Identifying QTLs in experimental crosses

Karl W. Broman
Department of Biostatistics
Johns Hopkins School of Public Health

kbroman@jhsph.edu
http://kbroman.homepage.com
Data

- n (100–1000) F$_2$ progeny
- $y_i$ = phenotype for individual i
- $g_{ij}$ = genotype for individual i at marker j (AA, AB or BB)
- Genetic map of the markers
- Phenotypes of parentals and F$_1$
Single QTL analysis

Marker D1M1

<table>
<thead>
<tr>
<th>genotype</th>
<th>Ave (SE)</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>54.0 (1.4)</td>
<td>8.7</td>
<td>22</td>
</tr>
<tr>
<td>AB</td>
<td>52.9 (0.8)</td>
<td>6.5</td>
<td>63</td>
</tr>
<tr>
<td>AA</td>
<td>47.5 (2.1)</td>
<td>9.8</td>
<td>37</td>
</tr>
<tr>
<td>Overall</td>
<td>52.3 (0.7)</td>
<td>8.1</td>
<td>122</td>
</tr>
</tbody>
</table>

Additive effect

\[ a = \frac{\mu(BB) - \mu(AA)}{2} \approx 3.3 \ (1.3) \]

Dominance deviation

\[ d = \mu(AB) - \frac{\mu(BB) + \mu(AA)}{2} \approx 2.2 \ (1.5) \]

Prop’n var explained

\[ h^2 = \frac{\text{var in } F_2 - \text{residual var}}{\text{var in } F_2} \approx \frac{1}{2} a^2 + \frac{1}{4} d^2 \]

\[ \approx \frac{1}{2} 3.3^2 + \frac{1}{4} 2.2^2 = 8\% \]
Single QTL analysis

Marker D2M2

<table>
<thead>
<tr>
<th>genotype</th>
<th>Ave (SE)</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>55.8 (1.2)</td>
<td>6.4</td>
<td>27</td>
</tr>
<tr>
<td>AB</td>
<td>54.0 (0.9)</td>
<td>7.6</td>
<td>65</td>
</tr>
<tr>
<td>AA</td>
<td>45.3 (1.3)</td>
<td>7.2</td>
<td>30</td>
</tr>
<tr>
<td>Overall</td>
<td>52.3 (0.7)</td>
<td>8.1</td>
<td>122</td>
</tr>
</tbody>
</table>

\[ a \approx 5.2 \ (0.9) \]
\[ d \approx 3.5 \ (1.3) \]
\[ h^2 \approx 25\% \]
Is it real?

\[ \text{LOD} = \log_{10} \left\{ \frac{\Pr(\text{data} \mid 1 \text{ QTL})}{\Pr(\text{data} \mid \text{no QTL})} \right\} \]

\[ \text{LOD}(D1M1) \approx 2.22 \]
\[ \text{LOD}(D2M2) \approx 7.54 \]

Hypothesis testing

Null hypothesis, \( H_0: \) no QTL

P-value = \( \Pr(\text{LOD} > \text{observed} \mid \text{no QTL}) \)

Small P (large LOD) \( \Rightarrow \) Reject \( H_0 \) (Good)
Large P (small LOD) \( \Rightarrow \) Fail to reject \( H_0 \) (Bad)

Generally want P < 0.05 or < 0.01

\[ P \approx 0.049 \equiv P \approx 0.051 \]
\[ \text{LOD} \approx 3.01 \equiv \text{LOD} \approx 2.99 \]
Multiple testing

→ We’re doing ~ 200 tests (one at each marker; correlated due to linkage)

→ Imagine the tests were uncorrelated, and that $H_0$ is true (there is no QTL)

  Toss 200 biased coins

  Heads $\equiv$ Reject $H_0$ (falsely conclude that there is a QTL)

  $\Pr(\text{Heads}) = 5\%$

Ave no. heads in 200 tosses $= 10$

Pr(at least one head in 200 tosses) $\approx 100\%$
Interval mapping
(Lander and Botstein 1989)

Interpolation between markers
At each point, imagine a putative QTL and maximize Pr(data | QTL, parameters)
Great for dealing with missing genotype data; important for widely spaced markers

Power

Power = Pr(Identify a QTL | there is a QTL)

Power depends on
- Sample size
- Size of QTL effect (relative to resid. var.)
- Marker density
- Level of statistical significance

Consider
- Pr(detect a particular locus)
- Pr(detect at least one locus)
If the power to detect a particular locus is not super high, its estimated effect (when it is identified) will be biased.

- \( \text{Power} \approx 90\% \Rightarrow \text{Bias} \approx 2\% \)
- \( \text{Power} \approx 45\% \Rightarrow \text{Bias} \approx 20\% \)
- \( \text{Power} \approx 5\% \Rightarrow \text{Bias} \approx 100\% \)
Multiple QTLs

It is often important to consider multiple QTLs simultaneously

- Increase power by reducing residual variation
- Separate linked loci
- Estimate epistatic effects

Analysis of single QTL:
- analysis of variance (ANOVA) or simple linear regression

Analysis of multiple QTL:
- multiple linear regression, possibly with interaction terms; possibly using tree-based models

A key issue: Things are more complicated than "Is there a QTL here or not?"

An example

Full model

<table>
<thead>
<tr>
<th>Locus 1</th>
<th>Locus 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td></td>
</tr>
<tr>
<td>BB</td>
<td></td>
</tr>
</tbody>
</table>

Additive model

<table>
<thead>
<tr>
<th>Locus 1</th>
<th>Locus 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td></td>
</tr>
<tr>
<td>BB</td>
<td></td>
</tr>
</tbody>
</table>
A tree-based model

Summary

- LOD scores
- Hypothesis testing
  - Null hypothesis
  - P-values
  - Significance levels
- Adjustment for multiple tests
- Power
  - To identify a particular locus
  - To identify at least one locus
- Selection bias
- Multiple QTLs
  - Increase power
  - Separate linked loci
  - Estimate epistasis
  - Things get complicated