Identifying sample mix-ups in eQTL data

Karl Broman

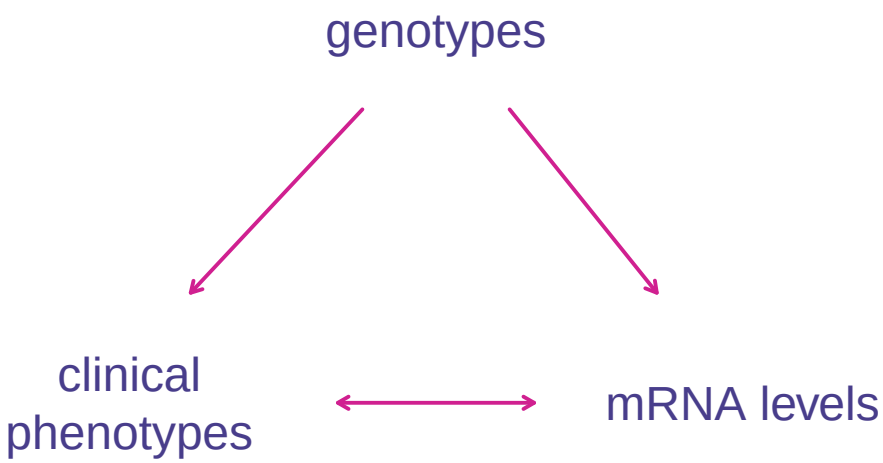
Biostatistics & Medical Informatics, Univ. Wisconsin–Madison

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/osga2021.pdf
Slides with notes: https://kbroman.org/Talk_OSGA2021/osga2021_notes.pdf

Associations in systems genetics

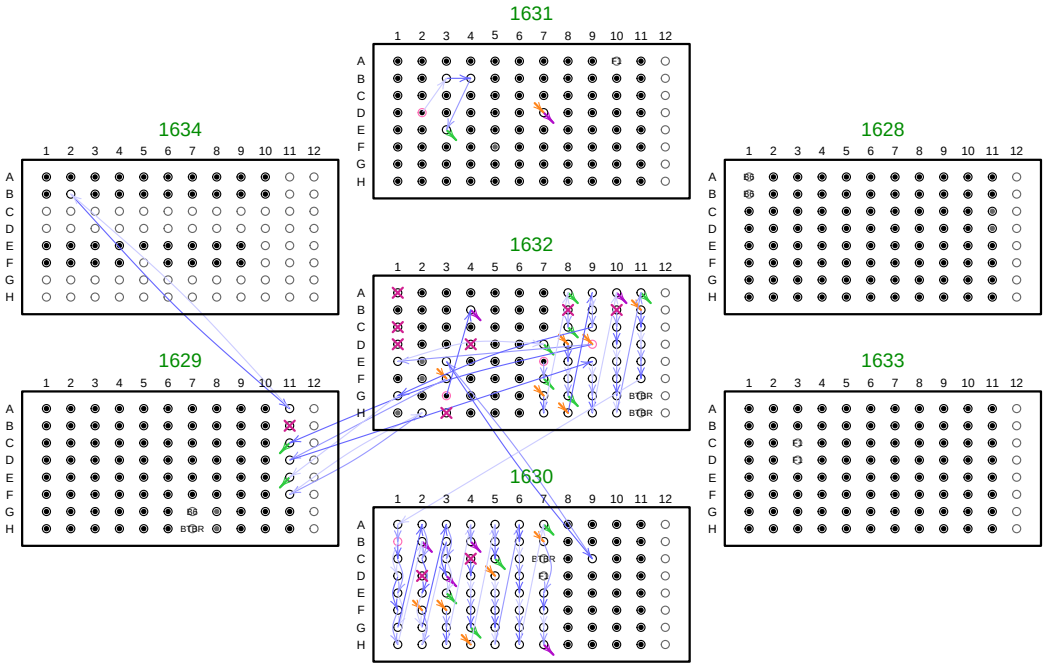


Systems genetics is all about associations between different datasets. It's critical, then, that the sample labels are correct for all data sets. As projects become larger and involve more groups of scientists, there's a greater chance for the introduction of errors in the sample labels.

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

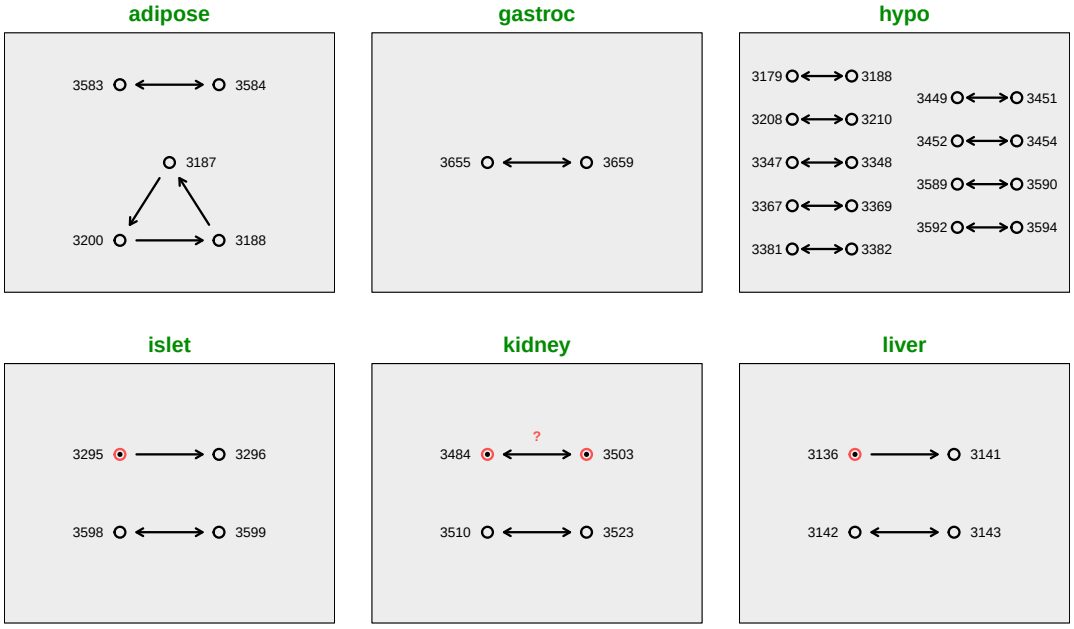


Broman et al. (2015) doi:10.1534/g3.115.019778

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
- Sample mix-ups:
  - mRNA $\leftrightarrow$ protein
  - mRNA $\leftrightarrow$ DNA
  - protein $\leftrightarrow$ DNA

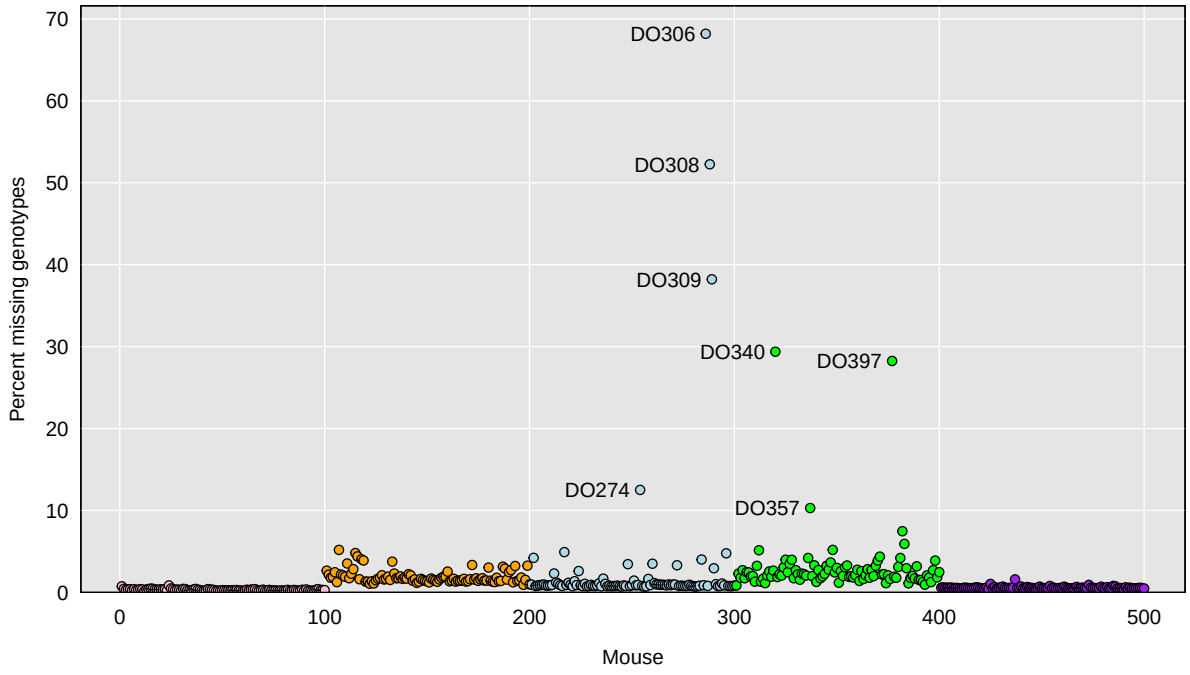
If you have high-throughput, low-level phenotypes, you should at least attempt to identify potential sample mix-ups. My goal in this talk is to make it clear how to do this, to help ensure that this becomes a routine part of the data cleaning procedures in eQTL analyses.

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.

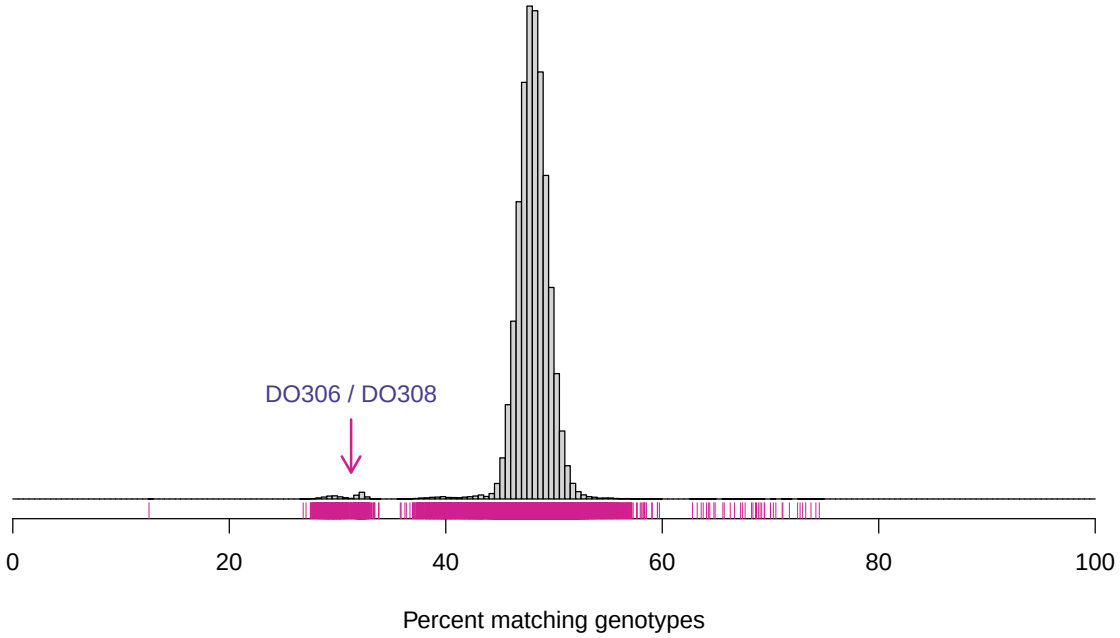
Note that I'll be using a variety of data in this talk, but I won't be explaining where it's from. But I thank my collaborators for the data.

Sample duplicates

The next thing to look for is sample duplicates. Are there pairs of individuals with too-similar genotypes?

These are pretty common. I don't know anything about monozygotic twins among mice, but we've always assumed that these are cases of sample duplication or contamination.

Percent matching genotypes

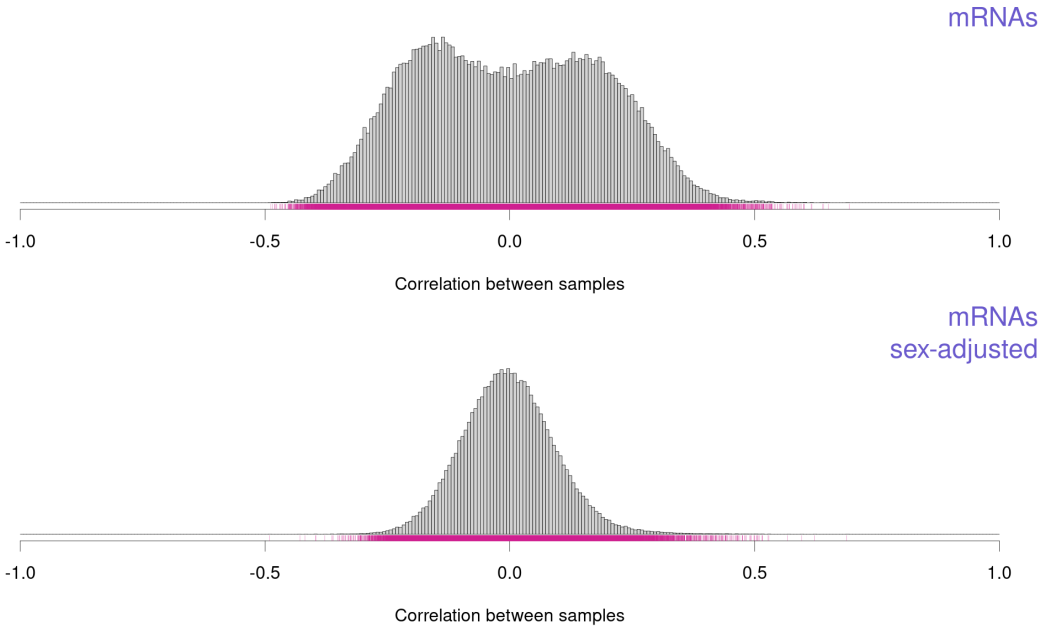


It's simple to look for duplicate DNA samples: just calculate the proportion of matching genotypes for each pair of samples, and look for pairs that have very similar data.

Here, we see no close matches. There's a group with rather low sharing, which are due to a couple of bad DNA samples, plus a group with somewhat above-normal sharing, which are likely siblings (these are again diversity outbred mice).

This technique only works well for organisms with a lot of chromosomes. It would be hard to do this in Drosophila, because the variation in the "just by chance" sharing would be really high and cover the full rather 0–100%.

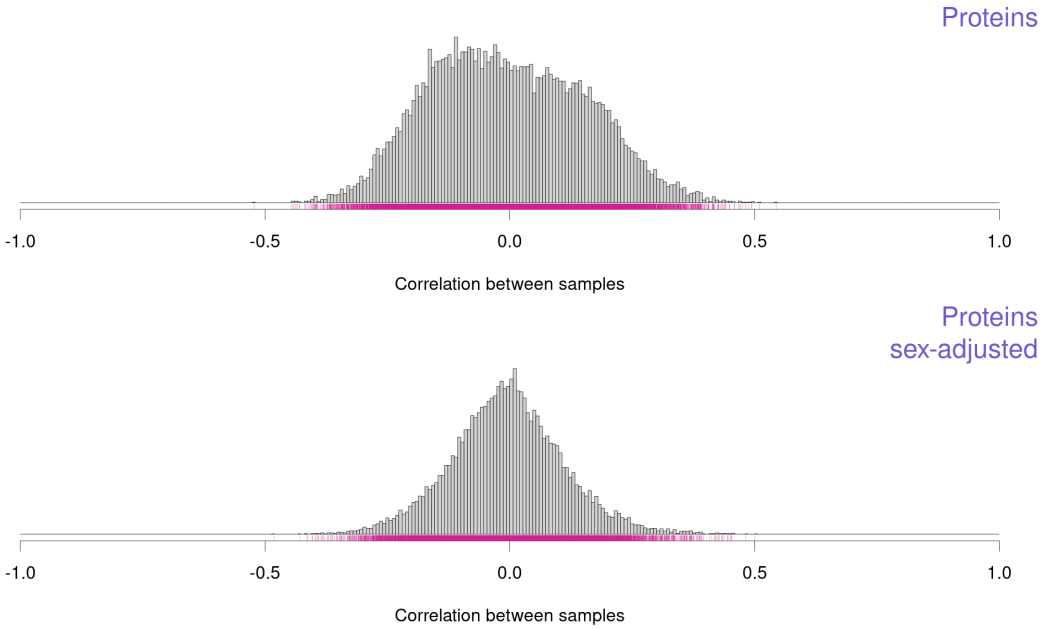
Correlation between mRNA samples



It seems like you should be able to do the same thing with mRNA or protein samples: just look at the correlation between samples, or perhaps the RMS difference. But I've not had much success finding duplicate samples this way. You maybe need to exclude genes that appear to not be expressed (and so are just noise).

These are histograms of the correlation between mRNA expression samples. The two sexes are anti-correlated. The lower histogram is for correlation between measurements after controlling for sex.

Correlation between protein samples

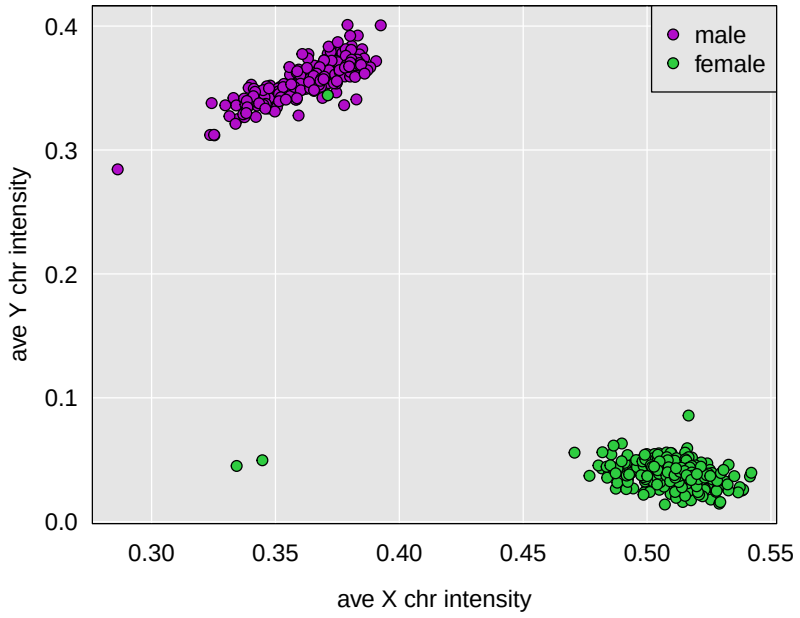


This is similar as the last slide, but with mass-spec-based protein measurements. The anti-correlation between sexes is not as strong but still present.

Sex verification

One way to identify sample mislabelings is by comparing the annotated sex to what you can infer from the genotypes or expression data.

X and Y genotype dosage



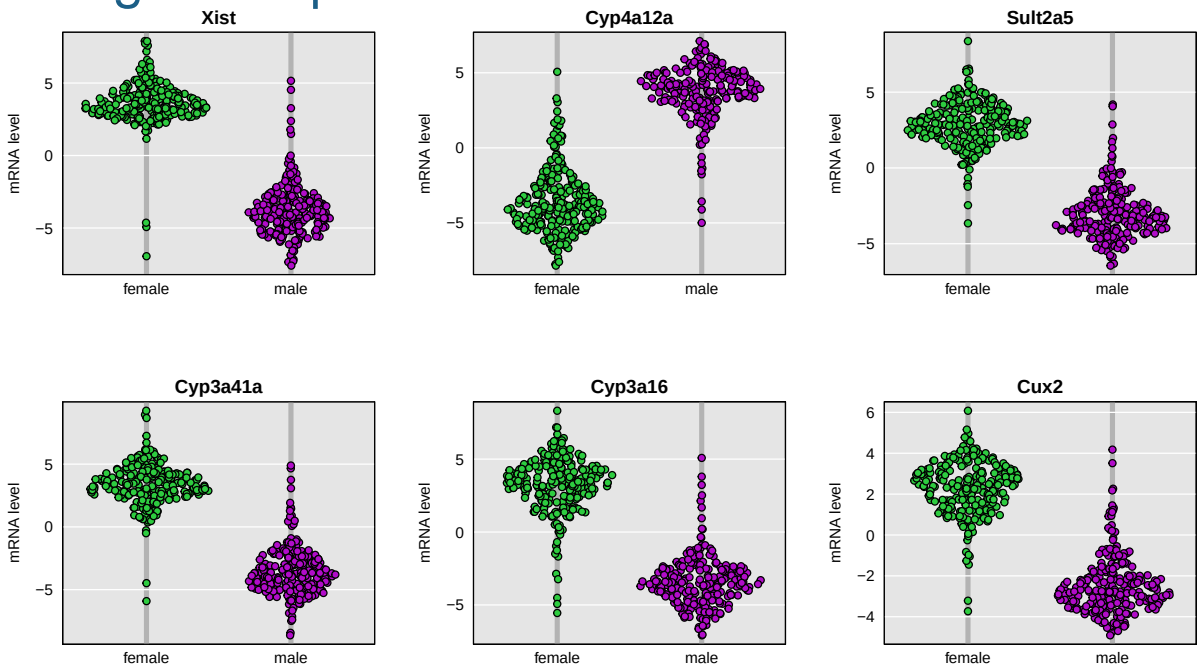
Historically, I would look at heterozygosity on the X chromosome to verify sex. But even better, for verifying sex in the genotype data, is to look at the dosage of X and Y chromosome markers (average intensity for microarray-based genotypes, or frequency of mapped reads for sequencing-based genotypes).

The x-axis is average intensity of SNPs on the X chromosome; the y-axis is average intensity of SNPs on the Y chromosome.

The green ball in the lower-right are females with two X chromosomes and no Y. The purple ball in the upper-left are males with one X and one Y. The points in the lower-left are maybe XO females.

We are looking for females in the upper-left (and there is one such) or males in the lower-right.

Sex and gene expression

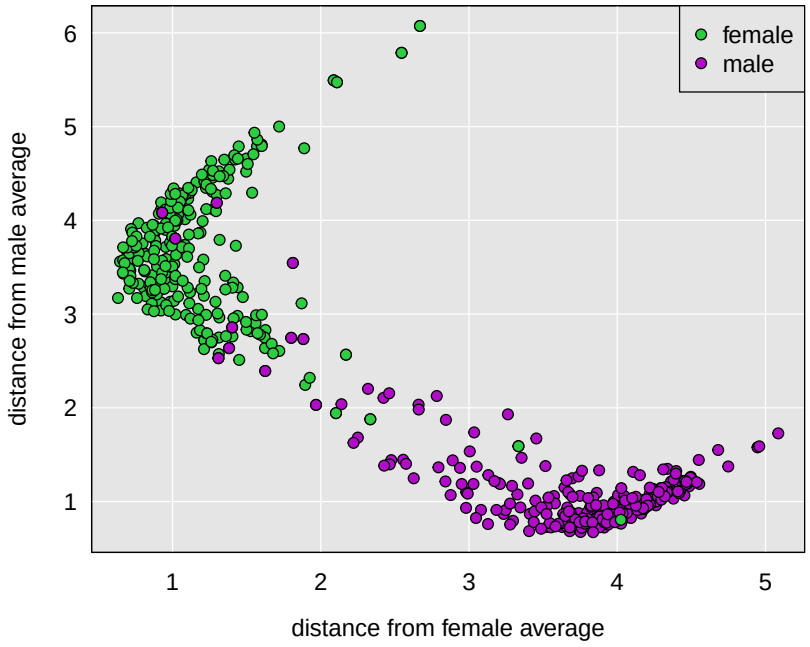


We should be able to do the same sort of thing, to verify the sex of the mRNA samples.

We could think hard about genes that should show a big sex difference (for example, Xist), or we could just do t-tests to identify a set of genes that have big sex differences. These are the top six genes, in terms of sex difference in expression. Of course, Xist is the first.

We could choose the top say 100 genes and use them to form a classifier for sex from gene expression. We want a method that can handle some misclassified samples.

Sex and gene expression

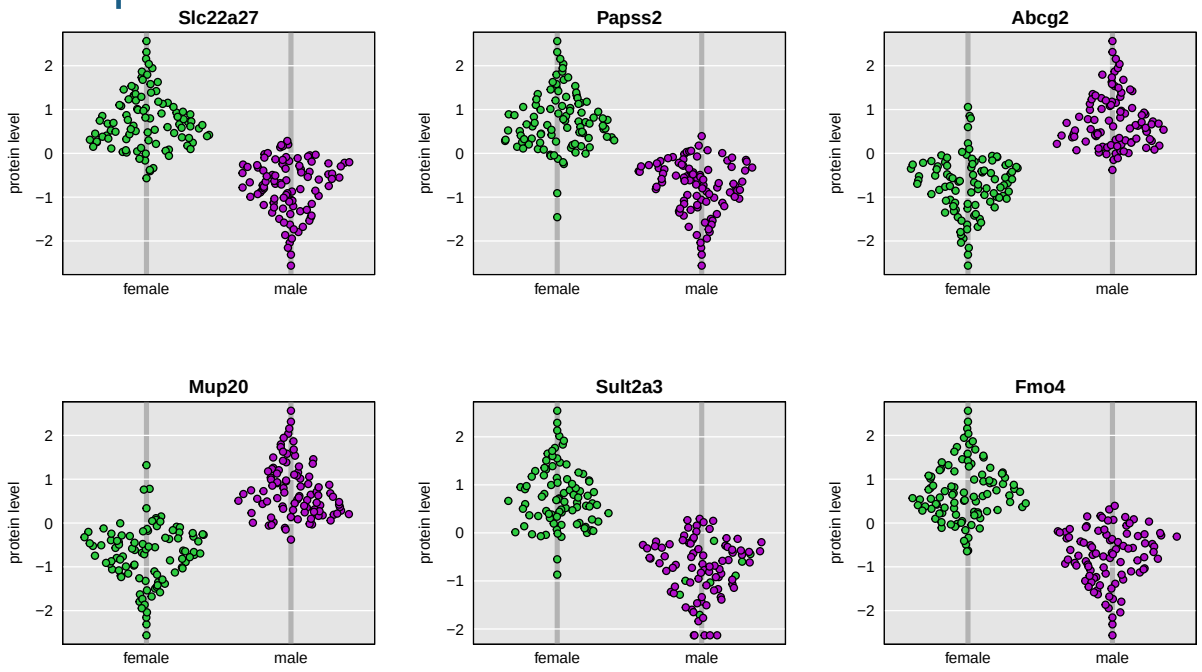


What I ended up doing was just pick the top 100 genes and then calculate the mean in males and females, and then for each sample, look at the RMS difference from the male means and from the female means. This is a plot of those.

Most of the samples have well-differentiated sex by this approach, but there are a bunch of samples in the middle, maybe due to batch differences that I've not accounted for?

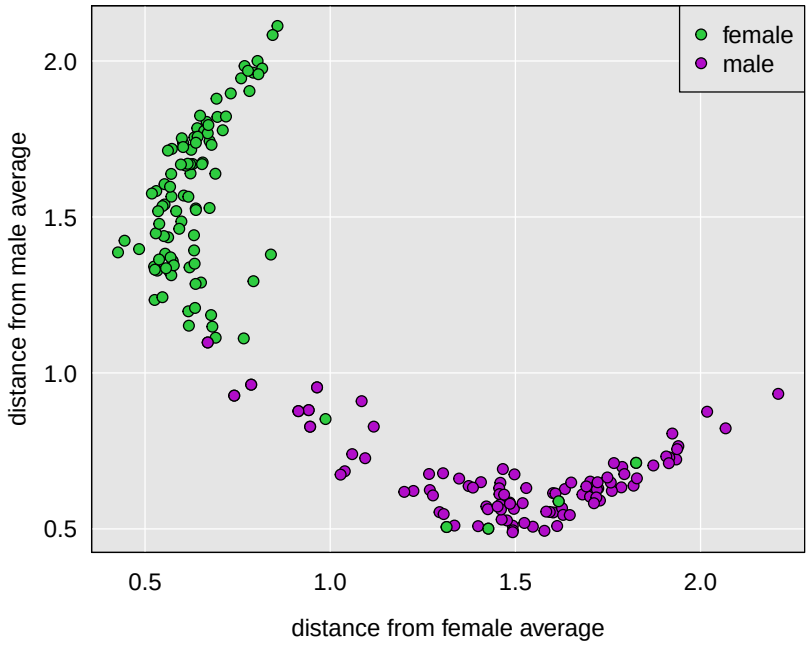
Anyway, while there are a number of samples in the middle that are unclear, there's also a set of males that look like females and one female that looks like male.

Sex and proteins



Protein levels also show big sex differences, so we can do the same thing again with proteins.

Sex and proteins



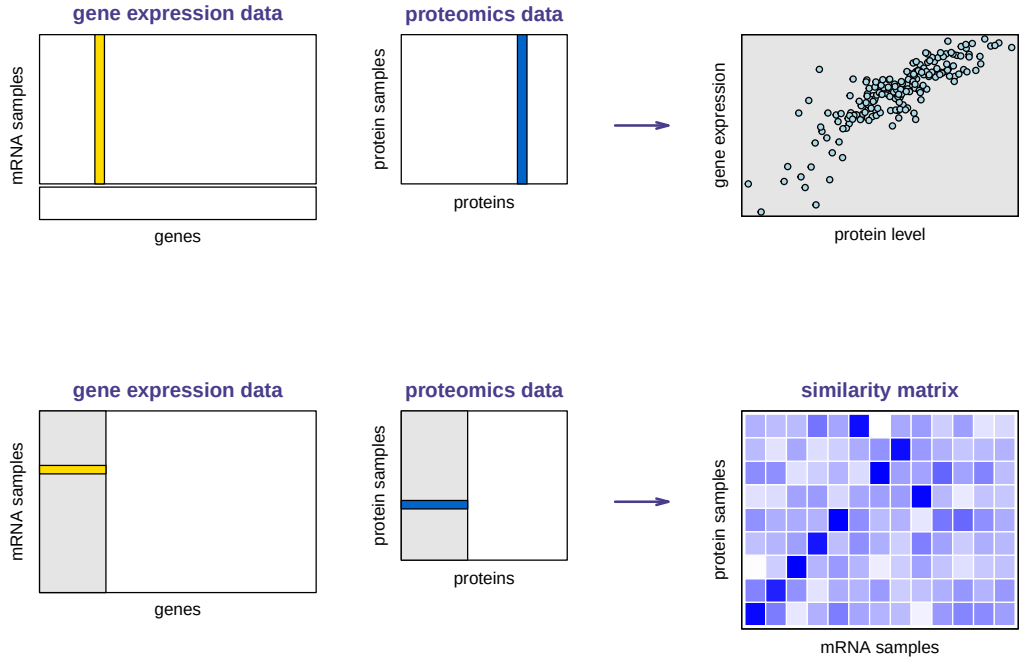
I did the same thing as with the mRNA data: pick the top 100 proteins by their sex difference. Calculate the average for males and for females, and then take the RMS difference for each sample from the male means and the female means.

For this particular data set, there are a number of females deep within the males.

Sample mix-ups

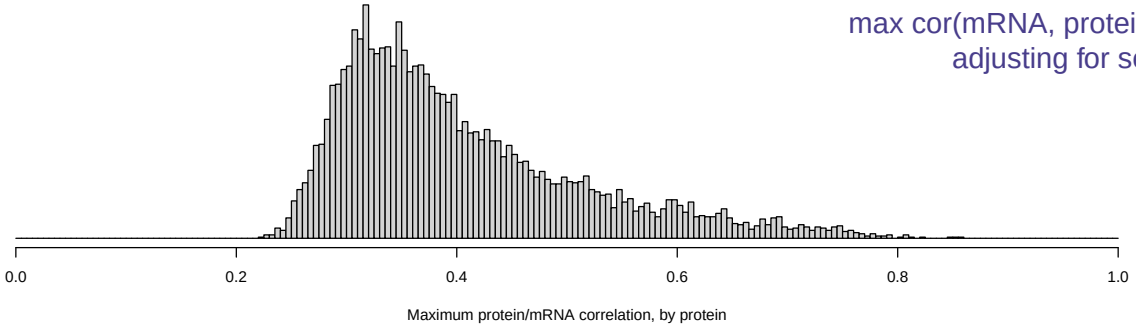
mRNA $\leftrightarrow$ protein

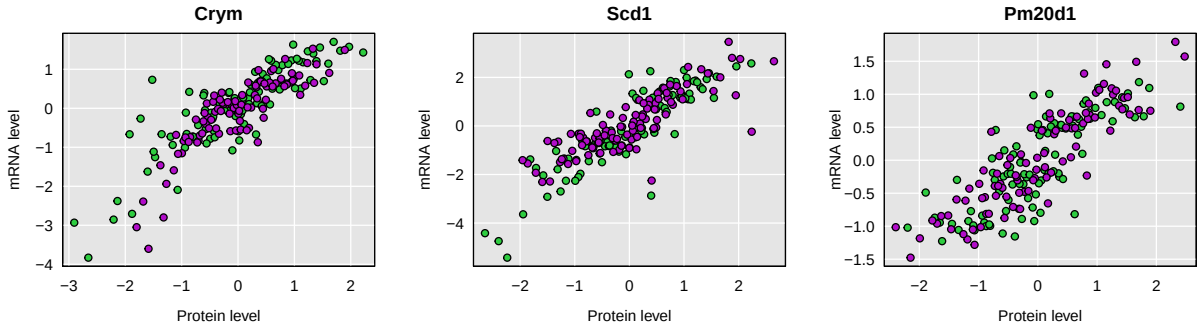
mRNA ↔ protein method



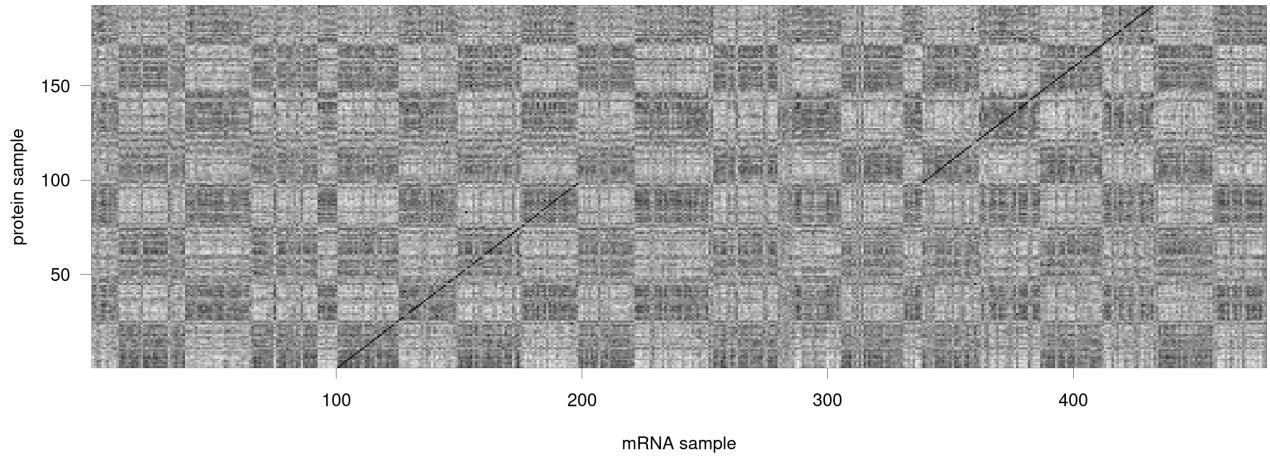
mRNA ↔ protein correlations

max cor(mRNA, protein)
adjusting for sex

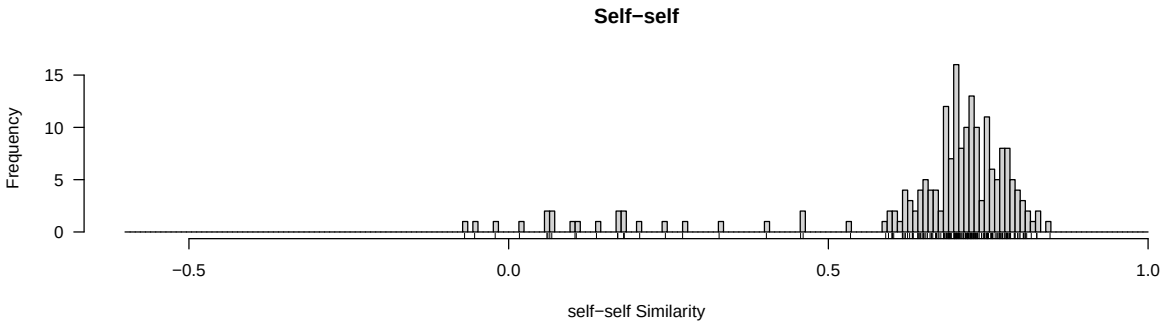




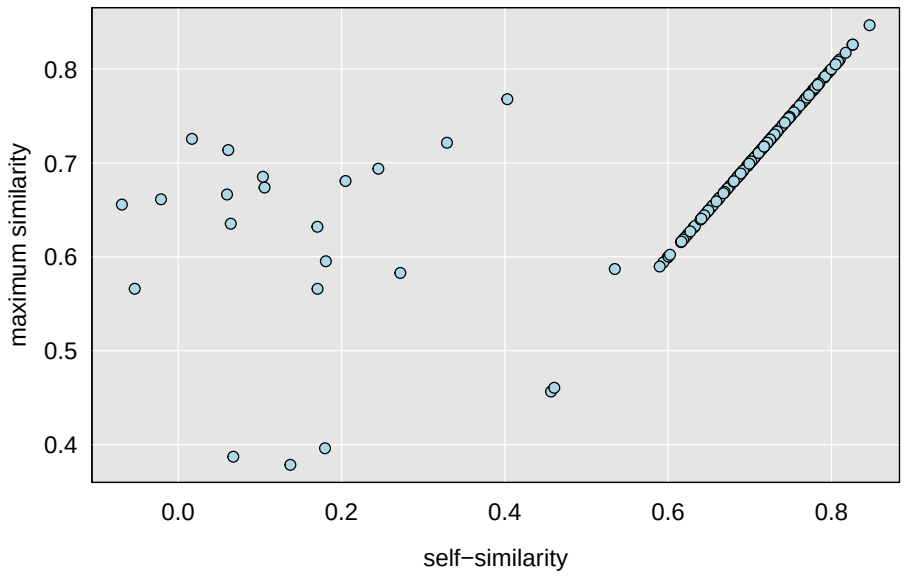
mRNA $\leftrightarrow$ protein similarity matrix



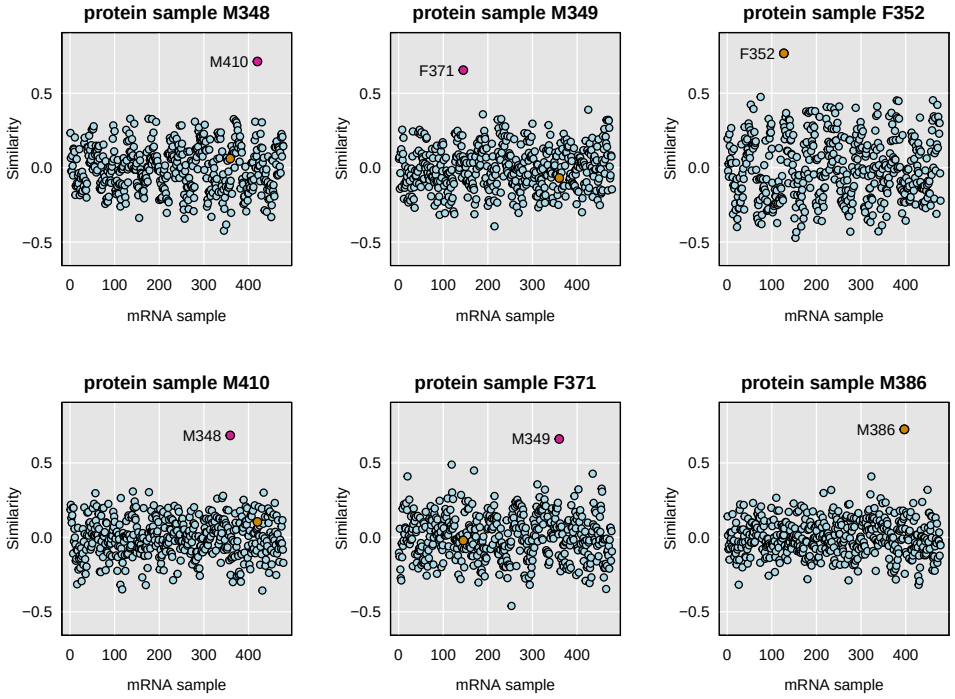
mRNA ↔ protein similarities



mRNA $\leftrightarrow$ protein: closest vs self



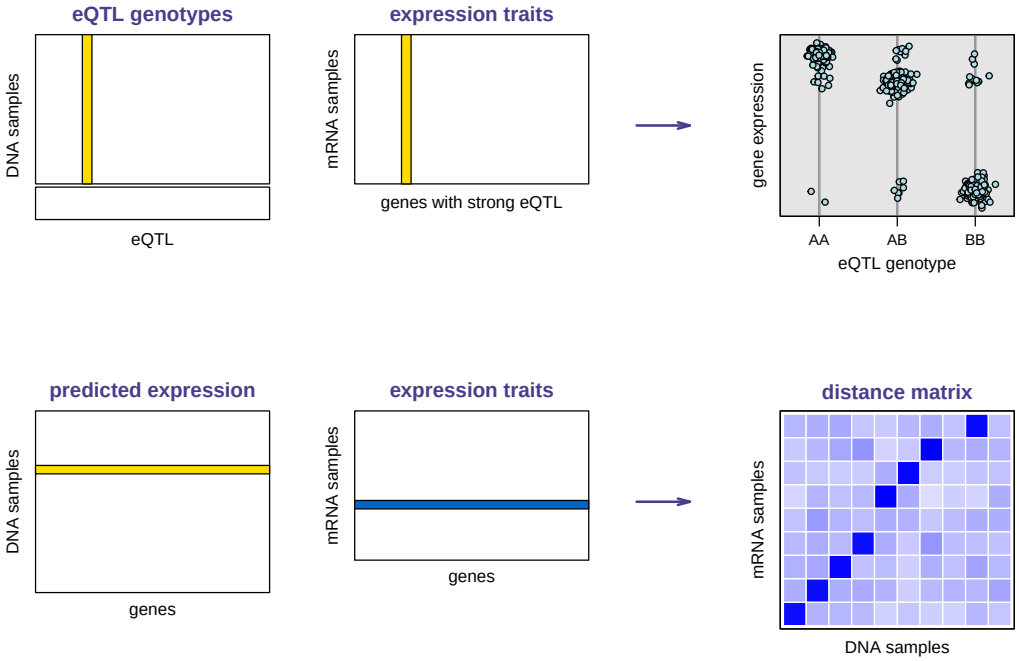
mRNA ↔ protein: selected samples



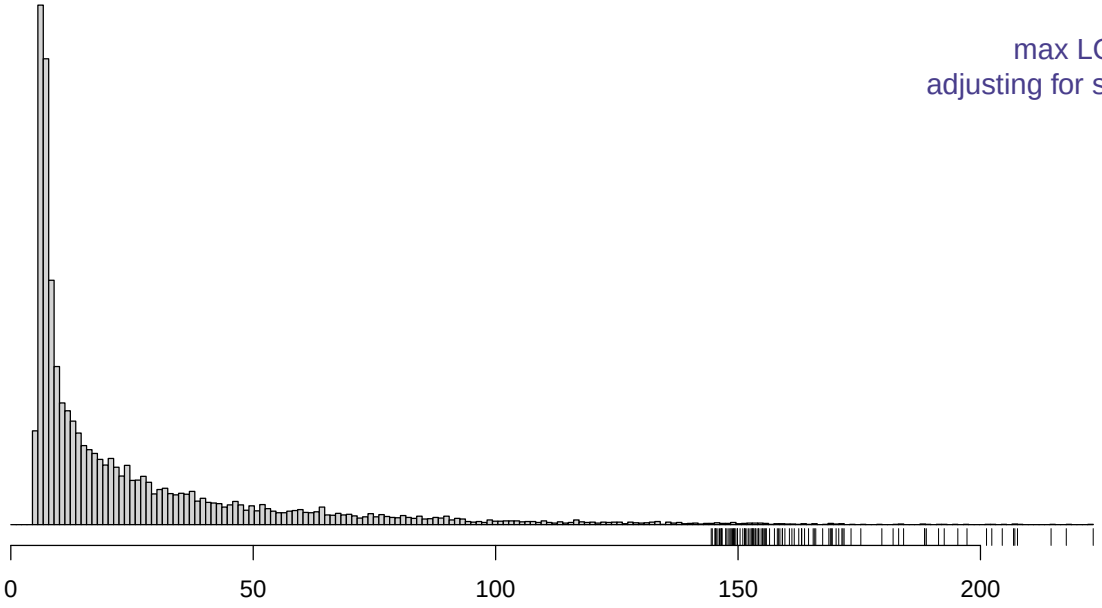
Sample mix-ups

DNA $\leftrightarrow$ mRNA

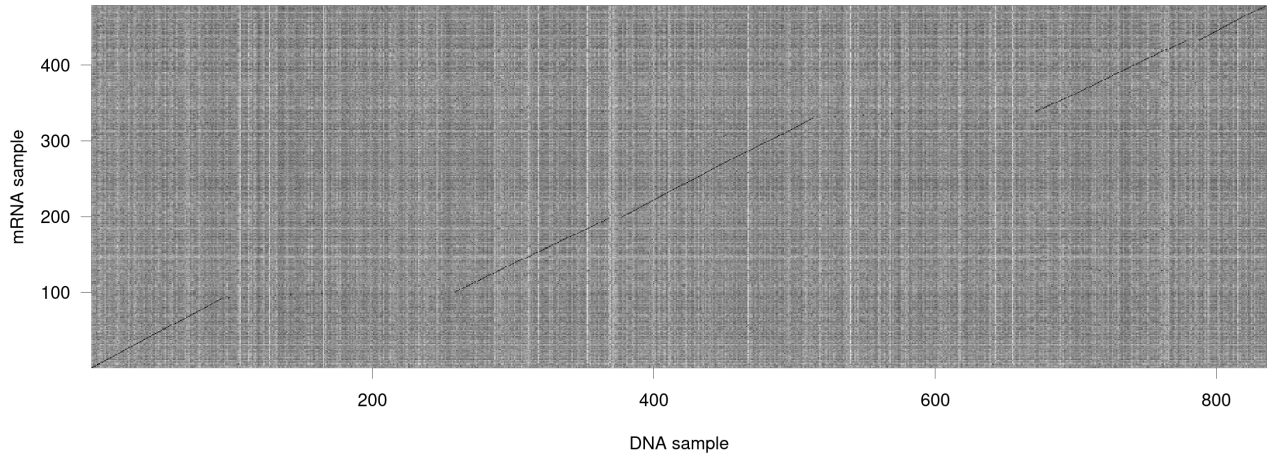
DNA ↔ mRNA method



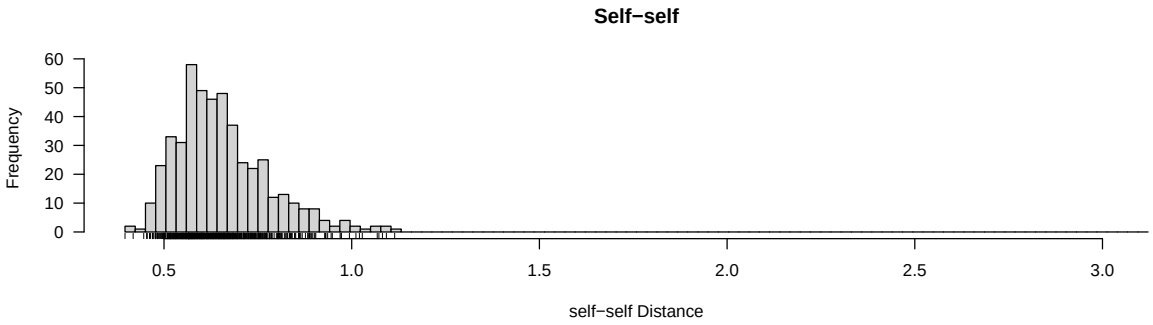
DNA $\leftrightarrow$ mRNA correlations



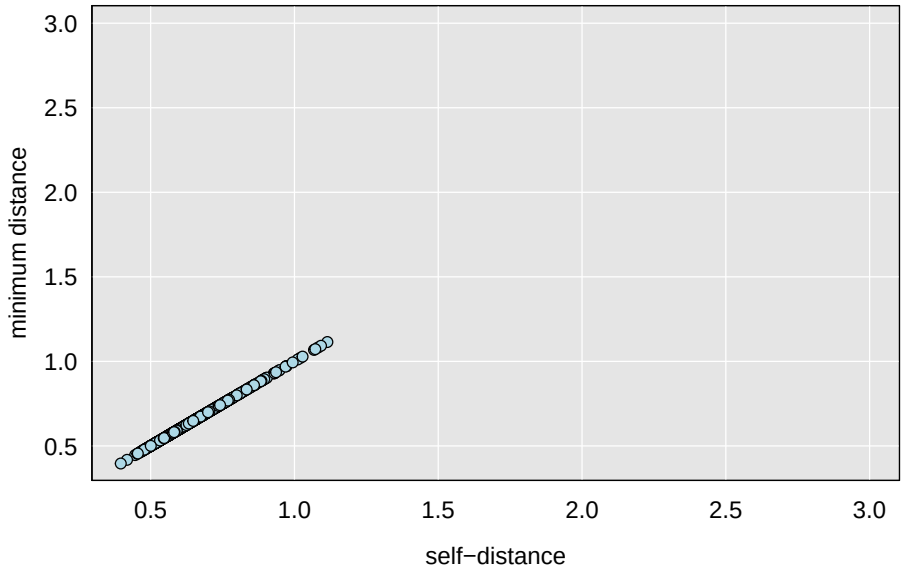
DNA $\leftrightarrow$ mRNA similarity matrix



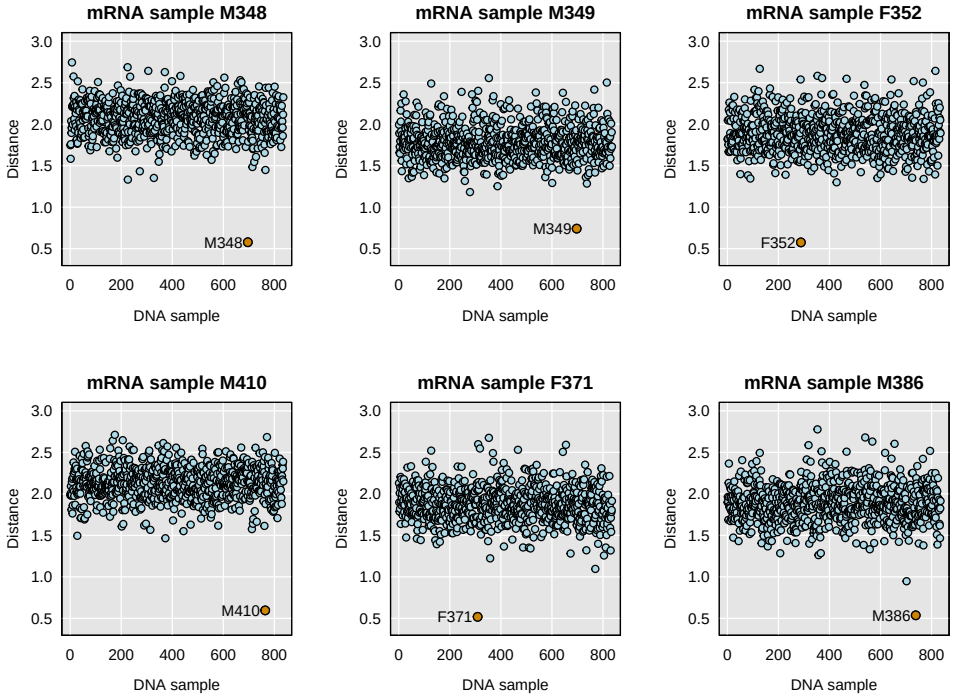
DNA $\leftrightarrow$ mRNA similarities



DNA $\leftrightarrow$ mRNA: closest vs self



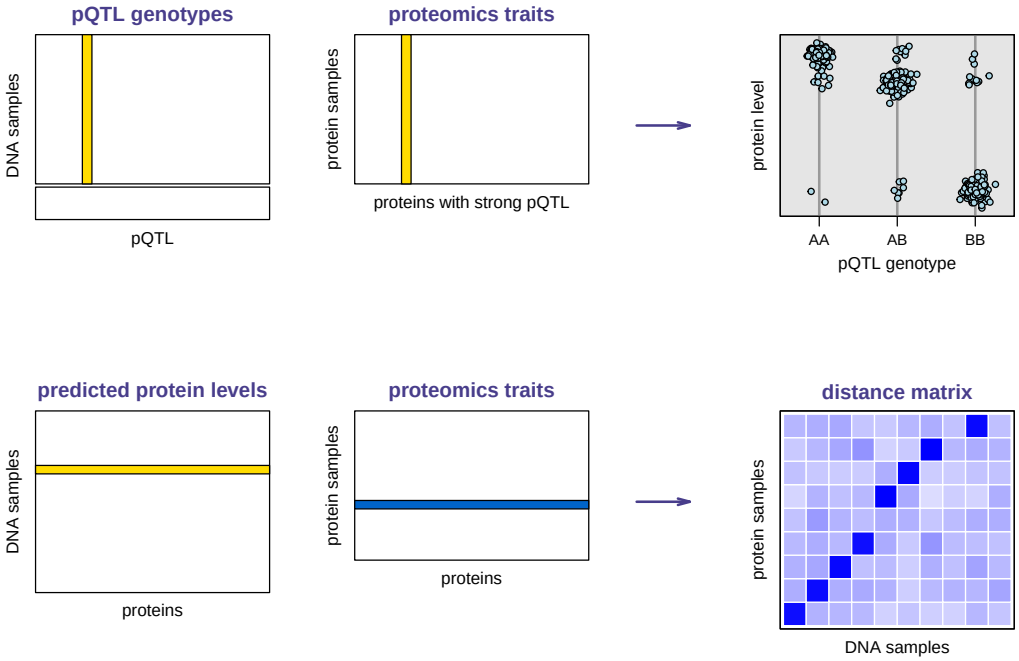
DNA ↔ mRNA: selected samples



Sample mix-ups

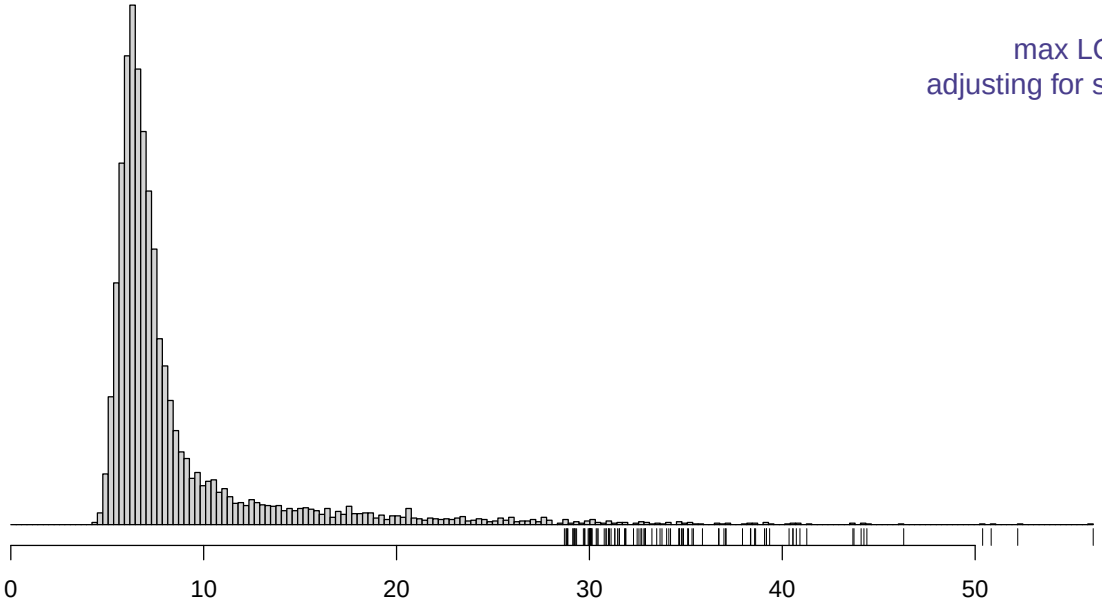
DNA $\leftrightarrow$ protein

DNA ↔ protein method

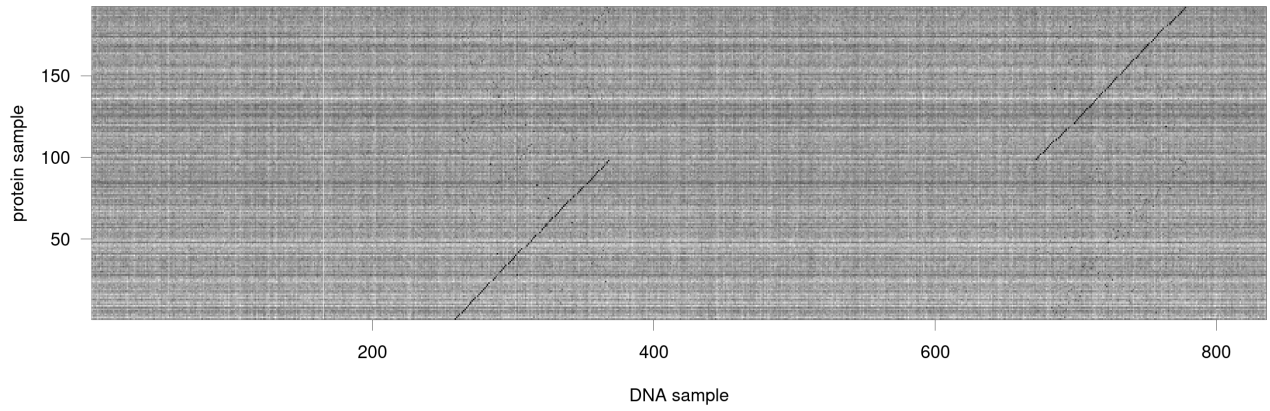


DNA $\leftrightarrow$ protein correlations

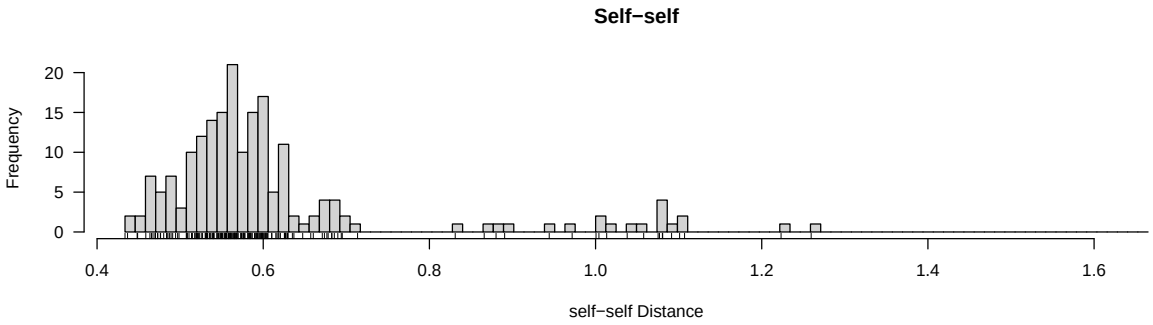
max LOD
adjusting for sex



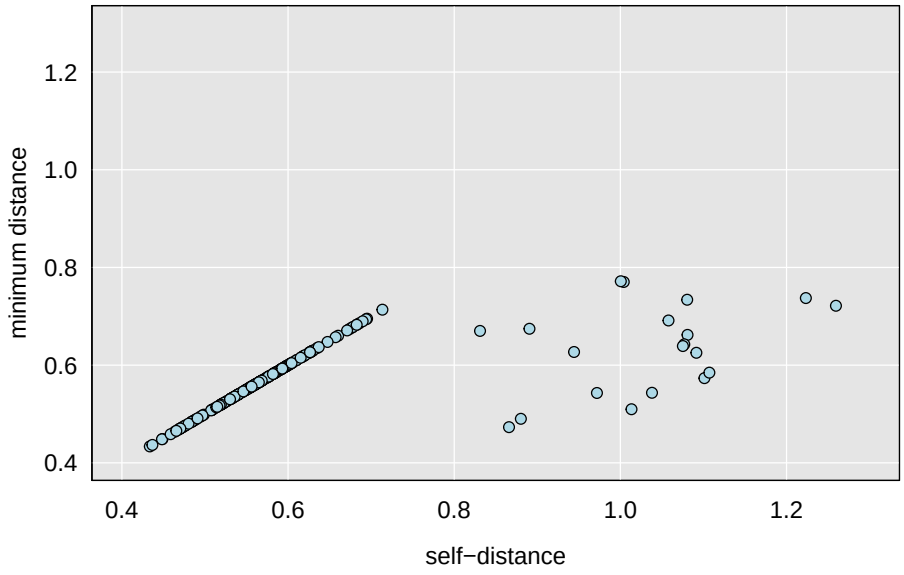
DNA $\leftrightarrow$ protein similarity matrix



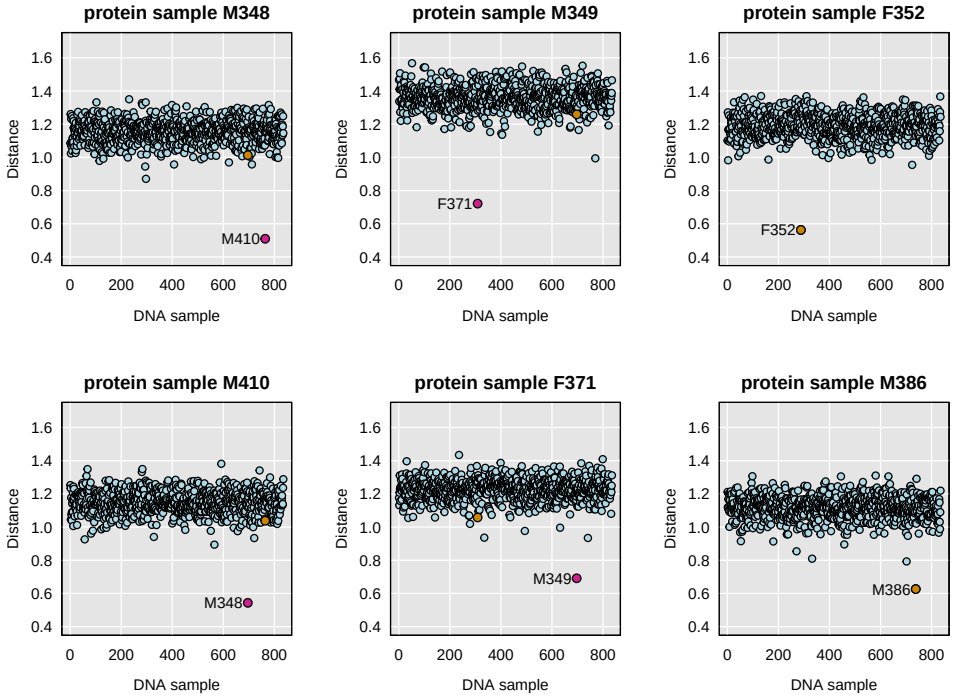
DNA $\leftrightarrow$ protein similarities



DNA $\leftrightarrow$ protein: closest vs self

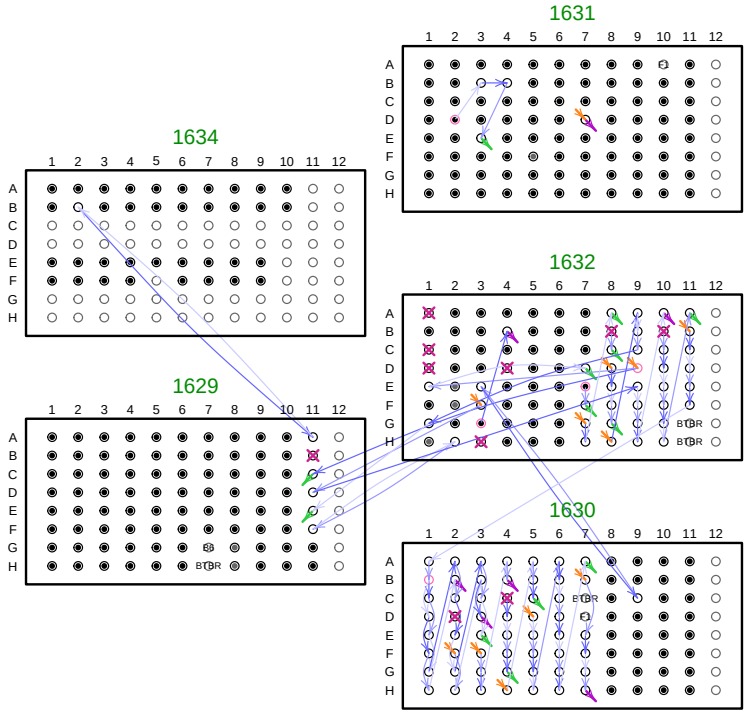


DNA ↔ protein: selected samples



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.



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.