

Received: 2019.04.13

Accepted: 2019.05.06

Published: 2019.06.10

Differentially Expressed Gene Screening, Biological Function Enrichment, and Correlation with Prognosis in Non-Small Cell Lung Cancer

Authors' Contribution:

Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

ABF **He Huang**
DG **Qingdong Huang**
E **Tingyu Tang**
BG **Xiaoxi Zhou**
B **Liang Gu**
AC **Xiaoling Lu**
BDG **Fang Liu**

Department of Respiratory, Zhejiang Hospital, Hangzhou, Zhejiang, P.R. China

Corresponding Author: Fang Liu, e-mai: liufangzju@126.com
Source of support: Departmental sources

Background: The aim of this study was to explore the differently expressed genes and pathways in non-small cell lung cancer (NSCLC) and their correlation with the prognosis.


Material/Methods: Gene expression data series of GSE19804, GSE101929, and GSE33532 were downloaded from the Gene Expression Omnibus (GEO) database. The overlapping differently expressed genes (DEGs) were identified from the above 3 data series. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) were used to analyze the biological functions and signal pathways of DEGs. The protein-protein interaction (PPI) was analyzed thorough Search Tool for the Retrieval of Interacting Gens (STRING). The relationship between the expression of hub genes and the prognosis of patients was analyzed by Kaplan-Meier Plotter online software.

Results: Twenty-nine DEGs were identified, with 22 upregulated genes and 7 downregulated genes. The enriched biological processes were mainly related to diet-induced thermogenesis and actin filament binding. The KEGG pathways were enriched in calcium signaling, regulation of lipolysis in adipocytes, and PPAR signaling. Two downregulated genes (*MMP1* and *SPP1*) were identified as hub genes by Cytohubba. Twenty-two dysregulated genes were correlated with patient prognosis.

Conclusions: Differentially expressed genes are common in NSCLC patients and can be used as biomarkers for patient prognosis.

MeSH Keywords: Lung Neoplasms • Microarray Analysis • Prognosis

Full-text PDF: <https://www.medscimonit.com/abstract/index/idArt/916962>

 1355

 3

 5

 26



Background

Lung cancer, including non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), is the leading cause of malignant tumor-related mortality [1]. Epidemiological studies show that more than 1 million new cases of lung cancer and more than 800 000 deaths occur every year [2,3]. The lung cancer epidemiology data from China demonstrate that the overall incidence of lung cancer in China is high, especially in Tianjin city in Dagang province and Xuanwei city in Yunnan province. The incidence of lung cancer in the above 2 areas is significantly higher than the overall global level [4,5]. It is reported that 75–80% of lung cancer is NSCLC, whose biological behavior and treatment methods are different from those of small cell lung cancer. At present, the molecular mechanism of the occurrence, development, invasion, and metastasis of NSCLC is still unclear.

In recent years, with the development of gene expression profiling chip and second-generation high-throughput sequencing technology, the amount of data on lung cancer expression profiles has greatly expanded, which provides the basis for the comprehensive study of differentially expressed genes and their biological functions in lung cancer [6]. In this study, 3 gene expression profiles of lung cancer were selected from the GEO (<https://www.ncbi.nlm.nih.gov/geo/>) [7] database, and we explored the function of DEGs in the development of lung cancer and its relationship with patient prognosis.

Material and Methods

Microarray data screening

Three gene expression data series – GSE19804 [8], GSE101929 [9], and GSE33532 [10] – relevant to lung cancer from the GEO database were identified and included for the present analysis. The original microarray data of the 3 data series were download. For GSE19804, 120 lung cancer specimens with 60 cancer tissues and paired 60 normal lung tissues were recognized with the platform of GPL570[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. A total of 41 non-small cell lung cancer cases were included in the data series of GSE101929 and the gene expression was detected by GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. For GSE33532, individual primary tumors and matched distant normal lung tissues (N) from 20 patients were used to establish gene expression patterns captured by GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array.

Data processing

The microarray data of the included 3 data series were first analyzed using R 3.4.4 statistical software (<https://www.r-project.org>),

then the identified dysregulated genes were further analyzed to find the overlapped genes of the 3 data series.

Biological function enrichment and pathway analysis

The biological function enrichment and pathways analysis were performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID, <http://david.ncifcr.gov>) [11]. This analysis included 2 aspects: one is gene ontology (GO) [12, 13] and the other is Kyoto Encyclopedia of Genes and Genomes (KEEG) [14]. The GO enrichment includes biological process (BP), cellular component (CC), and molecular function (MF).

Protein–protein network analysis and hub gene identification

The protein–protein network was built by the Search Tool for Retrieval of Interacting Genes (STRING) database with the criteria of: minimum required interaction score of 0.4 and active interaction sources of text mining, experiments, databases, co-expression, neighborhood, gene fusion, and co-occurrence. The target hub gene was selected with the criteria of top 10 genes according to 5 Cytohubba ranking method using Cytoscape software (<https://cytoscape.org/>) [15].

Survival analysis

The survival analysis of patients relevant to gene expression was expressed by the database of Kaplan-Meier Plotter (<http://kmpplot.com/analysis/index.php?p=background>) [16] through survival curves. According to the median expression of each gene in cancer tissues, the patients were divided into a high-expression group and a low-expression group. The overall survival (OS) was compared between the 2 groups for each included gene.

Results

Identification of differentially expressed genes

Datasets of GSE19804, GSE101929, and GSE33532 from the GEO database were included in our study. The DEGs were first screened from each dataset, and 40 overlapping differentially expressed genes ID were identified (Figure 1). However, the 40 gene IDs correspond to 30 genes with 10 duplicate genes, and 1 gene ID had no gene name. Finally, 29 genes were included for further analysis, of which 22 were upregulated and 7 downregulated (Table 1). The differentially expressed genes between cancer tissue and lung normal tissue are represented in a heat map in Figure 2.

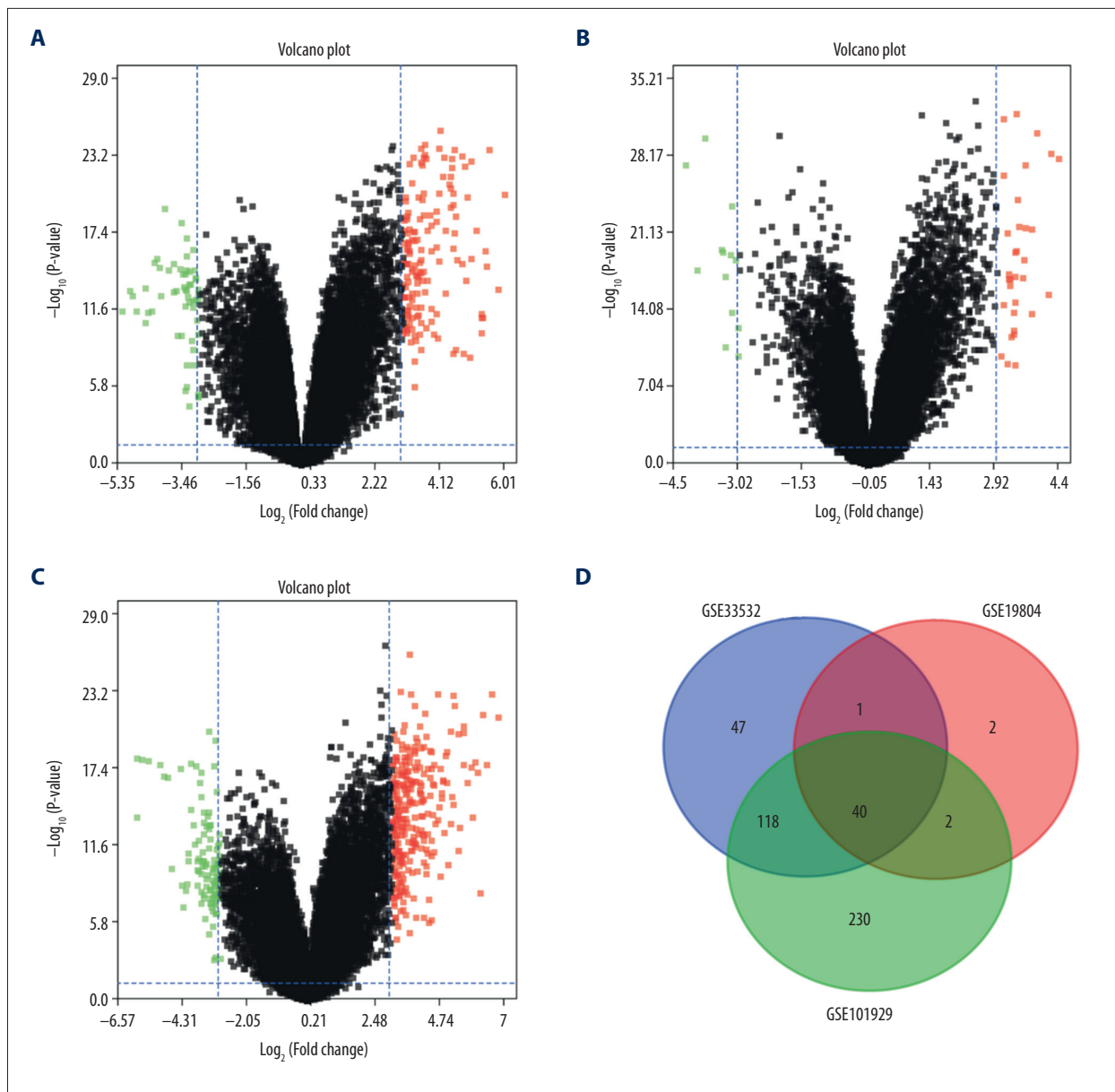


Figure 1. (A–D) Identification of differentially expressed genes from GSE33532, GSE19804, and GSE101929 data series (**A**: Volcano plot of GSE33532; **B**: Volcano plot of GSE19804; **C**: Volcano plot of GSE101929).

GO and KEGG analysis

The 29 dysregulated genes had gene ontology enrichment in terms of biological process (BP), cellular component (CC), and molecular function (MF). The enriched biological process was mainly related to diet-induced thermogenesis, ventricular cardiac muscle tissue morphogenesis, and brown fat cell differentiation. For the cellular component, the 29 genes were enriched in extracellular space, neuron projection, and plasma membrane. In the aspect of molecular function, only 1 term of actin filament binding was enriched. KEGG pathway analysis showed that the 29 dysregulated genes were enriched

in calcium signaling pathway, regulation of lipolysis in adipocytes, and PPAR signaling pathway (Table 2).

PPI network analysis of the 29 genes

The STRING database was used for PPI network analysis, showing 79 nodes and 336 edges, with the average node degree of 8.51 (Figure 3), and the local clustering coefficient was 0.648. We also use Cytohubba to select the hub genes, showing that 2 downregulated genes (*MMP1* and *SPP1*) were hub genes (Figure 4).

Table 1. The 29 included differentially expressed genes overlapping in GSE33532, GSE19804, and GSE101929 data series.

Gene ID	Gene symbol	Mean logFC (GSE19804)	Gene ID	Gene symbol	Mean logFC (GSE19804)
209612_s_at	ADH1B	3.36491817	204475_at	MMP1	-3.04218817
229309_at	ADRB1	3.21379367	204580_at	MMP12	-3.17540267
210081_at	AGER	3.21379367	239650_at	NCKAP5	3.105129
206209_s_at	CA4	3.86942117	230469_at	RTKN2	3.4138185
232578_at	CLDN18	4.16088183	205725_at	SCGB1A1	3.38117033
213317_at	CLIC5	3.45981	214387_x_at	SFTPC	3.2855315
204320_at	COL11A1	-3.32311183	242009_at	SLC6A4	3.59241617
225681_at	CTHRC1	-3.193161	213456_at	SOSTDC1	3.38619867
204273_at	EDNRB	3.190866	206239_s_at	SPINK1	-3.33911383
203980_at	FABP4	3.7473685	209875_s_at	SPP1	-4.503354
209074_s_at	FAM107A	3.4444825	230560_at	STXBP6	3.6274755
205866_at	FCN3	3.40527367	219230_at	TMEM100	3.56547367
238222_at	GKN2	3.25140117	209904_at	TNNC1	3.11718933
209469_at	GPM6A	3.61581183	204712_at	WIF1	3.77360267
230030_at	HS6ST2	-3.390935			

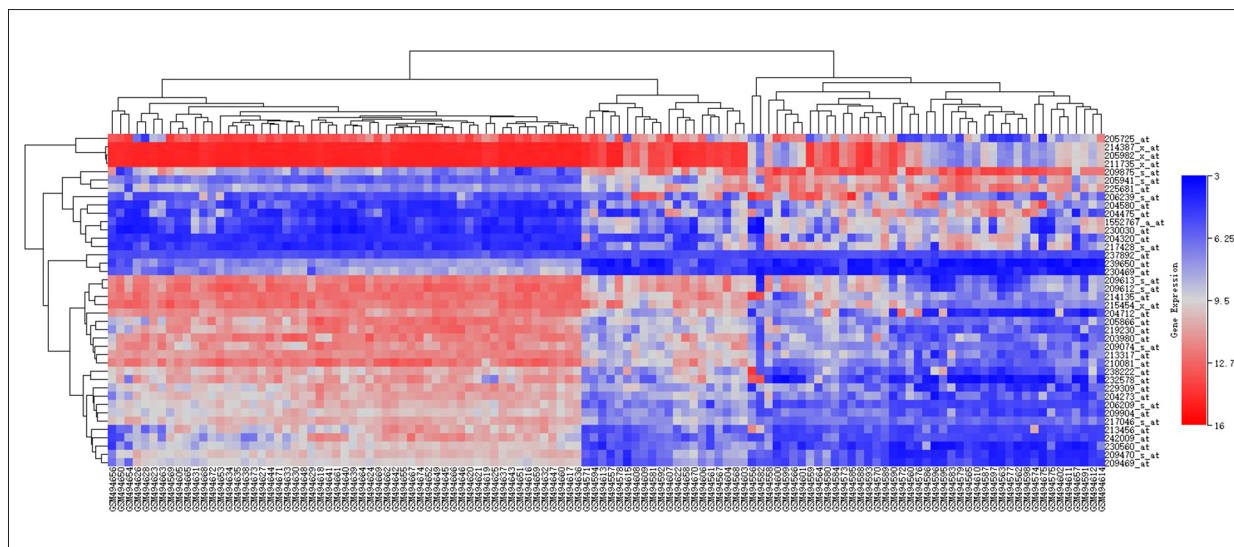


Figure 2. Heat map of the differentially expressed genes between cancer tissue and lung normal tissue.

Survival analysis

The prognostic significance of the 29 genes for NSCLC was analyzed in the Kaplan-Meier Plotter database. The significant difference in overall survival (OS) between upregulated and downregulated genes is shown in Figure 5. Twenty-two dysregulated genes were correlated with patient prognosis (Table 3).

Discussion

With the rapid development of bioinformatics, more and more microarrays and sequencing data can be publicly accessed [17]. These data are collected and stored in corresponding databases, such as GEO (<http://www.ncbi.nlm.nih.gov/geo/>), TCGA (<http://www.tcg.org/>), Kaplan-Meier Plotter, and STRING.

Table 2. GO and KEGG analysis of the differentially expressed genes between cancer tissue and lung normal tissue.

Category	Term	Count	P-value
GOTERM_BP_DIRECT	Diet-induced thermogenesis	2	9.9E-3
GOTERM_BP_DIRECT	Ventricular cardiac muscle tissue morphogenesis	2	3.4E-2
GOTERM_BP_DIRECT	Brown fat cell differentiation	2	4.2E-2
GOTERM_CC_DIRECT	Extracellular space	6	7.7E-3
GOTERM_CC_DIRECT	Neuron projection	3	1.3E-2
GOTERM_CC_DIRECT	Plasma membrane	8	2.5E-2
GOTERM_CC_DIRECT	Extracellular region	4	3.0E-2
GOTERM_CC_DIRECT	Collagen trimer	2	7.5E-2
GOTERM_MF_DIRECT	Actin filament binding	2	6.3E-2
KEGG_PATHWAY	Calcium signaling pathway	3	3.0E-2
KEGG_PATHWAY	Regulation of lipolysis in adipocytes	2	7.7E-2
KEGG_PATHWAY	PPAR signaling pathway	9.7E-2	

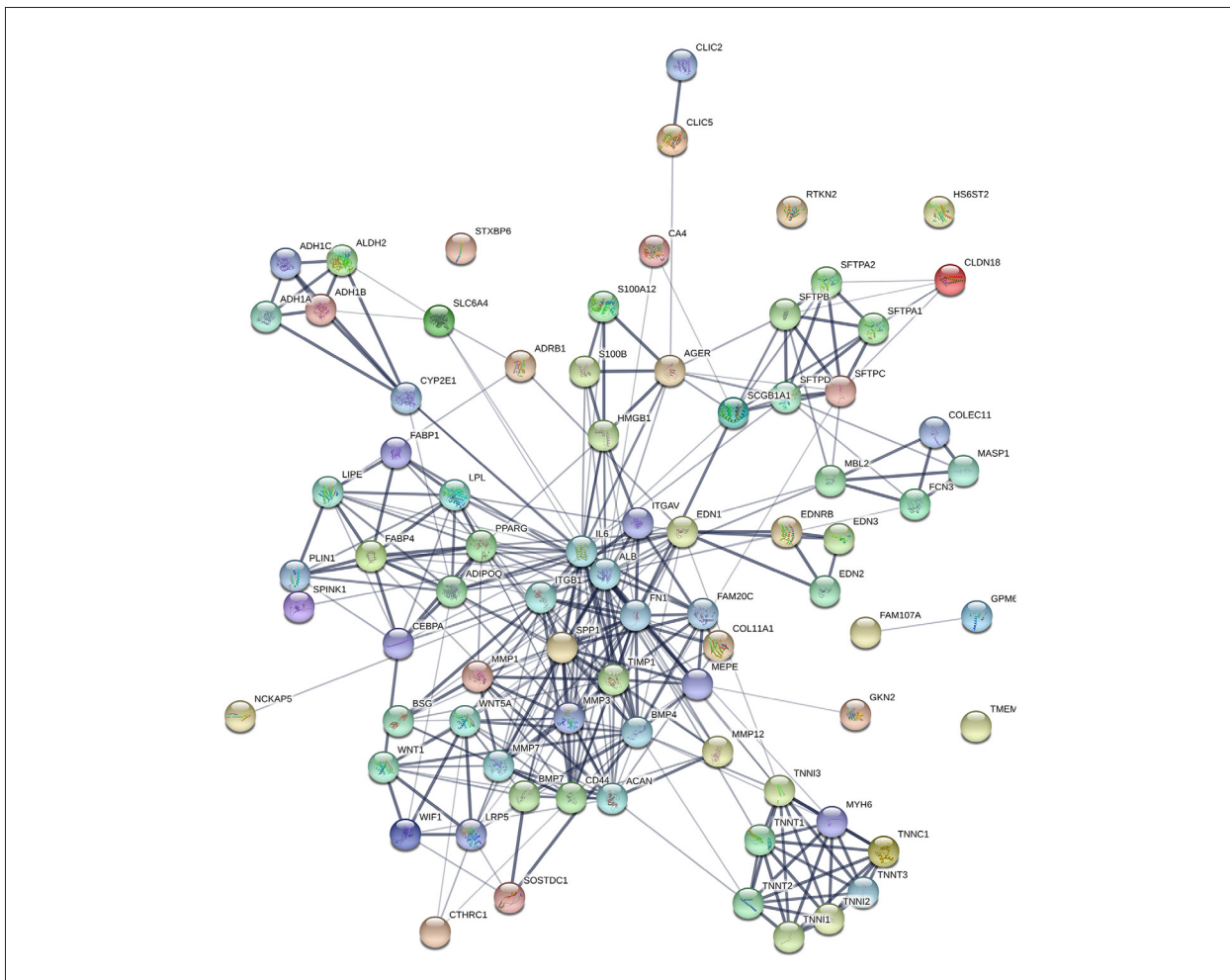


Figure 3. Protein-protein interaction (PPI) network of the 29 dysregulated genes.

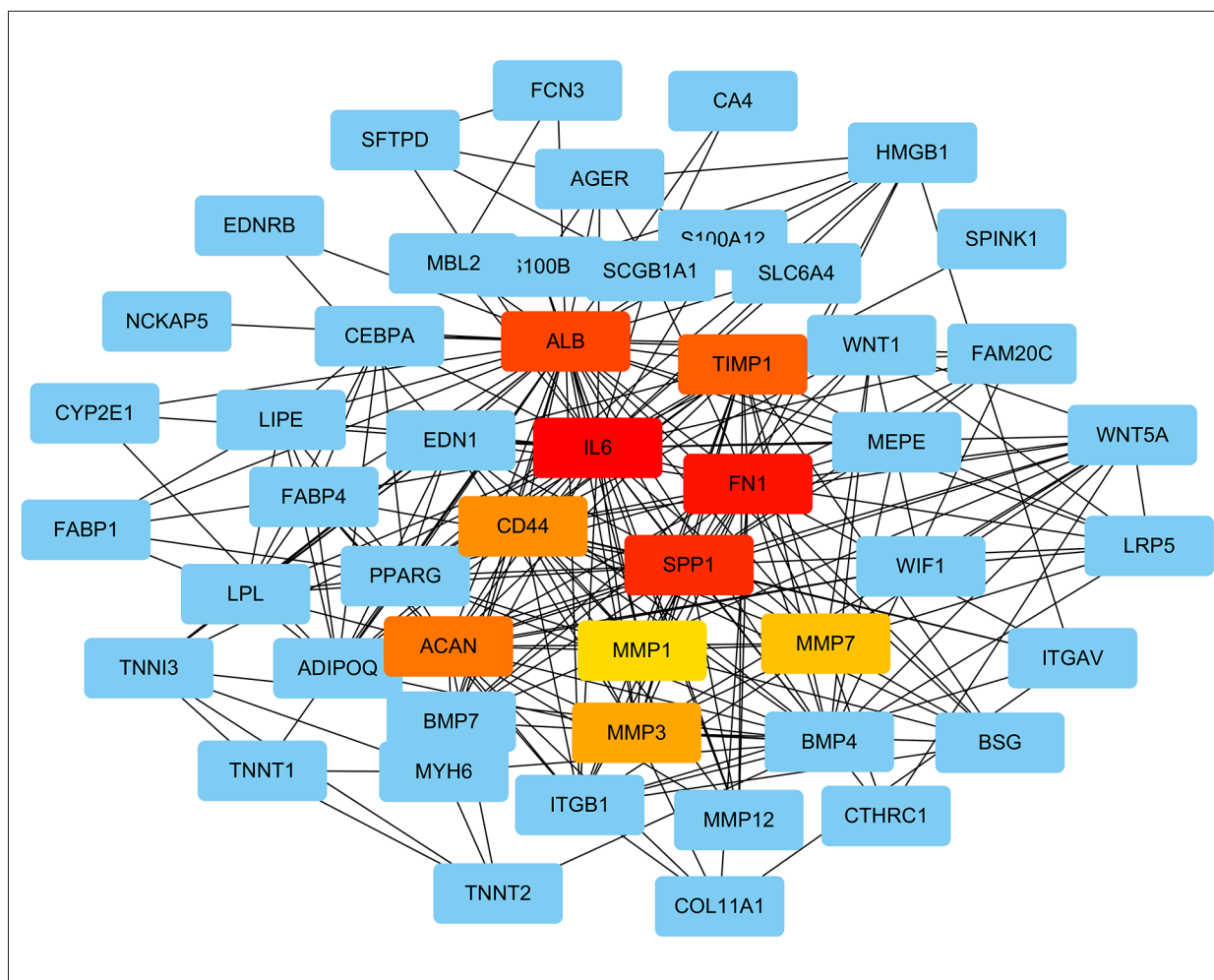


Figure 4. Hub gene identified by Cytohubba.

Clinical information (e.g., disease type, age, sex, and survival rate) and gene expression data can be freely downloaded or analyzed online, providing a reliable data platform for further data mining, analysis, and solving clinical problems [18,19].

The GEO database was established by the US National Library of Medicine in 2000. It is dedicated to the construction of gene expression databases and online analysis resources [20]. It mainly contains gene chip data and partial sequencing data of various tissues. At present, it is one of the most important databases in the field of bioinformatics data mining [7,21]. Fang et al. [22] performed integrative bioinformatics analysis, revealing potential long non-coding RNA biomarkers and analysis of function in non-smoking females with lung cancer. In that study, the authors found that 2 DEGs (LINC00968 and TBX5-AS1) were associated with unfavorable prognosis in never-smoking female lung cancer patients.

In our present work, we selected data on 3 gene chips relevant to differential expression between lung cancer tissues and

normal lung tissues of NSCLC patients in the GEO database. We finally identified 29 differentially expressed genes in 3 datasets and further analyzed them for biological function enrichment, pathways, and survival analysis. These 29 included dysregulated genes are mainly enriched in the biological function of diet-induced thermogenesis, ventricular cardiac muscle tissue morphogenesis, and actin filament binding. The KEGG pathway analysis showed that the 29 dysregulated genes were enriched in calcium signaling and regulation of lipolysis in adipocytes and in the PPAR signaling pathway. Further analysis showed that 2 genes (*MMP1* and *SPP1*) were hub genes. Matrix metalloproteinase-1 (*MMP-1*) is part of a cluster of *MMP* genes localized to chromosome 11q22.3. *MMP-1* is involved in the breakdown of extracellular matrix, which may play an important role in tumor metastasis by breaking down interstitial collagens types I, II, and III [23,24]. However, *SPP1* seems to have no correlation with cancer in terms of biological function enrichment [25,26].

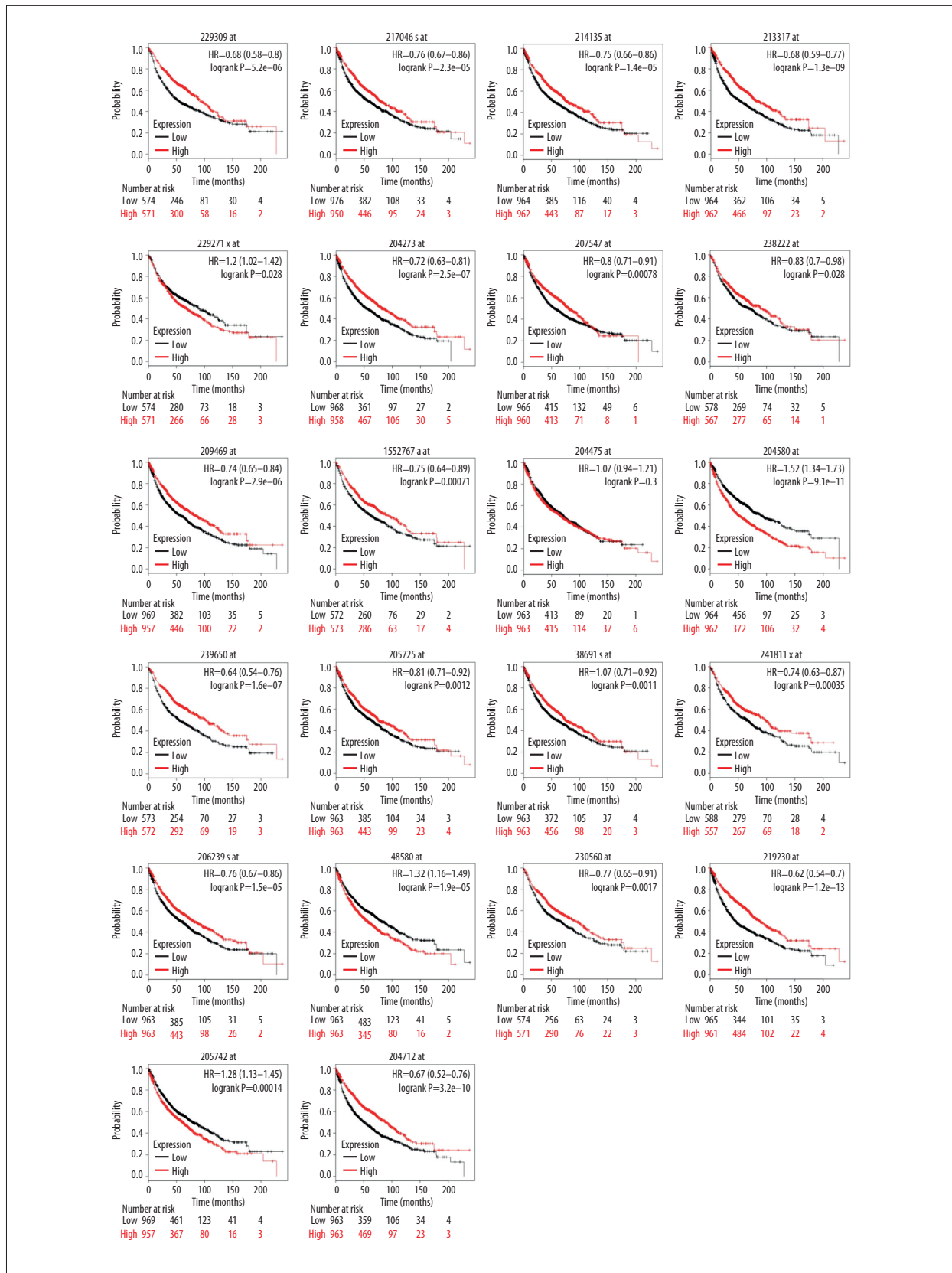


Figure 5. Survival curve of non-small cell lung cancer according to low and high expression of included genes.

Table 3. Survival analysis of the 29 included genes.

Gene ID	Gene symbol	HR (95% CI)	p-Value
209612_s_at	ADH1B	0.67 (0.59–0.76)	4.5E-10
229309_at	ADRB1	0.68 (0.58–0.80)	5.2e-6
210081_at	AGER	0.76 (0.67–0.86)	2.3e-5
206209_s_at	CA4	1.03 (0.9–1.165)	0.69
232578_at	CLDN18	0.75 (0.66–0.86)	1.4E-5
213317_at	CLIC5	0.68 (0.59–0.77)	1.3e-9
204320_at	COL11A1	1.2 (1.02–1.42)	0.028
225681_at	CTHRC1	1.11 (0.94–1.31)	0.21
204273_at	EDNRB	0.72 (0.36–0.81)	2.5e-7
203980_at	FABP4	1.02 (0.9–1.16)	0.78
209074_s_at	FAM107A	0.80 (0.71–0.91)	0.00078
205866_at	FCN3	0.99 (0.87–1.12)	0.88
238222_at	GKN2	0.83 (0.70–0.98)	0.028
209469_at	GPM6A	0.74 (0.65–0.84)	2.9e-6
230030_at	HS6ST2	0.75 (0.64–0.89)	0.00071
204475_at	MMP1	1.07 (0.94–1.21)	0.30
204580_at	MMP12	1.52 (1.34–1.73)	9.1e-11
239650_at	NCKAP5	0.64 (0.54–0.76)	1.6e-7
230469_at	RTKN2	1.02 (0.86–1.20)	0.85
205725_at	SCGB1A1	0.81 (0.71–0.92)	0.0012
214387_x_at	SFTPC	0.81 (0.71–0.92)	0.0011
242009_at	SLC6A4	0.74 (0.63–0.87)	0.00035
213456_at	SOSTDC1	1.07 (0.94–1.21)	0.32
206239_s_at	SPINK1	0.765 (0.67–0.86)	1.5e-5
209875_s_at	SPP1	1.32 (1.16–1.49)	1.9e-5
230560_at	STXBP6	0.77 (0.65–0.91)	0.0017
219230_at	TMEM100	0.62 (0.54–0.71)	1.2e-13
209904_at	TNNC1	1.28 (1.13–1.45)	0.00014
204712_at	WIF1	0.67 (0.59–0.76)	3.2e-10

Our survival analysis indicated that 22 of the 29 included dysregulated genes were correlated with patient prognosis, suggesting that these 22 genes could be used as biomarkers for patient prognosis.

Conclusions

Twenty-nine differently expressed genes were identified in the present work, which were enriched in the biological functions

of diet-induced thermogenesis, actin filament binding, and PPAR signaling pathway. Dysregulated genes were correlated with NSCLC patient survival and might be useful as biomarkers of prognosis. However, this conclusion needs further confirmation by laboratory experiments.

Conflict of interest

None.

References:

1. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2019. *Cancer J Clin*, 2019; 69: 7–34
2. Verghese C, Redko C, Fink B: Screening for lung cancer has limited effectiveness globally and distracts from much needed efforts to reduce the critical worldwide prevalence of smoking and related morbidity and mortality. *J Glob Oncol*, 2018; 4: 1–7
3. Torre LA, Bray F, Siegel RL et al: Global cancer statistics, 2012. *Cancer J Clin*, 2015; 65: 87–108
4. Chen W, Zheng R, Baade PD et al: Cancer statistics in China, 2015. *Cancer J Clin*, 2016; 66: 115–32
5. Cao M, Chen W: Epidemiology of lung cancer in China. *Thorac Cancer*, 2019; 10: 3–7
6. Azzawi H, Hou J, Xiang Y et al: Lung cancer prediction from microarray data by gene expression programming. *IET Syst Biol*, 2016; 10: 168–78
7. Barrett T, Suzek TO, Troup DB et al: NCBI GEO: Mining millions of expression profiles – database and tools. *Nucleic Acids Res*, 2005; 33: D562–66
8. Lu TP, Tsai MH, Lee JM et al: Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol Biomarkers Prev*, 2010; 19: 2590–97
9. Mitchell KA, Zingone A, Toulabi L et al: Comparative transcriptome profiling reveals coding and noncoding RNA differences in NSCLC from African Americans and European Americans. *Clin Cancer Res*, 2017; 23: 7412–25
10. Meister M, Belousov A, Xu EC et al: Intra-tumor heterogeneity of gene expression profiles in early stage non-small cell lung cancer. *Journal of Bioinformatics Research Studies*, 2014; 1(1): 1
11. Dennis G, Sherman BT, Hosack DA et al: DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol*, 2003; 4: P3
12. Kuznetsova I, Lugmayr A, Siira SJ et al: CirGO: An alternative circular way of visualising gene ontology terms. *BMC Bioinformatics*, 2019; 20: 84
13. Pomaznoy M, Ha B, Peters B: GOnet: A tool for interactive Gene Ontology analysis. *BMC Bioinformatics*, 2018; 19: 470
14. Ogata H, Goto S, Sato K et al: KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 1999; 27: 29–34
15. Doncheva NT, Morris JH, Gorodkin J et al: Cytoscape StringApp: Network analysis and visualization of proteomics data. *J Proteome Res*, 2019; 18: 623–32
16. Hou GX, Liu P, Yang J et al: Mining expression and prognosis of topoisomerase isoforms in non-small-cell lung cancer by using OncoPrint and Kaplan-Meier plotter. *PLoS One*, 2017; 12: e0174515
17. Hasan MM, Khatun MS, Kurata H: Large-scale assessment of bioinformatics tools for lysine succinylation sites. *Cell*, 2019; 8: pii: E95
18. Bris C, Goudenege D, Desquiret-Dumas V et al: Bioinformatics tools and databases to assess the pathogenicity of mitochondrial DNA variants in the field of next generation sequencing. *Front Genet*, 2018; 9: 632
19. Craveiro SAS, de Azevedo Medeiros LB, Agnez-Lima LF et al: Exploring seipin: From biochemistry to bioinformatics predictions. *Int J Cell Biol*, 2018; 2018: 5207608
20. Barrett T, Troup DB, Wilhite SE et al: NCBI GEO: Mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Res*, 2007; 35: D760–65
21. Barrett T, Edgar R: Reannotation of array probes at NCBI's GEO database. *Nat Methods*, 2008; 5: 117
22. Qiao F, Li N, Li W: Integrative bioinformatics analysis reveals potential long non-coding RNA biomarkers and analysis of function in non-smoking females with lung cancer. *Med Sci Monit*, 2018; 24: 5771–78
23. Ming XY, Zhang X, Cao TT et al: RHCG suppresses tumorigenicity and metastasis in esophageal squamous cell carcinoma via inhibiting NF-κB signaling and MMP1 expression. *Theranostics*, 2018; 8: 185–98
24. Liu M, Hu Y, Zhang MF et al: MMP1 promotes tumor growth and metastasis in esophageal squamous cell carcinoma. *Cancer Lett*, 2016; 377: 97–104
25. Acquaviva L, Drogat J, Dehé PM et al: Spp1 at the crossroads of H3K4me3 regulation and meiotic recombination. *Epigenetics*, 2013; 8: 355–60
26. Oliveira L, Tavares P, Alonso JC: Headful DNA packaging: Bacteriophage SPP1 as a model system. *Virus Res*, 2013; 173: 247–59