A model selection approach for the identification of quantitative trait loci in experimental crosses

Karl W Broman

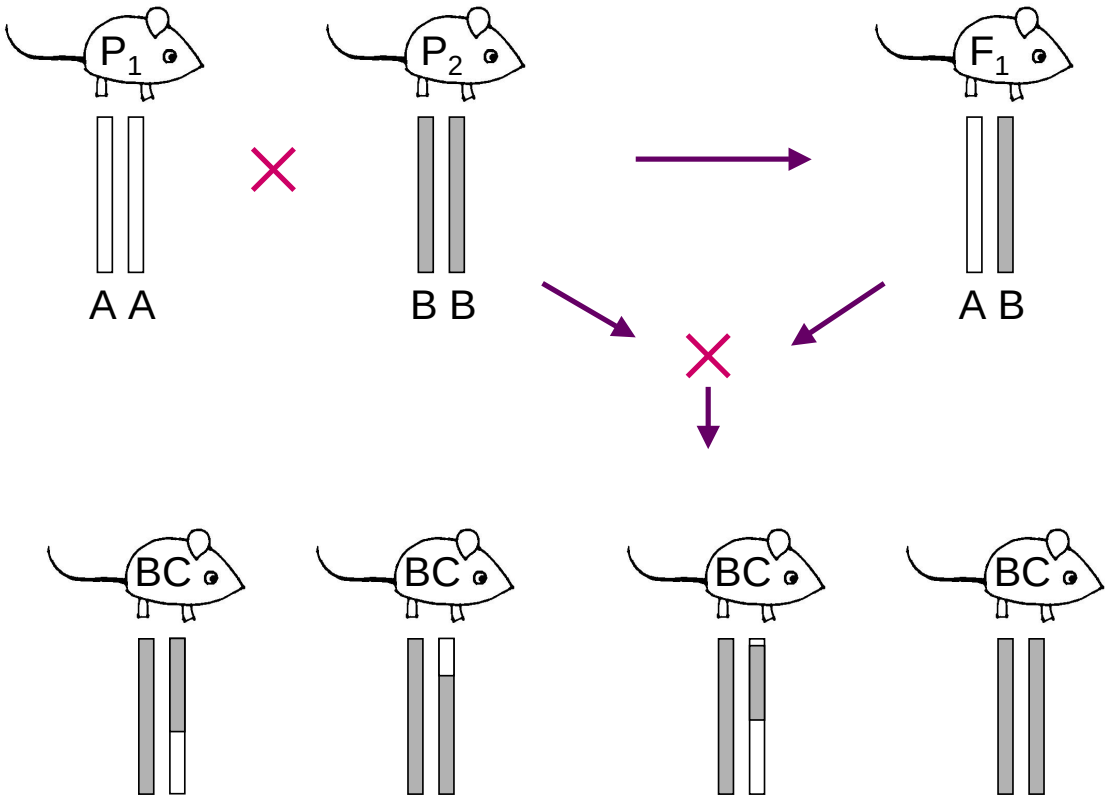
Department of Biostatistics, Johns Hopkins University

Terry Speed

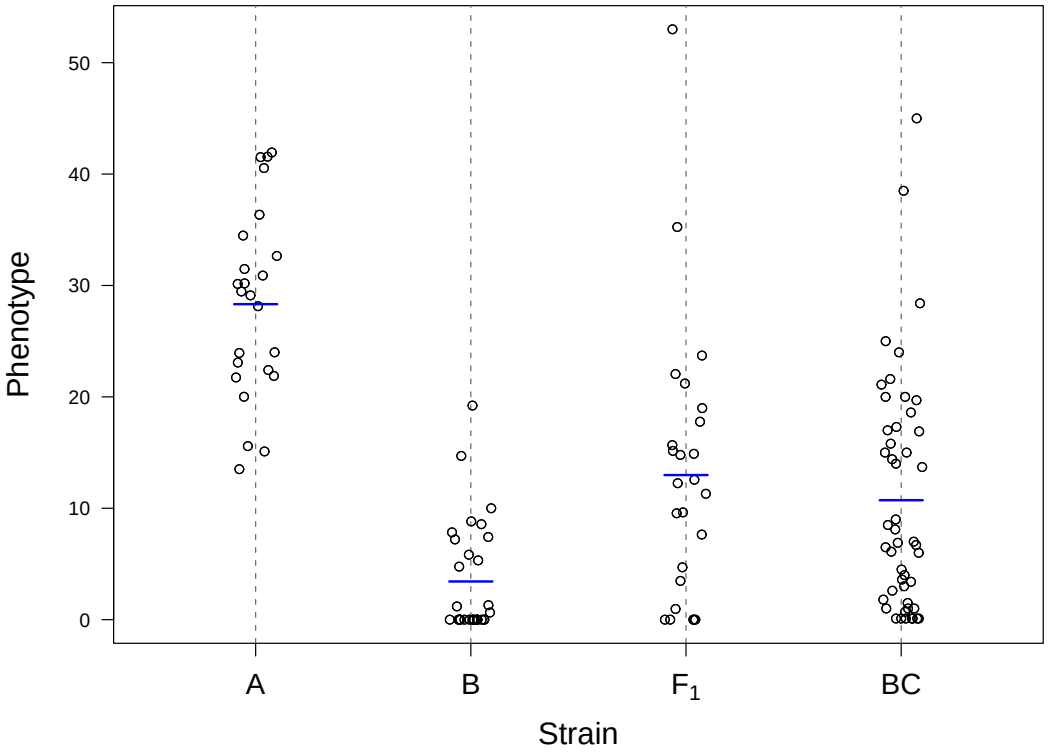
Department of Statistics, University of California, Berkeley

Walter and Eliza Hall Institute (Melbourne, Australia)

Backcross experiment



Trait distributions



Data and Goals

Phenotypes: y_i = trait value for mouse i

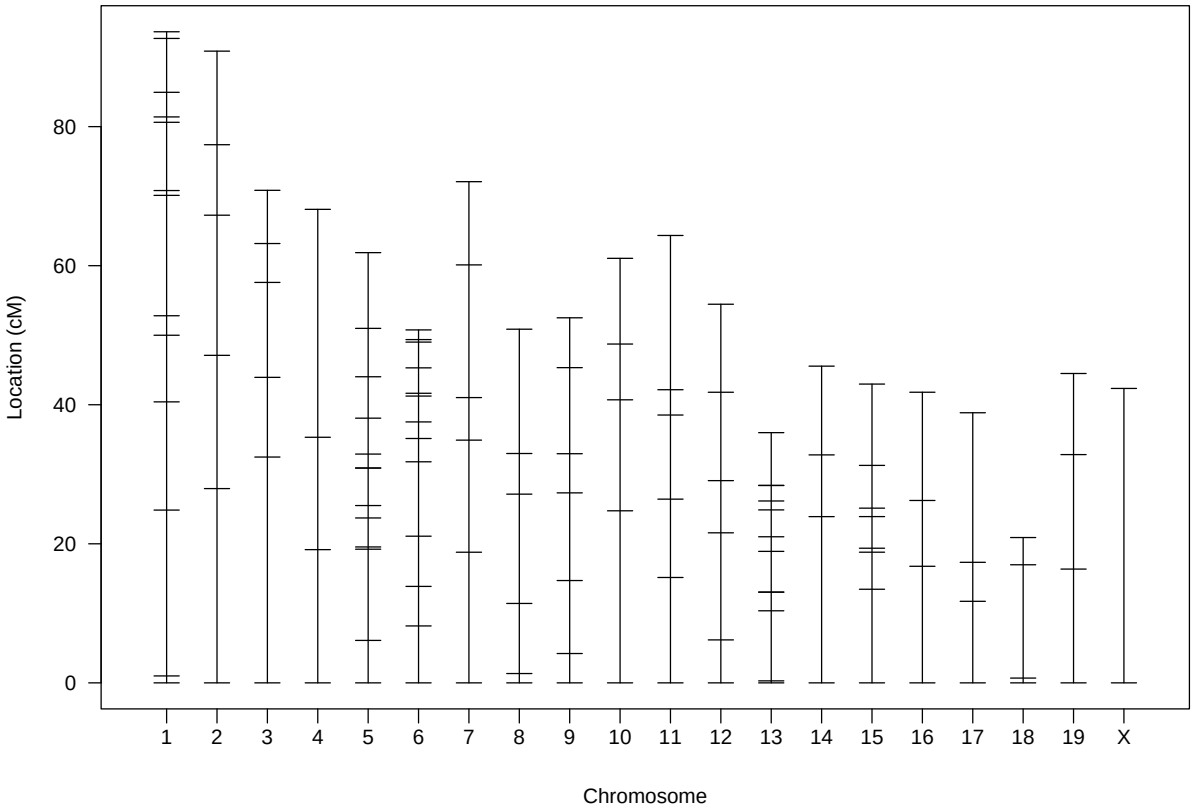
Genotypes: x_{ij} = 1/0 if mouse i is BB/AB at marker j (for a backcross)

Genetic map: Locations of markers

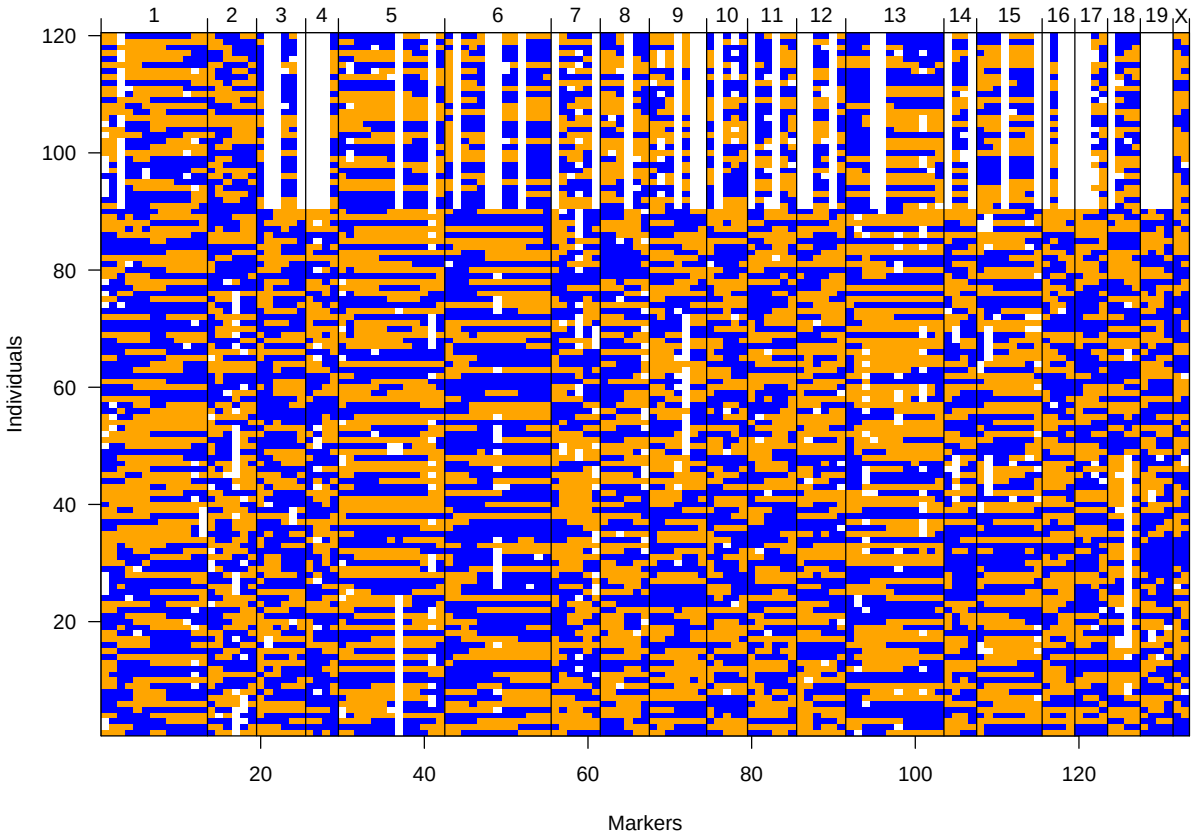
Goals:

- 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)

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

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

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.

Abstractions / simplifications

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

$\longrightarrow$ This work is not useful in practice
but serves to illustrate the key issues.

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

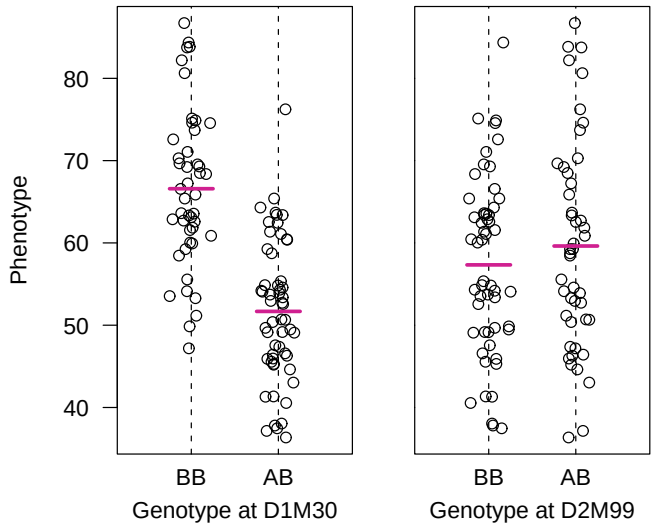
$$y_i = \mu + \sum_{j=1}^{M} \Delta_j \, x_{ij} + \epsilon_i$$

Which $\Delta_j \neq 0$?

Errors:
- Miss important loci
- Include extraneous loci

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

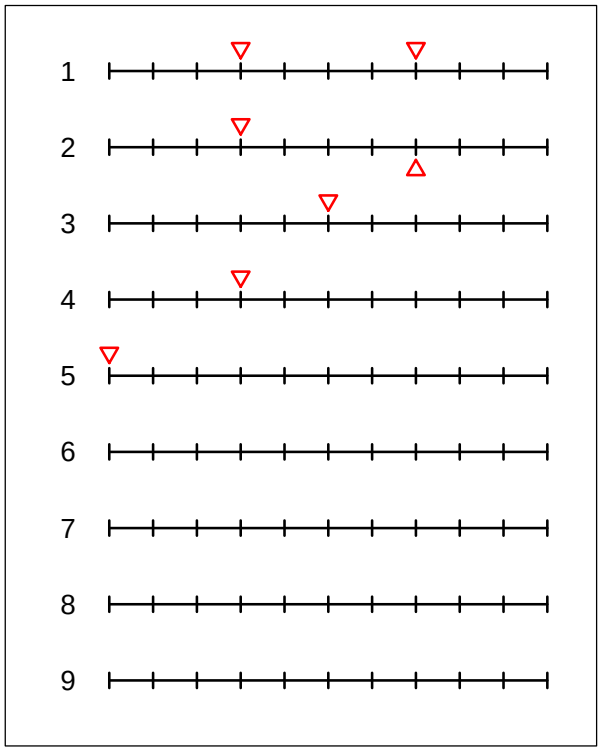


Model selection

- Select a class of models
  - Additive models
  - Up to pairwise interactions
  - Regression trees
- Compare models
  - Estimated prediction error
  - $\text{BIC}_\delta(\gamma) = \log \text{RSS}_\gamma + \delta|\gamma| \log n/n$
  - Permutation tests
- Search model space
  - Forward selection
  - Backward elimination
  - Stepwise selection
  - MCMC
- Assess the perfomance of a procedure

Simulations

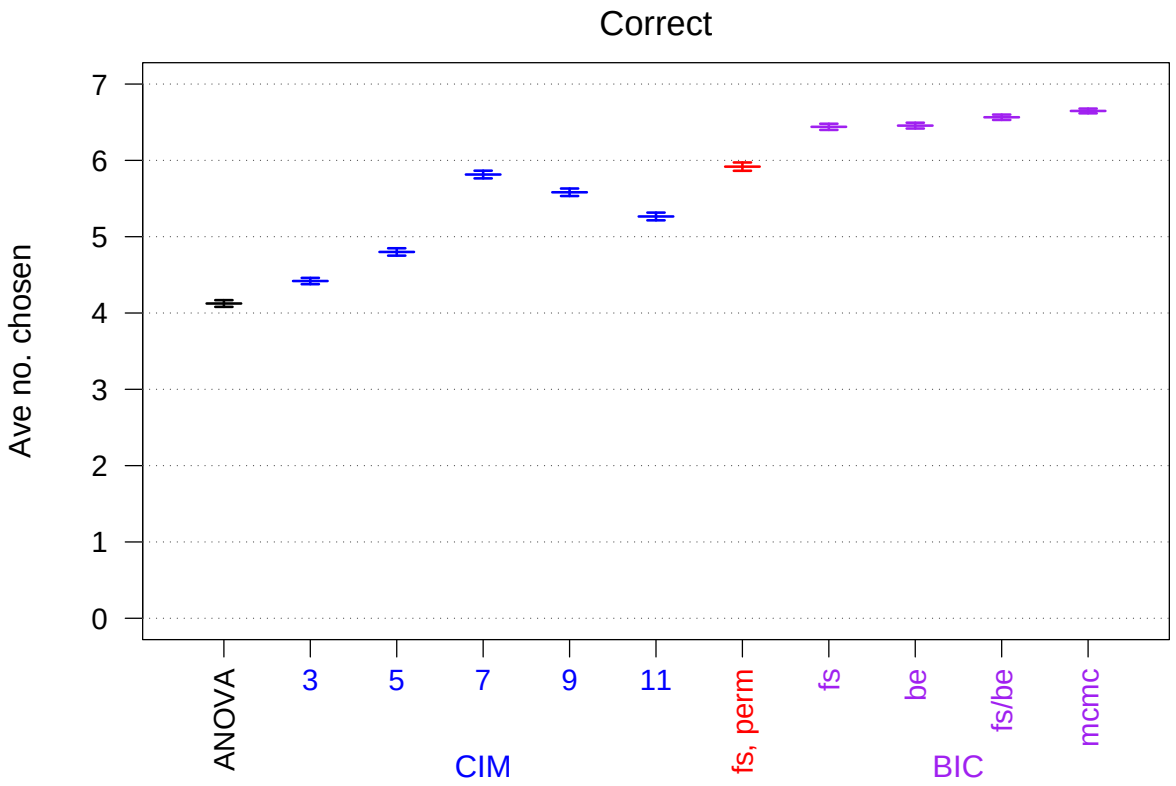
- Backcross with n=250
- No crossover interference
- 9 chr, each 100 cM
- Markers at 10 cM spacing; complete genotype data
- 7 QTLs
  - One pair in coupling
  - One pair in repulsion
  - Three unlinked QTLs
- Heritability = 50%
- 2000 simulation replicates

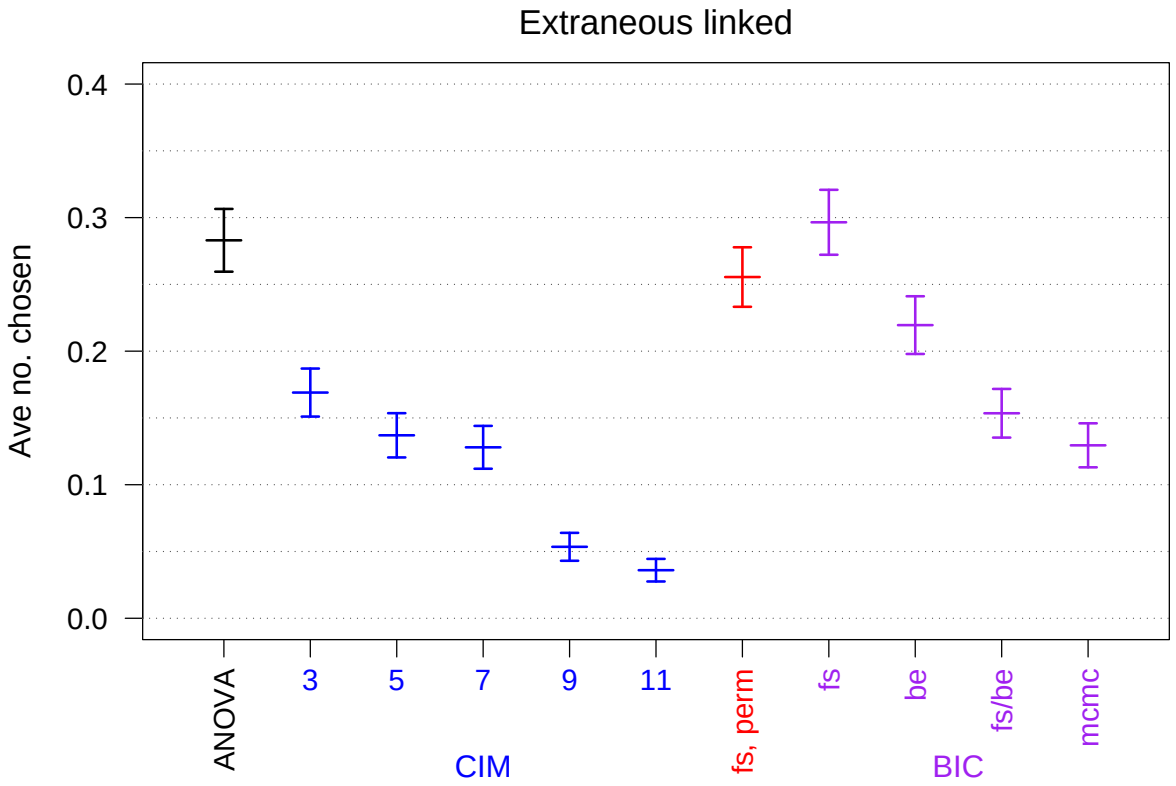


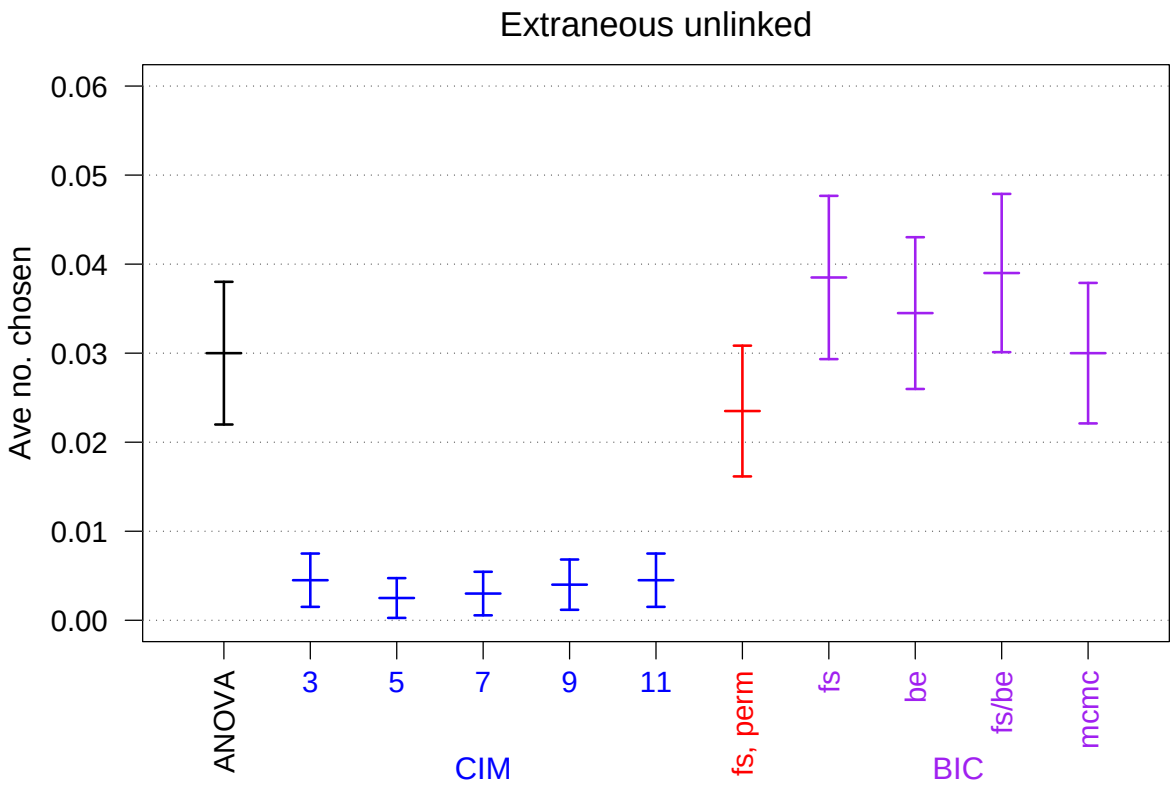
Methods

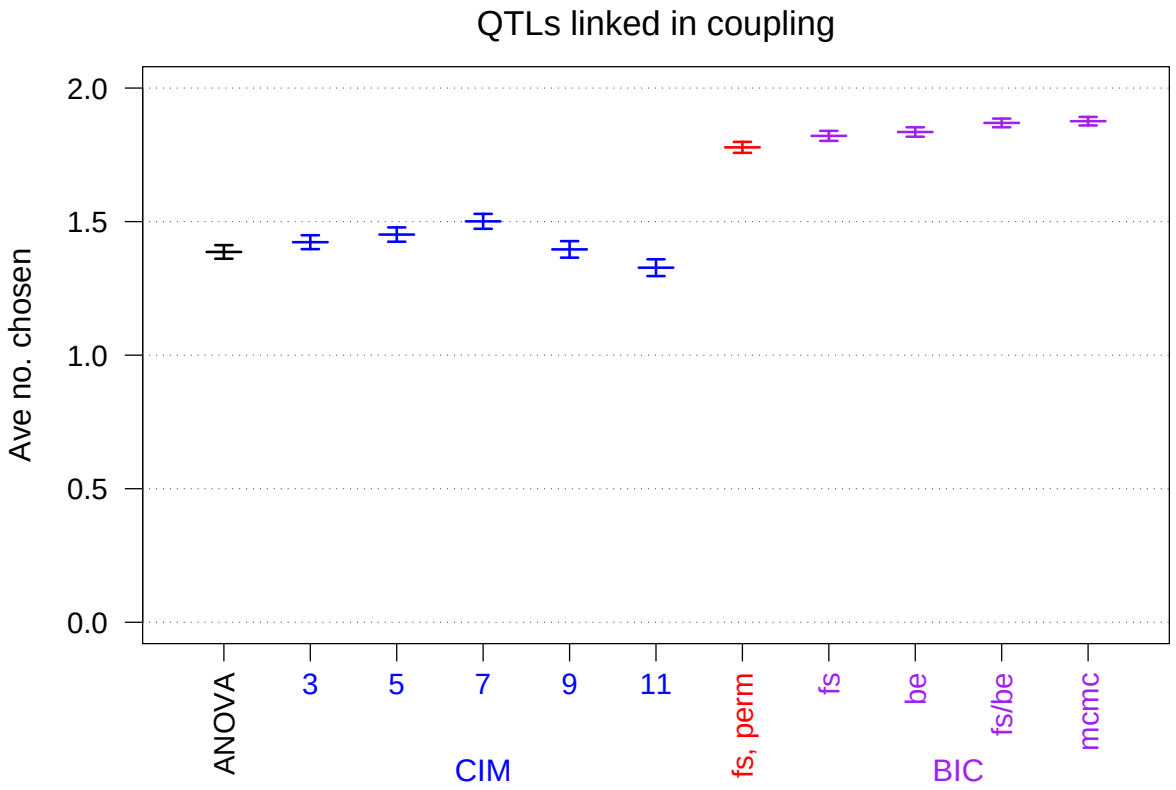
- ANOVA at marker loci
- Composite interval mapping (CIM)
- Forward selection with permutation tests
- Forward selection with BIC_δ
- Backward elimination with BIC_δ
- FS followed by BE with BIC_δ
- MCMC with BIC_δ

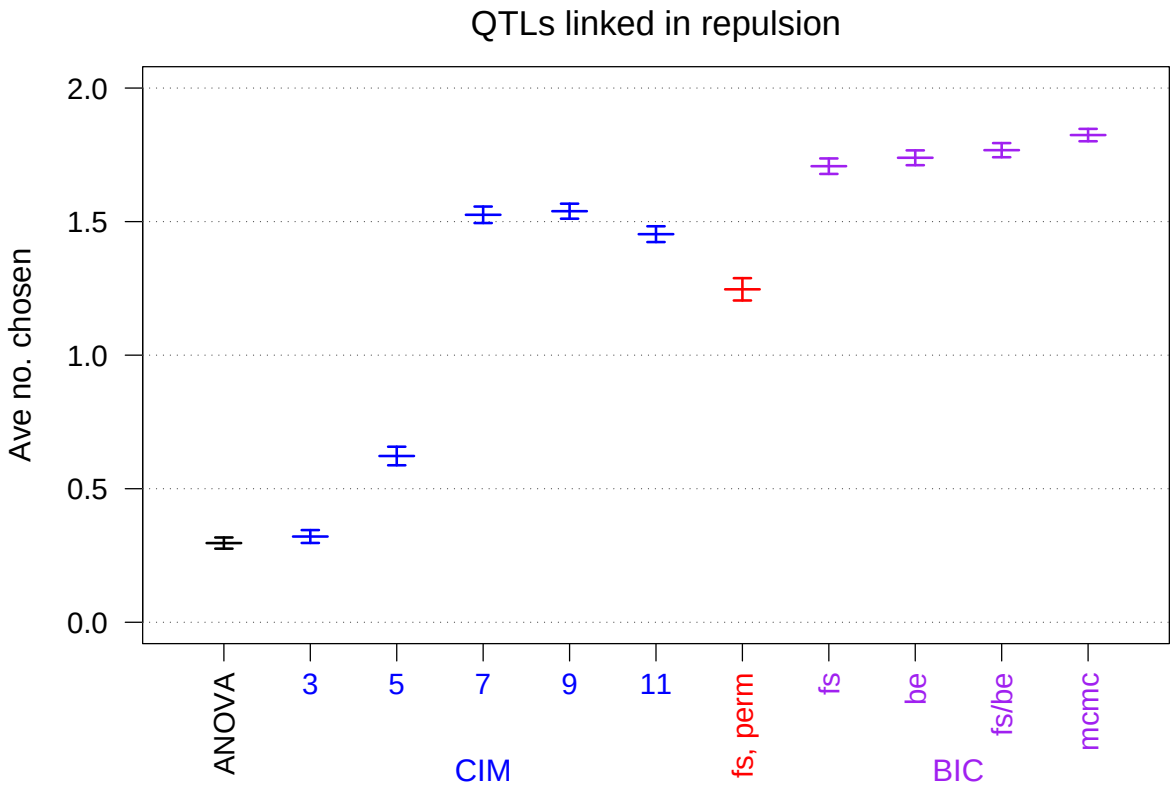
$\longrightarrow$ A selected marker is deemed correct if it is within 10 cM of a QTL (i.e., correct or adjacent)

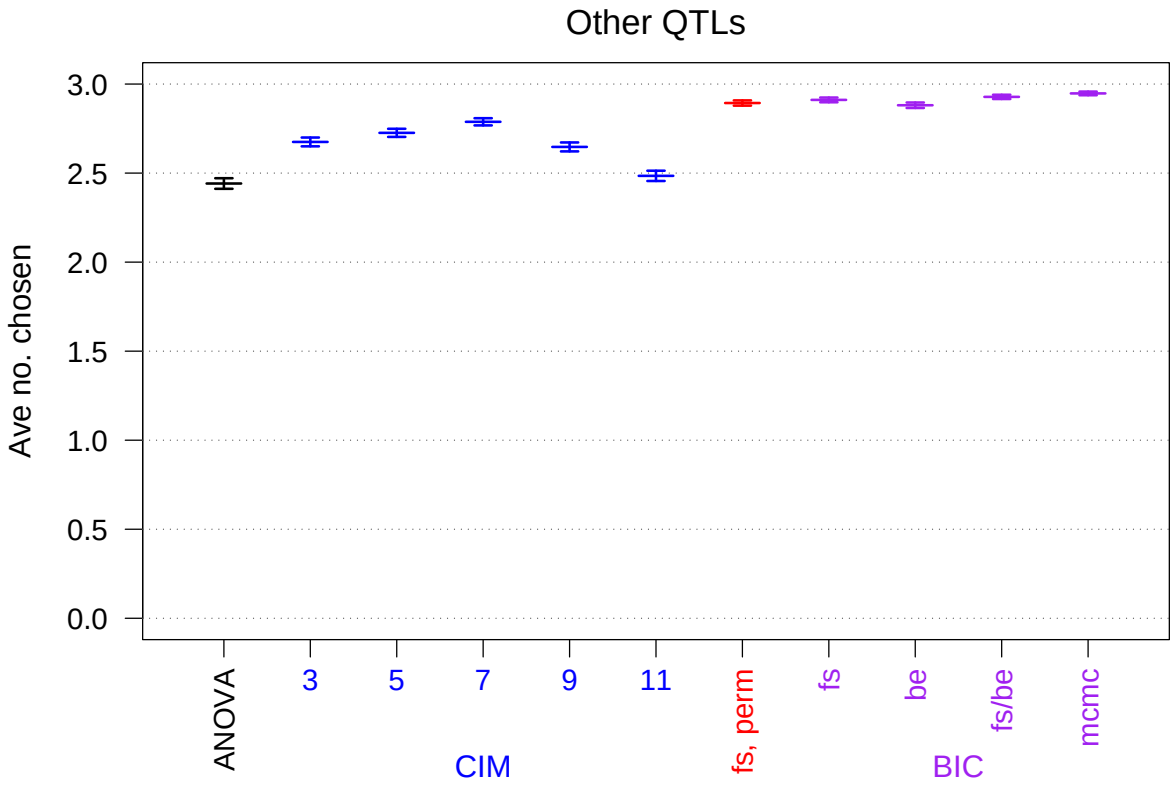

Correct
Ave no. chosen
0
1
2
3
4
5
6
7
ANOVA
3
5
7
9
11
fs, perm
fs
be
fs/be
mcmc
CIM
BIC


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
0.06
0.05
0.04
0.03
0.02
0.01
0.00
ANOVA
3
5
7
9
11
fs, perm
fs
be
fs/be
mcmc
CIM
BIC


QTLs linked in coupling
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


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

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.