Mapping multiple QTL in experimental crosses

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

www.biostat.wisc.edu/~kbroman
Human vs mouse
Backcross

P₁ x P₂

F₁

BC
Intercross

P₁

×

P₂

F₁

×

F₁

F₂
250 male mice from the backcross \((A \times B) \times B\)
Blood pressure after two weeks drinking water with 1% NaCl
Genotype data
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 $\leftrightarrow$ QTL

The model selection problem:
QTL, covariates $\rightarrow$ phenotype
• Split mice into groups according to genotype at a marker.
• Do a t-test / ANOVA.
• Repeat for each marker.
Lander & Botstein (1989)

- Assume a single QTL model.

- Consider each position in the genome, one at a time, as the location of the putative QTL.

- Let \( q = 0/1 \) if the (unobserved) QTL genotype is BB/AB. (Or 0/1/2 if the QTL genotype is AA/AB/BB in an intercross.)

Assume \( y \mid q \sim N(\mu_q, \sigma) \)

- Calculate \( p_q = Pr(q \mid \text{marker data}) \).

\[ y \mid \text{marker data} \sim \sum_q p_q \phi(y \mid \mu_q, \sigma) \]
LOD scores

$$\text{LOD}(\lambda) = \log_{10} \text{likelihood ratio comparing the hypothesis of a QTL at position } \lambda \text{ versus that of no QTL}$$

$$= \log_{10} \left\{ \frac{\Pr(y|\text{QTL at } \lambda, \hat{\mu}_q \lambda, \hat{\sigma}_\lambda)}{\Pr(y|\text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\}$$

$\hat{\mu}_q \lambda, \hat{\sigma}_\lambda$ are the MLEs, assuming a single QTL at position $\lambda$.

No QTL model: The phenotypes are iid $N(\mu, \sigma^2)$. 
Permutation test

markers

individuals

genotype data

phenotypes

LOD scores

maximum LOD score
Permutation results

Genome-wide maximum LOD score
LOD curves

LOD score vs. Chromosome

5% threshold
Estimated effects

- **Chr 1 @ 48 cM**
- **Chr 4 @ 30 cM**
- **Chr 6 @ 24 cM**
- **Chr 15 @ 20 cM**

Genotype and Blood Pressure Comparison:

- **Genotype**: BB, BA
- **Blood Pressure** values range from 98 to 106
Modeling multiple QTL

- Reduce residual variation $\rightarrow$ increased power
- Separate linked QTL
- Identify interactions among QTL (epistasis)
Estimated effects

1 x 4

Chr 1 genotype

Blood pressure

BB BA

Chr 4 genotype

6 x 15

Chr 6 genotype

BB BA

Chr 15 Genotype
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).
Model selection

- **Class of models**
  - Additive models
  - + pairwise interactions
  - + higher-order interactions
  - Regression trees

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

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

- **Model search**
  - Forward selection
  - Backward elimination
  - Stepwise selection
  - Randomized algorithms
• 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.

• More concerned about extraneous loci than about extraneous interactions.
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
• Assistance to the masses

• Understanding performance

• Many phenotypes
Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

\[ y = \mu + \sum \beta_j q_j + \epsilon \quad \text{which } \beta_j \neq 0? \]

\[ \text{pLOD}(\gamma) = \text{LOD}(\gamma) - T |\gamma| \]
Additive QTL

Simple situation:

- Dense markers
- Complete genotype data
- No epistasis

\[ y = \mu + \sum \beta_j q_j + \epsilon \quad \text{which } \beta_j \neq 0? \]

\[ \text{pLOD}(\gamma) = \text{LOD}(\gamma) - T |\gamma| \]

0 vs 1 QTL: \( \text{pLOD}(\emptyset) = 0 \)

\[ \text{pLOD}(\{\lambda\}) = \text{LOD}(\lambda) - T \]
Additive QTL

Simple situation:
- Dense markers
- Complete genotype data
- No epistasis

\[ y = \mu + \sum \beta_j q_j + \epsilon \]

which \( \beta_j \neq 0 \)?

\[ \text{pLOD}(\gamma) = \text{LOD}(\gamma) - T |\gamma| \]

For the mouse genome:
\[ T = 2.69 \text{ (BC)} \text{ or } 3.52 \text{ (F}_2\text{)} \]
Experience

- Controls rate of inclusion of extraneous terms
- Forward selection over-selects
- Forward selection followed by backward elimination works as well as MCMC
- Need to define performance criteria
- Need large-scale simulations

Broman & Speed, JRSS B 64:641-656, 2002
Epistasis

\[ y = \mu + \sum \beta_j q_j + \sum \gamma_{jk} q_j q_k + \epsilon \]

\[ pLOD(\gamma) = LOD(\gamma) - T_m |\gamma|_m - T_i |\gamma|_i \]

\[ T_m = \text{as chosen previously} \]

\[ T_i = ? \]
Imagine there are two additive QTL and consider a 2d, 2-QTL scan.

\[ T_i = 95\text{th percentile of the distribution of} \]
\[ \max \text{LOD}_f(s, t) - \max \text{LOD}_a(s, t) \]
Imagine there are two additive QTL and consider a 2d, 2-QTL scan.

\[ T_i = 95\text{th percentile of the distribution of} \]
\[ \max \text{LOD}_f(s, t) - \max \text{LOD}_a(s, t) \]

For the mouse genome:

\[ T_m = 2.69 \text{ (BC) or 3.52 (F}_2) \]
\[ T_i^H = 2.62 \text{ (BC) or 4.28 (F}_2) \]
Imagine there is one QTL and consider a 2d, 2-QTL scan.

\[ T_m + T_i = 95\text{th percentile of the distribution of } \max \text{ LOD}_f(s, t) - \max \text{ LOD}_1(s) \]
Imagine there is one QTL and consider a 2d, 2-QTL scan.

\[ T_m + T_i = 95\text{th percentile of the distribution of} \]
\[ \max \text{LOD}_f(s, t) - \max \text{LOD}_1(s) \]

For the mouse genome:

\[ T_m = 2.69 \text{ (BC)} \text{ or } 3.52 \text{ (F}_2\text{)} \]
\[ T_i^H = 2.62 \text{ (BC)} \text{ or } 4.28 \text{ (F}_2\text{)} \]
\[ T_i^L = 1.19 \text{ (BC)} \text{ or } 2.69 \text{ (F}_2\text{)} \]
Models as graphs
Results

LOD = 23.1
Results

$T_m = 2.69 \quad T_{iH} = 2.62 \quad T_{iL} = 1.19 \quad T_m + T_{iH} = 5.31 \quad T_m + T_{iL} = 3.88 \quad 2T_m = 5.38$

LOD = 23.1


### Results

<table>
<thead>
<tr>
<th></th>
<th>6.3</th>
<th>1</th>
<th>4</th>
<th>12.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3</td>
<td>7.9</td>
<td>6</td>
<td>15</td>
<td>7.1</td>
</tr>
<tr>
<td>5.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38
\]
Add an interaction?

\[ T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38 \]
Add an interaction?

\[ T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38 \]
Add an interaction?

\[ T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38 \]
Add an interaction?

\[ T_m = 2.69 \quad T^H_i = 2.62 \quad T^L_i = 1.19 \quad T_m + T^H_i = 5.31 \quad T_m + T^L_i = 3.88 \quad 2T_m = 5.38 \]
Add an interaction?

\[ T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38 \]
Add another QTL?

\[ T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38 \]
Add another QTL?

$T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38$
Add another QTL?

\[ T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38 \]
Add a pair of QTL?

\[ T_m = 2.69 \quad T_i^H = 2.62 \quad T_i^L = 1.19 \quad T_m + T_i^H = 5.31 \quad T_m + T_i^L = 3.88 \quad 2T_m = 5.38 \]
To do

- Improve search procedures
- X chromosome
- QTL $\times$ covariate interactions
- Measuring model uncertainty
- Measuring uncertainty in QTL location
• QTL mapping is a model selection problem

• The criterion for comparing models is most important

• We’re focusing on a penalized likelihood method and are close to a practiceable solution

Acknowledgments

Ani Manichaikul  University of Virginia
Gary Churchill  Jackson Laboratory
Šaunak Sen  University of California, San Francisco
Terry Speed  University of California, Berkeley
Brian Yandell  University of Wisconsin – Madison
Jee Young Moon  University of Wisconsin – Madison

Fumihiro Sugiyama  now at University of Tsukuba, Japan
Bev Paigen  Jackson Laboratory