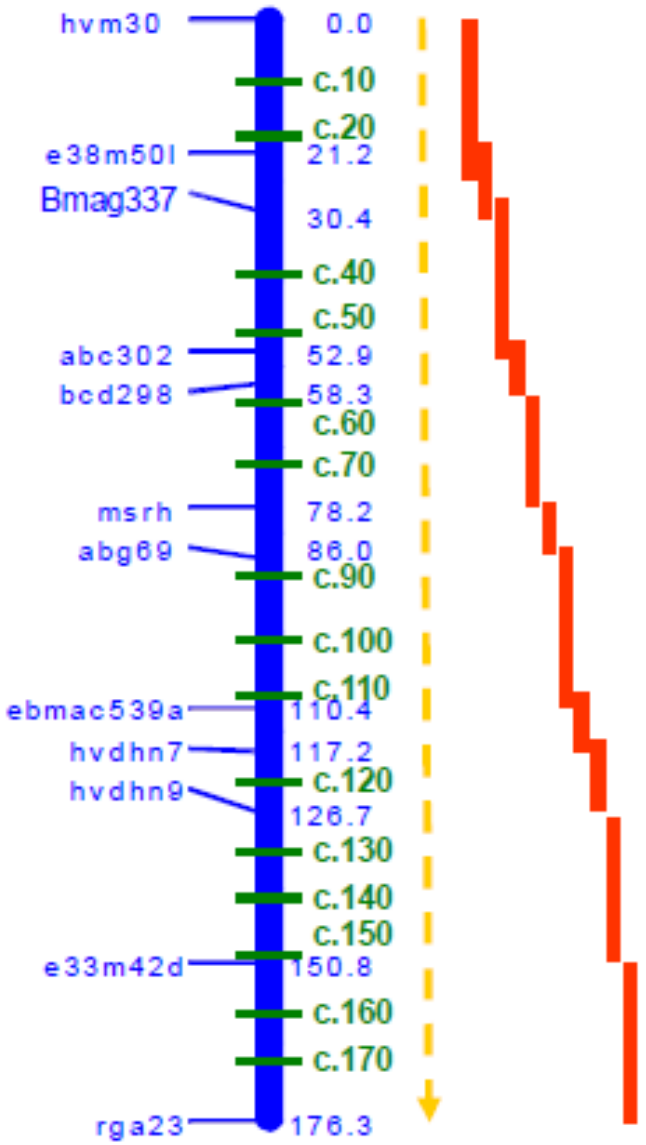
It is a mixture model combining: detection of position and estimation of effect

Marker-QTL associations are tested with Likelihood ratio test, that can be expressed as a LOD

The likelihood is a a multidimensional problem, function of the QTL means, variances, and map position

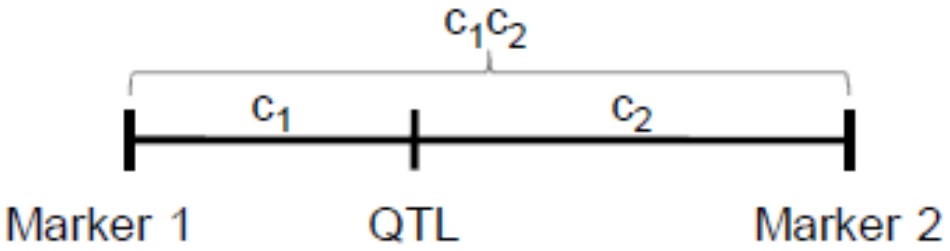
The LOD map projects the multidimensionality likelihood surface into a single dimension of the map position of the recombination events

# QTL mapping: 2. Interval Mapping



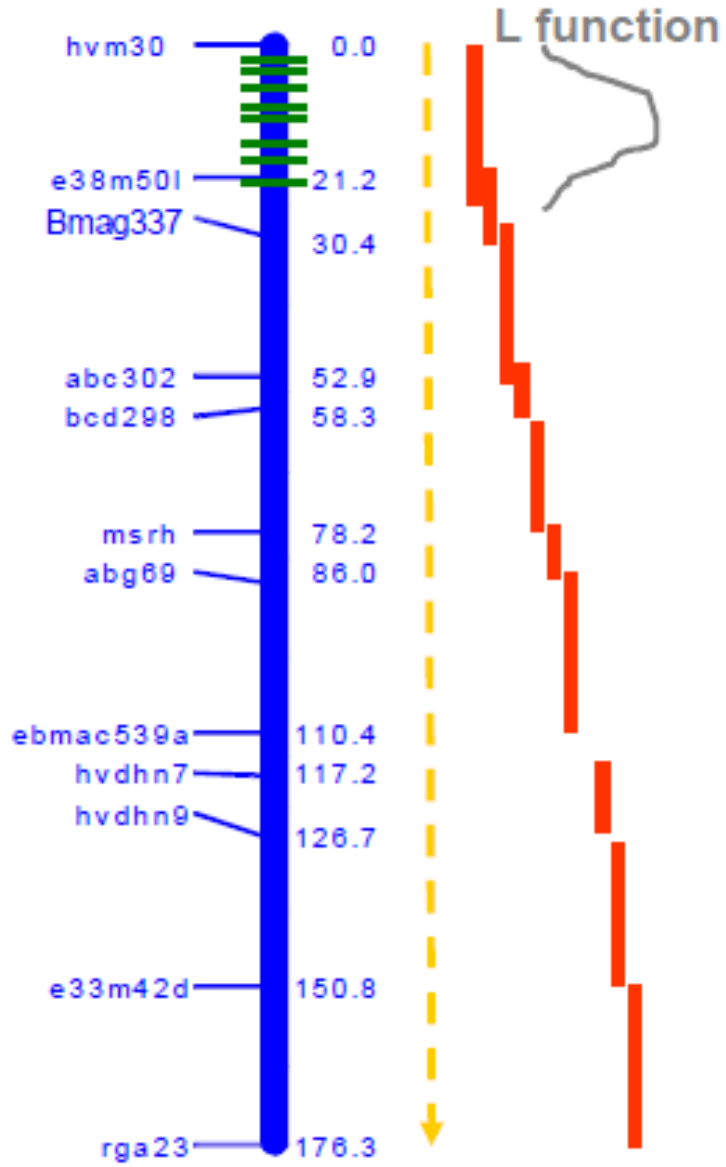

Broman and Sen, 2009

Uses contiguous marker information to improve the estimation of marker effects:



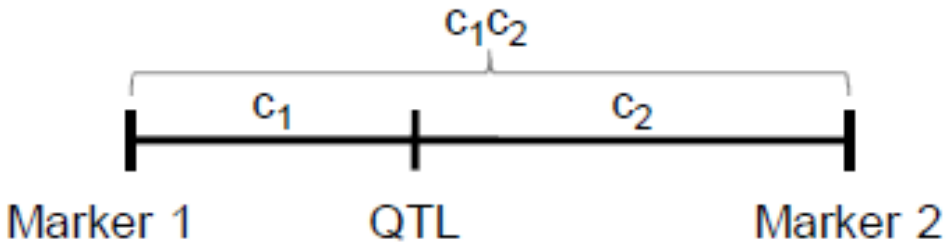

**Haley-Knott Regression:**
Uses the conditional probabilities calculated inside the interval defined by two markers directly as pseudo-markers and performs a regression on each point.

# QTL mapping: 2. Interval Mapping




Broman and Sen, 2009

Uses contiguous marker information to improve the estimation of marker effects:




**Maximum Likelihood Methods:**
Uses the likelihood function and the conditional probabilities inside the interval defined by two markers to determine the most likely position of the QTL inside the interval.

# QTL mapping: 2. Interval Mapping

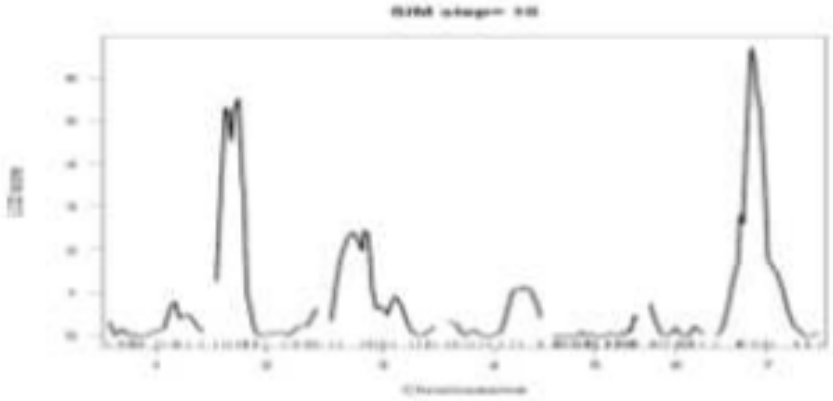
**Simple Interval Mapping (+):**

- Evaluation at and between markers
- Estimation of QTL position
- No specialized software (only for conditional distribution calculations)

**Simple Interval Mapping (-):**

- Single QTL model
- Loss of power due to residual variance caused by other QTL
- n-1* tests

With high marker density it is very similar to marker regression



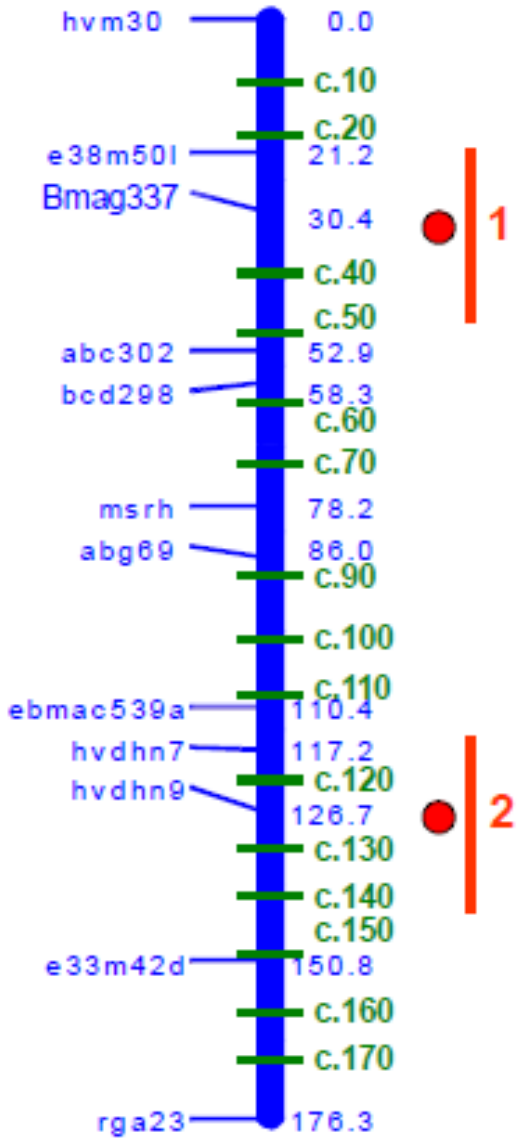
# QTL mapping: 3. Composite Interval

## COMPOSITE INTERVAL MAPPING (CIM):

**IDEA:** On top of using contiguous marker information, use background loci to get a better estimation of QTL effects. MR and SIM provides biased estimation when multiple QTL are close to a marker and have less power in general. The challenge is how to select the cofactors.

**WHEN TO USE IT?:** It is the preferred method because it has more power and decreases bias due to contiguous QTL.

# QTL mapping: 3. Composite Interval




Uses markers as cofactors to improve the estimation of genetic background interactions.

No cofactor is allowed within windows of specific size to avoid over fitting.

Conditional probabilities in-between markers are still used to improve estimations.

Outside both windows: $y_i = \mu + M_i + C_1 + C_2 + \varepsilon_i$

Inside window 1: $y_i = \mu + M_i + C_2 + \varepsilon_i$

Inside window 2: $y_i = \mu + M_i + C_1 + \varepsilon_i$



Broman and Sen, 2009

# QTL mapping: 3. Composite Interval

**Composite Interval Mapping (+):**

- Evaluation at and between markers
- Estimation of QTL position
- Control of genetic background interactions
- Multiple QTL screened

**Composite Interval Mapping (-):**

- How to select marker-cofactors?