Guidelines for Final Project
25% of Overall Grade in Course

I. Goals of this Assignment

The goal of this assignment is to integrate and demonstrate your understanding of RNA-seq data analysis using R and RMarkdown. By analyzing a publicly available NGS dataset, you will generate reproducible results, interpret findings, and communicate your scientific insights clearly through both code and narrative.

This assignment emphasizes:

- Hands-on experience with real RNA-seq datasets
- Proper use of differential expression tools (e.g., DESeq2)
- Reproducibility and transparency in bioinformatics workflows
- Clear scientific communication through well-structured reports and visualizations

II. Learning Objectives

By completing this assignment, students will be able to:

1. Describe the biological context of an RNA-seq study, including relevant background and rationale for the experimental design.
2. Perform RNA-seq data analysis using R, including data import, normalization, and differential expression analysis using DESeq2.
3. Interpret and visualize results
4. Structure a report using appropriate headers and formatting, ensuring it is accessible and useful to scientific peers.
5. Identify and reflect on analytical challenges or limitations and propose potential next steps or improvements.
6. Package and submit a complete, reproducible analysis project, including all necessary input files and a polished, readable HTML report.

III. General:

- All reports are due by Wednesday, April 30th by 11:59PM. The trail selected (*green*, *blue or black*) should continue to guide the overall analysis and presentation. In the class website (https://prodriguez19.github.io/MMG3320-5320/assignments/) I have posted examples of past project submissions for this assignment. Please note that the guidelines and rubric have changed for Spring 2025—these examples are provided only for general reference. They are intended to illustrate trail projects, not to serve as templates for formatting or content. Please refer to the Rmarkdown Source File Checklist at the end of this document.
- **If you are unable to submit this assignment by the date and time designated, please provide me with an excused absence from the dean or your medical provider.** Upon receiving this, you will be granted an extension for another specified date/time.

IV. Guidelines for Rproject_folder Submission:

a. Each student (or group) is responsible for emailing a completed Rproject_folder by the date and time specified.

b. The contents of the Rproject_folder should contain the following:

   i. One *annotated* Rmarkdown source file (.Rmd). This file should be clean, organized, and make use of **headings and subheadings** to clearly define sections. It should include:

   - Overview: Brief description of the project and dataset, including **GEO accession number** and **citation**
   - Libraries loaded: Include all libraries loaded
   - Data input: Brief explanation of what files are loaded
   - DE analysis: Include description of the **input design** for the `DESeqDataSet` object
   - **Visualizations**: All figures should be labeled and accompanied by a **Figure Legend**
   - Session info: Use `sessionInfo()`at the end of your file

   ii. One ***annotated, knitted*** Rmarkdown file (.html)

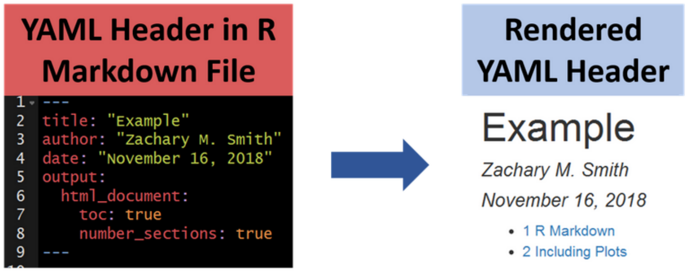
   - Same as above but in an .html format
   - ***The .html file should be easy-to-read and error-free*** i.e. does not contain excessive warnings including library warnings, contains headers and sub-headers, section breaks, and visible code.
   - ***Should display both code and output in a logical and clear way.***
   - You will be ****heavily graded**** on the overall readability and clarity of this file, as it serves as your scholarly report.

   iii. Input files: Include all files required to reproduce your analysis. This typically includes: normalized counts matrix, metafile, or any other input files required to run the .Rmd file. I should be able to replicate and generate a knitted html file on **<u>my</u>** laptop.

Rmarkdown Source File Checklist

_____Title/YAML Header

- Meaningful original title was provided
- Student full name(s) were given



```
---
title: "Example"
author: "Zachary M. Smith"
date: "November 16, 2018"
output:
  html_document:
    toc: true
    number_sections: true
---
```

Example

Zachary M. Smith

November 16, 2018

- 1 R Markdown
- 2 Including Plots

- Good blog post on rmarkdown_themes: https://rpubs.com/ranydc/rmarkdown_themes

_____General Information

- *About the dataset:* Overview of the authors research aim/hypothesis and description of the NGS- dataset collected by the authors. Description of the samples used for your NGS analysis. Often, you are using a subset of the original dataset - this should be stated clearly.
- *Citation:* AMA/APA format
- *GEO Accession Number*
- *Trail:* Indicate trail (Green, Blue, Black) and clearly state your overall objective

_____Data Input

- Metadata Input
- DE design (input for DESeqDataSet object)

_____Visualizations

- **Green Trail only:**
  - Your overall goal is to replicate figure(s) or figure panel in a primary research article and then change one parameter at the visualization stage. This could be altering colors, changing log2 fold change threshold, etc.

- **Blue Trail only:**
  - Your overall goal is to compare/contrast bioinformatic tools during the preprocessing stage and describe its impact on the data interpretation. Including MULTIQC outputs may be required to illustrate these points.

- **Black Trail only:**
  - Your overall goal is to test an original hypothesis, i.e. this should be *different* than the authors' aim/hypothesis. **In addition, you are tasked with generating an original figure(s) that are not found in the publication.**

_____Summary/Conclusions

- This section is completely up to you but must be included.
- Some thought questions you may try answering to conclude your presentation are:
  - During your analysis, did you hit any roadblocks?

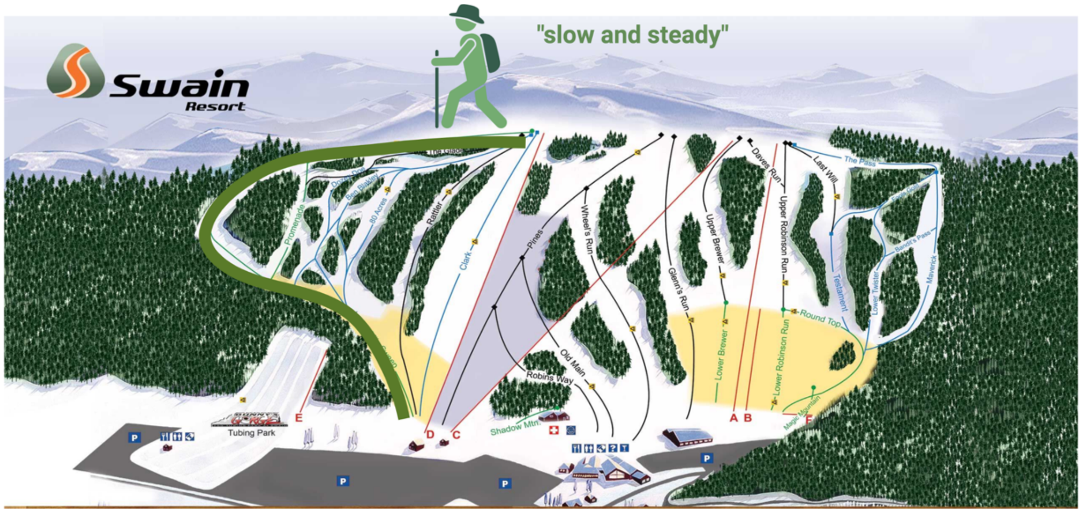
- ○ Did any of your findings surprise you?
- ○ Were they any red flags/limitations with how the samples were collected and subsequently sequenced?
- ○ What are the authors' next steps? What would be your next steps?

_____**Format and Grammar**

- Each section was clearly delineated
- Free of spelling and grammatical errors
- Dr. Rodriguez was able to recreate this analysis using the input files and RMD file provided

Trail 1: Green Mountain Trail

Replicate figure(s) in a primary research article and then change one parameter at the visualization stage



Challenge 1: Adjusting the Threshold for Differential Gene Expression (DEG) Investigate how changing the log2 fold change threshold (e.g., from 1 to 0.5) impacts the number and biological interpretation of differentially expressed genes. Discuss the trade-off between sensitivity and specificity in DEG analysis.

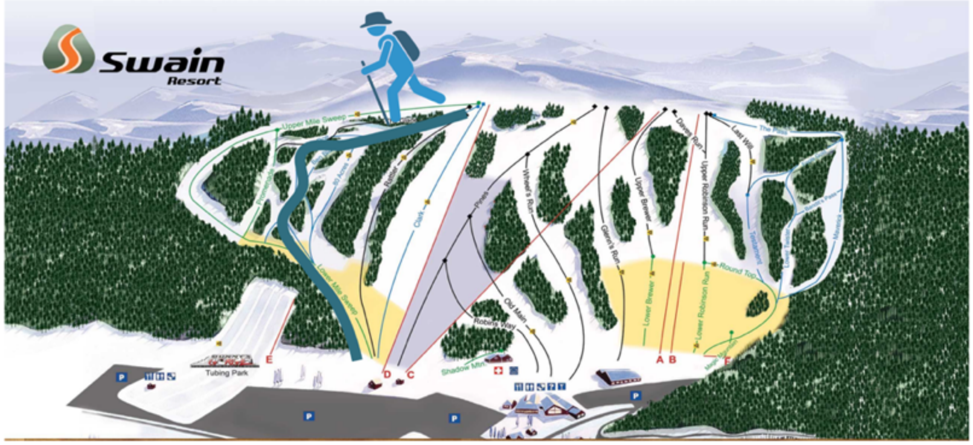
Challenge 2: Experimenting with Normalization Techniques
Compare visualizations of gene expression data normalized using two methods (e.g., TPM, CPM, vs. DESeq2's variance-stabilizing transformation). Assess how normalization affects downstream analyses like bar or box plots.

Challenge 3: Changing Color Schemes for Data Interpretation
Adjust the color scale of a heatmap (e.g., changing from a red-green to a blue-yellow color scheme) and evaluate how the choice of visualization colors influences the clarity of expression trends and ease of data interpretation.

The green trail guides students in understanding how to make ethically responsible decisions when visualizing NGS data.

Trail 2: Blue Sky Trail

Compare and Contrast bioinformatic tools during the preprocessing stage and describe its impact on the data interpretation



Challenge 1: Testing Different Alignment Tools
Align the RNA-Seq reads to the reference genome using two different aligners (e.g., HISAT2 vs. STAR). Compare metrics such as alignment rate, number of uniquely mapped reads, and runtime, and discuss how the choice of aligner might affect downstream analysis.

Challenge 2: Evaluating Reference Genome Versions
Map the RNA-Seq reads to two different versions of the reference genome (e.g., GRCh37 vs. GRCh38). Compare the alignment statistics and any differences in gene annotations. Discuss how the choice of reference genome might influence downstream results and biological interpretations.

Challenge 3: Comparing Count Generation Tools
Generate counts files using two different tools (e.g., HTSeq-count vs. featureCounts). Compare the total number of assigned reads, unassigned reads, and computational efficiency. Discuss how differences in counting strategies might influence downstream analyses such as differential expression.

The blue trail highlights the importance of tool selection during the preprocessing stage and its impact on the interpretation of RNA-Seq data.

Trail 3: Black Diamond Trail

“Process and Download an NGS dataset to test an original hypothesis”

Challenge 1: Creating Time-Series or Condition-Specific Plots

If your data includes multiple time points or conditions, create a figure (e.g., line plots or heatmaps) to visualize expression changes for key genes across these conditions. Highlight patterns or trends and discuss how they support or refute your biological hypothesis.

Challenge 2: Comparing Pathway Expression Across Groups

Use pathway analysis to identify key pathways enriched in a subset of your data. Create a customized plots (e.g. bar plots, dot plots, network graphs) to compare pathway activity between experimental groups not compared in the published work. Discuss how the visualization highlights the differences in pathway regulation.

Challenge 3: Annotating Single-Gene Expression Differences

Select a gene of interest from your dataset and create a violin plot or boxplot comparing its expression across conditions or groups. Customize the figure to include statistical annotations (e.g., p-values or fold changes) and explain why this gene is biologically significant.

For all black trail challenges you will be required to design a multi-panel figure that integrates multiple layers of analysis (e.g., a heatmap for expression patterns, a volcano plot for DEG results, and a GO enrichment bar chart). Explain how the combination of figures tells a cohesive story and enhances the overall interpretation of the data.

The black trail encourages students to think critically about data visualization while developing skills to create professional, publication-quality figures that clearly convey their <u>original</u> findings.