Data diagnostics

Cleaning genotype data in multi-parent populations

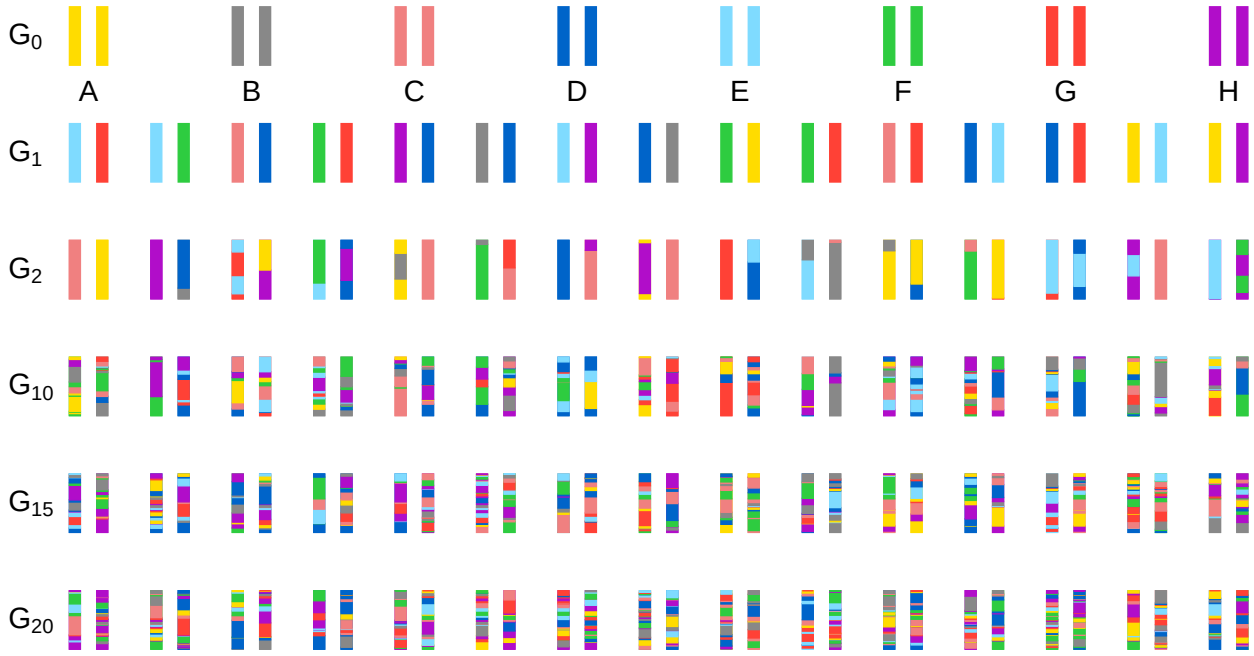
Karl Broman

Biostatistics & Medical Informatics, UW–Madison

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

In this lecture, we'll discuss data diagnostics: studying data to identify and hopefully correct problems. As usual for this class, we'll focus on a particular case study, of cleaning genotype data in diversity outbred mice.

Diversity outbred mice



I want to talk about data cleaning, and I'm going to do so focusing on a cleaning genotype data in Diversity Outbred mice. These mice are an advanced intercross derived from eight inbred founder strains.

Diversity outbred mouse data

- 500 DO mice
- GigaMUGA SNP arrays (114k SNPs)
- RNA-seq data on pancreatic islets
- Microbiome data (16S and shotgun sequencing)
- protein and lipid measurements by mass spec
- Collaboration with Alan Attie, Gary Churchill, Brian Yandell, Josh Coon, Federico Rey, and many others

The data we're looking at concerns a set of 500 DO mice, as part of a collaboration with a bunch of investigators at UW-Madison plus Gary Churchill at the Jackson Lab. We have dense genotypes from SNP arrays, plus RNA-seq data on one tissue and microbiome data, mass spec data, and loads of clinical measurements.

Principles

- What might have gone wrong?
- How could it be revealed?
- Also, just make a bunch of graphs.
- If you see something weird, try to figure it out.

Is data cleaning something that is totally specific to a given context, with no general rules? Or are there commonalities between cleaning genotypes in mice and cleaning say electronic health record data?

I declare that there are some general principles, and I think the first one is: imagine what could have gone wrong. And then next ask how it might be revealed in the data.

Further, you should just make a ton of graphs, and if you see something odd, try to figure out what's going on.

Possible problems

- Sample duplicates
- Sample mix-ups
- Bad samples
- Bad markers
- Genotyping errors in founders

For these genotype data, these are the main problems: sample duplicates or mix-ups. Sample mixtures, even.

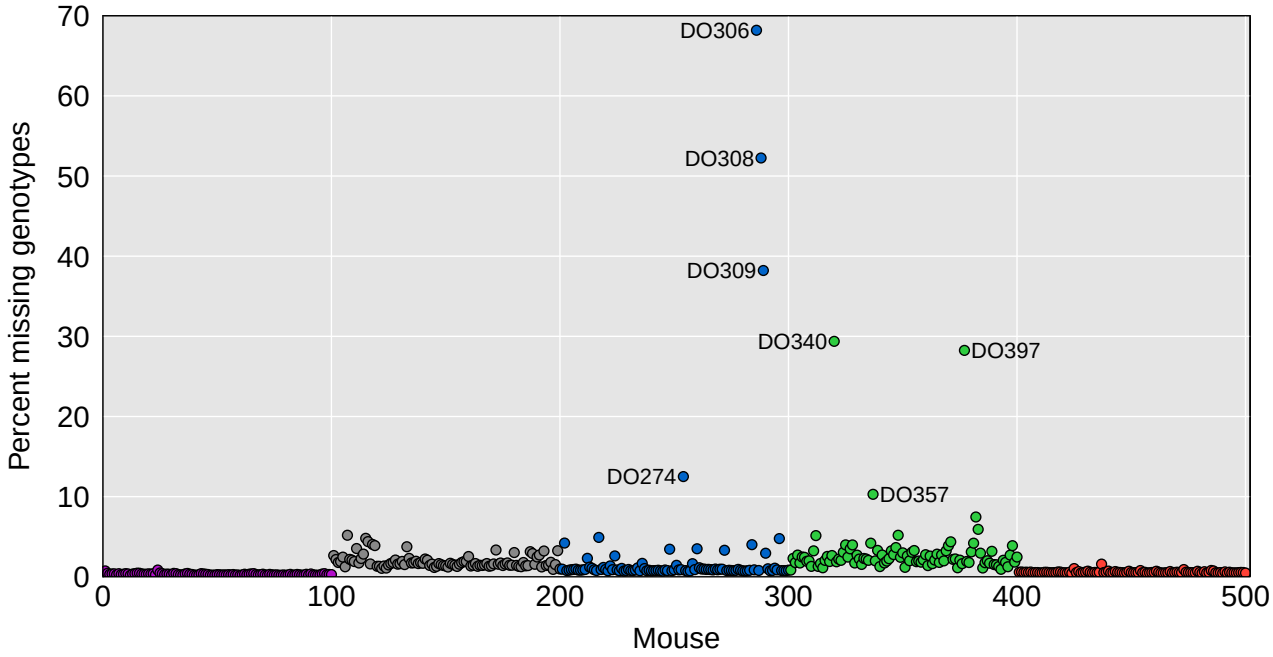
Also samples being bad, or genetic markers being bad. Also genotyping errors in the founder strains.

Could also be that markers are on the wrong chromosome.

What to look at first?

A key question is what to look at first? What are the most fundamental problems, and how might we find them?

Missing data per sample



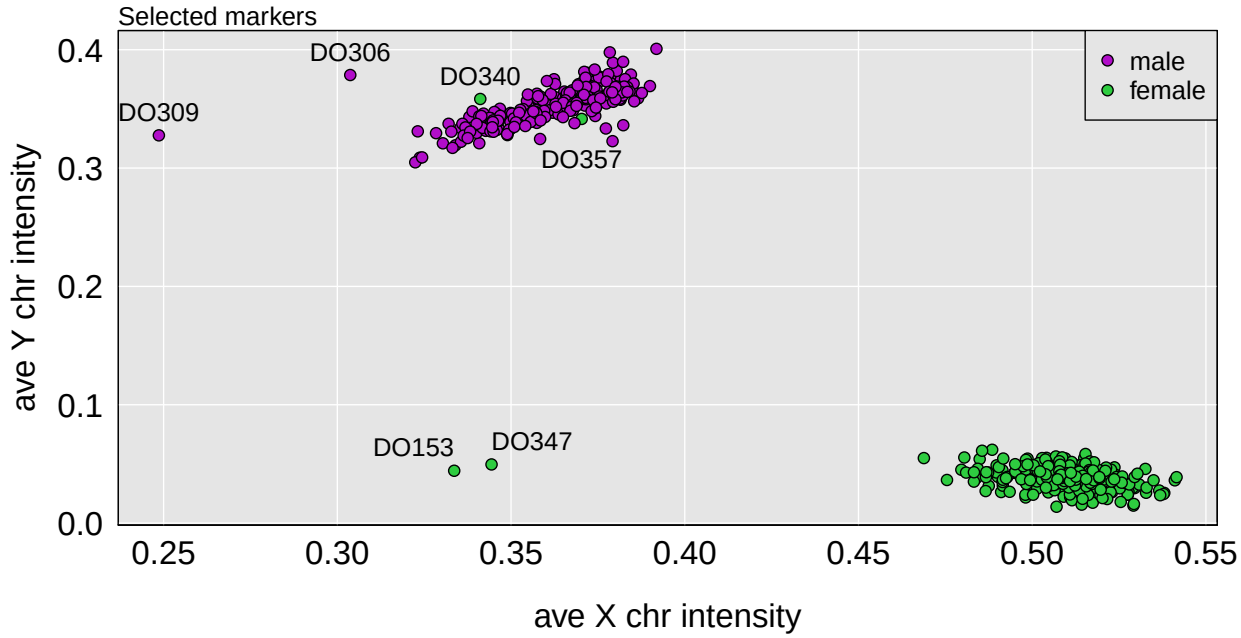
For genotype data, I think the first thing to look at is the amount of missing data. High proportions of missing data are usually an indication that a DNA sample was bad.

The 500 mice were gathered in batches of 100. There were clear differences in the amount of missing data by batch. But mostly, there are a set of 7 samples with >10% missing data, which should probably be tossed, as probably the remaining data is bad.

Swapped sex labels

The next thing that I always look at is whether the X chromosome genotypes for the mice matches what we would expect, given the reported sex.

Ave SNP intensity on X and Y chr

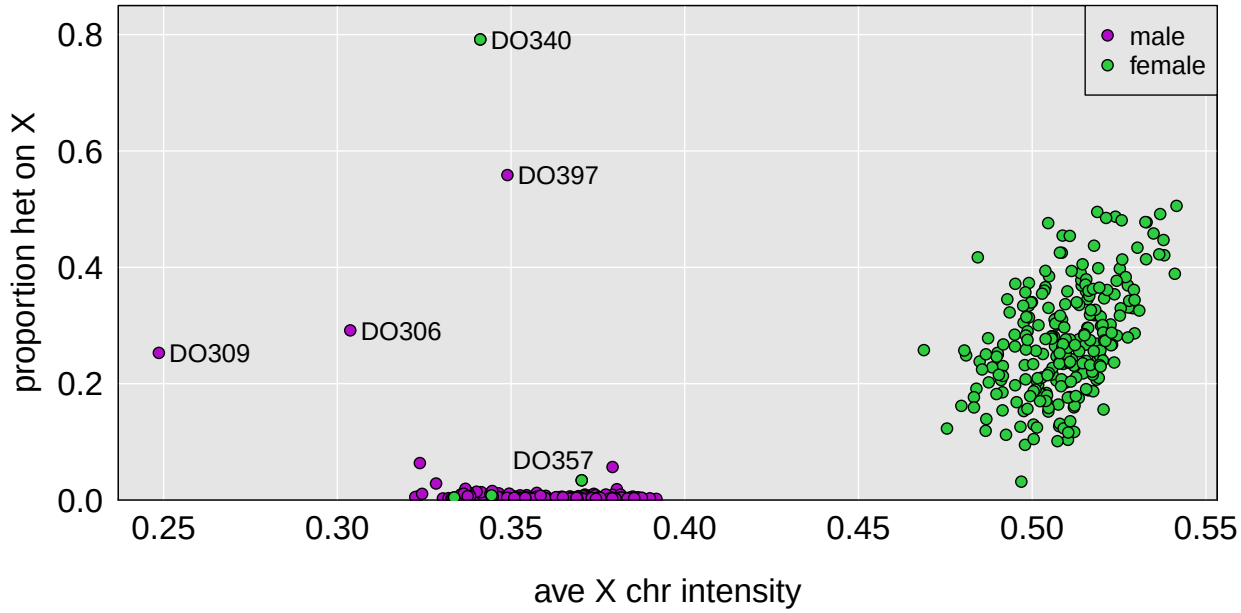


I've used a lot of different techniques for verifying sex from X chromosome genotypes. But for these SNP array data, and probably also for sequence-based genotypes, the best thing seems to be to look at the intensity of alleles on the X and Y chromosomes. This is because the chromosome dosage effects are really strong. Females have two Xs and no Y chromosomes, while males have one of each.

It turns out to be important to subset the X and Y chromosome markers to the ones that show a real sex difference. There seem to be a lot of markers that are annotated as being on the X or Y but don't actually have the expected dosage effects.

Having done that, we see two clear blobs: males with high intensity on Y and low intensity on X, and females with the opposite. There are a couple of clear females that are really males, plus a couple of mice that look like XO females.

Heterozygosity vs SNP intensity on X chr



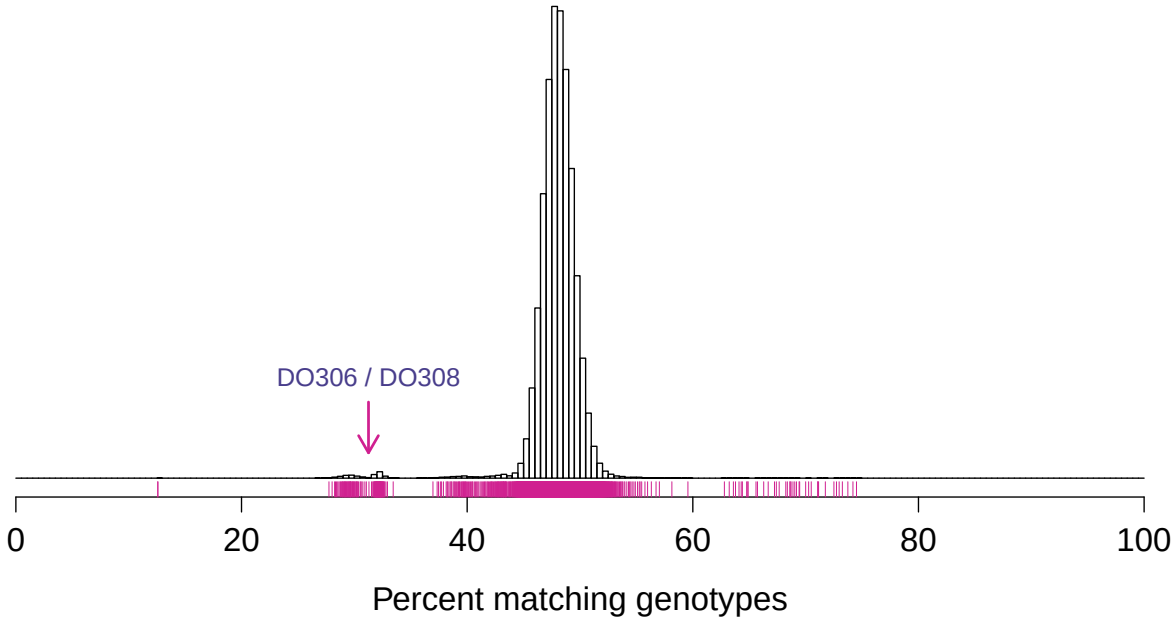
Traditionally, I would focus on heterozygosity on the X chromosome. That females will have some heterozygous calls and males should have none. This plot shows heterozygosity on X, vs X chr intensity as before. The idea is that the vertical axis here is informative, but not so informative as the just the X and Y chr intensities themselves.

It can take a while, and a lot of exploration, to come to the best diagnostics for a given problem.

Sample duplicates

The next thing to look for is sample duplication. Are there pairs of samples with very similar genotype data?

Percent matching genotypes between pairs



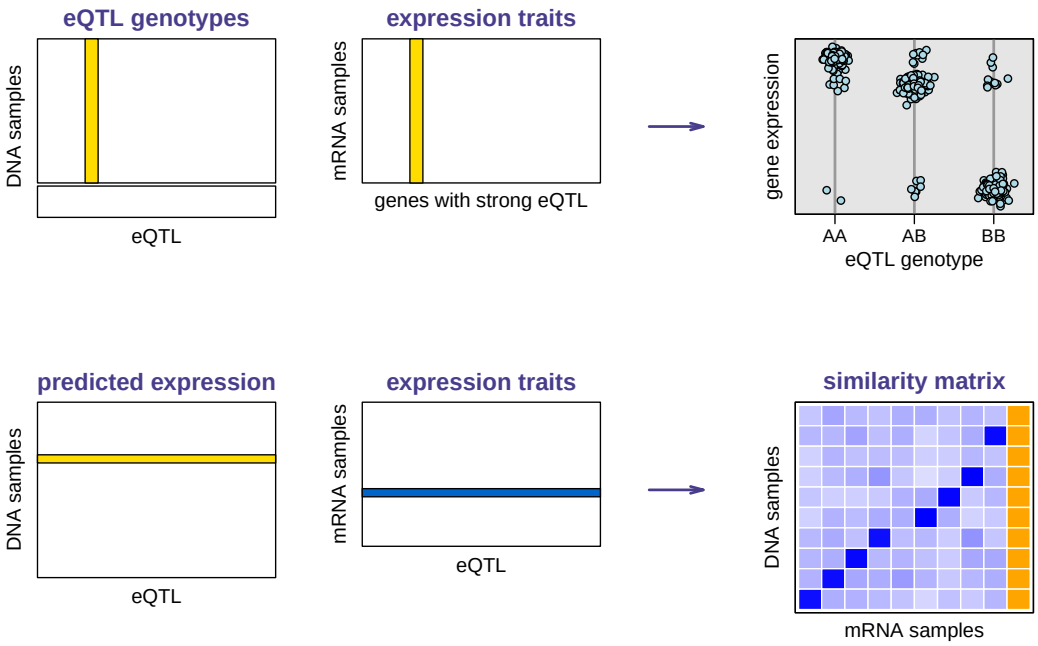
Here, I'll just look at all pairs of samples and calculation the proportion of matching genotypes. A histogram of that will many times show some pairs near 100%.

Here, the average is around 50%. There are some pairs that are especially low; these all involve either DO306 or DO308, and are just due to those two samples being crappy. There are also some pairs that are at around 70%. Those look to be siblings (which were supposed to be avoided here).

Sample mix-ups: RNA-seq data

Next I look at sample mix-ups, starting with the RNA-seq data. Do the RNA-seq results seem to match the genotype data?

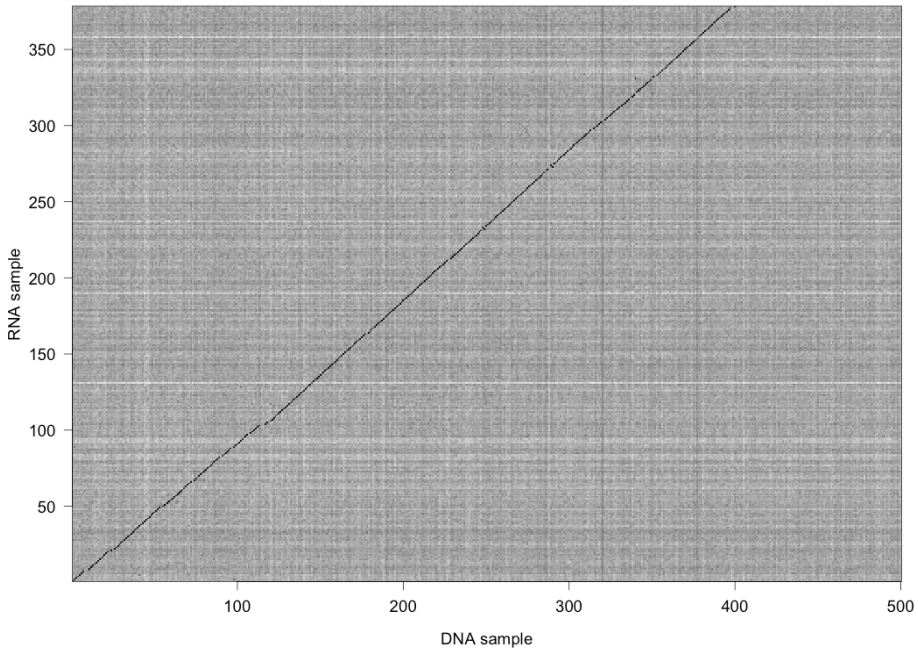
RNA-seq mix-ups



My scheme for looking at this is very similar to the lecture last week. First, for each expression trait, look for a strong eQTL.

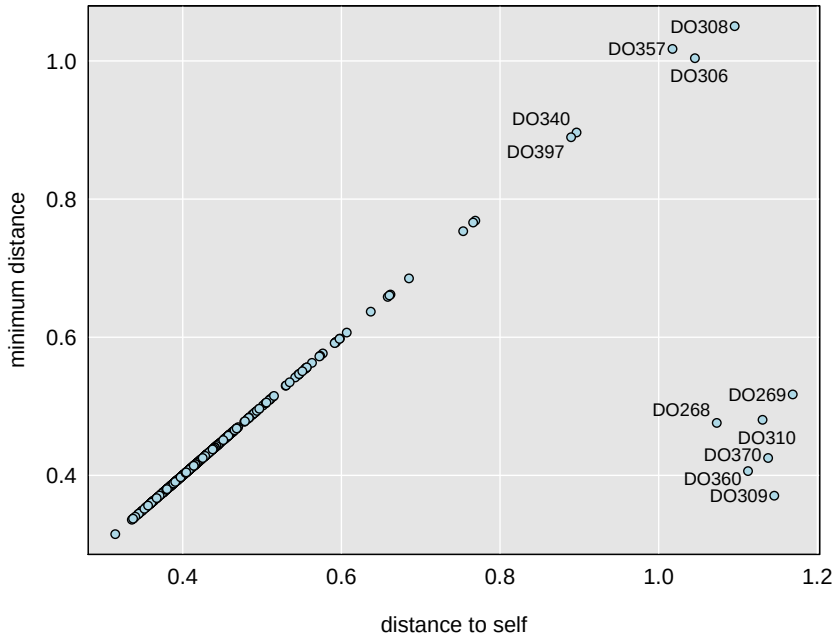
In a twist relative to what I'd done last week, instead of trying to predict genotype from expresssion, let's instead predict expression from genotype. So then we have, for each mouse, a set of predicted expression values. Compare those to the observed expression values and calculate the correlation as a measure of similarity (or maybe the RMS difference as a measure of dissimilarity).

Distance matrix



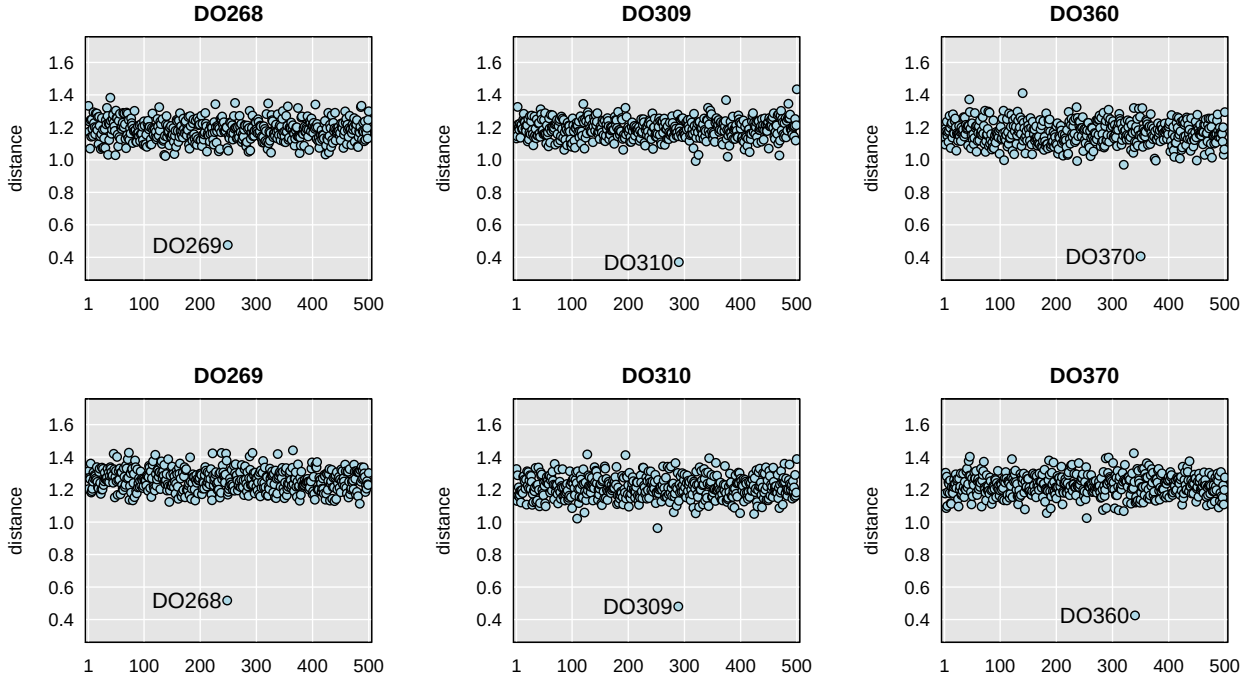
Turn those results into a distance matrix: for each RNA sample, how similar is it to each DNA sample? Here we have just under 400 RNA samples and 500 DNA samples. The leading diagonal should be saying the matching samples.

Min vs self distance



We can plot the minimum in each row vs the value on the diagonal, we get this plot. Values along the diagonal are presumed correct. Values in the lower-right corner are wrong but there's some other sample that their close to. If you look at the sample IDs, you can probably see that there are 3 sample swaps down there.

RNA-seq mix-ups, details

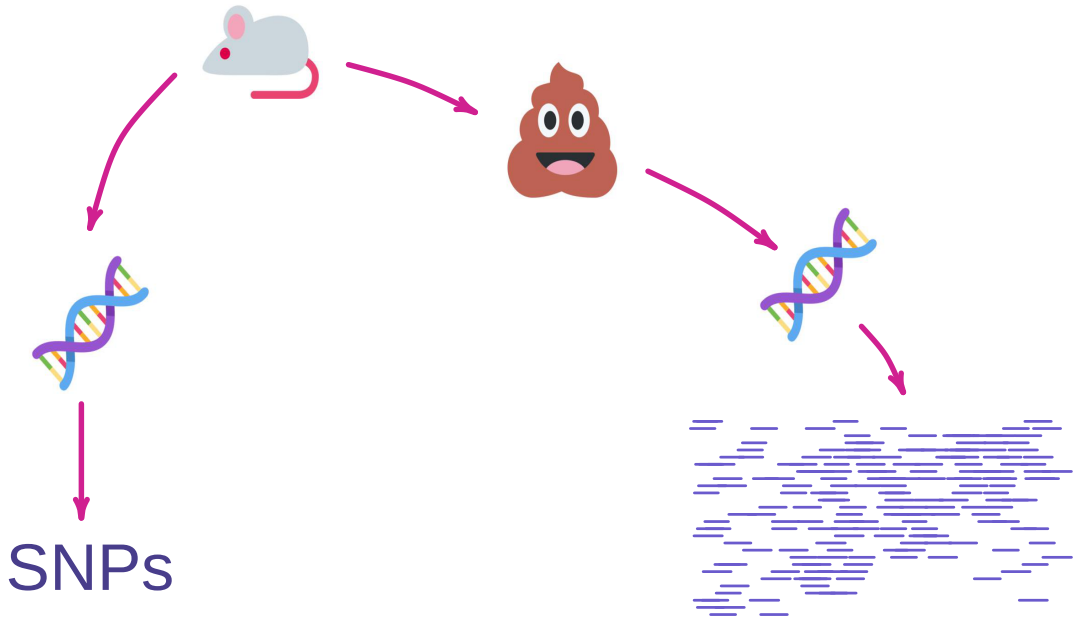


Here we look at those six problem samples in more detail, by plotting their distance to each other sample. We see that these are clearly three pairs of sample swaps. We can't be sure whether the problem is in the DNA or in the RNA.

Sample mix-ups: microbiome data

We also have a set of microbiome samples that could be used to investigate sample mix-ups.

Microbiome data



In this part of the study, they took mouse poop and extracted DNA and then did massive parallel sequencing. The goal was to characterize the populations of microorganisms in the gut of the mice, but the sequences also include many that were derived from the mouse house. The question is: do those mouse-derived sequences seem to match the SNP genotypes we have?

Sample mix-ups: Microbiome data

- Impute genotypes at all SNPs in DNA samples
- Map microbiome reads to mouse genome; find reads overlapping a SNP
- For each pair of samples (DNA + microbiome):
  - Focus on reads that overlap a SNP where that DNA sample is homozygous
  - Distance = proportion of reads where SNP allele doesn't match DNA sample's genotype

Our approach was first to impute the genome-wide SNP genotypes using the founder sequence data, and then to look at microbiome reads that overlap a SNP. For each pair of samples (genotypes, microbiome), we focused on reads where the genotype data said homozygous, and then looked at whether the reads had the same or a different allele.

Microbiome DO361 vs DNA DO361

	AA	BB
A	939,918	1,044
B	2,998	125,962

Here's a pair of samples both labeled DO361. The split up the SNPs into those where DO361 is homozygous for the major allele and those where it's homozygous for the minor allele. Then we look count the reads that overlap those SNPs, split according to whether they are showing the major or minor allele. Here we see nice correspondance. We can use the mismatches as an estimate of sequencing error, though it could also reflect errors in the SNP genotypes.

Microbiome DO360 vs DNA DO360

	AA	BB
A	2,661,645	190,188
B	427,685	202,335

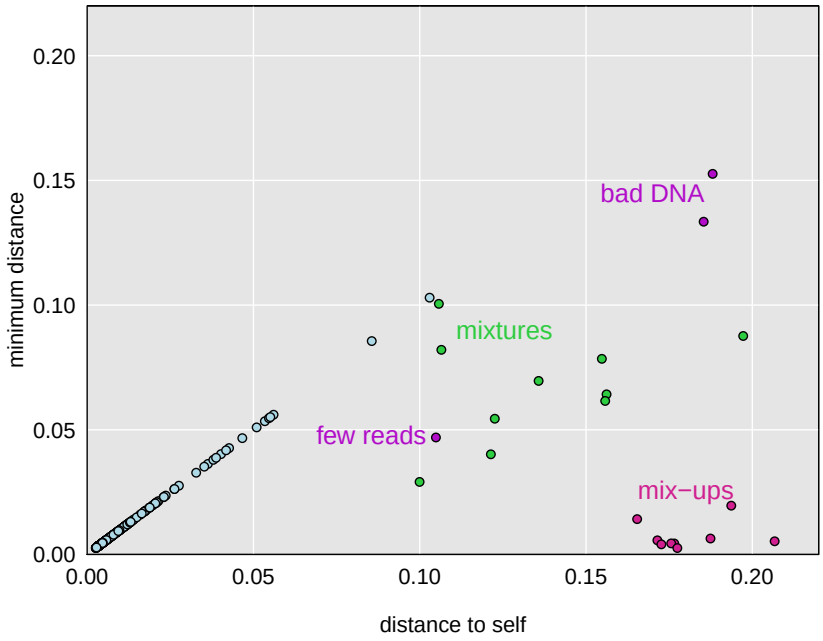
This sample, though, shows poor correspondance.

Microbiome DO360 vs DNA DO370

	AA	BB
A	3,137,751	1,475
B	7,461	310,369

If we compare that microbiome sample to a different DNA sample, though, we can find one with good correspondance. And note that this is one of the pairs that showed up as a mix-up when we looked at the RNA-seq data.

Microbiome mix-ups: min vs self distance



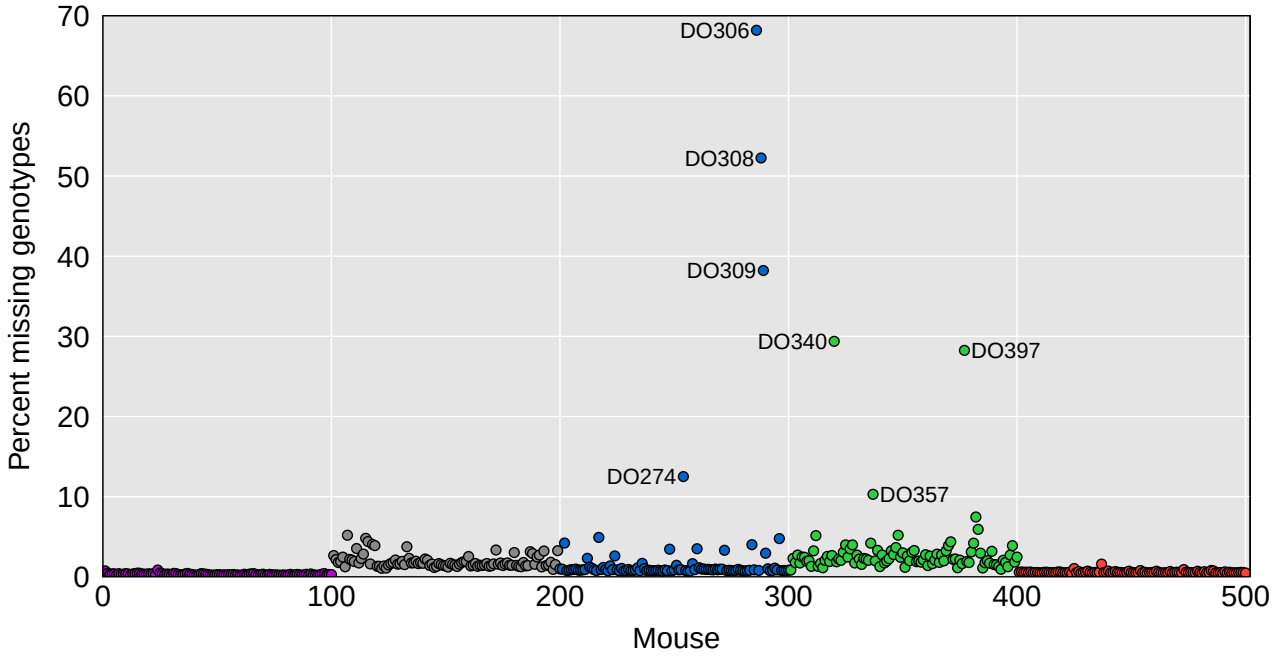
We can turn those frequencies of corresponding vs opposite alleles into a distance matrix, and then plot the value on the diagonal vs the minimum value in each row, as before. We find that most samples are fine, but there are also some clear mix-ups. And then there are a number of samples that are sort of in-between. Many of these turn out to look like mixtures. (And some of the good samples also really look like mixtures.)

Of the mix-ups, there was just one in common between the microbiome data and the RNA-seq data. But while we have 500 genotyped mice, only 400 were assayed for RNA-seq, and 400 were assayed for microbiome, with only 300 in common. And the other mix-ups are in the portion that can't be triagulated.

Sample quality

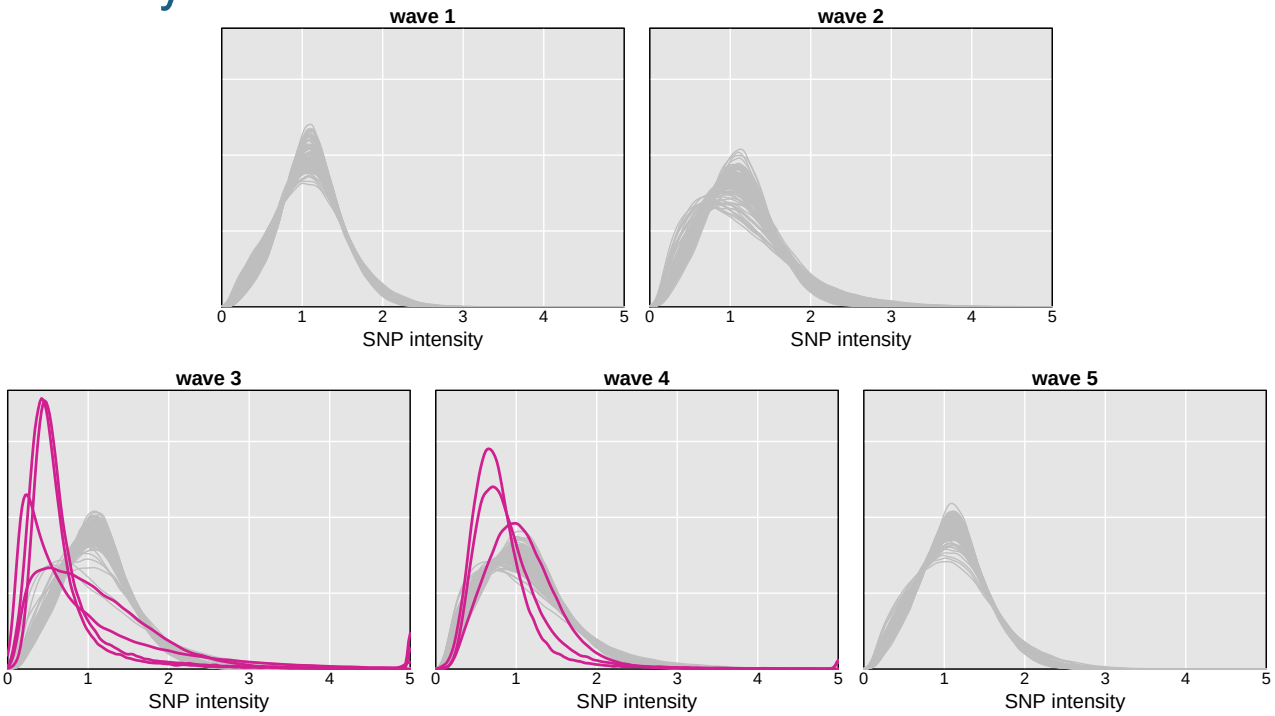
The next thing I look at is sample quality. Are there particular samples that are bad and should be omitted?

Missing data per sample



Again, the first thing to look at regarding sample quality is the amount of missing data per sample.

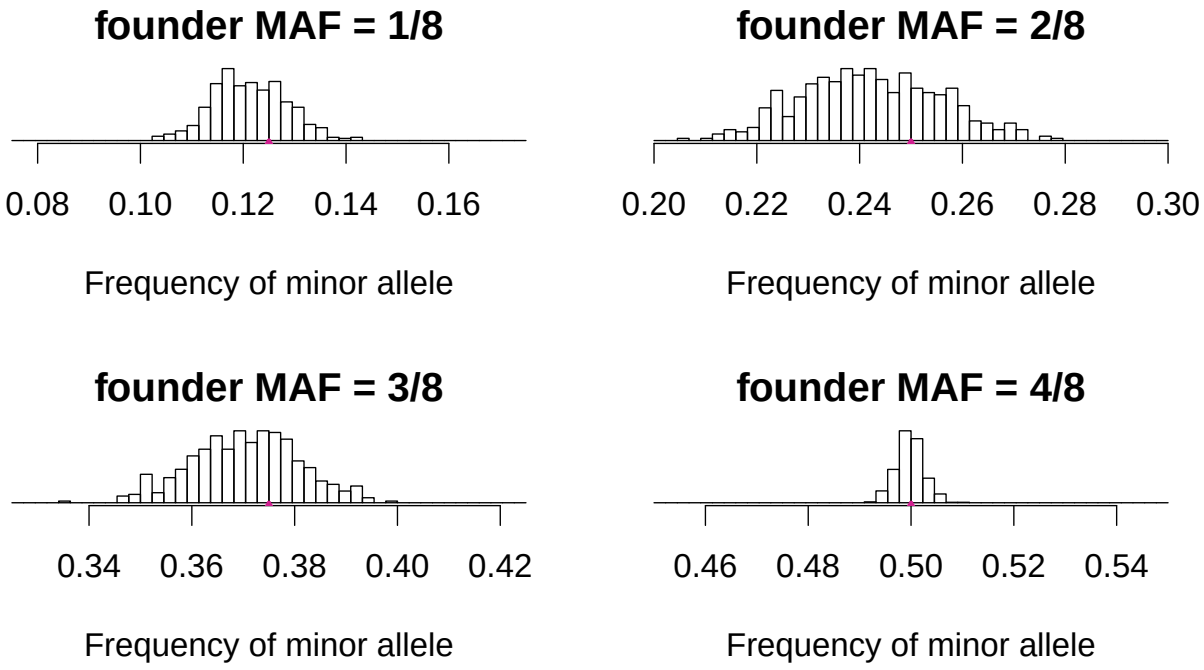
SNP array intensities



Also of interest is to look at the distribution of intensities on these SNP arrays. It's hard to look at 500 histograms, but actually you can get a pretty good picture of them by plotting density estimates. I've highlighted in pink the 7 samples that have very high rates of missing data. But in addition, you can sort of see two clusters of curves: a sort of typical curve (like all those in the first batch) plus curves that have somewhat reduced median and perhaps a heavy right tail.

To more precisely identify the groups of curves, you might make a scatterplot of say the 1st percentile against the 99th percentile.

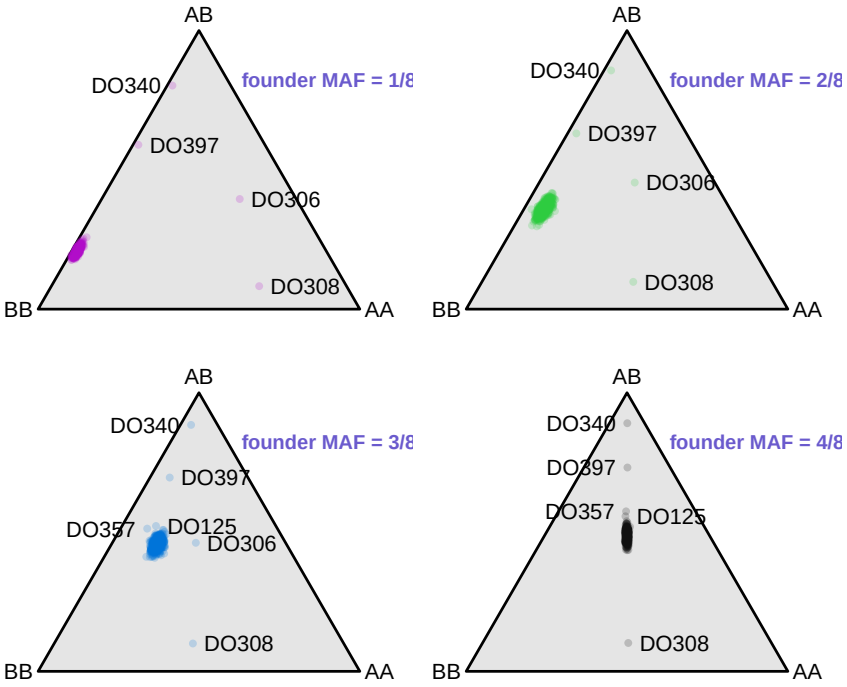
Allele frequencies by individual



The frequencies of the alleles at the SNPs can be an important clue about whether the samples are good quality and as expected. It's best to split up the markers according to the allele frequencies in the founders. Define the "minor allele" to be the less-common one in the founders, and split the markers according to their "minor allele frequency (MAF)".

In these data, there are some gross outliers, but the rest are reasonably well distributed around the expected frequency.

Genotype frequencies by individual



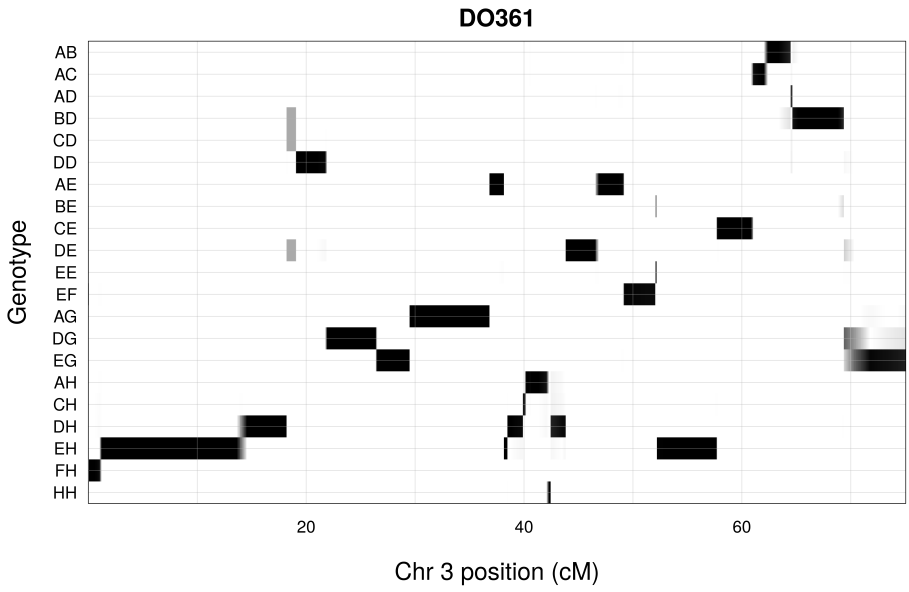
We can also look at the SNP genotype frequencies, again splitting the SNPs into four groups according to the minor allele frequency.

Here I make use of what's called a "ternary" diagram. It makes use of the fact that for any point within an equilateral triangle, the sum of the distances to the three sides is a constant. We can use that fact to create a correspondance between points in the triangle and trinomial probabilities.

In the upper-left panel, for MAF = 1/8, the points all have low frequency of the AA genotype, and so are close to the left edge. They are a bit closer to the lower edge (indicating low heterozygosity) and far from the right edge (indicating large probability of BB genotype. (Here "B" is the more frequent allele, which is the opposite of the notation I used when looking at the microbiome mix-ups.)

The outlying points are largely the samples with a lot of missing data, which we'd identified before. But there's also one mouse DO125, who has somewhat elevated heterozygosity.

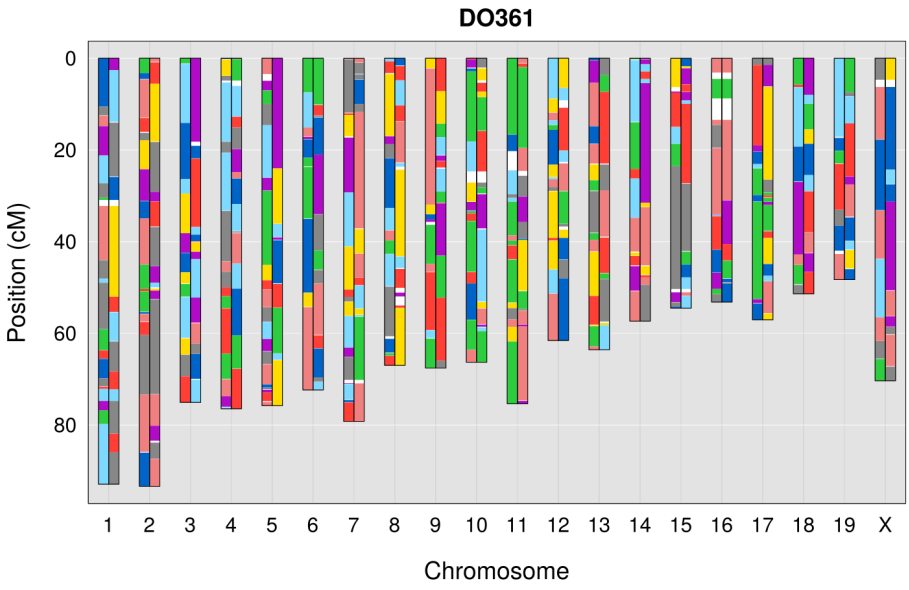
Genotype probabilities (one mouse on one chr)



Further diagnostics related to the sample quality make use of a reconstruction of the founder haplotypes in the DO offspring. I make use of a hidden Markov model, allowing for the presence of genotyping errors, to calculate the probability that each mouse has each of the 36 possible genotypes (AA, AB, BB, BC, ..., HH).

Here is a depiction of the genotype probabilities for one mouse on one chromosome.

Genotype reconstruction (one mouse)

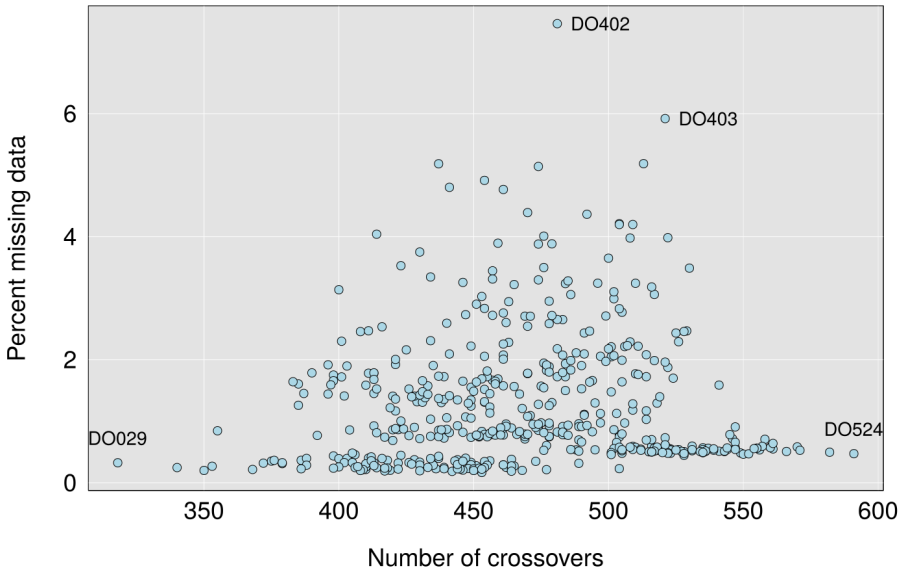


Here is a depiction of the genome reconstruction for one mouse. White bits are unknown; the other 8 colors correspond to inferred alleles.

There are two useful diagnostics that we can derive from this reconstruction. First, we can just count the total number of crossovers in the genome. Bad samples often will show an excess of apparent crossovers.

Second, we can use this reconstruction plus the SNP alleles in the founder strains to get predicted SNP genotypes for this mouse. We can then compare those to the observed SNP genotypes.

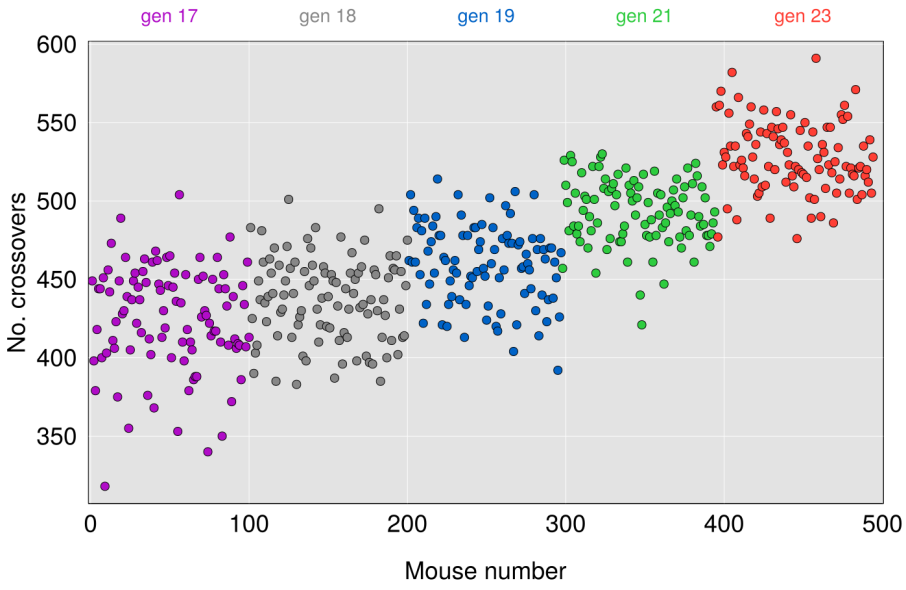
Percent missing vs. number of crossovers



We first look at the relationship between percent missing data and the observed number of crossovers. The mice with >10% missing data show an absurdly large number of crossovers, indicating that the remaining genotype data looks messed up.

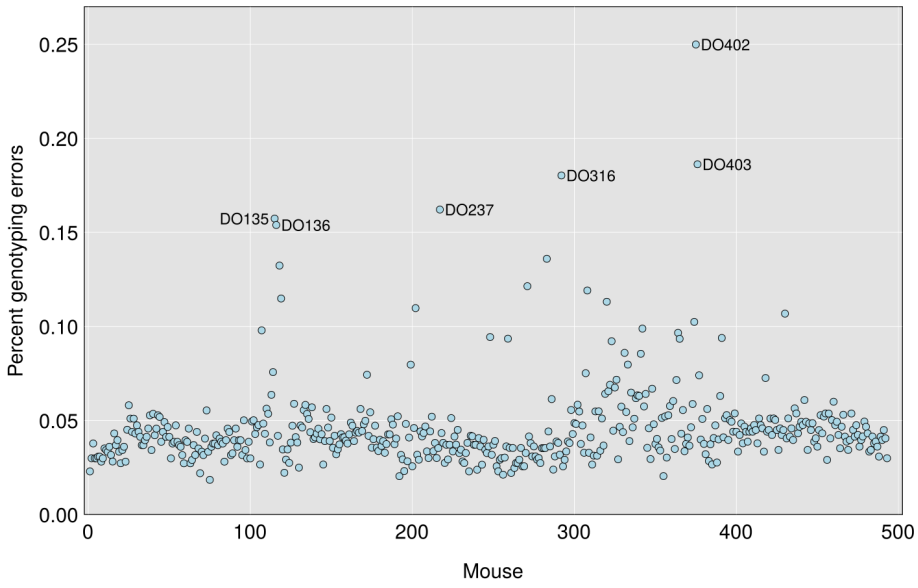
The other mice have aruond 300-600 crossovers, and there seems no relationship between the amount of missing data and the number of crossovers. Mice missing 4–6% genotypes seem to show a normal number of crossovers, indicating that their remaining genotypes are probably okay.

Crossovers by generation



The five batches of mice in this study are actually from different generations of the cross, and the number of crossovers seen are increasing with generation. So in considering the number of crossovers in a mouse, it's important to compare it to the other mice of its generation. Other than the seven mice with >10% missing genotypes (which are omitted from this figure), there are no worrisome outliers.

Percent genotyping errors

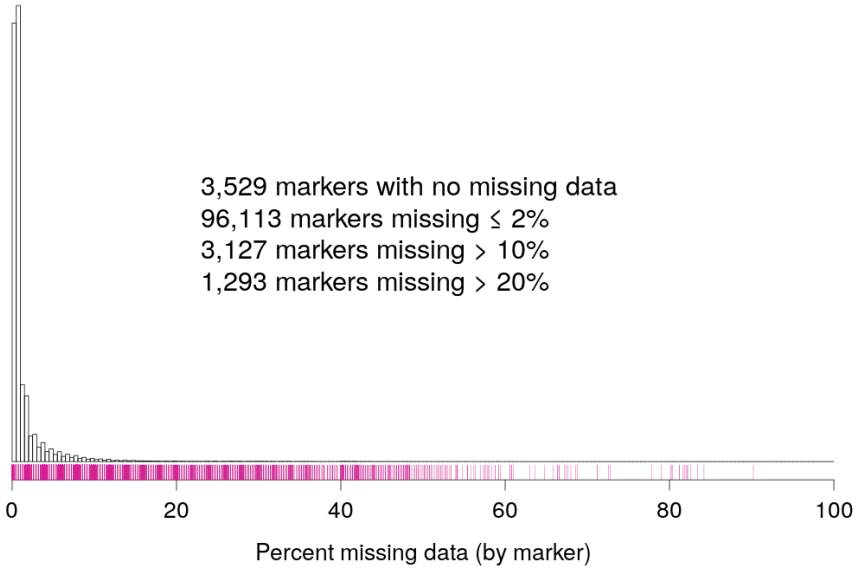


Finally, we compare the observed and predicted SNP genotypes to get an estimated genotyping error rate for each mouse. The mice with lots of missing data have high genotyping error rates. The others have rates that are on no more than like 1–2 per 1000, and most have error rate like <5 per 10,000. Given that they also show a reasonable number of crossovers, these are all acceptable levels of error.

Marker quality

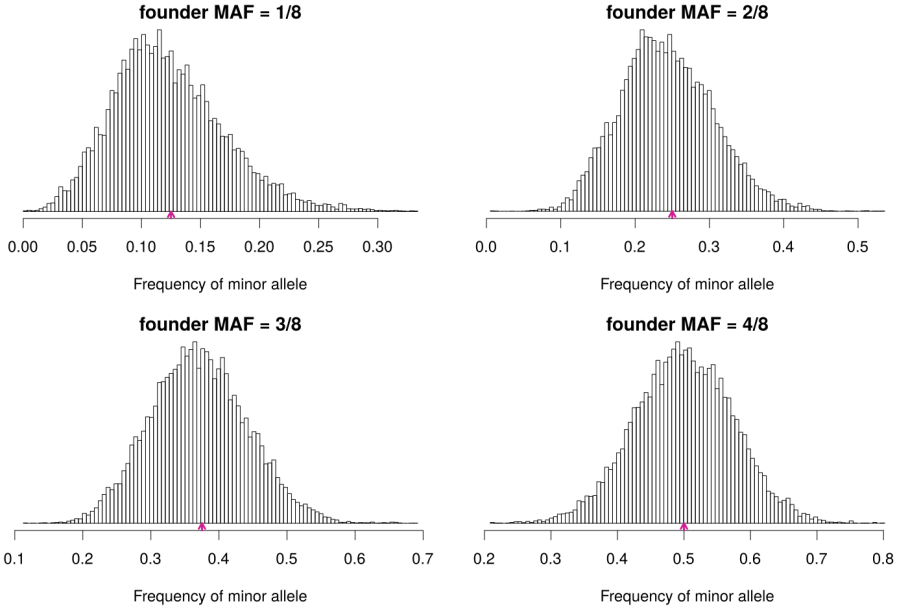
Looking for badly behaved markers is much like looking for bad samples, but we have many more markers to sift through (115,000 vs 500 samples), and less information about each.

Proportion missing data



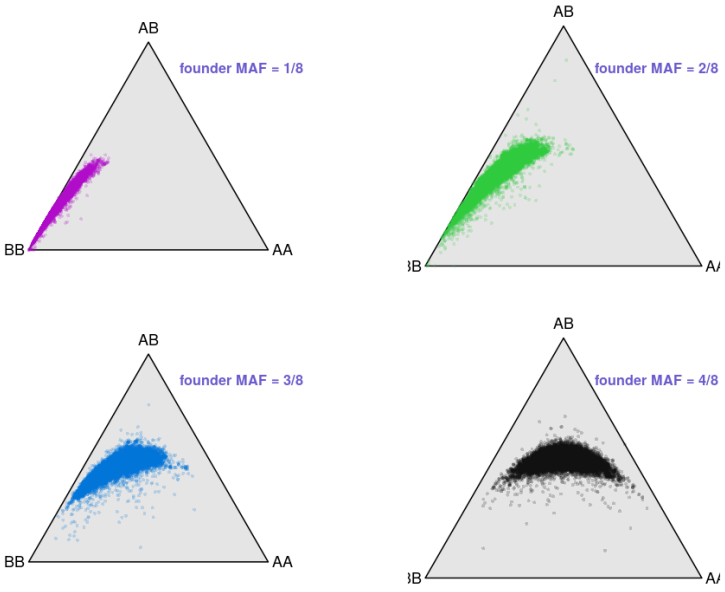
We can again start with the amount of missing data. Most markers have very little missing data, but there are 3,000 with >10% missing genotypes.

Allele frequencies by marker



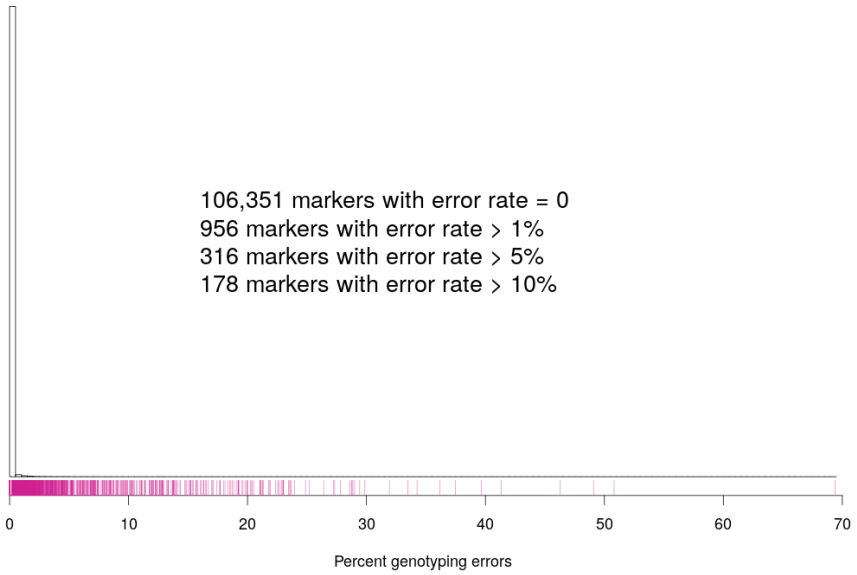
We can also look at the allele frequencies at the markers, again split by the allele frequencies in the founders. There's a lot of variation, but they on average they hit the target.

Genotype frequencies by marker



The genotype frequencies are similarly informative, but except in some extreme cases, it's hard to identify markers that are for sure bad from what could be natural variation.

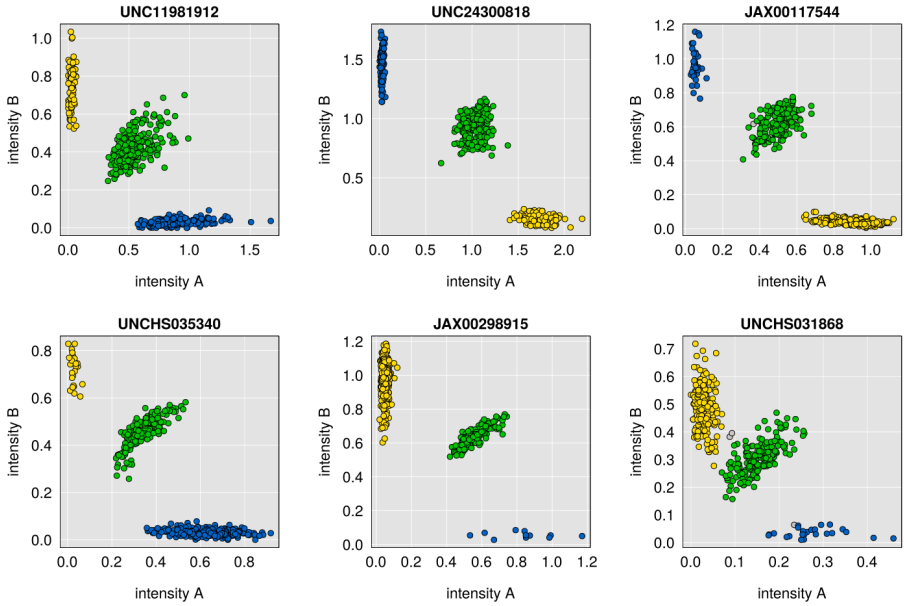
Genotyping error rates



Actually, what seems to be most useful is to use the genome reconstructions which then give predicted genotypes for each mouse. Just as we can use those to get estimated genotyping error rates for each mouse, we can also get estimated genotyping error rates for each marker. Does a marker's genotypes seem to correspond to what you would expect, given the surrounding markers?

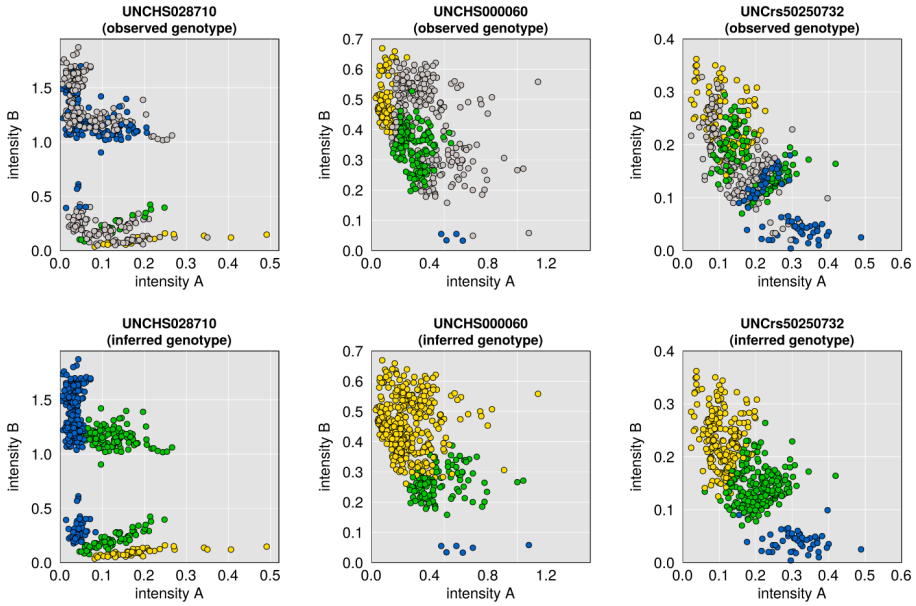
The vast majority of markers have precisely 0 apparent errors. There are less than 1000 with error rates above 1%, and just about 300 with really large error rates.

Nice markers



Most informative, for investigating marker quality, is to look at the allele intensities from the array. Here are a set of nicely behaved markers, with points colored by the genotype calls (green are heterozygotes; gray are missing). The genotype calls are based on plots like this: cluster analysis to define the three genotype clusters. All of these markers are nicely behaved and have three distinct clusters in the locations where you'd expect them to be.

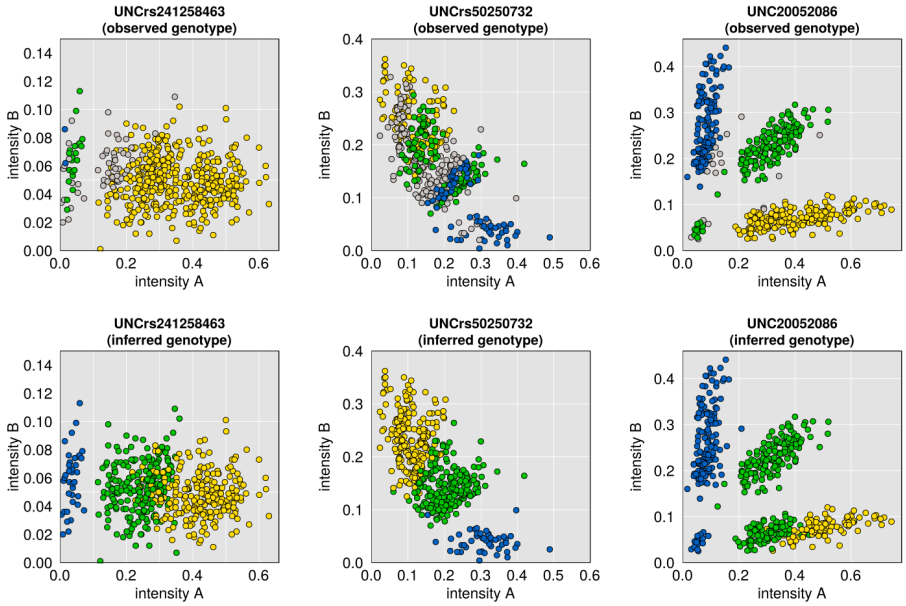
Crap markers



Here are a set of three badly behaved markers. In the top panel, the points are colored by the genotype calls. In the bottom panel, the points are colored by the predicted genotypes, using the genotype reconstructions and the SNP alleles in the founder strains.

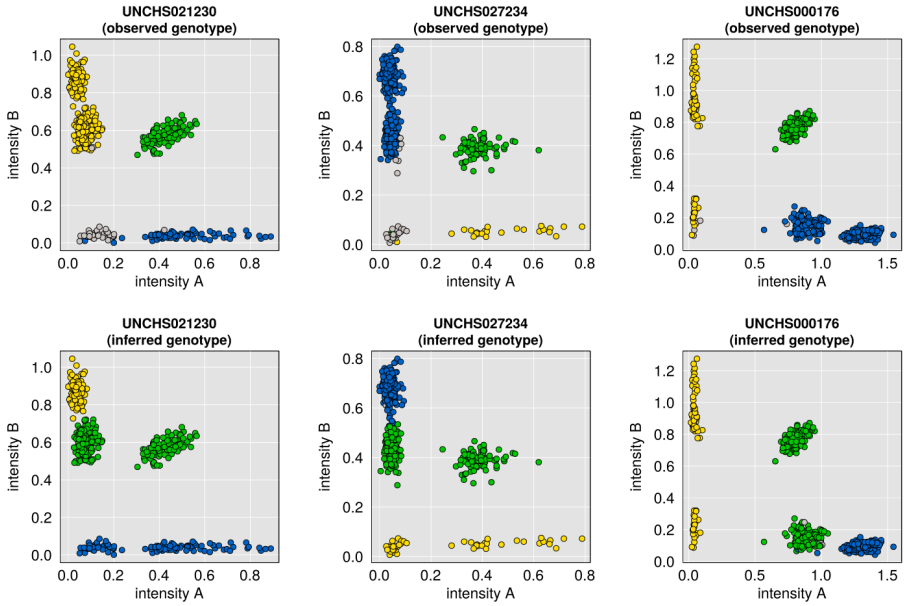
In each case, it's hard to identify the three clear clusters of points, and the genotype calling algorithm clearly did a bad job in the top panels. The bottom panels give some sense about what's really going on, and help to explain why the genotype calling went so badly.

More crap markers



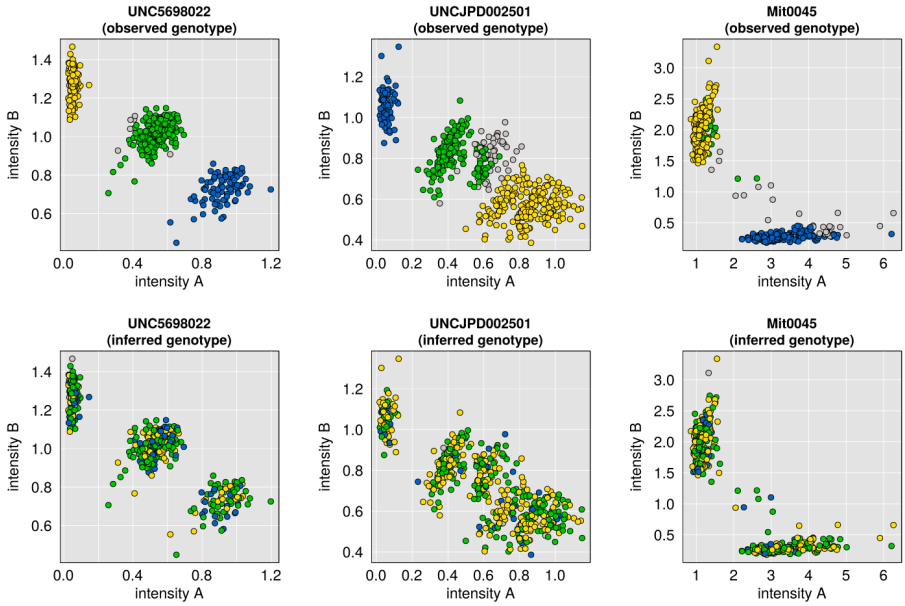
Here are three more badly behaved markers. Again, the lower panels with points colored by the predicted genotypes make some sense and point to the nature of the problems in the genotype calls in the top panels.

One bad blob



Here are a bunch of markers where the genotypes were generally called correctly, but there is one additional cluster of markers that got called incorrectly. These are likely cases where the microarray probe sequence had an additional SNP that was present in some of the founders.

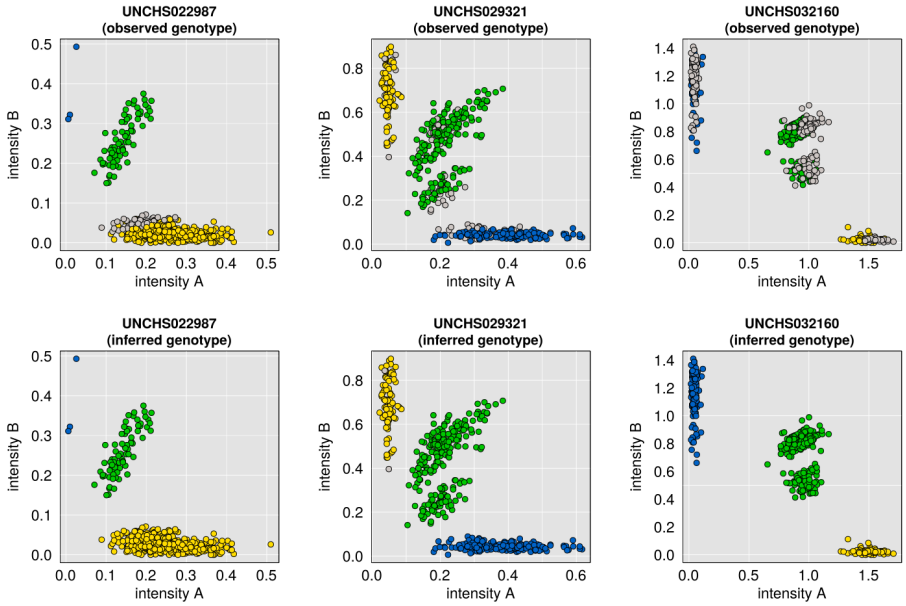
Wrong genomic position



Here are a set of markers that went wrong in a quite different way: the genomic coordinates in the marker annotation information is probably wrong. The marker genotypes are likely correct, but the marker has been placed at the wrong position in the genome, and so the predicted genotypes have no relation to the observed genotypes.

The marker on the right is an extreme case. This is a mitochondrial marker which for some reason was annotated as being on an autosome.

Puzzling no calls

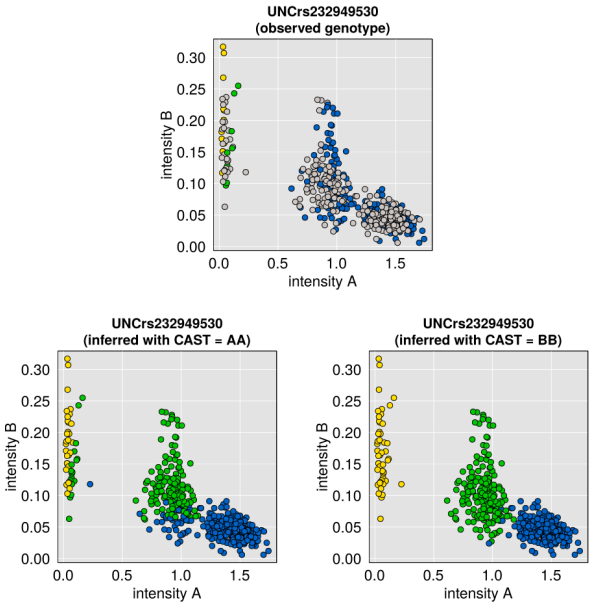


I'm not sure what to make of these markers. They have groups of "no calls" for no apparent reason; it seems like the calling algorithm should have been able to calls these genotypes correctly.

Founder genotyping errors

Another thing to look for is founder genotyping errors. The genotype reconstruction is strongly dependent on the founder genotypes. Can we identify any errors in that data?

One founder missing

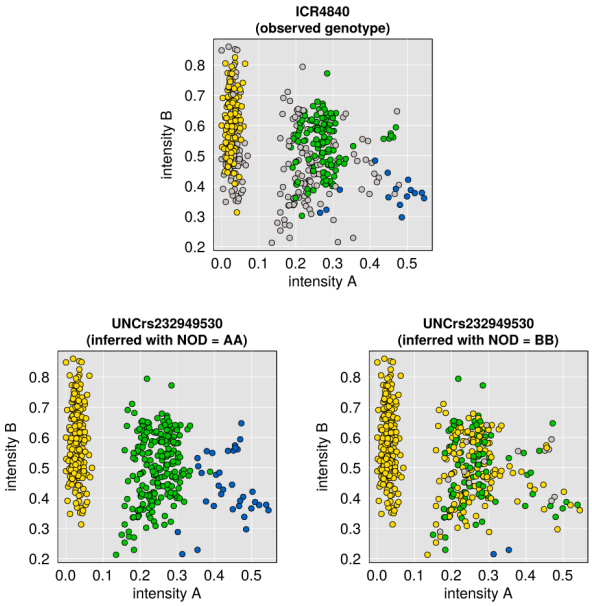


It turned out no; I couldn't find any cases where the founder genotypes looked to be wrong. But there were some cases where one of the founder genotypes were missing. For example, in this case the CAST founder had a missing gneotype.

The top panel shows the called genotypes. The bottom panels show the predicted genotypes if CAST is AA (left) and if CAST is BB (right). It seems clear that the right panel is best and so CAST much be BB at this locus.

And this also shows that we could, in principle, identify an error in the founder genotypes. If CAST had been called AA here, we could maybe see that it was wrong and should be BB.

Another case



Here's another case. The founder NOD is missing at this SNP. If it's AA we get the pattern on the left; if it's BB we get the pattern on the right. It seems clear that NOD is AA, and they if it were erroneously called BB we could in principle detect it.

Principles

- Think about what might have gone wrong, and how it might be revealed
- Order is important; cleaning one aspect may make it hard to see another
- Make lots of graphs
- If you see something weird, try to figure it out
- Don't trust; verify

While the methods I used here are very specific to the problem and data, I think the effort does point to some general principles for data diagnostics, the most important of which is to think about what might have gone wrong and how it might be revealed in the data.

The order in which you do things can be important. Cleaning one aspect of the data and obscure other kinds of problems. As you come up with new ideas of things to look for, you may need to go back to the beginning and start fresh with the raw data.

The other main technique is to just make lots of graphs, and to follow up any oddities you seen. Try to figure out the underlying cause, and whether it is something that could affect the final results or not.

In general, my basic principle is "Don't trust; verify." Cynical, but important.

Summary

- Amount of missing data, as main indicator of problem
- Sex swaps, sample duplicates, sample mix-ups
- Identifying bad samples most important
- Bad samples:
  - Missing data
  - Heterozygosity
  - Number of crossovers
  - Number of genotyping errors
- Bad markers:
  - Missing data
  - Number of genotyping errors
  - Observed vs inferred genotypes

Here's a summary of the things we did. Most important was identifying bad samples, as they have the biggest influence on the results. To look for bad samples, we ended up just focusing on four things: missing data, heterozygosity, number of crossovers, and number of genotyping errors.

Additional thoughts

- You often have to go back to the beginning and start over
- Interactive graphs can speed things up
- Do the work within a reproducible report

My final thoughts: expect that you'll need to be going back to the beginning and starting over. That is one of reasons that it's helpful to do this work within a reproducible report, in which you capture the full process (what you did, what you saw, how you interpreted it, why you made the decisions you made).

Interactive graphs can be super helpful for data cleaning: for identifying outliers, and for identifying which markers or samples you need to look at in more detail. There's nothing that you do with interactive graphs that you couldn't do with a series of static graphs, but that's like saying there's nothing that you can do on a computer that you couldn't do on a hand calculator. The increase in speed in exploring the data can ultimately give a qualitative improvement in the process.