Model selection for QTL mapping

Karl W Broman
Department of Biostatistics, Johns Hopkins University

Terry Speed
Department of Statistics, University of California, Berkeley
Walter and Eliza Hall Institute (Melbourne, Australia)

Backcross experiment

\[ P_1 \text{ (A A) } \times \text{ P}_2 \text{ (B B)} \rightarrow \text{ F}_1 \text{ (A B)} \]

\[ \text{BC} \text{ (BC)} \]
## Trait distributions

### Data and Goals

**Phenotypes:** $y_i =$ trait value 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 trait.
- Form confidence intervals for QTL locations.
- Estimate QTL effects.
**Models: Recombination**

We assume no crossover interference.

\[ \rightarrow \text{Points of exchange (crossovers) are according to a Poisson process.} \]

\[ \rightarrow \text{The } \{x_{ij}\} \text{ (marker genotypes) form a Markov chain} \]

---

**Models: Genotype \leftrightarrow Phenotype**

Let \( y = \text{phenotype} \)

\( g = \text{whole genome genotype} \)

Imagine a small number of QTLs with genotypes \( g_1, \ldots, g_p \).

\( (2^p \text{ distinct genotypes}) \)

\[ E(y|g) = \mu_{g_1,\ldots,g_p} \quad \text{var}(y|g) = \sigma^2_{g_1,\ldots,g_p} \]
Models: Genotype $\leftrightarrow$ Phenotype

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

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

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

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

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

**LOD threshold** = 95th percentile of distribution of max LOD, genome-wide, if there are no QTLs anywhere

**Derivation:**
- Analytical calculations
- Simulations
- Permutation tests

---

Multiple QTL methods

**Why consider multiple QTLs at once?**

- Reduce residual variation.
- Separate linked QTLs.
- Investigate interactions between QTLs (epistasis).
Abstractions / simplifications

- Complete marker data
- QTLs are at the marker loci
- QTLs act additively

→ This work is not useful in practice but serves to illustrate the key issues.

The problem

n backcross mice; M markers

\[ x_{ij} = \text{genotype (1/0) of mouse } i \text{ at marker } j \]
\[ y_i = \text{phenotype (trait value) of mouse } i \]

\[ y_i = \mu + \sum_{j=1}^{M} \Delta_j x_{ij} + \epsilon_i \quad \text{Which } \Delta_j \neq 0? \]

→ Model selection in regression
How is this problem different?

- Relationship among the x’s
- Find a good model vs. minimize prediction error
- Identify the major players vs. identify the model

Model selection

- Select a class of models
- Compare models
- Search model space
- Assess the performance of a procedure
Class of models

- Additive models
- Additive + pairwise interactions
- Additive + higher order interactions
- Regression trees

Model comparison

- Estimated prediction error
- $\text{BIC}_\delta = \log \text{RSS} + \text{no. markers} \times \left( \delta \times \frac{\log n}{n} \right)$
- Sequential permutation tests
Why $\text{BIC}_\delta$?

- For a fixed no. markers, letting $n \to \infty$, $\text{BIC}_\delta$ is consistent.

- There exists a prior (on models + coefficients) for which $\text{BIC}_\delta$ is the $-\log$ posterior.

- $\text{BIC}_\delta$ is essentially equivalent to use of a threshold on the conditional LOD score.

- It performs well.

$\text{BIC}_\delta \leftrightarrow \text{conditional LOD}$

Conditional LOD score:

$$\text{LOD}(x_k^* \mid x_1^*, \ldots, x_{k-1}^*) = \frac{n}{2} \log_{10} \left\{ \frac{\text{RSS}(x_1^*, \ldots, x_{k-1}^*)}{\text{RSS}(x_1^*, \ldots, x_k^*)} \right\}$$

Minimizing $\text{BIC}_\delta$ is approximately equivalent to choosing the largest $k$ such that

$$\text{LOD}(x_k^* \mid x_1^*, \ldots, x_{k-1}^*) \geq \frac{\delta}{2} \log_{10} n$$
Choice of $\delta$

Larger $\delta$: include more loci; higher false positive rate
Smaller $\delta$: include fewer loci; lower false positive rate

Let $L = 95\%$ genome-wide LOD threshold

Choose $\delta = 2 \frac{L}{\log_{10} n}$

With this choice of $\delta$, in the absence of QTLs, we’ll include at least one extraneous locus, $5\%$ of the time.

Model search

In the case of 100 markers, there are $2^{100} \approx 10^{30}$ possible models—far more than may be inspected individually.

Methods of searching through models:

- Forward selection (FS)
- Backward elimination (BE)
- FS followed by BE
- Randomized searches (e.g., MCMC)
Assessing performance

Once must balance
- missing important loci
- including extraneous loci

“Correctly identify a QTL:”
Choose a marker within 10 cM of the QTL.

One approach:
Control the false positive rate at 5%

The appropriate criterion depends on the goals of the experimenter

Simulations

- Backcross with n=250
- No crossover interference
- 9 chr, each 100 cM
- Markers at 10 cM spacing; complete genotype data
- 7 QTLs
  - One pair in coupling
  - One pair in repulsion
  - Three unlinked QTLs
- Heritability = 50%
- 2000 simulation replicates
Methods

- ANOVA at marker loci
- Composite interval mapping (CIM)
- Forward selection with permutation tests
- Forward selection with BIC_δ
- Backward elimination with BIC_δ
- FS followed by BE with BIC_δ
- MCMC with BIC_δ

A selected marker is deemed correct if it is within 10 cM of a QTL (i.e., correct or adjacent)

A simplified version of CIM

Select a set of markers, \( S \)

(e.g., by FS to a fixed number)

For each marker, \( x \), in the genome:

(a) If \( x \notin S \), calculate \( \text{LOD}(x \mid S) \)

(b) If \( x \in S \), calculate \( \text{LOD}(x \mid S \setminus \{x\}) \)

Compare to a genome-wide threshold.

(Take into account the choice of \( S \).)
Extraneous linked

Extraneous unlinked
QTLs linked in coupling

QTLs linked in repulsion
Summary

- QTL mapping is a **model selection** problem.
- Key issue: **the comparison of models**.
- Large-scale simulations are important.
- More refined procedures do not necessarily give improved results.
- **$\text{BIC}_\delta$** with forward selection followed by backward elimination works quite well.
What next?

- Software: R/qtl
- Intercrosses
- Missing genotype data / interval mapping
- Interactions