ORIGINAL ARTICLE

Identification of differential protein-coding gene expressions in early phase lung adenocarcinoma

Li-Na Zhou[1]* , Shi-Cheng Li[2]* , Xue-Ying Li[1], Hong Ge[1] & Hong-Mei Li[1]

1 Department of Cancer, The Affiliated Hospital of Qingdao University, Qingdao, China
2 Department of Thoracic Surgery, The Affiliated Hospital of Qingdao University, Qingdao, China

Abstract

Background: The diagnosis of early phase lung adenocarcinoma (LADC) is associated with therapeutic strategy, effect, and survival time. However, the sensitive biomarkers of early phase LADC are still unclear. This study aimed to identify protein-coding genes that can be used as biomarkers of early stage LADC.

Methods: Gene microarray analysis was performed to identify key hub genes that show different expression in lung adenocarcinoma compared to normal tissues. The microarray data of lung adenocarcinoma in stages IA, IB, IIA, IIB, and normal tissues (GSE10072) were downloaded from a free online database, Gene Expression Omnibus (GEO).

Results: A total of 572 differentially expressed genes (DEGs) were identified between early phase lung adenocarcinoma and normal tissues using R software. Database for Annotation, Visualization and Integrated Discovery online tools were used to obtain Gene Ontology analysis and the Kyoto Encyclopedia of Genes and Genomes was used to analyze DEGs. Cytoscape software was used to express the protein-protein interaction network. We found that some cancer-related Gene Ontology terms and pathways (e.g. cell adhesion, cell surface receptor signaling pathway, PI3K-Akt signaling pathway) were significantly enriched in DEGs.

Conclusion: Protein-coding genes *JUN*, *FYN*, *CAV1*, and *SFN* may play vital roles in the progress of early-stage lung adenocarcinoma. Consequently, through bioinformatics analysis, the key genes could be established to provide more potential references for the therapeutic targets of lung adenocarcinoma.

Introduction

Lung cancer is the most common cancer in the world, associated with significant morbidity and mortality.[1] The incidence of lung cancers continues to rise in China. Like other cancers, lung cancer is a heterogeneous disease that carries high rates of genetic somatic mutations, copy number alterations, and chromosomal rearrangement.[2] Lung adenocarcinoma (LADC) is a familiar histological type that accounts for 40% of all lung cancers and is one of the best genetically characterized human epithelial malignancies.[3] Although treatment methods, including surgery, chemotherapy, radiotherapy, and targeting therapy are improving, the five-year overall survival rate remains lower than 15%.[1] Consequently, elucidating the molecular mechanism involved in the proliferation of

LADC is important for the development of efficacious diagnosis and treatment strategies.

Gene expression analysis based on high-throughput techniques, such as microarrays, is widely used. This useful research method is increasingly being utilized as a promising tool to improve the molecular diagnosis and classification of cancers, as well as to determine new drug targets. Meanwhile, microarray technology can simultaneously detect thousands of genes to locate the differentially expressed genes (DEGs) involved in different pathways, molecular functions, or biological characteristics. Many DEGs play an important role in lung adenocarcinoma proliferation and progression, and could be valued as potential molecular targets or efficacious diagnosis markers. A recent study analyzed data with expression array analysis, Gene

Ontology (GO), and pathway analyses to selected genes, and after a series of trials, finally found that TRIM58 over-expression suppresses LADC cell proliferation and oncogenesis in vitro and in vivo, suggesting that this gene acts as a tumor suppressor gene (TSG) during the initiation and/or early progression stages of LADC.[4] Using microarray technology combining bioinformatics analysis, Griesing *et al.* demonstrated that miR-532-5p is directly regulated by TTF-1 and plays a vital role in tumor suppression by targeting KRAS and MKL2 in LADC.[5] To date, microarray screening has been utilized to identify DEGs in various types of cancer to better understand the mechanisms underlying cancer progression.

Original data (GSE10072) for our study was downloaded from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). We compared DEGs between LADC and normal samples, and performed function and pathway enrichment analysis. Additionally, a protein–protein interaction (PPI) network of all of the DEGs was constructed. Genes that showed higher degrees in the PPI networks were selected as key genes in LADC. Further, several transcription factors and their binding sites were studied.

Methods

Microarray data

The microarray data obtained for this study were available in the GEO database under the expression dataset profile of GSE10072, based on the GPL96 platform. The GSE10072 dataset consists of 107 ADC and non-tumor tissue samples, and includes tumor node metastasis (TNM) stages and survival data.

Identification of differentially expressed genes (DEGs)

Analysis of the GSE10072 dataset was carried out using Affymetrix Human Genome U133A Array (ThermoFisher Scientific, Waltham, MA, USA). A processed series matrix file was then downloaded for microarray data analysis. After the series matrix data was background normalized using the Robust Multichip Averaging (RMA) method, the data was subjected to DEG selection. The DEGs were defined by the Limma algorithm (R software, R Foundation for Statistical Computing, Vienna, Austria). A P value of < 0.05 and $|logFC| > 2$ were considered as significant.

Bioinformatic analysis

The GO (http://www.geneontology.org) is a widely used bioinformatics resource that supplies information about the annotation of genes, molecular functions, and cellular components.[6,7] The Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/) is an integrated database that provides the molecular functions of genes and proteins, genome sequences, and high-throughput data.[8] The Database for Annotation, Visualization and Integrated Discovery (DAVID, https://david.ncifcrf.gov/) is an essential foundation for visualization, annotation, and integrated findings.[9] We performed GO enrichment analysis and KEGG pathway enrichment for DEGs using the DAVID online tool. The Retrieval of Interacting Genes (STRING) database, was then used to evaluate protein–protein interaction (PPI) information to identify the interaction relationships between DEGs. Cytoscape software (National Institute of General Medical Sciences, Bethesda, MD, USA) was used to construct PPI networks for the identified hub genes between DEGs.

Survival analysis

Kaplan–Meier plots and Cox regression were employed to visualize the association between the genes under investigation and survival to explore the prognostic value of differentially expressed coding genes. Before analysis, the samples were classified into two groups: high expression or low expression. Survival analysis was then compared
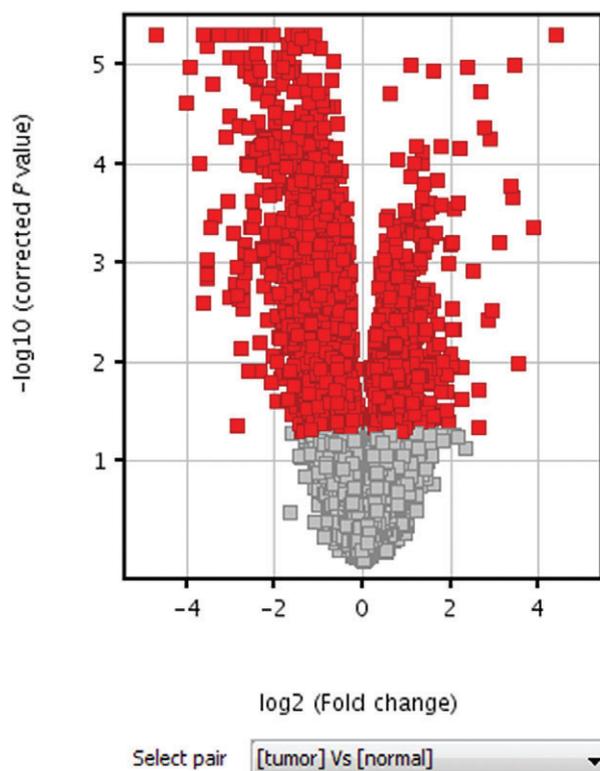

Figure 1 A volcano map using log2 as a standard to demonstrate the expression of differentially expressed genes (DEGs).
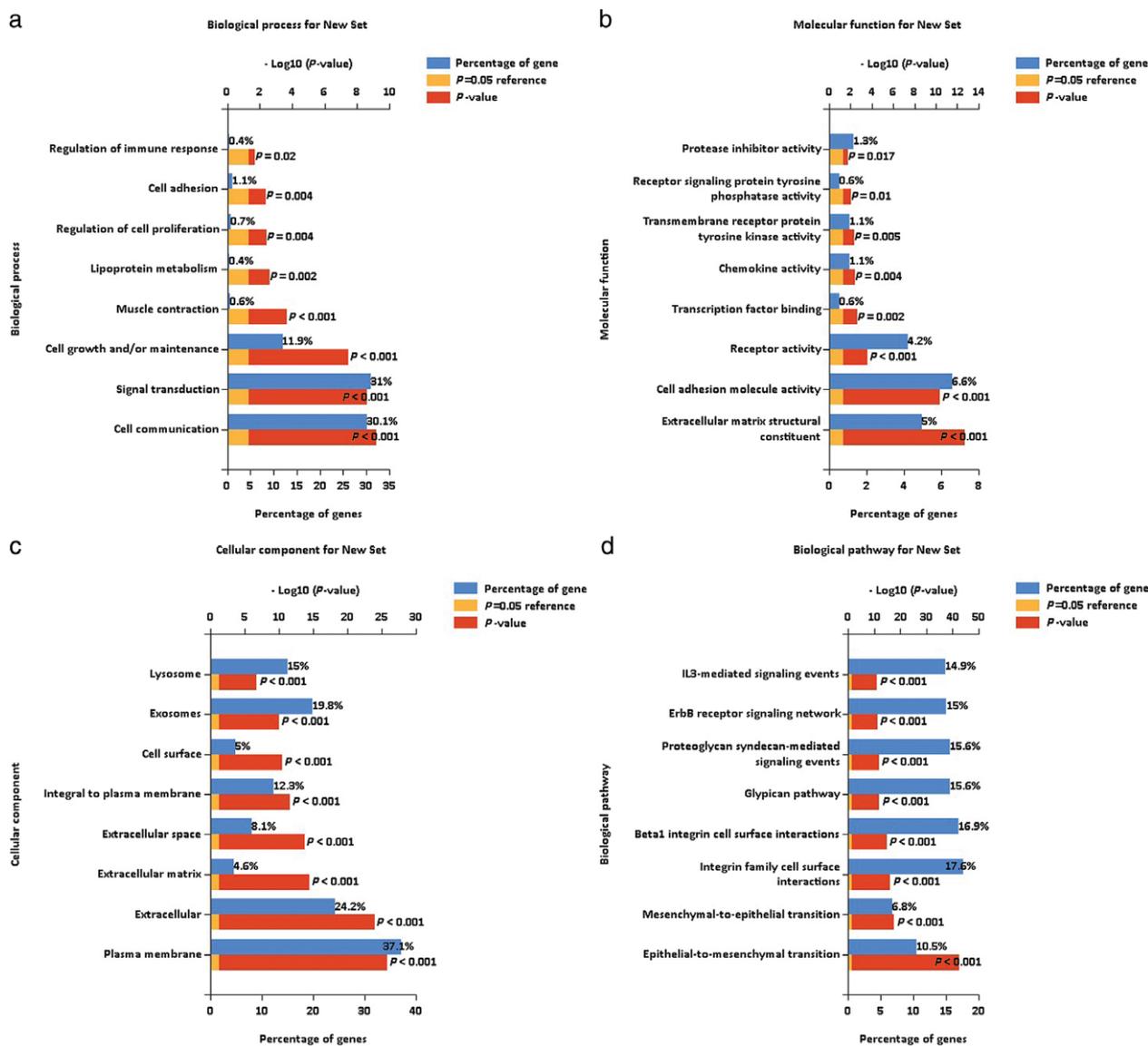
Figure 2 Gene Ontology analysis results showed that differentially expressed genes (DEGs) were significantly enriched in (**a**) biological processes (BP), (**b**) molecular functions (MF), and (**c**) cell components (CC). (**d**) Some cancer-related pathways were enriched in the DEGs, such as Beta1 integrin cell surface interactions, glypican pathways, and the ErbB receptor signaling network.

between the groups.[10] We conducted statistical analysis using ONCOMINE and the Kaplan–Meier Plotter database. Statistical significance was examined using the Student's *t*-test. *P* < 0.05 or *P*<0.01 were considered statistically significant.

Results

DEGs in lung adenocarcinoma

A total of 13 987 gene expression values were obtained from 107 samples after raw data pretreatment using the RMA method. R statistical analysis software was used to

define the DEGs. A total of 572 DEGs were revealed by comparing 31 gene expression profiles from normal tissues and 33 profiles from LADC, of which 420 were upregulated and 152 were downregulated in LADC. A volcano map of DEG expression in the samples is shown in Figure 1.

Functional enrichment analysis

GO or KEGG pathway enrichment analysis could reveal molecular functions or biological pathways and were performed for the DEGs (Fig 2). GO analysis showed that DEGs were significantly enriched in biological processes (BP), molecular functions (MF), and cell components
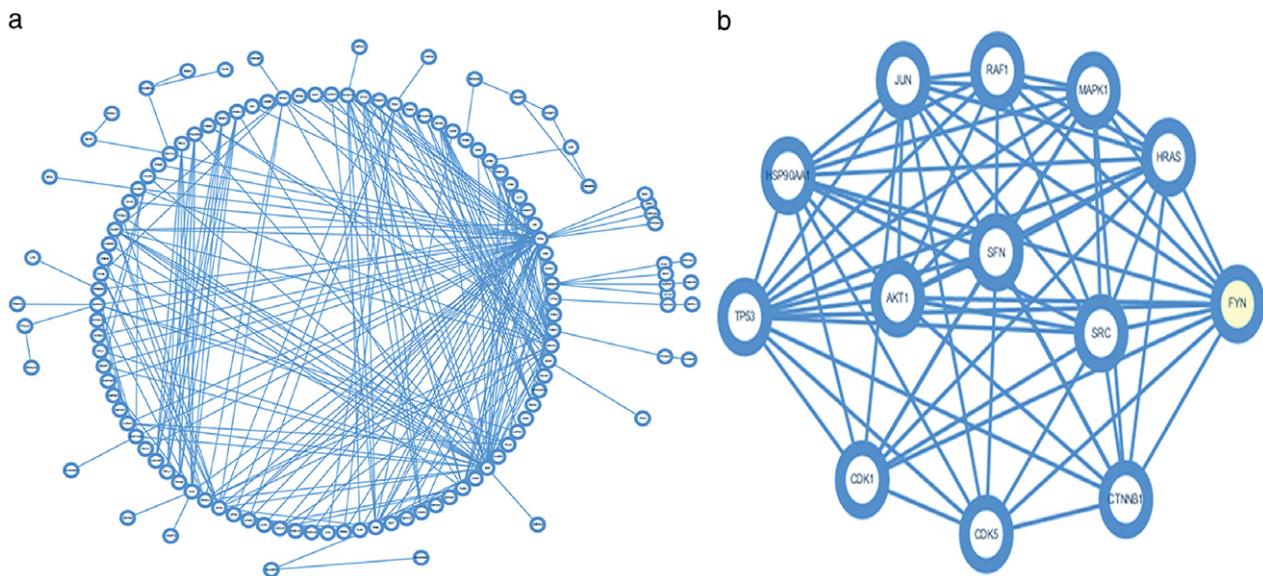
a

b

Figure 3 (**a**) The protein–protein network, which consisted of 423 nodes interacting at 3217 edges; and (**b**) the hub proteins and the number of interactions.

(CC), including cell communication, cell adhesion molecule activity, and plasma membrane. Some cancer-related pathways were enriched in the DEGs, such as Beta1 integrin cell surface interactions, glypican pathways, and the ErbB receptor signaling network (Fig 2).

Protein-protein interaction network construction and identification of hub genes

The STRING database and Cytoscape software were used to obtain further insight into the interaction between DEGs. The PPI network consisted of 423 nodes interacting at 3217 edges, while the remaining 149 DEGs did not match the PPI pairs. Meanwhile, the top four hub nodes with higher degrees were screened. These hub genes, *JUN*, *FYN*, *CAV1* and *SFN*, could code relative proteins, considered the hub proteins, which are more likely to play a vital role in the genesis and growth of cancer. The hub proteins and their interactions are depicted in Figure 3.

Survival analysis of hub genes

To test the potential prognostic roles of these hub genes, we analyzed whether there was any association to survival in lung cancer patients. Kaplan–Meier survival and univariate Cox proportional hazards regression analyses were performed for the hub genes, and results showed that expression levels of three genes, *JUN*, *CAV1* and *FYN*, were significantly associated with survival in lung cancer patients. High expression of these three genes led to

prolonged survival (Fig 4). To further confirm our findings, we consulted ONCOMINE. The results showed that *SFN* was significantly highly expressed in lung cancer tissues, thus high expression of SFN was associated with shorter survival (Fig 4).

Discussion

Lung cancer is a heterogeneous disease, a product of genetic somatic mutation, copy number alteration, and chromosomal rearrangement. LADC incidence is the most common among different pathological types. Therefore, understanding the molecular mechanism of LADC is vital for accurate diagnosis, treatment, and prognosis. Microarray and high-throughput analysis could provide expression levels of thousands of human genes and thus has been widely used to determine potential diagnosis biomarkers and therapeutic targets for LADC. In our study, we analyzed the microarray data from GSE10072 and identified 420 upregulated and 152 downregulated DEGs between LADC and normal tissue profiles. Using DAVID, bioinformation analysis showed that these DEGs were mainly involved in enzyme inhibitor activity, chemokine signaling pathways, and cytokine-cytokine receptor interaction. A PPI network was used to obtain further insight into the association between DEGs. The top four hub nodes with higher degrees were screened. These hub genes, *JUN*, *FYN*, *CAV1*, and *SFN*, may represent potential biomarkers or provide new ideas for therapeutic studies.

In the present study, gene expression of 107 samples was retrieved from the GEO dataset under the accession
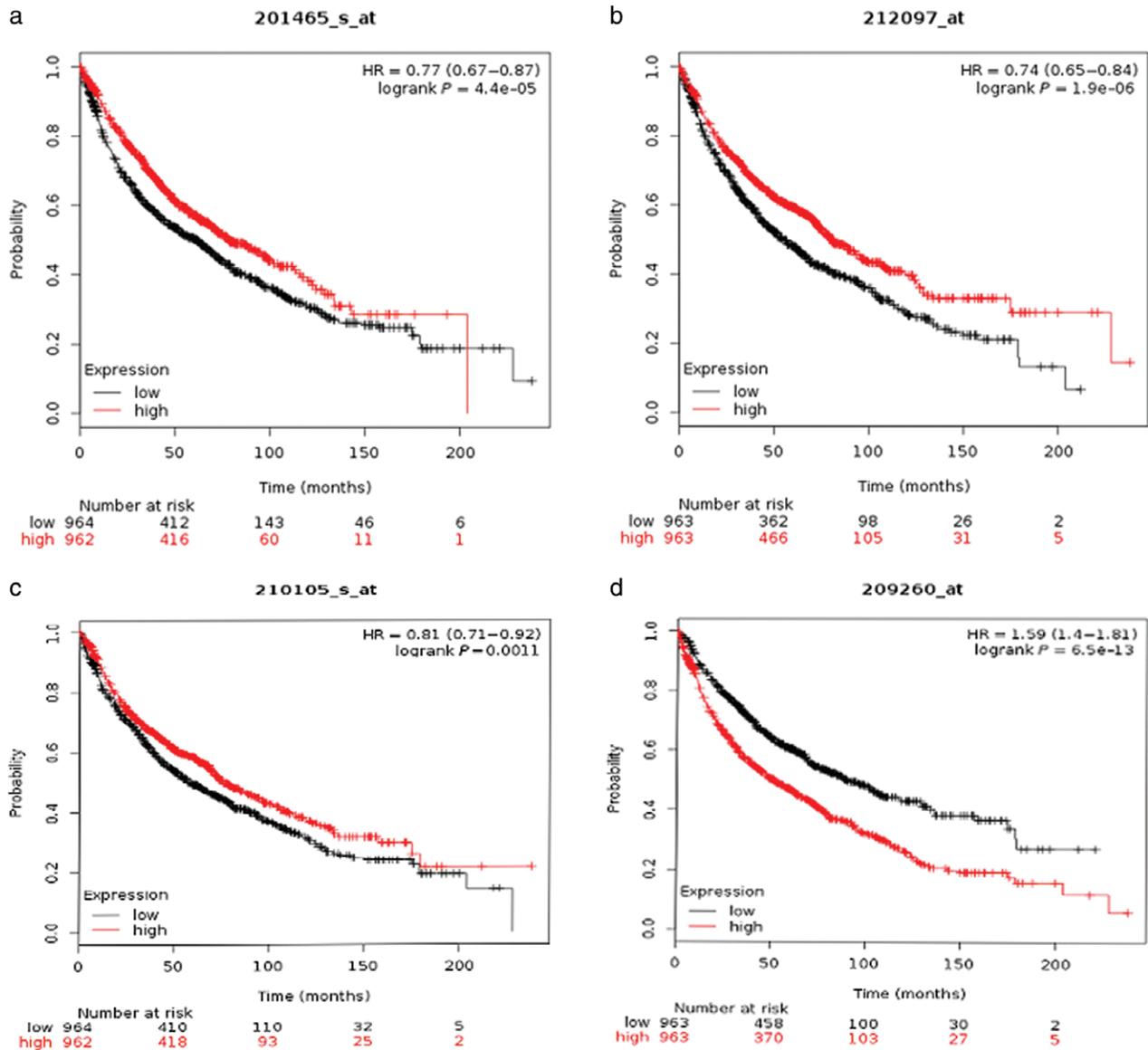
Figure 4 A patient with high (**a**) *JUN*, (**b**) *CAV1*, and (**c**) *FYN* expression can achieve longer survival. (**d**) A patient with high *SFN* expression would experience shorter survival. HR, hazard ratio.

GSE10072, including TNM stage and survival data. Using an RMA method, a total of 13 987 genes were obtained from these samples. After comparing the gene expression profiles of 31 normal tissues and 33 LADC, 572 DEGs were revealed, among which 420 were upregulated and 152 downregulated. Cumulative evidence has demonstrated that a group of genes with similar expression were made up of co-expression genes, which also take part in the biological process.[11] In order to gain deeper insight into the interaction of DEGs, we performed GO and KEGG pathway analyses.

The results of GO analysis showed that upregulated DEGs were significantly enriched in BP, including cellular component organization or biogenesis, biological regulation, and the developmental process; downregulated DEGs were significantly enriched in BP, including the metabolic process, cell communication, and cell cycle. This finding proved that cell component organization or biogenesis, the cell metabolic process and cell cycle is associated with tumorigenesis and prognosis. The cell cycle is a tightly integrated process that is frequently aberrant in non-small cell lung cancer (NSCLC).[12] It has been reported that for rapid cell growth and proliferation, tumors have high energetic and anabolic needs. Therefore, cell metabolic component organization and biogenesis is an important source of metabolic intermediates. Furthermore, KEGG analysis

showed that upregulated DEGs were significantly enriched in ECM-receptor interaction, PI3K-AKT signaling, and other cancer-related pathways. Previous studies have shown that EMC plays a vital role in maintaining stem cell properties and regulating stem cell differentiation. Dysregulation of EMC regulates the development and progression of tumors by promoting the tumor microenvironment.[13–15] Moreover, a disturbance in the PI3K-AKT signaling pathway is associated with resistance to EGFR-TKI therapy and lower survival in NSCLC patients.[16] Downregulated DEGs were associated with focal adhesion, the p53 signaling pathway, and other functions. Dy *et al.* reported that focal adhesion kinase is expressed in more than 50% of stage I NSCLC patients.[17] In addition, other articles have demonstrated that focal adhesion kinase is a promising factor to predict aggressive behavior and prognosis in NSCLC patients, particularly with the ADC subtype.[18] It has been reported that cisplatin causes p53 overexpression, which may represent a predictive marker for chemotherapy efficacy in LADC.[19] p53 is also involved in the immune response by regulating PD-L1.[20] Therefore, monitoring these singling pathways may improve the measure of diagnosis and the prediction of therapy.

After functional enrichment analysis, we constructed a PPI network with DEGs, and then determined the top degree hub genes: *JUN*, *FYN*, *CAV1*, and *SFY*. The first hub gene, *JUN*, plays an important role in cell proliferation, the suppression of apoptosis, and increased angiogenesis. *c-JUN* integrates signals of several developmental pathways, including EGFR-TKI and activin B-MAP3K1-JNK for cell proliferation and differentiation.[21] The second hub gene, *FYN*, belongs to the Src family kinase (SFKs), which is involved in signal transduction or the susceptibility of cancer cells to some anti-cancer treatments. In tamoxifen-resist breast cancer cell lines, *FYN* regulates and plays a key role in resistant mechanisms.[22] Recent studies have demonstrated that *FYN* together with other *SFK* members is a mediator of growth-factor induced anti-apoptotic activity of Akt/PKB. Several cancers, including breast, ovarian, and oral carcinomas are associated with *FYN*, as such *FYN* inhibitors are being investigated.[23] The last hub gene, *SFN*, a potent anti-cancer compound, inhibits cancer stem-like cell properties and enhances the therapeutic efficacy of cisplatin in human NSCLC.[24] Some articles have reported that *SFN* can induce apoptosis in a variety of cancers via multiple mechanisms. SFN-Cys has been used to inhibit invasion in human prostate cancer cells. Other trials have found that SFN-Cys triggers *ERK1/2* activation to mediate the mitochondria signaling pathway with upregulation and α-tubulin downregulation leading to NSCLC cell apoptosis.[25]

In conclusion, our data identified four key genes and some pathways related to LADC proliferation, progression, and apoptosis using a series of bioinformation analyses on DEGs between LADC and normal tissues. Previous studies have indicated that early-stage LADC corrects the CAM pathway.[26] The results of this study provide a series of useful targets for future investigation into the molecular mechanisms and biomarkers of LADC. Previous studies that used the complementary DNA microarray system found that numerous alterations of gene expression were present in early-stage cancer.[27] However, some of functions of those genes are still unclear, therefore, further molecular biological experiments are required to determine the biomarkers and potential therapeutic targets of early stage LADC.

Disclosure

No authors report any conflict of interest.

References

1 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA Cancer J Clin* 2016; **66**: 7–30.

2 Lawrence MS, Stojanov P, Polak P *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013; **499**: 214–8.

3 Bender E. Epidemiology: The dominant malignancy. *Nature* 2014; **513**: S2–3.

4 Kajiura K, Masuda K, Naruto T *et al.* Frequent silencing of the candidate tumor suppressor TRIM58 by promoter methylation in early-stage lung adenocarcinoma. *Oncotarget* 2017; **8**: 2890–905.

5 Griesing S, Kajino T, Tai MC *et al.* Thyroid transcription factor-1-regulated microRNA-532-5p targets KRAS and MKL2 oncogenes and induces apoptosis in lung adenocarcinoma cells. *Cancer Sci* 2017; **108**: 1394–404.

6 Gene Ontology Consortium. Gene ontology consortium: Going forward. *Nucleic Acids Res* 2015; **43**: D1049–56.

7 Gene Ontology Consortium. The gene ontology (GO) project in 2006. *Nucleic Acids Res* 2006; **34**: D322–6.

8 Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016; **44**: D457–62.

9 Dennis GJ, Sherman BT, Hosack DA *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003; **4**: P3.

10 Győrffy B, Surowiak P, Budczies J, Lánczky A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. (Published erratum appears in PLoS One 2014; 9: e111842.). *PLoS One* 2013; **8**: e82241.

11 Liang B, Li C, Zhao J. Identification of key pathways and genes in colorectal cancer using bioinformatics analysis. *Med Oncol* 2016; **33**: 111.

12 DeNicola GM, Chen PH, Mullarky E *et al*. NRF2 regulates serine biosynthesis in non-small cell lung cancer. *Nat Genet* 2015; **47**: 1475–81.

13 Shen Q, Wang Y, Kokovay E *et al*. Adult SVZ stem cells lie in a vascular niche: A quantitative analysis of niche cell–cell interactions. *Cell Stem Cell* 2008; **3**: 289–300.

14 Raymond K, Deugnier MA, Faraldo MM, Glukhova MA. Adhesion within the stem cell niches. *Curr Opin Cell Biol* 2009; **21**: 623–9.

15 Bhowmick NA, Neilson EG, Moses HL. Stromal fibroblasts in cancer initiation and progression. *Nature* 2004; **432**: 332–7.

16 Pérez-Ramírez C, Cañadas-Garre M, Molina MÁ, Faus-Dáder MJ, Calleja-Hernádnez MÁ. PTEN and PI3K/AKT in non-small-cell lung cancer. *Pharmacogenomics* 2015; **16**: 1843–62.

17 Dy GK, Ylagan L, Pokharel S *et al*. The prognostic significance of focal adhesion kinase expression in stage I non-small-cell lung cancer. *J Thorac Oncol* 2014; **9**: 1278–84.

18 Ji HF, Pang D, Fu SB *et al*. Overexpression of focal adhesion kinase correlates with increased lymph node metastasis and poor prognosis in non-small-cell lung cancer. *J Cancer Res Clin Oncol* 2013; **139**: 429–35.

19 Hirai Y, Shimizu Y, Oura S *et al*. [The importance of p53 and ras mutation as predictive markers for adjuvant chemotherapy in non-small-cell lung cancer]. *Gan To Kagaku Ryoho* 2012; **39**: 381–4. (In Japanese.)

20 Cortez MA, Ivan C, Valdecanas D *et al*. PDL1 regulation by p53 via miR-34. *J Natl Cancer Inst* 2016; **108**: pii: djv303.

21 Meng Q, Xia Y. c-Jun, at the crossroad of the signaling network. *Protein Cell* 2011; **2**: 889–98.

22 Elias D, Ditzel HJ. Fyn is an important molecule in cancer pathogenesis and drug resistance. *Pharmacol Res* 2015; **100**: 250–4.

23 Schenone S, Brullo C, Musumeci F, Biava M, Falchi F, Botta M. Fyn kinase in brain diseases and cancer: The search for inhibitors. *Curr Med Chem* 2011; **18**: 2921–42.

24 Li QQ, Xie YK, Wu Y *et al*. Sulforaphane inhibits cancer stem-like cell properties and cisplatin resistance through miR-214-mediated downregulation of c-MYC in non-small cell lung cancer. *Oncotarget* 2017; **8**: 12067–80.

25 Lin K, Yang R, Zheng Z *et al*. Sulforaphane-cysteine-induced apoptosis via phosphorylated ERK1/2-mediated maspin pathway in human non-small cell lung cancer cells. *Cell Death Discov* 2017; **3**: 17025.

26 Saji H, Tsuboi M, Shimada Y *et al*. Gene expression profiling and molecular pathway analysis for the identification of early-stage lung adenocarcinoma patients at risk for early recurrence. *Oncol Rep* 2013; **29**: 1902–6.

27 Nakamura H, Saji H, Ogata A *et al*. cDNA microarray analysis of gene expression in pathologic stage IA nonsmall cell lung carcinomas. *Cancer* 2003; **97**: 2798–805.