Multiple QTL mapping

Karl W Broman
Department of Biostatistics
Johns Hopkins University

www.biostat.jhsph.edu/~kbroman

[→ Teaching → Miscellaneous lectures]
Why?

- Reduce residual variation $\implies$ increased power
- Separate linked QTL
- Identify interactions among QTL
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 QTL are well supported?
  Is there evidence for QTL-QTL interactions?

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

The missing data problem: Markers ↔ QTL

The model selection problem: QTL, covariates → phenotype
Starting points

• Single-QTL scan (ie, interval mapping)
  – Loci with marginal effects should appear

• 2-dim, 2-QTL scan
  – Ability to separate linked QTL
  – Identify interacting loci
Example: 1d scan
Example: 2d scan
$LOD_i$ and $LOD_{fv1}$
$\text{LOD}_{av1}$ and $\text{LOD}_{fv1}$
Exploratory methods

- **Condition on a large-effect QTL**
  - Reduce residual variation
  - Conditional LOD score:
    \[
    \text{LOD}(q_2 | q_1) = \log_{10}\left\{ \frac{\Pr(\text{data} | q_1, q_2)}{\Pr(\text{data} | q_1)} \right\}
    \]

- **Piece together the putative QTL from the 1d and 2d scans**
  - Omit loci that no longer look interesting (drop-one-at-a-time analysis)
  - Study potential interactions among the identified loci
  - Scan for additional loci (perhaps allowing interactions), conditional on these
Condition on D3M3

The graph shows the LOD (Logarithm of the Odds) values across different chromosomes for two conditions: IM (blue) and condition on D3M3 (red). The LOD values are plotted along the y-axis, with chromosome numbers along the x-axis. The graph highlights the regions where the LOD values are significantly high, indicating potential markers of interest for each condition.
# Drop-one-at-a-time

<table>
<thead>
<tr>
<th>chr</th>
<th>pos</th>
<th>df</th>
<th>LOD</th>
<th>% var</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47.5</td>
<td>2</td>
<td>7.26</td>
<td>10.2</td>
</tr>
<tr>
<td>2</td>
<td>40.0</td>
<td>2</td>
<td>7.84</td>
<td>11.1</td>
</tr>
<tr>
<td>3</td>
<td>20.0</td>
<td>1</td>
<td>6.62</td>
<td>9.3</td>
</tr>
<tr>
<td>4</td>
<td>27.5</td>
<td>1</td>
<td>4.16</td>
<td>5.7</td>
</tr>
<tr>
<td>4</td>
<td>52.5</td>
<td>1</td>
<td>2.87</td>
<td>3.9</td>
</tr>
<tr>
<td>1 × 2</td>
<td>7.17</td>
<td>1</td>
<td>10.1</td>
<td></td>
</tr>
</tbody>
</table>

Overall: LOD = 18.2, % var = 28.5
## Refined positions

<table>
<thead>
<tr>
<th>chr</th>
<th>pos</th>
<th>df</th>
<th>LOD</th>
<th>% var</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50.0</td>
<td>2</td>
<td>7.57</td>
<td>10.6</td>
</tr>
<tr>
<td>2</td>
<td>40.0</td>
<td>2</td>
<td>8.21</td>
<td>11.6</td>
</tr>
<tr>
<td>3</td>
<td>22.5</td>
<td>1</td>
<td>6.90</td>
<td>9.6</td>
</tr>
<tr>
<td>4</td>
<td>30.0</td>
<td>1</td>
<td>4.69</td>
<td>6.4</td>
</tr>
<tr>
<td>4</td>
<td>52.5</td>
<td>1</td>
<td>3.30</td>
<td>4.4</td>
</tr>
<tr>
<td>1 × 2</td>
<td></td>
<td>1</td>
<td>7.51</td>
<td>10.5</td>
</tr>
</tbody>
</table>

Overall: LOD = 18.8, % var = 29.2
Scan for further QTL
Identify a set of markers, \( S = \{x_1, x_2, \ldots, x_k\} \), proxies for QTL

Scan the genome, using these markers as covariates
- What do we do in the case of missing marker genotype data?

At a position far from any of the marker covariates, compare \( S \cup q \) and \( S \)

Within some fixed window of a marker covariate, compare \( S \setminus \{x\} \cup q \) and \( S \setminus \{x\} \)

The key issue: How to select \( S \)?
- QTL Cartographer: forward selection to some fixed number of markers
CIM results

1 marker

Chromosome

lod

5 markers

Chromosome

lod

3 markers

Chromosome

lod

7 markers

Chromosome

lod
Perfect data situation

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

- Complete marker genotype data
- Markers are only putative QTL
- Normally distributed residuals

Example model (in a backcross):

$$y_i = \mu + \beta_1 q_{i1} + \beta_2 q_{i2} + \beta_3 q_{i3} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$q_j$ are 0/1 variable (QTL genotypes)

$\mu, \beta$’s are parameters, estimated by least squares

Fitted values: $\hat{y}_i = \hat{\mu} + \hat{\beta}_1 q_{i1} + \hat{\beta}_2 q_{i2} + \hat{\beta}_3 q_{i3}$

$RSS = \sum_i (y_i - \hat{y}_i)^2$ indicates model fit.
Consider all putative QTL and QTL × QTL interactions:

\[ y = \mu + \sum_{j} \beta_j q_j + \sum_{j<k} \gamma_{jk} q_j q_k + \epsilon \]

Which \( \beta_j \neq 0 \)?

Which \( \gamma_{jk} \neq 0 \)?
Model selection

- Class of models
  - Additive models
  - + pairwise interactions
  - + higher-order interactions

- Model fit
  - Maximum likelihood
  - Haley-Knott regression
  - extended Haley-Knott
  - Multiple imputation
  - MCMC

- Model comparison
  - Estimated prediction error
  - AIC, BIC, penalized likelihood
  - Bayes

- Model search
  - Forward selection
  - Backward elimination
  - Stepwise selection
  - Randomized algorithms
Intercross: class of models

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

  Try to distinguish additivities/dominance/recessiveness?

  \[ \begin{array}{ccc} 
  A & H & B \\
  \end{array} \]

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

  Try to distinguish ways to subdivide the \(3 \times 3\) table?

  \[ \begin{array}{ccc} 
  A & H & B \\
  A & & B \\
  H & & \\
  B & & \\
  \end{array} \]
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

q1

q2

A

H or B

A

H or B

100

80

20
Imagine you could fit all possible models; which would you like best?

This issue is like LOD thresholds, but more complex.

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.
The additive QTL case

n backcross mice; M markers

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

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

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

Which \( \beta_j \neq 0 \)?

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

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

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

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

Choose $\delta = 2T / \log_{10} n$

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

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

$$\propto - (\text{LOD} - \text{no. markers} \times T)$$
Model search

• 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.)
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
What is special here?

• Goal: identify the major players

• A continuum of ordinal-valued covariates (the genetic loci)

• Association among the covariates
  – Loci on different chromosomes are independent
  – Along chromosome, a very simple (and known) correlation structure
A simulation study

- Backcross with n=250
- No crossover interference
- 9 chr, each 100 cM
- Markers at 10 cM spacing; complete genotype data
- 7 QTL
  - One pair in **coupling**
  - One pair in **repulsion**
  - Three unlinked QTL
- 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 was deemed correct if it was within 10 cM of a QTL (i.e. correct or adjacent).
Correct

Ave no. chosen

ANOVA 3 5 7 9 11
CIM fs, perm fs be fs/be mcmc

BIC_δ
QTL in coupling

Ave no. chosen

ANOVA
3 5 7 9 11
CIM
fs, perm
fs
be
fs/be
mcmc
B IC δδ
QTL in repulsion

Ave no. chosen

ANOVA
fs, perm
fs
be
fs/be
mcmc

BIC_δ
Epistasis

- $\gamma = \text{model}$
  - $|\gamma|_m = \text{no. main effects}$
  - $|\gamma|_i = \text{no. interactions}$

- Additive QTL case:
  \[
  \text{LOD}(\gamma) - |\gamma|_m T_m
  \]

- With pairwise interactions:
  \[
  \text{LOD}(\gamma) - |\gamma|_m T_m - |\gamma|_i T_i
  \]

- Need a more complex penalty on interactions.
Models as graphs

A

B

C

D
Bayesian methods

- The likelihood function
  \[ L_\gamma(\theta) = \Pr(\text{data} \mid \theta, \gamma) \]
  \( \gamma = \text{model}, \theta = \text{parameters} \)

- Frequentists
  \[ L_\gamma = \max_\theta L_\gamma(\theta) \]
  Penalize model complexity

- Bayesians
  Prior: \( \Pr(\gamma), \Pr(\theta \mid \gamma) \)
  Posterior: \( \Pr(\gamma \mid \text{data}) = \int \Pr(\gamma) \Pr(\theta \mid \gamma) L_\gamma(\theta) \, d\theta \)
Summary

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

- Model selection =
  - Select a class of models
  - Select a method for fitting models
  - Selecting a criterion for comparing models
  - Select a method of searching model space

- 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 review of model selection in QTL mapping.

  A good book on model selection.

  A Bayesian approach for QTL mapping.

  A Bayesian approach for identifying interacting QTL.