QBIO 490: Directed Research - Multi-Omic Analysis
Methylation Homework [Spring 2025]

Due: Wednesday, March 12th (11:59 pm). Submit your GitHub link to Brightspace, with your R Notebook with all your code and code outputs in a folder called `week8_methylation` within your `qbio_490_name` repo. Please email extension requests (include the reason for your extension and a proposed new due date) to Mahija, Erika and Jeanne by **Friday, March 14th 11:59 pm**.

Purpose: What are the differences between metastatic and non-metastatic BRCA across the epigenome?

Exploration of Methylation Patterns: For each analysis, write 3-4 sentence description analyzing the plot and answering the following question: What conclusions can you and can you not draw about differences between metastatic and non-metastatic TCGA BRCA patients? Why?

1. Naive differential methylation between non-metastatic and metastatic patients (Volcano plot)
2. Direct comparison of methylation status to transcriptional activity across non-metastatic vs metastatic patients
3. Visualization of CpG sites and protein domains for 3 genes (use UCSC genome browser) for a few genes. Describe at least one academic article (research or review) that either supports or doesn't support your final conclusion for one of the genes. If previously published work doesn't support your analysis, explain why this might be the case.

Technical Tips:

- Metastasis status should be based on the rna_se@colData$definition column.
    - Only consider "Metastatic" or "Primary solid Tumor" samples
- Be careful about what "barcode" columns you use! The patient id, sample id, and sample barcode columns are all named slightly differently across the different dataframes. Double check that the columns you are using to match index values are correct!
- For DESeq2 data preprocessing:
    - Use the rna_se clinical data (rna_se@colData).
    - Filter out genes with a total expression across all patients of < 20
    - Threshold padj values at 0.05 and log2FoldChange at |1|