Homework 2: Population Genetics & RNA-seq

Submission Instructions

Answer the following questions in your own words upload a PDF of your code (for Part 1) and written answers to Gradescope. Make sure to write your name and PID at the start of your answers. This assignment is due 2/13/25 at 9:00AM.

See this document for instructions on creating a Jupyter notebook:

https://docs.google.com/document/d/15Ux16sVp9EsBU_tJiqCMJnrqqjuBor5s/edit?usp=sharing&ouid=102134196267797631717&rtpof=true&sd=true

Please use the cmm262-notebook:popgen environment for this assignment.

Part 1: Population genetics

1. (For this question, you can find the data files named *1000G_***_chr18** in the folder *~/public/hw2/data_chr18.* **Please include your code for this question in the PDF submitted on Gradescope.**) F_{ST} is a quantity that measures the divergence between two populations at a particular variant. The formula for F_{ST} is

$$\frac{\left(AAF_{Pop1} - AAF_{Pop2}\right)^2}{2p(1-p)}$$

where *AAF* is the "alternative allele frequency" with respect to Population 1 and *p* is the average AAF between population 1 and population 2.

What is the average F_{ST} across the first 5000 variants between all pairs of four populations using chromosome 18 SNP data from the 1000 Genomes Project (*1000G_***_chr18**) for European, African, East Asian, and American populations **(9 points)**? To simplify the calculation, please ignore any swaps between the .bim files. Simply calculate the alternative allele frequency for each population and compare their differences. Which pair of populations are the most genetically distant (large F_{ST}) **(1 point)**?

Part 2: RNA-seq

2. For the following main steps of a full differential expression RNA-seq analysis, explain the purpose of what is done computationally and give an example of a tool you would use (if relevant). (3 sentence limit for each step) (4 points)

a. QC the FASTQ files
b. Align the FASTQ files
c. Sort and index the aligned reads
d. QC the bam file
e. Count reads for every gene
f. Normalize the counts
g. Perform differential expression analysis
h. Visualize the results

3. Before loading your data into DESeq2 to perform your differential expression analysis, do you need to normalize your count matrix? If not, why not? (2 points)

4. Given the gene expression table below with read counts for Rep1, Rep2, and Rep3, which two samples are the ***most similar*** to one another? Make sure to show your work and final TPM values. (*Hint:* Calculate TPMs and sum the absolute value of TPM differences between every pair of samples. You can also use a scale factor of 10 instead of 1 million) (3 points)

Gene name	Rep 1	Rep 2	Rep 3
A (2kb)	15	3	30
B (4kb)	10	22	60
C (1kb)	5	10	15
D (10kb)	0	0	1

5. Your collaborator sent you some processed mouse data they wanted you to analyze, but forgot to include the sample annotations so you don't know what type of mice the samples are from. You do know that Gene_A and Gene_B are upregulated in the knockout (KO) mice compared to the wildtype (WT) mice. Based on the heatmap below, which samples do you think are from the WT and KO groups? Briefly explain your reasoning using the genes. (3 points)