Gene Expression Analysis of Lung Squamous Cell Carcinoma

Abdel-Salam Walid, Omar Sayed, Mahmoud Mohammed Abdel-Moneim, Zeyad Khaled

Systems and Biomedical Engineering, Faculty of Engineering, Cairo University

Gene Expression Analysis of Lung Squamous Cell Carcinoma

This paper is concerned with the analysis of gene expression levels (GE) of tissues infected with Lung Squamous Cell Carcinoma. The paper's main objectives are to compute the correlation coefficients between genes collected from healthy and infected tissues and report the set of differentially expressed genes (DEGs) before and after applying the false discovery rate (FDR) correction. All analysis was carried out in Python.

Data Collection

We were provided two tab-separated text files; one for genes taken from healthy tissues and the other for the same genes taken from infected tissues. Each file consisted of 19648 genes and each gene was sampled 50 times and on each time the GE level was noted. The data was paired meaning, they had the same order with the same number of cases.

Data Cleaning

The process of data cleaning consisted of replacing missing GE values of a gene with the mean of the samples unless it had more than 50% null samples, those missing above 50% led to gene removal in both files. In the end we had 17459 genes left to work with.

Data Analysis

We had two main objectives, the first was to compute the Pearson's correlation coefficients of the genes in healthy and infected tissues, rank them, and plot GE levels of the highest and lowest correlated genes. The second was to conduct hypothesis tests on the two

samples to infer the differentially expressed genes before and after applying the false discovery rate.

Correlation

Our goal was to find out whether genes in tissues infected with lung squamous cell carcinoma had different expression levels than genes in healthy tissues. We calculated Pearson's correlation coefficient of all genes in healthy and infected tissues. Pearson's correlation coefficient is given by:

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - x)^2 \sum(y_i - \overline{y})^2}}$$

$r = correlation\ coefficient$

$x_i = values\ of\ the\ x - variable\ in\ the\ sample$

$\overline{x} = mean\ of\ the\ x - variable$

$y_i = values\ of\ the\ y - variable\ in\ the\ sample$

$\overline{y} = mean\ of\ the\ x - variable$

The gene AREGB had the highest coefficients between healthy and infected tissues (0.969044), meanwhile, FAM22B had the lowest coefficients (-0.452807).

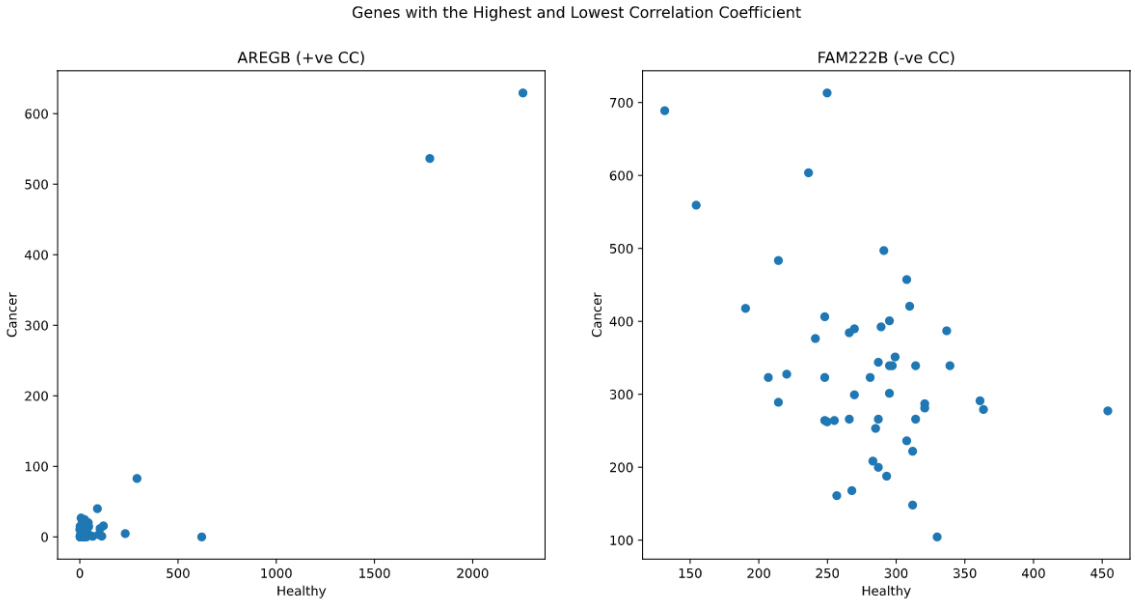

Figure 1

Hypothesis Testing

We carried out hypothesis testing to infer the DEGs. We Applied T tests on two separate cases with significance level at 5%, then computed the P value for every gene and reported the DEGs. Then we applied the FDR multiple tests correction method and compared the results before and after the correction method. In the first case, we considered the samples paired and in the second case we considered them independent. At last, we compared the paired and independent sets after the correction.

As expected, the FDR correction method reduced the number of DEGs, as it is more stringent to other methods.

	Paired	Independent
Raw	12787	12694
Corrected	12469	12380

Figure 2

Tools

All the analysis and data cleansing were carried out using python packages: Pandas, Numpy, Scipy, Statsmodels, Matplotlib. You can find the code here.