# Study to analyze gene expression data for (LUSC).

Team 4

January 2021

## 1 Introduction

We knew that the chromosomes take form of DNA inside the nucleus and the chromosomes divided into different sections called genes. So, the stem cell after protein synthesis is differentiated into different cells according to what genes are expressed. To know that the tissue is cancerous because of genetic disease or gene expression, we take RNA sample and compares it to a healthy one, if the sample is highly transicripted so, this gene is expressed.

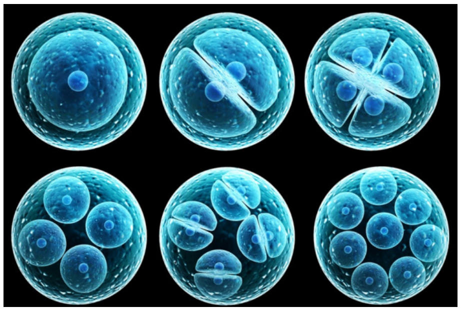

Figure 1: Gene expression

## 2 Methods

### 2.1 Packages Used

- scipy.stats
- numpy
- matplotlib.pyplot
- scipy
- statsmodels.stats.multitest
- tabulate

## 2.2 Steps

1. **Reading** the two txt files by open() method in python, and putting each row for every file in a list with .split() method, and using .append() method to make two lists of lists, one for the genes in healthy tissues and the other for genes in cancerous tissues (list_of_healthy_tissues , list_of_cancerous_tissues), each one contains a list for each row as a list.

2. **Filtration**

   (a) Define two counters, one for each list, and giving them initial values of 0 (Zeros,Zeroz) for healthy values and cancerous values respectively.

   (b) Creating an outer for loop that takes zip() for both list of lists in order to iterate on both lists in parallel

   (c) Creating an inner for loop that iterates on each element inside the lists and whenever an element is zero the counter for that list (Zeros/Zeroz) increments by one.

   (d) If both counters are less than or equal to 25 (which is half the number of columns) we take the list as it is and put it in a new list.

   (e) So now we have two new lists of lists that don't contain rows of zeroes.(NamesForHealthy , NamesForCancerous)

   (f) We made two other lists that don't have the first two elements in each list so we could have lists that contain only the values without the name and ID (list_of_healthy_tissues_without_Zeros , list_of_cancerous _tissues_without_Zeros).

   (g) Taking the values directly from the .split() method gave us string elements so we converted the elements in the previous two lists into float elements in order to correlate them.

3. **Correlation**

   (a) Creating a for loop that iterates on zip() of both list of lists (list_of_hea lthy_tissues_without_Zeros , list_of_cancerous_tissues_without_Zeros)

   (b) Applying correlation on each two lists together of the two list of lists.

   (c) Making a new list Correlations_list that contains the values of these correlations

   (d) Getting the maximum and minimum values of this list (Correlations_list) and also using the indexes of these values to get the names of the genes that had the maximum and minimum correlation values

   (e) Plotting the values of the list of the healthy gene with maximum correlation on the x-axis, and the values of the list of the cancerous gene with the maximum correlation value on the Y-axis to get the graph of the positive correlation

(f) Plotting the values of the list of the healthy gene with minimum correlation on the x-axis, and the values of the list of the cancerous gene with the minimum correlation value on the Y-axis to get the graph of the negative correlation

4. **Hypothesis**
Creating a for loop that iterates on zip() of both list of lists (list of healthy tissues without Zeros , list of cancerous tissues without Zeros) to get the p-values of tissues. It does so by apply stats.ttest_rel in paired samples case and stats.ttesy_ind in independent samples case. Apply FDR multiple test correction method on the p- values of paired and independent samples using multipletests function with returns both corrected values and reject boolean tells either they are greater than Alpha(False) or smaller (True). Detect either p-values before the correction are less than Alpha (True) or not (False). Compare the two reject boolean values of before and after FDR. We can compare the paired and independent DEGs sets after the correction, the values that remained True are common and the values that changed to False are distinct

# 3 Results and Discussion

## 3.1 Correlation

| | Correlation Value | Gene Name |
|---|---|---|
| Highest Positive | 0.9690441442970704 | AREGB |
| Lowest Negative | -0.45280727852470837 | FAM222B |

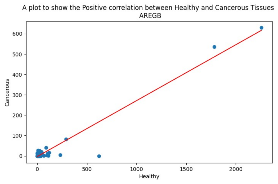

Figure 2: Positive Correlation

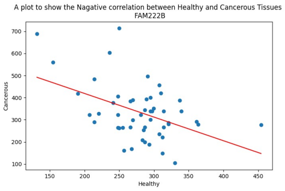

Figure 3: Negative Correlation

### 3.2 Hypothesis

Paired case: -before applying FDR method: number of genes is 12724 -after applying FDR method: –common genes number is 12410 –distinct genes number is 314

Independent case: -before applying FDR method: number of genes is 12631 -after applying FDR method: –common genes number is 12320 –distinct genes number is 311

| Common p-values | Distinct p-values |
|---|---|
| 1.45354e-07 | 0.0522795 |
| 0.000458941 | 0.0525569 |
| 2.45458e-11 | 0.0618698 |
| 8.13801e-06 | 0.0610962 |
| 4.55716e-05 | 0.056425 |
| 2.76867e-10 | 0.0646677 |

| Common p-values | Distinct p-values |
|---|---|
| 1.37807e-08 | 0.0650821 |
| 0.000172169 | 0.0643707 |
| 5.03496e-14 | 0.0618698 |
| 1.33719e-05 | 0.0530667 |
| 1.92174e-05 | 0.0544534 |
| 6.02266e-09 | 0.0678727 |

# 4 Conclusion

# 5 Members Contribution

| Name | Contribution |
|---|---|
| Khloud Abdelazeem | Hypothosis Testing |
| Mariam Osama | Correlation and Filteration |
| Meirna Kamal | Correlation and Filteration |
| Miran Mahmoud | Hypothosis Testing |