Mining Tcga Database to Screen Genes Valuable for Prognosis of Lung Cancer Microenvironment

Fuyong Bian*, Ming Zhang

Chuxiong Medical College, Chuxiong, Yunnan, 675005, China

Keywords: Tcga database, Microenvironmental prognosis of lung cancer, Transcription of genes, Mrna

Abstract: Objective: To screen lncRNA related to microenvironment prognosis of lung cancer by mining TCGA database. Methods: 230 lung adenocarcinoma patients with both clinical prognosis information and methylation data were included through data connection using lung adenocarcinoma prognosis data downloaded from TCGA website and whole genome methylation data based on Illumina Methylation 450 chip. The R language was used for data combination, standardization and difference analysis, and SPSS 20.0 software was used for data analysis of differentially expressed genes to screen out specifically expressed transcription genes significantly related to lung cancer survival. Results Spearman rank correlation analysis screened out differential genes related to prognosis, and Kaplan Meier performed Log-rank test to screen out 2 genes closely related to prognosis of lung cancer patients. Cox multivariate regression analysis showed that DLX6 was more specific. Conclusion: Through the excavation of TCGA database in this study, it is preliminarily found that the methylation level of the methylation site of KRI1 gene has an impact on the prognosis of lung adenocarcinoma, which can be used as a biomarker for further study of lung cancer prognosis.

1. Introduction

At present, lung cancer is one of the most common malignant tumors in human beings, and its morbidity and mortality are gradually increasing [1]. At present, researches believe that the occurrence and development of lung cancer is a process of multi-factors, multi-stages and multi-gene changes, and methylation inactivation of tumor suppressor genes is one of its important mechanisms [2]. Lung squamous cell carcinoma is the most common histologic subtype of lung cancer, and its morbidity and mortality are the highest among malignant tumors [3]. Due to the lack of early screening methods with high sensitivity at present, many lung cancer patients have progressed to the middle or even advanced stage at the time of treatment, resulting in a 5-year survival rate of less than 20% for lung cancer patients [4]. Improving the prognosis of lung cancer patients [5] is the key factor to improve the survival rate of lung cancer patients. Cancer Genome Map database is currently the largest tumor gene chip database and integrated data extraction platform in the world. Its integrated documents and chip data are widely recognized by the academic community with high quality. To explore the correlation between methylation in the whole genome and prognosis of lung adenocarcinoma, and find the mRNA related to methylation, and then the genes affecting prognosis of lung adenocarcinoma, so as to provide scientific basis for the research of prognosis-related markers of lung adenocarcinoma in the future.

2. Materials and Methods

2.1 Data Acquisition

The database used in this study is the clinical prognosis data of lung adenocarcinoma downloaded from TCGA website and the data of whole genome methylation at level3 detected by Methylation450 chip of Ill-umina Company in the United States. Through data connection, 230 lung adenocarcinoma patients with both clinical prognosis information and methylation data are

Copyright © (2020) Francis Academic Press, UK DOI: 10.25236/wccbm.2020.008

retained. The aim is to map the genomes of various tumors and their subtypes at various histological levels to explore the biological mechanisms in the process of tumor formation and development, and finally achieve the goal of improving the current situation of tumor prevention and treatment [6].

2.2 Screening of Differentially Expressed Genes

Using R language, TCGA data are merged and standardized, and then SPSS 20.0 software is used for statistical analysis. limma package is used to screen differentially expressed RNAs. On the detection platform, the Beta value corresponding to each sample will be mapped to the genome (methylation site/gene) and DNA methylation analysis will be performed. Then the Infinium II probe judges the methylation level of each base according to the fluorescence color of the base. If the DNA treated with bisulfite matches G base, the probe will show green fluorescence. If it can match with base A, it will show red fluorescence. Finally, 3731 lncRNA sites in cancer tissues and adjacent tissues of 45 lung cancer patients were used for differential expression analysis.

2.3 Screening of Genes Related to Microenvironment Prognosis of Lung Cancer

Spearman rank correlation analysis was performed on the expression value and survival time of differentially expressed RNAs, and the corresponding Spearman correlation coefficient and P value were calculated. Then the median value of each RNAs expression value of Spearman rank correlation analysis $P <0.05$ was calculated. The beta value of the methylation site corresponding to each sample will be mapped to the corresponding genome (the gene where the methylation site is located), and DNA methylation analysis will be performed. Different from the molecular model of a single marker, molecular tags not only take the function of a single gene as the research basis, but also pay more attention to the common coordination between genes, describing a specific biological characteristic from the overall and system levels [7]. The expression value of each RNAs greater than the respective median is defined as high expression, otherwise it is defined as low expression. thus, each RNAs is divided into two groups of high expression and low expression. KaplanMeie is used for Log-rank test, with $p <0.05$ having statistical significance.

2.4 Statistical Method

Continuous variables are expressed as mean ± standard deviation (± s), and subtype variables are expressed as sample size (composition ratio). The t-test method was used to compare the differences between the two samples. Analysis of the differences between the count data using χ 2 test or exact probability calculation method. Cox multivariate regression was performed on the screened gene expression data to analyze its expression in other types of tumors, and correlation coefficient and p value were calculated, with $p <0.05$ having statistical significance.

3. Result

3.1 Differential Expression Analysis

Limm package (r language) carries out differential expression analysis on RNAs count data of normal tissues and tumor tissues | log2fold-change | > 1. five variables including age, gender, race, smoking and pathological staging of lung adenocarcinoma are adjusted in the model. Methylation sites were sorted according to the p value of Cox survival analysis results from small to large, and the top 20 methylation sites with the lowest p value were selected. According to the characteristics of sequencing technology, the quality of sequencing fragment ends will be low, and the quality of large fragment library sequencing is lower than that of small fragment library sequencing. The difference was statistically significant ($p < 0.05$). the genes with significantly different expression were identified and volcano maps were drawn (Figure 1), from which transcription genes with differentially expressed Top100 were screened.

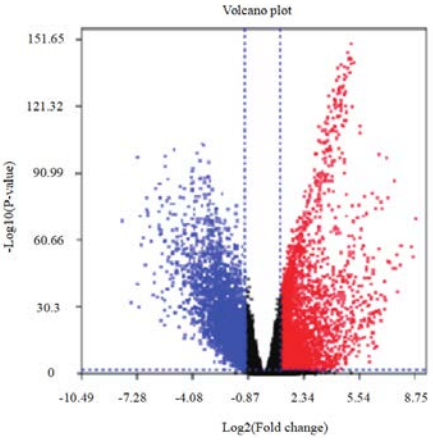

Fig.1 Volcanogram of Differentially Expressed Genes in Lung Cancer

3.2 Spearman Rank Correlation Analysis to Screen Rnas

Prognostic analysis of cg03955927 located at the methylation site of gene loc100132215/ehbp1 and lung adenocarcinoma showed that the p value was the smallest (p = 1.98× 10-7), and the corresponding HR value was 0.605 (0.501 ~ 0.731). Due to the random primers used in sequencing Illumina transcriptome, the GC content at the front end of the sequencing fragment is preferred. This fluctuation is normal and the normal GC content is about 50%. Spearman rank correlation analysis revealed that the expression of only four transcription genes FOXE1, DLX6, WIF1, TMPRSS11A were positively correlated with survival time.

3.3 Lncrna Differentially Expressed in Cancerous Tissue and Paracancerous Tissue

Paired T test was carried out on lncRNA sites in cancer and adjacent tissues of 45 patients with lung cancer. After analysis, 322 lncRNA with FDR&lt;0.05 and absolute value of multiple change ⩾3 were found, and all lncRNA were up-regulated in cancer tissues. Each column represents the tissue of a lung adenocarcinoma patient. Red indicates that the methylation level of the corresponding tissue sample increases, green indicates that the methylation level decreases, and the depth of color indicates the degree of difference between the methylation level in cancer and adjacent tissues. HR values corresponding to 17 methylation sites ranged from 0.61 to 0.65, suggesting that the hypermethylation level of these methylation sites may be a protective factor for the prognosis of lung adenocarcinoma. There were 47 studies on the results of statistical differences in SLC2A1 expression, of which 43 were increased and 4 were decreased. There are 8 studies with high expression and 0 studies with low expression in lung cancer. Unsupervised clustering analysis was carried out on differentially expressed lncRNA, and the results were presented in the form of a heat map, in which red indicates cancerous tissue and blue indicates paracancerous tissue, as shown in Figure 2.

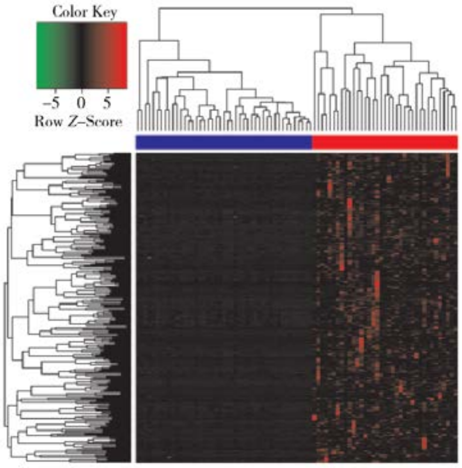

Fig.2 Analysis Results of Lncrna Unsupervised Clustering Thermogram

3.4 Expression Analysis of Dlx6 and Tmprss11a in Other Tumor Tissues

The GO enrichment analysis was carried out on the genes of 1291 methylation sites with significantly different methylation levels in tumor tissues and adjacent tissues of 25 lung adenocarcinoma patients. The biological functions of genes with different methylation sites were analyzed from three aspects of molecular composition, participation in cell processes and biological functions. The Oncolnc database was used to analyze the expression data of DLX6 and TMPRSS11A in 29 kinds of tumors and normal tissues. DLX6 was significantly different in renal clear cell carcinoma (KIRC) and (lung cancer), and the difference was statistically significant (p <0.05). The expression level of SLC2A1 gene in lung adenocarcinoma is about 8.37 times that of normal lung tissue. In order to verify this trend through clinical pathological analysis, the immunohistochemical staining results of the human protein mapping project were detected, and SLC2A1 was found to be strongly or moderately expressed in lung adenocarcinoma tissues, but weakly or negatively expressed in normal tissues. However, there was no significant difference in TMPRSS11A. The results show that DLX6 is more specific as a marker of lung cancer. See Tables 1 and 2.

Table 1 Results of Dlx6 Multivariate Cox Regression Analysis

Tumor type	Cox regression coefficient	P	FDR value	Median	Average value
KIRC	0.231	0.007	0.012	6.33	8.37
Lung cancer	-0138	0.003	0.721	108.34	166.24
STAD	0.117	0.1	0.275	3.62	33.28
LGG	-0.031	0.27	0.427	10.18	30.16
READ	-0.216	0.36	0.831	9.96	28.26
ESCA	0.716	0.71	0.772	7.14	29.37
UCEC	-0.004	0.51	0.382	8.47	70.14

Table 2 the Incidence Of Adverse Reactions between the Two Groups Was Compared (n,%)

Group	Quantity	Adverse reaction			Incidence of adverse reactions
		Vomiting	Diarrhea	Alanine aminotransferase(↑)	
Research group	115	3	4	2	7.41%
Control group	115	2	2	1	11.33%
P					>0.05

According to the above observation, the incidence rate of adverse reactions in the study group was 7.14%, while that in the control group was 11.43%. the difference between the two groups was statistically significant (p <0.05).

4. Discuss

In this study, through the excavation of TCGA database, 230 methylation sites with the strongest correlation with the prognosis of lung adenocarcinoma were initially found. Based on this, the correlation between these methylation sites and mRNA expression and the influence of mRNA expression level of corresponding genes on the prognosis of lung adenocarcinoma were analyzed. Previous studies have shown that, The dysfunction of many tumor suppressor genes and carcinogenic genes and the structural abnormalities of their products are involved in the process of the occurrence and development of lung cancer [8], but the exact mechanism is still unclear. Therefore, further exploration of etiological factors, molecular mechanisms and pathways of lung cancer is of great significance for optimizing diagnosis and treatment. COX regression analysis results show that the level of signature composed of three methylation sites in different models has significant correlation with the prognosis of lung adenocarcinoma. With the increase of methylation signature level, the death risk of lung adenocarcinoma patients gradually increases. When the population is divided into high expression group and low expression group according to the median of lncRNA molecular tags, the death risk of patients in high expression group is 2.14 times that of lung cancer patients in low expression group. Therefore, the higher the malignant degree of the tumor, the stronger its proliferative ability, and the stronger its glucose and glycolytic ability, the higher the abnormal expression of GLUT1 will be, which is also the case in lung adenocarcinoma [9]. Data analysis results of lung adenocarcinoma patients in TCGA database are similar. From the above results, it can be seen that the prediction model based on methylated signature has good performance and can effectively predict the prognosis of lung adenocarcinoma patients.

DNA methylation mainly regulates gene expression by methylation modification of cytosine of CpG sequence. Hypermethylation of CpG island of gene often leads to gene transcription silencing, which makes tumor suppressor genes, oncogenes and DNA regulatory genes lose their functions and causes chromosome instability, thus causing abnormal regulation of growth and differentiation of normal cells or tumor cells, leading to or inhibiting the occurrence of tumors [10]. Pyrosequencing is a new DNA sequence detection technology developed by Ronaghi et al. in 1987. after many optimizations, pyrosequencing has become an automated sequencing technology with high throughput, high stability and quantitative detection. The high expression of ncRNA KTN1-AS1 is a risk factor for the prognosis of head and neck squamous cell carcinoma, and the 3-lncRNA marker constructed by ncrnaktn1-as1 can better predict the survival of patients. In this paper, the method of data mining is used to analyze the expression of SLC2A1 gene and the survival data of corresponding patients in a large sample composed of different researchers collected in Oncomine and TCGA databases. Studies have found that hypermethylation levels of several genes can improve the prognosis of lung adenocarcinoma, and relevant functional studies support this correlation. For example, EHBP1 is EH domain binding protein 1, which contains 11 transcription factor binding sites and can regulate actin dynamics and clathrin-mediated endocytosis [11]. During DNA sequence detection, fluorescence labeling or electrophoresis is not required, and the base sequence can be directly detected, and the result is accurate, so that large sample DNA sequence can be quickly, quantitatively and efficiently detected. The high expression of LncRNA CTD-2555C10.3 is a prognostic risk factor for lung adenocarcinoma. The 7-dimensional transcriptome molecular tag constructed by LNC RNA CTD-2555C10.3 is a good indicator for the prognosis of lung [20] adenocarcinoma.

In order to further clarify the relationship between methylation levels of three methylation sites and gene expression, RT-PCR was used to detect gene expression levels in tumor tissues and adjacent tissues of 30 lung adenocarcinoma patients. Cg12013757, a new methylation site found in this study, is located in the KRI1 gene region. This gene is a key factor in the biosynthesis of 40S ribosome and is also a necessary gene to maintain cell activity [12]. In this study, LASSO Cox regression was used to screen lncRNA markers related to the microenvironment prognosis of lung cancer, avoiding multicollinearity caused by sample size far smaller than independent variables and reducing Class II errors. 59 cases of normal lung tissue and 515 cases of lung adenocarcinoma were screened from TCGA database. Statistical analysis showed that the expression level of SLC2A1 gene in lung adenocarcinoma was about 8.37 times of that in normal lung tissue. The results of the study on the relationship between the methylation level of methylation site and the expression of its gene show that there is a significant correlation between the methylation level of methylation site cg15386964 and the expression level of its gene HIST1H2BH. Some studies have shown that when the expression of KRI1 is shut down through experiments, the synthesis of polysomes and 40S ribosomes is reduced, suggesting that abnormal expression of KRI1 may have a greater impact on cell activity. It is suggested that methylation sites may affect the expression of genes or proteins related to lung adenocarcinoma by methylation of genes and changing gene expression levels, further affecting the prognosis of lung adenocarcinoma patients.

Of course, this study also has certain limitations. The prediction model constructed in this study only considers the expression of lncRNA, and does not consider the influence of other levels of biomarkers on the prognosis of lung cancer patients. There are inevitable statistical method errors in the data model, and false positives and false negatives may occur to a certain extent in the screening process. Therefore, similar data need to be supplemented and verified by subsequent further experiments, so as to provide more reliable basis for the screening and research of lung cancer-related molecular markers.

5. Conclusion

In a word, in this study, the existing TCGA database was used to preliminarily excavate the methylation sites related to lung adenocarcinoma prognosis at the whole genome level, and some new methylation sites related to lung adenocarcinoma prognosis were found. Through the excavation of TCGA transcriptome sequencing database, the lncRNA molecular tags screened and the prognosis prediction model constructed by LNC RNA molecular tags and clinical variables have better prediction value for the prognosis of lung cancer microenvironment, and have certain theoretical guiding significance for the prediction research of lung cancer prognosis in the future.

Acknowledgment

Scientific research fund project of Yunnan Provincial Education Department in 2019: Mining TC GA database to screen genes that are valuable for prognosis of lung cancer microenvironment (2019 J0413).

References

[1] Kou Ruihuan, Xie Yuning, Li Jiaying, et al. (2019). Relationship between cyclin 6 copy number variation and prognosis in patients with lung adenocarcinoma. Cancer Progress, no. 19, pp. 2259-2263.

[2] Wang Jin, Yu Xiaofan, Ouyang Nan, et al. (2019,). Methylation regulates the expression of SLIT3 and SPARCL1 genes in smoking-induced lung adenocarcinoma and its effect on patient prognosis. Chinese Journal of Medical Sciences,vol. 99, no. 20, pp. 1553-1557.

[3] Xiaomei Yang, Bi Yuxue. (2019). Screening of prognostic RNAs related to LUSC based on TCGA database. Electronic Journal of Clinical Medical Literature, vol. 6, no. 63, pp. 52-53.

[4] Ma Bangjing, Hao Xianglin, Han Fei, et al. (2018). SOX3O gene regulates desmosome gene DSC3 to inhibit lung adenocarcinoma proliferation and migration. Carcinogenesis. Distortion. Mutation, no. 3, pp. 165-170.

[5] Hu Xi, Mu Yudong, Wu Xiaoming, et al. (2018). Feature recognition of non-small cell lung cancer subtypes based on omics data analysis. Journal of Modern Oncology, no. 3, pp. 375-378.

[6] Wang Haijun, Chen Qiuyue, Song Na. (2019). Application of Oncomine Database in Cancer Research. Chinese Journal of Biochemistry and Molecular Biology, no. 10, pp. 1051-1057.

[7] Wang Wenjing, Wang Xiaodi, Liu Liying, et al. (2019). Abnormal expression of miR-338-3p in malignant tumors and its epigenetic histone modification sites. Journal of Xi'an Jiaotong University (Medical Edition), no. 05, pp. 706-710.

[8] Sun Lichun, Li Dandan, Li Jing, et al. (2019). Expressions and clinical significance of CCDC8 and TGF-β1 in non-small cell lung cancer. Journal of Practical Oncology, no. 3, pp. 215-221.

[9] Zhang Lanjun. (2016). Drug resistance in the treatment of non-small cell lung cancer cannot be underestimated. Chinese Medical Information Review, no. 6, pp. 16-16.

[10] Lin Kang, Pan Bei, Xueni Xu, et al. (2018). Construction of microRNAs risk model related to prognosis of lung adenocarcinoma based on TCGA database. Chinese Journal of Clinical Laboratory Management Electronics, vol. 6, no. 02, pp. 89-98 .

[11] Xiao Jinrong, Wang Ke, Liu Ying, et al. (2019). Mining long-chain non-coding RNA molecular tags related to prognosis of hepatocellular carcinoma based on public database. Chinese Journal of Epidemiology, vol. 40, no. 7, pp. 805- 809.

[12] Sun Meitao, Zi Jiaji, Chen Ying, et al. (2018). Prognostic significance of HIF-1α in gastric cancer using data mining technology. Journal of Dali University, no. 2, pp. 32-37.