

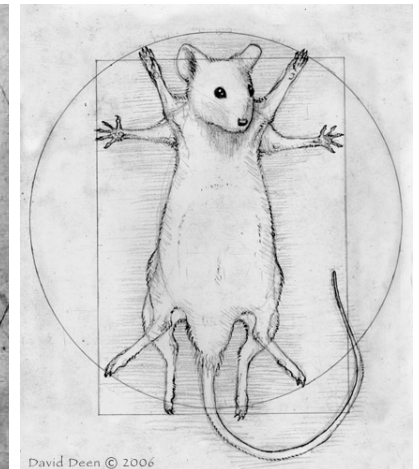
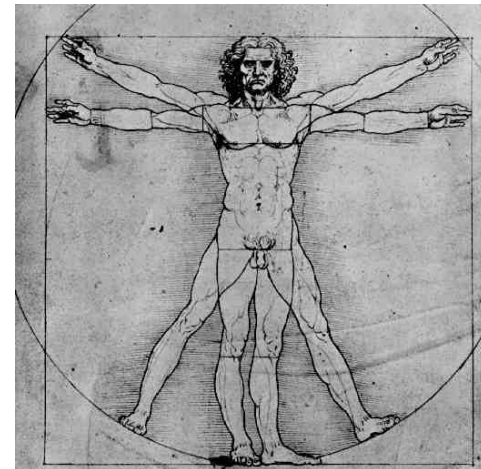
Introduction to QTL mapping in model organisms

Karl W Broman

Department of Biostatistics
Johns Hopkins University

www.biostat.jhsph.edu/~kbroman

[→ Teaching → Miscellaneous lectures]

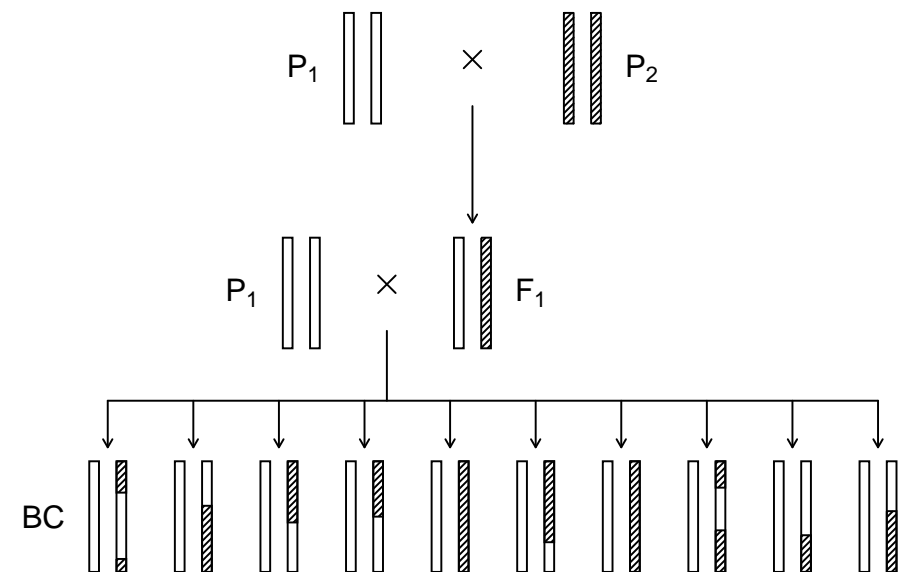


www.daviddeen.com

3

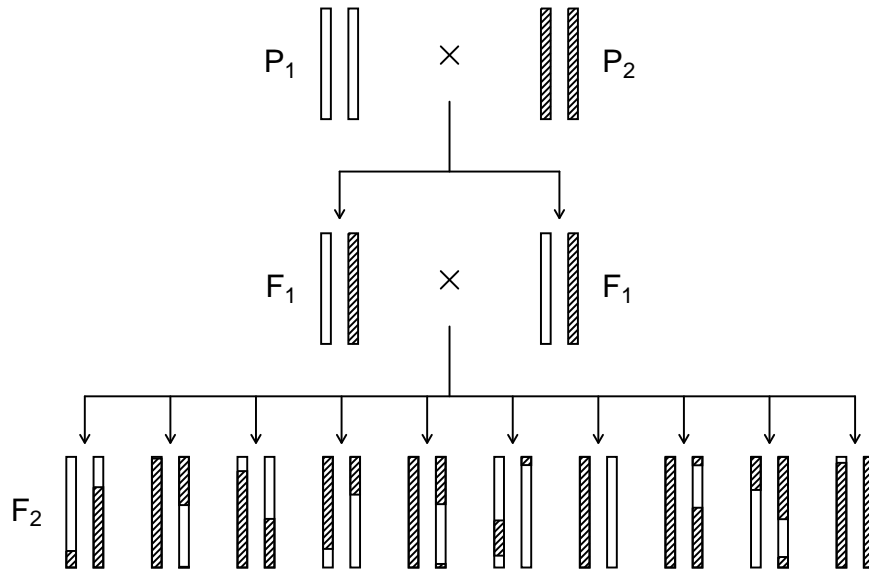


Backcross



4

Intercross



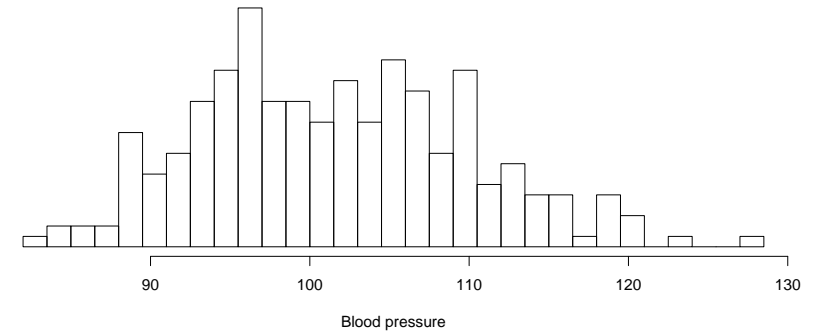
5

Phenotype data

Sugiyama et al. Genomics 71:70-77, 2001

250 male mice from the backcross (A × B) × B

Blood pressure after two weeks drinking water with 1% NaCl



7

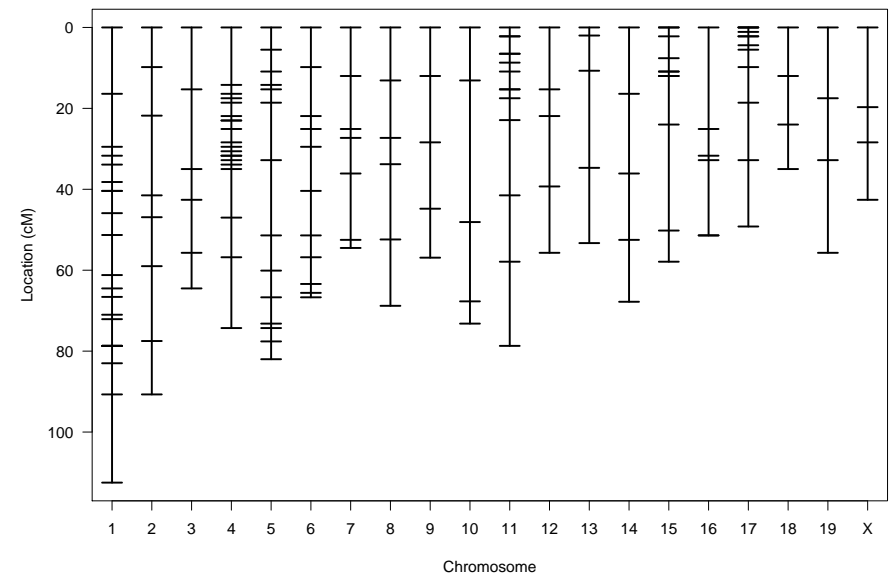
Data

Phenotypes: y_i = trait value for individual i

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

Genetic map: Locations of markers

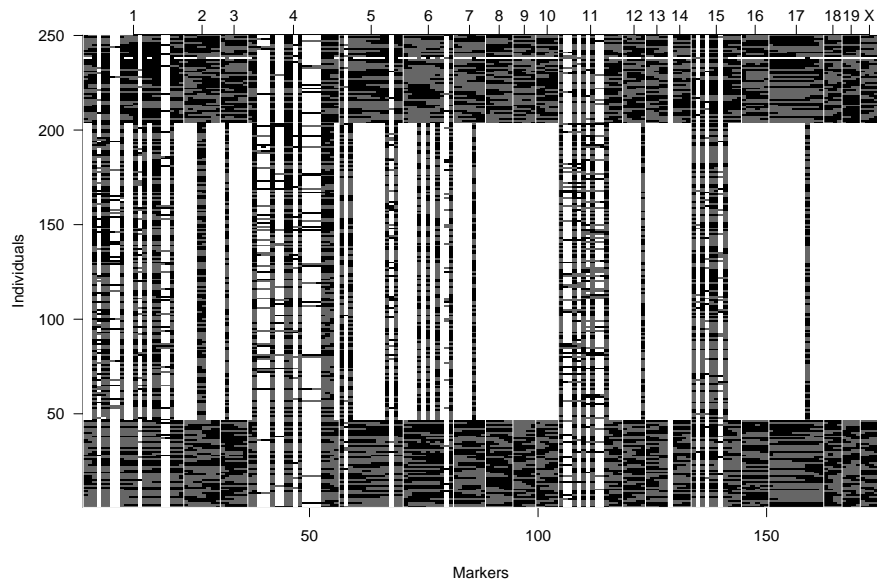
Genetic map



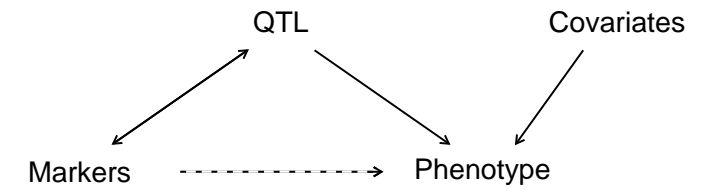
6

8

Genotype data



Statistical structure



The missing data problem:

Markers \longleftrightarrow QTL

The model selection problem:

QTL, covariates \longrightarrow phenotype

9

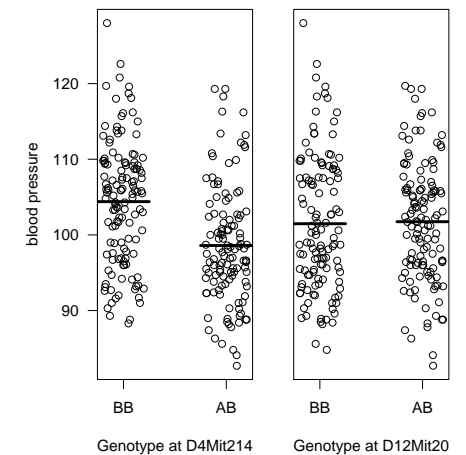
11

Goals

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

ANOVA at marker loci

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



10

12

ANOVA at marker loci

Advantages

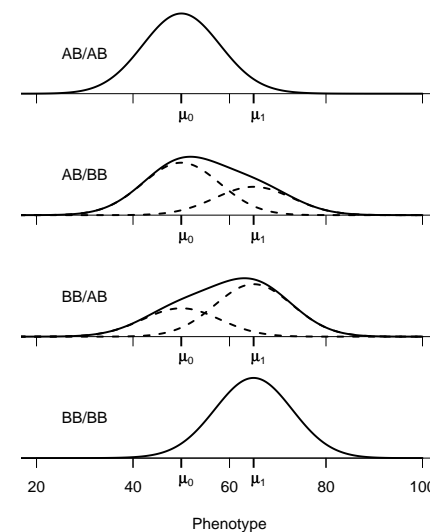
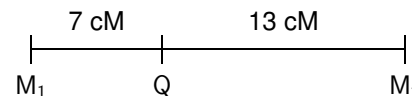
- Simple.
- Easily incorporates covariates.
- Easily extended to more complex models.
- Doesn't require a genetic map.

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.

13

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.

15

Interval mapping

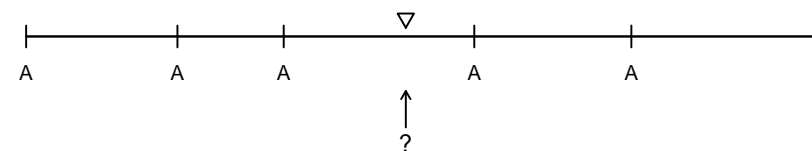
Lander & Botstein (1989)

- Assume a single QTL model.
- Each position in the genome, one at a time, is posited as the putative QTL.
- Let $q = 1/0$ if the (unobserved) QTL genotype is BB/AB.
(Or $2/1/0$ if the QTL genotype is BB/AB/AA in an intercross.)
Assume $y|q \sim N(\mu_q, \sigma)$
- Given genotypes at linked markers, $y \sim$ mixture of normal dist'ns with mixing proportions $\Pr(q | \text{marker data})$:

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	$r_L r_R/(1 - r)$	$(1 - r_L)(1 - r_R)/(1 - r)$

14

Genotype probabilities



Calculate $\Pr(q | \text{marker data})$, 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.)

16

Interval mapping

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

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

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

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

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.

17

EM algorithm

Dempster et al. (1977)

E step:

$$\begin{aligned} \text{Let } w_{ij}^{(k)} &= \Pr(q_i = j | y_i, \text{marker data}, \hat{\mu}_0^{(k-1)}, \hat{\mu}_1^{(k-1)}, \hat{\sigma}^{(k-1)}) \\ &= \frac{p_{ij} f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})}{\sum_j p_{ij} f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})} \end{aligned}$$

M step:

$$\begin{aligned} \text{Let } \hat{\mu}_j^{(k)} &= \sum_i y_i w_{ij}^{(k)} / \sum_i w_{ij}^{(k)} \\ \hat{\sigma}^{(k)} &= \sqrt{\sum_i \sum_j w_{ij}^{(k)} (y_i - \hat{\mu}_j^{(k)})^2 / n} \end{aligned}$$

The algorithm:

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

18

LOD scores

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

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

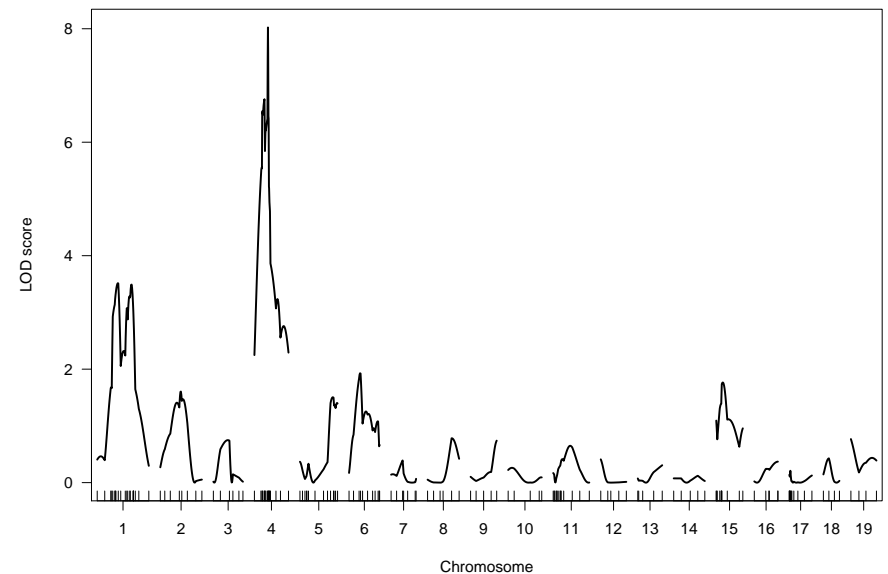
$$= \log_{10} \left\{ \frac{\Pr(y | \text{QTL at } \lambda, \hat{\mu}_{0\lambda}, \hat{\mu}_{1\lambda}, \hat{\sigma}_\lambda)}{\Pr(y | \text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\}$$

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

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

19

LOD curves



20

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.

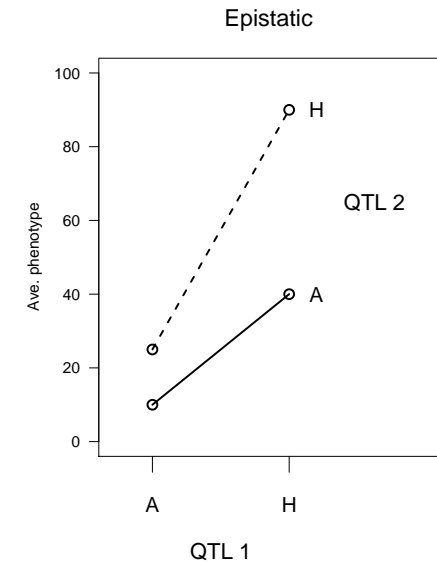
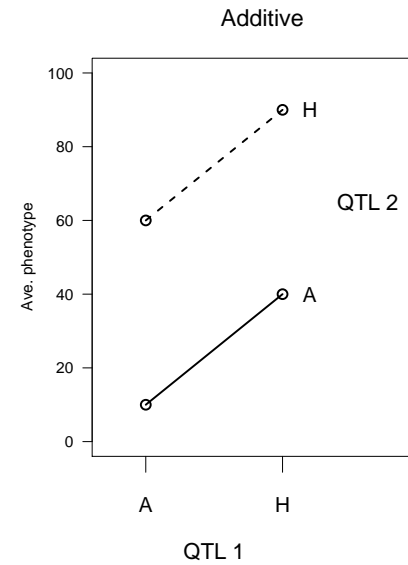
21

Modelling multiple QTL

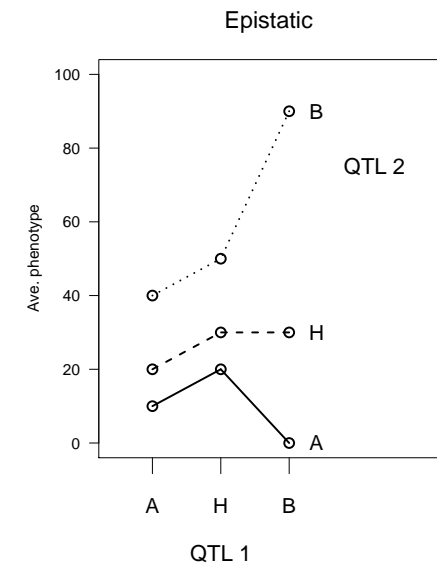
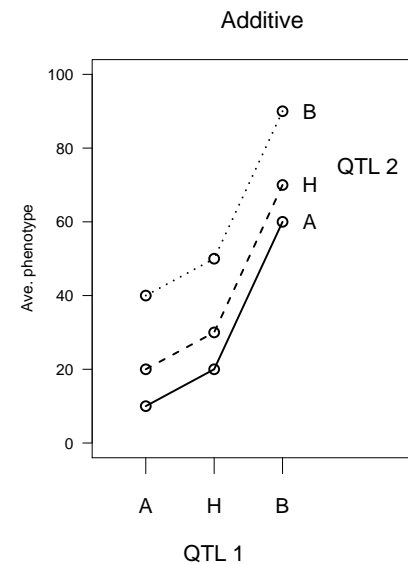
- Reduce residual variation \implies increased power
- Separate linked QTL
- Identify interactions among QTL

22

Epistasis in BC

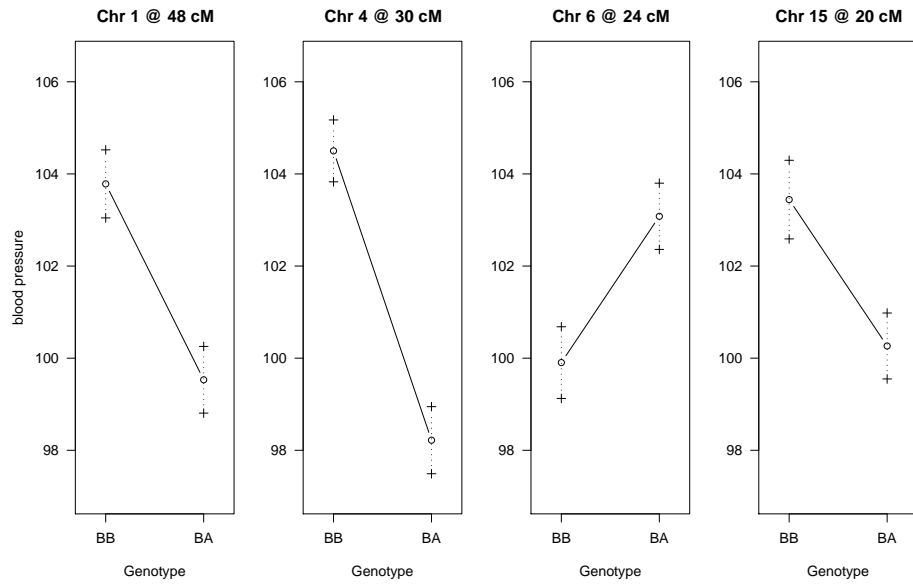


23



24

Our estimated effects



25

LOD thresholds

Large LOD scores indicate evidence for the presence of a QTL

Question: How large is large?

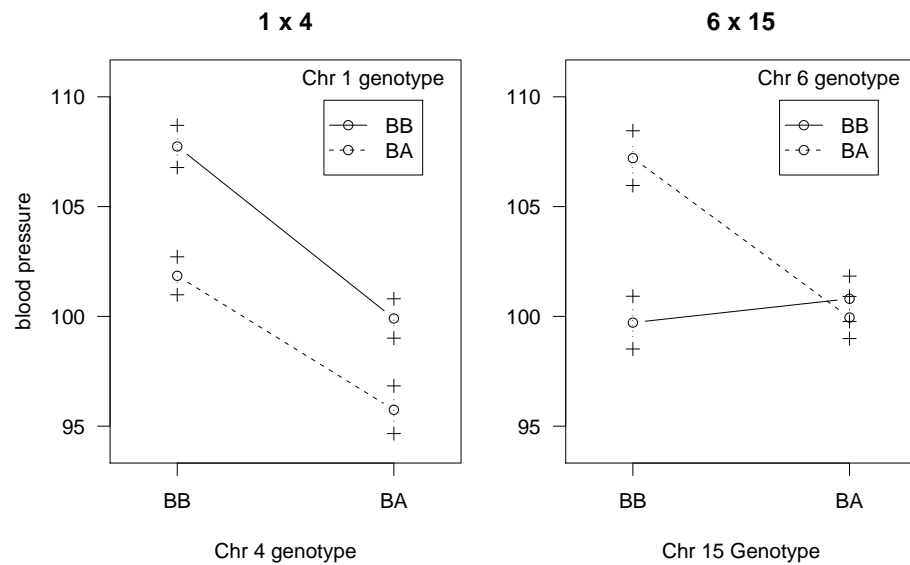
LOD threshold = 95 %ile of distr'n of max LOD, genome-wide, if there are no QTLs anywhere

Derivation:

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

27

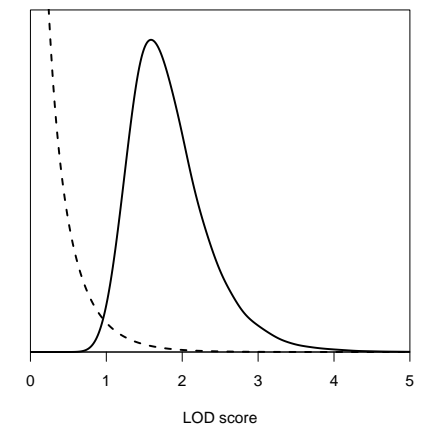
Our estimated effects



26

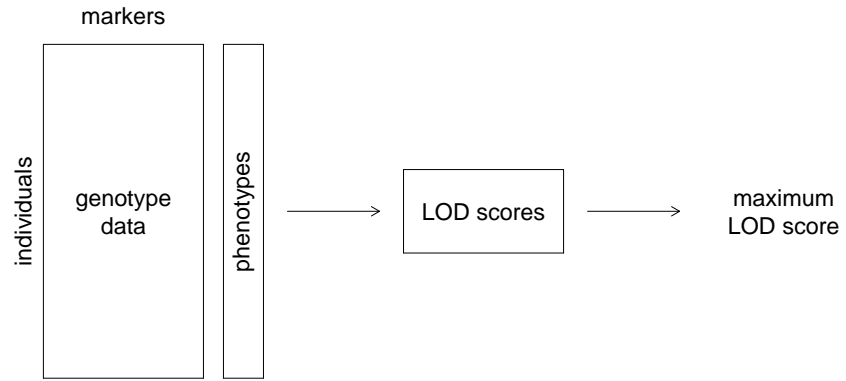
Null distribution of the LOD score

- 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.

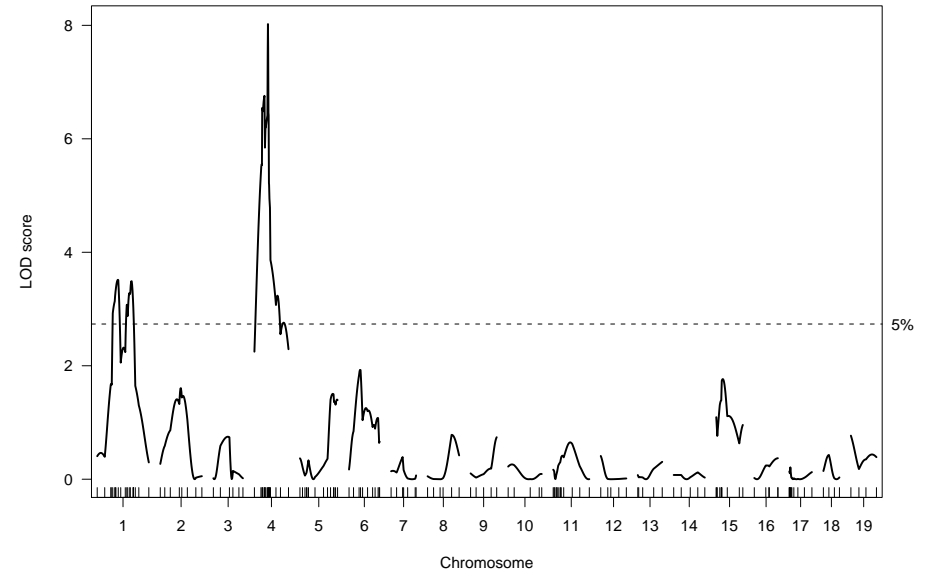


28

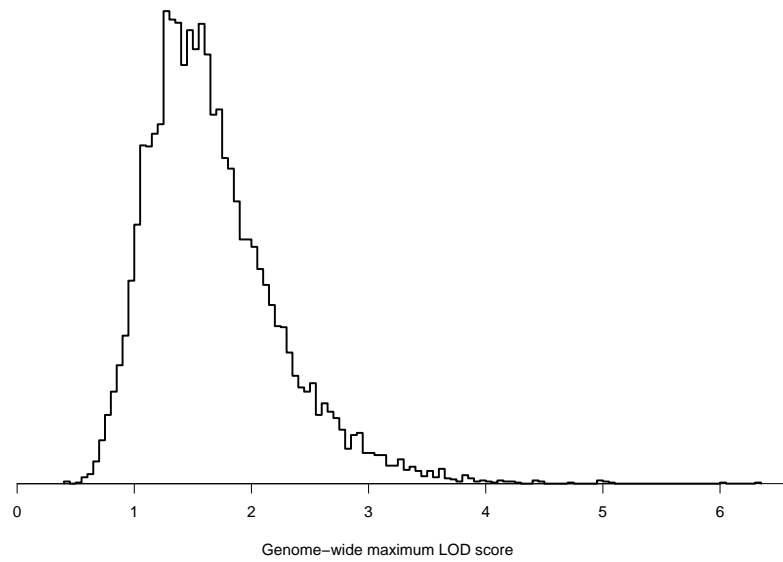
Permutation test



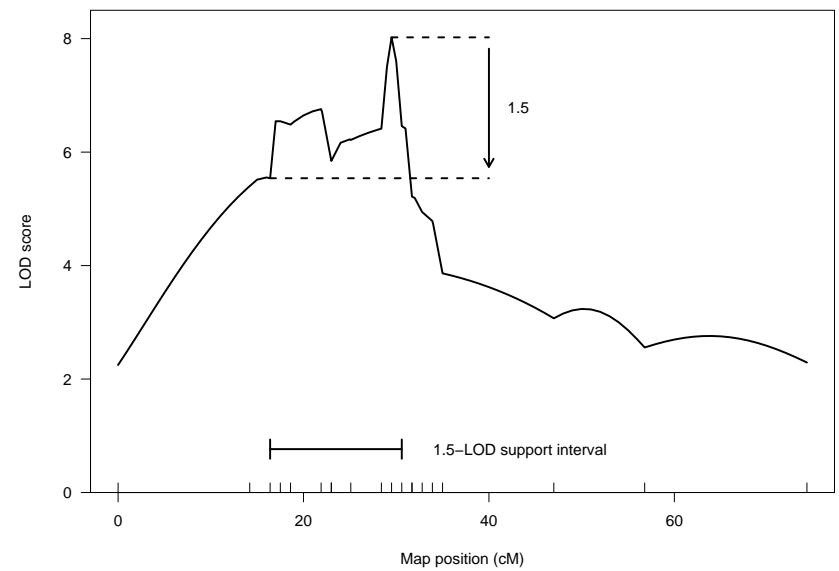
LOD curves



Permutation results

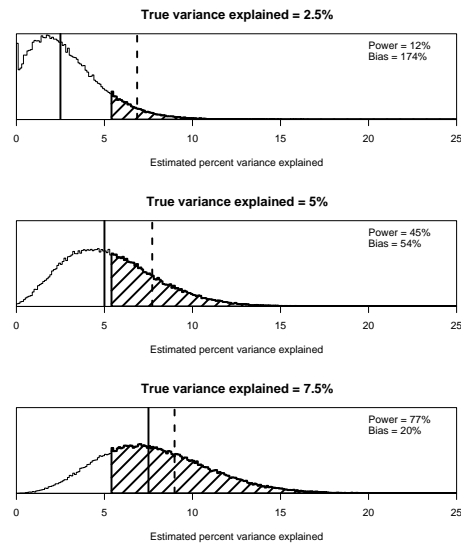


LOD support intervals



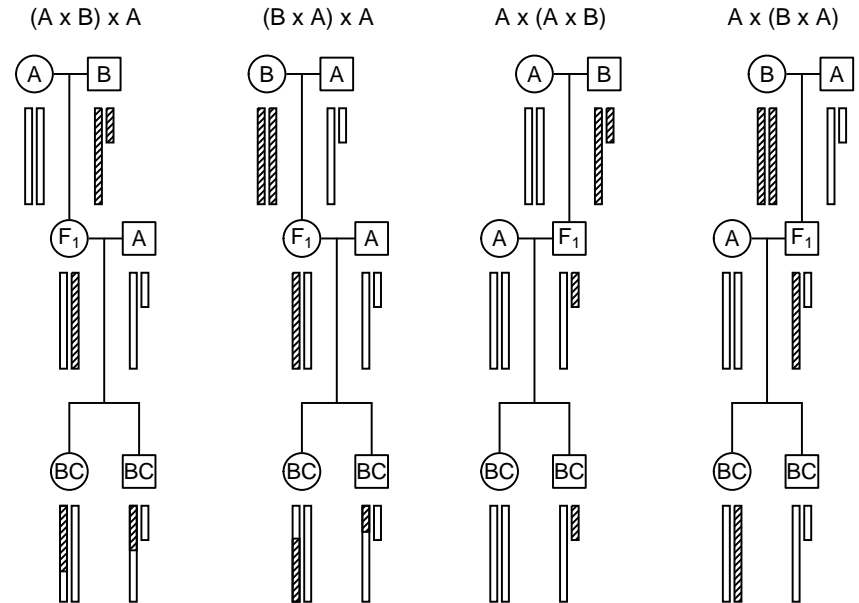
Selection bias

- 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.



33

X chr in a backcross

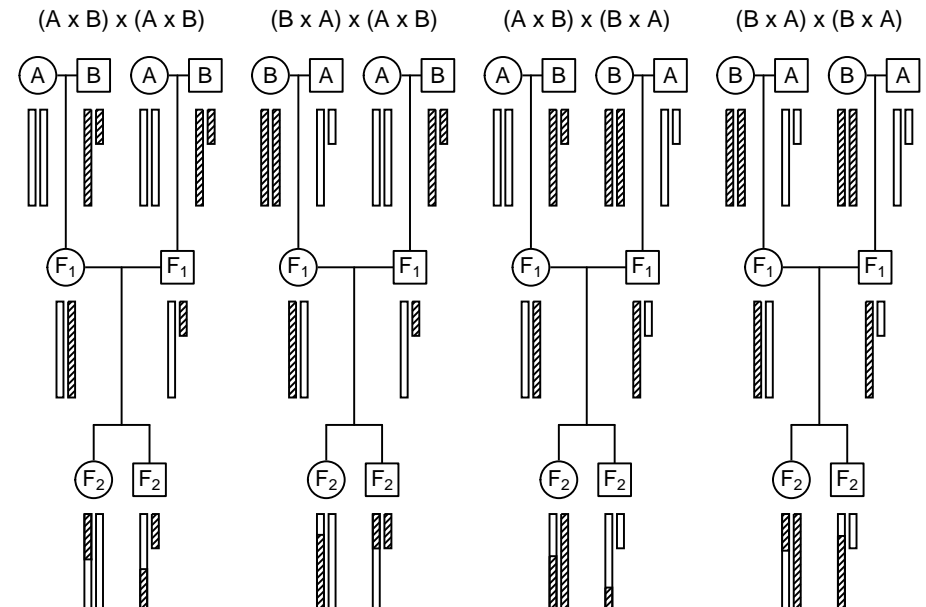


35

Implications

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

X chr in an intercross

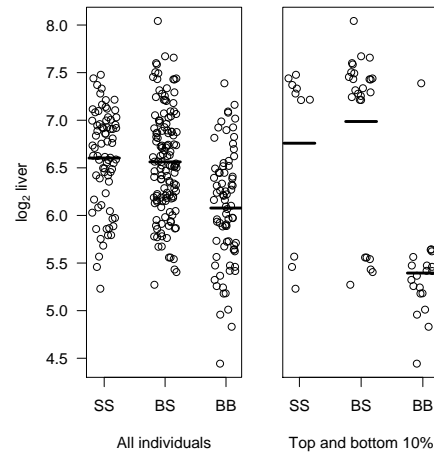


34

36

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.



37

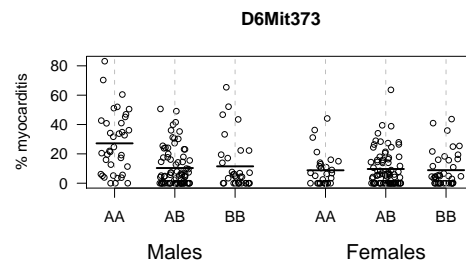
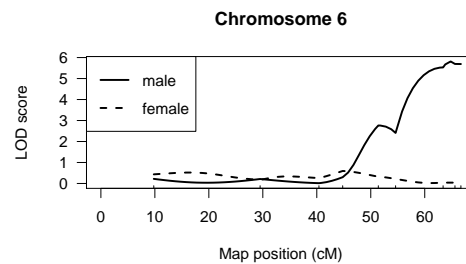
Non-normal traits

- 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)

39

Covariates

- Examples : treatment, sex, litter, lab, age.
- Control residual variation.
- Avoid confounding.
- Look for QTL \times covariate interactions



38

Data diagnostics

- Plot phenotypes
- Segregation distortion
- Genetic maps/marker positions
- Genotyping errors

40

References

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

41

- Kruglyak L, Lander ES (1995) A nonparametric approach for mapping quantitative trait loci. *Genetics* 139:1421–1428
- 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.
- Broman KW, Sen Ś, Owens SE, Manichaikul A, Southard-Smith EM, Churchill GA (2006) The X chromosome in quantitative trait locus mapping. *Genetics* 174:2151–2158
- 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.

42