Abstract. – BACKGROUND: Acute idiopathic pulmonary fibrosis (IPF) is a serious and progressive form of lung disease, and millions of people suffer from this disease in the world. To provide clues for getting a better understanding of the mechanism of this disease, we identified and further analyzed the differential expressed genes in IPF.

METHOD: In this study, we downloaded the gene expression microarray (GSE10667) from Gene Expression Omnibus (GEO) database. The dataset contained a total of 23 samples, including 15 normal controls and 8 diseases samples (IPF). Then, we identified the differentially expressed genes between normal and disease samples with packages in R language. Consequently, the PPI network was also constructed for the products of these DEGs, and modules in the network were analyzed by Cytoscape’s plugin Mcode and Bingo. Furthermore, enrichment analysis was performed by DAVID to illustrate the altered pathways in IPF. The drug compounds for PLK1 were screened in DrugBank.

RESULTS: A total of 349 genes were identified as differentially expressed genes between normal and disease samples, and we constructed a protein-protein interaction network which included 200 pairs of proteins. Then three modules were identified in our network. Function of these modules were predicted to be related to protein kinase binding, extracellular matrix structural and structural constituent of cytoskeleton, respectively. Finally, we focused on module A including 18 DEGs.

CONCLUSIONS: PLK1 (Polo like kinge-1) in this module was predicted as a marker gene in IPF, which was related to cell cycle pathway. Several compounds were found which may be the potential drug for IPF.

Key Words: Acute idiopathic pulmonary fibrosis, Differentially expressed genes, PPI network, Enrichment analysis.

Introduction

Acute respiratory distress syndrome (ARDS) is a serious lung syndrome which occurs in the produce of various non-cardiac diseases, such as infection, shock, injuries and burns. It is characterized by diffuse pulmonary interstitial edema and alveolar edema, which were caused by damages of lung capillary endothelial cells and alveolar epithelial cells1,2.

Acute idiopathic pulmonary fibrosis is also a main inducer accounting for ARDS. Idiopathic pulmonary fibrosis (IPF) is a chronic lung disease characterized by fibrosis of the supporting framework of the lungs3. Nowadays, IPF affects about 5 million people in the global. The development of this disease is typically progressive. Initially, the main symptoms is dry coughing and breathless. Then, significant disability of lung function will appear. At the end-stage, copulmonale often occurs, and in that case, many patients will die because of classic signs of right heart failure4. Moreover, according to a survey, the median survival of patient with IPF is less than five years from the time of diagnosis5.

Recently, many new molecular therapeutic targets and biology processes have been linked to IPF, such amesenchymal cell proliferation, excess extracellular matrix deposition, hypercoagulation status, epithelial injury, dysregulated repair and epithelial cell activation5,6. Matrix metalloproteinases (MMPs) are also reported to be associated with the IPF. The aberrant regulation of extracellular matrix (ECM) is mainly caused by the imbalance between the MMPs and their inhibitors, tissue inhibitor of metalloproteinases (TIMP), which can disrupt the degradation of pulmonary ECM, thereby, leading to excess ECM deposition in alveolar interstitium7,8.
Despite extensive investigation, the cause of IPF remains unclear. Medical therapy is ineffective in the treatment of IPF. However, research of screening the DEGs in IPF using bioinformatics methods remain rare. Moreover, genome-wide approaches have reshaped the landscape of the problem. Meta-analysis of multi-study data has allowed the identification of DEGs that may have biomarker potential.

In this study, a total of 349 genes were identified as differentially expressed genes (DEGs) between normal and disease samples based on GSE10667, and then the PPI network for the products of these DEGs was constructed. Consequently, Cytoscape and its plug-in Mcode were used to make the visualization of network and identify the modules in the network. Furthermore, GO (Gene Ontology) enrichment analysis was performed by DAVID to demonstrate the function of the genes in our modules. Finally, PLK1 (Polo like kinase-1) was chosen as a marker gene, and cell cycle pathway was found to be enriched in IPF. Moreover, we also screened the drug compounds in DrugBank for PLK1.

**Materials and Methods**

**Microarray Data**

The transcription profile of GSE106679,10 (a total of 23 samples, including 8 disease samples collected from patients with ARDS and 15 normal controls) was obtained from NCBI GEO database (http://www.ncbi.nlm.nih.gov/geo/) which is based on Agilent-014850 Whole Human Genome Microarray 4x44K G4112F.

**Differentially Expressed Analysis**

The format of dataset was transformed, and then missing value estimation was performed for these dataset for accurate estimation of missing data. Furthermore, these inputted missing data were normalized. Finally, the limma package11 in R language with multiple testing correction (BH: Benjamini Hochberg method) was applied to identify genes which were significantly differentially expressed between disease samples and normal control12. We defined $p$-value < 0.05, FDR (False Discovery Rate) < -0.05 and $|\log_{FC}| > 1$ to be statistically significant.

**Construction of PPI Network by String**

The protein-protein interaction (PPI) network for the products of DEGs was constructed by String13. String is a database containing known and predicted relationship of protein interactions. The interactions include direct (physical) and indirect (functional) associations. Here we used it to predict interaction between products of these DEGs.

**Analysis of Modules in Interaction Network**

Gene products in same module often have same or similar function, they work together to carry out their biological function. Therefore, Cytoscape and its plug-in Mcode were used to make the visualization of network and identify the modules in the network [parameters: Degree cutoff ≥ 2 (degrees of each nodes in module were larger than 2 at least), K-core ≥ 2 (subgraphs of each node in module were more than 2 at least)]. Then Bingo was used to determine which Gene Ontology (GO) terms were significantly overrepresented in the genes in our module (adj $p < 0.01$).

**GO Enrichment Analysis**

The online biological tool, DAVID (The Database for Annotation, Visualization and Integrated Discovery)14 was used to perform GO term enrichment analysis for each group of genes in our PPI network.

**Drug Discovery**

Using PLK1 as the target genes, we screened the drug compounds in DrugBank15. DrugBank is a database, which has been widely used to facilitate in silico drug target discovery, drug design, drug docking or screening, drug metabolism prediction, drug interaction prediction and general pharmaceutical education16.

**Results**

**Differential Expression Analysis**

In our criterion ($p$ value < 0.05, FDR < 0.05, $|\log_{FC}| > 1$), a total of 349 DEGs were identified by limma package in R language. Among them, 87 genes were down-regulated and other 262 genes were up-regulated. Some genes included 20 genes in the cytoskeletal category, 18 genes in the extracellular matrix category and 12 genes in the cell cycle category.
cludes 290 pairs of proteins (Figure 1). Among them, interaction relationships of 9 pairs were experimental verified and the others were predicted in STRING according to the sequencing, structure and characteristic of the proteins.

**Analysis of Modules in Interaction Network**

We identified the modules in our PPI network. Three modules were identified, which included 18, 5 and 4 genes, respectively (Figure 2). Then the function of each module was annotated based on GO (Table I). Interestingly, the genes in these three modules were mainly associated with protein kinase binding, extracellular matrix structural constituent and structural constituent of cytoskeleton, respectively.

**Function and Pathway Annotation**

To demonstrate the altered biology pathway in our modules, pathway enrichment analysis was performed by using DAVID. We defined *p*-value < 0.05 to be statistically significant. For genes in module B and C, no significantly enrichment was found. Otherwise genes in module A were identified to be enriched in cell cycle and p53 signaling pathway (Table II). Among them, the cell cycle pathway is mostly significant, and PLK1 was predicted to participate in the pathway.

**Drug Discovery**

Since PLK1 was predicted to be an important gene in this study, we also screened the drug compounds in DrugBank for it. Several compounds have been found, including nocodazole and some vitamins. These compounds are proved to be act on PLK1 by previous studies, and the related papers were listed in Table III.

**Discussion**

IPF is a serious lung disease, and until today, there is no effective way to treat it. However researching the underlying mechanism of IPF through bioinformatics may hint a new thought to the therapy of IPF.

![Figure 1. The PPI network for products of DEGs. Diamond represents confirmed interactions which were proved by references in String database; Round represents unconfirmed interactions which were predicted by String according to the sequence, structure and characteristic of the proteins.](image-url)
Differentially expressed genes in acute idiopathic pulmonary fibrosis

In this study, about 300 DEGs were identified in disease samples compared to normal control. Further the PPI network for these DEGs was constructed and modules in the network were identified. Three modules were identified (Figure 2), and among them module A were focused. The function of genes in cluster A was associated with protein kinases. We also performed enrichment analysis to demonstrate the altered pathways in IPF. Finally, PLK1 was predicted to be significantly enriched in cell cycle pathway.

Pololike kinase 1 (PLK1) is a serine/threonine protein kinase, in human which is 66kDA and consists of 603 amino acids17. PLK1 contains two conserved polo-box domain at C-terminus, and one kinase domain at N-terminus. Polo-box

Table I. The enriched GO categories of Module A, B and C (FDR < 0.05, Module A, only list the top 10 GO categories).

<table>
<thead>
<tr>
<th>Module</th>
<th>GO-ID</th>
<th>Description</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module A</td>
<td>19901</td>
<td>Protein kinase binding</td>
<td>1.11E-04</td>
</tr>
<tr>
<td>Module A</td>
<td>19900</td>
<td>Kinase binding</td>
<td>1.27E-04</td>
</tr>
<tr>
<td>Module A</td>
<td>19899</td>
<td>Enzyme binding</td>
<td>1.54E-03</td>
</tr>
<tr>
<td>Module A</td>
<td>3777</td>
<td>Microtubule motor activity</td>
<td>1.84E-03</td>
</tr>
<tr>
<td>Module A</td>
<td>5524</td>
<td>ATP binding</td>
<td>1.89E-03</td>
</tr>
<tr>
<td>Module A</td>
<td>32559</td>
<td>Adenyl ribonucleotide binding</td>
<td>1.89E-03</td>
</tr>
<tr>
<td>Module A</td>
<td>5515</td>
<td>Protein binding</td>
<td>1.92E-03</td>
</tr>
<tr>
<td>Module A</td>
<td>30554</td>
<td>Adenyl nucleotide binding</td>
<td>2.02E-03</td>
</tr>
<tr>
<td>Module A</td>
<td>1883</td>
<td>Purine nucleoside binding</td>
<td>2.02E-03</td>
</tr>
<tr>
<td>Module A</td>
<td>1882</td>
<td>Nucleoside binding</td>
<td>2.02E-03</td>
</tr>
<tr>
<td>Module B</td>
<td>5201</td>
<td>Extracellular matrix structural constituent</td>
<td>7.34E-08</td>
</tr>
<tr>
<td>Module B</td>
<td>48407</td>
<td>Platelet-derived growth factor binding</td>
<td>4.84E-05</td>
</tr>
<tr>
<td>Module B</td>
<td>5198</td>
<td>Structural molecule activity</td>
<td>7.96E-05</td>
</tr>
<tr>
<td>Module B</td>
<td>46332</td>
<td>SMAD binding</td>
<td>4.73E-04</td>
</tr>
<tr>
<td>Module B</td>
<td>19838</td>
<td>Growth factor binding</td>
<td>1.93E-03</td>
</tr>
<tr>
<td>Module C</td>
<td>5200</td>
<td>Structural constituent of cytoskeleton</td>
<td>2.63E-06</td>
</tr>
<tr>
<td>Module C</td>
<td>5198</td>
<td>Structural molecule activity</td>
<td>7.04E-06</td>
</tr>
<tr>
<td>Module C</td>
<td>30674</td>
<td>Protein binding, bridging</td>
<td>4.83E-02</td>
</tr>
</tbody>
</table>

Table II. The enriched pathways in Module A (p-value < 0.05).

<table>
<thead>
<tr>
<th>KEGG-ID</th>
<th>Description</th>
<th>p value</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa04110</td>
<td>Cell cycle</td>
<td>1.40E-04</td>
<td>PLK1</td>
</tr>
<tr>
<td>hsa04115</td>
<td>p53 signaling</td>
<td>0.001717002</td>
<td>CCNB1, RRM2, GTSE1</td>
</tr>
</tbody>
</table>
domain plays an important role in both subcellular localization and regulation of kinase activity. The GFP-PLK1 recombinant protein has been found to localize to centromere region during interphase, and PLK1 also plays an important role in early mitosis.

Moreover, PLK1 plays a key role in cell cycle pathway. It is recognized as an early trigger for G2/M transition. The cell cycle is series of events that take place in a cell leading to its division and duplication. In the course of evolution, cells develop and establish a serial of mechanism to ensure that the alternation of each phase of cell cycle is in order. Many regulators have been found to be related to the cell cycle. Cell cycle proteins, cyclin-dependent kinase, and cyclin-dependent kinase inhibitor are the three main categories of these regulators. PLK1 belongs to cyclin-dependent kinase.

In cell cycle pathway, PLK1 is phosphorylated and activated by CDK1, further the phosphorylated PLK1 can activate Cdc25C, a phosphatase that can dephosphorylate and activate the cyclinB/cdc2 complex. These enzymes are controlled by DNA structure checkpoints, which de-
lay the onset of mitosis in the presence of unreplicated or damaged DNA. The abnormal regulation of kinase in cell cycle pathway often leads to development of many diseases, such as cancer and IPF. More and more researchers begin to focus on the relationship between the cell cycle and IPF. For example, some researchers report that cell cycle genes encoding p53 are associated with IPF disease development and progression. Tumor protein 53 (p53) is a key regulator of apoptosis. It is up-regulated upon DNA damage and prevents damaged cells from becoming malignant by inducing growth arrest and cell death. Increased levels of p53 in the lungs of IPF patients are consistent with increased apoptosis. Moreover, P53-induced growth arrest is mediated by increased CDK proteins.

It is also been reported overexpression of FoxO3a in IPF fibroblasts increases the CDK inhibitor, p27 levels and suppresses the ability of IPF fibroblasts to proliferate on polymerized collagen.

Our findings consist with the previous studies, and these findings support the idea that cell cycle control plays a role in the pathology of IPF. Here, we firstly reported that PLK1 was related to IPF. Furthermore, according to our results, together with previous studies we are likely to suppose that PLK1 may be associated with IPF through CDK1.

We also screened the drug compounds in DrugBank for PLK1. Several compounds have been found to act on PLK1. Such as Nocodazole, it is an anti-neoplastic drug which is often used to interfere with microtubule polymerization. Here, we first reported it should be a potential drug for IPF. However, more experimental verifications are still needed to prove this hypothesis.

Conclusions

Molecular biology, genetics and other experimental techniques are continually updated and mature. Progress in these fields can provide us new methods and new thoughts in treatment of IPF. In this study, a total of 349 DEGs were identified, and among them PLK1 was identified as a marker gene in IPF and predicted to be involved in cell cycle pathway. What’s more, we also suggest that Nocodazole may become a drug for curing the IPF. Although, more experiments are needed to verify these results, our findings still increase cause for optimism in the treatment of IPF.

Conflict of Interest

The Authors declare that they have no conflict of interests.

References


