Practical 01 SG: Descriptive analysis of genetic markers

Quim Aguado, Marcel Cases

20-Nov-2021

SNP dataset

```
rm(list=ls())
```

Load *TSICHR22RAW* data into R.

```
filename <- url("http://www-eio.upc.es/~jan/data/bsg/TSICHR22RAW.raw")
df <- read.table(filename, header=TRUE)
```

Extract the variables individual ID (the second column IID) and the sex of the individual (the 5th column sex).

```
IID_SEX <- df[,c(2,5)]
```

Create a dataframe that only contains the genetic information that is in and beyond the 7th column.

```
gendata <- df[, 7:ncol(df)]
gendata[gendata==0] <- "AA"
gendata[gendata==1] <- "AB"
gendata[gendata==2] <- "BB"
```

Ex 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?

```
n <- nrow(gendata)
p <- ncol(gendata) # number of variants
p
```

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

```
perc.mis <- 100*sum(is.na(gendata))/(n*p) # returns true if a datapoint is missing; false otherwise
perc.mis
```

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

```
male <- length(which(IID_SEX == 1))
female <- length(which(IID_SEX == 2))
perc.male <- 100*male/(male+female)
perc.male
```

```
## [1] 51.96078
```

```
perc.female <- 100*female/(male+female)
perc.female
```

```
## [1] 48.03922
```

There are 20649 variants. There is a 0.1986% of missing data. 51.96% of the individuals are male, while 48.03% are female.

Ex 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?

```
gendata_poly <- gendata
monomorphics_index <- c()
for(i in 1:p) {       # for-loop over columns of gendata
  if (length(unique(gendata[!is.na(gendata[,i]),i]))==1) { # true if there is only one allele in the co
    monomorphics_index <- c(monomorphics_index, i)
  }
}
gendata_poly <- gendata[-c(monomorphics_index)] # we want to remove monomorphics and put polymorphics i
homomorphic.num <- ncol(gendata)-ncol(gendata_poly)
perc.homomorphic <- 100*homomorphic.num/(ncol(gendata))
perc.homomorphic
```

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

```
ncol(gendata_poly)
```

```
## [1] 18283
```

There is a 11.458% of monomorphic variants in the dataset. There are still 18283 variants in the dataset without monomorphics.

Ex 5 Report the genotype counts and the minor allele count of polymorphism rs8138488_C, and calculate the MAF of this variant.

```
rs8138488_C <- gendata_poly[,c("rs8138488_C")]
rs8138488_C.g <- genotype(rs8138488_C,sep="")
summary(rs8138488_C.g)
```

```
##
## Number of samples typed: 102 (100%)
##
## Allele Frequency: (2 alleles)
##   Count Proportion
## A   129       0.63
## B    75       0.37
##
##
## Genotype Frequency:
##     Count Proportion
## A/A    41       0.40
## A/B    47       0.46
## B/B    14       0.14
##
## Heterozygosity (Hu)  = 0.4672559
## Poly. Inf. Content   = 0.356869
```

Genotype counts: * A/A -> 41 * A/B -> 47 * B/B -> 14

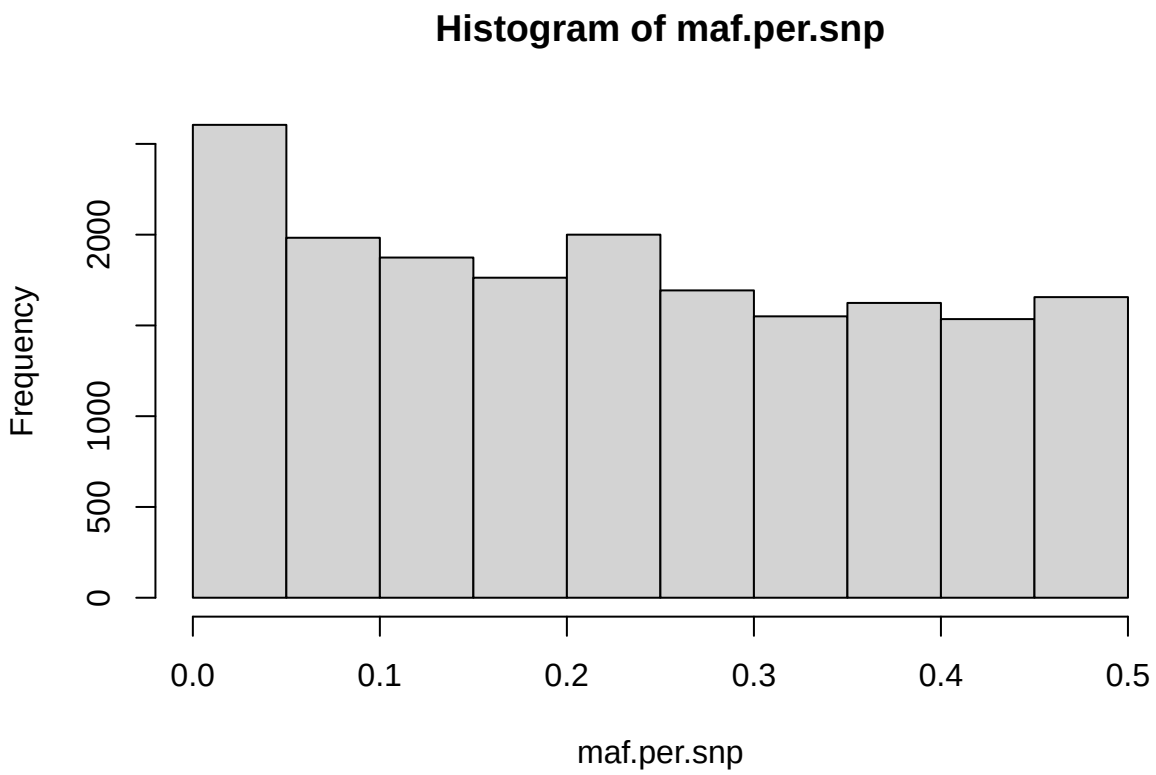
Minor allele count: 75 (allele B)

Minor allele frequency: * MAF = min(pA, pB) = 0.37

Ex 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?

```
n_poly <- nrow(gendata_poly)
nmis <- function(x) {
  y <- sum(is.na(x))
  return(y)
}
nmis.per.snp <- apply(gendata_poly,2,nmis)
pmis.per.snp <- 100*nmis.per.snp/n_poly
Y2 <- gendata_poly[,nmis.per.snp < nrow(gendata_poly)]
maf <- function(x){ # minor allele frequency
  x <- genotype(x,sep="")
  out <- summary(x)
  af1 <- min(out$allele.freq[,2],na.rm=TRUE)
  af1[af1==1] <- 0
  return(af1)
}
maf.per.snp <- apply(Y2,2,maf)
hist(maf.per.snp)
```



The MAF of all the markers follows a uniform distribution.

```
maf.per.snp_0_05 <- subset(maf.per.snp, maf.per.snp<0.05)
length(maf.per.snp_0_05)
```

```
## [1] 2601
```

```
maf.per.snp_0_05.perc <- 100*length(maf.per.snp_0_05)/(ncol(gendata_poly))
maf.per.snp_0_05.perc
```

```
## [1] 14.22633
```

```
maf.per.snp_0_01 <- subset(maf.per.snp, maf.per.snp<0.01)
length(maf.per.snp_0_01)
```

```
## [1] 865
```

```
maf.per.snp_0_01.perc <- 100*length(maf.per.snp_0_01)/(ncol(gendata_poly))
maf.per.snp_0_01.perc
```

```
## [1] 4.731171
```

- 14.22% of the markers have a MAF below 0.05
- 4.73% of the markers have a MAF below 0.01

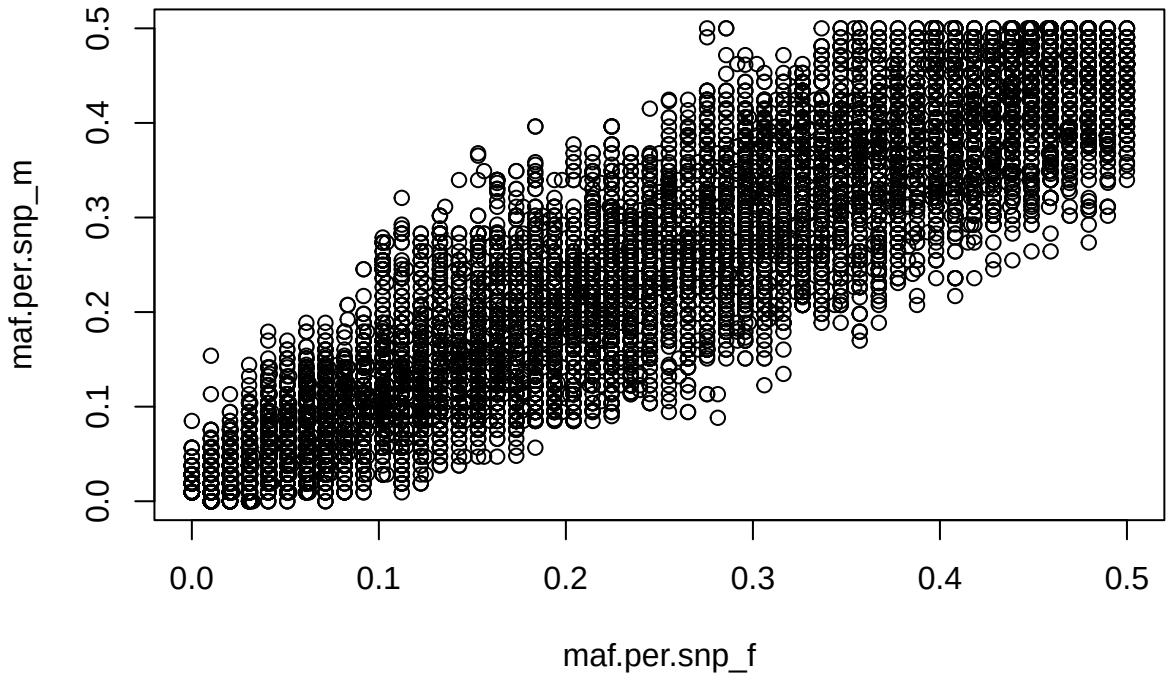
Ex 7 Calculate the minor allele frequency for males and for females and present a scatterplot of these variables. What do you observe? Calculate and report their correlation coefficient.

```
gendata_poly_fm <- gendata_poly
gendata_poly_fm <- cbind("SEX" = df[,c(5)], gendata_poly_fm)
gendata_poly_f <- subset(gendata_poly_fm, SEX==2) # female subset
gendata_poly_f <- gendata_poly_f[,2:ncol(gendata_poly_f)] # remove 'sex' column
gendata_poly_m <- subset(gendata_poly_fm, SEX==1) # male subset
gendata_poly_m <- gendata_poly_m[,2:ncol(gendata_poly_m)] # remove 'sex' column

n_poly <- nrow(gendata_poly)
nmis.per.snp_f <- apply(gendata_poly_f,2,nmis)
pmis.per.snp_f <- 100*nmis.per.snp_f/n_poly
Y2 <- gendata_poly_f[,nmis.per.snp_f < nrow(gendata_poly_f)]
maf.per.snp_f <- apply(Y2,2,maf)

nmis.per.snp_m <- apply(gendata_poly_m,2,nmis)
pmis.per.snp_m <- 100*nmis.per.snp_m/n_poly
Y2 <- gendata_poly_m[,nmis.per.snp_m < nrow(gendata_poly_m)]
maf.per.snp_m <- apply(Y2,2,maf)

plot(maf.per.snp_f, maf.per.snp_m)
```



The scatterplot shows a tendency: MAF for males and females from the genotype dataset are correlated.

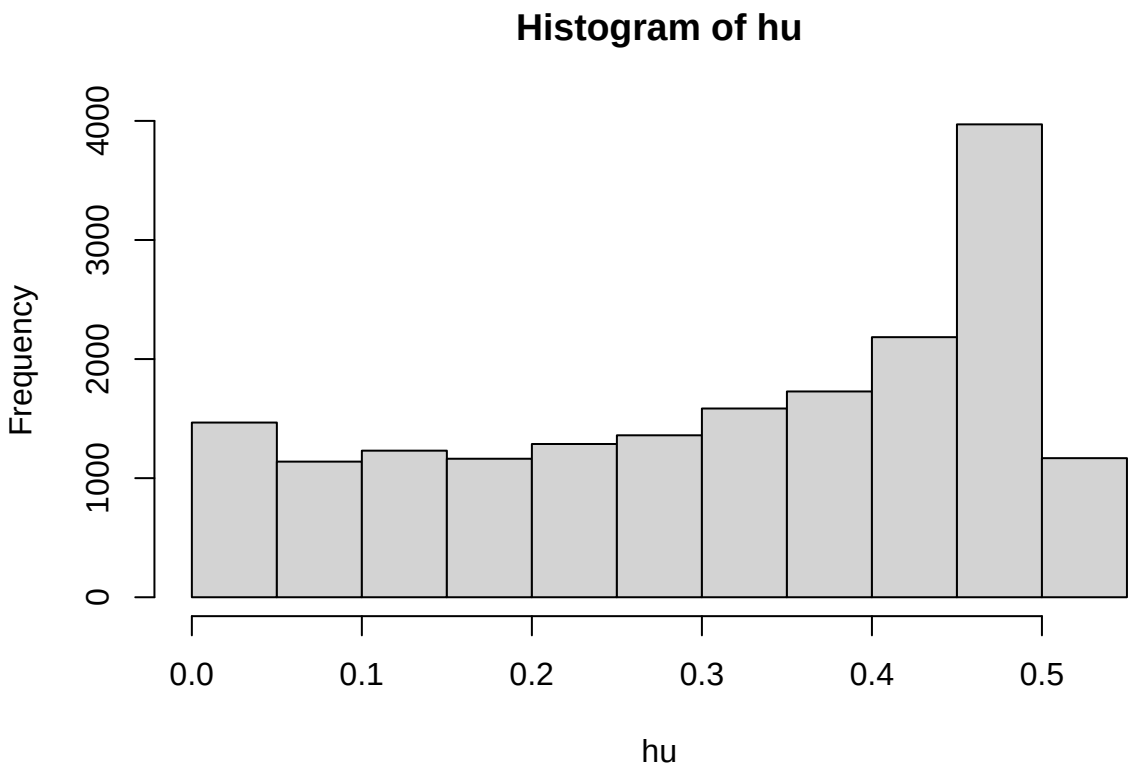
```
cor(maf.per.snp_f, maf.per.snp_m)
```

```
## [1] 0.9342241
```

R^2 between MAF for males and females is 0.934, which shows a close correlation.

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

```
hu <- c() # observed heterozygosity
for(i in 1:ncol(gendata_poly)) {       # for-loop over columns of gendata
  rs <- gendata_poly[,c(i)]
  rs.g <- genotype(rs,sep="")
  hu <- c(hu, summary(rs.g)$Hu) # retrieve heterozygosity and append to vector
}
hist(hu)
```




```
min(hu)
```

```
## [1] 0.009803922
```

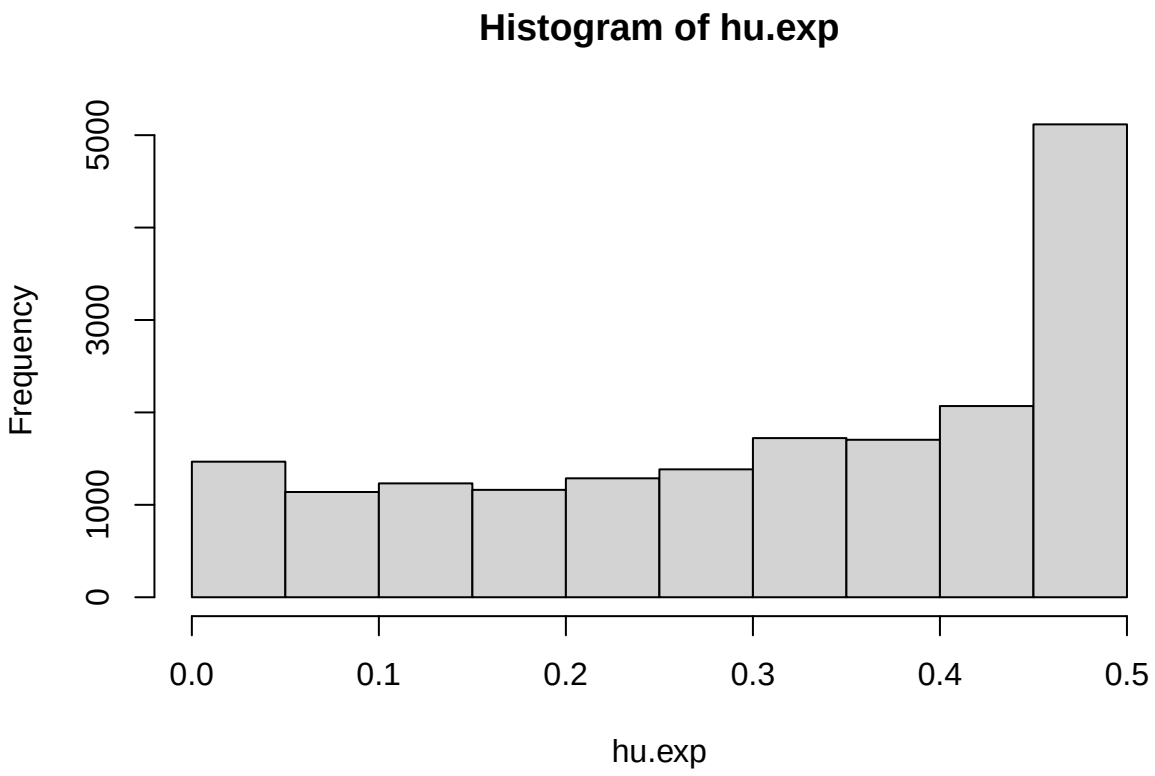
```
max(hu)
```

```
## [1] 0.5025381
```

The range of variation of the heterozygosity goes from 0.009803922 to 0.5025381.

Ex 9 Compute for each marker its expected heterozygosity (He)

```
hu.exp <- c() # expected heterozygosity
#summary(rs.g)
# <- length(summary(rs.g)$allele.names) # number of alleles
for(i in 1:ncol(gendata_poly)) {       # for-loop over columns of gendata
  rs <- gendata_poly[,c(i)]
  rs.g <- genotype(rs,sep="")
  pA <- summary(rs.g)$allele.freq[1,2] # retrieve frequency of allele A
  pB <- summary(rs.g)$allele.freq[2,2] # retrieve frequency of allele B
  val <- 1-pA^2-pB^2 # expected heterozygosity value
  hu.exp <- c(hu.exp, val)
}
hist(hu.exp)
```




```
min(hu.exp)
```

```
## [1] 0.009755863
```

```
max(hu.exp)
```

```
## [1] 0.5
```

```
mean(hu.exp)
```

```
## [1] 0.3115841
```

The range of variation of the heterozygosity goes from 0.009755863 to 0.5. The mean value of the expected heterozygosities is 0.3115841.

STR dataset

```
rm(list=ls())
```

Load *NistSTRs* data into R.

```
data(NistSTRs)
```

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

```
n <- nrow(NistSTRs) # number of individuals
p <- ncol(NistSTRs) # number of alleles
n
```

```
## [1] 361
```

```
p/2 # number of STRs
```

```
## [1] 29
```

The dataset has 361 individuals, each with 29 STRs (made up by two alleles each).

Ex 3 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).

```
compute.n.alleles.per.STR <- function(x1,x2) {
  STRi <- paste(NistSTRs[,x1],NistSTRs[,x2]) # concatenate
  numSTRi <- length(unique(STRi))
  return(numSTRi)
}
index.1 <- seq(1,length(NistSTRs),2) # takes odd values 1 3 5 ...
index.2 <- seq(2,length(NistSTRs),2) # takes even values 2 4 6 ...
n.alleles.per.STR <- c()
for(i in index.1) {
  n.alleles.per.STR <- c(n.alleles.per.STR, compute.n.alleles.per.STR(i,i+1))
}
n.alleles.per.STR
```

```
##  [1]  19  21  79  27  25  64  41  77  50  22  59  29  23  23  53  29  37  27  17
## [20]  14  54  17  27  39  92 195  21  17  28
```

```
mean(n.alleles.per.STR)
```

```
## [1] 42.27586
```

```
sd(n.alleles.per.STR)
```

```
## [1] 36.01973
```

```
median(n.alleles.per.STR)
```

```
## [1] 28
```

```
min(n.alleles.per.STR)
```

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

```
max(n.alleles.per.STR)
```

```
## [1] 195
```

Descriptive results of the alleles contained in *NistSTRs*: * Mean: 42.27586 * Standard deviation: 36.01973 * Median: 28 * Minimum: 14 * Maximum: 195

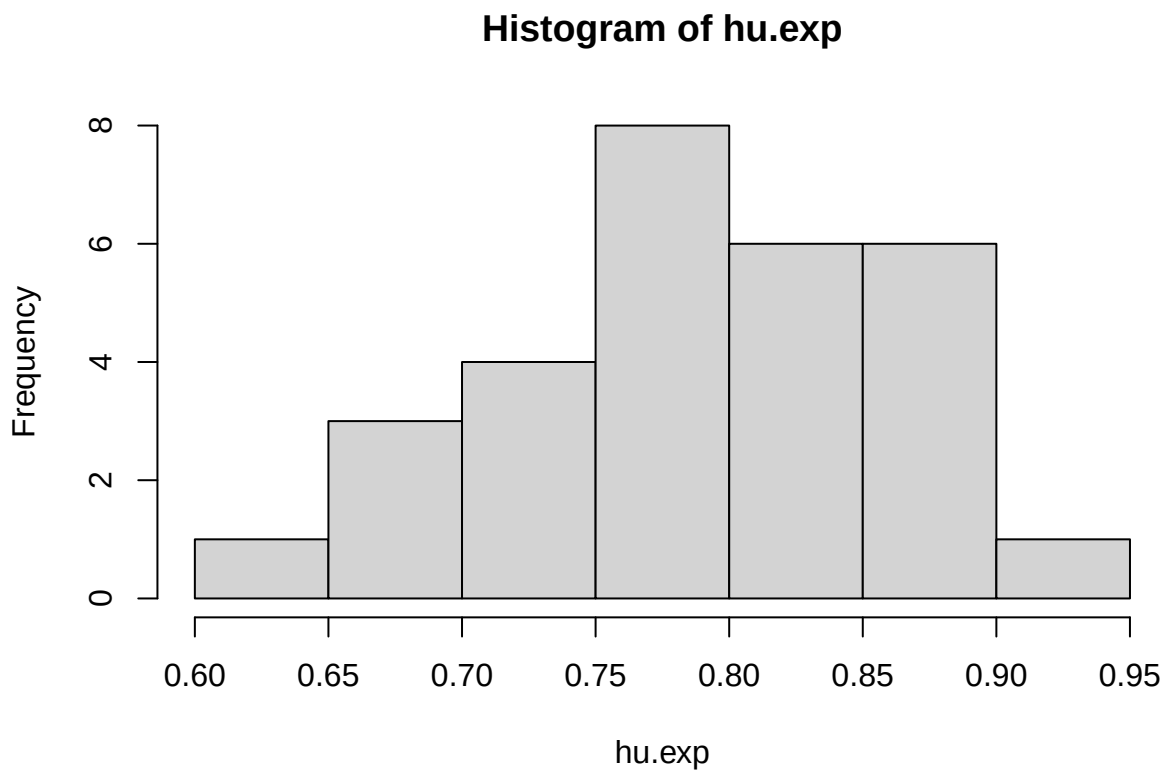
Ex 4 Make a table with the number of STRs for a given number of alleles and present a barplot of the number STRs in each category. What is the most common number of alleles for an STR?

Ex 5 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.

```
hu.exp <- c() # expected heterozygosity
index.1 <- seq(1,length(NistSTRs),2) # takes odd values 1 3 5 ...

for(i in index.1) {       # for-loop over columns of gendata
  STRi <- paste(NistSTRs[,i],NistSTRs[,i+1],sep="/") # concatenate
```

```
  NistSTRs.g <- genotype(STRi,sep="/") # separation is the space concatenation generates
  summary(NistSTRs.g)
  pX <- summary(NistSTRs.g)$allele.freq[,2] # retrieve frequency of all alleles
  pX
  val <- 1 # expected heterozygosity value
  for(i in pX) {
    val <- val - i^2
  }
  val
  hu.exp <- c(hu.exp, val)
}
hist(hu.exp)
```




```
mean(hu.exp)
```

```
## [1] 0.7904043
```

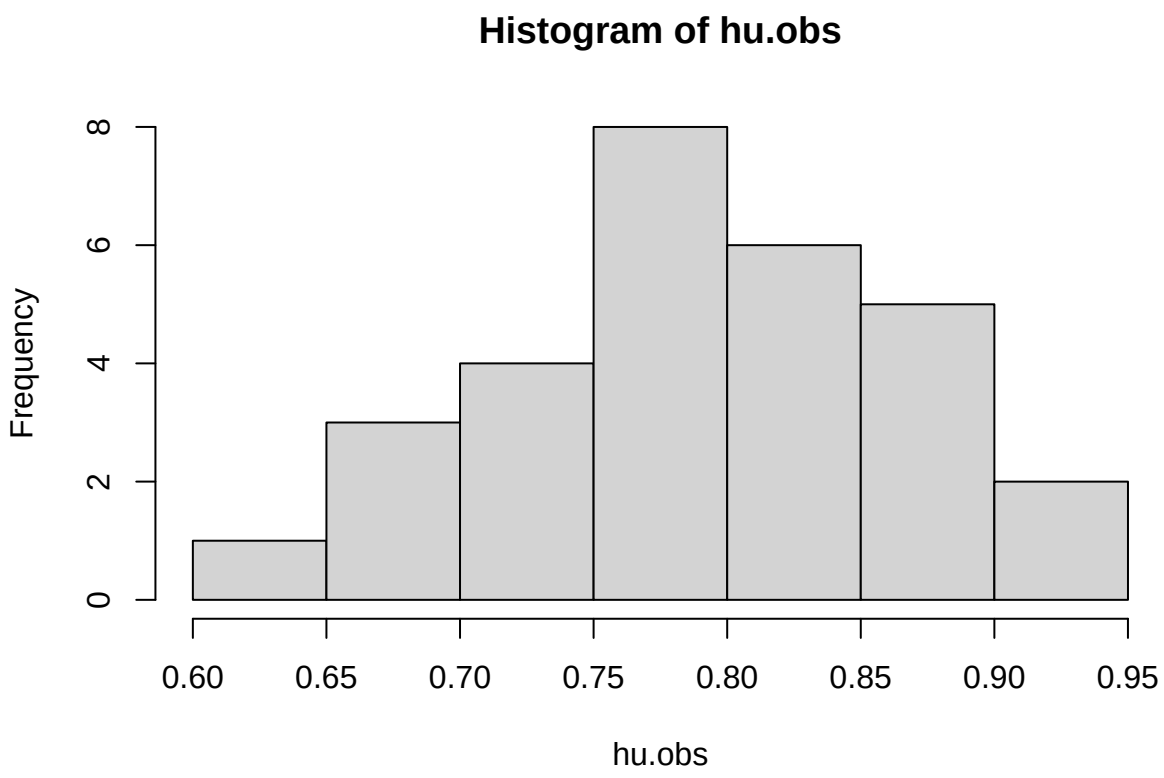
The average expected heterozygosity of the STR dataset is 0.79.

Ex 6 Calculate also the observed heterozygosity for each STR. Plot observed against expected heterozygosity, using all STRs. What do you observe?

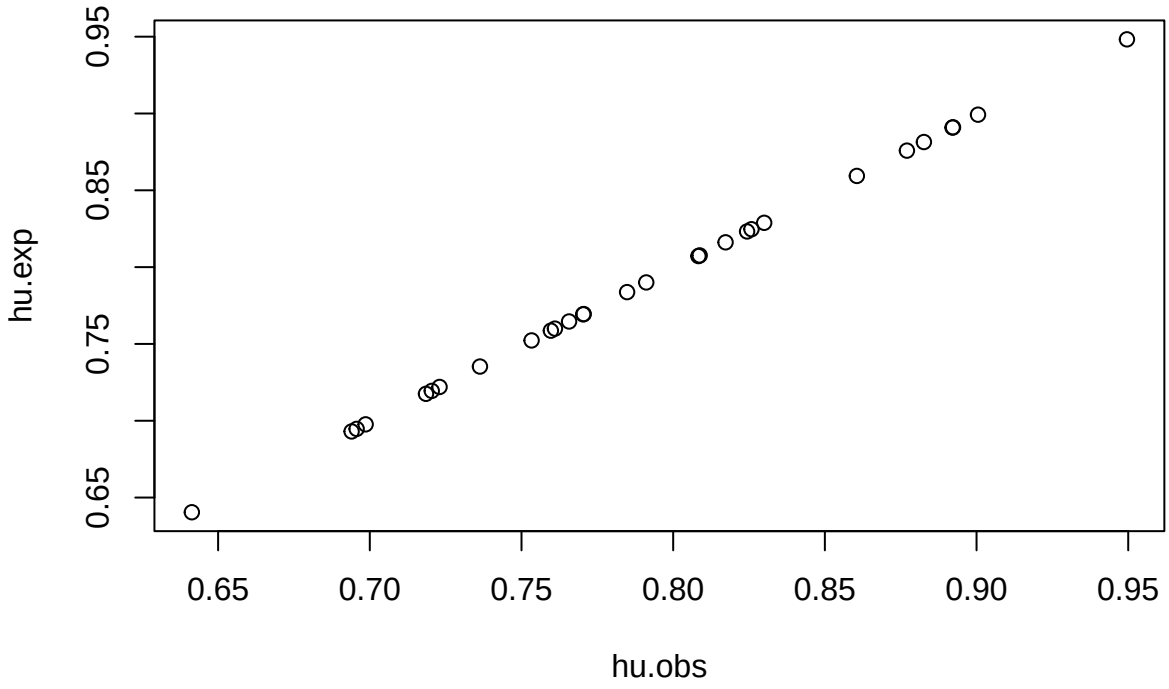
```
hu.obs <- c() # observed heterozygosity
index.1 <- seq(1,length(NistSTRs),2) # takes odd values 1 3 5 ...

for(i in index.1) {
  STRi <- paste(NistSTRs[,i],NistSTRs[,i+1],sep="/") # concatenate
  NistSTRs.g <- genotype(STRi,sep="/") # separation is the space concatenation generates
```

```
  hu.obs <- c(hu.obs, summary(NistSTRs.g)$Hu)
}
hist(hu.obs)
```




```
plot(hu.obs,hu.exp)
```




```
cor(hu.obs,hu.exp)
```

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

The observed and expected heterozygosities have a perfect correlation of R^2=1.

Ex 7 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?

SNP dataset has a more uniform heterozygosity, while the STR dataset follows a normal distribution, with a higher heterozygote value. This is a consequence of STRs having a faster mutation rate.