

# QTL Mapping II:

## Model Selection

---

Karl W Broman

Department of Biostatistics  
Johns Hopkins University

[kbroman@jhsph.edu](mailto:kbroman@jhsph.edu)

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

## Outline

---

- The model selection problem:
  - Class of models
  - Compare models
  - Search model space
  - Assess the performance of a procedure
- A simulation study

- 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**.

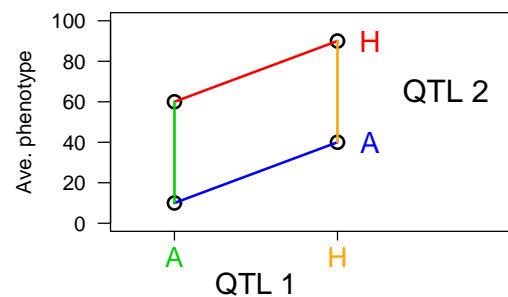
What set of QTLs are well-supported?

Is there evidence for QTL-QTL interactions?

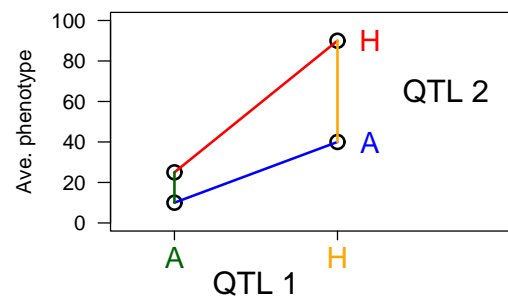
**Model** = a defined set of QTLs and QTL-QTL interactions (and possibly covariates and QTL-covariate interactions).

## Epistasis in a backcross

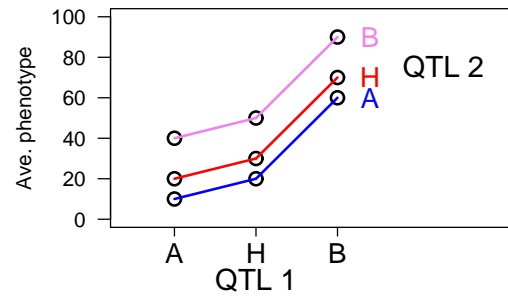
Additive QTLs



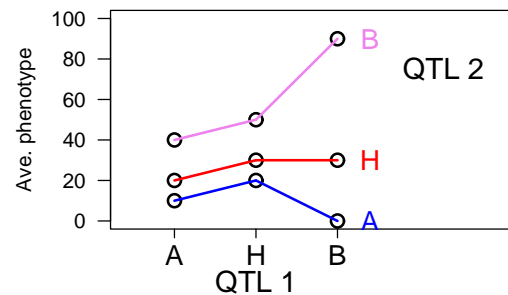
Interacting QTLs



## Additive QTLs



## Interacting QTLs



# Two-dimensional genome scan

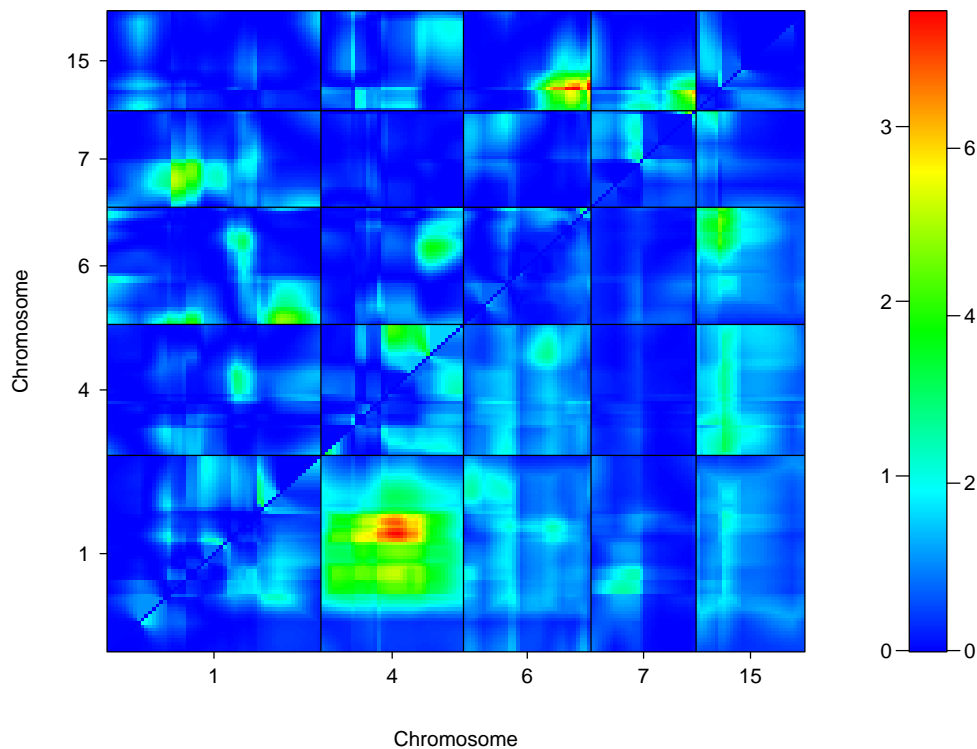
Consider each pair of positions,  $(\lambda_1, \lambda_2)$

### Models

- Full
- Additive
- QTL 1
- QTL 2
- Null

### Possible comparisons

- Full vs. null
- Full vs. additive
- Full vs. Best of QTL 1 & 2
- Add've vs. Best of QTL 1 & 2



## Perfect data situation

To ease discussion, we'll focus on a simple special case:

- Complete marker genotype data
- Markers are only putative QTLs
- Normally distributed residuals

Example model (in a backcross):

$$y_i = \mu + \Delta_1 q_{i1} + \Delta_2 q_{i2} + \Delta_3 q_{i3} + \epsilon_i \quad \epsilon_i \text{ are iid } N(0, \sigma^2)$$

$q_j$  are 0/1 variables (QTL genotypes)

$\mu, \Delta$ 's are parameters, estimated by least squares

**Fitted values:**  $\hat{y}_i = \hat{\mu} + \hat{\Delta}_1 q_{i1} + \hat{\Delta}_2 q_{i2} + \hat{\Delta}_3 q_{i3}$

**RSS =  $\sum_i (y_i - \hat{y}_i)^2$  indicates model fit.**

1. Class of models
2. Compare models
3. Search model space
4. Assess performance of a procedure

**Note:**

2 and 4 are much the same.

There might be a 0th item: Method for model fitting.  
(e.g., imputation, EM, etc.)

## Class of models

---

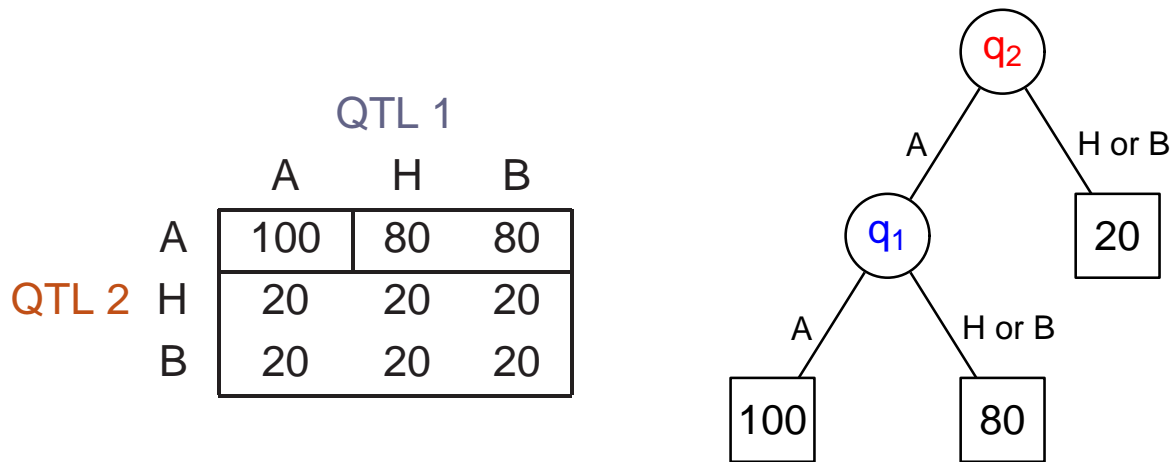
- Additive models

$$y = \mu + \sum_j \Delta_j q_j + \epsilon$$

- Also pairwise interactions  
Preserve hierarchy: if include interaction, also include both main effects.
- Also higher-order interactions  
Again, preserve hierarchy.
- Regression trees

# Example regression tree

11



## Intercrosses: class of models

12

- Always bring in both degrees of freedom with a QTL

or

Try to distinguish additivity/dominance/recessiveness?

A	H	B
---	---	---

- Always bring in all four d.f. with a QTL:QTL interaction

or

Try to distinguish ways to subdivide the 3×3 table?

	A	H	B
A			
H			
B			

## Compare models

---

- Imagine you could fit all possible models; which would you like best?
- This issue is like LOD thresholds, but more complicated.
- For models with the same number of parameters (QTLs and interactions), we prefer that with the “best fit” (smallest RSS or largest likelihood).
- If you fit more parameters, you’ll get a “better fit.”
  - How much better, before including additional terms?
  - I like a form of penalized likelihood.
- Note: Bayesians must also confront this issue (through the prior on models).

## The additive QTL case

---

n backcross mice; M markers

$x_{ij}$  = genotype (1/0) of mouse  $i$  at marker  $j$

$y_i$  = phenotype (trait value) of mouse  $i$

$$y_i = \mu + \sum_{j=1}^M \Delta_j x_{ij} + \epsilon_i \quad \text{Which } \Delta_j \neq 0?$$

$$\text{BIC}_\delta = \log \text{RSS} + \text{no. markers} \times \left( \delta \times \frac{\log n}{n} \right)$$

## Why $BIC_\delta$ ?

---

- For a fixed no. markers, letting  $n \rightarrow \infty$ ,  $BIC_\delta$  is consistent.
- There exists a prior (on models + coefficients) for which  $BIC_\delta$  is the  $-\log$  posterior.
- $BIC_\delta$  is essentially equivalent to use of a threshold on the conditional LOD score
- It performs well.

## $BIC_\delta \longleftrightarrow$ conditional LOD

---

Conditional LOD score:

$$\text{LOD}(x_k^* \mid x_1^*, \dots, x_{k-1}^*) = \frac{n}{2} \log_{10} \left\{ \frac{\text{RSS}(x_1^*, \dots, x_{k-1}^*)}{\text{RSS}(x_1^*, \dots, x_k^*)} \right\}$$

Minimizing  $BIC_\delta$  is approximately equivalent to choosing the largest **k** such that

$$\text{LOD}(x_k^* \mid x_1^*, \dots, x_{k-1}^*) \geq \frac{\delta}{2} \log_{10} n$$



## Choice of $\delta$

---

Smaller  $\delta$ : include more loci; higher false positive rate

Larger  $\delta$ : include fewer loci; lower false positive rate

Let  $L = 95\%$  genome-wide LOD threshold  
(compare single-QTL models to the null model)

Choose  $\delta = 2 L / \log_{10} n$

With this choice of  $\delta$ , in the absence of QTLs, we'll include at least one **extraneous** locus, 5% of the time.

Note that now we have

$$\text{BIC}_\delta = \log_{10} \text{RSS} + \text{no. markers} \times \left( \frac{2L}{n} \right)$$

## Search model space

---

- Consider the case of additive QTL models, with 100 putative QTLs.
- There are  $2^{100} \approx 10^{30}$  possible models, far more than can be inspected individually.
- Need a way to search through this space, to find the good ones.
- This is really a matter of “grunt work.” (More is better; the tradeoff is with computational time.)

- Forward selection
  - Find the best single-QTL model:  $q_1^*$ .
  - Find the best two-QTL model that includes  $q_1^*$ :  $(q_1^*, q_2^*)$ .
  - Find the best three-QTL model that includes  $q_1^*, q_2^*$ :  $(q_1^*, q_2^*, q_3^*)$ .
  - Etc.
- Backward elimination
- Forward selection followed by backward elimination
- Stepwise selection
- Randomized algorithms (e.g., MCMC, genetic algorithms, etc.)

## Assess performance

- 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**.
- You **can't know** the performance of your procedure with your data—you need to know the truth.
- You **can know**:
  - How a particular procedure performs in simulated cases
  - How a procedure performs in simulated data close to what you've inferred

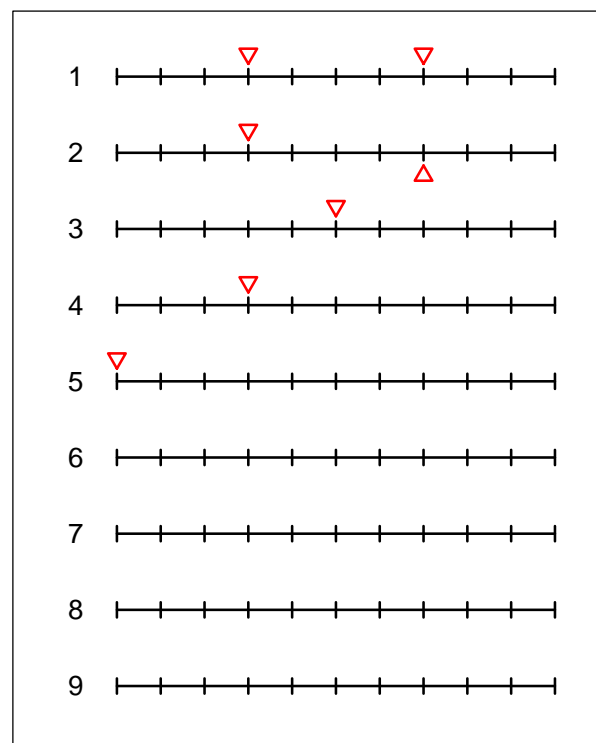
I've said here: pick a good model

Really:

- Want to guide future experiments
- Want some understanding of the uncertainty in different aspects of the chosen model

## A simulation study

- Backcross with  $n=250$
- No crossover interference
- 9 chr, each 100 cM
- Markers at 10 cM spacing; complete genotype data
- 7 QTLs
  - One pair in coupling
  - One pair in repulsion
  - Three unlinked QTLs
- Heritability = 50%
- 2000 simulation replicates



- ANOVA at marker loci
  - Composite interval mapping (CIM)
  - Forward selection with permutation tests
  - Forward selection with  $BIC_{\delta}$
  - Backward elimination with  $BIC_{\delta}$
  - FS followed by BE with  $BIC_{\delta}$
  - MCMC with  $BIC_{\delta}$
- A **selected marker** is deemed **correct** if it is within 10 cM of a QTL (i.e., correct or adjacent)

## A simplified version of CIM

Select a set of markers,  $S$

(e.g., by FS to a fixed number)

For each marker,  $x$ , in the genome:

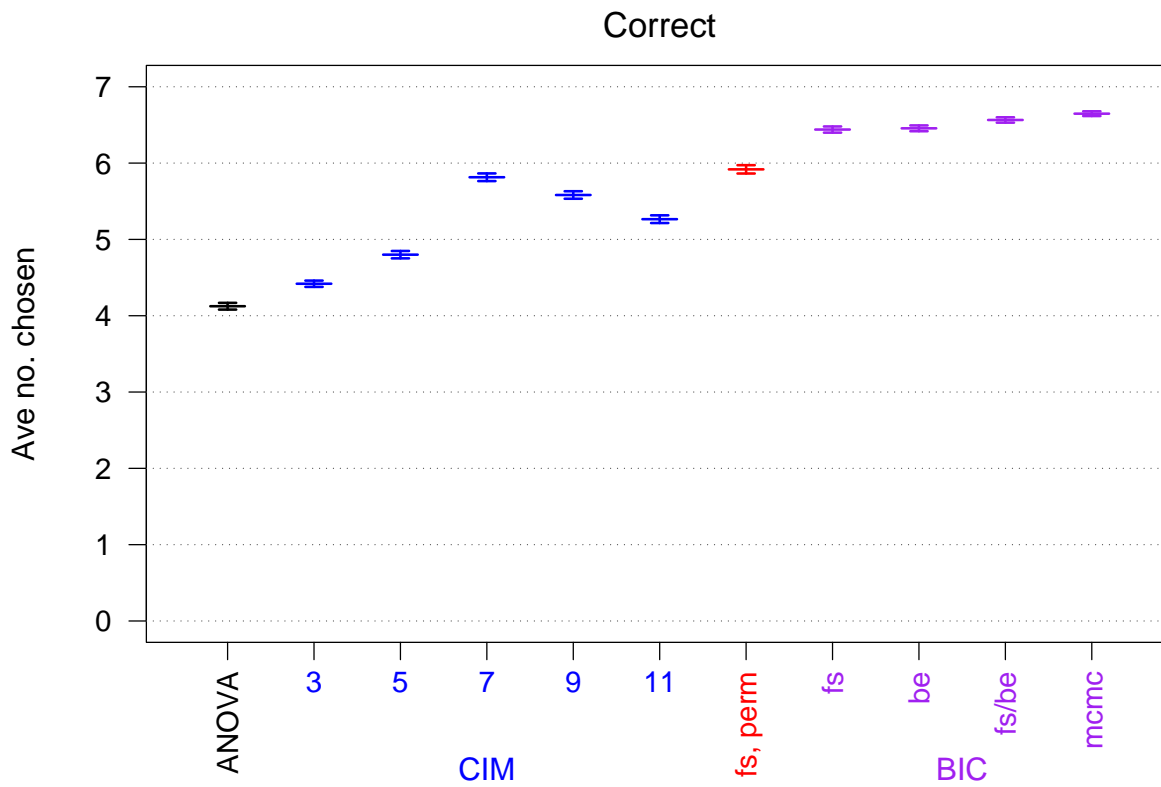
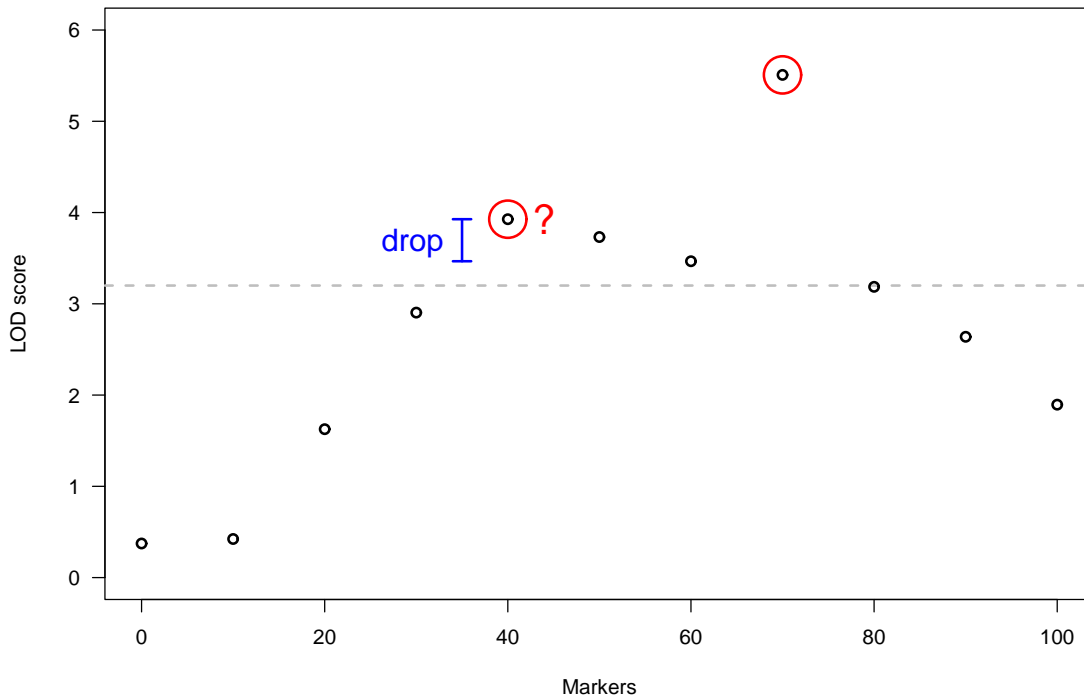
(a) If  $x \notin S$ , calculate  $LOD(x | S)$

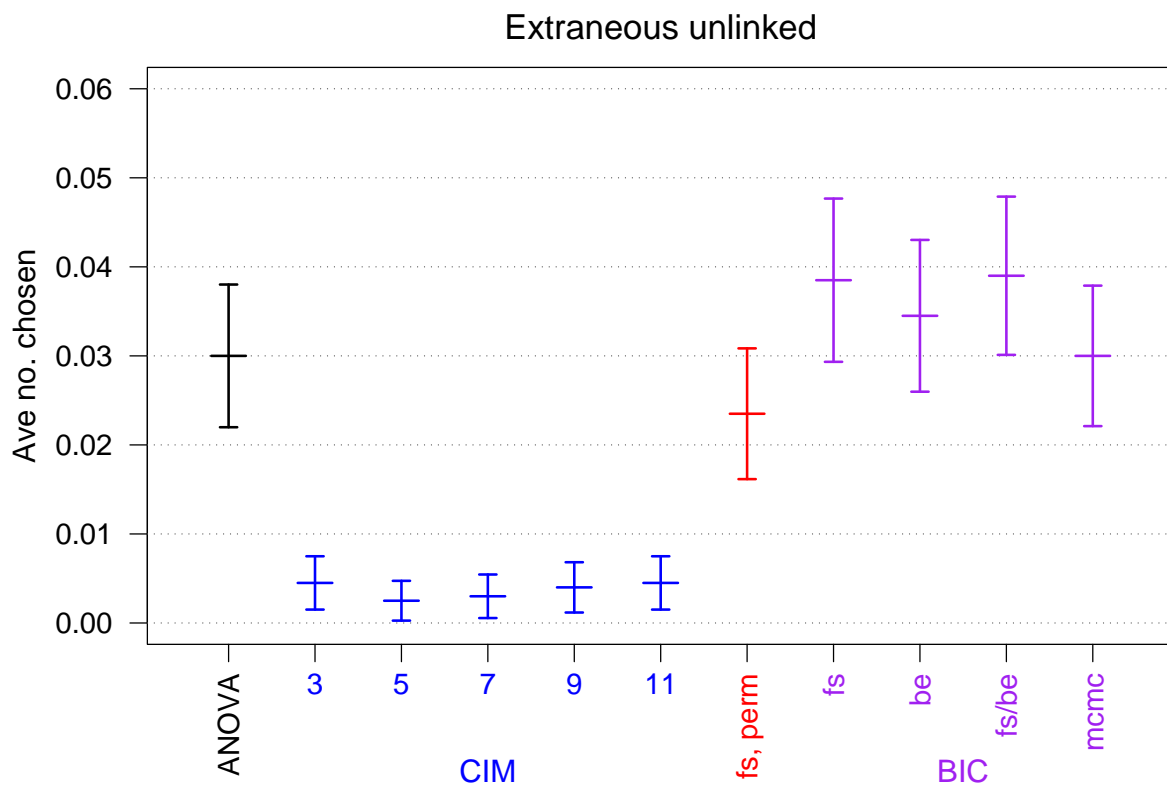
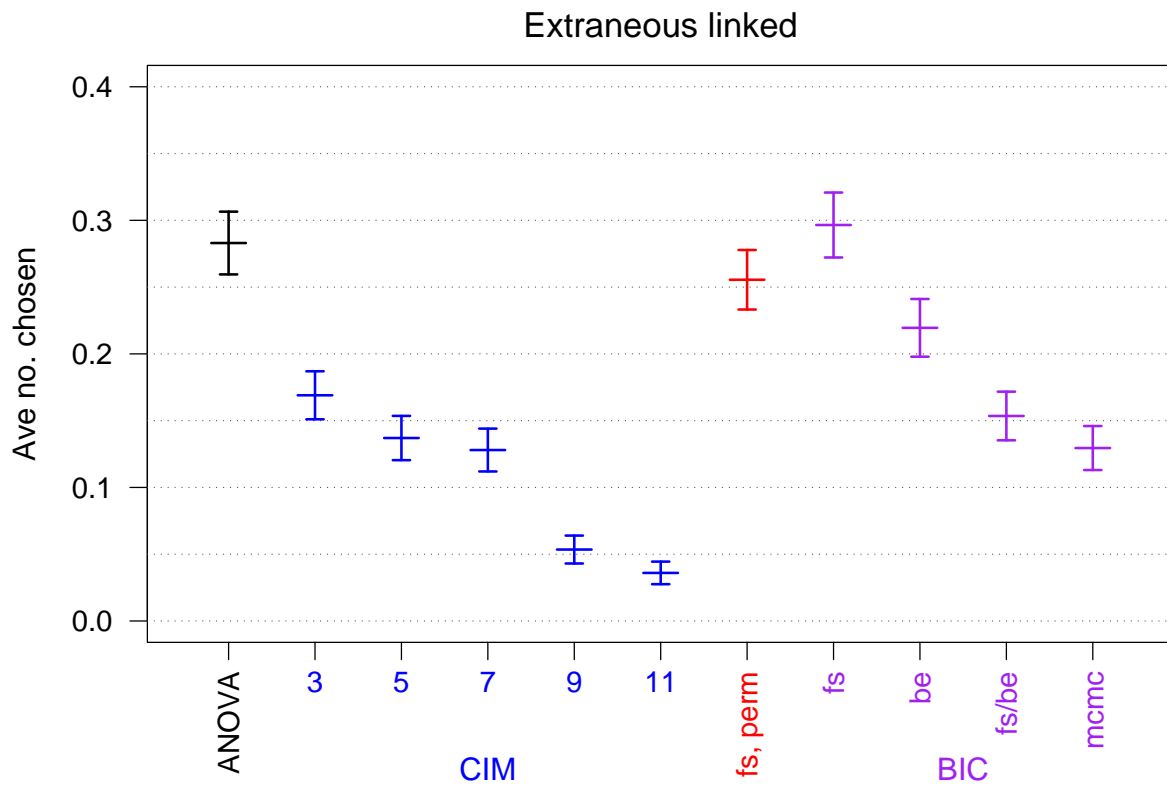
(b) If  $x \in S$ , calculate  $LOD(x | S \setminus \{x\})$

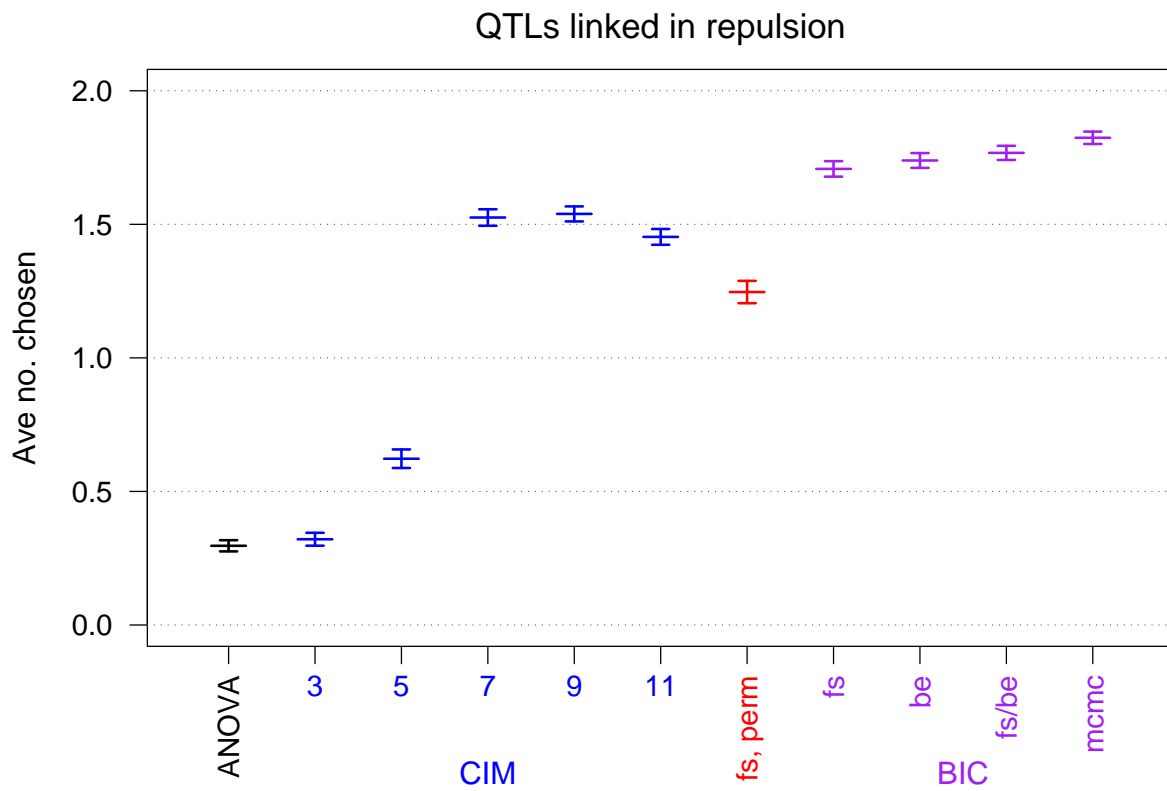
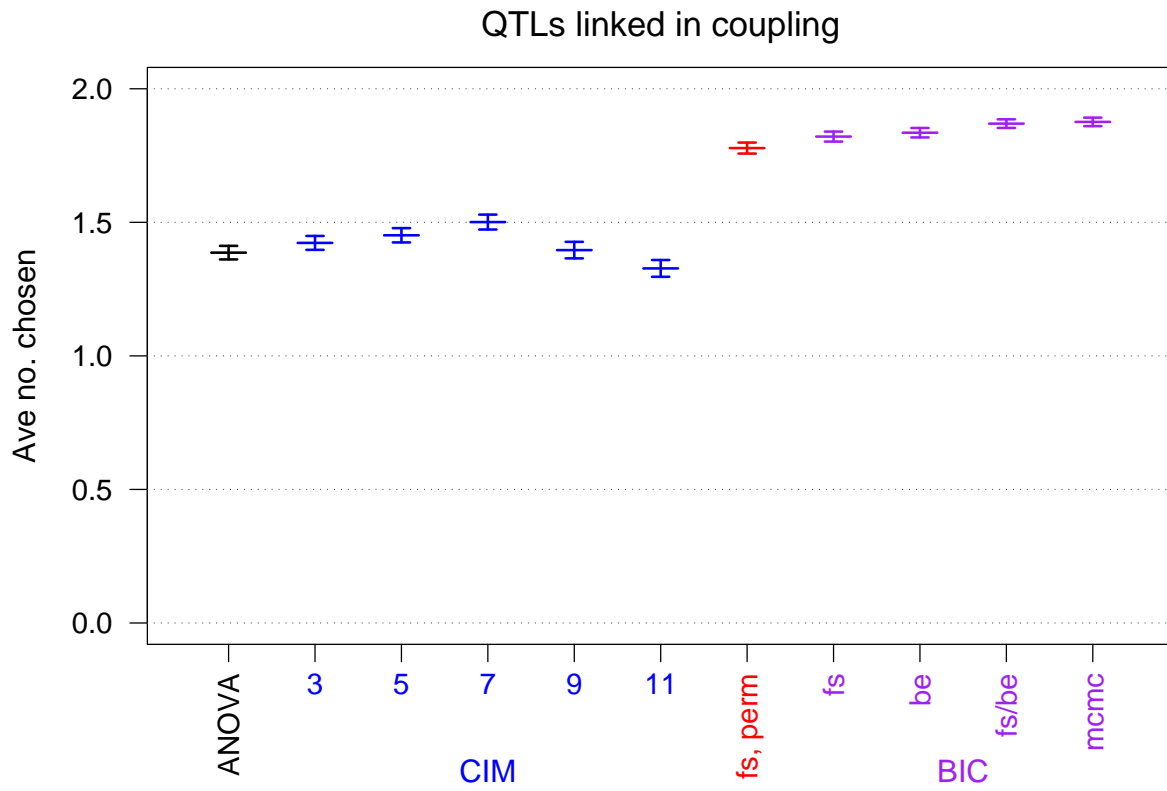
Compare to a genome-wide threshold.

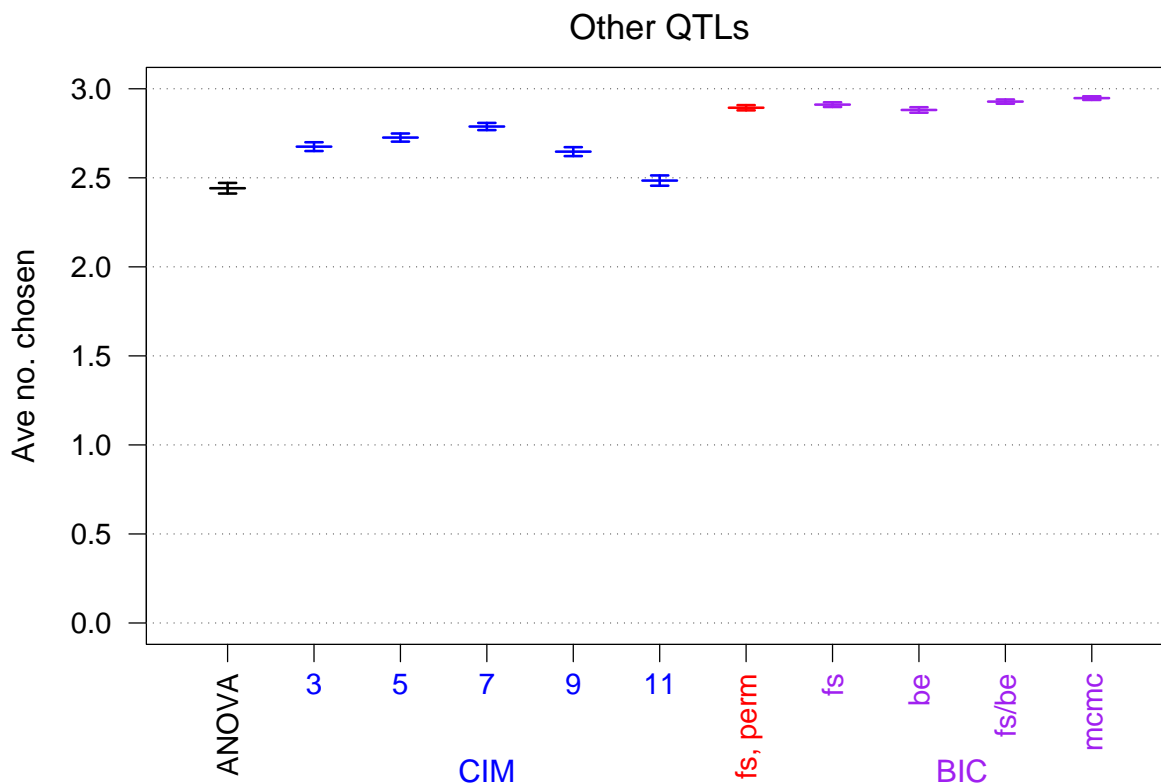
(Take into account the choice of  $S$ .)

# IM / CIM → model









## Summary

- QTL mapping is a model selection problem (rather than hypothesis testing).
- Model selection =
  - Select a class of models
  - Select a criterion for comparing models
  - Select a method of searching model space
  - Figure out how your procedure performs
- Key issue: the comparison of models.
- Large-scale computer simulations are necessary for assessing the performance of procedures



# References

---

- Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). *J Roy Stat Soc B* 64:641–656, 731–775  
Contains the simulation study described above.
- Zeng ZB, Kao CH, Basten CJ (1999) Estimating the genetic architecture of quantitative traits. *Genetical Research* 74: 279–289  
Another paper on the model selection aspects of QTL mapping.
- Miller AJ (2002) *Subset selection in regression*, 2nd edition. Chapman & Hall, New York  
A good book on model selection in regression.