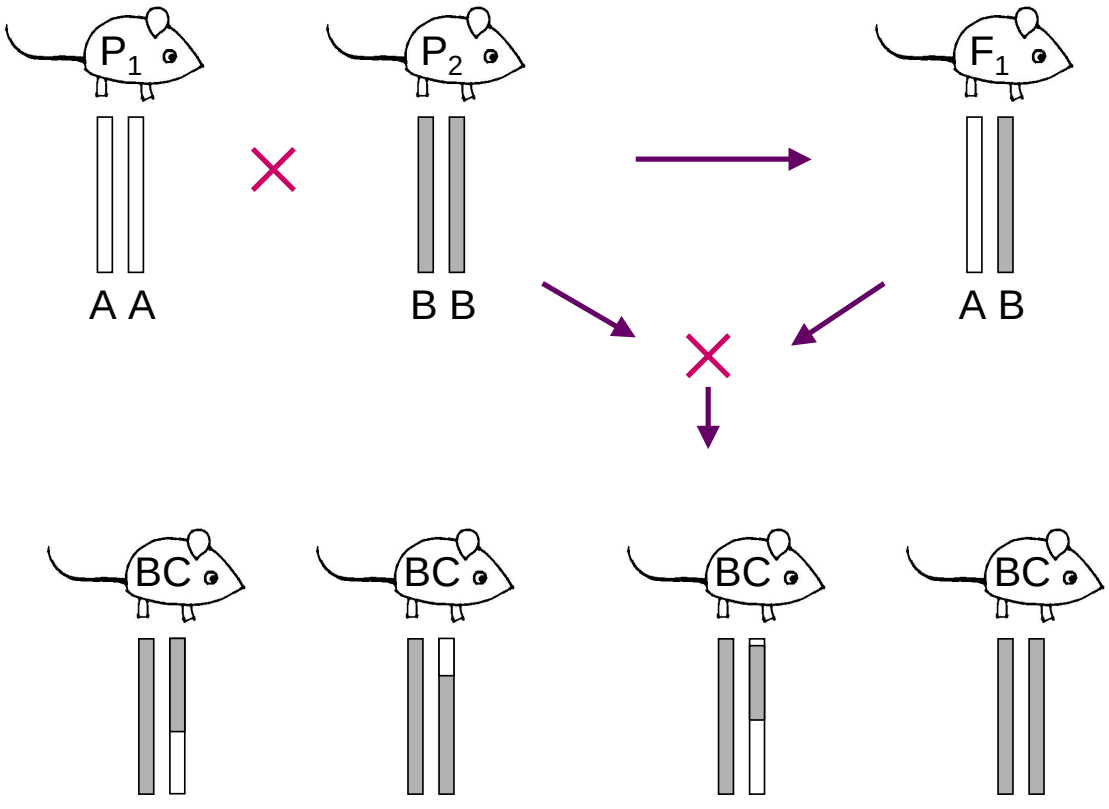
Model selection for QTL mapping

Karl W Broman

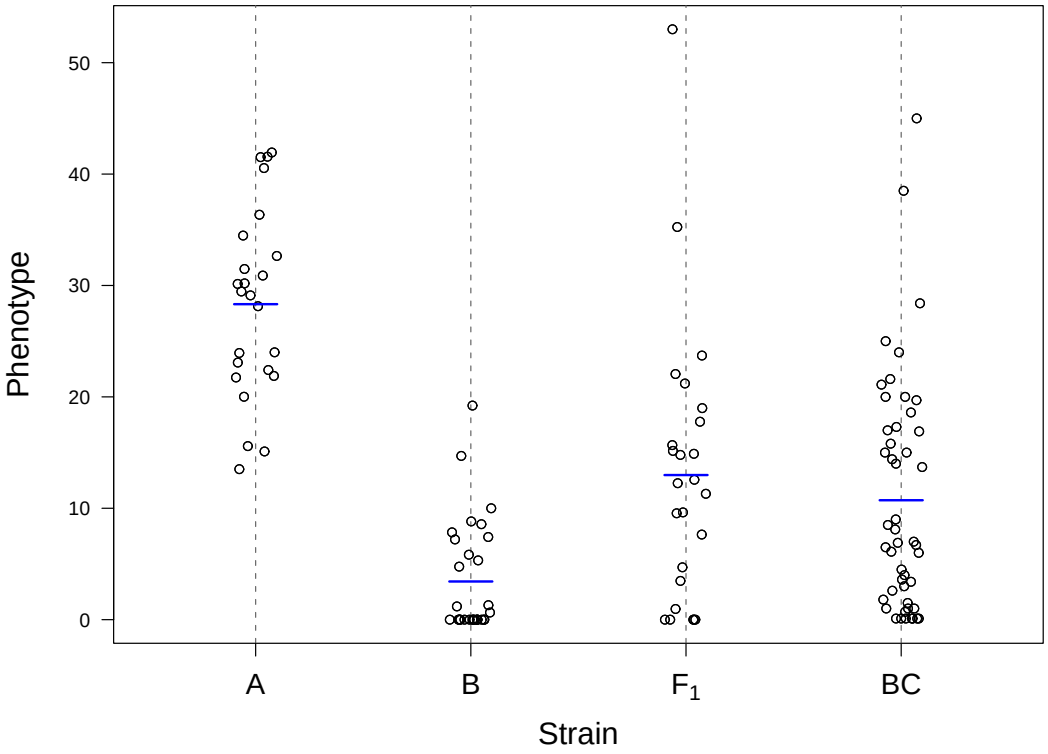
Department of Biostatistics
Johns Hopkins University

www.google.com

Backcross experiment



Trait distributions



Data and Goals

Phenotypes: y_i = trait value for mouse i

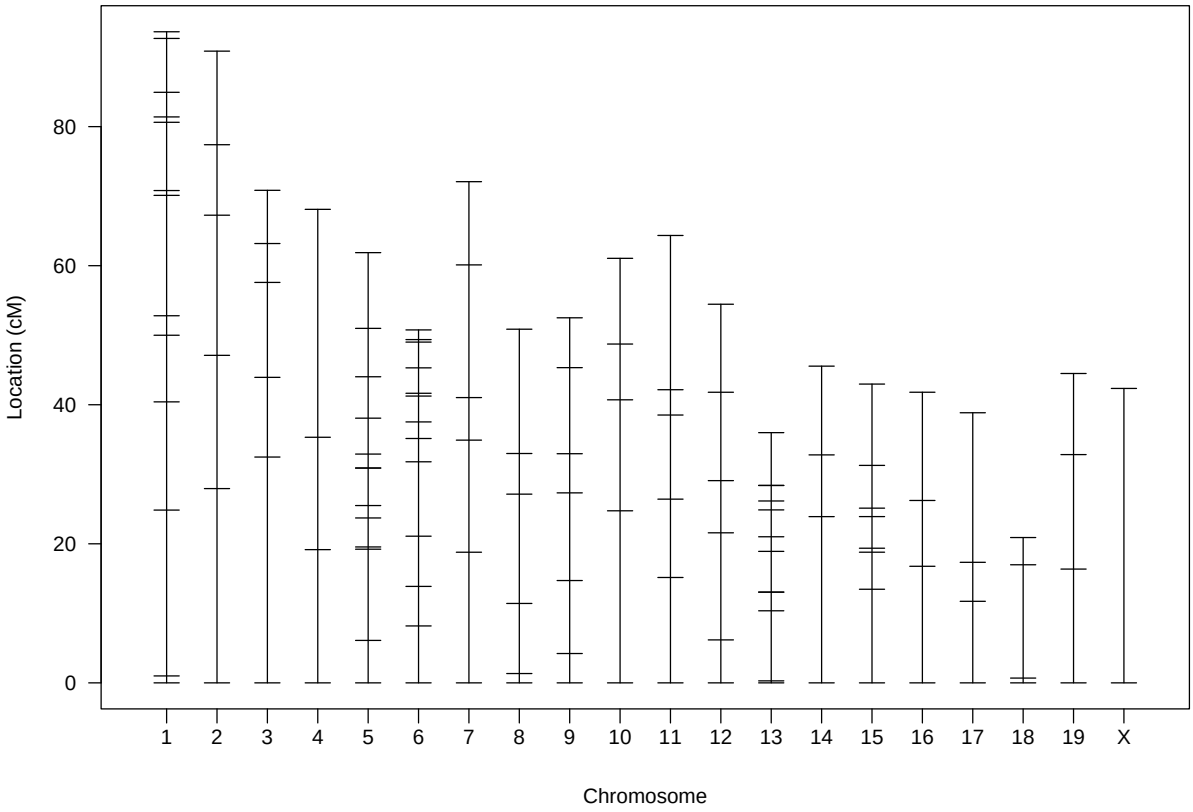
Genotypes: x_{ij} = 1/0 if mouse i is BB/AB at marker j (for a backcross)

Genetic map: Locations of markers

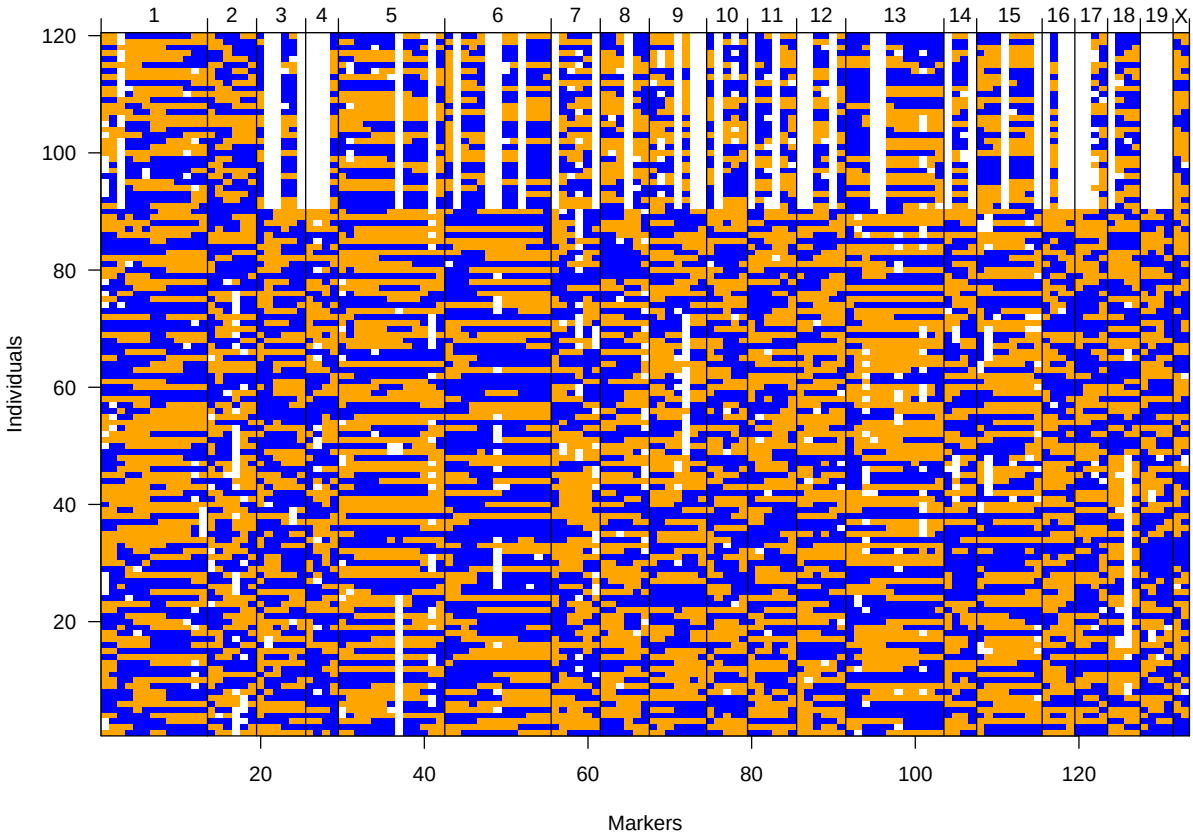
Goals:

- 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



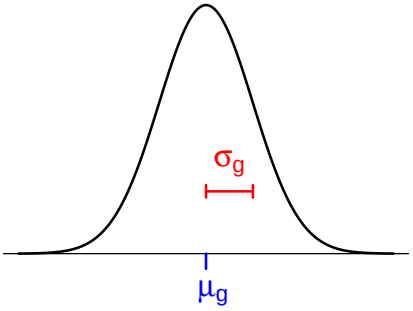
Models: Genotype $\longleftrightarrow$ Phenotype

Let y = phenotype
g = whole genome genotype

Consider all possible mice with a particular genome-type, g.

mean phenotype = μ_g

SD phenotype = σ_g



Suppose there are p QTLs, with genotypes denoted $g_1, \ldots, g_p$.

Then μ_g and σ_g depend only on $g_1, \ldots, g_p$.

There are 2^p distinct genotype groups.

Models: Genotype $\longleftrightarrow$ Phenotype

Simplifying assumptions:

Contant variance: $\sigma_g \equiv \sigma$

(environmental variation independent of genotype)

Normality: Given g, y is normal(μ_g, σ)

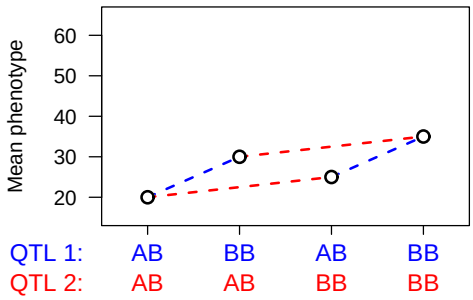
Additivity: $y = \mu + \sum_{j=1}^{p} \Delta_j \, z_j + \epsilon$

where z_j = 1/0 if g_j is AB/BB

Additivity vs. epistasis

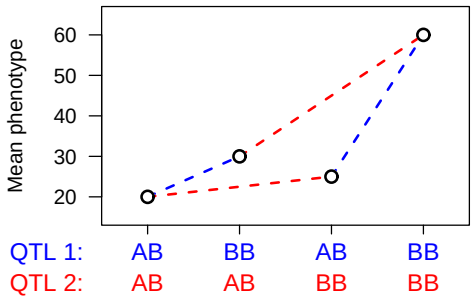
Additivity:

	QTL 1	
QTL 2	AB	BB
AB	20	30
BB	25	35



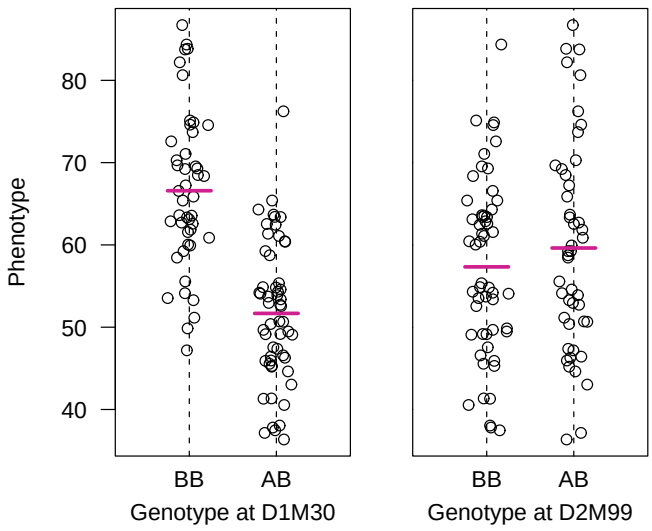
Epistasis:

	QTL 1	
QTL 2	AB	BB
AB	20	30
BB	25	**60**



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



ANOVA at marker loci

Advantages

- Simple.
- Easily incorporates covariates.
- Easily extended to more complex models.
- Doesn’t require a genetic map.

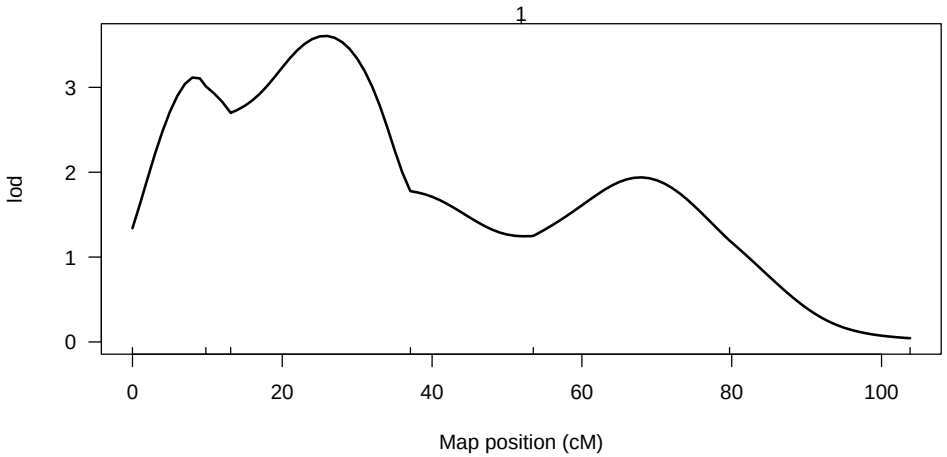
Disadvantages

- Must exclude individuals with missing genotype data.
- Imperfect information about QTL location.
- Suffers in low density scans.
- Only considers one QTL at a time.

Interval mapping (IM)

Lander & Botstein (1989)

- Take account of missing genotype data
- Interpolate between markers



Interval mapping

Advantages

- Takes proper account of missing data.
- Allows examination of positions between markers.
- Gives improved estimates of QTL effects.
- Provides pretty graphs.

Disadvantages

- Increased computation time.
- Requires specialized software.
- Difficult to generalize.
- Only considers one QTL at a time.

LOD scores

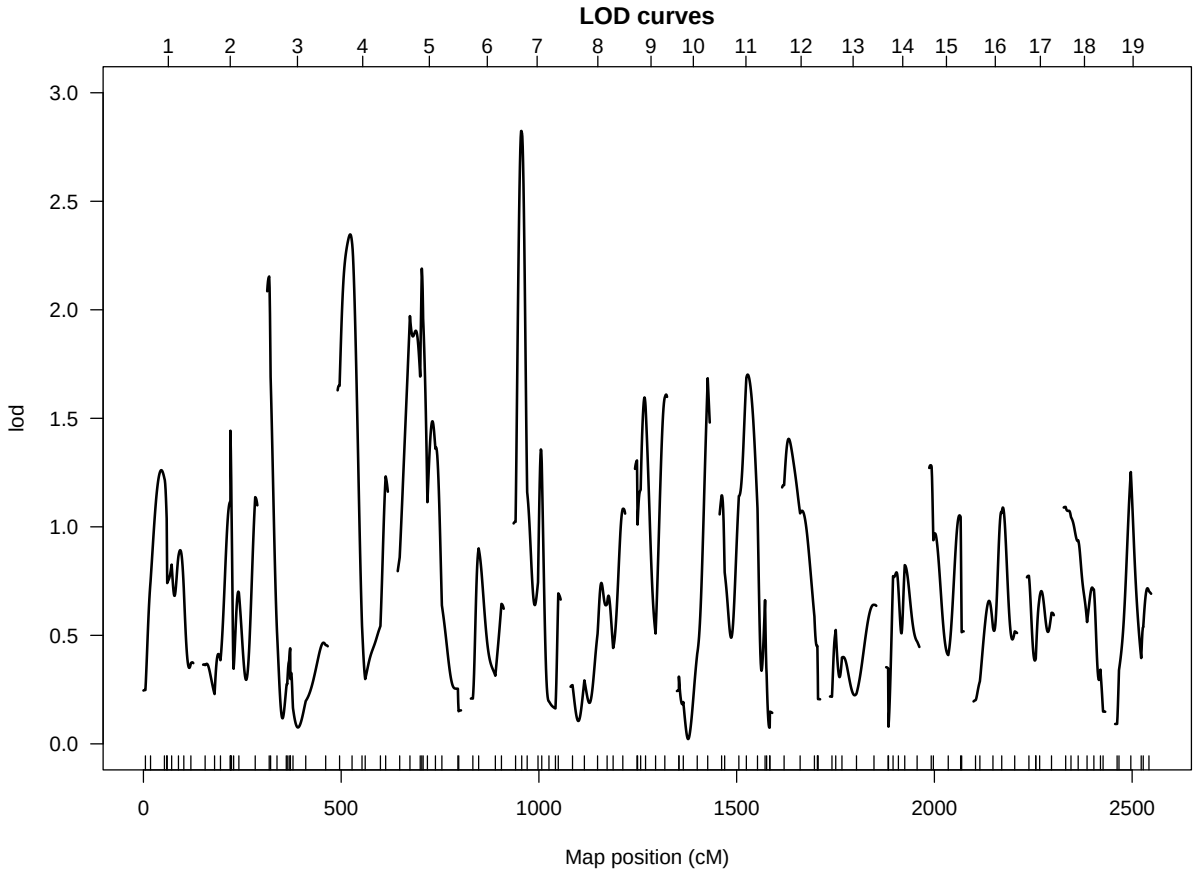
The LOD score is a measure of the strength of evidence for the presence of a QTL at a particular location.

LOD$(z) = \log_{10}$ likelihood ratio comparing the hypothesis of a QTL at position z versus that of no QTL

$$= \log_{10} \left\{ \frac{\Pr(y|\text{QTL at } z, \hat{\mu}_z, \hat{\Delta}_z, \hat{\sigma}_z)}{\Pr(y|\text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\}$$

$\hat{\mu}_z, \hat{\Delta}_z, \hat{\sigma}_z$ are the MLEs, assuming a single QTL at position z.

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



LOD thresholds

Large LOD scores indicate evidence for the presence of a QTL.

Q: How large is large?

$\rightarrow$ We consider the distribution of the LOD score under the null hypothesis of no QTL.

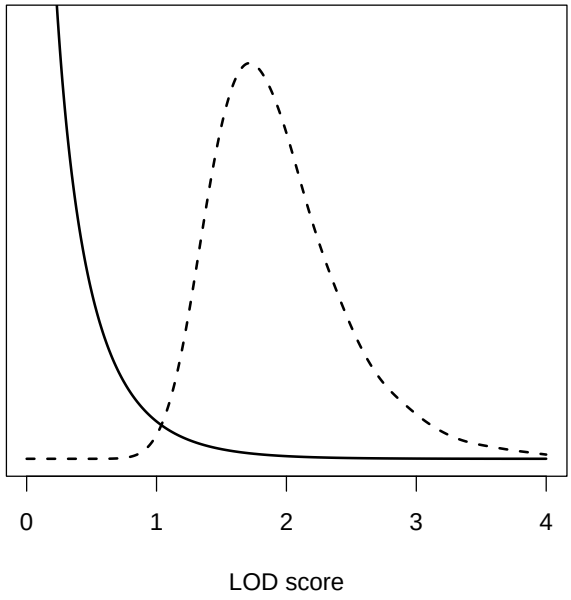
Key point: We must make some adjustment for our examination of multiple putative QTL locations.

$\rightarrow$ We seek the distribution of the *maximum* LOD score, genome-wide. The 95th %ile of this distribution serves as a genome-wide LOD threshold.

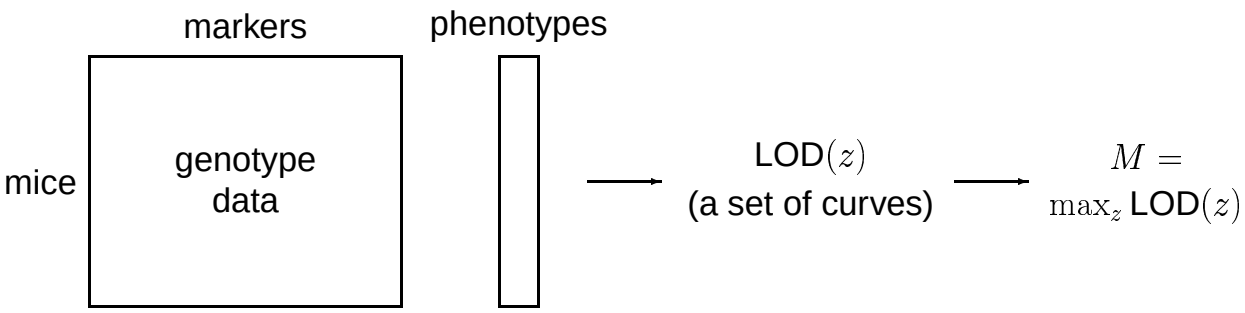
Estimating the threshold: simulations, analytical calculations, permutation (randomization) tests.

Null distribution of the LOD score

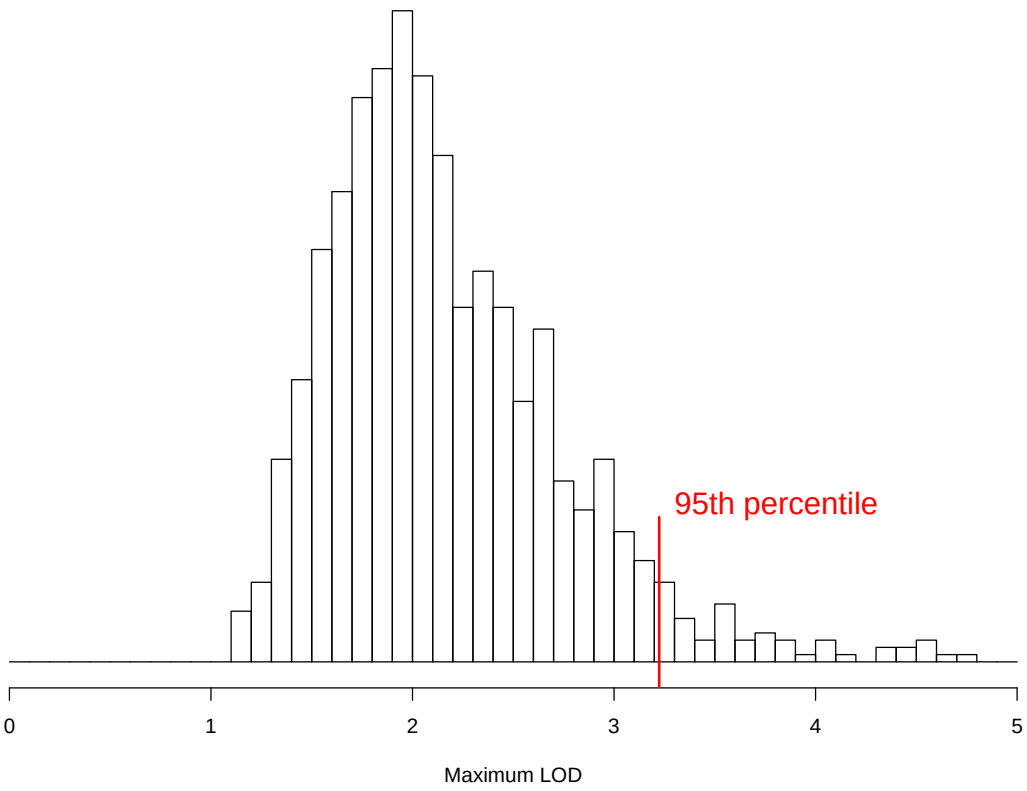
- Null distribution derived by computer simulation of backcross with genome of typical size.
- Solid curve: distribution of LOD score at any one point.
- Dashed curve: distribution of maximum LOD score, genome-wide.



Permutation tests



- Permute/shuffle the phenotypes; keep the genotype data intact.
- Calculate LOD$^\star(z) \longrightarrow M^\star = \max_z$ LOD$^\star(z)$
- We wish to compare the observed M to the distribution of $M^\star$.
- $\Pr(M^\star \geq M)$ is a genome-wide P-value.
- The 95th %ile of $M^\star$ is a genome-wide LOD threshold.
- We can't look at all $n!$ possible permutations, but a random set of 1000 is feasible and provides reasonable estimates of P-values and thresholds.
- Value: conditions on observed phenotypes, marker density, and pattern of missing data; doesn't rely on normality assumptions or asymptotics.



Multiple QTL methods

Why consider multiple QTLs at once?

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

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

Model selection

- Select a class of models
- Compare models
- Search model space
- Assess the perfomance of a procedure

Model fit

Model: $y = \mu + \Delta_3 x_3 + \Delta_7 x_7 + \Delta_9 x_9 + \epsilon$

Model fit: $\hat{\mu}$, $\hat{\Delta}_3$, $\hat{\Delta}_7$, $\hat{\Delta}_9$ by least squares

Fitted values: $\hat{y} = \hat{\mu} + \hat{\Delta}_3 x_3 + \hat{\Delta}_7 x_7 + \hat{\Delta}_9 x_9$

RSS = $\sum_i (y_i - \hat{y}_i)^2$ made as small as possible

Note: If you include an additional x, the RSS goes down.

Class of models

- Additive models
- Additive + pairwise interactions
- Additive + higher order interactions
- Regression trees

Model comparison

- Estimated prediction error
- $\text{BIC}_\delta = \log \text{RSS} + \delta \times$ no. markers $\times \frac{\log n}{n}$
- Sequential permutation tests

$\text{BIC}_\delta \longleftrightarrow$ conditional LOD

Minimizing BIC_δ is approximately equivalent to placing a threshold on the conditional LOD score:

$$\text{LOD}(x_k|x_1,\ldots,x_{k-1})$$

Choosing δ: We choose δ to correspond to a genome-wide LOD threshold.

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

Larger δ: include more loci; higher false positive rate

Smaller δ: include fewer loci; lower false positive rate

Model search

In the case of 100 markers, there are $2^{100} \approx 10^{30}$ possible models—far more than may be inspected individually.

Methods of searching through models:

- Forward selection (FS)
- Backward elimination (BE)
- FS followed by BE
- Randomized searches

Assessing performance

Once must balance

- missing important loci
- including extraneous loci

"Correctly identify a QTL:"

Choose a marker within 10 cM of the QTL.

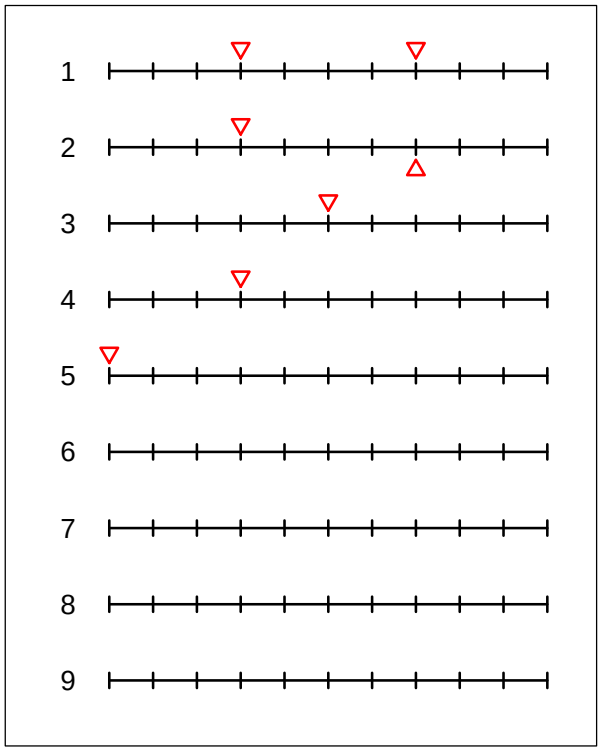
One approach:

Control the false positive rate at 5%

The appropriate criterion depends on the goals of the experimenter

Simulations

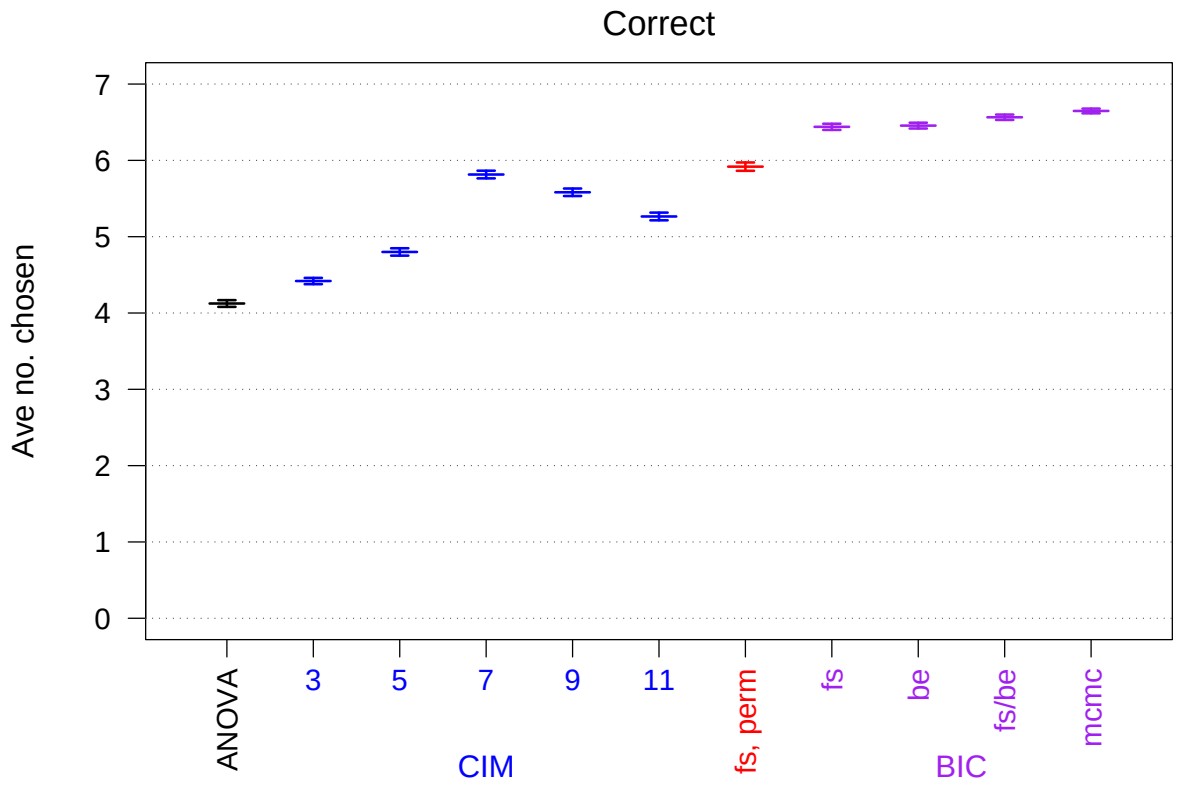
- Backcross with n=250
- No crossover interference
- 9 chr, each 100 cM
- Markers at 10 cM spacing; complete genotype data
- 7 QTLs
  - One pair in coupling
  - One pair in repulsion
  - Three unlinked QTLs
- Heritability = 50%
- 2000 simulation replicates

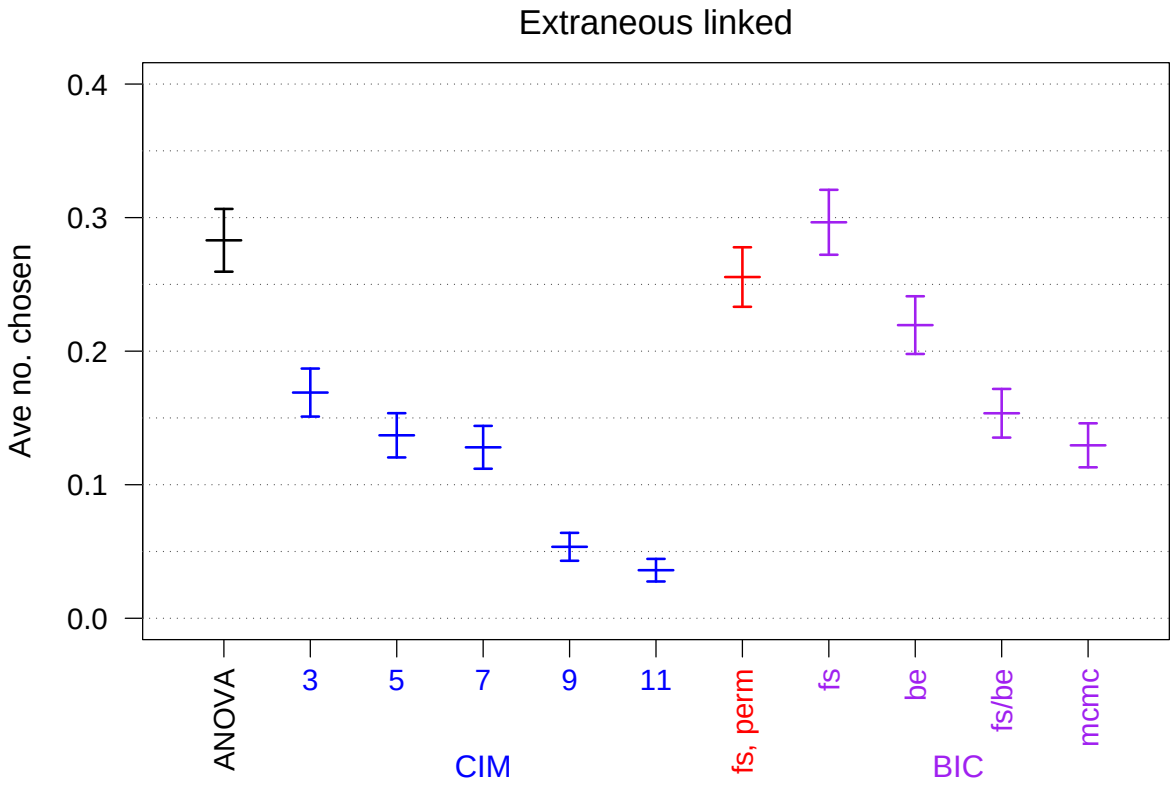


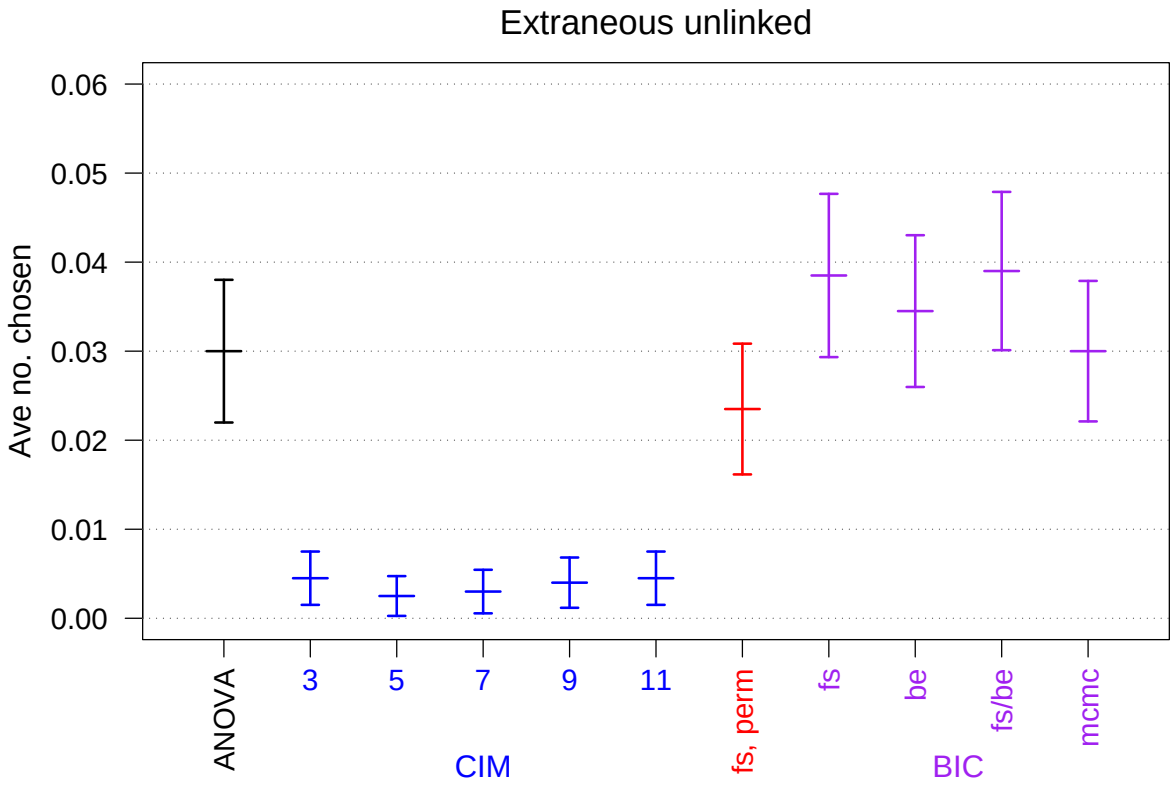
Methods

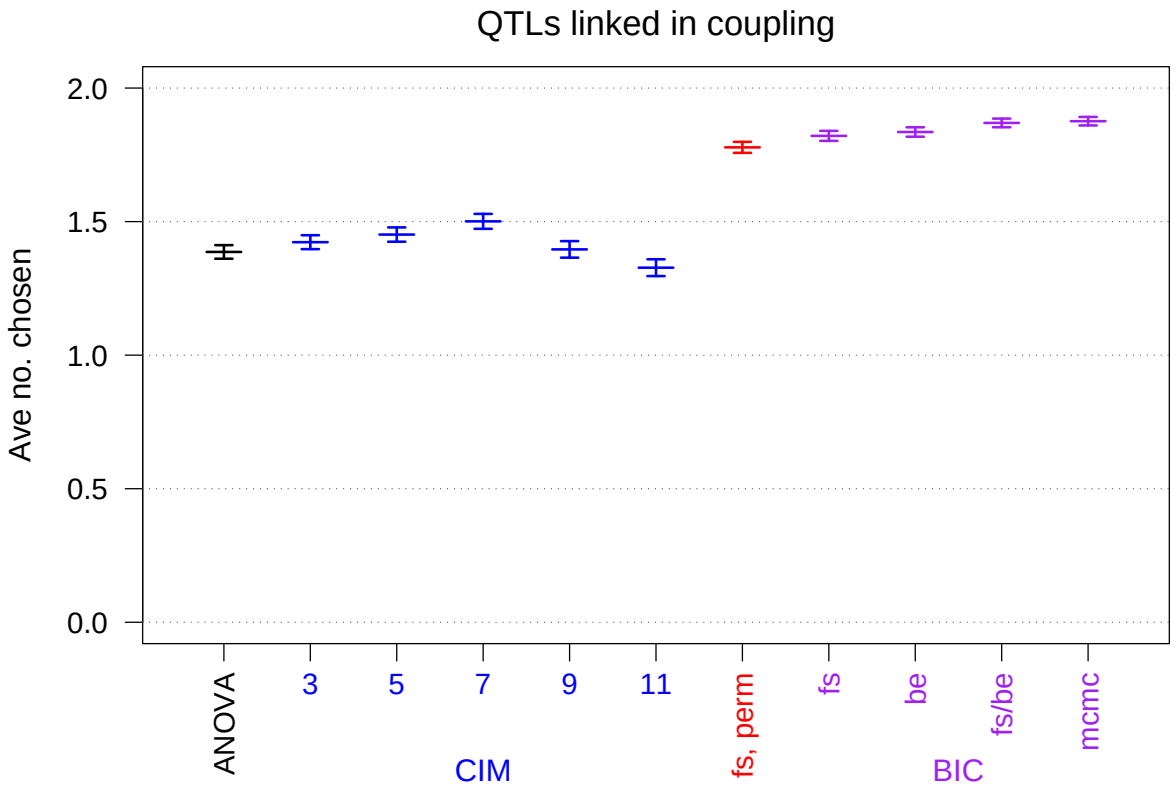
- ANOVA at marker loci
- Composite interval mapping (CIM)
- Forward selection with permutation tests
- Forward selection with BIC_δ
- Backward elimination with BIC_δ
- FS followed by BE with BIC_δ
- MCMC with BIC_δ

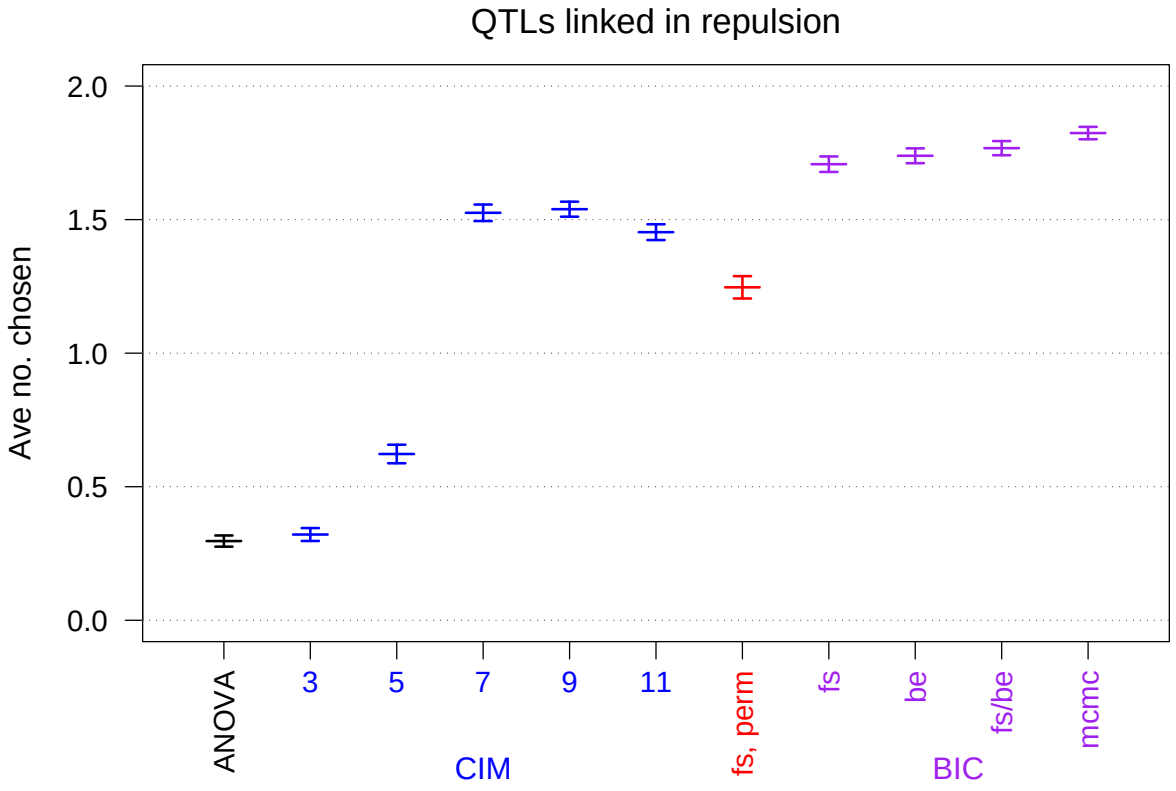
$\longrightarrow$ A selected marker is deemed correct if it is within 10 cM of a QTL (i.e., correct or adjacent)

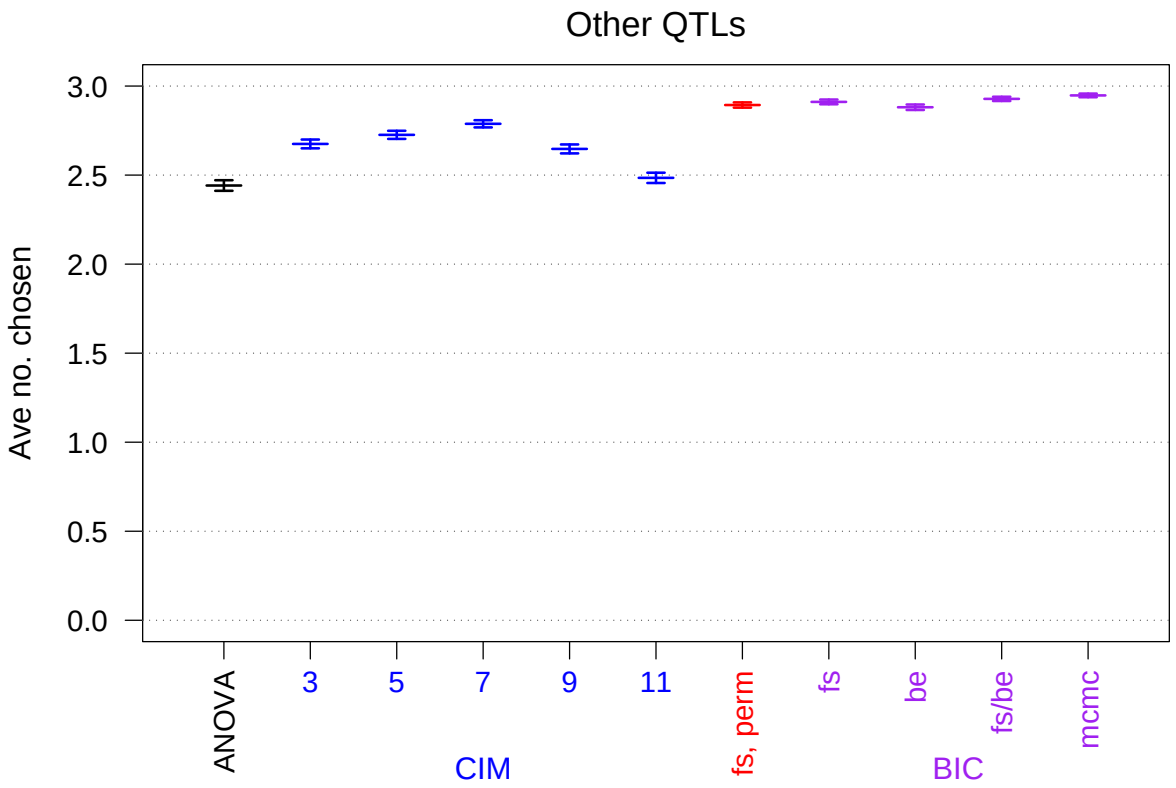



Extraneous linked
Ave no. chosen
0.0
0.1
0.2
0.3
0.4
ANOVA
3
5
7
9
11
fs, perm
fs
be
fs/be
mcmc
CIM
BIC


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


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



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.