18 years of R/qtl
developing, maintaining, and supporting
scientific software

Karl Broman
Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbrroman
Slides: bit.ly/UCSF2018
17 years of R/qtl

Lines of code vs Year

- R
- C
- doc
- idea
- svn
- git

Intercross
Data
QTL mapping

Chromosome LOD score

BB BR RR

0.8 0.9 1.0 1.1
Interactive plot

bit.ly/lod_and_effect
17 years of R/qtl

Lines of code vs. Year

- R
- C
- doc

Year:

Language:
- R
- C
- doc

Lines of code:
0 5000 10000 15000 20000 25000 30000 35000 40000
Good things
Good things

► some of the code
► basics of the user interface
► diagnostics and data visualization
► quite comprehensive
► quite flexible
Bad things
## Input file

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>liver</td>
<td>spleen</td>
<td>sex</td>
<td>pgm</td>
<td>D1Mit18</td>
<td>D1Mit80</td>
<td>D1Mit17</td>
<td>D2Mit379</td>
<td>D2Mit75</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27.3</td>
<td>51.4</td>
<td>110.4</td>
<td>38.3</td>
<td>48.1</td>
</tr>
<tr>
<td>4</td>
<td>61.92</td>
<td>153.16</td>
<td>m</td>
<td>1</td>
<td>BB</td>
<td>SB</td>
<td>SB</td>
<td>SB</td>
<td>SB</td>
</tr>
<tr>
<td>5</td>
<td>88.33</td>
<td>178.58</td>
<td>m</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>BB</td>
<td>BB</td>
</tr>
<tr>
<td>6</td>
<td>58</td>
<td>131.91</td>
<td>m</td>
<td>1</td>
<td>BB</td>
<td>SB</td>
<td>SB</td>
<td>SB</td>
<td>SB</td>
</tr>
<tr>
<td>7</td>
<td>78.06</td>
<td>126.13</td>
<td>m</td>
<td>1</td>
<td>SB</td>
<td>SB</td>
<td>BB</td>
<td>SS</td>
<td>SS</td>
</tr>
<tr>
<td>8</td>
<td>65.31</td>
<td>181.05</td>
<td>m</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>SB</td>
<td>SB</td>
</tr>
<tr>
<td>9</td>
<td>59.26</td>
<td>191.54</td>
<td>m</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>SS</td>
<td>SS</td>
</tr>
<tr>
<td>10</td>
<td>59.47</td>
<td>154.88</td>
<td>m</td>
<td>1</td>
<td>BB</td>
<td>BB</td>
<td>BB</td>
<td>SB</td>
<td>SB</td>
</tr>
<tr>
<td>11</td>
<td>65.63</td>
<td>184.12</td>
<td>m</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>SB</td>
<td>SB</td>
</tr>
<tr>
<td>12</td>
<td>38.64</td>
<td>133.05</td>
<td>m</td>
<td>1</td>
<td>SB</td>
<td>BB</td>
<td>SB</td>
<td>SB</td>
<td>SB</td>
</tr>
<tr>
<td>13</td>
<td>60.94</td>
<td>275.63</td>
<td>m</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>SB</td>
<td>BB</td>
</tr>
<tr>
<td>14</td>
<td>51.48</td>
<td>395.25</td>
<td>m</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>SB</td>
<td>BB</td>
</tr>
<tr>
<td>15</td>
<td>47.12</td>
<td>260.45</td>
<td>m</td>
<td>1</td>
<td>BB</td>
<td>SB</td>
<td>SB</td>
<td>BB</td>
<td>BB</td>
</tr>
</tbody>
</table>
n <- ncol(data)
temp <- rep(FALSE,n)
for(i in 1:n) {
  temp[i] <- all(data[2,1:i]=="")
  if(!temp[i]) break
}
if(!any(temp)) stop("...")
n.phe <- max(((1:n)[temp]))

kbroman.org/blog/2011/08/17/the-stupidest-r-code-ever
Open source means everyone can see my stupid mistakes.
Open source means everyone can see my stupid mistakes

Version control means everyone can see every stupid mistake I’ve ever made
Documentation
QTL mapping

Chromosome
LOD score

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
BB BR RR
0.8
0.9
1.0
1.1
Congenic line
Improving precision

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

G_0
A
B

G_1
A
B

G_2
ABCD

G_3

G_4

... 

G_\infty
Heterogeneous stock

<table>
<thead>
<tr>
<th>G0</th>
<th>G1</th>
<th>G2</th>
<th>G10</th>
<th>G15</th>
<th>G20</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>G1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>G2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>G15</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>G20</td>
</tr>
</tbody>
</table>
Genome-scale phenotypes
Challenges: diagnostics

kbroman.org/blog/2012/04/25/microarrays-suck
Challenges: diagnostics
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
Challenges: organizing, automating

- genotypes
- phenotypes
Challenges: organizing, automating

genotypes
phenotypes
Challenges: metadata

What the heck is "FAD_NAD SI 8.3_3.3G"?
What was the question again?
- High-density genotypes
- High-dimensional phenotypes
- Multi-parent populations
- Linear mixed models
R/qtl2: Let’s not make the same mistakes

- C++ and Rcpp
- Roxygen2 for documentation
- Unit tests
- A single “switch” for cross type
R/qtl2: Let’s not make the same mistakes

- C++ and Rcpp
- Roxygen2 for documentation
- Unit tests
- A single “switch” for cross type
- Yet another data input format
- Flatter data structures, but still complex
Sustainable academic software
Acknowledgments

Danny Arends
Gary Churchill
Nick Furlotte
Dan Gatti
Ritsert Jansen
Pjotr Prins
Śaunak Sen
Petr Simecek
Artem Tarasov
Hao Wu
Brian Yandell

Robert Corty
Timothée Flutre
Lars Ronnegard
Rohan Shah
Laura Shannon
Quoc Tran
Aaron Wolen

NIH/NIGMS
Slides: bit.ly/UCSF2018

kbroman.org

kbroman.org/qtl2

github.com/kbroman

@kwbroman