

Mapping multiple QTL in experimental crosses

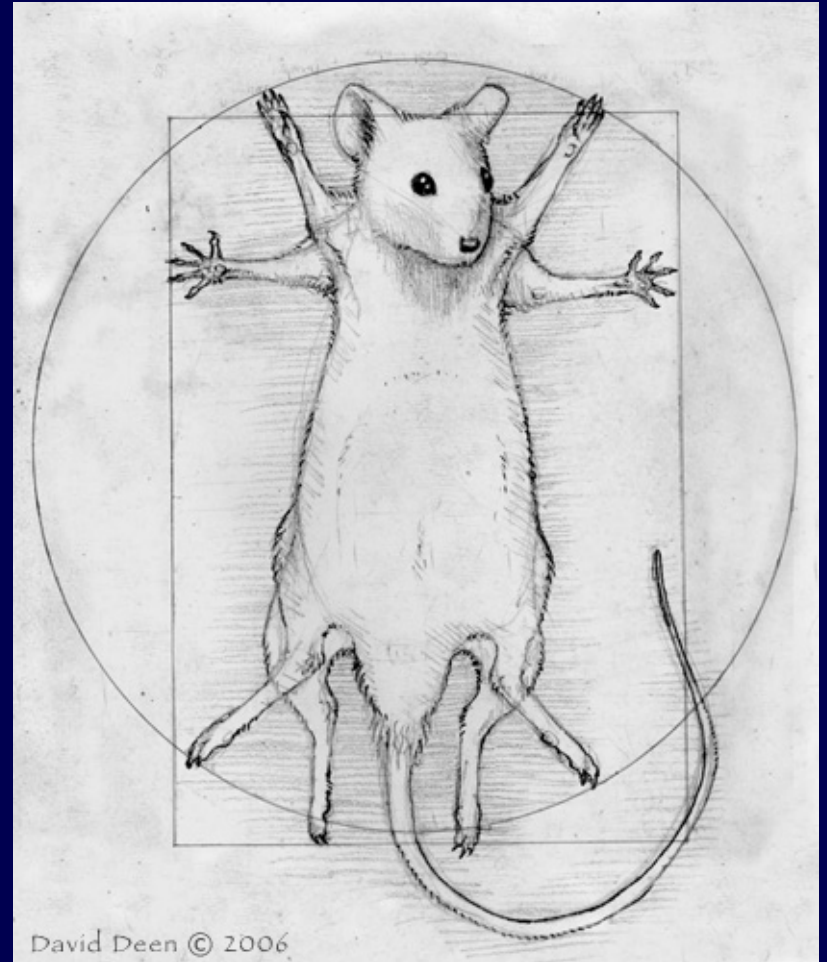
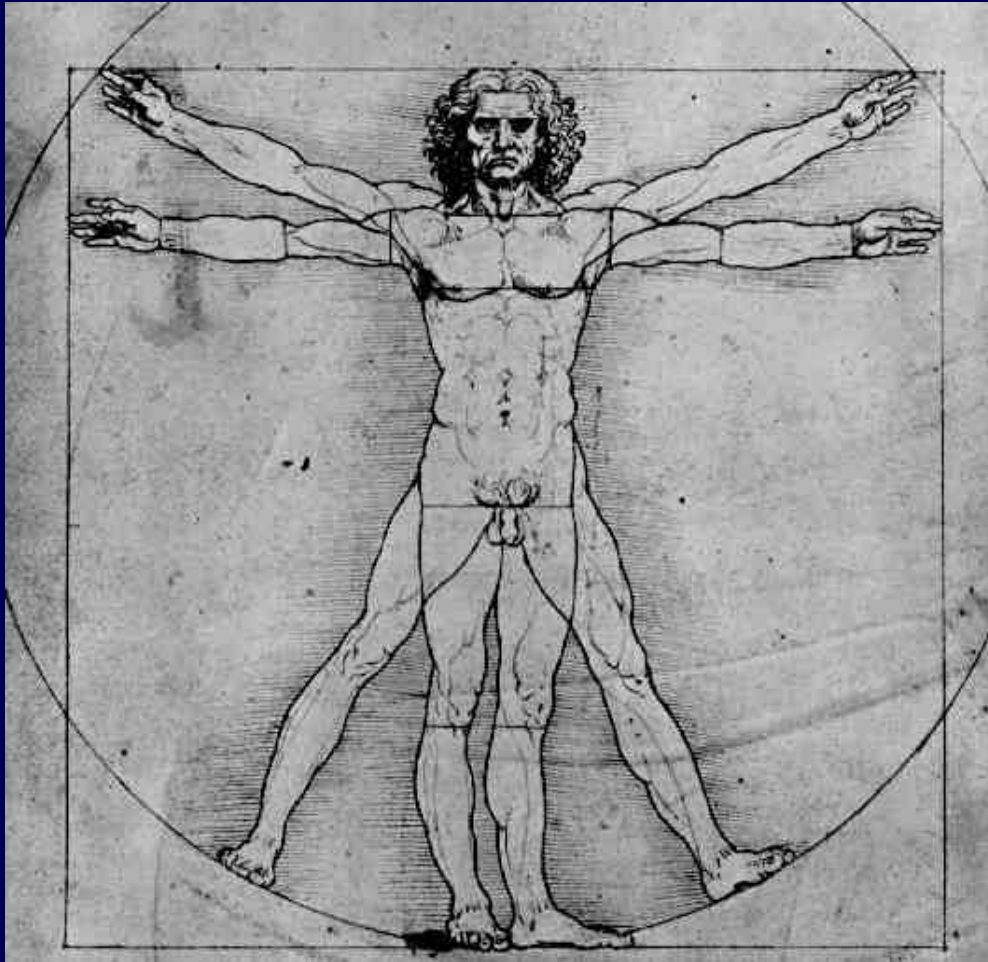
Karl W Broman

Department of Biostatistics
Johns Hopkins University

www.biostat.jhsph.edu/~kbroman

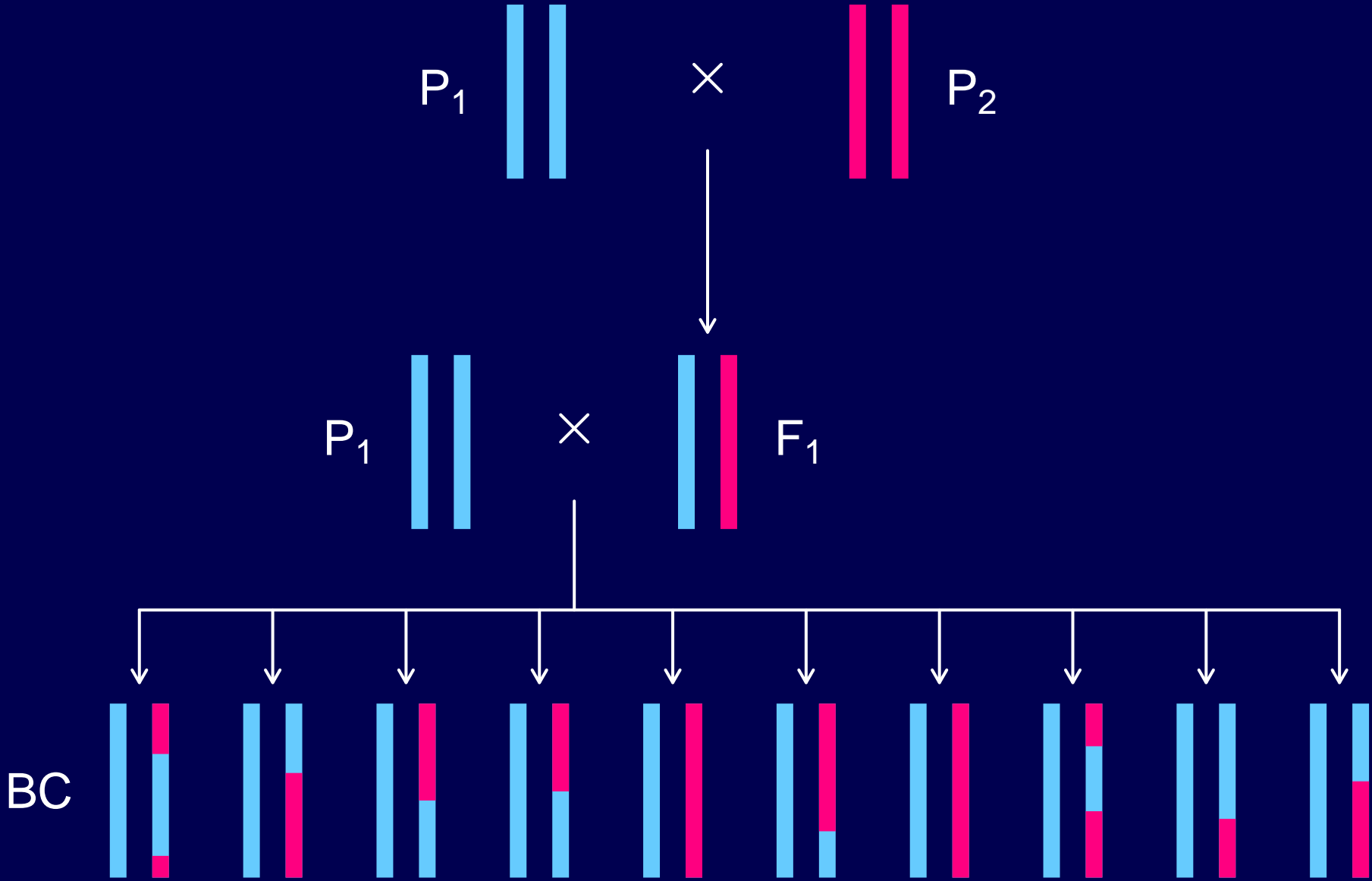


Human vs mouse

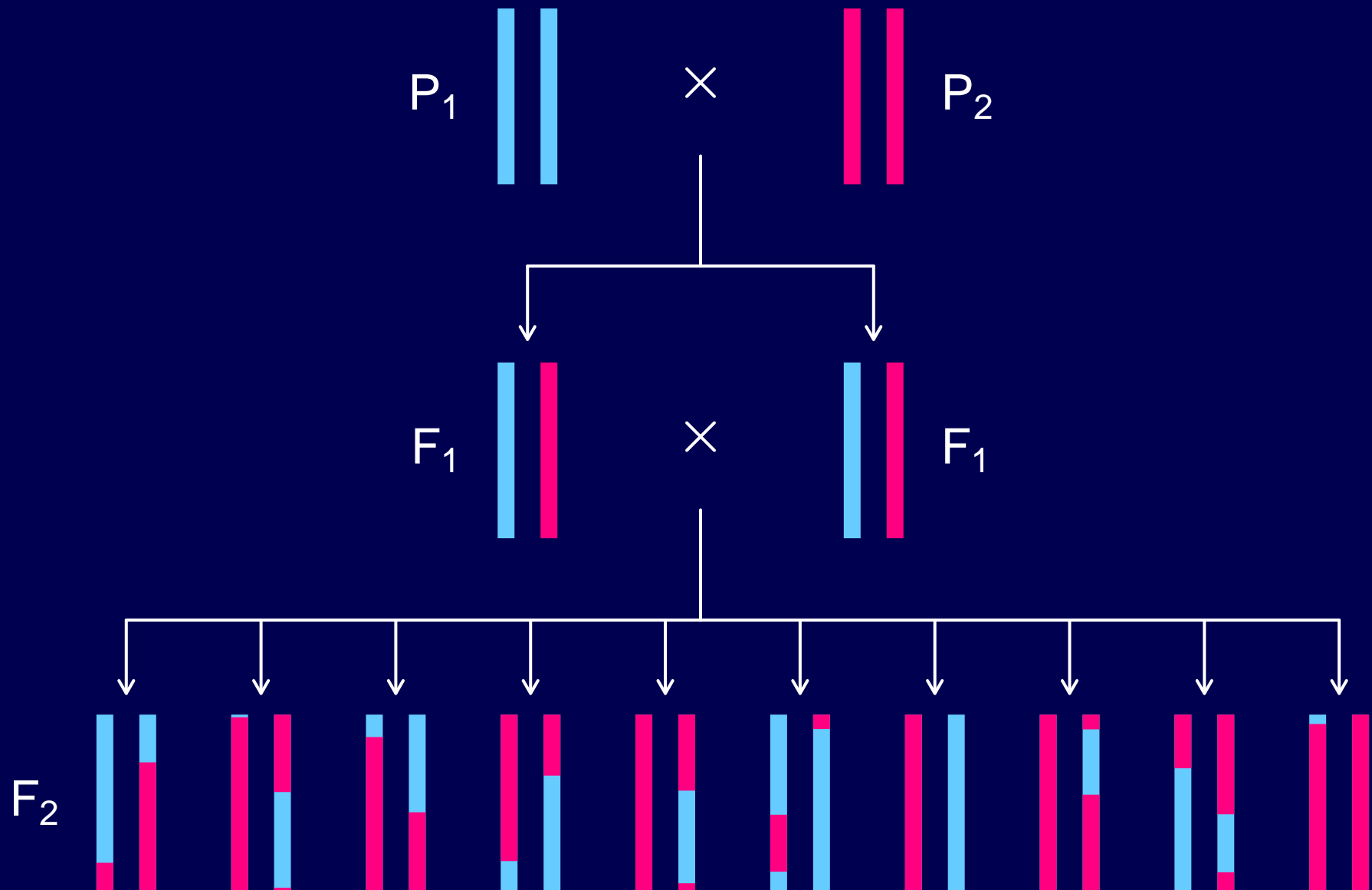


www.daviddeen.com

Backcross



Intercross

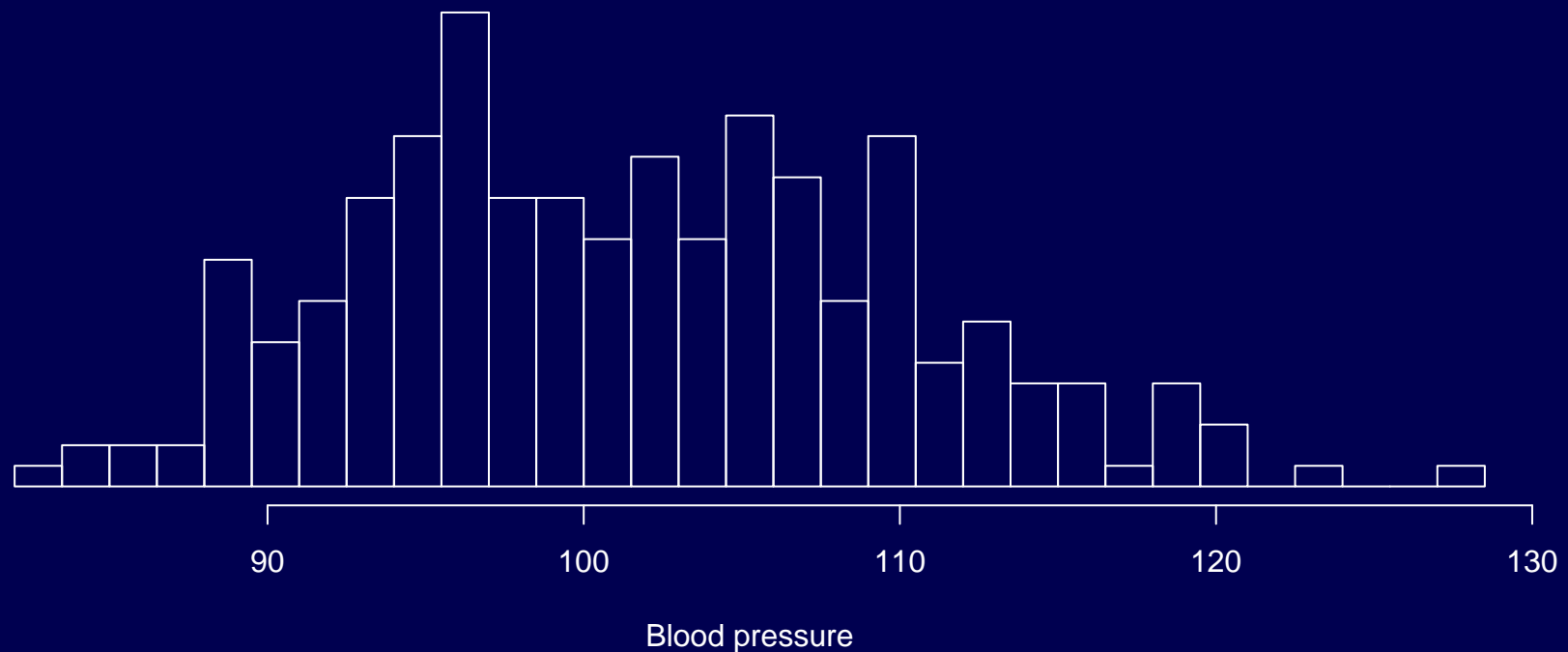


Phenotype data

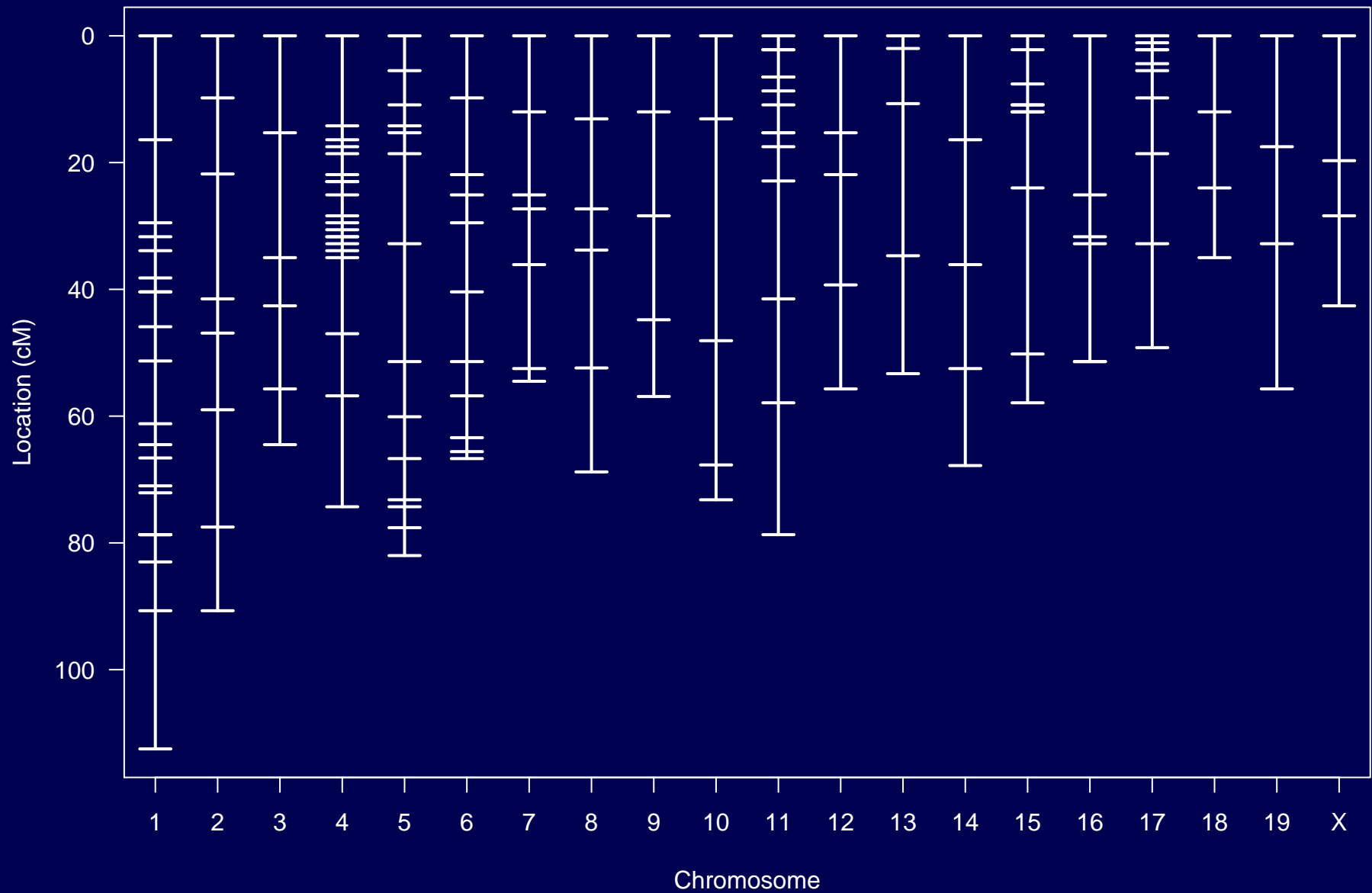
Sugiyama et al. Genomics 71:70-77, 2001

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

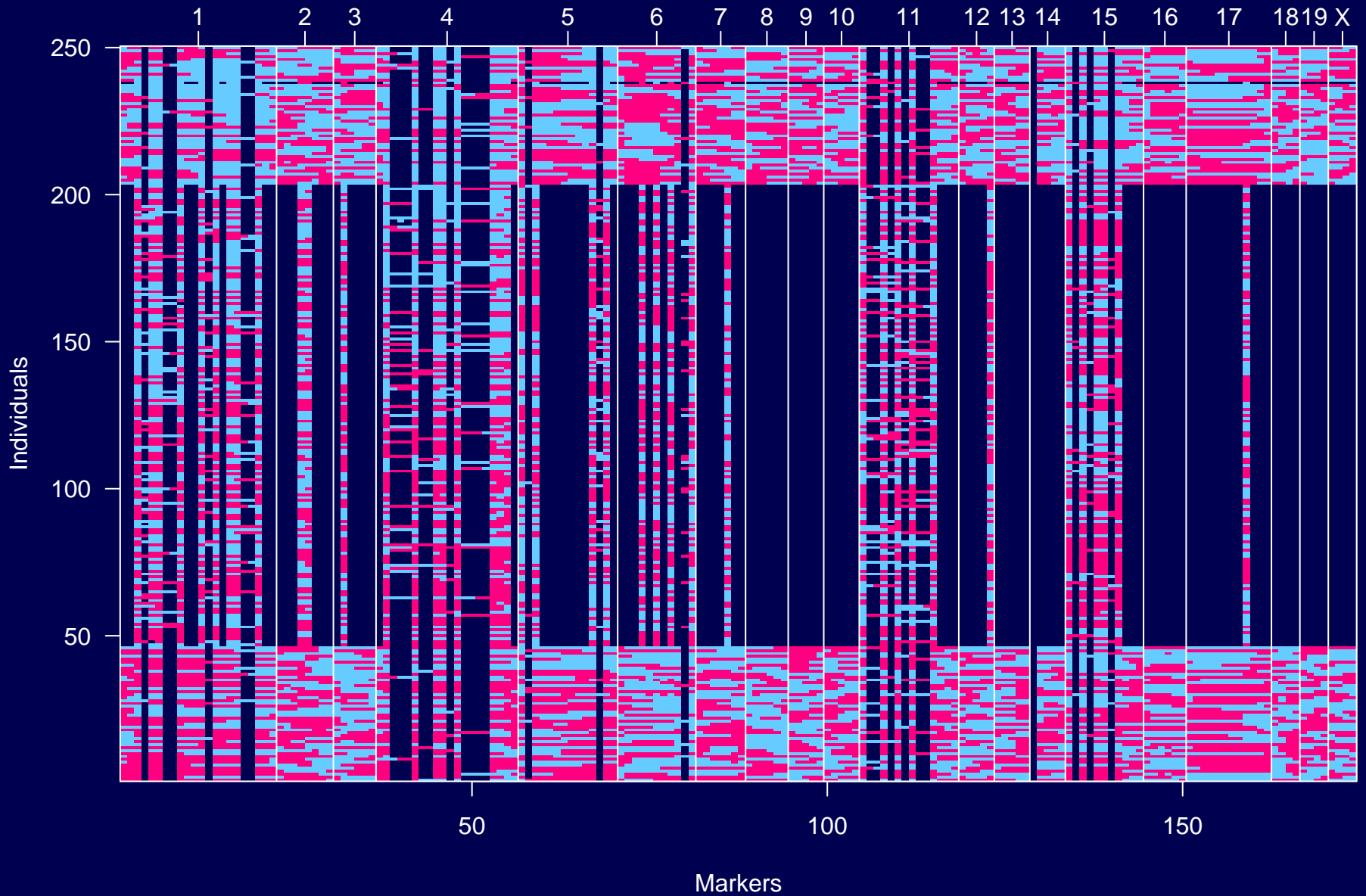
Blood pressure after two weeks drinking water with 1% NaCl



Genetic map



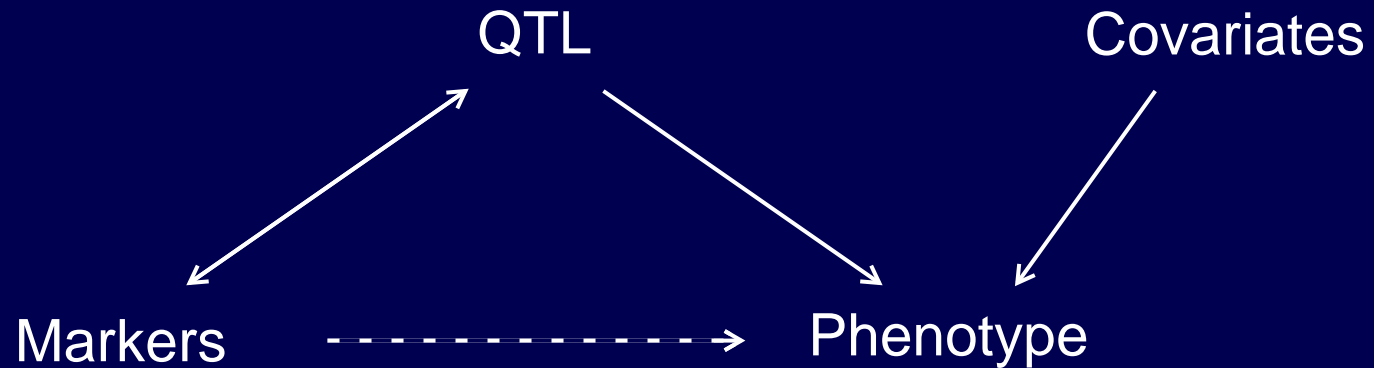
Genotype data



Goals

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

Statistical structure



The missing data problem:

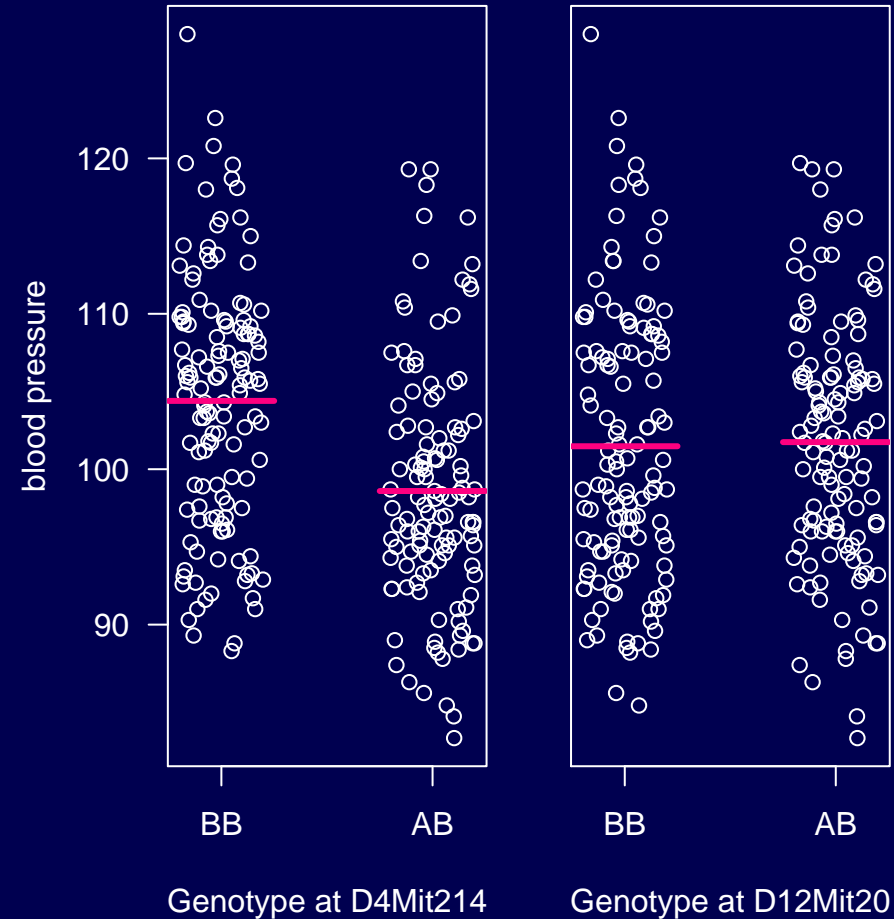
Markers \longleftrightarrow QTL

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

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

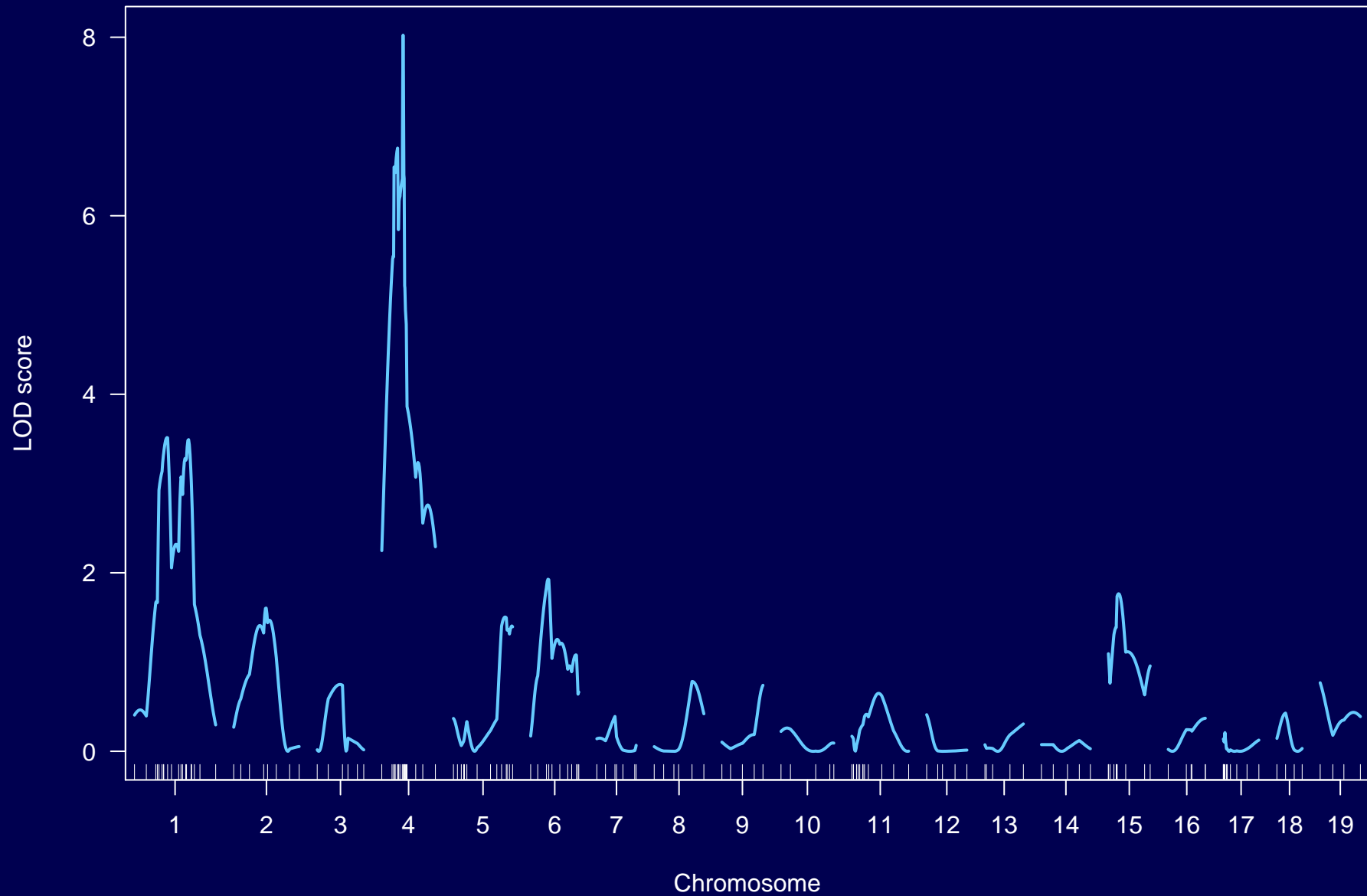
LOD scores

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

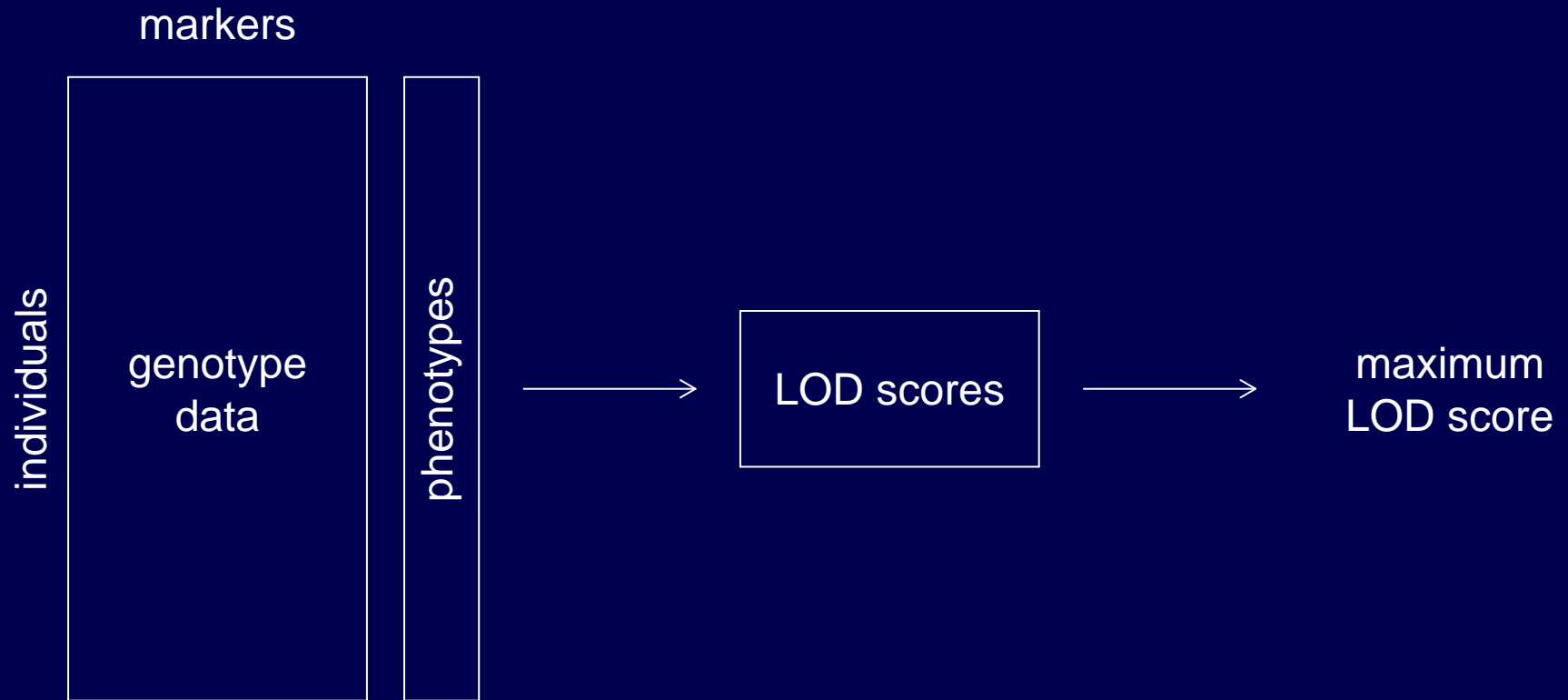
$\hat{\mu}_{q\lambda}, \hat{\sigma}_\lambda$ are the MLEs, assuming a single QTL at position λ .

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

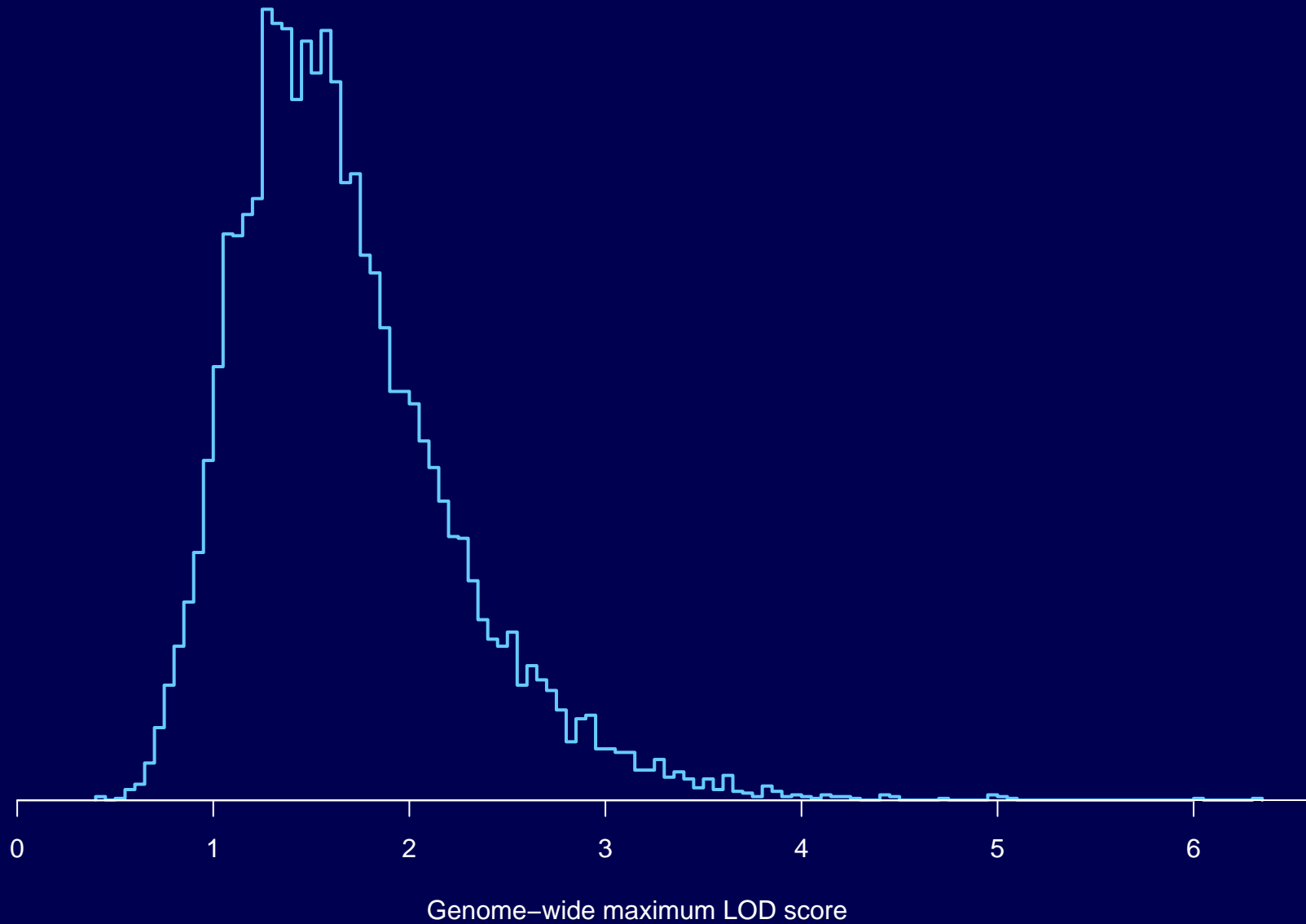
LOD curves



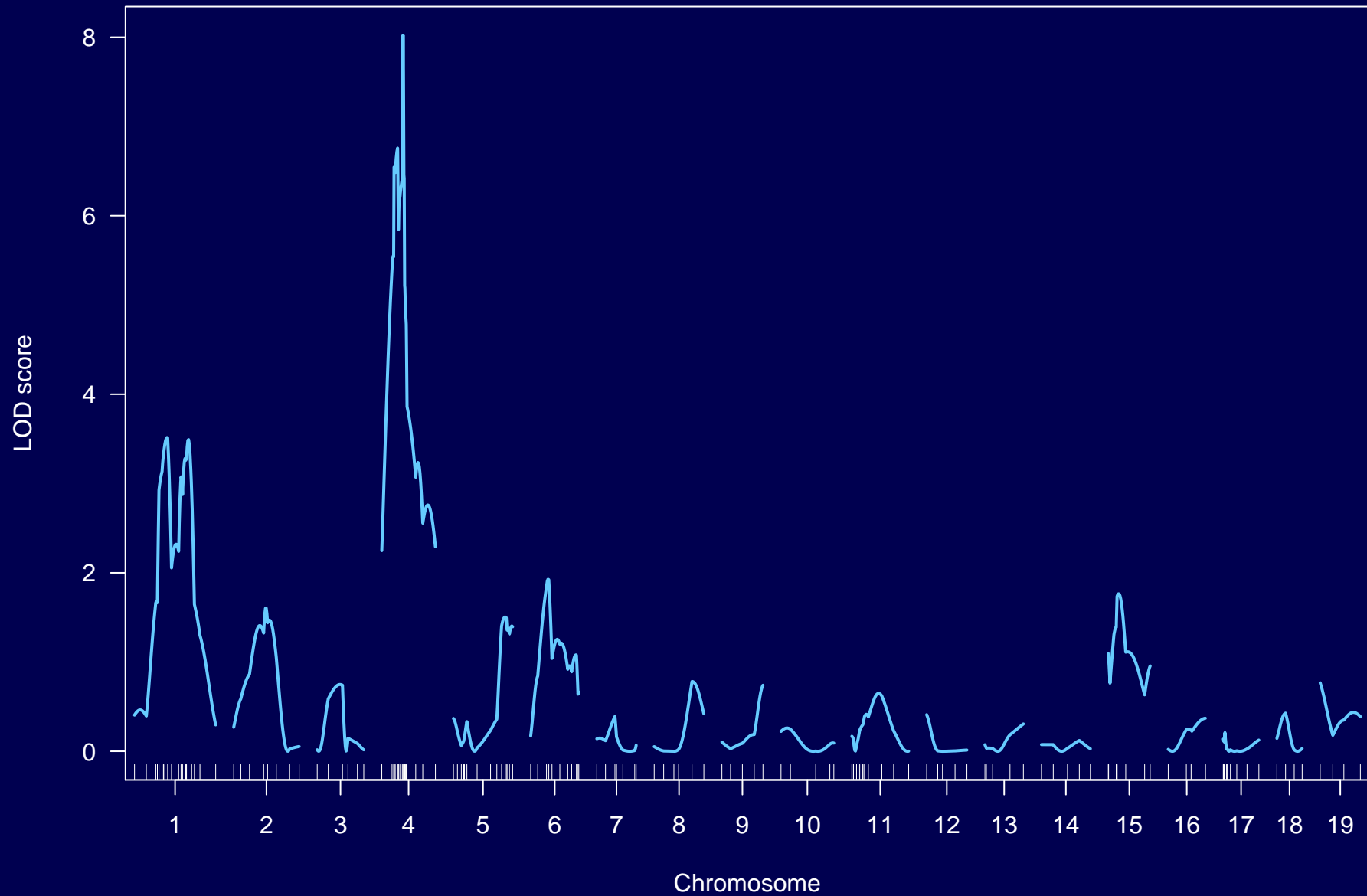
Permutation test



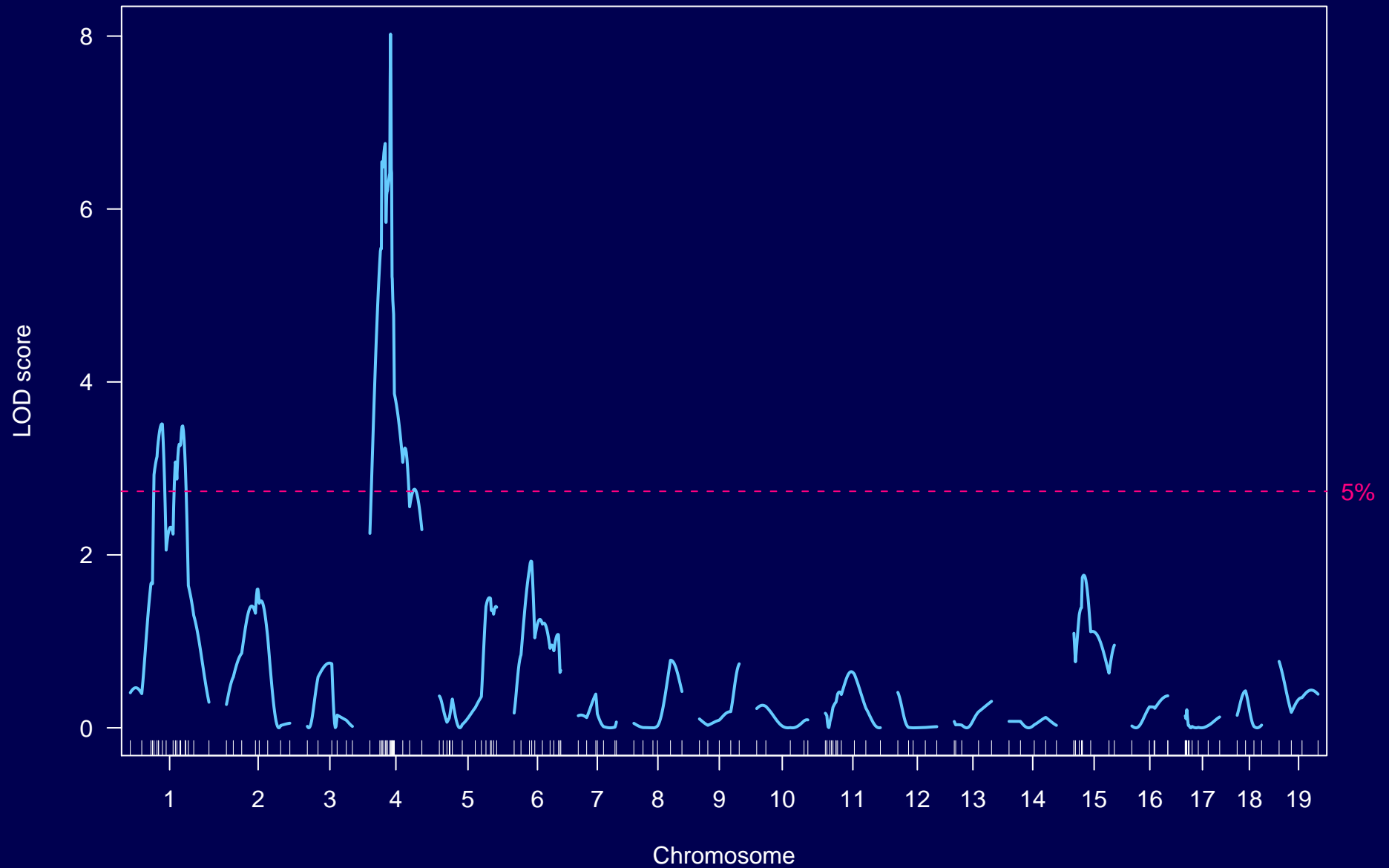
Permutation results



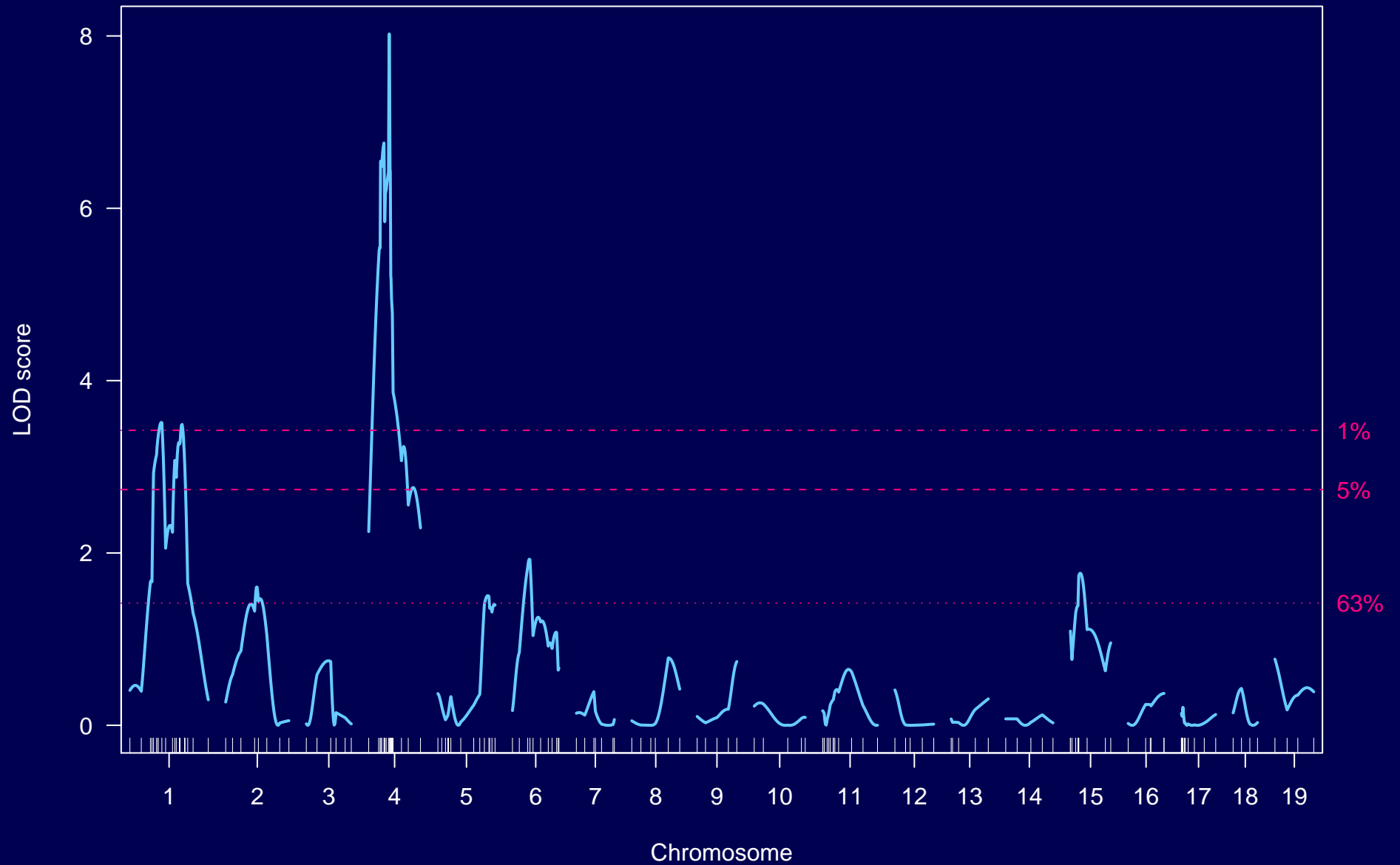
LOD curves



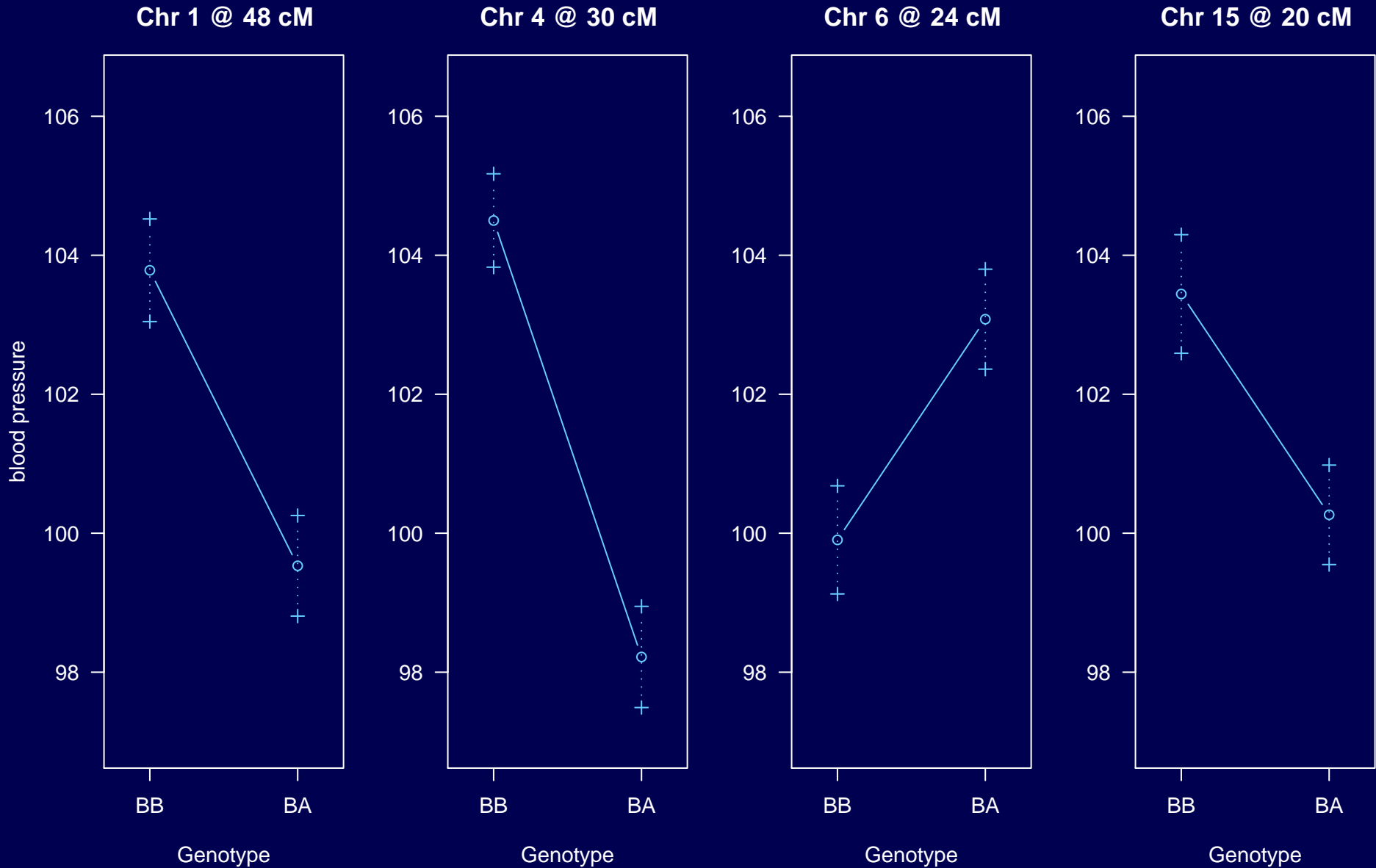
LOD curves



LOD curves



Estimated effects



Modeling multiple QTL

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

2-dim, 2-QTL scan

For each pair of positions, fit the following models:

$$H_f : y = \mu + \beta_1 q_1 + \beta_2 q_2 + \gamma q_1 q_2 + \epsilon$$

$$H_a : y = \mu + \beta_1 q_1 + \beta_2 q_2 + \epsilon$$

$$H_1 : y = \mu + \beta_1 q_1 + \epsilon$$

$$H_0 : y = \mu + \epsilon$$

2-dim, 2-QTL scan

For each pair of positions, fit the following models:

$$H_f : y = \mu + \beta_1 q_1 + \beta_2 q_2 + \gamma q_1 q_2 + \epsilon$$

$$H_a : y = \mu + \beta_1 q_1 + \beta_2 q_2 + \epsilon$$

$$H_1 : y = \mu + \beta_1 q_1 + \epsilon$$

$$H_0 : y = \mu + \epsilon$$

$$\text{LOD}_f(s, t) = \log_{10} \{ L_f(s, t) / L_0 \}$$

2-dim, 2-QTL scan

For each pair of positions, fit the following models:

$$H_f : y = \mu + \beta_1 q_1 + \beta_2 q_2 + \gamma q_1 q_2 + \epsilon$$

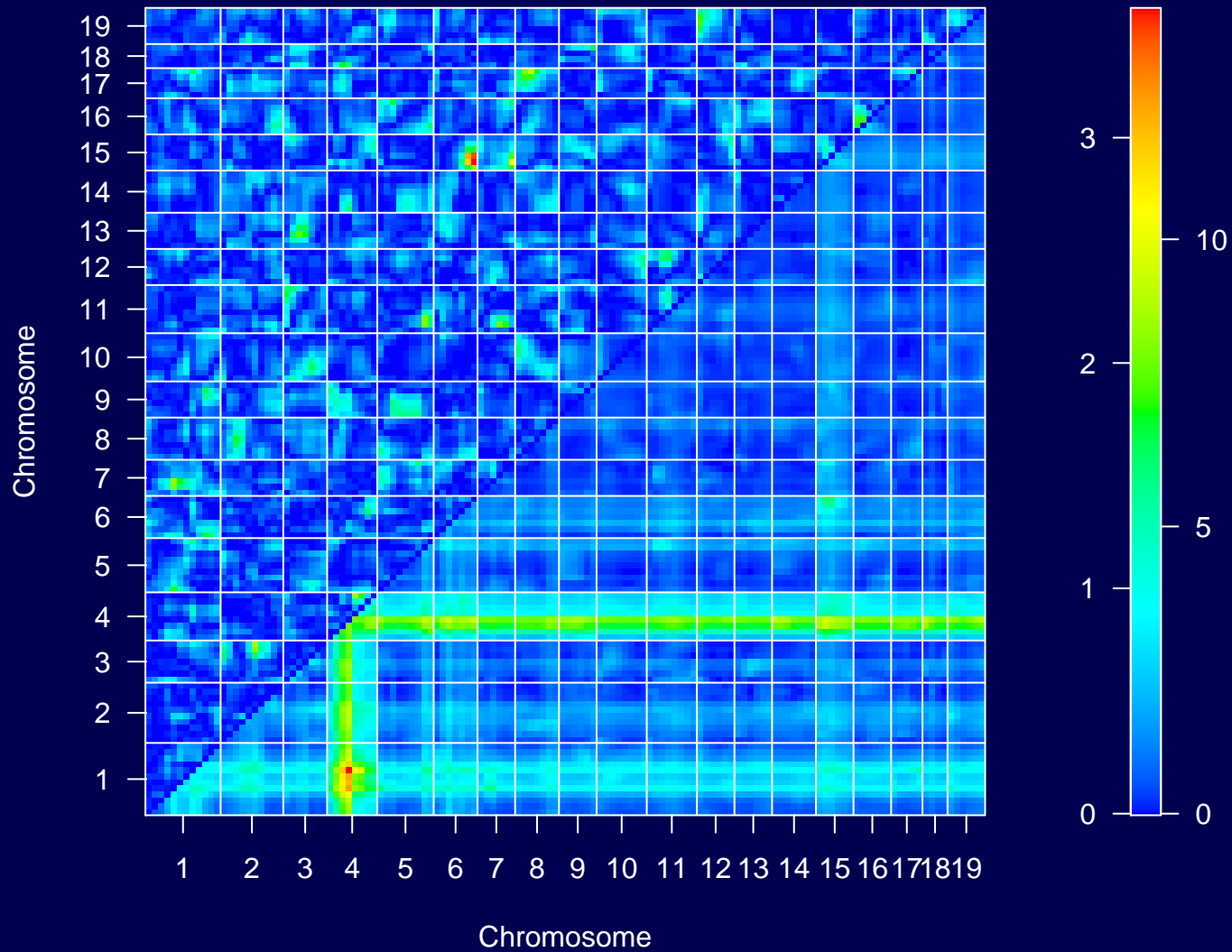
$$H_a : y = \mu + \beta_1 q_1 + \beta_2 q_2 + \epsilon$$

$$H_1 : y = \mu + \beta_1 q_1 + \epsilon$$

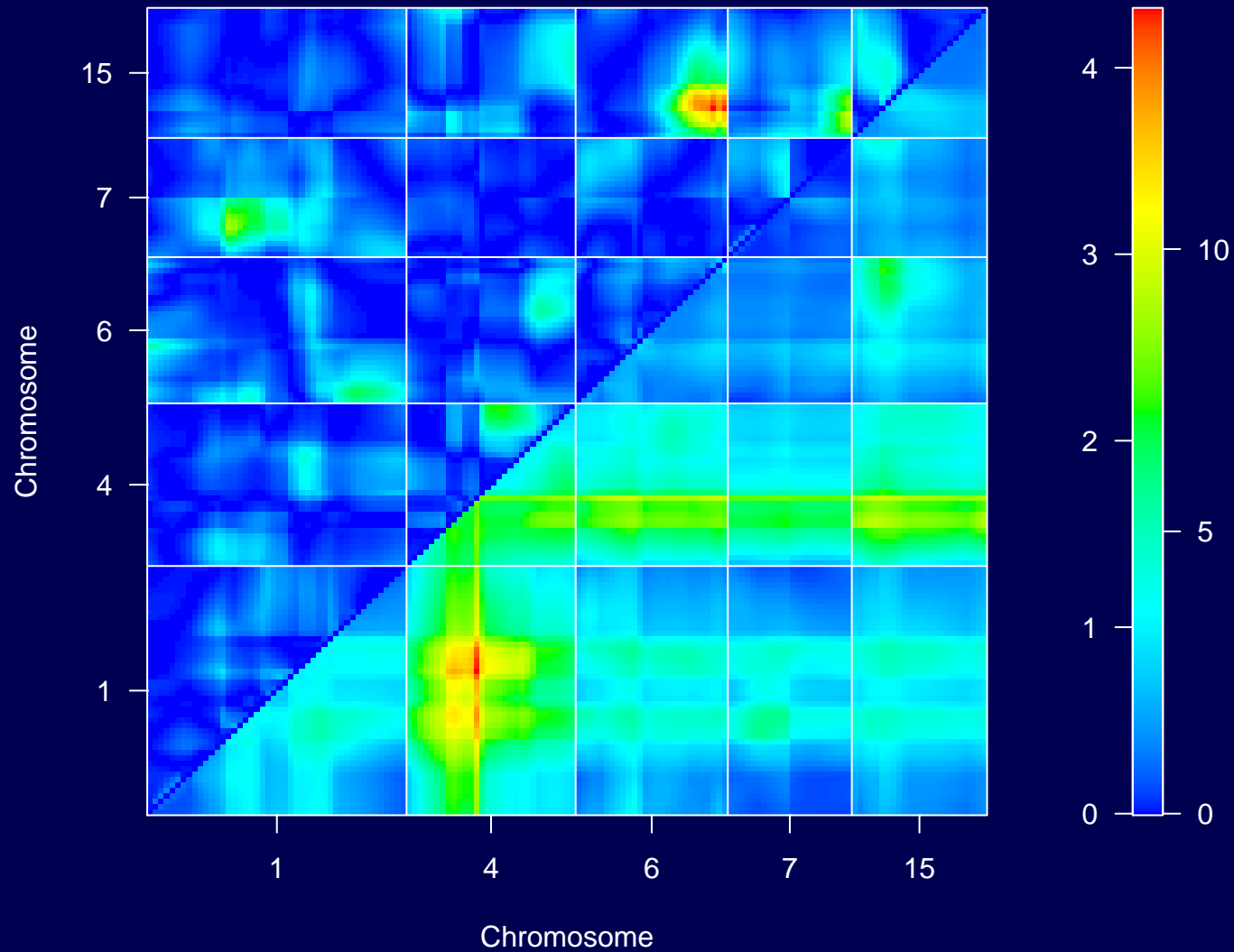
$$H_0 : y = \mu + \epsilon$$

$$\text{LOD}_i(s, t) = \log_{10} \{ L_f(s, t) / L_a(s, t) \}$$

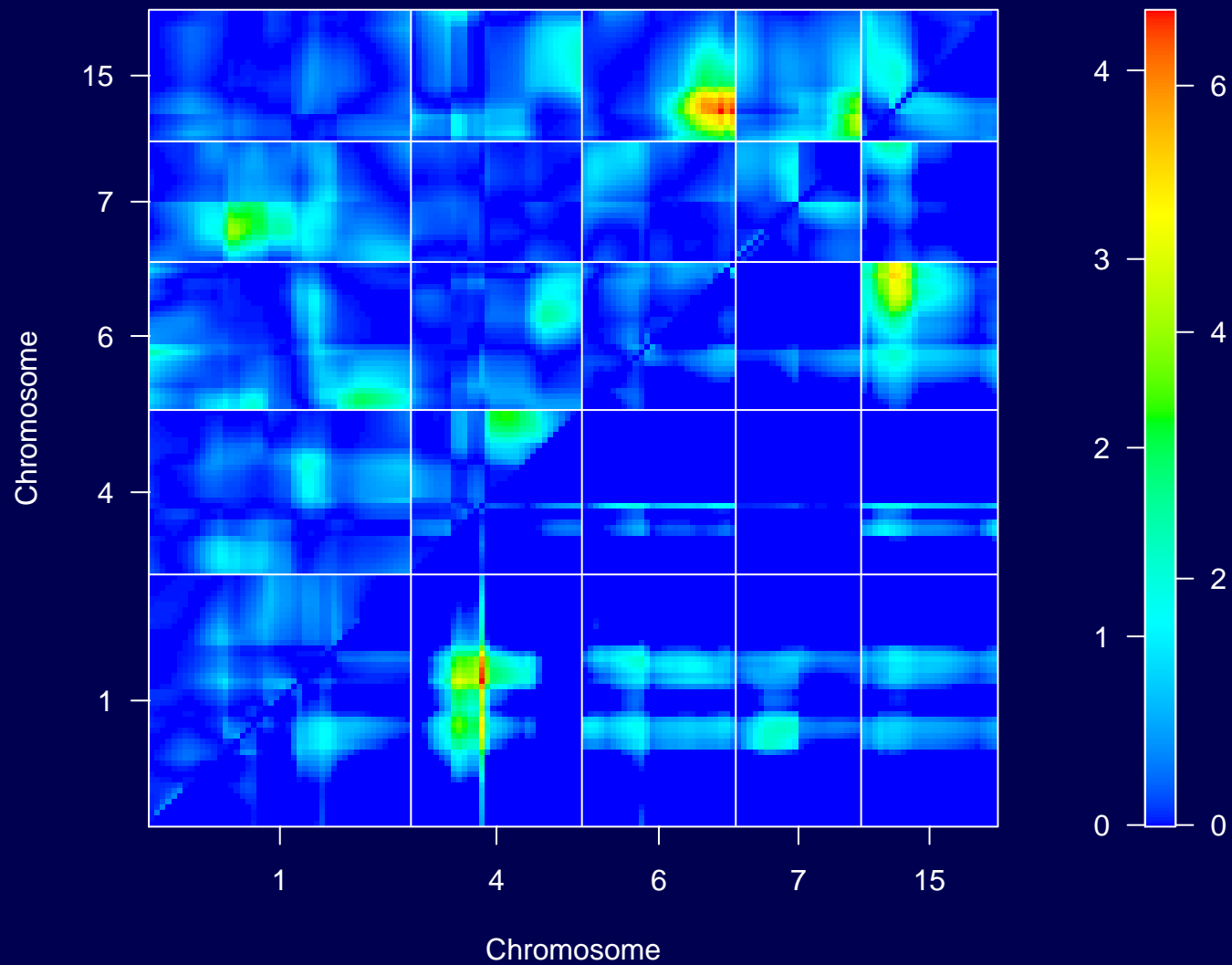
LOD_i and LOD_f



LOD_i and LOD_f

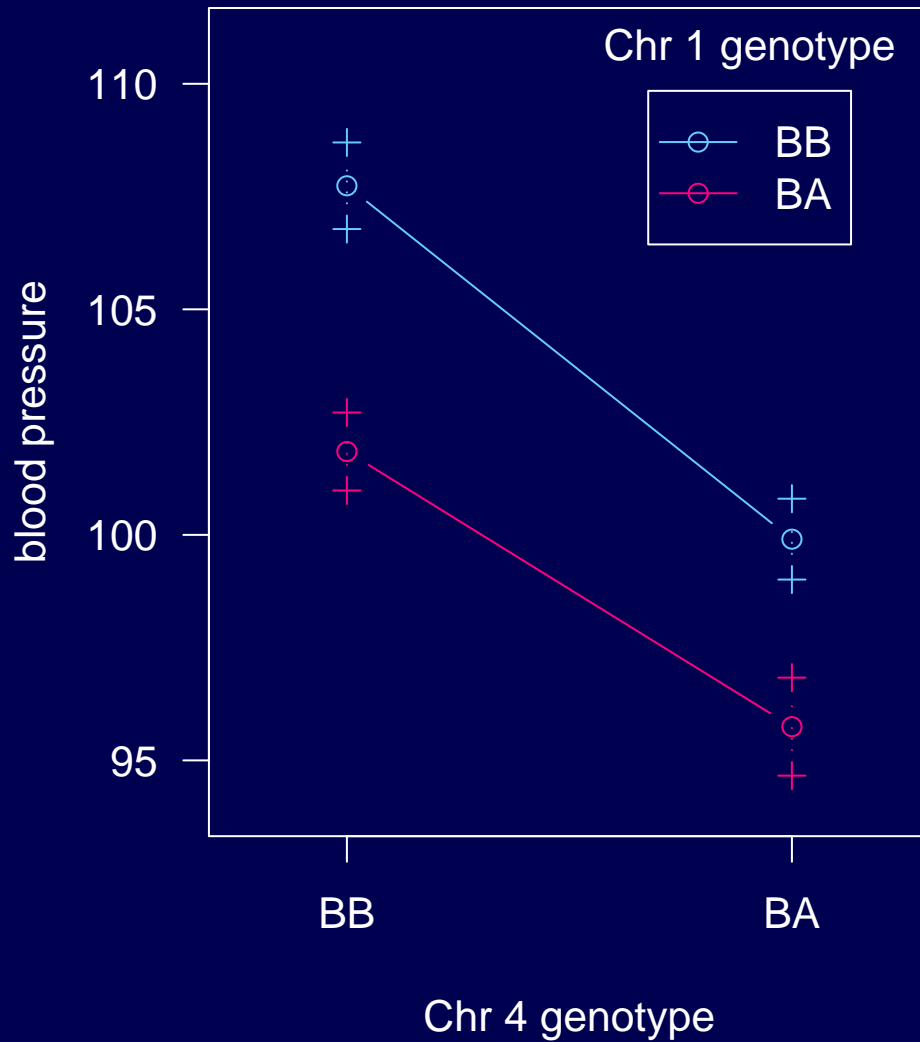


LOD_i and LOD_{fv1}

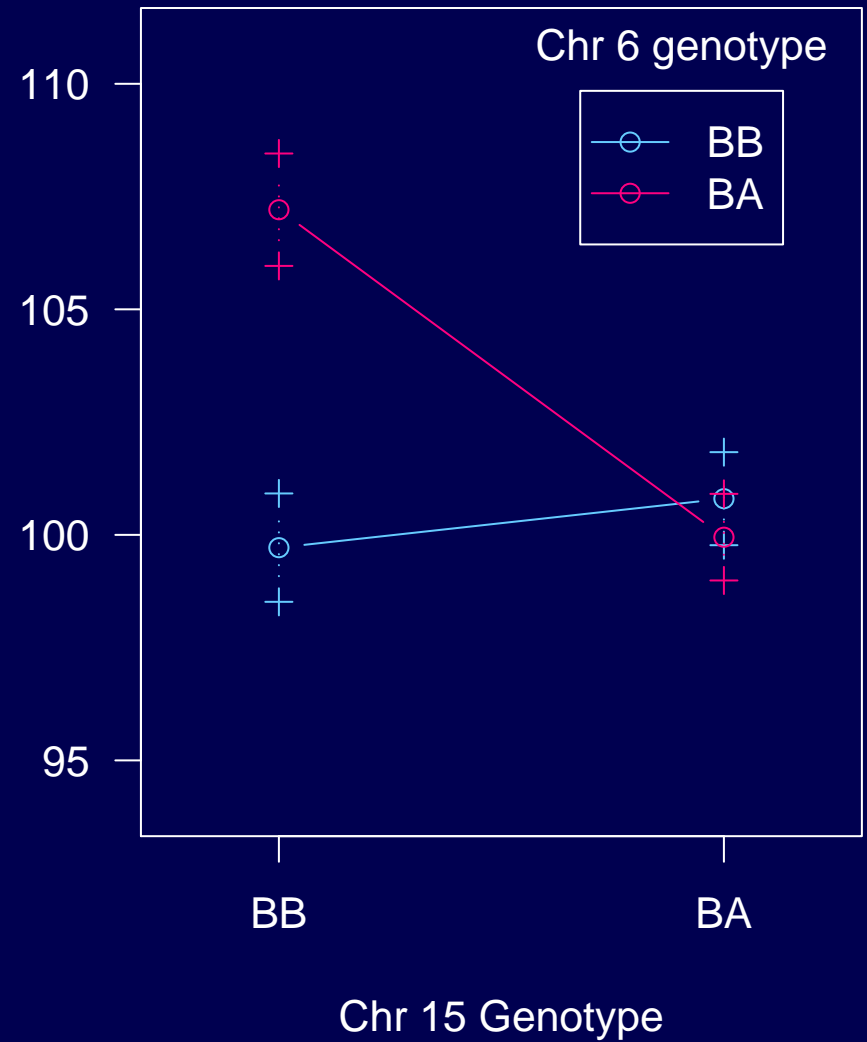


Estimated effects

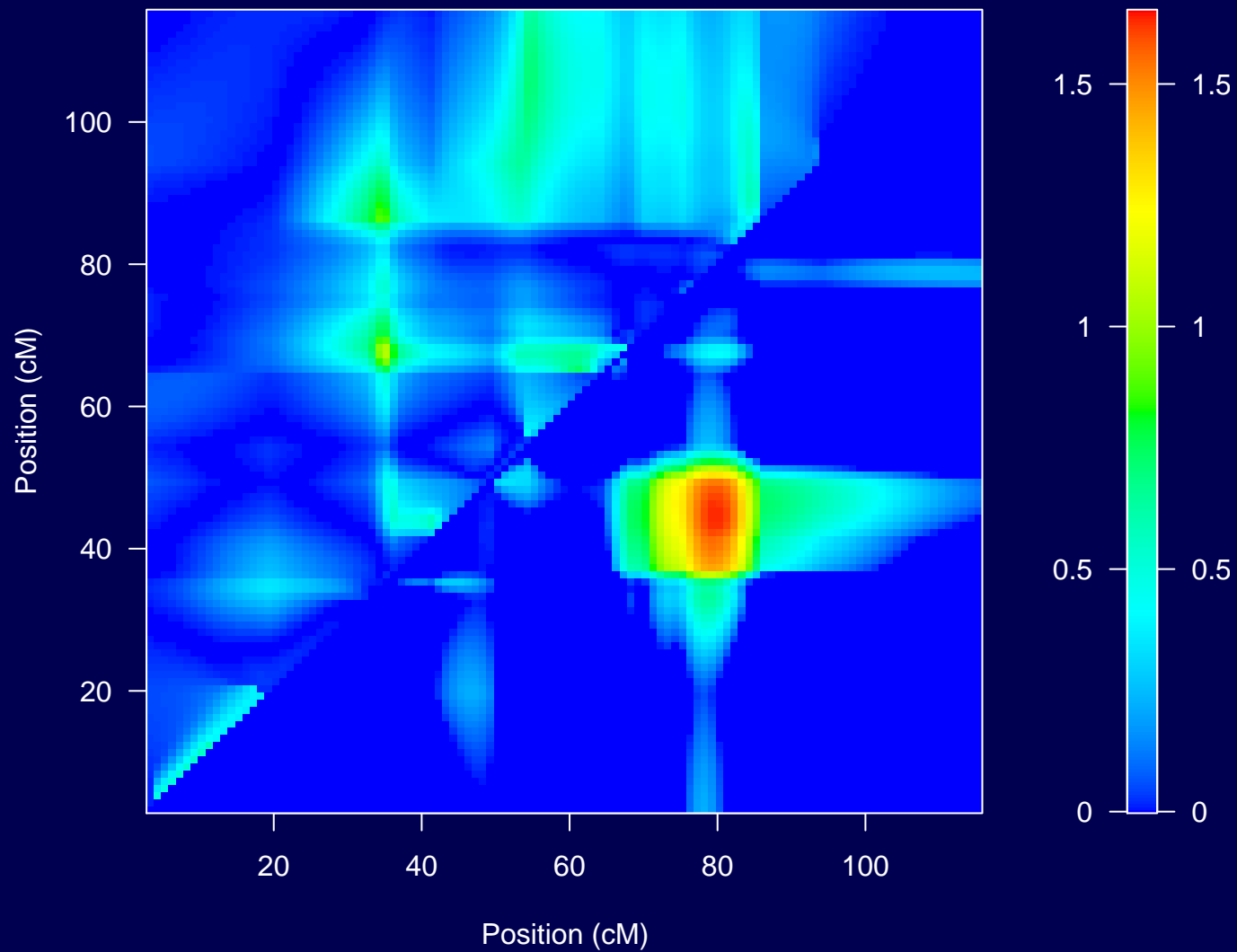
1 x 4



6 x 15



Chr 1: LOD_i and LOD_{av1}



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:

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

- 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

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.

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

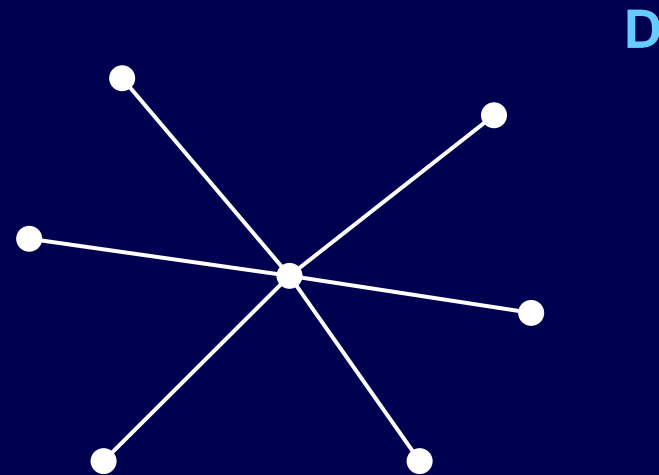
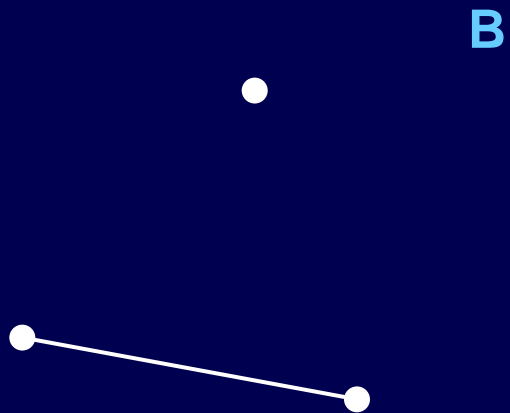
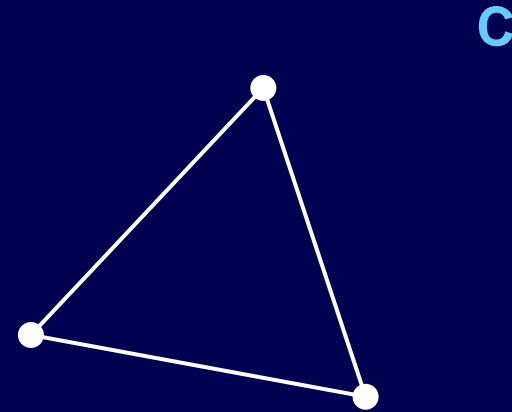
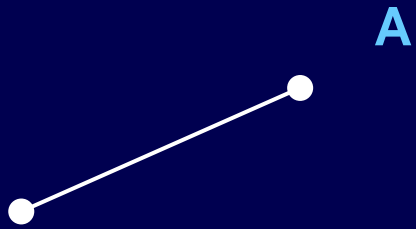
For the mouse genome:

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

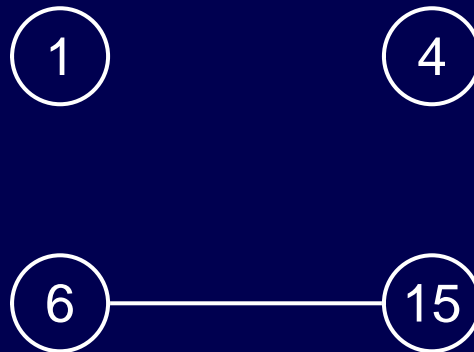
$$T_i^H = 2.62 \text{ (BC) or } 4.28 \text{ (F}_2\text{)}$$

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

Models as graphs

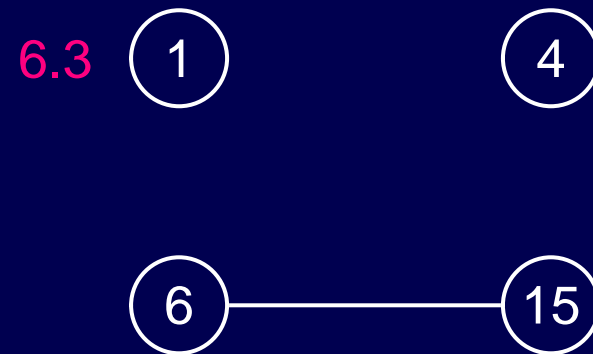


Results



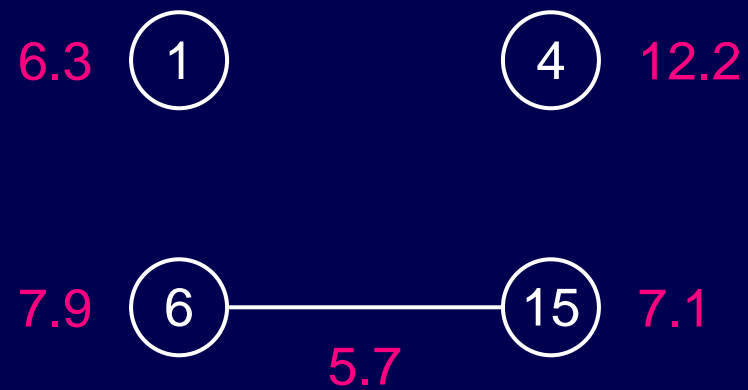
LOD = 23.1

Results



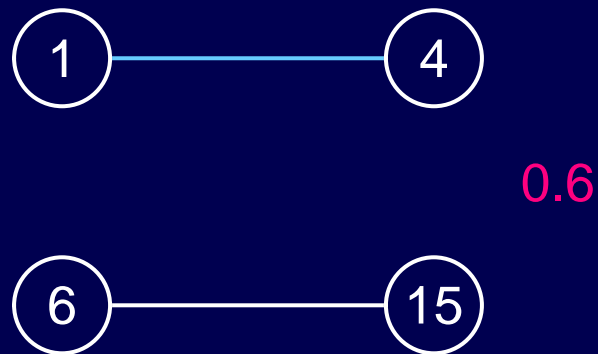
LOD = 23.1

Results



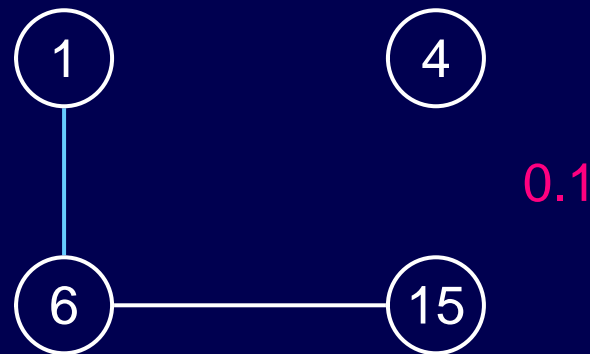
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88$$

Add an interaction?



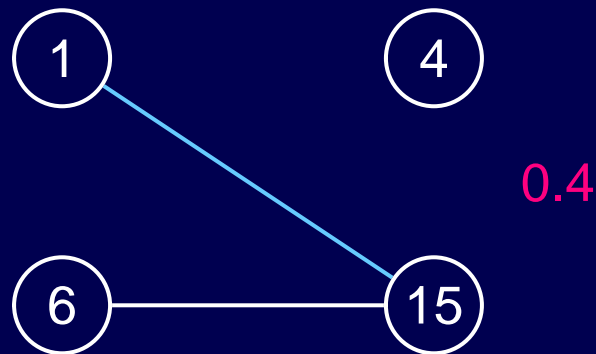
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88$$

Add an interaction?



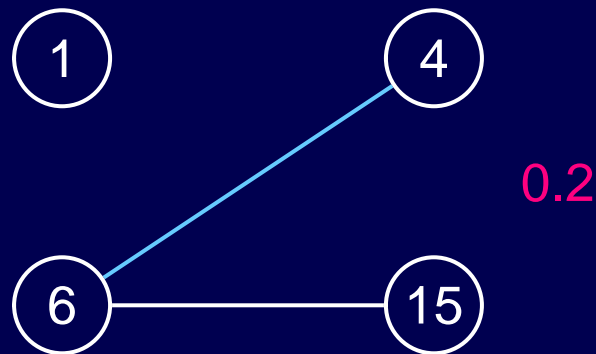
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88$$

Add an interaction?



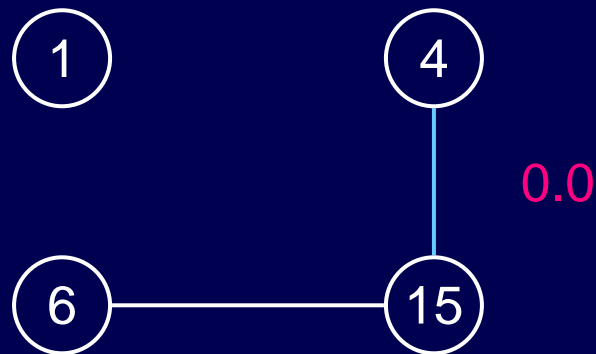
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88$$

Add an interaction?



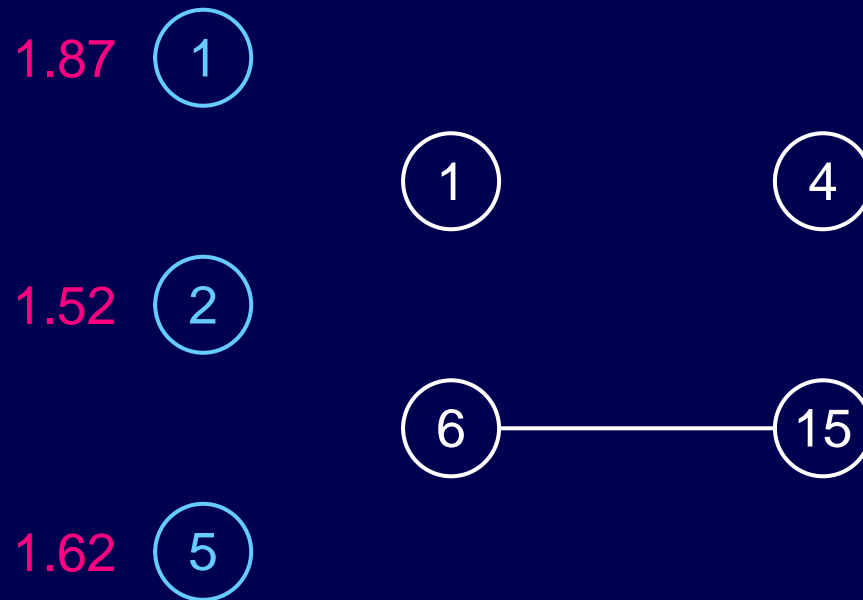
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88$$

Add an interaction?



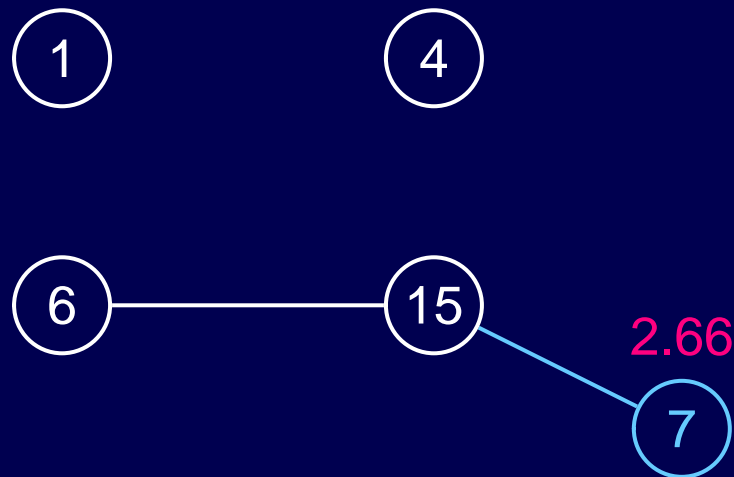
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88$$

Add another QTL?



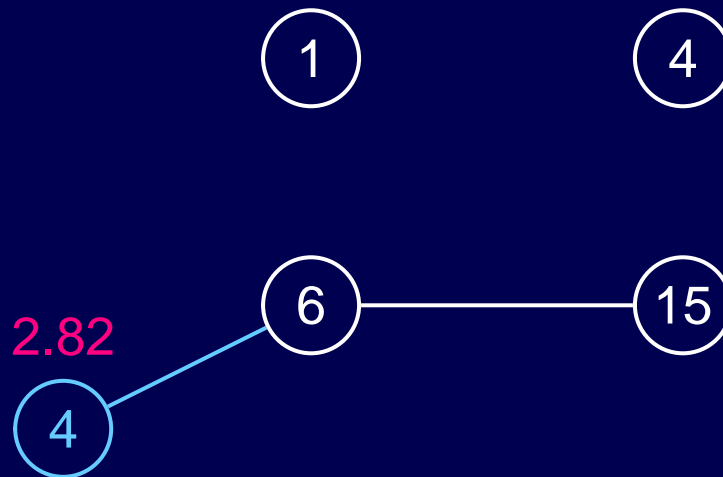
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88$$

Add another QTL?



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

Add another QTL?



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

To do

- Study performance
(especially relative to other approaches)
- Improve search procedures
- Measuring model uncertainty
- Measuring uncertainty in QTL location

Bayes/MCMC

Advantages

- All analysis aspects combined
- More fully captures uncertainty
- More clean expression of uncertainty in the inference

Disadvantages

- May require a specialist
- Prior specification is difficult
- Bayes factors can be difficult to interpret
- Can be difficult to assess performance

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

Acknowledgments

| | |
|-------------------|---|
| Ani Manichaikul | Johns Hopkins University |
| Gary Churchill | Jackson Laboratory |
| Śaunak Sen | University of California, San Francisco |
| Terry Speed | University of California, Berkeley |
| Brian Yandell | University of Wisconsin, Madison |
| Fumihiko Sugiyama | now at University of Tsukuba, Japan |
| Bev Paigen | Jackson Laboratory |