VCF Analysis

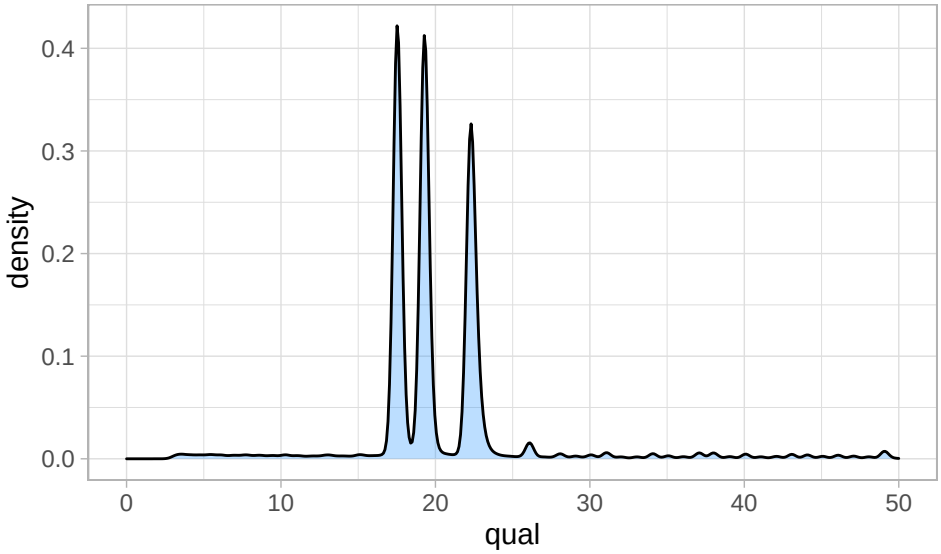
Austin Rosen

8/25/21

Variants were called using the samtools mpileup tool on our BAM files produced by HybPiper. This takes all reads at a given position and calls variants against the reference genome from the reads covering that position for all individuals. Using the known gene sequences for each loci as the reference genome, 899036 variants were obtained.

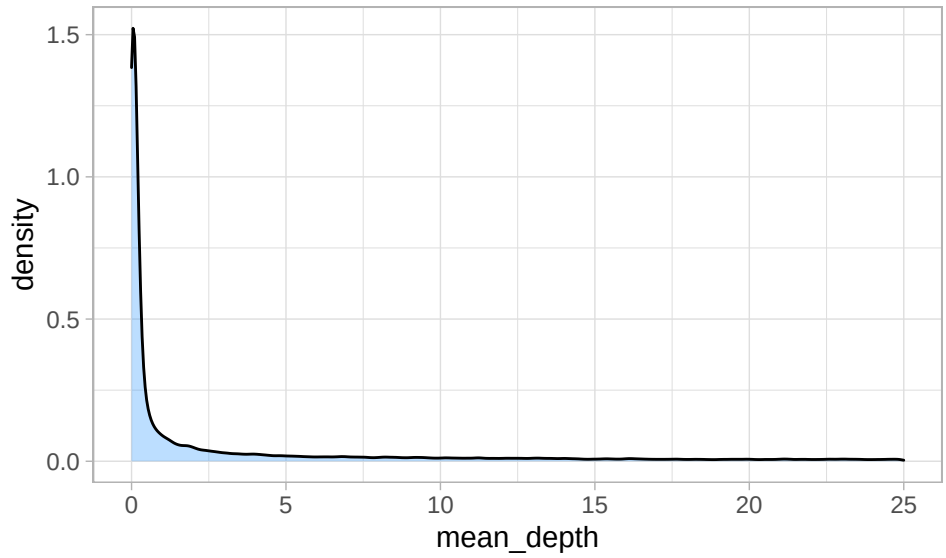
Chances are many of them are not useful for our analysis because they occur in too few individuals, their minor allele frequency is too low, they are covered by insufficient depth or are low quality sites. Lets's investigate our vcf (Variant Call Format) dataset.

#1. Check Variant Quality



Site quality scores are fairly low. Phred scores of 30 represents 99.9% accuracy, 20 represent 99% accuracy, and 10 represents 90% accuracy of the SNP call.

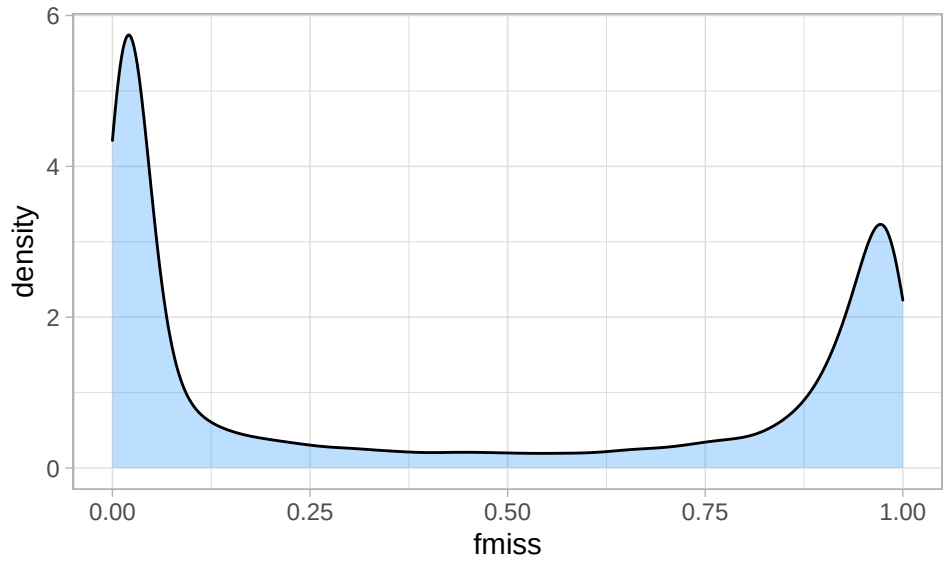
#2. Check variant mean depth



Mean of the read depth across all individuals represents the number of reads that have mapped to each position- it is for both alleles at a position and is not partitioned between the reference and the alternative.

Most of our variants sites clearly have very low coverage (< 2x). A minimum threshold of 10x is often recommended.

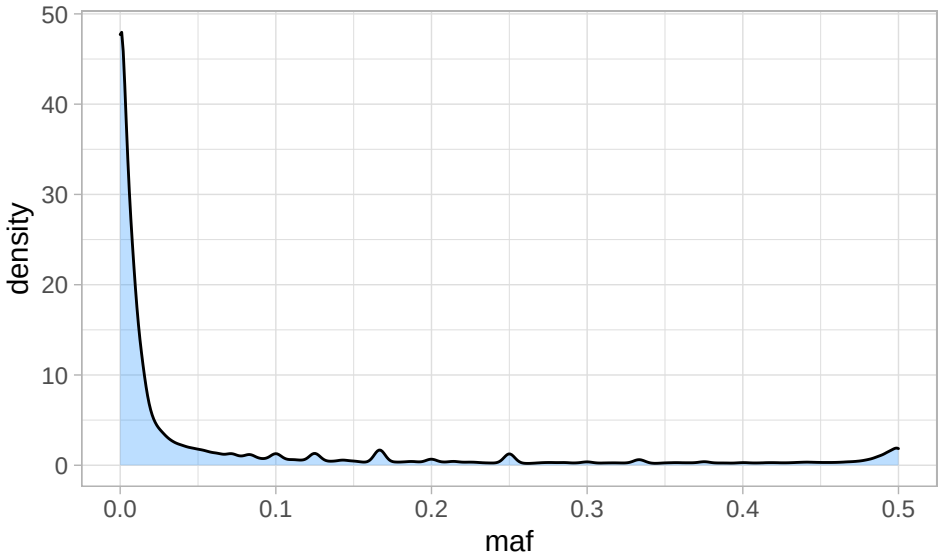
#3. Variant missingness



This is a measure of how many individuals lack a genotype at a call site.

This figure shows that there are many sites in which there is a call for almost every individual, but also many sites in which there are calls for only a very few individuals. Typically missingness of 75-95% is used.

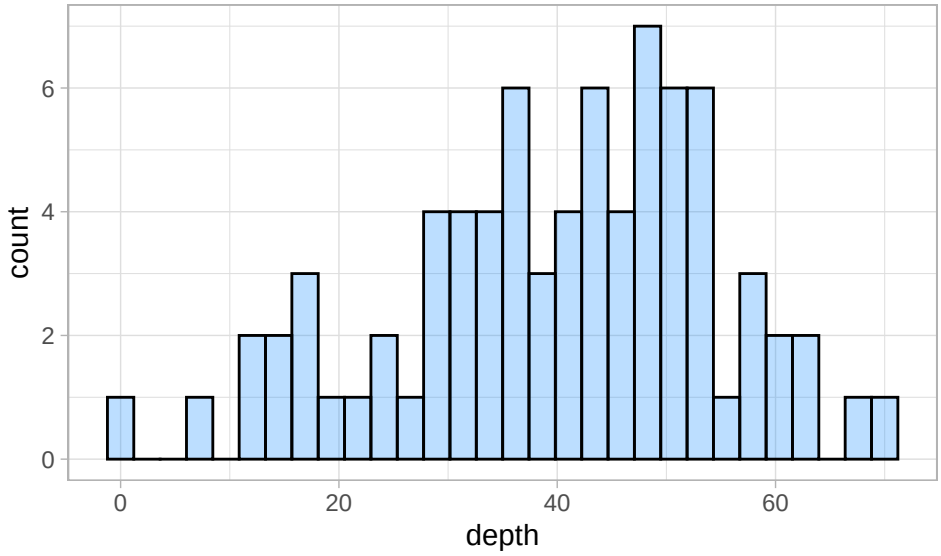
#4. Minor allele frequency




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0     0.0     0.0     0.1     0.1     0.5  777551
```

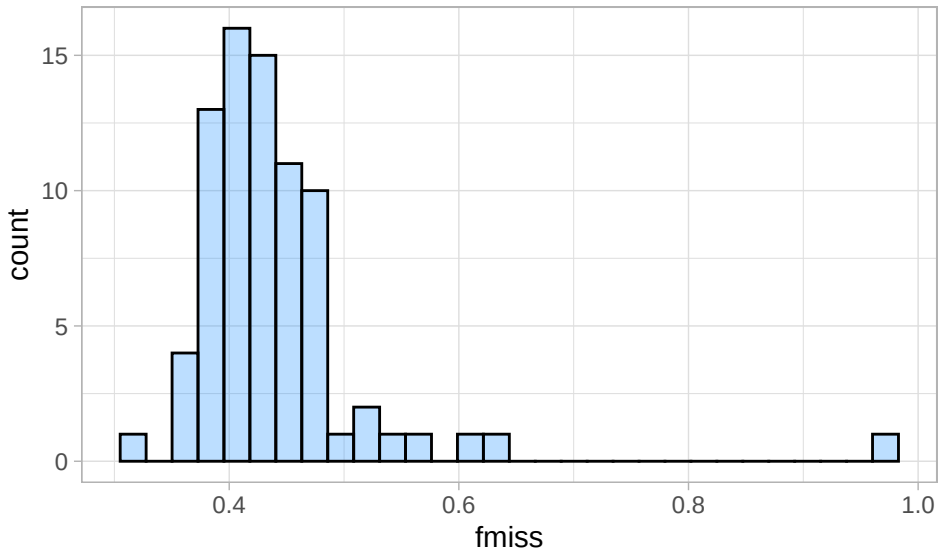
A large number of variants have low frequency alleles, which often result in uninformative loci.

#5. Mean depth per individual



This figure shows that most individuals were sequenced to a fairly acceptable depth, aside from perhaps 1 individual (probably individual #175 which didn't find any relevant sequences during HybPiper assembly).

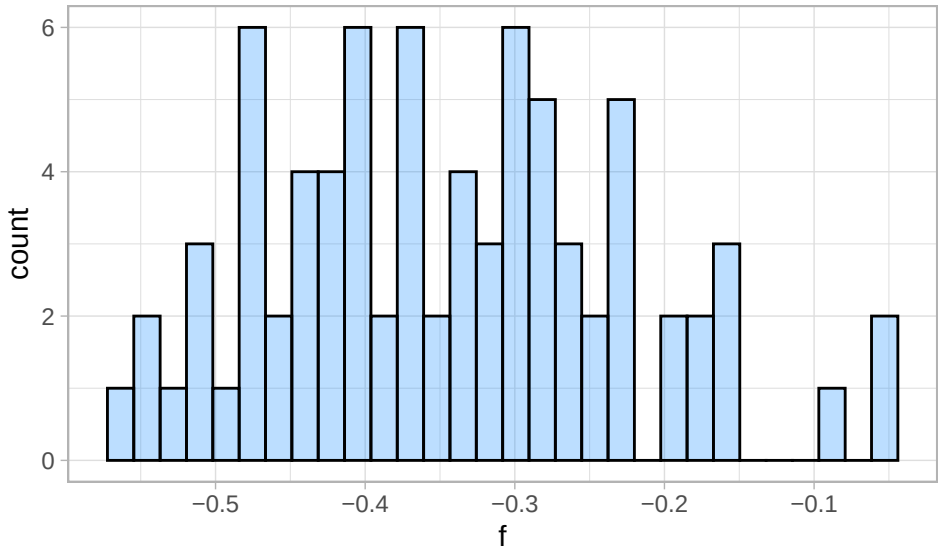
#6. Proportion of missing data per individual




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3237  0.4016  0.4247  0.4399  0.4605  0.9789
```

The proportion of missing data per individual is generally quite high, with a mean of 0.44 (44% missing data).

#7. Heterozygosity and inbreeding coefficient per individual




```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.56827 -0.43216 -0.34946 -0.34431 -0.26893 -0.05766
```

Strongly negative values of Inbreeding Coefficient (F) could mean there are too many heterozygotes and suggest a site with bad mapping/high levels of allelic dropout. A value of 0 suggests the site is in Hardy-Weinberg Equilibrium. Highly positive F values are indicative of DNA contamination.