

Introduction to QTL mapping in model organisms

Karl W Broman

Department of Biostatistics
Johns Hopkins University

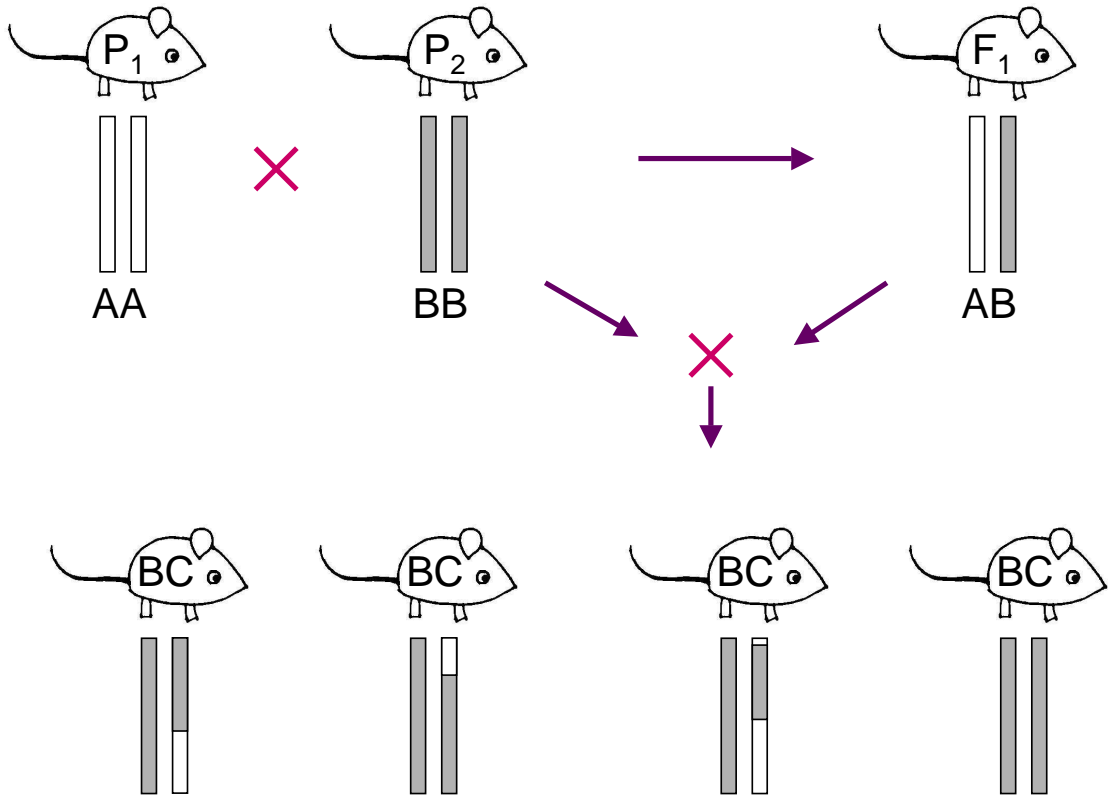
kbroman@jhsph.edu

www.biostat.jhsph.edu/~kbroman

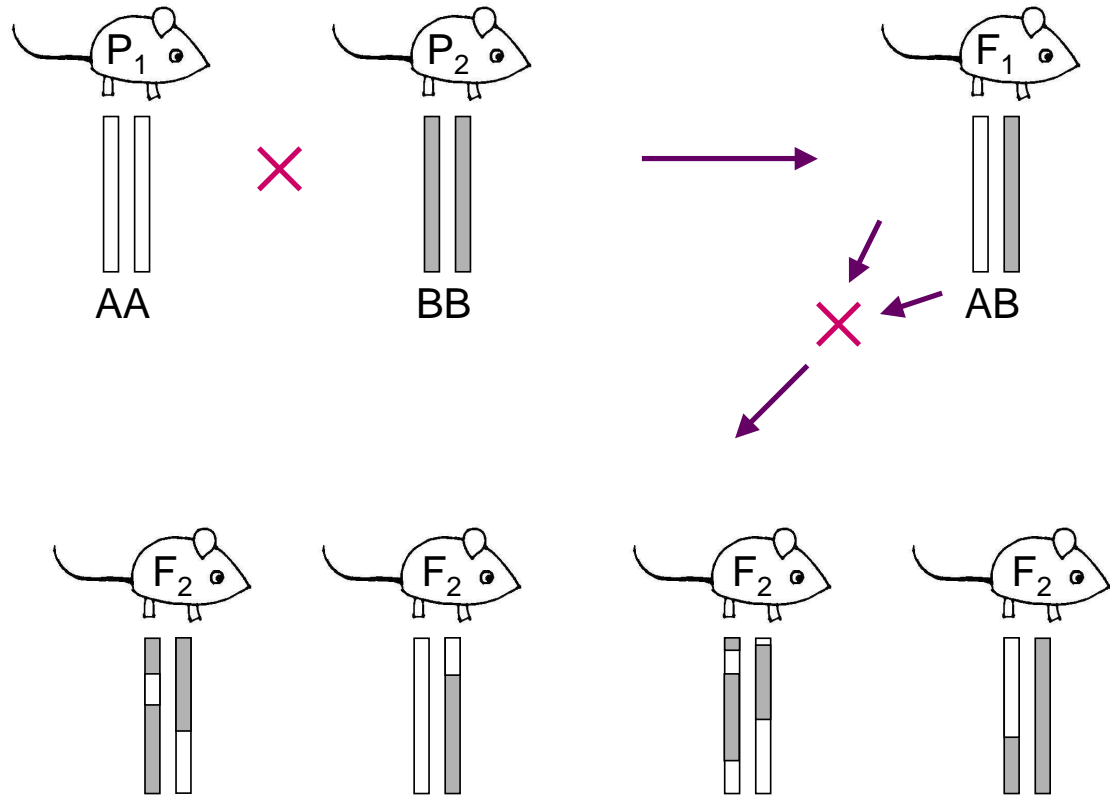
Outline

- Experiments and data
- Goals, statistical structure
- Models
- ANOVA and interval mapping
- LOD thresholds
- CIs for QTL location
- Selection bias
- Multiple QTLs, epistasis
- The X chromosome
- Selective genotyping
- Covariates
- Non-normal traits
- The need for good data

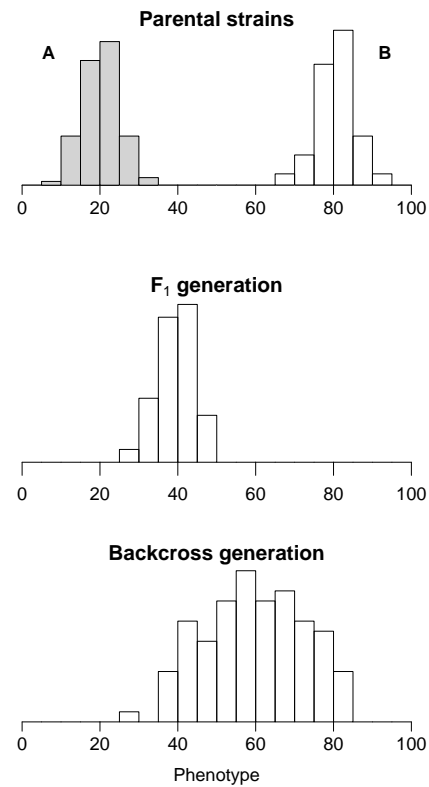
Backcross experiment



Intercross experiment



- Within each of the parental and F_1 strains, individuals are genetically identical.
- Environmental variation may or may not be constant with genotype.
- The backcross generation exhibits genetic as well as environmental variation.



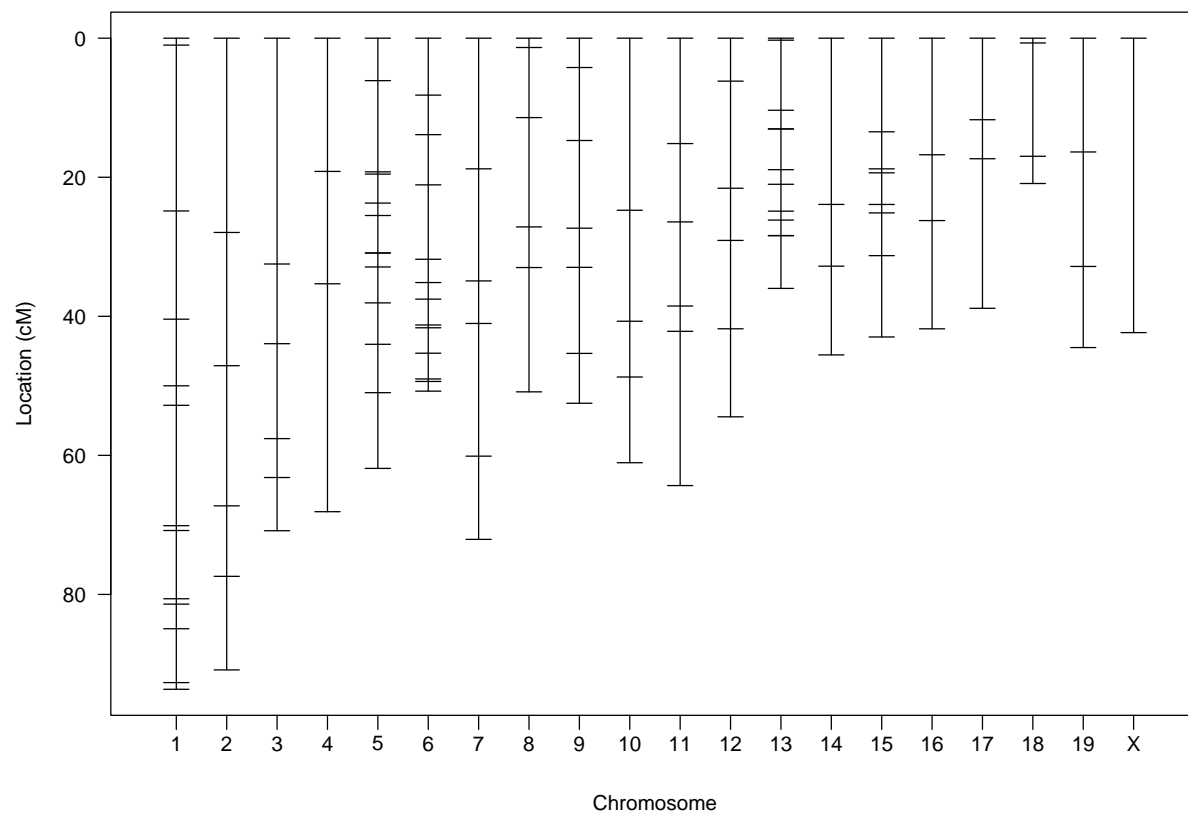
Data

Phenotypes: y_i = trait value for individual i

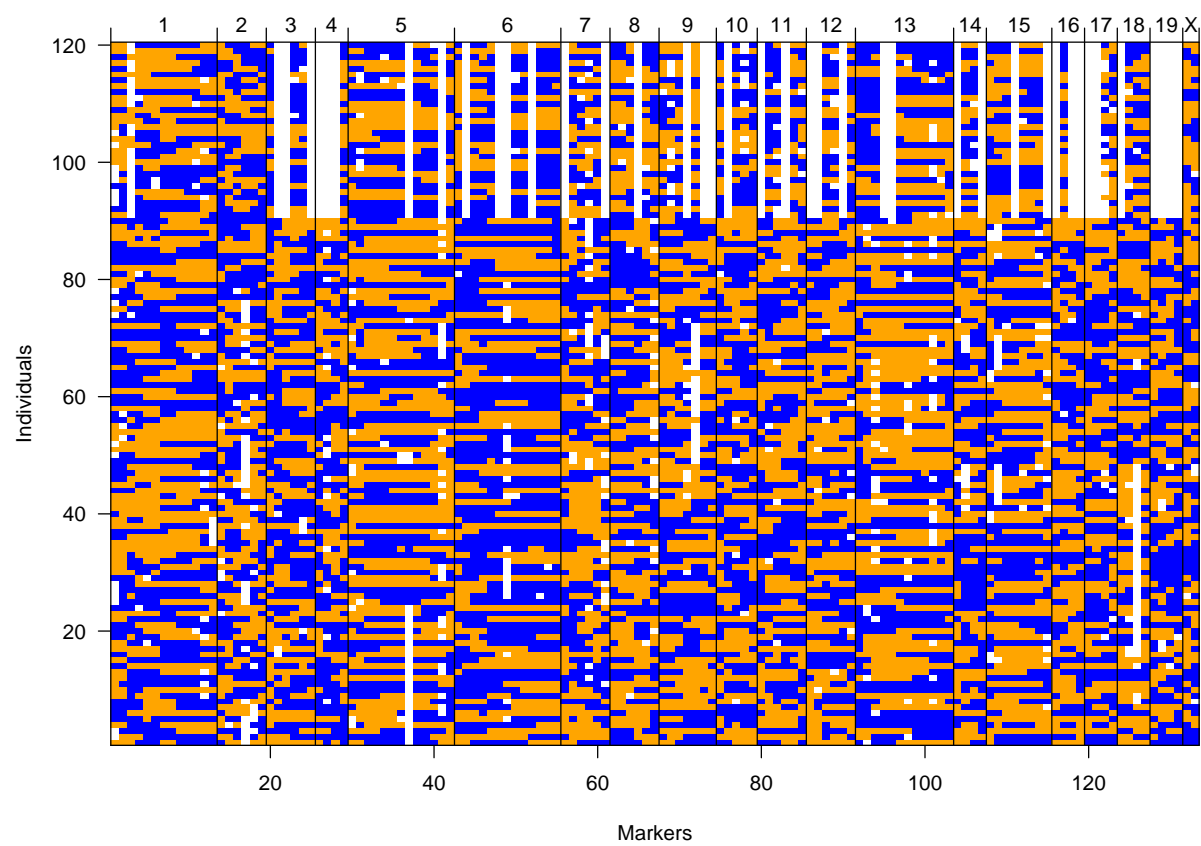
Genotypes: m_{ij} = 0/1 if mouse i is BB/AB at marker j
(or 0/1/2, in an intercross)

Genetic map: Locations of markers

Genetic map

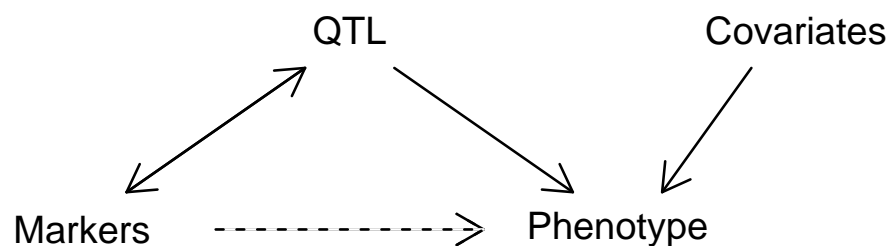


Genotype data



- Detect QTLs (and interactions between QTLs)
- Confidence intervals for QTL location
- Estimate QTL effects (effects of allelic substitution)

Statistical structure



The missing data problem:

Markers \longleftrightarrow QTL

The model selection problem:

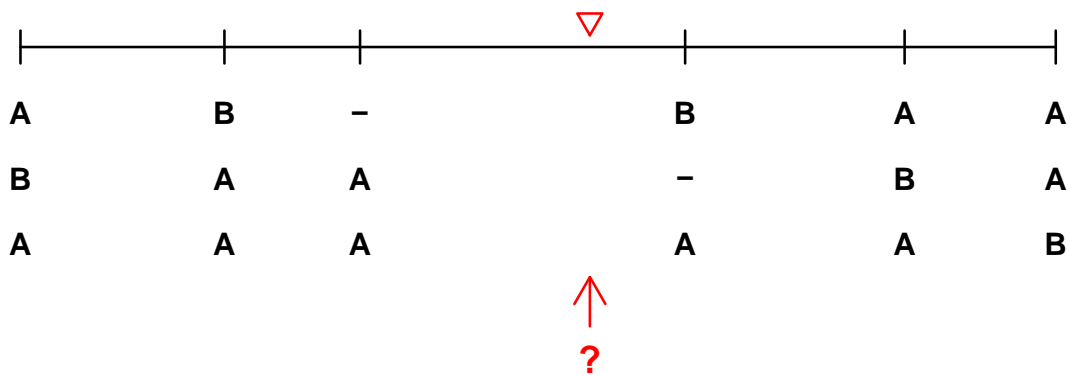
QTL, covariates \longrightarrow phenotype

We assume no crossover interference.

- ⇒ Points of exchange (crossovers) are according to a Poisson process.
- ⇒ The $\{m_{ij}\}$ (marker genotypes) form a **Markov chain**

Example

12



Models: Genotype \longleftrightarrow Phenotype

Let y = phenotype
 g = whole genome genotype

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

$$E(y|g) = \mu_{g_1, \dots, g_p} \quad \text{var}(y|g) = \sigma_{g_1, \dots, g_p}^2$$

Models: Genotype \longleftrightarrow Phenotype

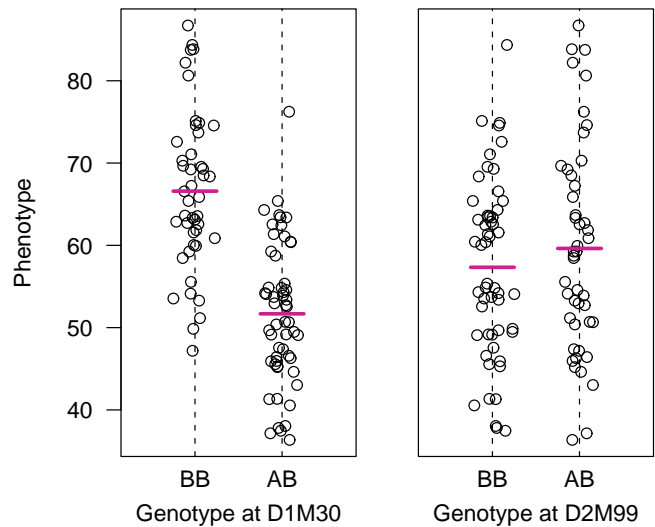
Homoscedasticity (constant variance): $\sigma_g^2 \equiv \sigma^2$

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

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

Epistasis: Any deviations from additivity.

- Also known as **marker regression**.
- 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 incorporates covariates.
- Easily extended to more complex models.
- Doesn't require a genetic map.

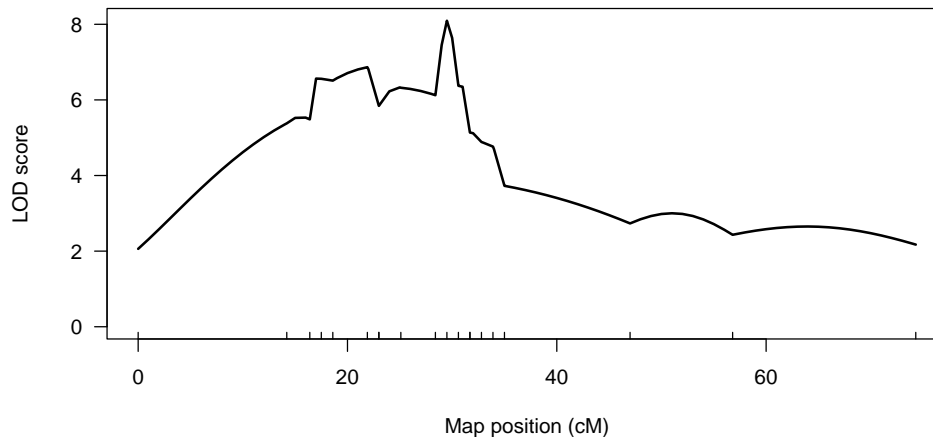
Disadvantages

- 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
- Maximum likelihood under a mixture model

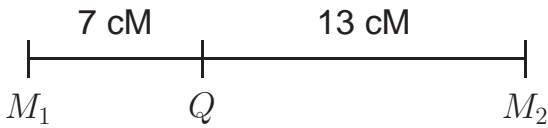


Interval mapping (IM)

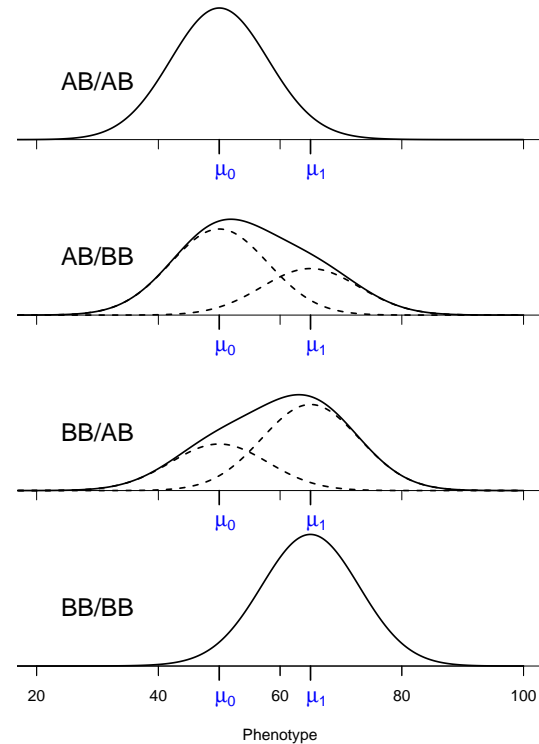
Lander & Botstein (1989)

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

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



- 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 $p_i = \Pr(q_i = 1 | \text{marker data})$

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

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

$$\text{where } f(y; \mu, \sigma) = \exp[-(y - \mu)^2 / (2\sigma^2)] / \sqrt{2\pi\sigma^2}$$

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 :

values for which $l(\mu_0, \mu_1, \sigma)$ is maximized.

Dempster et al. (1977)

E step:

$$\begin{aligned} \text{Let } w^{(k+1)} &= \Pr(z_i = 1 | y_i, \text{ marker data}, \hat{\mu}_0^{(k)}, \hat{\mu}_1^{(k)}, \hat{\sigma}^{(k)}) \\ &= \frac{p_i f(y_i; \hat{\mu}_1^{(k)}, \hat{\sigma}^{(k)})}{p_i f(y_i; \hat{\mu}_1^{(k)}, \hat{\sigma}^{(k)}) + (1-p_i) f(y_i; \hat{\mu}_0^{(k)}, \hat{\sigma}^{(k)})} \end{aligned}$$

M step:

$$\begin{aligned} \text{Let } \hat{\mu}_1^{(k+1)} &= \sum_i y_i w_i^{(k+1)} / \sum_i w_i^{(k+1)} \\ \hat{\mu}_0^{(k+1)} &= \sum_i y_i (1 - w_i^{(k+1)}) / \sum_i (1 - w_i^{(k+1)}) \\ \hat{\sigma}^{(k+1)} &= \sqrt{[\sum w_i (y - \hat{\mu}_1)^2 + \sum (1 - w_i) (y - \hat{\mu}_0)^2] / (n - 1)} \end{aligned}$$

The algorithm:

Start with $w_i^{(1)} = p_i$; iterate the E & M steps until convergence.

LOD scores

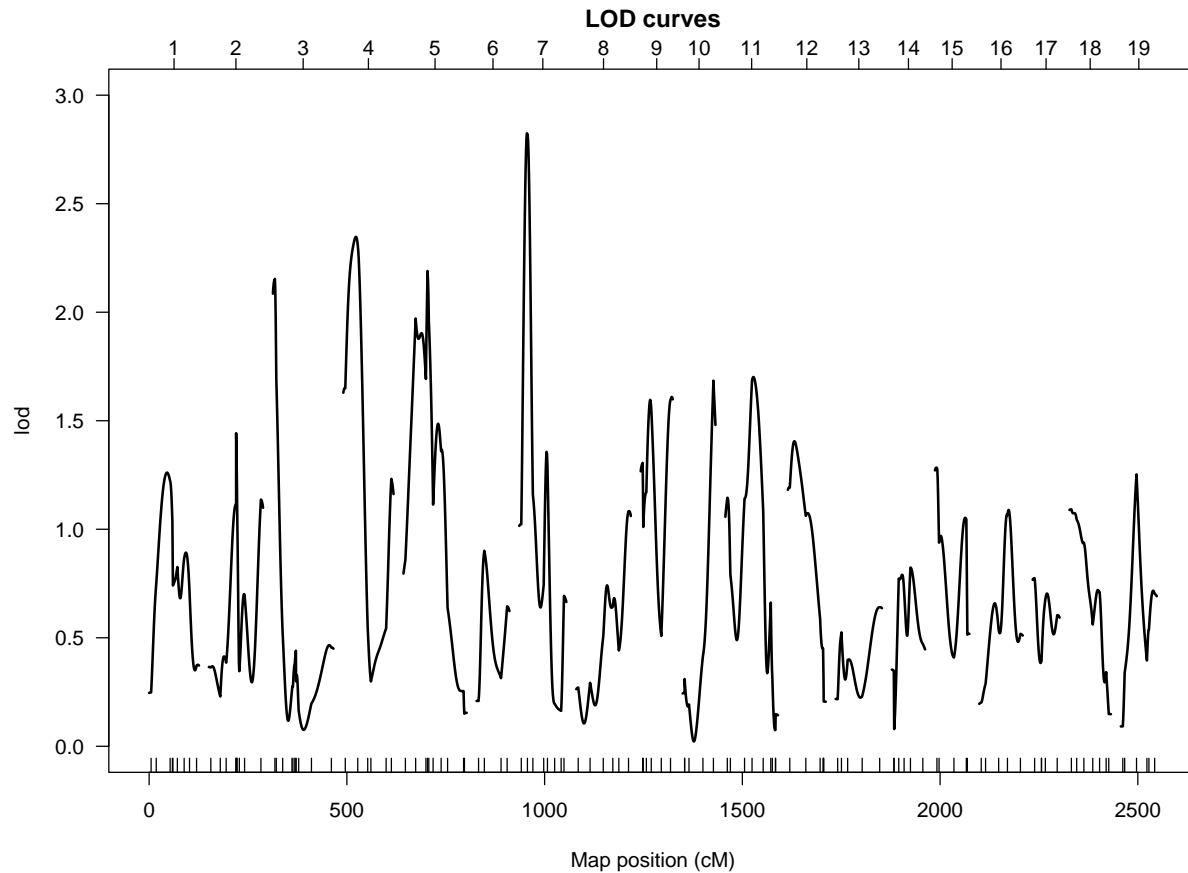
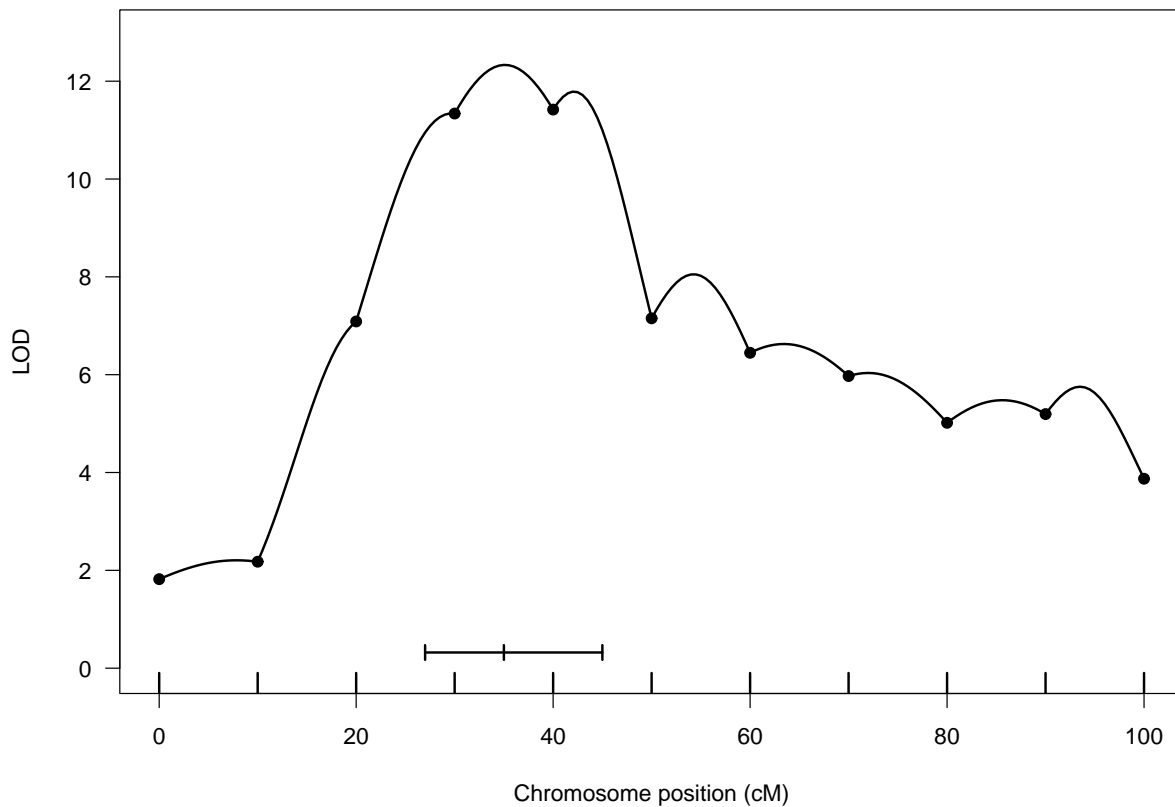
The LOD score is a measure of the **strength of evidence** for the presence of a QTL at a particular location.

$$\begin{aligned} \text{LOD}(\lambda) &= \log_{10} \text{ likelihood ratio comparing the hypothesis of a QTL at position } \lambda \text{ versus that of no QTL} \\ &= \log_{10} \left\{ \frac{\Pr(y | \text{QTL at } \lambda, \hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}, \hat{\sigma}_\lambda)}{\Pr(y | \text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\} \end{aligned}$$

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

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.**

Haley-Knott regression

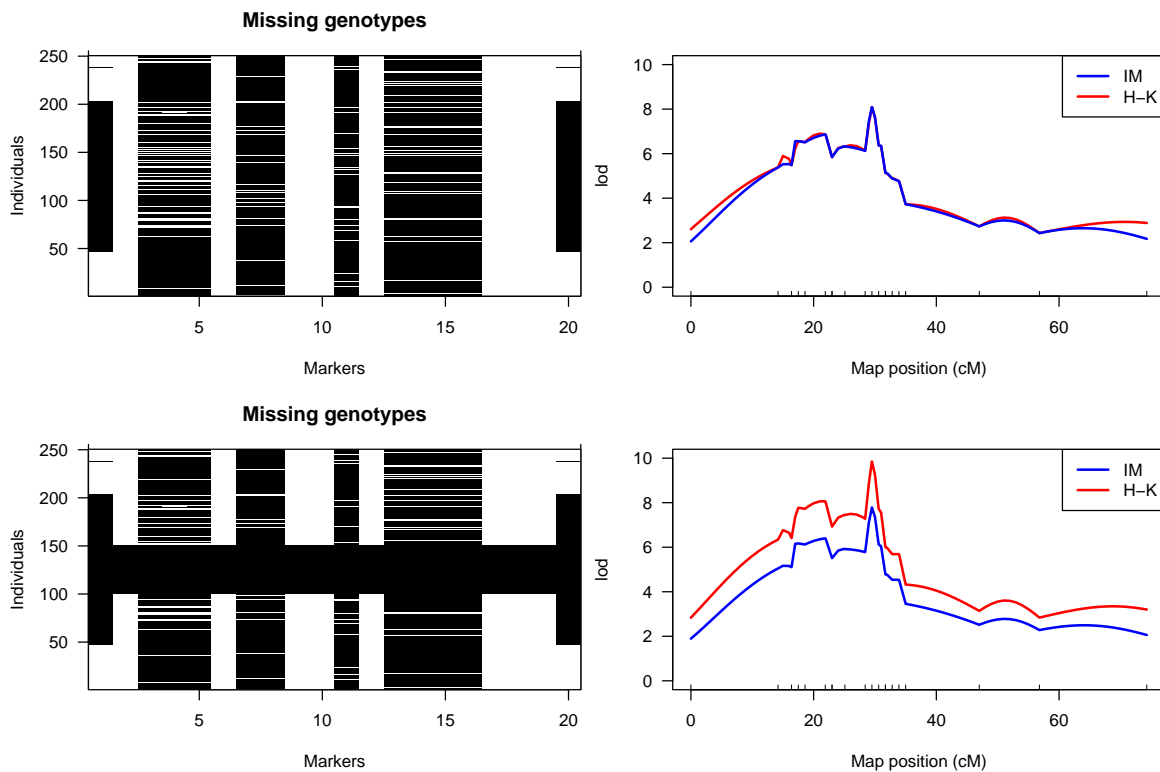
A quick approximation to Interval Mapping.

$$E(y \mid \text{QTL} = q) = \mu_0 + (\mu_1 - \mu_0) \mathbf{1}\{q = \text{AB}\}$$

$$E(y \mid \text{marker data}) = \mu_0 + (\mu_1 - \mu_0) \Pr(\text{QTL} = \text{AB} \mid \text{marker data})$$

- Regress y on $\Pr(\text{QTL} = \text{AB} \mid \text{marker data})$.
- **Pretend** that the residual variation is normally distributed.
- Calculate

$$\text{LOD}(\lambda) = (n/2) \log_{10} \left\{ \frac{\text{RSS}_0}{\text{RSS}_a(\lambda)} \right\}$$



LOD thresholds

Large LOD scores indicate evidence for the presence of a QTL

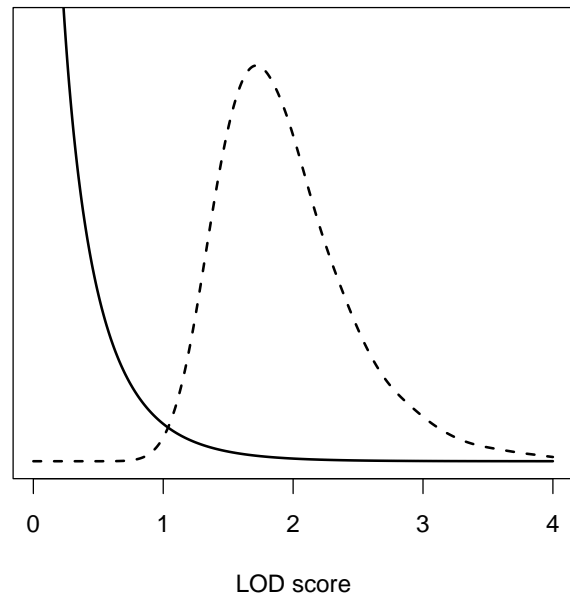
Question: How large is large?

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

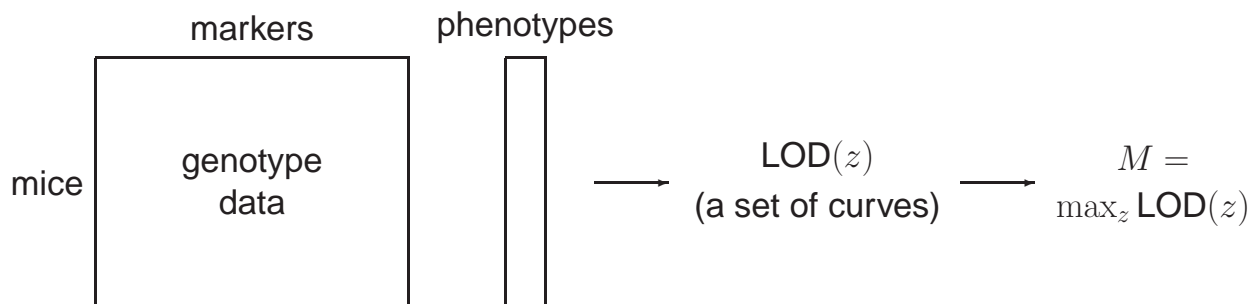
- Derivation:
- Analytical calculations (L & B 1989)
 - Simulations (L & B 1989)
 - Permutation tests (Churchill & Doerge 1994)

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.



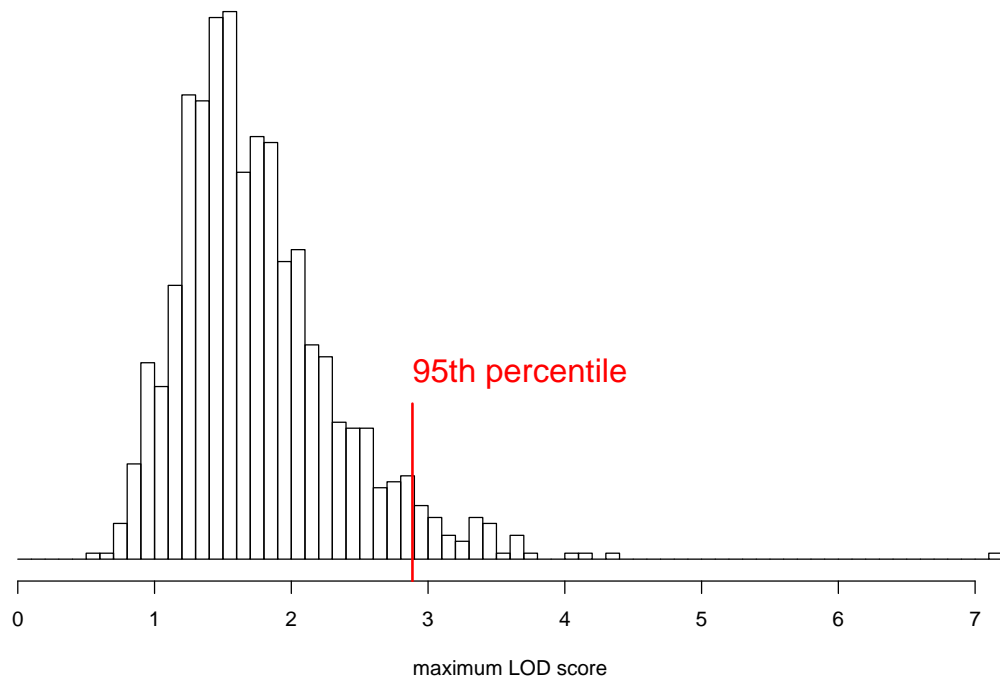
Permutation tests



- Permute/shuffle the phenotypes; keep the genotype data intact.
- Calculate $\text{LOD}^*(z) \rightarrow M^* = \max_z \text{LOD}^*(z)$
- We wish to compare the observed M to the distribution of M^* .
- $\Pr(M^* \geq M)$ is a genome-wide P-value.
- The 95th %ile of M^* 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.

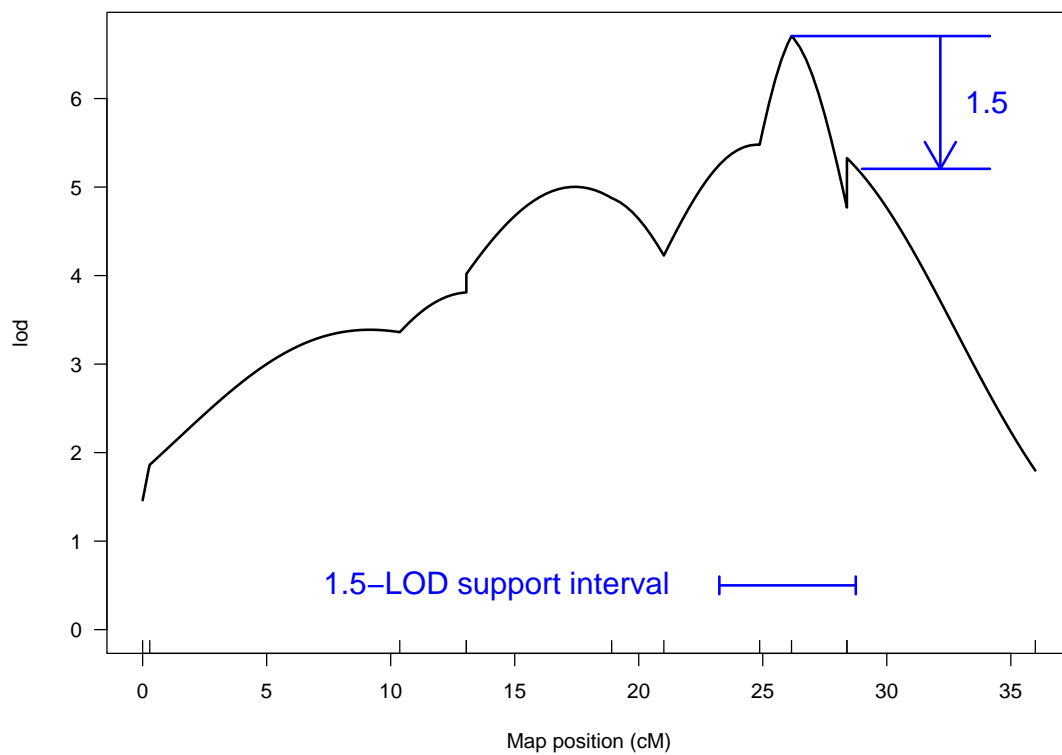
Permutation distribution

31

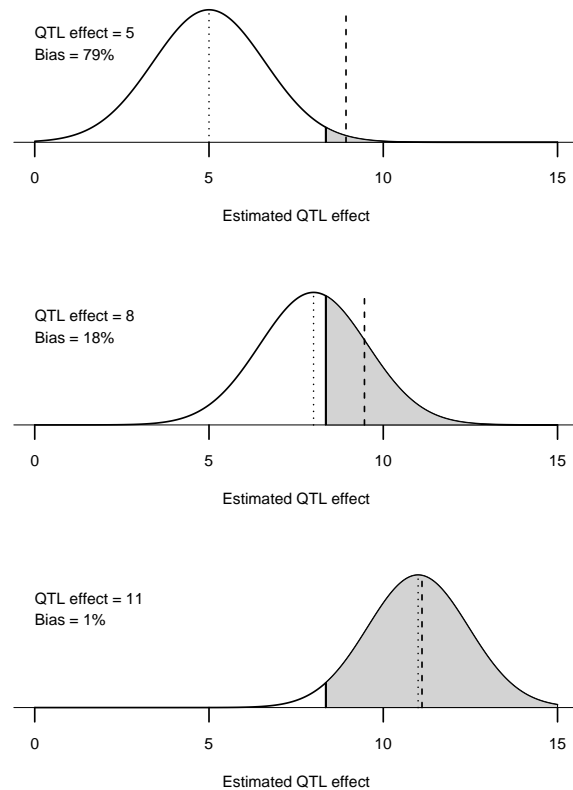


1.5-LOD support interval

32



- The estimated effect of a QTL will vary somewhat from its true effect.
- Only when the estimated effect is large will the QTL be detected.
- Among those experiments in which the QTL is detected, the estimated QTL effect will be, on average, larger than its true effect.
- This is **selection bias**.
- Selection bias is largest in QTLs with small or moderate effects.
- The true effects of QTLs that we identify are likely smaller than was observed.



Implications of selection bias

- Estimated % variance explained by identified QTLs
- Repeating an experiment
- Congenics
- Marker-assisted selection

Why consider multiple QTLs at once?

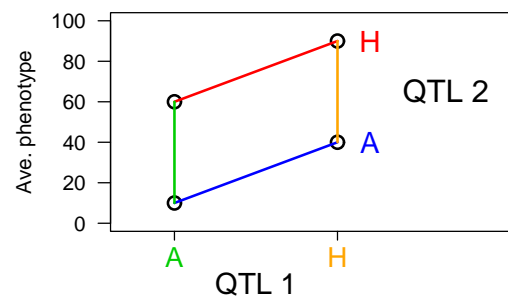
- Reduce residual variation.
- Separate linked QTLs.
- Investigate interactions between QTLs (epistasis).

Issues:

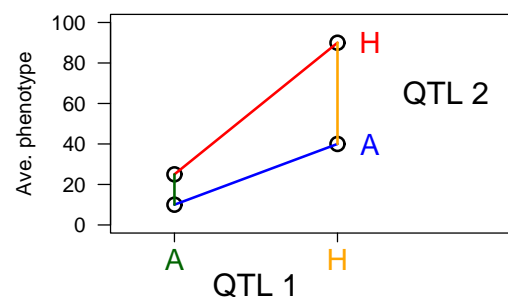
- Missing genotype information
- The model selection problem

Epistasis in a backcross

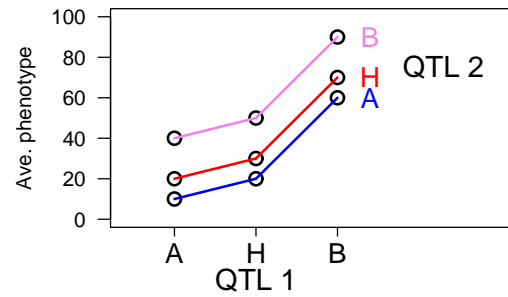
Additive QTLs



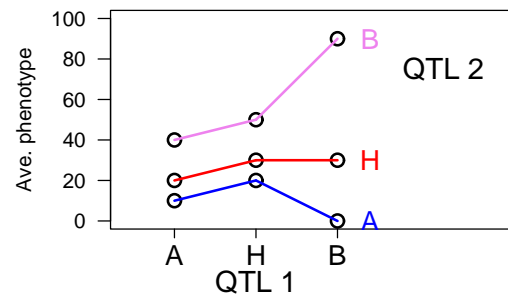
Interacting QTLs



Additive QTLs



Interacting QTLs



Two-dimensional genome scan

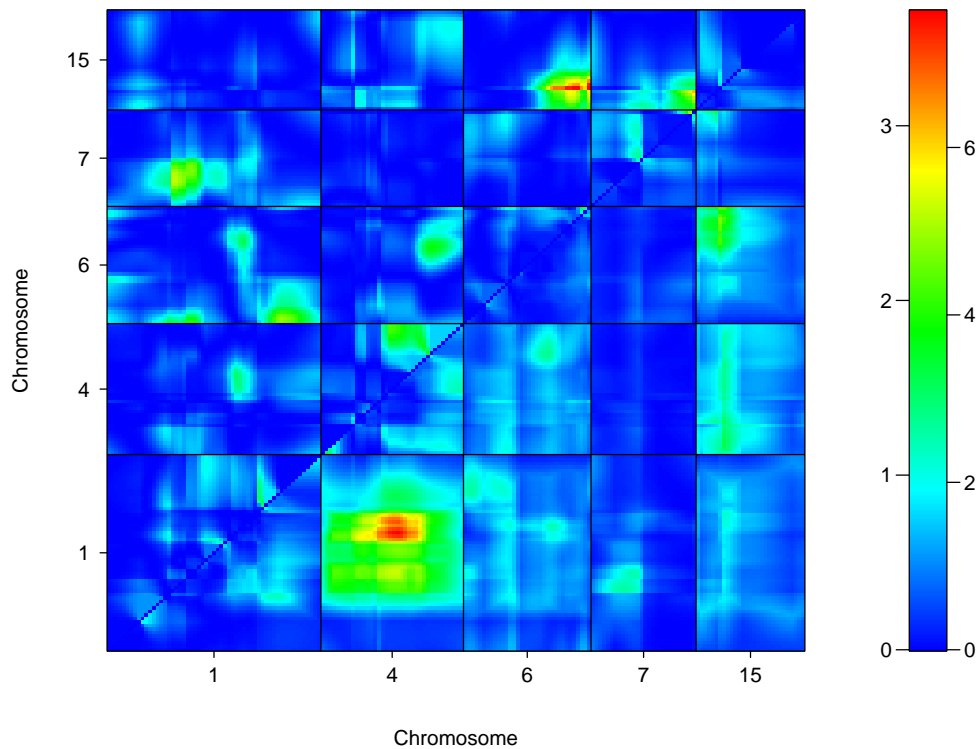
Consider each pair of positions, (γ_1, γ_2)

Models

- Full
- Additive
- QTL 1
- QTL 2
- Null

Possible comparisons

- Full vs. null
- Full vs. additive
- Full vs. Best of QTL 1 & 2
- Add've vs. Best of QTL 1 & 2



The X chromosome

In a backcross, the X chromosome may or may not be segregating.

$$(A \times B) \times A$$

Females: $X_{A \cdot B} X_A$

Males: $X_{A \cdot B} Y_A$

$$A \times (A \times B)$$

Females: $X_A X_A$

Males: $X_A Y_B$

The X chromosome

In an intercross, one must pay attention to the **paternal grandmother's genotype**.

$$(A \times B) \times (A \times B) \quad \text{or} \quad (B \times A) \times (A \times B)$$

Females: $X_{A \cdot B} X_A$

Males: $X_{A \cdot B} Y_B$

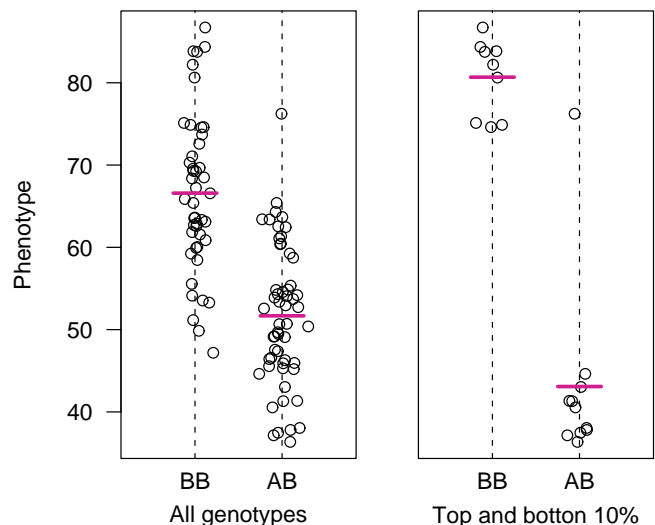
$$(A \times B) \times (B \times A) \quad \text{or} \quad (B \times A) \times (B \times A)$$

Females: $X_{A \cdot B} X_B$

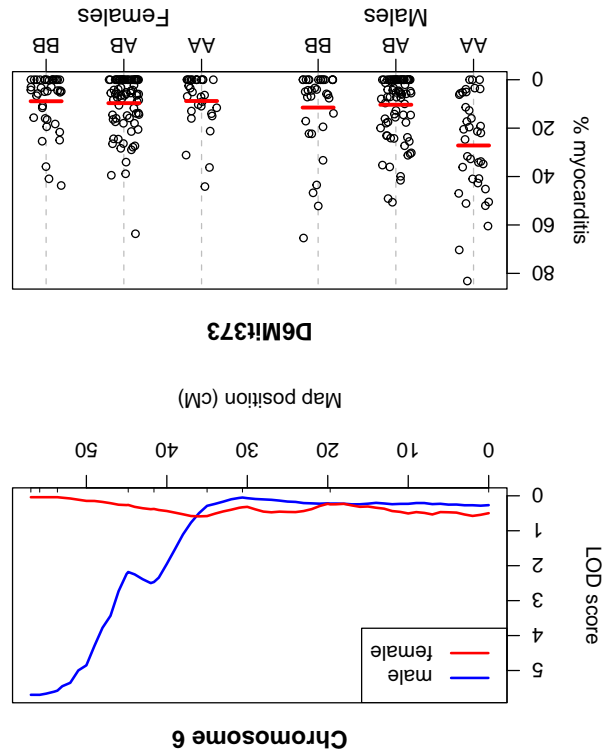
Males: $X_{A \cdot B} Y_A$

Selective genotyping

- Save effort by only typing the most informative individuals (say, top & bottom 10%).
- Useful in context of a **single, inexpensive** trait.
- Tricky to estimate the effects of QTLs: use IM with **all** phenotypes.
- Can't get at interactions.
- Likely better to also genotype some random portion of the rest of the individuals.



- **Examples:** treatment, sex, litter, lab, age.
- Control residual variation.
- Avoid confounding.
- Look for QTL \times environ't interactions
- Adjust before interval mapping (IM) versus adjust within IM.



Non-normal traits

- Standard interval mapping assumes normally distributed residual variation. (Thus the phenotype distribution is a mixture of normals.)
- **In reality:** we see dichotomous traits, counts, skewed distributions, outliers, and all sorts of odd things.
- Interval mapping, with LOD thresholds derived from permutation tests, generally performs just fine anyway.
- Alternatives to consider:
 - Nonparametric approaches (Kruglyak & Lander 1995)
 - Transformations (e.g., log, square root)
 - Specially-tailored models (e.g., a generalized linear model, the Cox proportional hazard model, and the model in Broman et al. 2000)

The success of QTL mapping depends crucially on the integrity of the data.

- Segregation distortion
- Genetic maps / marker positions
- Genotyping errors (tight double crossovers)
- Phenotype distribution / outliers
- Residual analysis

Summary I

- **ANOVA** at marker loci (aka marker regression) is simple and easily extended to include covariates or accommodate complex models.
- **Interval mapping** improves on ANOVA by allowing inference of QTLs to positions between markers and taking proper account of missing genotype data.
- ANOVA and IM consider only single-QTL models. **Multiple QTL methods** allow the better separation of linked QTLs and are necessary for the investigation of epistasis.
- Statistical significance of LOD peaks requires consideration of the maximum LOD score, genome-wide, under the null hypothesis of no QTLs. **Permutation tests** are extremely useful for this.
- **1.5-LOD support intervals** indicate the plausible location of a QTL.
- Estimates of QTL effects are subject to **selection bias**. Such estimated effects are often too large.

Summary II

- The **X chromosome** must be dealt with specially, and can be tricky.
- **Study your data**. Look for errors in the genetic map, genotyping errors and phenotype outliers. But don't worry about them too much.
- **Selective genotyping** can save you time and money, but proceed with caution.
- **Study your data**. The consideration of covariates may reveal extremely interesting phenomena.
- Interval mapping works reasonably well even with **non-normal traits**. But consider transformations or specially-tailored models. If interval mapping software is not available for your preferred model, start with some version of ANOVA.

References

- Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. *Lab Animal* 30:44–52
A review for non-statisticians.
- Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* 3:43–52
A very recent review.
- Doerge RW, Zeng Z-B, Weir BS (1997) Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science* 12:195–219
Review paper.
- Jansen RC (2001) Quantitative trait loci in inbred lines. In Balding DJ et al., *Handbook of statistical genetics*, John Wiley & Sons, New York, chapter 21
Review in an expensive but rather comprehensive and likely useful book.
- Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland, MA, chapter 15
Chapter on QTL mapping.
- 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.
- Kruglyak L, Lander ES (1995) A nonparametric approach for mapping quantitative trait loci. *Genetics* 139:1421–1428
Non-parameteric interval mapping.
- Broman KW (2003) Mapping quantitative trait loci in the case of a spike in the phenotype distribution. *Genetics* 163:1169–1175
QTL mapping with a special model for a non-normal phenotype.
- Miller AJ (2002) *Subset selection in regression*, 2nd edition. Chapman & Hall, New York.
A 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.