Non-small-cell lung cancer pathological subtype-related gene selection and bioinformatics analysis based on gene expression profiles

JIANGPENG CHEN1, XIAOQI DONG2, XUN LEI1, YINYIN XIA1, QING ZENG1, PING QUE1, XIAOYAN WEN1, SHAN HU1 and BIN PENG1

1School of Public Health and Management, Chongqing Medical University, Chongqing 400016; 2Department of Respiratory Diseases, The First Affiliated Hospital of Zhejiang University, Hangzhou, Zhejiang 310003, P.R. China

Received March 21, 2016; Accepted November 21, 2017

DOI: 10.3892/mco.2017.1516

Abstract. Lung cancer is one of the most common malignant diseases and a major threat to public health on a global scale. Non-small-cell lung cancer (NSCLC) has a higher degree of malignancy and a lower 5-year survival rate compared with that of small-cell lung cancer. NSCLC may be mainly divided into two pathological subtypes, adenocarcinoma and squamous cell carcinoma. The aim of the present study was to identify disease genes based on the gene expression profile and the shortest path analysis of weighted functional protein association networks with the existing protein-protein interaction data from the Search Tool for the Retrieval of Interacting Genes. The gene expression profile (GSE10245) was downloaded from the National Center for Biotechnology Information Gene Expression Omnibus database, including 40 lung adenocarcinoma and 18 lung squamous cell carcinoma tissues. A total of 8 disease genes were identified using Naïve Bayesian Classifier based on the Maximum Relevance Minimum Redundancy feature selection method following preprocessing. An additional 21 candidate genes were selected using the shortest path analysis with Dijkstra’s algorithm. The AURKA and SLC7A2 genes were selected three and two times in the shortest path analysis, respectively. All those genes participate in a number of important pathways, such as oocyte meiosis, cell cycle and cancer pathways with Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis. The present findings may provide novel insights into the pathogenesis of NSCLC and enable the development of novel therapeutic strategies. However, further investigation is required to confirm these findings.

Introduction

Lung cancer is one of the most common malignant diseases and a major threat to public health on a global scale. The main types of lung cancer are small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC). NSCLC has a higher degree of malignancy and a lower 5-year survival rate compared with SCLC, and may be divided into two major histopathological subtypes, namely adenocarcinoma (ADC) and squamous cell carcinoma (SCC). As NSCLC is a complex disease, the identification of disease genes has been among the main goals of biologists and clinicians. Although our understanding of lung cancer, particularly NSCLC, has improved significantly in recent years, several questions require further elucidation.

With the completion of the Human Genome Project, scientific research has entered the post-genome era. Furthermore, with the advances in biological research and the development of high throughput biotechnologies, the quantities of candidate genes and numerous genomic loci have been reported to play a vital role in the pathogenesis of NSCLC. However, the mechanism underlying this disease has not yet been fully elucidated. As is well-known, different analytical methods may result in different consequences due to the ‘small N large P’ problem in microarray data, i.e., the gene expression data usually come with only dozens of tissue samples, but with up to tens of thousands of gene features. This extreme sparseness is considered to significantly compromise the performance of a classifier (1). As a result, the ability to extract a subset of informative genes while removing irrelevant or redundant genes is crucial for accurate classification (2). Hence, the methodology research is of great importance and the reuse of microarray data is feasible and extremely valuable. It is meaningful to identify disease-specific genes to help provide a novel therapeutic target for tumor treatment and elucidate the mechanism of the disease.

The aim of the present study was to identify the NSCLC subtype-related genes based on the classification method and the shortest path analysis based on the protein-protein interaction (PPI) network, and to investigate their underlying functions by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis with the Database for Annotation, Visualization and Integrated
Discovery (DAVID). This workflow may be of value for identifying disease genes and may provide insight into the pathogenesis of malignant tumors. To the best of our knowledge, this investigation is the first of its type in lung cancer studies, and it may be useful for further lung cancer candidate gene verification, thus enabling a better understanding of the molecular mechanisms of NSCLC and the development of more effective diagnostic and intervention methods.

Materials and methods

Data source. The gene expression dataset (GDS3627, accession no. GSE10245) was downloaded from the National Center for Biotechnology Information Gene Expression Omnibus (GEO) from a research conducted by Kuner et al (3) on lung cancer. The data were processed with the Affymatrix GPL570 platform and generated from 58 tissue samples (40 ADC and 18 SCC). All the tissue samples underwent careful analysis of the histopathology, estimation of tumor and stromal content and subsequent macro-dissection of the tumor regions prior to enrollment in the microarray experiment. No ethics committee approval is required to obtain these data, since they are publicly available. In addition, only expression data are presented, and no personal patient information is revealed.

Preprocessing of microarray data. Data preprocessing, including quantile normalization and log_{2} transformation, was completed by the GEOquery package (4). In order to minimize false positives, the GeneSelector package in R programme was used to rank gene lists. Briefly, selection of differential expressed genes may be highly affected by the statistical methods used, and the overlap of genes selected using the same significance value may be quite low. GeneSelector ranks genes on the basis of how well they perform in a selected number of statistical tests and minimize the number of false positives (5). Ranked lists based on ordinary Student’s t-test, significance analysis of microarrays, and fold-change were first generated for each of the criteria. The Markov chain model within the GeneSelector package was used to aggregate the ranked lists of all criteria into a list of genes that are significant, independently of the method used. To reduce noise and due to the fact that prediction methods such as Naive Bayes Classifier prefer categorical data, the gene expression data were discretized using the respective μ (mean) and σ (standard deviation): Any data larger than μ + 2σ were transformed to state 1; any data between μ - 2σ and μ + 2σ were transformed to state 0; and any data smaller than μ - 2σ were transformed to state -1. These three states correspond to the overexpression, baseline, and underexpression of genes, respectively. According to the previous studies, discretization of gene expression data generally leads to better prediction accuracy.

Maximum relevance minimum redundancy (mRMR) feature selection method. The mRMR feature selection method was proposed by Ding et al (6). For discrete variables, the mutual information I of two variables x and y is defined based on their joint probabilistic distribution p(x,y) and the respective marginal probabilities p(x) and p(y).

\[ I(x,y) = \sum_{i,j} p(x_i,y_j) \log \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \]  

Let S denote the subset of features we are seeking. The maximum relevance condition is

\[ \max D(S,c), D = \frac{1}{|S|} \sum_{i \in S} I(x_i;c) \]  

The minimum redundancy condition is

\[ \min R(S,c), R = \frac{1}{|S|} \sum_{i,j \in S} I(x_i;x_j) \]  

where |S| is used to represent the number of features in S, and c represents the targeted classes.

The mutual information quotient criterion may be described as follows.

\[ \max \Phi(D,R), \Phi = D/R \]  

This optimization may be computed efficiently by heuristic algorithm

\[ \max V_{mqr}, V_{mqr} = \max \left\{ \frac{I(x_j;c)}{m-1} \sum_{i \neq j} I(x_i;x_j) \right\} \]  

where \( x_j \in X_q - S_{m-1}, X_q \) represents the original feature set.

Naive Bayes Classifier. The Naive Bayes Classifier is a simple probabilistic classifier based on applying Bayes’ theorem with strong independence assumptions between features. For a sample s with m gene expression levels \{g_{1}, g_{2}, ..., g_{m}\} for the m features, the posterior probability that s belongs to class \( c_k \) is

\[ p(c_k | s) \propto \prod_{i=1}^{m} p(g_i | c_k) \]  

where \( p(g_i | c_k) \) are conditional tables estimated from training examples.

The Naive Bayes Classifier was used after all the features ranked by the mRMR. A series of classifiers were then constructed when the features were added into the classifier one by one. For example, the first classifier was constructed by the first feature, the second classifier by the first two features, and the rest in the same manner.

Using incremental feature selection (IFS), the number N may be determined. Its idea is to compare prediction accuracy defined in the following selection among different \( N_s \)s, and select the one with the highest accuracy.

The prediction accuracy was formulated by

\[ \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \]  

where TP represents the number of true positives, TN represents the number of true negatives, FP represents the number of false positives, and FN represents the number of false negatives.

Shortest path analysis. The PPI data have been widely used for gene function prediction with the assumption that
interacting proteins share the same or have similar functions and, hence, may be involved in the same pathway (7–9). The initial weighted graph $G = (V, E)$, where $V$ represented all human proteins occurring in PPI retrieved from STRING (version 10.0) (http://string-db.org/) (10), and $E$ contained all pairs of nodes such that the corresponding proteins comprised a PPI in STRING, was constructed according to the information retrieved from STRING. In a network, shortest path measures the least nodes required to pass through from one node to another (11). The Dijkstra’s algorithm in the igraph package was applied to identify the shortest path connecting each pair of genes selected by the mRMR in the PPI network.

Table I. Differentially expressed genes selected based on the gene expression profiles.

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Ensembl gene ID</th>
<th>Description</th>
<th>Fold change*</th>
<th>$t$</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP63</td>
<td>ENSG00000073282</td>
<td>Tumor protein p63</td>
<td>0.158</td>
<td>2.63</td>
<td>8.77E-03</td>
</tr>
<tr>
<td>SPC24</td>
<td>ENSG00000161888</td>
<td>SPC24, NDC80 kinetochore complex component</td>
<td>0.343</td>
<td>5.60</td>
<td>1.11E-06</td>
</tr>
<tr>
<td>RGS12</td>
<td>ENSG00000159788</td>
<td>Regulator of G-protein signaling 12</td>
<td>0.034</td>
<td>0.94</td>
<td>1.79E-01</td>
</tr>
<tr>
<td>CTDSP1L2</td>
<td>ENSG00000137770</td>
<td>CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase-like 2</td>
<td>0.400</td>
<td>5.30</td>
<td>2.40E-06</td>
</tr>
<tr>
<td>TMEM62</td>
<td>ENSG00000137842</td>
<td>Transmembrane protein 62</td>
<td>0.347</td>
<td>8.20</td>
<td>8.62E-10</td>
</tr>
<tr>
<td>SDCCA1G</td>
<td>ENSG00000054282</td>
<td>Serologically defined colon cancer antigen 8</td>
<td>0.127</td>
<td>3.88</td>
<td>2.88E-04</td>
</tr>
<tr>
<td>LOC10508342</td>
<td></td>
<td>Embryonal Fyn-associated substrate</td>
<td>0.405</td>
<td>7.11</td>
<td>1.30E-09</td>
</tr>
</tbody>
</table>

*Fold change is calculated as the ratio of the mean expression levels of genes in lung adenocarcinoma and squamous cell carcinoma.

Table II. Genes selected by the shortest path analysis.

<table>
<thead>
<tr>
<th>Ensembl peptide ID</th>
<th>Ensembl gene ID</th>
<th>Gene symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSP0000004531</td>
<td>ENSG0000003989</td>
<td>SLC7A2</td>
<td>Solute carrier family 7 (cationic amino acid transporter, $\gamma^+$ system), member 2</td>
</tr>
<tr>
<td>ENSP0000005340</td>
<td>ENSG0000004975</td>
<td>DVL2</td>
<td>Dishevelled segment polarity protein 2</td>
</tr>
<tr>
<td>ENSP0000025754</td>
<td>ENSG0000011465</td>
<td>DCN</td>
<td>Decorin</td>
</tr>
<tr>
<td>ENSP00000160827</td>
<td>ENSG00000079616</td>
<td>KIF22</td>
<td>Kinesin family member 22</td>
</tr>
<tr>
<td>ENSP000001171887</td>
<td>ENSG0000079308</td>
<td>TNS1</td>
<td>Tensin 1</td>
</tr>
<tr>
<td>ENSP00000215904</td>
<td>ENSG00000241360</td>
<td>PDXP</td>
<td>Pyridoxal (pyridoxine, vitamin B6) phosphatase</td>
</tr>
<tr>
<td>ENSP00000216911</td>
<td>ENSG0000087586</td>
<td>AURKA</td>
<td>Aurora kinase A</td>
</tr>
<tr>
<td>ENSP00000222248</td>
<td>ENSG00000105641</td>
<td>SLC5A5</td>
<td>Solute carrier family 5 (sodium/iodide cotransporter), member 5</td>
</tr>
<tr>
<td>ENSP00000228928</td>
<td>ENSG00000111331</td>
<td>OAS3</td>
<td>2'-5'-oligoadenylate synthetase 3, 100 kDa</td>
</tr>
<tr>
<td>ENSP00000251337</td>
<td>ENSG00000134183</td>
<td>GNAT2</td>
<td>Guanine nucleotide binding protein (G protein), $\alpha$ transducing activity polypeptide 2</td>
</tr>
<tr>
<td>ENSP00000261884</td>
<td>ENSG00000103671</td>
<td>TRIP4</td>
<td>Thyroid hormone receptor interactor 4</td>
</tr>
<tr>
<td>ENSP00000264126</td>
<td>ENSG00000121957</td>
<td>GPSM2</td>
<td>G-protein signaling modulator 2</td>
</tr>
<tr>
<td>ENSP00000264977</td>
<td>ENSG0000073711</td>
<td>PPP2R3A</td>
<td>Protein phosphatase 2, regulatory subunit B', $\alpha$</td>
</tr>
<tr>
<td>ENSP00000266970</td>
<td>ENSG00000123374</td>
<td>CDK2</td>
<td>Cyclin-dependent kinase 2</td>
</tr>
<tr>
<td>ENSP00000288266</td>
<td>ENSG00000157500</td>
<td>APPL1</td>
<td>Adaptor protein, phosphotyrosine interaction, PH domain and leucine zipper containing 1</td>
</tr>
<tr>
<td>ENSP00000306330</td>
<td>ENSG00000170027</td>
<td>YWHAG</td>
<td>Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, $\gamma$</td>
</tr>
<tr>
<td>ENSP00000308339</td>
<td>ENSG00000120341</td>
<td>SEC16B</td>
<td>SEC16 homolog B (S. cerevisiae)</td>
</tr>
<tr>
<td>ENSP00000308450</td>
<td>ENSG00000117399</td>
<td>CDC20</td>
<td>Cell division cycle 20</td>
</tr>
<tr>
<td>ENSP00000322775</td>
<td>ENSG00000129667</td>
<td>RHBDF2</td>
<td>Rhomboid 5 homolog 2 (Drosophila)</td>
</tr>
<tr>
<td>ENSP00000332816</td>
<td>ENSG00000120899</td>
<td>PTK2B</td>
<td>Protein tyrosine kinase 2 $\beta$</td>
</tr>
<tr>
<td>ENSP00000362556</td>
<td>ENSG00000169155</td>
<td>ZBTB43</td>
<td>Zinc finger and BTB domain containing 43</td>
</tr>
</tbody>
</table>
The genes occurring in any shortest path were considered as candidate genes.

Function and pathway enrichment analysis. GO and KEGG pathway enrichment analysis was performed with the functional annotation tool DAVID (12). All the protein-coding genes in the human genome were taken as background during the enrichment analysis.

Results

Genes selected by Naïve Bayes Classifier. A total of 4,015 genes remained for the following analysis after preprocessing. Naïve Bayes Classifier with mRMR was used to select genes associated with the discrimination of ADC and SCC. The IFS result is provided in Fig. 1. In the IFS curve, the x-axis is the number of genes used for classification and the y-axis is the prediction accuracy. As shown, the accuracy reached the maximum (100%) when 8 features were included. Those genes are listed in Table I, along with their results of fold-change and t-test.

Shortest path analysis. Dijkstra's algorithm was applied to select genes in the shortest path between any two genes described in Table I. Finally, 21 genes were found in 21 shortest paths, where AURKA and SLC7A2 were selected three and two times, respectively. Those genes, along with their Ensembl peptide ID, Ensembl gene ID and descriptions, are listed in Table II.

Function and pathway enrichment analysis. Using the functional annotation tool provided by DAVID, GO and KEGG pathway enrichment analysis was performed on the genes selected by Naïve Bayes Classifier and the shortest path analysis. The results demonstrated that several genes are significantly enriched in several biological processes such as mitosis, nuclear division, M phase of the mitotic cell cycle and two cellular components, intracellular non-membrane-bound organelle and non-membrane-bound organelle. There was no significant enrichment in molecular function (Table III).

The KEGG pathway enrichment analysis revealed that these genes are significantly enriched in oocyte meiosis, cell cycle, and pathways in cancer (Table IV).
Discussion

Lung cancer is one of the leading causes of cancer-related mortality worldwide. The Naïve Bayes Classifier approach with mRMR was applied followed by IFS to a microarray data generated from 40 ADC and 18 SCC tissue samples. With this approach, 8 genes that could optimally discriminate between ADC and SCC were identified. Those genes not only included known NSCLC-related genes, such as RGS12 (13), TP63 (14) and EFS (15), but also included genes that were not found to be closely associated with NSCLC. Fortunately, several genes, such as TP63, were found to be differentially expressed between ADC and SCC (16). Previously reported experimental evidence is provided for the expression and functions of these inferred genes in the discrimination of NSCLC, indicating that our method is effective for the discovery of new candidate genes.

Liu et al (17) compared the protein profiles of LC5 cell lines with MASPIN overexpression and knockdown using comparative two-dimensional gel electrophoresis, and found that SDCCAG8 was unique in MASPIN-expressing cell lines, but absent in knockout cell lines. Therefore, the authors hypothesized that SDCCAG8 may play a significant role in the invasion of cancer cells. Jadoon et al (18) used the iterative threading refinement programme I-TASSER to match the TMEM62 sequence with known protein structures. The top 10 structural analogues and five enzyme homologues for the TMEM62 model in the Protein Data Bank were all lipooxygenases (LOXs). According to the previous studies, LOXs form a heterogeneous class of lipid peroxidizing enzymes and were found to be implicated in the pathogenesis of various cancers (19,20). Thus, TMEM62 may play a vital role in the diagnosis and treatment of NSCLC. Unlike the two genes discussed above, CTDSPL2 was reported in only two studies. Ma et al (21) preliminarily confirmed that the CTDSPL2 gene obviously improved the expression of ε- and γ-globin genes, which provides a new candidate target for effective treatment of sickle cell disease and β-thalassemia. Zhao et al (22) demonstrated that CTDSPL2, as a Smad phosphatase, plays a critical role in bone morphogenetic protein (BMP)-induced signaling and cellular functions. Experimental evidence demonstrated that the mRNA level of BMP5 was significantly higher in ADC tissues compared with that in SCC tissues, and the immunohistochemistry analysis confirmed this result at the protein level (23). However, a correlation between the CTDSPL2 expression level and the occurrence of NSCLC has not been reported to date; it may be a novel biomarker of NSCLC development and progression.

Ndc80, which is composed of two heterodimers, CDC1A-KNTC2 and SPC24-SPC25, has been shown to play an important role in stable microtubule-kinetochore attachment, chromosome alignment and spindle checkpoint activation mitosis. The study conducted by Hayama et al (24) reported that CDC1A and KNTC2 appear to belong in the category of cancer-testis antigens, and that their simultaneous upregulation is a frequent and important characteristic of cell growth/survival in lung cancer, and selective suppression of CDC1A or KNTC2 activity and/or inhibition of the CDC1A-KNTC2 complex formation may be a promising therapeutic target for the treatment of lung cancers. Kaneno et al (25) reported the mRNA overexpression of CDC1A, KNTC2, SPC24 and SPC25 was observed in colorectal and gastric cancers. Therefore, SPC24 may be associated with the development and discrimination of NSCLC based on the abovementioned experimental evidence.

Rather than the traditional approach to studying individual genes or loci, a systematic investigation of cancer proteins in the human PPI network may provide important biological information for uncovering the molecular mechanisms of cancer. Since the discrimination genes must have some common characteristics related to NSCLC, it is reasonable to identify novel genes based on the PPI network and may provide insights into the comprehensive biological systems. Therefore, molecular interaction networks were constructed based on the PPI data from the STRING database, and 21 genes were identified in the shortest paths among the genes identified with the Naïve Bayes Classifier with mRMR. The AURKA (fold-change = 0.168; t=2.59, P=0.006) and SLC7A2 (fold-change = 0.007; t=1.28, P=0.104) genes were selected three and two times, respectively, in the shortest path analysis.

In humans, the aurora kinase family consists of highly conserved serine-threonine kinases, which play a critical role in the regulation of mitotic events such as spindle assembly, function of centrosomes and cytokinesis and cytokinesis (26,27). Lo Iacono et al (28) confirmed that AURKA expression was significantly upregulated in tumor samples compared with matched lung tissue [P<0.01, mean log2 (FC) = 1.5]. Moreover, AURKA was principally upregulated in moderately and poorly differentiated lung cancers (P<0.01), as well as in SCC and ADC compared with the non-invasive bronchioloalveolar histotype (P=0.029). Only a limited number of studies have reported the association of SLC7A2 with the development of NSCLC and future experimental studies are required to investigate this association.

Additionally, KEGG pathway enrichment analysis revealed that all those genes participate in a number important pathways, such as oocyte meiosis, cell cycle and pathways in cancer, among which CDK2 was found to appear in three pathways simultaneously. In fact, a large number of previous studies have demonstrated the importance of the role of CDK family members, such as CDK2, in the cell cycle control of tumor cells (29,30).

These results are helpful for cancer candidate gene verification, biomarker discovery and understanding the etiology of cancer at the biological level. However, further experiments are required to confirm these findings.

Acknowledgements

The present study was supported by grants from the National Natural Science Foundation of China (no. 81373103) and the Science Committee Basic and Cutting-edge Research Program of Chongqing (no. cstc2013jcyjA10009).

References


