Salvaging an eQTL project
Identifying and correcting sample mix-ups

Karl W Broman
Department of Biostatistics & Medical Informatics
University of Wisconsin – Madison

www.biostat.wisc.edu/~kbroman
Attie project

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

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)

Considerable brain power
- Alan Attie, Karl Broman, Aimee Teo Broman, Danielle Greenawalt, Mark Keller, Christina Kendziorski, Amit Kulkarni, Eric Schadt, Brian Yandell, ...
Sex and the X chr

F₂ females: R/R or B/R
F₂ males: hemizygous B or R
Genotype mix-ups
E vs G

expression of 499541

Genotype at rs13476158
The image shows a scatter plot with the expression of gene 499541 on the y-axis and the genotypes at rs13476158 on the x-axis. The genotypes are represented as BB, BR, and RR. The plot demonstrates the distribution of expression values across different genotypes.
Genotype at rs6244221

E vs G

expression of 10004035488

expression of 518187
Figure 1: Genotype at rs13478402 and expression of 517583 vs. 502129.

- BB
- BR
- RR

E vs G
The figure shows a scatter plot comparing the expression of two genes, 502129 and 517583, with respect to the genotype at rs13478402. The genotype is represented as either BB, BR, or RR, with each genotype cluster shown in a different color. The expression values are plotted along the x and y axes, with expression of 502129 on the y-axis and expression of 517583 on the x-axis. The plot visually represents the correlation between the two gene expressions and the genotype.
Basic scheme

expression traits

mice

transcripts

mice

observed eQTL genotypes

mice

eQTL
Basic scheme

expression traits

mice

transcripts

observed eQTL genotypes

mice

eQTL

mice

inferred eQTL genotypes

eQTL
Basic scheme

Expression traits

Inferred eQTL genotypes

Observed eQTL genotypes

eQTL

mice

transcripts

mice
Basic scheme

expression traits

mice

transcripts

observed eQTL genotypes

mice

eQTL

inferred eQTL genotypes

mice

eQTL
Prop’n mismatches

Proportion of mismatches

Self–self

Self–nonself
Decisions

Self vs best

Next–best vs best

Fixable
Not found
Good
Genotype mix-ups
Plate 1631
Plates 1632 and 1630
E vs E

expression in islet

mice

transcripts

expression in liver

mice

transcripts
Expression in islet transcripts in mice

Expression in liver transcripts in mice
E vs E

expression in islet

mice

expression in liver

mice

transcripts

transcripts

liver expression

islet expression

Mouse3280
E vs E

expression in islet
transcripts
mice
expression in liver
transcripts
mice

transcript 512831
E vs E

expression in islet transcripts mice
expression in liver transcripts mice

transcript 507042

E vs E

expression in islet
transcripts
mice

expression in liver
transcripts
mice

transcript 507042


Expression in islet transcripts mice

Expression in liver transcripts mice
E vs E

expression in islet

transcripts

mice

expression in liver

transcripts

mice

Mouse3280

islet expression

liver 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

expression in liver

mice

transcripts

transcripts

Mouse3598

islet expression

liver expression
E vs E

expression in islet

transcripts
mice

expression in liver

transcripts
mice

Mouse3599 liver vs Mouse3598 islet

Mouse3598 islet expr

Mouse3599 liver expr
E vs E

expression in islet
mice
transcripts

expression in liver
mice
transcripts

Mouse3598 liver vs Mouse3599 islet

Mouse3598 liver expr

Mouse3599 islet expr
E vs E

Mouse3295 islet

Mouse3295 liver

Mouse3296 islet

Mouse3296 liver
Insulin QTL
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

• Very similar to MixupMapper (Westra et al., Bioinformatics, 2011)
  - Multiple tissues
  - Direct tissue-tissue comparisons
  - Predict genotype rather than expression phenotype
Acknowledgments

Alan Attie  Biochemistry, UW–Madison
Mark Keller

Brian Yandell  Statistics and Horticulture, UW–Madison

Christina Kendziorski  Biostatistics & Medical Informatics, UW–Madison
Aimee Teo Broman

Eric Schadt  Pacific Biosciences of California

Danielle Greenawalt  Merck & Co., Inc.
Amit Kulkarni

NIH: R01 GM074244, R01 DK066369