Practical 01 SG: Descriptive analysis of genetic markers

Carlos Moyano & Kleber Reyes

2022-11-07

SNP dataset

Questions about SNP dataset

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?

```
variants <- ncol(geneticData); variants # variants in the database
```

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

```
individuals <- nrow(geneticData)
perc.mis <- 100*sum(is.na(geneticData))/(variants*individuals); perc.mis # 0.1987%
```

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

```
# Let's assume values 1 is male and value 2 is female
perc.male <- 100*length(which(individualData$SEX == 1)) / individuals
perc.female <- 100*length(which(individualData$SEX == 2)) / individuals
perc.male; perc.female # 51.96% male - 48.04% female
```

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

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

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

```
cols <- which(colSums(geneticData == 1, na.rm = TRUE) > 0) # Non monomorphic (contains AB)
variants.poly <-length(cols); variants.poly # 18274 in db
```

```
## [1] 18274
```

```
variants.mono <- variants-variants.poly
perc.mono <- 100*variants.mono/variants; perc.mono # 11.50177
```

```
## [1] 11.50177
```

```
geneticData.poly <- geneticData[, cols]
```

5. Report the genotype counts and the minor allele count of polymorphism rs8138488_C, and calculate the MAF (Minor Allele Frequency) of this variant.

```
rs8138488_C <- dplyr::recode(geneticData.poly[, "rs8138488_C"], `0`="AA", `1`="AB", `2`="BB")
rs8138488_C.g <- genotype(rs8138488_C,sep="")
rs8138488_C.g.summary <- summary(rs8138488_C.g)
rs8138488_C.g.summary$genotype.freq
```

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

```
rs8138488_C.g.summary$allele.freq
```

```
##   Count Proportion
## A   129  0.6323529
## B    75  0.3676471
```

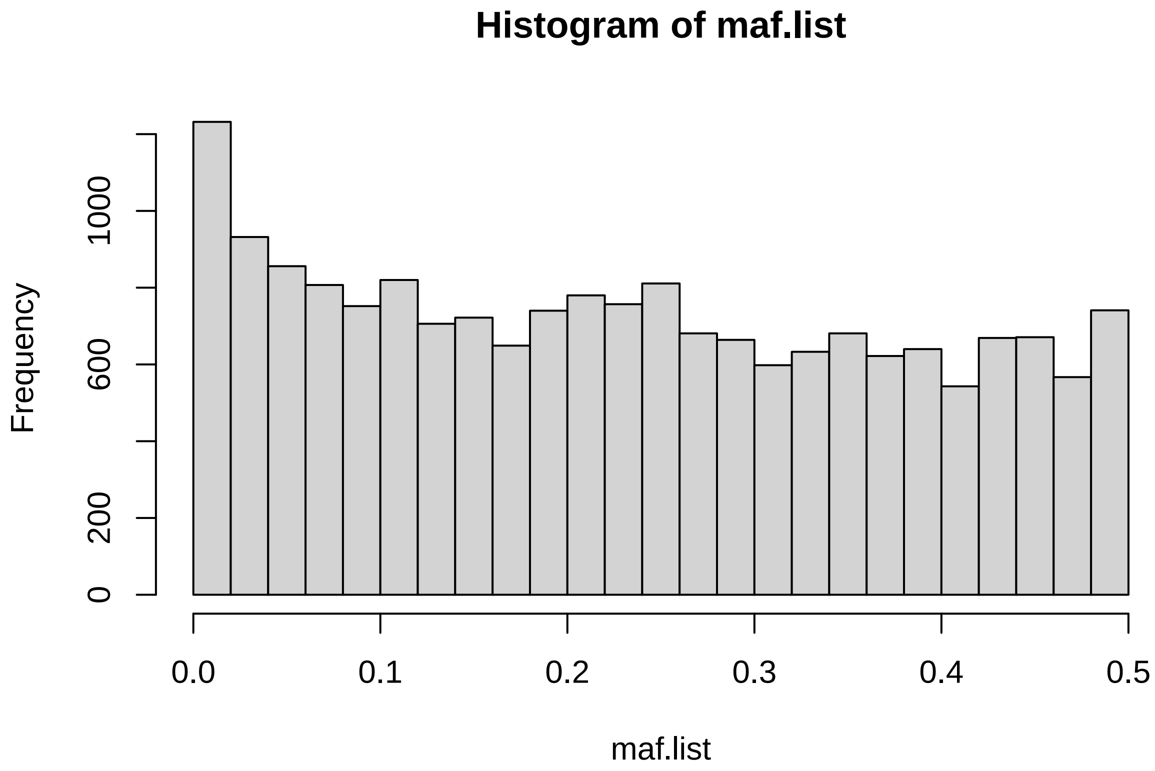
```
MAF = min(rs8138488_C.g.summary$allele.freq[,"Proportion"]); MAF
```

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

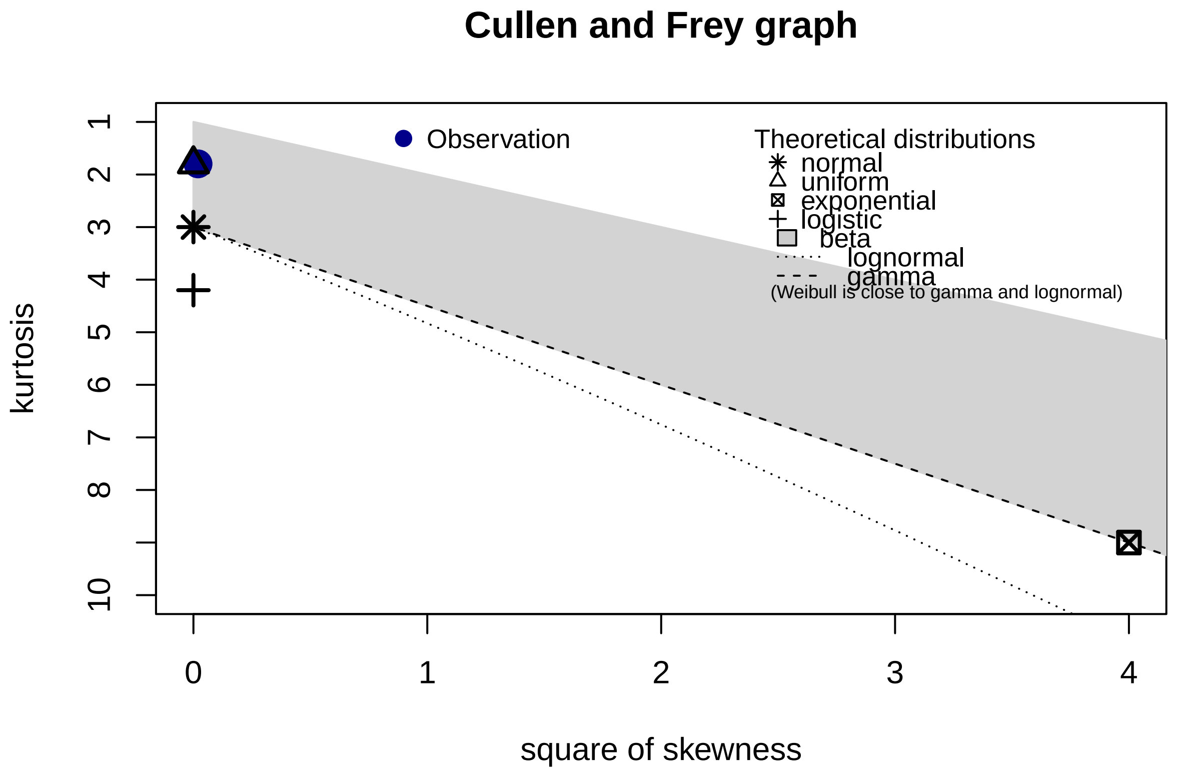
6. Compute the minor allele frequencies (MAF) for all markers, and make a histogram of it. Does the MAF follow a uniform distribution? What percentage of the markers have a MAF below 0.05? And below 0.01? Can you explain the observed pattern?

```
maf.list <- vector(mode="numeric", length=variants.poly)

for (i in 1:variants.poly) {
  variant <- dplyr::recode(geneticData.poly[, i], `0`="AA", `1`="AB", `2`="BB")
  variant.g <- genotype(variant,sep="")
  variant.g.summary <- summary(variant.g)
  maf.list[i] = min(variant.g.summary$allele.freq[,"Proportion"], na.rm = T)
}
```

```
hist(maf.list, breaks=20)
```

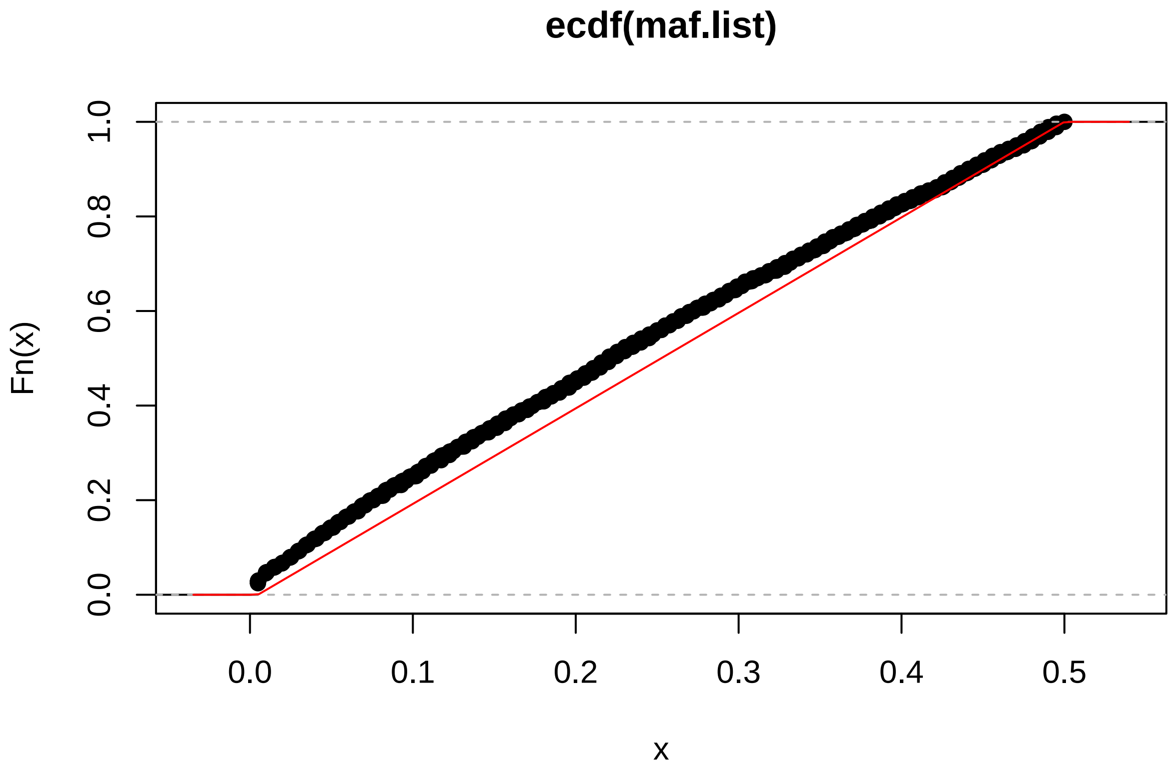


```
descdist(maf.list)
```



```
## summary statistics
## ------
## min:  0.004901961   max:  0.5
## median:  0.2205882
## mean:  0.2309362
## estimated sd:  0.1474513
## estimated skewness:  0.1407874
## estimated kurtosis:  1.797766
```

```
plot(ecdf(maf.list))
curve(punif(x, min(maf.list), max(maf.list)), add=TRUE, col="red")
```



```
maf.005 <- 100 * length(which(maf.list < 0.05)) / variants.poly; maf.005
```

```
## [1] 14.18409
```

```
maf.001 <- 100 * length(which(maf.list < 0.01)) / variants.poly; maf.001
```

```
## [1] 4.684251
```

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.

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

9. Compute for each marker its expected heterozygosity (He), where the expected heterozygosity for a bi-allelic marker is defined as 1 − E(from i=1 to k) pi^2 , where pi 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 He for this database?

STR dataset

Questions about STR dataset

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

```r
X <- NistSTRs
n <- nrow(X) # number of individuals
p <- ncol(X)/2 # number of STRs
n
```

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

```r
p
```

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

There are 361 individuals and 29 STRs.

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).

```r
# Function that determines the number of alleles for a STR.
n.alleles <- function(X, str.index) {
  allele.1 <- as.list(X[,str.index])
  allele.2 <- as.list(X[,(str.index+1)])
  return(length(table(unlist(c(allele.1, allele.2))))) # number of alleles
}

n.alleles.per.str.list <- list()
str.index <- 1
for (str.num in 1:p) {
  n.alleles.per.str.list  <- append(n.alleles.per.str.list, n.alleles(X, str.index))
  str.index <- str.index + 2
}
n.alleles.per.str <- unlist(n.alleles.per.str.list)
```

```
# Basic descriptive statistics of the number of alleles
mean(n.alleles.per.str)
```

```
## [1] 11.86207
```

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

```
## [1] 6.226236
```

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

```
## [1] 10
```

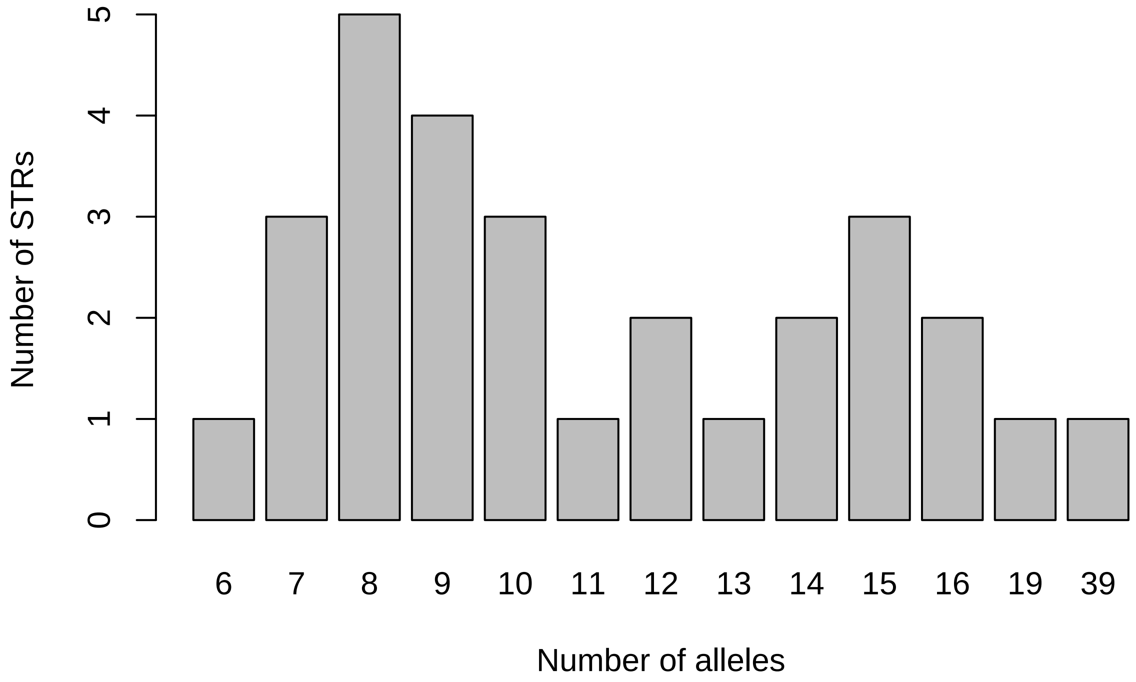
```
max(n.alleles.per.str)
```

```
## [1] 39
```

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

```
## [1] 6
```

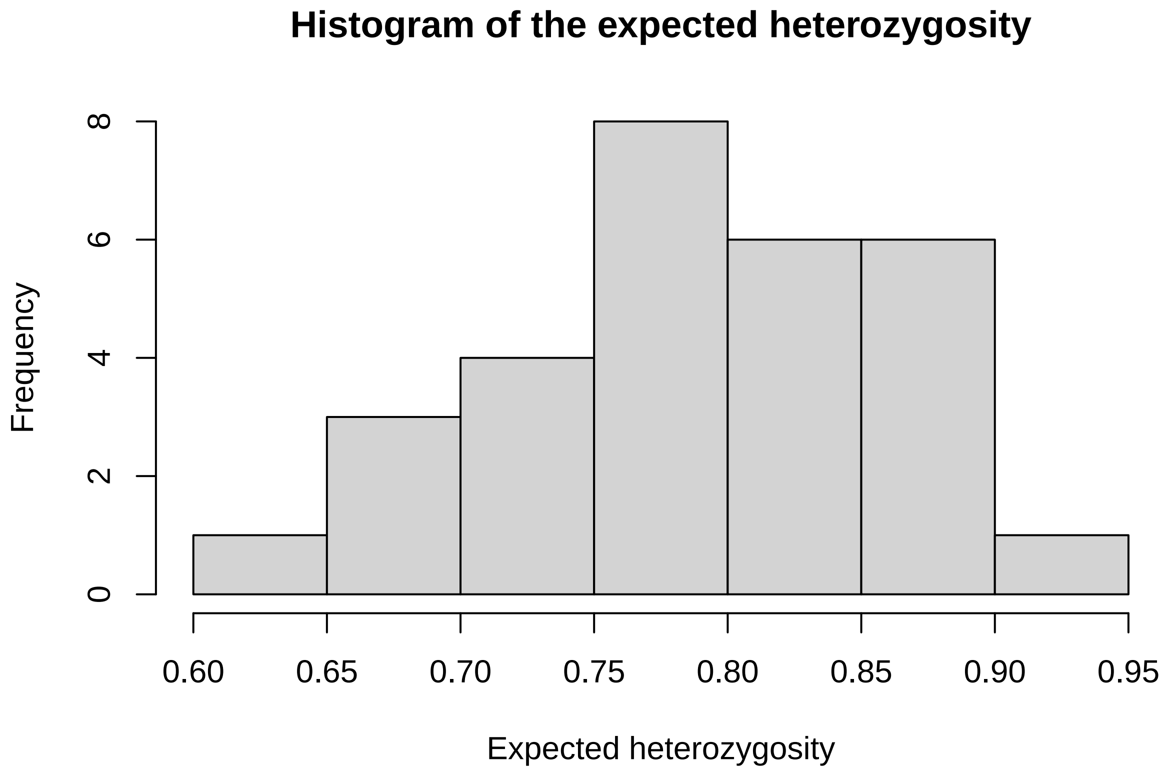
4. Make a table with the number of STRs for a given number of alleles and present a barplot of the number STRs in each category. What is the most common number of alleles for an STR?



The most common number of alleles for an STR is 8.

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.

```r
exp.heter <- function(X, str.index) {
  allele.1 <- as.list(X[,str.index])
  allele.2 <- as.list(X[,(str.index+1)])
  t <- table(unlist(c(allele.1, allele.2)))
  sum.t <- sum(unname(t)) # we sum the counts
  exp.heter <- round(1 - sum(sapply(unname(t), function(x) (x / sum.t)^2 )), 3)
  return(exp.heter) # expected heterozygosity formula
}

exp.heter.per.str.list <- list()
str.index <- 1
for (str.num in 1:p) {
  exp.heter.per.str.list  <- append(exp.heter.per.str.list, exp.heter(X, str.index))
  str.index <- str.index + 2
}
exp.heter.per.str <- unlist(exp.heter.per.str.list)

hist(exp.heter.per.str, xlab="Expected heterozygosity", main="Histogram of the expected heterozygosity")
```

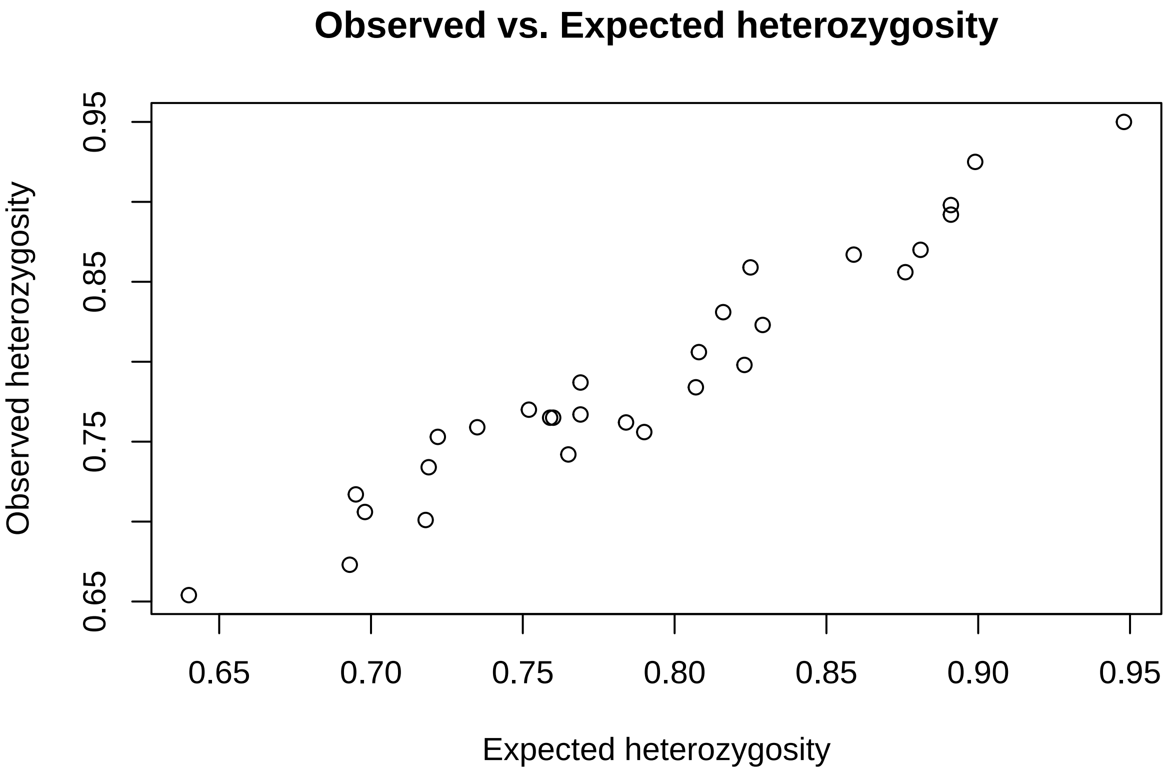


```
round(mean(exp.heter.per.str), 3) # average expected heterozygosity over all STRs
```

```
## [1] 0.79
```

6. Calculate also the observed heterozygosity for each STR. Plot observed against expected heterozygosity, using all STRs. What do you observe? (Ho = fAB)

```
obs.heter <- function(X, str.index) {

  allele.1 <- X[,str.index]
  allele.2 <- X[,str.index+1]
  allele.1n <- pmin(allele.1,allele.2)
  allele.2n <- pmax(allele.1,allele.2)

  index_different <- allele.1n != allele.2n

  individuals_heter <- paste(allele.1n[index_different], allele.2n[index_different],sep="/")
  individuals_heter

  individuals <- paste(allele.1n, allele.2n,sep="/")
  g.counts.sum <- sum(table(individuals))

  g.heter.counts.sum <- sum(table(individuals_heter))
  g.heter.counts.sum

  Ho <- round(g.heter.counts.sum / g.counts.sum, 3)

  return(Ho)
}

obs.heter.per.str.list <- list()
str.index <- 1
for (str.num in 1:p) {
  obs.heter.per.str.list  <- append(obs.heter.per.str.list, obs.heter(X, str.index))
  str.index <- str.index + 2
}
obs.heter.per.str <- unlist(obs.heter.per.str.list)
```



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?