Bioinformatics Final Project

Introduction:

Gene expression is the process that describes the DNA or genes usage either to proteins or any other of its forms. In all cases of the Normal human genome, it should be stable as the **differentiation** of its activity (**Transcription**) should always be zero. Some cases have a suitable range of variation as its activity may alter according to a change in human interference and development. For example, the differentiation of the INS gene which is responsible for insulin expression may increase or decrease according to the amount of sugar intake. However, this range of differentiation has a limit if exceeded by either increasing or decreasing it will cause malfunctions. It will be simple enough if the target is one gene as INS, however in respect to unknown malfunction of unknown gene a set of data should be used, and an approach should be used to define the suppressor or inheritor gene**(s)**, this approach can be statistical analysis such as Hypothesis test, or non-statistical such as fold change which is based on comparison of ratio score of average gene expression can be used. Either it is statistical Analysis or Non-statistical, in both cases, the expression levels for each gene should be provided. In our case the provided data was for the Lung Squamous Cell Carcinoma (LUSC) from the cancer genome atlas TCGA, we have 50 samples in both cases of cancerous and healthy tissue. In this project, we will use different methods to determine the differentially expressed genes DEGs.

Methods :

I. Hypothesis testing

First, we have filtered our data by dropping rows with more than or equal to 25 zero values. We checked the normality for the expression levels of the genes using the Shapiro-Wilk test, and we found that a number of genes have non-normal distributions, so we decided to use the Wilcoxon test to compute the p-value for the difference in the gene's expression levels. We had assumed 2 cases, the first one that the genes are from independent samples, and in the second case the samples are paired. In the case of independent samples, we used the Mann-Whitney U function of the SciPy package in python, and in the case of paired samples, we used the Wilcoxon function of SciPy. Then we applied an FDR correction method to both lists with $\alpha = 0.05$ and stored the new p values. We determined DEGs by comparing the p-values after FDR correction to the alpha value. If the p-value for a certain gene is less than alpha then it is considered DEG. Then we determined the common in the two cases of paired and independent samples and the distinct DEGs in each case.

II. Fold change method:

This method is dependent on gene expression levels as an input, and to extract the output, several steps should be taken:

- We should check that our data is not logged to the base 2, in this case a different pipeline should be used.
- Take the mean value for all healthy samples and for all tumor samples
- Take the log to the base 2 for the ratio between tumor means to the healthy means.
- Compare the logged values to a certain threshold, if the gene expression level significantly differs from healthy to tumor, so the absolute logged FC should exceed the threshold given.

III. Volcano plot:

This is done by taking the intersection between the DEGs determined by the statistical method (hypothesis testing) and the DEGs determined by the non statistical method (Fold change). So the number of DEGs in this step should not exceed the number of DEGs in each case of the 2 previous cases. Then we used the GSEA Software as a robust technique to extract some extra features like enrichment score.

III. The used Packages:

We have used the following python packages in this project:

- Scipy: used to perform Shapiro-Wilk, Wilcoxon, Mann-Whitney U tests
- Pandas: used for storing the data in data frames.
- Statsmodels: used for applying FDR correction.
- Matplotlib : used for plotting histograms for data rows.
- Bioinfokit : used for plotting volcano plot.

Results and Discussion:

I. Hypothesis testing:

In the case of independent samples, we have 13234 DEGs, and in the case of paired samples, we have 13141. The number of DEGs in the case of paired samples is less than in independent samples. The DEGs determined in case of independent cells are stored in “ind_hypo_degs.csv” , and in case of paired samples are stored in “paired_hypo_degs.csv” file. There are 12667 common DEGs between the two cases, 567 distinct DEGs in case of independent samples, and 474 distinct in case of paired samples.

II. Fold change method:

In this method we aimed to get the most significant 2000 genes, So using trial threshold we found that 2.2765 is the threshold which will gives us 2000 DEGs which their absolute logged FC will exceeds this threshold

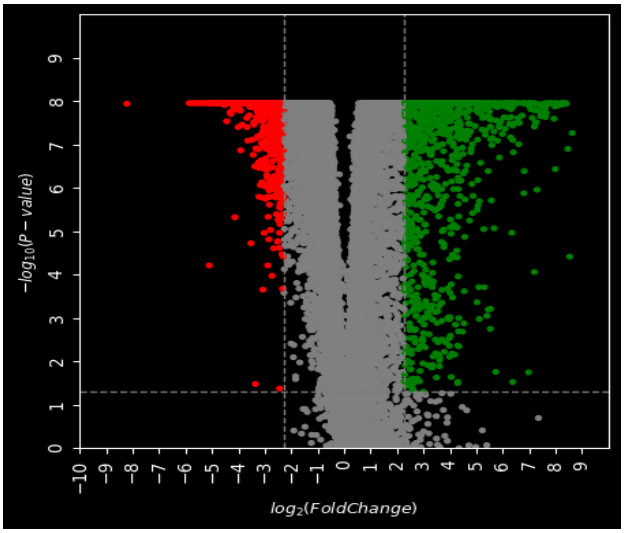
Hugo_Symbol	Mean healthy	Mean tumor	Log(FC)
GPX3	32299.1732	2821.4098	-3.517009
LRP2BP	96.6522	13.3372	-2.857347
AOC3	4360.0730	439.4352	-3.310630
SLC5A8	47.5624	8.9344	-2.412379
PNMA2	173.5604	32.9236	-2.398244

Fig.1 Example for 5 genes exceeds the threshold

For a full DEGs list, please check the resulting **fold_change.txt** file for the 2000 genes.

III. Volcano plot:

There are 2 volcano plots generated from the previous 2 methods.



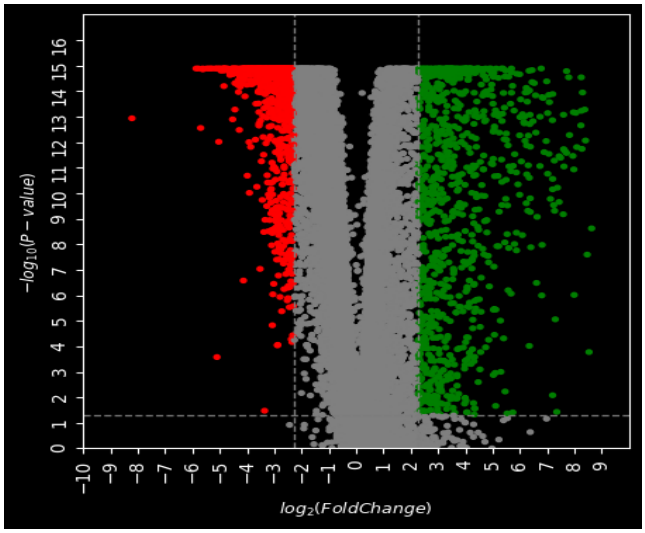


Fig.2 volcano plot for paired assumption

Fig.3 volcano plot for independent assumption

The common thing between the 2 plots is that the shape of the volcano plot generated is like an inverted volcano which is produced by p values.

Also we can see that the independent plot has higher amplitude (p values) than the paired one.

IV. GSEA Software:

Using the software for GSEA which needs as an input gct,cls and gmt files. The output was as follows:

- The number of markers for phenotype normal: 899 (46.5%)
- The number of markers for phenotype cancer: 1035 (53.5%)
- The global ES histogram is as the following image
- The correlation profile as follows:

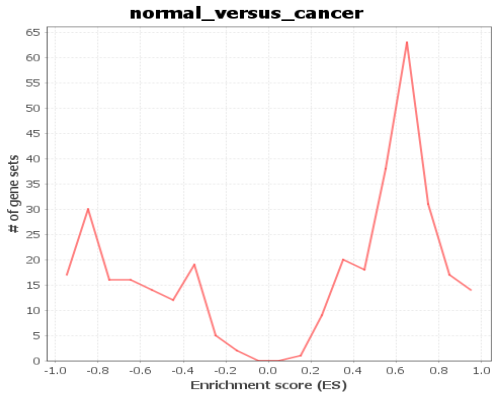


Fig.4 ES histogram

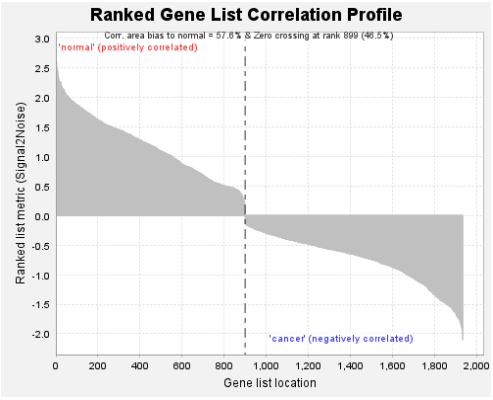


Fig.5 Correlation profile

Conclusion:

We have applied Fold change, hypothesis testing on TCGA dataset for Lung Squamous Cell Carcinoma (LUSC) in case of paired samples to determine the DEGs, then we used volcano plot method to get the common DEGs between the two methods. We have performed Gene Set Enrichment Analysis (GSEA) on the set of DEGs obtained by the volcano plot method, and we found that the number of markers for cancer phenotype was about 1035 genes which was similar to fold change results.

Contributions:

Member	Section	BN	Contribution
Abdallah Moahmmed Shehata	1	49	Fold change method, GSEA software
Ahmed Sayed Ahmed	1	4	Checking for normality ,Hypothesis Testing (paired, indep)
Ezz Eldien Ismail Ezz Eldien	1	50	Volcano plot