MMG3320
Spring 2025
Homework #6
150 points

Guidelines for this homework: All students have selected a dataset for their final project. Now, it's time to download the dataset from GEO and perform a FASTQC analysis. FASTQC is a quality control tool for high-throughput sequencing data. By identifying potential issues early, it helps researchers ensure data quality and improve downstream analysis.

- Please open a **new** Microsoft Word Document
- Font size: 12 pt
- Font: Times New Roman or Ariel
- Line spacing: 1.5
- Length: Minimum of 1-page. Maximum of 2-pages.
  I am **not** looking for the obvious, "This sample has 25 million reads". ***I would like to know if you could interpret the html report***. Therefore, I am looking for more insight, such as, "…this sample was determined to be potentially problematic as it contained 10 million less reads than any other in the dataset."

Due date: The due date for this homework is March 6th by 5:00pm.
PART A: The final MULTIQC output (.html) should be emailed to me. Brightspace will not accept .html files.
PART B: The .docx file containing your interpretation should be submitted via Brightspace.

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.

ASSIGNMENT:

PART A (60 points): Running FASTQC and MULTIQC

- Perform FASTQC on the FASTQ files downloaded from GEO.
- Generate a MULTIQC report (.html) summarizing the quality metrics for the final dataset.

PART B (60 points): Evaluation of MULTIQC output.

1. **Dataset Overview:** Describe the dataset/samples being analyzed.

2. **Sequencing Quality:**
   - How successful was the sequencing?
   - Do the FASTQC results indicate any technical issues?
   - Support your conclusions by interpreting relevant FASTQC metrics.

3. **Overrepresented Sequences:**
   - Are there any overrepresented sequences in the dataset?
   - If so, identify the sequence with the highest percentage and its likely source of contamination.

4. **Problematic Samples:**
   - Are there any samples that are a cause for concern? If so, how would you address these issues in the downstream analysis?

To aid in writing your analysis report, you begin by answering these questions:

Sample Overview:

- What kind of samples (human, mouse, tissue, etc) are being analyzed?
- How many biological replicates are included?
- Are these samples paired-end or single-end?

Read Quality & Composition:

- What is the GC content distribution across the samples?
- Do any samples show lower overall PHRED scores (sequence quality)?
- Is there adapter contamination present in any samples?
- Are there any sequences that appear in a large proportion of the reads (overrepresented sequences)?
- What will be your next steps to address any concerns?

PART C (30 points): Trimming & Filtering Class Exercises (L10)

Class Exercise #1: Following the prompt for `trimmomatic_exercise` and answer the following questions.

1) How many reads were trimmed in `SRR2589044_1` and `SRR2589044_2`, respectively. State which files were used to answer this question.

2) Did the per base sequence quality improve after trimming? If so, at which positions was the biggest improvement?

3) Was adapter content reduced or eliminated after trimming? In your answer, please state the name of the adapter that was altered.

Class Exercise #2: Following the prompt for `trim_galore_exercise` and answer the following questions.

1) Comparing the FASTQC reports before and after trimming, report how the sequence length changes.

2) Which FastQC modules showed the most improvement after trimming? Be sure to report which adapter was reduced/eliminated after trimming.

3) If you examine the Overrepresented Sequences table, do any contaminants remain?