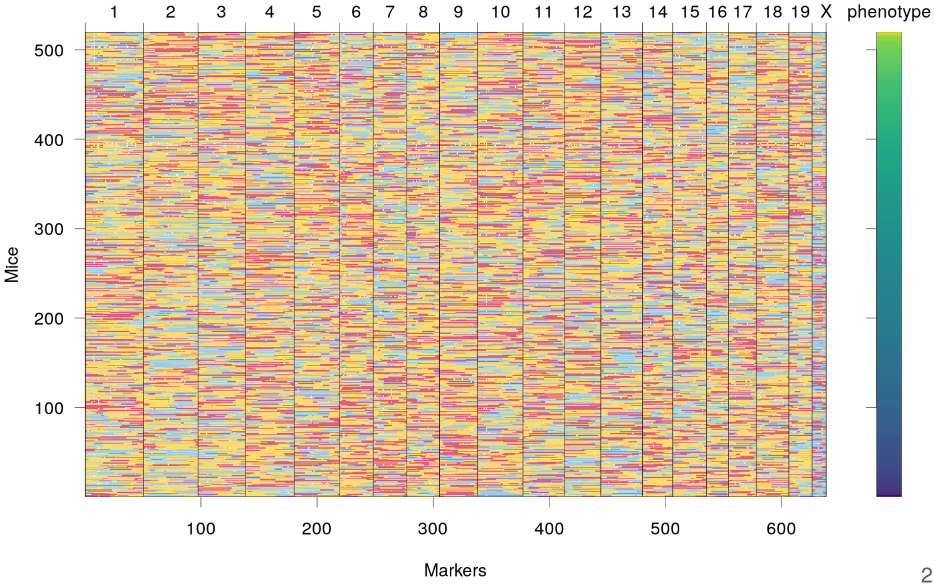
Sample mix-ups in eQTL data

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

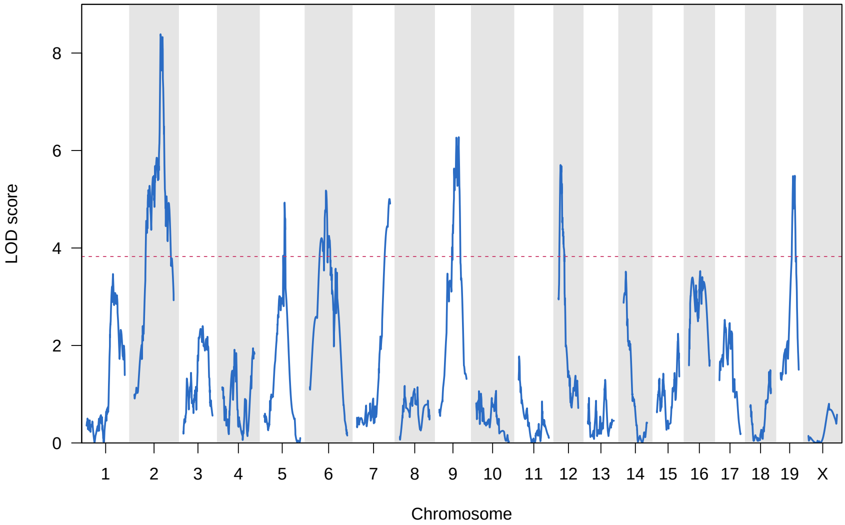
kbroman.org
github.com/kbroman
@kwbroman
Course web: kbroman.org/AdvData

Data

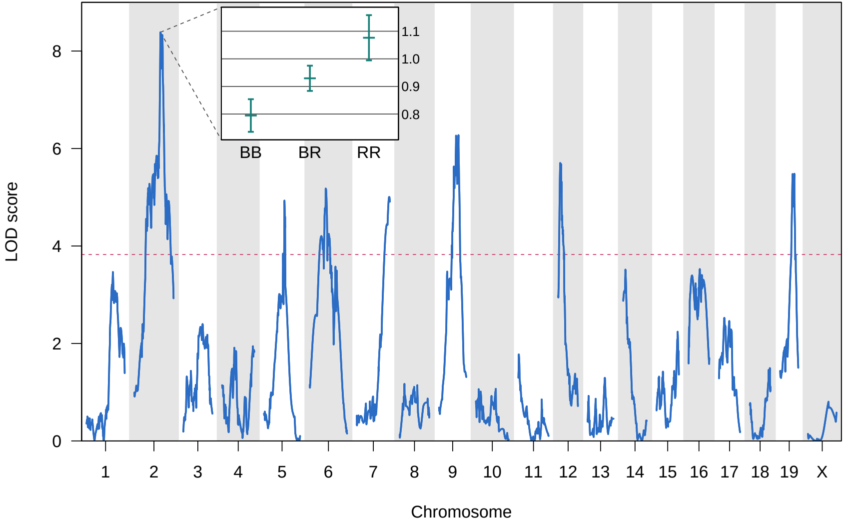


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
phenotype
500
400
300
200
100
Mice
100
200
300
400
500
600
Markers

QTL mapping


LOD score
0
2
4
6
8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Chromosome

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

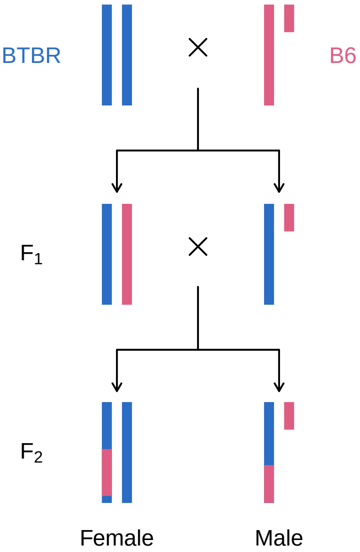
Attie project

$\sim$500 B6 $\times$ BTBR intercross mice, all ob/ob

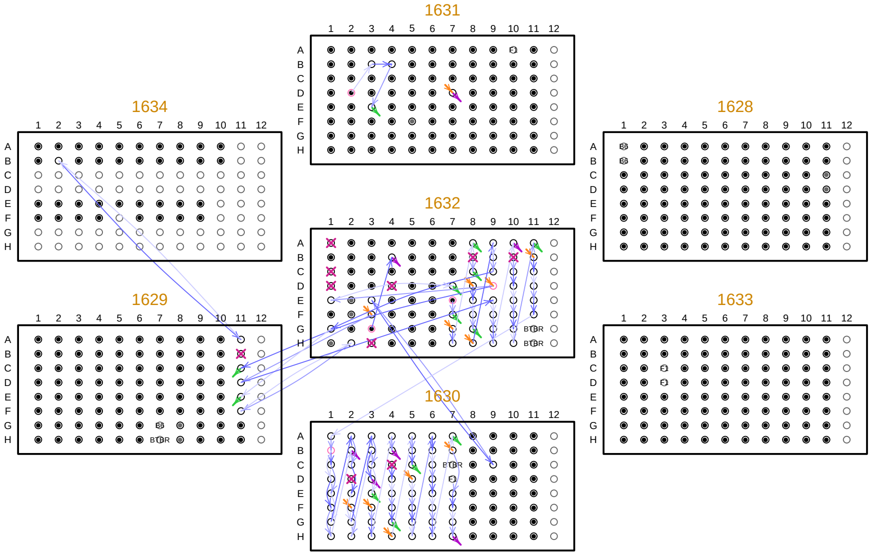
- Genotypes at 2057 SNPs (Affymetrix arrays)
- Gene expression in six tissues (Agilent arrays)
  - adipose
  - gastrocnemius muscle
  - hypothalamus
  - pancreatic islets
  - kidney
  - liver
- Numerous clinical phenotypes

  (e.g., body weight, insulin and glucose levels)

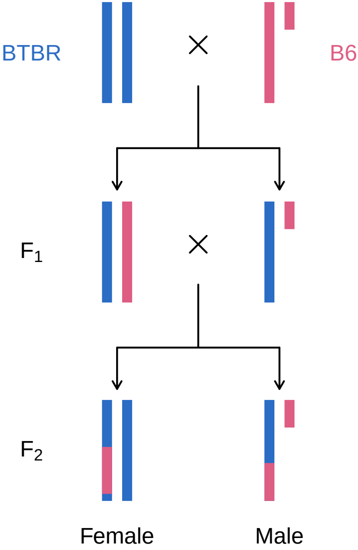
Sex and the X chr



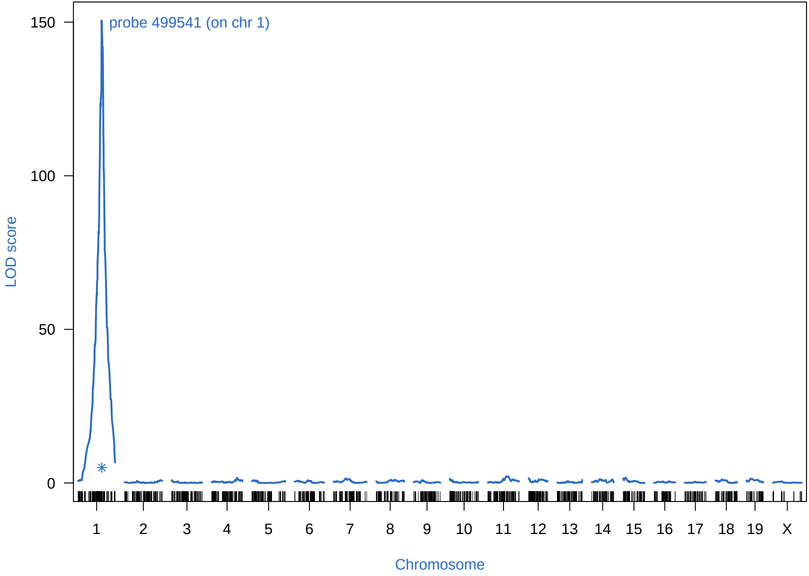
Genotype mix-ups



Sex and the X chr

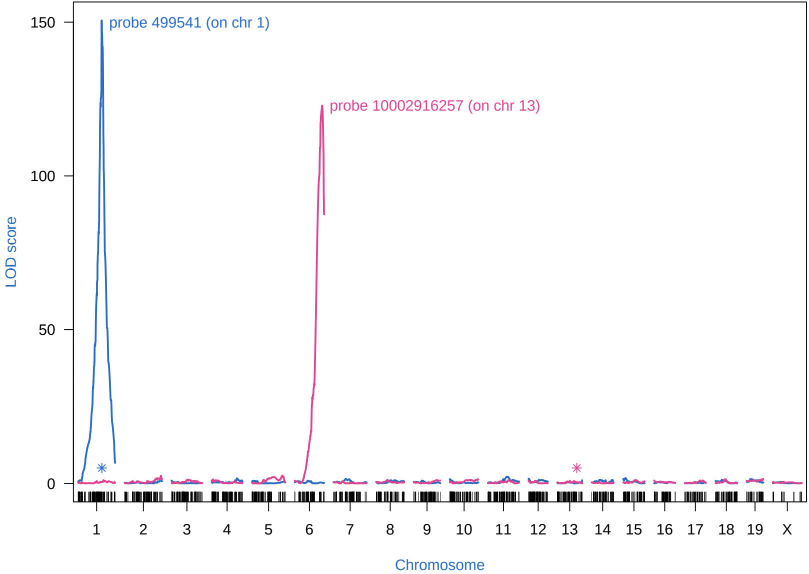


Strong eQTL



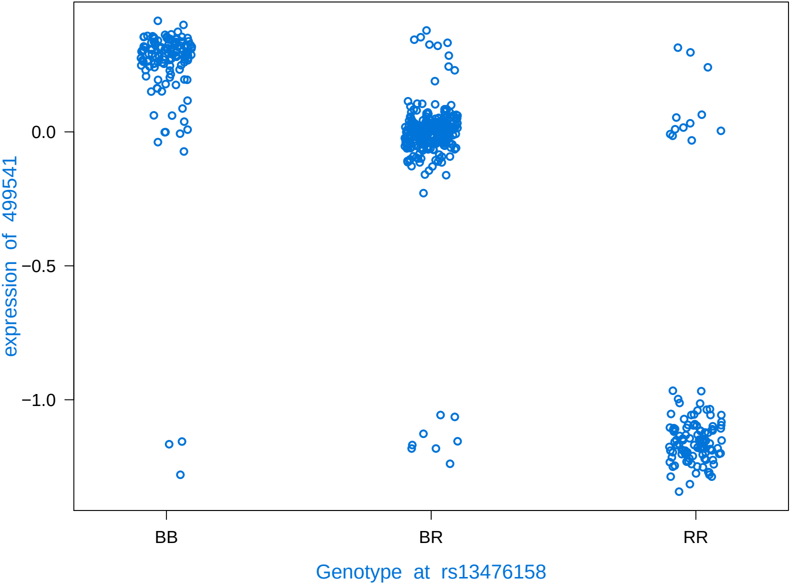
probe 499541 (on chr 1)
150
100
50
0
LOD score
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Chromosome

Strong eQTL



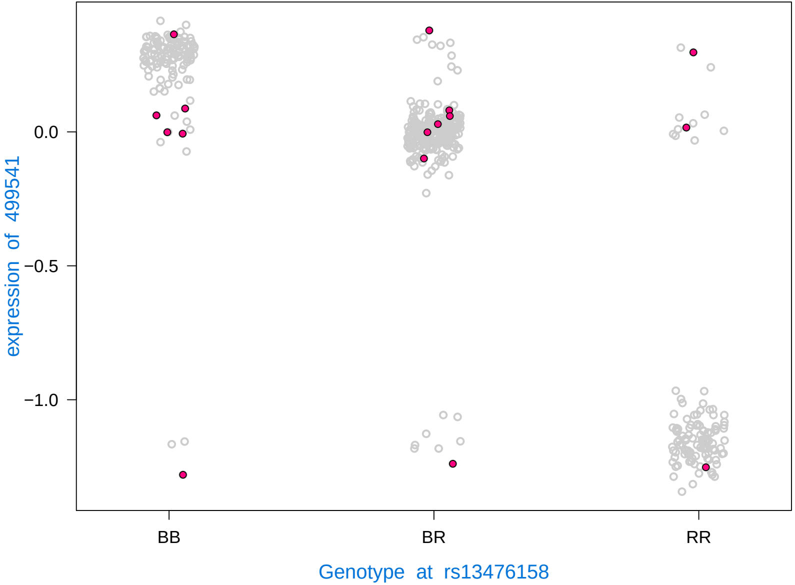
probe 499541 (on chr 1)
probe 10002916257 (on chr 13)
150
100
50
0
LOD score
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

E vs G



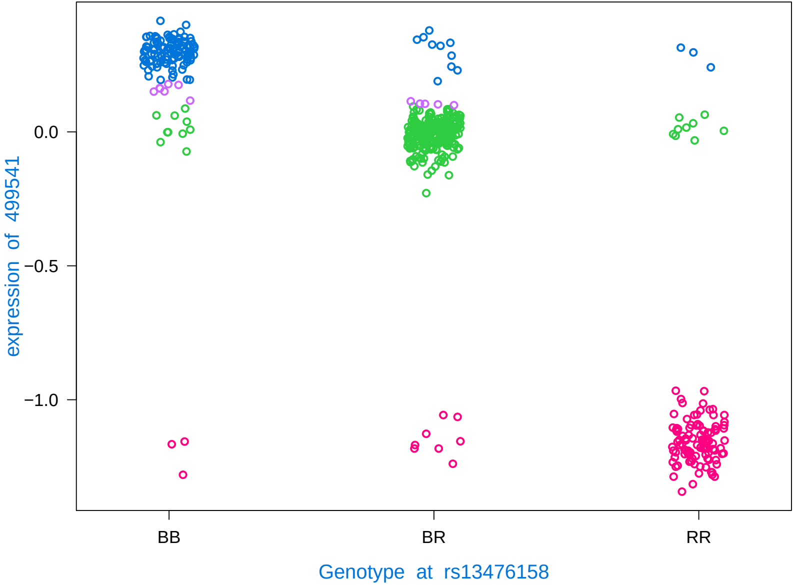
0.0
−0.5
−1.0
expression of 499541
BB
BR
RR
Genotype at rs13476158

E vs G

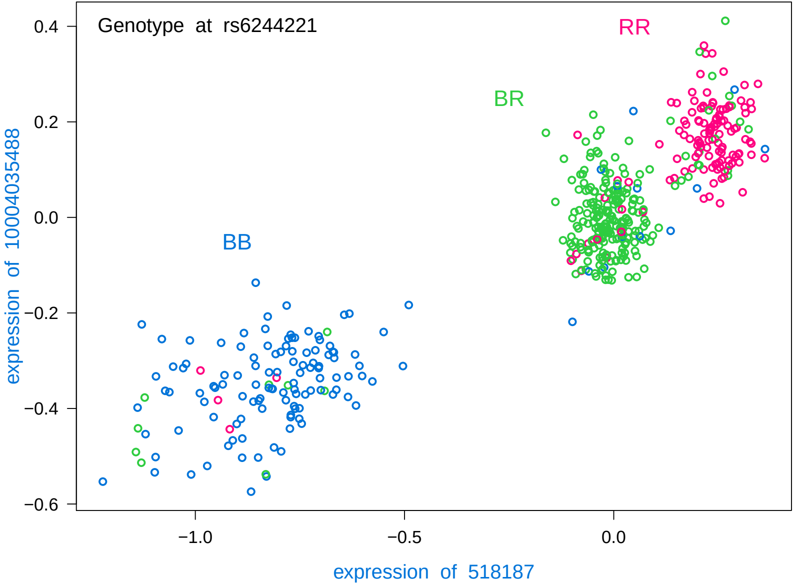


expression of 499541
0.0
−0.5
−1.0
BB
BR
RR
Genotype at rs13476158

kNN classifier

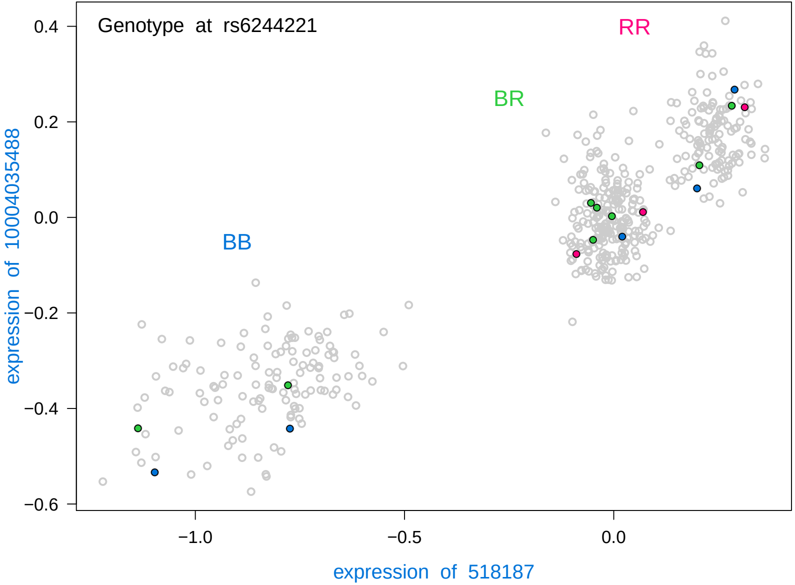


E vs G



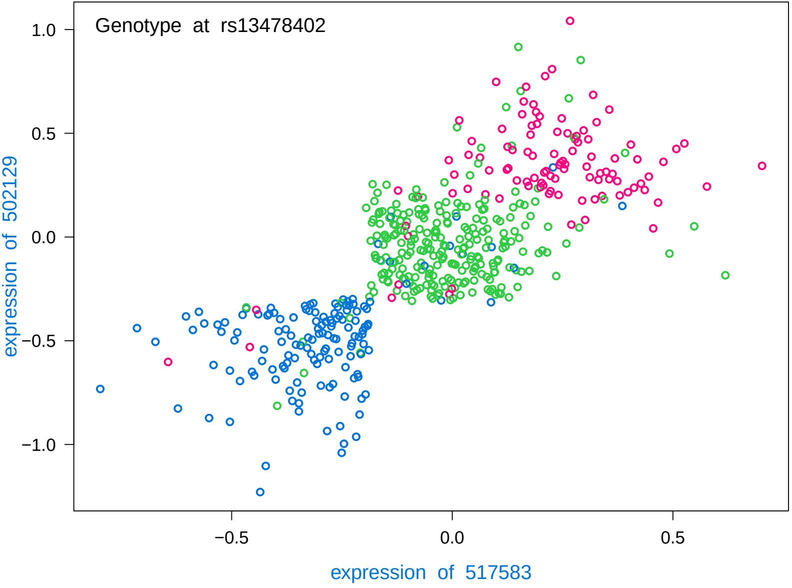
Genotype at rs6244221
RR
BR
BB
0.4
0.2
0.0
−0.2
−0.4
−0.6
expression of 10004035488
−1.0
−0.5
0.0
expression of 518187

E vs G

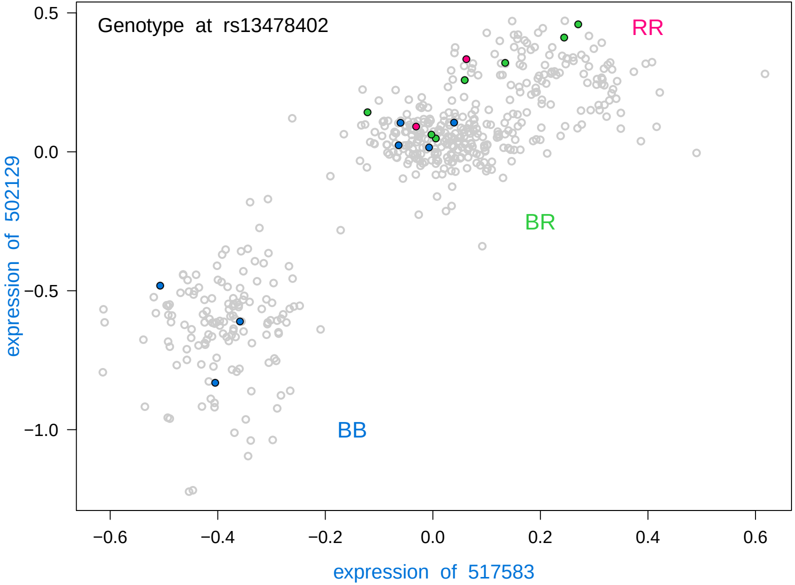


Genotype at rs6244221
RR
BR
BB
0.4
0.2
0.0
−0.2
−0.4
−0.6
−1.0
−0.5
0.0
expression of 10004035488
expression of 518187

E vs G

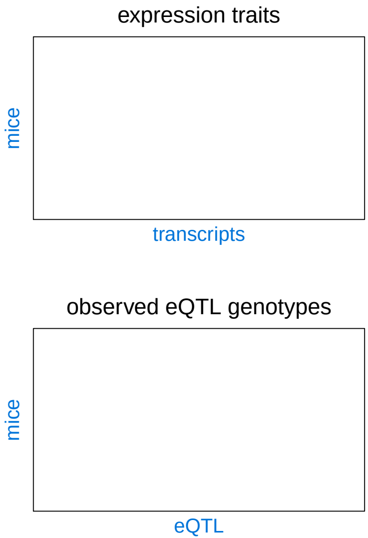

Genotype at rs13478402
expression of 502129
expression of 517583
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5

E vs G

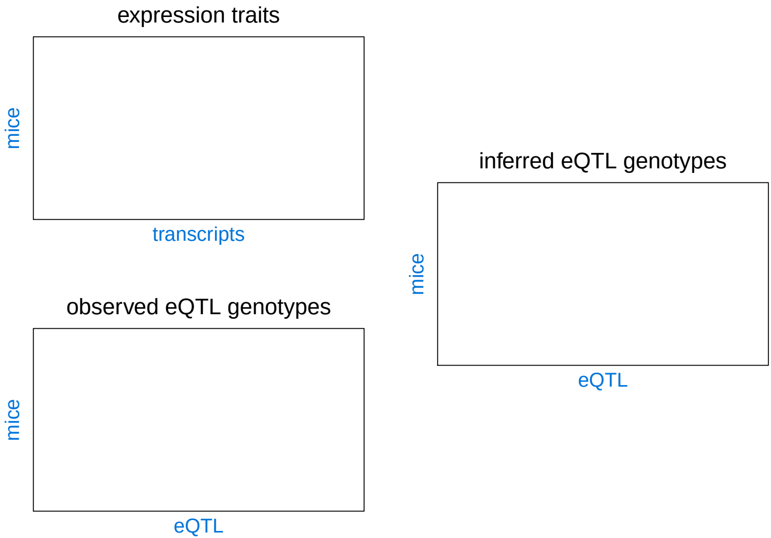


Genotype at rs13478402
RR
BR
BB
expression of 502129
expression of 517583
0.5
0.0
−0.5
−1.0
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

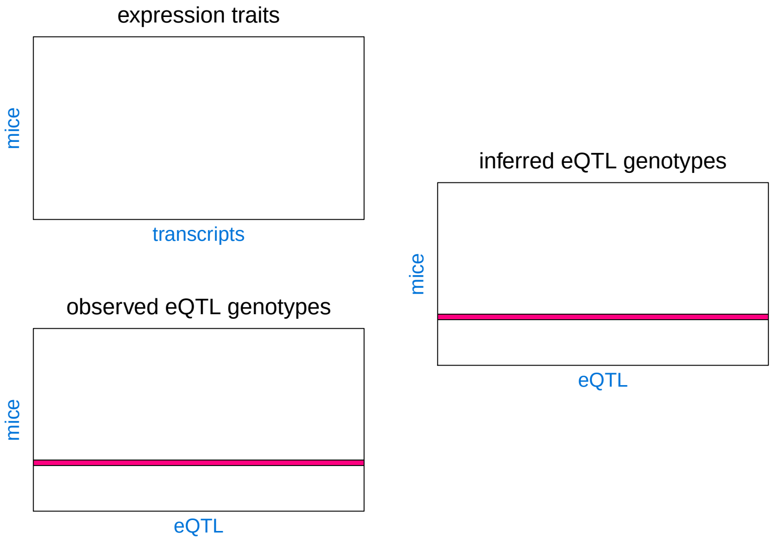
Basic scheme



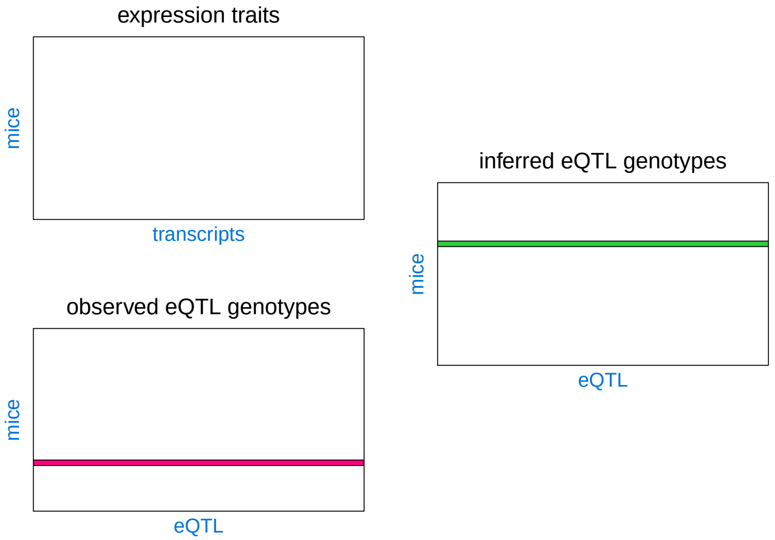
Basic scheme



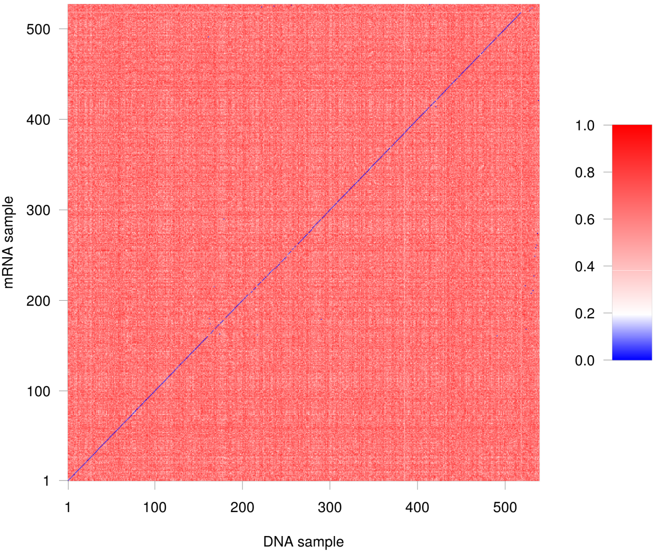
Basic scheme



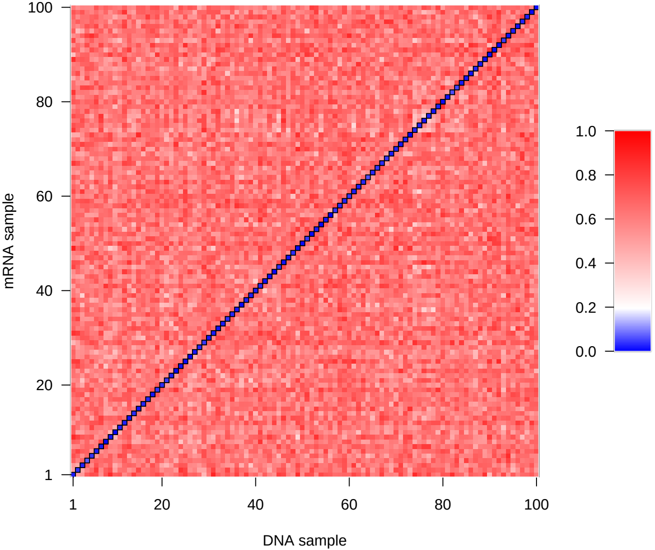
Basic scheme



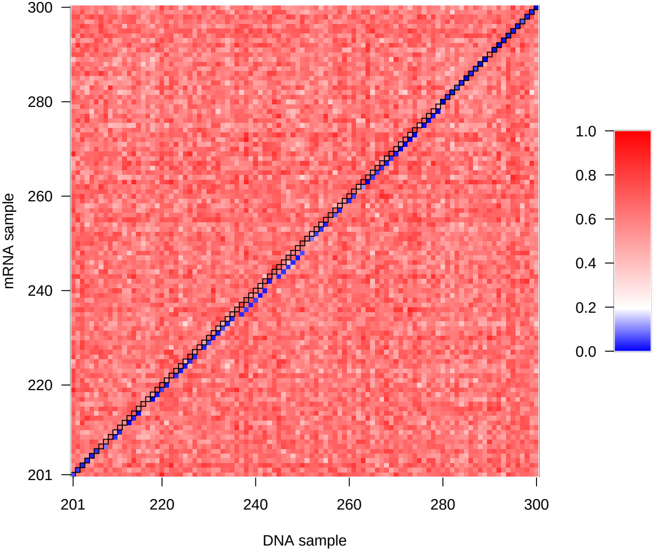
Prop'n mismatches



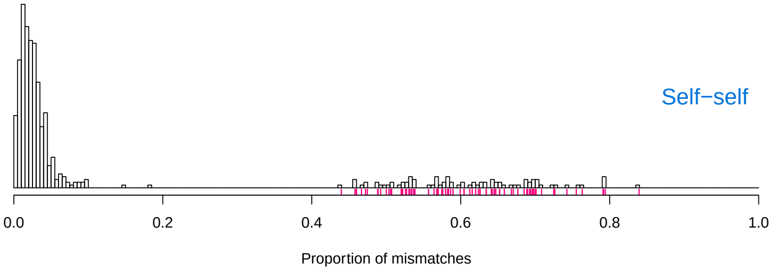
Prop'n mismatches

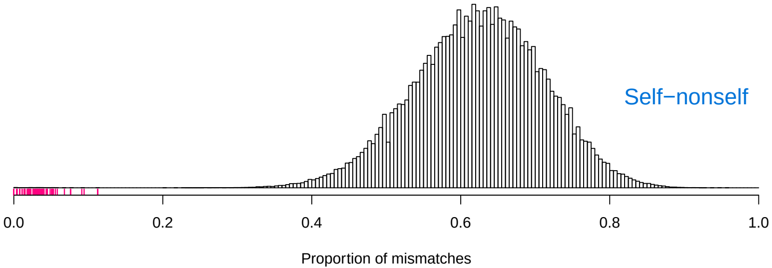


Prop'n mismatches

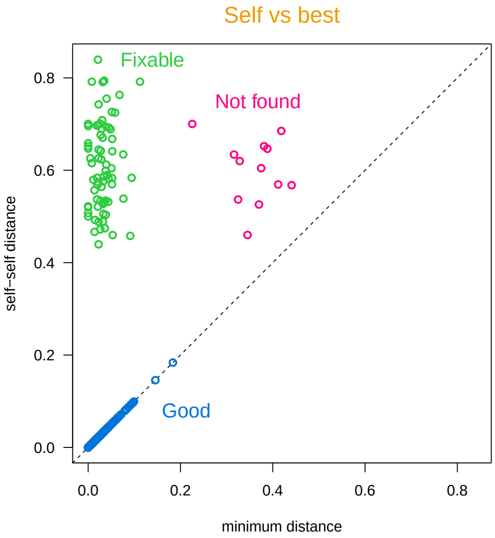


Prop’n mismatches





Decisions


Self vs best
Fixable
Not found
Good
0.8
0.6
0.4
0.2
0.0
self−self distance
0.0
0.2
0.4
0.6
0.8
minimum distance

Genotype mix-ups

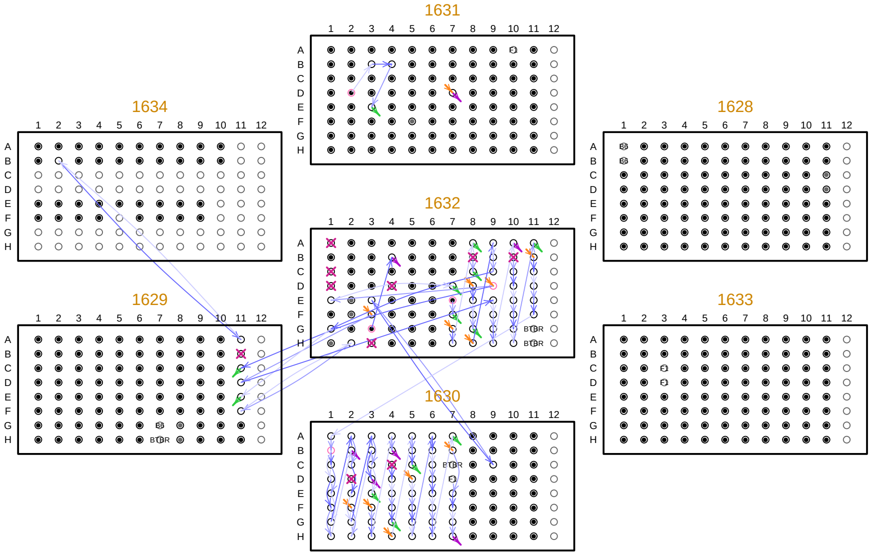
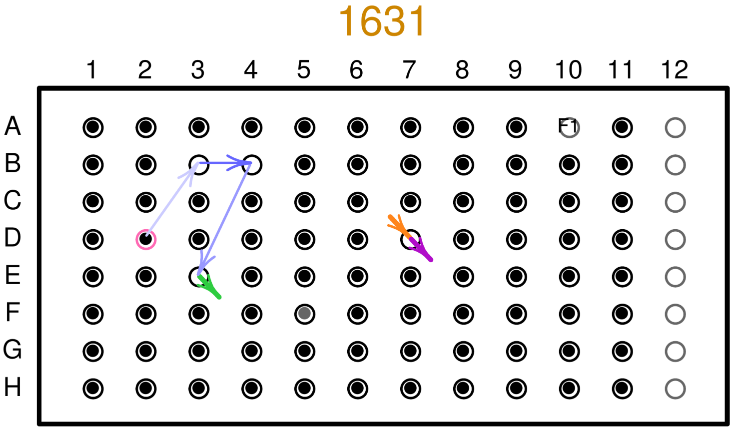


Plate 1631



Plates 1632 and 1630

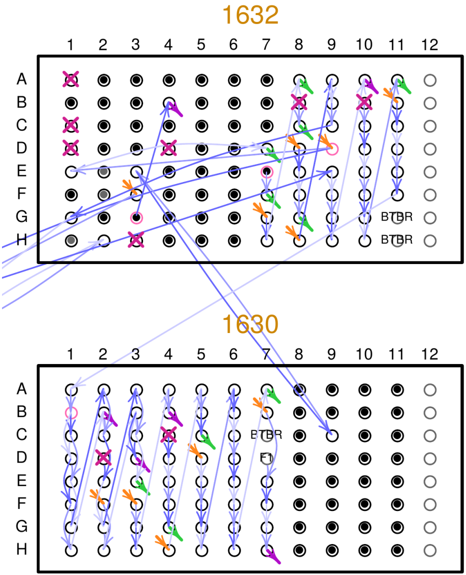
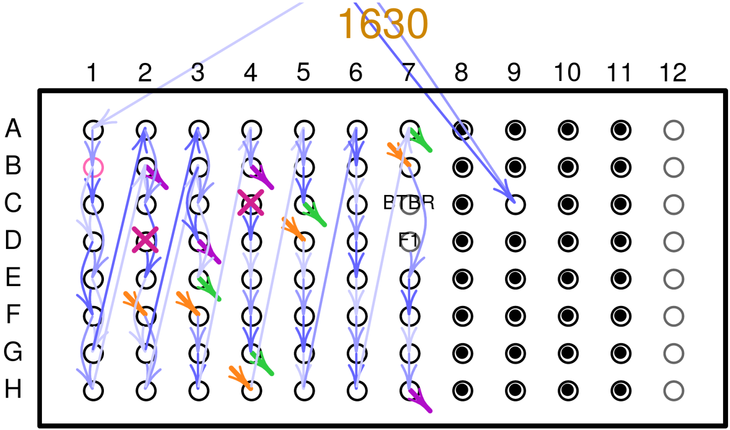


Plate 1630

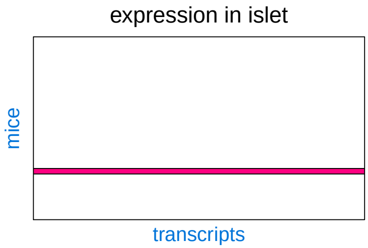


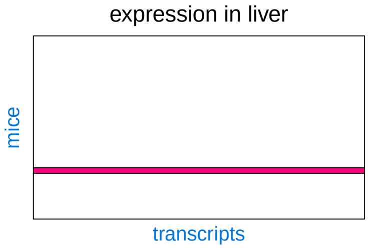
E vs E

expression in islet

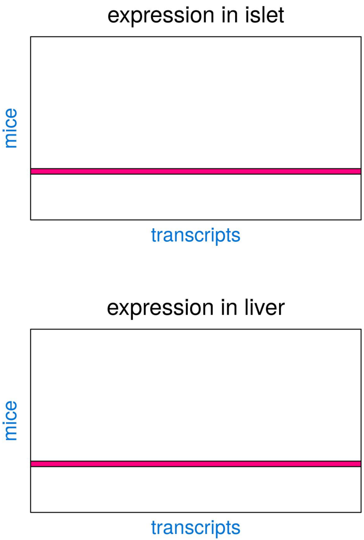
expression in liver

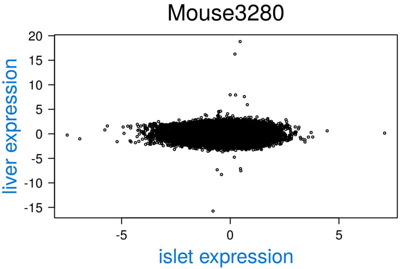
E vs E


expression in islet
mice
transcripts

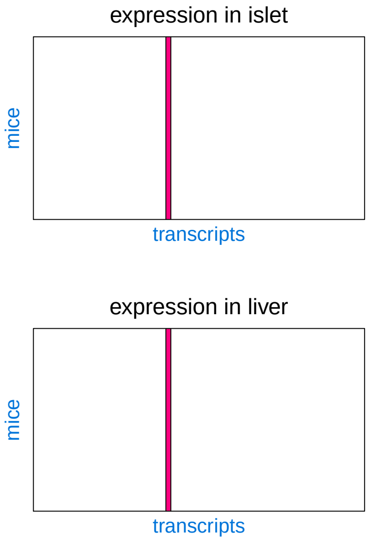

expression in liver
mice
transcripts

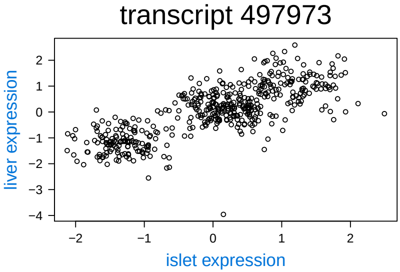
E vs E


expression in islet
mice
transcripts
expression in liver
mice
transcripts

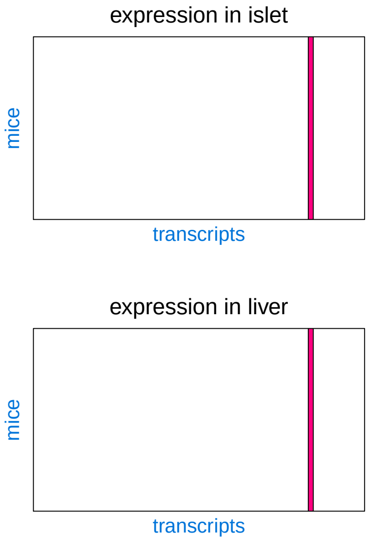

Mouse3280
liver expression
islet expression
20
15
10
5
0
-5
-10
-15
-5
0
5

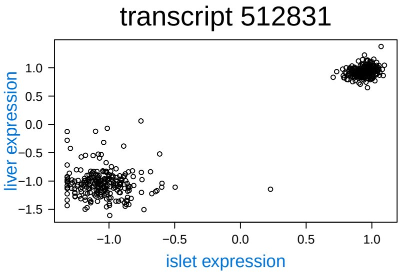
E vs E


expression in islet
mice
transcripts
expression in liver
mice
transcripts

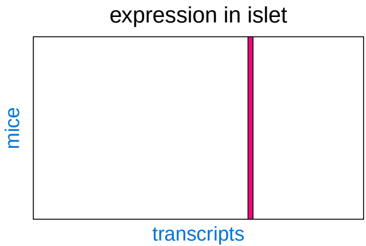

transcript 497973
liver expression
2
1
0
−1
−2
−3
−4
islet expression
−2
−1
0
1
2

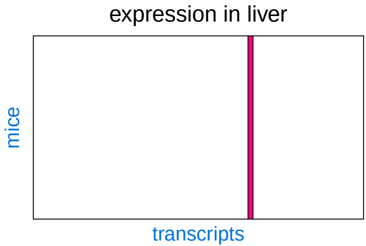
E vs E

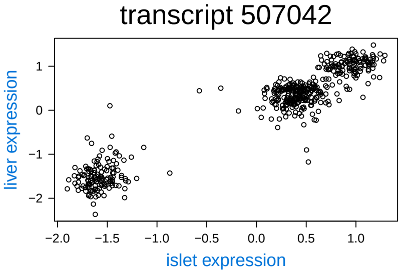

expression in islet
mice
transcripts
expression in liver
mice
transcripts


transcript 512831
liver expression
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
islet expression

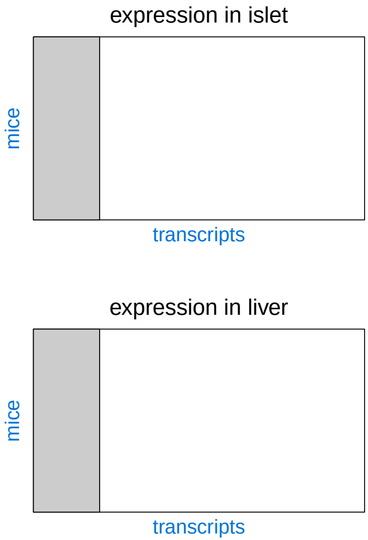
E vs E


expression in islet
mice
transcripts

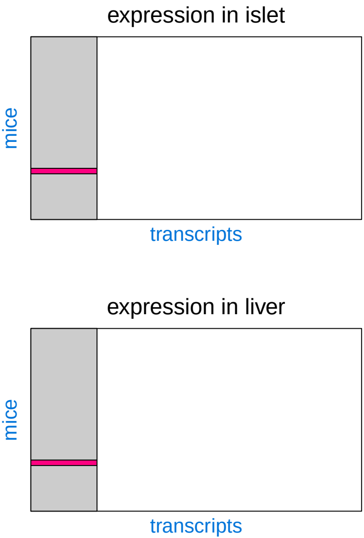

expression in liver
mice
transcripts

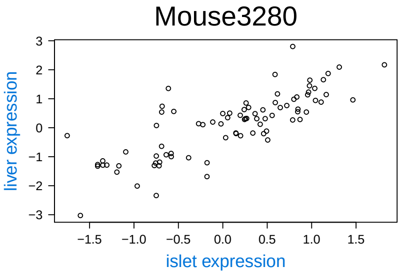

transcript 507042
liver expression
islet expression
1
0
−1
−2
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0

E vs E

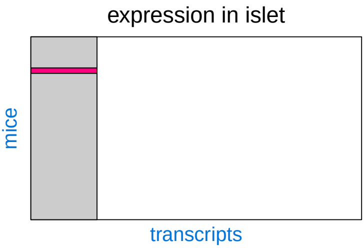

expression in islet
mice
transcripts
expression in liver
mice
transcripts

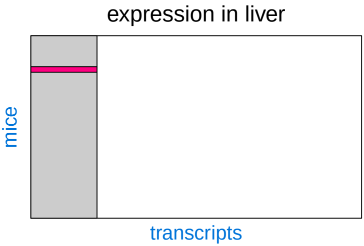
E vs E

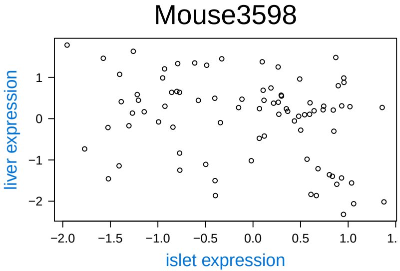

expression in islet
mice
transcripts
expression in liver
mice
transcripts


Mouse3280
liver expression
islet expression
3
2
1
0
−1
−2
−3
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

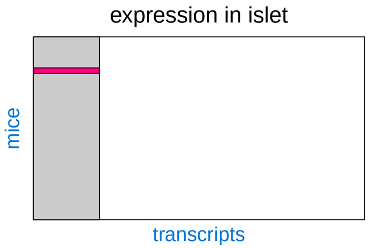
E vs E

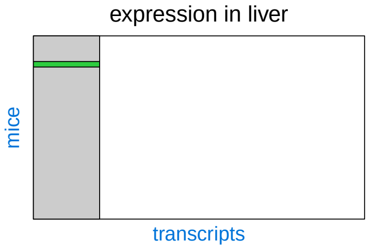

expression in islet
mice
transcripts

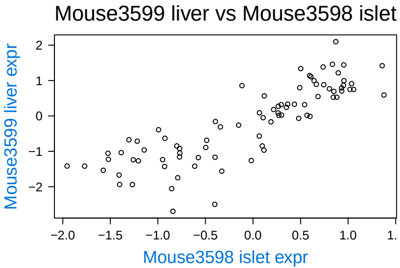

expression in liver
mice
transcripts


Mouse3598
liver expression
1
0
−1
−2
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
islet expression

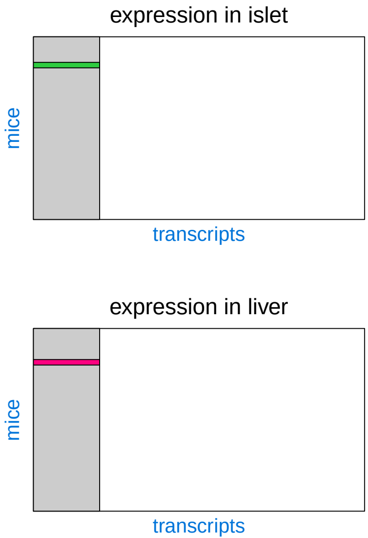
E vs E

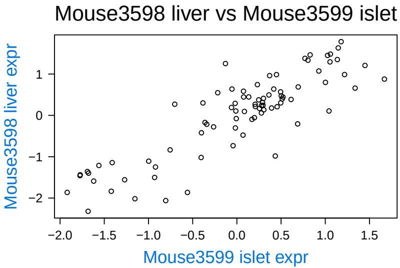

expression in islet
mice
transcripts


expression in liver
mice
transcripts

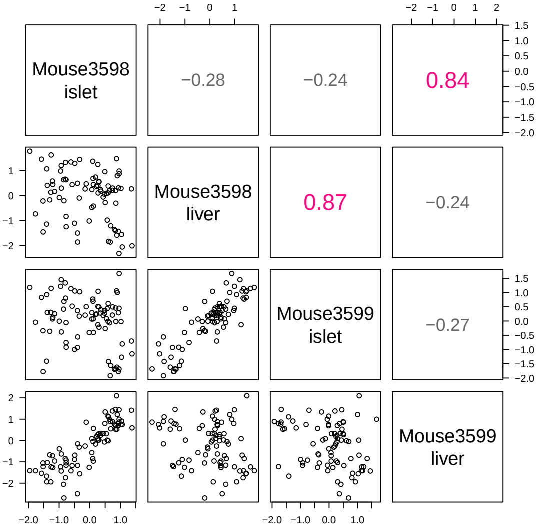

Mouse3599 liver vs Mouse3598 islet
Mouse3599 liver expr
Mouse3598 islet expr
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.
2
1
0
−1
−2

E vs E

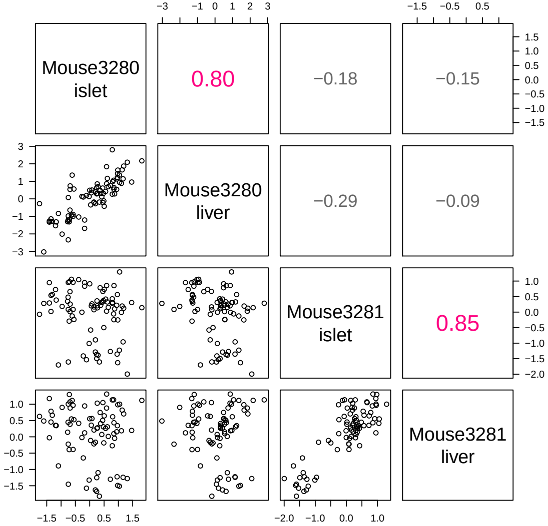

expression in islet
mice
transcripts
expression in liver
mice
transcripts


Mouse3598 liver vs Mouse3599 islet
Mouse3598 liver expr
Mouse3599 islet expr
1
0
−1
−2
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

E vs E

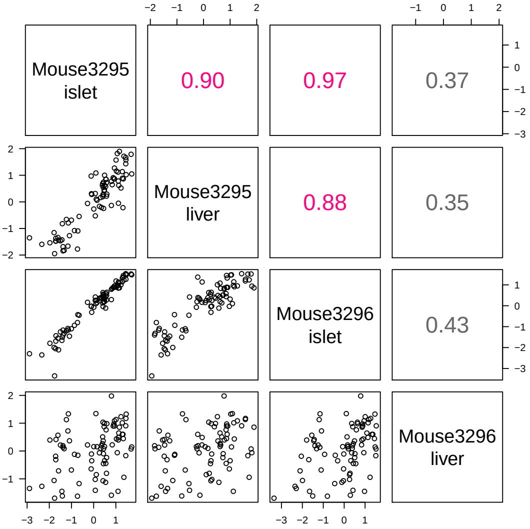

Mouse3598
islet
−0.28
−0.24
0.84
Mouse3598
liver
0.87
−0.24
Mouse3599
islet
−0.27
Mouse3599
liver

E vs E



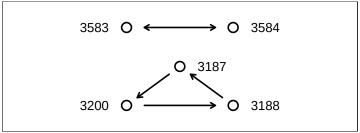
Mouse3280
islet
0.80
−0.18
−0.15
Mouse3280
liver
−0.29
−0.09
Mouse3281
islet
0.85
Mouse3281
liver

E vs E

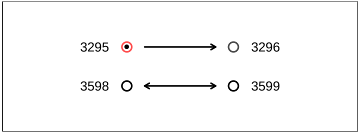

Mouse3295 islet
0.90
0.97
0.37
Mouse3295 liver
0.88
0.35
Mouse3296 islet
0.43
Mouse3296 liver

Expression mix-ups

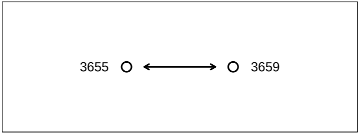
adipose



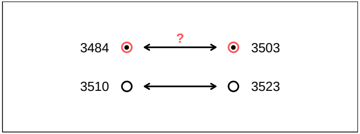
islet



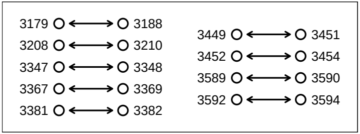
gastroc



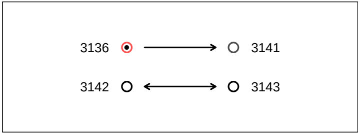
kidney



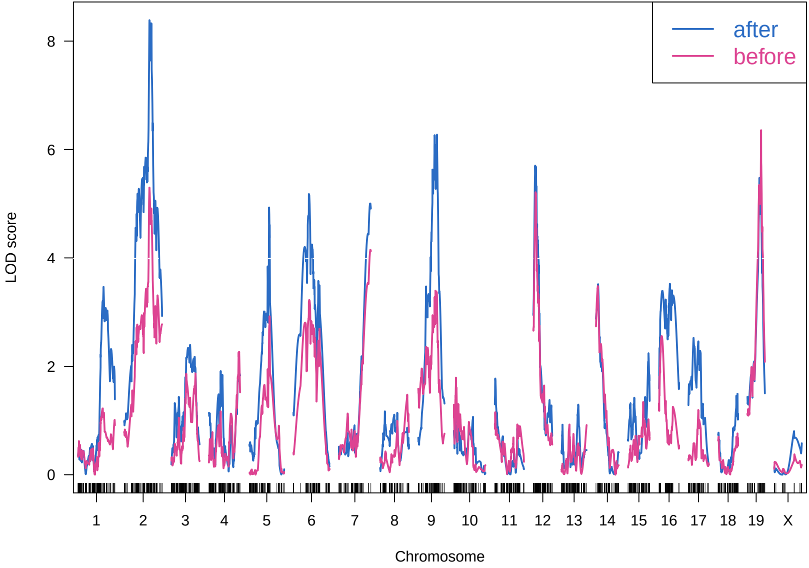
hypo



liver

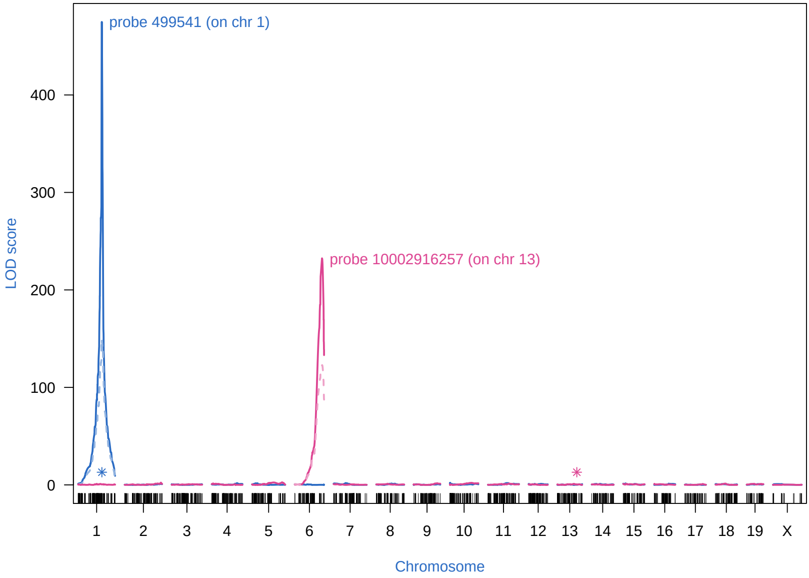


Insulin QTL



after
before
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

Strong eQTL


probe 499541 (on chr 1)
probe 10002916257 (on chr 13)
LOD score
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome

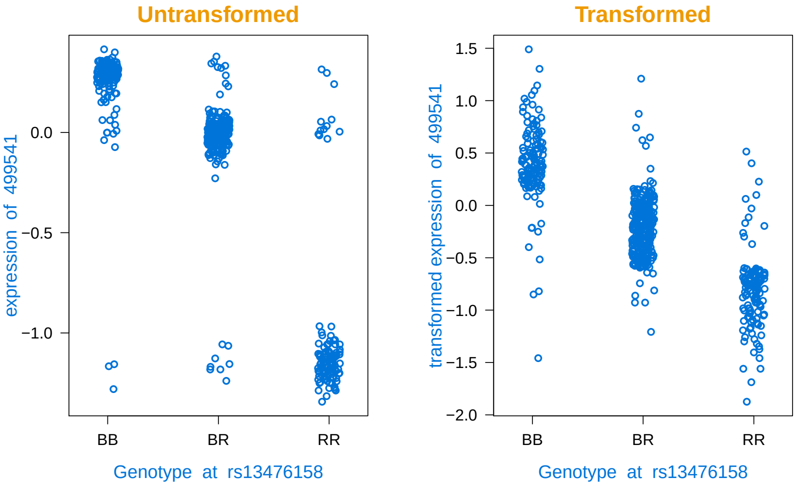
Summary

- Sample mix-ups happen
- With eQTL data, we can both identify and correct mix-ups
- There is great value in having expression on multiple tissues
- The general idea here has wide application for high-throughput data
- Broman et al. (2015) G3 5:2177-2186
  doi: 10.1534/g3.115.019778
- Related work:
  - Westra et al. (2011) Bioinformatics 27:2104–2111
  - Schadt et al. (2012) Nat Genet 44:603–608
  - Ekstrøm and Feenstra (2012) Stat Appl Genet Mol Biol 3:Article 13
  - Lynch et al. (2012) PLoS ONE 7:e41815

Lessons

- Don't fully trust anyone
  - Including yourself
- Make lots of plots
  - Don't rely on summary statistics, like LOD scores
  - Look at responses on the original scale
- Follow up all aberrations
- Take your time with data cleaning
  - A month, two months, a year?
- If you have big rectangles whose rows correspond, check that they *actually* correspond

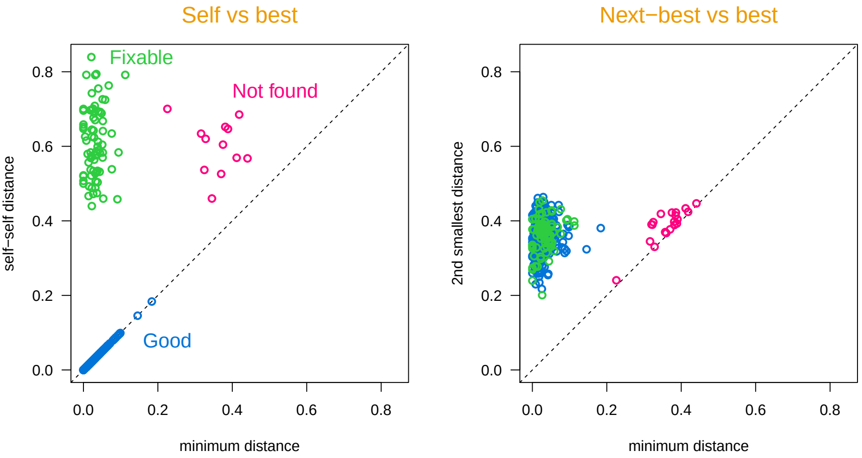
E vs G


Untransformed
Transformed
expression of 499541
transformed expression of 499541
0.0
−0.5
−1.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
BB
BR
RR
Genotype at rs13476158
Genotype at rs13476158

Lessons

- Don't fully trust anyone
  - Including yourself
- Make lots of plots
  - Don't rely on summary statistics, like LOD scores
  - Look at responses on the original scale
- Follow up all aberrations
- Take your time with data cleaning
  - A month, two months, a year?
- If you have big rectangles whose rows correspond, check that they *actually* correspond

Decisions



Self vs best
Fixable
Not found
Good
self−self distance
minimum distance
0.0
0.2
0.4
0.6
0.8
Next−best vs best
2nd smallest distance
minimum distance