davidsonii F2 mapping: VCF filtering: data NOT phased

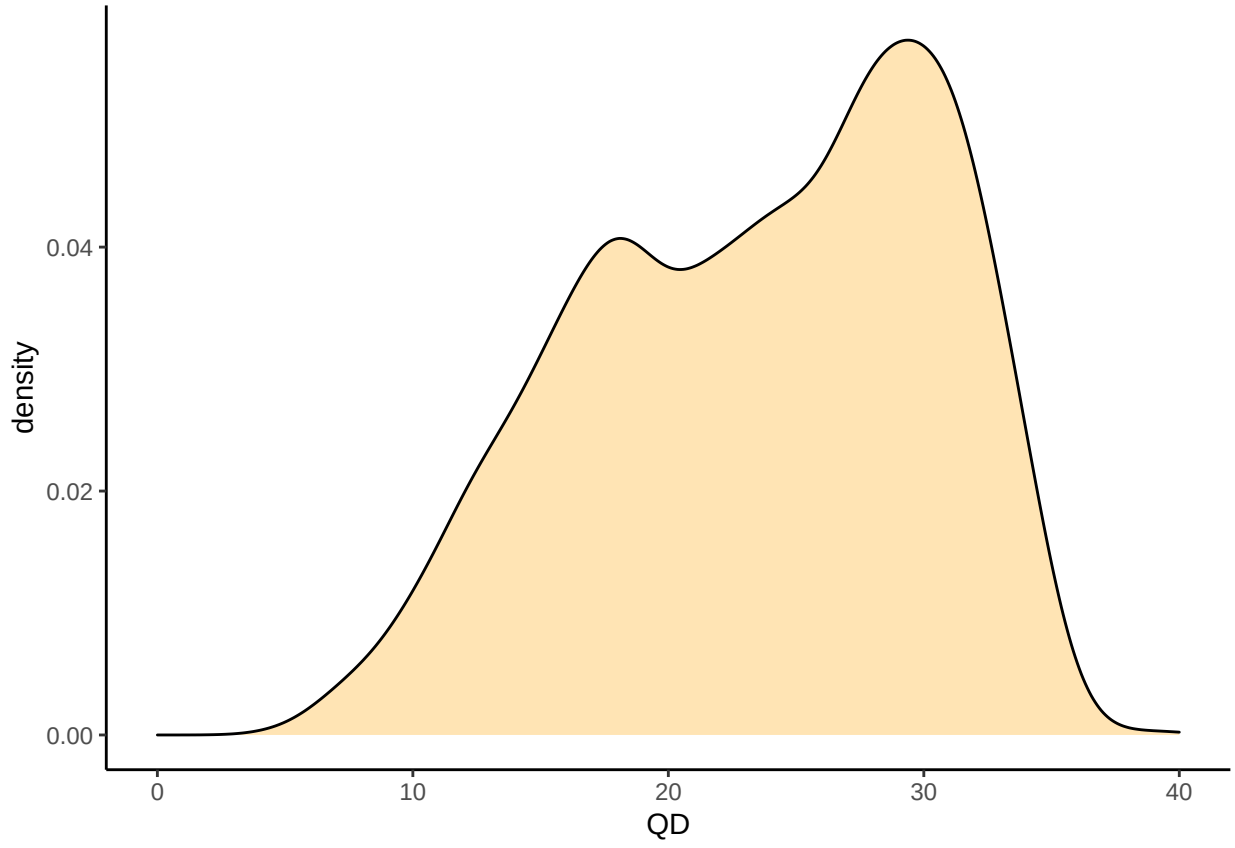
The VCF file was filtered in the following way:

- –do-not-run-physical-phasing option implemented in HaplotypeCaller
- minimum Mapping Quality is 30
- no more than ~40% missing data (50 or more individuals must be present)
- allele frequencies must be in hardy-weinberg proportions, and $0.3 \leq q \leq 0.7$
- Single SNP per 300 bp
- at least 8 individuals with minor allele

This resulted in a data set with 2648 SNPs (nearly double the phased output).

Quality by depth

GATK best practices recommend filtering QD < 2



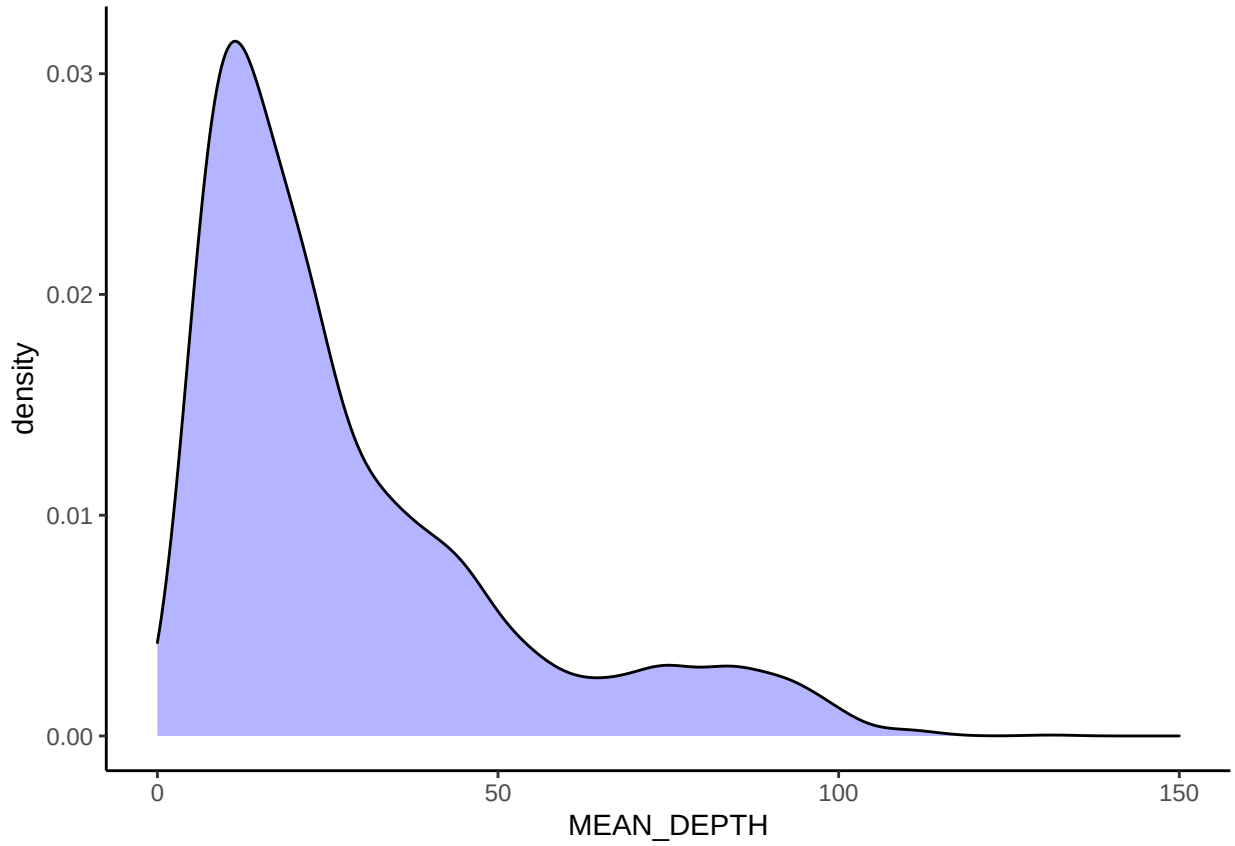
This looks very good. We have no low quality sites. After filtering, there are no sites with QD < 5:

```
length(which(t<5))
```

```
## [1] 0
```

Depth of Coverage

Higher coverage is better, obviously. But, reads with too high coverage could be mapping/assembly errors and/or repetitive regions. Ravinet & Meier suggest a good "rule of thumb" is filtering max depth > 2x mean depth, but I have seen less stringent filters elsewhere.



This looks pretty good. If we look for the proportion of reads > 2x mean depth...

```
length(which(t$MEAN_DEPTH > mean(t$MEAN_DEPTH)*2))/nrow(t)
```

```
## [1] 0.1257553
```

12.9% are higher than 2x mean. But none are particularly high coverage. Given this is ddrad data, nothing here screams mapping error to me. We also have only a few loci with low coverage:

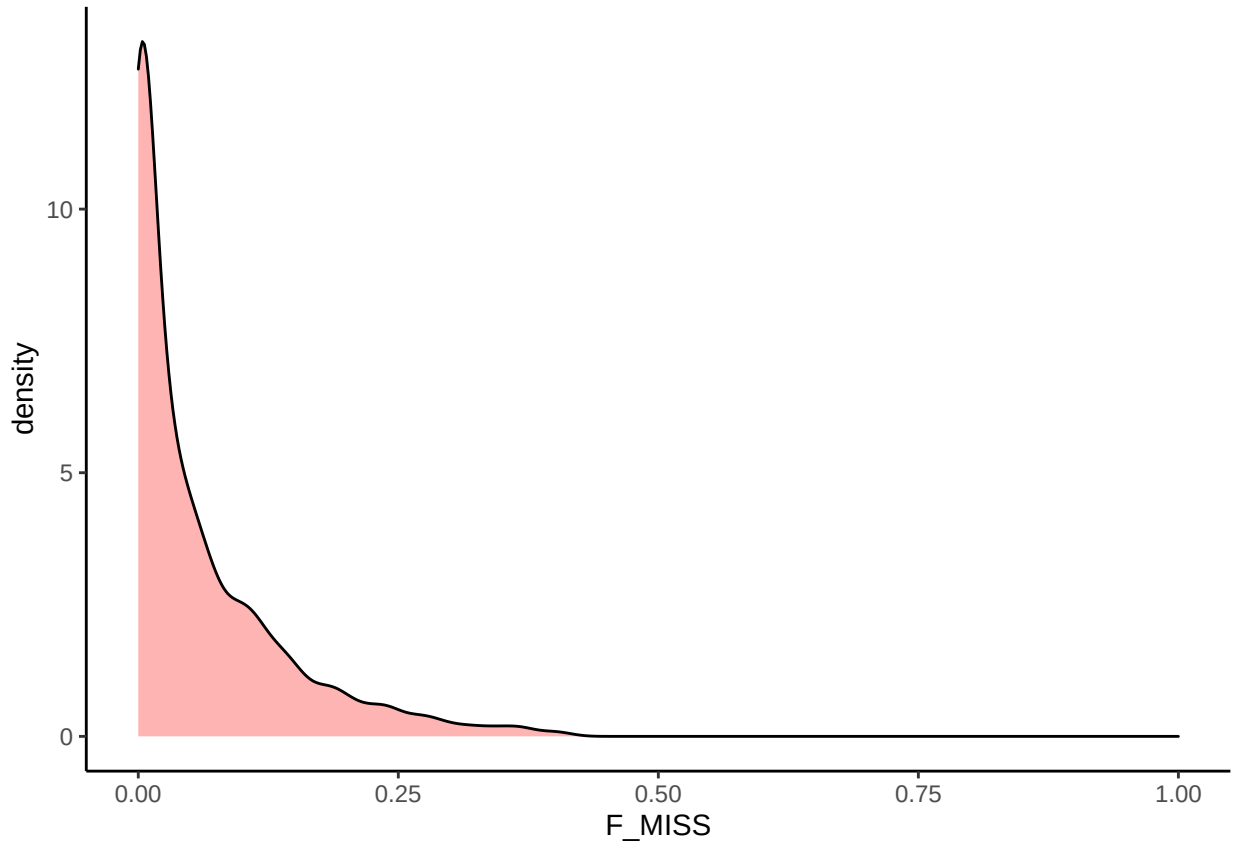
```
length(which(t$MEAN_DEPTH < 10))/nrow(t)
```

```
## [1] 0.189577
```

```
length(which(t$MEAN_DEPTH < 5))/nrow(t)
```

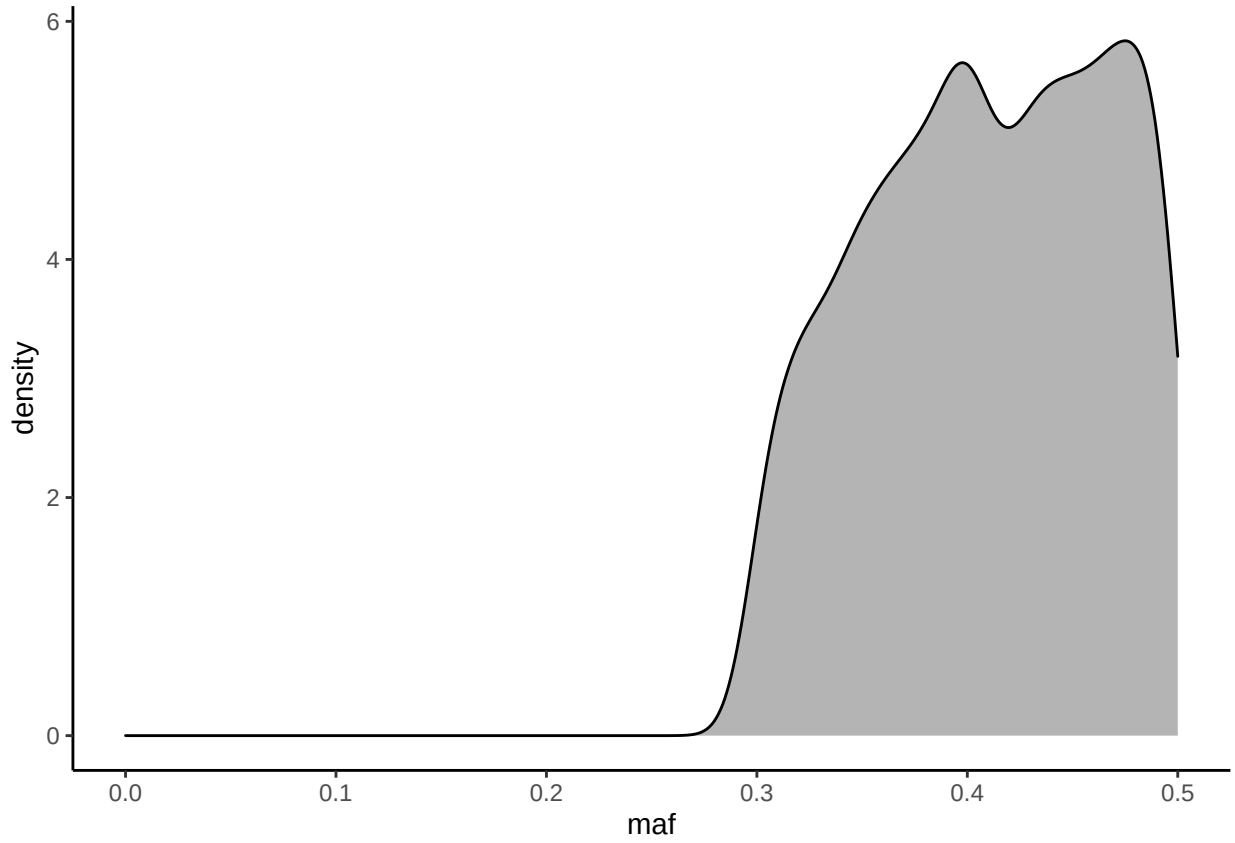
```
## [1] 0.02416918
```

Missing Data



Looks how we would expect: we filtered for no more than 33 individuals with missing data (~40%). Also, because we didn't implement GQ filters on this data, we didn't change reads to missing that didn't pass some quality threshold.

Minor Allele Frequency

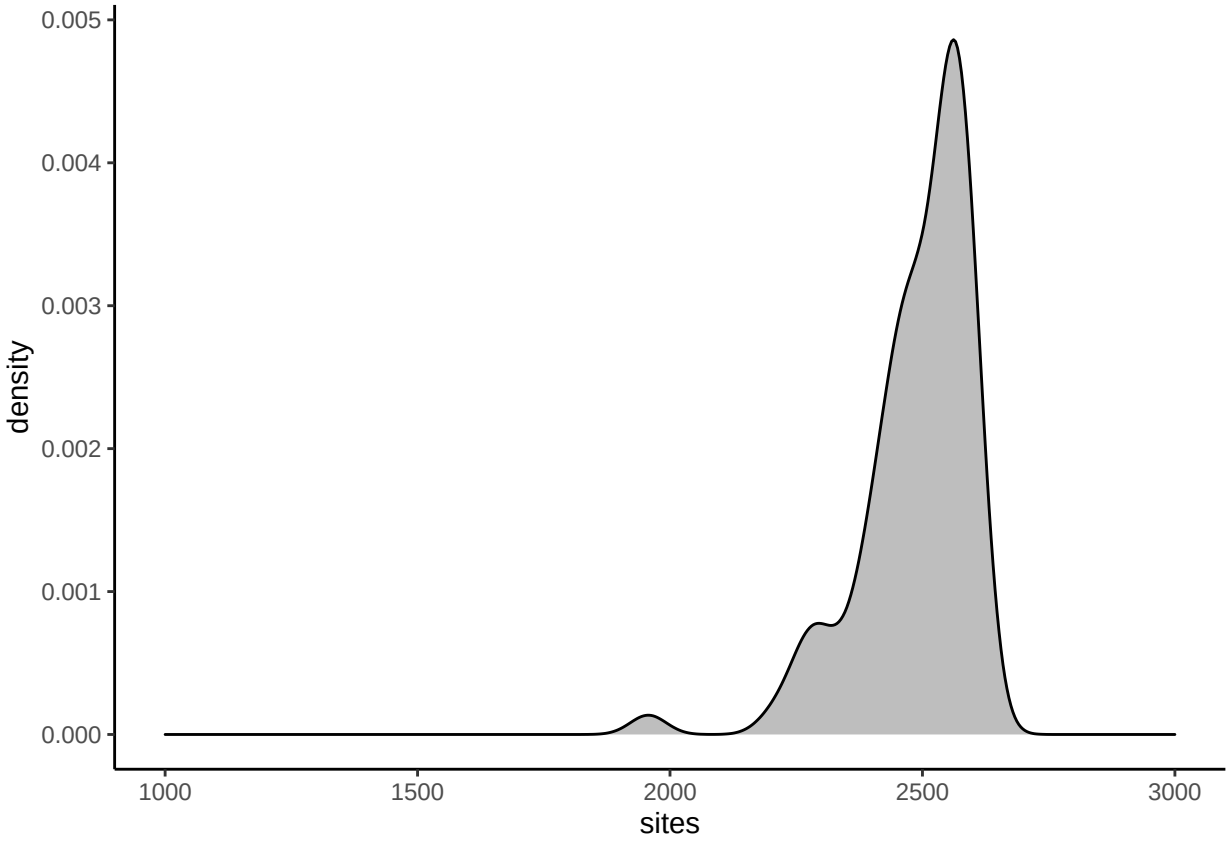


Again, we filtered this so that minor allele frequency is always > 0.3. So no surprise.

Heterozygosity

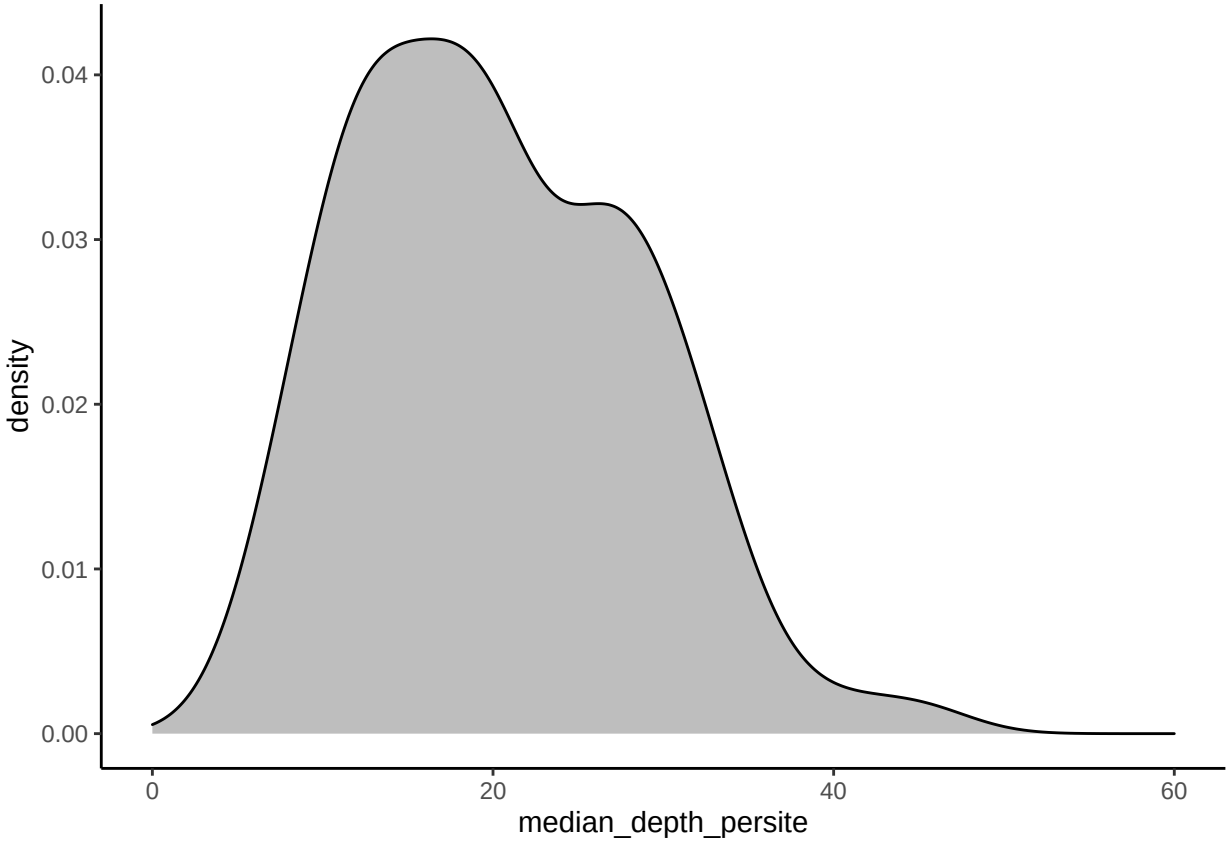
The remaining plots are generated from sites extracted from calc.sample.coverage.from.vcf.py.

Total sites



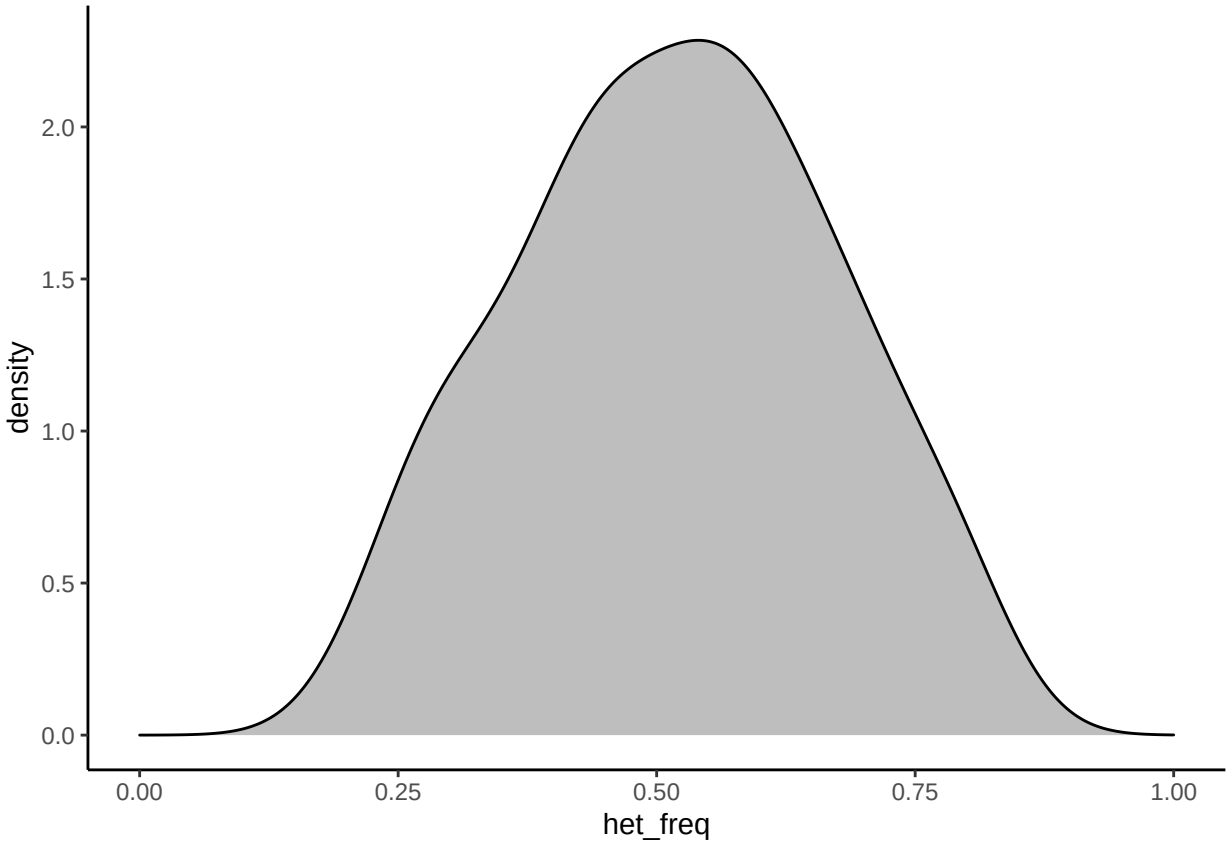
Given low proportions of missing data it isn't surprising to see that most individuals have ~ the same number of SNPs.

Median depth per site



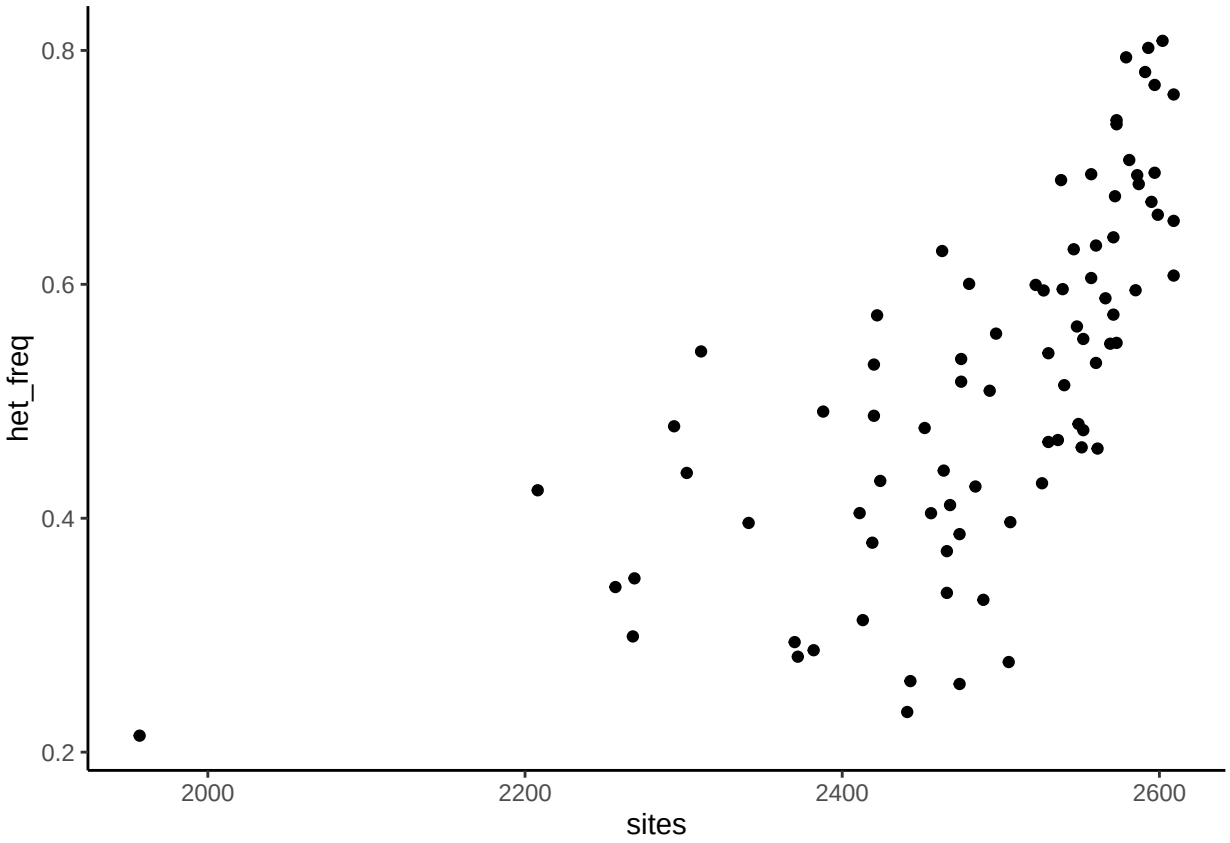
This plot is now shifted lower, so it looks like the filtering GQ did have an effect. The median depth/site now looks to be comparable to Carrie's example with *barbatus* and *neomexicanus* F2s.

Heterozygosity/sample



Bell curve centered around 50% heterozygote frequency (which we expect at these sites). Looks less skewed than when filtering for GQ and DP. However: there are still individuals that are heterozygous at $> 80\%$ of sites.

Heterozygosity by number of sites



Again: it is still clear that there are some individuals which are heterozygous at most sites. See above, and here:

```
## [1] 0.1084337
```

~11% of individuals are heterozygous at $> 70\%$ of sites. This is a decrease from before (it was ~18%). Conduct test to see whether this is outside expectations?