Identifying sample mix-ups in eQTL data

Karl Broman

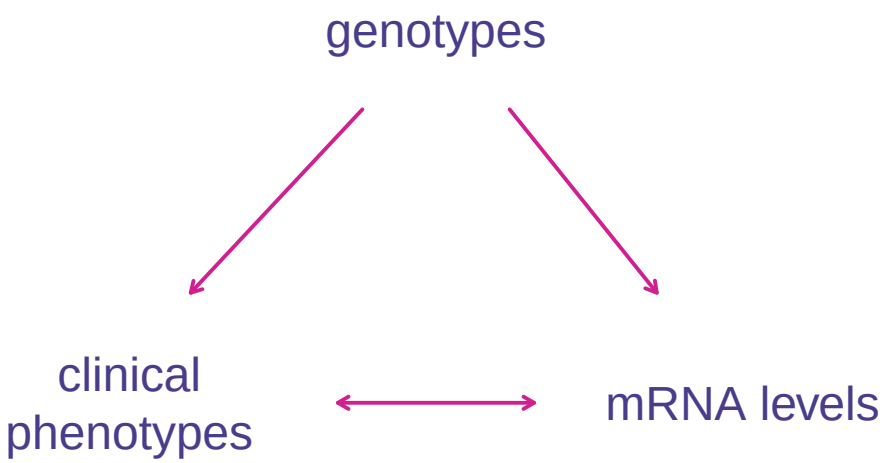
Biostatistics & Medical Informatics, Univ. Wisconsin–Madison

kbroman.org
github.com/kbroman
@kwbroman
Slides: kbroman.org/Talk_OSGA2021

These are slides for a talk for the OSGA seminar series on 11 June 2021.

Source: https://github.com/kbroman/Talk_OSGA2021
Slides: https://kbroman.org/Talk_OSGA2021

Associations in systems genetics

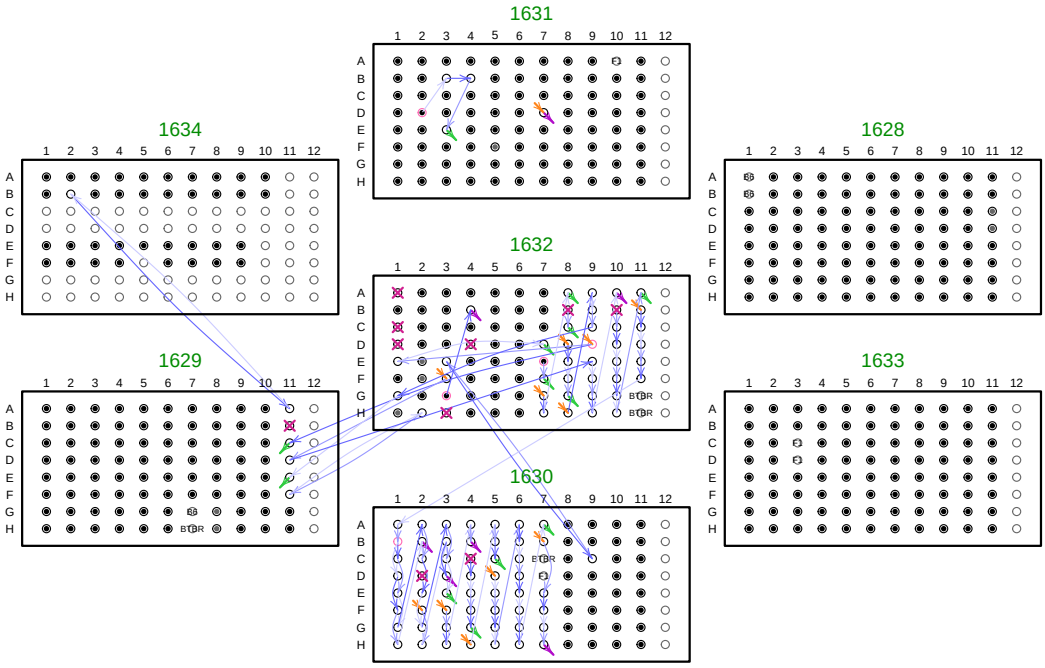


Systems genetics is all about associations between different datasets. It's critical, then, that the sample labels are correct for all data sets. As projects become larger and involve more groups of scientists, there's a greater chance for the introduction of errors in the sample labels.

Sample duplicates, mixtures, and mix-ups will all weaken associations and so reduce the quality of the study results.

On the other hand, with high-throughput genomic phenotypes, there is often the opportunity to both identify sample mix-ups and correct them.

Sample mix-ups

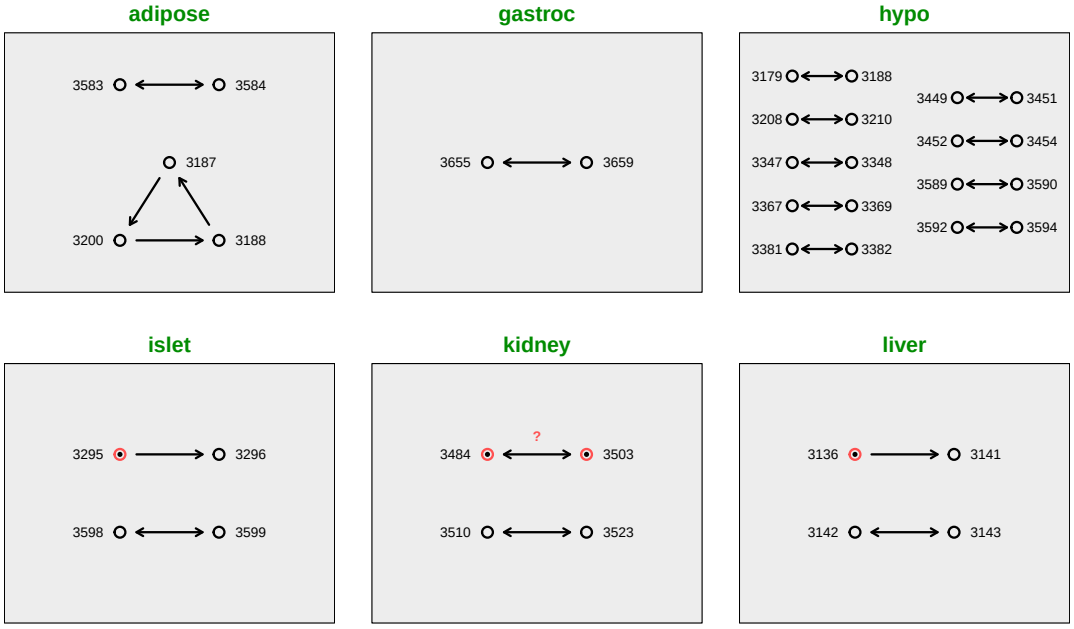


Broman et al. (2015) doi:10.1534/g3.115.019778

Here's an example of a set of mix-ups in the DNA samples for a project. In a mouse intercross with about 500 samples, there were nearly 20% mix-ups. The dots indicate that the correct sample was in the correct place. The arrows point from where a sample should have been to where it was actually found.

In this project, we had gene expression microarray data from six different tissues; that allowed us to identify and correct these errors.

More sample mix-ups



Broman et al. (2015) doi:10.1534/g3.115.019778

The mRNA samples had mix-ups, too. There were errors in each of the six tissues.

Westra et al. (2011)

Table 2. *Cis*-eQTL mapping and sample mix-up identification results

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

Westra et al. (2011) doi:10.1093/bioinformatics/btr323

Westra et al. (2011) was among the first to identify this potential problem and suggest a formal solution. They applied their approach to a number of public data sets and identified problems in most of them, including a study with 20% mix-ups.

Outline

- Sample duplicates
- Sex verification
- mRNA $\leftrightarrow$ protein
- mRNA $\leftrightarrow$ DNA
- protein $\leftrightarrow$ DNA

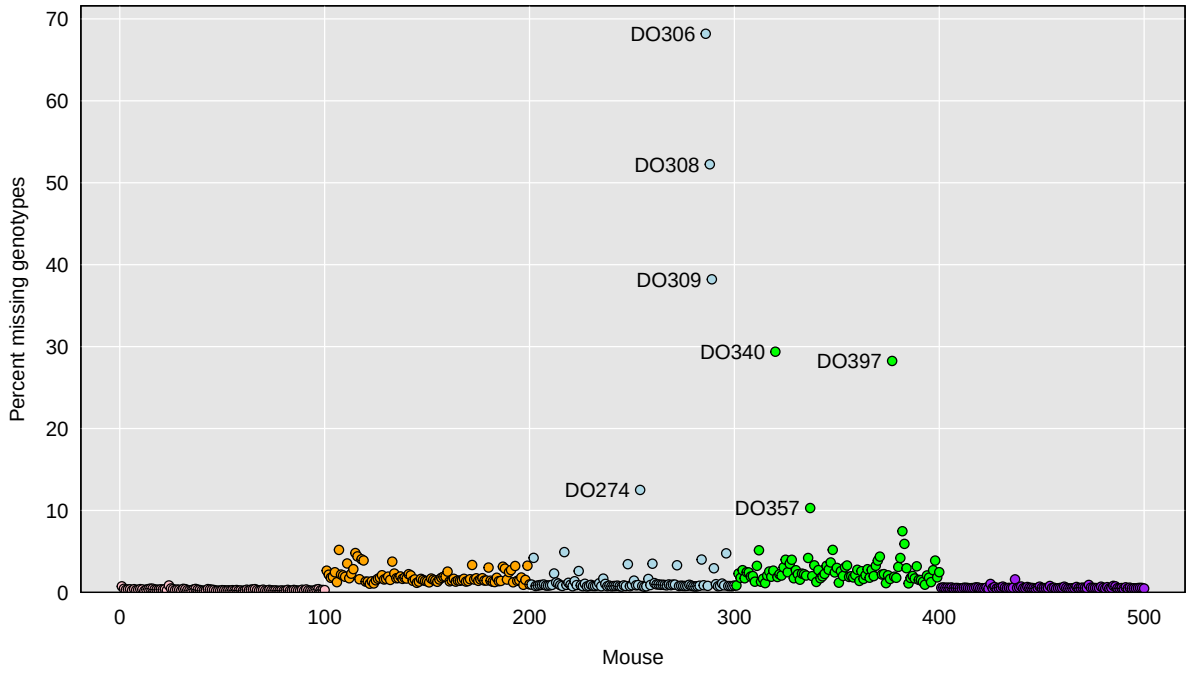
If you have high-throughput, low-level phenotypes, you should at least attempt to identify potential sample mix-ups. My goal is to make it clear how to do this, to help ensure that this becomes a routine part of the data cleaning procedures in eQTL analyses.

But first,

Missing Data

Before you do anything, you should look at the amount of missing data, as this is often an important indication of sample quality.

Percent missing genotypes



Here's a diversity outbred mouse project with 500 mice. Five samples had $> 25\%$ missing data and almost surely need to be omitted. A couple of samples have around 10% missing data and might be recoverable but are still worth watching.

References

- Westra et al. (2011) MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics 15:2104–2111 doi:10.1093/bioinformatics/btr323
- Lynch et al (2012) Calling sample mix-ups in cancer population studies. PLOS One 7:e41815 doi:10.1371/journal.pone.0041815
- Broman et al. (2015) Identification and correction of sample mix-ups in expression genetic data: A case study. G3 (Bethesda) 5:2177–2186 doi:10.1534/g3.115.019778
- Broman et al. (2019) Cleaning genotype data from Diversity Outbred mice. G3 (Bethesda) 9:1571–1579 doi:10.1534/g3.119.400165

Here are some relevant references. The Lynch et al. (2012) paper has some useful comments about experimental design.

Slides: kbroman.org/Talk_OSGA2021

kbroman.org

github.com/kbroman

@kwbroman

Here is where you can find me and my slides.