

MMG3320/5320

Spring 2025

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 first 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**).

- Graduate students *must* select either Blue Sky or Black Diamond trail
- One homework submission per student group will be accepted (*undergraduates only*)
- For Part A & B, open a new Microsoft Word document to answer:
  - 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 21<sup>st</sup> 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](mailto:princess.rodriguez@med.uvm.edu) if you have any questions.

## Part A: Logistics (10 points)

Please list a full citation of the primary peer-reviewed research article selected for the final project, group members (if any), AND trail selected (green, blue, black).

## Part B: Final Project Goals (40 points)

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

### Green Mountain trail only:

- You overall goal is to identify a research article and one Figure (ex. Figure 2) or Figure panel(s) (ex. Figure 2A-C) that you would like to replicate and learn more about.
- Provide a narrative for the Figure selected. This narrative should include:
  - Research question or hypothesis being tested for the Figure selected
  - Description of each figure panel (*i.e.* A, B, C, D...)
  - 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 only:

- The bioinformatic programs you will learn in class over the next few weeks are detailed below.

MMG3320/5320	What it does...
FASTQC	Quality control FASTQC files
Trimmomatic	Trim adaptors and low quality reads
HISAT2	Alignment to Genome
STAR	
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 include:
  - The research question or hypothesis being tested in the paper selected
  - 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 look into the supplement for more details on the methods.*
- Compare how the pipeline bioinformatic pipeline differs from what we will learn in class. In addition, please include what bioinformatic programs you are interested in comparing and contrasting for your final project and any visualizations you would like to create.

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

- 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 some “highlights” in terms of major findings as it pertains to the NGS dataset you plan to reanalyze.
- Clearly state the *original hypothesis* you will test using the NGS dataset provided from this primary article. In addition, if known, elaborate on any bioinformatic programs you are interested in using, visualizations you would like to create, and overall objectives for your final project. By providing me with as much information up front, I can guide and scale your project so that you are successful with your reanalysis.

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

- Select a primary research article that is *relevant* to your work and includes a *relevant* 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 you have created. In addition, if known, elaborate on any bioinformatic programs you are interested in using, visualizations you would like to create, and overall objectives for your final project. By providing me with as much information up front, I can guide and scale your project so that you are successful with your reanalysis.

## Part C: Creation of Metafile (50 points)


The primary article you have selected may contain *many more* sequencing datasets than 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: 2-3 biological replicates per condition is required
- Number of Samples to process **\*\*suggestion\*\***:
  - Minimum of 8 samples
  - Maximum of 16 samples
- Information found in the Gene Expression Omnibus (GEO) is going to 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**. Remember accession numbers?! 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:



Samples (26)	<a href="#">GSM6153193</a>	Th1 - WT1 (RNA-seq)
<a href="#">Less...</a>	<a href="#">GSM6153194</a>	Th1 - WT2 (RNA-seq)
	<a href="#">GSM6153195</a>	Th1 - WT3 (RNA-seq)
	<a href="#">GSM6153196</a>	Th1 - Ikzf3 KO1 (RNA-seq)
	<a href="#">GSM6153197</a>	Th1 - Ikzf3 KO2 (RNA-seq)
	<a href="#">GSM6153198</a>	Th1 - Ikzf3 KO3 (RNA-seq)

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:

<b>Data processing</b>	Sequence reads were trimmed to remove possible adapter sequences and nucleotides with poor quality using Trimmomatic v.0.36. The trimmed reads were mapped to the Mus musculus GRCm38 (mm10) reference genome available on ENSEMBL using the STAR aligner v.2.5.2b Using DESeq2, a comparison of gene expression between the customer-defined groups of samples was performed. The Wald test was used to generate p-values and log2 fold changes. Genes with an adjusted p-value < 0.05 and absolute log2 fold change > 1 were called as differentially expressed genes for each comparison. Assembly: Mus musculus GRCm38 (mm10) Supplementary files format and content: DESeq2 analysis including normalized counts, log2 fold changes, and p-values; .csv
------------------------	---