Data diagnostics

Cleaning genotype data in multi-parent populations

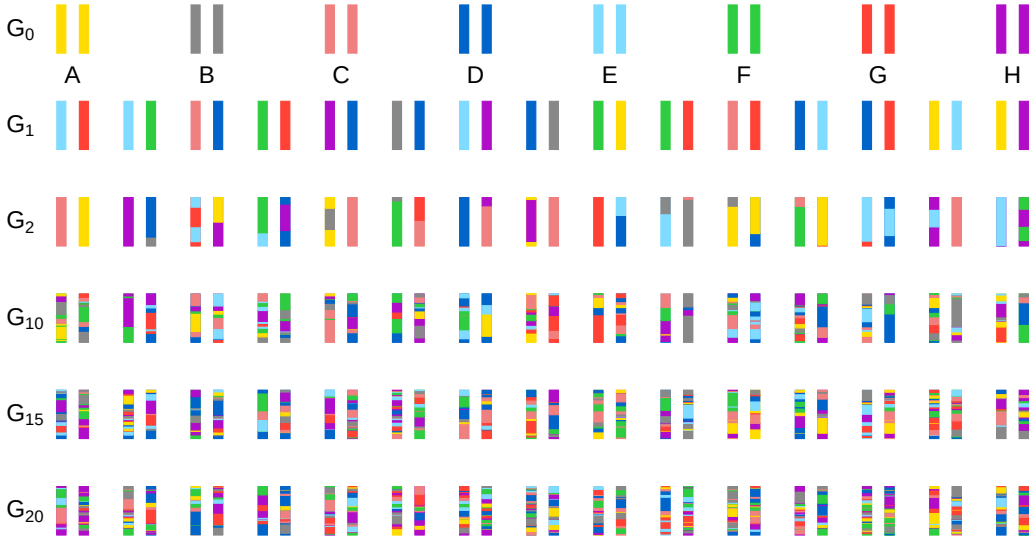
Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman

Course web: kbroman.org/AdvData

Diversity outbred mice



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

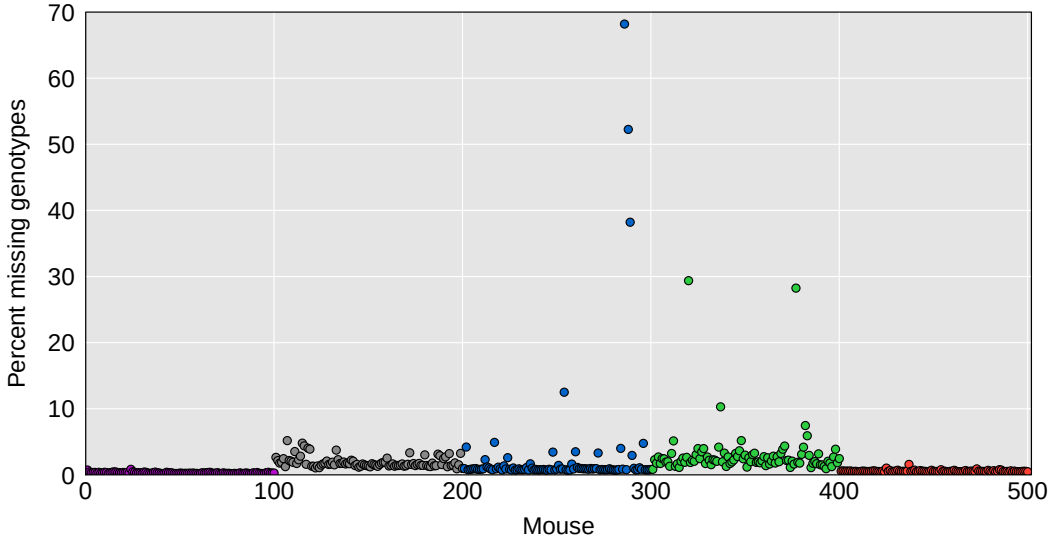
- What might have gone wrong?
- How could it be revealed?
- Also, just make a bunch of graphs.
- If you see something weird, try to figure it out.

Possible problems

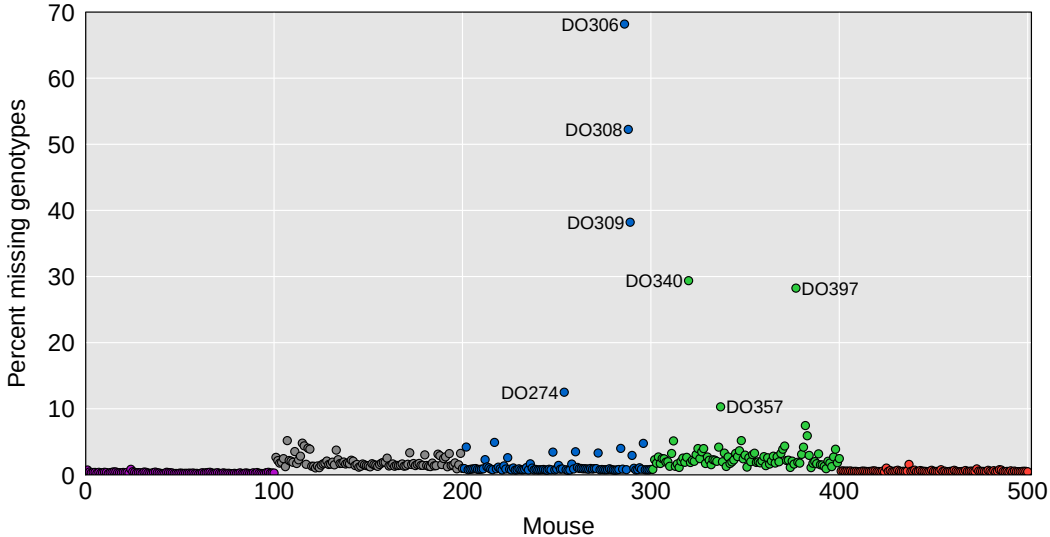
- Sample duplicates
- Sample mix-ups
- Bad samples
- Bad markers
- Genotyping errors in founders

What to look at first?

Missing data per sample

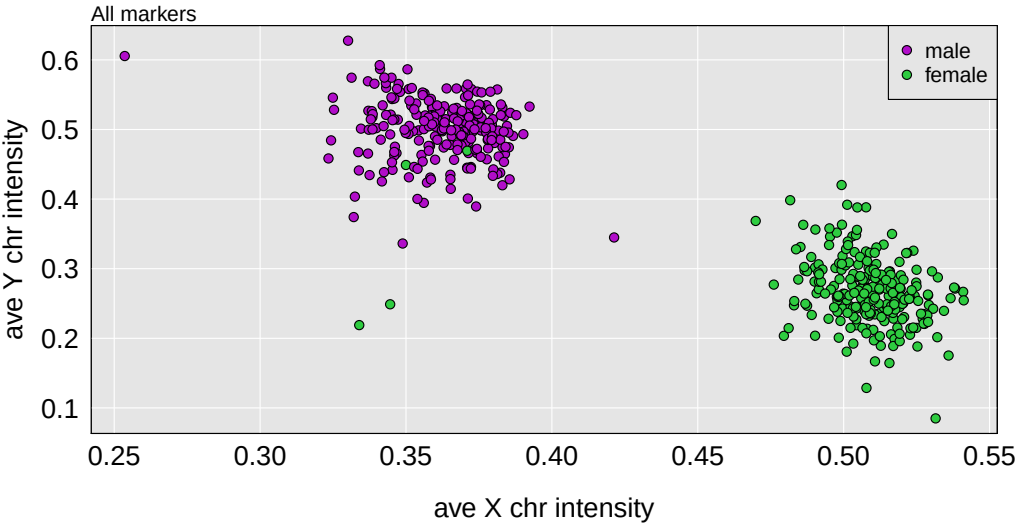


Missing data per sample

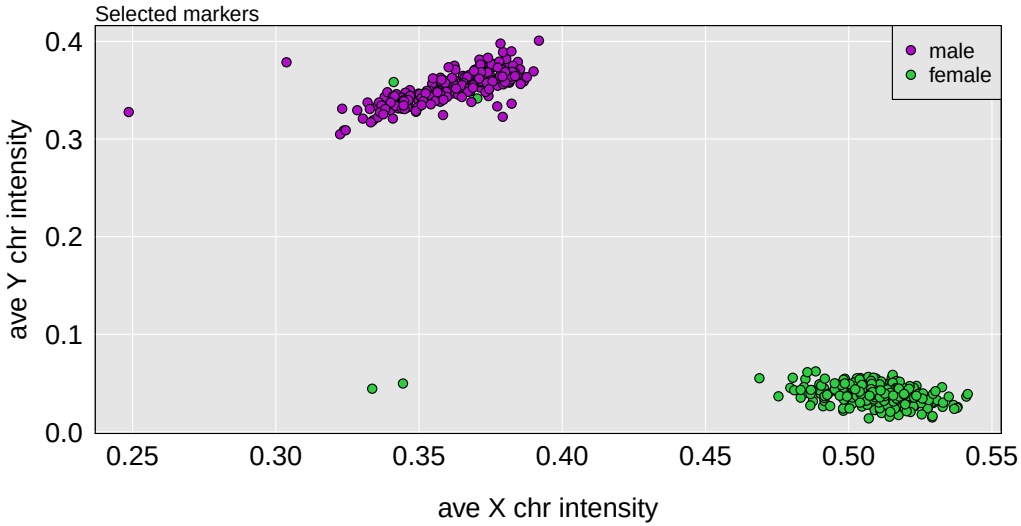


Swapped sex labels

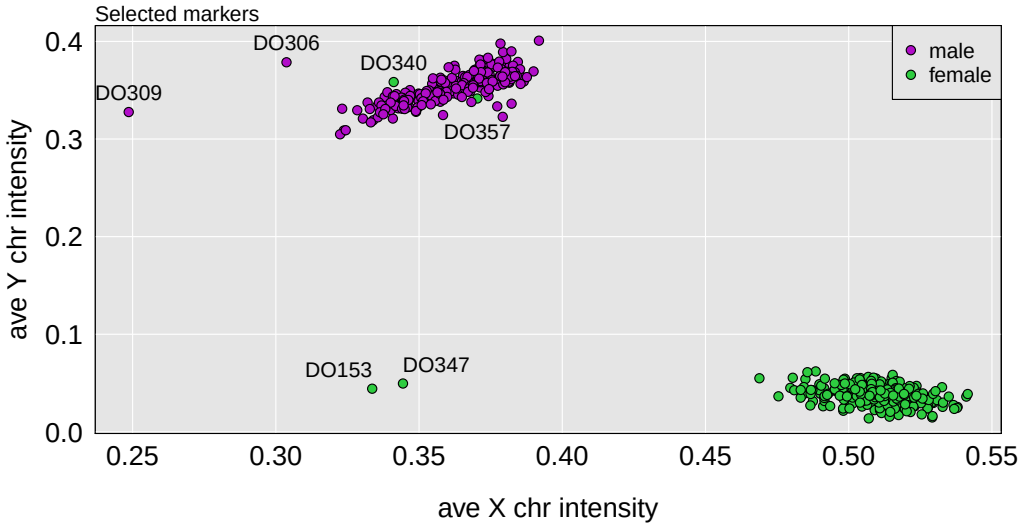
Ave SNP intensity on X and Y chr



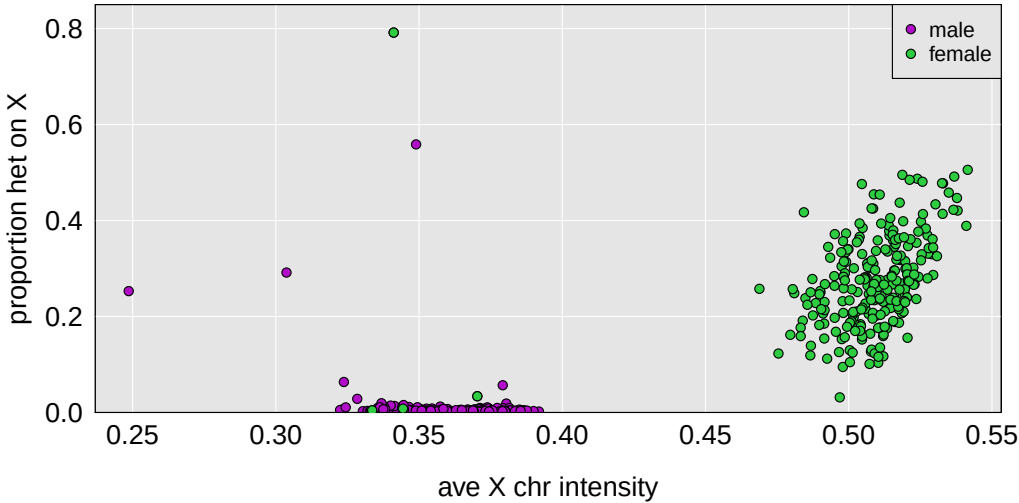
Ave SNP intensity on X and Y chr



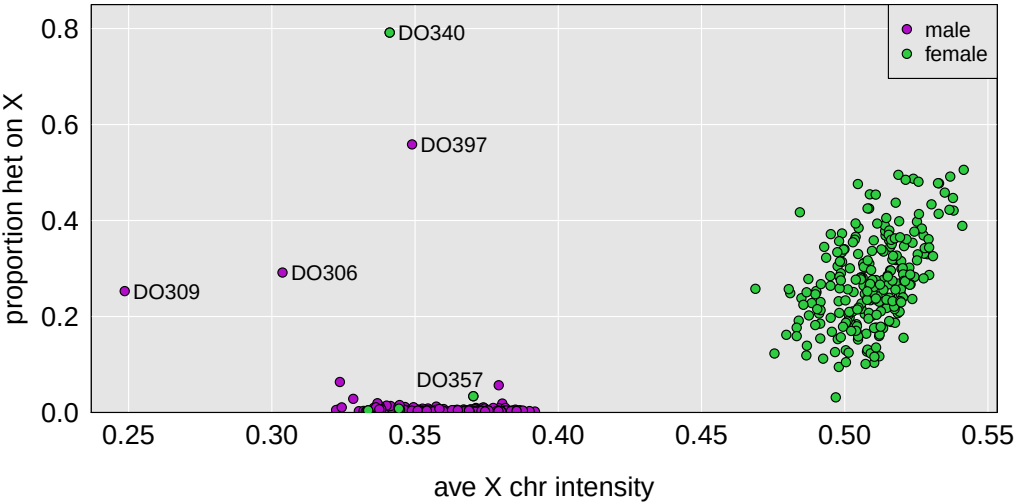
Ave SNP intensity on X and Y chr



Heterozygosity vs SNP intensity on X chr

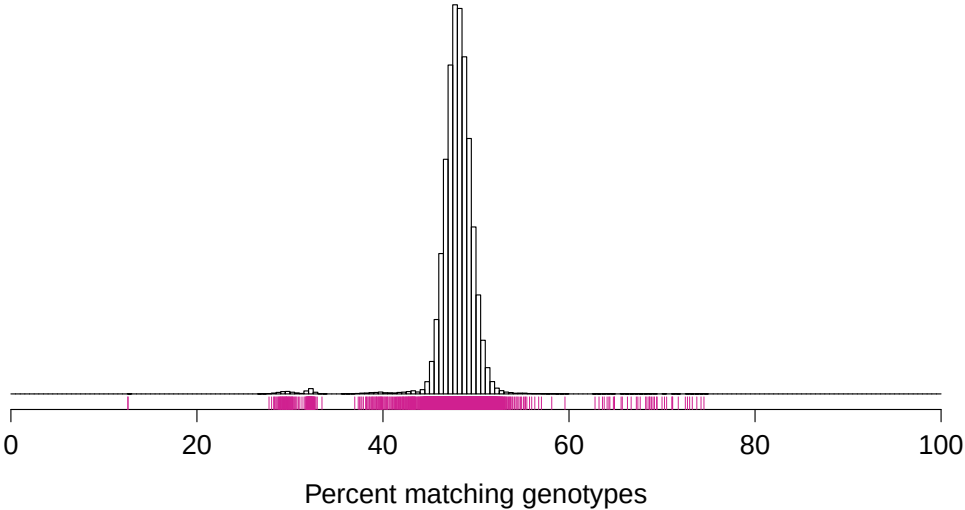


Heterozygosity vs SNP intensity on X chr

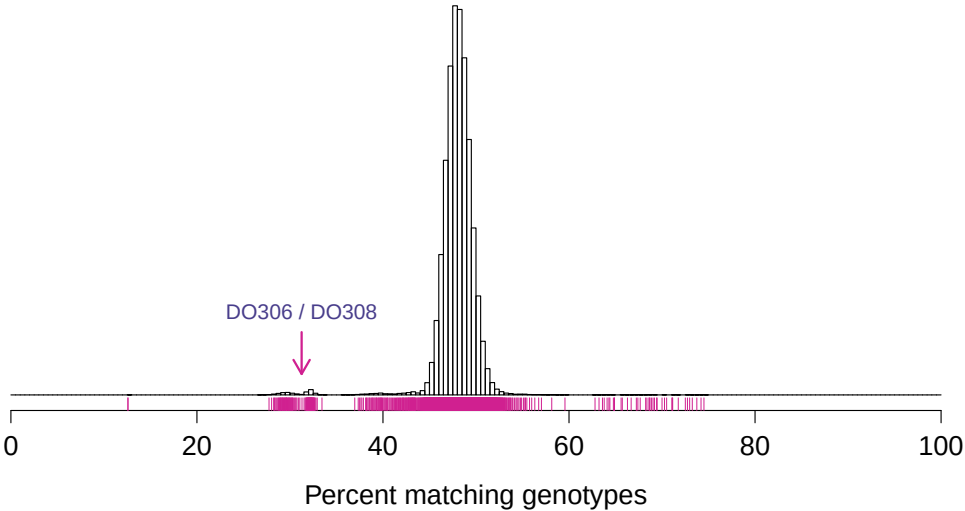


Sample duplicates

Percent matching genotypes between pairs

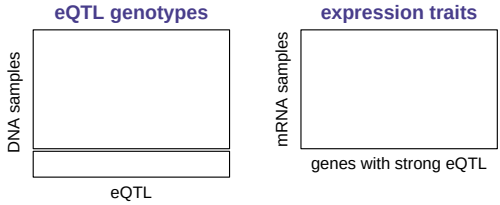


Percent matching genotypes between pairs

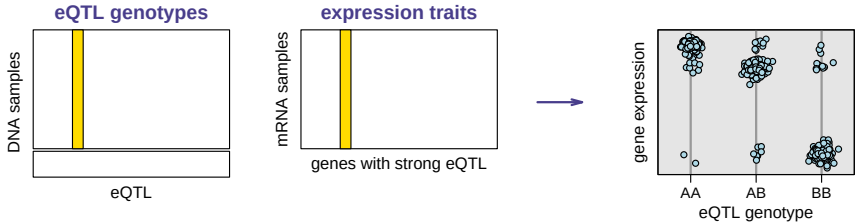


Sample mix-ups: RNA-seq data

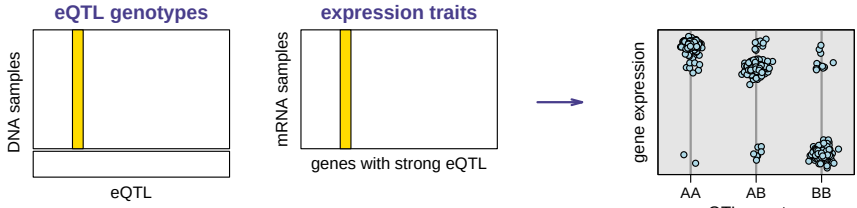
RNA-seq mix-ups

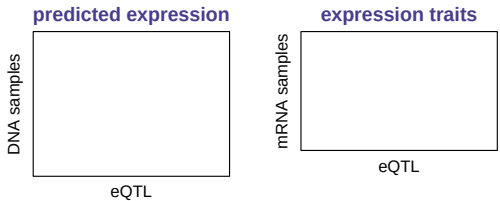


RNA-seq mix-ups

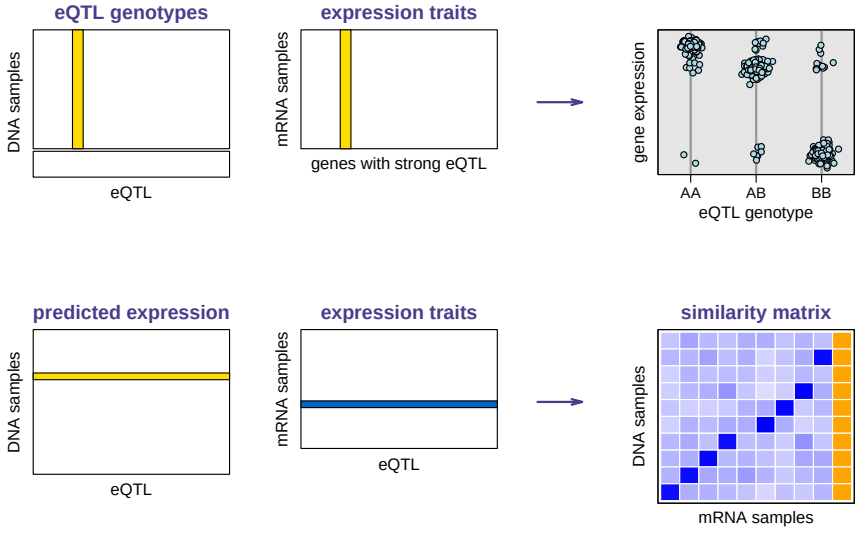


RNA-seq mix-ups

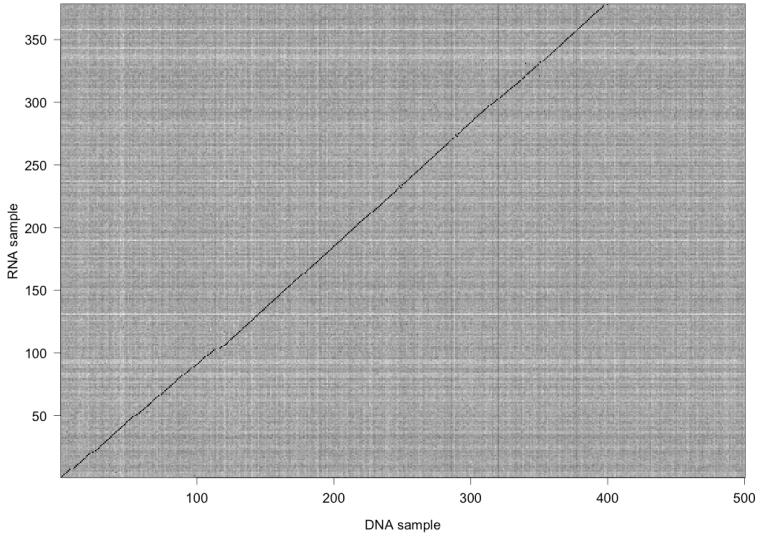




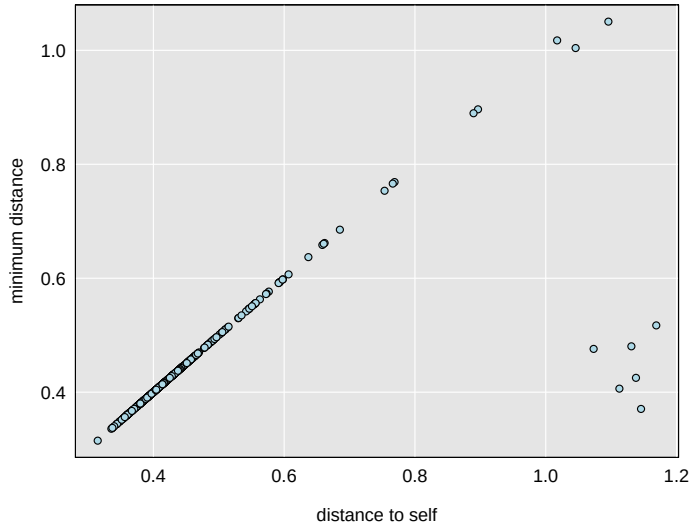
RNA-seq mix-ups



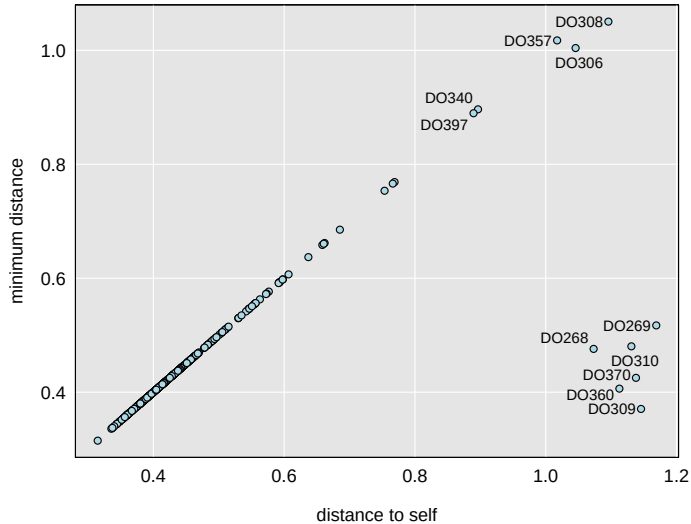
Distance matrix



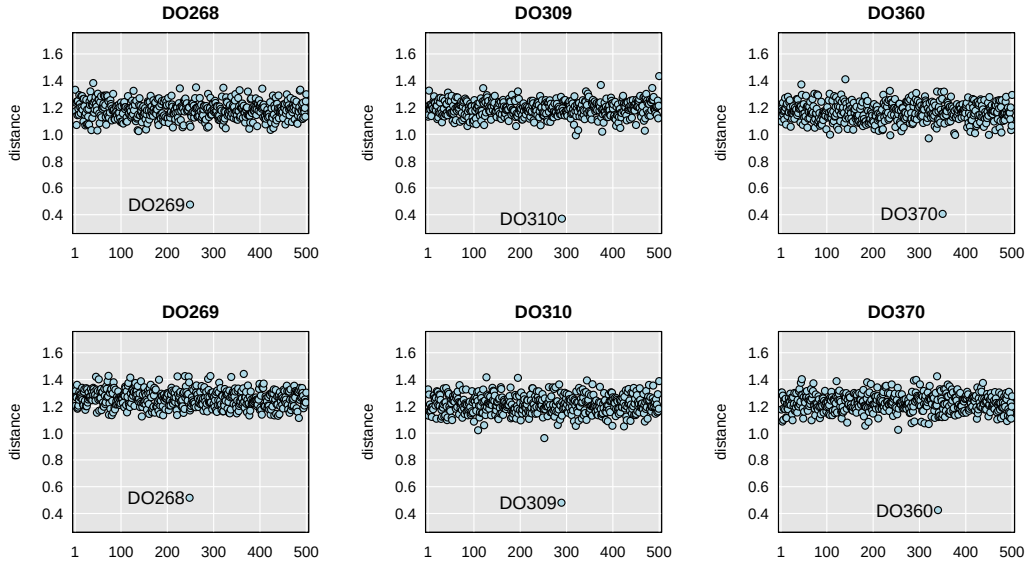
Min vs self distance



Min vs self distance

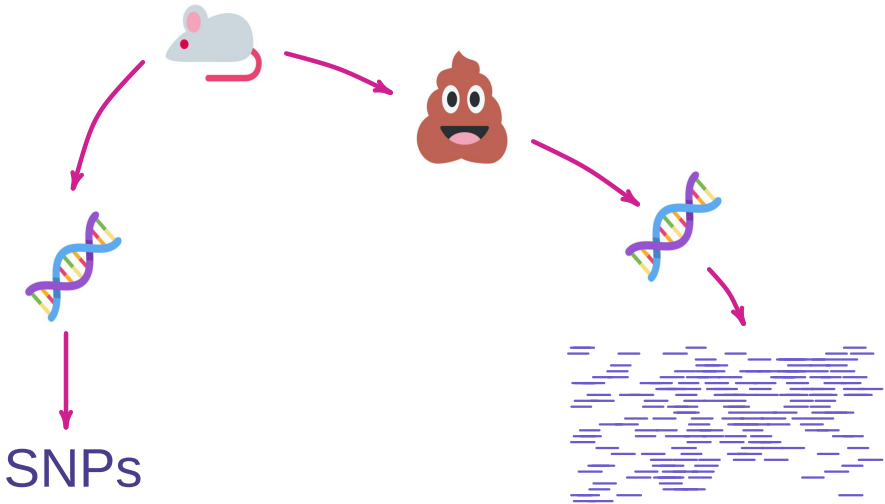


RNA-seq mix-ups, details



Sample mix-ups: microbiome data

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

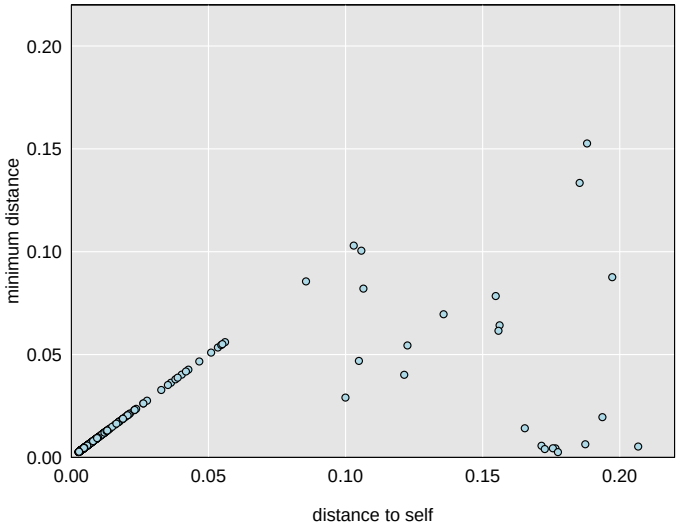
Microbiome DO360 vs DNA DO360

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

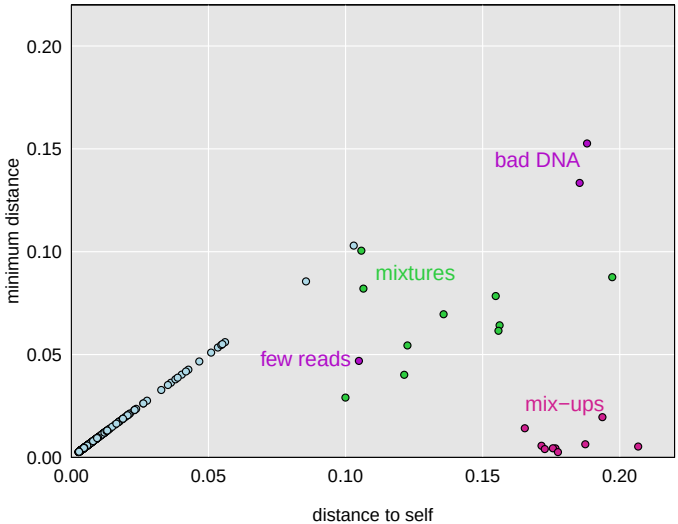
Microbiome DO360 vs DNA DO370

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

Microbiome mix-ups: min vs self distance

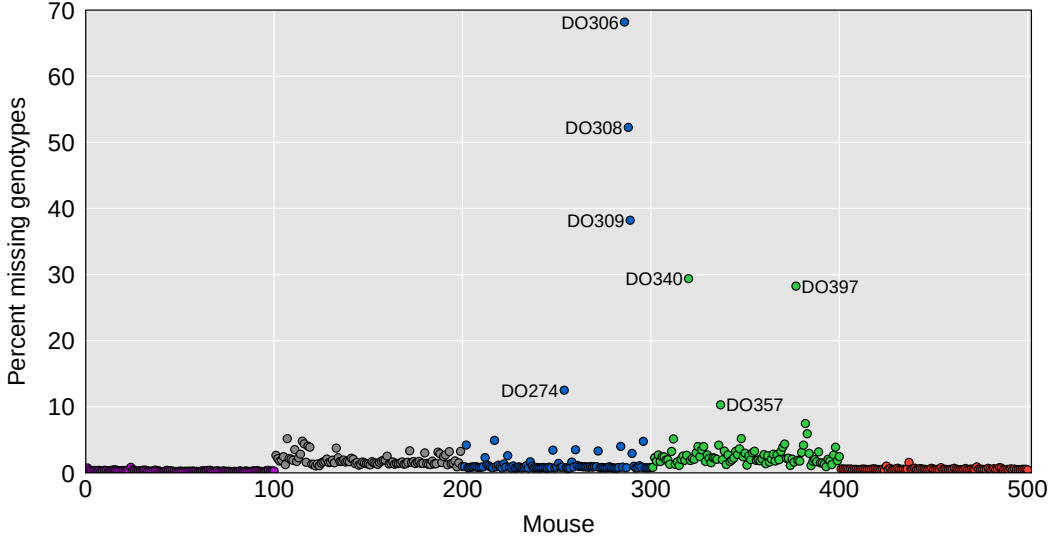


Microbiome mix-ups: min vs self distance

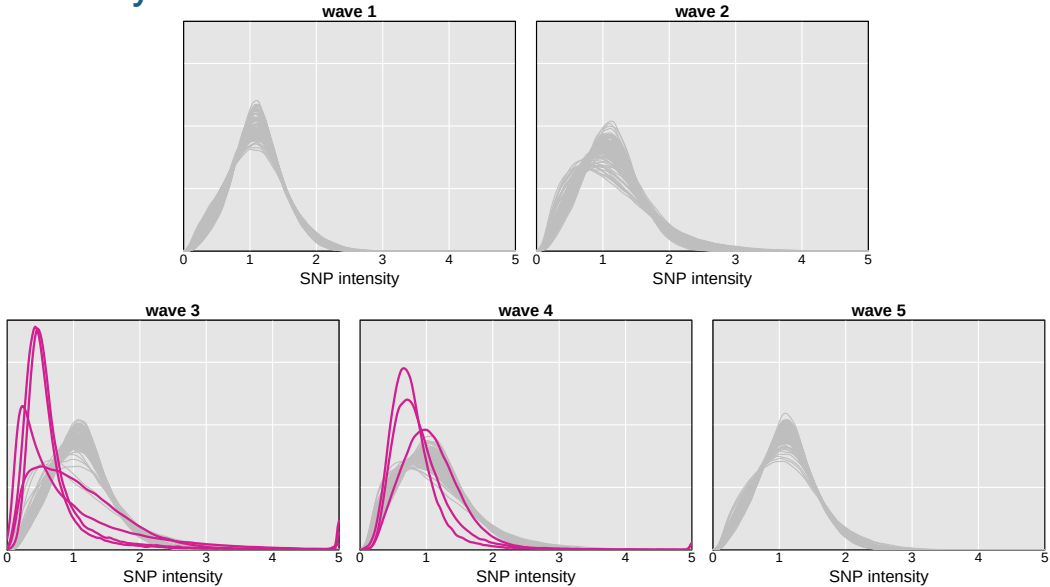


Sample quality

Missing data per sample

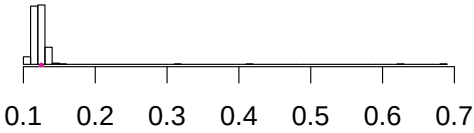


SNP array intensities

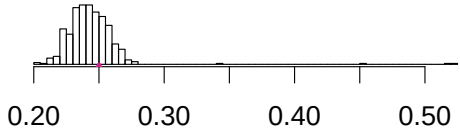


Allele frequencies by individual

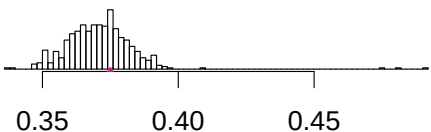
founder MAF = 1/8



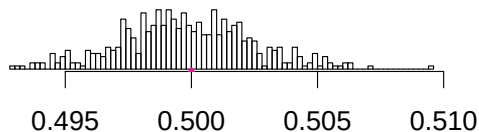
founder MAF = 2/8



founder MAF = 3/8

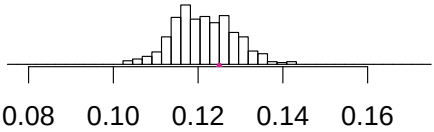


founder MAF = 4/8

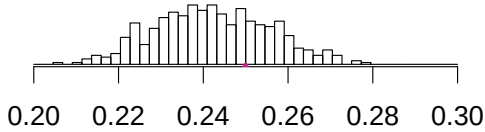


Allele frequencies by individual

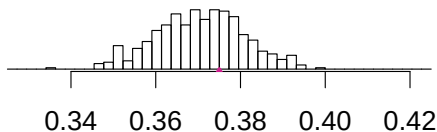
founder MAF = 1/8



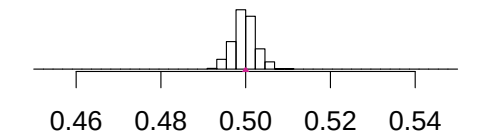
founder MAF = 2/8



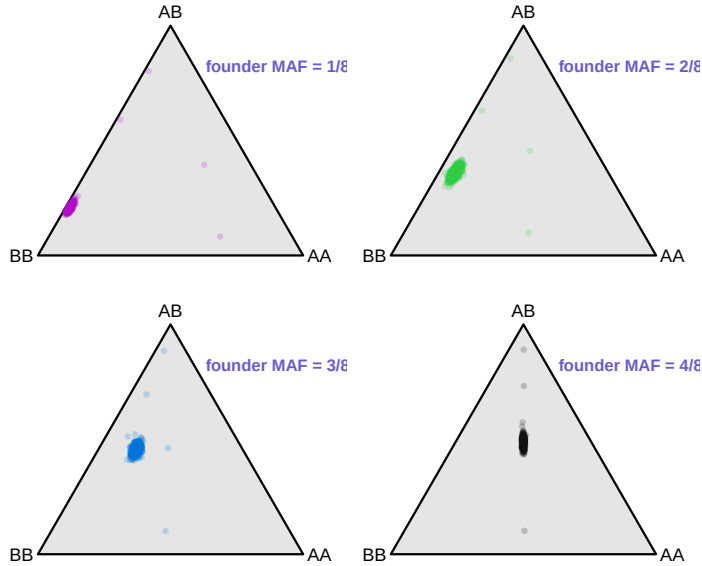
founder MAF = 3/8



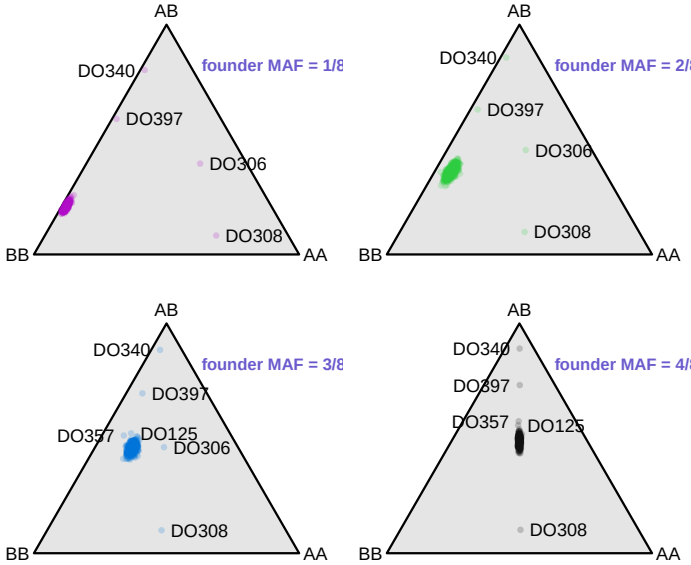
founder MAF = 4/8



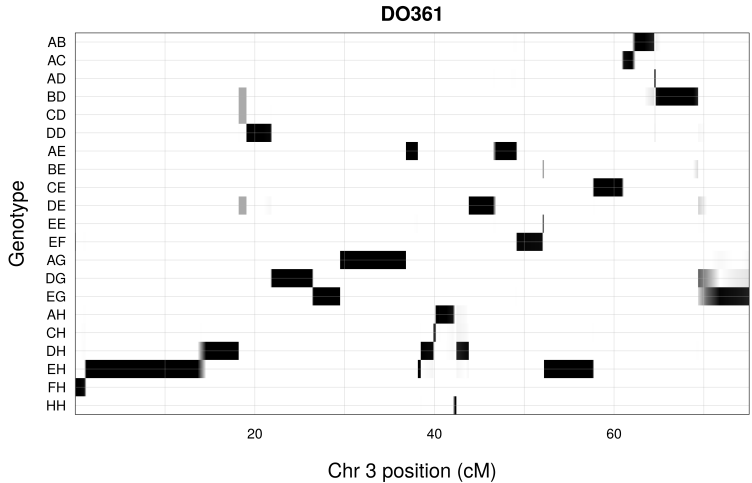
Genotype frequencies by individual



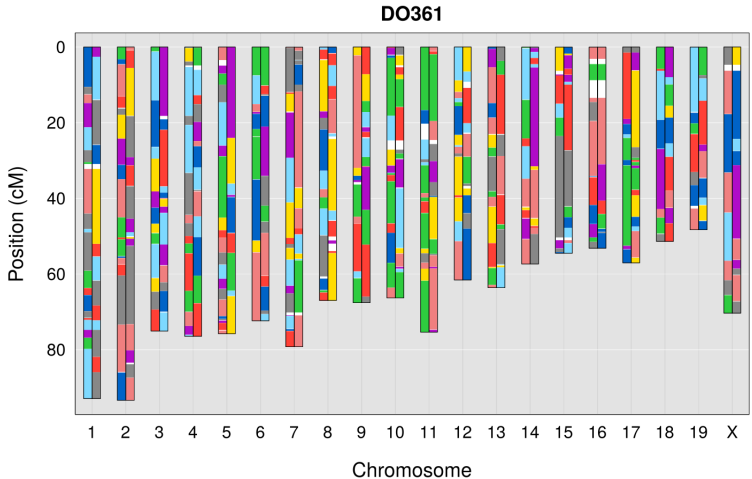
Genotype frequencies by individual



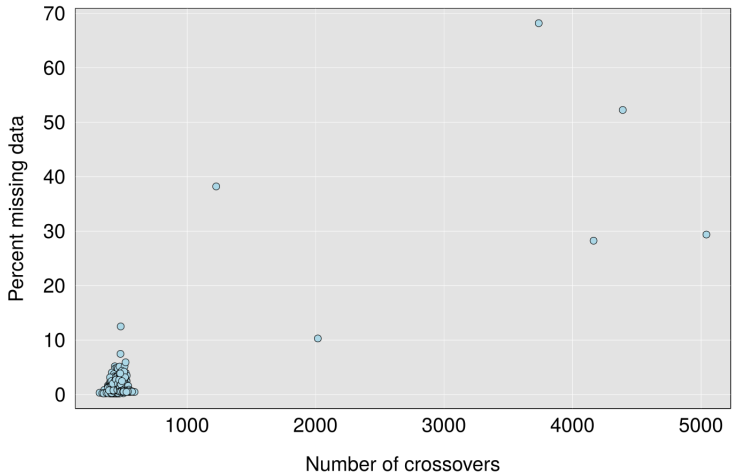
Genotype probabilities (one mouse on one chr)



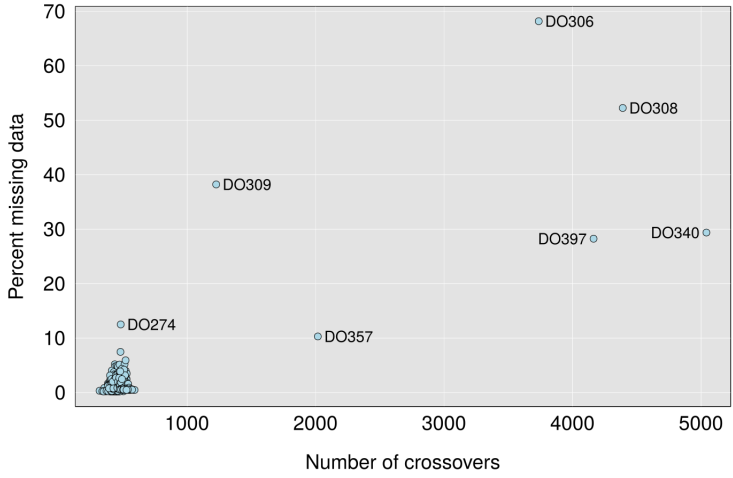
Genotype reconstruction (one mouse)



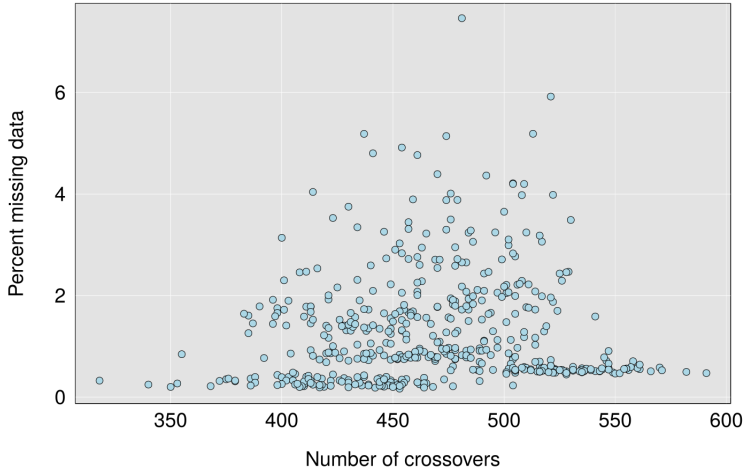
Percent missing vs. number of crossovers



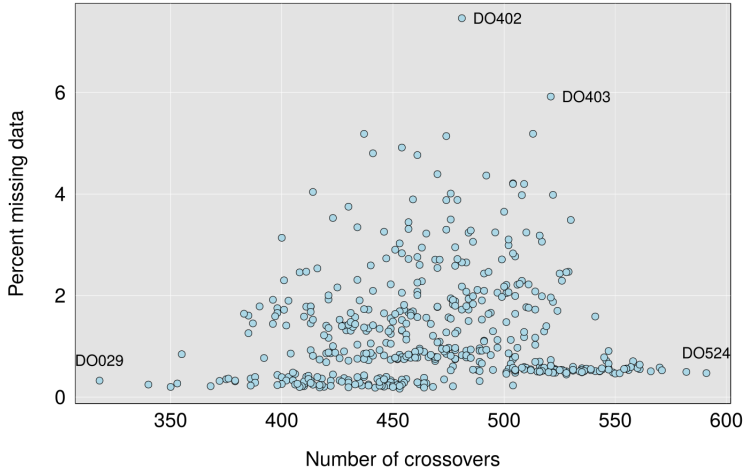
Percent missing vs. number of crossovers



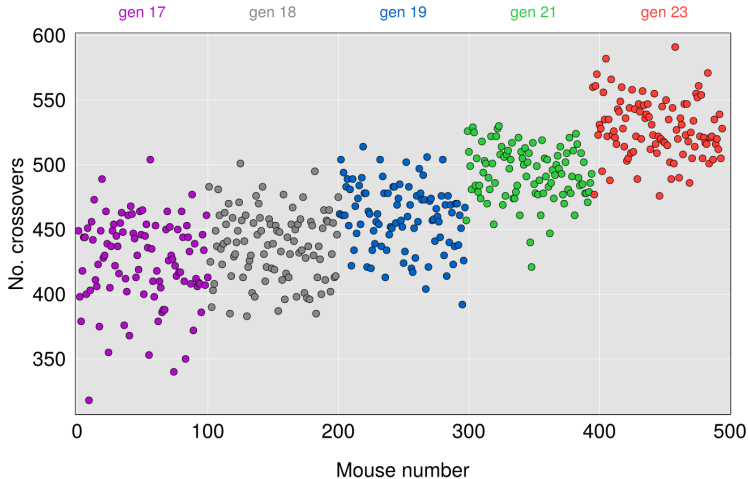
Percent missing vs. number of crossovers



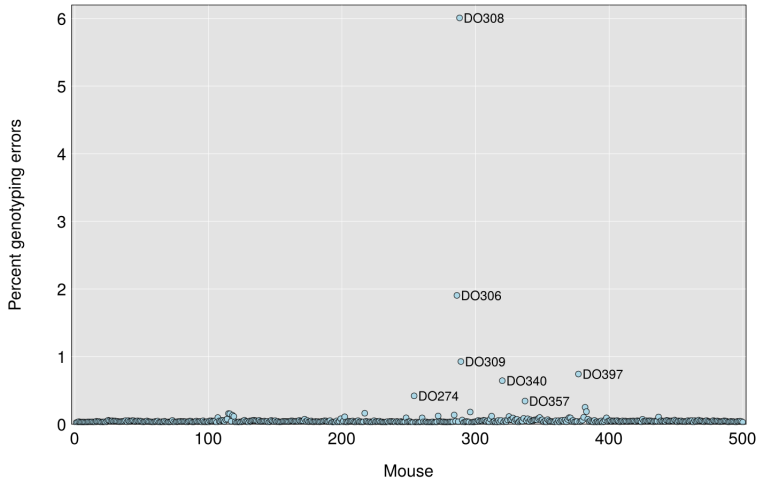
Percent missing vs. number of crossovers



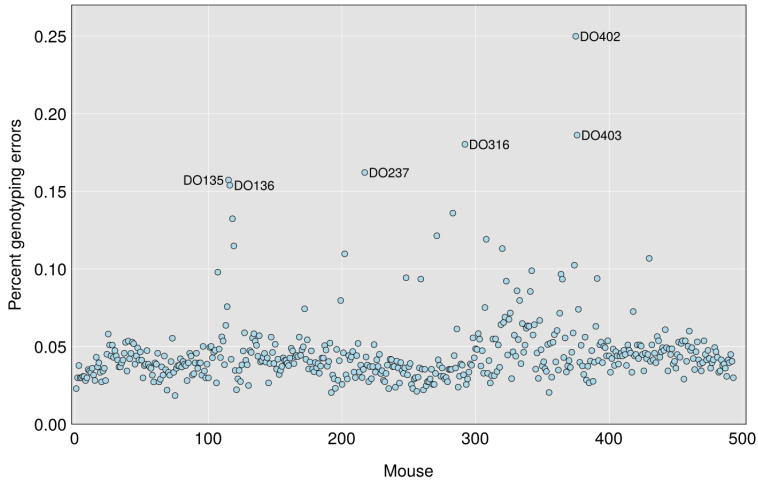
Crossovers by generation



Percent genotyping errors

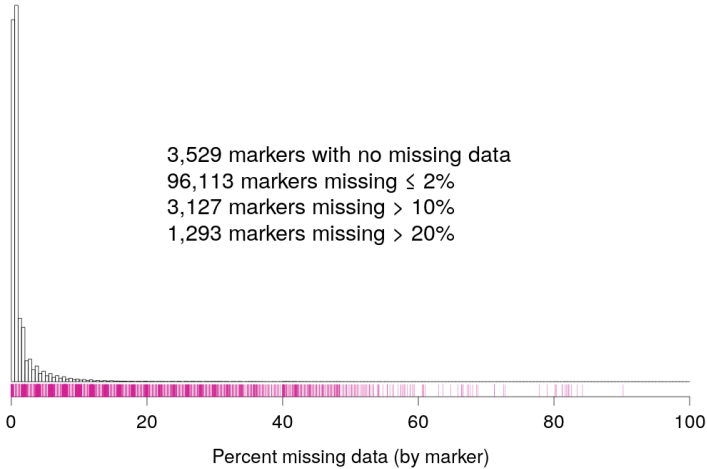


Percent genotyping errors

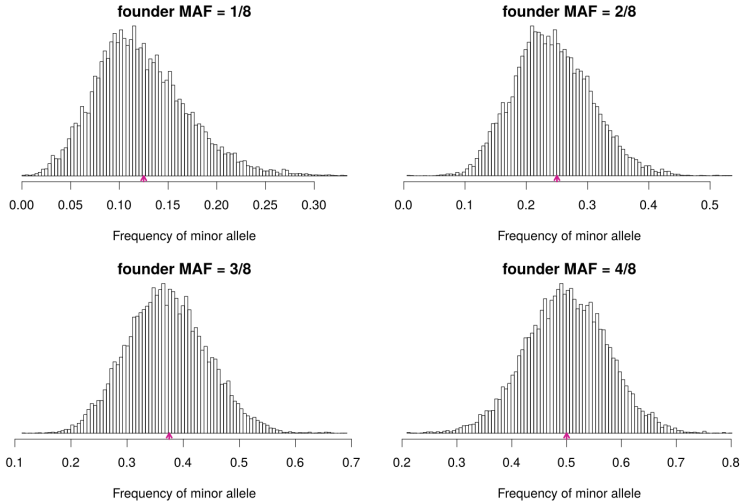


Marker quality

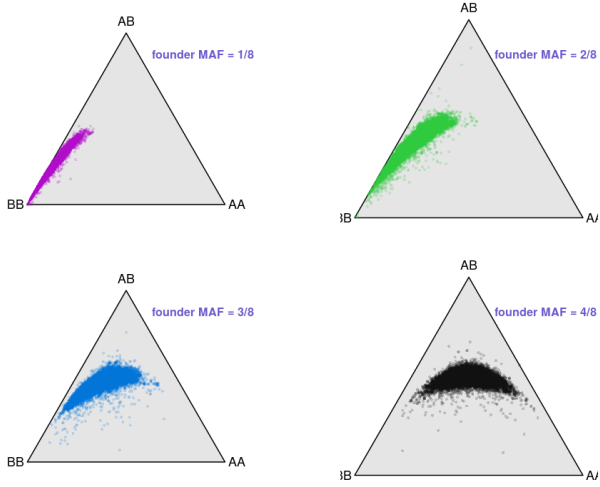
Proportion missing data



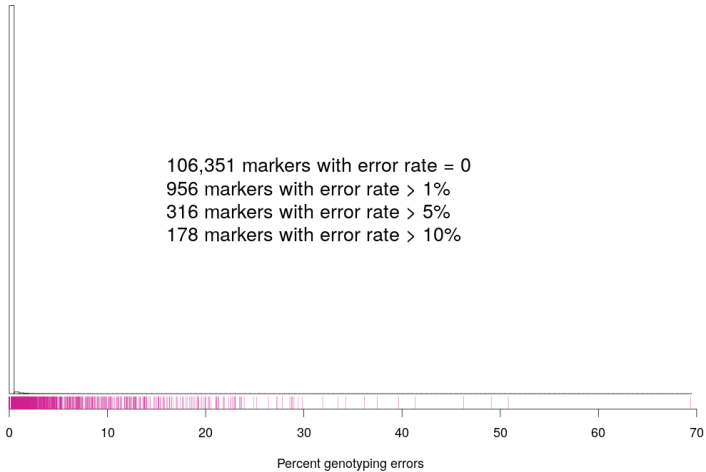
Allele frequencies by marker



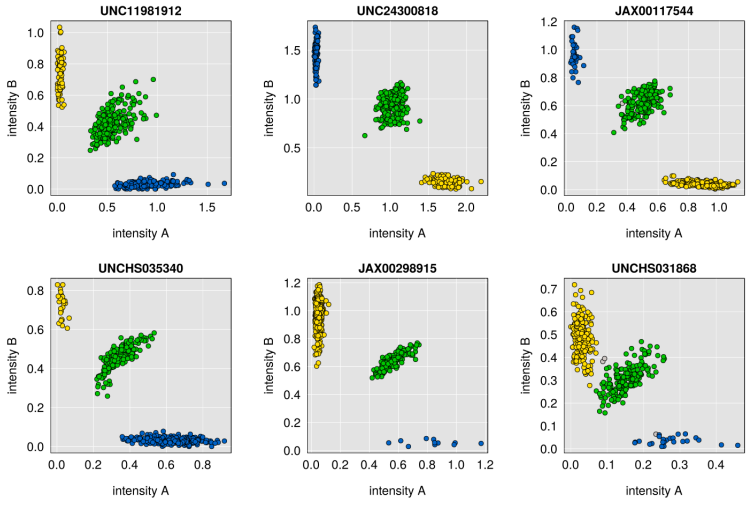
Genotype frequencies by marker



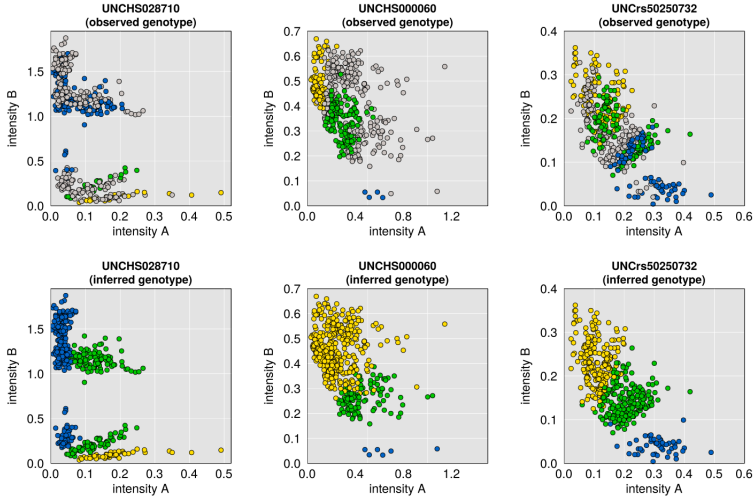
Genotyping error rates



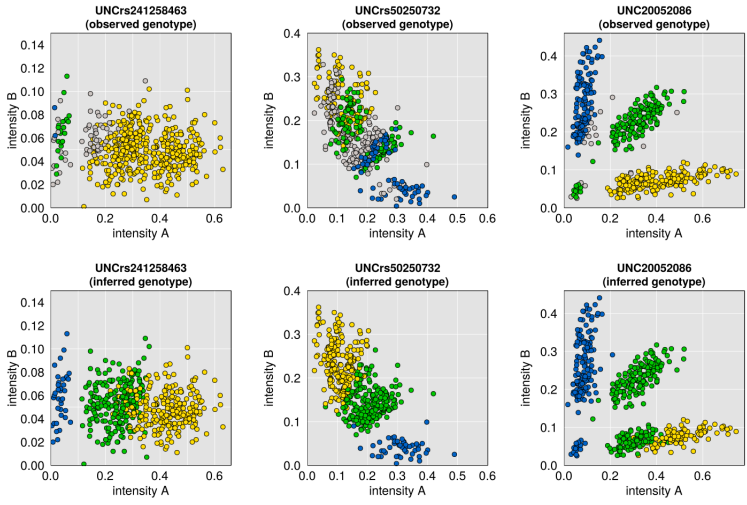
Nice markers



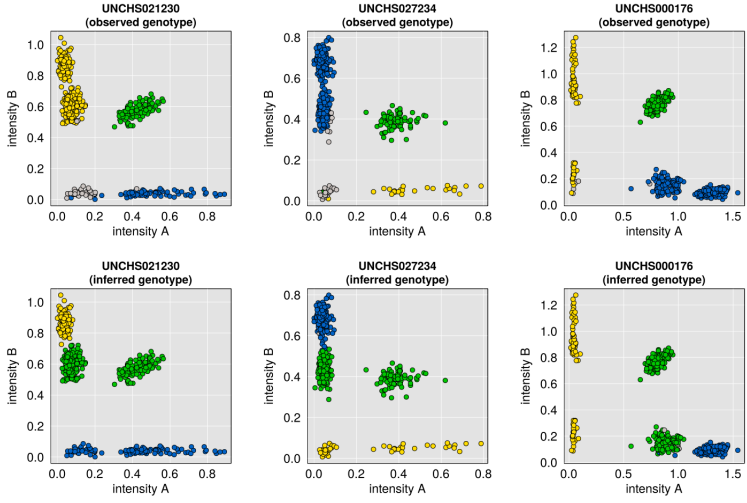
Crap markers



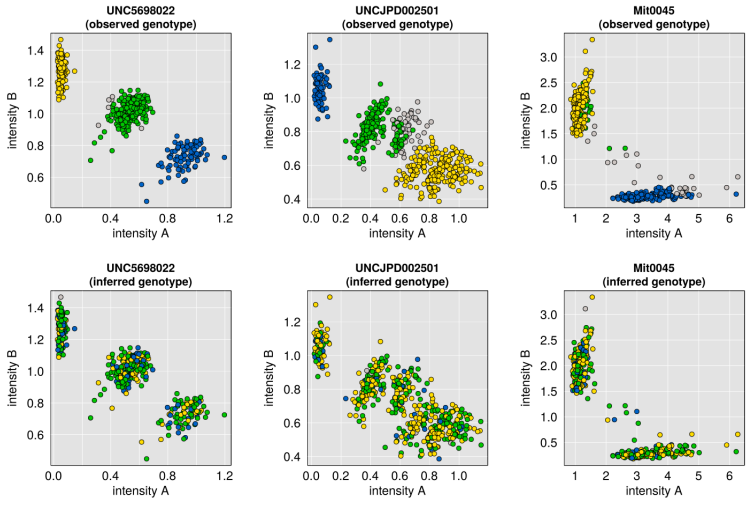
More crap markers



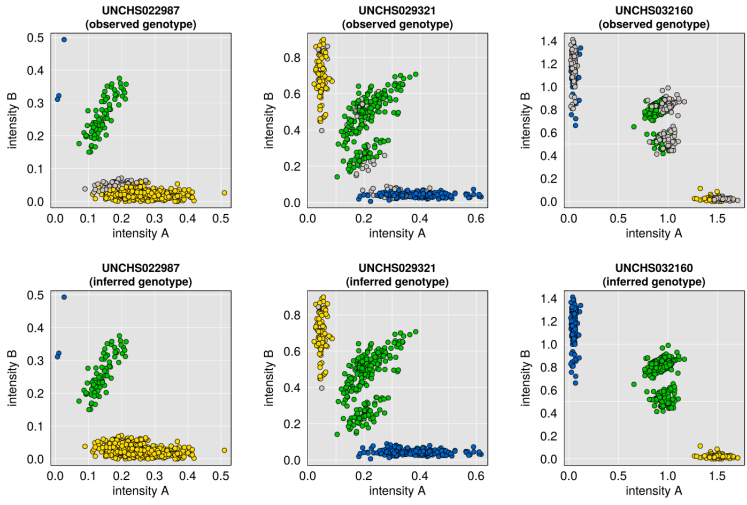
One bad blob


UNCHS021230
(observed genotype)
UNCHS027234
(observed genotype)
UNCHS000176
(observed genotype)
UNCHS021230
(inferred genotype)
UNCHS027234
(inferred genotype)
UNCHS000176
(inferred genotype)
intensity A
intensity B

Wrong genomic position

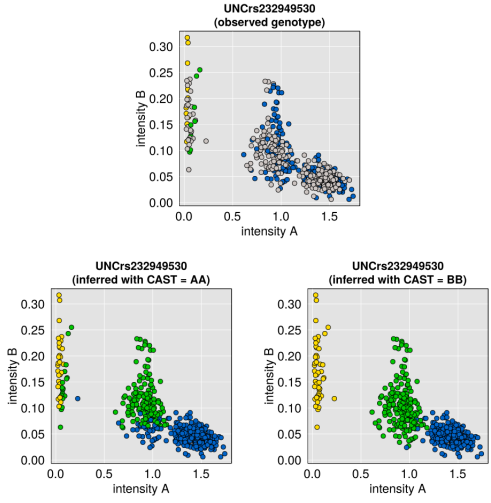


Puzzling no calls

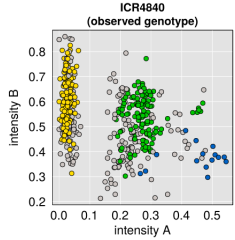


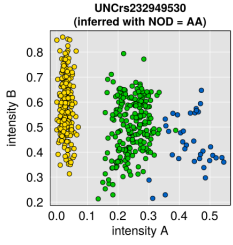
Founder genotyping errors

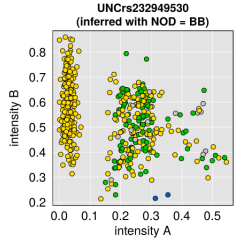
One founder missing



Another case







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

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

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