Assignment 6: RNA Biology (Start early, this assignment may take some time)

Experimental data

Dr. B is interested in understanding the function of transcriptional regulator SnoN in the mammalian brain – cerebellum. SnoN plays critical roles in the development of postmitotic granule neurons in the developing cerebellum. It promotes the growth and stability of granule neuron parallel fiber axons in the rodent cerebellar cortex and regulates granule neuron migration and dendrite branching. Whereas the roles and mechanisms of SnoN have been characterized in postmitotic granule neurons, SnoN functions in granule neuron precursors have remained unknown.

Dr. B created conditional KO of SnoN in granule neuron precursors using the Math1-Cre driver. Because granule neuron precursors proliferate and differentiate in the external granule layer (EGL), they deployed a laser capture microdissection (LCM) approach to focus the analyses of SnoN specifically on the EGL of the developing mouse cerebellum. To better understand the transcriptional regulatory function of SnoN, they subjected the EGL from conditional SnoN KO and control littermate mice to RNA-seq after laser capture microdissection of EGL (LCMRNA-seq) at postnatal age of P6. (PMID: 30425119)

1. The data file (SnowKO_HTseq_table.txt) contains RNA-seq count data generated from mouse samples taken from wildtype (WT) and SnoN knockout (KO) mice cerebellum. There are 6 samples from wild type mice and 6 samples from SnoN knockout mice.

2. The metadata file (SnowKO_metadata.txt) contains sample information such as genotype, RIN number etc.

3. A template R script (Bio5488_Homework_RNA_seq.Rmd) was provided with hints for what's needed to complete the analysis. You will need to fill in the important details to complete the analysis to get genes differentially expressed between wildtype and SnoN knockout mice.

4. The proteinCodingMouseGenes.Rda contains the names of mouse coding genes.

5. The RIS_RStudio_Instructions.docx file contains the instructions on how to run RStudio on RIS using Open OnDemand.

Copy the above 4 files from the /storage1/fs1/workshops/Active/BIO5488/SP2026.L41.BIOL.5488.01/Assignments/A06/Assignment_Data/ folder to your home directory or your work directory /storage1/fs1/workshops/Active/BIO5488/SP2026.L41.BIOL.5488.01/Assignments/A06/Users/<wustlkey>.

Follow the instructions in the file RIS_RStudio_Instructions.docx file to run RStudio and open the Bio5488_Homework_RNA_seq.Rmd file on RIS using Open OnDemand. It also contains the instructions on how to install DESeq2 and other packages.

Bio5488, Spring 2026 Assignment 6 RNA-seq

There are prompts in Bio5488_Homework_RNA_seq.Rmd file for where you need to complete the assigned work. Look for TODO list sign and comments in the script as shown below that will tell you what you need to do.)
============================ TODOs ============================

Open up Bio5488_Homework_RNA_seq.Rmd and read through first to get a sense of what all you will need to do. Follow the steps shown below and write code in the script (Bio5488_Homework_RNA_seq.Rmd) to make sure you complete all parts of the assignment. Remember to comment your code.

Part 1 — Understand your data and put it into the correct format

When start a new project always check your starting data to make sure the data is correct and in the right format.

As input, the count-based statistical methods, such as DESeq2, edgeR, limma with the voom method etc. expect input data in the form of a matrix of un-normalized counts. The value in the i-th row and the j-th column of the matrix tells how many reads (or fragments, for paired-end RNA-seq) can be assigned to gene i in sample j. Analogously, for other types of assays, the rows of the matrix might correspond e.g., to binding regions (with ChIP-Seq), or peptide sequences (with quantitative mass spectrometry).

The values in the matrix should be counts or estimated counts of sequencing reads/fragments. This is important for DESeq2's statistical model to hold, as only counts allow assessing the measurement precision correctly. It is important to never provide counts that were pre-normalized for sequencing depth/library size, as the statistical model is most powerful when applied to un-normalized counts, and is designed to account for library size differences internally. If you want to understand how the count matrix was generated you can read more about it here (https://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html) .

1. After reading in the count matrix file, your data will be stored in a data frame with rows being genes and columns being different samples. The values will be the expression levels of that gene found in each sample in your study. You need to put the data into the correct format in terms of rows, columns, and matching the information between the count table and the metadata table.

Question 1:

Find our which rows in the input table do you need to remove and remove them from the count matrix.

take a look at the first 10 lines of the table. What was the content?

take a look at the last 10 lines of the table. What was the content?

Take a look at the first 10 lines of the count matrix after fixing all the problems to see what the correct format looks like.

2. Coding genes vs. non-coding genes: Dr. B is only interested in coding genes. He decided to remove non-coding genes from downstream analysis.
3. Read in the metadata file and take a look at the content. Understand what are the rows and columns. Which column contains the treatment conditions that you want to compare.
4. It is absolutely critical that the column names of the count matrix and the row names of the metadata table (information about samples) are in the same order with exactly the same name. There is code in the template script to examine the count matrix and metadata to see if they are consistent in terms of sample name and order. As they are not the same, we need to re-name the column names one or the row names so that they are consistent in terms of sample name and order. Check before and after to see what's the difference.

Part 2 — create DESeq2 data object

Read the section "3 The DESeqDataSet object, sample information and the design formula" following the link below to understand the data structure and data types required for appropriate downstream analysis.
https://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html

Question 2:
Construct DESeq object using "DESeqDataSetFromMatrix" function

it is preferred in R that the first level of a factor be the reference level (e.g. control, or untreated samples). In this case WT is the control.

Question 3:
Find out the current levels of the two conditions. If "WT" is not the reference level use "relevel" function to change the conditions so that "WT" is the reference level.

Part 3 — filtering, normalization

Advantages of pre-filtering by removing rows in which there are very few reads:
1. Reduce the memory size of the dds data object, and increase the speed of count modeling within DESeq2.
2. Improve visualizations, as features with no information for differential expression are not plotted in dispersion plots or MA-plots.
3. to improve power of detection

Here we perform pre-filtering to keep only rows that have a count of at least 10 for a minimal number of samples. The count of 10 is a reasonable choice for bulk RNA-seq.

Question 4:
pre-filtering to keep only rows that have at least 10 reads total. How many genes are left after removing genes with zero expression in all samples?

Read the section "4.2 The variance stabilizing transformation and the rlog" to understand the two transformation methods. rlog is more robust in the case when the size factors vary widely. The transformation is useful when checking for outliers or as input for machine learning techniques such as clustering or linear discriminant analysis.

Part 4 — Sample distances and outlier detection

Read the section "4.3 Sample distances" and "4.4 PCA plot" to understand how to calculate sample distances and plot heatmap and PCA plot.

Question 5:
use rlog transformation to calculate the Euclidean distance between samples and visualize the distances in a heatmap. Save your distance heatmap as pheatmap_before_outlier_removal.png

Question 6:
use PCA plot to visualize the distances. Save your distance heatmap as PCA_Plot_before_outlier_removal.png

It seems some samples were not behaving correctly such that WT and KO samples were mixed up. This suggest that some samples could be outliers due to technical or biological reasons. In Chen et al. BMC Bioinformatics, 2020 paper we show that PcaGrid with default parameters on rlog transformation normalized data achieved 100% sensitivity and 100% specificity in outlier detection for all tests.

Question 7:
use rld and PcaGrid to detect outliers in the dataset. What are the samples that were considered as "outliers"? Plot PCA diagnostic plot (or outlier map) and save it as "PCA_diagnostic_plot.png".

```
pcaG <- PcaGrid(t(assay(rlog(dds))), k=2)
which(pcaG@flag=='FALSE')
plot(pcaG)
```

Question 8:
remove outliers from the count matrix table and from the design table.
Question 9:
Construct DESeq2 object and perform filtering and normalization as you did above.

Part 5 — Differential expression analysis

Read the section “Differential expression analysis” to understand how to perform differential expression analysis, multiple testing, p value vs. adjusted p value,

Question 10:
Perform differential expression analysis and generate the result table comparing SnoN KO vs. WT with default parameters. What’s the default parameters for p values and log 2 fold changes? How many differentially expressed genes are there? Plot the counts plot for gene "ENSMUSG00000046152". Save the file as "Count_Plot_ENSMUSG00000046152.png”

Bonus question :
draw a heatmap to show the differentially expressed genes and save it as “DEG_Heatmap.png”.

What to turn in (*submission folder:
/storage1/fs1/workshops/Active/BIO5488/SP2025.L41.BIOL.5488.01/Assignments/A06/Users/<wustlkey>)

1. Edited script: Bio5488_Homework_RNA_seq.Rmd
2. Knitted html from Bio5488_Homework_RNA_seq.Rmd
3. Output files
   a. pheatmap_before_outlier_removal.png
   b. PCA_Plot_before_outlier_removal.png
   c. PCA_diagnostic_plot.png
   d. Count_Plot_ENSMUSG00000046152.png

For bonus question:
DEG_Heatmap.png

Note:
You can only cd or ls directly into your own submission and work directories with your wustlkey as students do not have access to the Users directory. Make sure you use the complete directory path. Also, we strongly encourage you to not change the names of your submission files so we can expedite the grading process.