Problem set 5

This problem set covers the content from week 6-7: RNA sequencing demo and single-cell transcriptomics.

Tips and rules:

- You can answer in English or in Thai.
- There can be more than one correct answer. What I am looking for from you is not just the correct answer but the rationale for your answer.
- Please provide evidence of how you think and what sources of information you used.
- AI such as ChatGPT may be used. You can also work together with friends. But you must write the answer in your own words.
- Any incidence of plagiarism and copying of another student's work will be reported to the Graduate Affairs.

Differential expression analysis using *sleuth*

For this problem set, we identify genes that are affected by MOV10 overexpression (https://www.genecards.org/cgi-bin/carddisp.pl?gene=MOV10).

The processed transcriptomics data from *salmon* can be downloaded from https://figshare.com/articles/dataset/Processed_overexpression_RNA-seq_salmon_/24182664.

There are 6 samples: 3 control and 3 MOV10 overexpression.

Q1: Create a metadata table describing sample names, conditions, and path to the data. Show your metadata table here.

Q2: Edit **run_sleuth.R** from the in-class demo to analyze this dataset. Explain how you modify the script here.

Now, perform differential expression analysis and use the outputs from sleuth to answer **Q3-Q6**.

Q3: How many transcripts are differentially expressed at q-value cutoff of 0.05? How many transcripts are differentially expressed at q-value cutoff of 0.01?

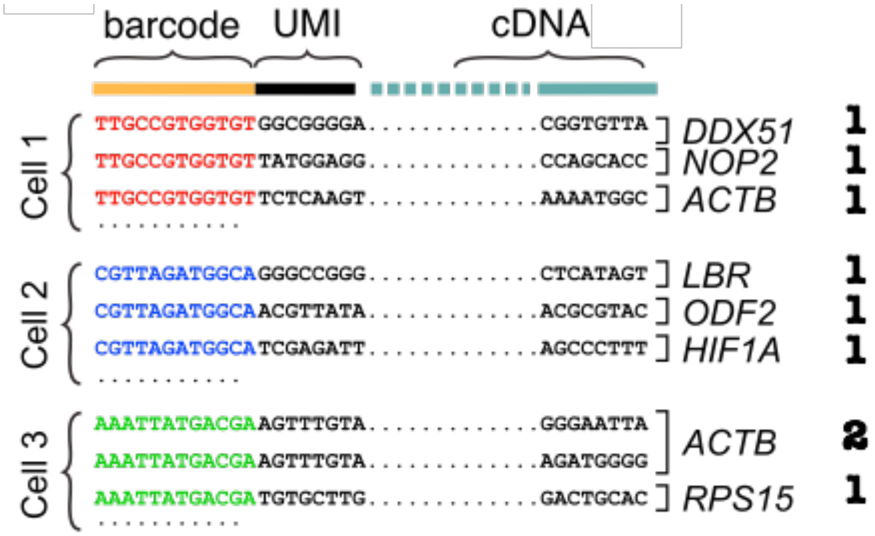
Q4: What are the top 3 most significantly up-regulated transcripts? What are the top 3 most significantly down-regulated transcripts?

Q5: Identify the gene symbols for the transcripts in **Q5** and **Q6**. Discuss whether the differential expression analysis result makes sense given the experimental conditions.

Q6: Visualize the boxplots for the TPMs of the most significantly up-regulated transcript (one transcript) and the most significantly down-regulated transcript (1 transcript).

Single-cell transcriptomics

Here is a diagram of a single-cell sequencing adapter. Use this diagram to answer **Q7-Q9**.


barcode
UMI
cDNA
Cell 1
TTGCCGTGGTGT GGCGGGGA.............CGGTGTTA
TTGCCGTGGTGT TATGGAGG.............CCAGCACC
TTGCCGTGGTGT TCTCAAGT.............AAAATGGC
DDX51 1
NOP2 1
ACTB 1
Cell 2
CGTTAGATGGCA GGGCCGGG.............CTCATAGT
CGTTAGATGGCA ACGTTATA.............ACGCGTAC
CGTTAGATGGCA TCGAGATT.............AGCCCTTT
LBR 1
ODF2 1
HIF1A 1
Cell 3
AAATTATGACGA AGTTTGTA.............GGGAATTA
AAATTATGACGA AGTTTGTA.............AGATGGGG
AAATTATGACGA TGTGCTTG.............GACTGCAC
ACTB 2
RPS15 1

Q7: Explain what barcode and UMI are.

Let's assume that a new biotechnology startup invented a new single-cell preparation protocol with new barcode and UMI designs as shown above. This company then hires you to develop a bioinformatics pipeline for read data produced by their platform.

Q8: What do you need to know about the characteristics of the barcode and the UMI from this company in order to extract them from each read?

Q9: Propose a **conceptual** bioinformatics pipeline for quantifying the expression level of each gene from this data. The output should be a table of expression data like *kallisto*'s output.

Hint: What should be done after you extracted barcode and UMI from each read? Adapt what you learned about the processing of sequencing data.

Q10: Study this single-cell paper: https://www.nature.com/articles/s41588-022-01100-4. What kind of filters were used to select high-quality cells and samples? For each filter, explain what it removes from the data.

Hint: There are several filters. The details may span across multiple paragraphs.

Q11: Although single-cell technique provides much more information than bulk RNA-seq, it also consumes much more resources. Propose a scenario where you think bulk RNA-seq should be performed over single-cell sequencing. Provide your reasoning.