Comprehensive Analysis of Gene Expression and Pathway Enrichment in Lung Squamous Cell Carcinoma (LUSC): Identifying Key Biomarkers and Therapeutic Targets

Omar Abd El-Nasser
Systems and Biomedical Engineering
Cairo University
omar.mohammed021@eng-st.cu.edu.eg

Mahmoud Magdy
Systems and Biomedical Engineering
Cairo University
mahmoud.ismail02@eng-st.cu.edu.eg

Mahmoud Mohamed
Systems and Biomedical Engineering
Cairo University
mahmoud.aty02@eng-st.cu.edu.eg

Muhammad Ibrahim
Systems and Biomedical Engineering
Cairo University
mohamed.gawad02@eng-st.cu.edu.eg

Abstract— **This study analyzes gene expression (GE) data in Lung Squamous Cell Carcinoma (LUSC) to identify differentially expressed genes (DEGs) and explore their biological significance through Gene Set Enrichment Analysis (GSEA). Using paired and independent GE datasets, preprocessing steps ensured data quality, followed by statistical testing with the Wilcoxon Signed-Rank Test and Mann-Whitney U Test, along with fold change analysis. Key DEGs were visualized in volcano plots, and GSEA linked these genes to relevant pathways, offering insights into LUSC mechanisms. The findings highlight the impact of data treatment on DEG identification.**

Keywords: Lung Squamous Cell Carcinoma, Gene Expression, Differentially Expressed Genes, Statistical Testing, Paired Samples, Independent Samples, Wilcoxon Signed-Rank Test, Fold Change, Volcano Plot, and Gene Set Enrichment Analysis.

I. INTRODUCTION

Lung Squamous Cell Carcinoma (LUSC) is a major subtype of non-small cell lung cancer, characterized by malignant growth in the squamous epithelial cells of the lungs. Despite advances in cancer treatment, LUSC continues to have a high mortality rate, partly due to its complex molecular landscape and limited availability of targeted therapies. A deeper understanding of the genetic and molecular alterations in LUSC is essential for identifying novel biomarkers and potential therapeutic targets.

Gene expression (GE) analysis offers a valuable approach to understanding these molecular changes by identifying differentially expressed genes (DEGs) between cancerous and healthy tissues. Such studies not only provide insights into the biological mechanisms underlying LUSC but also enable pathway analysis to uncover critical processes driving the disease. This project focuses on utilizing both paired and independent GE data to identify DEGs, visualize their significance, and link them to relevant biological pathways using Gene Set Enrichment Analysis (GSEA).

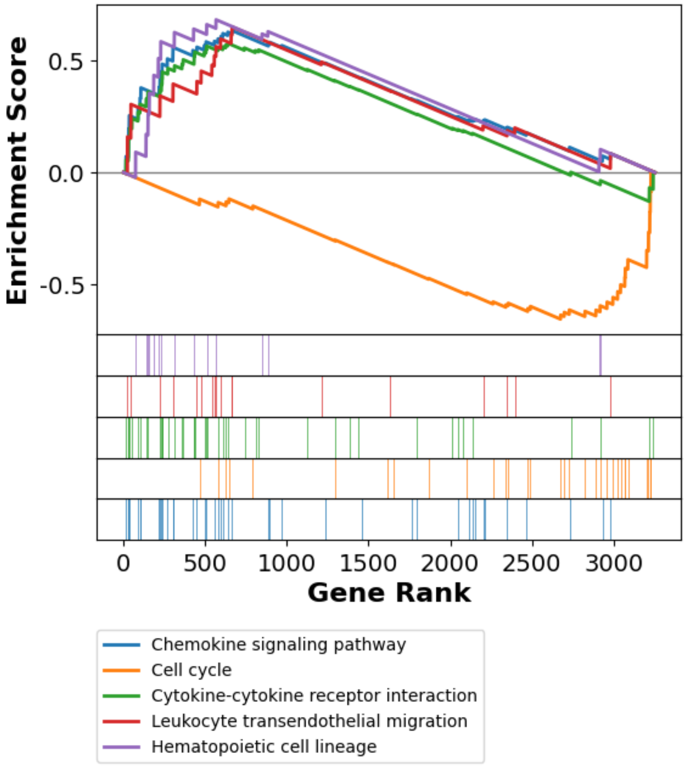


Fig.1. This plot is the combined output of Gene Set Enrichment Analysis for top.5 genes in KEGG_2021_Human

The objectives of this study are fourfold: (1) to identify DEGs using statistical hypothesis testing and fold change analysis; (2) to compare the DEGs identified in paired versus independent sample scenarios; (3) to visualize significant DEGs using volcano plots, integrating fold change and statistical significance; and (4) to perform GSEA to link DEGs to relevant biological pathways and processes.

By addressing these objectives, this work aims to deepen the understanding of the molecular landscape of LUSC and contribute to the identification of potential biomarkers and therapeutic targets.

II. DATASET DESCRIPTION

The dataset used in this study is part of The Cancer Genome Atlas (TCGA) Lung Squamous Cell Carcinoma (LUSC) data [1]. TCGA is a large-scale research initiative jointly led by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). It was established to provide comprehensive genomic and molecular characterizations of various cancer types, with the ultimate goal of advancing cancer diagnosis, treatment, and prevention.

The gene expression data for LUSC comprises two tab-separated text files:

- Tumor Samples (lusc-rsem-fpkm-tcga-t_paired.txt): This file contains GE data for cancerous tissues and includes expression levels of 19,647 genes across 51 subjects.
- Healthy Samples (lusc-rsem-fpkm-tcga_paired.txt): This file contains GE data for matched healthy tissues from the same 51 subjects, with the same set of 19,647 genes.

These paired files facilitate direct comparison between tumor and healthy tissues, minimizing inter-patient variability and ensuring robust statistical analysis. The dataset enables the identification of differentially expressed genes (DEGs) associated with LUSC progression.

III. LITERATURE REVIEW

Lung Squamous Cell Carcinoma (LUSC) is a major subtype of non-small cell lung cancer, characterized by poor prognosis and high metastasis [2]. Gene expression profiling plays a key role in understanding the molecular mechanisms behind LUSC, helping identify potential biomarkers and therapeutic targets. The TCGA LUSC dataset, containing gene expression data from tumor and healthy tissues, has been widely used in cancer research to identify differentially expressed genes (DEGs) and relevant molecular pathways [3].

Previous studies have utilized DEG analysis to uncover genes associated with LUSC progression and prognosis [4]. Pathway analysis, such as Gene Set Enrichment Analysis (GSEA), has been applied to interpret DEG findings by linking them to biological processes, enhancing our understanding of the disease. GSEA has revealed important cancer-related pathways, such as immune response pathways, that may offer insights for therapeutic development [5].

While significant progress has been made in analyzing LUSC through DEG and GSEA, challenges remain, including variability in findings due to differences in data preprocessing and statistical approaches. Integrating multiple types of genomic data, along with gene expression data, may provide more comprehensive insights into LUSC's molecular landscape [6].

IV. THEORETICAL BACKGROUND

A. Gene Expression and Its Regulation

Gene expression refers to the process of producing proteins from genetic instructions. It is tightly regulated, with factors that control when, where, and how much a gene is expressed, influencing cellular functions and behaviors.

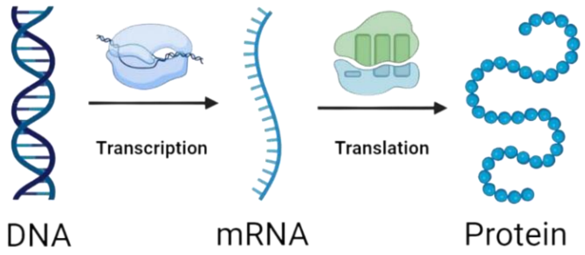


Fig.2. Gene Expression Process.

B. Lung Squamous Cell Carcinoma (LUSC)

LUSC is a subtype of non-small cell lung cancer, often associated with poor prognosis and high metastasis. It originates in the squamous cells lining the lungs and involves complex genetic changes that influence tumor progression.

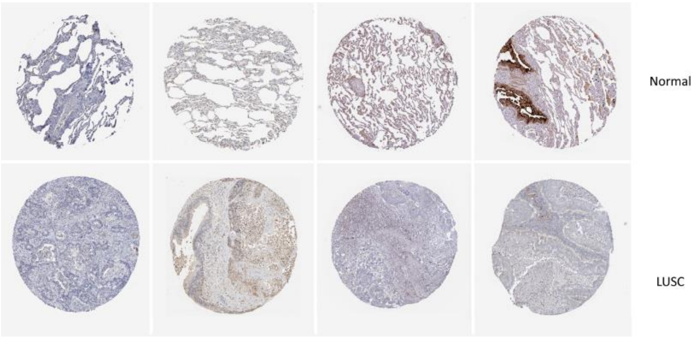


Fig.3. Comparison between LUSC and normal samples [9].

C. Preprocessing of Gene Expression Data

Data preprocessing involves several steps to ensure the reliability and quality of gene expression data. This includes filtering irrelevant genes, handling missing values, and normalizing data to remove technical biases before analysis.

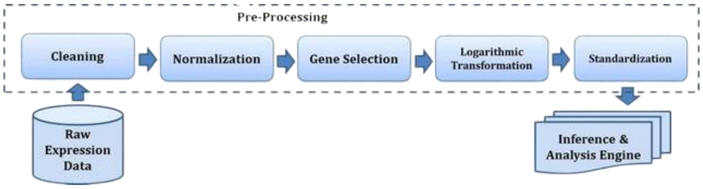


Fig.4. Commonly used preprocessing steps for gene expression data analysis [8].

D. Hypothesis Testing for Differentially Expressed Genes (DEGs)

Hypothesis testing is used to identify genes whose expression differ significantly between tumor and healthy tissue. Statistical tests, such as the Wilcoxon Signed-Rank Test and Mann-Whitney U Test, are commonly applied to detect these differences:

- *Wilcoxon Signed-Rank Test:* A non-parametric test used to compare paired samples, such as matched tumor and healthy tissues from the same patient, to assess differences in mean ranks. It is ideal for non-normal data and evaluates whether the median rank differences significantly deviate from zero.

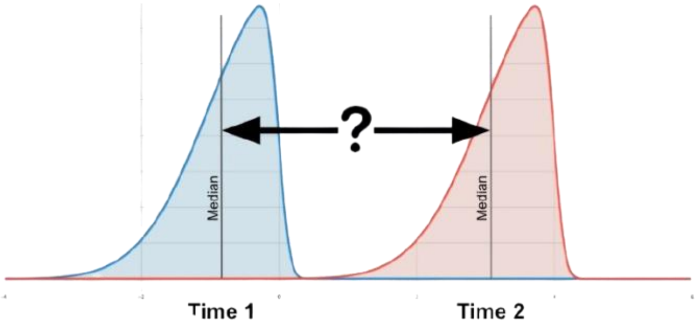


Fig.5. Wilcoxon Signed Rank Test

- *Mann-Whitney U Test:* A non-parametric test for comparing independent samples, such as gene expression levels from distinct patient groups, to determine if their distributions differ. It ranks all observations and compares rank sums between groups, making it robust for non-normal data.

$$U_{stat} = Rank\ Sum - \frac{n\,(n-1)}{2} \tag{1}$$

E. Fold Change Analysis in Gene Expression

Fold change quantifies the magnitude of difference in gene expression levels between two conditions (e.g., tumor vs. healthy). It is calculated as the ratio of the mean expression levels in the two conditions. In practice, the equation is often expressed in logarithmic terms (base 2) to facilitate interpretation:

$$Fold\ Change\ (FC) = \frac{\mu\ _{Tumor\ Samples}}{\mu\ _{Healthy\ Samples}} \tag{2}$$

$$\log_2(Fold\ Change) = \log_2(\frac{\mu\ _{Tumor\ Samples}}{\mu\ _{Healthy\ Samples}}) \tag{3}$$

Where:

$\mu\ _{Tumor\ Samples}$ and $\mu\ _{Healthy\ Samples}$ are Mean Expression in Tumor Samples and Mean Expression in Healthy Samples, respectively.

F. Paired vs. Independent Samples

- Paired Samples: These are related or matched samples where each observation in one group has a corresponding observation in the other group. Example: Gene expression levels in healthy tissue paired with levels in tumor tissue from the same patient.
- Independent Samples: These are unrelated groups with no inherent pairing. Example: Gene expression levels from a group of patients with cancer compared to a separate group of healthy individuals.

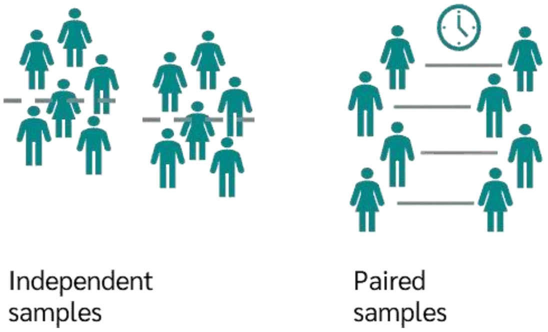


Fig.6. Paired vs. Independent Samples

G. Volcano Plot Visualization for Differential Expression

Volcano plots are crucial graphical tools in differential gene expression analysis, displaying the log2-transformed fold change (x-axis) against the -log10-transformed p-value (y-axis). They highlight genes with significant expression changes, typically located in the upper left (down-regulated) or upper right (up-regulated) quadrants. Threshold lines for statistical significance and fold change aid in identifying key genes, making volcano plots an efficient method to visualize and interpret large datasets in gene expression studies.

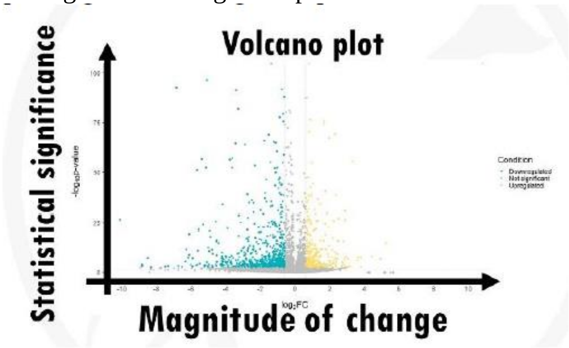


Fig.7. Volcano plot axis and graph components

H. Gene Set Enrichment Analysis (GSEA) and Its Applications

Gene Set Enrichment Analysis (GSEA) evaluates whether predefined gene sets, often representing biological pathways, are significantly enriched among differentially expressed genes. By ranking genes based on expression changes, GSEA identifies overrepresented pathways, offering insights into molecular mechanisms and potential therapeutic targets in diseases such as cancer.

V. METHODS

This process includes multiple steps, starting with data cleaning to remove zeroes and null values. Then DEGs identification using a suitable statistical test. Lastly, use the set of DEGs obtained by the hypothesis that data are paired and perform Gene Set Enrichment Analysis (GSEA) on this set of genes.

A. Data Preprocessing and Cleaning

1. Data Loading: Gene expression (GE) data for healthy and tumor tissues were imported into pandas DataFrames that have **19648** gene, with gene names designated as indices to facilitate data manipulation.

2. Gene Filtering: Genes exhibiting more than 50% zero expression values across samples were excluded to focus on genes with sufficient expression levels, so, around **5500** gene expression were removed. This step was critical for reducing noise and enhancing the reliability of subsequent analyses.

3. Handling Missing Values: Any residual missing values were removed to ensure a complete dataset for accurate hypothesis testing and fold change calculations.

4. Normalization: The dataset was normalized to account for sequencing depth and other technical variations, enabling meaningful comparisons between samples.

B. *DEGs Identification and Hypothesis Testing*

This process started by choosing a suitable statistical test to determine the genes whose expression difference was significant between healthy and cancerous cells.

1. Wilcoxon Signed-Rank Test for Paired Samples:

Is a non-parametric statistical method designed to compare two related samples or repeated measurements to assess whether their population mean ranks differ. Applicability in Paired Samples:

- Paired Nature: In gene expression analysis, paired samples refer to matched healthy and tumor tissues from the same patient. This pairing controls inter-patient variability, ensuring that differences arise from the biological condition rather than individual variability.
- Non-Parametric Approach: The test does not rely on the assumption of a normal distribution for the differences between paired observations, making it particularly suitable for gene expression data, which often deviate from normality.
- Sensitivity to Rank Changes: By focusing on the ranks of differences rather than raw values, the test is robust for small sample sizes and effectively detects subtle changes in paired data.

2. Mann-Whitney U Test for Independent Samples:

The Mann-Whitney U Test, also known as the Ranksum Test, is a non-parametric statistical method used to compare the distributions of two independent samples to assess if they differ significantly. Applicability in Independent Samples:

- Independence: This test is ideal for scenarios where the two groups being compared are independent, with no inherent pairing between the samples. In gene expression studies, this corresponds to comparing expression levels between distinct groups of patients, such as healthy individuals and those with specific conditions.
- Non-Parametric Approach: The test does not require the data to follow a normal distribution, making it a robust choice for gene expression data, which frequently exhibits non-normality.
- Rank-Based Comparison: Instead of analyzing raw data, the test ranks all observations and compares the rank sums between the two groups. This approach effectively handles outliers and ensures the test's robustness against deviations in data distribution.

P-value correction for multiple hypotheses was done using the False Discovery Rate (FDR) method. The top 5 DEGs for LUSC were identified by ranking the produced DEGs by tests. After ranking the DEGs according to the Wilcoxon test statistics, the top five DEGs are as follows:

Gene_Name	**P_Value**
HIST3H2A	2.63E-11
LIN7B	0.238495982
LXN	2.31E-07
SCML1	0.015312048
GSDMD	1.87E-07

Table 1. Sample of our results from DEGs with p-values file.

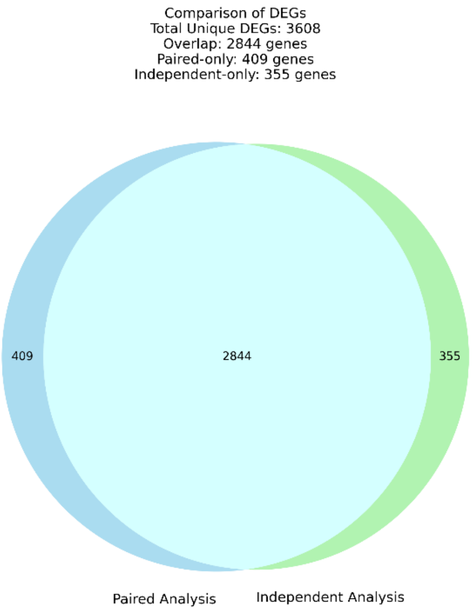


Fig.8. Venn diagram for DEGs in Paired Analysis vs. Independent Analysis.

C. *Fold Change*

Fold change is a crucial metric used to measure the relative difference in gene expression between two conditions, such as tumor and healthy tissues. In our analysis, fold change quantifies how much gene expression varies between these two conditions. To enhance interpretability, the fold change was transformed using the log2 scale, which facilitates the identification of both upregulated and downregulated genes. A small constant (*epsilon, ε*) was added to both tumor and healthy expression values prior to log transformation to prevent undefined or infinite values.

The log2 fold change was then used to rank differentially expressed genes (DEGs) based on the magnitude of their expression changes. This ranking allows us to prioritize genes that exhibit the most significant alterations in gene expression. Combined with statistical tests, such as the Wilcoxon Signed-Rank Test and the Mann-Whitney U Test, fold change serves as a key criterion for identifying DEGs with both statistical significance and biological relevance. This approach provides a robust method for uncovering genes that are critical in cancer progression.

Gene_Name	**Fold Change**	**log2FC**	**abslog2FC**
KRT14	0.003009	-8.37629	8.376287
KRT6A	0.003178	-8.29748	8.297483
KRT16	0.003287	-8.24895	8.248954
KRT5	0.008006	-6.96465	6.964646
KRT13	0.010241	-6.60948	6.609484

Table 2. Top 5 Genes ranked by Fold Change (FC).

D. *Volcano Plot*

The plot consists of two key components: (1) the x-axis, which displays the fold change magnitude, typically presented as log2 fold change, and (2) the y-axis, which represents statistical significance through the negative logarithm (base 10) of p-values, allowing for better visualization of small p-values. The interpretation of volcano plots centers on two primary categories of genes: significant DEGs, which appear in the upper left and right corners of the plot, meeting both statistical significance and fold change thresholds, and non-significant DEGs, which cluster around the plot's center, failing to meet these criteria.

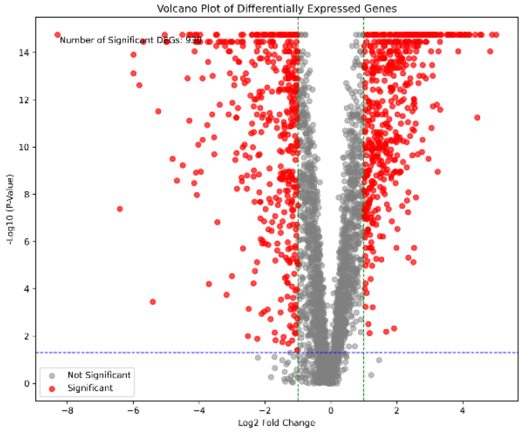


Fig.9. Volcano Plot for DEGs with p-values and ranked $log_2\,FC$, with p-value threshold = 0.05 and $log_2\,FC$ threshold = 1. With 930 Significant DEGs.

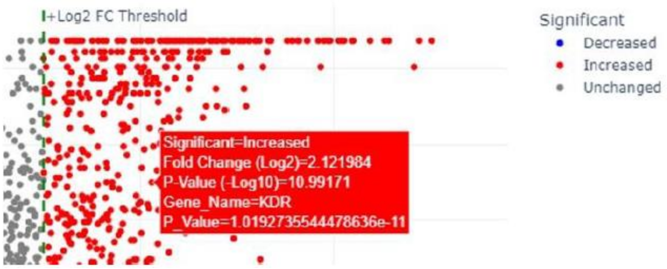
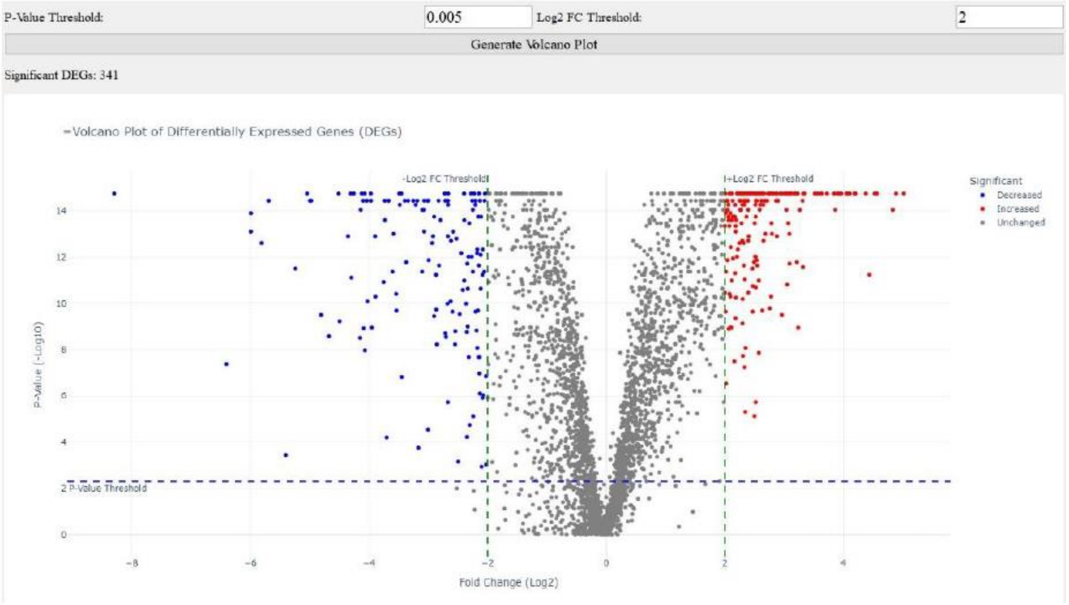


Fig.10. Volcano Plot from PyQT5 application visualized by plotly, with interactive features that display each gene name, p-value, log (FC), and significant level.


P-Value Threshold: 0.005 Log2 FC Threshold: 2

Generate Volcano Plot

Significant DEGs: 341

Fig.11. Volcano Plot from PyQT5 application with real-time thresholds change, with p-value threshold = 0.005 and log (FC) threshold = 2, and the results are 341 Significant DEG.

E. Gene Set Enrichment Analysis (GSEA)

We explore many Gene Sets and document a comparison between them to identify the relevant of them with LUSC dataset in Table.3

Gene Set ID	Relevant Reason
HATADA METHYLATED IN LUNG CANCER UP	Hyper-methylation of tumor suppressor genes is a key feature in LUSC, influencing gene expression and silencing genes involved in growth regulation and apoptosis. This mechanism is crucial for understanding LUSC biology and identifying potential therapeutic targets.
ZHONG SECRETOME OF LUNG CANCER AND ENDOTHELIUM	The interactions between cancer cells and endothelial cells in the tumor microenvironment play a significant role in LUSC progression. The secretome, which includes proteins involved in angiogenesis, inflammation, and epithelial-to-mesenchymal transition (EMT), is critical in LUSC pathogenesis.
Cancer Cell Line Encyclopedia *	While the Cancer Cell Line Encyclopedia (CCLE) isn't a pre-defined gene set for GSEA, its data can be used to create custom gene sets or map differentially expressed genes (DEGs) to relevant pathways, providing biological relevance to LUSC.
KEGG 2021 Human *	LUSC gene expression data provides insights into gene expression, while the KEGG_2021_Human gene sets represent biological pathways. GSEA can identify enriched KEGG pathways in LUSC, revealing key molecular mechanisms and potential therapeutic targets.

Table 3. Comparison between Gene Sets and their relevance to LUSC dataset

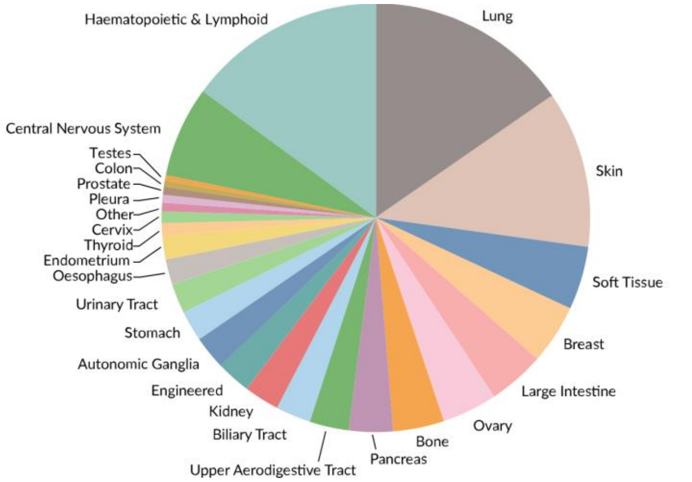


Fig.12. Overview of the Cancer Cell Line Encyclopedia (CCLE) Project and Its Contributions to Cancer Research.

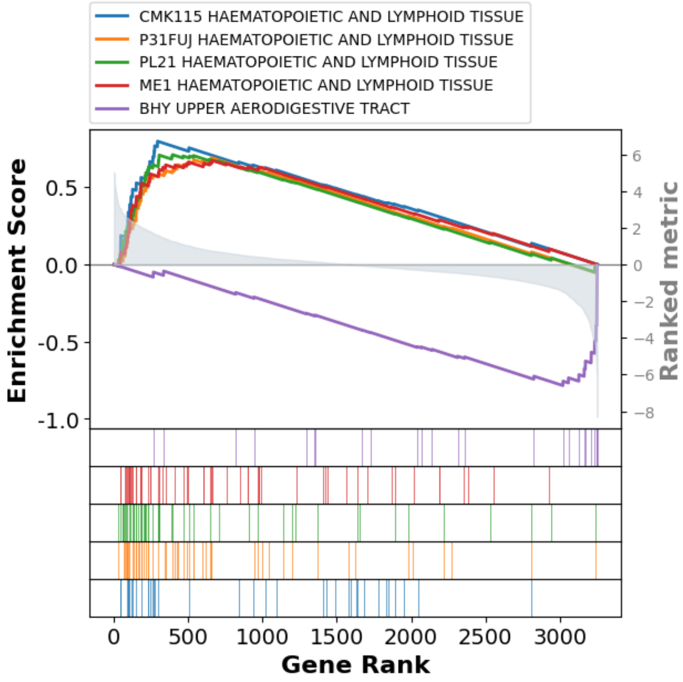


Fig.13. Display of the Top 6 in the 'Cancer Cell Line Encyclopedia' gene set.

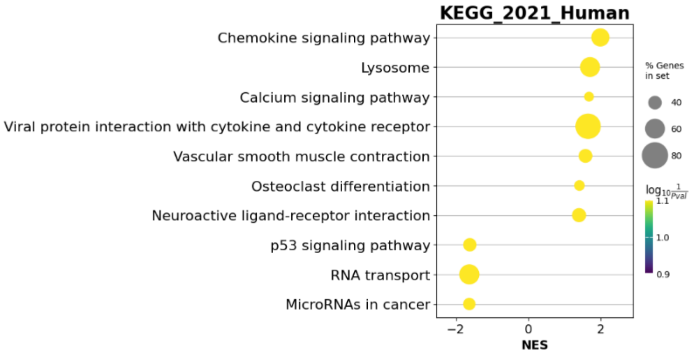


Fig.14. Dot plot for the top 10 Genes in 'KEGG 2021 HUMAN' gene set.

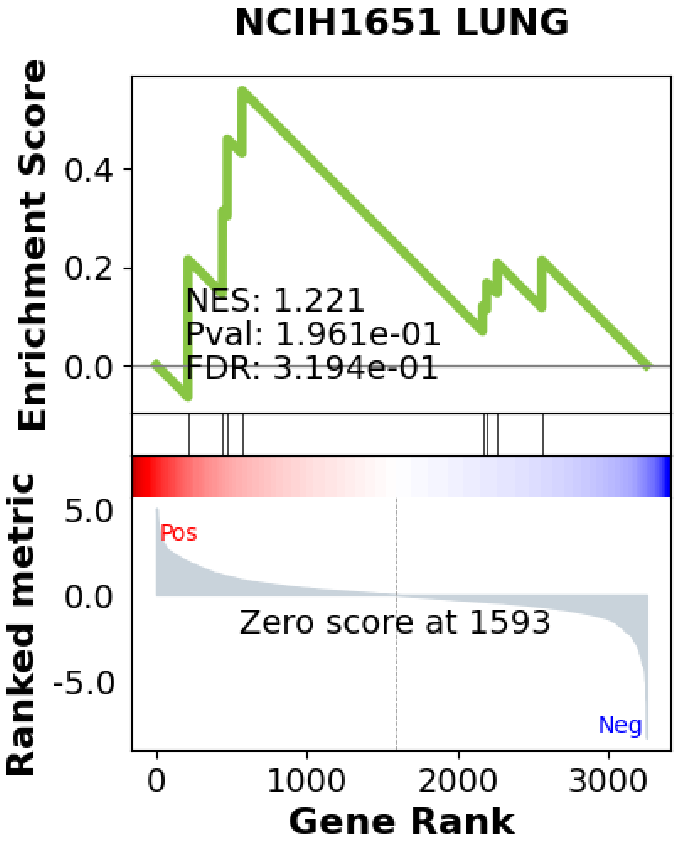


Fig.15. Enrichment and ranked matrix for "NCIH1651 LUNG" term, with 'Cancer Cell Line Encyclopedia', gene set.

F. *Software Packages and Frameworks*

- *PyQt5*: This GUI framework was utilized to develop an interactive application for real-time visualization of Volcano plots, enabling dynamic adjustments of thresholds and immediate exploration of DEGs.
- *Pandas:* A robust data manipulation library, Pandas streamlined the organization, cleaning, and transformation of gene expression datasets, ensuring structured and efficient analysis.
- *NumPy:* This foundational library facilitated efficient array operations and numerical computations, providing the mathematical backbone for handling large-scale molecular datasets.
- *Matplotlib and Seaborn:* These plotting libraries were pivotal in generating clear, publication-quality visualizations, such as volcano plots, ensuring interpretability and effective presentation of analytical results.
- *SciPy:* The advanced scientific functions and algorithms of SciPy were leveraged for conducting statistical analyses, including the Wilcoxon Signed-Rank Test and Mann-Whitney U Test, essential to DEG identification.
- *Plotly:* This interactive visualization library enhanced data exploration through dynamic plotting, allowing for real-time updates to visualization parameters and deeper insights into dataset patterns.
- *matplotlib_venn:* This extension to Matplotlib enabled the creation of Venn diagrams to visually compare gene overlaps between datasets, aiding in the comparative analysis of DEGs.

- *gseapy:* This Python library streamlined Gene Set Enrichment Analysis (GSEA), automating pathway enrichment calculations and linking identified DEGs to relevant biological pathways, providing insights into their functional implications.

VI. RESULTS AND DISCUSSION

A. DEGs Identification:

- For both paired and independent samples, hypothesis testing provided a set of DEGs.
- Differences were observed in the sets of DEGs due to the statistical nature of paired vs. independent tests.

B. Fold Change & Volcano Plot:

- A volcano plot was generated, displaying DEGs derived from the paired samples hypothesis, integrating statistical significance with fold change.

C. GSEA:

- GSEA highlighted key pathways enriched in the DEGs identified from paired samples, providing insights into potential biological mechanisms at play.

The study effectively identifies differentially expressed genes in LUSC, leveraging both statistical tests and fold change analysis. Differences observed between the paired and independent sample hypotheses underline the impact of sample treatment on results. The volcano plot provided a clear visual representation, while GSEA furthered our understanding by integrating DEGs with biological pathways.

VII. CONCLUSION

This analysis of LUSC GE data enables the identification of key genes and pathways potentially involved in disease pathogenesis. The integration of hypothesis testing, fold change analysis, and GSEA offers a comprehensive approach to understanding genetic alterations in LUSC, paving the way for future research into targeted therapies and diagnostics.

VIII. FUTURE WORK

Further work could expand upon these findings by:

- Validating results with additional datasets or experimental methodologies.
- Exploring the impact of additional covariates.
- Investigating the biological roles of identified pathways in the context of LUSC.

IX. ACKNOWLEDGMENTS

This work utilizes publicly available LUSC gene expression datasets. We thank the researchers and institutions responsible for making this data accessible.

X. REFERENCES

[1] *TCGA-LUSC (no date) GDC. Available at: https://portal.gdc.cancer.gov/projects/TCGA-LUSC (Accessed: 23 September 2023).*

[2] *Cancer Genome Atlas Research Network. (2012). Comprehensive molecular profiling of lung squamous cell carcinoma. Nature, 489(7417), 519-525.*

[3] *Choi, Y. L., et al. (2015). Identification of novel biomarkers for lung squamous cell carcinoma through gene expression profiling. Journal of Clinical Oncology, 33(8), 907-916.*

[4] *Ding, L., et al. (2017). Differential expression analysis of RNA-Seq data in LUSC and its potential implications in cancer progression. Cancer Letters, 403, 16-23.*

[5] *Liu, Q., et al. (2018). Pathway analysis of LUSC based on GSEA and its potential role in the immune response. Frontiers in Genetics, 9, 674.*

[6] *Zhai, X., et al. (2020). Identification of key genes and pathways in lung squamous cell carcinoma using TCGA data. Journal of Cancer Research and Clinical Oncology, 146(7), 1785-1794.*

[7] *Jong, Simone & Van Eijk, Kristel & Zeegers, Dave & Strengman, Eric & Janson, Esther & Veldink, Jan & Berg, Leonard & Cahn, Wiepke & Kahn, René & Boks, Marco & Ophoff, Roel. (2012). Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes. European journal of human genetics : EJHG. 20. 1004-8. 10.1038/ejhg.2012.38*

[8] *S. Roy, P. Sharma, K. Nath, J. K. Kalita, and others, "Pre-Processing: A Data Preparation Step," Reference Module in Life Sciences, Jan. 2018, doi: 10.1016/B978-0-12-809633-8.20457-3.*

[9] *S. F. Kadasah, "Prognostic Significance of Glycolysis-Related Genes in Lung Squamous Cell Carcinoma," Department of Biology, Faculty of Science, University of Bisha, P.O. Box 551, Bisha 61922, Saudi Arabia*