Int J Clin Exp Med 2019;12(11):12702-12710
www.ijcem.com /ISSN:1940-5901/IJCEM0099076

Original Article

Identification of differentially-expressed genes in lung squamous cell carcinoma and correlation levels with prognosis through integrated bioinformatics analysis

Licui Zhang[1*], Chen Zhong[2*], Yang Gu[1], Yajing Ma[1], Xinliang Ming[3], Xin Su[1], Min Liu[1]

[1]Department of Clinical Laboratory, The First Affiliated Hospital of The Medical College, Shihezi University, Shihezi 832000, Xinjiang Autonomous Region, PR China; [2]Department of Gastrointestinal & Thyroid Surgery, The First Affiliated Hospital of The Medical College, Shihezi University, Shihezi 832000, Xinjiang Autonomous Region, PR China; [3]Department of Genetic Diagnosis Center, Zhongnan Hospital of Wuhan University, Wuhan 430000, PR China. [*]Equal contributors.

Received June 30, 2019; Accepted September 3, 2019; Epub November 15, 2019; Published November 30, 2019

Abstract: Background: The aim of the current study was to screen differentially-expressed genes (DEGs) relevant to cancer progression and prognosis of squamous cell lung carcinoma (SqCLC). Methods: DEGs mRNA expression data of SqCLC was screened from the Oncomine database. This data was further analyzed by comparing tumor tissues to normal tissues. Prognostic values of DEGs relevant to SqCLC were investigated using the "Kaplan-Meier Plotter" (KM plotter) database. Bioinformation for included genes was analyzed by gene GO and KEGG enrichment, aiming to explain the potential roles of identified genes in SqCLC. Protein-protein interaction (PPI) of the genes was evaluated using the STRING database. Results: Four independent microarray datasets relevant to SqCLC were identified in the Oncomine database, with the top 10 consistently upregulated and top 10 consistently downregulated genes included in the present analysis. Significant differences of overall survival (OS) were correlated with *SMC4, HIST2H2AA3, GMPS, CKS1B, POLR2H, PDCD10, PLOD2, DVL3, C-type CLEC3B, TNNC1, FAM107A, FYR, MEF2C, SLIT3, CX3CR1, C17orf91*, *LIM, and LIMCH1* (all $P < 0.05$). Possible protein-protein interaction analysis of the top 20 dysregulated genes showed that proteins of *SMC4, POLR2H,* and *NCBP2* in upregulated genes and *TNNC1* and *MEF2C* in downregulated genes interacted with more than 5 other proteins. This may play an important role in the development of SqCLC. Conclusion: *MC4, POLR2H, TNNC1,* and *MEF2C* genes were dysregulated in SqCLC. Thus, they may play an essential role in the development of SqCLC, as biomarkers for patient prognosis.

Keywords: Squamous cell lung carcinoma, data mining, prognosis, Oncomine database

Introduction

Lung cancer is one of the most diagnosed malignant tumors, clinically, and the leading cause of carcinoma related deaths [1]. Epidemiology studies have indicated that lung cancer ranks in the top 3 for cancer incidence rates in the U.S [2, 3]. Therefore, lung cancer has become a big problem for humans, placing a heavy burden on the government [4]. Squamous cell lung carcinoma (SqCLC) is one of the major subtypes of non-small cell lung cancer (NSCLC), accounting for 40% of all diagnosed lung carcinomas [5, 6]. Prognosis of SqCLC remains poor. SqCLC is difficult to manage, as it is not sensitive to chemotherapy and radiation [7, 8].

Differentially-expressed genes (DEGs) are common in malignant tumors. They play an essential role in cancer cell development. Liu et al. [9] evaluated genes that correlated with lung adenocarcinoma progression and prognosis using Oncomine and Cancer Genome Atlas databases. They found eighty DEGs between cancer tissues and corresponding normal lung tissues. Results suggested that AGER and CCNB1 can be used as potential biomarkers for lung adenocarcinoma diagnosis and prognosis. Dysregulated expression of cancer related genes is common in SqCLC, playing an important role in development and prognosis [10]. However, identification of dysregulated genes and analysis of the correlation and relationship with SqCLC patients prognosis remain difficult in

Table 1. General information of included data sets

Data sets	Method	Sample size	Genes measured	Data type	Year
Bhattacharjee	Human Genome U95A-Av2 Array	203	8,603	mRNA	2001
Hou	Human Genome U133 Plus 2.0	156	19,574	mRNA	2010
Talbot	Human Genome U95A-Av2 Array	93	8,603	mRNA	2005
Yamagata	Platform not pre-defined	31	2,509	mRNA	2003

many datasets [11]. The present study provides a method of screening dysregulated genes relevant to SqCLC, examining biological function and correlation levels with patient prognosis.

Material and methods

Database selection

The Oncomine database (https://www.oncomine.org/) provides differential expression analyses. It compares most major types of cancer with respective normal tissues, as well as a variety of cancer subtypes, providing clinical-based and pathology-based analyses. Dysregulated genes between cancer and normal tissues of SqCLC patients were first screened through the Oncomine database (https://www.oncomine.org/). The top 120 dysregulated genes were included and 20 constantly upregulated or downregulated genes were further analyzed for prognosis (**Table 1**). Biological information of the included 20 gene were enriched through ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). Protein-protein interaction (PPI) of the 20 constantly dysregulated genes were investigated using the STRING database.

Data extraction from Oncomine database and analysis

Using the Oncomine database, the current study screened all dysregulated genes between cancer tissues and normal tissues of SqCLC patients. Data screen restrictions was as follows: (1) Cancer type: squamous cell lung carcinoma; (2) Tissue comparison: cancer tissue versus normal tissue; (3) Data type: mRNA; (4) mRNA fold change: more than two fold; and (5) Significance: $P < 0.001$. The top 120 dysregulated genes (60 upregulated and 60 downregulated) were included in the biological information analysis via GO and KEGG enrichment. Of the included 120 genes, 20 constantly dysregulated genes (10 upregulated and 10 downregulated genes) were selected for survival and PPI analysis.

Survival analysis by kaplan-meier plotter

Prognosis (overall survival, OS) significance of the 20 constantly dysregulated genes of SqCLC patients was analyzed using the Kaplan-Meier Plotter database. According to the median mRNA expression level, the data was divided into two groups. These included a high expression and low expression group for each of the included 20 genes.

Bioinformatics analyses of the databases

Online analysis software DAVID (https://david.ncifcrf.gov/) was used to enrich the biological information of the top 120 genes screened in the Oncomine database. Two-tailed $P < 0.05$ was set as the cut-off in enrichment of GO and KEGG.

Results

Dysregulated genes for SqCLC

Four independent microarray datasets, relevant to SqCLC, were identified in the Oncomine database [12-15]. The current study identified the 120 most dysregulated genes, comparing tumor tissues to normal tissues of SqCLC patients ($P < 0.05$). Included patients had comparable demographic characteristics. Regarding the 120 dysregulated genes, 60 were upregulated (**Figure 1**) and 60 were downregulated (**Figure 2**). Of the 120 most dysregulated genes, 10 consistently upregulated and 10 consistently downregulated genes were selected as objective genes for further analysis.

Functional enrichment analyses

The 20 dysregulated genes contained three enriched gene ontology categories, including BP: Biological process, CC: Cellular component,

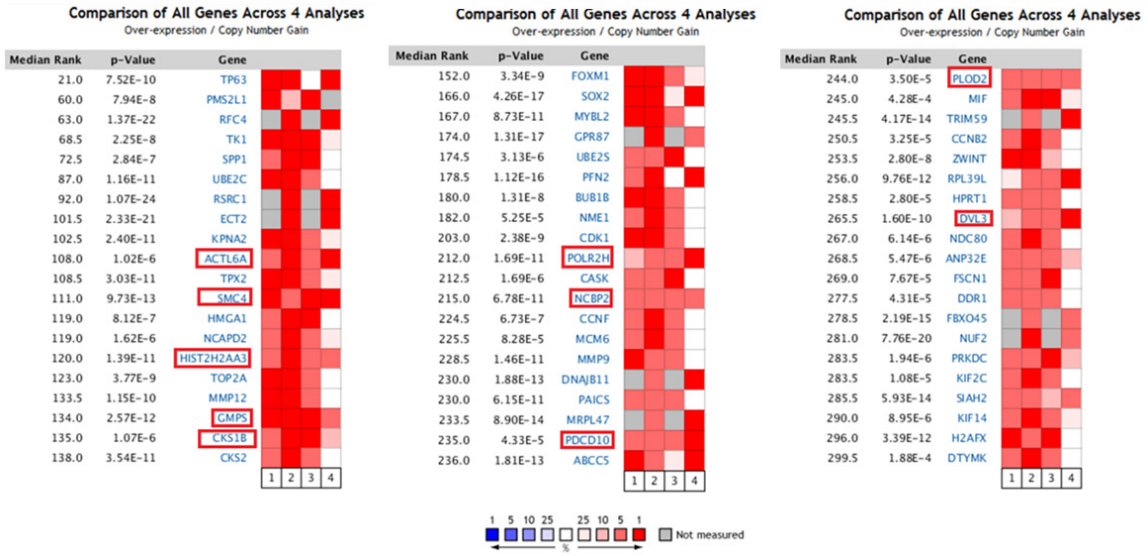


Figure 1. The 60 most upregulated genes, comparing tumor tissues to normal tissues of SqCLC patients screened in the Oncomine database.

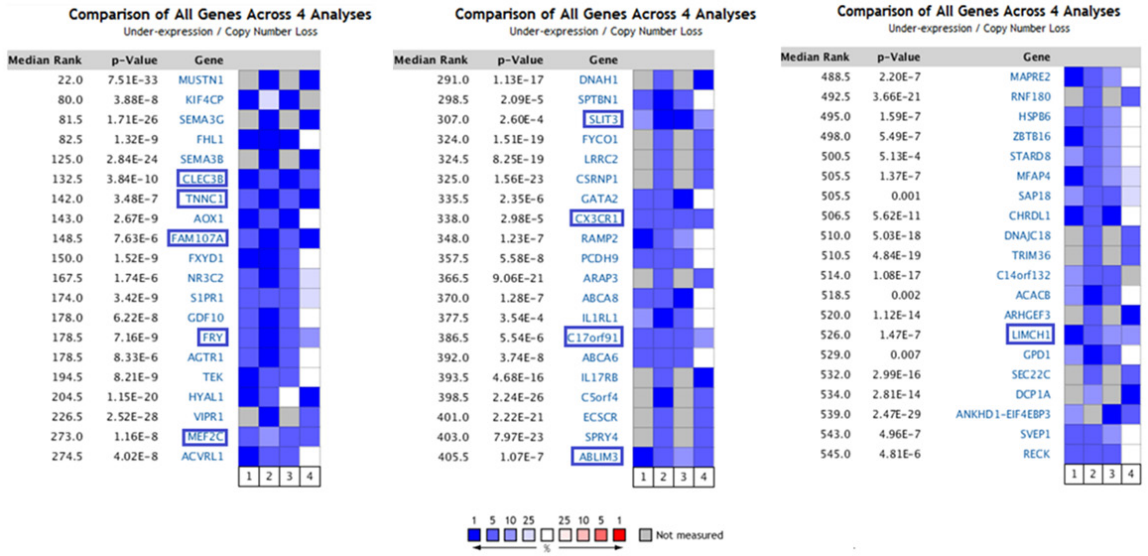


Figure 2. The 60 most downregulated genes, comparing tumor tissues to normal tissues of SqCLC patients screened in the Oncomine database.

and MF: molecular function. For the 10 upregulated genes, enriched biological processes were mainly related to positive regulation of transcription from RNA polymerase II promoter, negative regulation of the apoptotic process, activation of protein kinase activity, and protein homotetramerization (**Figure 3A**). In the CC category, the membrane, cytosol, nucleoplasm, nucleus, and cytoplasm were mainly enriched in the cellular component of the 10 upregulated genes (**Figure 3B**). Regarding MF, ATP binding, chromatin binding, transcription regulatory region DNA binding, histone binding, and ubiquitin conjugating enzyme activity were the top 4 enrichments (**Figure 3C**). Ten upregulated genes were enriched in cell cycle, purine metabolism, pyrimidine metabolism, and drug metabolism-other enzymes (**Figure 3D**).

Regarding the 10 downregulated genes, BP enrichment included negative regulation of gene expression, angiogenesis, positive regulation of angiogenesis, and positive regulation of GTPase activity (**Figure 4A**). For the cellular

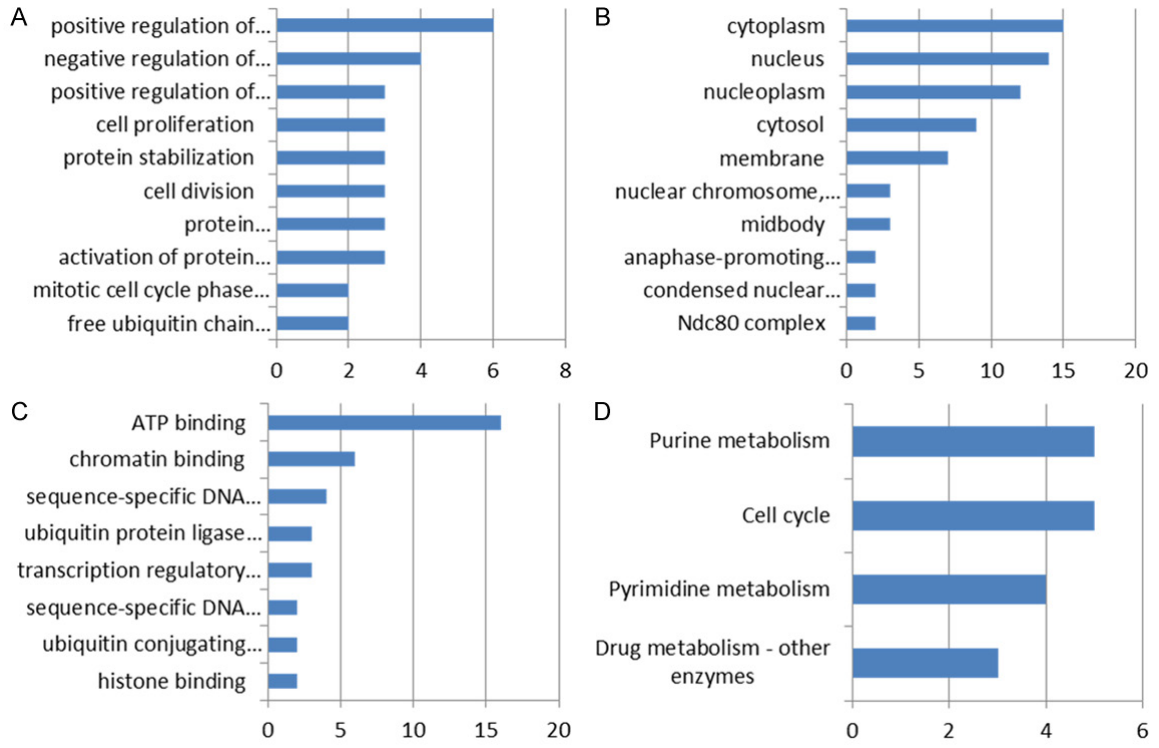


Figure 3. Go and KEGG analysis of the 10 upregulated genes (A: Biological process; B: Cellular component; C: Molecular function; D: KEGG).

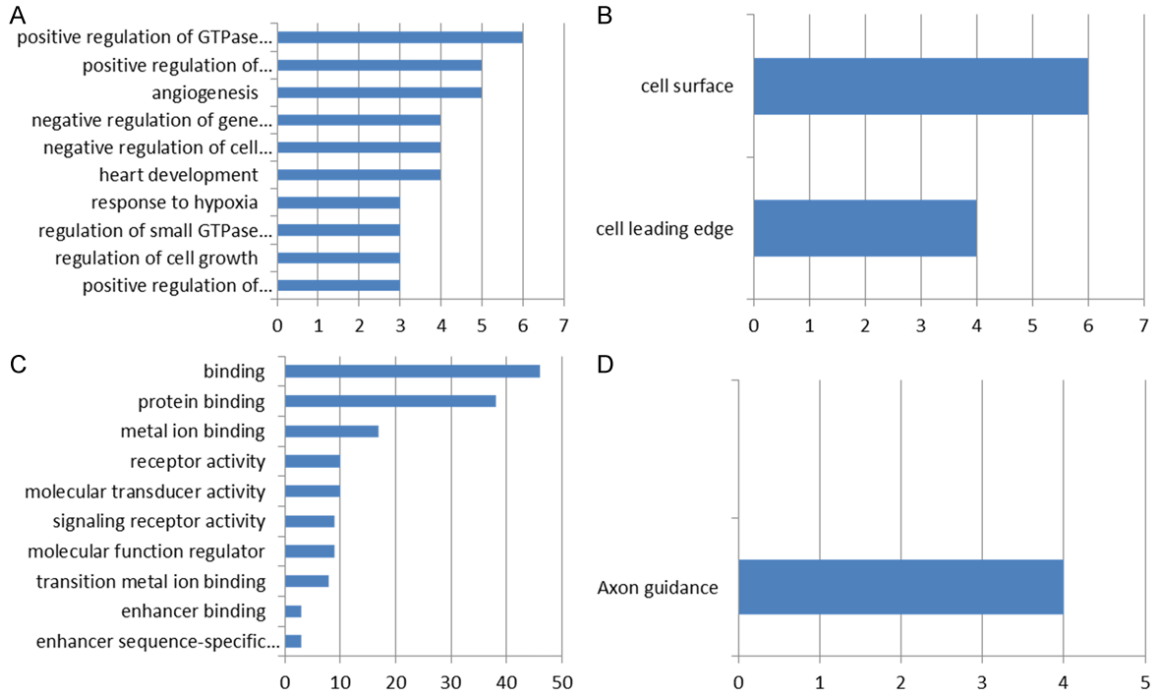
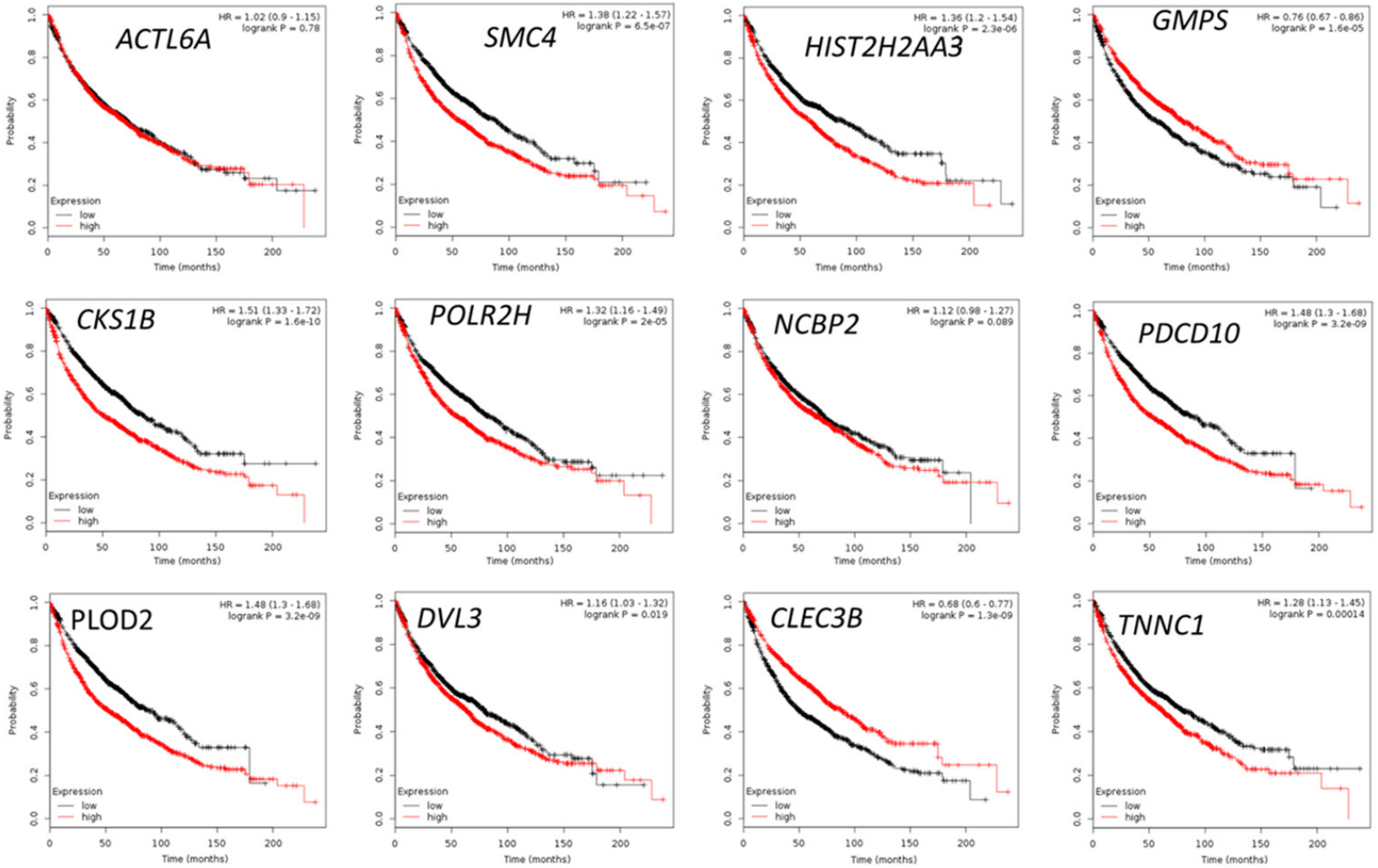


Figure 4. Go and KEGG analysis of the 10 downregulated genes. Y-axis demonstrates the gene function and pathway enrichment and X-axis indicates counts. (A: Biological process; B: Cellular component; C: Molecular function; D: KEGG).


ACTL6A
HR = 1.02 (0.9 - 1.15)
logrank P = 0.78
SMC4
HR = 1.38 (1.22 - 1.57)
logrank P = 6.5e-07
HIST2H2AA3
HR = 1.36 (1.2 - 1.54)
logrank P = 2.3e-06
GMPS
HR = 0.76 (0.67 - 0.86)
logrank P = 1.6e-05
CKS1B
HR = 1.51 (1.33 - 1.72)
logrank P = 1.6e-10
POLR2H
HR = 1.32 (1.16 - 1.49)
logrank P = 2e-05
NCBP2
HR = 1.12 (0.98 - 1.27)
logrank P = 0.089
PDCD10
HR = 1.48 (1.3 - 1.68)
logrank P = 3.2e-09
PLOD2
HR = 1.48 (1.3 - 1.68)
logrank P = 3.2e-09
DVL3
HR = 1.16 (1.03 - 1.32)
logrank P = 0.019
CLEC3B
HR = 0.68 (0.6 - 0.77)
logrank P = 1.3e-09
TNNC1
HR = 1.28 (1.13 - 1.45)
logrank P = 0.00014
Probability
Expression
low
high
Time (months)

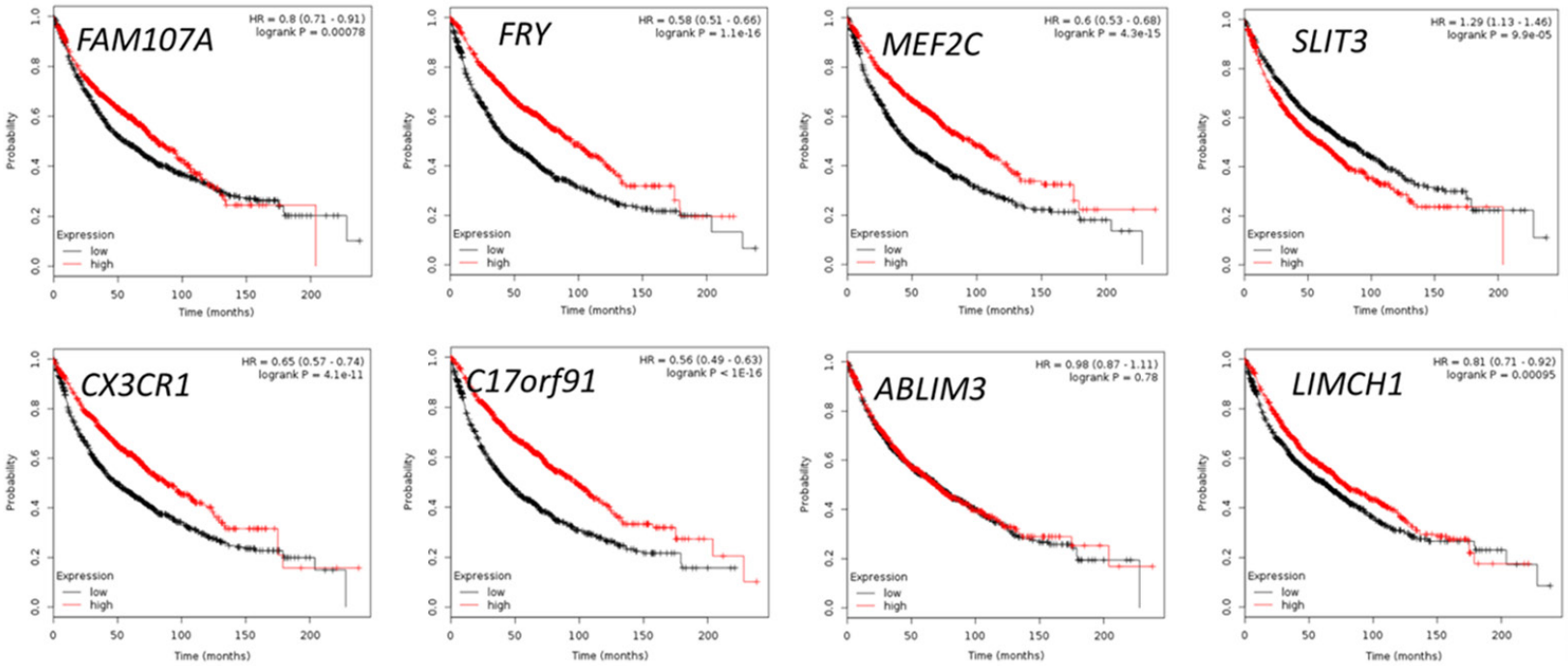


Figure 5. Survival curve comparing low and high expression of the 20 genes.

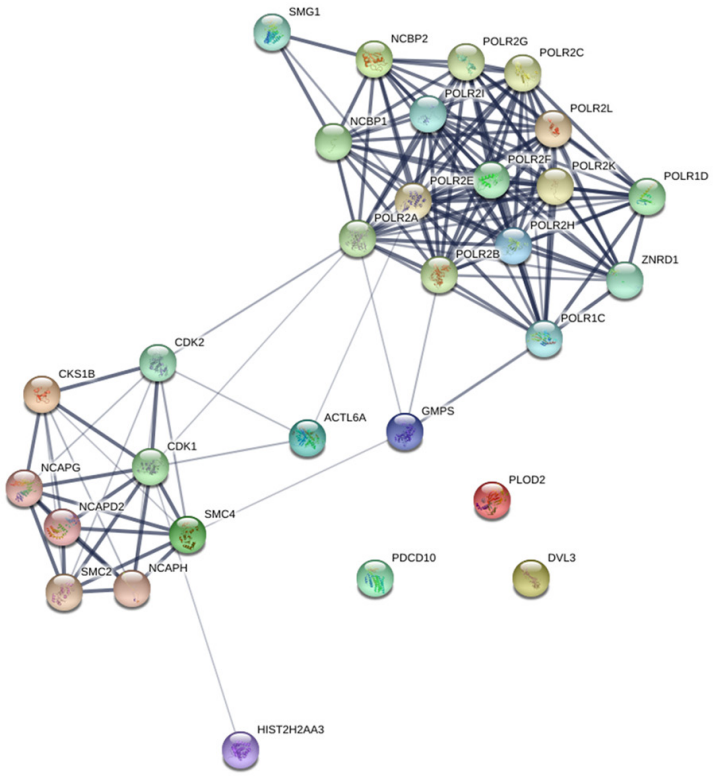


Figure 6. Protein-protein interaction of the 10 upregulated genes. Each node indicates the relevant protein.

component, two items were enriched, cell leading edge and cell surface (**Figure 4B**). Binding, protein binding, metal ion binding, and molecular transducer activity were enriched in the category of molecular function enrichment (**Figure 4C**). Concerning KEGG analysis, only axon guidance was included (**Figure 4D**).

Potential roles of genes in SqCLC progression

The prognostic significance of mRNA expression levels of the 20 genes and prognosis of lung cancer patients was analyzed via online software. Differences in OS between upregulated and down regulated genes are shown in **Figure 5**. OS was correlated with *SMC4, HIST2H2AA3, GMPS, CKS1B, POLR2H, PDCD10, PLOD2, DVL3, CLEC3B, TNNC1, FAM107A, FYR, MEF2C, SLIT3, CX3CR1, C17orf91,* and *LIMCH1* (all $p < 0.05$).

Protein-protein interaction of the 20 dysregulated genes

The possible protein-protein interaction of the top 20 dysregulated genes was analyzed using the STRING database. Results revealed that proteins of SMC4, POLR2H, and NCBP2 in upregulated genes (**Figure 6**) and TNNC1 and MEF2C in downregulated genes (**Figure 7**) interacted with more than 5 other proteins.

Discussion

In recent years, bioinformation anlaysis has developed quickly. Thus, more and more microarray and sequence datasets can be used in open databases, including Oncomine, GEO (http://www.ncbi.nlm.nih.gov/geo), TGCA (http://www.tcga.org/), and Kaplan-Meier Plotter [16, 17]. Relevant data, such as gene expression and clinical information (disease type, age, gender, survival), can be freely downloaded or analyzed. This data may be further used for clinical practices or deep data mining.

In the Oncomine database, gene expression data can be screened and mined according to the needs of researchers [18, 19]. Moreover, researchers can discover genes highly ranked by over-expression or DNA copy gain in medulloblastoma clinical specimens. This database allows researchers to explore interest genes in the largest normal tissue panel and survey a growing collection of TCGA gene expression and DNA copy number datasets [17, 20-22].

Long et al. [23] identified DEGs and enriched pathways in lung carcinoma using bioinformatics methods. In their study, the author screened GEO databases and selected GSE19804 as the study data. They identified DEGs and relevant pathways. They also provided survival data.

The present study screened the Oncomine database, finding the top 120 dysregulated genes relevant to SqCLC. Of the 120 included genes, 60 upregulated genes were enriched in the biological functions of positive regulation of transcription from RNA polymerase II promoter, negative regulation of apoptotic process, activation of protein kinase activity, protein homotetramerization with the KEGG pathway enrichment of cell cycle, purine metabolism, pyrimidine metabolism, and drug metabolism-other enzymes. Regarding the top 60

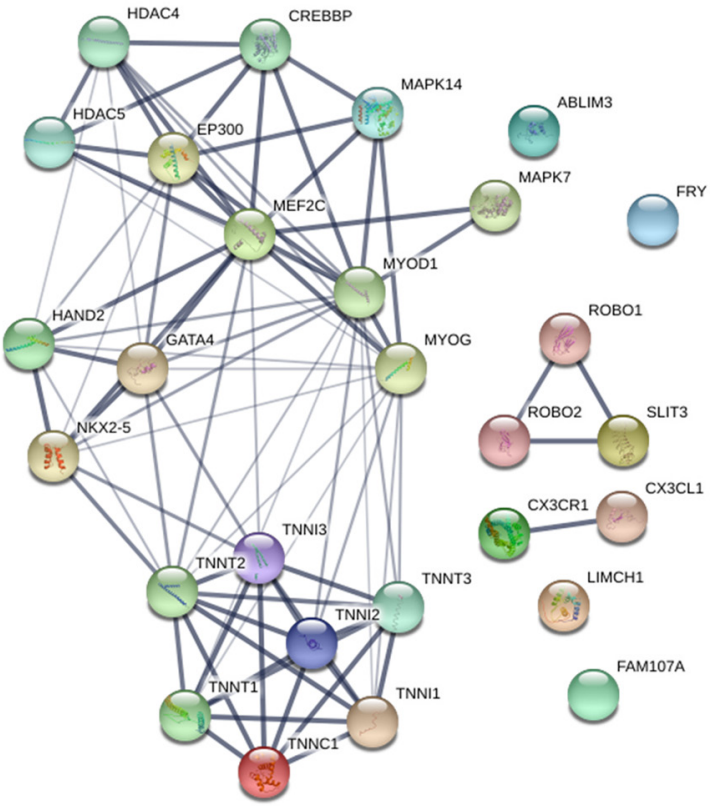


Figure 7. Protein-protein interaction of the 10 downregulated genes.

downregulated genes, biological functions and KEGG pathways were enriched in the aspects of negative regulation of gene expression, angiogenesis, positive regulation of angiogenesis, positive regulation of GTPase activity, and axon guidance. These functions or pathways are correlated with cancer development or metastasis.

Of the top 120 dysregulated genes, 20 genes were constantly upregulated or downregulated in the 4 datasets [12-15]. The current study further analyzed the prognostic significance of the 20 constantly dysregulated genes. Results indicated that *SMC4, HIST2H2AA3, GMPS, CKS1B, POLR2H, PDCD10, PLOD2, DVL3, CLEC3B, TNNC1, FAM107A, FYR, MEF2C, SLIT3, CX3CR1, C17orf91,* and *LIMCH1* were correlated with SqCLC patient overall survival, according to Kaplan-Meier survival curves (all $P < 0.05$).

High expression of *SMC4, HIST2H2AA3, CKS1B, POLR2H, PDCD10, PLOD2, DVL3, TNNC1,* and *SLIT3* was correlated with poor prognosis for SqCLC patients. However, for *GMPS, CLEC3B, FAM107A, FYR, MEF2C, CX3CR1, C17orf91,* and *LIMCH1* genes, high expression of mRNA was correlated with good prognosis. Results suggest that the above gene expression levels can be used as biomarkers for SqCLC patient prognosis. The current study also screened PubMed databases, but did not discover any studies that used the above genes coding protein as biomarkers for SqCLC prognosis. Therefore, present findings suggest that the survival biomarkers of the above genes may be applied to mRNA levels. PPI analysis of the top 20 dysregulated genes demonstrated that *SMC4, POLR2H, NCBP2, TNNC1,* and *MEF2C* interacted with more than 5 other proteins. Results suggest that these 5 genes may play important roles in SqCLC development.

Conclusion

In summary, the current study screened the Oncomine database. A total of 4 microarray datasets were included and analyzed in the present work. After deep data mining, it was found that *MC4, POLR2H, TNNC1,* and *MEF2C* genes were dysregulated in SqCLC. Thus, they may play essential roles in SqCLC development. They may be used as biomarkers of SqCLC prognosis. However, present results require further validation in future studies.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Min Liu, Department of Clinical Laboratory, The First Affiliated Hospital of The Medical College, Shihezi University, No. 107 Beier Road, Shihezi 832008, Xinjiang Autonomous Region, PR China. E-mail: jm37wc@163.com

References

[1] Torre LA, Siegel RL and Jemal A. Lung cancer statistics. Adv Exp Med Biol 2016; 893: 1-19.
[2] Siegel RL, Miller KD and Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018; 68: 7-30.
[3] Siegel RL, Miller KD and Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019; 69: 7-34.

[4] Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu XQ and He J. Cancer statistics in China, 2015. CA Cancer J Clin 2016; 66: 115-32.
[5] Fu XJ, Lu P, Ye GJ and Wang CY. Analysis of the risk factors of peripherally inserted central catheter-associated venous thrombosis after chemotherapy in patients with lung cancer. Int J Clin Exp Med 2019; 12: 5852-5859..
[6] Cao M and Chen W. Epidemiology of lung cancer in China. Thorac Cancer 2019; 10: 3-7.
[7] Xia R, Xu G, Huang Y, Sheng X, Xu X and Lu H. Silencing of ILT4 suppresses migration and invasion of non-small cell lung cancer cells by inhibiting MMP-2. Int J Clin Exp Med 2019; 12: 5306-5314.
[8] Zhang Y, Zhou H and Zhang L. Which is the optimal immunotherapy for advanced squamous non-small-cell lung cancer in combination with chemotherapy: anti-PD-1 or anti-PD-L1. J Immunother Cancer 2018; 6: 135.
[9] Liu W, Ouyang S, Zhou Z, Wang M, Wang T, Qi Y, Zhao C, Chen K and Dai L. Identification of genes associated with cancer progression and prognosis in lung adenocarcinoma: analyses based on microarray from oncomine and the cancer genome atlas databases. Mol Genet Genomic Med 2019; 7: e00528.
[10] Huang H, Huang Q, Tang T, Zhou X, Gu L, Lu X and Liu F. Differentially expressed gene screening, biological function enrichment, and correlation with prognosis in non-small cell lung cancer. Med Sci Monit 2019; 25: 4333-4341.
[11] Tian ZQ, Li ZH, Wen SW, Zhang YF, Li Y, Cheng JG and Wang GY. Identification of commonly dysregulated genes in non-small-cell lung cancer by integrated analysis of microarray data and qRT-PCR validation. Lung 2015; 193: 583-592.
[12] Yamagata N, Shyr Y, Yanagisawa K, Edgerton M, Dang TP, Gonzalez A, Nadaf S, Larsen P, Roberts JR, Nesbitt JC, Jensen R, Levy S, Moore JH, Minna JD and Carbone DP. A training-testing approach to the molecular classification of resected non-small cell lung cancer. Clin Cancer Res 2003; 9: 4695-704.
[13] Hou J, Aerts J, den Hamer B, van Ijcken W, den Bakker M, Riegman P, van der Leest C, van der Spek P, Foekens JA, Hoogsteden HC, Grosveld F and Philipsen S. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS One 2010; 5: e10312.
[14] Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ and Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 2001; 98: 13790-5.
[15] Talbot SG, Estilo C, Maghami E, Sarkaria IS, Pham DK, O-charoenrat P, Socci ND, Ngai I, Carlson D, Ghossein R, Viale A, Park BJ, Rusch VW and Singh B. Gene expression profiling allows distinction between primary and metastatic squamous cell carcinomas in the lung. Cancer Res 2005; 65: 3063-71.
[16] Deng Y, He R, Zhang R, Gan B, Zhang Y, Chen G and Hu X. The expression of HOXA13 in lung adenocarcinoma and its clinical significance: a study based on the cancer genome atlas, oncomine and reverse transcription-quantitative polymerase chain reaction. Oncol Lett 2018; 15: 8556-8572.
[17] Xie ZC, Dang YW, Wei DM, Chen P, Tang RX, Huang Q, Liu JH and Luo DZ. Clinical significance and prospective molecular mechanism of MALAT1 in pancreatic cancer exploration: a comprehensive study based on the genechip, GEO, oncomine, and TCGA databases. Onco Targets Ther 2017; 10: 3991-4005.
[18] Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, Varambally S, Ghosh D and Chinnaiyan AM. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 2007; 9: 166-80.
[19] Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A and Chinnaiyan AM. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 2004; 6: 1-6.
[20] Wilson BJ and Giguère V. Identification of novel pathway partners of p68 and p72 RNA helicases through oncomine meta-analysis. BMC Genomics 2007; 8: 419.
[21] Hou GX, Liu P, Yang J and Wen S. Mining expression and prognosis of topoisomerase isoforms in non-small-cell lung cancer by using oncomine and Kaplan-Meier plotter. PLoS One 2017; 12: e0174515.
[22] Shin S, Kim Y, Chul Oh S, Yu N, Lee ST, Rak Choi J and Lee KA. Validation and optimization of the Ion torrent S5 XL sequencer and oncomine workflow for BRCA1 and BRCA2 genetic testing. Oncotarget 2017; 8: 34858-34866.
[23] Long T, Liu Z, Zhou X, Yu S, Tian H and Bao Y. Identification of differentially expressed genes and enriched pathways in lung cancer using bioinformatics analysis. Mol Med Rep 2019; 19: 2029-2040.