Genetic analysis of high-throughput phenotypes
Challenges and opportunities

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman
Slides: bit.ly/sgn2017
Intercross

\[
\begin{array}{c}
\text{P}_1 \\
\downarrow \\
\text{F}_1 \\
\downarrow \\
\text{F}_2 \\
\downarrow \\
\text{F}_1 \\
\downarrow \\
\text{P}_2 \\
\end{array}
\]
QTL mapping
Genome scan
Permutation test
Permutation test
Histogram of permutation results

Frequency

Maximum LOD score

1 2 3 4 5 6 7

0
100
200
300
400
Modeling multiple QTL

- Reduce residual variation $\rightarrow$ increased power
- Separate linked QTL
- Identify interactions among QTL (epistasis)
Epistasis in $F_2$

Additive

Epistatic

QTL 1

QTL 2

Ave. phenotype

A H B

0 20 40 60 80 100

A H B

0 20 40 60 80 100

A H B
Improving precision

- more recombinations
- more individuals
- more precise phenotype
- lower-level phenotypes
  - transcripts, proteins, metabolites
Advanced intercross lines

P
A
B
F2
F3
F4
F7
F10
Recombinant inbred lines
Collaborative Cross

$G_0$

A B

$G_1$

A B

$G_2$

ABCD

$G_3$

$G_4$

$G_\infty$

EFGH
Heterogeneous stock
Genome-scale phenotypes
Challenges: diagnostics
Challenges: diagnostics

www.biostat.wisc.edu/~kbroman/D3/manyboxplots
Challenges: diagnostics

- What might have gone wrong?
- How might it be revealed?
- Make lots of graphs
- Follow up artifacts
Challenges: scale of results

- genotypes
- phenotypes
Challenges: scale of results

- genotypes
- phenotypes
- results
Challenges: organizing, automating

genotypes
phenotypes
Challenges: organizing, automating
Challenges: organizing, automating

- genotypes
- phenotypes
Challenges: organizing, automating

- genotypes
- phenotypes
Challenges: organizing, automating

genotypes
phenotypes
Challenges: organizing, automating

<table>
<thead>
<tr>
<th>genotypes</th>
<th>phenotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Challenges: organizing, automating

<table>
<thead>
<tr>
<th>genotypes</th>
<th>phenotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Challenges: metadata

What the heck is "FAD_NAD SI 8.3_3.3G"?
What was the question again?
The ridiculome
Pleiotropy?
Multivariate phenotypes
Multivariate phenotypes
Composite phenotypes

share more data, sooner
Are your results reproducible?

cf Baggerly & Coombes (2009)
projecteuclid.org/euclid.aoas/1267453942
Karl -- this is very interesting, however you used an old version of the data (n=143 rather than n=226).

I'm really sorry you did all that work on the incomplete dataset.

Bruce
Steps toward reproducible research

1. Organize your data & code
2. Everything with a script
3. Automate the process (GNU Make)
4. Turn scripts into reproducible reports
5. Turn repeated code into functions
6. Create a package/module
7. Use version control (git/GitHub)
8. Pick a license, any license
Slides: bit.ly/sgn2017

kbroman.org

github.com/kbroman

@kwbroman