Bioinformatics and Statistical Genetic

Descriptive analysis of genetic markers

Leonardo Ortoleva & Egon Ferri

Contents


```
require(genetics, quietly = T)
require(viridis, quietly = T)
require(knitr, quietly = T)
require(purrr)
```

1 SNP dataset

2 Load and clean the data

```
data<- read.table('CHDCHR22RAW.raw', header = TRUE)
```

```
data[c(1, 3, 4, 6)] <- list(NULL)
```

```
kable(data[1:10,1:5])
```

IID	SEX	rs11089128_G	rs7288972_C	rs2032141_A
NA17970	2	1	1	0
NA17977	2	0	1	0
NA17981	2	0	1	0
NA17993	2	0	0	0
NA18101	2	1	0	0
NA18105	2	0	1	0
NA18109	2	0	0	0
NA18129	2	1	0	0
NA18135	2	0	0	0
NA18139	2	1	1	0

3 How many variants are there in this database? What percentage of the data is missing? How many individuals in the database are males and how many are females?

```
## [1] "The number of variants is 16393"
```

```
## [1] "Percentage of the data is missing is 0%"
```

```
data$SEX[data$SEX==1]<- 'M'
data$SEX[data$SEX==2]<- 'F'
data[data==0]<-'AA'
data[data==1]<-'AB'
data[data==2]<-'BB'
```

```
kable(table(data$SEX))
```

Var1	Freq
F	59
M	50

4 Calculate the percentage of monomorphic variants. Exclude all monomorphics from the database for all posterior computations of the practical. How many variants do remain in your database?

```
is_mono<- function(snp){
  res<- summary(genotype(snp, sep=''))
  return(max(res$genotype.freq)==109)
}
```

```
onlygeno<- data
onlygeno[c(1, 2)] <- list(NULL)
```

```
y <- apply(onlygeno, 2, is_mono)
```

```
## [1] "Monomorphic percentage is 19.53%"
```

```
monos<-y[y==T]
data_nomo<- data
data_nomo[names(monos)]<- list(NULL)
onlygeno[names(monos)]<-list(NULL)
```

```
## [1] "There are 13192 variants still in the database"
```

5 Report the genotype counts and the minor allele count of polymorphism rs3729688_G, and calculate the MAF of this variant.

```
res<-summary(genotype(data$rs3729688_G, sep=''))
res
```

```
##
## Number of samples typed: 109 (100%)
##
## Allele Frequency: (2 alleles)
##   Count Proportion
## A   119       0.55
## B    99       0.45
##
##
## Genotype Frequency:
##     Count Proportion
## A/A    29       0.27
## A/B    61       0.56
## B/B    19       0.17
##
## Heterozygosity (Hu)  = 0.4980764
## Poly. Inf. Content   = 0.3728869
```

```
MAF<-min(res$allele.freq)
MAC<-min(res$allele.freq[,1])
```

```
## A/A A/B B/B
##  29  61  19
```

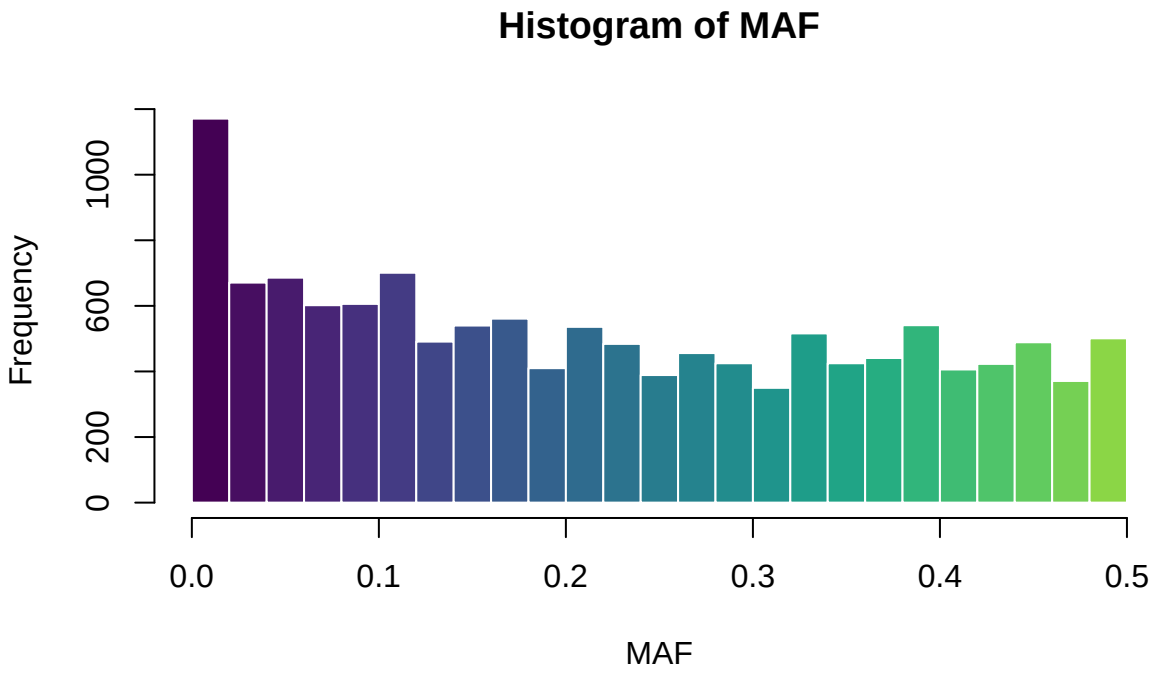
```
## [1] "Minor allele count is 99, minimum allele frequency is 0.454"
```

6 Compute the minor allele frequencies (MAF) for all markers, and make a histogram of it. Does the MAF follow a uniform distribution? What percentage of the markers have a MAF below 0.05? And below 0.01? Can you explain the observed pattern?

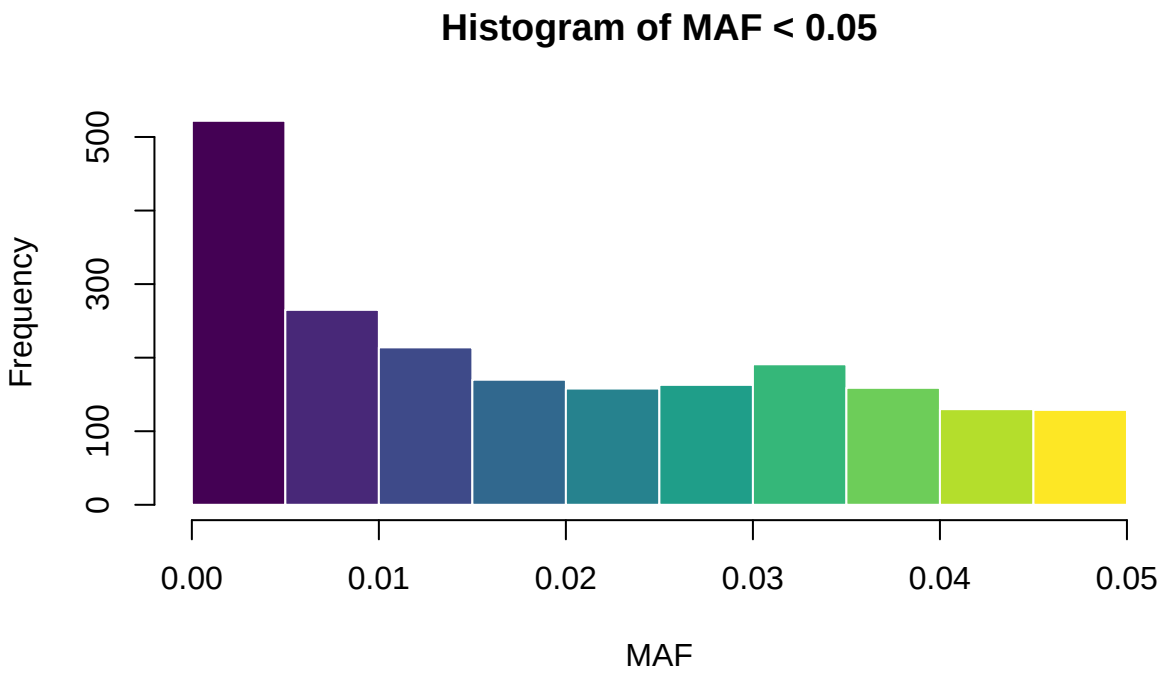
```
MAF<- function(snp){
  res<- summary(genotype(snp, sep=''))
  return(min(res$allele.freq))
}
```

```r
mafs<- apply(onlygeno, 2, MAF)
```

```r
hist(mafs, breaks = 30, col=viridis(30), border='white',
     main=title("Histogram of MAF"),
     xlab = "MAF")
```




```r
mafs_small<-mafs[mafs<0.05]
hist(mafs_small, breaks = 10, col=viridis(10), border='white',
     main = title("Histogram of MAF < 0.05"),
     xlab = "MAF")
```




```
mafs_small_perc=length(mafs[mafs<0.05])/length(mafs)*100
mafs_very_small_perc=length(mafs[mafs<0.01])/length(mafs)*100
```

```
## [1] "15.926% of the markers have a MAF below 0.05, 5.966% below 0.01"
```

We have a lot of very small minor allele frequencies, because probably there are a lot of snippets where variation are very rare, so rare that we can hardly define the genotype as polymorphic (the term polymorphism is sometimes reserved for marker where the most common allele has a frequency below 99%). So, the MAF doesn't follow an uniform distribution because there are a lot SNPs which are almost monomorphic (with a MAF near 0).

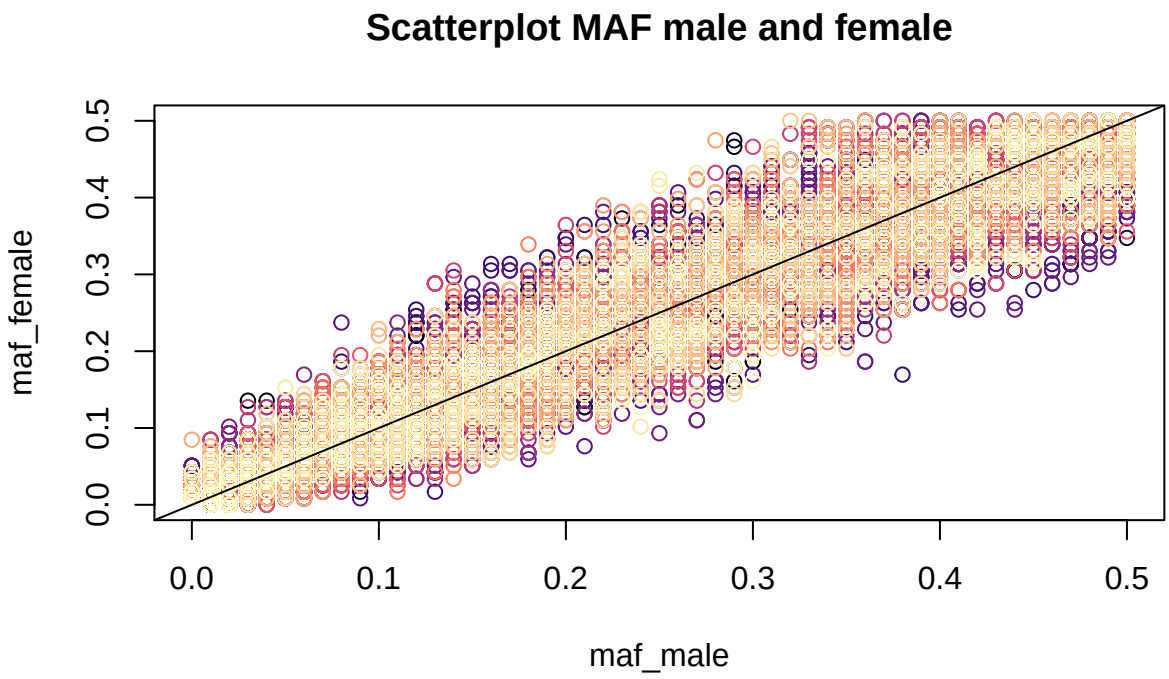
7 Calculate the minor allele frequency for males and for females and present a scatterplot of these variables. What do you observe? Calculate and report their correlation coefficient

```
onlymales<- subset(data_nomo, SEX =='M')
onlyfemales<- subset(data_nomo, SEX =='F')

onlymales[c(1,2)] <- list(NULL)
onlyfemales[c(1,2)] <- list(NULL)
```

```
maf_male=apply(onlymales, 2, MAF)
maf_female=apply(onlyfemales, 2, MAF)
```

```
maf_male[maf_male==1]<-0
maf_female[maf_female==1]<-0
plot(maf_male, maf_female, col=magma(13192),
     main = title("Scatterplot MAF male and female"))
abline(a=0, b=1)
```



In this scatterplot we can see that the gender doesn't influence so much the final MAF, in fact they are strictly correlated and there aren't no systematic differences.

```
correlation<-cor(maf_male, maf_female)
```

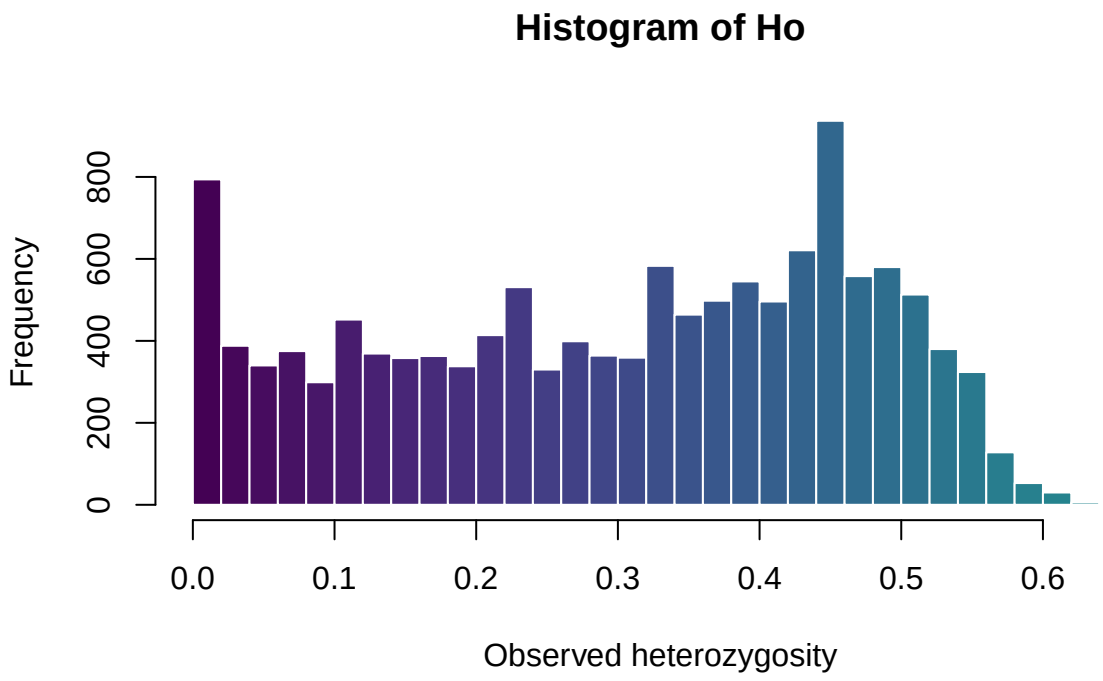
```
## [1] "The correlation coefficient is: 0.95"
```

8 Calculate the observed heterozygosity H_o, and make a histogram of it. What is, theoretically, the range of variation of this statistic?

```
observed_hetero<- function(snp){
  res<- summary(genotype(snp, sep=''))
  return(res$genotype.freq[2,2])
}

obs<- apply(onlygeno, 2, observed_hetero)
```

```
hist(obs, breaks = 30, col=viridis(68), border='white',
     xlab = 'Observed heterozygosity', main = "Histogram of Ho")
```



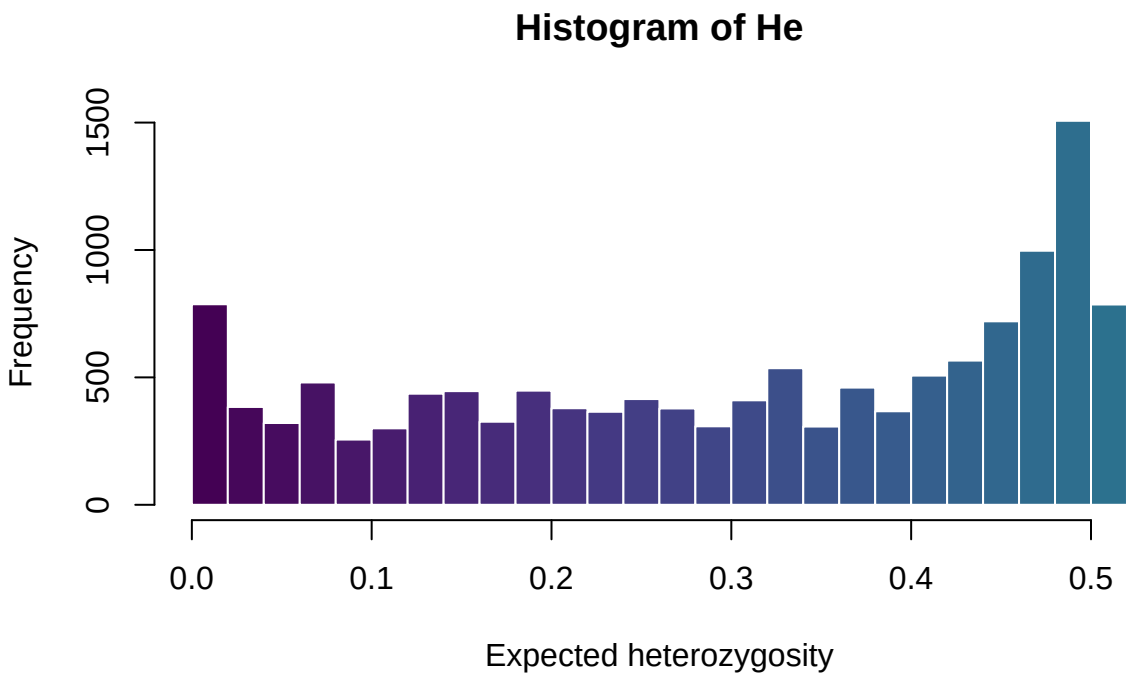
Theoretically, we should expect that the range of this statistic should be from 0 to 1, but we observe only values that are at maximium 0.6513.

9 Compute for each marker its expected heterozygosity H_e. Make a histogram of the expected heterozygosity. What is, theoretically, the range of variation of this statistic? What is the average of H_e for this database?

```
expected_hetero<- function(snp){
  res<- summary(genotype(snp, sep=''))
  return(res$Hu)
}
```

```
expect<- apply(onlygeno, 2, expected_hetero)
```

```
hist(expect, breaks = 30, col=viridis(68), border='white',
     xlab = 'Expected heterozygosity',
     main=title("Histogram of He"))
```



The expected heterozygosity for a two-allele system is described by a concave down parabola that starts and ends at zero (when we have only homozygous) and goes to a maximum at 0.5 (with no homozygous), so the range of variation is [0, 0.5]. For example with two alleles A and B, we can have H_e=0 when p_a=0 and p_b=1 or when p_a=1 and p_b=0; while we obtain H_e=0.5 when $p_a = p_b = 0.5$.

```
## [1] "The average expected heterozigosity for this database is: 0.2999"
```

2 STR dataset

1. The file FrenchStrs.dat contains genotype information (STRs) of individuals from a French population. STR data starts at the second column. Load this data into the R environment.

```
data<- read.table('FrenchSTRs.dat', header = TRUE)
```

```
kable(data[1:10,1:5])
```

	Individual	GTTTT002P_1	MFD424.TTTA003_1	GATA23G09_1	D1S1612
165	511	150	292	146	176
166	511	165	296	150	184
167	512	155	264	138	156
168	512	155	284	142	180
169	513	155	264	-9	176
170	513	155	292	-9	184
171	514	155	264	138	168
172	514	160	296	146	180
173	515	155	280	130	156
174	515	155	296	146	168

2 How many individuals and how many STRs contains the database?

```
## [1]  58 679
```

```
## [1] "We got 29 individuals, with each genotype splitted in two rows, and 678 STRs"
```

3 The value -9 indicates a missing value. Replace all missing values by NA. What percentage of the total amount of datavalues is missing?

```
onlySTR <- data.frame(data[2:length(data)])

onlySTR[onlySTR==-9] <- NA

perMis = sum(is.na(onlySTR))/(length(onlySTR)*nrow(onlySTR))*100
```

```
## [1] "Percentage of missing: 4.21%"
```

4 Write a function that determines the number of alleles for a STR. Determine the number of alleles for each STR in the database. Compute basic descriptive statistics of the number of alleles (mean, standard deviation, median, minimum, maximum).

```
nAlleles <- function(x) {
  y <- length(unique(x[!is.na(x)]))
  return(y)
}

nAll <- apply(onlySTR, 2, nAlleles)
basic_stats<- nAll
kable(summary(data.frame(basic_stats)))
```

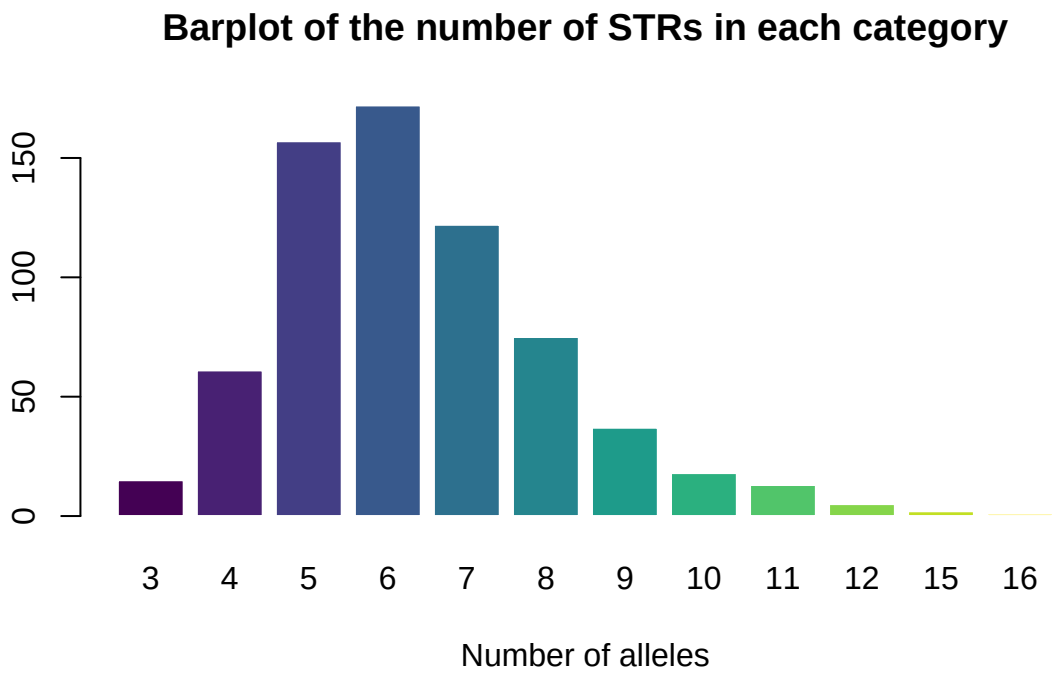
basic_stats
Min. : 3.000
1st Qu.: 5.000
Median : 6.000
Mean : 6.375
3rd Qu.: 7.000
Max. :16.000

```
print(paste0("Standard deviation: ", round(sd(nAll), 2)))
```

```
## [1] "Standard deviation: 1.82"
```

5 Make a table with the number of STRs for a given number of alleles and present a barplot of the number STRs in each category. What is the most common number of alleles for an STR?

```
barplot(table(nAll), col=viridis(12), border='white',
        main = 'Barplot of the number of STRs in each category',
        xlab = 'Number of alleles')
```



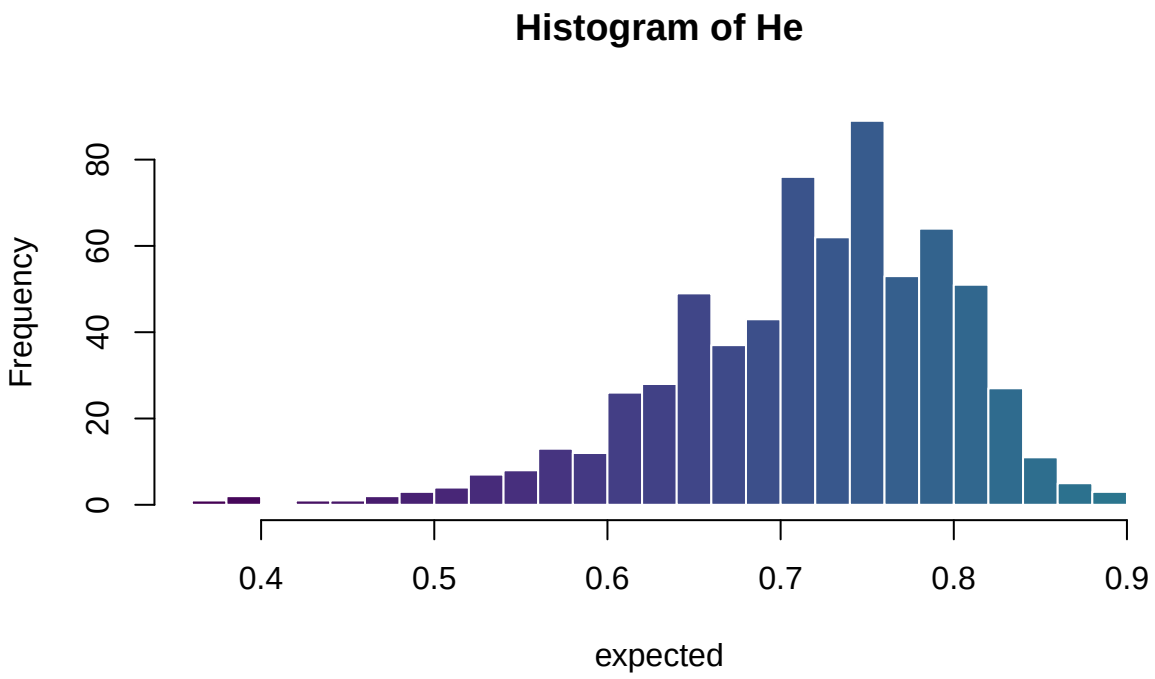
The most common number of alleles for STRs is 6.

6 Compute the expected heterozygosity for each STR. Make a histogram of the expected heterozygosity over all STRs. Compute the average expected heterozygosity over all STRs.

```
exp_eter<-function(x){
 1-sum(prop.table(table(x))^2)
}

expected<- apply(onlySTR, 2, exp_eter)
```

```
hist(expected, breaks = 19, col=viridis(68), border='white', main = 'Histogram of He')
```




```
## [1] "The average expected heterozygosity over all STRs is 0.717"
```

7 Compare the results you obtained for the SNP database with those you obtained for the STR database. What differences do you observe between these two types of genetic markers?

We have a lot of differences.

First of all, we have some differences in the datasets; the sample size of the first dataset is a lot bigger, and it does not have missing values (instead the second small dataset has a 4% of NAs).

While in the first dataset we have only 2 possible alleles, in the second one we have a lot of them (between 3 and 16, with a peak on 6), this obviously affect the computation of the expected heterozygosity, that is a lot bigger in the second one than in the first one.