davidsonii F2 mapping: VCF filtering

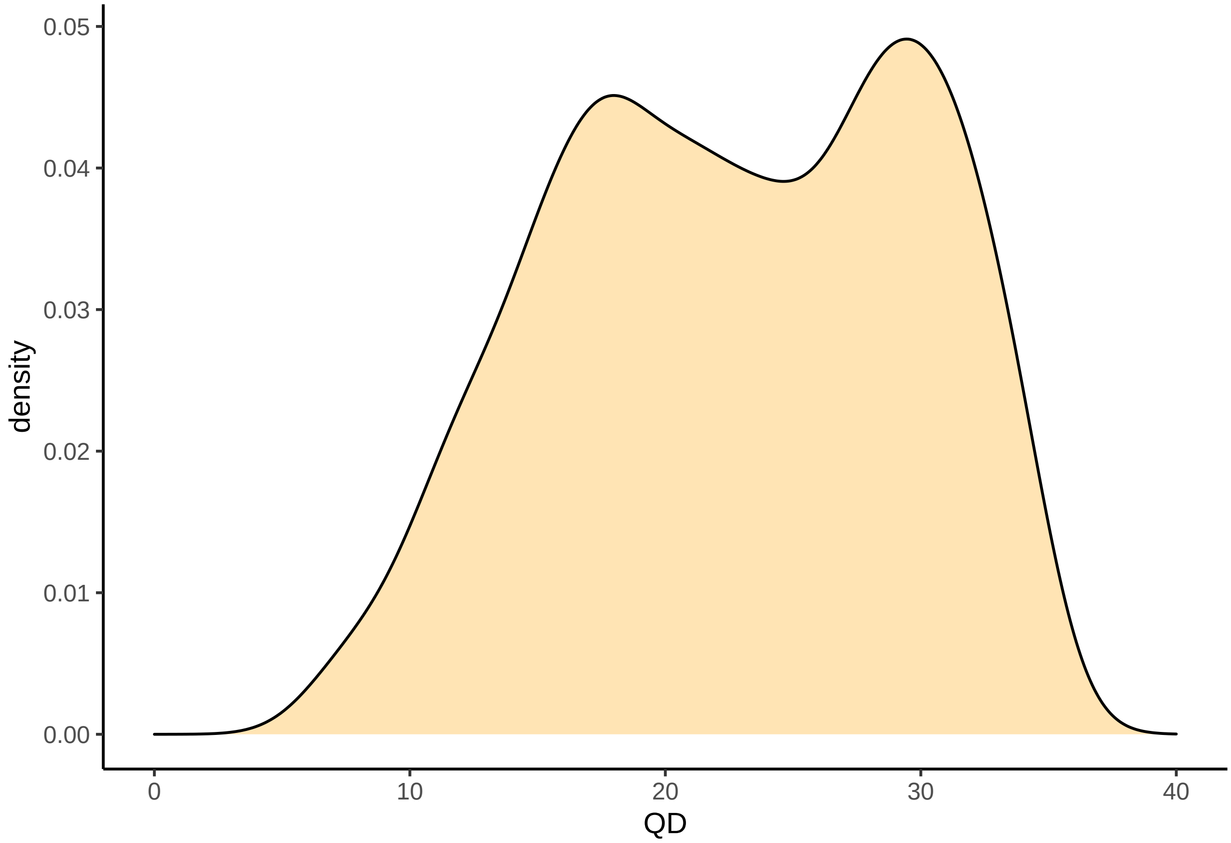
This report is for min50_hwe.30_300bp. The VCF file was filtered in the following way:

- minimum Genotype quality is 20 (99% accuracy)
- minimum allele depth is 4
- 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 1427 SNPs.

Quality by depth

GATK best practices recommend filtering QD < 2



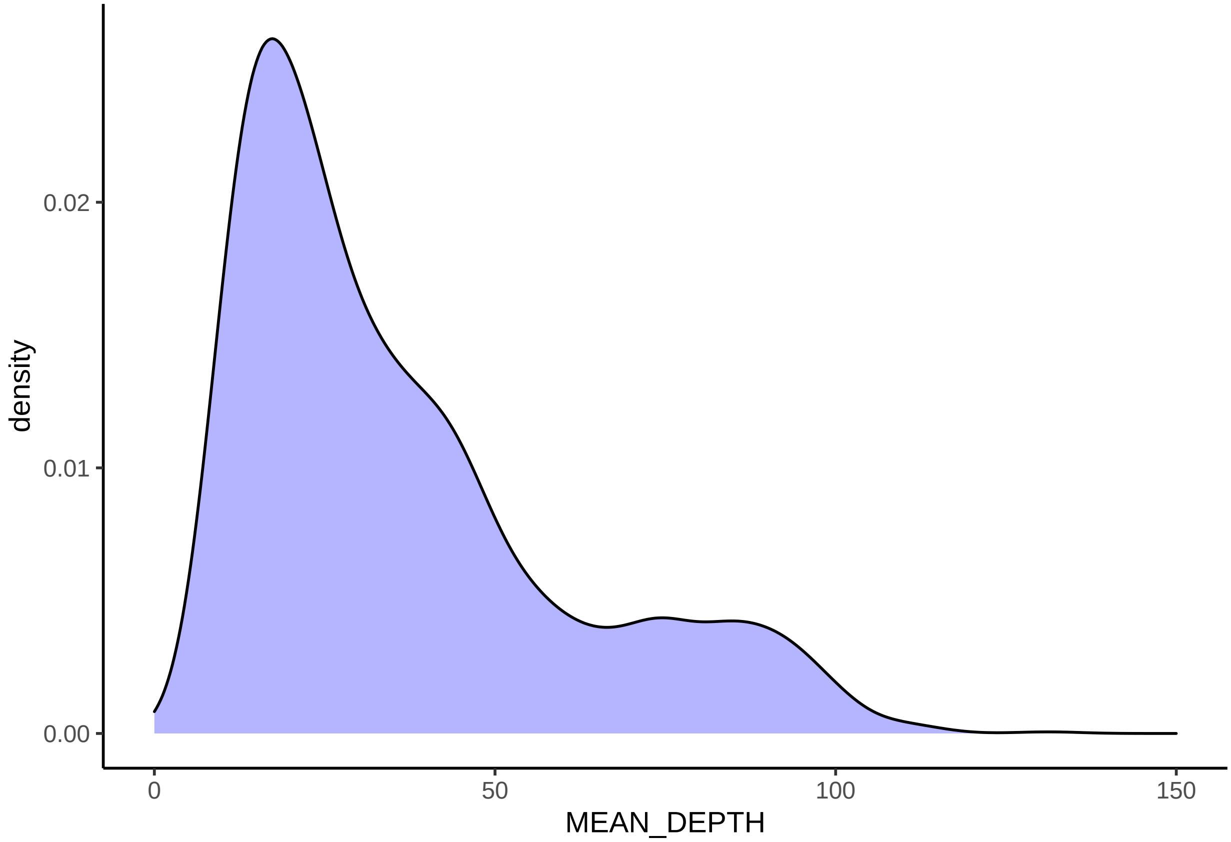
This looks very good. We have no low quality sites. After filtering, there are no sites with QD < 5:

```
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] 0.1226349
```

12.3% 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 nothing with particularly low coverage:

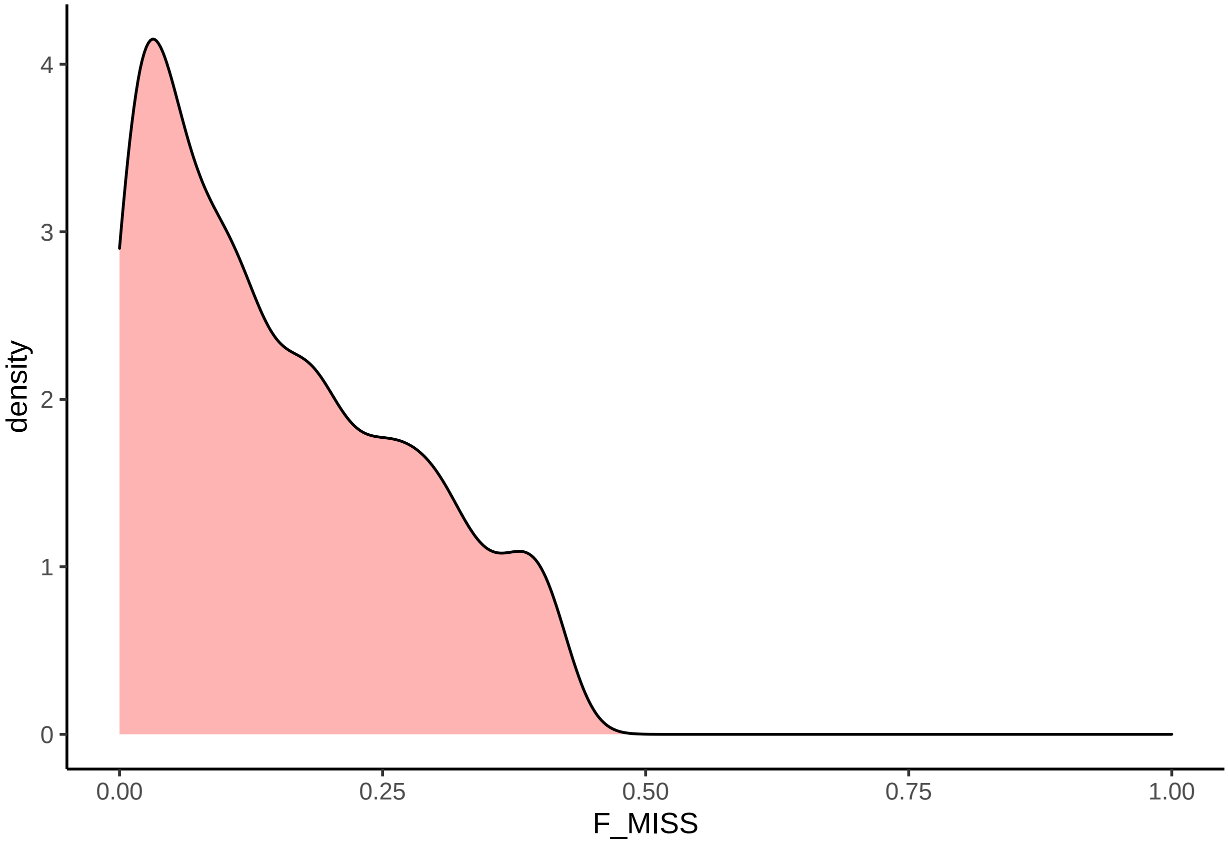
```
length(which(t$MEAN_DEPTH < 10))/nrow(t)
```

```
## [1] 0.02733006
```

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

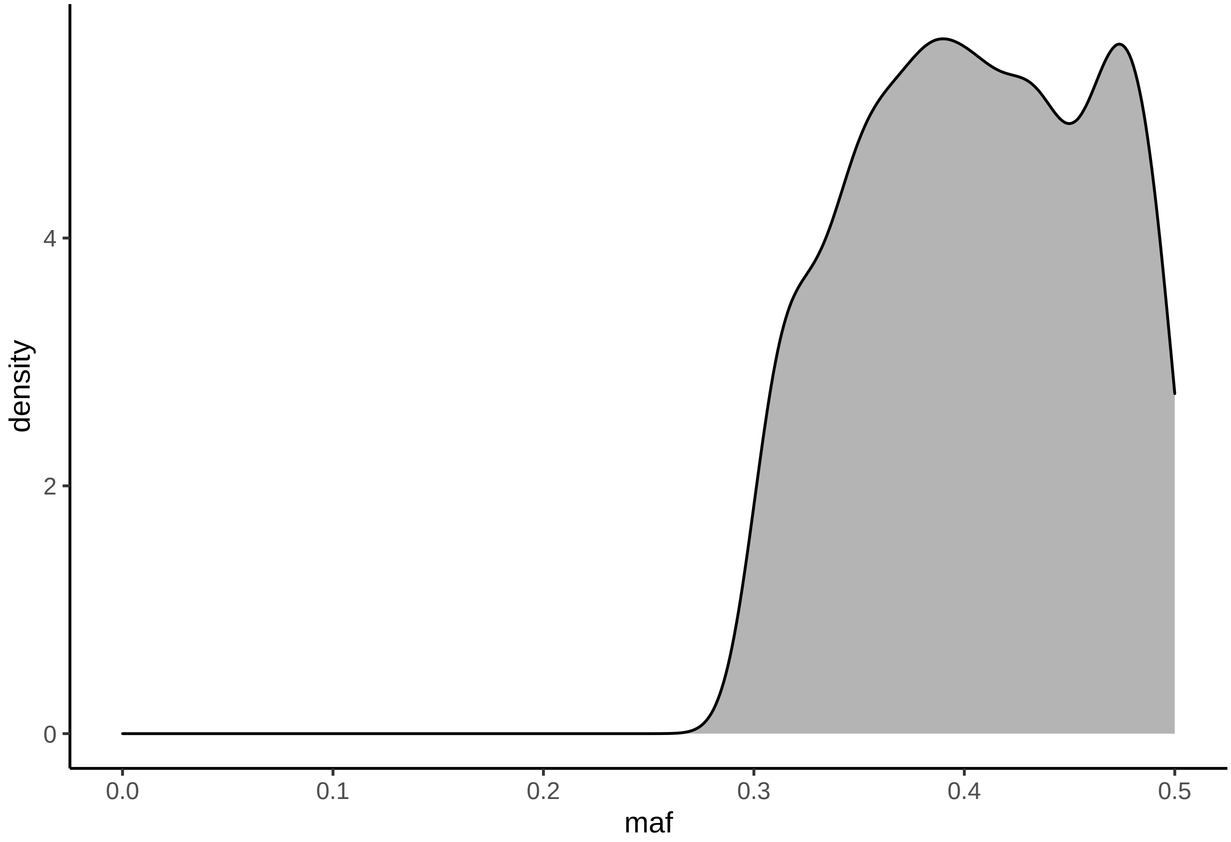
```
## [1] 0
```

Missing Data



Looks how we would expect: we filtered for no more than 33 individuals with missing data (~40%)

Minor Allele Frequency

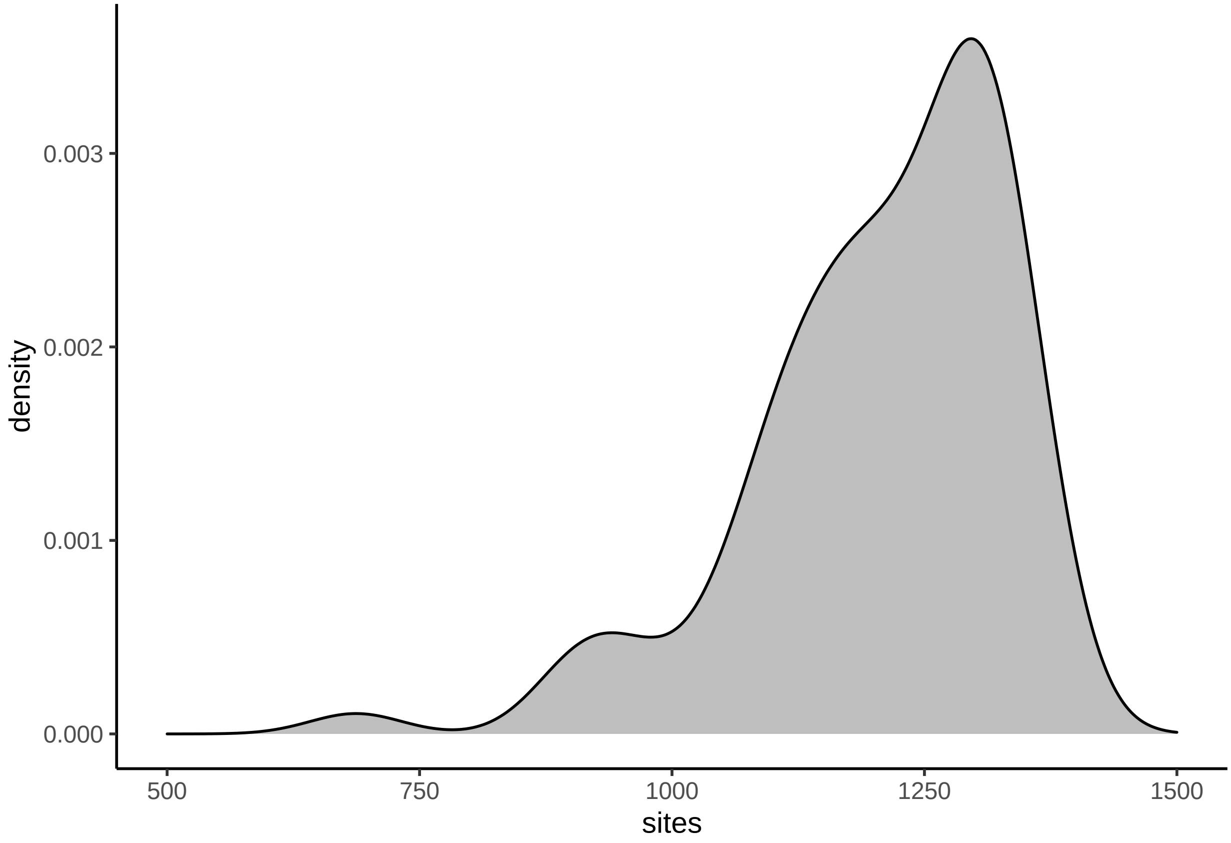


Again, we filtered this so that minor allele frequency is always > 0.3. So no surprise.

Heterozygosity

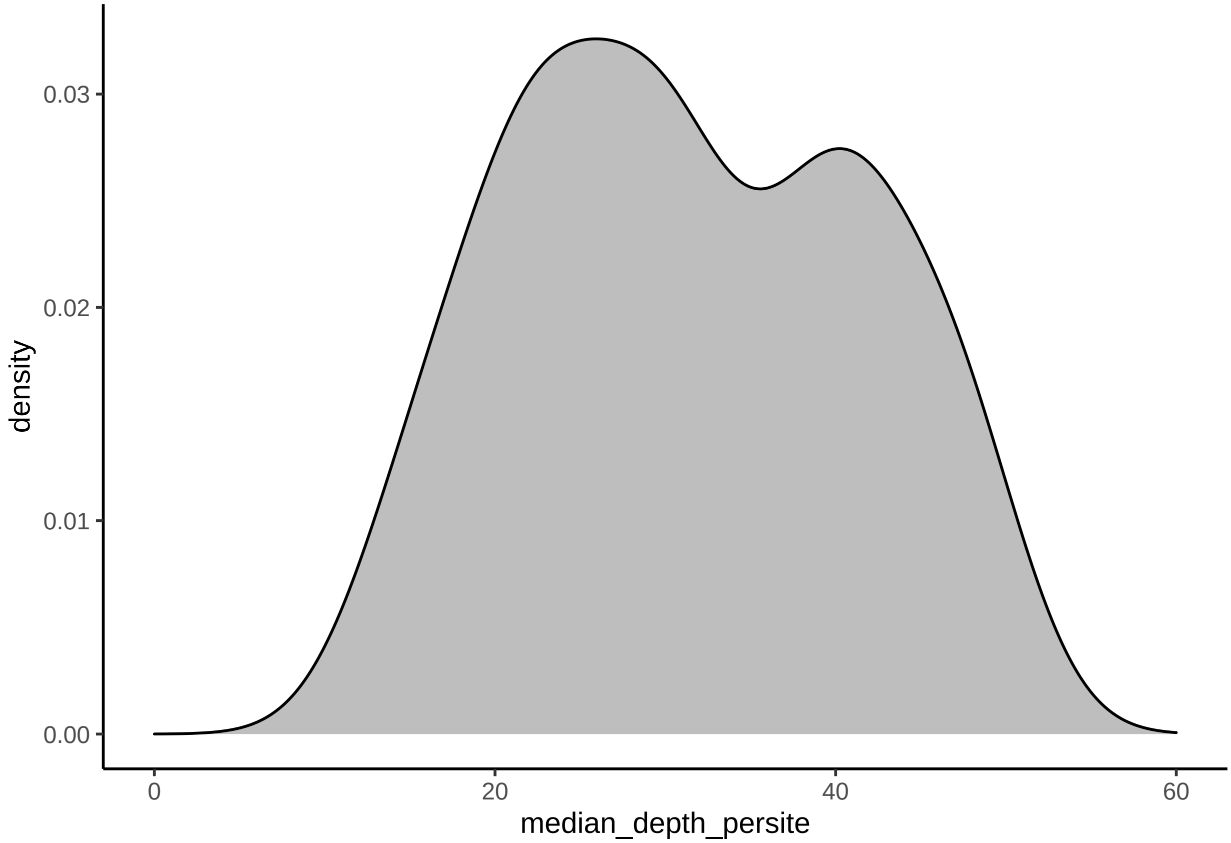
The remaining plots are generated from sites extracted from calc.sample.coverage.from.vcf.py.

Total sites



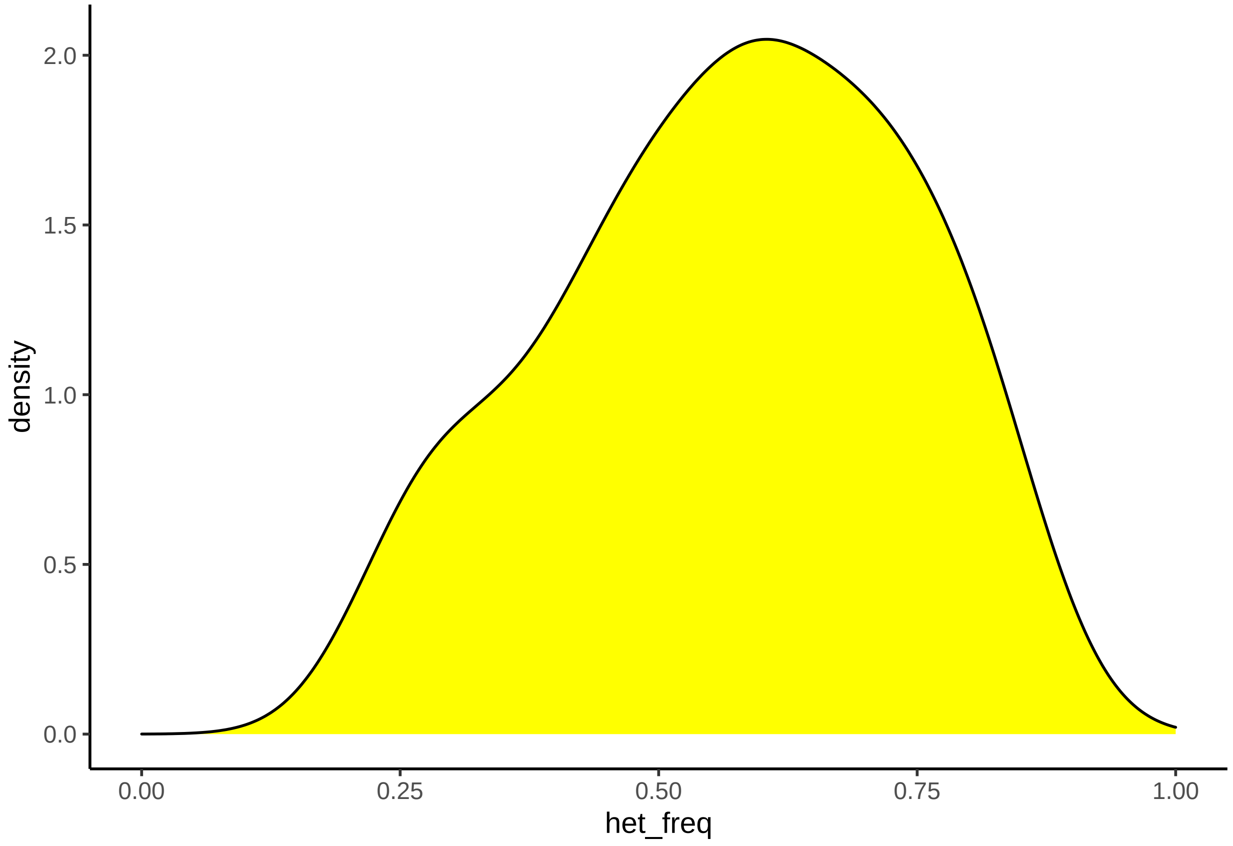
The majority of individuals have > 1000 sites.

Median depth per site



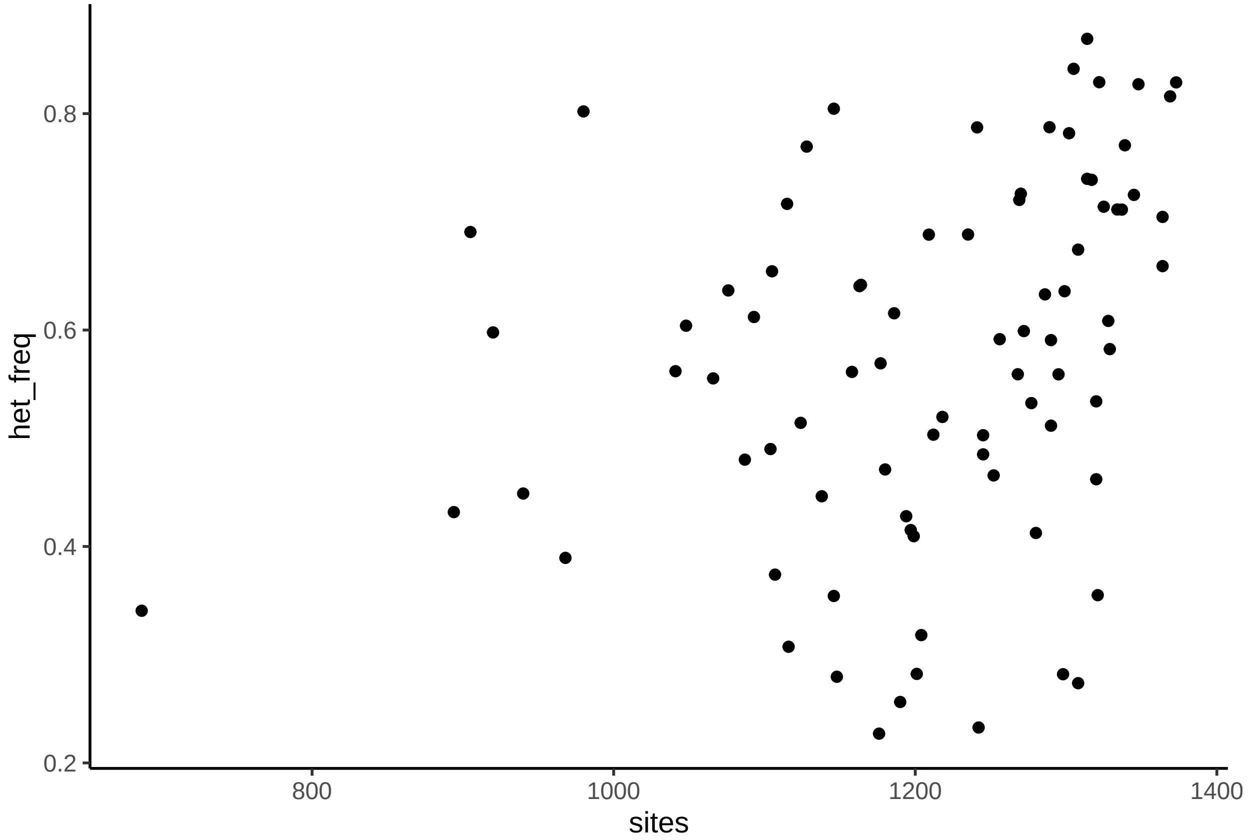
The mediant depth/site is much higher than Carrie's example with *barbatus neomexicanus* F2s. This is presumably because I filtered GQ < 20. With a less stringent filter there we could return much more data

Heterozygosity/sample



A nice bell curve centered around ~50%. There are no individuals with < 20% heterozygosity.

Heterozygosity by number of sites



Don't see a similar issue with Carrie's data set regarding relationship between number of reads/heterozygosity. Given the strict filters we imposed this was likely to be the case. However this also leaves us with fewer data. Relaxing thresholds on missing data and on genotype quality (QC) would increase the number of SNPs.