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.

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; tcca-data.ncl.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/bioconductor/html/edgeR.html) (21). The DEGs between the data sets were obtained using $|\text{log}_2\text{-fold change}| \geq 2.0$ 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 survivalROC (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 ($|\text{log}_2\text{-fold change}| \geq 2.0$ 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
The top 10 GO functions and corresponding genes are 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 identified.
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 survival.
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...
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 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    | 6.82x10^{-22}    | HIST1H4C, HIST1H4B, HIST2H2AB, HIST1H4E, HIST1H2BB, HIST1H4D, HIST1H2BI, HIST1H3B, HIST1H2AB, HIST1H2AJ, HIST1H2BL, HIST1H2AH, HIST1H2BE, HIST1H3C, HIST1H2BM, HIST2H2AC, HIST1H3F, HIST1H4A, HIST1H2BI, HIST1H3J, HIST1H3A, HIST1H2AL, HIST1H3I, HIST1H2BO, HIST2H3D, HIST1H4F, HIST1H4L, HIST1H2BF, HIST1H2AD, HIST1H2BA |
| hsa05034   | Alcoholism                                   | 33    | 1.65x10^{-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    | 1.21x10^{-5}     | CYP17A1, HSD3B2, CYP1B1, CYP2A2, CYP1B2, AKR1C4, UGT2A1, UGT1A8, HSD3B1, UGT1A7                  |
| hsa05203   | Viral carcinogenesis                         | 16    | 8.13x10^{-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, CYP1B1, CYP2A2, 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            | NKKX2-2, 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, CYP2A1, CYP1B2, MC2R, CALML5, HSD3B1                                            |
| hsa03320   | Peroxisome proliferator-activated receptor signaling pathway | 6     | 0.049            | FABP1, APOA1, APOA2, FABP7, APOC3, ADIPOQ                                                    |

Hsa, homo sapiens.
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...
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.
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, Harn 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 1a. 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 1a mice receiving gene therapy. Hum Mol Genet 26: 1890-1899, 2017.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.