Bioinformatics Resources Project

Execute the following tasks and write a report describing and commenting the results you obtain. The code you generate should be provided as R script or you can use frameworks like R markdown.

Select one among the RData files available in folder “Project” in our gdrive/moodle representing RNA-seq count data extracted from different TCGA* cancer datasets. From the original TCGA data only 50 cases (tumor samples) and 50 controls (normal samples) were selected.

1. Load the RData in RStudio. The following three dataframes are available:
    - a *raw_counts_df* = contains the raw RNA-seq counts
    - b *c_anno_df* = contains sample name and condition (Case and Control).
    - c *r_anno_df* = contains the ENSEMBL genes ids, the length of the genes and the genes symbols

2. Update *raw_count_df* and *r_anno_df* extracting only protein coding genes.
    - a Use biomaRt package to retrieve the needed information
    - b Next tasks should use the new data frames you have created

3. Perform differential expression analysis using edgeR package and select up- and down-regulated genes using a p-value cutoff of 0.01, a log fold change ratio >1.5 for up-regulated genes and < (-1.5) for down-regulated genes and a log2 CPM >1. Relax the thresholds if no or few results are available.
    - a Use the workflow we developed during the course
    - b Filter raw counts data retaining only genes with a raw count >20 in at least 1 Case and 1 Control sample
    - c Create a volcano plot of your results
    - d Create an annotated heatmap focusing only on up- and downregulated genes

4. Perform gene set enrichment analysis using clusterProfiler R package.
    - a Perform both GO (BP and MF) and KEGG analysis

   b Report the top 10 enriched GO terms and the top 10 enriched KEGG pathways resulting from both up- and down-regulated gene lists

5. Use the pathview R package to visualize one pathway you find enriched using the up-regulated gene list.

6. Identify which transcription factors (TFs) have enriched scores in the promoters of all up-regulated (or down-regulated if you prefer) genes.
   a use a window of 500 nucleotides upstream each gene

7. Select one among the top enriched TFs, compute the empirical distributions of scores for all PWMs that you find in MotifDB for the selected TF and determine for all of them the distribution (log2) threshold cutoff at 99.5%.

8. Identify which up-regulated (or down-regulated depending on the choice you made at point 7) genes have a region in their promoter (defined as previously) with binding scores above the computed thresholds for any of the previously selected PWMs.
   a Use pattern matching as done during the course

9. Use STRING database to find PPI interactions among differentially expressed genes and export the network in TSV format.

10. Import the network in R and using igraph package determine which is the largest connected component.

* https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga