

QTL Mapping II:

Introduction

Karl W Broman

Department of Biostatistics
Johns Hopkins University

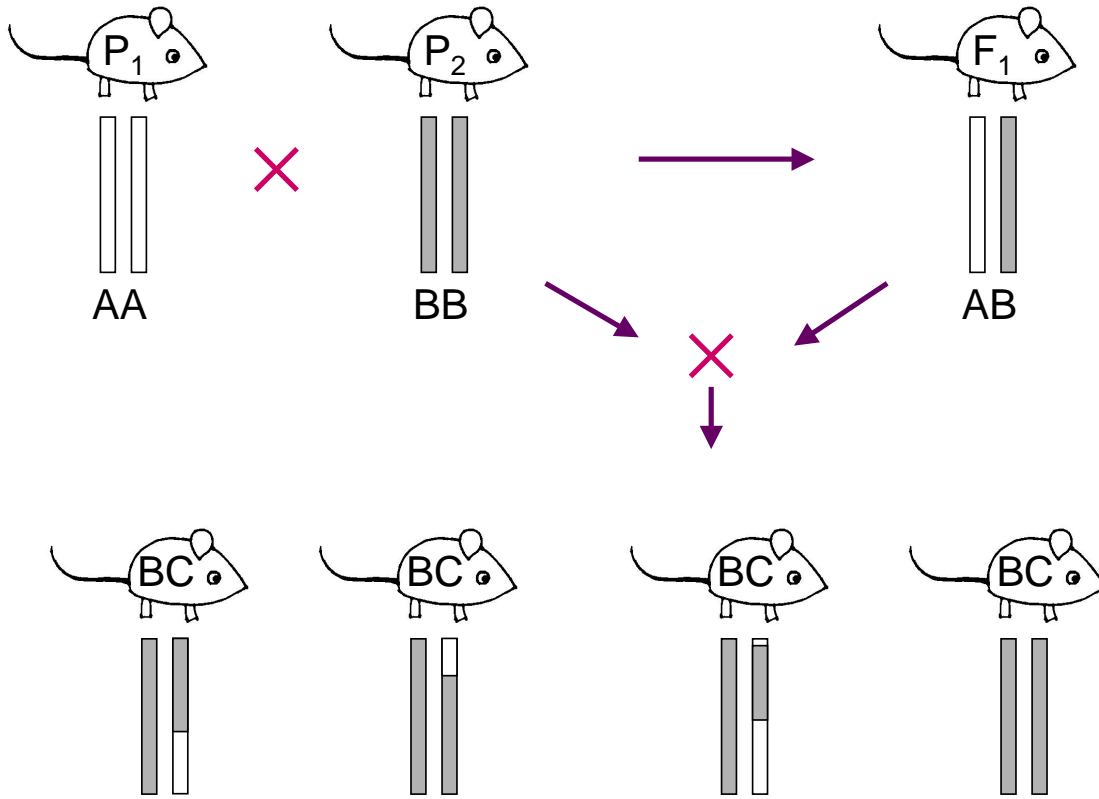
kbroman@jhsph.edu

www.biostat.jhsph.edu/~kbroman

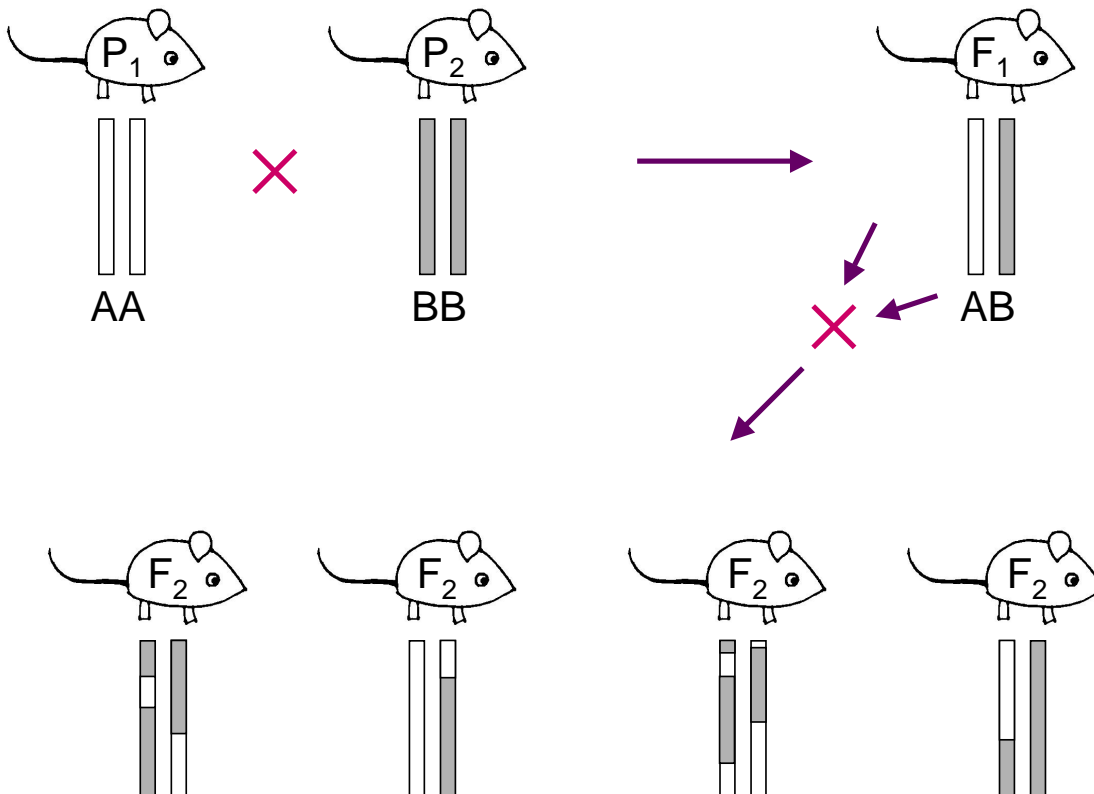
Outline

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

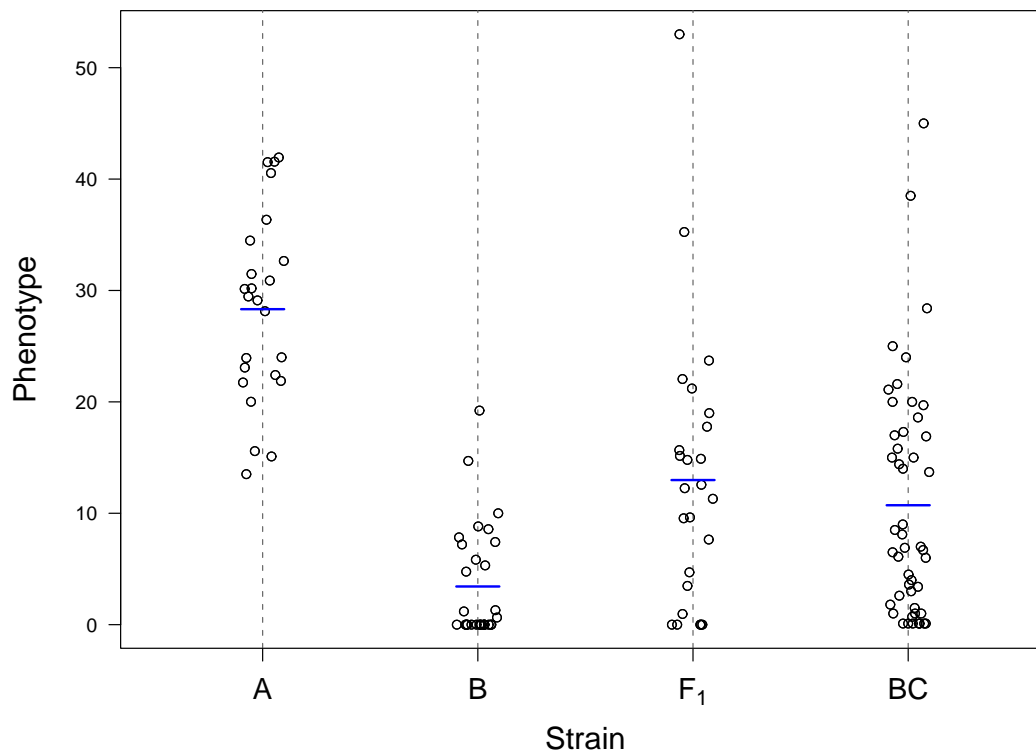
Backcross experiment



Intercross experiment



Trait distributions



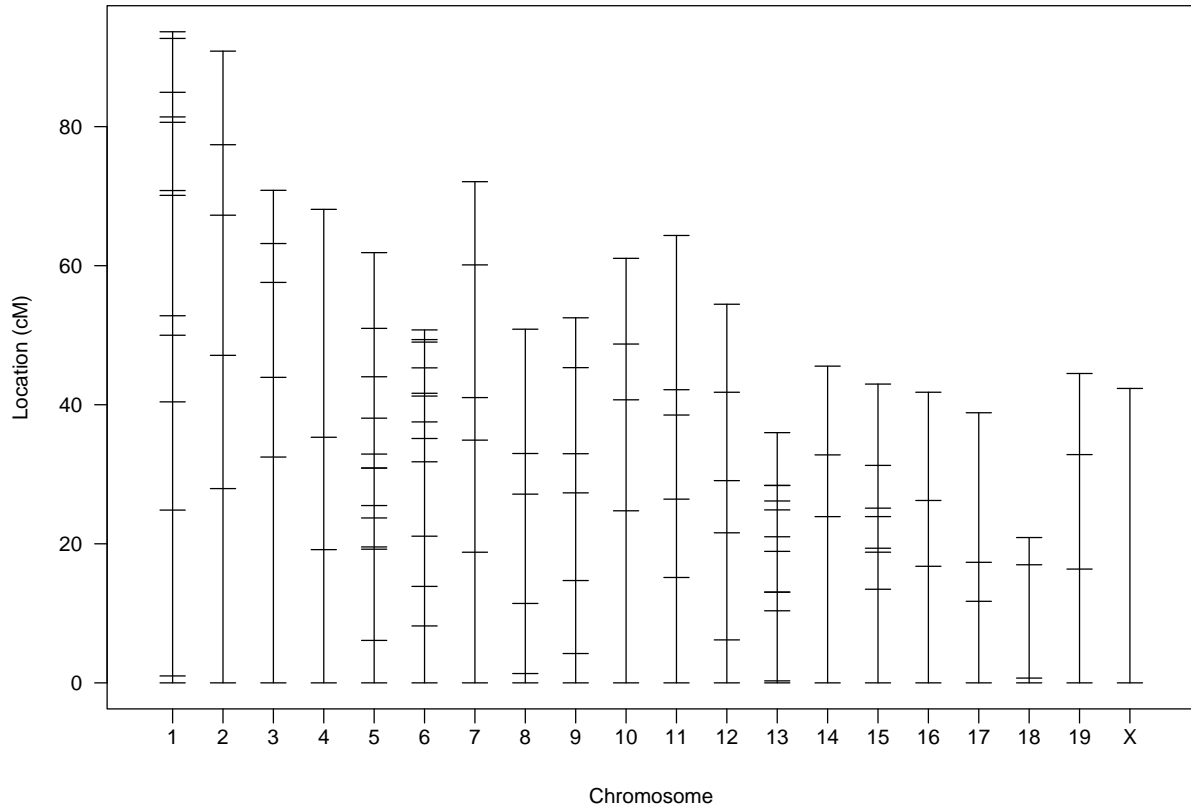
Data

Phenotypes: y_i = trait value for individual i

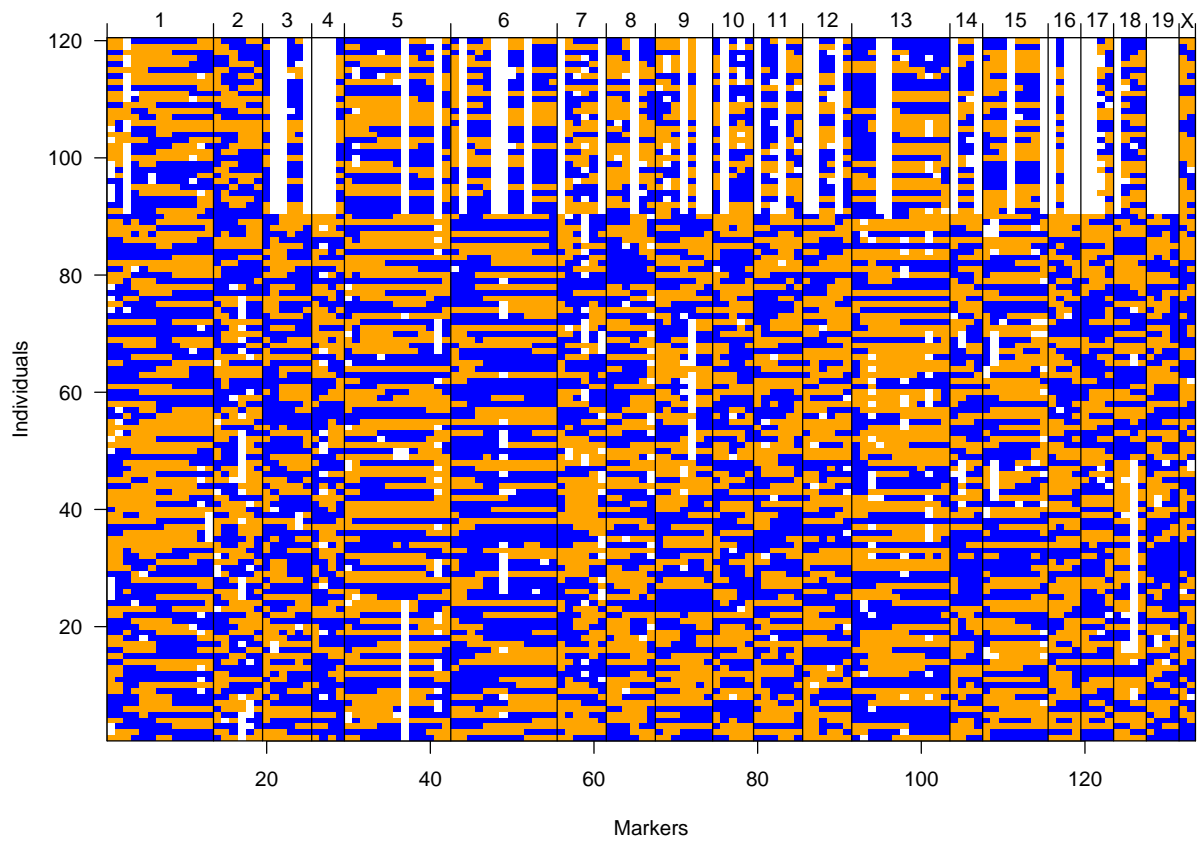
Genotypes: x_{ij} = 0/1 if mouse i is BB/AB at marker j
(or 0/1/2, in an intercross)

Genetic map: Locations of markers

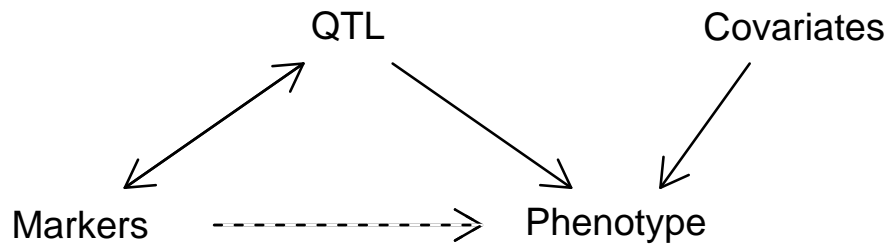
Genetic map



Genotype data



Statistical structure



The missing data problem:

Markers \longleftrightarrow QTL

The model selection problem:

QTL, covariates \longrightarrow phenotype

Goals

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

Models: Recombination

We assume no crossover interference.

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

Crossover locations



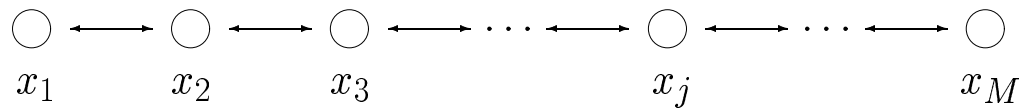
Crossover locations follow a Poisson process

1. Number of XOs \sim Poisson($L / 100$)
Locations of XOs (given number) \sim iid uniform
2. Inter-XO distances are iid exponential(mean = 100 cM)

Note: In reality, crossovers are more spread out, but this no interference (NI) model is extremely convenient.

Marker genotypes

Under the no interference model,
the marker genotypes follow a **Markov chain**.



$$\Pr(x_{j+1} \mid x_j, x_{j-1}, \dots, x_1) = \Pr(x_{j+1} \mid x_j)$$

The **past** and **future** are **conditionally independent**,
given the **present**.

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)

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

Models: Genotype \longleftrightarrow Phenotype

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

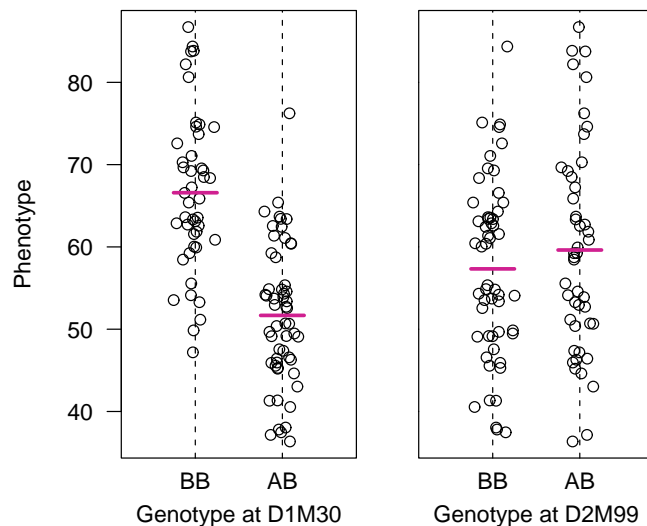
Normally distributed residual variation: $y|g \sim N(\mu_g, \sigma^2)$.

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.

The simplest method: ANOVA

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



LOD score = \log_{10} likelihood ratio comparing single-QTL model to “no QTL anywhere.”

ANOVA at marker loci

Advantages

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

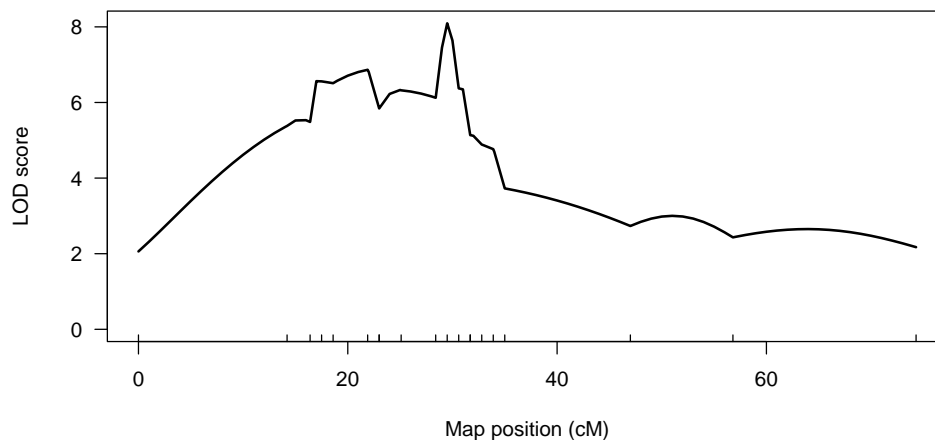
Disadvantages

- 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



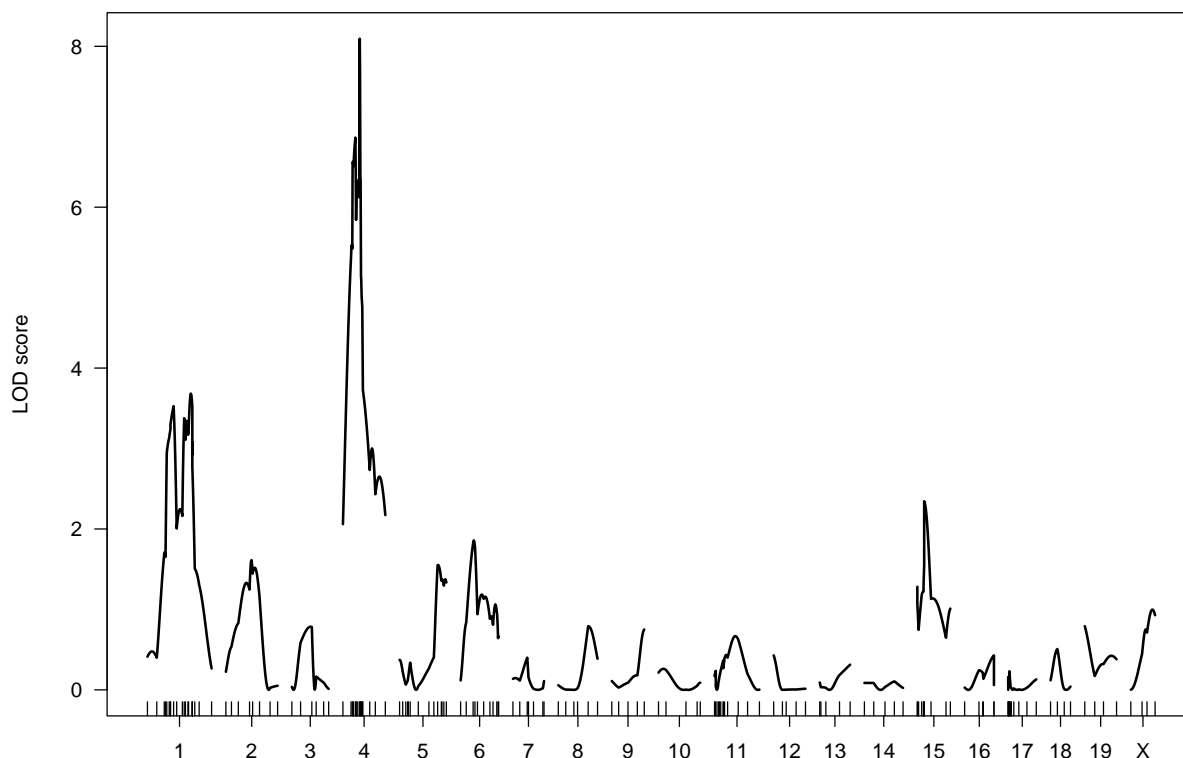
Interval mapping

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.**



LOD thresholds

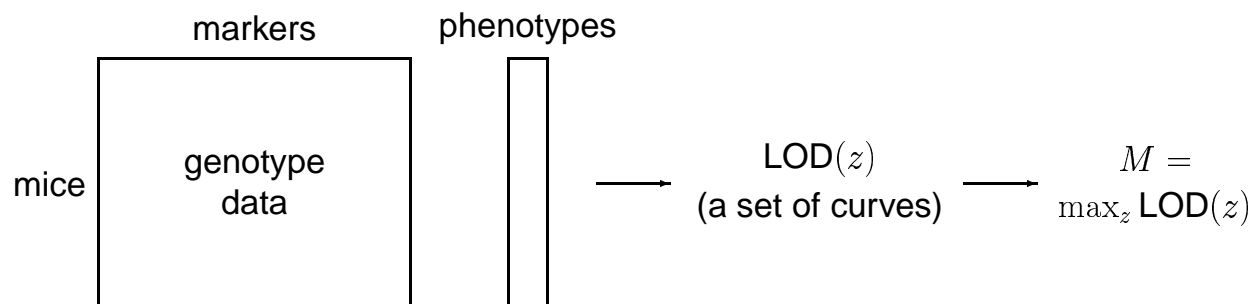
Large LOD scores indicate evidence for the presence of a QTL

Question: How large is large?

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

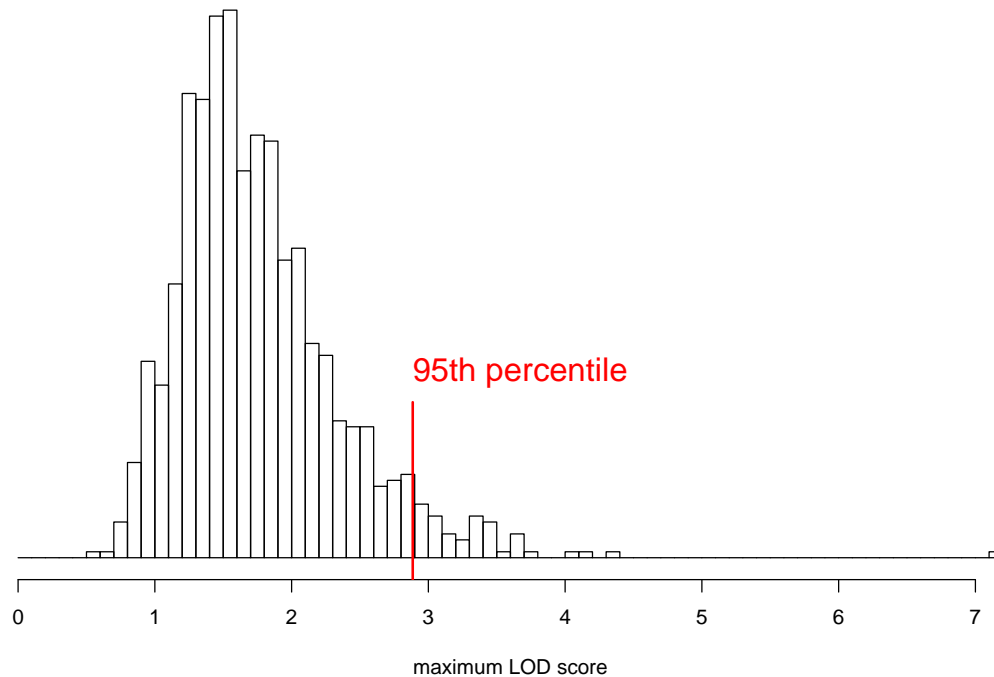
- Derivation:**
- Analytical calculations (L & B '89)
 - Simulations (L & B '89)
 - Permutation tests (Churchill & Doerge 1994)

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



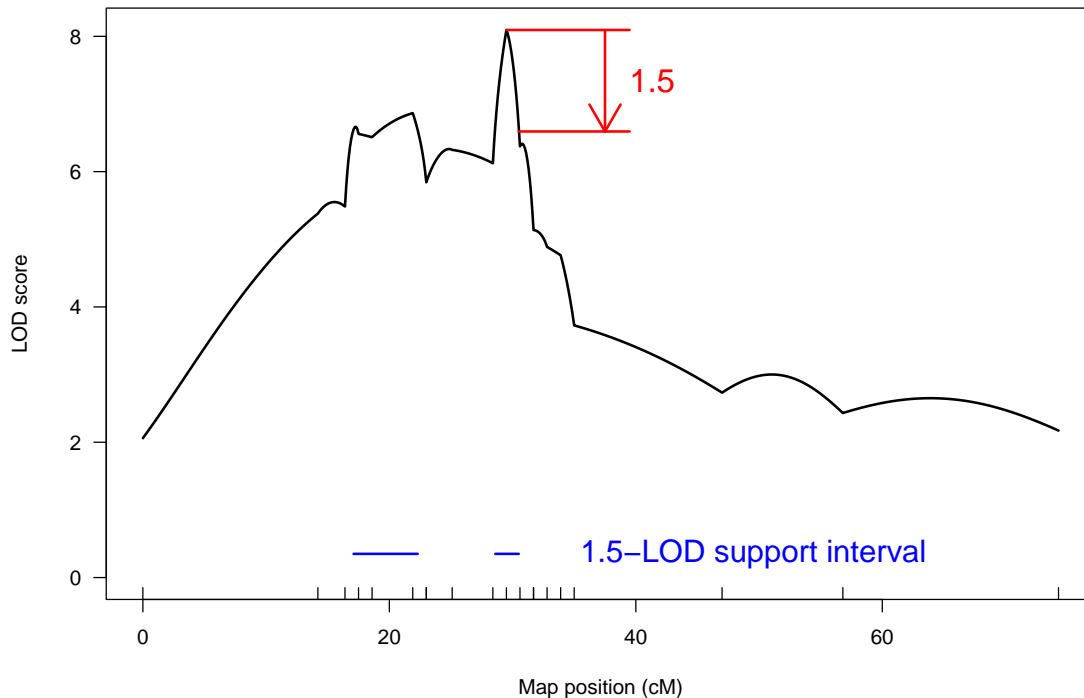
Confidence intervals for QTL location

Confidence interval indicates plausible location of a QTL.

Derivation:

- **LOD support intervals**
(I prefer 1.5-LOD intervals.)
- **Bootstrap**
(Resample individuals with replacement)

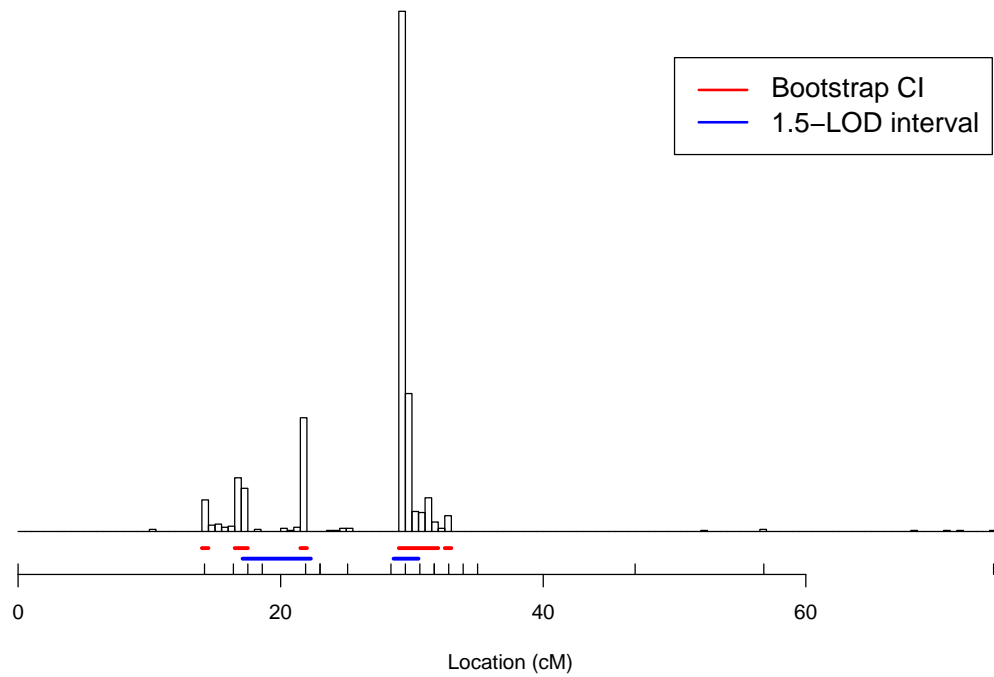
1.5-LOD support interval



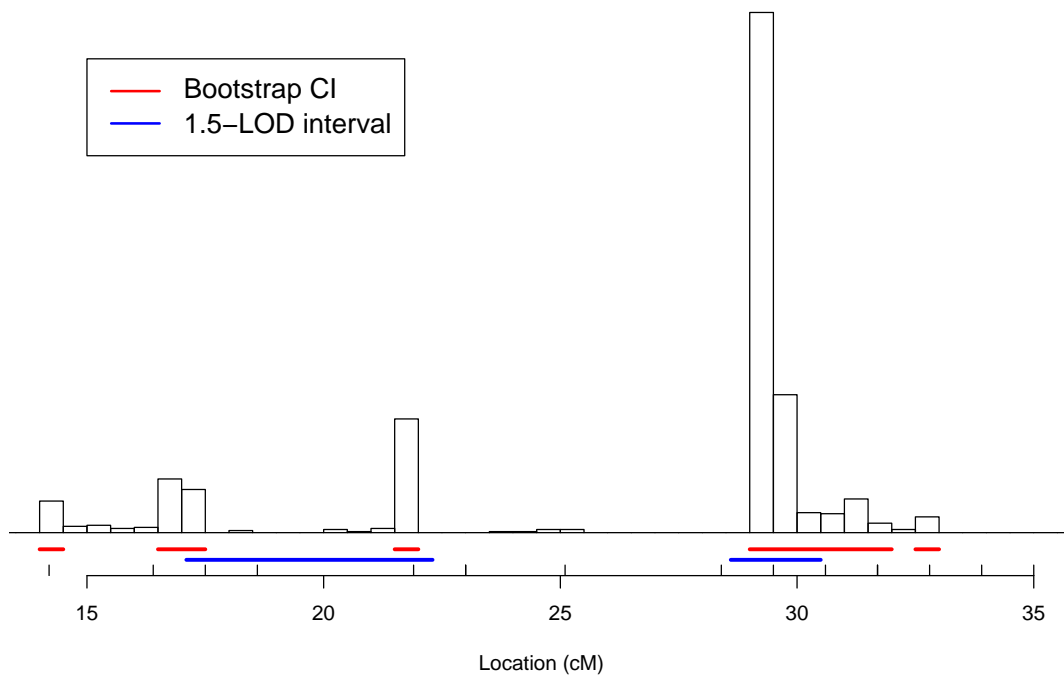
Bootstrap-based confidence intervals

- **Consider** a set of n individuals (e.g., $n = 124$)
- **Sample, with replacement**, n individuals
 - Some individuals duplicated
 - Some individuals missing
- Using the sampled individuals, **re-calculate the LOD curve** for the chromosome of interest
- **Identify the location of the maximum LOD score**
- **Repeat many times** (e.g., 250)
- **Bootstrap CI** = (2.5 %ile to 97.5 %ile) of locations of maximum LOD.

Example of bootstrap CI

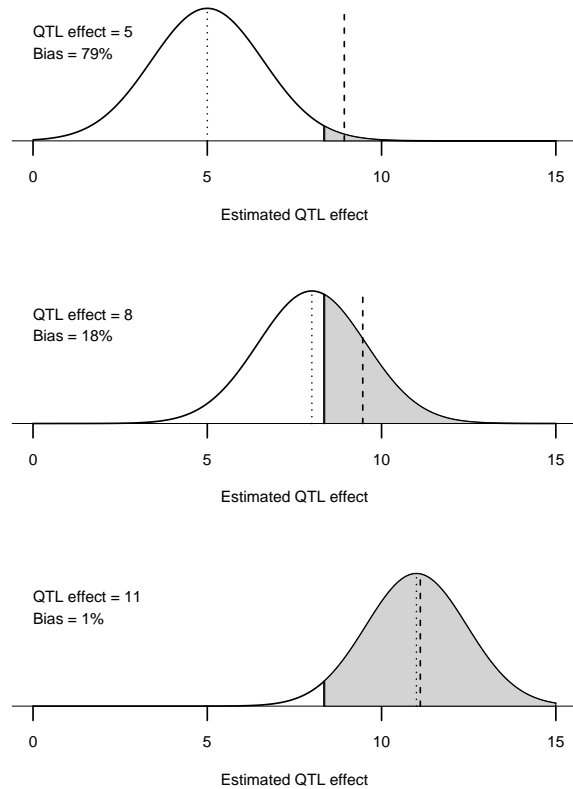


A closer view



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 of selection bias

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

Multiple QTL methods

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

Data structure

y = phenotypes

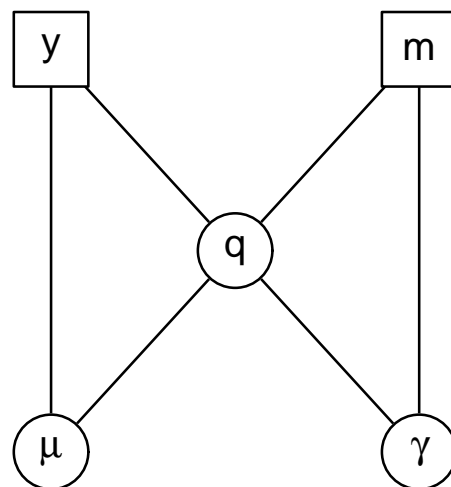
m = observed marker genotypes

q = unobserved QTL genotypes

μ = model parameters

γ = QTL locations

H = QTL model



Likelihood vs. Bayes

$$l(H, \mu, \gamma | y, m) = \sum_q \Pr(y, q | m, \mu, \gamma, H)$$

Likelihood:

$$l(H | y, m) = \max_{\mu, \gamma} l(H, \mu, \gamma | y, m)$$

Bayes:

$$\Pr(H, \mu, \gamma | y, m) \propto l(H, \mu, \gamma | y, m) \Pr(\mu, \gamma | H) \Pr(H)$$

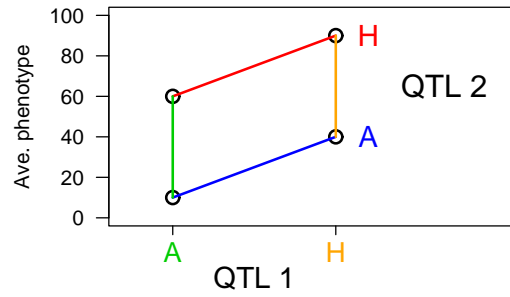
$$\Pr(H | y, m) = \iint \Pr(H, \mu, \gamma | y, m) d\mu d\gamma$$

Model selection

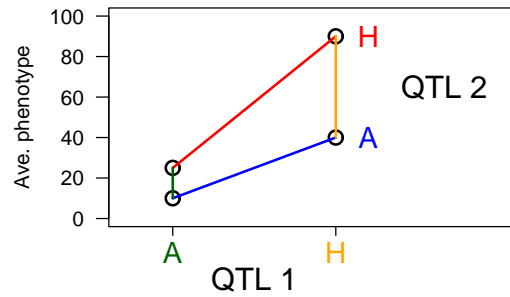
- **Select class of models**
 - Additive models
 - Add've plus pairwise interactions
 - Regression trees
- **Search model space**
 - Forward selection (FS)
 - Backward elimination (BE)
 - FS followed by BE
 - MCMC
- **Compare models**
 - Penalized likelihood (e.g., AIC, BIC)
 - Sequential permutation tests
 - Bayes (posterior probability)
- **Assess performance**
 - Maximize no. QTLs found; control false positive rate

Epistasis in a backcross

Additive QTLs

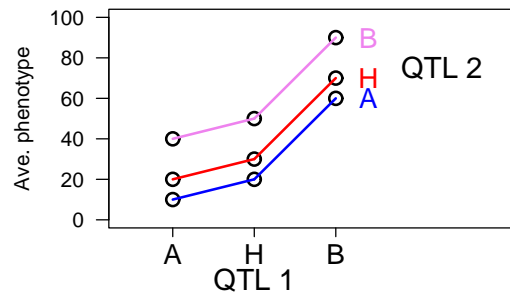


Interacting QTLs

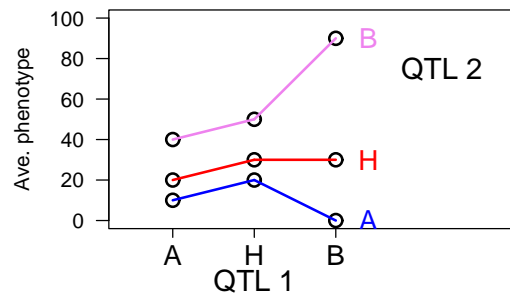


Epistasis in an intercross

Additive QTLs



Interacting QTLs



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: $X_{A \cdot B} Y_B$

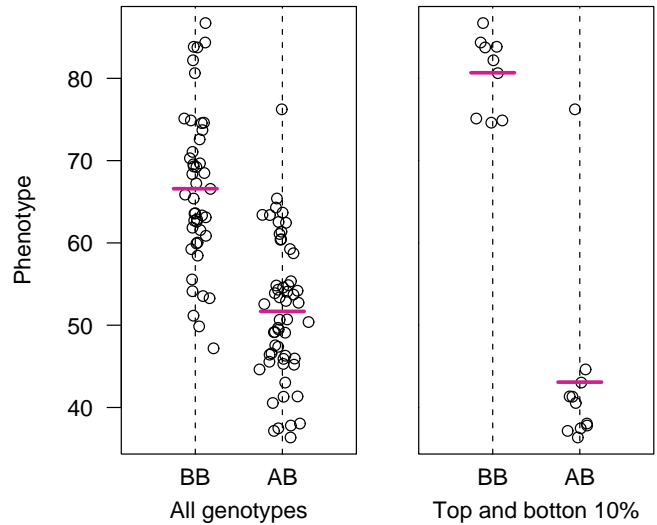
$$(A \times B) \times (B \times A) \quad \text{or} \quad (B \times A) \times (B \times A)$$

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

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

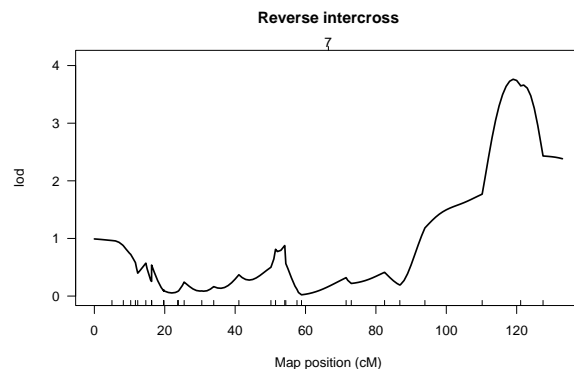
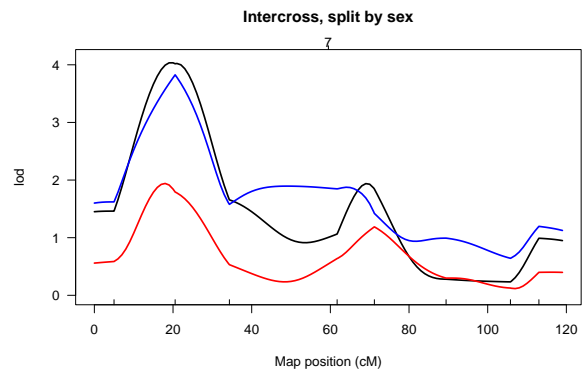
Selective genotyping

- 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.



Covariates

- **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)

Check data integrity

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. Alternatively, perform a **bootstrap**.
- 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.](#)
- Boyartchuk VL, Broman KW, Mosher RE, D’Orazio S, Starnbach M, Dietrich WF (2001) Multigenic control of *Listeria monocytogenes* susceptibility in mice. *Nat Genet* 27:259–260
Broman KW, Boyartchuk VL, Dietrich WF (2000) Mapping time-to-death quantitative trait loci in a mouse cross with high survival rates. Technical Report MS00-04, Department of Biostatistics, Johns Hopkins University, Baltimore, MD
[QTL mapping with a special model for a non-normal phenotype.](#)
- Visscher PM, Thompson R, Haley CS (1996) Confidence intervals in QTL mapping by bootstrapping. *Genetics* 143:1013–1020
[Bootstrap confidence intervals.](#)

- Miller AJ (2002) *Subset selection in regression*, 2nd edition. Chapman & Hall, New York.

A good book on model selection in regression; this new edition has just been printed.

- 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.