Practical 1: Descriptive analysis of genetic markers

Krishna Kalyan

Working with SNP dataset

1. The file CHBChr2.rda contains genotype information (10.000 SNPs) of 45 individuals of a Chinese population of unrelated individuals. Load this data into the R environment.

```
load('CHBChr2.rda')
```

2. Recode all "NN" genotypes as missing values (NA). What percentage of the data is missing?

```
## [1] 36.428
```

3. For how many SNPs the genotype information is completely missing? Remove these SNPs from the database.

```
## [1] 2975
```

4. What is, on the average, the percentage of missing information per individual, after fully missing SNPs have been removed?

```
## [1] 9.50605
```

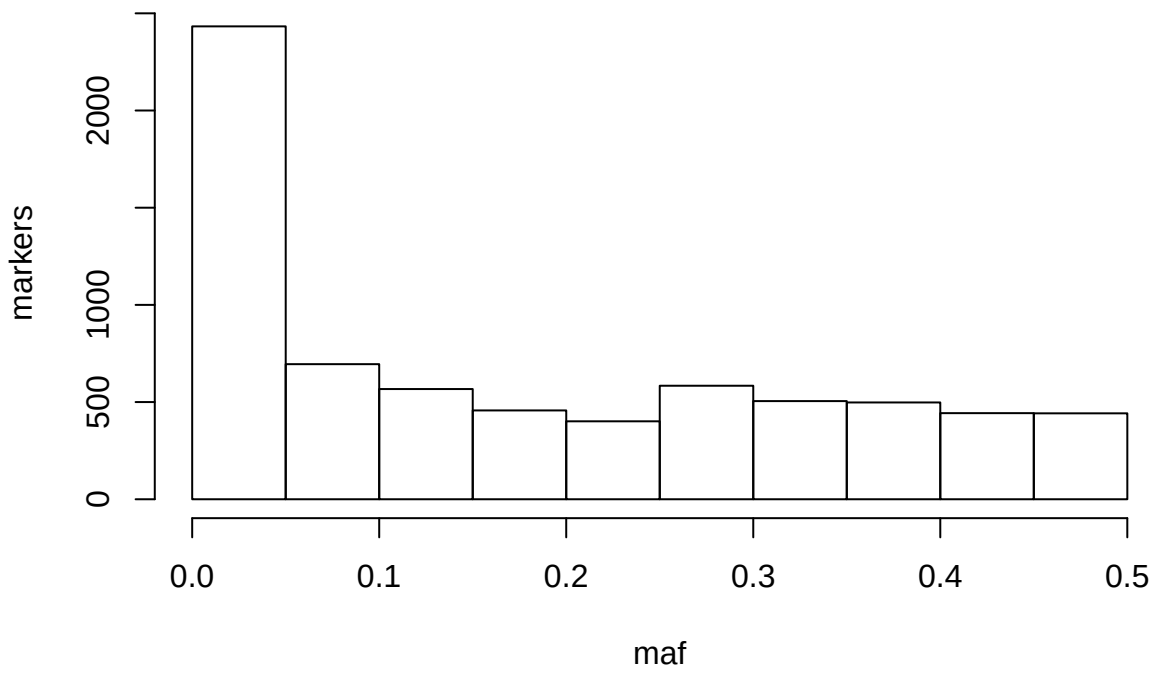
5. How many markers are monomorphic?

```
## [1] 1860
```

6. Write a function to compute the minor allele frequency. Make sure the function also produces sensible answers for markers that consist of missing values only, or markers that are monomorphic. Include the source code of your function here.

```
maf <- function(x){
  if(mean(is.na(x))==1){
    return(NA)
  }
  alleles <- table(unlist(strsplit(as.character(unlist(unname(x))),split="")))
  af <- alleles/sum(alleles)
  min.maf <- min(af)
  if(min.maf == 1){
    return(0)
  }else{
    return(min.maf)
  }
}
```

7. Compute the minor allele frequencies for all markers, and make a histogram of it.



8. What percentage of the markers have a maf below 0.05? And below 0.01?

```
mean(maf < 0.05) * 100
```

```
## [1] 34.51957
```

```
mean(maf < 0.01) * 100
```

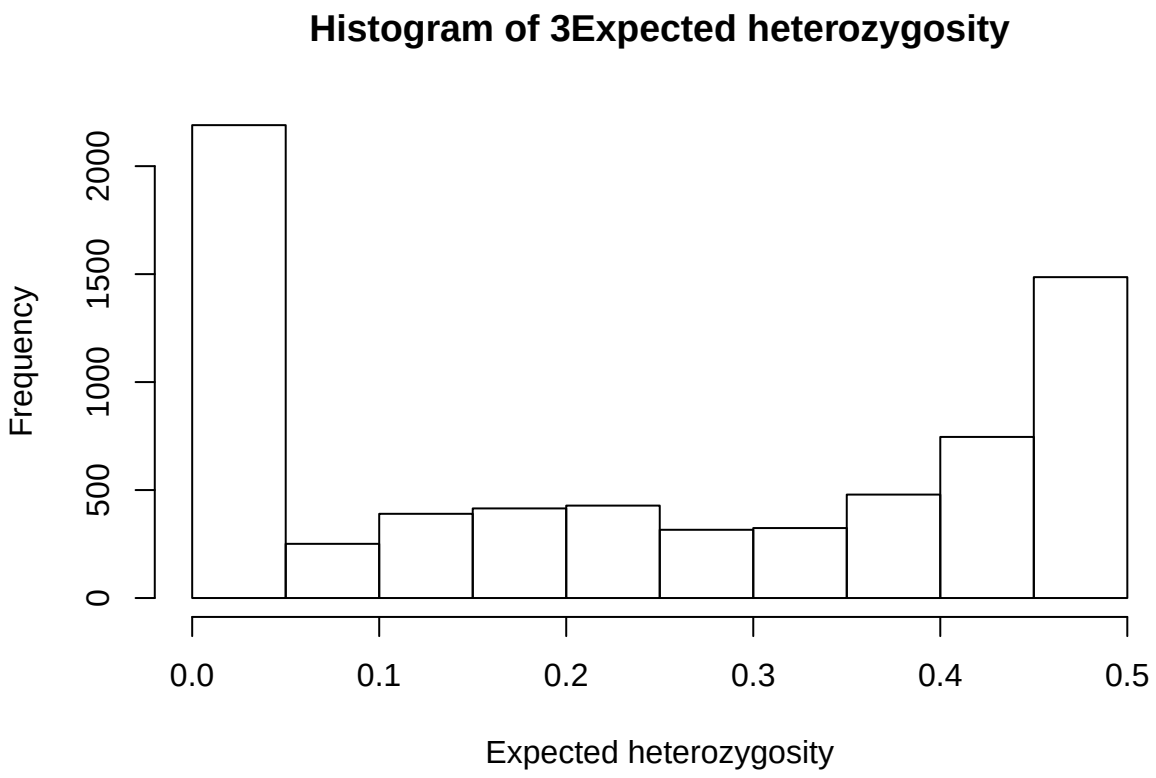
```
## [1] 26.03559
```

9. Compute for each marker its expected heterozygosity, where the expected heterozygosity for a bi-allelic markers is defined as 1 Pk i=1 p2i, where pi is the frequency of the ith allele. Compute the average expected heterozygosity over all markers. Make a histogram of the expected heterozygosity.

```
heterozygosity <- function(x){
  alleles <- table(unlist(strsplit(as.character(unlist(unname(x))),split="")))
  af <-  alleles/sum(alleles)
  het <- 1- sum(af^2)
  return(het)
}
heter <- apply(X.clean,2,heterozygosity)
mean(heter)
```

```
## [1] 0.2353208
```

```r
hist(heter , main= "Histogram of 3Expected heterozygosity", xlab = "Expected heterozygosity")
```



Working with STR dataset

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

```r
data = read.table("FrenchSTRs.dat")
dim(data)
```

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

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

```
## [1] "Individuals 29"
```

```
## [1] "STRs  678"
```

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

```
## [1] 4.206083
```

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).

```
n.alleles <- function(x) {
  y <- length(unique(x[!is.na(x)]))
  return(y)
}
alleles.str = as.data.frame(apply(X, 2, n.alleles))
names(alleles.str) = c('alleles')
lapply(alleles.str, mean)
```

```
## $alleles
## [1] 6.374631
```

```
# sd
lapply(alleles.str, sd)
```

```
## $alleles
## [1] 1.823385
```

```
# medain
lapply(alleles.str, median)
```

```
## $alleles
## [1] 6
```

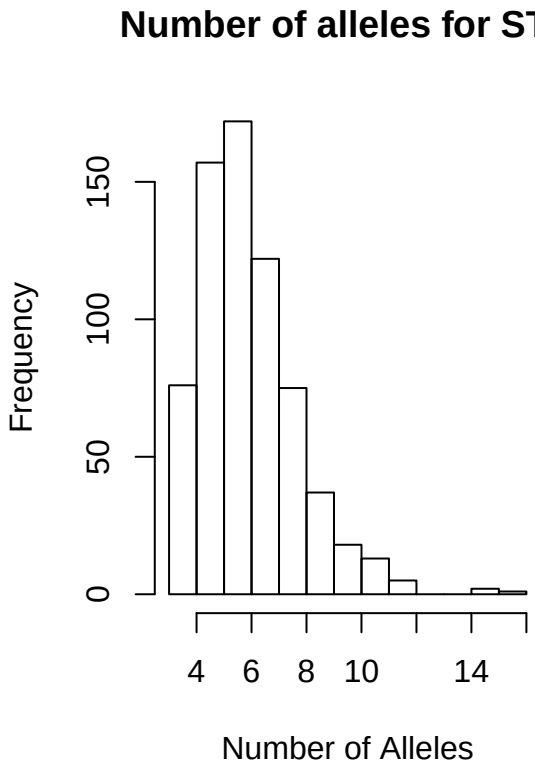
```
# max
lapply(alleles.str, max)
```

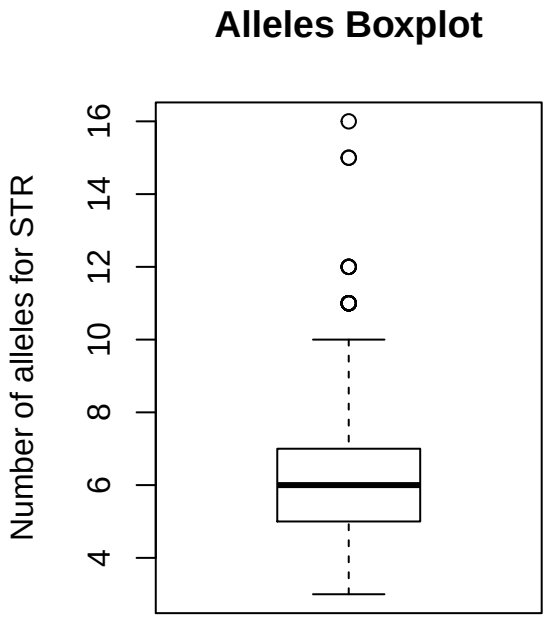
```
## $alleles
## [1] 16
```

```
# min
lapply(alleles.str, min)
```

```
## $alleles
## [1] 3
```

5. Make a boxplot and a histogram of the number of alleles per STR. What is the most common number of alleles for an STR?



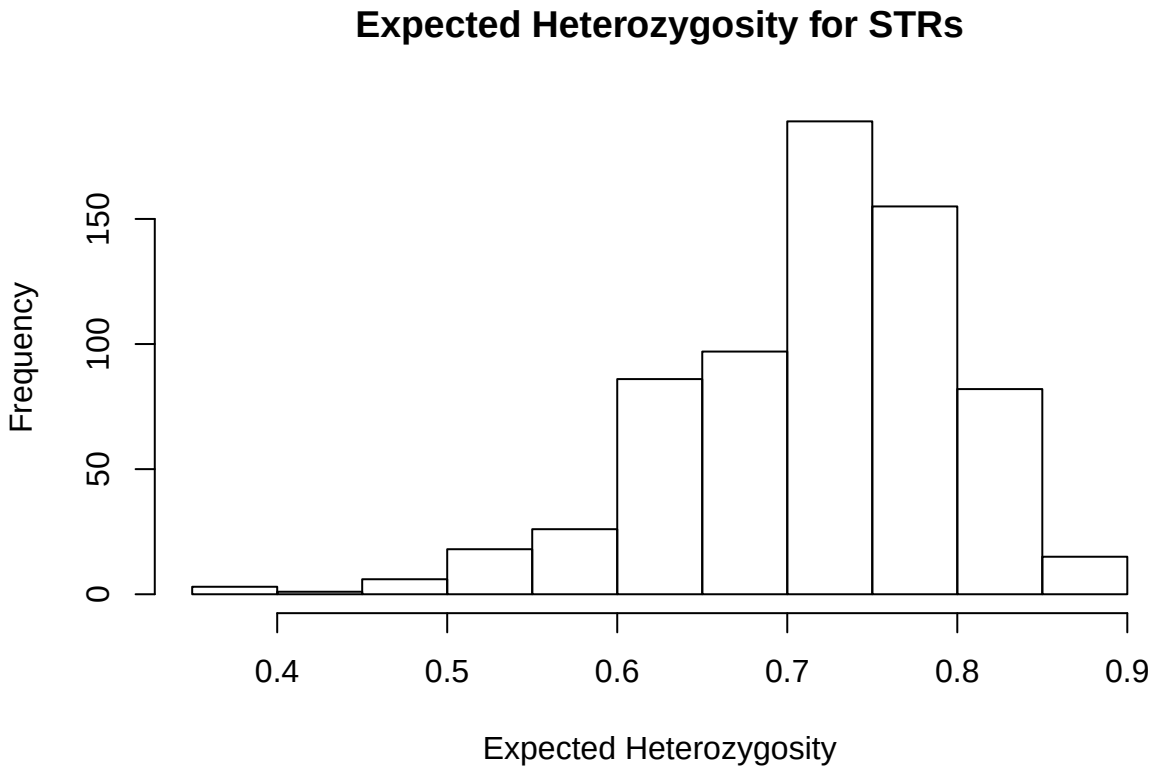


6. Compute the expected heterozygosity for each STR. Make a histogram of the expected het-erozygosity over all STRS. Compute the average expected heterozygosity over all STRs.

```
heterozygosity <- function(x){
  freq <- table(x)/sum(table(x))
  1 - sum(freq^2)
}
allels.het = apply(X, 2, heterozygosity)
mean(allels.het)
```

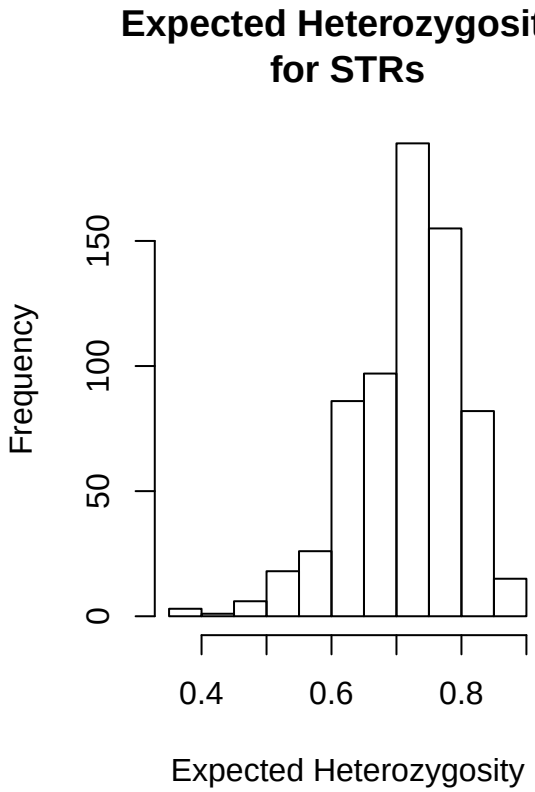
```
## [1] 0.7172662
```

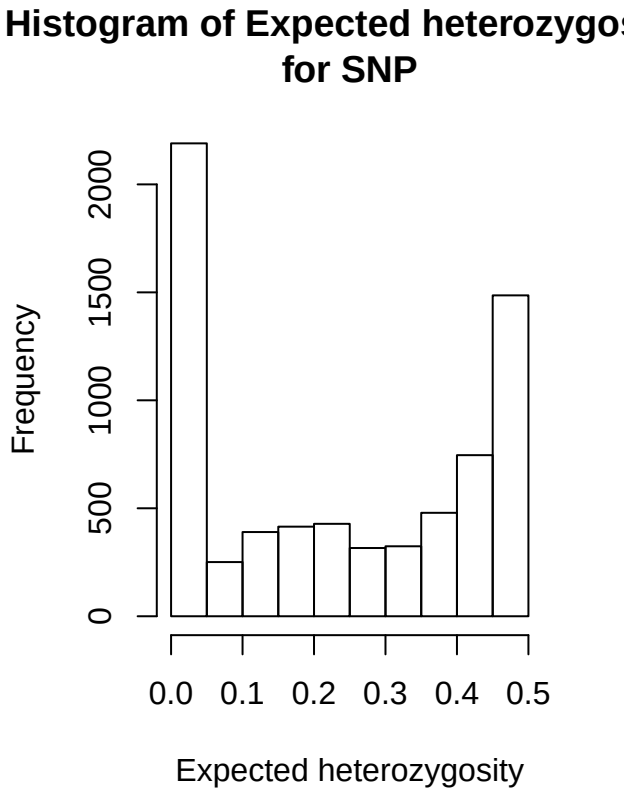
```
hist(allels.het,main="Expected Heterozygosity for STRs", xlab="Expected Heterozygosity")
```



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?

```
par(mfrow=c(1,2))
hist(allels.het,main="Expected Heterozygosity \nfor STRs", xlab="Expected Heterozygosity")
hist(heter , main= "Histogram of Expected heterozygosity \nfor SNP", xlab = "Expected heterozygosity")
```





We can compare two plots above we see that expected heterozygosity for STR range between 0 and 0.9, where as for SNP the range is between 0 and 0.5. The frequency value for SNP tends to be larger (0-2000) than STRs(0-200).