Identifying sample mix-ups in eQTL data

Karl Broman

Biostatistics & Medical Informatics, Univ. Wisconsin–Madison

kbroman.org
github.com/kbroman
@kwbroman

Slides: kbroman.org/Talk_OSGA2021
Associations in systems genetics

- Genotypes
- Clinical phenotypes
- mRNA levels
Sample mix-ups

Broman et al. (2015) doi:10.1534/g3.115.019778
More sample mix-ups

**adipose**

- 3583 ↔ 3584
- 3200 ↔ 3188
- 3187

**gastroc**

- 3655 ↔ 3659

**hypo**

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

**islet**

- 3295 ↔ 3296
- 3598 ↔ 3599

**kidney**

- 3484 ↔ 3503
- ?
- 3510 ↔ 3523

**liver**

- 3136 ↔ 3141
- 3142 ↔ 3143

Broman et al. (2015) doi:10.1534/g3.115.019778
Table 2. Cis-eQTL mapping and sample mix-up identification results

<table>
<thead>
<tr>
<th>Stud</th>
<th>Population</th>
<th>Sample-size</th>
<th>Initial cis-eQTLs</th>
<th>Mix-ups detected(^a) n (%)</th>
<th>Sample-size after correction n (%)</th>
<th>cis-eQTLs after correction n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choy et al. (2008)</td>
<td>CHB+JP</td>
<td>87</td>
<td>138</td>
<td>20 (23)</td>
<td>79 (90)</td>
<td>418 (+203)</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>84</td>
<td>558</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>YR</td>
<td>85</td>
<td>274</td>
<td>2 (2)</td>
<td>83 (97)</td>
<td>287 (+5)</td>
</tr>
<tr>
<td>Stranger et al. (2007)</td>
<td>CHB+JP</td>
<td>90</td>
<td>1511</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>CE</td>
<td>90</td>
<td>903</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>YR</td>
<td>90</td>
<td>663</td>
<td>1 (1)</td>
<td>89 (99)</td>
<td>667 (+1)</td>
</tr>
<tr>
<td>Zhang et al. (2009)</td>
<td>CE</td>
<td>87</td>
<td>2581</td>
<td>NA</td>
<td>89 (100)</td>
<td>1635 (+12)</td>
</tr>
<tr>
<td></td>
<td>YR</td>
<td>89</td>
<td>1454</td>
<td>2 (2)</td>
<td>89 (100)</td>
<td>1635 (+12)</td>
</tr>
<tr>
<td>Webster et al. (2009)</td>
<td>Brai</td>
<td>36</td>
<td>1284</td>
<td>16 (4)</td>
<td>356 (98)</td>
<td>1367 (+6)</td>
</tr>
<tr>
<td>Heinzen et al. (2008)</td>
<td>Brai</td>
<td>93</td>
<td>349</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>PBMC</td>
<td>80</td>
<td>297</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Sample duplicates

Sex verification

Sample mix-ups:
- mRNA ↔ protein
- mRNA ↔ DNA
- protein ↔ DNA
But first

Missing Data
Percent missing genotypes

Mouse

Percent missing genotypes

0 100 200 300 400 500

0

10

20

30

40

50

60

70

DO274

DO306

DO308

DO309

DO340

DO357

DO397
Heterogeneous Stock/Diversity Outbreds

G0  A  B  C  D  E  F  G  H

G1

G2

G10

G15

G20
Sample duplicates
Correlation between mRNA samples

Correlation between samples

mRNAs

mRNAs sex-adjusted

Correlation between samples
Correlation between protein samples

Proteins

Proteins sex-adjusted
Sex verification
X and Y genotype dosage

![Graph showing the average X and Y chromosome intensity for males and females. The graph plots the average X chromosome intensity on the x-axis and the average Y chromosome intensity on the y-axis. The graph distinguishes between males (purple) and females (green).]
Sex and gene expression

- **Xist mRNA level**
- **Cyp4a12a mRNA level**
- **Sult2a5 mRNA level**
- **Cyp3a41a mRNA level**
- **Cyp3a16 mRNA level**
- **Cux2 mRNA level**
Sex and gene expression

[Graph showing the distance from female and male averages with data points for both females and males.]
Sex and proteins

- **Slc22a27**
- **Papss2**
- **Abcg2**
- **Mup20**
- **Sult2a3**
- **Fmo4**

![Graphs showing protein levels for different proteins in female and male individuals.](image-url)
Sex and proteins

![Graph showing the relationship between distance from female average and distance from male average. The graph displays a scatter plot with data points for both female and male categories.](image)

- Female
- Male
Sample mix-ups

mRNA ↔ protein
mRNA ↔ protein method

gene expression data

proteomics data

mRNA samples

genes

protein samples

proteins
mRNA ↔ protein method

gene expression data

mRNA samples
genes

proteomics data

protein samples
proteins

gene expression
protein level
mRNA ↔ protein method

gene expression data

proteomics data

mRNA samples

genome

protein samples

protein level

mRNA samples

genome

protein samples

protein level
mRNA ↔ protein method

gene expression data

mRNA samples
genes

proteomics data

protein samples
proteins

gene expression
protein level

mRNA samples
genes
proteomics data
protein samples
proteins

similarity matrix

protein samples
mRNA samples
mRNA ↔ protein correlations

max cor(mRNA, protein)
adjusting for sex
mRNA ↔ protein similarity matrix
mRNA ↔ protein similarities
mRNA ↔ protein: closest vs self
mRNA ↔ protein: selected samples
Sample mix-ups

DNA ↔ mRNA
DNA ↔ mRNA method

eQTL genotypes

DNA samples

eQTL

mRNA samples

genes with strong eQTL

expression traits
DNA ↔ mRNA method

eQTL genotypes

DNA samples

mRNA samples

genes with strong eQTL

eQTL genotype
AA AB BB

gene expression

expression traits

eQTL genotypes

DNA samples

mRNA samples

genes with strong eQTL

eQTL genotype
AA AB BB

gene expression
DNA ↔ mRNA method

eQTL genotypes

expression traits

predicted expression

expression traits

DNA samples

mRNA samples

genes with strong eQTL

gene expression

eQTL genotype

DNA samples

mRNA samples

genes

predicted expression

expression traits

DNA samples

mRNA samples

genes
DNA ↔ mRNA method

eQTL genotypes

DNA samples

mRNA samples

genes with strong eQTL

gene expression

AA
AB
BB

eQTL genotype

predicted expression

DNA samples

mRNA samples

genes

distance matrix

DNA samples

genes

expression traits

mRNA samples

genes

distance matrix

DNA samples
DNA ↔ mRNA LOD scores

max LOD
adjusting for sex
DNA ↔ mRNA distance matrix
DNA ↔ mRNA distances

Self–self

Self–nonself

self–self Distance

self–nonself Distance

Frequency
DNA ↔ mRNA: closest vs self
DNA ↔ mRNA: selected samples

mRNA sample M348

mRNA sample M349

mRNA sample F352

mRNA sample M410

mRNA sample F371

mRNA sample M386
Sample mix-ups

DNA ↔ protein
DNA ↔ protein method

- **pQTL genotypes**
  - DNA samples
  - pQTL

- **proteomics traits**
  - protein samples
  - proteins with strong pQTL

- **predicted protein levels**
  - DNA samples
  - proteins

- **proteomics traits**
  - protein samples
  - proteins

- **distance matrix**
  - DNA samples
  - protein samples

The diagram illustrates the relationship between DNA genotypes and protein levels, highlighting proteins with strong pQTL effects.
DNA ↔ protein correlations

max LOD
adjusting for sex

Maximum LOD score, by prot
DNA ↔ protein distance matrix
DNA ↔ protein distances

Self–self

Self–nonself

Frequency

self–self Distance

self–nonself Distance

Frequency
DNA ↔ protein: closest vs self
DNA ↔ protein: selected samples

- **protein sample M348**
  - DNA sample
  - Distance
  - M410

- **protein sample M349**
  - DNA sample
  - Distance
  - F371

- **protein sample F352**
  - DNA sample
  - Distance
  - F352

- **protein sample M410**
  - DNA sample
  - Distance
  - M348

- **protein sample F371**
  - DNA sample
  - Distance
  - M349

- **protein sample M386**
  - DNA sample
  - Distance
  - M386
Summary

- This shouldn’t happen.
- But if it does, you should find it.
- If two data sets have rows that correspond, you should check that they do correspond.


Slides: kbroman.org/Talk_OSGA2021

kbroman.org

github.com/kbroman

@kwgbromam
DNA ↔ protein: best vs 2nd-best