

# Mapping multiple QTL in experimental crosses

---

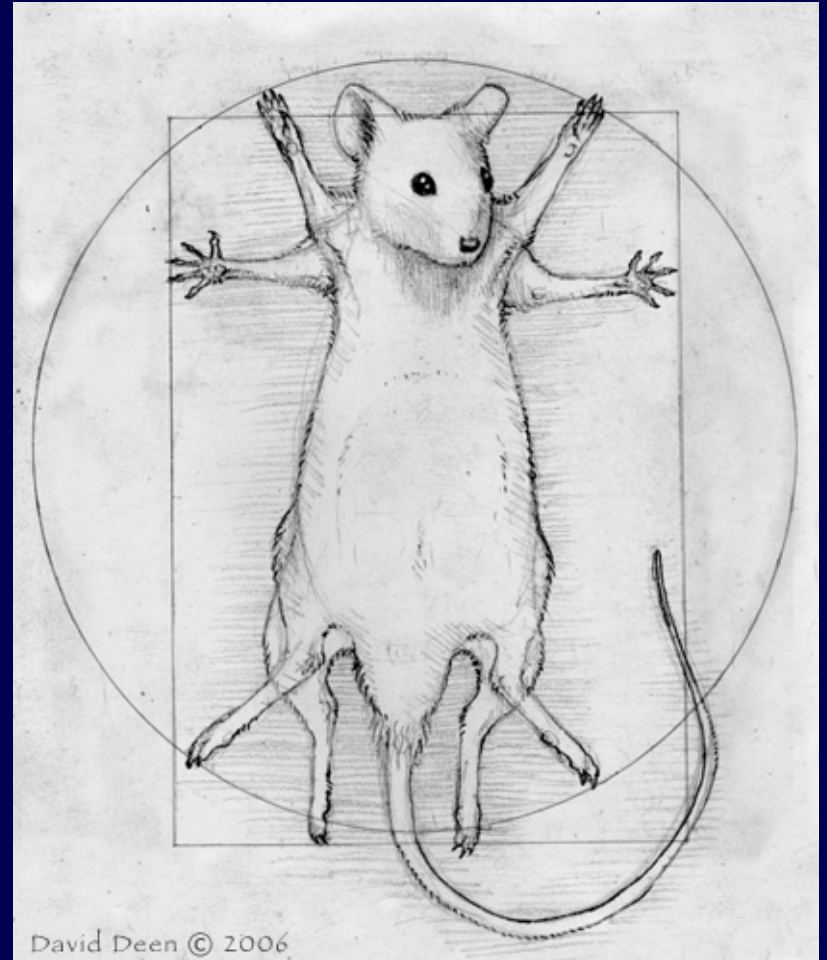
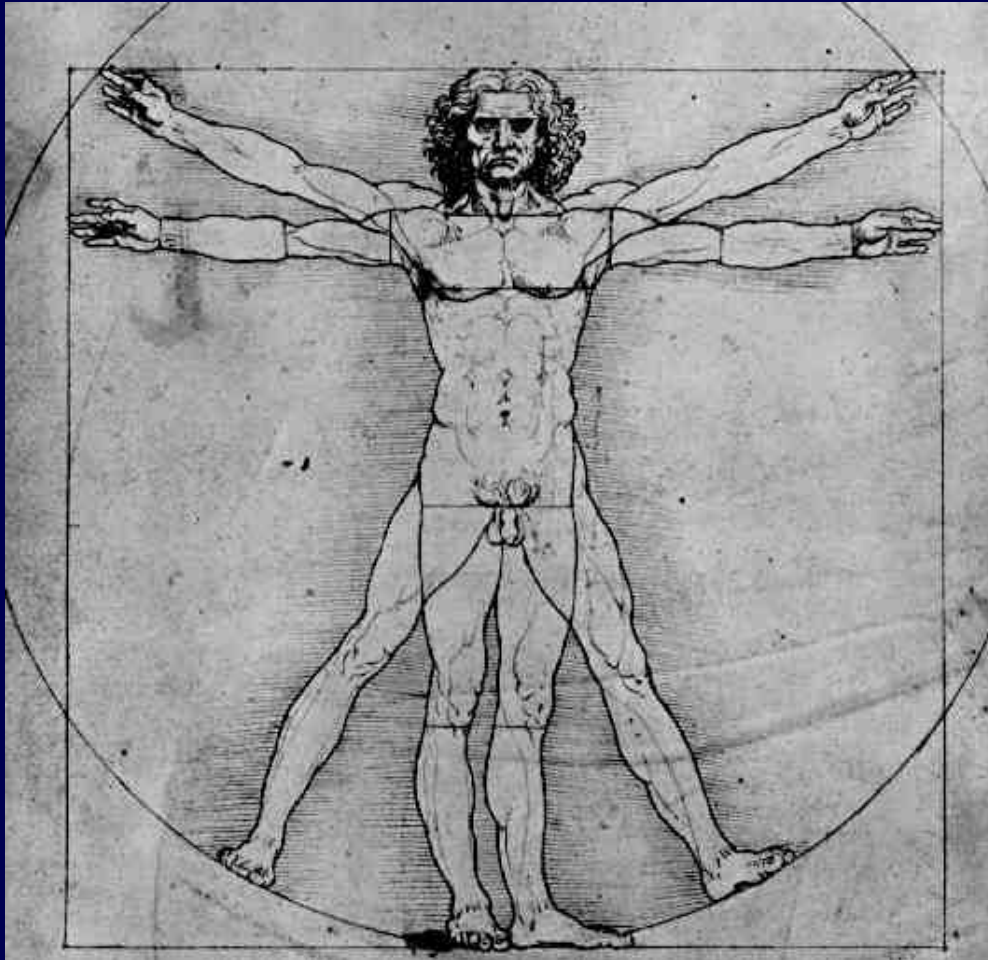
Karl W Broman

Department of Biostatistics & Medical Informatics  
University of Wisconsin – Madison

[www.biostat.wisc.edu/~kbroman](http://www.biostat.wisc.edu/~kbroman)

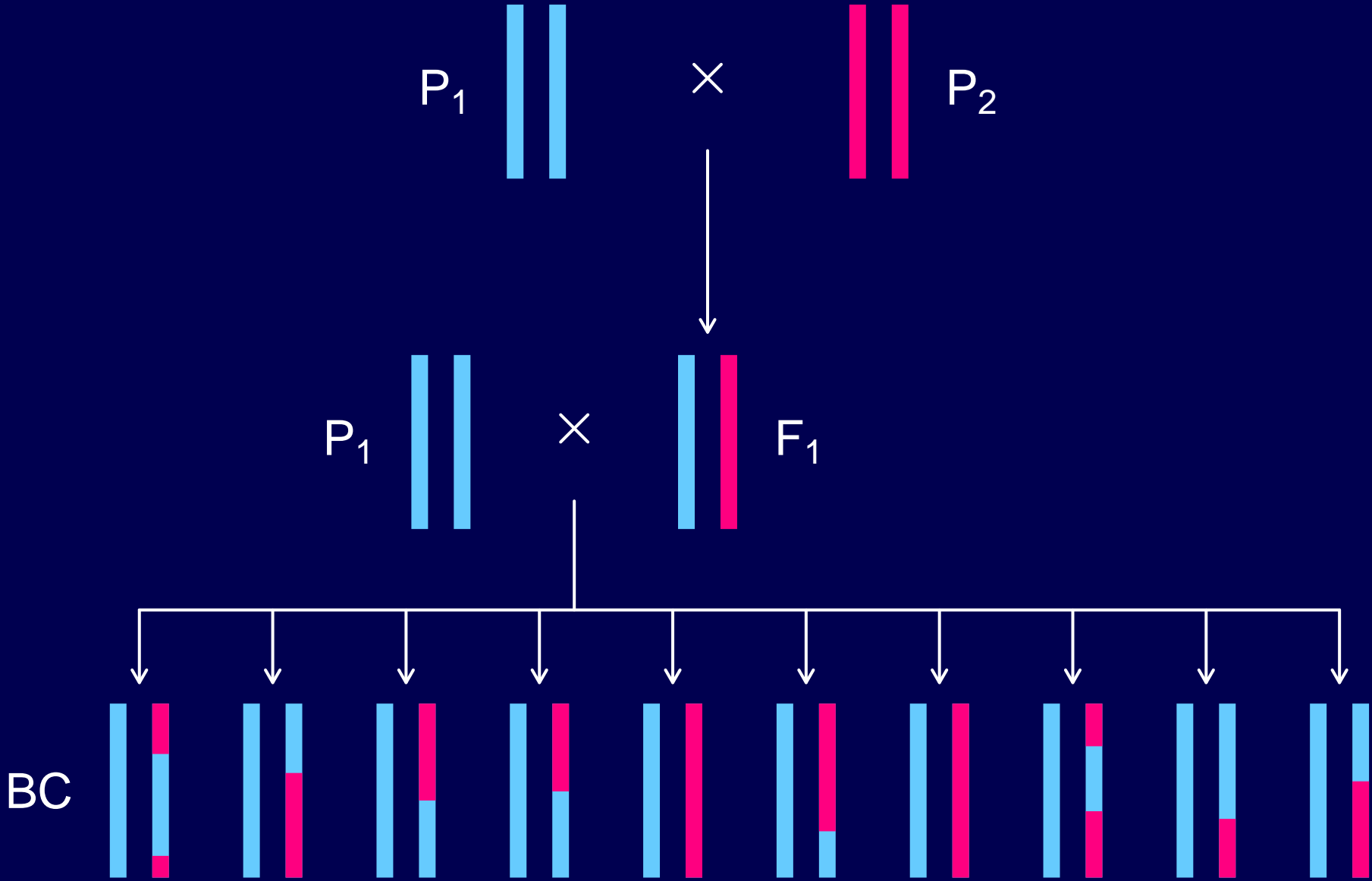


# Human vs mouse

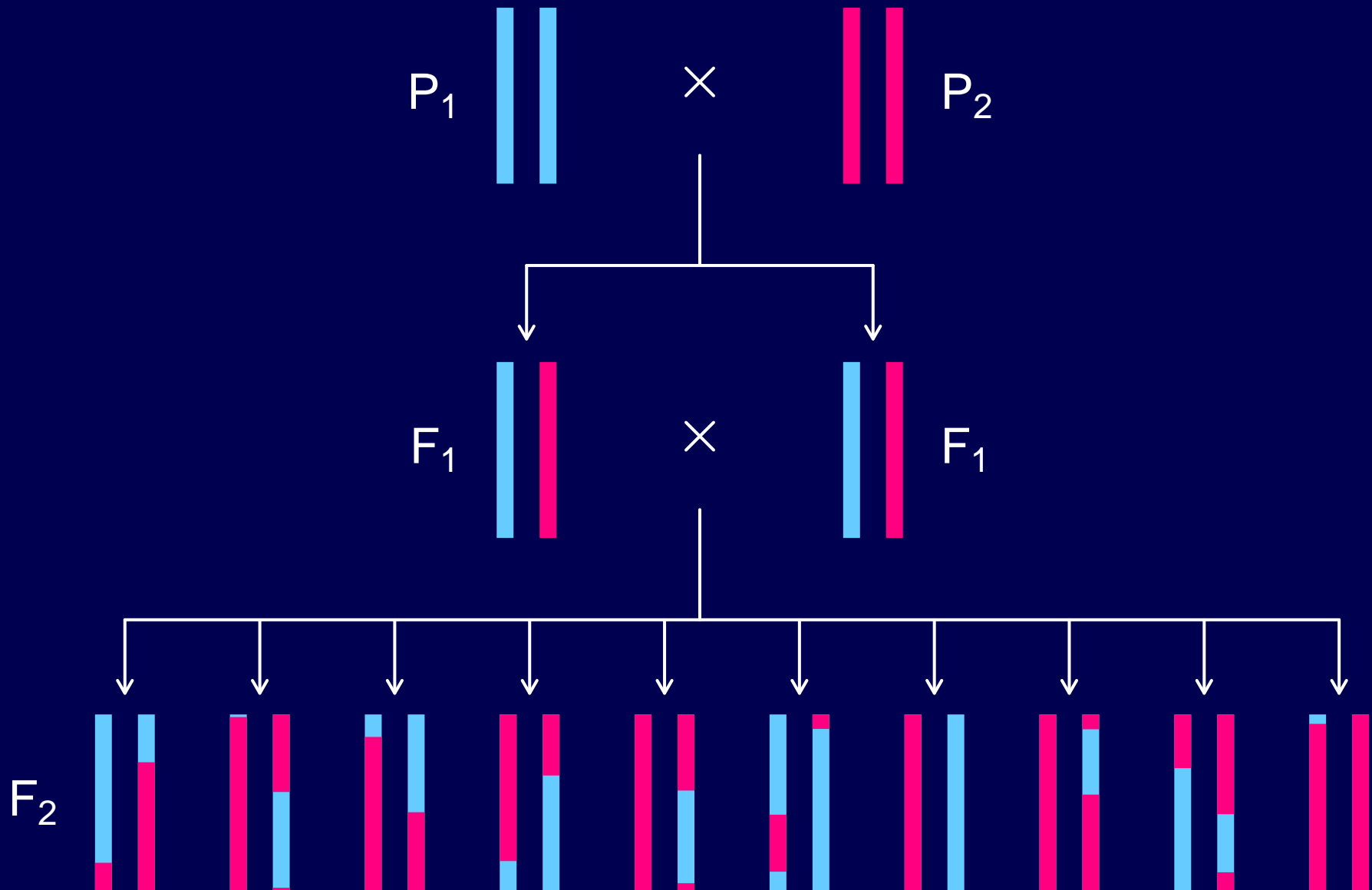


[www.daviddeen.com](http://www.daviddeen.com)

# Backcross



# Intercross

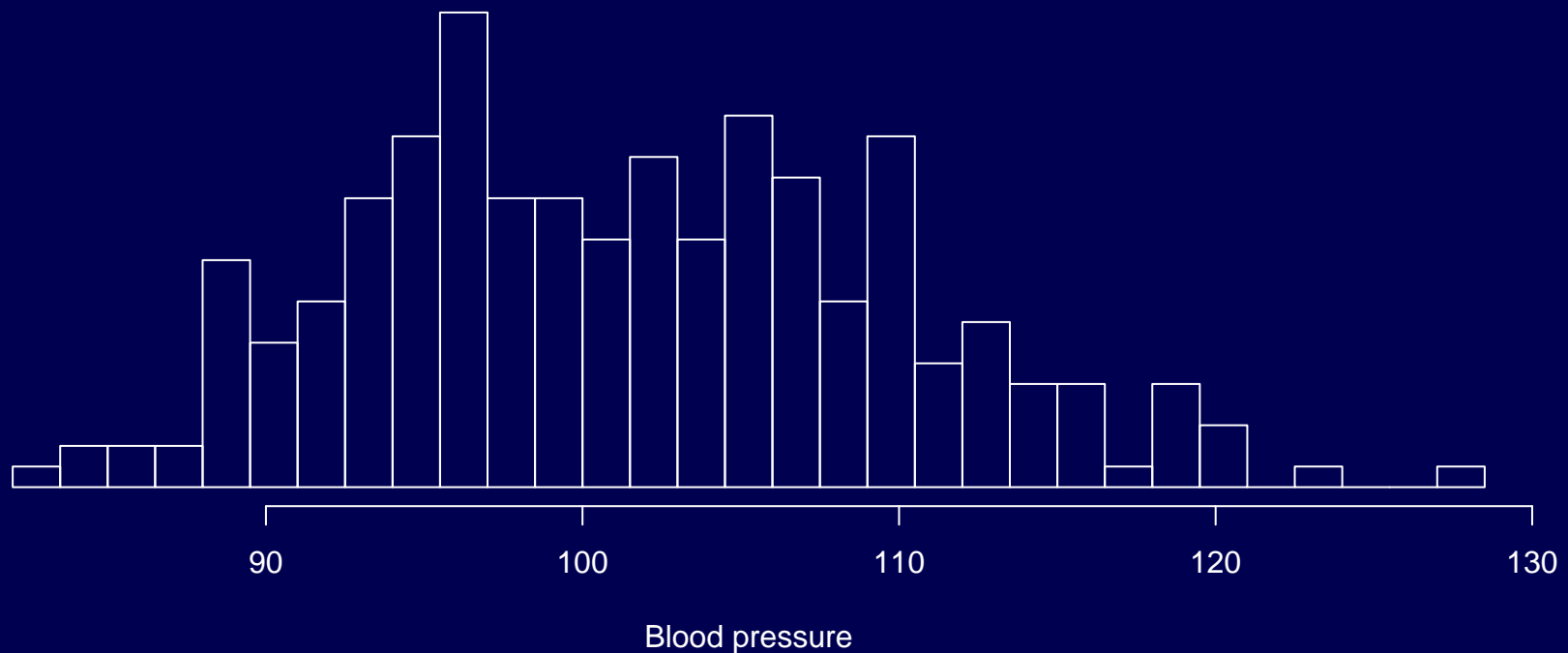


# Phenotype data

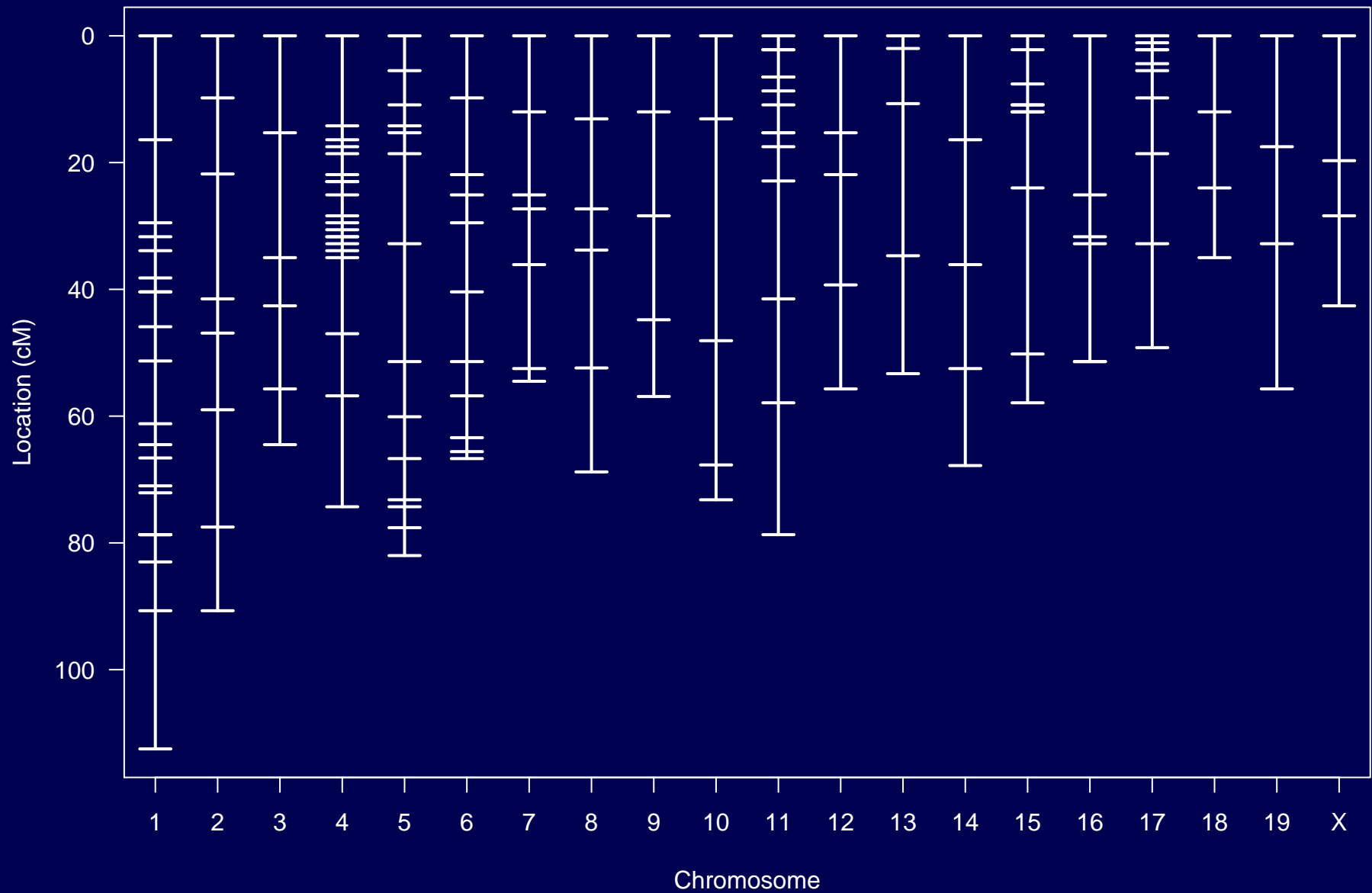
Sugiyama et al. Genomics 71:70-77, 2001

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

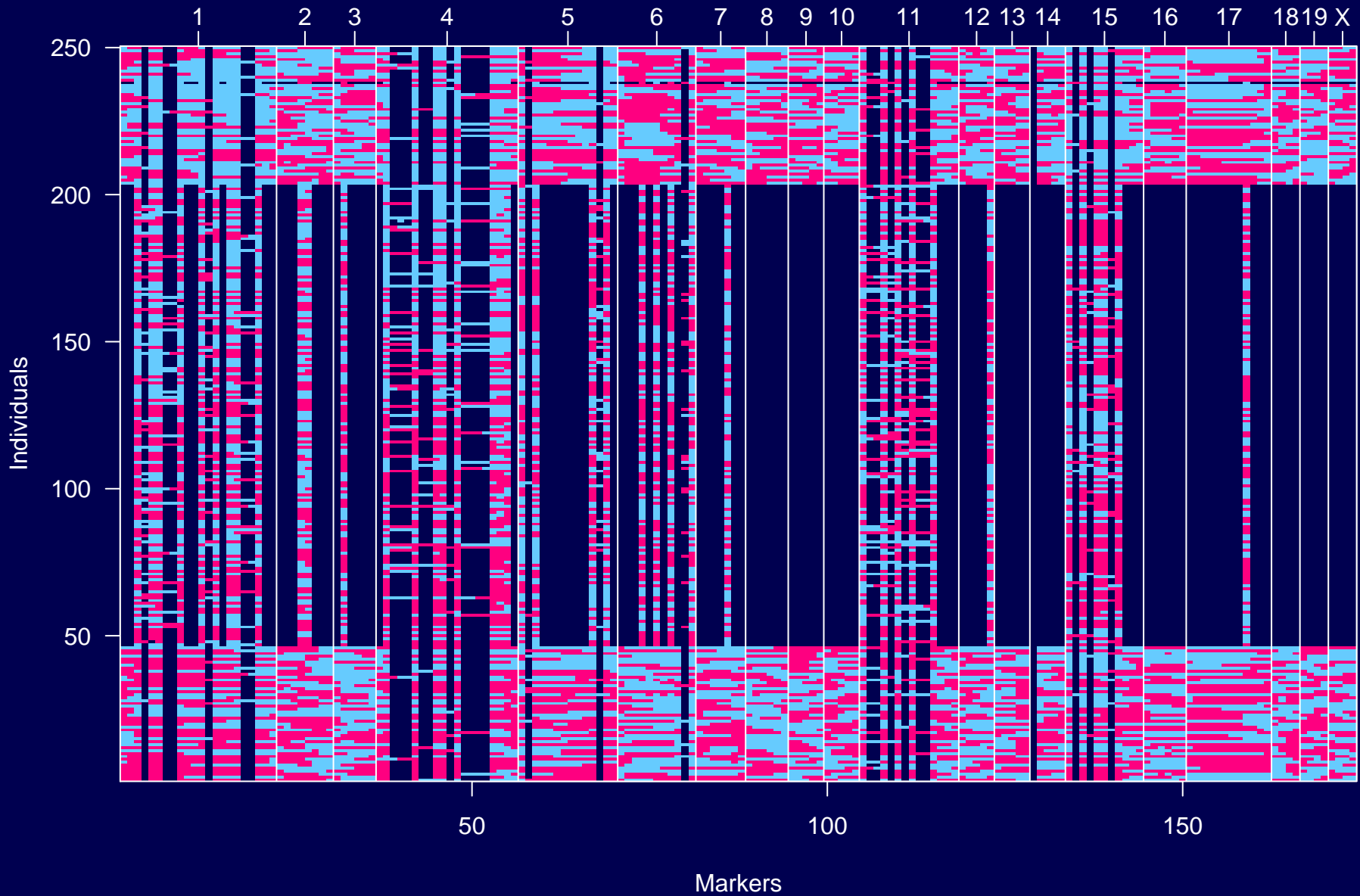
Blood pressure after two weeks drinking water with 1% NaCl



# Genetic map



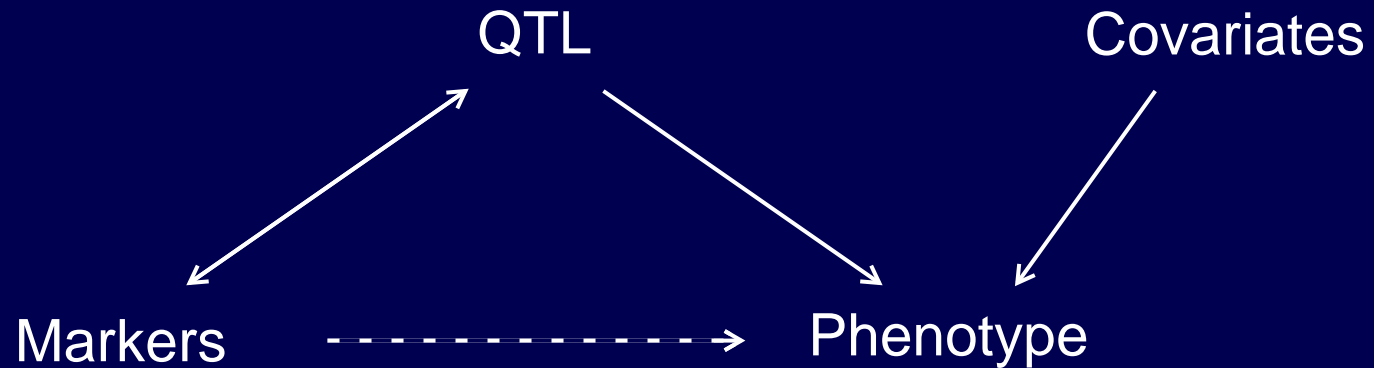
# Genotype data



# Goals

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

# Statistical structure



The missing data problem:

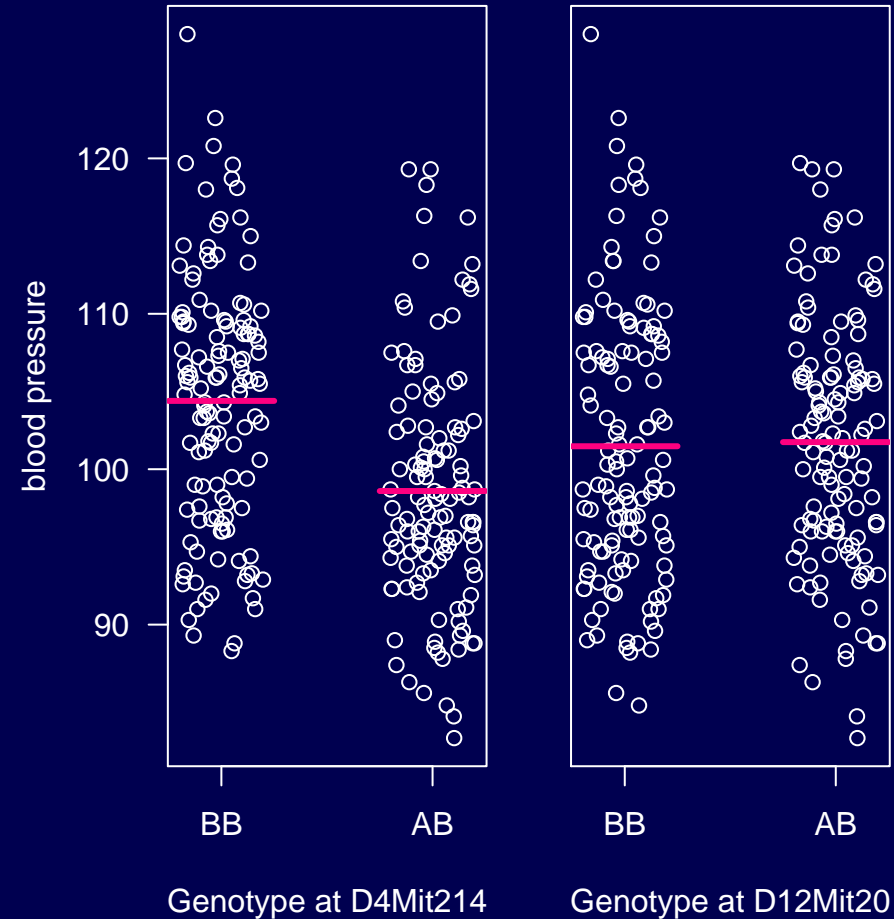
Markers  $\longleftrightarrow$  QTL

The model selection problem:

QTL, covariates  $\longrightarrow$  phenotype

# ANOVA at marker loci

- Split mice into groups according to genotype at a marker.
- Do a t-test / ANOVA.
- Repeat for each marker.



# Interval mapping

## Lander & Botstein (1989)

- Assume a **single** QTL model.
- Consider each position in the genome, one at a time, as the location of the putative QTL.
- Let  $q = 0/1$  if the (unobserved) QTL genotype is BB/AB.  
(Or 0/1/2 if the QTL genotype is AA/AB/BB in an intercross.)

Assume  $y \mid q \sim N(\mu_q, \sigma)$

- Calculate  $p_q = \Pr(q \mid \text{marker data})$ .

$$y \mid \text{marker data} \sim \sum_q p_q \phi(y \mid \mu_q, \sigma)$$

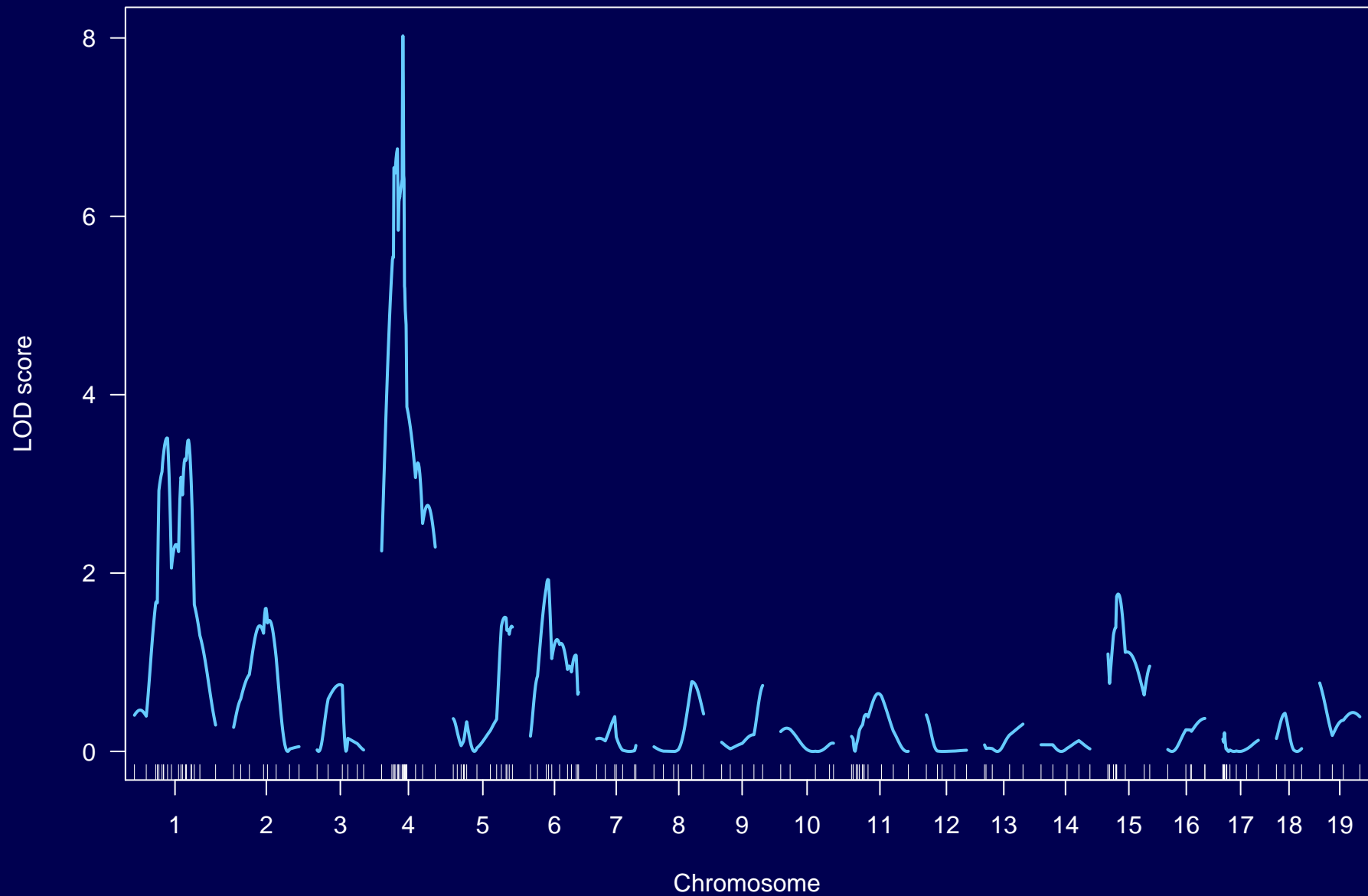
# LOD scores

$$\begin{aligned}\text{LOD}(\lambda) &= \log_{10} \text{likelihood ratio comparing the hypothesis of a} \\ &\quad \text{QTL at position } \lambda \text{ versus that of no QTL} \\ &= \log_{10} \left\{ \frac{\text{Pr}(y|\text{QTL at } \lambda, \hat{\mu}_{q\lambda}, \hat{\sigma}_\lambda)}{\text{Pr}(y|\text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\}\end{aligned}$$

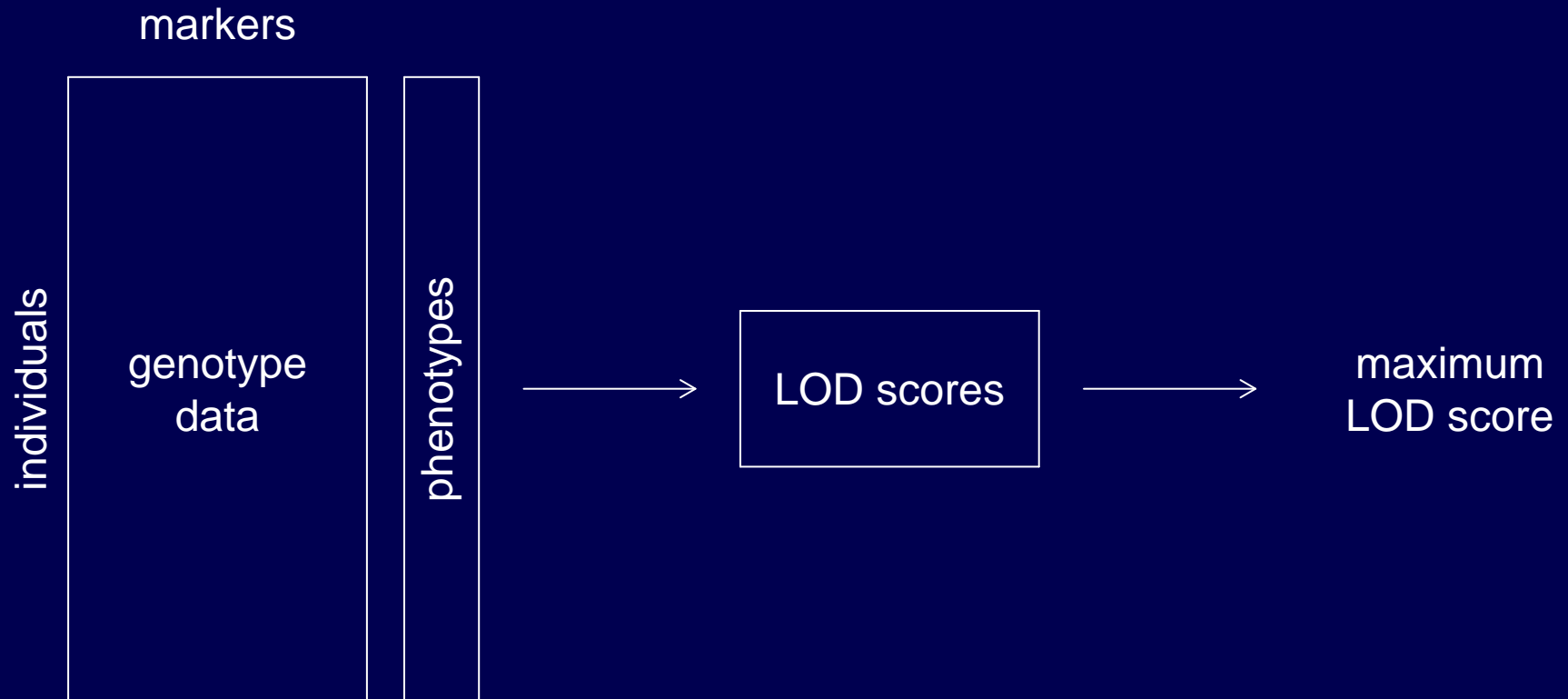
$\hat{\mu}_{q\lambda}, \hat{\sigma}_\lambda$  are the MLEs, assuming a single QTL at position  $\lambda$ .

No QTL model: The phenotypes are iid  $N(\mu, \sigma^2)$ .

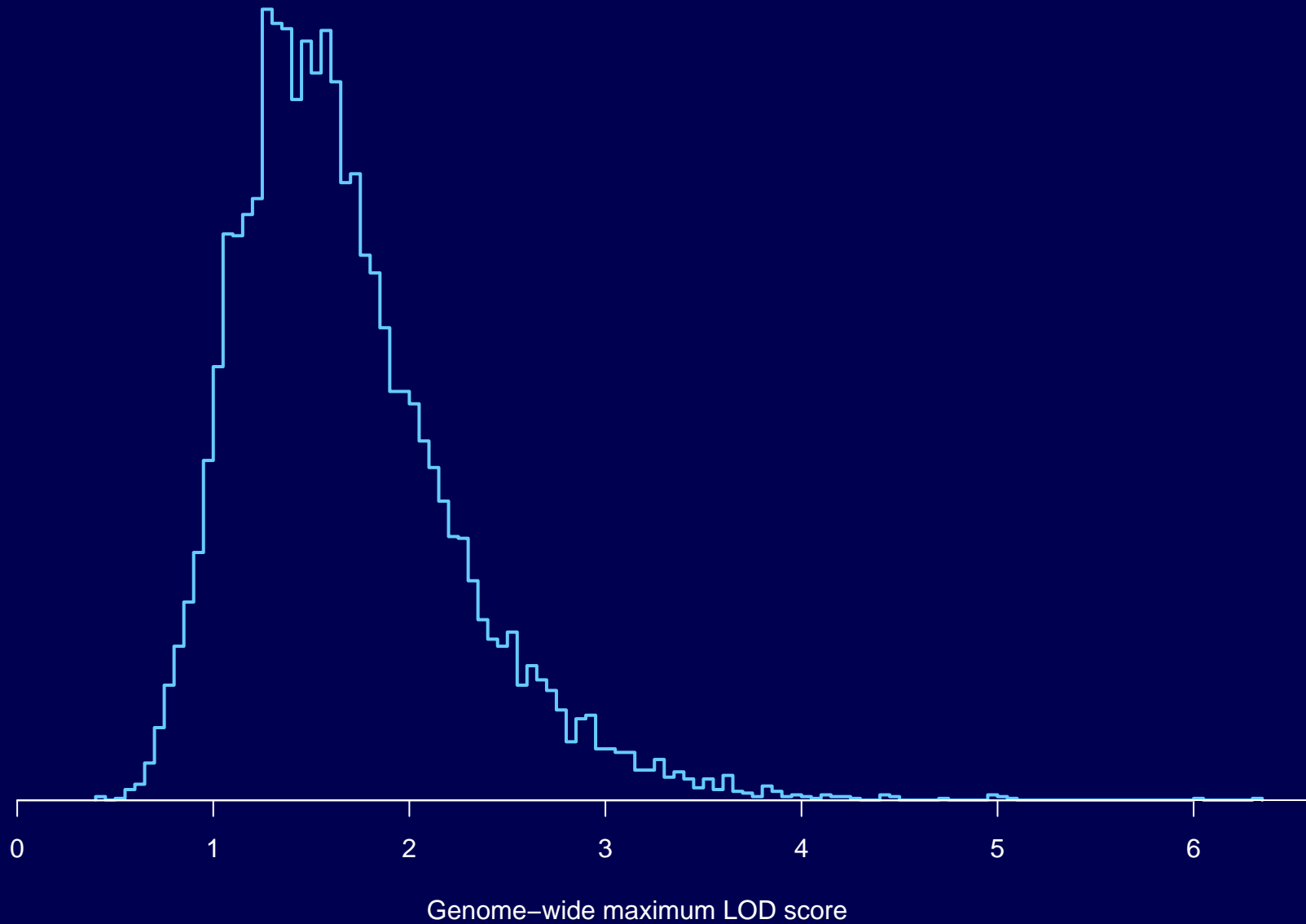
# LOD curves



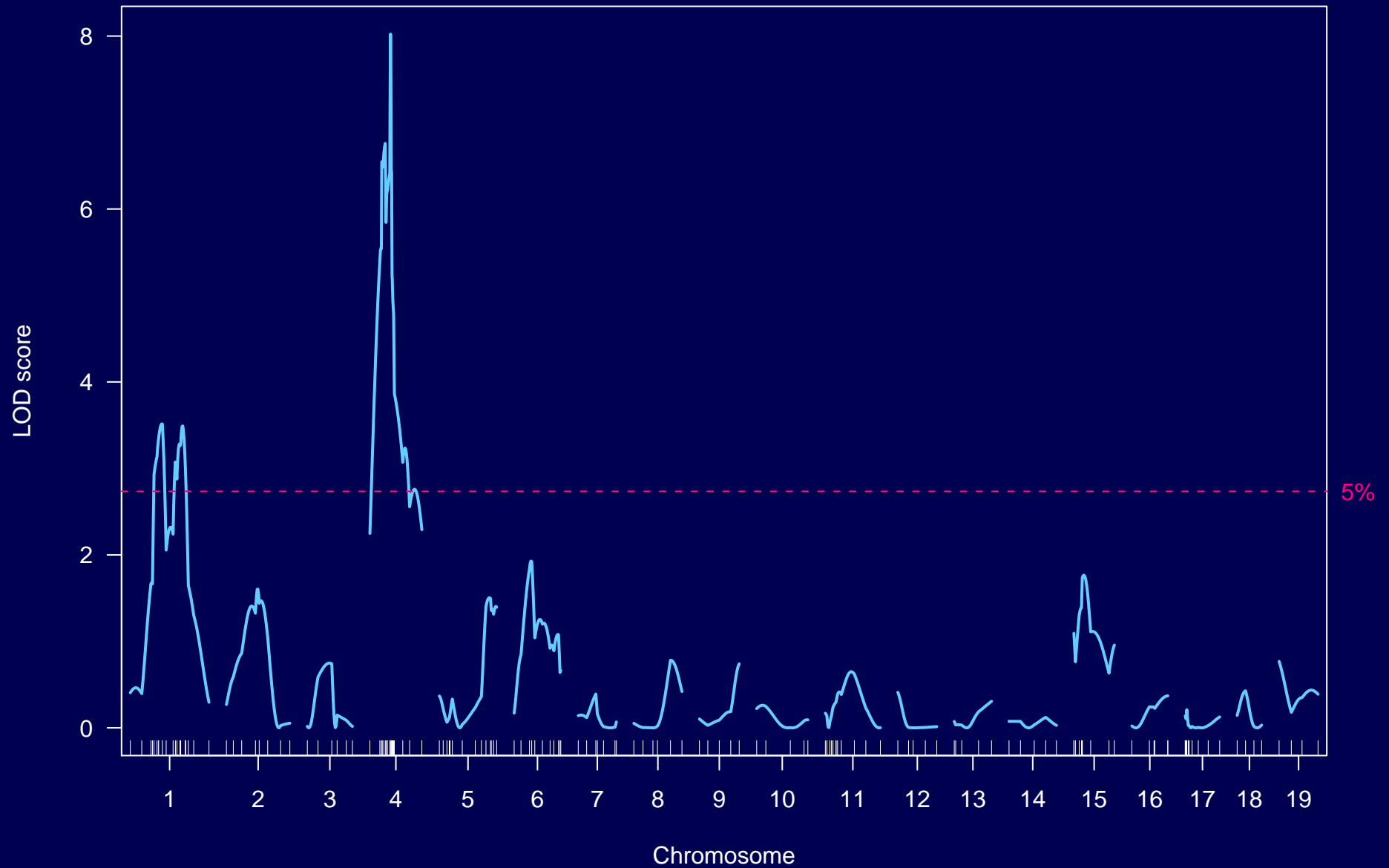
# Permutation test



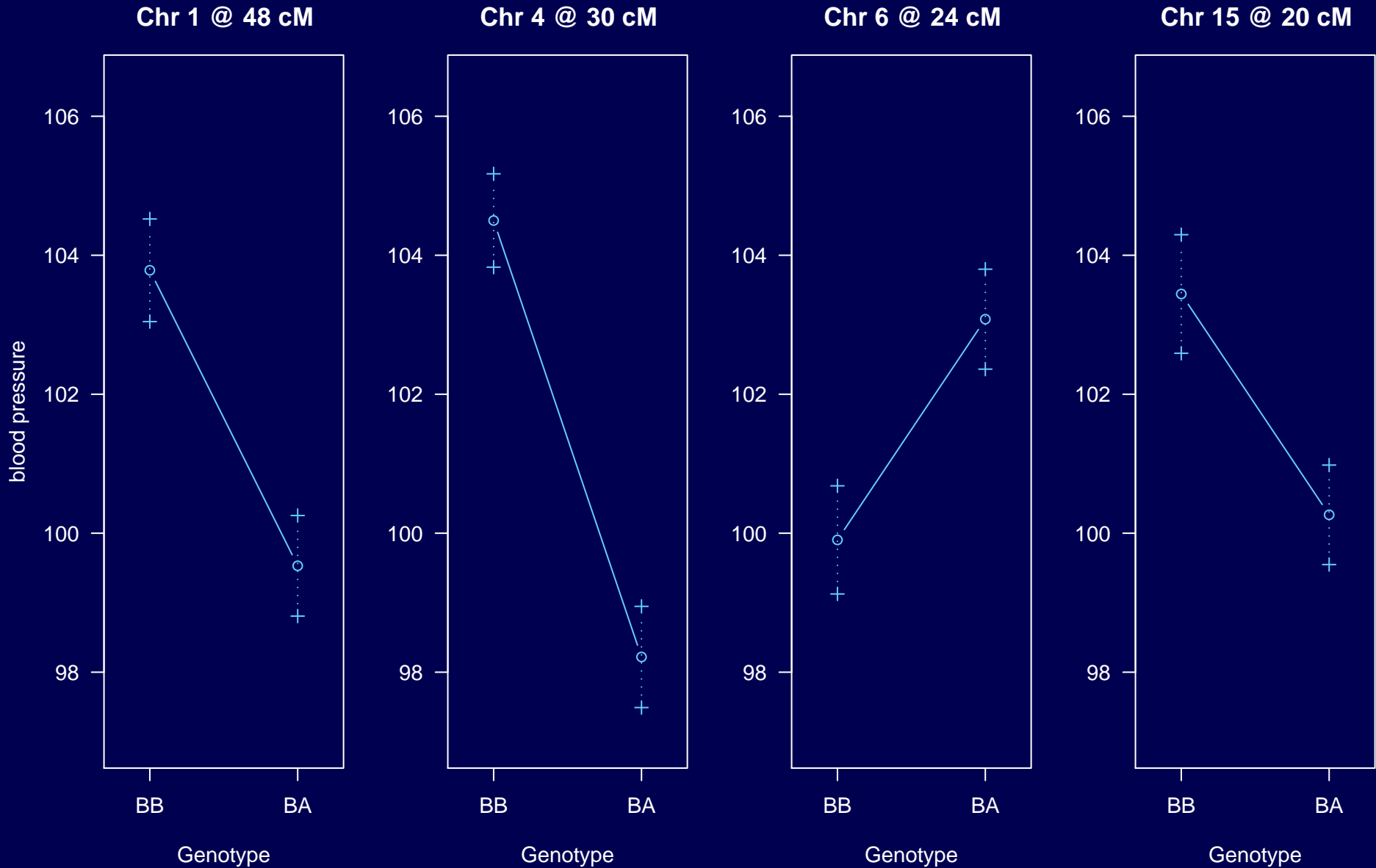
# Permutation results



# LOD curves



# Estimated effects

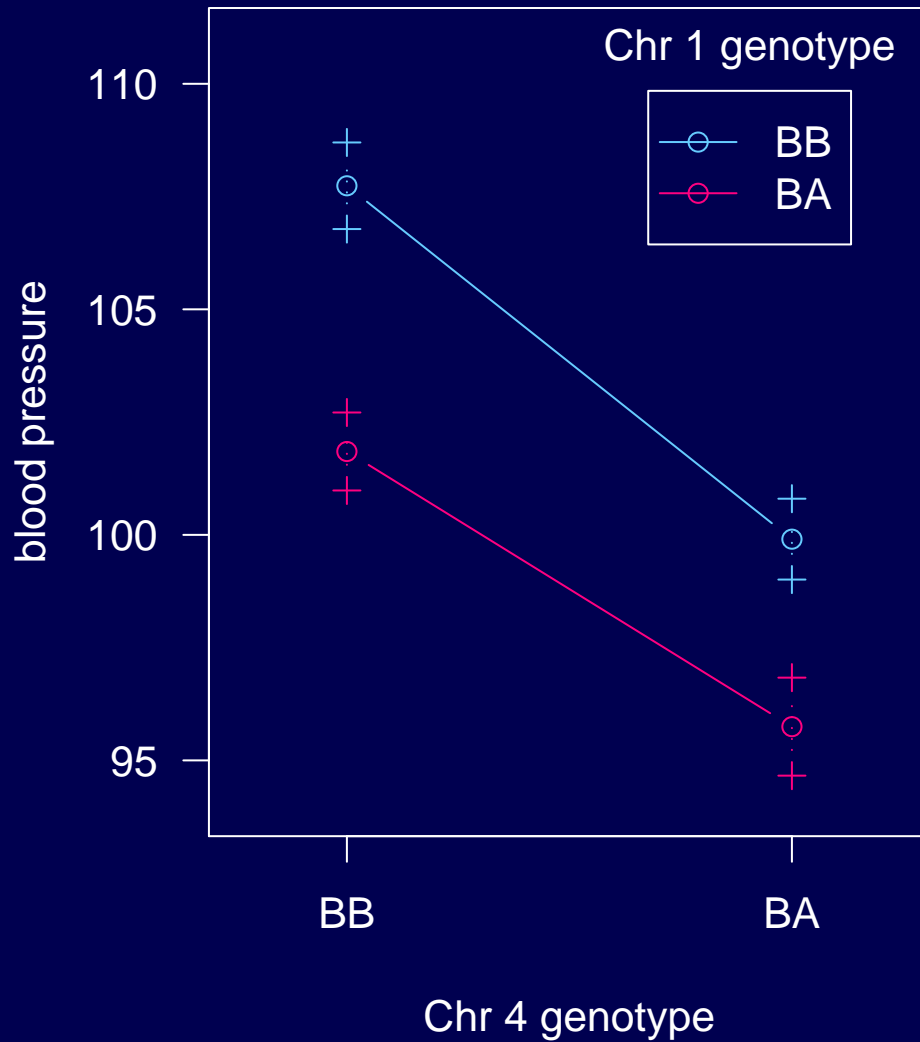


# Modeling multiple QTL

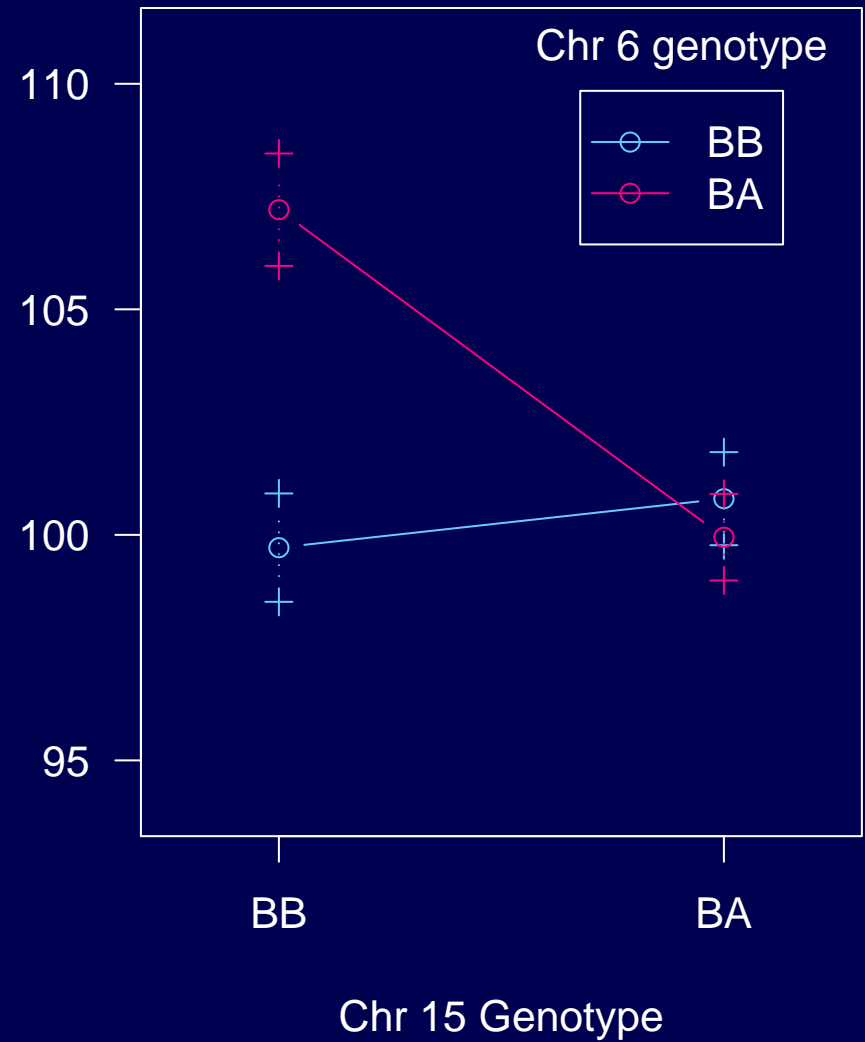
- Reduce residual variation  $\longrightarrow$  increased power
- Separate linked QTL
- Identify interactions among QTL (epistasis)

# 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

# Automation

- Assistance to the masses
- Understanding performance
- Many phenotypes

# Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

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

$$\mathbf{BIC}(\gamma) = \log \mathbf{RSS}(\gamma) + \left( \frac{\log n}{n} \right) |\gamma|$$

# Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

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

$$\mathbf{BIC}_\delta(\gamma) = \log \mathbf{RSS}(\gamma) + \left( \delta \cdot \frac{\log n}{n} \right) |\gamma|$$

# Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

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

$$\text{LOD}_\delta(\gamma) = \text{LOD}(\gamma) - (2\delta \log n) |\gamma|$$

# Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

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

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

# Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

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

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

$$0 \text{ vs } 1 \text{ QTL: } \text{LOD}_\delta(\emptyset) = 0$$

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

# Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

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

$$\text{LOD}_\delta(\gamma) = \text{LOD}(\gamma) - T |\gamma|$$

For the mouse genome:

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

# 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

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

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

$T_m$  = 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.

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

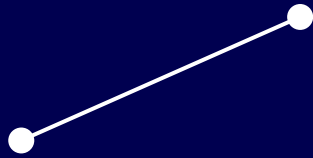
For the mouse genome:

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

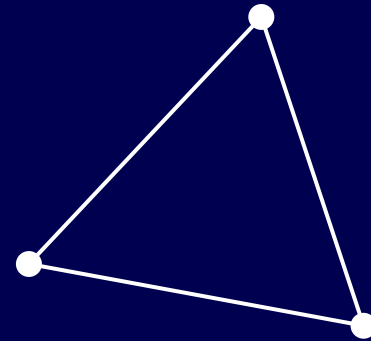
$$T_i^H = 2.62 \text{ (BC) or } 4.28 \text{ (F}_2\text{)}$$

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

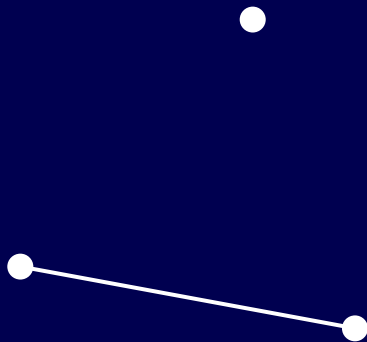
# Models as graphs



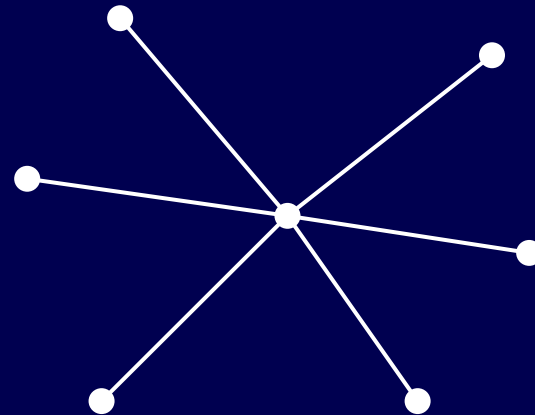
A



C

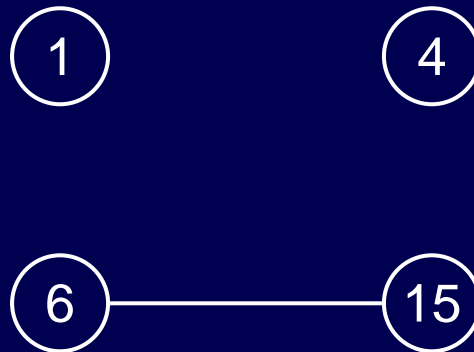


B



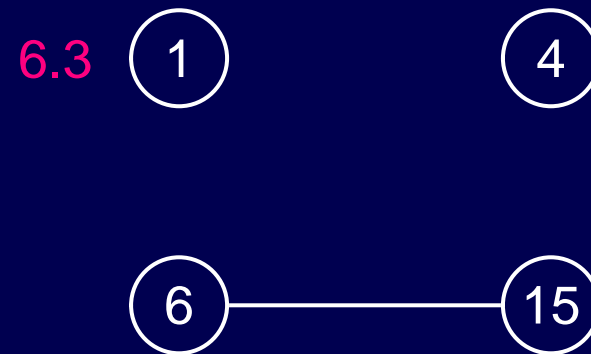
D

# Results



LOD = 23.1

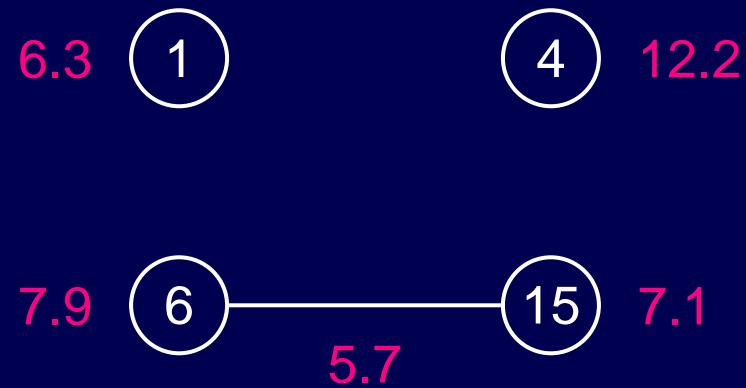
# Results



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



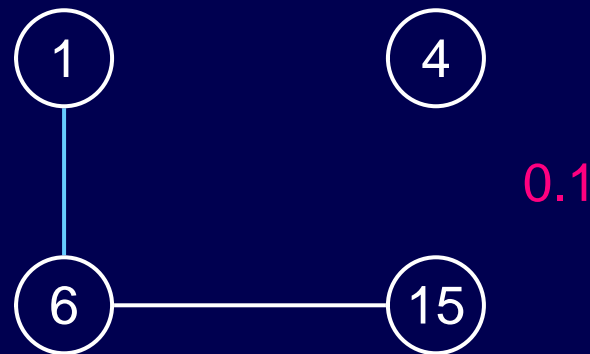
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

# Add an interaction?



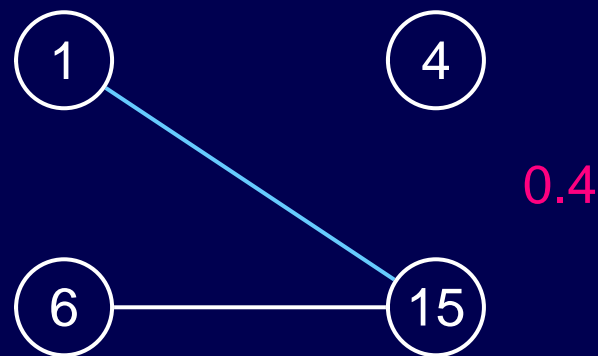
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

# Add an interaction?



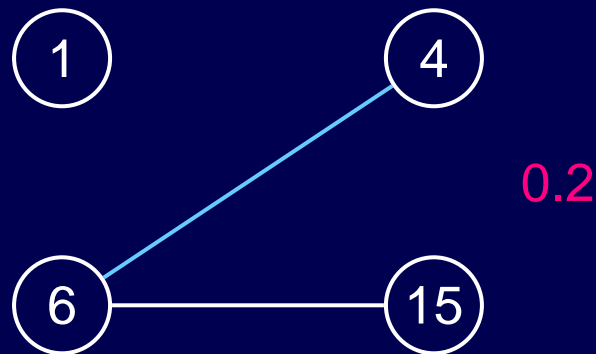
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

# Add an interaction?



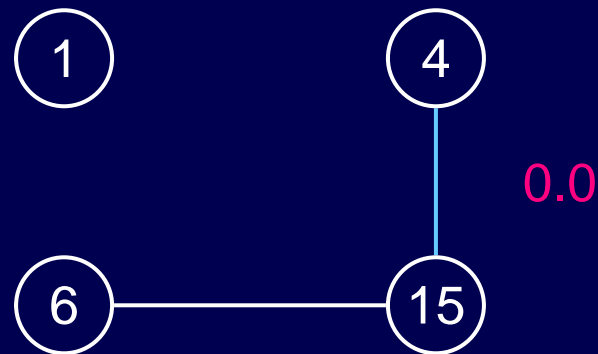
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

# Add an interaction?



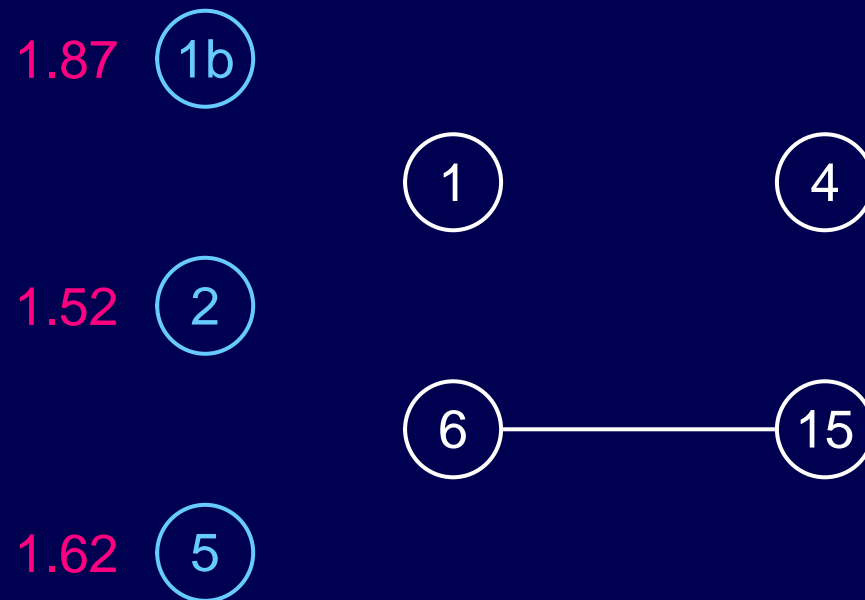
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

# Add an interaction?



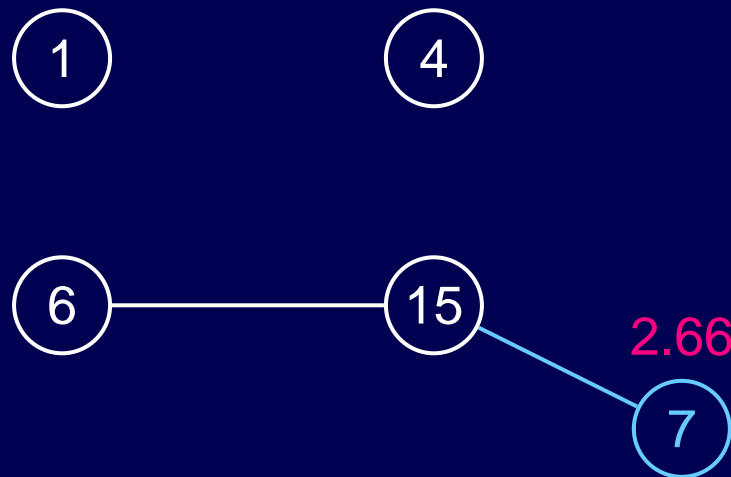
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

# Add another QTL?



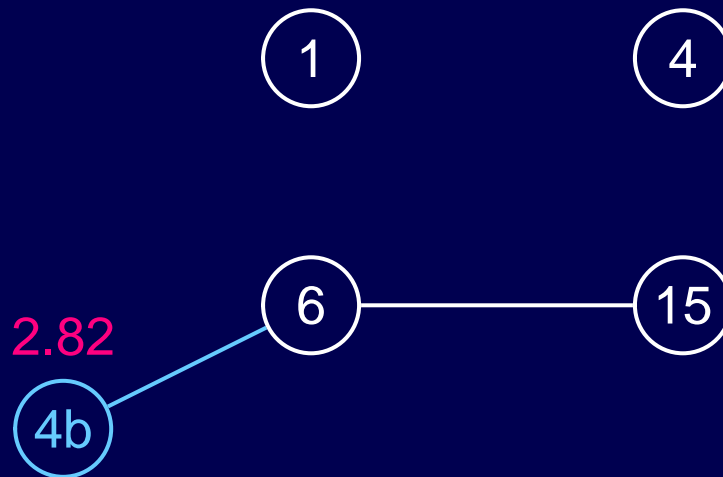
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

# Add another QTL?



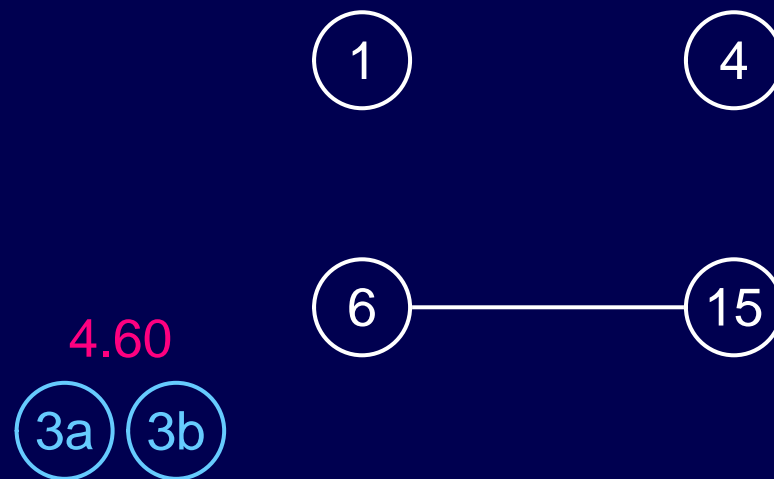
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

# 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?



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

# To do

- Improve search procedures
- X chromosome
- QTL  $\times$  covariate interactions
- Measuring model uncertainty
- Measuring uncertainty in QTL location

# Summary

- QTL mapping is a model selection problem
- The criterion for comparing models is most important
- We're focusing on a penalized likelihood method and are close to a practiceable solution

Manichaikul A, Moon JY, Sen Ś, Yandell BS, Broman KW (2009) A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics*, to appear

# Acknowledgments

Ani Manichaikul	University of Virginia
Gary Churchill	Jackson Laboratory
Śaunak Sen	University of California, San Francisco
Terry Speed	University of California, Berkeley
Brian Yandell	University of Wisconsin – Madison
Jee Young Moon	University of Wisconsin – Madison
Fumihiko Sugiyama	now at University of Tsukuba, Japan
Bev Paigen	Jackson Laboratory