Identification of genes and analysis of prognostic values in nonsmoking females with non-small cell lung carcinoma by bioinformatics analyses

Guangda Yang¹
Qianya Chen¹
Jieming Xiao²
Hailiang Zhang¹
Zhichao Wang¹
Xiangan Lin³

¹Department of Cancer
Chemotherapy, Zengcheng District
People’s Hospital of Guangzhou
(Boji-Affiliated Hospital of Sun
Yat-Sen University), Guangzhou,
China; ²Department of Emergency,
Zengcheng District People’s Hospital
of Guangzhou (Boji-Affiliated Hospital
of Sun Yat-Sen University), Guangzhou,
China; ³Department of Cancer
Chemotherapy, Sun Yat-sen Memorial
Hospital of Sun Yat-sen University,
Guangzhou, China

Background: This study was performed to identify disease-related genes and analyze prognostic values in nonsmoking females with non-small cell lung carcinoma (NSCLC).

Materials and methods: Gene expression profile GSE19804 was downloaded from the Gene Expression Omnibus (GEO) database and analyzed by using GEO2R. Gene Ontology and Kyoto Encyclopedia of Genes and Genomes were used for the functional and pathway enrichment analysis. Then, the Search Tool for the Retrieval of Interacting Genes, Cytoscape, and Molecular Complex Detection were used to construct the protein–protein interaction (PPI) network and identify hub genes. Finally, the Kaplan–Meier plotter online tool was used for the overall survival analysis of hub genes.

Results: A cohort of 699 differentially expressed genes was screened, and they were mainly enriched in the terms of ECM–receptor interaction, focal adhesion, and cell adhesion molecules. A PPI network was constructed, and 15 hub genes were identified based on the subset of PPI network. Then, two significant modules were detected, and several genes were found to be associated with the cell cycle pathway. Finally, nine hub genes (UBE2C, DLGAP5, TPX2, CCNB2, BIRC5, KIF20A, TOP2A, GNG11, and ANXA1) expressions were found to be associated with the prognosis of the patients.

Conclusion: Overall, we propose that the cell cycle pathway may play an important role in nonsmoking females with NSCLC and the nine hub genes may be further explored as potential targets for NSCLC diagnosis and treatment.

Keywords: non-small cell lung carcinoma, NSCLC, nonsmoking females, GEO2R, prognostic biomarkers, Kaplan–Meier plotter

Introduction
Lung cancer is one of the leading causes of cancer mortality worldwide.¹ Non-small-cell lung cancer (NSCLC), the main type of lung cancer, accounts for 80%–85% of all cases, and small cell lung cancer (SCLC) accounts for around 20% of all cases. Although smoking is the major risk factor for lung cancer, only 7% of female patients with lung cancer have a history of cigarette smoking in Taiwan.³ In addition, other factors, such as environmental exposure⁴,⁵ or hereditary factors,⁶ have been reported in association with nonsmoking lung cancer patients. However, the molecular mechanisms of NSCLC among nonsmoking women are unclear, although some genes, such as EML4-ALK,⁷ EGFR,⁸ TP53,⁹ and PIK3CA,¹⁰ have been found to be associated with lung cancer in never smokers. Thus, it is important to explore the molecular mechanisms involved in NSCLC’s onset and progression among nonsmoking females and identify effective biomarkers.

Recently, gene microarray and bioinformatics analysis were widely used to identify potential biomarkers of cancer. Interestingly, studies were performed to identify
disease-related genes, which were associated with prognosis of breast cancer, by using integrated bioinformatics methods. Similarly, some studies were performed to find important key genes in lung cancer. But there are limited studies on nonsmoking females with NSCLC.

In 2010, Lu et al obtained a panel of differentially expressed genes (DEGs) from nonsmoking female NSCLC patients and normal samples. Using the same data, the present study was performed to further screen the DEGs and predict their underlying function by functional and pathway enrichment analyses. Then, protein–protein interaction network (PPI) networks and modules of PPI network were performed to identify hub genes. More importantly, the prognostic values of the hub genes were further confirmed by survival analysis in nonsmoking females with NSCLC.

Materials and methods

Microarray data

Microarray expression profiles of GSE19804 were obtained from the Gene Expression Omnibus (GEO) database. GSE19804, which was based on the platform of the GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array, Santa Clara, CA, USA), consisted of 60 samples from nonsmoking female NSCLC patients and 60 normal samples.

Identification of DEGs

The DEGs between NSCLC samples and normal controls were identified using GEO2R, which is a web tool that is applied to screen DEGs by comparing two groups of samples. |log FC| > 1.5 and P < 0.01 were selected as the cutoff criterion.

Functional and pathway enrichment analysis

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of DEGs were performed using Database for Annotation, Visualization, and Integrated Discovery (DAVID), which is a comprehensive set of functional annotation tools. P < 0.05 was set as the cutoff criterion.

PPI network analysis of DEGs and modules selection

The Search Tool for the Retrieval of Interacting Genes (STRING) database was used to construct a PPI network for DEGs. Then, the modules of PPI network with significant gene pairs (combined score > 0.8) was obtained by the Molecular Complex Detection (MCODE) plugin in Cytoscape. MCODE scores > 14 and the number of nodes > 15 were used as the cutoff values. Moreover, the function and pathway enrichment analysis of genes in each module was performed using DAVID.

Survival analysis of the hub genes

Kaplan–Meier plotter is a web tool that predicts the prognostic values of genes in some cancer patients, including breast, ovarian, lung, and gastric cancer patients. The nonsmoking female patients with NSCLC were divided into two groups according to the particular gene expression level (high vs low expression). Based on these categories, overall survival (OS) analysis of the two patient groups was then compared by the tool. The HR with 95% CIs and log-rank P-value were calculated and showed.

Results

Identification of DEGs

The total number of samples included 60 NSCLC samples and 60 normal samples. A total of 699 genes were identified, including 161 upregulated and 538 downregulated genes.

GO and KEGG pathway enrichment analysis

We used the DAVID for GO and KEGG pathway enrichment analysis based upon DEGs. The top five GO terms of DEGs are shown in Table 1. As to biological process, the DEGs were significantly enriched in the regulation of cell proliferation, cell adhesion, biological adhesion, vasculature development, and blood vessel development. About cellular component, the DEGs were significantly enriched in the extracellular region part, extracellular region, extracellular space, cell surface, and proteinaceous extracellular matrix. In addition, the most enriched GO terms in molecular function were growth factor binding, carbohydrate binding, calcium ion binding, pattern binding, and polysaccharide binding. On the other hand, the most enriched KEGG pathway terms were as follows: ECM–receptor interaction, focal adhesion, cell adhesion, biological adhesion, vasculature development, and blood vessel development. About cellular component, the DEGs were significantly enriched in the extracellular region part, extracellular region, extracellular space, cell surface, and proteinaceous extracellular matrix. In addition, the most enriched GO terms in molecular function were growth factor binding, carbohydrate binding, calcium ion binding, pattern binding, and polysaccharide binding. On the other hand, the most enriched KEGG pathway terms were as follows: ECM–receptor interaction, focal adhesion, cell adhesion molecules, neuroactive ligand-receptor interaction, and complement and coagulation cascades.

Construction of PPI network and module identification

To predict the interactions of the identified DEGs from the protein level, a PPI network from STRING was constructed. In total, 439 nodes and 1,056 edges were shown in the PPI network.
Identification of genes and prognostic values in nonsmoking NSCLC females

network (Figure 1). The subset of PPI network for the DEGs with a combined score >0.8 was performed to determine the hub genes. The top 15 genes were selected as hub genes, which were GNG11, UBE2C, CCNB2, NMU, ANXA1, FPR2, KIF20A, MELK, NUSAP1, DLGAP5, CENPF, BIRC5, TOP2A, TPX2, and HMMR (Table 2). Subsequently, two modules with MCODE scores >14 and the number of nodes >15 were selected (Figure 2), and enrichment analysis showed that the genes in the modules were mainly associated with cell cycle, chemokine signaling pathway, cytokine–cytokine receptor interaction, and neuroactive ligand-receptor interaction (Table 3).

Survival analysis
Kaplan–Meier plotter was used to predict the prognostic value of 15 identified hub genes. Our results showed that high expression of UBE2C was associated with worse OS for NSCLC patients, as well as GNG11 (P<0.05) (Figure 3). Additionally, low expression of ANXA1 was associated with poorer OS for NSCLC patients, as well as GNG11 (P<0.05) (Figure 3).

Discussion
In this study, a total of 699 DEGs were screened, including 161 upregulated genes and 538 downregulated genes. Moreover, we selected two significant modules with several key DEGs in the regulatory network of nonsmoking female patients with NSCLC. Then, survival analysis of these genes determined that seven overexpressed genes (UBE2C, DLGAP5, TPX2, CCNB2, BIRC5, KIF20A, and TOP2A) and two downregulated genes (GNG11 and ANXA1) were significantly correlated with worse OS of nonsmoking female patients with NSCLC.

The data showed that the seven overexpressed genes are involved in “module 1” of the subnetwork, which is enriched in the cell cycle pathway. As we know that, ubiquitin-conjugating enzyme E2C (UBE2C) is a key regulator of cell cycle progression. High expression of UBE2C is associated with aggressive progression and

### Table 1: Functional and pathway enrichment analysis of DEGs in nonsmoking females with NSCLC

<table>
<thead>
<tr>
<th>GO_function</th>
<th>Description</th>
<th>Count</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO-BP:0042127</td>
<td>Regulation of cell proliferation</td>
<td>55</td>
<td>2.29E-10</td>
</tr>
<tr>
<td>GO-BP:0007155</td>
<td>Cell adhesion</td>
<td>51</td>
<td>2.94E-10</td>
</tr>
<tr>
<td>GO-BP:0022610</td>
<td>Biological adhesion</td>
<td>51</td>
<td>3.04E-10</td>
</tr>
<tr>
<td>GO-BP:0001944</td>
<td>Vasculature development</td>
<td>28</td>
<td>1.03E-09</td>
</tr>
<tr>
<td>GO-BP:0001568</td>
<td>Blood vessel development</td>
<td>27</td>
<td>2.83E-09</td>
</tr>
<tr>
<td>GO-CC:0044421</td>
<td>Extracellular region part</td>
<td>87</td>
<td>5.32E-22</td>
</tr>
<tr>
<td>GO-CC:0005576</td>
<td>Extracellular region</td>
<td>130</td>
<td>2.84E-20</td>
</tr>
<tr>
<td>GO-CC:0005615</td>
<td>Extracellular space</td>
<td>61</td>
<td>8.88E-15</td>
</tr>
<tr>
<td>GO-CC:0009986</td>
<td>Cell surface</td>
<td>36</td>
<td>1.30E-10</td>
</tr>
<tr>
<td>GO-CC:0005578</td>
<td>Proteinaceous extracellular matrix</td>
<td>33</td>
<td>9.70E-10</td>
</tr>
<tr>
<td>GO-MF:0019838</td>
<td>Growth factor binding</td>
<td>15</td>
<td>6.47E-07</td>
</tr>
<tr>
<td>GO-MF:0030246</td>
<td>Carbohydrate binding</td>
<td>27</td>
<td>2.57E-06</td>
</tr>
<tr>
<td>GO-MF:0005509</td>
<td>Calcium ion binding</td>
<td>48</td>
<td>8.91E-06</td>
</tr>
<tr>
<td>GO-MF:0001871</td>
<td>Pattern binding</td>
<td>16</td>
<td>1.39E-05</td>
</tr>
<tr>
<td>GO-MF:0030247</td>
<td>Polysaccharide binding</td>
<td>16</td>
<td>1.39E-05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>KEGG_PATHWAY</th>
<th>Description</th>
<th>Count</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa04512</td>
<td>Ecm-receptor interaction</td>
<td>16</td>
<td>3.74E-08</td>
</tr>
<tr>
<td>hsa04510</td>
<td>Focal adhesion</td>
<td>16</td>
<td>0.001596</td>
</tr>
<tr>
<td>hsa04514</td>
<td>CAMs</td>
<td>11</td>
<td>0.008625</td>
</tr>
<tr>
<td>hsa04080</td>
<td>Neuroactive ligand-receptor interaction</td>
<td>16</td>
<td>0.014903</td>
</tr>
<tr>
<td>hsa04610</td>
<td>Complement and coagulation cascades</td>
<td>7</td>
<td>0.021259</td>
</tr>
</tbody>
</table>

Note: Top five terms were selected according to P-value.

**Abbreviations:** BP, biological process; CAMs, cell adhesion molecules; CC, cellular component; DEGs, differentially expressed genes; ECM, extracellular matrix; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; MF, molecular function; NSCLC, non-small-cell lung cancer.
poor outcome of malignant glioma,\textsuperscript{21} and UBE2C has been identified as a prognostic protein marker in bladder cancer.\textsuperscript{22} Moreover, UBE2C, DLGAP5, and TPX2 are associated with the progression and prognosis of pancreatic carcinoma.\textsuperscript{23} Additionally, both DLGAP5 and TPX2 are mitosis-associated genes and correlated with poor prognosis in NSCLC patients.\textsuperscript{24} TPX2 overexpression is associated with poor survival in gastric cancer.\textsuperscript{25} A study also found that TPX2 promotes glioma cell proliferation and invasion via activation of the AKT signaling pathway in glioblastoma multiforme.\textsuperscript{26} Elevated DLGAP5 expression was negatively correlated with both OS and relapse-free survival of lung cancer.\textsuperscript{27} A study found that silencing of DLGAP5 by siRNA significantly inhibits the proliferation and invasion of hepatocellular carcinoma cells.\textsuperscript{28} Cyclin B2 (CCNB2) is a member of cyclin family proteins. Using the same data (GEO:GSE19804), Qian et al further demonstrated that both mRNA and protein expressions of CCNB2 were higher in NSCLC than in normal lung tissues and CCNB2 overexpression is a poor prognostic biomarker in Chinese NSCLC patients.\textsuperscript{29} BIRC5, which codes for survivin, can participate in cellular survival functions, such as cell cycle progression.\textsuperscript{30} Moreover, BIRC5 (survivin) is a pejorative prognostic marker in stage II/III breast cancer.\textsuperscript{31} Similarly, survivin is upregulated in lung cancer tissues and high expression of BIRC5 is associated with poor

Figure 1 PPI network of DEGs.

Abbreviations: DEGs, differentially expressed genes; PPI, protein–protein interaction.
Identification of genes and prognostic values in nonsmoking NSCLC females

32,33 Kinesin family member 20A (KIF20A), also known as RAB6KIFL, plays an important role in cytokinesis.34 In addition, it has been reported that high expression of KIF20A is associated with poor OS in cervical squamous cell carcinoma.35 Similarly, a study has found that positive expression of KIF20A indicates poor prognosis of glioma patients.36 High expression of TOP2A, which is involved in DNA synthesis and transcription, is associated with recurrence and metastasis of prostate cancer.37 In addition, high mRNA levels of TOP2A are independent predictors of poor outcome in renal cell carcinoma patients.38 Moreover, another study has found that TOP2A is associated with worse prognosis in NSCLC patients.39 Taken together, these data suggest that UBE2C, DLGAP5, TPX2, CCNB2, BIRC5, KIF20A, and TOP2A are involved in the cell cycle pathway and play a significant role in cancer development, which supports our findings.

On the other hand, GNG11 and ANXA1, which are found in “module 2”, were associated with chemokine signaling pathway, cytokine-cytokine receptor interaction, and neuroactive ligand-receptor interaction. GNG11, a lipid-anchored protein, acts as a tumor suppressor in lung adenocarcinoma.40 Annexin A1 (ANXA1) is a calcium- and phospholipid-binding protein. Some studies found that low ANXA1 expression was associated with better OS in gastric cancer41 and esophageal squamous cell carcinoma.42 However, low expression of ANXA1 was correlated with worse OS in breast cancer patients.43 Downregulation of ANXA1 is correlated with radioreistance in nasopharyngeal carcinoma.44 Moreover, decreased expression of ANXA1 was correlated with poor survival of pancreatic ductal adenocarcinoma (PDAC) patients, and ANXA1 knockdown inhibited cell proliferation, induced G1 phase cell cycle arrest, and increased PDAC cell migration and invasion capacity compared with controls.45 Similarly, ANXA1 knockdown suppressed the proliferation, migration, and invasion of NSCLC cells.46 Together, we speculate that GNG11 and ANXA1 may play a crucial role in NSCLC. ANXA1 could regulate the gastric cancer cell invasion through the formyl peptide receptor (FPR)/extracellular signal-regulated kinase/integrin beta-1-binding protein pathway, and all three FPRs (FPR1 through FPR3) were involved in the regulation process.41 FPR2 promotes invasion and metastasis of gastric cancer cells and predicts the prognosis of patients.47 A study also found that FPR2 was overexpressed in epithelial ovarian cancer (EOC) tissues and FPR2 overexpression indicated poor prognosis of EOC patients.48 However, in the present study, we found that

<table>
<thead>
<tr>
<th>Gene</th>
<th>Regulation</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNG11</td>
<td>Down</td>
<td>29</td>
</tr>
<tr>
<td>UBE2C</td>
<td>Up</td>
<td>28</td>
</tr>
<tr>
<td>CCNB2</td>
<td>Up</td>
<td>23</td>
</tr>
<tr>
<td>NMU</td>
<td>Up</td>
<td>22</td>
</tr>
<tr>
<td>ANXA1</td>
<td>Down</td>
<td>22</td>
</tr>
<tr>
<td>FPR2</td>
<td>Down</td>
<td>22</td>
</tr>
<tr>
<td>KIF20A</td>
<td>Up</td>
<td>22</td>
</tr>
<tr>
<td>MELK</td>
<td>Up</td>
<td>21</td>
</tr>
<tr>
<td>NUSAP1</td>
<td>Up</td>
<td>21</td>
</tr>
<tr>
<td>DLGAP5</td>
<td>Up</td>
<td>21</td>
</tr>
<tr>
<td>CENPF</td>
<td>Up</td>
<td>21</td>
</tr>
<tr>
<td>BIRC5</td>
<td>Up</td>
<td>21</td>
</tr>
<tr>
<td>TOP2A</td>
<td>Up</td>
<td>21</td>
</tr>
<tr>
<td>TPX2</td>
<td>Up</td>
<td>21</td>
</tr>
<tr>
<td>HMMR</td>
<td>Up</td>
<td>21</td>
</tr>
</tbody>
</table>

Abbreviation: PPI, protein–protein interaction.
FPR2 was downregulated and low expression indicated better prognosis of NSCLC patients. Similarly, we found that Neuromedin U (Nmu), a secreted neuropeptide, was upregulated and high expression indicated better prognosis of NSCLC patients. But in HNSCC, the expression levels of Nmu in primary tumors with regional metastasis were higher, compared with those without metastasis, and overexpression of Nmu may be involved in the process of regional metastasis of HNSCC.49 So, the roles of FPR2 and Nmu need to be further investigated.

In summary, the present study was intended to identify the potential biomarkers and analyze the prognostic values in nonsmoking females with NSCLC by bioinformatics analysis. We found that hub genes of complex networks, such as UBE2C, DLGAP5, TPX2, CCNB2, BIRC5, KIF20A, TOP2A, GNG11, and ANXA1, may act as potential biomarkers for nonsmoking females with NSCLC. However, the current study was performed by bioinformatics analysis and the conclusions remain to be confirmed by corresponding experiments. Therefore,
Identification of genes and prognostic values in nonsmoking NSCLC females

Figure 3 Kaplan–Meier curves depicting OS in nonsmoking females with NSCLC with high and low expression of ANXA1 (A), GNG11 (B), BIRC5 (C), CCNB2 (D), DLGAP5 (E), KIF20A (F), TOP2A (G), TPX2 (H), and UBE2C (I).

Abbreviations: NSCLC, non-small cell lung carcinoma; OS, overall survival.

The authors report no conflicts of interest in this work.

Disclosure

Further investigation is required to verify our findings and determine the potential clinical value of these as biomarkers.

References
