Multiple QTL mapping

Karl W Broman

Department of Biostatistics
Johns Hopkins University

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

[→ Teaching → Miscellaneous lectures]

Why?

- Reduce residual variation $\Longrightarrow$ increased power
- Separate linked QTL
- Identify interactions among QTL

Hypothesis testing?

- In the past, QTL mapping has been regarded as a task of hypothesis testing.

  Is this a QTL?

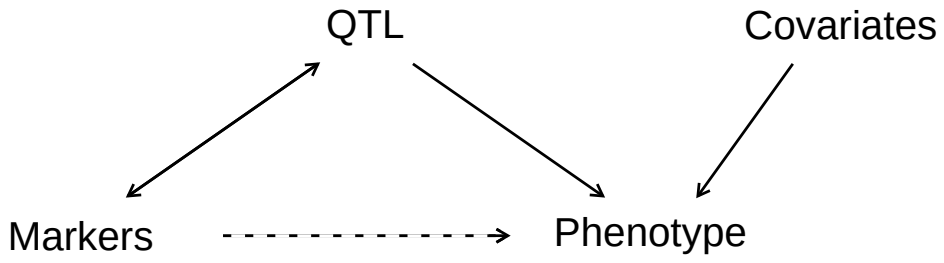
  Much of the focus has been on adjusting for test multiplicity.

- It is better to view the problem as one of model selection.

  What set of QTL are well supported?
  Is there evidence for QTL-QTL interactions?

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

Statistical structure



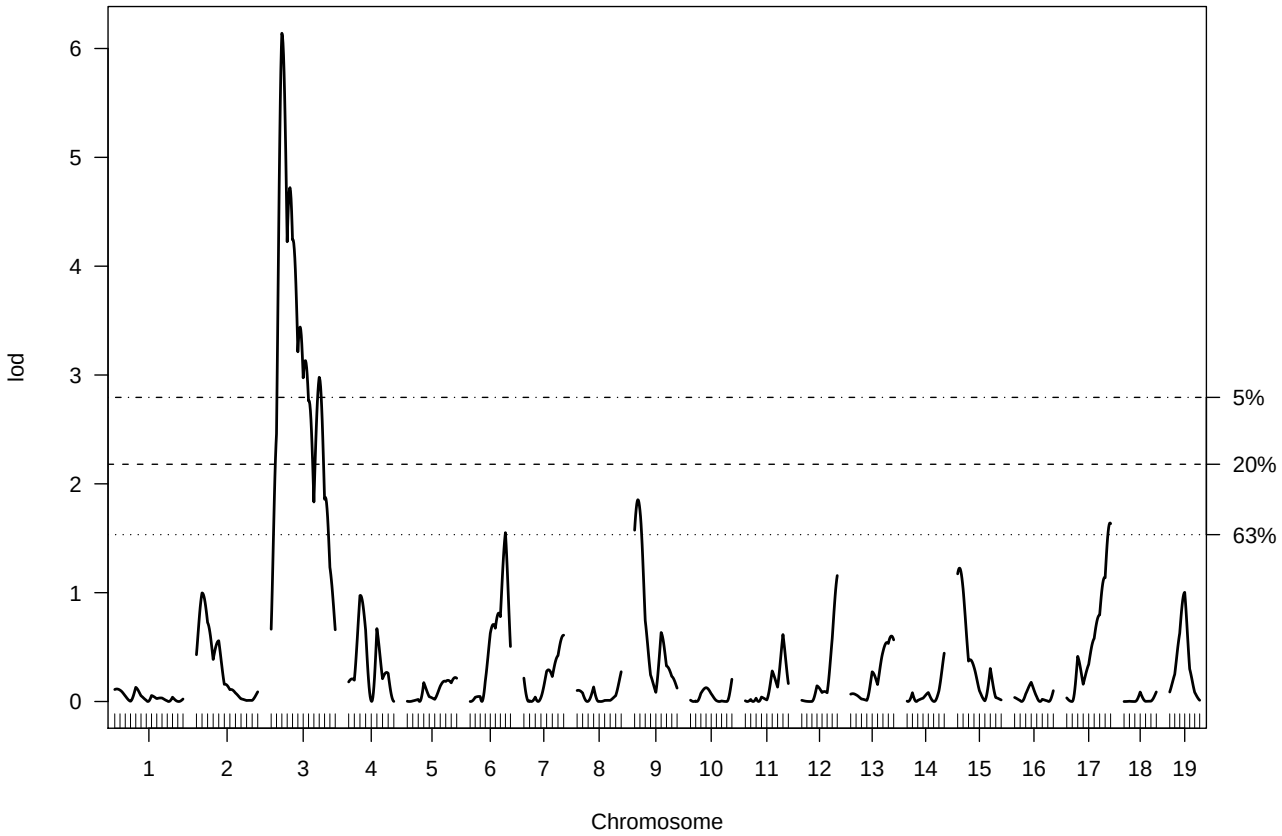
The missing data problem: Markers $\longleftrightarrow$ QTL

The model selection problem: QTL, covariates $\longrightarrow$ phenotype

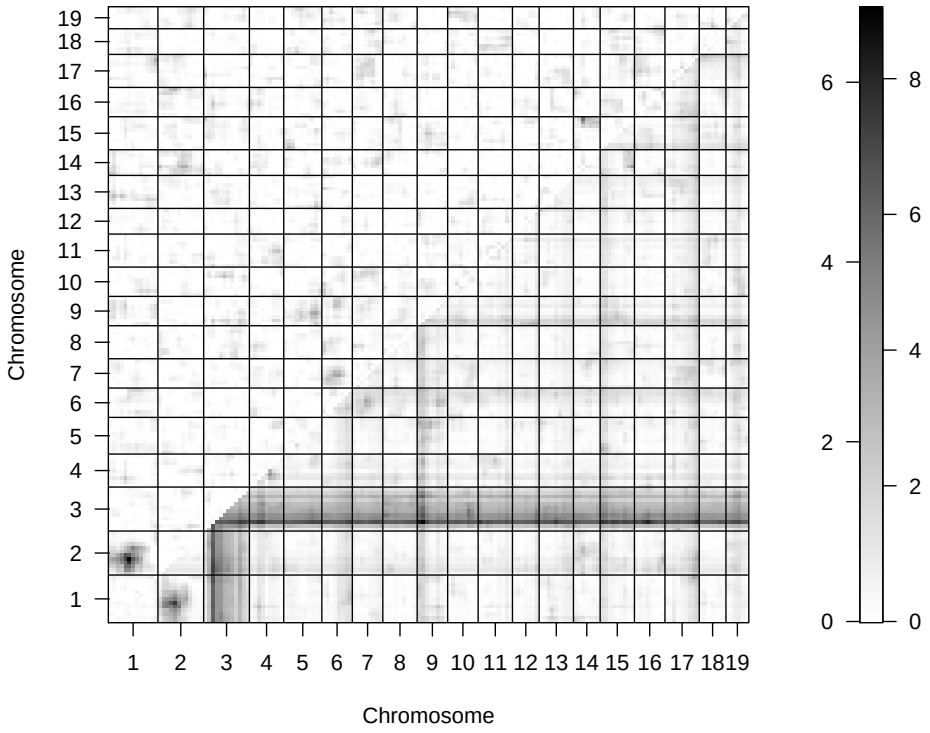
Starting points

- Single-QTL scan (ie, interval mapping)
  - Loci with marginal effects should appear
- 2-dim, 2-QTL scan
  - Ability to separate linked QTL
  - Identify interacting loci

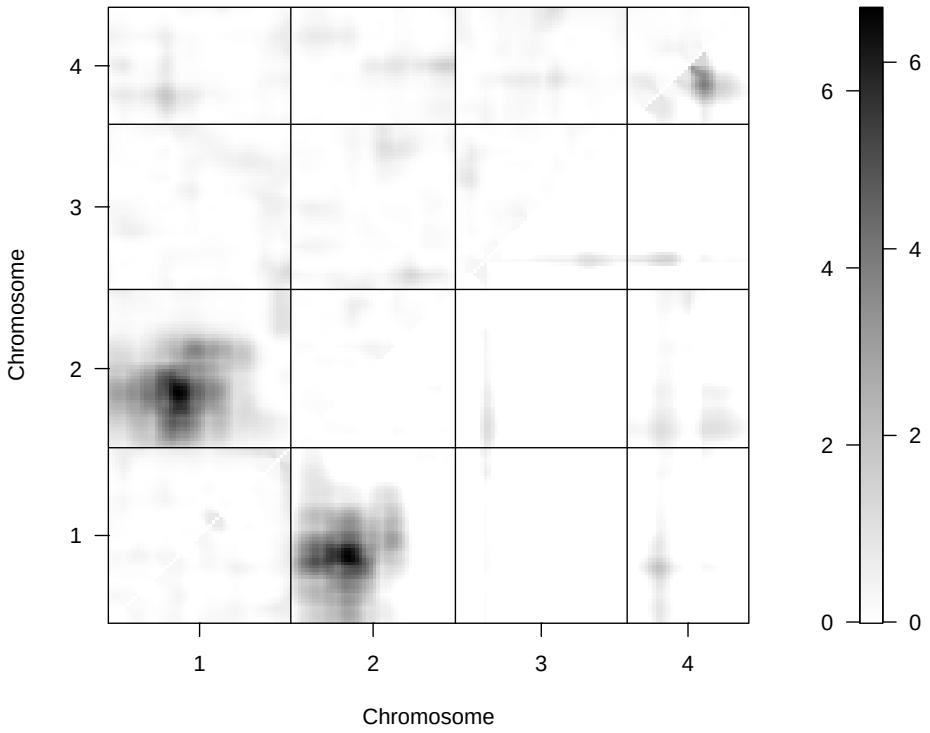
Example: 1d scan



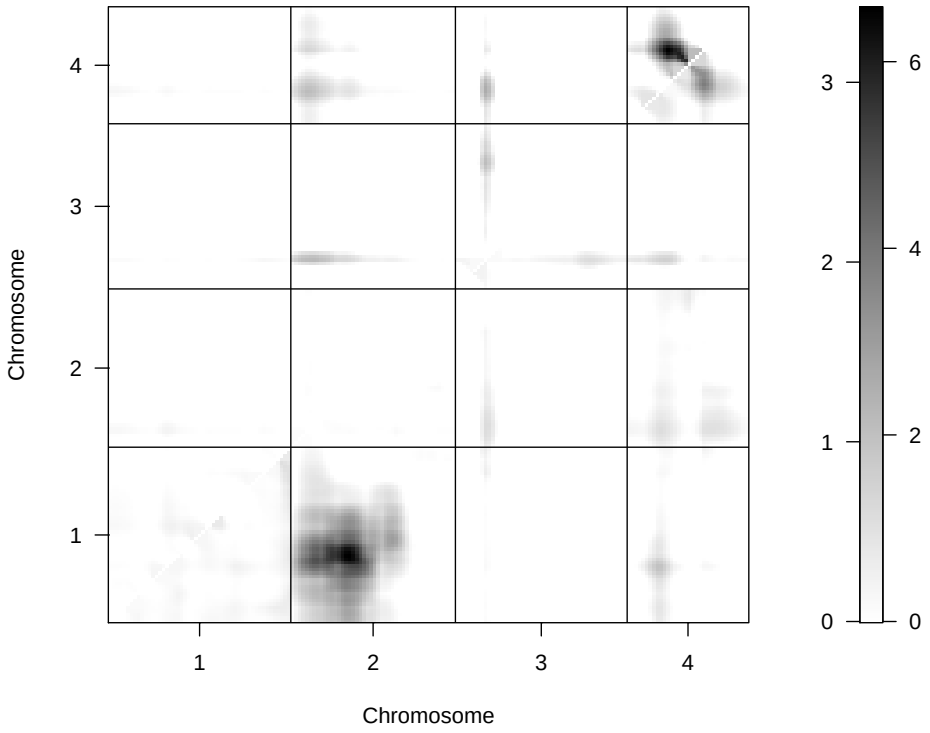
Example: 2d scan



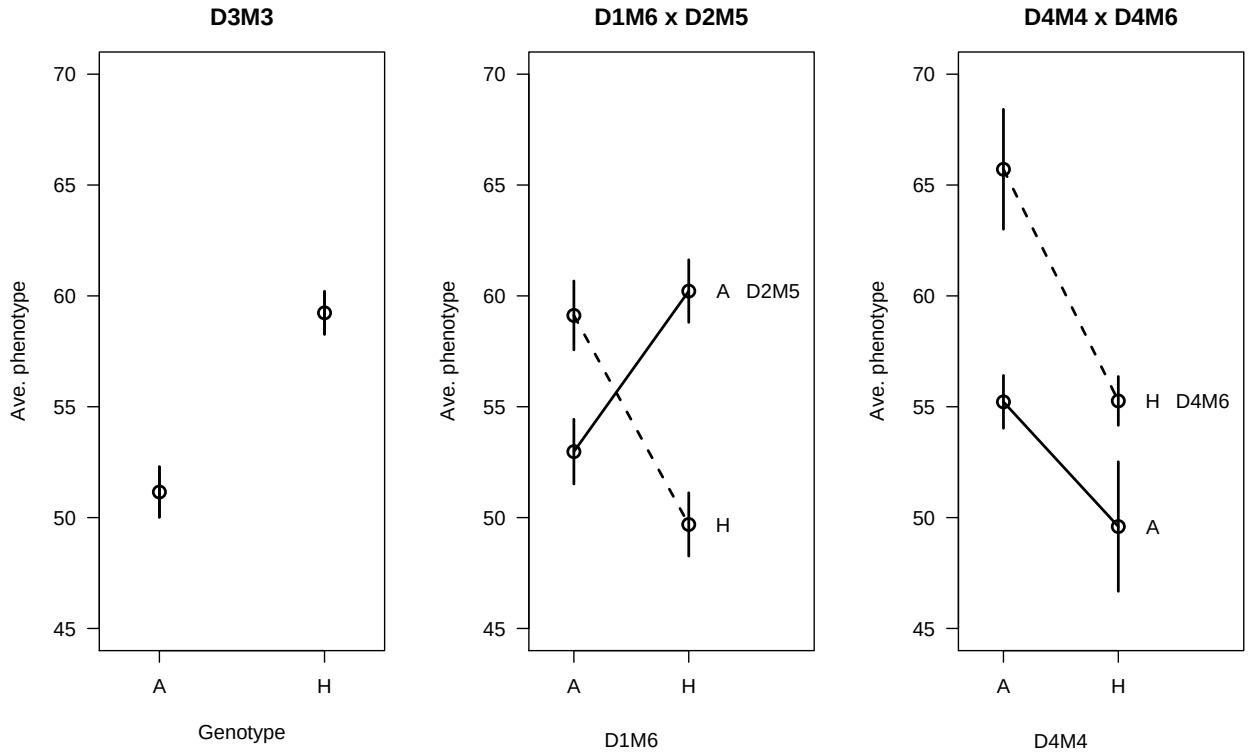
LOD_i and LOD_{fv1}



LOD_{av1} and LOD_{fv1}



QTL effects



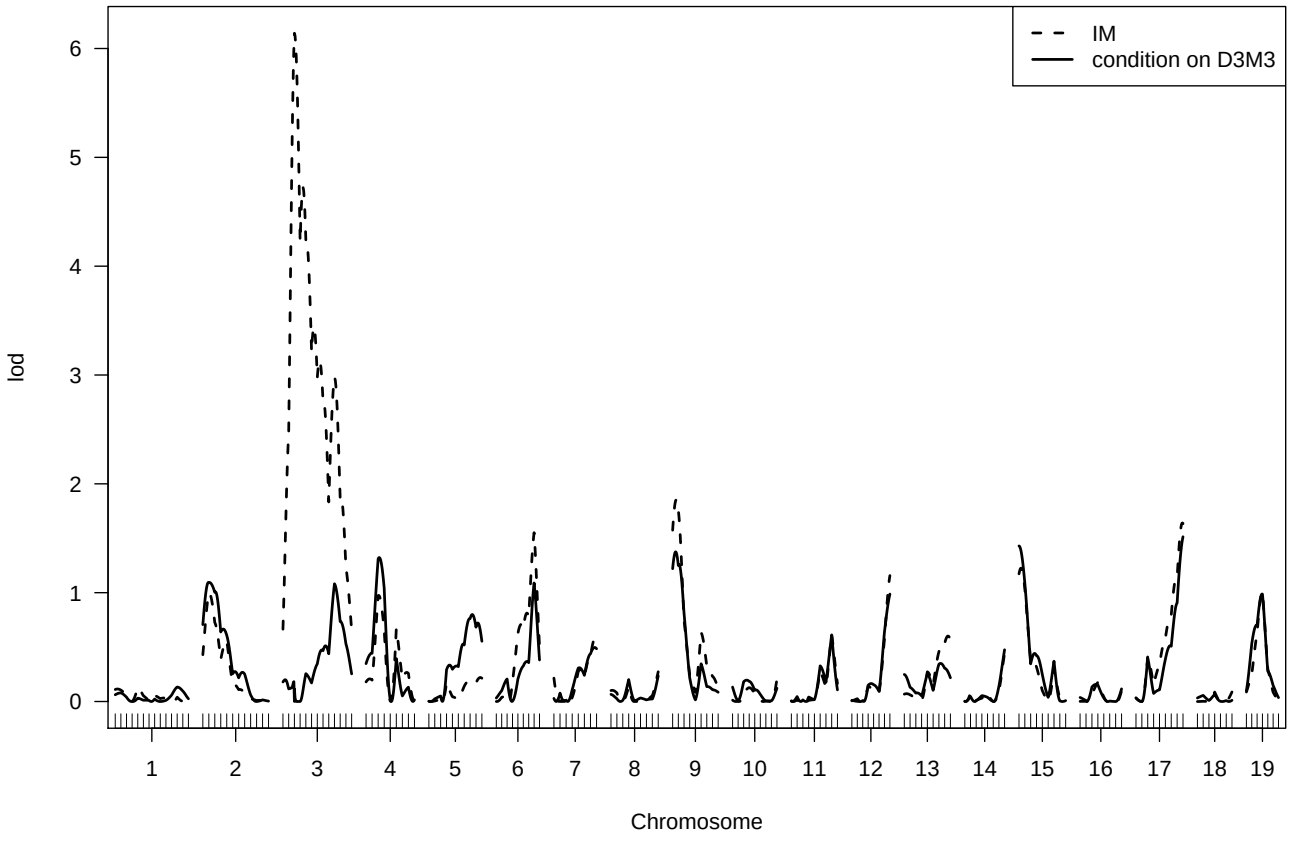
Exploratory methods

- Condition on a large-effect QTL
  - Reduce residual variation
  - Conditional LOD score:

$$\text{LOD}(q_2 \mid q_1) = \log_{10}\left\{\frac{\Pr(\text{data} \mid q_1, q_2)}{\Pr(\text{data} \mid q_1)}\right\}$$

- Piece together the putative QTL from the 1d and 2d scans
  - Omit loci that no longer look interesting (drop-one-at-a-time analysis)
  - Study potential interactions among the identified loci
  - Scan for additional loci (perhaps allowing interactions), conditional on these

Condition on D3M3



Drop-one-at-a-time

chr	pos	df	LOD	% var
1	47.5	2	7.26	10.2
2	40.0	2	7.84	11.1
3	20.0	1	6.62	9.3
4	27.5	1	4.16	5.7
4	52.5	1	2.87	3.9
1×2		1	7.17	10.1

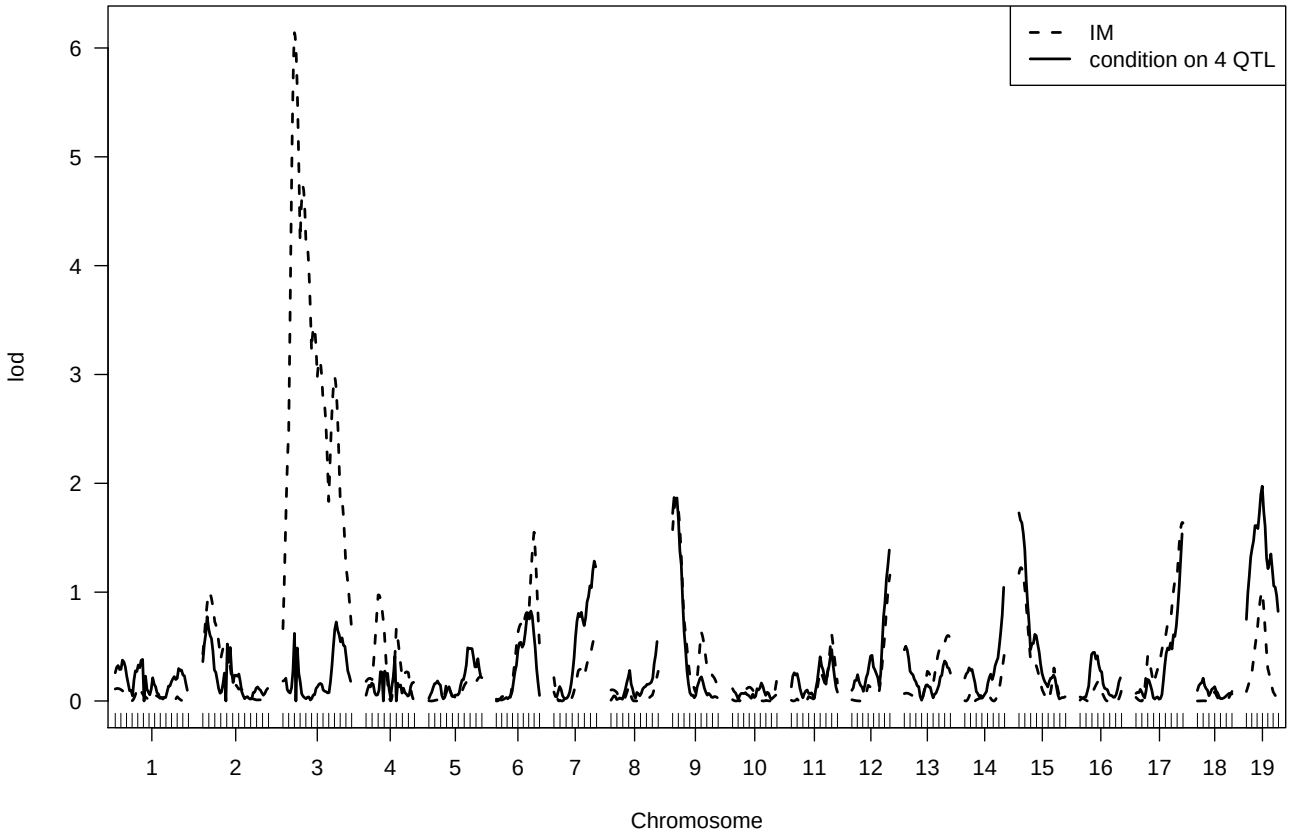
Overall: LOD = 18.2, % var = 28.5

Refined positions

chr	pos	df	LOD	% var
1	50.0	2	7.57	10.6
2	40.0	2	8.21	11.6
3	22.5	1	6.90	9.6
4	30.0	1	4.69	6.4
4	52.5	1	3.30	4.4
1×2		1	7.51	10.5

Overall: LOD = 18.8, % var = 29.2

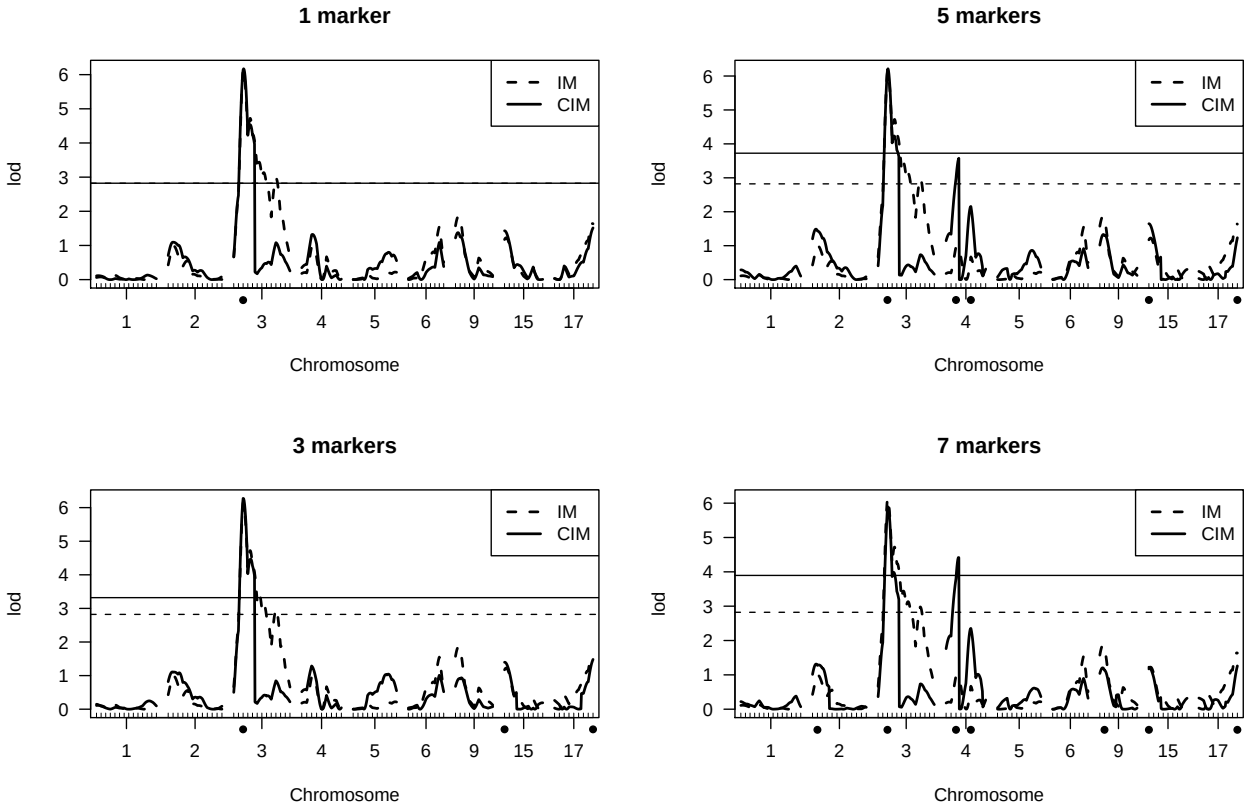
Scan for further QTL



Composite interval mapping

- Identify a set of markers, $S = \{x_1, x_2, \ldots, x_k\}$, proxies for QTL
- Scan the genome, using these markers as covariates
  - What do we do in the case of missing marker genotype data?
- At a position far from any of the marker covariates, compare
  $S \cup q$ and S
- Within some fixed window of a marker covariate, compare
  $S \setminus \{x\} \cup q$ and $S \setminus \{x\}$
- The key issue: How to select S?
  - QTL Cartographer: forward selection to some fixed number of markers

CIM results



Perfect data situation

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

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

Example model (in a backcross):

$$y_i = \mu + \beta_1 q_{i1} + \beta_2 q_{i2} + \beta_3 q_{i3} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

q_j are 0/1 variable (QTL genotypes)

μ, β's are parameters, estimated by least squares

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

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

The problem

Consider all putative QTL and QTL$\times$QTL interactions:

$$y = \mu + \sum_j \beta_j q_j + \sum_{j<k} \gamma_{jk} q_j q_k + \epsilon$$

Which $\beta_j \neq 0$?

Which $\gamma_{jk} \neq 0$?

Model selection

- Class of models
  - Additive models
  - + pairwise interactions
  - + higher-order interactions
- Model fit
  - Maximum likelihood
  - Haley-Knott regression
  - extended Haley-Knott
  - Multiple imputation
  - MCMC
- Model comparison
  - Estimated prediction error
  - AIC, BIC, penalized likelihood
  - Bayes
- Model search
  - Forward selection
  - Backward elimination
  - Stepwise selection
  - Randomized algorithms

Intercross: class of models

- 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

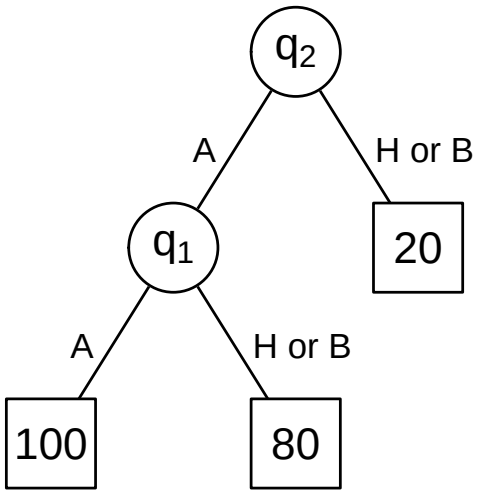
  or

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

	A	H	B
A			
H			
B			

Regression tree

		QTL 1		
		A	H	B
	A	100	80	80
QTL 2	H	20	20	20
	B	20	20	20



Model comparison

- Imagine you could fit all possible models; which would you like best?
- This issue is like LOD thresholds, but more complex.
- 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.

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

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

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

Choice of δ

Smaller δ: include more loci; higher false positive rate

Larger δ: include fewer loci; lower false positive rate

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

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

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

Note that now we have

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

$$\propto\; -(\mathrm{LOD} - \text{no. markers} \;\times \mathrm{T})$$

Model search

- 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.)

Target

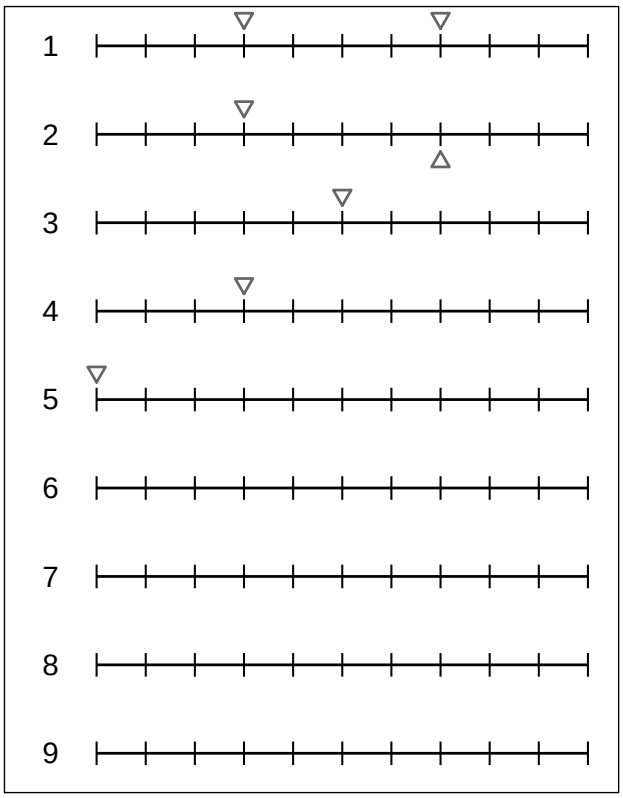
- Selection of a model includes two types of errors:
  - Miss important terms (QTLs or interactions)
  - Include extraneous terms
- Unlike in hypothesis testing, we can make both errors at the same time!
- Identify as many correct terms as possible, while controlling the rate of inclusion of extraneous 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

What is special here?

- Goal: identify the major players
- A continuum of ordinal-valued covariates (the genetic loci)
- Association among the covariates
  - Loci on different chromosomes are independent
  - Along chromosome, a very simple (and known) correlation structure

A simulation study

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

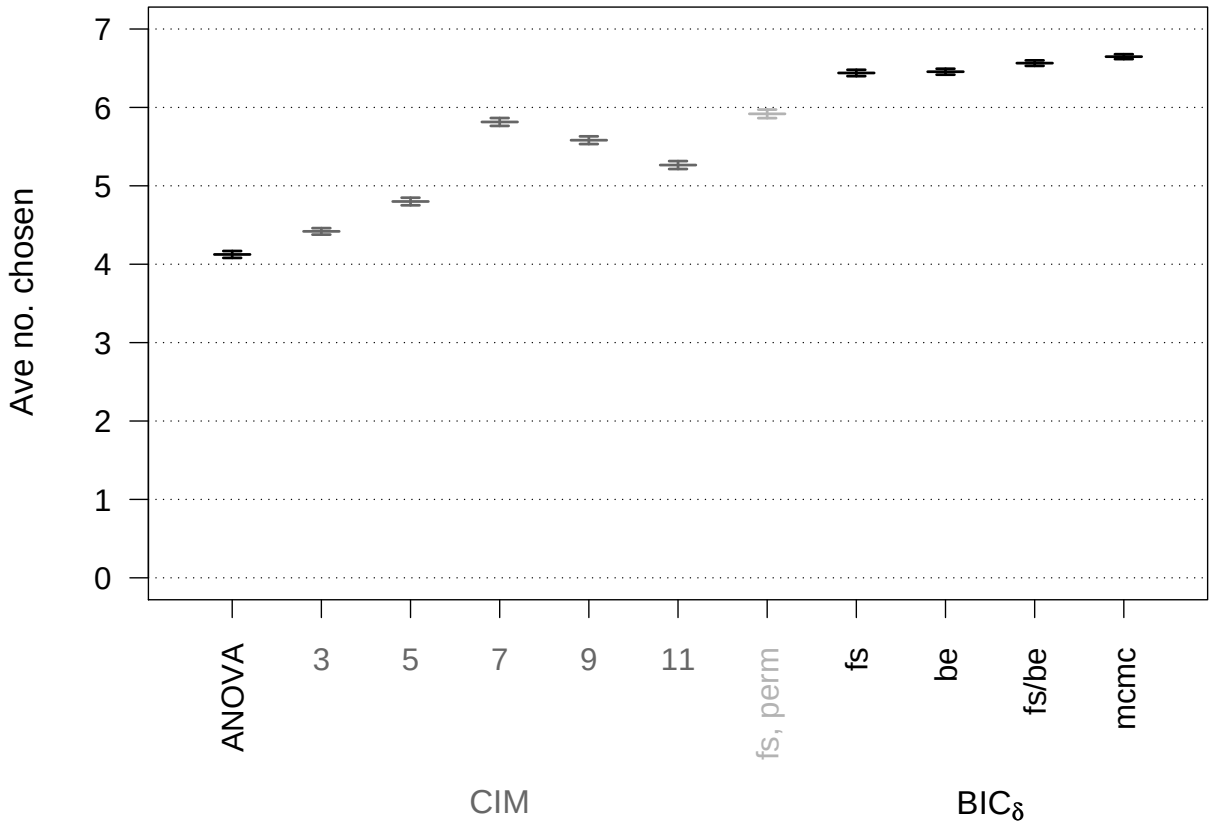


Methods

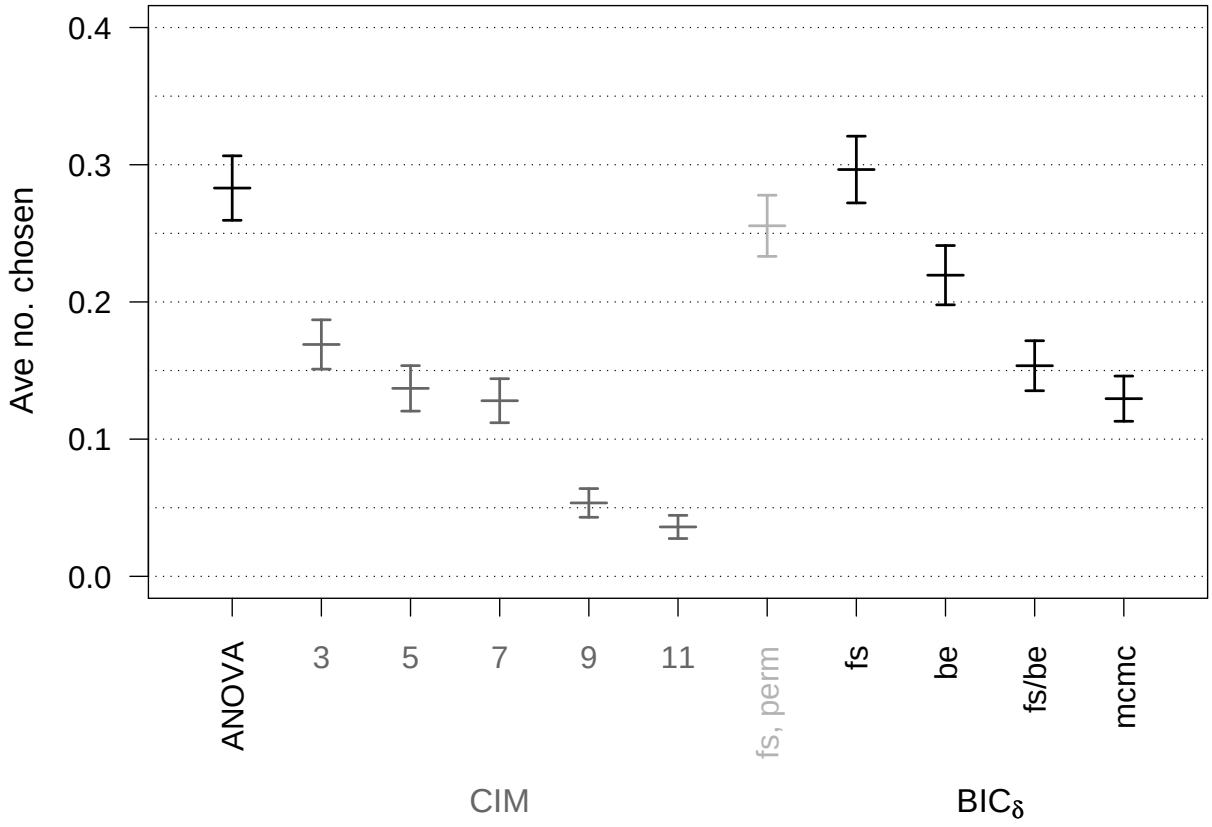
- ANOVA at marker loci
- Composite interval mapping (CIM)
- Forward selection with permutation tests
- Forward selection with BIC_δ
- Backward elimination with BIC_δ
- FS followed by BE with BIC_δ
- MCMC with BIC_δ

A selected marker was deemed correct if it was within 10 cM of a QTL (i.e. correct or adjacent).

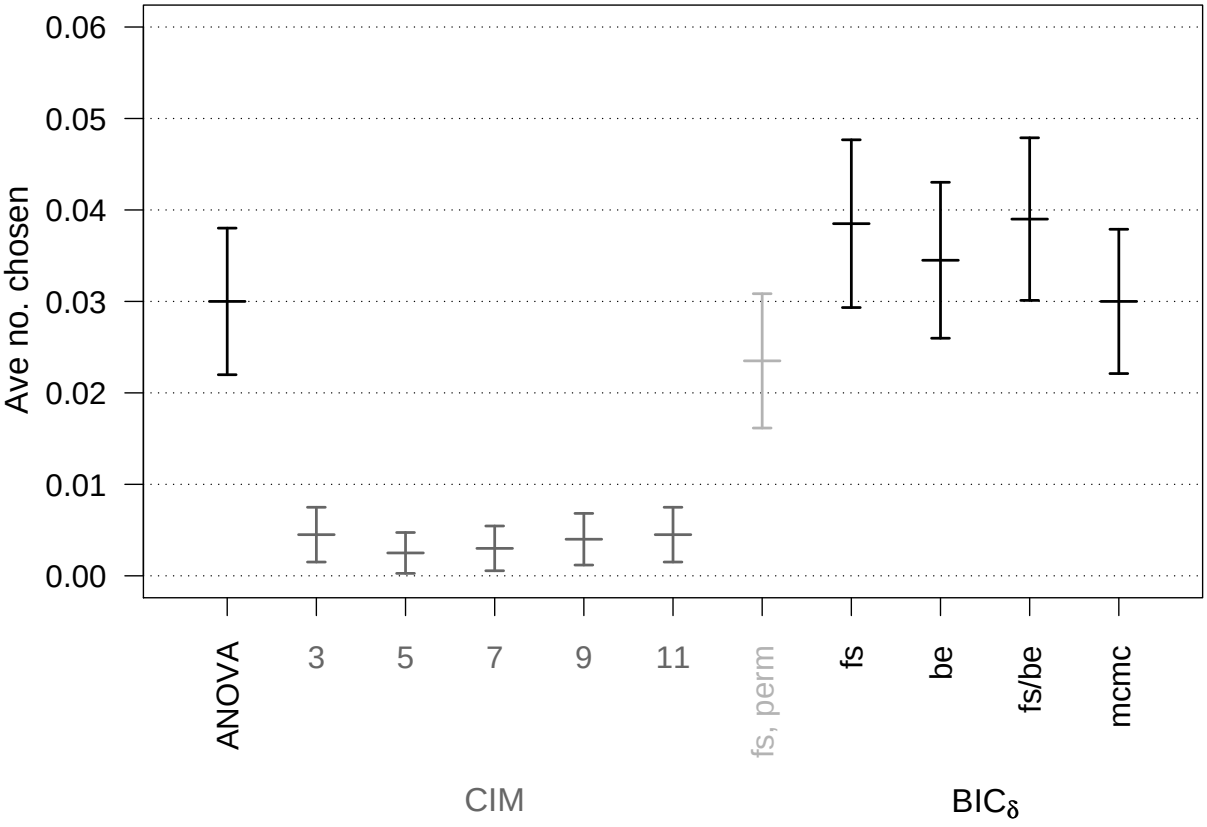
Correct



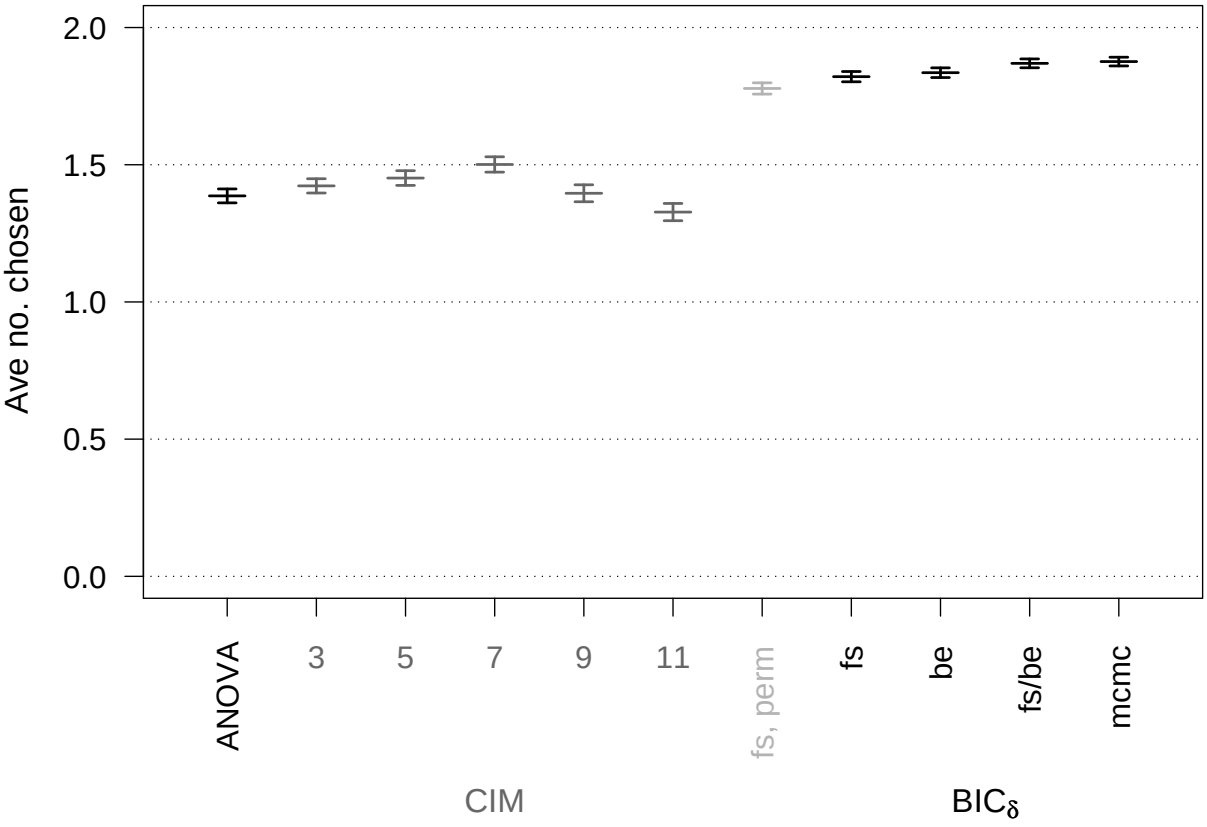
Extraneous linked



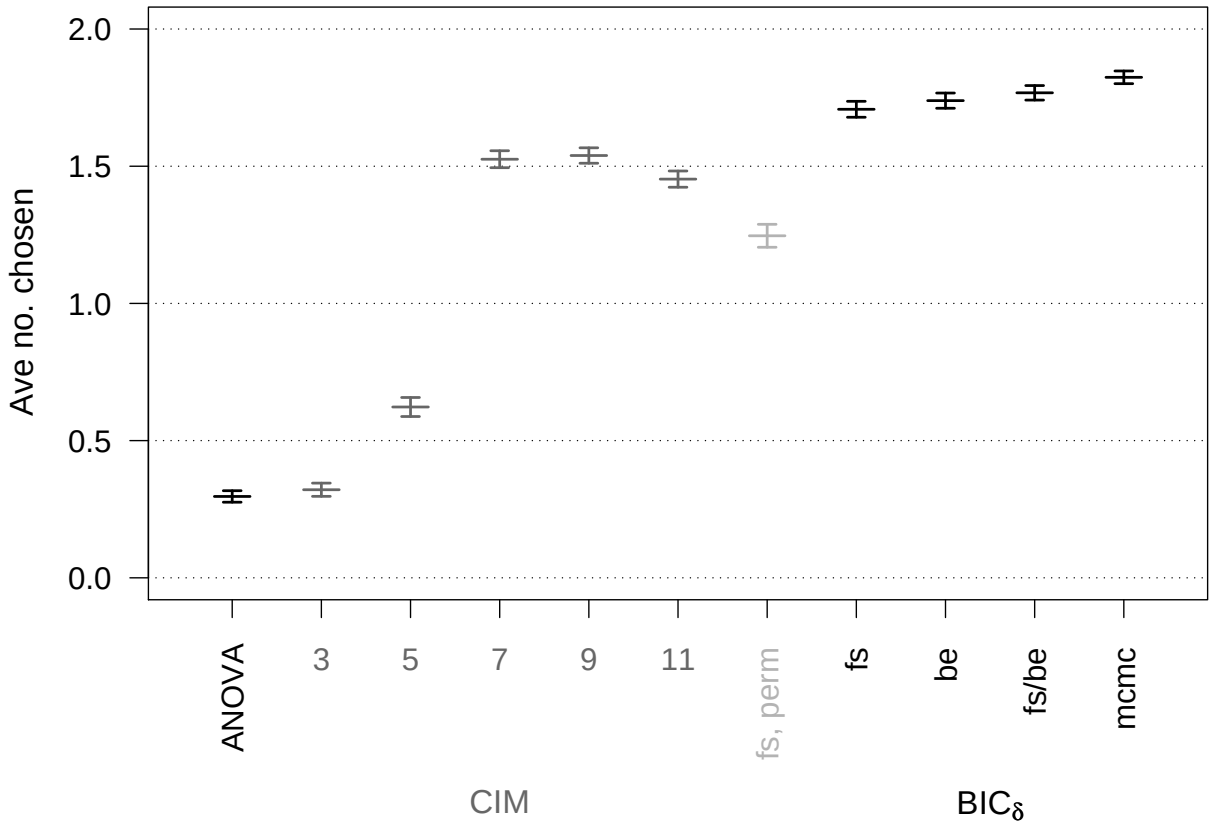
Extraneous unlinked



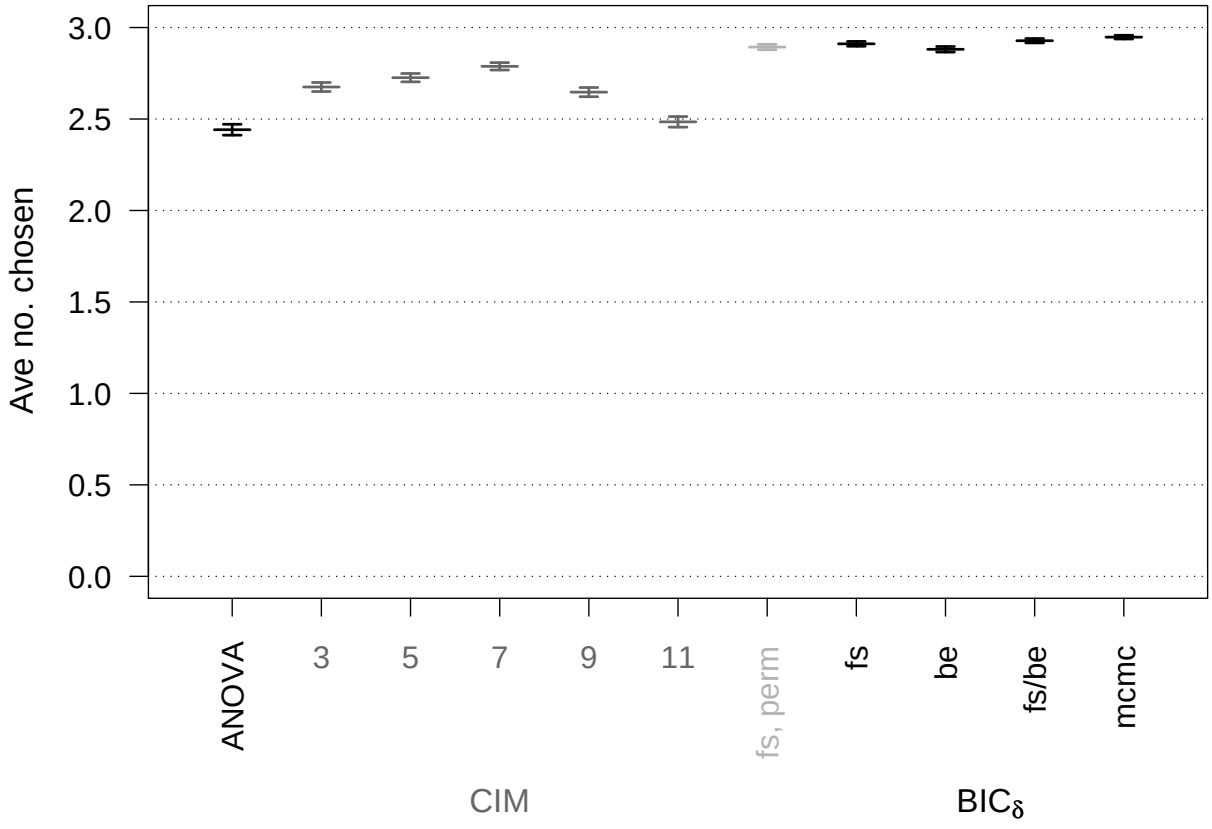
QTL in coupling



QTL in repulsion



Unlinked QTL



Epistasis

- γ = model

  $|\gamma|_m$ = no. main effects $|\gamma|_i$ = no. interactions

- Additive QTL case:

$$\text{LOD}(\gamma) - |\gamma|_m T_m$$

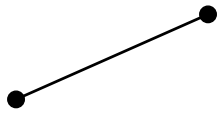
- With pairwise interactions:

$$\text{LOD}(\gamma) - |\gamma|_m T_m - |\gamma|_i T_i$$

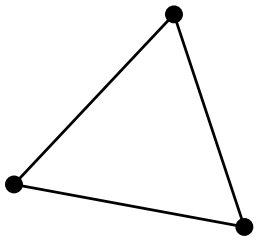
- Need a more complex penalty on interactions.

Models as graphs

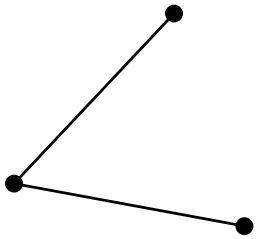
A



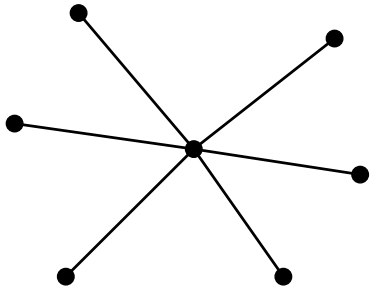
C



B



D



Bayesian methods

- The likelihood function

$$\mathsf{L}_\gamma(\theta) = \Pr(\text{data} \mid \theta, \gamma)$$

γ = model, θ = parameters

- Frequentists

$$\mathsf{L}_\gamma = \max_\theta \mathsf{L}_\gamma(\theta)$$

Penalize model complexity

- Bayesians

Prior $\Pr(\gamma)$, $\Pr(\theta \mid \gamma)$

Posterior $\Pr(\gamma \mid \text{data}) = \int \Pr(\gamma)\ \Pr(\theta \mid \gamma)\, \mathsf{L}_\gamma(\theta)\, \mathrm{d}\theta$

Summary

- QTL mapping is a model selection problem (rather than hypothesis testing).
- Model selection =
  - Select a class of models
  - Select a method for fitting models
  - Selecting a criterion for comparing models
  - Select a method of searching model space
- 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 R Statist Soc B 64:641–656, 737–775

  Contains the simulation study described above.

- Zeng Z-B, Kao C-H, Basten CJ (1999) Estimating the genetic architecture of quantitative traits. Genet Res 74:279–289

  Another paper on the model selection aspects of QTL mapping.

- Sillanpaa MJ, Corander J (2002) Model choice in gene mapping: what and why. Trends Genet 18:301–307

  A good review of model selection in QTL mapping.

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

  A good book on model selection.

- Yi N (2004) A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. Genetics 167:967–975

  A Bayeian approach for QTL mapping.

- Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D (2005) Bayesian model selection for genome-wide epistatic QTL analysis. Genetics 170:1333–1344

  A Bayesian approach for identifying interacting QTL.