Identification of genes associated with lung cancer by bioinformatics analysis

J. Li¹, H. Yu¹, Y.-F. Ma¹, M. Zhao¹, J. Tang²

¹Department of Thoracic Surgery, General Hospital of People’s Liberation Army, Beijing, China
²Department of Cardio-thoracic Surgery, First Affiliated Hospital of Chinese PLA General Hospital, Beijing, China

Abstract. – OBJECTIVE: This study was aimed to explore the underlying genes associated with lung cancer (LC) by bioinformatics analysis.

DATA AND METHODS: Gene expression profile GSE2514 was downloaded from the Gene Expression Omnibus database. Twenty lung and nineteen para-carcinoma tissue samples were used to identify the differentially expressed genes (DEGs) by paired t-test. Pathway enrichment analysis of DEGs was performed, followed by the construction of protein-protein interaction (PPI) network. Functional enrichment analysis of the module identified from PPI network was performed, and the enriched term with the highest enrichment scores was selected for pathway enrichment analysis.

RESULTS: Total 257 DEGs including 179 up-regulated DEGs such as monoamine oxidase A (MAOA) and intercellular adhesion molecule 2 (ICAM2), and 78 down-regulated DEGs such as thrombospondin-2 (THBS2) were identified. Up-regulated DEGs were enriched in 7 pathways, such as drug metabolism, tyrosine metabolism and cell adhesion molecules (CAMs). Down-regulated DEGs were enriched in extracellular matrix receptor interaction and focal adhesion pathways. In the PPI network, interleukin-6 (IL6) had the highest connectivity degree of 39. Module 1 with the highest functional enrichment scores of 5.457 containing 13 hub genes such as KIAA0101.

CONCLUSIONS: DEGs of LC were mainly enriched in the pathways related to metabolism and cell adhesion. The DEGs such as MAOA, ICAM2, IL6, THBS2 and KIAA0101 may be the potential targets for LC diagnosis and treatment.

Key Words:
Lung cancer, Molecular mechanism, Differentially expressed genes, Pathway enrichment analysis.

Introduction

Lung cancer (LC) is one of the most common fatal malignancies in economically developed countries and the second leading cause of death in developing countries¹. More than 1 million people die of LC annually². Although diagnosis at an early stage is increasing with the introduction of some novel technologies, LC is still a devastating disease which has a very poor prognosis³. The poor prognosis for LC patients is generally owing to the inability to treat the resultant late stage of it⁴.

Numerous studies have made contributions to explore the mechanism and therapeutic method for LC. Achievements have been obtained in exploring the pathological mechanism underlying the LC development. Some genes and pathways have been found to play key roles in the LC progression. For example, somatic mutations in p53 gene play an important role in the pathogenesis of early stage LC⁵. Amplified and increased expression of the v-myc avian myelocytomatosis viral oncogene homolog (myc) family of protooncogenes has been described to be associated with rapid proliferation in LC cells⁶. The dysregulation or deletion of one gene might affect the function of many other genes since there are complex interactions among various genes, their effectors, and the pathways in which they involve⁷. In addition, pathways have been identified to control important aspects of DNA repair and genomic stability⁸. At present, some critical pathways have been found to be related to LC, such as mitogen-activated protein kinase kinase and phosphatidylinositol 3-kinase pathway⁹. The development of gene expression profiling technology¹⁰ is becoming more and more important to our understanding of LC biology. Although progresses have been achieved about the pathogenesis of LC, the genetic mechanisms of LC are far from being clear.

In the present study, we downloaded GSE2514 and identified the differentially expressed genes (DEGs) between the LC and normal donors (ND)
samples to explore the molecular mechanisms of LC. Besides, we performed pathway enrichment analysis, protein-protein interaction (PPI) networks and module analysis to study and identify the target genes for diagnosis and treatment of LC. Findings of this study might play important roles in lung tumor genesis and may potentially serve as biomarkers in both diagnosis and prognosis of human LC.

**Data and Methods**

**Affymetrix Microarray Data**

The microarray expression profile dataset GSE25141 was downloaded from National Center of Biotechnology Information (NCBI) Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) database which is based on the GPL8300 [HG_U95Av2] Affymetrix Human Genome U95A Version 2 Array platform. The dataset contains 39 samples including 20 lung and 19 para-carcinoma tissue samples. In this paper, the samples were analyzed by bioinformatics.

**Data Preprocessing and Differential Expression Analysis**

The original expression data in CEL files were firstly converted into expression measures and the quartile data were normalized by robust multi-array average (RMA) in R Affy package. Then the samples were divided into two groups: LC and ND. The paired t-test based on the limma package in R was used to identify DEGs in LC and ND samples. Then log2-fold change (logFC) was calculated. Only genes with logFCI > 1.0 and adjusted p-value < 0.05 were regarded as DEGs.

**Pathway Enrichment Analysis of DEGs**

The Database for Annotation, Visualization and Integrated Discovery (DAVID) that consists of an integrated biological knowledgebase and analytic tools is used for systematic and integrative analysis of large gene protein lists. Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of online databases used for dealing with genomes, enzymatic pathways and biological chemicals. KEGG pathway enrichment analysis for the identified DEGs was performed by DAVID online tool with the threshold of p-value < 0.05.

**Protein-Protein Interactions Analysis and Network Construction**

The Search Tool for the Retrieval of Interacting Genes (STRING) database is a precomputed global resource which has been designed to evaluate the PPI information. In this paper, the STRING online tool was applied to analyze the PPI of DEGs and only those experimentally validated interactions with a combined score > 0.4 was selected as significant.

Cytoscape is a general bioinformatics package used for visualizing biological network and integrating data. PPI networks were constructed using the Cytoscape software based on the PPI relationships obtained. From the previous studies on biological network obtained, most of the PPI networks obeyed the scale-free attribution. So

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up-regulated DEGs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa04610: Complement and coagulation cascades</td>
<td>8</td>
<td>7.7E-05</td>
</tr>
<tr>
<td>hsa04270: Vascular smooth muscle contraction</td>
<td>9</td>
<td>2.8E-04</td>
</tr>
<tr>
<td>hsa00982: Drug metabolism</td>
<td>5</td>
<td>1.5E-02</td>
</tr>
<tr>
<td>hsa04514: Cell adhesion molecules (CAMs)</td>
<td>7</td>
<td>1.5E-02</td>
</tr>
<tr>
<td>hsa05416: Viral myocarditis</td>
<td>5</td>
<td>2.3E-02</td>
</tr>
<tr>
<td>hsa00350: Tyrosine metabolism</td>
<td>4</td>
<td>2.9E-02</td>
</tr>
<tr>
<td>hsa00360: Phenylalanine metabolism</td>
<td>3</td>
<td>4.4E-02</td>
</tr>
<tr>
<td>Down-regulated DEGs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsa04512: ECM-receptor interaction</td>
<td>3</td>
<td>1.7E-03</td>
</tr>
<tr>
<td>hsa04510: Focal adhesion</td>
<td>7</td>
<td>3.6E-02</td>
</tr>
</tbody>
</table>

DEGs stands for differentially expressed genes; Count stands for the number of enriched DEGs.

Table 1. The enriched KEGG pathways for the up- and down-regulated DEGs.
Identification of genes associated with lung cancer by bioinformatics analysis

networks with 169 nodes and 448 edges were obtained (Figure 1). The genes interleukin-6 (IL6) (degree = 39), FBJ murine osteosarcoma viral oncogene homolog (FOS) (degree = 28), matrix metallopeptidase 9 (degree = 24), coagulation factor III (degree = 17), von Willebrand factor (degree = 17) and KIAA0101 (degree = 16) were selected as top 6 hub nodes (Table II).

Module Analysis
Module 1 had 13 nodes and 70 edges (Figure 2). After GO functional enrichment analysis of module 1, three enriched terms were obtained and their functional enrichment scores were 5.457, 2.384 and 1.998. The enriched terms with the highest enrichment score (5.457) was related to cell cycle. The other two enriched terms were related to nucleotide binding and cellular component respectively (Table III).

Discussion
The analysis of gene expression profile revealed the underlying gene activity changes related to LC and enabled the identification of targets for management policy. In the present study, a total of 257 DEGs were identified between LC and ND samples through gene expression profile of GSE2514. The up-regulated DEGs were enriched in the pathways related to metabolism and cell adhesion, such as drug metabolism, tyrosine metabolism and CAMs, while the down-regulated DEGs were enriched in ECM receptor interaction and focal adhesion. In the PPI network, IL6 had the highest connectivity degree. Module 1 contained 13 hub genes such as KIAA0101. This result suggested that these genes and pathways may play an important role in the progression of LC.

A previous study of the RNA-seq analysis showed that the DEGs between lung cancer smoker and normal controls were mainly involved in cellular processes, which suggested that smoking might influence the normal cellular process and contributed to lung cancer progres-
Another microarray data analysis for lung cancer indicated that DEGs were enriched in functions such as cell cycle, DNA replication and immune function. In our work, we obtained the similar results that cell cycle was the significant GO term that represented by DEGs in modular 1. Thus, the normal cell cycle was dysregulated in the development of lung cancer.

Additionally, it has been confirmed that genes and proteins involved in cellular metabolism played a crucial role in the development and progression of cancer. In this study, MAOs enriching in several metabolism signaling pathways such as drug metabolism, tyrosine metabolism were up-regulated. MAOs are flavoprotein enzymes located in the mitochondrial outer membranes. Their substrates include dietary amines and neurotransmitters which are biogenic monoamines. Recent evidence supports that neurotransmitters can influence tumor growth and progression. The enzymes have 2 forms, MAOA and MAOB. MAOA preferentially oxidizes serotonin and norepinephrine, whereas MAOB has a greater affinity for benzyamine and phenylethylamine. Dopamine is a common substrate of MAOA and MAOB. Dopamine can inhibit tumour growth by interfering with vascular endothelial growth factor signaling in endothelial cells and blocking its angiogenic functions. MAOA produces a by-product, hydrogen peroxide which is a major source of reactive oxygen species (ROS). ROS can causes DNA damage and tumor initiation and progression. Kaminski et al investigated MAOs in serum of patients.
Identification of genes associated with lung cancer by bioinformatics analysis

We identified genes associated with lung cancer (LC) through bioinformatics analysis and found that the MAOs activities of serum increased notably in some cases of LC. We also found that up-regulated DEGs were enriched in the CAMs pathways. CAMs belonging to the integrin, cadherin and immunoglobulin superfamilies play a key role in cell-cell adhesion, and a number of these molecules have been implicated in cancer progression and metastasis. The cellular processes are modulated by the interaction of cells with each other and with their microenvironment. In normal tissue, CAM expression is tightly regulated. However, deregulation of CAMs disrupts normal cell-cell and cell-matrix interactions, which results in freeing cells from normal checkpoints and constraints, and facilitating tumour formation and metastasis. In this study, ICAM2 enriched in CAMs signaling pathway was up-regulated. ICAM2 is a member of immunoglobulin families. Another member, ICAM1 has been reported to be associated with primary tumors and metastases and interfering with its expression inhibits experimental metastasis by melanomas in nude mice. ICAM2 has structural and functional homology to ICAM1. As a cell surface adhesion protein involved in leukocyte recruitment, ICAM2 is expressed at low levels on most leukocytes including T and B lymphocytes. Recent gene expression profiling indicated that patients with large B cell lymphomas presented an elevated expression of ICAM2.

In this study, down regulated DEGs such as thrombospondin-2 (THBS2) were found to enrich in the ECM-receptor interaction pathways. ECM is a complex structure formed by distinct molecular networks which interact with specific cell receptors. It has been demonstrated to play a crucial role in cell proliferation, migration, differentiation, apoptosis as well as carcinogenesis. The protein encoded by THBS2 has been shown to mediate cell-to-cell and cell-to-matrix interactions and inhibit the angiogenic activity, mitogenesis and formation of focal adhesions in endothelial cells. This protein has been shown to function as...
a potent inhibitor of tumor growth and angiogenesis. Alvarez et al. suggested that the reduction of THBS2 expression may be coupled with the development of a malignant phenotype. Czekierdowsk et al. reported the down-regulated expression of THBS2 gene in ovarian cancer. In conclusion, metabolism, cell adhesion and ECM-receptor interaction signaling pathways were closely associated with LC. As a result, DEGs related to these pathways may be used as potential targets for LC treatment.

A recent study for screening lung cancer related genes with RNA-seq indicated that IL6 was a significant DEG between smokers with lung cancer and normal smokers. It is consistent with our results that IL6 as a significant DEG was also a hub gene with the highest connectivity degree of 39 in PPI network. IL6 is a multifunctional cytokine which functions in defense response, cell survival, proliferation and apoptosis. It is reported that IL6 and its major effector signal transducer and activator of transcription 3 play a central role in the epigenetic switch from non-transformed epithelia to cancer cells. Yanagawa et al. found that patients with LC have elevated levels of IL6 and its expression is related with poor prognosis of lung cancer patients. IL6 is expressed by LC cells and acts in an autocrine or paracrine mechanisms to stimulate cancer cell proliferation and migration. Taken together, these data suggest that IL6 may be a candidate molecular marker associated with LC.

In addition, we also found KIAA0101 was a hub gene in the PPI network with a higher connectivity degree of 16. Module analysis of the PPI network revealed that KIAA0101 was also enriched in module 1. KIAA0101 is a proliferating cell nuclear antigen (PCNA) associated factor and was identified as p15 in 2001. PCNA is an essential scaffold molecule for DNA repair through interactions with several DNA replication proteins. In mammalian cells, the over-expression of KIAA0101 is found to protect cells from UV-induced cell death, which supports that KIAA0101 participates in the regulation of DNA repair, apoptosis and cell cycle progression. In recent years, several reports have demonstrated that the over-expression of KIAA0101 occurred in a variety of human malignancies including LC and involved in tumor progression. Cancer patients with elevated KIAA0101 expression show a higher recurrence rate and significantly declined survival rate. The over-expressed KIAA0101 has the potential to facilitate cell proliferation under both physiological conditions and human cancers, such as LC and thyroid carcinoma. Besides, the elevated KIAA0101 level was concerned with more advanced tumor stages so the assessment of KIAA0101 expression may be helpful in determining the most appropriate duration and intensity of treatment.

Conclusions

Our data provide a comprehensive bioinformatics analysis of DEGs which may be involved in LC. The findings in current work may contribute to our understanding of the underlying molecular mechanisms of LC. DEGs such as MAOA, ICAM2, THBS2, IL6 and KIAA0101 have the potential to be used as targets for LC diagnosis and treatment.

However, this study has a few limitations. In the process of data analysis, only a few cases of adenocarcinoma of lung and para-carcinomatissues samples based on one platform were used for the analyses. This may cause a high rate of false positive result. Further genetic investigations with larger sample size are still needed to confirm our observation.

Conflict of Interest

The Authors declare that they have no conflict of interests.

References

5) CHIBA I, TAKAHASHI T, NAU M, D’AMICO D, CURIEL D, MITSUDOMI T, BUCHHAGEN D, CARBONE D, PANTADOSI S, KOGA H, ET AL. Mutations in the p53 gene are


39) THBS H, Kit TE. Mybiosource. Com.


