Bioinformatics and functional analyses of key genes in smoking-associated lung adenocarcinoma

DAJIE ZHOU[1,2], YILIN SUN[3], YANFEI JIA[1], DUANRUI LIU[1], JING WANG[1], XIAOWEI CHEN[1], YUJIE ZHANG[2] and XIAOLI MA[1]

[1]Central Laboratory, Jinan Central Hospital Affiliated to Shandong University, Jinan, Shandong 250013; [2]Department of Medical Laboratory, Weifang Medical University, Weifang, Shandong 261053; [3]College of Science, Northwest A&F University, Yangling, Shaanxi 712100, P.R. China

Received February 1, 2019; Accepted July 12, 2019

DOI: 10.3892/ol.2019.10733

Abstract. Smoking is one of the most important factors associated with the development of lung cancer. However, the signaling pathways and driver genes in smoking-associated lung adenocarcinoma remain unknown. The present study analyzed 433 samples of smoking-associated lung adenocarcinoma and 75 samples of non-smoking lung adenocarcinoma from the Cancer Genome Atlas database. Gene Ontology (GO) analysis was performed using the Database for Annotation, Visualization and Integrated Discovery and the ggplot2 R/Bioconductor package. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was performed using the R packages RSQLite and org.Hs.eg.db. Multivariate Cox regression analysis was performed to screen factors associated with patient survival. Kaplan-Meier and receiver operating characteristic curves were used to analyze the potential clinical significance of the identified biomarkers as molecular prognostic markers for the five-year overall survival time. A total of 373 differentially expressed genes (DEGs; |log2-fold change|≥2.0 and P<0.01) were identified, of which 71 were downregulated and 302 were upregulated. These DEGs were associated with 28 significant GO functions and 11 significant KEGG pathways (false discovery rate <0.05). Two hundred thirty-eight proteins were associated with the 373 differentially expressed genes, and a protein-protein interaction network was constructed. Multivariate regression analysis revealed that 7 mRNAs, cytochrome P450 family 17 subfamily A member 1, PKHD1 like 1, retinoid isomerohydrolase RPE65, neurotensin receptor 1, fetuin B, insulin-like growth factor binding protein 1 and glucose-6-phosphatase catalytic subunit, significantly distinguished between non-smoking and smoking-associated adenocarcinomas. Kaplan-Meier analysis demonstrated that patients in the 7 mRNAs-high-risk group had a significantly worse prognosis than those of the low-risk group. The data obtained in the current study suggested that these genes may serve as potential novel prognostic biomarkers of smoking-associated lung adenocarcinoma.

Correspondence to: Professor Xiaoli Ma, Central Laboratory, Jinan Central Hospital Affiliated to Shandong University, 105 Jiefang Road, Jinan, Shandong 250013, P.R. China
E-mail: mxl7125@126.com

Key words: differentially expressed genes, prognostic value, smoking, lung adenocarcinoma, bioinformatics analysis

Introduction

Lung cancer is one of the most prevalent malignancies worldwide. The incidence of lung cancer was 234,030 cases in 2018 (accounting for 27% of new cancer cases), with 154,050 mortalities in 2018 (accounting for 51% of cancer-associated mortalities) (1). The five-year net survival rate of patients with lung cancer was typically low (10-20% in most nations) (2,3). Smoking is a major risk factor for lung cancer. Studies have revealed that lung cancer morbidity and mortality increases with smoking in a dose-dependent manner (4-6). Meanwhile, secondhand smoke exposure results in >41,000 mortalities among non-smoking adults each year (7).

Although the majority of lung cancer cases were the result of smoking, until 2008 10-30% of lung cancer cases worldwide were not due to tobacco use (8,9). The development of lung cancer in people who have never smoked (defined as <100 cigarettes in their lifetime) is becoming a growing health problem. Tumors from patients who had never smoked have significant gender, geography, histopathological, molecular and clinical differences when compared with smoking-induced lung cancer tumors (10). However, the genome-wide similarities and differences between smoking-associated and non-smoking lung adenocarcinoma are largely unknown. Lung adenocarcinoma has surpassed squamous cell carcinoma as the most common histologic subtype in various nations (11,12). Therefore, a deeper understanding of the biological characteristics and differences between smoking and non-smoking lung adenocarcinoma may improve the treatment and screening options for patients.

In recent years, several mRNAs, long non-coding RNAs and microRNAs have been identified as biomarkers for the non-invasive detection of various types of cancer, including lung, breast, ovarian, prostate and endometrial cancer (13-17). The current study performed an analysis of smoking and

non-smoking lung adenocarcinoma in The Cancer Genome Atlas (TCGA) database to identify differentially expressed genes (DEGs) and associated signaling pathways. Multivariate regression analysis showed that seven mRNAs, cytochrome P450 family 17 subfamily A member 1 (CYP17A1), PKHD1 like 1 (PKHD1L1), retinoid isomerohydrolase RPE65 (RPE65), neurotensin receptor 1 (NTSR1), fetuin B (FETUB), insulin-like growth factor binding protein 1 (IGFBP1) and glucose-6-phosphatase catalytic subunit (G6PC), significantly distinguished between non-smoking and smoking adenocarcinomas. These genes may serve as potential non-invasive biomarkers for the diagnosis of smoking-associated lung adenocarcinoma.

Materials and methods

Lung adenocarcinoma patient datasets. The mRNA expression information and corresponding clinical information of patients with lung adenocarcinoma was obtained from The Cancer Genome Atlas (TCGA; tcga-data.nci.nih.gov/tcga). The chosen cohort contained 522 lung adenocarcinoma sample tissues, comprising 433 samples of smoking-associated lung adenocarcinoma, 75 samples of non-smoking lung adenocarcinoma and 14 samples where smoking information was not available generated by the TCGA Research Network (https://www.cancer.gov/tcga). A sample was considered as non-smoking adenocarcinoma if the patient had never smoked or smoked <100 cigarettes in their lifetime (18). Samples from past and current smokers were pooled together as smoking-associated adenocarcinoma (19,20).

Identification of DEGs between smoking and non-smoking lung adenocarcinoma. Differential mRNA expression between smoking and non-smoking lung adenocarcinoma was evaluated using the edgeR package in R/Bio conductor (version 3.26.5; http://www.bioconductor.org/packages/release/bioc/html/edgeR.html) (21). The DEGs between the data sets were obtained using |log2-fold change|≥2.0 and P<0.01 as cut-off criteria.

Function and pathway enrichment analysis of differentially expressed mRNAs. To understand the DEGs underlying biological processes and pathways, Gene Ontology (GO; geneontology.org) and Kyoto Encyclopedia of Genes and Genomes (KEGG; www.genome.jp/kegg) pathway analysis were conducted using R software and the Database for Annotation, Visualization and Integrated Discovery (DAVID version 6.8; david.ncifcrf.gov). GO enrichment results were visualized using the R packages digest (version 0.6.20; CRAN.R-project.org/package=digest) and ggplot2 (version 3.2.0; CRAN.R-project.org/package=ggplot2). KEGG enrichment results were analyzed by the R packages RSQLite (version 2.1.1; CRAN.R-project.org/package=RSQLite) and org.Hs.eg.db (version 3.8.2; bioconductor.org/packages/org.Hs.eg.db) along with ActivePerl software (version 5.24.3; https://www.activestate.com/products/activeperl). GO terms and KEGG pathways were selected with a false discovery rate (FDR)<0.05.

Construction of DEG protein-protein interaction (PPI) networks and hub genes association networks. The online protein interaction Search Tool for the Retrieval of Interacting Genes/Proteins (version 11.0; STRING; string-db.org) was used to identify the human proteins associated with the DEGs and to establish a PPI network (22). Only the interactions with a combined score >0.4 were chosen for the PPI network (23). The PPI network was visualized using Cytoscape software (version 3.6.1) (24) and the association between the proteins and DEGs was analyzed. The tight link hub genes in the PPI network were calculated using MCODE (version 1.5.1; http://apps.cytoscape.org/apps/mcode) using default parameters.

Cox proportional hazard regression model. After integrating clinical data and differential gene expression data, 19 of 433 patients with smoking lung adenocarcinoma were deleted because of no overall survival clinical data. Therefore, 414 patients were used for further analysis. The clinical survival information and DEG data were combined and a univariate Cox proportional hazard analysis was performed to identify target biomarkers (P<0.001) and candidate genes associated with patient survival time. Multivariate Cox regression analysis was subsequently performed to further screen for factors associated with patient survival time. Using the median of the prognostic risk score as a critical point (0.94), smoking-related lung adenocarcinomas were classified as high-risk (n=207) or low-risk (n=207). Kaplan-Meier and receiver operating characteristic (ROC) curves were used to analyze the potential clinical significance of these biomarkers as molecular prognostic markers for the five-year overall survival. Kaplan-Meier curves were constructed using the R package survival (CRAN.R-project.org/package=survival. ROC curves were constructed using the R package survival ROC (version 1.0.3; CRAN.R-project.org/package=survivalROC). The risk heat map was constructed using the R package pheatmap (version 1.0.12; CRAN.R-project.org/package=pheatmap) and had a significant impact on survival.

Results

Differentially expressed mRNAs in smoking-associated lung adenocarcinoma compared with non-smoking lung adenocarcinoma. Analysis of TCGA transcription data from 433 smoking-associated lung adenocarcinoma samples and 75 non-smoking lung adenocarcinoma samples revealed that 373 mRNAs were differentially expressed (|log2-fold change|≥2.0 and P<0.01). Of these DEGs, 71 mRNAs were downregulated while 302 mRNAs were upregulated. These results demonstrated that the gene profiles of smoking and non-smoking lung adenocarcinomas were significantly different. The DEGs are displayed in a heat map and a volcano map (Fig. 1A and B). Detailed differential mRNA expression levels are presented in Table I.

GO functional predictions of DEGs in smoking-associated adenocarcinoma. To predict the function of aberrantly expressed genes, GO functional data were downloaded from DAVID. Differential mRNA expression analysis was performed with three functional assemblies: Biological process, cellular component and molecular function (Fig. 2A and B). A total of 28 significant GO functions with an FDR<0.05 were

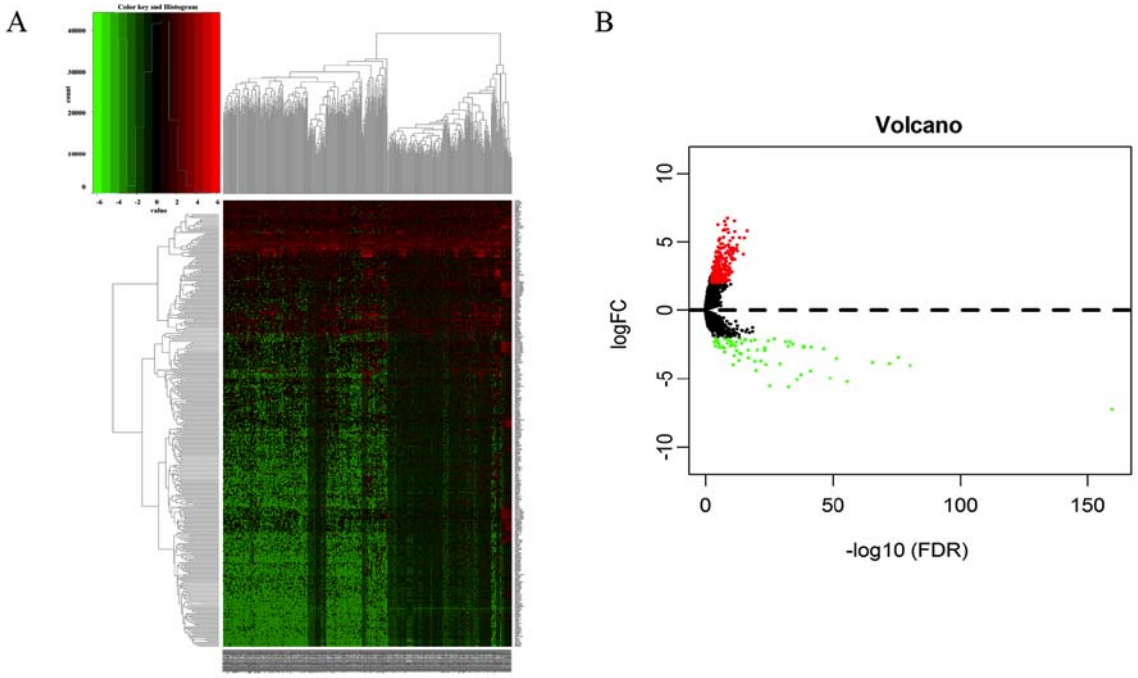

Figure 1. Analysis of differentially expressed mRNAs in smoking-associated adenocarcinomas compared with non-smoking lung adenocarcinomas. (A) Heatmap displaying the expression levels of the differentially expressed genes. (B) Volcano plot of the log2FC and -log10 (FDR). Significant RNA expression differences in smoking and non-smoking lung adenocarcinoma are presented (upregulated genes in red and downregulated genes in green). FC, fold change; FDR, false discovery rate.

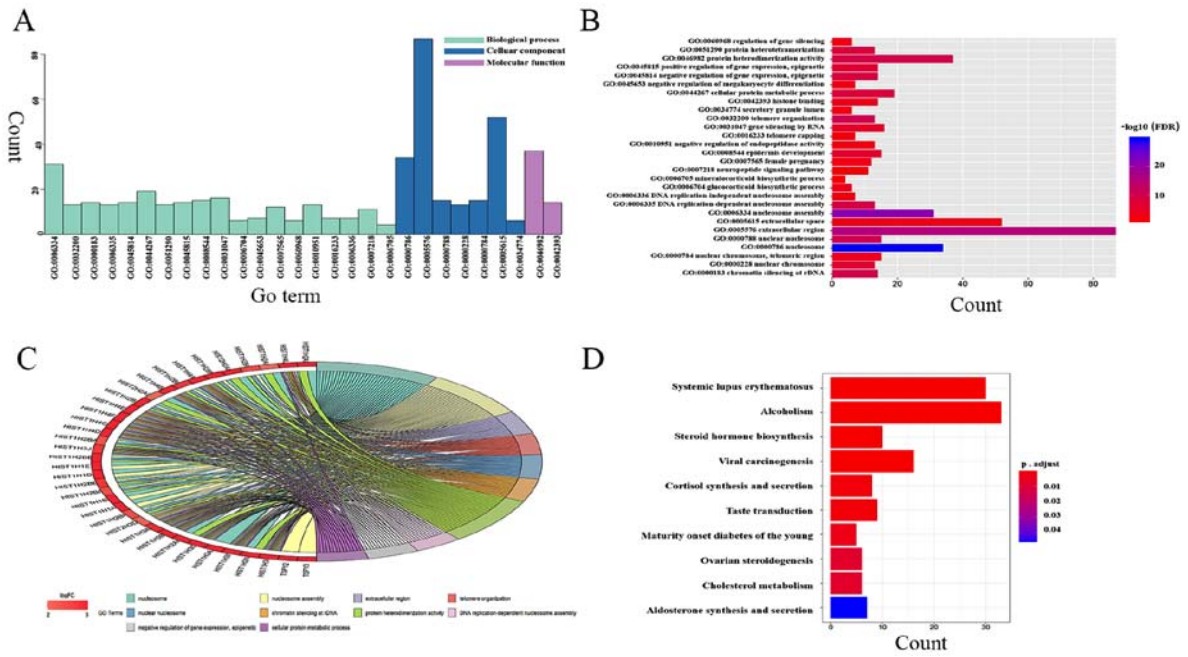

Figure 2. Significantly enriched GO terms and KEGG pathway analysis in smoking-associated lung adenocarcinoma. (A) GO analysis classified the DEGs by biological process, cellular component and molecular function. (B) Significantly enriched GO terms for the DEGs in smoking lung adenocarcinoma (functions). (C) Top 10 significant GO terms and associated hub genes. The color key represents the corresponding GO (D) KEGG pathway analysis of significantly enriched genes and hub gene counts. For each term, the number of enriched genes is indicated by the bar size; while the level of significance is represented by the color. Blue indicates low significance while red represents high significance (FDR<0.05). GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes; FC, fold change; FDR, false discovery rate.

identified. The top 10 GO functions and corresponding genes are presented in Fig. 2C. Detailed GO results are presented in Table II. The present study demonstrated that 'nucleosomes' was the most significant GO term for the identified DEGs.

KEGG pathway enrichment of differentially expressed mRNAs. To predict the KEGG pathway enrichment for the identified DEGs, pathway enrichment data were downloaded from KEGG. A total of 11 significantly KEGG pathways

Table I. Differentially expressed genes in smoking-associated lung adenocarcinoma compared with non-smoking adenocarcinoma.

A, Upregulated genes
CALB1, HIST1H4C, HIST1H1E, HIST1H1B, POU5F2, HIST1H4B, HIST2H2AB, HIST1H4E, HIST1H2BB, WFDC5, HIST1H4D, HIST1H1D, HIST1H2BI, PNMA5, HIST1H3B, HIST1H2AB, WFDC12, HIST1H2AJ, TEX19, KIR2DL1, HIST1H2BL, MSTN, HIST1H2AH, HIST1H2BE, GPR22, HIST1H3C, TAS2R30, NNAT, NTS, APOA1, GPR52, DHRS2, HIST1H2BM, HIST2H2AC, HIST1H3F, PRH2, HIST1H4A, HIST1H2BH, HIST1H3J, LRRC38, APOA2, AFP, HIST1H1A, HIST1H3A, HIST1H2AL, HIST1H3I, PRB4, HIST1H2BO, HIST2H3D, NECAB2, PRB3, CHGA, HRG, INSM1, TAC3, IFNK, MYT1, MAEL, SCG2, HIST1H4F, PRSS48, ACTN3, HIST1H4L, C10orf113, NSG2, HIST1H2BF, VTN, IRX4, SPIC, LRRTM2, TAS2R13, GAL, DPPA2, PSG11, FABP7, TKTL1, SEZ6, ZPBP2, NKX2-3, PSG1, KCNH6, ADGRB1, GABRA2, TAS2R46, TUBA3E, ADAM20, PSG8, STXBP5L, 4-Mar, OR6T1, ANGPTL3, ZP2, PSG5, F2, TAGLN3, PSG3, HBE1, FXYD4, SERPINB13, TDRD12, PNMA6E, SPATA21, CDK5R2, BOLL, RPE65, SPINK4, HIST1H2AD, PTPRN, HMX2, SPRR2E, PBOV1, SLC14A2, SPRR2G, MAB21L2, CT45A1, AKR1C4, RNF113B, BHMT, PSG2, AMBP, PRSS56, HRH3, PI3, KRT14, TSPYL6, SLC1A6, CHRNB2, RBM46, TDRD15, MPC1L, XKR7, ACTL6B, NOS1, CLCA4, PSG7, FGF4, LIPF, KIR3DL2, EPHA5, KRT13, KCNJ13, C12orf40, OR4A16, FEV, GC, SBSN, DPPA5, CXorf67, LRTM2, CGA, APOC3, TSPY2, PSG6, KNG1, NEUROD4, FRG2C, NKX2-2, TAS2R50, CNGA3, KRT5, TAS2R3, CDH9, GCG, APOB, HHLA1, HEPACAM2, KLK13, VSX2, KRT31, NEUROG3, NTSR1, ADH7, CA6, SLC7A14, MSMB, KRT33A, C6orf10, FOXI1, VGLL2, SNX31, PTF1A, DKK4, LGALS14, UGT2A1, CLEC2A, TSPY3, DEFA5, KRT83, BANF2, FETUB, PRB1, TMIGD1, LCE3D, KRT77, TEX13B, CBLN1, OR51B5, CRISP1, SERPINA11, FAM83C, MYBPC1, NRSN1, RAX, SPRR2A, KPRP, H3.Y, SCG3, NPY, NLRP11, PPP1R3A, CALY, PAH, FGF3, DSPP, PSG4, MUC2, CACNG7, AMBN, SOHLH1, INS, SLC6A2, TUNAR, FAM205C, GPR50, BPIFB4, IGFBP1, G6PC, SPINT4, TAS2R43, KRT9, TMPRSS11A, ALB, CRYBA2, GMNC, HSD3B1, SLC6A19, ADAMTS19, MORC1, SLC6A5, RBP3, ADGRG7, SULT1C3, PNMA6F, PAQR9, PRLHR, UCN3, NEUROD1, HDGFL1, SPRR2D, SRARP, TLE7, FGF21, CERS3, CT45A10, LUZP4, CLCA1, TAC1, FRG2, S100A7, ZNF560, ZMAT4, SAGE1, SLC17A6, HIST1H2BA, CACNG2, UGT3A1, AMELY, NTSR2, LCN9, LIN28A, C10orf99, TFAP2B, OR13H1, GNAT3, UGT1A7, HAO1, TAAR1, LGALS13, DSG3, MAGEA11, CPLX2, OTX2, RBFOX1, CRH, STRA8, TSPY1, GLRA4, NR0B1, PCSK2, ST8SIA3, ASCL1, NLRP13, BLID, KRT76, CRYGD, AMELX, PRODH2, DMRTB1, CT47B1, SPRR2B, CALCA, AC187653.1, OR56A3
B, Downregulated genes
ITLN1, PRG4, MYRFL, CYP17A1, STAR, HSD3B2, MYL2, TNMD, PKHD1L1, ASIC2, FAM9C, BMX, C21orf62, EBF3, GPR26, FAM9A, PDZRN4, RSPO1, CYP11B1, SLC3A1, CRB2, CYP4F8, AXDND1, SPAG11B, CYP21A2, CYP11B2, SERTM1, MYH7, RHAG, MC2R, SSX3, ANKRD1, FABP1, FBN2, EMX2, CALN1, HPR, STAC2, SORCS3, PCDH8, TUSC5, BARHL2, PRSS38, CEACAM18, OLFM4, DCX, SULT2A1, SCGB2A2, SPAG11A, AGXT2, CASR, C1orf94, BTNL3, HOXA13, VCX3B, BNC1, CRABP1, SNTG1, REG3A, DPCR1, REG3G, REG4, SPANXD, SPANXC, MUC17, ADIPOQ, UGT1A8, SLC2A2, CALML5, TRIM48, FTHL17

with an FDR<0.05 were identified and R software was used to analyze downloaded data. The KEGG pathways analyzed included: 'Systemic lupus erythematosus', 'alcoholism', 'steroid hormone biosynthesis', 'viral carcinogenesis', 'cortisol synthesis and secretion', 'taste transduction', 'maturity-onset diabetes of the young', 'ovarian steroidogenesis', 'cholesterol metabolism', 'aldosterone synthesis and secretion' and 'peroxisome proliferator-activated receptor signaling pathway' (Fig. 2D and Table III). The majority of the DEGs were significantly enriched in the 'systemic lupus erythematosus' pathway. Notably, genes associated with histones, which are an important part of nucleosomes, were identified in this pathway.

Construction of a PPI network using the DEGs. PPI network analysis was performed using the STRING online database and Cytoscape software. A total of 238 proteins were analyzed (Fig. 3) and the tightly linked hub genes in the PPI network were calculated using MCODE. The top 5 most significant gene clusters were identified (Table IV). These genes may serve an important role in the development of smoking-associated lung adenocarcinoma.

Cox proportional hazards regression model. The R/Bioconductor packages survival, survivalROC and pheatmap were used to calculate the prognostic survival of patients in the smoking-associated lung adenocarcinoma group. Seven mRNAs were significantly associated with overall survival, including CYP17A1, PKHD1L1, RPE65, NTSR1, FETUB, IGFBP1, and G6PC. Using the median of the prognostic risk score (0.94) as a cut-off point, these 7 mRNAs were assigned to each patient in the high-risk (n=207) or low-risk (n=207) smoking-associated lung adenocarcinoma groups. The Kaplan-Meier estimate was used to calculate the high-risk and low-risk patient cohort overall survival for the 7 mRNA signatures in patients. Patients in the high-risk group had a significantly worse prognosis compared with the low-risk group (P<0.001; Fig. 4A). ROC analysis was used to assess the sensitivity and specificity of the 7 mRNA markers for the prediction of the five-year

Table II. Significant GO enrichment analysis of differentially expressed genes in smoking-associated lung adenocarcinoma.

TERM ID	Term	Count	False discovery rate
GO:0000786	Nucleosome	34	$2.66x10^{-30}$
GO:0006334	Nucleosome assembly	31	$2.32x10^{-22}$
GO:0005576	Extracellular region	87	$1.87x10^{-16}$
GO:0032200	Telomere organization	13	$3.83x10^{-11}$
GO:0000788	Nuclear nucleosome	15	$4.70x10^{-11}$
GO:0000183	Chromatin silencing at rDNA	14	$1.23x10^{-10}$
GO:0046982	Protein heterodimerization activity	37	$3.58x10^{-10}$
GO:0006335	DNA replication-dependent nucleosome assembly	13	$4.57x10^{-10}$
GO:0045814	Negative regulation of gene expression, epigenetic	14	$9.80x10^{-9}$
GO:0044267	Cellular protein metabolic process	19	$1.41x10^{-8}$
GO:0051290	Protein heterotetramerization	13	$1.89x10^{-8}$
GO:0045815	Positive regulation of gene expression, epigenetic	14	$1.87x10^{-7}$
GO:0000228	Nuclear chromosome	13	$2.76x10^{-7}$
GO:0008544	Epidermis development	15	$1.07x10^{-6}$
GO:0031047	Gene silencing by RNA	16	$4.51x10^{-6}$
GO:0000784	Nuclear chromosome, telomeric region	15	$2.08x10^{-4}$
GO:0006704	Glucocorticoid biosynthetic process	6	$4.59x10^{-4}$
GO:0042393	Histone binding	14	$5.15x10^{-4}$
GO:0045653	Negative regulation of megakaryocyte differentiation	7	0.001
GO:0007565	Female pregnancy	12	0.001
GO:0060968	Regulation of gene silencing	6	0.001
GO:0005615	Extracellular space	52	0.001
GO:0034774	Secretory granule lumen	6	0.002
GO:0010951	Negative regulation of endopeptidase activity	13	0.005
GO:0016233	Telomere capping	7	0.005
GO:0006336	DNA replication-independent nucleosome assembly	7	0.012
GO:0007218	Neuropeptide signaling pathway	11	0.035
GO:0006705	Mineralocorticoid biosynthetic process	4	0.043

GO, Gene Ontology.

overall survival. The area under the curve (AUC) was 0.769 [95% confidence interval (CI), 0.70-0.83], which indicated that the 7 mRNAs had high sensitivity and specificity (Fig. 4B). Therefore, the model exhibits a high predictive power that could be used to predict the overall survival of patients with smoking-associated lung adenocarcinoma. To better understand the association between the expression of these 7 mRNAs and the survival time of patients, a risk heat map of these mRNAs in combination with clinical survival data was generated (Fig. 4C).

Discussion

Lung cancer is the main cause of oncogenic mortality in males and females worldwide. In spite of improved understanding of oncogenic drivers, few studies have identified genes that are differentially expressed between smoking and non-smoking lung adenocarcinoma. The elucidation of the mechanisms underlying the pathogenesis of smoking-associated lung adenocarcinoma is a challenging task. The current study used bioinformatics methods to analyze 433 samples of smoking-associated lung adenocarcinoma and 75 samples of non-smoking lung adenocarcinoma. A total 373 mRNAs that were differentially expressed between the two groups were identified. Of these, 71 mRNAs were downregulated and 302 mRNAs were upregulated. To predict the function of aberrantly expressed genes, pathway analysis was performed and 28 significant GO functions and 11 significantly enriched KEGG pathways were identified. The Cox proportional hazards regression model suggested that 7 mRNAs may be used as prognostic indicators: CYP17A1, PKHD1L1, RPE65, NTSR1, FETUB, IGFBP1 and G6PC. The AUC of the 7 mRNAs analyzed was 0.769 (95% CI, 0.70-0.83), which indicated that the model had a good predictive value (25).

CYP17A1 is a qualitative regulator of human steroid biosynthesis (26). It is a potential non-small cell lung cancer (NSCLC) susceptibility candidate gene, which converts testosterone to estradiol in hormone-associated cancers (27). Olivo-Marston *et al* (28) revealed a small yet significant association between the CYP17A1 rs743572 polymorphism and lower serum estrogen and improved survival of patients with NSCLC. While Zhang *et al* (29) demonstrated that

Table III. Significant Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis of differentially expressed genes in smoking-associated lung adenocarcinoma.

Pathway ID	Pathway	Count	P-value (adjust)	Genes
hsa05322	Systemic lupus erythematosus	30	6.82×10^{-22}	HIST1H4C, HIST1H4B, HIST2H2AB, HIST1H4E, HIST1H2BB, HIST1H4D, HIST1H2BI, HIST1H3B, HIST1H2AB, HIST1H2AJ, HIST1H2BL, HIST1H2AH, HIST1H2BE, HIST1H3C, HIST1H2BM, HIST2H2AC, HIST1H3F, HIST1H4A, HIST1H2BH, HIST1H3J, HIST1H3A, HIST1H2AL, HIST1H3I, HIST1H2BO, HIST2H3D, HIST1H4F, HIST1H4L, HIST1H2BF, HIST1H2AD, HIST1H2BA
hsa05034	Alcoholism	33	1.65×10^{-21}	HIST1H4C, HIST1H4B, HIST2H2AB, HIST1H4E, HIST1H2BB, HIST1H4D, HIST1H2BI, HIST1H3B, HIST1H2AB, HIST1H2AJ, HIST1H2BL, HIST1H2AH, HIST1H2BE, HIST1H3C, HIST1H2BM, HIST2H2AC, HIST1H3F, HIST1H4A, HIST1H2BH, HIST1H3J, HIST1H3A, HIST1H2AL, HIST1H3I, HIST1H2BO, HIST2H3D, HIST1H4F, HIST1H4L, HIST1H2BF, HIST1H2AD, NPY, CALML5, HIST1H2BA, CRH
hsa00140	Steroid hormone biosynthesis	10	1.21×10^{-5}	CYP17A1, HSD3B2, CYP11B1, CYP21A2, CYP11B2, AKR1C4, UGT2A1, UGT1A8, HSD3B1, UGT1A7
hsa05203	Viral carcinogenesis	16	8.13×10^{-5}	HIST1H4C, HIST1H4B, HIST1H4E, HIST1H2BB, HIST1H4D, HIST1H2BI, HIST1H2BL, HIST1H2BE, HIST1H2BM, HIST1H4A, HIST1H2BH, HIST1H2BO, HIST1H4F, HIST1H4L, HIST1H2BF, HIST1H2BA
hsa04927	Cortisol synthesis and secretion	8	<0.001	CYP17A1, STAR, HSD3B2, CYP11B1, CYP21A2, MC2R, HSD3B1, NR0B1
hsa04742	Taste transduction	9	<0.001	ASIC2, TAS2R30, TAS2R13, GABRA2, TAS2R46, TAS2R50, TAS2R3, TAS2R43, GNAT3
hsa04950	Maturity onset diabetes of the young	5	0.003	NKX2-2, NEUROG3, SLC2A2, INS, NEUROD1
hsa04913	Ovarian steroidogenesis	6	0.007	CYP17A1, STAR, HSD3B2, CGA, INS, HSD3B1
hsa04979	Cholesterol metabolism	6	0.007	STAR, APOA1, APOA2, ANGPTL3, APOC3, APOB
hsa04925	Aldosterone synthesis and secretion	7	0.046	STAR, HSD3B2, CYP21A2, CYP11B2, MC2R, CALML5, HSD3B1
hsa03320	Peroxisome proliferator-activated receptor signaling pathway	6	0.049	FABP1, APOA1, APOA2, FABP7, APOC3, ADIPOQ

Hsa, *homo sapiens*.

CYP17A1 polymorphisms were not associated with NSCLC development in Asian patients. PKHD1L1 has been implicated in lymph node metastasis in endometrial cancer (30). Mutation of PKHD1L1 served an important role in patients with early high-grade serous ovarian cancer (31). RPE65 is highly expressed in the retinal pigment epithelium and encodes an isomerohydrolase that is required for converting all-trans-retinyl esters into 11-cis-retinal, the natural ligand and chromophore for the opsins in rod and cone photoreceptor cells (32). NTSR1 and its ligand neurotensin are frequently overexpressed in tumors of epithelial origins. This ligand/receptor complex contributes to the progression of several tumor types, such as liver cancer or prostate cancer, via the activation of the biological processes involved in tumor progression (33,34). The monoclonal antibody against NTSR1 restores sensitivity to platinum-based therapy and decreases metastasis in lung cancer (35). FETUB, a liver-derived plasma protein, has recently been reported to influence glucose metabolism (36). FETUB copy number amplification in human esophageal cancer, head and neck

Table IV. Top five most significant gene clusters analyzed by MCODE in the protein-protein interaction network.

Cluster	Nodes number	Edges number	Genes
1	30	420	HIST1H4C, HIST1H3F, HIST1H4D, HIST1H4L, HIST1H4E, HIST1H3A, HIST1H4F, HIST1H3I, HIST1H2AH, HIST1H4B, HIST2H2AC, HIST2H2AB, HIST1H2BH, HIST1H2AB, HIST1H2AJ, HIST2H3D, HIST1H2BM, HIST1H4A, HIST1H2BL, HIST1H2BA, HIST1H2BF, HIST1H2BB, HIST1H2BO, HIST1H2AD, HIST1H3J, HIST1H3B, HIST1H3C, HIST1H2BE, HIST1H2AL, HIST1H2BI
2	16	89	NPY, GAL, TAS2R13, ALB, KNG1, GCG, TAS2R46, GNAT3, HRH3, TAS2R43, TAS2R3, TAC1, CASR, TAS2R30, NTS, TAS2R50
3	23	81	HSD3B2, RHAG, HBE1, APOA2, CYP11B2, CALCA, CYP17A1, SULT2A1, AMBP, CRH, MC2R, STAR, CYP21A2, IGFBP1, NR0B1, APOB, APOA1, TAC3, AFP, CYP11B1, NTSR2, NTSR1, APOC3
4	5	10	LGALS13, PSG2, PSG1, PSG3, PSG6
5	4	6	SPINT4, SPAG11B, SPAG11A, CRISP1

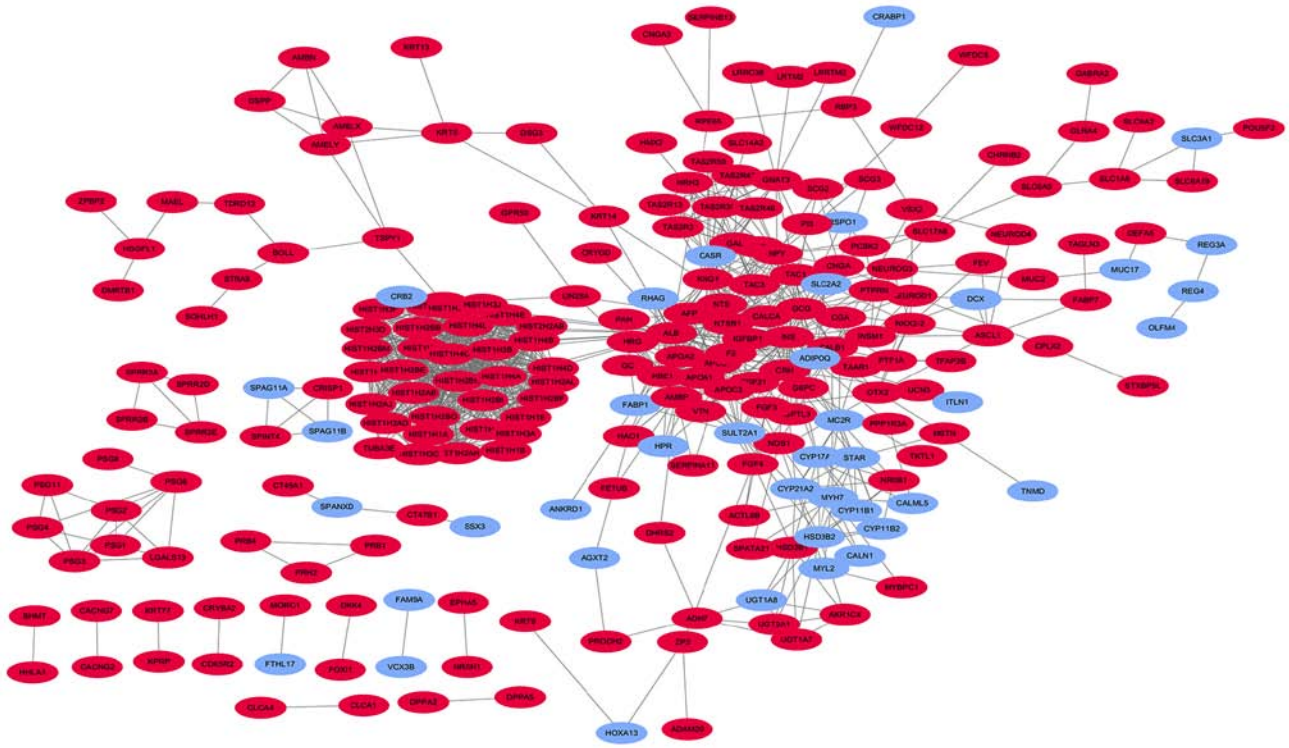

Figure 3. DEG protein-protein interaction network and hub gene analysis. A total of 238 DEGs were filtered into a PPI network containing 360 nodes and 1116 edges. Upregulated proteins are shown in red, and downregulated proteins are shown in blue. DEG, differentially expressed gene; PPI, protein-protein interaction.

squamous cell carcinoma was at least 10-23% (37). FETUB was associated with decreased lung function in patients with chronic obstructive pulmonary disease (COPD), and predicted the occurrence of acute exacerbation or frequent acute exacerbation (38). FETUB, in combination with other markers, may have diagnostic and prognostic value in COPD.

IGFBP1-6 are high-affinity regulators of insulin-like growth factor (IGF) activity and modulate important biological processes, including cell proliferation, survival, migration, senescence, autophagy, angiogenesis, differentiation and apoptosis (39,40). Apart from inhibiting the actions of IGF by inhibiting binding to the IGF-1 receptor, IGFBP1 also performs IGF-independent actions, including the modulation of other growth factors, nuclear localization, transcriptional regulation and binding to non-IGF molecules involved in tumorigenesis, growth, progression and metastasis (41). The expression and function of IGFBP1 in stimulating or inhibiting lung cancer growth have yet to be elucidated (39). G6PC catabolizes glucose-6-phosphate (G6P) to glucose and inorganic phosphate, thereby preventing the accumulation

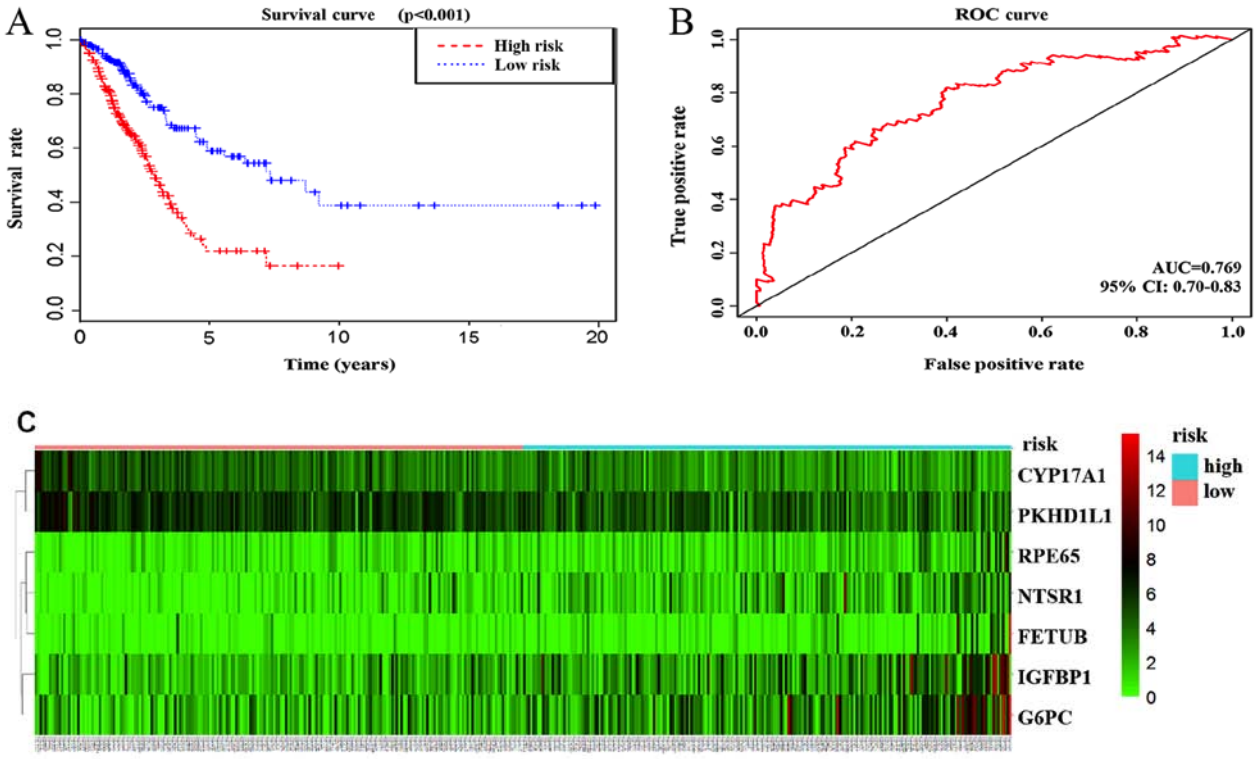

Figure 4. Cox proportional hazards regression model. (A) Kaplan-Meier curves for the analysis of overall survival differences in low and high-risk patients (P<0.001). (B) ROC curves of the sensitivity and specificity of 7 mRNAs in overall survival prediction in patients. (C) A risk heat map constructed from 7 mRNAs that had a significant impact on survival from 414 patients. The risk value gradually increases from left to right. ROC, receiver operating characteristic; AUC, area under the curve; CI, confidence interval; CYP17A1, cytochrome P450 family 17 subfamily A member 1; FETUB, fetuin B; G6PC, glucose-6-phosphatase catalytic subunit; IGFBP1, insulin-like growth factor binding protein 1; NTSR1, neurotensin receptor 1; PKHD1L1, PKHD1 like 1; RPE65, retinoid isomerohydrolase RPE65.

of G6P, which regulates oxidative metabolism of cancer cells (42).

While primarily thought of as an hepatic enzyme that serves a major role in glucose homeostasis, G6PC is dysregulated in an array of human tumor types, such as ovarian cancer (43). Lack of G6PC expression decreased liver cell immunity and promoted tumor development in patients with glycogen storage disease (44,45).

In conclusion, the present study evaluated the mRNA expression of 433 patients with smoking-associated lung adenocarcinoma and 75 patients with non-smoking lung adenocarcinoma. A total of seven genes were identified to have high diagnostic sensitivity and specificity associated with overall survival of patients with smoking-associated lung adenocarcinoma patients. The lack of experimental data to verify these findings is a limitation of the present study. It will be interesting to further explore the roles of CYP17A1, NTSR1, FETUB, IGFBP1 and G6PC in the development of smoking-associated lung adenocarcinoma.

Acknowledgements

Not applicable.

Funding

The current study was supported by the Natural Science Foundation of Shandong Province (grant no. ZR2018MH021), Shandong Medical and Health Science and Technology Development Project (grant no. 2016WS0144) and the National Natural Science Foundation of China (grant no. 81602593).

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions

DZ and XM designed the study. YS, YJ, DL, JW, XC and YZ contributed to the analysis and interpretation of data. DZ and XM wrote the initial draft of the manuscript. DZ, YJ and XM revised the paper. All authors approved the final version manuscript.

Ethical approval and consent to participate

Not applicable.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Siegel RL, Miller KD and Jemal A: Cancer statistics, 2018. CA Cancer J Clin 68: 7-30, 2018.
2. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Nikšić M, Bonaventure A, Valkov M, Johnson CJ, Estève J, *et al*: Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): Analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet 391: 1023-1075, 2018.
3. Wei S, Zhang ZY, Fu SL, Xie JG, Liu XS, Xu YJ, Zhao JP and Xiong WN: Hsa-miR-623 suppresses tumor progression in human lung adenocarcinoma. Cell Death Dis 7: e2388, 2016.
4. Promotion H: Let's make the next generation tobacco-free: Your guide to the 50th anniversary Surgeon General's Report on Smoking and Health. Health Promotion, 1964.
5. Prevention NCfCD, Smoking HPOo, Health. The Health Consequences of Smoking-50 Years of Progress: A report of the surgeon general. Usnational Library of Medicine, 2014.
6. USA USDoH, Services H. How tobacco smoke causes disease: The biology and behavioral basis for smoking-attributable disease: A report of the Surgeon General, 2010.
7. Homa DM, Neff LJ, King BA, Caraballo RS, Bunnell RE, Babb SD, Garrett BE, Sosnoff CS and Wang L; Centers for Disease Control and Prevention (CDC): Vital signs: Disparities in nonsmokers' exposure to secondhand smoke-United States, 1999-2012. MMWR Morb Mortal Wkly Rep 64: 103-108, 2015.
8. Subramanian J and Govindan R: Lung cancer in never smokers: A review. J Clin Oncol 25: 561-570, 2007.
9. Casal-Mouriño A, Valdés L, Barros-Dios JM and Ruano-Ravina A: Lung cancer survival among never smokers. Cancer Lett 451: 142-149, 2019.
10. Sun S, Schiller JH and Gazdar AF: Lung cancer in never smokers-a different disease. Nat Rev Cancer 7: 778-790, 2007.
11. Patel MI, Cheng I and Gomez SL: US lung cancer trends by histologic type. Cancer 121: 1150-1152, 2015.
12. Ryan BM: Lung cancer health disparities. Carcinogenesis 39: 741-751, 2018.
13. Liu J, Wang Y, Liu X, Yuan Q, Zhang Y and Li Y: Novel molecularly imprinted polymer (MIP) multiple sensors for endogenous redox couples determination and their applications in lung cancer diagnosis. Talanta 199: 573-580, 2019.
14. Jiao ZY, Tian Q, Li N, Wang HB and Li KZ: Plasma long non-coding RNAs (lncRNAs) serve as potential biomarkers for predicting breast cancer. Eur Rev Med Pharmacol Sci 22: 1994-1999, 2018.
15. Yang Y, Wu L, Shu X, Lu Y, Shu XO, Cai Q, Beeghly-Fadiel A, Li B, Ye F, Berchuck A, *et al*: Genetic data from nearly 63,000 women of European descent predicts DNA methylation biomarkers and epithelial ovarian cancer risk. Cancer Res 79: 505-517, 2019.
16. Carleton NM, Zhu G, Gorbounov M, Miller MC, Pienta KJ, Resar LMS and Veltri RW: PBOV1 as a potential biomarker for more advanced prostate cancer based on protein and digital histomorphometric analysis. Prostate 78: 547-559, 2018.
17. Zhou Q, Eldakhakhny S, Conforti F, Crosbie EJ, Melino G and Sayan BS: Pir2/Rnf144b is a potential endometrial cancer biomarker that promotes cell proliferation. Cell death Dis 9: 504, 2018.
18. Irimie AI, Braicu C, Cojocneanu R, Magdo L, Onaciu A, Ciocan C, Mehterov N, Dudea D, Buduru S and Berindan-Neagoe I: Differential effect of smoking on gene expression in head and neck cancer patients. Int J Environ Res Public Health 15: pii: E1558, 2018.
19. Li X, Li J, Wu P, Zhou L, Lu B, Ying K, Chen E, Lu Y and Liu P: Smoker and non-smoker lung adenocarcinoma is characterized by distinct tumor immune microenvironments. Oncoimmunology 7: e1494677, 2018.
20. Mathewos T, Yingling CM, Yushi L, Tellez CS, Leander VN, Baylin SS and Belinsky SA: Genome-wide unmasking of epigenetically silenced genes in lung adenocarcinoma from smokers and never smokers. Carcinogenesis 35: 1248-1257, 2014.
21. Robinson MD, McCarthy DJ and Smyth GK: edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140, 2010.
22. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C and Jensen LJ: STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41 (Database Issue): D808-D815, 2013.
23. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, *et al*: STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47: D607-D613, 2019.
24. Rajput A, Thakur A, Sharma S and Kumar M: aBiofilm: A resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. Nucleic Acids Res 46: D894-D900, 2018.
25. Hillis SL: Equivalence of binormal likelihood-ratio and bi-chi-squared ROC curve models. Stat Med 35: 2031-2057, 2016.
26. Xiao F, Yang M, Xu Y and Vongsangnak W: Comparisons of prostate cancer inhibitors abiraterone and TOK-001 binding with CYP17A1 through molecular dynamics. Comput Struct Biotec 13: 520-527, 2015.
27. Gomez L, Kovac JR and Lamb DJ: CYP17A1 inhibitors in castration-resistant prostate cancer. Steroids 95: 80-87, 2015.
28. Olivo-Marston SE, Mechanic LE, Mollerup S, Bowman ED, Remaley AT, Forman MR, Skaug V, Zheng YL, Haugen A and Harris CC: Serum estrogen and tumor-positive estrogen receptor-alpha are strong prognostic classifiers of non-small-cell lung cancer survival in both men and women. Carcinogenesis 31: 1778-1786, 2010.
29. Zhang Y, Hua S, Zhang A, Kong X, Jiang C, Deng D and Wenlong B: Association between polymorphisms in COMT, PLCH1, and CYP17A1, and Non-small-cell lung cancer risk in Chinese nonsmokers. Clin Lung Cancer 14: 45-49, 2013.
30. Kang S, Thompson Z, McClung EC, Abdallah R, Lee JK, Gonzalez-Bosquet J, Wenham RM and Chon HS: Gene expression signature-based prediction of lymph node metastasis in patients with endometrioid endometrial cancer. Int J Gynecol Cancer 28: 260-266, 2018.
31. Chien J, Sicotte H, Fan JB, Humphray S, Cunningham JM, Kalli KR, Oberg AL, Hart SN, Li Y, Davila JI, *et al*: TP53 mutations, tetraploidy and homologous recombination repair defects in early stage high-grade serous ovarian cancer. Nucleic Acids Res 43: 6945-6958, 2015.
32. Harrison EH: Mechanisms of transport and delivery of vitamin A and carotenoids to the retinal pigment epithelium. Mol Nutr Food Res: e1801046, 2019 doi: 10.1002/mnfr.201801046 (Epub ahead of print).
33. Wu Z, Galmiche A, Liu J, Stadler N, Wendum D, Segal-Bendirdjian E, Paradis V and Forgez P: Neurotensin regulation induces overexpression and activation of EGFR in HCC and restores response to erlotinib and sorafenib. Cancer Lett 388: 73-84, 2017.
34. Zhu S, Tian H, Niu X, Wang J, Li X, Jiang N, Wen S, Chen X, Ren S, Xu C, *et al*: Neurotensin and its receptors mediate neuroendocrine transdifferentiation in prostate cancer. Oncogene 38: 4875-4884, 2019.
35. Wu Z, Fournel L, Stadler N, Liu J, Boullier A, Hoyeau N, Fléjou JF, Duchatelle V, Djebrani-Oussedik N, Agopiantz M, *et al*: Modulation of lung cancer cell plasticity and heterogeneity with the restoration of cisplatin sensitivity by neurotensin antibody. Cancer Lett 444: 147-161, 2019.
36. Kralisch S, Hoffmann A, Lössner U, Kratzsch J, Blüher M, Stumvoll M, Fasshauer M and Ebert T: Regulation of the novel adipokines/hepatokines fetuin A and fetuin B in gestational diabetes mellitus. Metabolism 68: 88-94, 2017.
37. Khammanivong A, Anandharaj A, Qian X, Song JM, Upadhyaya P, Balbo S, Bandyopadhyay D, Dickerson EB, Hecht SS and Kassie F: Transcriptome profiling in oral cavity and esophagus tissues from (S)-N'-nitrosonornicotine-treated rats reveals candidate genes involved in human oral cavity and esophageal carcinogenesis. Mol Carcinog 55: 2168-2182, 2016.
38. Diao WQ, Shen N, Du YP, Liu BB, Sun XY, Xu M and He B: Fetuin-B (FETUB): A plasma biomarker candidate related to the severity of lung function in COPD. Sci Rep 6: 30045, 2016.
39. Zheng F, Tang Q, Zheng XH, Wu J, Huang H, Zhang H and Hann SS: Inactivation of Stat3 and crosstalk of miRNA155-5p and FOXO3a contribute to the induction of IGFBP1 expression by beta-elemene in human lung cancer. Exp Mol Med 50: 121, 2018.

40. Major JM, Laughlin GA, Kritz-Silverstein D, Wingard DL and Barrett-Connor E: Insulin-like growth factor-I and cancer mortality in older men. J Clin Endocrinol Metab 95: 1054-1059, 2010.
41. Tang Q, Wu J, Zheng F, Hann SS and Chen Y: Emodin increases expression of insulin-like growth factor binding protein 1 through activation of MEK/ERK/AMPKα and interaction of PPARγ and Sp1 in lung cancer. Cell Physiol Biochem 41: 339-357, 2017.
42. Nyce JW: Detection of a novel, primate-specific 'kill switch' tumor suppression mechanism that may fundamentally control cancer risk in humans: An unexpected twist in the basic biology of TP53. Endocr Relat Cancer 25: R497-R517, 2018.
43. Guo T, Chen T, Gu C, Li B and Xu C: Genetic and molecular analyses reveal G6PC as a key element connecting glucose metabolism and cell cycle control in ovarian cancer. Tumor Biol 36: 7649-7658, 2015.
44. Gjorgjieva M, Calderaro J, Monteillet L, Silva M, Raffin M, Brevet M, Romestaing C, Roussel D, Zucman-Rossi J, Mithieux G, *et al*: Dietary exacerbation of metabolic stress leads to accelerated hepatic carcinogenesis in glycogen storage disease type Ia. J Hepatol 69: 1074-1087, 2018.
45. Kim GY, Kwon JH, Cho J-H, Zhang L, Mansfield BC and Chou JY: Downregulation of pathways implicated in liver inflammation and tumorigenesis of glycogen storage disease type Ia mice receiving gene therapy. Hum Mol Genet 26: 1890-1899, 2017.