Problem set 4

RNA-seq data analysis

For this problem set, we will adapt *sleuth* R script and online tools for functional enrichment analysis to identify genes and biological functions that are affected by MOV10 (https://www.genecards.org/cgi-bin/carddisp.pl?gene=MOV10) overexpression.

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

There are 6 samples, 3 control replicates and 3 MOV10 overexpression replicates.

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

Perform differential expression analysis between MOV10 overexpression and control.

Q2: How many transcripts are significantly differentially expressed at q-value cutoff of 0.05?

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

Q4: What are the top 3 most significantly up-regulated transcripts (sorted by q-value)?

Q5: What are the top 3 most significantly down-regulated transcripts (sorted by q-value)?

Q6: Identify the gene symbols for the top differentially expressed transcripts in **Q4** and **Q5**. Do they correspond to different genes, or are they isoforms of the same genes?

Q7: Do the top differentially expressed genes in **Q6** make sense given what you know about this dataset?

Visualize the TPM expression level of the most significantly up-regulated transcript (one transcript) and the most significantly down-regulated transcript (1 transcript).

Q8: Show your boxplots here.

At q-value cutoff of 0.05, enter differentially expressed genes into DAVID (https://david.ncifcrf.gov/home.jsp) to identify annotations that are enriched.

Q9: What is the appropriate "Identifier" for gene/transcripts in this dataset?

Q10: Summarize your findings and attach some screenshots.

Perform GSEA analysis using WebGestalt (http://www.webgestalt.org/) to identify KEGG pathways that are up-regulated and down-regulated in this dataset.

Q11: Summarize your finding and attach some screenshots.

Let's answer some more design and conceptual questions.

Q12: In addition to Likelihood Ratio Test (LRT), *sleuth* can also perform Wald test (https://pachterlab.github.io/sleuth/docs/sleuth_wt.html) to identify differentially expressed genes/transcripts. Explain conceptual differences between these two approaches. *Hint: These tests aim at identifying differentially expressed transcripts, but through different formulation of null and alternative hypotheses.*

Q13: How do we know that Poisson distribution poorly explain RNA-seq read data?

Q14: Explain two factors that make pseudo-alignment of RNA-seq to a reference transcriptome much faster than alignment to a reference genome.

Q15: When using DAVID or WebGestalt to perform overrepresentation analysis (ORA mode), you may notice that there are keywords "background" and "reference gene list". Explain what they are and what are their impact on the functional enrichment analysis.

We want to identify non-coding RNAs that are not typically expressed in normal tissues but are up-regulated in liver cancers. How would you acquire and analyze the data?

Q16: How would you prepare RNA samples? Which sequencing platform would you use?

Q17: What are the pros and cons of aligning RNA-seq reads to the reference human genome or the reference human transcriptome?