SBE304 – Biostatistics Final Project

Ahmed Mohamed Abdelfattah Sec:1 BN:5
Anas Mohamed Abdelrahman Sec:1 BN:17
Mostafa Mohamed Essam Yehia Sec:2 BN:27
Mohamed Khalaf Sec:2 BN:48

I. Introduction

We have conducted a study to analyze gene expression data for the cancer type Lung Squamous Cell Carcinoma. We used two paired gene expression data sets, one for tissues in a healthy state and another with cancer

II. Methods

We conducted our test using Python, which is very useful when dealing with statistical tests. As our code shows, the steps were as follow:

1) We opened the data sets' files using functions from **Pandas** library and stored them as dataframes
2) We filtered the data sets from rows that have zero values with more than 50% of the values
3) We calculated Pearson's correlation coefficient using **pearsonr** function in **scipy.stats** module and the in-built **min()** and **max()** functions in Python
4) We plotted our results using functions from **matplotlib.pyplot** and **numpy** libraries
5) We conducted the hypothesis test using **ttest_rel** function for paired-sample case and **ttest_ind** function for independent-sample case from **scipy.stats** module
6) We applied then the FDR correction method using **multitests** function from **statsmodels.stats** module
7) We then compared the common genes which were not affected after the FDR correction with the distinct ones which were affected
8) We used the **tabulate** library to format the dataframe output in a nice way

III. Results and Discussion

A. Correlation

We found that the gene, which has the highest positive correlation coefficient, was **AREGB** and had a correlation coefficient of 0.9690441442970706. The gene with the highest negative correlation coefficient was **FAM222B**, with a correlation coefficient of -0.4528072785247083.

```
(base) ta7a@ta7a-ubuntu:~/Biostats Project$ python -u "/home/ta7a/Biostats Project/project.py"
Highest CC is:
0.9690441442970706
Its name is:
AREGB
Lowest CC is:
-0.4528072785247083
Its name is:
FAM222B
```

Fig. 1. Results

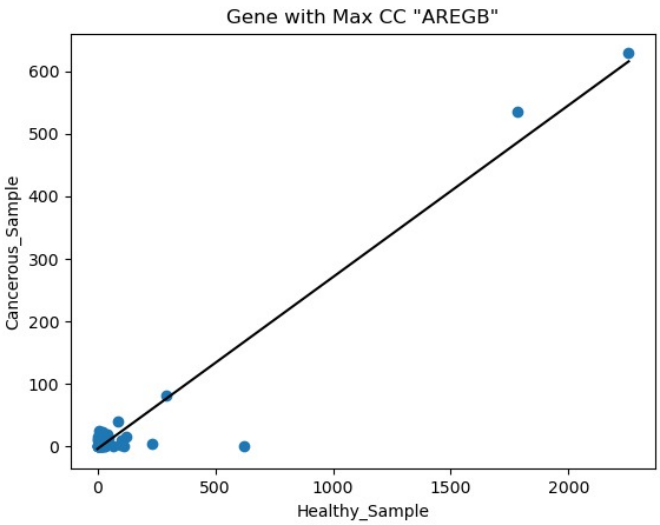

Fig. 2. Highest correlation coefficient

Fig. 3. Lowest correlation coefficient

B. Hypothesis test

We found that the genes that satisfied our hypothesis (the genes whose expression level differ from a healthy state to being cancerous) were 12410 genes in the paired-sample case and 12320 genes when the samples were independent. These were the results after running the FDR correction method. There were 314 rejected genes from the hypothesis after

running the FDR correction in the paired-sample and 311 genes when they were independent.

Common genes - Paired

	Gene_name	p_values	p_values_fdr
0	HIST3H2A	4.04361e-08	1.45354e-07
2	LXN	0.000232237	0.000458941
3	CNKSR2	3.42058e-12	2.45458e-11
6	GSDMD	3.04172e-06	8.13801e-06
7	AKR1C1	1.93857e-05	4.55716e-05
8	C3orf62	4.76856e-11	2.76867e-10
9	CRISPLD2	1.37661e-05	3.31661e-05
11	SLC33A1	7.58445e-10	3.60449e-09
12	GLI1	0.0213018	0.0308529
13	STK17B	4.51183e-13	3.78614e-12

Fig. 4. A sample of the common genes in case the samples were paired

Common genes - Independent

	Gene_name	p_values	p_values_fdr
0	HIST3H2A	3.60714e-09	1.37807e-08
2	LXN	8.16404e-05	0.000172169
3	CNKSR2	6.37465e-15	5.03496e-14
6	GSDMD	5.34429e-06	1.33719e-05
7	AKR1C1	7.85788e-06	1.92174e-05
8	C3orf62	1.50072e-09	6.02266e-09
9	CRISPLD2	9.74275e-05	0.000203384
11	SLC33A1	6.38206e-11	3.02559e-10
12	GLI1	0.0239804	0.0347499
13	STK17B	5.02413e-13	3.10862e-12

Fig. 5. A sample of the common genes in case the samples were independent

Distinct genes - Paired

	Gene_name	p_values	p_values_fdr
34	SHKBP1	0.0375699	0.0522795
147	TDRKH	0.0377832	0.0525569
161	BAI1	0.0478737	0.065379
165	DZIP1	0.0445121	0.0610962
208	ELP3	0.0408123	0.056425
235	PPDPF	0.047312	0.0646677
308	CLEC11A	0.0481667	0.065748
393	CERK	0.047102	0.0644012
533	NMBR	0.0452342	0.0620039
534	GPATCH8	0.0458589	0.0628155

Fig. 6. A sample of the distinct genes in case the samples were paired

Distinct genes - Independet

	Gene_name	p_values	p_values_fdr
4	SCML1	0.0472659	0.0650821
10	DOCK5	0.0467158	0.0643707
161	BAI1	0.0447545	0.0618698
165	DZIP1	0.0379153	0.0530667
180	PARP11	0.038994	0.0544534
278	MFAP5	0.0494296	0.0678727
353	ZNF354A	0.0379478	0.0531037
393	CERK	0.0469104	0.0646183
534	GPATCH8	0.0358471	0.050408
548	MTPN	0.0361542	0.0508111

Fig. 7. A sample of the distinct genes in case the samples were independent

IV. Conclusion

To sum up, Most of the genes satisfied the hypothesis test and some of the genes showed a negative correlation coefficient, which means that the expressions of these genes were changed when the healthy tissues became cancerous.

V. Members Contribution

We thought that it was not necessary to make each member work on a part of the project on his own. So, we worked as a team in writing the code, searching for the proper libraries and modules, and making the presentation slides. Ahmed Abdelfattah helped in writing the report using **LaTeX**