Genomics and Big Data (NWI-BP031)
R EXAM
27 March 2023, 12.45-15.45

Please read the following exam instructions carefully:

- This is an at-home exam. You are responsible for ensuring that you have stable internet, Brightspace access, and backup computers if necessary, during the exam.
- At the end of the exam, you must submit your R script as a Brightspace assignment. You are responsible for submitting your script on time. Late submissions will not be graded.
- You can submit only one file, with the extension of `.R`. Name your file by your student number (`s123456.R` where `s123456` is your student number).
- If you have any questions during the exam, you can find a TA in the following zoom link. Only questions related to clarification of the exam will be answered. We MAY provide limited technical support.
    - https://radbouduniversity.zoom.us/j/89776798287?pwd=Y000VzB5cmNNMmVTYlJqQm5OWnQ4QT09
    - Meeting ID: 897 7679 8287
    - Passcode: 702771
- This exam is open-book and open-internet. You can use your notes, Google, and any other websites. If you copy or adapt code from a public source, credit your source in a comment.
- This is an individual exam. You are not allowed to communicate or share information with anyone in any way. This includes emails, chats, direct messages, file sharing, and any other means of communication.
- Use the files given in this exam to answer the questions.
- Mark the questions clearly with comments, using the `#` character.
- For each question, use R to get the answer, and then write the answer again in comments.
- Comment your code liberally. If your code has bugs, we MAY use your comments to give you part marks.

Question 1: Transcriptomics

Please find the following three files: `transcriptomics_data.csv`, `transcriptomics_sample.csv`, and `transcriptomics_gene.csv`.
Expression of genes of samples are documented in `transcriptomics_data.csv`.
Detailed descriptions about samples used in this experiments could be found in `transcriptomics_sample.csv`.
A complete list for genes is in `transcriptomics_gene.csv`.

1. Let's first check `transcriptomics_data.csv` and `transcriptomics_sample.csv`. How many samples did researchers use for this experiment?
2. How many genotypes did researchers use for this experiment?
3. You probably notice that not all genes are expressed across all samples. For example, the gene with geneID "**100009614**"is not expressed in all samples. For how many genes no transcripts are detected across all samples?
4. Now you also notice geneID is not human readable. Let's check `transcriptomics_gene.csv`, you will find gene symbol of each geneID. Can you add the gene symbol information to the data frame so that people can easily understand the expression of each genes?
5. What are the ten most abundant genes for WT and KO samples (Note. WT and KO stands for wild type and knock out, respectively)?
6. The expression level in this data set is called transcript count, which is normalized across samples. Do you feel the digits of number are too long to read, right? To make life easy, we always convert the transcript count to transcripts per million (TPM) by simply dividing the expression values with 1,000,000. Can you make a new data set of TPM but without losing any geneID and symbol information?

Question 2: Proteomics

Please find the following three files: `proteomics_data.csv`, `proteomics_sample.csv`, and `proteomics_gene.csv`.
Abundance of proteins of samples are documented in `proteomics_data.csv`.
Detailed descriptions about samples used in this experiment could be found in `proteomics_sample.csv`.
A complete list for genes/proteins is in `proteomics_gene.csv`.

In this experiment, researchers perform "Differential expression analysis" (DEA) between two groups: Primed vs Naïve. When the abundance of a protein in Primed group is significantly changed in relative to Naïve group, the log2 fold change (`log2FC`) of this protein is larger or less than 1 (That is 2 times change), with a p value (`p_value`) less than 0.05.

1. Let first check `proteomics_data.csv`. You must notice that the GENEID column. As it is not human readable, could you add `SYMBOL` information from `proteomics_gene.csv` into the data set?
2. Could you tell us how many proteins with significant changes in their abundances?
   (Note. Log2FC $<$ -1 or $>$ 1 with p value $<$ 0.05)
3. We want to apply a stricter filter to identify important proteins. Can you tell us how many proteins with significant changes by 4 times in their abundances?
   (Note. Log2FC $<$ -2 and Log2FC $>$ 2 with p value $<$ 0.05)?
4. You may notice p value is not easy to read (too many 0!). Why don't we transform the p value to a human readable format? Could you transform the p value with the function "`-log10()`" (negative log10) so that it is easier to read?
5. Now since we have log2FC and -log10 P value, we can finally make a classical plot: Volcano plots. Volcano plot is a typical x-y scatter plot, with x axis is log2FC and y axis is -log10 P value. Can you draw the volcano plot?
6. On this volcano plot, can you label the gene symbol with the highest and lowest Log2FC as well as with p value $<$ 0.05?
7. **Bonus question:** Change the colour of all significant (p $<$ 0.05) proteins with a log2FC $<$ -2 (blue) OR log2FC $>$ 2 (red). Hint: add an extra column to your data frame which mentions which rows match the requirements.