

# Elucidating the Predictive Capacity of Gene Expression in Lung Carcinogenesis

Xilei Yan

School of Medical Economic Management, Nanjing University of Chinese Medicine, Nanjing, China  
012421118@njucm.edu.cn

**Abstract.** Lung cancer remains the leading cause of cancer-related mortality globally, with the highest incidence rates reported worldwide. This pervasive disease is significantly influenced by factors such as smoking, air pollution, and occupational exposure, with varying incidence rates across different regions and demographics. Advances in genomic technology have ushered in a new era of high-throughput gene analysis, enhancing our understanding of lung cancer's molecular underpinnings. This study focuses on the gene expression profiles of 551 patients, employing techniques such as volcano plots and binary regression analysis to identify key genetic markers associated with lung cancer. Our findings highlight five critical genes: CLDN18, GKN2, LYVE-1, GPIHBP1, and CLIC5, which not only play significant roles in the early detection of lung cancer but also hold potential as targets for future therapeutic interventions. By leveraging gene expression for early screening, this research contributes to improving lung cancer prognosis and paves the way for developing targeted treatment strategies, thereby underscoring the importance of integrating molecular genetics into clinical practice for personalized medicine approaches in lung cancer management.

**Keywords:** Lung cancer, gene expression.

## 1. Introduction

Lung cancer continues to be a major public health challenge and is recognized as one of the most prevalent cancers globally [1, 2]. According to data from the World Health Organization, lung cancer exhibits the highest incidence and mortality rates among all cancers, with significantly higher rates observed in men compared to women. However, the incidence among women is progressively increasing, particularly in those with smoking habits [3, 4]. The variation in lung cancer rates across different countries and regions is largely attributed to local factors such as smoking habits, air pollution, and occupational exposures. Notably, China, as one of the largest developing nations, accounts for approximately 40% of all global mortality from lung cancer and over 30% of newly diagnosed cases [5].

Advancements in genomic technology have enabled scientists to perform high-throughput analyses on a multitude of genes, aiding in the identification of genetic variations that are linked to lung cancer. This technological progress has shifted the focus from external risk factors, such as smoking and air pollution, to exploring the predictive capabilities of gene expression in the onset of lung cancer. Understanding gene expression patterns can facilitate early screening and potentially improve the prognosis for lung cancer patients. The investigation into the genomic underpinnings of lung cancer not only reveals the molecular mechanisms of the disease but also aids in the development of targeted therapeutic interventions.

This study contributes significantly to the field of lung cancer research by employing sophisticated genomic analysis techniques to elucidate the role of gene expression in the development of lung cancer. Key contributions of this research include:

Through detailed analysis, this study identifies critical genes associated with lung cancer, providing valuable insights into the biological mechanisms driving the disease.

By leveraging gene expression data, we have developed methodologies that can potentially screen lung cancer at its early stages, significantly improving patient outcomes.

The genetic insights gained from this study lay the groundwork for the creation of targeted therapies that are customized to the genetic profile of individual patients, moving towards more personalized and effective treatment strategies.

Overall, this research enhances our understanding of the genetic factors contributing to lung cancer, supports the early diagnosis and precision treatment of the disease, and paves the way for the advancement of personalized medicine in oncology.

## **2. Definition of Basic Concepts**

### **2.1. Lung Cancer**

A cancerous growth that starts in the lungs is called lung cancer. typically formed by abnormal proliferation of cells in lung tissue. It is one of the most common cancers and a leading cause of cancer-related deaths. Lung cancer can be divided into several main types, non-small cell lung cancer, which accounts for around 85% of all occurrences of lung cancer, is the most prevalent kind of the disease. Lung cancer with small cells: Although this kind of lung cancer is not very common, but develops rapidly and easily spreads, often closely related to smoking. Due to late detection and lack of effective treatment options, lung cancer is also one of the leading causes of cancer-related deaths. Therefore, identifying potential high-risk cancer patients is crucial for early detection. In addition to well-known risk factors such as smoking, age, and genetics, other factors such as alcohol consumption and air pollution should also be considered. In addition, certain symptoms such as coughing up blood, wheezing, and nail pain are important indicators of lung cancer [6].

### **2.2. Gene Expression**

Gene expression refers to the process of converting genetic information into functional products. This process involves two main steps: transcription and translation. Transcription is a process in which DNA converts genetic information to messenger RNA, which takes place in the nucleus. Translation refers to the transformation of mRNA genes into proteins, which takes place in the cytoplasm of ribosome. Gene expression is based on DNA to RNA, and then to protein. Transcription converts gene information on DNA into mRNA, while translation converts information on mRNA into proteins. This process is the foundation of various functions and activities within cells, therefore it is crucial for the life activities of cells and the functions of tissues.

## **3. Methods**

The dataset contains 551 patients. Each patient has 56907 different expressed genes, with a total of 49 patients diagnosed as healthy and the remaining 502 patients diagnosed as cancer in this dataset. This study used R studio as the main data processing method and utilized volcano plots to demonstrate the differential expression between healthy patients and cancer patients. Use 80% of the data to make a volcano map and 20% of the data to predict the outcome. By displaying both the fold difference and P-value on a single graph, it is possible to more intuitively screen for genes with significant differential expression between two sample groups. In addition, a binary system is used to label whether patients are healthy or have lung cancer. The numbers 1 and 0 represent disease or health conditions, respectively, and two models were created: Logistic regression and Random forest, to predict gene expression during lung cancer development.

3.1. Volcano Map

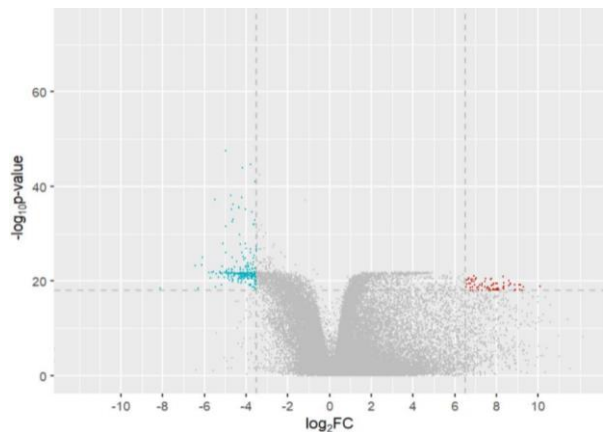


Figure 1. Volcano map (Photo credit: Original)

From the above Fig. 1, it can be seen that the x-axis represents Fold change, with values deviating from the center. The greater the variation in differences. The y-axis represents p-value, The larger the value along the vertical value, the more significant the difference. In the Figure, the red dense dots represent upregulated genes, and the green dense dots represent downregulated genes.

3.2. Logistic Regression

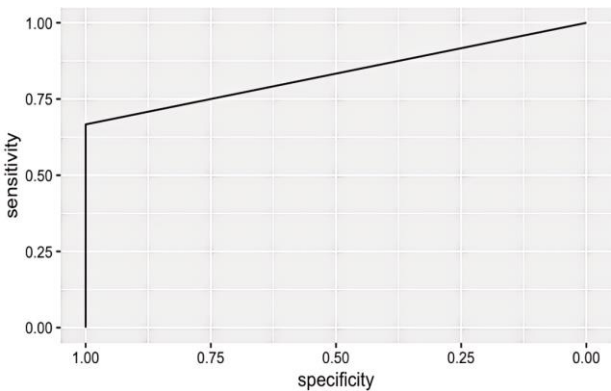


Figure 2. AUC map (Photo credit: Original)

From the above Fig. 2, it can be seen that this is a graph related to logistic regression and specificity. The horizontal and vertical axes represent the value of specificity, reflecting the variation from 0 to 1. The area under the curve is AUC. In logistic regression models, AUC is usually used to evaluate the performance of the model. The AUC model in the above Figure measures the ability of the model to distinguish between diseased and non diseased individuals. If the AUC value is close to 1, it indicates that the model has strong discriminative ability. If the AUC model is close to 0.5, it indicates that the model has almost no discriminative ability, similar to random guessing. The AUC value in the above Figure is 0.83, with an accuracy of 64.7%. Therefore, through this model, we can obtain good results to distinguish the ability of individuals with and without disease.

Table 1. Datagram

| Accuracy  | Kappa     | AccuracySD | KappaSD   |
|-----------|-----------|------------|-----------|
| <dbl>     | <dbl>     | <dbl>      | <dbl>     |
| 0.9701374 | 0.7850606 | 0.03265688 | 0.2106893 |

From the above Table 1, the accuracy value is about 0.97, indicating the accuracy of gene measurement and prediction; The kappa value is approximately 0.79, used to evaluate the degree of gene consistency and similarity; The value of accuracy SD is about 0.03, which is the standard deviation of accuracy and reflects the degree of dispersion of accuracy; The value of kappa SD is about 0.21, which is the standard deviation of kappa and also reflects the dispersion of kappa.

## 4. Discussion

**Table 2.** Number of affected genes

[illegible]

As show in the table 2. A total of 75 unique genes were selected as influencing genes, with most genes only being selected as influencing genes once. The CLDN18 gene was randomly selected 4 times; Genes GKN2, LYVE-1, GPIHBP1, and CLIC5 were randomly selected three times each; Genes SFTPA1, PRG4, SLC39A8, AC097537.1, AL078624.1, AC093909.5, ADGRD1, ADH1B, ANGPT4, CHIA, CHIAP2, DES, F11, FHL1, GDF10, LHFPL3AS2, 1, SLC46A2, TNXB, VEPH1, INMT, C1QTNF7, PGM5P4 were randomly selected twice. Among these influencing genes, some genes are positively correlated with whether a patient has cancer, while others are negatively correlated.

It was found in the study, CLDN18 has two subtypes: CLDN18.1 specifically expressed in lung tissue and CLDN18.2 expressed in gastric tissue. The CLDN18 gene can serve as an oncogene or tumor suppressor gene in different tumor types [7]. There is evidence to suggest that GKN2 deficiency is one of the mechanisms underlying the pathogenesis of gastric cancer, and suggesting that restoring GKN2 expression can be a strategy for inhibiting gastric cancer progression [8, 9]. It can be speculated that GKN2 expression may also be absent in GCSCs, and restoring GKN2 expression may have a similar inhibitory effect on GCSCs [10]. Through this study, it was also found that the GKN2 gene has a certain inhibitory effect on lung cancer.

## 5. Conclusion

This study has successfully identified significant genes such as CLDN18, GKN2, LYVE-1, GPIHBP1, and CLIC5 that play pivotal roles in the diagnosis of lung cancer. The investigation of these genes through high-throughput genomic technologies has not only underscored their influence in the progression of lung cancer but also emphasized their potential as biomarkers for early detection. Importantly, the analysis revealed that a set of 28 genes, including these five, are strongly associated with lung cancer and may serve as reliable indicators for the presence of the disease. These findings open the door to developing targeted therapeutic drugs, potentially revolutionizing the treatment of lung cancer patients by enabling more personalized medicine approaches.

Despite the promising discoveries, the study faces limitations that future research must address to enhance the accuracy and applicability of the findings. The current model's weighting method for calculating the AUC value requires refinement to ensure more rigorous and precise outcomes. Additionally, utilizing only 20% of patients as a test dataset has posed challenges in adequately training the predictive model, suggesting the need for larger datasets for validation. Future studies should focus on expanding the gene pool examined and incorporating more comprehensive patient data to improve the predictive accuracy. Moreover, translating these genomic discoveries into clinical practice will involve overcoming significant translational and regulatory hurdles. The continued exploration of gene expression in lung cancer will be crucial for advancing understanding and treatment of this complex disease, ultimately aiming to reduce its burden globally.

## References

- [1] Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*, 2015, 136 (5): E359-386.
- [2] Li X. Lung cancer risk prediction and feature importance analysis with machine learning algorithm. The 5th International Conference on Computing and Data Science (CONF-CDS 2023), Macao China, 2023.
- [3] Liu C, Shi J, Wang H, et al. Population-level economic burden of lung cancer in China: Provisional prevalence-based estimations, 2017-2030. *Chinese Journal of Cancer Research*, 2021, 33 (01): 79-96.
- [4] Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*, 2021, 71 (3): 209-249.
- [5] Zhang Z, Xue H, Dong Y, et al. GKN2 promotes oxidative stress-induced gastric cancer cell apoptosis via the Hsc70 pathway. *J Exp Clin Cancer Res*, 2019, 38 (1): 338.
- [6] Zhu, X., Guo, C., Feng, H., Huang, Y., Feng, Y., Wang, X., & Wang, R. (2024). A Review of Key Technologies for Emotion Analysis Using Multimodal Information. *Cognitive Computation*, 1-27.
- [7] Wang R., Zhu J., Wang S., Wang T., Huang J., Zhu X. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. In: *International Journal of Multimedia Information Retrieval*, 2024
- [8] Luo Jia. Correlation analysis between PNCK and EGFR expression and EGFR gene expression status in non-small cell lung cancer, 2023.
- [9] Tan Suyu, Xie Changsheng. CLDN18.2 Progress in advanced gastric cancer. *The Journal of Oncology*, 2024, 30 (05): 407-413.
- [10] Wei Fei Ran. Study on the role and molecular mechanism of EDNRB gene in lung adenocarcinoma, 2020.