Identifying sample mix-ups in eQTL data

Karl Broman

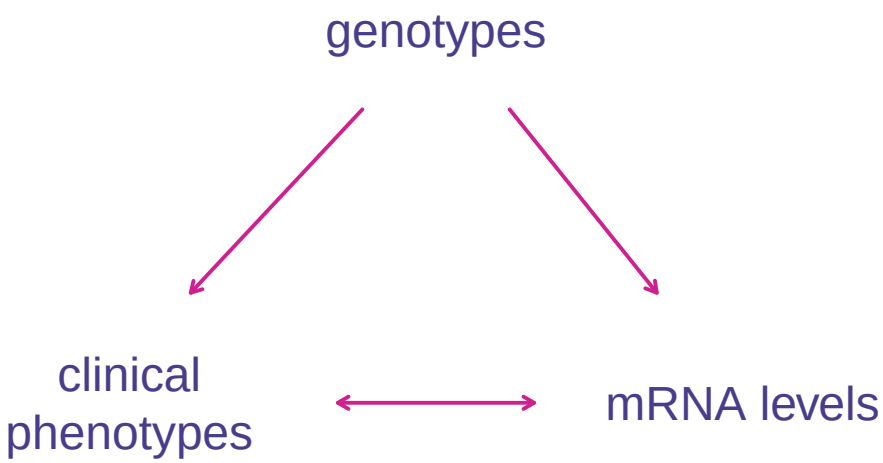
Biostatistics & Medical Informatics, Univ. Wisconsin–Madison

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

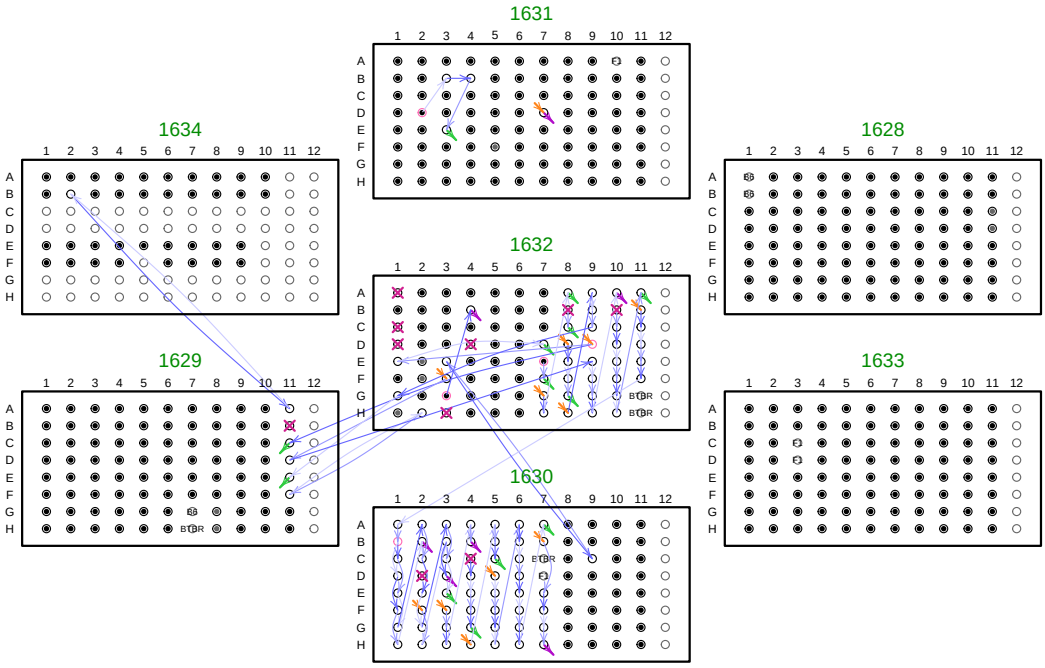


Systems genetics is all about associations between different datasets. It's critical, then, that the sample labels are correct for all data sets. As projects become larger and involve more groups of scientists, there's a greater chance for the introduction of errors in the sample labels.

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

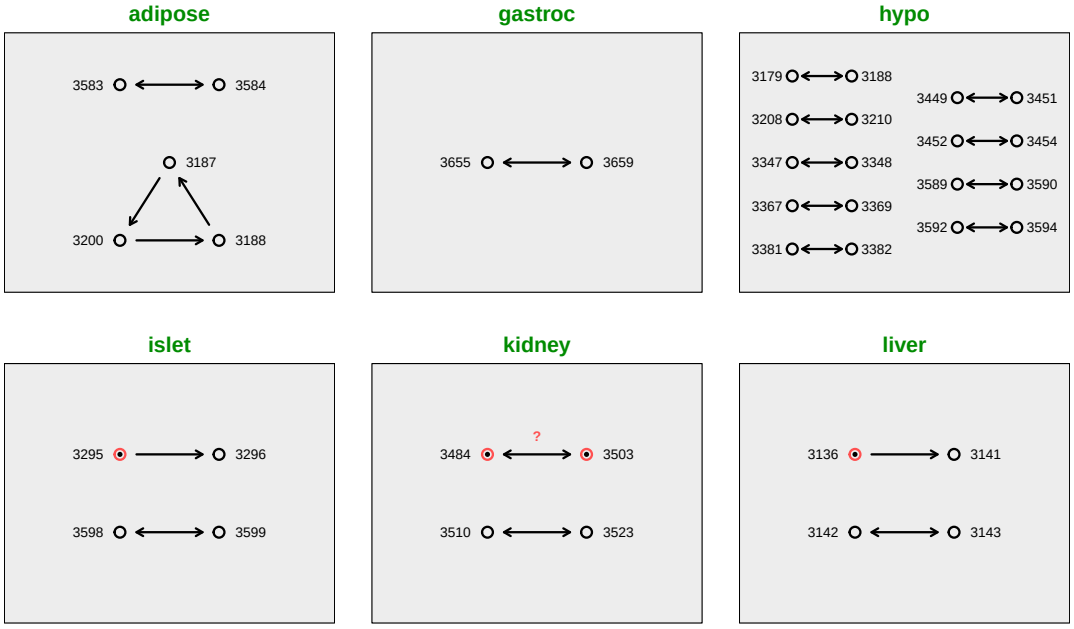


Broman et al. (2015) doi:10.1534/g3.115.019778

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

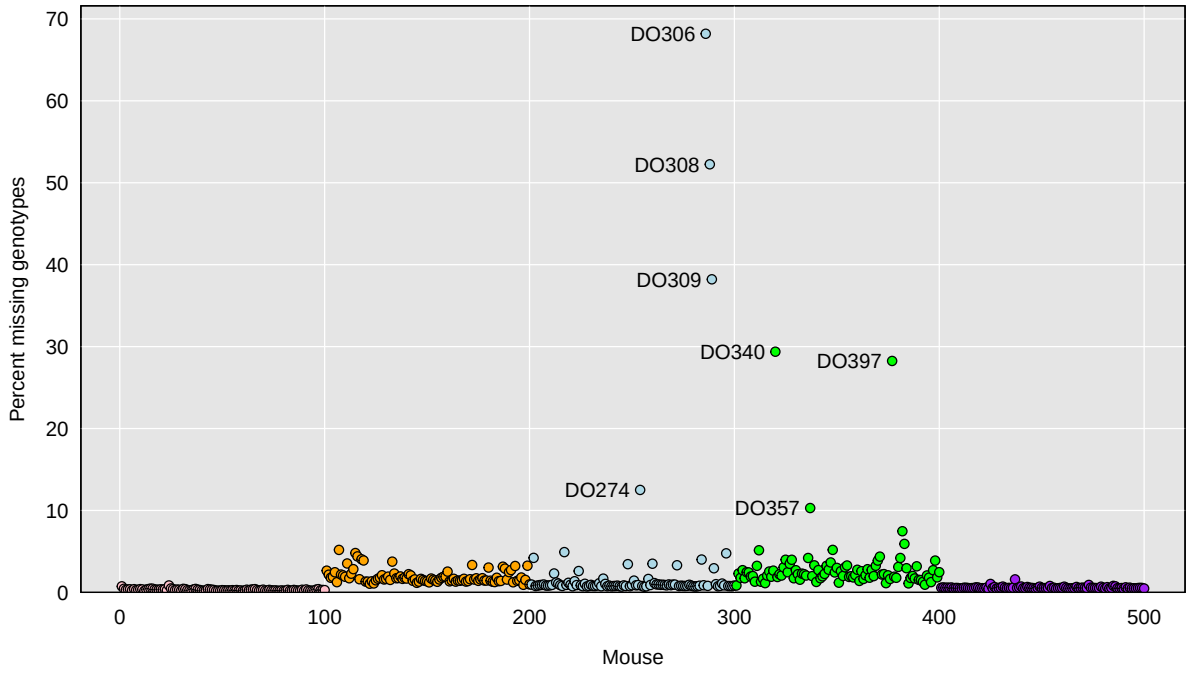
If you have high-throughput, low-level phenotypes, you should at least attempt to identify potential sample mix-ups. My goal in this talk is to make it clear how to do this, to help ensure that this becomes a routine part of the data cleaning procedures in eQTL analyses.

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.

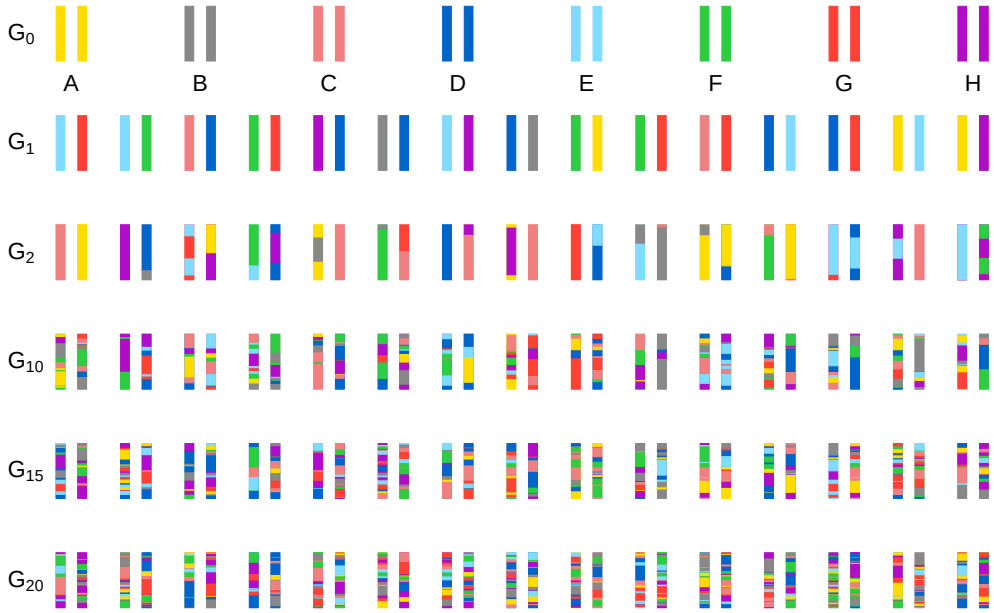
Percent missing genotypes



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.

Heterogeneous Stock/Diversity Outbreds



I'm considering two datasets here. I won't say anything about where they're from, but they're both on diversity outbred mice, one with about 500 individuals and the other with 800 (though only about 500 with gene expression data and 200 with proteomics data).

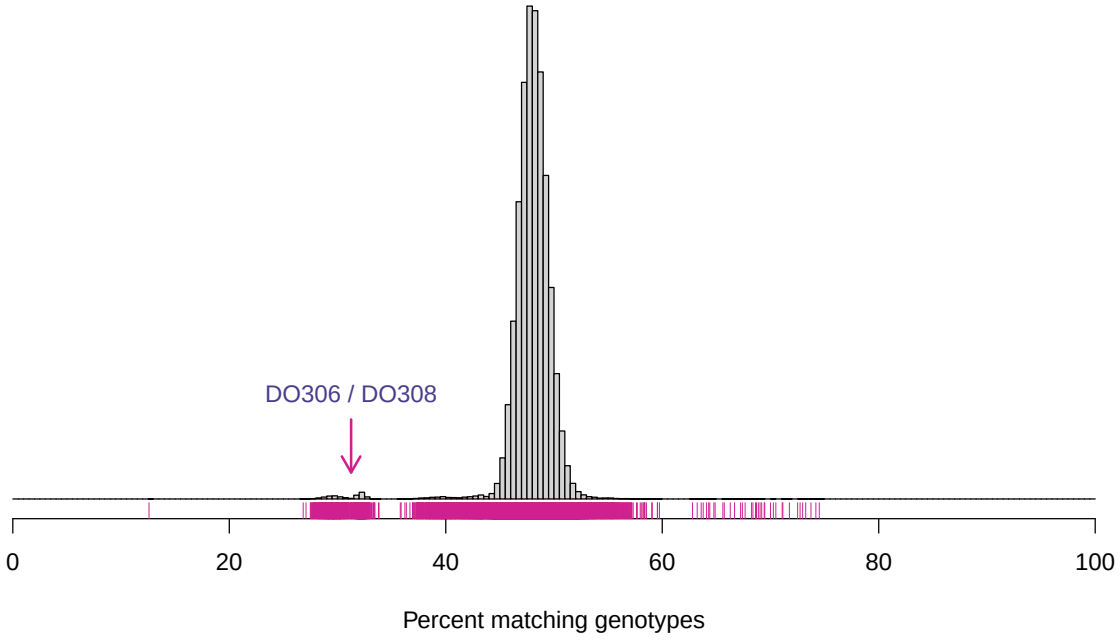
You maybe can tell the sources, but let's not talk about the particular datasets right now. I'm working on a tutorial on identifying sample mix-ups but have agreed to wait to post it; I'll post it once an article correction appears.

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

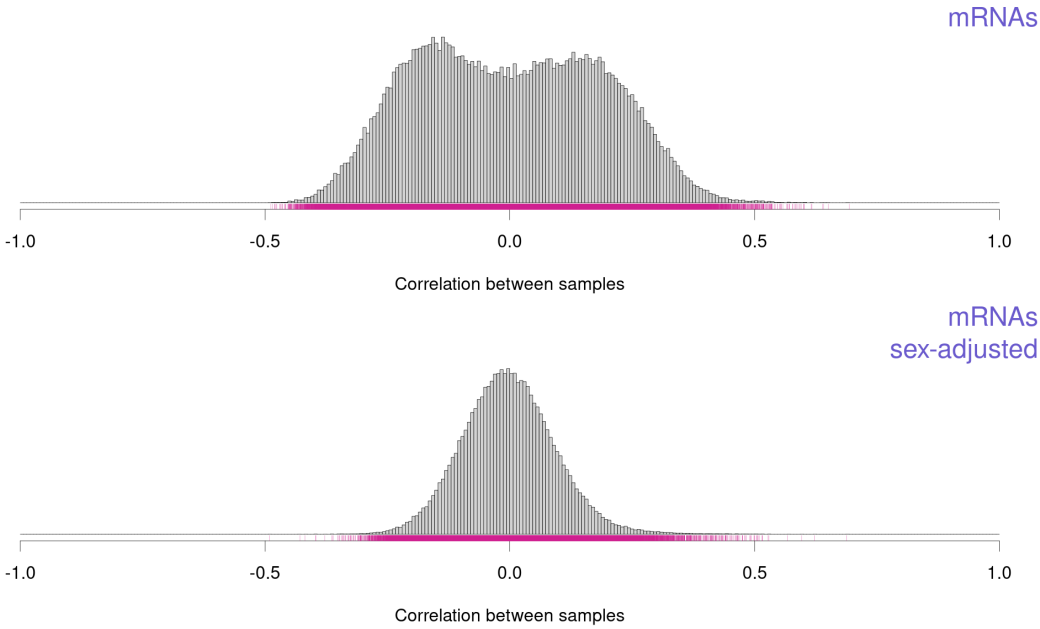


It's simple to look for duplicate DNA samples: just calculate the proportion of matching genotypes for each pair of samples, and look for pairs that have very similar data.

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%.

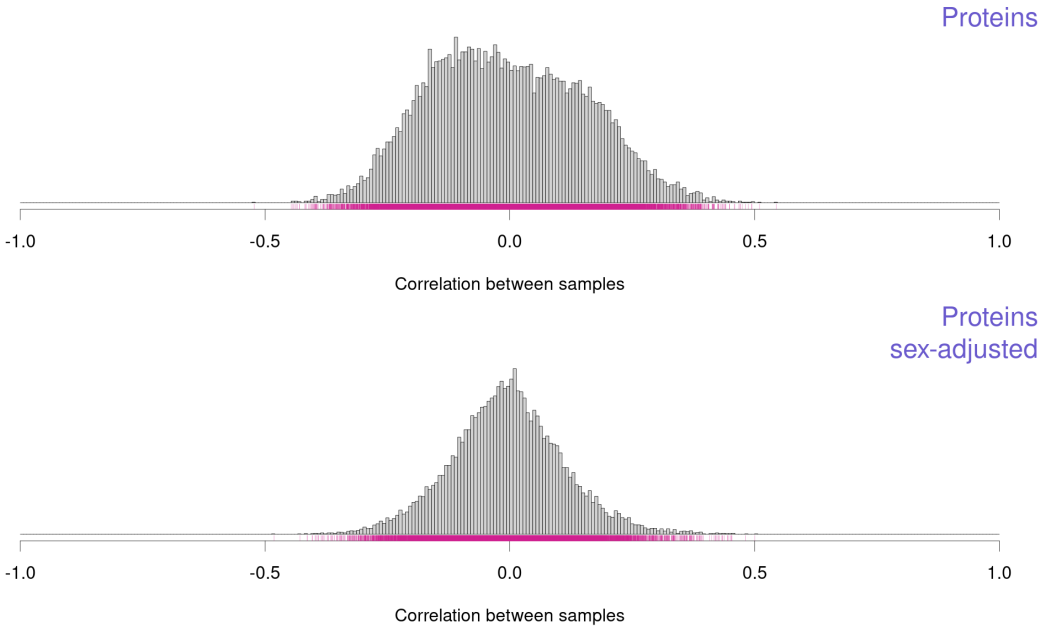
Correlation between mRNA samples



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

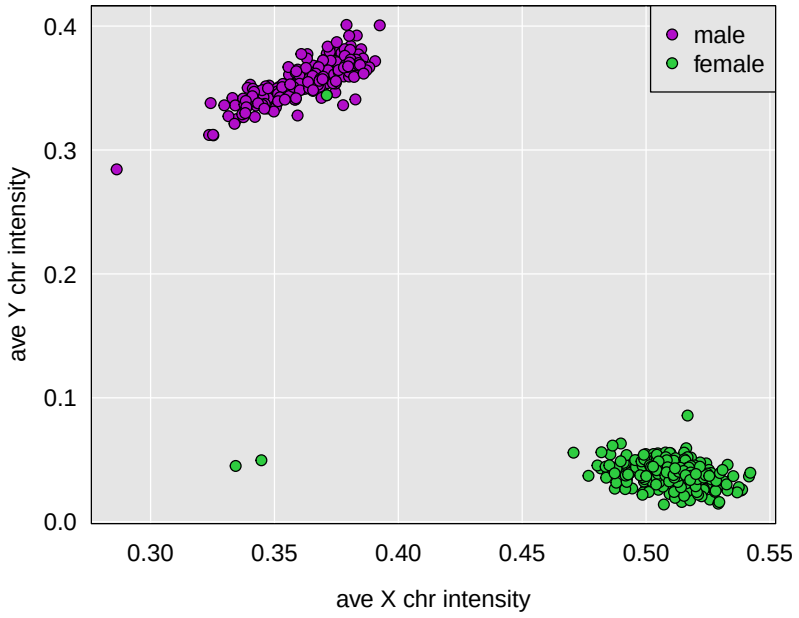


This is similar as the last slide, but with mass-spec-based protein measurements. The anti-correlation between sexes is not as strong but still present.

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



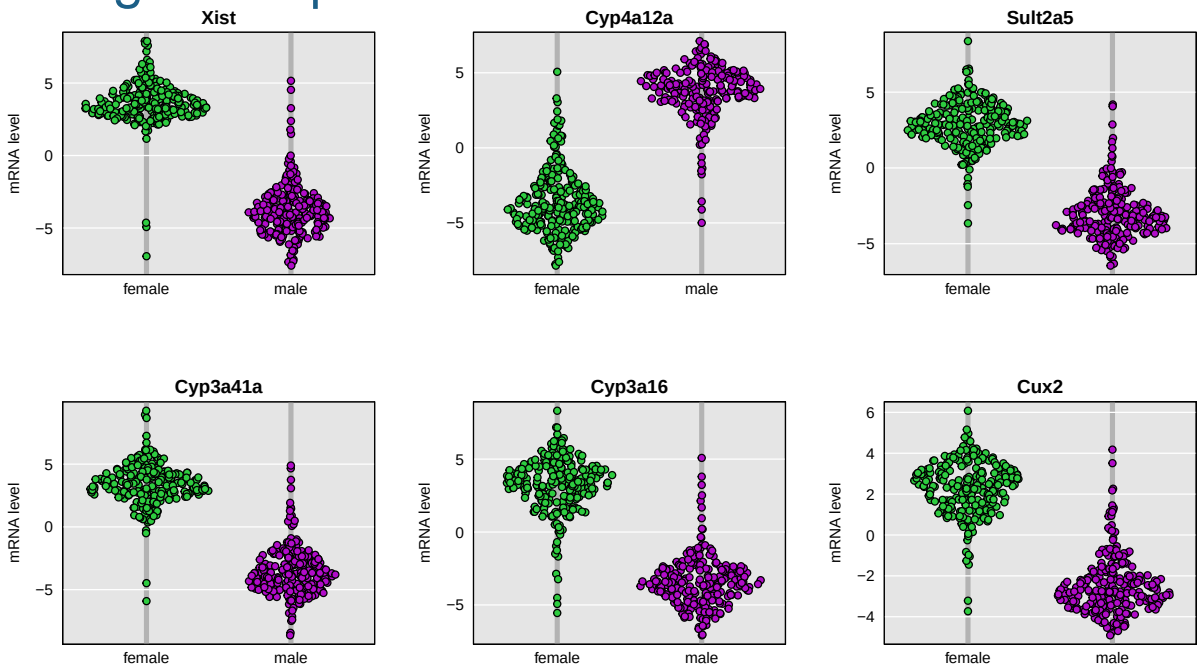
Historically, I would look at heterozygosity on the X chromosome to verify sex. But even better, for verifying sex in the genotype data, is to look at the dosage of X and Y chromosome markers (average intensity for microarray-based genotypes, or frequency of mapped reads for sequencing-based genotypes).

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

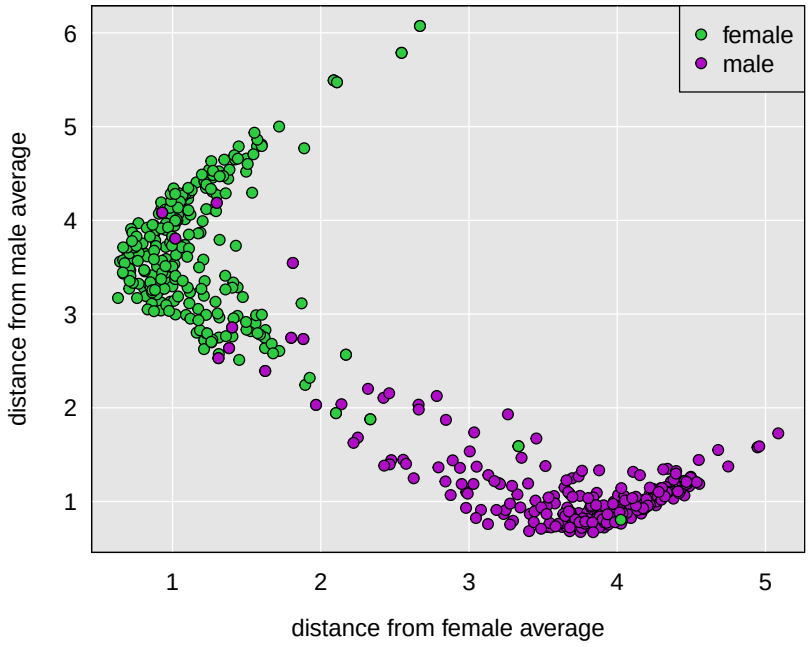


We should be able to do the same sort of thing, to verify the sex of the mRNA samples.

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

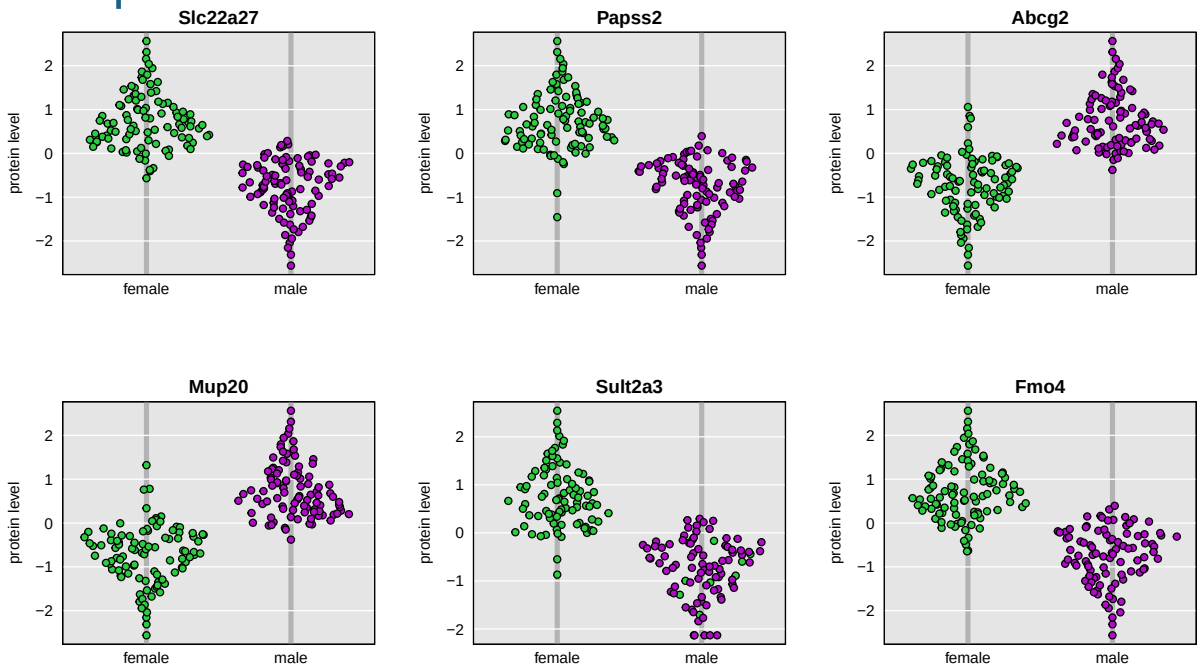


What I ended up doing was just pick the top 100 genes and then calculate the mean in males and females, and then for each sample, look at the RMS difference from the male means and from the female means. This is a plot of those.

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?

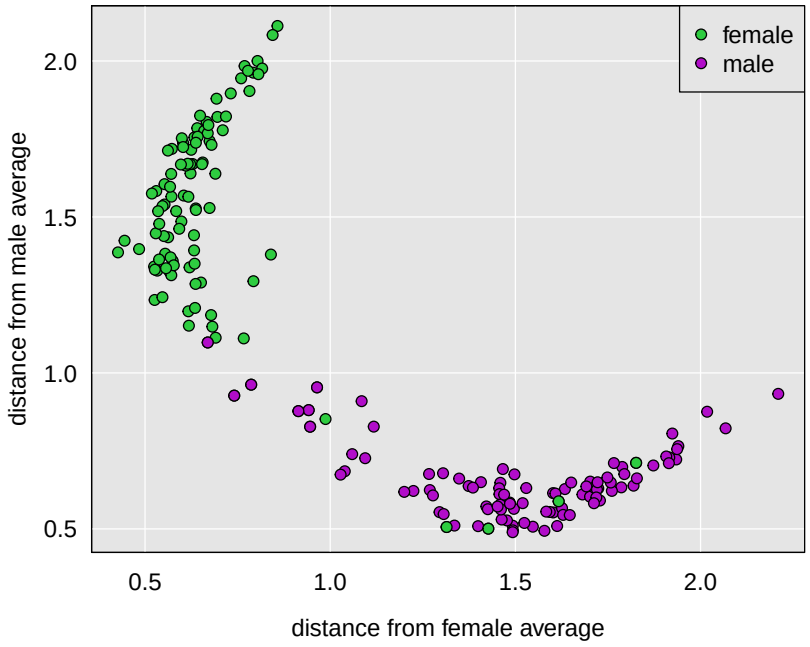
Anyway, while there are a number of samples in the middle that are unclear, there's also a set of males that look like females and one female that looks like male.

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.

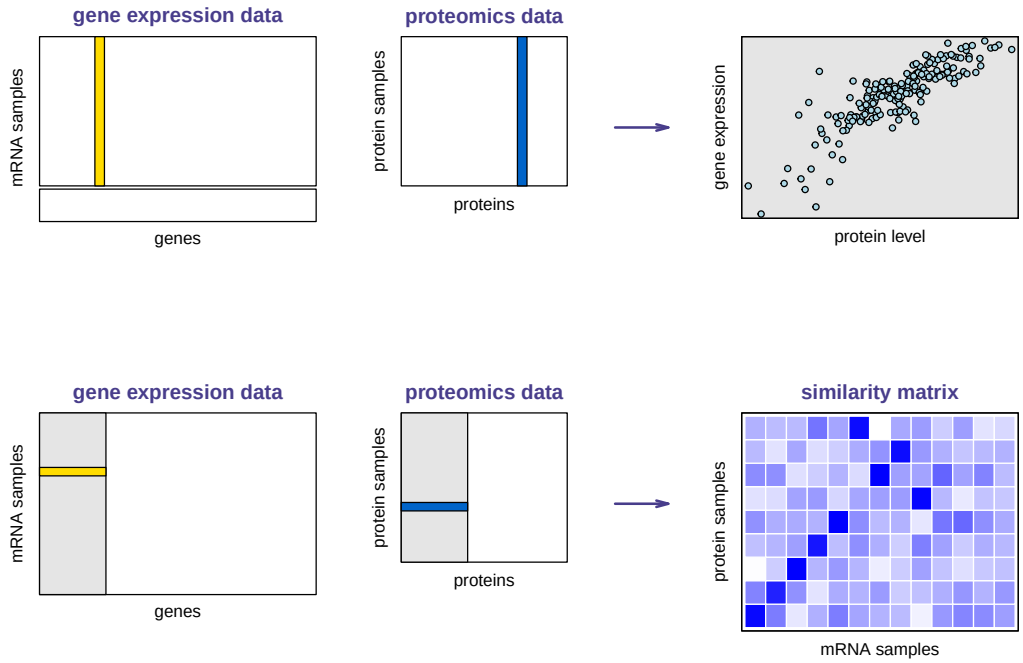
For this particular data set, there are a few females deep within the males.

Sample mix-ups

mRNA $\leftrightarrow$ protein

We now turn to the question of sample mix-ups more directly. We'll start by comparing a set of mRNA expression data to a set of proteomics data.

mRNA ↔ protein method

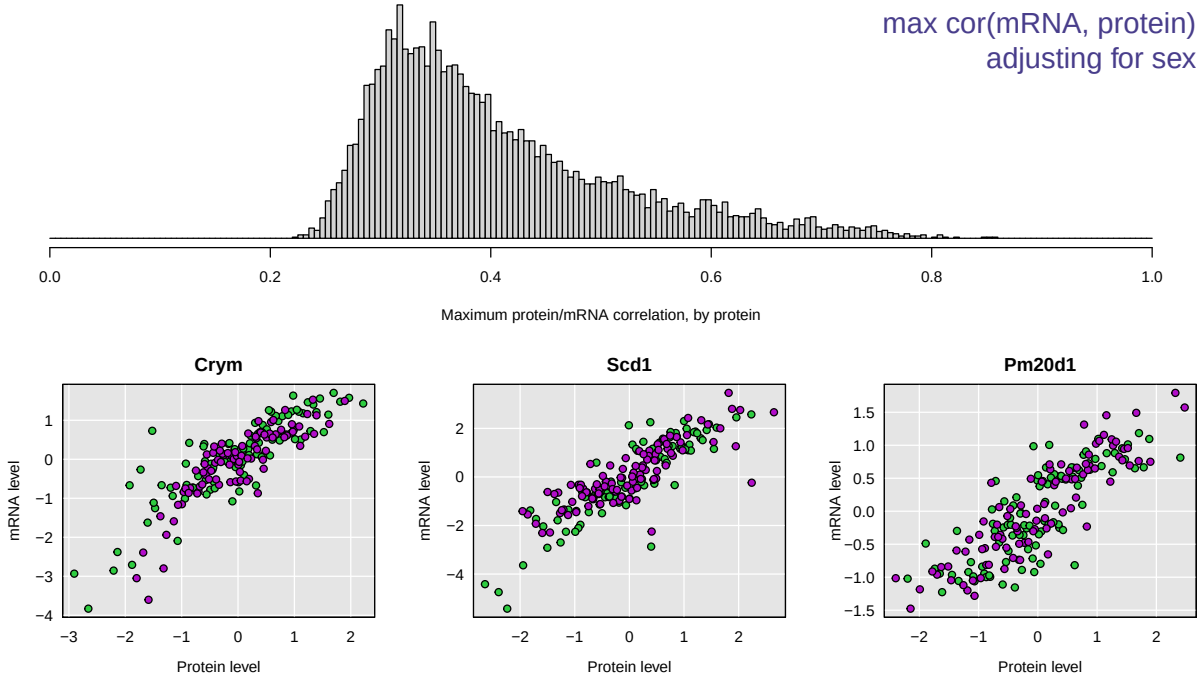


We have two rectangles of data where some of the rows are supposed to correspond.

We first look for gene/protein pairs that are highly correlated. We then focus on those and calculate the correlation between rows, among the highly correlated gene/protein pairs.

The between-sample correlations can be viewed as a similarity matrix. We're hoping to see a single large value in each row, for the samples that are supposed to correspond.

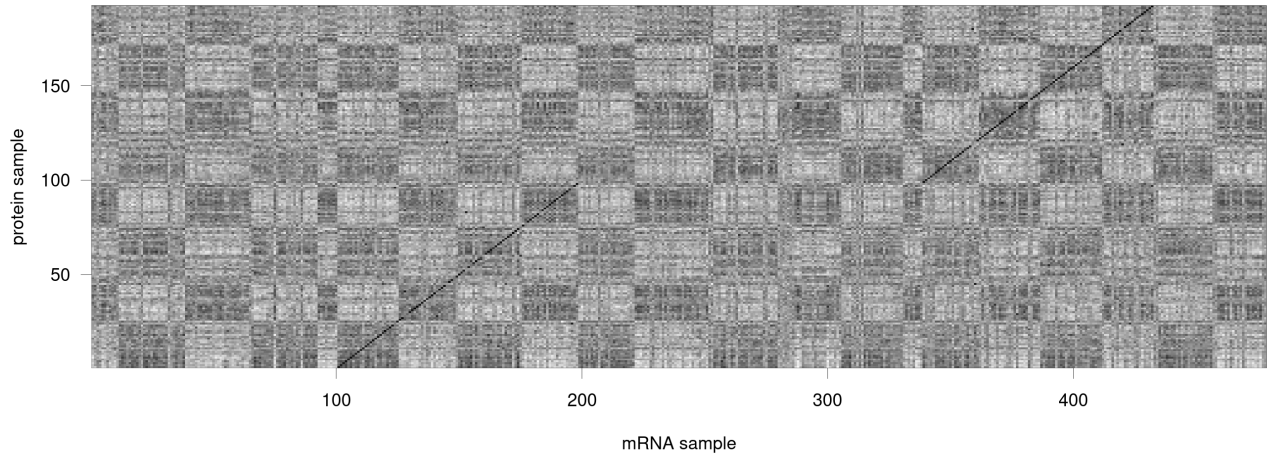
mRNA ↔ protein correlations



For a set of gene expression data and proteomics data, I considered each protein and found the most-correlated gene, adjusting for sex. (Whether to adjust for sex is not totally clear to me, but it makes the subsequent figures easier to understand.)

I show the top protein/gene pair (which all correspond to an mRNA and its protein product).

mRNA ↔ protein similarity matrix

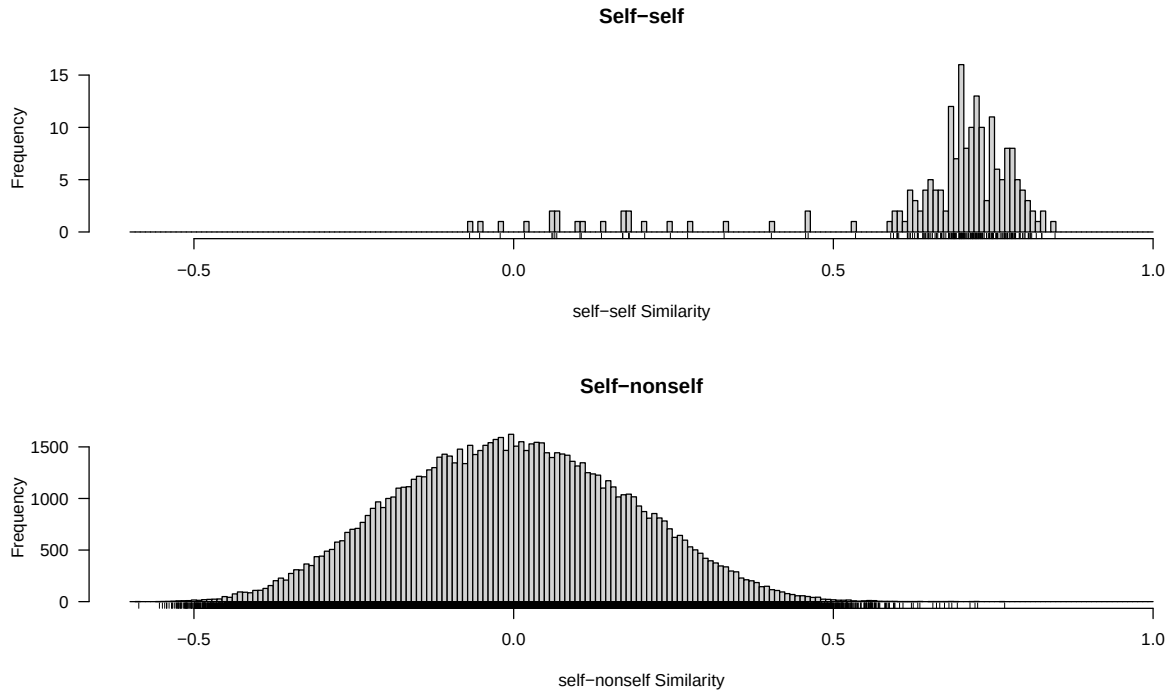


We take the top 100 gene/protein pairs and use them to calculate the correlation between samples. This is that correlation matrix. There are not quite 200 protein samples and a bit over 400 mRNA samples.

Black means more similar, and white less similar. The checkboard pattern is probably some sort of batch effects. If I hand't controlled for sex, there would be two big squares here, for the two sexes.

You can see a couple of diagonal lines, for the samples that correspond.

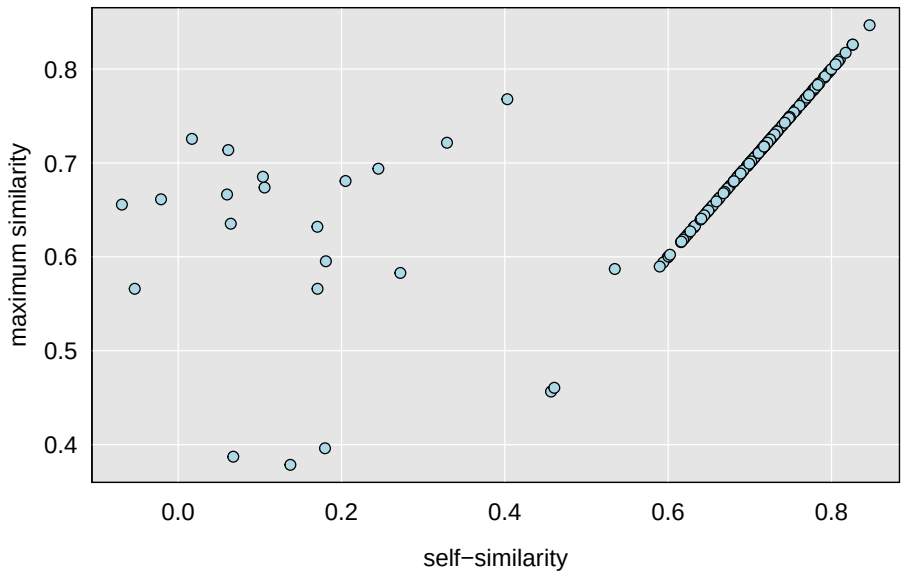
mRNA $\leftrightarrow$ protein similarities



If we separate the values for samples that are supposed to correspond and the values for samples that are not supposed to, we ideally would have two non-overlapping distributions.

The bulk of the samples are like that: large values for the self-self similarities, and small values for the self-nonself similarities. But there are a bunch of self-self values that are small, and a few self-nonself values that are large.

mRNA $\leftrightarrow$ protein: closest vs self

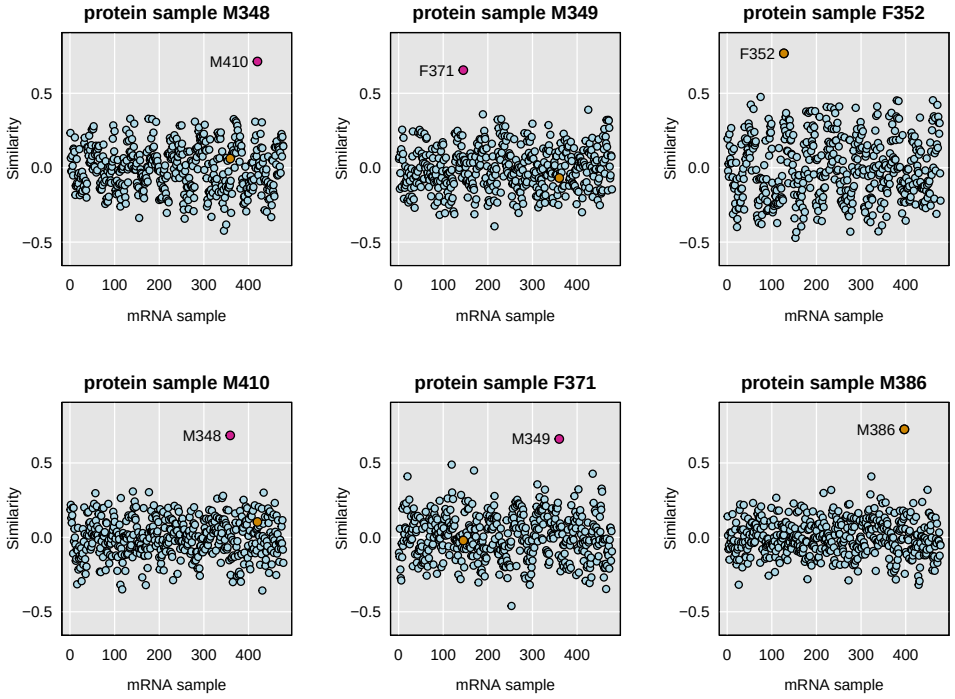


Here's a scatterplot where for each protein sample I find the maximum similarity in that row and plot it against the similarity for the mRNA sample that is supposed to correspond.

For most samples, these two values are the same, and they are large. (That's the diagonal line in the top-right.)

But there are a bunch of samples where the self-similarity is small, indicating that it seems to have the wrong label. Many of these are largely similar to some other sample, but a few are not similar to anything.

mRNA ↔ protein: selected samples



Here I have focused on a selected set of six protein samples, and just plot all of the values in that row of the similarity matrix. For each of these protein samples, I highlight the closest mRNA sample.

The pair on the left looks like a sample swap: M348 is closest to M410 and vice versa. Similarly, the pair in the middle looks like a sample swap, with F371 being closest to M349 and vice versa. The two samples on the right look to be correctly labeled.

Overall, there were about 9 sample swaps and then a few other protein samples that didn't seem to correspond to any mRNA sample.

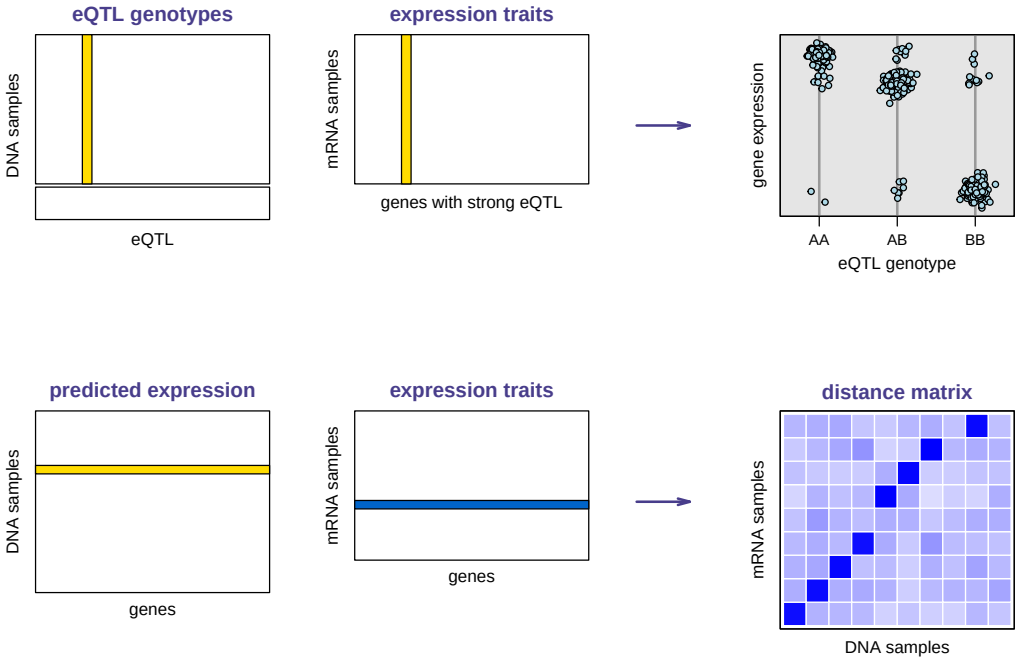
With these sample swaps, we can't tell, from these data alone, whether the problem is with the mRNA samples, the protein samples, or some of each.

Sample mix-ups

DNA $\leftrightarrow$ mRNA

Let's now turn to the same sort of thing, but comparing DNA-based genotypes to mRNA expression.

DNA ↔ mRNA method

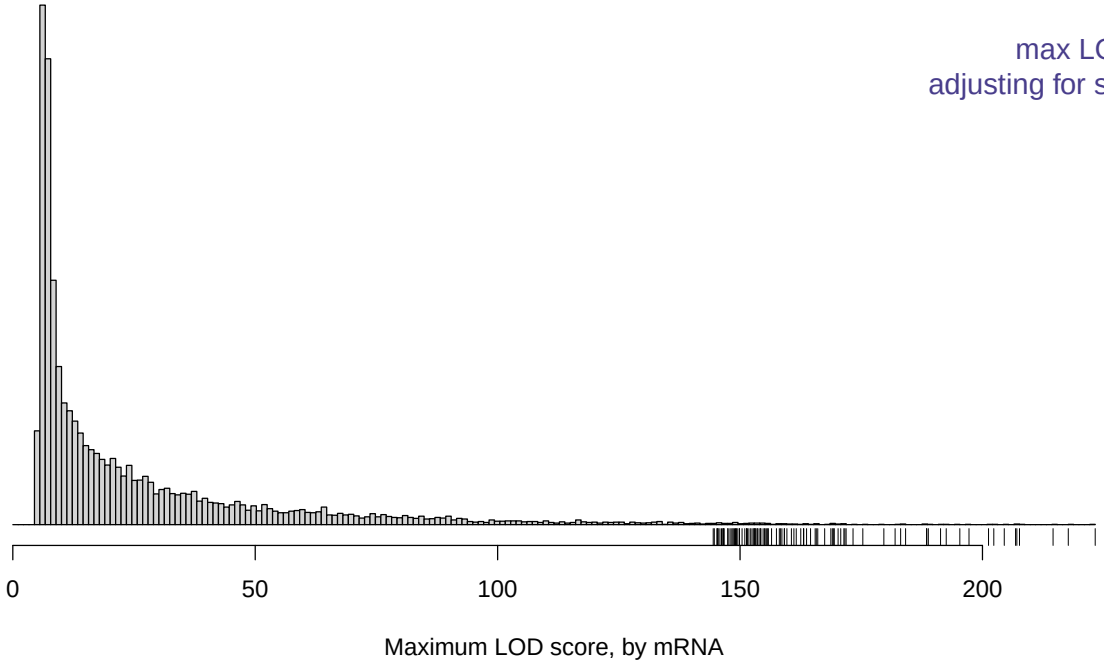


The approach we take starts similarly: we identify a set of expression traits with strong eQTL, and then pull out the genotypes at those eQTL.

The simplest thing is to use the genotype/expression correspondence to get predected expression values for each trait, based on the observed eQTL genotypes. Basically just calculate the average trait value for each genotype and assign that as the fitted value.

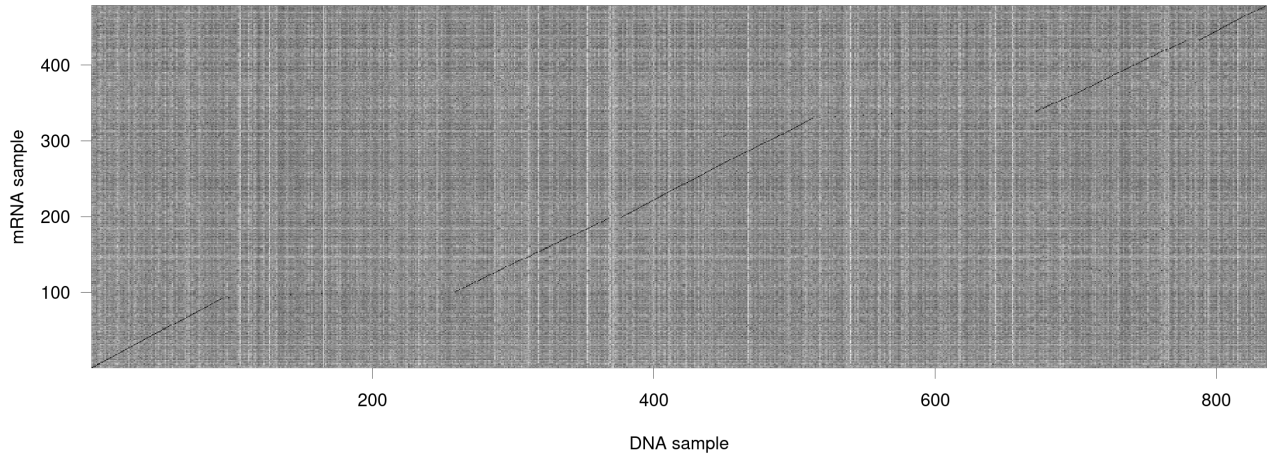
We can then take the root-mean-square (RMS) difference between the predicted expression values for a DNA sample and the observed expression values for an mRNA sample, and used that as a distance matrix, comparing the DNA and mRNA samples.

DNA $\leftrightarrow$ mRNA LOD scores



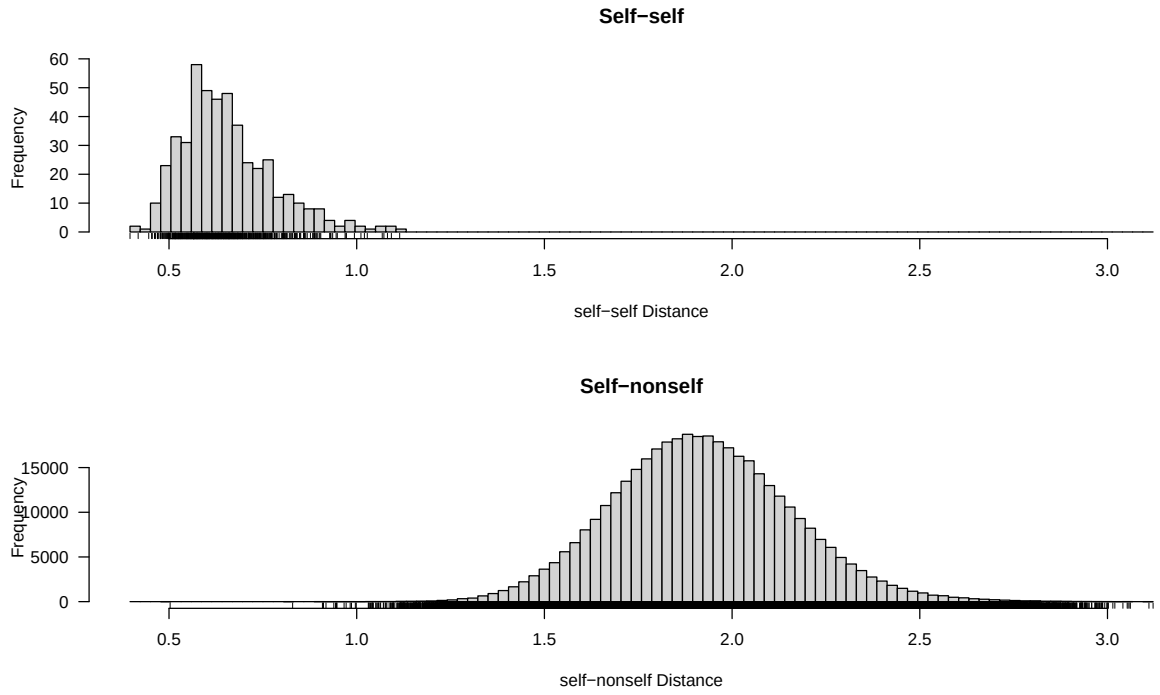
For a set of RNA-seq based expression data, here are the LOD scores. For each expression trait, I did a genome scan adjusting for the sex and picked out the single genome-wide maximum value. There are 100 genes with LOD about 150 and higher.

DNA $\leftrightarrow$ mRNA distance matrix



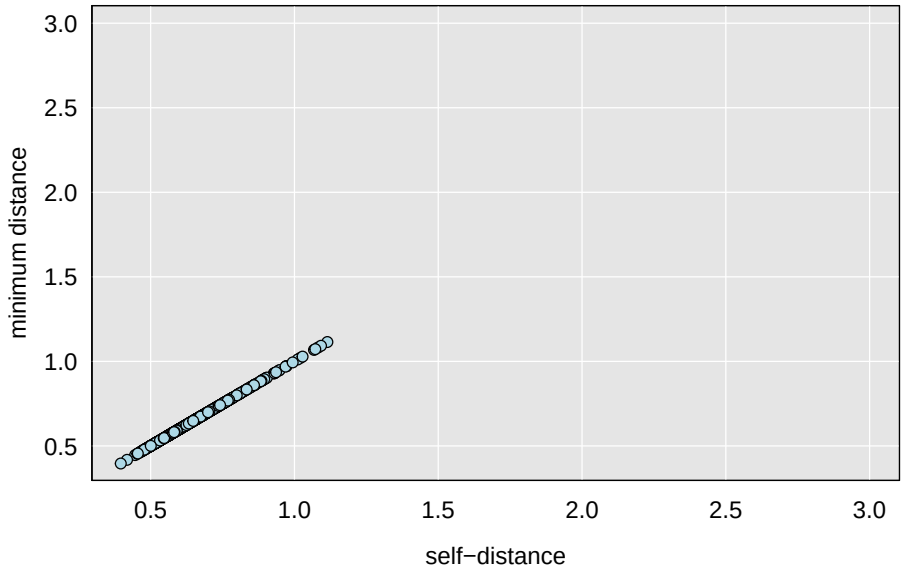
Using those 100 genes, I calculated predicted expression values for each DNA sample, and then calculated this distance matrix. Black again means close. You can see a few black diagonal lines in here. There are about 450 mRNA samples and a bit over 800 DNA samples.

DNA $\leftrightarrow$ mRNA distances



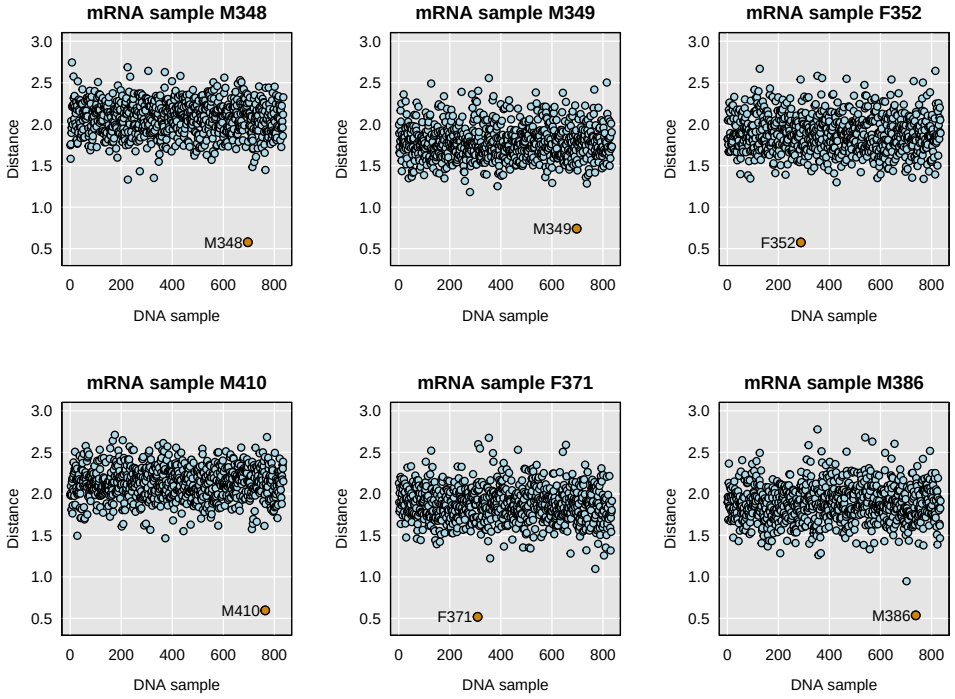
Here, if I pull apart the self-self and self-nonself distances, I find that there's almost no overlap.

DNA $\leftrightarrow$ mRNA: closest vs self



If I calculate the minimum distance in each row and plot it against the self-self distance, I found that each mRNA sample is actually closest to itself than to any other sample. There don't seem to be any problems here.

DNA ↔ mRNA: selected samples



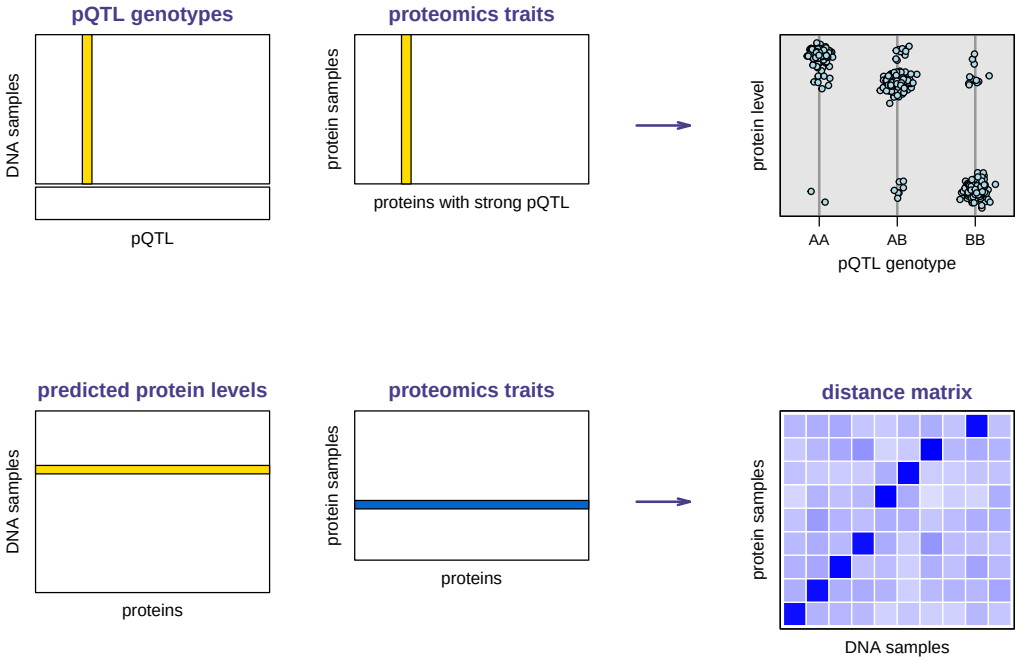
Here are those six selected samples again. In each case, the samples look to have the correct labels.

Sample mix-ups

DNA $\leftrightarrow$ protein

Let's turn to a comparison between DNA-based genotypes and proteomics data.

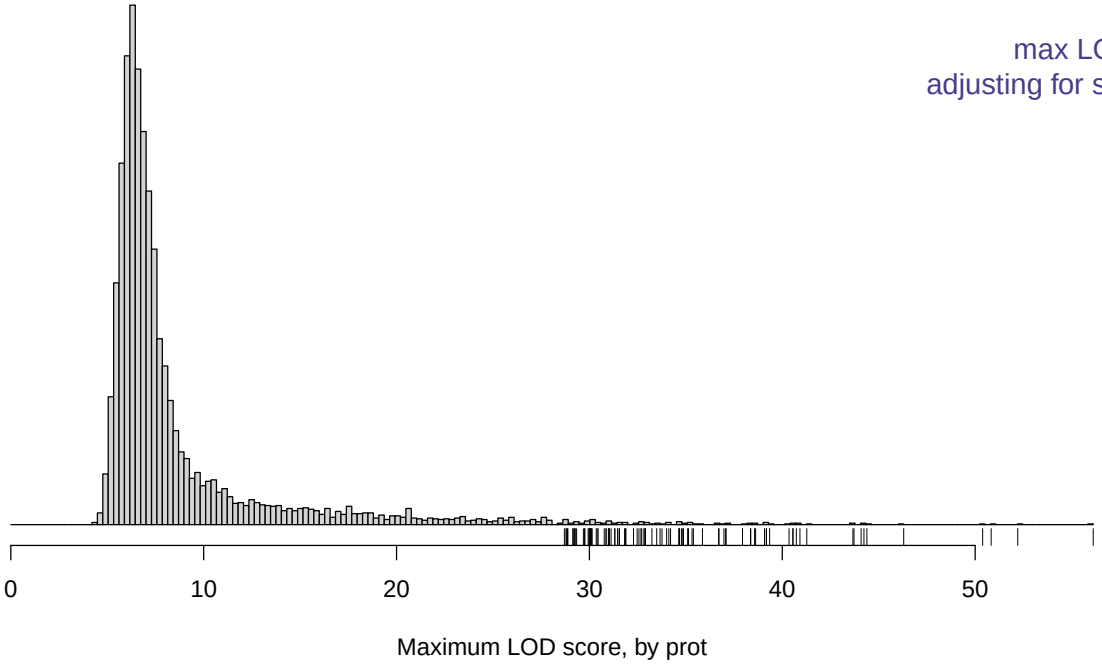
DNA $\leftrightarrow$ protein method



The approach is the same as for the comparison of DNA and mRNA. For each protein, I do a genome scan and look for the strongest pQTL, and I pull out the proteins that have strong pQTL.

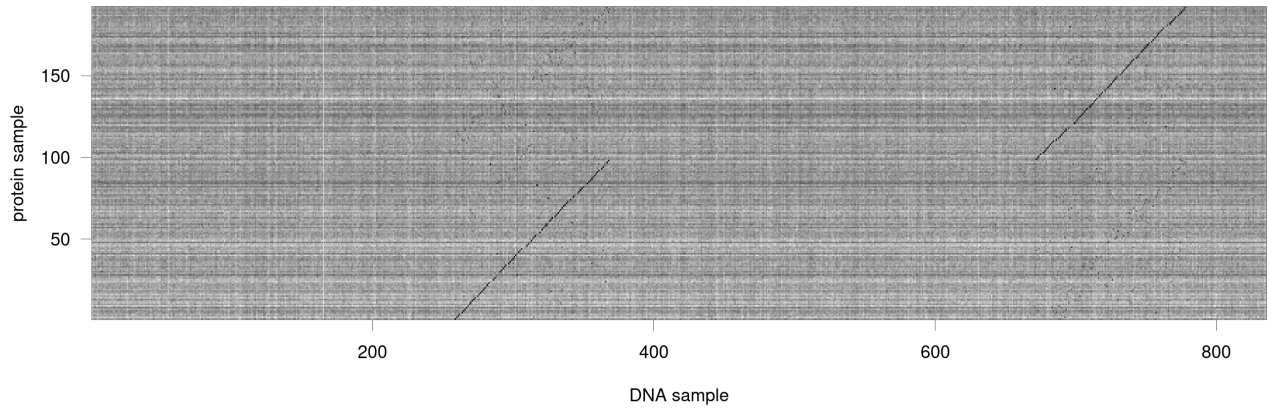
I then use the association between genotype and protein level to get predicted protein levels for each DNA sample. I then calculate the RMS difference between the predicted protein levels for a DNA sample and observed protein levels for some proteomics sample, and use this as a distance matrix.

DNA $\leftrightarrow$ protein correlations



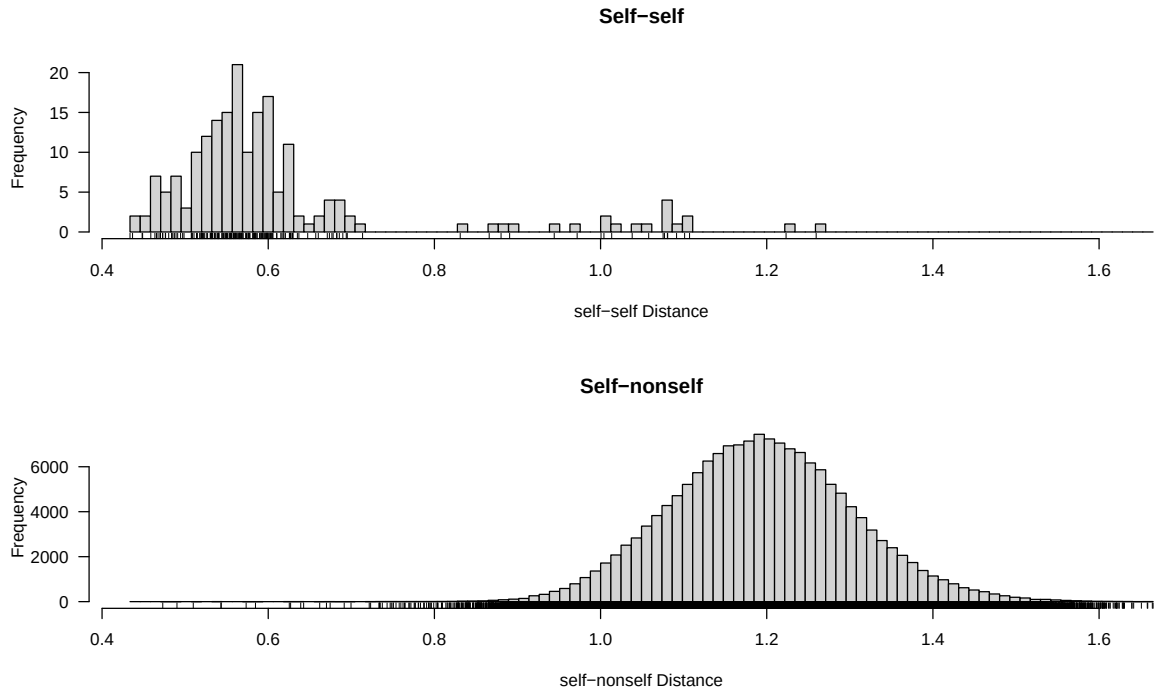
There are many proteins with very strong pQTL. I'll focus on the top 100 proteins, with LOD scores about 30 and up.

DNA $\leftrightarrow$ protein distance matrix



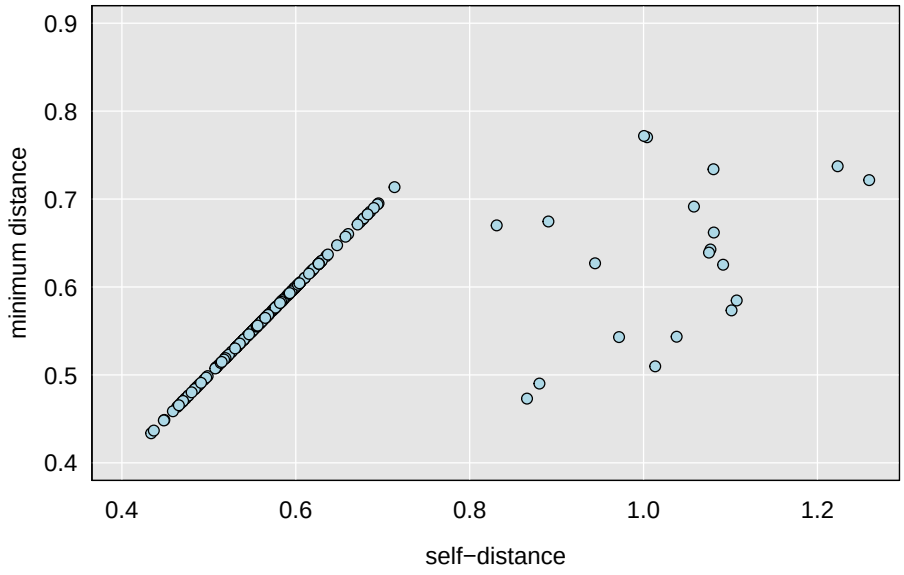
Here's the distance matrix comparing just under 200 proteomics samples to just over 800 DNA samples. Black means close; you can see two prominent diagonal lines.

DNA $\leftrightarrow$ protein distances



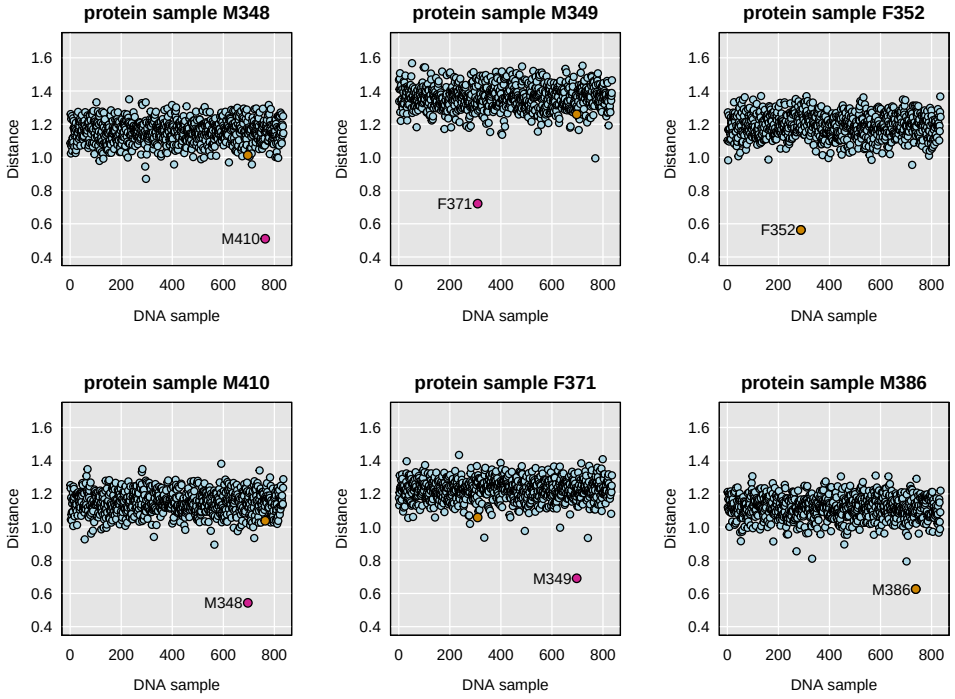
If we separate the self-self distances from the self-nonself ones, we again see a bunch of self-self distances that are overly large, and self-nonself distances that are too small.

DNA $\leftrightarrow$ protein: closest vs self



Taking the minimum distance in each row and plotting that against the distance for the sample with the same label, we see that most samples look to be correctly labeled, but there's a big group that are not close to the corresponding sample but are close to some other sample.

DNA $\leftrightarrow$ protein: selected samples



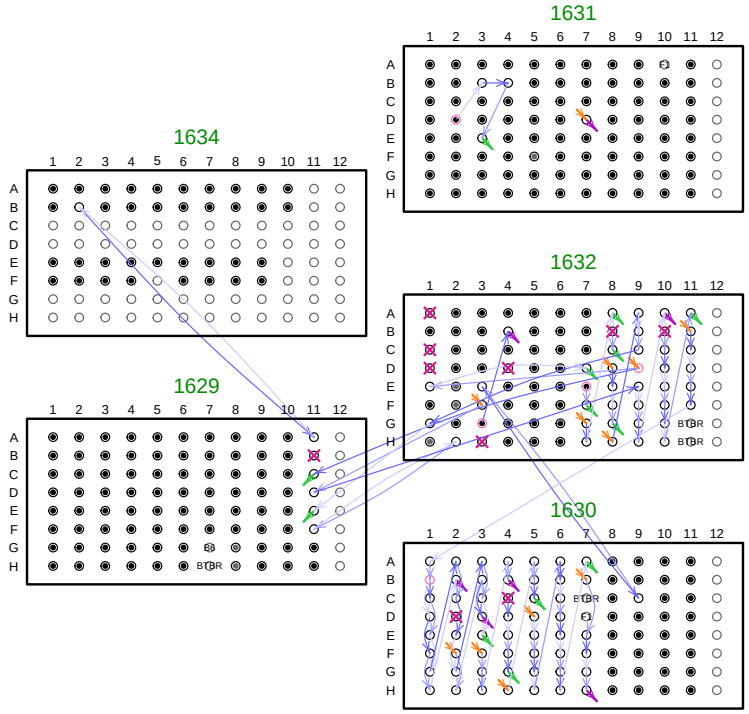
Here are those six selected samples again. A couple of sample swaps and a pair that are correct.

Combining what we've learned from the DNA/mRNA/protein, we can conclude that the mislabeling problems were in the protein samples.

Overall, we identified 9 sample swaps and two additional mislabeled samples. There were 192 total samples, so this is just above 10% problems.

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.



In summary: sample mix-ups shouldn't happen; we should do our best to avoid them. But if the do happen, we should find them.

It's not too hard to look for sample mix-ups. Pull out pairs of highly correlated variables and then use them to establish the similarity or distance between samples.

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

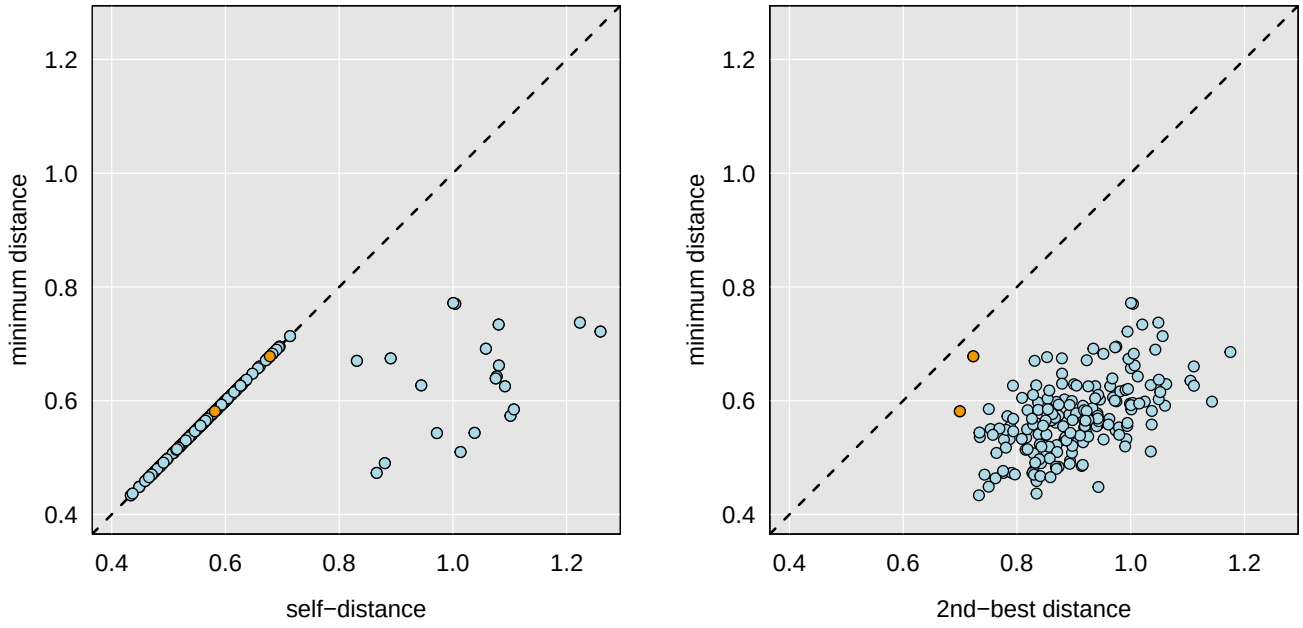
kbroman.org

github.com/kbroman

@kwbroman

Here is where you can find me and my slides.

DNA ↔ protein: best vs 2nd-best



Here's an extra slide showing (on the right) the best distance vs the 2nd-best distance. The panel on the left we'd seen before, of the best distance vs the minimum distance.

This is useful for assessing the evidence that the sample with the best distance is the real sample identity. If the best distance is quite far from the second-best distance, than it seems like we can identify the true sample. If they're similar, we don't have much evidence to relabel the sample.

I've highlighted a couple of points in orange, where the second-best distance is quite close to the best distance. But for both of these, the best distance is the same as the self distance, and so there's no reason to question their identity.