Identifying and correcting sample mix-ups in high-dimensional data

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

www.biostat.wisc.edu/~kbroman
Intercross

P₁ × P₂

F₁ × F₁

F₂
Alan 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

- **F₂ females**: R/R or B/R
- **F₂ males**: hemizygous B or R
Genotype mix-ups
Sex and the X chr

F₂ females: R/R or B/R
F₂ males: hemizygous B or R
Strong eQTL

probe 499541 (on chr 1)
Strong eQTL

probe 499541 (on chr 1)
probe 10002916257 (on chr 13)
E vs G

expression of 499541

Genotype at rs13476158
E vs G

expression of 499541

Genotype at rs13476158

BB BR RR
kNN classifier

Genotype at rs13476158

expression of 499541

BB
BR
RR
Basic scheme

expression traits

mice

transcripts

observed eQTL genotypes

mice

eQTL
Basic scheme

- Expression traits
- Transcripts
- Observed eQTL genotypes
- Mice
- Inferred eQTL genotypes
- Mice
- eQTL
Basic scheme

expression traits

transcripts

mice

observed eQTL genotypes

inferred eQTL genotypes

eQTL

mice

eQTL
mRNA sample

DNA sample

Prop’n mismatches
Prop’n mismatches

DNA sample vs. mRNA sample
Prop’n mismatches

DNA sample

mRNA sample

201
201
0.0
0.2
0.4
0.6
0.8
1.0
Prop’n mismatches

Self–self

Proportion of mismatches

Self–nonself

Proportion of mismatches
Decisions

Self vs best

- Fixable
- Not found
- Good

Minimum distance
Self−self distance
Genotype mix-ups
Plates 1632 and 1630
Plate 1630
E vs E

expression in islet transcripts mice

expression in liver transcripts mice
E vs E

expression in islet

expression in liver

mice

transcripts

mice

transcripts
E vs E

expression in islet

expression in liver

mice

transcripts

transcript 497973

liver expression

islet expression
E vs E

expression in islet

expression in liver

mice

transcripts
E vs E

expression in islet
expression in liver

Mouse3280

islet expression
liver expression

transcripts
mice
transcripts
mice
expression in islet expression in liver

Mouse3598

islet expression
liver expression
E vs E

expression in islet

transcripts

mice

expression in liver

transcripts

mice

Mouse3599 liver vs Mouse3598 islet

Mouse3599 liver expr

Mouse3598 islet expr
E vs E

expression in islet transcripts mice

expression in liver transcripts mice

Mouse3598 liver vs Mouse3599 islet

Mouse3598 liver expr

Mouse3599 islet expr
E vs E

Mouse3295

islet

−2 −1 0 1 2

0.90 0.97 0.37

Mouse3295

liver

0.88 0.35

Mouse3296

islet

0.43

Mouse3296

liver
• 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 27:2104–2111, 2011)
  – Multiple tissues
  – Direct tissue-tissue comparisons
  – Predict genotype rather than expression phenotype
Acknowledgments

Alan Attie  
Mark Keller  

Biochemistry, UW–Madison

Brian Yandell  

Statistics and Horticulture, UW–Madison

Christina Kendziorski  
Aimee Teo Broman  

Biostatistics & Medical Informatics, UW–Madison

Eric Schadt  

Pacific Biosciences of California

Danielle Greenawalt  
Amit Kulkarni  

Merck & Co., Inc.

Śaunak Sen  

University of California, San Francisco

NIH: R01 GM074244, R01 DK066369