SBE304 – Biostatistics.
Final Project.
Deadline: Tuesday – January 19th, 2021, at 8pm.
Presentation: Wednesday – January 20th, 2021.

- You can submit and present earlier. Please arrange with me for that.
- No late submissions.
- Only one member from each group should submit the project files to Google Classroom.

Teams: Each team should have 3-4 students.

<u>Project Description:</u>

You are required to conduct a study to analyze gene expression (GE) data for the cancer type Lung Squamous Cell Carcinoma (LUSC).

There are two files for the GE data (sent with this statement) for this cancer type:

1. "lusc-rsem-fpkm-tcga-t_paired.txt": GE data for tissues with cancer,
2. "lusc-rsem-fpkm-tcga_paired.txt" : GE data for tissues in a healthy case.

- Each row in the two files represent a gene, and the columns represent the expression level of this gene in different samples.
- Data are paired meaning that a healthy sample and a diseased sample are taken from the same subject. So, both GE files have the same number of cases with the same order.
- Files are tab-separated.

Requirements:

- **<u>Correlation.</u>** Compute the correlation between the normal samples and the diseased samples for each gene.
  - Rank genes based on their correlation coefficient (CC) and report the highest positive CC and the lowest negative CC and the names of these two genes.
  - Plot the expression levels of the above two genes.

- **<u>Hypothesis Testing.</u>** Infer the differentially expressed genes (DEGs); the genes whose expression level differ from one condition (healthy) to another (diseased).
  - Apply the appropriate test statistic for the following two pairing cases:
    1. Samples are paired,
    2. Samples are independent.
  - Apply the FDR multiple tests correction method.
  - Report the set of DEGs before and after the FDR correction for each of the above two pairing cases.
  - Compare the two DEGs sets (paired and independent) after the FDR correction in terms of the common and distinct genes.

- **Useful python functions.** Search for these Python functions, and if you think they are useful to your project, you can use one or all of them beside any other function not mentioned here.
  - scipy.stats.pearsonr
  - scipy.stats.spearmanr
  - scipy.stats.ttest_ind

    - scipy.stats.ttest_rel
    - statsmodels.stats.multitest.multipletests

- Support your findings/results/conclusions with figures.

- You have to deliver the following:
    - All the code scripts you used for your analysis,
        - Comments are a must.
    - Project report:
        - It should look like a research paper. It should have the following sections:
            - Introduction,
            - Methods: describe all the steps carefully and include all the used software packages,
            - Results and Discussion: report your results in details and discuss them,
                - You can augment your results with textual files or spreadsheets.
            - Conclusion: list the overall findings of your analysis.
            - **Members Contribution: list in details what each member in your group did in this project. Each member in the group may receive a different grade based on the contribution weight.**

- Presentation:
    - You will be given a few minutes to represent your work online.
    - Prepare yourself for discussing your analysis and findings.

Good luck!