ISSN 2959-409X

Genetic selection for lung adenocarcinoma: a statistical approach based on pathological genetic data

Yuhao Wang

Abstract:

Lung adenocarcinoma (LUAD) is the most common subtype of non-small cell lung cancer (NSCLC), with high morbidity and mortality worldwide, posing a serious threat to human health. Due to the complex molecular mechanism of lung adenocarcinoma, which involves a variety of genetic and environmental factors, this makes early diagnosis and treatment a major challenge. This research aims to identify the key pathogenic genes of lung adenocarcinoma to solve the challenges in early diagnosis and treatment. In this paper, three differential analysis methods, limma, DESeq2 and edgeR, were used to select 3315 differential genes. Then, combined with Lasso regression and XGboost algorithm, the pathological gene data were analyzed in depth, and through these advanced statistical and machine learning techniques, the genes related to lung adenocarcinoma were screened from a large number of gene expression data, and 17 characteristic genes of the highest importance were found GO enrichment analysis and KEGG enrichment analysis were performed on these characteristic genes to clarify the biological functions of the genes. Finally, 8 significant pathogenic genes were identified by Cox survival analysis. The expression levels of these genes are closely related to the prognosis of lung adenocarcinoma patients, providing a new perspective for understanding the disease mechanism. The conclusions of this study not only improve the understanding of the molecular mechanism of lung adenocarcinoma, but also provide potential biomarkers for clinical diagnosis and treatment. The discovery of these genes is expected to promote the development of early diagnosis technology and guide the formulation of personalized treatment plans, thereby improving the treatment efficacy and quality of life of lung adenocarcinoma patients. In the future, these genes may become key targets for the development of new drugs and therapeutic strategies.

Keywords: Lung adenocarcinoma; difference analysis; Lasso returns; XGboost; enrichment analysis; Cox survival analysis;

I. Introduction

Cancer is a serious disease with high mortality rate and low cure rate, which poses a serious threat to human life and health worldwide, and the pathogenic mechanism of cancer makes it difficult for human beings to adopt universal and effective treatment methods. Lung cancer is one of the malignant cancers with the highest incidence and mortality among cancers, and according to data from the Chinese National Health Commission in 2024, lung cancer has the highest incidence and mortality among all cancers in China. Lung cancer can be divided into lung adenocarcinoma and lung squamous cell carcinoma, among which lung adenocarcinoma is a common subtype of non-small cell lung cancer (NSCLC), and unlike lung squamous cell carcinoma, the incidence of lung adenocarcinoma is relatively less affected by external influences (especially soot) and may be caused by abnormal genes. According to the World Health Organization (WHO), more than one million people are diagnosed with lung adenocarcinoma every year. Despite advances in early diagnosis and treatment in recent years, the five-year survival rate for lung adenocarcinoma remains relatively low, which has led to an urgent need for researchers to conduct more in-depth research on lung adenocarcinoma.

The molecular mechanism of lung adenocarcinoma is complex and involves a variety of genetic and environmental factors. With the establishment of the TCGA database[15] and the continuous breakthrough of RNA sequencing technology [17], statistical analysis and identification of genes have become easier. The identification of pathogenic genes can help reveal the pathogenesis of lung adenocarcinoma, provide targets for the development of new therapeutic strategies, and formulate more effective early intervention strategies, and the development of genomics and transcriptomics provides powerful tools for studying gene expression patterns in lung adenocarcinoma [9].

A number of studies have identified some genes associated with lung adenocarcinoma, but most have focused on known driver genes, such as EGFR and ALK [8,18,19,21]. Gao and Deng (2018) investigated the correlation between gene expression and location distribution of metastatic

organs; Yang et al. (2022) investigated the relationship between gene mutations and imaging and pathological features; Yang et al. (2021) investigated the driving principles of these genes in the early stage of lung adenocarcinoma, but some non-significant and indirect pathogenic genes are not considered enough; Zubair and Bandyopadhyay (2023) explored the therapeutic mechanism of EGFR inhibitors. However, due to the heterogeneity of lung adenocarcinoma, gene expression patterns may vary significantly between different patients, which increases the complexity of pathogenic gene identification.

At the same time, some studies have combined genetic and clinical data to deeply discuss the development and evolution of lung cancer, as well as the corresponding treatment and prognosis studies [1,2,3,7,11,13,14]. Frankell et al. (2023) explored the relationship between different driver genes and clinical data; Al Bakir et al. (2023) identified the need to develop a targeted metastasis strategy; Martínez-Ruiz et al. (2023) discovered the interaction of genome and transcriptome in influencing ITH, lung cancer evolution, and metastasis; Boumelha et al. (2023) used mouse simulated ERV response expression to predict ICB expression in human lung adenocarcinoma. Abbosh et al. (2023) developed a novel circulating tumor DNA (ctDNA) method for the detection and analysis of residual tumor cells that persist after curative therapy; Al-Sawaf et al. (2023) explored the association between body composition and body weight and survival, and delineated the underlying biological processes and mediators that contribute to the development of CAC; Karasaki et al. (2023) showed that different gene frequencies were seen in tumors with different histological features, suggesting the existence of different evolutionary trajectories. With the discovery of these studies, a huge breakthrough has been made in the clinical diagnosis and treatment of lung cancer.

However, these studies often ignore other potentially important factors in gene expression data, and although some progress has been made in existing studies, there are still many challenges in identifying and validating the genes that cause lung adenocarcinoma. For example, most studies focus on known driver genes and ignore potential novel disease-causing genes. In addition, there may be inconsistencies in results between different studies, which may be related to differences in sample selection, data analysis methods, or statistical methods. This study aims to fill this gap by identifying pathogenic genes for lung adenocarcinoma from pathological genetic data through the comprehensive application of advanced statistical methods and machine learning techniques. Our goal is to improve the accuracy of pathogenic gene identification and provide new biomarkers for early diagnosis, personalized treatment and prognosis evaluation of lung adenocarcinoma. Future studies need to adopt more rigorous experimental design and statistical methods to improve the reliability and generalization ability of pathogenic gene identification.

This research used three differential analysis methods: DESeq2, edgeR and limma, to identify genes with significant differential expression. These methods have different statistical advantages, and through their combined use, the reliability of the results can be improved. In addition, we introduce Lasso regression and XGboost algorithms, two machine learning methods that excel in processing high-dimensional data and discovering complex patterns. Through Cox survival analysis, we further determined the impact of these genes on the survival of lung adenocarcinoma patients. These findings not only contribute to the understanding of the molecular mechanisms of lung adenocarcinoma, but may also have a direct impact on clinical practice, such as guiding the development of personalized treatment strategies.

The article is organized as follows: In section II, we detail the data sources, data wrangling and cleaning processes, as well as the selection and application of variance analysis methods. In section III, we show the screening process and results of pathogenic genes, as well as the detailed steps and conclusions of Cox survival analysis. Finally, in section IV, we discuss the potential implications of the findings of this study for lung adenocarcinoma research and clinical practice, and propose directions for future research.

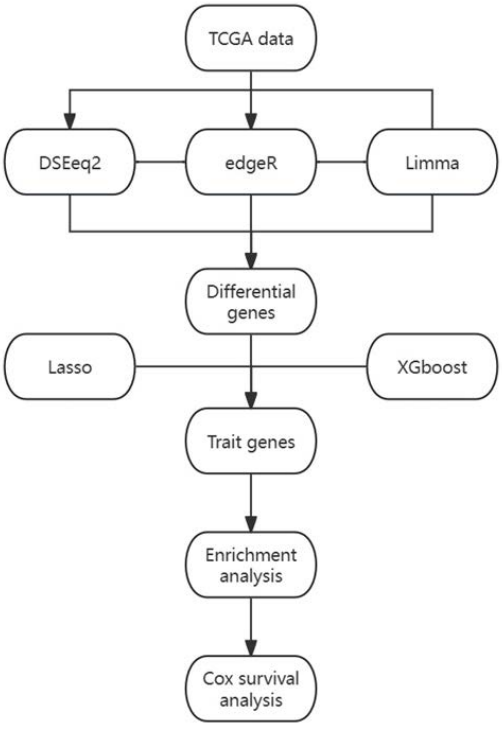


Figure 1.1.1 Process analysis

II. Data Processing

In this paper, the genetic status and clinical data of lung adenocarcinoma patients were selected from the TCGA database, and then the three commonly used differential gene analysis methods in R language, namely limma, DESeq2 and edgeR, were used for specific processing. These methods are widely used in transcriptomics research, and each has its own unique algorithm and application scenarios, so these methods were chosen to identify differentially expressed genes between lung adenocarcinoma and normal tissues.

Data Sources and Processing

The data for this study were derived from The Cancer Genome Atlas (TCGA) project, which provided a large number of gene expression data and clinical information for patients with lung adenocarcinoma. We collected gene expression data samples from the TCGA database, including lung adenocarcinoma patients and normal controls, and 522 samples of relevant clinical information (with duplicates).

Data processing is a critical step in ensuring data quality and the reliability of analytical results. First, the data was processed with the following steps:

1. Data merging: For multiple sample data downloaded from TCGA, we read and merge them into matrix form through R language.
2. Data integrity check and cleaning: ensure that the gene expression data of all samples are complete and there are no missing values.
3. Data type conversion: Convert the raw data into a format suitable for subsequent analysis, such as converting the gene expression matrix into CSV or R data frame format.
4. Data filtering: When processing the transcriptome data matrix, the expression level of the gene in many samples is 0, which means that the gene is very likely to be irrelevant to the problem studied in this paper, and is in the statistical consideration of big data, the study selects 10% of the total number of samples (i.e., 60 cases) as the threshold, if the expression level of the gene in no more than 60 samples is not 0, the gene is removed, and finally the expression matrix we need is obtained.
5. Data grouping: Through classification, 59 normal samples and 541 diseased samples were identified.

Differential genes

After data cleaning and filtering, there were 18662 genes in the expression matrix, which greatly affected the research and analysis of pathogenic genes for lung adenocarcinoma.

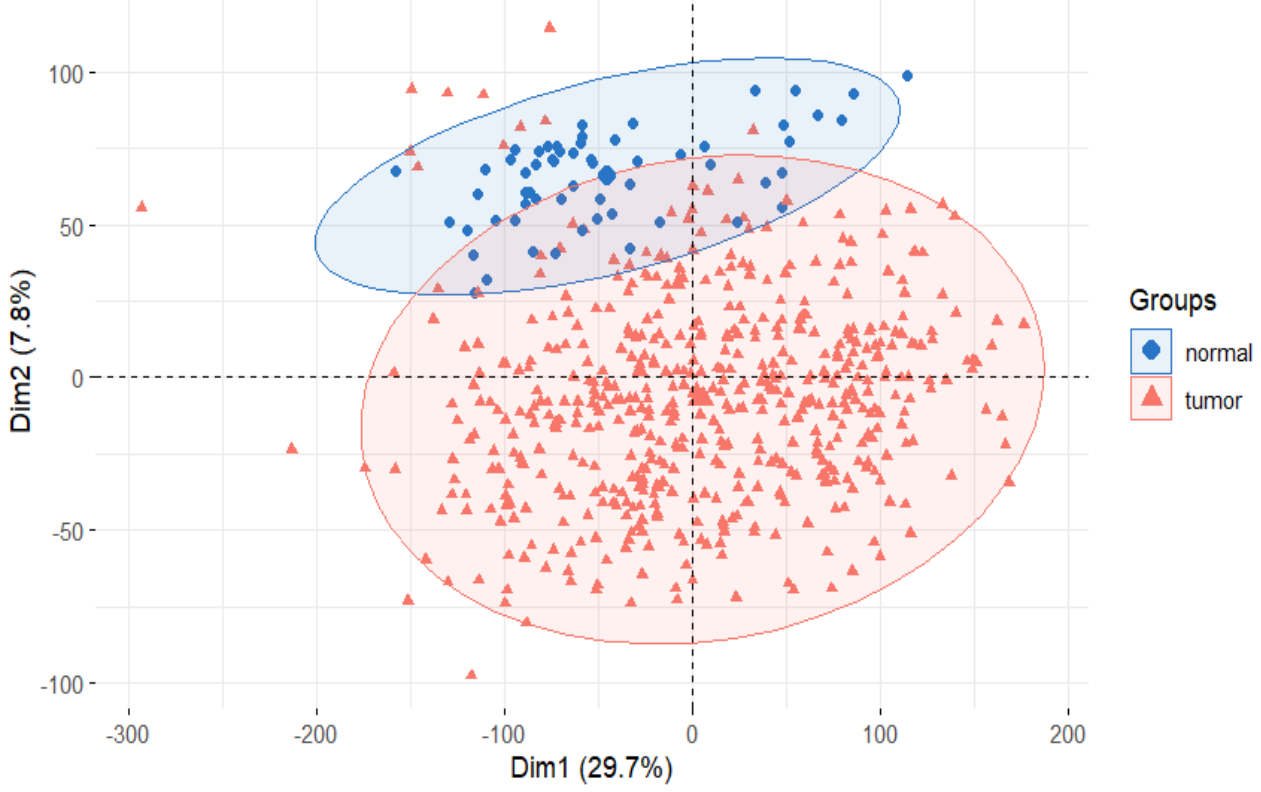


Figure 2.2.1 PCA result

For default reasons, we finalized the up-regulation and down-regulation criteria for differentially impartially muted genes as: $mean(|log_2FC| + 2sd|log_2FC|)$ and $p < 0.05$ as a threshold. The results of the selection of the three differential genes are as follows:

1. DESeq2: 127 down-regulated genes, 847 up-regulated genes, and 17688 non-difference genes were obtained.
2. edgeR: 74 down-regulated genes, 879 up-regulated genes, and 17709 non-difference genes were obtained.
3. Limma: 583 down-regulated genes, 440 up-regulated genes, and 17639 non-differential genes were obtained.

Based on the above three difference analysis methods, their respective heat maps and volcano maps are obtained as follows:

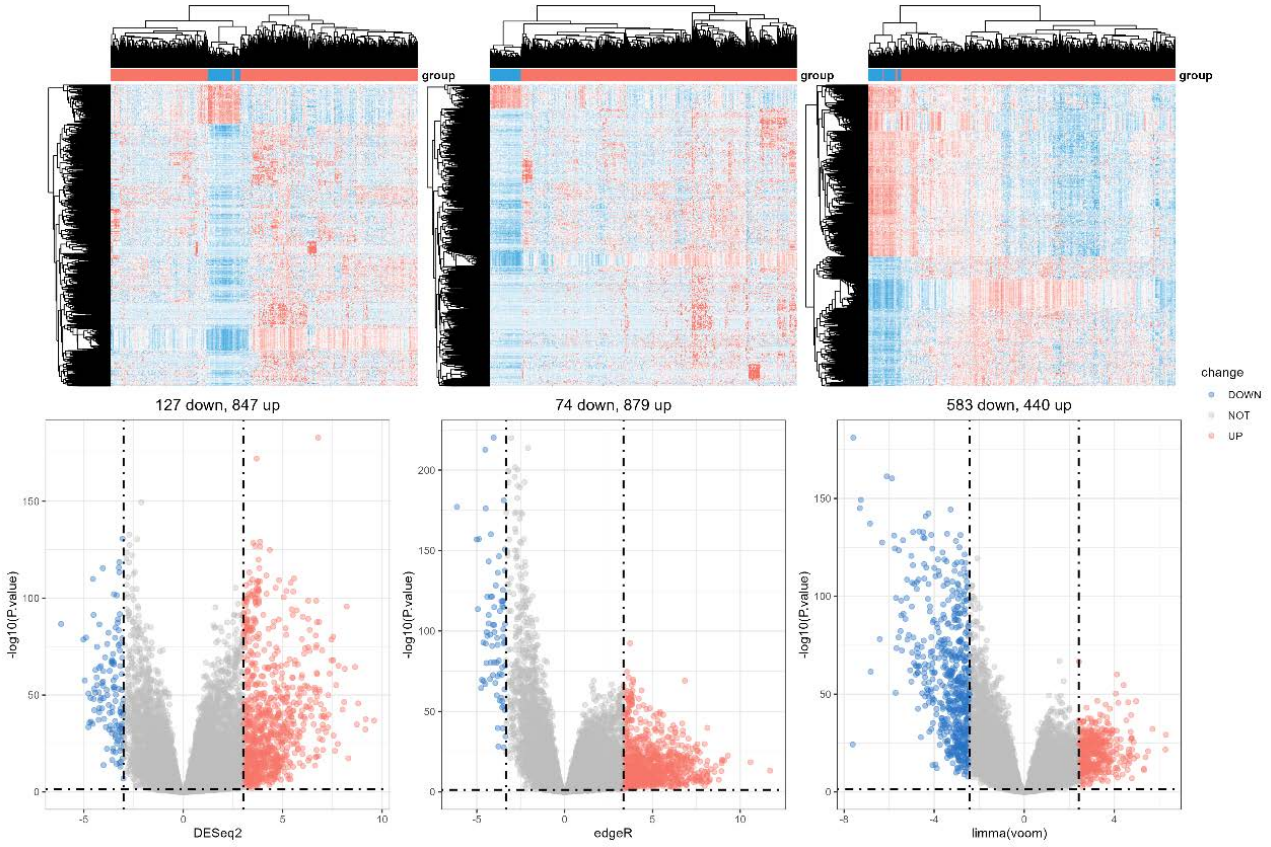


Figure 2.2.2 Three major differences analysis heat map and volcano map

Finally, if the differential gene is selected so that the gene is treated as a differential gene in any of the differential analysis methods, the result is 302 up-regulated genes and 73 down-regulated genes, with a total of 375 differential genes.

However, this screening method leads to a small number of differential genes, which will delete some important differential genes, which is not conducive to the subsequent selection of characteristic genes and the determination of the final pathogenic genes. At the same time, in view of the advantages and disadvantages of the above three differential analysis methods in differential gene selection, we believe that as long as it is up-regulated or down-regulated in a differential analysis, it is considered to be the required differential gene. By this screening method, 3315 differential genes without duplications were selected.

III. Selection of disease-causing genes

This section employs two statistical or machine learning methods, Lasso regression and XGboost methods, to further screen and validate differential genes, which are uniquely advantageous in processing high-dimensional data and discovering complex patterns, and then picking out the trait genes needed for the study. After obtaining the identified characteristic genes, we further evaluated the relationship between gene expression and survival time of lung adenocarcinoma patients through enrichment analysis and Cox survival analysis, and obtained the final pathogenic genes, and at the same time plotted the corresponding survival curves.

Application of Lasso regression and XGboost method

Lasso regression is a linear regression model with L1 regularization terms, capable of feature selection, i.e., automatic exclusion of unimportant variables during model training, as follows:

$$\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right) + \lambda\sum_{j=1}^{k}|\beta_j|$$

Compared with other regression analysis methods, Lasso's L1 regular term can completely remove some coefficients, which is very useful in high-dimensional processing and feature selection, and can effectively prevent overfitting.

Through the analysis of the gene sample expression matrix in Section 2, Lasso regression was carried out, and the highest degree of fitting was obtained when $\lambda = 0.001062$, and then the upper and lower bounds of λ were selected and re-introduced into the regression model through cross-validation, and finally 8 characteristic genes were obtained: GAPDH, AUNIP, IGF2BP1, CDC25A, FAM83B, RGS20, GIMAP5, and DKK1. At the same time, the ROC curve was used to evaluate the fitting ability of the model, and the $AUC = 0.657$ was obtained.

However, due to the strong punitive nature of Lasso, its ability to predict the model is relatively weak, so the next machine learning method will be used to further improve the prediction ability of the model, and also provide another method for selecting characteristic genes.

XGboost is an ensemble learning method based on gradient boosting decision trees, which excels in handling unbalanced data and improving prediction accuracy.

We use the default parameters of XGboost in the R language package to analyze the same data matrix, and in the selection of the training set and the test set, we divide into two methods. Method 1: simply select the test set and the training set in a ratio of 1:4, take importance as the main parameter, and obtain 6 characteristic genes: IGF2BP1, PCP4, PKIB, KIAA1549L, CDH22, CALB1, and use the ROC curve to evaluate the fitting ability of the model, and obtain $AUC = 0.742$; Method 2: On the basis of method 1, after the average split of the training set, after comprehensive training in two small training sets, fitting in the original test set, similarly obtained 6 characteristic genes: DKK1, ACSM6, VTCN1, TMC2, DYDC1, KRT81, and using the ROC curve to evaluate the fitting ability of the model, and obtained $AUC = 0.918$.

Finally, by combining Lasso regression and XGboost methods, 18 characteristic genes were selected after removing duplicate genes. In terms of cancer prediction, the XGboost method is significantly better than Lasso, and the combination of the two methods can further improve the fitting ability of the model by improving the selection of characteristic genes.

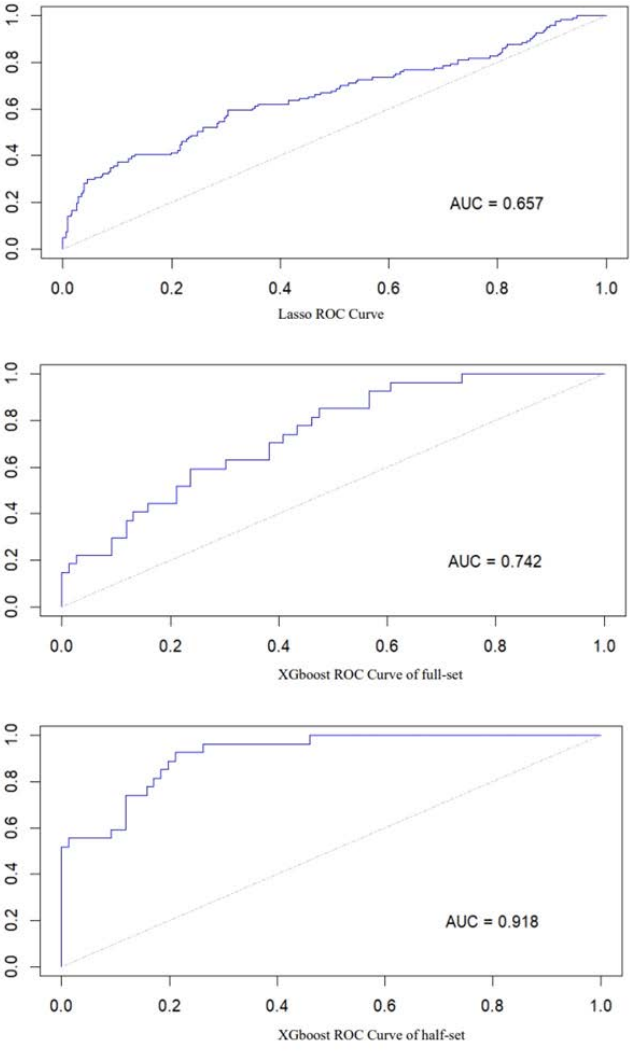


Figure 3.1.1 Simulated fitting ROC curves of the three models

Enrichment Analysis

GO (Gene Ontology) enrichment analysis is a widely used method to determine the statistical significance of genes in a gene set in terms of cellular composition (CC), biological processes (BP), and molecular function (MF). In this way, it is possible to discover the correlation between gene function and disease states. GO enrichment analysis of the above 18 characteristic genes yielded 15 significant results for cell composition, 20 significant results for biological processes and 10 significant results for molecular function ($p < 0.05$). These genes regulate the formation of cellular components, control their division, and thus affect the outcome of lung adenocarcinoma.

KEGG (Kyoto Encyclopedia) enrichment assays focus

on identifying whether genes in a gene set are involved in specific metabolic pathways or signaling pathways. The KEGG database provides a wealth of biological information, including metabolic pathways, disease-related genes, and drug mechanisms of action. Through KEGG enrichment analysis, it is possible to understand the interaction of genes and their network relationships in the occurrence and development of diseases. The results show that these genes play a key role in regulating metabolism, energy balance, and cellular stress response.

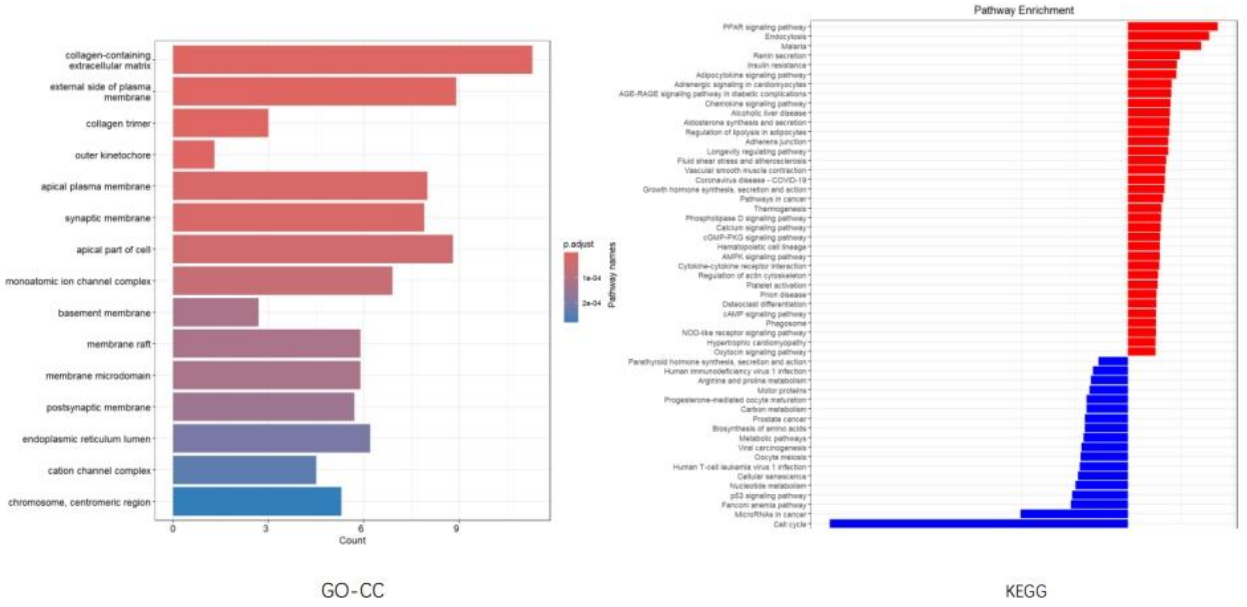


Figure 3.2.1 The results of the enrichment analysis are partially displayed

Identification and conclusion of pathogenic genes

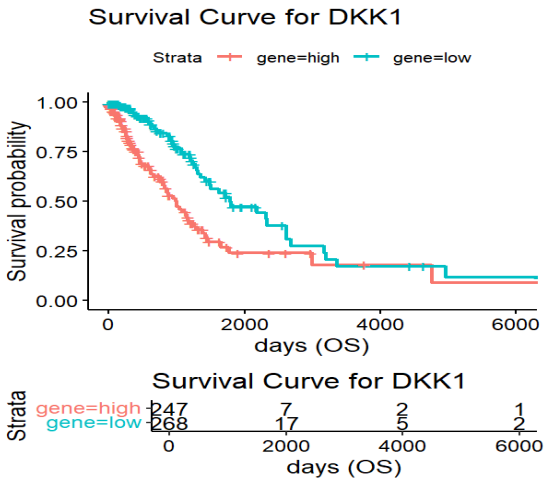
After the 18 characteristic genes were selected based on Lasso regression and XGboost method, the pathogenic genes were further screened out, and the main basis of the study was Cox proportional hazards regression, which has significant advantages in dealing with the effects of multiple covariates and data censoring. Therefore, combined with clinical data, in order to exclude the significant expression results caused by non-lung adenocarcinoma direct factors such as age, sex and race, 8 significant genes were finally obtained from Cox analysis: DKK1, FAM83B, KIAA1549, KRTB1, GAPDH, PKIB, IGF2BP1, and RGS20 ($p<0.05$).

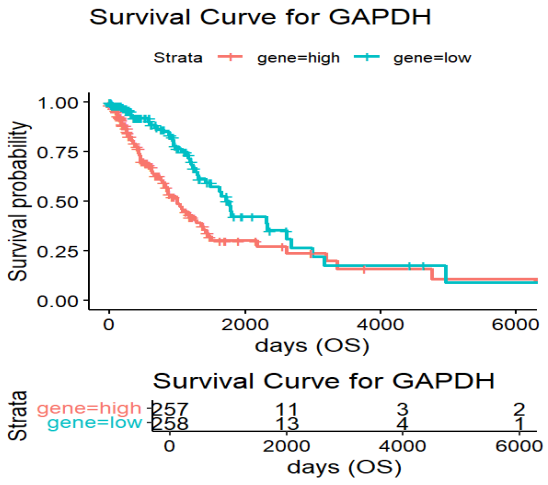
The results of the Cox analysis are as follows:

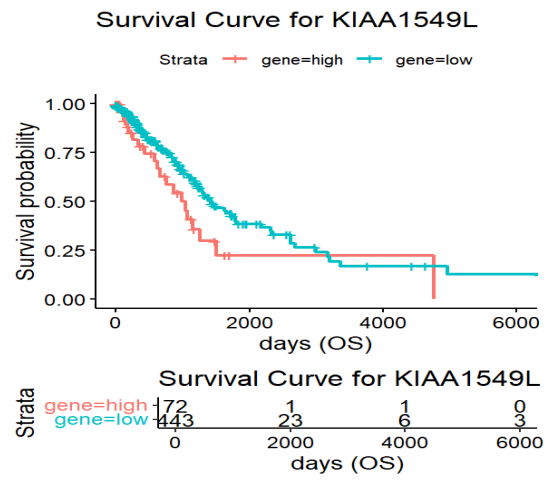
Gene	coef	se	z	p	HR	HRCILL	HRCIUL
GAPDH	0.673	0.184	3.651	0.000	1.961	1.366	2.815
DKK1	0.812	0.186	4.370	0.000	2.251	1.564	3.240
KIAA1459L	0.626	0.184	3.394	0.001	1.870	1.303	2.685
KRT81	0.588	0.189	3.114	0.002	1.801	1.244	2.608
PKIB	0.506	0.185	2.727	0.006	1.685	1.153	2.385
FAM83B	0.490	0.182	2.697	0.007	1.633	1.143	2.332
RGS20	0.494	0.184	2.683	0.007	1.639	1.142	2.350
IGF2BP1	0.474	0.184	2.576	0.010	1.606	1.120	2.304

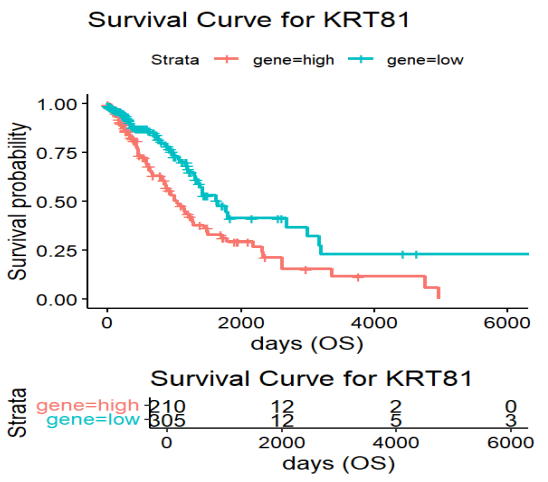
Table 3.3.1 Cox analysis of salient genes

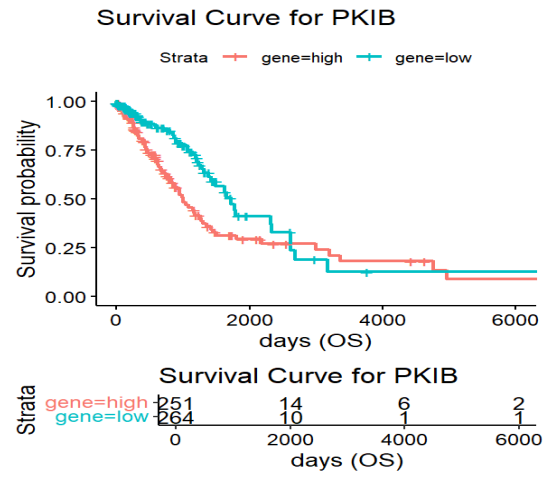
Next, the survival curves of these 8 characteristic genes were plotted to visualize the results:

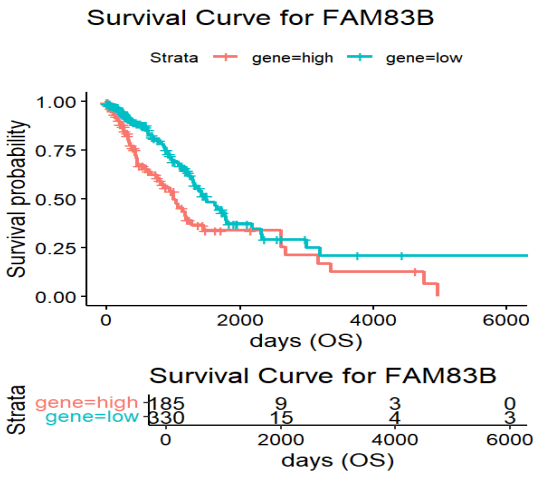

Survival Curve for DKK1
Strata gene=high gene=low
Survival probability
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
days (OS)
Survival Curve for DKK1
Strata
gene=high 247 7 2 1
gene=low 268 17 5 2
0
2000
4000
6000
days (OS)


Survival Curve for GAPDH
Strata gene=high gene=low
Survival probability
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
days (OS)
Survival Curve for GAPDH
Strata
gene=high 257 11 3 2
gene=low 258 13 4 1
0
2000
4000
6000
days (OS)


Survival Curve for KIAA1549L
Strata gene=high gene=low
Survival probability
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
days (OS)
Survival Curve for KIAA1549L
Strata
gene=high 172 1 1 0
gene=low 443 23 6 3
0
2000
4000
6000
days (OS)


Survival Curve for KRT81
Strata gene=high gene=low
Survival probability
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
days (OS)
Survival Curve for KRT81
Strata
gene=high 210 12 2 0
gene=low 305 12 5 3
0
2000
4000
6000
days (OS)


Survival Curve for PKIB
Strata gene=high gene=low
Survival probability
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
days (OS)
Survival Curve for PKIB
Strata
gene=high 251 14 6 2
gene=low 264 10 1 1
0
2000
4000
6000
days (OS)


Survival Curve for FAM83B
Strata gene=high gene=low
Survival probability
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
days (OS)
Survival Curve for FAM83B
Strata
gene=high 185 9 3 0
gene=low 330 15 4 3
0
2000
4000
6000
days (OS)

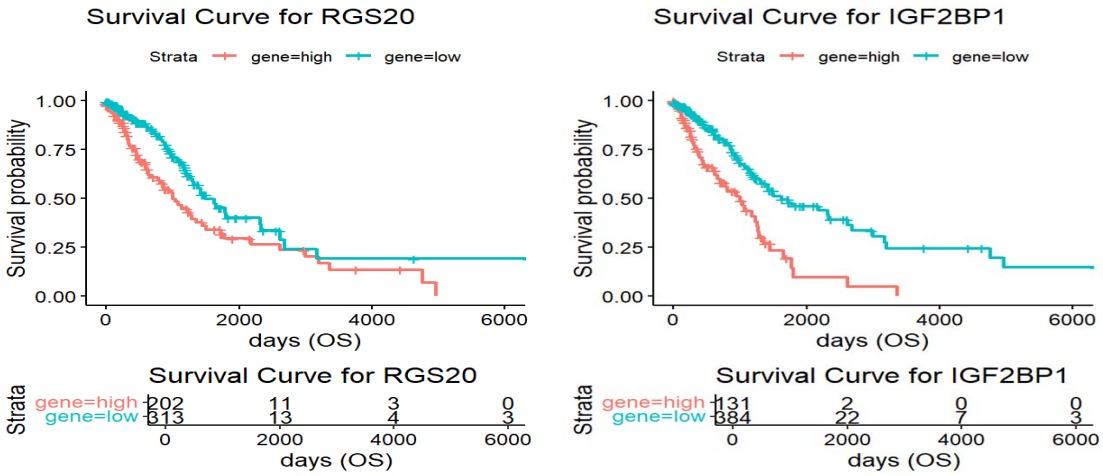


Figure 3.3.1 Survival curves of 8 characteristic genes

From Table 3.3.1 and Figure 3.3.1, it can be found that the final selection of eight pathogenic genes does play an important role in the occurrence and subsequent impact of lung adenocarcinoma, and at the same time, through literature search, we can understand their pathogenic mechanism from a biological sense, and have a deeper understanding of whether these genes can become prognostic markers and targets for lung adenocarcinoma immunotherapy: DKK1[12]; FAM83B[16]; KRT8[4]; GAPDH[10]; PKIB[6]; IGF2BP1[20]; RGS20[5].

IV. Discussion and Conclusion

Discussion

As the most common subtype of non-small cell lung cancer, lung adenocarcinoma has a high incidence and mortality rate worldwide, posing a serious threat to human health. In view of the complex molecular mechanism of lung adenocarcinoma, in-depth study of its pathogenic genes is of great significance to reveal the nature of the disease, improve the early diagnosis rate and develop new treatment strategies. In this study, 3315 differential genes were screened by comprehensive application of limma, DESeq2 and edgeR. On this basis, 18 pathogenic genes with the highest importance were successfully screened by Lasso regression and XGboost algorithm, and their pathogenic mechanisms were found by GO and KEGG enrichment analysis. Further Cox survival analysis identified 8 significant pathogenic genes, providing a new perspective for the molecular mechanism research and clinical treatment of lung adenocarcinoma. Compared to previous studies, this study not only identified known driver genes, such as EGFR and ALK, but also identified some new potential pathogenic genes. These findings expand our understanding of the gene expression profile of lung adenocarcinoma and provide new research directions for future research. The screening of pathogenic genes not only helps to deepen the understanding of the molecular mechanisms of lung adenocarcinoma, but may also have a direct impact on clinical practice, such as guiding the development of personalized treatment strategies and the improvement of prognostic assessment. Although this study has made some progress in the identification of pathogenic genes, there are still some limitations. For example, limitations in the sample size of studies downloaded from the TCGA database may have affected the generalizability of the results; The screening method of differential genes leads to the problem of multicollinearity in the sample, which makes the fitting effect of the subsequent model unsatisfactory. In addition, this study mainly focuses on the changes at the gene expression level, and the in-depth analysis of gene function and regulatory network is insufficient. The findings of this study have potential clinical significance for early diagnosis, personalized treatment and prognosis evaluation of lung adenocarcinoma. In the future, these pathogenic genes can be used as new therapeutic targets or diagnostic markers to further promote the development of precision medicine for lung adenocarcinoma.

Conclusion

In this study, advanced statistical and machine learning techniques were used to screen 17 key genes associated with lung adenocarcinoma from pathological data, and 8 genes were confirmed to have a significant impact on patient survival through Cox analysis. These findings provide new biomarkers for the early diagnosis and treatment of lung adenocarcinoma, which is expected to improve the prognosis of patients.

In the future, these genes may become new targets for clinical treatment, which will help develop personalized medical protocols and improve patient survival. Despite the limitations of the research, the results have opened up a new path for the precision treatment of lung adenocarci-

noma and the improvement of quality of life.

References

[1] *Abbosh, C., Frankell, A.M., Harrison, T. et al. Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA. Nature* 616, *553–562 (2023). https://doi.org/10.1038/s41586-023-05776-4.*

[2] *Al Bakir, M., Huebner, A., Martínez-Ruiz, C. et al. The evolution of non-small cell lung cancer metastases in TRACERx. Nature* 616, *534–542 (2023).* https://doi.org/10.1038/s41586-023-05729-x.

[3] *Al-Sawaf, O., Weiss, J., Skrzypski, M. et al. Body composition and lung cancer-associated cachexia in TRACERx. Nat Med* 29, *846–858 (2023). https://doi.org/10.1038/s41591-023-02232-8.*

[4] *Campayo M, Navarro A, Viñolas N, Tejero R, Muñoz C, Diaz T, Marrades R, Cabanas ML, Gimferrer JM, Gascon P, Ramirez J, Monzo M. A dual role for KRT81: a miR-SNP associated with recurrence in non-small-cell lung cancer and a novel marker of squamous cell lung carcinoma. PLoS One. 2011;6(7):e22509. doi: 10.1371/journal.pone.0022509. Epub 2011 Jul 25. PMID: 21799879; PMCID: PMC3143163.*

[5] *Ding, X., Li, X., Jiang, Y. et al. RGS20 promotes non-small cell lung carcinoma proliferation via autophagy activation and inhibition of the PKA-Hippo signaling pathway. Cancer Cell Int 24, 93 (2024). https://doi.org/10.1186/s12935-024-03282-9.*

[6] *Dou P, Zhang D, Cheng Z, Zhou G, Zhang L. PKIB promotes cell proliferation and the invasion-metastasis cascade through the PI3K/Akt pathway in NSCLC cells. Exp Biol Med (Maywood). 2016 Nov;241(17):1911-1918. doi: 10.1177/1535370216655908. Epub 2016 Jun 20. PMID: 27325557; PMCID: PMC5068460*

[7] *Frankell, A.M., Dietzen, M., Al Bakir, M. et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. Nature* 616,*525–533(2023). https://doi.org/10.1038/s41586-023-05783-5.*

[8] *Gao G, Deng L. [Association between EGFR, ALK and KRAS Gene Status and Synchronous Distant Organ Metastasis in Non-small Cell Lung Cancer]. Zhongguo Fei Ai Za Zhi. 2018 Jul 20;21(7):536-542. Chinese. doi: 10.3779/j.issn.1009-3419.2018.07.06. PMID: 30037374; PMCID: PMC6058661.*

[9] *Han G, Sinjab A, Rahal Z, Lynch AM, Treekitkarnmongkol W, Liu Y, Serrano AG, Feng J, Liang K, Khan K, Lu W, Hernandez SD, Liu Y, Cao X, Dai E, Pei G, Hu J, Abaya C, Gomez-Bolanos LI, Peng F, Chen M, Parra ER, Cascone T, Sepesi B, Moghaddam SJ, Scheet P, Negrao MV, Heymach JV, Li M, Dubinett SM, Stevenson CS, Spira AE, Fujimoto J, Solis LM, Wistuba II, Chen J, Wang L, Kadara H. An atlas of epithelial cell states and plasticity in lung adenocarcinoma. Nature. 2024 Mar;627(8004):656-663. doi: 10.1038/s41586-024-07113-9. Epub 2024 Feb 28. Erratum in: Nature. 2024 Apr;628(8006):E1. doi: 10.1038/s41586-024-07277-4. PMID: 38418883; PMCID: PMC10954546.*

[10] *Jin Wang, Xueting Yu, Xiyuan Cao, Lirong Tan, Beibei Jia, Rui Chen, Jianxiang Li,GAPDH: A common housekeeping gene with an oncogenic role in pan-cancer,Computational and Structural Biotechnology Journal,Volume 21,2023,Pages 4056-4069,.*

[11] *Karasaki, T., Moore, D.A., Veeriah, S. et al. Evolutionary characterization of lung adenocarcinoma morphology in TRACERx. Nat Med* 29, *833–845 (2023). https://doi.org/10.1038/s41591-023-02230-w.*

[12] *Lu R, Li Y, Xieyidai A, Yu T, Feng Y. [Exploring the Role of DKK1 in the Occurrence of Lung Adenocarcinoma Based on the Analysis of Bioinformatics]. Zhongguo Fei Ai Za Zhi. 2023 Aug 20;26(8):579-590. Chinese. doi: 10.3779/j.issn.1009-3419.2023.101.22. PMID: 37752538; PMCID: PMC10558759.*

[13] *Martínez-Ruiz, C., Black, J.R.M., Puttick, C. et al. Genomic–transcriptomic evolution in lung cancer and metastasis. Nature* 616, *543–552 (2023).* https://doi.org/10.1038/s41586-023-05706-4.

[14] *Ng, K.W., Boumelha, J., Enfield, K.S.S. et al. Antibodies against endogenous retroviruses promote lung cancer immunotherapy. Nature* 616, *563–573 (2023).* https://doi.org/10.1038/s41586-023-05771-9.

[15] *Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68-77. doi: 10.5114/wo.2014.47136. PMID: 25691825; PMCID: PMC4322527.*

[16] *Wang J, Li P, Sun L, Zhang J, Yue K, Wang Y, Wu X. FAM83B regulates mitochondrial metabolism and anti-apoptotic activity in pulmonary adenocarcinoma. Apoptosis. 2024 Jun;29(5-6):743-756. doi: 10.1007/s10495-024-01944-7. Epub 2024 Mar 13. PMID: 38478170.*

[17] *Wang Z, Li Z, Zhou K, Wang C, Jiang L, Zhang L, Yang Y, Luo W, Qiao W, Wang G, Ni Y, Dai S, Guo T, Ji G, Xu M, Liu Y, Su Z, Che G, Li W. Deciphering cell lineage specification of human lung adenocarcinoma with single-cell RNA sequencing. Nat Commun. 2021 Nov 11;12(1):6500. doi: 10.1038/s41467-021-26770-2. PMID: 34764257; PMCID: PMC8586023.*

[18] *Yang H, Liu Z, Wang H, Chen L, Wang J, Wen W, Xu X, Zhu Q. [Relationship between EGFR, ALK Gene Mutation and Imaging and Pathological Features in Invasive Lung Adenocarcinoma]. Zhongguo Fei Ai Za Zhi. 2022 Mar 20;25(3):147-155. Chinese. doi: 10.3779/j.issn.1009-3419.2022.101.10. PMID: 35340157; PMCID: PMC8976203.*

[19] *Yang R, Liao X, Lei Y, et al. Advances in the Key Oncogene and Tumor-suppressor Gene in Early Lung Adenocarcinoma [J]. Chinese General Practice, 2021, 24(29): 3774-3780.* DOI: 10.12114/j.issn.1007-9572.2021.00.542..

[20] *Zhu Q, Zhang C, Qu T, Lu X, He X, Li W, Yin D, Han L, Guo R, Zhang E. MNX1-AS1 Promotes Phase Separation of IGF2BP1 to Drive c-Myc-Mediated Cell-Cycle Progression*

and Proliferation in Lung Cancer. Cancer Res. 2022 Dec 2;82(23):4340-4358. doi: 10.1158/0008-5472.CAN-22-1289. PMID: 36214649.

[21] *Zubair T, Bandyopadhyay D. Small Molecule EGFR Inhibitors as Anti-Cancer Agents: Discovery, Mechanisms of Action, and Opportunities. Int J Mol Sci. 2023 Jan 31;24(3):2651. doi: 10.3390/ijms24032651. PMID: 36768973; PMCID: PMC9916655.*