Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/17/20240704

Analysis of key genes in the development of lung cancer

Qin Yuan

Hainan International College, Communication University of China, No.1 Dingfuzhuang East Street, Chaoyang District, Beijing, China, 100024

yuanqin_phoebe2023@163.com

Abstract. Lung cancer is the cancer with the highest incidence rate. Due to the insidious early symptoms, more than 80% of lung cancer patients are diagnosed at an advanced stage. If certain key genes that have a main effect on oncogenic ability can be identified, researchers will be able to use precisely targeted therapies to predict and treat the cancer. In order to identify the key genes that play an important role in the development of lung cancer, this study collected gene expression data respectively from lung cancer patients and paraneoplastic samples, and used statistical modeling methods such as Pearson correlation analysis and significance test to screen out typical genes that have a high degree of variability in cancer and paraneoplastic samples and a high correlation with the survival time of patients. The study obtained 1487 genes that were significantly up-regulated in cancer samples and 447 significantly down-regulated genes. Additionally, these key genes were enriched for pathways relying on the GO and KEGG databases to obtain the common functions of these genes and to explore the biological mechanisms of lung cancer development.

Keywords: Lung Cancer, Gene, Differential Analysis, Correlation Analysis, Functional Enrichment.

1. Introduction

Cancer is now one of the leading causes of death globally and is a major public health problem worldwide [1] . It is estimated that in 2018, 9.6 million people died from cancer. About one in six deaths worldwide is caused by cancer. In recent years, the increasing pollution of the human living environment and people's contact with carcinogenic factors have led to a continuous increase in the incidence of cancer globally, and lung cancer is the cancer with the highest incidence and mortality rate in the world [2], with a 5-year survival rate of less than 20% [3, 4]. Lung adenocarcinoma (LUAD) is the most common histologic subtype of non-small cell lung cancer (NSCLC), accounting for approximately 40% of lung malignancies [5]. Early surgical resection is the most common and effective treatment for lung cancer at this stage, but due to the insidious symptoms of lung cancer, which are not easy to be detected, and the limitations of conventional X-ray chest examination, which is unable to detect small early lesions, more than 80% of patients with lung cancer have already been diagnosed in the middle or late stage, and have missed the optimal time for surgical treatment, and the situation is grim.

Genes are transcribed and translated to form proteins, which are involved in regulating homeostasis in living organisms as the executors of gene functions. For example, there are certain genes that can be translated into proteins that play specific roles in the regulation of cell growth, cell cycle, and DNA replication. This study is mainly based on the gene expression data of the TCGA database, using

DOI: 10.54254/2753-8818/17/20240704

differential analysis and correlation analysis to identify the key genes in the development of lung cancer, and to understand the molecular mechanism of cancer through the functional enrichment analysis of the key genes. A personalized medicine model based on the identification of key genes can accurately predict the population that will benefit from a disease prevention and treatment program, thus maximizing therapeutic efficacy and minimizing medical damage.

2. Methodology

2.1. Acquisition of gene expression profile and clinical data

The lung cancer gene expression data and clinical data used in the study were obtained from the TCGA website, a cancer research program established by the National Cancer Institute (NCI) in collaboration with the National Human Genome Research Institute (NHGRI), which collects and organizes a wide range of genomic data related to cancer and provides a large-scale reference database for cancer research, covering genomic, transcriptomic, epigenomic, and proteomic data. The data were downloaded using the R package TCGAbiolinks, which is a third-party tool recommended by the GDC, and the data in this study were downloaded through the official API of the GDC to ensure the timeliness and accuracy of the data.

2.2. Data preprocessing

The gene expression FPKM data of lung cancer patients were obtained in TCGA download, and the genome-wide G expression data of different samples N constitute a $G \times N$ data matrix M, usually $G >> N$, where each element gene expression data represents the expression level value of the gene i in sample j. Since a single patient may have had multiple genomes sequenced in TCGA, this study used the average value of the patient's genes. The gene expression profiles extracted from the TCGA database contain expression data for more than 60,000 genes, however, according to existing studies, not all genes express proteins, and some genes are not involved in protein expression and may exercise regulatory roles [6]. Therefore, this study screened the genes involved in protein expression in the gene expression profiles of lung cancer patients (Tumor group) and non-lung cancer patients (Normal group) by extracting protein-coding genes for subsequent analysis. As seen from the expression profiles, a considerable number of genes had zero expression in most of the samples, and genes with zero expression values were removed from more than 50% of the samples under each cancer type to ensure the accuracy of the results as well as their clinical significance.

2.3. Variation analysis and correlation analysis

The two main strategies used in this study to analyze differences are the multiplicative method of differences and the statistical test. The multiplicative method (fold change) is generally used to measure the gene expression level by COUNT, TPM, or FPKM, so the gene expression value must be non-negative, then the value of FOLD CHANGE belongs to $(0, +\infty)$. In order to make the degree of difference more intuitive, this study uses the $\log_2$ fold change. For example, when $\text{expr(A)} < \text{expr(B)}$, B's fold change to A is greater than 1, $\log_2$ fold change is greater than 0, and B is up-regulated relative to A; when $\text{expr(A)} > \text{expr(B)}$, B's fold change with respect to A is less than 1, $\log_2$fold change is less than 0. In this study, Equation (1) was used to determine whether the calculated genes belonged to up-regulated or down-regulated genes

$$log_2\left(\frac{avg(T)}{avg(N)}\right) \tag{1}$$

The multiplicative method does not take into account the statistical significance of differential expression, and in order to ensure the accuracy of the results, this study introduces the rank sum test to assess the significance of the results. A significant level is used to indicate the ability of groups to be distinguished from each other [7] and correlation analysis measures the closeness of the correlation

DOI: 10.54254/2753-8818/17/20240704

between the factors of the two variables [8]. In this paper, the Pearson method is used to study the correlation between gene expression and patient survival time, and the expression is as follows:

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X] \cdot Var[Y]}} \quad (2)$$

2.4. Functional enrichment

When performing differential expression analysis, a lot of differentially expressed genes can be obtained. It is difficult to find a pattern that shows the relationship between these genes if they are arranged only according to their names. In order to see the function of these genes clearly, this paper uses the enrichment analysis method. GO and KEGG are databases of gene-related functions stored based on different classification ideas [9]. GO database, the full name of which is Gene Ontology, divides the functions of genes into three parts: cellular component (CC), molecular function (MF), and biological process (biological process, BP). Through the GO database, it can know which target genes are mainly associated with the CC, MF and BP levels. Besides the annotation of the function of the genes themselves, the genes are involved in various pathways in the human body, and the database based on the human body's pathways is the KEGG database. GO term is a pure set of genes, and there is no definition of the interrelationships of the genes in it. KEGG is not only a set of genes, but also defines the complex interrelationships between genes and metabolites, which is why it is called a pathway.

3. Results Analysis

3.1. Identifying significant differences between cancer and para-cancer genes

To ensure the accuracy as well as the clinical significance of the results, genes with zero expression values in more than 50% of the samples under each cancer type were removed from this study. For each gene obtained after screening, the mean value of its expression in the Tumor and Normal groups was calculated separately, and the two means were divided to obtain the multiplicity of differences (foldchange) in the expression of individual genes in lung cancer samples versus normal samples. In order to ensure the credibility of the analysis results, so that the genes under study differed only between the lung cancer group and the Normal, rather than in both randomized groups, the study tested the significance of the difference between the expression of each gene in the Tumor group and the Normal group using the rank-sum test, and the p value of each gene was obtained. According to the $P < 0.05$ and the $\log 2(\text{Foldchange}) > 2$ criteria, were screened to obtain 1487 genes that were significantly up-regulated in cancer samples. According to the $P < 0.05$, $\log2(\text{Foldchange}) < -2$ criteria, 447 significantly down-regulated genes were selected as shown in Figure 1. The X-coordinate of the volcano plot indicates the multiplicity of differences in gene expression between lung cancer samples and normal samples, the Y-coordinate indicates the degree of significance of the genes, and the dotted line is the screening criterion of whether the genes are significant for the study. From Figure 1, it can be clearly visualized the size of the number of up-regulated genes (red dots) compared with the number of down-regulated genes (blue dots). There are significantly more up-regulated genes than down-regulated genes. This shows that the genes with abnormal expression in lung cancer patients are more often significantly up-regulated.

DOI: 10.54254/2753-8818/17/20240704

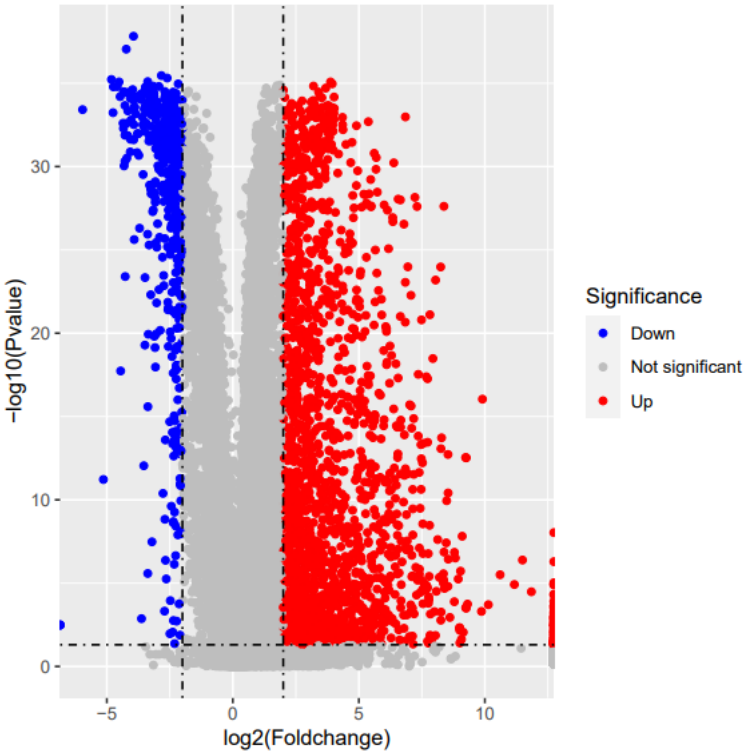

Figure 1. A Volcano Plot of the significance of genes.

3.2. Analysis of genes associated with patient survival time

The clinical data tables of lung cancer patients in the TCGA database were first read to screen out samples of deceased lung cancer patients in order to facilitate the determination of the survival time of the patients, and samples with both survival information and gene expression data were selected for subsequent correlation analysis. By analyzing the gene expression and survival time of lung cancer samples, it can measure the closeness of the correlation between the two variable factors. In this study, the expression of each protein-coding gene was analyzed in correlation with the number of days of survival of the patients, and the correlation coefficients, as well as the corresponding significance were obtained, thus the correlation coefficients and the degree of significance of each lung cancer sample can be obtained.

As shown in Table 1, through the difference analysis and correlation analysis, this study obtained important genes with high multiplicity of difference, high degree of correlation, and high significance, which can generally understand the level of difference between cancer samples and normal samples, the degree of correlation between the expression of each gene and the survival time, as well as the significance levels, so as to obtain a specific, detailed and clear data table. This is convenient to study the role of each gene. Positive correlation coefficients mean the expression of the gene is positively correlated with the number of days of survival, with the greater the expression the longer the survival. Negative correlation coefficients indicate that the gene is detrimental.

Based on the analysis of correlation and variance, the study found that the gene FAM83A was significantly different in cancer and paraneoplastic samples and was significantly associated with the survival time of patients, consistent with previous studies [10]. The expression of FAM83A in cancer and paraneoplastic samples is shown in Figure 2, where each dot corresponds to one sample, the position of the dot corresponds to the size of the gene expression in the sample, and the red dots are the mean values of FAM83A gene expression in the group of samples.

DOI: 10.54254/2753-8818/17/20240704

Table 1. A list of important genes.

	Foldchange	Pvalue	corP	corE
ENSG00000167578	0.97	0.09	0.10	0.11
ENSG00000078237	1.07	0.63	0.99	-0.08E-02
ENSG00000146083	1.41	7.86E-11	0.13	-0.10
ENSG00000158486	0.71	0.71	0.89	-0.09E-01
ENSG00000198242	1.47	2.41	0.46	-0.05
ENSG00000134108	0.86	3.85	0.05E-01	0.20
ENSG00000172137	0.96	3.36	0.35	-0.06
ENSG00000167700	2.38	4.44	0.75	-0.02
ENSG00000060642	1.21	0.06E-01	0.22	0.09

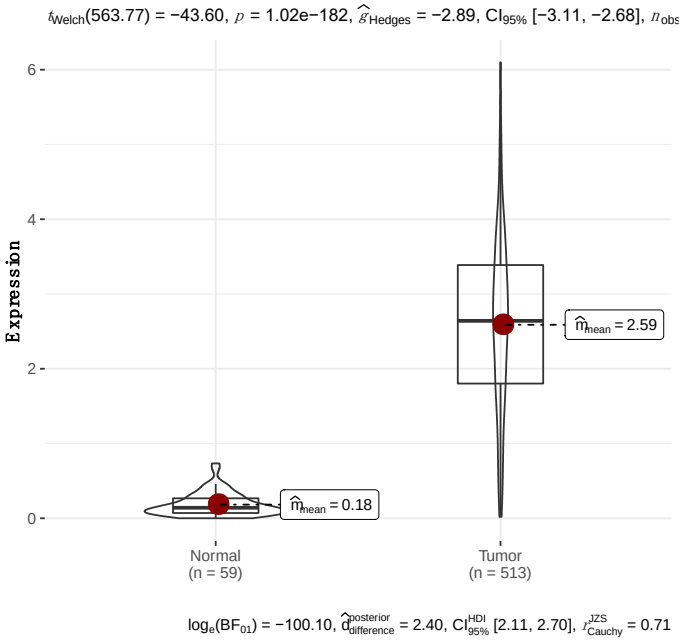

Figure 2. Box Plot of FAM83A Expression.

3.3. Functional enrichment of key genes

Through differential analysis and correlation analysis, only the expression differences of individual genes in different samples, the correlation links between individual gene expression and survival time can be obtained, and finally the differences and correlation levels between them can be obtained. However, genes in an organism exercise their functions together, and it is difficult to find a pattern to illustrate the connection between genes if these genes are presented only by gene name. Therefore, in order to systematically analyze individual genes and study the functions that genes can perform together (e.g., apoptosis) at a systematic and holistic level so as to know the relationship between diseases and gene functions instead of studying the relationship between genes and diseases at the level of individual genes, the present study was based on the two databases, GO and KEGG, for pathway enrichment of genes. The 147 genes with a high degree of difference and high correlation were screened for functional enrichment in the previously obtained pairwise total analysis table. This study mainly relied on two R packages (org.Hs.eg.db and clusterProfiler) to realize the conversion of gene ids and functional enrichment. The results are shown in Figure 3 and Figure 4.

DOI: 10.54254/2753-8818/17/20240704

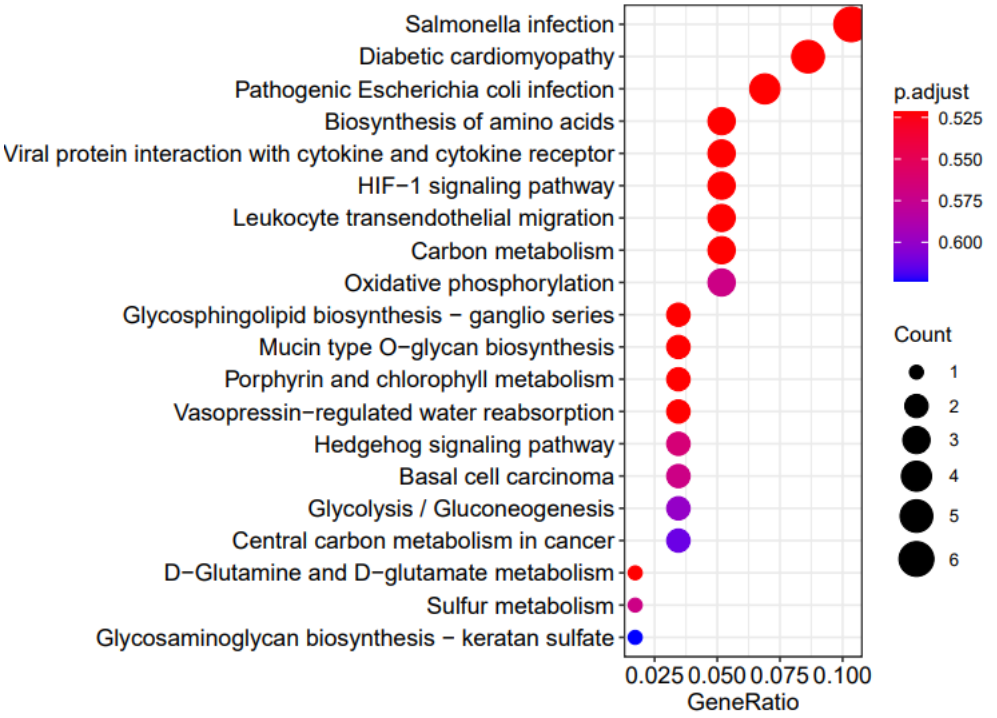

Figure 3. KEGG Enrichment Result.

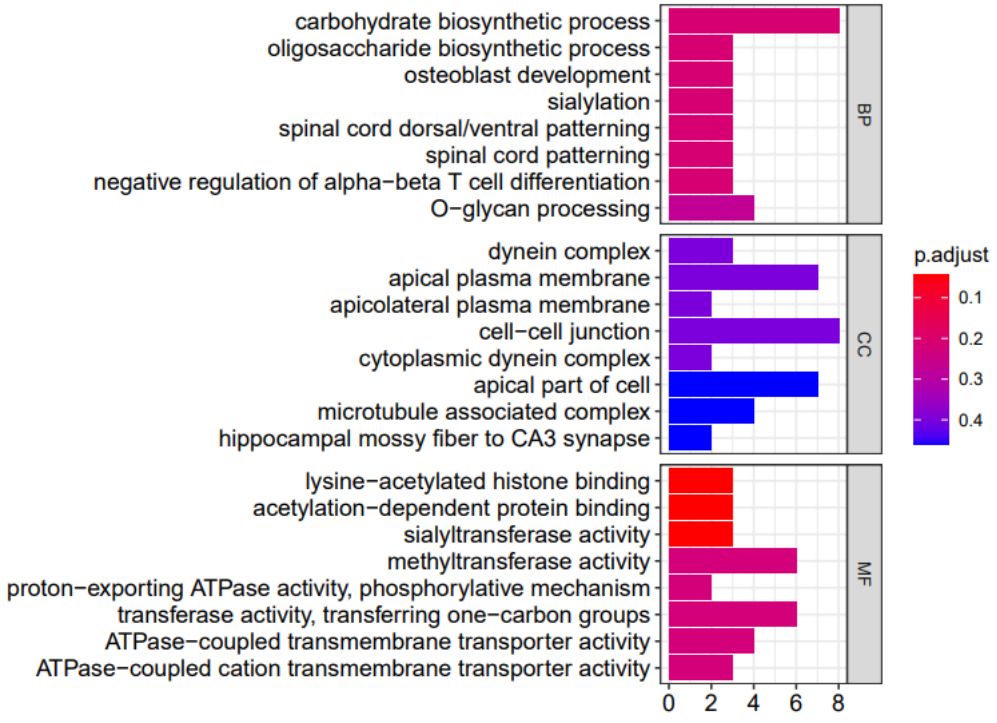

Figure 4. GO Enrichment Result.

Figure 3 shows the KEGG-based functional enrichment results, each point is an enrichment result and the redder the color of the point the smaller the p-value, which indicates its higher significance.

DOI: 10.54254/2753-8818/17/20240704

Figure 4 is the GO enrichment result. The above two figures visually present the enrichment of 147 genes identified for some immune-related, biosynthesis, and metabolic pathways, such as Leukocyte transendothelial migration, Biosynthesis of amino acids, Glycosaminoglycan Biosynthesis of amino acids, Glycosaminoglycan and Carbon metabolism pathways. Thus, it can be seen that functions such as immune and energy-related pathways greatly influence the occurrence and development of lung cancer. Taking immune function as an example, abnormal expression of genes exercising immune function leads to dysregulation of immune function and decrease of human immunity, which leads to untimely processing of cancer cells, resulting in the occurrence and deterioration of lung cancer.

4. Conclusion

This paper adopted the differential expression analysis, survival correlation analysis, and pathway enrichment of human genes, screened out 147 key lung cancer genes with high differential expression and at the same time high correlation with survival days, and analyzed the functions that these genes exercised together through pathway enrichment. It was found that there were significantly more abnormally up-regulated genes than down-regulated genes in lung cancer patients. Meanwhile, by analyzing and comparing the expression of the screened key genes, the characteristics of the expression of these key genes and their strong correlation with the survival time of patients are presented.

These findings can be applied to clinical medicine, and these genes can be used as drug targets to develop new drugs and accelerate the speed of drug development. Meanwhile, modeling can be used to predict the future survival of patients, which can greatly help cancer researchers improve the efficiency of cancer prevention, diagnosis, and treatment.

The data used in the study came from the TCGA website, whose data sources were mainly taken from European and American patients, and the data results may be somewhat different from those of Asians due to their different living backgrounds. In the future, while more clinical data on Asians is still a direction that needs to be continuously explored. Besides, the study only analyzed the correlation between gene expression and survival days, and more clinical manifestations, such as living habits, environmental factors, and other conditions, also need to be continuously explored.

References

[1] Siegel R L, Miller K D, Jemal A. Cancer statistics. (2019). CA: Cancer J Clin, 69(1), 7-34.
[2] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, CA Cancer J Clin, 71(3), 209-249.
[3] P. Saintigny, J.A. Burger. (2012). Recent advances in non-small cell lung cancer biology and clinical management, Discov Med 13(71), 287-97.
[4] R.S. Herbst, D. Morgensztern, C. Boshoff. (2018). The biology and management of non-small cell lung cancer, Nature 553(7689), 446-454.
[5] Z. Chen, C.M. Fillmore, P.S. Hammerman, C.F. Kim, K.K. Wong. (2014). Non-small-cell lung cancers: a heterogeneous set of diseases, Nat Rev Cancer 14(8), 535-46.
[6] Katherine R. Mattaini. (2020). Introduction to Molecular and Cell Biology. Regulation of gene expression, 17.
[7] Bhandari, P. (2023). An Easy Introduction to Statistical Significance (With Examples). Scribbr.
[8] Bhandari, P. (2023). Correlation Coefficient | Types, Formulas & Examples. Scribbr.
[9] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res 27(1), 29-34.
[10] Y.W. Zheng, Z.H. Li, L. Lei, C.C. Liu, Z. Wang, L.R. Fei, M.Q. Yang, W.J. Huang, H.T. Xu. (2020). FAM83A Promotes Lung Cancer Progression by Regulating the Wnt and Hippo Signaling Pathways and Indicates Poor Prognosis, Front Oncol 10, 180.