Introduction to Bioinformatics (236523)

HW 2 – Spring 2021

**General instructions:**

• Deadline: 24/05/21 23:59.

• Submission in pairs only.

• The submission is via the course website.

• You should submit the markdown file (.Rmd) and its word/pdf version in a .ZIP format named according to next format:

<HW#>\_<ID1>\_<ID2>.zip . Please decide which ID number of the two partners will be first and keep this convention for the rest of the semester. For example: HW2\_398837676\_234543234.zip.

**Detailed instruction for the HW:**

The answers to the questions should be submitted as a single markdown file. Please make sure that the output of the markdown file as doc or pdf (applying the Knit) option will give us the question number in bold large font.

**Question 1:**

* 1. Download the GSE88741 dataset count matrix from <https://maayanlab.cloud/archs4/series/GSE88741>. The downloaded file is in the .gz format and should be unzipped.
  2. Search for the information regarding the GSE88741 dataset at GEO database. Describe shortly what this dataset represents.
  3. Create a coldata table containing the GSM IDs (such as GSM2344965), cell line type and the condition: normal/cancer. To understand which GSM ID stands for normal and which for cancer, you’ve got 2 options:

1. You can browse the information at GEO at the GSE88741 page by GSM sample ID.
2. You can open the GSE88741 file. Use Ctrl+O or File->Open file at the RStudio window.
   1. Create a subset of 8 samples for analysis. Choose 2 samples from normal cell line samples and 2 samples for **each** cancer cell line. Overall, your dataset must contain 8 samples. Perform differential expression analysis on the dataset of cancer vs normal samples.
   2. Create a volcano plot for the differentially expressed genes. The volcano plot will present the log2fold-change(log2FC) vs. the −log10 adjusted p-value.
   3. Create a heatmap using the pheatmap() function for the read counts of top 30 differentially expressed genes. These genes must be statistically significant at FDR<0.05 and **downregulated** in cancer by at least 8x. The heatmap must be built from the scaled read counts for each of the genes in all the 8 samples.

The scaling of the data must be performed by row.

Hint: You can generate the volcano plot and the heatmap using the relevant code chunks in tutorial 5 and 6 as an example.

* 1. Choose one gene among the top 30 most upregulated genes and one gene among the top 30 most downregulated genes and briefly describe the known functions for each of the two genes.

**Question 2:**

1. Why the mapped reads are mapped preferentially to exons?
2. Sometimes if we map RNA-seq reads of certain tissue samples and we look on certain genes we will see that reads are also mapped to the intronic region. What can be biological cause/reason for such an observation?
3. Open the <https://www.genecards.org/> database and search for MARCH7 gene at this site.
   1. What are additional names (aliases) of this gene? Which databases these gene names were extracted from?
   2. What can be a possible problem if you will paste the name of this gene manually into Excel sheet ?

**Question 3:**

Use the expression data stored in sdy420.rds file from tutorial 6.

1. Cluster the samples data using three different clustering techniques – hierarchical clustering, k-means and dbscan. Use the `hclust`, `kmeans` and `dbscan` packages (Pay attention that you cluster by samples, not the genes, transpose the expression matrix if needed).

The dbscan installation instructions and information regarding package can be found at <https://github.com/mhahsler/dbscan>.

1. Try different parameters for each of the techniques to see how it affects the clusters. To visualize the clustering results using the different techniques you may use the `Rtsne` package and the following code:

tsne = Rtsne(t(sdy$expr))

plot(tsne$Y, col=clusters)

*\*tSNE allows to visualize the multidimensional data in 2D.*

1. For the `hclust` try different clustering methods as was shown in class. You can visualize the dendogram using the plot() function on the results of hclust. Which method in hclust() seems to give more accurate clustering according to the tSNE plot ?
2. Did you get the same or different clusters using the three techniques? Find ways to present the similarities and differences between the clusterings. If the clusters are different what can explain the difference?
3. The demographic data of each of the samples can be found in the sample\_info.csv file from tutorial 6. Can demographic variables explain the way the data is clustered? Choose one of your clustering results (that is not trivial) and find the features that can explain the differences between the clusters.

Make sure that the answer to the question includes the code you use to experiment with parameters and to look at the effect of the demographic features on clustering.