Introduction to QTL mapping in model organisms

Karl W Broman
Department of Biostatistics and Medical Informatics
University of Wisconsin – Madison

www.biostat.wisc.edu/~kbroman

[→ Teaching → Miscellaneous lectures]

Backcross
Intercross

\[ P_1 \times P_2 \]

\[ F_1 \times F_1 \]

Phenotype data

\[ \log_2 \text{ liver} \]

\[ \log_2 \text{ spleen} \]
Goals

- Identify quantitative trait loci (QTL) (and interactions among QTL)
- Interval estimates of QTL location
- Estimated QTL effects

Statistical structure

The missing data problem: Markers ←→ QTL

The model selection problem: QTL, covariates → phenotype
Also known as marker regression.

Split mice into groups according to genotype at a marker.

Do a t-test / ANOVA.

Repeat for each marker.

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

Lander & Botstein (1989)

- Assume a single QTL model.
- Each position in the genome, one at a time, is posited as the putative QTL.
- Let \( q = 1/0 \) if the (unobserved) QTL genotype is BB/AB.
  (Or 2/1/0 if the QTL genotype is BB/AB/AA in an intercross.)
  Assume \( y \mid q \sim N(\mu_q, \sigma) \)
- Given genotypes at linked markers, \( y \sim \) mixture of normal dist’ns with mixing proportions \( \Pr(q \mid \text{marker data}) \):

<table>
<thead>
<tr>
<th>( M_1 )</th>
<th>( M_2 )</th>
<th>QTL genotype</th>
<th>( BB )</th>
<th>( AB )</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>BB</td>
<td>(1 - ( r_L ))(1 - ( r_R ))/(1 - ( r ))</td>
<td>( r_{LR}/(1 - r) )</td>
<td></td>
</tr>
<tr>
<td>BB</td>
<td>AB</td>
<td>(1 - ( r_L ))( r_R )/r</td>
<td>( r_L(1 - r_R)/r )</td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td>BB</td>
<td>( r_L(1 - r_R)/r )</td>
<td>(1 - ( r_L ))( r_R )/r</td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td>AB</td>
<td>( r_{LR}/(1 - r) )</td>
<td>(1 - ( r_L ))(1 - ( r_R ))/(1 - ( r ))</td>
<td></td>
</tr>
</tbody>
</table>

Genotype probabilities

Calculate \( \Pr(q \mid \text{marker data}) \), assuming
- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology
- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)
The normal mixtures

- Two markers separated by 20 cM, with the QTL closer to the left marker.
- The figure at right shows 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

Let $p_{ij} = \Pr(q_i = j | \text{marker data})$

$y_i | q_i \sim \mathcal{N}(\mu_{q_i}, \sigma^2)$

$\Pr(y_i | \text{marker data}, \mu_0, \mu_1, \sigma) = \sum_j p_{ij} f(y_i; \mu_j, \sigma)$

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

Log likelihood: $l(\mu_0, \mu_1, \sigma) = \sum_i \log \Pr(y_i | \text{marker data}, \mu_0, \mu_1, \sigma)$

Maximum likelihood estimates (MLEs) of $\mu_0, \mu_1, \sigma$: values for which $l(\mu_0, \mu_1, \sigma)$ is maximized.
EM algorithm

Dempster et al. (1977)

E step:
Let \( w_{ij}^{(k)} = \Pr(q_i = j|y_i, \text{marker data}, \hat{\mu}_0^{(k-1)}, \hat{\mu}_1^{(k-1)}, \hat{\sigma}^{(k-1)}) \)
\[
= \frac{p_{ij} f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})}{\sum_j p_{ij} f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})}
\]

M step:
Let \( \hat{\mu}_j^{(k)} = \frac{\sum_i y_i w_{ij}^{(k)}}{\sum_i w_{ij}^{(k)}} \)
\( \hat{\sigma}^{(k)} = \sqrt{\frac{\sum_i \sum_j w_{ij}^{(k)} (y_i - \hat{\mu}_j^{(k)})^2}{n}} \)

The algorithm:
Start with \( w_{ij}^{(1)} = p_{ij} \); 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.

\[
\text{LOD}(\lambda) = \log_{10} \frac{\Pr(y|QTL \text{ at } \lambda, \hat{\mu}_0\lambda, \hat{\mu}_1\lambda, \hat{\sigma}_\lambda)}{\Pr(y|\text{no QTL, } \hat{\mu}, \hat{\sigma})}
\]

\( \hat{\mu}_0\lambda, \hat{\mu}_1\lambda, \hat{\sigma}_\lambda \) are the MLEs, assuming a single QTL at position \( \lambda \).

No QTL model: The phenotypes are independent and identically distributed (iid) \( N(\mu, \sigma^2) \).
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.
Large LOD scores indicate evidence for the presence of a QTL

Question: How large is large?

LOD threshold = 95 \text{th} \text{ile} \text{ of distr'}n \text{ of max LOD, genome-wide, if there are no QTLs anywhere}

Derivation: • Analytical calculations (L & B 1989)
• Simulations (L & B 1989)
• Permutation tests (Churchill & Doerge 1994)

Null distribution of the LOD score

• Null distribution derived by computer simulation of backcross with genome of typical size.
• Dashed curve: distribution of LOD score at any one point.
• Solid curve: distribution of maximum LOD score, genome-wide.
Permutation test

Permutation results

Genome-wide maximum LOD score

0 1 2 3 4 5 6
Modelling multiple QTL

- Reduce residual variation $\implies$ increased power
- Separate linked QTL
- Identify interactions among QTL
Epistasis in BC

Epistasis in F₂
References

A review for non-statisticians.

Chapter on QTL mapping.

The seminal paper.

LOD thresholds by permutation tests.

An old but excellent general genetics textbook with a very interesting discussion of epistasis.