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

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

www.biostat.wisc.edu/~kbroman
Human vs mouse
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)
Sex and the X chr

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

Genotype at rs13476158

expression of 499541

−1.0
−0.5
0.0
Genotype at rs13476158

E vs G

Expression of 499541
kNN classifier

Genotype at rs13476158

expression of 499541
E vs G

Genotype at rs6244221

expression of 518187

expression of 10004035488

BB

BR

RR
Genotype at rs6244221

Expression of 10004035488

Expression of 518187
Basic scheme

expression traits

mice

transcripts

observed eQTL genotypes

mice

eQTL
Basic scheme

expression traits

mice

transcripts

observed eQTL genotypes

mice

inferred eQTL genotypes

mice

eQTL
Basic scheme

expression traits

mice

transcripts

observed eQTL genotypes

mice

eQTL

mice

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
Prop’n mismatches

Self−self

Self−nonself

Proportion of mismatches

0.0 0.2 0.4 0.6 0.8 1.0

Proportion of mismatches

0.0 0.2 0.4 0.6 0.8 1.0

24
Genotype mix-ups
Plate 1631

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1631
Plates 1632 and 1630
Plate 1630
E vs E
expression in islet transcripts mice
expression in liver transcripts mice
expression in islet
mice
transcripts
E vs E

expression in islet
mice
transcripts

expression in liver
mice
transcripts
expression in islet transcripts mice

expression in liver transcripts mice

transcript 497973
E vs E

expression in islet

expression in liver

transcript 512831

islet expression

liver expression

mice

transcripts

mice

transcripts
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

Mouse3598

expression in islet

expression in liver

mice

transcripts

transcripts

islet expression

liver expression
E vs E

expression in islet

expression in liver

Mouse3599 liver vs Mouse3598 islet

Mouse3599 liver expr

Mouse3598 islet expr
expression in islet transcripts mice

expression in liver transcripts mice

Mouse3598 liver vs Mouse3599 islet

E vs E
Insulin QTL

Chromosome

LOD score

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

after

before
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

- 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?

- Have a system for keeping track of everything
  - Files, versions of files, analyses, …
  - Like a lab notebook
Genotype at rs13476158

expression of 499541
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?

• Have a system for keeping track of everything
  – Files, versions of files, analyses, …
  – Like a lab notebook
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

Šaunak Sen  UCSF

NIH: R01 GM074244, R01 DK066369
Decisions

Self vs best

Next–best vs best

Self–self distance

Fixable

Not found

Good

2nd smallest distance

minimum distance

minimum distance