Introduction to QTL mapping in experimental crosses

Karl W Broman

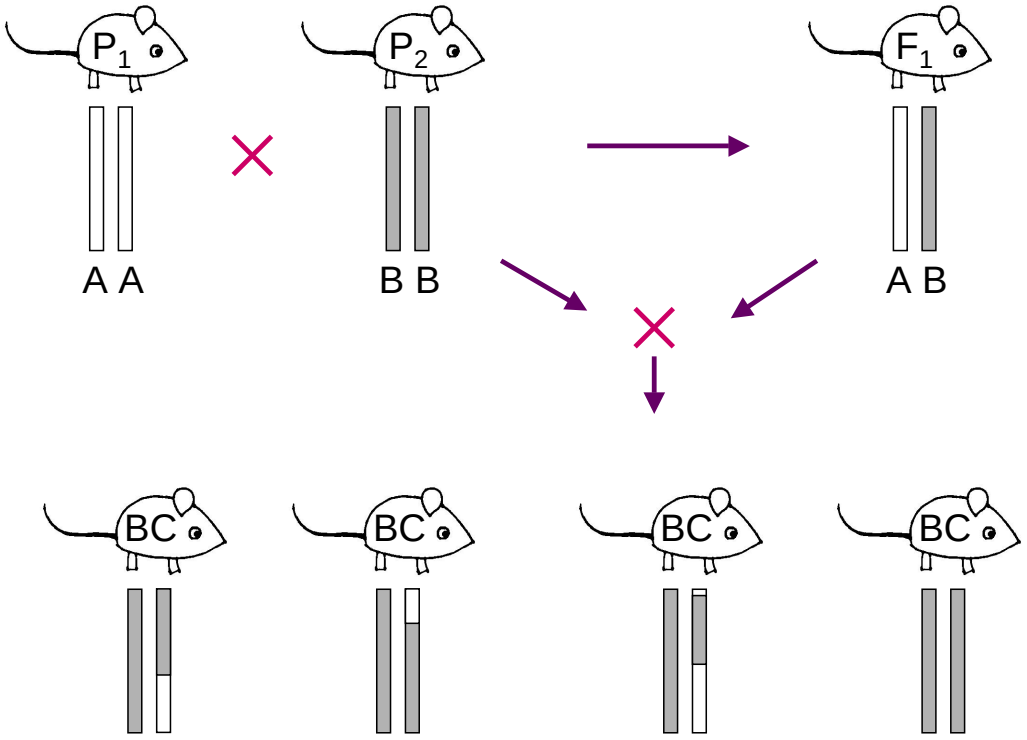
Department of Biostatistics
The Johns Hopkins University

http://biosun01.biostat.jhsph.edu/~kbroman

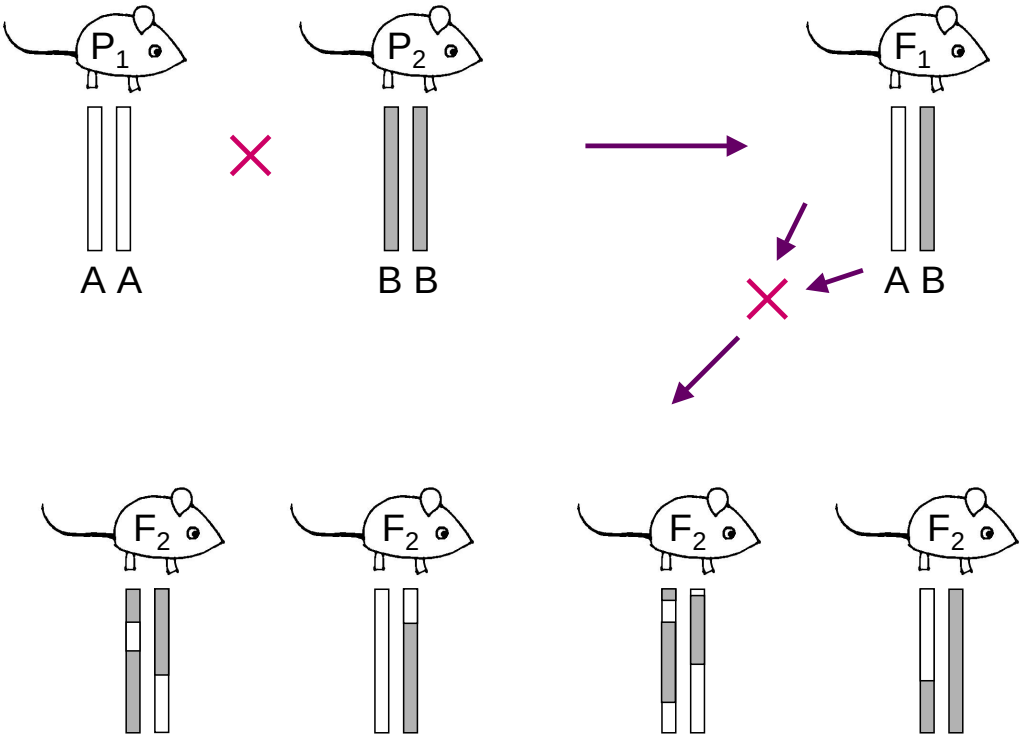
Outline

- Experiments and data
- Models
- ANOVA at marker loci
- Interval mapping
- LOD thresholds
- LOD support intervals
- Power to detect QTLs
- How many markers/mice?
- Selection bias
- Errors in the map
- Genotyping errors
- Selective genotyping
- Covariates
- Non-normal traits

Backcross experiment

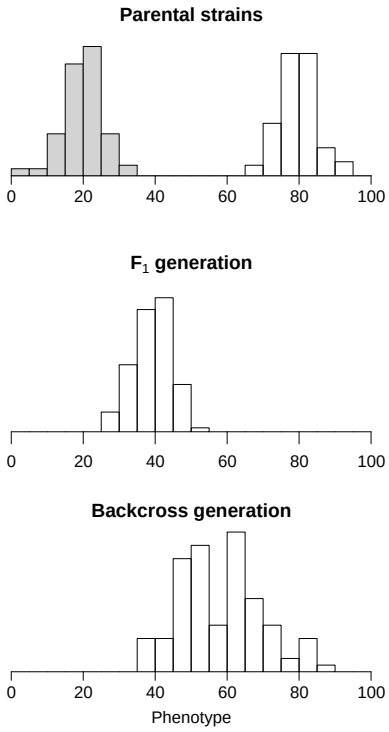


Intercross experiment



Phenotype distributions

- 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 and Goals

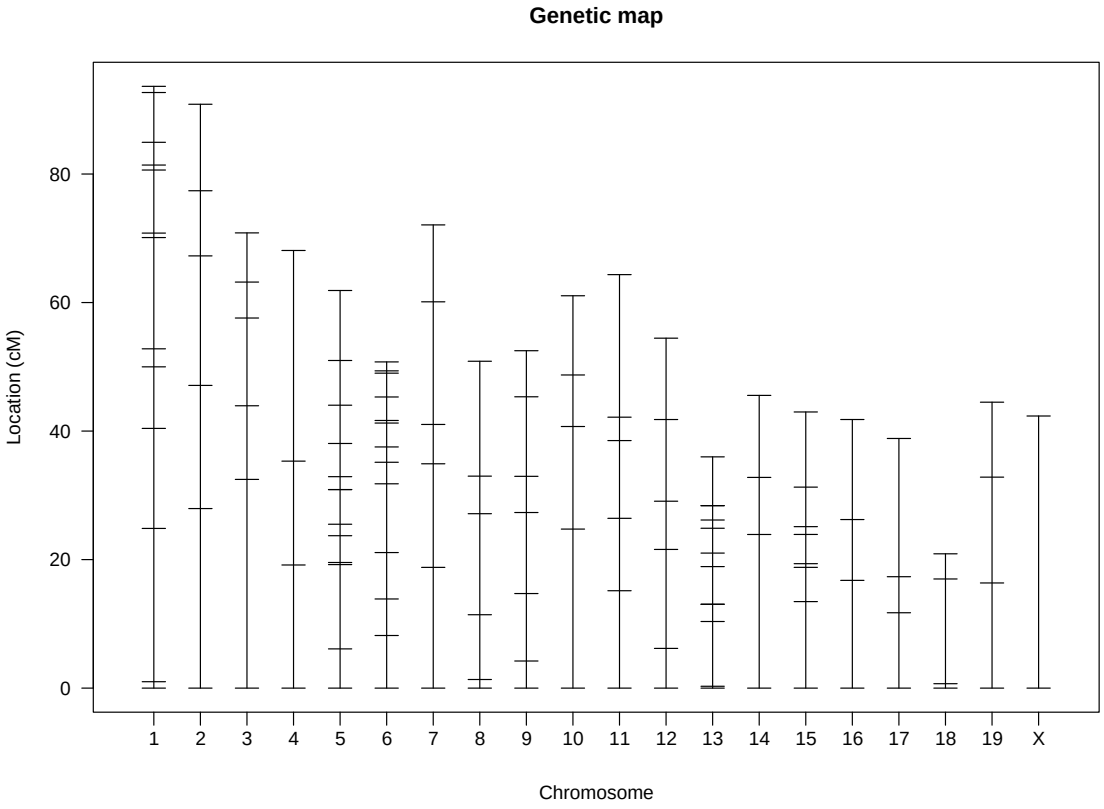
Phenotypes: y_i = phenotype for mouse i

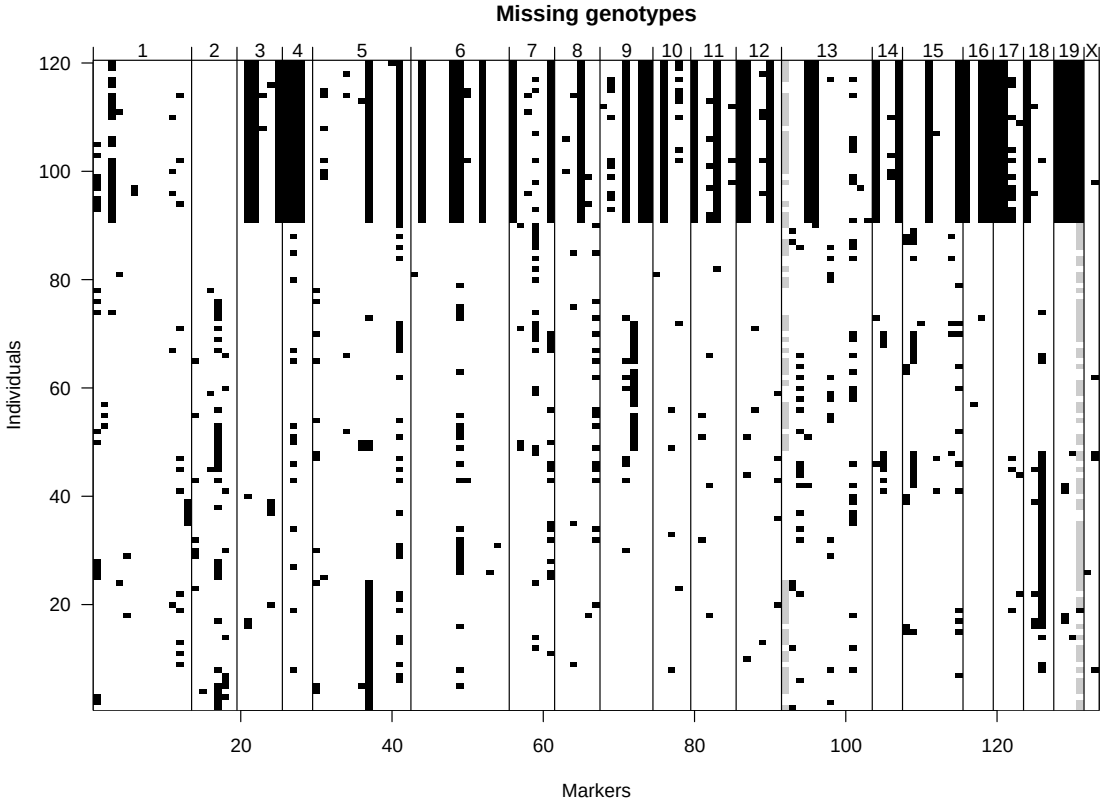
Genotypes: x_{ij} = 1/0 if mouse i is BB/AB at marker j (for a backcross)

Genetic map: Locations of markers

Goals:

- Identify the (or at least one) genomic regions (QTLs) that contribute to variation in the phenotype.
- Form confidence intervals for QTL locations.
- Estimate QTL effects.


Genetic map
Location (cM)
0
20
40
60
80
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome


Missing genotypes
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Individuals
20
40
60
80
100
120
Markers

Models: Recombination

We assume: Mendel's rules
No crossover interference

$\Rightarrow \Pr(x_{ij} = 0) = \Pr(x_{ij} = 1) = 1/2$

Locations of crossovers are according to a Poisson process.

$\Rightarrow \{x_{ij}\}$ form a Markov chain with transition probabilities:

$$\Pr(x_{i,j+1} = 1 | x_{ij} = 0) = \Pr(x_{i,j+1} = 0 | x_{ij} = 1) = r_j$$

r_j = recombination fraction = $(1 - e^{-2d_j})/2$

d_j is the genetic distance in Morgans.

Markov chain: $\Pr(x_{i,j+1} | x_{ij}, x_{i,j-1}, x_{i,j-2}, \ldots) = \Pr(x_{i,j+1} | x_{ij})$

Models: Genotype $\longleftrightarrow$ Phenotype

Let y = phenotype
g = whole genome genotype

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

$$\mathrm{E}(y|g) = \mu_{g_1,\ldots,g_p} \qquad \mathrm{var}(y|g) = \sigma^2_{g_1,\ldots,g_p}$$

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

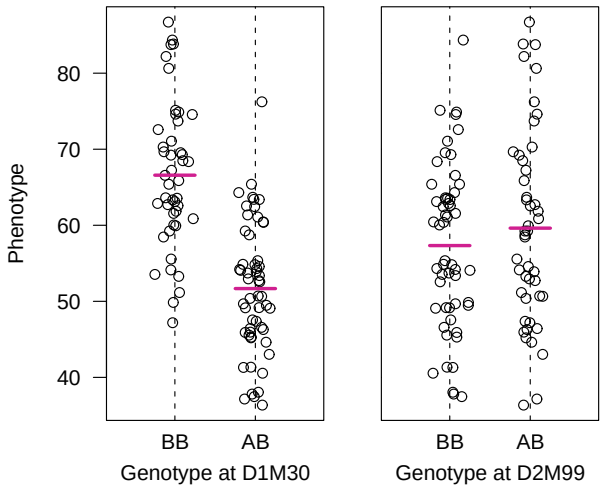
Normally distributed residual variation: $y|g \sim N(\mu_g, \sigma^2)$.

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

Epistasis: Any deviations from additivity.

The simplest method: ANOVA

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



Effect at a marker

Consider the case of a single QTL with effect $\Delta = \mu_{BB} - \mu_{AB}$.

Consider a marker linked to the QTL, with r = recomb. frac.

Of individuals with marker genotype BB, mean phenotype is:

$$\mu_{BB}\,(1-r) + \mu_{AB}\,r = \mu_{BB} - r\,\Delta$$

Of individuals with marker genotype AB, mean phenotype is:

$$\mu_{AB}\,(1-r) + \mu_{BB}\,r = \mu_{AB} + r\,\Delta$$

Difference: $(\mu_{BB} - r\,\Delta) - (\mu_{AB} + r\,\Delta) = \Delta\,(1-2r)$

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.

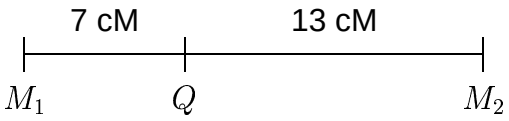
Interval mapping (IM)

Lander & Botstein (1989)

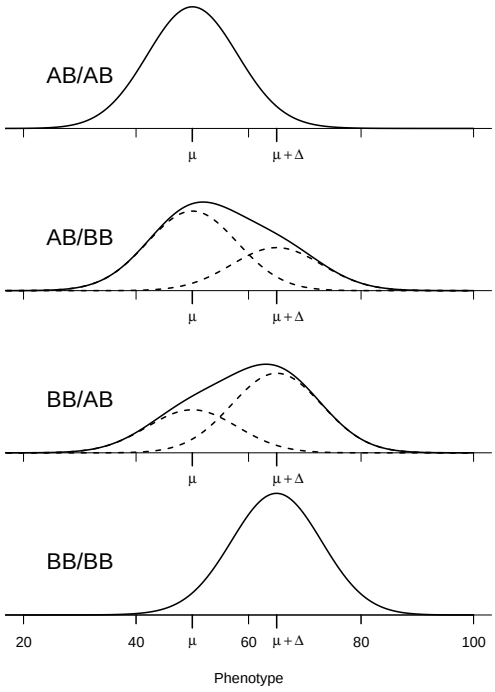
- Assume a single QTL model.
- Each position in the genome, one at a time, is posited as the putative QTL.
- Let $z = 1/0$ if the (unobserved) QTL genotype is BB/AB. Assume $y = \mu + \Delta\, z + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$.
- Given genotypes at linked markers, $y \sim$ mixture of normal dist'ns with mixing proportion $\Pr(z = 1|\text{marker data})$:

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

The normal mixtures



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

$y_i|z_i \sim N(\mu + \Delta\, z_i, \sigma^2)$

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

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

Log likelihood: $l(\mu, \Delta, \sigma) = \sum_i \log \Pr(y_i|\text{marker data}, \mu, \Delta, \sigma)$

Maximum likelihood estimates (MLEs) of μ, Δ, σ:

values for which $l(\mu, \Delta, \sigma)$ is maximized.

EM algorithm

Dempster et al. (1977)

E step:

Let $w^{(k+1)} = \Pr(z_i = 1 | y_i, \text{marker data}, \hat{\mu}^{(k)}, \hat{\Delta}^{(k)}, \hat{\sigma}^{(k)})$

$$= \frac{p_i f(y_i; \hat{\mu}^{(k)} + \hat{\Delta}^{(k)}, \hat{\sigma}^{(k)})}{p_i f(y_i; \hat{\mu}^{(k)} + \hat{\Delta}^{(k)}, \hat{\sigma}^{(k)}) + (1-p_i) f(y_i; \hat{\mu}^{(k)}, \hat{\sigma}^{(k)})}$$

M step:

Let $\hat{\mu}^{(k+1)} = \sum_i y_i (1 - w_i^{(k+1)}) / \sum_i (1 - w_i^{(k+1)})$

$\hat{\Delta}^{(k+1)} = \sum_i y_i w_i^{(k+1)} / \sum_i w_i^{(k+1)} - \hat{\mu}^{(k+1)}$

$\hat{\sigma}^{(k+1)} =$ a bit complicated

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.

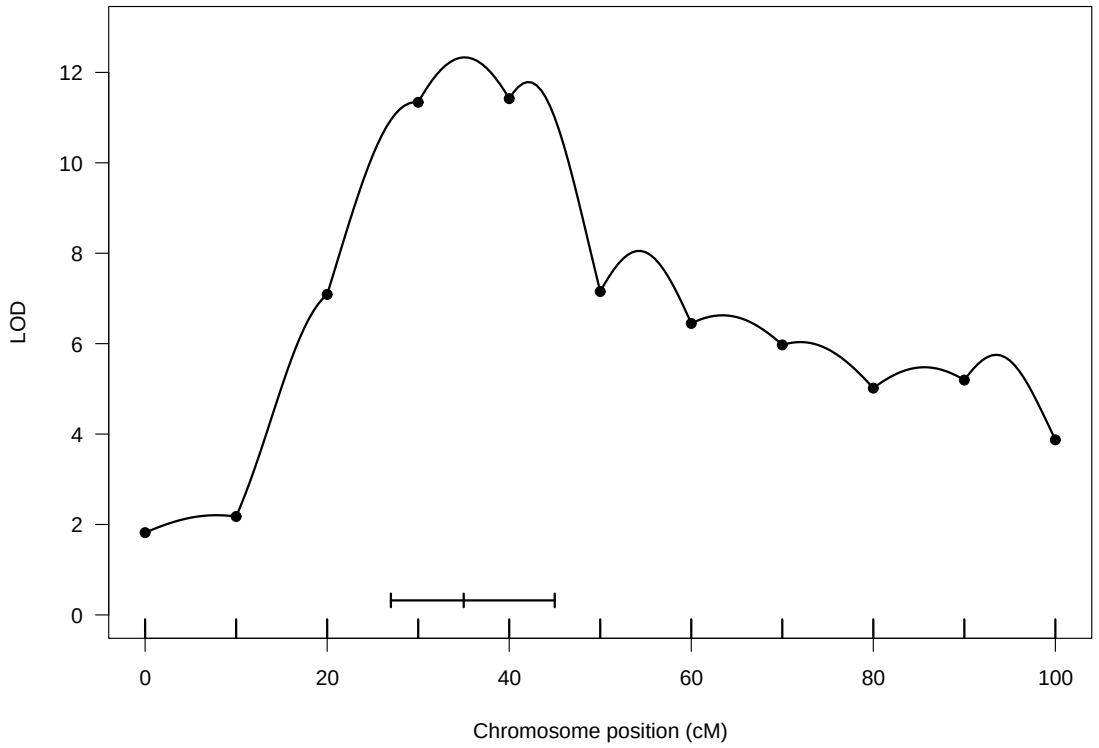
LOD$(z) =$ $\log_{10}$ likelihood ratio comparing the hypothesis of a QTL at position z versus that of no QTL

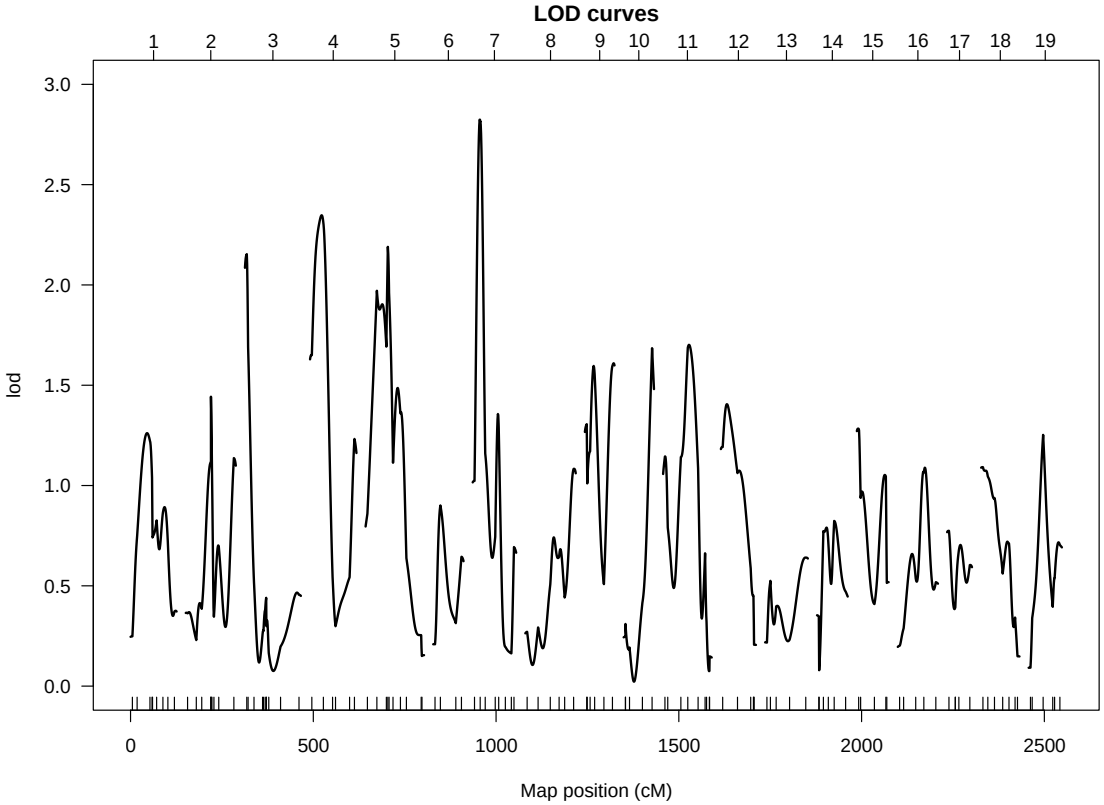
$$= \log_{10} \left\{ \frac{\Pr(y | \text{QTL at } z, \hat{\mu}_z, \hat{\Delta}_z, \hat{\sigma}_z)}{\Pr(y | \text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\}$$

$\hat{\mu}_z, \hat{\Delta}_z, \hat{\sigma}_z$ are the MLEs, assuming a single QTL at position z.

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.

Multiple QTL methods

Why consider multiple QTLs at once?

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

LOD thresholds

Large LOD scores indicate evidence for the presence of a QTL.

Q: How large is large?

$\rightarrow$ We consider the distribution of the LOD score under the null hypothesis of no QTL.

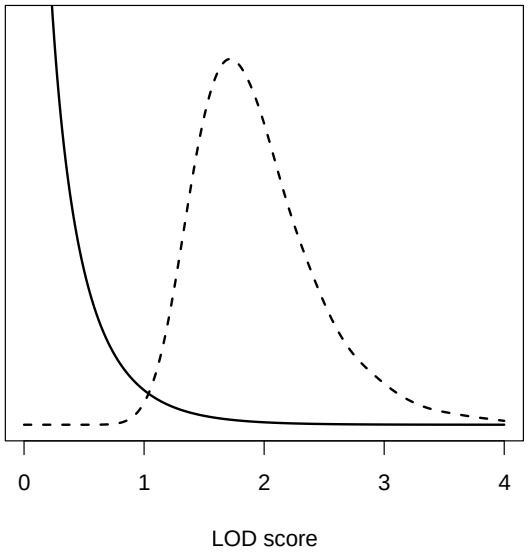
Key point: We must make some adjustment for our examination of multiple putative QTL locations.

$\rightarrow$ We seek the distribution of the *maximum* LOD score, genome-wide. The 95th %ile of this distribution serves as a genome-wide LOD threshold.

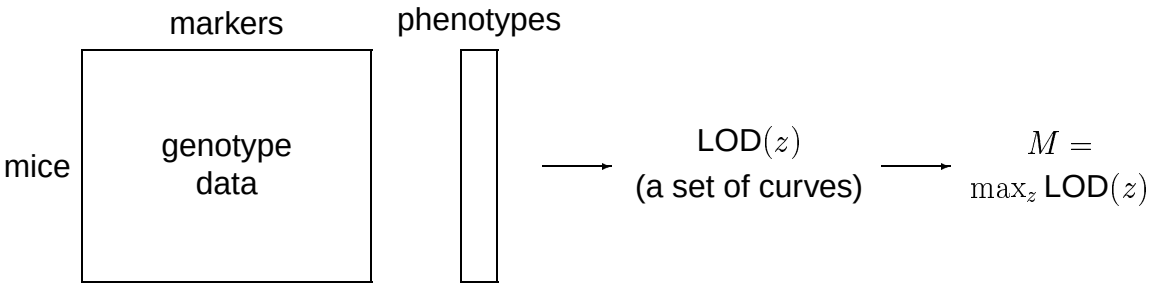
Estimating the threshold: simulations, analytical calculations, permutation (randomization) tests.

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

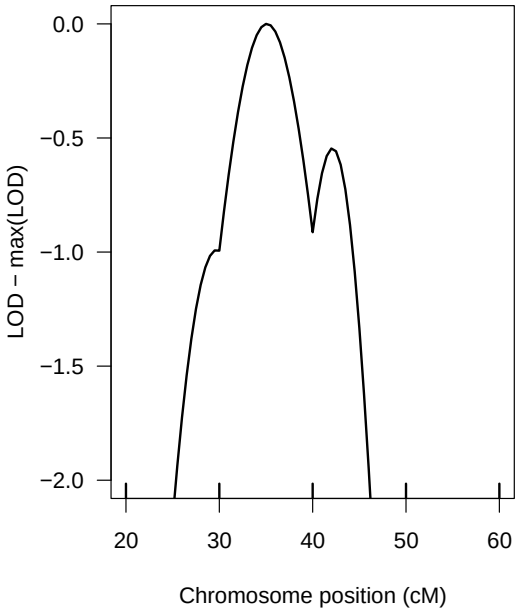


- Permute/shuffle the phenotypes; keep the genotype data intact.
- Calculate LOD$^\star(z) \longrightarrow M^\star = \max_z \text{LOD}^\star(z)$
- We wish to compare the observed M to the distribution of $M^\star$.
- $\Pr(M^\star \geq M)$ is a genome-wide P-value.
- The 95th %ile of $M^\star$ is a genome-wide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

LOD support intervals

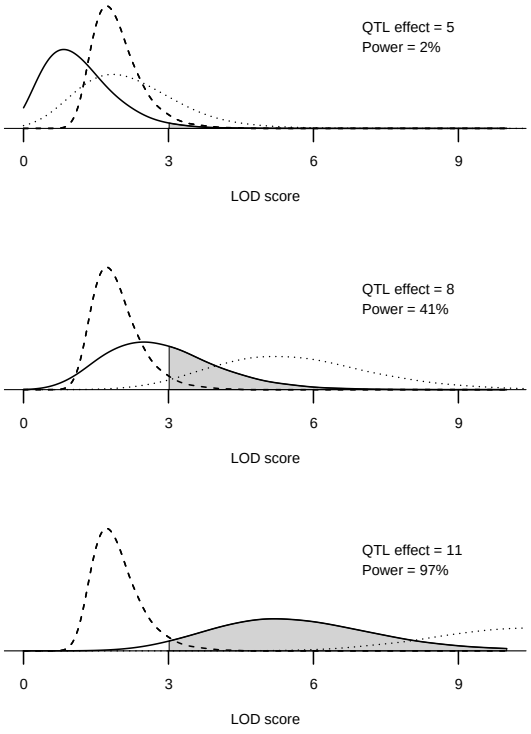
L-LOD support interval

- Chromosomal region for which the LOD score is within L of its maximum.
- Generally $L = 1$ or 2; I prefer $L = 1.5$.
- Plot of LOD $- \max\{\text{LOD}\}$ depicts evidence for QTL location.



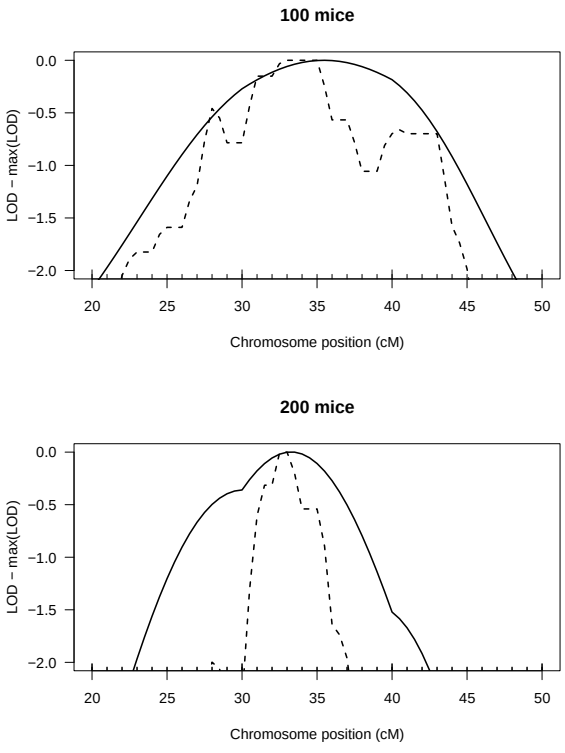
Power to detect QTLs

- The power to detect a QTL is the chance that its LOD score exceeds the genome-wide threshold.
- Power depends on
  - Size of the QTL effect.
  - Number of progeny.
  - Type of cross.
  - Density of markers.
  - Stringency of the LOD threshold.
- At right:
  - Dashed curve: dist'n of max LOD under null hypothesis.
  - Solid curve: dist'n of LOD score at QTL, with $n = 100$.
  - Dotted curve: dist'n of LOD score at QTL, with $n = 200$.



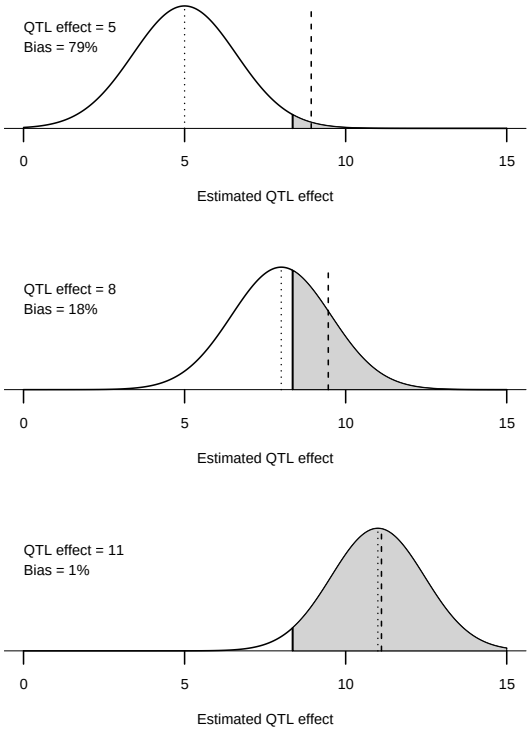
How many markers/mice?

- At right:
  - Top: $n = 100$
  - Bottom: $n = 200$
  - Solid: 10 cM spacing
  - Dashed: 1 cM spacing
- More mice:
  - More recombination breakpoints.
  - Reduced sampling variation.
- More markers:
  - More detailed genotype information.
  - Not necessarily increased precision (depends on number of mice, size of QTL effect, and luck).
- Note: The figures should be taken with a grain of salt.



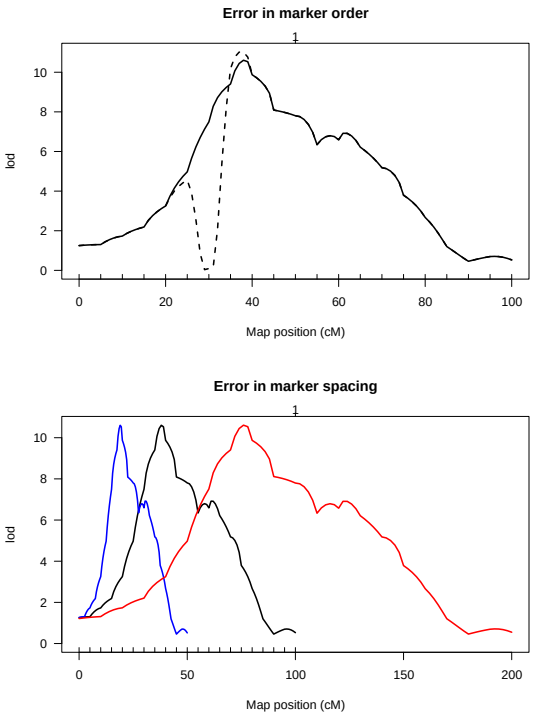
Selection bias

- 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.



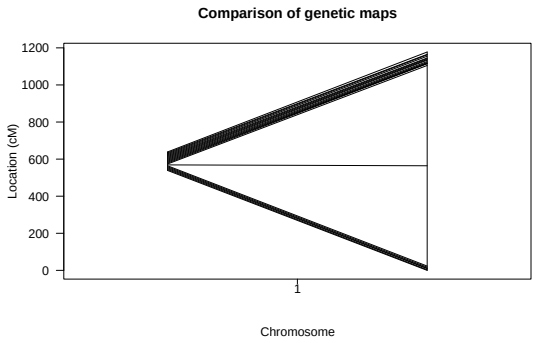
The genetic map: effects of errors

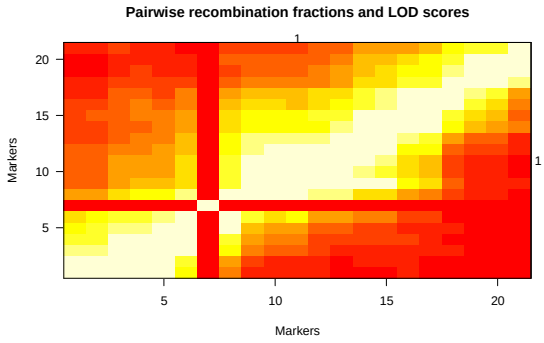
- Marker order
  - Causes wiggly LOD curves.
  - Shouldn't completely eliminate a signal.
- Map distances
  - Doesn't seem to make much difference.
  - Makes a big difference in perceived length of LOD support intervals.
- Greater effects of errors with appreciable missing data



The genetic map: finding problems

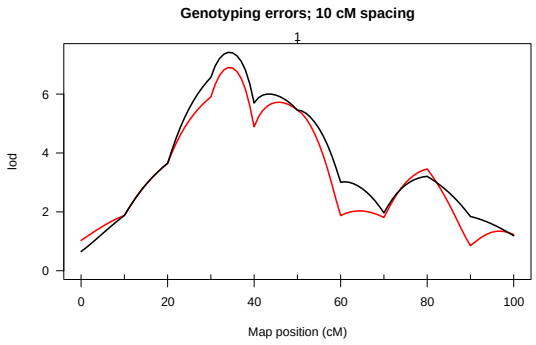
- Consider:
  - Estimated genetic map
  - Pairwise recombination fractions
- Misplaced markers cause:
  - Big gaps in the map
  - Large recombination fractions

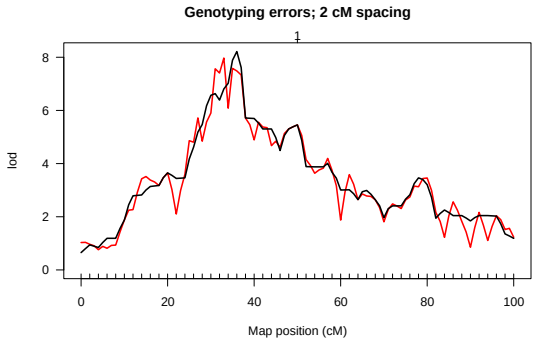




Genotyping errors: effects

- With genotyping errors, individuals are placed in the wrong genotype group.
- With widely spaced markers, there is little effect.
- With dense markers, errors make the LOD curve have more dips.





Identifying genotyping errors

- Look for tight double crossovers.

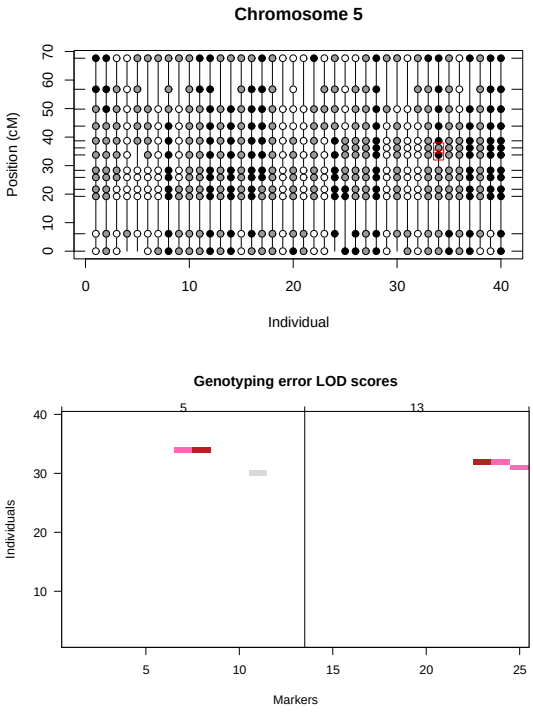
  (Crossover interference is often strong.)

- Error LOD scores

  (Lincoln & Lander 1992)

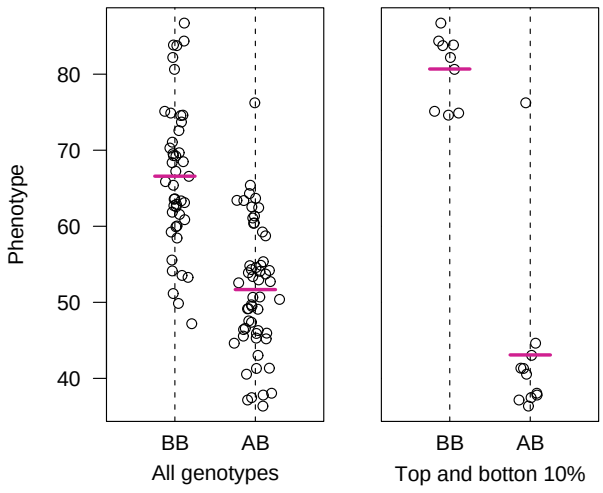
  - Model for genotyping errors.
  - Assumed error rate.
  - Assumption of no interference.
  - At marker j in mouse i, LOD =

$$\log_{10}\left\{\frac{\Pr(g_{ij} \text{ in error} \mid \text{marker data}, \epsilon)}{\Pr(g_{ij} \text{ correct} \mid \text{marker data}, \epsilon)}\right\}$$



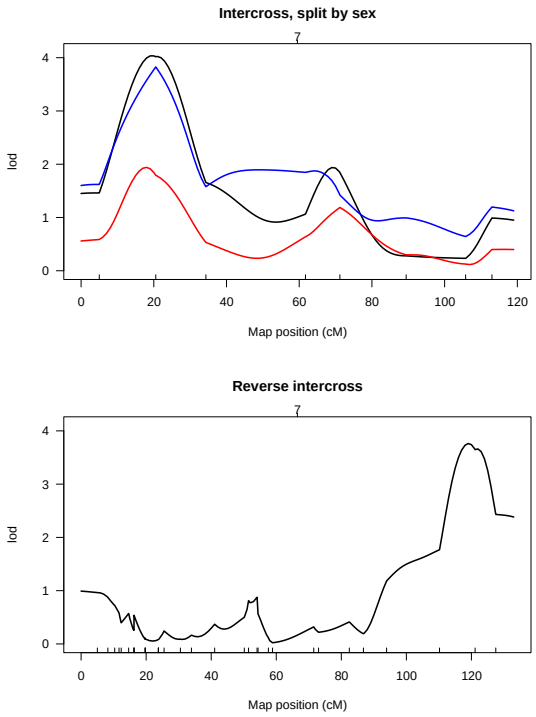
Selective genotyping

- 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 × 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)

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. A plot of the LOD curve, re-centered so that its maximum is at 0, is a valuable tool for depicting evidence for QTL location.
- Once you've achieved a 10 cM marker spacing, more mice will probably be more important than more markers. But this depends on the number of mice, the size of the QTL effect, and luck.

Summary II

- Estimates of QTL effects are subject to selection bias. Such estimated effects are often too large.
- 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.