

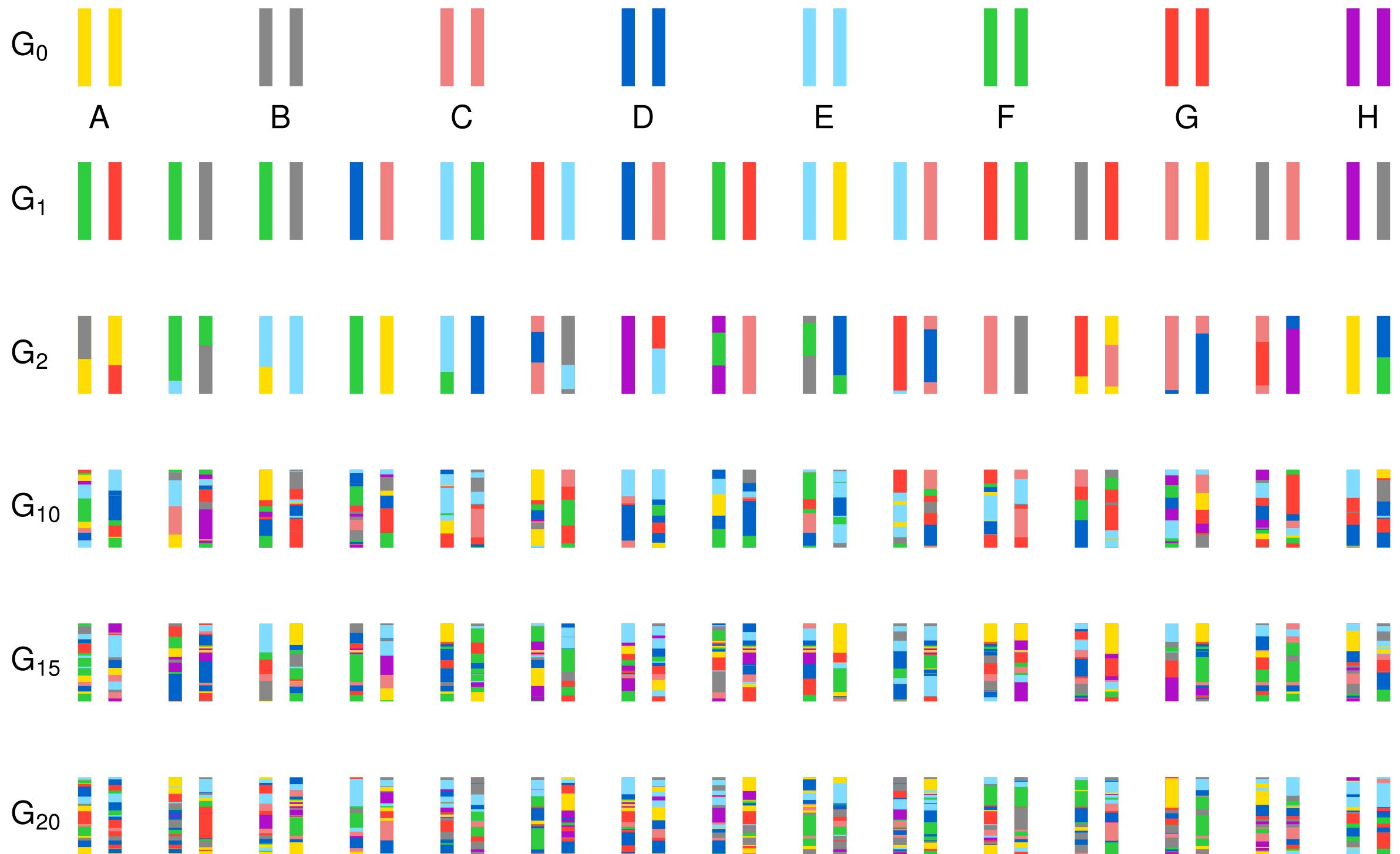
Cleaning genotype data for diversity outbred mice

Karl Broman

Biostatistics & Medical Informatics
University of Wisconsin–Madison

kbroman.org
github.com/kbroman
[@kwbroman](https://twitter.com/kwbroman)
Slides: bit.ly/2018ctc

Heterogeneous stock



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

Principles

What might have gone wrong?

How could it be revealed?

Principles

What might have gone wrong?

How could it be revealed?

Also, just make a bunch of graphs.

Principles

What might have gone wrong?

How could it be revealed?

Also, just make a bunch of graphs.

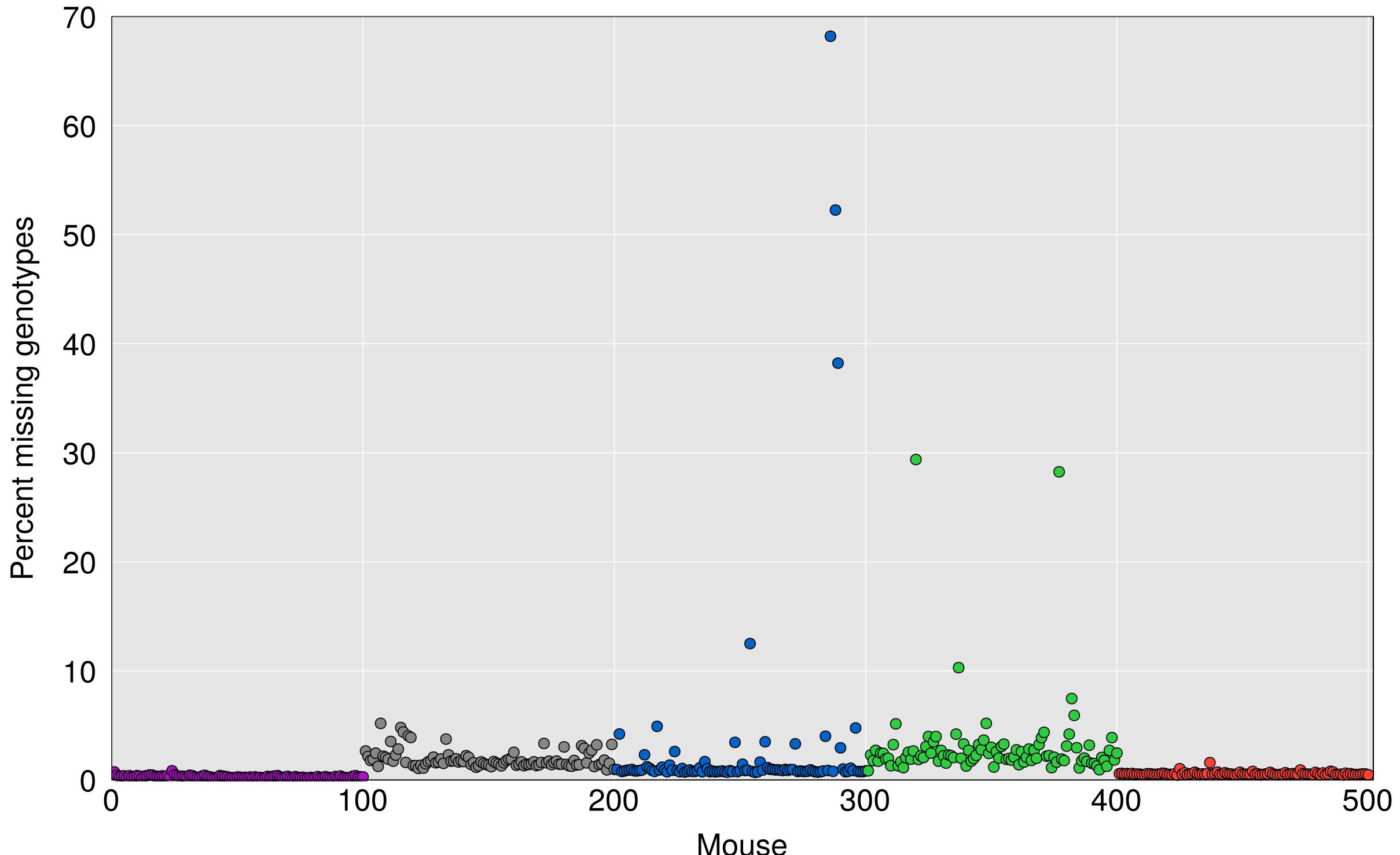
If you see something weird, try to figure it out.

Possible problems

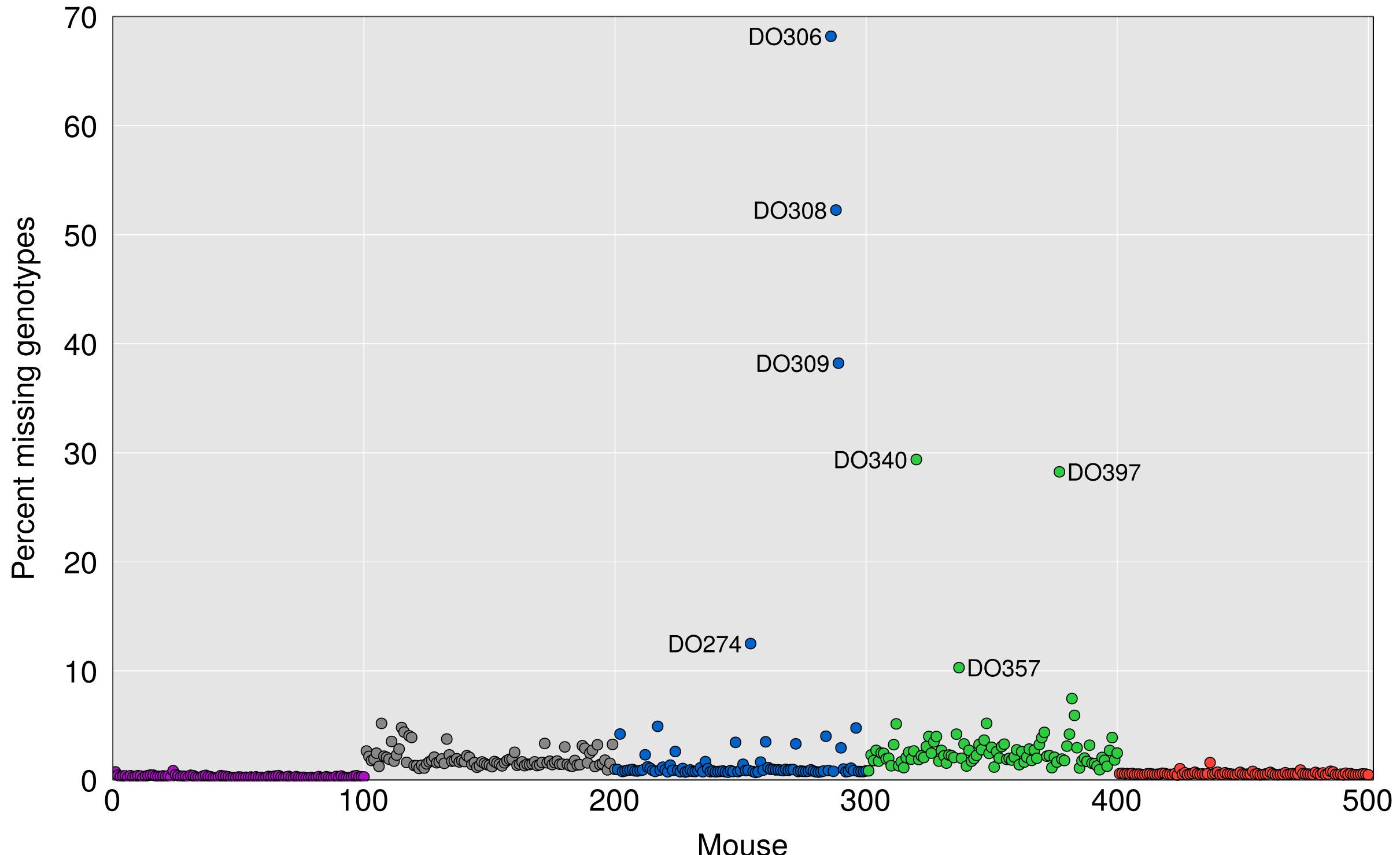
- Sample duplicates
- Sample mix-ups
- Bad samples
- Bad markers
- Genotyping errors in founders

What to look at first?

Missing data per sample

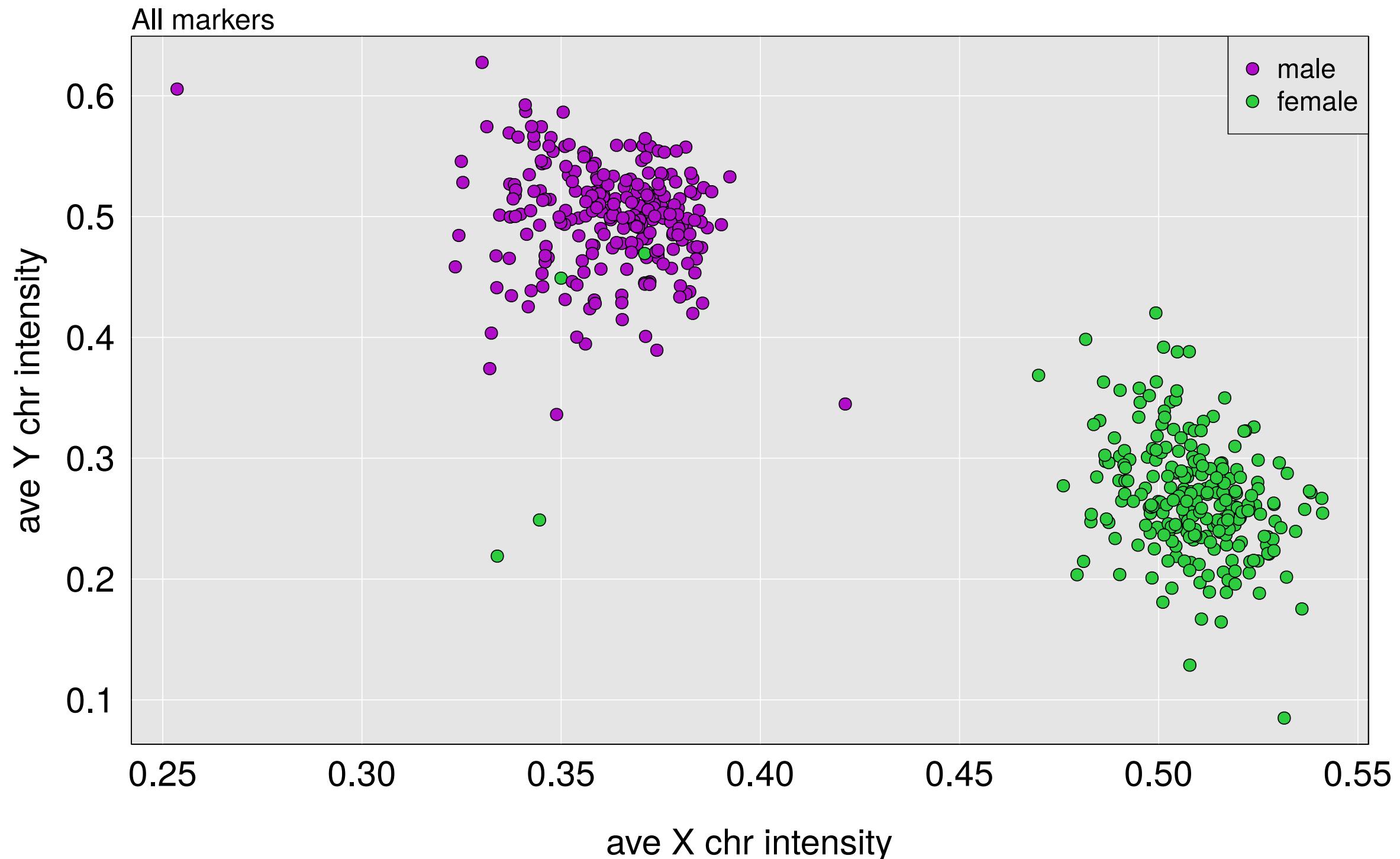


Missing data per sample

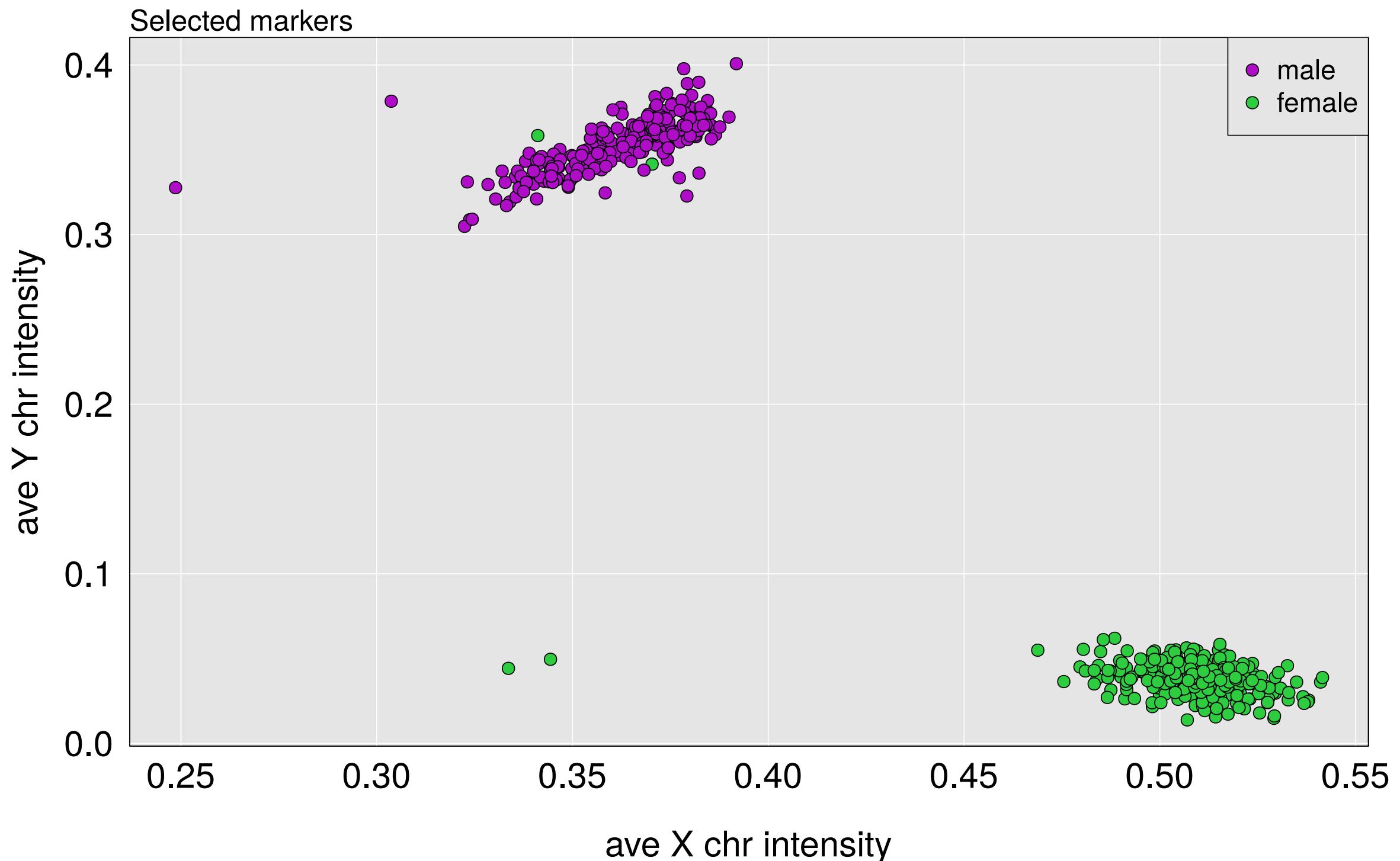


Swapped sex labels

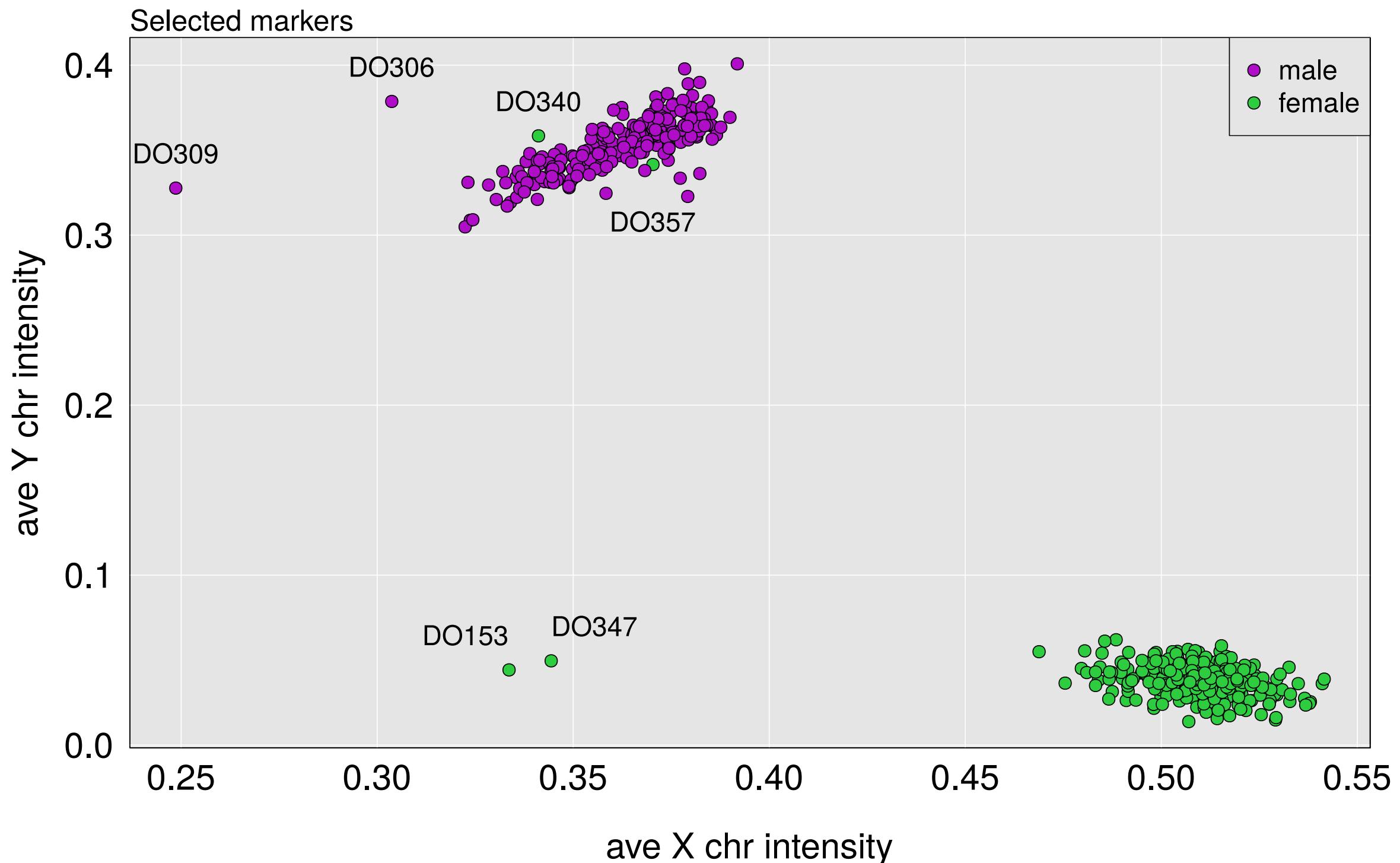
Average SNP intensity on X and Y chr



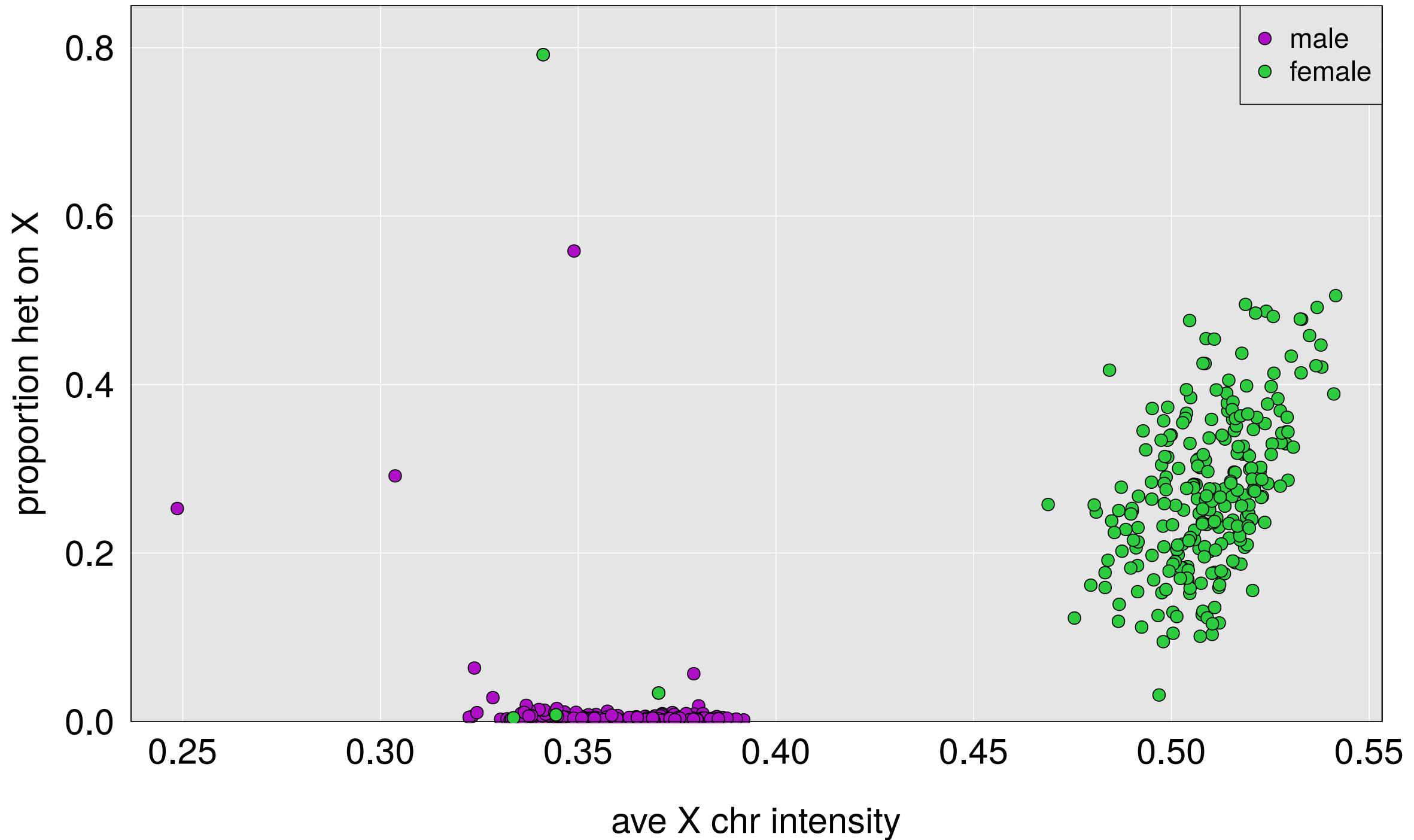
Average SNP intensity on X and Y chr



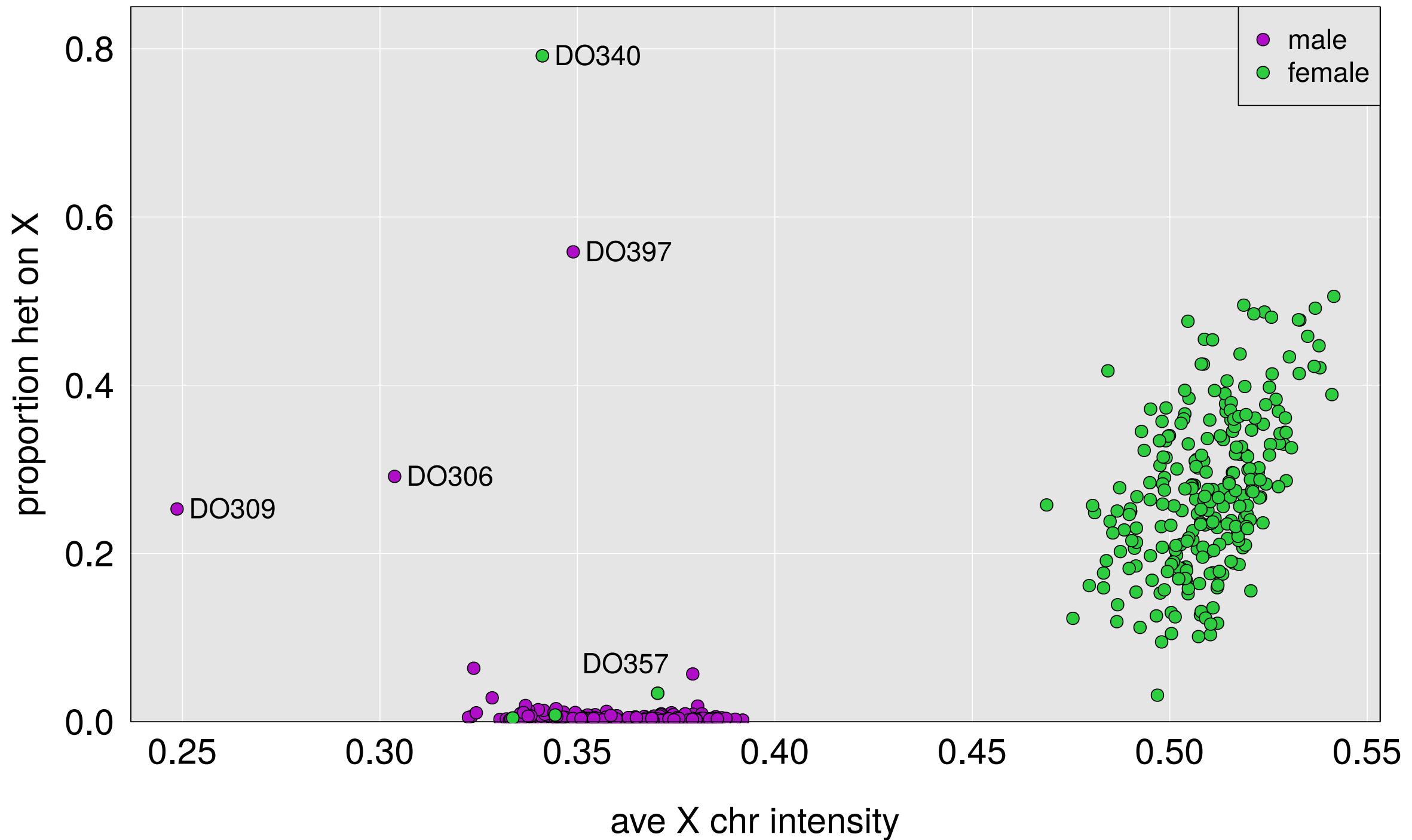
Average SNP intensity on X and Y chr



Heterozygosity vs SNP intensity on X chr

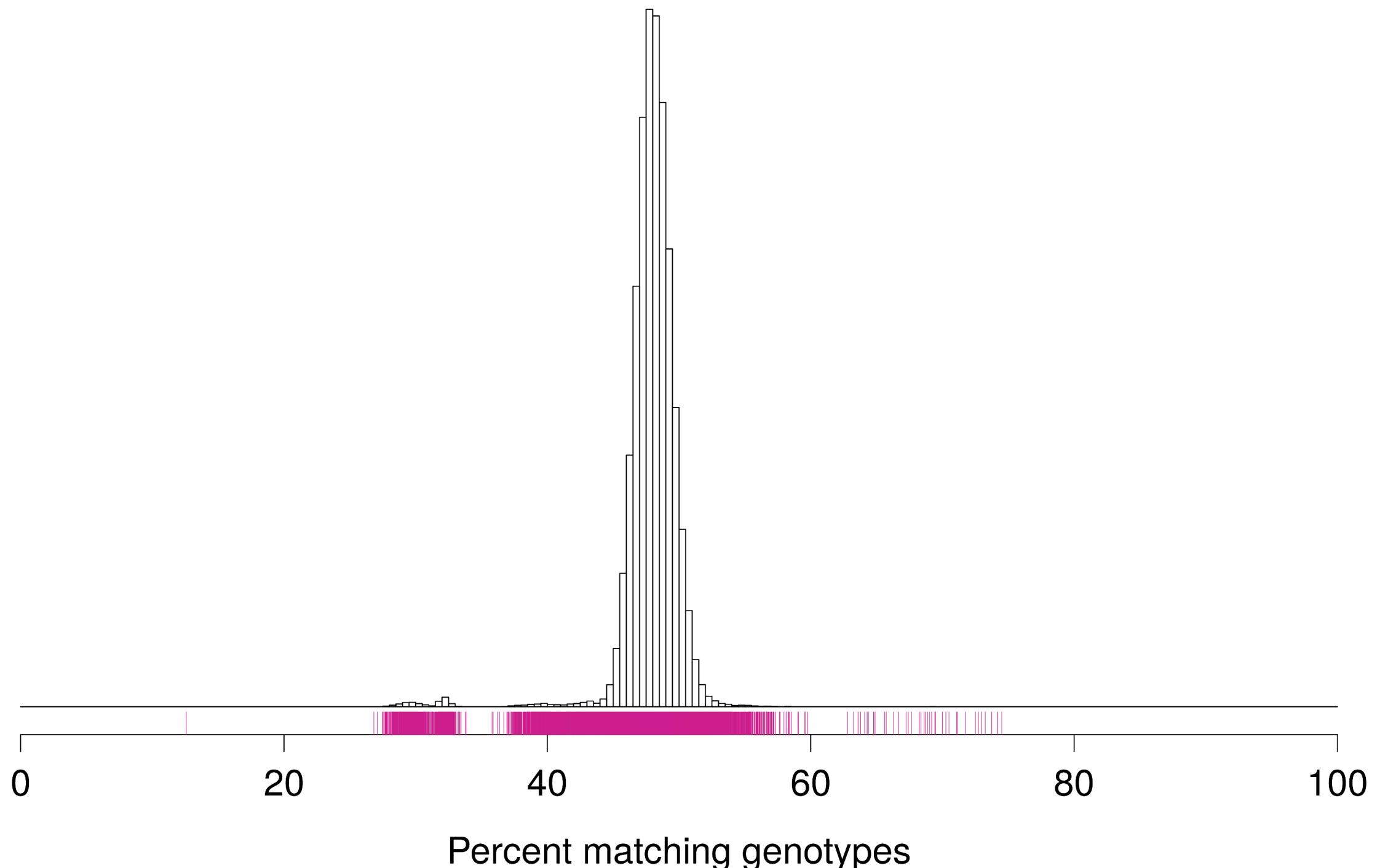


Heterozygosity vs SNP intensity on X chr

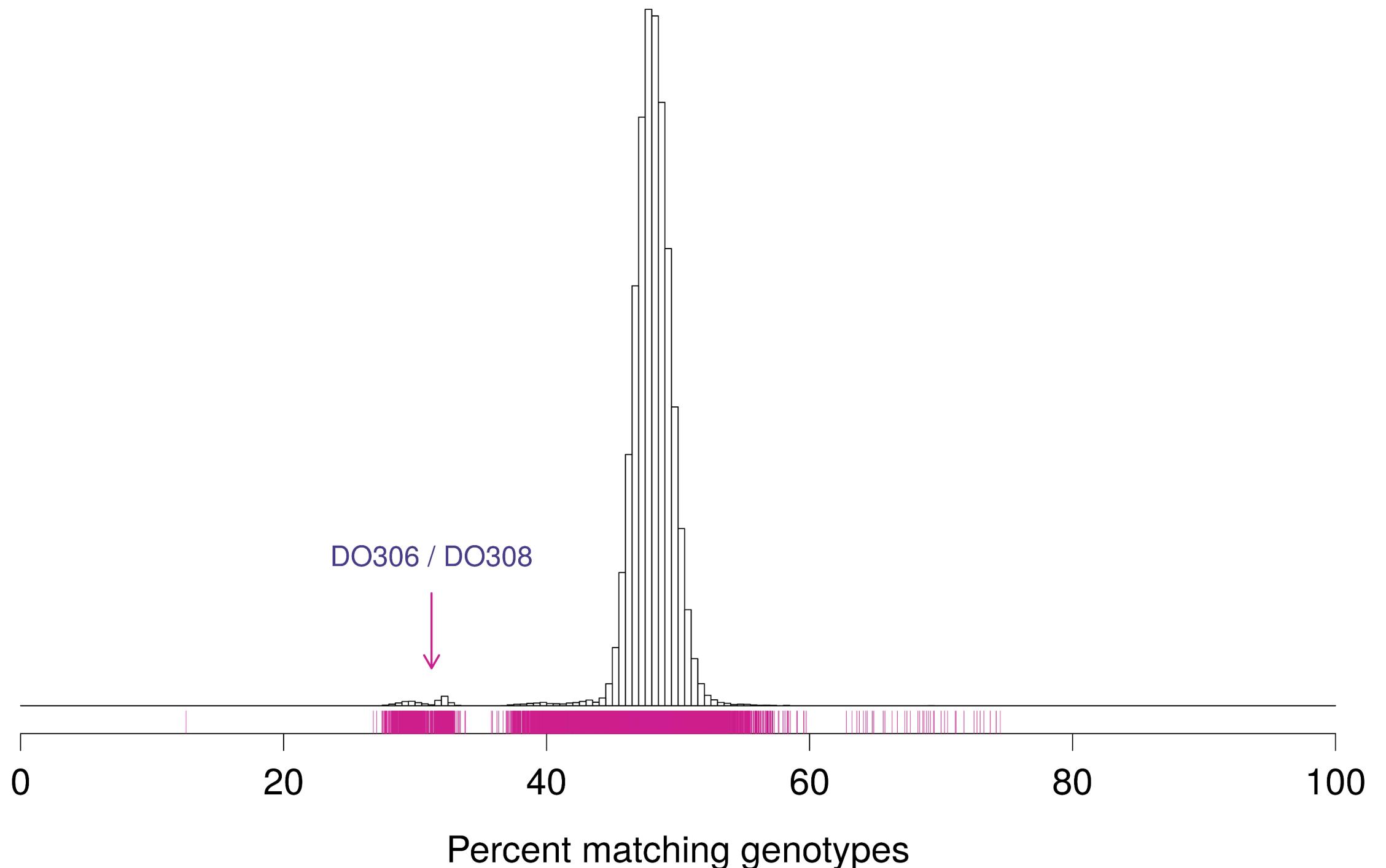


Sample duplicates

Percent matching genotypes between pairs

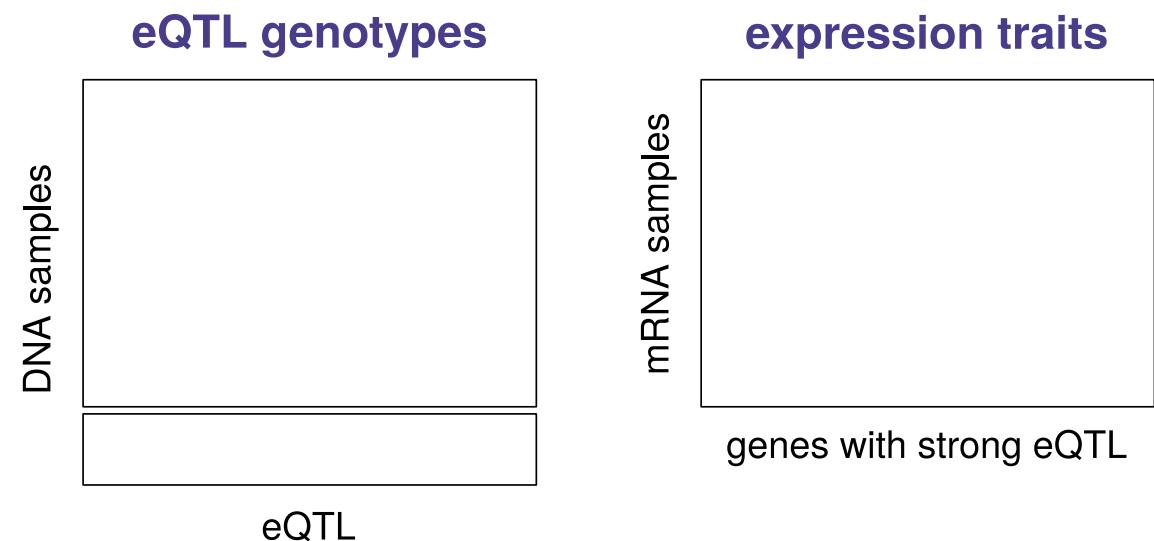


Percent matching genotypes between pairs

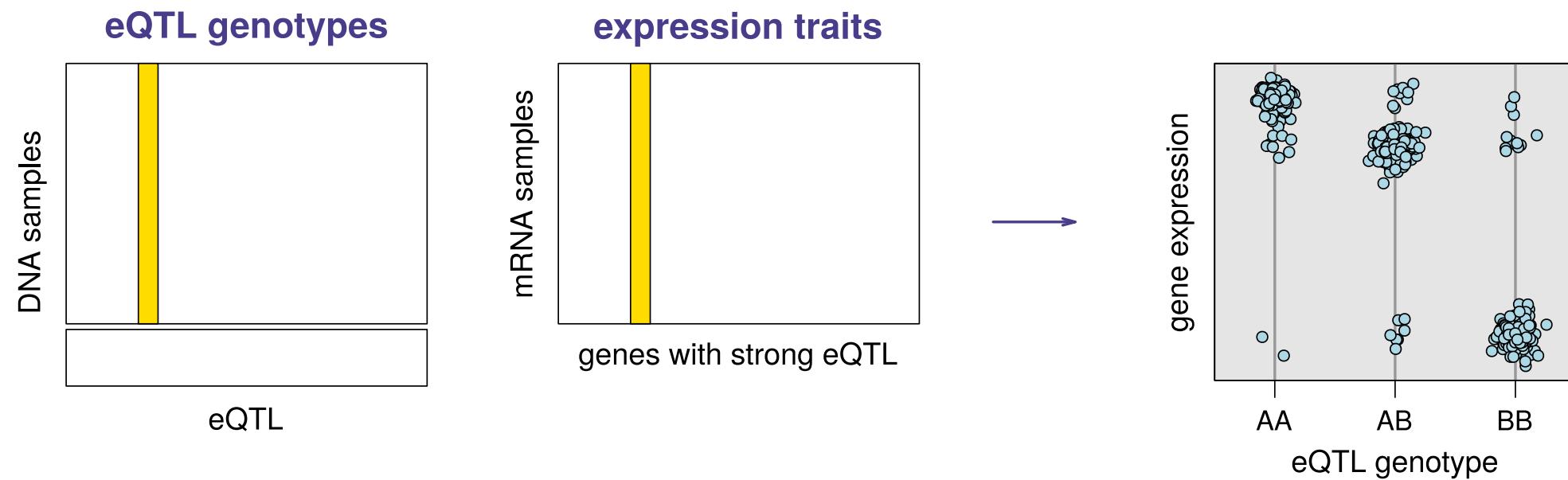


Sample mix-ups

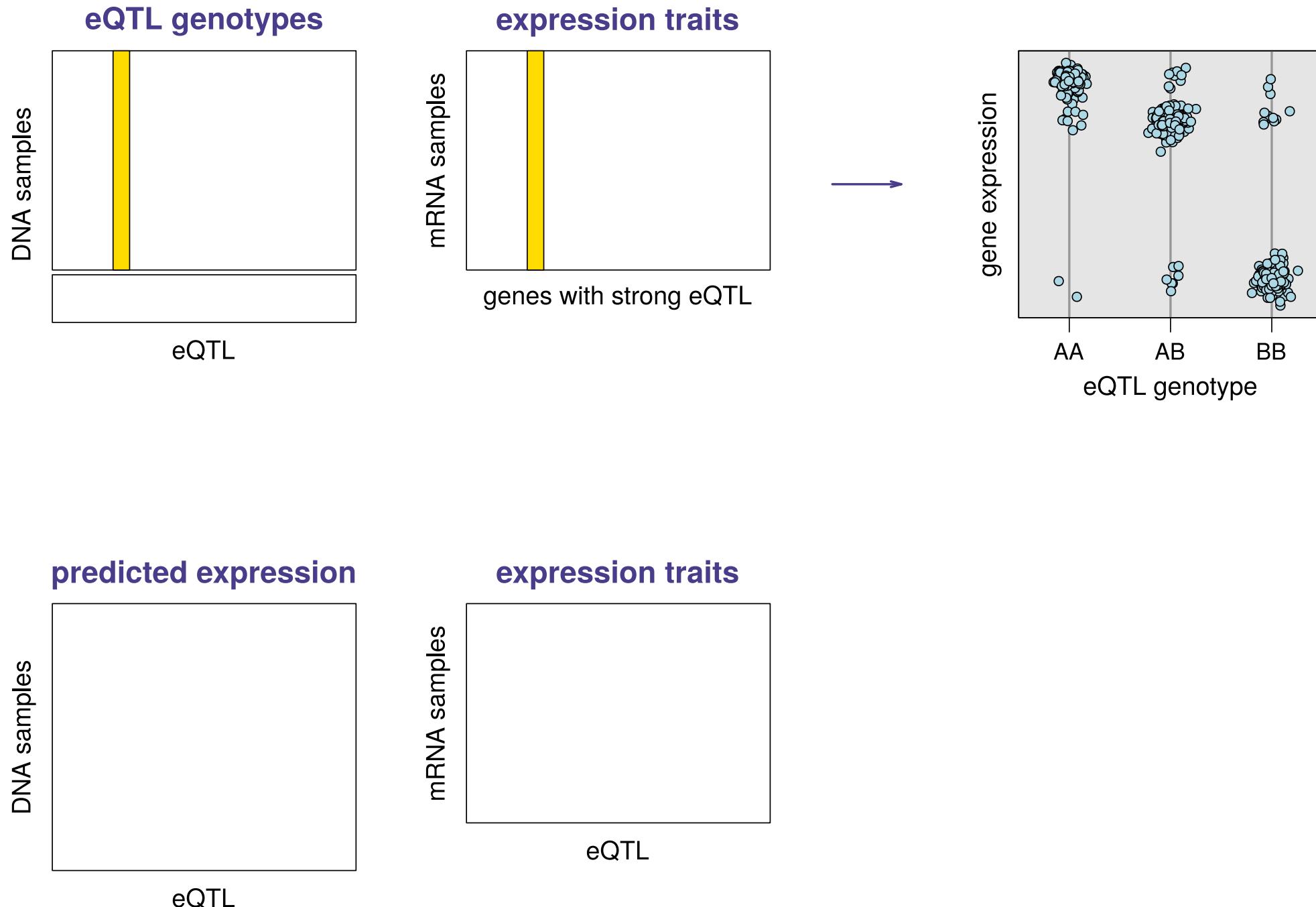
Sample mix-ups



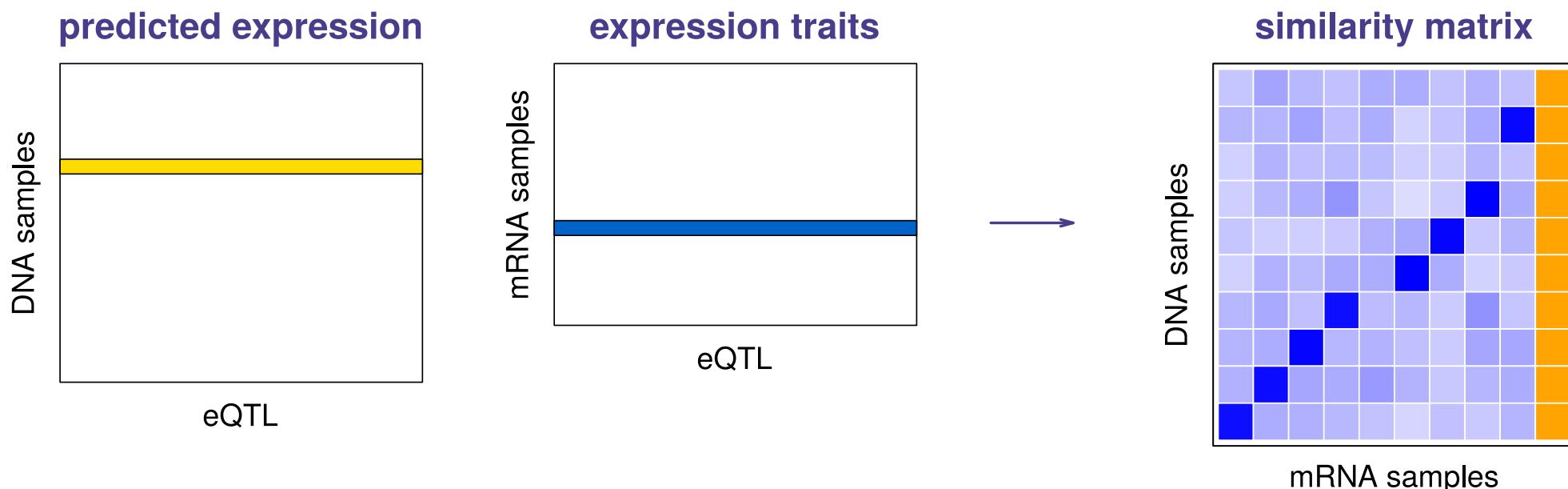
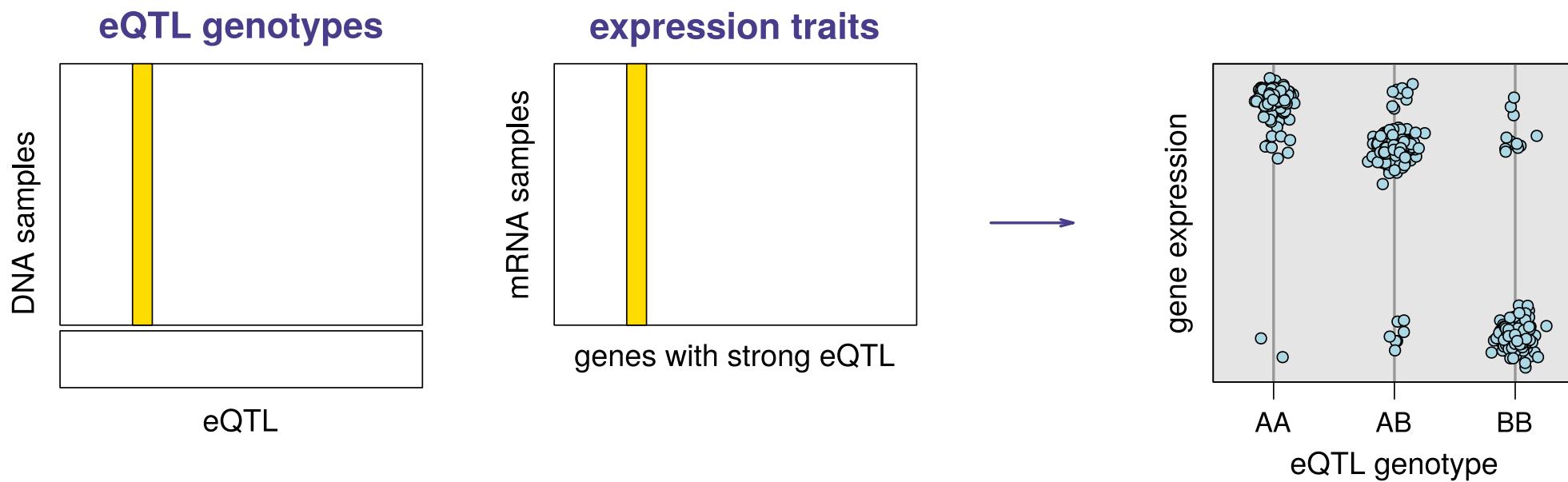
Sample mix-ups



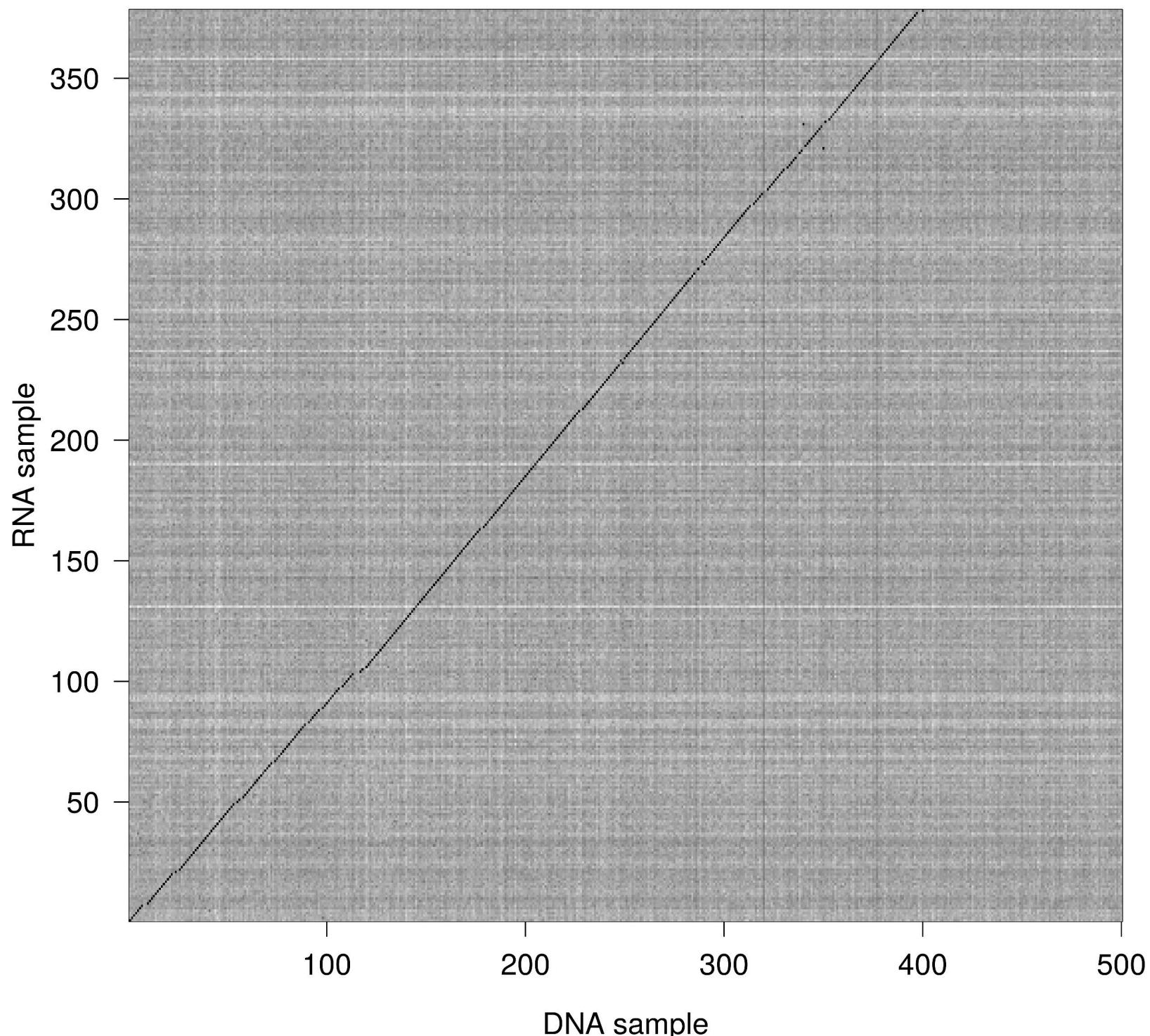
Sample mix-ups



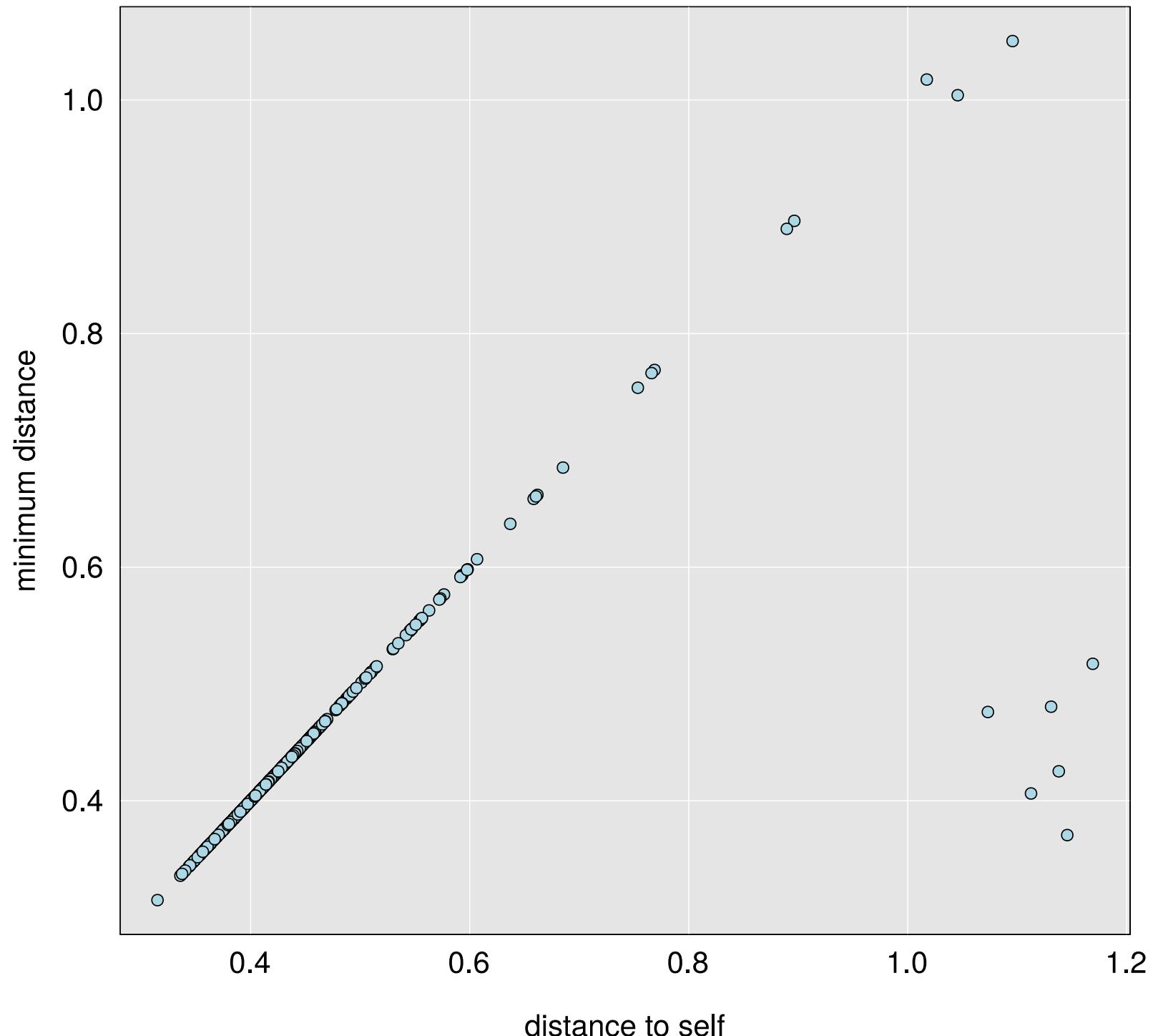
Sample mix-ups



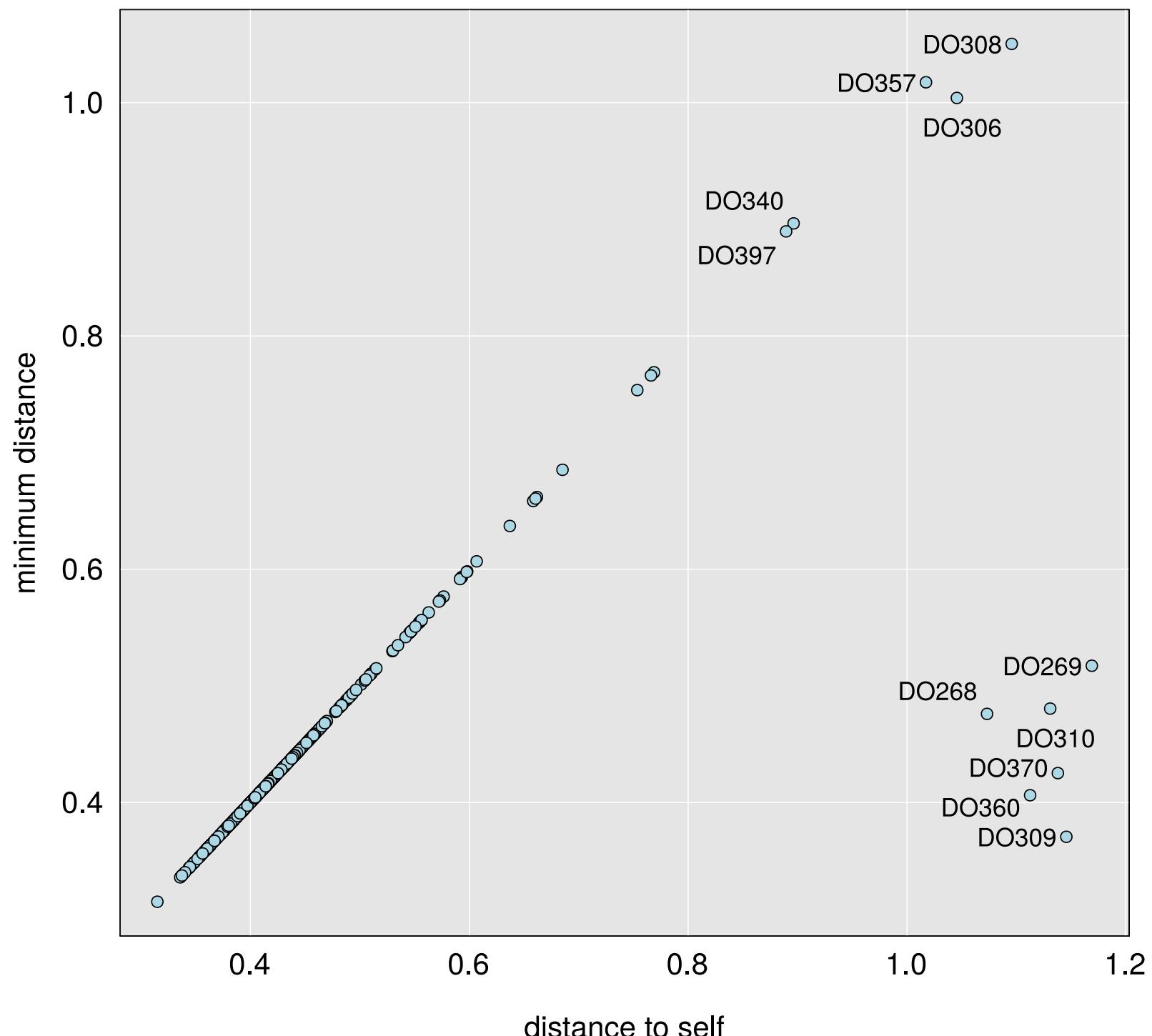
RNA-seq sample mix-ups: distance matrix



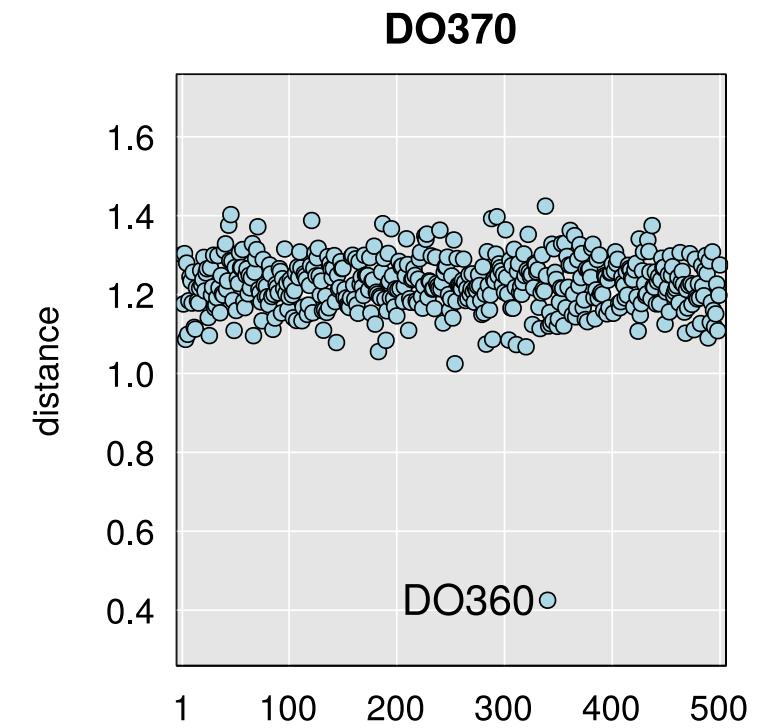
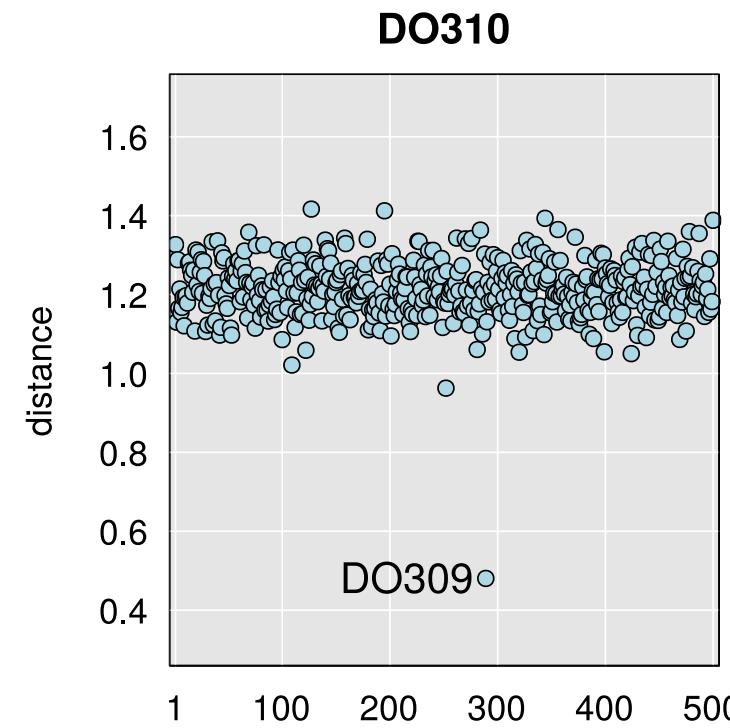
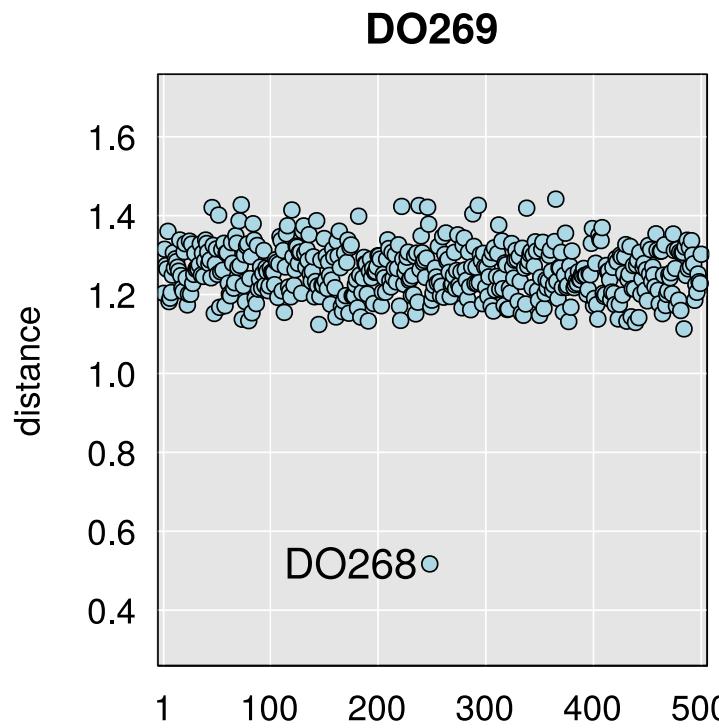
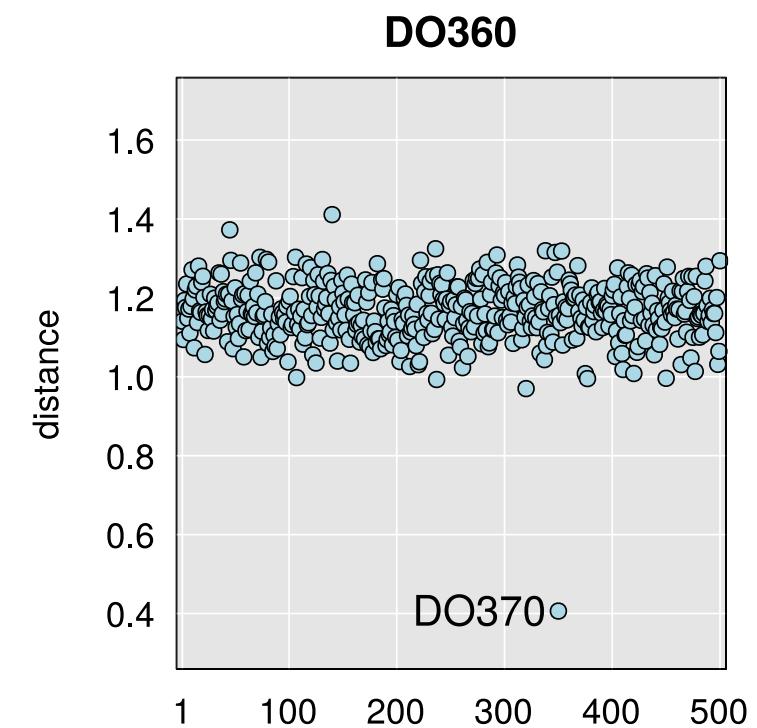
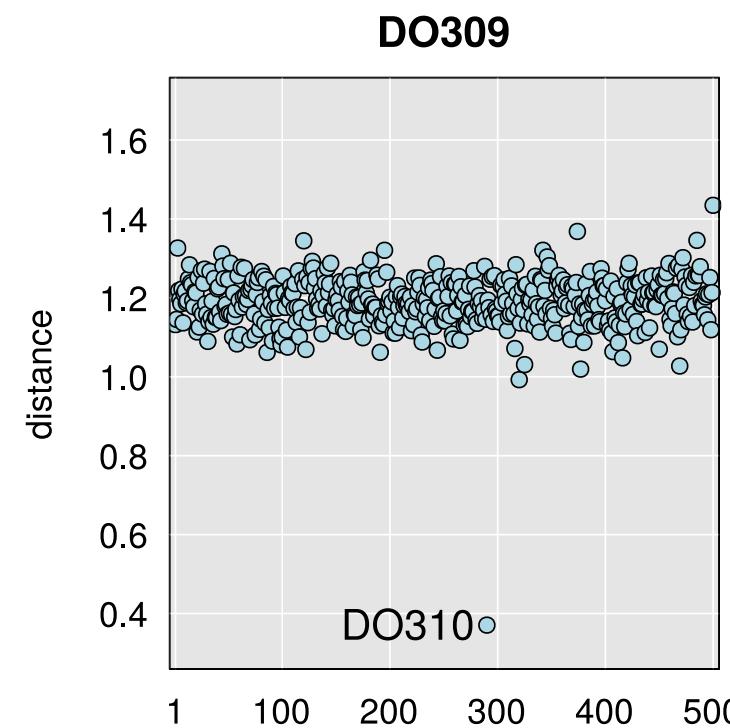
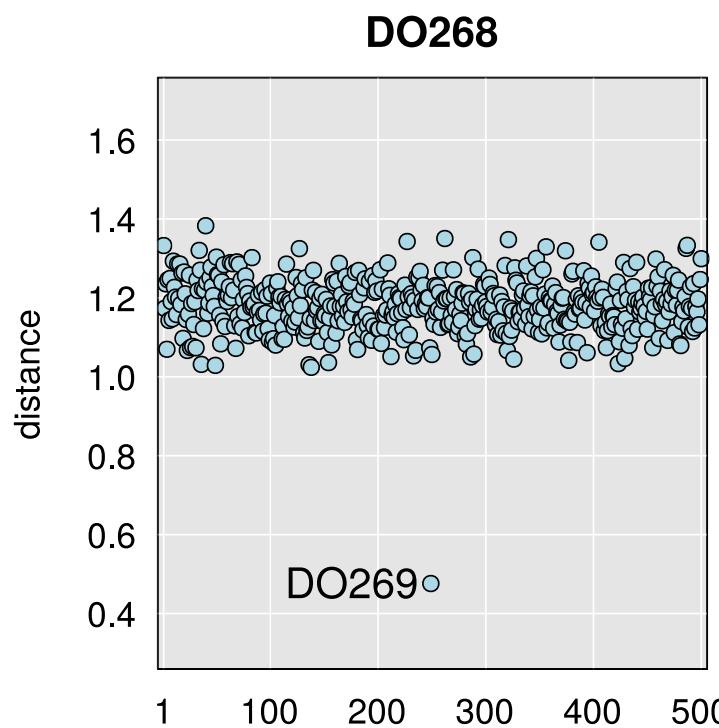
RNA-seq sample mix-ups: min vs self distance



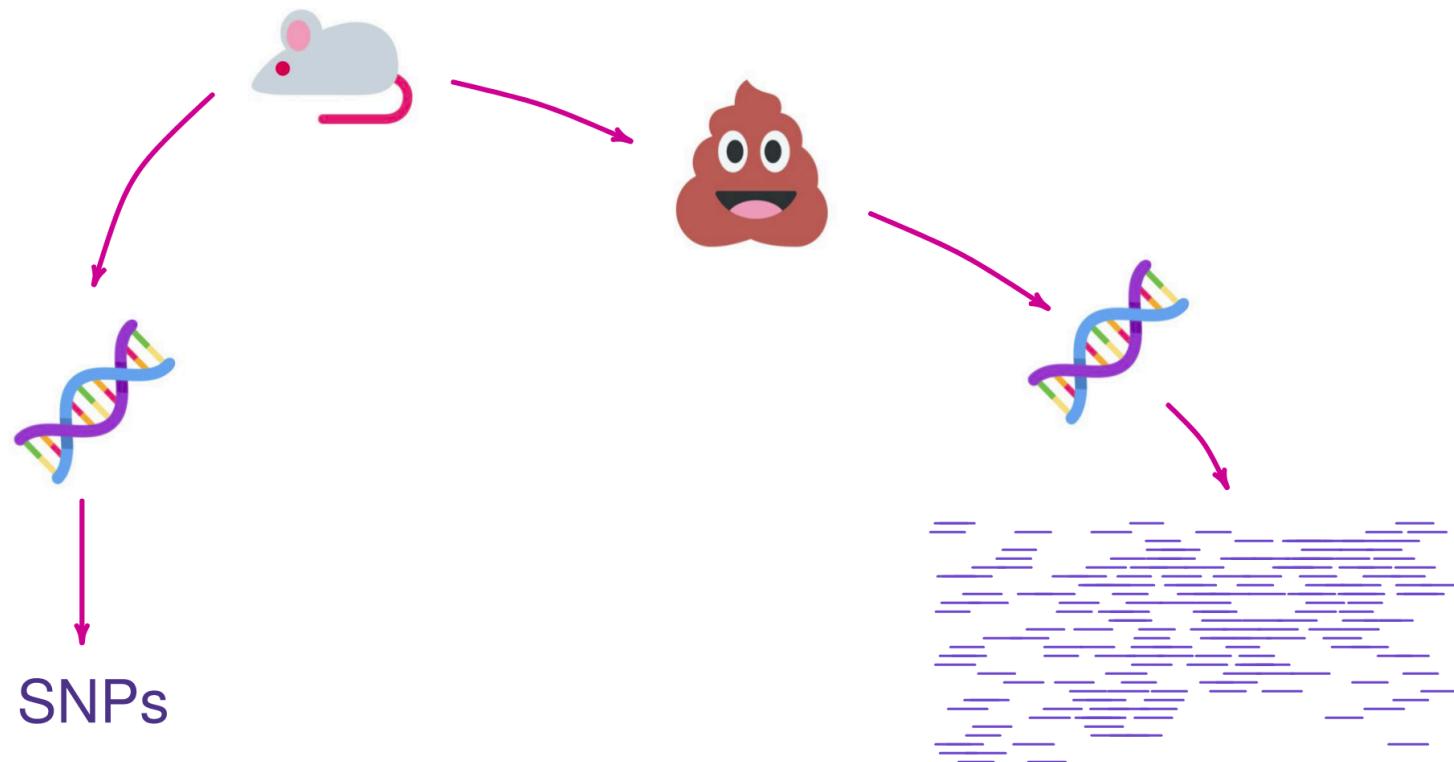
RNA-seq sample mix-ups: min vs self distance



RNA-seq sample mix-ups: detail



Microbiome data



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

Microbiome DO361 vs DNA DO361

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

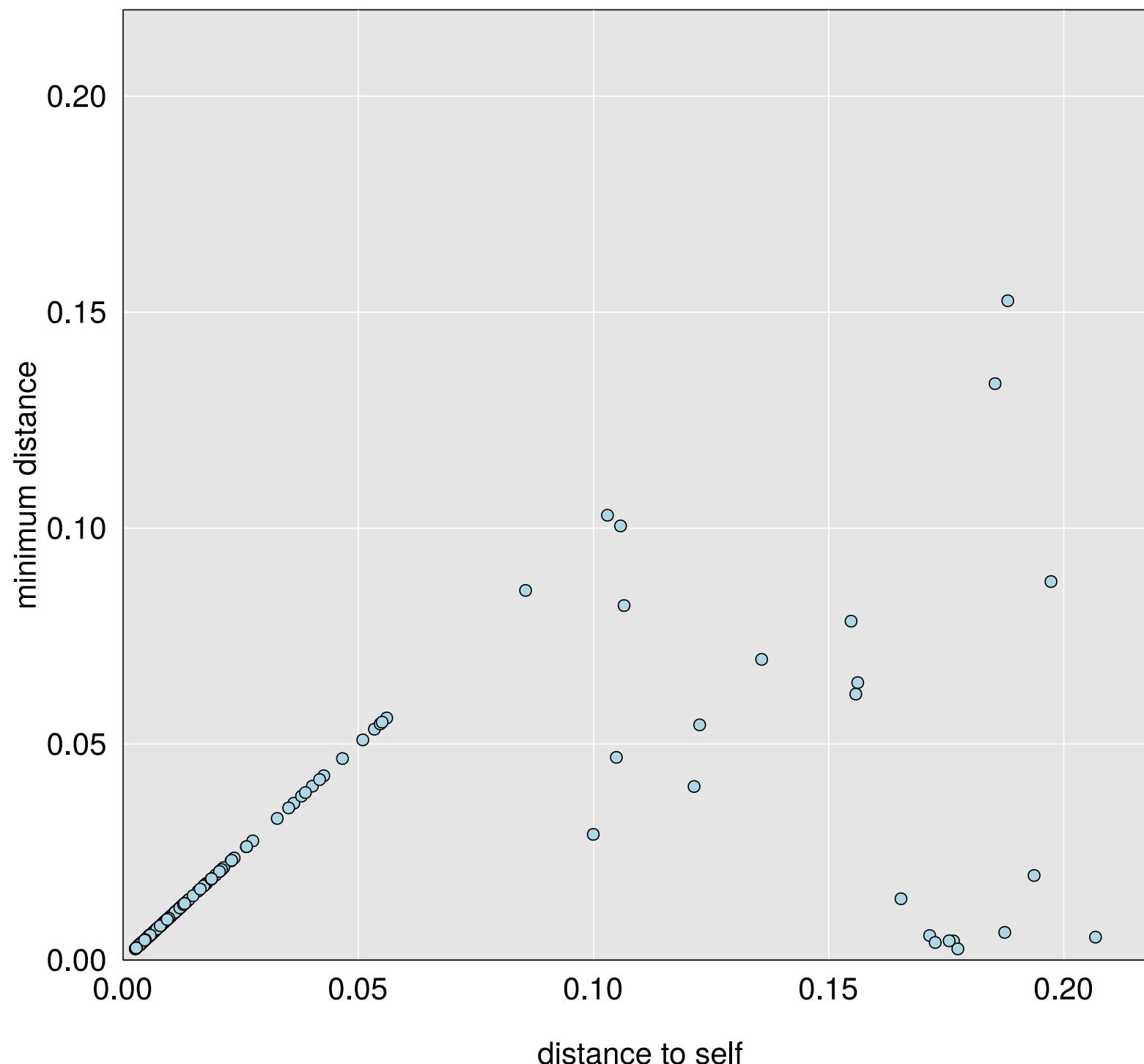
Microbiome DO360 vs DNA DO360

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

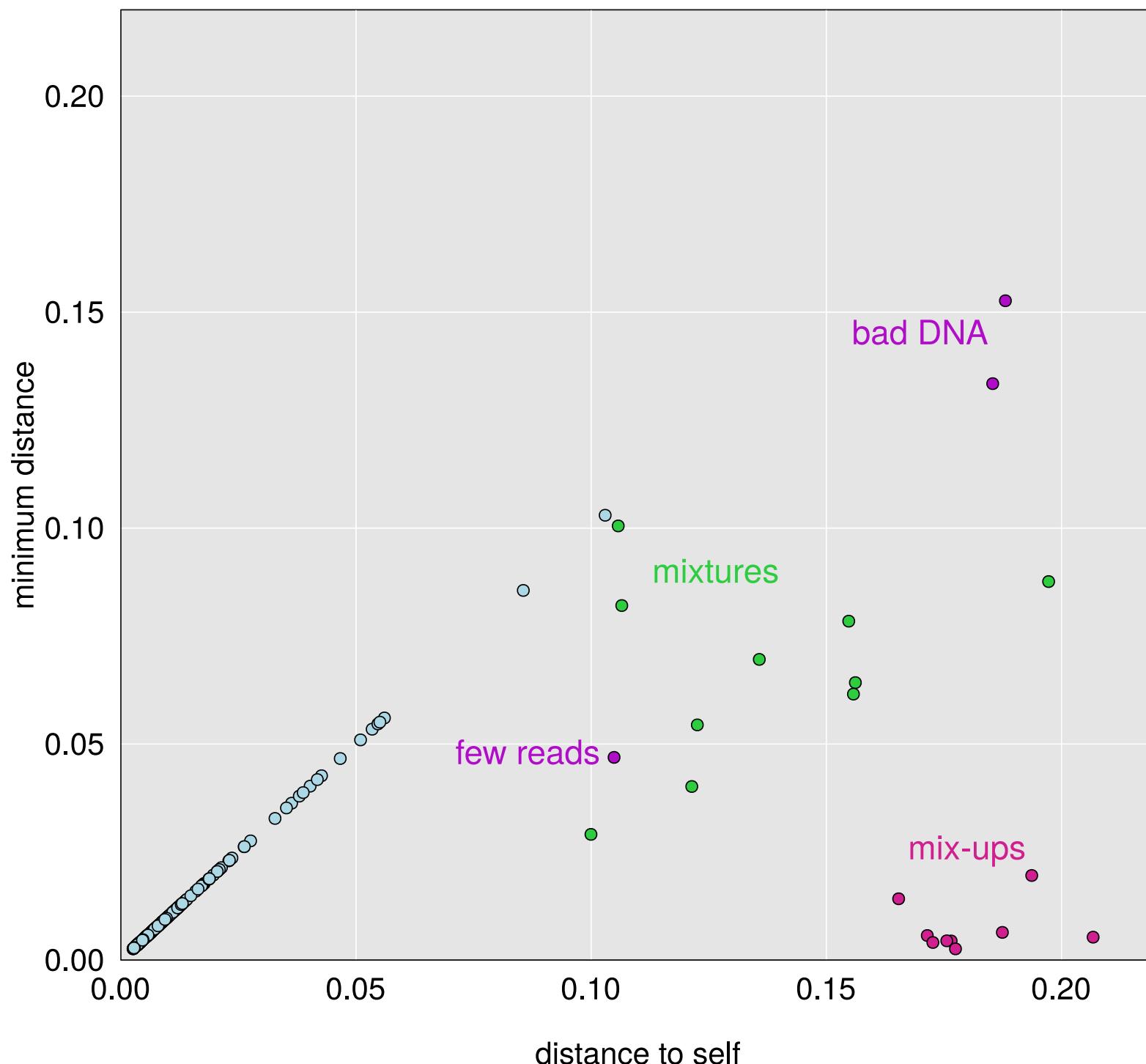
Microbiome DO360 vs DNA DO370

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

Microbiome mix-ups: min vs self distance

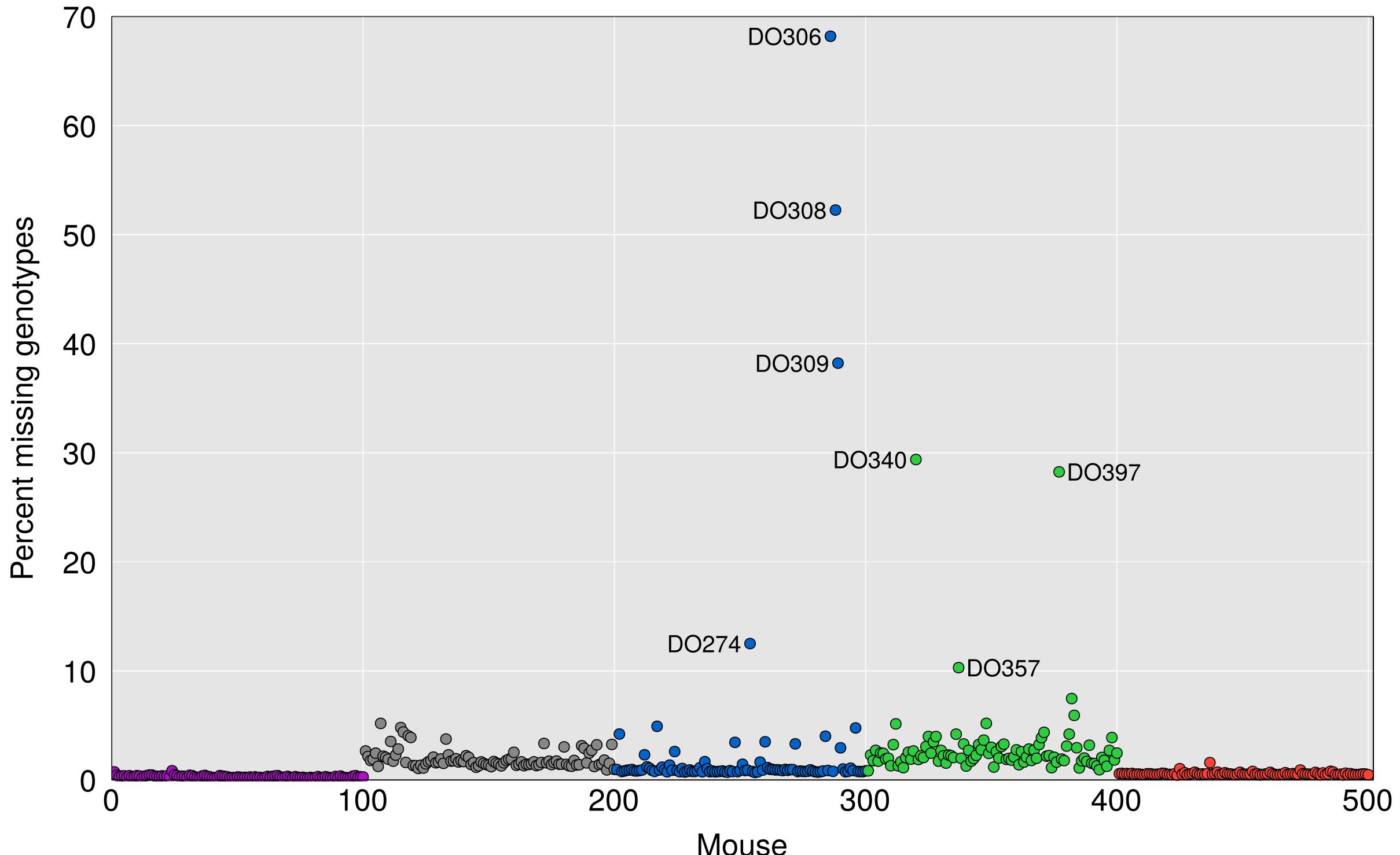


Microbiome mix-ups: min vs self distance

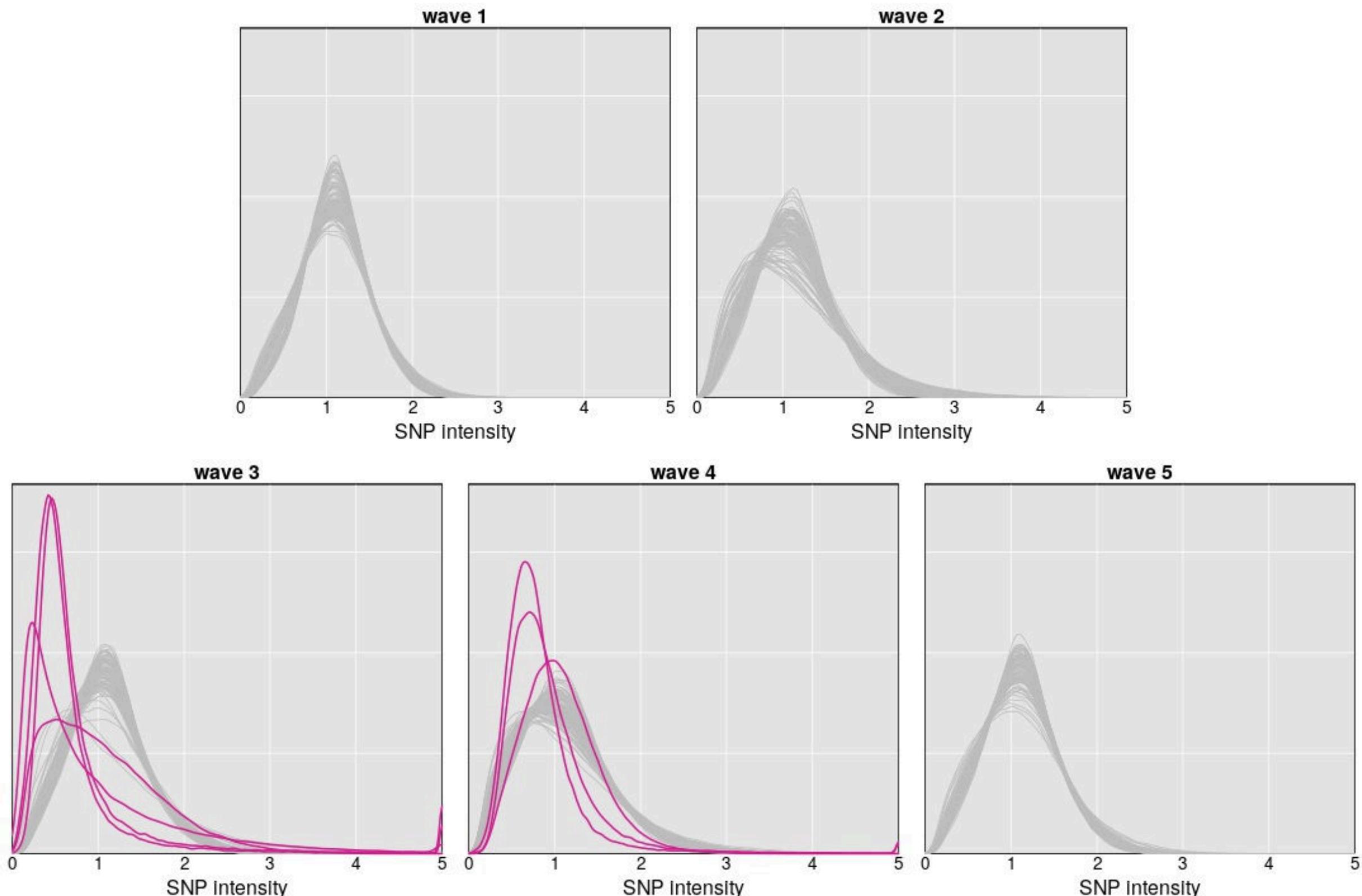


Sample quality

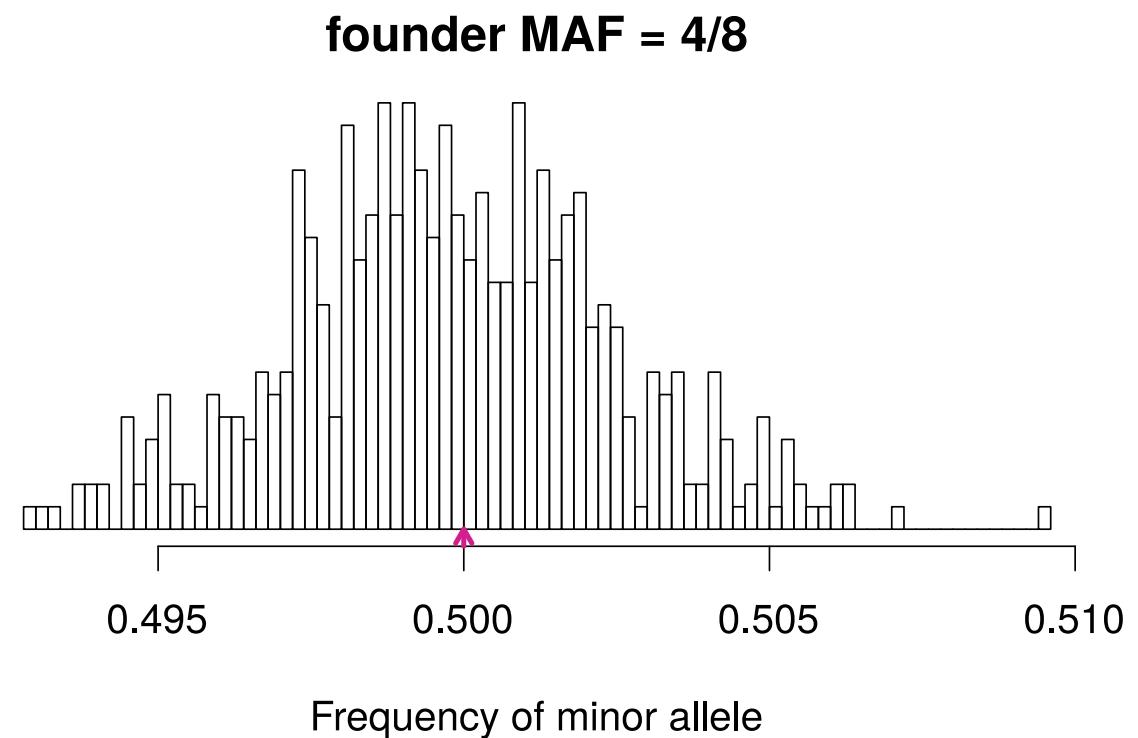
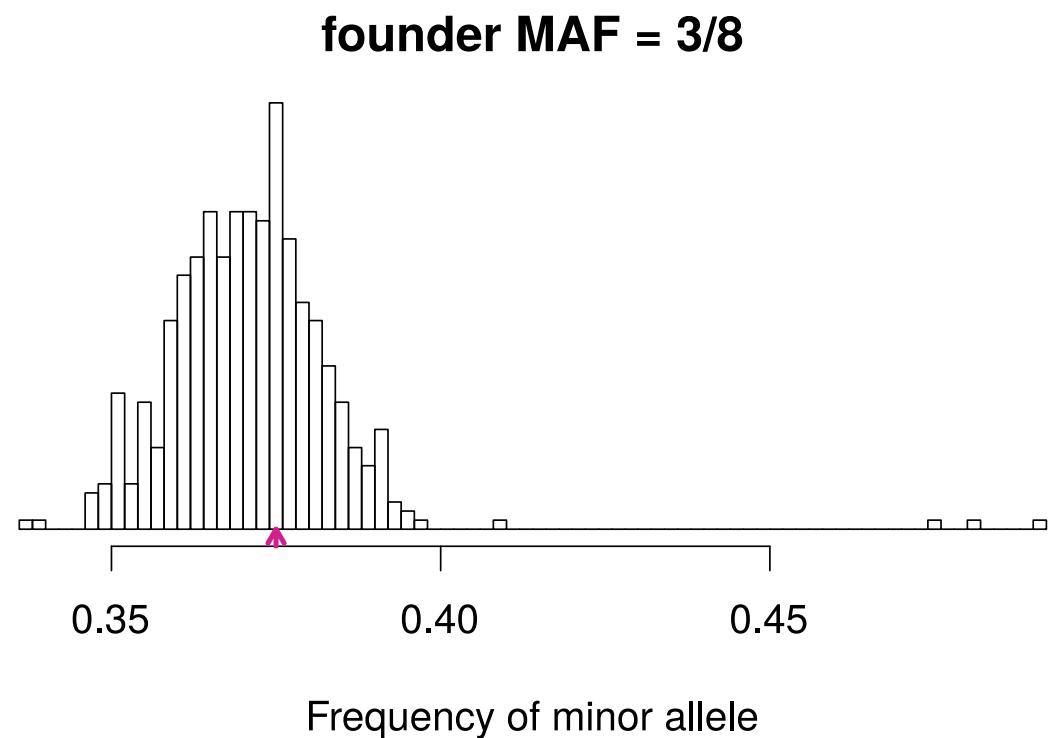
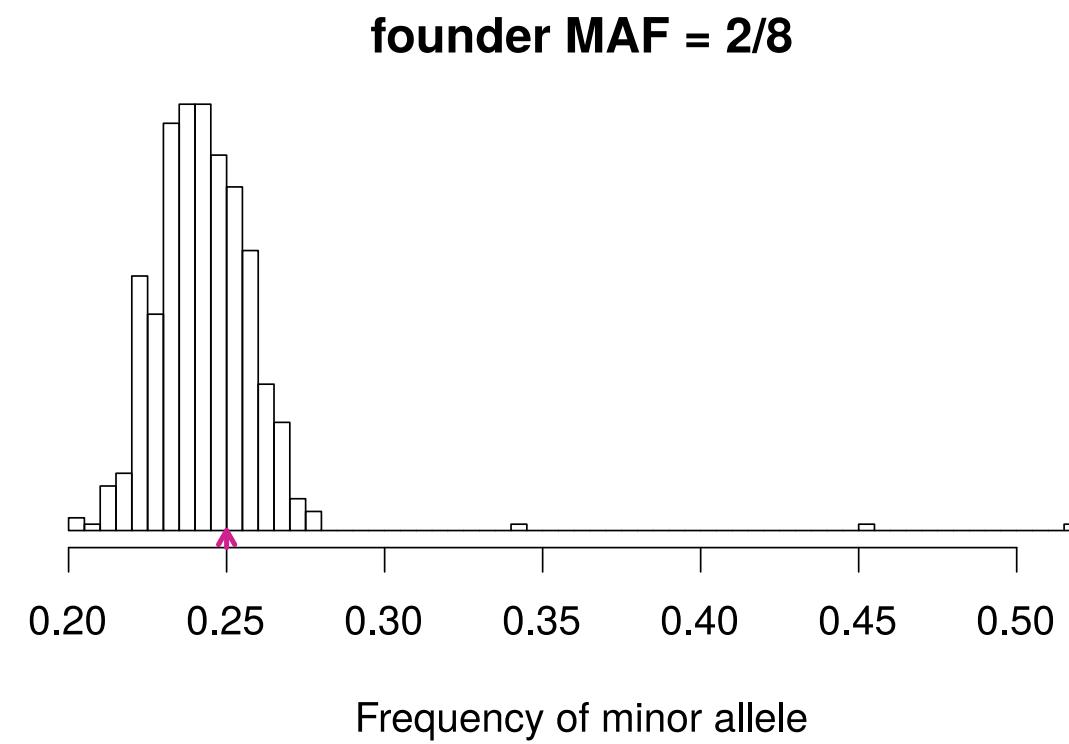
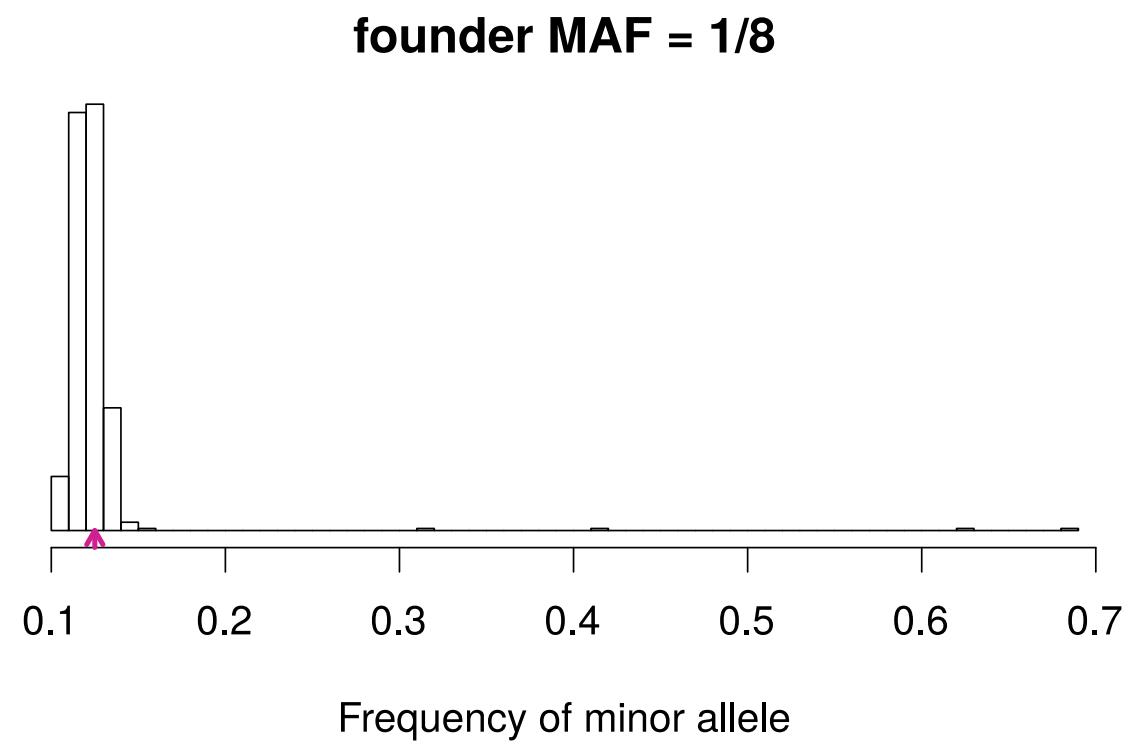
Missing data per sample



Array intensities

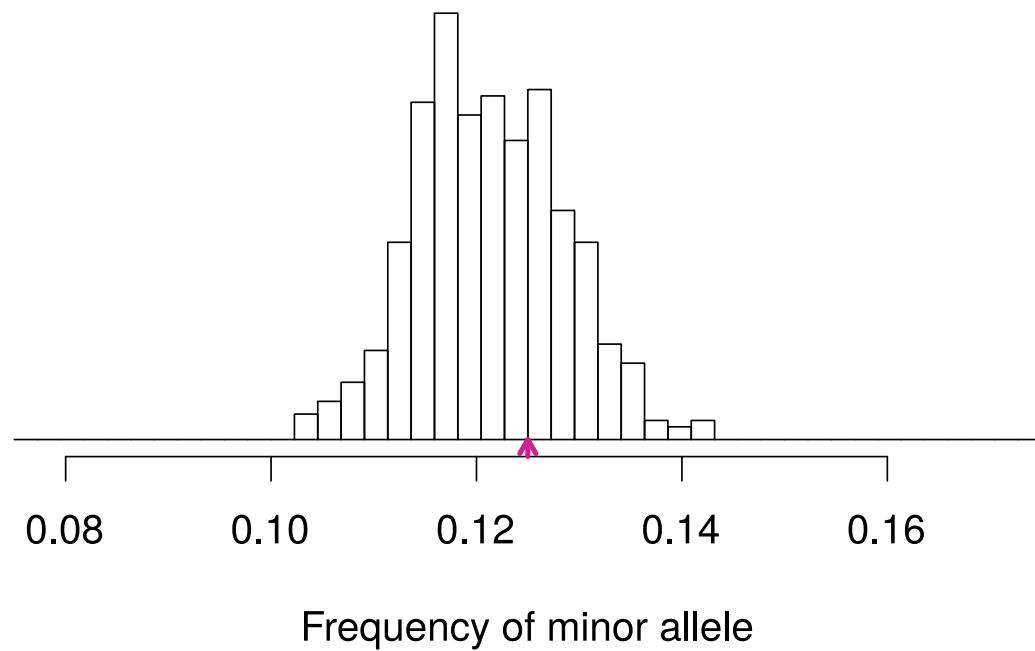


Allele frequencies, by individual

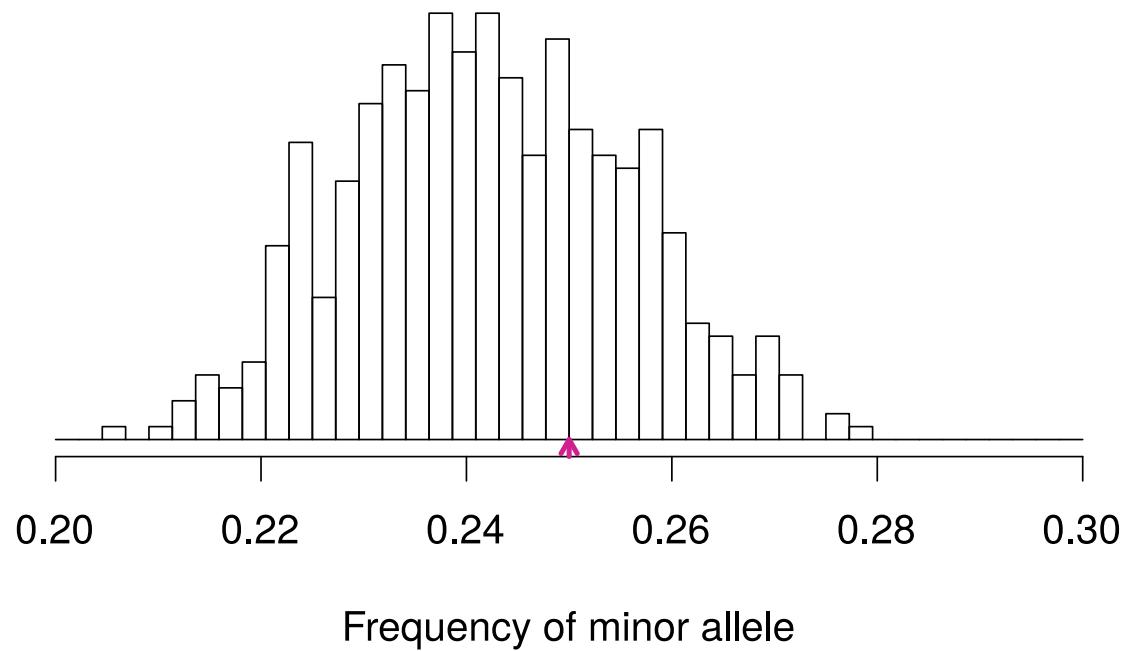


Allele frequencies, by individual

founder MAF = 1/8



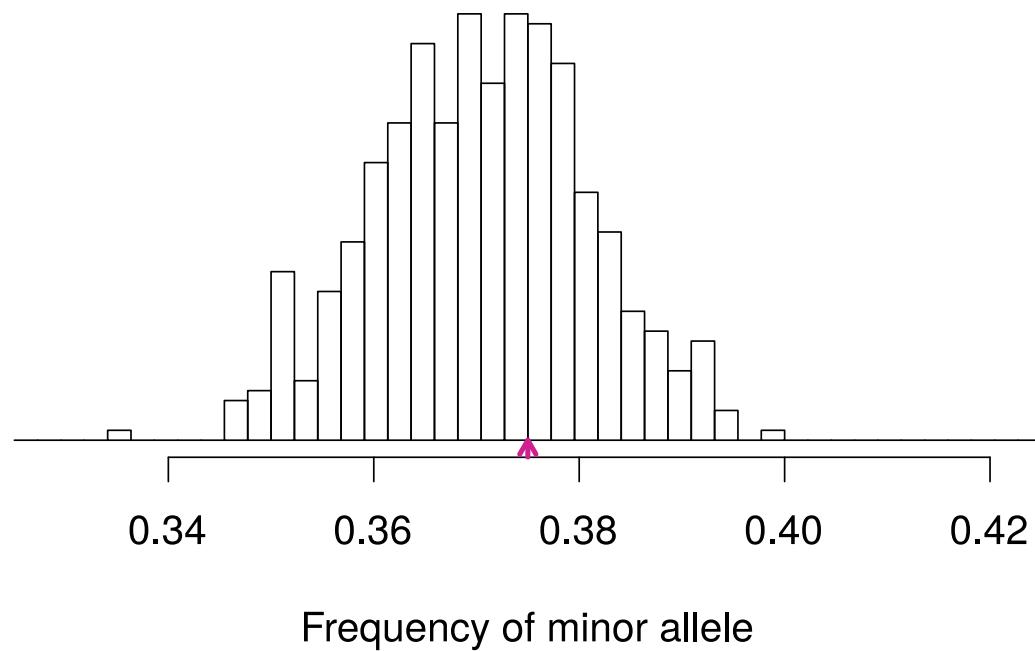
founder MAF = 2/8



Frequency of minor allele

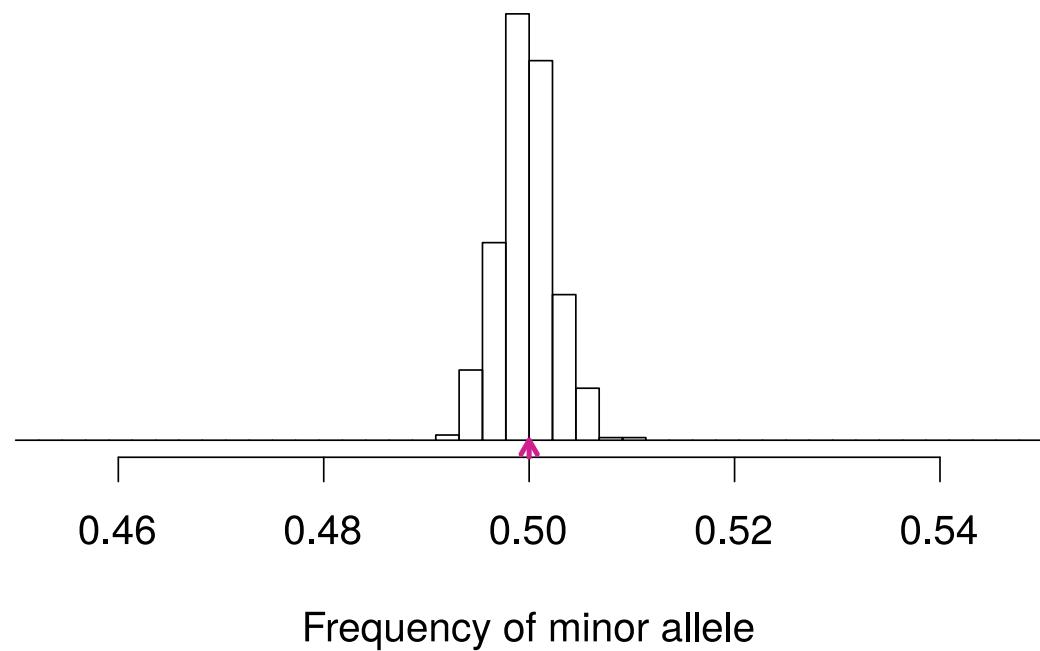
Frequency of minor allele

founder MAF = 3/8



Frequency of minor allele

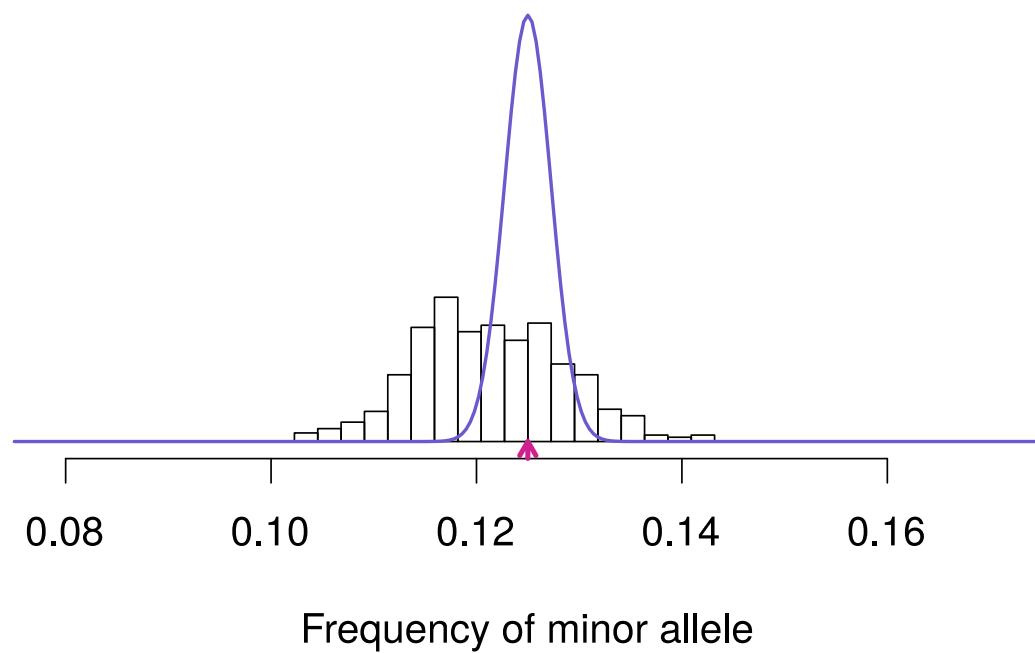
founder MAF = 4/8



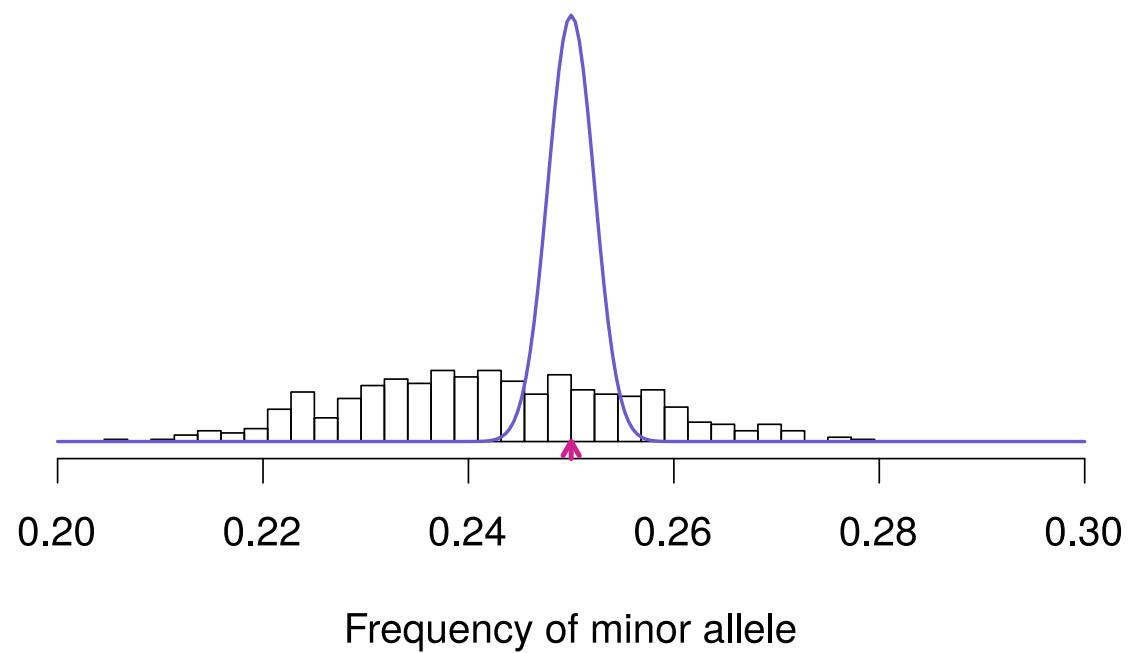
Frequency of minor allele

Allele frequencies, by individual

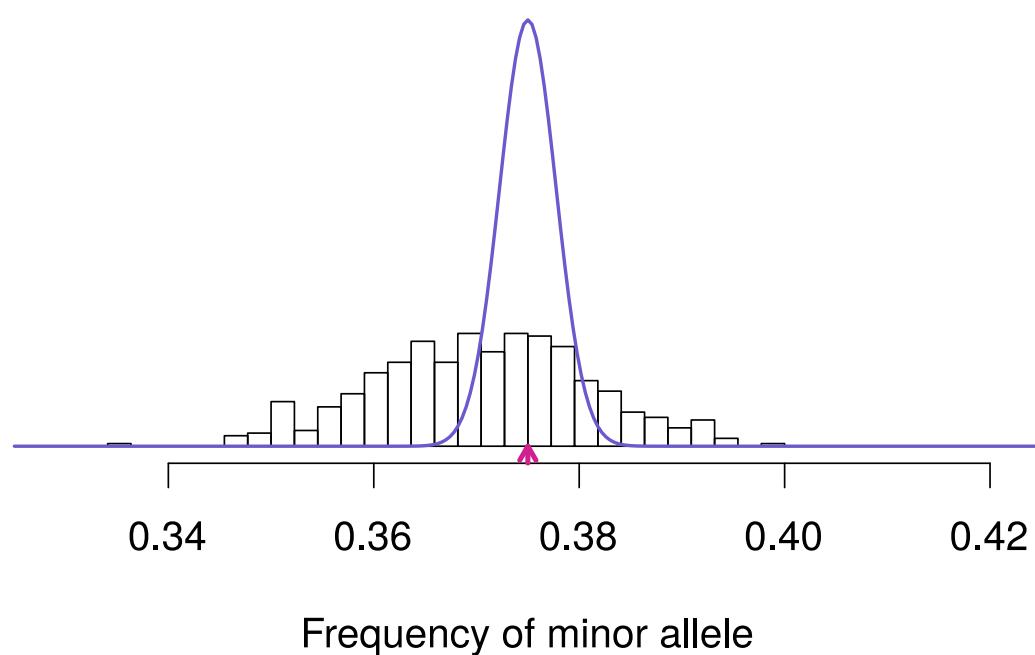
founder MAF = 1/8



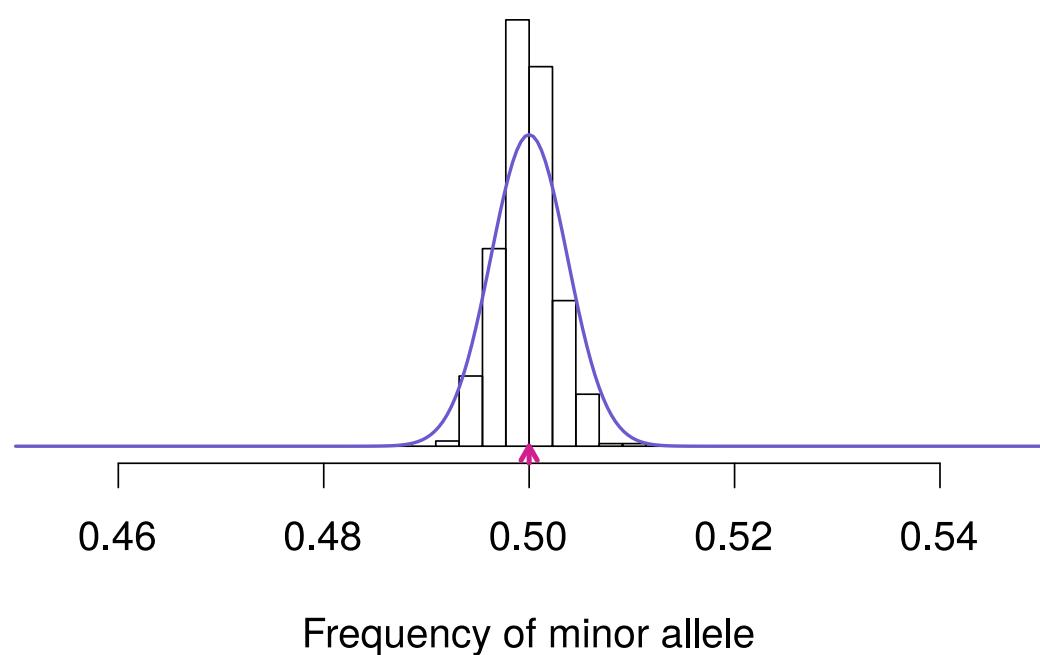
founder MAF = 2/8



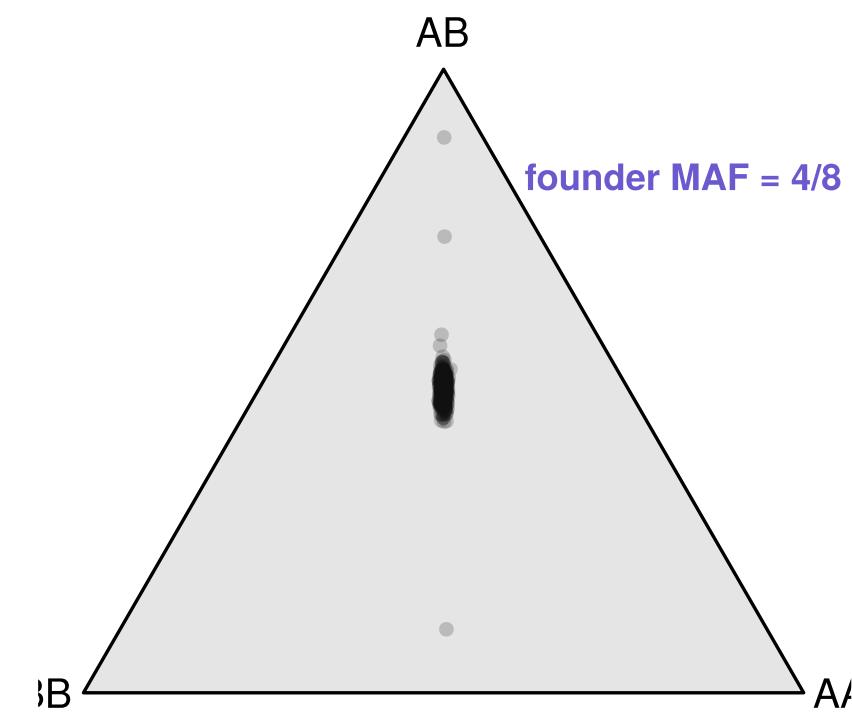
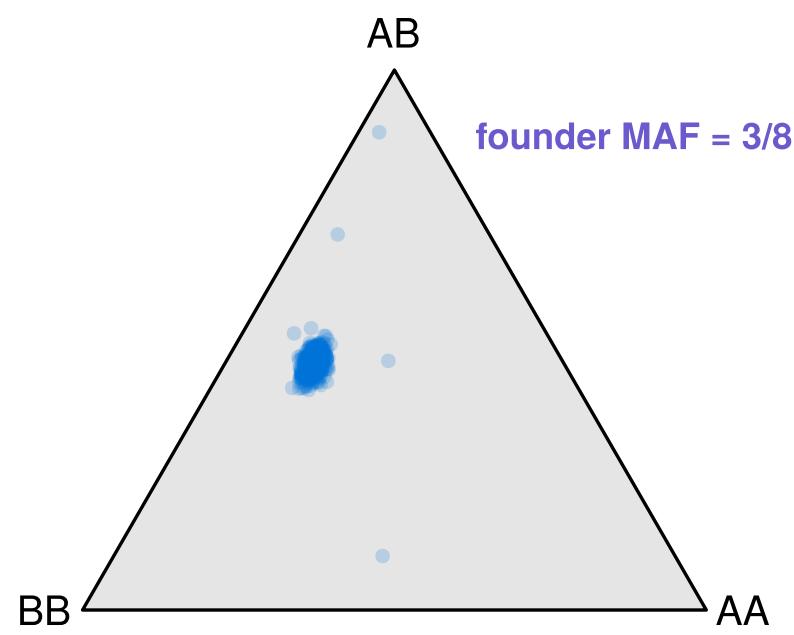
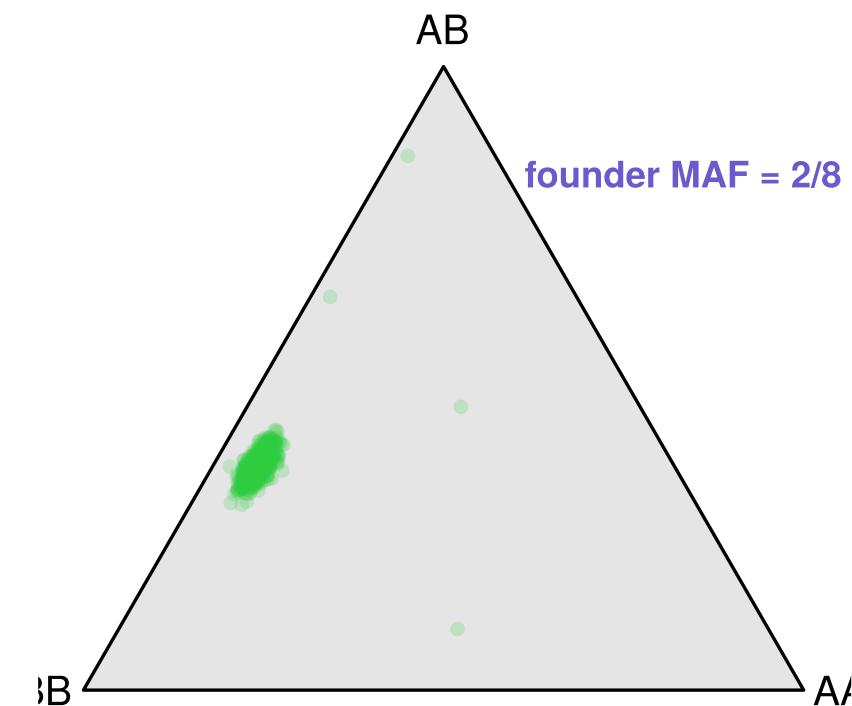
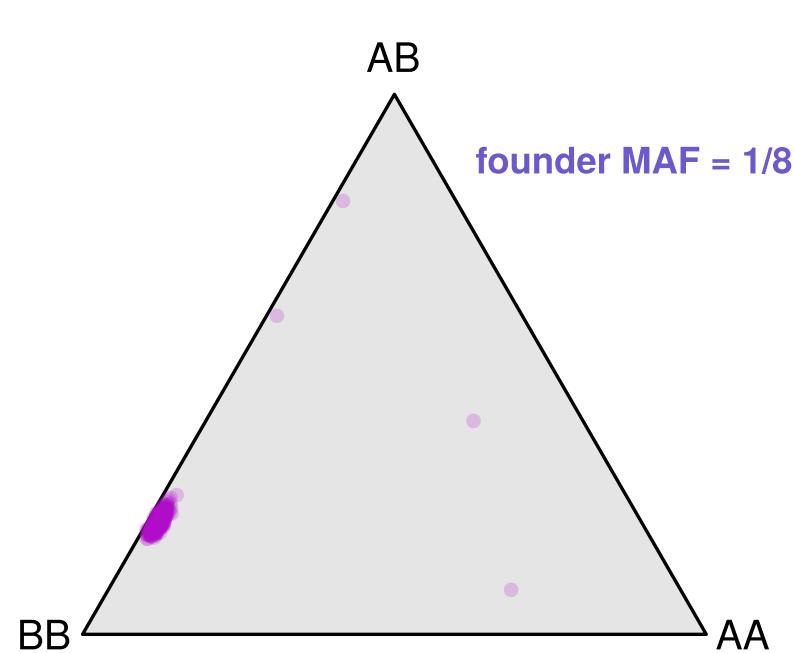
founder MAF = 3/8



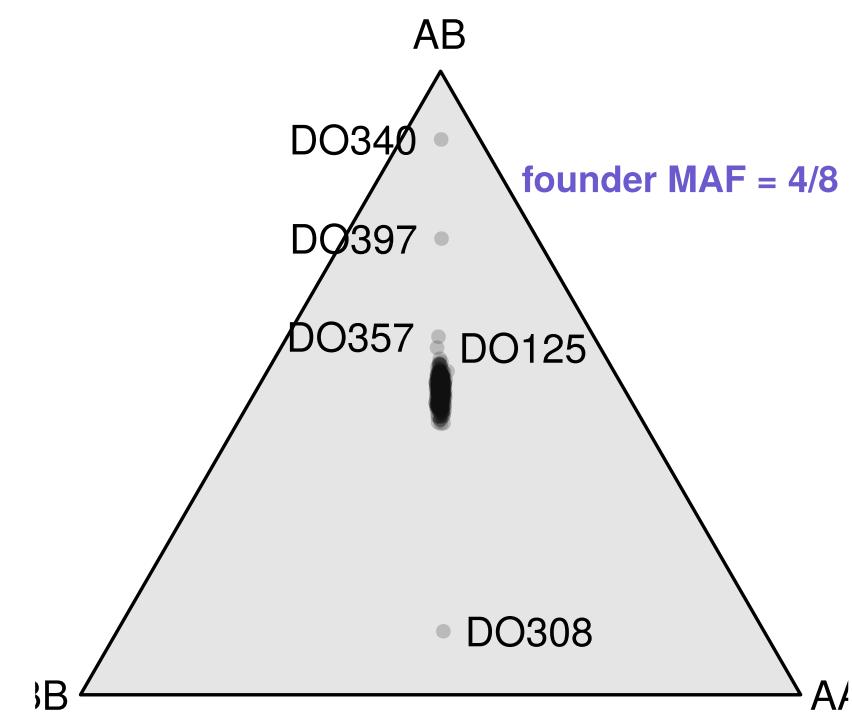
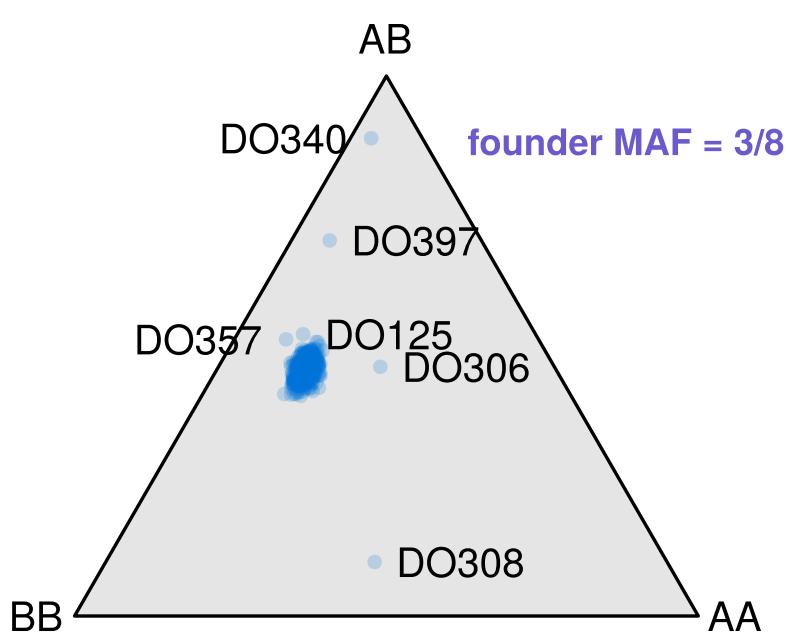
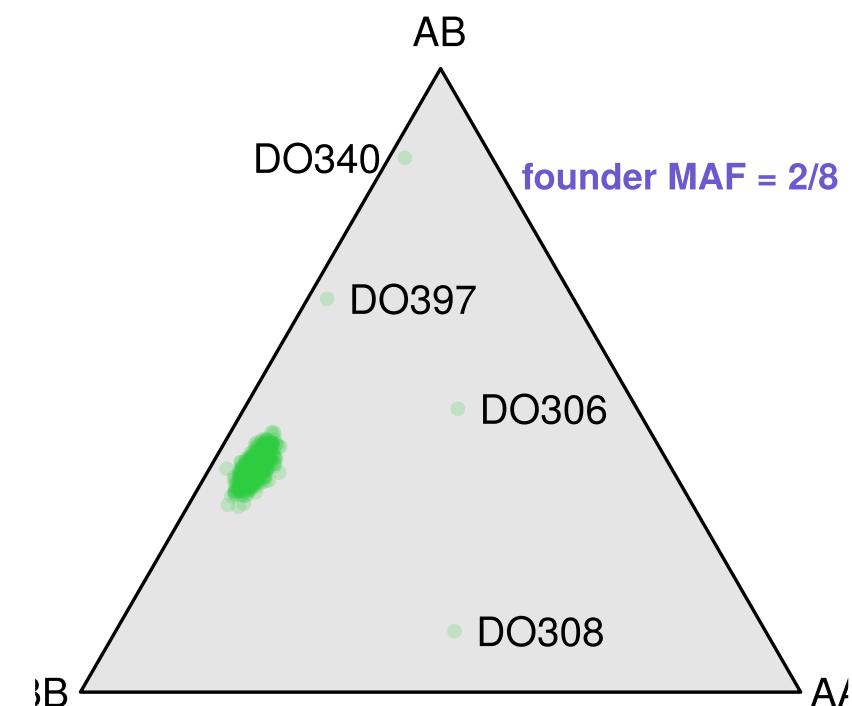
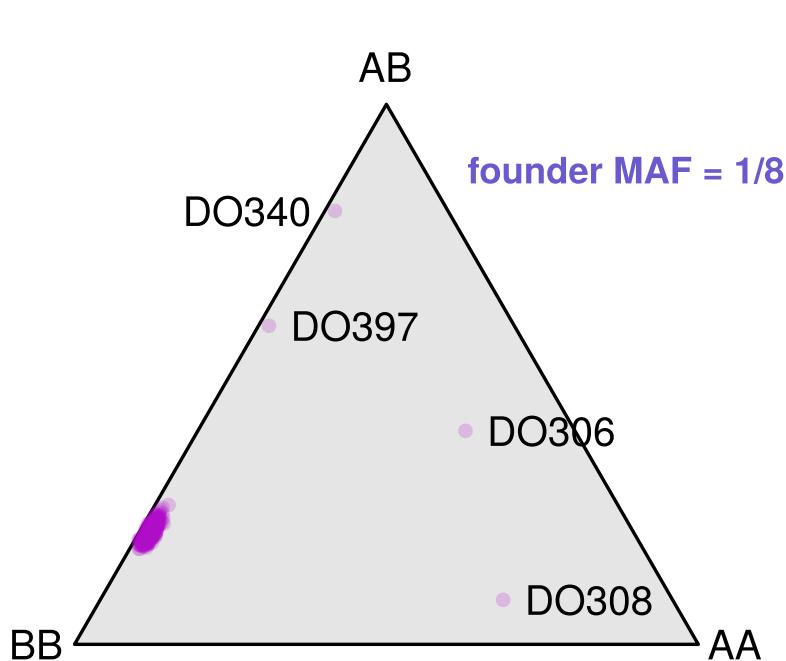
founder MAF = 4/8



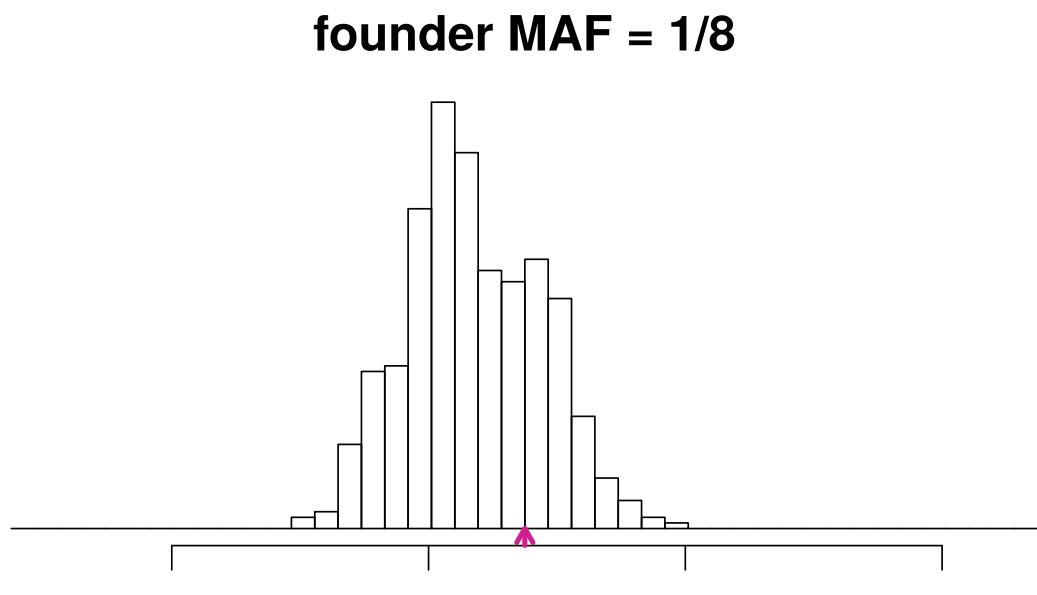
Genotype frequencies, by individual



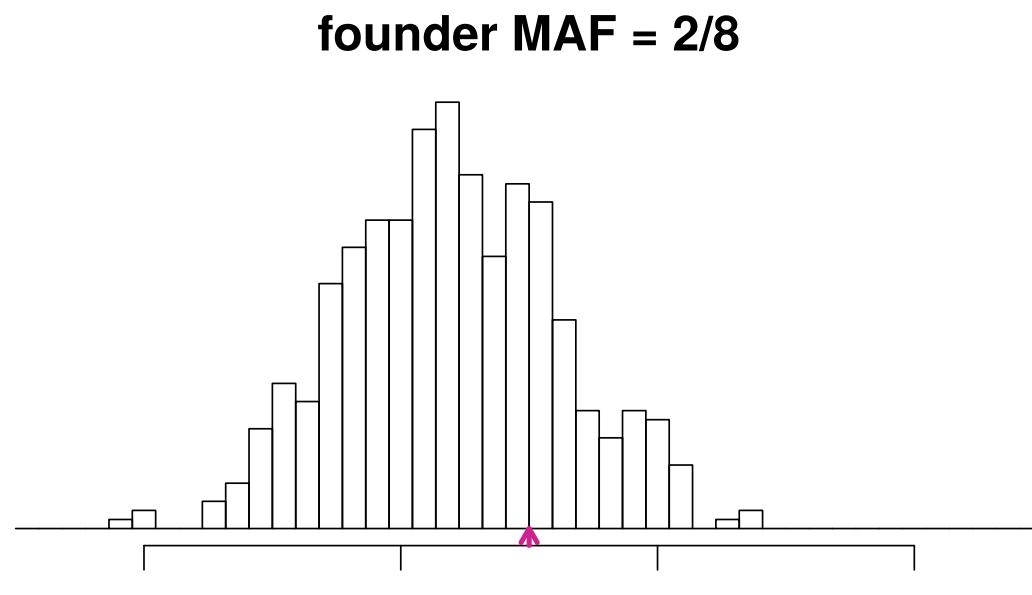
Genotype frequencies, by individual



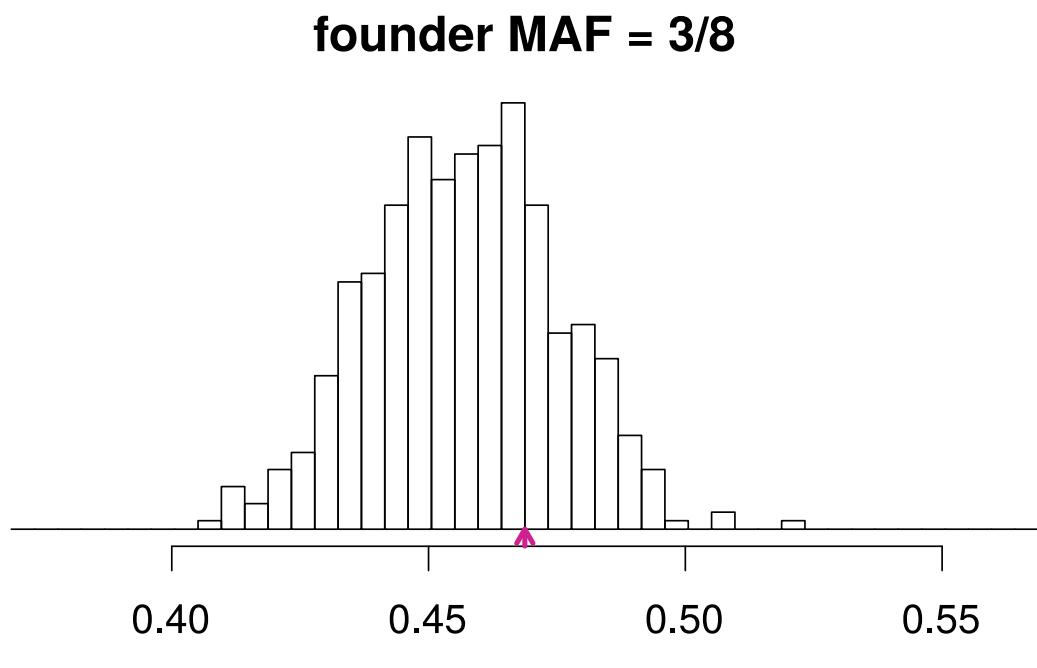
Heterozygosities, by individual



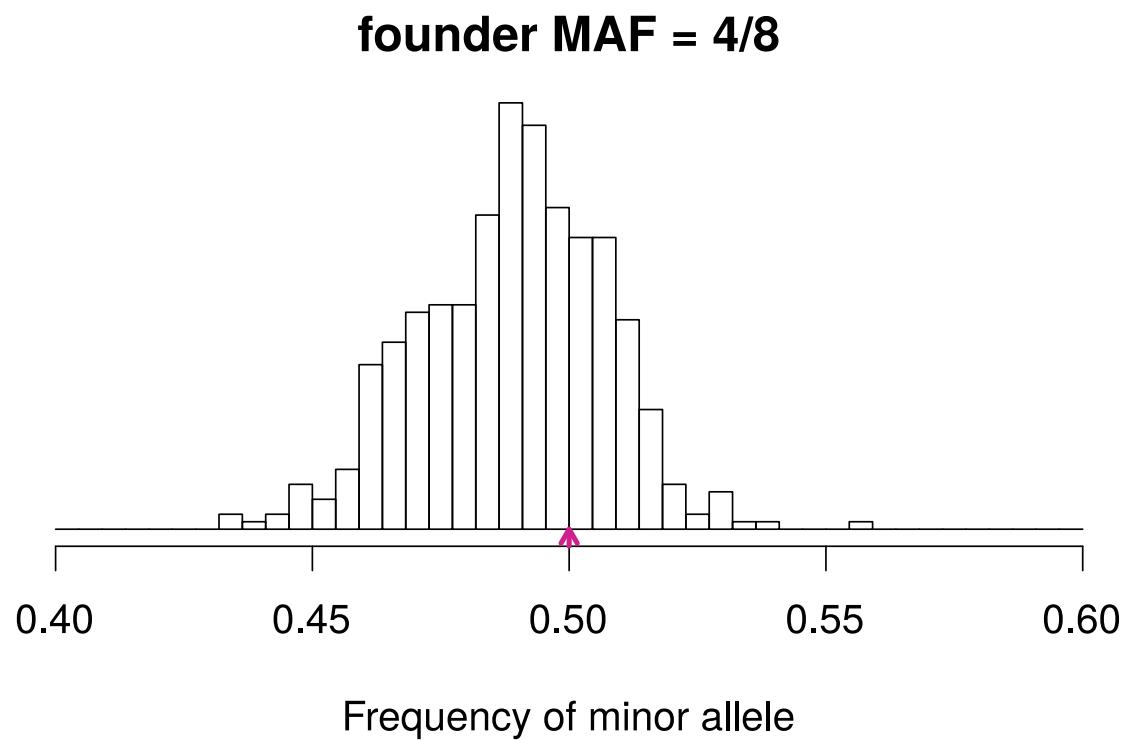
Frequency of minor allele



Frequency of minor allele



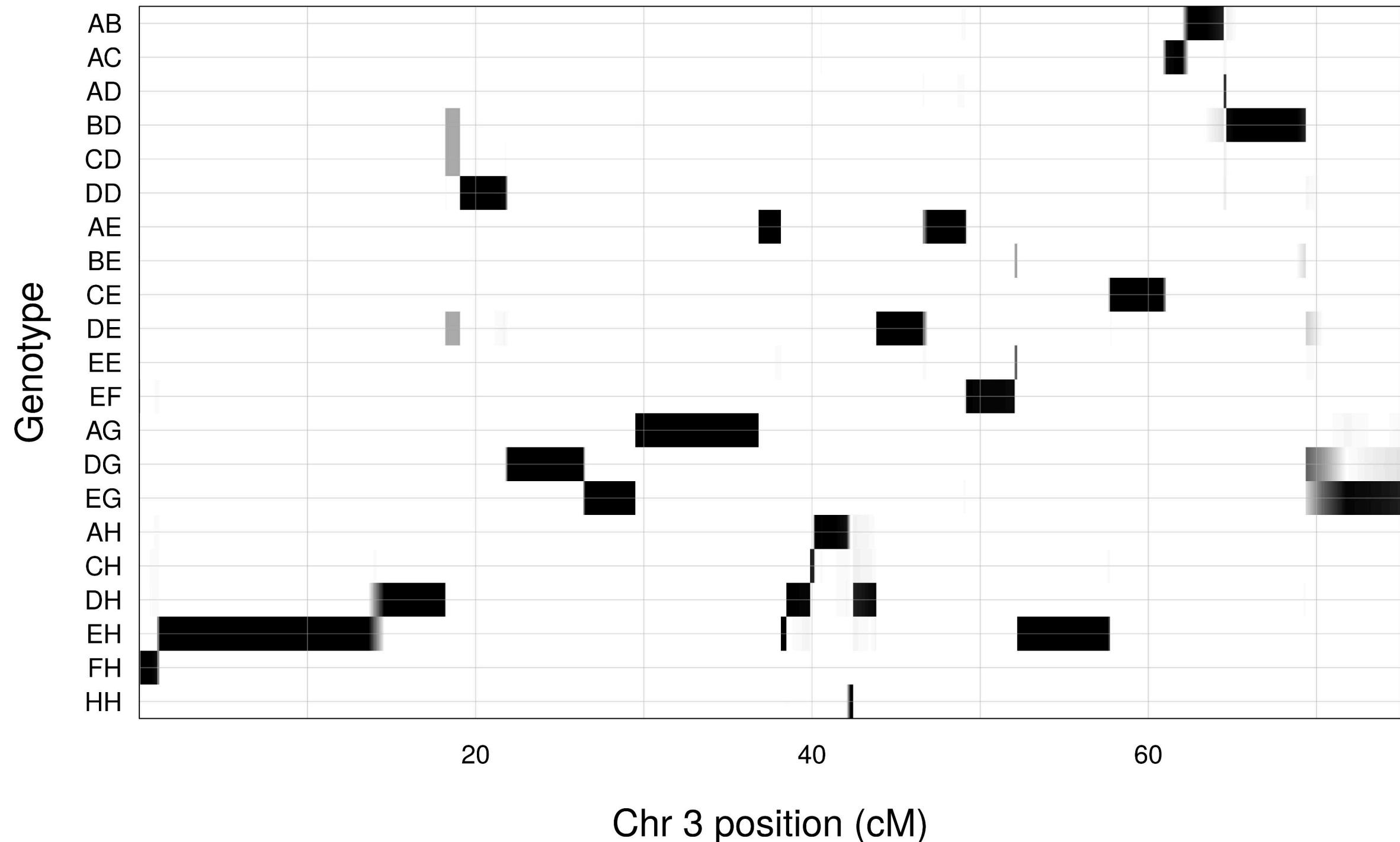
Frequency of minor allele



Frequency of minor allele

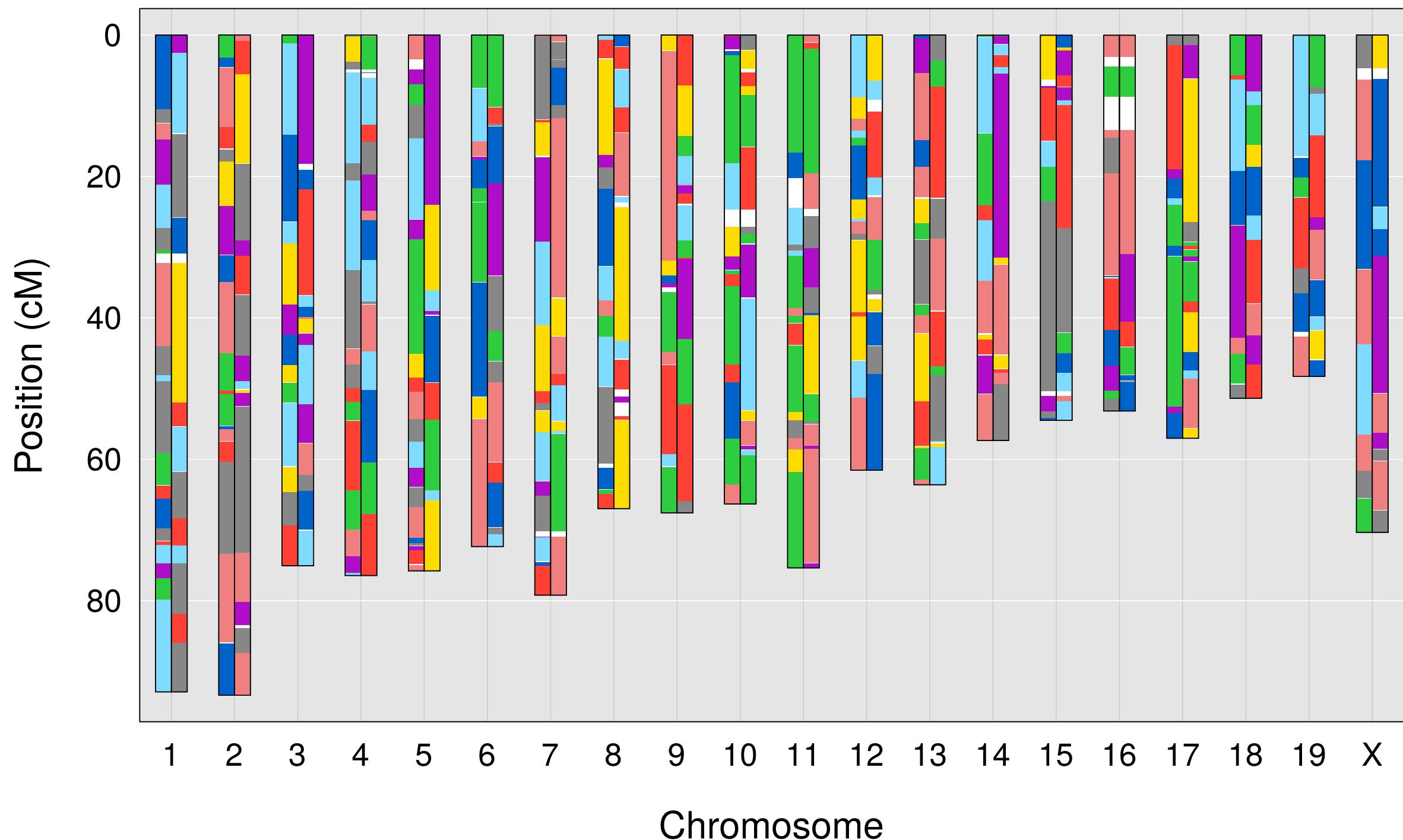
Genotype probabilities (one mouse on one chr)

DO361

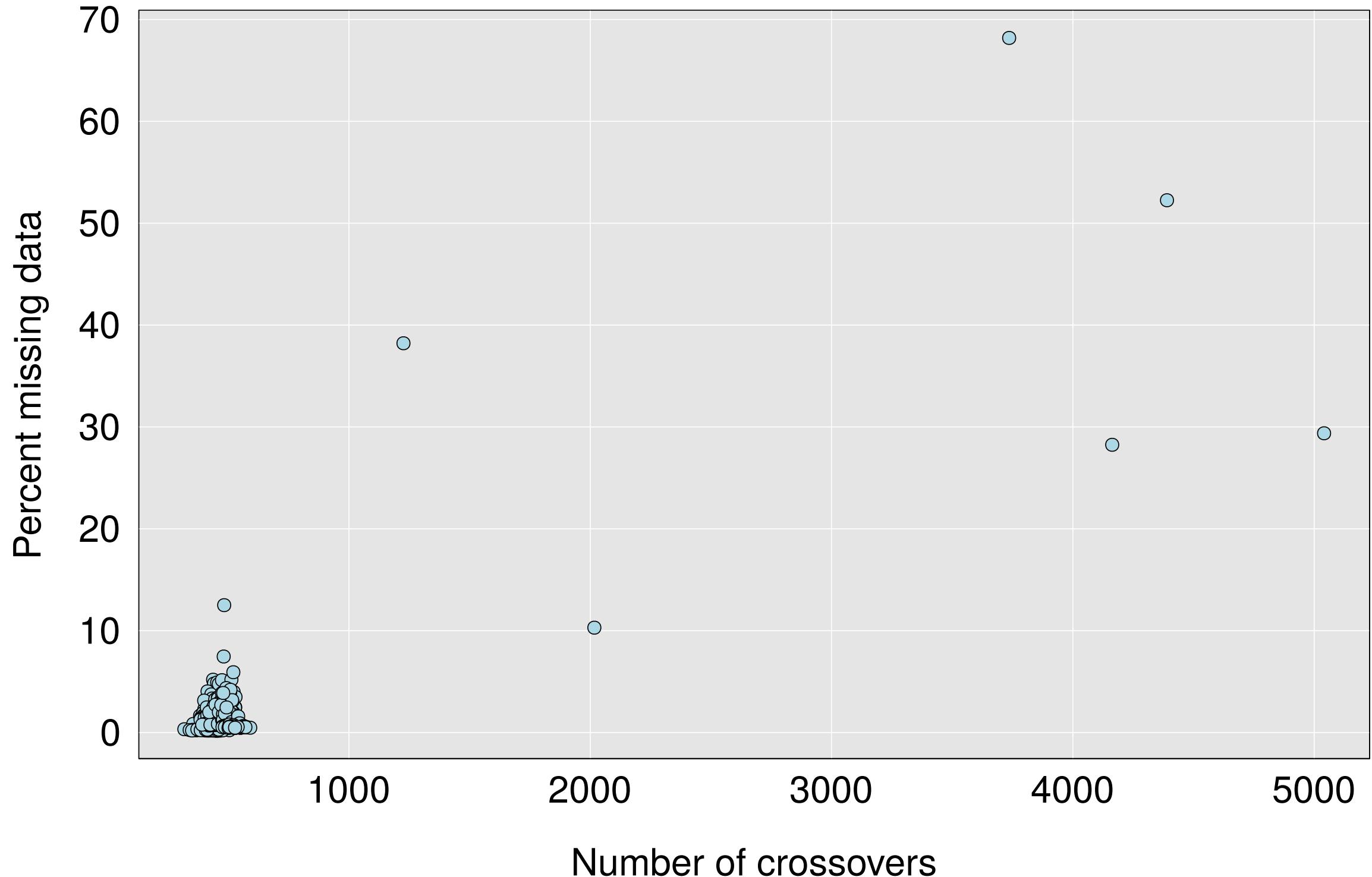


Genome reconstruction (one mouse)

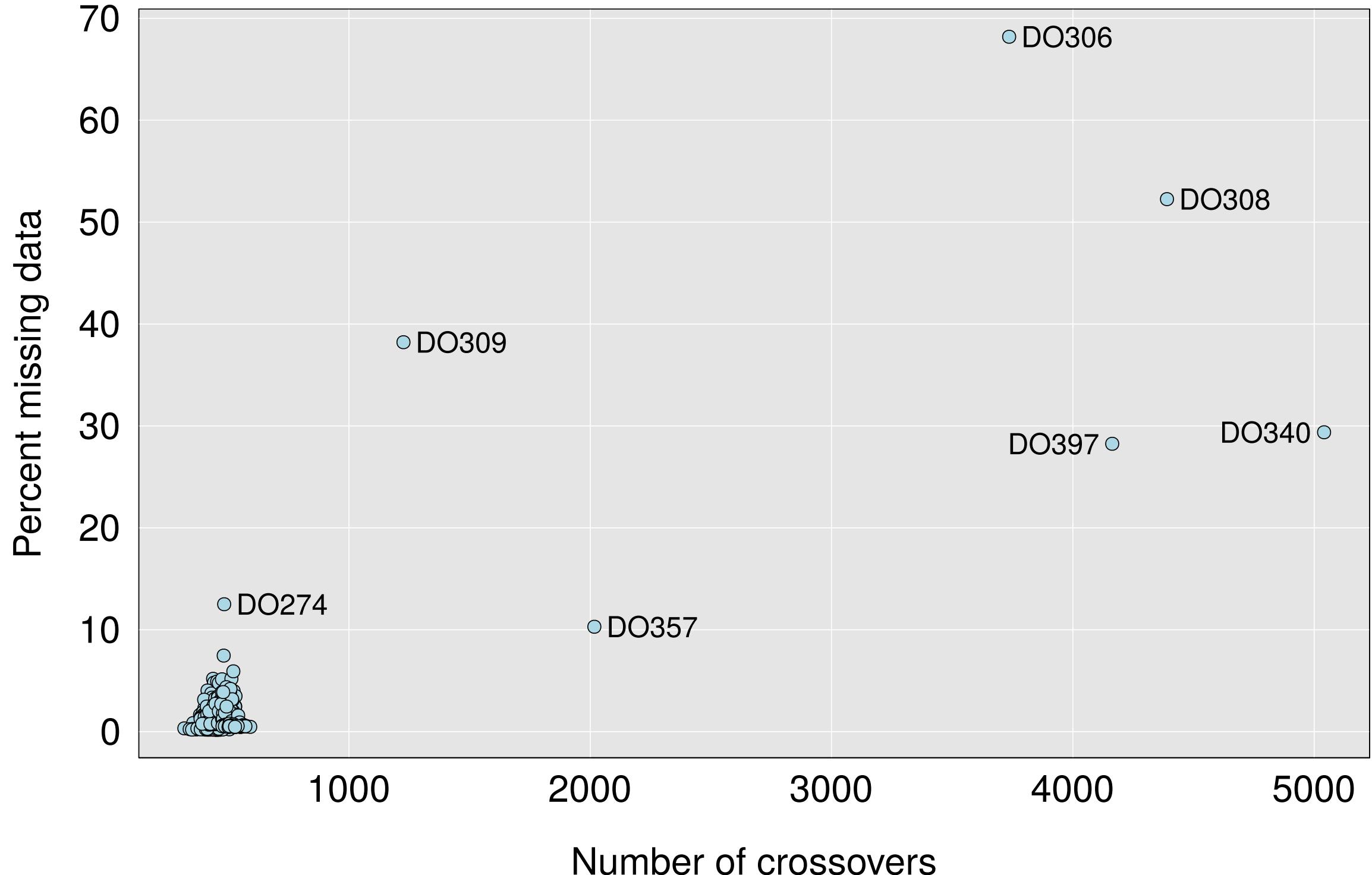
DO361



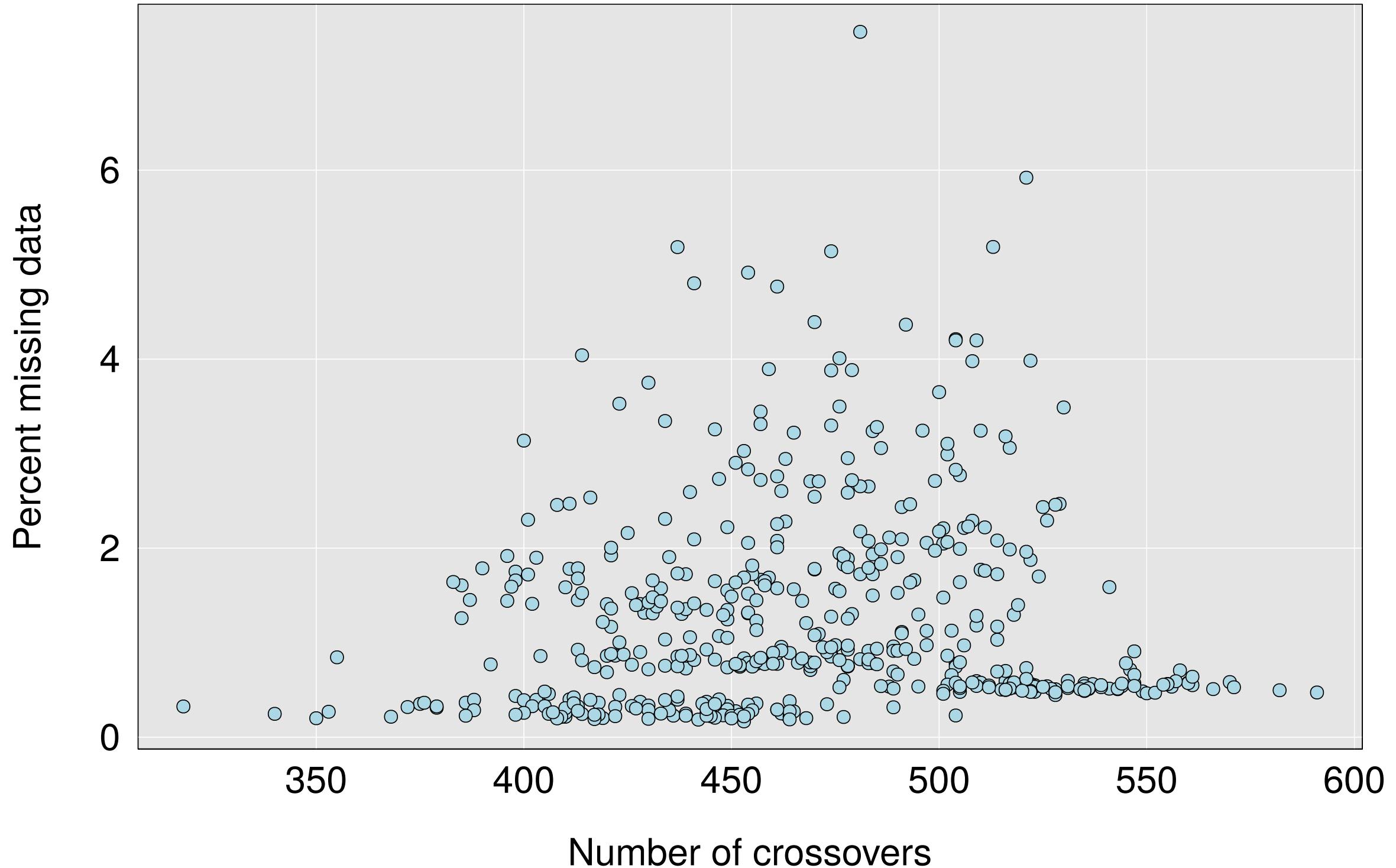
Percent missing vs number of crossovers



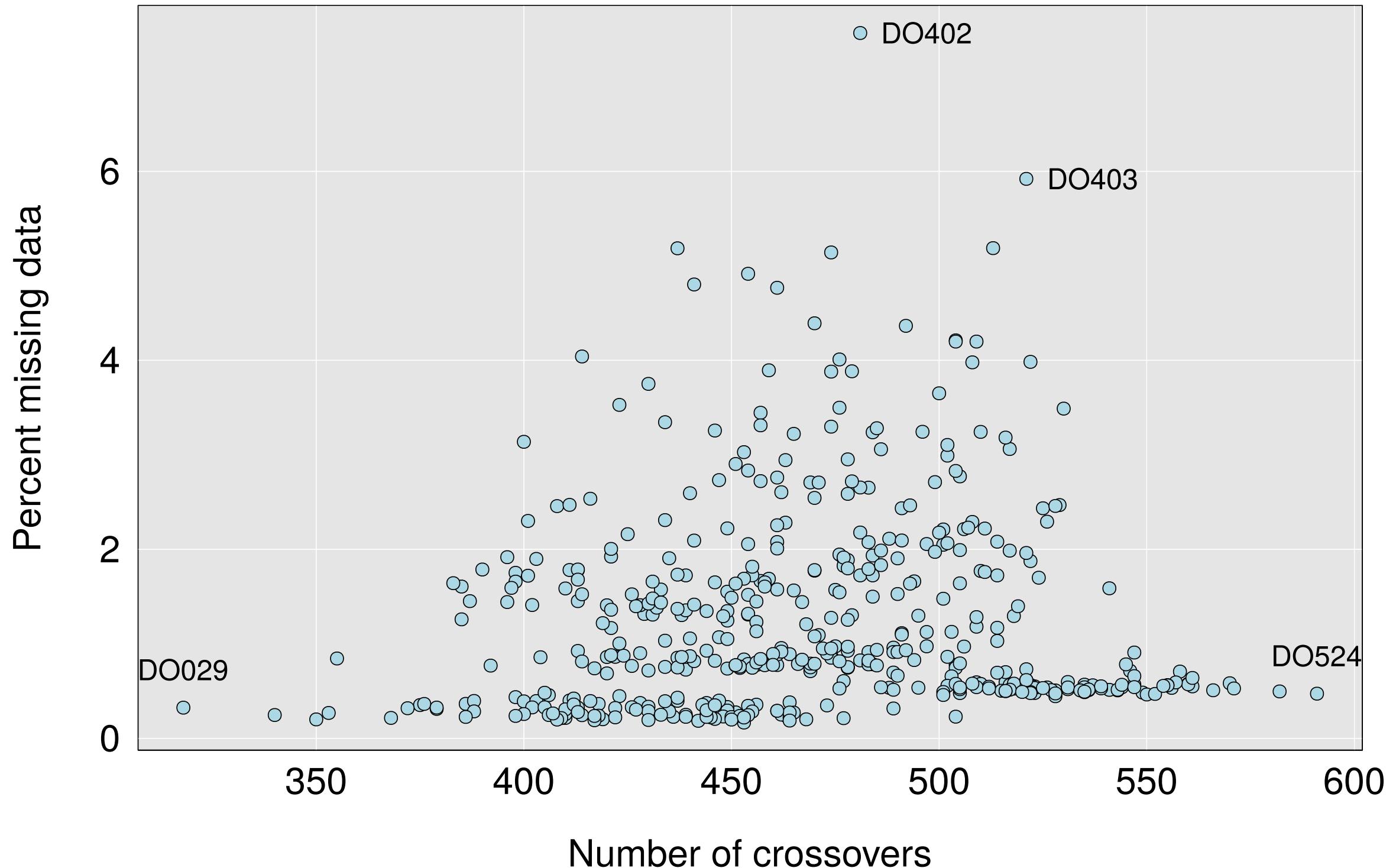
Percent missing vs number of crossovers



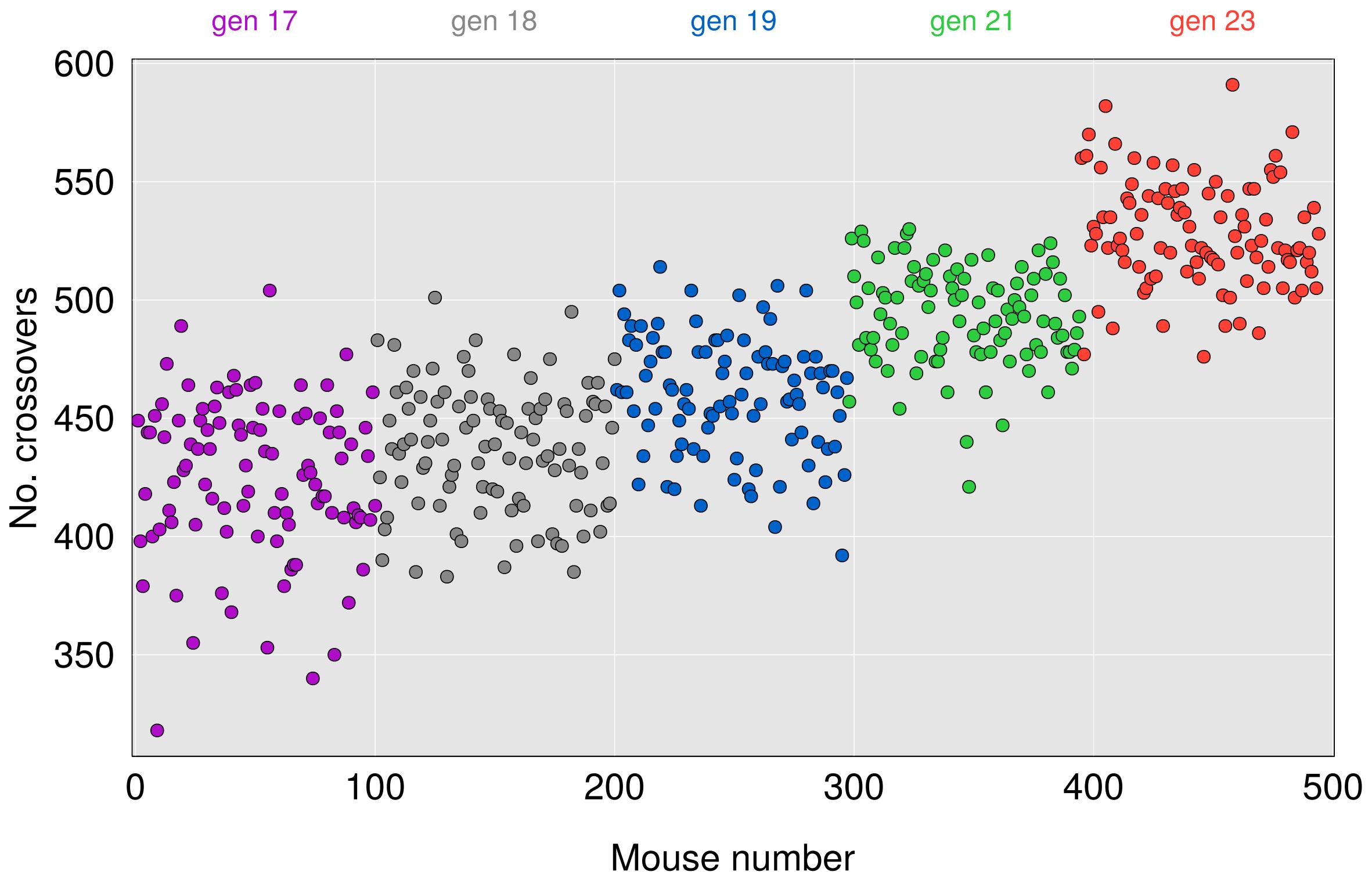
Percent missing vs number of crossovers



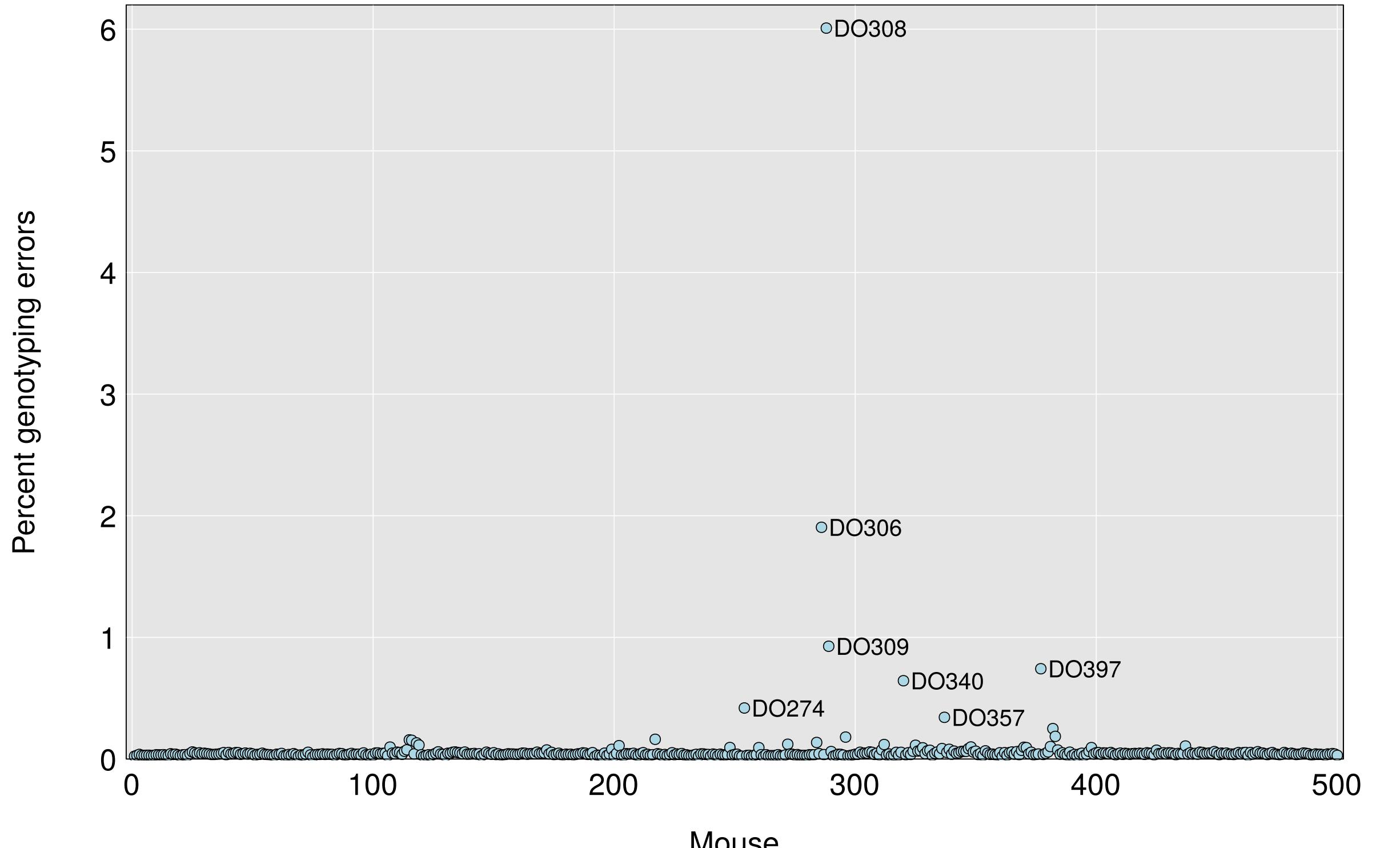
Percent missing vs number of crossovers



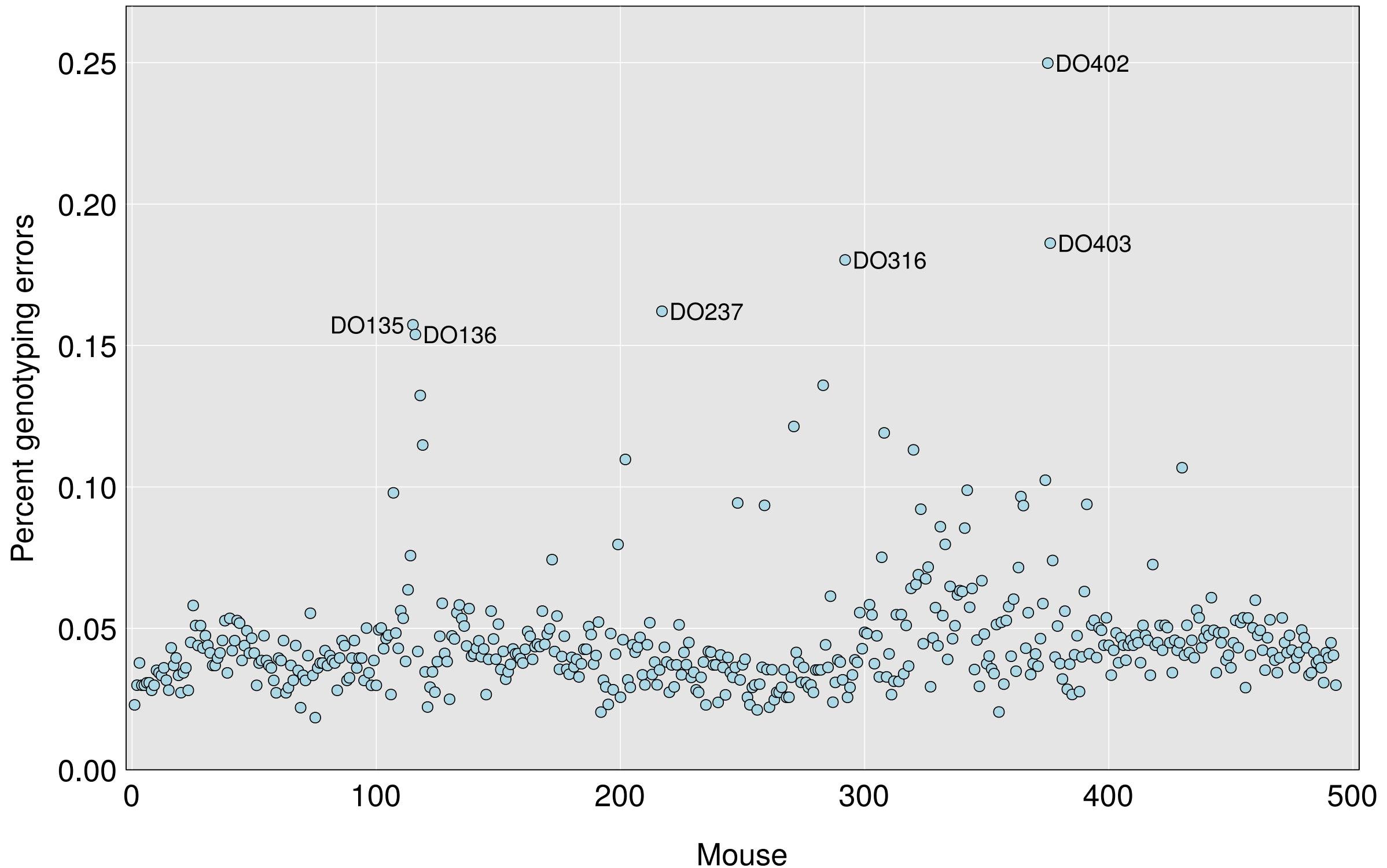
No. crossovers by generation



Estimated percent of genotyping errors

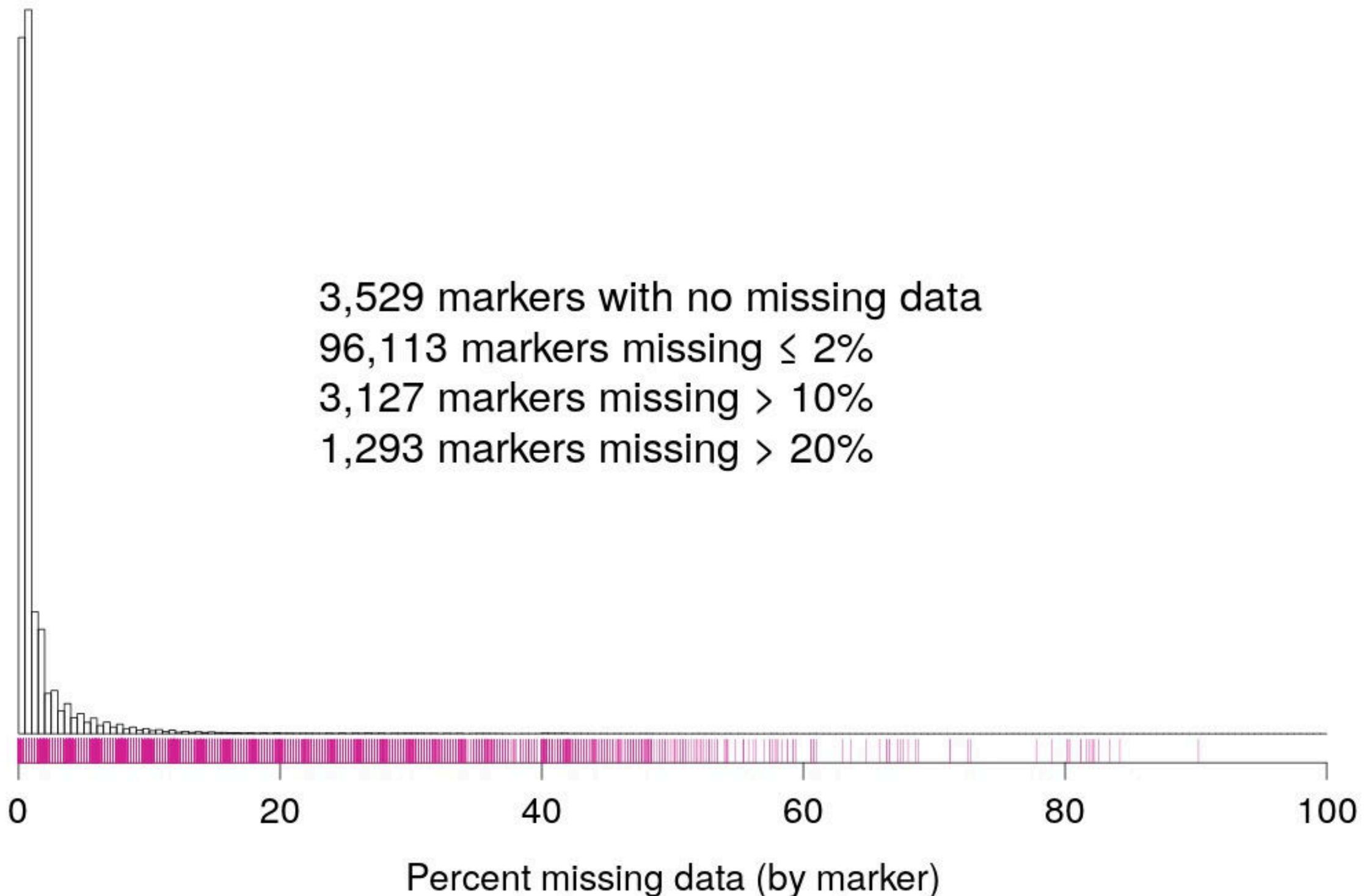


Estimated percent of genotyping errors

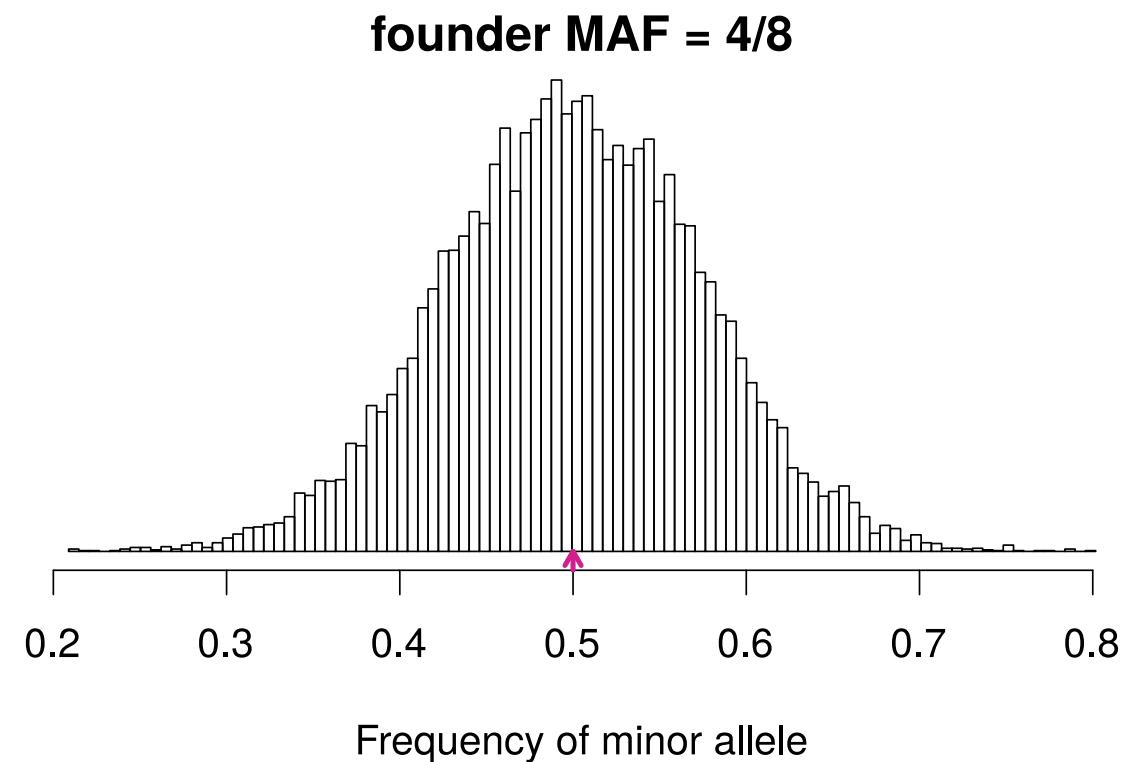
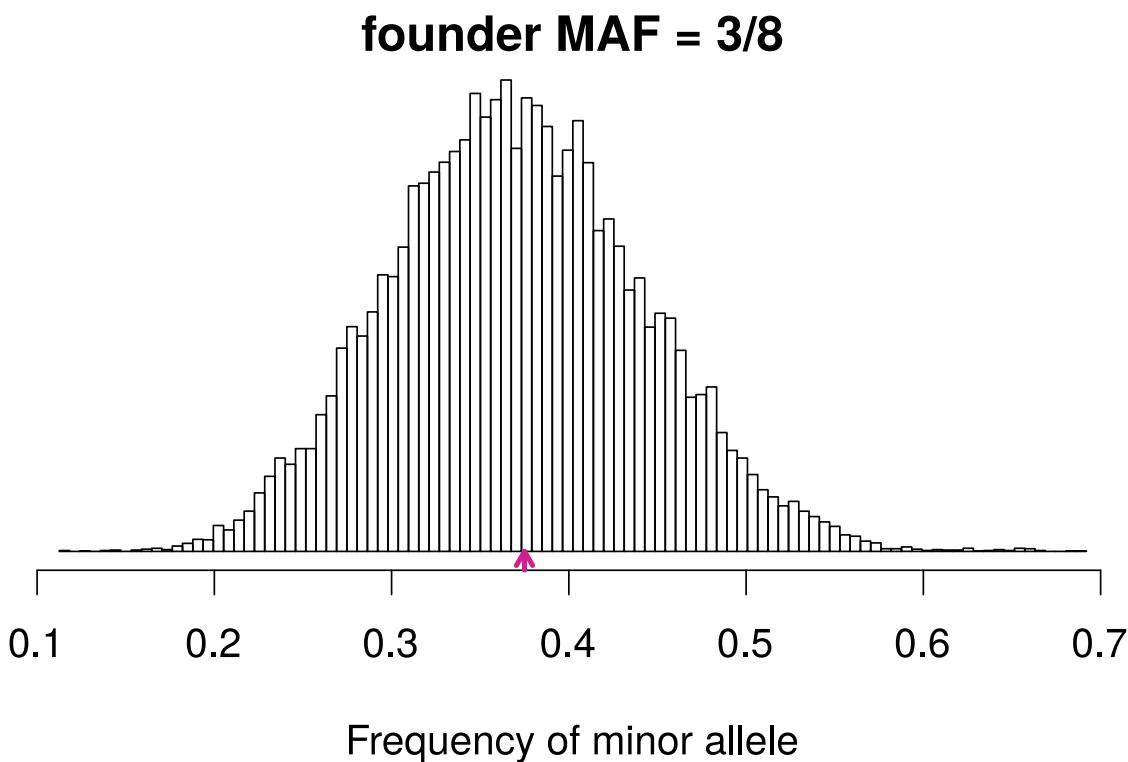
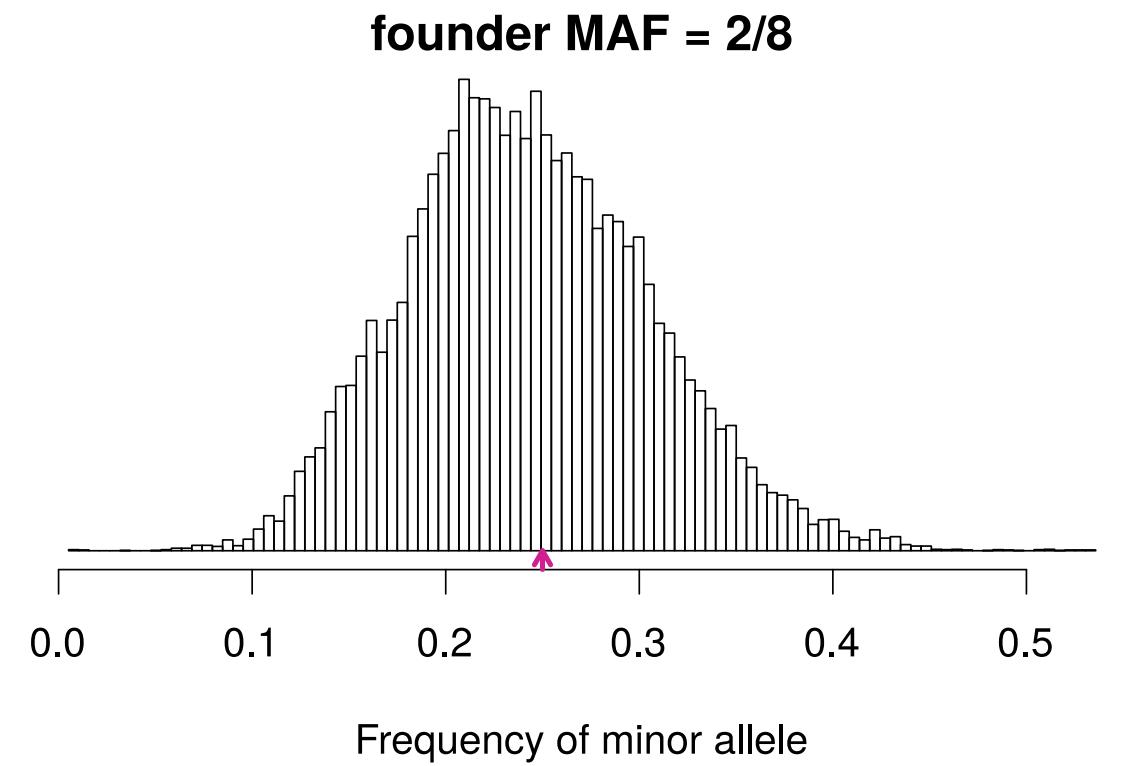
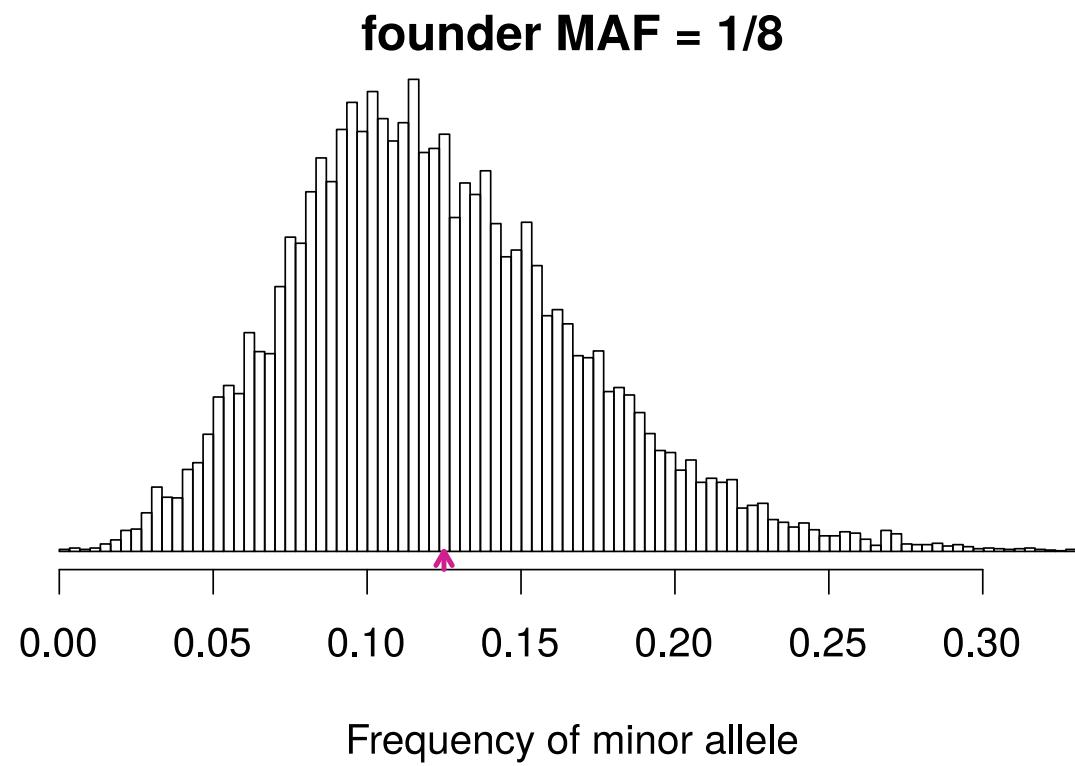


Marker quality

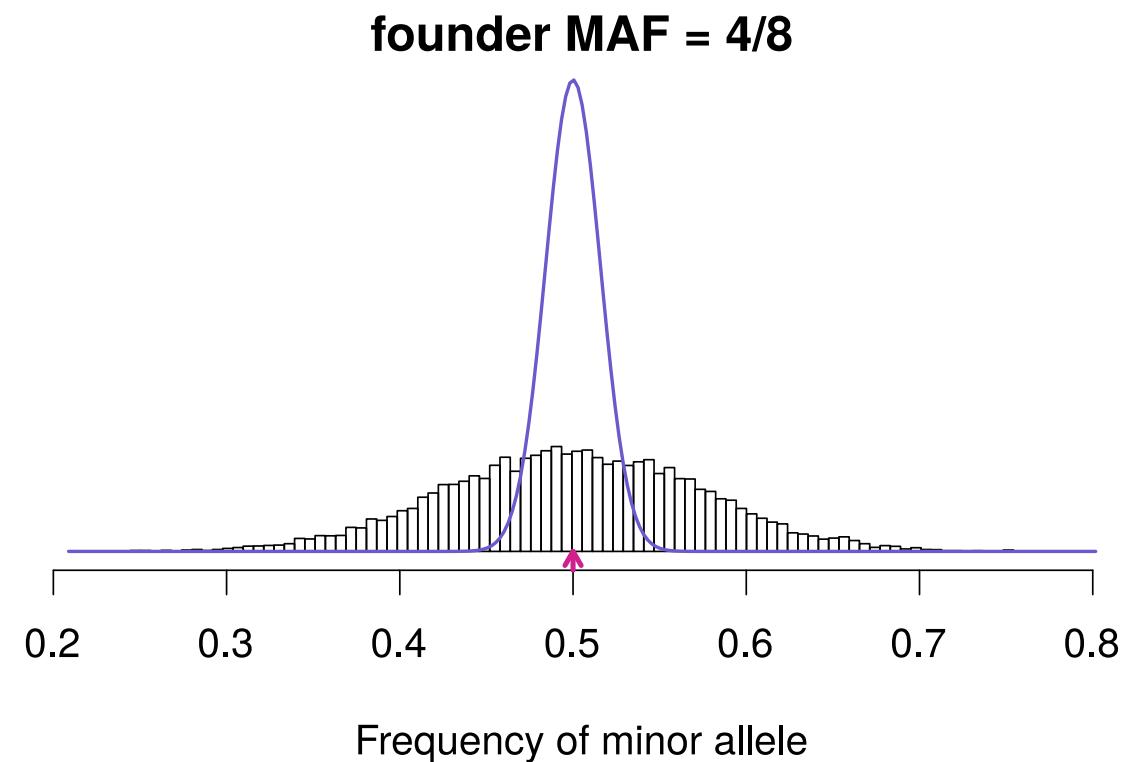
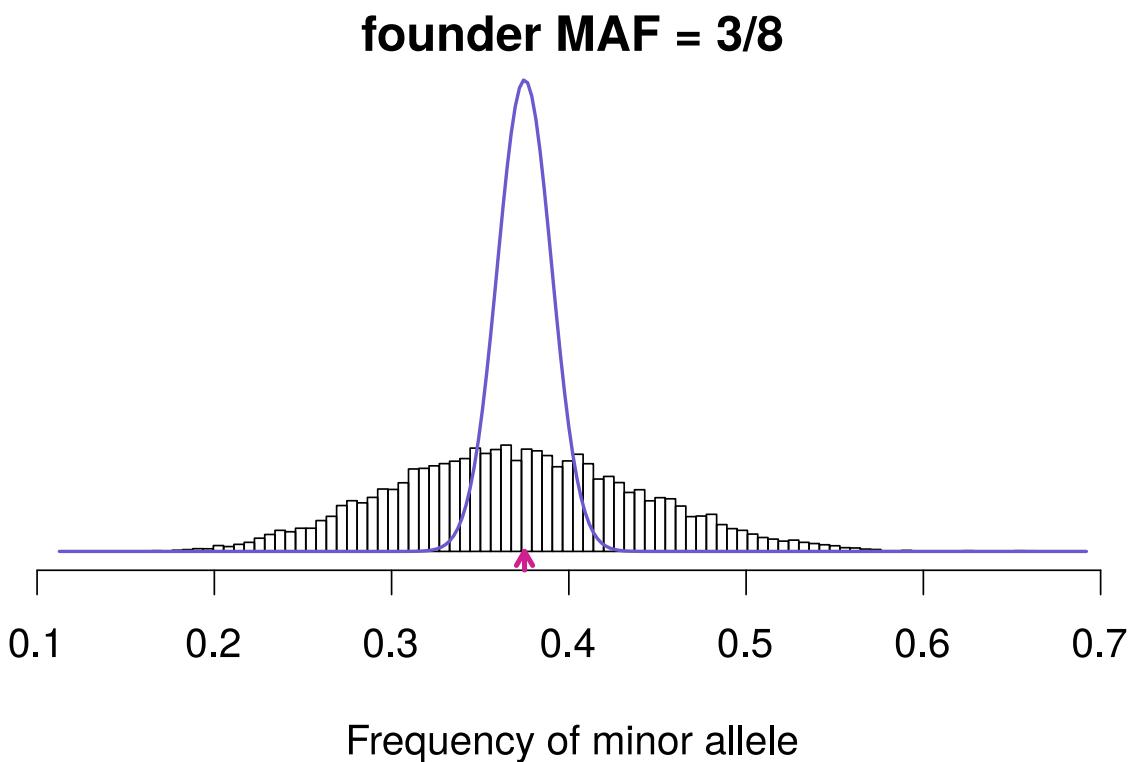
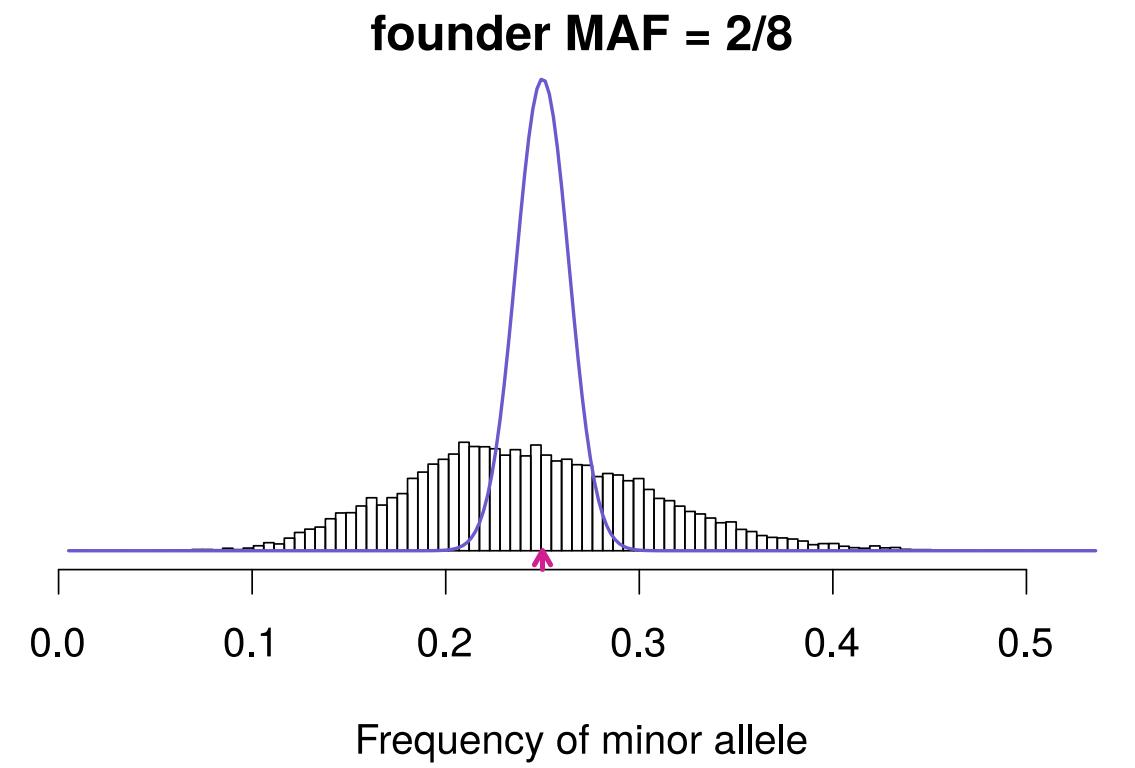
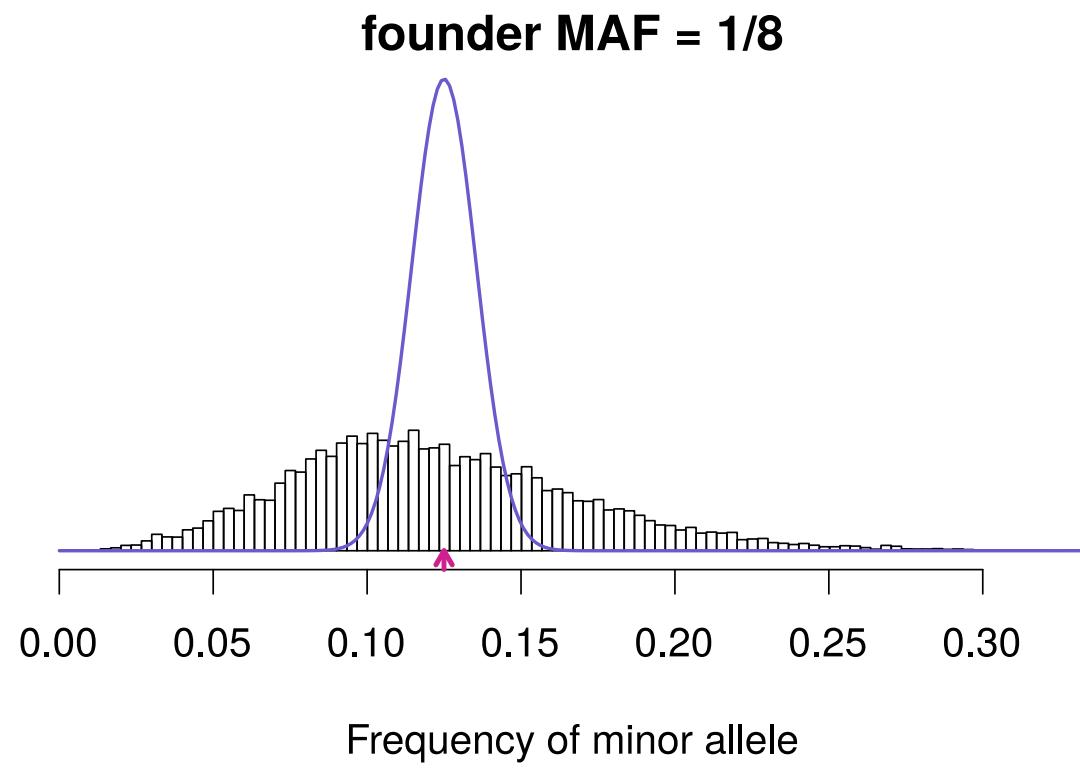
Proportion missing data



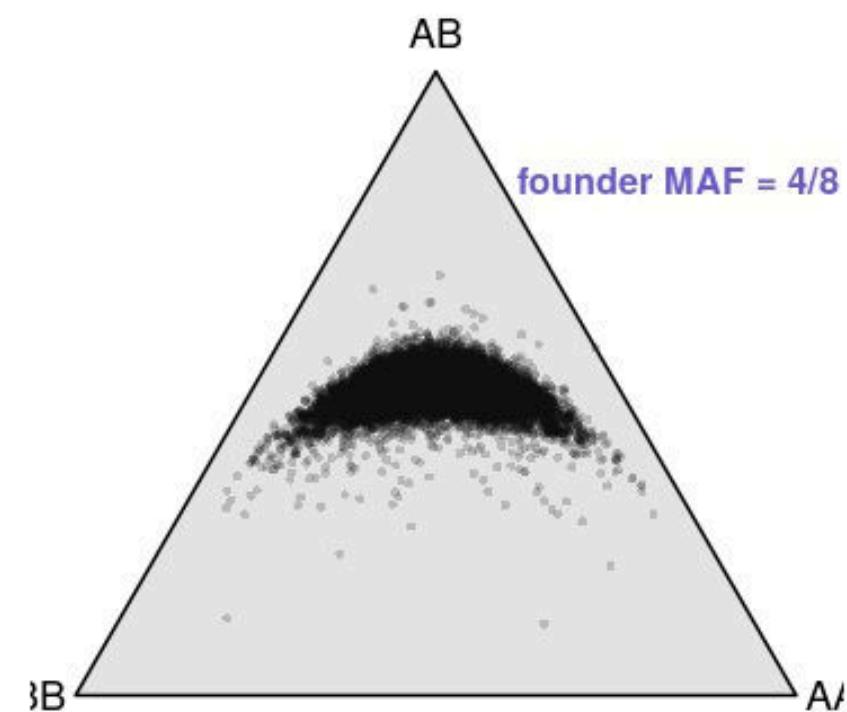
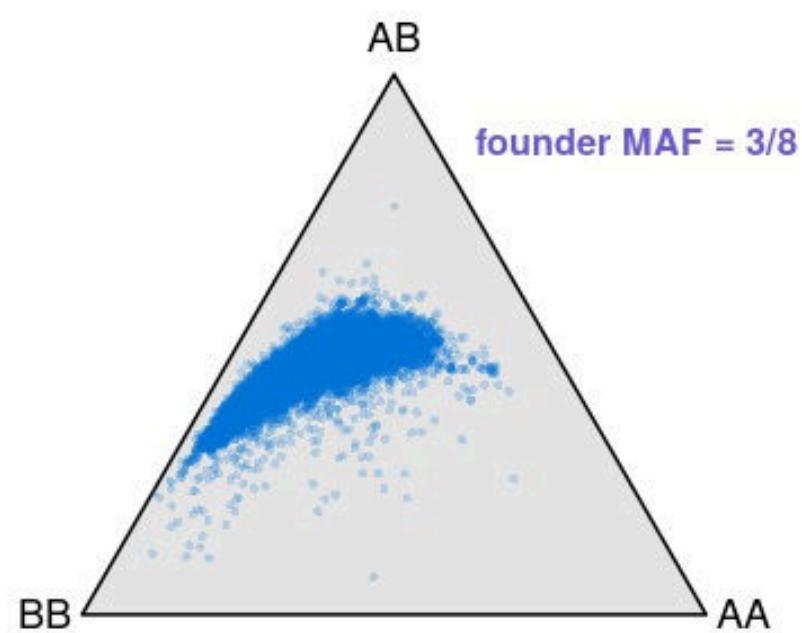
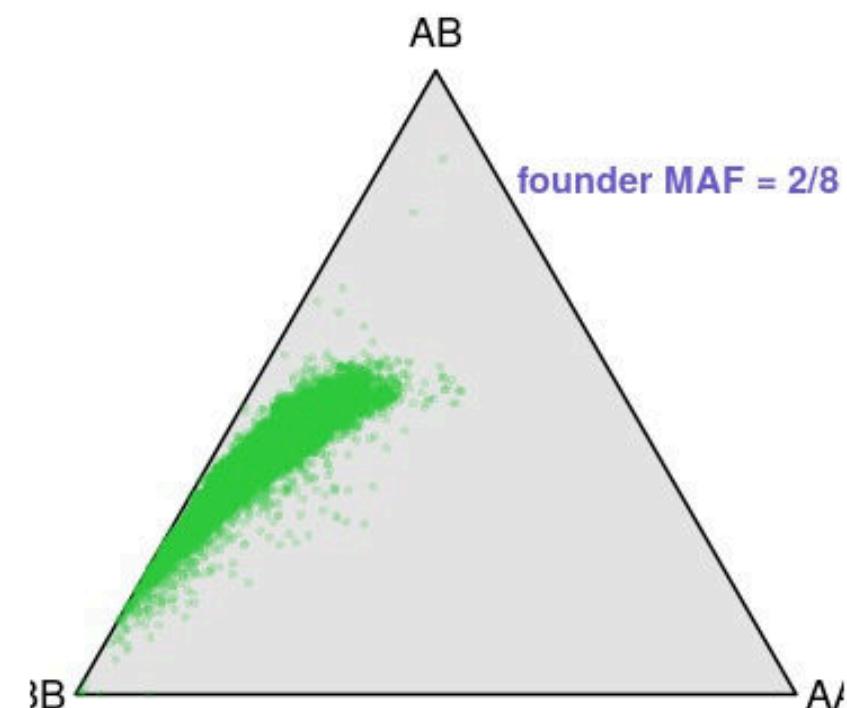
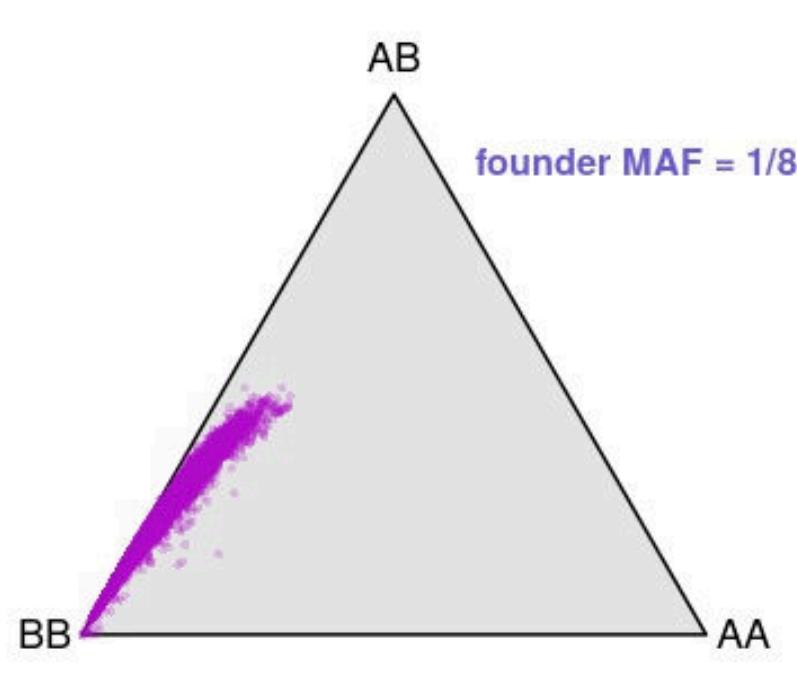
Allele frequencies, by marker



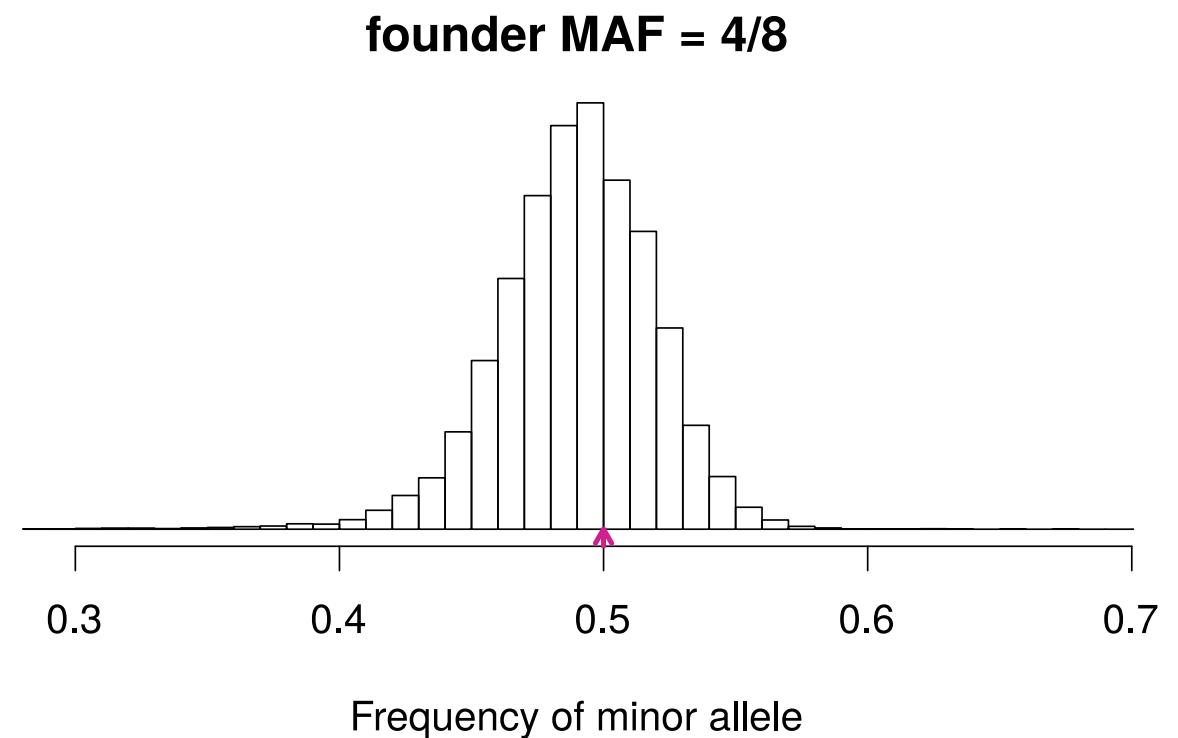
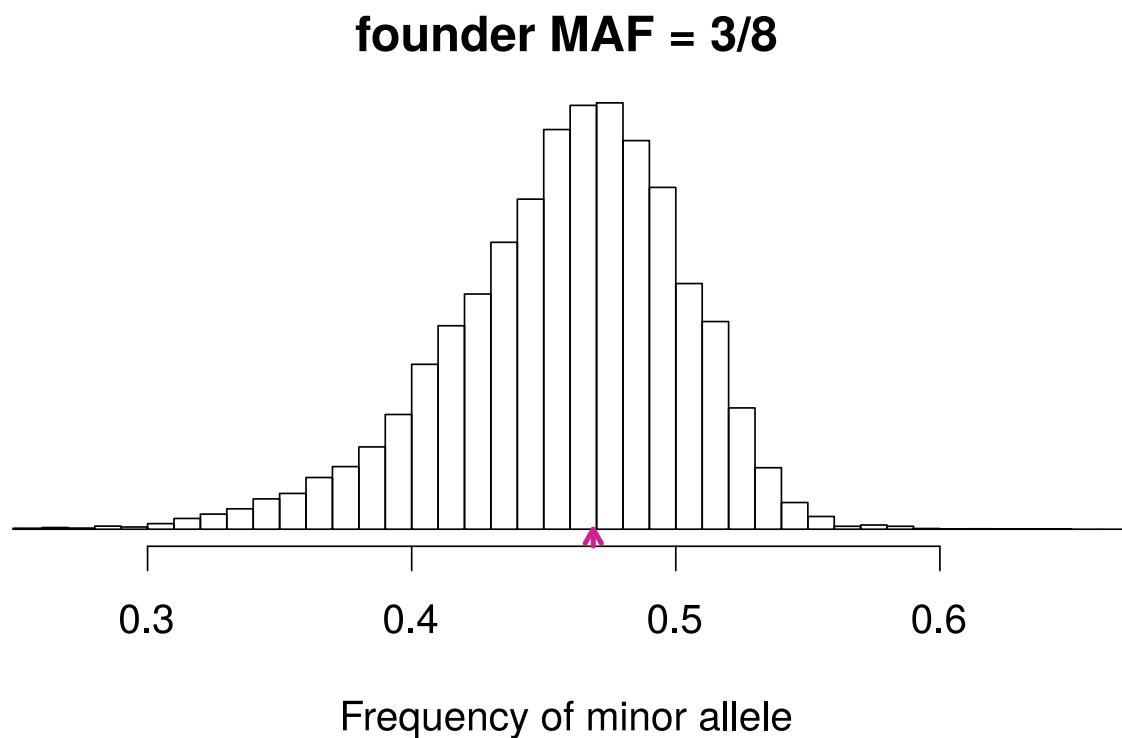
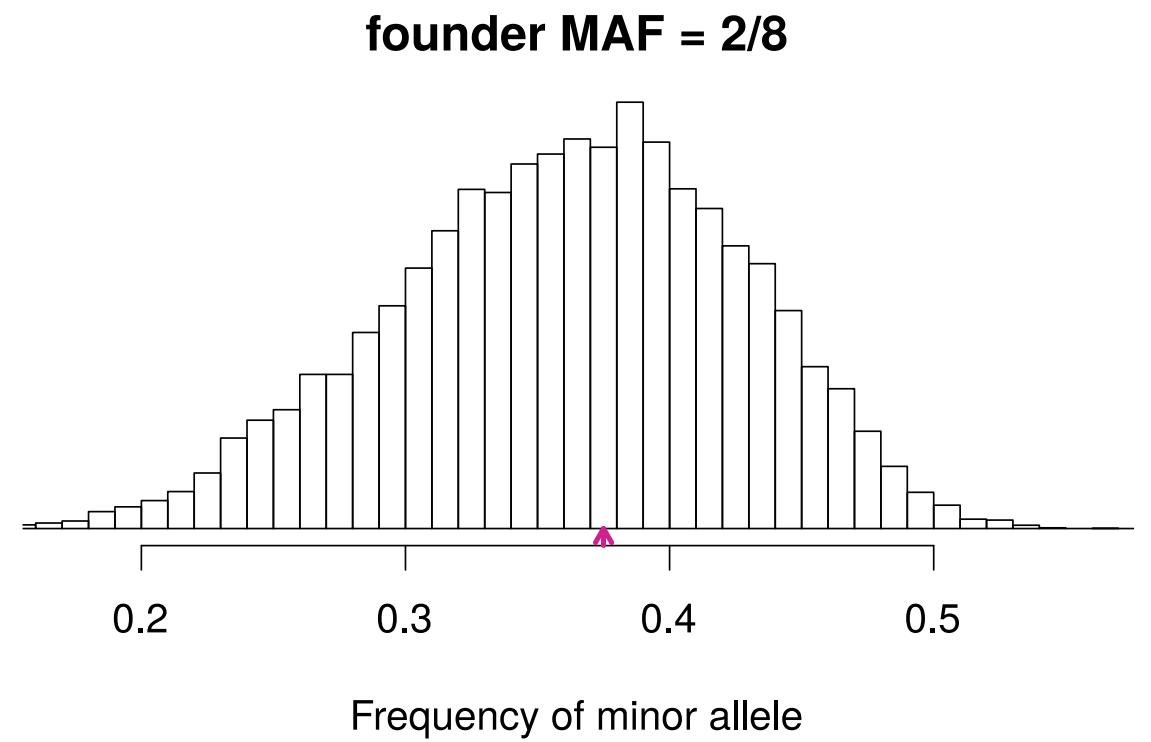
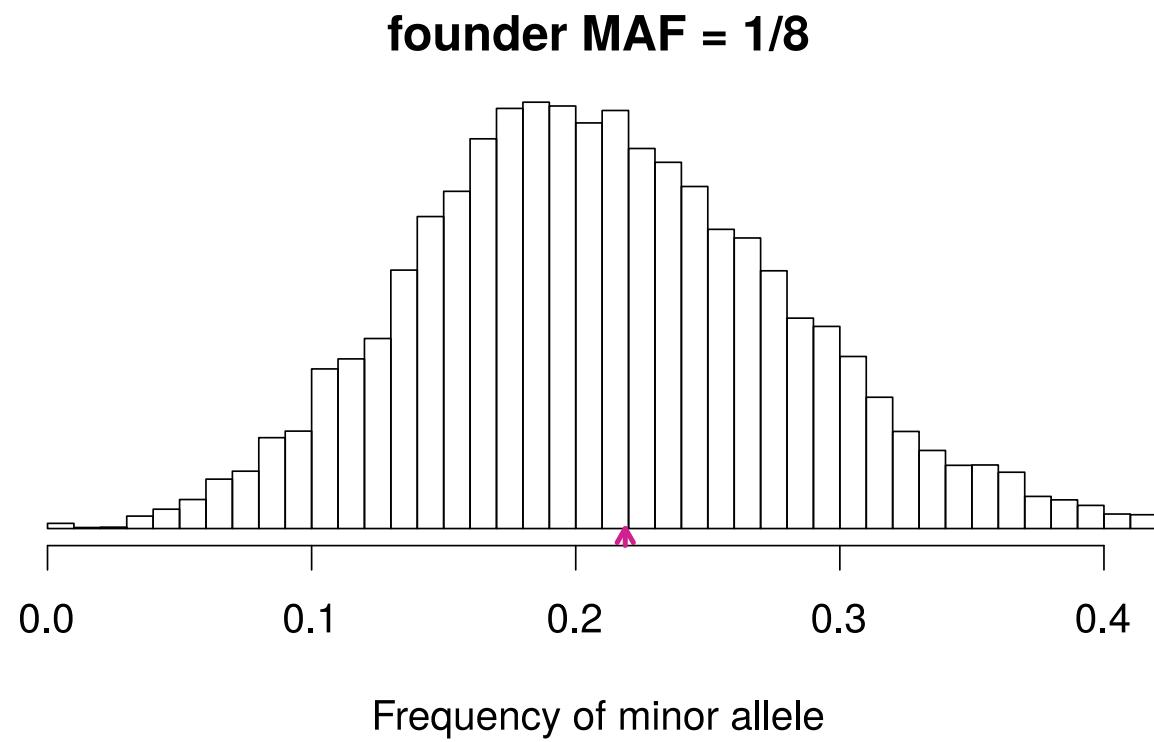
Allele frequencies, by marker



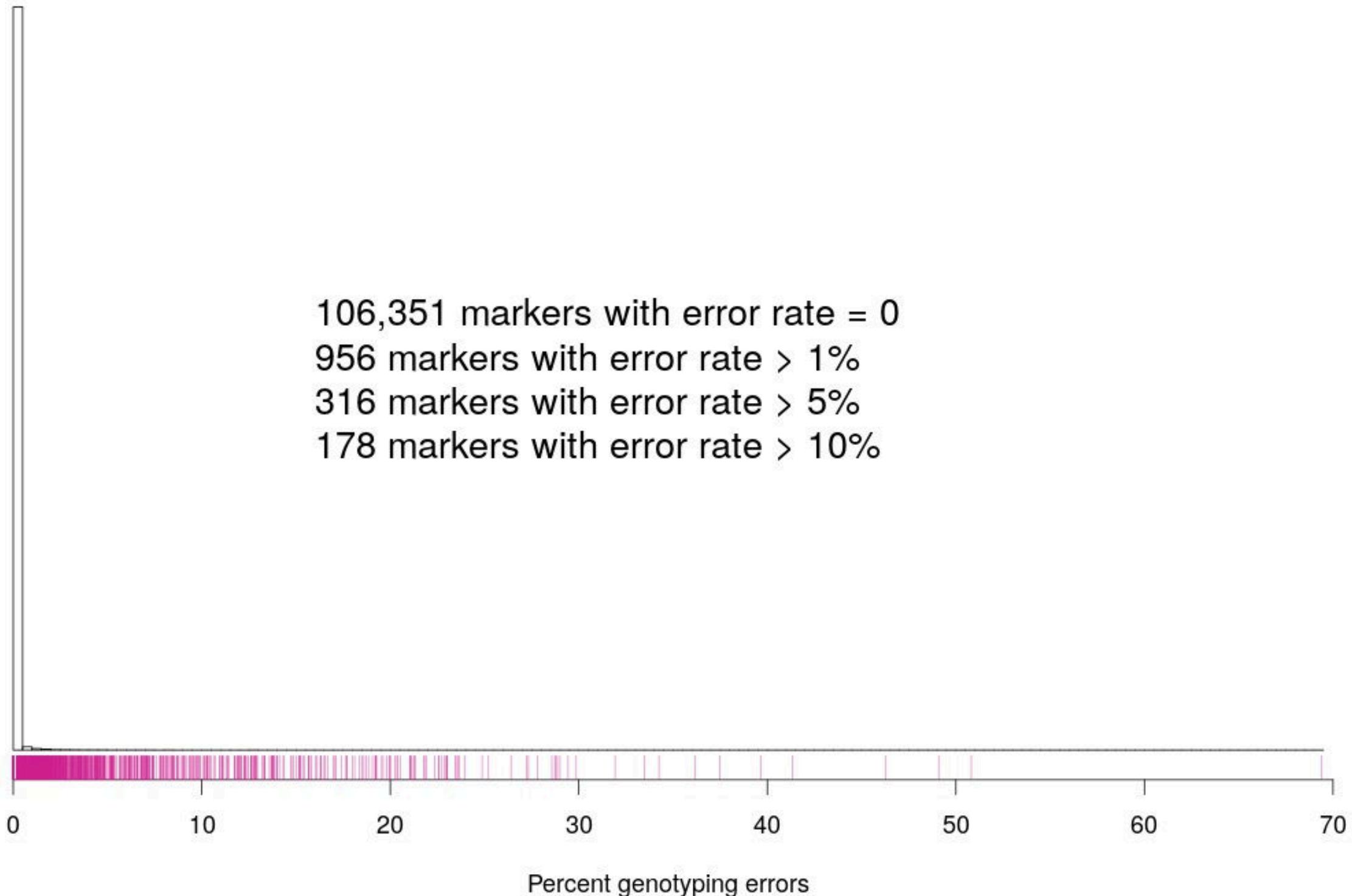
Genotype frequencies, by marker



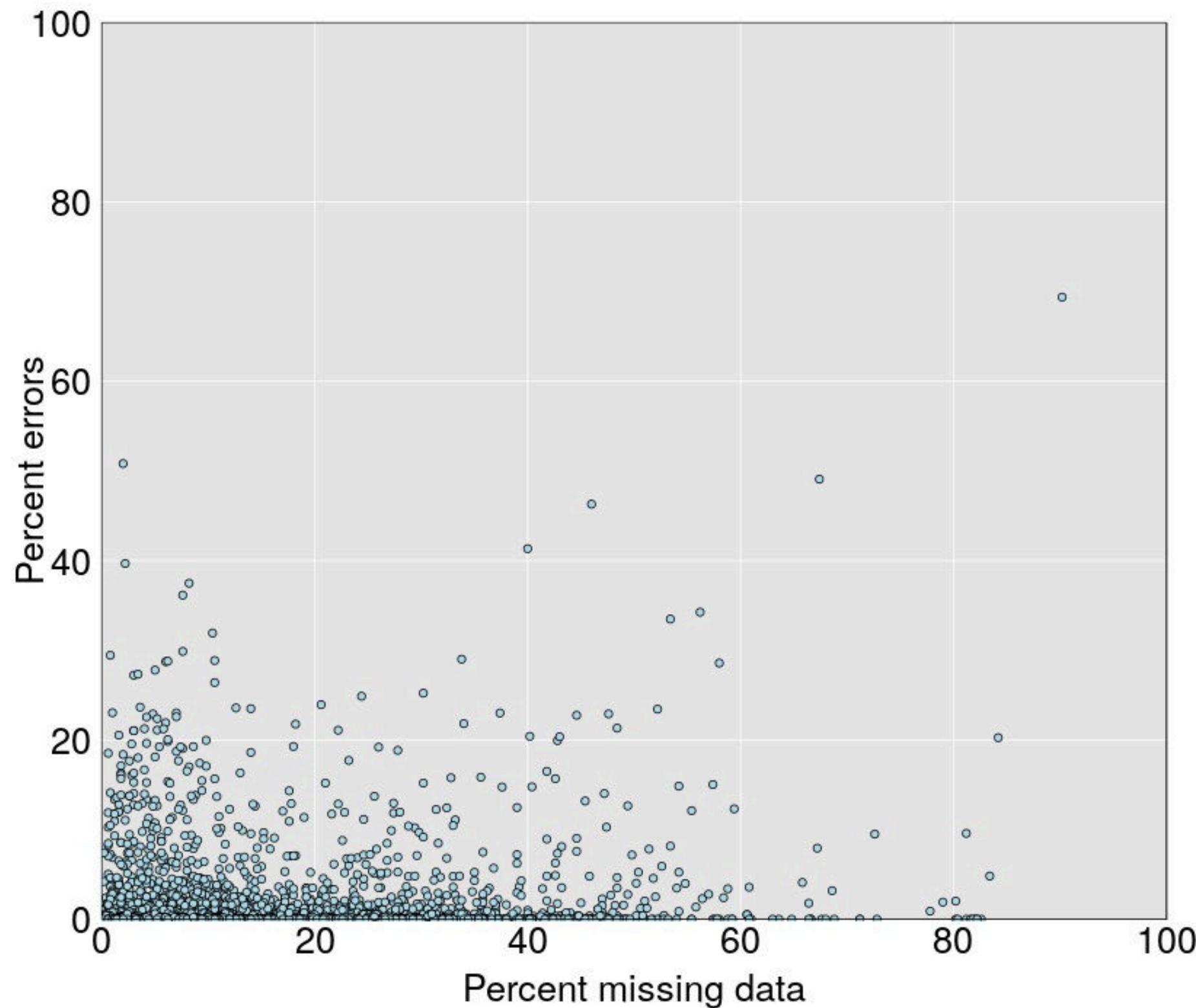
Heterozygosities, by marker



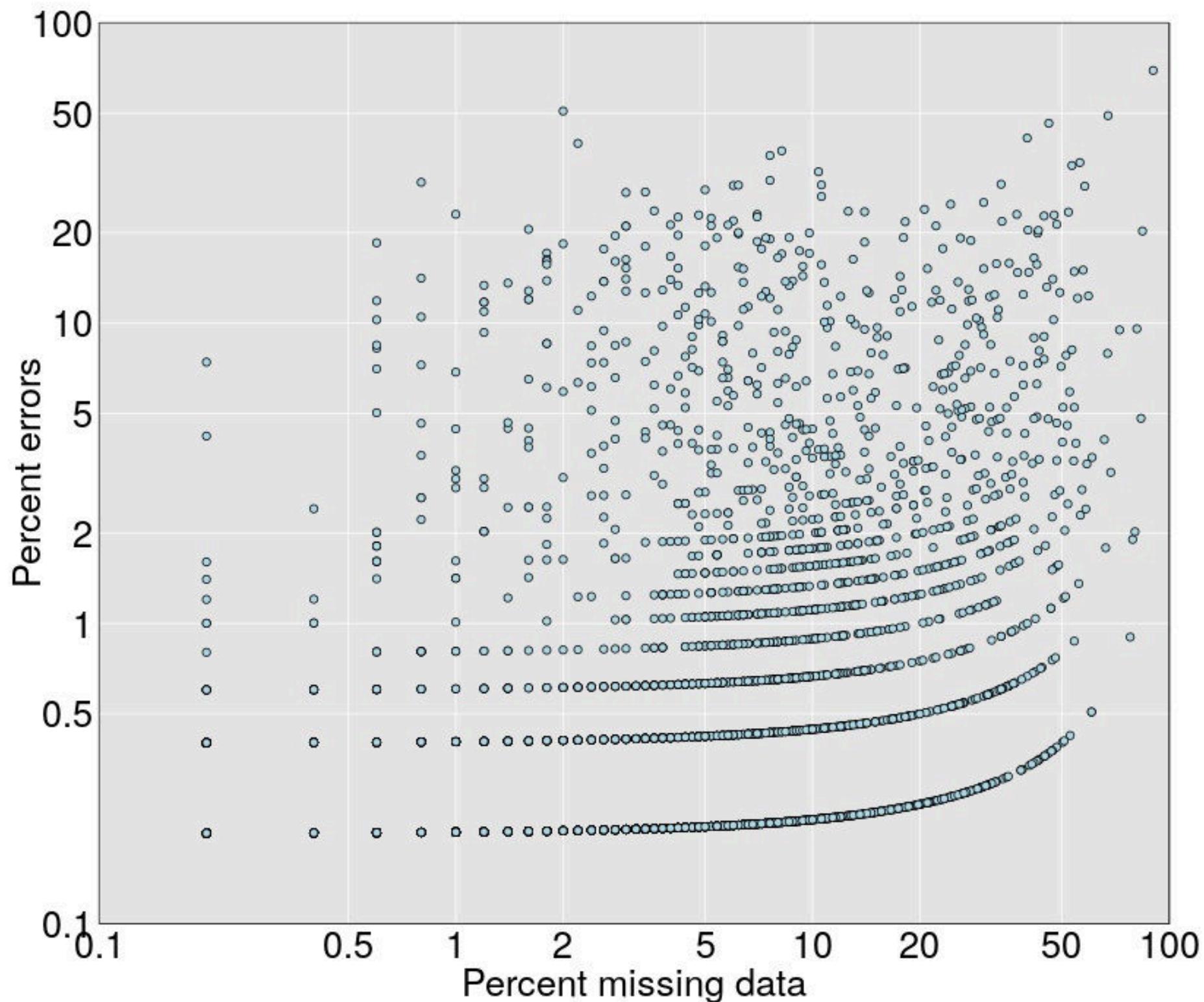
Genotyping error rates



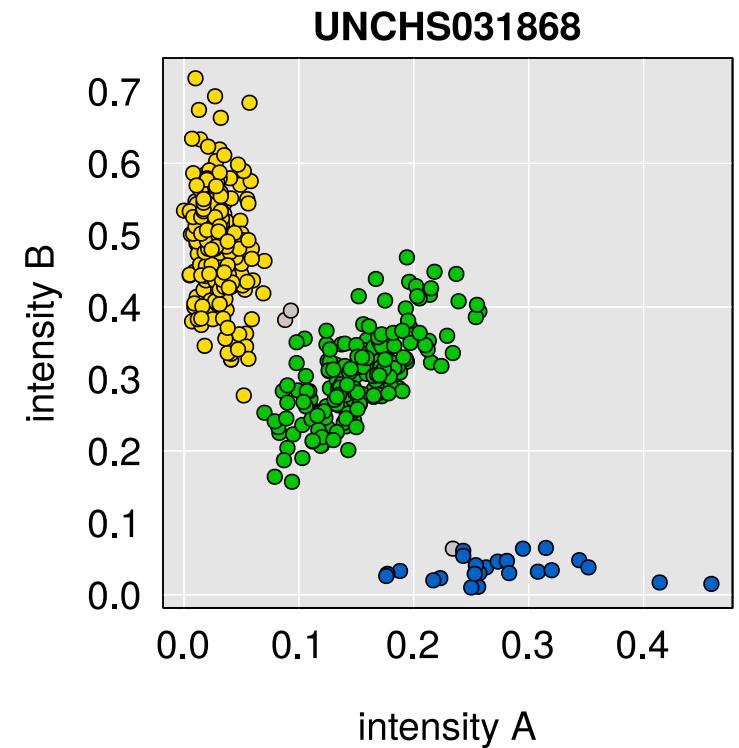
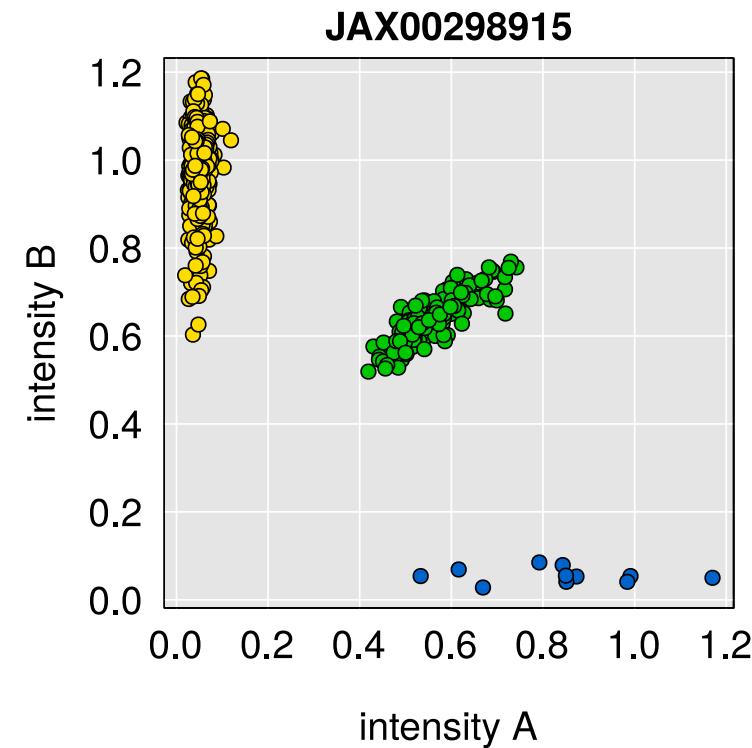
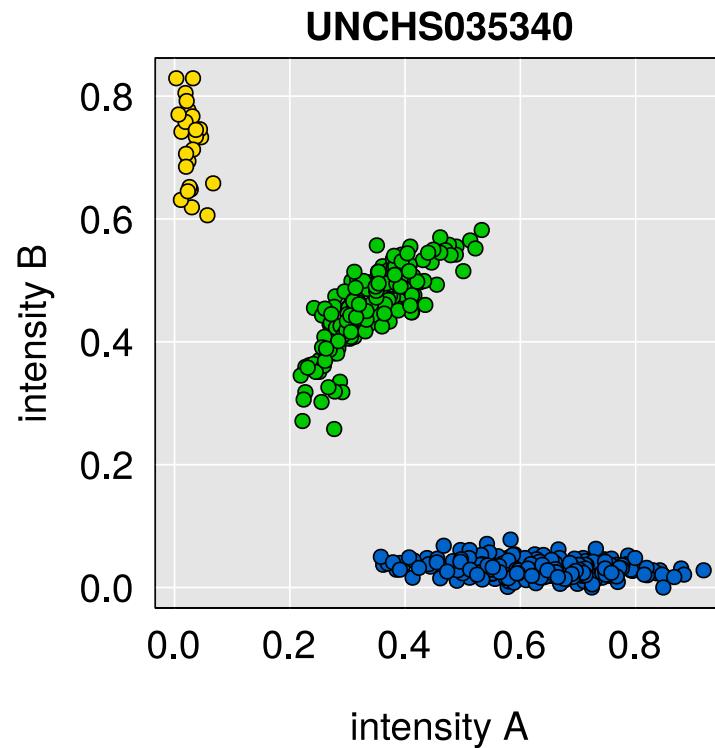
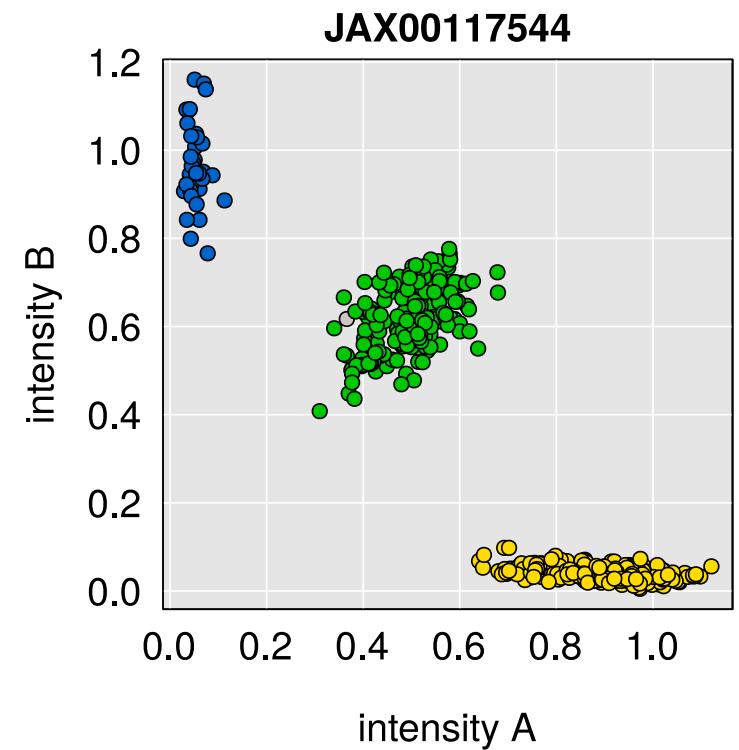
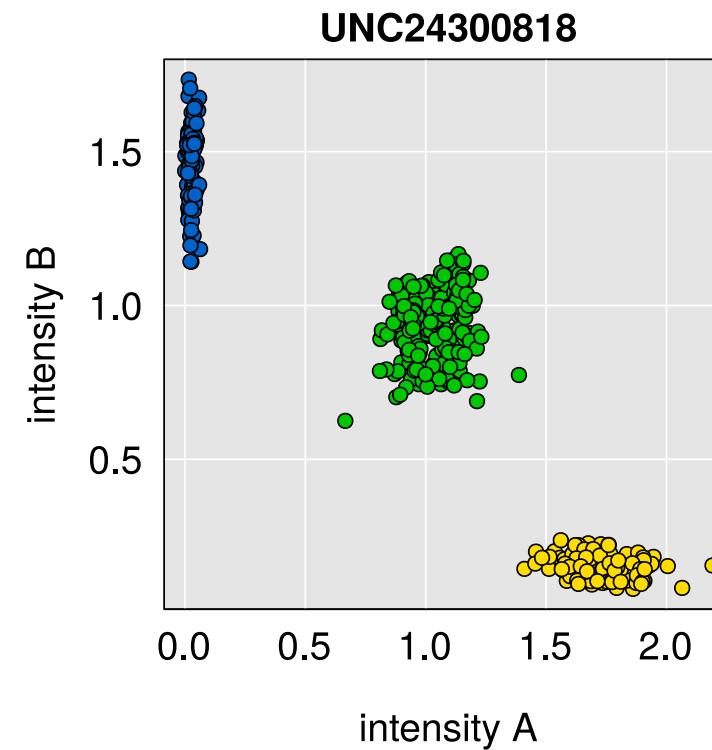
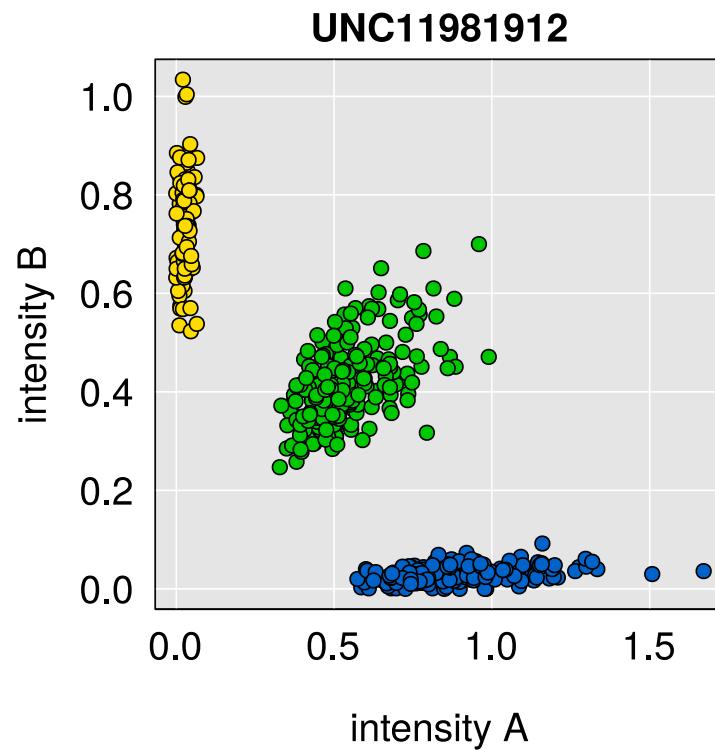
Genotyping error rate vs percent missing



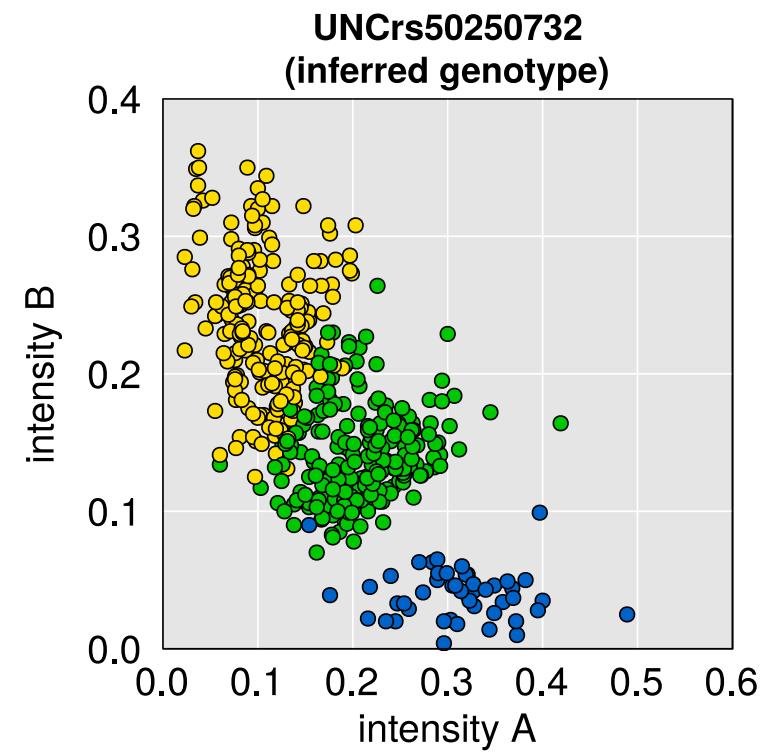
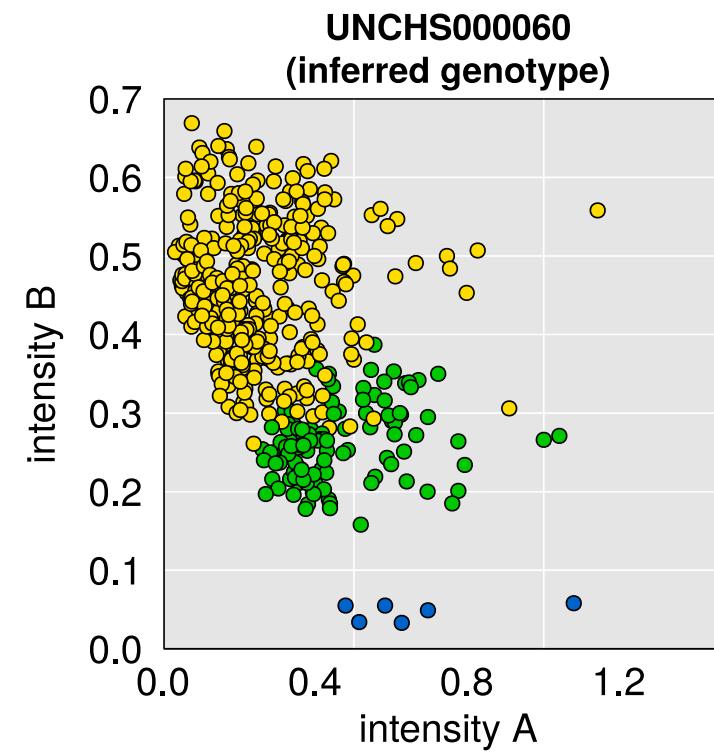
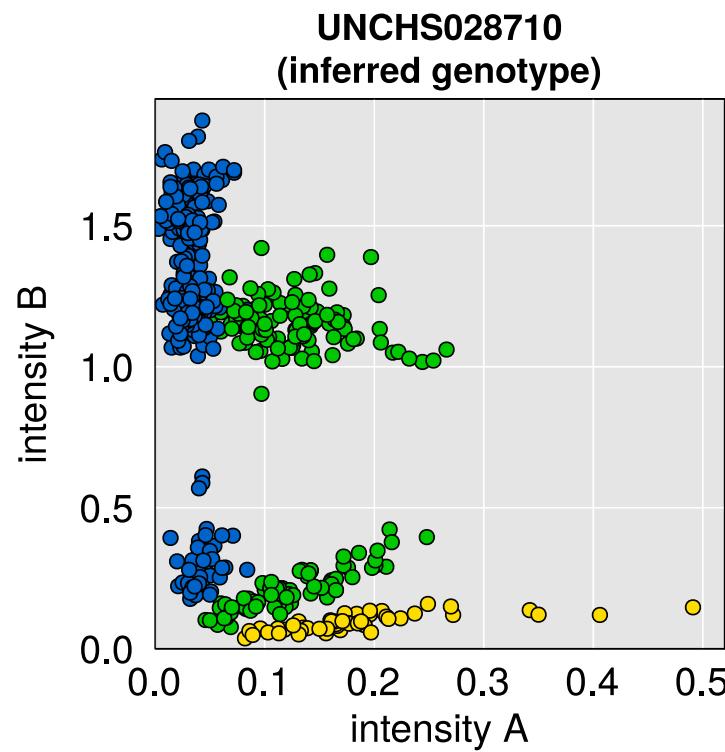
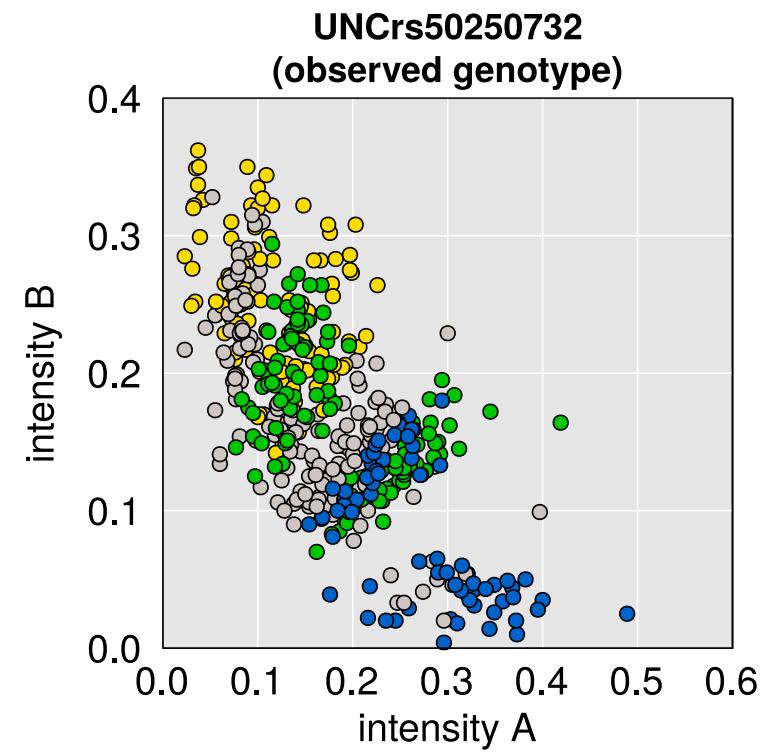
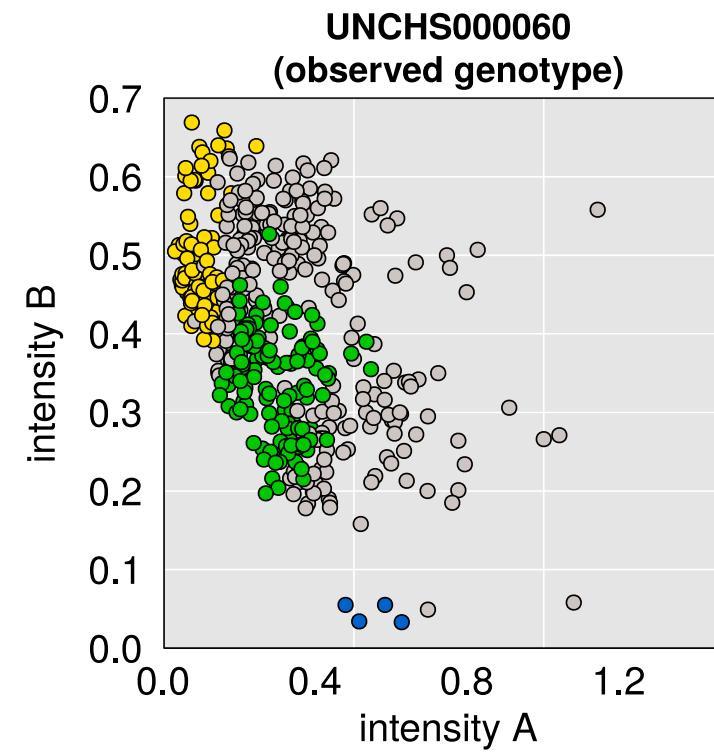
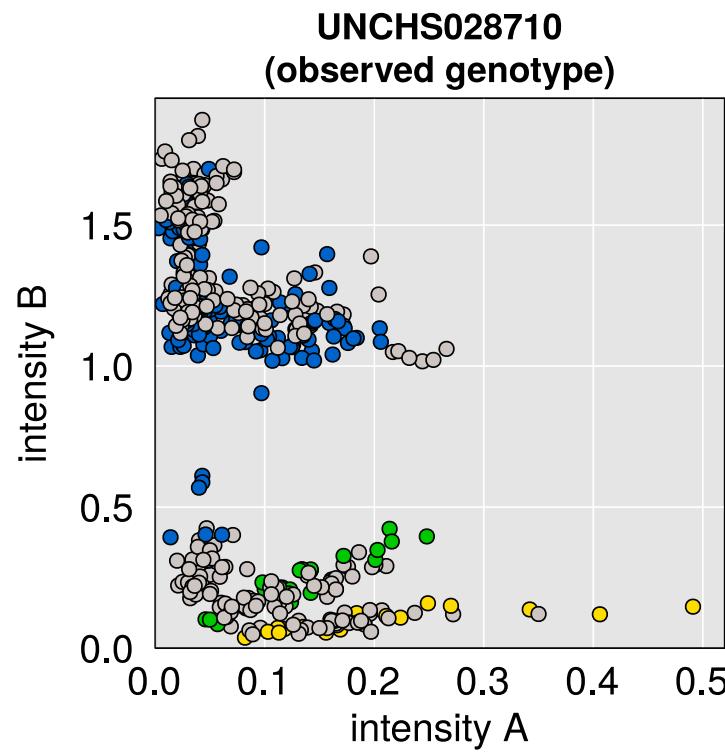
Genotyping error rate vs percent missing



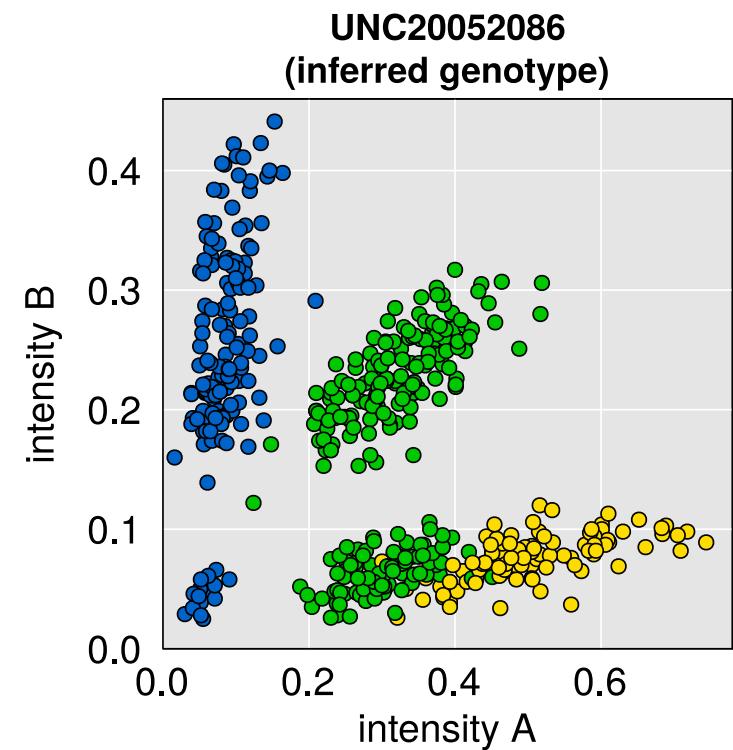
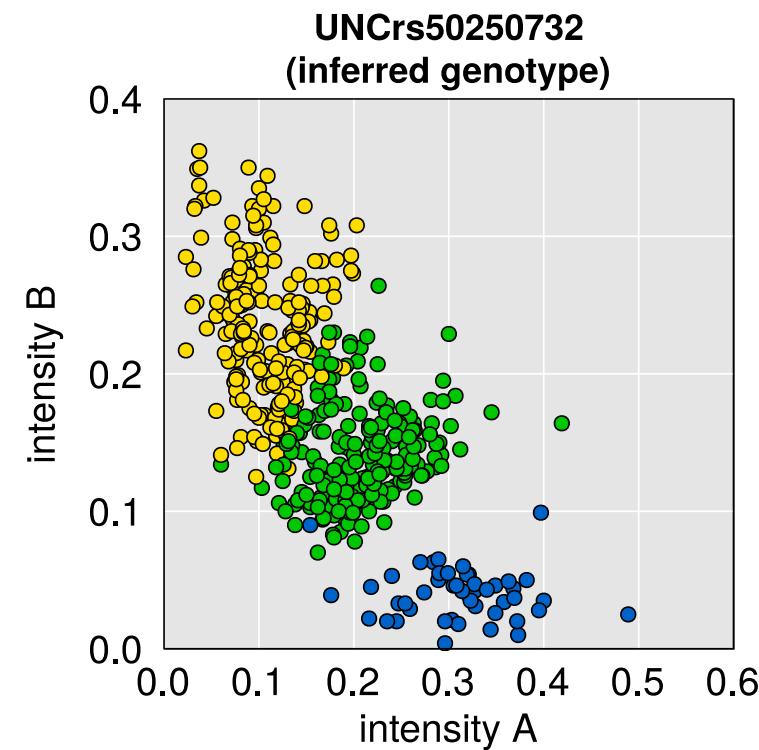
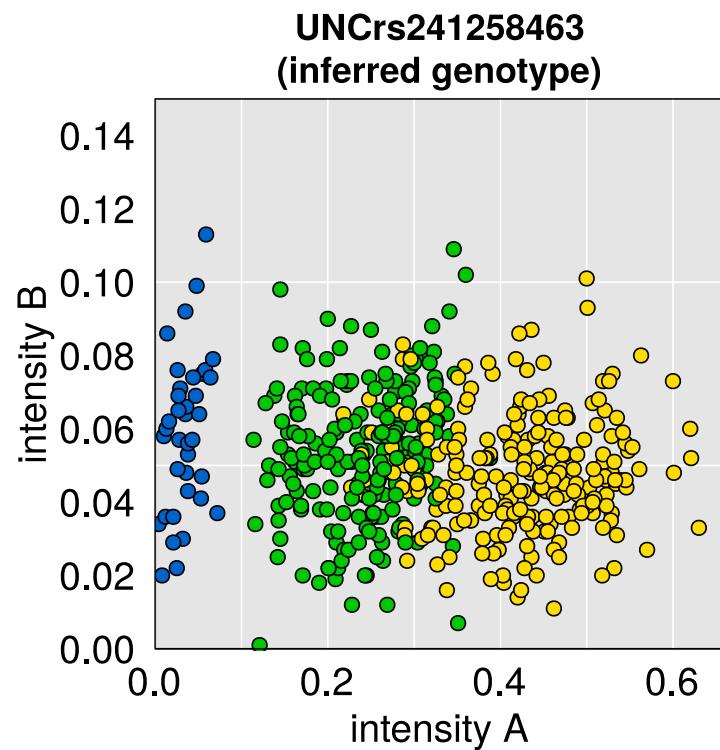
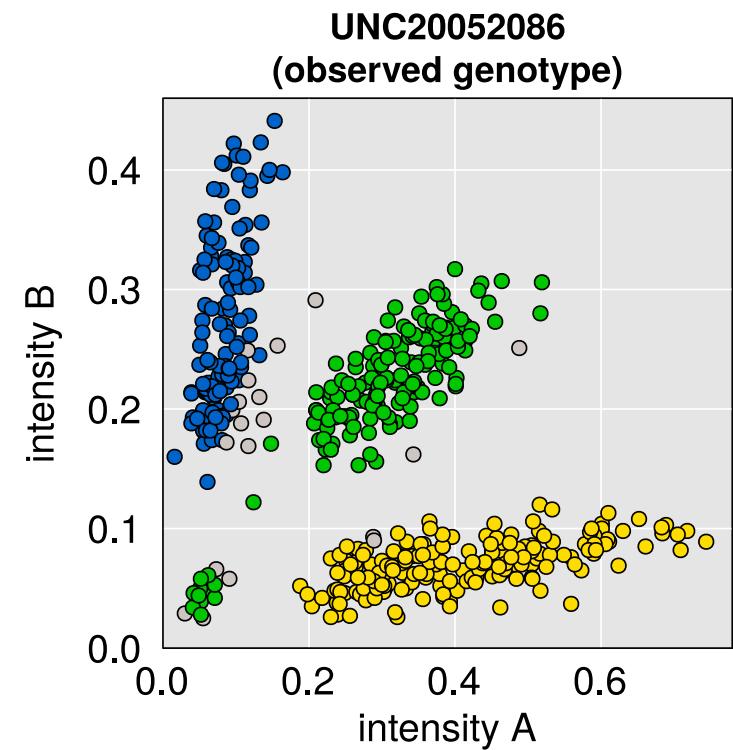
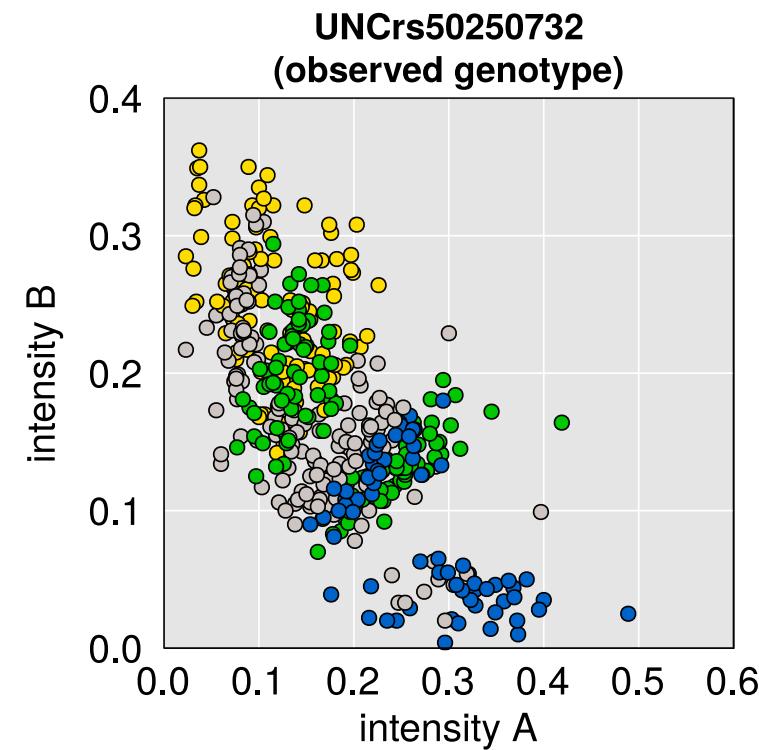
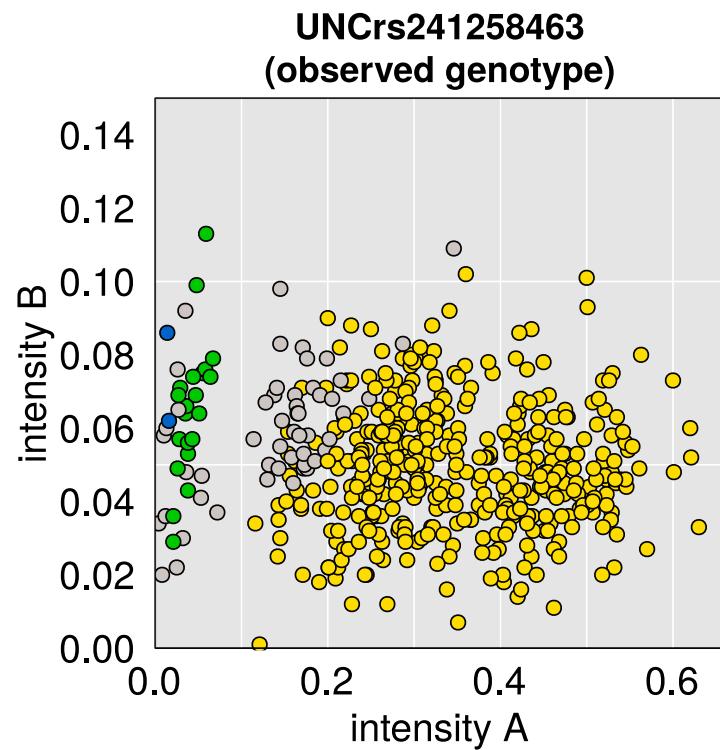
Nice markers



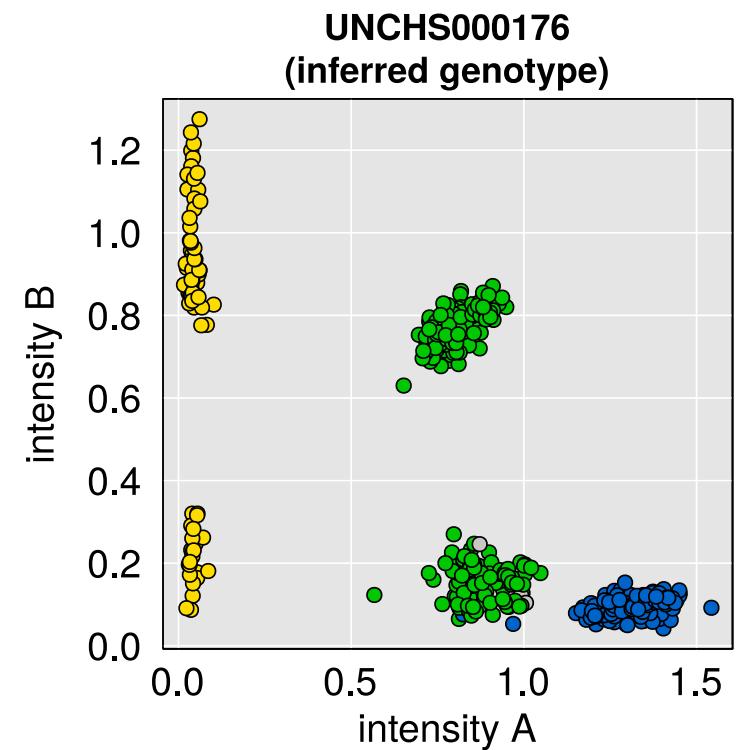
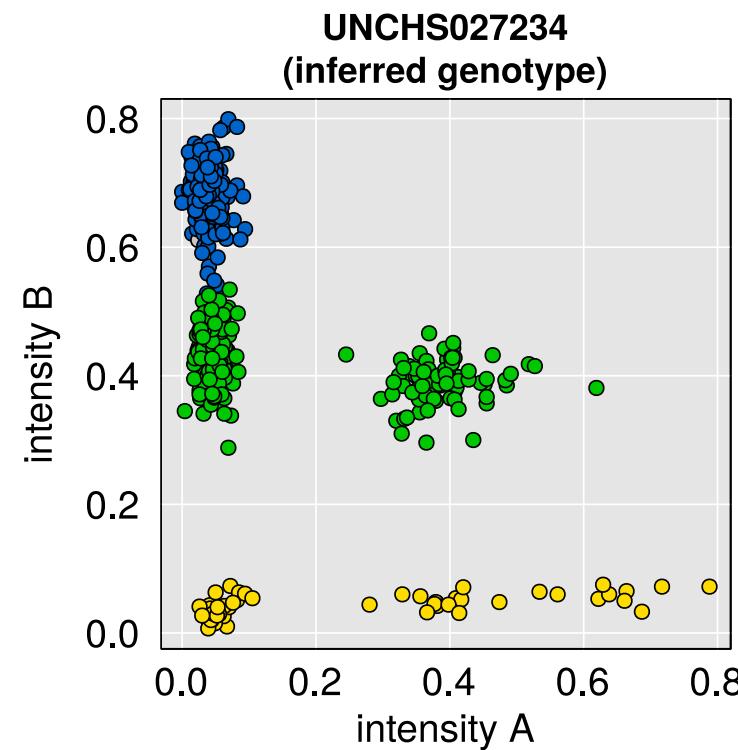
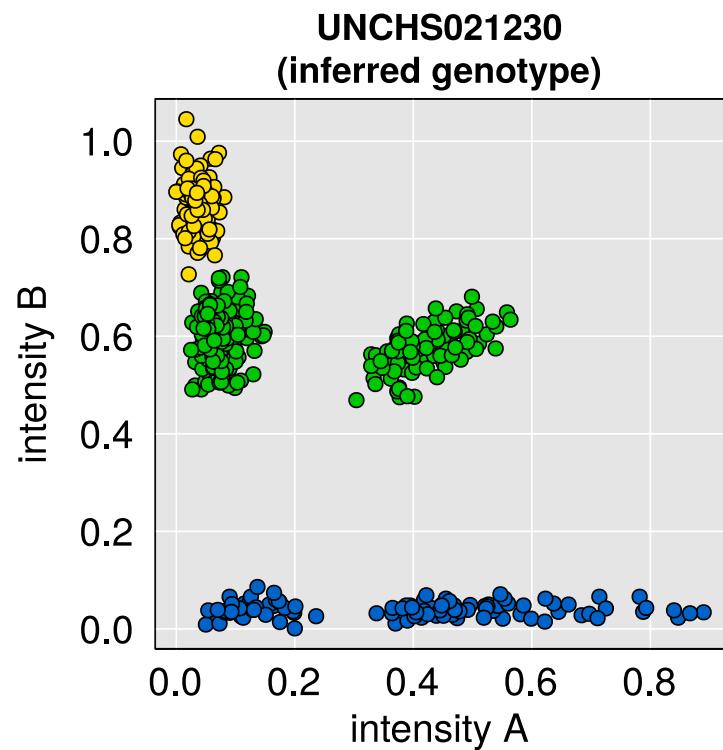
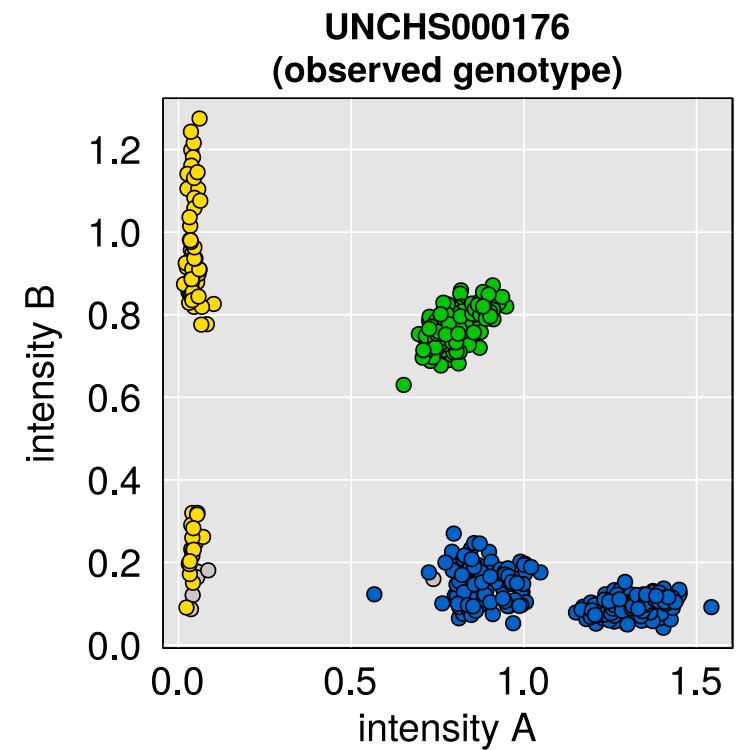
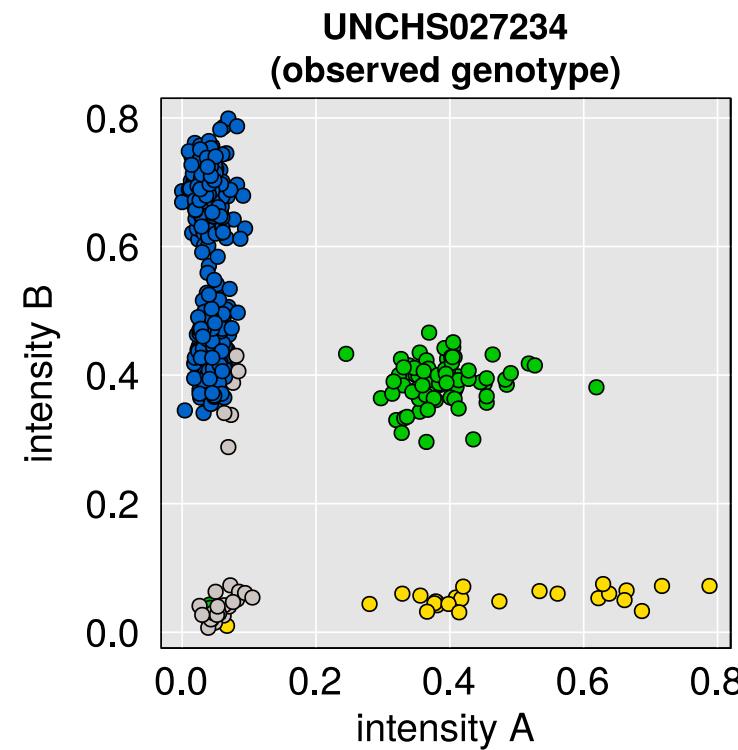
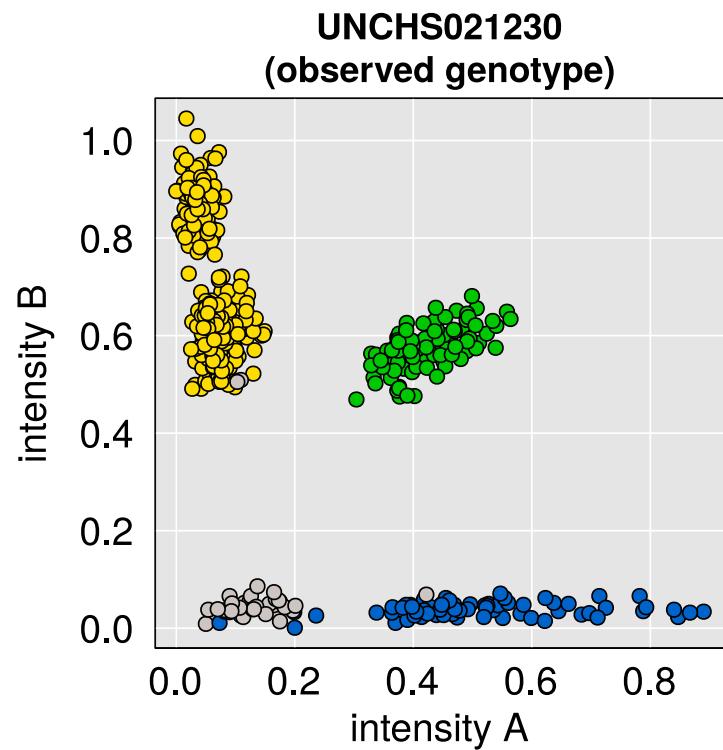
Crap markers



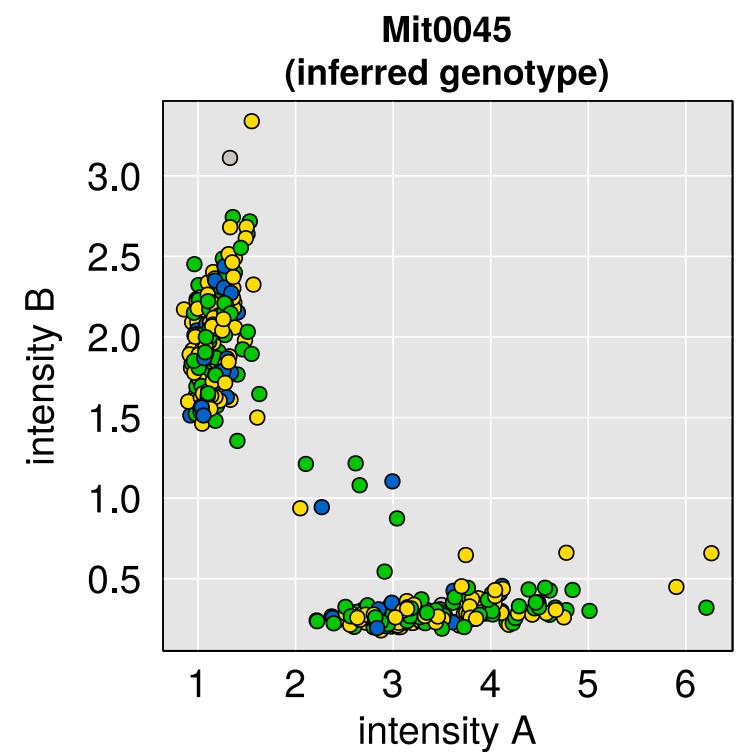
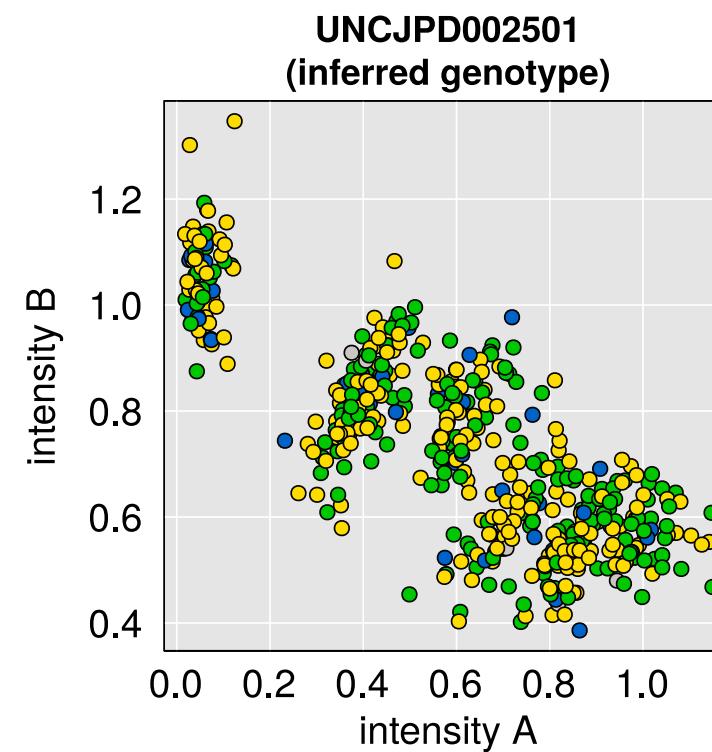
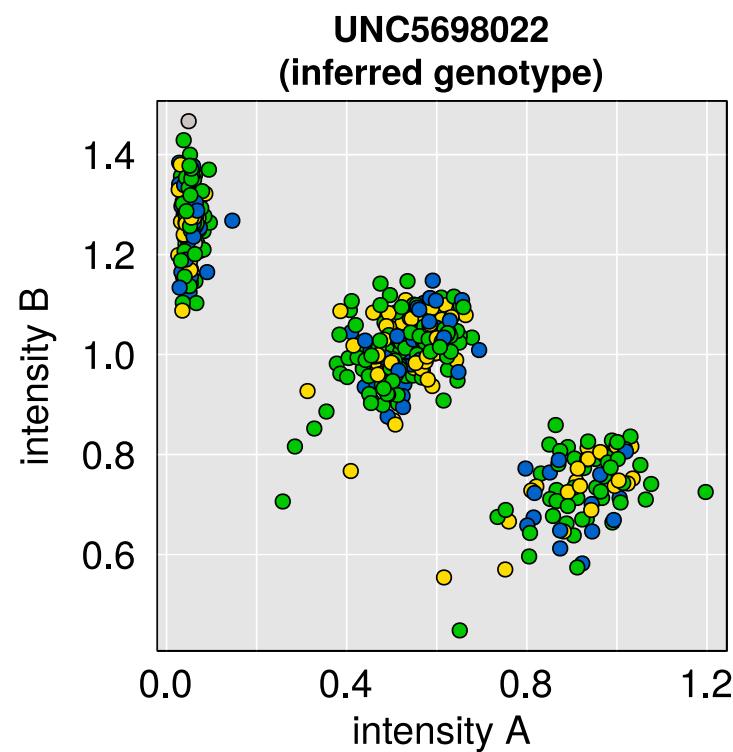
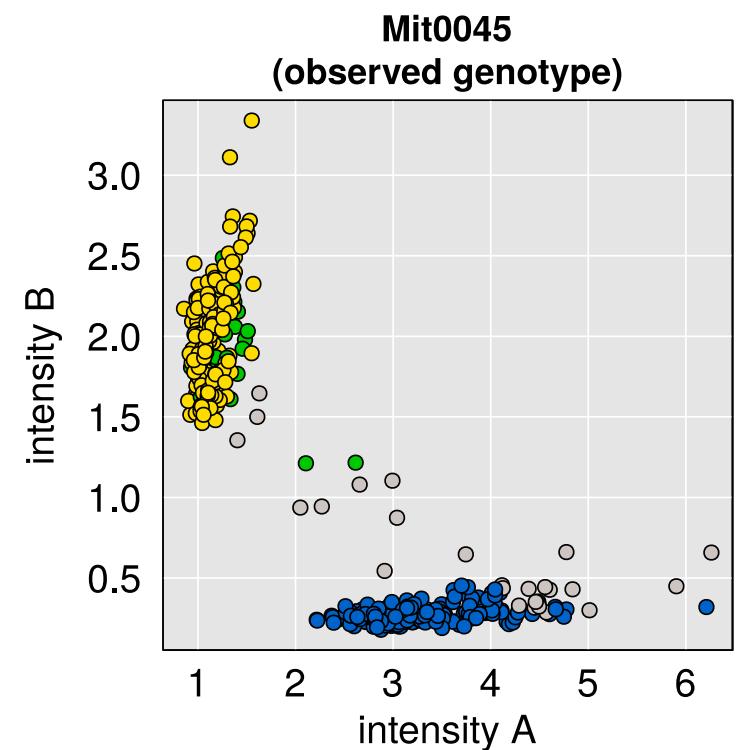
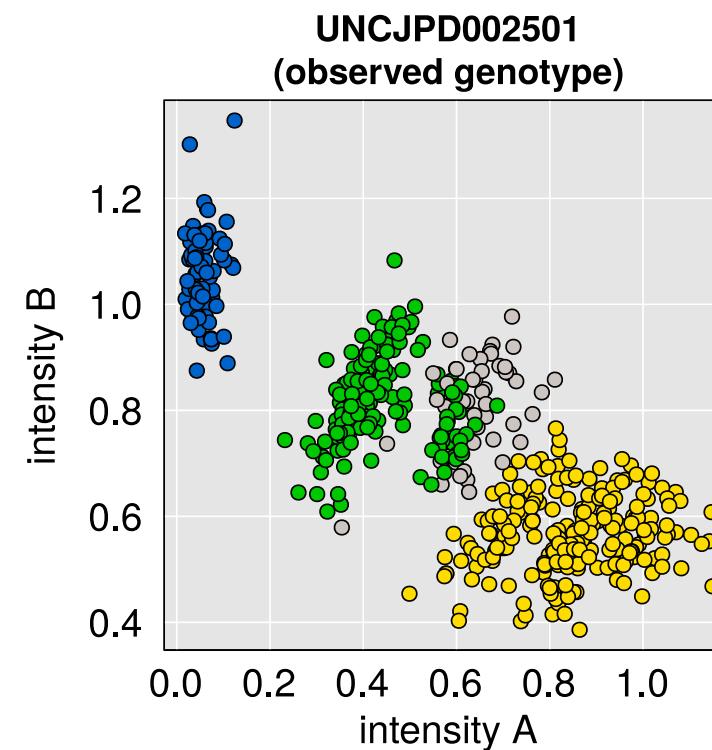
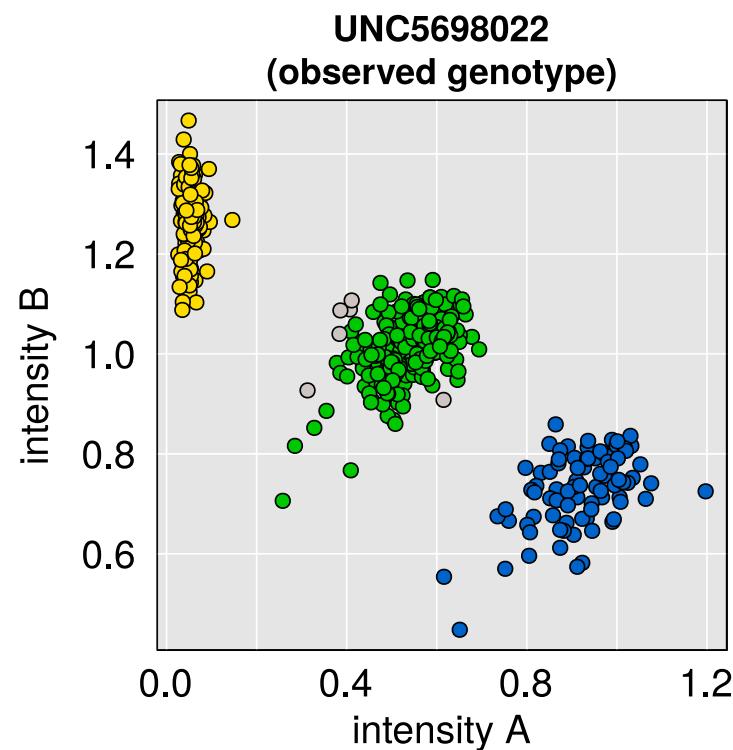
More crap markers



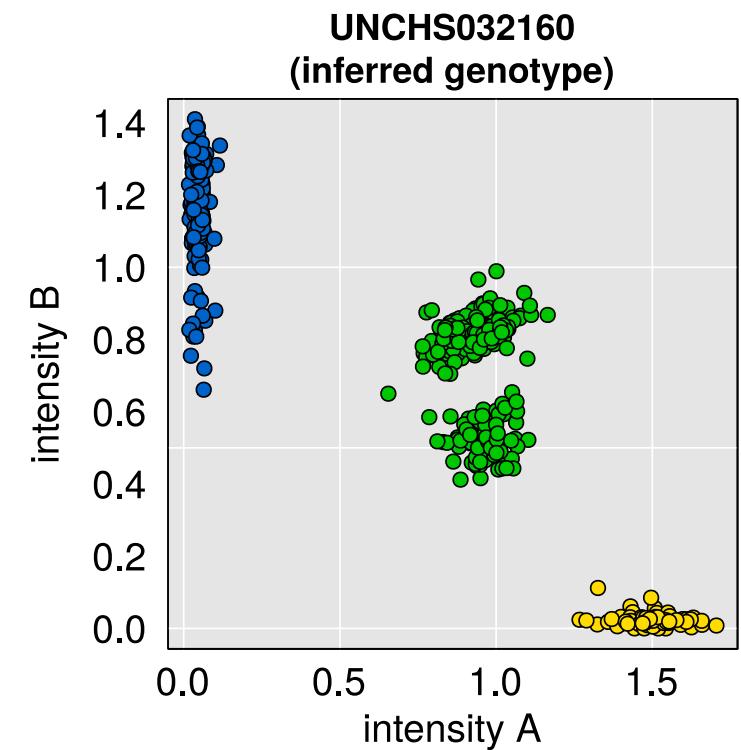
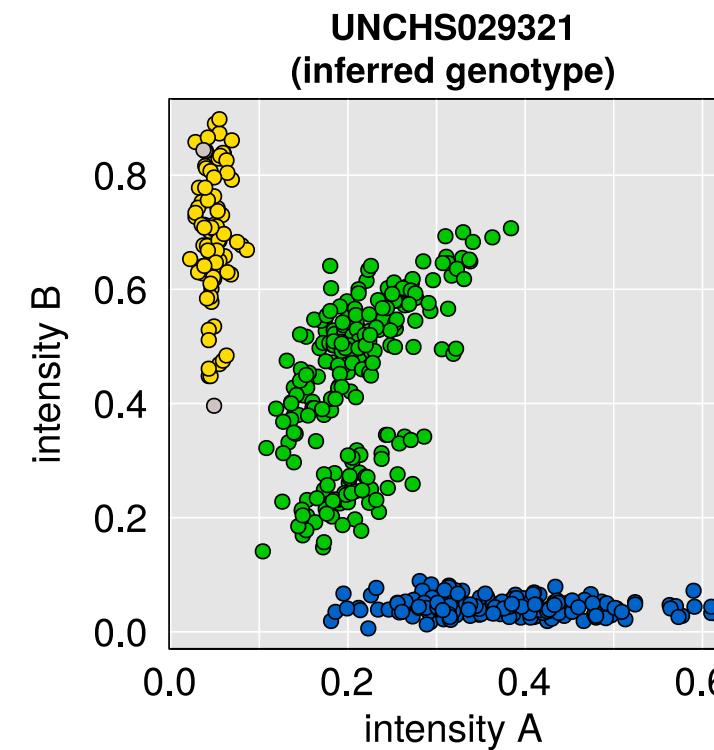
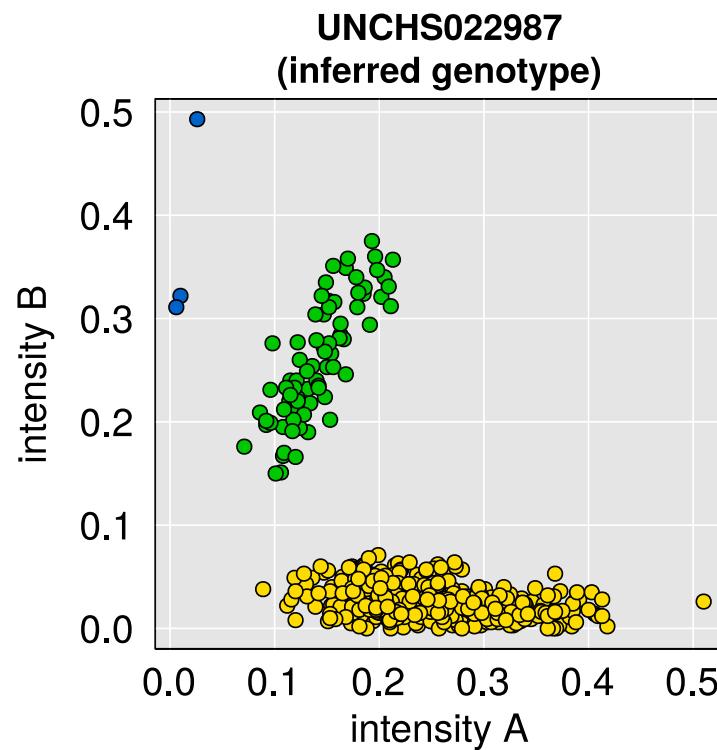
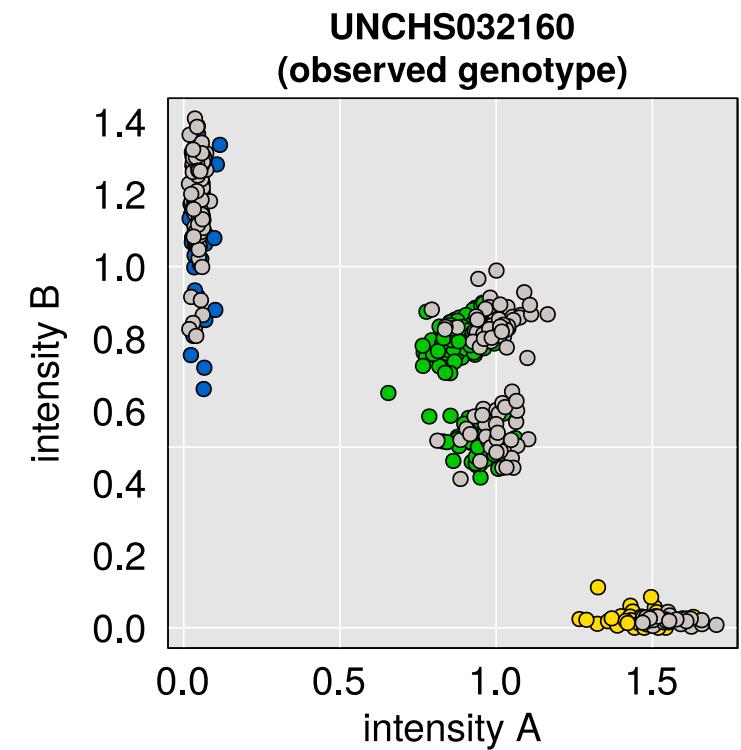
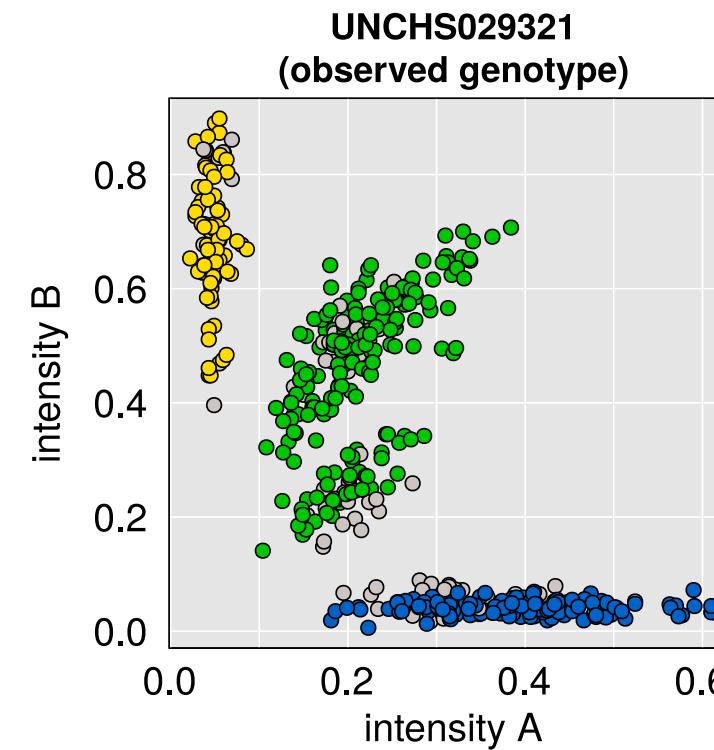
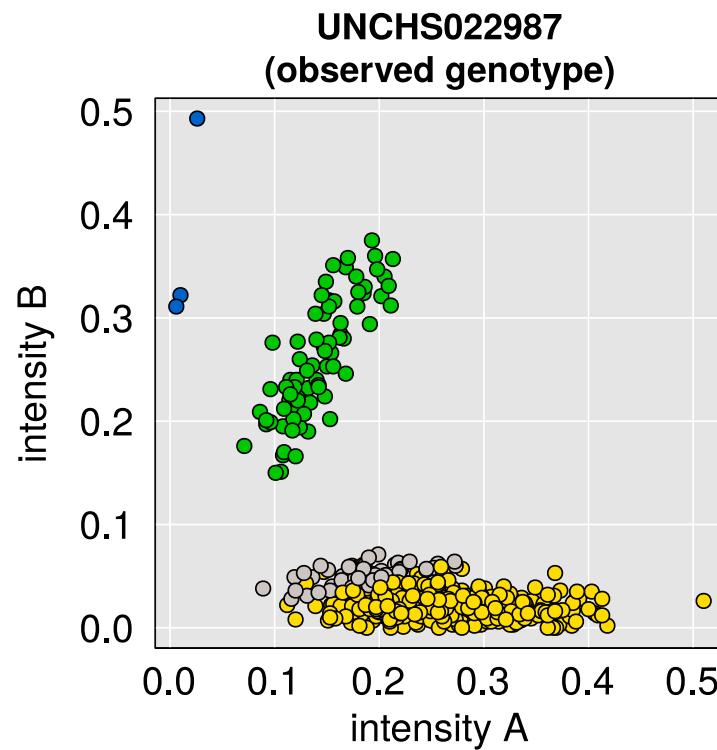
One bad blob



Wrong genomic coordinates

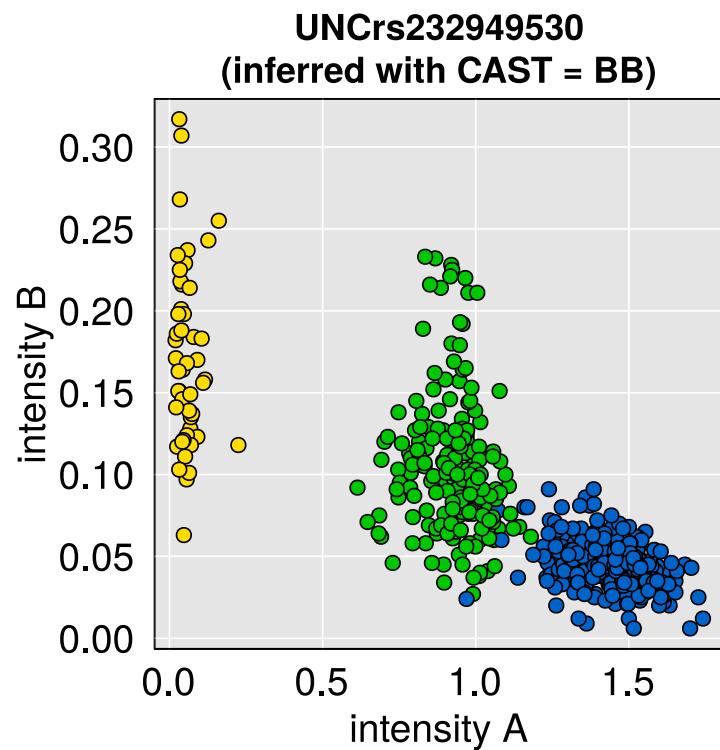
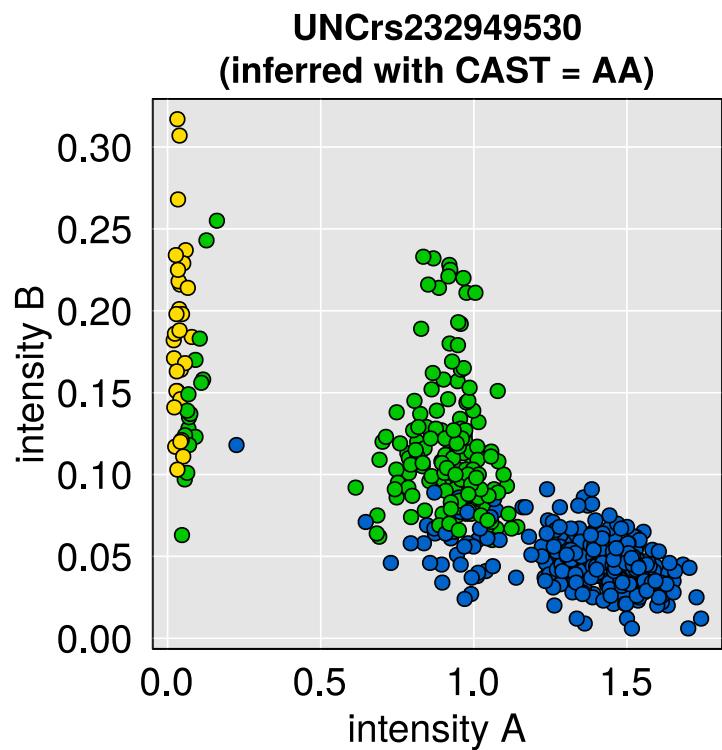
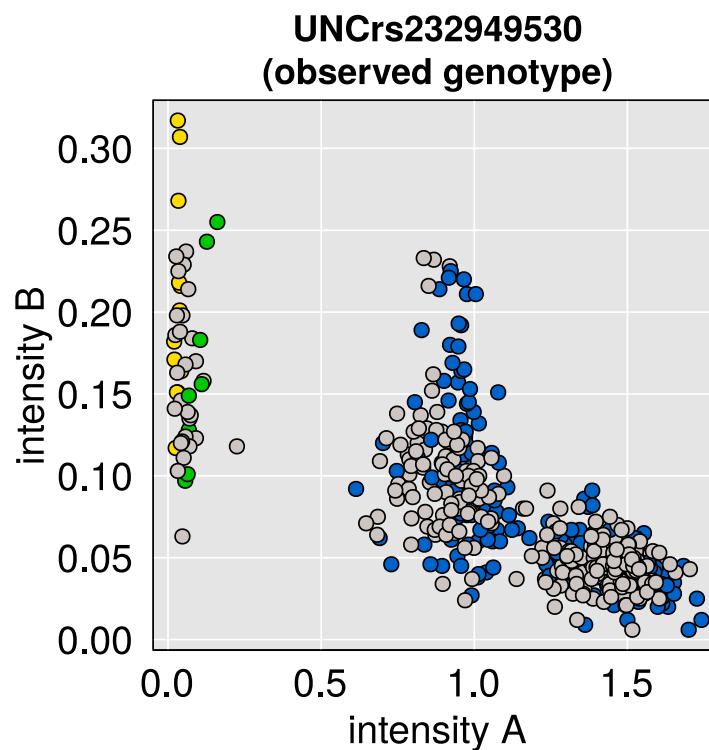


Puzzling no calls

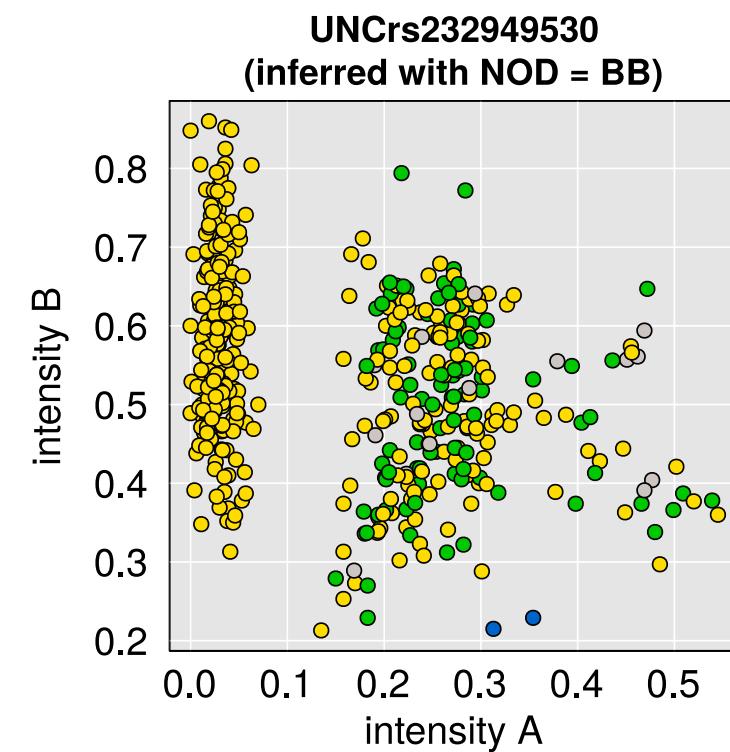
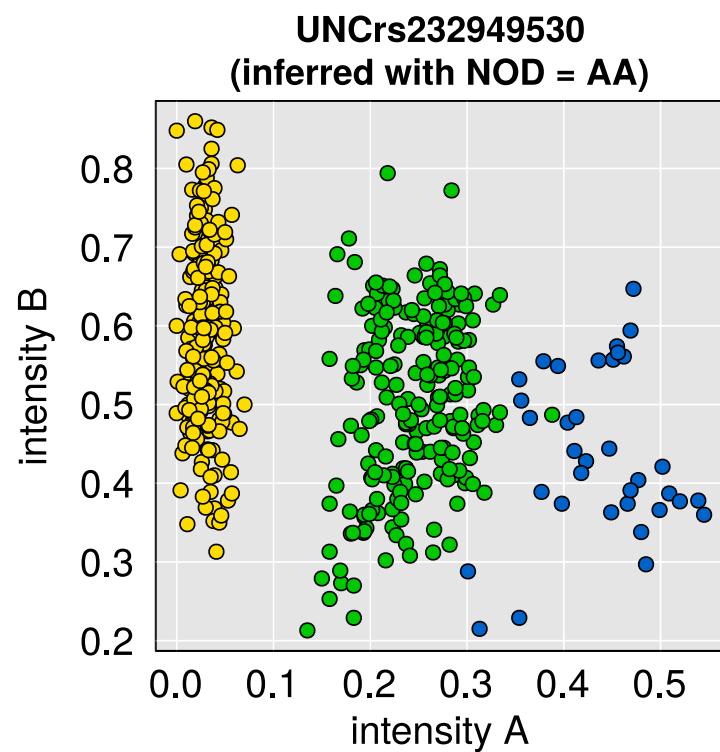
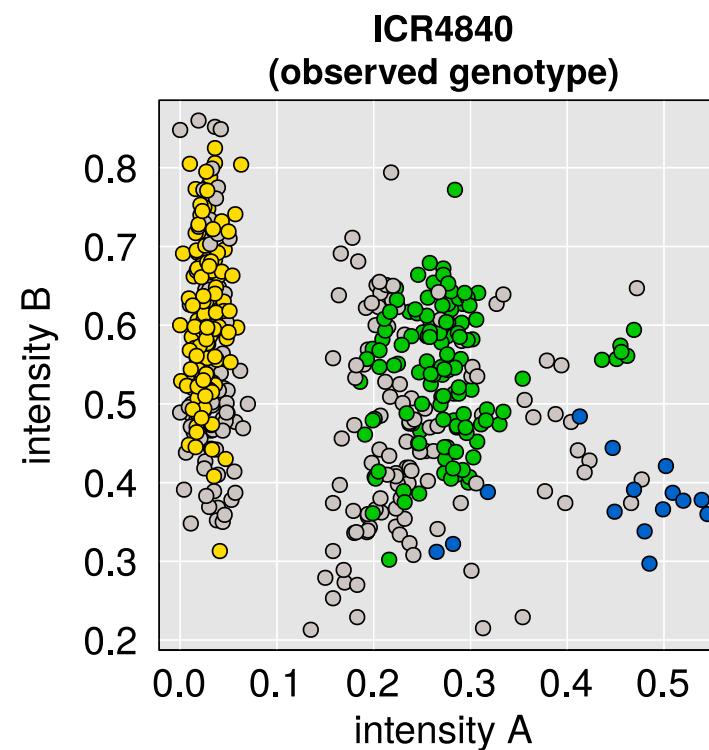


Founder genotyping errors

One founder missing



Another case with one founder missing



Summary

- Quality of results depends on quality of data
- Think about what might have gone wrong, and how it might be revealed
- Pulling out the bad samples is the most important thing
- Sex swaps: look at array intensities
- Look for sample duplicates, and if possible sample mix-ups
- Samples: missing data, array intensities, crossovers, errors
- Markers: lots of reasons for the bad ones

Acknowledgments

Alan Attie
Gary Churchill
Dan Gatti
Alexandra Lobo
Federico Rey
Saunak Sen
Lindsay Traeger
Brian Yandell

NIH/NIGMS, NIH/NIDDK

Slides: bit.ly/2018ctc



kbroman.org

github.com/kbroman

@kwbroman