Practical on allele and genotype frequencies

Consider the following data set of 40 DNA sequences of 10kbp from 20 brown bears. The DNA data is represented in the .csv called `bears.csv` where each row corresponds to an individual chromosome (one copy of one individual) and each column to a genomic position in the 10kbp locus. Each individual as two chromosomes (copies), meaning that individual 1 has data on rows 1 and 2, individual 2 has data on rows 3 and 4, and so on.

From this data, you want to make some inferences about how variable the genome of brown bears is and whether it is an inbred population or not. For doing this, you need to calculate several quantities of population genetics.

Write a script in R that complete all the following tasks:

1. identify which positions are SNPs (polymorphic, meaning that they have more than one allele)
2. calculate, print and visualise allele frequencies for each SNP
3. calculate and print genotype frequencies for each SNP
4. calculate (observed) homozygosity and heterozygosity for each SNP
5. calculate expected genotype counts for each SNP and test for HWE
6. calculate, print and visualise inbreeding coefficient for each SNP deviating from HWE

Ideally, you would like to implement a single R function that takes some genomic data (e.g. from a .csv file) and performs all these tasks. However, this is not required. Some of the functions and lines implemented here will be used for the next practicals.

Some considerations for each task:

1. only mono- or di-allelic sites are present in this data set; after you identify which site are SNPs, it may be convenient to reduce the data set to only the polymorphic positions (ignore the monomorphic sites); to read the data you can use the R function `read.csv` making sure you specify the option `stringsAsFactors` as FALSE and `colClasses` as a vector of characters (otherwise the T nucleotide will be read as a boolean variable), as in `data <- read.csv("bears.csv", stringsAsFactors=F, header=F, colClasses=rep("character", 10000))`

2. you may want to choose a "reference" allele for each SNP and calculate frequency for this allele;

3. for calculating genotype frequencies, remember that each individual has two chromosomes, so two rows: e.g. the first individual has data on rows 1 and 2, the second on rows 3 and 4 etc etc; at each SNP you can detect the two alleles and then count each possible genotype with regard to these two alleles;

4. you can reuse part of the code for point 3 to quickly calculate the observed heterozygosity and homozygosity;

5. you can reuse code for points 2 and 3; in R you can obtain a p-value for a chi^2 test with one degree of freedom with `1-pchisq(chi, df=1)` with `chi` being the statistic calculated from observed and expected genotype counts;

6. you should reuse most of the code for point 5 with the addition of the calculation of inbreeding coefficient for SNPs deviating from HWE.

Is this population inbred or not? Do you have enough data to provide a definite answer in your opinion?