BNFO 262 2025
Homework 3

Instructions
Answer the following questions in your own words using a Jupyter notebook and upload a PDF of your answers to Gradescope. Make sure to write your name and PID at the start of your answers. This assignment is due **3/6/25 at 9:00AM**.

Part 0: (2 Points) Please make sure you select the correct pages for your answer to each question on Gradescope. You will get full points if you assign the pages correctly for all questions.

Part 1: ChIP-seq

Before you start:

- Launch ghcr.io/biom262/cmm262-notebook:chipseq DataHub environment for this section
- Open a terminal and run `source activate chipseq`
- The following questions refer to these two files:
  - ~/public/hw3/hw-f1.bam
  - ~/public/hw3/hw-f2.bam
  - Note: you don't need to copy over these files, you can just use the path above

1. You received two BAM files from a colleague (hw-f1.bam and hw-f2.bam) but there was a mess up, and the files were not annotated properly. Which UNIX command(s) might you use to investigate which species the data was aligned to? Justify why your command(s) would work. (2 points)

2. Find the two BAM files in the hw3 directory and then run your suggested commands from the previous question. Which species were the two BAM files aligned to? (1 point)

3. What chromosome does the data in these two files come from? Note, some reads can align to multiple chromosomes, so make sure that your answer is not impacted by outlier reads. (1 point)

4. You were told that the two files are from a ChIP-seq experiment and you are trying to figure out what is in each file: input or ChIP of DNA-binding protein. Which file is the input and which file comes from the ChIP-seq of the DNA-binding protein? What is the target protein/histone modification? Explain the steps you took and the commands that you used to reach this decision and add screenshots of any visualization tools. Hint: use two consecutive steps of the eleven in the ChIP-seq module. (3 points)

Part 2: Single cell RNA-seq

5. Describe the experimental differences between droplet based methods and physical separation methods. Start by explaining each. Then discuss when it might be better to use one or the other. Limit 5 sentences. (3 points)

6. When analyzing scRNA-seq data, how do you know which cell a read comes from? Limit 5 sentences. (1 point)

7. When preprocessing scRNA-seq data, there are multiple metrics to perform QC on. Please list three QC metrics and explain what each controls for? Limit 5 sentences. (3 points)

8. Come up with an experiment in which you'd prefer to use scRNA-seq rather than RNA-seq. Justify why you believe this assay would be preferable for this experiment. Limit 5 sentences. (2 points)

Part 3: Variant Calling

9. In your own words, explain the difference between sequencing depth and breadth. Describe one potential variant calling analysis where you'd prefer higher depth and one experiment where you'd prefer higher breadth. Limit 5 sentences. (2 points)

10. Why would you prefer to use a microarray over short-read sequencing for variant calling? (also why would you prefer short-read sequencing over microarrays?) Limit 3 sentences. (2 points)

11. What are the pros and cons of using long read sequencing to call variants. Limit 3 sentences. (2 points)

12. Your friend is interested in knowing the eye color of his preborn son. You know that the presence of two G alleles at SNP rs12913832 is strongly predictive of having blue eye color. Your friend has provided you with variants called from a prenatal microarray, but rs12913832 isn't present in the microarray dataset. Fortunately, rs12913832 is present in the 1000 Genomes project, a large dataset of known variants. How can you use the 1000 Genomes dataset to determine whether your friend's son has both G alleles? (1 point)