

R/qtl workshop

(part 2)

Karl Broman

Biostatistics and Medical Informatics
University of Wisconsin – Madison

kbroman.org

github.com/kbroman

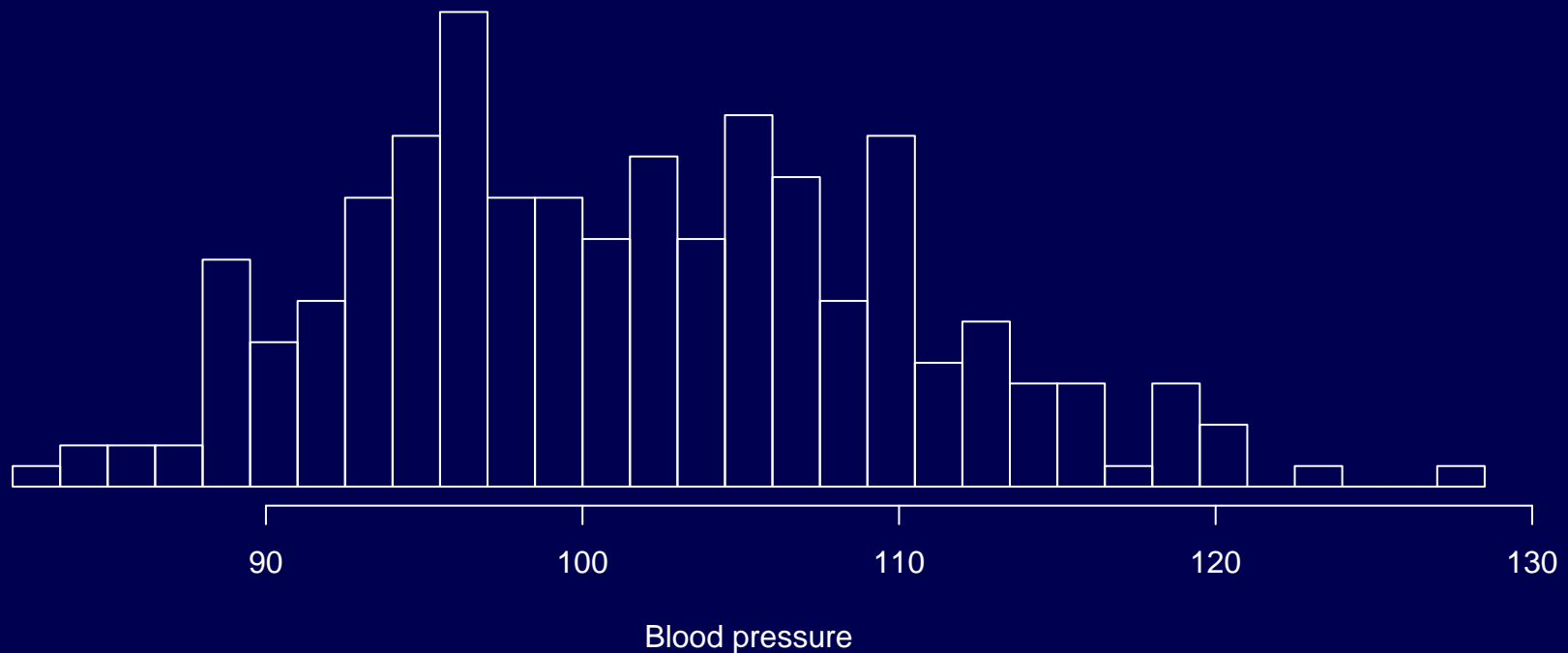
@kwbroman

Example

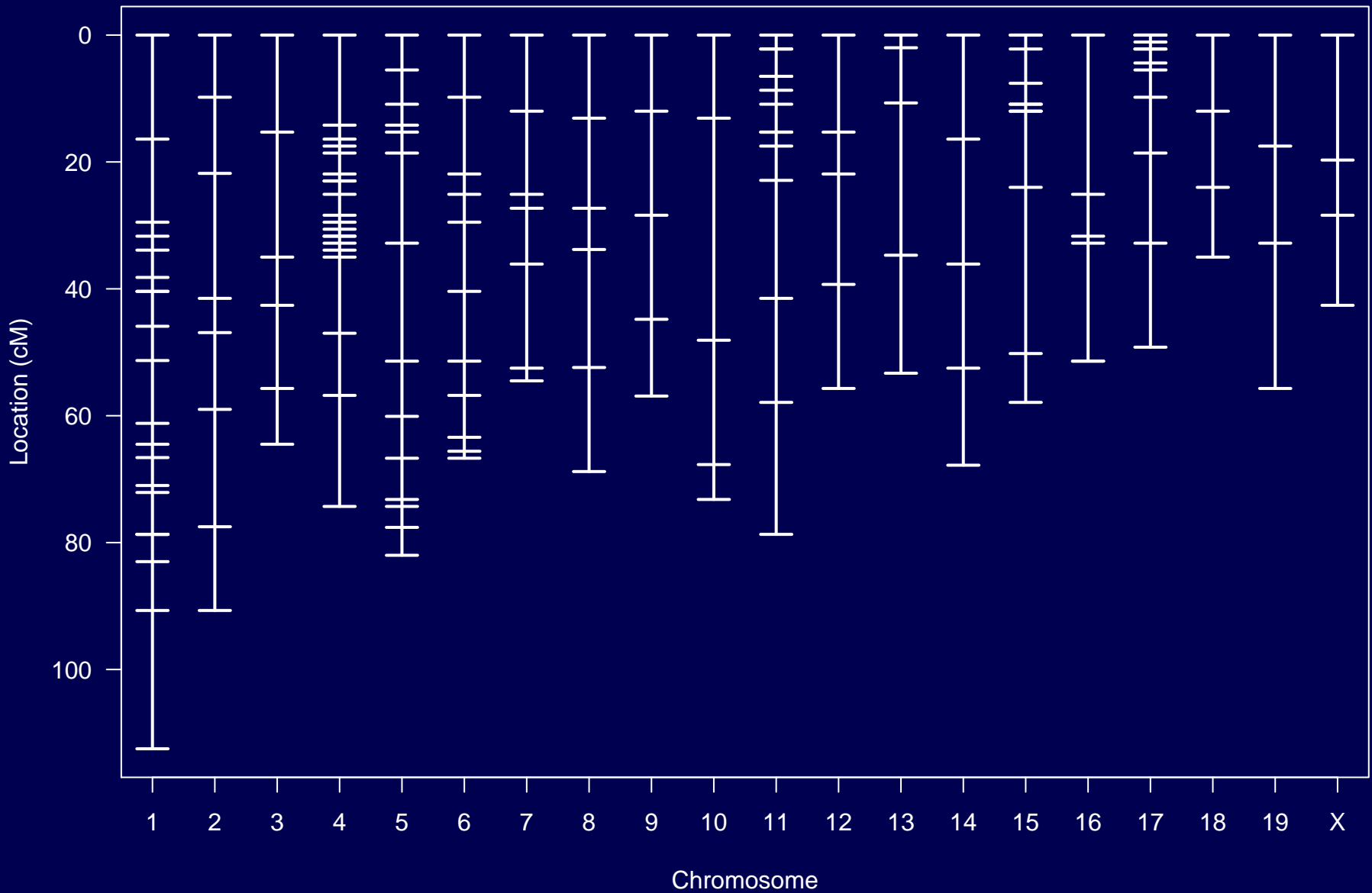
Sugiyama et al. Genomics 71:70-77, 2001

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

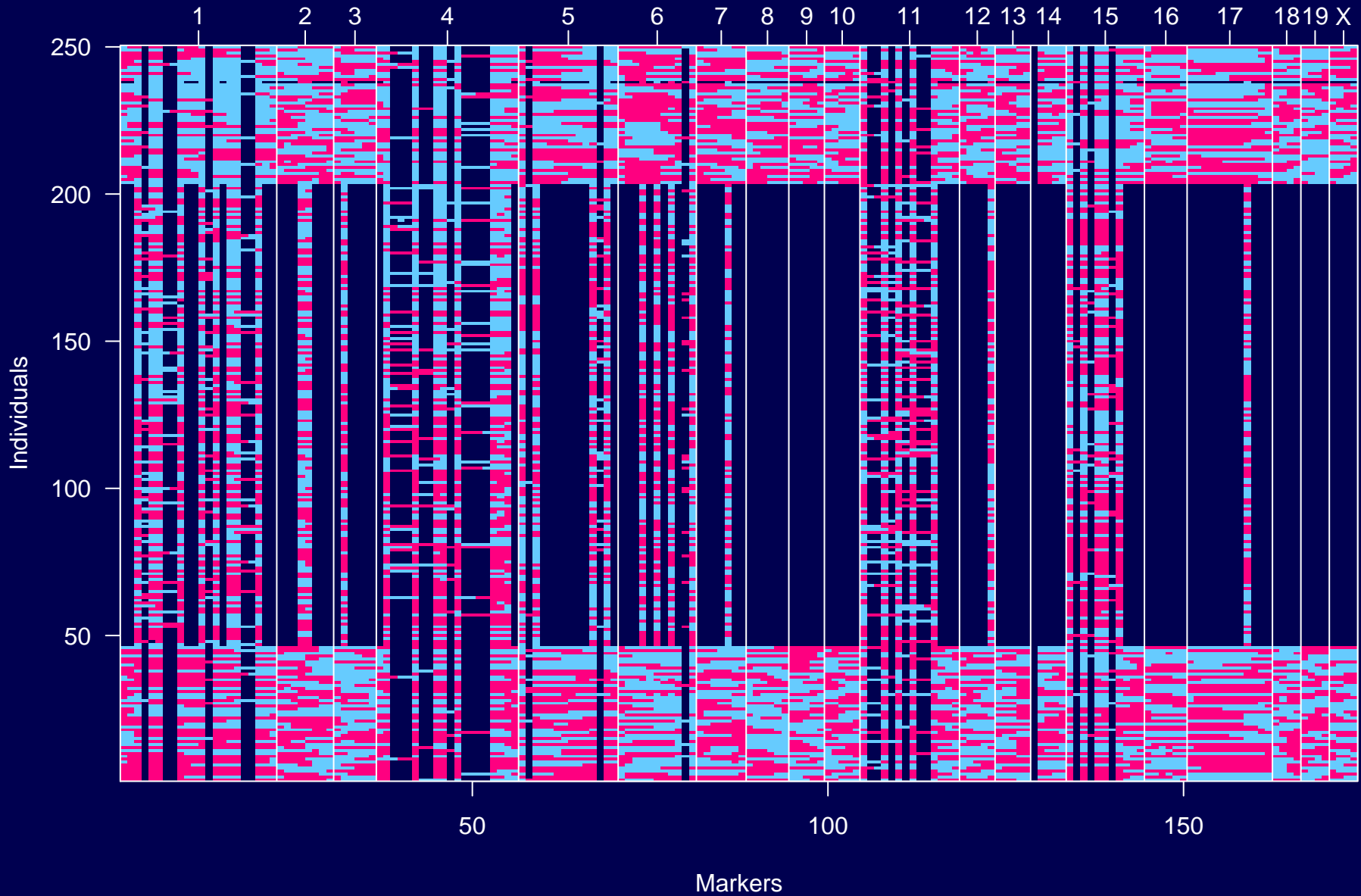
Blood pressure after two weeks drinking water with 1% NaCl



Genetic map



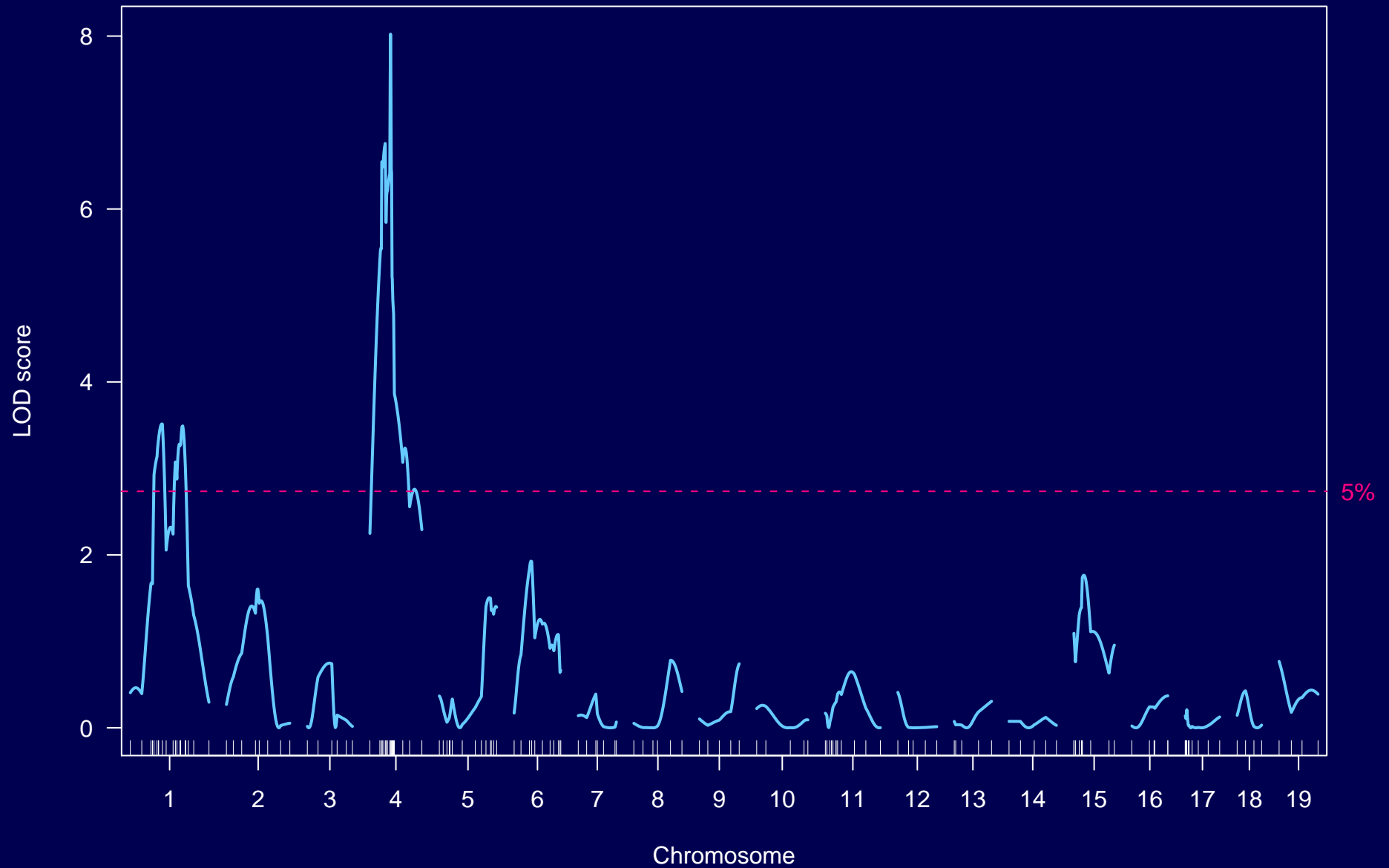
Genotype data



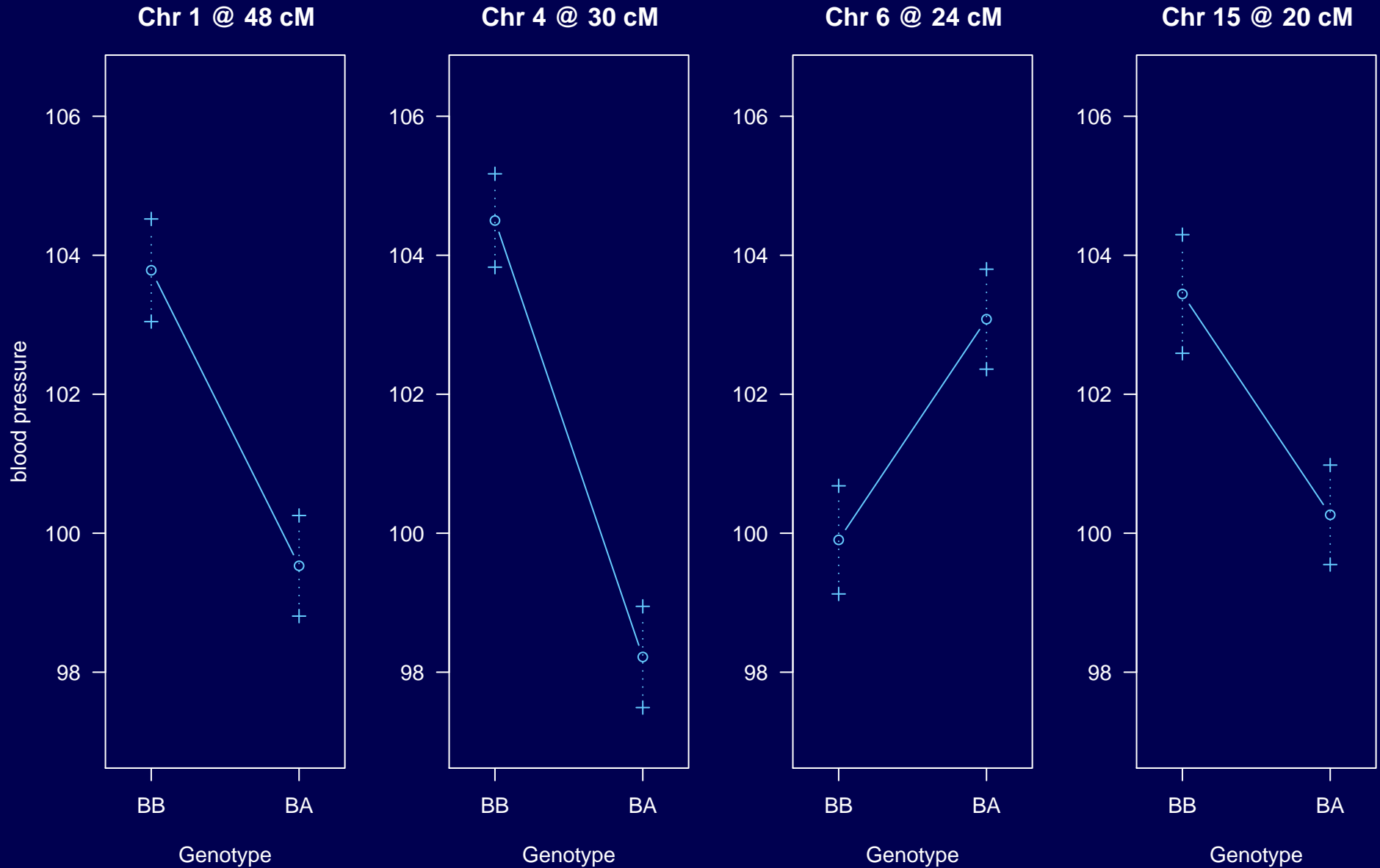
Goals

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

2-dim, 2-QTL scan

For all pairs of positions, fit the following models:

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

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

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

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

\log_{10} likelihoods:

$$l_f(s, t)$$

$$l_a(s, t)$$

$$l_1(s)$$

$$l_0$$

2-dim, 2-QTL scan

LOD scores:

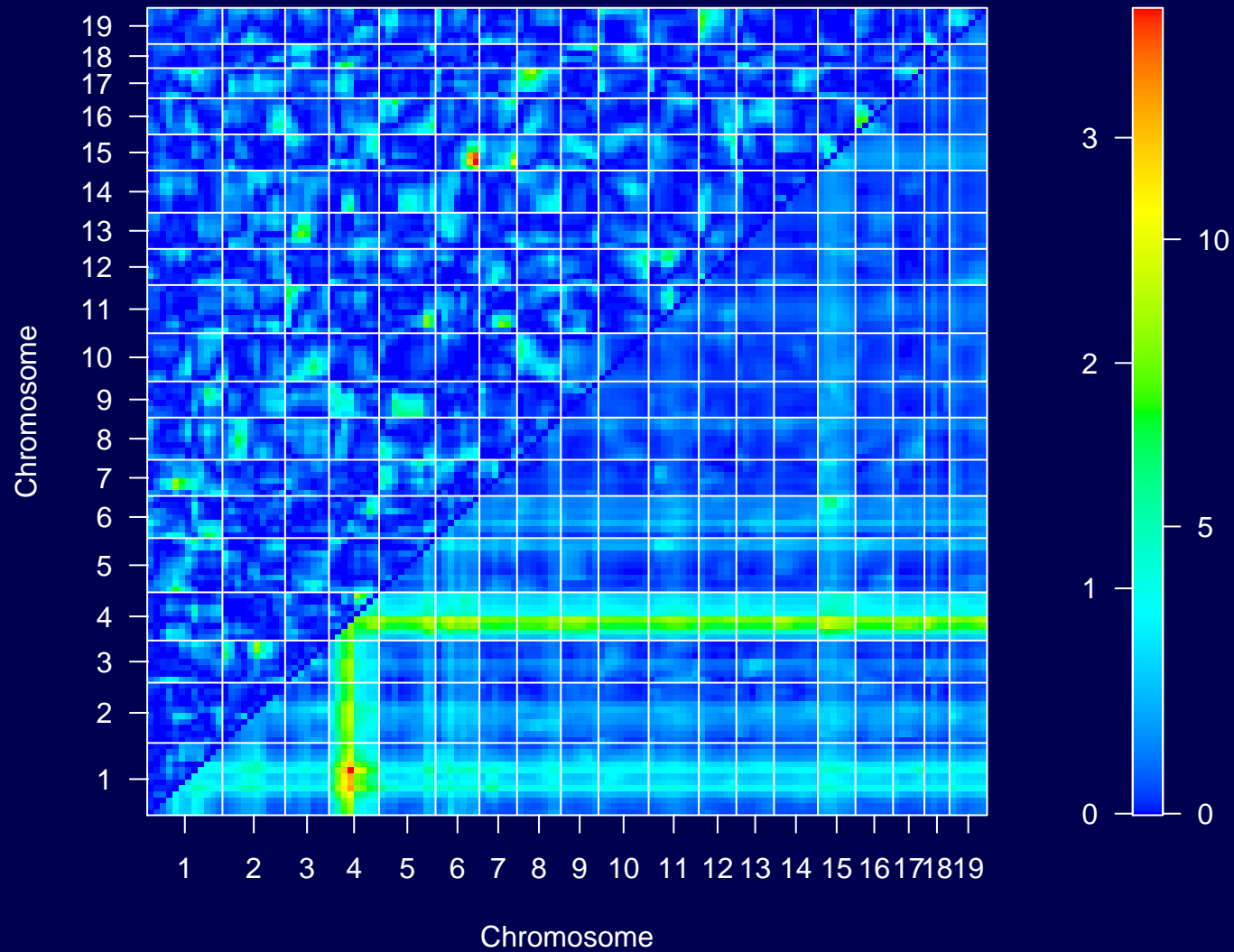
$$\text{LOD}_f(s, t) = l_f(s, t) - l_0$$

$$\text{LOD}_a(s, t) = l_a(s, t) - l_0$$

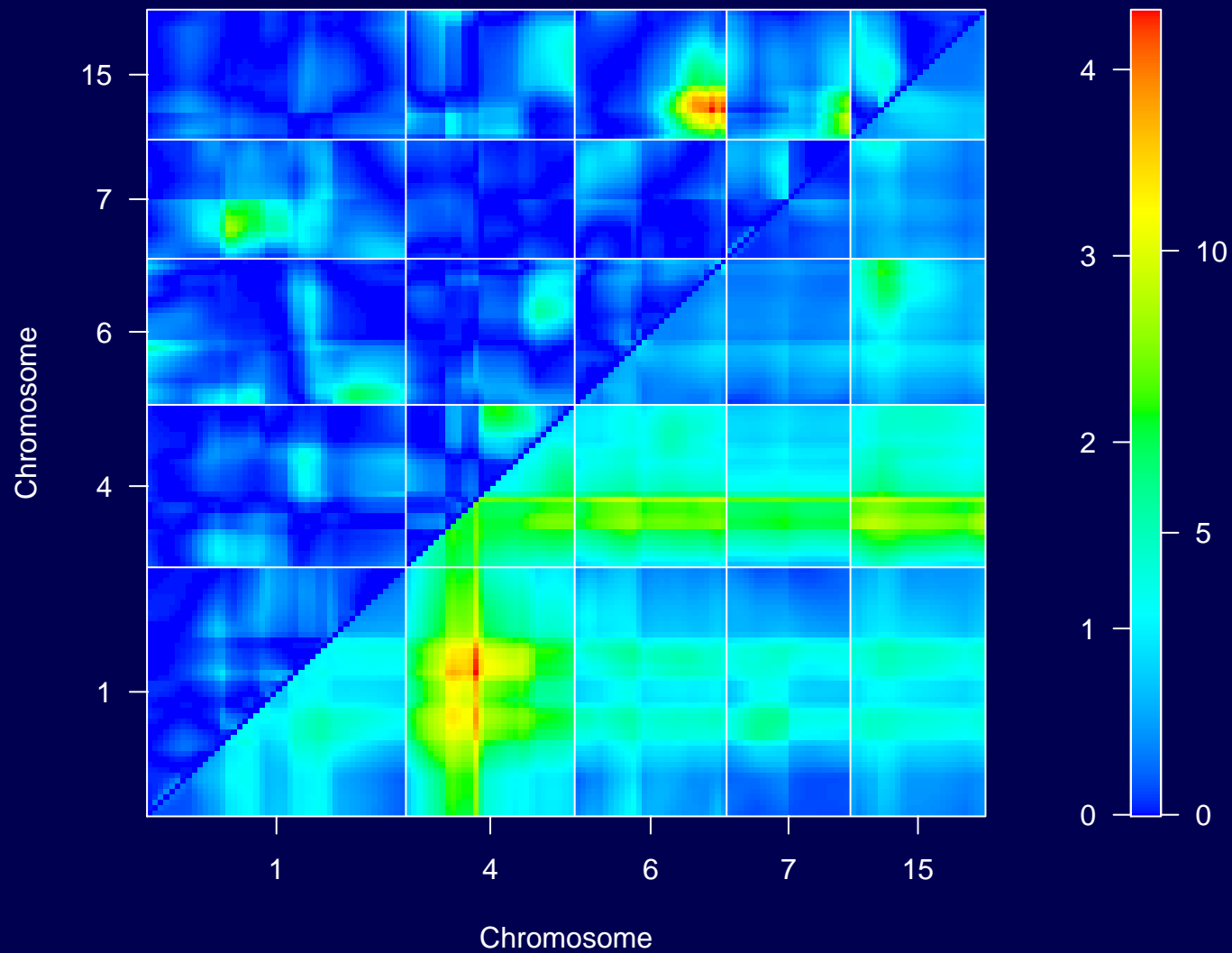
$$\text{LOD}_i(s, t) = l_f(s, t) - l_a(s, t)$$

$$\text{LOD}_1(s) = l_1(s) - l_0$$

Results: LOD_i and LOD_f



Results: LOD_i and LOD_f



Summaries

Consider each pair of chromosomes, (j, k) ,
and let $c(s)$ denote the chromosome for position s .

$$M_f(j, k) = \max_{c(s)=j, c(t)=k} \text{LOD}_f(s, t)$$

$$M_a(j, k) = \max_{c(s)=j, c(t)=k} \text{LOD}_a(s, t)$$

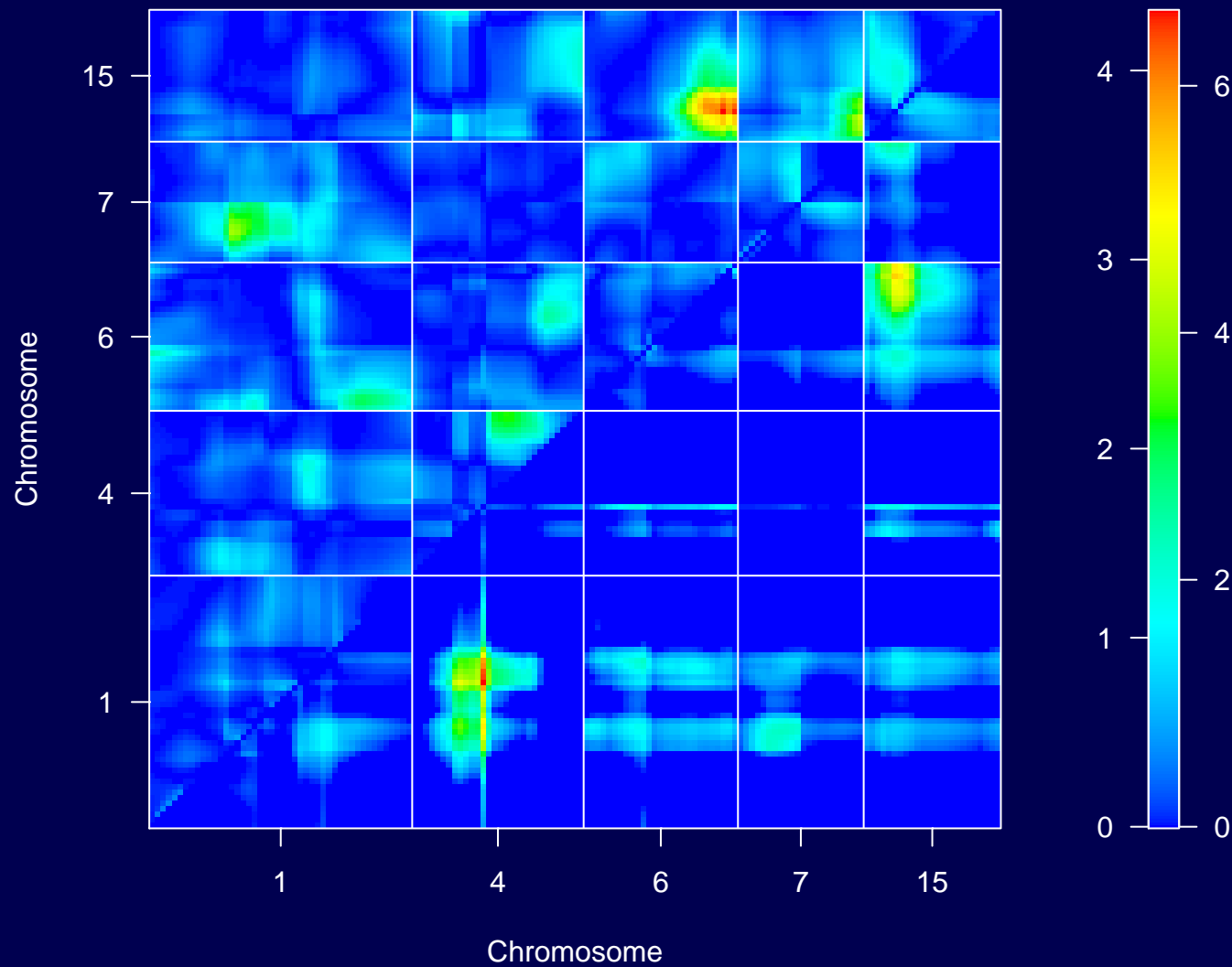
$$M_1(j, k) = \max_{c(s)=j \text{ or } k} \text{LOD}_1(s)$$

$$M_i(j, k) = M_f(j, k) - M_a(j, k)$$

$$M_{fv1}(j, k) = M_f(j, k) - M_1(j, k)$$

$$M_{av1}(j, k) = M_a(j, k) - M_1(j, k)$$

Results: LOD_i and LOD_{fv1}



[R/qtI]

Thresholds

A pair of chromosomes (j, k) is considered interesting if:

$$M_f(j, k) > T_f \quad \text{and} \quad \{ M_{fv1}(j, k) > T_{fv1} \text{ or } M_i(j, k) > T_i \}$$

or

$$M_a(j, k) > T_a \quad \text{and} \quad M_{av1}(j, k) > T_{av1}$$

where the thresholds $(T_f, T_{fv1}, T_i, T_a, T_{av1})$ are determined by a permutation test with a 2d scan

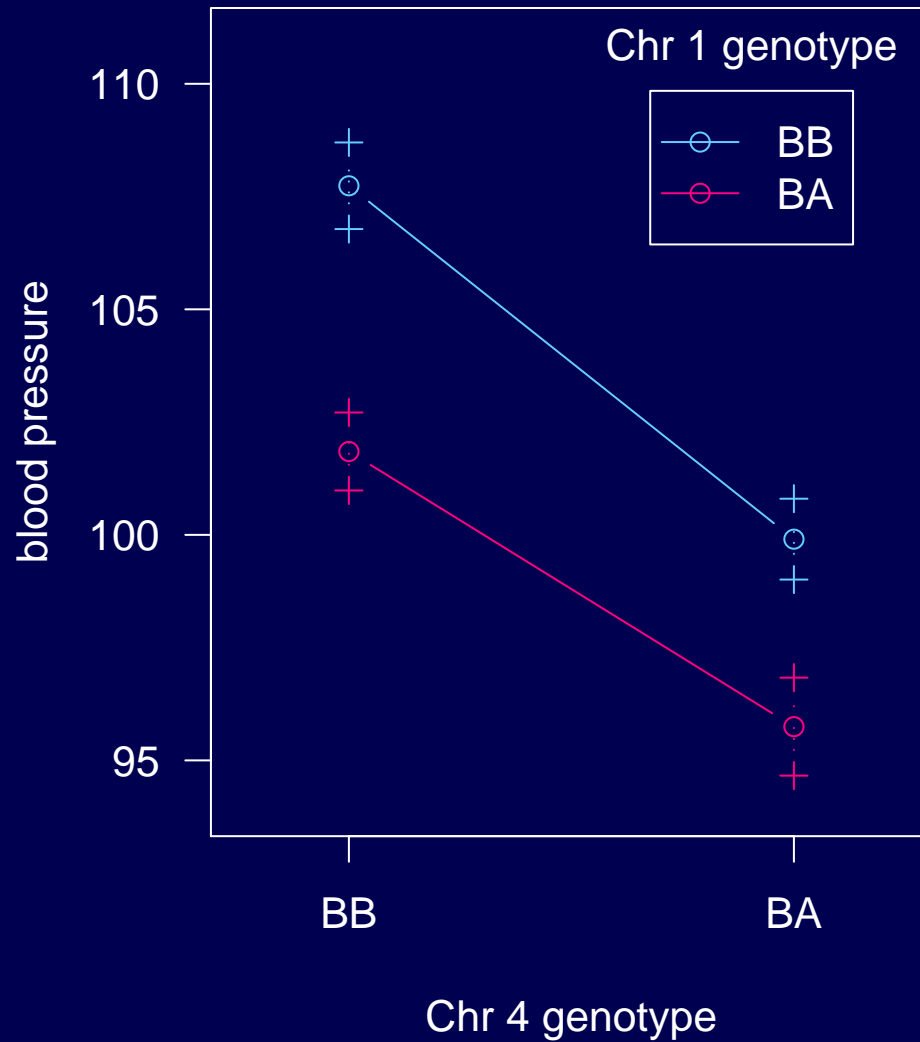
2d scan summary

	pos1f	pos2f	lod.full	lod.fv1	lod.int
c1:c4	71.3	30.0	14.36	6.78	0.27
c6:c15	55.0	20.5	6.91	4.95	2.92
c1:c1	39.3	78.3	5.10	1.58	0.09

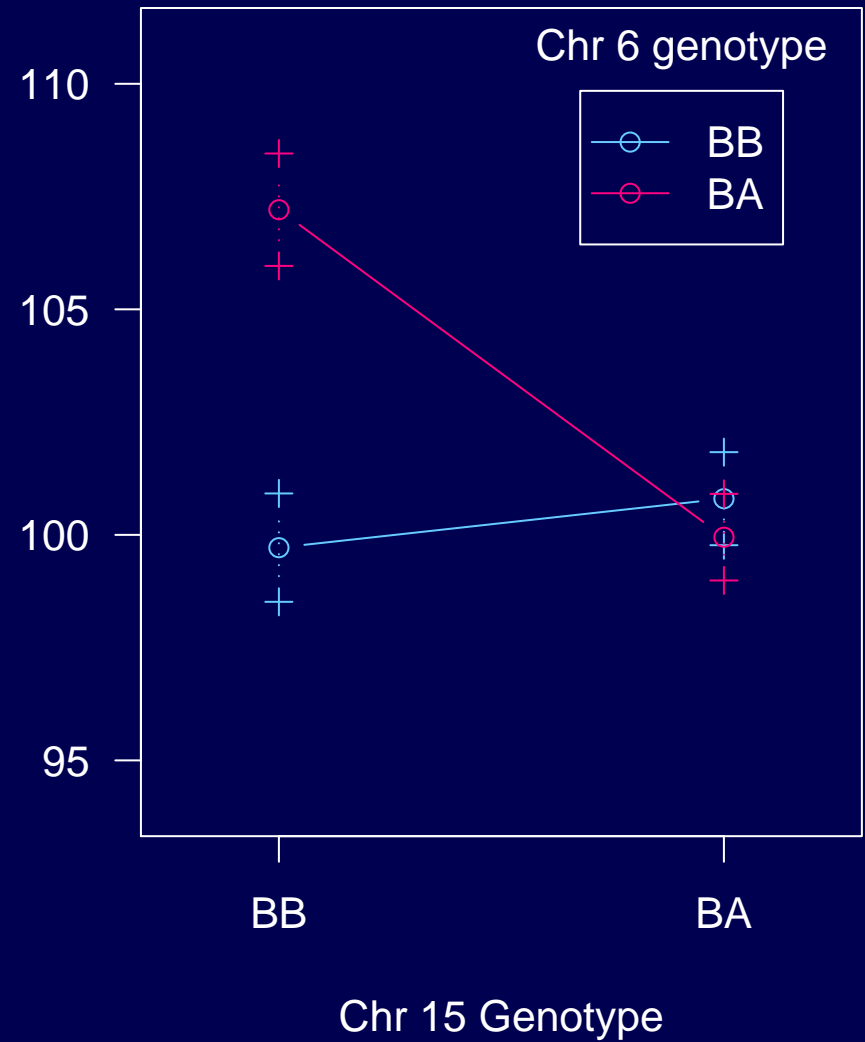
	pos1a	pos2a	lod.add	lod.av1
c1:c4	68.3	30.0	14.09	6.50
c6:c15	24.0	22.5	3.99	2.03
c1:c1	48.3	79.3	5.02	1.50

Estimated effects

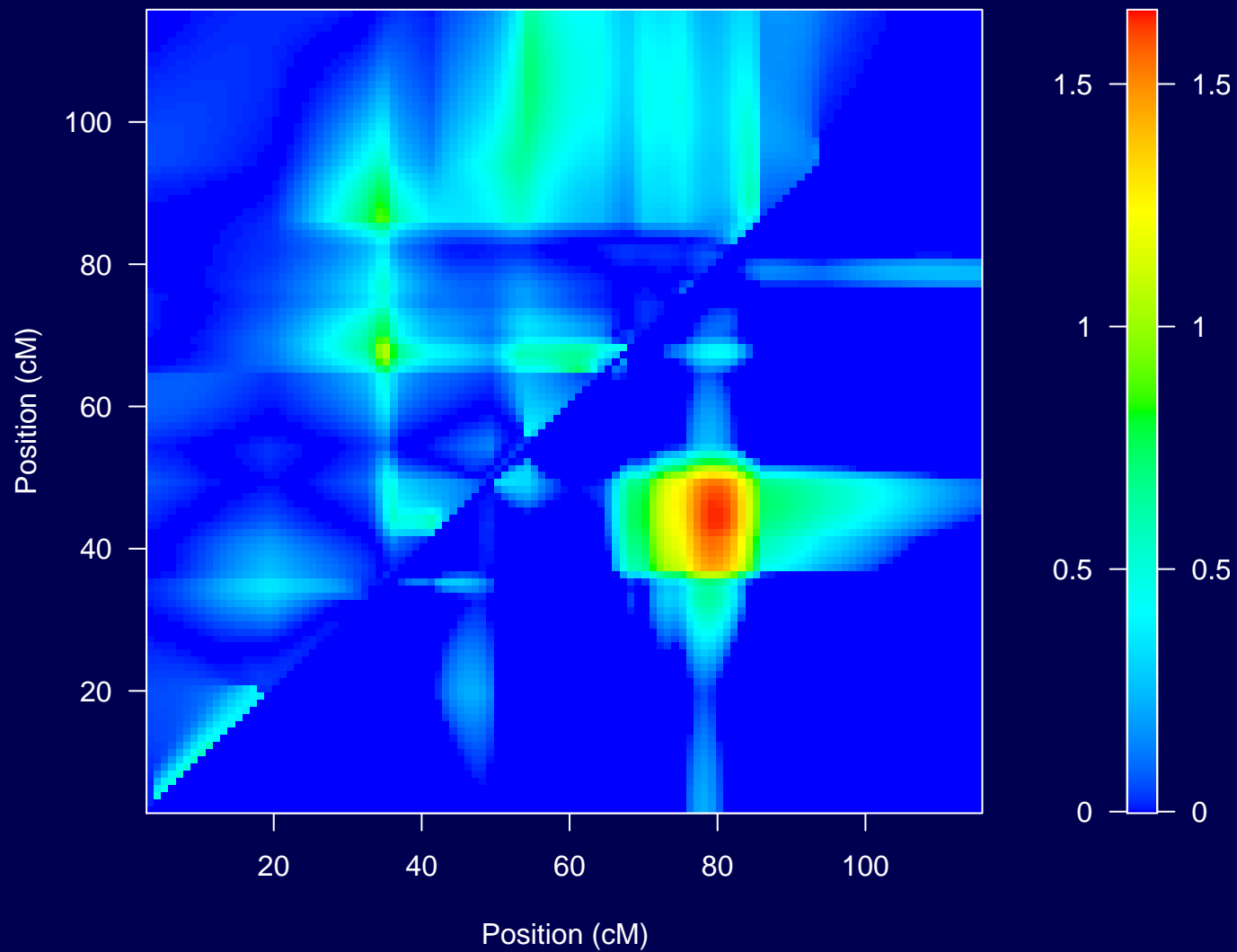
1 x 4



6 x 15



Chr 1: LOD_i and LOD_{av1}



[R/qtI]

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

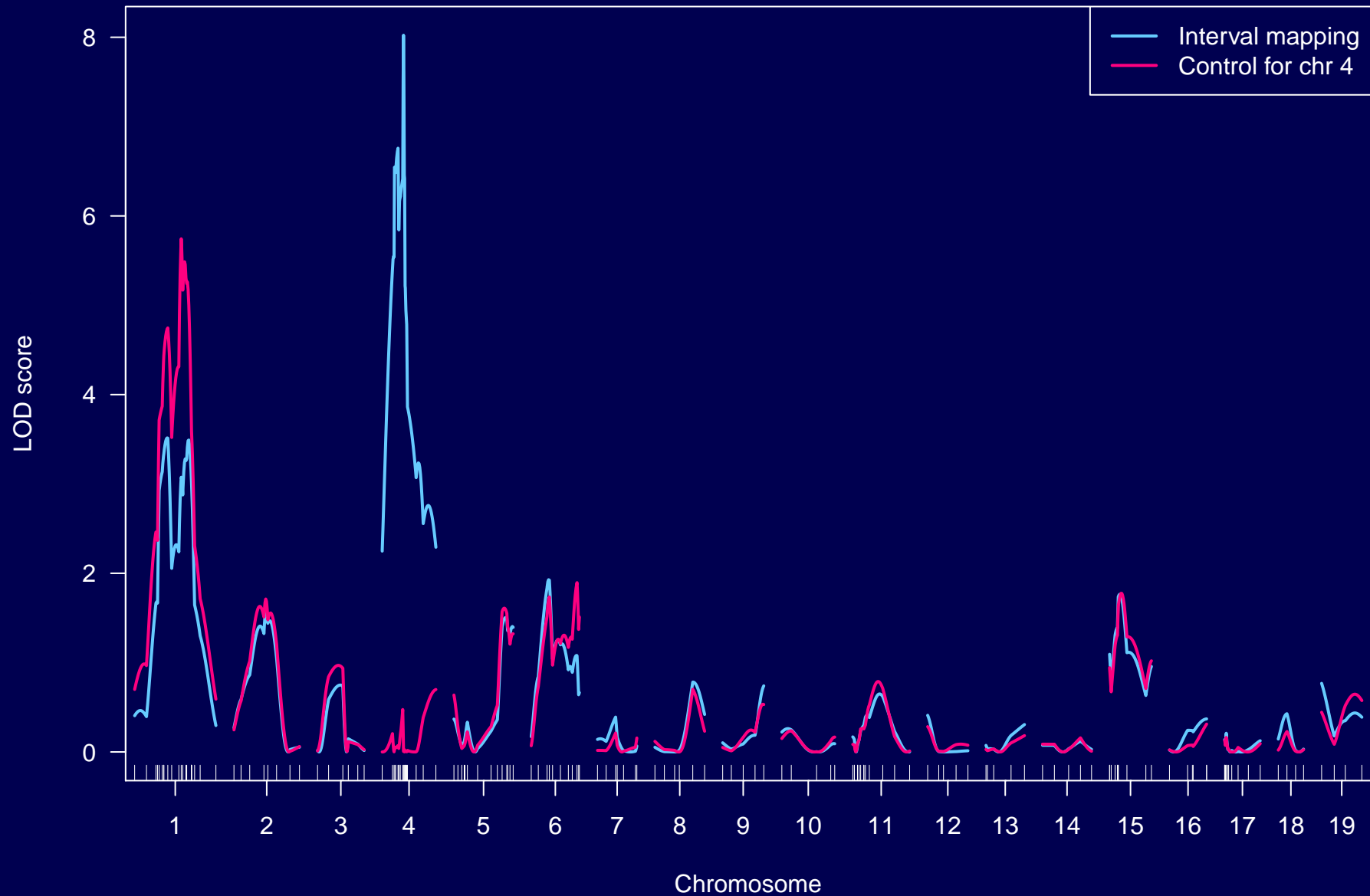
Exploratory methods

- 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

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

[R/ql]

Automation

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

Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

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

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

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

[R/ql]

Epistasis

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

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

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.

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

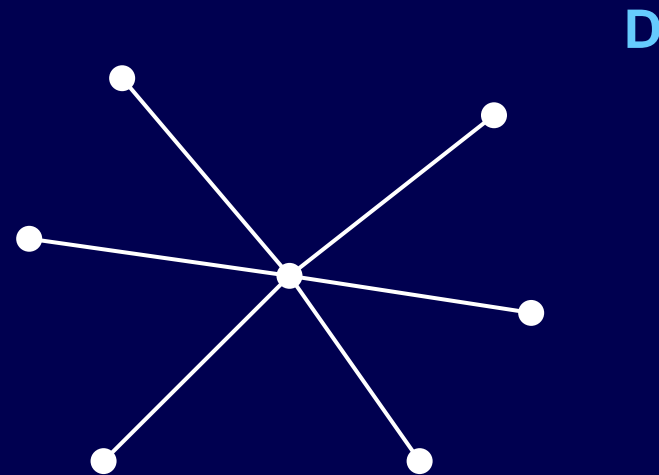
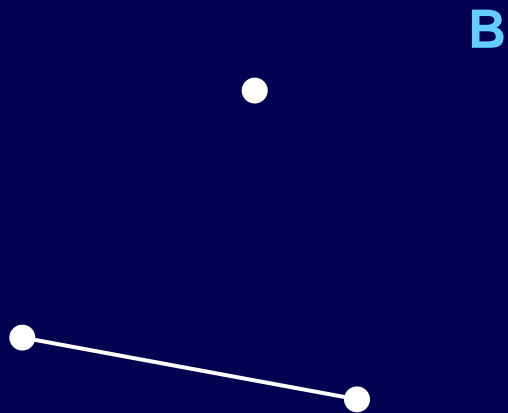
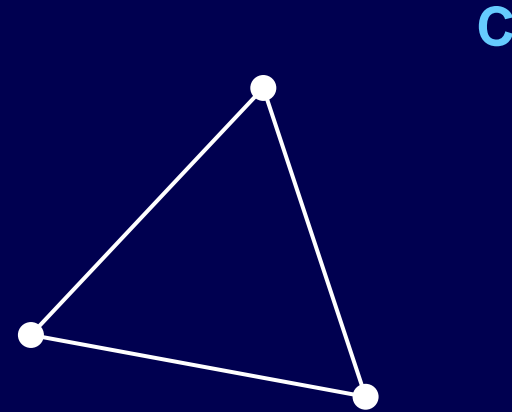
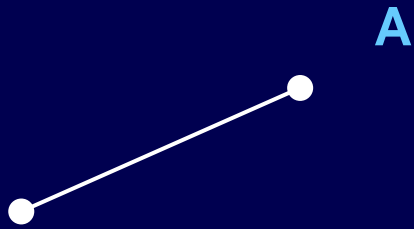
For the mouse genome:

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

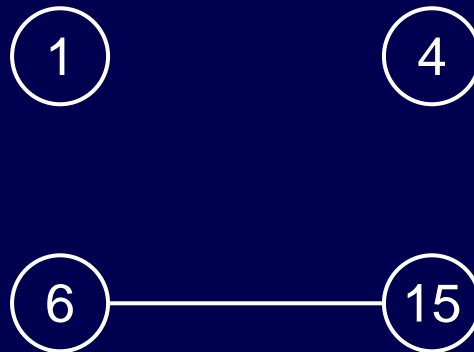
$$T_i^H = 2.62 \text{ (BC) or } 4.28 \text{ (F}_2\text{)}$$

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

Models as graphs

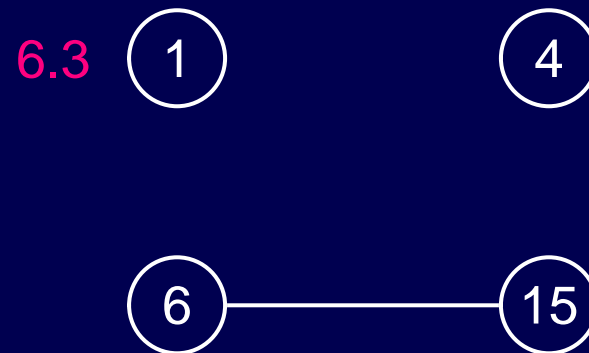


Results



LOD = 23.1

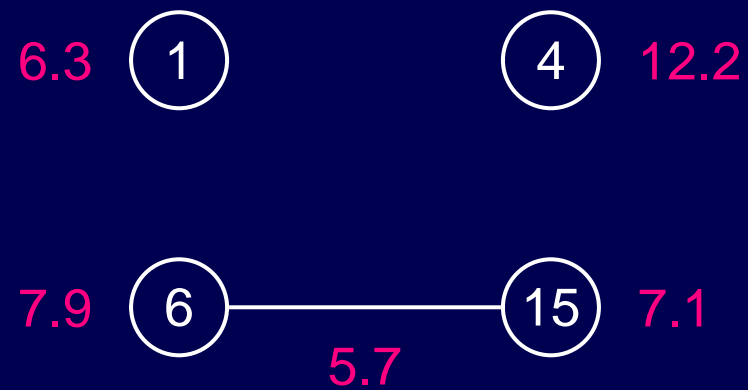
Results



LOD = 23.1

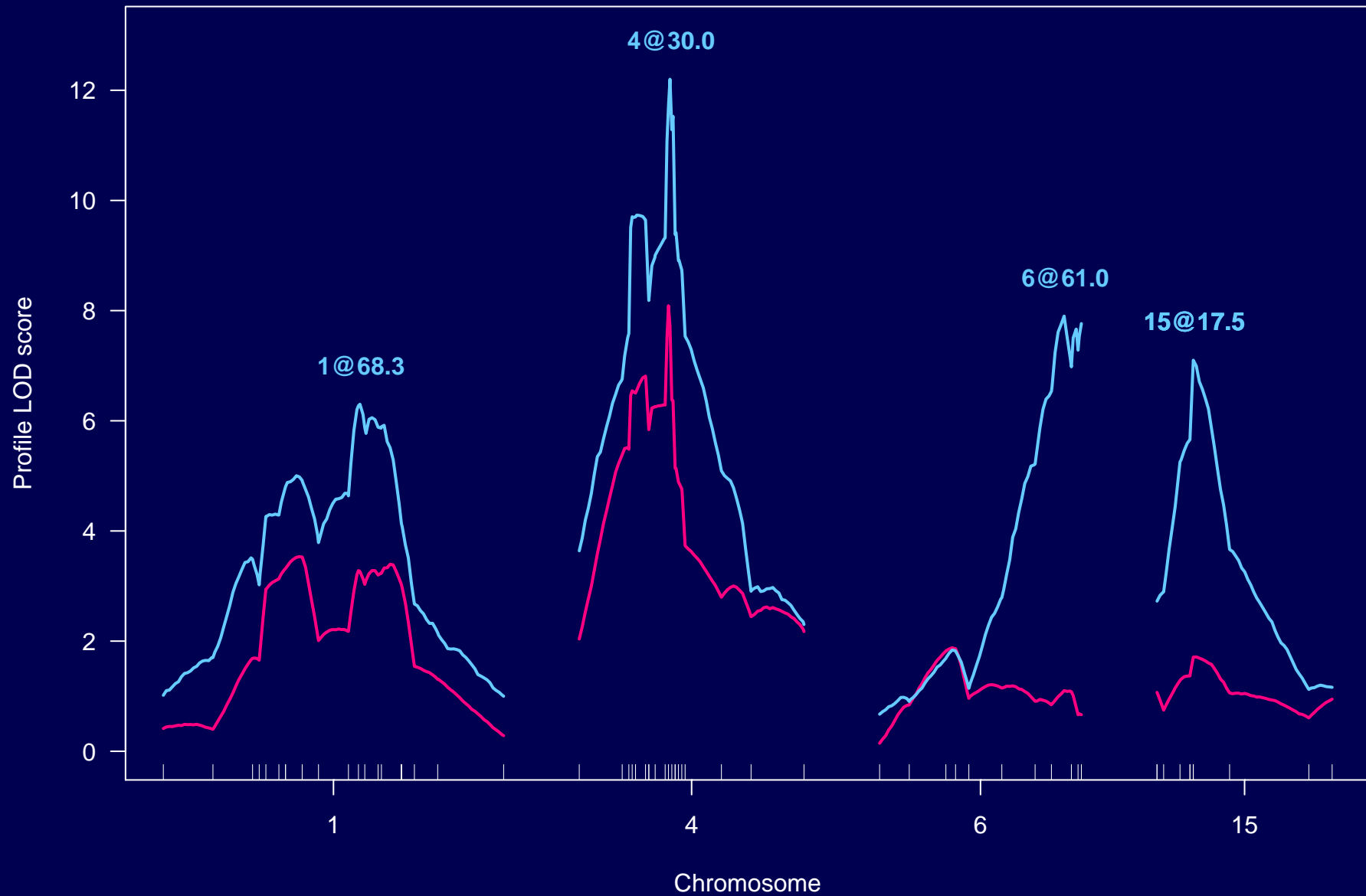
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Results



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

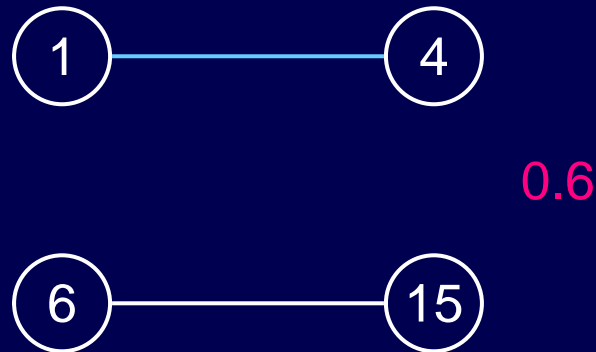
Profile LOD curves



Drop-one-QTL table

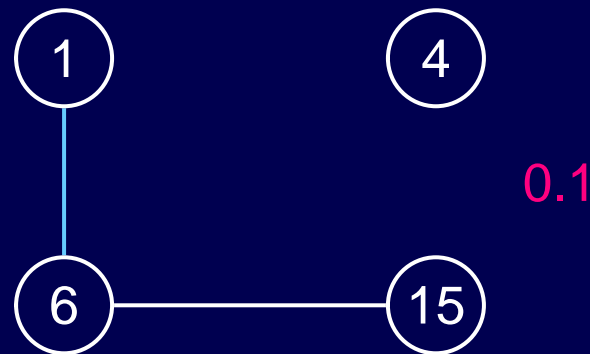
	df	LOD	%var
1@68.3	1	6.30	11.0
4@30.0	1	12.21	20.1
6@61.0	2	7.93	13.6
15@17.5	2	7.14	12.3
6@61.0 : 15@17.5	1	5.68	9.9

Add an interaction?



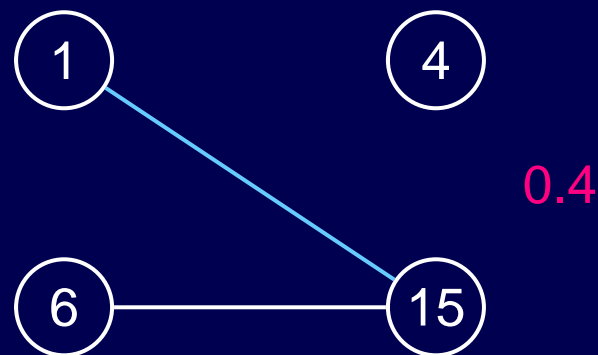
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add an interaction?



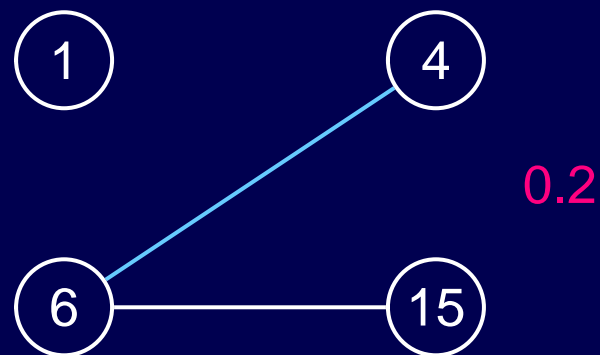
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add an interaction?



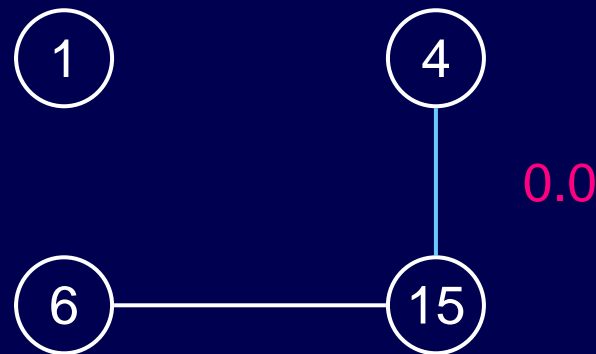
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add an interaction?



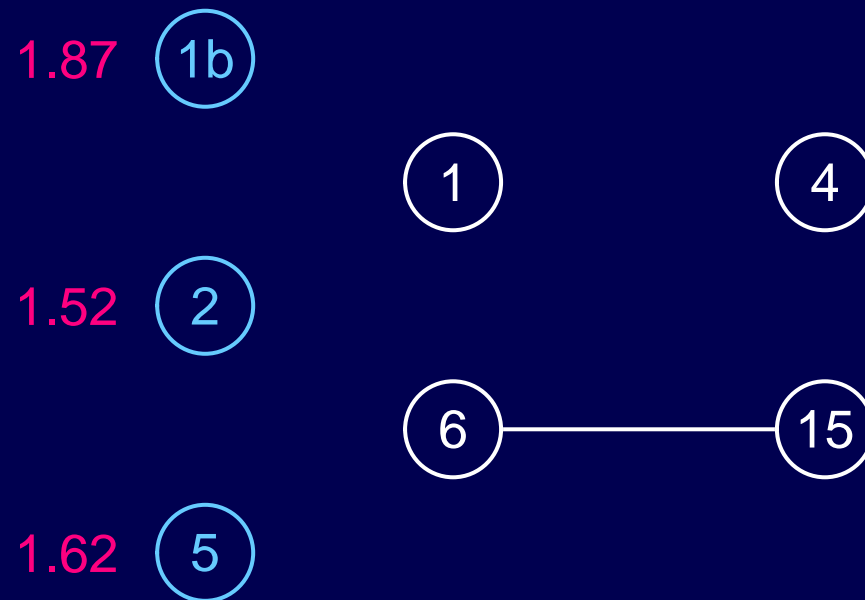
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add an interaction?



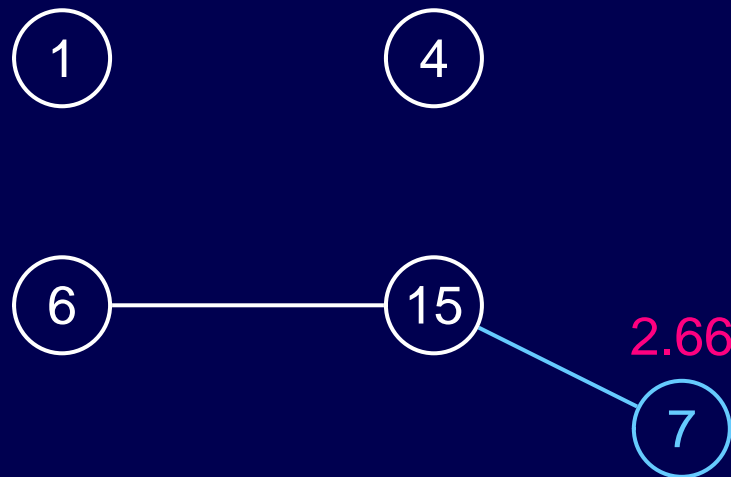
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add another QTL?



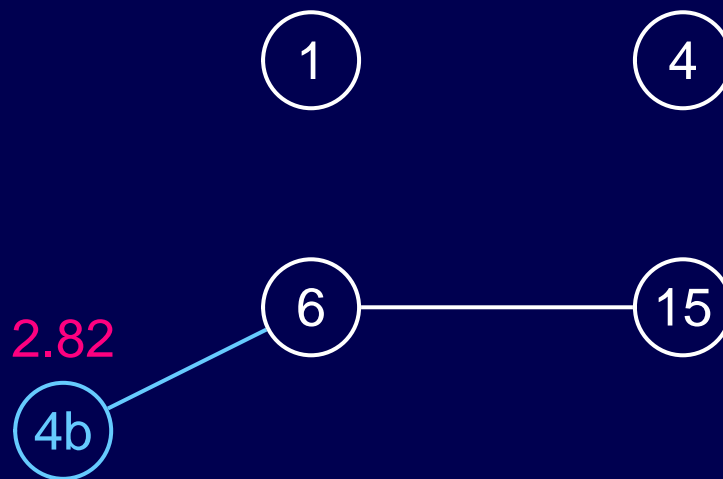
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

Add another QTL?



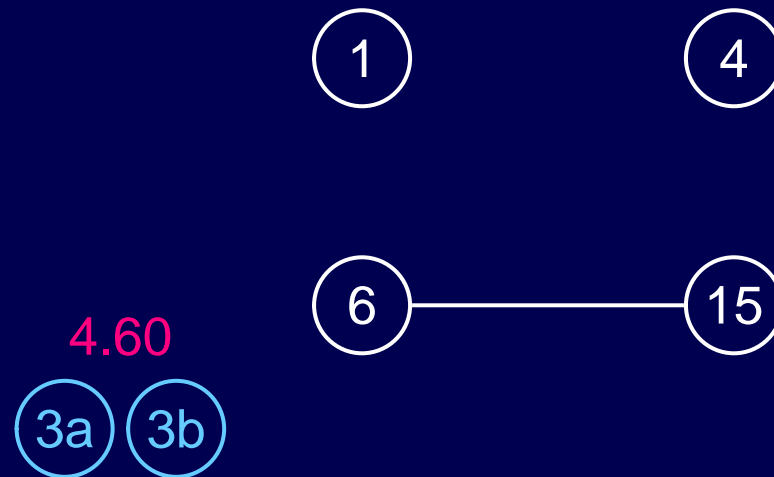
$$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$$

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?



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

Summary

- QTL mapping is a model selection problem
- 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