QTL Mapping II: Model Selection

Karl W Broman
Department of Biostatistics
Johns Hopkins University
kbroman@jhsph.edu
www.biostat.jhsph.edu/~kbroman

Outline

- The model selection problem:
  - Class of models
  - Compare models
  - Search model space
  - Assess the performance of a procedure
- A simulation study
Hypothesis testing?

• In the past, QTL mapping has been regarded as a task of hypothesis testing.

Is this a QTL?

Much of the focus has been on adjusting for test multiplicity.

• It is better to view the problem as one of model selection.

What set of QTLs are well-supported?

Is there evidence for QTL-QTL interactions?

Model = a defined set of QTLs and QTL-QTL interactions (and possibly covariates and QTL-covariate interactions).

Perfect data situation

To ease discussion, we’ll focus on a simple special case:

• Complete marker genotype data
• Markers are only putative QTLs
• Normally distributed residuals

Example model (in a backcross):

\[ y_i = \mu + \Delta_1 q_{i1} + \Delta_2 q_{i2} + \Delta_3 q_{i3} + \epsilon_i \]

\( \epsilon_i \) are iid \( N(0, \sigma^2) \)

\( q_j \) are 0/1 variables (QTL genotypes)

\( \mu, \Delta \)'s are parameters, estimated by least squares

Fitted values: \( \hat{y}_i = \hat{\mu} + \hat{\Delta}_1 q_{i1} + \hat{\Delta}_2 q_{i2} + \hat{\Delta}_3 q_{i3} \)

\[ RSS = \sum_i (y_i - \hat{y}_i)^2 \] indicates model fit.
Model selection

1. Class of models
2. Compare models
3. Search model space
4. Assess performance of a procedure

Note:

2 and 4 are much the same.

There might be a 0th item: Method for model fitting.
(e.g., imputation, EM, etc.)

Class of models

- Additive models
  \[ y = \mu + \sum_j \Delta_j q_j + \epsilon \]

- Also pairwise interactions
  Preserve hierarchy: if include interaction, also include both main effects.

- Also higher-order interactions
  Again, preserve hierarchy.

- Regression trees
**Example regression tree**

QTL 1

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>H</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>H</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>B</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

QTL 2

**Intercrosses: class of models**

- Always bring in both degrees of freedom with a QTL
  
  or

  Try to distinguish additivity/dominance/recessiveness?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>H</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Always bring in all four d.f. with a QTL:QTL interaction
  
  or

  Try to distinguish ways to subdivide the $3 \times 3$ table?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>H</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Compare models

- Imagine you could fit all possible models; which would you like best?

- This issue is like LOD thresholds, but more complicated.

- For models with the same number of parameters (QTLs and interactions), we prefer that with the “best fit” (smallest RSS or largest likelihood).

- If you fit more parameters, you’ll get a “better fit.”
  - How much better, before including additional terms?
  - I like a form of penalized likelihood.

- Note: Bayesians must also confront this issue (through the prior on models).

The additive QTL case

n backcross mice; M markers

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

\[ y_i = \text{phenotype (trait value) of mouse } i \]

\[ y_i = \mu + \sum_{j=1}^{M} \Delta_j x_{ij} + \epsilon_i \]

Which \( \Delta_j \neq 0 \)?

\[ \text{BIC}_\delta = \log\text{ RSS} + \text{no. markers} \times \left( \delta \times \frac{\log n}{n} \right) \]
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.

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

Conditional LOD score:

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

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

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

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

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

Let $L = 95\%$ genome-wide LOD threshold
(compare single-QTL models to the null model)

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.

Note that now we have

$$BIC_\delta = \log_{10} RSS + \text{no. markers} \times \left( \frac{2L}{n} \right)$$

Search model space

- Consider the case of additive QTL models, with 100 putative QTLs.
- There are $2^{100} \approx 10^{30}$ possible models, far more than can be inspected individually.
- Need a way to search through this space, to find the good ones.
- This is really a matter of “grunt work.” (More is better; the tradeoff is with computational time.)
Methods of model search

- Forward selection
  - Find the best single-QTL model: \( q_1^* \).
  - Find the best two-QTL model that includes \( q_1^*, q_2^* \).
  - Find the best three-QTL model that includes \( q_1^*, q_2^*, q_3^* \).
  - Etc.

- Backward elimination

- Forward selection followed by backward elimination

- Stepwise selection

- Randomized algorithms (e.g., MCMC, genetic algorithms, etc.)

Assess performance

- Selection of a model includes two types of errors:
  - Miss important terms (QTLs or interactions)
  - Include extraneous terms

- Unlike in hypothesis testing, we can make both errors at the same time!

- Identify as many correct terms as possible, while controlling the rate of inclusion of extraneous terms.

- You can’t know the performance of your procedure with your data—you need to know the truth.

- You can know:
  - How a particular procedure performs in simulated cases
  - How a procedure performs in simulated data close to what you’ve inferred
Science isn’t really model selection

I’ve said here: **pick a good model**

*Really:*

- Want to guide future experiments
- Want some understanding of the **uncertainty** in different aspects of the chosen model

---

**A simulation study**

- 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 $\text{BIC}_\delta$
- Backward elimination with $\text{BIC}_\delta$
- FS followed by BE with $\text{BIC}_\delta$
- MCMC with $\text{BIC}_\delta$

→ 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 $x \not\in S$, 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 $S$.)
IM / CIM → model

![Graph showing marker LOD scores and correct average number of chosen markers for different models. The graph includes a dot plot for LOD scores and a bar plot for average number of chosen markers with categories such as ANOVA, CIM, fs, perm, fs, be, fs/be, and mcmc.](image-url)
Extraneous linked

Extraneous unlinked
QTLs linked in coupling

QTLs linked in repulsion
Summary

- QTL mapping is a model selection problem (rather than hypothesis testing).

- Model selection =
  - Select a class of models
  - Select a criterion for comparing models
  - Select a method of searching model space
  - Figure out how your procedure performs

- Key issue: the comparison of models.

- Large-scale computer simulations are necessary for assessing the performance of procedures.
References

  Contains the simulation study described above.

  Another paper on the model selection aspects of QTL mapping.

  A good book on model selection in regression.