Research Paper

Identifying lncRNAs and mRNAs related to survival of NSCLC based on bioinformatic analysis and machine learning

Wei Yue[1], Jing Wang[1], Bo Lin[1,2], Yongping Fu[3]

[1]Innovation Centre for Information, Binjiang Institute of Zhejiang University, Hangzhou 310053, China
[2]College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[3]Department of Cardiovascular Medicine, Affiliated Hospital of Shaoxing University, Shaoxing 312099, China

Correspondence to: Bo Lin, Yongping Fu; **email:** rainbowlin@zju.edu.cn, fuyongping@usx.edu.cn

ABSTRACT

Non-small cell lung cancer (NSCLC) is the most common histopathological type, and it is purposeful for screening potential prognostic biomarkers for NSCLC. This study aims to identify the lncRNAs and mRNAs related to survival of non-small cell lung cancer (NSCLC). The expression profile data of lung adenocarcinoma and lung squamous cell carcinoma were downloaded in The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) dataset. A total of eight survival related long non-coding RNAs (lncRNAs) and 262 survival related mRNAs were filtered. By gene set enrichment analysis, 17 significantly correlated Gene Ontology signal pathways and 14 Kyoto Encyclopedia of Genes and Genomes signal pathways were screened. Based on the clinical survival and prognosis information of the samples, we screened eight lncRNAs and 193 mRNAs by single factor Cox regression analysis. Further single and multifactor Cox regression analysis were performed, 30 independent prognostication-related mRNAs were obtained. The PPI network was further constructed. We then performed the machine learning algorithms (Least absolute shrinkage and selection operator, Recursive feature elimination, and Random forest) to screen the optimized DEGs combination, and a total of 17 overlapping mRNAs were obtained. Based on the 17 characteristic mRNAs obtained, we firstly built a Nomogram prediction model, and the ROC values of training set and testing set were 0.835 and 0.767, respectively. By overlapping the 17 characteristic mRNAs and PPI network hub genes, three genes were obtained: CDC6, CEP55, TYMS, which were considered as key factors associated with survival of NSCLC. The *in vitro* experiments were performed to examine the effect of CDC6, CEP55, and TYMS on NSCLC cells. Finally, the lncRNAs-mRNAs networks were constructed.

INTRODUCTION

Lung cancer is one of the malignant tumors with the highest case fatality rate in the world. The incidence rate and mortality of lung cancer have been ascended globally, especially in a number of developing countries [1]. Non-small-cell lung cancer (NSCLC) is the most common histopathological type, accounting for about 85% of lung cancer. Lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) are two primary subtypes of NSCLC [2]. NSCLC begins with concealment and progresses rapidly. Most patients are diagnosed in local late stage or advanced stage, often losing the opportunity of the first radical resection. Due to the progress of comprehensive treatment methods for NSCLC, in addition to traditional radiotherapy, chemotherapy and other treatment methods, targeted therapy, immunotherapy and a variety of combined treatment modes have flourished. These strategies demonstrate extensive and significant clinical efficacy.

However, the survival rate of NSCLC is still at a relatively low level. It has been reported that from 2012 to 2015, the survival rate of NSCLC in Chinese male patients was only 16.8%, 62.5% lower than that of thyroid cancer with the highest survival rate [3]. The survival rate of Chinese women with NSCLC is 25.1%, which is also classified as low survival rate [4]. Therefore, it is urgent to explore new strategies that can improve the clinical therapeutic effect of NSCLC.

Machine learning is a multi-disciplinary and inter-disciplinary discipline, covering knowledge of probability theory, statistics, approximate theory and complex algorithms [5]. It uses computers as tools and is committed to simulating human learning methods, dividing existing content into knowledge structures to effectively improve learning efficiency, and integrating computer science and statistics into medical problems [6]. By improving algorithms, absorbing input data, applying computer analysis to predict the output value within the acceptable accuracy range, identifying the patterns and trends in the data, and finally learning from previous experience, the development of machine learning brings a new direction to the diagnosis and treatment of lung cancer [7]. Nomogram is the RMS package in the R statistical software, based on the LNR settings [8, 9]. The consistency index (C-index) and the calibration map were used to measure the performance of the model [10]. The consistency index is generally between 0.5-1. When the C index value is closer to 1, it is larger, but also indicates that the consistency of the model is better, that is, the prediction performance of the model is better.

In this study, we synthesized the expression profile data of LUAD and LUSC samples, screened the lncRNAs and mRNAs that are significantly related to survival, further screened the characteristic genes through different machine learning algorithms, and constructed the survival status classification model of the samples.

MATERIALS AND METHODS

Data processing

The LUAD and LUSC gene expression level data were downloaded from Xena Database (https://xenabrowser.net/datapages/), including 585 and 550 samples, respectively. The detection platform was Illumina HiSeq 2000 RNA Sequencing. According to the sample clinical information downloaded at the same time, the LUAD and LUSC samples with survival and prognosis information were retained. A total of 994 NSCLC tumor samples and 107 normal control samples were included in this analysis. Data from TCGA samples were used for training data sets. Since LUAD and LUSC are gene

expression level data from different batches, we first use the R3.6.1 sva package [11] version 3.38.0 (http://www.bioconductor.org/packages/release/bioc/html/sva.html) to remove the batch effect of LUAD and LUSC expression profile data.

At the same time, data in GSE37745 [12, 13] of NSCLC expression profile were downloaded from the NCBI GEO database (https://www.ncbi.nlm.nih.gov/), including 196 NSCLC tumor samples with clinical prognosis information. The detection platform is GPL570 Affinemetrix Human Genome U133 Plus 2.0 Array. This data set was used as a validation set.

After downloading and obtaining the expression profile data, we annotated the detected lncRNAs and mRNAs according to the Transcript ID in the Illumina HiSeq 2000 RNA sequencing annotation platform for the TCGA platform detection data. For the NCBI GEO dataset, we download the detailed annotation information (including probe, gene symbol, RNA type and other information) of the platform involved in the corresponding platform from the Ensemble gene browser 96 database Biomart [14] (http://asia.ensembl.org/index.html), and then re-annotated the detection probe to obtain the corresponding expression level of the detected lncRNA and mRNA.

Screening of differentially expressed RNAs (DERs)

In the expression profile data set after combining LUAD and LUSC, the samples were first divided into Tumor vs. Control comparison groups according to the sample source, and then in the tumor samples, the samples were divided into Dead vs. Alive comparison groups according to the survival status of the samples [15–17]. Later, we used the limma package Version 3.34.7 (https://bioconductor.org/packages/release/bioc/html/limma.html) to screen DERs in the two comparison groups by R3.6.1 language, and FDR<0.05 and | log2FC |>0.5 were selected as the threshold for screening significant factors.

Finally, we compared the DERs filtered in the comparison groups of Tumor vs. Control and Dead vs. Alive, and overlapping part was obtained. Based on DAVID version 6.8 [18, 19] (https://david.ncifcrf.gov/), GO Biological Process and KEGG signal pathway enrichment analysis was performed, and FDR<0.05 was selected as the threshold of enrichment significance.

Identifying DERs with a significant association of prognosis

For the selected DERs, combined with the clinical survival and prognosis information of the samples, the single-factor Cox regression analysis of the

survival package Version 2.41-1 in R3.6.1 (http://bioconductor.org/packages/survivalr/) was used to screen the significantly different expressions of lncRNAs and mRNAs that were significantly related to the survival and prognosis of NSCLC [20]. The NSCLC prognosis-related mRNAs were further analyzed by multivariate Cox regression, and the independent prognostic related mRNAs were selected. P-value less than 0.05 was selected as the threshold for screening significant correlation.

Construction of protein-protein interaction (PPI) network and analysis of network topology

The STRING database [21] (Version:11.0, http://string-db.org/) was used to search for the interaction relationship between mRNAs gene product proteins that were significantly related to survival and prognosis, and the interaction network was constructed. The network was visualized and the network topology structure was analyzed through Cytoscape Version 3.6.1 [22] (http://www.cytoscape.org/).

Optimal mRNAs screening and nomogram diagnostic model construction

Based on the expression level of mRNAs that were significantly related to the independent prognosis obtained from the previous screening, three different optimization algorithms [LASSO (least absolute shrinkage and selection operator) [23], RFE (recursive feature elimination) [24], and RF (random forest) [25]] were used to screen the characteristic factors. We subsequently compared the results of the three algorithms and selected the overlapping part as the final feature mRNAs combination.

Construction and verification of nomogram diagnostic model

We used R3.6.1 rms package (https://cran.r-project.org/web/packages/rms/index.html) Version 5.1-2 to build the Nomogram model [26], and analyzed the model with a line chart, using C index as a parameter to measure the fit between the model and the actual. Based on the selected characteristic mRNA factors, we used the R3.6.1 language rmda package [27] (https://cran.r-project.org/web/packages/rmda/index.html) Version 1.6 to observe the model yield. Finally, in the validation data from GSE37745, the Nomogram model was also constructed based on the characteristic mRNA factors obtained as previously screened to validate the diagnostic model efficacy. The ROC curves of the Nomogram model were assessed by R3.6.1 pROC v1.18.0 package [28] (https://cran.r-project.org/web/packages/pROC/index.html).

Screening of the key mRNAs

We compared the selected characteristic genes for constructing survival diagnosis model with the important link hub genes in PPI network, and selected the overlapping part as the important factor. In the combined TCGA training set and GSE37745 validation data set, the Kaplan-Meier curve method in the survival package Version 2.41-1 in R3.6.1 [20] was used to analyze and display the correlation between the expression level of important genes and survival prognosis.

Co-expression network of lncRNAs-mRNAs

Based on the characteristic factors in the diagnosis model of lncRNAs and Nomogram, which were significantly correlated with independent prognosis, the Pearson correlation coefficient between them was calculated by using the cor function (http://77.66.12.57/R-help/cor.test.html) in R3.6.1, and the co-expression network of lncRNAs-mRNAs with independent prognosis was constructed. The network was displayed through Cytoscape Version 3.6.1.

Cell lines

The NSCLC cell lines (A549 and H1299) were purchased from the Cell Bank of Chinese Academy of Medical Science (Shanghai, China). A549 and H1299 cells were separately cultured in the Dulbecco's modified eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 100 U/mL penicillin and 100 mg/mL streptomycin and incubated in an incubator at $37°$ C and 5% CO_2 conditions.

Cell transfection

A549 and H1299 cells were harvested to prepare cell suspension with the serum-free DMEM, with the final cell concentration of 1×10^6 cells/mL. Thereafter, the cell suspension (1 ml/well) was added into the 6-well plates. Then, the siRNAs targeted *CDC6* (siCDC6-1, siCDC6-2), *CEP55* (siCEP55-1, siCEP55-2), *TYMS* (siTYMS-1, siTYMS-2), and negative control (NC, Genechem, Shanghai, China) were transfected into A549 and H1299 cells, respectively, following the manuals. Lipofectamine 2000 (Thermo Fisher Scientific, Waltham, MA, USA) was used in cell transfection, and all cells were transfected for 8 h under $37°$ C and 5% CO_2 conditions. Afterwards, DMEM supplemented with 10% FBS was added into each well to culture cells.

RT-PCR

The extraction of total RNA in A549 and H1299 cells was performed by using the Trizol Reagent

(Life Technologies, Shanghai, China) according to the instructions. The total RNA was reverse transcribed into the cDNA template by using the PrimeScript™ RTreagent kit (Takara, Beijing, China). The expression of *CDC6*, *CEP55*, and *TYMS* mRNA was scrutinized by qPCR with the SYBR Premix Ex Taq™ kit (Takara, Dalian, China) on the ABI 7900HT Real-Time PCR system (Applied Biosystems, Foster City, CA, USA). The relative mRNA expression of *CDC6*, *CEP55*, and *TYMS* was estimated by the $2^{-\Delta\Delta Ct}$ method and normalized to GAPDH.

Western blot

The total proteins in A549 and H1299 cells were extracted by applying the RIPA lysis buffer (Beyotime, Shanghai, China) in the light of the instructions. The concentration of total protein samples was scrutinized by using the BCA kit (Beyotime, Shanghai, China). A weight of 20 μg of the total protein sample were collected and separated with 10% SDS-PAGE. After the blockage by 5% skimmed milk for 1 h at 25° C, the proteins were transferred to the polyvinylidene difluoride (PVDF) membranes. Rabbit anti-primary antibodies were then dropped onto the PVDF membranes to probe the proteins for 12 h at 4° C, including anti-CDC6 (1:1000, ab109315, Abcam, Shanghai, China), anti-CEP55 (1:1000, ab170414, Abcam), anti-TYMS (1:1000, CSB-PA025393GA01HU, CUSABIO, Wuhan, China) and anti-GAPDH (1:1000, CSB-MA000071, CUSABIO). Followed by this, goat anti-rabbit secondary antibody (1:2000, A21020, AmyJet Scientific, Wuhan, China) was utilized for 2 h treatment of the proteins at room temperature. The enhanced chemiluminescent (ECL) kit (AmyJet Scientific, Wuhan, China) was applied for the visualization of the specific protein blots according to the directions. The quantification of proteins was determined by Image Lab software 3.0 (Bio-Rad Laboratories, Hercules, CA, USA).

Cell counting kit-8 (CCK-8) assay

CCK-8 assay was performed to estimate cell viability of A549 and H1299 cells. In brief, 1×10^4 A549 and H1299 cells were respectively seeded into the 96-well plates containing 100 μL DMEM supplemented with 10% FBS, and maintained under 37° C and 5% CO_2 conditions. After 24, 48, 72 and 96 h of culture, the 96-well plates were taken out from the incubator. Then, CCK-8 solution (10 μL/well) was added into each well to incubate cells for 4 h at 37° C. The absorbance (OD) value of each well was measured at 450 nm using a microplate reader.

Cell migration and invasion

The 24-well Transwell chambers (Litchi Biotechnology, Shanghai, China) were purchased for evaluating cell migration and invasion. A549 and H1299 cells were suspended into 300 μL non-serum DMEM, followed by being seeded into the upper chambers. DMEM containing 10% FBS was added into the lower chambers. After 24 h of incubation, the migration cells were sequentially fixed by 4% paraformaldehyde and stained by 1% crystal violet for 10 min. The number of migration cells was counted under the microscope (IX81, Olympus, Tokyo, Japan). For cell invasion, 100 μL Matrigel (Litchi Biotechnology) was pre-coated into the upper chambers before cell seeding.

Cell apoptosis

A549 and H1299 cells were harvested and washed by PBS twice. A549 and H1299 cells were subsequently resuspended into 100 μL 1× Binding Buffer. Then 5 μL FITC and 10 μL propidium iodide solution were added and gently mixed. Cells were placed at 25° C for 15 min. A total of 400 μL 1× Binding Buffer was then added to treated cells for 15 min on ice. Cell apoptosis was analyzed by the FACSCalibur flow cytometer (BD Biosciences, San Jose, CA, USA).

Statistical analysis

The experiments in the present study were carried out in triplicates. Statistical analysis was implemented using GraphPad Prism 10 software. The data were displayed as mean ± standard deviation. Paired Student's t-test was executed to analyze the difference between the two groups. One-way analysis of variance and Tukey's post hoc test were employed for the data comparison in more than two groups. $P < 0.05$ revealed a statistically significant difference.

Availability of data and material

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

RESULTS

Screening of factors with a significant association of prognosis

We downloaded the expression profile data corresponding to LUAD and LUSC in TCGA, and then combined them into a data set. The sample relationship before and after batch effect removal was shown in Figure 1A. In the expression profile data set after combining LUAD and LUSC, we first divided the samples into Tumor (n = 994) vs Control (n = 107) comparison group, and then divided the samples into Dead (n = 394) vs Alive (n = 600) comparison group. In the comparison group of Tumor

vs. Control, 55 significantly differentially expressed lncRNAs and 2287 significantly differentially expressed mRNAs were screened (Figure 1B). In the Dead vs. Alive comparison group, 22 significantly differentially expressed lncRNAs and 459 significantly differentially expressed mRNAs were screened (Figure 1B). We further compared the DERs filtered in Tumor vs. Control and Dead vs. Alive, and a total of eight overlapping lncRNAs and 262 mRNAs were filtered (Figure 1C). The information list was shown in Supplementary Table 1. Finally, the enrichment analysis of GO function and KEGG signal pathway based on DAVID was carried out for the overlapped mRNAs with significant differential expression. A total of 17 significantly correlated GO signal pathways and 14 KEGG signal pathways were screened, which were displayed in Figure 1D–1F.

Identifying of DERs with a significant association of prognosis and PPI network construction

Based on the clinical survival and prognosis information of the samples, eight lncRNAs and 262 mRNAs by single

factor Cox regression analysis were screened. Eight lncRNAs and 193 mRNAs that were significantly related to survival and prognosis were obtained. The eight lncRNAs were: *FEZF1-AS1*, *SNHG12*, *BANCR*, *SNHG3*, *HLA-DQB1-AS1*, *SH3BP5-AS1*, *VIM-AS1*, and *FAM83A-AS1*. Further multifactor Cox regression analysis was performed on 193 prognostication-related mRNAs, and the independent prognostication-related mRNAs were selected. A total of 30 independent prognostication-related mRNAs were obtained, which was shown in Supplementary Table 2. The STRING database was used to search for the interaction relationship between 262 mRNAs product proteins that are significantly related to survival and prognosis. The PPI network was constructed, as shown in Figure 2. The network contained 145 gene nodes in total.

Optimal mRNA marker excavation and nomogram diagnostic model construction

Based on the expression level of 30 independent prognosis significantly correlated mRNAs obtained in
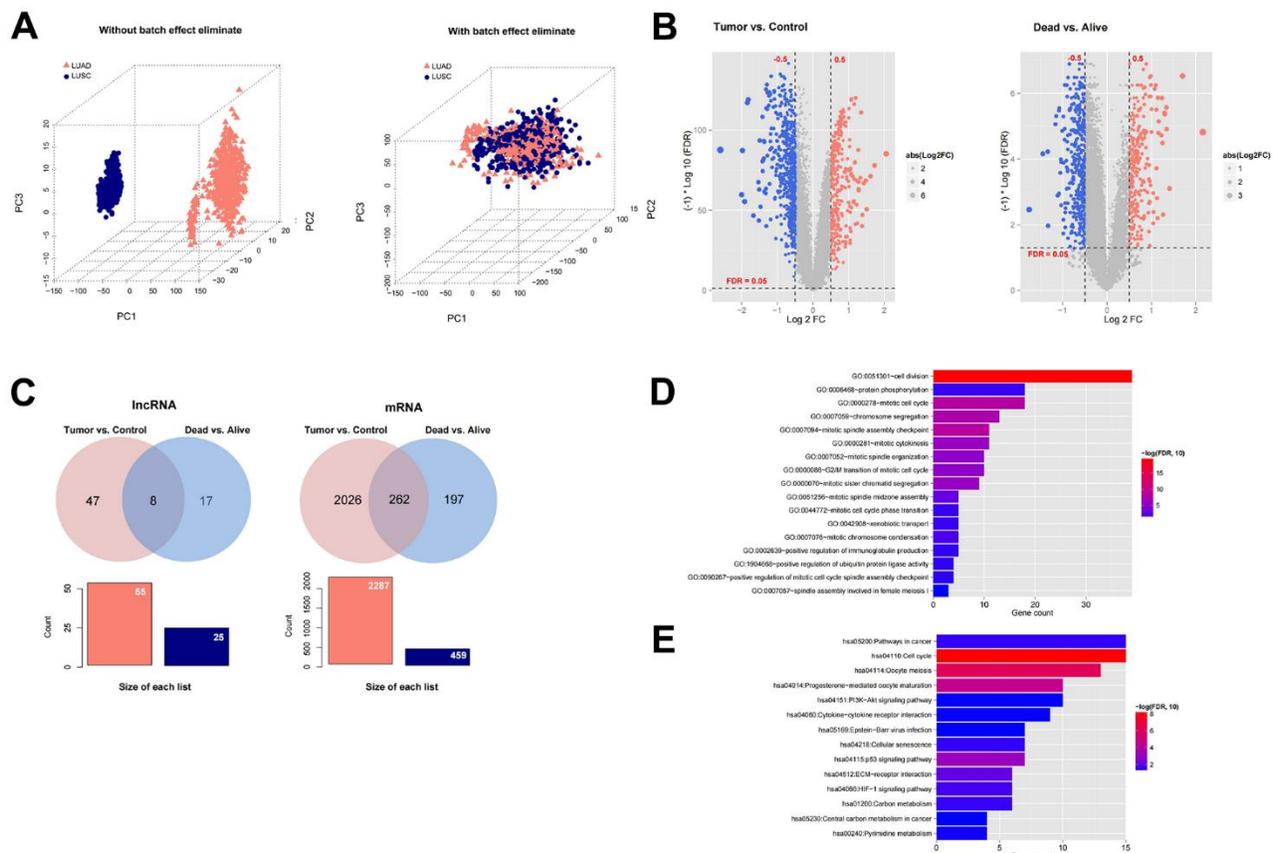
Figure 1. Screening prognosis-related mRNAs and lncRNAs based on TCGA data. (**A**) The sample relationship before and after batch effect removal. (**B**) Heatmap of differentially expressed mRNAs and lncRNAs in Tumor (994) vs Control (107) comparison group and Dead (394) vs Alive (600) comparison group. (**C**) A total of eight overlapping lncRNAs and 262 mRNAs were filtered. (**D**–**E**) The enrichment analysis of GO function and KEGG signal pathway based on DAVID was carried out for the overlapped mRNAs with significant differential expression.

the previous step in the TCGA combined data set, the LASSO, RFE and RF algorithms were used to screen the optimized DEGs combination, and the parameter diagram of algorithm filtering is shown in Figure 3A–3C). In LASSO, RFE and RF algorithms, we screened 22, 27 and 25 mRNAs respectively. Comparing these three mRNAs sets, a total of 17 overlapping mRNAs were obtained (Figure 3D). The 17 mRNAs obtained as the final optimized mRNAs combination: *ADRB2*, *ATP13A4*, *CDC6*, *CEP55*, *CLIC6*, *COL4A3*, *CPED1*, *DEPDC1B*, *DNASE1L3*, *E2F7*, *FAM83A*, *FSTL3*, *IER5L*, *LAMC2*, *SFTA3*, *TOX*, *TYMS*.

The diagnostic nomogram model construction and validation

Based on the 17 characteristic mRNAs obtained by screening, we constructed a Nomogram prediction model according to the expression level of each factor, as shown in Figure 4A. Then, the Nomogram diagnostic model was analyzed by line graph, as shown in Figure 4B, from which the Cindex value was 0.765. After that, the decision curve analysis was carried out on the model to observe the net return rate of the sample diagnosis results of the model, as shown in Figure 4D. In addition, the ROC curve of the model was analyzed, and the results are shown in Figure 4C. Finally, in the validation data set GSE37745, the Nomogram model was also built based on the 17 mRNAs factors screened previously to verify the effectiveness of the diagnostic model. The results were shown in Figure 5A–5D). The

ROC values of training set and testing set were 0.835 and 0.767, respectively.

We visualized the expression level of 17 genes in the combined TCGA training data set and validation data set (GSE37745). As displayed in Figure 6A, 6B), the expression level of 17 genes in the GSE37745 validation data set was completely consistent with the direction of the expression difference in the combined TCGA training data set. The expression level of 13 genes, including *ADRB2*, *ATP13A4*, *CDC6*, *CEP55*, *CLIC6*, *COL4A3*, *CPED1*, *DEPDC1B*, *FAM83A*, *FSTL3*, *SFTA3*, *TOX*, *TYMS*, was significantly different in the group comparison.

Screening of key genes

By comparing the 17 characteristic genes selected to construct the survival diagnosis model with the important link hub genes in the PPI network, the overlapping part was selected as an important factor, and a total of 3 genes were obtained: *CDC6*, *CEP55*, *TYMS*. In the combined TCGA training set and GSE37745 validation data set, the samples were divided into low-volume (expression level lower than the median value) and high-volume expression group (expression level higher than or equal to the median value) according to the respective expression level of the three genes. The Kaplan-Meier curve method was used to analyze and display the correlation between the expression level of important genes and survival and
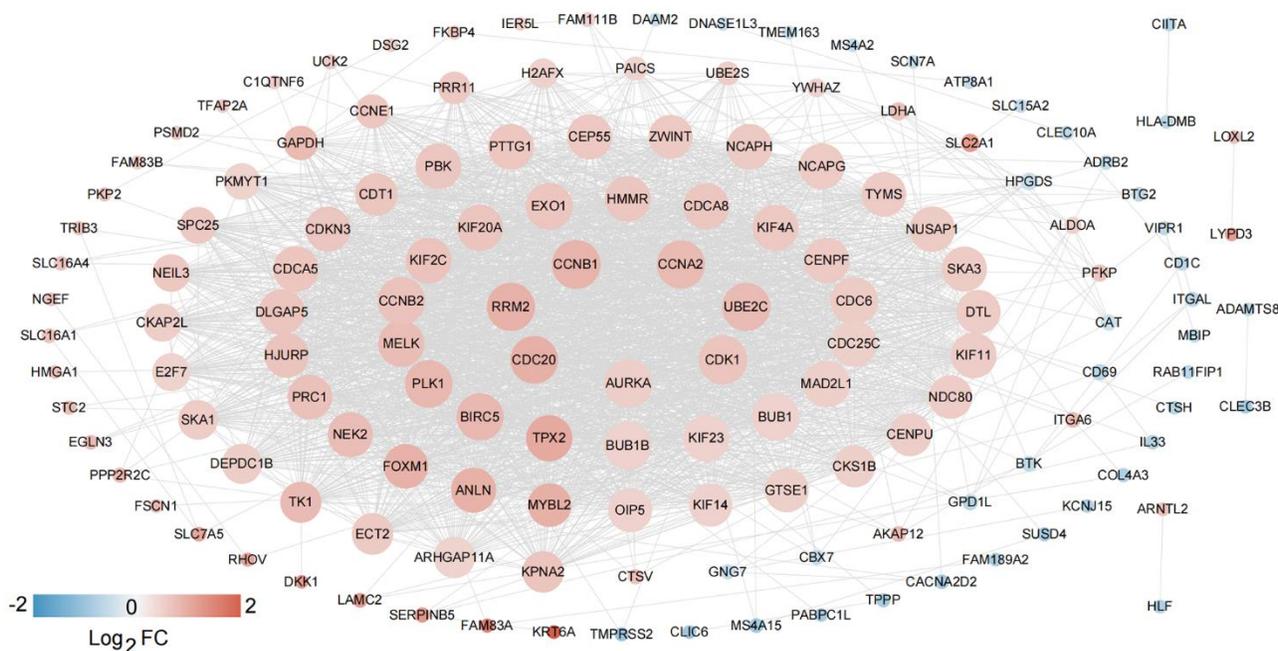
Figure 2. PPI network construction. The network contained 145 gene nodes in total.

prognosis. The results are shown in Figure 7A, 7B), high expression of *CDC6*, *CEP55*, and *TYMS* predicted poor prognosis.

Construction of a co-expression network based on characteristic mRNAs and lncRNAs

Based on the expression level of eight lncRNAs that are significantly related to independent prognosis and 17 important characteristic genes related to survival status diagnosis, the expression correlation between them was

calculated. After retaining the action pairs with significant correlation $P < 0.05$, a total of 79 pairs of relationship pairs were screened, and the relationship connection network is shown in Figure 8.

CDC6, CEP55, and TYMS affected cell activities in NSCLC cells

NSCLC cells were transfected with siCDC6, siCEP55, and siTYMS, and the efficiency of transfection was examined by RT-PCR and western blot
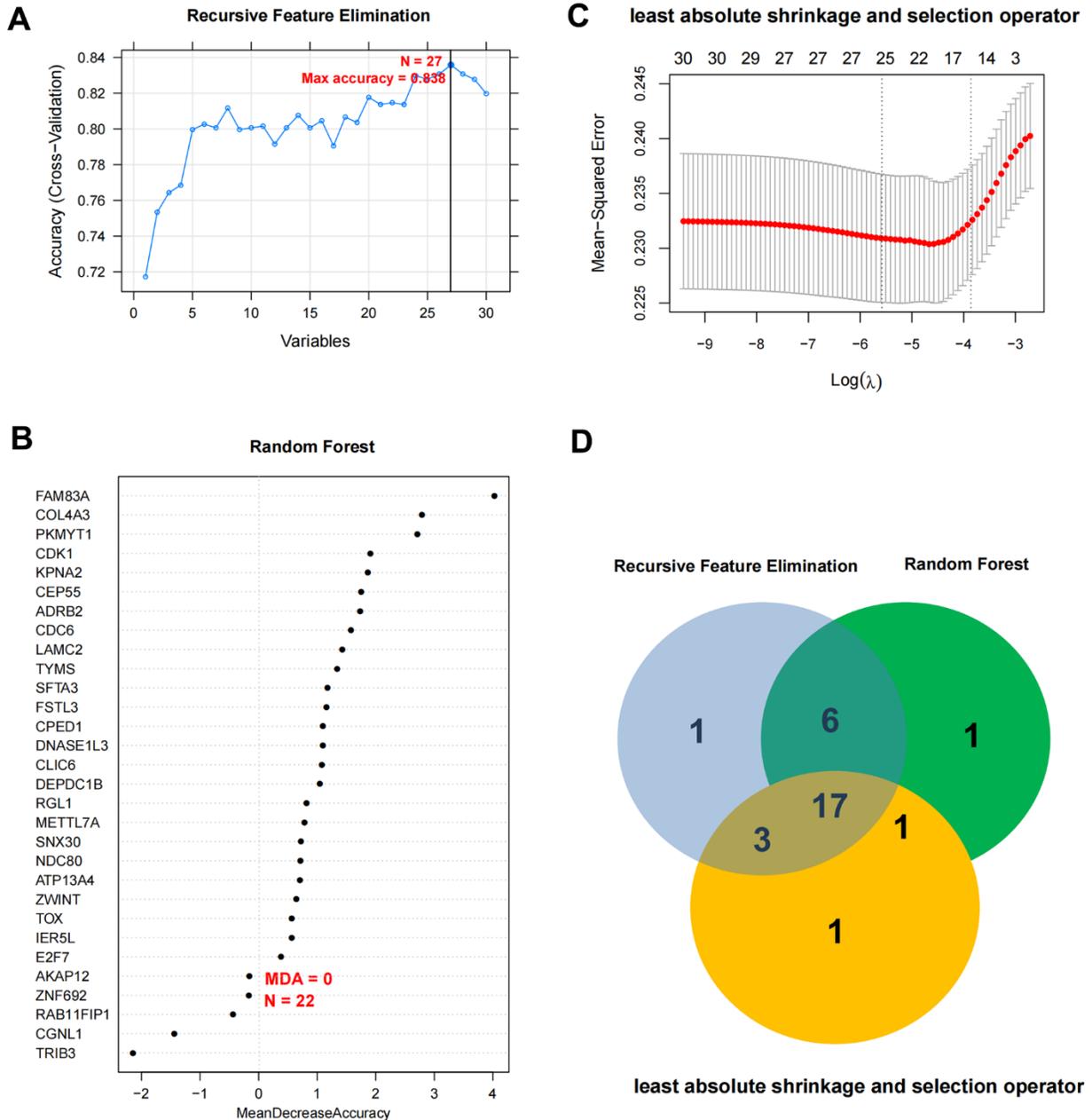
Figure 3. Optimal mRNA marker excavation and nomogram diagnostic model construction. (**A**–**C**) Filter characteristic mRNAs parameter diagram of RFE, RF, and LASSO. (**D**) Comparison chart of characteristic mRNAs combinations filtered by RFE, RF, and LASSO.
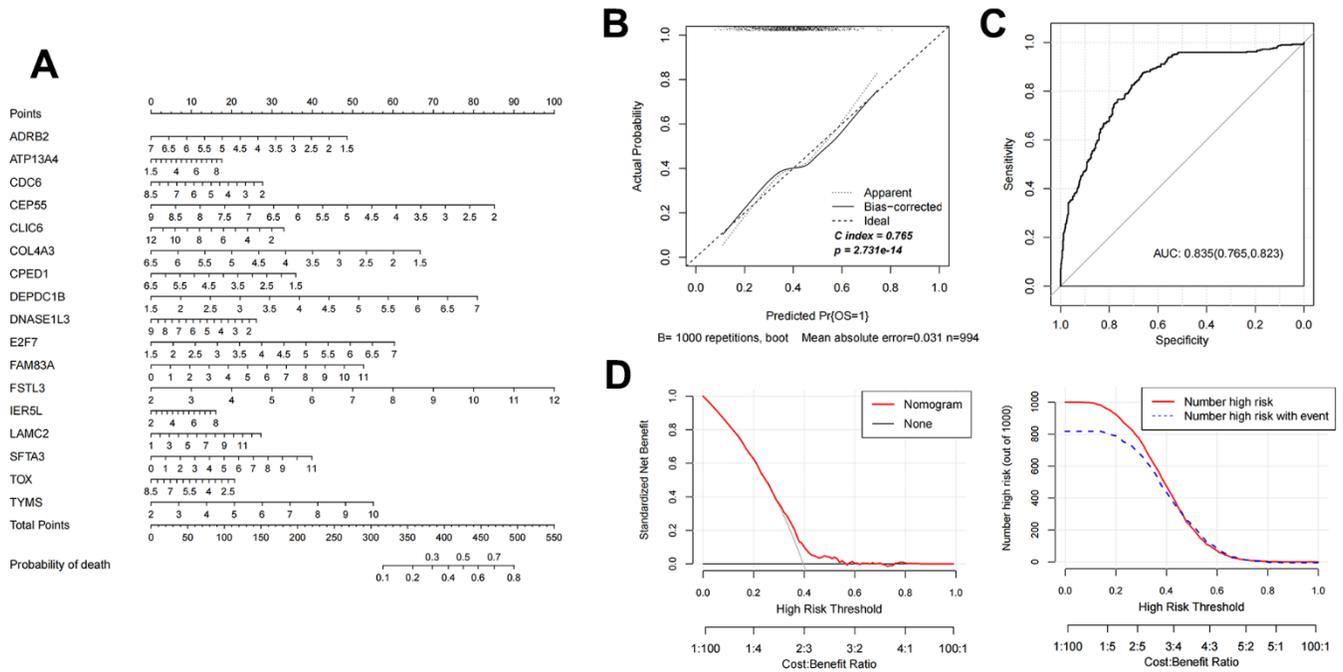
Figure 4. Nomogram diagnostic model construction and evaluation. (**A**) Nomogram model diagram based on the expression level of 17 characteristic mRNAs in the combined training data set. (**B**) Nomogram diagnostic model line chart. (**C**) The ROC value was calculated. (**D**) Model decision line diagram.
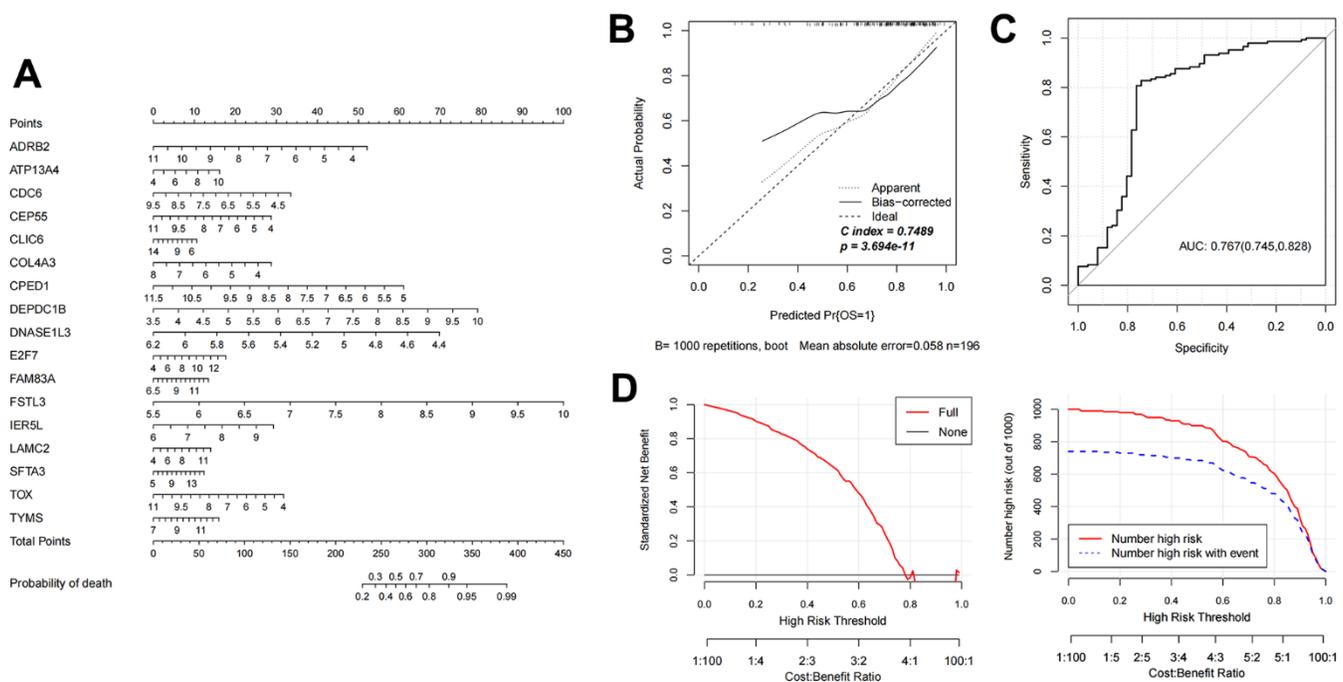
Figure 5. Evaluation of nomogram diagnostic model in GSE37745 dataset. (**A**) Nomogram model diagram of expression level of mRNAs in GSE37745 validation data set based on 17 features. (**B**) Nomogram diagnostic model line chart. (**C**) The ROC value was calculated. (**D**) Model decision line diagram.

A

TCGA

B
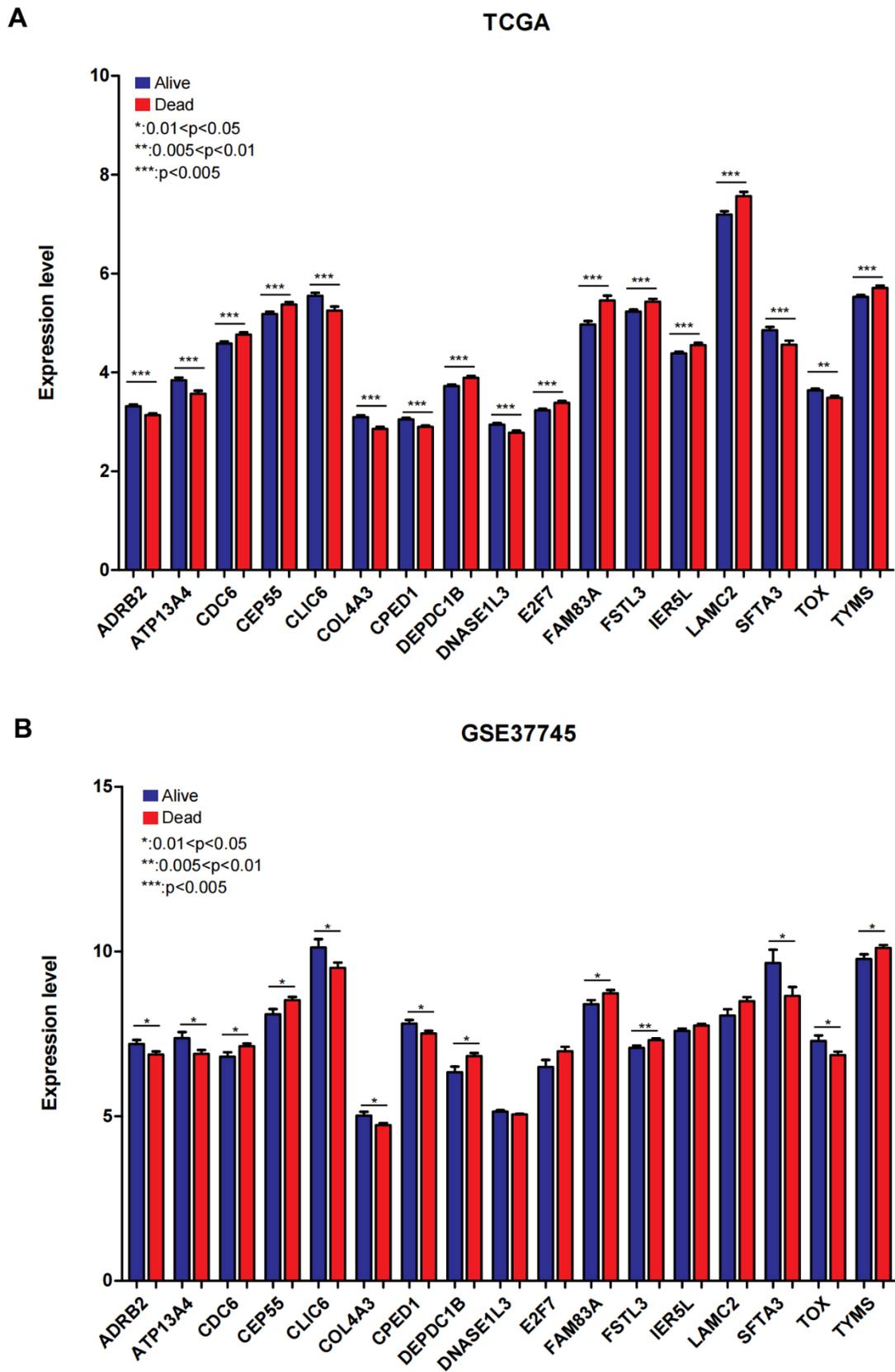
GSE37745

Figure 6. The expression of 17 mRNAs in combined TCGA training set and GSE37745 testing dataset. (**A**) The expression of 17 mRNAs in combined TCGA training set. (**B**) The expression of 17 mRNAs in GSE37745 dataset. 0.01<*P<0.05; 0.005< **P<0.01; ***P<0.005.

(Supplementary Figure 1). The siCDC6-1 (siCDC6), siCEP55-1 (siCEP55), and siTYMS-1 (siTYMS) showed favorable transfection efficiency, which were used for the further experiments. CCK8, Transwell and FCM were employed to test cell proliferation, migration, invasion and cell apoptosis. The result showed that siCDC6, siCEP55, and siTYMS inhibited cell proliferation, migration, and invasion in NSCLC cells, and promoted cell apoptosis in NSCLC cells (Supplementary Figures 2–4).

DISCUSSION

The prognosis of lung cancer is poor, and the 5-year survival rate after diagnosis is only 16.2% [29]. Because

of the lack of clinical symptoms in the early stage of the disease, when symptoms appear, the preferably treatment opportunity has been missed. In China, on the account of economic conditions and national awareness, there are very few lung cancer patients who can be diagnosed in the early stage [30, 31]. Therefore, early screening of NSCLC has great scientific significance. How to improve the accurate diagnosis, treatment and survival prognosis of NSCLC is particularly important. The TCGA project was started in 2006 by National Cancer Institute and National Human Genome Research Institute. It is an epoch-making project in the field of cancer genomics, which contains research data from different disciplines and institutions. TCGA has identified more than 20000 primary cancers at the
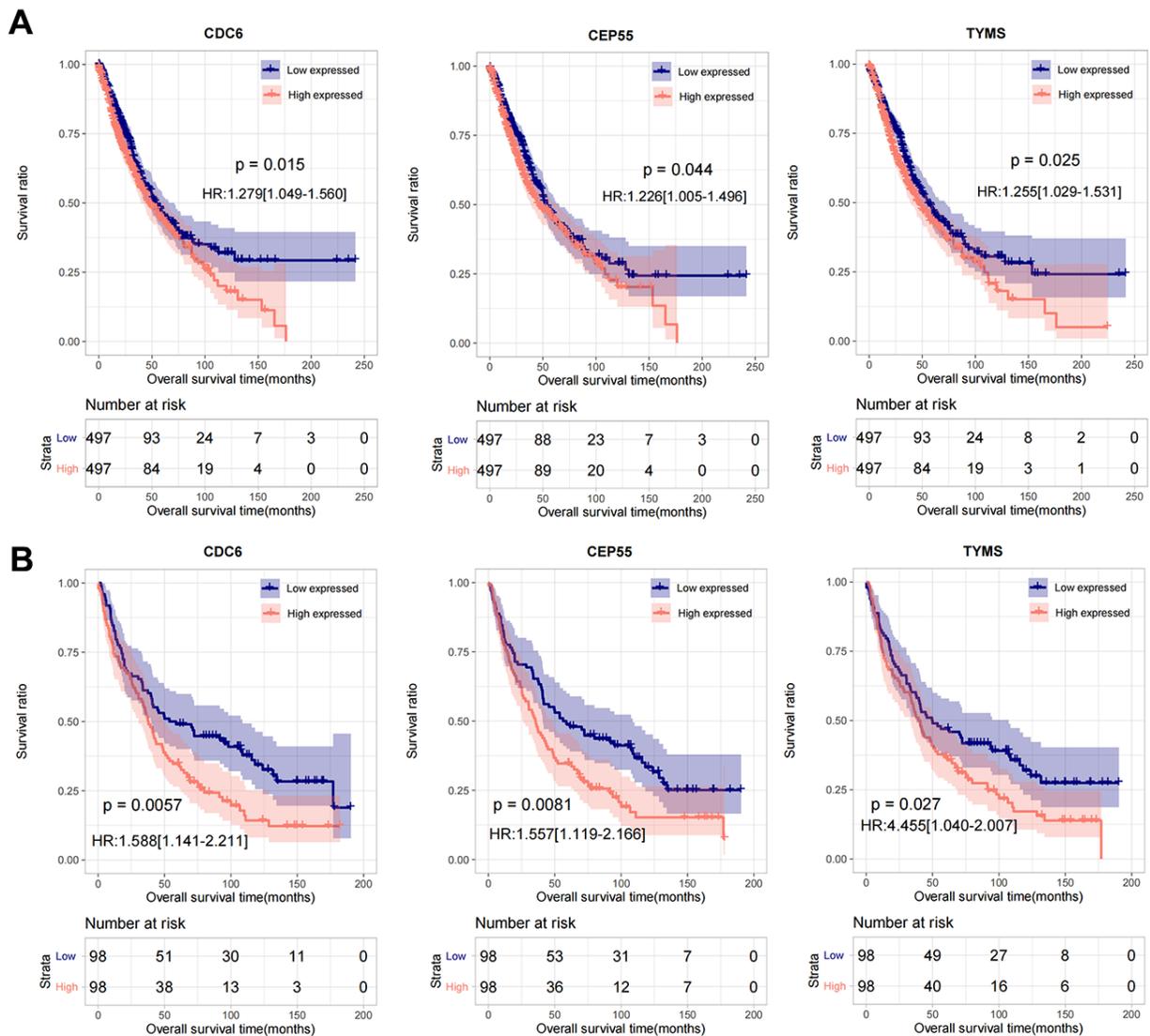
Figure 7. The prognostic analysis of CDC6, CEP55, and TYMS in TCGA and GSE37745. (**A**) Kaplan-Meier used for prognostic analysis of CDC6, CEP55, and TYMS in combined TCGA training set. (**B**) Kaplan-Meier used for prognostic analysis of CDC6, CEP55, and TYMS in GSE37745 validation data set.

molecular level and matched 33 normal tissue samples of cancer [32, 33]. Through the comprehensive analysis of genome, transcriptome and proteome data, valuable biological information about tumor molecular changes can be obtained.

We downloaded the RNA sequence data of LUAD and LUSC from TCGA database. By preliminary analysis, a total of eight survival related long non-coding RNAs (lncRNAs) and 262 survival related mRNAs were filtered. By gene set enrichment analysis, 17 significantly correlated GO pathways and 14 KEGG signal pathways were screened. The GO pathways like GO:0051301-cell division [34], GO:0006468-protein phosphorylation [35], and GO:0000278-mitotic cell cycle [36], which have been reported to exhibit an important role in the occurrence, development, drug resistance and metastasis of NSCLC. The KEGG pathways including hsa05200-Pathways in cancer [37], hsa04110-Cell cycle [38], and hsa04151-PI3K-Akt [39] signaling pathways are also reported to regulate the occurrence and development of NSCLC.

In view of the clinical survival and prognosis information of the samples, we screened eight lncRNAs and 193 mRNAs by single factor Cox regression analysis. Further single and multifactor Cox regression analysis were performed, 30 independent prognostication-related mRNAs were obtained. The PPI network was further constructed. The top ten hub genes were *CDK1*, *CCNB1*, *UBE2C*, *RRM2*, *CCNA2*, *AURKA*,

PLK1, *CDC6*, and *CDC20*, most mRNAs of which have been reported to be involved in the regulation of cell cycle in lung cancer cells [40–43]. The machine learning algorithms (LASSO, RFE, and RF) were employed to screen the optimized DEGs combination, and a total of 17 overlapping mRNAs were obtained. Based on the 17 characteristic mRNAs obtained, we firstly built a Nomogram prediction model. The ROC values of training set and testing set were 0.835 and 0.767 respectively, which suggested that Nomogram prediction model represented favourable performance. By overlapping the 17 characteristic mRNAs and PPI network hub genes, three genes were obtained: *CDC6*, *CEP55*, *TYMS*, which was considered as key factors associated with survival of NSCLC. Allera-Moreau et al. reported that CDC6 was associated with overall, disease-free and relapse-free survival in NSCLC [44]. Another study indicated that *CDC6* was involved in the replication licensing and the proliferation, migration, and invasion of lung cancer cells mediated by miR-26a and miR-26b, and *CDC6* represented potential cancer diagnostic and prognostic markers as well as anti-cancer targets [45]. Centrosome-associated protein 55 kDa (CEP55) is a member of the coiled-coil protein family. Its main function is to anchor microtubule polymerization-associated protein, participate in spindle formation, and then regulate cell proliferation [46]. The protein is expressed in normal tissues and tumor cells, and CEP55 can be coupled with the centrosome and intermediates in the cell cycle. After phosphorylation, it plays a role in regulating the cell cycle. The
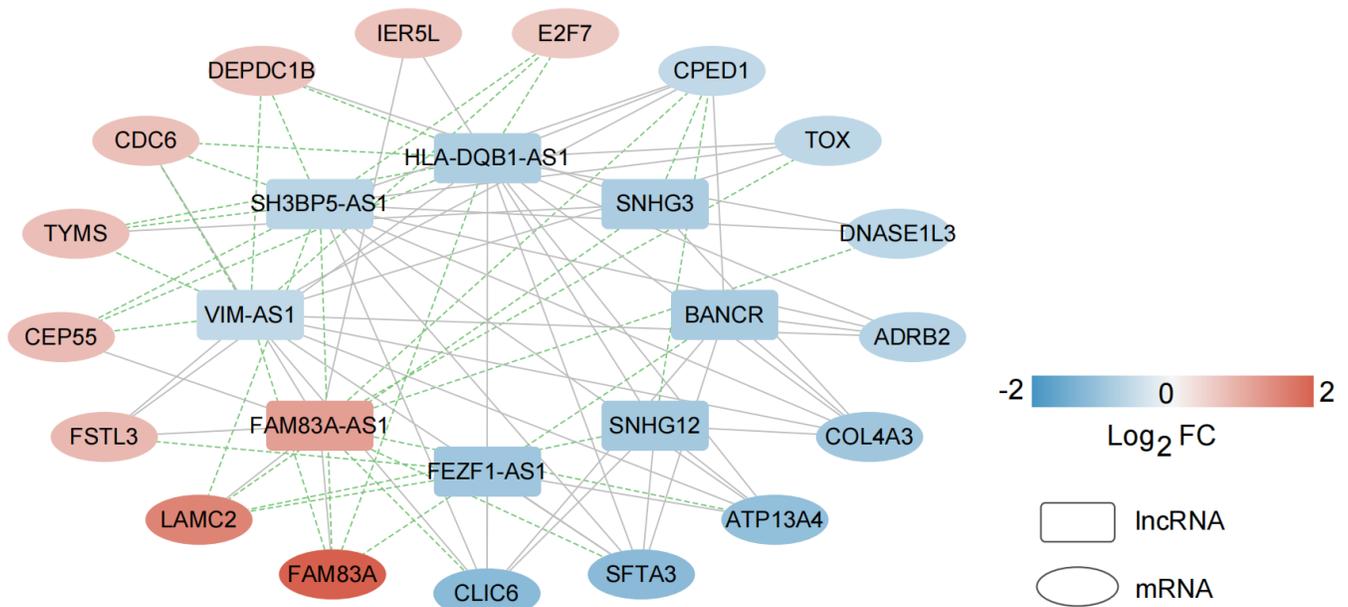
Figure 8. Construction of a co-expression network based on characteristic mRNAs and lncRNAs. A total of 79 pairs of relationship pairs were screened, and the relationship connection network was constructed.

overexpression of *CEP55* is significantly correlated with the tumor stage, invasion and metastasis of many malignant tumors [47]. Jiang et al. found that *CEP55* expression was commonly elevated in NSCLC tissues and overexpression of *CEP55* was correlated with unfavorable prognosis in the patients with NSCLC [48]. Fan et al. reported that *CEP55* expression affected the survival and prognosis of patients with NSCLC, and participated in the process of tumor immune response [49]. Moreover, thymidylate synthetase (TYMS) silencing was reported to increase the pemetrexed sensitivity of NSCLC cells [50]. Zhang et al. demonstrated that significant correlation was observed in *TYMS* expression and clinical features, especially histology in NSCLC [51]. Tsyganov et al. demonstrated that exploring *TYMS* expression could contribute to the personalized chemotherapy, which can improve treatment efficacy and reduce unnecessary toxicity [52]. In the present study, A549 and H1299 cells were transfected with siCDC6, siCEP55, and siTYMS, respectively. Cell proliferation, migration, invasion and apoptosis were examined. The results presented that silencing of *CDC6*, *CEP55*, *TYMS* showed carcinostatic effect on NSCLC cells. Finally, the lncRNAs-mRNAs networks were constructed, and a total of 79 pairs of relationship pairs were screened. There are some limitations in the present study. The relationship between lncRNAs and mRNAs and immune cells, the analysis of drug sensitivity data, and the analysis of microRNAs data based on public databases are not studied in the present study. These works can be part of our future work.

In conclusion, this study explored the lncRNAs and mRNAs related to survival of NSCLC based on bioinformatic analysis and machine learning. We firstly built a Nomogram prediction model, which exhibited favourable performance. *CDC6*, *CEP55*, and *TYMS* are considered as key factors associated with survival of NSCLC. This paper provides a new idea for the early screening of NSCLC.

AUTHOR CONTRIBUTIONS

Wei Yue: substantial contributions to conception and design, data acquisition, drafting the article; Jing Wang: data acquisition, drafting the article; Bo Lin: data acquisition; reviewing the article; Yongping Fu: experiment performing, data analysis. All the authors took part in the experiment. All the authors read and approved the manuscript.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Basumallik N, Agarwal M. Small Cell Lung Cancer. 2023. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2024.
   PMID:29494065

2. Roy-Chowdhuri S. Molecular Pathology of Lung Cancer. Surg Pathol Clin. 2021; 14:369–77.
   https://doi.org/10.1016/j.path.2021.05.002
   PMID:34373089

3. Xia C, Dong X, Li H, Cao M, Sun D, He S, Yang F, Yan X, Zhang S, Li N, Chen W. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chin Med J (Engl). 2022; 135:584–90.
   https://doi.org/10.1097/CM9.0000000000002108
   PMID:35143424

4. Sun D, Li H, Cao M, He S, Lei L, Peng J, Chen W. Cancer burden in China: trends, risk factors and prevention. Cancer Biol Med. 2020; 17:879–95.
   https://doi.org/10.20892/j.issn.2095-3941.2020.0387
   PMID:33299641

5. Li Y, Wu X, Yang P, Jiang G, Luo Y. Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis. Genomics Proteomics Bioinformatics. 2022; 20:850–66.
   https://doi.org/10.1016/j.gpb.2022.11.003
   PMID:36462630

6. Gould MK, Huang BZ, Tammemagi MC, Kinar Y, Shiff R. Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data. Am J Respir Crit Care Med. 2021; 204:445–53.
   https://doi.org/10.1164/rccm.202007-2791OC
   PMID:33823116

7. Cheng J, Pan Y, Huang W, Huang K, Cui Y, Hong W, Wang L, Ni D, Tan P. Differentiation between immune checkpoint inhibitor-related and radiation pneumonitis in lung cancer by CT radiomics and machine learning. Med Phys. 2022; 49:1547–58.
   https://doi.org/10.1002/mp.15451 PMID:35026041

8. Wu J, Lu L, Chen H, Lin Y, Zhang H, Chen E, Lin W, Li J, Chen X. Prognostic nomogram to predict the overall survival of patients with early-onset colorectal cancer: a population-based analysis. Int J Colorectal Dis. 2021; 36:1981–93.
   https://doi.org/10.1007/s00384-021-03992-w
   PMID:34322745

9. Mao W, Fu Z, Wang K, Wu J, Xu B, Chen M. Prognostic nomogram for patients with lung metastatic renal cell carcinoma: a SEER-based study. Ann Palliat Med. 2021; 10:2791–804.
https://doi.org/10.21037/apm-20-1488
PMID:33549017

10. Yue CY, Gao JP, Zhang CY, Ni YH, Ying CM. Development and validation of a nomogram for the early prediction of preeclampsia in pregnant Chinese women. Hypertens Res. 2021; 44:417–25.
https://doi.org/10.1038/s41440-020-00558-1
PMID:33060833

11. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012; 28:882–3.
https://doi.org/10.1093/bioinformatics/bts034
PMID:22257669

12. Jabs V, Edlund K, König H, Grinberg M, Madjar K, Rahnenführer J, Ekman S, Bergkvist M, Holmberg L, Ickstadt K, Botling J, Hengstler JG, Micke P. Integrative analysis of genome-wide gene copy number changes and gene expression in non-small cell lung cancer. PLoS One. 2017; 12:e0187246.
https://doi.org/10.1371/journal.pone.0187246
PMID:29112949

13. Lohr M, Hellwig B, Edlund K, Mattsson JS, Botling J, Schmidt M, Hengstler JG, Micke P, Rahnenführer J. Identification of sample annotation errors in gene expression datasets. Arch Toxicol. 2015; 89:2265–72.
https://doi.org/10.1007/s00204-015-1632-4
PMID:26608184

14. Newman V, Moore B, Sparrow H, Perry E. The Ensembl Genome Browser: Strategies for Accessing Eukaryotic Genome Data. Methods Mol Biol. 2018; 1757:115–39.
https://doi.org/10.1007/978-1-4939-7737-6_6
PMID:29761458

15. Sun Y, Luo J, Qian C, Luo L, Xu M, Min H, Cen Y. The Value of Nutritional Status in the Prognostic Analysis of Patients with AIDS-Related Lymphoma. Infect Drug Resist. 2021; 14:1105–13.
https://doi.org/10.2147/IDR.S295077
PMID:33776456

16. Yan LP, Liu ZB, Wu M, Ge YP, Zhang Q. Effect of lncRNA MALAT1 expression on survival status of elderly patients with severe pneumonia. Eur Rev Med Pharmacol Sci. 2020; 24:3959–64.
https://doi.org/10.26355/eurrev_202004_20865
PMID:32329872

17. Zhu Y, Du Z, Zhu Y, Li W, Miao H, Li Z. Evaluation of organ function in patients with severe COVID-19 infections. Med Clin (Engl Ed). 2020; 155:191–6.
https://doi.org/10.1016/j.medcle.2020.05.015
PMID:32984539

18. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4:44–57.
https://doi.org/10.1038/nprot.2008.211
PMID:19131956

19. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37:1–13.
https://doi.org/10.1093/nar/gkn923 PMID:19033363

20. Wang P, Wang Y, Hang B, Zou X, Mao JH. A novel gene expression-based prognostic scoring system to predict survival in gastric cancer. Oncotarget. 2016; 7:55343–51.
https://doi.org/10.18632/oncotarget.10533
PMID:27419373

21. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017; 45:D362–8.
https://doi.org/10.1093/nar/gkw937 PMID:27924014

22. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–504.
https://doi.org/10.1101/gr.1239303 PMID:14597658

23. Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. Biom J. 2010; 52:70–84.
https://doi.org/10.1002/bimj.200900028
PMID:19937997

24. Deist TM, Dankers FJ, Valdes G, Wijsman R, Hsu IC, Oberije C, Lustberg T, van Soest J, Hoebers F, Jochems A, El Naqa I, Wee L, Morin O, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. Med Phys. 2018; 45:3449–59.
https://doi.org/10.1002/mp.12967 PMID:29763967

25. Tolosi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics. 2011; 27:1986–94.
https://doi.org/10.1093/bioinformatics/btr300
PMID:21576180

26. Wu J, Zhang H, Li L, Hu M, Chen L, Xu B, Song Q. A nomogram for predicting overall survival in patients with low-grade endometrial stromal sarcoma: A population-based analysis. Cancer Commun (Lond). 2020; 40:301–12.

https://doi.org/10.1002/cac2.12067
PMID:32558385

27. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996; 15:361–87.
https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4
PMID:8668867

28. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12:77.
https://doi.org/10.1186/1471-2105-12-77
PMID:21414208

29. Brody H. Lung cancer. Nature. 2020; 587:S7.
https://doi.org/10.1038/d41586-020-03152-0
PMID:33208969

30. Li W, Liu JB, Hou LK, Yu F, Zhang J, Wu W, Tang XM, Sun F, Lu HM, Deng J, Bai J, Li J, Wu CY, et al. Liquid biopsy in lung cancer: significance in diagnostics, prediction, and treatment monitoring. Mol Cancer. 2022; 21:25.
https://doi.org/10.1186/s12943-022-01505-z
PMID:35057806

31. Luo YH, Chiu CH, Scott Kuo CH, Chou TY, Yeh YC, Hsu HS, Yen SH, Wu YH, Yang JC, Liao BC, Hsia TC, Chen YM. Lung Cancer in Republic of China. J Thorac Oncol. 2021; 16:519–27.
https://doi.org/10.1016/j.jtho.2020.10.155
PMID:33781442

32. Blum A, Wang P, Zenklusen JC. SnapShot: TCGA-Analyzed Tumors. Cell. 2018; 173:530.
https://doi.org/10.1016/j.cell.2018.03.059
PMID:29625059

33. Travaglino A, Raffone A, Mascolo M, Guida M, Insabato L, Zannoni GF, Zullo F. TCGA Molecular Subgroups in Endometrial Undifferentiated/Dedifferentiated Carcinoma. Pathol Oncol Res. 2020; 26:1411–6.
https://doi.org/10.1007/s12253-019-00784-0
PMID:31811476

34. Canman JC, Cabernard C. Mechanics of cell division and cytokinesis. Mol Biol Cell. 2018; 29:685–6.
https://doi.org/10.1091/mbc.E17-11-0671
PMID:29535170

35. Humphrey SJ, James DE, Mann M. Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. Trends Endocrinol Metab. 2015; 26:676–87.
https://doi.org/10.1016/j.tem.2015.09.013
PMID:26498855

36. Ni L, Zhu X, Gong C, Luo Y, Wang L, Zhou W, Zhu S, Li Y. Trichosanthes kirilowii fruits inhibit non-small cell lung cancer cell growth through mitotic cell-cycle arrest. Am J Chin Med. 2015; 43:349–64.
https://doi.org/10.1142/S0192415X15500238
PMID:25779643

37. Liu Q, Li A, Tian Y, Wu JD, Liu Y, Li T, Chen Y, Han X, Wu K. The CXCL8-CXCR1/2 pathways in cancer. Cytokine Growth Factor Rev. 2016; 31:61–71.
https://doi.org/10.1016/j.cytogfr.2016.08.002
PMID:27578214

38. Fariha A, Hami I, Tonmoy MI, Akter S, Al Reza H, Bahadur NM, Rahaman MM, Hossain MS. Cell cycle associated miRNAs as target and therapeutics in lung cancer treatment. Heliyon. 2022; 8:e11081.
https://doi.org/10.1016/j.heliyon.2022.e11081
PMID:36303933

39. Ik, Pothongsrisit S, Pongrakhananon V. Targeting the PI3K/AKT/mTOR Signaling Pathway in Lung Cancer: An Update Regarding Potential Drugs and Natural Products. Molecules. 2021; 26:4100.
https://doi.org/10.3390/molecules26134100
PMID:34279440

40. Zhang L, Peng R, Sun Y, Wang J, Chong X, Zhang Z. Identification of key genes in non-small cell lung cancer by bioinformatics analysis. PeerJ. 2019; 7:e8215.
https://doi.org/10.7717/peerj.8215 PMID:31844590

41. Hu Y, Chen G. Pathogenic mechanisms of lung adenocarcinoma in smokers and non-smokers determined by gene expression interrogation. Oncol Lett. 2015; 10:1350–70.
https://doi.org/10.3892/ol.2015.3462 PMID:26622675

42. Yuan B, Wang X, Fan C, You J, Liu Y, Weber JD, Zhong H, Zhang Y. DHX33 Transcriptionally Controls Genes Involved in the Cell Cycle. Mol Cell Biol. 2016; 36:2903–17.
https://doi.org/10.1128/MCB.00314-16
PMID:27601587

43. Mengyan X, Kun D, Xinming J, Yutian W, Yongqian S. Identification and verification of hub genes associated with the progression of non-small cell lung cancer by integrated analysis. Front Pharmacol. 2022; 13:997842.
https://doi.org/10.3389/fphar.2022.997842
PMID:36176446

44. Allera-Moreau C, Rouquette I, Lepage B, Oumouhou N, Walschaerts M, Leconte E, Schilling V, Gordien K, Brouchet L, Delisle MB, Mazieres J, Hoffmann JS, Pasero P, Cazaux C. DNA replication stress response involving PLK1, CDC6, POLQ, RAD51 and CLASPIN upregulation prognoses the outcome of early/mid-stage non-small cell lung cancer patients. Oncogenesis. 2012; 1:e30.

https://doi.org/10.1038/oncsis.2012.29
PMID:23552402

45. Zhang X, Xiao D, Wang Z, Zou Y, Huang L, Lin W, Deng Q, Pan H, Zhou J, Liang C, He J. MicroRNA-26a/b regulate DNA replication licensing, tumorigenesis, and prognosis by targeting CDC6 in lung cancer. Mol Cancer Res. 2014; 12:1535–46.
https://doi.org/10.1158/1541-7786.MCR-13-0641
PMID:25100863

46. Wu L, Chen Y, Wan L, Wen Z, Liu R, Li L, Song Y, Wang L. Identification of unique transcriptomic signatures and key genes through RNA sequencing and integrated WGCNA and PPI network analysis in HIV infected lung cancer. Cancer Med. 2023; 12:949–60.
https://doi.org/10.1002/cam4.4853 PMID:35608130

47. Liu L, Mei Q, Zhao J, Dai Y, Fu Q. Suppression of CEP55 reduces cell viability and induces apoptosis in human lung cancer. Oncol Rep. 2016; 36:1939–45.
https://doi.org/10.3892/or.2016.5059 PMID:27633074

48. Jiang C, Zhang Y, Li Y, Lu J, Huang Q, Xu R, Feng Y, Yan S. High CEP55 expression is associated with poor prognosis in non-small-cell lung cancer. Onco Targets Ther. 2018; 11:4979–90.
https://doi.org/10.2147/OTT.S165750 PMID:30154666

49. Fan H, Zhang J, Zou B, He Z. The Role of CEP55 Expression in Tumor Immune Response and Prognosis of Patients with Non-small Cell lung Cancer. Arch Iran Med. 2022; 25:432–42.
https://doi.org/10.34172/aim.2022.72 PMID:36404510

50. Agulló-Ortuño MT, García-Ruiz I, Díaz-García CV, Enguita AB, Pardo-Marqués V, Prieto-García E, Ponce S, Iglesias L, Zugazagoitia J, López-Martín JA, Paz-Ares L, Nuñez JA. Blood mRNA expression of REV3L and TYMS as potential predictive biomarkers from platinum-based chemotherapy plus pemetrexed in non-small cell lung cancer patients. Cancer Chemother Pharmacol. 2020; 85:525–35.
https://doi.org/10.1007/s00280-019-04008-9
PMID:31832811

51. Zhang Q, Sun T, Kang P, Qian K, Deng B, Zhou J, Wang R, Jiang B, Li K, Liu F, Wu S, Tan Q. Combined analysis of rearrangement of ALK, ROS1, somatic mutation of EGFR, KRAS, BRAF, PIK3CA, and mRNA expression of ERCC1, TYMS, RRM1, TUBB3, EGFR in patients with non-small cell lung cancer and their clinical significance. Cancer Chemother Pharmacol. 2016; 77:583–93.
https://doi.org/10.1007/s00280-016-2969-y
PMID:26842788

52. Tsyganov MM, Rodionov EO, Ibragimova MK, Miller SV, Cheremisina OV, Frolova IG, Tuzikov SA, Litviakov NV. Personalized Prescription of Chemotherapy Based on Assessment of mRNA Expression of BRCA1, RRM1, ERCC1, TOP1, TOP2α, TUBβ3, TYMS, and GSTP1 Genes in Tumors Compared to Standard Chemotherapy in the Treatment of Non-Small-Cell Lung Cancer. J Pers Med. 2022; 12:1647.
https://doi.org/10.3390/jpm12101647
PMID:36294786
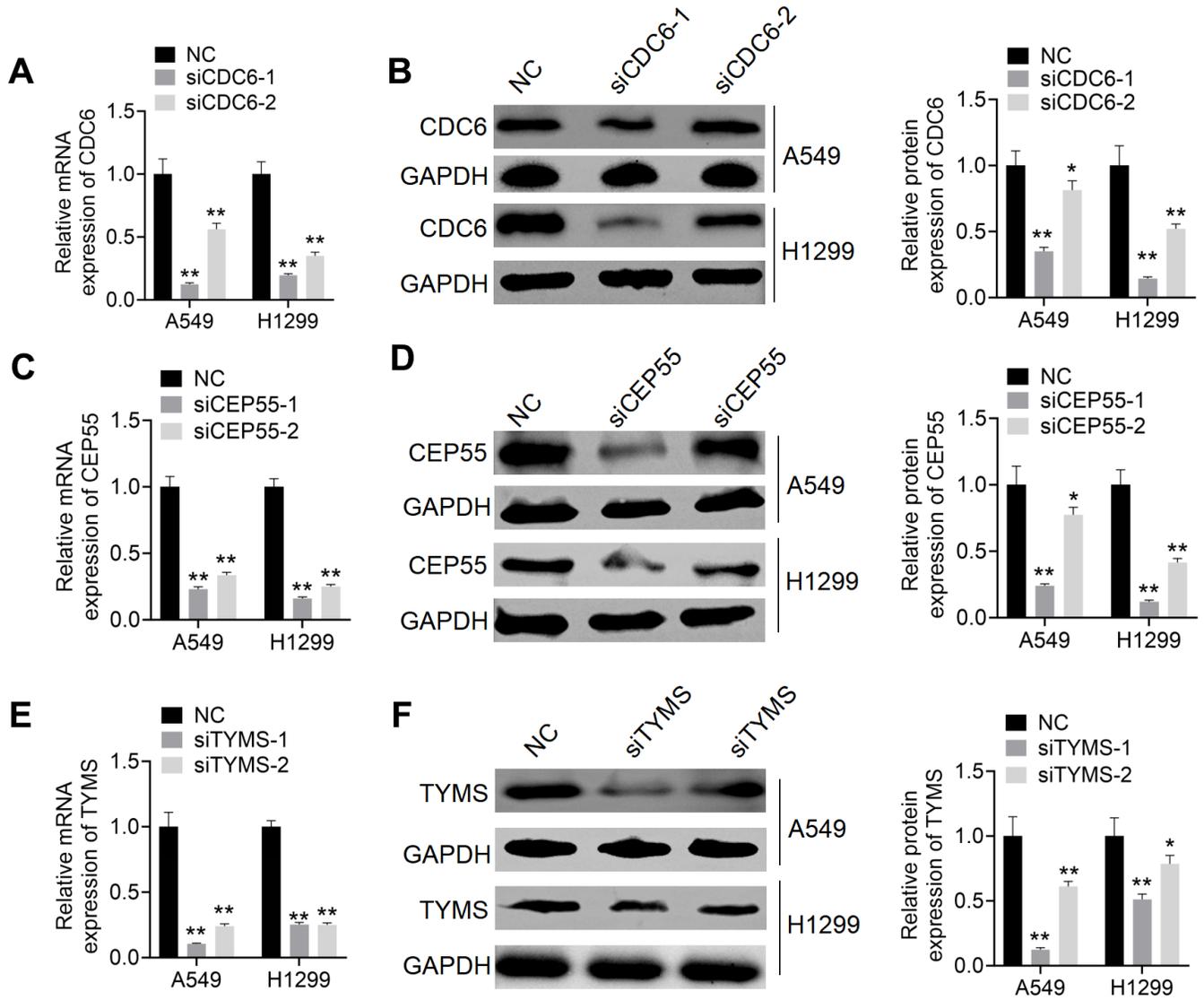
Supplementary Figure 1. Transfection efficiency was examined by PCR and western blot. (**A**, **B**) RT-PCR and western blot were performed to examine the transfection efficiency of CDC6 knockdown. (**C**, **D**) RT-PCR and western blot were performed to examine the transfection efficiency of CEP55 knockdown. (**E**, **F**) RT-PCR and western blot were performed to examine the transfection efficiency of TYMS knockdown. *$P<0.05$, **$P<0.01$ vs NC group.
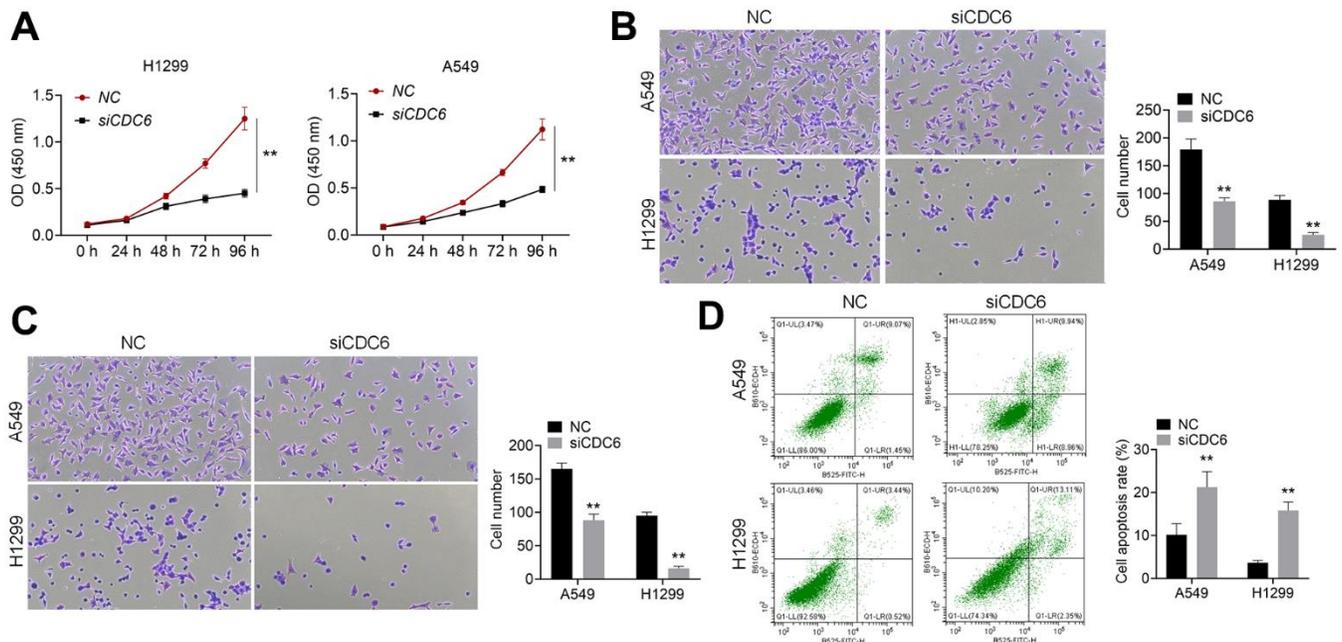
Supplementary Figure 2. Effect of CDC6 knockdown on malignant oncology phenotype in NSCLC cells. (**A**) Cell proliferation was examined by CCK8. (**B**, **C**) Cell migration and invasion were tested by Transwell assay. (**D**) Cell apoptosis was evaluated by FITC. *$P<0.05$, **$P<0.01$ vs NC group.
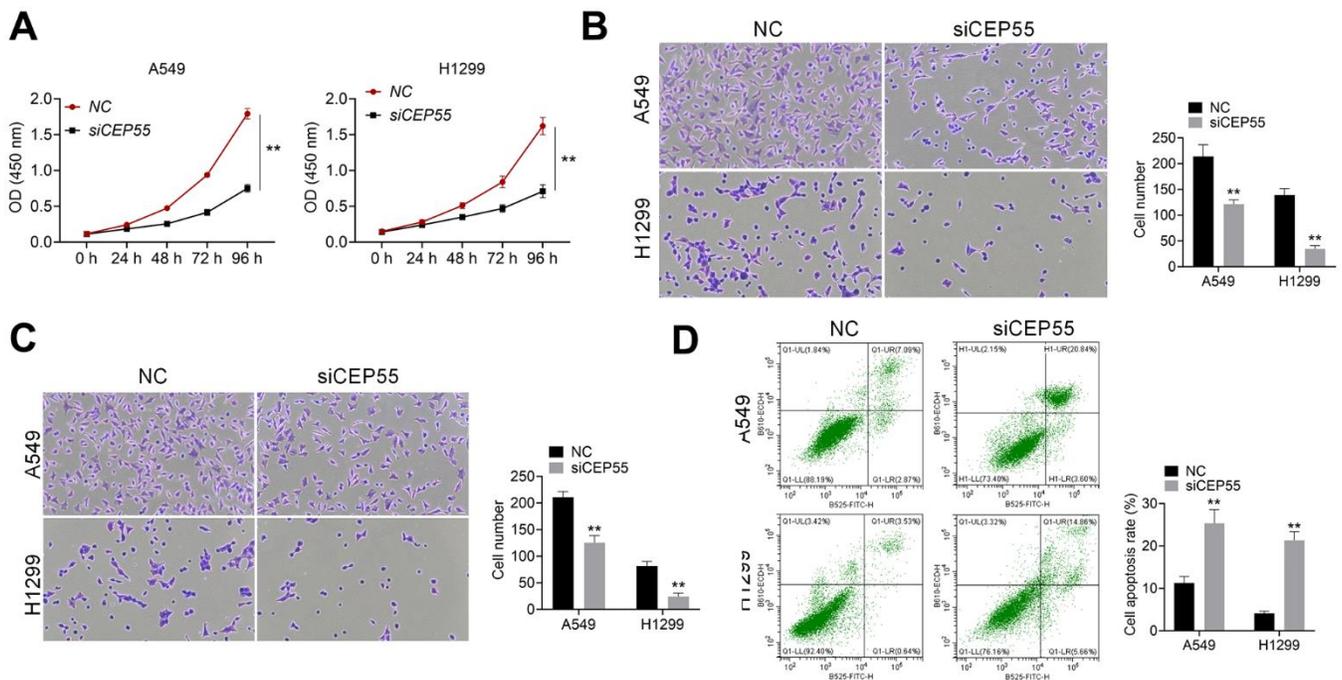
Supplementary Figure 3. Effect of CEP55 knockdown on malignant oncology phenotype in NSCLC cells. (**A**) Cell proliferation was examined by CCK8. (**B**, **C**) Cell migration and invasion were tested by Transwell assay. (**D**) Cell apoptosis was evaluated by FITC. *$P<0.05$, **$P<0.01$ vs NC group.
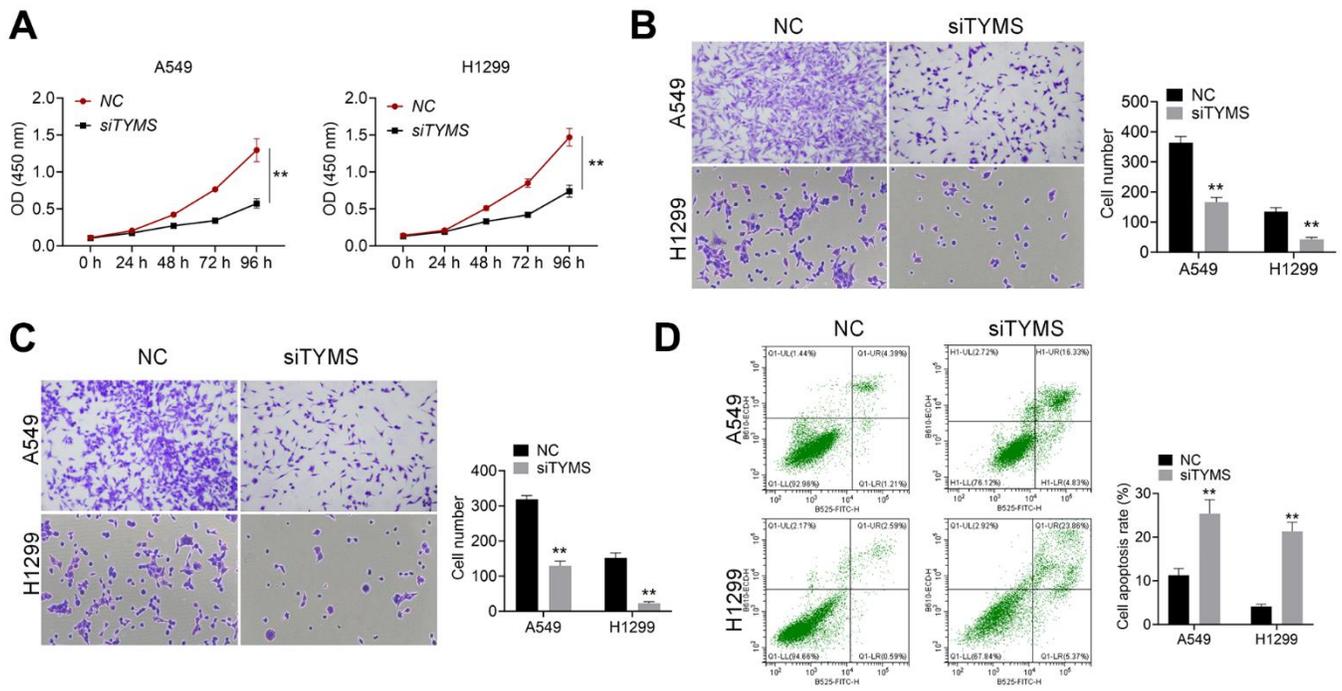
Supplementary Figure 4. Effect of TYMS knockdown on malignant oncology phenotype in NSCLC cells. (**A**) Cell proliferation was examined by CCK8. (**B**, **C**) Cell migration and invasion were tested by Transwell assay. (**D**) Cell apoptosis was evaluated by FITC. *$P<0.05$, **$P<0.01$ vs NC group.

Supplementary Tables

Please browse the Full Text version to see the data of Supplementary Tables 1, 2.

Supplementary Table 1. DERs filtered in tumor vs. control and dead vs. alive.

Supplementary Table 2. 30 independent prognostication-related mRNAs.