



Genetic Links Between Common Lung Diseases and Lung Cancer Progression: Bioinformatics and Machine Learning Insights

Md Ali Hossain ^{1, 2}, Tania Akter Asa ^{2, 3}, Md. Zulfiker Mahmud ³, AKM Azad ⁴,
Mohammad Zahidur Rahman ^{1*}, Mohammad Ali Moni ^{5, 6*}, Ahmed Moustafa ^{7, 8*}

¹ Department of Computer Science & Engineering, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh.

² Health Informatics Lab, Department of Computer Science & Engineering, Daffodil International University, Dhaka 1216, Bangladesh.

³ Department of Computer Science and Engineering, Jagannath University, Dhaka 1100, Bangladesh.

⁴ Department of Mathematics & Statistics, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 13318, Saudi Arabia.

⁵ Artificial Intelligence and Cyber Futures Institute, Charles Sturt University, Bathurst 2795, Australia.

⁶ Rural Health Research Institute, Charles Sturt University, Orange 2800, Australia.

⁷ Department of Human Anatomy and Physiology, Faculty of Health Sciences, University of Johannesburg, Doornfontein, 2094, South Africa.

⁸ Centre for Data Analytics and School of Psychology, Bond University, Gold Coast, Queensland, 4229, Australia.

Abstract

Lung cancer (LC) is one of the most frequently diagnosed cancers and remains the leading cause of cancer-related mortality worldwide, representing a significant global health challenge. While numerous common lung diseases (CLDs) are implicated in LC development, the underlying causes of LC originating from CLDs remain inadequately elucidated. A thorough exploration of LC's progression from CLDs is essential; our approach integrated bioinformatics and machine learning, utilizing data from GEO and TCGA databases. We began by identifying differentially expressed genes (DEGs) in LC and CLDs, and our gene-disease network revealed for the first time shared DEGs (LC shares significant genes with TB (36), asthma (10), pneumonia (17), COPD (18), and Idiopathic Pulmonary Fibrosis (IPF) (78)), providing insights into potential connections of LC with CLDs. This analysis not only broadened our understanding of their associations but also identified significant pathways and hub proteins (SPTBN1, KCNA4, SCN7A, KCNQ3, GRIA1, and SDC1) through a protein-protein interaction network (PPI). Furthermore, RNA-seq and clinical data were obtained from the cBioPortal portal for shared DEGs of LC and CLDs, assessing their impact on LC patient survival. Integrated mRNA-Seq and clinical data were analyzed via univariate and multivariate Cox Proportional Hazard models to elucidate the influence of significant genes on survival. Furthermore, we developed and deployed a predictive model leveraging the identified hub genes, which demonstrated high accuracy in predicting LC progression. The identified biomarkers and pathways hold promise for further translational research and potential therapeutic targets, advancing understanding of LC development from CLDs. Additionally, co-expression networks among common genes were explored using the Weighted Gene Co-expression Network Analysis (WGCNA). Finally, the hub genes were validated using the Human Protein Atlas (HPA) database and evaluated through various classification algorithms to ascertain their predictive power and diagnostic potential.

Keywords:

Lung Cancer;
Commonly Lung Disorders;
Survival Curve;
COX PH Model;
Classification Algorithms;
PPI Network;
Molecular Pathways.

Article History:

Received:	23	September	2024
Revised:	27	February	2025
Accepted:	04	March	2025
Published:	01	April	2025

1- Introduction

Lung cancer (LC) is the most often diagnosed disease and the most common cause of cancer death in both men and women worldwide, killing about 1.8 million people each year [1]. Specifically, LC is characterized by the uncontrolled proliferation of abnormal cells within the lungs, primarily in the cells lining the airways. Furthermore, the disease is

* **CONTACT:** rmzahid@juniv.edu; m.moni@uq.edu.au; ahmedhalimo@gmail.com

DOI: <http://dx.doi.org/10.28991/ESJ-2025-09-02-021>

© 2025 by the authors. Licensee ESJ, Italy. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<https://creativecommons.org/licenses/by/4.0/>).

classified into two primary types: small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC) [2]. Notably, NSCLC accounts for around 85% of all cases and includes histological subtypes such as adenocarcinoma, squamous cell carcinoma, and large cell carcinoma [3]. On the other hand, SCLC is more aggressive than NSCLC, with faster doubling times and a higher propensity for early metastasis [4]. The progression of LC is influenced by a variety of risk factors, among which common lung diseases (CLDs) are particularly noteworthy. Moreover, several studies have identified five CLDs—tuberculosis (TB), asthma, idiopathic pulmonary fibrosis (IPF), chronic obstructive pulmonary disease (COPD), and pneumonias critical risk factors for LC [5-9]. For instance, COPD, particularly chronic bronchitis, is a prevalent lung disease primarily caused by smoking and is associated with breathing difficulties. Studies have demonstrated that chronic bronchitis significantly elevates LC risk [6]. Similarly, TB, a bacterial infection caused by *Mycobacterium tuberculosis* (MTB), predominantly affects the lungs and spreads through the air when infected individuals cough, spit, speak, or sneeze. TB has also been shown to heighten LC risk [5, 10].

Furthermore, we found in one study that asthma increases the risk of LC [11]. Another note-worthy risk factor is IPF, a progressive disease that not only increases LC risk but also presents challenges in treating LC among IPF patients due to the potential exacerbation of fibrosis [12-14]. Despite the high prevalence of lung cancer in IPF patients, the pathogenesis and treatment strategies remain unclear [12]. Lastly, pneumonia, an acute respiratory infection, has also been implicated in LC risk, further emphasizing the role of CLDs as precursors to LC [7]. Indeed, several recent studies found that CLDs increase the risk of LC [11, 15, 16]. A recent study by Miron et. al. [17] provides an overview of the relationship between chronic lung diseases and lung cancer, discussing common risk factors, biological mechanisms, and their impact on patient prognosis.

In addition, bioinformatics and machine learning approaches have been increasingly used to explore the molecular mechanisms underlying LC progression from CLDs in recent years. For instance, researchers have identified shared differentially expressed genes (DEGs) and transcriptional regulators between LC and other lung diseases, offering insights into potential mechanisms of cancer development in patients with pre-existing lung conditions [18, 19]. Yao et al. applied bioinformatics approaches to identify potential therapeutic targets involved in the progression of IPF to NSCLC [19]. Similarly, Dasgupta utilized bioinformatics techniques to uncover potential targets for therapeutic intervention in the context of interstitial lung disease and LC [20]. Furthermore, a recent study by Ali et al. identified key genetic pathways linking lung cancer, smoking, and COVID-19, highlighting therapeutic targets using bioinformatics and machine learning [21]. Notably, machine learning techniques have been instrumental in analyzing large-scale datasets, enhancing lung cancer diagnosis, prognosis prediction, and treatment planning [19]. Additionally, some recent studies have highlighted the association between CLDs and LC [11, 15-17, 21-23]. However, while these investigations provide critical insights into the molecular mechanisms underlying LC in the context of CLDs, the precise mechanisms and the extent to which CLDs contribute to LC risk remain unclear. Furthermore, current studies have yet to establish how commonly CLDs increase the likelihood of LC development. The genetic associations between significant CLDs and LC need to be validated, and causal links between these conditions must be elucidated. Consequently, there is an urgent need for integrative bioinformatics and machine learning models that can explore shared pathways and genetic mechanisms between CLDs and LC. Such efforts will not only enhance our understanding of disease progression but also open new avenues for personalized care and therapeutic interventions.

In the present study, bioinformatics and machine learning techniques were utilized to identify genes associated with the risk of developing LC in the presence of various common lung diseases such as TB, asthma, IPF, COPD, and pneumonia. The analysis involved high-throughput transcriptomics data analysis, PPI sub-network reconstruction, gene ontologies, and molecular pathways using a network-based "multi-omics" approach to understand the genetic influence of these factors on the progression of LC. To investigate the effect of clinical factors and disease marker gene expression on LC patient survival, standard Cox Proportional Hazard (PH) models were used for univariate and multivariate analyses. The study started by identifying differentially expressed genes (DEGs) of LC and 5 CLDs, followed by identifying common genes between LC and the 5 CLDs. A gene-disease association network was constructed using these shared genes to see the association of LC with CLDs.

The Weighted Gene Coexpression Network Analysis (WGCNA), a widely used method for identifying clusters of correlated genes [24], was applied using the WGCNA R package to construct a correlation network and assess the interrelationships among common genes [25]. We used publicly accessible clinical data from the Broad Institute Cancer Genome Atlas (TCGA) datasets as well as LC gene transcription profiles to correlate LC patient survival and other clinical variables with gene expression to identify new LC biomarkers that predict patient mortality. Furthermore, the study identified the mRNA of common genes of LC and CLDs by comparing these genes with the mRNA data of LC obtained from TCGA. These clinical and mRNA data were integrated and subjected to univariate

and multivariate analyses to identify genes that affect the survival of LC. Finally, common pathway and GO ontology analyses were performed on the commonly identified genes of LC with five CLDs. We used machine learning approaches to check the validity of the identified biomarker genes among the diseases on the lung cancer dataset. Furthermore, we used the HPA database to explore the protein expression levels of significant biomarker genes in normal tissues and lung cancer tissues through immunohistochemistry (IHC) testing. We subjected these genes to evaluation through classification algorithms. Additionally, we developed and deployed a predictive model using these identified hub genes to assess LC progression, offering a practical application of our findings in clinical settings. This deployment emphasizes the translational potential of our study, bridging the gap between molecular research and real-world clinical implementation.

The study employs bioinformatics and machine learning techniques to identify genes associated with lung cancer risk amidst common lung diseases like TB, asthma, IPF, COPD, and pneumonia. Biomarkers are validated through multi-omics analysis and machine learning, while protein expression levels are examined for robustness using the Human Protein Atlas database. This comprehensive approach enhances understanding of genetic influences on lung cancer progression and mortality prediction.

2- Materials and Methods

In this study, we analyzed publicly available microarray datasets from the NCBI Gene Expression Omnibus (GEO) database (<http://www.ncbi.nlm.nih.gov/geo/>) focusing on lung diseases such as LC, TB, asthma, COPD, and pneumonia. Additionally, mRNA-Seq and clinical data of LC were accessed from the Cancer Genome Atlas (TCGA) via the TCGA Genome Data Analysis Center (<http://gdac.broadinstitute.org/>), providing a comprehensive resource for investigating shared molecular mechanisms.

Figure 1 shows the work steps used in our work. Our research work follows the following steps:

- 1) DEG Identification: We identified DEGs for each disease (LC and common CLDs) using the limma R package for the microarray dataset.
- 2) Shared DEG Identification: We identified the shared significant DEGs between LC and CLDs by overlapping the DEGs obtained from the datasets of LC and CLDs.
- 3) Enrichment Analysis: We conducted pathway and Gene Ontology analyses on the shared significant DEGs to reveal shared biological pathways and functional categories.
- 4) Construction of PPI Network and Identification of Key Hub Genes: Using Cytoscape and the Cyto-Hubba plugin, we constructed a PPI network around the shared genes to identify highly connected hub proteins. We applied four algorithms (degree, EPC, MCC, and MNC) and identified shared hub genes.
- 5) Utilization of TCGA Data: mRNA sequencing data for lung cancer (LC) was sourced from the TCGA Genome Data Analysis Center, encompassing 510 cases and gene expression profiles for 20,510 genes.
- 6) Data Preprocessing: We categorized normal, tumor, and control samples using the TCGA barcode. We removed samples with missing values and lower read counts ($counts < 100$) and normalized the data using the FPKM method.
- 7) Identification of shared Gene FPKM: FPKM values for shared genes between LC and CLDs, as well as hub genes, were extracted from the LC dataset.
- 8) Survival Analysis: Univariate and multivariate Cox Proportional Hazard regression analyses were conducted to identify biomarker genes that significantly impact the survival of LC patients.
- 9) Performance Assessment: Classification algorithms were applied to the extracted FPKM values of shared and hub genes to evaluate their effectiveness in prediction. We also performed ROC curve and heat map of key hub genes.
- 10) Deployment: A machine learning-based prediction model was deployed to enable real-time prediction of LC progression using hub gene expression profiles.
- 11) We also checked the expression level of significant genes from the HPA database.
- 12) WGCNA Analysis: We performed WGCNA analysis on significant genes to explore correlations among them using the WGCNA package.

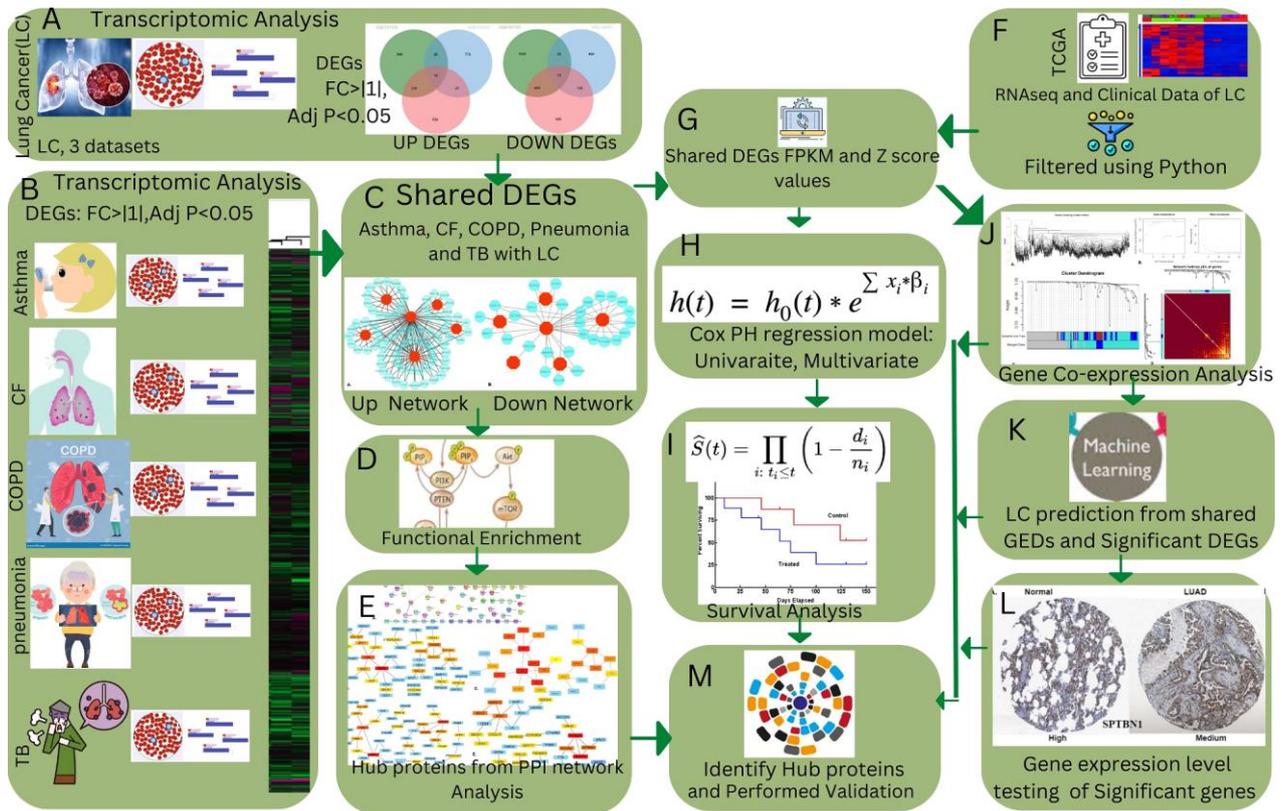


Figure 1. Flowchart of the research work. A to C. DEGs of each disease used in this study were identified, and then common genes between lung cancer (LC) and 5 CLD (Common Lung Diseases) were identified as well as a Gene- Disease association network of LC with 5 CLDs was built. D to E. Hub proteins and key pathways that are linked with comorbidity of LC with 5 CLDs were identified using the PPI network, common pathway, and GO ontology analysis. F to G. Following that, LC RNAseq data was converted to FPKM and then to Z score values, after that, mRNA of shared genes were identified, and next, the mRNA data and clinical data were combined by utilizing patient IDs, allowing for the integration of molecular and clinical information to gain a comprehensive understanding of LC and its associated factors. H. Univariate, and multivariate, Cox Proportional Hazard regression analyses were conducted to identify biomarker genes that significantly impact the survival of LC patients. We utilized TCGA's mRNA data to procure and process LC information, identifying shared and hub genes. K. These genes underwent assessment via classification algorithms. I, J and L. Concurrently, we performed survival analysis of significant genes, WGCNA Analysis to unveil inter-gene correlations and also checked the expression level of significant genes. M. Identified significant Hub proteins through validation. This comprehensive investigation shed light on the intricate relationships among LC and CLDs, augmenting our comprehension of their intricate interplay.

2-1- Materials

For investigating the association of LC and significant lung diseases at the molecular level, we used seven gene expression microarray datasets with accession numbers GSE89039 and GSE136043 (LC), GSE62525 (tuberculosis), GSE35716 (Pneumoniae), GSE43696 (Asthma), GSE24206 (Idiopathic Pulmonary Fibrosis), and GSE76925 (COPD) obtained from the NCBI (<http://www.ncbi.nlm.nih.gov/geo/>). In this study, we applied standard statistical procedures, including filtering, normalization, and Student's unpaired t-test using the Limma package, to analyze transcriptomics datasets and identify differentially expressed genes (DEGs) based on a threshold of $|\log FC| \geq 1$ and $p\text{-value} < 0.05$. There were eight LUAD samples and eight normal samples in the GSE89039 dataset and in the GSE136043 dataset, the number of LUAD samples is 5 and normal samples is 5. The information on datasets used in this study is given in Table 1.

Table 1. Dataset descriptions of used LC and CLDs diseases

Diseases Name	GSE Number	Case	Control
Lung Cancer	GSE89039	Lung carcinoma tissue: 8	Non-cancerous lung Tissues: 8
Gray lung cancer	GSE136043	Lung cancer tissues: 5	Non-tumor tissues:5
Tuberculosis (TB)	GSE62525	Active TB: 14 and latent TB:14	Healthy Subject: 14
Gray pneumonia	GSE35716	Bacterial pneumonia: 10	Healthy controls (n=18)
Asthma	GSE43696	Severe asthmatic (SA) patients: 38	Normal control (NC): 20
Gray Idiopathic pulmonary fibrosis (IPF) or chronic fibrosing (CF)	GSE24206	Idiopathic pulmonary fibrosis: 17	Healthy controls: 6
COPD	GSE76925	COPD case: 111	Healthy controls: 40

For further analysis to obtain the significant genes, we used the mRNA data of LC that we obtained from the TCGA genome data analysis centre (<http://gdac.broadinstitute.org/>). We acquired the anonymised clinical data and mRNAseq data for LC (Lung Adenocarcinoma, TCGA, PanCancer Atlas) from the cBioPortal to investigate a specific topic of interest, survival analysis of LC on clinical and genetic determinants. The clinical dataset we used for our analysis contained 566 cases and 57 features. Out of these, 510 cases had mRNA gene expression data available, which included information on 20510 genes. We picked 3 crucial clinical factors patients ID, Overall Survival and Overall Survival status. Additionally, we identified significant genes associated with LC and lung diseases. We studied only one outcome variable, which was LC-specific survival, using the aforementioned data. By matching the patient IDs in both clinical and mRNA expression datasets, we discovered 510 patients with data available in both sets, and we used the same 3 clinical variables described above. We computed z-scores for RNAseq data using the process that was described in the work of Hossain et al. [26]. Utilizing the TCGA barcode, we categorized samples into normal and tumor groups. The sample type is indicated by the two digits at positions 14-15 of the barcode. Normal samples are denoted by digits ranging from 10 to 19, tumor samples from 01 to 09, and control samples from 20 to 29. We remove the missing value samples and lower the read count (total read count < 100). We then normalized the modified dataset with the FPKM method for the performance evaluation of the significant genes with classification algorithms and WGCNA analysis. To identify samples with gene expression that was either over-expressed or under-expressed, we used z-score values. We considered a sample to be altered if its z-score was equal to or greater than a particular threshold value (e.g., $z=2$). Accordingly, we defined altered samples as those with z-scores greater than or equal to 2, and normal samples as those with z-scores less than 2.

2-2- Methods

2-2-1- Significant DEGs Identification

We analyzed these datasets to find genes that were differently expressed in patients compared to normal samples. To normalize the datasets, we first use the Limma package in R to conduct the \log_2 transformation using combinatorial statistical approaches. To determine significant genes, the raw p-values were adjusted using the Benjamini-Hochberg method, and significance was assessed with an unpaired Student's t-test. A threshold of absolute \log_2 fold change value as at least 1 and an Adjusted p - value < 0.05 were set.

2-2-2- Diseasome Network Construction

We started by identifying DEGs for each of the disorders, and then we looked for genes that were shared by LC and five major lung diseases. After that, we used the concordant genes to create the LC diseasome network with five CLDs. For the diseasome network, topological and neighbourhood-based benchmark [27, 28] approaches were utilized, which were more suited to our networks. Using Cytoscape, we created the LC diseasome network [27].

2-2-3- Construction of PPI Networks and Identification of Hub Genes

The STRING database (<https://string-db.org/>) was used to obtain the protein-protein interactions of the overlapped DEGs of LC with CLDs [27, 29]. To ensure comprehensive analysis, we utilized interaction data from the STRING database, which incorporated information from PubMed abstracts, co-expression patterns, gene fusions, and genomic neighborhood associations. Furthermore, a combined interaction score with a medium confidence threshold (> 0.4) was applied as the cut-off value. For clearer visualization, the protein-protein interaction (PPI) network was displayed using Cytoscape (v3.9.1) in combination with the Cyto-Hubba plugin [27, 30]. Notably, Cyto-Hubba was instrumental in identifying highly connected hub proteins by applying algorithms such as degree, edge percolated component (EPC), maximal clique centrality (MCC), and maximum neighborhood component (MNC) [31]. The top 20 nodes with the degree, EPC, MCC, and MNC were chosen, and the hub proteins were determined by taking the intersection of the four algorithms.

2-2-4- Hub Genes Expression Levels Validation using the HPA Database

The HPA database was used to confirm the protein expression levels of hub genes in normal tissues and lung cancer tissues through immunohistochemistry (IHC) testing [32].

2-2-5- Machine Learning Models for Important Proteins Identification and Survival Predictions

The analysis conducted on patients with LC involved several techniques. Firstly, the product-limit estimator was used to estimate the survival function. Subsequently, the log-rank test was performed to determine if there were any significant differences between the two groups, namely patients with altered gene expression and those with unaltered gene

expression. Finally, Cox Proportional Hazards regression models were used to identify significant genes and clinical factors. The study used important clinical variables described in the data collection section, as well as significant shared genes associated with LC and common lung diseases. Each gene's Z-score value was converted to either an altered or normal category, depending on whether it exceeded the threshold value of Z-score ($z > 2$), as described in the data section. The analysis then proceeded to perform univariate analysis (examining each gene individually) and multivariate analysis (examining all genes simultaneously). We use the Cox proportional hazards (Cox PH) model for the above two analyses. The Cox Proportional Hazards (Cox PH) model is a regression model used to analyze the relationship between survival times and one or more predictor variables.

The formula for the univariate Cox PH model is:

$$h(t|x) = h_0(t) \cdot e^{\beta x} \quad (1)$$

where $h(t|x)$ represents the hazard function for an individual with covariate values x , $h_0(t)$ is the baseline hazard function, and β is the estimated coefficient for the predictor variable x .

The formula for the multivariate Cox PH model is:

$$h(t|x_1, x_2, \dots, x_p) = h_0(t) \cdot e^{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p} \quad (2)$$

where $h(t|x_1, x_2, \dots, x_p)$ represents the hazard function for an individual with covariate values x_1, x_2, \dots, x_p , $h_0(t)$ is the baseline hazard function, and $\beta_1, \beta_2, \dots, \beta_p$ are the estimated coefficients for the predictor variables x_1, x_2, \dots, x_p respectively.

2-2-6- Performance Evaluation of the Significant Genes with Classification Algorithms

The study evaluated the reliability of the identified shared genes of LC with five CLDs by using four popular classification algorithms, namely Bayesian Network, support vector machine (SVM), random forest (RF), and Logistic Regression. For this purpose, mRNA data of LC were normalized with the FPKM method and then the shared gene's normalized values of LC samples and normal samples.

We used 10-fold cross-check validation for the four algorithms. The aim was to determine the effectiveness of these hub genes in classifying different disease states using machine-learning techniques. The study used two performance measures to evaluate the effectiveness of the identified shared genes, which were accuracy, and area under the ROC curve (AUC). Furthermore, the best-performing classification model, based on accuracy and AUC, was deployed as a real-time prediction tool. This tool allows for the classification of LC progression by leveraging the expression profiles of significant hub genes.

2-2-7- Weighted Gene Co-expression Networks Construction

The endeavour to unravel clustering trends between shared genes identified from the LC and CLDS led us to employ the WGCNA package [24] in R. This approach facilitated the identification of weighted gene co-expression networks interlinking these genes. To embark on this journey, we initially addressed the prospect of outlier samples. By constructing a sample cluster dendrogram using the `hclust` function in R for both datasets, potential outlier samples were pruned. Subsequently, the `pickSoftThreshold` function within R guided us in exploring numerous soft thresholding powers (β) across R^2 , eventually pinpointing the value of β that yielded the highest R^2 .

The process further involved the construction of an adjacency matrix and a Topological Overlap Matrix (TOM), leveraging the transformed gene expression matrix. This matrix was pivotal in encapsulating the interconnectedness of the genes. Additionally, the Dissimilarity of TOM (dis-sTOM) was harnessed, affording a network heatmap plot that not only visually captured the relationships but also facilitated subsequent analytical endeavours. This concerted approach within the WGCNA framework illuminated the underlying co-expression patterns among the common genes, offering valuable insights into their potential functional collaborations and interactions.

2-2-8- Pathways and Ontology Analysis for Common DEGs of LC and Significant Lung Diseases

We used the Enrichr bioinformatics tool (<http://amp.pharm.mssm.edu/Enrichr/>) and the KEGG pathways database to perform pathway and gene ontology analysis to learn more about the molecular pathways of LC that overlap with tuberculosis (TB), Asthma, IPF, COPD and Pneumonia. We used adjusted p -value < 0.05 as the threshold Adjusted p -value for significant enrichment results.

3- Results and Discussion

3-1- Results

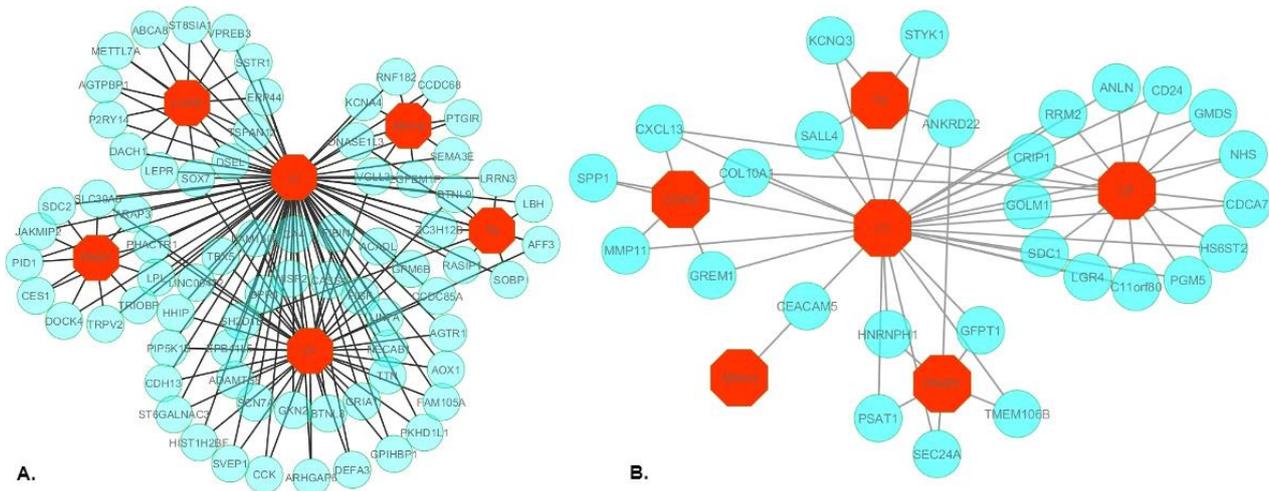
3-1-1- Infectome and Disease Analysis

We conducted a comprehensive global transcriptome analysis to examine gene expression patterns in lung cancer (LC) patient tissues, comparing these with those of normal subjects.

We used two datasets of LC and found 1652 differentially expressed genes (DEGs) (765 down-regulated, 887 up-regulated) in GSE136043 and 4028 DEGs (2659 up-regulated, 4028 down-regulated) in GSE89039. Next, a cross-comparison analysis was conducted between the two LC datasets, which identified 681 significant differentially expressed genes (DEGs), comprising 197 down-regulated and 484 up-regulated genes.

To observe the association of LC with the other 5 lung diseases, we have collected microarray raw data associated with each disease. After several rounds of statistical analysis, we identified the most significantly over- and underexpressed genes for each disease. Our analyses revealed a considerable number of differentially expressed genes: 1977 (421 downregulated and 1556 upregulated) in TB, 1277 (441 downregulated and 836 upregulated) in Pneumonia, 643 (54 downregulated and 589 upregulated) in COPD, 155 (51 downregulated and 104 upregulated) in Asthma, and 819 (440 downregulated and 379 upregulated) in IPF.

We also performed cross-comparative analysis to find the common significant genes between each disease and LC. We observed that LC shares 36, 10, 17, 18, and 78 significant genes with common LDIs TB, asthma, pneumonia, COPD, and IPF, respectively. To identify statistically significant connections between these infections and diseases, we constructed an infectome-disease relationship network focused on LC. In this network, two diseases were considered comorbid if they shared one or more associated genes (Figure 2). Notably, 2 significant genes, *BTNL9* and *AFF3*, are commonly dysregulated among LC, TB, and IPF; 3 significant genes, *CXCL13*, *COL10A1* and *SOX7* are commonly dysregulated among LC, COPD, and IPF; and 9 significant genes, *LRRN3*, *LBH*, *SH2D1B*, *SOBP*, *ZC3H12B*, *RASIP1*, *KCNQ3*, *SALL4* and *STYK1*, are commonly dysregulated among LC, TB, and ARDS; 1 significant gene (*DNASE1L3*) is commonly dysregulated among LC, TB, and asthma; and 1 significant gene (*NAPRT*) is commonly dysregulated among LC, TB, and IPF. Interestingly, only 1 gene (*VGLL3*) is common among LC, asthma and IPF. However, 1 gene (*DSEL*) plays an important role and is differentially expressed among LC, TB, IPF and COPD; and 1 gene (*ANKRD22*) among LC, Pneumonia, and TB.



through WGCNA, we determined the optimal soft thresholding power ($=9$) based on scale-free topology criteria (Figure 3-B). Using this power value, we constructed a co-expression network and identified four modules with the Dynamic Tree Cut technique, employing $deepS\ split = 2$ and $minClusterSize = 10$ parameters. These modules contained 28, 10, 37, and 43 genes in the blue, brown, grey, and turquoise modules, respectively. Additionally, three modules (comprising 8, 48, and 62 genes in the blue, grey, and turquoise modules) were identified using the Auto-merged algorithm. The module dendrogram plots for both the Dynamic Tree Cut and Auto-merged algorithm are presented in Figure 3-C. Figure 3-D displays the network heatmap of all genes within these three modules identified through the Auto-merged algorithm. Figure 3 confirms that the 122 genes selected through statistical models exhibit a similar nature within the lung cancer and CLDs datasets.

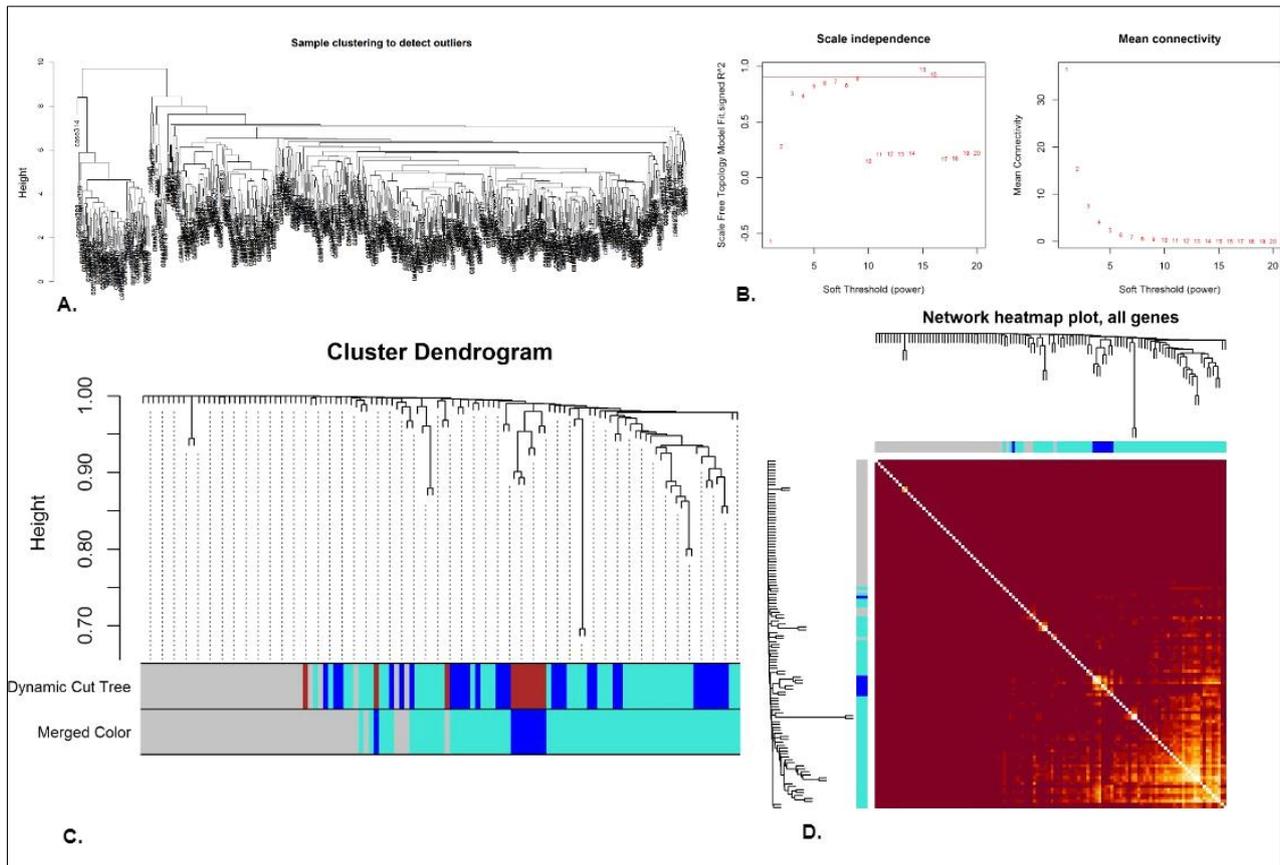


Figure 3. Correlation analysis using WGCNA for the common significant genes in LC and CLDs. This analysis aimed to understand the genetic interplay between lung cancer (LC) and common lung diseases (CLDs). Key findings include the identification of 122 common differentially expressed genes (DEGs) through WGCNA analysis, revealing robust clustering patterns and the absence of outlier samples (3-A). Optimal soft thresholding power ($=9$) was determined for module identification based on scale-free topology criteria (3-B), resulting in the identification of four modules containing distinct gene sets (blue, brown, grey, and turquoise) through both Dynamic Tree Cut and Auto-merged algorithms (3-C). A network heatmap (3-D) confirms the similarity in nature of the identified genes within LC and CLDs datasets.

3-1-3- PPI Network Construction and Hub Genes Identification

We identified 122 significant common DEGs among CLDs and LC. The STRING database [27] was used to obtain the PPI (shown in Figure 4-A) of the overlapped 122 DEGs of LC with CLDs. Next, the PPI network was displayed using Cytoscape and Cyto-Hubba plug-in was used to identify the hub proteins. To enhance the dependability of the hub genes, we combined the outcomes of four algorithms (Degree (Figure 4-B), EPC (Figure 4-C), MCC (Figure 4-D), and MNC (Figure 4-E) in our analysis. For each algorithm the top 20 proteins were chosen (see Table 2), after that we identified shared 6 hub proteins (SPTBN1, KCNA4, SCN7A, KCNQ3, GRIA1, and SDC1).

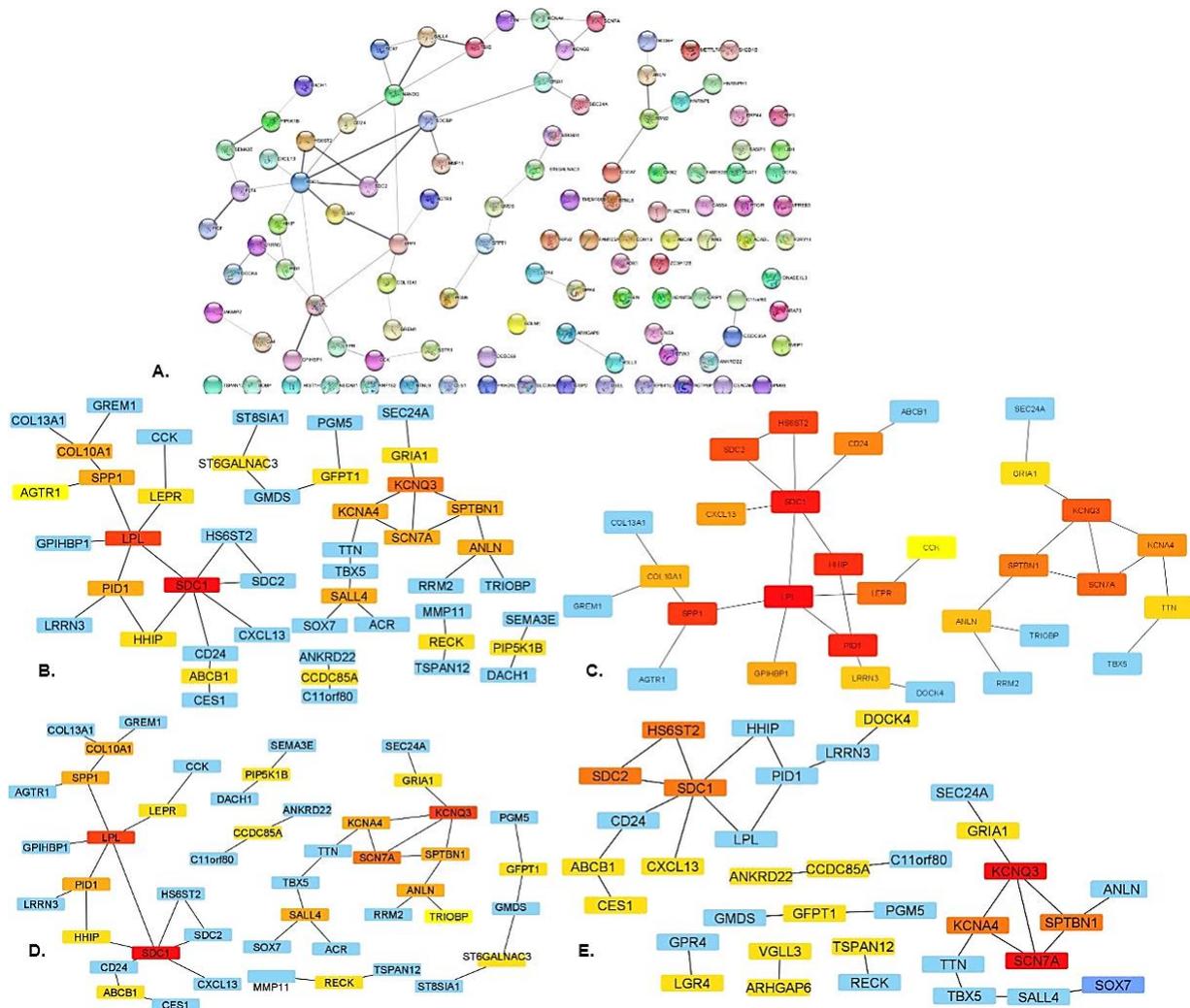


Figure 4. PPI Network for Shared Significant genes of LC and Significant lung diseases. **A.** PPI network of significant genes **B.** 20 hub proteins identification using Degree, EPC, MCC and MNC algorithms (B to E, respectively)

Table 2. Top 20 genes of four algorithms (Degree, EPC, MCC, and MNC)

MCC	MNC	DEGREE	EPC
RECK	SDC1	KCNA4	CXCL13
Gray C11orf80	VGLL3	SCN7A	SDC1
CCDC85A	ARHGAP6	KCNQ3	SDC2
Gray ANLN	TSPAN12	SPTBN1	HS6ST2
SSL4	LGR4	GRIA1	CD24
Gray SPTBN1	KCNA4	ANLN	HHIP
KCNA4	SCN7A	ST6GALNAC3	PID1
Gray SCN7A	KCNQ3	GFPT1	LPL
KCNQ3	SPTBN1	SALL4	LEPR
Gray TBX5	GRIA1	ABCB1	SPP1
COL10A1	DOCK4	COL10A1	COL10A1
Gray SPP1	SDC2	SPP1	GPISPBI
LEPR	HS6ST2	LEPR	LRRN3
Gray LPL	CXCL13	LPL	ANLN
PID1	ABCB1	PID1	SPTBN1
Gray ABCB1	CES1	SDC1	KCNA4
PIP5K1B	ANKRD22	HHIP	SCN7A
Gray GRIA1	GFPT1	PIP5K1B	KCNQ3
ST6GALNAC3	CCDC85A	RECK	TTN
Gray GFPT1		CCDC85A	GRIA1

3-1-4- Protein Expression Levels of Hub Genes in HPA Database

The HPA database did not contain any immunohistochemical information regarding KCNQ3. Figure 5 shows the protein expression levels of several hub genes (SPTBN1, KCNA4, SDC1, GRIA1, and SCN7A) obtained from HPA database. It is interesting to note that the SDC1 gene was not detected in normal lung tissues but was highly expressed in lung cancer tissues. On the other hand, SCN7A showed low expression levels in normal lung tissues as well as in lung cancer tissues. Interestingly, GRIA1 was not detected in both normal lung tissues and lung cancer tissues. However, in the case of the SPTBN1 gene, high expression was found in normal lung tissues and medium expression in lung cancer tissues. KCNA4 gene expression was not detected in normal lung tissues and medium expression in lung cancer tissues.

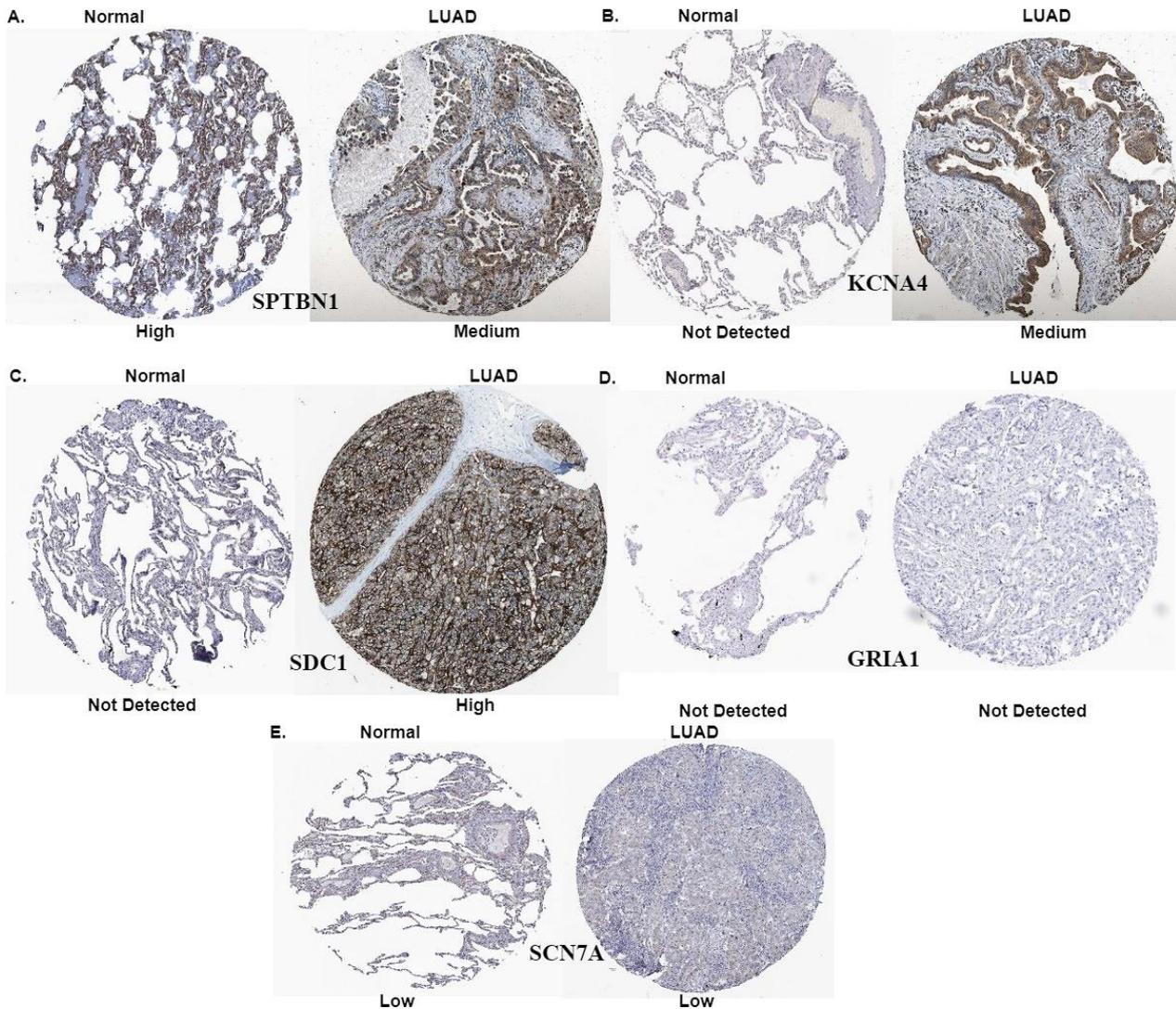


Figure 5. Immunohistochemistry images of hub genes in normal lung tissues and LUAD tissues were obtained from the HPA database. The images included protein representations of SPTBN1, KCNA4, SDC1, GRIA1, and SCN7A (labelled as A-E), with HPA standing for Human Protein Atlas.

These findings may have implications for understanding the molecular mechanisms underlying lung cancer and identifying potential targets for therapeutic interventions. However, it is important to note that these results are based on immunohistochemical data from the HPA database and may not necessarily reflect the protein expression levels in all cases of lung cancer. Further studies are needed to confirm these findings and determine their clinical significance.

3-1-5- Survival Prediction of the Significant Genes that were Common in LC and Commonly Lung Diseases (CLDs)

The analysis is crucial for understanding the genetic and clinical factors influencing patient survival in LC prognosis. The study obtained RNA-seq data and clinical information related to LC from the cBioPortal. This comprehensive dataset allows for a focused exploration of the impact of these factors on patient outcomes using machine learning techniques, particularly in the context of survival prediction for the significant genes common to LC and CLDs, thus potentially identifying novel genes with prognostic relevance. We identified 122 significant shared genes of LC with CLDs. We obtained 510 patients with 122 shared genes' mRNA values and three clinical variable values, described in the data section. The datasets included information from 566 cases, with 57 different features. Specifically, 510 cases

had gene expression data, which consisted of 20,440 gene expression values. After comparing 122 genes commonly associated with LC and 5 CLDs, only 122 genes had mRNA data available. Out of the 57 clinical features, only three were selected for analysis. We then integrated the clinical and RNA-seq data using patient IDs and utilized machine learning techniques to examine how these factors influenced patient survival in LC prognosis.

We used the product limit estimator to construct survival curves for 122 genes. The survival curves were used to compare the survival patterns between the two groups, which were altered and normal (non-altered). Genes that showed a significant difference in their survival patterns between the two groups were included in the analysis. The significance of a gene's role was determined by its p-value, which indicated the difference in survival pattern based on its expression level in the two categories. The analysis identified 27 genes (CASS4, EPB41L5, PKHD1L1, GNAZ, KCNA4, ANLN, LRRN3, TBX5, GRIA1, ST8SIA1, ADM2, CCK, SCN7A, P2RY14, TMEM106B, DNASE1L3, ADAMTS8, VEGFD, LBH, KCNQ3, LGR4, GPIHBP1, C1ORF21, PHACTR1, AFF3, and LEPR) that had a significant impact on patient survival. These genes had a lower p-value and their altered expression was associated with a lower likelihood of survival compared to the non-altered group, as demonstrated in Figures 6 and 7. The figures or graphs showed that the blue line represented altered gene expression, while the red line represented normal/unaltered gene expression. In Figure 6, ADAMTS8 showed the highest hazard ratio (HR) of 2.97, with a confidence interval (CI) of 1.1 to 8.02, indicating a significantly increased risk for individuals with this gene alteration within the analyzed set (ADM2, ADAMTS8, AFF3, ANLN, C1ORF21, CASS4, CCK, DNASE1L3, EPB41L5, GNAZ, and GPIHBP1). In contrast, both CCK and ADM2 had CIs entirely below 1, suggesting these genes may have a protective effect on survival. These results highlight ADAMTS8 as a strong prognostic marker and CCK and ADM2 as possible protective factors. For Figure 7, GRIA1 exhibited the highest hazard ratio (HR) of 4.5 with a confidence interval (CI) ranging from 1.44 to 14.08, underscoring its strong association with increased risk among the studied genes (GRIA1, KCNA4, KCNQ3, LBBH, LEPR, LGR4, LRRN3, PHACTR1, PKHD1L1, P2RY14, SCN7A, ST8SIA1, TBX5, TMEM106B, and VEGFD). In contrast, KCNQ3 showed a lower HR of 1.37, with a CI of 1.01 to 1.84, suggesting a modestly increased risk but statistically significant association with survival.

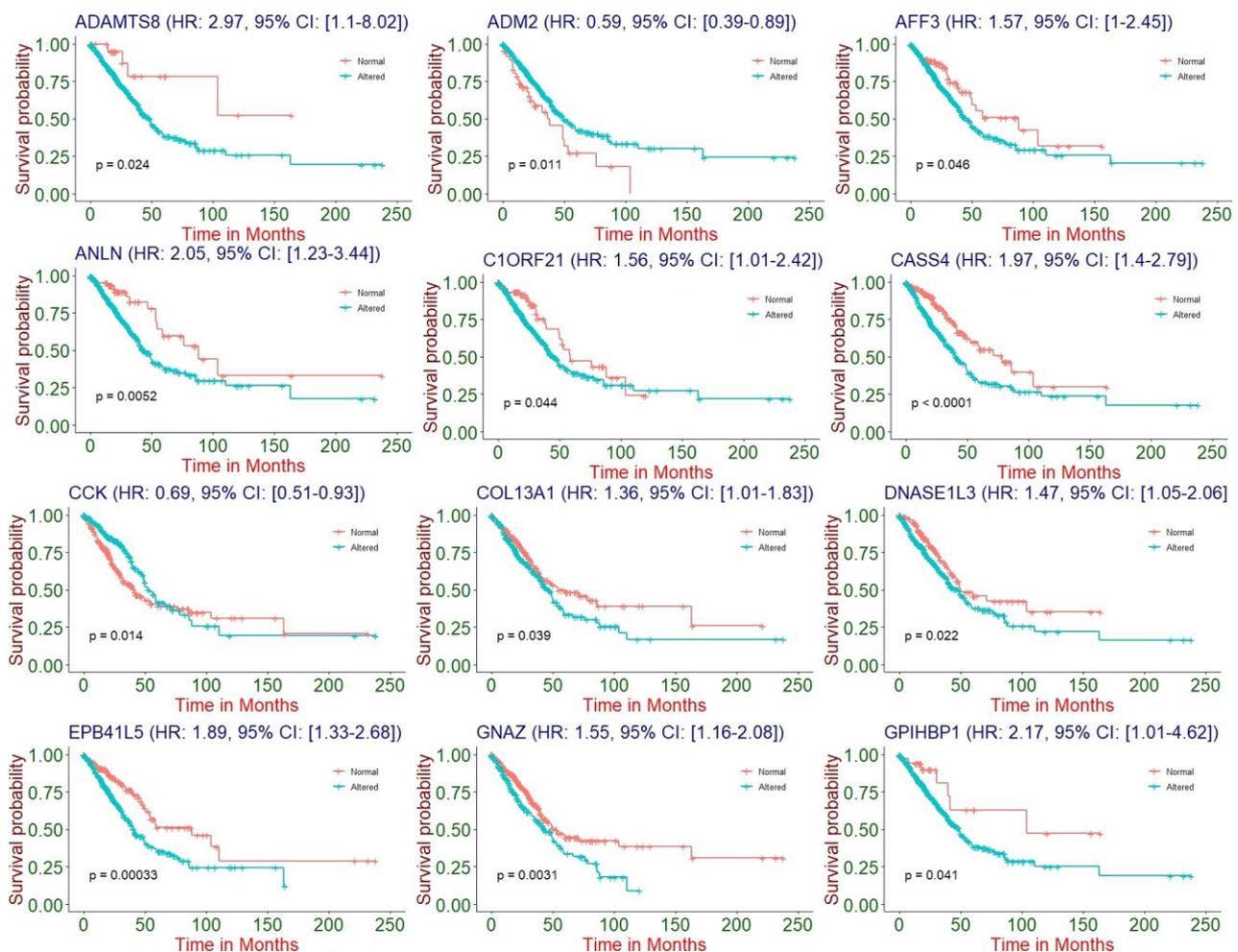


Figure 6. Survival curves of significant genes (ADM2, ADAMTS8, AFF3, ANLN, C1ORF21, CASS4, CCK, DNASE1L3, EPB41L5, GNAZ, and GPIHBP1) illustrating overall survival probabilities based on gene alteration status. Hazard ratios (HR) with 95% confidence intervals (CI) are shown for each gene, indicating the relative risk associated with altered expression. The inclusion of HR and CI values highlights the prognostic importance of these genes and their potential roles in influencing survival outcomes in disease progression.

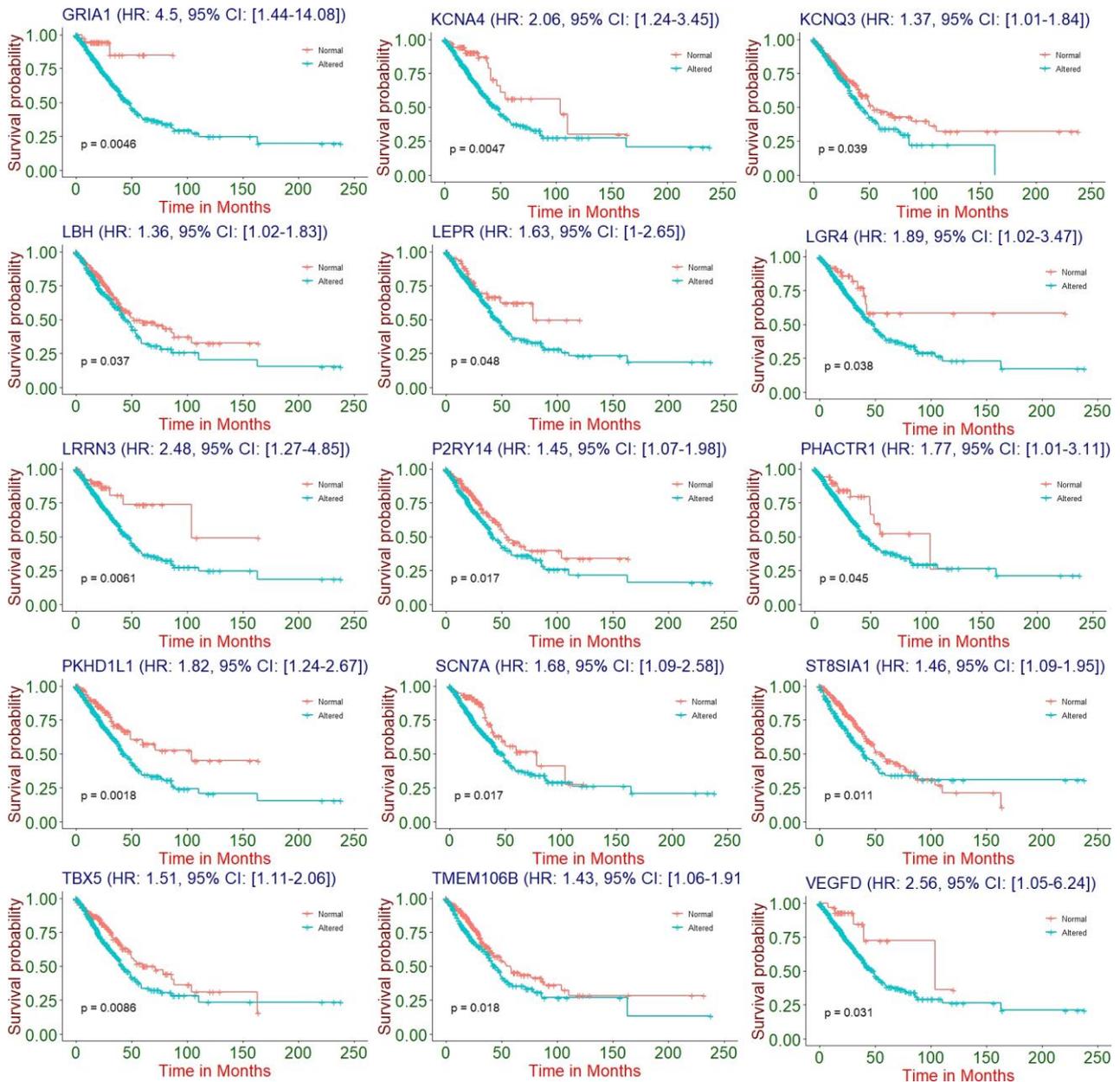


Figure 7. Survival curves for significant genes (GRIA1, KCNA4, KCNQ3, LBBH, LEPR, LGR4, LRRN3, PHACTR1, PKHD1L1, P2RY14, SCN7A, ST8SIA1, TBX5, TMEM106B, and VEGFD), illustrating overall survival based on gene alteration status. Hazard ratios (HR) and 95% confidence intervals (CI) are provided for each gene, showing the relative risk associated with gene expression changes. These insights into survival probability underscore the prognostic relevance of these genes in disease progression.

3-1-6- Modeling the Hazard Risk on the mRNA Data

To determine which genes are most significant for patient survival in LC, we utilized the Cox proportional hazards (PH) regression model to measure the relative likelihood of the risk of death for each gene separately (Univariate analysis) and simultaneously for all genes (multivariate analysis). Both univariate and multivariate analyses were conducted for the 122 genes. Tables 3 and 4 display only the genes with a p-value less than 0.05, along with their corresponding estimated coefficient (β), hazard ratio (HR), and p-value. From the univariate analysis, we identified 26 genes (CASS4, EPB41L5, PKHD1L1, GNAZ, KCNA4, ANLN, LRRN3, TBX5, GRIA1, ST8SIA1, ADM2, CCK, SCN7A, P2RY14, TMEM106B, DNASE1L3, ADAMTS8, VEGFD, LBH, KCNQ3, LGR4, GPIHBP1, C1ORF21, PHACTR1, AFF3, and LEPR) with a p-value less than 0.05 (see Table 3). Additionally, in the multivariate analysis, we found 14 significant genes (VGLL3, ADM2, GNAZ, RNF182, CCK, CASS4, HHIP, ST8SIA1, CDH13, LEPR, ANLN, EPB41L5, GMDS, GRIA1, and PKHD1L1) (see Table 4). Among these, only nine genes (CASS4, EPB41L5, PKHD1L1, GNAZ, ANLN, GRIA1, ST8SIA1, ADM2, CCK, and LEPR) were significant in both univariate and multivariate analyses.

Table 3. Significant genes associated with lung cancer are identified while assessing the relative likelihood of death risk for each gene separately, considering their individual associations with the outcome variable. Statistical metrics (coef, Z-Score, P-values) quantify gene-level associations with patient survival in univariate analysis.

Gene Name	coef	exp.coef	P-Values
CASS4	0.68	1.97	1.15E-04
EPB41L5	0.63	1.89	4.13E-04
Gray PKHD1L1	0.60	1.82	2.15E-03
GNAZ	0.44	1.55	3.39E-03
Gray KCNA4	0.73	2.06	5.68E-03
ANLN	0.72	2.05	6.26E-03
Gray LRRN3	0.91	2.48	8.05E-03
TBX5	0.41	1.51	9.05E-03
Gray GRIA1	1.50	4.50	9.85E-03
ST8SIA1	0.38	1.46	1.15E-02
Gray ADM2	-0.54	0.59	1.16E-02
CCK	-0.38	0.69	1.44E-02
Gray SCN7A	0.52	1.68	1.79E-02
P2RY14	0.37	1.45	1.81E-02
Gray TMEM106B	0.35	1.43	1.85E-02
DNASE1L3	0.39	1.47	2.31E-02
Gray ADAMTS8	1.09	2.97	3.16E-02
VEGFD	0.94	2.56	3.79E-02
Gray LBH	0.31	1.36	3.81E-02
KCNQ3	0.31	1.37	4.02E-02
Gray LGR4	0.63	1.89	4.15E-02
GPIHBP1	0.77	2.17	4.57E-02
Gray C1ORF21	0.45	1.56	4.62E-02
PHACTR1	0.57	1.77	4.76E-02
Gray AFF3	0.45	1.57	4.84E-02
LEPR	0.49	1.63	5.00E-02

Table 4. Significant genes in Multivariate analysis. In multivariate analysis, significant genes associated with lung cancer are identified by simultaneously considering multiple variables, allowing for a comprehensive assessment of their combined impact on the risk of death, thus providing a thorough understanding of genetic factors influencing patient survival. Statistical metrics (coef, Z-Score, P-values) quantify gene-level associations with patient survival in multivariate analysis.

Gene Name	coef	Z-Score	P-Values
Gray VGLL3	-1.17	-3.63	2.82E-04
ADM2	-1.11	-3.14	1.72E-03
Gray GNAZ	0.77	3.12	1.81E-03
RNF182	-1.04	-3.01	2.61E-03
Gray CCK	-0.61	-2.90	3.72E-03
CASS4	0.89	2.87	4.06E-03
Gray HHIP	-0.92	-2.77	5.62E-03
ST8SIA1	0.59	2.50	1.24E-02
Gray CDH13	0.80	2.44	1.49E-02
LEPR	0.88	2.34	1.93E-02
Gray ANLN	0.96	2.19	2.87E-02
EPB41L5	0.57	2.17	2.99E-02
Gray GMDS	-0.47	-2.04	4.17E-02
GRIA1	1.68	2.01	4.47E-02
Gray PKHD1L1	0.67	2.00	4.54E-02

3-1-7- Performance Evaluation and Deployment of Significant Hub Genes

We evaluated the effectiveness of hub genes in classifying different disease states using machine learning techniques. We measured the performance using two evaluation metrics: accuracy and area under the ROC curve (AUC). The analysis was conducted on two datasets: one containing 122 shared genes between LC and five CLDs and another comprising significant genes obtained from hub genes and those identified in univariate and multivariate analyses (see Table 5). mRNA expression data were normalized using the FPKM method, and gene values for LC samples and normal samples were used for model training and testing.

Table 5. Performance evaluation of the common genes and significant genes

Accuracy	33 Significant genes		122 common genes	
	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
Bayesian Network	1.000	1.000	1.000	1.000
Logistic Regression	1.000	0.995	1.000	0.993
Random Forest	1.000	1.00	1.000	0.993
SVM (Linear)	1.000	0.997	1.000	0.993

Using a 10-fold cross-validation strategy, the results demonstrated high accuracy and effectiveness for the classification task. All machine learning models, including SVM, achieved excellent performance, as reflected in both train and test accuracies (see Table 5). The SVM algorithm displayed robust classification capabilities for both datasets, with high AUC values. The ROC curve (Figure 8) further illustrates the model's ability to distinguish LC samples from normal samples effectively. These results validate the reliability of the shared and significant genes in predicting LC progression. Six key hub genes (SPTBN1, KCNA4, SCN7A, KCNQ3, GRIA1, and SDC1) were validated using the AUC values computed from ROC curves, utilizing the TCGA lung cancer dataset (see Figure 9). Figure 9-a. displays the ROC curve of six key candidate genes along with their corresponding AUC values for the TCGA lung cancer dataset: KCNA4 (AUC: 0.94, 95% CI: 0.935 to 0.985), GRIA1 (AUC: 0.99, 95% CI: 0.963 to 0.986), SCN7A (AUC: 0.98, 95% CI: 0.922 to 0.969), KCNQ3 (AUC: 0.82, 95% CI: 0.416 to 0.738), SPTBN1 (AUC: 0.98, 95% CI: 0.938 to 0.980), and SDC1 (AUC: 0.86, 95% CI: 0.708 to 0.809). Figure 9-b. Shows the heatmap of hub genes. Finally, the SVM model was deployed (see Figure 10) for real-world applications, enabling users to input gene expression profiles and receive immediate predictions of disease states. This deployment underscores the potential of machine-learning approaches to aid in clinical diagnostics and personalized treatment planning for lung cancer. The deployment of the machine learning model for lung cancer prediction using significant hub genes is available on GitHub for reproducibility and further analysis. Access it here: *Lung Cancer Prediction* (<https://github.com/Ali-bd/Lung-Cancer-Prediction.git>).

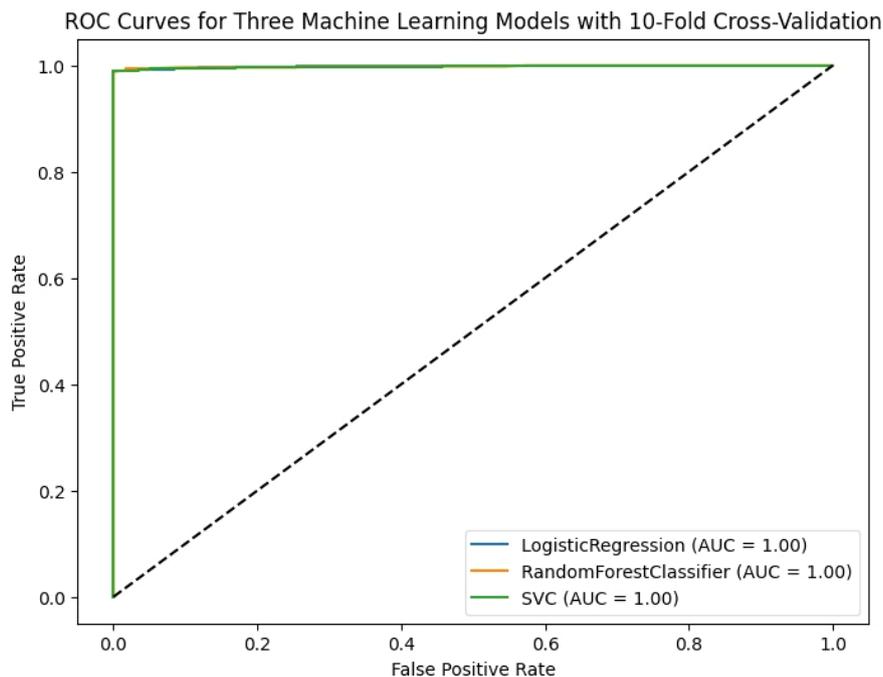


Figure 8. ROC curve of Significant genes. ROC curves of significant genes were generated to assess the effectiveness of hub genes in predicting or detecting lung cancer from shared genes

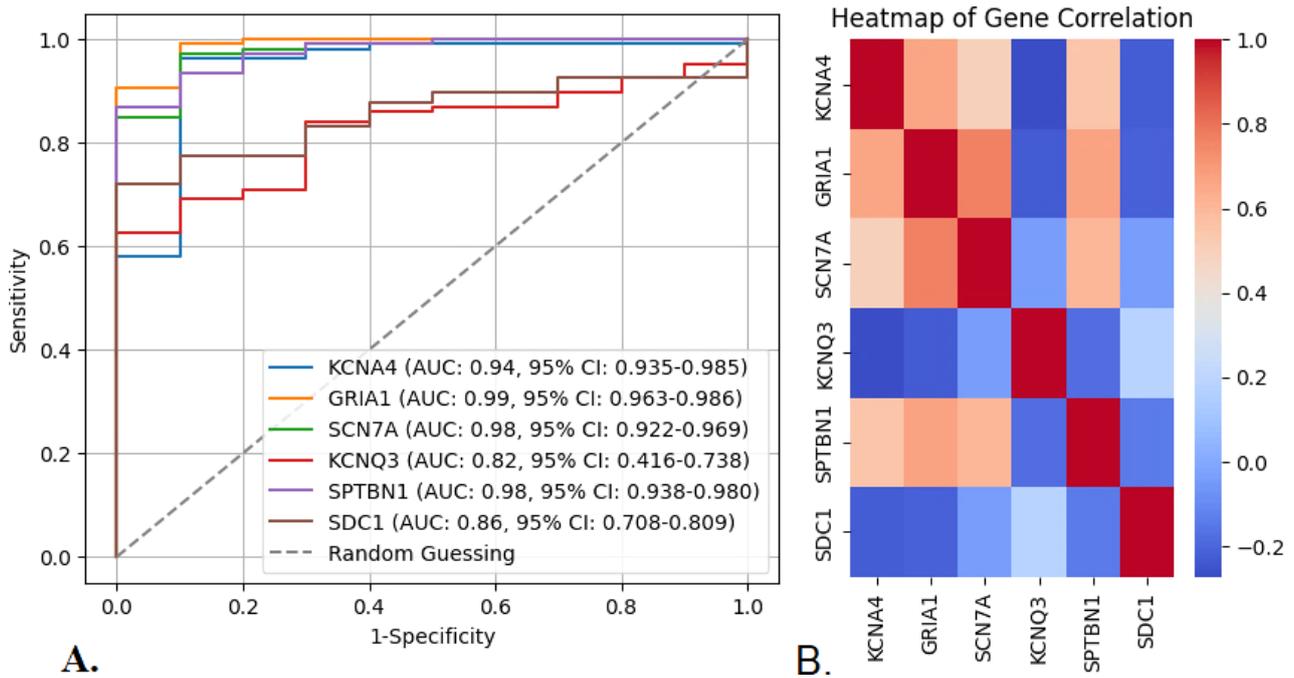


Figure 9. ROC curve of Heatmap of 6 hub genes. ROC curves of significant genes were generated to assess the effectiveness of hub genes in predicting or detecting lung cancer from shared genes and Heatmap to see the correlation of these genes.

Lung Cancer Prediction

Select Input Method:

Manually Enter Values

Upload PDF

KCNA4:

GRIA1:

SCN7A:

KCNQ3:

SPTBN1:

SDC1:

Submit

Prediction Result

FEATURE	VALUE
KCNA4	0.02358
GRIA1	0.021605
SCN7A	1.35792
KCNQ3	2.022098
SPTBN1	4.417103
SDC1	6.955016

Prediction: Positive

Figure 10. Deployment of the machine learning model using SVM for lung cancer prediction. The system allows the input of normalized gene expression profiles (FPKM values) and provides predictions on disease states (LC or normal) based on the trained model.

3-1-8- Pathway and Functional Correlation Analysis

Considering 122 significant shared DEGs of LC with CLDs, we performed pathway analysis and gene ontology enrichment analysis using the Enrichr (KEGG pathway) database and biological process (<http://amp.pharm.mssm.edu/Enrichr/enrich>), respectively. We observed that 8 significant pathways were associated with the 122 significant DEGs shown in Figure 11-A. Significant pathways are vitamin B6 metabolism, neuroactive ligand-receptor interaction, drug metabolism, ABC transporters, glycosphingolipid biosynthesis, amino sugar and nucleotide sugar metabolism, malaria, and the cAMP signaling pathway. Genes associated with these pathways are presented in Figure 11-A. We also observed that 15 significant gene ontology groups (shown in Figure 11-B) are associated with the significant shared DEGs.

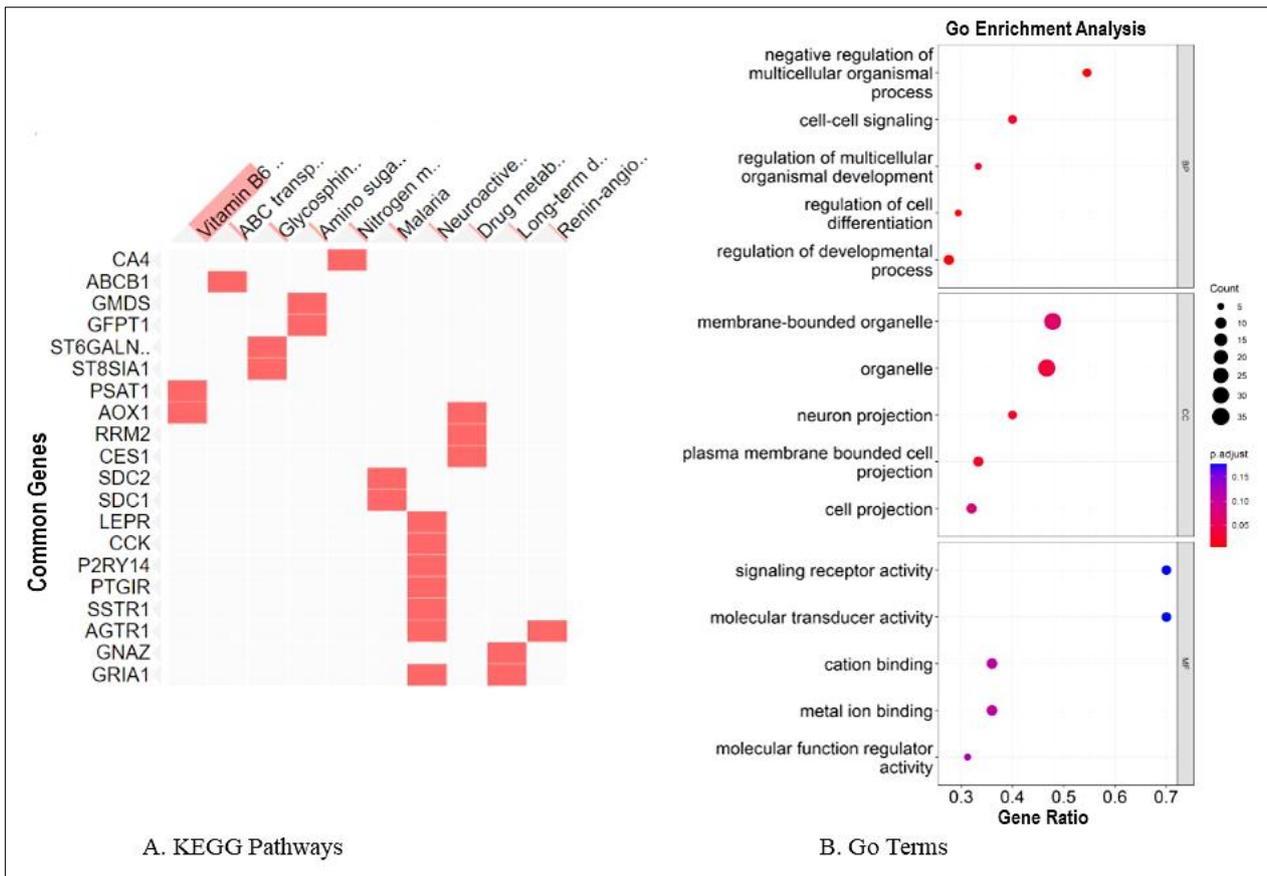


Figure 11. A. 8 Pathways associated with the significant Common DEGs between CLDs and LC. B. 15 Gene ontologies (BP: Biological Processes; MF: Molecular Functions; CC: Cellular Components) associated with the significant Common DEGs between CLD and LC.

3-2-Discussion

In this study, we have investigated lung comorbid diseases' influence on lung cancer development and progression through integrated network analysis of gene expression and pathways relevant to lung cancer and commonly found lung disorders. We used a variety of bioinformatics techniques to decode multi-relational infectome, disease comorbidity, and disease linkages between LC and 5 CLD disorders. LC has 36, 10, 17, 18, and 78 important shared genes with TB, asthma, pneumonia, COPD, and IPF, respectively, according to the infectome-disease network. The WGCNA analysis provided valuable insights into the regulatory networks and coexpression patterns of common genes associated with LC and its comorbidities. This analysis highlighted potential molecular mechanisms underlying disease progression and identified key gene modules involved in LC pathogenesis. Furthermore, the performance evaluation of common genes using classification algorithms underscores their potential as diagnostic and predictive biomarkers for LC. By assessing their performance across various algorithms, we gained a comprehensive understanding of their utility in clinical practice.

From the infectomedisease analysis, we identified that LC most resembles pulmonary infections such as IPF and TB. After analyzing the PPI network constructed around the common DEGs between lung disorders and LC, we identify 6 common hub proteins (SPTBN1, KCNA4, SCN7A, KCNQ3, GRIA1, and SDC1) using the four algorithms (degree, EPC, MCC, and MNC) in the Cytoscape plug-in and provide their potential significance in the pathogenesis of LC and its associated risk factors. The sharing of hub proteins between LC and other respiratory disorders, such as IPF, asthma, and TB, suggests common molecular mechanisms underlying these conditions. In particular, it was discovered that LC and IPF shared SCN7A, GRIA1, and SDC1; that asthma shared KCNA4; and that TB shared SPTBN1 and KCNQ3. Moreover, the outcomes of the univariate and multivariate studies shed new light on how these hub proteins function in the development of LC. While SCN7A, KCNA4, and KCNQ3 genes were significant only in the univariate analysis, GRIA1 was shown to be significant in both analyses. This suggests unique functions for SCN7A, KCNA4, and KCNQ3 in determining disease prognosis and underscores the potential significance of GRIA1 as a major regulator in LC development and progression. In the survival analysis, GRIA1 exhibited the highest risk with an HR of 4.5 (95% CI: 1.44 to 14.08), indicating a significant association with poorer survival. KCNA4 followed, showing an increased risk with an HR of 2.07 (95% CI: 1.24 to 3.45). KCNQ3 had a moderate effect on survival, with an HR of 1.37 (95% CI:

1.01 to 1.84). SCN7A presented an HR of 1.68 (95% CI: 1.09 to 2.58), also indicating increased risk. These genes highlight important prognostic markers, with potential applications in risk stratification and therapeutic targeting. The significance of these hub proteins' functional role and their potential as therapeutic targets and diagnostic markers are underscored by these findings. SPTBN1-ALK fusion was identified by Gu et al. as a putative LC biomarker [33]. In individuals with gliomas, KCNA4 methylation rises as the tumor grade progresses and is linked to a bad prognosis [34]. Additionally, it is increased in patients with gastric cancer [35]. Nevertheless, the relationship between channel expression and KCNA4 hypermethylation, as well as the impact on cancer cells, remains unclear [36].

Liu et al. reported SCN7A as a potential biomarker of LC [37]. Underexpression of SCN7A has been associated with poor prognosis in LC patients [38]. SCN7A may inhibit lung cancer cell proliferation and migration, and its expression correlates with immune cell infiltration and immune checkpoint expression [38]. GRIA1 is a gene that is strongly correlated with the prognosis of LC [39]. GRIA1, also known as glutamate ionotropic receptor AMPA type subunit 1, is a protein-coding gene that encodes a subunit of the AMPA receptor, which is involved in synaptic plasticity and neurotransmission. Studies have shown that overexpression of GRIA1 is associated with poor prognosis in LUAD patients [39]. Al-Dherasi et al. reported GRIA1 as a potential signature of LC [40]. The nuclear translocation of SDC-1 (Syndecan-1) can have different effects on different types of cells. In human B6FS fibrosarcoma cells, it was found to facilitate the elimination of mesenchymal and invasive characteristics, indicating that SDC-1 may act as a tumor suppressor [34, 41]. However, in human A549 lung cancer cells undergoing TGF-1-induced EMT (epithelial-mesenchymal transition), loss of nuclear SDC-1 was associated with cell elongation and a switch from E-cadherin to N-cadherin, which are markers of a more mesenchymal and invasive phenotype [41]. In lung cancer, syndecan-1 serves as a favorable indicator. This is because higher levels of syndecan-1 in lung cancer cells are linked with increased chances of survival [42, 43]. Moreover, as the histologic grade of lung cancer increases, the expression of syndecan-1 decreases [43]. Parimon et al. confirm that syndecan-1 expression could potentially serve as a predictive factor for the prognosis of lung cancer patients [44].

The association of SPTBN1 gene expression with lung cancer has been previously reported in the literature by Gu et al. [33] and Zhai et al. [45]. Our findings are consistent with these studies, which showed the upregulation of SPTBN1 expression in lung cancer tissues. Moreover, the prognostic value of SPTBN1 expression has also been demonstrated in other cancer types, including ovarian cancer, colorectal cancer, breast cancer, and gastric cancer [46-48]. Our study extends these findings to lung cancer and suggests that high expression of SPTBN1 may serve as a prognostic biomarker for lung cancer patients. In this study, several genes (CASS4, EPB41L5, PKHD1L1, GNAZ, ANLN, GRIA1, ST8SIA1, ADM2, CCK, and LEPR) were found to be significant in both univariate and multivariate analyses. Notably, CASS4, EPB41L5, PKHD1L1, ANLN, GRIA1, and CCK are shared with IPF, while ST8SIA1 and LEPR are linked to chronic obstructive pulmonary disease (COPD). Among these, only GRIA1 was shared among the six hub genes, underscoring its unique role as an independent prognostic marker within the network associated with lung cancer risk. Meanwhile, ADM2 and CCK, although significant in initial analyses, had confidence intervals below 1, suggesting they may not contribute to increased risk. Among them, the gene CASS4 overexpression promotes invasion in non-small cell lung cancer (NSCLC) by activating the AKT signaling pathway and inhibiting E-cadherin expression [49]. PKHD1L1 expression is significantly lower in lung adenocarcinoma compared to normal tissues, and its decreased expression is associated with unfavorable overall survival [50]. Overexpression of ANLN at both RNA and protein levels is associated with poor prognosis and metastasis in LUAD patients [51, 52]. The ST8SIA1 gene is highly expressed across multiple cancers, including lung cancer, where its role in promoting tumor progression and metastasis highlights it as a potential target for cancer therapies [53]. Tang et al. [54] suggested that LEPR polymorphisms may serve as biomarkers for both risk assessment and disease progression in NSCLC, as they have been associated with increased susceptibility and metastasis risk.

The pathway of interaction between neuroactive ligands and receptors in the brain is crucially involved in nicotine addiction, which significantly contributes to the development of lung cancer [55]. Additionally, the low expression levels of the enzyme pyridoxal kinase (PDXK), responsible for generating the bioactive form of vitamin B6, have been linked to poor disease outcomes in non-small cell lung cancer (NSCLC) patients. This finding underscores the critical role of vitamin B6 metabolism in sensitizing cancer cells to chemotherapy-induced apoptosis [56]. Furthermore, a recent study confirms the association between heightened vitamin B6 catabolism and an increased risk of lung cancer, attributed to inflammation and immune activation [57]. Similarly, the neuroactive ligand-receptor interaction pathway is connected to nicotine dependence, which remains a significant factor in the heightened risk of lung cancer development [55]. Moreover, the cAMP signaling pathway has demonstrated a strong correlation with lung cancer [58], while drug metabolism pathways also show a notable association with the disease [59]. In addition, glycosphingolipid (GSL) biosynthesis plays a pivotal role in lung cancer transformation and progression. Specific GSLs, such as -GalCer, NeuGcGM3, and GM2, have been identified as potential immunotherapy targets, influencing tumor growth, metastasis, and treatment resistance [60]. Finally, the dysregulated expression and activity of amino sugar and nucleotide sugar metabolism enzymes, including LAT1 (SLC7A5) and LAT2 (SLC7A8), as well as their associated regulatory factors, contribute to altered nutrient uptake and utilization in lung cancer. These findings highlight their significance as therapeutic targets [61].

Although this study provides valuable insights into the molecular mechanisms underlying LC and its associated risk factors, it has several limitations. First, the analysis relied on publicly available transcriptomic data, which may vary in quality and may not fully represent the complexity of the disease. Second, the lack of experimental validation, such as qPCR, of the identified genes limits the robustness of the findings. Additionally, the retrospective nature of the study and the use of clinical data from public databases may introduce bias. Furthermore, the study focused on the analysis of gene expression data related to LC and common lung diseases (CLDs), which may not encompass all relevant conditions contributing to LC development. Despite these limitations, our integrated approach sheds light on the intricate interplay between lung cancer and common lung diseases, highlighting shared genetic mechanisms and pathways. Moving forward, transcriptomic and pathway-based personalized medicine holds promise for enhancing our understanding of disease mechanisms and opening new avenues for diagnosis, therapy, and prevention of disease comorbidities.

4- Conclusion

In conclusion, our study provides valuable insights into the molecular mechanisms underlying LC and its associated risk factors, shedding light on the intricate interplay between lung cancer and common lung diseases. In our analysis, it is observed that LC and IPF shared the highest number of common DEGs between them. So, a patient of IPF has a high possibility of becoming a patient of LC. Moreover, significant hub genes (SPTBN1, KCNA4, SCN7A, KCNQ3, GRIA1, and SDC1) and significant pathways associated with LC and commonly found lung disorders suggest their (i.e., hub genes and pathways) critical roles in LC development. Furthermore, we conducted WGCNA analysis on the identified common genes, unraveling intricate coexpression patterns and potential regulatory networks associated with LC and its related risk factors. The diagnostic and predictive potential of these genes was further validated through classification algorithms, revealing their capability to accurately distinguish LC cases from controls. The deployment of a predictive model based on these six genes highlights their clinical relevance, offering a tool to assess LC risk and progression in patients. This model can facilitate early detection and improve prognostic evaluations, paving the way for personalized therapeutic strategies. Through rigorous univariate and multivariate analyses, we explored the individual and combined effects of these genes on LC prognosis, enhancing our understanding of their clinical relevance. Moreover, we investigated the protein expression levels of the identified hub genes in the HPA database, shedding light on their potential functional roles in LC progression. These findings underscore the importance of an integrated bioinformatics and machine learning approach in uncovering disease relationships and driving personalized therapeutic strategies. By bridging the gap between molecular mechanisms and clinical applications, this study lays the foundation for precision medicine, offering new opportunities for improved diagnosis, targeted therapy, and prevention of LC and its comorbidities.

5- Declarations

5-1- Author Contributions

Conceptualization, M.A.H. and M.A.M.; methodology, M.A.H.; software, M.A.H.; validation, M.A.H. and T.A.A.; formal analysis, M.A.H. and T.A.A.; investigation, M.A.H. and T.A.A.; resources, M.A.H. and T.A.A.; data curation, M.A.H.; writing—original draft preparation, M.A.H. and T.A.A.; writing—review and editing, M.A.H., T.A.A., Z.M., A.Z., M.Z.M., M.A.M., and A.M.; visualization, M.A.H. and T.A.A.; supervision, M.Z.M., M.A.M., and A.M.; project administration, M.A.H., M.Z.M., M.A.M., and A.M.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

5-2- Data Availability Statement

We obtained our datasets from NCBI GEO and TCGA. Processed and analyzed data from our study will be readily available for any inquiries. We utilized microarray datasets of Tuberculosis, Pneumoniae, Asthma, Idiopathic Pulmonary Fibrosis, and COPD with the accession numbers GSE62525, GSE35716, GSE43696, GSE24206, and GSE76925, respectively. Additionally, we utilized two microarray datasets of LC with the accession numbers GSE89039 and GSE136043. We utilized mRNAseq data of Lung Adenocarcinoma (LC) from The Cancer Genome Atlas (TCGA) through the TCGA genome data analysis center (<http://gdac.broadinstitute.org/>). The source code for the machine learning model used in this study is available on GitHub: Lung Cancer Prediction.

5-3- Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

5-4- Institutional Review Board Statement

Not applicable.

5-5- Informed Consent Statement

Not applicable.

5-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

6- References

- [1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6), 394–424. doi:10.3322/caac.21492.
- [2] Lemjabbar-Alaoui, H., Hassan, O. U., Yang, Y.-W., & Buchanan, P. (2015). Lung cancer: Biology and treatment options. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1856(2), 189–210. doi:10.1016/j.bbcan.2015.08.002.
- [3] Gridelli, C., Rossi, A., Carbone, D. P., Guarize, J., Karachaliou, N., Mok, T., Petrella, F., Spaggiari, L., & Rosell, R. (2015). Non-small-cell lung cancer. *Nature Reviews Disease Primers*, 1(1). doi:10.1038/nrdp.2015.9.
- [4] Sher, T., Dy, G. K., & Adjei, A. A. (2008). Small Cell Lung Cancer. *Mayo Clinic Proceedings*, 83(3), 355–367. doi:10.4065/83.3.355.
- [5] Navarro-Torné, A., Vidal, M., Trzaska, D. K., Passante, L., Crisafulli, A., Laang, H., van de Loo, J.-W., Berkouk, K., & Draghia-Akli, R. (2015). Chronic respiratory diseases and lung cancer research: a perspective from the European Union. *European Respiratory Journal*, 46(5), 1270–1280. doi:10.1183/13993003.00395-2015.
- [6] Sekine, Y., Katsura, H., Koh, E., Hiroshima, K., & Fujisawa, T. (2012). Early detection of COPD is important for lung cancer surveillance. *European Respiratory Journal*, 39(5), 1230–1240. doi:10.1183/09031936.00126011.
- [7] Young, A.Y., Shannon, V.R. (2020). Acute Respiratory Distress Syndrome in Cancer Patients. *Oncologic Critical Care*. Springer, Cham, Switzerland. doi:10.1007/978-3-319-74588-6_48.
- [8] Zhan, P., Suo, L. J., Qian, Q., Shen, X. K., Qiu, L. X., Yu, L. K., & Song, Y. (2011). Chlamydia pneumoniae infection and lung cancer risk: A meta-analysis. *European Journal of Cancer*, 47(5), 742–747. doi:10.1016/j.ejca.2010.11.003.
- [9] Xu, X., Liu, Z., Xiong, W., Qiu, M., Kang, S., Xu, Q., Cai, L., & He, F. (2020). Combined and interaction effect of chlamydia pneumoniae infection and smoking on lung cancer: A case-control study in Southeast China. *BMC Cancer*, 20(1), 1–10. doi:10.1186/s12885-020-07418-8.
- [10] Zhang, K., Qi, S., Cai, J., Zhao, D., Yu, T., Yue, Y., Yao, Y., & Qian, W. (2022). Content-based image retrieval with a Convolutional Siamese Neural Network: Distinguishing lung cancer and tuberculosis in CT images. *Computers in Biology and Medicine*, 140, 105096. doi:10.1016/j.combiomed.2021.105096.
- [11] Qu, Y. L., Liu, J., Zhang, L. X., Wu, C. M., Chu, A. J., Wen, B. L., Ma, C., Yan, X. yan, Zhang, X., Wang, D. M., Lv, X., & Hou, S. J. (2017). Asthma and the risk of lung cancer: A meta-analysis. *Oncotarget*, 8(7), 11614–11620. doi:10.18632/oncotarget.14595.
- [12] Wang, Y., Liu, Y., Zhang, Q., Gao, R., & Wang, K. (2017). IPF and lung cancer: Homologous but different endings, the progress in the correlation research. *International Journal of Clinical and Experimental Medicine*, 10(3), 4319–4329.
- [13] Antoniou, K. M., Tomassetti, S., Tsitoura, E., & Vancheri, C. (2015). Idiopathic pulmonary fibrosis and lung cancer. *Current Opinion in Pulmonary Medicine*, 21(6), 626–633. doi:10.1097/mcp.0000000000000217.
- [14] Qubo, A. A., Numan, J., Snijder, J., Padilla, M., Austin, J. H. M., Capaccione, K. M., Pernia, M., Bustamante, J., O’connor, T., & Salvatore, M. M. (2022). Idiopathic pulmonary fibrosis and lung cancer: future directions and challenges. *Breathe*, 18(4). doi:10.1183/20734735.0147-2022.
- [15] Ang, L., Ghosh, P., & Seow, W. J. (2021). Association between previous lung diseases and lung cancer risk: a systematic review and meta-analysis. *Carcinogenesis*, 42(12), 1461–1474. doi:10.1093/carcin/bgab082.
- [16] Li, J., Zhang, J. T., Jiang, X., Shi, X., Shen, J., Feng, F., Chen, J., Liu, G., He, P., Jiang, J., Tsang, L. L., Wang, Y., Rosell, R., Jiang, L., He, J., & Chan, H. C. (2015). The cystic fibrosis transmembrane conductance regulator as a biomarker in non-small cell lung cancer. *International Journal of Oncology*, 46(5), 2107–2115. doi:10.3892/ijo.2015.2921.
- [17] Miron, O., Afrasanie, V. A., Paduraru, M. I., Trandafir, L. M., & Miron, L. (2020). The relationship between chronic lung diseases and lung cancer-a narrative review. *J BUON*, 25, 1687-92.
- [18] Otálora-Otálora, B. A., Florez, M., López-Kleine, L., Canas Arboleda, A., Grajales Urrego, D. M., & Rojas, A. (2019). Joint Transcriptomic Analysis of Lung Cancer and Other Lung Diseases. *Frontiers in Genetics*, 10, 1260. doi:10.3389/fgene.2019.01260.

- [19] Yao, Y., Li, Z., & Gao, W. (2022). Identification of Hub Genes in Idiopathic Pulmonary Fibrosis and NSCLC Progression: Evidence From Bioinformatics Analysis. *Frontiers in Genetics*, 13, 855789. doi:10.3389/fgene.2022.855789.
- [20] Dasgupta, S. (2024). Identification of Hub Genes in Interstitial Lung Disease and their Association With Lung Cancer: An In-Silico Analysis. *ICCECE 2024 - International Conference on Computer, Electrical and Communication Engineering*, 1–8. doi:10.1109/ICCECE58645.2024.10497218.
- [21] Hossain, M. A., Rahman, M. Z., Bhuiyan, T., & Moni, M. A. (2024). Identification of Biomarkers and Molecular Pathways Implicated in Smoking and COVID-19 Associated Lung Cancer Using Bioinformatics and Machine Learning Approaches. *International Journal of Environmental Research and Public Health*, 21(11), 1392. doi:10.3390/ijerph21111392.
- [22] Ding, X., Liu, H., Xu, Q., Ji, T., Chen, R., Liu, Z., & Dai, J. (2024). Shared biomarkers and mechanisms in idiopathic pulmonary fibrosis and non-small cell lung cancer. *International Immunopharmacology*, 134, 112162. doi:10.1016/j.intimp.2024.112162.
- [23] Dasgupta, S. (2024). Thinking Beyond Disease Silos: Dysregulated Genes Common in Tuberculosis and Lung Cancer as Identified by Systems Biology and Machine Learning. *OMICS A Journal of Integrative Biology*, 28(7), 347–356. doi:10.1089/omi.2024.0116.
- [24] Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 1–13. doi:10.1186/1471-2105-9-559.
- [25] Ali Hossain, M., Asa, T. A., Rabiul Auwul, M., Aktaruzzaman, M., Mahfizur Rahman, M., Rahman, M. Z., & Moni, M. A. (2023). The pathogenetic influence of smoking on SARS-CoV-2 infection: Integrative transcriptome and regulomics analysis of lung epithelial cells. *Computers in Biology and Medicine*, 159, 106885. doi:10.1016/j.combiomed.2023.106885.
- [26] Hossain, Md. A., Saiful Islam, S. M., Quinn, J. M. W., Huq, F., & Moni, M. A. (2019). Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. *Journal of Biomedical Informatics*, 100, 103313. doi:10.1016/j.jbi.2019.103313.
- [27] Hossain, M. A., Asa, T. A., Rahman, M. M., Uddin, S., Moustafa, A. A., Quinn, J. M. W., & Moni, M. A. (2020). Network-based genetic profiling reveals cellular pathway differences between follicular thyroid carcinoma and follicular thyroid adenoma. *International Journal of Environmental Research and Public Health*, 17(4), 1373. doi:10.3390/ijerph17041373.
- [28] Xu, H., Moni, M. A., & Liò, P. (2015). Network regularised Cox regression and multiplex network models to predict disease comorbidities and survival of cancer. *Computational Biology and Chemistry*, 59, 15–31. doi:10.1016/j.compbiolchem.2015.08.010.
- [29] Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Von Mering, C. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1), D607–D613. doi:10.1093/nar/gky1131.
- [30] Chen, S. H., Chin, C. H., Wu, H. H., Ho, C. W., Ko, M. T., & Lin, C. Y. (2009). cyto-Hubba: A Cytoscape plug-in for hub object analysis in network biology. 20th international conference on genome informatics. 14-16, December, 2009, Yokohama, Japan.
- [31] Ma, Z., Xu, J., Ru, L., & Zhu, W. (2021). Identification of pivotal genes associated with the prognosis of gastric carcinoma through integrated analysis. *Bioscience Reports*, 41(4). doi:10.1042/BSR20203676.
- [32] Wei, J., Wang, Y., Shi, K., & Wang, Y. (2020). Identification of Core Prognosis- Related Candidate Genes in Cervical Cancer via Integrated Bioinformatical Analysis. *BioMed Research International*, 8959210. doi:10.1155/2020/8959210.
- [33] Gu, F. F., Zhang, Y., Liu, Y. Y., Hong, X. H., Liang, J. Y., Tong, F., Yang, J. S., & Liu, L. (2016). Lung adenocarcinoma harboring concomitant SPTBN1-ALK fusion, c-Met overexpression, and HER-2 amplification with inherent resistance to crizotinib, chemotherapy, and radiotherapy. *Journal of Hematology & Oncology*, 9(1), 1–3. doi:10.1186/s13045-016-0296-8.
- [34] Guo, S., Wu, X. Y., Lei, T., Zhong, R., Wang, Y. R., Zhang, L., Zhao, Q. Y., Huang, Y., Shi, Y., & Wu, L. (2022). The Role and Therapeutic Value of Syndecan-1 in Cancer Metastasis and Drug Resistance. *Frontiers in Cell and Developmental Biology*, 9, 3663. doi:10.3389/fcell.2021.784983.
- [35] Zheng, Y., Chen, L., Li, J., Yu, B., Su, L., Chen, X., Yu, Y., Yan, M., Liu, B., & Zhu, Z. (2011). Hypermethylated DNA as potential biomarkers for gastric cancer diagnosis. *Clinical Biochemistry*, 44(17–18), 1405–1411. doi:10.1016/j.clinbiochem.2011.09.006.
- [36] Angi, B., Muccioli, S., Szabò, I., & Leanza, L. (2023). A Meta-Analysis Study to Infer Voltage-Gated K⁺ Channels Prognostic Value in Different Cancer Types. *Antioxidants*, 12(3), 573. doi:10.3390/antiox12030573.
- [37] Liu, Y., Li, X., Chang, R., Chen, Y., & Gao, Y. (2020). PLEK2 and SCN7A: novel biomarkers of non-small cell lung cancer. Preprint: Research Square, 1-19. doi:10.21203/rs.3.rs-16005/v1.
- [38] Lv, H., Song, H., Qin, Z., Xing, R., & Chen, Y. (2023). Underexpression of SCN7A is associated with poor prognosis in lung adenocarcinoma. *Gene & Protein in Disease*, 2(1), 363. doi:10.36922/gpd.363.

- [39] Xu, C., Song, L., Yang, Y., Liu, Y., Pei, D., Liu, J., Guo, J., Liu, N., Li, X., Liu, Y., Li, X., Yao, L., & Kang, Z. (2022). Clinical M2 Macrophage-Related Genes Can Serve as a Reliable Predictor of Lung Adenocarcinoma. *Frontiers in Oncology*, 12, 919899. doi:10.3389/fonc.2022.919899.
- [40] Al-Dherasi, A., Huang, Q. T., Liao, Y., Al-Mosaib, S., Hua, R., Wang, Y., Yu, Y., Zhang, Y., Zhang, X., Huang, C., Mousa, H., Ge, D., Sufiyan, S., Bai, W., Liu, R., Shao, Y., Li, Y., Zhang, J., Shi, L., ... Liu, Q. (2021). A seven-gene prognostic signature predicts overall survival of patients with lung adenocarcinoma (LUAD). *Cancer Cell International*, 21(1), 1–16. doi:10.1186/s12935-021-01975-z.
- [41] Kumar-Singh, A., Parniewska, M. M., Giotopoulou, N., Javadi, J., Sun, W., Szatmári, T., Dobra, K., Hjerpe, A., & Fuxe, J. (2021). Nuclear syndecan-1 regulates epithelial-mesenchymal plasticity in tumor cells. *Biology*, 10(6), 521. doi:10.3390/biology10060521.
- [42] Shah, L., Walter, K. L., Borczuk, A. C., Kawut, S. M., Sonett, J. R., Gorenstein, L. A., Ginsburg, M. E., Steinglass, K. M., & Powell, C. A. (2004). Expression of syndecan-1 and expression of epidermal growth factor receptor are associated with survival in patients with nonsmall cell lung carcinoma. *Cancer*, 101(7), 1632–1638. doi:10.1002/cncr.20542.
- [43] Anttonen, A., Heikkilä, P., Kajanti, M., Jalkanen, M., & Joensuu, H. (2001). High syndecan-1 expression is associated with favourable outcome in squamous cell lung carcinoma treated with radical surgery. *Lung Cancer*, 32(3), 297–305. doi:10.1016/S0169-5002(00)00230-0.
- [44] Parimon, T., Brauer, R., Schlesinger, S. Y., Xie, T., Jiang, D., Ge, L., Huang, Y., Birkland, T. P., Parks, W. C., Habel, D. M., Hogaboam, C. M., Gharib, S. A., Deng, N., Liu, Z., & Chen, P. (2018). Syndecan-1 Controls Lung Tumorigenesis by Regulating miRNAs Packaged in Exosomes. *The American Journal of Pathology*, 188(4), 1094–1103. doi:10.1016/j.ajpath.2017.12.009.
- [45] Zhai, Y., Chen, Y., Jiang, Y., & Li, Q. (2019). Weighted Gene Co-expression Network Analysis of Gene Modules for Lung Adenocarcinoma. 2018 5th International Conference on Systems and Informatics, ICSAI 2018, 473–477. doi:10.1109/ICSAI.2018.8599411.
- [46] Chen, M., Zeng, J., Chen, S., Li, J., Wu, H., Dong, X., Lei, Y., Zhi, X., & Yao, L. (2020). SPTBN1 suppresses the progression of epithelial ovarian cancer via SOCS3-mediated blockade of the JAK/STAT3 signaling pathway. *Aging*, 12(11), 10896–10911. doi:10.18632/aging.103303.
- [47] Ying, J., Lin, C., Wu, J., Guo, L., Qiu, T., Ling, Y., Shan, L., Zhou, H., Zhao, D., Wang, J., Liang, J., Zhao, J., Jiao, Y., Lu, N., & Zhao, H. (2015). Anaplastic lymphoma kinase rearrangement in digestive tract cancer: Implication for targeted therapy in Chinese population. *PLoS ONE*, 10(12), 144731. doi:10.1371/journal.pone.0144731.
- [48] Yang, P., Yang, Y., Sun, P., Tian, Y., Gao, F., Wang, C., Zong, T., Li, M., Zhang, Y., Yu, T., & Jiang, Z. (2020). β II spectrin (SPTBN1): Biological function and clinical potential in cancer and other diseases. *International Journal of Biological Sciences*, 17(1), 32–49. doi:10.7150/ijbs.52375.
- [49] Li, A., Zhang, W., Xia, H., Miao, Y., Zhou, H., Zhang, X., Dong, Q., Li, Q., Qiu, X., & Wang, E. (2016). Overexpression of CASS4 promotes invasion in non-small cell lung cancer by activating the AKT signaling pathway and inhibiting E-cadherin expression. *Tumor Biology*, 37(11), 15157–15164. doi:10.1007/s13277-016-5411-5.
- [50] Kang, J. Y., Yang, J., Lee, H., Park, S., Gil, M., & Kim, K. E. (2024). Systematic Multiomic Analysis of PKHD1L1 Gene Expression and Its Role as a Predicting Biomarker for Immune Cell Infiltration in Skin Cutaneous Melanoma and Lung Adenocarcinoma. *International Journal of Molecular Sciences*, 25(1), 359. doi:10.3390/ijms25010359.
- [51] Xu, J., Zheng, H., Yuan, S., Zhou, B., Zhao, W., Pan, Y., & Qi, D. (2019). Overexpression of ANLN in lung adenocarcinoma is associated with metastasis. *Thoracic Cancer*, 10(8), 1702–1709. doi:10.1111/1759-7714.13135.
- [52] Ru, M., Liu, Y.-B., Tian, Z.-H., GODJE, I. S. G., & Li, J.-Q. (2021). ANLN and Lung Adenocarcinoma Prognosis and Immune Infiltration Research. Preprint: Research Square, 1–21. doi:10.21203/rs.3.rs-826813/v1.
- [53] Kasprovicz, A., Sophie, G. D., Lagadec, C., & Delannoy, P. (2022). Role of GD3 Synthase ST8Sia I in Cancers. *Cancers*, 14(5), 1299. doi:10.3390/cancers14051299.
- [54] Tang, W., Wang, J., Dai, T., Qiu, H., Liu, C., Chen, S., & Hu, Z. (2024). Association of leptin receptor polymorphisms with susceptibility of non-small cell lung cancer: Evidence from 2249 subjects. *Cancer Medicine*, 13(8), 7178. doi:10.1002/cam4.7178.
- [55] Ji, X., Bossé, Y., Landi, M. T., Gui, J., Xiao, X., Qian, D., Joubert, P., Lamontagne, M., Li, Y., Gorlov, I., de Biasi, M., Han, Y., Gorlova, O., Hung, R. J., Wu, X., McKay, J., Zong, X., Carreras-Torres, R., Christiani, D. C., ... Amos, C. I. (2018). Identification of susceptibility pathways for the role of chromosome 15q25.1 in modifying lung cancer risk. *Nature Communications*, 9(1). doi:10.1038/s41467-018-05074-y.

- [56] Galluzzi, L., Vitale, I., Senovilla, L., Olaussen, K. A., Pinna, G., Eisenberg, T., Goubar, A., Martins, I., Michels, J., Kratassiouk, G., Carmona-Gutierrez, D., Scoazec, M., Vacchelli, E., Schlemmer, F., Kepp, O., Shen, S., Tailler, M., Niso-Santano, M., Morselli, E., ... Kroemer, G. (2012). Prognostic Impact of Vitamin B6 Metabolism in Lung Cancer. *Cell Reports*, 2(2), 257–269. doi:10.1016/j.celrep.2012.06.017.
- [57] Zuo, H., Ueland, P. M., Midttun, Ø., Tell, G. S., Fanidi, A., Zheng, W., Shu, X., Xiang, Y., Wu, J., Prentice, R., Pettinger, M., Thomson, C. A., Giles, G. G., Hodge, A., Cai, Q., Blot, W. J., Johansson, M., Hultdin, J., Grankvist, K., ... Ulvik, A. (2019). Vitamin B6 catabolism and lung cancer risk: results from the Lung Cancer Cohort Consortium (LC3). *Annals of Oncology*, 30(3), 478–485. doi:10.1093/annonc/mdz002.
- [58] Ahmed, M. B., Alghamdi, A. A. A., Islam, S. U., Lee, J.-S., & Lee, Y.-S. (2022). cAMP Signaling in Cancer: A PKA-CREB and EPAC-Centric Approach. *Cells*, 11(13), 2020. doi:10.3390/cells11132020.
- [59] Han, K., Wang, J., Qian, K., Zhao, T., & Zhang, Y. (2021). Establishment of non-small-cell lung cancer risk prediction model based on prognosis-associated ADME genes. *Bioscience Reports*, 41(10), BSR20211433. doi:10.1042/BSR20211433.
- [60] Zhuo, D., Li, X., & Guan, F. (2018). Biological Roles of Aberrantly Expressed Glycosphingolipids and Related Enzymes in Human Cancer Development and Progression. *Frontiers in Physiology*, 9. doi:10.3389/fphys.2018.00466.
- [61] Wei, Z., Liu, X., Cheng, C., Yu, W., & Yi, P. (2021). Metabolism of Amino Acids in Cancer. *Frontiers in Cell and Developmental Biology*, 8. doi:10.3389/fcell.2020.603837.