Topics in Statistical Genetics

Karl W Broman

Department of Biostatistics
Johns Hopkins School of Public Health
Past

- QTLs in experimental crosses
- CEPH data \{ 8 families (~ 130 people), >8000 STRPs \}
  - Genetic maps
  - Recombinational variation
  - Autozygosity
  - Crossover interference
- Genotyping/pedigree errors
- Dog genetic maps
- Various disease projects
Traditional genetics: **binary** traits
  e.g. disease or no disease

**Quantitative** traits
  e.g. yield of tomato crop
    # bristles on a fly’s abdomen
  – Many genes
  – Environmental variation

**Why?**
  – Biochemical basis of trait
  – Selection experiments
  – Evolution

**Goal:** find (some of) the genes
Backcross experiment

LL × HH → HL

Backcross progeny (LL or HL)
Genetic markers: STRPs or microsatellites

GATAGATA \ldots GATA
CTATCTAT \ldots CTAT
Data

Phenotypes (trait values)

\[ y_i = \text{phenotype for individual } i \]

Marker genotypes

\[ x_{ij} = 1/0 \text{ if } i \text{ is HL/LL at marker } j \]

Genetic map

Locations of markers

Models

Recombination: No interference

Phenotype/genotype connection

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

\[ \epsilon \sim \text{Normal}(0, \sigma^2) \]
Problem

100 to 1000 backcross progeny
100 to 400 markers

\[ y = \mu + \sum \beta_j x_j + \varepsilon \]

Find the \( x \)'s with \( \beta_j \neq 0 \)

Errors:

- Miss important loci
- Include extraneous loci
Discussion

• Identifying QTLs is model selection

• Simulation studies are necessary
  – compare procedures
  – understand a procedure’s performance

• Different situations will require different procedures
Human data

- www.marshmed.org/genetics
- 8 CEPH families
  - three generations
  - 11 to 15 progeny
  - 92 meioses
- ~8,000 STRP markers
  - 90 ± 7 % typed
- Average spacing
  - female: 0.6 ± 1.2 cM
  - male: 0.4 ± 1.0 cM
  - sex-ave: 0.5 ± 0.9 cM
- Data cleaning
  - Removed 764/964,425 (~0.08%) genotypes resulting in tight double recombinants
Meiosis
Interference

- **Strand choice**
  \[\rightarrow\text{ Chromatid interference}\]

- **Spacing**
  \[\rightarrow\text{ Chiasma (crossover) interference}\]
Genetic distance

distance (cM) = average # crossovers
              in 100 meiotic products

per Morgan  \{ 2 chiasmata on 4-strand bundle
             1 crossover on meiotic product

Map function

recombination fraction as a function of genetic distance

Haldane \quad r(d) = \frac{1}{2} \left[ 1 - \exp(-2d) \right]

Kosambi \quad r(d) = \frac{1}{2} \tanh(2d)
Crossovers in Mother

Total Number of Crossovers
Chromosome 1

Female/Male Distance Ratio vs Sex-averaged Location
Chromosome 4

Sex-averaged Location

Female/Male Distance Ratio

0

50

100

150

200

0

0.5

1.0

1.5

2.0

2.5

3.0

3.5

4.0

4.5

5.0

5.5

6.0
CEPH: homozygous regions

<table>
<thead>
<tr>
<th>Individual</th>
<th>Number of regions</th>
<th>Length (cM)</th>
<th>Total length (cM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>884-01 (father)</td>
<td>10</td>
<td>2 – 23</td>
<td>116</td>
</tr>
<tr>
<td>884-02 (mother)</td>
<td>11</td>
<td>5 – 40</td>
<td>165</td>
</tr>
<tr>
<td>884: grandparents</td>
<td>2</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>7 – 10</td>
<td>1 – 29</td>
<td>58 – 120</td>
</tr>
<tr>
<td>884: 12 progeny</td>
<td>5 – 18</td>
<td>2 – 36</td>
<td>62 – 204</td>
</tr>
<tr>
<td>102: 14 progeny</td>
<td>4 – 13</td>
<td>1 – 39</td>
<td>85 – 254</td>
</tr>
<tr>
<td>1331-12 (GP)</td>
<td>2</td>
<td>1 – 6</td>
<td>7</td>
</tr>
<tr>
<td>1416-14 (GP)</td>
<td>4</td>
<td>6 – 29</td>
<td>76</td>
</tr>
<tr>
<td>35/100 of the others</td>
<td>1 – 2</td>
<td>0 – 7</td>
<td></td>
</tr>
</tbody>
</table>
Present

• Finish the past
• Yellowstone wolves
• Sib pairs where parents are unavailable
• Various disease projects
Future

• Large genetic/epidemiological studies
  – Many people
  – Many phenotypes
  – Many environmental covariates

• Highly computational statistical genetic methods

• Various disease projects