QTL mapping in mice

Karl W Broman

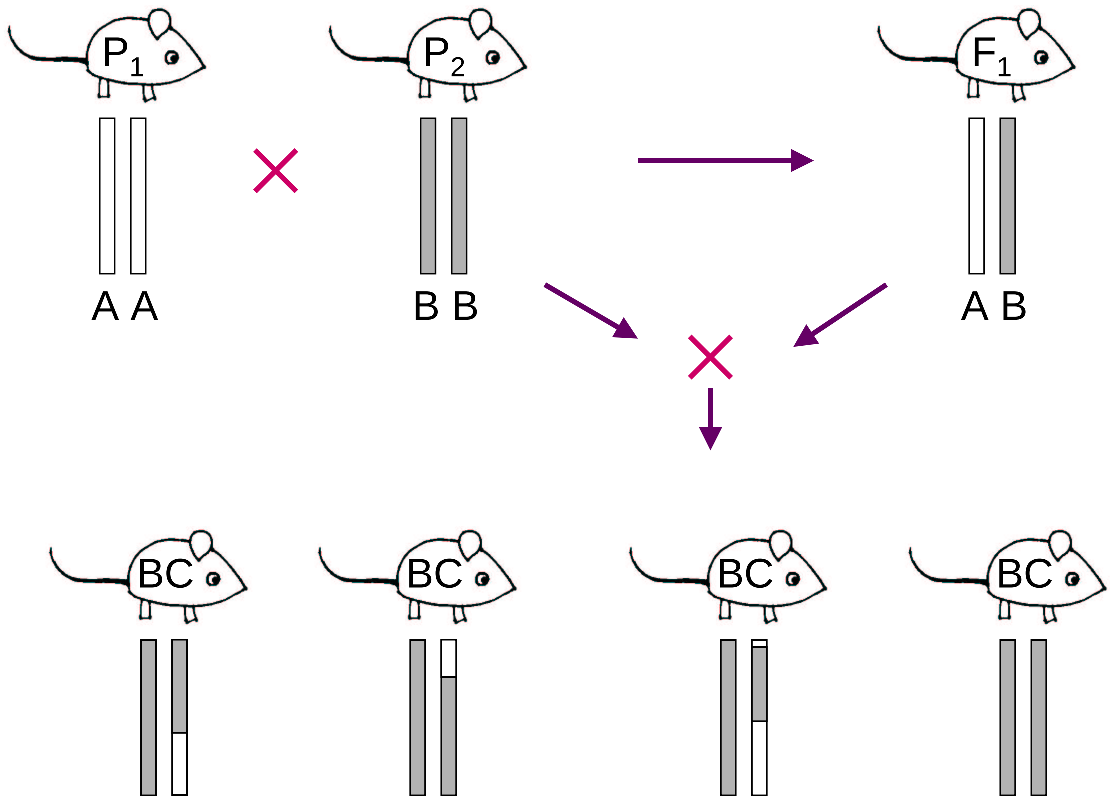
Department of Biostatistics
Johns Hopkins University
Baltimore, Maryland, USA

`www.biostat.jhsph.edu/~kbroman`

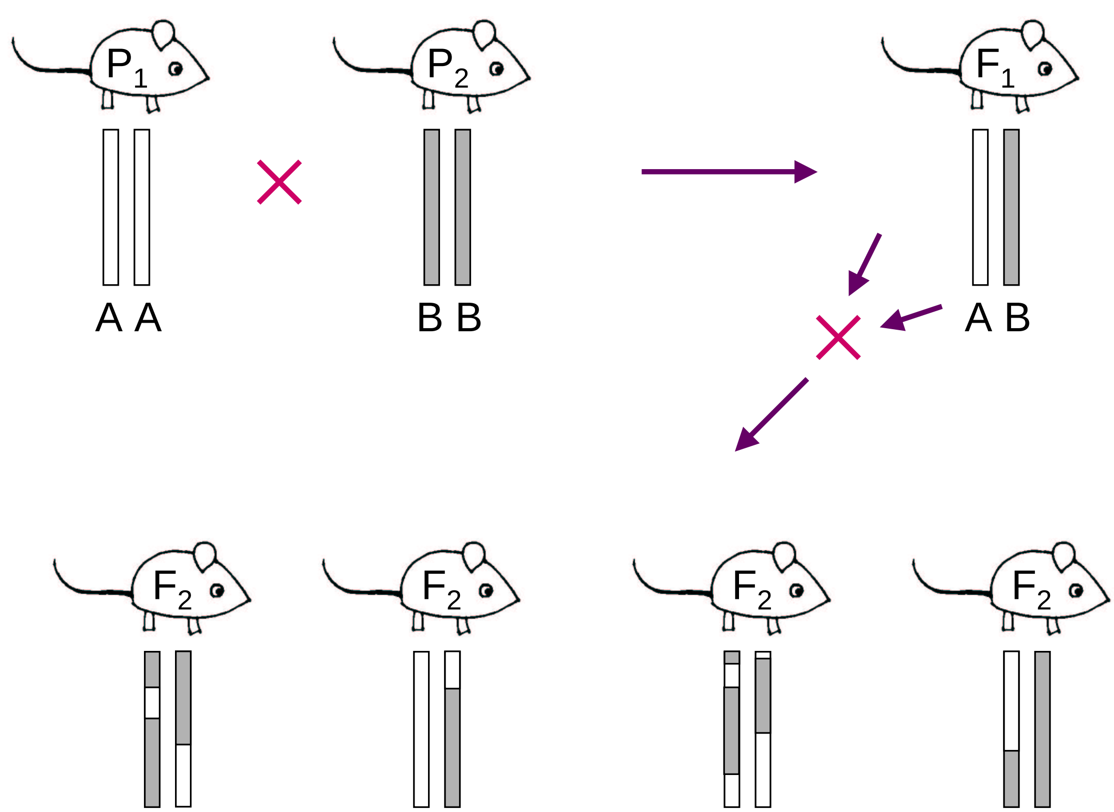
Outline

- Experiments, data, and goals
- Models
- ANOVA at marker loci
- Interval mapping
- LOD scores, LOD thresholds
- Mapping multiple QTLs
- Simulations

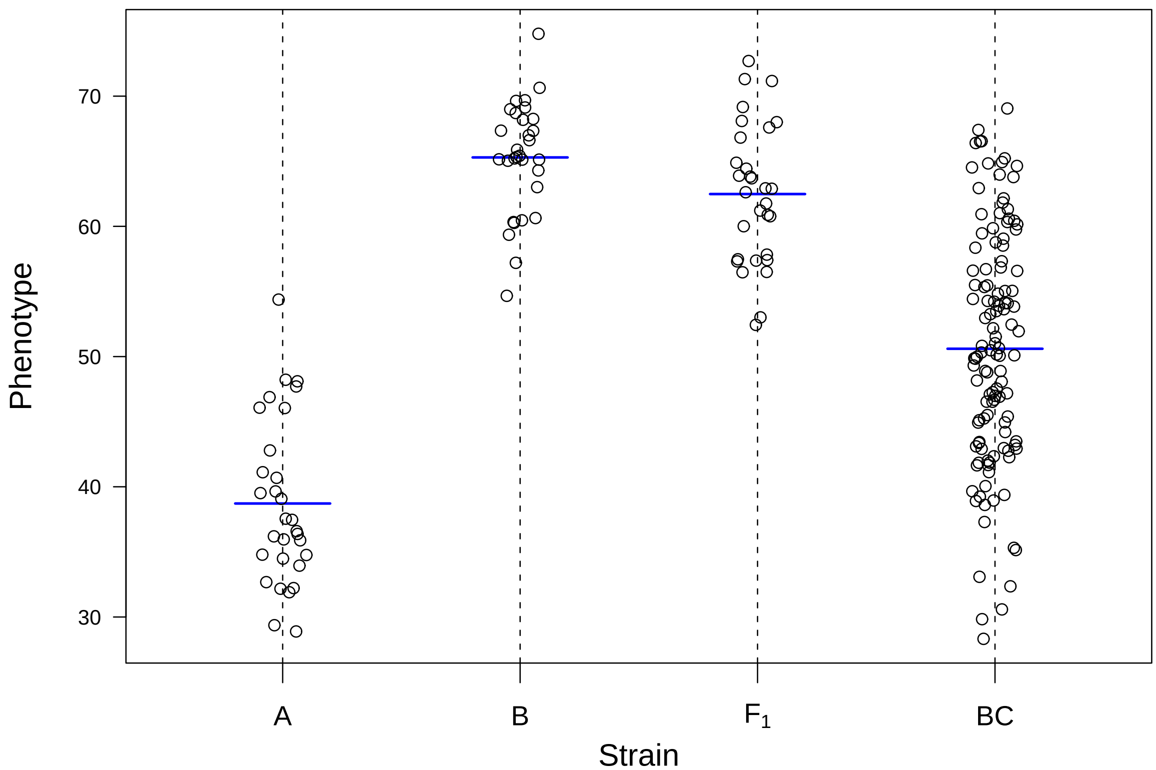
Backcross experiment



Intercross 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.

Note: QTL = "quantitative trait locus"

Why?

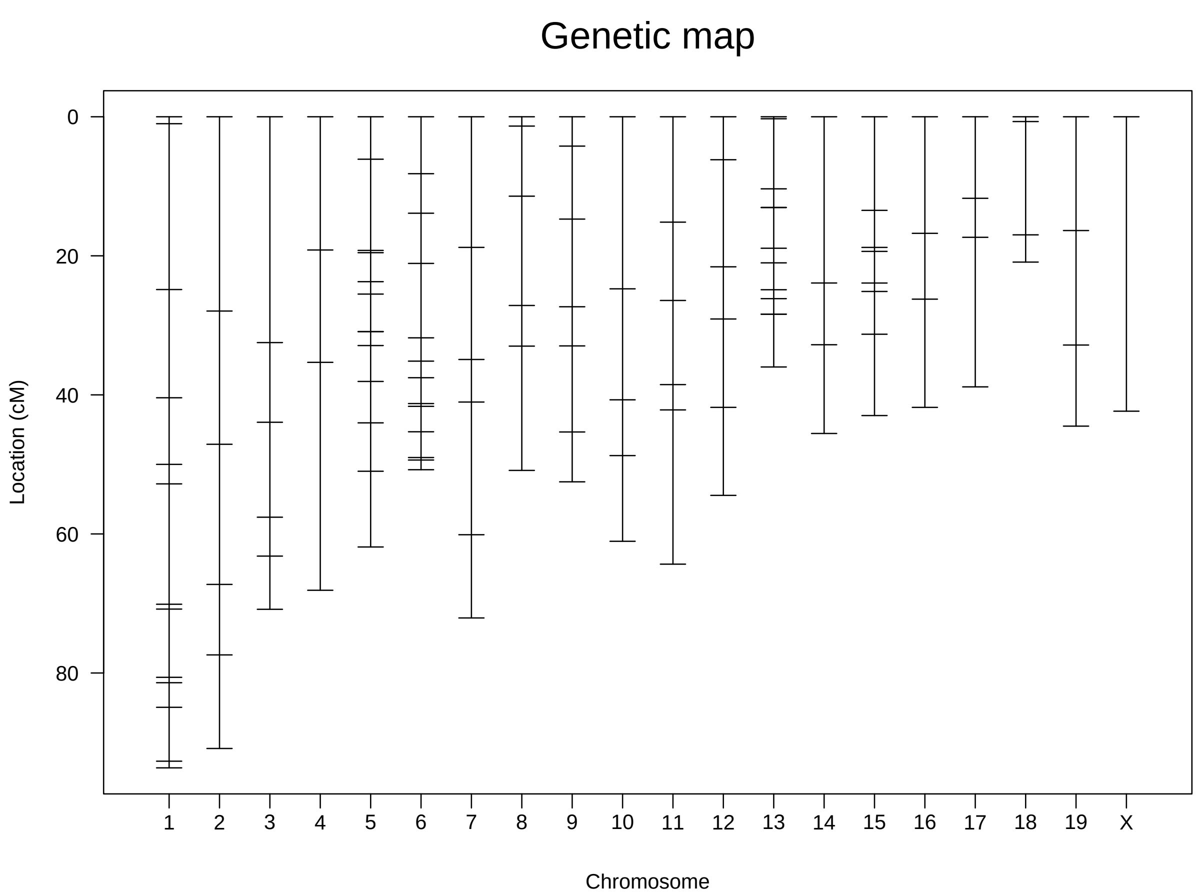
Mice: Find gene

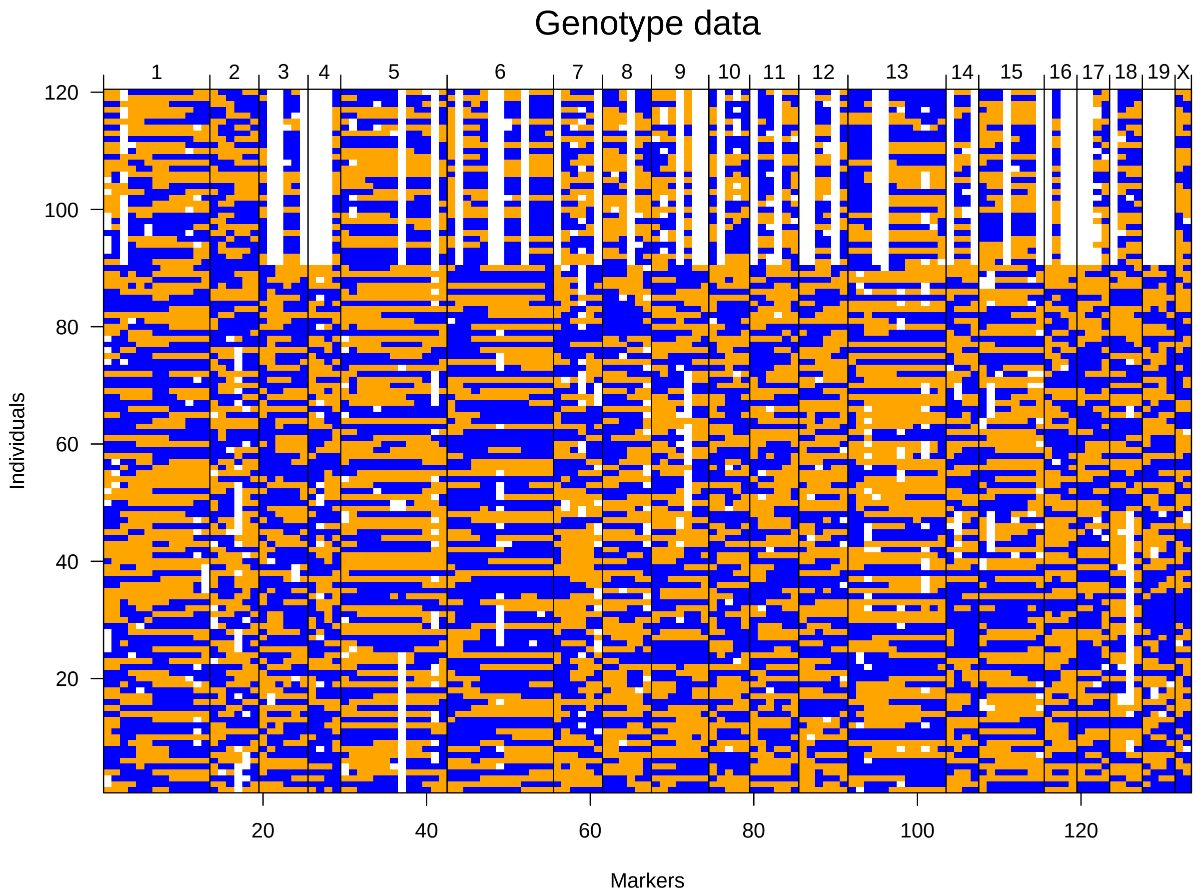
$\longrightarrow$ Drug targets, biochemical basis

Agronomy: Selection for improvement

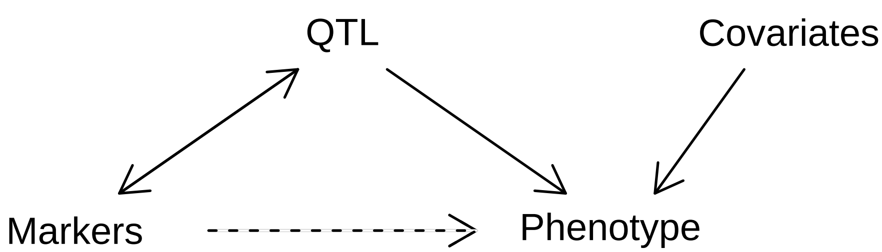
Flies: Genetic architecture

$\longrightarrow$ Evolution





Statistical structure



The missing data problem:

$$\text{Markers} \longleftrightarrow \text{QTL}$$

The model selection problem:

$$\text{QTL, covariates} \longrightarrow \text{phenotype}$$

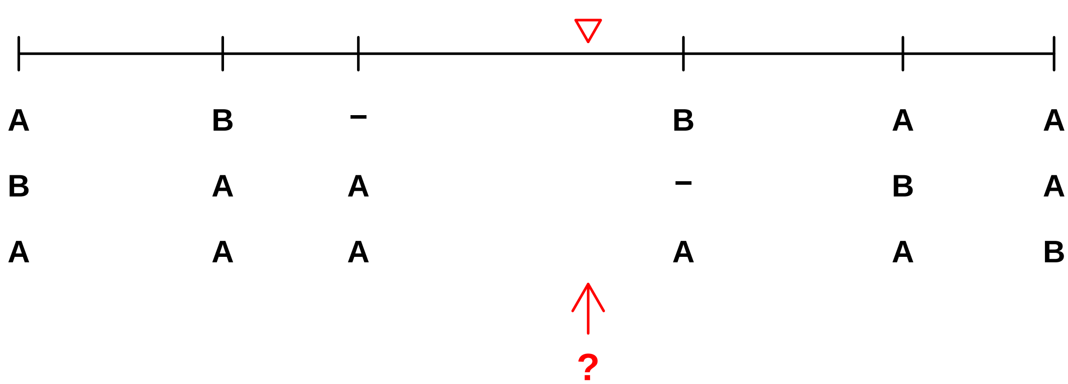
Models: Recombination

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

Example



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)

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

Models: Genotype $\longleftrightarrow$ Phenotype

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

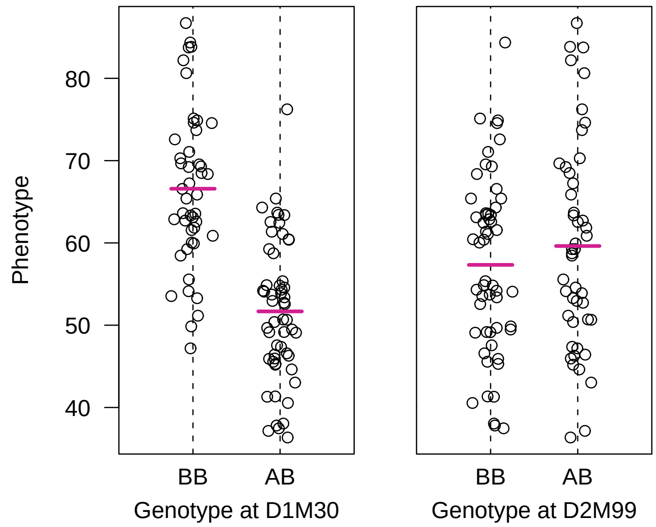
Normally distributed residual variation: $y|g \sim N(\mu_g, \sigma^2)$.

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.

The simplest method: ANOVA

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



ANOVA at marker loci

Advantages

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

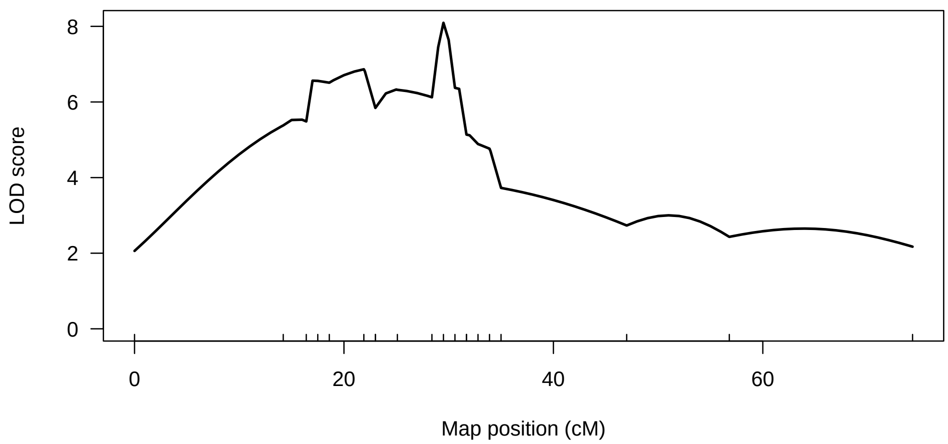
Disadvantages

- 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
- Maximum likelihood under a mixture model



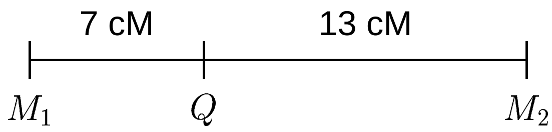
Interval mapping (IM)

Lander & Botstein (1989)

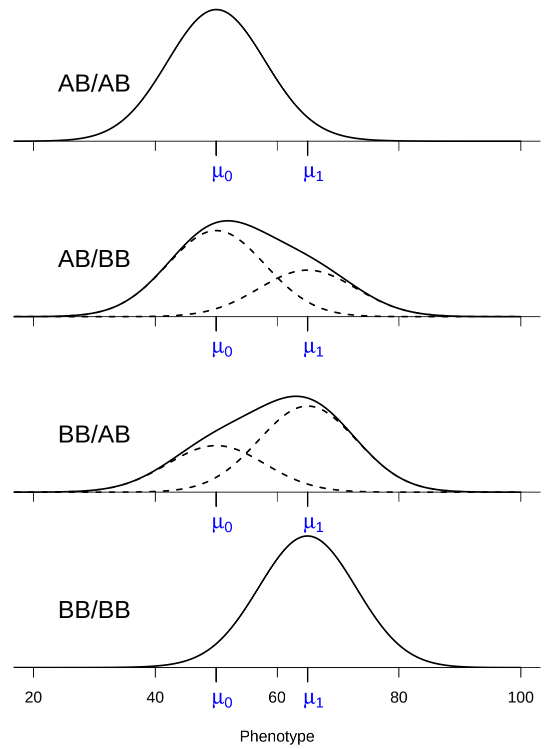
- Assume a single QTL model.
- Each position in the genome, one at a time, is posited as the putative QTL.
- Let $z =$ 1/0 if the (unobserved) QTL genotype is BB/AB. Assume $y \sim N(\mu_z, \sigma)$
- Given genotypes at linked markers, $y \sim$ mixture of normal dist'ns with mixing proportion $\Pr(z = 1|\text{marker data})$:

		QTL genotype	
M_1	M_2	BB	AB
BB	BB	$(1-r_L)(1-r_R)/(1-r)$	$r_L r_R/(1-r)$
BB	AB	$(1-r_L)r_R/r$	$r_L(1-r_R)/r$
AB	BB	$r_L(1-r_R)/r$	$(1-r_L)r_R/r$
AB	AB	$r_L r_R/(1-r)$	$(1-r_L)(1-r_R)/(1-r)$

The normal mixtures



- Two markers separated by 20 cM, with the QTL closer to the left marker.
- The figure at right show the distributions of the phenotype conditional on the genotypes at the two markers.
- The dashed curves correspond to the components of the mixtures.



Interval mapping (continued)

Let $p_i = \Pr(z_i = 1|\text{marker data})$

$y_i|z_i \sim N(\mu_{z_i}, \sigma^2)$

$\Pr(y_i|\text{marker data}, \mu_0, \mu_1, \sigma) = p_i f(y_i; \mu_1, \sigma) + (1 - p_i) f(y_i; \mu_0, \sigma)$

where $f(y; \mu, \sigma) =$ density of normal distribution

Log likelihood: $l(\mu_0, \mu_1, \sigma) = \sum_i \log \Pr(y_i|\text{marker data}, \mu_0, \mu_1, \sigma)$

Maximum likelihood estimates (MLEs) of μ_0, μ_1, σ:

EM algorithm.

LOD scores

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

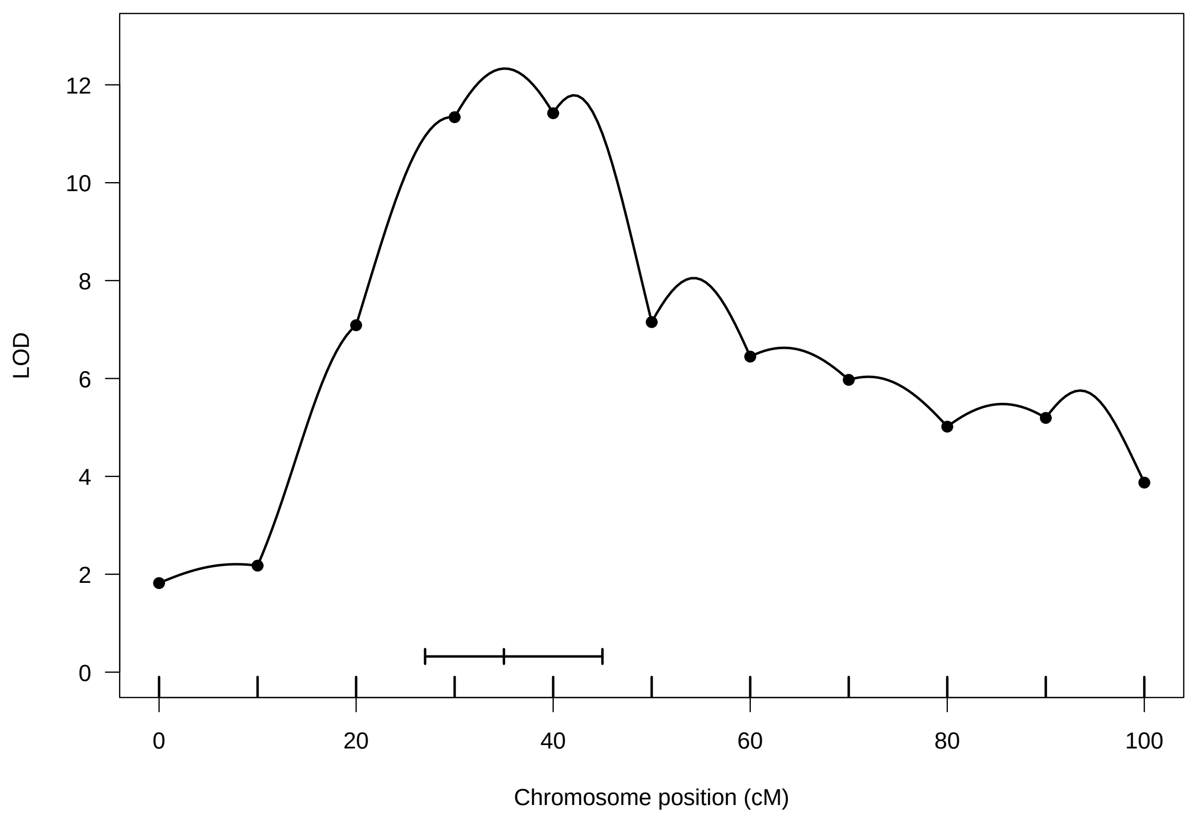
LOD$(z) =$ $\log_{10}$ likelihood ratio comparing the hypothesis of a QTL at position z versus that of no QTL

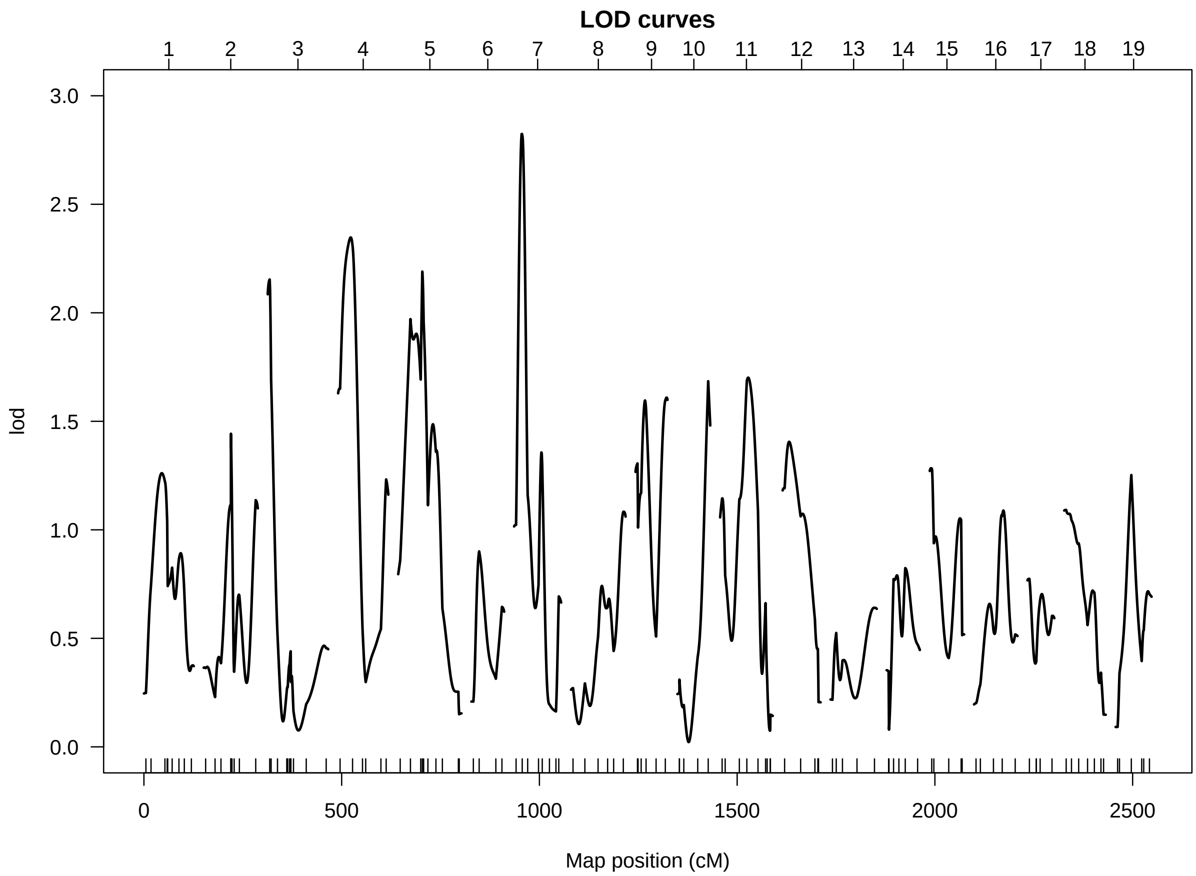
$$= \log_{10} \left\{ \frac{\Pr(y|\text{QTL at } z, \hat{\mu}_{0z}, \hat{\mu}_{1z}, \hat{\sigma}_z)}{\Pr(y|\text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\}$$

$\hat{\mu}_{0z}, \hat{\mu}_{1z}, \hat{\sigma}_z$ are the MLEs, assuming a single QTL at position z.

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

An example LOD curve





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 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.

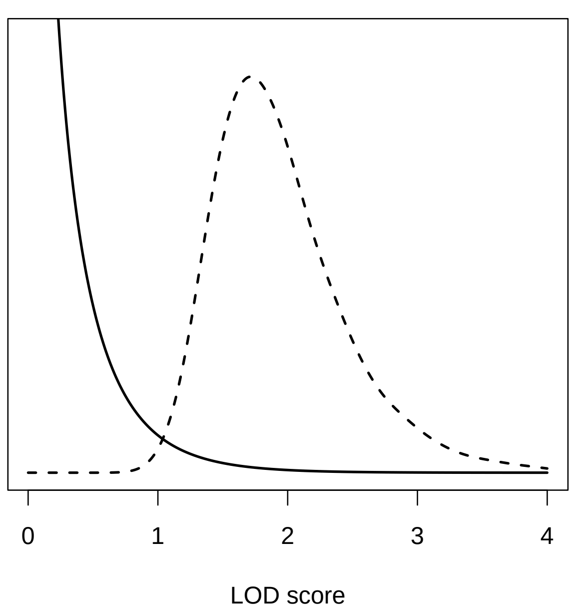
Key point: We must make some adjustment for our examination of multiple putative QTL locations.

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

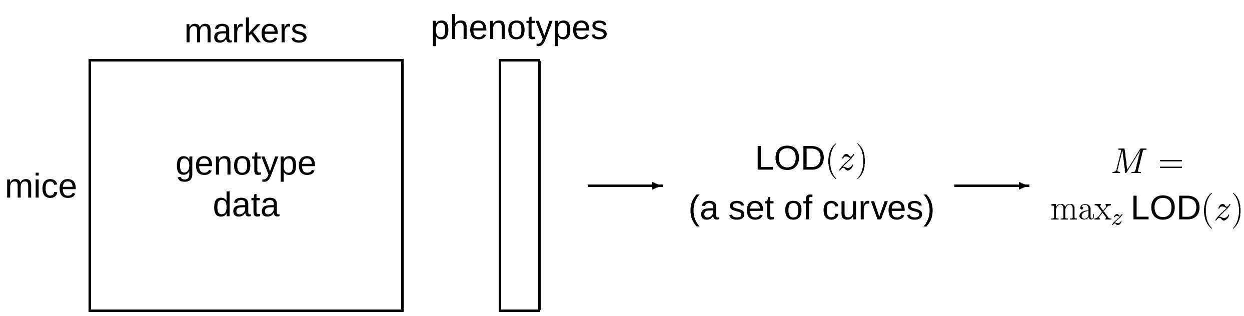
Estimating the threshold: simulations, analytical calculations, permutation (randomization) tests.

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.

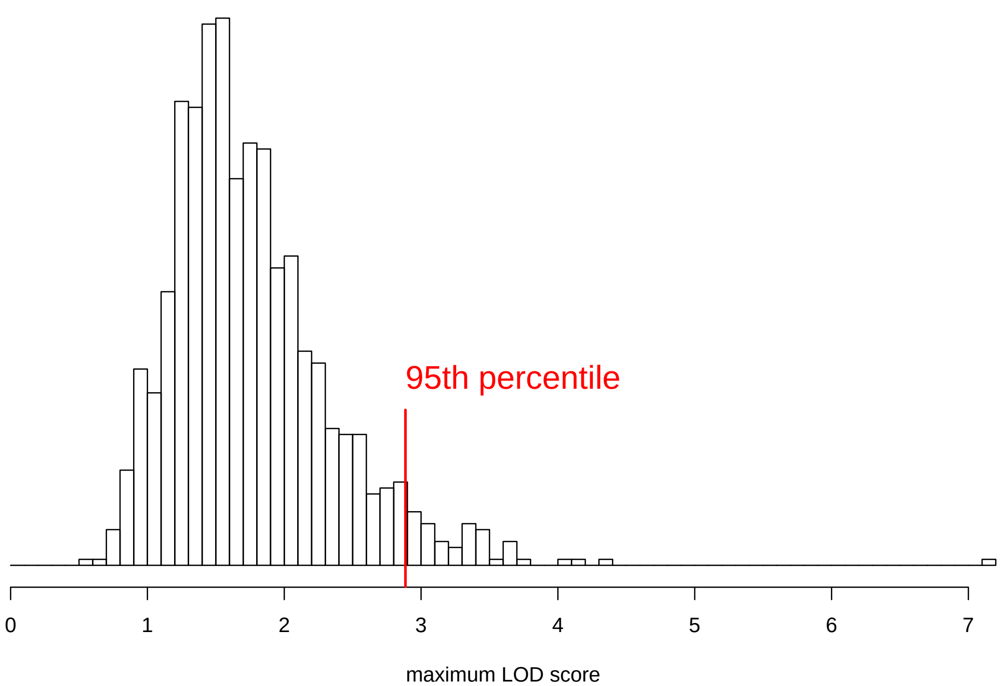


Permutation tests



- Permute/shuffle the phenotypes; keep the genotype data intact.
- Calculate $\text{LOD}^\star(z) \longrightarrow M^\star = \max_z \text{LOD}^\star(z)$
- 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.

Permutation distribution



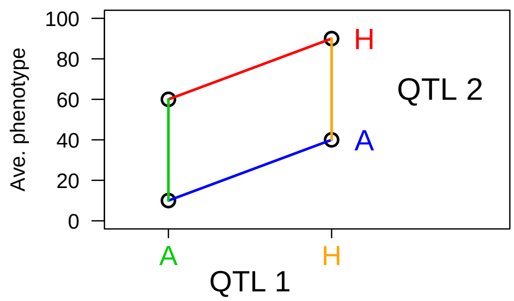
Multiple QTL methods

Why consider multiple QTLs at once?

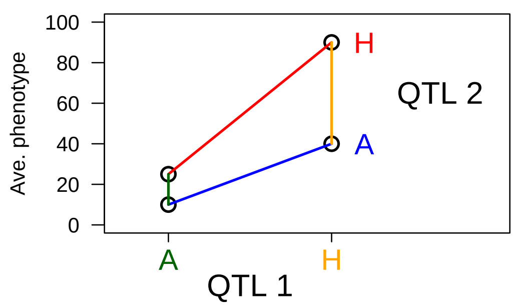
- Reduce residual variation.
- Separate linked QTLs.
- Investigate interactions between QTLs (epistasis).

Epistasis in a backcross

Additive QTLs

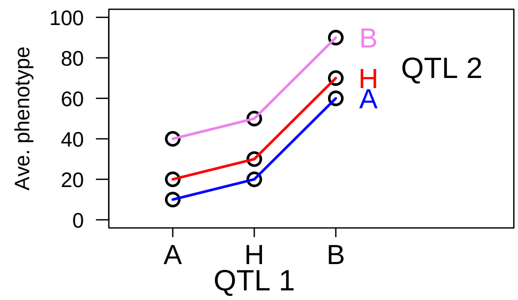


Interacting QTLs

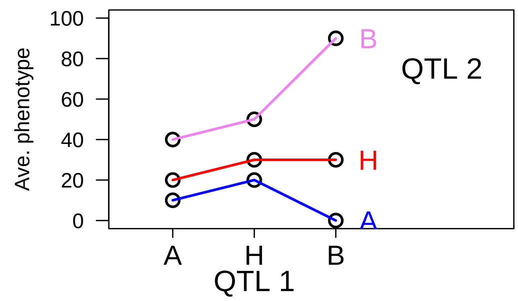


Epistasis in an intercross

Additive QTLs



Interacting QTLs



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
  - Estimate of prediction error
- Assess performance
  - Maximize no. QTLs found; control false positive rate

Why BIC_δ?

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

Choice of δ

Smaller δ: include more loci; higher false positive rate

Larger δ: include fewer loci; lower false positive rate

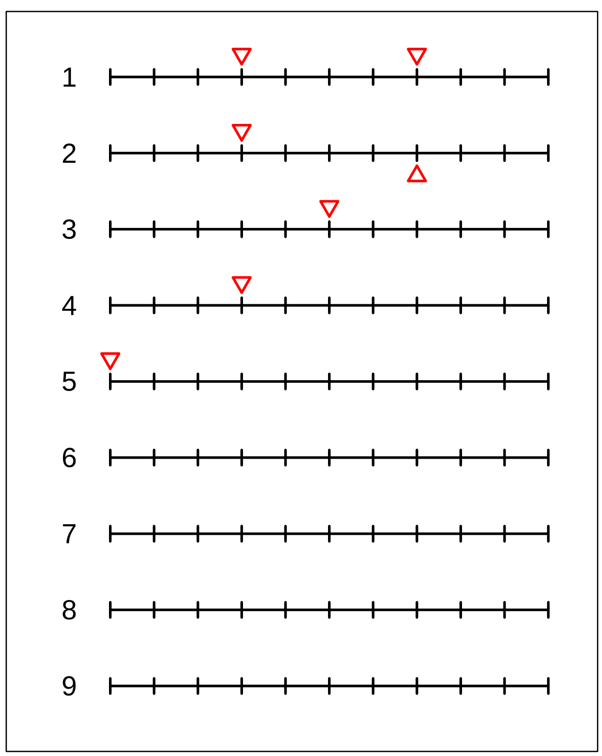
Let L = 95% genome-wide LOD threshold
(compare single-QTL models to the null model)

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

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

Simulations

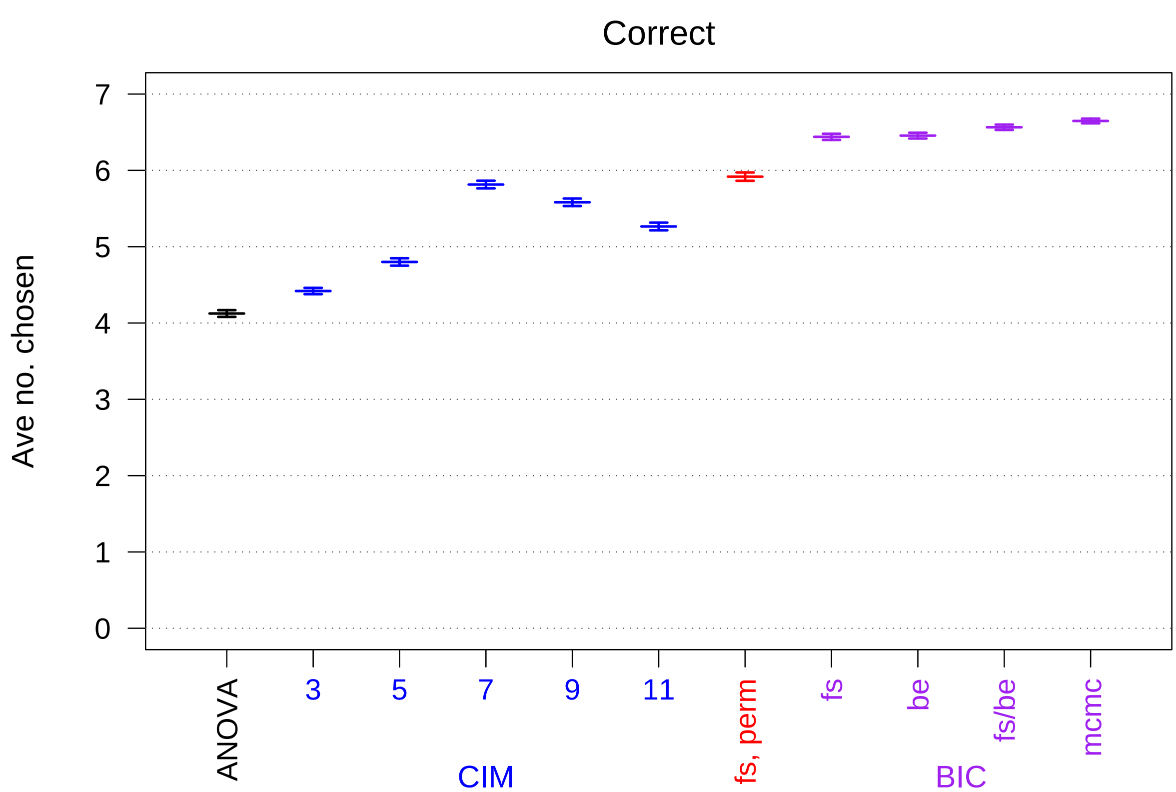
- Backcross with n=250
- No crossover interference
- 9 chr, each 100 cM
- Markers at 10 cM spacing; complete genotype data
- 7 QTLs
  - One pair in coupling
  - One pair in repulsion
  - Three unlinked QTLs
- Heritability = 50%
- 2000 simulation replicates

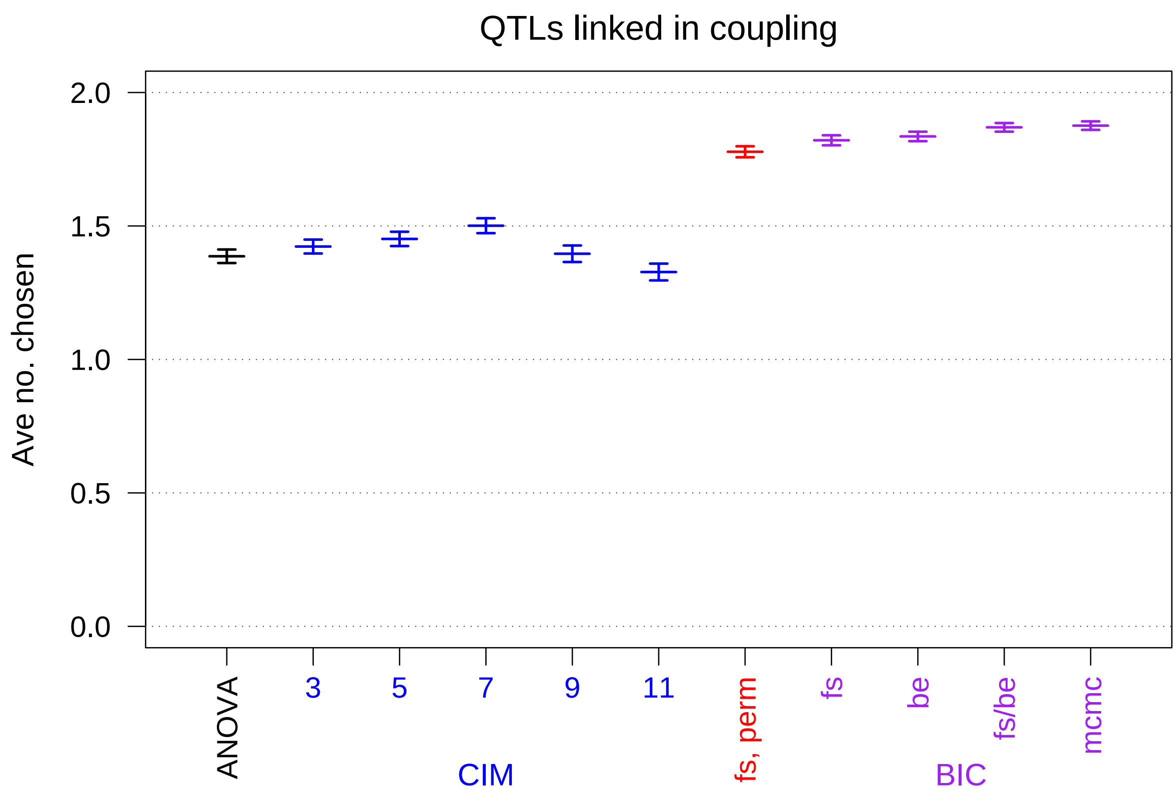


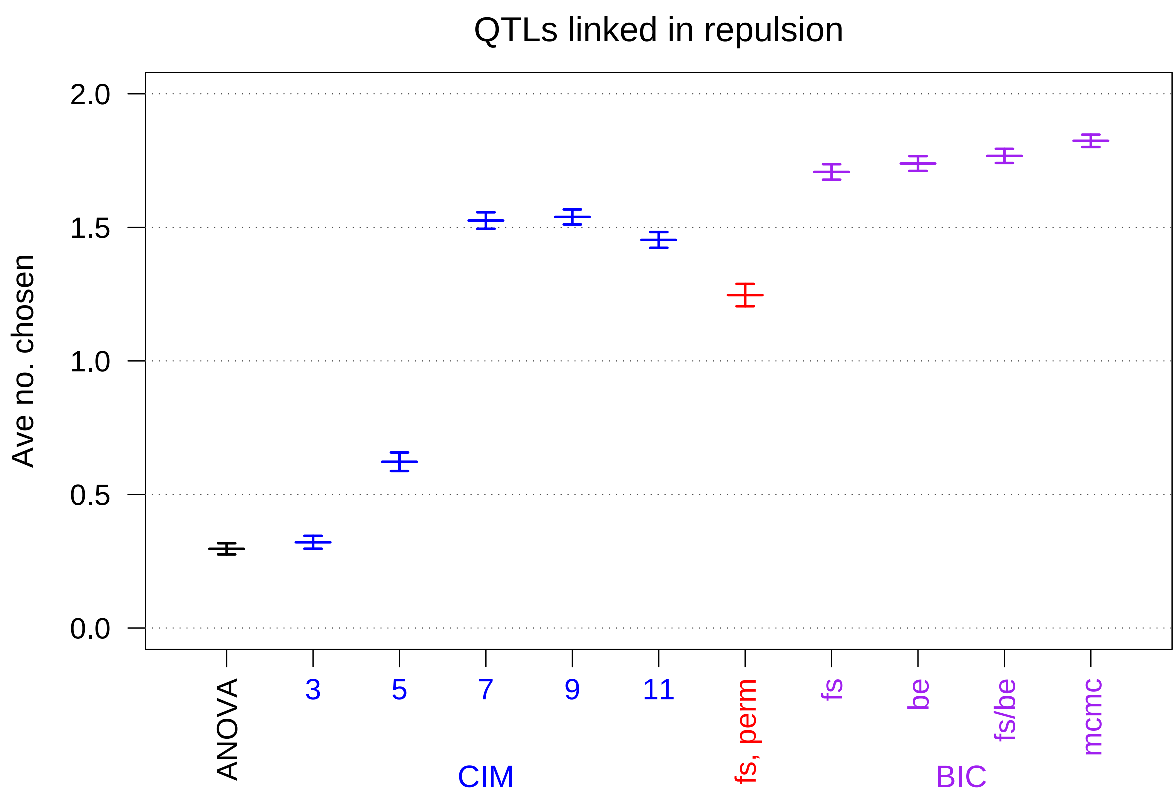
Methods

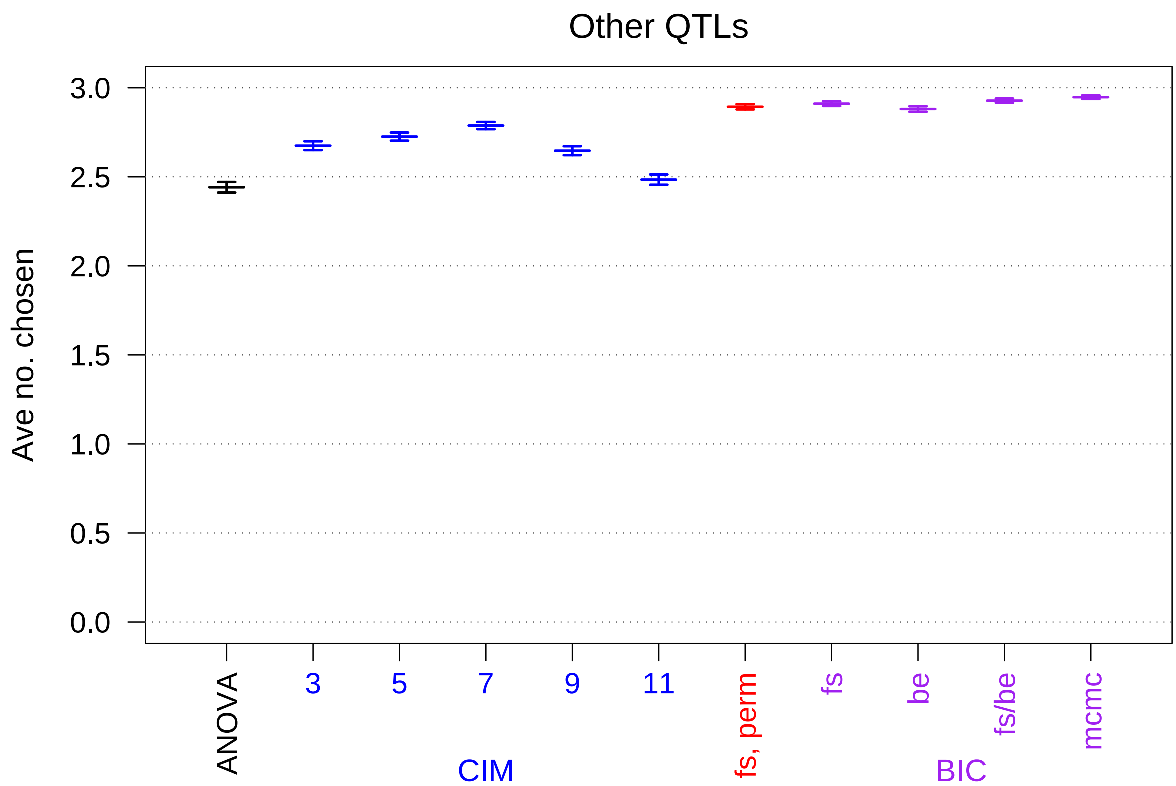
- ANOVA at marker loci
- Composite interval mapping (CIM)
- Forward selection with permutation tests
- Forward selection with BIC_δ
- Backward elimination with BIC_δ
- FS followed by BE with BIC_δ
- MCMC with BIC_δ

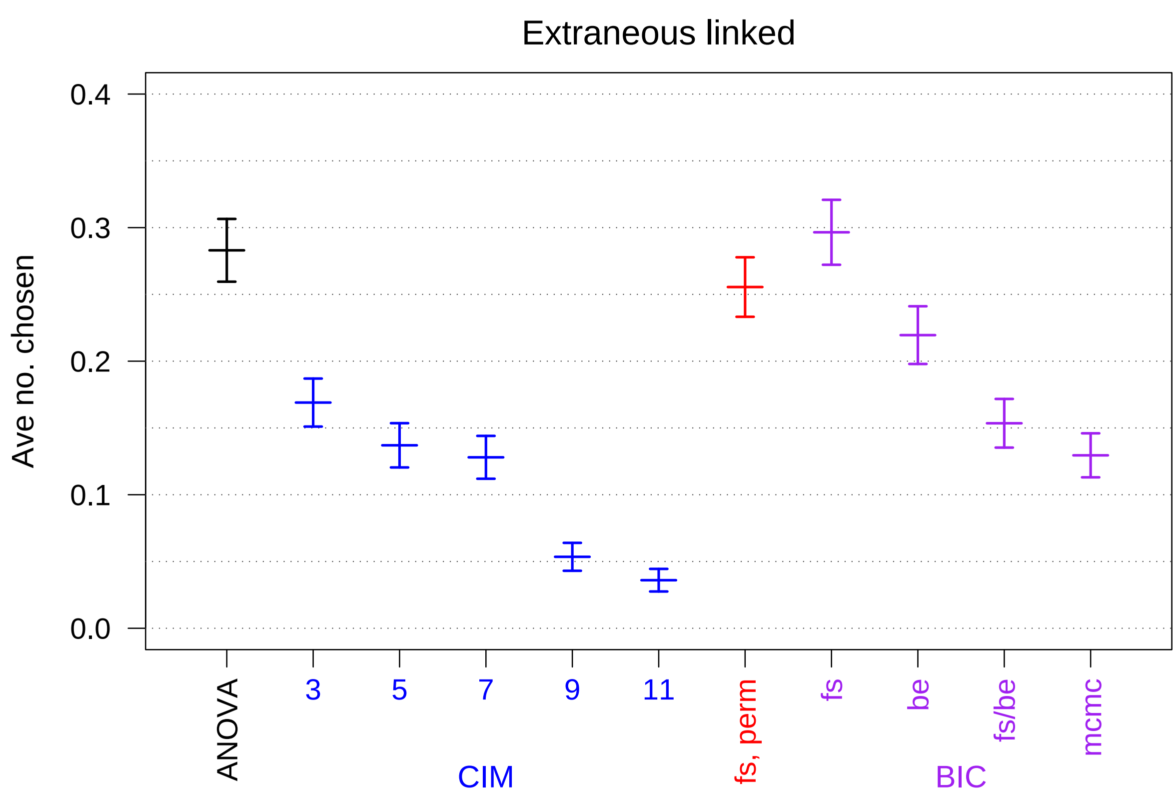
$\longrightarrow$ A selected marker is deemed correct if it is within 10 cM of a QTL (i.e., correct or adjacent)

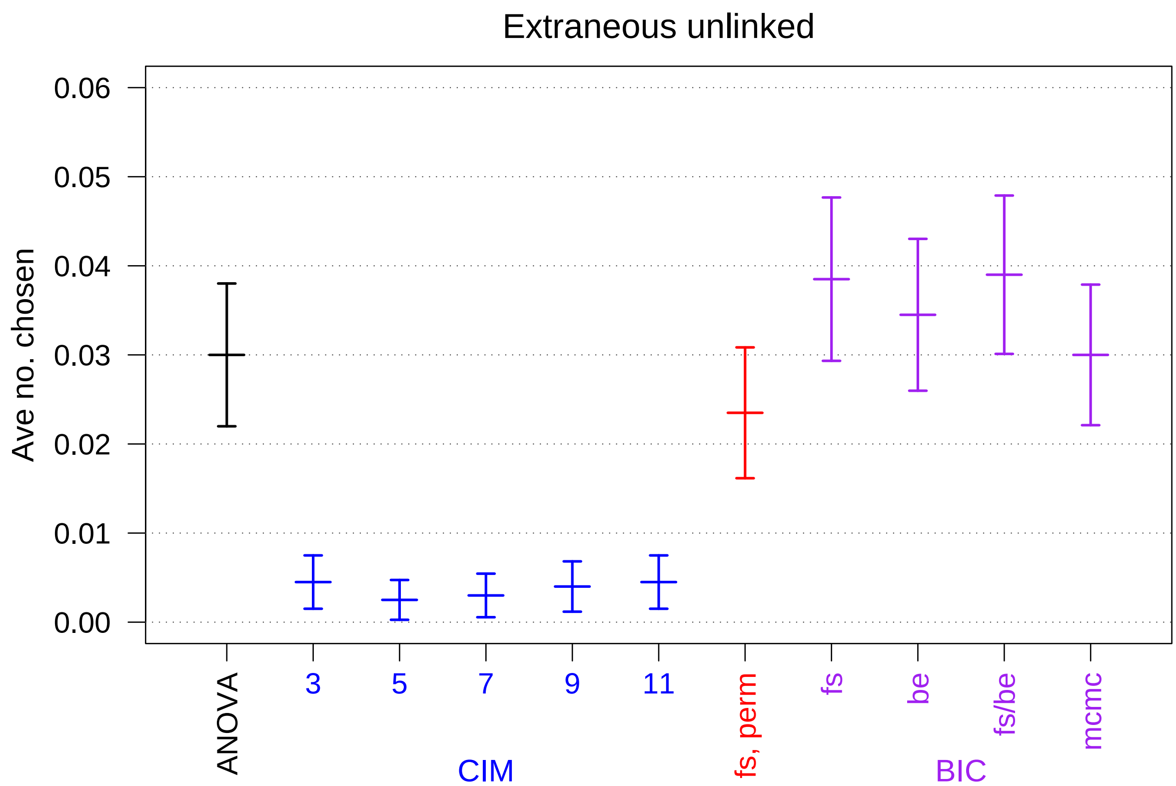



QTLs linked in coupling
Ave no. chosen
0.0
0.5
1.0
1.5
2.0
ANOVA
3
5
7
9
11
CIM
fs, perm
fs
be
fs/be
mcmc
BIC


QTLs linked in repulsion
Ave no. chosen
0.0
0.5
1.0
1.5
2.0
ANOVA
3
5
7
9
11
CIM
fs, perm
fs
be
fs/be
mcmc
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
0.4
0.3
0.2
0.1
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 (in the case of additive QTLs).

Acknowledgements

Terry Speed, University of California, Berkeley, and WEHI

Gary Churchill, The Jackson Laboratory

Śaunak Sen, University of California, San Francisco

References

- Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. Lab Animal 30(7):44–52

  Review for non-statisticians

- Broman KW, Speed TP (1999) A review of methods for identifying QTLs in experimental crosses. In: Seillier-Moiseiwitch F (ed) Statistics in Molecular Biology. IMS Lecture Notes—Monograph Series. Vol 33, pp. 114–142

  Older, more statistical review.

- Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

  The seminal paper.

- Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971

  LOD thresholds by permutation tests.

- Miller AJ (2002) *Subset selection in regression*, 2nd edition. Chapman & Hall, New York.

  A reasonably good book on model selection in regression.

- Strickberger MW (1985) *Genetics*, 3rd edition. Macmillan, New York, chapter 11.

  An old but excellent general genetics textbook with a very interesting discussion of epistasis.

- 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, 737–775

  Contains the simulation study described above.