MMG3320
Spring 2026
Homework #5
100 points

Guidelines for this homework: GEO (Gene Expression Omnibus) is an international public repository that archives and freely distributes microarray, next-generation sequencing (NGS), and other forms of high-throughput functional genomics data submitted by the research community. It contains 100's of organisms and thousands of different expression analysis datasets. Each dataset that has been deposited to GEO is from an original peer-reviewed research article. For your final project, you will be asked to reanalyze a NGS dataset of your choosing.

For this homework assignment, please identify the primary research article and samples you would like to perform this bioinformatic reanalysis on. Keep in mind that each reanalysis will be performed with a specific, larger "goal" in mind. These goals are specific to the trail selected and can be broadly summarized as: 1) to replicate the findings from the authors (**Green Mountain**), 2) alter the bioinformatic pipeline and understand how this impacts the final findings (**Blue Sky**), or 3) use the dataset to test an original hypothesis (**Black Diamond**). Please select a trail that best matches your background, curiosity, and comfort level.

- For Part A & B, open a new Microsoft Word document:
  - Include Name, Course, Date, and Assignment as first (4) lines
  - Acceptable fonts = Arial or Times New Roman
  - Font size = 12pt
- For Part C, please complete and submit the Microsoft Excel template – **sample_metafile.xlsx**. Instructions and a step-by-step guide for GEO are found below.
- References must be listed in APA or AMA format. Instructions can be found here: https://researchguides.uvm.edu/c.php?g=290226&p=1934958

Due date: **The due date for this homework is Friday, February 20th by 5pm**. Please upload Part A and B as a .pdf or .docx onto Brightspace (50 points). Then please upload Part C separately as a .xlsx file onto Brightspace (50 points). Late homework will be docked 10% of the overall grade for every day that the assignment is late. An assignment is considered late if it is not submitted by the date and time specified. Three days past the due date (weekend included), the assignment will no longer be accepted, and the student(s) will receive a ZERO.

Please email princess.rodriguez@med.uvm.edu if you have any questions.

Part A: Logistics (10 points)

Please the following:

- full citation of the primary peer-reviewed research article selected for the final project
- trail selected (green, blue, black)
- GEO Accession Number
- PDF of primary paper (include with submission)

Part B: Final Project Goals (40 points)

Answer the following prompt according to the trail selected. **This should be ~1 page.**

Green Mountain trail:

The Green Mountain trail focuses on learning how raw sequencing data are transformed into interpretable biology results by carefully reproducing a published analysis.

- Identify a research article and one figure that you would like to replicate and genuinely want to understand better.
- Provide a narrative for the Figure selected. This narrative should include:
  - Description of each figure panel *(i.e. A, B ,C, D…)*
    - Describe what data type is shown in each panel (i.e. counts, clustering, enrichment), not just what it looks like.
  - What specific comparison or biological signal is this figure designed to reveal?
  - Notable findings or overall conclusions for the Figure selected
  - An image of the Figure you would like to recreate.
  - Also include why you selected this paper at the end of the summary.

Blue Sky trail:

The Blue Sky trail is designed for students who are curious about how different bioinformatic decisions influence results. Rather than strictly reproducing an analysis, you will focus on understanding and modifying the computational workflow.

Below is the core RNA-Seq pipeline you will learn in this course. Your goal is to understand how this pipeline compares to the one used in your selected paper.

MMG3320	What it does…
FASTQC	Quality control FASTQC files
Trimmomatic	Trim adaptors and low quality reads
HISAT2 STAR	Alignment to Genome
SAMtools	SAM to BAM
HT-Seq-count	Create counts files

Provide a narrative of the experimental design as it pertains to the NGS dataset that is to be reanalyzed for this project. This narrative should:

- Focus on elements that directly influence downstream analysis (i.e. strandedness, read length, read depth, genome reference, aligner choice, etc).
- Provide a complete overview of methods: sample collection, number of samples analyzed, genome reference and bioinformatic programs used. Include versions and parameters if provided. *You may need to investigate the supplement for more details on the methods*.
- Which steps are identical or differ from what we will learn in class?
- By altering the pipeline strategy, what differences do you expect and how will this translate into your downstream results?

Black Diamond trail: Follow this set of instructions if you plan to analyze data from a research article.

The Black Diamond trail will emphasize scientific ownership, feasibility, and ambition. Projects in this trial will be closely mentored to ensure that the scope is appropriate and achievable within the course timeline.

- Summarize the rationale/importance of the research being conducted in the primary article selected. Clearly state the research question and/or hypothesis **being tested by the authors.** Finally, summarize major findings as it pertains to the NGS dataset you plan to reanalyze.
- Clearly state the *original hypothesis* you will test. Your hypothesis should be specific and testable using the available data. In addition, elaborate on the types of analysis you believe are needed (ex. isoform usage, splicing, pathway analysis). What will a successful outcome for this project look like?

Black Diamond trail: Follow this set of instructions if you plan to analyze YOUR data that has not been published.

- Select a primary research article that is *relevant* to your work and includes a NGS dataset. Summarize the rationale/importance of the research being conducted in the primary article selected. Clearly state the research question and/or hypothesis tested by the authors.
- Clearly state the *original hypothesis* you will test using the NGS dataset your lab has created. Your hypothesis should be specific and testable using the available data. In addition, elaborate on the types of analysis you believe are needed (ex. isoform usage, splicing, pathway analysis). What will a successful outcome for this project look like?

Part C: Creation of Metafile (50 points)

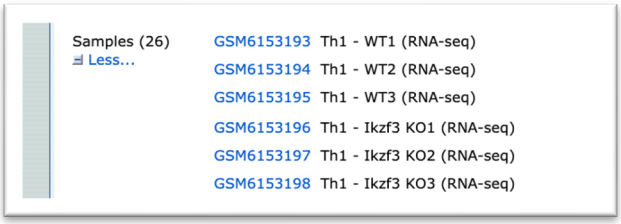
The primary article you have selected may contain *many more* sequencing datasets then what you plan on analyzing. Fill out and submit sample_metafile.xlsx for only the samples you plan on reanalyzing for the final project.

Considerations:

- Replicates: 3 biological replicates per condition is required
- Number of Samples to process:
  - Minimum of 8 samples
  - Maximum of 24 samples
- Information found in the Gene Expression Omnibus (GEO) is going be extremely helpful for filling out this table.

Below is a step-by-step guide for using GEO: *You will not be allowed to use the RNA-Seq data from this paper for your final project.*

1. Find your research article of interest.
   a. https://www.nature.com/articles/s41467-023-37420-0
2. Scroll until you find the **Data availability section.** Click on the accession number that corresponds with the dataset of interest.
3. This will open the GEO page: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE203066
4. Scroll until you see the Samples:



5. Each sample comes with its own Sample Accession Number that starts with GSM. Click on the first one to open. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM6153193
6. Under Data Processing, is the bioinformatic pipeline used to analyze this sample: