Filtering protocol

Aurora García-Berro

4/16/2021

We will do 4 filterings:

- We will remove individuals of bad quality, based on *missing data per individual*
- Then we will select the *NHE and SH populations*, which are the better represented ones, for a reduced dataset, and we will work from now on with both the *original and the reduced* dataset.
- *Minor allele frequency*: we will make 4 classes based on cutoffs (plus the non filtered dataset). We will thus have 4 classes for the original and 4 for the reduced, plus 1 non-filtered; that is 9 datasets in total.
- Last, we will filter by *missing data*, and will make 2 datasets (plus 1 non-filtered). Total: 2x9=18 datasets, + 1 non-filtered: 19 datasets.

First, I did a subset of my vcf:

vcfrandomsample cardui_migrdiv_all_snps.vcf -r 0.012 > cardui_migrdiv_subset.vcf

Then I loaded it into https://bmedeiros.shinyapps.io/matrix_condenser

And then I downloaded the matrix in a csv file. It has 3 columns: 1) Sample name 2) Missing data 3) Population

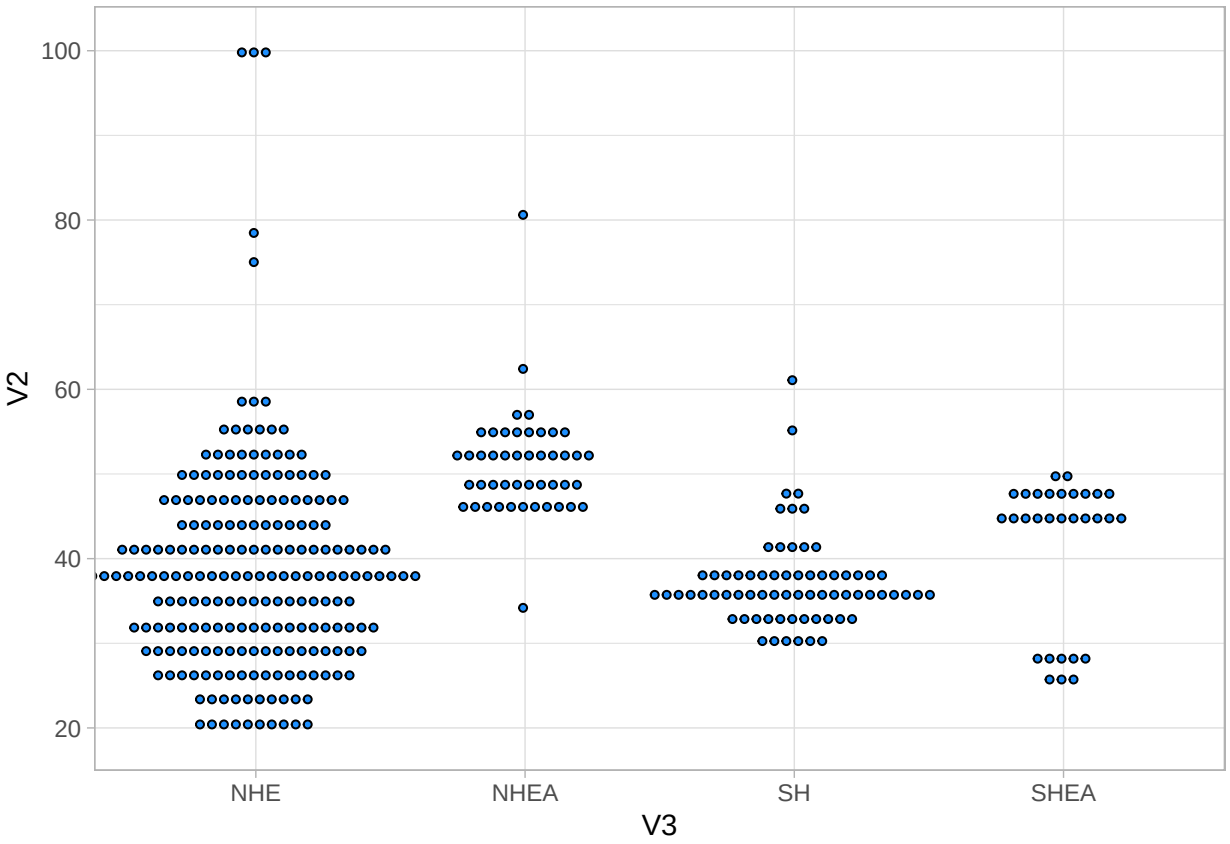
Missing data per individual

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
setwd("/Users/Aurora/Desktop/Migratory_divide_files/")
all_samples <- read.csv("all_samples.csv", header = F)
plot2<- ggplot(all_samples, aes(x=V3, y=V2)) + geom_dotplot(fill = "dodgerblue1", binaxis='y', stackdir=
plot2 + theme_light()
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



Based on the plot we decide to remove the 8 worst individuals (outliers in the top part). Or, what is equivalent, to set a threshold of 60% maximum missing data per individual. The following individuals will be removed:

Sample Missing_data
Ethiopia_14U295 62.41501
Ethiopia_14U299 80.60961
South_14B030 61.07855
Benin_14E446 75.02931
Estonia_10A746 99.67175
Israel_16E191 99.92966
Russia_16A391 99.92966
Russia_16E557 78.47597

There is no option in vcftools to filter miss data per individual. Instead we do this:

We first create a list of individuals to remove, so we store the first column containing the sample name into a file "lowDP.indv". We can feed that directly into VCFtools for filtering:

cd /proj/uppstore2017185/b2014034_nobackup/Aurora/Migratory_divide/vcfs/vcf_filtering/ vcftools –gzvcf cardui_migrdiv_all_snps.vcf.gz –remove lowDP.indv –recode –stdout | gzip -c > cardui_migrdiv_indv.vcf.gz

We could also obtain it from the vcf like this:
awk '$5 > 0.6' cardui_migrdiv.imiss | cut -f1 > lowDP6.indv

Reading the log, we see that: After filtering, kept 347 out of 355 Individuals
Outputting VCF file…
After filtering, kept 138553 out of a possible 138553 Sites

Reduced and original datasets

I need to get the samples that will be removed using the same method.

awk '$3 =="SHEA" ' all_samples.txt | cut -f1 > selectedpop.indv (30 individuals)
awk '$3 =="NHEA" ' all_samples.txt | cut -f1 >> selectedpop.indv (46 individuals)

Now I have a list of 76 individuals and will feed it to vcf like the previous point:

vcftools –gzvcf cardui_migrdiv_indv.vcf.gz –remove-indels –remove selectedpop.indv –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_NHE_SH.vcf.gz

Minor allele frequency

It's the num of times an allele appears over all indv at that site divided by total number of alleles at that site. For the V. atalanta dataset we used MAC (without dividing by total number of alleles), and used a minimum of 3.

Based on how the plot looks like the 5 classes will be:

0) all variants (no filtering) (probably will be ignored)
1) very rare variants (0.001 MAF) (probably will be ignored)
2) rare variants (0.001 MAF < 0.010)
3) low-frequency variants (0.010 MAF < 0.05)
4) common variants (MAF 0.05)

And will apply it to vcf tools, to the original dataset…

kept 193 out of a possible 138553 Sites

vcftools –gzvcf cardui_migrdiv_indv.vcf.gz –max-maf 0.001 –out 3_original_veryrare –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_veryrarev.vcf.gz

kept 27526 out of a possible 138553 Sites

vcftools –gzvcf cardui_migrdiv_indv.vcf.gz –maf 0.001 –out 3_original_rarev –max-maf 0.01 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_rarev.vcf.gz

kept 30888 out of a possible 138553 Sites

vcftools –gzvcf cardui_migrdiv_indv.vcf.gz –maf 0.01 –max-maf 0.05 –out 3_original_lowfreqv –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_lowfreqv.vcf.gz

kept 80742 out of a possible 138553 Sites

vcftools –gzvcf cardui_migrdiv_indv.vcf.gz –maf 0.05 –out 3_original_commonv –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_commonv.vcf.gz

and the reduced:

kept 8271 out of a possible 138553 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH.vcf.gz –out 3_reduced_veryrare –max-maf 0.001 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_NHE_SH_veryrarev.vcf.gz

kept 23757 out of a possible 138553 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH.vcf.gz –out 3_reduced_rarev –maf 0.001 –max-maf 0.01 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_NHE_SH_rarev.vcf.gz

kept 29339 out of a possible 138553 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH.vcf.gz –out 3_reduced_lowfreqv –maf 0.01 –max-maf 0.05 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_NHE_SH_lowfreqv.vcf.gz

kept 77931 out of a possible 138553 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH.vcf.gz –out 3_reduced_commonv –maf 0.05 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_NHE_SH_commonv.vcf.gz

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

```
## -- Attaching packages ----------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.0.0     v dplyr   1.0.0
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.3
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

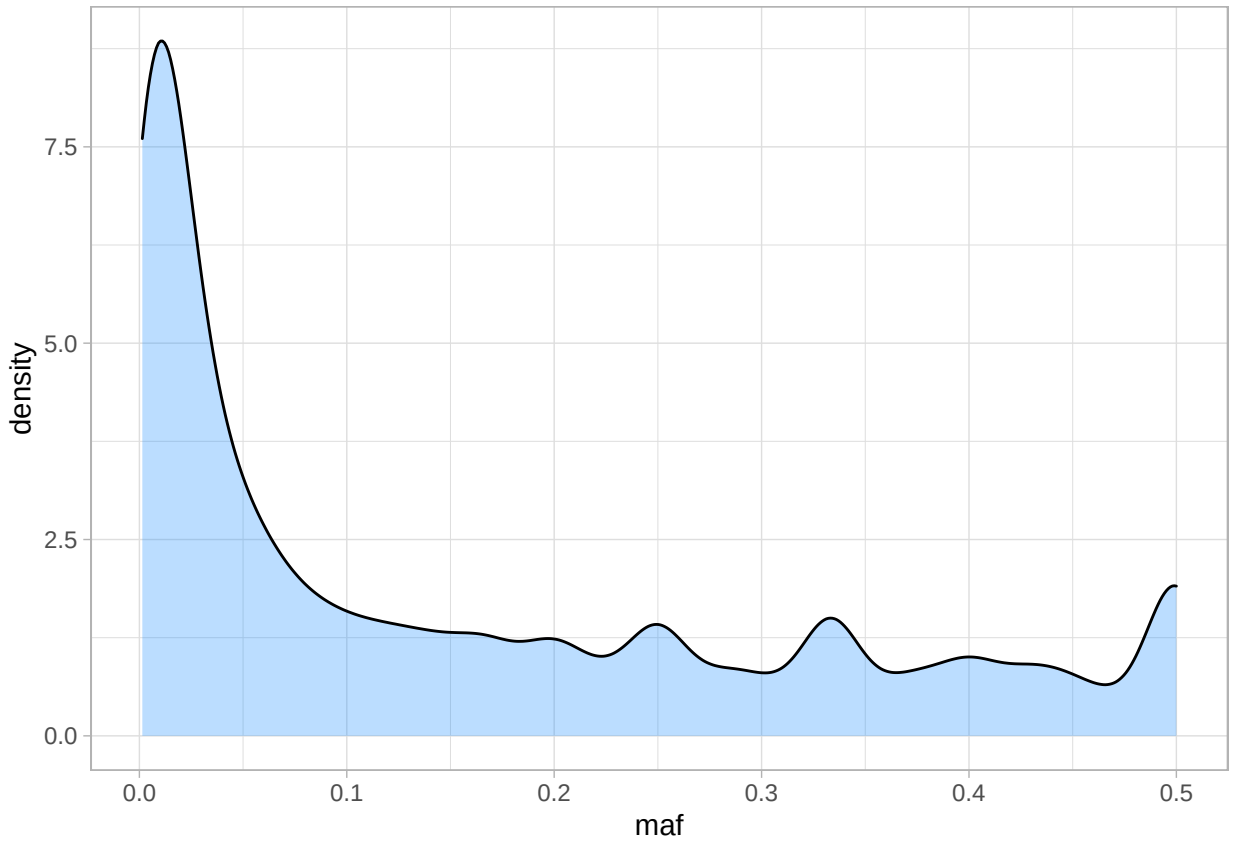
```
## Warning: package 'forcats' was built under R version 3.5.2
```

```
## -- Conflicts -------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
var_freq <- read_delim("/Users/aurora/Desktop/Migratory_divide_files/Migratory_divide/cardui_migrdiv.fr
```

```
## Parsed with column specification:
## cols(
##   chr = col_character(),
##   pos = col_double(),
##   nalleles = col_double(),
##   nchr = col_double(),
##   a1 = col_double(),
##   a2 = col_double()
## )
```

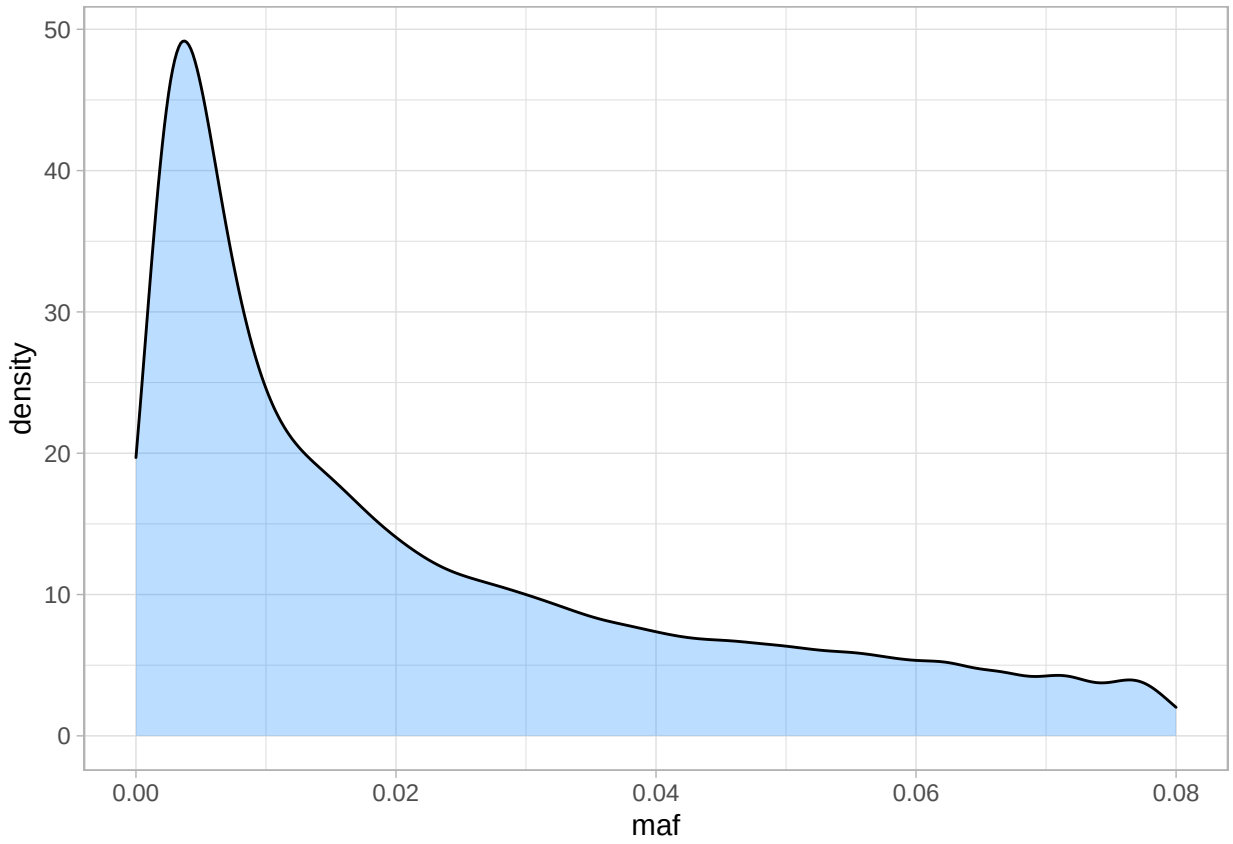
```
var_freq$maf <- var_freq %>%
  select(a1, a2) %>%
  apply(1, function(z) min(z))
freq_plot <- ggplot(var_freq, aes(maf)) + geom_density(fill = "dodgerblue1", colour = "black", alpha = 
freq_plot + theme_light()
```



If we take a closer look:

```
freq_plot + theme_light() + xlim(0, 0.08)
```

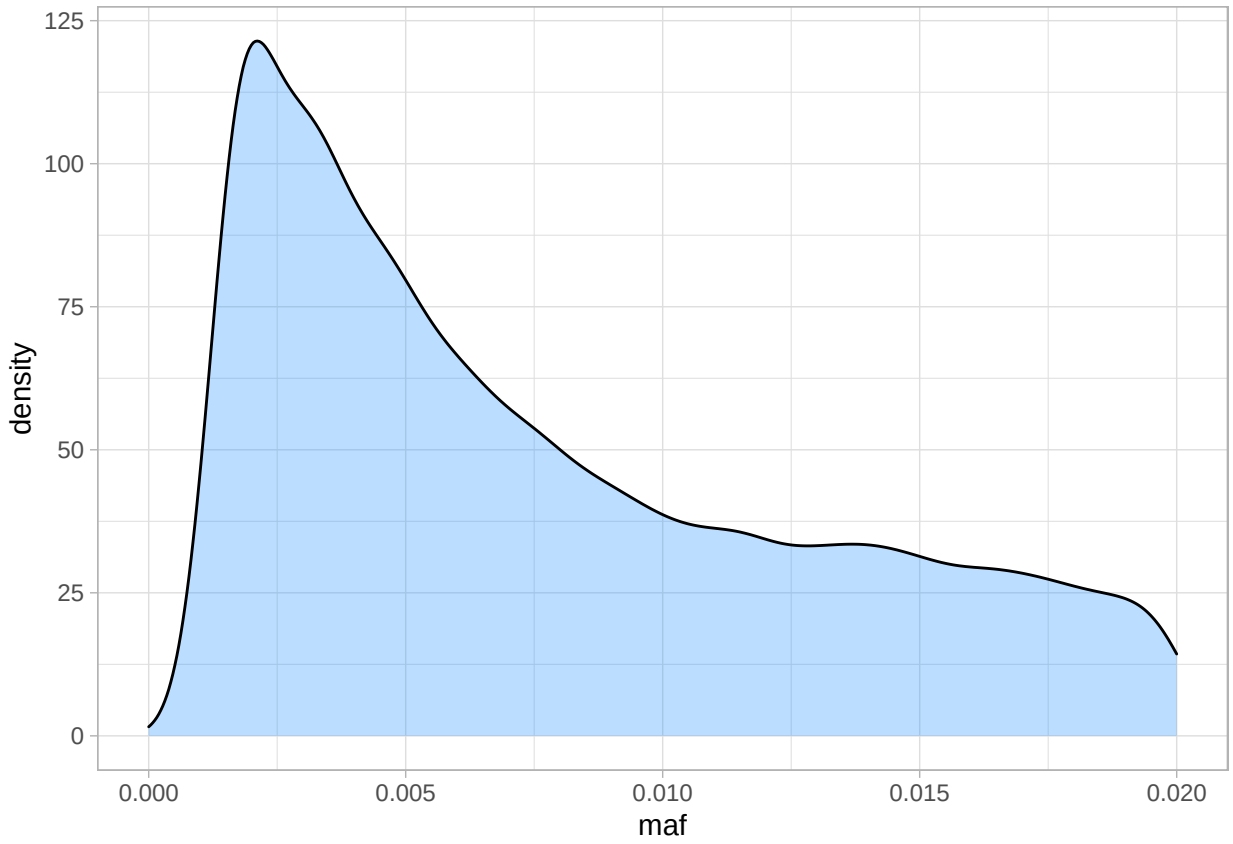
```
## Warning: Removed 70515 rows containing non-finite values (stat_density).
```



And closer:

```
freq_plot + theme_light() + xlim(0, 0.02)
```

```
## Warning: Removed 98609 rows containing non-finite values (stat_density).
```



Missing data

Minimum number of sample/locus - loci - missing data 10% 6913 55% 20% 5200 45% 30% 4243 39% 40% 3470 33% 50% 2693 26% 60% 2051 21%

Based on this we choose 45% missing data and 25% missing data:

kept 31 out of a possible 193 Sites

vcftools –gzvcf cardui_migrdiv_indv_veryrarev.vcf.gz –out 4_original_veryrare_45md –max-missing 0.55 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_veryrarev_md45.vcf.gz

kept 14 out of a possible 193 Sites

vcftools –gzvcf cardui_migrdiv_indv_veryrarev.vcf.gz –out 4_original_veryrare_25md –max-missing 0.75 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_veryrarev_md25.vcf.gz

kept 16203 out of a possible 27526 Sites

vcftools –gzvcf cardui_migrdiv_indv_rarev.vcf.gz –out 4_original_rare_45md –max-missing 0.55 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_rarev_md45.vcf.gz

kept 9040 out of a possible 27526 Sites

vcftools –gzvcf cardui_migrdiv_indv_rarev.vcf.gz –out 4_original_rare_25md –max-missing 0.75 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_rarev_md25.vcf.gz

kept 5837 out of a possible 30888 Sites

vcftools –gzvcf cardui_migrdiv_indv_lowfreqv.vcf.gz –out 4_original_lowfreq_45md –max-missing 0.55 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_lowfreqv_md45.vcf.gz

kept 2345 out of a possible 30888 Sites

vcftools –gzvcf cardui_migrdiv_indv_lowfreqv.vcf.gz –out 4_original_lowfreq_25md –max-missing 0.75 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_lowfreqv_md25.vcf.gz

kept 1646 out of a possible 80742 Sites

vcftools –gzvcf cardui_migrdiv_indv_commonv.vcf.gz –out 4_original_common_45md –max-missing 0.55 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_lowfreqv_md45.vcf.gz

kept 604 out of a possible 80742 Sites

vcftools –gzvcf cardui_migrdiv_indv_commonv.vcf.gz –out 4_original_common_25md –max-missing 0.75 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_lowfreqv_md25.vcf.gz

and the reduced:

kept 25657 out of a possible 138553 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH.vcf.gz –out 4_reduced_veryrare_45md –max-missing 0.55 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_veryrarev_md45.vcf.gz

kept 13014 out of a possible 138553 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH.vcf.gz –out 4_reduced_veryrare_25md –max-missing 0.75 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_veryrarev_md25.vcf.gz

kept 15534 out of a possible 23757 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH_rarev.vcf.gz –out 4_reduced_rarev_45md –max-missing 0.55 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_NHE_SH_rarev_md45.vcf.gz

kept 8692 out of a possible 23757 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH_rarev.vcf.gz –out 4_reduced_rarev_25md –max-missing 0.75 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_NHE_SH_rarev_md25.vcf.gz

kept 6676 out of a possible 29339 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH_lowfreqv.vcf.gz –out 4_reduced_lowfreqv_45md –max-missing 0.55 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_NHE_SH_lowfreqv_md45.vcf.gz

kept 2665 out of a possible 29339 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH_lowfreqv.vcf.gz –out 4_reduced_lowfreqv_25md –max-missing 0.75 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_NHE_SH_lowfreqv_md25.vcf.gz

kept 660 out of a possible 77931 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH_commonv.vcf.gz –out 4_reduced_commonv_45md –max-missing 0.55 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_NHE_SH_commonv_md45.vcf.gz

kept 1821 out of a possible 77931 Sites

vcftools –gzvcf cardui_migrdiv_indv_NHE_SH_commonv.vcf.gz –out 4_reduced_commonv_25md –max-missing 0.75 –recode –recode-INFO-all –stdout | gzip -c > cardui_migrdiv_indv_NHE_SH_commonv_md25.vcf.gz