Dean&Francis
ISSN 2959-409X

Comprehensive Bioinformatics Analysis of Differential Gene Expression Differences between Lung Adenocarcinoma and Lung Squamous Cell Carcinoma.

Duoduo Qian

Abstract:

Lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) are the two most prevalent histological subtypes of non-small cell lung cancer (NSCLC), each of which exhibits distinct molecular characteristics and requires subtype-specific treatment strategies. This study aimed to comprehensively analyze the differential gene expression profiles between LUAD and LUSC to identify subtype-specific biomarkers and pathogenic pathways. Using RNA-seq data from The Cancer Genome Atlas (TCGA), this study performed differential gene expression analysis with DESeq2, disease-free survival analysis via GEPIA, and pathway enrichment analysis using Enrichr with the KEGG 2021 Human database. A total of 29,052 and 28,866 significant differentially expressed genes (DEGs) were identified in LUAD and LUSC respectively. Key up-regulated genes in LUAD, including *FAM83A* and *FAM83A-AS1*, were significantly associated with poor disease-free survival (HR = 1.5, p = 0.0048; HR = 1.6, p = 0.0024), while down-regulated *PECAM1* showed protective effects (HR = 0.71, p = 0.03). Pathway analysis revealed significant dysregulation of the cell cycle pathway in both subtypes, with LUAD showing stronger association. The DNA replication pathway was notably prominent in LUSC, while a unique malaria pathway was identified in LUAD. The research findings underscore the molecular heterogeneity between LUAD and LUSC, highlighting potential prognostic biomarkers like *FAM83A* and subtype-specific pathogenic pathways, which could inform the development of precise therapeutic interventions between LUAD and LUSC.

Keywords: TCGA, non-small-cell lung cancer (NSCLC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), bioinformatics, differential gene expression, disease-free survival (DFS), pathway enrichment, KEGG, GEPIA

ISSN 2959-409X

1 Introduction

Lung cancer is one of the leading causes of cancer-related death worldwide, with approximately 1.8 million deaths each year. It is a serious health problem that can lead to injury and death. Symptoms of lung cancer include a persistent cough, chest pain and shortness of breath. [1]

As with other cancers, the fundamental abnormality that leads to the development of lung cancer is the continuous uncontrolled proliferation of cancer cells, forming a tumor. [2] NSCLC is the most common type of lung cancer, constituting approximately 80% to 85% of all lung cancer cases worldwide. It is primarily categorized into two histological subtypes: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Lung adenocarcinoma (LUAD) is the most prevalent NSCLC subtype, accounting for about 40% of all lung cancers. [3]

Lung cancer is often diagnosed at advanced stages when treatment options are limited. Therefore, despite the advances in diagnosis and treatment for lung cancer, the prognosis for patients with NSCLC remains suboptimal, with a relatively low 5-year survival rate of approximately 26.7%. [4] This highlights the need for a deep understanding of the molecular differences between LUAD and LUSC in order to facilitate the development of more effective subtype-specific treatments against lung cancer.

Differential gene expression (DGE) analysis is a key tool used to elucidate the molecular basis of cancer. A gene is a section of DNA that codes for a specific protein that is essential for body functions. Cancer may begin when certain genes malfunction, leading to uncontrolled cell growth. [5] Gene expression is the process where genes are activated to produce RNA or proteins [6], whereas differential gene expression refers to variations in gene activity between different conditions. Studying the differentially expressed genes (DEGs) between the two NSCLC subtypes can help to identify and understand the differences in the molecular basis of cancers.

Despite the advancements in cancer research, there is a need for a more comprehensive understanding of the differential gene expression profiles between NSCLC subtypes. Therefore, this research will delve into the field of cancer genomics and bioinformatics, aiming to conduct a comprehensive bioinformatics analysis to delineate different patterns of DEGs between LUAD and LUSC. This will be achieved by utilizing The Cancer Genome Atlas (TCGA) and KEGG (Kyoto Encyclopedia of Genes and Genomes) 2021 Human database for data acquisition, and by performing differential gene expression analysis to identify DEGs, conducting survival analysis to assess the prognostic significance of DEGs, and carrying out pathway enrichment analysis to find key pathway differences between LUAD and LUSC.

2 Literature Review

2.1 Genomics: Role of Genes in Cancer Development

Cancer genomics, which is defined as the study of the complete sequence of DNA and its expression in tumor cells. [7]

2.1.1 Relationship of gene and cancer

"Cancer is a genetic disease—that is, it is caused by changes to genes that control the way our cells function, especially how they grow and divide." [4] As mentioned, genetic changes are the key reason for the occurrence of cancer. These changes may arise from environmental exposures or spontaneous errors during DNA replication. [8] Gene mutation is one of the typical causes of cancer. Several studies, specifically Vogelstein, B. et al. (2013) have shown that mutation in genes such as *TP53* and *KRAS* leads to uncontrolled proliferation and survival of cancer cells. On the other hand, studies have shown that environmental stimuli can cause epigenetic changes, which alters gene expression levels without changing the underlying DNA sequence through regulating DNA methylation or histone modification. [9, 10] Research in fields of epigenetics shows that non-mutational changes can also be a potential cause of cancer. This statement could be further evidenced by Bradly, D.P. et al. (2012) and Li, Y. et al. (2020), proving that hypermethylation of the *CDKN2A* gene can silence the expression of the p16 tumor suppressor gene, leading to formation of tumor in lung cancer. All of the above studies have shown how genetic changes could contribute to the occurrence of cancer, but also indicate that the relationship of genes and cancer is very complex, involving intricate interactions between gene mutations, epigenetic modifications, environmental factors, which collectively contribute to cancer development and progression.

2.1.2 Cancer related genes

Different types of genes play a critical role in regulating cellular processes such as cell division, apoptosis, and DNA repair. [11] Oncogenes, tumor suppressor genes, and DNA repair genes are 3 main types of genes that may lead to cancer development when they are mutated or abnormally expressed.

2.1.3 Genomic differences between LUAD and LUSC

Genomic differences refer to variations in the genetic material like DNA sequences, copy number alterations or chromosomal rearrangements. For instance, mutations

in oncogenes, tumor suppressor genes, and DNA repair genes are frequently observed in lung cancer but with distinct patterns between LUAD and LUSC. According to The Cancer Genome Atlas Research Network (2014), LUAD is frequently characterized by mutations in EGFR, KRAS, and rearrangements in ALK gene that drive oncogenic signaling pathways. In contrast, another study has shown that LUSC frequently shows alterations in TP53, CDKN2A, and PIKCA, along with amplifications in SOX2 and FGFR1. These findings present the genomic difference between the different variations in genetic materials of LUAD and LUSC (The Cancer Genome Atlas Research Network, 2012).

2.2 Bioinformatics: Bioinformatics Approaches in NSCLC Research

Bioinformatics is an interdisciplinary science that combines biology, computer science, and statistics to collect, store and analyze large-scale biological data using high-throughput technologies. Researchers are able to study systematically in the fields of genomics, transcriptomics, and proteomics. [12, 13]

2.2.1 Introduction to The Cancer Genome Atlas (TCGA) database

The Cancer Genome Atlas (TCGA) database provides various types of biological data such as transcriptome profiling, simple nucleotide variation, copy number variation, DNA methylation, RNA sequencing, clinical and biospecimen data of about 84,392 cancer patients. [14] With no doubt, TCGA has produced rich data sets of immeasurable value and promoted substantial development in the bioinformatics field. Its program has molecularly characterized over 20,000 primary cancer and matched to normal samples from 33 cancer types, bringing together researchers from different disciplines and multiple institutions. [15, 13]

2.2.2 Application of bioinformatics in NSCLC research

Bioinformatics has escalated cancer research by making the integration and analysis of large-scale data sets more feasible. Many NSCLC studies are based on TCGA, GEO, and KEGG databases. These bioinformatics databases are important for discovering NSCLC subtype-specific molecular features, key driver mutations, gene expression patterns, pathway alterations and therapeutic targets. [16] At the same time, bioinformatics tools also play a significant role in NSCLC research. For instance, Dai, B., Ren, L., Han, X. and Liu, D. (2019) used DAVID for pathway enrichment analysis and the Geo2R tool in GEO database for DGE analysis. This study revealed new biomarkers related to diagnosis and prognosis in the pathogenesis of NSCLC, and identified potential therapeutic targets through bioinformatics analysis without the use of clinical trials. These applications highlight the transformative role of bioinformatics in improving cancer diagnosis and treatment, helping researchers to further understand the occurrence and development of NSCLC, translating these rich data sets into biological insights and clinical applications.

2.3 Differential Gene Expression (DGE) Analysis

2.3.1 Overview of differential gene expression analysis-its meaning and significance

Differential gene expression analysis is the statistical analysis of gene expression data between the test group (such as the tumor group) and the control group (usually the healthy group or normal group) under specific conditions. It works by filtering out the genes with significant expression changes. [17] Gene expression levels could be up-regulated where genes are over-expressed, or down-regulated where genes are under-expressed. DGE analysis reveals the subtype-specific gene expression patterns of NSCLC. It is important to identify DEGs of NSCLC and understand their role in the molecular mechanisms, resulting in advancement in diagnosis and treatment of NSCLC.

2.3.2 Criteria for significance in DGE analysis

There are specific criteria to define these “significant expression changes” in DGE analysis to distinguish between true biological differences and random events. The criteria are primarily based on statistical thresholds, which include the p-value, adjusted p-value, and Log2 fold change.

The p-value measures the likelihood that observed differences in gene expression occurred by chance. A p-value < 0.05 is commonly used as a cutoff of showing significance.[18]

The adjusted p-value (adjp) or the false discovery rate (FDR) corrects for multiple testing to reduce false positives by using statistic method. One common method is known as the x procedure, which adjusts p-values to account for the number of tests performed. [19] A false discovery rate (FDR) < 0.05 is often seen as showing significance.

Log2 fold change measures the magnitude of expression differences in two conditions. A $\log2FC > 0$ indicates that genes are up-regulated, while a $\log2FC < 0$ indicates that genes are down-regulated. [20] A common threshold for significance is $|\log2FC| > 1$. This threshold ensures that only genes with substantial changes are considered.

In order for the analysis to be both statistically and biologically meaningful, the expression level of a gene will also be considered. Genes with low expression may be filtered out to focus on biologically relevant changes.

ISSN 2959-409X

2.3.3 Introducing to the DESeq2 method

DESeq2 is a widely used R or Bioconductor package tool for DGE analysis for RNA-seq data. For instance, Zhai, Y. et al, (2021) conducted a study using DESeq2 for DGE analysis in LUAD and have identified 1654 differentially expressed immune genes (DEIGs) including 436 prognostic genes, verifying specific genes expression level. This study demonstrates the utility of DESeq2 in identifying DEGs and pathways critical to LUAD development.

2.3.4 Key differentially expressed genes (DEGs) in LUAD and LUSC

Differential gene expression (DGE) analysis was extremely useful in revealing the molecular distinctions between LUAD and LUSC. In LUAD, several studies have identified *EGFR* and *KRAS* as significantly upregulated genes, which drive tumor progression through activation of pathways like PI3K-AKT and MAPK signaling. [16] These genes are often mutated in LUAD, leading to uncontrolled proliferation and survival of cancer cells. Additionally, a transcription factor *TTF1*, is a hallmark of LUAD which further distinguishing it from LUSC. [21] In contrast, LUSC is characterized by the up-regulation of *TP63* and *SOX2*, which promote squamous differentiation and tumor growth. *TP63*, a member of the p53 family, is particularly critical in maintaining the squamous phenotype, while *SOX2* is involved in tumor aggressiveness. [22] Furthermore, *FGFR1* amplifications are more common in LUSC and are associated with poor prognosis, highlighting its potential as a therapeutic target. [23] Overall, DGE analysis not only elucidates the distinct molecular mechanisms underlying LUAD and LUSC but also guides the development of subtype-specific treatment strategies.

3 Methodology

This primary research aims to identify and analyze the differential gene expression patterns between lung adenocarcinoma and lung squamous cell carcinoma using bioinformatics tools including DESeq2, ggplot2, GEPIA, and Enrichr as well as publicly available databases including TCGA and KEGG. In order to accomplish the aim, the comprehensive bioinformatics analysis was divided into three steps: Differential Gene Expression (DGE) analysis, disease-free survival (DFS) analysis, and pathway enrichment analysis.

3.1 Differential Gene Expression (DGE) Analysis

This analysis was performed to achieve the first objective-- to identify key differentially expressed genes in LUAD and LUSC. TCGA database was selected in this study because it has sufficient records of LUAD and LUSC patients to provide data sample for analysis. To date, numerous studies on NSCLC have utilized the TCGA database, which demonstrates its credibility and usability. This widespread adoption further supports the selection of the TCGA database utilized for the DGE analysis conducted in this study. To identify DEGs for both LUAD and LUSC, RNA-seq data for LUAD and LUSC were obtained from The Cancer Genome Atlas (TCGA) database through the use of TCGAbiolinks R package. This requires programming in the environment in RStudio. The generated data include raw read counts for each gene in tumor and normal samples and were further processed by filtering out low-expression genes and normalize read counts by the negative binomial model using the DESeq2 package in R. A final document of a table with differentially expressed genes of LUAD and LUSC was saved to the chosen working folder when programming. To obtain a result table of DEGs arranged in ascending order of adjusted p-value, the document of DEGs table was modified manually using Numbers in Mac Os system. The results of the DGE analysis were further visualized by conducting volcano plots and PCA plots using the ggplot2 package in R. The ggplot2 package is a very useful drawing tool that can convert results of DGE analysis (table of differentially expressed genes with statistical numbers) into visualization results (such as Volcano plots and PCA plots). The DESeq2 package in R was the main tool to perform the differential expression analysis. DESeq2 tool was selected as the analytical tool due to its accessibility and enables quick learning and application in DEG analysis. The package is publicly available and the detailed tutorial is accessible on the Bioconductor website.

3.2 Disease-Free Survival (DFS) Analysis

The second step was to evaluate the prognostic significance of the identified DEGs in LUAD and LUSC patients, in order to determine the real significance of the DEGs. Disease-free survival (DFS) analysis was selected over overall survival (OS) analysis because it specifically measures cancer recurrence by excluding deaths from unrelated causes. It provides a clearer evaluation of the contribution of genetic factors to cancer development and progression. Single-gene disease-free survival analysis was performed using GEPIA (Gene Expression Profiling Interactive Analysis), an interactive web-based tool for survival analysis using TCGA data. [25] The GEPIA platform offers a highly efficient and user-friendly interface for generating single-gene disease-free survival Kaplan-Meier (KM) plots. It only requires simple operation by clicking: GoPIA > Survival > Survival Plots, followed by the selection of the specific gene to analyze and the

customization of cutoff values. This platform carries out statistical calibration and generates visualized survival plots without any additional manual operation. Eight significant DEGs identified on the volcano plots were selected to perform the disease-free survival analysis.

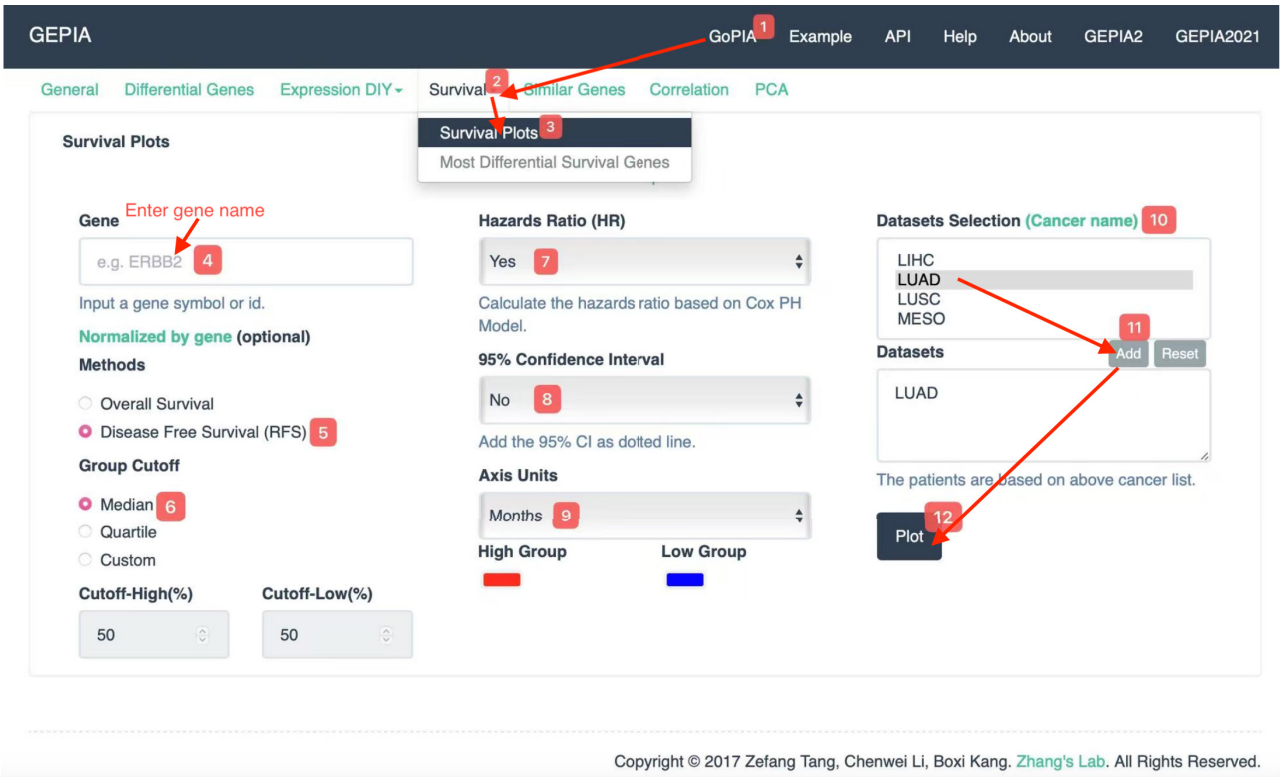

Figure 1 A screenshot showing the detailed procedure of using GEPIA

3.3 Pathway Enrichment Analysis

To obtain a conclusion, the final step was to identify key biological pathways and processes associated with the most significant DEGs in LUAD and LUSC. Pathway enrichment analysis was performed using Enrichr. This tool is free and easy to operate, no programming skills are required; users can simply input the gene list into the tool and it automatically analyze the genes through the database selected, producing a list of biological pathways linked to the gene set. The KEGG 2021 Human database was selected for pathway enrichment analysis. The database was widely recognized and credible in pathway enrichment analysis. It features recent updates with high-quality pathway maps, broad coverage of biological processes and diseases, and the ability to link genes to their functional roles in pathways, which is suitable for identifying key pathway differences of LUAD and LUSC using significant DEGs.

The top 1000 DEGs of LUAD and LUSC with ascending order of adjp were entered into the Enrichr gene list. However, the data table of DEGs was first processed, removing any non protein-coding genes. This is because the KEGG database is primarily designed for protein-coding genes. Including other gene types such as non-coding RNAs (IncRNA) may lead to incomplete or inaccurate pathway mapping. Enrichr performed over-representation analysis (ORA) to identify pathways significantly enriched in the DEGs. The top significant pathways were visualized by simple bar graph and table.

ISSN 2959-409X

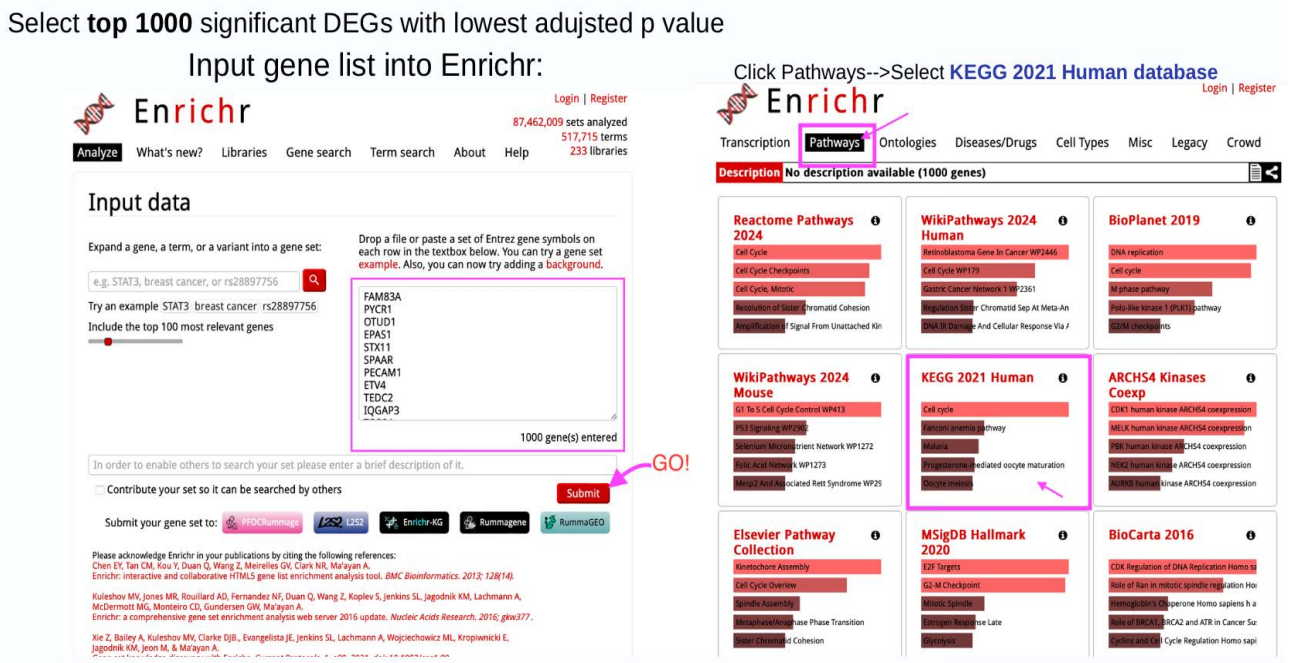

Figure 2 A screenshot showing the detailed procedure of using Enrichr

3.4 Data Availability

All data used in this research are credible and accessible for anyone to use for research purposes. The TCGA database provides publicly available data that can be accessed through the Genomic Data Commons (GDC) portal. The KEGG database is freely accessible at https://www.genome.jp/kegg/ and is open to the public for research purposes.

3.5 Software Package Versions

The analysis was performed in the RStudio environment (Version 2024.12.0+467) using the R base package (version 4.4.2), DESeq2 (version 1.46.0), and ggplot2 (version 3.5.1).

3.6 Ethical Approval

This research was carried out entirely on a computer, using computational methods and publicly available datasets. No animal or human tissue samples were used, and no experiments involving living organisms were performed. Therefore, ethical approval or patient consent was not required for this study.

4 Results

4.1 Differential Gene Expression Analysis

For LUAD, a total of 57628 DEGs were generated using R, including 29052 DEGs with an adjusted p-value (adjp) <0.05. The top20 most significant protein-coding DEGs with lowest adjp (in ascending order) include: *FAM83A, PYCR1, OTUD1, EPAS1, STX11, SPAAR, ETV4, TEDC2, IQGAP3,TOP2A, S1PR1, B3GNT3, ACVRL1, RGCC, EMP2, SEMA3G, SAPCD2, RTKN2, TEK,* and *PTPN21.* For LUSC, a total of 57489 DEGs were generated, including 28866 DEGs with an adjusted p-value<0.05. The top 20 most significant protein-coding DEGs with lowest adjp (in ascending order) include: *TPX2, KIF4A, TTK, CENPA, NEK2, HJURP, TROAP, KIF2C, CDCA5, UBE2C, CDC20, EXO1, KIF23, CCNB2, MYBL2, NCAPH, PLK1, BIRC5, BUB1B,* and *FAM83B.* Surprisingly, the adjp of the top10 genes generated were 0.

4.1.1 Volcano plots

These 20 DEGs displayed in the volcano plot were selected from the DEG table by simultaneously meeting the significance thresholds for log2 fold change (expression level) and adjusted p-value.

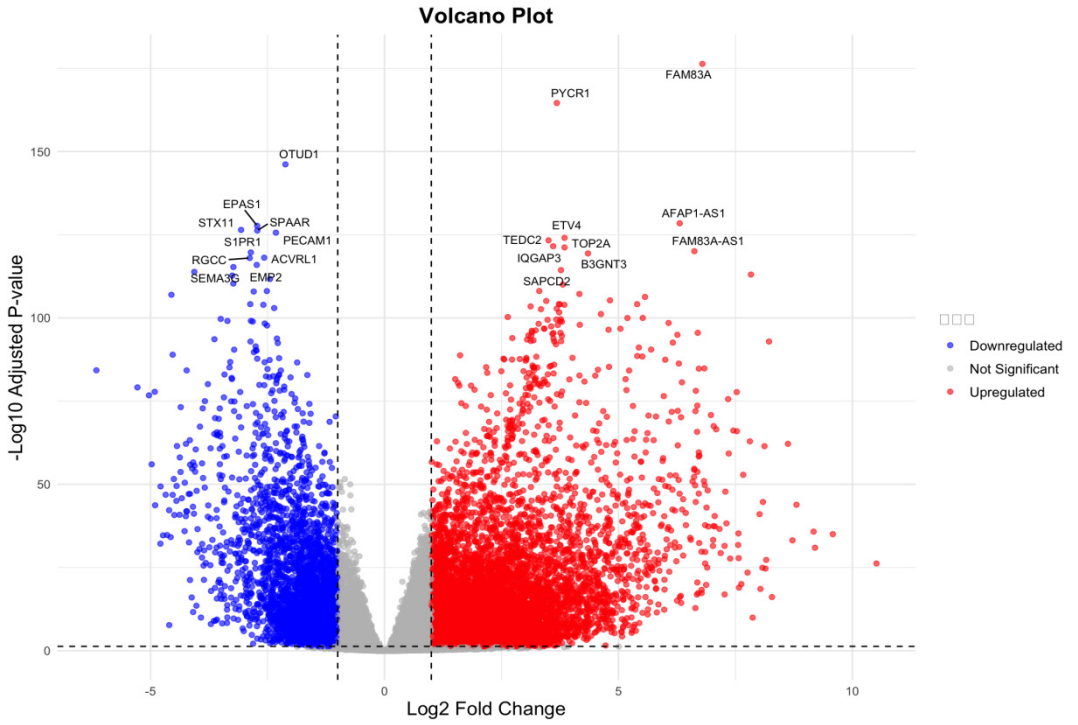

Figure 3 Volcano plot (LUAD)

Figure 3 displays 10 up-regulated and down-regulated DEGs of LUAD. Up-regulated DEGs include: *FAM83A, PYCR1, AFAP1-AS1, ETV4, TEDC2, TOP2A, FAM83A-AS1, IQGAP3, B3GNT3,* and *SAPCD2.* Down-regulated DEGs include: *OTUD1, EPAS1, SPAAR, STX11, S1PR1, PECAM1, RGCC, ACVRL1, SEMA3G,* and *EMP2.* The most significant up-regulated DEG is clearly identified as *FAM83A*, and the most significant down-regulated DEG is *OTUD1.*

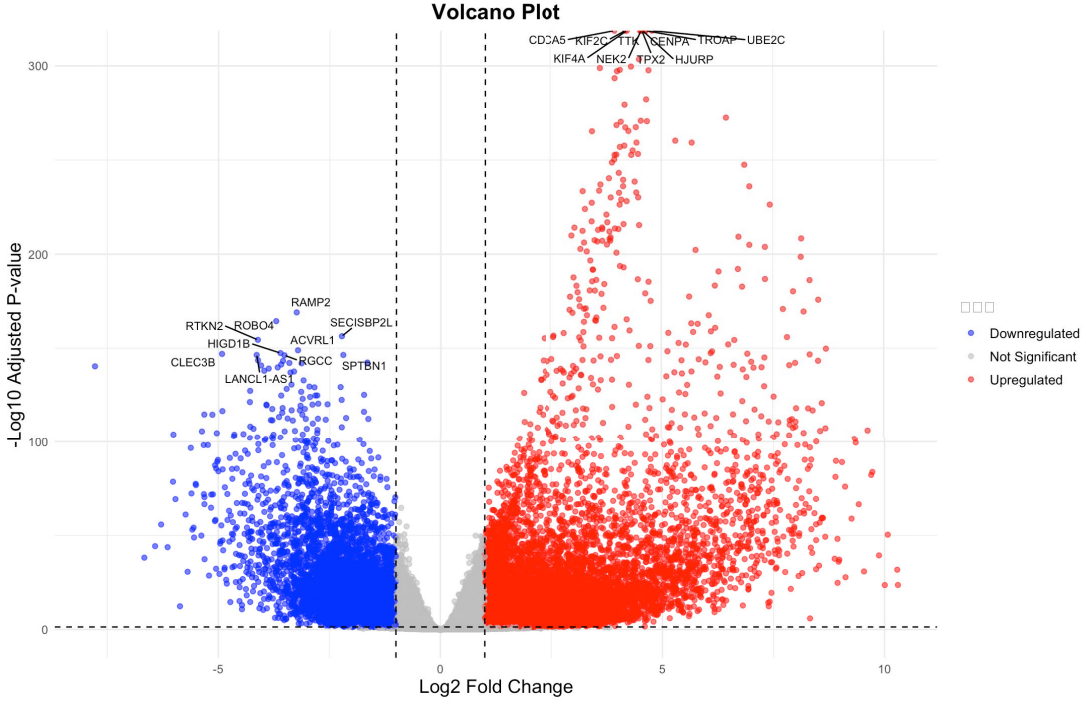

Figure 4 Volcano plot (LUSC)

ISSN 2959-409X

Figure 4 displays the top 10 up-regulated and down-regulated DEGs of LUSC. Up-regulated DEGs include: CDCA5, *KIF2C, TTK, CENPA, TROAP, UBE2C, KIF4A,NEK2, TPX2,* and *HJURP.* Down-regulated DEGs include: *RAMP2, ROBO4, RTKN2, HIGD1B, CLEC3B, LANCL1-AS1, RGCC, SPTBN1, ACVRL1* and *SECISBP2L*. The top 10 up-regulated DEGs show minimal variation due to their extremely low adjusted p-values. The most significant down-regulated DEG identified is *RAMP2*.

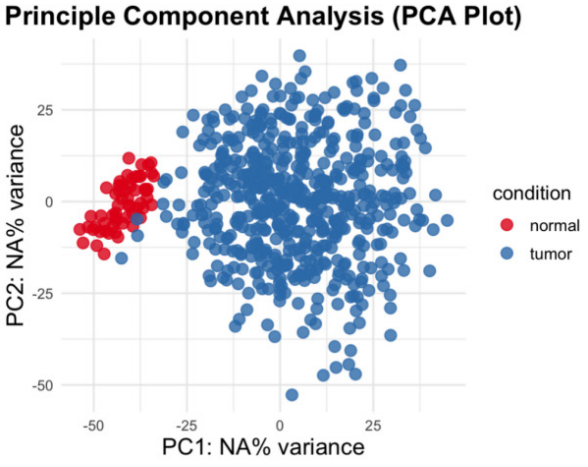

Figure 5 PCA Plot (LUAD)

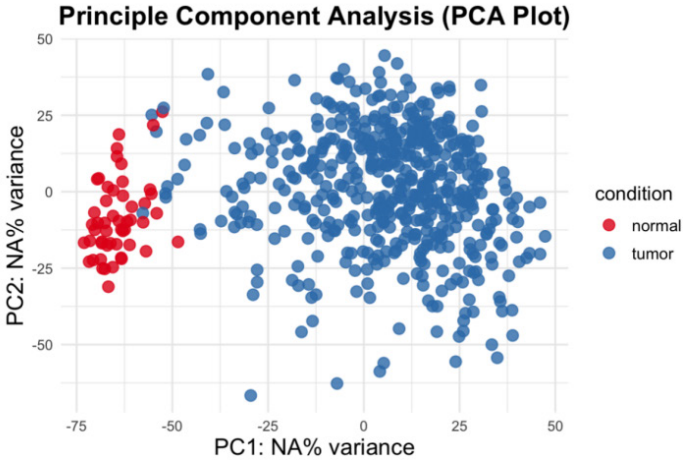

Figure 6 PCA Plot (LUSC)

In Figure 5 LUAD PCA plot, tumor samples (colored blue) and normal samples (colored red) form distinct clusters. This clear separation indicates that tumor and normal tissues have very different gene expression patterns. Similarly, in Figure 6 LUSC PCA plot, tumor samples and normal samples also cluster separately, suggesting that LUSC tumors have a distinct gene expression profile compared to normal lung tissue.

While both LUAD and LUSC show clear separation between tumor and normal samples, the pattern of clustering differs slightly. For example, LUAD tumors cluster more tightly and regularly, suggesting more homogeneous gene expression, while LUSC tumors show more spread, indicating greater heterogeneity. The difference in clustering could reflect the distinct molecular characteristics of

RGCC was identified as down-regulated DEG in both LUAD and LUSC, with similar expression levels and adjusted p-value. Therefore *RGCC* is not considered to represent a differential gene expression difference between the subtypes.

4.1.2 PCA plots

The plots show how samples cluster based on their gene expression profiles.

LUAD and LUSC.

4.2 Disease-Free Survival Analysis

In this GEPIA single-gene disease-free survival analysis, 8 DEGs (4 LUAD and 4 LUSC) were selected and analyzed. Among them, 4 DEGs were found to have no prognostic significance, which include: LUAD down-regulated gene SEMA3G, LUSC up-regulated genes *TPX2* and *TTK*, and LUSC down-regulated gene *ROBO4*.

4.2.1 Kaplan-Meier (KM) survival graphs

Kaplan-Meier (KM) curves illustrate the probability of survival over time for high and low expression groups.

LUAD

Up-regulated genes: *FAM83A* & *FAM83A-AS1*

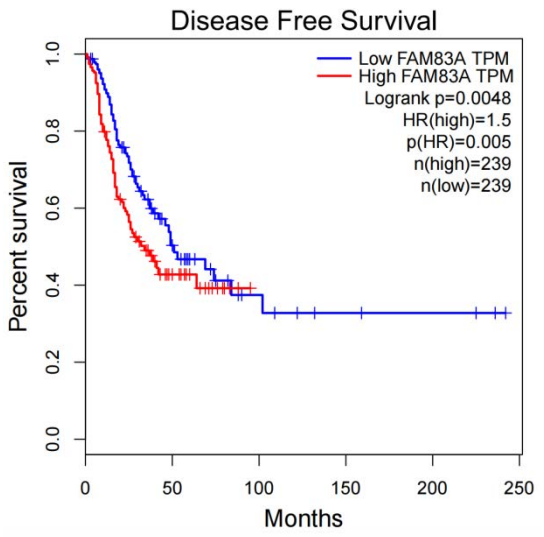

Figure 7 KM graph for FAM83A

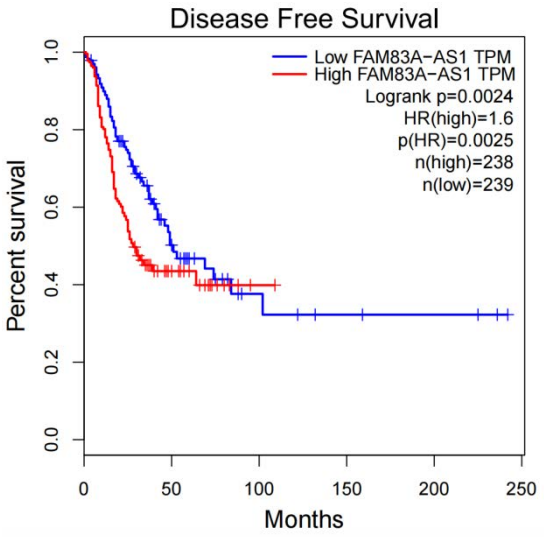

Figure 8 KM graph for FAM83A-AS1

Figure 7 shows that for *FAM83A*, the gene has a very low log-rank p-value of 0.0048, with a hazard ratio (HR) of 1.5. This indicates that patients with high *FAM83A* expression have a 1.5 times greater risk of cancer recurrence compared to those with low expression. Similarly, for *FAM83A-AS1*, the log-rank p-value is 0.0024 with an HR of 1.6; the low log-rank p-value suggests an even stronger association between high expression and poor disease free survival. More specifically, the red curve for *FAM83A* ends at approximately 100 months, while *FAM83A-AS1* ends at around 125 months. In contrast, the blue curve continues throughout the observed timeline, indicating better survival for patients with low gene expression in the long term. These findings demonstrate that *FAM83A* and *FAM83A-AS1* have prognostic meaning in LUAD, with high expression levels associating with worse clinical outcomes.

LUAD

Down-regulated genes: *PECAM1* & *SEMA3G*

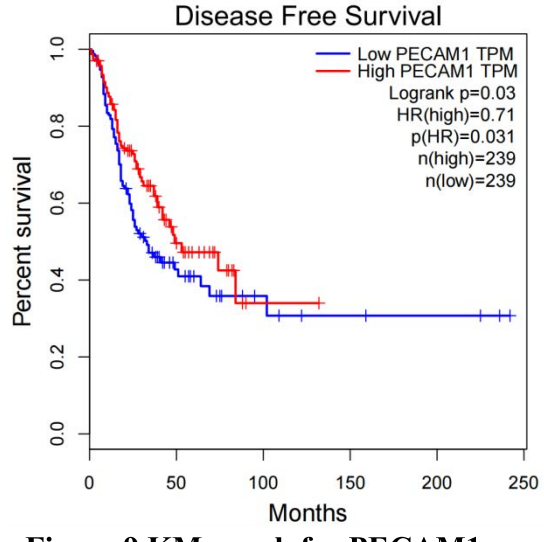

Figure 9 KM graph for PECAM1

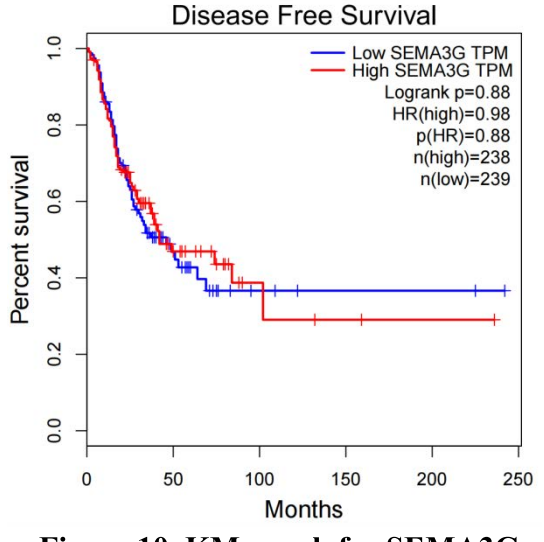

Figure 10 KM graph for SEMA3G

Figure 9 shows a low log-rank p-value of 0.03 for *PECAM1*, with a hazard ratio (HR) of 0.71, meaning that patients with high *PECAM1* expression have a 29% lower risk of recurrence or progression compared to those with low expression. Since this gene is down-regulated in LUAD, it indicates that its lower expression is linked to poor survival, demonstrating its prognostic significance. In contrast, looking at Figure 10, *SEMA3G* has a high log-rank p-value of 0.88 and a HR of 0.98 , both values approach 1. This means that there is no significant association between *SEMA3G* expression and disease free survival. Moreover, the overlapping of the KM curves for

ISSN 2959-409X

SEMA3G in multiple regions further indicates that its expression levels do not impact patient outcomes in LUAD. To further validate this, the blue curve for low expression of *SEMA3G* continues throughout the timeline similarly to the red curve, meaning the gene does not impact survival.

LUSC

Upregulated genes: *TPX2* & *TTK*

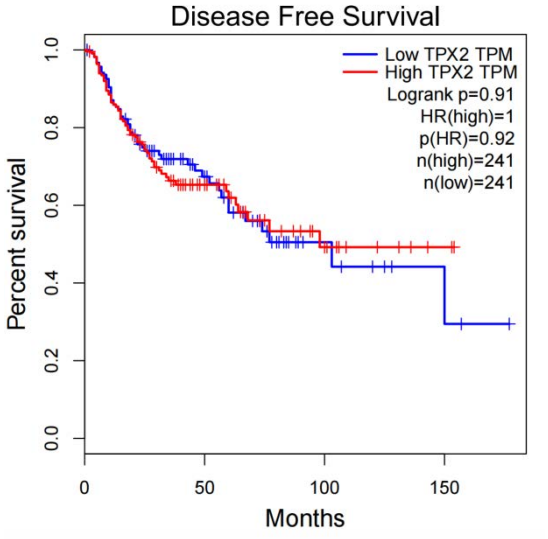

Figure 11 KM graph for TPX2

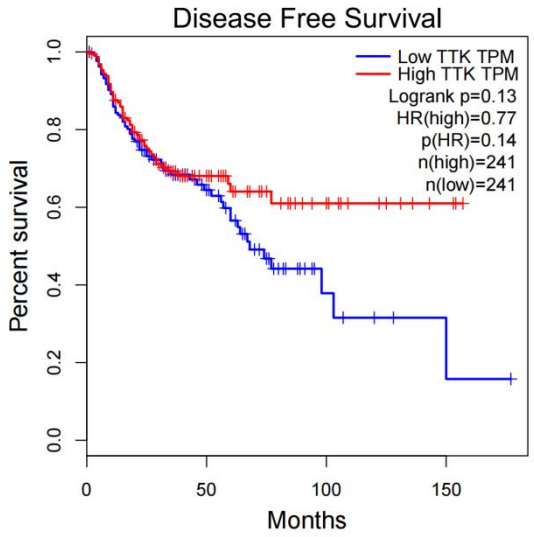

Figure 12 KM graph for TTK

For up-regulated genes in LUSC, both *TPX2* and *TTK* show no significant association with disease-free survival (DFS), considering its prognostic significance. *TPX2* has a log-rank p-value of 0.91 and a HR of 1 (p(HR) = 0.92). For *TTK*, the log-rank p-value is 0.13, with an HR of 0.77 (p(HR) = 0.14). Log-rank p-values for both gene are greater than 0.05, confirming the lack of prognostic significance.

LUSC

Down regulated genes: *RAMP2* & *ROBO4*

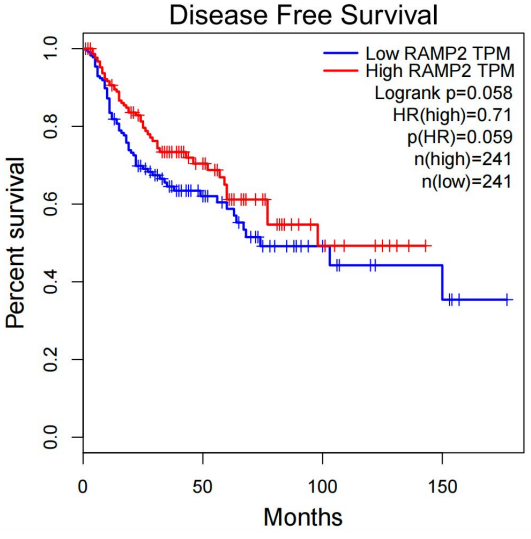

Figure 13 KM graph for RAMP2

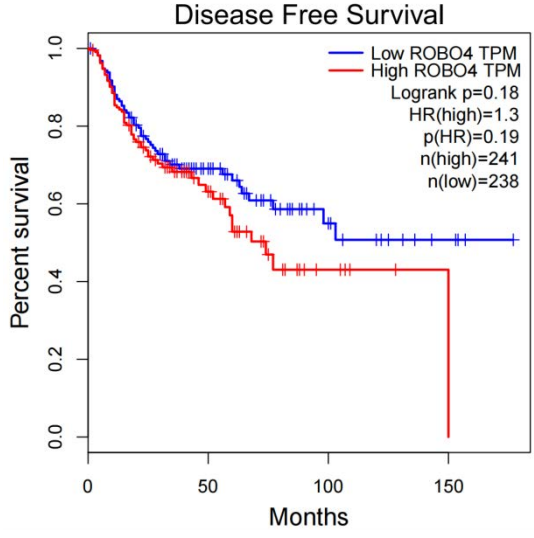

Figure 14 KM graph for ROBO4

For *RAMP2* in Figure 13, the log-rank p-value is 0.058, with an HR of 0.71, suggesting a marginal trend toward better outcomes in patients with high *RAMP2* expression. Although the log-rank p-value exceeds the conventional 0.05 threshold for significance, the HR below 1 indicates that low *RAMP2* expression is associated with worse survival. Upon careful interpretation of the KM curve, it is reasonable to suggest that *RAMP2* shows some prognostic significance, as the curves generally do not intersect; a low expression of *RAMP2* influences survival and is associated with poorer outcome, a trend which is consistent with what is expected. For *ROBO4* presented in Figure

14, the log-rank p-value is 0.18, greater than 0.05, and a HR of 1.3 (greater than 1). It is obvious to see a non-significant trend toward worse outcomes in patients with low *ROBO4* expression.

4.3 Pathway Enrichment Analysis

Several significant pathways were identified using the KEGG 2021 Human database in this pathway enrichment analysis. Key findings show that the cell cycle pathway was the most significant in both subtypes, and the DNA replication pathway was much more significant in LUSC compared to LUAD. The Fanconi anemia pathway and progesterone-mediate oocyte maturation pathway were significant (adjp < 0.1) in both subtypes. Other prominent pathways include the malaria pathway in LUAD and the p53 signaling pathway in LUSC.

4.3.1 LUAD key pathways bar graph and table

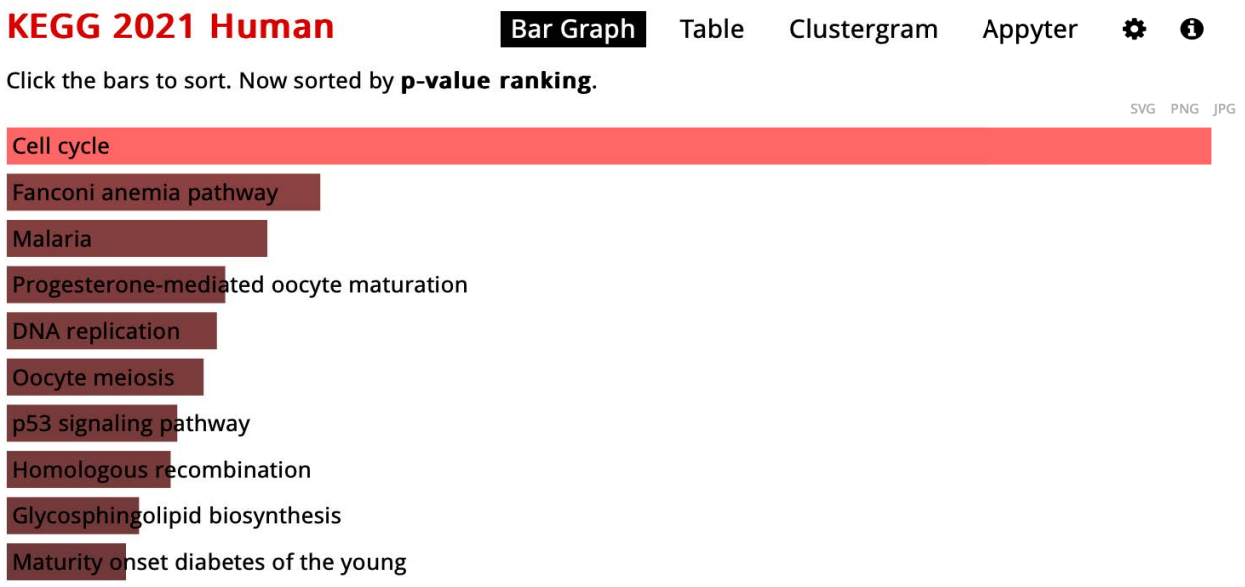

Figure 15 KEGG 2021 Human pathway bar graph (LUAD)

ISSN 2959-409X

KEGG 2021 Human Bar Graph **Table** Clustergram Appyter

Hover each row to see the overlapping genes.

25 entries per page Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Cell cycle	3.904e-10	1.089e-7	5.15	111.54
2	Fanconi anemia pathway	0.0003001	0.04186	4.35	35.30
3	Malaria	0.0007411	0.06892	4.20	30.27
4	Progesterone-mediated oocyte maturation	0.001419	0.09896	2.86	18.78
5	Oocyte meiosis	0.001991	0.1111	2.52	15.69
6	p53 signaling pathway	0.003313	0.1541	3.04	17.34
7	Homologous recombination	0.003873	0.1544	3.93	21.84
8	Glycosphingolipid biosynthesis	0.006578	0.2294	3.52	17.67
9	DNA replication	0.008277	0.2361	3.82	18.30
10	Maturity onset diabetes of the young	0.008463	0.2361	4.54	21.67
11	Human T-cell leukemia virus 1 infection	0.01405	0.3564	1.82	7.76
12	Bladder cancer	0.01548	0.3599	3.27	13.63
13	MicroRNAs in cancer	0.02293	0.4526	1.61	6.07
14	ABC transporters	0.02372	0.4526	2.93	10.98
15	Cellular senescence	0.02475	0.4526	1.89	6.97
16	Complement and coagulation cascades	0.02595	0.4526	2.26	8.26
17	ECM-receptor interaction	0.03160	0.4995	2.18	7.51
18	Mucin type O-glycan biosynthesis	0.03223	0.4995	3.07	10.56
19	Alanine, aspartate and glutamate metabolism	0.03578	0.5254	2.98	9.92
20	Axon guidance	0.03984	0.5558	1.72	5.53

Figure 16 KEGG 2021 Human pathway table (LUAD)

In LUAD, the cell cycle pathway was the most significant, as indicated by an adjusted p-value of 1.089*E^-7 in Figure 16. This extremely low adjusted p-value suggests that dysregulation of the cell cycle pathway is associated with aggressive cell division and rapid tumor growth. Other significant pathways (adjp < 0.1) include the Fanconi anemia pathway (adjp = 0.04186), the malaria pathway (adjp = 0.06892), and the progesterone-mediated oocyte maturation pathway (adjp = 0.09896).

4.3.2 LUSC key pathways bar graph and table

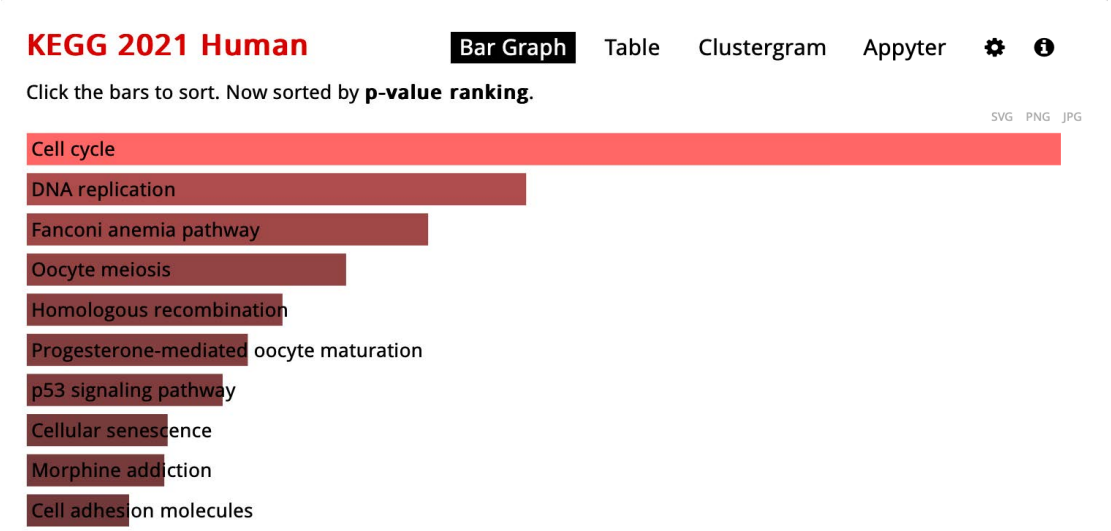

Figure 17 KEGG 2021 Human pathway bar graph (LUSC)

KEGG 2021 Human Bar Graph Table Clustergram Appyter

Hover each row to see the overlapping genes.

25 entries per page Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Cell cycle	7.927e-15	2.227e-12	6.79	220.59
2	DNA replication	9.384e-8	0.00001319	9.60	155.40
3	Fanconi anemia pathway	0.000001848	0.0001731	6.09	80.40
4	Oocyte meiosis	0.00002243	0.001576	3.33	35.61
5	Homologous recombination	0.0001551	0.008718	5.38	47.22
6	Progesterone-mediated oocyte maturation	0.0004450	0.02084	3.12	24.10
7	p53 signaling pathway	0.0009559	0.03837	3.40	23.62
8	Cellular senescence	0.005099	0.1761	2.19	11.56
9	Morphine addiction	0.005639	0.1761	2.63	13.62
10	Cell adhesion molecules	0.01645	0.4623	2.00	8.21
11	Drug metabolism	0.01922	0.4649	2.17	8.57
12	Estrogen signaling pathway	0.01985	0.4649	2.01	7.86
13	Base excision repair	0.02292	0.4710	3.40	12.86
14	Adherens junction	0.02498	0.4710	2.42	8.94
15	Mismatch repair	0.02573	0.4710	4.01	14.68
16	Human T-cell leukemia virus 1 infection	0.02682	0.4710	1.71	6.20
17	GABAergic synapse	0.03366	0.5440	2.15	7.28
18	Rap1 signaling pathway	0.03485	0.5440	1.69	5.66
19	Phenylalanine metabolism	0.05018	0.7321	4.08	12.21
20	Renin secretion	0.05633	0.7321	2.15	6.19

Figure 18 KEGG 2021 Human pathway table (LUSC)

In LUSC, the cell cycle pathway was highly significant, with an adjusted p-value of 2.227*E^-12, which is significantly lower than in LUAD. Additionally, the DNA replication pathway was more significant in LUSC with an adjusted p-value of 0.00001319 compared to a non-significant adjusted p-value of 0.2361 in LUAD. LUSC also contained a greater number of pathways with adjp lower than 0.1, including the Fanconi anemia pathway (adjp =

ISSN 2959-409X

0.0001371), oocyte meiosis pathway(adjp = 0.001576), homologous recombination pathway (adjp = 0.008718), progesterone-mediated oocyte maturation pathway(adjp = 0.02084), and the p53 signaling pathway (adjp = 0.03837).

5 Discussion

5.1 Limitations

This study exist several limitations. Firstly, all the results obtained still need to be rigorously evaluated in clinical practice. Secondly, the selection of databases may influence the outcomes, as different databases contain different datasets, which could lead to discrepancies in the findings. The disease-free survival analysis was conducted on a limited number of genes using single-gene survival analysis. Performing a multi-gene survival analysis could enhance the precision of the results. Moreover, the differentially expressed genes without prognostic significance were included for pathway enrichment analysis, which may dilute the relevance of the identified pathways. To improve the current study, non-prognostic DEGs like *SEMA3G*, *TPX2*, *TTK*, and *ROBO4* could be excluded to obtain more biologically meaningful pathways.

5.2 Explanation of DGE Visualizing Plots

5.2.1 Volcano plots

Volcano plots show the relationship between gene expression changes and statistical significance. Genes on the top left and top right quadrants are significant DEGs with large fold changes and extremely low p-values. In Figure 3, *FAM83A* is located in the top right quadrant, meaning that this gene has both a high positive log2 fold change (high expression in tumor tissue) and a highly significant low adjusted p-value, which reveals its significance in LUAD. Genes clustered near the center of the plot (around log2 fold change = 0) means that they are not significantly differentially expressed. These genes have a low Log2 fold change meaning that they have low expression and therefore not counted as DEGs. Comparing the distribution of genes in Figure 3 and 4, the genes are similarly expressed but in Figure 4, DEGs locates more to the top, the labeled 10 up-regulated DEGs of LUSC has much lower adjp which shows higher statistical significance. This suggests that the LUSC up-regulated genes are more consistently and strongly associated with the tumor phenotype compared to LUAD. The top 20 DEGs visualized on the volcano plots for both LUAD and LUSC are characterized by extremely low adjusted p-values (approaching zero), indicating a high degree of statistical significance.

5.2.2 PCA plots

Samples with similar expression profiles cluster together, which helps to identify patterns. The separation between tumor and normal samples in both Figure 5 and Figure 6 supports the idea that cancer could induces widespread changes in gene expression, including up-regulation of genes involved in cell proliferation or immune evasion, as well as down-regulation of genes involved in cell apoptosis and the regulation of normal cell functions. To validate, Zengin, T. and Önal-Süzek, T. (2021) state that LUAD signature genes plays role in immune-related pathways that are different from those in LUSC. Although LUAD and LUAC have overlapping signature gene pathways and share similar differential expression pathways, including a total of 2106 DEGs, they cluster separately in PCA plot. The difference in clustering also highlights the potential for using gene expression profiles to distinguish tumor from normal tissue, which could improve diagnosis and treatment for LUAD and LUSC.

5.3 Evaluating Prognostic Significance of DEGs

5.3.1 Further interpretation of KM graphs

Kaplan-Meier (KM) survival graphs were generated for significant DEGs from the volcano plots, stratifying patients into high expression (red curve) and low expression groups (blue curve) based on median expression. A clear gap between the blue and red curves indicates that the DEG impacts survival. For instant, the KM curves for the up-regulated genes *FAM83A* and *FAM83A-AS1* in LUAD (Figure 7 and 8) exhibit a clear gap with no intersection before 100 months. As a counter-example, the KM graph for the up-regulated gene *TTK* in LUSC shows overlapping survival curve before 50 months, with the high-expression (red) curve even showing higher survival than the low-expression (blue) curve, which is opposite to the expected result of high-expression linked with worse survival. The KM curves of both *TTK* and *TPX2* showed overlapping survival curves, reinforcing the weak prognostic significance of *TPX2* and *TTK* in LUSC.

5.3.2 Biological and clinical implications

As the LUAD up-regulated genes FAM83A and FAM83A-AS1 have strong prognostic significance, they are suitable as potential therapeutic targets for LUAD treatment. This is supported by Zheng, Y.-W. et al. (2020) , who state that the over-expression of *FAM83A* enhanced the proliferation, colony formation, and invasion of lung cancer cells, and is associated with poor prognosis. Similarly, increased *FAM83A-AS1* expression leads to LUAD cell proliferation and metastasis, further promoting NSCLC progression via the ERK signaling pathway. [24] Therefore, inhibiting *FAM83A* and *FAM83A-AS1* expres-

sion may improve patient survival by reducing tumor aggressiveness, consistent with their characterization as oncogenes. Similarly, the LUAD down-regulated gene PECAM1 has protective role when higher expressed, suggesting that enhancing its expression or activity could be a new strategy for LUAD treatment. This is validated by Cao, S. et al. (2021), who carried out a survival analysis showing that high expression of *PECAM1* was associated with improved survival. Furthermore, the study found that the overall survival of the *PECAM1* high-expression group of postoperative patients with lung cancer shows a better trend than the low-expression group. In LUSC the statistically significant up-regulated genes TPX2 and TTK are not associated with patient survival, suggesting that LUSC mechanism relies less on tumor aggressiveness. The lack of prognostic significance of TPX2 and TTK also emphasizes the molecular heterogeneity of this subtype and the need for subtype-specific therapeutic strategies. Lastly, the marginal significance of LUSC down-regulated gene RAMP2 requires further investigation to determine its real prognostic value in LUSC.

5.4 Key Pathway Differences and Processes

The pathway enrichment analysis revealed various distinct molecular pathways in LUAD and LUSC. The cell cycle pathway is a biological process by which cells grow, replicate DNA, and undergo mitosis. It was the most significant pathway in both subtypes. However, by comparing the bar graphs in Figure 15 and 17, cell cycle pathway in LUAD shows a higher proportion of significance than LUSC. This suggests that dysregulation of the cell cycle pathway is a critical driver of tumor growth in both subtypes, but LUAD may rely more heavily on cell cycle dysregulation for its progression. Furthermore, this discovery aligns with the disease-free survival analysis and previous studies showing that LUAD often exhibits over-expression of genes involved in cell cycle regulation, such as *FAM83A* and *FAM83A-AS1*, which promote cell proliferation and are associated with worsened survival.

Another notable difference was that the DNA replication pathway was much more significant in LUSC (adjp = 0.00001319) than in LUAD (adjp = 0.2361), meaning that the DEGs involved in this biological process are highly active. This suggests that LUSC tumors may rely more heavily on rapid DNA replication for growth compared to LUAD tumors, which rely more on cell cycle dysregulation. Thus, this could indicate that LUSC tumors are rapidly replicating their DNA, potentially becoming more aggressive or resistant to treatment.

The malaria pathway was a unique pathway in LUAD with a relatively high significance (adjp = 0.06892). This pathway refers to the biological process by which the malaria parasite interacts with human cells during infection. While this pathway is not directly related to cancer, it could provide unique insights into immune evasion and stress responses, which are also critical for cancer cell survival.

6 Conclusion

In conclusion, this comprehensive bioinformatics analysis of lung adenocarcinoma and lung squamous cell carcinoma reveals significant differences in their gene expression profiles and molecular pathways, providing further insights into their distinct biological behaviors. To achieve the aim of this research, pathway enrichment analysis was carried out to reveal the differential gene expression difference of two subtypes. Key findings show difference in proportion of the most significant pathway-cell cycle pathway in both subtypes, with LUAD showing a stronger reliance on cell cycle dysregulation, supported by the over-expression of oncogenes like *FAM83A* and *FAM83A-AS1*, which promote tumor growth and worsen survival outcomes. In contrast, LUSC exhibited greater significance in the DNA Replication pathway, suggesting a reliance on rapid DNA replication, which may contribute to its aggressiveness or treatment resistance. Unique pathway- Malaria pathway have been found in LUAD, providing new insights into immune evasion and stress responses. these discoveries show the difference in molecular basis and mechanisms of LUAD and LUSC. Additionally, the disease-free survival analysis further validated the prognostic significance of key genes, with *FAM83A* and *FAM83A-AS1* in LUAD and the marginal significance of *RAMP2* in LUSC providing potential therapeutic targets. The clear separation of tumor and normal samples in PCA plots and the distinct clustering patterns in volcano plots emphasize the widespread changes in gene expression caused by cancer. These key findings not only emphasize the molecular heterogeneity between LUAD and LUSC but also highlight the importance of subtype-specific approaches in diagnosis, prognosis, and treatment of NSCLC subtypes. By integrating differential gene expression, survival data and pathway analysis, this study contributes to a deeper understanding of the molecular mechanisms of LUAD and LUSC and the exploration of more precise and effective therapeutic strategies.

References

[1] World Health Organization, 2023. *Lung Cancer.* [online] World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/lung-cancer [Accessed 4 Jan.

ISSN 2959-409X

2025].

[2] Cooper, G.M., 2000. *The Development and Causes of Cancer*. [online] Nih.gov. Available at: https://www.ncbi.nlm.nih.gov/books/NBK9963/ [Accessed 6 Jan. 2025].

[3] Myers, D.J. and Wallen, J.M., 2023. *Lung adenocarcinoma*. [online] Nih.gov. Available at: https://www.ncbi.nlm.nih.gov/books/NBK519578/ [Accessed 20 Jan. 2025]

[4] National Cancer Institute, 2024. *Cancer of the Lung and Bronchus - Cancer Stat Facts.* [online] National Cancer Institute. Available at: https://seer.cancer.gov/statfacts/html/lungb.html [Accessed 16 Nov. 2024].

[5] American Cancer Society, 2022. *Oncogenes, Tumor Suppressor Genes, and DNA Repair Genes*. [online]www.cancer.org. Available at: https://www.cancer.org/cancer/understanding-cancer/genes-and-cancer/oncogenes-tumor-suppressor-genes.html [Accessed 11 Feb. 2025].

[6] Wikipedia Contributors, 2019. *Gene*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Gene [Accessed 11 Jan. 2025].

[7] Piña-Sánchez, P. et al., 2021. Cancer Biology, Epidemiology, and Treatment in the 21st Century: Current Status and Future Challenges From a Biomedical Perspective. *Cancer Control : Journal of the Moffitt Cancer Center*, 28, p.10732748211038735. doi:10.1177/10732748211038735.

[8] National Cancer Institute, 2018. *Cancer-Causing Substances*. [online] National Cancer Institute. Available at: https://www.cancer.gov/about-cancer/causes-prevention/risk/substances [Accessed 9 Feb. 2025].

[9] Weinhold, B., 2006. Epigenetics: The Science of Change. *Environmental Health Perspectives*, 114(3). doi:https://doi.org/10.1289/ehp.114-a160.

[10] Mukherjee, S., Dasgupta, S., Mishra, P.K. and Chaudhury, K., 2021. Air pollution-induced epigenetic changes: disease development and a possible link with hypersensitivity pneumonitis. *Environmental Science and Pollution Research International*, 28(40), pp.55981–56002. doi:https://doi.org/10.1007/s11356-021-16056-x.

[11] admin-science, 2023. *Genetics*. [online] Genetics. Available at:https://scienceofbiogenetics.com/articles/genes-in-cells-understanding-the-fundamental-building-blocks-of-life [Accessed 10 Feb. 2025].

[12] Chen, C., Huang, H. and Wu, C.H., 2017. Protein Bioinformatics Databases and Resources. *Methods in molecular biology (Clifton, N.J.)*, 1558, pp.3–39. doi:https://doi.org/10.1007/978-1-4939-6783-4_1.

[13] National Cancer Institute, 2022. *The Cancer Genome Atlas Program (TCGA) - NCI.* [online] www.cancer.gov. Available at: https://www.cancer.gov/ccg/research/genome-sequencing/tcga [Accessed 15 Feb. 2025].

[14] Zengin, T. and Önal-Süzek, T., 2021. Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma. *Journal of Personalized Medicine*, 11(2), p.154. doi:https://doi.org/10.3390/jpm11020154.

[15] Center for Cancer Genomics, 2019. *About the Program - NCI*. [online] www.cancer.gov. Available at: https://www.cancer.gov/ccg/research/genome-sequencing/tcga/history [Accessed 15 Feb. 2025].

[16] The Cancer Genome Atlas Research Network, 2012. Comprehensive genomic characterization of squamous cell lung cancers. Nature, 489(7417), pp.519–525. doi:https://doi.org/10.1038/nature11404

[17] Rosati, D. et al., 2024. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A Review. *Computational and Structural Biotechnology Journal*, 23(1154–1168). doi:https://doi.org/10.1016/j.csbj.2024.02.018.

[18] Shreffler, J. and Huecker, M.R., 2023. *Hypothesis Testing, P Values, Confidence Intervals, and Significance.* [online] PubMed. Available at: https://www.ncbi.nlm.nih.gov/books/NBK557421/. [Accessed 15 Feb. 2025]

[19] Jafari, M. and Ansari-Pour, N., 2019. Why, When and How to Adjust Your P Values? *Cell Journal (Yakhteh)*, 20(4), pp.604–607. doi:https://doi.org/10.22074/cellj.2019.5992.

[20] Renesh Bedre, 2024. *What is Log2 Fold Change in Bioinformatics.* [online] RS Blog. Available at: https://www.reneshbedre.com/blog/log-fold-change.html [Accessed 15 Feb. 2025].

[21] Wang, T., Zhang, L., Tian, P. and Tian, S., 2017. Identification of differentially-expressed genes between early-stage adenocarcinoma and squamous cell carcinoma lung cancer using meta-analysis methods. *Oncology Letters*, 13(5), pp.3314–3322. doi:https://doi.org/10.3892/ol.2017.5838.

[22] The Cancer Genome Atlas Research Network, 2014. Comprehensive molecular profiling of lung adenocarcinoma. Nature, 511(7511), pp.543–550. doi:https://doi.org/10.1038/nature13385.

[23] Seo, A.N. et al., 2014. FGFR1 amplification is associated with poor prognosis and smoking in non-small-cell lung cancer. *Virchows Archiv*, 465(5), pp.547–558. doi:https://doi.org/10.1007/s00428-014-1634-2.

[24] Wang, W. et al., 2021. LncRNA FAM83A-AS1 promotes lung adenocarcinoma progression by enhancing the pre-mRNA stability of FAM83A. *Thoracic cancer*, 12(10), pp.1495–1502. doi:https://doi.org/10.1111/1759-7714.13928.

[25] Tang, Z. et al., 2017. *GEPIA About*. [online] gepia.cancer-pku.cn. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res, 10.1093/nar/gkx247. Available at: http://gepia.cancer-pku.cn/about.html. [Accessed 8 Feb. 2025].