Bioinformatics and Statistical Genetics	Descriptive analysis of genetic markers
Jan Graffelman	November 15, 2017

Name: Sten-Oliver Salumaa

Name: Denis Kovalenko

Perform the computations and make the graphics that are asked for in the practical below. You can write your answers on this sheet, and attach graphics at the end. Give each graphic a title, and clearly label *x* and *y* axes. Send your solution to jan.graffelman@upc.edu no later than 23rd of November 2017. You can make use of the R-package **genetics** (and other packages) to compute your answers. The first part of the practical is dedicated to the descriptive analysis of SNP data, whereas the second part is dedicated to the analysis of STR data. The datasets can be downloaded by clicking on their file names given below.

1 SNP dataset

1. The file JPT22.rda contains genotype counts in a generic notation (AA,AB,BB) for genetic variants on chromosome 22 of 104 individuals of a Japanese population of unrelated individuals. This data has been extracted from the 1000 genomes project at www.internationalgenome.org/.

The datafile contains a dataframe with the following columns: the RS identifier of the variant, rs, the position of the variant in base pair units, pos, the genotype counts for males, (mAA, mAB, mBB) and for females (fAA, fAB and fAB), the total genotype counts (AA, AB, BB) and the number of missing genotype results for the variant, nmis. Load this data into the R environment.

2. (1p) How many variants are there in this database?

   1100472

   What percentage of the data is missing?

   0.012 %

   How many individuals in the database are males and how many are females?

   48 women, 56 men

3. (1p) For how many SNPs the genotype information is completely missing?

   1

   What is the average percentage of missings per variant?

   1.2678 %

   Delete, for all posterior calculations, those variants that have 50% or more missing values. How many variants remain in your database?

   1100403

4. (1p) 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?

   173611

5. (1p) Write a function to compute the minor allele frequency on the basis of a vector of genotype counts. 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.

   R handles cases with all NAs (returns NA).

```r
maf_fun <- function(x){
  prob <- (x[1]+x[2]/2)/sum(x)
  res <- min(prob, 1-prob)
  return(res)
}
```

6. (1p) Compute the minor allele frequencies (MAF) for all markers, and make a histogram of it. Does the MAF follow a uniform distribution?

   No, the distribution is exponentially diminishing.

   What percentage of the markers have a maf below 0.05?

   48.7135 %

   And below 0.01?

   31.2509 %

7. (1p) One might expect that the minor allele in males is the same as the minor allele of the females. For what percentage of the variants is this not the case?
   9.2857 %

8. (2p) Calculate the minor allele frequency for males and for females and present a scatterplot of these variables. Calculate and report their correlation coefficient.

   Correlation: 0.9641

9. (2p) Compute for each marker its **expected heterozygosity**. Compute the average expected heterozygosity over all markers. Make a histogram of the expected heterozygosity.

   Histogram in the last part of the paper under 'Part 1 Exercise 9'

2 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.

2. (1p) How many individuals and how many STRs contains the database?
There are 29 individuals - 58 rows, 2 for each individual ID.
678 STRs - there are 679 columns in total, 1 of which is for individual's ID

3. (1p) The value −9 indicates a missing value. Replace all missing values by NA. What percentage of the total amount of datavalues is missing?

4.2%

4. (2p) 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).

Mean - 6.375, standard deviation: 1.823385, median - 6, minimum - 3, maximum - 16.

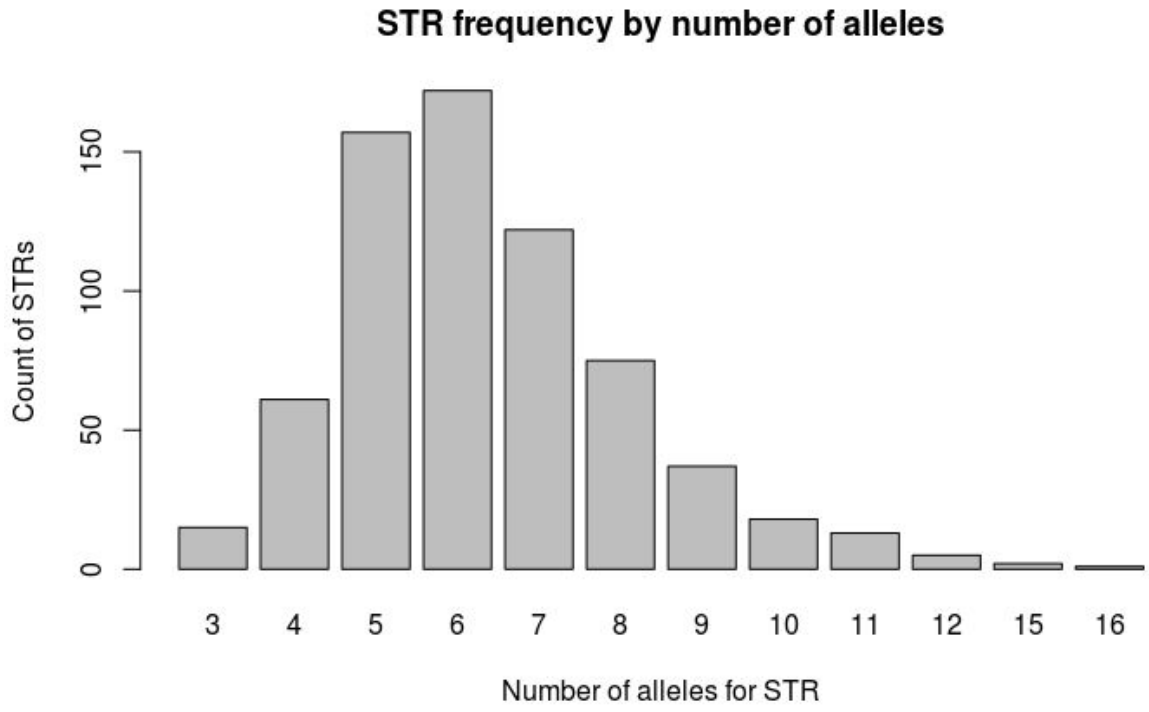
5. (2p) Make a table with the number of STRs for a given number of alleles and present a barplot of the number STRs in each category. What is the most common number of alleles for an STR?

Table with with frequencies by number of alleles.

STR_COUNT	freq
3	15
4	61
5	157
6	172
7	122
8	75
9	37
10	18

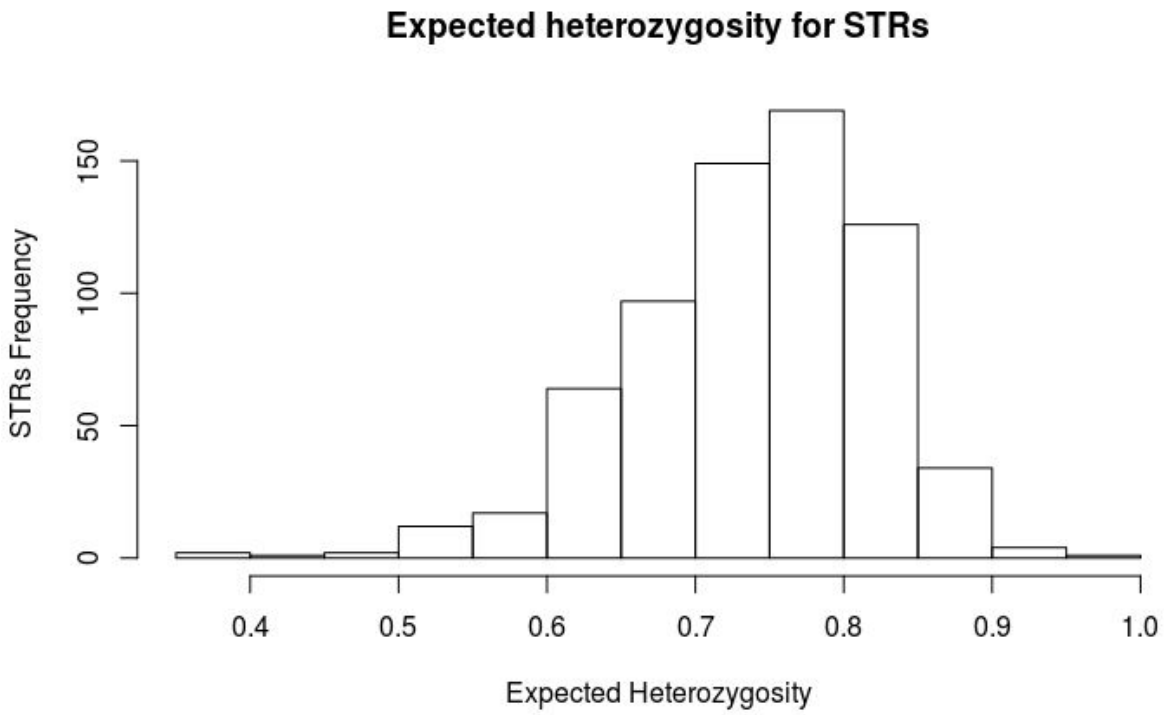
11	13
12	5
15	2
16	1

Bar plot of frequencies of STR grouped by number of alleles



6 is most common number of alleles in this dataset

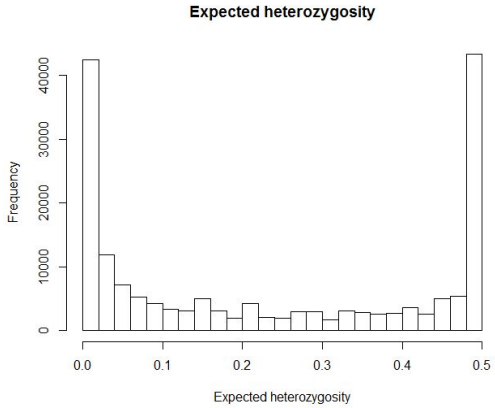
6. (2p) Compute the expected heterozygosity for each STR. Make a histogram of the expected heterozygosity over all STRS. Compute the average expected heterozygosity over all STRs.

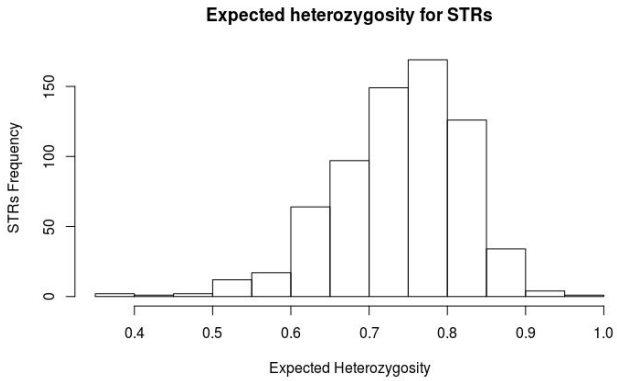


Average expected heterozygosity: 0.7393344

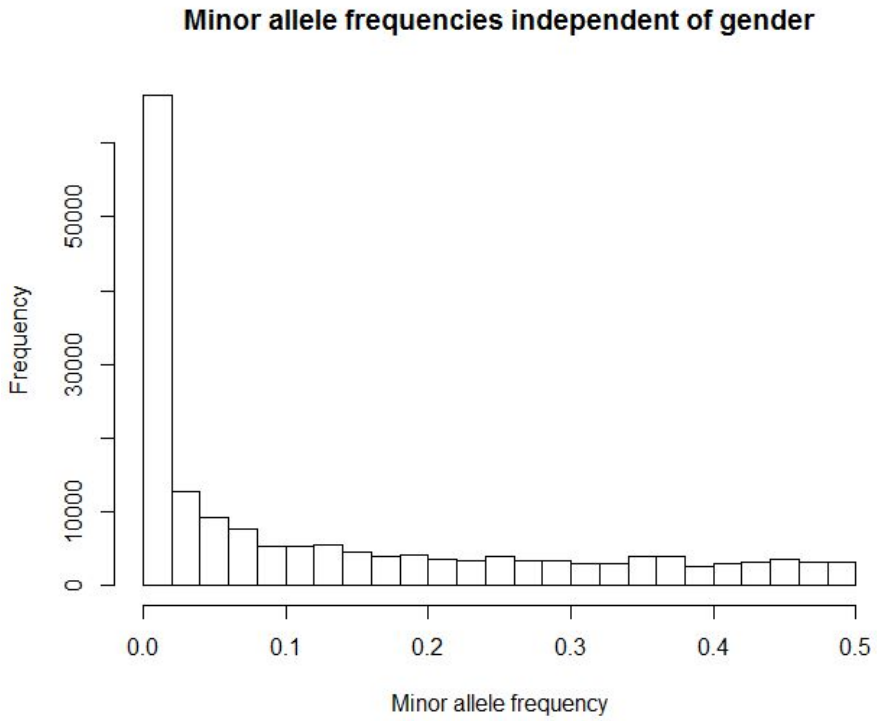
7. (2p) Compare the results you obtained for the SNP database with those you obtained for the STR database. What differences do you observe between these two types of genetic markers?

STR have more alleles on average, so considering expected heterozygosity, when comparing 2 charts we can see that for SNP majority of has either low heterozygosity (which means that for this SNP one allele is highly dominant by frequency), or very high, which means that each individual allele has low frequency in dataset. For STR, on the other hand, majority is in the "middle" - few alleles have high frequency, and others relatively low.

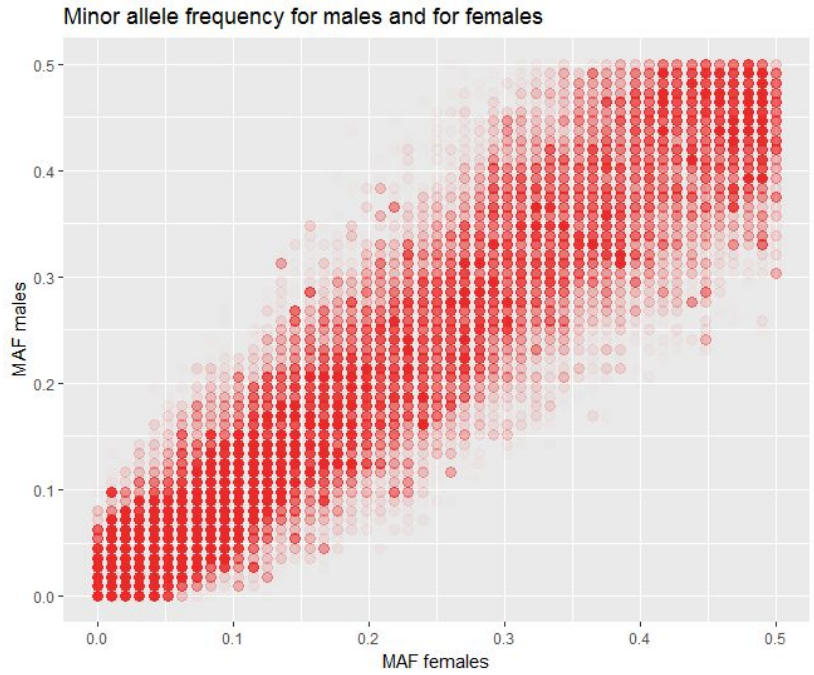

Expected heterozygosity
Frequency
0
10000
20000
30000
40000
0.0
0.1
0.2
0.3
0.4
0.5
Expected heterozygosity


Expected heterozygosity for STRs
STRs Frequency
0
50
100
150
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Expected Heterozygosity

Part 1 Exercise 6



Part 1 Exercise 8



Part 1 Exercise 9