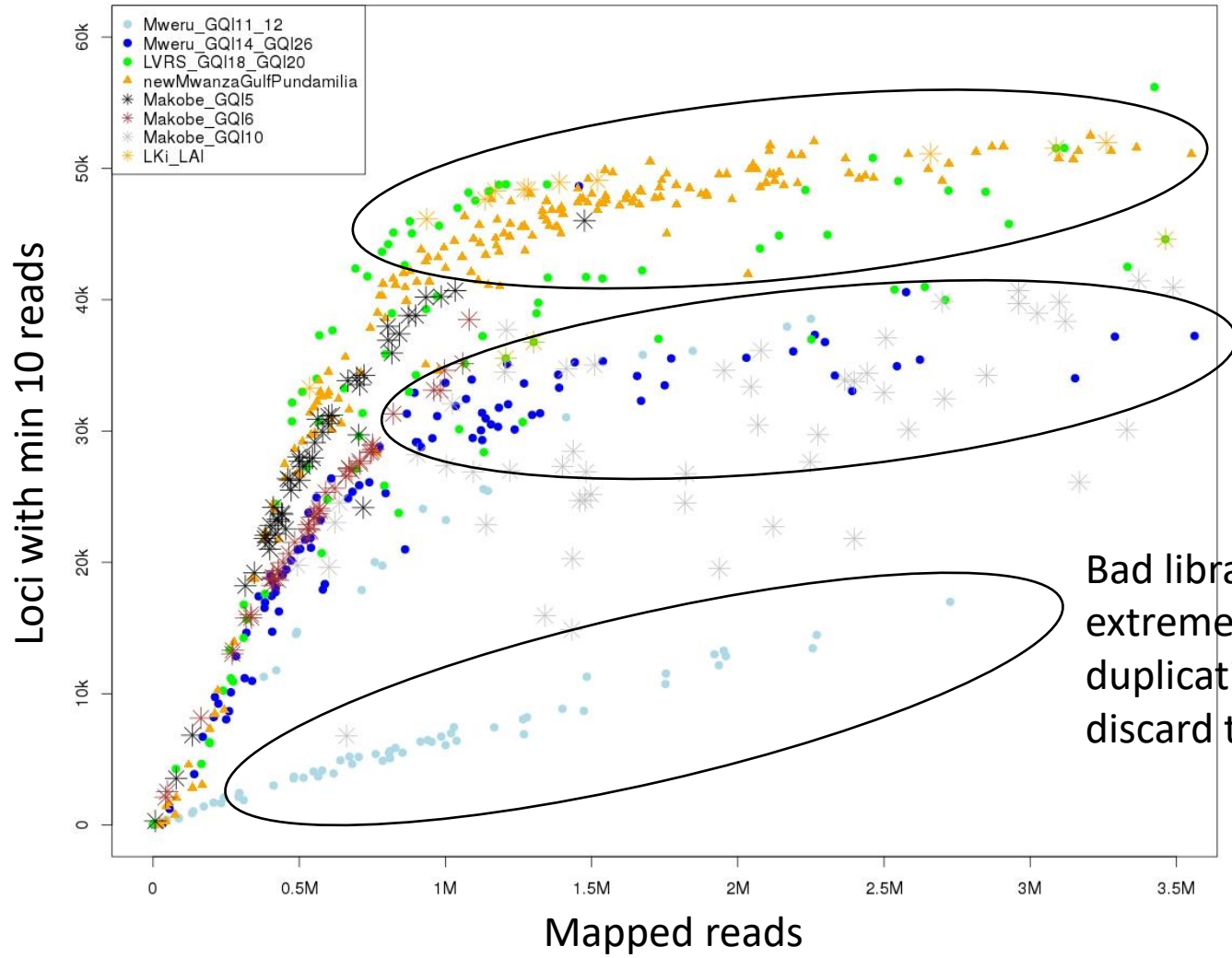
Detecting high levels of PCR duplication and contamination

Joana Meier

Checking PCR duplication levels of single-end sequenced RAD/GBS/UCE libraries



Great samples

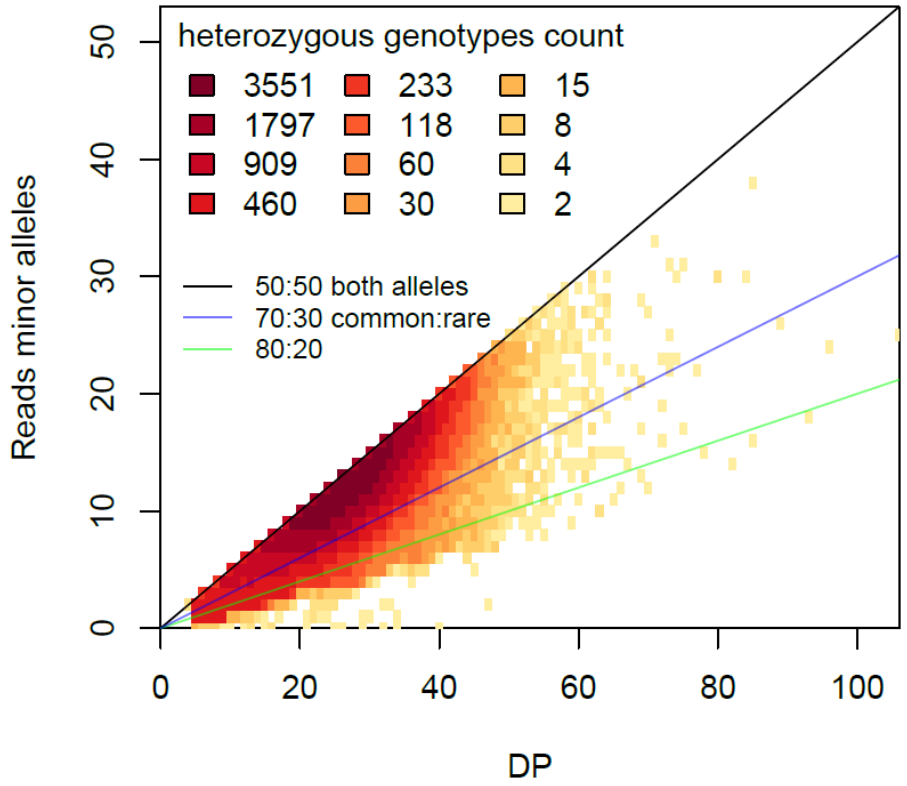
Samples at higher divergence from the reference genome

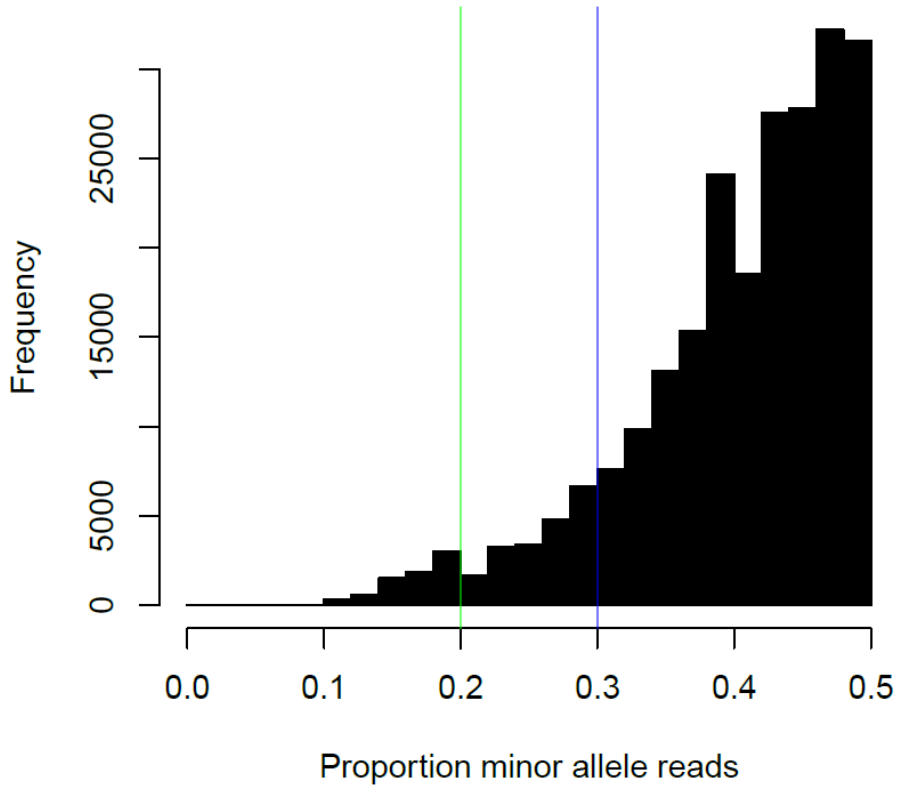
Bad library with extremely high PCR duplication levels -> discard the data!

Heterozygote positions are informative about potential contamination, Illumina barcode switching or PCR errors

If reads truly independent sequences of the same genotype (from different cells of the same individual), heterozygotes are expected to have a roughly equal number of reads supporting each allele.

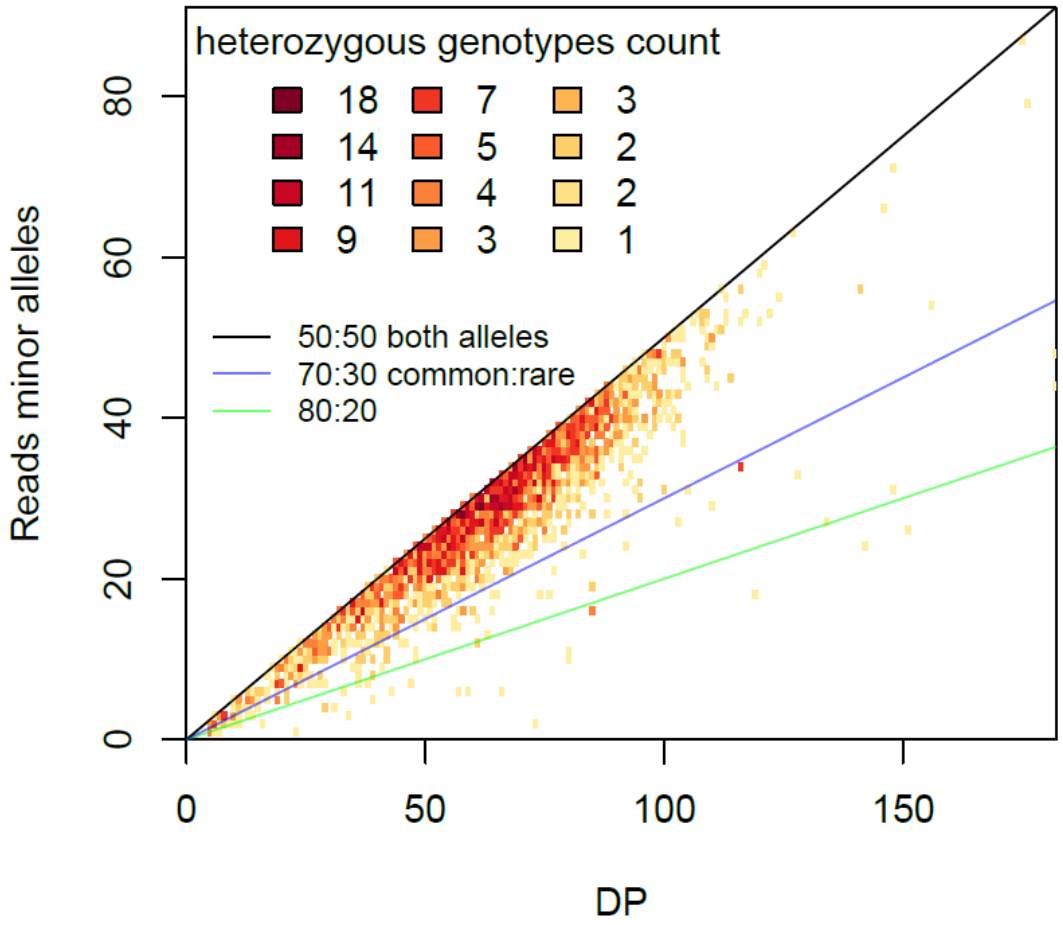
Example from a good wgs sample

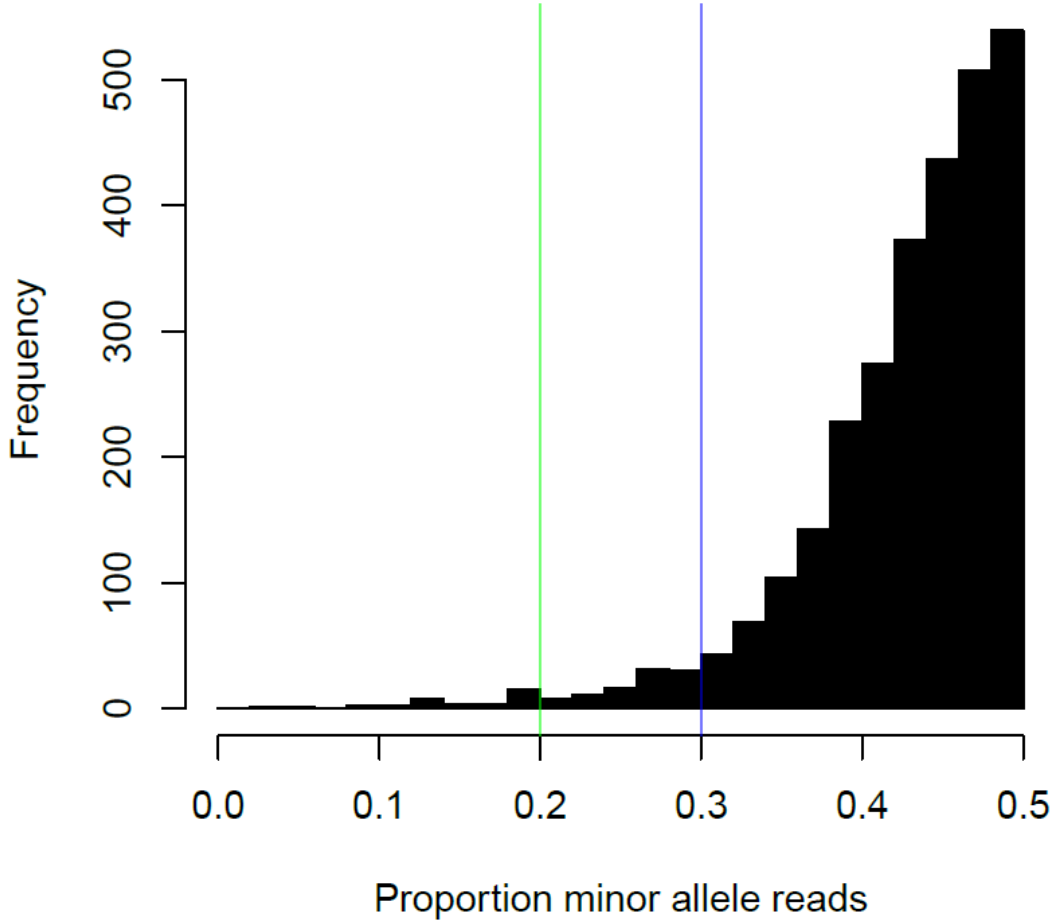




Heterozygote positions are informative about potential contamination, Illumina barcode switching or PCR errors

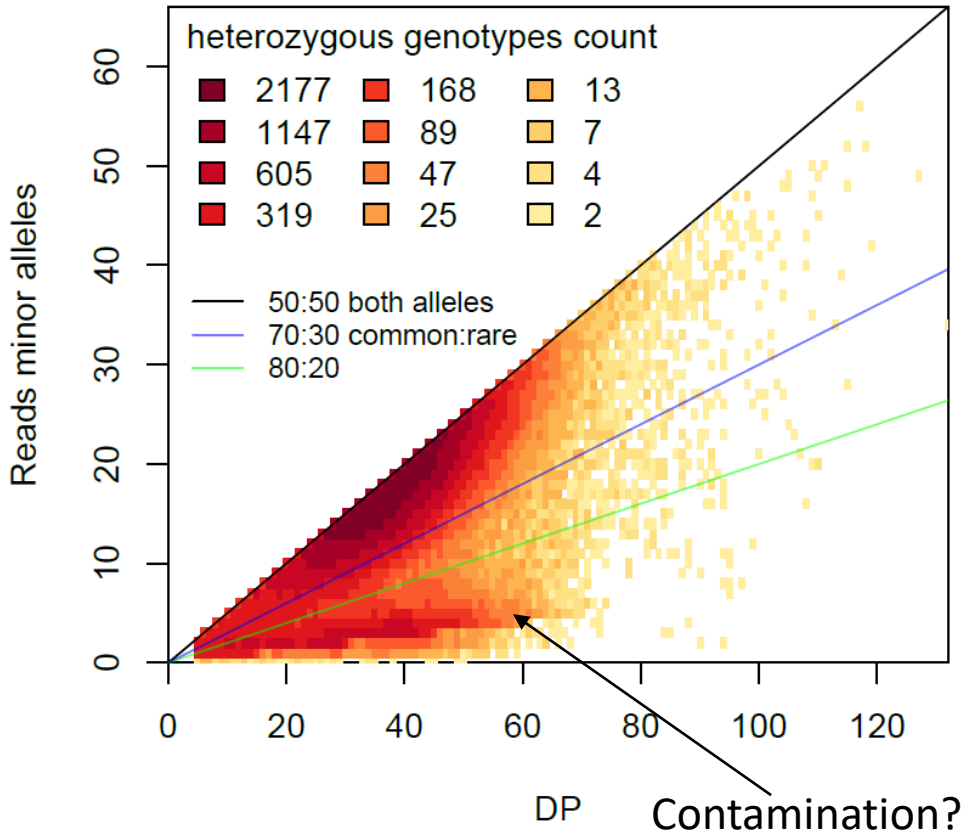
Example from a good RAD library

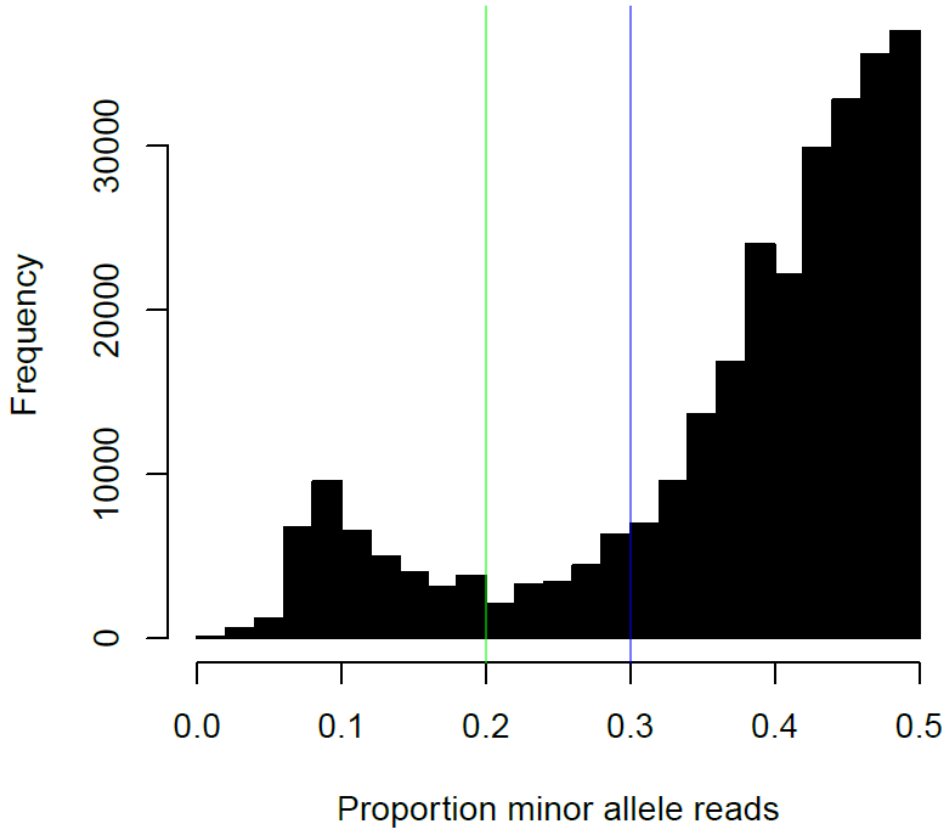




Heterozygote positions are informative about potential contamination, Illumina barcode switching or PCR errors

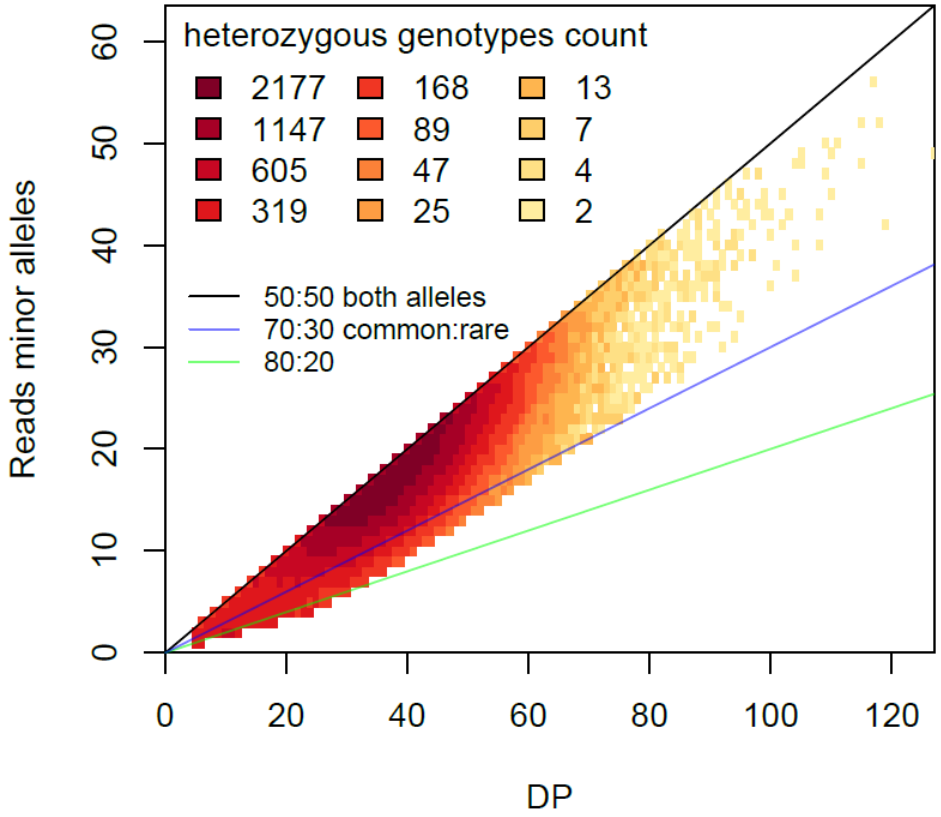
Example from a contaminated wgs sample

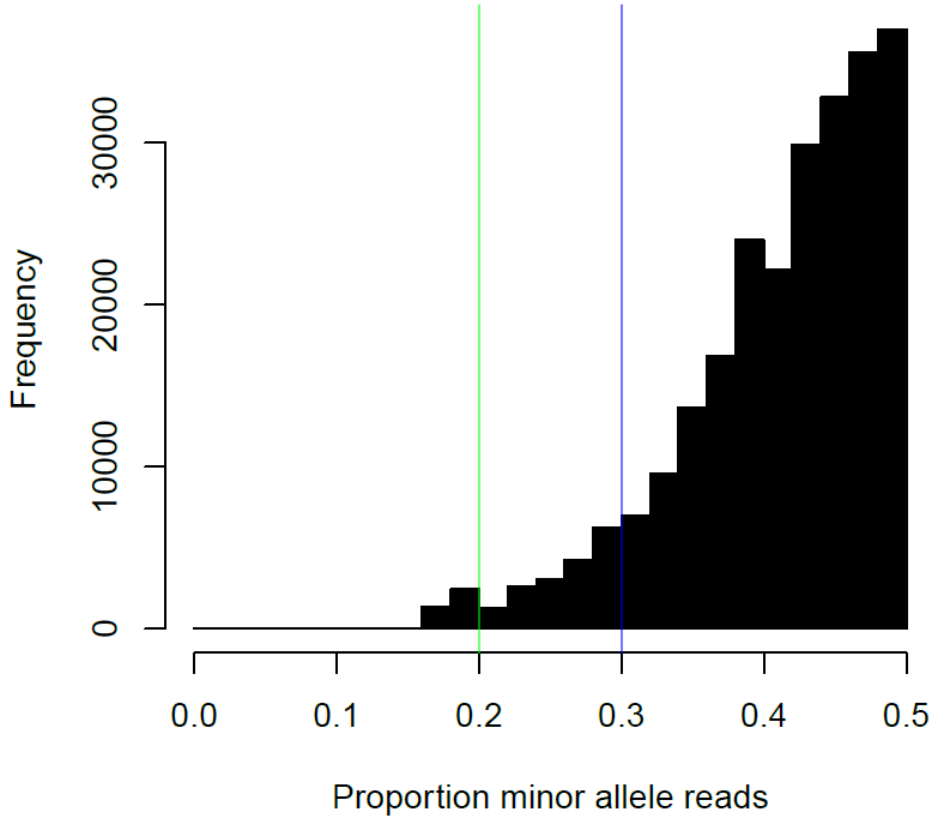




Contamination?
Illumina barcode switching?
PCR errors?

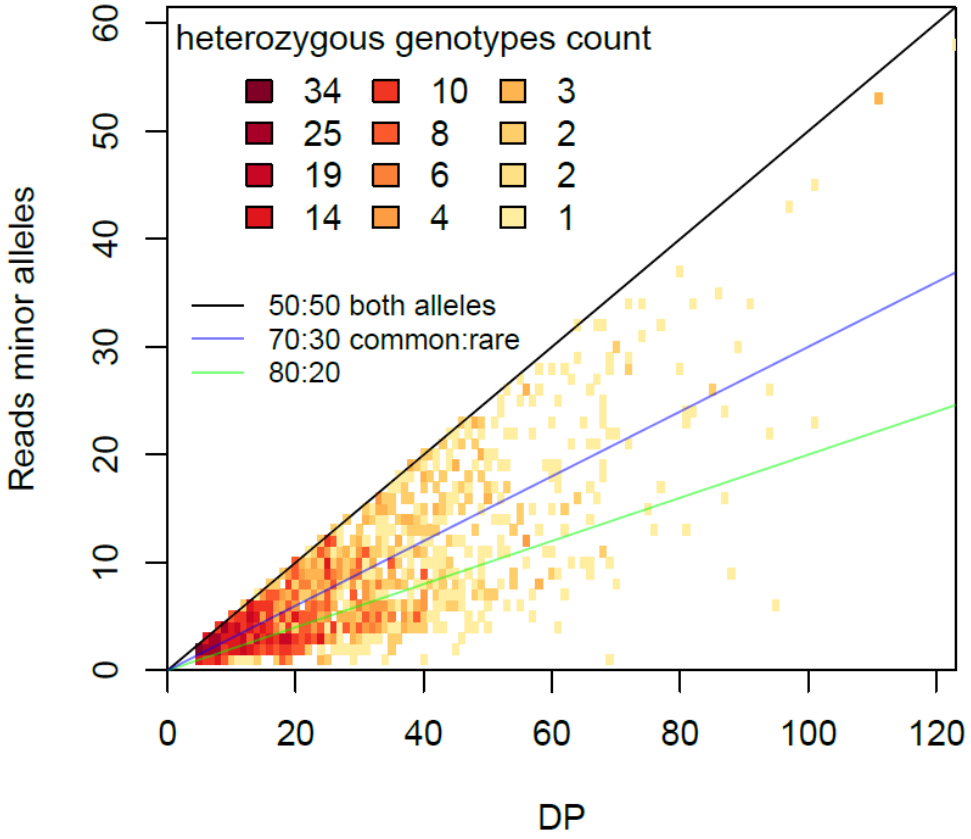
AllelicBalance.py to remove heterozygote positions failing a binomial test

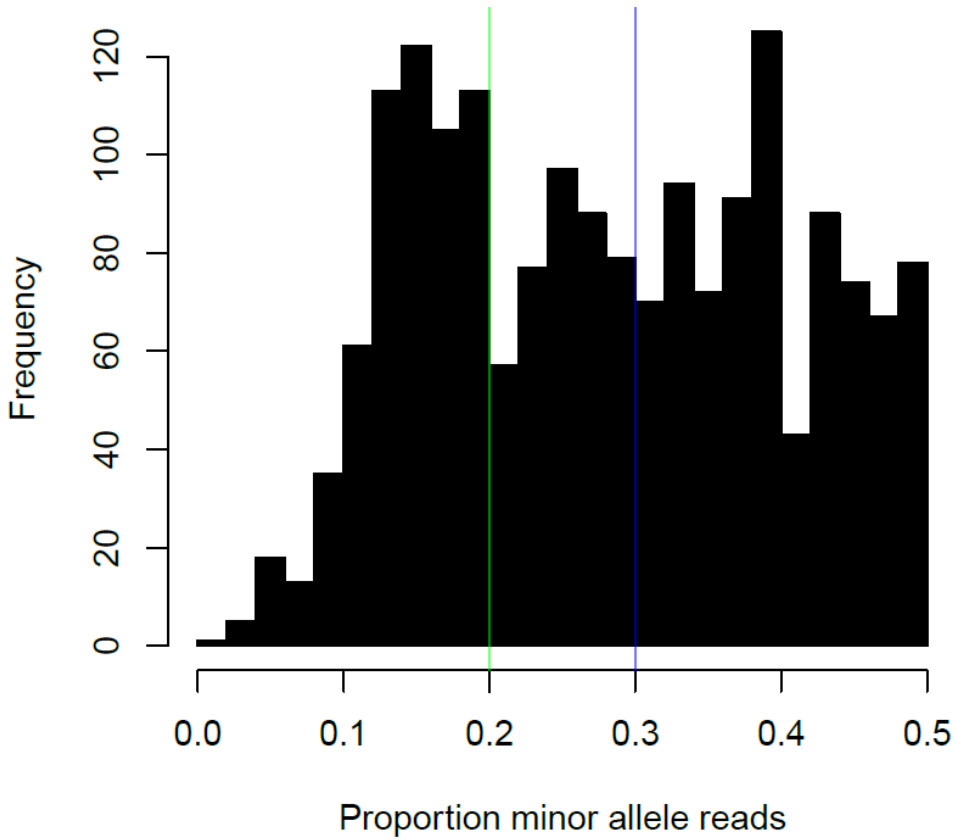




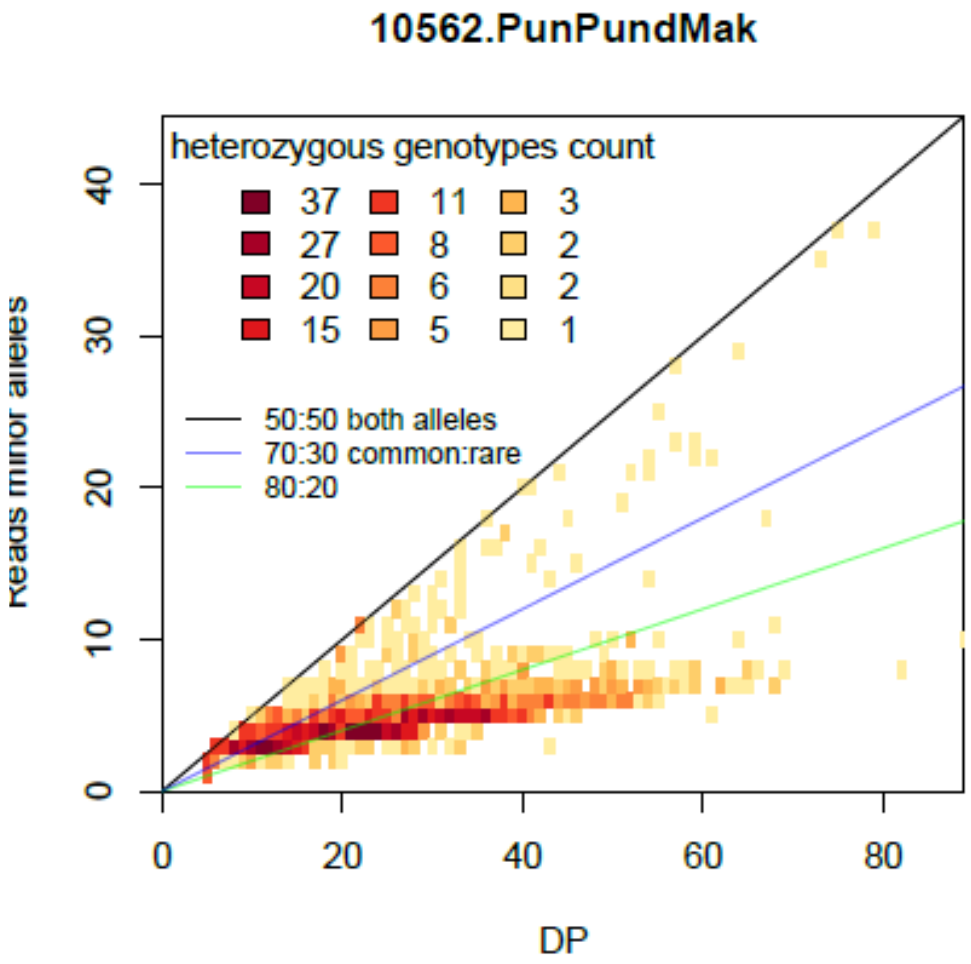
Heterozygote positions are informative about potential contamination, Illumina barcode switching or PCR errors

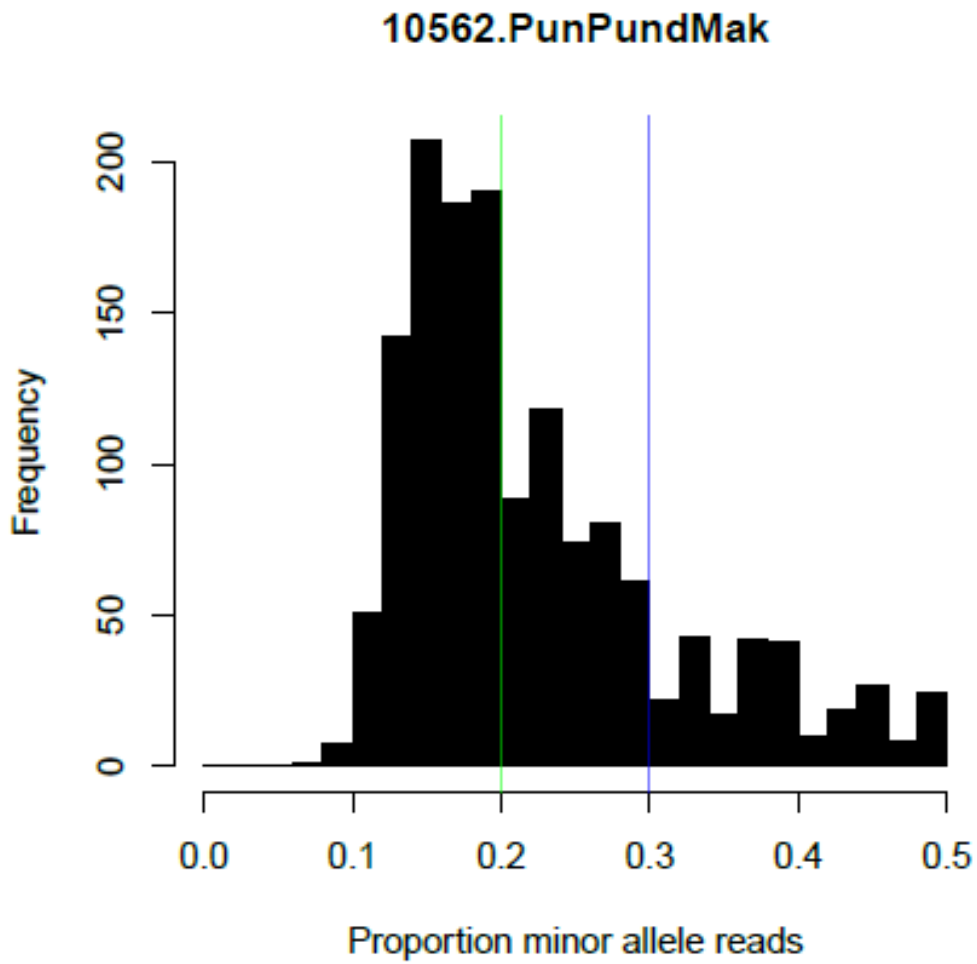
Example from a very bad RAD library





This individual looked ok but when looking only at sites with mac 1 across all individuals, PCR errors became very obvious





How to distinguish the different sources?

- **Illumina barcode switching**: Affects samples that are most divergent (wrongly assigned reads are most likely to cause SNPs) and samples that do not have good DNA quality, allelic balance should be better in singletons
- **PCR errors in libraries with PCR duplicates**: allelic balance worst in singletons, usually decreased heterozygosity due to allelic dropout, generally high variance of reads supporting each allele even at high sequencing depth -> make sure to remove singletons
- **Contamination**: increased heterozygosity expected, may appear to be a hybrid, no signal in singletons