03_check_snp_stats

2024-01-03

Prep:

```
rm(list=ls())

knitr::opts_knit$set(root.dir = "~/Dropbox/killer_whale_genomics/snps2/02-snps_stats")

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)
```

Import unfiltered vcf information (.info file from bcftools):

```
vcfInfo <- read_table2("killerwhale2_snps.ID.info")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `#` = col_character(),
##   `[1]CHROM` = col_double(),
##   `[2]POS` = col_character(),
##   `[3]REF` = col_character(),
##   `[4]ALT` = col_double(),
##   `[5]QUAL` = col_double(),
##   `[6]MQ` = col_double(),
##   `[7]QD` = col_double(),
##   `[8]TC` = col_character(),
##   `[9]FR` = col_character()
## )
```

```
colnames(vcfInfo) <- c("CHROM", "POS", "REF", "ALT", "QUAL", "MQ", "QD", "TC", "FR")
```

Get stats:

```
# allele freq, missingness, hwe from vcftools
var_frq <- read_delim("killerwhale2_snps.ID.frq", delim="\t",col_names=c("CHR", "POS", "N_ALLELES", "N_
```

```
## Rows: 10126268 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): CHR
## dbl (5): POS, N_ALLELES, N_CHR, A1, A2
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
var_miss <- read_delim("killerwhale2_snps.ID.lmiss", delim="\t")
```

```
## Rows: 10202871 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): CHR
## dbl (5): POS, N_DATA, N_GENOTYPE_FILTERED, N_MISS, F_MISS
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ind_miss <- read_delim("killerwhale2_snps.ID.imiss", delim="\t")
```

```
## Rows: 87 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): INDV
## dbl (4): N_DATA, N_GENOTYPES_FILTERED, N_MISS, F_MISS
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
hwe <- read_delim("killerwhale2_snps.ID.hwe", delim="\t", col_names=c("CHR", "POS", "OBS.HOM1.HET.HOM2"
```

```
## Rows: 10126268 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (3): CHR, OBS.HOM1.HET.HOM2, E.HOM1.HET.HOM2
## dbl (5): POS, ChiSq_HWE, P_HWE, P_HET_DEFICIT, P_HET_EXCESS
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

View distributions:

```
# QUAL
min(vcfInfo$QUAL)
```

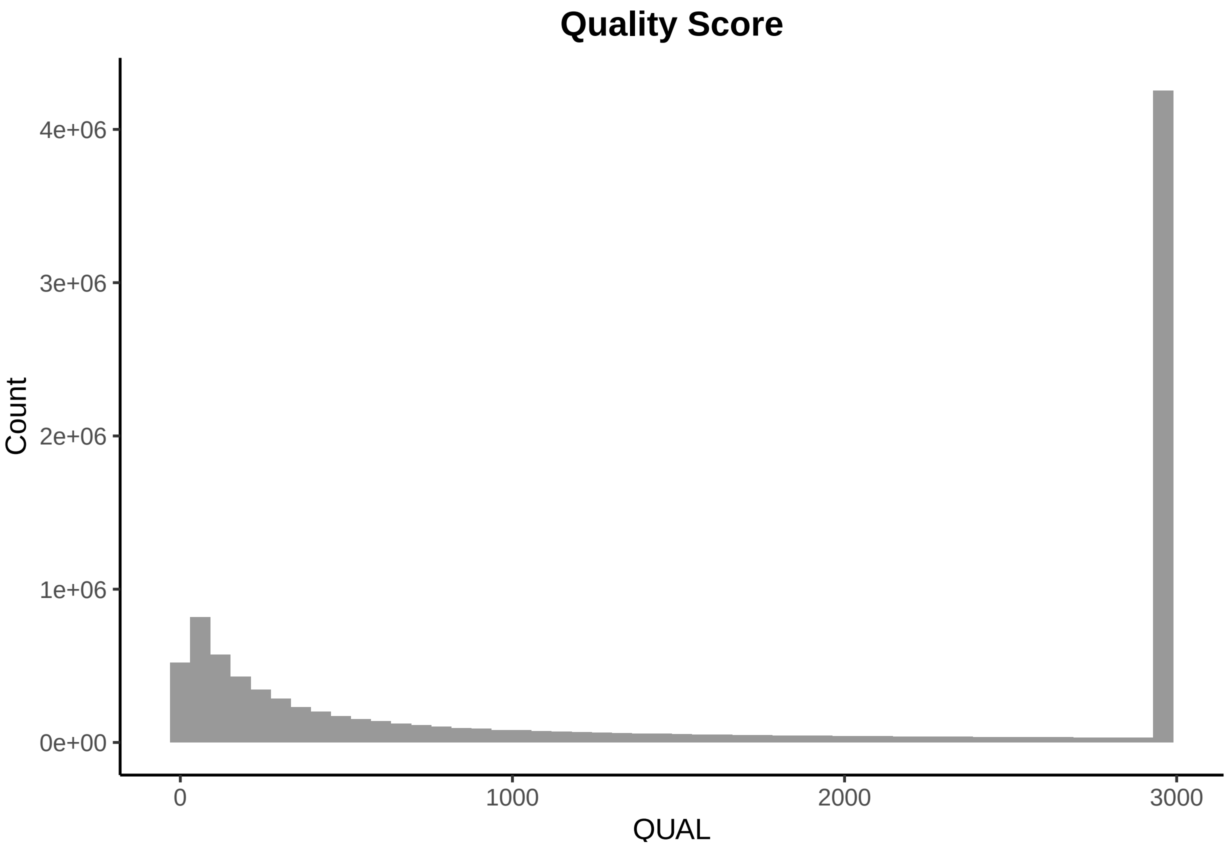
```
## [1] 5
```

```
max(vcfInfo$QUAL)
```

```
## [1] 2965
```

```
qual <- ggplot(vcfInfo, aes(x=QUAL))+
  geom_histogram(fill="gray60", bins=50)+
  theme_classic()+
  ggtitle("Quality Score")+
  theme(plot.title=element_text(hjust=0.5, face="bold"))+
 # geom_vline(xintercept=20, col="red")+
  ylab("Count")
```

```
qual
```



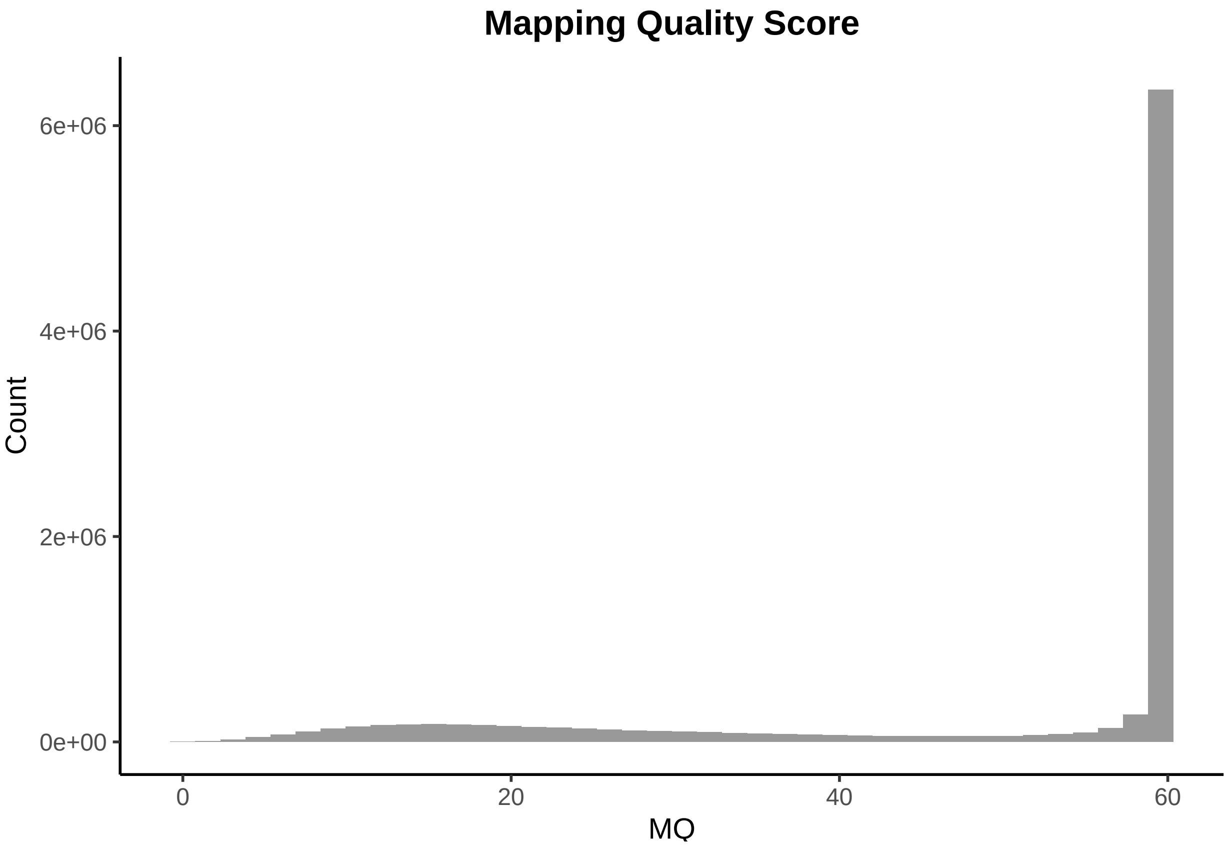
```
# MQ
min(vcfInfo$MQ, na.rm=TRUE)
```

```
## [1] 0.42
```

```
max(vcfInfo$MQ, na.rm=TRUE)
```

```
## [1] 60
```

```
mq <- ggplot(vcfInfo, aes(x=MQ))+
  geom_histogram(fill="gray60", bins=40)+
  theme_classic()+
  ggtitle("Mapping Quality Score")+
  theme(plot.title=element_text(hjust=0.5, face="bold"))+
#  geom_vline(xintercept=30, col="red")+
  ylab("Count")

mq
```



```
# QD
min(vcfInfo$QD, na.rm=TRUE)
```

```
## [1] 0.00646965
```

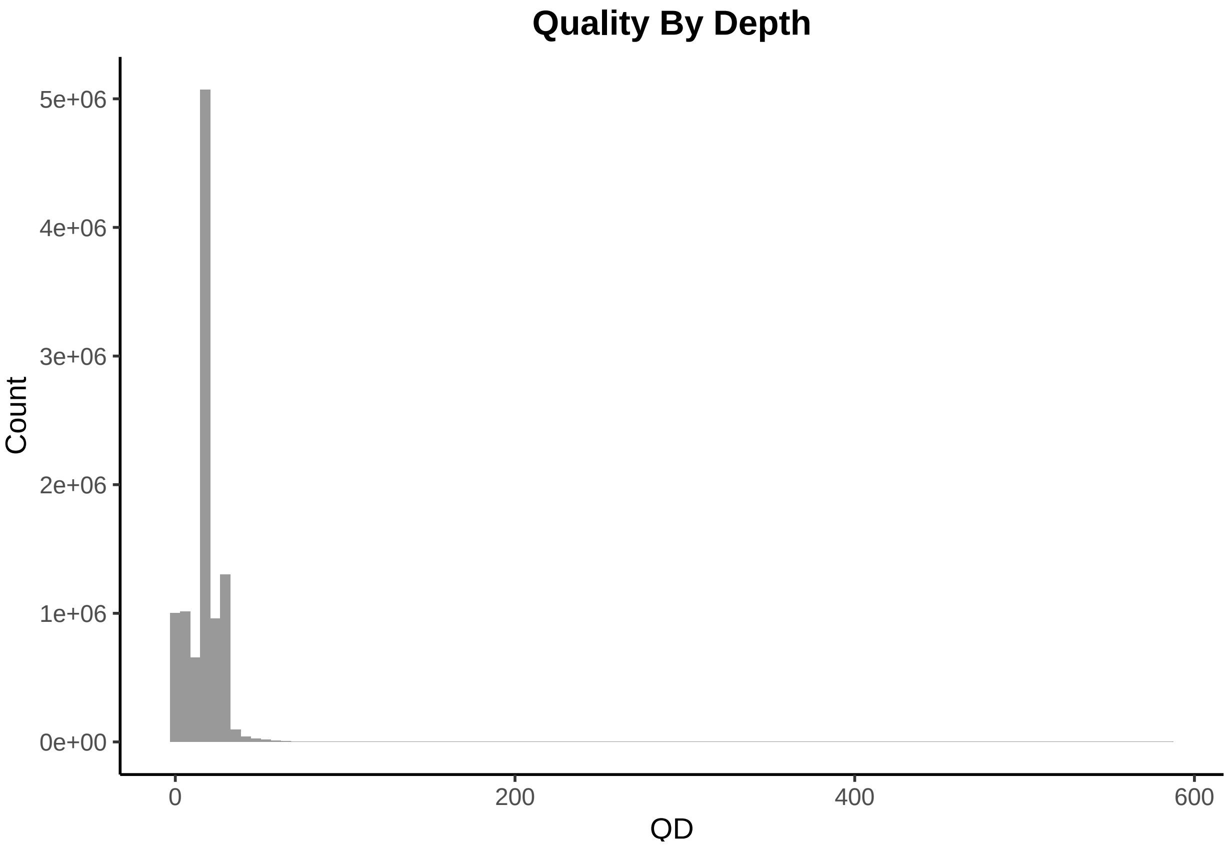
```
max(vcfInfo$QD, na.rm=TRUE)
```

```
## [1] 584.693
```

```
summary(vcfInfo$QD)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0065  13.6254  20.0000  18.0492  20.0000 584.6930
```

```
qd <- ggplot(vcfInfo, aes(x=QD))+
  geom_histogram(fill="gray60", bins=100)+
  theme_classic()+
  ggtitle("Quality By Depth")+
  theme(plot.title=element_text(hjust=0.5, face="bold"))+
 # geom_vline(xintercept=2, col="red")+
  ylab("Count")

qd
```

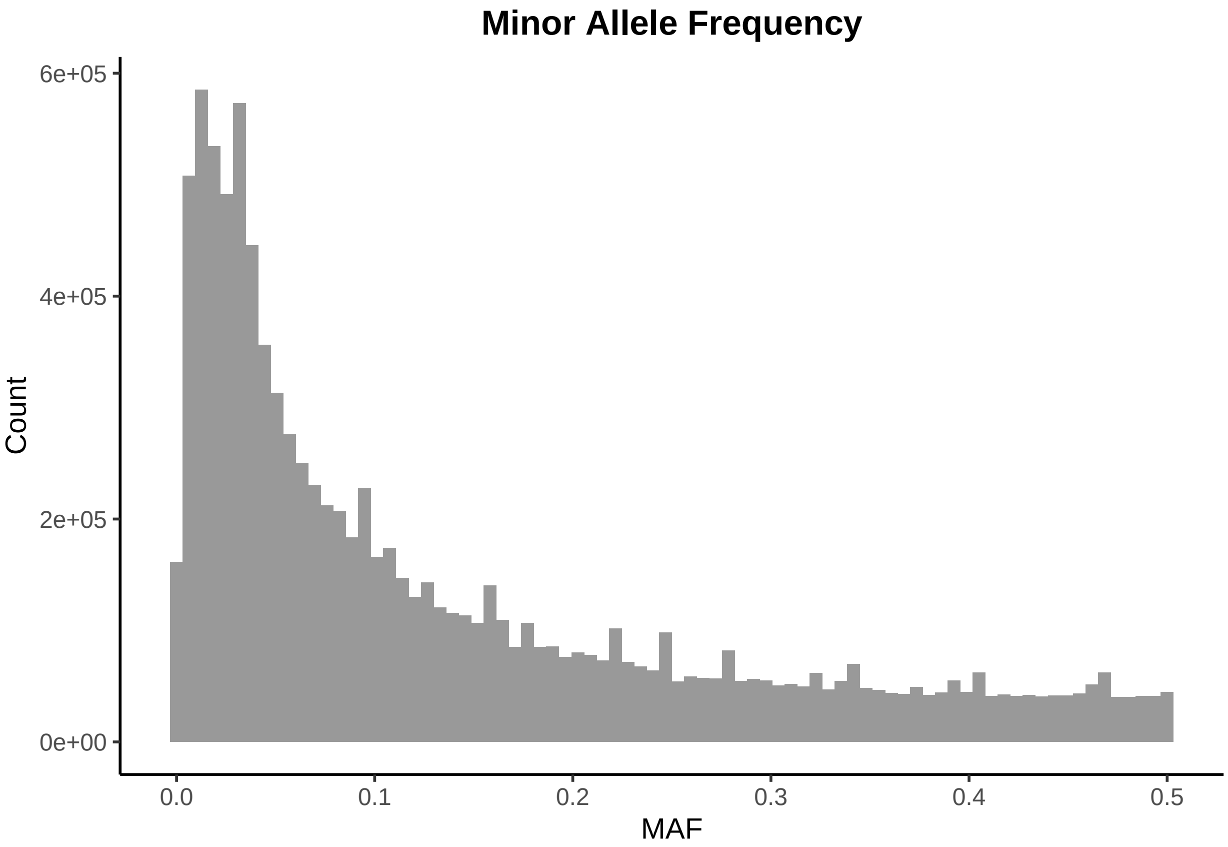


```r
# Minor allele frequency
var_frq$MAF <- var_frq %>% select(A1, A2) %>% apply(1, function(z) min(z))
summary(var_frq$MAF)
```

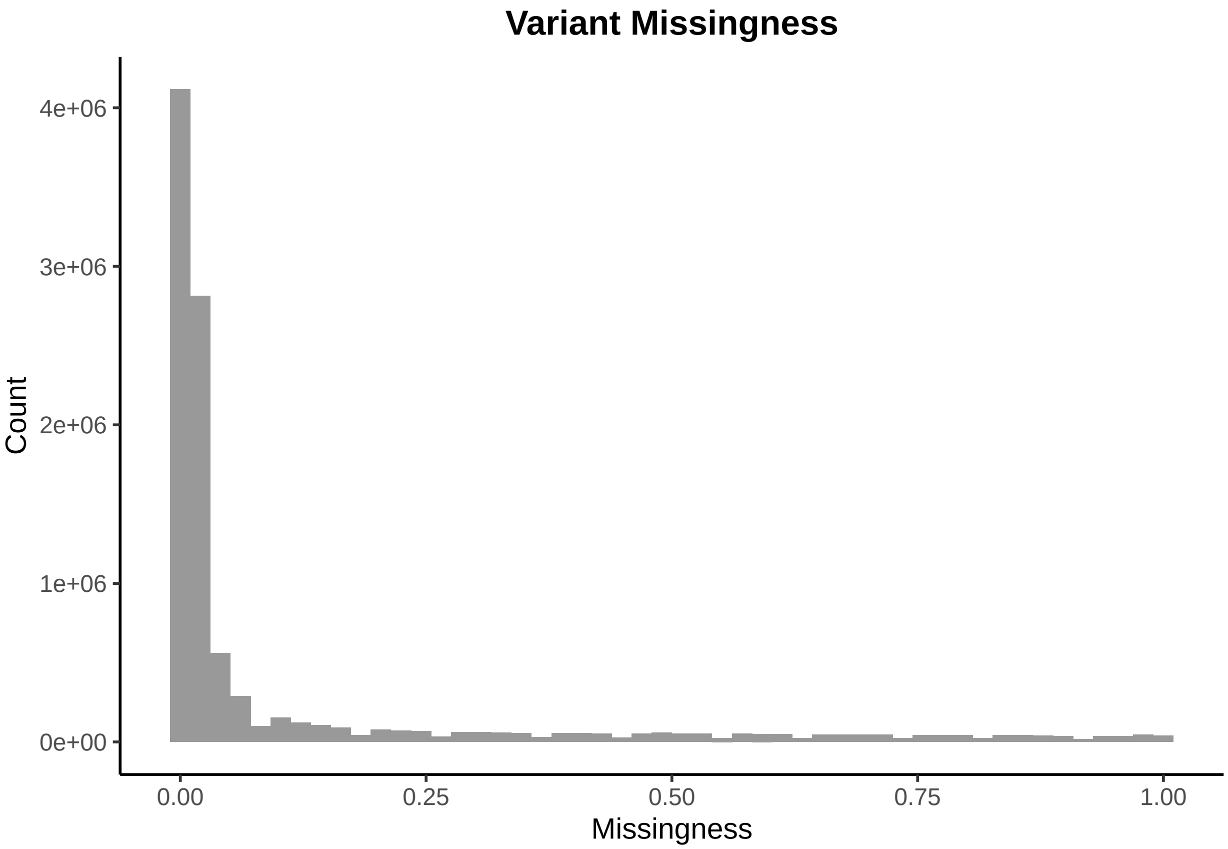
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.03    0.08    0.14    0.21    0.50   40977
```

```r
maf <- ggplot(var_frq, aes(x=MAF))+
  geom_histogram(fill="gray60", bins=80)+
  theme_classic()+
  ggtitle("Minor Allele Frequency")+
  theme(plot.title=element_text(hjust=0.5, face="bold"))+
 # geom_vline(xintercept=0.041, col="red")+
  ylab("Count")

maf
```

```
## Warning: Removed 40977 rows containing non-finite values (`stat_bin()`).
```



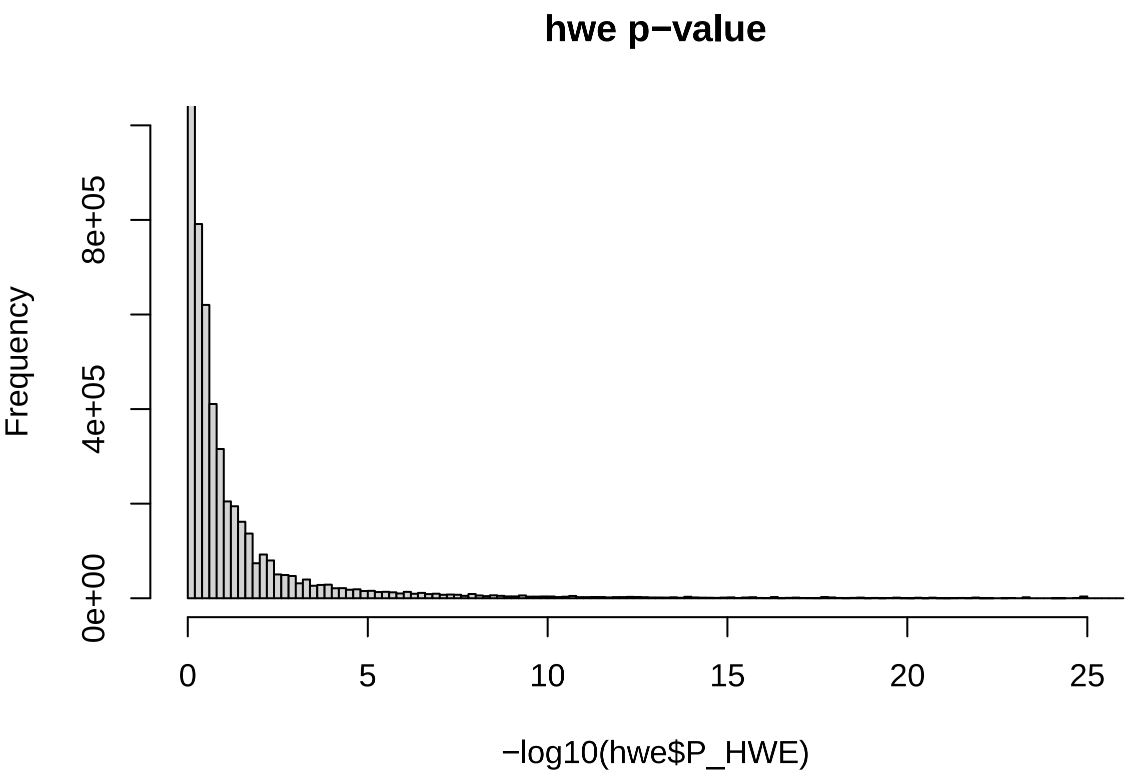
```r
# Variant missingness
variant_miss <- ggplot(var_miss, aes(x=F_MISS))+
  geom_histogram(fill="gray60", bins=50)+
  theme_classic()+
  ggtitle("Variant Missingness")+
  theme(plot.title=element_text(hjust=0.5, face="bold"))+
 # geom_vline(xintercept=0.4, col="red")+
  xlab("Missingness")+
  ylab("Count")

variant_miss
```



```
# HWE for heterozygosity
summary(-log10(hwe$P_HWE))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.6658  0.5244 25.9173
```

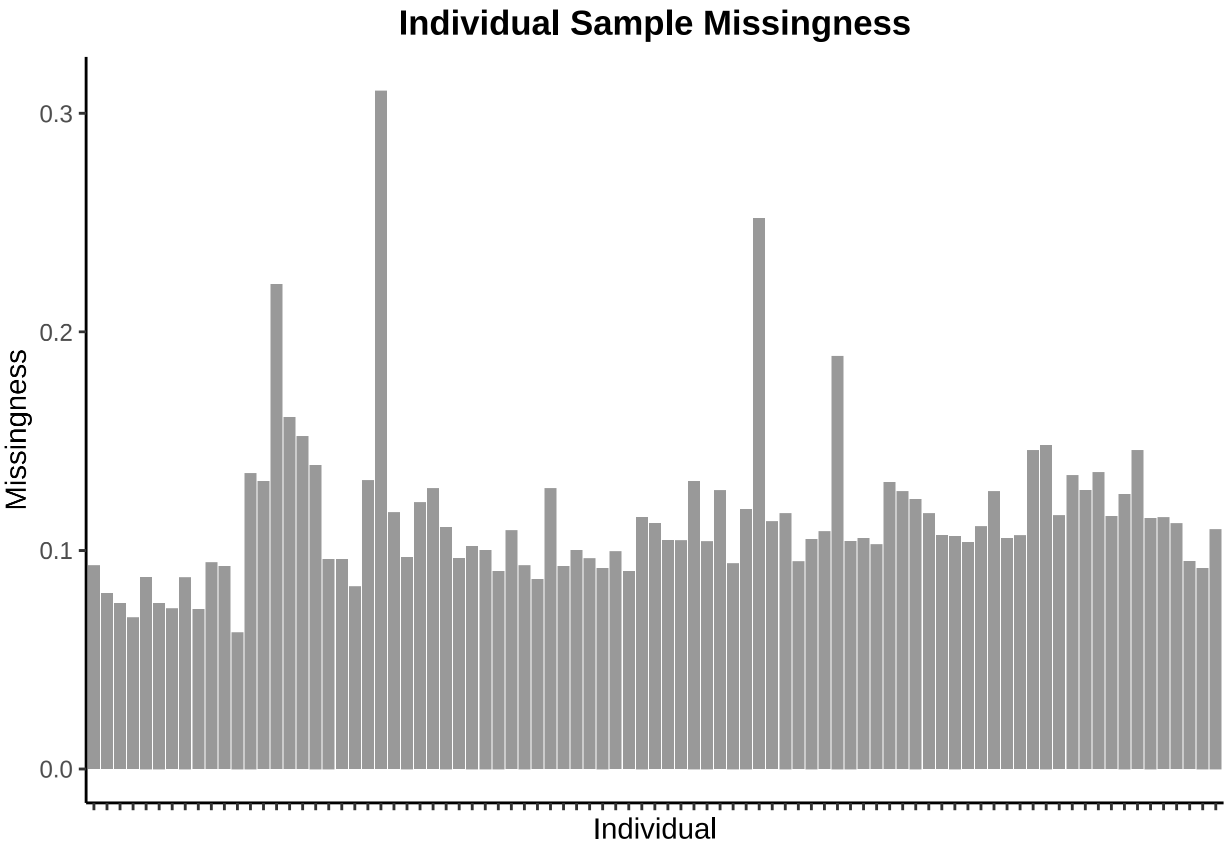
```
hist(-log10(hwe$P_HWE), breaks=100, ylim=c(0,1000000), main="hwe p-value")
```



```
# Individual missingness
summary(ind_miss$N_MISS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  638235  969532 1091756 1171346 1295926 3165625
```

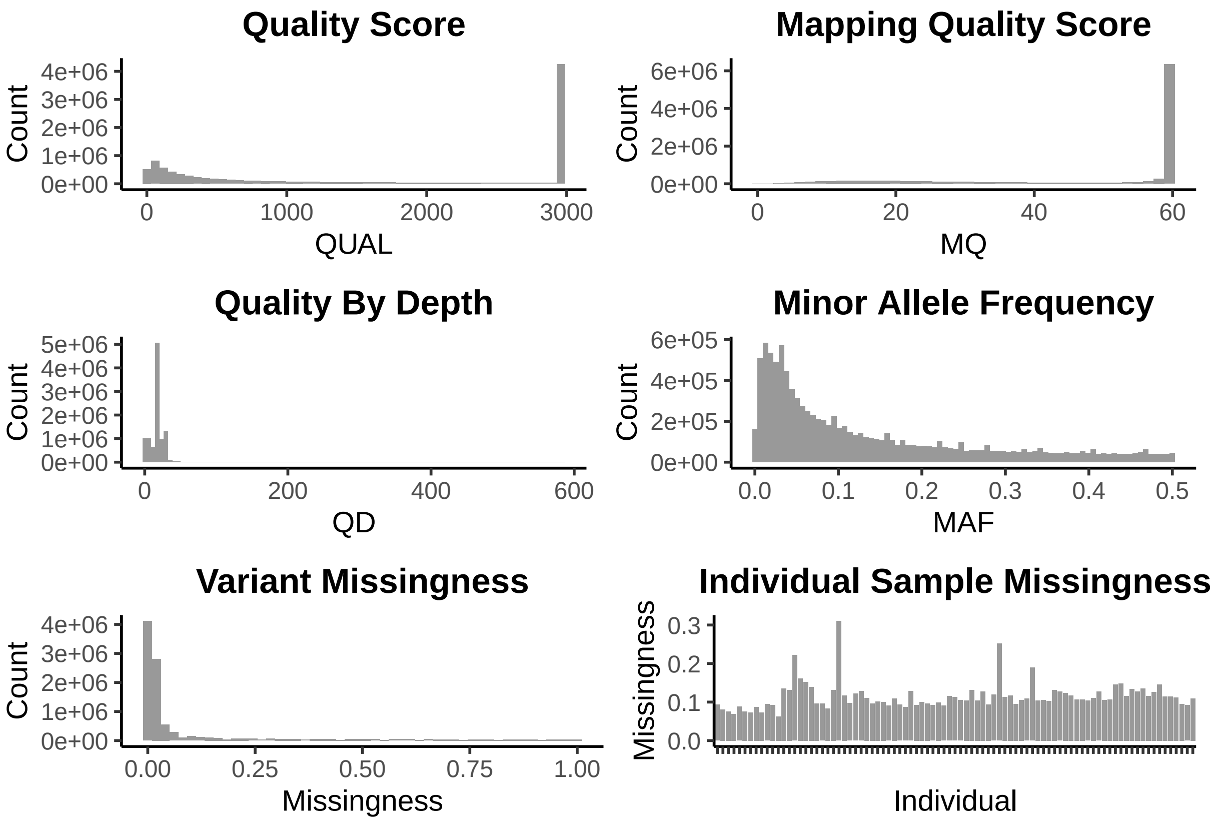
```
indiv_miss <- ggplot(data=ind_miss, aes(x=factor(INDV), y=F_MISS))+
  geom_bar(stat="identity", fill="gray60")+
  theme_classic()+
  xlab("Individual")+
  ylab("Missingness")+
  ggtitle("Individual Sample Missingness")+
  theme(plot.title=element_text(hjust=0.5, face="bold"),axis.text.x=element_blank())
  #geom_hline(yintercept=0.2, col="red")

indiv_miss
```



Plot everything into one figure:

```
library(patchwork)

(qual | mq ) / (qd | maf ) / (variant_miss | indiv_miss)
```

```
## Warning: Removed 40977 rows containing non-finite values (`stat_bin()`).
```



```
#ggsave("KW2_SNPS_prefilter.png", width = 9, height = 9, dpi = 300)
```