ECRONICON OPEN ACCESS

EC PULMONOLOGY AND RESPIRATORY MEDICINE

Research Article

Identification of Key Biomarkers and Analysis of Prognostic Values in Early-Stage Lung Adenocarcinoma: Evidence from Integrating Bioinformatics Approach with Clinical Indices

Cheng Jiang[1], Wenqian Ma[1], Liuliu Yang[1], You Zhou[1], Hui Li[1], Lang Guo[2], Dong Zhang[3] and Wei Zhang[1]*

[1]Department of Respiratory Medicine, The First Affiliated Hospital of Guangzhou University of Traditional Chinese Medicine, Guangzhou, China

[2]Department of Urology Surgery, Guangdong Provincial Hospital of Chinese Medicine, Guangzhou University of Chinese Medicine, Guangzhou, Guangdong Province, China

[3]Institute of Gastrointestinal Research, Guangzhou University of Chinese Medicine, Guangzhou, Guangdong Province, China

***Corresponding Author**: Wei Zhang, Department of Respiratory Medicine, The First Affiliated Hospital of Guangzhou University of Traditional Chinese Medicine, Guangzhou, China.

Received: August 29, 2019; **Published:** October 23, 2019

Abstract

Background: Lung cancer contributes more deaths worldwide compared with the other top three cancers together. So that identification of additional biomarkers that can assist in early diagnosis and “tailor” specific therapeutic protocols would vastly improve the death rate for this appalling cancer.

Methods: Three-gene expression datasets (GSE7670, GSE10072, GSE31547) were downloaded from the Gene Expression Omnibus repository. The differentially expressed genes filter of microarray data were analyzed via multiple open source R/Bioconductor software packages. We also performed functional and pathway enrichment analysis, protein-protein interaction network and module construction as well as gene expression level comparison. We then estimated and analyzed overall survival (OS) and hazard ratios of key biomarkers in terms of histology, stage and smoking history by using KM plotter in lung cancer.

Results: A total of 442 differentially expressed genes were ultimately obtained, including 123 upregulated genes and 319 downregulated genes. We identified six key genes with a high degree of connectivity in the PPI network, namely, *CCNB1, MAD2L1, CDK1, ZWINT, RRM2* and *TOP2A*. Prognostic analysis demonstrated that high expression of each key genes was significantly correlated to worse OS for Lung adenocarcinoma(LUAD) patients, while not for lung squamous cell carcinoma patients. Notably, high expression of these genes was associated with negative OS in clinical stage I LUAD patients, but not in stage II. In addition, only the increased mRNA expression of *CCNB1* and *MAD2L1* was related to worse OS in LUAD patients with smoking history.

Conclusion: Our bioinformatics analysis unveiled *CCNB1, MAD2L1, CDK1, ZWINT, RRM2* and *TOP2A* as putative key biomarkers, and they may have potentially used as independent factors for the early diagnosis, adjuvant therapy and accurate prognosis of LUAD patients.

Keywords: *Lung Adenocarcinoma; Bioinformatics; Differentially Expressed Genes; Key Biomarkers; Survival*

Introduction

Lung cancer ranks as the predominant cause of cancer-associated mortality in the United States, accounting for nearly 25% of all cancer deaths [1]. Approximately 85% of lung cancer cases are non-small cell lung carcinoma (NSCLC), of which LUAD is the mainly diagnosed histological subtype [2]. LUAD is more likely to occur among never smokers compared with lung squamous cell carcinoma (LUSC) [3]. Although treatment for lung cancer has shifted from the use of cytotoxic therapy to the current age of personalized treatment based on molecular alterations, the 5-year relative survival rate for this neoplasm is consistently low (around 15%). Owning to a lack of particular clinical symptoms, up to two-thirds of LUAD patients is typically diagnosed at an advanced stage, leaving little opportunity for effective treatment [4]. Thus, exploring novel key biomarkers could assist in early diagnosis and deliver specific treatment protocols for LUAD patients.

Development of putative biomarkers associated with the pathogenesis and prognosis of LUAD calls for a holistic understanding of the underlying bioinformatics [5]. In this regard, the accessibility of next-generation sequencing (NGS) in the last decade has resulted in the generation of high-throughput genomic profiling and multiplex genotyping that underpin the quantitative assessment of Transcriptomic profiling. Notably, The most significant driver mutations found in LUAD, such as sensitizing EGFR mutations, BRAF mutations as well as ALK and ROS1 rearrangements, have approved by the US Food and Drug Administration (FDA) [6]. However, LUAD is a highly heterogeneous disease that often harbors genetic mutations and deficiencies in tumor suppressor genes [7]. Furthermore, the majority of previous studies primarily focused on different microarray platforms and overlooked the batch effects that arise from technical variation between independent studies, all of which could significantly hamper downstream analyses [8-10]. Therefore, the identification of additional genes altered in LUAD is urgently needed to fulfill clinical requirements.

Herein, we first performed background normalization, batch effect correction, and differentially expressed genes (DEGs) filter of microarray data via multiple open source R /Bioconductor software packages. Then, analysis of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways of the DEGs were implemented. Subsequently, a protein-protein interaction (PPI) network and screening of active modules were constructed. In addition, We compared the gene expression level of key biomarkers in LUAD, LUSC tissues, and normal tissues via The Cancer Genome Atlas (TCGA) cohort, respectively. Finally, we estimated and analyzed overall survival (OS) and hazard ratios (HR) of key biomarkers in terms of histology, stage as well as smoking history by using KM plotter in lung cancer. Together, our study aimed to identify key biomarkers that may have potentially used to improve LUAD patient outcomes in the near future, paving the way for precision medicine.

Materials and Methods

Microarray data

Three gene expression datasets (GSE7670, GSE10072, GSE31547) were downloaded from Gene Expression Omnibus (GEO) repository by the National Center of Biotechnology Information using the keywords "microarray and lung adenocarcinoma". All included datasets were further screened as the following criteria: (1) the selected tissue samples were originated from Homo sapiens and contained LUAD and corresponding adjacent or normal tissues. (2) each dataset included at least 40 samples. (3) the experiments of all datasets were performed by GPL96 (Affymetrix Human Genome U133A Array) platform, doing this not only minimizes batch effects that result from different microarray platforms but also allows the annotation of the same set of genes with the same probes. The data provided by GEO is public and did not require the approval of a local ethics committee.

Integrated analysis of microarray datasets

We processed the microarray datasets using multiple open source R/Bioconductor software packages. First, affy package [11] was used to perform quality control of all of the microarray CEL files and the RMA method [12] was applied to the raw data for background correction, normalization, and probe-to-gene mapping. Next, the mean value was calculated using the aggregate function as the expression value of that particular gene when multiple probes corresponded to the same gene symbol. K-nearest neighbor (KNN) method was employed to supplement missing values when the expression value of the probe was absent. Then, using the Combat algorithm in sva package [13], we corrected for batch effects that arise from technical variation between independent studies. Subsequently, the DEGs of the samples between tumor and normal were identified via the limma package [14] by applying the following statistical criteria: (1) log 2 fold change ≥ 1; (2) adjusted P-value < 0.01, Such stringent cutoff thresholds generate only a handful of significant genes that distinguish tumors from tumor-free lung tissues. All DEGs were visualized in a volcano plot produced using the ggplot2 package. Finally, Disease Ontology Semantic and Enrichment analysis (DOES) package [15] was applied to verify the identification of DEGs association with lung diseases due to DO is an important annotation in translating molecular findings from high-throughput data to clinical relevance.

Functional enrichment analysis of DEGs

Gene ontology (GO) analysis, which included molecular function (MF), biological process (BP) and cellular component (CC), is increasingly applied for annotating genes and gene products and for identifying characteristic biological attributes of high-throughput genome or transcriptome data [16]. Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.kegg.jp/), which links genomic information with higher-order functional information, is a well-known database for biological interpretation of genome sequences and other high-throughput data [17]. DAVID (https://david.ncifcrf.gov/) is an essential online tool for high-throughput gene functional analysis, which provides the functionality to perform simultaneous GO and KEGG analysis for DEGs [18]. P < 0.05 was set as the cut-off criterion for significant enrichment. The results of the functional enrichment analysis of upregulated and downregulated genes were visualized via R/RStudio software.

Construction of the PPI network and screening of active modules

The Search Tool for the Retrieval of Interacting Genes (STRING) database (http://string-db.org/) is online software containing comprehensive interactions of lists of proteins and genes pertaining to homo sapiens [19]. Cytoscape (version 3.7.0) is an open-source tool for visualizing molecular interaction networks [20]. In the current study, DEGs were uploaded to STRING to build a PPI network and a combined score of ≥ 0.4 was used as the cut-off value. The Cytoscape plugin Molecular Complex Detection (MCODE) was applied to identify notable modules in this PPI network (Bader and Hogue, 2003) with degree cutoff = 2, node score cutoff = 0.2, k-core = 2, and max. depth = 100 [21]. Moreover, the Maximal Clique Centrality (MCC) method in Cytoscape plugin cytoHubba was also applied to identify notable module since MCC has a better performance on the precision of predicting essential proteins from the yeast PPI network among the 11 methods [22]. Subsequently, the enrichment analysis of the module was conducted by the clusterProfiler package in R [23].

Expression level analysis of hub genes

The Gene Expression Profiling Interactive Analysis (GEPIA, http://gepia.cancer-pku.cn/index.html), which applies a standard processing pipeline, is a web-based tool to provide key interactive and customizable functions based on the Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) data [24]. In the present study, we demonstrated the gene expression level of key genes in LUAD and LUSC tissues and normal tissues via GEPIA, Then the box plots were generated to visualize the relationship.

Survival analysis and hazard ratios estimation of hub genes

The correlation of key biomarkers expression with OS was analyzed using an online database, which was established using gene expression data and survival information of non-small cell lung cancer (NSCLC) patients downloaded from the Cancer Biomedical Informatics Grid (caBIG), GEO and TCGA repositories [25]. In addition, the clinical data of NSCLC patients contained histology, stage, grade, gender, and smoking history, and treatment groups comprise surgery, chemotherapy, and radiotherapy. In this study, array quality control was selected "exclude biased arrays". The cut-off points of individual key biomarkers expression and other clinicopathological parameters including histology subtypes, smoking history, clinical stages were assessed according to their median mRNA levels among the selected lung cancer samples via the Kaplan-Meier plotter (http://kmplot.com/analysis/index.php?p = service&cancer = lung). The Log-rank p-value and hazard ratio (HR) with 95% confidence intervals (CI) were calculated and displayed automatically on the webpage. HR and 95% CI > 1 were considered as a poor prognostic indicator of LUAD and P value < 0.01 was considered statistically significant to reduce the false positive rate.

Results

Data preprocessing and identification of DEGs

There were 116 LUAD samples and 97 normal samples in this study (Table 1). The effect of sva package for removing batch effects and other unwanted variation was explicitly demonstrated on three preprocessed datasets (Figures 1), 442 DEGs were ultimately obtained in LUAD samples compared with normal samples, including 123 upregulated and 319 downregulated genes (Figure 2A). DOSE package analysis showed that the DEGs were closely related to lung disease and non-small cell lung carcinoma (Figure 2B), thus verified our previous finding.

GEO	PMID	Platform	Normal	Tumor	Reference
GSE7670	17540040	GPL96[HG–U133A] Affymetrix Human Genome U133A Array	28	28	Su LJ., *et al*. 2007
GSE10072	18297132	GPL96[HG–U133A] Affymetrix Human Genome U133A Array	49	58	Landi M., *et al*. 2008
GSE31547	15701842	GPL96[HG–U133A] Affymetrix Human Genome U133A Array	20	30	Dobbin KK., *et al*. 2005

***Table 1:** The gene expression profile data characteristics.*

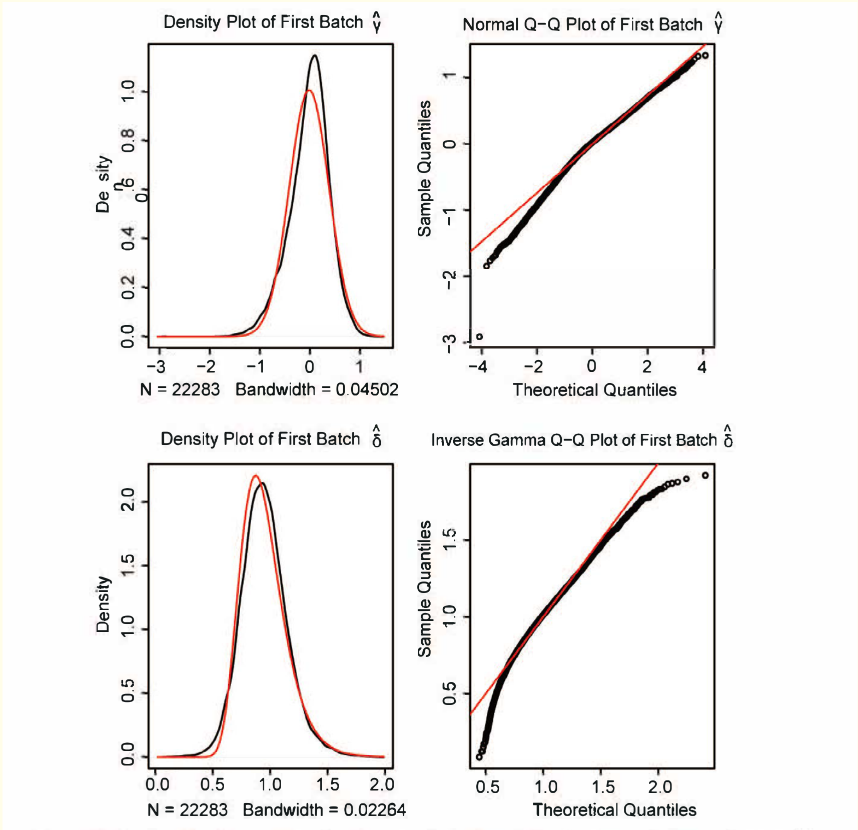


***Figure 1:** The batch effects that arise from technical variation between independent studies were detected using the Combat algorithm in sva package.*

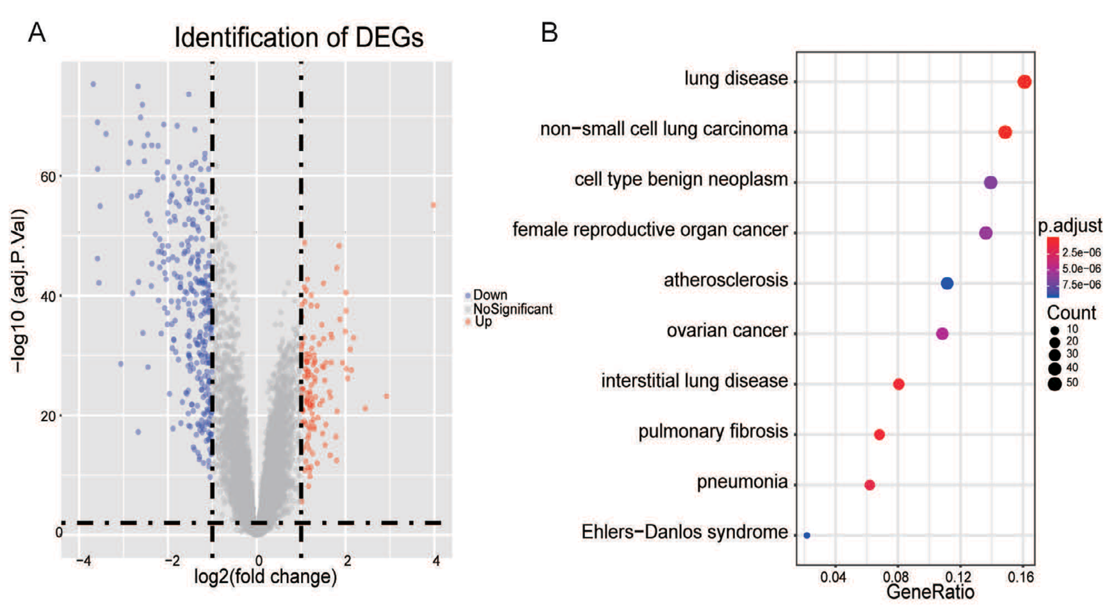


Figure 2: *Identification and verification of DEGs. (A) Volcano Plot visualizing the DEGs. The vertical lines demark the fold-change values. The right vertical line corresponds to log2FC > 1 changes, while the left vertical line corresponds to log2FC < -1 changes. The horizontal line marks adjusted P-value < 0.01; (B) DOES package was applied to annotate the identification of DEGs. DEGs: differentially expressed genes; DOES: Disease Ontology Semantic and Enrichment analysis.*

Functional enrichment analysis of DEGs

For a thorough understanding of the DEGs, GO terms and KEGG pathways were applied to analyze upregulated and downregulated DEGs, respectively. For BP, GO analysis showed that the upregulated DEGs were significantly associated with collagen catabolic process, extracellular matrix organization, and extracellular structure organization, while the downregulated DEGs were associated with regulation of inflammatory response, vasculature development and angiogenesis (Figure 3BP). For MF, the upregulated DEGs were mainly enriched in extracellular matrix structural constituent, protease binding, and platelet-derived growth factor binding, while the downregulated DEGs enriched in glycosaminoglycan binding, enzyme inhibitor activity and growth factor binding (Figure 3MF). For CC, the upregulated DEGs were chiefly enriched in fibrillar collagen trimer, banded collagen fibril, and proteinaceous extracellular matrix, and downregulated DEGs enriched in the extracellular matrix, proteinaceous extracellular matrix and lamellar body (Figure 3CC). In addition. KEGG pathway analysis revealed that the upregulated DEGs were mainly enriched in Protein digestion and absorption and p53 signaling pathway, while the downregulated DEGs were mainly enriched in Complement and coagulation cascades and Fluid shear stress and atherosclerosis (Table 2).

PPI network construction and key genes screening

we acquired the PPI network of a total of 1484 protein pairs corresponding to 441 nodes by mapping into STRING. Our results showed that most of the DEGs present in notable modules were upregulated genes rather than downregulated genes. The most significant module with score > 5 was obtained by MCODE (Supplementary Figure S1). Then, we also used the MCC algorithm in CytoHubba plugin to search and explore the PPI network and the top six genes among the two subsets regard as the key genes (Figure 4A). *CCNB1, MAD2L1, CDK1, ZWINT, RRM2 and TOP2A* were identified as key genes with higher node degrees both in MCODE module and MCC algorithm (Table 3). Furthermore, KEGG pathway enrichment analysis showed that module 1 was mainly associated with the mitotic cell cycle pathway, Oocyte meiosis and p53 signaling pathway (Figure 4B).

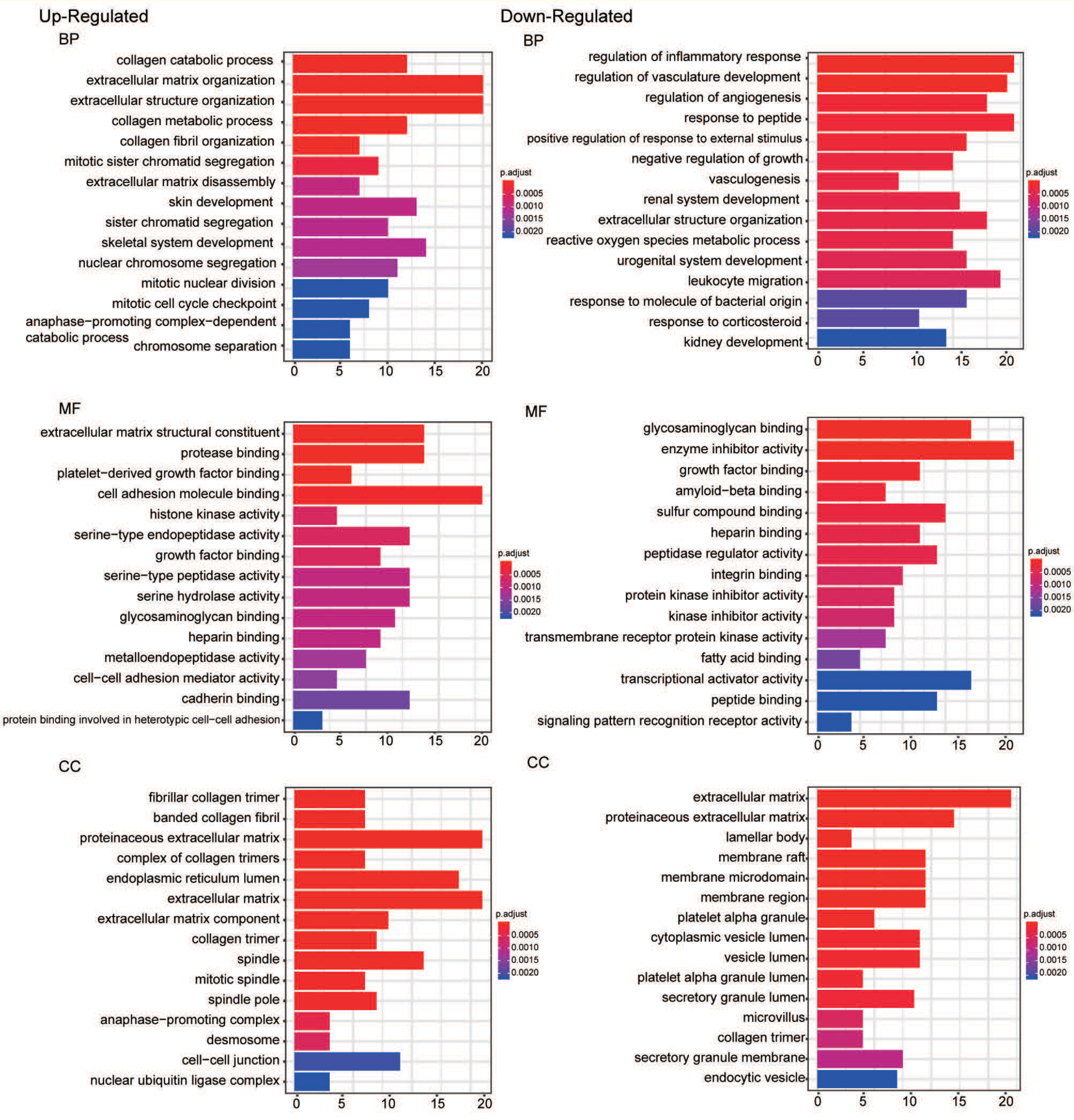
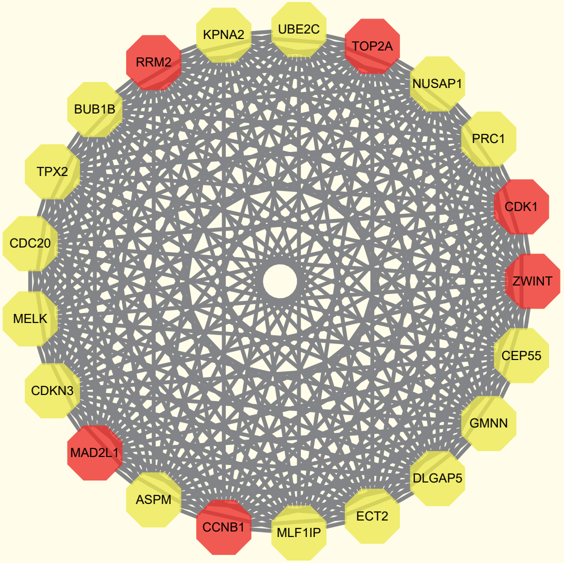


Figure 3: *Top 15 Gene Ontology analysis of up-regulated and down-regulated differentially expressed genes associated with LUAD. The left represents up-regulated DEGs, while the right represents down-regulated DEGs. BP: Biological Processes; MF: Molecular Function; CC: Cellular Component.*



Supplementary Figure S1: *MCODE was applied to identify notable module in this PPI network with degree cutoff = 2, node score cutoff= 0.2, k-core = 2, and max. depth = 100. MCODE: Molecular Complex Detection.*

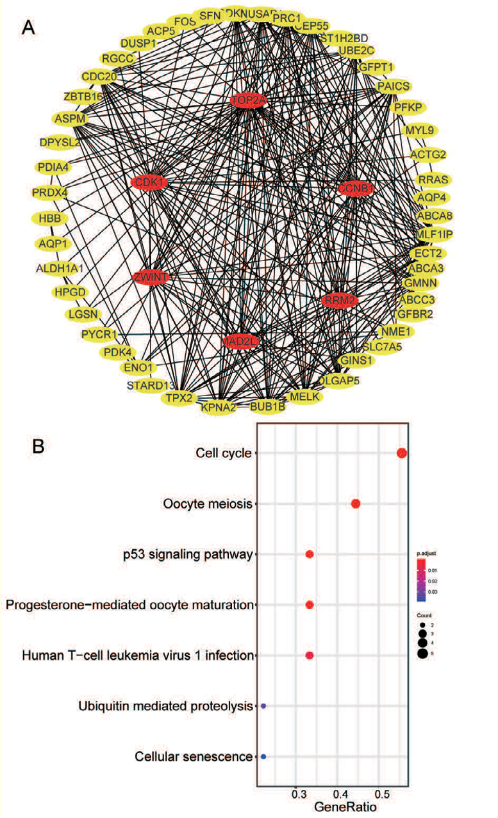


Figure 4: *(A) MCC algorithm in CytoHubba plugin to explore key genes; (B) pathway analysis of Module 1.*

ID	Description	P.adjust	Genes
The top 4 enriched KEGG pathways of upregulated genes			
hsa04974	Protein digestion and absorption	7.72E−4	*COL10A1, COL11A1, COL1A1, COL3A1, COL5A1, COL5A2, COL1A2, KCNN4*
hsa04115	P53 signaling pathway	2.42E−3	*RRM2, IGFBP3, CCNB1, CDK1, SFN, PERP*
hsa04512	ECM−receptor interaction	2.41E−2	*SPP1, THBS2, COL1A1, COMP, COL1A2*
hsa04110	Cell cycle	2.43E−2	*CDC20, CCNB1, CDK1, BUB1B, SFN, MAD2L1*
The top 4 enriched KEGG pathways of downregulated genes			
hsa04610	Complement and coagulation cascades	2.04E−07	*C14ORF13, C1QA, C1QB, SERPING1, C4BPA, C7, CLU, CFD, PROS1, THBD, VWF, VSIG4, CPB2*
hsa05418	Fluid shear stress and atherosclerosis	1.94E−2	*CAV1, CAV2, CDH5, DUSP1, EDN1, FOS, NCF2, PECAM1, SELE, THBD, KLF2*
hsa04670	Leukocyte transendothelial migration	3.62E−2	*CDH5, NCF2, JAM2, PECAM1, CLDN5, CLDN18, JAM3, MYL9, CXCL12*
hsa05020	Prion diseases	3.62E−2	*C1QA, C1QB, C7, EGR1, IL6*

Table 2: *The top 4 enriched KEGG pathways of differentially expressed genes associated with LUAD.*

ID=identification number of KEGG pathway. Description represents the name of KEGG pathway.

Gene	Type	MCODE		MCC	
		Degree	MCODE Score	Rank	Score
TOP2A	Up	31	20	1	2.43E-18
CCNB1	Up	25	20	1	2.43E-18
CDK1	Up	24	20	1	2.43E-18
RRM2	Up	23	20	1	2.43E-18
ZWINT	Up	22	20	1	2.43E-18
MAD2L1	Up	22	20	1	2.43E-18

Table 3: *Hub genes with high degree of connectivity.*

MCODE: Molecular Complex Detection; MCC: Maximal Clique Centrality.

Expression level and prognostic values of key genes

GEPIA revealed that key genes expression were markedly higher both in LUAD and LUSC tissues as compared with normal tissues (P < 0.01) based on 969 tumors and 109 normal samples from the TCGA databases (Figure 5). The prognostic values of each key genes in LUAD were examined in Kaplan Meier-plotter. The Affymetrix IDs are valid: 214710_s_at (*CCNB1*), 210559_s_at (*CDK1*), 1554768_a_at (*MAD2L1*), 209773_s_at (*RRM2*), 201292_at (*TOP2A*), 204026_s_at (*ZWINT*). It was found that high expression of *CCNB1* [HR 2.04 (1.6-2.61), P = 5.4e-09], *CDK1* [HR 2.5 (1.94-3.21), P = 1.7e-13], *MAD2L1* [HR2.41 (1.86-3.14), P = 1.3e-11], *RRM2* [HR1.95 (1.53-2.49), P = 3.2e-08], *TOP2A* [HR1.76 (1.38-2.23), P = 3e-06], *ZWINT* [HR1.35 (1.07-1.71), P = 0.011] was significantly correlated to worse OS for LUAD patients (n = 720) (Figure 6). However, higher mRNA expressions of key genes (*CCNB1* [HR1 (0.79-1.27), P = 0.99], *CDK1* [HR0.92 (0.73-1.17), P = 0.51], *MAD2L1* [HR1.29 (0.94-1.76), P = 0.12], *RRM2* [HR0.98 (0.77-1.24), P = 0.87], *TOP2A* [HR0.97(0.77-1.23), P = 0.8], *ZWINT* [HR0.99 (0.78-1.25), P = 0.91])were not associated with OS for LUSC patients (n = 720) (Table 4).

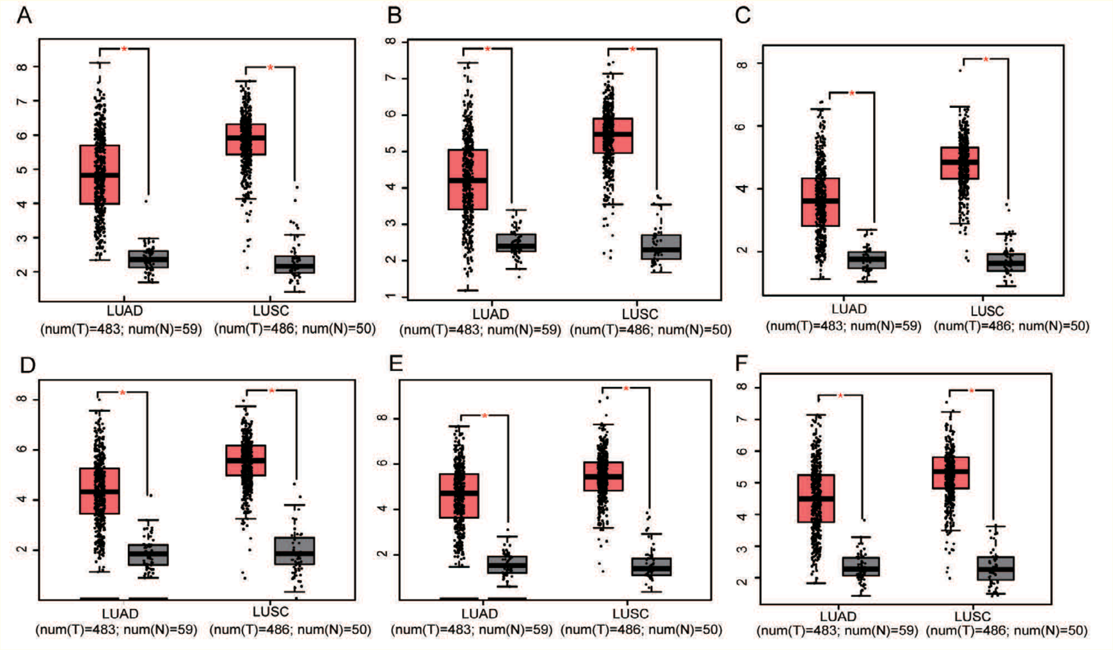


Figure 5: *Analysis of six hub genes expression level both in LUAD and LUSC tissues as compared with normal tissues (P < 0.01) based on 969 tumors and 109 normal samples from the TCGA databases. The red and gray boxes represent cancer and normal tissues, respectively. (A) CCNB1; (B) CDK1; (C) MAD2L1; (D) RRM2; (E) TOP2A;(F) ZWINT; LUAD: Lung Adenocarcinoma; LUSC: Lung Squamous Cell Carcinomas.*

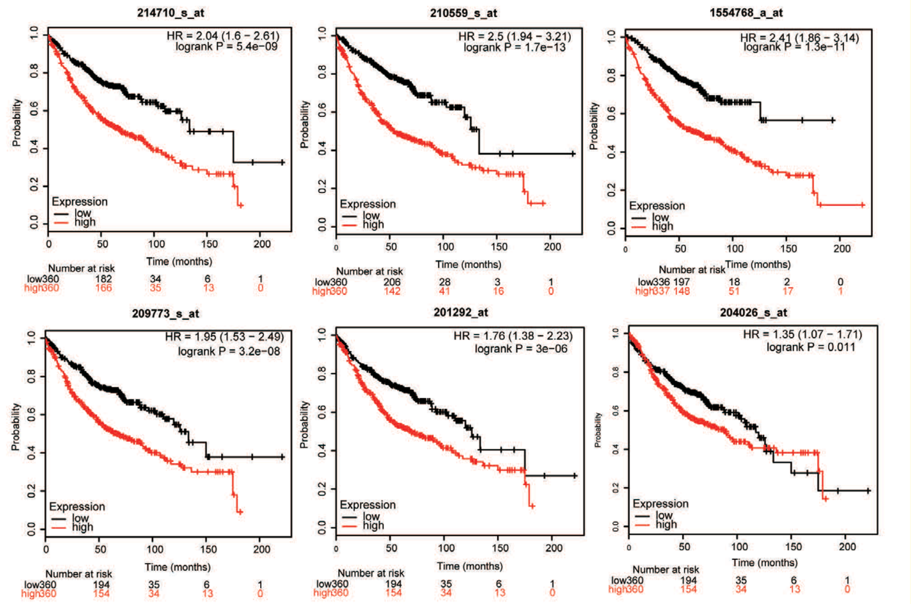


Figure 6: *Prognostic roles of six hub genes in the LUAD patients. Survival curves are plotted for LUAD cancer patients. The Affymetrix IDs are valid: 214710_s_at (CCNB1); 210559_s_at (CDK1); 1554768_a_at (MAD2L1); 209773_s_at (RRM2); 201292_at (TOP2A); 204026_s_at (ZWINT).*

Biomarkers	Histology	Cases	HR (95% CI)	P value
CCNB1	LUSC	524	1 (0.79 - 1.27)	0.99
MAD2L1	LUSC	524	1.29 (0.94 - 1.76)	0.12
CDK1	LUSC	524	0.92 (0.73 - 1.17)	0.51
ZWINT	LUSC	524	0.99 (0.78 - 1.25)	0.91
RRM2	LUSC	524	0.98 (0.77 - 1.24)	0.87
TOP2A	LUSC	524	0.97(0.77 - 1.23)	0.8

Table 4: *Correlation of key genes expression with overall survival in LUSC patients.*

For further assess the relationship between key genes and other clinicopathological parameters, we investigated the correlation with the patients' smoking status (Table 5) and clinical stages (Table 6). Table 5 showed that high mRNA expression of *CCNB1* [HR 2.35 (1.42 - 3.89), P = 0.00059], *MAD2L1* [HR 1.87(1.14 - 3.08), P = 0.012] correlated with worse OS in LUAD patients with smoking history(n = 246),but not with nonsmoking history (n = 143). In addition, *CDK1, ZWINT, RRM2* and *TOP2A* expression were not linked to OS in patients with and without smoking history. From table 6, elevated *CCNB1* [HR2.85 (1.84 - 4.42), P = 1e-06], *MAD2L1* [HR4.88 (2.92 - 8.17), P = 2.5e-11], *CDK1* [HR 2.16 (1.42 - 3.28), P = 0.00022], *ZWINT* [HR 1.82 (1.21 - 2.74), P = 0.0037], *RRM2* [HR1.95 (1.3 - 2.94), P = 0.0011], and *TOP2A* [HR1.88 (1.25 -2.82), P = 0.002] mRNA expression was correlated with a worse OS in clinical stage I LUAD patients (n = 346). However, the expression of these mRNA had no effect on OS in clinical stage II LUAD patients (n = 136).

Biomarkers	Smoking status	Cases	HR (95% CI)	P value
CCNB1	Smoked	246	2.35 (1.42 - 3.89)	0.00059*
	Never smoked	143	1.37 (0.61 - 3.09)	0.45
MAD2L1	Smoked	246	1.87 (1.14 - 3.08)	0.012*
	Never smoked	143	1.23 (0.54 - 2.79)	0.62
CDK1	Smoked	246	1.16 (0.73 - 1.85)	0.53
	Never smoked	143	1.76 (0.76 - 4.03)	0.18
ZWINT	Smoked	246	1.31 (0.82 - 2.09)	0.26
	Never smoked	143	1.73 (0.76 - 3.96)	0.19
RRM2	Smoked	246	1.21 (0.54 - 2.71)	0.64
	Never smoked	143	1.21 (0.54 - 2.71)	0.64
TOP2A	Smoked	246	1.48 (0.92 - 2.37)	0.1
	Never smoked	143	2.02 (0.86 - 4.74)	0.098

Table 5: *Correlation of key genes expression with smoking status of LUAD patients.*

**: P < 0.05.*

Biomarkers	Clinical stages	Cases	HR (95% CI)	P value
CCNB1	I	346	2.85 (1.84 - 4.42)	1e - 06*
	II	136	1.08 (0.67 - 1.74)	0.76
MAD2L1	I	346	4.88 (2.92 - 8.17)	2.5e - 11*
	II	136	1.22 (0.72 - 2.07)	0.47
CDK1	I	346	2.16 (1.42 - 3.28)	0.00022*
	II	136	1.57 (0.97 - 2.54)	0.065

ZWINT	I	346	1.82 (1.21 - 2.74)	0.0037*
	II	136	0.74 (0.46 - 1.2)	0.22
RRM2	I	346	1.95 (1.30 - 2.94)	0.0011*
	II	136	1.10 (0.68 - 1.78)	0.7
TOP2A	I	346	1.88 (1.25 - 2.82)	0.002*
	II	136	1.08 (0.67 - 1.74)	0.76

Table 6: *Correlation of key genes expression with OS in different clinical stage LUAD patients.*

**: P < 0.05.*

Discussion and Conclusion

Lung cancer contributes more deaths worldwide compared with the other top three cancers together [1]. Biomarkers that can aid to early detection and "tailor" specific therapeutic regimen would tremendously ameliorate the death rate for this devastating cancer. In the present study, we incorporated into 116 LUAD and corresponding 97 normal samples from the three microarray databases. A total of 442 DEGs were ultimately obtained, including 123 upregulated genes and 319 downregulated genes. Function enrichment analysis suggested that these gene signatures were significantly associated with the carcinogenesis of LUAD, such as extracellular matrix organization, p53 signaling pathway as well as mitotic cell cycle, which is in line with the previous studies [26-28]. We also identified six key genes with a high degree of connectivity in the PPI network, namely, *CCNB1, MAD2L1, CDK1, ZWINT, RRM2* and *TOP2A* and uniformly all of them were upregulated genes in LUAD. GEPIA revealed that these genes expression level were markedly higher both in LUAD and LUSC tissues as compared with normal tissues based on TCGA cohort. Prognostic analysis demonstrated that high expression of each key genes was significantly correlated to worse OS for LUAD patients, while not for LUSC patients. Notably, high expression of these genes was associated with negative OS in clinical stage I LUAD patients, but not in stage II. In addition, only the increased mRNA expression of *CCNB1* and *MAD2L1* was related to worse OS in LUAD patients with smoking history. Collectively, our bioinformatics analysis unveiled *CCNB1, MAD2L1, CDK1, ZWINT, RRM2* and *TOP2A* as key biomarkers, and they may be crucial in the development and prognosis of LUAD.

Function enrichment analysis provides crucial complementary information into the collective biological characters of a panel of genes. Extracellular matrix (ECM), which could influence cellular events both at the physical and molecular level, plays a crucial function in tumor progression [29]. A recent study suggested that elevated *ZEB1* causes LOXL2-mediated collagen deposition in the ECM to actuate lung cancer progression [30]. The p53 pathway is a pivotal factor that serves as an internal sentinel by preventing mutations caused by DNA damage or cellular stress [31]. Mutations in p53 are related to genomic instability and an increased susceptivity to cancer, and it is found that up to 50% of all cancers involve p53-inactivating mutations [32]. Unfortunately, p53 research has not yet generated extensive applications on cancer supervision and therapy due to the complexity and versatility in their biological effects [33]. The cell cycle consists of four sequential phases and dysregulation of its engine may trigger the cell proliferation that leads to cancer [34]. Analysis of the cell cycle machinery may provide a promising diagnostic and therapeutic interventions target in cancer as it locates downstream at the integration point of intricate oncogenic signaling networks [35].

Despite informative, pathways analyses do not provide a holistic and systematic view of a biological response or disease process. At present, a multitude of cell cycle associated genes has been demonstrated to be involved in the initiation and progression of lung cancer. Our study identified six key biomarkers via PPI network construction, of which *TOP2A* harbored higher node degree. *TOP2A*, an isoform in the topoisomerase II family, is cell cycle-dependent and mediates the topologic states of DNA during transcription [36]. Elevated expression of *TOP2A* was confirmed to be associated with the development and progression of NSCLC [37]. In addition, Labbé DP., *et al*. [38] found that *TOP2A* could be used as a marker for early detection of a subset of prostate cancer patients with aggressive potential. *CDK1*, a highly conserved small protein, is a critical determinant of mitotic progression. Aberrant activation of *CDK1* is involved in the unbounded proliferation and apoptosis of ovarian cancer cells since the dysregulations of the upstream signaling pathway [39]. High

CCNB1 expression is also detected in NSCLC and generally associated with negative prognosis for patients with early-stage NSCLC [40]. *CCNB1* and its associated partner *CDK1* provide a host of timely biofuel for cell cycle such as G2/M transition, whereas overexpression in the tumor may lead to uncontrolled cell growth that attenuates the efficacy of anti-cancer therapy [41]. The attenuated function of *MAD2L1* might lead to reduced spindle checkpoint during mitosis that confers susceptibility to the development of lung cancer [42]. *RRM2* is a member of Ribonucleotide reductase and frequently overexpressed in NSCLC patients, thus, it plays an important role in tumor progression [43]. *ZWINT* overexpression is also found in various malignant cancers, including breast [44] and ovarian cancers [45].

Comprehensive analysis of expression level and prognosis of key markers provides a better understanding of heterogeneity and complexity of LUAD on molecular biology. Our results demonstrate that over-expression of each key markers is an independent poor prognostic factor in early-stage LUAD, but not in LUSC. The different prognostic roles exhibited by key markers in LUAD and LUSC underscores the heterogeneity among these two histologic subtypes. Nicotine, one of the carcinogens, can motivate several signaling pathways that can induce the mutation, disrupt cell proliferation, apoptosis, angiogenesis, and promote a tumor-supporting microenvironment [46,47]. However, approximately 25% of lung cancer cases worldwide are not associated with tobacco smoking and most of them are LUAD subtypes [48,49]. In this study, we find that high expression of *CCNB1, MAD2L1,* but not that *CDK1, ZWINT, RRM2* and *TOP2A* are correlated with the smoking status of LUAD patients.

At present, several studies have been conducted to explore key genes correlated with diagnosis and prognosis of NSCLC, creating new opportunities for precision medicine. Tang., *et al.* [50] identified nine genes from one GEO dataset via bioinformatics analysis including identification of DEGs using R, GO enrichment analysis, PPI network construction, survival analysis. Xiao., *et al.* [51] explored 195 DEGs by analyzing 4 GEO datasets from different platforms and identified 5 hub genes associated with poor OS based on the PPI network analysis. Wang., *et al.* [52] determined *CCND1* mRNA as the putative prognostic biomarkers using GSEA analysis and constructed microRNA-mRNA regulatory networks in NSCLC. Piao., *et al.* [53] targeted 16 hub genes and further found 14 of them were related to prognosis of NSCLC patients by integrated bioinformatics approach incorporating GO and KEGG analysis, PPI network development and OS analysis. Compared to previous studies, the merits of the present work are primarily embodied in the following points: First, our study concentrates on the same chip platform and utilize sva package to remove batch effects arising from technical variation between independent studies. Secondly, LUAD is a highly heterogeneous disease, while most previous studies mainly focused on NSCLC patients rather than LUAD patients. Our results demonstrate that over-expression of each key markers is an independent negative prognostic factor in early-stage LUAD, but not in LUSC. Finally, we validate the gene expression level of key biomarkers in LUAD, LUSC tissues, and normal tissues via the TCGA cohort and analysis of prognostic values with various clinicopathologic characteristics. However, our study was mainly focused on the integrating bioinformatics approach with clinical indices, further experiments are needed to validate the candidate genes which we disclosed in LUAD.

Acknowledgements

We are grateful to the researchers who provided their data for this analysis, and it is our pleasure to acknowledge their contributions.

Bibliography

1. Siegel RL., *et al*. “Cancer statistics, 2018”. *CA: A Cancer Journal for Clinicians* 68.1 (2018): 7-30.

2. Travis WD., *et al*. “The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification”. *Journal of Thoracic Oncology* 10.9 (2015): 1243-1260.

3. Khuder SA. “Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis”. *Lung Cancer* 31.2-3 (2001): 139-148.

4. Goss GD and Spaans JN. “Epidermal Growth Factor Receptor Inhibition in the Management of Squamous Cell Carcinoma of the Lung”. *Oncologist* 21.2 (2016): 205-213.

5. Hurgobin B., *et al*. “Insights into respiratory disease through bioinformatics”. *Respirology* 23.12 (2018): 1117-1126.

6. Ahmadzada T., *et al*. “An Update on Predictive Biomarkers for Treatment Selection in Non-Small Cell Lung Cancer”. *Journal of Clinical Medicine* 7.6 (2018): 153.

7. Imielinski M., *et al*. “Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing”. *Cell* 150.6 (2012): 1107-1120.

8. Goh WWB., *et al*. “Why Batch Effects Matter in Omics Data, and How to Avoid Them”. *Trends in Biotechnology* 35.6 (2017): 498-507.

9. Nyamundanda G., *et al*. “A Novel Statistical Method to Diagnose, Quantify and Correct Batch Effects in Genomic Studies”. *Scientific Reports* 7.1 (2017): 10849.

10. Lim SB., *et al*. “A merged lung cancer transcriptome dataset for clinical predictive modeling”. *Scientific Data* 5 (2018): 180136.

11. Gautier L., *et al*. “affy--analysis of Affymetrix GeneChip data at the probe level”. *Bioinformatics* 20.3 (2004): 307-315.

12. McCall MN., *et al*. “Frozen robust multiarray analysis for Affymetrix Exon and Gene ST arrays”. *Bioinformatics* 28.23 (2012): 3153-3154.

13. Leek JT., *et al*. “The sva package for removing batch effects and other unwanted variation in high-throughput experiments”. *Bioinformatics* 28.6 (2012): 882-883.

14. Ritchie ME., *et al*. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. *Nucleic Acids Research* 43.7 (2015): e47.

15. Yu G., *et al*. “DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis”. *Bioinformatics* 31.4 (2015): 608-609.

16. Gene Ontology Consortium. “Gene Ontology Consortium: going forward”. *Nucleic Acids Research* 43 (2015): D1049-D1056.

17. Kanehisa M., *et al*. “New approach for understanding genome variations in KEGG”. *Nucleic Acids Research* (2019).

18. Huang DW., *et al*. “DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists”. *Nucleic Acids Research* 35 (2007): W169-W175.

19. Szklarczyk D., *et al*. “STRING v10: protein-protein interaction networks, integrated over the tree of life”. *Nucleic Acids Research* 43 (2015): D447-D452.

20. Shannon P., *et al*. “Cytoscape: a software environment for integrated models of biomolecular interaction networks”. *Genome Research* 13.11 (2003): 2498-2504.

21. Bader GD and Hogue CW. “An automated method for finding molecular complexes in large protein interaction networks”. *BMC Bioinformatics* 4 (2003): 2.

22. Chin CH., *et al*. “cytoHubba: identifying hub objects and sub-networks from complex interactome”. *BMC Systems Biology* 8.4 (2014): S11.

23. Yu G., *et al*. “clusterProfiler: an R package for comparing biological themes among gene clusters”. *OMICS* 16.5 (2012): 284-287.

24. Tang Z., *et al*. “GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses”. *Nucleic Acids Research* 45.W1 (2017): W98-W102.

25. Gyorffy B., *et al*. "Online Survival Analysis Software to Assess the Prognostic Value of Biomarkers Using Transcriptomic Data in Non-Small-Cell Lung Cancer". *PLoS ONE* 8.12 (2013): e82241.

26. Shi YX., *et al*. "Prognostic and predictive values of CDK1 and MAD2L1 in lung adenocarcinoma". *Oncotarget* 7.51 (2016): 85235-85243.

27. Ni M., *et al*. "Identification of Candidate Biomarkers Correlated With the Pathogenesis and Prognosis of Non-small Cell Lung Cancer via Integrated Bioinformatics Analysis". *Frontiers in Genetics* 9 (2018): 469.

28. Yang G., *et al*. "Identification of genes and analysis of prognostic values in nonsmoking females with non-small cell lung carcinoma by bioinformatics analyses". *Cancer Management and Research* (2018).

29. Multhaupt HA., *et al*. "Extracellular matrix component signaling in cancer". *Advanced Drug Delivery Reviews* 97 (2016): 28-40.

30. Peng DH., *et al*. "ZEB1 induces LOXL2-mediated collagen stabilization and deposition in the extracellular matrix to drive lung cancer invasion and metastasis". *Oncogene* 36.14 (2017): 1925-1938.

31. Vousden KH and Lane DP. "p53 in health and disease". *Nature Reviews Molecular Cell Biology* 8.4 (2007): 275-283.

32. Stegh AH. "Targeting the p53 signaling pathway in cancer therapy - the promises, challenges and perils". *Expert Opinion on Therapeutic Targets* 16.1 (2012): 67-83.

33. Olivier M., *et al*. "TP53 mutations in human cancers: origins, consequences, and clinical use". *Cold Spring Harbor Perspectives in Biology* 2.1 (2010): a001008.

34. Champeris Tsaniras S., *et al*. "Licensing of DNA replication, cancer, pluripotency and differentiation: an interlinked world?" *Seminars in Cell and Developmental Biology* 30 (2014): 174-180.

35. Williams GH and Stoeber K. "The cell cycle and cancer". *Journal of Pathology* 226.2 (2012): 352-364.

36. Li Y., *et al*. "EGCG regulates the cross-talk between JWA and topoisomerase IIα in non-small-cell lung cancer (NSCLC) cells". *Scientific Reports* 5 (2015): 11009.

37. Hou GX., *et al*. "Mining expression and prognosis of topoisomerase isoforms in non-small-cell lung cancer by using Oncomine and Kaplan-Meier plotter". *PLoS One* 12.3 (2017): e0174515.

38. Labbé DP., *et al*. "TOP2A and EZH2 Provide Early Detection of an Aggressive Prostate Cancer Subgroup". *Clinical Cancer Research* 23.22 (2017): 7072-7083.

39. Zhang R., *et al*. "The aberrant upstream pathway regulations of CDK1 protein were implicated in the proliferation and apoptosis of ovarian cancer cells". *Journal of Ovarian Research* 10.1 (2017): 60.

40. Soria JC., *et al*. "Overexpression of cyclin B1 in early-stage non-small cell lung cancer and its clinical implication". *Cancer Research* 60.15 (2000): 4000-4004.

41. Xie B., *et al*. "Cyclin B1/CDK1-regulated mitochondrial bioenergetics in cell cycle progression and tumor resistance". *Cancer Letters* 443 (2019): 56-66.

42. Guo Y., *et al*. "Functional evaluation of missense variations in the human MAD1L1 and MAD2L1 genes and their impact on susceptibility to lung cancer". *Journal of Medical Genetics* 47.9 (2010): 616-622.

43. Rahman MA., *et al*. "RRM2 regulates Bcl-2 in head and neck and lung cancers: a potential target for cancer therapy". *Clinical Cancer Research* 19.13 (2013): 3416-3428.

44. Woo Seo D., *et al*. "Zwint-1 is required for spindle assembly checkpoint function and kinetochore-microtubule attachment during oocyte meiosis". *Scientific Reports* 5 (2015): 15431.

45. Wu X., *et al*. "Nucleolar and spindle associated protein 1 promotes the aggressiveness of astrocytoma by activating the Hedgehog signaling pathway". *Journal of Experimental and Clinical Cancer Research* 36.1 (2017): 127.

46. Cardinale A., *et al*. "Nicotine: specific role in angiogenesis, proliferation and apoptosis". *Critical Reviews in Toxicology* 42.1 (2012): 68-89.

47. Grando SA. "Connections of nicotine to cancer". *Nature Reviews Cancer* 14.6 (2014): 419-429.

48. Sun S., *et al*. "Lung cancer in never smokers--a different disease". *Nature Reviews Cancer* 7.10 (2007): 778-790.

49. Toh CK., *et al*. "Never-smokers with lung cancer: epidemiologic evidence of a distinct disease entity". *Journal of Clinical Oncology* 24.15 (2006): 2245-2251.

50. Tang Q., *et al*. "Hub genes and key pathways of non-small lung cancer identified using bioinformatics". *Oncology Letters* 16.2 (2018): 2344-2354.

51. Xiao Y., *et al*. "Identification of key differentially expressed genes associated with non-small cell lung cancer by bioinformatics analyses". *Molecular Medicine Reports* 17.5 (2018): 6379-6386.

52. Wang J., *et al*. "Identification of key genes and construction of microRNA-mRNA regulatory networks in non-small cell lung cancer". *Cancer Genetics* (2018).

53. Piao J., *et al*. "Target gene screening and evaluation of prognostic values in non-small cell lung cancers by bioinformatics analysis". *Gene* (2018).

Volume 8 Issue 11 November 2019