R/qtl workshop

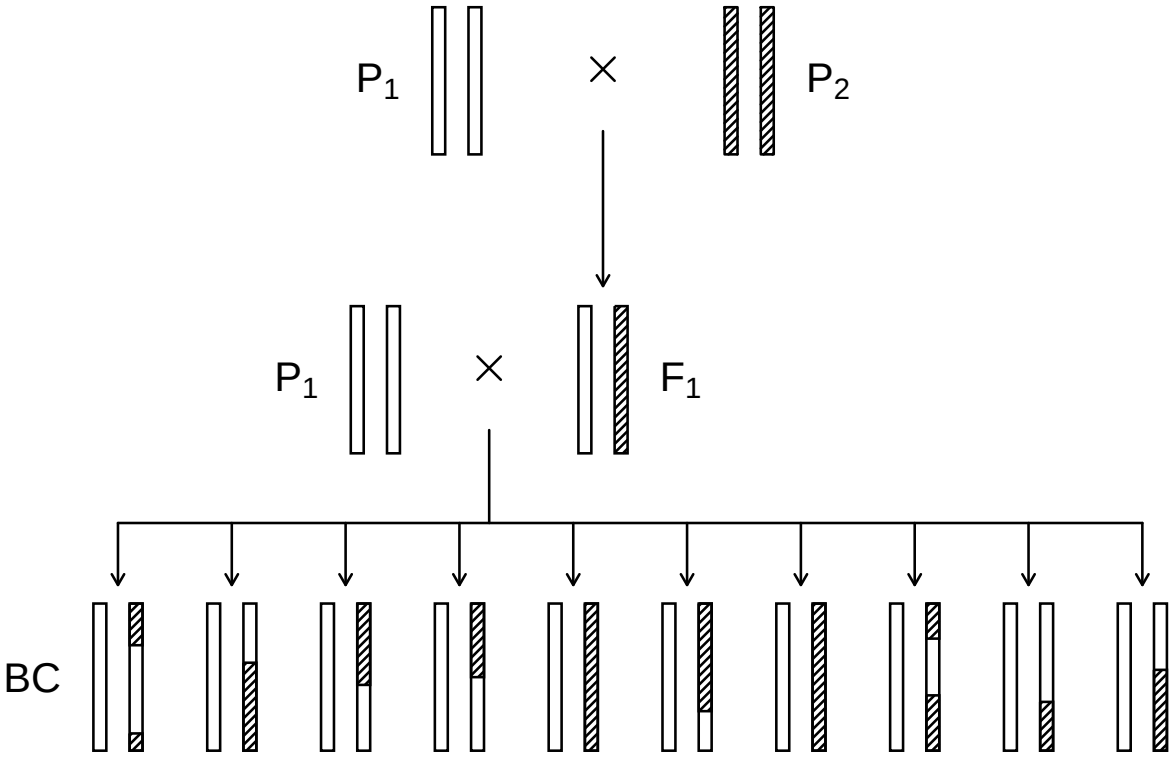
(part 1)

Karl Broman

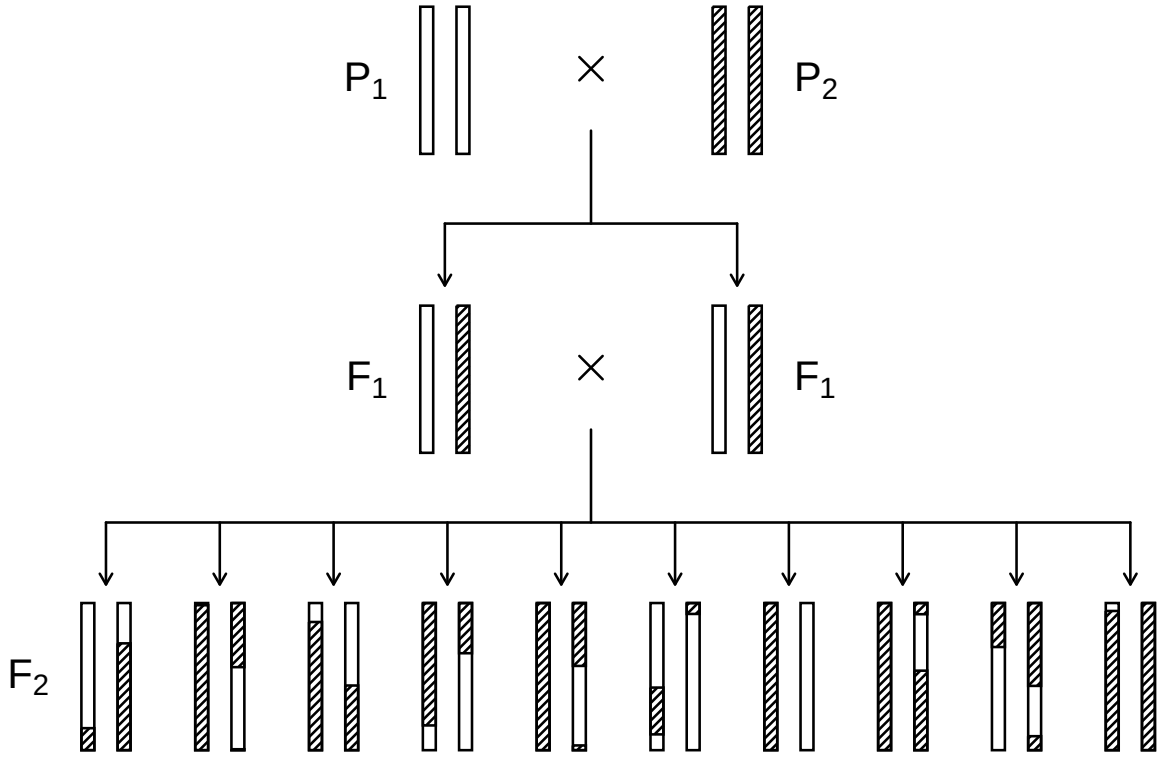
Biostatistics and Medical Informatics
University of Wisconsin – Madison

kbroman.org
github.com/kbroman
@kwbroman

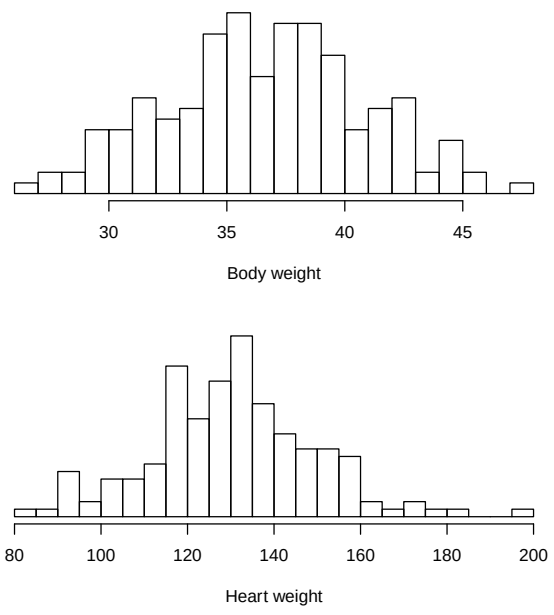
Backcross

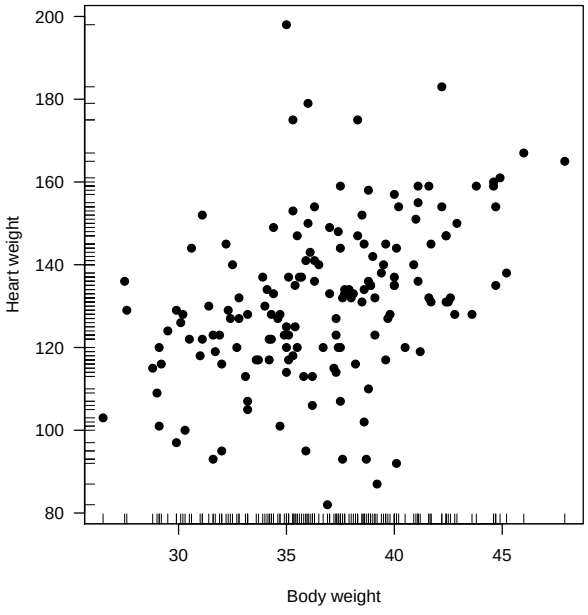


Intercross



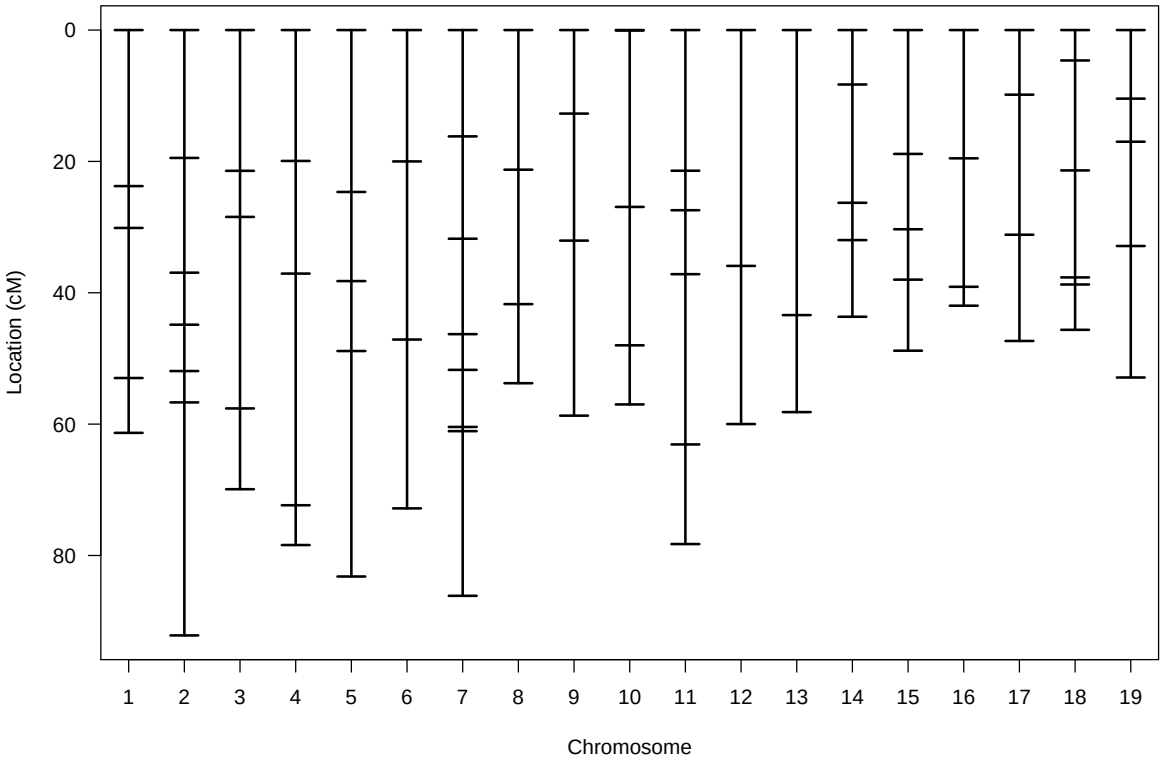
Phenotype data



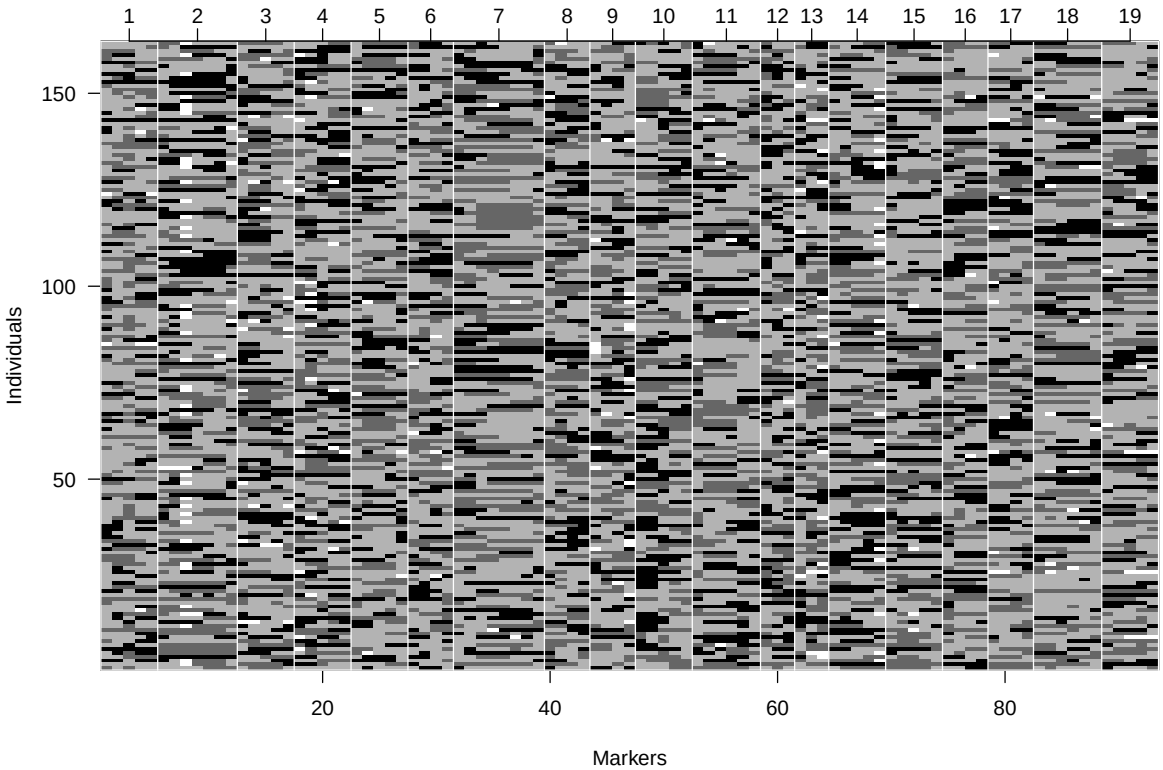


Sugiyama et al. (2002) Physiol Genomics 10:5–12

Genetic map



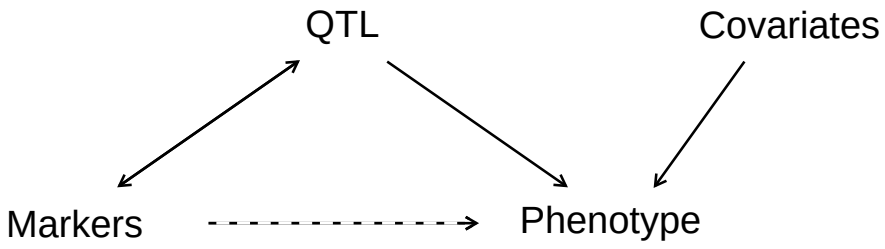
Genotype data



Goals

- Identify quantitative trait loci (QTL)
  (and interactions among QTL)
- Interval estimates of QTL location
- Estimated QTL effects

Statistical structure

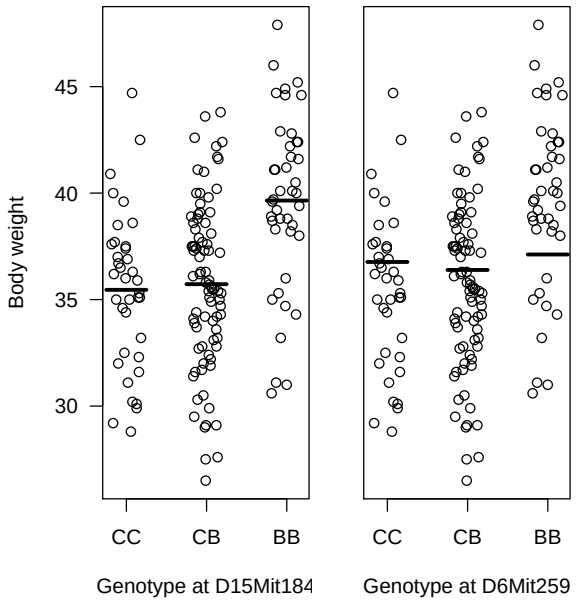


The missing data problem: Markers $\longleftrightarrow$ QTL

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

ANOVA at marker loci

- 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

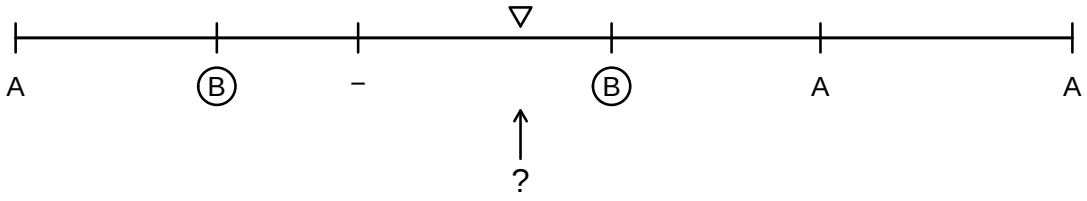
Lander & Botstein (1989)

- Assume a single QTL model.
- Each position in the genome, one at a time, is posited as the putative QTL.
- Let $\mathsf{q} = 1/0$ if the (unobserved) QTL genotype is BB/AB.

  (Or 2/1/0 if the QTL genotype is BB/AB/AA in an intercross.)

  Assume $\mathsf{y}|\mathsf{q} \sim \mathsf{N}(\mu_{\mathsf{q}}, \sigma)$
- Given genotypes at linked markers, $\mathsf{y} \sim$ mixture of normal dist'ns with mixing proportions $\Pr(\mathsf{q} \mid \text{marker data})$:

Genotype probabilities



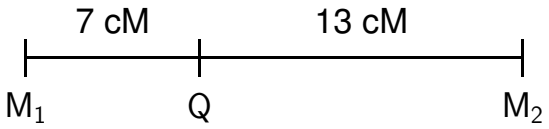
Calculate $\Pr(\mathsf{q} \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

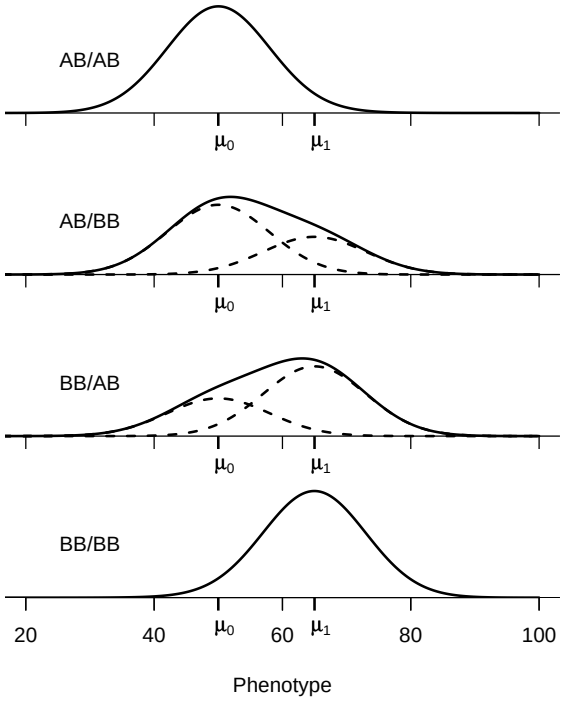
Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)

The normal mixtures



- Two markers separated by 20 cM, with the QTL closer to the left marker.
- The figure at right shows 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

Let $p_{ij} = \Pr(q_i = j|\text{marker data})$

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

$\Pr(y_i|\text{marker data}, \mu_0, \mu_1, \sigma) = \sum_j p_{ij}\, f(y_i; \mu_j, \sigma)$

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.

EM algorithm

Dempster et al. (1977)

E step:

Let $w_{ij}^{(k)} = \Pr(q_i = j | y_i, \text{marker data}, \hat{\mu}_0^{(k-1)}, \hat{\mu}_1^{(k-1)}, \hat{\sigma}^{(k-1)})$

$$= \frac{p_{ij}\, f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})}{\sum_j p_{ij}\, f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})}$$

M step:

Let $\hat{\mu}_j^{(k)} = \sum_i y_i w_{ij}^{(k)} / \sum_i w_{ij}^{(k)}$

$$\hat{\sigma}^{(k)} = \sqrt{\sum_i \sum_j w_{ij}^{(k)} (y_i - \hat{\mu}_j^{(k)})^2 / n}$$

The algorithm:

Start with $w_{ij}^{(1)} = p_{ij}$; 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.

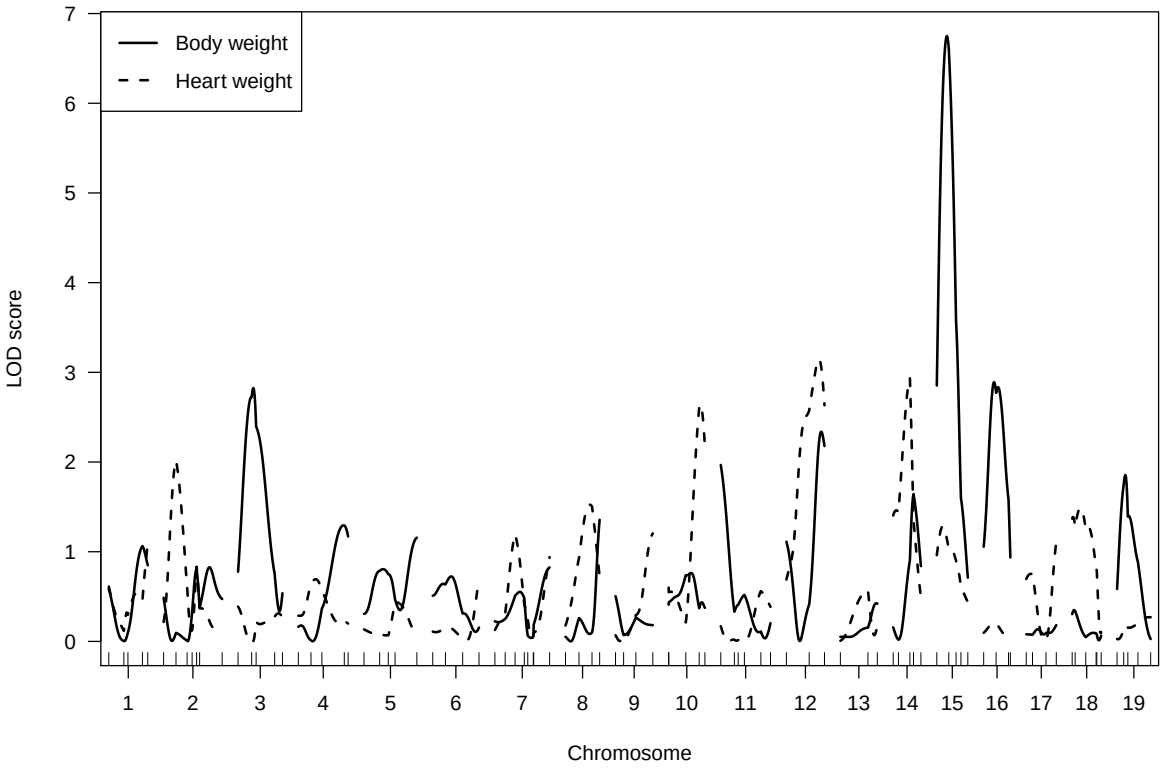
LOD(λ) $= \log_{10}$ likelihood ratio comparing the hypothesis of a QTL at position λ versus that of no QTL

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

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

No QTL model: The phenotypes are independent and identically distributed (iid) N(μ, σ^2).

LOD curves



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

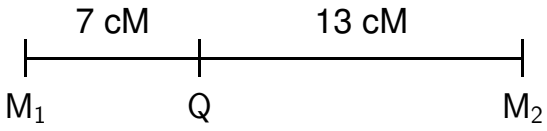
A quick approximation to Interval Mapping.

$$\begin{aligned} E(y_i|q_i) &= \mu_q \\ E(y_i|M_i) &= E[\, E(y_i|q_i) \,|M_i] = \textstyle\sum_j \Pr(q = j|M_i)\mu_j \\ &= \textstyle\sum_j p_{ij}\mu_j \end{aligned}$$

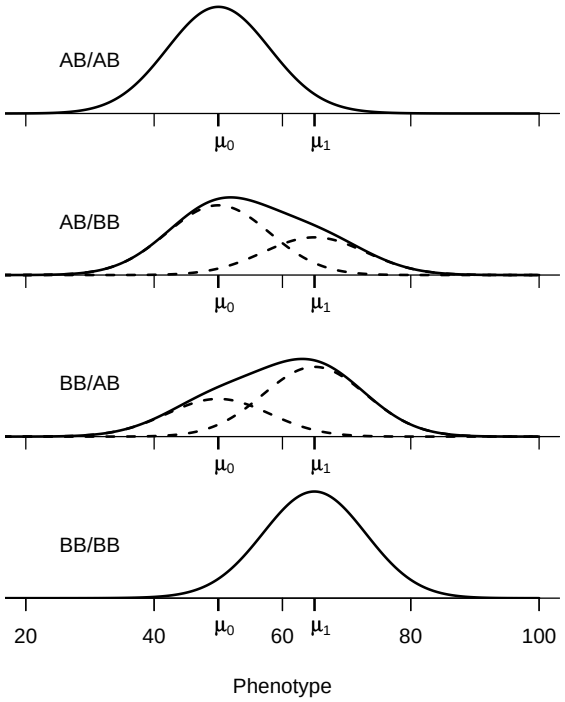
Regress y on p_i, pretending the residual variation is normally distributed (with constant variance).

$$\text{LOD} = \frac{n}{2}\log_{10}\left(\frac{RSS_0}{RSS_1}\right)$$

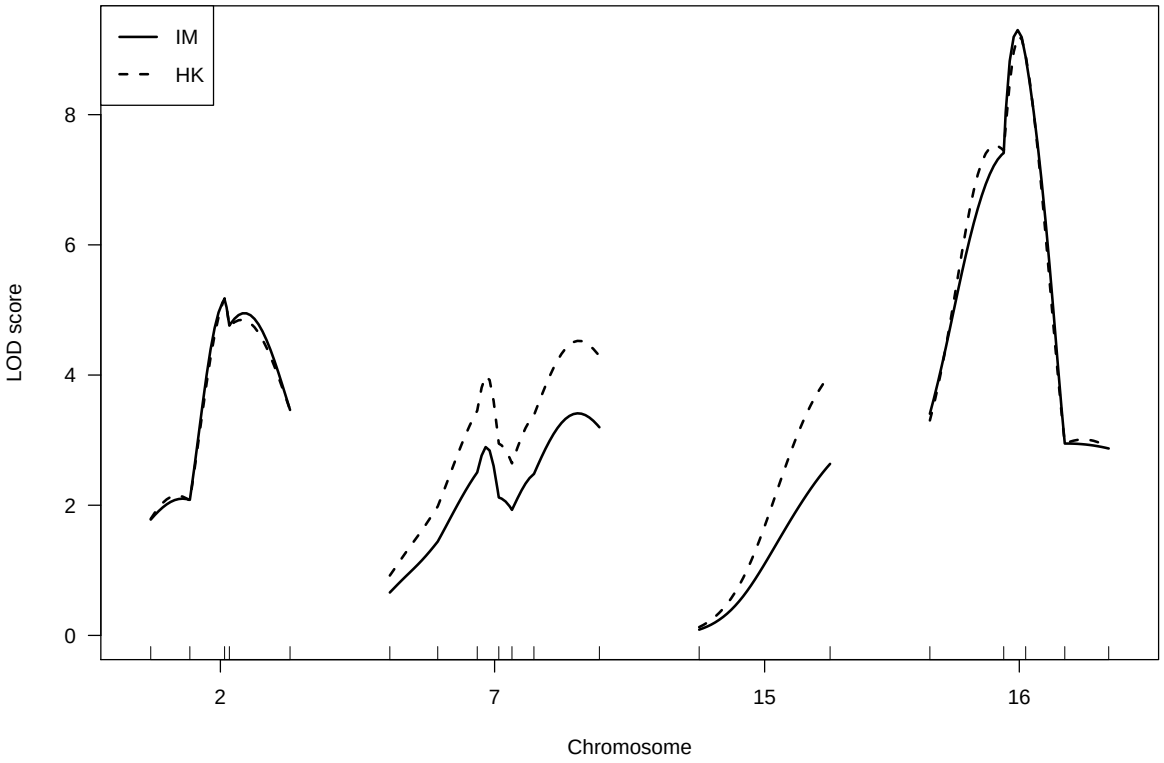
The normal mixtures



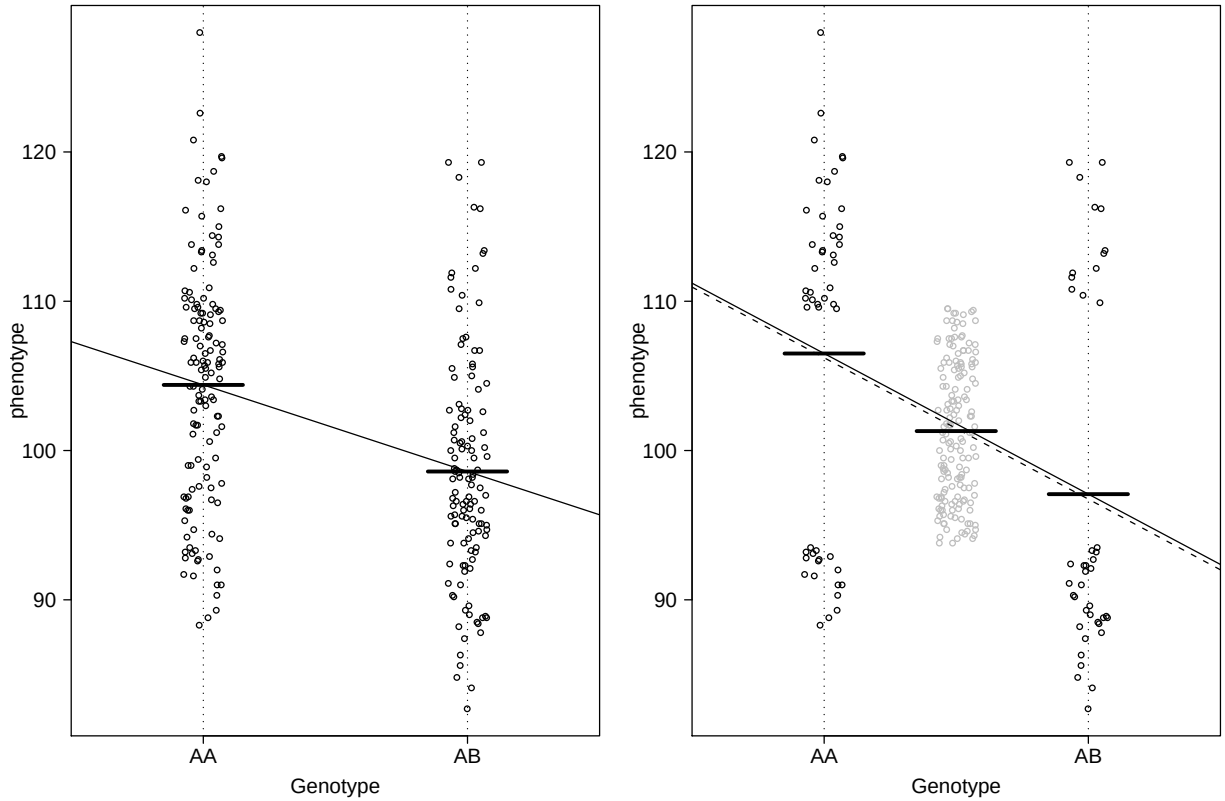
- Two markers separated by 20 cM, with the QTL closer to the left marker.
- The figure at right shows the distributions of the phenotype conditional on the genotypes at the two markers.
- The dashed curves correspond to the components of the mixtures.



Haley-Knott results



H-K with selective genotyping



LOD thresholds

Large LOD scores indicate evidence for the presence of a QTL

Question: How large is large?

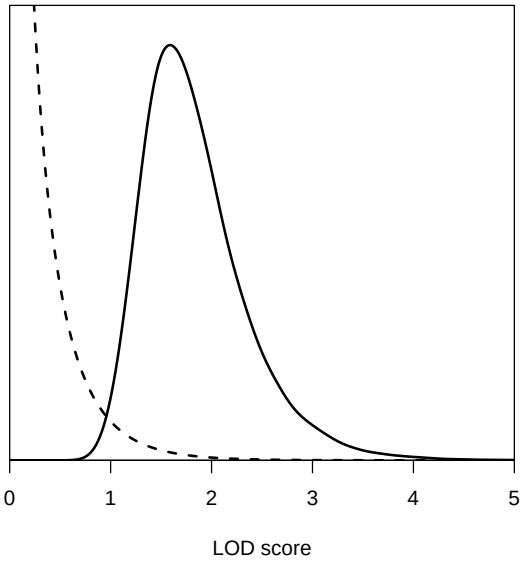
LOD threshold = 95 %ile of distr'n of max LOD, genome-wide, if there are no QTLs anywhere

Derivation:

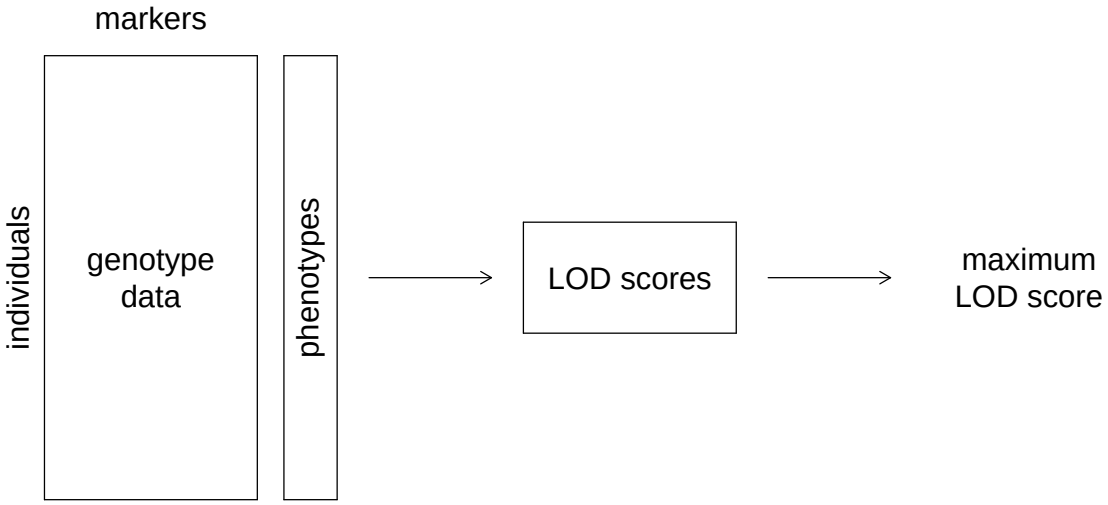
- Analytical calculations (L & B 1989)
- Simulations (L & B 1989)
- Permutation tests (Churchill & Doerge 1994)

Null distribution of the LOD score

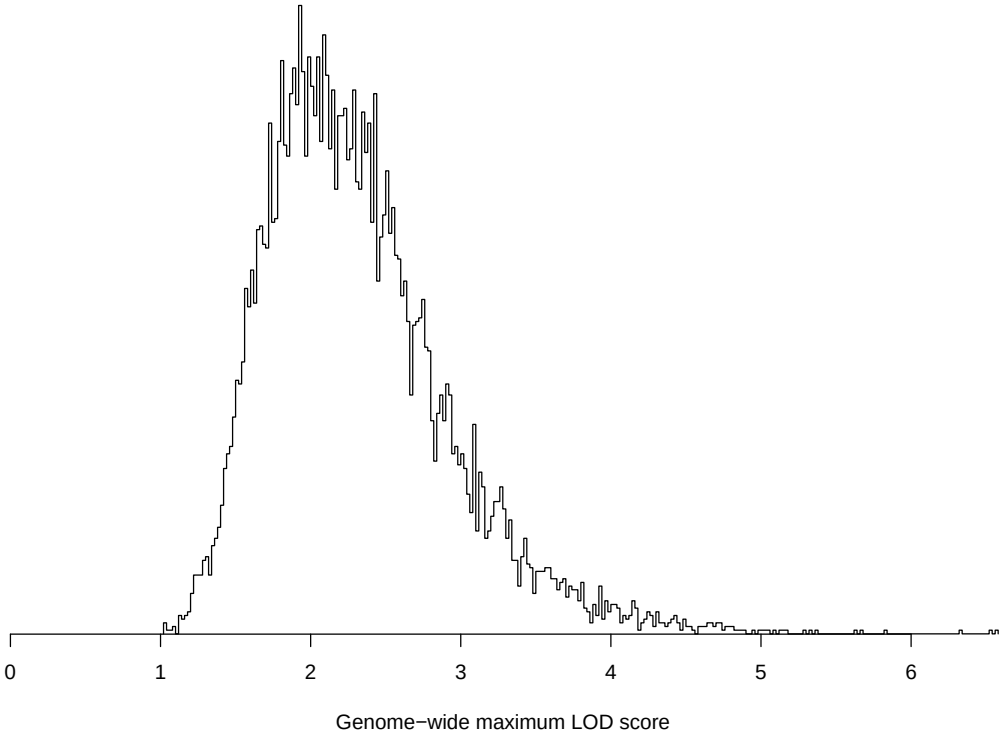
- Null distribution derived by computer simulation of backcross with genome of typical size.
- Dashed curve: distribution of LOD score at any one point.
- Solid curve: distribution of maximum LOD score, genome-wide.



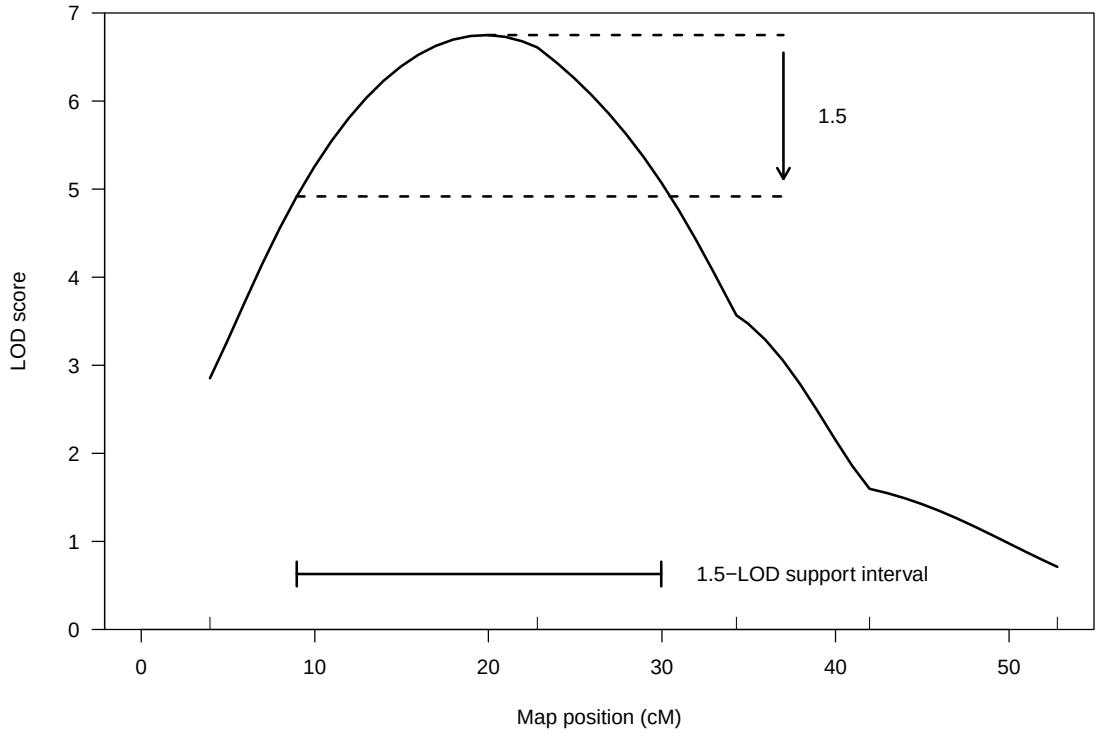
Permutation test



Permutation results



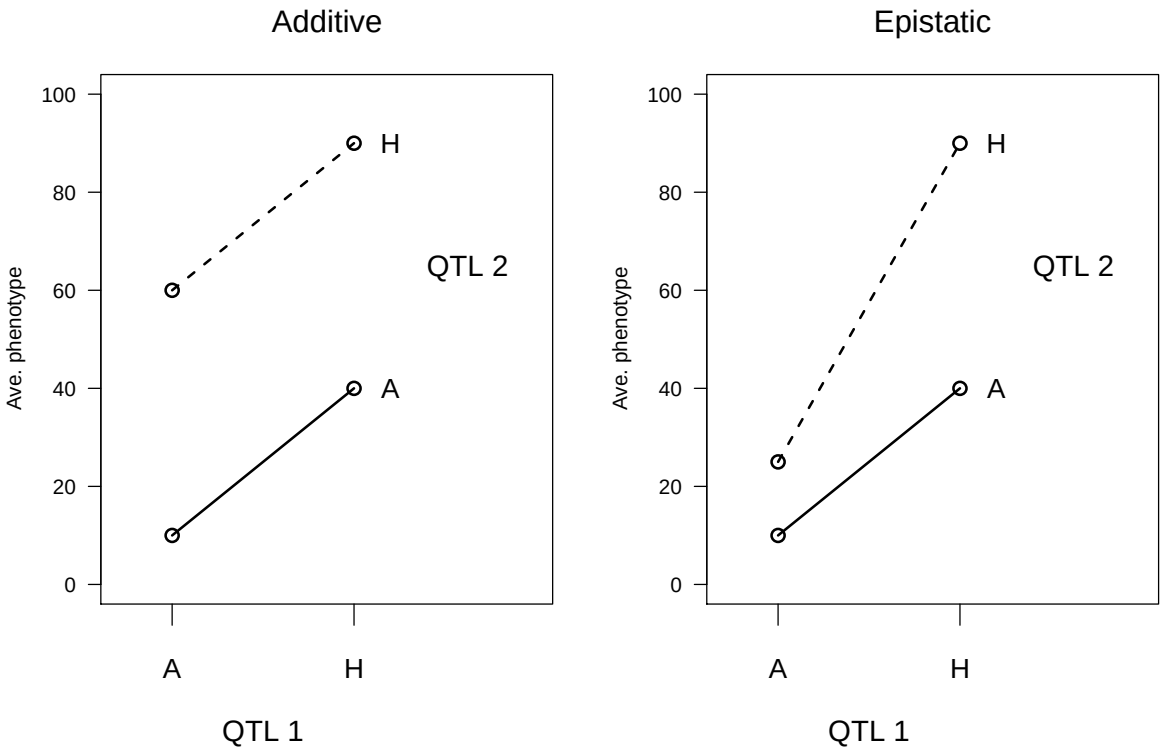
LOD support intervals



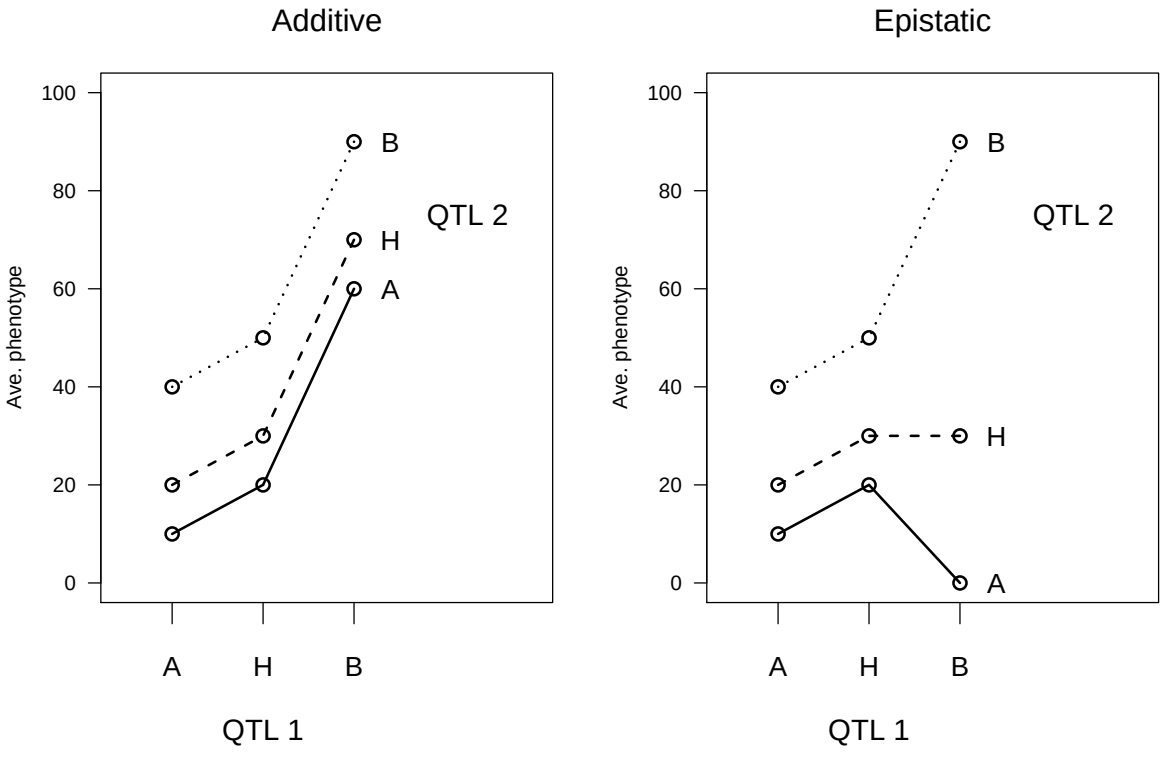
Modelling multiple QTL

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

Epistasis in BC



Epistasis in F_2



References

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

  A review for non-statisticians.

- 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.

- 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.

URLs

`http://kbroman.org`

`http://kbroman.org/pages/teaching.html`

`http://www.biostat.wisc.edu/~kbroman/D3/em_alg`

`http://www.biostat.wisc.edu/~kbroman/D3/lod_and_effect`

`http://www.biostat.wisc.edu/~kbroman/D3/lod_random`