Group work 1

The file SNPdata is a tab delimited file that you can download from moodle.
It has chromosome_position (e.g. “chr1_11169676”) as row names, whereas on the columns you can read the IDs of the subjects: 1200 patients in column 1:1200, and 800 control subjects in column 1201:2000.
In the file

- 0 indicates homozygous genotype AA (0 mutated alleles)
- 1 indicates heterozygous genotype Aa (1 mutated allele)
- 2 indicates homozygous genotype aa (2 mutated alleles)

You are asked to submit an R script which implement 3 functions described in the next slides.

NOTE:

- the name of the function and of the name of the input parameters is MANDATORY. Do not change it!
- The R script must not contain any other code but the 3 functions

qcalculation

Use the indicated names for the function and the parameters
This is mandatory and you will be penalized otherwise

The first function:

```
qcalculation <- function(SNPdata) {
    BODY OF THE FUNCTION
}
```

1. Take as input a numeric data matrix that is supposed to have the same format of the genetic data provided in stem

2. Calculates the minor allele frequency q for each SNP

3. Return the minor allele frequency of each SNP as a vector of numeric values with names corresponding to the SNP IDs (chromosome_position, e.g. “chr1_11169676”) with the same order they had in the input matrix

Suggestions:

- *It might be useful to use the function table() and to convert it in a data.frame().*
- *Sometime you might have 0 subject with genotype aa… consider this possibility*

HWEtest

Use the indicated names for the function and the parameters
This is mandatory and you will be penalized otherwise

The second function:

```
HWEtest <- function(SNPdata) {
    BODY OF THE FUNCTION
}
```

1. Take as input a <u>numeric data matrix</u> that is supposed to have the same format of the genetic data provided in stem

2. Compute a <u>HWE test for each SNP</u> given as input
   - By calculating the χ^2_{obs} from the data
   - By computing the p value using the function pchisq (DO NOT use directly the chisq.test() function)

3. Return the HWE test p-values for each SNP as a vector of numeric values with names corresponding to the SNP IDs (chromosome_position, e.g. "chr1_11169676") with the same order they had in the input matrix

Suggestion: be careful when you use pchisq(). The probability it gives as output by default is $P[X \leq \chi^2_{obs}]$

VARIANTanalysis (1/2)

Use the indicated names for the function and the parameters
This is mandatory and you will be penalized otherwise

The third function:

```
VARIANTanalysis <- function(filepath, indCTRL, MAFth=0.01, HWEalpha=0.01) {
    BODY OF THE FUNCTION
}
```

1. Take as input
   - a file name (entire path). The file is supposed to have the same format of the genetic data provided in stem
   - a vector of indexes indicating in which columns the input file has data from control subjects
   - a threshold to filter SNPs with lower minor allele frequency (default 0.01)
   - a significance level alpha to be used to filter SNPs with lower p-values because possibly not in HWE (default 0.01)

2. The function VARIANTanalysis() read the file and analyse the different variants
   - filtering out those with (MAF < MAFth) OR (HWE-p-value < HWEalpha)
   - calculating the χ^2_{obs} from the data
   - computing the p value using the function pchisq (DO NOT use directly the chisq.test() function)

VARIANTanalysis (2/2)

Use the indicated names for the function and the parameters
This is mandatory and you will be penalized otherwise

The third function:

```
VARIANTanalysis <- function(filepath, indCTRL, MAFth=0.01, HWEalpha=0.01) {
    BODY OF THE FUNCTION
}
```

3. Compute the q-value for each SNP using the Benjamini-Hockberg procedure

4. Return a matrix with chromosome_position (e.g. “chr1_11169676”) as row names (not all the SNPs will be given as output but only those passing the MAF and HWE filter) and 8 columns with names c("AA_ctrl","Aa_ctrl","aa_ctrl","AA_case","Aa_case","aa_case","pval","qval"). In each row, i.e. for each SNP, the matrix reports the number of occurrences of each genotype for controls and cases respectively (first six columns), the p-value an the q-value.

Suggestion:
Remember that the HWE test should be applied to controls only
be careful when you use pchisq(). The probability it gives as output by default is $P[X \leq \chi^2_{obs}]$

I will test your solution

Only one student from each group has to submit the code on behalf of his/her mates as representative of the group.

I will test your code with different inputs… this is not a programming course so a valid submission means that your code must be submitted fully working… I am not going to correct coding and programming bugs…
Therefore, before submitting it, try running your code placing the input files in a directory different from the working directory and cleaning the workspace before running your functions.

To assign a grade I will evaluate if the MAF calculation, the HWE test and the association test perform correctly and give appropriate an reasonable results