Identification of a lncRNA-mRNA Regulatory Module for Prognosis, Classification and Potential Treatment of Lung Adenocarcinoma

Qin-Yu Zhao
The Third Xiangya Hospital of Central South University
Le-Ping Liu
The Third Xiangya Hospital of Central South University
Lu Lu
The Third Xiangya Hospital of Central South University
Rong Gui (guirong@csu.edu.cn)
The Third Xiangya Hospital of Central South University
Yan-Wei Luo
The Third Xiangya Hospital of Central South University

Research Article

Keywords: lung adenocarcinoma, lncRNA-mRNA regulatory module, survival prediction, machine learning, the connectivity map, molecular docking

Posted Date: June 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-567242/v1

Abstract

Background: Lung cancer remains the leading cause of cancer death worldwide, with lung adenocarcinoma (LUAD) being the most prevalent subtype of lung cancer. This study aimed to identify a lncRNA-mRNA regulatory module related to the prognosis, classification, and potential treatment of LUAD.

Methods: Publicly available gene expression data of three cohorts were downloaded from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) databases. Differential expression analysis between LUAD and normal samples, as well as the survival analysis, was performed. Protein-protein interaction (PPI) network and co-expression analyses were conducted to further identify key genes. A least absolute shrinkage and selection operator (LASSO) Cox regression model was developed to predict overall survival. Five machine learning models, including logistic regression, K-nearest neighbor (KNN), support vector machine (SVM), random forest, and extremely gradient boosting (XGBOOST), were trained to distinguish early-stage or epidermal growth factor receptor (EGFR)-mutation LUAD from others. Furthermore, connectivity map (CMap) and molecular docking analyses were performed to identify compounds with the ability to reverse the expression profiles of the key genes.

Results: A cohort comprised of 535 LUAD and 59 normal samples in TCGA was used as the training set, while the GSE31210 and GSE30219 datasets were used as validation sets. 189 mRNAs and 11 lncRNAs were differentially expressed and associated with the overall survival. 43 hub mRNAs were further identified from the PPI network, and 3 lncRNAs were significantly correlated to the expression of hub mRNAs. Six genes with nonzero coefficients were selected by using the LASSO COX regression analysis, and the corresponding risk score was derived. The time-dependent ROC and Kaplan Meier analysis demonstrated that the risk score accurately discriminates the patients with a high or low risk. KNN and XGBOOST were the best models to recognize early-stage and EGFR-mutation LUAD, respectively. Purvalanol-a and Etoposide obtained a score of -99.98 in CMap and were successfully docked with three key genes.

Conclusions: A lncRNA-mRNA regulatory module, including 4 mRNAs and 2 lncRNAs, was identified, and this module can facilitate the exploration of pathogenesis of LUAD and speed up the development of new treatments for LUAD patients.

Background

Despite the significant and accelerated decline in the mortality of lung cancer, it still remains the leading worldwide cause of cancer death in the combined population of women and men (1). According to the latest cancer statistics, 361 Americans per day are estimated to die from lung cancer in 2021; the expected deaths exceed the ones of breast cancer, prostate cancer and colorectal cancer combined (1). Small cell lung cancer and non-small cell lung cancer (NSCLC) are the two main subtypes of lung cancer. Lung adenocarcinoma (LUAD), the most prevalent subtype of NSCLC (2), comprises half of all lung cancer cases and is characterized by its low 5-year survival rate (3). These facts have highlighted the

need to identify novel risk genes with prognostic value and explore new potential therapeutic targets for LUAD.

In the last decade, numerous studies have attempted to predict and stratify prognosis of LUAD. An increasing number of genes associated with the survival of LUAD patients have been identified and validated, including immune-related genes (4, 5), glycolysis-related genes (6), hypoxia-related genes (7) and ferroptosis-related genes (8). Besides, long non-coding RNAs (lncRNAs) were also proved to play a critical role in the pathogenesis of LUAD and other various cancers (9–12). However, the research on integrated analysis of messenger RNA (mRNA) and lncRNA in LUAD still remains limited, although they are closely related in involved biological processes. Several studies attempted to develop a combined signature but some of them failed to validate their results in independent datasets (13, 14), while some only used machine learning methods to predict survival, limiting thus their clinical applicability (15, 16). Moreover, most of the studies aiming to identify prognosis-related gene signatures failed to explore the potential treatment of LUAD, partly limiting the clinical impact of their research findings. Therefore, deeper research is essential to promote the understanding of lncRNA-mRNA interactions and explore potential treatments for LUAD.

In this study, we aimed to use bioinformatic analysis and machine learning to identify a lncRNA-mRNA regulatory module associated with prognosis, classification and potential treatment of LUAD. As shown in Fig. 1, publicly available gene expression data of three cohorts were downloaded from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) databases. Differential expression analysis and survival analysis were performed to identify differentially expressed genes (DEGs) with prognostic value. Then, mRNAs and lncRNAs were further screened by using protein-protein interaction (PPI) network and co-expression analyses. A least absolute shrinkage and selection operator (LASSO) Cox regression model was developed to predict survival. Five machine learning models, including logistic regression, K-nearest neighbor (KNN), support vector machine (SVM), random forest and extremely gradient boosting (XGBOOST), were trained to distinguish early-stage or epidermal growth factor receptor (EGFR)-mutation LUAD from others. Furthermore, connectivity map (CMap) and molecular docking analyses were performed to identify several drugs that can reverse the expression profiles of the key genes demonstrating thus that they can serve as a potential treatment for LUAD.

Methods

Datasets

Raw read count data of gene expressions and clinical information were downloaded from TCGA, including 535 LUAD and 59 normal samples (17). Then, the read count data were normalized by using the voom function in the limma R package (version 3.42.2) (18). Besides, two independent cohorts from GEO were downloaded and analyzed for validation (GSE31210, GSE30219). mRNAs and lncRNAs were identified according to Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13), and only common genes in the three cohorts were analyzed.

Differential expression and survival analyses

The differentially expressed mRNAs (DEmRNAs) and lncRNAs (DElncRNAs) between LUAD and normal groups in the TCGA cohort were identified using the limma R package (18). Those with | log2 fold change | > 2 and adjusted p-value < 0.01 were prioritized as DEGs. Heatmaps and volcano maps were drawn to visualize the DEGs.

A univariate COX regression model was developed using the TCGA LUAD cohort to assess the prognostic value of each DEmRNA and DElncRNA using the survival R package (version 3.2-7). The genes significantly associated with the prognosis of LUAD ($p<0.01$), were selected for the subsequent analysis.

Additional filtering of mRNAs and lncRNAs

The DEmRNAs with prognostic value were uploaded to the STRING database (19) and a PPI network was constructed. We visualized the network using the Cytoscape software (20). In the network, each node represents a protein or a gene and an edge represents the interaction between two nodes. The degree of a node is the number of its connections with other nodes. Based on the PPI network, the MCODE algorithm was used to cluster the network into important modules and then the cytoHubba tool was used to calculate the degree value of each node. In line with the previous research, only the nodes with a degree value > 20 were considered as hub mRNAs (21).

The co-expression analysis was conducted by assessing the Pearson correlation between mRNAs and DElncRNAs with prognostic value for LUAD. A lncRNA-mRNA co-expression network was derived according to the criteria of | Correlation Coefficient | > 0.4 and P < 0.01 using the limma R package (18). The DElncRNAs that correlated to the hub mRNAs were screened out.

The hub mRNAs selected from the PPI network and the lncRNAs co-expressed with the hub mRNAs were then combined. It is noteworthy that these genes had been screened by univariate COX regression, and therefore, were associated with prognosis.

COX regression

A LASSO Cox regression model was developed based on the combined genes using the glmnet R package (version 4.0-2). The genes with a regression coefficient of zero were removed, and then a risk score was generated from the regression model. The genes included in the risk score were considered as the final lncRNA-mRNA regulatory module. The formula of the risk score is:

$$RiskScore = coef_1 \times gene_1 + coef_2 \times gene_2 + coef_3 \times gene_3 + \cdots$$

where the gene represents the normalized expression of a given mRNA or lncRNA in the final model and the coef represents its coefficient in the LASSO COX regression model.

LUAD patients were split into high- and low-risk groups according to whether their risk scores were greater than the median risk score in TCGA. The Kaplan Meier analysis with a log-rank test was applied to evaluate the prognostic difference between the two groups. The time-dependent receiver operator characteristic curve (ROC) was used to assess the accuracy of the suggested risk score. Univariate and multivariate COX regression analyses were successively performed among the risk score and available clinical variables to determine whether the risk score is an independent predictor of survival. In addition, the risk score was validated in two GEO cohorts.

Further, gene set enrichment analysis using the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) was conducted to assess the DEGs between high- and low-risk groups, using the clusterProfiler R package (version 3.14.3) (22). The infiltrating score of 16 immune cells and the activity of 13 immune-related pathways were calculated with the single-sample gene set enrichment analysis (ssGSEA) (23) using the GSVA R package (version 1.34.0) (24). The annotated gene set file was obtained from previous research (25).

Machine learning

We designed two classification tasks to explore the ability of the final regulatory module to discriminate different types of LUAD by using machine learning models. Five representative models were assessed, including logistic regression (26), KNN (27), SVM (28), random forest (29) and XGBOOST (30, 31). In the first task, these five machine learning models were applied in the TCGA cohort to build classifiers for the discrimination of early-stage (pathological Stage I-II) LUAD from others (pathological Stage III-IV). In the second task, these models were trained to recognize LUAD with EGFR mutation, based on the GSE31210 cohort. Only the genes of the final regulatory module were used as features in both tasks.

Ten-fold cross validation was performed taking into consideration the limited sample size (31), randomly splitting the dataset into 10 subsets, and using in each iteration, 9 of them to train the models and last 1 for validation. After 10 iterations, each subset had been validated and the validation results were combined to robustly assess the model performance. Then, the SHapley Additive exPlanation (SHAP) values were used to illustrate the classification results of the XGBOOST model, according to a game theory approach (32).

A web-based tool to use our score and models

A web-based tool was developed using HTML and javascript language, and the back-end of this tool was managed using Django python package. When a user enters the normalized expression data of genes, the risk score and classification results of machine learning models are calculated and provided. This tool will allow clinicians or cancer patients to learn more about the condition and benefit from the results of our study.

Connectivity map-based drug screening and molecular docking

The L1000-based next-generation CMap (33) (Touchstone v1) was used to assess the functional connections between drugs and genes. The L1000-based CMap expanded the original CMap by using nearly 28,000 perturbagens including over 19,000 small molecules and about 7000 genetic modulations (33). The hub mRNAs were split into upregulated and downregulated gene groups, and were uploaded to the CMap web server. The LUAD cell line A549 and all compounds in the database were chosen for this analysis, and a list of compounds with connectivity scores ranging from –100 to 100 was obtained. In brief, a positive score indicates the similarity between a compound's signature and the signature of the query, while a negative score indicates that the two signatures are opposing (21, 34). The magnitude of the score corresponds to the magnitude of similarity or dissimilarity. Therefore, compounds with a connectivity score of -90 or lower were considered as potential drugs for LUAD.

Molecular docking was used to verify the targeting association between the compounds and the revealed proteins. The 3-dimension models of targeted proteins and compounds were downloaded from the RCSB PDB and the ZINC database, respectively. Pre-processing such as removing solvent, adding hydrogens and choosing torsion, was conducted using the AutoDock Tool (version 1.5.6). Molecular docking was simulated by the AutoDock Vina program (version 1.1.2) (35). The interaction energy between the ligand and the receptor was calculated for the entire binding site and expressed as affinity. In line with previous research, a docking result with an affinity score < -5 kcal/mol is considered a strong blinding between the compound and protein (36). The docking results were visualized by the PyMOL program (version 2.4).

Results

Differentially expressed mRNAs and lncRNAs associated with prognosis

Expression data and clinical information of three cohorts were downloaded from TCGA and GEO. The baseline characteristics were summarized in Table 1. The first cohort consisted of 535 LUAD and 59 normal samples from TCGA, and was used for differential expression analysis. 17122 mRNAs and 1826 lncRNAs were screened, and 1456 mRNAs and 92 lncRNAs were differentially expressed between LUAD and normal groups. The expression profiles of the DEmRNAs and DElncRNAs were visualized in Fig. 2 in the form of heatmaps and volcano maps. As shown in Fig. 2A&C, 606 and 850 mRNAs were upregulated and downregulated in LUAD samples respectively while Fig. 2B&D shows the 49 upregulated and the 43 downregulated lncRNAs respectively.

189 DEmRNAs and 11 DElncRNAs were associated with the overall survival of LUAD patients using univariate COX regression. The genes with the 10 highest and the 10 lowest Hazard Ratios (HRs) were reported in Table S1 (Supplementary Materials).

Additional filtering of mRNAs and lncRNAs

A PPI network was constructed by uploading the 189 prognosis-associated DEmRNAs on the STRING database. Fig. 3A depicts the derived PPI network, which was visualized by the Cytoscape program and contained 77 nodes and 919 edges. Then, 2 modules were identified using the MCODE algorithm (shown

in Fig. 3B-C). 43 mRNAs were further characterized as hub genes by the cytoHubba tool. Co-expression analysis of mRNAs and lncRNAs was also performed, as shown in Fig. 3D. 3 lncRNAs were significantly correlated to the expression of hub mRNAs. As a result, 46 genes (43 hub mRNAs and 3 lncRNAs) were used for the following risk modeling analysis.

Risk scores derived from COX regression

6 genes with nonzero coefficients were identified in the LASSO COX regression model. The risk score was calculated for each sample using the following formula:

RiskScore= 0.08443×DLGAP5 + 0.02479×MKI67 + 0.00070×TOP2A + 0.09877×CCNB1 + 0.05551×PICSAR + 0.10420×SCAT1

Samples were classified into high- and low-risk groups by comparing their risk scores to the median score in the TCGA cohort. As shown in Fig. 4A-C, the areas under the time-dependent ROC (AUROC) of TCGA are 0.690, 0.687, 0.673 and 0.726 for 1-, 2-, 3- and 4-years survival, respectively. The predicted risk scores presented better performance in the GSE31210 cohort (AUROC: 0.710, 0.696, 0.723 and 0.749 for 5, 6, 7 and 8 years, respectively) and the GSE30219 cohort (AUROC: 0.699, 0.696, 0.731 and 0.726 for 1, 2, 3 and 4 years, respectively). It is noteworthy that a different time interval was assessed for GSE31210 because LUAD was relatively mild (mostly in stage I) and survival time was generally longer in this cohort. Additionally, the Kaplan Meier analysis with log-rank tests revealed a significant difference in the overall survival ($p<0.01$) between the two groups in the TCGA cohort as well as in the two validation cohorts (shown in Fig. 4D-F). Moreover, as shown in Fig. 4G-I, the proposed risk score was proven to be an independent prognostic factor for LUAD ($HR>1$, $p<0.05$) by successively using univariate and multivariate COX regression.

GO and KEGG pathway enrichment analyses were performed on the DEGs between the high- and low-risk groups to elucidate the biological functions and pathways associated with the risk score. As shown in Fig. 5A-B, the DEGs between different risk groups are mainly enriched in the functions associated with mitosis and cell cycle, such as DNA replication, chromosome segregation and spindle. KEGG pathway enrichment analysis (Fig. 5D-E) also revealed that DNA replication and cell cycle pathways were enriched. Interestingly, the DEGs are significantly enriched in pathways of neurodegeneration and related diseases, such as Huntington disease and Alzheimer disease.

Fig. 5C&F shows the correlation between the risk score and immune status. As seen, the scores of aDCs, pDCs, APC co-stimulation, and MHC class I, are significantly different between the low-risk and high-risk groups. Most of the immune functions are significantly higher expressed in the high-risk group with the exception of the Type II IFN Response.

Machine learning modeling

The distributions of gene expression and risk scores in the different stages of LUAD are presented in Fig. 6A. Gene expression and risk scores were all significantly higher in pathological Stage III-IV LUAD,

suggesting poorer survival for these LUAD stages. Five machine learning models, including logistic regression, KNN, SVM, random forest and XGBOOST, were applied to classify early-stage LUAD (pathological Stage I-II) from others (pathological Stage III-IV) based on the 6 risk genes. The AUROCs were assessed and are shown in Fig. 6B. The KNN model presented the highest AUROC (0.790) outperforming the other machine learning models. Fig. 6C shows the SHAP illustration of the XGBOOST model. Blue through red color indicates low to high expression level. SHAP values depict the effect of gene expression on the classification model. A positive SHAP value means that gene expression leads to an increase in the possibility of being Stage I-II, while a negative value represents that gene expression reduces this possibility.

Interestingly, the 6 genes' expression and risk scores were lower in EGFR mutation than other LUAD. Five machine learning models were trained to recognize EGFR mutation, and the XGBOOST model was the one that presented the best performance (AUROC: 0.778). The AUROCs and SHAP values are shown in Fig. 6F&G.

The web-based tool

A web-based tool was developed (http://www.aimedicallab.com/tool/aiml-luad.html), and its interface is shown in Fig. 6D&H. Users only need to enter the normalized expression data of the 6 genes for each patient and click the 'predict' button. Then, the risk score and classification results of the machine learning models will be calculated and shown. Note that the early-stage classification was performed by KNN and the EGFR mutation was recognized by XGBOOST, since these were the best performing models in the two classification problems, respectively. Fig. 6D shows the input and results of a high-risk patient from the TCGA cohort; he was 71 years old with stage-III LUAD and survived only about 2 months in the study. In contrast, Fig. 6E is an example of low-risk patients. Her condition was more moderate (Stage I) and she survived more than 6 years.

CMap analysis and molecular docking

The forty-three hub mRNAs were all upregulated in LUAD samples and were uploaded to the CMap web server. A connectivity list of thousand compounds was obtained, with 20 of them at the top or bottom presented in Table S2 (Supplementary Materials). Four compounds had a score of 100 or -100, including SA-792541, doxorubicin, JW-7-24-1 and CD-437. The scores of mirtazapine, teniposide were 99.99 and -99.99 respectively. The other 14 compounds, including quizartinib, terreic-acid, SN-38 and purvalanol-a, presented a score of 99.98 or -99.98.

Molecular docking was conducted using the AutoDock Vina. The affinity scores of TOP2A-Purvalanola, Ki67-Etoposide and CCNB1-Etoposide were -6.7, -6.4 and -6.8 kcal/mol, respectively. These scores were all less than -5 kcal/mol suggesting, therefore, a strong blinding. The docking results are visualized using the PyMOL program and they are presented in Fig. 7.

Discussion

A lncRNA-mRNA regulatory module associated with prognosis, staging, EGFR mutation, and a potential therapeutic target for LUAD, was identified in the present study. The risk score derived from this module presented high AUROCs across the three independent cohorts mined from TCGA and GEO. LUAD patients were classified into low- and high-risk groups which presented a significant difference in the overall survival. Early-stage and EGFR mutation were accurately recognized by the 5 machine learning models. CMap and molecular docking analysis provided a prioritized list of drugs that may be useful for treating LUAD.

The identified genes, including 4 mRNAs (DLGAP5, MKI67, TOP2A, and CCNB1) and 2 lncRNAs (PICSAR and SCAT1), are significantly related to the prognosis of LUAD. Previous studies have demonstrated that these genes are involved in the pathogenesis of LUAD. For instance, the prognostic value of DLGAP5 has been assessed in other cancers, such as colorectal cancer (37), pancreatic cancer (38) and breast cancer (39). Prior studies on LUAD have also identified prognostic signatures including DLGAP5 (40); however, little is known about how this gene affects the development of LUAD (41, 42). MKI67 is the protein coding gene for the Ki67 protein, and this protein is known for decades to be a proliferation marker for human tumor cells. In fact, Ki-67 has roles in both interphase and mitotic cells (43) and was found to be associated with survival outcomes in lung cancer (44, 45). TOP2A was found to be broadly expressed in many types of cancers (46). Several studies have reported that TOP2A expression was aberrantly upregulated in LUAD, targeting cell proliferation, migration and invasion (46, 47), and associated with poor prognosis (48). Wenxia Ma, et al. have found that TOP2A can potentially regulate the development of NSCLC working together with TPX2 (49). Shweta Arora, et al. reported that CCNB1 and other genes target cell cycle and immunosurveillance mechanisms via HMGA2 and E2F7 in NSCLC (50). Besides, Fan Kou, et al reported that inhibition of TOP2A reduces the expression levels of CCNB1 and CCNB2, which indicates that TOP2A targeting CCNB1 and CCNB2 promotes GLC82 and A549 cells proliferation and metastasis. Our study confirmed the important role that the 4 mRNAs have in LUAD development.

The two lncRNAs included in our signature, namely, PICSAR and SCAT1 were also associated with the development and prognosis of cancers. It was reported that PICSAR promotes the growth of cutaneous squamous cell carcinoma by regulating ERK1/2 activity (51) and PICSAR/miR-588/EIF6 axis regulates tumorigenesis of hepatocellular carcinoma by activating PI3K/AKT/mTOR signaling pathway (52). Besides, Chaoqi Zhang, et al have identified a three-lncRNA signature from pretreatment biopsies, including SCAT1 to predict pathological response and outcome in esophageal squamous cell carcinoma with neoadjuvant chemoradiotherapy (53). However, PICSAR and SCAT1 have received much less attention in LUAD research. In the present study, two lncRNAs were found to have the potential to be prognostic biomarkers and act as potential therapeutic targets for LUAD. This result should be further validated in prospective studies and may help in the stratification and treatment of LUAD patients.

A risk score was developed to predict LUAD survival based on the selected prognostic genes and the cohorts were accordingly divided into high- and low-risk groups. Kaplan Meier analysis demonstrated that there was a significant difference in the overall survival between the two risk groups. The time-dependent ROC and COX regression proved that the risk score was a valuable and independent prognostic factor for

LUAD. Therefore, this risk score can help predict and stratify the prognosis for LUAD, and also assist other future studies.

GO and KEGG pathway enrichment analyses were performed to further elucidate the differences between high- and low-risk patients. DNA replication, chromosome segregation, cell cycle G2/M phase transition were significantly enriched functions, as shown in Fig. 5. KEGG enrichment analysis also confirmed that homologous recombination, DNA replication, cell cycle were the significantly enriched pathways. Interestingly, pathways of neurodegeneration, Huntington disease and Alzheimer disease were also enriched. In fact, epidemiological studies indicated that patients suffering from Alzheimer's disease had a lower risk of developing lung cancer (54, 55). Jon Sánchez-Valle, et al have found inverse patterns of expression in Alzheimer's disease and lung cancer by using transcriptomic meta-analyses (56). However, little is understood about the inverse relationship between lung cancer and neurodegenerative disease. More research is needed to shed light on it.

Additionally, the correlation between the risk score and immune status was assessed. Compared with low-risk group, the high-risk group had significantly higher scores of B cells, CD8 + T cells, Macrophages, pDCs, Tfh, Th1 cells, Th2 cells, and Treg. In contrast, the scores of aDCs, Mast cells, and Neutrophils were significantly lower in the high-risk group. The present study confirmed the conclusion of previous research that tumor-infiltrating immune cells affect LUAD outcome (57–59). Besides, the scores of immune functions were significantly higher in the high-risk group, with the exception of cytolytic activity, HLA and Type II IFN response. These findings suggest that immune status was aberrantly changed in high-risk LUAD compared to low-risk group. As the future work, immunotherapy guided by risk scores targeting the abnormal immune status can be a promising treatment for LUAD patients.

Besides risk stratification, the regulatory module was found to be differentially expressed in different stages of LUAD or LUAD with EGFR mutation, in relation to pathogenesis and prognosis. In this study, five machine learning models were developed and applied to recognize early-stage LUAD or EGFR mutation in LUAD, based on the 6 selected genes. Despite using a limited gene number, the models presented a remarkable performance (AUROC > 0.75). Models based on broader sets of genes were also developed to further assess the potential of machine learning. The highest AUROCs of the 5 models were 0.815 for early-stage classification and 0.806 for EGFR mutation recognition based on the 43 hub mRNAs. The AUROCs exceeded 0.85 in both classification tasks based on all DEmRNAs and DElncRNAs with prognostic potential. However, these models and results are not shown in the present study, in consideration of the inconvenience and high cost in clinical practice. Currently, gene signatures comprised of several or a dozen of genes are still the research focus. Integrated analysis and clinical application of genomics using advanced algorithms are awaiting technical development and more studies.

CMap and molecular docking were performed to find potential drugs with the ability to reverse the expression profiles of the key genes. Among the drugs with the top connectivity scores, several have already been studied in cancer treatment. For example, doxorubicin has been used to treat various types

of cancers. Studies were conducted to overcome the toxic side effects and multidrug resistance during doxorubicin chemotherapy by using nanomedicine and combination treatment (60, 61). Teniposide and etoposide have been established as active compounds in the treatment of small cell lung cancer (62). The occurrence of these drugs in the table showed that the CMap results were reasonable. Furthermore, since the affinity scores were all below − 5 kcal/mol, this indicated a strong blinding between the compounds and selected genes.

Several limitations of this study should be considered. Firstly, data used in the bioinformatic analysis were obtained from public datasets, without verification of *in vitro* or *in vivo* biochemical experiments. Thus, the robustness of the six-gene signature requires further validation in large-scale prospective investments. Secondly, only mRNA and lncRNA were analyzed in our study. Multi-omics data may help us deeply understand the pathogenesis and better predict survival for LUAD. Lastly, potential drugs were selected by using CMap and molecular docking, which cannot replace biochemical experiments or clinical trials. There is still a long way to go before using these potential drugs.

Conclusions

In this study, a lncRNA-mRNA regulatory module, including 4 mRNAs and 2 lncRNAs, was identified, which is related to prognosis, staging, EGFR mutation and potential treatment of LUAD.

Abbreviations

NSCLC: small cell lung cancer and non-small cell lung cancer; LUAD: lung adenocarcinoma; lncRNAs: long non-coding RNAs; mRNA: messenger RNA; TCGA: the cancer genome atlas; GEO: gene expression omnibus; DEG: differentially expressed gene; PPI: protein-protein interaction; LASSO: least absolute shrinkage and selection operator; KNN: k-nearest neighbor; SVM: support vector machine; XGBOOST: extremely gradient boosting; EGFR: epidermal growth factor receptor; CMap: connectivity map; GRCh38.p13: genome reference consortium human build 38 patch release 13; DEmRNAs: differentially expressed mRNAs; DElncRNAs: differentially expressed lncRNAs; ROC: receiver operator characteristic curve; GO: gene ontology; KEGG: kyoto encyclopedia of genes and genomes; ssGSEA: single-sample gene set enrichment analysis; SHAP: shapley additive explanation; HRs: hazard ratios; AUROC: the areas under receiver operator characteristic curve.

Declarations

Acknowledgements

The authors would like to express their gratitude to EditSprings (https://www.editsprings.com/) for the expert linguistic services provided.

Funding

This study was supported by the National Natural Science Foundation of China (No. 81573091 and No. 81802668), the Natural Science Foundation of Hunan Province (No. 2018JJ3776 and No. 2017JJ3467), and the Fundamental Research Funds for the Central Universities of Central South University under Grant (No. 2020zzts892).

Authors' contributions

(I) Conception and design: Qin-Yu Zhao and Le-Ping Liu

(II) Administrative support: Rong Gui and Yan-Wei Luo

(III) Collection and assembly of data: Qin-Yu Zhao

(IV) Data analysis and interpretation: Qin-Yu Zhao and Le-Ping Liu

(V) Manuscript writing: All authors

(VI) Final approval of manuscript: All authors

Availability of data and materials

TCGA data were available on the project website at https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga, while the GEO datasets were available at https://www.ncbi.nlm.nih.gov/geo/.

Ethics approval and consent to participate

The study was an analysis of third-party anonymized publicly available datasets with pre-existing institutional review board (IRB) approvals.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA Cancer J Clin. 2021;71(1):7-33.
2. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyo D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24(10):1559-67.

3. Denisenko TV, Budkevich IN, Zhivotovsky B. Cell death-based treatment of lung adenocarcinoma. Cell Death Dis. 2018;9(2):117.
4. Zhang M, Zhu K, Pu H, Wang Z, Zhao H, Zhang J, et al. An Immune-Related Signature Predicts Survival in Patients With Lung Adenocarcinoma. Front Oncol. 2019;9:1314.
5. Shi X, Li R, Dong X, Chen AM, Liu X, Lu D, et al. IRGS: an immune-related gene classifier for lung adenocarcinoma prognosis. J Transl Med. 2020;18(1):55.
6. Zhang L, Zhang Z, Yu Z. Identification of a novel glycolysis-related gene signature for predicting metastasis and survival in patients with lung adenocarcinoma. J Transl Med. 2019;17(1):423.
7. Mo Z, Yu L, Cao Z, Hu H, Luo S, Zhang S. Identification of a Hypoxia-Associated Signature for Lung Adenocarcinoma. Front Genet. 2020;11:647.
8. Gao X, Tang M, Tian S, Li J, Liu W. A ferroptosis-related gene signature predicts overall survival in patients with lung adenocarcinoma. Future Oncol. 2021.
9. Dong HX, Wang R, Jin XY, Zeng J, Pan J. LncRNA DGCR5 promotes lung adenocarcinoma (LUAD) progression via inhibiting hsa-mir-22-3p. J Cell Physiol. 2018;233(5):4126-36.
10. Pan J, Fang S, Tian H, Zhou C, Zhao X, Tian H, et al. lncRNA JPX/miR-33a-5p/Twist1 axis regulates tumorigenesis and metastasis of lung cancer by activating Wnt/beta-catenin signaling. Mol Cancer. 2020;19(1):9.
11. Guo H, Feng Y, Yu H, Xie Y, Luo F, Wang Y. A novel lncRNA, loc107985872, promotes lung adenocarcinoma progression via the notch1 signaling pathway with exposure to traffic-originated PM2.5 organic extract. Environ Pollut. 2020;266(Pt 1):115307.
12. Zhou H, Zhang H, Chen J, Cao J, Liu L, Guo C, et al. A seven-long noncoding RNA signature predicts relapse in patients with early-stage lung adenocarcinoma. J Cell Biochem. 2019;120(9):15730-9.
13. Li L, Peng M, Xue W, Fan Z, Wang T, Lian J, et al. Integrated analysis of dysregulated long non-coding RNAs/microRNAs/mRNAs in metastasis of lung adenocarcinoma. J Transl Med. 2018;16(1):372.
14. Li S, Cui Z, Zhao Y, Ma S, Sun Y, Li H, et al. Candidate lncRNA-microRNA-mRNA networks in predicting non-small cell lung cancer and related prognosis analysis. J Cancer Res Clin Oncol. 2020;146(4):883-96.
15. He SY, Xi WJ, Wang X, Xu CH, Cheng L, Liu SY, et al. Identification of a Combined RNA Prognostic Signature in Adenocarcinoma of the Lung. Med Sci Monit. 2019;25:3941-56.
16. Zhao J, Cheng W, He X, Liu Y, Li J, Sun J, et al. Construction of a specific SVM classifier and identification of molecular markers for lung adenocarcinoma based on lncRNA-miRNA-mRNA network. Onco Targets Ther. 2018;11:3129-40.
17. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68-77.
18. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

19. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45(D1):D362-D8.
20. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27(3):431-2.
21. Zhang J, Liu H, Zhang W, Li Y, Fan Z, Jiang H, et al. Identification of lncRNA-mRNA Regulatory Module to Explore the Pathogenesis and Prognosis of Melanoma. Front Cell Dev Biol. 2020;8:615671.
22. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284-7.
23. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015;160(1-2):48-61.
24. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
25. Liang JY, Wang DS, Lin HC, Chen XX, Yang H, Zheng Y, et al. A Novel Ferroptosis-related Gene Signature for Overall Survival Prediction in Patients with Hepatocellular Carcinoma. Int J Biol Sci. 2020;16(13):2430-41.
26. Zhang Z. Model building strategy for logistic regression: purposeful selection. Ann Transl Med. 2016;4(6):111.
27. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016;4(11):218.
28. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565-7.
29. Sapir-Pichhadze R, Kaplan B. Seeing the Forest for the Trees: Random Forest Models for Predicting Survival in Kidney Transplant Recipients. Transplantation. 2020;104(5):905-6.
30. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O, written on behalf of AMEB-DCTCG. Predictive analytics with gradient boosting in clinical medicine. Ann Transl Med. 2019;7(7):152.
31. Badillo S, Banfai B, Birzele F, Davydov, II, Hutchinson L, Kam-Thong T, et al. An Introduction to Machine Learning. Clin Pharmacol Ther. 2020;107(4):871-85.
32. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell. 2020;2(1):56-67.
33. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6):1437-52 e17.
34. Zhang L, Zhou Y, Zhang J, Chang A, Zhuo X. Screening of hub genes and prediction of putative drugs in arsenic-related bladder carcinoma: An in silico study. J Trace Elem Med Biol. 2020;62:126609.
35. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455-61.
36. Mishra A, Dey S. Molecular Docking Studies of a Cyclic Octapeptide-Cyclosaplin from Sandalwood. Biomolecules. 2019;9(11).

37. Branchi V, Garcia SA, Radhakrishnan P, Gyorffy B, Hissa B, Schneider M, et al. Prognostic value of DLGAP5 in colorectal cancer. Int J Colorectal Dis. 2019;34(8):1455-65.
38. Ke MJ, Ji LD, Li YX. Bioinformatics analysis combined with experiments to explore potential prognostic factors for pancreatic cancer. Cancer Cell Int. 2020;20:382.
39. Xu T, Dong M, Li H, Zhang R, Li X. Elevated mRNA expression levels of DLGAP5 are associated with poor prognosis in breast cancer. Oncol Lett. 2020;19(6):4053-65.
40. Li S, Xuan Y, Gao B, Sun X, Miao S, Lu T, et al. Identification of an eight-gene prognostic signature for lung adenocarcinoma. Cancer Manag Res. 2018;10:3383-92.
41. Schneider MA, Christopoulos P, Muley T, Warth A, Klingmueller U, Thomas M, et al. AURKA, DLGAP5, TPX2, KIF11 and CKAP5: Five specific mitosis-associated genes correlate with poor prognosis for non-small cell lung cancer patients. Int J Oncol. 2017;50(2):365-72.
42. Shi YX, Yin JY, Shen Y, Zhang W, Zhou HH, Liu ZQ. Genome-scale analysis identifies NEK2, DLGAP5 and ECT2 as promising diagnostic and prognostic biomarkers in human lung cancer. Sci Rep. 2017;7(1):8072.
43. Sun X, Kaufman PD. Ki-67: more than a proliferation marker. Chromosoma. 2018;127(2):175-86.
44. Grant L, Banerji S, Murphy L, Dawe DE, Harlos C, Myal Y, et al. Androgen Receptor and Ki67 Expression and Survival Outcomes in Non-small Cell Lung Cancer. Horm Cancer. 2018;9(4):288-94.
45. Wei DM, Chen WJ, Meng RM, Zhao N, Zhang XY, Liao DY, et al. Augmented expression of Ki-67 is correlated with clinicopathological characteristics and prognosis for lung cancer patients: an updated systematic review and meta-analysis with 108 studies and 14,732 patients. Respir Res. 2018;19(1):150.
46. Kou F, Sun H, Wu L, Li B, Zhang B, Wang X, et al. TOP2A Promotes Lung Adenocarcinoma Cells' Malignant Progression and Predicts Poor Prognosis in Lung Adenocarcinoma. J Cancer. 2020;11(9):2496-508.
47. Du X, Xue Z, Lv J, Wang H. Expression of the Topoisomerase II Alpha (TOP2A) Gene in Lung Adenocarcinoma Cells and the Association with Patient Outcomes. Med Sci Monit. 2020;26:e929120.
48. Guo W, Sun S, Guo L, Song P, Xue X, Zhang H, et al. Elevated TOP2A and UBE2C expressions correlate with poor prognosis in patients with surgically resected lung adenocarcinoma: a study based on immunohistochemical analysis and bioinformatics. J Cancer Res Clin Oncol. 2020;146(4):821-41.
49. Ma W, Wang B, Zhang Y, Wang Z, Niu D, Chen S, et al. Prognostic significance of TOP2A in non-small cell lung cancer revealed by bioinformatic analysis. Cancer Cell Int. 2019;19:239.
50. Arora S, Singh P, Rahmani AH, Almatroodi SA, Dohare R, Syed MA. Unravelling the Role of miR-20b-5p, CCNB1, HMGA2 and E2F7 in Development and Progression of Non-Small Cell Lung Cancer (NSCLC). Biology (Basel). 2020;9(8).
51. Piipponen M, Nissinen L, Farshchian M, Riihila P, Kivisaari A, Kallajoki M, et al. Long Noncoding RNA PICSAR Promotes Growth of Cutaneous Squamous Cell Carcinoma by Regulating ERK1/2 Activity. J Invest Dermatol. 2016;136(8):1701-10.

52. Liu Z, Mo H, Sun L, Wang L, Chen T, Yao B, et al. Long noncoding RNA PICSAR/miR-588/EIF6 axis regulates tumorigenesis of hepatocellular carcinoma by activating PI3K/AKT/mTOR signaling pathway. Cancer Sci. 2020;111(11):4118-28.
53. Zhang C, Zhang Z, Zhang G, Xue L, Yang H, Luo Y, et al. A three-lncRNA signature of pretreatment biopsies predicts pathological response and outcome in esophageal squamous cell carcinoma with neoadjuvant chemoradiotherapy. Clin Transl Med. 2020;10(4):e156.
54. Musicco M, Adorni F, Di Santo S, Prinelli F, Pettenati C, Caltagirone C, et al. Inverse occurrence of cancer and Alzheimer disease: a population-based incidence study. Neurology. 2013;81(4):322-8.
55. Ou SM, Lee YJ, Hu YW, Liu CJ, Chen TJ, Fuh JL, et al. Does Alzheimer's disease protect against cancers? A nationwide population-based study. Neuroepidemiology. 2013;40(1):42-9.
56. Sanchez-Valle J, Tejero H, Ibanez K, Portero JL, Krallinger M, Al-Shahrour F, et al. A molecular hypothesis to explain direct and inverse co-morbidities between Alzheimer's Disease, Glioblastoma and Lung cancer. Sci Rep. 2017;7(1):4474.
57. Mansuet-Lupo A, Alifano M, Pecuchet N, Biton J, Becht E, Goc J, et al. Intratumoral Immune Cell Densities Are Associated with Lung Adenocarcinoma Gene Alterations. Am J Respir Crit Care Med. 2016;194(11):1403-12.
58. Lavin Y, Kobayashi S, Leader A, Amir ED, Elefant N, Bigenwald C, et al. Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses. Cell. 2017;169(4):750-65 e17.
59. Joshi NS, Akama-Garren EH, Lu Y, Lee DY, Chang GP, Li A, et al. Regulatory T Cells in Tumor-Associated Tertiary Lymphoid Structures Suppress Anti-tumor T Cell Responses. Immunity. 2015;43(3):579-90.
60. Hong Y, Che S, Hui B, Yang Y, Wang X, Zhang X, et al. Lung cancer therapy using doxorubicin and curcumin combination: Targeted prodrug based, pH sensitive nanomedicine. Biomed Pharmacother. 2019;112:108614.
61. Lu G, Cao L, Zhu C, Xie H, Hao K, Xia N, et al. Improving lung cancer treatment: Hyaluronic acidmodified and glutathioneresponsive amphiphilic TPGSdoxorubicin prodrugentrapped nanoparticles. Oncol Rep. 2019;42(1):361-9.
62. Splinter TA, Sahmoud T, Festen J, van Zandwijk N, Sorenson S, Clerico M, et al. Two schedules of teniposide with or without cisplatin in advanced non-small-cell lung cancer: a randomized study of the European Organization for Research and Treatment of Cancer Lung Cancer Cooperative Group. J Clin Oncol. 1996;14(1):127-34.

Tables

Table 1. Clinical characteristics of the three cohorts included in this study

		TCGA	GSE31210	GSE30219
LUAD Samples		535	226	85
Age, mean (SD)		65.56 (10.09)	59.58 (7.40)	61.49 (9.28)
Gender, n (%)	Male	249 (46.54)	105 (46.46)	66 (77.65)
Ever-smoker, n (%)	0.0	392 (82.35)	115 (50.88)	-
	1.0	84 (17.65)	111 (49.12)	-
Stage, n (%)	I	294 (55.79)	168 (74.34)	71 (83.53)
	II	123 (23.34)	58 (25.66)	13 (15.29)
	III	84 (15.94)	0 (0.00)	1 (1.18)
	IV	26 (4.93)	0 (0.00)	0 (0.00)
T, n (%)	1	175 (32.71)	-	71 (83.53)
	2	289 (54.02)	-	12 (14.12)
	3	49 (9.16)	-	2 (2.35)
	4	19 (3.55)	-	0 (0.00)
	X	3 (0.56)	-	0 (0.00)
N, n (%)	0	348 (65.17)	-	82 (96.47)
	1	95 (17.79)	-	3 (3.53)
	2	74 (13.86)	-	0 (0.00)
	3	2 (0.37)	-	0 (0.00)
	X	15 (2.81)	-	0 (0.00)
M, n (%)	0	361 (68.24)	-	85 (100.00)
	1	25 (4.73)	-	0 (0.00)
	X	143 (27.03)	-	0 (0.00)

TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; LUAD, lung adenocarcinoma; SD, standard deviation.

Figures

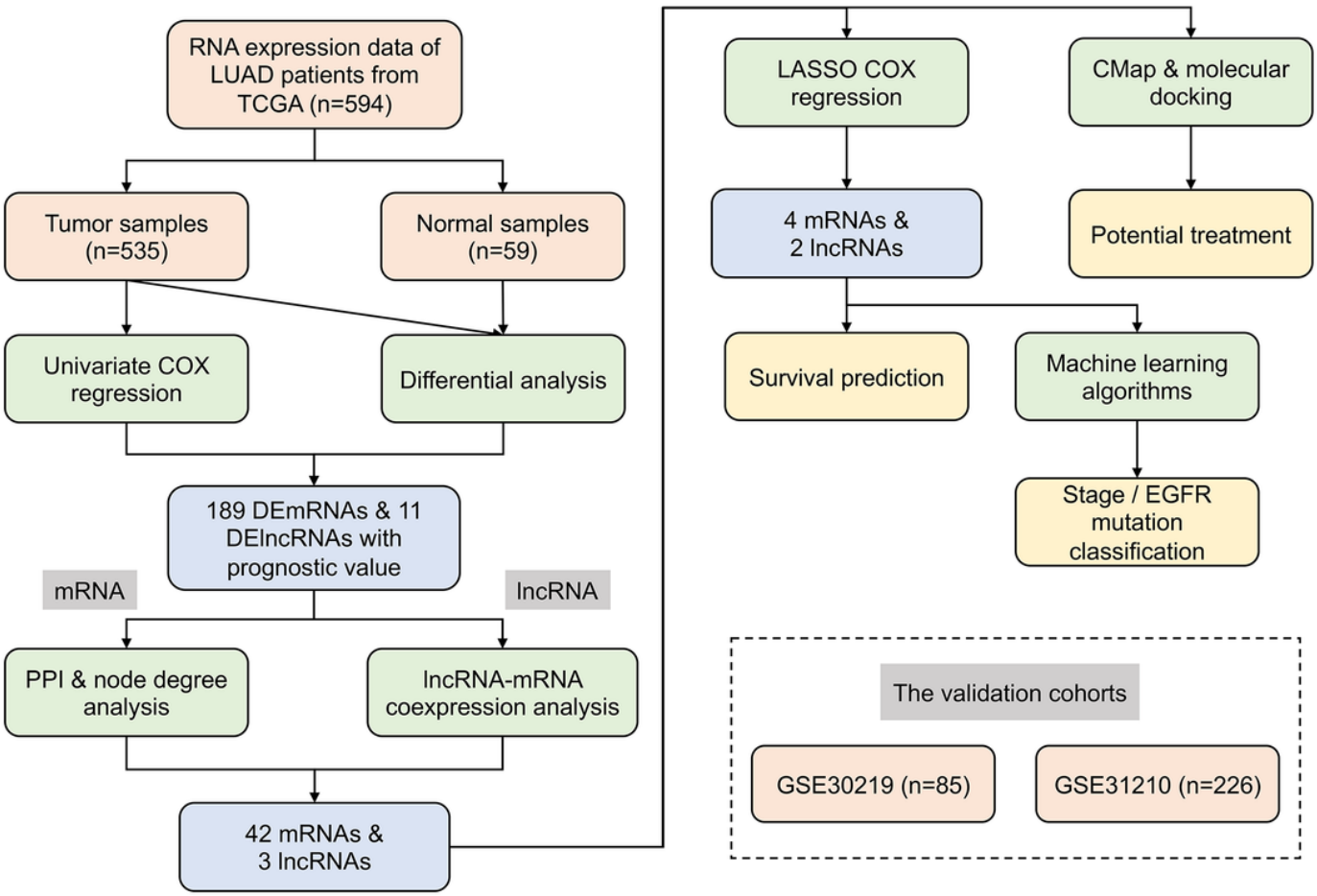

Figure 1

Flow chart of the design of the present study. LUAD, lung adenocarcinoma; TCGA, The Cancer Genome Atlas; DEmRNAs, differentially expressed mRNAs; DElncRNAs, differentially expressed lncRNAs; PPI, protein-protein interaction; LASSO, least absolute shrinkage and selection operator; CMAP, the connectivity map; EGFR, epidermal growth factor receptor; GSE, Gene Expression Omnibus.

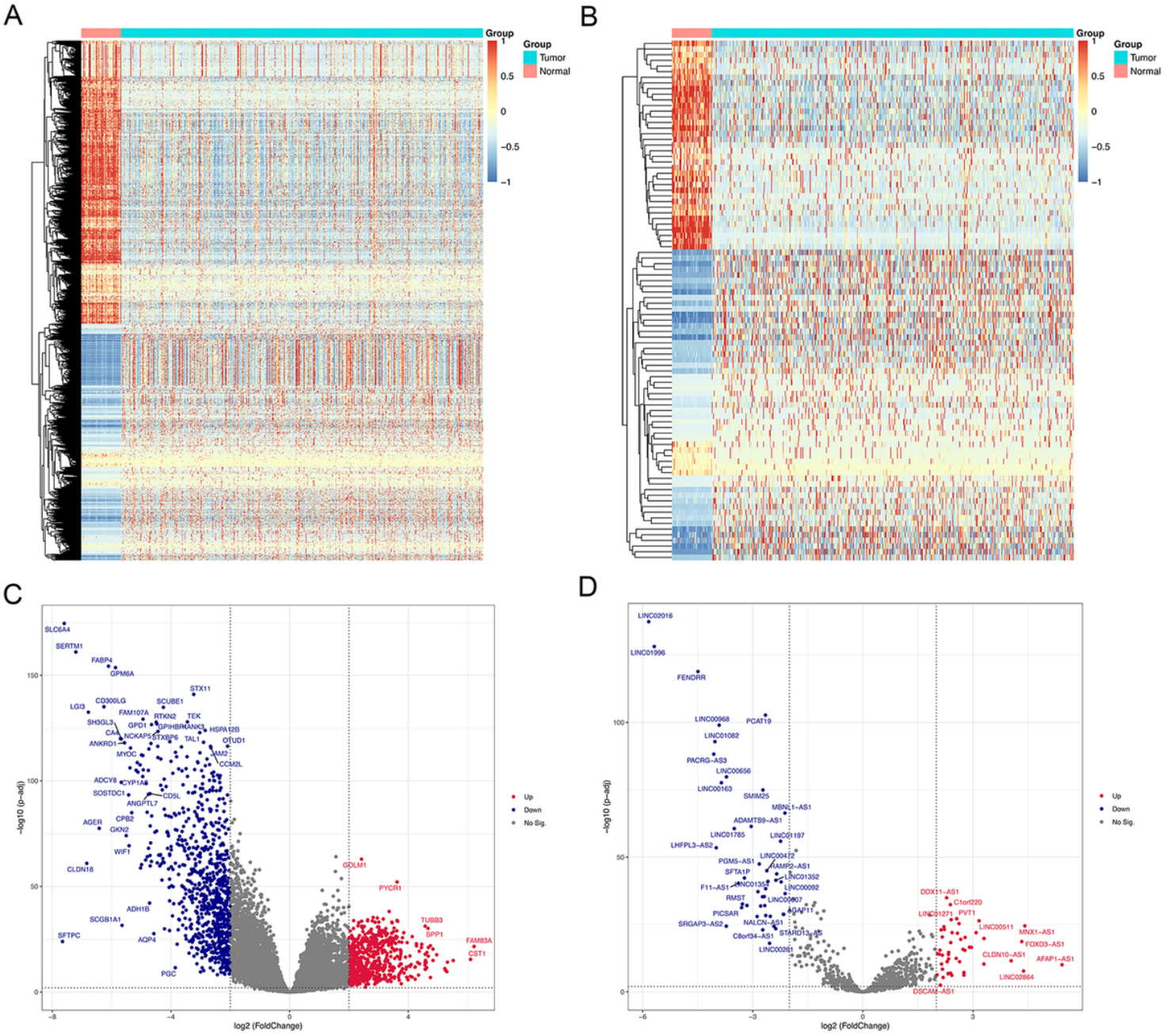

Figure 2

Differentially expressed genes between LUAD and normal samples (| log2 fold change | >2, adjusted p-value < 0.01). 606 mRNAs were upregulated and 850 downregulated in LUAD samples, while 49 lncRNAs were upregulated and 43 downregulated in comparison to normal samples. (A) and (B) shows differentially expressed mRNAs and lncRNAs, respectively. Blue through red color indicates low to high expression level. (C) and (D) are volcano maps for differentially expressed mRNAs and lncRNAs, respectively. Red dots represent significantly upregulated expressed genes and navy-blue dots represent significantly downregulated expressed genes.

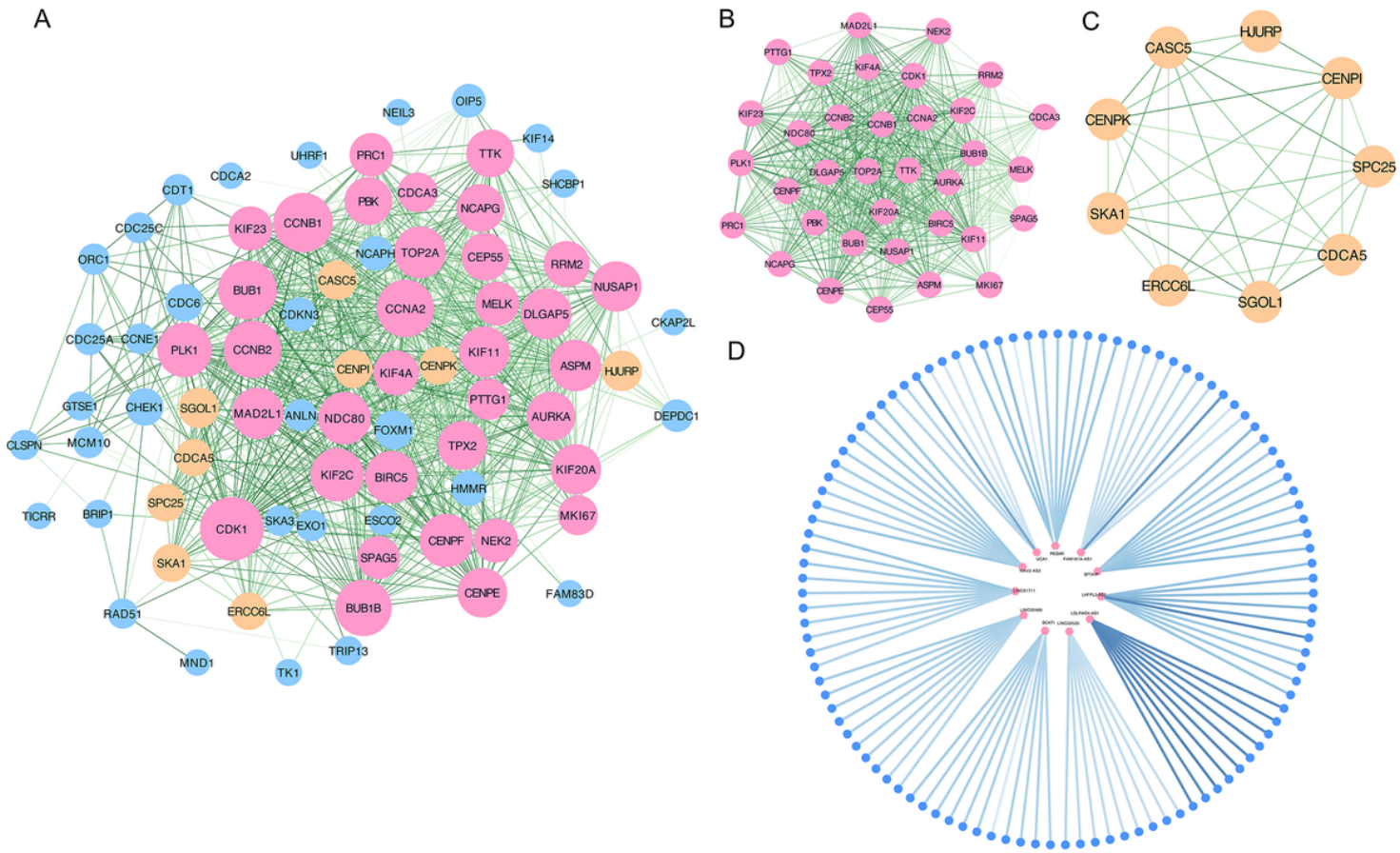

Figure 3

PPI network and co-expression analysis. (A) A PPI network with 77 nodes and 919 edges visualizes the interactions between differentially expressed mRNAs with prognostic value. (B) and (C) display two significant modules identified by MCODE, respectively. (D) co-expression relationships between differentially expressed lncRNAs with prognostic value (red points) and other mRNAs.

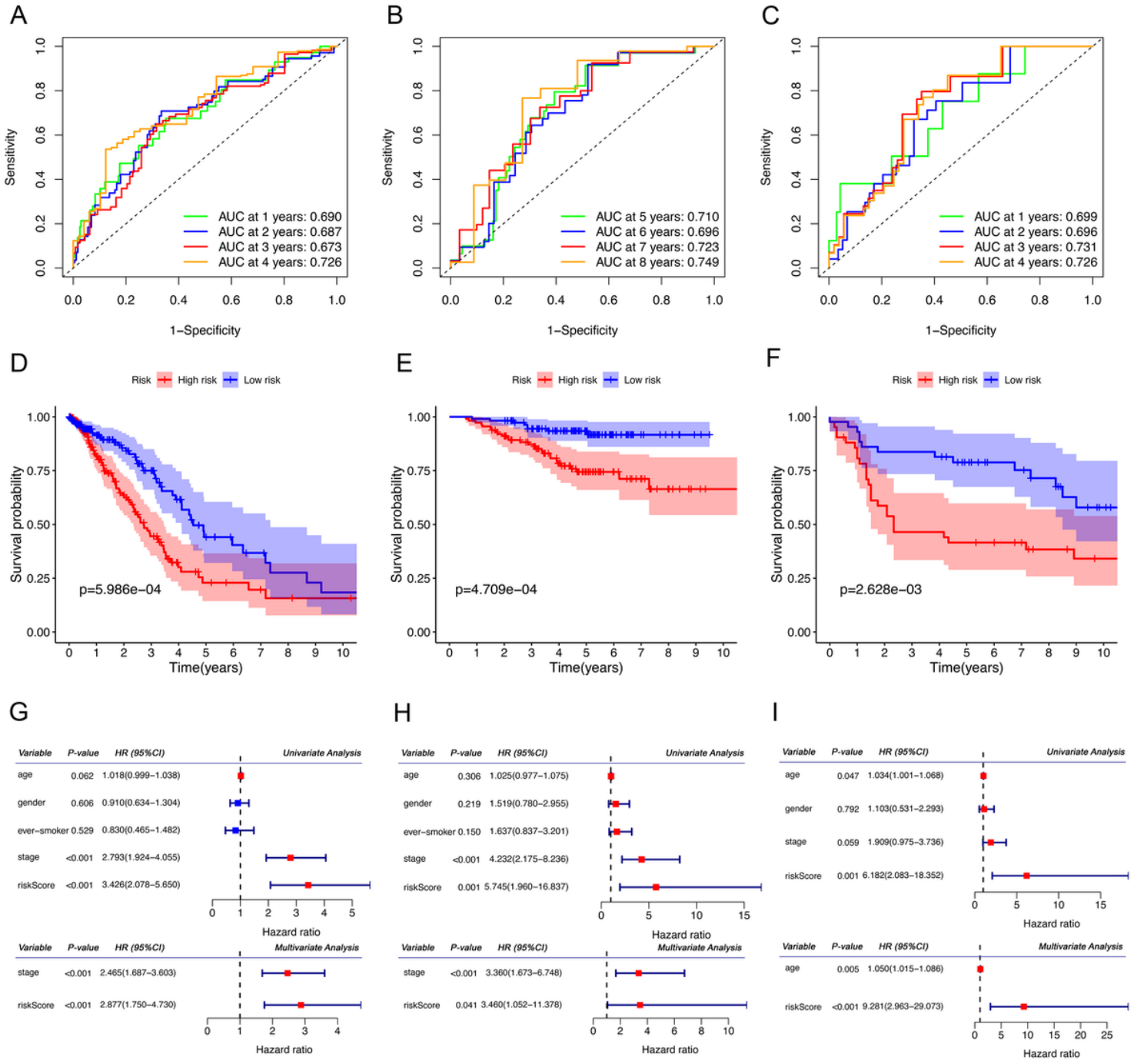

Figure 4

Prognostic analysis of the risk score in the three cohorts. (A-C) Time-dependent receiver operating characteristic curves assess the prognostic performance of the risk score. (D-F) Kaplan-Meier curves compare the overall survival of patients in the high- and low-risk groups in the three cohorts. (A) and (D) the TCGA cohort; (B) and (E) the GSE31210 cohort; (C) and (F) the GSE30219 cohort. (G-I) display the results of the univariate and multivariate Cox regression analyses regarding the overall survival in the TCGA cohort (G), the GSE31210 cohort (H), and the GSE30219 cohort (I).

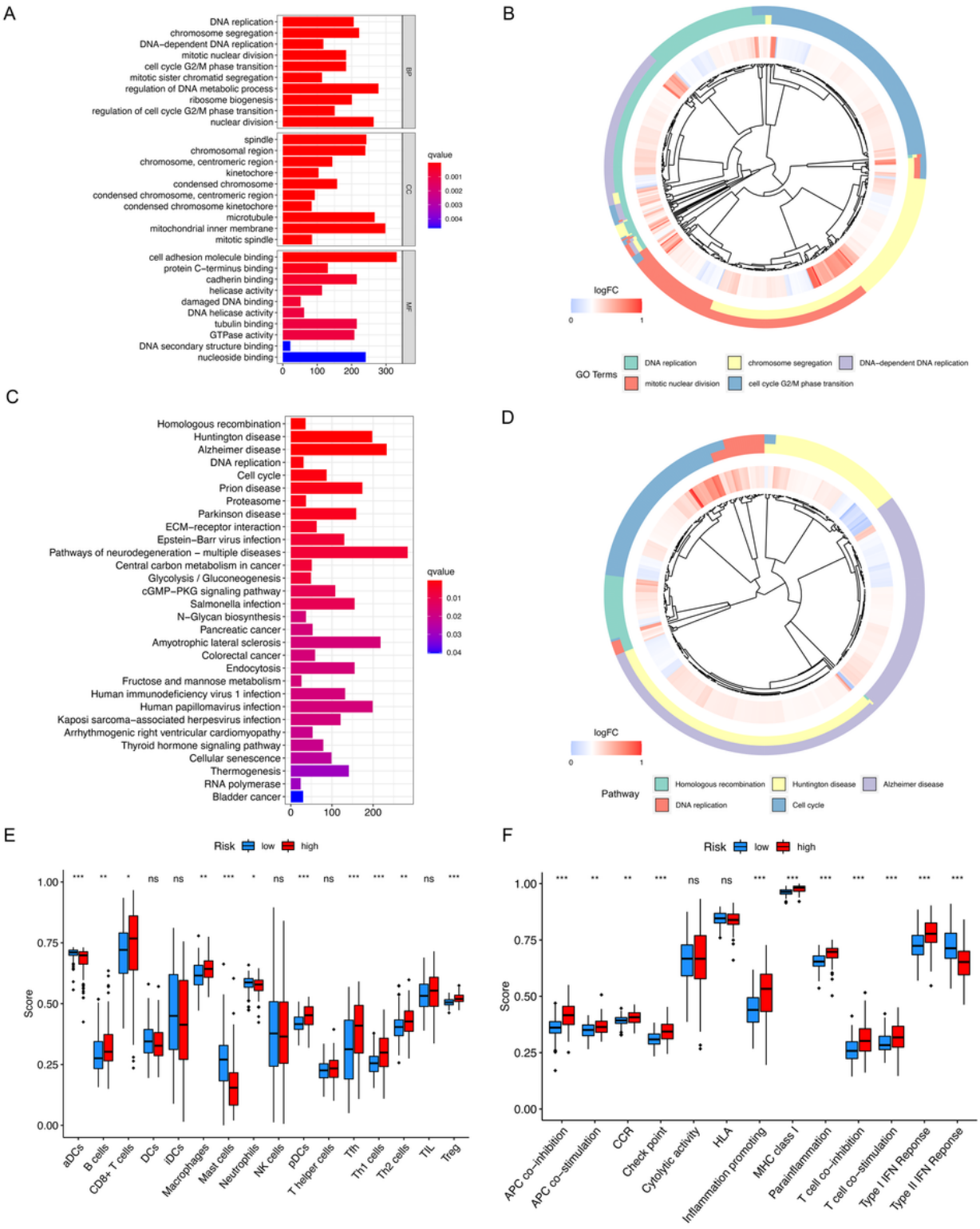

Figure 5

Further comparison between high- and low-risk groups. (A) and (B) show the bar plot and cluster plot of significant GO functional items, respectively. (C) and (D) show the bar plot and cluster plot of significant KEGG pathways, respectively. (E) and (F) demonstrate the comparison of the ssGSEA scores between different risk groups in the GSE31210 cohort. The scores of 16 immune cells (E) and 13 immune-related functions (F) are displayed in boxplots.

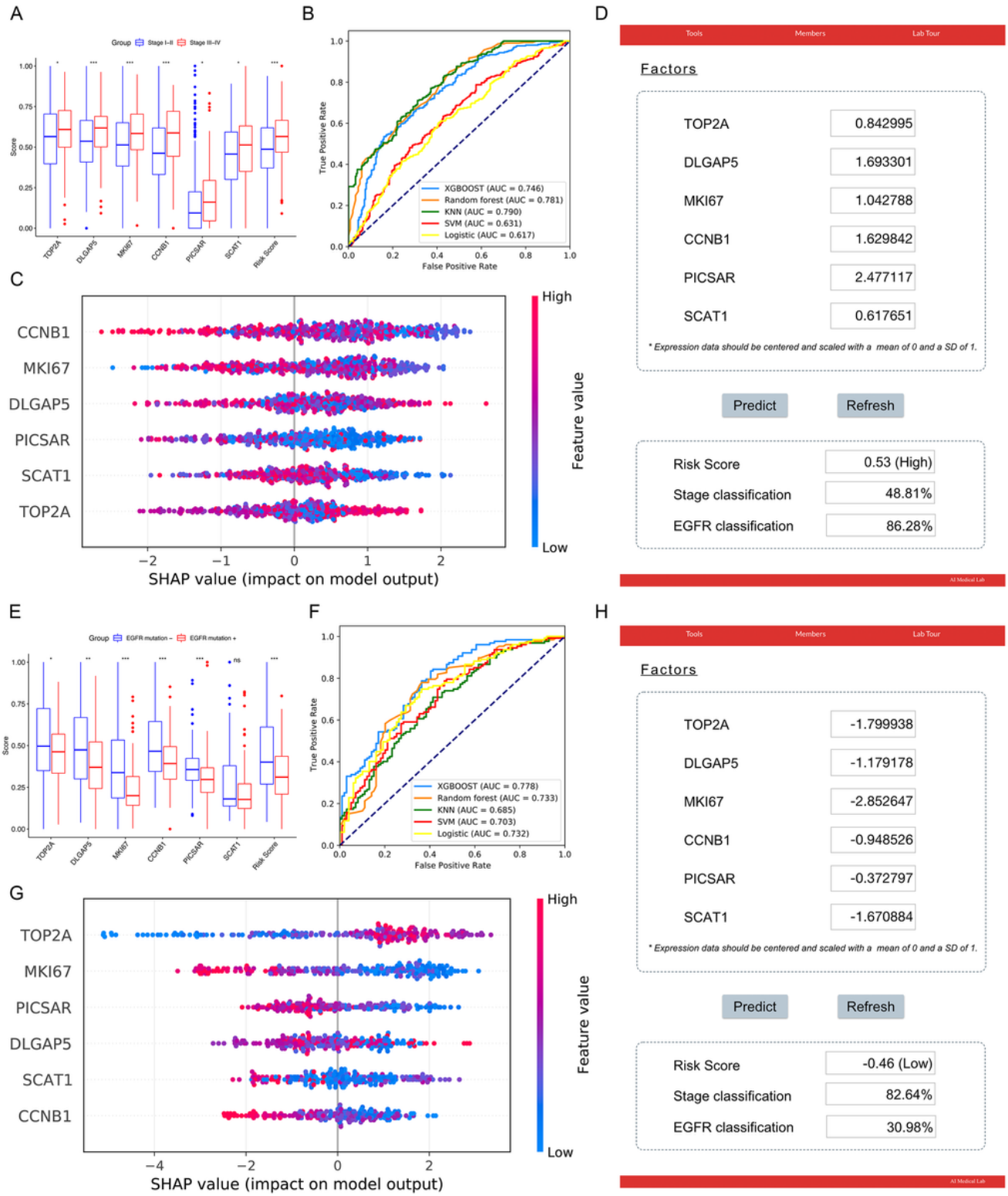

Figure 6

Development of machine learning models and a web-based tool. (A) is a boxplot that compares the distributions of six genes and the risk score between the pathological Stage I-II and pathological Stage III-IV groups. (B) shows the receiver operating characteristic curves (ROCs) of the five utilized machine learning models to discriminate early-stage (pathological Stage I-II) LUAD from other stages LUAD patients. (C) illustrates the classification results by using the SHAP values. (E) compares the distributions

of six genes and the risk score between the EGFR mutation + and EGFR mutation – groups. (F) shows the ROCs of 5 machine learning models to classify LUAD with or without EGFR mutation +. (G) is the SHAP plot to explain the results for EGFR classification. (D) and (H) show two examples of using the web-based tool. A user enters the normalized expression data of the six genes, clicking the 'Predict' button. Then, the risk score, the possibility of being early-stage predicted by the KNN model, and the possibility of being with EGFR mutation predicted by the XGBOOST model will be assessed and provided. KNN, K-nearest neighbor; SVM, support vector machine; XGBOOST, extremely gradient boosting; SHAP, SHapley Additive exPlanation; EGFR, epidermal growth factor receptor.

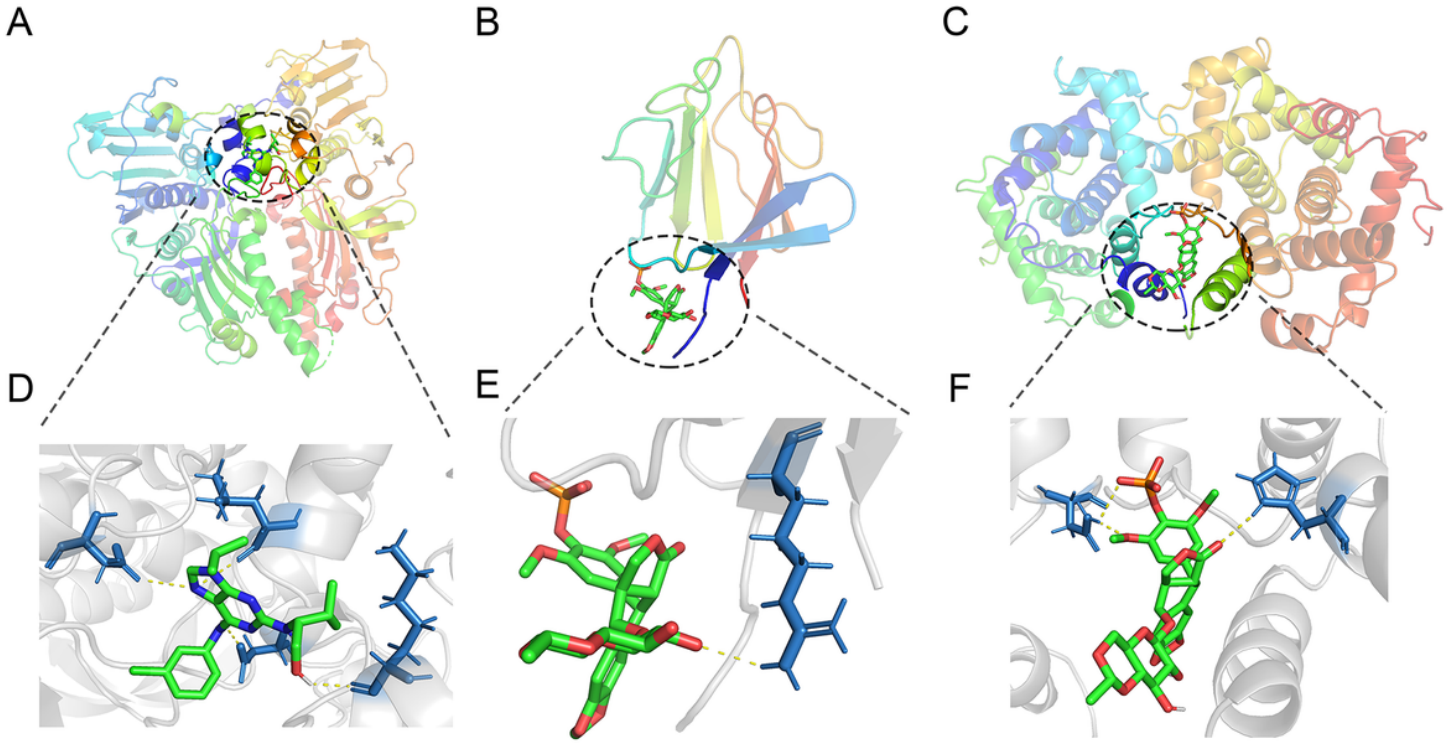

Figure 7

Three examples of the molecular docking results. (A) and (D) show the docking results of TOP2A and Purvalanol-a with an affinity of -6.7 kcal/mol. (B) and (E) are the docking results of Ki67 and Etoposide with an affinity of -6.4 kcal/mol. (C) and (F) are the docking results of CCNB1 and Etoposide with an affinity of -6.8 kcal/mol.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.docx