Guidelines for Final Project
25% of Overall Grade in Course

I. Upon completion of this project, the student will demonstrate:

a. An ability to effectively communicate complex scientific information to an audience
b. The aptitude to communicate complex analysis using R/Rstudio; providing enough detail that another scholar can repeat their entire NGS analysis.

II. General:

Each student (or group) will give an oral presentation describing their analysis of an NGS dataset. These presentations will be given by the student(s) on April 28th or April 30th. All graduate students are expected to attend both sessions and provide feedback to peer presenters. The trail selected (*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.

1. **If you are unable to present for any reason, please provide me with an excused absence from the dean or your medical provider.** Upon receiving this, you will be able to reschedule your presentation for another date/time.
2. Please utilize the Oral Presentation Checklist found at the very end of this document!
3. I'm including guidelines about the blue and black trail posted on January 22, 2025.

III. Guidelines for Rproject_folder Submission:

a. The Rproject_folder submission is worth 50% of final project grade
b. Each student (or group) is responsible for emailing a completed Rproject_folder, one hour before their scheduled presentation to princess.rodriguez@med.uvm.edu
c. 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 **my** laptop.
iv. ***If you are analyzing a non-RNA-Seq dataset,*** please follow guidelines provided by Dr. Rodriguez directly to you.

IV. Guidelines for Oral Presentation:

a. The oral presentation is worth 50% of final project grade
b. Presentations should be completed using Microsoft PowerPoint
c. Each student or group is responsible for emailing a completed PowerPoint presentation, one hour before their scheduled presentation to princess.rodriguez@med.uvm.edu
d. Presentations will begin at 2:00pm on dates specified, see schedule.
e. Each student or group will have ~15 min to present and answer questions from the audience.
f. Carefully proofread any text on your slides
g. Reference page must be included in APA or AMA citation format
h. Please follow the oral presentation checklist provided below.

Presentation Schedule:

Time	Student Presenter Monday, April 28th
2:00 - 2:15	Emma Danza
2:15 - 2:30	Emma Steenbergen
2:30-2:45	Chika Ikpechukwu
2:45-3:00	Keegan Smith
3:00-3:15	Sara Jaffrani
Time	**Student Presenter Wednesday April 30th**
2:00 - 2:15	Yusuf and Rob
2:15 - 2:30	Jaime Boisoneau
2:30-2:45	Tyler Hogan
2:45-3:00	Brailey and Owen
3:00-3:15	Simran Rai

Oral Presentation Checklist

_____Title

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

_____Introduction

- *Overall, the student delivered necessary background to the audience so that they understood the importance of the research conducted.*
    - Note: I am not the only person in this room. You should be tailoring your presentation to an audience of your peers.
- *Adequate general background was provided.*
    - What is the disease/abnormality/complication being studied? Who does it affect? (provide stats, mortality, overall impact)
    - What are the perturbations being used? For example:
        - If it is a *drug* we need to know what it does and how it works.
        - If it is a *mutation* we need to know what it affects (signaling pathway?)
        - If it is a *knockout* we need to know how this gene normally functions
        - If it's a *cell state* this needs to be defined (ex. senescent, hypoxia, etc)
- *Your current research.*
    - It would also be great for the audience to hear which laboratory you are in, the sort of research this laboratory performs, and how long you have worked here (~1 slide).

_____Experimental Design

- 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. More often than not, you are using a subset of the original dataset - this should be stated clearly.

_____Major Findings

- Description of DE design i.e. input for DESeqDataSet object. Be sure to provide total numbers of DEGs, distinguishing up- vs down-regulated genes.
- A comprehensive overview of any data visualization, including but not limited to: PCA plot, Volcano plot, Heatmap, Functional Analysis, Box plots

- **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

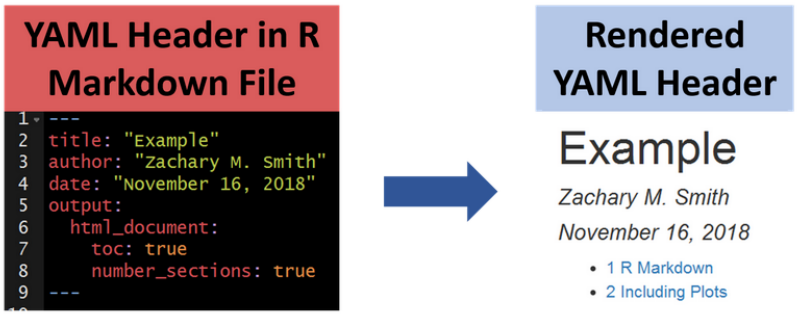
- This section is completely up to you but must be included.
- Some thought questions you may try answering to conclude your presentation are:
    - During your analysis, did you hit any roadblocks?
    - Did any of your findings surprise you?
    - Were they any red flags/limitations with how the samples were collected and subsequently sequenced?
    - What are the authors' next steps? What would be your next steps?

_____References

- One slide dedicated to references should be provided at the end in APA or AMA format.

_____Format, Grammar, and Delivery

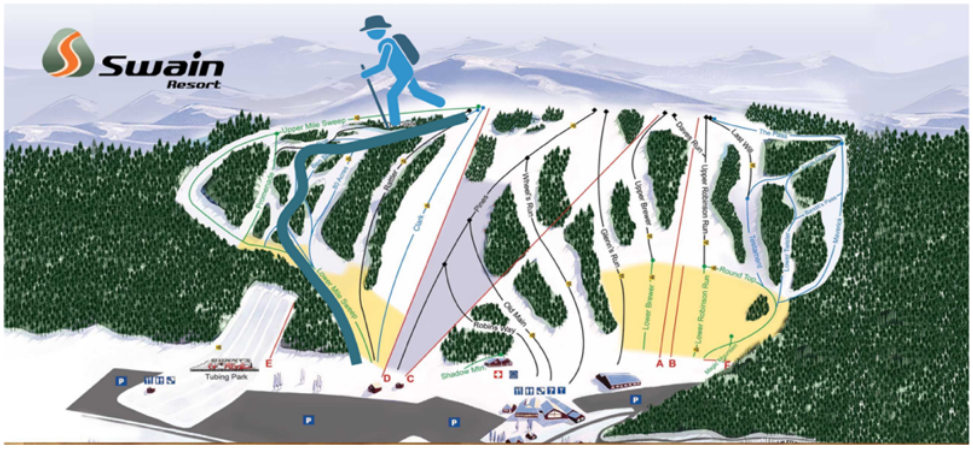
- Each section was clearly delineated
- Free of spelling and grammatical errors
- Smooth delivery and tone
- Stayed on time (not too long, not too short)
- Overall knowledgeable about topic
- Made a valid attempt to answer questions



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

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.

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