R/qtl2

high-dimensional data & multi-parent populations

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

`kbroman.org`
`github.com/kbroman`
`@kwbroman`

Slides: `bit.ly/uncc2017`

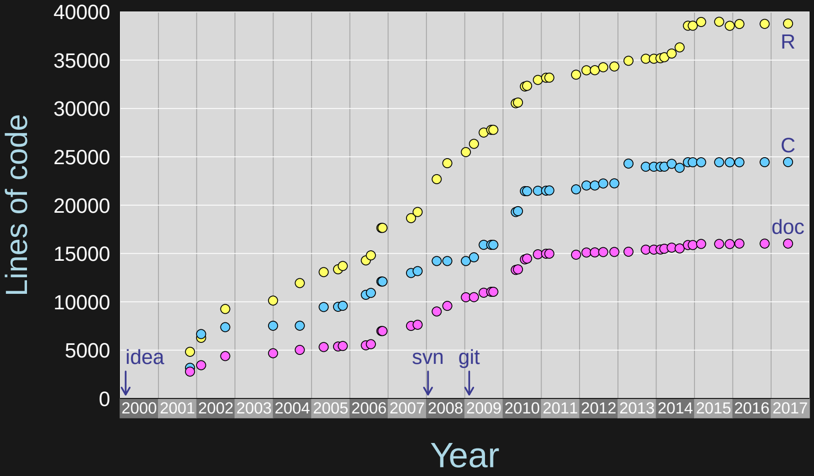
New

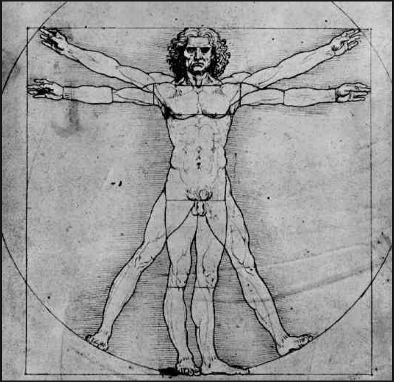
University of Wisconsin–Madison

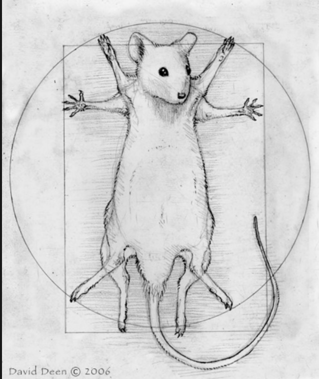
PhD in Biomedical Data Science

`bit.ly/MadBDS`

17 years of R/qtl

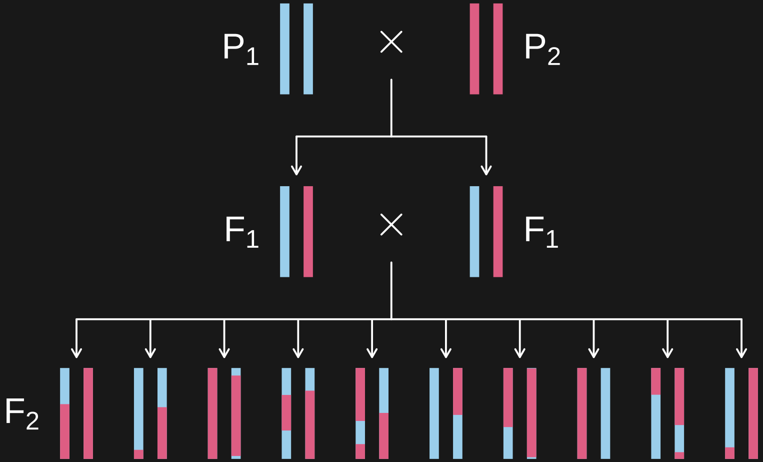




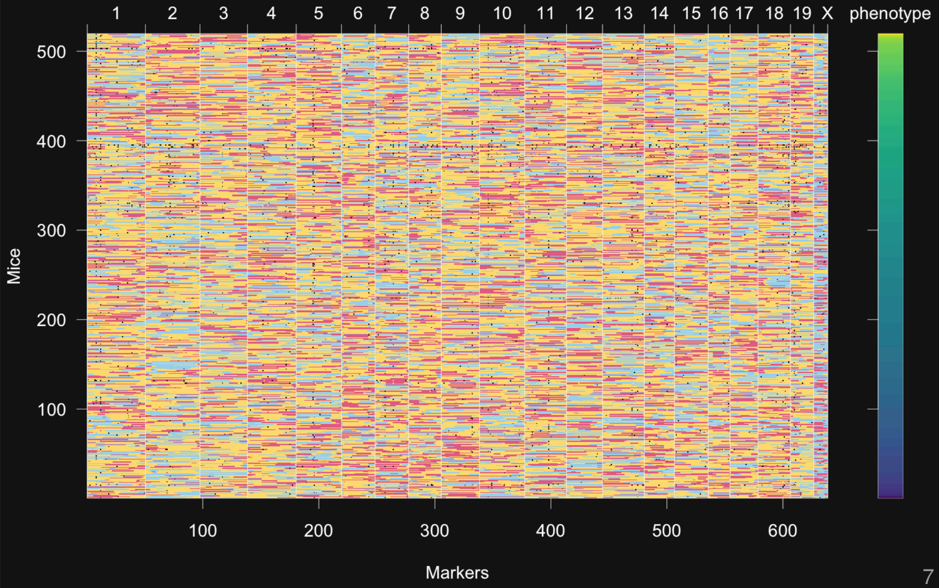


daviddeen.com

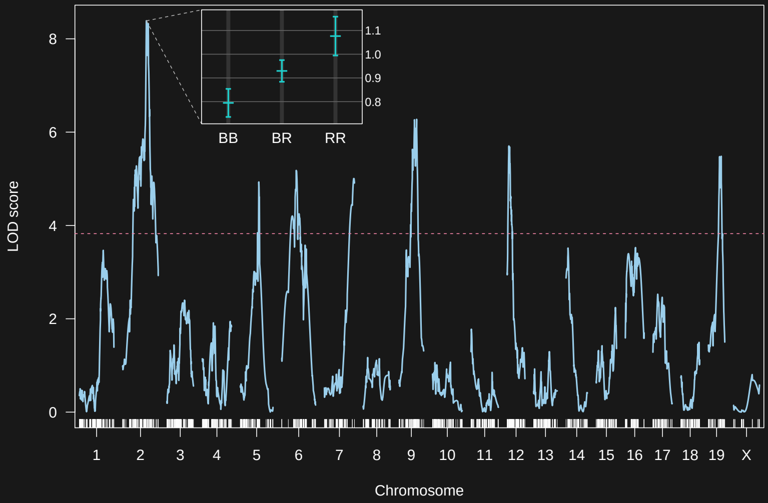
Intercross



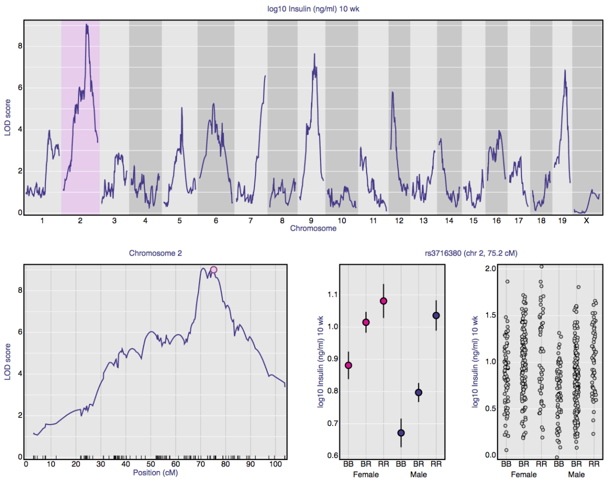
Data


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
phenotype
500
400
300
200
100
Mice
100
200
300
400
500
600
Markers

QTL mapping

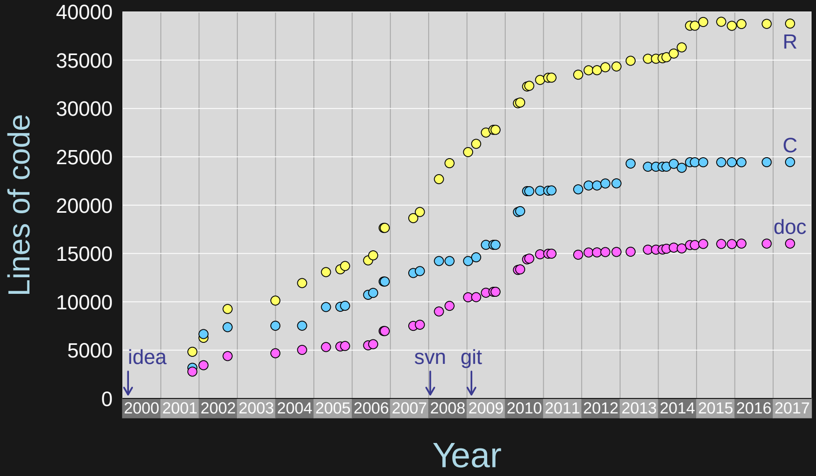

LOD score
0
2
4
6
8
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
BB
BR
RR
0.8
0.9
1.0
1.1

Interactive plot



bit.ly/lod_and_effect

17 years of R/qtl



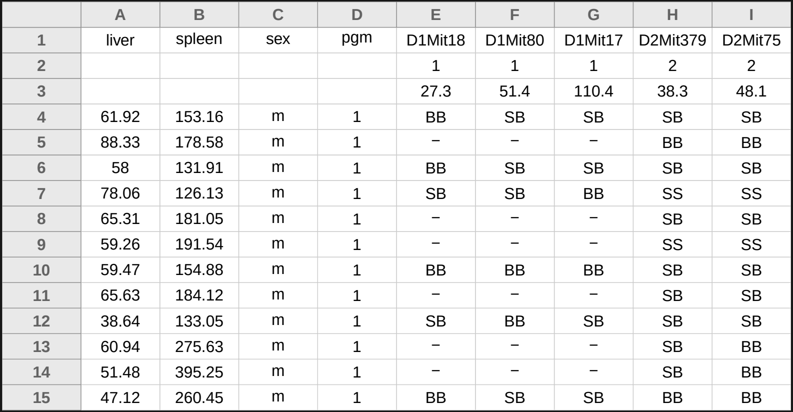
Good things

Good things

- some of the code
- basics of the user interface
- diagnostics and data visualization
- quite comprehensive
- quite flexible

Bad things

Input file



	A	B	C	D	E	F	G	H	I
1	liver	spleen	sex	pgm	D1Mit18	D1Mit80	D1Mit17	D2Mit379	D2Mit75
2					1	1	1	2	2
3					27.3	51.4	110.4	38.3	48.1
4	61.92	153.16	m	1	BB	SB	SB	SB	SB
5	88.33	178.58	m	1	–	–	–	BB	BB
6	58	131.91	m	1	BB	SB	SB	SB	SB
7	78.06	126.13	m	1	SB	SB	BB	SS	SS
8	65.31	181.05	m	1	–	–	–	SB	SB
9	59.26	191.54	m	1	–	–	–	SS	SS
10	59.47	154.88	m	1	BB	BB	BB	SB	SB
11	65.63	184.12	m	1	–	–	–	SB	SB
12	38.64	133.05	m	1	SB	BB	SB	SB	SB
13	60.94	275.63	m	1	–	–	–	SB	BB
14	51.48	395.25	m	1	–	–	–	SB	BB
15	47.12	260.45	m	1	BB	SB	SB	BB	BB

Stupidest code ever

```
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

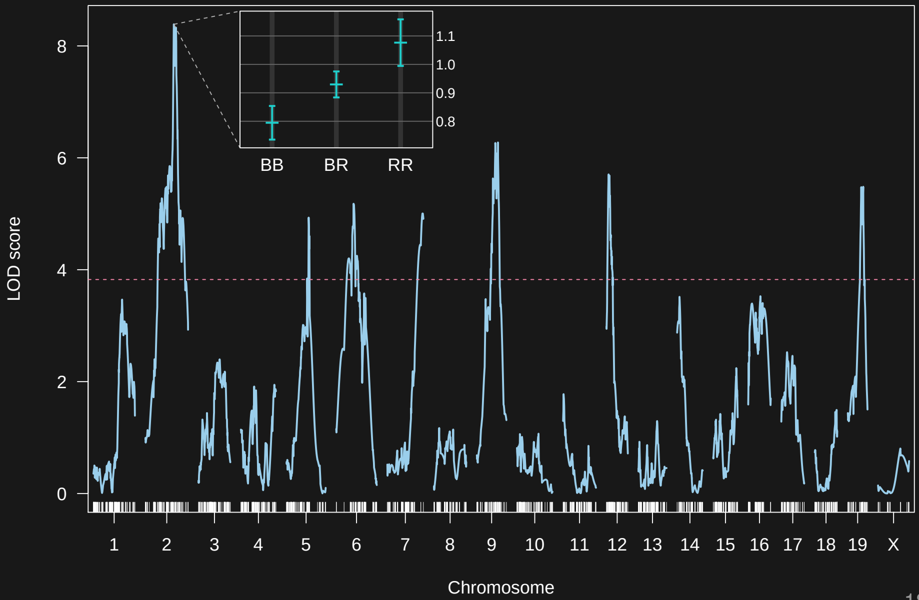
Open source means
everyone can see my stupid mistakes

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

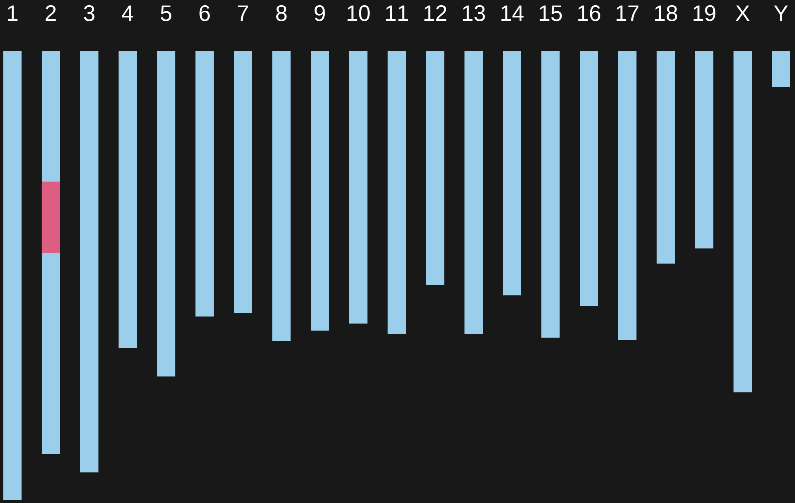
Documentation

Support

QTL mapping


LOD score
8
6
4
2
0
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
BB
BR
RR
1.1
1.0
0.9
0.8

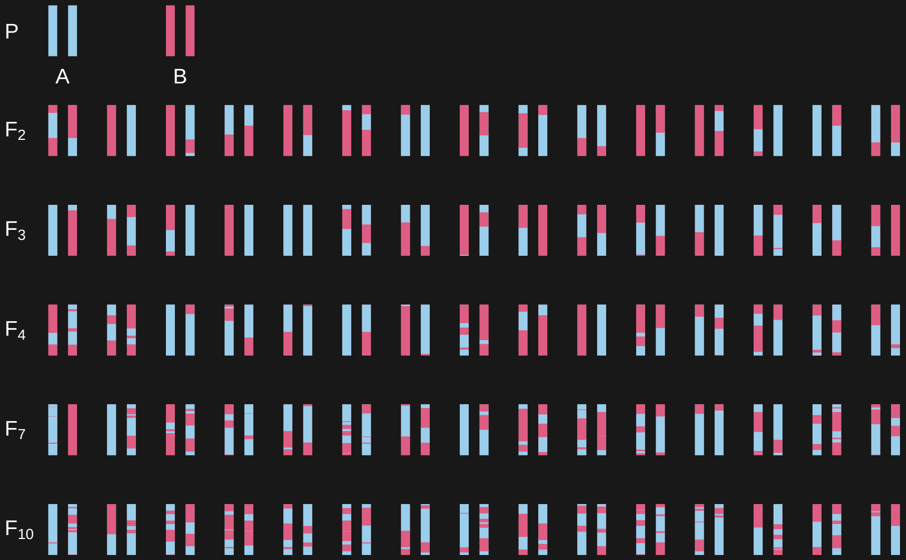
Congenic line



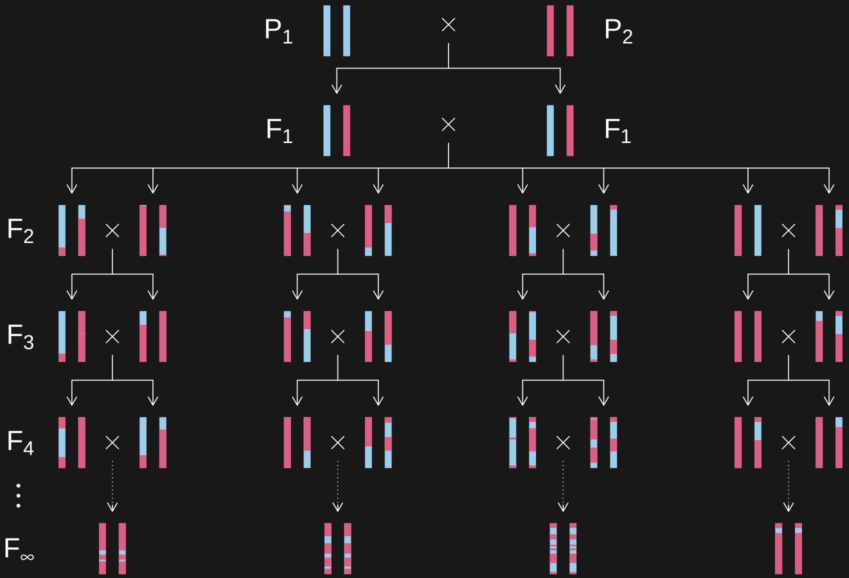
Improving precision

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

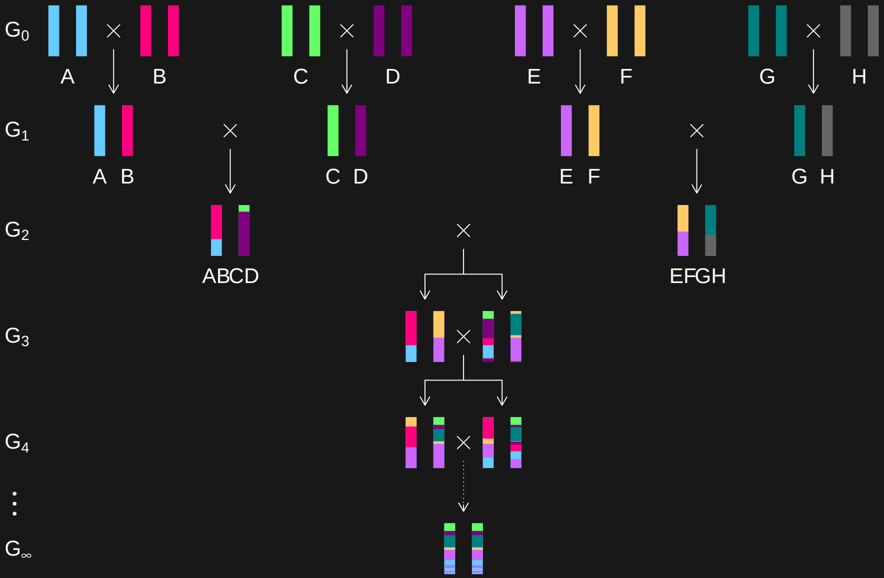
Advanced intercross lines



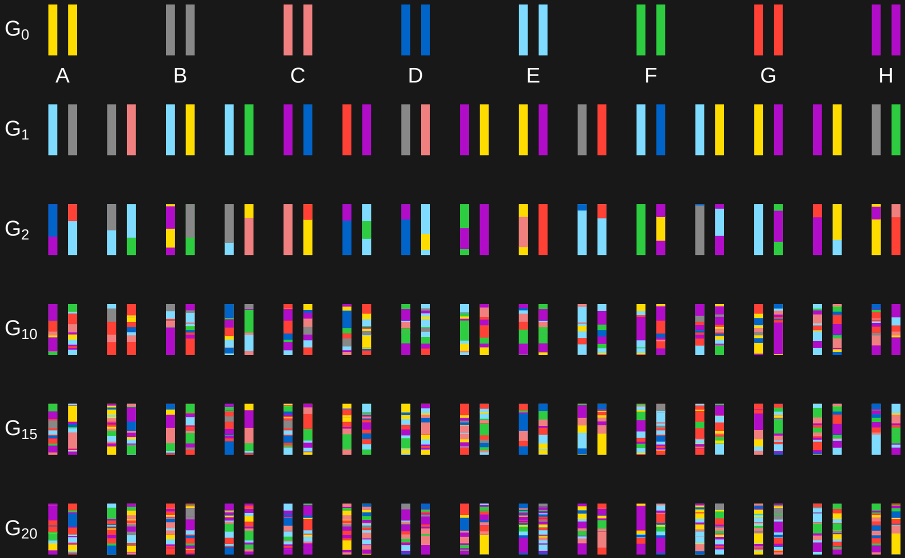
Recombinant inbred lines



Collaborative Cross



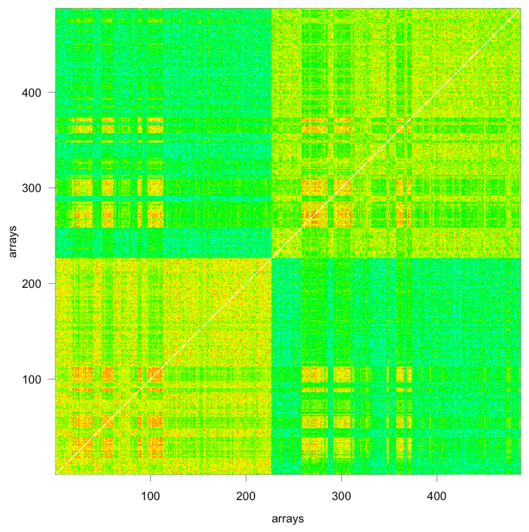
Heterogeneous stock



Genome-scale phenotypes

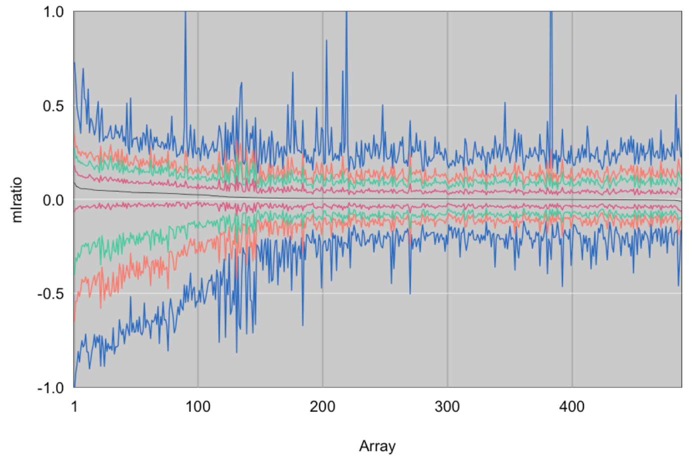
Alan Attie

Challenges: diagnostics



kbroman.org/blog/2012/04/25/microarrays-suck

Challenges: diagnostics

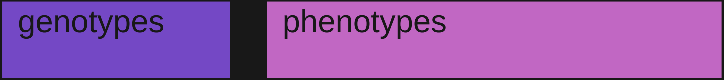


bit.ly/many_boxplots

Challenges: diagnostics

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

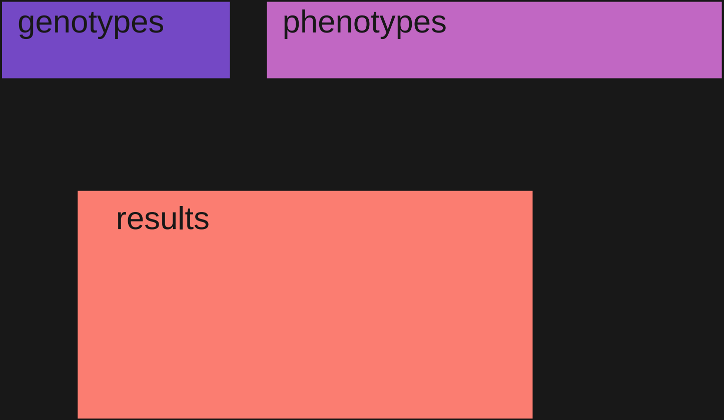
Challenges: scale of results



genotypes

phenotypes

Challenges: scale of results



genotypes

phenotypes

results

Challenges: organizing, automating

genotypes

phenotypes

Challenges: organizing, automating

genotypes

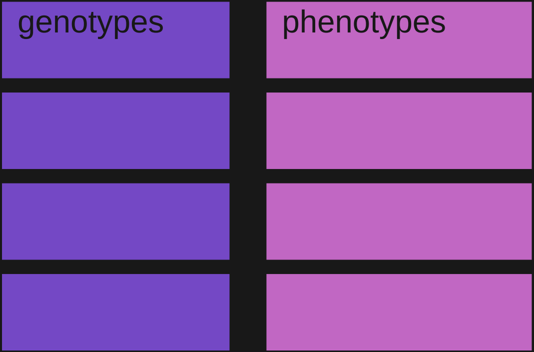
phenotypes

Challenges: organizing, automating

genotypes

phenotypes

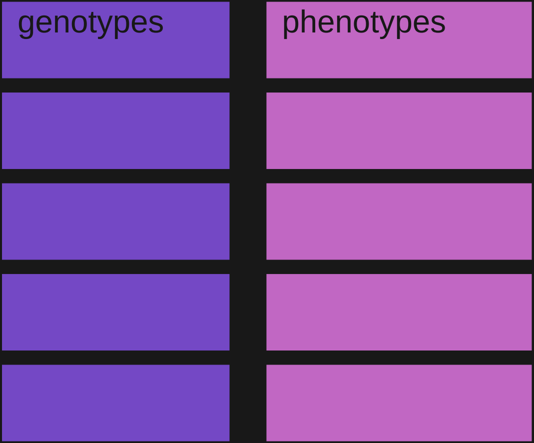
Challenges: organizing, automating



genotypes

phenotypes

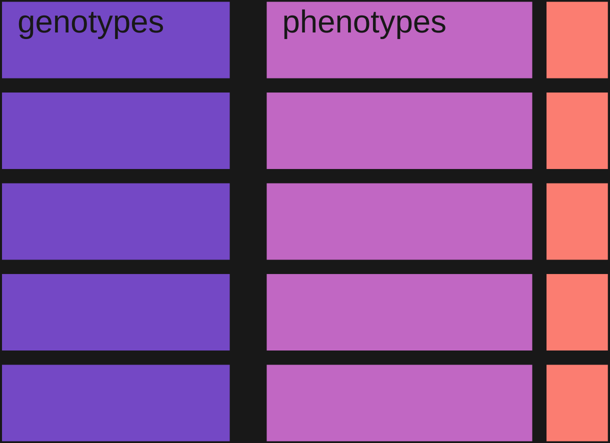
Challenges: organizing, automating



genotypes

phenotypes

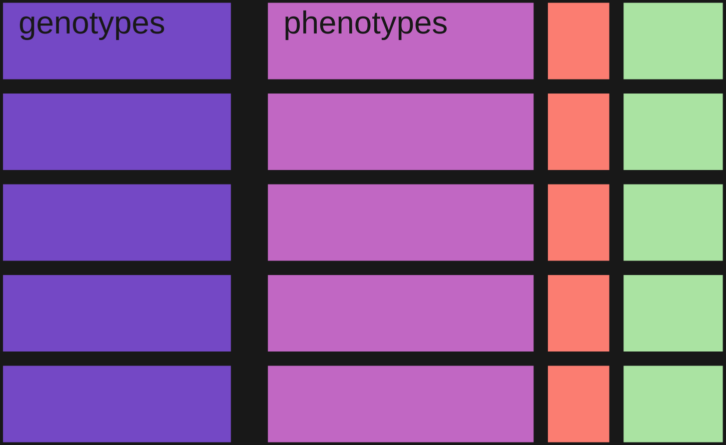
Challenges: organizing, automating



genotypes

phenotypes

Challenges: organizing, automating



Challenges: metadata

What the heck is `"FAD_NAD SI 8.3_3.3G"`?

What was the question again?

ropenscilabs.github.io/miner_book

R/qtl2

- High-density genotypes
- High-dimensional phenotypes
- Multi-parent populations
- Linear mixed models

`kbroman.org/qtl2`

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
- Split into multiple packages
- 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/uncc2017`

`kbroman.org`

`kbroman.org/qtl2`

`github.com/kbroman`

`@kwbroman`