CMM 262 2023
Homework 3: RNA-seq and Variant Calling (Priya Pantham and Kyle Gaulton)

Instructions
Answer the following questions in your own words 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 **3/9/23 at 9:00AM**.

Part 1: RNA-seq

1. 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

2. Before loading your data into DESeq2 to perform your differential expression analysis, do you need to normalize your count matrix? If not, why not? (1 point)

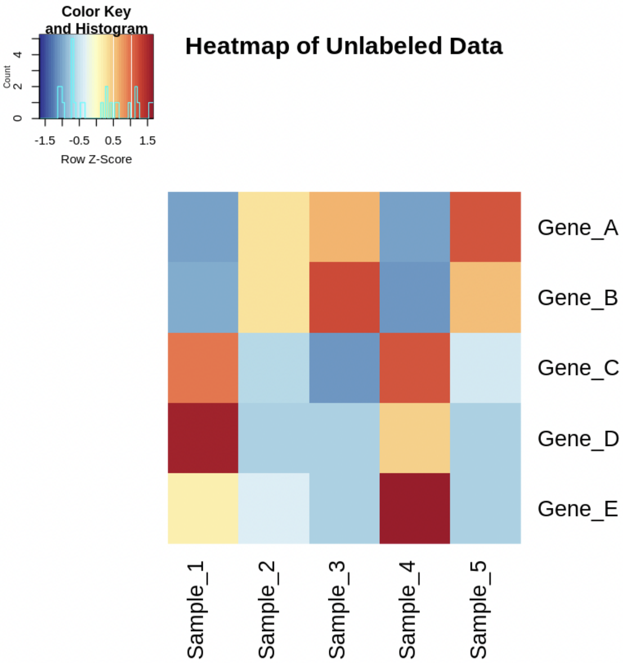
3. Say you are looking at two genes in an experiment, *Gene A* and *Gene B*. If the log2 fold change of *Gene A* to *Gene B* is 3 ($log2(\frac{Gene\ A}{Gene\ B}) = 3$), how many times the expression of A is the expression of gene B? (1 point)

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? (*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. (2 points)



6. You have a list of Ensembl human gene IDs that you want more information on so you decide to use biomaRt to get that information.

    a. What code would you use to connect to the Ensembl gene mart? (Hint: use the `mart<-useDataset()` function) (2 points)

    b. What function would you use to see the attributes available for the mart variable? (1 point)

c. After looking through the available attributes, we want the following information: Ensembl gene ID, gene description, chromosome name, and the gene start and end. Using the attributes table generated by the previous command (saved in ~/public/hw3/attributes.txt for you to view), provide the code to generate the biomaRt gene list with the necessary information. Assume your gene list is stored in a variable `gene_list`. (Hint: use the `getBM()` function) (2 points)

Part 2: Variant Calling

7. In your own words, explain the difference between sequencing depth and breadth. Describe one potential experiment where you'd prefer higher depth and one experiment where you'd prefer higher breadth. (4 points)

8. 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?) (2 points)

9. What are the pros and cons of using long read sequencing to call variants. (2 points)

10. 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)

11. List and describe three VCF fields commonly used to filter out poor variant calls or recalibrate variant quality scores. Assume you are working with short germline variants from GATK's HaplotypeCaller. (3 points)

12. Another command line tool which can be used to work with variant call datasets is plink. In this question you'll be introduced to plink and use it to filter variant calls from an individual from the 1000 Genomes Project. In a datahub terminal window, create symlinks to the vcf and tbi files we'll use in this question with these commands:

a. The collaborator who sent you this file failed to explain anything about the sample. Go to the 1000 Genomes data portal (https://www.internationalgenome.org/data-portal/sample) and search for relevant information on sample HG02106. What is the sex of the person who donated this sample and what ancestry group are they a part of? (1 point)

b. Convert your vcf into a plink file set, treating any half-calls as missing data. See the plink documentation page on input data types for help (https://www.cog-genomics.org/plink/1.9/input) and to run plink use `/opt/conda/envs/r-bio/bin/plink`. Write the command you used here. (1 point)

c. Plink outputs multiple files (bed, bim, fam, log, nosex), what information is contained in each file? You can read through the plin documentation to find this information, but make sure to rephrase it in your own words. (2 points)

d. Filter your plink file set to only contain variants on chromosome 20 with a minor allele frequency (MAF) ≥ 0.05. See the plink documentation page on data filtering for help (https://www.cog-genomics.org/plink/1.9/filter) and to run plink use `/opt/conda/envs/r-bio/bin/plink`. Write the command you used here. (1 point)

e. How many variants remain after applying the filter in part d? (1 point)