Model selection for QTL mapping

Karl W Broman

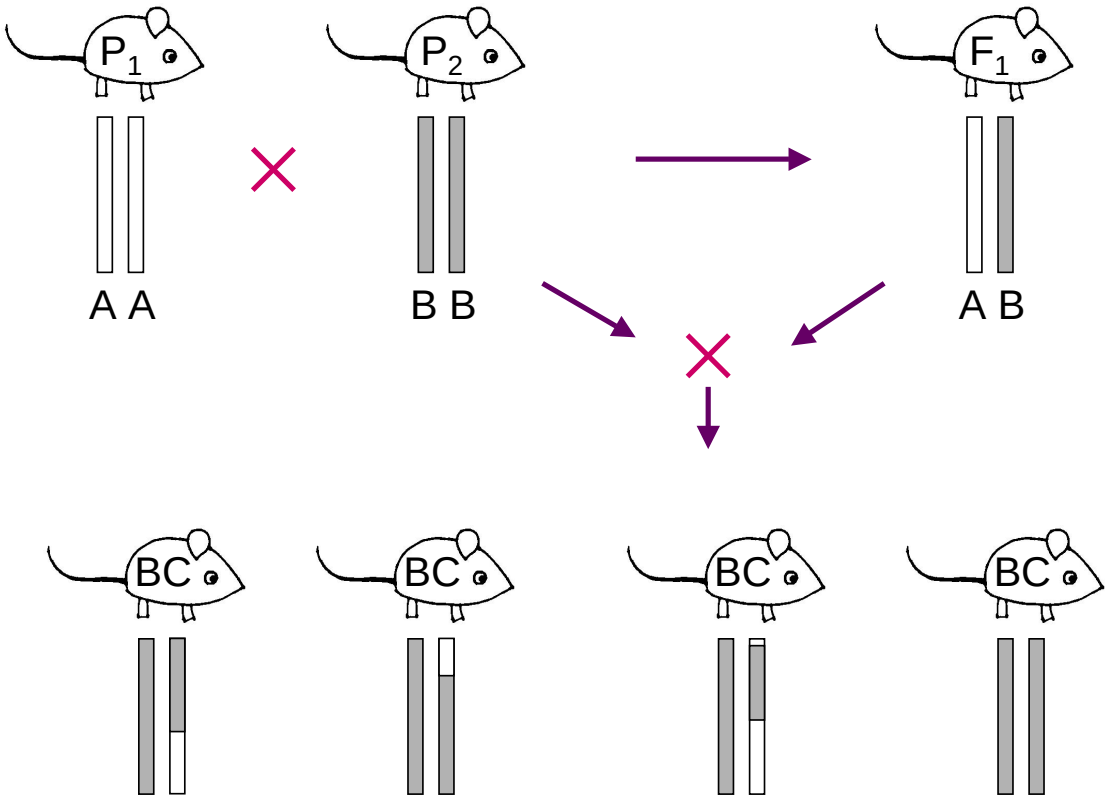
Department of Biostatistics, Johns Hopkins University

Terry Speed

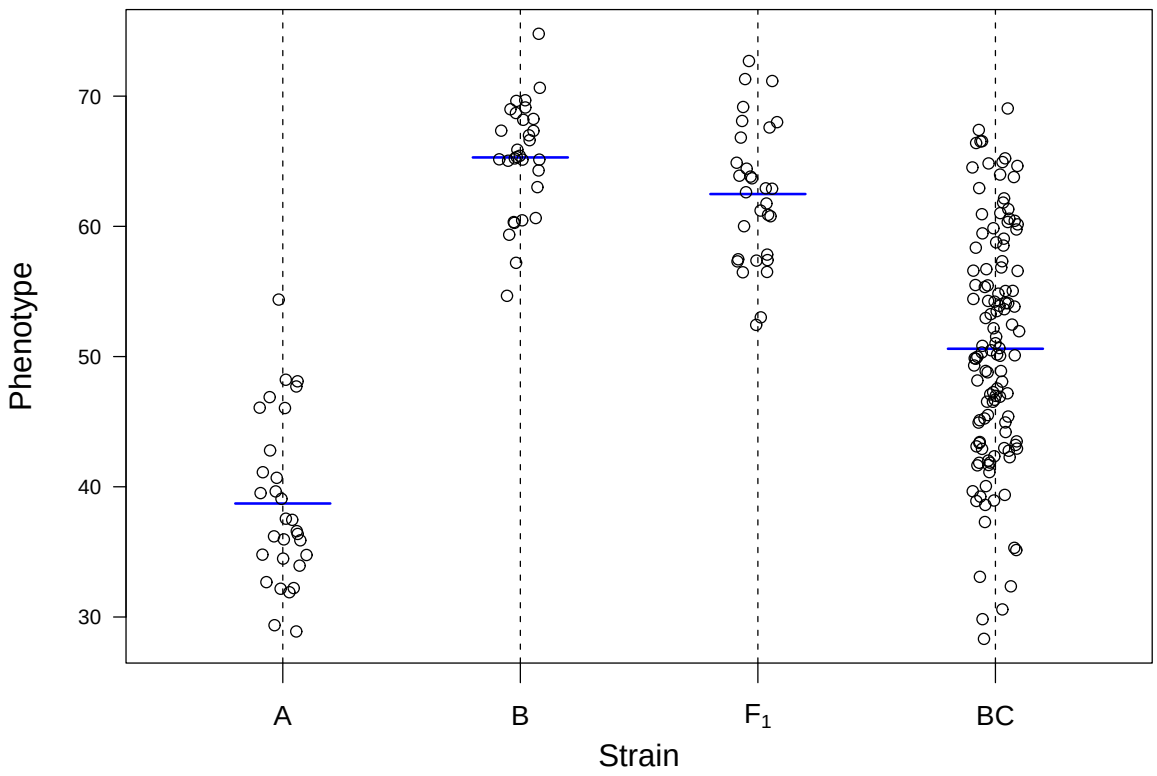
Department of Statistics, University of California, Berkeley

Walter and Eliza Hall Institute (Melbourne, Australia)

Backcross experiment



Trait distributions



Data and Goals

Phenotypes: y_i = trait value for mouse i

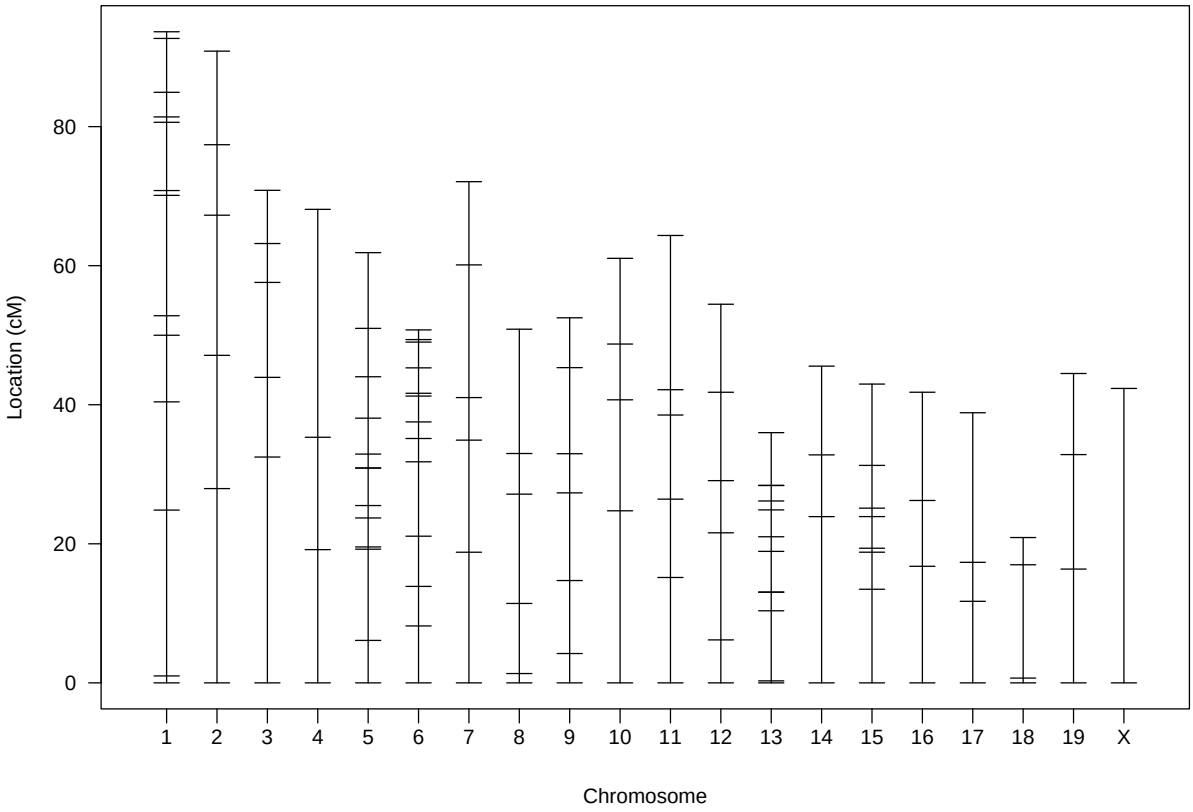
Genotypes: x_{ij} = 1/0 if mouse i is BB/AB at marker j (for a backcross)

Genetic map: Locations of markers

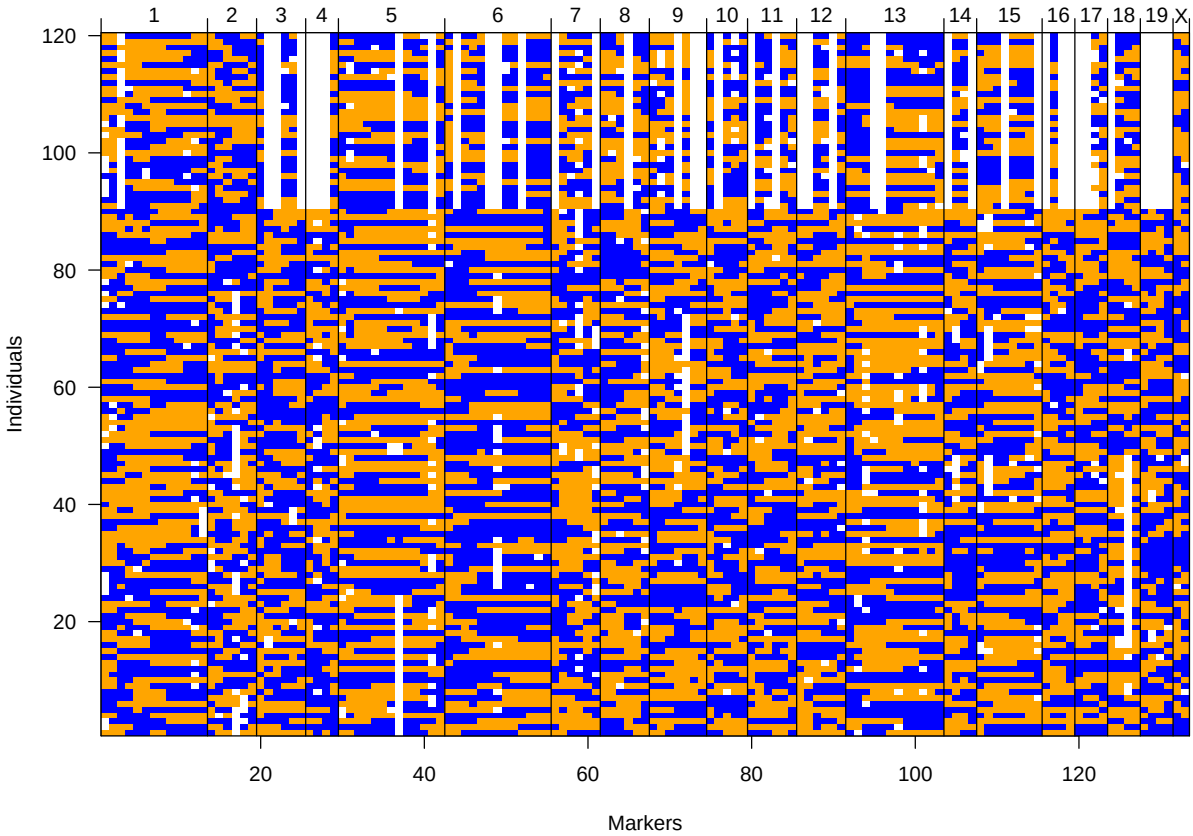
Goals:

- Identify the (or at least one) genomic regions (QTLs) that contribute to variation in the trait.
- Form confidence intervals for QTL locations.
- Estimate QTL effects.

Genetic map



Genotype data



Models: Recombination

We assume no crossover interference.

$\Longrightarrow$ Points of exchange (crossovers) are according to a Poisson process.

$\Longrightarrow$ The $\{x_{ij}\}$ (marker genotypes) form a Markov chain

Models: Genotype $\longleftrightarrow$ Phenotype

Let y = phenotype
g = whole genome genotype

Imagine a small number of QTLs with genotypes $g_1, \ldots, g_p$.
(2^p distinct genotypes)

$$\mathrm{E}(y|g) = \mu_{g_1,\ldots,g_p} \qquad \mathrm{var}(y|g) = \sigma^2_{g_1,\ldots,g_p}$$

Models: Genotype $\longleftrightarrow$ Phenotype

Homoscedasticity (constant variance): $\sigma_g^2 \equiv \sigma^2$

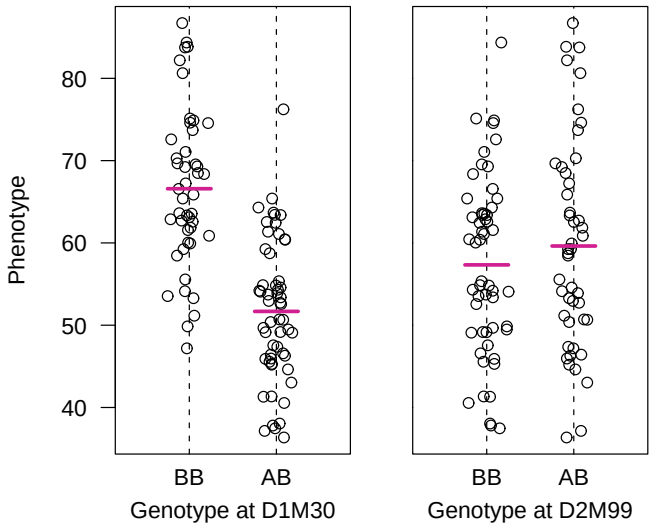
Normally distributed residual variation: $y|g \sim N(\mu_g, \sigma^2)$.

Additivity: $\mu_{g_1,\ldots,g_p} = \mu + \sum_{j=1}^{p} \Delta_j \, g_j$ ($g_j = 1$ or 0)

Epistasis: Any deviations from additivity.

The simplest method: ANOVA

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

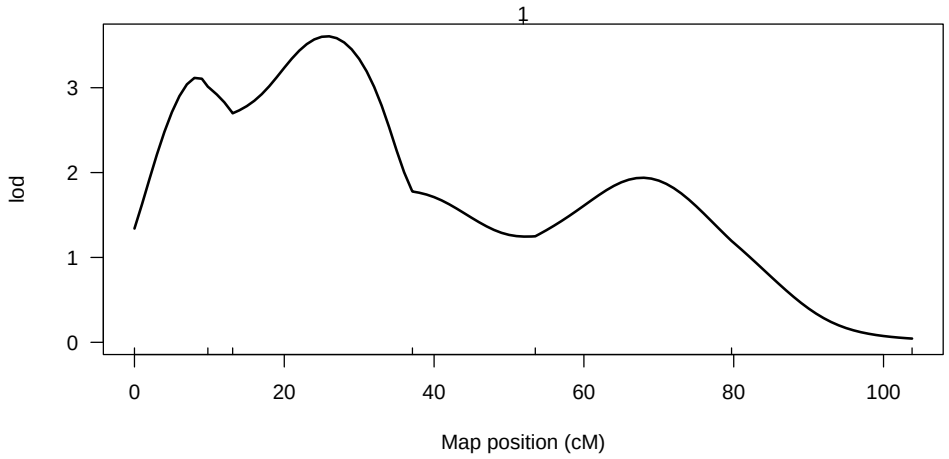


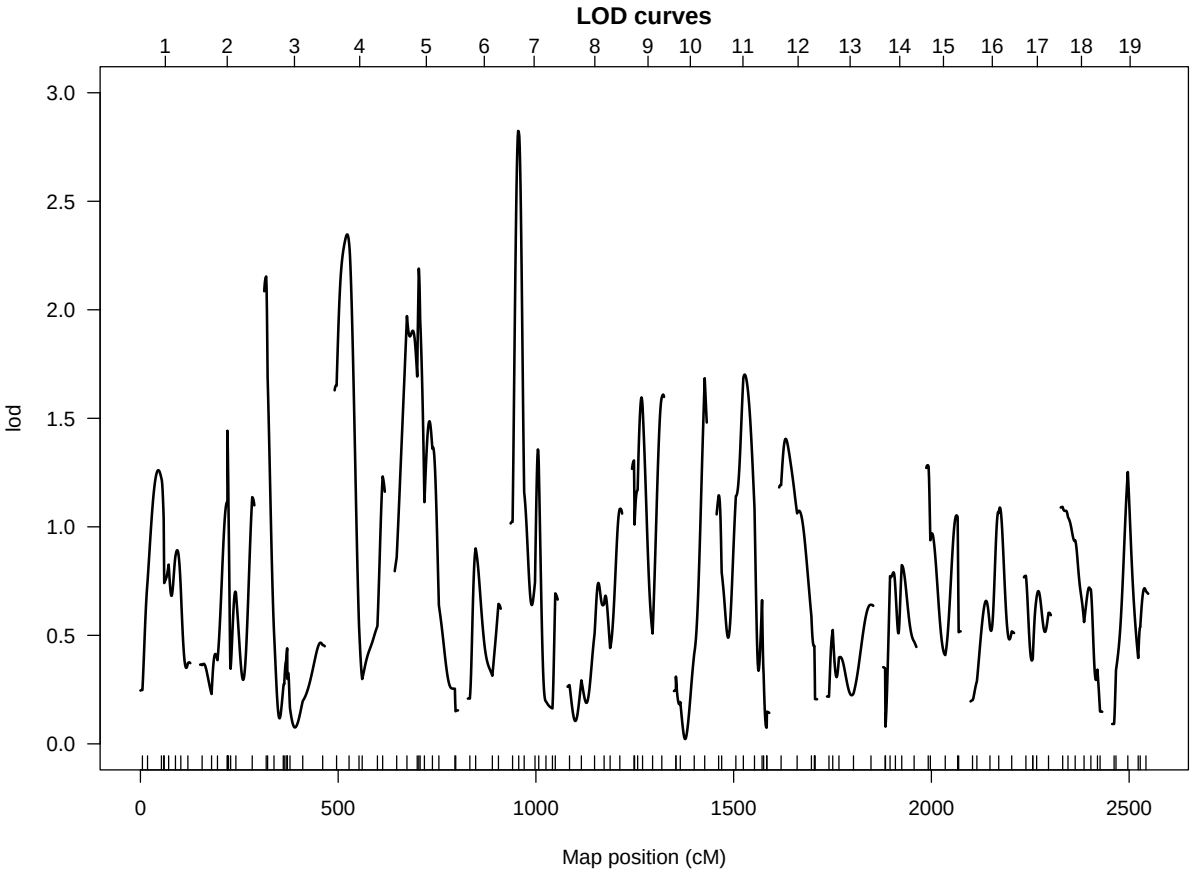
LOD score = $\log_{10}$ likelihood ratio comparing single-QTL model to "no QTL anywhere."

Interval mapping (IM)

Lander & Botstein (1989)

- Take account of missing genotype data
- Interpolate between markers
- Maximum likelihood under a mixture model





LOD thresholds

LOD threshold = 95 %ile of distr'n of max LOD, genome-wide, if there are no QTLs anywhere

Derivation:
- Analytical calculations
- Simulations
- Permutation tests

Multiple QTL methods

Why consider multiple QTLs at once?

- Reduce residual variation.
- Separate linked QTLs.
- Investigate interactions between QTLs (epistasis).

Abstractions / simplifications

- Complete marker data
- QTLs are at the marker loci
- QTLs act additively

The problem

n backcross mice; M markers

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

y_i = phenotype (trait value) of mouse i

$$y_i = \mu + \sum_{j=1}^{M} \Delta_j \, x_{ij} + \epsilon_i$$

Which $\Delta_j \neq 0$?

$\longrightarrow$ **Model selection in regression**

How is this problem different?

- Relationship among the x's
- Find a good model vs. minimize prediction error

Model selection

- Select class of models
  - Additive models
  - Add've plus pairwise interactions
  - Regression trees
- Search model space
  - Forward selection (FS)
  - Backward elimination (BE)
  - FS followed by BE
  - MCMC
- Compare models
  - $\text{BIC}_\delta(\gamma) = \log \text{RSS}(\gamma) + |\gamma| \, (\delta \, \frac{\log n}{n})$
  - Sequential permutation tests
- Assess performance
  - Maximize no. QTLs found; control false positive rate

Why BIC$_\delta$?

- For a fixed no. markers, letting $n \to \infty$, 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.

Choice of δ

Larger δ: include more loci; higher false positive rate

Smaller δ: include fewer loci; lower false positive rate

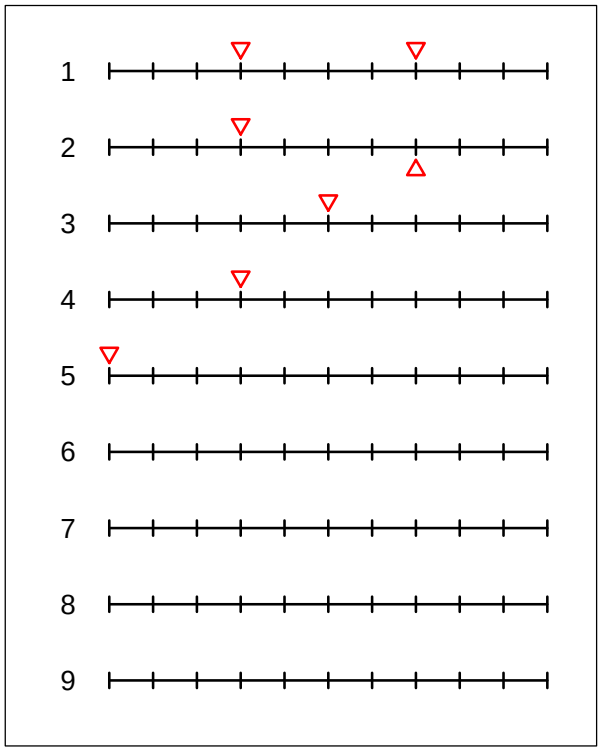
Let *L* = 95% genome-wide LOD threshold
(compare single-QTL models to the null model)

Choose δ = 2 *L* / $\log_{10} n$

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

Simulations

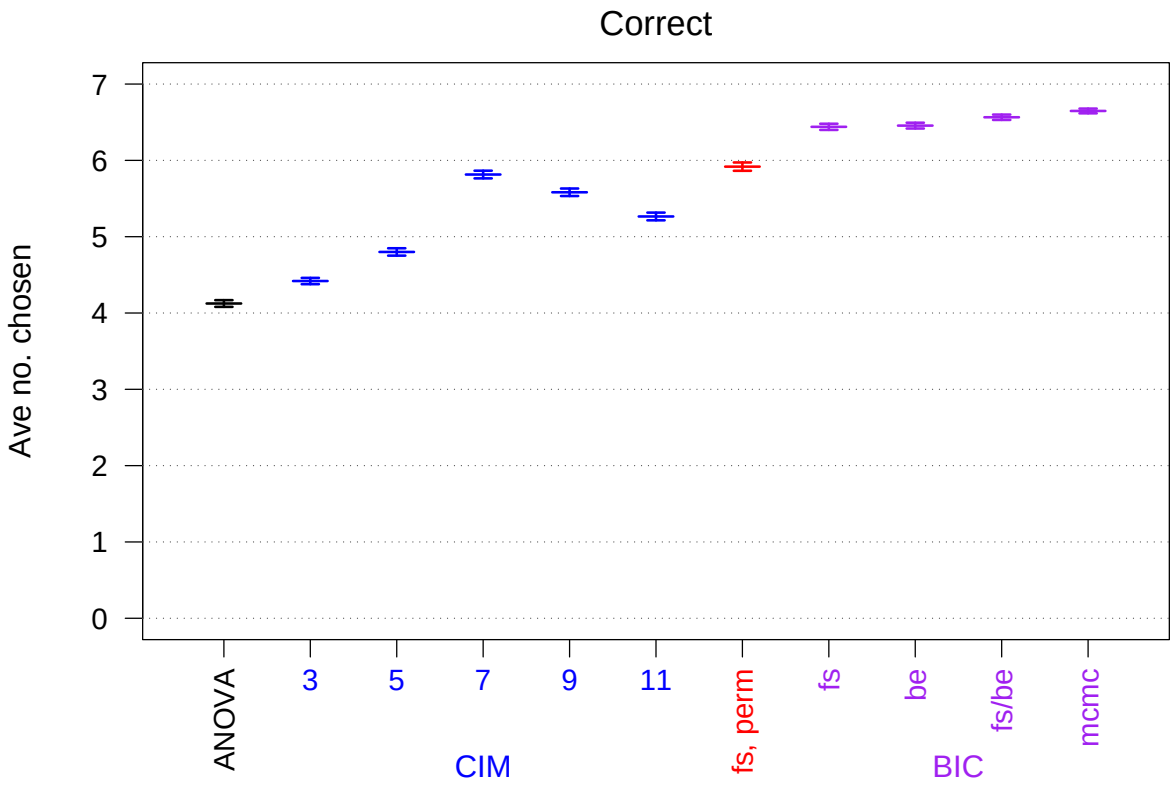
- Backcross with n=250
- No crossover interference
- 9 chr, each 100 cM
- Markers at 10 cM spacing; complete genotype data
- 7 QTLs
  - One pair in coupling
  - One pair in repulsion
  - Three unlinked QTLs
- Heritability = 50%
- 2000 simulation replicates

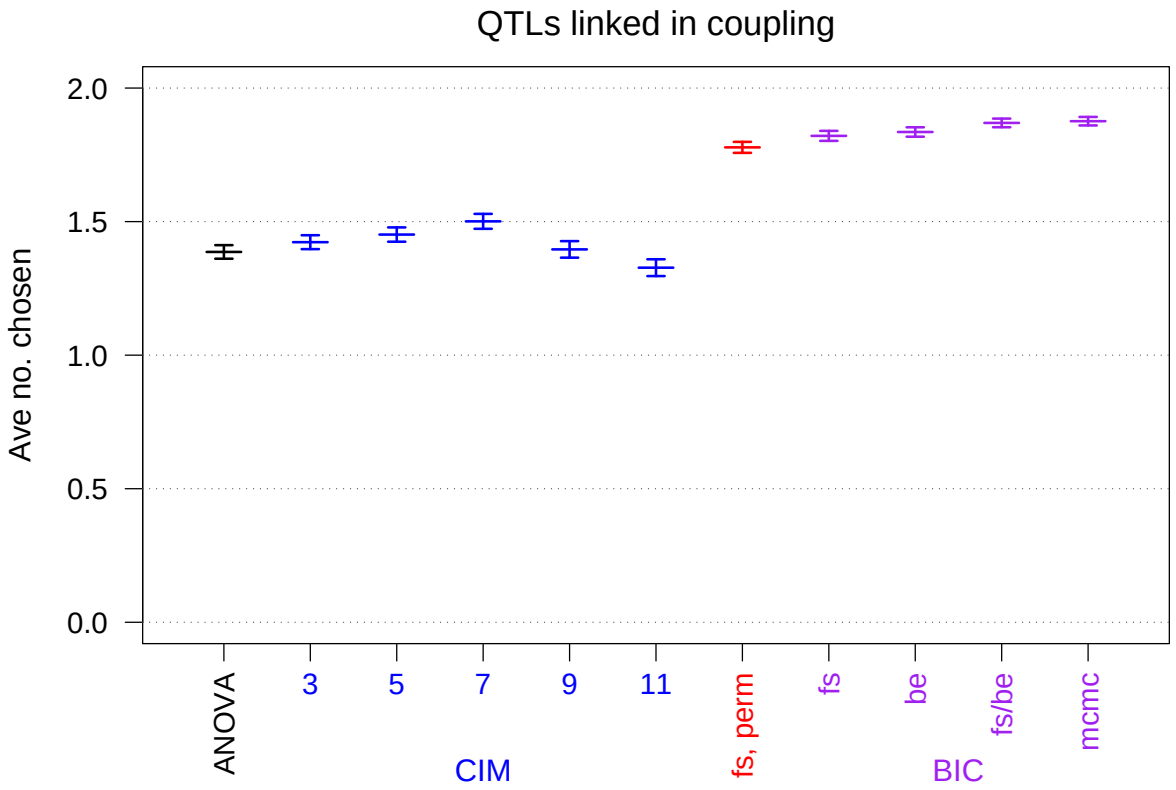


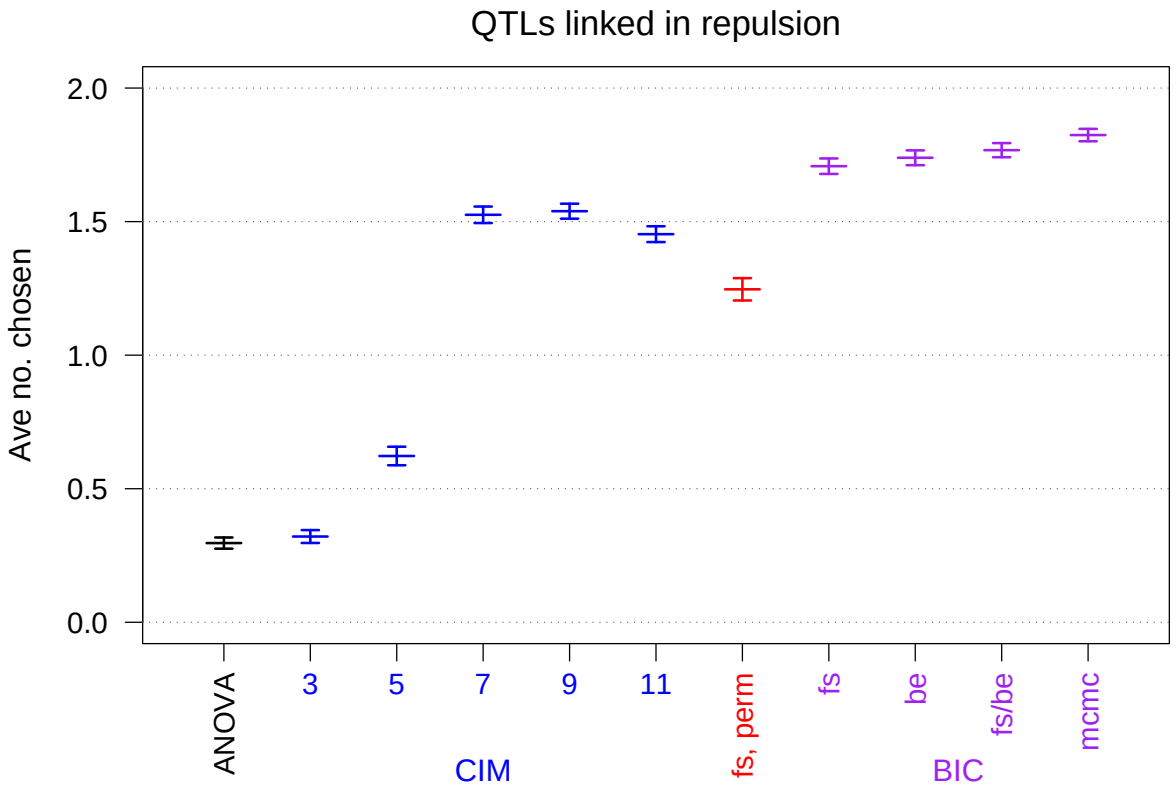
Methods

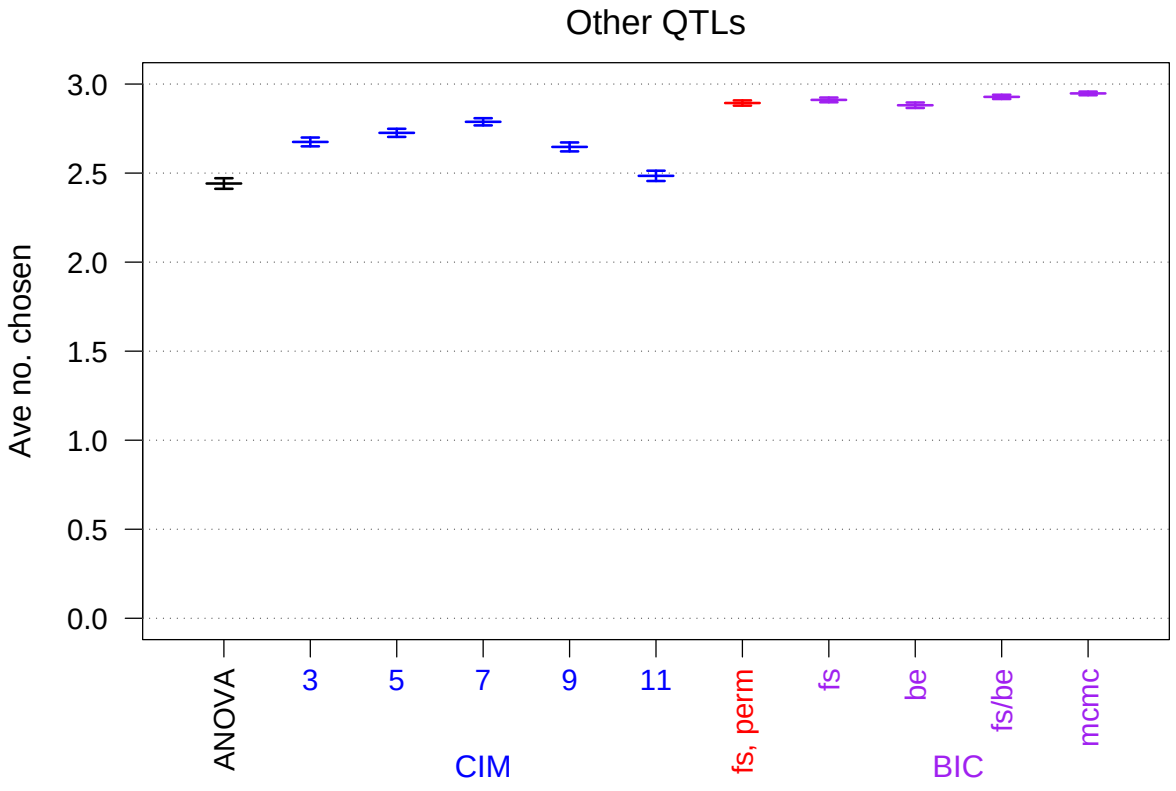
- ANOVA at marker loci
- Composite interval mapping (CIM)
- Forward selection with permutation tests
- Forward selection with BIC_δ
- Backward elimination with BIC_δ
- FS followed by BE with BIC_δ
- MCMC with BIC_δ

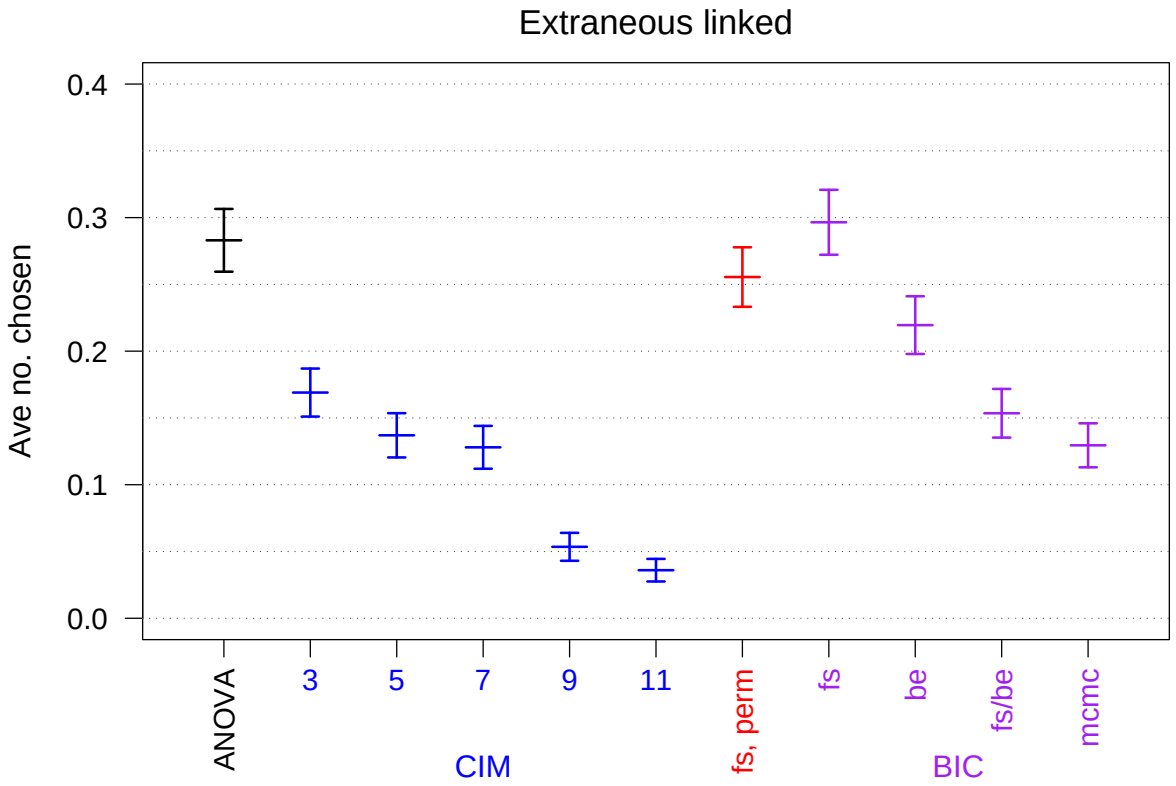
$\longrightarrow$ A selected marker is deemed correct if it is within 10 cM of a QTL (i.e., correct or adjacent)

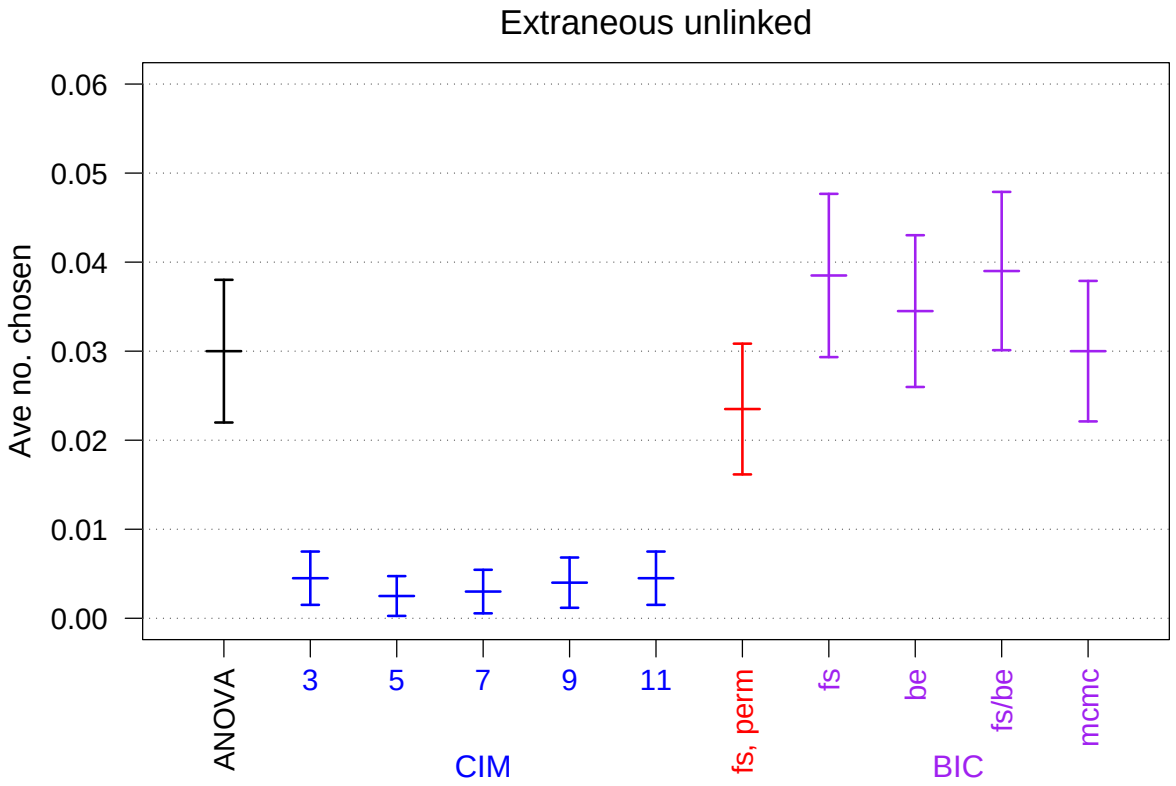

Correct
Ave no. chosen
0
1
2
3
4
5
6
7
ANOVA
3
5
7
9
11
fs, perm
fs
be
fs/be
mcmc
CIM
BIC


QTLs linked in coupling
Ave no. chosen
0.0
0.5
1.0
1.5
2.0
ANOVA
3
5
7
9
11
fs, perm
fs
be
fs/be
mcmc
CIM
BIC


QTLs linked in repulsion
Ave no. chosen
2.0
1.5
1.0
0.5
0.0
ANOVA
3
5
7
9
11
fs, perm
fs
be
fs/be
mcmc
CIM
BIC


Other QTLs
Ave no. chosen
3.0
2.5
2.0
1.5
1.0
0.5
0.0
ANOVA
3
5
7
9
11
fs, perm
fs
be
fs/be
mcmc
CIM
BIC


Extraneous linked
Ave no. chosen
0.0
0.1
0.2
0.3
0.4
ANOVA
3
5
7
9
11
fs, perm
fs
be
fs/be
mcmc
CIM
BIC


Extraneous unlinked
Ave no. chosen
0.00
0.01
0.02
0.03
0.04
0.05
0.06
ANOVA
3
5
7
9
11
fs, perm
fs
be
fs/be
mcmc
CIM
BIC

Summary

- QTL mapping is a model selection problem.
- Key issue: the comparison of models.
- Large-scale simulations are important.
- More refined procedures do not necessarily give improved results.

- BIC_δ with forward selection followed by backward elimination works quite well.