Cairo University

Faculty of Engineering

Systems and Biomedical Department

First Semester - 2024/2025

Study Analysis for Lung Squamous Cell Carcinoma (LUSC)

Team Members

Reem Adel

Hesham Tamer

Mariam Magdy

Mina Adel

Supervised by: Dr. Ibrahim Mohamed

Contents

Introduction

Lung squamous cell carcinoma (LUSC) is a major subtype of non-small cell lung cancer (NSCLC), characterized by distinct molecular and pathological features. Gene expression (GE) analysis provides crucial insights into the molecular mechanisms underlying LUSC and aids in identifying differentially expressed genes (DEGs) that could serve as potential biomarkers or therapeutic targets. This study aims to perform a comprehensive analysis of paired GE data from cancerous and healthy tissues in LUSC patients, utilizing statistical and computational approaches to identify DEGs and perform gene set enrichment analysis (GSEA).

Medical Background

Genomes and Molecular Structures

Genomes

The genome is the complete set of an organism's DNA, including all its genes. It contains instructions for cell structure, function, and regulation. In humans, the genome consists of about 3 billion DNA base pairs organized into 23 chromosome pairs within the nucleus of each cell. These chromosomes carry all the genetic information needed for life processes.

DNA (deoxyribonucleic acid) is a molecule that stores genetic information in a double-helix structure. Each unit of DNA, called a nucleotide, includes a sugar, a phosphate group, and one of four bases: adenine (A), thymine (T), cytosine (C), or guanine (G). These bases pair specifically (A-T, C-G), helping DNA replicate accurately during cell division and pass genetic information to offspring.

Genes

DNA segments called genes contain instructions for producing proteins, which are essential for cellular functions. The process by which genes are utilized to produce proteins is known as gene expression. Transcription is the process of creating RNA from DNA, while translation is the process of creating proteins from RNA. To guarantee that proteins are synthesized at the appropriate time and location, this process is meticulously controlled. Gene alterations or changes in how genes are regulated can raise the chance of developing diseases like cancer. have a better understanding of the interactions between genes and their regulatory components.

Gene Expressions (GEs)

The dynamic process of gene expression involves the transcription and translation of genetic information into useful proteins. It is highly controlled and differs depending on the cell type and environment. One of the characteristics of cancer is dysregulated gene expression.
Tumors suppressor genes, which prevent proliferation, may be downregulated, whilst oncogenes, which stimulate cell growth, may be overexpressed. Examining patterns of gene expression provide information on the molecular processes that underlie malignant behavior.

Lung Squamous Cell Carcinoma (LUSC)

The epithelial cells that line the airways are where lung squamous cell carcinoma (LUSC) begins. Usually, these cells are scale-like and flat. Knowing the genetic changes unique to LUSC is essential since it clarifies the causes of its aggressive character and helps find possible therapeutic targets.

Methods

Data Preprocessing

Data Description

Two tab-separated GE datasets were analyzed:

- **Cancer Tissue Data**: "lusc-rsem-fpkm-tcga-t_paired.txt"
- **Healthy Tissue Data**: "lusc-rsem-fpkm-tcga_paired.txt"

Pre-processing

Ensuring the quality and integrity of the dataset is crucial during the data preparation process. Our method entails carefully examining every dataset to find and fix any missing information. This is crucial since incomplete data might generate biases and errors in subsequent studies.

Identification of Missing Values and Handling Values

Initially, we verify that the names and IDs of the genes are consistent throughout the files.
Each data point is examined by our code as it iterates through the dataset to find any instances where information is absent (50% zero data or more).
Genes that contain more than or equal to 50% zeros are excluded from additional examination. This prevents genes with a large percentage of missing data from unnecessarily affecting the outcomes.

Differential Expression Analysis

One crucial step in identifying the genes that significantly contribute to the development of cancer is Differential Expression Analysis (DEA). A **specialized** method was used because our samples were paired.

Paired Sample Analysis

We used a hypothesis testing strategy that was peculiar to this design since the samples were paired, meaning that each malignant sample matched a particular sample of healthy tissue. This guarantees that the dependence between matched samples is appropriately taken into account in the analysis. The objective is to use a strong statistical method of hypothesis testing to find Differentially Expressed Genes (DEGs).

Independent Sample Analysis

In the alternative scenario where samples are independent (i.e., no specific pairing between cancerous and healthy tissue samples), a different approach is necessary. We will conduct hypothesis testing designed for independent samples. This involves assessing the normality of both the normal and cancerous datasets for lung cancer.

Hypothesis Testing

Identifying the Differentially Expressed Genes (DEGs). We do this by employing hypothesis testing, which is a robust statistical approach.

Choice of Statistical Test

We must precisely identify the DEGs for this important stage. We use hypothesis testing techniques to accomplish this. In particular, we employ the Shapiro-Wilk test for normality. In our analysis, this test is essential. It assists us in determining if the data on gene expression follows a normal distribution. This normality assumption is crucial, particularly for some parametric tests.

Paired Sample Analysis

Our samples are uniquely paired, with each cancerous sample directly linked to a corresponding healthy tissue sample. This pairing creates a dependency that requires a specialized analytical approach to account for these relationships.

To address this, we employ the Wilcoxon signed-rank test. This non-parametric test is particularly effective for comparing the distribution of differences between paired samples. Unlike parametric tests, it does not assume a normal distribution, making it especially suitable when the data's normality cannot be guaranteed. By using this method, we ensure that the paired nature of the data is appropriately considered, resulting in more accurate and meaningful insights.

Independent Sample Analysis

In cases where the samples are independent—such as when no specific pairing exists between cancerous and healthy tissue samples—our analytical approach must address this lack of dependency. A methodology designed for independent datasets is essential to produce valid comparisons.

For this analysis, we use the Wilcoxon rank-sum test. This non-parametric test is well-suited for comparing the distributions of two independent groups. Its strength lies in its robustness to deviations from normality, making it an ideal choice when the data do not meet parametric assumptions. This approach was chosen to ensure the analysis remains valid and reliable across diverse data distributions, providing a robust framework for comparing independent samples.

Significance Assessment

Our method goes beyond only detecting differences in order to understand the variations in gene expression between malignant and healthy tissues. Through accurate computations of test statistics and p-values, we lay a high priority on thoroughly assessing the significance of these discrepancies.

We modify our techniques according to the distributional properties of the data to guarantee the validity of our conclusions. We use the Shapiro-Wilk test to confirm this assumption for datasets that show a normal distribution. We rely on non-parametric techniques that are specific to the analysis when the data deviates from normality.

The Wilcoxon signed-rank test, which is very useful for evaluating the distribution of differences between paired samples, is what we use for paired sample analysis. It is the best option in this situation because of its capacity to manage data that is not normally distributed.

The Wilcoxon rank-sum test, which is excellent at comparing distributions between independent groups without making the assumption of normality, is used in independent sample analysis. By precisely measuring the probability that observed differences are the result of chance, these techniques give us a solid and trustworthy framework for determining significance that is adapted to the particulars of both paired and independent datasets.

Fold Change

Fold change is a crucial statistic in bioinformatics that helps determine how much gene expression levels vary between experimental settings. In order to have a better understanding of the magnitude of these changes, our analysis extends beyond simply detecting gene differentials. Fold change is essential to this endeavor.

Fundamentally, fold change measures the proportion of gene expression levels in two circumstances, such as the healthy tissues in our study and the malignant tissues. We can ascertain not only whether a gene is differentially expressed but also the degree of its variation thanks to this metric, which provides a clear and interpretable estimate of the magnitude of change.

A key technique in locating Differentially Expressed Genes (DEGs) is fold change. Combining it with statistical techniques, such the Wilcoxon rank-sum test for independent data or the Wilcoxon signed-rank test for paired samples, enables us to identify genes that show both physiologically important changes and statistically significant differences.

Both normal and malignant expression levels are given an epsilon (ε) number before the log2 transformation to guarantee the accuracy of fold change computations. This modification guarantees that the metric is still valid even at very low expression levels by avoiding infinite or undefinable results.

By using fold change in our study, we can better understand the biological significance of changes in gene expression. Significant physiological effects could result from small but steady changes in several genes. Thus, fold change turns into a powerful tool for understanding molecular dynamics, improving the scope and accuracy of our bioinformatics research.

Volcano Plots

Volcano plots are a cornerstone of bioinformatics, providing a comprehensive visualization of gene expression patterns by plotting statistical significance against fold change. Each data point represents a gene, with the most extreme values—resembling the peaks of a volcano—highlighting genes of particular interest.

These plots are invaluable for identifying Differentially Expressed Genes (DEGs) by simultaneously showcasing both statistical significance and the magnitude of biological changes. Genes positioned at the extremes of the plot exhibit large fold changes and high statistical reliability, enabling researchers to focus on genes that demonstrate both meaningful alterations and substantial biological impact.

The capacity of volcano plots to guide focused research into the molecular mechanisms underlying particular circumstances is what gives them their actual value. Finding genes that are both statistically and biologically significant allows researchers to quickly rank candidates for further investigation. This focused method improves comprehension of the molecular dynamics underlying changes in gene expression while streamlining analysis.

Volcano plots are essentially effective navigational aids that direct researchers to the most significant and pertinent genes within intricate datasets. They aid in the prioritization and interpretation of DEGs in addition to their visualization function, which eventually leads to a deeper comprehension of the biological processes involved.

Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is a powerful computational method used to determine whether a predefined set of genes shows statistically significant enrichment at the top or bottom of a ranked list of genes. This section details the GSEA conducted for gene sets relevant to lung cancer.

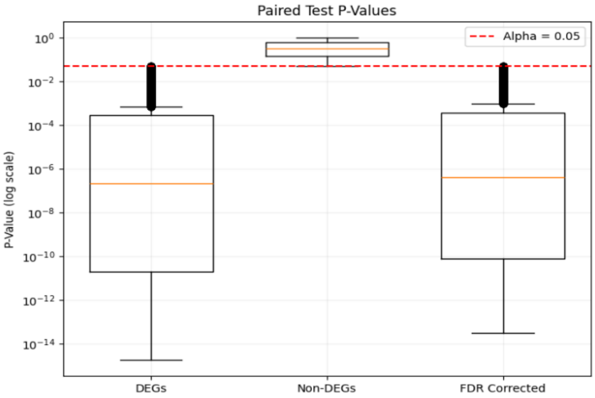
The goal of this analysis is to identify biologically significant pathways or processes associated with upregulated or downregulated genes in the dataset. GSEA evaluates whether a particular gene set (e.g., related to lung cancer cell lines) is overrepresented in the most upregulated or downregulated genes.

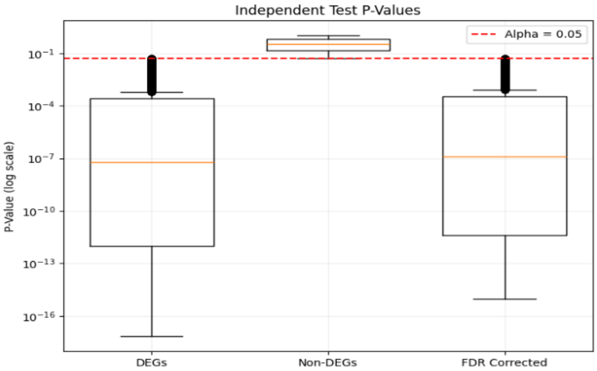
Key Metrics:

1. Enrichment Score (ES):
   - Measures the overrepresentation of a gene set in the ranked list.
   - Positive ES indicates enrichment at the top (upregulated genes).
   - Negative ES indicates enrichment at the bottom (downregulated genes).
2. Normalized Enrichment Score (NES):
   - Normalizes the ES to account for differences in gene set size and list variability, allowing comparisons between gene sets
3. P-value:
   - The probability of observing the given ES (or more extreme) by chance.
4. False Discovery Rate (FDR):
   - Adjusted P-value to control for multiple testing, ensuring robust statistical significance.

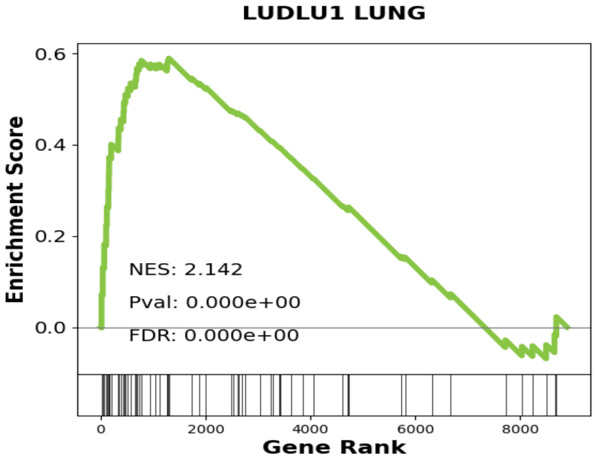
Results:

- The figures below show the P-values of DEGs and Non-DEGs before and after FDR correction for paired and independent data.

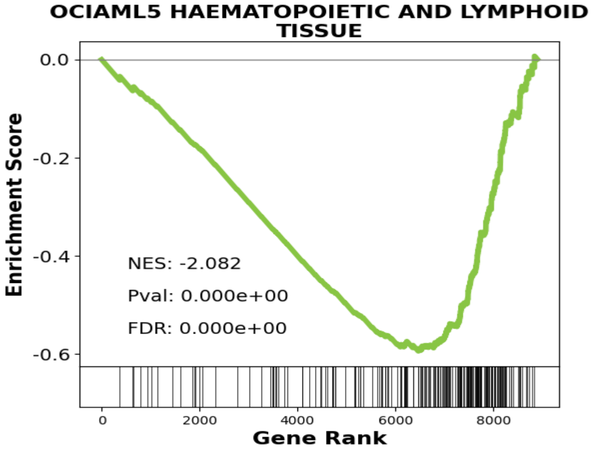




- **Comparison of GSEA Results:**



The **LUDLU1 LUNG** gene set shows significant upregulation, highlighting lung-specific processes potentially driving the observed condition.



The **OCIAML5 HAEMATOPOIETIC AND LYMPHOID TISSUE** gene set shows significant downregulation, with a focus on suppressed immune or blood cell pathways.

Software Packages Utilized

Python: Python served as the foundation of our analysis due to its versatility and extensive library ecosystem. Its robust tools and frameworks enabled efficient data processing, statistical analysis, and visualization.

Pandas: This essential data manipulation library allowed us to seamlessly organize, clean, and transform the gene expression datasets. It provided a framework for handling large, structured data efficiently.

NumPy: NumPy was used for advanced numerical operations, including array handling and mathematical computations, forming the backbone of our data calculations and transformations.

SciPy: SciPy's statistical functions were pivotal in conducting hypothesis tests, including the Shapiro-Wilk test, Wilcoxon signed-rank test, and rank-sum test. Its advanced tools ensured rigorous statistical analysis.

Matplotlib and Seaborn: These visualization libraries were utilized to generate informative and aesthetically pleasing plots, including volcano plots and heatmaps. They were critical in effectively communicating the results of our analysis.

Statsmodels: The multitest module from Statsmodels was employed for multiple hypothesis testing, ensuring the statistical validity of results when handling large datasets.

PrettyTable: PrettyTable allowed us to create well-structured and visually appealing tables for presenting data summaries and test results in an organized format.

Matplotlib-Venn: This library was used to create Venn diagrams, visually illustrating the overlap between sets of differentially expressed genes identified under paired and independent sample analyses.

GSEApy: GSEApy, a Python library for Gene Set Enrichment Analysis (GSEA), was crucial in performing GSEA on the identified differentially expressed genes. Its built-in plotting capabilities enhanced the clarity of enrichment results.

CSV Module: Python's CSV module was utilized for reading and writing CSV files, ensuring seamless data input and output handling.

Member Contribution:

1. **Mina Adel:** Data Pre-processing, Normality test and volcano plot

2. **Mariem Magdy:** Data Pre-processing, Normality test and volcano plot

3. **Hesham Tamer:** Hypothesis testing, Fold change and GSEA

4. **Reem Adel:** Hypothesis testing, Fold change and GSEA