Identification of key protein-coding genes in lung adenocarcinomas based on bioinformatic analysis

Ruixue Yao¹#, Xiaoming Chen²#, Luyao Wang¹#, Yuanyong Wang³, Shaoli Chi², Na Li⁴, Xuejun Tian⁵, Nan Li², Jia Liu¹

¹Department of Pharmacology, School of Pharmacy, Qingdao University, Qingdao 266000, China; ²The Third Department of Cadre's Ward, Navy 971 Hospital, Qingdao 266071, China; ³Department of Thoracic Surgery, Affiliated Hospital of Qingdao University, Qingdao 266000, China; ⁴The Department of Nuclear Medicine, Navy 971 Hospital, Qingdao 266071, China; ⁵Key Laboratory of Neuropsychiatric Drug Research of Zhejiang Province, Institute of Materia Medica, Zhejiang Academy of Medical Sciences, Hangzhou 310013, China

Contributions: (I) Conception and design: R Yao; (II) Administrative support: J Liu, N Li; (III) Provision of study materials or patients: L Wang; (IV) Collection and assembly of data: R Yao, X Chen, L Wang; (V) Data analysis and interpretation: R Yao, S Chi, N Li; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

#These authors contributed equally to this work.

Correspondence to: Nan Li, MD. Professor, The Third Department of Cadre’s Ward, Navy 971 Hospital, 266071 Qingdao, China. Email: li_nan1900@126.com; Jia Liu, MD, PhD. Professor, Department of Pharmacology, School of Pharmacy, Qingdao University, Qingdao 266000, China. Email: dadaliujia@qdu.edu.cn.

Background: Lung cancer is one of the most common cancers and the primary cause of cancer-related deaths in the world. The 5-year survival of lung cancer patients is lower than 15%. As a common subtype of lung cancer, lung adenocarcinoma still has a high morbidity and mortality, although many strategies have been made, such as surgical operation, chemotherapy, targeted therapy. The use of gene expression microarray has provided a feasible and effective approach for the study on lung cancer. However, the biomarkers and potential therapeutic targets of lung adenocarcinomas are still not completely identified. Our study is aimed to find biomarkers and therapeutic targets of lung adenocarcinomas by identifying the key protein-coding gene in lung adenocarcinomas by bioinformatical approaches.

Methods: We selected and obtained messenger RNA microarray datasets from Gene Expression Omnibus database to identify differentially expressed genes between lung adenocarcinomas and normal lung tissue. The differentially expressed genes were clarified by Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, the protein-protein interaction (PPI) network and statistical analyses. Subsequently, quantitative real-time PCR was used to verify the results of bioinformatic analysis.

Results: We obtained 1,264, 896 and 408 differentially expressed genes from GSE32863, GSE43458 and GSE63459, respectively. The 242 common differentially expressed genes in three datasets were related to cell adhesion molecules, ECM-receptor interaction, Leukocyte transendothelial migration according to KEGG analysis. GO analysis showed that these common differentially expressed genes were enriched in tumor-related functions. ASPM, CCNB2, CDC20, CDC45, MELK, TOP2A and UBE2T and KIAA0101 have the strongest protein-protein interaction relationships based on protein-protein interaction networks. Survival analysis showed that these nine genes were closely related to the survival of lung adenocarcinomas. The further qRT-PCR assays indicated that seven key genes (ASPM, CCNB2, CDC20, CDC45, MELK, TOP2A and UBE2T) display differential profile between clinical lung adenocarcinoma specimens and their matched normal tissues.

Conclusions: ASPM, CCNB2, CDC20, CDC45, MELK, TOP2A and UBE2T may be key protein coding genes in lung adenocarcinoma, and deserve further study to verify their feasibility and effectiveness as biomarkers and therapeutic targets for lung adenocarcinomas.

Keywords: Lung adenocarcinoma; key protein-coding gene; biomarker
Introduction

Nowadays, lung cancer remains a major type of malignancies with high death rate all around the word (1). Small cell lung cancer and non-small cell lung cancer are two histologic subtypes of lung carcinoma, and non-small cell carcinoma accounts for >85% of lung cancer (2). Lung adenocarcinomas is featured with frequent recurrence, and even at the first diagnosis, metastasis is easily detected in the patients (3). The prognosis is still poor and not satisfying, although many therapeutic modalities have been applied to the treatment of lung adenocarcinomas in recent years (4). The identification of biomarkers is of great significance for lung adenocarcinomas. Therefore, it is greatly needed to identify biomarkers and therapeutic targets in lung adenocarcinomas for better diagnose and treatment. The bioinformatics analyses is an effective way for detecting genome-wide gene expression, and The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) are Cancer-related public genome data databases used to re-analyze gene expression data from different research institute. So the bioinformatics analysis is effective for finding potential biomarker closely related to disease by studying some expression profile datasets of some disease online, using different bioinformatics analysis tools (such as GO analysis and KEGG analysis). In this study, we extracted data from GSE32863, GSE43458 and GSE63459 datasets, and screened 242 common differentially expressed genes between lung adenocarcinomas and normal pulmonary tissues. The function and pathway enrichment of DEGS were analyzed on GO and KEGG online databases, and their protein-protein interaction network was constructed by STRING online database. Furthermore, quantitative real-time PCR was used to identify the bioinformatic analysis.

Methods

Our study was approved by the Ethics Committee of the Affiliated Hospital of Qingdao University (No.2018015). Each patient was informed of the use of their tissue specimens.

Identification of feature genes

The expression profile datasets were gained from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/), which is a public genomics data repository (5,6). We selected three gene expression profile datasets (GSE32863, GSE43458 and GSE63459). GES63459 consists of 32 normal samples and 33 lung adenocarcinoma samples. GSE43458 consists of 30 normal samples and 80 lung adenocarcinoma samples, GSE32863 consists of 58 normal samples and 58 lung adenocarcinoma samples. The GEO2R (http://www.ncbi.nlm.nih.gov/geo/geo2r/) is a tool which can compare different groups of samples and identify DEGs in a GEO database. To identify the DEGs in lung adenocarcinomas, GEO data were divided into two groups, and GEO2R was used to compare them. P-value (adj. P-value) was adjusted using Benjamini-Hochberg method. |logFC| >1 and <0.05 were considered as the criteria to identify DEGs between normal tissue samples and lung adenocarcinoma samples.

Co-expression network construction

The online tool STRING database and Cytoscape software (www.cytoscape.org) (7) were used to establish the protein-protein interaction (PPI) network. Input of differentially expressed genes on the STRING database to predict their interactions. Hub genes were identified from PPIs, and then we gained the edge length between nodes from these hub genes. Hub genes are a small part of a protein that interacts with other proteins.

Bioinformatic analysis

Based on the Database for Annotation, Visualization and Integrated Discovery (DAVID) (https://david.ncifcrf.gov), Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) were used to analyze the biological function of DEGs. The enrichment threshold values were considered significant at P<0.01.

Survival analysis of patients

To identify the clinical value of differentially expressed genes, Kaplan-Meier curve and Cox regression analysis were performed. Kaplan-Meier is an online tool, and can predicts the effect of genes on survival while submit interested genes. The samples were divided into high expression group and low expression group. The ONCOMINE database (http://www.cytoscape.org) and Kaplan-Meier Plotter database are employed to perform all of the statistical analysis (8). Statistically significant was assessed using student’s t-test, P<0.05 was considered significance.
Quantitative real-time PCR

The expression levels of ASPM, CCNB2, CDC20, CDC45, MELK, PRC1, TOP2A, UBE2T and KIAA0101 in lung adenocarcinoma tissues were detected by qRT-PCR. We collected seven pairs of lung cancer and adjacent normal tissues. The specimens were excised and immediately frozen in liquid nitrogen and stored at −80 °C. Total RNA was retrieved into cDNA by reverse transcription kit (Tiangen, Beijing). qRT-PCR was performed using a SYBR Green PCR Kit (Tiangen, Beijing) and ABI StepOne Plus thermal circulator. The primer sequences are shown in Table 1. All specimens were analyzed in three copies. Data were calculated based on the comparative 2-ΔΔCT method.

Results

Identification of DEGs and bioinformatic analysis

We obtained 1,264, 896 and 408 DGEs in the datasets GSE32863, GSE43458 and GSE63459, respectively, identified 242 genes expressed in all the three datasets (Figure 1), and the up-regulated and down-regulated genes were shown in the volcano plot (Figure 2), logFC>1 and P<0.05 were considered as significantly up-regulated genes, logFC<−1 and P<0.05 were considered as significantly down-regulated genes. Gene symbols of the three databases are list in the http://cdn.amegroups.cn/static/application/bb0fb2e8a974b80b905d240f0bb48d8d/tcr.2019.10.45-1.pdf. The KEGG analysis showed that these differential genes were enriched in many biological pathways, including cell adhesion molecules, ECM-receptor interaction, leukocyte transendothelial migration and protein digestion, absorption, malaria and PPAR signaling pathway (Figure 3). The result of GO analysis demonstrated that DEGs were associated with many molecular function, such as calcium ion binding, heparin binding, collagen binding, glycosaminoglycan binding, extracellular matrix binding and transforming growth factor beta binding. In cellular component ontology, DEGs were most enriched in extracellular space, extracellular region, proteinaceous extracellular matrix, extracellular matrix, transforming matrix, membrane raft and collagen trimer. These DEGs were further analyzed in the biological, and they are mainly enriched in cell adhesion, extracellular matrix organization, angiogenesis transforming growth factor beta receptor signaling pathway, cellular response to hormone stimulus and response to cAMP (Figure 4).
Figure 2 Volcano plot of gene expression profile data in lung adenocarcinomas tissues and normal ones. (A) Volcano plot of GSE32863, (B) volcano plot of GSE43458, (C) volcano plot of GSE63459, Green represents down-regulation, red represents up-regulation.

Figure 3 The top six terms of biological processes of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway.
Two hundred forty-two common DEGs of three datasets were uploaded to STRING online database, and then Cytoscape software was used to identify the cluster. Among the 242 differentially expressed genes, ASPM, CCNB2, CDC20, CDC45, CDC47, MELK, PRC1, TOP2A, UBE2T and KIAA0101 had the closest correlation. Figure 5 showed the PPI work of the ten genes.

**Survival analysis**

In order to further evaluate the prognostic value of the ten significantly interacted genes, we conducted Kaplan-Meier curve using the Kaplan-Meier-plotter database. The results showed that ASPM, CCNB2, CDC20, CDC45, MELK, PRC1, TOP2A, UBE2T and KIAA0101 were all significantly associated with the survival of patients with lung adenocarcinomas (hazard ratio =1.76, 1.78, 2.13, 1.67, 1.95, 2.05, 1.86, 1.59, 1.71 respectively) (Figure 6). However, CDCA7 was not associated with poor survival in patients. Furthermore, the ONCOMINE database showed that the expression of ASPM, CCNB2, CDC20, CDC45, MELK, PRC1, TOP2A, UBE2T and KIAA0101 were higher in lung adenocarcinoma tissues, compared with normal counterparts (Figure 7). Although the analysis with online database revealed that the 9-gene signature was differentially expressed in pulmonary malignancy and associated with the prognosis, the above predicted genes were required for further confirmation in clinical specimen derived from lung adenocarcinoma.
Figure 6 Kaplan-Meier plots for ten genes in lung adenocarcinomas from http://www.oncolnc.org/. Log-rank P values and hazard ratios (HR, 95% confidence intervals in parentheses) are shown.
Figure 7 Figures were derived from gene expression data in the ONCOMINE database comparing expression levels in normal (left plot) and cancer tissues (right plot) in nine key genes.
Clinical verification based on qRT-PCR

Subsequently, to verify our bioinformatic analysis, qRT-PCR was used to confirm the expression profile of ASPM, CCNB2, CDC20, CDC45, MELK, PRC1, TOP2A, UBE2T and KIAA0101 in lung adenocarcinoma patients. The result of qRT-RCR showed, as compared to normal tissues, ASPM, CCNB2, CDC20, CDC45, MELK, TOP2A, UBE2T were obviously upregulated in lung adenocarcinoma tissues (Figure 8). However, there was no distinct difference in the expression of PRC1 and KIAA0101 between lung adenocarcinoma and normal tissues. Therefore, these genes may be potential biomarkers for predicting survival rates in patients with lung adenocarcinomas.

Discussion

Although chemotherapy and targeted therapy have made great benefits on patients, lung cancer remains a major cause for cancer-related deaths. Our study is aimed at identifying reliable biomarkers and therapeutic targets in lung adenocarcinomas to improve diagnosis and survival. In our study, we obtained three datasets from GEO database (GSE32863, GSE43458 and GSE63459), and then, compared gene expression profiles between lung cancer and normal tissues (n=291). Two hundred forty-two common DEGs were obtained in three datasets. The result of GO analysis showed that DEGs were associated with many the molecular functions, the cellular component
ontology and the biological process. The KEGG analysis indicated that the DEGs were enriched in many biological pathways. In our PPI network, ten key hub genes (ASPM, CCNB2, CDC20, CDC45, CCNB2, CDC20, CDC45, MELK, TOP2A, UBE2T and KIAA0101) were identified markedly interacted. Survival analysis also showed these nine hug genes (ASPM, CCNB2, CDC20, CDC45, MELK, TOP2A, and UBE2T) were higher expression in lung adenocarcinoma tissues. The small sample size may be one of the major reasons the profile of PRC1 and KIAA0101 were barely noticeable difference. Therefore, ASPM, CCNB2, CDC20, CDC45, MELK, TOP2A, and UBE2T may be responsible for lung adenocarcinomas, and they may be potential molecular targets for the treatment of lung adenocarcinomas. CCNB2, also called cyclin B2, is a member of cell cycle regulator family. CCNB2 interacted with cyclin-dependent protein kinase (CDK) family in eukaryotic cell cycle regulation. CCNB2 is mainly distributed in the Golgi apparatus (9), and is expressed during the all stages of cell division. Moreover, CCNB2 lays a key role in regulating eukaryotic cell cycle in during G2/M transition and mitosis by activating CDK1 kinase (10). The overexpression of CCNB2 has been verified in many human tumors (11), and its overexpression affects the development and prognosis of non-small cell lung cancer.

Cell division cycle associated 7 (CDC47) is a nuclear protein which contains a zinc finger domain in the C-terminus (12,13), and is a myc-regulated target associated with cancer transformation (14). Overexpression of CDC47 may promote myc-mediated tumorigenesis, which is often found in human tumors (15). One study has shown that mutations in CDC47 and HELLS can lead to immunodeficiency-centromeric instability-facial anomalies syndrome (16). CDC47 participates in hematopoietic stem cell emergence by recruiting the RBPJ factor, especially Notch1 in aorta-gonad-mesonephros cells (17).

Abnormal spindle microtubule assembly (ASPM) is a microtubule-related protein found in the mitotic spindle pole, and is essential for cell division. ASPM plays an important role in the orientation and maintenance of mitotic spindle, especially in nerve stem/progenitor cells (18). Its mutation may promote apoptosis of cells and deplete self-renewal of neural stem cells and progenitor cells, perhaps causing microcephaly (19). ASPM has recently been found as an important modulator of Wnt signaling transduction and cancer stem cells in hormone-refractory prostate cancer (PCA) (20). Its overexpression in glioblastoma cell accelerates the proliferation of the tumor cells (21).

DNA topoisomerase II alpha (TOP2A) is an enzyme that regulates DNA topology during transcription. The overexpression and mutation of TOP2A, and its interaction with other protein partners result in the progression of tumors (22). The expression of TOP2A, which peaks at G2 phase of cell cycle (23), is related to the growth of tumors (24). A study indicated that TOP2A is expressed in more than 70% of oncocytomas, and the expression of TOP2A is related to the size and invasiveness of tumors (25). In pancreatic adenocarcinoma, TOP2A may affect the prognosis and progression of cancer by regulating cell cycle and p53 signaling pathway (26). Since the expression of TOP2A is closely related to the prognosis of hepatocellular carcinoma, TOP2A may be an important biomarker of hepatocellular carcinoma (27).

Cell division cycle 20 (CDC20) is an essential regulatory protein for mitotic cell cycle, and interacts with several other proteins at multiple points during the cell cycle (28). And CDC20 is a key co-factor for promoting E3 ubiquitin ligase of compounds or late bodies (APC/C), so as to regulate the APC/C activity of ubiquitin on specific substrates (29). CDC20, which is significantly overexpressed while the development of malignant tumors, is related to the poor prognosis of various carcinomas. It has been proved that CDC20 may be a potential biomarker of pancreatic ductal adenocarcinoma (PDAC) (30).

Protein regulator of cytokinesis 1 (PRC1) is an important substrate of tubulin (31), and participates in the formation of the spindle through interacting with multiple proteins (32). PRC1 plays an important role in cytoplasmic separation, and is closely related to the occurrence of various tumors (33), especially in breast cancer (34). The TCGA database suggests that the expression of PRC1 is significantly higher in breast cancer tissues than in adjacent tissues.

KIAA0101 is a key biological factor regulating cell proliferation, and belongs to proliferating cell nuclear antigen binding factor family (35). Studies showed that KIAA0101 is overexpressed in cancer, and is involved in regulating DNA replication and repair, cell cycle progression, and cell proliferation (36). Silencing KIAA0101 gene in breast could promote the formation of P35/Sp1 complex and the inhibition of cell cycle (36).

Ubiquitin conjugating enzyme E2 T (UBE2T) is a
member of the E2 family of ubiquitin-binding enzymes. \( UBE2T \) has attracted more and more attention due to its repairing effect on DNA damage in Fanconi anemia (37). \( UBE2T \) regulates the level of mono-ubiquitination of \( FANCD2 \) and affects the repairing process of DNA damage within cells (38). This enzyme plays its oncogenic role by inhibiting the expression of \( BRCA1 \) gene (39). \( UBE2T \) also participates in cell proliferation, and apoptosis, and is up-regulated in osteosarcoma (40), bladder cancer (41), prostate cancer (42), nasopharyngeal cancer (41) and breast cancer (39).

Maternal embryonic leucine zipper kinase (\( MELK \)), cell cycle-dependent protein kinase, belongs to the KIN1/PAR-1/MARK family (43). \( MELK \) is an indispensable protein kinase of regulating cell cycle, cell proliferation, apoptosis and embryonic development (44). \( MELK \) is involved in many protein-protein interactions, and affects many stages of tumorigenesis (45).

Cell division cycle 45 (\( CDC45 \)) is an important component of DNA helicase complex in eukaryotic DNA replication to initiate DNA synthesis (46). It has been found that increased origin firing by up-regulate \( CDC45 \) could ultimately lead to cell apoptosis (47) Mutations in \( CDC45 \) may be a major cause for Meier-Gorlin syndrome and skull syndromes (48).

In our study, we identified 242 genes expressed in all the three datasets, and nine key genes (\( ASPM, CCNB2, CDC20, CDC45, MELK, PRC1, TOP2A, UBE2T \) and \( KIAA0101 \)) were associated with survival of lung adenocarcinoma patients by bioinformatics analysis, because there are many other factors (age and comorbidity) that affect the survival rate of lung adenocarcinoma patients. The result of our study may be a new indicators for further research on lung adenocarcinoma. We used qRT-PCR to further verify the relationship between these nine hug genes and lung adenocarcinoma. As shown Figure 8, \( ASPM, CCNB2, CDC20, CDC45, MELK, TOP2A \) and \( UBE2T \) were more significantly upregulated than \( PRC1 \) and \( KIAA0101 \) in lung adenocarcinoma patients tissue. Except for the effect of sample size, which cannot cover total races and regions, which may affect gene expression in lung adenocarcinoma. And another limitation is that we did not observes further the protein expression level of these genes via Western Blot or immunohistochemistry to validate the transcriptomic finding. Therefore, we will further study their relationship at the protein level in the further study.

At present, some studies about key genes related to lung adenocarcinoma in the database have been published. For example, Gao et al. identified 11 key genes, among which 10 genes were associated with development of NSCLC patients, from 4 GEO datasets by a bioinformatics method incorporating functional enrichment analysis, PPI network, the module analysis and overall survival analysis (49). Piao et al. found 14 hub genes play an important role in prognosis of NSCLC patients from 3 Gene expression profiles and 1 miRNA expression profile via using data preprocessing, functional and pathway enrichment, the protein-protein interaction network and overall survival analysis. Compared to previous works, we not only screened out the key genes and analyzed the functional enrichment, but verified the relationship between the expression of these genes and lung adenocarcinoma patients at the transcriptional level based on qRT-PCR.

**Conclusions**

In summary, we screened out seven key genes (\( ASPM, CCNB2, CDC20, CDC45, MELK, TOP2A \) and \( UBE2T \)) out of 242 DEGs from GSE32863, GSE43458 and GSE63459, which may play significant role during the progression of lung adenocarcinomas, and be potential biomarkers and therapeutic targets for lung adenocarcinomas. Nevertheless, further studies are needed to elucidate the molecular mechanisms underlying the identified genes in the pathogenesis of lung adenocarcinomas.

**Acknowledgments**

**Funding:** This work was supported by Zhejiang Provincial Natural Science Foundation of China (LQ16C090001).

**Footnote**

**Conflicts of Interest:** The authors have no conflicts of interest to declare.

**Ethical Statement:** The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Our study was approved by the Ethics Committee of the Affiliated Hospital of Qingdao University (No.2018015). Each patients was informed of the use of their tissue specimens.

**References**

1. Gao C, Zhuang J, Li He, et al. Exploration of methylation-


