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
QTL mapping

![Graph showing QTL mapping with LOD scores on the y-axis and chromosome numbers on the x-axis. Peaks indicate regions of high LOD scores, with a threshold line at LOD = 4.]

Chromosome

LOD score

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

LOD score

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

Chromosome
QTL mapping
Attie project

~500 B6 × 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

BTBR  ×  B6
   /   \
F1  ×  
   /   \
F2  

Female  Male
Genotype mix-ups
Sex and the X chr

![Genetic diagram showing the crosses between BTBR and B6 genotypes leading to F1 and F2 generations. The diagram illustrates the sex and genetic traits resulting from these crosses.]

- BTBR (blue) × B6 (red) → F1
- F1 (blue and red) × F1 (blue and red) → F2
- F2 generations are divided into Female and Male categories, showing the genetic characteristics of each sex.
Strong eQTL

probe 499541 (on chr 1)
Strong eQTL

LOD score

probe 499541 (on chr 1)

probe 10002916257 (on chr 13)
E vs G

Genotype at rs13476158

expression of 499541

BB BR RR
E vs G

Genotype at rs13476158

expression of 499541
kNN classifier

Genotype at rs13476158

expression of 499541
E vs G

Genotype at rs6244221

expression of 518187

expression of 10004035488

Genotype at rs6244221

BB

BR

RR
E vs G

Genotype at rs6244221

expression of 10004035488

expression of 518187
E vs G

Genotype at rs13478402

expression of 517583

expression of 502129
E vs G

Genotype at rs13478402

expression of 502129

expression of 517583
Basic scheme

- expression traits
- transcripts
- observed eQTL genotypes
- eQTL
- mice
Basic scheme

expression traits

transcripts

observed eQTL genotypes

mice

inferred eQTL genotypes

mice

eQTL
Basic scheme

expression traits

mice

transcripts

observed eQTL genotypes

mice

eQTL

inferred eQTL genotypes

mice

eQTL
Basic scheme

expression traits

mice

transcripts

observed eQTL genotypes

mice

eQTL

inferred eQTL genotypes

mice

eQTL
Prop’n mismatches
Prop’n mismatches

DNA sample

mRNA sample

0.0
0.2
0.4
0.6
0.8
1.0

0.0
0.2
0.4
0.6
0.8
1.0

16
Prop’n mismatches

Self–self

Self–nonself
Decisions

Self vs best

- Fixable
- Not found
- Good
Genotype mix-ups
Plates 1632 and 1630
Plate 1630
E vs E

expression in islet

expression in liver

mice

transcripts

mice

transcripts
E vs E

expression in islet

expression in liver

mice

transcripts
E vs E

expression in islet

mice

transcripts

expression in liver

mice

transcripts

Mouse3280

liver expression

islet expression
expression in islet transcripts mice

expression in liver transcripts mice

transcript 497973

islet expression

liver expression
E vs E

expression in islet

mice
transcripts

expression in liver

mice
transcripts

transcript 512831

islet expression

liver expression

−1.5 −1.0 −0.5 0.0 0.5 1.0

transcript 512831

−1.5 −1.0 −0.5 0.0 0.5 1.0

−1.5 −1.0 −0.5 0.0 0.5 1.0

27
E vs E

expression in islet

transcripts

mice

expression in liver

transcripts

mice

transcript 507042
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

islet expression

liver expression
E vs E

Expression in islet transcripts mice

Expression in liver transcripts mice

Mouse3598

-2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5

-2 -1 0 1

-2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.0

Islet expression

Liver expression
E vs E

expression in islet

expression in liver

Mouse3599 liver vs Mouse3598 islet

Mouse3598 islet expr

Mouse3599 liver expr
E vs E

expression in islet

expression in liver

Mouse3598 liver vs Mouse3599 islet
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

Mouse3280

Mouse3281

Mouse3281

islet

islet

liver

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

3583 → 3584
3200 → 3187 → 3188

islet

3295 → 3296
3598 ↔ 3599

gastroc

3655 ↔ 3659

kidney

3484 ↔ ? ↔ 3503
3510 ↔ 3523

hypo

3179 ↔ 3188
3208 ↔ 3210
3347 ↔ 3348
3367 ↔ 3369
3381 ↔ 3382
3449 ↔ 3451
3452 ↔ 3454
3589 ↔ 3590
3592 ↔ 3594

liver

3136 → 3141
3142 ↔ 3143
Insulin QTL
Strong eQTL

probe 499541 (on chr 1)

probe 10002916257 (on chr 13)
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
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

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

Next–best vs best

- Fixable
- Not found
- Good