Introduction to QTL mapping in model organisms

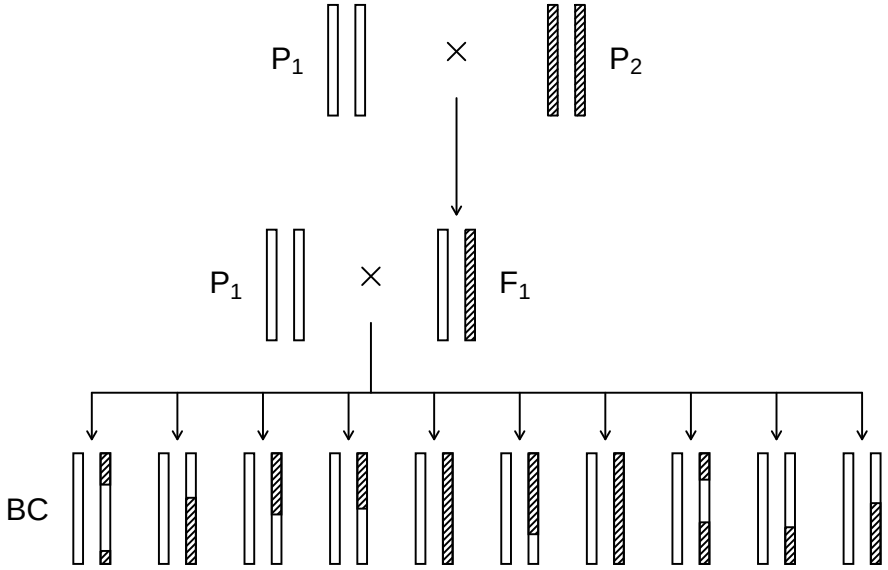
Karl W Broman

Department of Biostatistics and Medical Informatics
University of Wisconsin – Madison

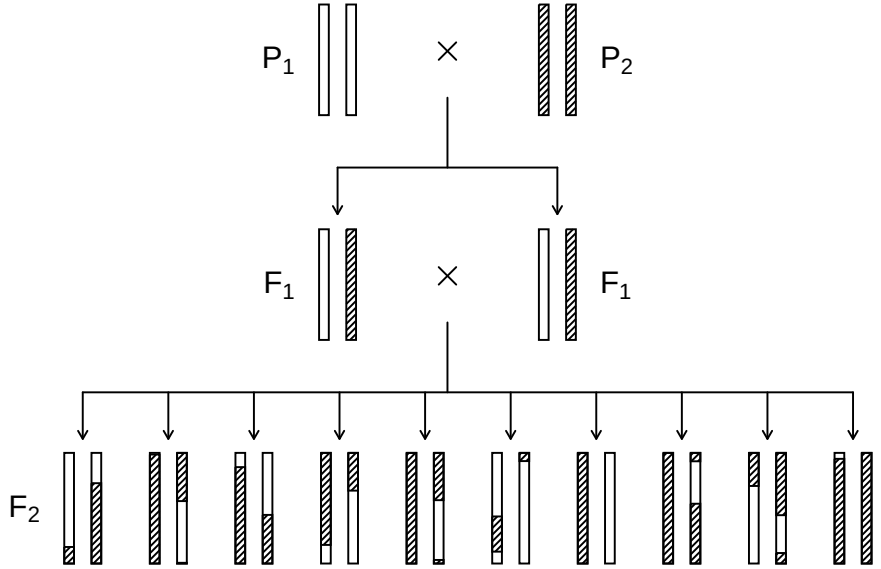
`www.biostat.wisc.edu/~kbroman`

[→ Teaching → Miscellaneous lectures]

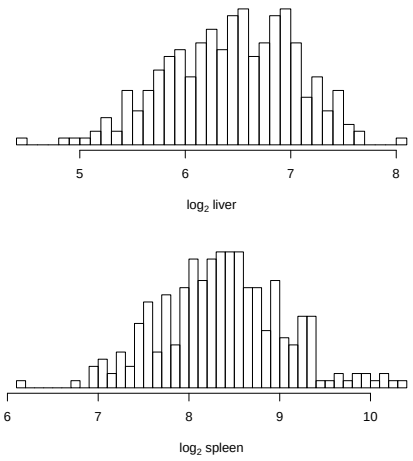
Backcross

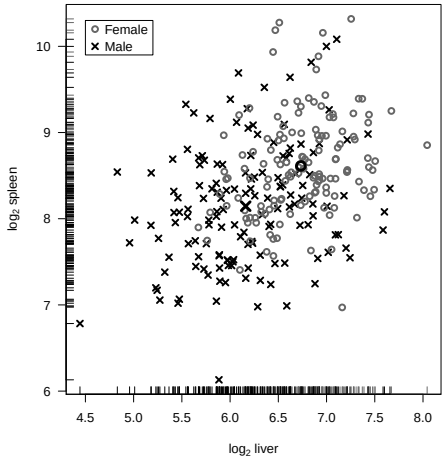


Intercross

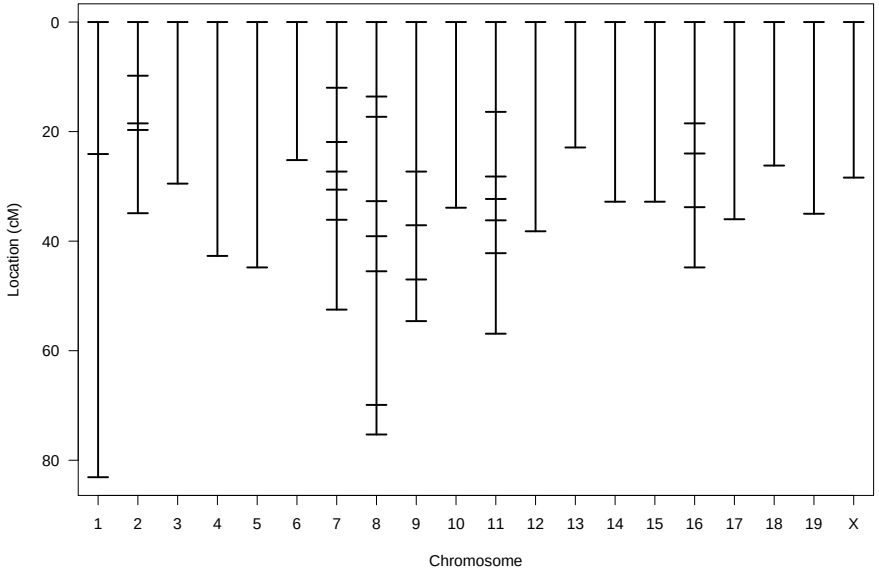


Phenotype data

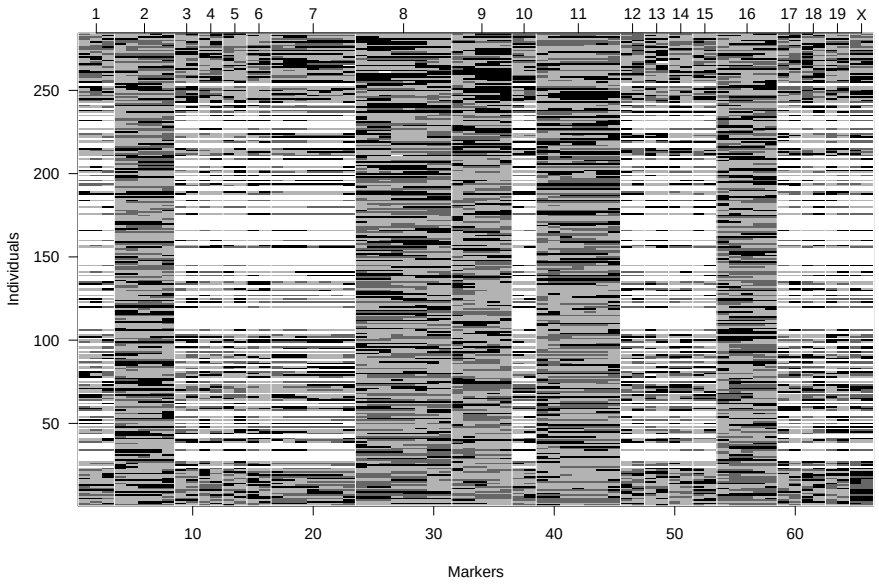




Genetic map



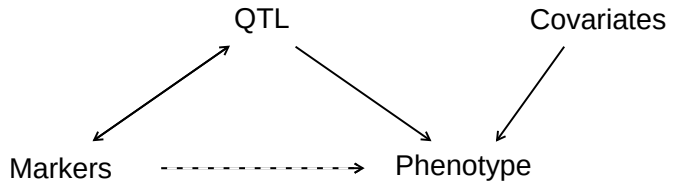
Genotype data



Goals

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

Statistical structure

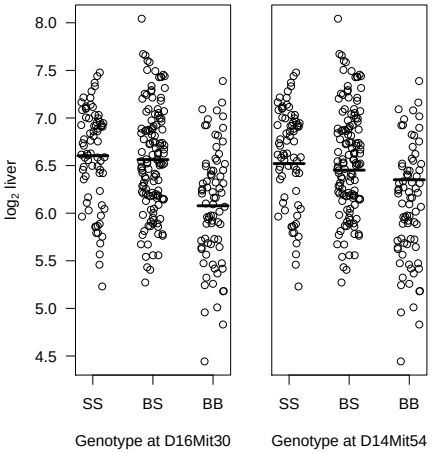


The missing data problem: Markers $\longleftrightarrow$ QTL

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.

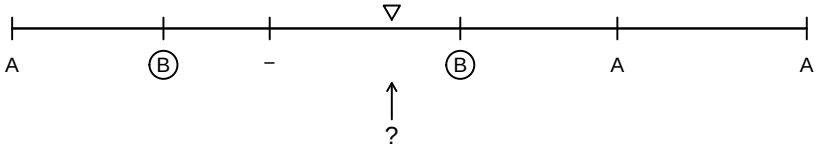
Interval mapping

Lander & Botstein (1989)

- Assume a single QTL model.
- Each position in the genome, one at a time, is posited as the putative QTL.
- Let $q = 1/0$ 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 $\Pr(q \mid \text{marker data})$:

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

Genotype probabilities



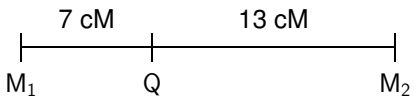
Calculate $\Pr(q \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

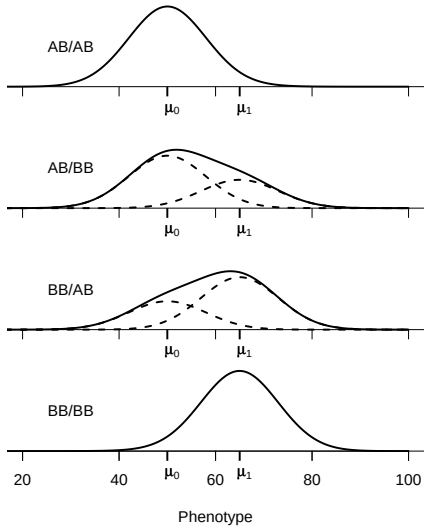
Or use the hidden Markov model (HMM) technology

- 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.



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.

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.

LOD scores

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

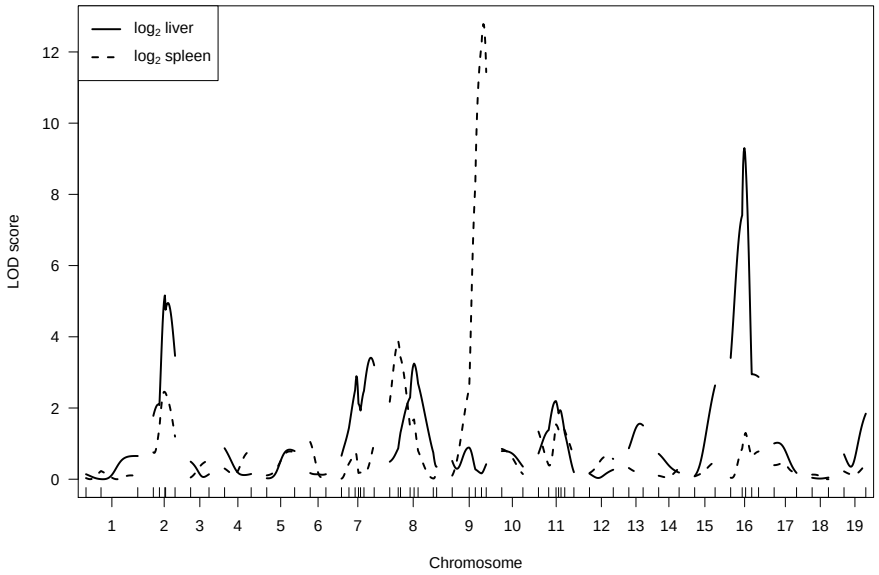
$\text{LOD}(\lambda) = \log_{10}$ likelihood ratio comparing the hypothesis of a QTL at position λ versus that of no QTL

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

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

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

LOD curves



LOD ↔ F

$$F = \left(10^{\frac{2}{n}LOD} - 1\right)\left(\frac{n - df - 1}{df}\right)$$

$$LOD = \frac{n}{2}\log_{10}\left[F\left(\frac{df}{n - df - 1}\right) + 1\right]$$

$$\text{estimated \% var explained} = 1 - 10^{-\frac{2}{n}LOD}$$

Interval mapping

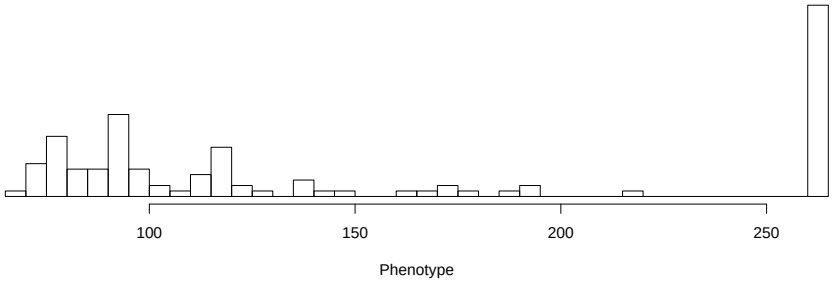
Advantages

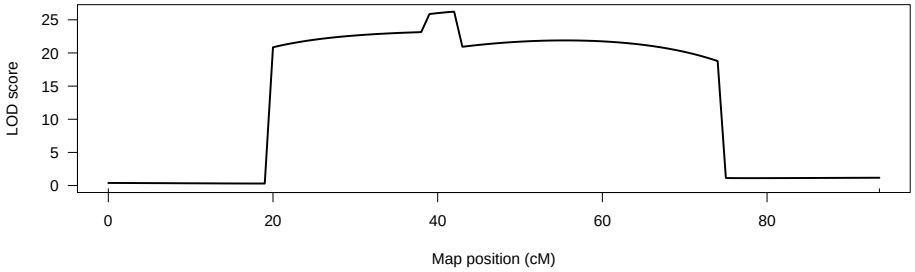
- Takes proper account of missing data.
- Allows examination of positions between markers.
- Gives improved estimates of QTL effects.
- Provides pretty graphs.

Disadvantages

- Increased computation time.
- Requires specialized software.
- Difficult to generalize.
- Can give spurious peaks.
- Only considers one QTL at a time.

Spurious LOD peak





LOD thresholds

Large LOD scores indicate evidence for the presence of a QTL

Question: How large is large?

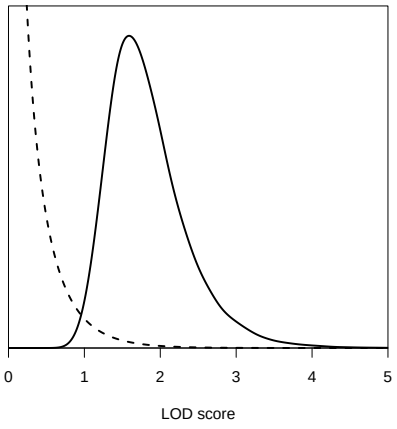
LOD threshold = 95 %ile of distr'n of max LOD, genome-wide, if there are no QTLs anywhere

Derivation:

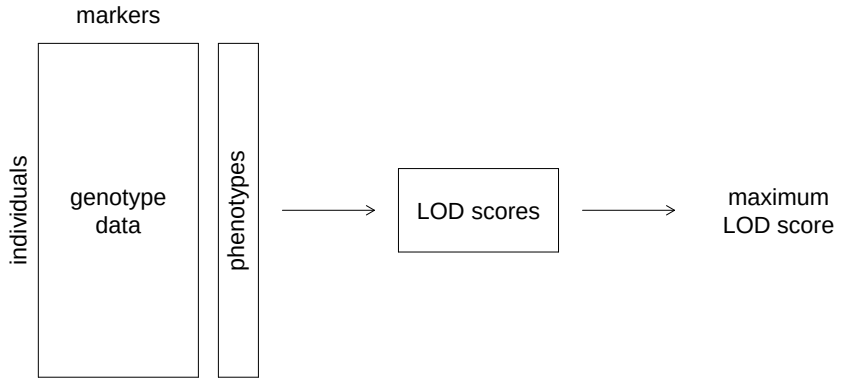
- Analytical calculations (L & B 1989)
- Simulations (L & B 1989)
- Permutation tests (Churchill & Doerge 1994)

Null distribution of the LOD score

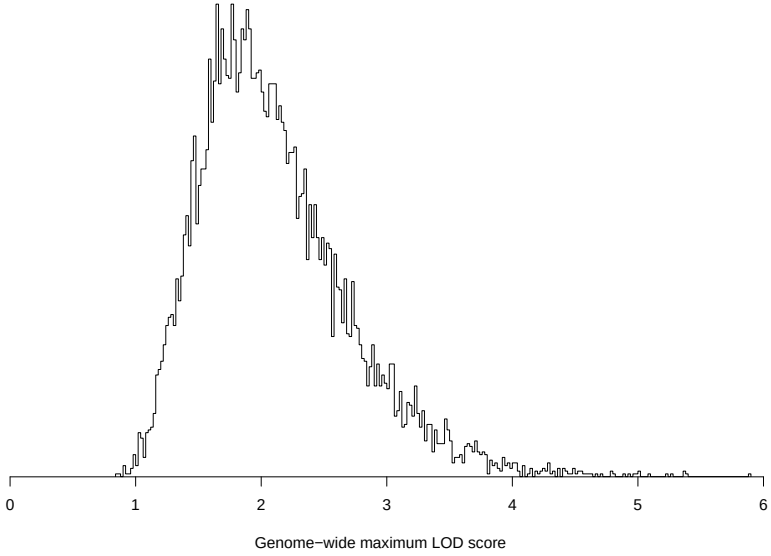
- Null distribution derived by computer simulation of backcross with genome of typical size.
- Dashed curve: distribution of LOD score at any one point.
- Solid curve: distribution of maximum LOD score, genome-wide.



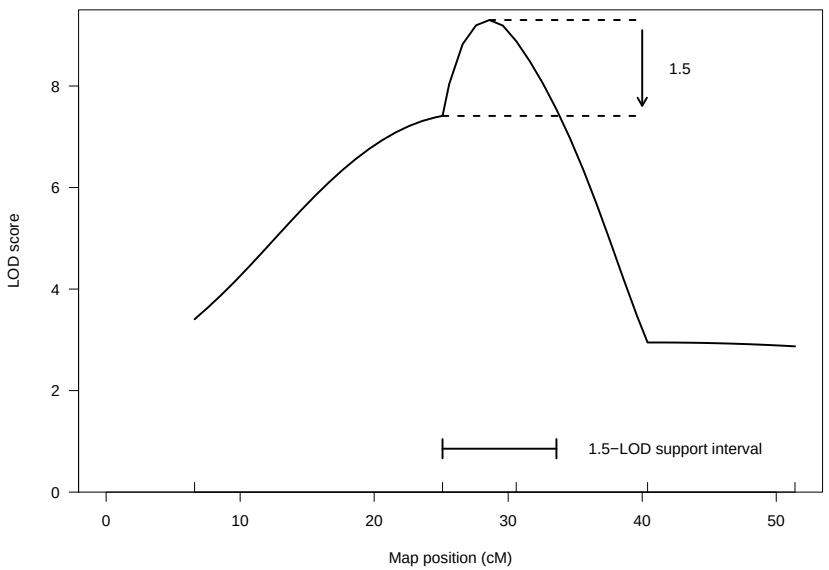
Permutation test



Permutation results



LOD support intervals



Haley-Knott regression

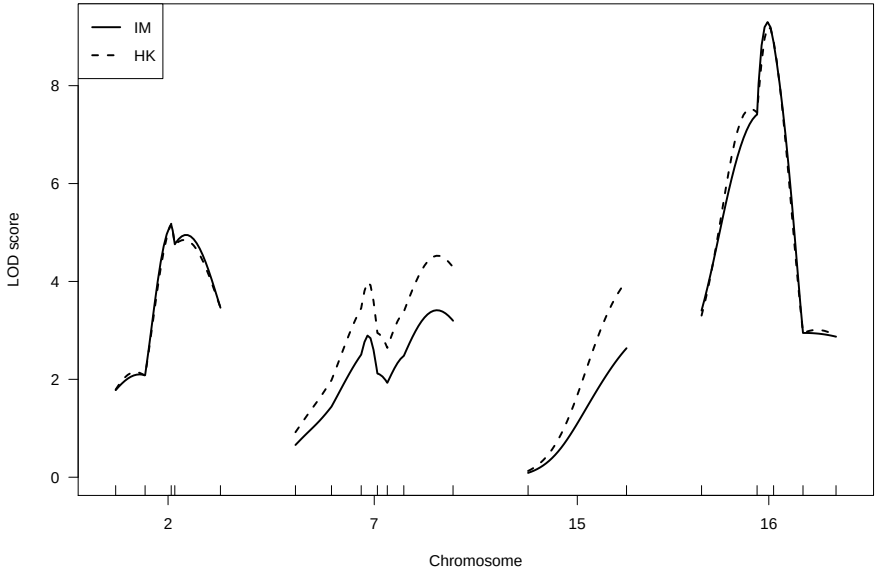
A quick approximation to Interval Mapping.

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

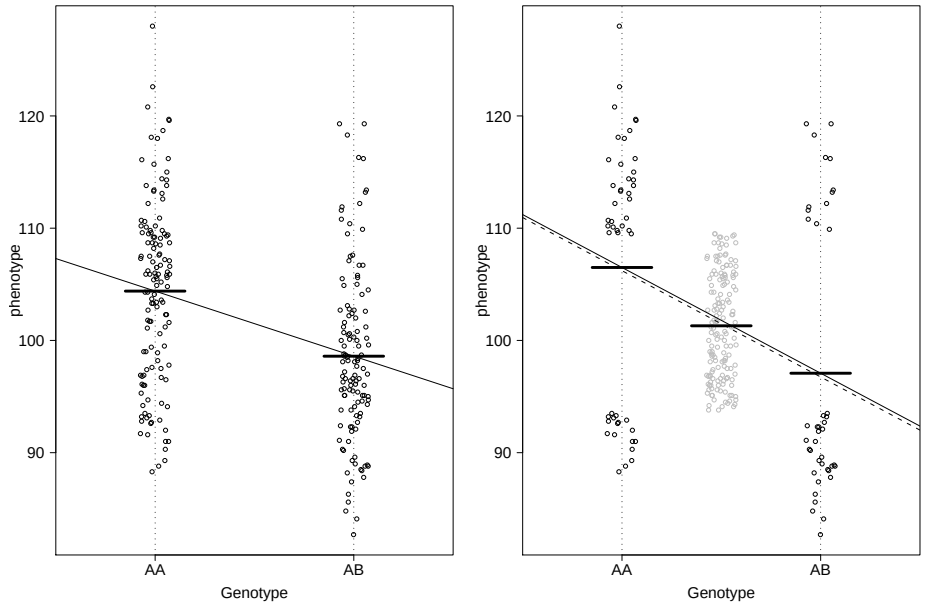
Regress y on p_i, pretending the residual variation is normally distributed (with constant variance).

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

Haley-Knott results



H-K with selective genotyping



Extended Haley-Knott

Like H-K, but also take account of the variances.

$$\text{Let} \quad m_i \equiv E(y_i|M_i) = \sum_j p_{ij}\mu_j$$

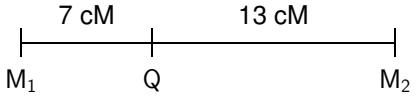
$$v_i \equiv \text{var}(y_i|M_i) = E[\text{ var}(y_i|q_i) \,|M_i] + \text{var}[\text{ } E(y_i|q_i) \,|M_i]$$
$$= \sigma^2 + \sum_j p_{ij}(\mu_j - m_i)^2$$

IM: $y_i|M_i \sim$ mixture of normals

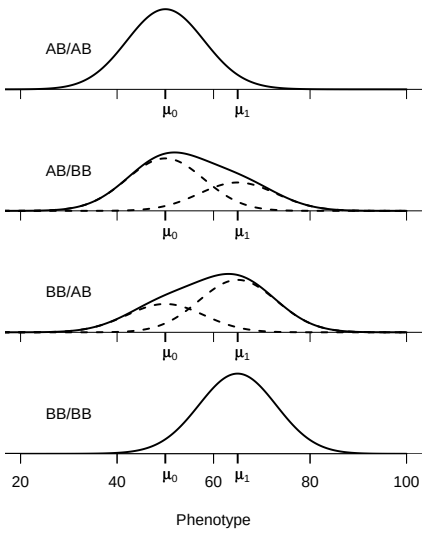
H-K: pretend $y_i|M_i \sim N(m_i, \sigma^2)$

eHK: pretend $y_i|M_i \sim N(m_i, v_i)$ (Again need an iterative algorithm.)

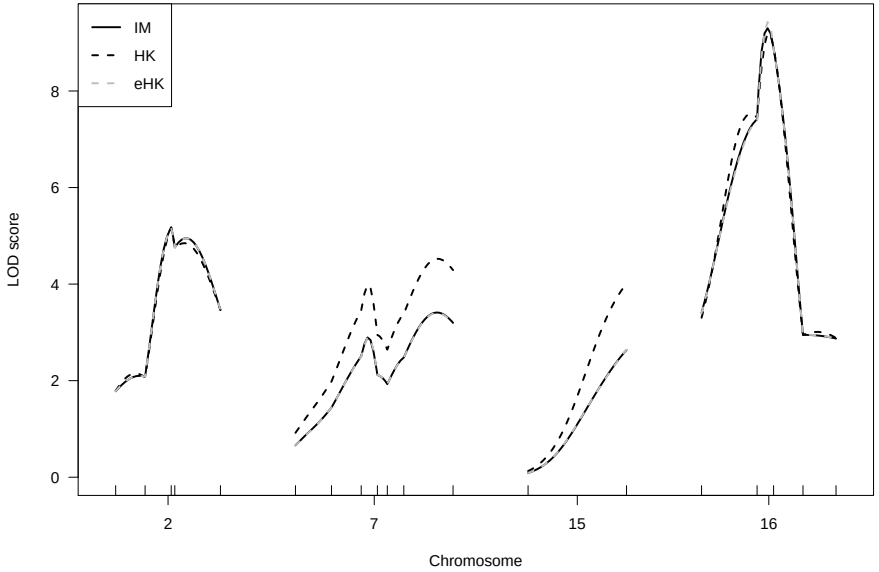
The normal mixtures, again



- 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.



eHK results

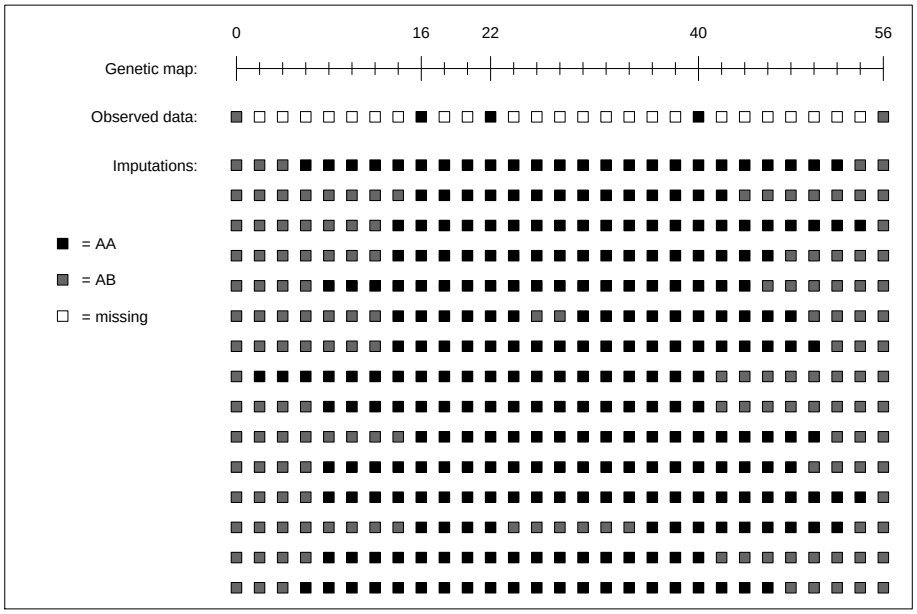


Multiple imputation

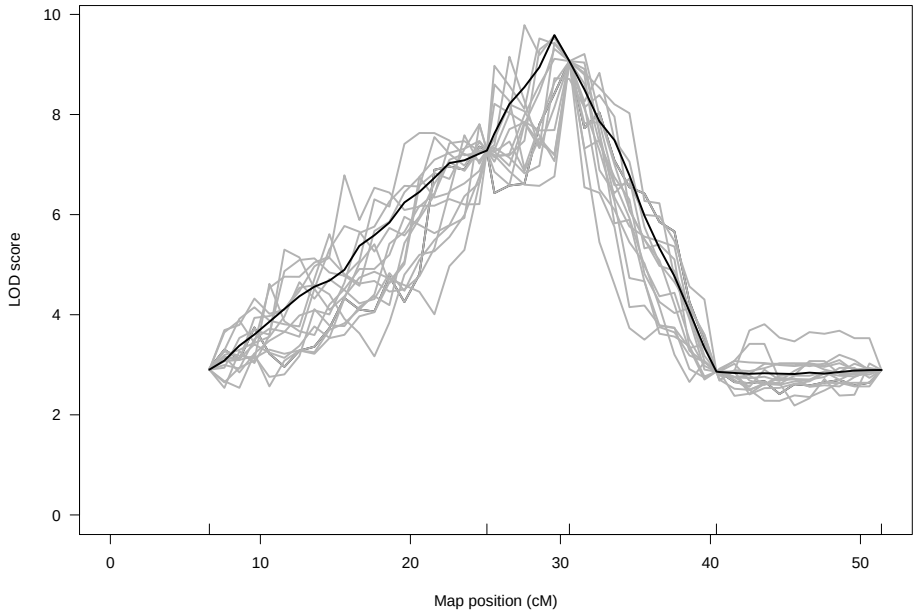
If we had complete genotype data, analysis would be easy, so:

- Fill in missing genotype data (at random, conditional on observed data)
- Calculate LOD curve
- Repeat many times
- Average the LOD curves

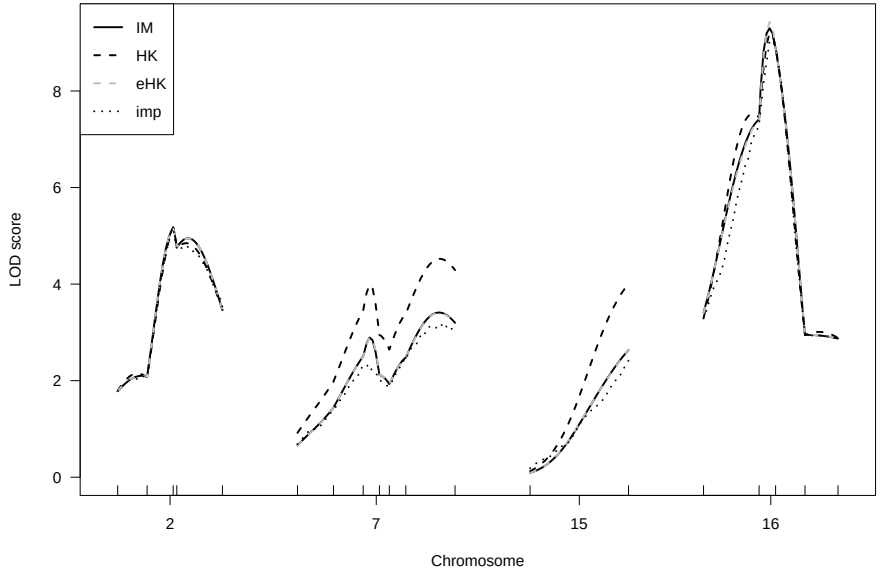
The imputations



Imputation LOD curves



Imputation results



Summary

- ANOVA

  Easy and fast, but must omit ungenotyped individuals.

- standard interval mapping

  Gold standard; but hard to extend, not fast, and can give spurious LOD peaks.

- Haley-Knott regression

  Easy and fast, but performs poorly in the case of selective genotyping.

- extended Haley-Knott

  More robust than IM, better approx'n than H-K, but not fast and not easy to extend.

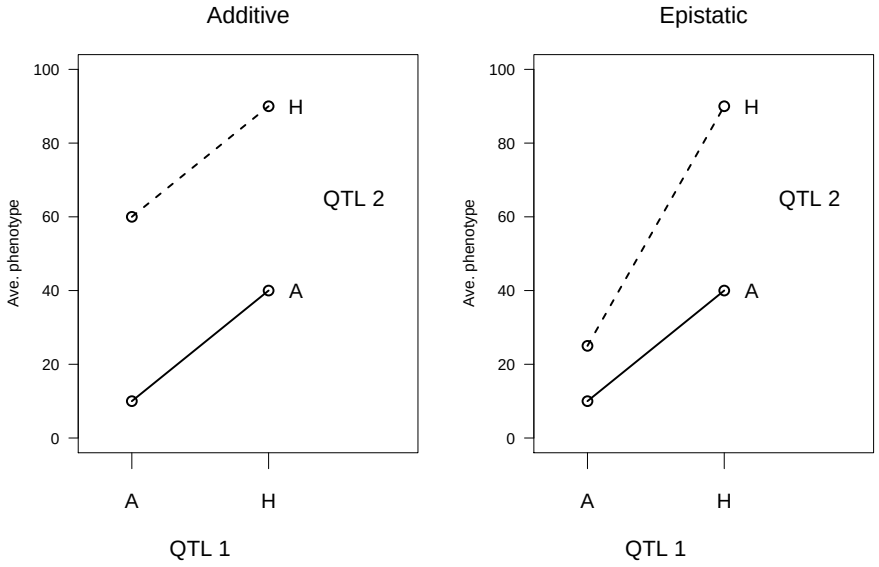
- multiple imputation

  Slow, but easy to extend; especially good for multiple QTL models.

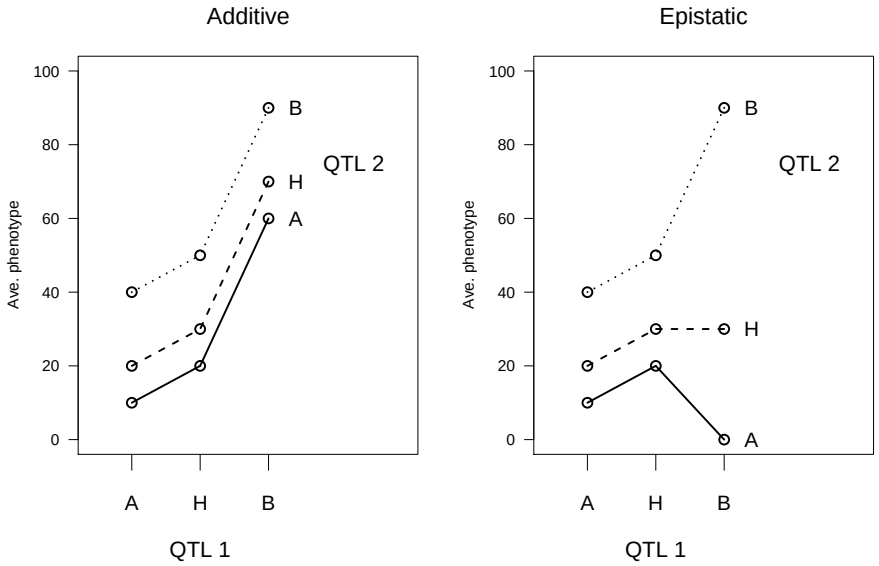
Modelling multiple QTL

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

Epistasis in BC



Epistasis in F_2



2-dim, 2-QTL scan

For all pairs of positions, fit the following models:

$$H_f : y = \mu + \beta_1 q_1 + \beta_2 q_2 + \gamma q_1 q_2 + \epsilon$$

$$H_a : y = \mu + \beta_1 q_1 + \beta_2 q_2 + +\epsilon$$

$$H_1 : y = \mu + \beta_1 q_1 + \epsilon$$

$$H_0 : y = \mu + \epsilon$$

$\log_{10}$ likelihoods:

$l_f(s,t)$ $\quad$ $l_a(s,t)$ $\quad$ $l_1(s)$ $\quad$ l_0

2-dim, 2-QTL scan

LOD scores:

$$\mathrm{LOD}_f(s,t) = l_f(s,t) - l_0$$

$$\mathrm{LOD}_a(s,t) = l_a(s,t) - l_0$$

$$\mathrm{LOD}_i(s,t) = l_f(s,t) - l_a(s,t)$$

$$\mathrm{LOD}_1(s) = l_1(s) - l_0$$

Summaries

Consider each pair of chromosomes, (j,k),
and let $c(s)$ denote the chromosome for position s.

$$M_f(j,k) = \max_{c(s)=j,c(t)=k} \mathrm{LOD}_f(s,t)$$

$$M_a(j,k) = \max_{c(s)=j,c(t)=k} \mathrm{LOD}_a(s,t)$$

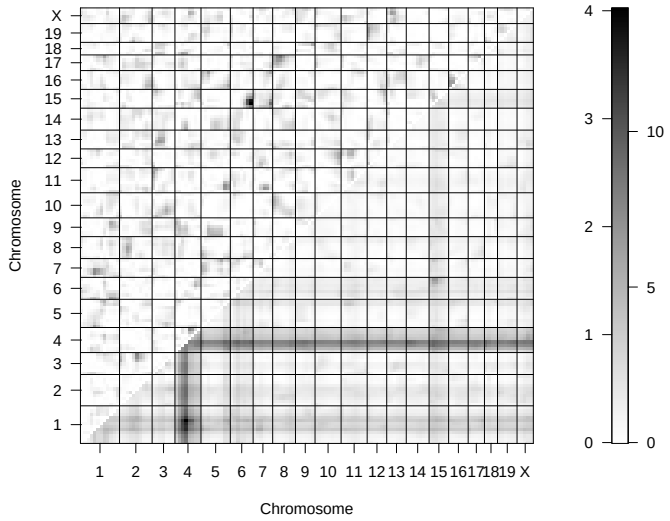
$$M_1(j,k) = \max_{c(s)=j \text{ or } k} \mathrm{LOD}_1(s)$$

$$M_i(j,k) = M_f(j,k) - M_a(j,k)$$

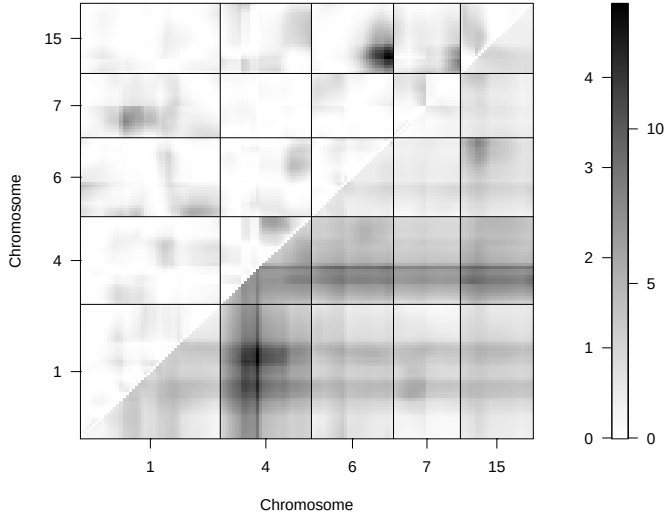
$$M_{fv1}(j,k) = M_f(j,k) - M_1(j,k)$$

$$M_{av1}(j,k) = M_a(j,k) - M_1(j,k)$$

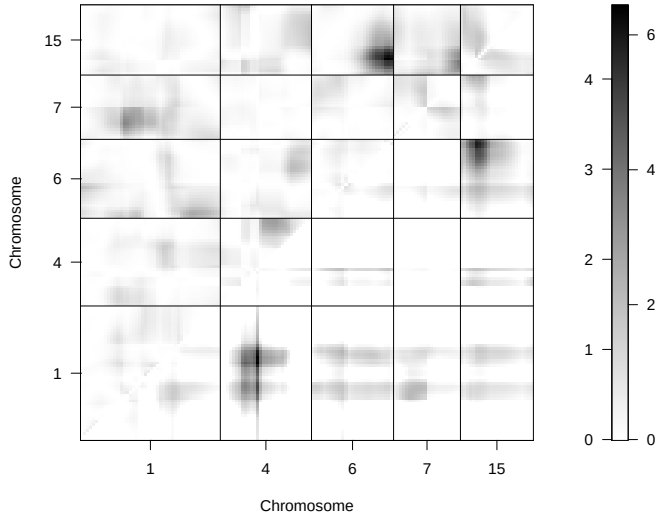
Results: LOD_i and LOD_f



Results: LOD_i and LOD_f

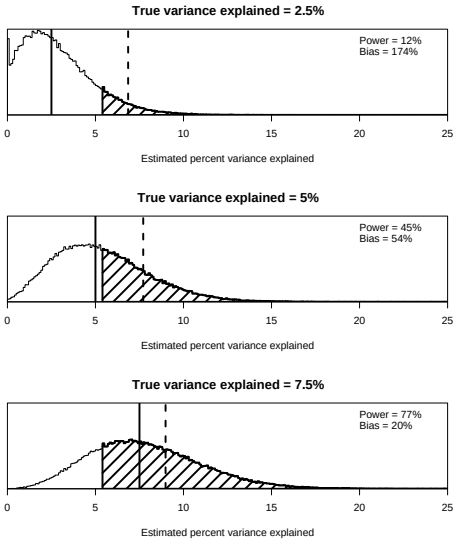


Results: LOD_i and LOD_{fv1}



Selection bias

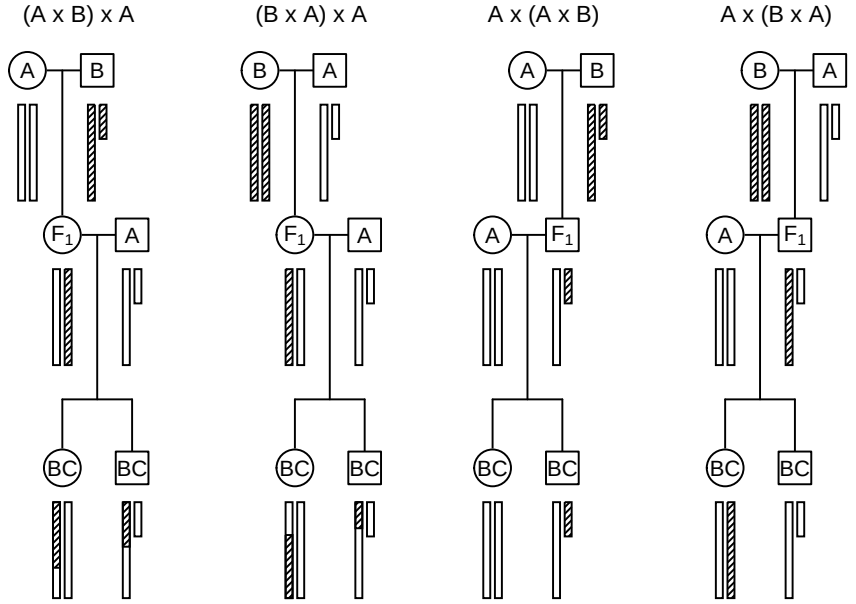
- The estimated effect of a QTL will vary somewhat from its true effect.
- Only when the estimated effect is large will the QTL be detected.
- Among those experiments in which the QTL is detected, the estimated QTL effect will be, on average, larger than its true effect.
- This is selection bias.
- Selection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



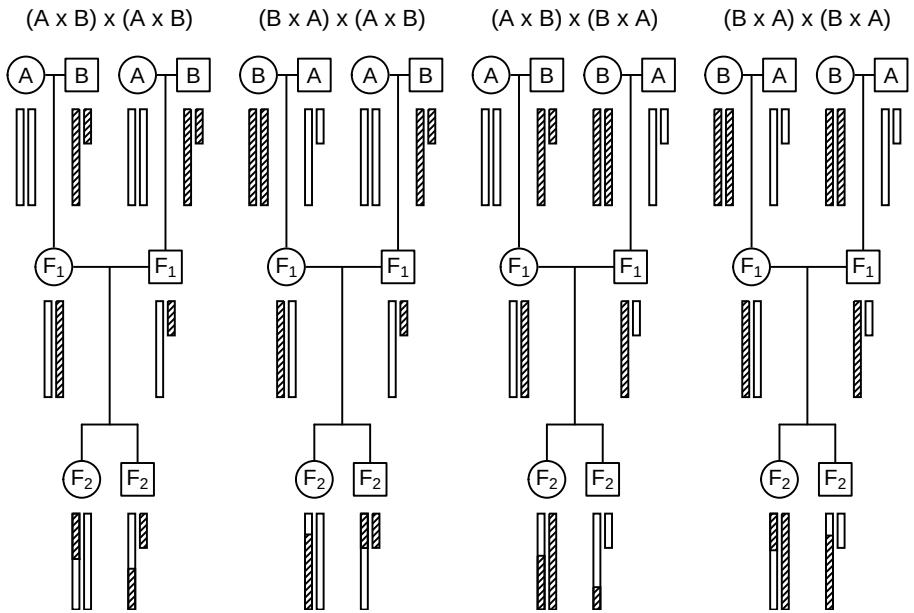
Implications

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

X chr in a backcross

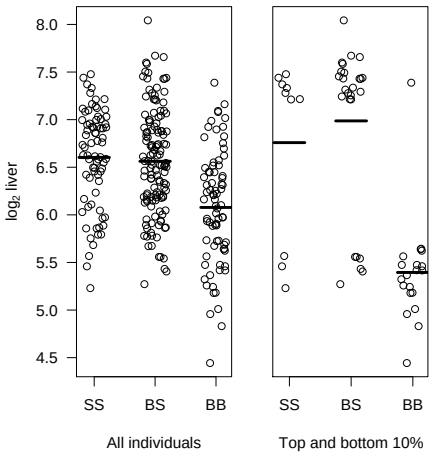


X chr in an intercross



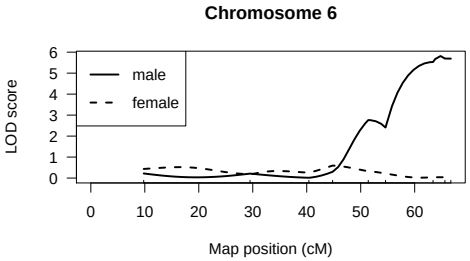
Selective genotyping

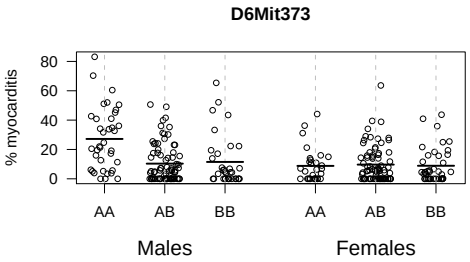
- Save effort by only typing the most informative individuals (say, top & bottom 10%).
- Useful in context of a single, inexpensive trait.
- Tricky to estimate the effects of QTLs: use IM with all phenotypes.
- Can't get at interactions.
- Likely better to also genotype some random portion of the rest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



Covariates

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





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)

Data diagnostics

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

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.
- Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–314

- Feenstra B, Skovgaard IM, Broman KW (2006) Mapping quantitative trait loci by an extension of the Haley-Knott regression method using estimating equations. Genetics 173:2269–2282
- Sen Ś, Churchill GA (2001) A statistical framework for quantitative trait mapping. Genetics 159:371–387
  The multiple imputation method.
- 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.