Genetic analysis of high-throughput phenotypes

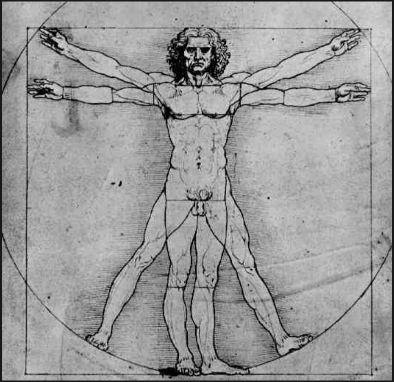
Challenges and opportunities

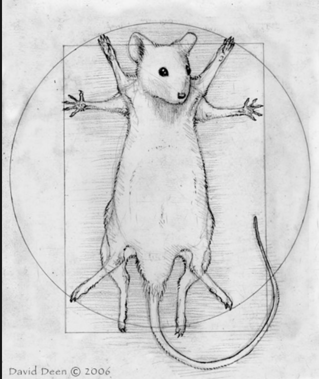
Karl Broman

Biostatistics & Medical Informatics, UW–Madison

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

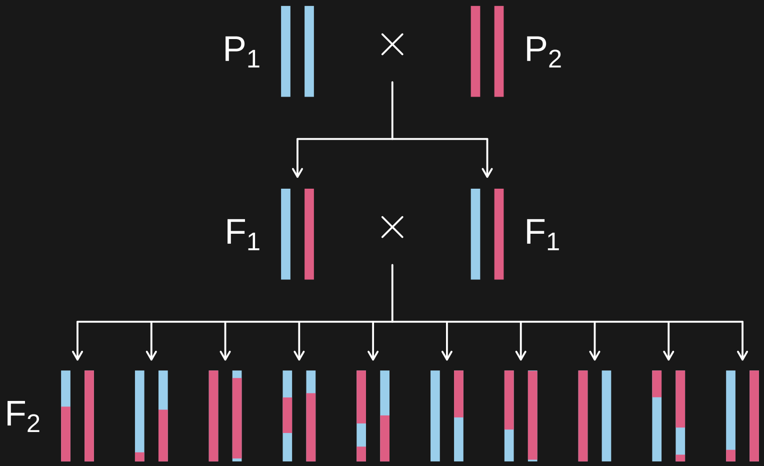
PUBLIC DOMAIN



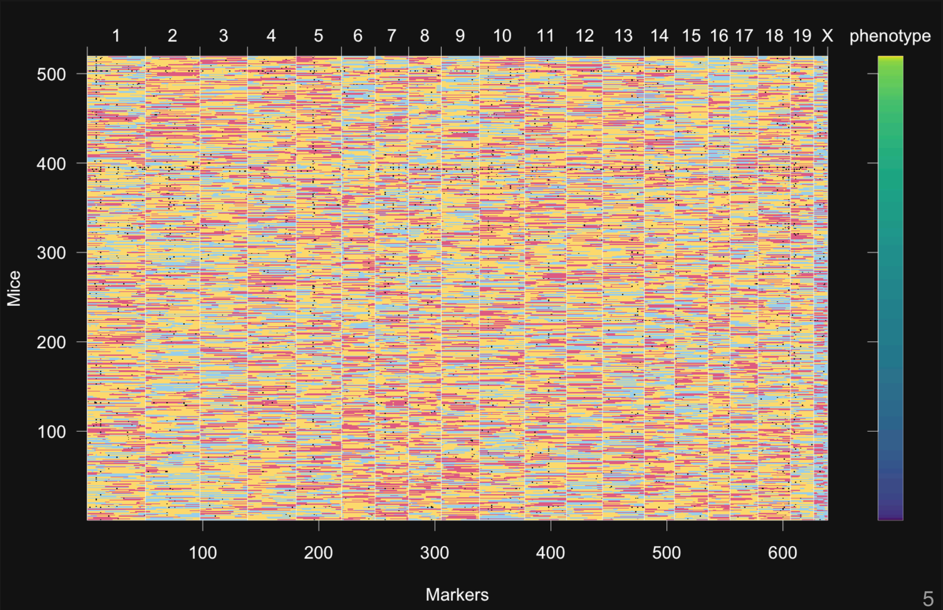


daviddeen.com

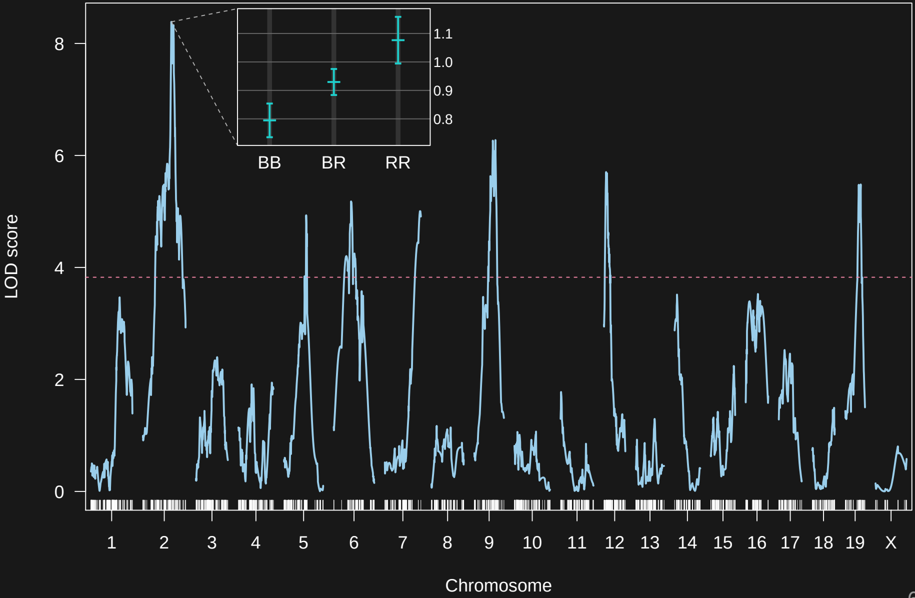
Intercross



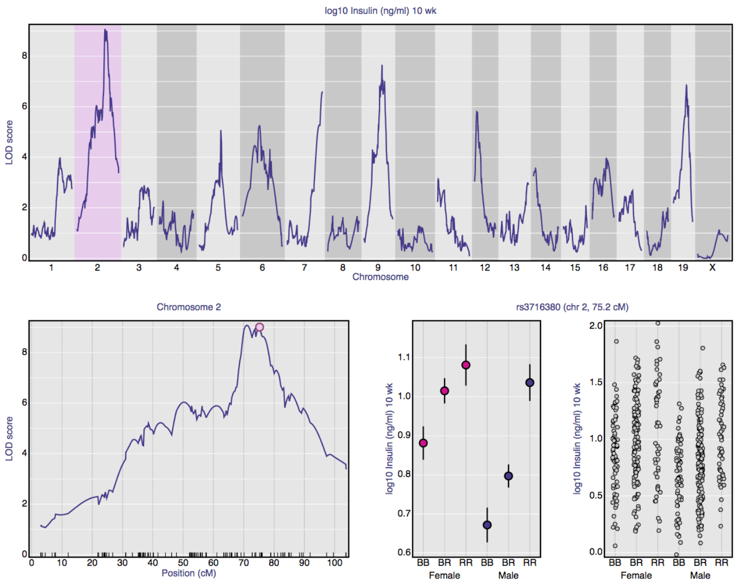
Data


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

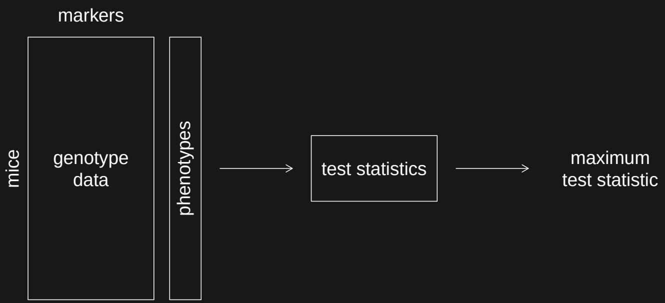
QTL mapping


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
1.1
1.0
0.9
0.8
BB
BR
RR

Genome scan


log10 Insulin (ng/ml) 10 wk
LOD score
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome
Chromosome 2
LOD score
0
2
4
6
8
0
10
20
30
40
50
60
70
80
90
100
Position (cM)
rs3716380 (chr 2, 75.2 cM)
log10 Insulin (ng/ml) 10 wk
0.6
0.7
0.8
0.9
1.0
1.1
BB
BR
RR
BB
BR
RR
Female
Male
log10 Insulin (ng/ml) 10 wk
0.0
0.5
1.0
1.5
2.0
BB
BR
RR
BB
BR
RR
Female
Male

Permutation test



markers

mice

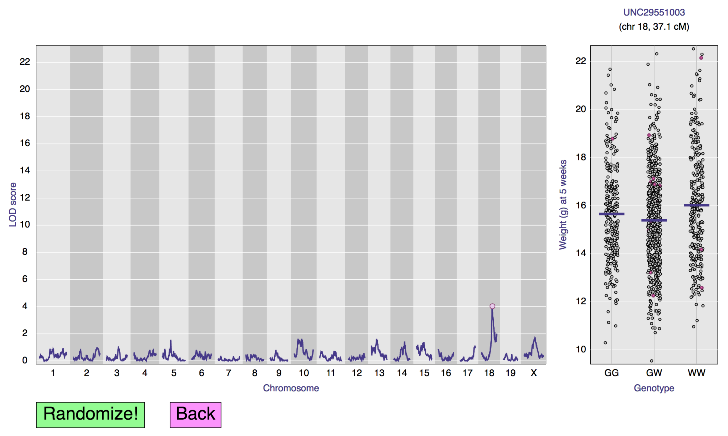
genotype
data

phenotypes

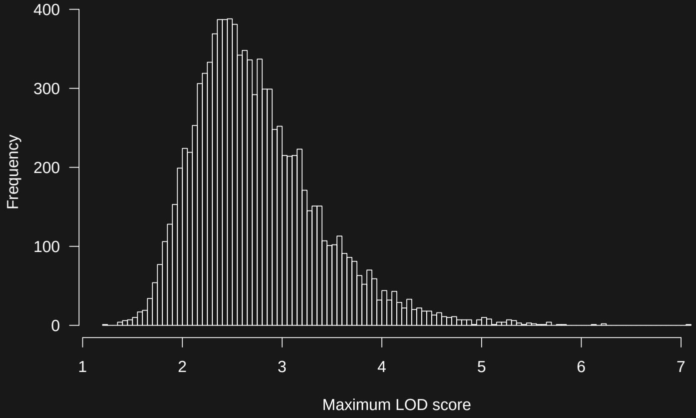
test statistics

maximum
test statistic

Permutation test



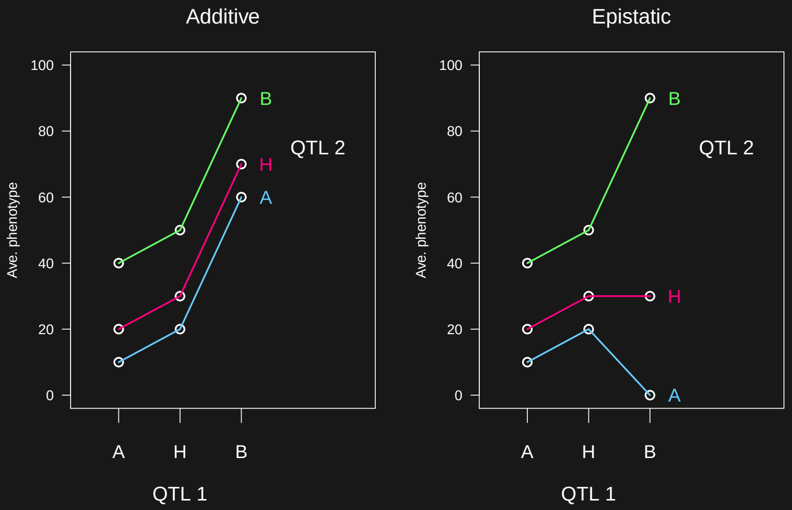
Histogram of permutation results



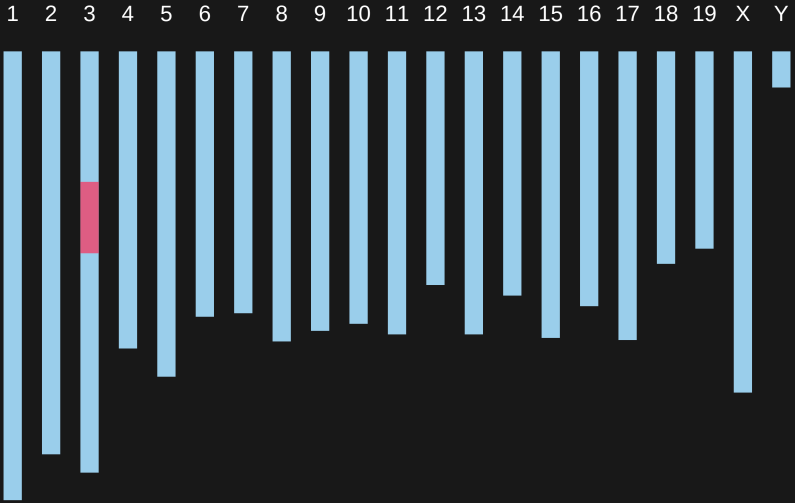
Modeling multiple QTL

- Reduce residual variation ⟶ increased power
- Separate linked QTL
- Identify interactions among QTL (epistasis)

Epistasis in F_2



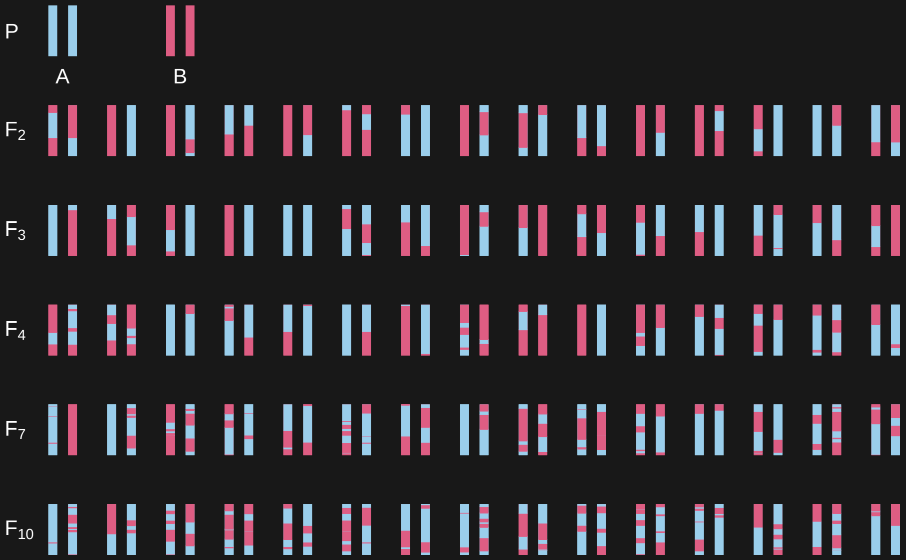
Congenic line



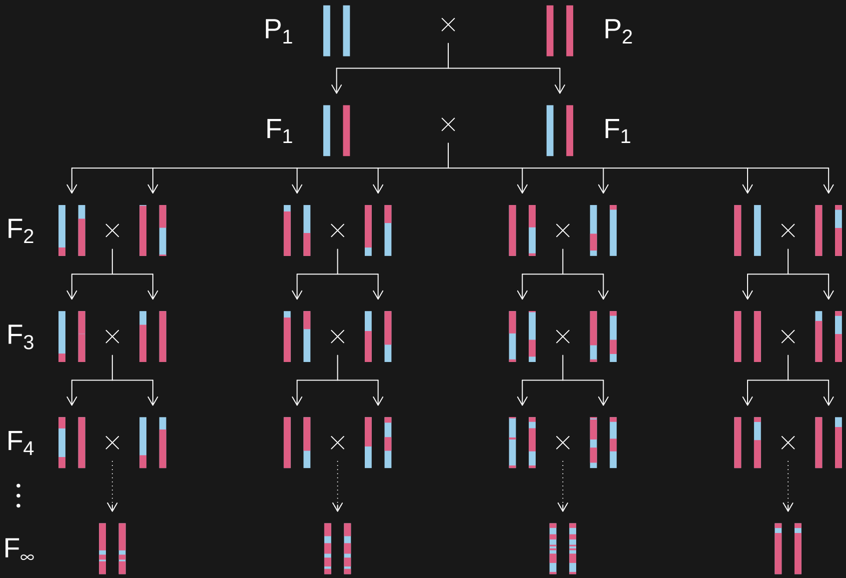
Improving precision

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

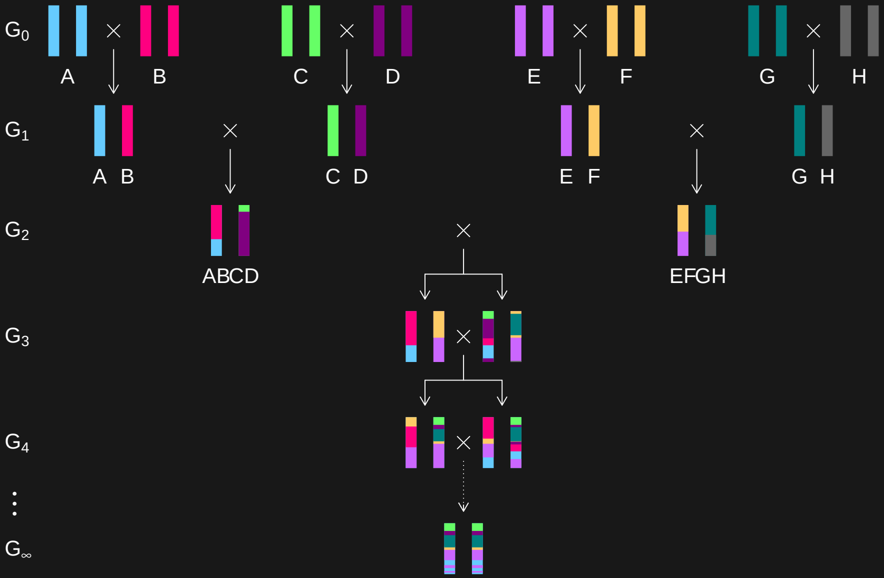
Advanced intercross lines



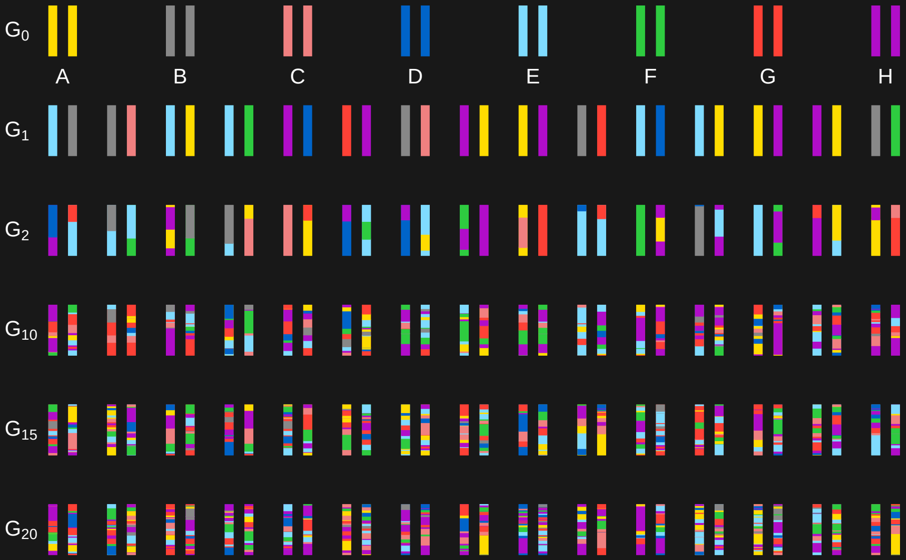
Recombinant inbred lines



Collaborative Cross



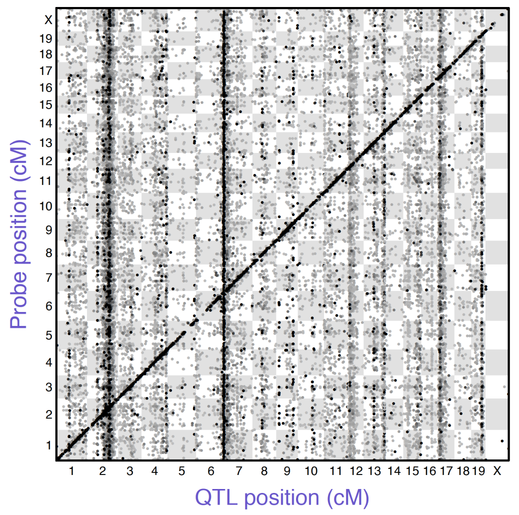
Heterogeneous stock



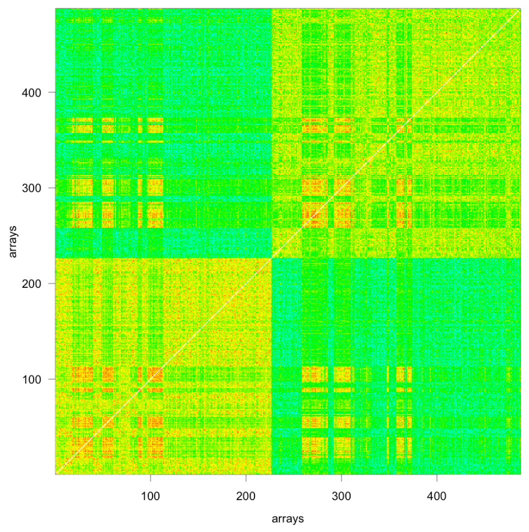
Genome-scale phenotypes

Alan Attie

eQTL

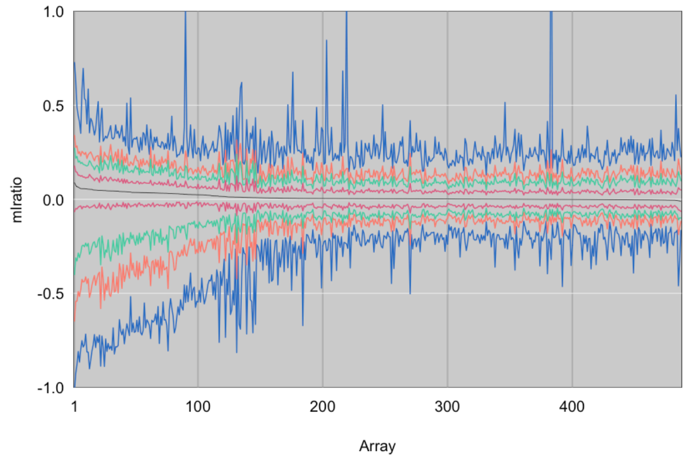

Probe position (cM)
X
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
QTL position (cM)

Challenges: diagnostics



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

Challenges: diagnostics



www.biostat.wisc.edu/~kbroman/D3/manyboxplots

Challenges: diagnostics

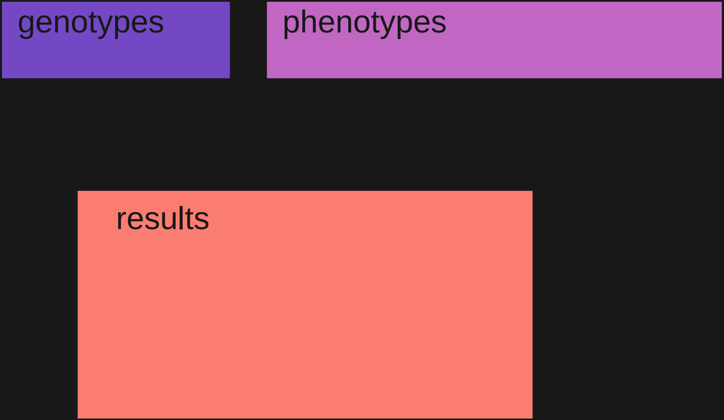
- What might have gone wrong?
- How might it be revealed?
- Make lots of graphs
- Follow up artifacts

Challenges: scale of results

genotypes

phenotypes

Challenges: scale of results

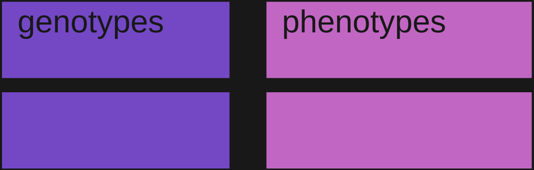


Challenges: organizing, automating

genotypes

phenotypes

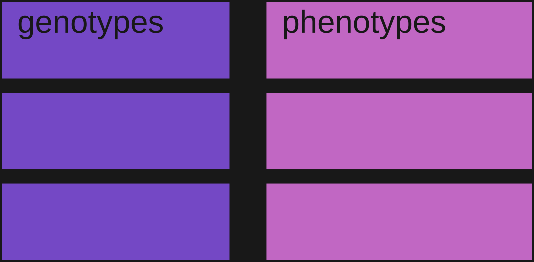
Challenges: organizing, automating



genotypes

phenotypes

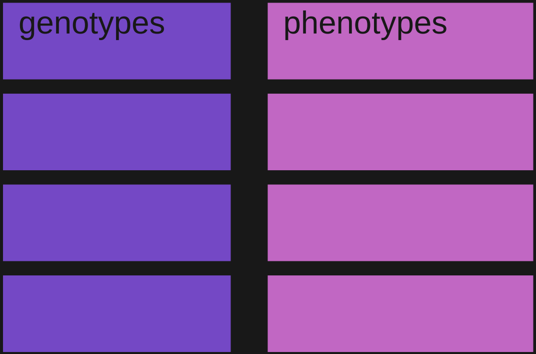
Challenges: organizing, automating



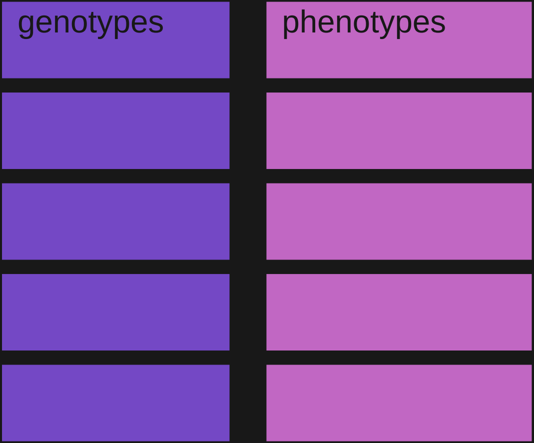
genotypes

phenotypes

Challenges: organizing, automating



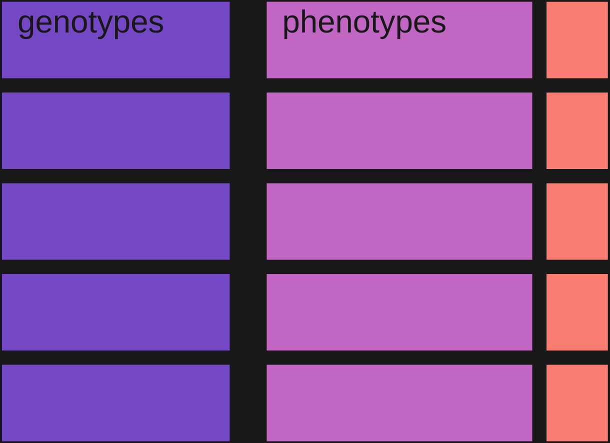
Challenges: organizing, automating



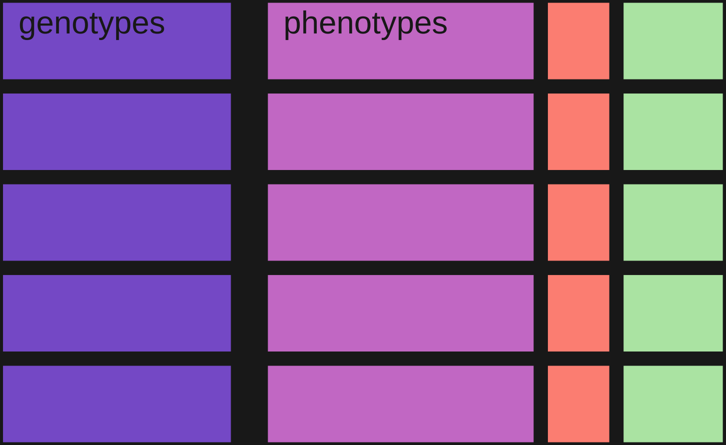
genotypes

phenotypes

Challenges: organizing, automating



Challenges: organizing, automating



genotypes

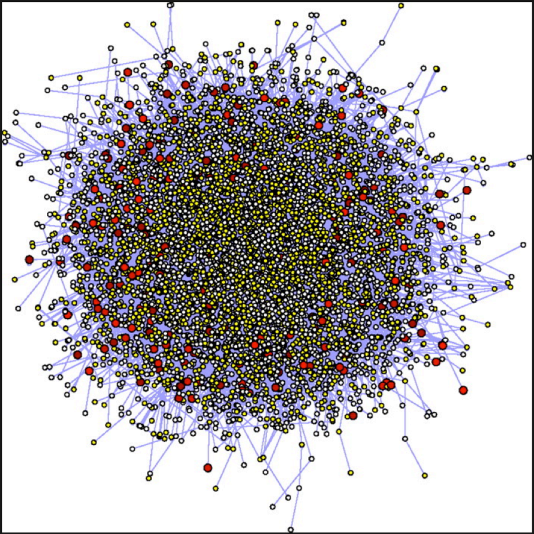
phenotypes

Challenges: metadata

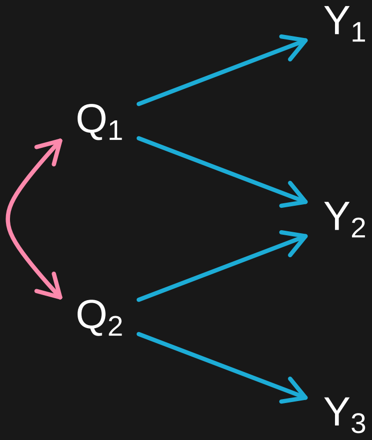
What the heck is `"FAD_NAD SI 8.3_3.3G"`?

What was the question again?

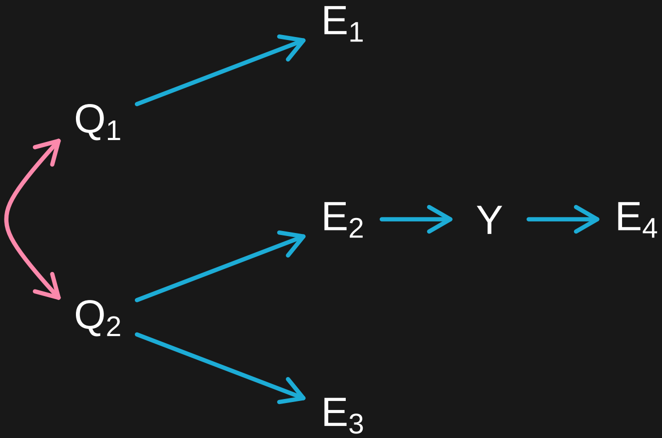
The ridiculome



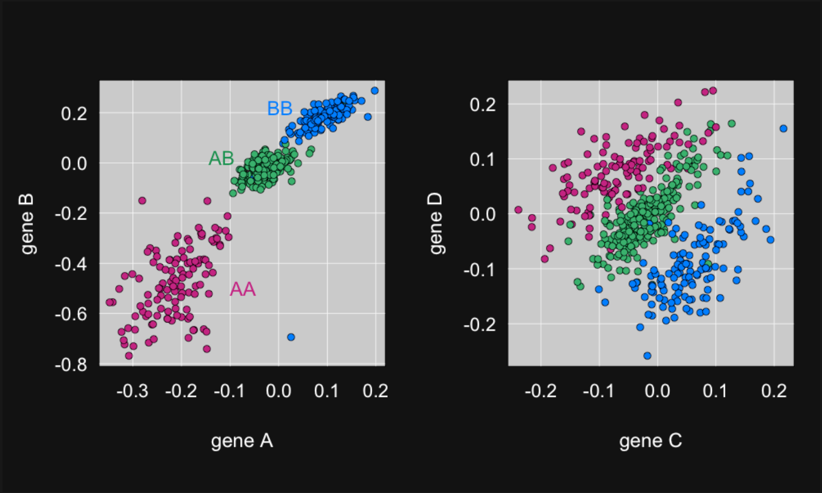
Pleiotropy?



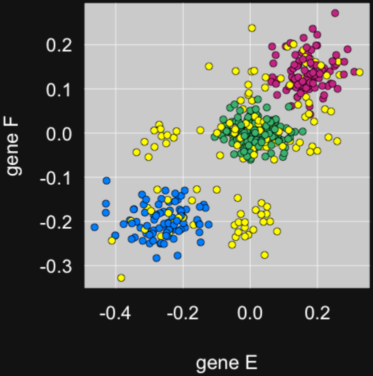
Causal?



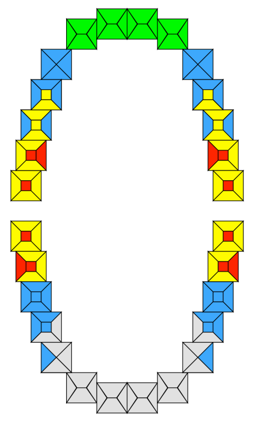
Multivariate phenotypes



Multivariate phenotypes



Composite phenotypes



Shaffer et al. (2013) J Dent Res 92:32-37 share more data, sooner

Are your results reproducible?

cf Baggerly & Coombes (2009)

`projecteuclid.org/euclid.aoas/1267453942`

Reproducibility

```
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

`kbroman.org/steps2rr`

Slides: `bit.ly/sgn2017`

PUBLIC DOMAIN

`kbroman.org`

`github.com/kbroman`

`@kwbroman`