

Sample mix-ups and mixtures in microbiome data in DO mice

Karl Broman

Biostatistics & Medical Informatics
Univ. Wisconsin–Madison

kbroman.org

github.com/kbroman

@kwbroman

Slides: bit.ly/2019CTC



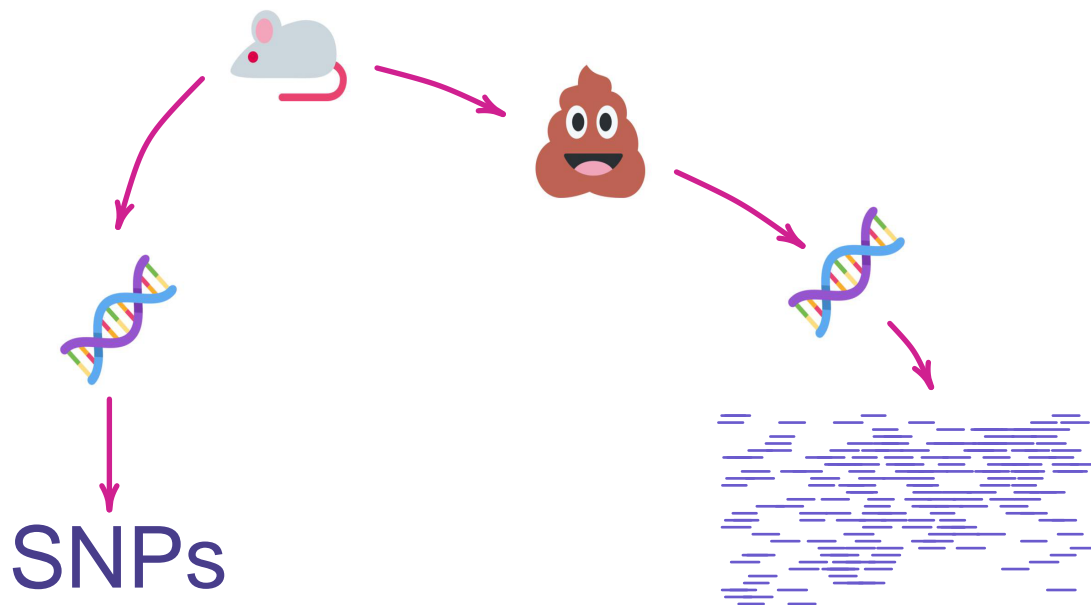
These are slides for a talk I gave at the Complex Trait Community meeting (<http://ratgenes.org/ctc2019>) in San Diego on 10 June 2019.

Source: https://github.com/kbroman/Talk_CTC2019

Slides: <https://bit.ly/2019CTC>

Slides with notes: https://bit.ly/2019CTC_notes

Microbiome genetics data



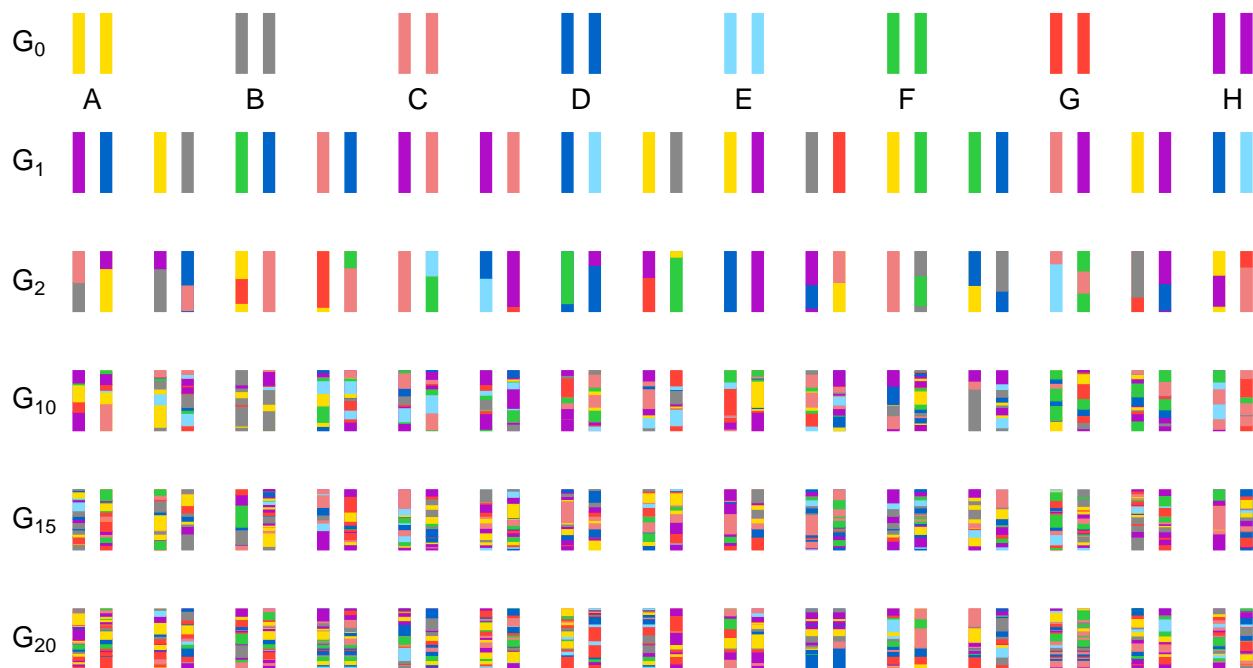
2

As part of a larger project seeking to understand the genetics of diabetes and obesity, we're looking at genetic effects on the microbiome. We have a set of about 500 diversity outbred mice, with SNP genotypes from GigaMUGA arrays. For 300 mice, we have shotgun sequencing data on DNA extracted from mouse poop.

The goal of the microbiome sequencing was to characterize the bacteria in the gut of the mice, but the data also include reads derived from the mouse host. This offers an opportunity to check for sample mix-ups between the microbiome samples and the genomic DNA samples.

We want to relate the genotypes at SNPs across the genome to the alleles observed in the microbiome sequence reads.

Multi-parent advanced intercross population

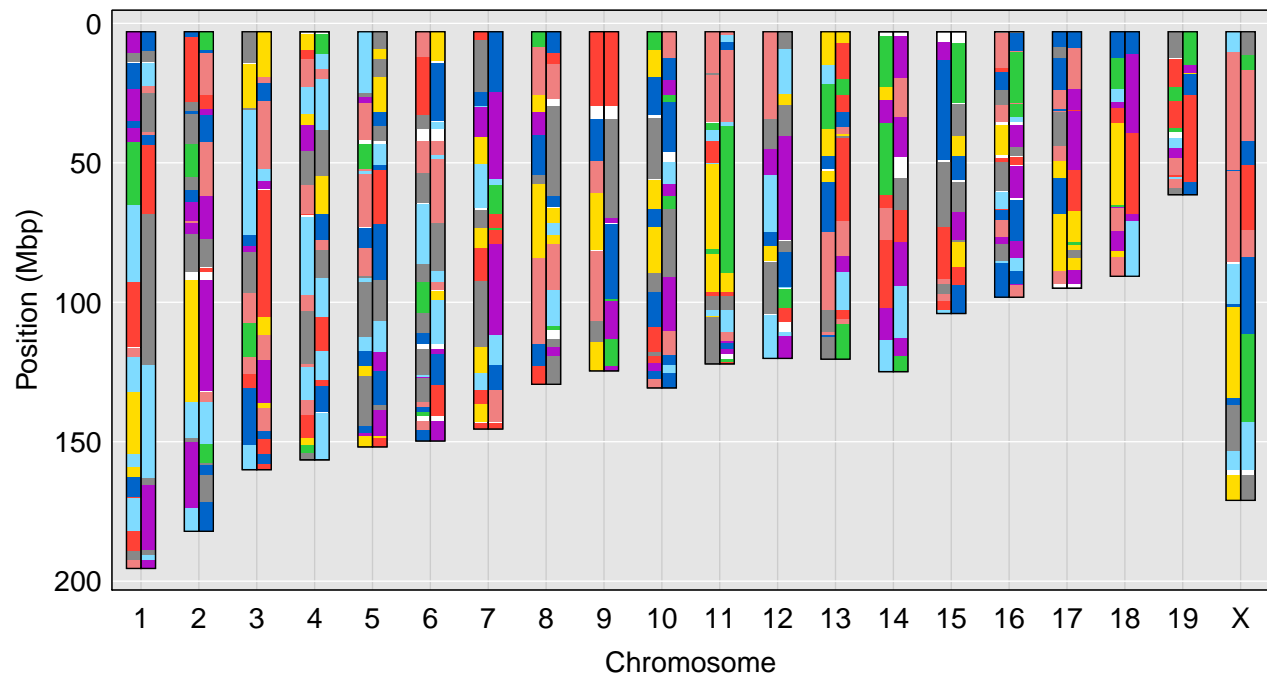


3

Diversity outbred mice, like heterogeneous stock, are an example of a multi-parent advanced intercross population. Eight founder strains were inter-bred for many generations, maintaining as large a population as is feasible at each generation and avoiding crosses between siblings, to avoid inbreeding and genetic drift.

The resulting population is a heterogeneous mixture of the initial eight strains, with the chromosomes broken down into small pieces.

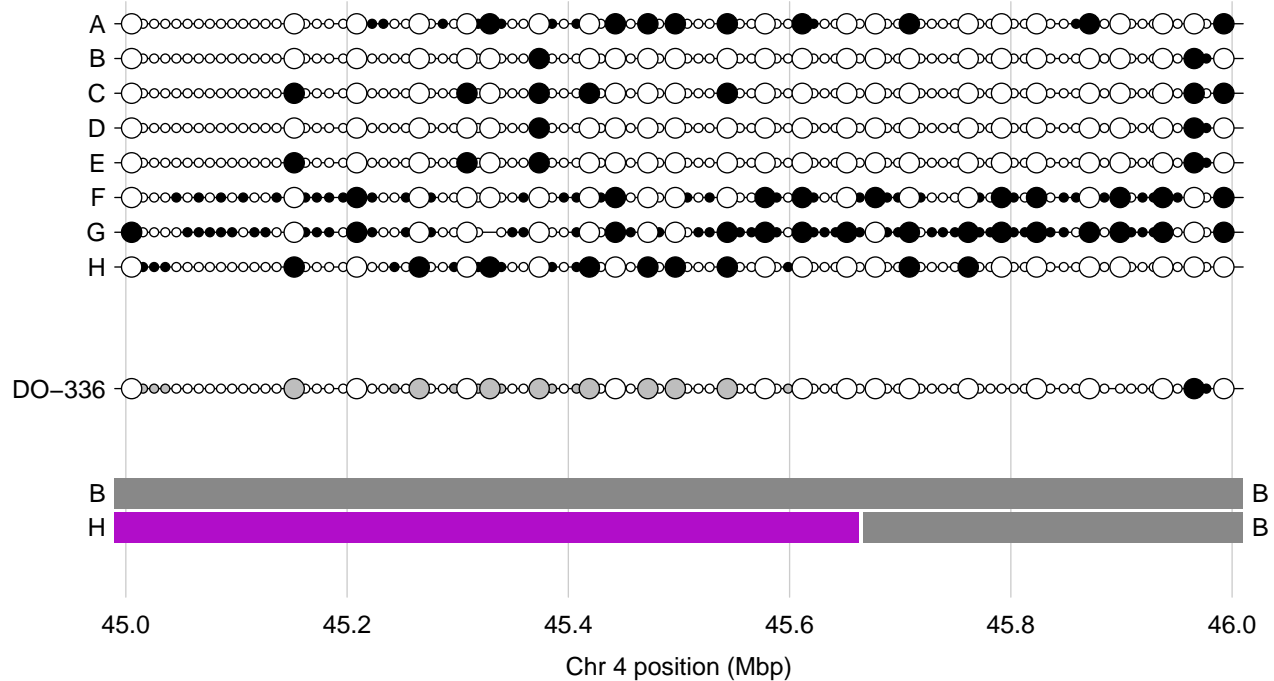
Genome of a diversity outbred mouse



4

This is an example of the genome of one DO mouse. At any one position, they have one of 36 possible genotypes. The white patches are regions where genotype is uncertain.

Genotype reconstruction



5

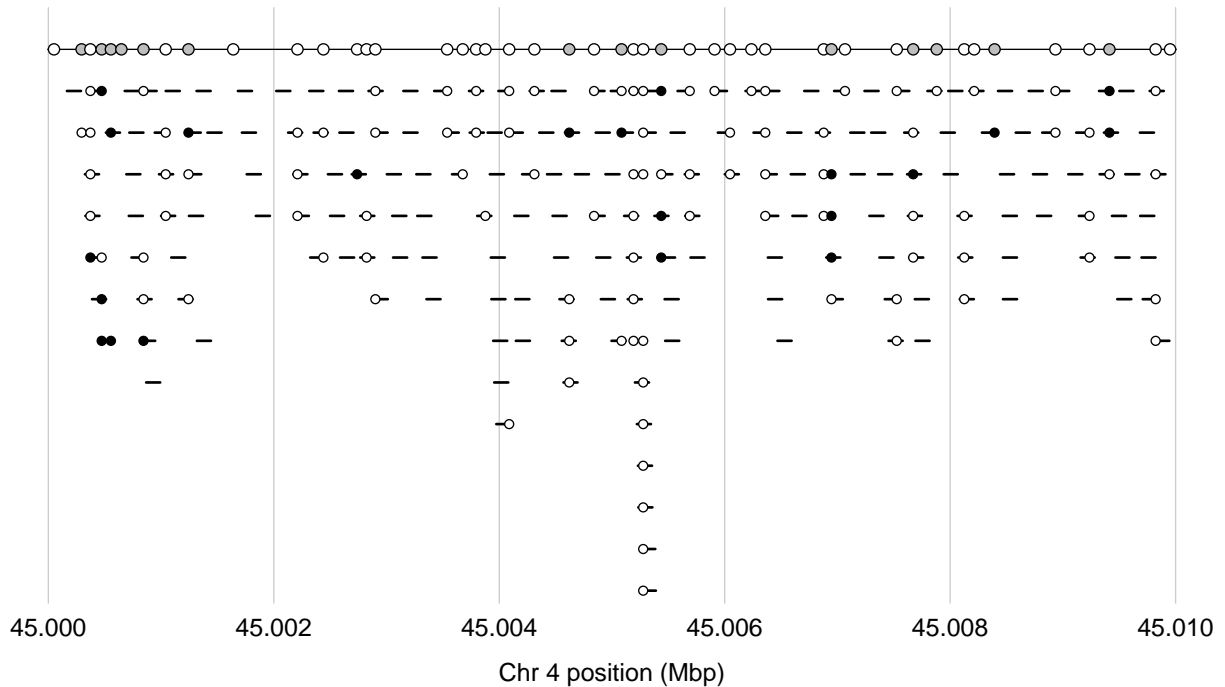
Our first step is to reconstruct the founder genotypes along the genome of each DO mouse, using SNP data on DO mice and the eight founder strains. We use a hidden Markov model for this purpose, which allows for the presence of genotyping errors.

The white, gray, and black circles indicate SNP genotypes AA, AB, and BB, with A being the major allele and B being the minor allele (based on frequency in the eight founder strains).

The eight founder strains have been sequenced, and so we know their genotype at about 40 million polymorphisms in the genome. We can use those genotypes plus the DO genotype reconstructions to infer the DO mouse genotypes at all SNPs.

This is important because we want to compare sequence reads to the SNP genotypes, and while many sequence reads will overlap a SNP, few will overlap one of the SNPs on the GigaMUGA array, for which we have direct DO genotype data.

Mapped reads



6

The second part of the process is to map sequence reads to the mouse genome. We then look at the SNPs in a region, identify which reads overlap a SNP, and count the observed alleles at the different SNPs.

If a DO mouse really corresponds to that microbiome sample, all reads should be A wherever the mouse has genotype AA, and B where the mouse is BB. At SNPs where the mouse is AB, half the reads should be A and half should be B.

There will be some errors in the sequence reads, such as the black dot in the interval 45.002–45.004, corresponding to an allele B in a read at a SNP where the mouse is AA.

Genomic DNA vs microbiome reads

genomic DO-381 vs microbiome DO-381

genomic DNA	microbiome DNA	
	A	B
AA	2,762,341	7,303
AB	606,312	578,017
BB	2,128	375,559

percent mismatch = 0.3%

7

Our main strategy is to split all SNPs according to their genotype in a genomic DNA sample, with AA meaning homozygous for the major allele (among the eight founding strains). For a microbiome sample, we count the number of overlapping reads with the A vs. the B allele.

If the microbiome sample really comes from that mouse, the reads should be largely A at the AA SNPs, and largely B at the BB SNPs, and about 50:50 A vs B at the heterozygous SNPs.

In this particular case, the percent discordant reads in the homozygous SNPs is 0.3%. You could take that as an estimate of the sequencing error rate, assuming there's no mix-up here, and that the genomic DNA genotypes are correct.

Genomic DNA vs microbiome reads

genomic DO-360 vs microbiome DO-360

genomic DNA	microbiome DNA	
	A	B
AA	8,863,572	1,520,169
AB	2,870,063	1,075,126
BB	671,722	536,010

percent mismatch = 19%

This case, of DO-360 genomic DNA vs DO-360 microbiome, shows a clearly different pattern: 19% discordant reads in the homozygous SNPs, and the heterozygous SNPs show A:B like 3:1.

Genomic DNA vs microbiome reads

genomic DO-370 vs microbiome DO-360

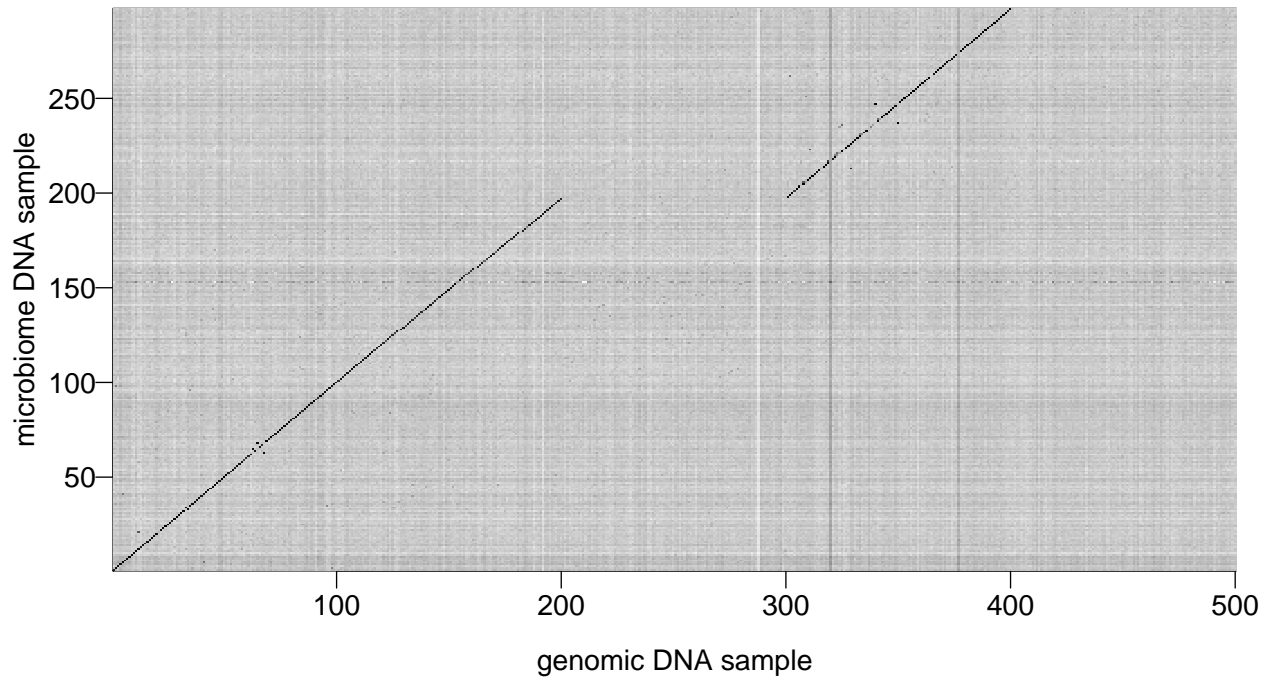
genomic DNA	microbiome DNA	
	A	B
AA	10,324,265	23,256
AB	2,083,947	1,986,380
BB	5,347	1,117,994

percent mismatch = 0.2%

If we compare the DO-360 microbiome to genomic DNA for DO-370, however, we see nice concordance.

While the heterozygous category does contain information, we're going to just focus on homozygous SNPs and use the percent discordant reads as a measure of distance.

Distance matrix

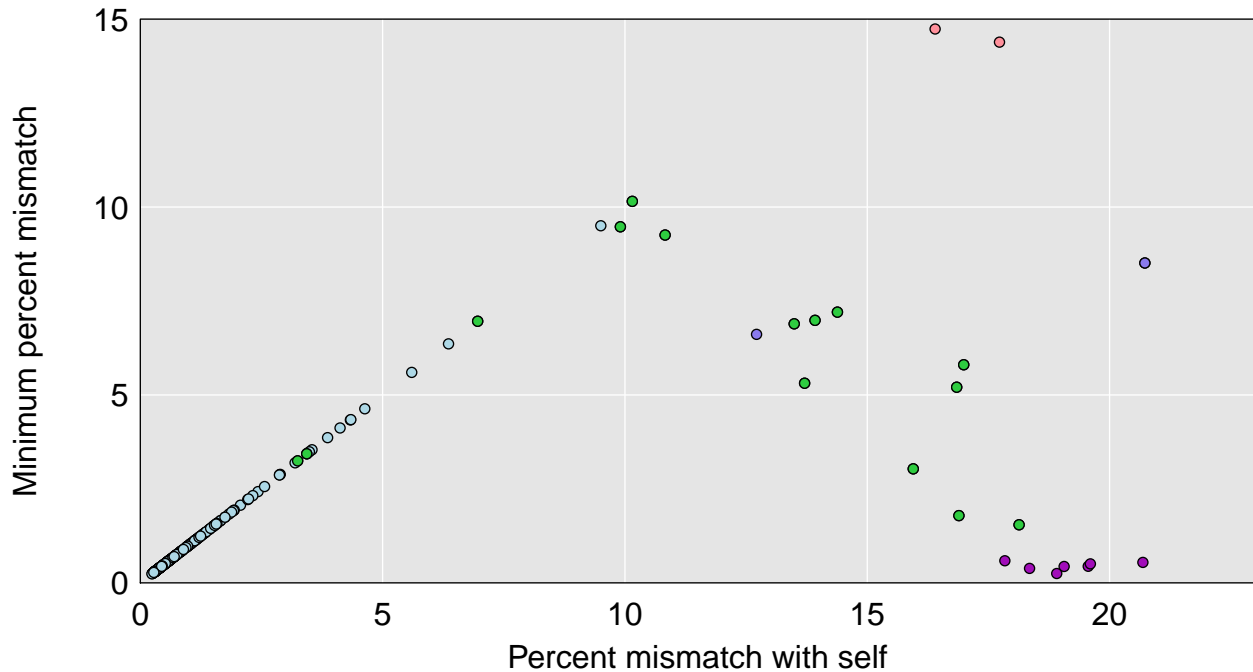


10

This shows the distances between microbiome samples (on y-axis) and genomic DNA samples (on x-axis), with black indicating the samples are similar, and white indicating they are different.

There were a total of 500 DO mice, in five batches of 100 mice each. The microbiome study included the first, second, and fourth batches. The black diagonal line indicates that most of the samples are correct; but if you look closely you can see some problems.

Minimum vs. self distance



11

Here, we plot the minimum distance for each microbiome sample (the minimum value in each row in the distance matrix) vs. the self-self distance (the diagonal in the distance matrix).

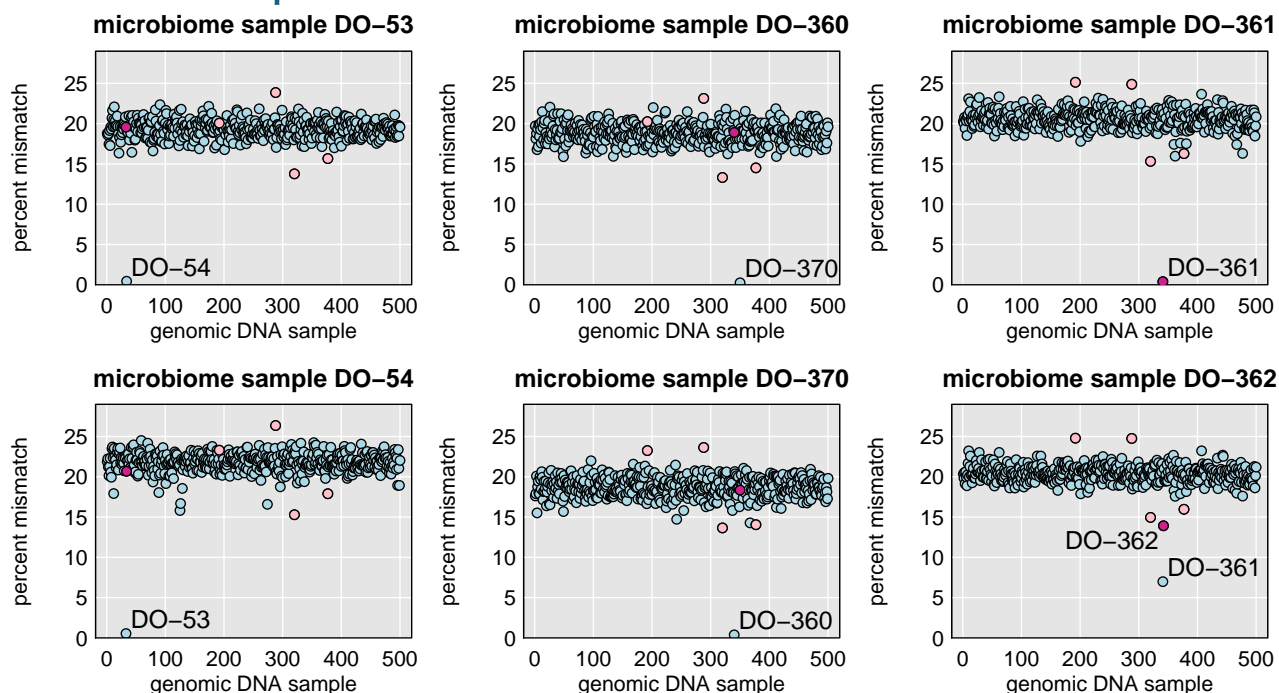
Samples on the diagonal here (in light blue) look to be correctly labeled: they are similar to the corresponding genomic DNA sample and no other sample is closer.

Samples in the lower-right corner (purple) are mix-ups. They are very different from the correspondingly labeled sample, but there is another sample that is quite close.

Samples in green turn out to be mixtures (more on this below).

Samples in pink (upper right) have bad genomic DNA samples. Samples in light purple have low read counts.

Selected samples



12

These are the results for selected samples, of the distance between a microbiome sample and all 500 genomic DNA samples (the values along a row of the distance matrix).

The dark pink dot is for the genomic DNA sample with the sample label. The light pink dots are a set of low-quality genomic DNA samples.

The left panels are a clear mix-up between DO-53 and DO-54. The center panels are a clear mix-up between DO-360 and DO-370. We can't tell whether the mix-ups are in the microbiome samples or the genomic DNA samples, though for DO-360 and DO-370, we have RNA-seq data, and the sample swap is seen there, too. Thus, we can conclude for DO-360 and DO-370, the mix-up was in the genomic DNAs.

The upper-right panel shows that DO-361 is well behaved.

The lower-right panel is for the microbiome sample DO-362. It is most similar to genomic DNA sample DO-361, but it's not too similar, and the second-most similar genomic DNA sample is the one with the same label as the microbiome sample, DO-362. This suggested that the microbiome sample is perhaps a mixture between DO-361 and DO-362.

Genotype pair vs microbiome reads

genomic DO-362 and DO-361 vs microbiome DO-362

DO-362: AA			DO-362: AB			DO-362: BB		
DO-361	A	B	DO-361	A	B	DO-361	A	B
AA	99.7%	0.3%	AA	84.0%	16.0%	AA	67.7%	32.3%
AB	67.6%	32.4%	AB	51.3%	48.7%	AB	34.3%	65.7%
BB	34.7%	65.3%	BB	17.1%	82.9%	BB	0.6%	99.4%

13

To assess whether the microbiome sample DO-362 is in fact a mixture, we can divide SNPs into nine groups according to the genotypes of DO-362 and DO-361, and then count the number of A and B reads in each group.

It should be that the frequency of A vs. B reads in the microbiome sample DO-362 only depends on the genotype of DO-362 and not on the genotype of DO-361. But as shown on this slide, there is a very strong dependence on the genotype of DO-361.

For example, when DO-362 is AA, there should be a small proportion of B reads, irrespective of what the genotype of DO-361 is. But when DO-361 is BB, there are about 65% B reads. It seems clear that the DO-362 microbiome sample is contaminated with DNA from DO-361, and in fact it could be about 65% from DO-361.

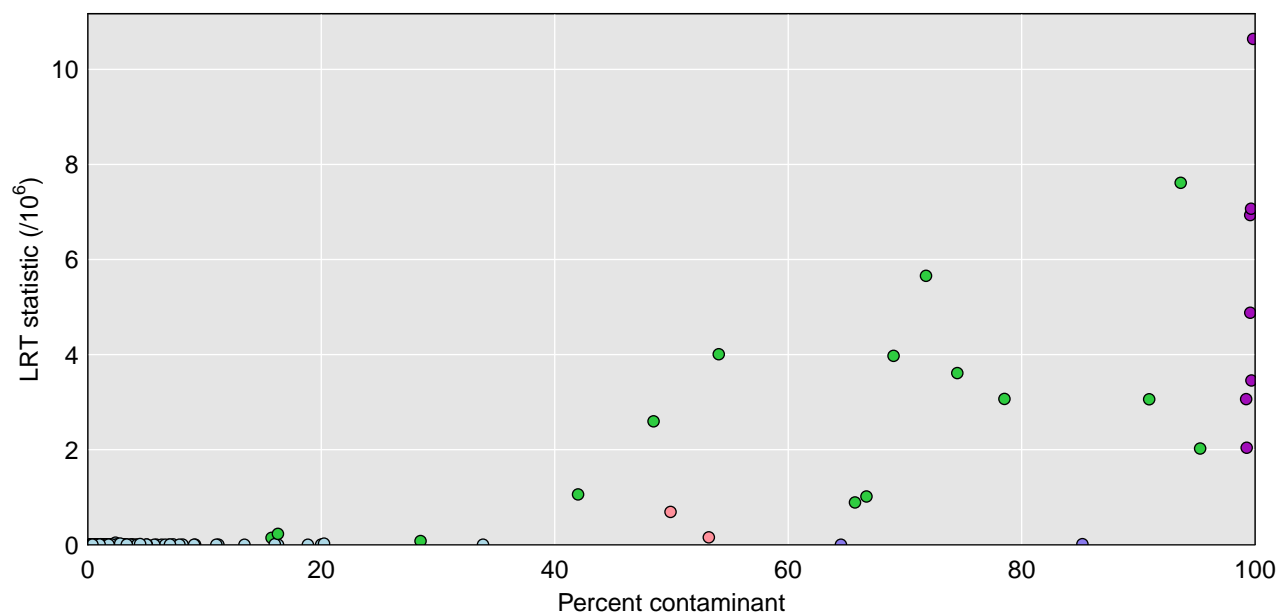
Genotype pair vs microbiome reads

genomic DO-361 and DO-362 vs microbiome DO-361

DO-361: AA			DO-361: AB			DO-361: BB		
DO-362	A	B	DO-362	A	B	DO-362	A	B
AA	99.7%	0.3%	AA	51.4%	48.6%	AA	1.0%	99.0%
AB	99.6%	0.4%	AB	50.9%	49.1%	AB	0.8%	99.2%
BB	99.5%	0.5%	BB	50.6%	49.4%	BB	0.7%	99.3%

Here's a similar table for the DO-361 microbiome sample. This is what a clean microbiome sample should look like. But note that there is still some small association with the genotype of the DO-362 sample.

log likelihood vs percent contaminant



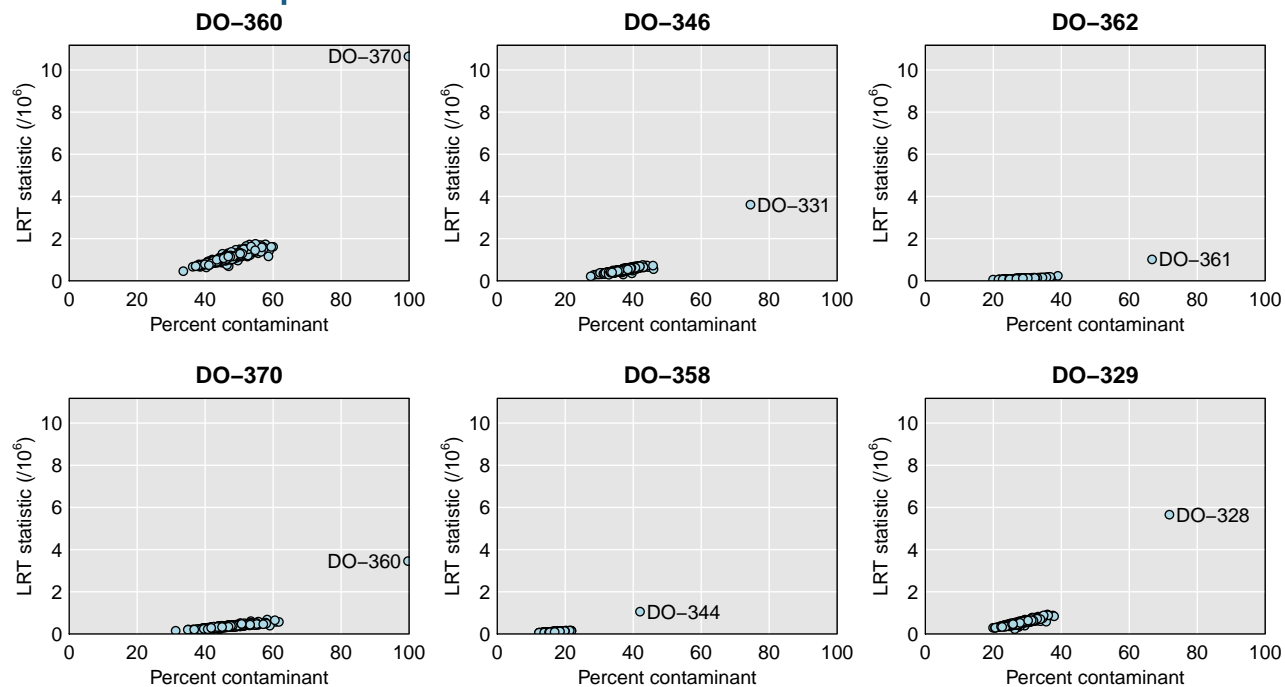
15

To evaluate the potential for mixtures more formally, we fit a model where we assume that a microbiome sample is a mixture of its own DNA and some proportion p of DNA from one other sample, and that the number of A and B microbiome reads are binomial counts with sequencing error rate e .

This figure shows the estimated percent contaminant (on the x-axis) and the likelihood ratio statistic for the test of $p = 0$ (on the y-axis). On the far right (in purple) are the sample mix-ups. In green are a variety of mixtures, most with $> 50\%$ coming from the contaminant.

Note that the likelihood ratio statistics are huge. We're looking for something like 10, and we're getting values > 1 million.

Selected samples



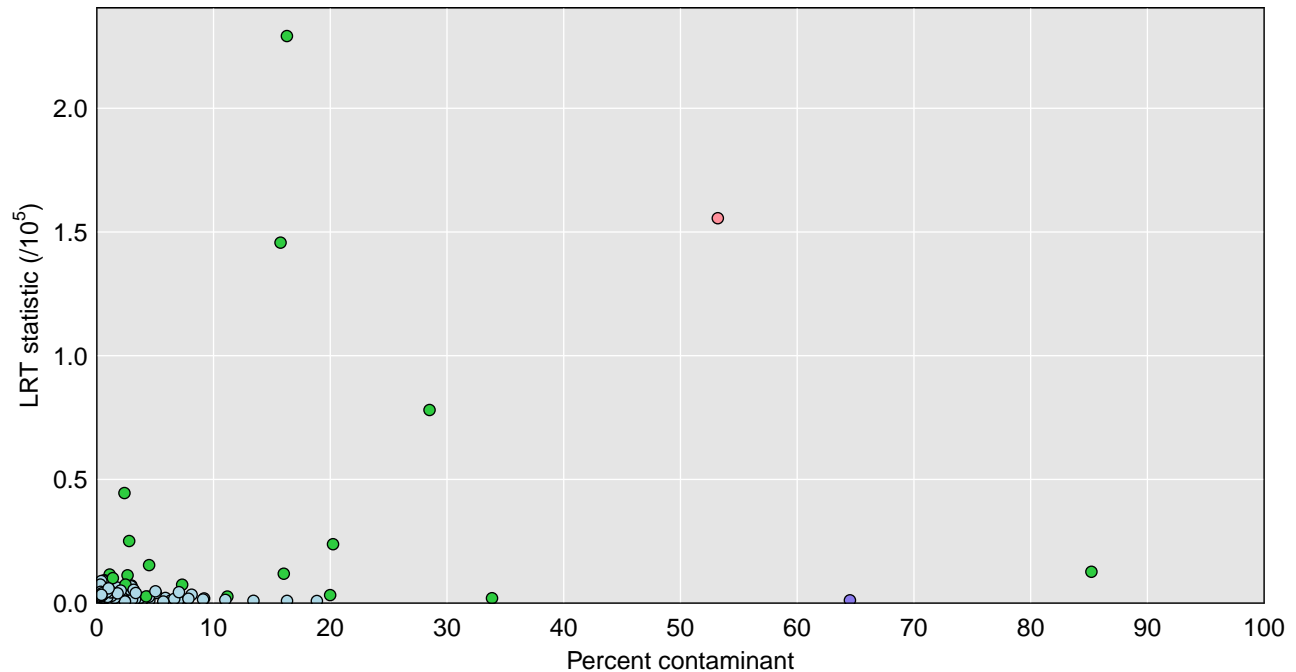
16

Here are results for a selected set of microbiome samples. Each sample considers the microbiome sample as a mixture of the corresponding genomic DNA sample and one contaminant, and each point is one of the 499 possible contaminants.

Each panel shows the likelihood ratio test statistic for the test of $p = 0$ (on the y-axis) vs. the estimated percent contaminant (on the x-axis). In all cases, a single contaminant stands out.

The left panels are for one of the mix-ups. The other four panels are for different apparent mixtures.

Is everything contaminated?



17

If we expand the lower-left corner of the overview figure, we find lots of samples that have strong evidence for being mixtures. Note that the LRT statistics on the y-axis have been divided by 10,000.

It's hard to draw a line on what is a real mixture and what is just noise, and we probably don't need to worry about cases that are < 5% contaminant; for studying the genetics of the microbiome composition, that could be viewed as acceptable phenotypic noise.

Summary

- ▶ Microbiome shotgun reads include reads from the host
- ▶ With such data, sample mix-ups can be identified
- ▶ Simple method:
 - Impute genotype at all SNPs
 - Count alleles in reads overlapping SNPs
 - Focus on homozygous SNPs and calculate percent discordant reads
- ▶ We also saw strong evidence for many samples being mixtures

It is always important to provide a summary.

Acknowledgments

- ▶ Lindsay Traeger
- ▶ Alexandra Lobo
- ▶ Federico Rey
- ▶ Alan Attie, Mark Keller, Gary Churchill, Brian Yandell
- ▶ NIH: NIDDK, NIGMS

Lindsay had the idea to look for these mix-ups. (She was a postdoc with Federico Rey in Microbiology at UW–Madison.) Alexandra did most of the work. (She was a summer student with me and now is a graduate student in the Biomedical Data Science PhD program at UW–Madison.)

This is part of a larger project looking at the genetics of diabetes, obesity, and related traits.

Slides: bit.ly/2019CTC



bit.ly/2019CTC_notes

bioRxiv manuscript: doi.org/10.1101/529040

kbroman.org

github.com/kbroman

@kwbroman

Here's where you can find me, as well as the slides for this talk.

Also note that there is a bioRxiv manuscript describing the details of this work.