QTL mapping in experimental crosses

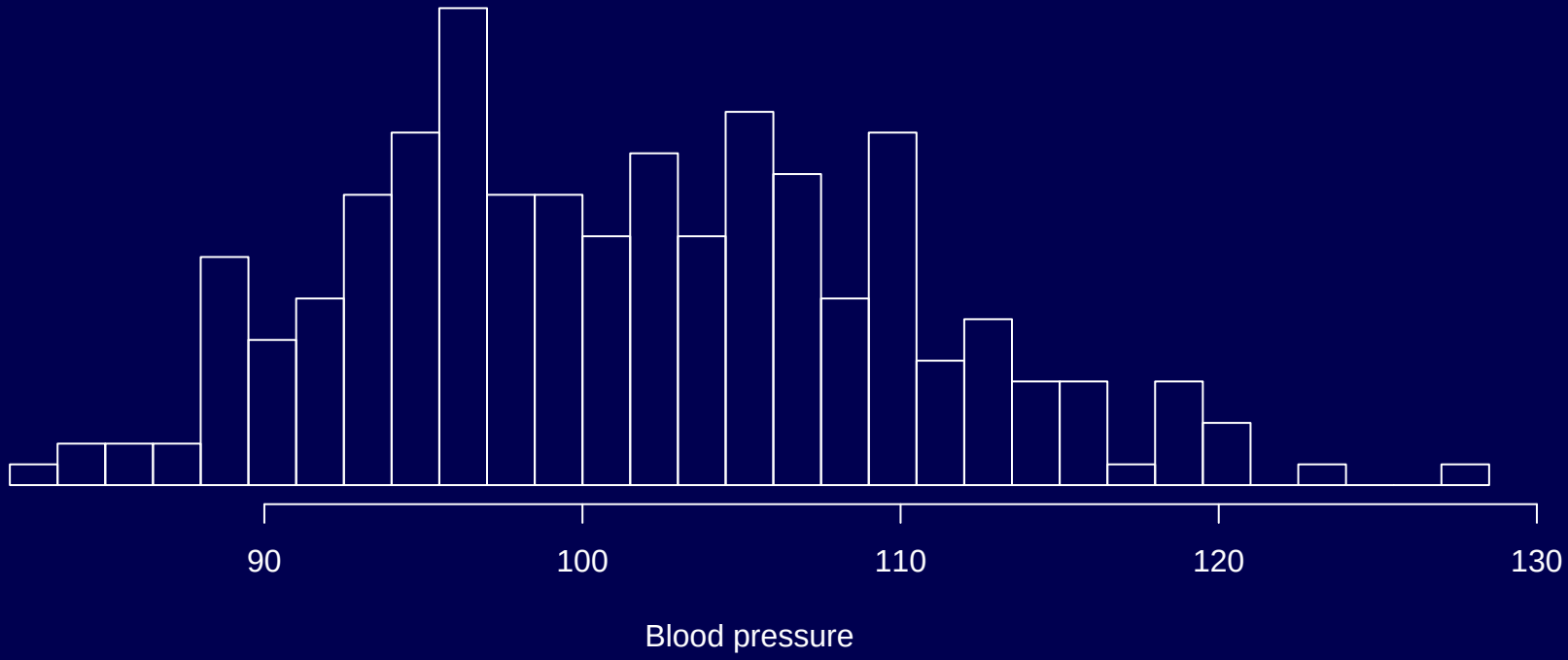
Part II

Karl W Broman
kbroman.org
github.com/kbroman
@kwbroman

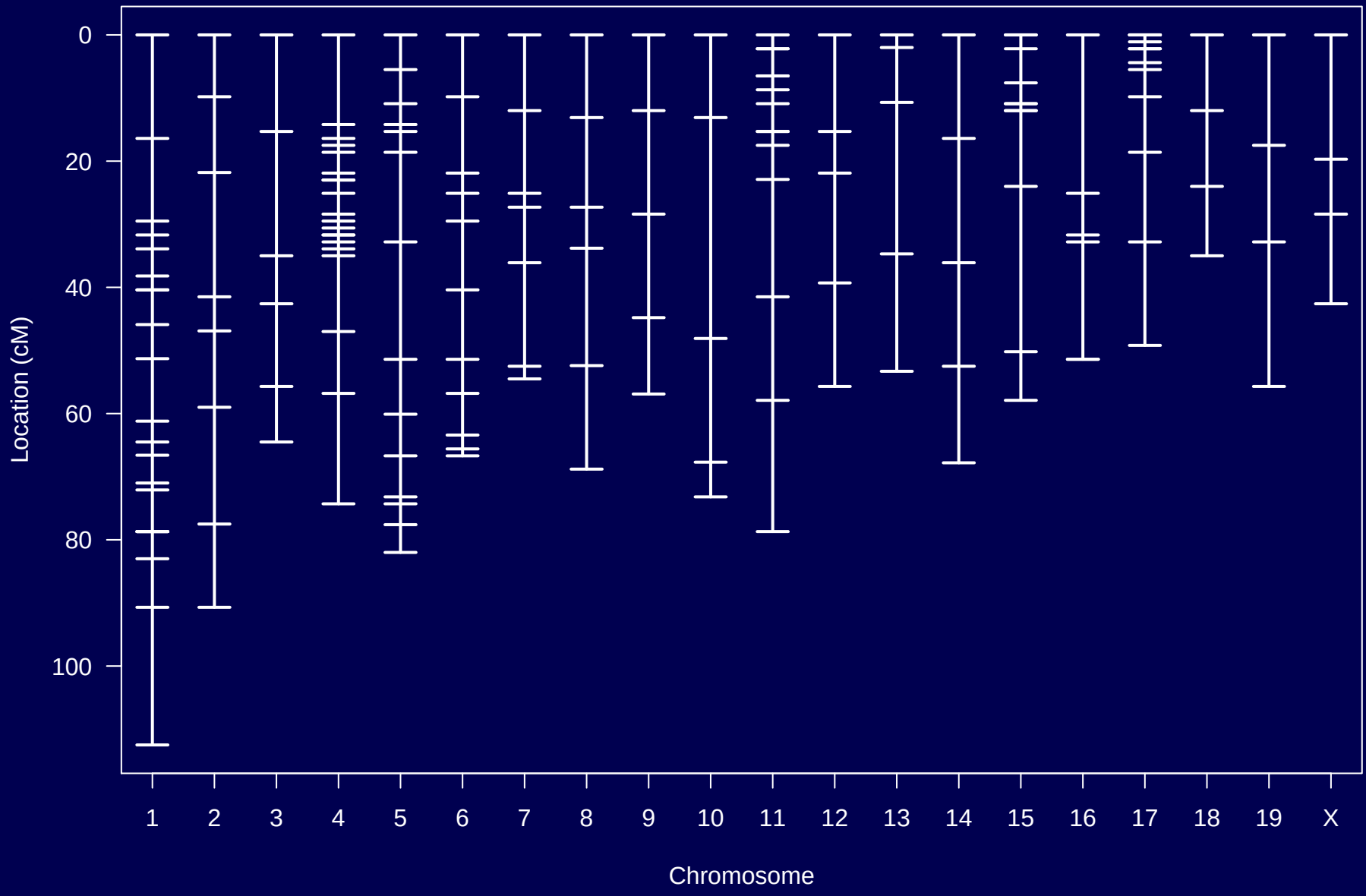
Example

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

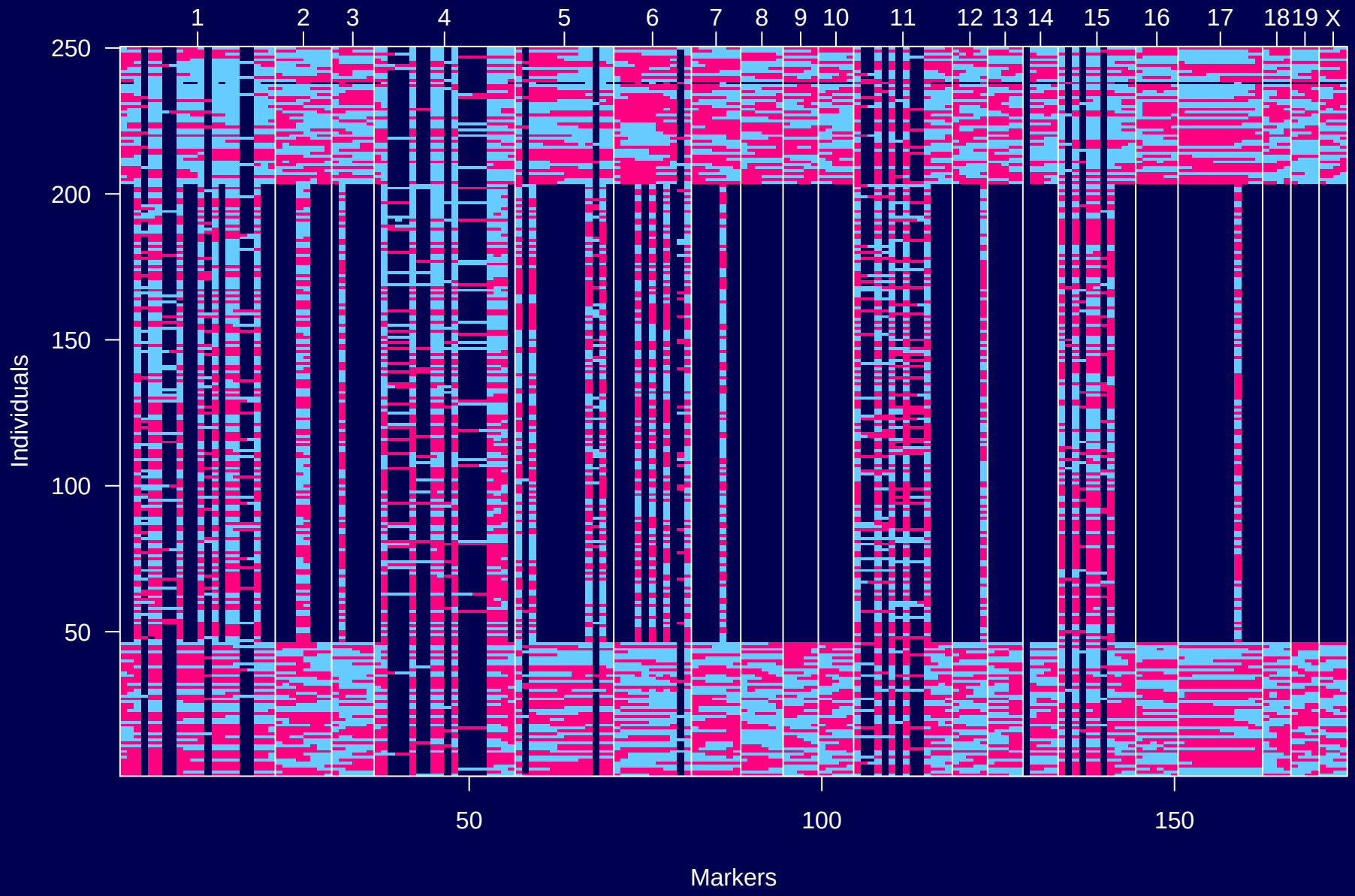
250 male mice from the backcross (A $\times$ B) $\times$ B
Blood pressure after two weeks drinking water with 1% NaCl



Genetic map


Location (cM)
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

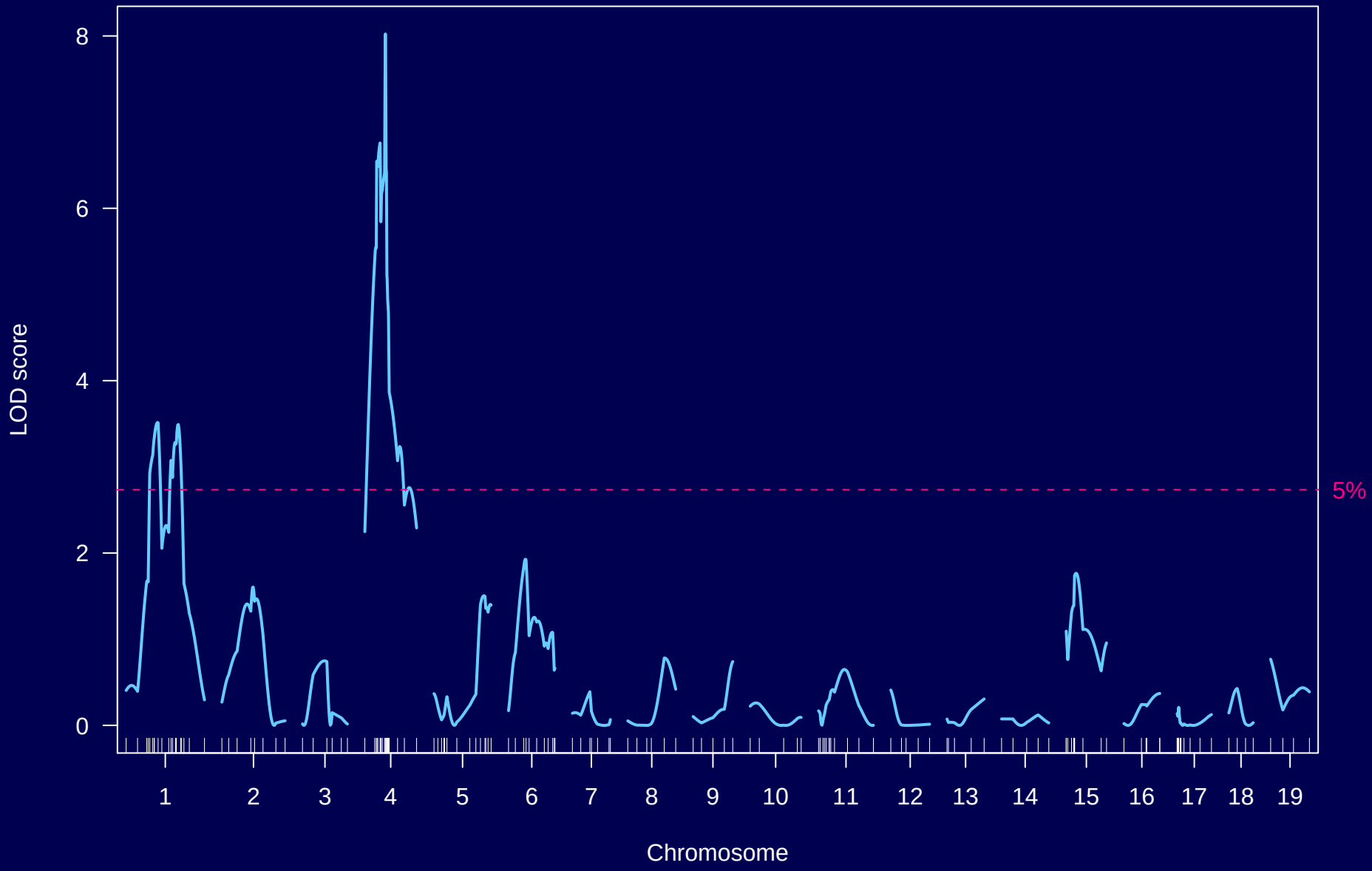
Genotype data



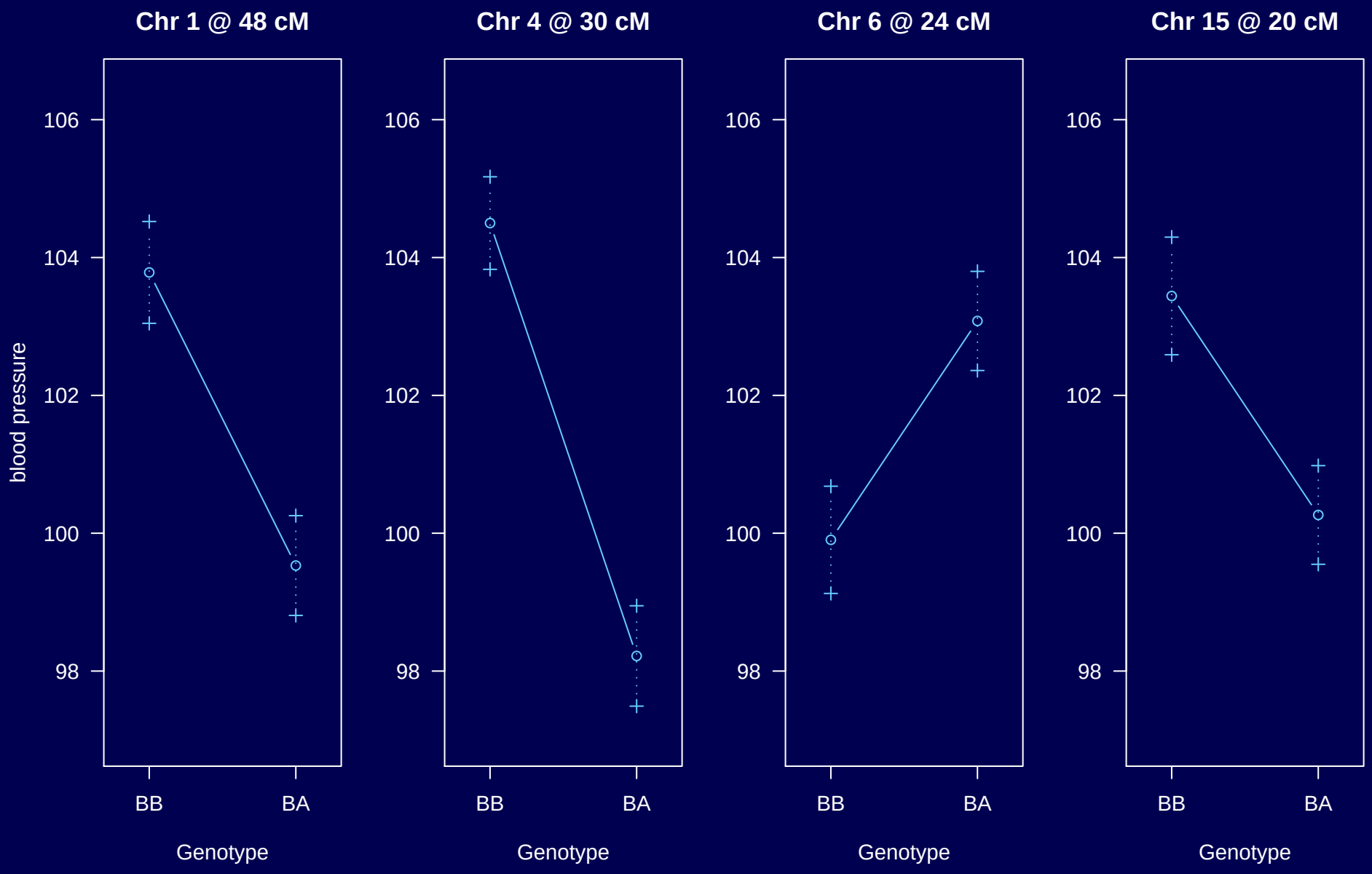
Goals

- 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

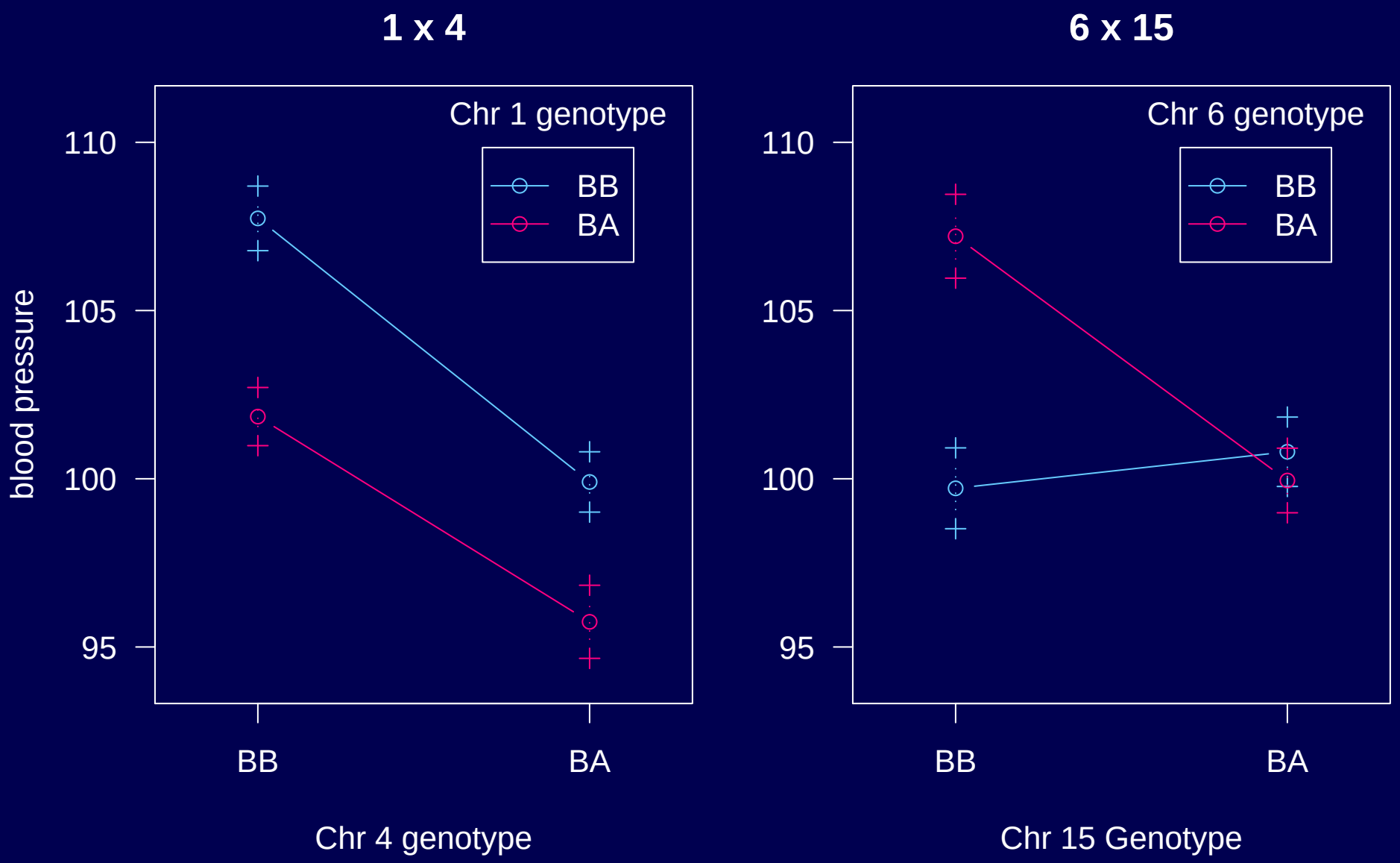
Estimated effects



Modeling multiple QTL

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

Estimated effects



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:

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

- 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 non-specialists
- Understanding performance
- Many phenotypes

Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

$$\mathsf{y} = \mu + \sum \beta_{\mathsf{j}}\, \mathsf{q}_{\mathsf{j}} + \epsilon \qquad \text{which } \beta_{\mathsf{j}} \neq 0?$$

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

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

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:

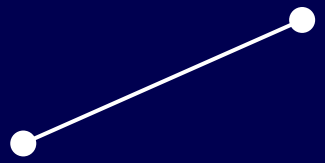
T_m = 2.69 (BC) or 3.52 (F_2)

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

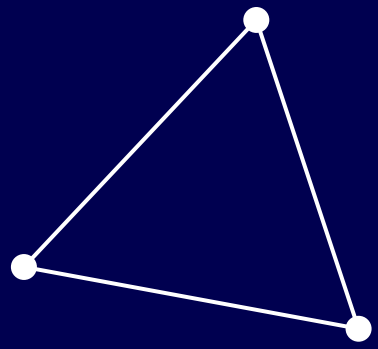
T_i^L = 1.19 (BC) or 2.69 (F_2)

Models as graphs

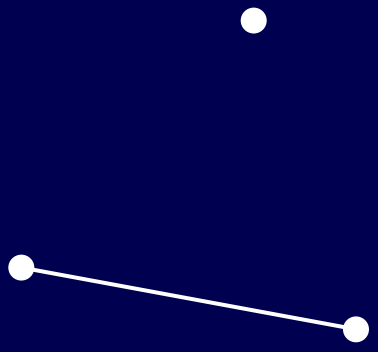
A



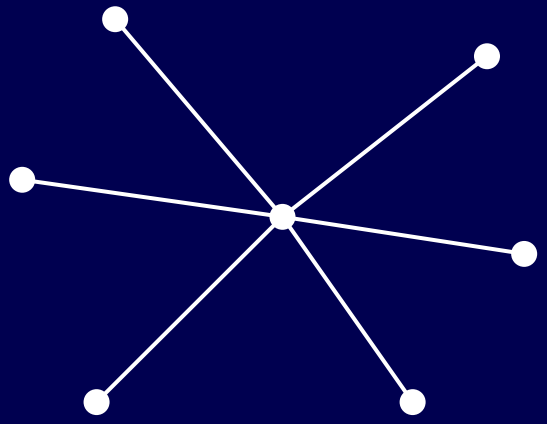
C



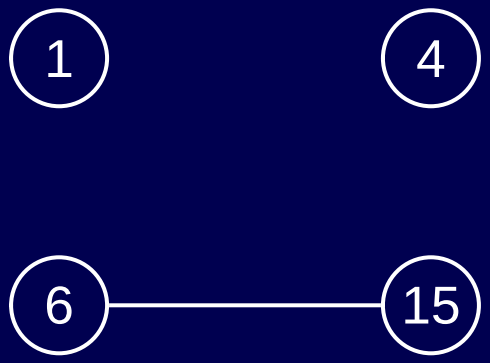
B



D

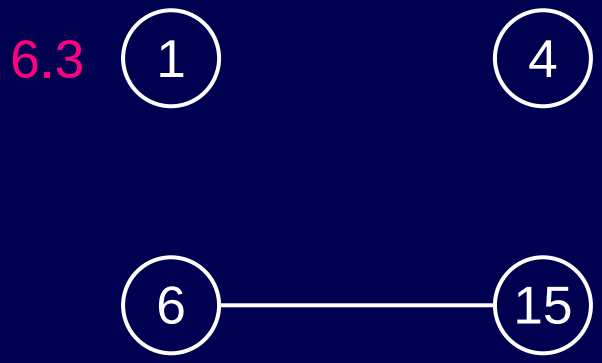


Results



LOD = 23.1

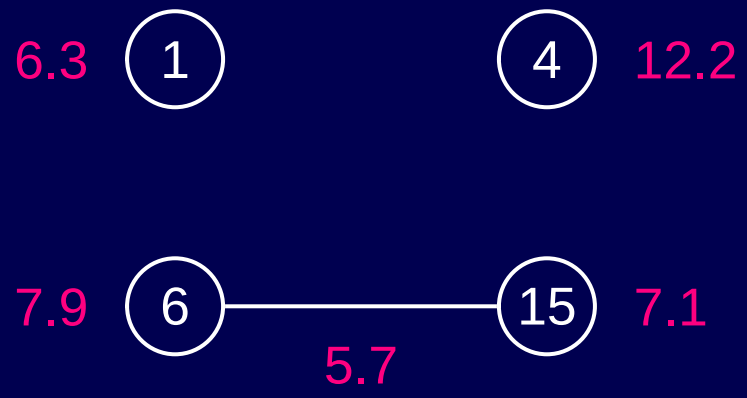
Results



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

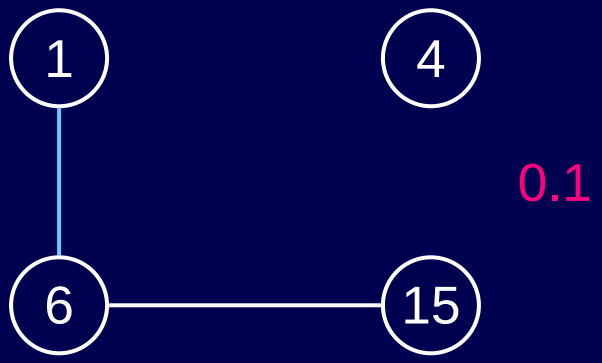


$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

Add an interaction?

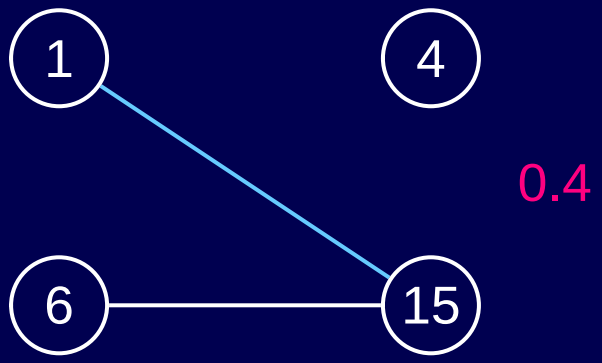
$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

Add an interaction?



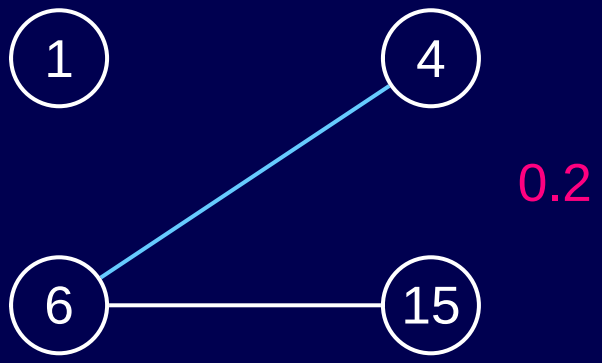
$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

Add an interaction?



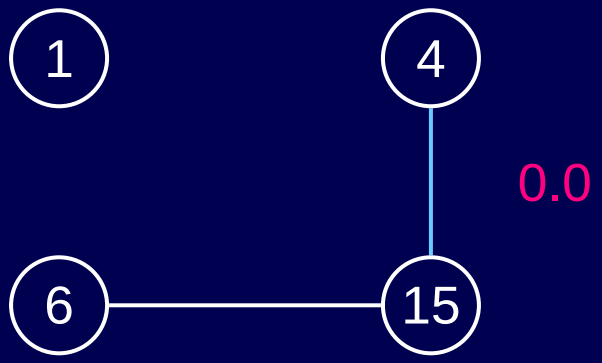
$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

Add an interaction?



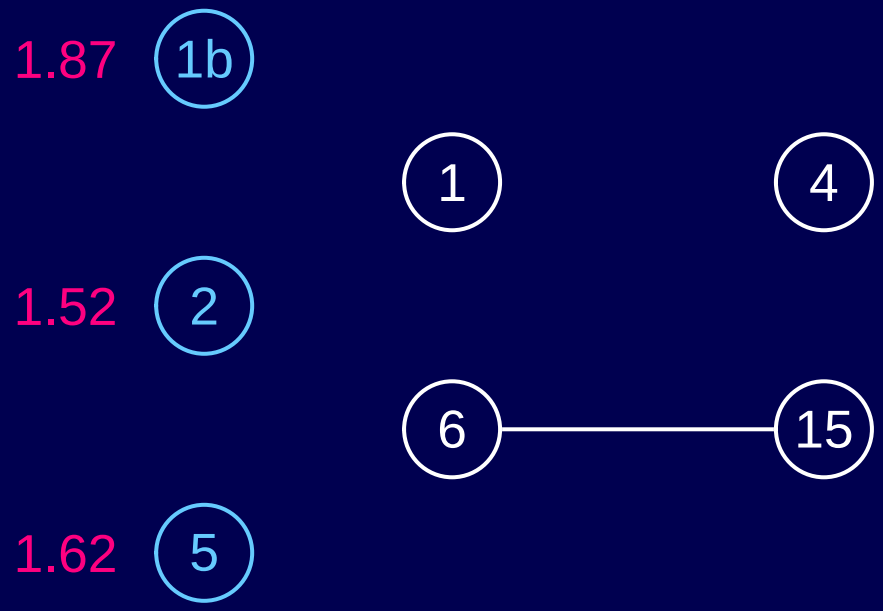
$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

Add an interaction?



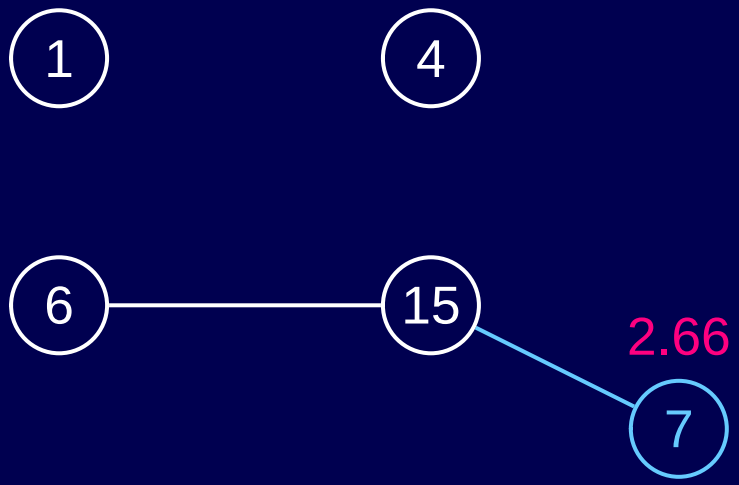
$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

Add another QTL?



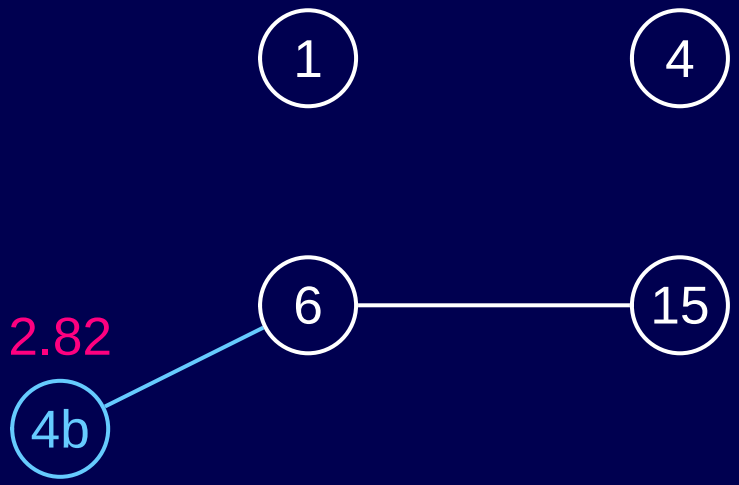
$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

Add another QTL?



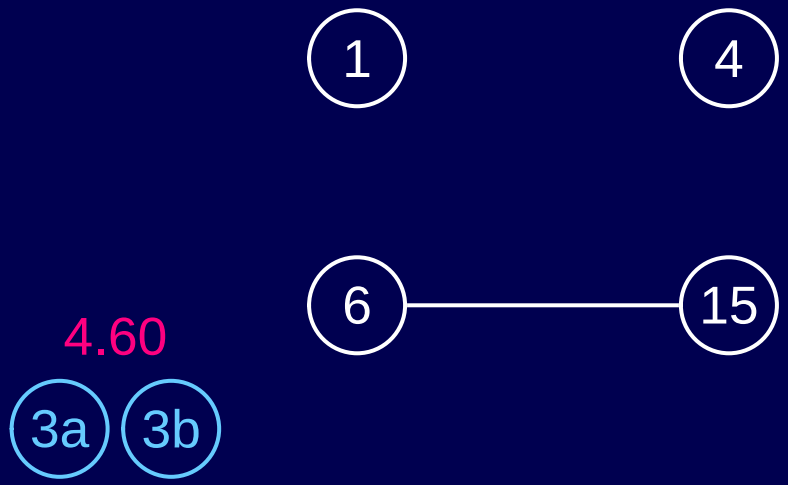
$T_m = 2.69$ $T_i^H = 2.62$ $T_i^L = 1.19$ $T_m + T_i^H = 5.31$ $T_m + T_i^L = 3.88$ $2T_m = 5.38$

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 a pair of QTL?



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

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