Sample mix-ups in eQTL data

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

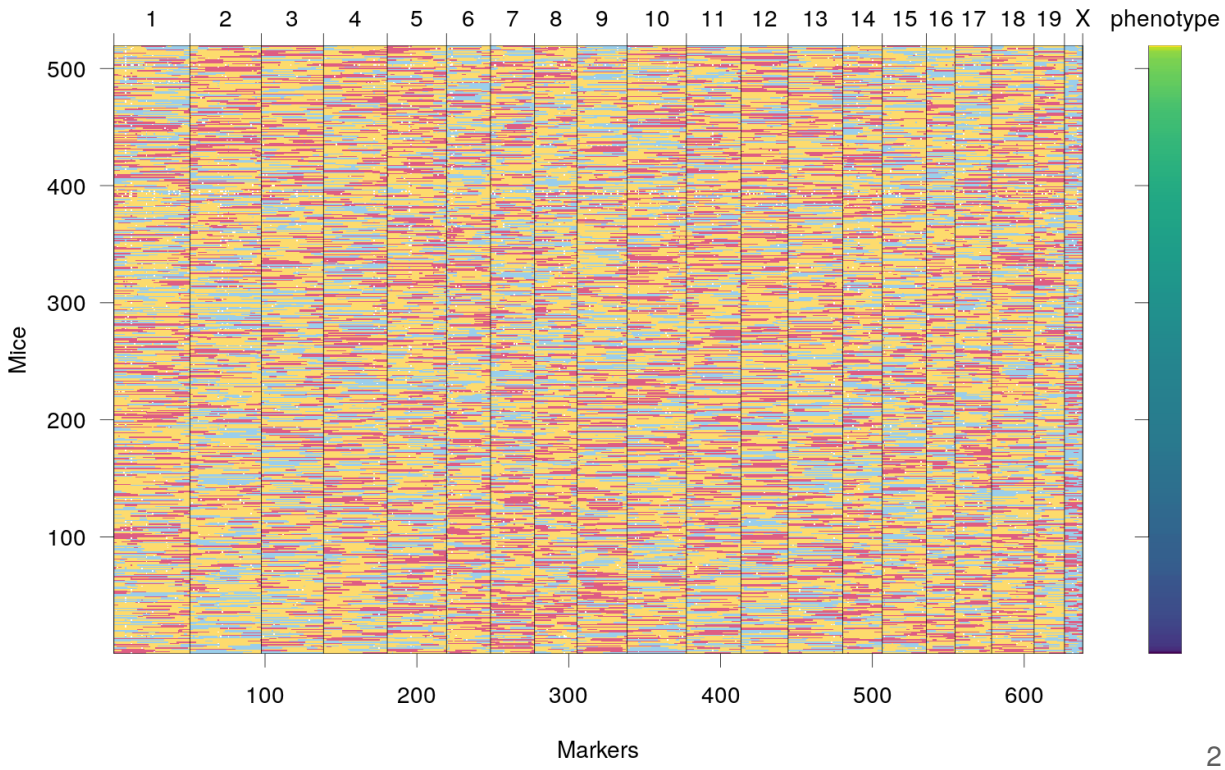
kbroman.org
github.com/kbroman
@kwbroman
Course web: kbroman.org/AdvData

In this case study, I'll talk about a QTL mapping experiment where I discovered that like 18% of the genotyped samples were mixed up.

A weakness of QTL mapping has been the poor precision in estimated QTL location; it's very hard to identify the underlying genes. One strategy to deal with this weakness is to also measure intermediate phenotypes, such as the mRNA expression of all genes in a relevant tissue. We then seek to identify genetic loci (called expression quantitative trait loci, eQTL) that affect mRNA expression, and to find genes for which genotype is associated with mRNA expression and also the clinical trait.

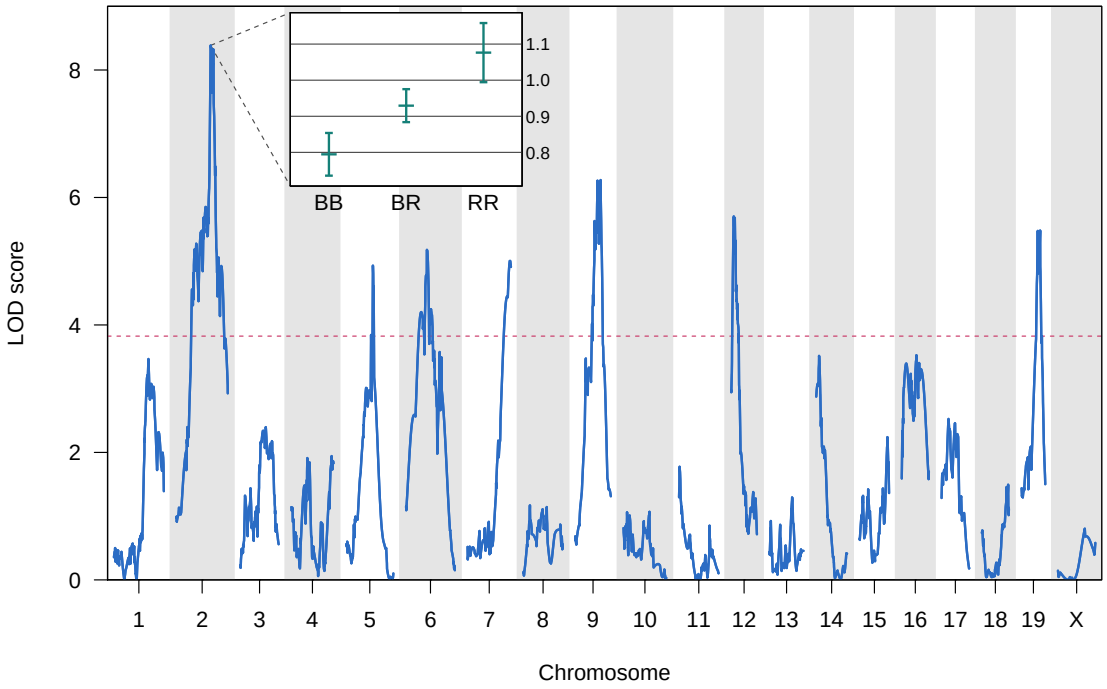
In a recent study with 500 intercross mice and gene expression microarray data on six tissues, I identified a large number of sample mix-ups in the genotype data and a smaller number of mix-ups in each set of microarrays. I'll describe how I found and corrected these problems. In a nutshell: the expression of some genes is so strongly associated with genotype that the expression data can effectively serve as a DNA fingerprint for establishing individuals' identities.

Data



Again, I'm talking about QTL mapping. The data consist of genotypes at a set of markers across the genome, plus some quantitative phenotype for each mouse. The goal is to identify regions of the genome where the genotype is associated with the phenotype.

QTL mapping



Our goal is to identify quantitative trait loci (QTL): regions of the genome for which genotype is associated with the phenotype.

The basic analysis is to consider each locus, one at a time, split the mice into the three genotype groups, and perform analysis of variance.

We then plot a test statistic that indicates the strength of the genotype-phenotype association. For historical reasons, we calculate a LOD score as the test statistic: the $\log_{10}$ likelihood ratio comparing the hypothesis that there's a QTL at that position to the null hypothesis of no QTL anywhere.

Large LOD scores indicate evidence for QTL and correspond to there being a difference in the phenotype average for the three genotype groups.

Attie project

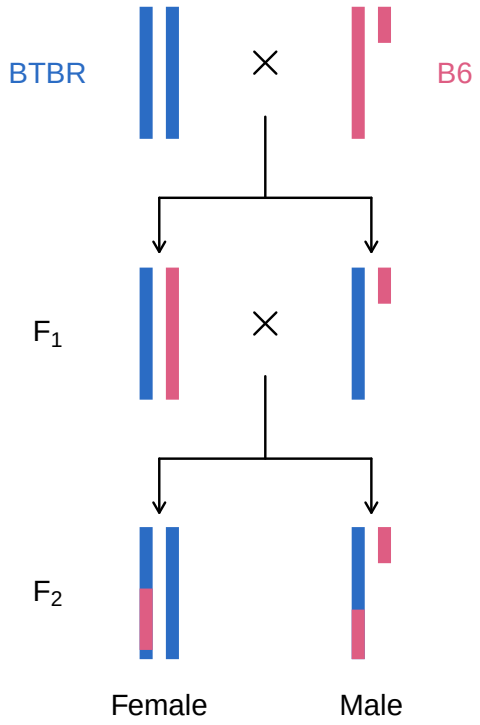
~500 B6 × BTBR intercross mice, all ob/ob

- 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)

When I first got to UW-Madison, I joined a collaboration that was carrying out a very large QTL mapping experiment that included about 500 mice with dense genotype data and numerous clinical phenotypes, but also with gene expression data in six different tissues.

I had mostly been in the back of the room, heckling. But a couple of years into the project, I agreed to write the first paper.

Sex and the X chr



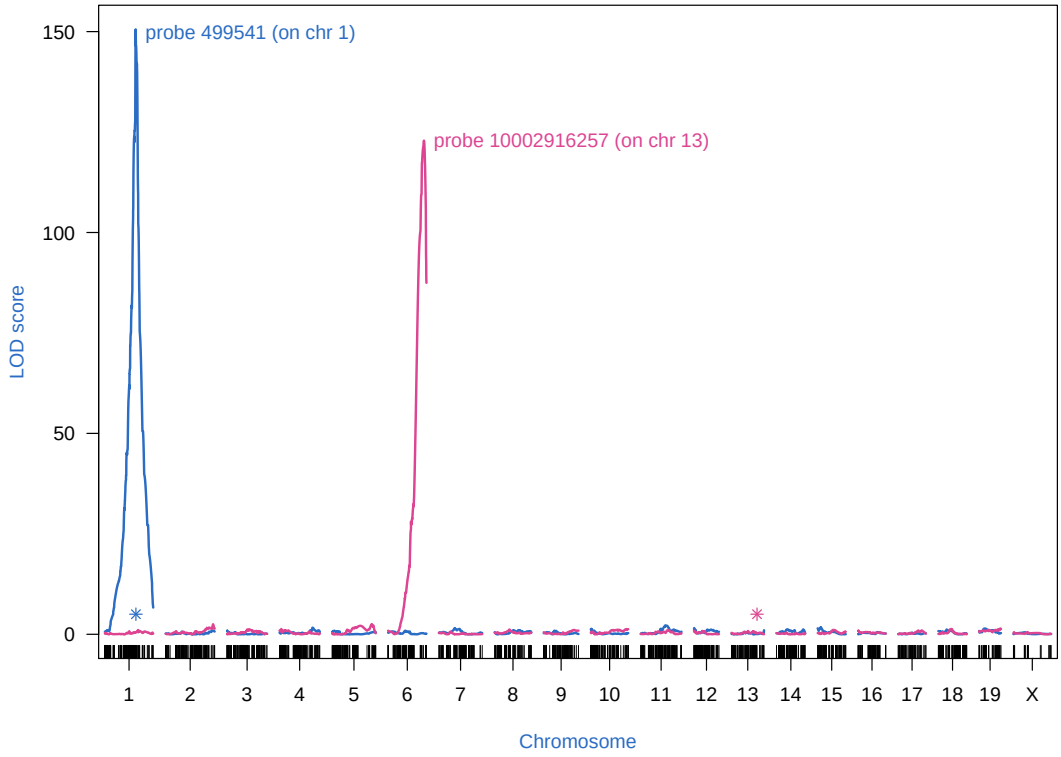
In getting ready to prepare that first paper, I decided to go back to the basics and really check that all of the data were in good order, starting from the raw genotype files.

I noticed that there were a number of mice whose X chromosome genotype data did not match their sex. The way the cross was carried out, female F_2 mice will be homozygous BTBR or heterozygous, and male F_2 mice will be hemizygous (and so look like homyzogous). But there were a number of females who were homozygous B6 on the X, and a number males who were heterozygous. (Previously, these incompatible genotypes had just been omitted.)

The number of mice with this problem ($\sim$16 out of 500) was not large, but it was more than I'd expected, and I sat and pondered how to figure out which was correct: sex or genotype.

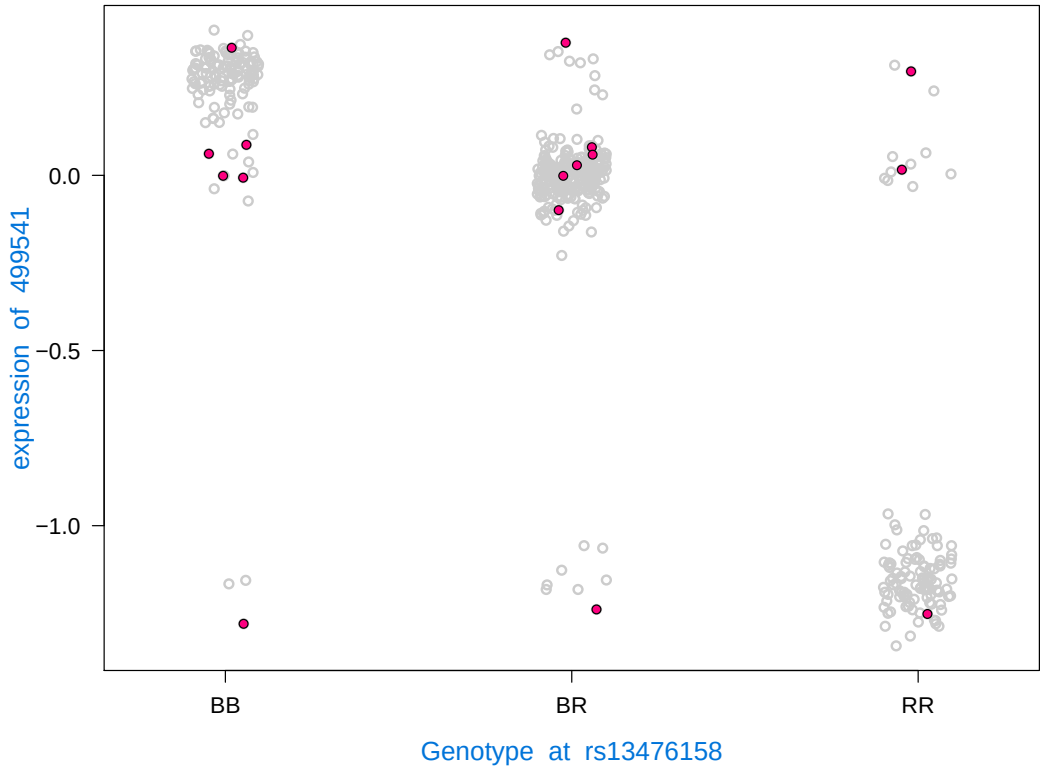
I realized that I could maybe use the gene expression data to help.

Strong eQTL



In many cases the gene expression traits have very strong genetic effects. In particular, for many genes the expression level is strongly affected by genotype right at the location of the gene. For other genes, expression is strongly affected by genotype at some other location. A locus that effects gene expression is called an expression QTL or eQTL.

E vs G



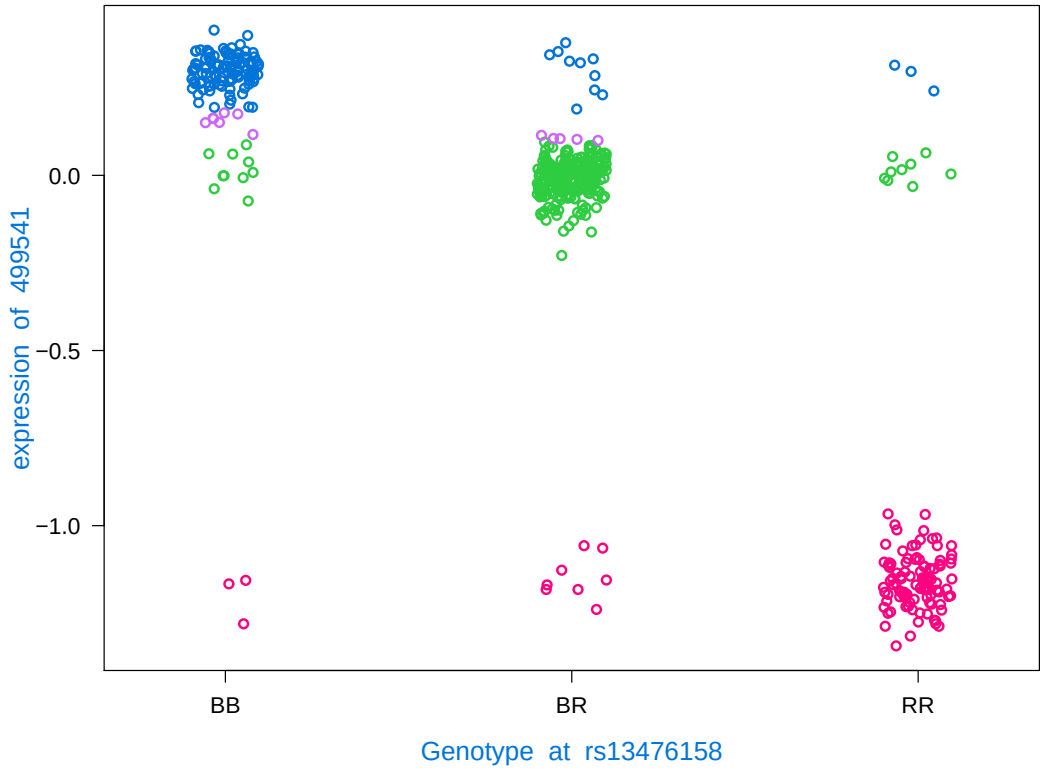
I looked at the gene expression versus genotype at one of these eQTL and saw a very strange pattern. There was a very strong association, but there were also a lot of mice whose gene expression seemed to not match their genotype.

I mean, there are basically three kinds of mice, expression-wise: low, high, or very high. And the low-expression mice are mostly RR, while the very-high mice are mostly BB, with the high-expression mice being BR. Except there are a bunch of mice that seem to be in the wrong ball, expression-wise. And the 16 six-swapped mice include 9 that are in the wrong ball.

It's like the sex-swapped mice had been assigned to a random genotype. If the genotypes are in the proportions 1:2:1, then we'd expected 3/8 to be correct just by chance, which is very similar to the 7/16 we see in these data.

And note that there are 43 mice that look to be in the wrong ball. If they are all being assigned genotypes at random, that would suggest that there are like $43 \times (8/3) \approx 115$ problem mice.

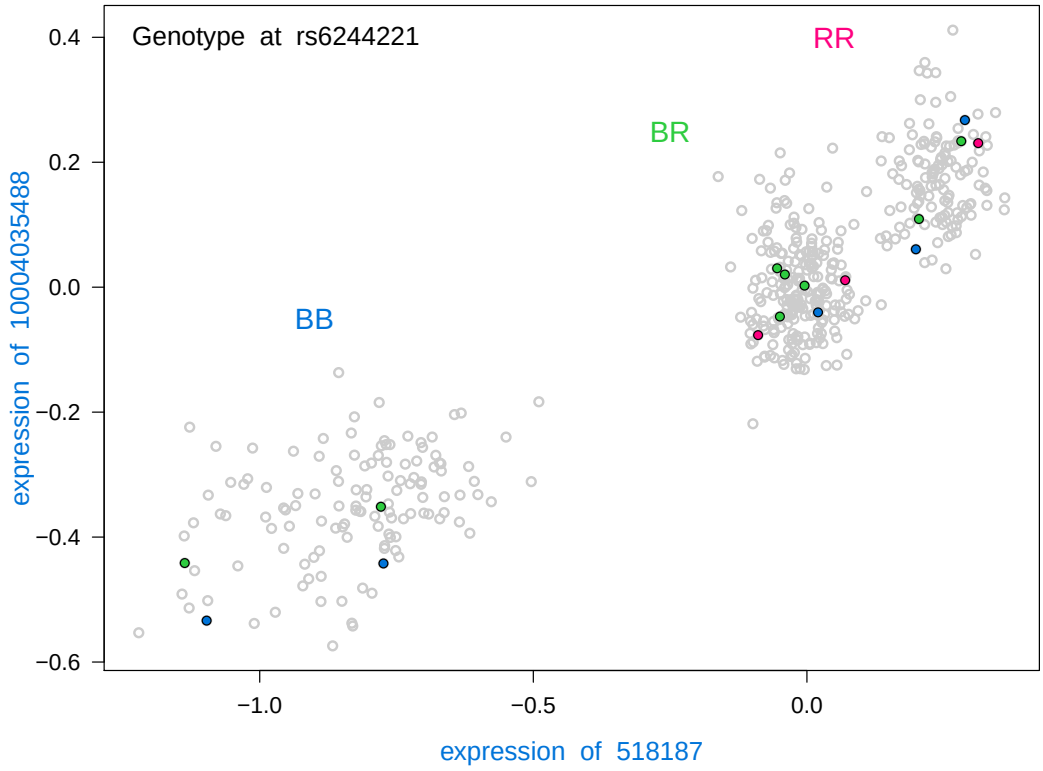
kNN classifier



But we can use the gene expression data to figure out what we think each mouse's genotype at this location really is. For example, we can create a k-nearest-neighbor classifier, for predicting genotype from gene expression.

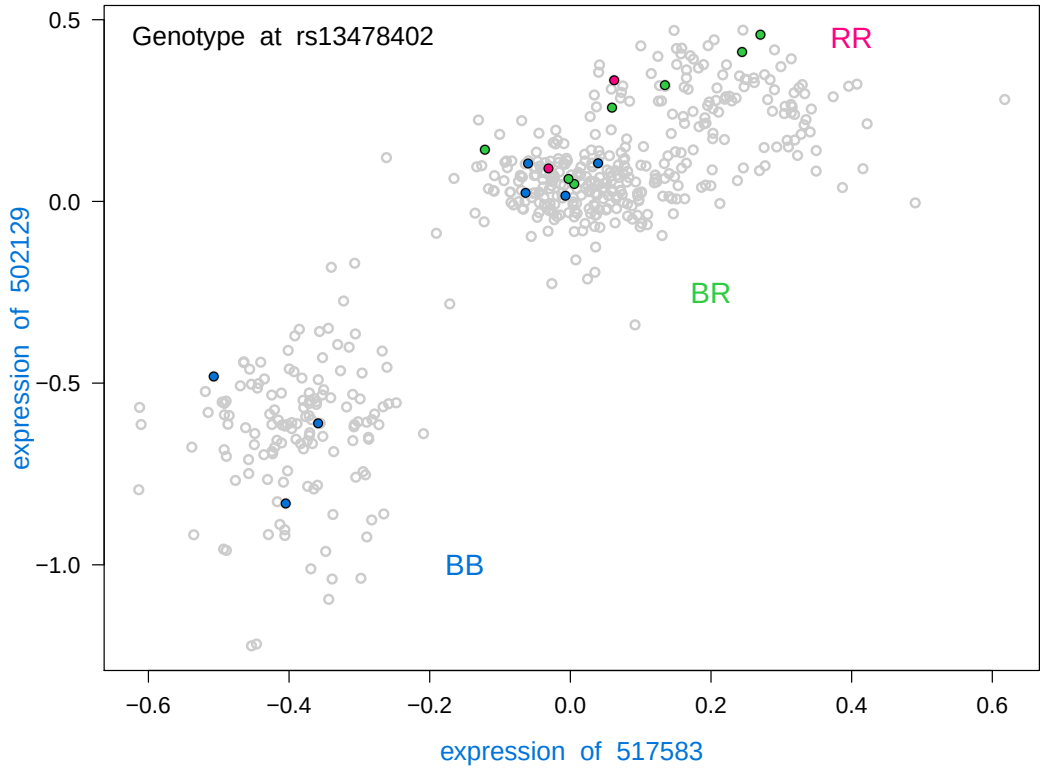
If we do this at many strong eQTL, we could potentially reconstruct the true genotypes for each mouse, from their expression data.

E vs G



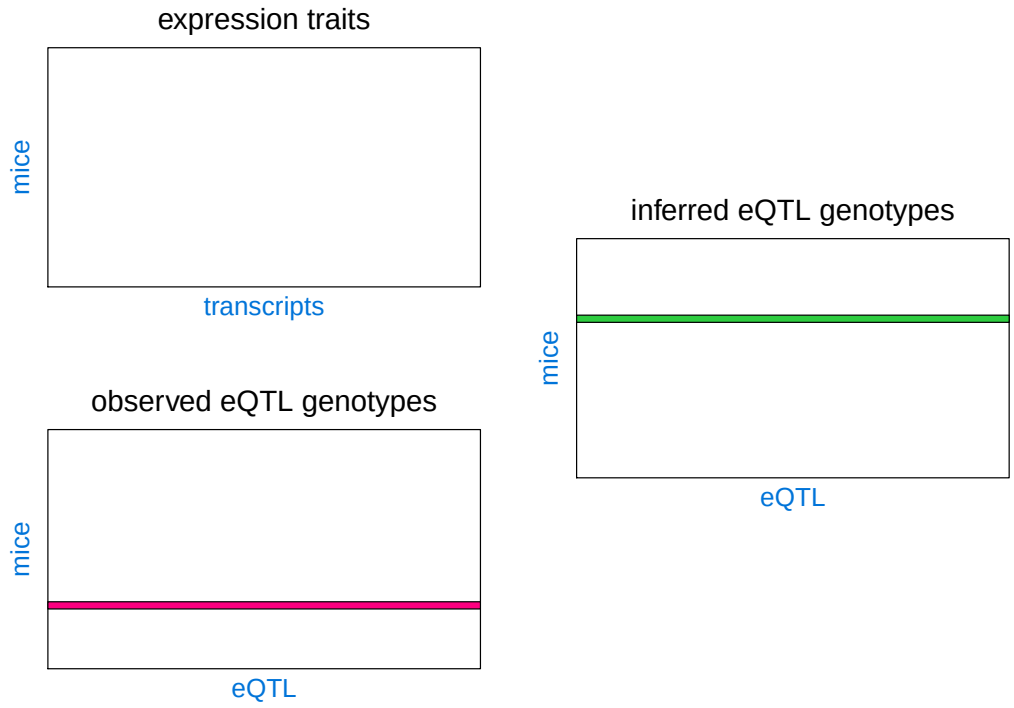
Many times there will be two different genes whose expression maps to a common location. We can look at their expression jointly. In many cases, the gene expression clusters are even more clear. And again the sex-swapped mice are seen in the wrong ball with frequency like 9/16.

E vs G



Here's one more case. Again, the sex-swapped mice are in the wrong ball with frequency like 9/16. The particular mice that are correct or not different among the eQTL.

Basic scheme

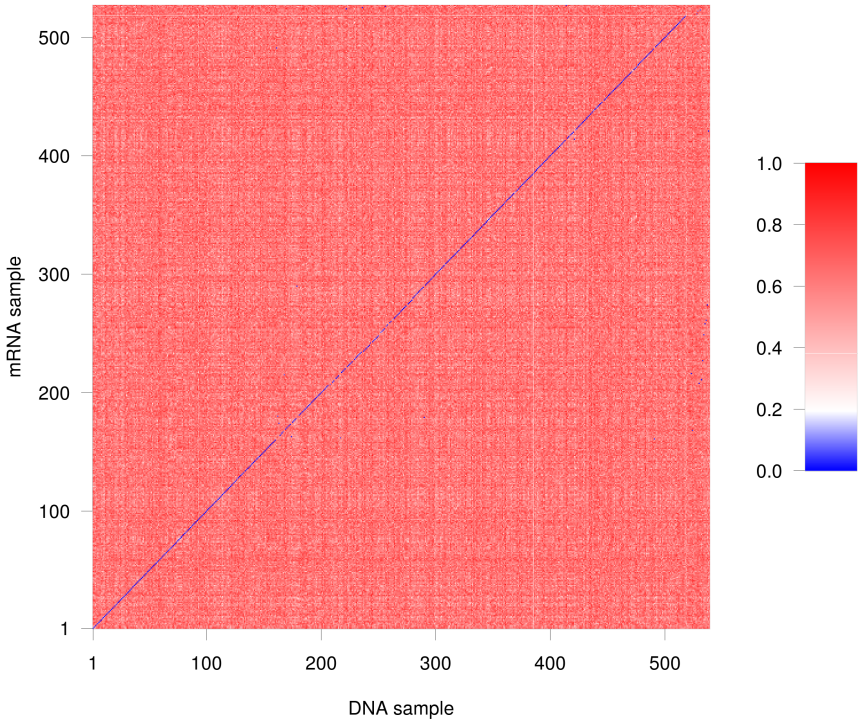


So this leads to our basic scheme for identifying (and correcting) the sample mix-ups.

We first identify a set of expression traits with very strong eQTL. We use the expression and corresponding eQTL genotypes to form classifiers for predicting eQTL genotype from gene expression. This gives us a matrix of inferred eQTL genotypes.

We then compare the inferred eQTL genotypes to the observed eQTL genotypes. If a sample's observed eQTL genotypes don't match its inferred eQTL genotype, we conclude that the labels for one or the other are incorrect. And we might be able to find another row in the inferred eQTL genotypes that matches its observed genotypes.

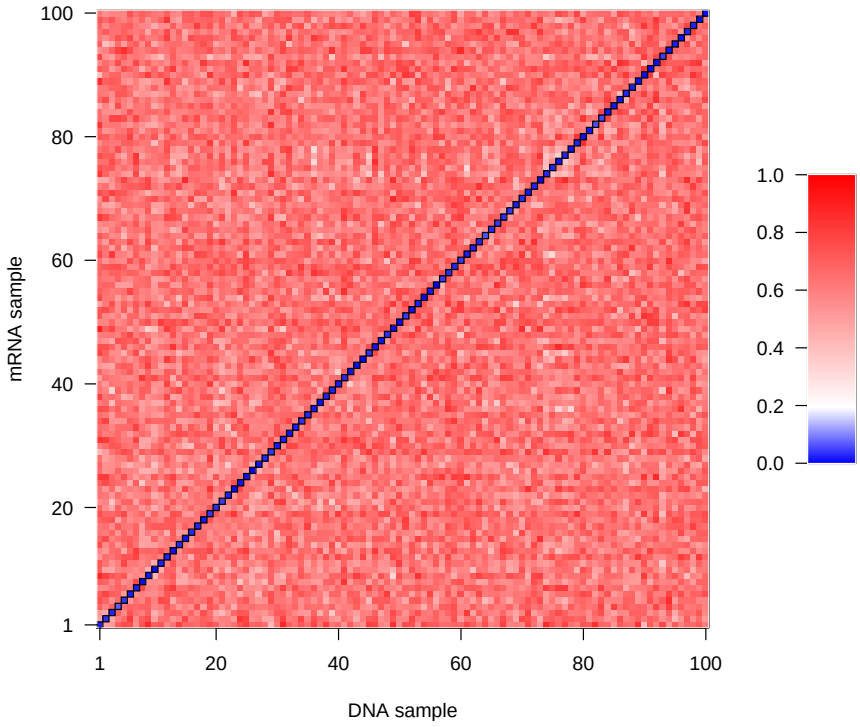
Prop'n mismatches



For each pair of samples, one DNA (genotype) sample and one RNA (gene expression) sample, we get a measure of distance as the proportion of mismatches between the observed eQTL genotypes and the inferred eQTL genotypes.

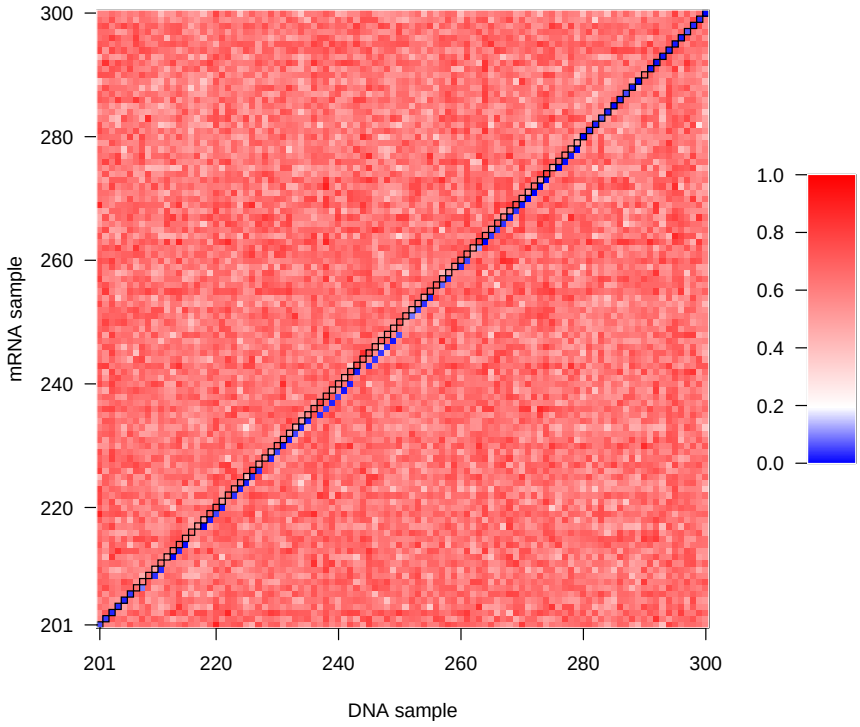
Here's a picture of this distance matrix. It should be blue along the diagonal and red everywhere else.

Prop'n mismatches



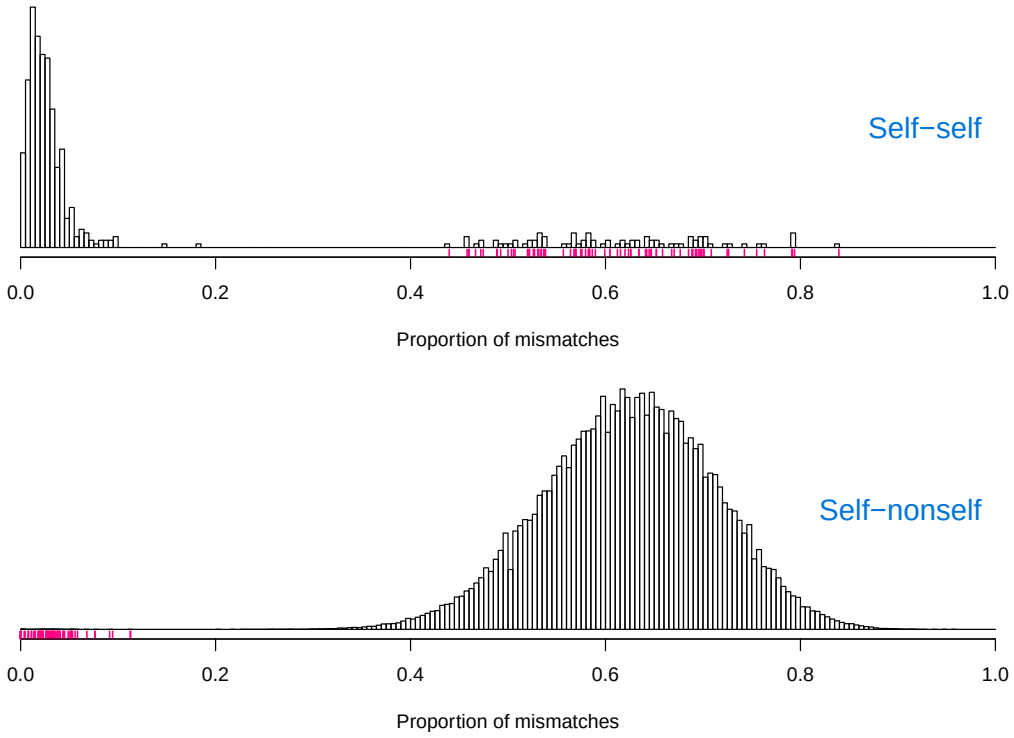
And if we look at the first 100 samples, that's exactly what we see: the samples are close to themselves and not to anyone else.

Prop'n mismatches



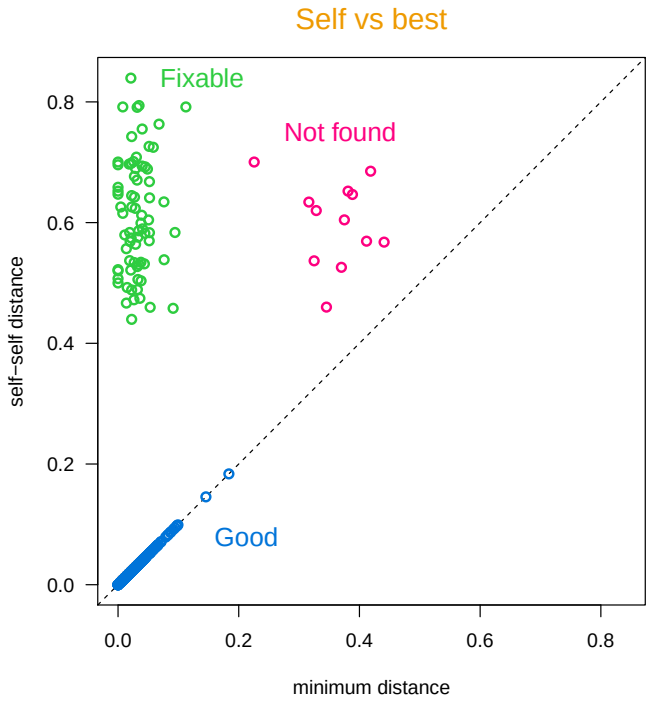
But if we look at the middle 100 samples, we find a whole bunch of off-by-one and off-by-two errors. The samples are quite different from the corresponding one, but their close to the one next to it or two over.

Prop'n mismatches



If we look at histograms of the diagonal of the distance matrix (top panel) and the off-diagonal values (lower panel), we find that most samples are correct, but there are a bunch of values on the diagonal that really are non-matching, and a bunch of values off the diagonal that are indicative of matches.

Decisions



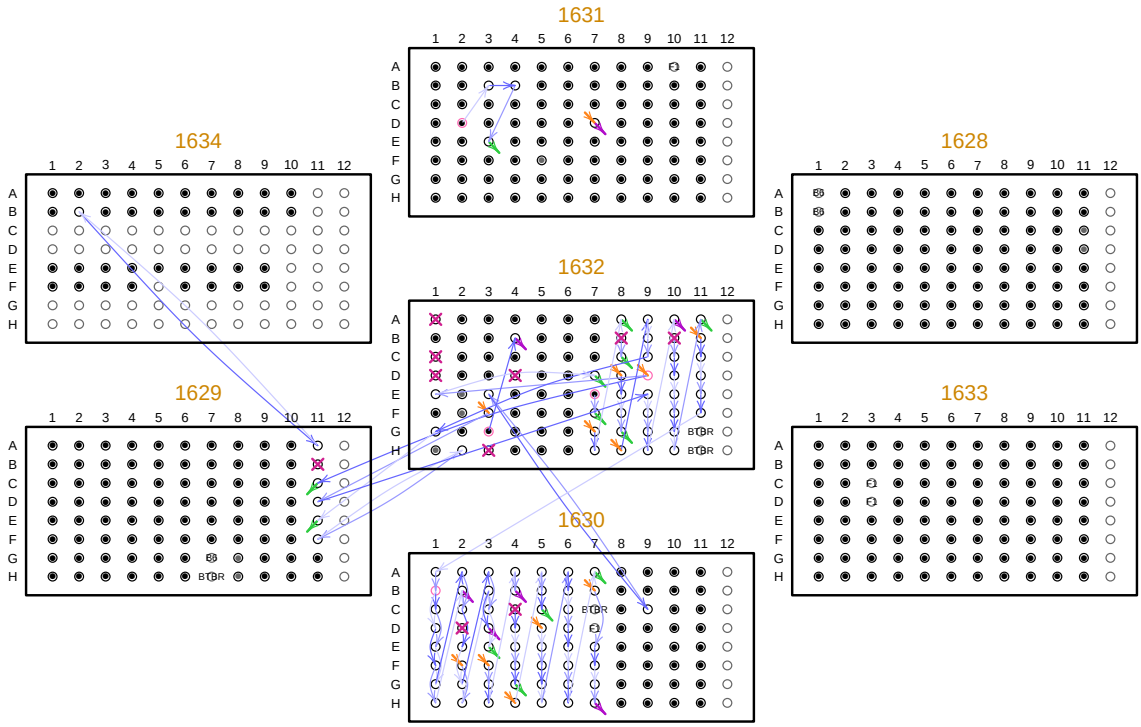
If for each row of the distance matrix we take the value on the diagonal (the self-self distance) and plot it against the minimum value in that row, we find a bunch of samples that look correct (in blue in the lower-left corner), as they are closest to themselves and that distance is small.

There are a number of samples that are wrong but "fixable" (in green), as they are not close to themselves but they are close to some other sample.

Then there are samples "not found" (in red) that are not close to anything. There were actually 550 total mice (good to have backups in case one dies), but only about 500 had gene expression data in any one tissue, and some of the DNA samples were apparently lost.

We don't know, from these results, whether it is the DNA samples that were mislabelled or the mRNA samples, but because we have six sets of mRNA samples, for six different, we can compare the DNA to each of the mRNA samples and in doing so it is clear that it's the DNA that was wrong.

Genotype mix-ups



Even more incriminating, though, is the information about the locations of the DNA samples. DNA samples were arrayed in a set of six 8×12 plates. In this figure, the black dots indicate the correct DNA sample was placed in the correct well, while the arrows point from where a DNA sample should have been to where it actually ended up.

Two of the plates look fine, while half of each of two plates are entirely messed up.

Plate 1631

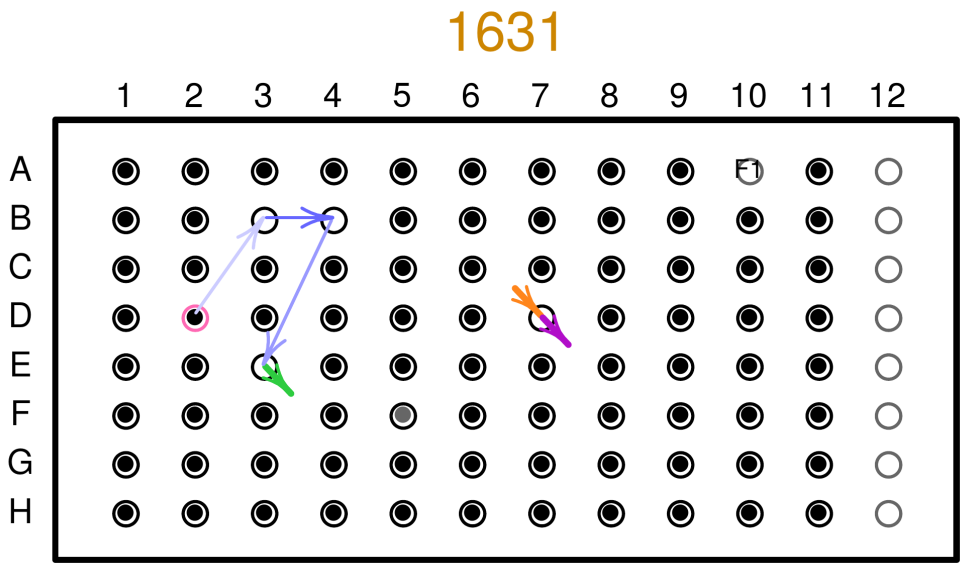


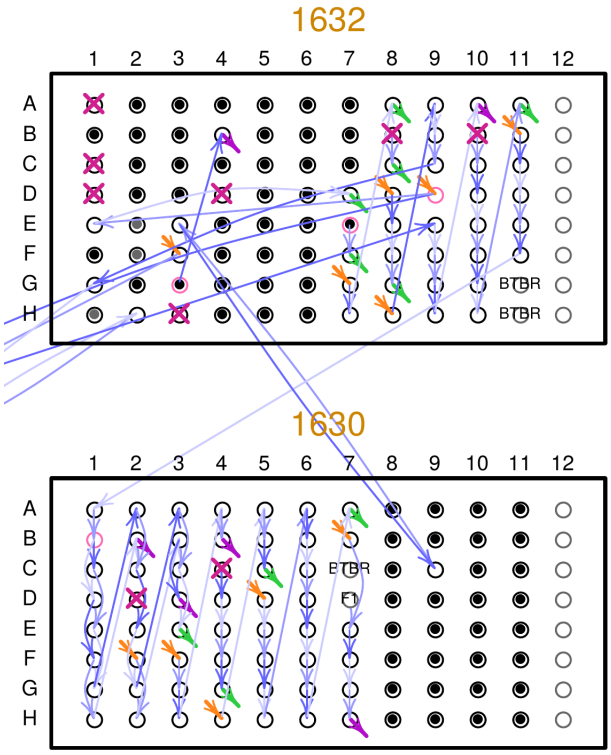
Plate 1631 is a good example. Again, black dots indicate that the correct DNA was placed in the correct well.

The little orange and purple arrow heads indicate that sample in well D7 is of unknown origin, and the sample that should have been there was lost.

The pink circle around D2 indicates that that sample was duplicated: it was placed in the correct well (the black dot), but it was also placed in well B3. The sample that was supposed to be in B3 was placed in B4, the sample that was supposed to be in B4 was in E3, and the sample that was supposed to be in E3 was lost.

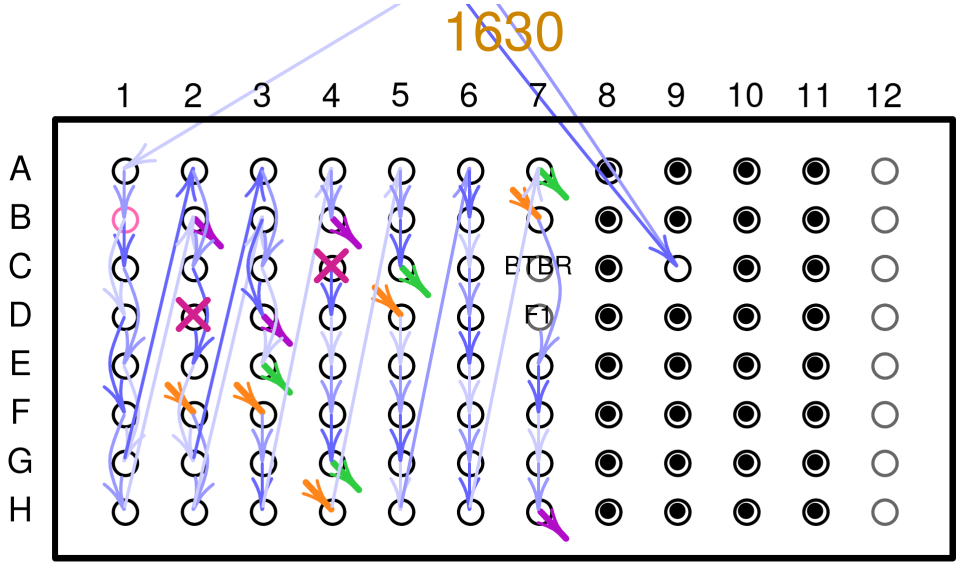
(The purple arrow head for D7 means that the DNA was lost but that there is expression data for that sample, while the green arrow head for E3 means that the DNA was lost but there is no expression data for that sample.)

Plates 1632 and 1630



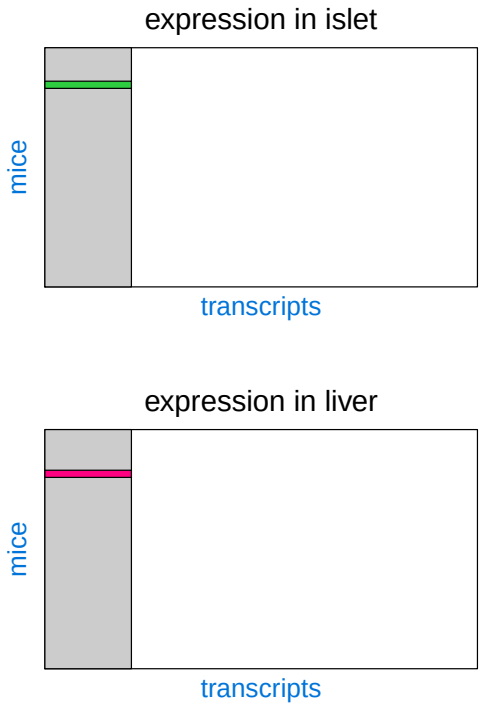
Plates 1632 and 1630 are where most of the problems are. There are some long-range swaps and other misplacements of samples, but most of the problems are due to a series of off-by-one and off-by two errors. Note that the red X's indicate DNAs that were omitted due as being of bad quality (possibly mixtures).

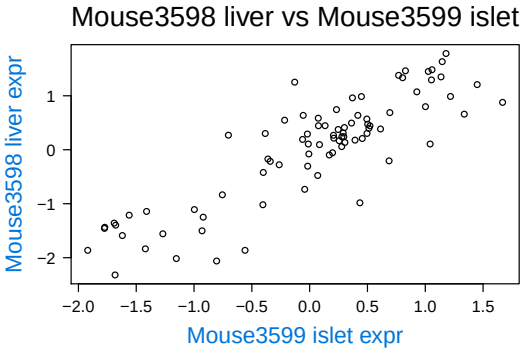
Plate 1630



Consider plate 1630. The sample found in A1 really belonged back on plate 1632. The sample that was supposed to be in A1 was found in B1. The sample that was supposed to be in B1 was duplicated, in both C1 and D1. So then we're off by two for a while: the sample that should have been in C1 was in E1, and the sample that should have been in D1 was in F1, etc. At well F5 we're back to being off by one again, and then a DNA was lost at H7 and we're back to being correct.

E vs E





We can use the same trick to look for mix-ups among the gene expression data sets.

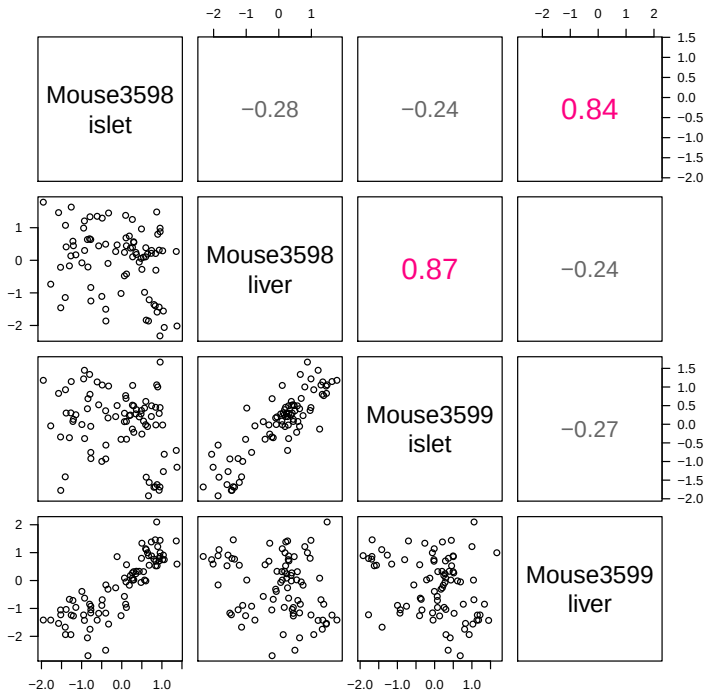
The basic scheme is to first identify a subset of expression traits that are highly correlated between two tissues.

Then look at the correlation between samples, using just that subset of expression traits.

When a sample is correctly labeled in both tissues, the expression values should be correlated. If not, we may find another sample in one tissue that is correlated, to indicate the true label.

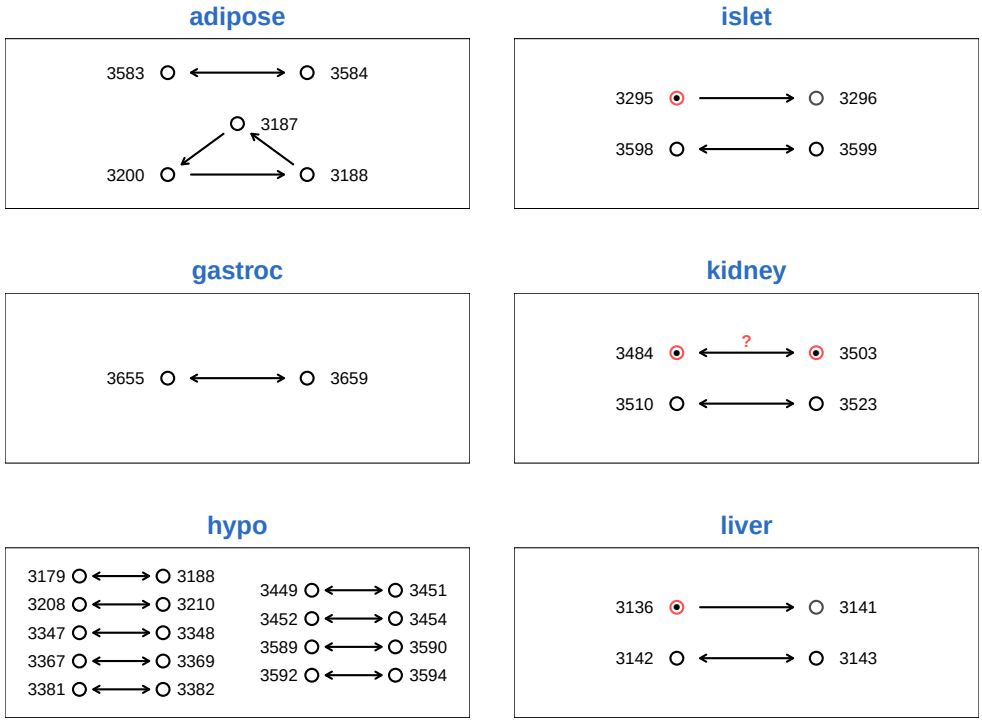
Again, we make use of the multiple tissues to figure out the truth. If we had just two tissues we could see that they were mixed up but not which was the correct label.

E vs E



Here's an example of a mislabelling: Mouse3598 liver looks more like Mouse3599 islet, and Mouse3598 islet looks more like Mouse3599 liver. We use expression in the other four tissues to decide which is right.

Expression mix-ups

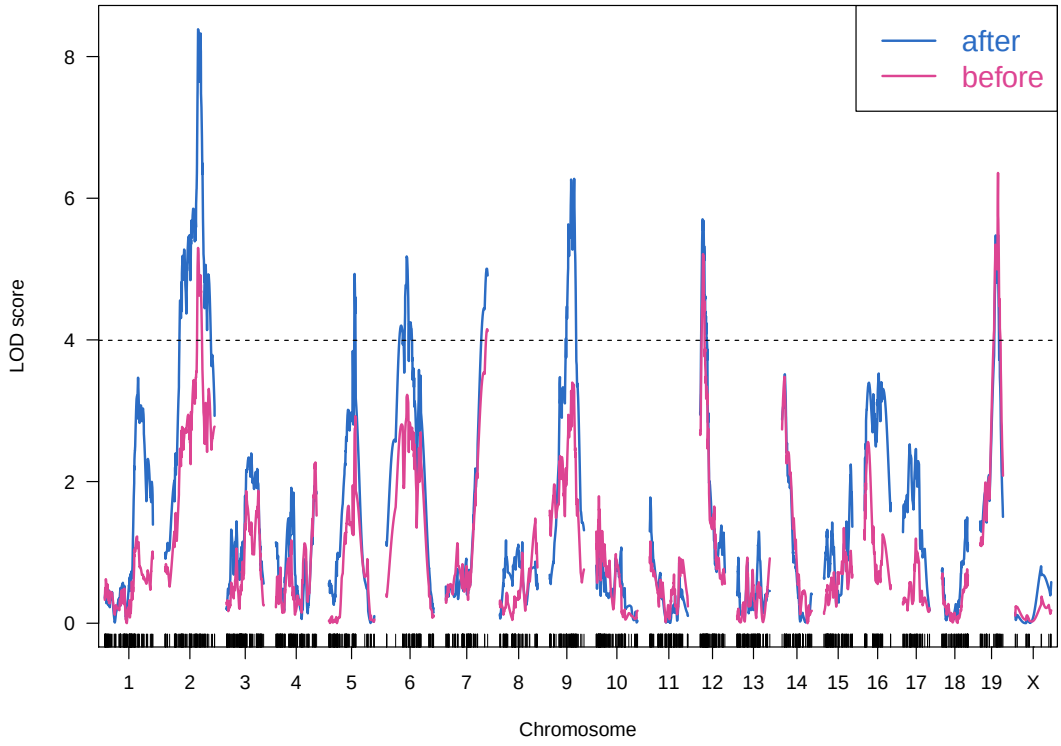


Here are the set of mix-ups I found in the expression data. The arrows point from the correct label to how it appeared.

Each tissue had some mistakes; hypothalmous was the worst. The pink circles indicate a sample duplicate. So, for example, in islet sample 3295 was correctly labeled but also appeared in duplicate with one sample labelled as 3296. The 3296 islet sample was lost.

Adipose had a 3-way swap. 3187 was labelled as 3200 which was labelled as 3188 which was labelled as 3187. Note that most of the problems concern sample numbers that are close (but not necessarily immediately adjacent) in number.

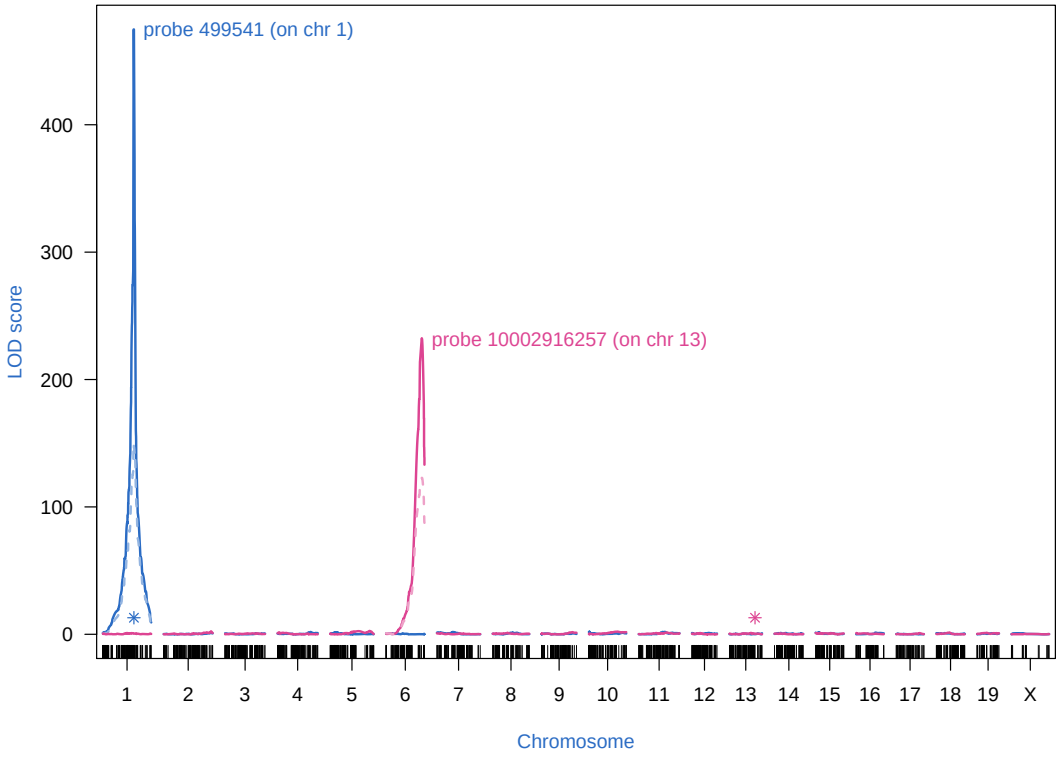
Insulin QTL



This shows the genome scan results for insulin (one of the more important clinical traits) before and after the 18% sample mix-ups were corrected. With the mix-ups in the data, we did see four QTL (chr 2, 7, 12, and 19), but after correcting the mix-ups, the strength of evidence for the chr 2 locus increased considerably, and we see additional significant QTL on chromosomes 5, 6, and 9.

We had been studying these data for a couple of years without noticing any problems. And it is sort of remarkable that with $\sim$20% of the genotypes mis-labelled, you can still get good evidence for QTL. The evidence of course improves greatly when you correct the mix-ups, but it's not as drammatic as you might have expected.

Strong eQTL



The two strong eQTL that I had shown before also show dramatic increases in LOD score after correcting the sample mix-ups, for example the gene on chr 1 had LOD score like 150 and now has LOD score over 450.

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
- Broman et al. (2015) G3 5:2177-2186
  doi: 10.1534/g3.115.019778
- Related work:
  - Westra et al. (2011) Bioinformatics 27:2104–2111
  - Schadt et al. (2012) Nat Genet 44:603–608
  - Ekstrøm and Feenstra (2012) Stat Appl Genet Mol Biol 3:Article 13
  - Lynch et al. (2012) PLoS ONE 7:e41815

In summary, sample mix-ups happen. With eQTL data you can both identify and correct mix-ups. There was great value in having expression data from multiple tissues, in identifying the source of the problems.

The general idea here has wide application for high-throughput data, generally. If you have mutiple rectangles of data whose rows are supposed to correspond, you should check to see if they do correspond. The strategy we used for aligning two expression datasets could work with little change in much broader contexts.

The article describing this work was published in 2015. A number of others happened upon similar problems and similar solutions at about the same time that I did, but published much sooner (2011 and 2012). They're all interesting reads.

Lessons

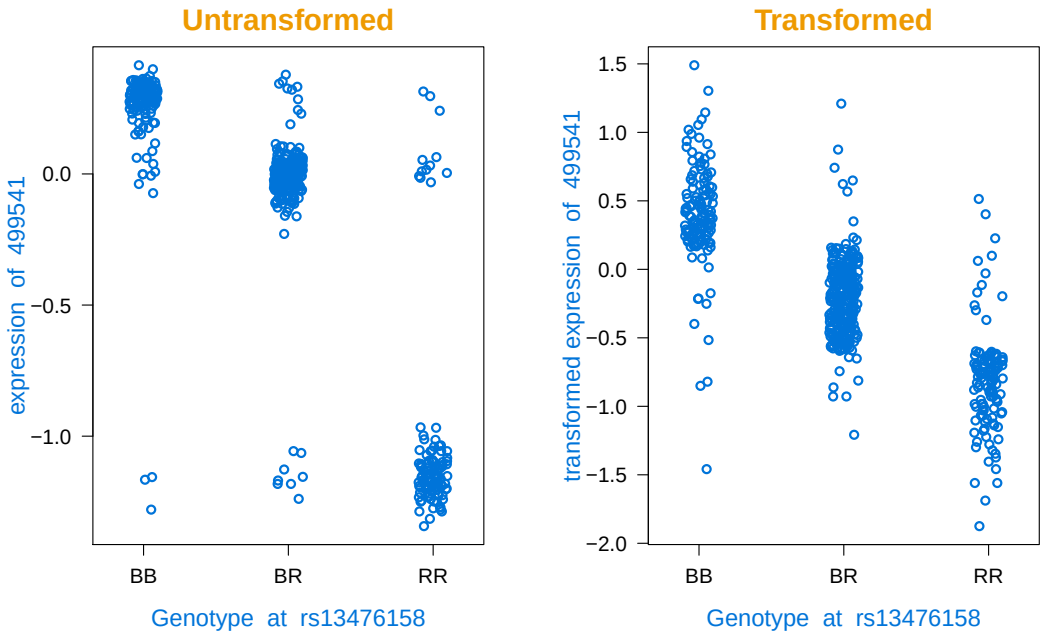
- Don't fully trust anyone
  - Including yourself
- Make lots of plots
  - Don't rely on summary statistics, like LOD scores
  - Look at responses on the original scale
- Follow up all aberrations
- Take your time with data cleaning
  - A month, two months, a year?
- If you have big rectangles whose rows correspond, check that they actually correspond

There are a number of important lessons to draw from this work. First, don't fully trush anyone, even yourself. That seems overly cynical, but really: if you care about the results, take the time to double-check that previous data cleaning efforts (perhaps by you six months ago) didn't skip over some critical problem.

Also, make lots of plots. The long delay in us identifying problems was partly due to the fact that we had mostly focused on summary plots like the LOD curves. You can't really see the problems until you look at the phenotype/genotype relationships. Also, we had transformed all of the expression phenotypes by taking ranks and then converting them to normal quantiles. This was great for eliminating the effects of outliers, but it made it hard to identify problems.

[go to the next slide for an illustration of this point]

E vs G



For example, the panel on the left here is the plot I showed before, for expression vs eQTL genotype. This is the one that had indicated to me that there was a problem.

But we hadn't been looking at the plot on the left, with the untransformed expression values, but rather the plot on the right, in which the expression values were ranked and then transformed to normal quantiles.

The odd pattern on the left is made less odd by the transformation. The plot on the right is a little weird, but it looks more like three overlapping normal distributions.

Transformations are great, but for diagnostic purposes, and to assess the effects of covariates like QTL, it is best to return to the original scale of measurement, as transformations can obscure important features.

Lessons

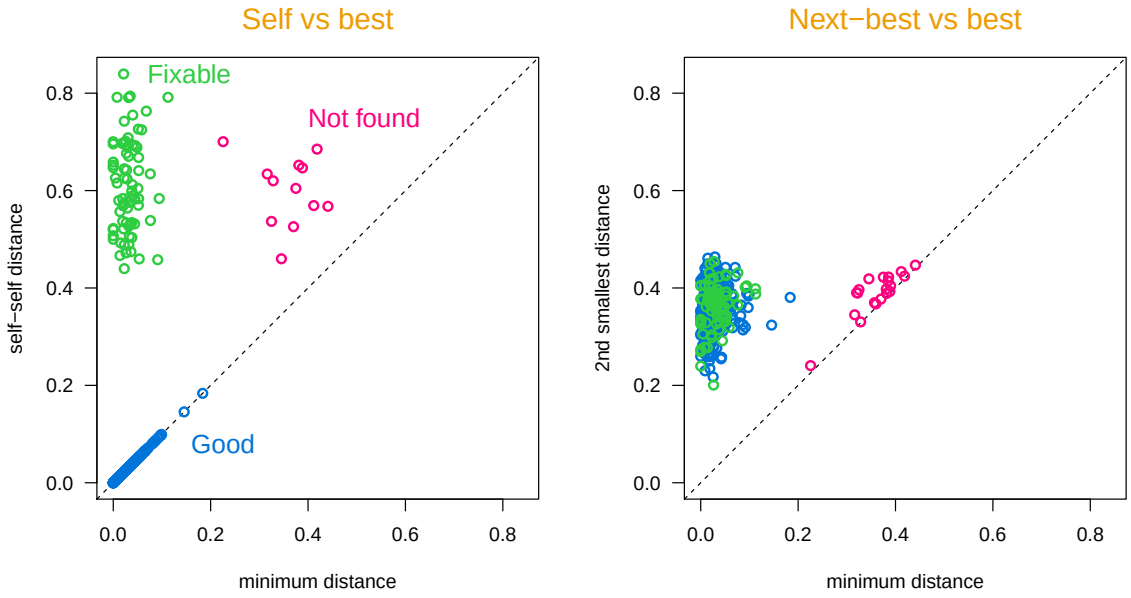
- Don't fully trust anyone
  - Including yourself
- Make lots of plots
  - Don't rely on summary statistics, like LOD scores
  - Look at responses on the original scale
- Follow up all aberrations
- Take your time with data cleaning
  - A month, two months, a year?
- If you have big rectangles whose rows correspond, check that they actually correspond

Returning to our lessons from this case study, we need to emphasize again to follow up all aberrations. I came to the realization of these sample mix-ups on the basis of just 16/500 mice whose sex didn't match their X chromosome genotypes. We might have just omitted the samples and moved on, but it was only by really puzzling through the cause of the problem that I was able to identify the much larger issue.

Related to that: take your time in data cleaning. If you spend $5 million dollars gathering data, isn't it reasonable to spend a month, two months, even a year on the data cleaning? Sometimes it seems like my collaborators think that the more money you spend, the faster you should get results. But if you really care about getting the right answers, you should be willing to spend time verifying that the data are in good order.

Finally, again, if you have mutiple rectangles of data whose rows are supposed to correspond, you should check to see if they actually do correspond.

Decisions



This is an extra slide to show that the evidence for the mixed-up samples' true identities is strong. On the left is the plot we saw before: the self-self distance vs the minimum distance.

On the right is the 2nd biggest distance vs the minimum distance. That the blue and green points are well away from the diagonal indicates that we can tell which sample is really the smallest; there's not some nearby next-most-similar to confuse things.