Comprehensive bioinformatics analysis identifies several potential diagnostic markers and potential roles of cyclin family members in lung adenocarcinoma

Li-Wei Gao1
Guo-Liang Wang2

1Department of Oncology, General Hospital of Pingmei Shenma Medical Group, Pingdingshan, Henan 467000, China; 2Department of Research & Development, Henan Zhongping Genetic Technology Co, Ltd, Zhengzhou, Henan 450000, China

Purpose: The aim of this study was to identify critical genes in lung cancer progression.

Methods: We downloaded and reanalyzed gene expression profiles from different public datasets using comprehensive bioinformatics analysis. Differentially expressed genes (DEGs) were identified in lung adenocarcinoma tissues compared with adjacent nonmalignant lung tissues. The overlapping DEGs identified from different datasets were used for functional and pathway enrichment analyses and protein–protein interaction (PPI) analysis. Moreover, transcription factors (TFs) and miRNAs that regulated the overlapping DEGs were predicted, followed by a TF–miRNA–target network construction. Furthermore, survival analysis of genes was performed. Several genes were further validated by quantitative real-time PCR (qRT-PCR).

Results: A total of 647 overlapping upregulated genes and 979 overlapping downregulated genes were identified. The overlapping upregulated genes and downregulated genes were involved in different functions, such as cell cycle, p53 signaling pathway, immune response, and cell adhesion molecules (CAMs). Several genes belonging to the cyclin family, including CCNB1, CCNB2, and CCNA2, were hubs of the PPI network and TF–miRNA–target network. Additionally, genes, including NPAS2, GNG7, CHIA, and SLC2A1, were predicted to be prognosis-related DEGs. Gene expression profiles determined by bioinformatics analysis and qRT-PCR were highly comparable.

Conclusion: CCNB1, CCNB2, CCNA2, NPAS2, GNG7, CHIA, and SLC2A1 are promising targets for the clinical diagnosis and therapy of lung adenocarcinoma.

Keywords: lung cancer, differentially expressed genes, transcription factor, prognosis

Introduction

Lung carcinomas are becoming a prevalent disease worldwide and are the leading cause of cancer-associated mortality among men and women.1 Lung cancer is mainly classified as small cell lung cancer and non-small-cell lung cancer (NSCLC), which accounts for approximately 85% of all lung cancers.2 Lung adenocarcinoma is the predominant pathological type of NSCLC. The average 5-year survival rate of NSCLC is around 15%, and adenocarcinoma makes a significant contribution.3 Current standard treatments for lung cancer include surgical resection, targeted therapies, and platinum-based dual chemotherapy.4 Although tremendous progress has been made concerning treatment, overall prognosis of lung cancer patients is still dismal.5 Thus, it is essential to discover effective diagnostic biomarkers and better understand the mechanisms underlying lung cancer.

Correspondence: Guo-Liang Wang
Department of Research & Development, Henan Zhongping Genetic Technology Co, Ltd, No 20, Shangwuwhuan Road, Zhengzhou, Henan 450000, China
Tel:+86 152 9086 1800
Fax:+86 0371 8998 6303
Email: biojeremy@163.com
Gene expression profiling has become a new and powerful approach to identify molecular targets for prognostic markers and therapy. A large-scale meta-analysis of lung cancer gene expression defined protein tyrosine kinase 7 (Inactive) (PTK7) as a specifically expressed gene in lung adenocarcinoma. C-X-C motif chemokine receptor 2 (CXCR2) expression was found to promote invasion and metastasis in lung adenocarcinoma. Konstantinidou et al showed that the RHOA–FAK signaling axis was critical for the maintenance of lung adenocarcinomas carrying activating mutations of the proto-oncogene KRAS. Recently, Lenka et al demonstrated differential expression of STXB6 in lung adenocarcinoma, which may be used as a novel biomarker for the prognosis of lung adenocarcinoma. Nevertheless, the significant genes associated with the progression of lung adenocarcinoma are still not well known.

In the present study, we downloaded and reanalyzed gene expression profiles from different public datasets using comprehensive bioinformatics analysis. Differentially expressed genes (DEGs) were identified in lung adenocarcinoma tissues compared with adjacent nonmalignant lung tissues. The overlapping DEGs identified from different datasets were used for functional and pathway enrichment analyses and protein–protein interaction (PPI) analysis. Moreover, transcription factors (TFs) and miRNAs that regulated the overlapping DEGs were predicted, followed by a TF–miRNA–target network construction. Furthermore, survival analysis of genes was performed. Several genes were further validated by quantitative real-time PCR (qRT-PCR). This study aimed to identify critical genes in lung cancer progression, which may lead to the discovery of candidate biomarkers and novel therapeutic targets for lung adenocarcinoma.

Methods

Data acquisition

Microarray data GSE75037 deposited by Girard et al, which contained a total of 166 related samples, including 83 lung adenocarcinoma tissues and 83 adjacent nonmalignant lung tissues, was downloaded from an international public repository, Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). The data platform is the GPL6884 Illumina HumanWG-6 v3.0 expression beadchip.

In addition, mRNA-seq and clinical data from lung tumor and paired normal tissues were downloaded from FireBrowse (http://firebrowse.org/), which is a portal to cull and analyze data generated by The Cancer Genome Atlas (TCGA). We finally obtained 515 tumor samples and 59 adjacent nontumor control samples from 515 patients with lung adenocarcinoma (average age: 65 years; 277 females and 238 males). The detailed patient information is found in Table S1.

Screening of DEGs

For the GSE75037 data from the GEO database, the Student’s t-test was implemented in the limma package of R/Bioconductor to screen DEGs in the tumor group vs in the normal group. Benjamini–Hochberg (BH) procedure was used for multiple test adjustment. Finally, we chose the adjusted P-value of <0.05 and the \(\log_{10}\) fold change (FC) of >1 as the significant threshold to screen DEGs.

The R-package edgeR (Version 3.4, http://www.bioconductor.org/packages/release/bioc/html/edgeR.html) was used to identify DEGs for the TCGA mRNA-seq data. Briefly, raw read count was processed using the Trimmed Mean of M-values (TMM) normalization, which is the default normalization method of the edgeR package. A generalized linear model was applied, and Voom’s precision weights were used to calculate the mean–variance relationship of the log-counts. Afterward, differential expression analysis was performed on mRNA data in the tumor group vs in the normal group using the t-test method implemented in the limma package. The P-value was also corrected by BH multiple correction test, resulting in an adjusted P-value. The threshold for the screening of DEGs was an adjusted P-value of <0.05 and the \(\log_{10}\) FC of >1.

Finally, the overlapped (common) genes between the DEGs from the GEO data and the TCGA data were collected, which were considered as significant genes associated with lung adenocarcinomas.

Functional and pathway enrichment analyses

The popular enrichment analysis tool DAVID (https://david-d.ncifcrf.gov/) was used to analyze the Gene Ontology (GO) function and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways of the upregulated and downregulated genes, respectively. The enriched gene count of ≥2 and the hypergeometric test P-value of <0.05 were set as the significant threshold.

Construction of PPI network

Based on human PPIs in the STRING (Version: 10.0, http://www.string-db.org/) database, PPIs for DEGs were analyzed using the default parameters. Required confidence (combined score) >0.9 was selected as the threshold for PPI analysis.
Through the PPI relationship obtained in the previous step, Cytoscape software\(^\text{21}\) was used to construct the network diagram. CytoNCA plugin (Version 2.1.6, http://apps.cytoscape.org/apps/cytonca)\(^\text{22}\) was used to perform node topology analysis with the following parameter: without weight. By ranking the scores of each node, including degree centrality (DC), betweenness centrality (BC), and closeness centrality (CC), important nodes in the PPI network were obtained, namely hub protein.\(^\text{21}\) Moreover, significant modules were extracted from the PPI network using the MCODE clustering algorithm\(^\text{24}\) with the score threshold of $>10$.

**Prediction of miRNAs and TFs that regulate DEGs**

Enrichr (http://amp.pharm.mssm.edu/Enrichr/) is a web-based comprehensive gene set enrichment analysis resource that accumulates biological knowledge for further biological discoveries.\(^\text{25}\) Enrichr uses the TargetScan database to predict miRNAs of genes.\(^\text{26}\) In this study, the Enrichr tool was used to identify lung cancer-associated miRNAs. Significant enrichment results with $P$-values $<0.05$ were screened out.

Transcriptional regulation data were identified using the ITFP database (http://itfp.biosino.org/itfp) and TRANSFAC database (http://www.gene-regulation.com/pub/databases.html). We downloaded all the predicted TF–target pairs from these two databases, screened the DEGs regulated by TFs by local Perl script, and further screened the TFs for DEGs.

Cytoscape software\(^\text{21}\) was used to integrate the miRNA–target gene network and the TF–target network, and the TF–miRNA–target network was constructed to find the major regulated genes.

**Survival analysis**

Clinical data of lung cancer patients were obtained from the TCGA data portal, including the overall survival (OS) time and vital status. The patients were divided into the following two groups according to the median expression value of the identified overlapping DEGs in the tumor group: high expression group and low expression group. The Kaplan–Meier curves of the two groups were drawn, and the log-rank statistical test was performed to analyze the relationship of gene and patient’s prognosis. $P$-value $<0.05$ was set as a statistically significant threshold.

**qRT-PCR validation**

To validate the results identified by the abovementioned analysis, the expression patterns of the top four prognosis-related DEGs were determined by qRT-PCR. The total RNA was extracted from a cultured human NSCLC cell line A549 and a human normal bronchial epithelial cell line BEAS-2B (as control) using the TRIzol reagent (Takara, Kyoto, Japan) according to the manufacturer’s protocol. Purity and concentration of isolated total RNA from cells were measured by the TECAN infinite M100 PRO Biotek microplate reader (Tecan Group, Ltd., Männedorf, Switzerland). First-strand cDNA was generated from 0.5 µg of total RNA using PrimerScript™ RT Master Mix (Takara Biotechnology Co., Ltd., Dalian, *China) according to the manufacturer’s instructions, and quantitative PCR was performed using the PowerUp SYBR Green Master Mix Kit (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer’s protocol. The total reaction volume was 20 µL, including 10 µL of SYBR Premix Ex Taq (2×), 1 µL of PCR Forward Primer (10 µM), 1 µL of PCR Reverse Primer (10 µM), and 8 µL of cDNA (diluted in double-distilled water). The qRT-PCR was set at an initial step of 3 minutes at 50°C and 3 minutes at 95°C, followed by 40 cycles at 95°C for 10 seconds and 60°C for 30 seconds. All experiments were done in triplicate, and all samples were normalized to GAPDH. The expression levels were calculated using 2$^{−ΔΔCt}$ methods. All primer sequences are listed in Table 1. Statistical analysis was performed with SPSS 22.0 software (IBM Corporation, Armonk, NY, USA) using Student’s $t$-test. Data are expressed as the mean ± SD from three independent experiments. A $P$-value of $<0.05$ was considered as statistically significant.

**Results**

**Identification of DEGs**

After analysis of the GSE75037 data, 1,324 upregulated genes and 1,542 downregulated genes were obtained. For the TCGA mRNA-seq data, 2,189 upregulated genes and 2,701 downregulated genes were obtained. The intersection of DEGs from both GSE75037 and the TCGA mRNA-seq data were considered as our focus. Thus, 647 overlapping upregulated genes and 979 overlapping downregulated genes were identified.

**Functional annotation of DEGs**

The overlapping upregulated genes and downregulated genes were found to be involved in different GO terms and pathways. The top five GO biological process (BP) terms and pathways of overlapping upregulated genes and downregulated genes are shown in Table 2, respectively. From the results, we found that the overlapping upregulated genes were mainly enriched in functions associated with the cell cycle, protein digestion and absorption, and p53 signaling.
Table 2: The enriched GO BP and pathway for overlapping upregulated and downregulated genes, respectively

<table>
<thead>
<tr>
<th>Expression</th>
<th>Category</th>
<th>Term</th>
<th>Count</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upregulated</td>
<td>G0TERM_BP_FAT</td>
<td>GO:0007067–mitotic nuclear division</td>
<td>33</td>
<td>2.34E–11</td>
</tr>
<tr>
<td></td>
<td>G0TERM_BP_FAT</td>
<td>GO:0051301–cell division</td>
<td>38</td>
<td>2.32E–10</td>
</tr>
<tr>
<td></td>
<td>G0TERM_BP_FAT</td>
<td>GO:0006260–Dna replication</td>
<td>20</td>
<td>6.23E–07</td>
</tr>
<tr>
<td></td>
<td>G0TERM_BP_FAT</td>
<td>GO:0030574–collagen catabolic process</td>
<td>13</td>
<td>8.05E–07</td>
</tr>
<tr>
<td></td>
<td>G0TERM_BP_FAT</td>
<td>GO:0000082–g1/s transition of mitotic cell cycle</td>
<td>14</td>
<td>2.42E–05</td>
</tr>
<tr>
<td></td>
<td>KEEG_PATHWAY</td>
<td>hasa04110:Cell cycle</td>
<td>18</td>
<td>1.31E–06</td>
</tr>
<tr>
<td></td>
<td>KEEG_PATHWAY</td>
<td>hasa04974:Protein digestion and absorption</td>
<td>12</td>
<td>2.37E–04</td>
</tr>
<tr>
<td></td>
<td>KEEG_PATHWAY</td>
<td>hasa01230:Biosynthesis of amino acids</td>
<td>11</td>
<td>2.38E–04</td>
</tr>
<tr>
<td></td>
<td>KEEG_PATHWAY</td>
<td>hasa04115:Complement and coagulation cascade</td>
<td>10</td>
<td>5.10E–04</td>
</tr>
<tr>
<td></td>
<td>KEEG_PATHWAY</td>
<td>hasa0250:Alanine, aspartate, and glutamate metabolism</td>
<td>7</td>
<td>0.001224098</td>
</tr>
<tr>
<td>Downregulated</td>
<td>G0TERM_BP_FAT</td>
<td>GO:0006955–immune response</td>
<td>99</td>
<td>3.92E–19</td>
</tr>
<tr>
<td></td>
<td>G0TERM_BP_FAT</td>
<td>GO:0006952–defense response</td>
<td>91</td>
<td>1.81E–18</td>
</tr>
<tr>
<td></td>
<td>G0TERM_BP_FAT</td>
<td>GO:0009611–response to wounding</td>
<td>80</td>
<td>8.98E–17</td>
</tr>
<tr>
<td></td>
<td>G0TERM_BP_FAT</td>
<td>GO:001944–vasculature development</td>
<td>52</td>
<td>1.11E–16</td>
</tr>
<tr>
<td></td>
<td>G0TERM_BP_FAT</td>
<td>GO:001568–blood vessel development</td>
<td>50</td>
<td>1.08E–15</td>
</tr>
<tr>
<td></td>
<td>KEEG_PATHWAY</td>
<td>hasa04610:Complement and coagulation cascade</td>
<td>18</td>
<td>1.09E–06</td>
</tr>
<tr>
<td></td>
<td>KEEG_PATHWAY</td>
<td>hasa05332:Graft-versus-host disease</td>
<td>13</td>
<td>3.51E–06</td>
</tr>
<tr>
<td></td>
<td>KEEG_PATHWAY</td>
<td>hasa04514:CAMs</td>
<td>22</td>
<td>9.33E–05</td>
</tr>
<tr>
<td></td>
<td>KEEG_PATHWAY</td>
<td>hasa04940:Type 1 diabetes mellitus</td>
<td>11</td>
<td>2.56E–04</td>
</tr>
<tr>
<td></td>
<td>KEEG_PATHWAY</td>
<td>hasa04270:Vascular smooth muscle contraction</td>
<td>19</td>
<td>2.59E–04</td>
</tr>
</tbody>
</table>

Abbreviations: BP, biological process; CAMs, cell adhesion molecules; GO, Gene Ontology.

pathway. The overlapping downregulated genes were mainly associated with the immune response and cell adhesion molecules (CAMs).

### Construction of the PPI network

The PPI network of overlapping DEGs is shown in Figure S1. The top 10 nodes with higher degree in the PPI network included G protein subunit gamma transducin 1 (GNGT1, upregulated, degree = 50), G protein subunit gamma 7 (GNG7, downregulated, degree = 48), G protein subunit gamma transducin 2 (GNGT2, downregulated, degree = 48), G protein subunit gamma 4 (GNG4, upregulated, degree = 48), cyclin B1 (CCNB1, upregulated, degree = 47), G protein subunit gamma 11 (GNG11, downregulated, degree = 46), angiotensinogen (AGT, upregulated, degree = 45), cyclin A2 (CCNA2, upregulated, degree = 43), cell division cycle 20 (CDC20, upregulated, degree = 43), and Jun proto-oncogene, AP-1 TF subunit (JUN, downregulated, degree = 41). Moreover, four significant modules were identified from the PPI network (Figure 1). We found that the genes in module B were upregulated including CCNB1 and cyclin B2 (CCNB2).

### Integrated analysis of the miRNA–target gene network and TF–target network

A total of two miRNAs (hsa-miR-4804-5p and hsa-miR-4790-5p) and four TFs (Forkhead Box M1 [FOXM1], TF TFDP1, E2F4, and SIN3A) were predicted to regulate the DEGs. Among them, FOXM1 was also differentially expressed in this study. Then, the TF–miRNA–target network was constructed (Figure 2). We found that CCNA2 and CCNB1 can be targeted by these four TFs. CCNB2 can be regulated by a miRNA and three TFs.

### Survival analysis

A total of 232 prognosis-related DEGs were identified from the overlapping DEGs. The top four prognosis-related DEGs
Significant genes in lung cancer

Figure 1 Four modules extracted from the PPI network. Note: Red node indicates upregulated genes, and green node represents downregulated genes. Abbreviation: PPI, protein–protein interaction.

Figure 2 TF–miRNA–target network of the overlapping DEGs. Notes: Red node indicates upregulated genes, and green node indicates downregulated genes. White triangle represents TF, and white quadrilateral represents miRNA. Among them, the DEG FOXM1 is also a TF. Abbreviations: DEGs, differentially expressed genes; TF, transcription factor.

(qranked by $P$-value) were neuronal PAS domain protein 2 ($NPAS2$, upregulated), G-protein subunit gamma 7 ($GNG7$) (downregulated), chitinase, acidic ($CHIA$, downregulated), and solute carrier family 2 member 1 ($SLC2A1$, upregulated). Survival curves of these four genes are shown in Figure 3.

qRT-PCR validation

We examined the expression of the top four prognosis-related DEGs ($NPAS2$, $GNG7$, $CHIA$ and $SLC2A1$) in lung cancer cell lines using qRT-PCR to validate the results of gene profiles of which $NPAS2$ and $SLC2A1$ were upregulated and
**Figure 3** Survival curves.

Notes: (A) Survival curves of NPAS2 expression. (B) Survival curves of GNG7 expression. (C) Survival curves of CHIA expression. (D) Survival curves of SLC2A1 expression.

**Discussion**

In this study, we used comprehensive bioinformatics analysis to identify DEGs in human lung adenocarcinomas, by investigating the microarray data from the GEO database and data from the TCGA database. Members of the cyclin family function as regulators of the cell cycle by activating cyclin-dependent kinase (CDK) enzymes.\(^27\) Deregulated expression of cyclin D1 is found frequently in cancer, and it has been demonstrated to be a biomarker of cancer progression.\(^28\) Activation of CCNB1 was found to promote cell proliferation and tumor growth in colorectal cancer cells.\(^29\) A study has demonstrated that CCNB2 may be used as a very

**GNG7** and **CHIA** were downregulated. As shown in Figure 4, two genes (NPAS2 and GNG7) yielded concordance with the results of bioinformatics analysis; NPAS2 was indeed significantly upregulated and GNG7 was significantly downregulated when compared with normal control cells. **CHIA** was downregulated without significant difference. However, **SLC2A1** showed a slight expression change in the opposite direction to our initial result, which may be affected by the heterogeneity between the samples used in the downloaded datasets and qRT-PCR. Nevertheless, gene expression profiles determined by bioinformatics analysis and qRT-PCR were highly comparable.
In addition, the findings of Guo et al showed that CCNA2 was a biomarker for breast cancer prognosis. In this study, CCNB1, CCNB2, and CCNA2 were identified as overlapping DEGs using data from two databases. Moreover, they were hubs in the PPI network and the TF–miRNA–target network, indicating their possible significant roles. In this context, we suggest that CCNB1, CCNB2, and CCNA2 may play critical roles in the progression of lung adenocarcinoma, which warrants further verification.

Additionally, NPAS2, GNG7, CHIA, and SLC2A1 were predicted to be prognosis-related DEGs. NPAS2 is a TF expressed primarily in the mammalian forebrain and is implicated in the regulation of the circadian rhythm. Evidence indicates that circadian rhythm disruption promotes lung tumorigenesis. Dysregulation of GNG7 was found in various types of cancer. The protein encoded by CHIA degrades chitin and chitotriose, which have relation to immune response regulation. A study had indicated the potential roles of chitin, chitinases, and chitinase-like proteins in lung diseases. Furthermore, the differential expression of SLC2A1 was demonstrated to promote tumor cell proliferation and metastasis. In addition, the differential expression profile of these four candidate genes was further validated by qRT-PCR in lung cancer cells due to the current lack of tissue samples. Therefore, SLC2A1 showed an opposite expression direction between the initial result and experiments, which may be affected by the heterogeneity between the samples used in the downloaded datasets and qRT-PCR. Taken together, the results suggest NPAS2, GNG7, CHIA, and SLC2A1 as promising diagnostic markers for patients with lung adenocarcinoma; therefore, conducting further investigation is important.

**Conclusion**
Our integrated approach showed several potential targets for lung adenocarcinoma. CCNB1, CCNB2, and CCNA2 may play critical roles in the progression of lung adenocarcinoma. NPAS2, GNG7, CHIA, and SLC2A1 may be promising diagnostic markers for patients with lung adenocarcinoma. However, additional research is required to clarify whether these genes are involved in lung adenocarcinoma and to
promote the development of targets for the clinical diagnosis and therapy of lung cancer.

Disclosure
The authors report no conflicts of interest in this work.

References
Supplementary material

Figure S1  PPI network of the overlapping DEGs.
Abbreviations: DEGs, differentially expressed genes; PI, protein–protein interaction.