# Building regression models

## Mapping multiple QTL

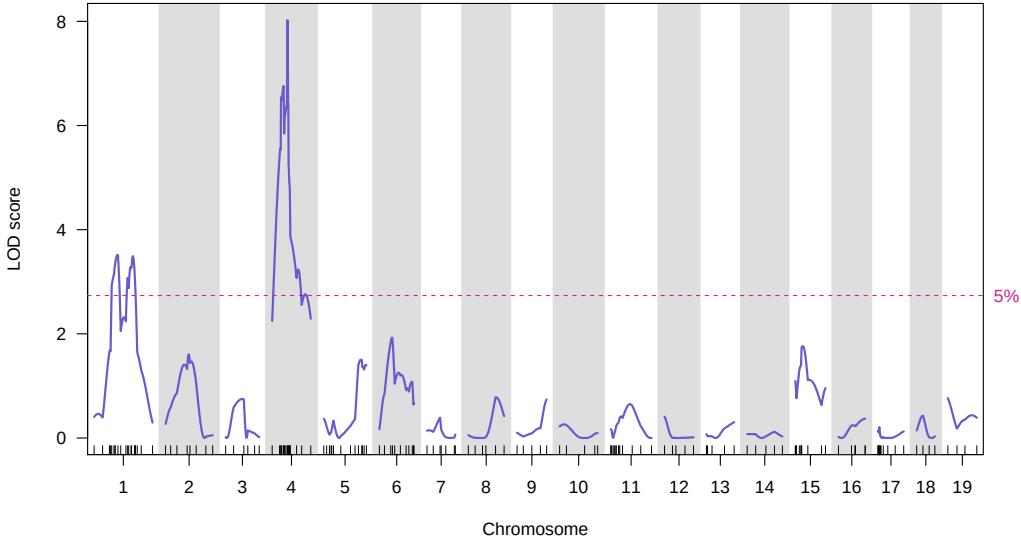
Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman

Course web: kbroman.org/AdvData

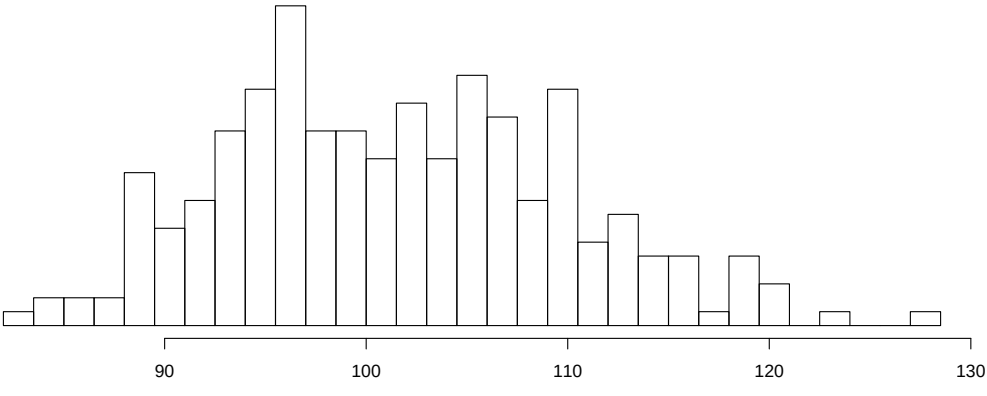
# LOD curves


LOD score
0
2
4
6
8
5%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Chromosome

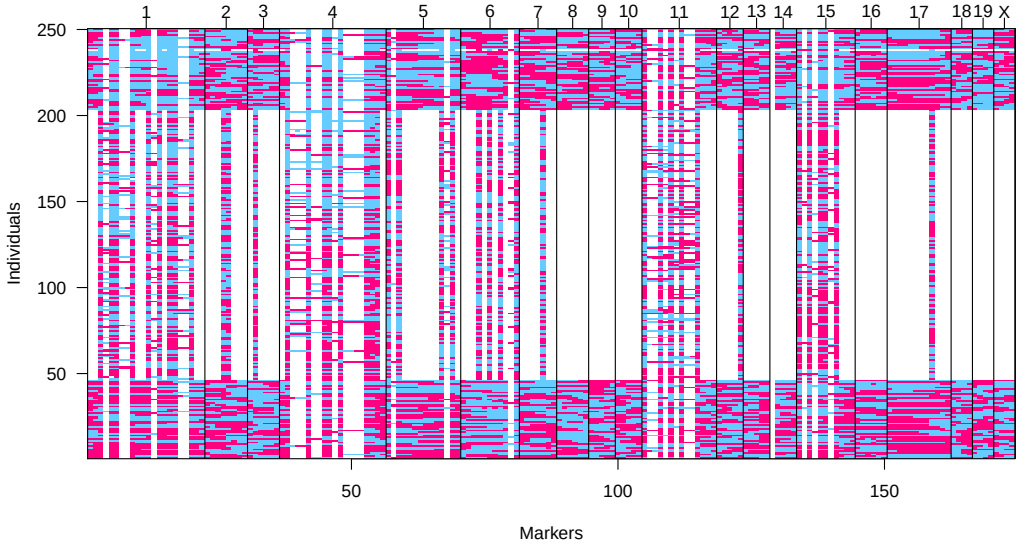
# Example

Sugiyama et al. Genomics 71:70-77, 2001

- 250 male mice from the backcross (A $\times$ B) $\times$ B
- Blood pressure after two weeks drinking water with 1% NaCl



# Genotype data

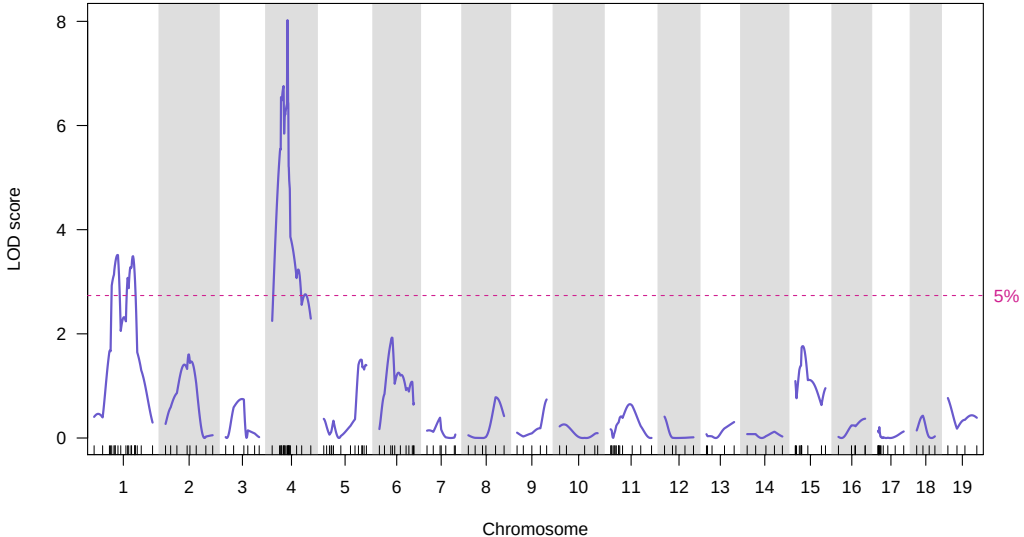


# Goals

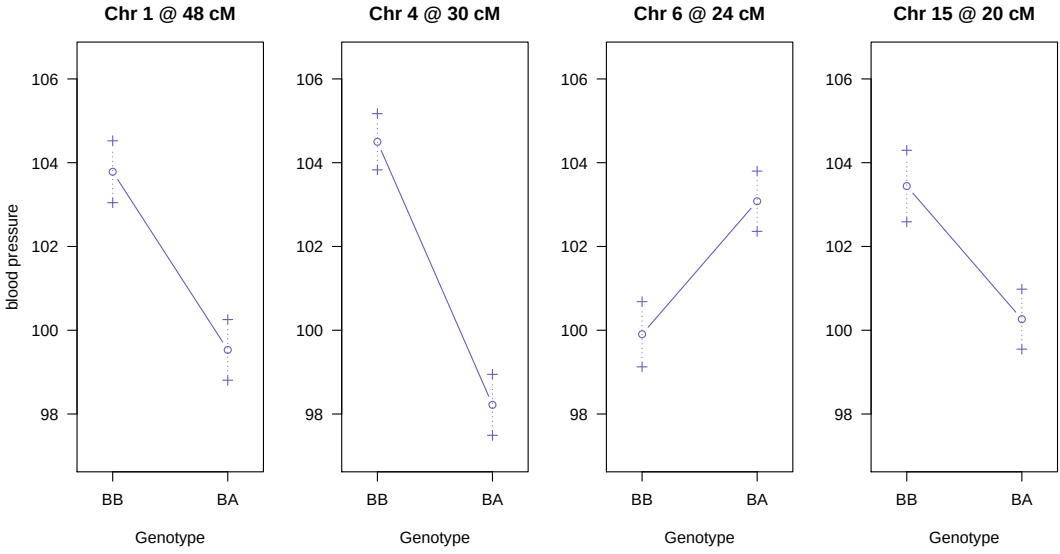
- Identify quantitative trait loci (QTL)

  (and interactions among QTL)
- Interval estimates of QTL location
- Estimated QTL effects

# LOD curves


LOD score
0
2
4
6
8
5%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Chromosome

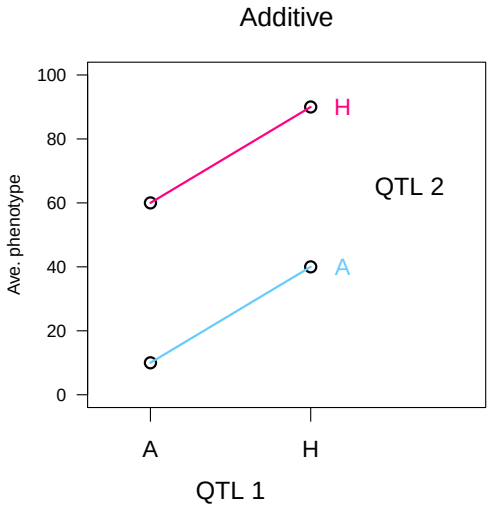
# Estimated effects

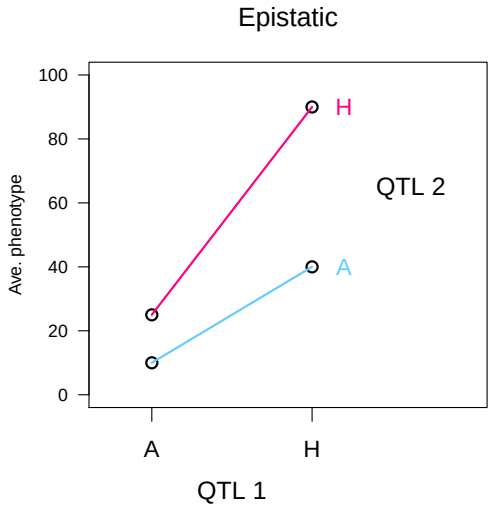


# Modeling multiple QTL

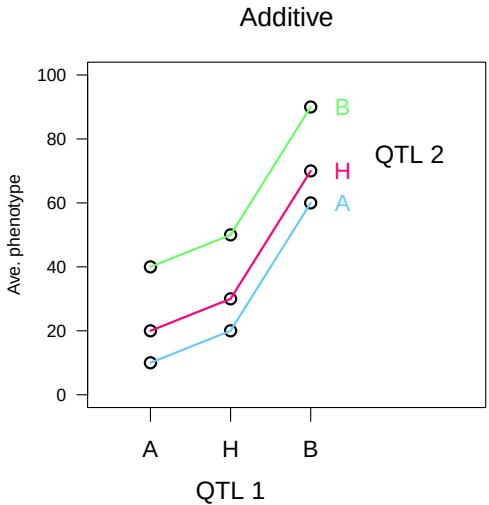
- Reduce residual variation $\longrightarrow$ increased power
- Separate linked QTL
- Identify interactions among QTL (epistasis)

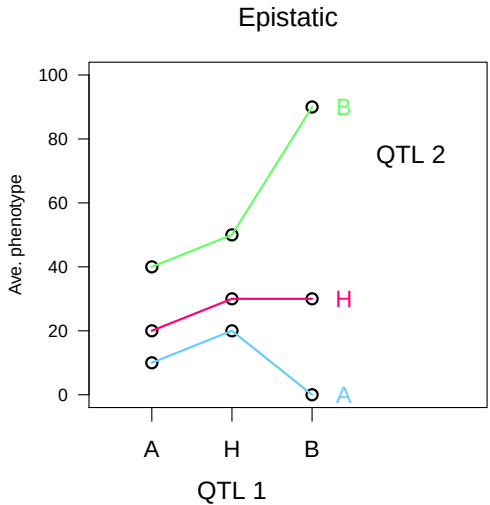
# Epistasis in BC



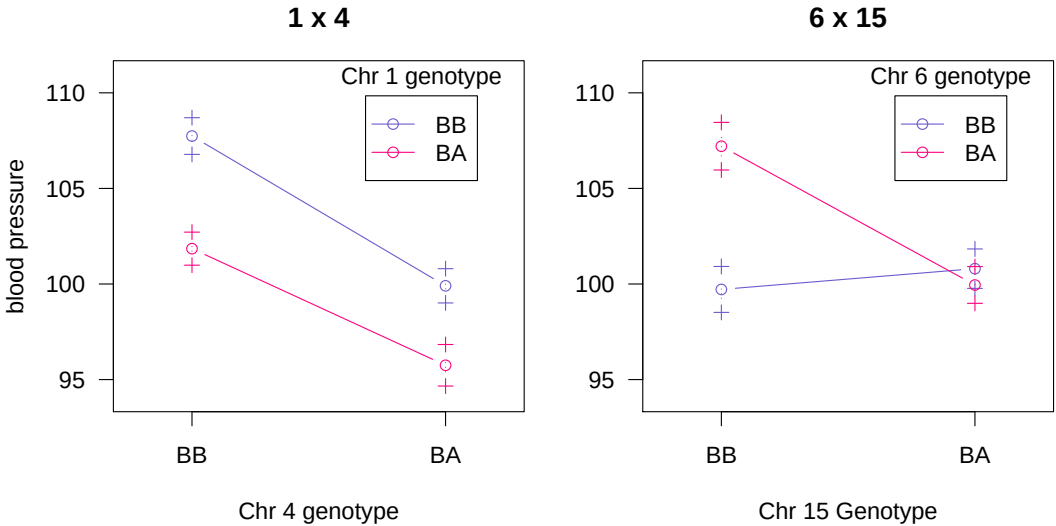



# Epistasis in $F_2$



# Estimated effects



# Model selection (or variable selection)

- Subset selection
- $L_1$-penalized regression (the LASSO)
- regression forests
- Bayes
- ...

# Model selection

- Class of models
  - Additive models
  - + pairwise interactions
  - + higher-order interactions
  - Regression trees
- Model fit
  - Maximum likelihood
  - Haley-Knott regression
  - extended Haley-Knott
  - Multiple imputation
  - MCMC
- Model comparison
  - Estimated prediction error
  - AIC, BIC, penalized likelihood
  - Bayes
- Model search
  - Forward selection
  - Backward elimination
  - Stepwise selection
  - Randomized algorithms

# Target

- Selection of a model includes two types of errors:
  - Miss important terms (QTLs or interactions)
  - Include extraneous terms
- Unlike in hypothesis testing, we can make both errors at the same time.
- Identify as many correct terms as possible, while controlling the rate of inclusion of extraneous terms.

# What is special here?

- Goal: identify the major players
  - not prediction
- A continuum of ordinal-valued covariates (the genetic loci)
- Association among the covariates
  - Loci on different chromosomes are independent
  - Along chromosome, a very simple (and known) correlation structure

# Exploratory methods

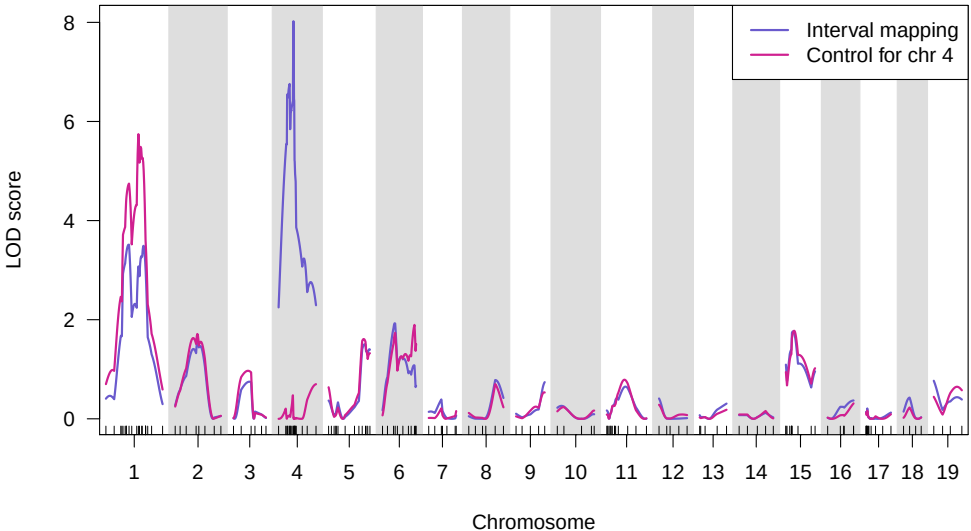
- Condition on a large-effect QTL
  - Reduce residual variation
  - Conditional LOD score:

$$\text{LOD}(q_2 \mid q_1) = \log_{10}\left\{\frac{\Pr(\text{data} \mid q_1, q_2)}{\Pr(\text{data} \mid q_1)}\right\}$$

- Two-dimensional, two-QTL scan to investigate linked loci or interactions.

- Piece together the putative QTL from the 1d and 2d scans
  - Omit loci that no longer look interesting (drop-one-at-a-time analysis)
  - Study potential interactions among the identified loci
  - Scan for additional loci (perhaps allowing interactions), conditional on these

# Controlling for chr 4



# Drop-one-QTL table

| | df | LOD | %var |
|---|---|---|---|
| 1@68.3 | 1 | 6.30 | 11.0 |
| 4@30.0 | 1 | 12.21 | 20.1 |
| 6@61.0 | 2 | 7.93 | 13.6 |
| 15@17.5 | 2 | 7.14 | 12.3 |
| 6@61.0 : 15@17.5 | 1 | 5.68 | 9.9 |

# Automation

- Assistance to non-specialists
- Understanding performance
- Many phenotypes

# Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

$$y = \mu + \sum \beta_j q_j + \epsilon \qquad \text{which } \beta_j \neq 0?$$

$$\text{pLOD}(\gamma) = \text{LOD}(\gamma) - T\,|\gamma|$$

# Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

$$y = \mu + \sum \beta_j q_j + \epsilon \qquad \text{which } \beta_j \neq 0?$$

$$\text{pLOD}(\gamma) = \text{LOD}(\gamma) - T\,|\gamma|$$

0 vs 1 QTL:

$$\text{pLOD}(\emptyset) = 0$$
$$\text{pLOD}(\{\lambda\}) = \text{LOD}(\lambda) - T$$

# Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

$$y = \mu + \sum \beta_j q_j + \epsilon \qquad \text{which } \beta_j \neq 0?$$

$$\text{pLOD}(\gamma) = \text{LOD}(\gamma) - T\,|\gamma|$$

For the mouse genome:

$T$ = 2.69 (BC) or 3.52 ($F_2$)

# Experience

- Controls rate of inclusion of extraneous terms
- Forward selection over-selects
- Forward selection followed by backward elimination works as well as MCMC
- Need to define performance criteria
- Need large-scale simulations

Broman & Speed, JRSS B 64:641-656, 2002
10.1111/1467-9868.00354

# Epistasis

$$y = \mu + \sum \beta_j q_j + \sum \gamma_{jk} q_j q_k + \epsilon$$

$$\text{pLOD}(\gamma) = \text{LOD}(\gamma) - T_m |\gamma|_m - T_i |\gamma|_i$$

$T_m$ = as chosen previously

$T_i$ = ?

# Idea 1

Imagine there are two additive QTL and consider a 2d, 2-QTL scan.

$T_i$ = 95th percentile of the distribution of

$$\max \mathrm{LOD}_f(s,t) - \max \mathrm{LOD}_a(s,t)$$

# Idea 1

Imagine there are two additive QTL and consider a 2d, 2-QTL scan.

$T_i$ = 95th percentile of the distribution of

$$\max \mathrm{LOD}_f(s,t) - \max \mathrm{LOD}_a(s,t)$$

For the mouse genome:

$T_m$ = 2.69 (BC) or 3.52 ($F_2$)

$T_i^H$ = 2.62 (BC) or 4.28 ($F_2$)

# Idea 2

Imagine there is one QTL and consider a 2d, 2-QTL scan.

$T_m + T_i$ = 95th percentile of the distribution of

$$\max \mathrm{LOD}_f(s,t) - \max \mathrm{LOD}_1(s)$$

# Idea 2

Imagine there is one QTL and consider a 2d, 2-QTL scan.

$T_m + T_i$ = 95th percentile of the distribution of

$$\max \mathrm{LOD}_f(s,t) - \max \mathrm{LOD}_1(s)$$

For the mouse genome:

$T_m$ = 2.69 (BC) or 3.52 ($F_2$)

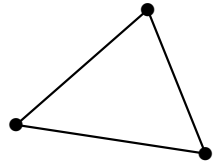
$T_i^H$ = 2.62 (BC) or 4.28 ($F_2$)

$T_i^L$ = 1.19 (BC) or 2.69 ($F_2$)

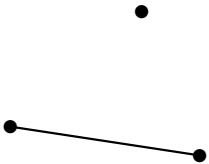
# Models as graphs

A

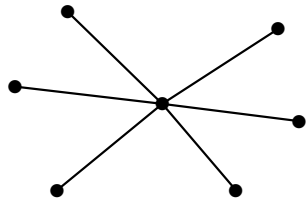
C



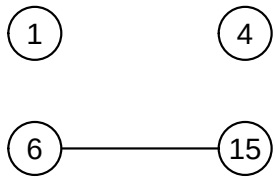
B



D

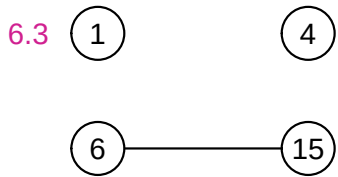


# Results




LOD = 23.1

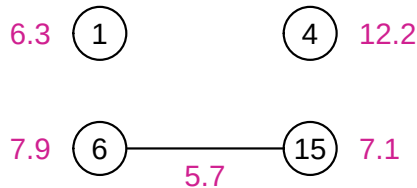
# Results




LOD = 23.1

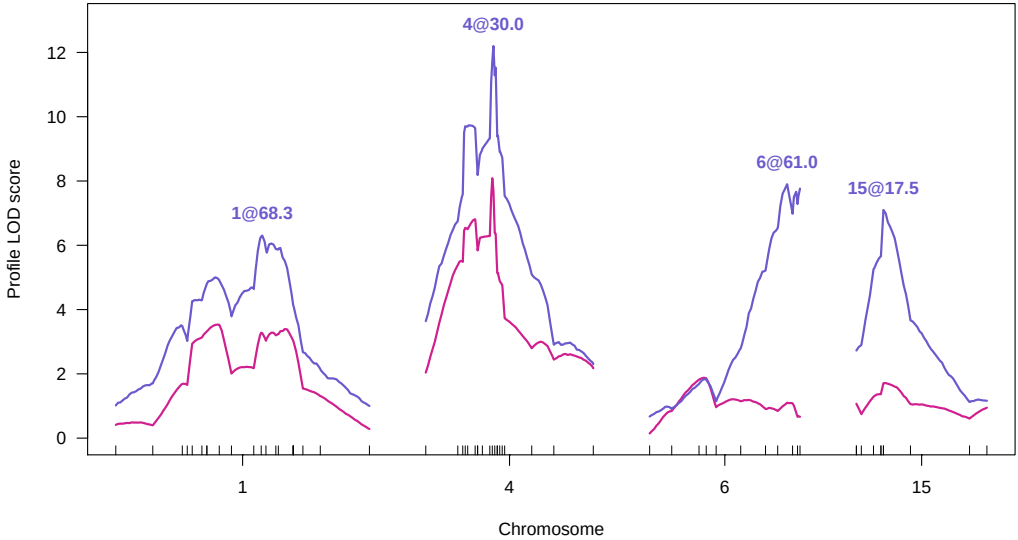
$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

# Results




$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

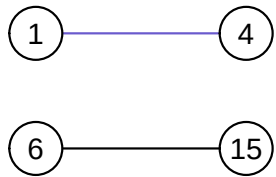
# Profile LOD curves



# Drop-one-QTL table

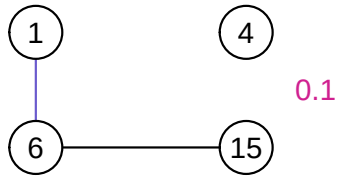
| | df | LOD | %var |
|---|---|---|---|
| 1@68.3 | 1 | 6.30 | 11.0 |
| 4@30.0 | 1 | 12.21 | 20.1 |
| 6@61.0 | 2 | 7.93 | 13.6 |
| 15@17.5 | 2 | 7.14 | 12.3 |
| 6@61.0 : 15@17.5 | 1 | 5.68 | 9.9 |

# Add an interaction?



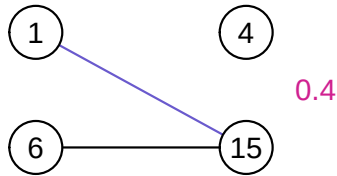

$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

# Add an interaction?



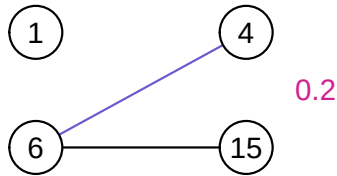

$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

# Add an interaction?



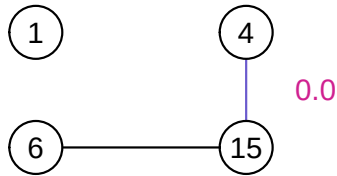

$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

# Add an interaction?



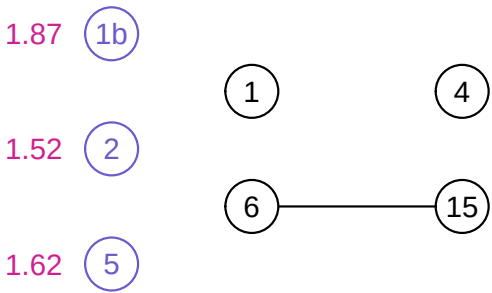

$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

# Add an interaction?



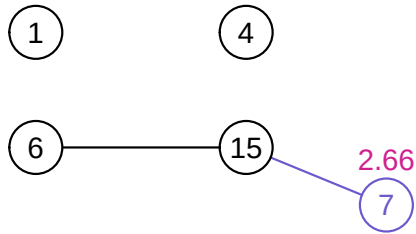

$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

# Add another QTL?



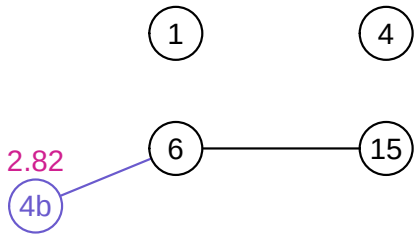

$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

# Add another QTL?



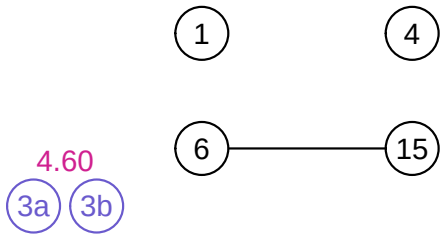

$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

# Add another QTL?




$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

# Add another QTL?




$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

# Summary

- QTL mapping is a model selection problem
- The problem is finding the major players, not minimizing prediction error
- The criterion for comparing models is most important
- We're focusing on a penalized likelihood method, with penalties derived from permutation tests with 1d and 2d scans
- Manichaikul et al., Genetics 181:1077–1086, 2009
  doi:10.1534/genetics.108.094565