# Hendrik Wellmann

# Antonin Rosa

# Assigment BSG Stats 1

## Fisrt data set

### Question 1

```r
file <-"TSICHR22RAW.raw"
data <- read.table(file, header=FALSE, sep=" ")
genetic_data <- data[, 7:ncol(data)]
first_elements <- sapply(genetic_data, function(x) x[1])
rs_columns <- grepl("^rs", first_elements)
genetic_data_rs <- genetic_data[, rs_columns]
print(dim(genetic_data_rs))
```

```
## [1]   103 20649
```

```r
num_variants <- ncol(genetic_data_rs)
missing_percentage <- mean(is.na(genetic_data_rs)) * 100

cat("Number of variants in the database:", num_variants, "\n")
```

```
## Number of variants in the database: 20649
```

```r
cat("Percentage of missing data:", missing_percentage, "%\n")
```

```
## Percentage of missing data: 0.1967231 %
```

For the missing, we decided to compute of the value which are equal to NA under all the number of value present in the table.

### Question 2

```r
monomorphic_variants <- sapply(genetic_data_rs, function(col) {
  cleaned_col <- col[-1]  # Remove the first element (variant name)
  length(unique(cleaned_col[!is.na(cleaned_col)])) == 1
})
percentage_monomorphic <- (sum(monomorphic_variants) / length(monomorphic_variants)) * 100
non_monomorphic_data <- genetic_data_rs[, !monomorphic_variants]
num_remaining_variants <- ncol(non_monomorphic_data)

cat("Percentage of monomorphic variants:", percentage_monomorphic, "%\n")
```

```
## Percentage of monomorphic variants: 11.45818 %
```

```r
cat("Number of remaining variants:", num_remaining_variants, "\n")
```

```
## Number of remaining variants: 18283
```

### Question 3

```r
column_index <- which(sapply(genetic_data_rs, function(col) col[1] == "rs8138488_C"))
rs8138488_column <- genetic_data_rs[, column_index]
genotype_counts <- table(rs8138488_column[-1])
```

```r
genotype_0 <- genotype_counts["0"]
genotype_1 <- genotype_counts["1"]
genotype_2 <- genotype_counts["2"]
minor_allele_count <- min(genotype_0, genotype_1, genotype_2)
total_individuals <- length(rs8138488_column) - 1
MAF <- minor_allele_count / (2 * total_individuals)
cat("Genotype counts for rs8138488_C:\n")
```

```
## Genotype counts for rs8138488_C:
```

```r
print(genotype_counts)
```

```
##
##  0  1  2
## 41 47 14
```

```r
cat("Minor Allele Count:", minor_allele_count, "\n")
```

```
## Minor Allele Count: 14
```

```r
cat("Minor Allele Frequency (MAF):", MAF, "\n")
```

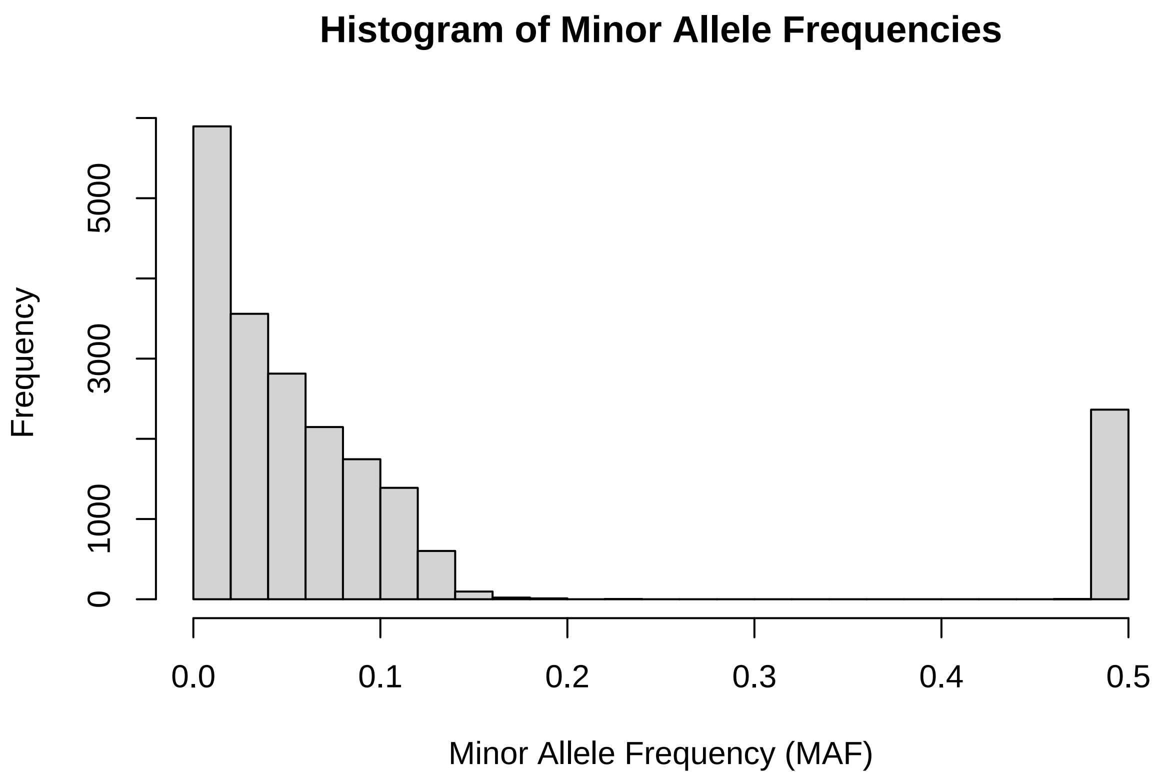
```
## Minor Allele Frequency (MAF): 0.06862745
```

### Question 4

```r
MAF_values <- numeric(length(genetic_data_rs))

for (i in 1:length(genetic_data_rs)) {
  column <- genetic_data_rs[[i]]
  allele_counts <- table(column[-1])
  minor_allele_count <- min(allele_counts)
  total_individuals <- length(column) - 1
  MAF <- minor_allele_count / (2 * total_individuals)
  MAF_values[i] <- MAF
}

hist(MAF_values, breaks = 20, xlab = "Minor Allele Frequency (MAF)", main = "Histogram of Minor Allele
```



```
percentage_below_005 <- sum(MAF_values < 0.05) / length(MAF_values) * 100
percentage_below_001 <- sum(MAF_values < 0.01) / length(MAF_values) * 100

cat("Percentage of markers with MAF below 0.05:", percentage_below_005, "%\n")
```

```
## Percentage of markers with MAF below 0.05: 52.87423 %
```

```
cat("Percentage of markers with MAF below 0.01:", percentage_below_001, "%\n")
```

```
## Percentage of markers with MAF below 0.01: 17.74904 %
```

The MAF does not follow an uniform distribution.

A possible explaination for this pattern could that most geneteric variants appear infrequently because they are unlikely to change in human population. The peak at 0.5 means that this genetic variant is very common in population. So it doesn't seem to be harmful for the population.

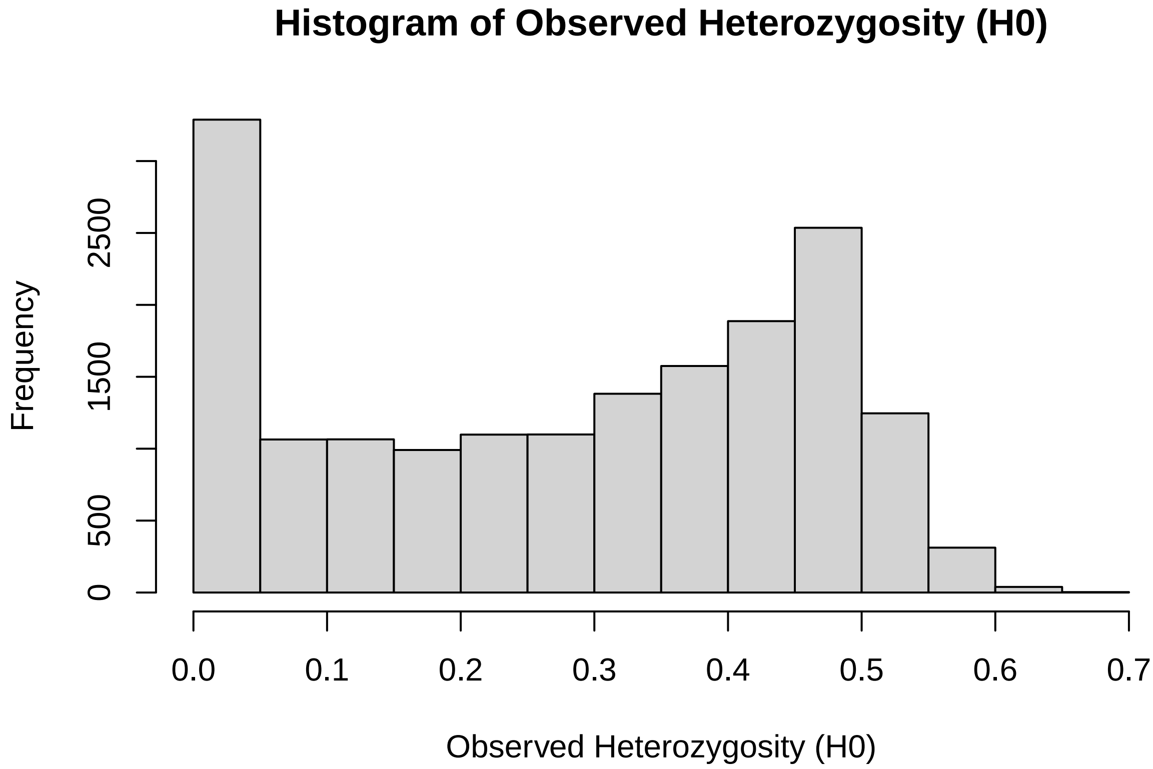
### Question 5

```
HO_values <- numeric(length(genetic_data_rs))

for (i in 1:length(genetic_data_rs)) {
  column <- genetic_data_rs[[i]]
  genotypes <- column[-1]
  num_heterozygous <- sum(genotypes == 1)
  total_individuals <- length(genotypes)
  HO <- num_heterozygous / total_individuals
  HO_values[i] <- HO
```

```
}

hist(H0_values, breaks = 20, xlab = "Observed Heterozygosity (H0)", main = "Histogram of Observed Hetero
```



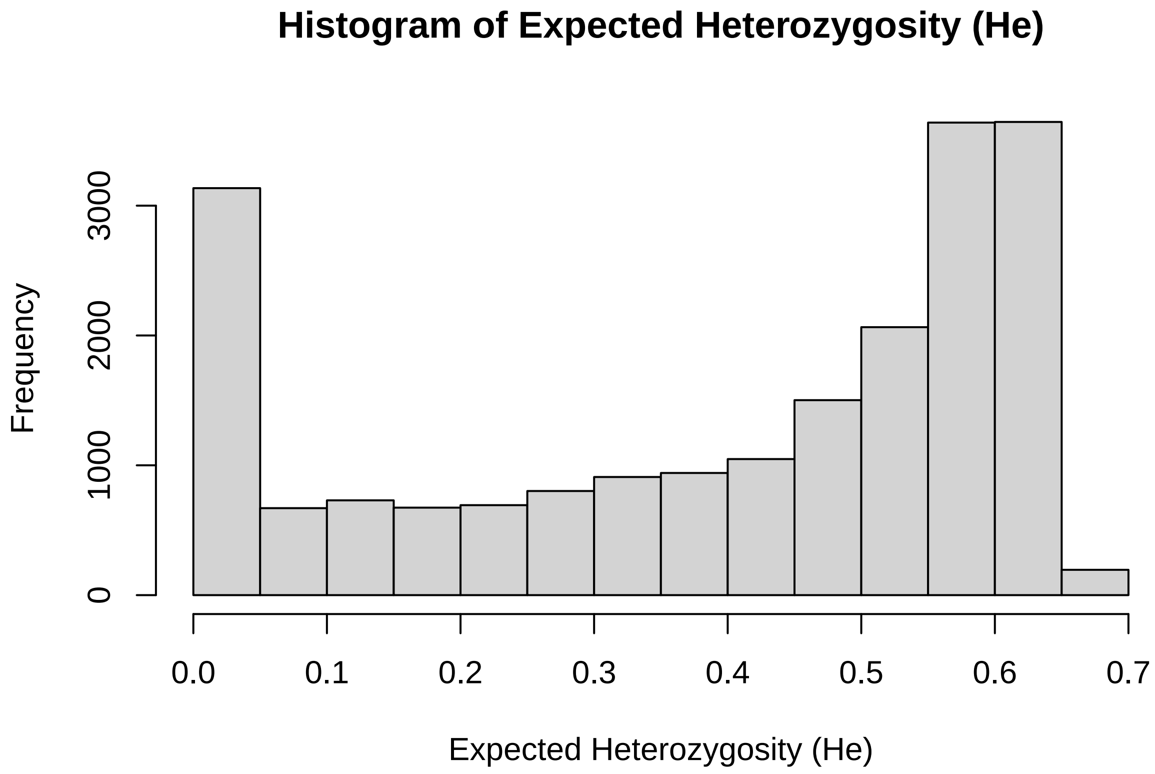
Theoretical range of variation for H0 is [0, 0.5]

**Question 6**

```
He_values <- numeric(length(genetic_data_rs))

for (i in 1:length(genetic_data_rs)) {
  column <- genetic_data_rs[[i]]
  genotypes <- column[-1]
  allele_freq <- table(genotypes) / length(genotypes)
  He <- 1 - sum(allele_freq^2)
  He_values[i] <- He
}

hist(He_values, breaks = 20, xlab = "Expected Heterozygosity (He)", main = "Histogram of Expected Hetero
```



```r
average_He <- mean(He_values)
cat("Average Expected Heterozygosity (He) for this database:", average_He, "\n")
```

```
## Average Expected Heterozygosity (He) for this database: 0.3904951
```

Theoretical range of variation for H0 is [0, 0.5]

## Second data set

### Question 1

```r
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
```

```r
data(NistSTRs)
dimensions <- dim(NistSTRs)
num_individuals <- dimensions[1]
num_STRs <- dimensions[2] / 2
cat("Number of individuals in the database:", num_individuals, "\n")
```

```
## Number of individuals in the database: 361
```

```r
cat("Number of STRs in the database:", num_STRs, "\n")
```

```
## Number of STRs in the database: 29
```

### Question 2

```r
count_alleles <- function(STR_locus) {
  unique_alleles <- unique(STR_locus)
  num_alleles <- length(unique_alleles)
  return(num_alleles)
}
num_alleles_list <- sapply(NistSTRs, count_alleles)

mean_num_alleles <- mean(num_alleles_list)
sd_num_alleles <- sd(num_alleles_list)
median_num_alleles <- median(num_alleles_list)
min_num_alleles <- min(num_alleles_list)
max_num_alleles <- max(num_alleles_list)

cat("Descriptive statistics of the number of alleles:\n")
```

```
## Descriptive statistics of the number of alleles:
```

```r
cat("Mean:", mean_num_alleles, "\n")
```

```
## Mean: 9.482759
```

```r
cat("Standard Deviation:", sd_num_alleles, "\n")
```

```
## Standard Deviation: 5.006106
```

```r
cat("Median:", median_num_alleles, "\n")
```

```
## Median: 8
```

```r
cat("Minimum:", min_num_alleles, "\n")
```

```
## Minimum: 5
```

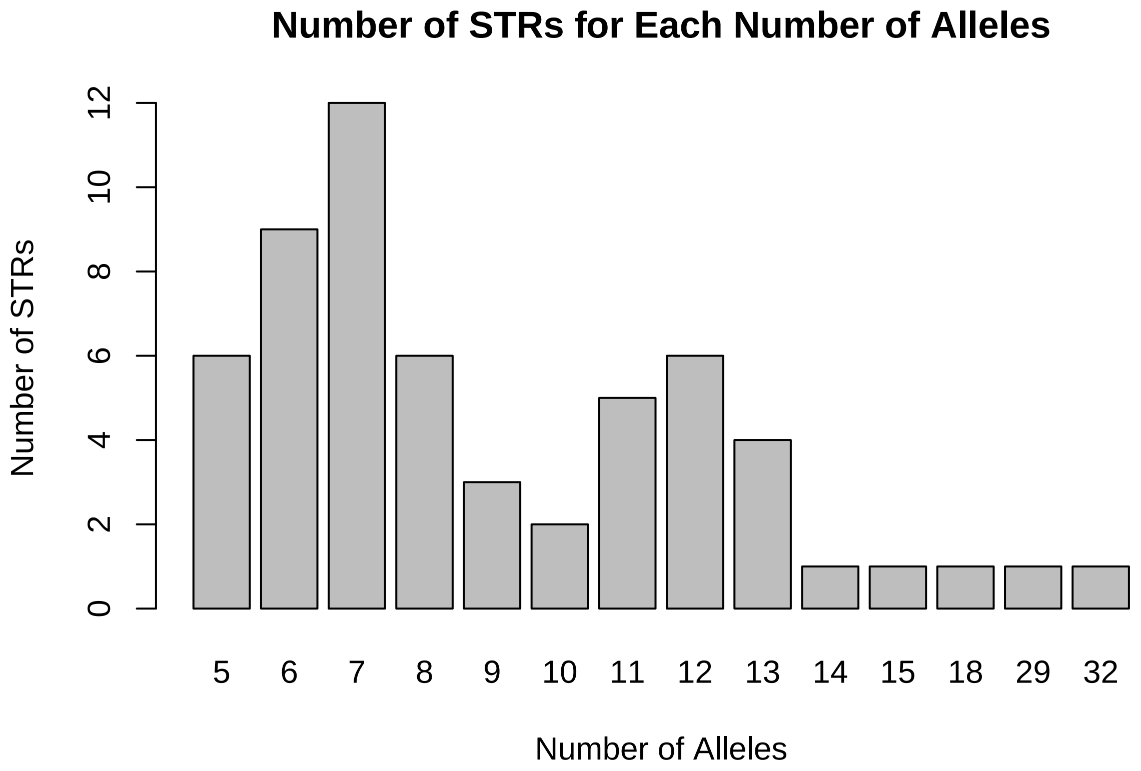
```r
cat("Maximum:", max_num_alleles, "\n")
```

```
## Maximum: 32
```

### Question 3

```r
table_num_alleles <- table(num_alleles_list)

barplot(table_num_alleles, xlab = "Number of Alleles", ylab = "Number of STRs", main = "Number of STRs
```



```
most_common_alleles <- names(table_num_alleles)[which.max(table_num_alleles)]
cat("The most common number of alleles for an STR is:", most_common_alleles, "\n")
```

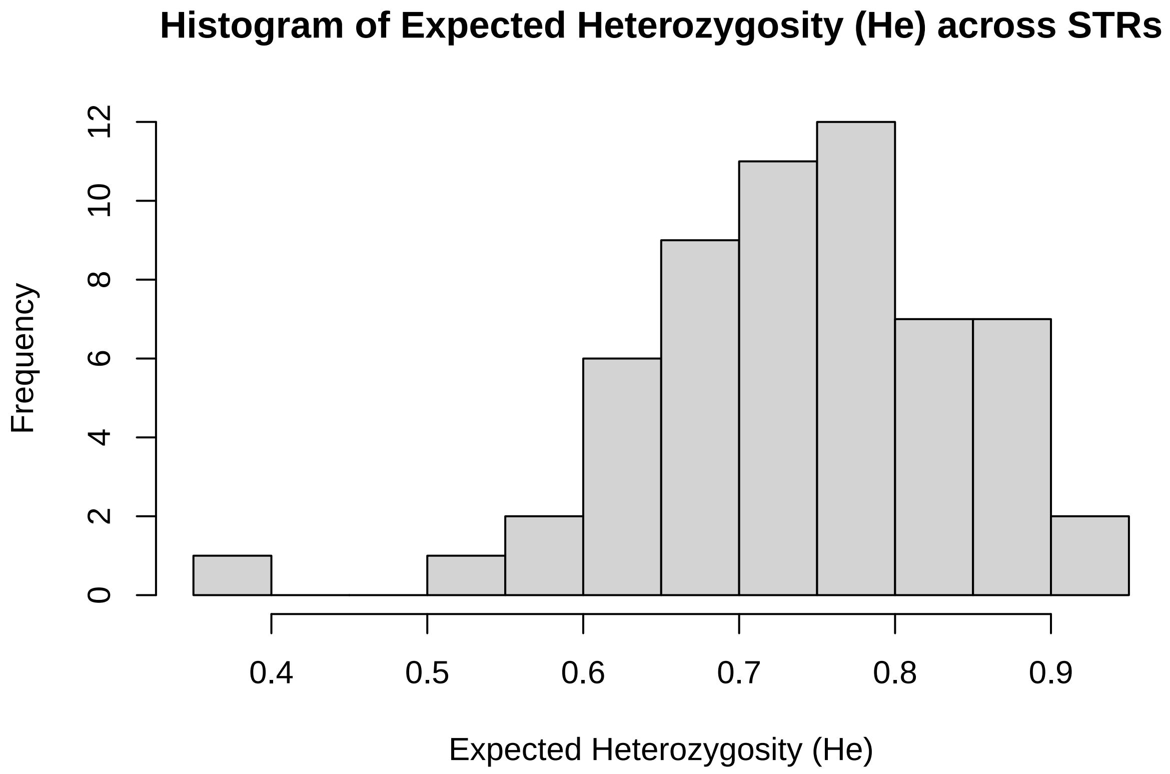
```
## The most common number of alleles for an STR is: 7
```

**Question 4**

```
calculate_He <- function(STR_locus) {
  unique_alleles <- unique(STR_locus)
  allele_freq <- table(STR_locus) / length(STR_locus)
  He <- 1 - sum((allele_freq / sum(allele_freq))^2)
  return(He)
}

He_values <- sapply(NistSTRs, calculate_He)

hist(He_values, breaks = 20, xlab = "Expected Heterozygosity (He)", main = "Histogram of Expected Heter
```



```
average_He <- mean(He_values)
cat("Average Expected Heterozygosity over all STRs:", average_He, "\n")
```

```
## Average Expected Heterozygosity over all STRs: 0.7404483
```

**Question 5**

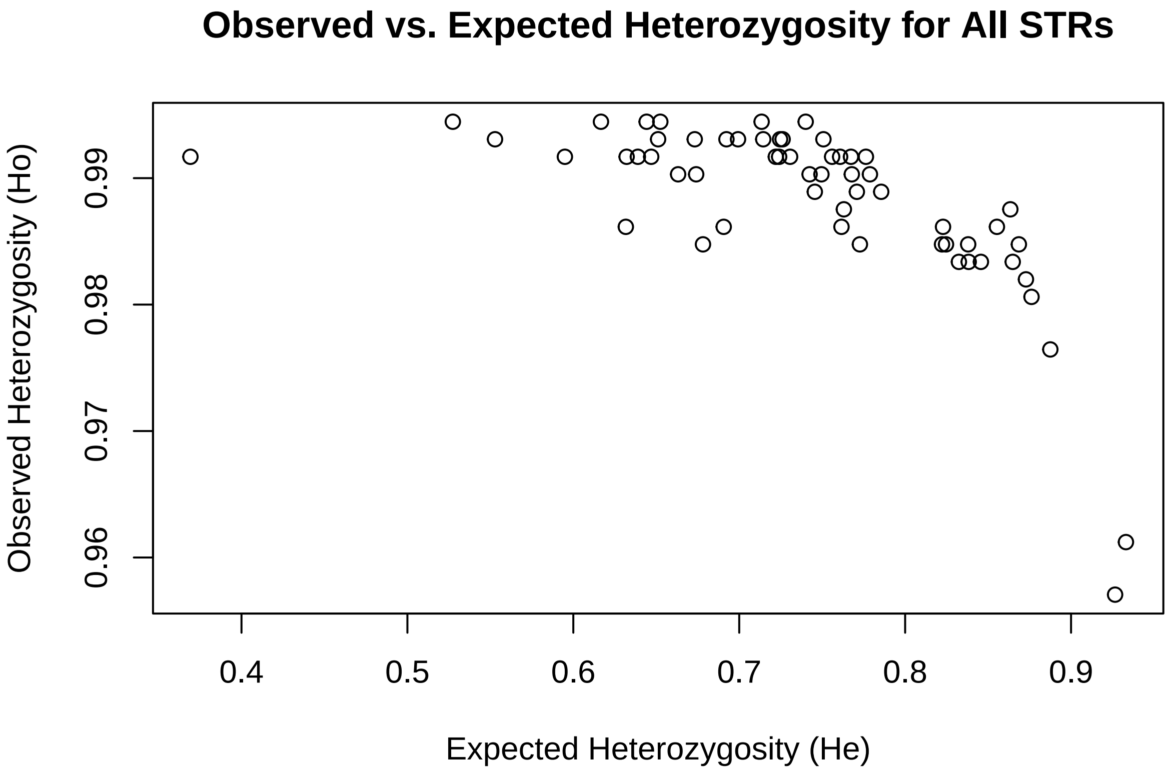
```
calculate_Ho <- function(STR_locus) {
  num_unique_alleles <- length(unique(STR_locus))
  total_individuals <- length(STR_locus)
  Ho <- 1 - (num_unique_alleles - 1) / (2 * total_individuals)
  return(Ho)
}

Ho_values <- sapply(NistSTRs, calculate_Ho)

heterozygosity_data <- data.frame(He = He_values, Ho = Ho_values)
print(heterozygosity_data)
```

```
##                     He        Ho
## CSF1PO-1    0.6731686 0.9930748
## CSF1PO-2    0.6441786 0.9944598
## D10S1248-1 0.6511000 0.9930748
## D10S1248-2 0.7608137 0.9916898
## D12S391-1  0.8456657 0.9833795
## D12S391-2  0.8761750 0.9806094
## D13S317-1  0.7507462 0.9930748
## D13S317-2  0.7424130 0.9903047
## D16S539-1  0.7245954 0.9930748
## D16S539-2  0.6992733 0.9930748
## D18S51-1   0.8324061 0.9833795
## D18S51-2   0.8553188 0.9861496
## D19S433-1  0.6906331 0.9861496
## D19S433-2  0.7727227 0.9847645
## D1S1656-1  0.8648491 0.9833795
## D1S1656-2  0.8728141 0.9819945
## D21S11-1   0.7616271 0.9861496
## D21S11-2   0.8246253 0.9847645
## D22S1045-1 0.7220479 0.9916898
## D22S1045-2 0.5949156 0.9916898
## D2S1338-1  0.8228298 0.9861496
## D2S1338-2  0.8634065 0.9875346
## D2S441-1   0.6631625 0.9903047
## D2S441-2   0.7455437 0.9889197
## D3S1358-1  0.7240583 0.9916898
## D3S1358-2  0.7559948 0.9916898
## D5S818-1   0.6388993 0.9916898
## D5S818-2   0.6321314 0.9916898
## D6S1043-1  0.6781409 0.9847645
## D6S1043-2  0.8685477 0.9847645
## D7S820-1   0.7763599 0.9916898
## D7S820-2   0.7673360 0.9916898
## D8S1179-1  0.7855679 0.9889197
## D8S1179-2  0.7678425 0.9903047
## F13A01-1   0.7134844 0.9944598
## F13A01-2   0.6316403 0.9861496
## F13B-1     0.7400496 0.9944598
## F13B-2     0.5528042 0.9930748
## FESFPS-1   0.6166466 0.9944598
## FESFPS-2   0.6524351 0.9944598
## FGA-1      0.8222159 0.9847645
## FGA-2      0.8379923 0.9847645
## LPL-1      0.5273747 0.9944598
## LPL-2      0.6922906 0.9930748
## Penta_C-1  0.6740433 0.9903047
## Penta_C-2  0.7306574 0.9916898
## Penta_D-1  0.7630543 0.9875346
## Penta_D-2  0.7708965 0.9889197
## Penta_E-1  0.8382993 0.9833795
## Penta_E-2  0.8875009 0.9764543
## SE33-1     0.9330806 0.9612188
## SE33-2     0.9265429 0.9570637
## TH01-1     0.7261608 0.9930748
## TH01-2     0.6469410 0.9916898
## TPOX-1     0.3692114 0.9916898
## TPOX-2     0.7144666 0.9930748
## vWA-1      0.7787693 0.9903047
## vWA-2      0.7495338 0.9903047
```

```r
plot(heterozygosity_data$He, heterozygosity_data$Ho, xlab = "Expected Heterozygosity (He)", ylab = "Obs
     main = "Observed vs. Expected Heterozygosity for All STRs")
```



He values are very different from H0 values and it's quite surprising to observe that H0 values is superior to 0.96 for all STR locus and even He is superior to 0.5.

### Question 6

There seem to be big differences in heterozygosity levels between the SNP and STR databases. The STR database has much higher heterozygosity levels compared to the SNP database. This could be due to the inherent characteristics of the markers, the size of the databases, or both.