Bioinformatics Resources Project

Please complete the following tasks and provide commentary on the results obtained. You can choose between two formats for your submission:

1. Provide an R script containing the code along with a separate PDF report that describes your analyses and comments the results you obtained.
2. Use frameworks like R Markdown to combine both the report and the code in a single document. In this case, submit the final document in a readable format, such as PDF or HTML.

Overview: Select one of the available RData files (ensure that no more than two groups select the same dataset) representing RNA-seq count data extracted from different cancer datasets in The Cancer Genome Atlas (TCGA). The dataset includes both tumor (case) and normal (control) samples, which have been randomly selected from the original TCGA data.

1. **Load the RData file:** The following three data frames are available:
    - *raw_counts_df*: Contains the raw RNA-seq counts.
    - *c_anno_df*: Contains sample names and conditions (case or control).
    - *r_anno_df*: Contains ENSEMBL gene IDs, gene lengths, and gene symbols.

2. **Update *raw_counts_df* and *r_anno_df* extracting only protein-coding genes:**
    - Use the biomaRt package to retrieve the necessary information.
    - The following tasks should be performed using the new data frames created in this step.

3. **Differential Expression Analysis using the edgeR package:**
    - Annotate up- and down-regulated genes using the following criteria:
        - Adjusted p-value cutoff of 0.05.
        - Log fold change > 1.5 for up-regulated genes and < -1.5 for down-regulated genes.
        - Log CPM > 1.
        - If no or few results are obtained, relax the thresholds.
    - Follow the workflow developed during the course:
        - Filter the raw counts data to retain only genes with a raw count > 20 in at least 5 case samples or 5 control samples.
    - Create a volcano plot of your results.
    - Create an annotated heatmap focusing only on up- and down-regulated genes.

4. **Gene Set Enrichment Analysis using the fgsea R package:**
    - Focus on the C4 and C6 gene set collections, performing the analysis separately for

each collection.
- Report the top 10 up- and top 10 down-regulated gene sets and plot the GSEA table for each collection.
- Select one up-regulated and one down-regulated gene set and plot the enrichment score.

5. **Identify Transcription Factors (TFs):**
    - Identify any TFs with enriched PWM scores in the promoters of the genes in the top 10 up- and down-regulated gene sets, as defined in task 4:
        - Use a window of 500 nucleotides upstream of each gene.
        - Focus only on genes annotated as up- or down-regulated in task 3.

6. **Select one of the top-enriched TFs:**
    - Compute the empirical distributions of scores for all PWMs found in MotifDB for the selected TF.
    - Determine the distribution (log2) threshold cutoff at 99.75% (relax the threshold if necessary).

7. **Protein-Protein Interaction (PPI) Network using the STRING database:**
    - Use the STRING database to find PPI interactions among the differentially expressed genes (as defined in task 3).
    - Export the network in TSV format.

8. **Network Analysis in R using the igraph package:**
    - Import the network into R and perform the following:
        - Calculate and plot the network degree distribution.
        - Calculate degree, closeness, and betweenness centrality indexes for all genes in the network and identify the genes showing the top values for those indexes.
        - Identify and characterize the largest connected component in the network.