frontiers in Genetics

ORIGINAL RESEARCH
published: 21 December 2020
doi: 10.3389/fgene.2020.565206

Identification and Validation of Two Lung Adenocarcinoma-Development Characteristic Gene Sets for Diagnosing Lung Adenocarcinoma and Predicting Prognosis

Cheng Liu[1]†, Xiang Li[2†], Hua Shao[2] and Dan Li[2]*

[1]Department of Thoracic Surgery, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China, [2]Department of Neurology, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Doron Levy,
University of Maryland, College Park, United States

Reviewed by:
Padhmanand Sudhakar, Katholieke Universiteit (KU) Leuven, Belgium
Priyanka Baloni, Institute for Systems Biology (ISB), United States

***Correspondence:**
Cheng Liu
liuchengdoctor@163.com

†These authors share first authorship

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: 01 July 2020
Accepted: 26 November 2020
Published: 21 December 2020

Citation:
Liu C, Li X, Shao H and Li D (2020) Identification and Validation of Two Lung Adenocarcinoma-Development Characteristic Gene Sets for Diagnosing Lung Adenocarcinoma and Predicting Prognosis. Front. Genet. 11:565206. doi: 10.3389/fgene.2020.565206

Background: Lung adenocarcinoma (LUAD) is one of the main types of lung cancer. Because of its low early diagnosis rate, poor late prognosis, and high mortality, it is of great significance to find biomarkers for diagnosis and prognosis.

Methods: Five hundred and twelve LUADs from The Cancer Genome Atlas were used for differential expression analysis and short time-series expression miner (STEM) analysis to identify the LUAD-development characteristic genes. Survival analysis was used to identify the LUAD-unfavorable genes and LUAD-favorable genes. Gene set variation analysis (GSVA) was used to score individual samples against the two gene sets. Receiver operating characteristic (ROC) curve analysis and univariate and multivariate Cox regression analysis were used to explore the diagnostic and prognostic ability of the two GSVA score systems. Two independent data sets from Gene Expression Omnibus (GEO) were used for verifying the results. Functional enrichment analysis was used to explore the potential biological functions of LUAD-unfavorable genes.

Results: With the development of LUAD, 185 differentially expressed genes (DEGs) were gradually upregulated, of which 84 genes were associated with LUAD survival and named as LUAD-unfavorable gene set. While 237 DEGs were gradually downregulated, of which 39 genes were associated with LUAD survival and named as LUAD-favorable gene set. ROC curve analysis and univariate/multivariate Cox proportional hazards analyses indicated both of LUAD-unfavorable GSVA score and LUAD-favorable GSVA score were a biomarker of LUAD. Moreover, both of these two GSVA score systems were an independent factor for LUAD prognosis. The LUAD-unfavorable genes were significantly involved in p53 signaling pathway, Oocyte meiosis, and Cell cycle.

Conclusion: We identified and validated two LUAD-development characteristic gene sets that not only have diagnostic value but also prognostic value. It may provide new insight for further research on LUAD.

Keywords: lung adenocarcinoma, prognostic stratification system, The Cancer Genome Atlas, gene set variation analysis score, predicting prognosis

INTRODUCTION

Lung cancer is the most common cancer (11.6% of the total cases) among men and women in the world, which is also the main cause of cancer death (18.4% of the total cancer deaths; Bray et al., 2018). Non-small cell lung cancer (NSCLC) accounts for 85% of all lung cancer cases (Govindan et al., 2006), and lung adenocarcinoma (LUAD) is one of the main subtypes of NSCLC. Smoking is currently considered to be the main cause of lung cancer. However, LUAD is more likely to occur in women who do not smoke, and the age of patients tends to become younger (Hecht, 1999; Donner et al., 2018). Early, LUAD can be treated by surgery; however, most patients with LUAD are often diagnosed with advanced cancer (Ding et al., 2008). Although target therapy is effective for selected advanced LUAD, the overall survival of patients is poor due to the emergence of drug resistance. Therefore, it has become one of the hot spots in clinical research to find the diagnosis and prognosis indexes of LUAD.

In recent years, high-throughput sequencing technology and gene database have been widely used in the study of cancer diagnosis and prognosis (Feng et al., 2016; Dama et al., 2017; Zhao et al., 2018a; He et al., 2019). For example, EGFR, KRAS, BRAF, and ERBB 2 have been shown to be associated with treatment efficacy and prognosis (Naoki et al., 2002; Mendelsohn and Baselga, 2003; Guan et al., 2013). Moreover, DGCR 5 has been found to be a prognostic indicator and therapeutic target for the diagnosis and treatment of LUAD (Dong et al., 2018). Overexpression of Rcc 2 induces epithelial-mesenchymal metastasis in LUAD, enhances cell mobility, and promotes tumor metastasis (Pang et al., 2017). Overexpression of KIF20A confers malignant phenotype of LUAD by promoting cell proliferation and inhibiting apoptosis (Zhao et al., 2018b). However, most studies do not take the simultaneous changes of multiple genes into account. Moreover, there are few studies on the LUAD-development characteristic gene sets.

In present study, we identified two LUAD-development characteristic gene sets named as LUAD-unfavorable gene set and LUAD-favorable gene set. Gene set variation analysis (GSVA) was used to score individual samples against the two gene sets. Survival analysis and receiver operating characteristic (ROC) curve analysis were used to identify the diagnostic and prognostic capabilities of two gene sets GSVA score, respectively. Both of LUAD-unfavorable GSVA score and LUAD-favorable GSVA score were reliable biomarkers for diagnosing LUAD and independent biomarkers for predicting prognosis.

MATERIALS AND METHODS

The Cancer Genome Atlas (TCGA; Tomczak et al., 2015)[1] and Gene Expression Omnibus (GEO; Barrett et al., 2013)[2] are the international genetic databases, which are publicly accessible and freely available to researchers. In our study, a total of 512 LUAD samples and 57 healthy lung tissue samples were downloaded from TCGA, including 281 stage I LUADs, 121 stage II LUADs, 84 stage III LUADs, and 26 stage IV LUADs. In addition, GSE10072 based on GPL96 platform was downloaded from GEO, including 58 LUAD samples and 49 healthy lung tissue samples. GSE31210 based on GPL570 platform was downloaded from GEO, including 226 LUAD samples and 20 healthy lung tissue samples. The two data sets were used to verify the prognostic value. The "normalizeBetweenArrays" function in the limma package (Ritchie et al., 2015) was used to normalize the gene expression profiles. If a gene responds to multiple probes, the average value of these probes is considered to be the expression value of the corresponding gene. The flow of this study is shown in **Figure 1**.

Differential Expression Analysis and Short Time-Series Expression Miner

In TCGA, the RNA sequencing expression profile of LUAD was displayed as read counts, which was subsequently normalized by voom function (Law et al., 2014) in limma package. Differentially expressed genes (DEGs) in four stages of LUAD were identified using limma package, respectively. $p < 0.01$ adjusted by the false discovery rate (FDR) and $|\log$ fold change(FC)$| > 1.5$ were considered as significance. In the developing of LUAD, if a DEG was gradually upregulated (logFCstage I vs. control < logFCstage II vs. control < logFCstage III vs. control < logFCstage IV vs. control) or gradually downregulated (logFCstage I vs. control > logFCstage II vs. control > logFCstage III vs. control > logFCstage IV vs. control), and then it was considered to be LUAD-development characteristic gene. These genes were organized into different clusters based on expression patterns using short time-series expression miner (STEM; Ernst and Bar-Joseph, 2006).

Survival Analysis and LUAD-Development Characteristic Gene Set

We used the median expression value of each LUAD-development characteristic gene as the cutoff point to dichotomize patients into high-expression group and low-expression group. Moreover, Kaplan Meier survival analysis and log rank method were performed to explore whether the expression level of the LUAD-development characteristic gene is related to the overall survival (OS) time. Survival analysis was performed using survival package[3] in R, and $p < 0.01$ was considered as significance. In our study, a LUAD-development characteristic gene which gradually upregulated with the development of LUAD and associated with poor prognosis of LUAD was considered to be LUAD-unfavorable gene. On the contrary, a LUAD-development characteristic gene which gradually downregulated with the development of LUAD and associated with good prognosis of LUAD was considered to be LUAD-favorable gene.

[1]https://www.cancer.gov/
[2]https://www.ncbi.nlm.nih.gov/geo/
[3]https://CRAN.R-project.org/package=survival

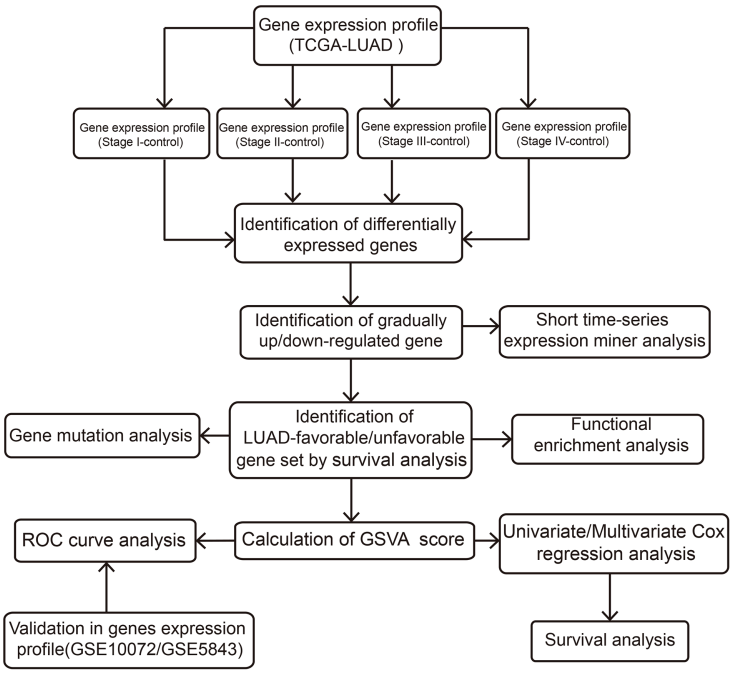

FIGURE 1 | Flowchart of this study.

LUAD-unfavorable genes and LUAD-favorable genes constituted LUAD-unfavorable genes set and LUAD-favorable genes set, respectively.

Calculation of LUAD-Development Characteristic GSVA Score

Gene set variation analysis is a popular method of scoring individual samples for molecular characteristics or gene sets. GSVA package (Hanzelmann et al., 2013) in R was used to calculate LUAD-unfavorable GSVA score and LUAD-favorable GSVA score for individual samples.

ROC Curve Analysis and Univariate/Multivariate Cox Proportional Hazards Analyses

The pROC package (Robin et al., 2011) was used to conduct ROC curve analysis of LUAD-unfavorable GSVA score and LUAD-unfavorable GSVA score to evaluate their ability to diagnose LUAD. Univariate/multivariate Cox proportional hazards analyses were used to compare the relative prognostic value of the two GSVA score systems with that of routine clinicopathological features.

Functional Enrichment Analysis

To further explore the biological function of LUAD-unfavorable genes, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis were performed using the clusterProfiler package (Yu et al., 2012) in R. $p < 0.05$ was considered as significance.

Gene Mutation Analysis and Validation of Differential Expression of LUAD-Unfavorable Genes at Protein Level

In order to explore the potential mechanism about differential expression of LUAD-unfavorable genes, the TCGAbiolinks package (Mounir et al., 2019) was used to download and scan the alteration statuses of LUAD-unfavorable genes. In addition, we randomly selected 10 genes from LUAD-unfavorable gene set and scanned the Human Protein Atlas[4] (Colwill et al., 2011) web tool to validate whether the LUAD-unfavorable genes are upregulated at protein level, compared with normal lung tissue.

RESULTS

Multiple Genes Were Defined as LUAD-Development Characteristic Genes

Compared to normal lung tissue samples, there were 3,082 DEGs in stage I LUADs, 3,437 DEGs in stage II LUADs, 3,518 DEGs in stage III LUADs, and 3,510 DEGs stage IV LUADs (**Figure 2A**). It indicated that the gene expression patterns were various with the development of LUAD. A total of 2,658 common DEGs was in stage I-IV LUADs (**Figure 2B**). Among of them, 185 DEGs were gradually upregulated and 237 DEGs were gradually downregulated with the development of LUAD, which maybe play a crucial role in the LUAD development. The result of STEM demonstrated that two gene clusters were

[4]https://v15.proteinatlas.org/

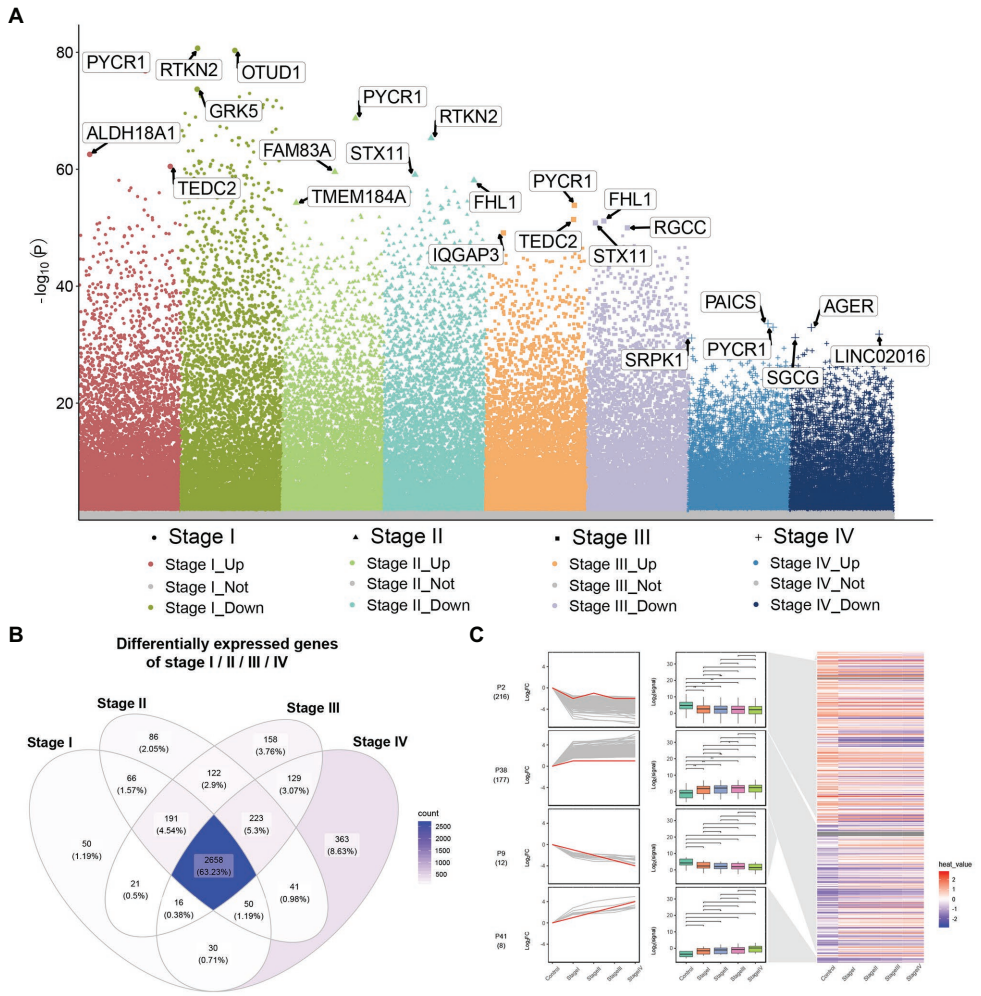

FIGURE 2 | Differential expression gene analysis and short time-series expression miner (STEM) analysis. **(A)** Manhattan plot showed differentially expressed genes (DEGs) in different stage of lung adenocarcinoma (LUAD). Genes with significant differences are highlighted. **(B)** Common DEGs in LUAD stage I–IV. **(C)** STEM results. Line plots (left panels) and box plots (right panels) are used to show fold changes (log2 scale) and absolute expression levels (log2 scale), respectively. In each line plot, one representative gene is highlighted in red.

significantly upregulated, while two gene clusters were significantly downregulated with the development of LUAD (**Figure 2C**).

LUAD-Development Characteristic Genes Were Associated With LUAD Prognosis

The result of survival analysis showed a total of 84 LUAD-development characteristic genes that are gradually upregulated with the development of LUAD and associated with poor prognosis, while a total of 39 LUAD-development characteristic genes that are gradually downregulated with the development of LUAD and associated with good prognosis (**Table 1**). This means that not all LUAD-development characteristic genes are associated with the prognosis of LUAD. In the LUAD-unfavorable gene set, NEK2, CENPK, CDC25C, PLK4, LYPD3, FAM72D, NEIL3, GTSE1, CDK1, and KIF14 were the ten genes with most significant association with poor prognosis (**Figure 3A**). While in the LUAD-favorable gene set, OR7E47P, MS4A2, RAB44, BMP5, ARHGEF6, JAML, TRPC2, HPGDS, HPSE2, and KLK11 were the ten genes with most significant association with good prognosis (**Figure 3B**).

LUAD-Unfavorable GSVA Score and LUAD-Favorable GSVA Score Are Biomarker of LUAD and LUAD Prognosis

As shown in **Figure 4A**, LUAD-favorable GSVA score was gradually downregulated with the development of LUAD, while LUAD-unfavorable GSVA score was gradually upregulated with the development of LUAD. Moreover, the result of ROC curve analysis indicated that both LUAD-unfavorable GSVA score and LUAD-favorable GSVA score are a biomarker of LUAD with AUC = 0.982 and AUC = 0.994, respectively

(**Figure 4B**). Furthermore, the two GSVA score systems were also validated in GSE10072 (**Figure 4C**) and GSE31210 (**Figure 4D**), respectively. According to median GSVA score, all LUAD patients in TCGA were separated into low-score group and high-score group. And both the two GSVA score systems were significantly associated with LUAD prognosis (**Figure 4E**). Patients with high LUAD-unfavorable GSVA score had worse prognosis, while patients with high LUAD-favorable GSVA score had better prognosis. Univariate and multivariate Cox analysis showed that the two GSVA score systems were the independent factors for LUAD prognosis compared with clinicopathological features (**Tables 2** and **3**). Moreover, the two GSVA score systems were also significantly associated with LUAD prognosis in GSE31210 (**Figure 4F**).

TABLE 1 | LUAD-unfavorable gene set and LUAD-favorable gene set.

Gene set	Gene symbol
LUAD-unfavorable gene set	ARHGAP11A, ASPM, BLM, C5orf34, CA9, CCNA2, CDC25C, CDC6, CDCA2, CDK1, CENPF, CENPK, CHAF1B, CLSPN, DDX11-AS1, DEPDC1, DNMT3B, DTL, E2F7, ECT2, EGLN3, ESCO2, EXO1, FAM111B, FAM57B, FAM72D, FAM83D, FANCI, FBXO43, GAL, GTSE1, HASPIN, HELLS, HMMR, KIF11, KIF14, KIFC1, KNL1, KNTC1, KREMEN2, KRT6A, KRT81, LINC01269, LOC101929128, LYPD3, MAD2L1, MELK, MIR924HG, MKI67, MYO19, NCAPG, NDC80, NEIL3, NEK2, NUF2, NUSAP1, OIP5, ORC1, ORC6, PAICS, PARPBP, PCLAF, PIMREG, PLK1, PLK4, POLQ, PRC1, PTPRN, RAD51, RRM2, SGO1, SLC2A1-AS1, SPAG5, SPOCK1, TEDC2, TESMIN, TGFBR3L
LUAD-favorable gene set	TICRR, TROAP, TTK, TYMS, UBE2T, UCA1, ZWINT, ACKR1, ADAMTS8, ADGRF5, ARHGEF6, ATP13A4, BMP5, CASS4, CCDC69, CLEC3B, COL6A6, CTSG, FAM189A2, FBP1, FCER1A, FLI1, GCSAML, GIMAP4, GIMAP7, HPGDS, HPSE2, INMT, JAML, KLK11, LSAMP, LY86, MAL, MS4A2, OR7E47P, P2RY12, RAB44, RTN1, SCN2B, SIGLEC17P, SLCO4C1, SPN, TM6SF1, TRPC2, UNC45B, ZEB2

The Differential Expression of LUAD-Unfavorable Gene May Not Result From Mutation

Only 52 (9.17%) of 567 samples had an alteration in one or several LUAD-unfavorable genes and most samples did not have genetical alteration (**Figure 5A**). Moreover, compared with normal lung tissue, ten genes (ASPM, BLM, CDC25C, CDK1, DEPDC1, KIF11, KIF14, LYPD3, NEK2, and PLK4) of LUAD-unfavorable gene set were included in The Human Protein Atlas and were highly expressed in LUAD (**Figure 5B**).

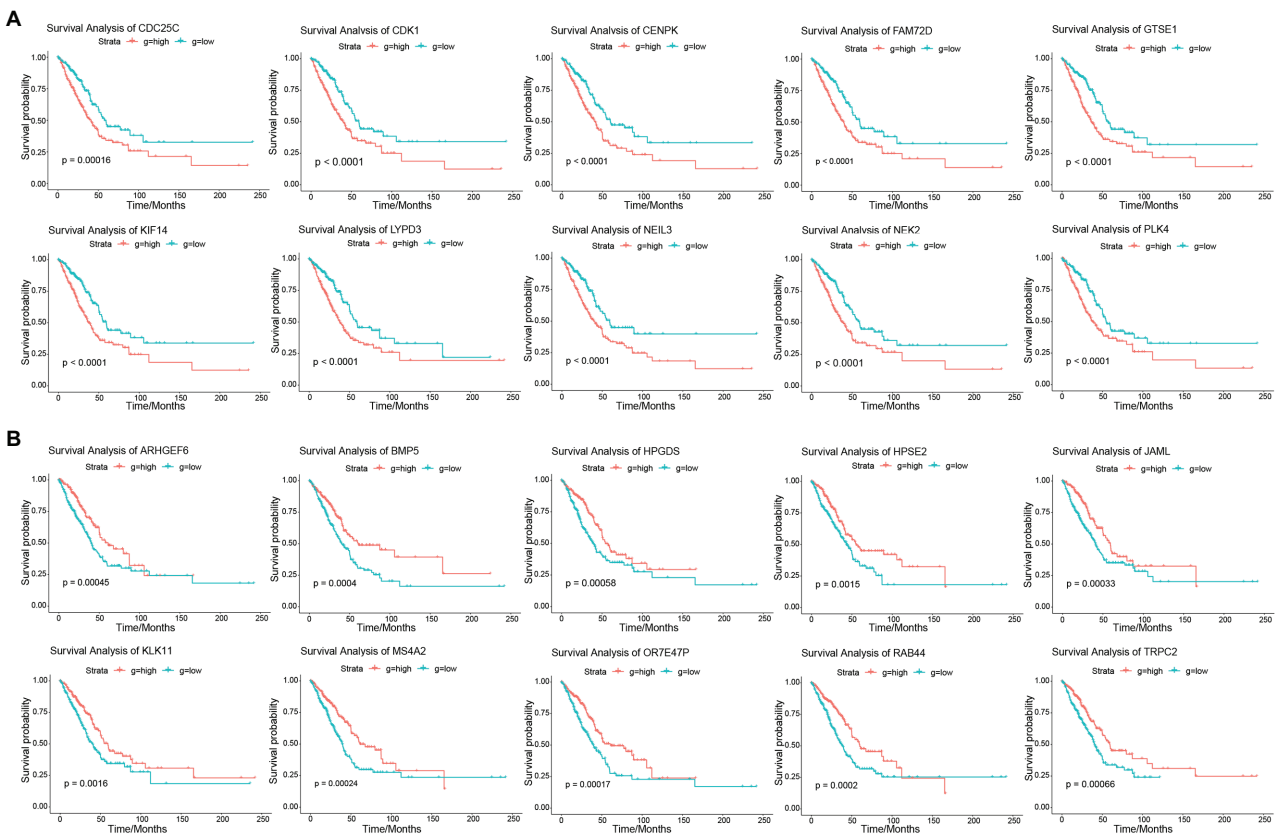

FIGURE 3 | Survival analysis. **(A)** Survival curves of 10 genes most significantly correlated with LUAD in LUAD-unfavorable gene set. **(B)** Survival curves of 10 genes most significantly correlated with LUAD in LUAD-favorable gene set.

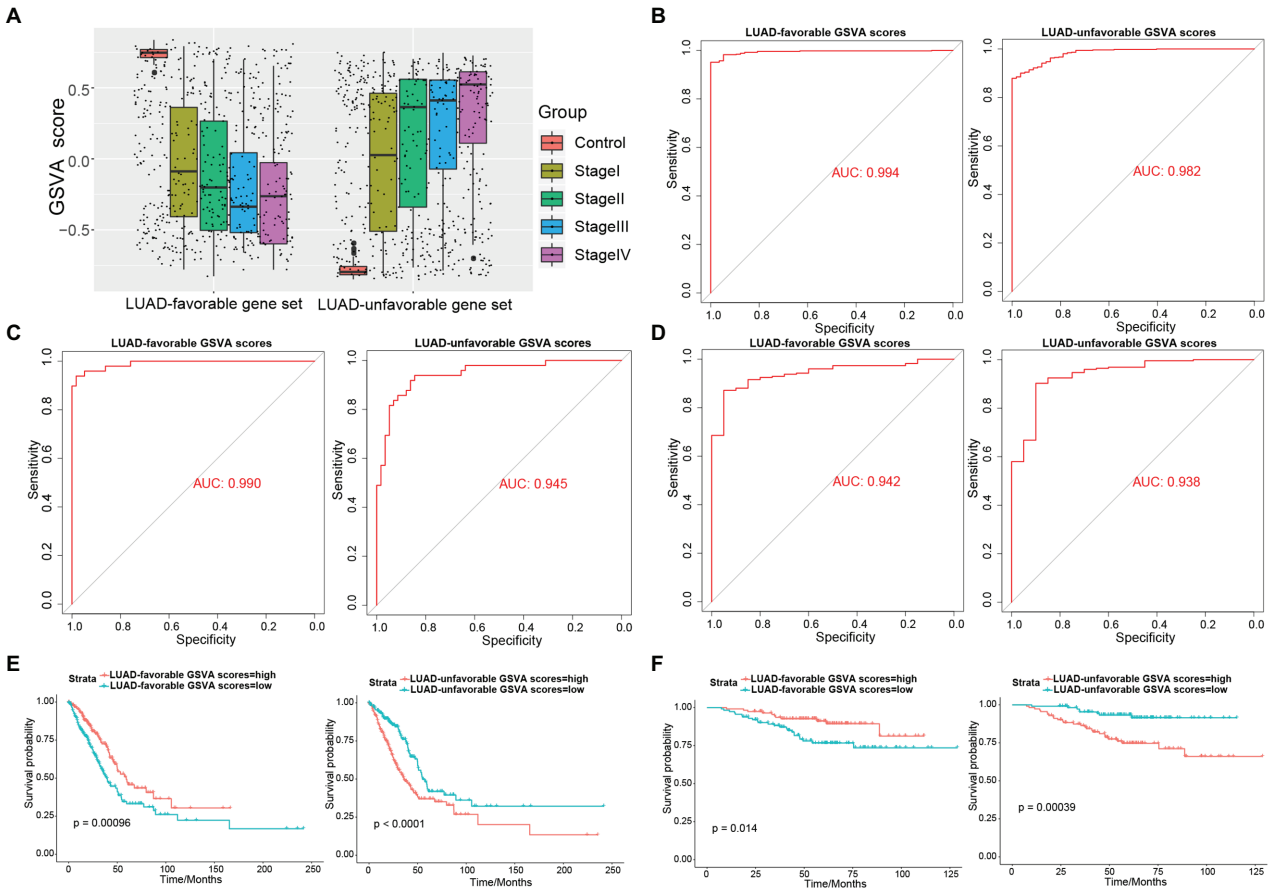

FIGURE 4 | Exploring the diagnostic and prognostic abilities of LUAD-unfavorable gene set and LUAD-favorable gene set. **(A)** The box plots of LUAD-unfavorable gene set gene set variation analysis (GSVA) score and LUAD-favorable gene set GSVA score. **(B)** Receiver operating characteristic (ROC) curves analysis of LUAD-unfavorable gene set GSVA score and LUAD-favorable gene set GSVA score. **(C)** ROC curves analysis of LUAD-unfavorable GSVA score and LUAD-favorable GSVA score in GSE10072. **(D)** ROC curves analysis of LUAD-unfavorable GSVA score and LUAD-favorable GSVA score in GSE31210. **(E)** Survival analysis of LUAD-unfavorable gene set GSVA score and LUAD-favorable gene set GSVA score in LUAD from The Cancer Genome Atlas (TCGA). **(F)** Survival analysis of LUAD-unfavorable GSVA score and LUAD-favorable GSVA score in GSE31210.

TABLE 2 | Univariate and multivariate analyses of LUAD-unfavorable GSVA score.

Factor	Univariate Cox analysis			Multivariate Cox analysis		
	β	*p*	HR (95% CI)	*β*	*p*	HR (95% CI)
Gender (female/male)	0.025	0.867	0.763–1.378			
Age (>65 years/≦65 years)	0.178	0.243	0.886–1.610			
T stage (T3–4/T1–2)	0.821	0.000	1.543–3.346	0.589	0.016	1.114–2.914
Lymph node stage (N2–3/N0–1)	0.818	0.000	1.582–3.243	0.121	0.757	0.523–2.437
Metastasis (M1/M0)	0.749	0.006	1.234–3.626	0.109	0.799	0.482–2.580
Pathological stage (III–IV/I–II)	0.967	0.000	1.924–3.592	0.509	0.211	0.749–3.695
LUAD-unfavorable GSVA score (high/low)	0.614	0.000	1.365–2.500	0.575	0.002	1.230–2.565

LUAD-Unfavorable Genes Involved in Multiple Cancer-Related Pathways

In order to explore the biological functions of LUAD-unfavorable genes, LUAD-unfavorable genes were performed functional enrichment analysis. The results showed that these genes are mainly related to nuclear division, organelle fission, mitotic nuclear division, nuclear chromosome segregation, and chromosome segregation (**Figure 6A**). Moreover, LUAD-unfavorable genes were significantly involved in many pathways, such as Fanconi anemia pathway, p53 signaling pathway, Oocyte meiosis, Cell cycle, and Progesterone-mediated oocyte maturation (**Figure 6B**).

TABLE 3 | Univariate and multivariate analyses of LUAD-favorable GSVA score.

Factor	Univariate Cox analysis			Multivariate Cox analysis		
	β	*p*	HR (95% CI)	*β*	*p*	HR (95% CI)
Gender (female/male)	0.025	0.867	0.763–1.378			
Age (>65 years/≦65 years)	0.178	0.243	0.886–1.610			
T stage (T3–4/T1–2)	0.821	0.000	1.543–3.346	0.557	0.028	1.062–2.869
Lymph node stage (N2–3/N0–1)	0.818	0.000	1.582–3.243	0.420	0.254	0.739–3.134
Metastasis (M1/M0)	0.749	0.006	1.234–3.626	0.263	0.518	0.586–2.886
Pathological stage (III–IV/I–II)	0.967	0.000	1.924–3.592	0.365	0.361	0.659–3.152
LUAD-favorable GSVA score (high/low)	−0.489	0.001	0.454–0.828	−0.434	0.017	0.453–0.926

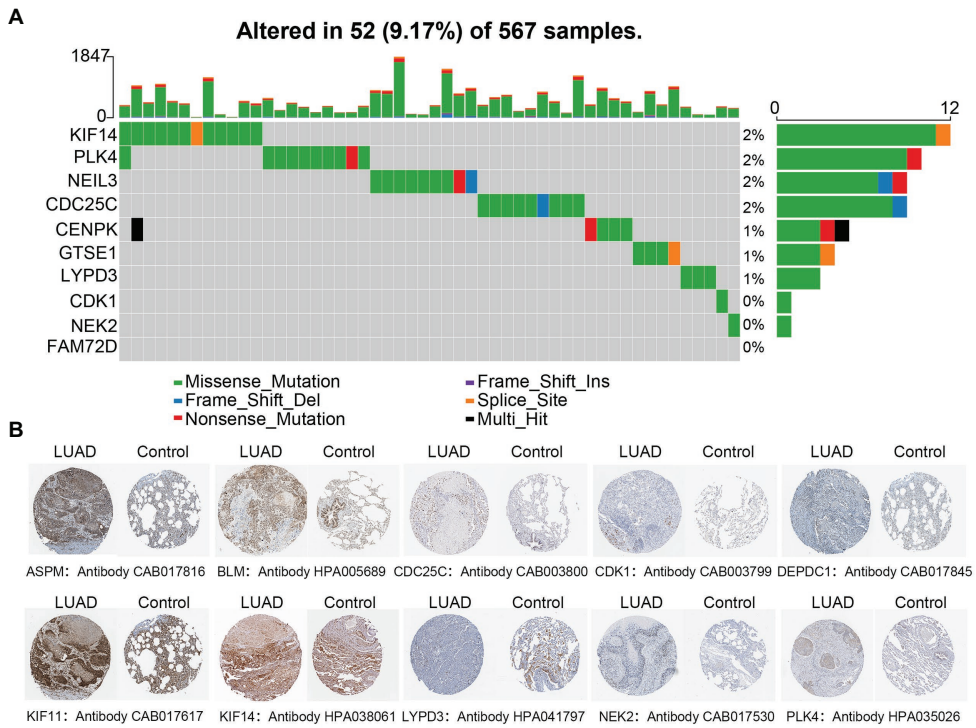

FIGURE 5 | Genetical alteration analysis and immunohistochemistry analysis. **(A)** Genetical alteration analysis of LUAD-unfavorable genes. **(B)** Identification of LUAD-unfavorable gene at protein level. Lung cancer samples are on the left and normal lung tissue samples are on the right.

DISCUSSION

In the world, lung cancer is the main cause of cancer-related death. Even with surgical treatment, the recurrence rate of lung cancer still is very high (Scott et al., 2007). Therefore, it is of great significance to explore biomarkers which can accurately diagnose lung cancer and predict prognosis for the treatment and management of lung cancer. A large number of studies have shown that abnormal expression of genes in lung cancer (including LUAD) is closely related to prognosis, and can be used as a potential biomarker of prognosis (Xu et al., 2013; Cui et al., 2015; Giatromanolaki et al., 2015).

In the present study, we found a number of genes were differentially expressed in LUAD different stages. This indicated gene expression patterns were various with the LUAD development. Compared to normal lung tissue, a gene may be differentially expressed in early LUAD but not in advanced stage. We identified 422 LUAD-development characteristic genes, including 185 genes gradually upregulated and 237 genes gradually downregulated with LUAD-development. The development of LUAD results from synergistic effects of multiple genes. Notably, not all LUAD-development characteristic genes are associated with the prognosis of LUAD. LUAD-unfavorable gene set contained 84 gradually upregulated DEGs and LUAD-favorable gene set contained 39 gradually downregulated DEGs. Unsurprisingly, previous studies have suggested that some of them are associated with LUAD development. NEK2 is overexpressed in a variety of malignant tumors and is closely related to tumor drug resistance, rapid recurrence, and poor prognosis (Zhou et al., 2013; Fang and Zhang, 2016; Li et al., 2017). KIF14 has also been found to

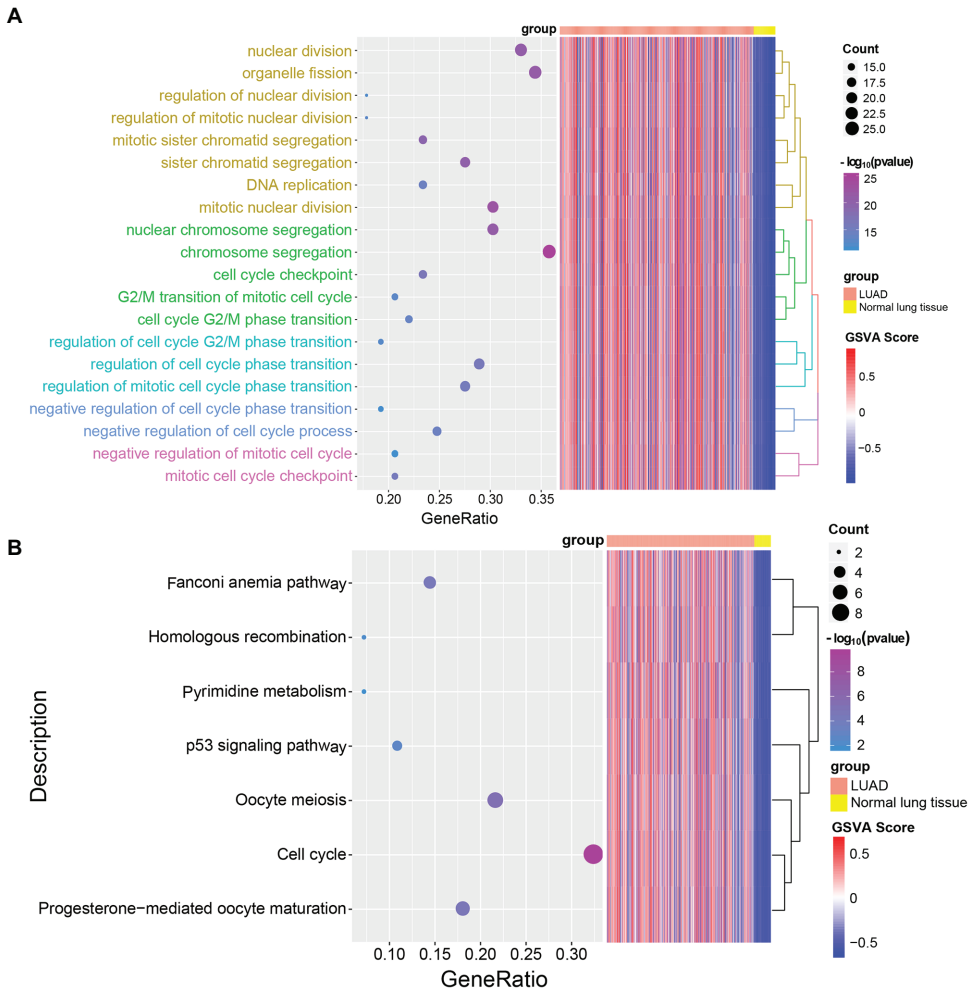

FIGURE 6 | Gene Ontology (GO) biological process (BP) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analysis of LUAD-unfavorable genes. **(A)** Biological process of LUAD-unfavorable genes. **(B)** KEGG pathway analysis of LUAD-unfavorable genes.

be associated with poor prognosis in a variety of cancers (O'Hare et al., 2016; Zhang et al., 2017). While in the LUAD-favorable gene set, genes which were significantly associated with LUAD survival included OR7E47P, MS4A2, RAB44, BMP5, ARHGEF6, and KLK11. Among them, KLK11 was found to be a diagnostic and prognostic indicator of NSCLC (Xu et al., 2014). These result confirmed the possibility that the LUAD-unfavorable gene set and LUAD-unfavorable gene set can be used as a prognostic model for LUAD.

All samples were calculated LUAD-unfavorable GSVA scores and LUAD-favorable GSVA scores. This is obviously different from the gene signatures in other previous studies (Li et al., 2014; Shi et al., 2018; Liu et al., 2019). In the previous studies, a gene often got a coefficient from a Cox regression analysis or other method in the training set. However, due to the limitations of the sample size and the heterogeneity of the tumor, we may never know the true coefficient of a gene. Therefore, GSVA was used to score individual samples against gene sets (LUAD-unfavorable gene set and LUAD-favorable gene set) in our study. ROC curve analysis suggested that both LUAD-unfavorable GSVA score and LUAD-favorable GSVA score exhibited strong diagnostic capacity of LUAD and which was verified in other two independent data sets. Univariate and multivariate Cox regression analysis suggested that LUAD-unfavorable GSVA score and LUAD-unfavorable gene set were independent prognostic factors for LUAD's overall survival. This result was also verified in an independent data set.

Moreover, we found that the mutation rate of most genes is very low, indicating that the differential expression of genes may not be caused by mutation. Additionally, functional enrichment analysis indicates that LUAD-unfavorable genes are significantly involved in p53 signaling pathway, Cell cycle, and other pathways. It is suggested that LUAD-unfavorable genes may be involved in the occurrence and development of LUAD through these pathways. However, further studies are needed to investigate and validate the functions of these genes.

In the present study, although we provided new insights into the LUAD prognostic stratification system, several limitations were notable. Firstly, the two gene sets may be too large. Their application to the clinic still needs to wait for further decline in sequencing costs. Secondly, the synergy between the genes of these two gene sets to promote LUAD development still requires molecular experimental validation.

CONCLUSION

In conclusion, we identified and validated two LUAD-development characteristic gene sets that not only have diagnostic value but also prognostic value. It may provide new insight for further research on LUAD.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

CL and XL conducted the experiments. HS and DL designed the experiments and wrote the paper. All authors contributed to the article and approved the submitted version.

FUNDING

The study was supported by Subject of Education Department of Heilongjiang Provincial (no. 12531258 and 12511264).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at https://www.researchsquare.com/article/rs-12465/v1 (Liu et al., 2020). We would like to thank the Bioinforcraftsman Technology Research and Development Co., Ltd for generously assisting with science research experience and bioinformatics analysis.

REFERENCES

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Colwill, K., Renewable Protein Binder Working Group, and Graslund, S. (2011). A roadmap to generate renewable protein binders to the human proteome. *Nat. Methods* 8, 551–558. doi: 10.1038/nmeth.1607

Cui, Y., Liu, J., Yin, H. B., Liu, Y. F., and Liu, J. H. (2015). Fibulin-1 functions as a prognostic factor in lung adenocarcinoma. *Jpn. J. Clin. Oncol.* 45, 854–859. doi: 10.1093/jjco/hyv094

Dama, E., Melocchi, V., Dezi, F., Pirroni, S., Carletti, R. M., Brambilla, D., et al. (2017). An aggressive subtype of stage I lung adenocarcinoma with molecular and prognostic characteristics typical of advanced lung cancers. *Clin. Cancer Res.* 23, 62–72. doi: 10.1158/1078-0432.CCR-15-3005

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., Mclellan, M. D., Cibulskis, K., et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 455, 1069–1075. doi: 10.1038/nature07423

Dong, H. X., Wang, R., Jin, X. Y., Zeng, J., and Pan, J. (2018). LncRNA DGCR5 promotes lung adenocarcinoma (LUAD) progression via inhibiting hsa-mir-22-3p. *J. Cell. Physiol.* 233, 4126–4136. doi: 10.1002/jcp.26215

Donner, I., Katainen, R., Sipila, L. J., Aavikko, M., Pukkala, E., and Aaltonen, L. A. (2018). Germline mutations in young non-smoking women with lung adenocarcinoma. *Lung Cancer* 122, 76–82. doi: 10.1016/j.lungcan.2018.05.027

Ernst, J., and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* 7:191. doi: 10.1186/1471-2105-7-191

Fang, Y., and Zhang, X. (2016). Targeting NEK2 as a promising therapeutic approach for cancer treatment. *Cell Cycle* 15, 895–907. doi: 10.1080/15384101.2016.1152430

Feng, A., Tu, Z., and Yin, B. (2016). The effect of HMGB1 on the clinicopathological and prognostic features of non-small cell lung cancer. *Oncotarget* 7, 20507–20519. doi: 10.18632/oncotarget.7050

Giatromanolaki, A., Kalamida, D., Sivridis, E., Karagounis, I. V., Gatter, K. C., Harris, A. L., et al. (2015). Increased expression of transcription factor EB (TFEB) is associated with autophagy, migratory phenotype and poor prognosis in non-small cell lung cancer. *Lung Cancer* 90, 98–105. doi: 10.1016/j.lungcan.2015.07.008

Govindan, R., Page, N., Morgensztern, D., Read, W., Tierney, R., Vlahiotis, A., et al. (2006). Changing epidemiology of small-cell lung cancer in the United States over the last 30 years: analysis of the surveillance, epidemiologic, and end results database. *J. Clin. Oncol.* 24, 4539–4544. doi: 10.1200/JCO.2005.04.4859

Guan, J. L., Zhong, W. Z., An, S. J., Yang, J. J., Su, J., Chen, Z. H., et al. (2013). KRAS mutation in patients with lung cancer: a predictor for poor prognosis but not for EGFR-TKIs or chemotherapy. *Ann. Surg. Oncol.* 20, 1381–1388. doi: 10.1245/s10434-012-2754-z

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7. doi: 10.1186/1471-2105-14-7

He, S. Y., Xi, W. J., Wang, X., Xu, C. H., Cheng, L., Liu, S. Y., et al. (2019). Identification of a combined RNA prognostic signature in adenocarcinoma of the lung. *Med. Sci. Monit.* 25, 3941–3956. doi: 10.12659/MSM.913727

Hecht, S. S. (1999). Tobacco smoke carcinogens and lung cancer. *J. Natl. Cancer Inst.* 91, 1194–1210. doi: 10.1093/jnci/91.14.1194

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15:R29. doi: 10.1186/gb-2014-15-2-r29

Li, X., Shi, Y., Yin, Z., Xue, X., and Zhou, B. (2014). An eight-miRNA signature as a potential biomarker for predicting survival in lung adenocarcinoma. *J. Transl. Med.* 12:159. doi: 10.1186/1479-5876-12-159

Li, G., Zhong, Y., Shen, Q., Zhou, Y., Deng, X., Li, C., et al. (2017). NEK2 serves as a prognostic biomarker for hepatocellular carcinoma. *Int. J. Oncol.* 50, 405–413. doi: 10.3892/ijo.2017.3837

Liu, C., Li, X., Shao, H., and Li, D. (2020). Identification and validation of two LUAD-development characteristic gene sets for diagnosing lung adenocarcinoma and predicting prognosis [Preprint]. doi: 10.21203/rs.2.21884/v1

Liu, C., Li, Y., Wei, M., Zhao, L., Yu, Y., and Li, G. (2019). Identification of a novel glycolysis-related gene signature that can predict the survival of patients with lung adenocarcinoma. *Cell Cycle* 18, 568–579. doi: 10.1080/15384101.2019.1578146

Mendelsohn, J., and Baselga, J. (2003). Status of epidermal growth factor receptor antagonists in the biology and treatment of cancer. *J. Clin. Oncol.* 21, 2787–2799. doi: 10.1200/JCO.2003.01.504

Mounir, M., Lucchetta, M., Silva, T. C., Olsen, C., Bontempi, G., Chen, X., et al. (2019). New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput. Biol.* 15:e1006701. doi: 10.1371/journal.pcbi.1006701

Naoki, K., Chen, T. H., Richards, W. G., Sugarbaker, D. J., and Meyerson, M. (2002). Missense mutations of the BRAF gene in human lung adenocarcinoma. *Cancer Res.* 62, 7001–7003.

O'Hare, M., Shadmand, M., Sulaiman, R. S., Sishtla, K., Sakisaka, T., and Corson, T. W. (2016). Kif14 overexpression accelerates murine retinoblastoma development. *Int. J. Cancer* 139, 1752–1758. doi: 10.1002/ijc.30221

Pang, B., Wu, N., Guan, R., Pang, L., Li, X., Li, S., et al. (2017). Overexpression of RCC2 enhances cell motility and promotes tumor metastasis in lung adenocarcinoma by inducing epithelial-mesenchymal transition. *Clin. Cancer Res.* 23, 5598–5610. doi: 10.1158/1078-0432.CCR-16-2909

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Scott, W. J., Howington, J., Feigenberg, S., Movsas, B., and Pisters, K., and American College of Chest Physicians (2007). Treatment of non-small cell lung cancer stage I and stage II: ACCP evidence-based clinical practice guidelines (2nd edition). *Chest* 132, 234S–242S. doi: 10.1378/chest.07-1378

Shi, X., Tan, H., Le, X., Xian, H., Li, X., Huang, K., et al. (2018). An expression signature model to predict lung adenocarcinoma-specific survival. *Cancer Manag. Res.* 10, 3717–3732. doi: 10.2147/CMAR.S159563

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136

Xu, P., Liu, L., Wang, J., Zhang, K., Hong, X., Deng, Q., et al. (2013). Genetic variation in BCL2 3'-UTR was associated with lung cancer risk and prognosis in male Chinese population. *PLoS One* 8:e72197. doi: 10.1371/journal.pone.0072197

Xu, C. H., Zhang, Y., and Yu, L. K. (2014). The diagnostic and prognostic value of serum human kallikrein-related peptidases 11 in non-small cell lung cancer. *Tumour Biol.* 35, 5199–5203. doi: 10.1007/s13277-014-1674-x

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, Y., Yuan, Y., Liang, P., Zhang, Z., Guo, X., Xia, L., et al. (2017). Overexpression of a novel candidate oncogene KIF14 correlates with tumor progression and poor prognosis in prostate cancer. *Oncotarget* 8, 45459–45469. doi: 10.18632/oncotarget.17564

Zhao, K., Li, Z., and Tian, H. (2018a). Twenty-gene-based prognostic model predicts lung adenocarcinoma survival. *OncoTargets Ther.* 11, 3415–3424. doi: 10.2147/OTT.S158638

Zhao, X., Zhou, L. L., Li, X., Ni, J., Chen, P., Ma, R., et al. (2018b). Overexpression of KIF20A confers malignant phenotype of lung adenocarcinoma by promoting cell proliferation and inhibiting apoptosis. *Cancer Med.* 7, 4678–4689. doi: 10.1002/cam4.1710

Zhou, W., Yang, Y., Xia, J., Wang, H., Salama, M. E., Xiong, W., et al. (2013). NEK2 induces drug resistance mainly through activation of efflux drug pumps and is associated with poor prognosis in myeloma and other cancers. *Cancer Cell* 23, 48–62. doi: 10.1016/j.ccr.2012.12.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.