doi: 10.4149/gpb_2020037

Identification of differentially expressed genes associated with lung adenocarcinoma *via* bioinformatics analysis

Xinmeng Yang[1], Qingchuan Feng[1], Jianan Jing[1], Jiahui Yan[1], Zhaoshu Zeng[2], Hao Zheng[1] and Xiaoli Cheng[1]

[1] *Department of Medical Genetics and Cell Biology, School of Basic Medical Sciences, Zhengzhou University, Zhengzhou, Henan, China*
[2] *Department of Forensic Medicine, School of Basic Medical Sciences, Zhengzhou University, Zhengzhou, Henan, China*

Abstract. Lung adenocarcinoma (LUAD) with extremely high morbidity as well as mortality is still in the exploration stage of pathogenesis and treatment. This study aimed to screen and identify differentially expressed genes (DEGs) associated with LUAD *via* bioinformatics analysis. Three LUAD microarray datasets, GSE116959, GSE68571 and GSE40791, were selected from the Gene Expression Omnibus (GEO) database to analyze the DEGs. 128 DEGs were identified in all, incorporating 36 upregulated and 92 downregulated. Function and pathway enrichment analyses showed that metabolic pathways were their main signaling pathways. After that, seven hub genes including *VWF, SPP1, PECAM1, TOP2A, CDK1, UBE2C* and *KIF23* were mined by the protein-protein interaction (PPI) network. Gene expression analysis, TNM and survival analysis of these hub genes were performed *via* Gene Expression Profiling Interactive Analysis (GEPIA) online database. Further analysis indicated that *TOP2A, CDK1, UBE2C* and *KIF23* were related to the stage of LUAD patients and overall survival. Then, we verified the relative expression levels of *TOP2A, CDK1, UBE2C* and *KIF23* in LUAD cell lines by qRT-PCR. In conclusion, this study indicated that the four hub genes screened out by bioinformatics analysis were differentially expressed in LUAD compared to normal sample and might be prognostic markers of LUAD.

Key words: Lung adenocarcinoma — Differentially expressed genes — Microarray datasets — Bioinformatics analysis — Hub genes

Abbreviations: BP, biological processes; CC, cellular component; DAVID, Database annotation for visualization and integrated discovery; DEGs, differentially expressed genes; GEO, Gene Expression Omnibus; GEPIA, Gene expression profiling interactive analysis; GO, gene ontology; GTEx, Genotype-Tissue Expression; KEGG, Kyoto encyclopedia of genes and genomes; LUAD, lung adenocarcinoma; MF, molecular function; MNC, Maximum Neighborhood Component; NSCLC, non-small cell lung cancer; OS, overall survival; PPI, protein-protein interaction; qRT-PCR, quantitative reverse transcriptase real time PCR; SCLC, small cell lung cancer; STRING, Search Tool for the Retrieval of Interacting Genes; TCGA, The Cancer Genome Atlas.

Electronic supplementary material. The online version of this article (doi: 10.4149/gpb_2020037) contains Supplementary Material.

Correspondence to: Xiaoli Cheng, Department of Medical Genetics and Cell Biology, School of Basic Medical Sciences, Zhengzhou University, Zhengzhou 450001, Henan, China
E-mail: chengxiaolizzu@163.com

Introduction

Lung cancer is a general malignant tumor which has a leading morbidity and mortality worldwide. According to statistics, the number of freshly diagnosed lung cancer patients around the world in 2018 was 2.094 million, and the number of lung cancer deaths worldwide was 1.761 million, ranking first in cancer (Bray et al. 2018). Lung cancer consists of small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). The latter can be divided into squamous cell lung carcinoma, lung adenocarcinoma (LUAD) and large cell lung carcinoma, accounting for 80 to 85% of lung cancer (Wood. 2015). LUAD is the most common histological subtype of primary lung cancer. Nowadays, the cause of NSCLC is not clear, and the early symptoms are not obvious. When patients are diagnosed, they are already in the middle and advanced stages of cancer, especially LUAD. Although humans have made some achievements in the pathogenesis and treatment of LUAD, the average overall survival (OS) of LUAD patients is less than five years (Siegel et al. 2017; Denisenko et al. 2018). Therefore, it is of great significance to further understand the pathogenesis of LUAD at the molecular level.

With the development of bioinformatics analysis methods, the Gene Expression Omnibus (GEO) online public databases has become common tools to select DEGs, which is helpful for studying molecular signals and their relationships as well as building gene interaction networks (Clough et al. 2016). So far, through comprehensive analysis of single or multiple microarray datasets in the GEO database, some studies have identified some genes closely related to the development of LUAD. For example, Zhang et al. (2018b) found that *AURKA*, *CDC20* and *TPX2* could be potential biomarkers for predicting poor prognosis of smoking-related LUAD. Zhou et al. (2018) found that protein-encoding genes, including *JUN*, *FYN*, *CAV1* and *SFN* were associated with the progression of early LUAD. Guo et al. (2019) found that *MYH10*, *METTL7A*, *FCER1G* and *TMOD1* might play an important role in the occurrence and development of LUAD. However, the LUAD development mechanism is still not comprehensive and systematic, and further research is needed.

In this study, a total of three microarray datasets, GSE116959, GSE68571 as well as GSE40791 were obtained from the GEO database for bioinformatics analysis, including 121 normal tissues and 237 tumor tissues. Then 128 differentially expressed genes (DEGs) were screened totally for gene ontology (GO) functional annotation analysis along with Kyoto encyclopedia of genes and genomes (KEGG) pathway enrichment analysis. After that, a protein-protein interaction (PPI) network was constructed by Search Tool for the Retrieval of Interacting Genes (STRING) database to screen hub genes using Cytoscape software. At last, the results were verified with GEPIA database and quantitative reverse transcriptase real time PCR (qRT-PCR), which provided clues and basis for further studying the mechanism of LUAD development.

Materials and Methods

Data collection

The LUAD related microarray datasets were obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). The search terms included "lung adenocarcinoma", "homo sapiens", "expression profiling by array", and we chose the datasets including normal tissues and tumor tissues. GSE116959 (Moreno et al. 2019) was a relatively new dataset that had not yet been used for bioinformatics analysis. It studied transcriptome profiling of 57 LUAD samples and 11 peritumoral normal lung tissues; GSE68571 (Beer et al. 2002) identified a set of genes that predict survival in early-stage LUAD, which contained 86 tumor tissues and 10 normal tissues. GSE40791 (Zhang et al. 2012) had a large sample size, which included 100 non-neoplastic (N) lung samples, and 69 stage I, 12 stage II, 13 stage III LUAD frozen tissues. The above three datasets met the basic screening conditions, with a total of 121 normal tissues and 237 tumor tissues and the sample diversity is rich. Therefore, we chose these three datasets for further analysis.

Identification of DEGs

DEGs were identified from three GEO microarray datasets using GEO2R (https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html) online analysis tool which is based on GEOquery as well as Limma R packages of Bioconductor project. We defined DEGs that met the two screening conditions of adjusted (adj.) p value < 0.05 as well as |log2 fold change (log2FC)| > 1.5 were statistically significant. p value is corrected by the method of Benjamini & Hochberg false discovery rate, which is the most commonly used adjustment of microarray data, and provides a good balance between the discovery of statistically significant genes and the limitation of false positives (Benjamini et al. 2001). Venny 2.1.0 (http://bioinfogp.cnb.csic.es/tools/venny/index.html) was used to intersect the screening results of the three GEO microarray datasets to obtain the DEGs that were commonly upregulated and downregulated.

Functional annotation and enrichment analysis of DEGs

To understand the biological functions of DEGs, DAVID (version 6.7, https://david-d.ncifcrf.gov/) database was used for GO functional annotation (Huang et al. 2007), and results with $p < 0.05$ were statistically significant. The threshold of EASE Score, a modified Fisher Exact p value, for gene-enrichment analysis and p value is corrected by the

method of Benjamini. Then, KOBAS 3.0 (http://kobas.cbi.pku.edu.cn/kobas3) was used for KEGG enrichment analysis (Kanehisa et al. 2017). Finally, we visualized the results by R package ggplot2. Corrected p value < 0.05 indicated statistical significance. p value is corrected by the method of Benjamini & Hochberg.

PPI network construction and hub genes identification

STRING (version 11.0, https://string-db.org/) database, which has 5090 organisms and 24.6 million proteins (Szklarczyk et al. 2019), was used for PPI network construction PPI relationship analysis between DEGs. The minimum required interaction was set as "medium confidence = 0.4" and the max number of interactors was set as "none". Imported the analyzed data into Cytoscape (version 3.6.1, http: //www.cytoscape.org/) software, using the Maximum Neighborhood Component (MNC) and Degree algorithms to select the hub genes. MNC and Degree are two topology-based scoring methods for screening hub genes in Cytoscape's cytohubba plugin. Among them, MNC is the score of a gene node v, which is defined to be the size of the maximum connected component of subnetwork N(v); the subnetwork N(v) is induced by the nodes adjacent to gene node v. Degree is the number of nodes directly connected to a gene node v. Hub gene are defined as genes with high correlation in candidate modules. High connectivity means that the connectivity ranked at top 10%. To facilitate calculations, we selected the top 10 DEGs of MNC and Degree, and then taking the intersection to obtain the hub genes.

Validation of hub genes

GEPIA (http://gepia.cancer-pku.cn/) database online analysis tool, which is based on The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) database, providing differential expression analysis, profiling plotting, survival analysis and so on (Tang et al. 2017), was used for hub genes expression analysis and survival analysis.

Cell lines and cell culture

MRC-5 fibroblasts (human fetal lung) as well as Human LUAD cell lines A549 and H1650 were obtained from Chinese Academy of Sciences Cell Bank/Stem Cell Bank (Shanghai, China). MRC-5 cell line was cultured in EMEM medium supplemented with 10% fetal bovine serum and 1% penicillin and streptomycin at 37°C in a humidified atmosphere of 5% CO_2. A549 and H1650 were cultured in DMEM medium and RPMI 1640 medium, respectively. 500,000 MRC-5 and A549 as well as H1650 cells were respectively seeded in the wells of the six-well plate for RNA extraction, and the experiment was repeated three times.

RNA extraction and qRT-PCR

Total RNA was extracted from the cultured cells using Trizol Reagent following the manufacturer's protocol. cDNA was synthesized with FastKing RT Kit (with gDNase), and qRT-PCR was performed with TB Green TM Premix Ex Taq TM (Tli RNaseH Plus). All experiments were performed in triplicate. Data were normalized to *β-actin* and were analyzed using the comparative Ct ($2^{-\Delta\Delta Ct}$) method for quantification.

Immunohistochemical analysis

The Human Protein Atlas database (http://www.proteinatlas.org) was used to analyze the protein expression of DEGs in LUAD tissue. According to the staining intensity of the protein in the tissue and the percentage of stained cells, compare the difference in protein expression of the DEGs in normal tissue and tumor tissue, and intercept representative immunostaining images.

Results

Screening of DEGs

GEO is a public repository that archives and freely distributes comprehensive sets of microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. In addition to data storage, a collection of web-based interfaces and applications are available to help users query and download the studies and gene expression patterns stored in GEO. This study selected three GEO microarray datasets, GSE116959, GSE68571 and GSE40791 (Table 1 showed the detailed information of

Table 1. GEO datasets information

Dataset	GSE116959	GSE68571	GSE40791	Total
Platform	GPL17077	GPL80	GPL570	–
Year	2019	2015	2013	–
Normal Tissues	11	10	100	121
Tumor Tissues	57	86	94	237
Total Tissues	68	96	194	358
Genes	32080	5601	22189	59870
Upregulated DEGs	303	116	668	1087
Downregulated DEGs	674	156	1262	2092
Total DEGs	977	272	1930	3179

GEO, Gene Expression Omnibus; DEGs, differentially expressed genes.

these GEO microarray datasets). There were 358 samples totally, including 237 tumor tissues and 121 normal tissues. Next, we used GEO2R tool to analyze the GEO microarray datasets and we filtered DEGs by the two conditions of "adj. p value < 0.05" and "|log2FC| > 1.5". According to the statistics, GSE116959 contains 32,080 genes and 977 DEGs were screened, including 303 upregulated and 674 downregulated; GSE68571 contains 5601 gene and 272 DEGs were screened, including 116 upregulated and 156 downregulated; GSE40791 contains 22189 gene and 1930 DEGs were screened, including 668 upregulation and 1262 downregulated. The DEGs volcano plots of three GEO datasets were shown in Figure 1A–C. As shown in Figure 1D and E, through Venny 2.1.0, the upregulated and downregulated DEGs of three microarray datasets were analyzed, and 36 common upregulated DEGs and 92 common downregulated DEGs were obtained and Table 2 showed the details of 128 DEGs in three GEO datasets. Then we performed enrichment analysis on the DEGs that have been screened out to predict which pathways these genes were mainly concentrated in.

GO functional annotation and KEGG enrichment analysis of DEGs

The GO/KEGG enrichment analysis provided important clues for exploring the mechanism of cancer development. GO functional annotation of DEGs was performed by DAVID database which included biological process (BP), cellular component (CC) as well as molecular function (MF) in total three parts. The top 12 upregulated GO terms and the top 15 downregulated GO terms were summarized in Table 3. The result of GO functional annotation indicated that the upregulated DEGs were mostly concentrated in regulation of apoptosis (BP), nuclear lumen and organelle lumen (CC), as well as metalloendopeptidase and metallopeptidase activity (MF), as shown in Figure 2A. The downregulated DEGs were mostly concentrated in regulation of cell proliferation (BP), plasma membrane (CC) and calcium ion binding (MF), as shown in Figure 2B. GO functional annotation showed that the upregulated genes played a role in LUAD by regulating the biological process of apoptosis, while downregulated genes promoted the occurrence and

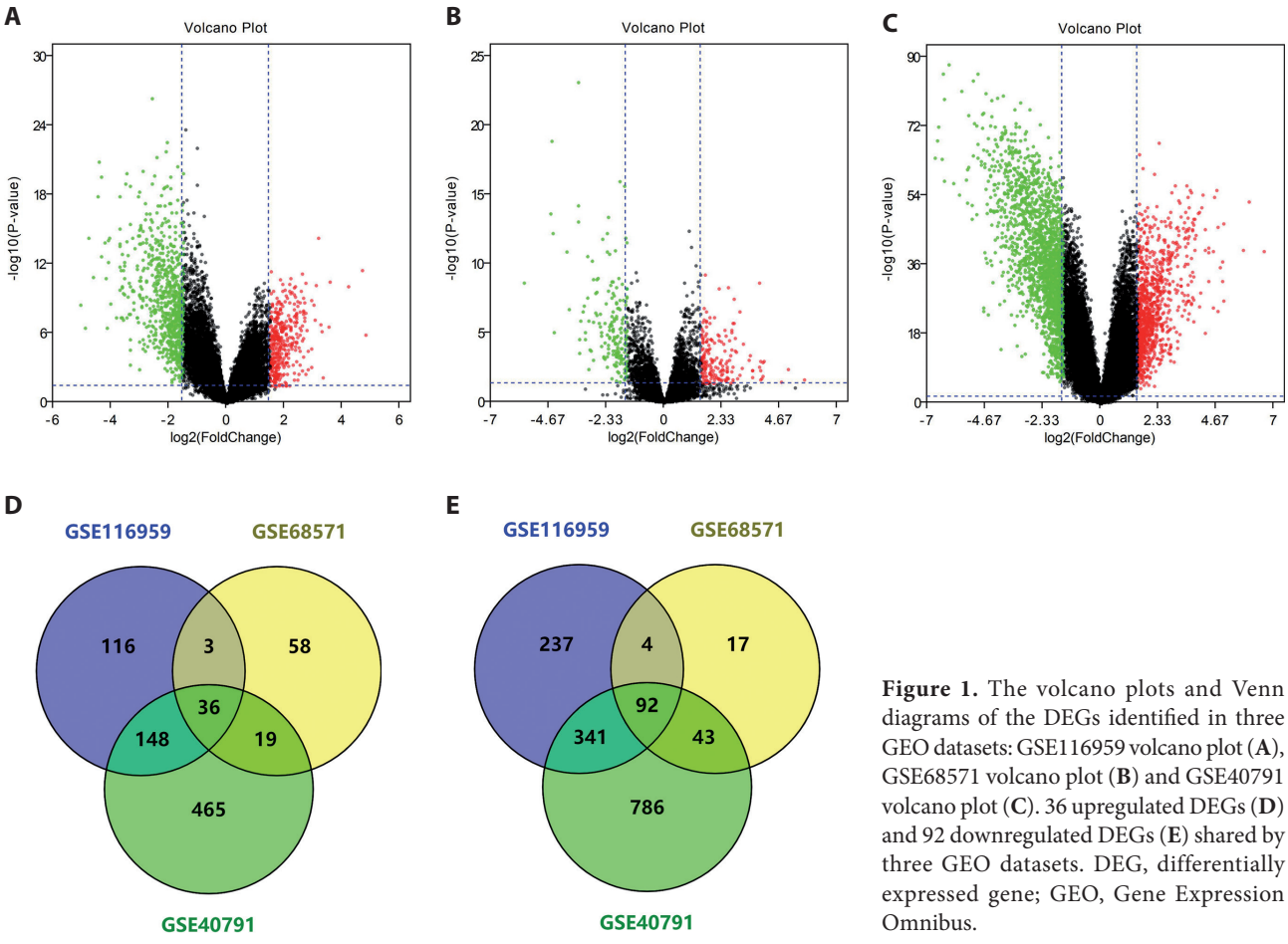


Figure 1. The volcano plots and Venn diagrams of the DEGs identified in three GEO datasets: GSE116959 volcano plot (**A**), GSE68571 volcano plot (**B**) and GSE40791 volcano plot (**C**). 36 upregulated DEGs (**D**) and 92 downregulated DEGs (**E**) shared by three GEO datasets. DEG, differentially expressed gene; GEO, Gene Expression Omnibus.

Table 2. The information of upregulated and downregulated DEGs in GSE116959, GSE68571 and GSE40791

DEGs	GSE116959			GSE68571			GSE40791		
	p	adj. *p*	logFC	*p*	adj. *p*	logFC	*p*	adj. *p*	logFC
Upregulated									
SPP1	8.49E−11	1.16E−08	4.21	2.07E−06	9.01E−05	2.78	5.65E−53	8.77E−51	5.98
FHL2	6.08E−10	6.15E−08	2.05	9.13E−04	1.01E−02	1.53	2.78E−20	3.18E−19	1.86
MDK	8.16E−10	7.82E−08	2.65	5.44E−09	9.01E−07	2.60	1.64E−44	1.25E−42	2.38
LAD1	2.81E−09	2.30E−07	1.85	3.50E−06	1.40E−04	1.71	1.22E−28	2.61E−27	1.71
BIRC5	3.87E−09	3.00E−07	3.04	3.76E−04	5.27E−03	1.59	2.56E−44	1.92E−42	3.87
PAFAH1B3	8.04E−09	5.68E−07	1.85	6.8E−09	1.05E−06	2.23	2.47E−46	2.17E−44	2.07
EZH2	1.69E−08	1.07E−06	1.78	2.26E−03	1.96E−02	1.77	3.16E−57	7.36E−55	3.46
TROAP	2.62E−08	1.55E−06	2.39	1.29E−03	1.29E−02	4.01	3.98E−24	6.13E−23	1.71
KIAA0101	3.37E−08	1.91E−06	2.33	1.03E−03	1.11E−02	1.71	9.03E−48	8.97E−46	3.39
COL10A1	4.29E−08	2.34E−06	2.69	8.52E−04	9.67E−03	2.02	6.08E−41	3.42E−39	4.34
UBE2C	7.03E−08	3.55E−06	3.04	7.86E−06	2.56E−04	1.55	4.11E−44	3.03E−42	3.58
KIF14	1.13E−07	5.30E−06	1.91	6.35E−04	7.84E−03	1.54	1.71E−42	1.09E−40	3.06
MMP1	2.68E−07	1.09E−05	3.54	1.18E−03	1.22E−02	2.55	1.32E−24	2.11E−23	4.73
SLC2A1	3.44E−07	1.35E−05	2.56	6.31E−10	1.75E−07	1.67	1.23E−33	3.85E−32	2.43
CCNB1	3.66E−07	1.42E−05	2.00	2.83E−07	2.05E−05	3.05	1.21E−45	1E−43	3.14
TOP2A	6.44E−07	2.31E−05	1.94	1.46E−05	4.36E−04	1.53	4.93E−56	1.01E−53	4.71
BIK	6.57E−07	2.34E−05	1.88	4.15E−05	9.36E−04	1.69	5.54E−22	7.22E−21	1.95
HMGA1	9.26E−07	3.15E−05	1.76	2.78E−05	7.02E−04	1.91	1.49E−28	3.17E−27	1.62
PFKP	1.51E−06	4.71E−05	1.64	3.80E−03	2.86E−02	1.55	2.47E−30	6.01E−29	1.75
MELK	2.01E−06	5.93E−05	1.78	6.86E−04	8.24E−03	1.89	8.12E−46	6.83E−44	3.47
MMP11	2.11E−06	6.19E−05	2.26	1.57E−04	2.64E−03	1.52	8.2E−22	1.05E−20	2.64
CDK1	2.43E−06	6.93E−05	1.78	2.55E−05	6.69E−04	2.86	3.22E−37	1.31E−35	2.79
IGFBP3	7.46E−06	1.74E−04	1.92	8.01E−07	4.39E−05	1.86	1.74E−17	1.59E−16	1.74
KIF23	1.05E−05	2.32E−04	1.62	1.13E−06	5.93E−05	1.62	1.73E−30	4.26E−29	2.27
MMP12	1.92E−05	3.80E−04	2.50	1.13E−03	1.19E−02	4.05	2.41E−40	1.29E−38	5.75
CP	2.16E−05	4.17E−04	1.90	5.78E−06	2.05E−04	2.73	1.07E−26	1.99E−25	3.65
CDKN3	2.46E−05	4.66E−04	1.75	1.01E−05	3.15E−04	1.69	6.56E−39	3.08E−37	3.49
SIX1	3.38E−04	3.91E−03	1.89	6.09E−04	7.58E−03	1.63	2.93E−23	4.21E−22	3.31
FAP	5.26E−04	5.59E−03	1.55	1.35E−03	1.33E−02	1.64	5.66E−19	5.8E−18	2.07
DUSP4	6.23E−04	6.45E−03	1.58	2.71E−03	2.26E−02	1.62	6.63E−18	6.26E−17	1.74
EEF1A2	8.40E−04	8.24E−03	2.94	2.36E−05	6.37E−04	1.58	2.66E−14	1.83E−13	2.11
COMP	9.98E−04	9.44E−03	1.71	5.08E−04	6.61E−03	1.58	4.19E−13	2.59E−12	2.19
S100P	1.58E−03	1.37E−02	2.82	4.11E−04	5.61E−03	2.69	9.77E−22	1.25E−20	3.74
CEACAM5	1.80E−03	1.51E−02	2.63	1.12E−03	1.18E−02	3.90	1.47E−25	2.52E−24	4.56
PITX2	2.41E−03	1.90E−02	1.58	8.65E−04	9.75E−03	3.41	1.01E−15	7.96E−15	2.17
IGF2BP3	5.91E−03	3.82E−02	1.68	1.28E−03	1.28E−02	1.95	7.76E−29	1.68E−27	3.52
Downregulated									
GPM6A	4.31E−27	2.18E−22	−2.58	1.86E−09	4.04E−07	−2.10	8.1E−82	8.86E−78	−5.62
SLC6A4	3.02E−23	5.1E−19	−2.08	1.67E−05	4.85E−04	−1.90	1.86E−64	1.14E−61	−6.71
CA4	1.35E−20	6.28E−17	−3.45	1.5E−12	8.01E−10	−2.48	2.31E−86	4.2E−82	−6.39
ADH1B	6.05E−18	9.46E−15	−2.84	1.28E−11	5.25E−09	−3.94	2.52E−53	3.99E−51	−3.98
FABP4	4.12E−17	4.73E−14	−2.66	1.27E−19	4.41E−16	−4.53	3.47E−52	5E−50	−5.22
EMP2	4.71E−16	4.04E−13	−2.98	1.07E−11	4.65E−09	−1.96	1.13E−64	7.16E−62	−2.95

(continued on page 36)

Table 2. (continued)

DEGs	GSE116959			GSE68571			GSE40791		
	p	adj. *p*	logFC	*p*	adj. *p*	logFC	*p*	adj. *p*	logFC
FOXF1	4.86E−16	4.1E−13	−2.25	1.2E−09	3.10E−07	−2.67	3.65E−69	4.43E−66	−3.76
CALCRL	7.47E−16	5.64E−13	−2.36	2.65E−12	1.31E−09	−1.51	5.93E−63	3E−60	−3.85
VIPR1	9.28E−16	6.91E−13	−2.98	3.01E−04	4.38E−03	−2.26	1.16E−52	1.75E−50	−4.20
FMO2	9.51E−16	6.98E−13	−3.70	9.44E−14	8.2E−11	−3.47	1.06E−51	1.47E−49	−3.63
SPOCK2	3.41E−15	2.15E−12	−3.03	4.33E−08	5.14E−06	−1.66	3.07E−67	3E−64	−3.58
GPM6B	7.89E−15	4.29E−12	−2.45	5.88E−05	1.23E−03	−1.99	7.62E−58	1.93E−55	−3.42
AGER	9.18E−15	4.84E−12	−4.09	8.2E−24	5.7E−20	−3.46	1.14E−88	6.25E−84	−6.16
CAV1	1.82E−14	8.62E−12	−3.70	2.77E−11	9.61E−09	−3.12	4.99E−62	2.24E−59	−4.06
GDF10	2.58E−14	1.1E−11	−3.69	6.69E−04	8.09E−03	−2.44	6.45E−76	2.35E−72	−4.77
GYPC	3.79E−14	1.5E−11	−1.59	1.69E−06	7.94E−05	−1.53	2.17E−30	5.3E−29	−1.72
MYH10	9.54E−14	3.47E−11	−1.73	2.70E−08	3.46E−06	−1.52	1.8E−55	3.47E−53	−2.08
ADRB2	2.04E−13	6.44E−11	−2.68	2.85E−05	7.08E−04	−1.58	3.01E−57	7.03E−55	−3.45
ITGA8	3.79E−13	1.12E−10	−2.38	1.29E−12	7.49E−10	−1.58	7.33E−50	8.46E−48	−3.31
AOC3	4.24E−13	1.22E−10	−2.91	1.36E−09	3.26E−07	−2.88	5.72E−55	1.05E−52	−3.46
HOXA5	4.83E−13	1.35E−10	−1.69	6.14E−06	2.13E−04	−2.09	1.79E−45	1.47E−43	−3.15
CD36	8.35E−13	2.17E−10	−2.69	1.39E−11	5.37E−09	−2.31	1.41E−59	4.5E−57	−4.46
ADARB1	1.85E−12	4.37E−10	−1.79	1.29E−08	1.79E−06	−2.10	3.83E−53	6.01E−51	−2.76
GRK5	1.9E−12	4.43E−10	−2.40	1.03E−09	2.74E−07	−1.83	2.31E−71	4.21E−68	−3.03
PECAM1	3.41E−12	7.26E−10	−1.95	5.96E−08	6.68E−06	−1.93	3.66E−65	2.47E−62	−2.50
TEK	3.49E−12	7.36E−10	−2.91	1.81E−06	8.11E−05	−2.42	9.5E−72	1.79E−68	−3.87
BCHE	4.23E−12	8.62E−10	−2.07	1.19E−03	1.22E−02	−1.74	1.34E−50	1.69E−48	−4.44
PTPRB	5.9E−12	1.15E−09	−2.01	1.82E−06	8.11E−05	−2.64	8.54E−81	6.67E−77	−3.96
COX7A1	7.07E−12	1.33E−09	−2.14	1.2E−16	2.79E−13	−1.81	8.3E−50	9.56E−48	−2.67
ANGPT1	1.19E−11	2.13E−09	−2.29	6.56E−13	4.15E−10	−2.30	1.01E−46	9.11E−45	−3.74
ASPA	1.22E−11	2.18E−09	−1.77	1.20E−04	2.14E−03	−1.91	3.24E−69	4.03E−66	−3.05
FEZ1	1.37E−11	2.39E−09	−1.91	4.05E−07	2.68E−05	−2.59	2.89E−47	2.74E−45	−3.14
TNNC1	4.13E−11	6.3E−09	−3.39	2.48E−16	4.3E−13	−1.63	1.69E−56	3.65E−54	−4.65
TTN	4.43E−11	6.61E−09	−2.38	3.19E−04	4.57E−03	−2.07	4.99E−19	5.14E−18	−1.57
S1PR1	6.03E−11	8.71E−09	−1.96	6.31E−05	1.31E−03	−1.88	9.77E−67	9.21E−64	−2.96
AQP4	6.13E−11	8.79E−09	−2.42	3.71E−06	1.44E−04	−2.20	5.87E−33	1.74E−31	−3.64
MYL9	8.18E−11	1.12E−08	−1.88	6.45E−07	3.83E−05	−1.58	1.09E−44	8.35E−43	−2.55
GPC3	9.89E−11	1.32E−08	−2.27	5.65E−11	1.87E−08	−2.92	1.71E−52	2.55E−50	−3.87
CFD	1.76E−10	2.12E−08	−2.46	1.25E−09	3.10E−07	−2.74	2.94E−55	5.53E−53	−2.97
MFAP4	2.48E−10	2.81E−08	−3.24	7.93E−07	4.39E−05	−1.67	1.32E−50	1.67E−48	−4.16
MS4A2	2.9E−10	3.22E−08	−2.05	4.66E−05	1.03E−03	−2.39	7.52E−42	4.58E−40	−3.03
PLA2G1B	3.76E−10	4.03E−08	−3.59	2.83E−05	7.07E−04	−3.22	1.03E−37	4.38E−36	−3.59
SPARCL1	5.06E−10	5.23E−08	−2.24	4.24E−14	4.21E−11	−2.25	3.49E−41	2.01E−39	−2.23
FAM189A2	9.97E−10	9.19E−08	−2.89	1.52E−04	2.58E−03	−2.99	1.5E−51	2.07E−49	−3.40
GAS6	1.41E−09	1.25E−07	−1.52	2.06E−04	3.24E−03	−2.05	1.76E−44	1.34E−42	−2.44
SEPP1	1.76E−09	1.53E−07	−1.94	2.84E−09	5.19E−07	−1.72	6.62E−43	4.39E−41	−2.54
CYP4B1	2.66E−09	2.21E−07	−4.03	1.11E−07	1.09E−05	−3.25	1.24E−36	4.81E−35	−4.40
PROS1	3.09E−09	2.52E−07	−1.59	1.27E−04	2.22E−03	−1.72	5.33E−40	2.76E−38	−1.80
VWF	3.47E−09	2.76E−07	−2.11	7.90E−08	8.32E−06	−1.93	1.73E−54	3.06E−52	−2.92
LMO2	3.6E−09	2.84E−07	−2.05	2.50E−08	3.29E−06	−1.53	3.35E−55	6.25E−53	−2.22

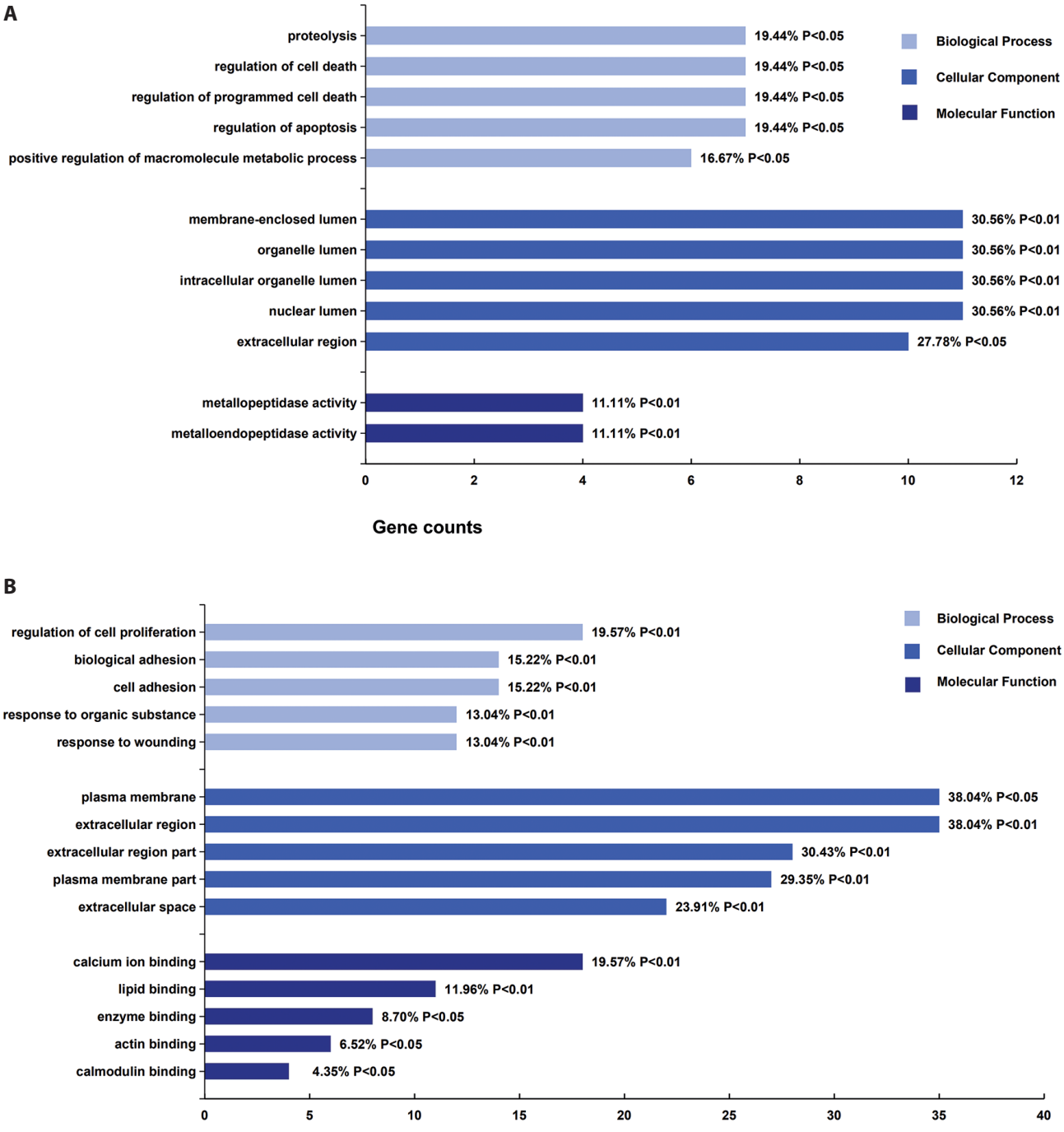
(continued on page 37)

Table 2. (continued)

DEGs	GSE116959			GSE68571			GSE40791		
	p	adj. *p*	logFC	*p*	adj. *p*	logFC	*p*	adj. *p*	logFC
GPX3	4.48E−09	3.39E−07	−2.55	9.79E−08	9.72E−06	−1.58	9.47E−57	2.08E−54	−3.22
FRY	4.74E−09	3.57E−07	−1.77	2.47E−05	6.59E−04	−1.74	6.49E−61	2.41E−58	−2.56
THBD	4.85E−09	3.63E−07	−2.14	2.65E−05	6.82E−04	−2.04	3.21E−54	5.53E−52	−3.01
PMP22	6.23E−09	4.54E−07	−1.65	5.30E−08	6.14E−06	−1.55	3.75E−40	1.98E−38	−1.60
STAC	9.99E−09	6.84E−07	−1.67	2.77E−04	4.12E−03	−1.92	1.34E−26	2.48E−25	−2.09
MRC1	1.22E−08	8.14E−07	−2.55	1.11E−06	5.85E−05	−1.68	2E−24	3.13E−23	−1.69
ALOX5	1.46E−08	9.54E−07	−2.24	1.50E−05	4.46E−04	−1.50	1.64E−26	3.01E−25	−1.64
SRPX	1.62E−08	1.03E−06	−2.05	5.85E−07	3.63E−05	−2.12	7.91E−41	4.41E−39	−2.91
FABP5	2.04E−08	1.26E−06	−2.05	1.56E−07	1.43E−05	−1.85	1.07E−34	3.58E−33	−1.86
EFEMP1	2.13E−08	1.31E−06	−1.88	2.47E−04	3.76E−03	−1.84	2.48E−38	1.11E−36	−1.96
FMO3	6.04E−08	3.14E−06	−1.77	1.57E−04	2.63E−03	−1.89	3.34E−30	8.05E−29	−2.56
CXCL12	8.57E−08	4.22E−06	−1.86	4.61E−06	1.72E−04	−1.75	1.23E−29	2.83E−28	−2.24
PPBP	1.11E−07	5.23E−06	−1.88	7.16E−04	8.50E−03	−1.76	4.78E−31	1.24E−29	−3.74
WASF3	1.11E−07	5.23E−06	−1.50	1.11E−03	1.17E−02	−1.51	4.33E−37	1.73E−35	−2.34
CLDN5	1.68E−07	7.46E−06	−2.22	7.68E−08	8.22E−06	−2.34	2.44E−57	5.81E−55	−3.13
ID4	1.76E−07	7.78E−06	−1.96	3.22E−04	4.60E−03	−1.97	3.46E−45	2.76E−43	−2.06
CPA3	2.15E−07	9.07E−06	−2.04	2.00E−08	2.67E−06	−2.05	3.28E−29	7.29E−28	−2.77
A2M	2.18E−07	9.17E−06	−2.00	1.73E−07	1.50E−05	−1.91	2.89E−45	2.32E−43	−1.61
LPL	2.91E−07	1.18E−05	−2.54	2.42E−06	1.02E−04	−2.15	5.93E−34	1.89E−32	−2.86
KRT4	2.97E−07	1.20E−05	−2.31	1.75E−05	5.02E−04	−2.56	4.34E−29	9.57E−28	−2.60
SCGB1A1	3.34E−07	1.32E−05	−4.92	8.28E−06	2.68E−04	−4.45	8.15E−15	5.87E−14	−4.05
HYAL1	4.06E−07	1.56E−05	−2.00	4.04E−04	5.55E−03	−1.88	4.51E−44	3.31E−42	−3.87
MYLK	6.78E−07	2.40E−05	−1.58	2.19E−06	9.39E−05	−2.17	2.84E−38	1.26E−36	−1.50
CD52	8.83E−07	3.02E−05	−2.13	2.74E−08	3.46E−06	−2.11	5.63E−26	9.88E−25	−1.67
DPT	8.96E−07	3.06E−05	−1.96	1.23E−04	2.18E−03	−1.75	3.86E−28	8.01E−27	−2.55
ABCA3	1.23E−06	3.99E−05	−2.28	6.22E−07	3.75E−05	−1.96	2.07E−29	4.69E−28	−2.57
MGP	1.27E−06	4.10E−05	−2.03	2.12E−05	5.84E−04	−1.58	1.41E−27	2.79E−26	−1.79
RBP4	1.47E−06	4.61E−05	−2.44	5.48E−04	7.02E−03	−1.92	6.62E−29	1.44E−27	−2.83
S100A3	1.91E−06	5.70E−05	−1.78	3.72E−03	2.82E−02	−1.54	4.18E−45	3.3E−43	−3.15
CA2	2.35E−06	6.74E−05	−2.18	2.49E−03	2.11E−02	−1.92	1.77E−38	8.01E−37	−2.84
OASL	4.37E−06	1.13E−04	−1.65	7.86E−03	4.85E−02	−1.57	4.22E−17	3.73E−16	−1.63
C7	5.67E−06	1.39E−04	−1.87	3.02E−07	2.08E−05	−3.27	9.63E−30	2.23E−28	−2.79
SLPI	6.66E−06	1.59E−04	−2.53	1.22E−04	2.16E−03	−2.25	1.48E−19	1.6E−18	−1.99
TSPAN7	8.61E−06	1.96E−04	−1.52	4.77E−04	6.28E−03	−1.95	9.64E−35	3.24E−33	−3.38
GATA6	1.24E−05	2.65E−04	−1.56	1.91E−04	3.08E−03	−1.96	1.15E−40	6.33E−39	−2.26
IGFBP6	1.38E−05	2.89E−04	−1.91	7.19E−05	1.42E−03	−2.27	6.87E−19	6.99E−18	−1.82
MYH11	2.18E−05	4.21E−04	−1.95	3.37E−07	2.30E−05	−2.80	1.9E−42	1.21E−40	−3.59
AQP1	2.82E−05	5.21E−04	−2.08	1.88E−05	5.33E−04	−1.96	3.1E−34	1.01E−32	−2.81
PTGDS	6.70E−05	1.05E−03	−1.88	5.43E−05	1.15E−03	−1.55	1.11E−35	3.97E−34	−2.35
FCER1A	2.02E−04	2.57E−03	−1.60	3.00E−04	4.37E−03	−1.88	6.14E−19	6.27E−18	−2.35
SFTPC	5.93E−04	6.18E−03	−3.34	2.56E−09	4.82E−07	−5.65	1.23E−49	1.41E−47	−5.04
PGC	1.36E−03	1.21E−02	−3.03	4.56E−03	3.29E−02	−3.05	1.58E−13	1.01E−12	−3.08

GEO, Gene Expression Omnibus, DEGs, differentially expressed genes; adj. *p*, adjusted *p* value.

development of tumors mainly by regulating proliferation. This was not counterintuitive; proliferation and apoptosis were two different pathways and mechanisms. When a gene plays a role in proliferation, it does not necessarily have a direct effect on the apoptosis pathway, and *vice versa*. Cell proliferation and apoptosis are the basic life activities of any multicellular life. Various tissue cells in the organism maintain the balance of cell numbers through proliferation and apoptosis. Once this balance is broken, it will cause some diseases, such as cancer. KEGG enrichment analysis of DEGs was performed by KOBAS online analysis tool and following the bubble diagram was drawn by R package ggplot2. As shown in Figure 2C, KEGG pathway of DEGs was mainly concentrated in metabolic pathways, focal adhesion, PI3K-Akt signaling pathway and so on. Top 20 enriched pathways of DEGs were showed in Table 4.



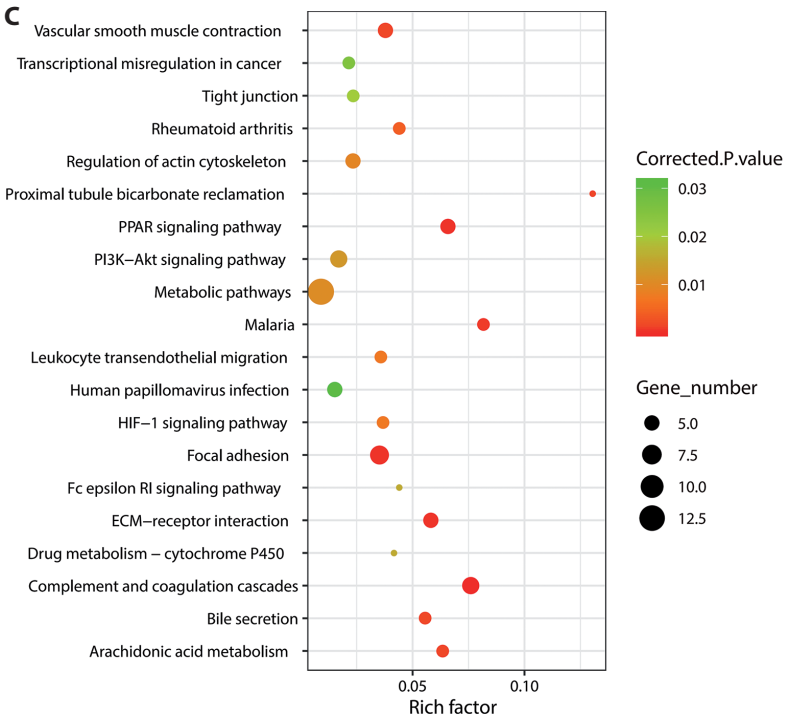


Figure 2. The GO functional annotation and KEGG enrichment analysis of the DEGs. **A.** GO functional annotation of upregulated (**A**) and downregulated (B) DEGs. **C.** KEGG pathway enrichment analysis of DEGs. The percentage indicates the ratio of the number of genes enriched in a GO term to the total number of all annotated DEGs. Rich factor: the ratio of the number of DEGs under a KEGG pathway to the number of all genes annotated to KEGG pathway. The larger the value, the greater the degree of enrichment. The GO term was used in enrichment analysis with GO level 1. GO, Gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEG, differentially expressed gene.

Table 3. Enriched Gene Ontology terms of the upregulated and downregulated DEGs

Category	Term	Count	*p*
Upregulated DEGs			
BP	regulation of apoptosis	7	1.04E–02
	regulation of programmed cell death	7	1.09E–02
	regulation of cell death	7	1.11E–02
	proteolysis	7	3.47E–02
	skeletal system development	6	8.43E–04
CC	nuclear lumen	11	1.61E–03
	intracellular organelle lumen	11	7.29E–03
	organelle lumen	11	8.56E–03
	membrane-enclosed lumen	11	9.82E–03
	nucleoplasm	10	1.72E–04
MF	metalloendopeptidase activity	4	1.73E–03
	metallopeptidase activity	4	8.46E–03
Downregulated DEGs			
BP	regulation of cell proliferation	18	6.41E–06
	cell adhesion	14	4.01E–04
	biological adhesion	14	4.06E–04
	response to wounding	12	4.64E–04
	response to organic substance	12	5.33E–03
CC	extracellular region	35	4.11E–08
	plasma membrane	35	2.07E–02
	extracellular region part	28	3.64E–11
	plasma membrane part	27	1.43E–03
	extracellular space	22	2.03E–09
MF	calcium ion binding	18	1.73E–05
	lipid binding	11	2.79E–04
	enzyme binding	8	3.31E–02
	actin binding	6	4.21E–02
	calmodulin binding	4	4.88E–02

DEGs, differentially expressed genes; BP, biological process; CC, cellular component; MF, molecular function.

Table 4. Top 20 enriched pathways of DEGs

ID	Pathway	Gene No.	Corr. *p*	Gene
hsa01100	Metabolic pathways	13	1.09E−02	*PLA2G1B, AOC3, CA4, CA2, GPX3, HYAL1, COX7A1, ASPA, PFKP, PTGDS, ADH1B, ALOX5, PAFAH1B3*
hsa04510	Focal adhesion	7	3.83E−04	*ITGA8, CAV1, COMP, VWF, MYL9, MYLK, SPP1*
hsa04610	Complement and coagulation cascades	6	6.13E−05	*VWF, PROS1, THBD, C7, A2M, CFD*
hsa04151	PI3K-Akt signaling pathway	6	1.30E−02	*ITGA8, COMP, VWF, ANGPT1, SPP1, TEK*
hsa03320	PPAR signaling pathway	5	3.83E−04	*MMP1, FABP4, FABP5, LPL, CD36*
hsa04512	ECM-receptor interaction	5	5.07E−04	*ITGA8, VWF, COMP, SPP1, CD36*
hsa04270	Vascular smooth muscle contraction	5	1.81E−03	*MYL9, MYH11, PLA2G1B, MYLK, CALCRL*
hsa04810	Regulation of actin cytoskeleton	5	9.63E−03	*MYH10, ITGA8, CXCL12, MYLK, MYL9*
hsa05165	Human papillomavirus infection	5	3.13E−02	*ITGA8, COMP, SPP1, VWF, OASL*
hsa05144	Malaria	4	8.87E−04	*COMP, PECAM1, CD36, GYPC*
hsa00590	Arachidonic acid metabolism	4	1.81E−03	*PLA2G1B, PTGDS, ALOX5, GPX3*
hsa04976	Bile secretion	4	2.04E−03	*AQP1, SLC2A1, CA2, AQP4*
hsa05323	Rheumatoid arthritis	4	4.33E−03	*MMP1, CXCL12, ANGPT1, TEK*
hsa04066	HIF-1 signaling pathway	4	7.61E−03	*PFKP, SLC2A1, ANGPT1, TEK*
hsa04670	Leukocyte transendothelial migration	4	7.70E−03	*MYL9, CXCL12, CLDN5, PECAM1*
hsa04530	Tight junction	4	2.06E−02	*MYH10, MYH11, MYL9, CLDN5*
hsa05202	Transcriptional misregulation in cancer	4	2.56E−02	*SIX1, TSPAN7, LMO2, IGFBP3*
hsa04964	Proximal tubule bicarbonate reclamation	3	1.81E−03	*CA4, CA2, AQP1*
hsa04664	Fc epsilon RI signaling pathway	3	1.60E−02	*ALOX5, MS4A2, FCER1A*
hsa00982	Drug metabolism - cytochrome P450	3	1.60E−02	*FMO3, FMO2, ADH1B*

DEGs, differentially expressed genes; corr. *p*, corrected *p* value.

PPI network construction and hub genes identification

In order to reveal the protein relationships of 128 DEGs (36 upregulated and 92 downregulated), we constructed a PPI network by STRING online database and Cytoscape 3.6.1 software, as shown in Figure 3A. It contained 128 nodes which interconnected by 308 edges. Based on MNC and Degree algorithms in cytoHubba plugin of Cytoscape, we selected the top 10 DEGs of MNC and Degree algorithms respectively, and then obtained a total of seven hub genes by intersecting them, which were *VWF, SPP1, PECAM1, TOP2A, CDK1, UBE2C*, and *KIF23*, as shown in Figure 3B,C and Table 5. Hub genes are defined as genes with high correlation in candidate modules. In order to further verify the level expression of the selected hub genes in LUAD, we performed gene expression analysis, TNM and survival analysis on the 7 hub genes.

Validation of hub genes

First, we verified the Gene expression analysis, TNM and survival analysis of seven hub genes in GSE116959, GSE68571 and GSE40791, respectively. As shown in Figure S1–S4, the seven hub genes were indeed differentially expressed in the three datasets and were related to tumor stage. Since the original data of GSE116959 and GSE40791 do not give survival time and survival outcome, there is no survival analysis on these two datasets. From Figure S5, we found that in GSE68571, the high or low expression of the seven hub genes and the survival outcome were not statistically significant. This might be due to the relatively small sample size, so we used TCGA data to verify seven hub genes. Then, we used GEPIA to perform gene expression analysis, TNM and survival analysis of seven hub genes selected above. GEPIA is a web server for analyzing the RNA sequencing

Table 5. Top10 hub gene in MNC and Degree

MNC			Degree			Common gene
Rank	Name	Score	Rank	Name	Score	
1	*VWF*	23	1	*VWF*	23	*VWF*
2	*SPP1*	19	1	*SPP1*	23	*SPP1*
3	*PECAM1*	14	3	*PECAM1*	18	*PECAM1*
4	*TOP2A*	12	4	*CDK1*	13	*TOP2A*
5	*CDK1*	11	4	*KIAA0101*	13	*CDK1*
5	*UBE2C*	11	4	*EZH2*	13	*UBE2C*
5	*CCNB1*	11	7	*CAV1*	12	*KIF23*
5	*BIRC5*	11	7	*UBE2C*	12	
5	*KIF23*	11	7	*TOP2A*	12	
5	*CDKN3*	11	7	*KIF23*	12	

MNC, Maximum Neighborhood Component.

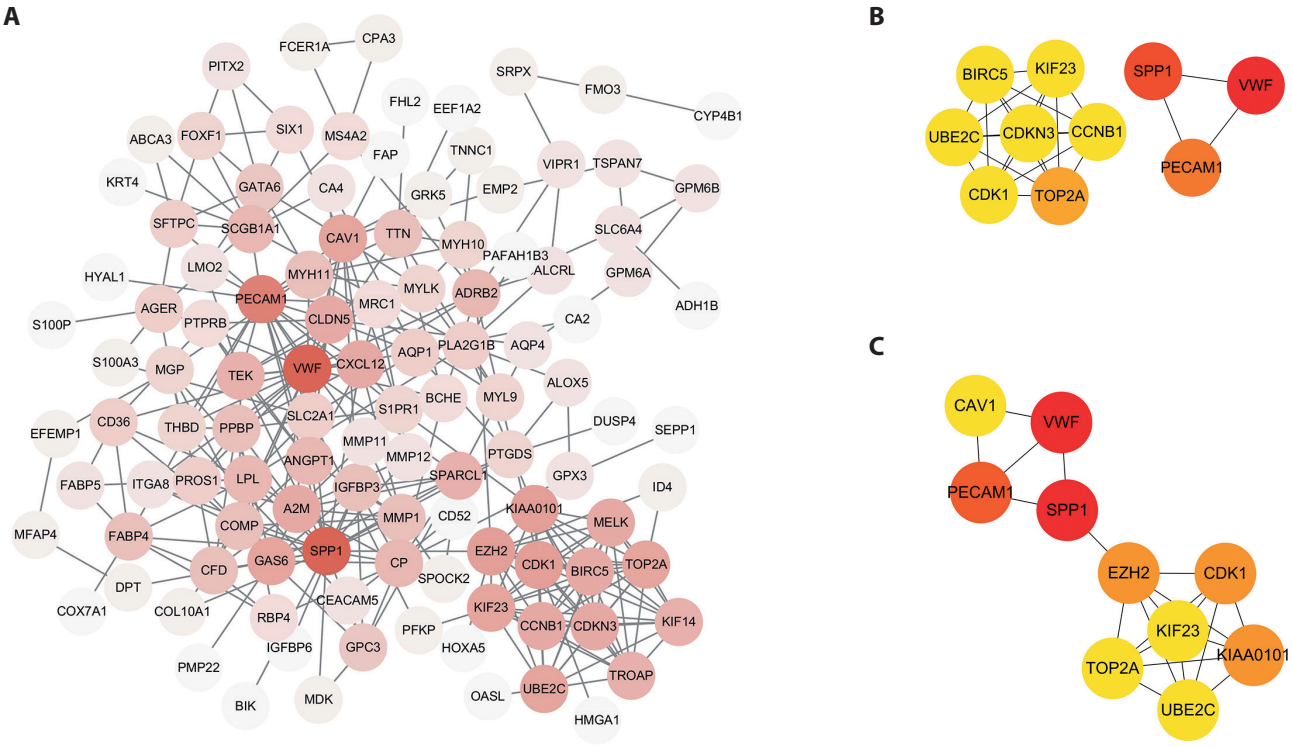


Figure 3. The protein-protein interaction (PPI) network and hub genes. **A**. The PPI network constructed for 128 DEGs. **B.** Top 10 genes with MNC. **C.** Top 10 genes with Degree. A coloring scheme was used to display the ranking score of each node. When the node is redder, the ranking of it will be higher. Solid lines indicate the connected nodes interact to each other (direct connections). DEG, differentially expressed gene; MNC, Maximum Neighborhood Component. (See online version for color figure.)

expression data of 9,736 tumors and 8,587 normal samples from the TCGA and the GTEx projects, using a standard processing pipeline. For LUAD, GEPIA uses 483 TCGA tumor samples with paired adjacent 59 TCGA normal samples and 288 GTEx normal samples. As shown in Figure 4, the relative expression levels of *VWF* and *PECAM1* in LUAD were significantly decreased, while the others were significantly increased, compared with normal tissues. These results were consistent with the expression trend of these seven DEGs in the three datasets. As shown in Figure 5, the LUAD patients with clinic Stage II, Stage III or Stage IV had a higher expression level of *TOP2A*, *CDK1*, *UBE2C* and *KIF23* than Stage I, which indicated that these upregulated hub genes might be linked to tumor progression positively. As shown in Figure 6, the high expression of *SPP1*, *TOP2A*, *CDK1*, *UBE2C*, *KIF23* as well as the low expression of *PECAM1* were associated with the worse OS in LUAD patients, which revealed that *SPP1*, *TOP2A*, *CDK1*, *UBE2C*, *KIF23* were associated with LUAD tumor progression and might be used as tumor progression predictors for LUAD patients. *TOP2A*, *CDK1*, *UBE2C* and *KIF23* were differentially expressed in normal tissues and tumor tissues, which were associated with stage and survival analysis of LUAD, so we detected the relative expression levels of *TOP2A*, *CDK1*, *UBE2C* and *KIF23* in LUAD cell lines by qRT-PCR. The primer sequences of each gene were shown in Table 6. As shown in Figure 7, compared

Table 6. The primer sequences

Gene	The forward primer	The reversed primer
TOP2A	AAGATTCATTGAAGACGCTTCG	GCTGTAAAATGCCATTTCTTGC
CDK1	CACAAAACTACAGGTCAAGTGG	GAGAAATTTCCCGAATTGCAGT
UBE2C	CAACCTTTTCAAATGGGTAGGG	CAGGATGTCCAGGCATATGTTA
KIF23	AGACAGAAGGCGAGGGATG	GGAGACGAATTGGTGGTGC
β-actin	CCTGGCACCCAGCACAAT	GGGCCGGACTCGTCATAC

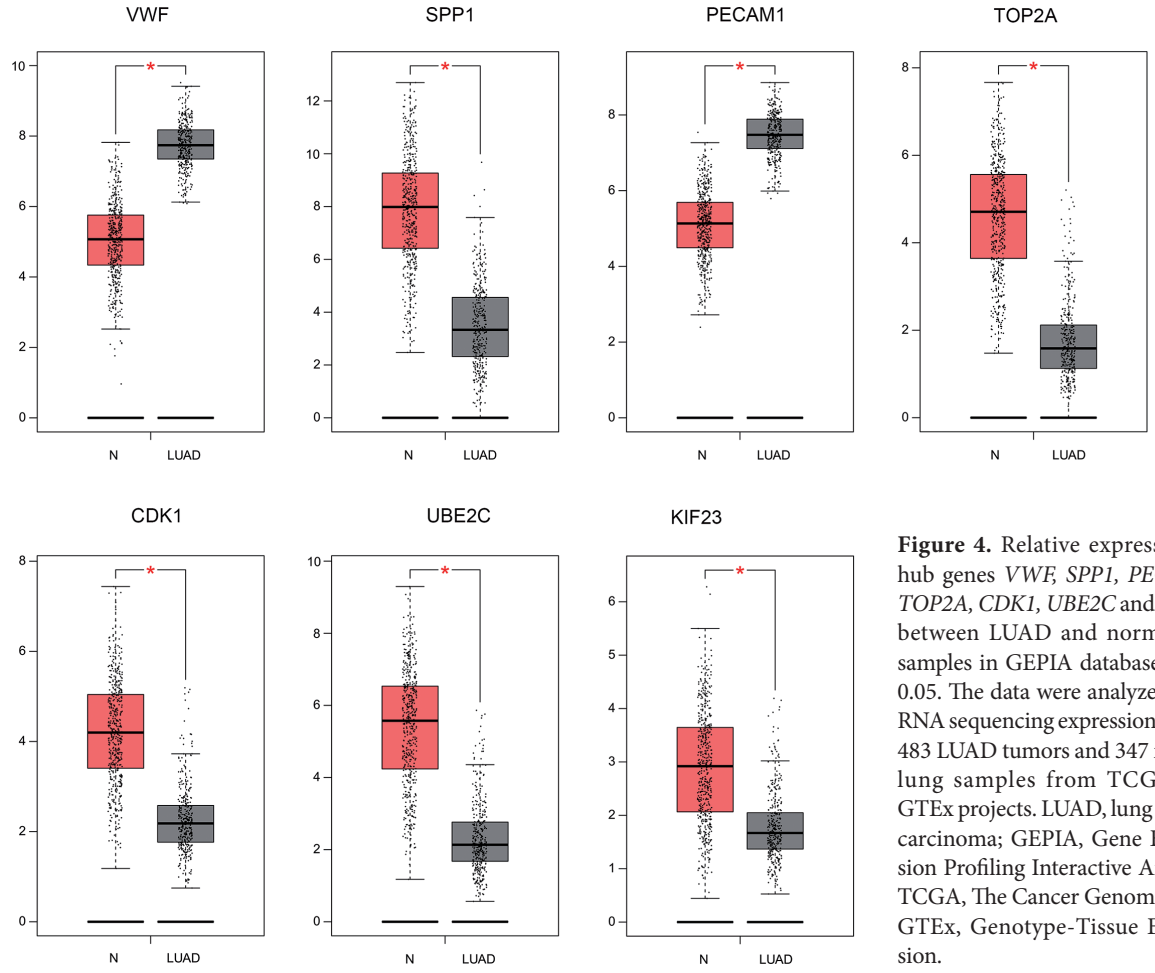


Figure 4. Relative expression of hub genes *VWF, SPP1, PECAM1, TOP2A, CDK1, UBE2C* and *KIF23*, between LUAD and normal (N) samples in GEPIA database. * $p < 0.05$. The data were analyzed from RNA sequencing expression data of 483 LUAD tumors and 347 normal lung samples from TCGA and GTEx projects. LUAD, lung adenocarcinoma; GEPIA, Gene Expression Profiling Interactive Analysis; TCGA, The Cancer Genome Atlas; GTEx, Genotype-Tissue Expression.

with normal lung cell, *TOP2A*, *CDK1* and *UBE2C* were significantly up-regulated in A549 cell and H1650 cell, but *KIF23* was only significantly up-regulated in A549 cell. The results of qRT-PCR showed that the four genes were indeed differentially expressed in LUAD compared with normal cell. Immunohistochemical analysis of the 4 genes, *TOP2A*, *CDK1*, *UBE2C* and *KIF23* through The Human Protein Atlas database revealed that these genes were all positive in lung cancer tissues. The antibody HPA026773 was used to detect *TOP2A* at low intensity in normal lung tissues, while in LUAD tissues, it showed medium intensity staining, with the proportion of stained cells ranging from 25 to 75%. The antibody CAB003387 did not detect *CDK1* in normal lung tissues, and showed moderate intensity staining in LUAD tissues, with the proportion of stained cells >75%. The antibody CAB03599 was used to detect *UBE2C* at low intensity in normal lung tissue, and it showed medium intensity staining in LUAD tissue, with the proportion of stained cells >75%. The antibody CAB01041 was used to detect *KIF23* with high intensity in normal lung tissues, and it was stained with high intensity in LUAD tissues, and the proportion of stained cells was >75%. The results are shown in Figure 8.

Discussion

In this study, we selected three GEO microarray datasets about LUAD, GSE116959, GSE68571 and GSE40791, which were related to the stage and survival analysis of LUAD. Then, we used GEO2R to analyze the DEGs in the three datasets and got the intersection of DEGs by Venn diagrams. The analysis found that although the number of DEGs selected by the three datasets was different, the DEGs of one dataset were not covered by another dataset and the results of the Venn diagram can confirm this. In the end, 128 DEGs were selected, including 36 upregulated genes and 92 downregulated genes. Then we performed GO functional annotation and KEGG enrichment analysis of DEGs. GO functional annotation showed that the upregulated genes were mainly related to the regulation of apoptosis process

and the downregulated genes were mainly related to the regulation of cell proliferation process. KEGG enrichment analysis revealed that 128 DEGs were mainly enriched in metabolic pathways. Apoptosis is a unique morphological and biochemical form of programmed cell death as well as dysregulation of apoptosis is associated with the pathogenesis of many human diseases, especially cancer. And cancer can cause uncontrolled cell proliferation. The development of tumor depends on the reprogramming of cell metabolism, which plays an increasingly important role in cancer (Hanahan et al. 2011). Cancer cells use a variety of metabolic pathways to proliferate constantly, such as glucose, amino acids, serine/glycine and lipid metabolism, etc. (Pavlova et al. 2016). In short, apoptosis, cell proliferation, and metabolic pathways are closely related to cancer. Next, by constructing a PPI network and using MNC and Degree algorithms of Cytoscape to process it, seven hub genes were finally selected, which were *VWF*, *SPP1*, *PECAM1*, *TOP2A*, *CDK1*, *UBE2C* and *KIF23*. So, we verified the Gene expression analysis, TNM and survival analysis of seven hub genes

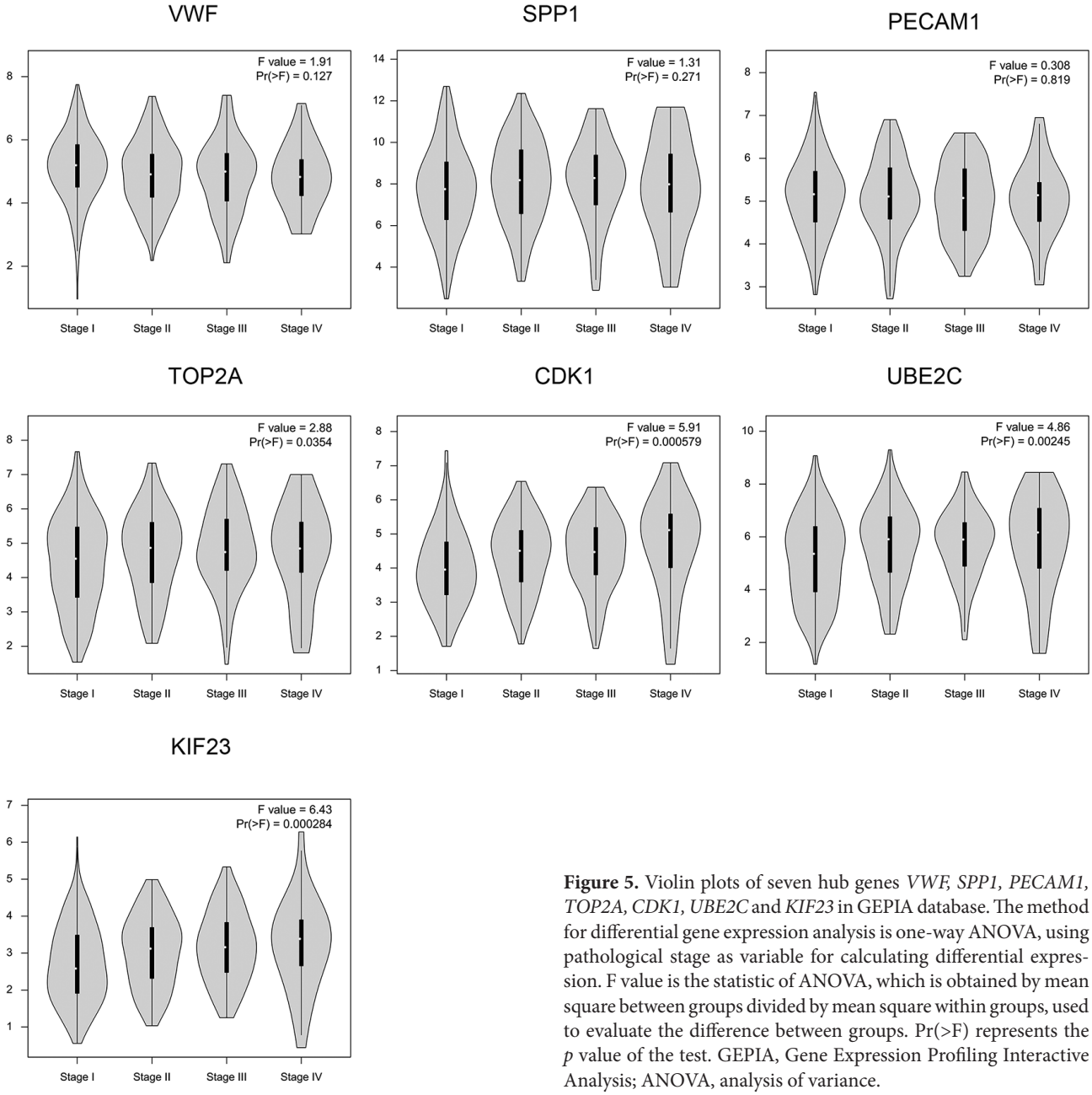


Figure 5. Violin plots of seven hub genes *VWF, SPP1, PECAM1, TOP2A, CDK1, UBE2C* and *KIF23* in GEPIA database. The method for differential gene expression analysis is one-way ANOVA, using pathological stage as variable for calculating differential expression. F value is the statistic of ANOVA, which is obtained by mean square between groups divided by mean square within groups, used to evaluate the difference between groups. Pr(>F) represents the *p* value of the test. GEPIA, Gene Expression Profiling Interactive Analysis; ANOVA, analysis of variance.

in GSE116959, GSE68571 and GSE40791, respectively. The seven hub genes were differentially expressed in the three datasets and were related to tumor stage. However, they were not related to survival analysis which may be caused by too small sample size. And then, we used TCGA data to verify seven hub genes. GEPIA analysis revealed that the relative expression levels of *VWF* and *PECAM1* in LUAD patients were significantly decreased, and the relative expression levels of *SPP1*, *TOP2A*, *CDK1*, *UBE2C*, and *KIF23* were significantly increased compared with normal controls. *TOP2A*, *CDK1*, *UBE2C* and *KIF23* were related to the stage of LUAD patients, indicating that the above four genes might be participated in the occurrence and development of LUAD. *SPP1*, *TOP2A*, *CDK1*, *UBE2C*, *KIF23* and *PECAM1* were associated with the survival analysis in LUAD patients. However, the number of patients in each stage was not marked in the stage analysis chart generated by GEPIA in Figure 5, which is not scientific enough. Although the figures

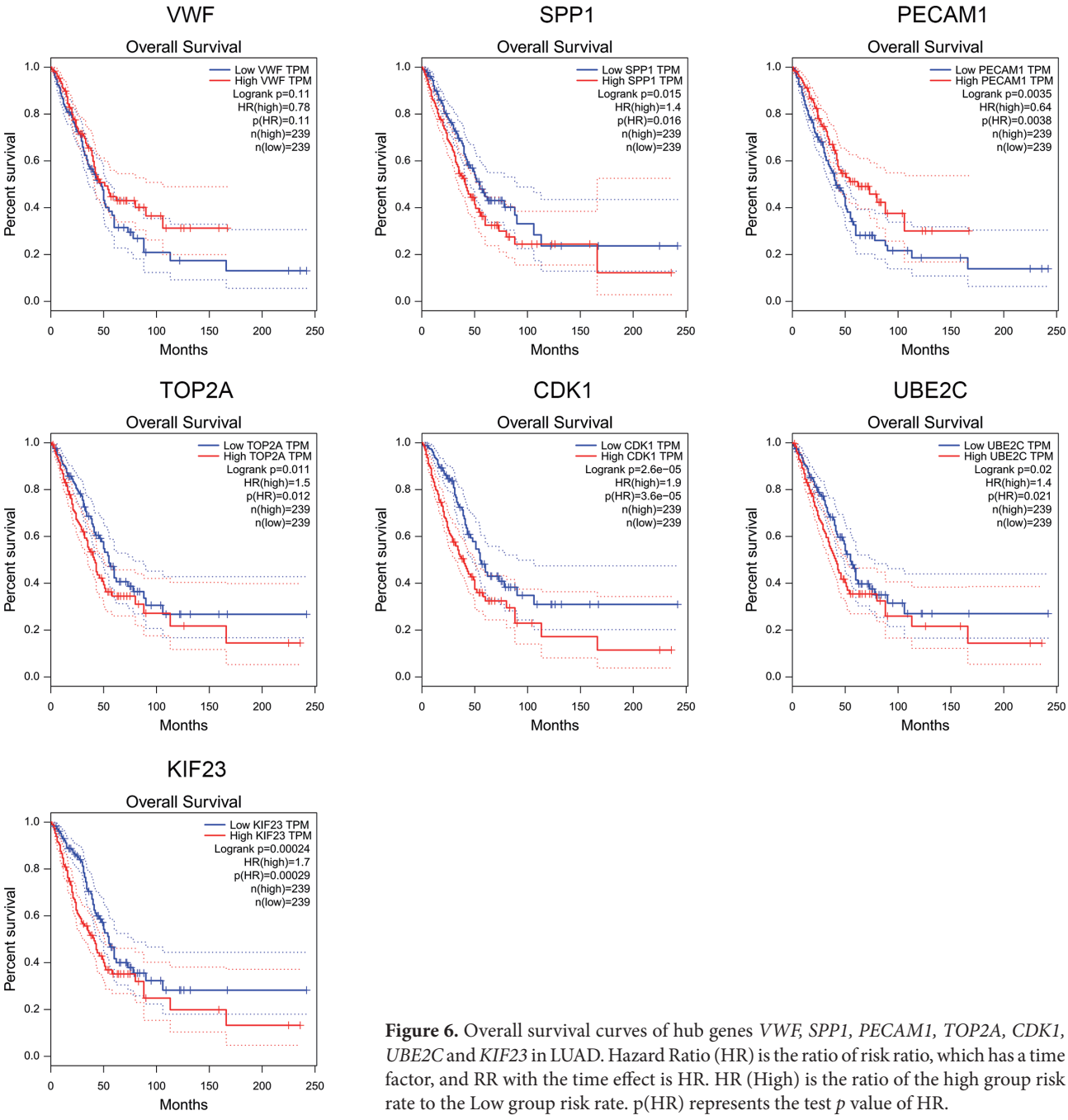


Figure 6. Overall survival curves of hub genes *VWF, SPP1, PECAM1, TOP2A, CDK1, UBE2C* and *KIF23* in LUAD. Hazard Ratio (HR) is the ratio of risk ratio, which has a time factor, and RR with the time effect is HR. HR (High) is the ratio of the high group risk rate to the Low group risk rate. p(HR) represents the test *p* value of HR.

obtained by the GEPIA online analysis tool do not show the detailed information of the data, GEPIA is supported by a large amount of raw data from TCGA and GTEx, which is convenient to operate and has a large number of users. There were also many papers published based on GEPIA. Therefore, the data has high credibility. Then we verified the relative expression levels of *TOP2A*, *CDK1*, *UBE2C* and *KIF23* in LUAD cell lines by qRT-PCR. The results showed that *TOP2A*, *CDK1*, *UBE2C* and *KIF23* were significantly upregulation in A549 cell and *TOP2A*, *CDK1*, as well as *UBE2C* were significantly upregulated in H1650 cell. The results of immunohistochemistry showed that *TOP2A* and *UBE2C* were lowly expressed in normal lung tissues, but moderately expressed in LUAD tissues; *CDK1* was not de-

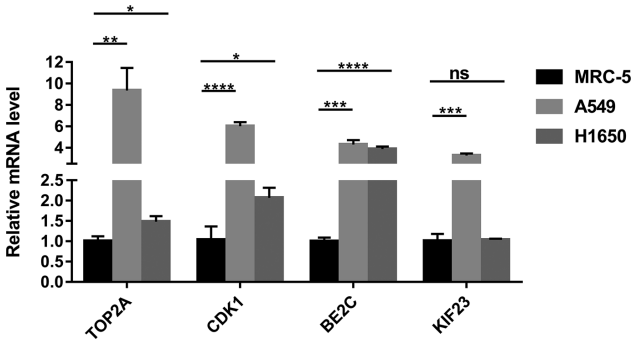


Figure 7. The relative mRNA level of *TOP2A, CDK1, UBE2C* and *KIF23* in LUAD cell lines (MRC-5, A549 and H1650). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

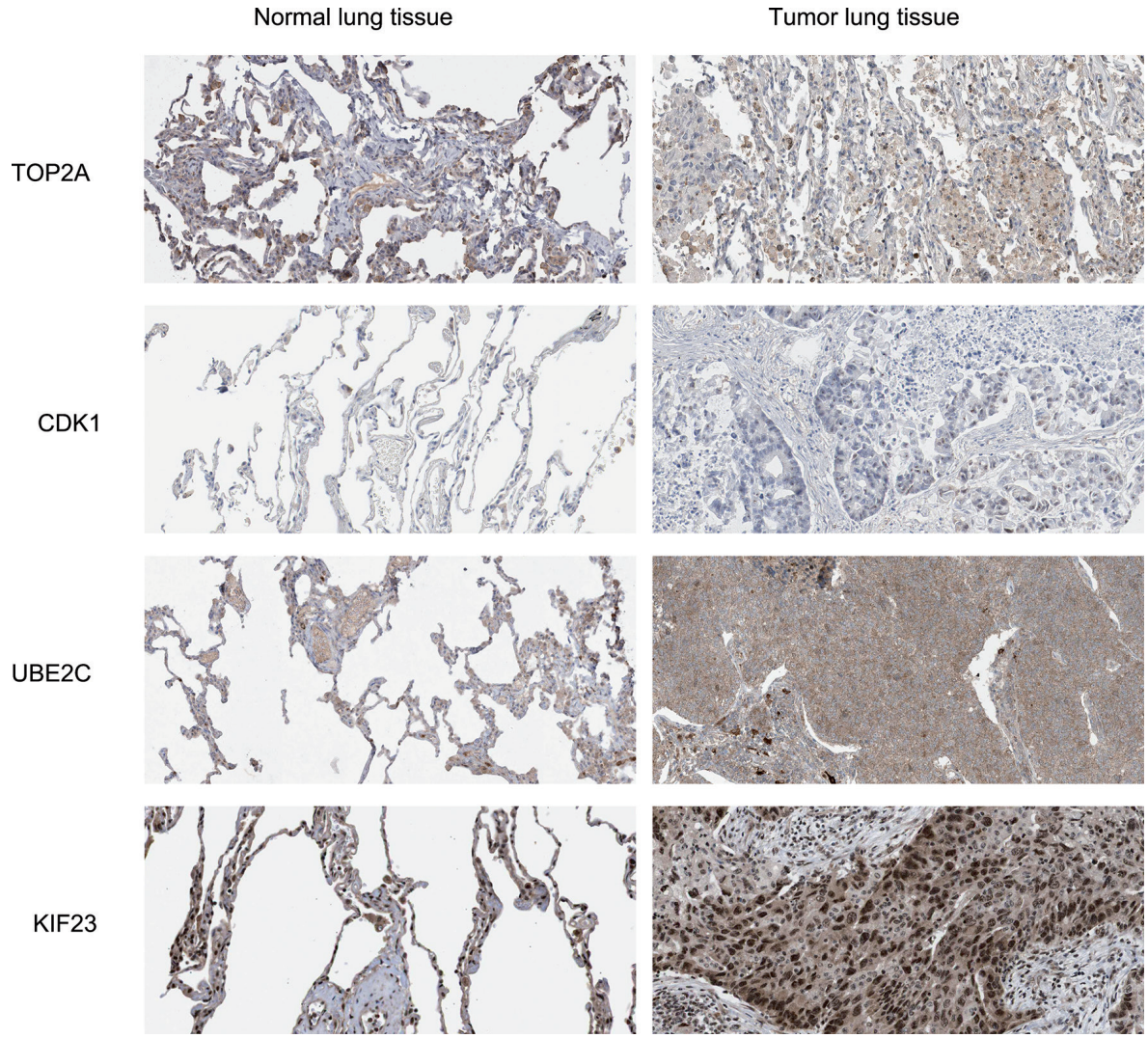


Figure 8. The representative immunohistochemical results of *TOP2A, CDK1, UBE2C* and *KIF23* in LUAD tissues. Judge the degree of positivity based on the intensity of staining and the percentage of stained cells.

tected in normal lung tissues, but was medium expression in LUAD tissues; *KIF23* was highly expressed in normal lung tissues and LUAD tissues. Through bioinformatics analysis, it was found that the four genes were related to the prognosis and stage of LUAD. So, they can be knocked down or overexpressed to verify their function in the occurrence and development of LUAD in future studies.

VWF (von Willebrand factor) encodes a hemostasis-associated glycoprotein whose mutation causes von Willebrand disease, an inherited bleeding disorder. We found that *VWF* was lowly expressed in LUAD. Moreover, another study based on bioinformatics analysis indicated that high expression of *VWF* predict longer survival in NSCLC (Piao et al. 2018). However, a recent study reported that *VWF* was high expression in LUAD tissues (Xu et al. 2017), which was inconsistent with our results. Therefore, further verification is needed.

The protein encoded by *SPP1* (secreted phosphoprotein 1) can increase expression of interferon-gamma and interleukin-12 and is related to the attachment of osteoclasts to the mineralized bone matrix. The overexpression of *SPP1* inhibited autophagy and apoptosis and promoted proliferation of SCLC cells (Liu et al. 2019). We found that *SPP1* was overexpression in LUAD and was related to poor OS. Moreover, a recent research also demonstrated that the higher expression of *SPP1* in LUAD tissues predicted unfavorable prognosis of stage T1, N0 and N1 patients and worse survival (Li et al. 2018).

The protein encoded by *PECAM1* (platelet and endothelial cell adhesion molecule 1) belongs to the immunoglobulin superfamily and may be related to leukocyte migration, angiogenesis and integrin activation. Higher expression of PECAM-1 was found in stromal tissues of primary NSCLC lesions, metastatic lymph nodes and brain lesions; In contrast, lower expression was detected in the non-cancerous lung stromal tissues (Kuang et al. 2013). The present study found that *PECAM1* was lowly expressed in LUAD and was associated with poor OS. Moreover, another study also found that *PECAM1* was lowly expressed in LUAD by bioinformatic analysis (Yu et al. 2020). In short, the role of *PECAM1* in the development of LUAD needs further study.

The DNA topoisomerase encoded by *TOP2A* can alter topological states of DNA and be the marker of anticancers. It was reported that *TOP2A* was also related to cell proliferation (Pommier et al. 2010). *TOP2A* has potential prognostic biomarkers in breast cancer (El Rebey et al. 2016), NSCLC (Guo et al. 2020), prostate cancer (de Resende et al. 2013), and so on. Our study found that *TOP2A* was highly expressed in LUAD, which was related to TNM stage and poor OS, so *TOP2A* might be a prognostic biomarker for LUAD.

The protein encoded by *CDK1* (cyclin dependent kinase 1) belongs to the Ser/Thr protein kinase family and can control G1-S and G2-M phase transitions of eukaryotic cell cycle. *CDK1* promotes G2/M progression through phosphorylation of mitochondrial substrates(Wang et al. 2014)(Wang et al. 2014). Inhibition of *CDK1* resulted in a G2 phase cell cycle arrest and decreased proliferation of TNBC cells (Reese et al. 2017). Our study demonstrated that *CDK1* was highly expressed in LUAD and was related to TNM stage as well as poor OS. Consistent with our results, another study also indicated *CDK1* was upregulated in LUAD and that its up-regulation was associated with unfavorable overall survival (Shi et al. 2016).

The protein encoded by *UBE2C* (ubiquitin conjugating enzyme E2 C) belongs to the E2 ubiquitin-conjugating enzyme family, which is necessary for cell cycle progression and for destroying mitotic cyclins and may be related to cancer progression. Downregulation of *UBE2C* suppressed the ovarian cancer cells proliferation and induced G2/M arrest, cell apoptosis as well as cisplatin resistance reversal *via* downregulating *CDK1* (Li et al. 2020). Our study found that *UBE2C* was highly expressed in LUAD and was related to TNM stage as well as poor OS, and another report also showed that high expression of *UBE2C* may be indicative of poor survival rates in LUAD patients (Guo et al. 2020).

The protein encoded by *KIF23* (kinesin family member 23) belongs to the kinesin-like protein family and the protein is involved in driving microtubule movement *in vitro*. In lung cancer cells, the depletion of *KIF23* led to the emergence of giant multinucleated cells followed by the death of apoptotic cells (Iltzsche et al. 2017). The current study showed that *KIF23* was high expression in LUAD and was associated with TNM stage and poor prognosis, similarly another study indicated that *KIF23*, by RT-PCR, was upregulated in lung cancer tissues and knockdown of *KIF23* could decrease the NSCLC cells proliferation (Kato et al. 2016).

In this study, we carefully selected three LUAD GEO microarray datasets, of which GSE116959 was the first time to be used for bioinformatics analysis. In addition, we used two methods, MNC and Degree, to identify the hub genes of the PPI network, not just degree. We also used qRT-PCR to verify the differential expression of *TOP2A*, *CDK1, UBE2C and KIF23*, not just the method of biological information. Interestingly, we found that *AFF3* was up-regulated in GSE68571, but down-regulated in GSE116959 and GSE40791. Then a search on GEPIA database showed that *AFF3* was down-regulated in lung adenocarcinoma samples, and finally, we found *AFF3* was low expression in lung adenocarcinoma samples by searching the literature (Zhang et al. 2018). *AFF3* (AF4/FMR2 family member 3, or LAF4) was first considered a lymphoid-specific gene; it is expressed and locate in the nuclear of B cells.

In summary, by bioinformatics analysis including GO functional annotation and KEGG enrichment analysis, PPI network as well as hub genes identification and validation, this study found that *TOP2A*, *CDK1*, *UBE2C* and

KIF23 were all differentially expressed in LUAD and were associated with stage and survival analysis of LUAD, which were expected to be prognostic factor of LUAD. Obviously, much deeper studies are certainly needed to confirm their clinical values.

Acknowledgments. We sincerely thank the researchers for providing their GEO microarray datasets information online, it is our pleasure to acknowledge their contributions. We also sincerely thank the supporting of DAVID, STRING, GEPIA public databases, and Cytoscape, R packege software. This work was supported by Henan Joint Program, National Natural Science Foundation of China (No. U1804194).

Conflict of interest. The authors declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Author contributions. YXM designed the study. YXM, JJN and YJH contributed to the literature search. YXM performed this study and wrote the initial draft of the manuscript. CXL, FQC, ZZS and ZH reviewed and edited the manuscript. All authors read and approved the manuscript.

References

Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani IJB (2001): Controlling the false discovery rate in behavior genetics research. Behav. Brain Res. **125,** 279-284 https://doi.org/10.1016/S0166-4328(01)00297-2

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018): Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **68,** 394-424 https://doi.org/10.3322/caac.21492

Clough E, Barrett T (2016): The Gene Expression Omnibus Database. Methods Mol. Biol. **1418,** 93-110 https://doi.org/10.1007/978-1-4939-3578-9_5

de Resende MF, Vieira S, Chinen LT, Chiappelli F, da Fonseca FP, Guimarães GC, Soares FA, Neves I, Pagotty S, Pellionisz PA, et al. (2013): Prognostication of prostate cancer based on TOP2A protein and gene assessment: TOP2A in prostate cancer. J. Transl. Med. **11,** 36 https://doi.org/10.1186/1479-5876-11-36

Denisenko TV, Budkevich IN, Zhivotovsky B (2018): Cell death-based treatment of lung adenocarcinoma. Cell Death Dis. **9,** 117 https://doi.org/10.1038/s41419-017-0063-y

El Rebey HS, Aiad HA, Abulkheir IL, Asaad NY, El-Wahed MM, Abulkasem FM, Mahmoud SF (2016): The predictive and prognostic role of topoisomerase IIα and tissue inhibitor of metalloproteinases 1 expression in locally advanced breast carcinoma of Egyptian patients treated with anthracycline-based neoadjuvant chemotherapy. Appl. Immunohistochem. Mol. Morphol. **24,** 167-178 https://doi.org/10.1097/PAI.0000000000000154

Guo T, Ma H, Zhou Y (2019): Bioinformatics analysis of microarray data to identify the candidate biomarkers of lung adenocarcinoma. PeerJ **7,** e7313 https://doi.org/10.7717/peerj.7313

Guo W, Sun S, Guo L, Song P, Xue X, Zhang H, Zhang G, Wang Z, Qiu B, Tan F, et al. (2020): Elevated TOP2A and UBE2C expressions correlate with poor prognosis in patients with surgically resected lung adenocarcinoma: a study based on immunohistochemical analysis and bioinformatics. J. Cancer Res. Clin. Oncol. **146,** 821-841 https://doi.org/10.1007/s00432-020-03147-4

Hanahan D, Weinberg Robert A (2011): Hallmarks of cancer: The next generation. Cell **144,** 646-674 https://doi.org/10.1016/j.cell.2011.02.013

Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, et al. (2007): DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. **35,** W169-175 https://doi.org/10.1093/nar/gkm415

Iltzsche F, Simon K, Stopp S, Pattschull G, Francke S, Wolter P, Hauser S, Murphy DJ, Garcia P, Rosenwald A, et al. (2017): An important role for Myb-MuvB and its target gene KIF23 in a mouse model of lung adenocarcinoma. Oncogene **36,** 110-121 https://doi.org/10.1038/onc.2016.181

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima KJN (2017): KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. **45,** D353-361 https://doi.org/10.1093/nar/gkw1092

Kato T, Wada H, Patel P, Hu HP, Lee D, Ujiie H, Hirohashi K, Nakajima T, Sato M, Kaji M, et al. (2016): Overexpression of KIF23 predicts clinical outcome in primary lung cancer patients. Lung Cancer **92,** 53-61 https://doi.org/10.1016/j.lungcan.2015.11.018

Kuang BH, Wen XZ, Ding Y, Peng RQ, Cai PQ, Zhang MQ, Jiang F, Zhang XS, Zhang X (2013): The prognostic value of platelet endothelial cell adhesion molecule-1 in non-small-cell lung cancer patients. Med. Oncol. **30,** 536 https://doi.org/10.1007/s12032-013-0536-5

Li J, Zhi X, Shen X, Chen C, Yuan L, Dong X, Zhu C, Yao L, Chen M (2020): Depletion of UBE2C reduces ovarian cancer malignancy and reverses cisplatin resistance via downregulating CDK1. Biochem. Biophys. Res. Commun. **523,** 434-440 https://doi.org/10.1016/j.bbrc.2019.12.058

Li S, Yang R, Sun X, Miao S, Lu T, Wang Y, Wo Y, Jiao W (2018): Identification of SPP1 as a promising biomarker to predict clinical outcome of lung adenocarcinoma individuals. Gene **679,** 398-404 https://doi.org/10.1016/j.gene.2018.09.030

Liu H, Wei S, Zhang L, Yuan C, Duan Y, Wang Q (2019): Secreted phosphoprotein 1 promotes the development of small cell lung cancer cells by inhibiting autophagy and apoptosis. Pathol. Oncol. Res. **25,** 1487-1495 https://doi.org/10.1007/s12253-018-0504-7

Pavlova NN, Thompson CB (2016): The emerging hallmarks of cancer metabolism. Cell Metab. **23,** 27-47 https://doi.org/10.1016/j.cmet.2015.12.006

Piao J, Sun J, Yang Y, Jin T, Chen L, Lin Z (2018): Target gene screening and evaluation of prognostic values in non-small cell lung cancers by bioinformatics analysis. Gene **647,** 306-311 https://doi.org/10.1016/j.gene.2018.01.003

Pommier Y, Leo E, Zhang H, Marchand C (2010): DNA topoisomerases and their poisoning by anticancer and antibacterial drugs. Chem. Biol. **17,** 421-433 https://doi.org/10.1016/j.chembiol.2010.04.012

Reese JM, Bruinsma ES, Monroe DG, Negron V, Suman VJ, Ingle JN, Goetz MP, Hawse JR (2017): ERβ inhibits cyclin dependent kinases 1 and 7 in triple negative breast cancer. Oncotarget **8,** 96506-96521 https://doi.org/10.18632/oncotarget.21787

Shi YX, Zhu T, Zou T, Zhuo W, Chen YX, Huang MS, Zheng W, Wang CJ, Li X, Mao XY, et al. (2016): Prognostic and predictive values of CDK1 and MAD2L1 in lung adenocarcinoma. Oncotarget **7,** 85235-85243 https://doi.org/10.18632/oncotarget.13252

Siegel RL, Miller KD, Jemal A (2017): Cancer Statistics, 2017. CA Cancer J. Clin. **67,** 7-30 https://doi.org/10.3322/caac.21387

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. (2019): STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. **47,** D607-613 https://doi.org/10.1093/nar/gky1131

Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z (2017): GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. **45,** W98-102 https://doi.org/10.1093/nar/gkx247

Wang Z, Fan M, Candas D, Zhang TQ, Qin L, Eldridge A, Wachsmann-Hogiu S, Ahmed KM, Chromy BA, Nantajit D, et al. (2014): Cyclin B1/Cdk1 coordinates mitochondrial respiration for cell-cycle G2/M progression. Dev. Cell **29,** 217-232 https://doi.org/10.1016/j.devcel.2014.03.012

Wood DE (2015): National Comprehensive Cancer Network (NCCN) clinical practice guidelines for lung cancer screening. Thorac. Surg. Clin. **25,** 185-197 https://doi.org/10.1016/j.thorsurg.2014.12.003

Xu Y, Pan S, Liu J, Dong F, Cheng Z, Zhang J, Qi R, Zang Q, Zhang C, Wang X, et al. (2017): GATA3-induced vWF upregulation in the lung adenocarcinoma vasculature. Oncotarget **8,** 110517-110529 https://doi.org/10.18632/oncotarget.22806

Yu DH, Huang JY, Liu XP, Ruan XL, Chen C, Hu WD, Li S (2020): Effects of hub genes on the clinicopathological and prognostic features of lung adenocarcinoma. Oncol. Lett. **19,** 1203-1214 https://doi.org/10.3892/ol.2019.11193

Zhang D, Qu L, Ma L, Zhou Y, Wang G, Zhao X, Zhang C, Zhang Y, Wang M, Zhang M, et al. (2018a): Genome-wide identification of transcription factors that are critical to non-small cell lung Cancer Lett. **434,** 132-143 https://doi.org/10.1016/j.canlet.2018.07.020

Zhang MY, Liu X.X, Li H, Li R, Liu X, Qu YQ (2018b): Elevated mRNA levels of AURKA, CDC20 and TPX2 are associated with poor prognosis of smoking related lung adenocarcinoma using bioinformatics analysis. Int. J. Med. Sci. **15,** 1676-1685 https://doi.org/10.7150/ijms.28728

Zhou LN, Li SC, Li XY, Ge H, Li HM (2018): Identification of differential protein-coding gene expressions in early phase lung adenocarcinoma. Thorac. Cancer **9,** 234-240 https://doi.org/10.1111/1759-7714.12569

Received: June 13, 2020
Final version accepted: September 19, 2020

doi: 10.4149/gpb_2020037

Supplementary Material

Identification of differentially expressed genes associated with lung adenocarcinoma *via* bioinformatics analysis

Xinmeng Yang[1], Qingchuan Feng[1], Jianan Jing[1], Jiahui Yan[1], Zhaoshu Zeng[2], Hao Zheng[1] and Xiaoli Cheng[1]

[1] *Department of Medical Genetics and Cell Biology, School of Basic Medical Sciences, Zhengzhou University, Zhengzhou, Henan, China*
[2] *Department of Forensic Medicine, School of Basic Medical Sciences, Zhengzhou University, Zhengzhou, Henan, China*

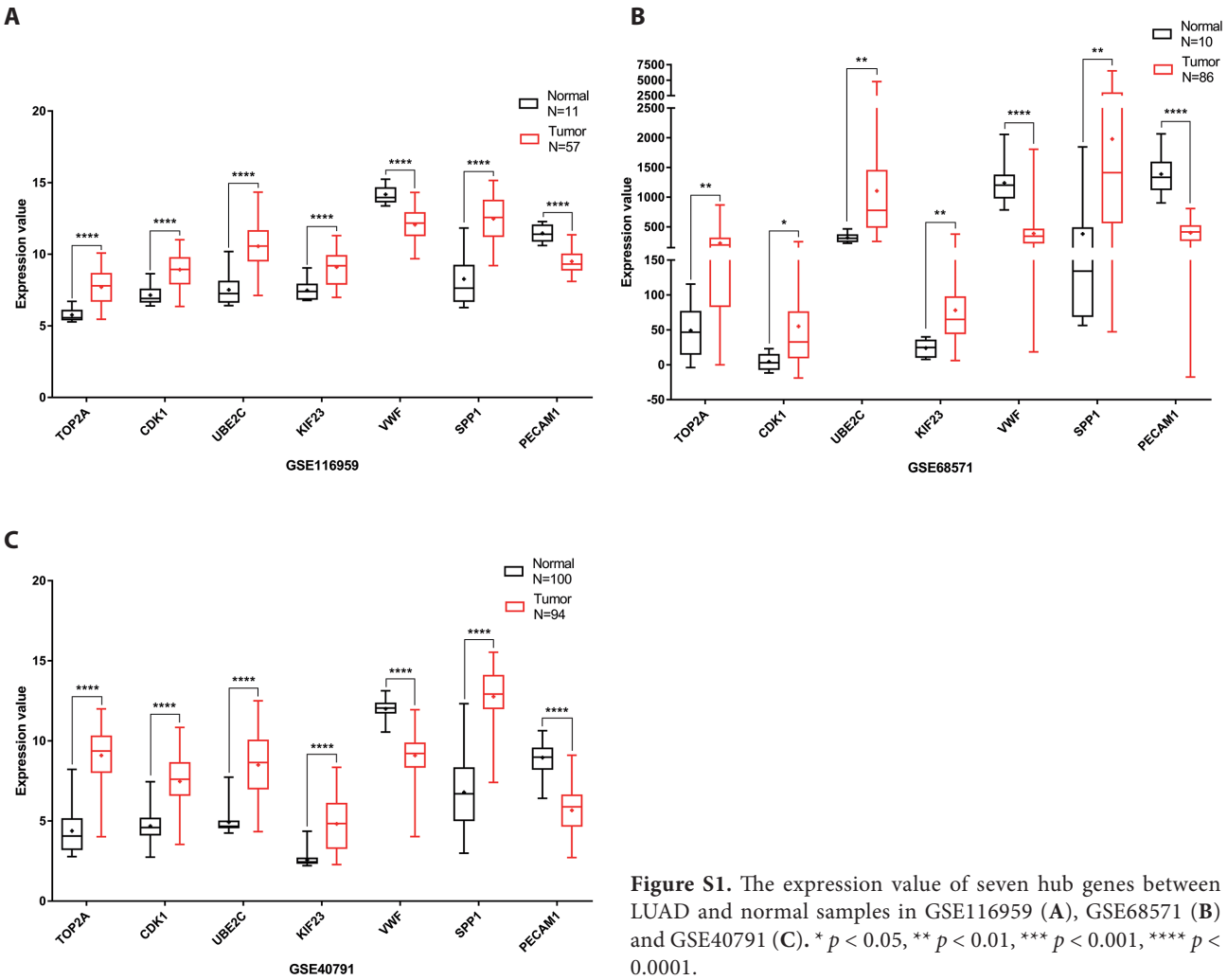


Figure S1. The expression value of seven hub genes between LUAD and normal samples in GSE116959 (**A**), GSE68571 (**B**) and GSE40791 (**C**). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

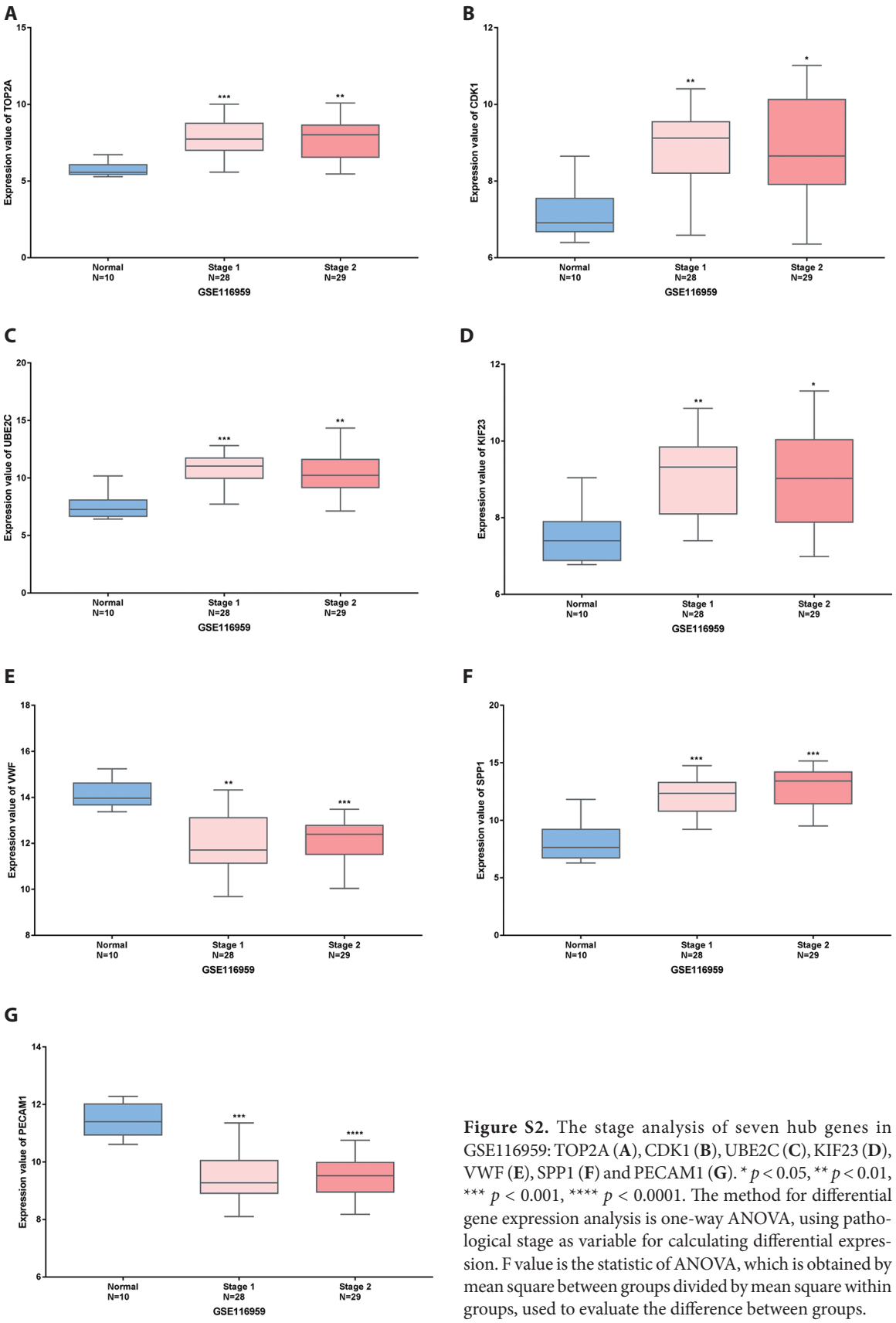


Figure S2. The stage analysis of seven hub genes in GSE116959: TOP2A (**A**), CDK1 (**B**), UBE2C (**C**), KIF23 (**D**), VWF (**E**), SPP1 (**F**) and PECAM1 (**G**). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$. The method for differential gene expression analysis is one-way ANOVA, using pathological stage as variable for calculating differential expression. F value is the statistic of ANOVA, which is obtained by mean square between groups divided by mean square within groups, used to evaluate the difference between groups.

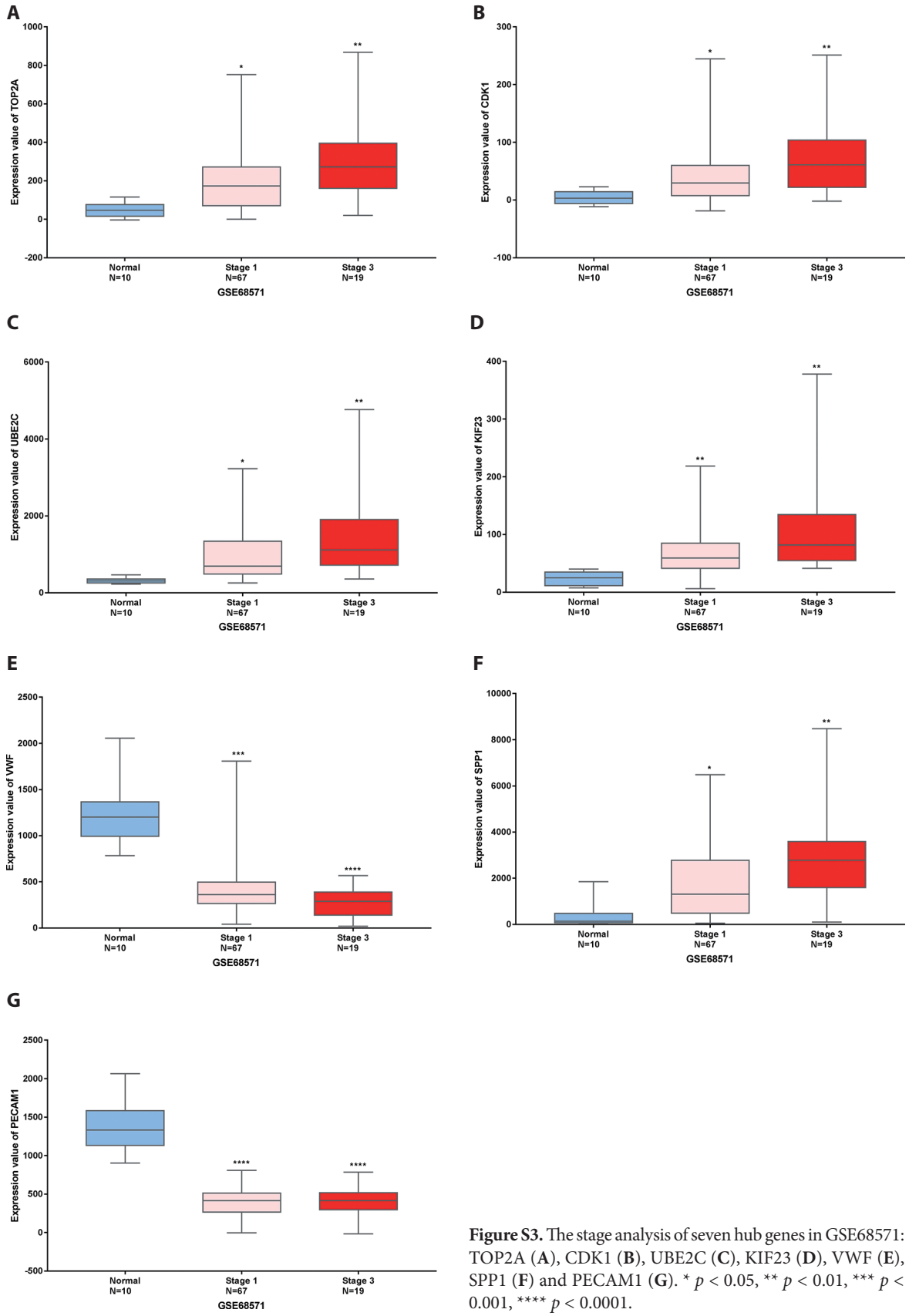


Figure S3. The stage analysis of seven hub genes in GSE68571: TOP2A (**A**), CDK1 (**B**), UBE2C (**C**), KIF23 (**D**), VWF (**E**), SPP1 (**F**) and PECAM1 (**G**). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

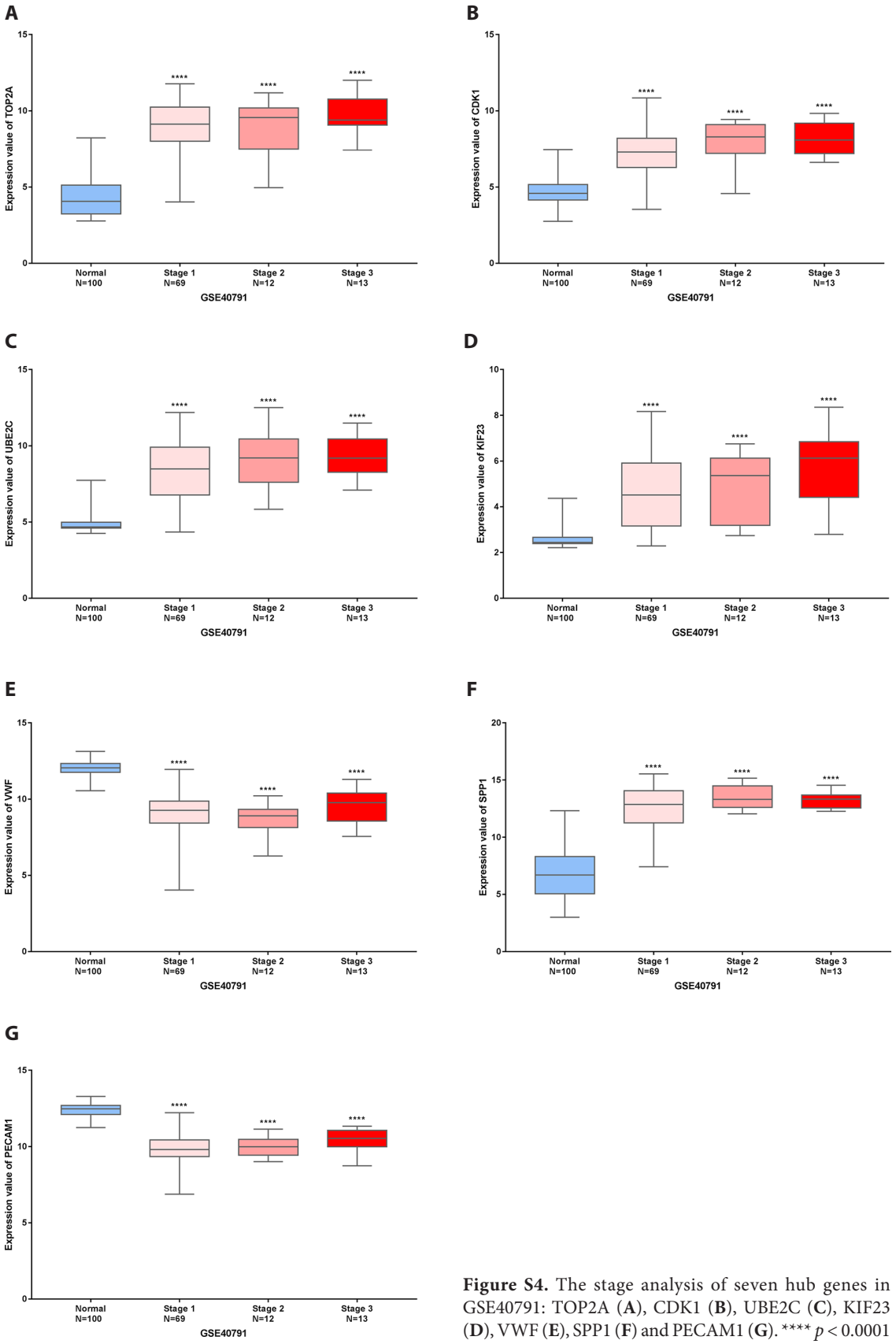


Figure S4. The stage analysis of seven hub genes in GSE40791: TOP2A (**A**), CDK1 (**B**), UBE2C (**C**), KIF23 (**D**), VWF (**E**), SPP1 (**F**) and PECAM1 (**G**). **** $p < 0.0001$

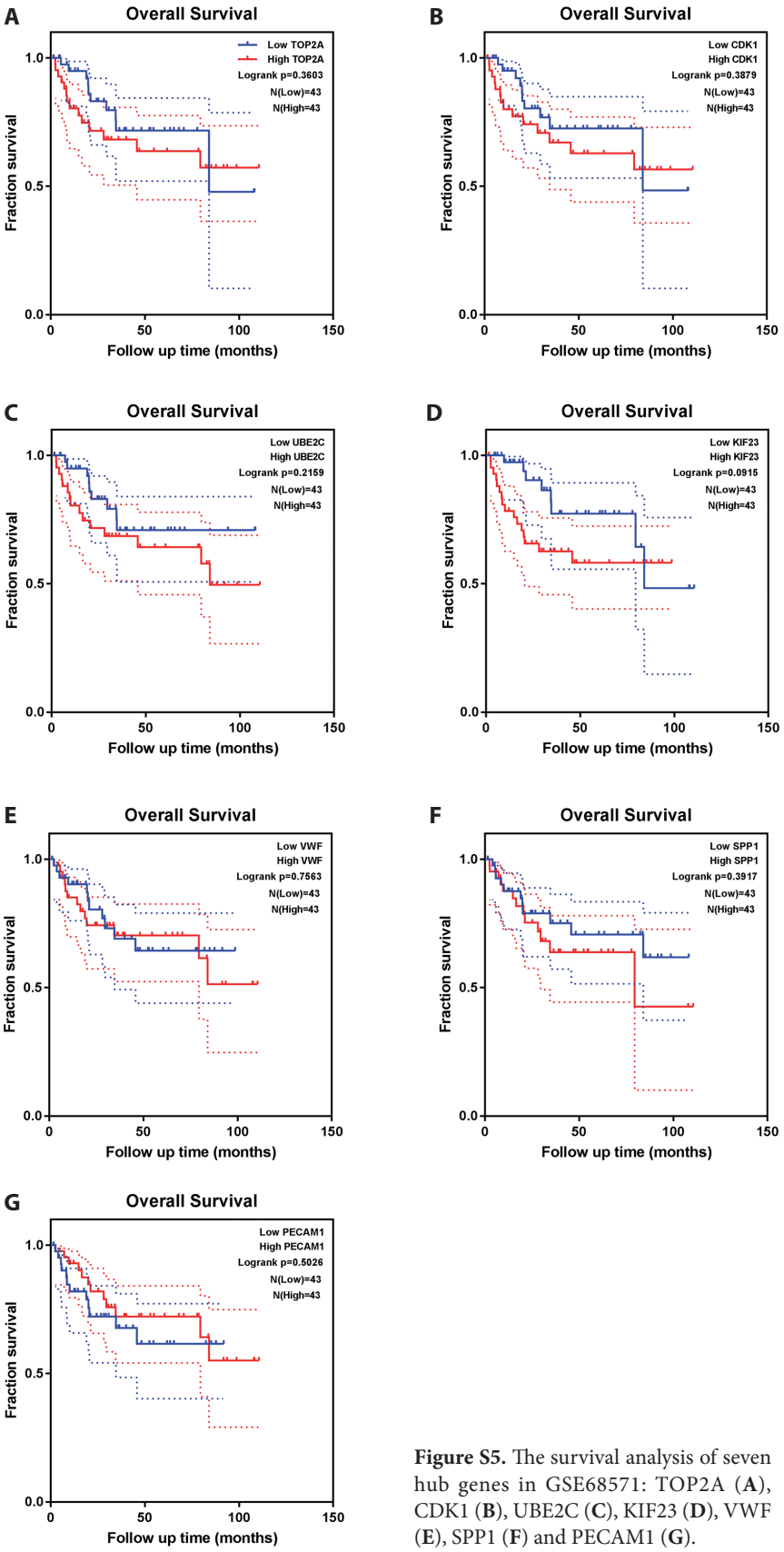


Figure S5. The survival analysis of seven hub genes in GSE68571: TOP2A (**A**), CDK1 (**B**), UBE2C (**C**), KIF23 (**D**), VWF (**E**), SPP1 (**F**) and PECAM1 (**G**).