

QTL mapping in experimental crosses

Part II

Karl W Broman

kbroman.org

github.com/kbroman

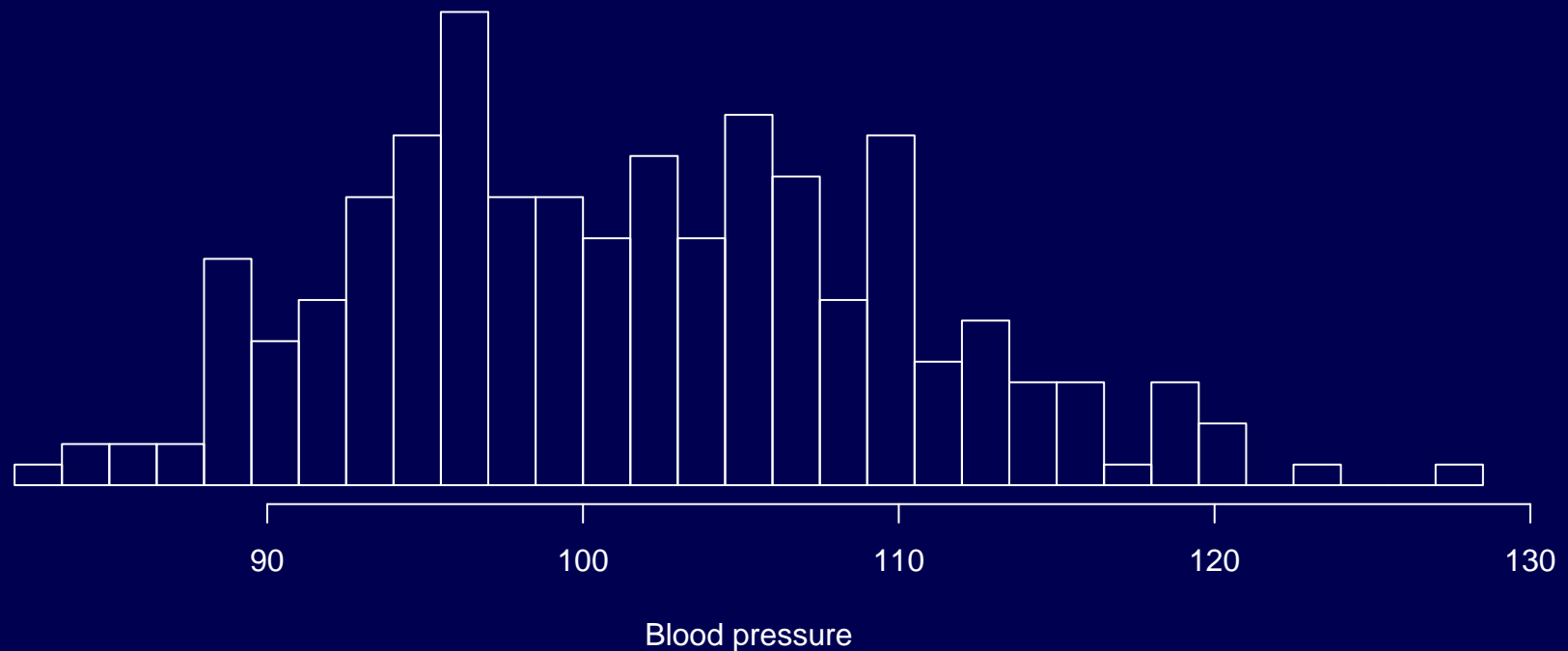
@kwbroman

Example

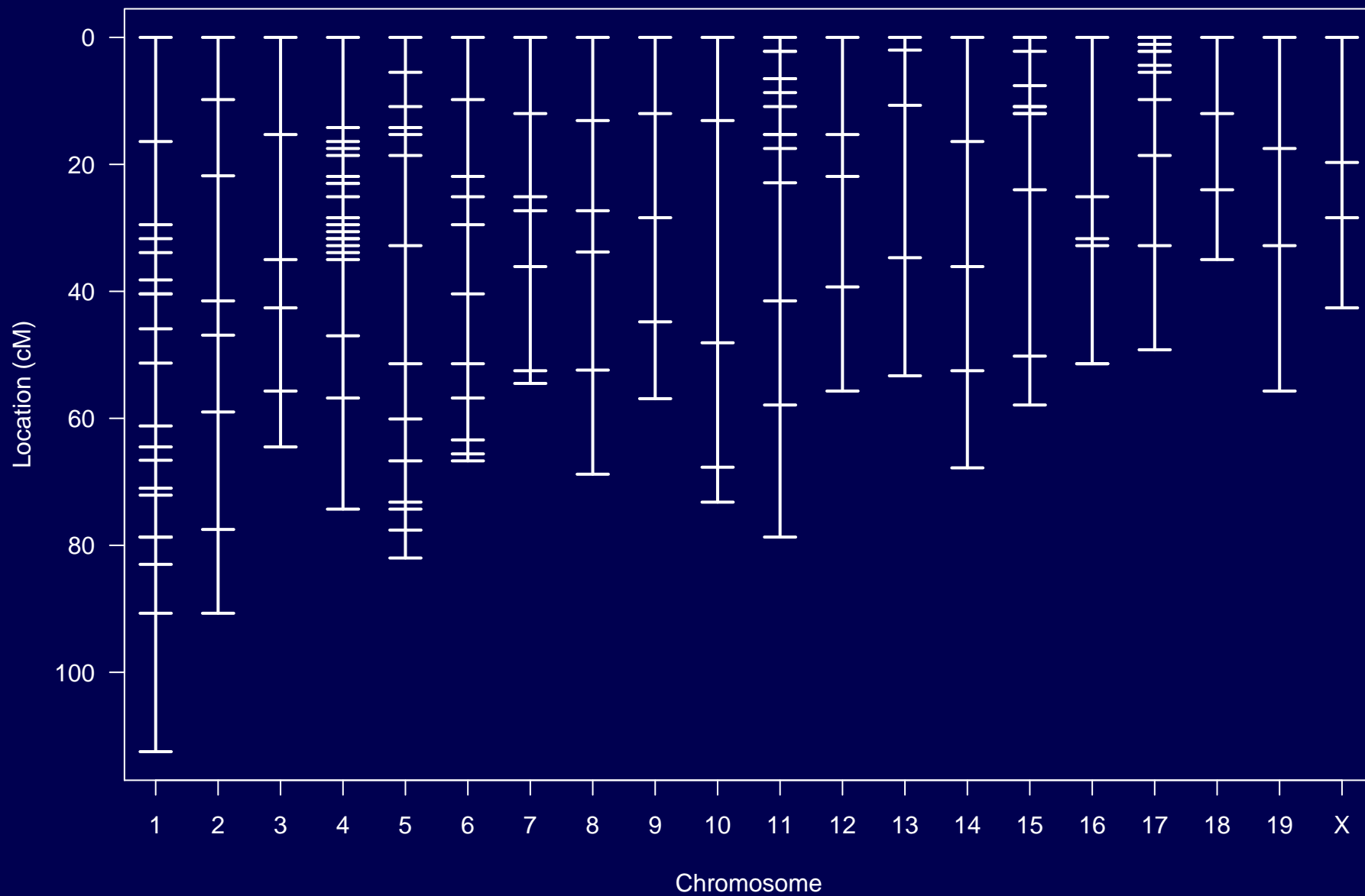
Sugiyama et al. Genomics 71:70-77, 2001

250 male mice from the backcross $(A \times B) \times B$

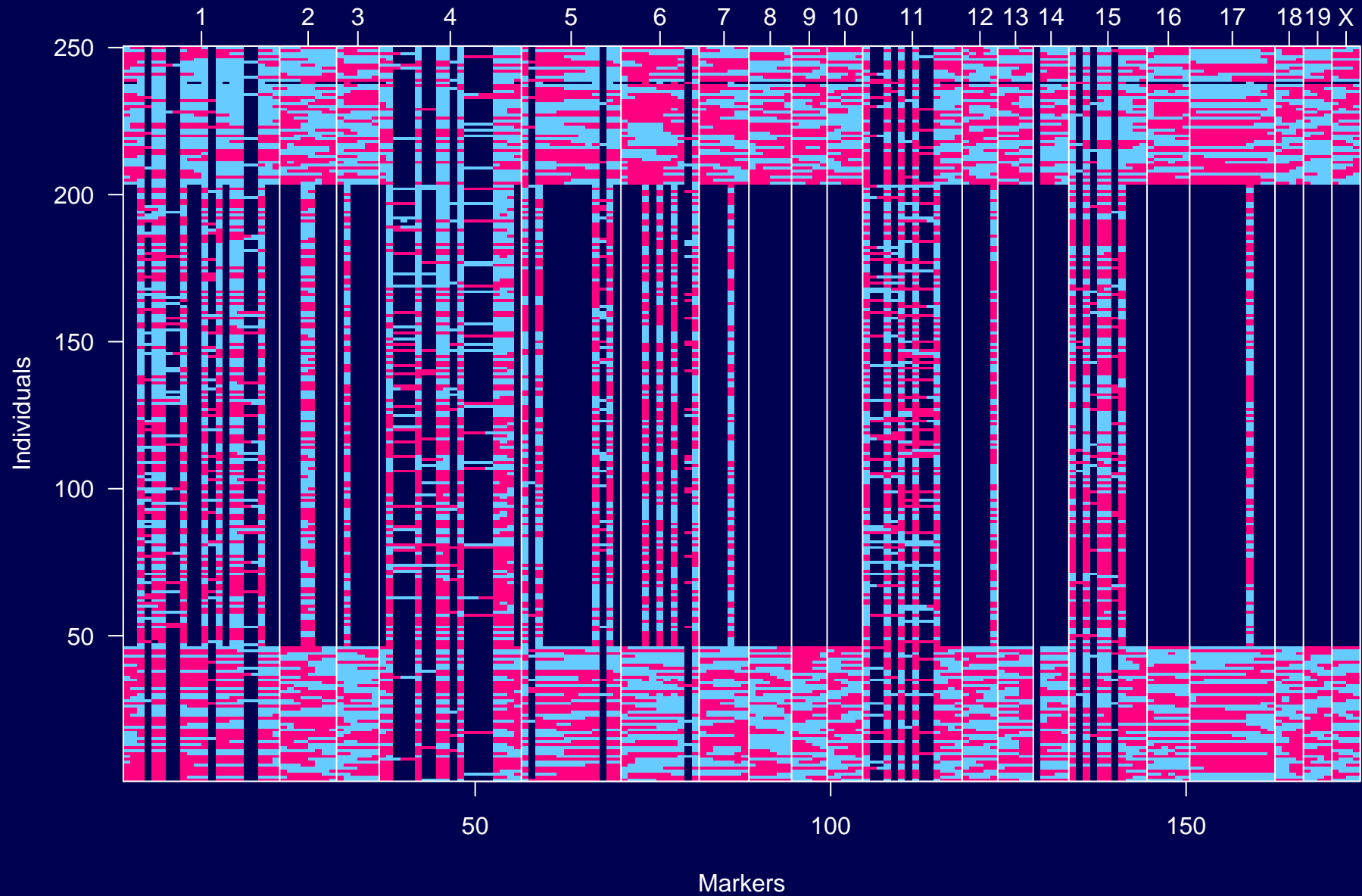
Blood pressure after two weeks drinking water with 1% NaCl



Genetic map



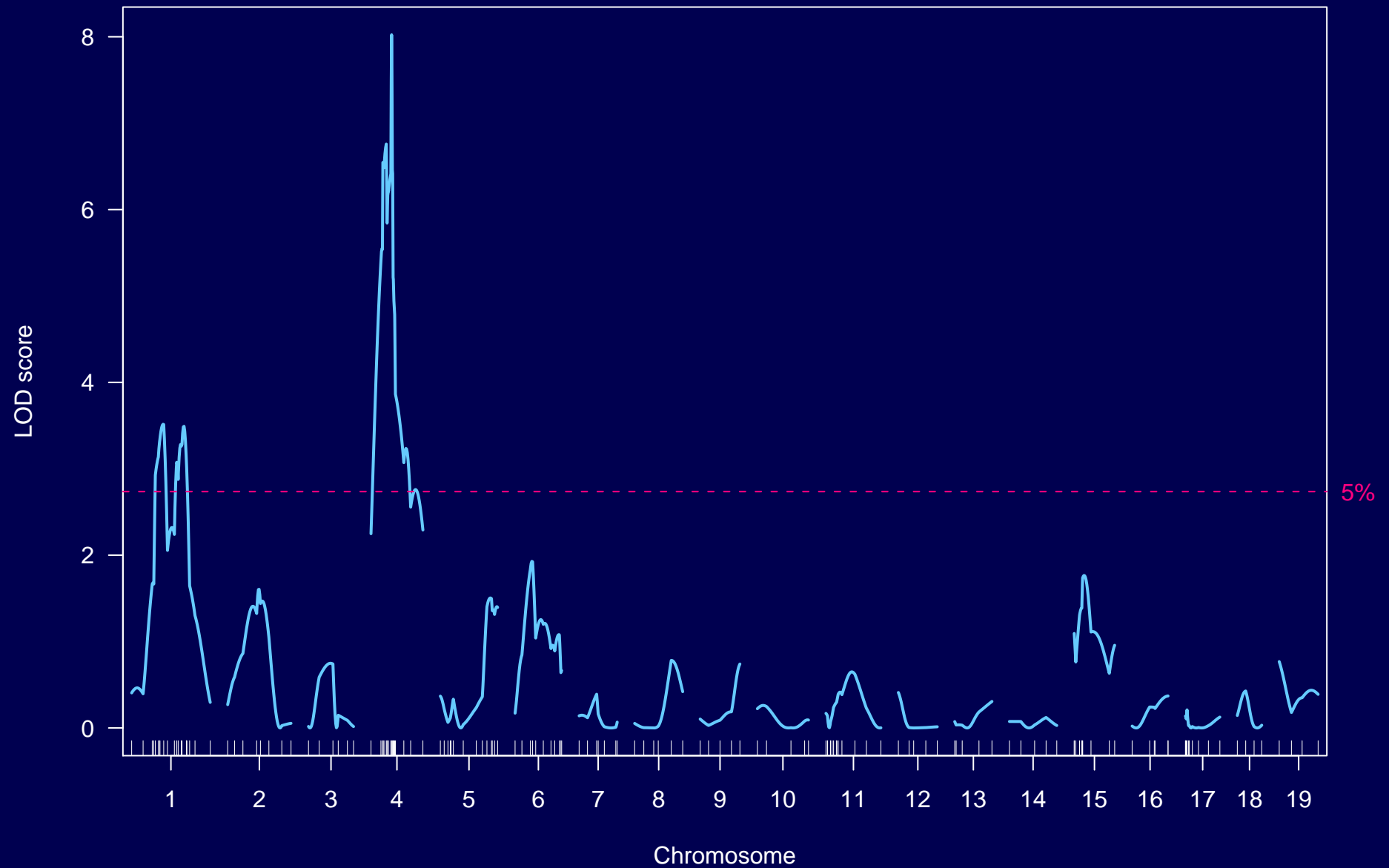
Genotype data



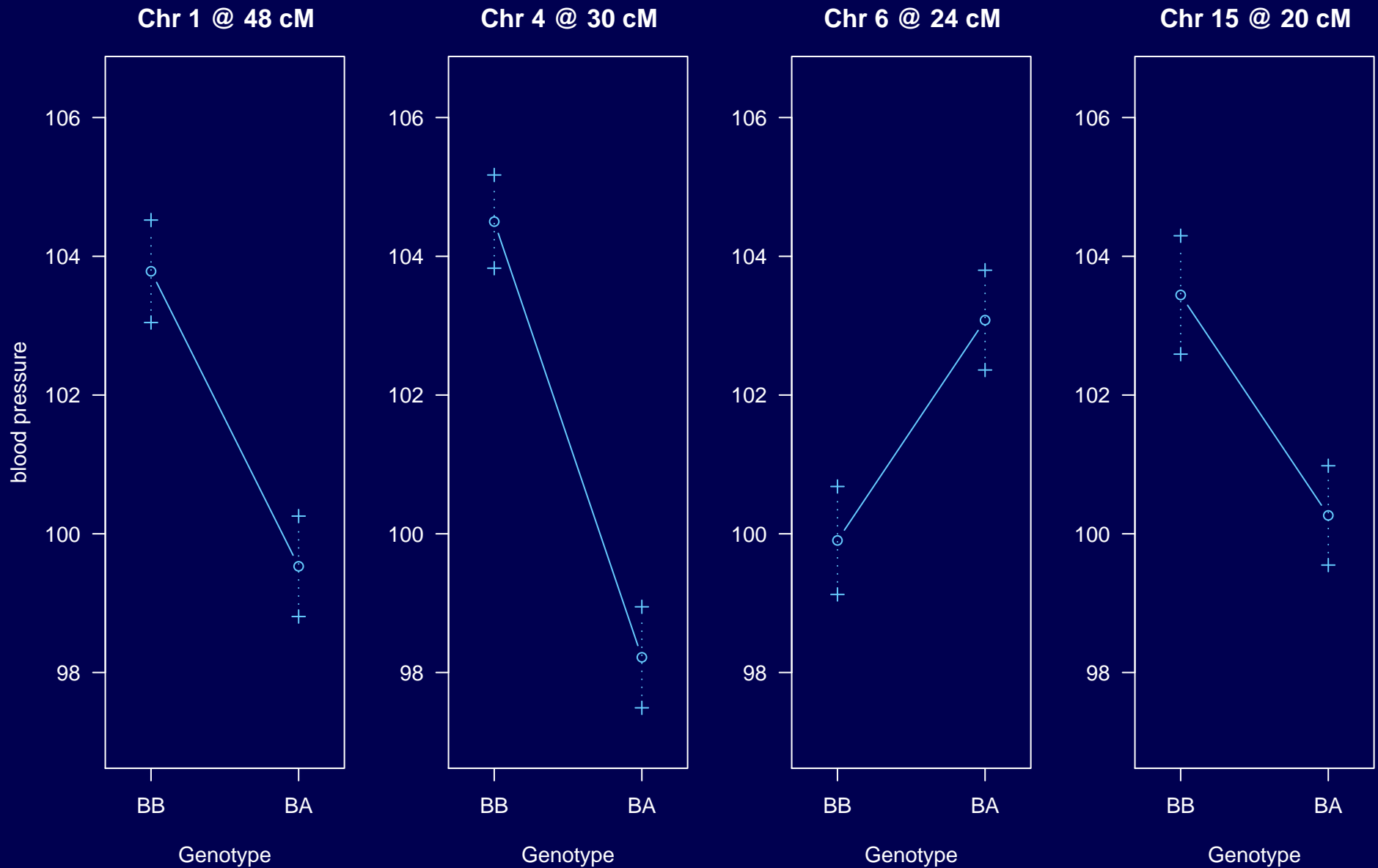
Goals

- Identify quantitative trait loci (QTL)
(and interactions among QTL)
- Interval estimates of QTL location
- Estimated QTL effects

LOD curves



Estimated effects

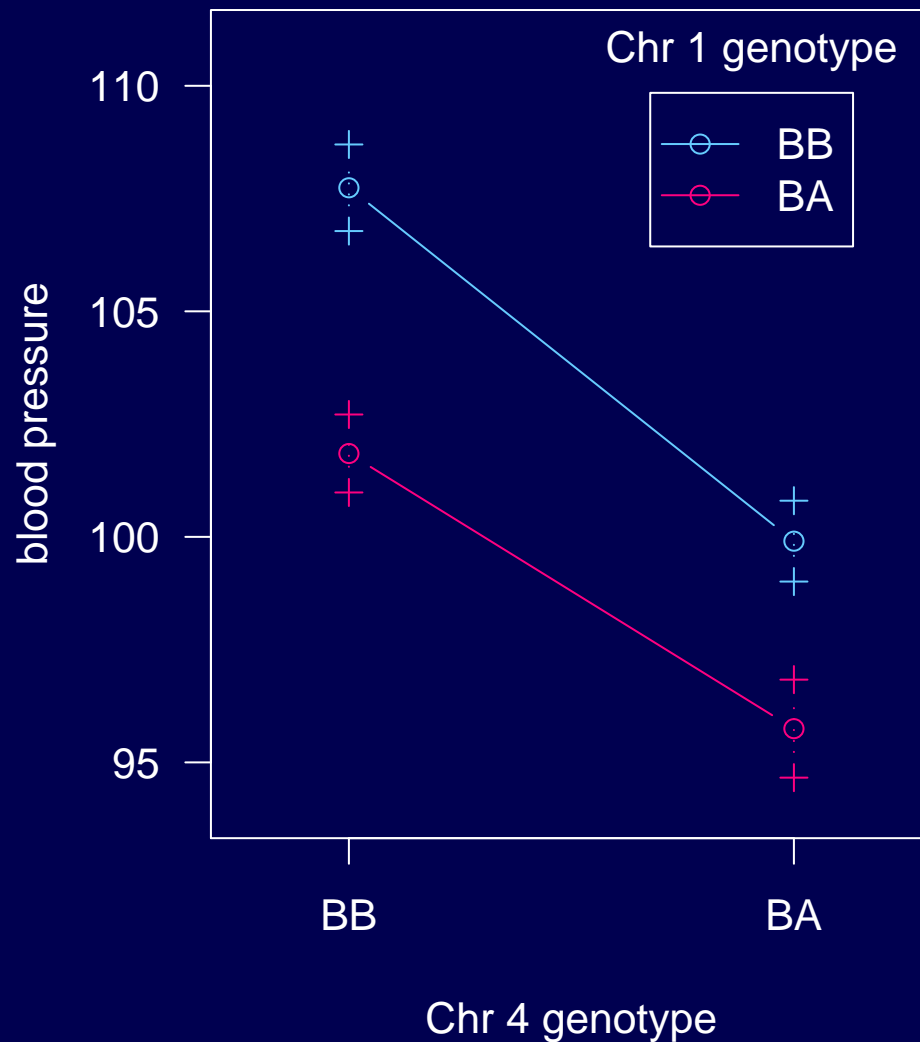


Modeling multiple QTL

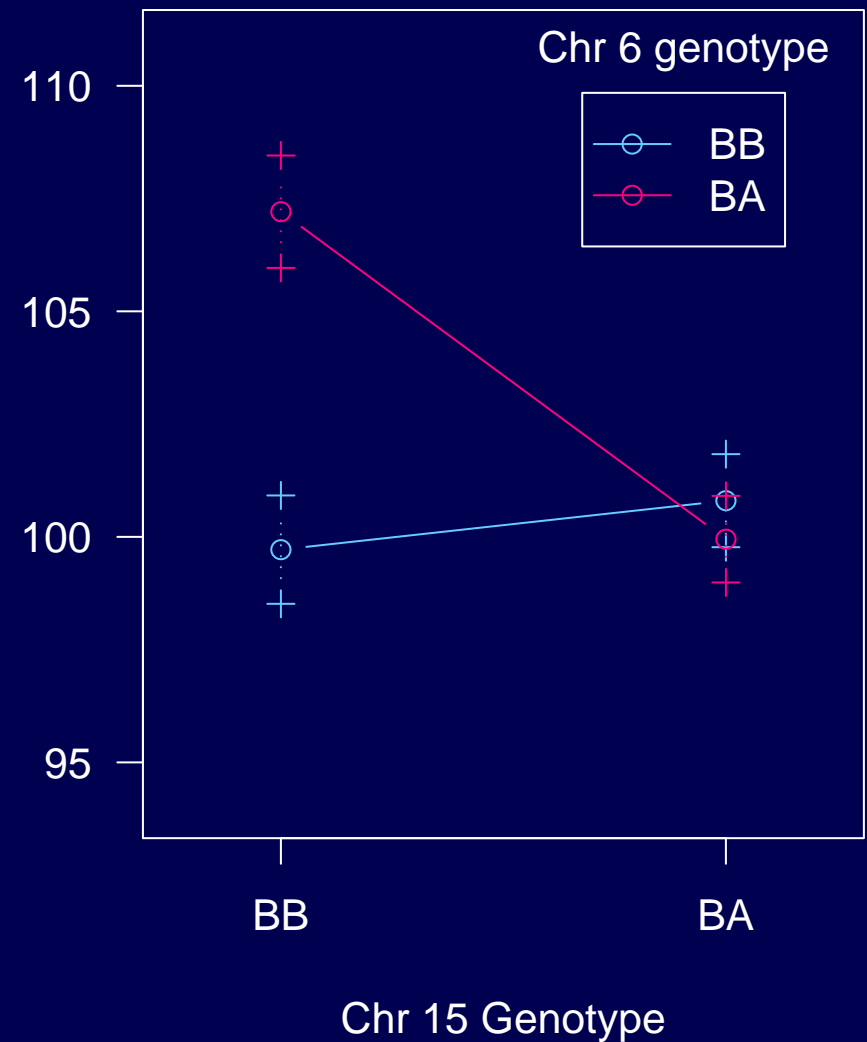
- Reduce residual variation → increased power
- Separate linked QTL
- Identify interactions among QTL (epistasis)

Estimated effects

1 x 4



6 x 15



Hypothesis testing?

- In the past, QTL mapping has been regarded as a task of hypothesis testing.

Is this a QTL?

Much of the focus has been on adjusting for test multiplicity.

- It is better to view the problem as one of model selection.

What set of QTL are well supported?

Is there evidence for QTL-QTL interactions?

Model = a defined set of QTL and QTL-QTL interactions
(and possibly covariates and QTL-covariate interactions).

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
- 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:

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

- Piece together the putative QTL from initial exploration
 - 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

Automation

- Assistance to the masses
- Understanding performance
- Many phenotypes

Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

$$y = \mu + \sum \beta_j \mathbf{q}_j + \epsilon \quad \text{which } \beta_j \neq 0?$$

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

Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

$$y = \mu + \sum \beta_j \mathbf{q}_j + \epsilon \quad \text{which } \beta_j \neq 0?$$

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

$$0 \text{ vs } 1 \text{ QTL: } \text{pLOD}(\emptyset) = 0$$

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

Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

$$y = \mu + \sum \beta_j \mathbf{q}_j + \epsilon \quad \text{which } \beta_j \neq 0?$$

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

For the mouse genome:

$$\mathbf{T} = 2.69 \text{ (BC) or } 3.52 \text{ (F}_2\text{)}$$

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

Epistasis

$$y = \mu + \sum \beta_j \mathbf{q}_j + \sum \gamma_{jk} \mathbf{q}_j \mathbf{q}_k + \epsilon$$

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

T_m = as chosen previously

T_i = ?

Idea 1

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

$$T_i = 95\text{th percentile of the distribution of} \\ \max \text{LOD}_f(s, t) - \max \text{LOD}_a(s, t)$$

Idea 1

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

$$T_i = 95\text{th percentile of the distribution of} \\ \max \text{LOD}_f(s, t) - \max \text{LOD}_a(s, t)$$

For the mouse genome:

$$T_m = 2.69 \text{ (BC) or } 3.52 \text{ (F}_2\text{)}$$

$$T_i^H = 2.62 \text{ (BC) or } 4.28 \text{ (F}_2\text{)}$$

Idea 2

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

$$T_m + T_i = 95\text{th percentile of the distribution of} \\ \max \text{LOD}_f(s, t) - \max \text{LOD}_1(s)$$

Idea 2

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

$$T_m + T_i = 95\text{th percentile of the distribution of} \\ \max \text{LOD}_f(s, t) - \max \text{LOD}_1(s)$$

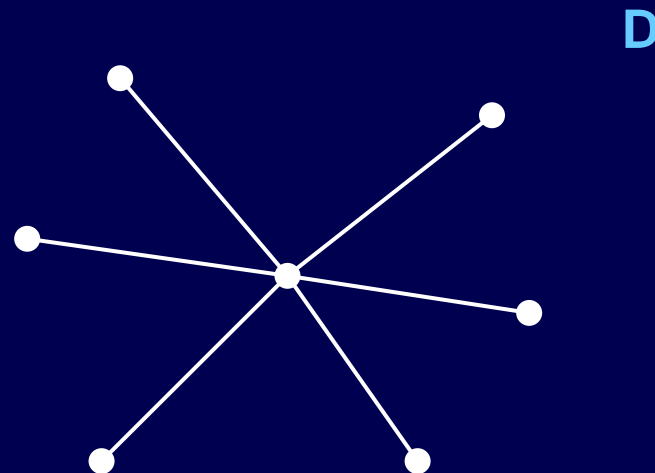
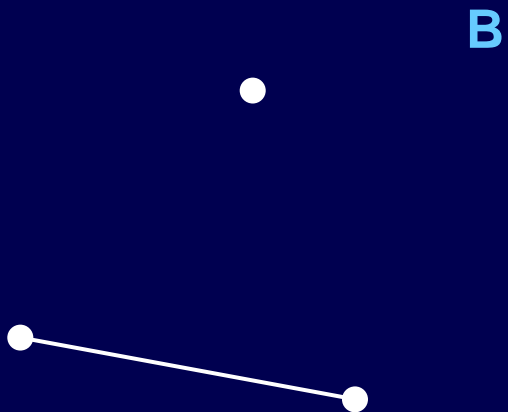
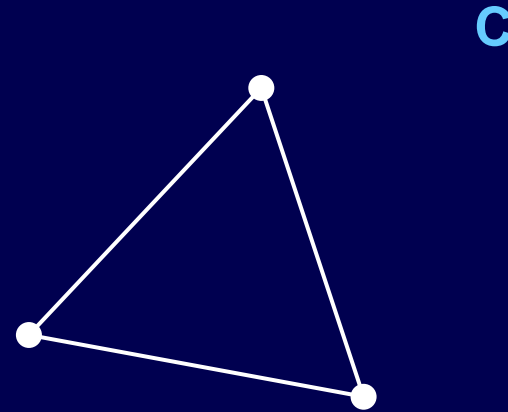
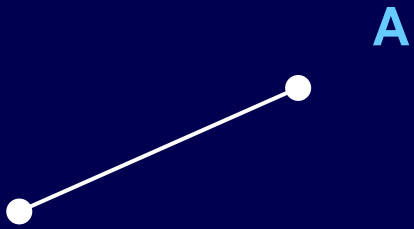
For the mouse genome:

$$T_m = 2.69 \text{ (BC) or } 3.52 \text{ (F}_2\text{)}$$

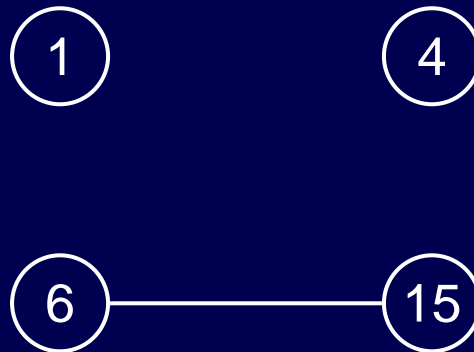
$$T_i^H = 2.62 \text{ (BC) or } 4.28 \text{ (F}_2\text{)}$$

$$T_i^L = 1.19 \text{ (BC) or } 2.69 \text{ (F}_2\text{)}$$

Models as graphs

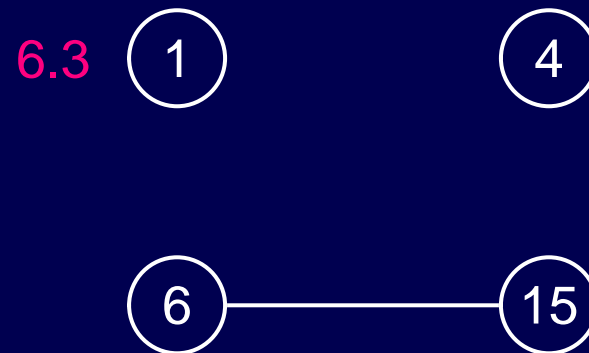


Results



LOD = 23.1

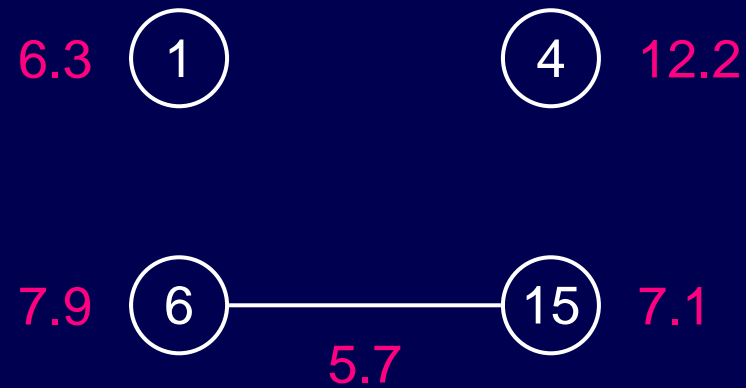
Results



LOD = 23.1

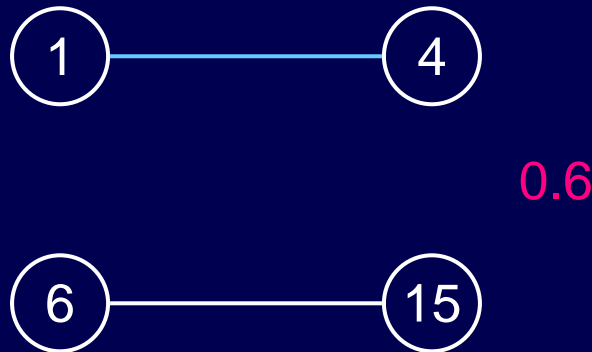
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Results



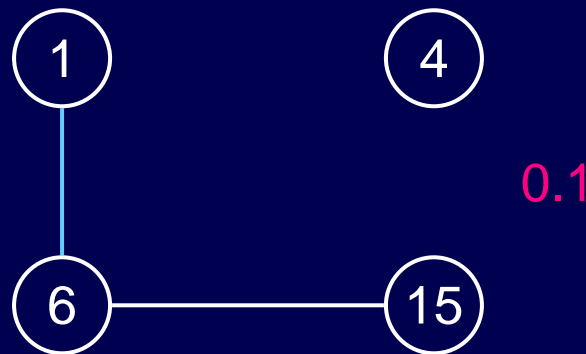
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add an interaction?



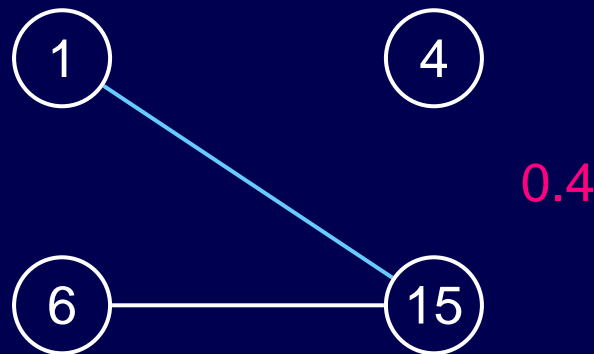
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add an interaction?



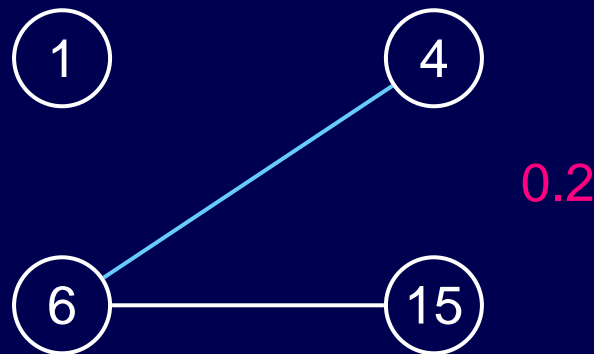
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add an interaction?



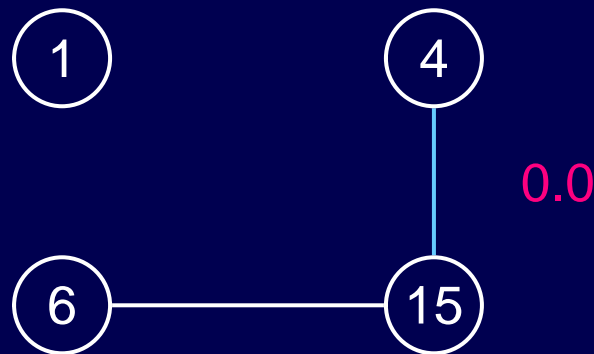
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add an interaction?



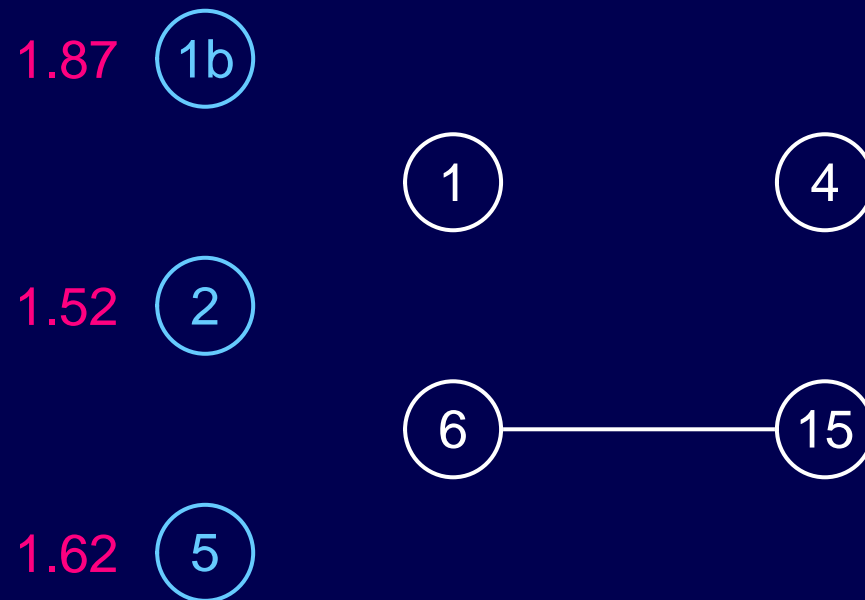
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add an interaction?



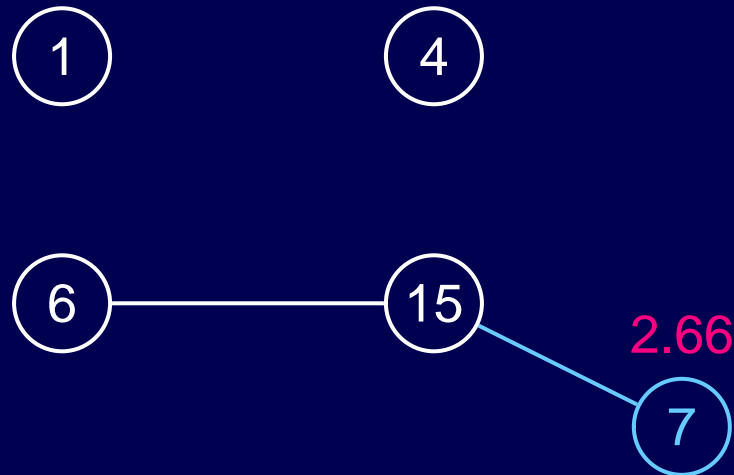
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add another QTL?



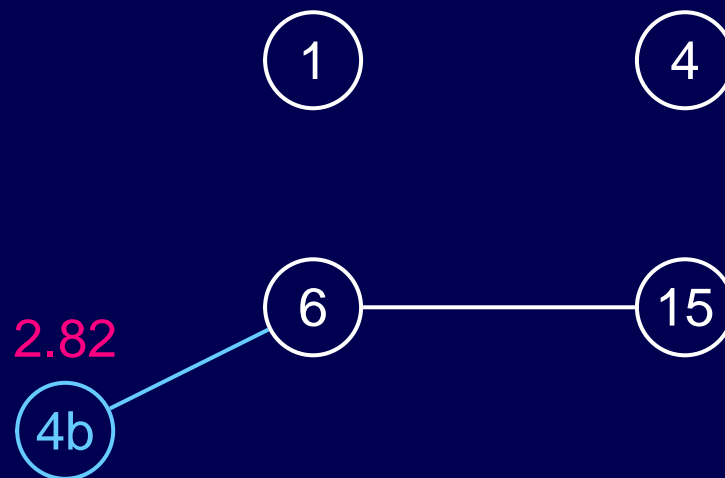
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add another QTL?



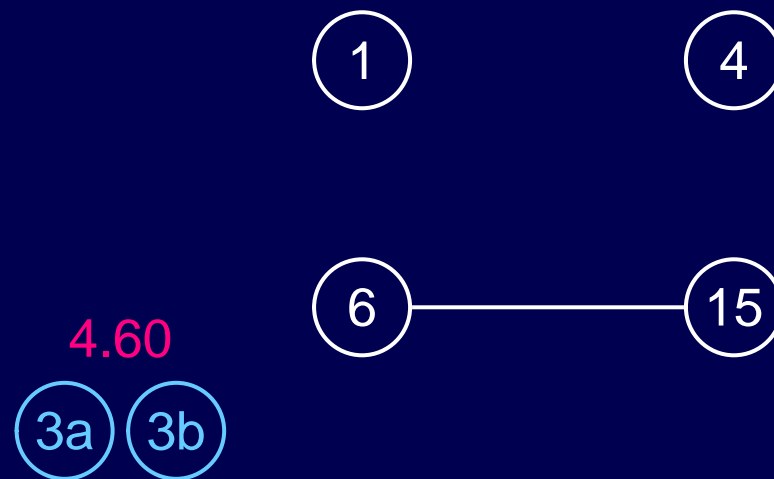
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add another QTL?



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

Add a pair of QTL?



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

Open problems

- Improve search procedures
- Measuring model uncertainty
- Measuring uncertainty in QTL location

Open problems

- Improve search procedures
- Measuring model uncertainty
- Measuring uncertainty in QTL location
- Multi-parent populations
- High-throughput phenotypes (e.g. expression, proteins, microbiome)
- QTL \times environment interactions

Summary

- QTL mapping is a model selection problem
- The criterion for comparing models is most important
- I've been focusing on a penalized likelihood method and have a reasonably practiceable solution
- Broman & Speed, JRSS B 64:641-656, 2002
Manichaikul et al., Genetics 181:1077–1086, 2009