CGS4144 Bioinformatics - Assignment 2

Directions: *Following the steps below, analyze the RNA-seq data you selected in Assignment 1. Save each image/table you create and write a short summary (3-5 sentences) of how they were made and what you discovered.* ***You may work as a team, but this is an individual assignment. Submit individually.*** *Create a work document with your team's name at the top followed by your results, to be submitted on Canvas.*

You will submit a document (doc/docx/pdf) that lists the names of everyone on your team, your team's GitHub repository with all code used in the assignment, and each image/table plus their short summaries. These writeups will be used to create your final project report, so it is to your advantage to do a thorough job now.

Team:
Add the full names, email addresses, and GitHub handles for each team member (3-5 people total) to this table:

Name	***Email***	***GitHub***
<student1 name>	<student1@ufl.edu>	<github.com/student1>
<student2 name>	<student2@ufl.edu>	<github.com/student2>
<student3 name>	<student3@ufl.edu>	<github.com/student3>
<student4 name>	<student4@ufl.edu>	<github.com/student4>
<student5 name>	<student5@ufl.edu>	<github.com/student5>

Data: <Link to your dataset >

Scientific Question: < Question you will answer using the data you've selected >

GitHub Repository for Project: < Link to your team GitHub repository >

1. *Download the expression data and matching metadata from Refine.Bio that you selected in Assignment 1.*
    a. *You should have a matrix of samples by genes expression data*
    b. *If your matrix has Ensembl IDs (e.g. ENSG00000141510) instead of Hugo gene names (e.g. TP53), convert the names to Hugo gene names. Here are some guides:*
        i. *alexslemonade.github.io/refinebio-examples/03-rnaseq/gene-id-annotation_rnaseq_01_ensembl.html*
        ii. *bioconductor.org/help/course-materials/2019/BSS2019/05_Annotations.html - org.hs.eg.db*
    c. *Load the data into your chosen programming language (R or python recommended). What size is your expression matrix? How many genes does it include? How much variation do you see in the data? To answer these questions, log-scale the data, calculate per-gene median expression ranges, then make a density plot showing those results. Summarize your findings.*
2. *Now that you have loaded the expression data, generate a PCA plot:*
    a. *You can do this using any PCA implementation. Here is a guide to using the DESeq2 function plotPCA() to generate your plot (see here).*
    b. *Color your PCA plot by the 2 groups you identified in assignment 1 (e.g., cancer vs normal)*
    c. *Make sure you include a legend and label the axes!*
    d. *Also generate t-SNE and UMAP plots, making sure to color code and label each plot.*
        i. *t-SNE (example here)*
        ii. *UMAP (example here)*
    e. *Summarize the differences and similarities between your three plots.*
    f. *Save your plot(s) and summarize your findings.*
3. *Perform differential analysis on the samples from your two groups.*

    a. *A tutorial for this: alexslemonade.github.io/refinebio-examples/03-rnaseq/differential-expression_rnaseq_01.html*
    b. *Create a volcano plot of your data. Make sure to label the axes and provide a legend.*
    c. *Create a table of the top 50 differentially expressed genes. Add this to your assignment writeup. Include the full table of all results in a results folder in your GitHub repository.*
    d. *Save and summarize your findings.*
4. *Extract the list of significantly differentially expressed genes, and generate a heatmap showing only those genes*
    a. *Example using ComplexHeatmap. Package reference (https://jokergoo.github.io/ComplexHeatmap-reference/book/)*
    b. *Add a side bar colored by sample groupings (cancer vs not, etc.)*
5. *Extract the list of differentially expressed genes and run gene set enrichment analysis.* ***Each student in your team should run a different combination of method and ontology*** *(e.g., if there are 4 students on the team, there should be results for 4 applications in your assignment writeup).*
    a. *Choose a method:*
        i. *topGO*
        ii. *clustProfiler*
        iii. *gProfiler2*
        iv. *GenomicSuperSignature*
        v. *PyDESeq2 (BioStars example here)*
        vi. *Wilcoxon rank-sum test*
    b. *Choose an ontology (e.g. Disease Ontology, Gene Ontology)*
    c. *Run enrichment analysis on your data using your selected method and ontology*
    d. *Create a table of these results. Add it to the results folder in your GitHub repository.*
6. *Create a table showing statistically significantly enriched terms (and any characteristics) shared by the method you used (e.g., q-value, p-value, log fold change). Include the full tables in a results folder in your GitHub repository. This will be a joint table of all your team's step 5 results. Each unique gene set / term should have 1 row and all results from each method in that row. Add a column indicating how many of the methods found this term to be significantly enriched in your analysis, and how many methods included this term in the analysis. Add the full table to your GitHub repository results Folder.*
7. *Using the table created in step 6, create a combined table that shows the top 10 terms enriched in all (or most) methods. Include this in your assignment writeup.*
8. *Write a short summary to go with each plot/table you create. Describe what you did, what parameters you used (if any) and an interesting result from it.*
9. *Combine all results into a single file, submit on Canvas. Each student on the team must submit their own work separately. Make sure that all your code and results are added to your GitHub repository.*