Mid-term exam 1

The data set for your exam is located in a Github repository. It is part of the results of a transcriptomic study that addressed how genes in the brain express differently in macaques, bonobos and humans. The data set contains the selection of genes that were considered human-specific according to their differential patterns of gene expression in different brain regions and cell types. The file is named `human_specific.csv`.

Run the following code to download your data set to your current working directory:

```
download.file("https://raw.githubusercontent.com/clauw87/IB_practicals_R/main/human_specific.csv",
destfile = "human_specific.csv")
```

Exercise 1

1a. Import the data set into an R data frame stored in a variable called `data`. (you can previously use the function *readLines()* with argument n=2, to figure out the separator and whether there is a header or not, and the function *head()* afterwards to verify that you have imported your data set properly)

1b. Explore your data frame to see how many columns, and how many rows does it contain and what types of variables are stored in the columns?

1c. What different tissues (`tissue`) and what different cell types (`cell_type`) are represented in the data set?

Exercise 2A

The variable avg_logFC (log fold change) is a symmetrically transformed variable that quantifies the estimated level of human specific fold change expression of a gene: this is, how much more (+ values) or less (− values) is the gene expressed in humans compared to the average of the two non-human primates. The genes in this data set are those considered to have human specificity and account in part for what makes us humans.

2a. Try the following plot. It is called a vulcano plot and shows how the genes in a transcriptomic study distribute according to the magnitude and the direction of their expression change (`avg_logFC`) and the significance of the test to consider them as significantly changed (`p_val_adj`).
You will see that there is a gap in the middle of the vulcano and a clear-cut base; that corroborates that the data set includes only the significant results, for which the researchers have put certain thresholds on `avg_logFC` and `p_val_adj` for what they regard as a reliable change in gene expression to consider a gene change "human-specific".

```
plot(data$avg_logFC, -log10(data$p_val_adj))
```

2b. Explore the distribution of the values of the variables `avg_logFC` and `p_val_adj` in the data set graphically with histograms and numerically with the five-number summary.

2b-1. Can you figure out/ approximate what are the thresholds mentioned in 2a.?

2b-2. Do human-specific changes in gene expression comprise more gene over-expression (`avg_logFC` > 0), gene under-expression (`avg_logFC` < 0), or are similarly distributed on both directions?

2c. What is the most over-expressed gene in the data set (the one with the highest positive `avg_logFC`)? Print its HUGO symbol (hgnc_symbol).

2d. Some genes might have resulted significantly human-specific in just one or in more than one tissue and/ or cell types. How many total (different) human-specific genes (`gene`) did the study find in total?

Exercise 2B

The astrocytes in the Caudate Nucleus may have played a role in adding flexibility in functions that determined our differentiation from other primates: for things like the planning, learning and memory, motivation, emotion and romantic interaction. In the data set Caudate Nucleus (`tissue`) is coded as "CN" while astrocytes (`cell_type`) is coded as"Ast".

2e. Create a new data frame called `cnast` to store only the subset of data corresponding to the Caudate nucleus astrocytes by subsetting the data frame based on those two conditions on the columns `tissue` and `cell_type`, respectively.

2f. How many human-specific genes were found in this specific combination of brain tissue and cell type?

2g How many human-specific genes are over-expressed (avg_logFC > 0), and how many are under-expressed (avg_logFC < 0) in the Caudate Nucleus astrocytes?

2h Create a new column in `cnast` data frame called `direction` that takes value "up" if the gene is over-expressed and takes value "down" if the gene is under-expressed.

Exercise 3

The results in the data set show that in this study there were found 8 human-specific genes from the mitochondrial chromosome (MT) out of the total of 1268 human-specific genes found.

Knowing that the total of genes in mitochondria that they could study was 37 and the total of genes from all chromosomes was 16000:

3a. How likely would it be to get 8 or more mitochondrial genes in a random sample of 1268 genes from such those 16000 genes. Does the result give you reasons to suspect that the mitochondrial genes in the tissues and cell types studied must have had a specially important role for human specificity?

Knitting your exam

Run the following code by copying and pasting it in the console. (Do not uncomment it!!!!) It will generate an HTML with all your code and its execution.

```
install.packages("markdown")
install.packages("knitr")
library(markdown)
library(knitr)
spin("./exama.R", precious=TRUE)
```