BSG-MDS practical 1 Statistical Genetics

Biel Caballero and Gerard Gomez

07/11/2023, submission deadline 14/11/2023

```
library(genetics)
```

```
## Loading required package: combinat

##
## Attaching package: 'combinat'

## The following object is masked from 'package:utils':
##
##     combn

## Loading required package: gdata

##
## Attaching package: 'gdata'

## The following object is masked from 'package:stats':
##
##     nobs

## The following object is masked from 'package:utils':
##
##     object.size

## The following object is masked from 'package:base':
##
##     startsWith

## Loading required package: gtools

## Loading required package: MASS

## Loading required package: mvtnorm

##

## NOTE: THIS PACKAGE IS NOW OBSOLETE.
```

```
##

##   The R-Genetics project has developed an set of enhanced genetics

##   packages to replace 'genetics'. Please visit the project homepage

##   at http://rgenetics.org for informtion.

##


##
## Attaching package: 'genetics'

## The following objects are masked from 'package:base':
##
##     %in%, as.factor, order
```

```
library(HardyWeinberg)
```

```
## Loading required package: mice

##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind

## Loading required package: Rsolnp

## Loading required package: nnet
```

SNP

Create dataframe with only genetic information

```
data <- read.table("TSICHR22RAW.raw", header = TRUE)
genetic_data<-data[,c(7:length(data))]
```

1. How many variants are there in this database? What percentage of the data is missing?

```r
#Numver of variants
num_variants<-dim(genetic_data)[2]
num_variants
```

```
## [1] 20649
```

```r
#Percentage of missing data
total<-prod(dim(genetic_data))
missing<-sum(is.na(genetic_data))
percentage <- (missing * 100)/(total)
percentage
```

```
## [1] 0.1986518
```

There are 20649 variants and 0.2% of data is missing.

2. 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?

```r
#Percentage of monomorphic
monomorphic<-sum(colSums(genetic_data,na.rm = TRUE)==0)
percentage<-(monomorphic * 100)/(length(genetic_data))
percentage
```

```
## [1] 11.45818
```

```r
#New dataset
new_dataset<-genetic_data[,-(which(colSums(genetic_data,na.rm = TRUE)==0))]
```

The percentage of monomorphic variants is 11.46% over all the dataset. There remains 18283 variants in our dataset.

3. Report the genotype counts and the minor allele count of polymorphism rs8138488 C, and calculate the MAF of this variant.

```r
#Genotype counts
nAA<-length(which(new_dataset["rs8138488_C"]==0))
nAA
```

```
## [1] 41
```

```r
nAB<-length(which(new_dataset["rs8138488_C"]==1))
nAB
```

```
## [1] 47
```

```r
nBB<-length(which(new_dataset["rs8138488_C"]==2))
nBB
```

```
## [1] 14
```

```r
#Minor allele count
nA<-nAA*2+nAB
nA
```

```
## [1] 129
```

```r
nB<-nBB*2+nAB
nB
```

```
## [1] 75
```

```r
#MAF
pA<-(nAA+1/2*nAB)/102
pA
```

```
## [1] 0.6323529
```

```r
pB<-(nBB+1/2*nAB)/102
pB
```

```
## [1] 0.3676471
```

```r
maf<-min(pA,pB)
maf
```

```
## [1] 0.3676471
```

The genotype counts are: - nAA = 41 - nAB = 47 - nBB = 14

The minor allele counts are: - nA = 129 - nB = 75

The MAF is the minimum between the probability of allele A and allele B, in this case the MAF is the probability of allele B that is 0.37

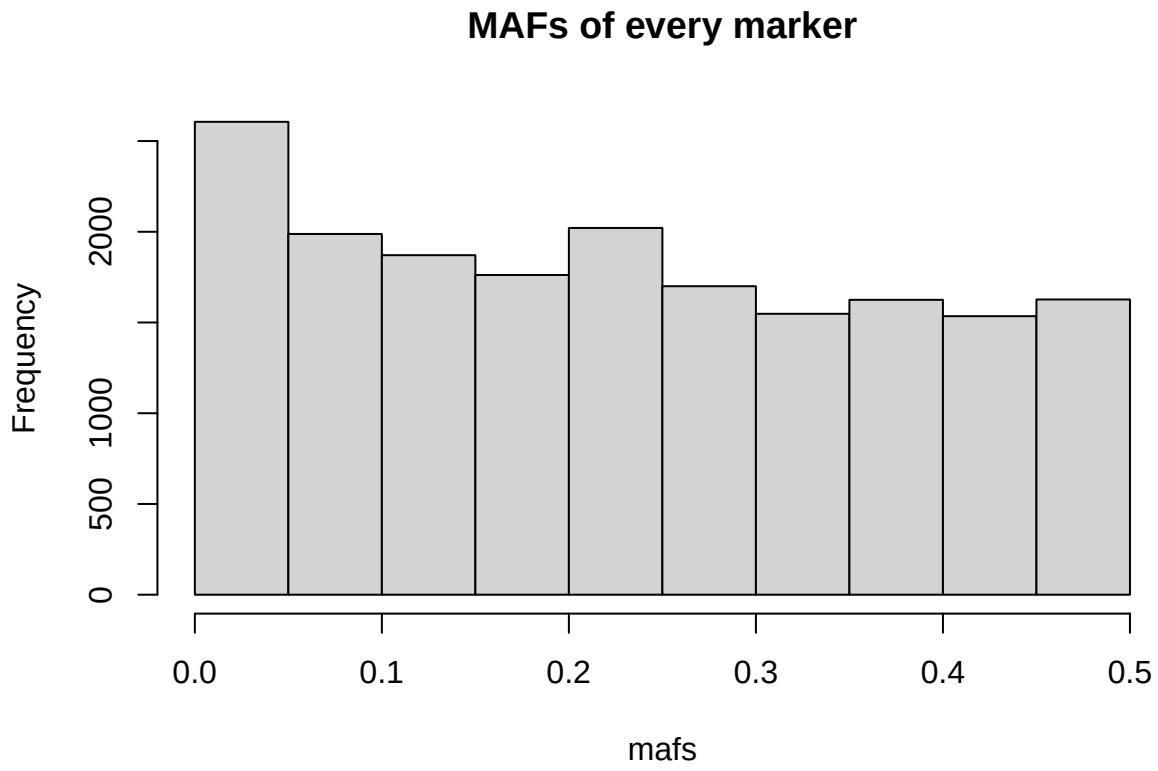
4. Compute the minor allele frequencies (MAF) for all markers, and make a histogram of it. Does the MAF follow a uniform distribution? What percentage of the markers have a MAF below 0.05? And below 0.01? Can you explain the observed pattern?

```r
#MAFs histogram
maf_list<-list()
for (x in colnames(new_dataset)){
  #genotype count
  nAA<-length(which(new_dataset[x]==0))
  nAB<-length(which(new_dataset[x]==1))
```

```
  nBB<-length(which(new_dataset[x]==2))

  #allele probabilities
  pA<-(nAA+1/2*nAB)/102
  pB<-(nBB+1/2*nAB)/102

  #MAF
  maf<-min(pA,pB)
  maf_list<-append(maf_list,maf)
}
mafs <- unlist(maf_list, use.names = FALSE)
hist(mafs, main = "MAFs of every marker")
```




```
#Percentage of MAFs
mafs0.05<-which(mafs<0.05)
(length(mafs0.05)/length(mafs))*100
```

```
## [1] 14.25368
```

```
mafs0.01<-which(mafs<0.01)
(length(mafs0.01)/length(mafs))*100
```

```
## [1] 4.791336
```

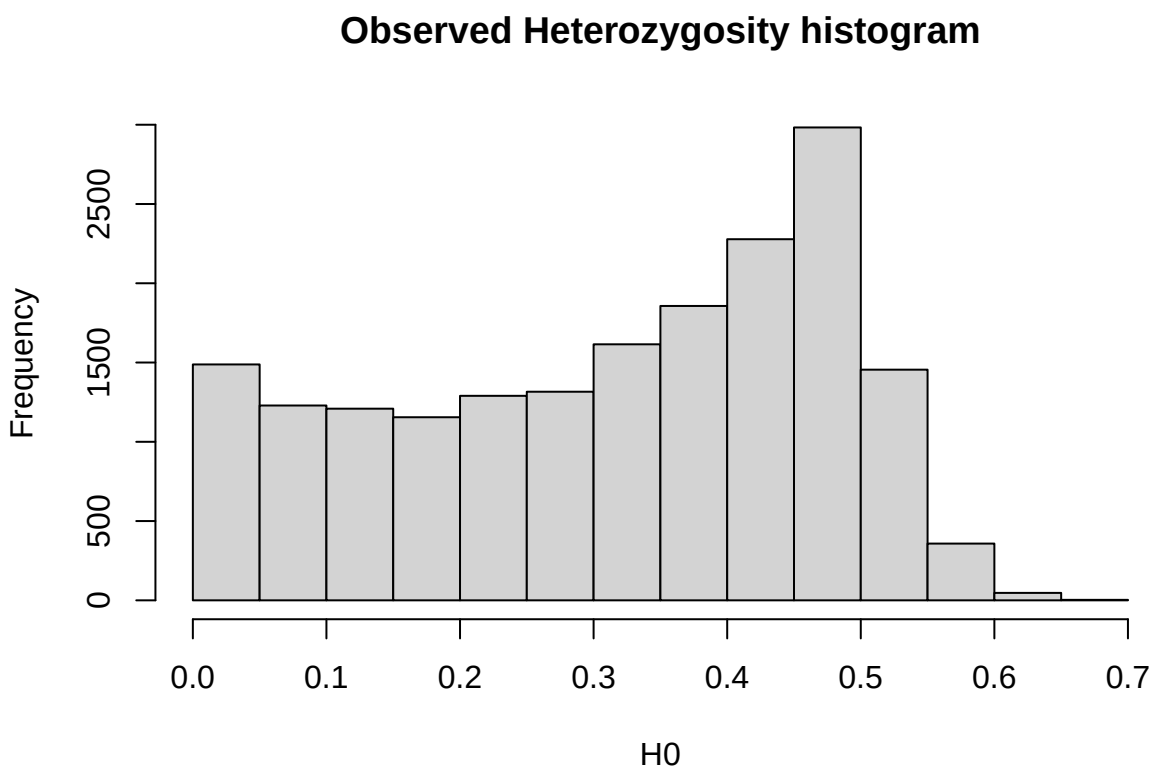
The MAF follows a uniform distribution indeed. The percentage of markers having a MAF below 0.05 is 14.25% and the percentage of markers having a MAF below 0.01 is 4.79%. The pattern that we can observe is that the most repeated case is when there is a high probability of allele A and almost a null probability of allele B, being this probaility below 0.05. This means that in most cases, the most probable allele is alelle A.

5. Calculate the observed heterozygosity H_0, and make a histogram of it. What is, theoretically, the range of variation of this statistic?

```
H0_list<-list()
for (x in colnames(new_dataset)){
  #genotype count
  nAB<-length(which(new_dataset[x]==1))

  #genotype frequencies
  fAB<-nAB/102

  H0_list<-append(H0_list,fAB)
}
H0 <- unlist(H0_list, use.names = FALSE)
hist(H0, main = "Observed Heterozygosity histogram")
```



```r
print(paste("The range of the H0 statistics is",min(H0),"and",round(max(H0),2)))
```

```
## [1] "The range of the H0 statistics is 0 and 0.68"
```

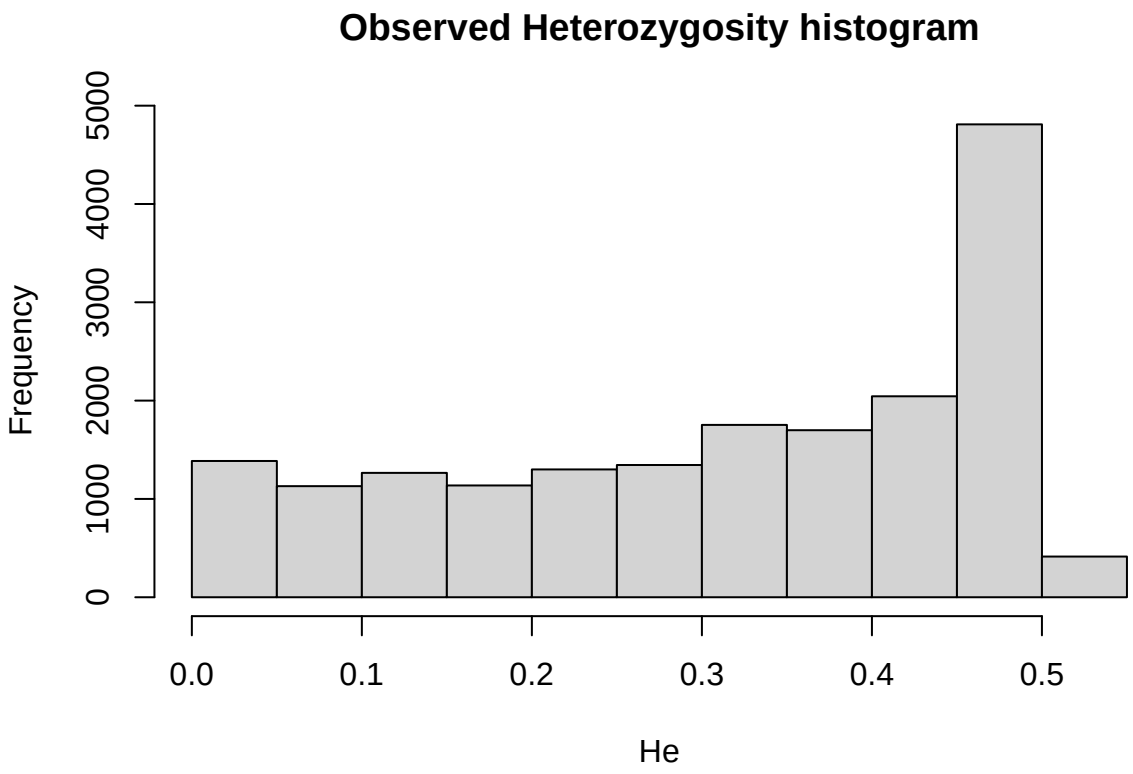
The range of variation of this statistics goes from 0 to 0.68.

6. Compute for each marker its expected heterozygosity (He), where the expected heterozygosity for a bi-allelic marker is defined as $1 - \sum_{i=1}^{k} p_i^2$, where p_i^2 is the frequency of the ith allele. 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?

```r
He_list<-list()
for (x in colnames(new_dataset)){
  #genotype count
  nAA<-length(which(new_dataset[x]==0))
  nAB<-length(which(new_dataset[x]==1))
  nBB<-length(which(new_dataset[x]==2))

  #allele probabilities
  pA<-(nAA+1/2*nAB)/102
  pB<-(nBB+1/2*nAB)/102

  #Compute He
  he<-1-(pA*pA+pB*pB)
  He_list<-append(He_list,he)
}
He <- unlist(He_list, use.names = FALSE)
hist(He, main = "Observed Heterozygosity histogram")
```




```
print(paste("The range of the He statistics is",round(min(He),3),"and",round(max(He),2),".","The averag
```

```
## [1] "The range of the He statistics is 0.01 and 0.55 . The average of He is 0.314"
```

The range of variation of this statistics goes from 0.01 and 0.55. The average of the He statistics is 0.314.

STR

```
data("NistSTRs")
```

1. How many individuals and how many STRs contains the database?

```
print(paste0("There are ",dim(NistSTRs)[1]," individuals"))
```

```
## [1] "There are 361 individuals"
```

```
print(paste0("There are ",dim(NistSTRs)[2]," STRs"))
```

```
## [1] "There are 58 STRs"
```

2. 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).

```
numberAlleles <- function(STR){
  return(length(unique(STR)))
}

alleles <- numeric(ncol(NistSTRs))
for(i in 1:ncol(NistSTRs)){
  alleles[i] <- numberAlleles(NistSTRs[,i])
}

print("This is the number of alleles in each STR:")
```

```
## [1] "This is the number of alleles in each STR:"
```

```
print(alleles)
```

```
##  [1]  6  5  6  7 13 15  6  8  6  6 13 11 11 12 13 14 11 12  7  7 11 10  8  9  7
## [26]  7  7  7 12 12  7  7  9  8  5 11  5  6  5  5 12 12  5  6  8  7 10  9 13 18
## [51] 29 32  6  7  7  6  8  8
```

```
print(paste0("The mean of alleles is ",mean(alleles)))
```

```
## [1] "The mean of alleles is 9.48275862068965"
```

```
print(paste0("The standard deviation of the number of alleles is ",sd(alleles)))
```

```
## [1] "The standard deviation of the number of alleles is 5.0061063740629"
```

```
print(paste0("The median of alleles is ",median(alleles)))
```

```
## [1] "The median of alleles is 8"
```

```
print(paste0("The minimum of alleles is ",min(alleles)))
```

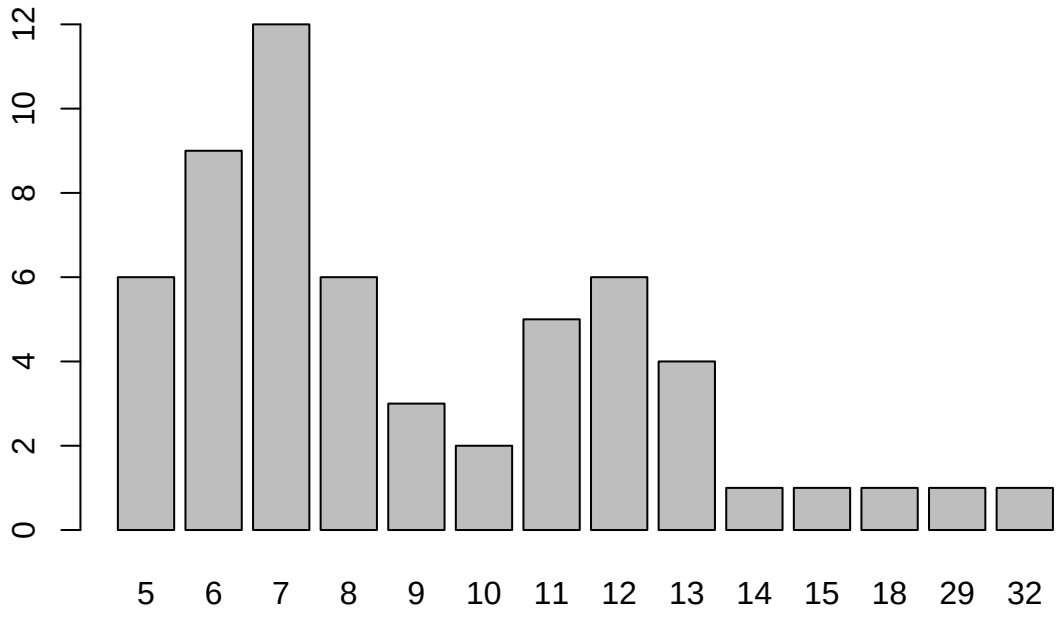
```
## [1] "The minimum of alleles is 5"
```

```
print(paste0("The maximum of alleles is ",max(alleles)))
```

```
## [1] "The maximum of alleles is 32"
```

3. 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?

```
tbl <- table(alleles)
barplot(tbl)
```



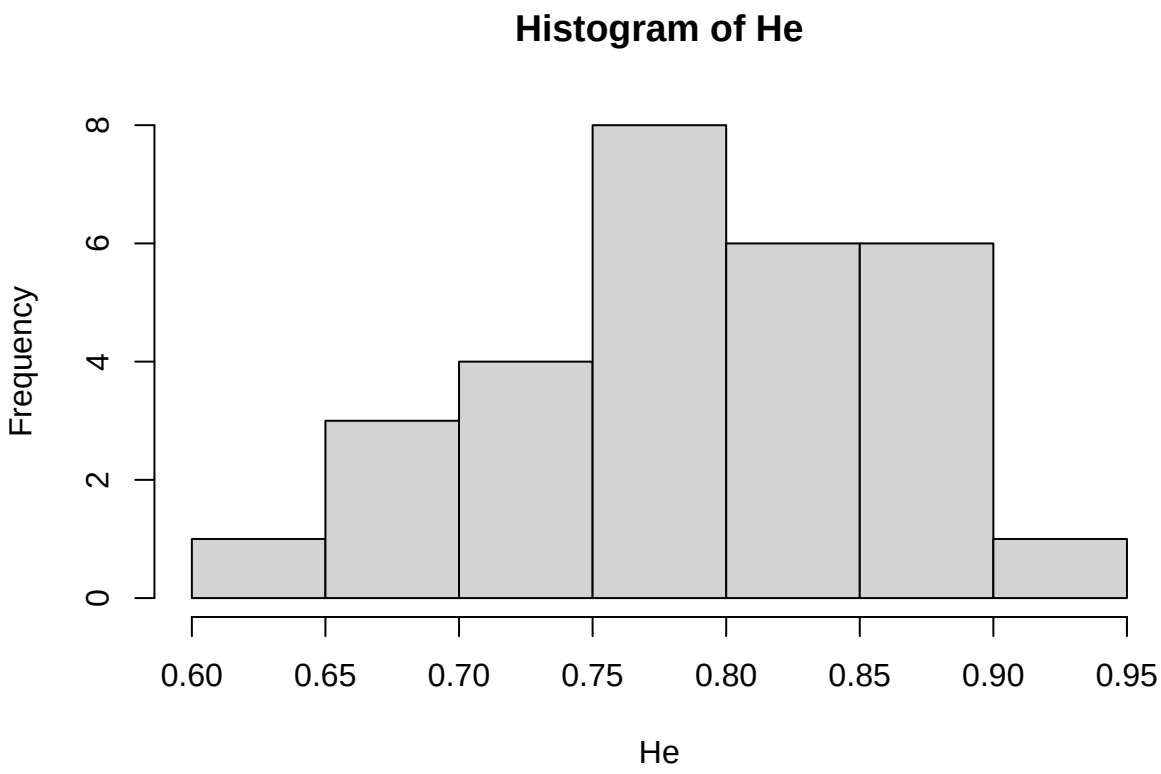
As we can see in the barplot, the most common number of alleles for an STR is 7

4. 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.

```
He <- numeric(ncol(NistSTRs)/2-1)

for(i in 0:(ncol(NistSTRs)/2-1)){
  chr1 <- NistSTRs[,i*2+1]
  chr2 <- NistSTRs[,i*2+2]
  chrC <- unique(append(chr1,chr2))
  prob <- numeric(length(chrC))
  for(j in 1:length(chr1)){
    if(chr1[j] == chr2[j]){
      prob[which(chr1[j]==chrC)] = prob[which(chr1[j]==chrC)] + 1
    } else {
      prob[which(chr1[j]==chrC)] = prob[which(chr1[j]==chrC)] + 0.5
      prob[which(chr2[j]==chrC)] = prob[which(chr2[j]==chrC)] + 0.5
    }
  }
  prob <- (prob/nrow(NistSTRs))^2
```

```
  He[i+1] <- 1-sum(prob)
}
hist(He)
```



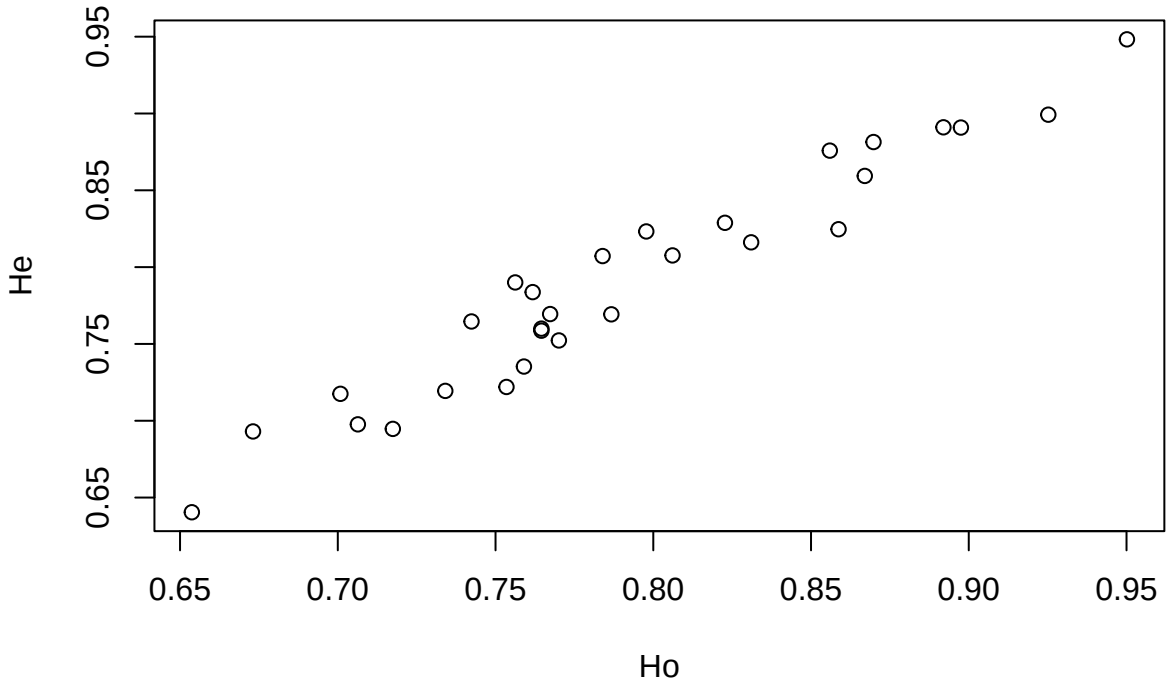

```
print(paste0("The average expected heterozygosity is ", mean(He)))
```

```
## [1] "The average expected heterozygosity is 0.790404277607362"
```

5. Calculate also the observed heterozygosity for each STR. Plot observed against expected heterozy- gosity, using all STRs. What do you observe?

```
Ho <- numeric(ncol(NistSTRs)/2-1)
for(i in 0:(ncol(NistSTRs)/2-1)){
  chr1 <- NistSTRs[,i*2+1]
  chr2 <- NistSTRs[,i*2+2]
  tbl <- table(chr1==chr2)
  Ho[i+1] <- tbl[1]/(sum(tbl))
}

plot(Ho,He)
```



We can see that the expected heterogeneity is very close to the one finally observed

6. Compare, overall, 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?

The differences between the results obtained from SNP and STR are that the average He is higher for STR than for SNP. The differences we can observe between these two types of genetic markers is that the number of alleles is higher for STR than for SNP as for SNP there is only one change of base and for STR there are 1 to 6 bp involved. The final remark would be that the reason that STR has higher values for H0 and He is the higher amount of different alleles when comparing it to the SNP case.