BSG Lab 7

Sara Montese

14 November 2023

1 Part 1

1. In the database there are 20649 genetic variants. 0.19865178867324% of the data is missing.
2. Percentage of monomorphic variants: 11.45818%
   Number of monomorphic variants excluded: 2366
   Number of remaining variants: 18283
3. Genotype counts for rs8138488_C are shown in Table 1.

Table 1:

	Count	Proportion
B/B	41	0.4019608
B/A	47	0.4607843
A/A	14	0.1372549

Minor Allele Count (MAC) for rs8138488_C: 75
Minor Allele Frequency (MAF) for rs8138488_C: 0.3676471

4. From the histogram 1, it is shown a non-uniform distribution of minor allele frequencies (MAF). The distribution appears to be skewed towards lower MAF values, resembling a exponential diminishing pattern.
   Percentage of markers with MAF below 0.05: 14.22633 %
   Percentage of markers with MAF below 0.01: 4.731171 %
   Observed Pattern: The non-uniform distribution with high values for low MAFs and an exponential diminishing pattern suggests a population with a prevalence of rare alleles.

Figure 1:
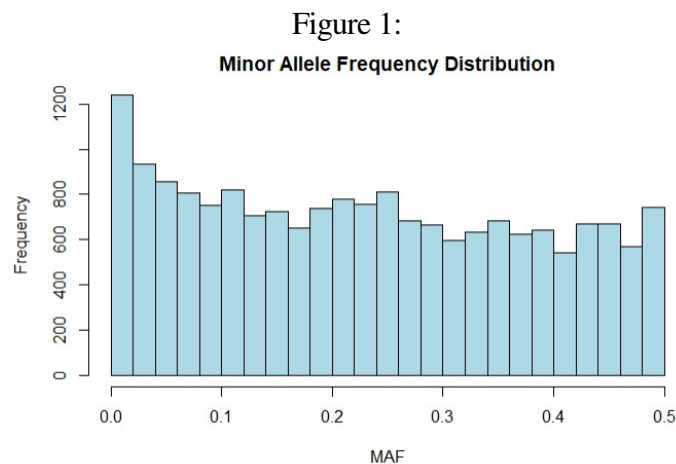
5. The histogram is shown in Fig. 2. The range of variation for heterozygosity theoretically falls between 0 (no heterozygosity, all individuals are homozygous) and 0.5 (maximum heterozygosity, all individuals are heterozygous). So, the theoretical range of variation for H0: [0, 0.5].
6. The histogram is shown in Fig. 3. Theoretical range of variation for He: [0, 1].
   Average of He: 0.3115841.

2 Part 2

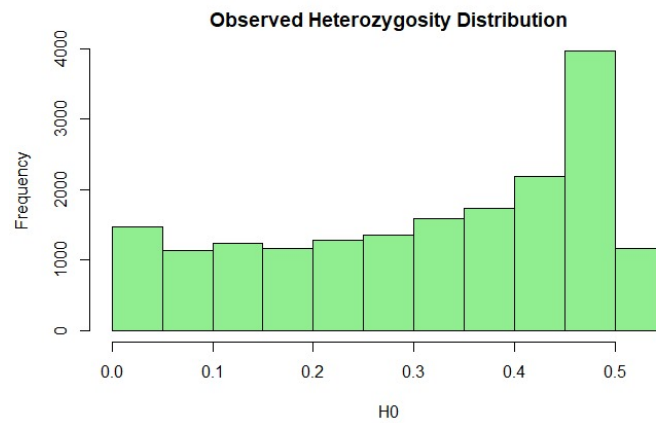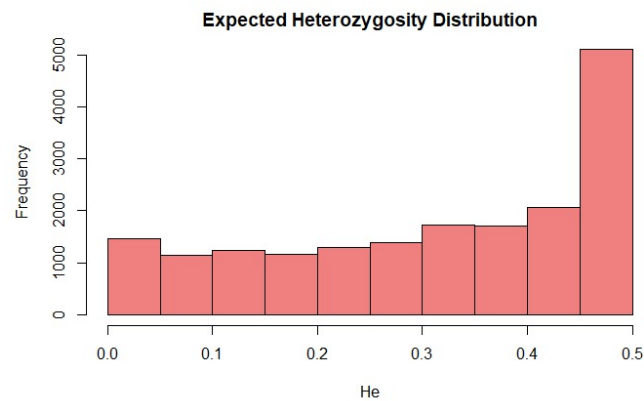1. Number of individuals: 361
   Number of STRs: 29

Figure 2:

Observed Heterozygosity Distribution

Figure 3:

Expected Heterozygosity Distribution

Table 2:

Mean	SD	Median	Min	Max
11.862069	6.226236	10.000000	6.000000	39.000000

2. Basic descriptive statistics of the number of alleles (mean, standard deviation, median, minimum, maximum) in Table 2

Number of alleles for each STR in the database is shown in Fig. 4.

Figure 4:

```
CSF1PO D10S1248 D12S391 D13S317 D16S539 D18S51 D19S433 D1S1656 D21S11
     7         9      16        8        7     15      15      15      16
D22S1045 D2S1338 D2S441 D3S1358 D5S818 D6S1043 D7S820 D8S1179 F13A01 F13B
       8      12     11        9        9      14      9      10      12    6
FESFPS FGA LPL Penta_C Penta_D Penta_E SE33 TH01 TPOX vWA
     7  14   8      10      13      19   39    8    8  10
```

Function that determines the number of alleles for a STR:

```
calculate_set_lengths <- function(dataset) {
  column_names <- colnames(dataset)
  set_lengths <- list()

  for (i in seq(2, ncol(dataset), by = 2) - 1) {
      # Calculate↩
          length of set of unique elements for each pair of columns
      str <-   unlist(strsplit(column_names[i], "-"))[1]
```

```
    set_lengths↩
        [str] <- length(unique(c(dataset[, i], dataset[, i+1])))
    }

    return(set_lengths)
}
```
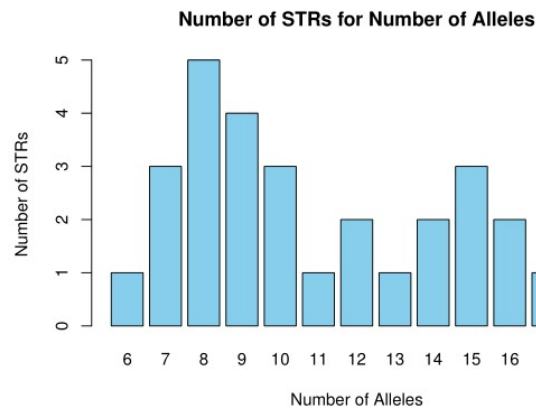
3. The table with the number of STRs for a given number of alleles is shown in Table 3

Table 3:

N. Alleles	6	7	8	9	10	11	12	13	14	15	16	19	39	
N. STRs	1	3	5	4	3	1	2	1	2	3	2	1	1	height

The barplot of the number STRs in each category is shown in Fig. 5. The most common number of alleles

Figure 5:

Number of STRs for Number of Alleles
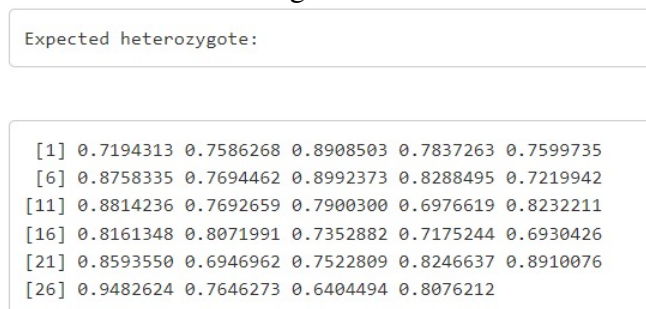
for an STR is 8.

4. The expected heterozygosity for each STR is shown in Fig. 6. The histogram of the expected heterozygosity over

Figure 6:

```
Expected heterozygote:


 [1] 0.7194313 0.7586268 0.8908503 0.7837263 0.7599735
 [6] 0.8758335 0.7694462 0.8992373 0.8288495 0.7219942
[11] 0.8814236 0.7692659 0.7900300 0.6976619 0.8232211
[16] 0.8161348 0.8071991 0.7352882 0.7175244 0.6930426
[21] 0.8593550 0.6946962 0.7522809 0.8246637 0.8910076
[26] 0.9482624 0.7646273 0.6404494 0.8076212
```

all STRS is shown in Fig. 7.
The average expected heterozygosity over all STRs is 0.7904043.

5. The observed heterozygosity for each STR is shown in Fig. 8. The plot of observed against expected heterozygosity, using all STRs, is shown in Fig. 9.Observations: From the plot and the values it can be seen that the observed heterozygosity is what would be expected, resulting in a perfect linear relationship between observed and expected heterozygosity. This suggests a population that is in genetic equilibrium with no deviations from the expected patterns of heterozygosity.

6. The main difference I noticed is that SNPs are typically bi-allelic, meaning there are only two possible alleles at a given locus. As a result, the observed heterozygosity for SNPs can only range from 0 to 0.5, differently from the one for STRs.
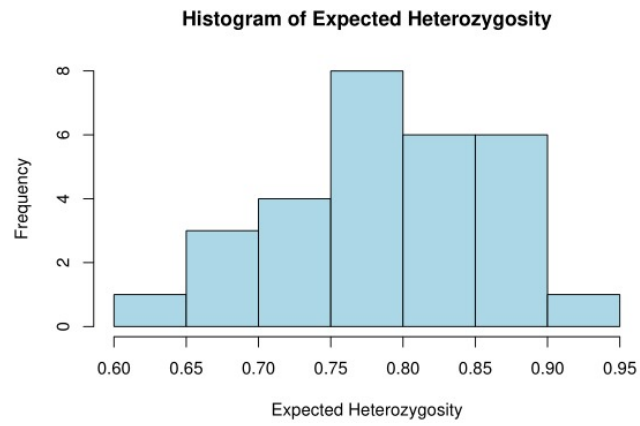
Figure 7:

Histogram of Expected Heterozygosity

Figure 8:

```
Observed heterozygote:
```

```
 [1] 0.7204291 0.7596790 0.8920859 0.7848133 0.7610275
 [6] 0.8770483 0.7705134 0.9004845 0.8299991 0.7229955
[11] 0.8826461 0.7703328 0.7911257 0.6986296 0.8243629
[16] 0.8172667 0.8083187 0.7363081 0.7185196 0.6940038
[21] 0.8605469 0.6956597 0.7533243 0.8258075 0.8922434
[26] 0.9495776 0.7656879 0.6413376 0.8087413
```

Figure 9:

Observed vs Expected Heterozygosity