Mapping multiple QTL in experimental crosses

Karl W Broman

Department of Biostatistics and Medical Informatics
University of Wisconsin – Madison

www.biostat.wisc.edu/~kbroman

[→ Teaching → Miscellaneous lectures]

250 male mice from the backcross (A × B) × B
Blood pressure after two weeks drinking water with 1% NaCl
Genotype data

<table>
<thead>
<tr>
<th>Markers</th>
<th>Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>X</td>
<td>20</td>
</tr>
</tbody>
</table>

The diagram shows a genotype analysis with markers ranging from 50 to 250 and individuals from 1 to 19. The color intensity variation likely indicates different genotypes or allele frequencies.
Goals

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

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

blood pressure

Genotype

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

blood pressure

Genotype

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

blood pressure

Genotype
Modeling multiple QTL

- Reduce residual variation $\rightarrow$ increased power
- Separate linked QTL
- Identify interactions among QTL (epistasis)
Epistasis in BC

Additive

Epistatic

QTL 1

QTL 2

A

H

A

H

A

H

A

H

QTL 2

Ave. phenotype

Ave. phenotype

Ave. phenotype

Ave. phenotype

0

20

40

60

80

100

0

20

40

60

80

100

0

20

40

60

80

100
Epistasis in $F_2$
2-dim, 2-QTL scan

For all pairs of positions, fit the following models:

\[ H_f : y = \mu + \beta_1 q_1 + \beta_2 q_2 + \gamma q_1 q_2 + \epsilon \]

\[ H_a : y = \mu + \beta_1 q_1 + \beta_2 q_2 + \epsilon \]

\[ H_1 : y = \mu + \beta_1 q_1 + \epsilon \]

\[ H_0 : y = \mu + \epsilon \]

log_{10} likelihoods:

\[ l_f(s, t) \quad l_a(s, t) \quad l_1(s) \quad l_0 \]
2-dim, 2-QTL scan

LOD scores:

\[ \text{LOD}_f(s, t) = l_f(s, t) - l_0 \]

\[ \text{LOD}_a(s, t) = l_a(s, t) - l_0 \]

\[ \text{LOD}_i(s, t) = l_f(s, t) - l_a(s, t) \]

\[ \text{LOD}_1(s) = l_1(s) - l_0 \]
Results: $\text{LOD}_i$ and $\text{LOD}_f$
Results: LOD$_i$ and LOD$_f$
Consider each pair of chromosomes, \((j, k)\), and let \(c(s)\) denote the chromosome for position \(s\).

\[
M_f(j, k) = \max_{c(s)=j, c(t)=k} \text{LOD}_f(s, t)
\]

\[
M_a(j, k) = \max_{c(s)=j, c(t)=k} \text{LOD}_a(s, t)
\]

\[
M_1(j, k) = \max_{c(s)=j \text{ or } k} \text{LOD}_1(s)
\]

\[
M_i(j, k) = M_f(j, k) - M_a(j, k)
\]

\[
M_{fv1}(j, k) = M_f(j, k) - M_1(j, k)
\]

\[
M_{av1}(j, k) = M_a(j, k) - M_1(j, k)
\]
Results: $LOD_i$ and $LOD_{fv1}$
A pair of chromosomes \((j, k)\) is considered interesting if:

\[
M_f(j, k) > T_f \quad \text{and} \quad \{ M_{fv1}(j, k) > T_{fv1} \text{ or } M_i(j, k) > T_i \}
\]

or

\[
M_a(j, k) > T_a \quad \text{and} \quad M_{av1}(j, k) > T_{av1}
\]

where the thresholds \((T_f, T_{fv1}, T_i, T_a, T_{av1})\) are determined by a permutation test with a 2d scan.
## 2d scan summary

<table>
<thead>
<tr>
<th></th>
<th>pos1f</th>
<th>pos2f</th>
<th>lod.full</th>
<th>lod.fv1</th>
<th>lod.int</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1:c4</td>
<td>71.3</td>
<td>30.0</td>
<td>14.36</td>
<td>6.78</td>
<td>0.27</td>
</tr>
<tr>
<td>c6:c15</td>
<td>55.0</td>
<td>20.5</td>
<td>6.91</td>
<td>4.95</td>
<td>2.92</td>
</tr>
<tr>
<td>c1:c1</td>
<td>39.3</td>
<td>78.3</td>
<td>5.10</td>
<td>1.58</td>
<td>0.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>pos1a</th>
<th>pos2a</th>
<th>lod.add</th>
<th>lod.av1</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1:c4</td>
<td>68.3</td>
<td>30.0</td>
<td>14.09</td>
<td>6.50</td>
</tr>
<tr>
<td>c6:c15</td>
<td>24.0</td>
<td>22.5</td>
<td>3.99</td>
<td>2.03</td>
</tr>
<tr>
<td>c1:c1</td>
<td>48.3</td>
<td>79.3</td>
<td>5.02</td>
<td>1.50</td>
</tr>
</tbody>
</table>
Estimated effects

Chr 1 genotype

Chr 4 genotype

Chr 15 Genotype
Chr 1: $\text{LOD}_i$ and $\text{LOD}_{av1}$
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

- **Model fit**
  - Maximum likelihood
  - Haley-Knott regression
  - extended Haley-Knott
  - Multiple imputation
  - Markov chain Monte Carlo

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

- **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.
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
Exploratory methods

- Condition on a large-effect QTL
  - Reduce residual variation
  - Conditional LOD score:

\[
\text{LOD}(q_2 \mid q_1) = \log_{10} \left\{ \frac{\Pr(\text{data} \mid q_1, q_2)}{\Pr(\text{data} \mid q_1)} \right\}
\]

- 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
Controlling for chr 4

![Graph showing LOD score against chromosome number with intervals mapped for chr 4 controlled.](image-url)
## Drop-one-QTL table

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>LOD</th>
<th>%var</th>
</tr>
</thead>
<tbody>
<tr>
<td>1068.3</td>
<td>1</td>
<td>6.30</td>
<td>11.0</td>
</tr>
<tr>
<td>4030.0</td>
<td>1</td>
<td>12.21</td>
<td>20.1</td>
</tr>
<tr>
<td>6061.0</td>
<td>2</td>
<td>7.93</td>
<td>13.6</td>
</tr>
<tr>
<td>15017.5</td>
<td>2</td>
<td>7.14</td>
<td>12.3</td>
</tr>
<tr>
<td>6061.0 : 15017.5</td>
<td>1</td>
<td>5.68</td>
<td>9.9</td>
</tr>
</tbody>
</table>
Automation

- Assist inexperienced analysts
- Understand 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

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

which \( \beta_j \neq 0 \)?

\[ pLOD(\gamma) = LOD(\gamma) - T \left| \gamma \right| \]

0 vs 1 QTL: \( pLOD(\emptyset) = 0 \)

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

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

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

which \( \beta_j \neq 0? \)

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

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

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

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

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

\[ T_i = ? \]
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:

\[ T_m = 2.69 \text{ (BC)} \text{ or } 3.52 \text{ (F}_2\text{)} \]
\[ T_i^H = 2.62 \text{ (BC)} \text{ or } 4.28 \text{ (F}_2\text{)} \]
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.

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

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

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

LOD = 23.1
Results

\[ T_m = 2.69 \]
\[ T_i^H = 2.62 \]
\[ T_i^L = 1.19 \]
\[ T_m + T_i^H = 5.31 \]
\[ T_m + T_i^L = 3.88 \]
\[ 2T_m = 5.38 \]
## Drop-one-QTL table

<table>
<thead>
<tr>
<th>df</th>
<th>LOD</th>
<th>%var</th>
</tr>
</thead>
<tbody>
<tr>
<td>1@68.3</td>
<td>6.30</td>
<td>11.0</td>
</tr>
<tr>
<td>4@30.0</td>
<td>12.21</td>
<td>20.1</td>
</tr>
<tr>
<td>6@61.0</td>
<td>7.93</td>
<td>13.6</td>
</tr>
<tr>
<td>15@17.5</td>
<td>7.14</td>
<td>12.3</td>
</tr>
<tr>
<td>6@61.0 : 15@17.5</td>
<td>5.68</td>
<td>9.9</td>
</tr>
</tbody>
</table>
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?

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

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

- QTL mapping is a model selection problem
- The criterion for comparing models is most important
- We’re focusing on a penalized likelihood method, with penalties derived from permutation tests with 1d and 2d scans
  Manichaikul et al., Genetics 181:1077–1086, 2009