BMC Cancer

RESEARCH ARTICLE

Open Access

Five long non-coding RNAs establish a prognostic nomogram and construct a competing endogenous RNA network in the progression of non-small cell lung cancer

Yong Yu[1] and Kaiming Ren[2*]

Abstract

Background: Accumulating evidence has revealed that long non-coding RNAs (lncRNAs) play vital roles in the progression of non-small cell lung cancer (NSCLC). But the relationship between lncRNAs and survival outcome of NSCLC remains to be explored. Therefore, we attempt to figure out their survival roles and molecular connection in NSCLC.

Methods: By analyzing the transcriptome profiling of NSCLC from TCGA databases, we divided patients into three groups, and identified differentially expressed lncRNAs (DELs) of each group. Next, we explored the prognostic roles of common DELs by univariate and multivariate Cox analysis, Lasson, and Kaplan-Meier analysis. Additionally, we assessed and compared the prognostic accuracy of 5 lncRNAs through ROC curves and AUC values. Ultimately, we detected their potential function by enrichment analysis and molecular connection through establishing a competing endogenous RNA (ceRNA) network.

Results: One hundred ninety-seven common DELs were spotted. And we successfully screened out 5 lncRNAs related to the patient's survival, including LINC01833, AC112206.2, FAM83A-AS1, BANCR, and HOTAIR. Combing with age and AJCC stage, we constructed a nomogram that prognostic prediction was superior to the traditional parameters. Furthermore, 275 qualified mRNAs related to 5 lncRNAs were spotted. Functional analysis indicates that these lncRNAs act key roles in the progression of NSCLC, such as P53 and cell cycle signaling pathway. And ceRNA network also suggests that these lncRNAs are tightly connected with tumor progression.

Conclusions: A nomogram and ceRNA network based on 5 lncRNAs indicate that there can effectively predict the overall survival of NSCLC and potentially serve as a therapeutic guide for NSCLC.

Keywords: NSCLC, Long non-coding RNA, Nomogram, ceRNA, Overall survival

* Correspondence: renkmcmu@163.com
[2]Department of Thoracic Surgery, Shengjing Hospital of China Medical
University, Shenyang 110004, Liaoning, China
Full list of author information is available at the end of the article

Background

Non-small cell lung cancer (NSCLC) is one of the most common and deadly cancers in the world. Despite advances in treatments, only 19% of patients with NSCLC survival for more than 5 years [1, 2]. What's worse, due to lacking specific symptoms in the early stage, most patients seek treatment at an advanced stage, which misses the best timing for a radical operation [3]. Therefore, it is urgent to establish a prognostic risk-score model for NSCLC patients, thereby providing a therapeutic guide for NSCLC.

Long non-coding RNAs (lncRNAs) refers to non-protein-coding transcripts over 200 nucleotides in length [4]. Although lncRNAs do not directly encode RNA, it can regulate protein expression at various stages of transcription [5, 6]. Based on the ceRNA hypothesis, lncRNAs, messenger RNAs, and pseudogenes can "talk" to each other using miRNA response elements (MREs) and assemble as a ceRNA network [7]. In this network, lncRNA act as "sponges" to absorb and bind miRNA, thereby weakening their binding ability to mRNA and regulating gene expression. Accumulating evidence has backed that lncRNAs were involved in the ceRNA network of many types of cancers, including pancreatic cancer, gastric cancer, as well as NSCLC [8–10]. Notably, LncRNAs have great advantages as biomarkers because they are stable, highly tissue-specific, and easy to detect in body fluids. Besides, some lncRNAs have been recognized as a novel biomarker, for instance, BANCR in gastric carcinoma, HOTAIR in colorectal carcinoma, and MALAT1 in lung cancer [11–13]. Therefore, it is needful to detect the prognostic relationship between lncRNAs and NSCLC.

In this study, we performed a large sample analysis to find out survival-related lncRNAs and validated it using univariate, Lasson, and multivariate Cox proportional hazards regression (CPHR). Moreover, combing with age and AJCC stage, we constructed a nomogram based on lncRNAs, which performed better prognostic prediction than clinical factors, and we successfully assessed its efficiency in LUAD and LUSC groups. Functional analysis indicated that these lncRNAs also act important roles in the progression of NSCLC, for example, the P53 signaling pathway and cell cycle pathway. Next, we successfully constructed a ceRNA network related to 5 lncRNAs. Those results indicated that those lncRNAs not only effectively predict the prognosis of NSCLC patients but also take part in the progression of NSCLC and potentially serve as a therapeutic target.

Methods

Data selection and process

As is well know that the majority of NSCLC is composed of lung squamous carcinoma (LUSC) and lung adenocarcinoma (LUAD). So we obtained the raw counts of transcriptome profiling and clinical data from TCGA-LUAD including 535 cancer samples and 59 non-tumor tissues, and data from TCGA-LUSC including 502 cancer samples and 49 non-tumor tissues. We obtained the RNA-Seq data for lncRNA and mRNA analysis and downloaded the miRNA-Seq data for miRNA analysis.

Since there was only one patient in the TCGA-MESO database, we included TCGA_LUAD and TCGA_LUSC in the NSCLC group with 1145 patients. Then we divided those data into three groups including the NSCLC group, LUAD group, and LUSC group. All the raw count data of lncRNAs, miRNAs, and mRNAs were obtained from The Cancer Genome Atlas (TCGA) database and were annotated through the Ensemble database (Homo_sapiens.GRCh38.99) [14, 15]. Stepwise, all the data were log2 (x + 1) transformed and normalized using the "LIMMA" package [16]. The "LIMMA" package is a widely used tool for calculating gene differential expression [17, 18]. This calculation process is performed automatically by R software. Additionally, to avoid bias caused by low expression genes, we removed differential genes with low average expression. Then, we screened out differentially expressed miRNAs (DEMIs) in three groups by the "LIMMA" package with the criteria of |log2FC| > 1, average expression > 1, and adjust P-value < 0.05, and identified differentially expressed lncRNAs (DELs) and differentially expressed mRNAs (DEMs) in three groups with the criteria of |log2FC| > 2, average expression > 2, and adjust P-value < 0.05, respectively. Finally, we selected the DELs, DEMs and DEMIs that both expressed in three groups for further analysis to reduce the bias caused by a single database.

Survival analysis

Additionally, we removed samples without survival time or survival time less than 7 days to improve the reliability of our study. We first estimated the association between overall survival (OS) and clinical parameters through univariate and multivariate CPHR analysis. Furthermore, we evaluated the relationship between survival time and common DELs expression through Kaplan Meier analysis and univariate CPHR method. Only DELs that their P-value was lower than 0.05 and their expression consistent with prognosis were regarded as candidate survival-related lncRNAs. Combing with clinical risk factors, we performed Lasson Cox regression analysis to obtain the best fitting variables. After selecting the best-fit of OS-related variables by the calculation mentioned above, we further verified their prognostic value through multivariate CPHR analysis. And only variables that P-value was lower than 0.05 in univariate and

multivariate CPHR calculations were selected as qualified lncRNAs and were chose for the next analysis.

Constructing a risk score formula and nomogram model

Combining with clinical risk variables and qualified lncRNAs, we performed univariate and multivariate CPHR analysis to identify OS-related biomarkers in 970 patients with NSCLC. Also, we calculated the prognostic risk score of each patient through multivariate CPHR analysis according to the formula as follows: risk score = $X_1\alpha_1 + X_2\alpha_2 + X_3\alpha_3 + ... + X_n\alpha_n$. And patients in the NSCLC group were divided into high-risk groups and low-risk groups based on the median value of risk score. C-indexes were performed to assess the predictive performances of our risk score formula. Next, we evaluated the prognostic differences in high-risk and low-risk groups by Kaplan Meier analysis and T-test. We assessed the prediction performance of the risk formula through ROC curves at 3 and 5 years and computed their AUC values in three groups. Moreover, we established a nomogram to vividly depict the predictive relationship among clinical factors, lncRNAs, and OS. Calibration curves of 3 and 5 years were calculated to assess the reliability of OS prediction between predicted performance and actual ability. All the analyses mentioned above were conducted in NSCLC, LUAD, and LUSC groups, respectively.

Functional enrichment analysis

We performed the Gene Ontology (GO) terms and Kyoto Encyclopedia of Gene and Genomes (KEGG) pathways enrichment analysis to elucidate the potential functions of lncRNAs in the nomogram. The common DEMs in the three groups were first identified. Next, the correlation coefficient of each lncRNA with common DEMs was calculated, respectively. To obtain an accurate result, we only selected DEMs with a correlation coefficient greater than 0.2 for further enrichment analysis. Then, we performed the GO and KEGG analysis of lncRNA-related DEMs via the DAVID database, and the Enrichr database [19, 20]. And we depicted the top 10 enriched GO terms and KEGG pathways through a bar plot with the criteria of adjusting *P*-value < 0.05.

Establishing a ceRNA network

To explore the potential interaction between 5 lincRNAs and miRNAs, the LncBase database that provided miR-NAs and lncRNAs interactions according to MREs sites was applied to predict the downstream miRNA of 5 lncRNAs with the criteria of Prediction score > 0.8 [21]. Only miRNAs expressed on the NSCLC group, LUAD group, LUSC group, and LncBase database were chosen as qualified miRNAs. Additionally, the miRDB database and the miRTarBase database are a widely-used tool for

miRNA target prediction that were employed to find out potential mRNAs binding to qualified miRNAs [22, 23]. Only miRNAs that validated in miRDB database, miR-TarBase database, and DEMs were considered as candidate mRNAs. Finally, to vividly display the interaction of 5 lncRNAs with qualified miRNAs and candidate mRNAs, we constructed a ceRNA network by Cytoscape software (Version 3.7.2) [24].

Statistical analysis

We performed all the statistical analyses mentioned above using R software (version 3.6.1). Briefly, OS was analyzed using the Kaplan–Meier test, and the log-rank T-test was applied to calculate the statistical significance. Univariate and multivariate CPHR analyses were conducted through "Survival" packages (Version 3.2–7) [25]. Lasson CPHR was performed using "Survival" and "glmnet" packages (Version 4.0–2) [26]. A nomogram was constructed by "Survival" and "rms" packages (Version 6.1–0) [27]. And a time-dependent ROC analysis was performed by the "survivalROC" package (Version 1.0.3) [28], C-index by "survival" package, and calibration curve by "rms" package. The ceRNA network was constructed by Cytoscape software. And we set a *p*-value lower than 0.05 as s statistical significance.

Results

Screening differentially expressed RNAs

A total of 1145 NSCLC patients were enrolled from the TCGA database. We divided those data into three groups including the NSCLC group, LUAD group, and LUSC group. Then, we identified the DELs and DEMs in three groups by standards of |log2FC| > 2, average expression > 2, and adjust *P* value < 0.05, respectively. There are a total of 426 DELs in the NSCLC group, 312 DELs in the LUAD group, 687 DELs in the LUSC group, and 197 common DELs in three groups (Fig. 1a–d). There is a total of 1905 DEMs in the NSCLC group, 1434 DEMs in the LUAD group, 2641 DEMs in the LUSC group, and 1131 common DEMs in three groups (Supplementary Figure 1A–D). And there is a total of 130 DEMIs in the NSCLC group, 133 DEMIs in the LUAD group, 163 DEMIs in the LUSC group, and 89 common DEMIs in three groups (Supplementary Figure 5A–D) with the criteria of |log2FC| > 1, average expression > 1, and adjust *P*-value < 0.05. Also, we present the top 10 up and down-regulated DELs spotted in three groups (Fig. 1e). Common DELs in three groups were selected for the next survival analysis (Supplementary Table 5).

Survival analysis

After excluding insufficient survival data or survival time less than 7 days, 970 NSCLC patients remained in our
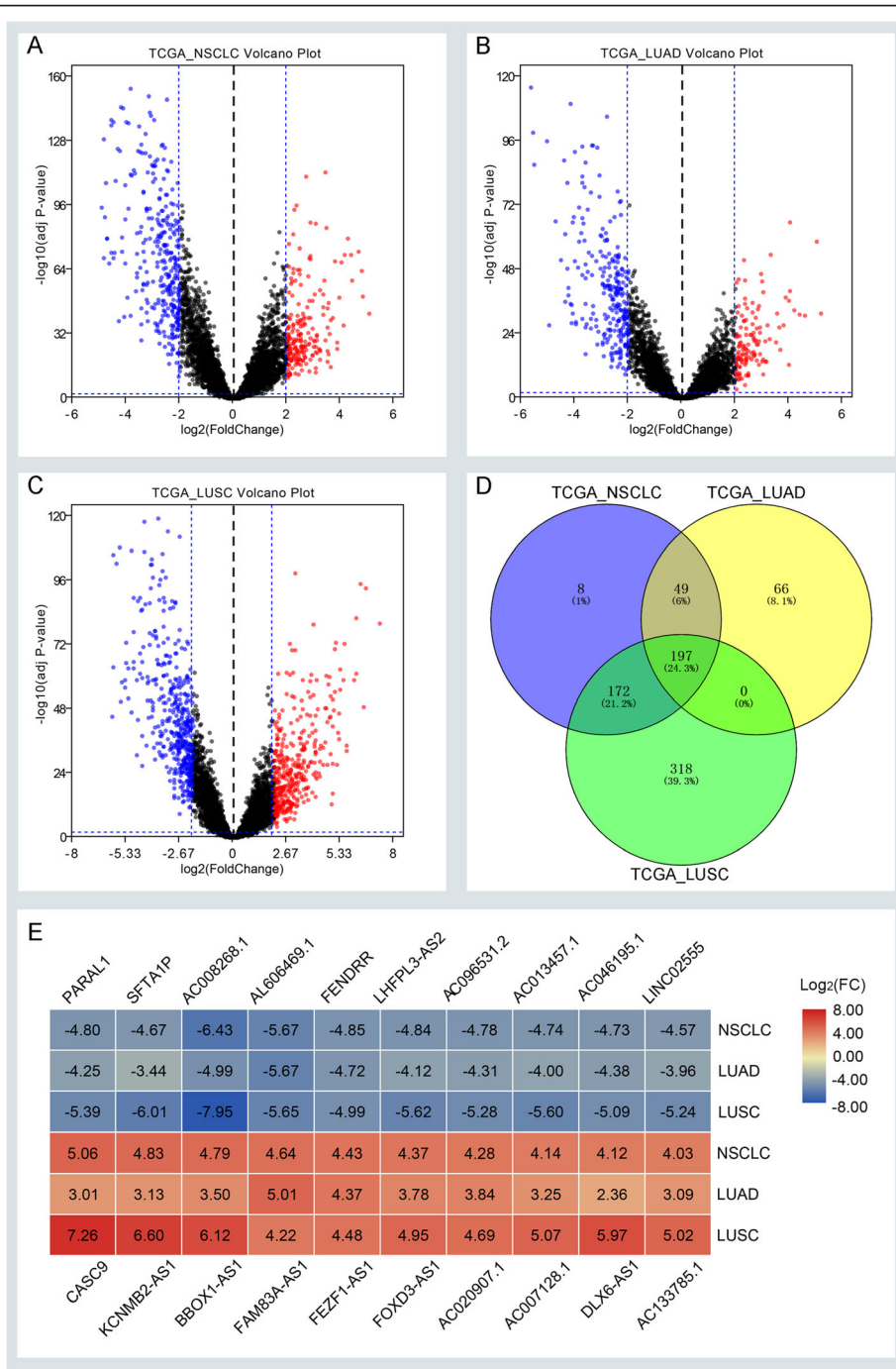
Fig. 1 Screening differentially expressed lncRNAs (DELs) in three groups. **a–c** The volcano plots of DELs in the TCGA_NSCLC group, TCGA_LUAD group, and TCGA_LUSC group with thresholds of |log2FC| > 2, average expression > 2, and adjust *P*-value < 0.05, respectively. The red dots and blue dots represent the up-regulated and down-regulated DELs, separately. **d** The intersection of DELs in three groups. **e** The top 10 up and down DELs identified in three groups. FC, fold change

study. We displayed the clinical information of each group in Supplementary Table 1.

Briefly, the age of our cohort ranges from 33 to 88. There were 746 smoking patines among 970 NSCLC patients. And there were 23 EGFR mutations in the 98

patients and 24 KRAS mutations in the 68 patients. Additional details are available in Supplementary Table 1. We first performed univariate and multivariate CPHR to figure out the association between clinical parameters and OS. The results indicated that age and AJCC stage

are independent survival factors as depicted in Supplementary Table 2. Next, univariable CPHR and Kaplan-Meier analysis were performed to assess the survival roles of common DELs. There a total of 14 lncRNAs that are not only related to OS but also show a statistical difference in univariable CPHR analysis with a *P*-value < 0.05. Combing with age and AJCC stage, a Lasson CPHR analysis was further applied to assess the best fitting variables in our cohort. The results suggested that only 12 variables that are best suitable for our further analysis as displayed in Fig. 2a-b. Furthermore, we assess the survival value of 12 variables through multivariable CPHR. Only 7 variables show statistic difference with a *P*-value < 0.05 as shown in Fig. 2c. We depicted the expression value and prognostic roles of 5 lncRNAs in Fig. 3a-e. There are 3 lncRNAs (LINC01833, FAM83A-AS1, and HOTAIR) that are not only high expression in the NSCLC group but also show worse survival. And there are only BANCR and AC112206.2 that are low expression with better survival. Ultimately, we identify five lncRNAs combined with age and AJCC stage which could predict the survival outcome of NSCLC patients.

Developing a risk score formula and prognostic nomogram

Using multivariable CPHR analysis, we divided NSCLC patients into high-risk and low-risk groups based on the value of the median risk score. And we developed a risk score formula to understand the relationship between overall survival and lncRNAs with clinical variables. The formula was as shown as follows: Risk Score = (0.152 × age) + (0.370 × AJCC) + (0.127 × Expression LINC01833) + (0.250 × Expression FAM83A-AS1) + (0.127 × Expression HOTAIR) - (0.124 × Expression BANCR) - (0.283 × Expression AC112206.2). We depicted the distributions of risk score and the status of overall survival in the NSCLC group (Fig. 4a). Also, we validate the distributions of risk score and the status of overall survival in the LUAD group (Supplementary Figure 2A) and LUSC group (Supplementary Figure 2B). Furthermore, we found that the high-risk group of NSCLC patients had a worse prognostic outcome than patients with lower risk scores (Fig. 4b). And this tendency also appeared in the LUAD group (Supplementary Figure 2C) and LUSC group (Supplementary Figure 2D). We performed the time-related ROC curves to compare the sensitivities and specificities of predictive formula in NSCLC. The result suggested that the AUC value of 3 and 5 years was 0.703, 0.667, respectively (Fig. 4c). But the AUC value of clinical variables was only 0.632 and 0.618, which indicate that the prognostic prediction of our nomogram was better than the traditional age and AJCC stage (Fig. 4d). Also, we assessed the predictive ability of formula in LUAD and LUSC group. In the
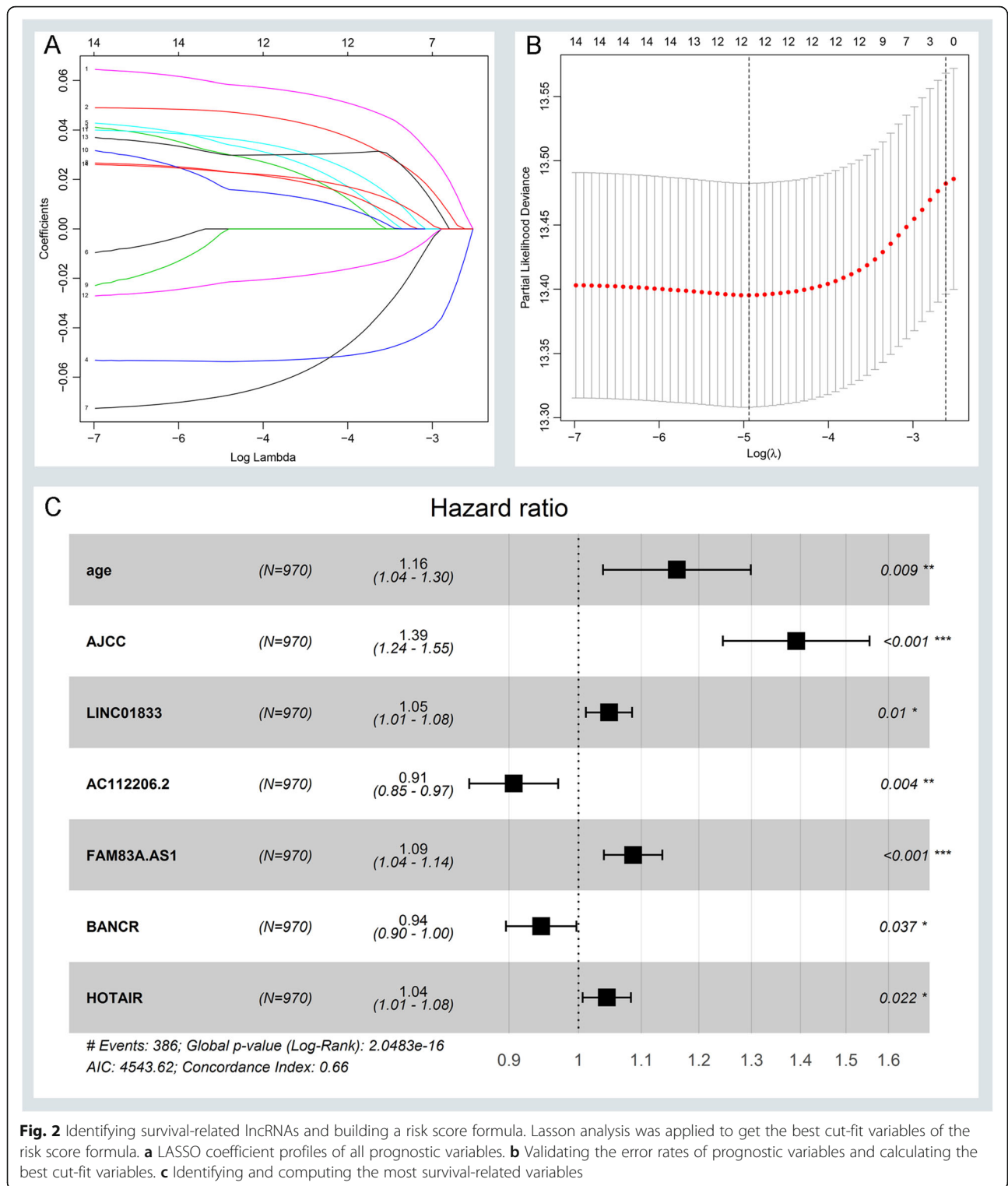
LUAD group, the AUC value of 3 and 5 years was 0.749, 0.735, respectively (Supplementary Figure 2E). The AUC value of 3 and 5 years was 0.65, 0.619 in the LUSC group, respectively (Supplementary Figure 2F). To vividly displayed the prognostic performance of OS-related variables, a nomogram was established. As shown in Fig. 5a, the nomogram could usefully predict the prognosis of 3 years and 5 years in NSCLC patients (Fig. 5a). Additionally, calibration curves suggested that the nomogram had a superior agreement between the predicted and actual OS of 3-year and 5-year in the NSCLC group (Fig. 5b-c) as well as in the LUAD group (Supplementary Figure 3A-B) and LUSC group (Supplementary Figure 3C-D). Moreover, we evaluated the relationship between our risk models and clinical variables. Our result indicated that the high-risk score of our risk model was related to age, male, smoking status, AJCC stage, and EGFR mutation with a statistical difference (*P* < 0.05). But there was no statistically significant difference in KRAS mutation (Supplementary Figure 6).

Functional enrichment analysis

We performed the GO term and KEGG pathway analysis to detect the functional roles of 5 lncRNAs. We first calculated the correlation coefficient of 5 lncRNAs and selected DEMs with a correlation coefficient greater than 0.2 for enrichment analysis. There are 275 qualified DEMs in the correlation analysis of five lncRNAs. There are a total of 183 GO terms and 11 KEGG pathways that are primarily enriched in those DEMs including 119 terms in biological processes (BP), 37 terms in cellular components (CC), 27 terms in molecular functions (MF). We demonstrated the top ten BP, CC, MF, and KEGG Pathway in Supplementary Figure 4A–D. Briefly, the most enriched terms of GO were DNA metabolic process in BP, spindle in CC, and protein kinase binding in MF (Supplementary Figure 4A–C). Additionally, several cancer-related pathways were detected in the KEGG pathway analysis, for example, the p53 signaling pathway, MicroRNAs in cancer pathway, and Cell cycle pathway (Supplementary Figure 4D). Overall, the results from functional enrichment analysis were tightly linked with NSCLC.

Construction of a ceRNA network

Based on the miRNA prediction from 5 lncRNAs, we identified 294 downstream miRNAs by the LncBase database (Supplementary Table 3). And we evaluated the expression of 294 miRNAs in the NSCLC group, LUAD group, and LUSC group. There are a total of 20 common miRNAs in four groups (Supplementary Figure 5D). Next, miRDB and miRTarBase databases were employed for screening miRNA-linked mRNAs.

Fig. 2 Identifying survival-related lncRNAs and building a risk score formula. Lasson analysis was applied to get the best cut-fit variables of the risk score formula. **a** LASSO coefficient profiles of all prognostic variables. **b** Validating the error rates of prognostic variables and calculating the best cut-fit variables. **c** Identifying and computing the most survival-related variables

And we validated their expression in the DEMs of three groups. There are a total of 91 common mRNAs in four groups. According to prediction in Supplementary Table 4, we have found out 22 pairs of lncRNA-miRNA interactions, 145 pairs of miRNA-

mRNA interactions. Finally, we constructed a ceRNA network in Fig. 6 to vividly display the interactions of 5 lncRNAs with 20 miRNAs and 91 mRNAs. Overall, this evidence has revealed that those lncRNAs not only effectively predict the survival outcome of NSCL
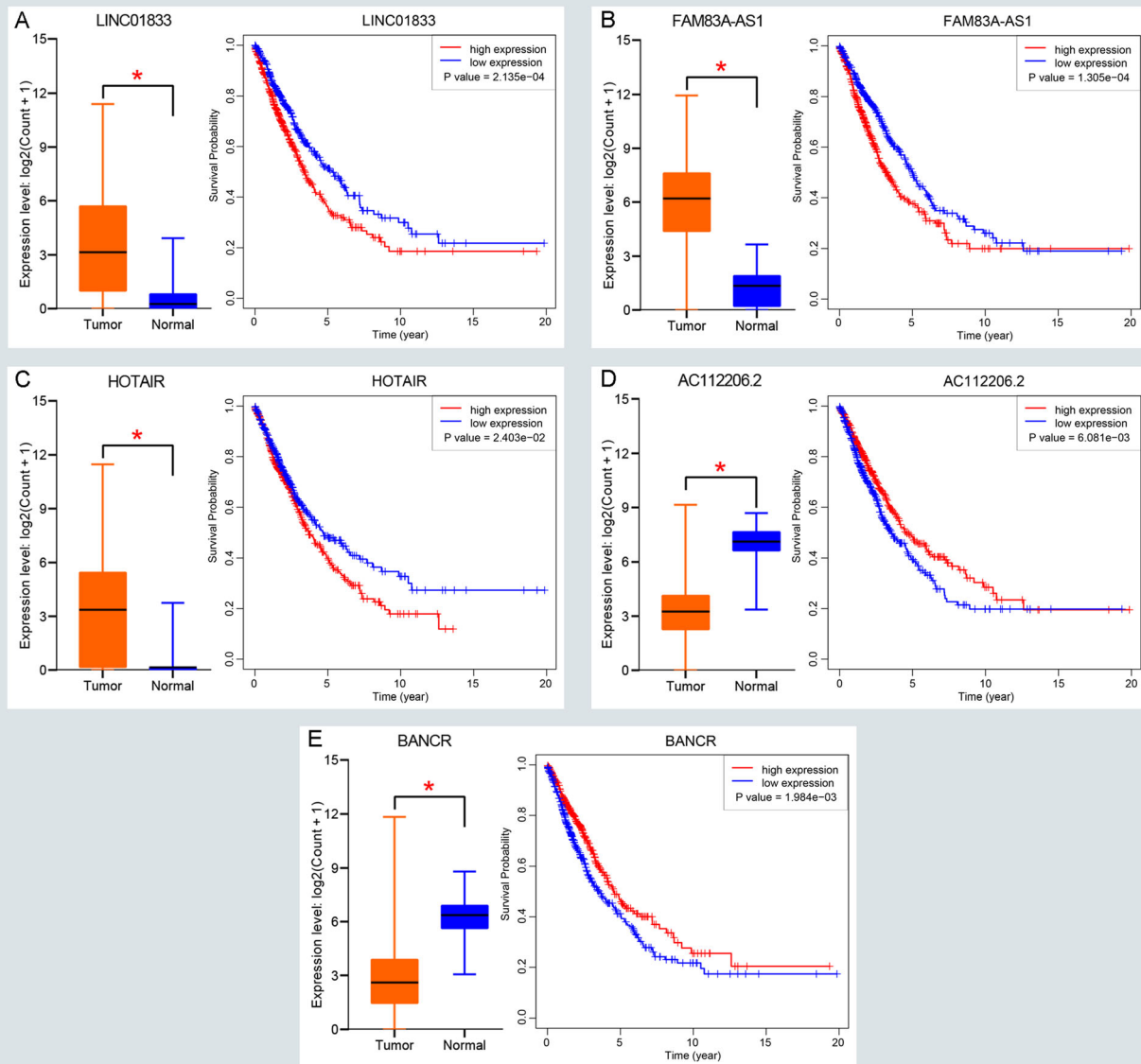
Fig. 3 Screening and validating the expression roles and prognosis values of survival-related lncRNAs in NSCLC. **a-e** Validating expression roles and prognosis values of LINC01833, FAM83A-AS1, HOTAIR, AC112206.2, and BANCR in the NSCLC database, respectively. (*$P < 0.05$)

C patients but also take part in the progression of NSCLC and potentially serve as a therapeutic target.

Discussion

It is well acknowledged that the AJCC stage has been extensively used to estimate the survival outcome of tumor patients [29]. However, some limitations of the AJCC stage can be found in our clinical practice. For example, patients with similar anatomic sites and AJCC staging can exhibit variable responses to treatment and different survival outcomes. This difference may result from tumor heterogeneity, which partly arises from genetic mutations [30, 31]. Furthermore, recent studies have

indicated that age and gender are also effective predictors for OS [32, 33]. So, we attempted to establish a new staging system that combines clinical variables with genetic mutations. Also, some evidence has indicated that lncRNAs not only act a regulatory role in the progression of NSCLC, but also have great potentials as biomarkers because they are stable, highly tissue-specific, and easy to detect in body fluids. For example, serum exosomal MALAT-1 was identified as a diagnostic predictor for NSCLC patients when they are in early-phase or metastasis [13]. High expression of HOTAIR was closely related to progressive disease, worse survival, and more potential in tumor recurrence after radical
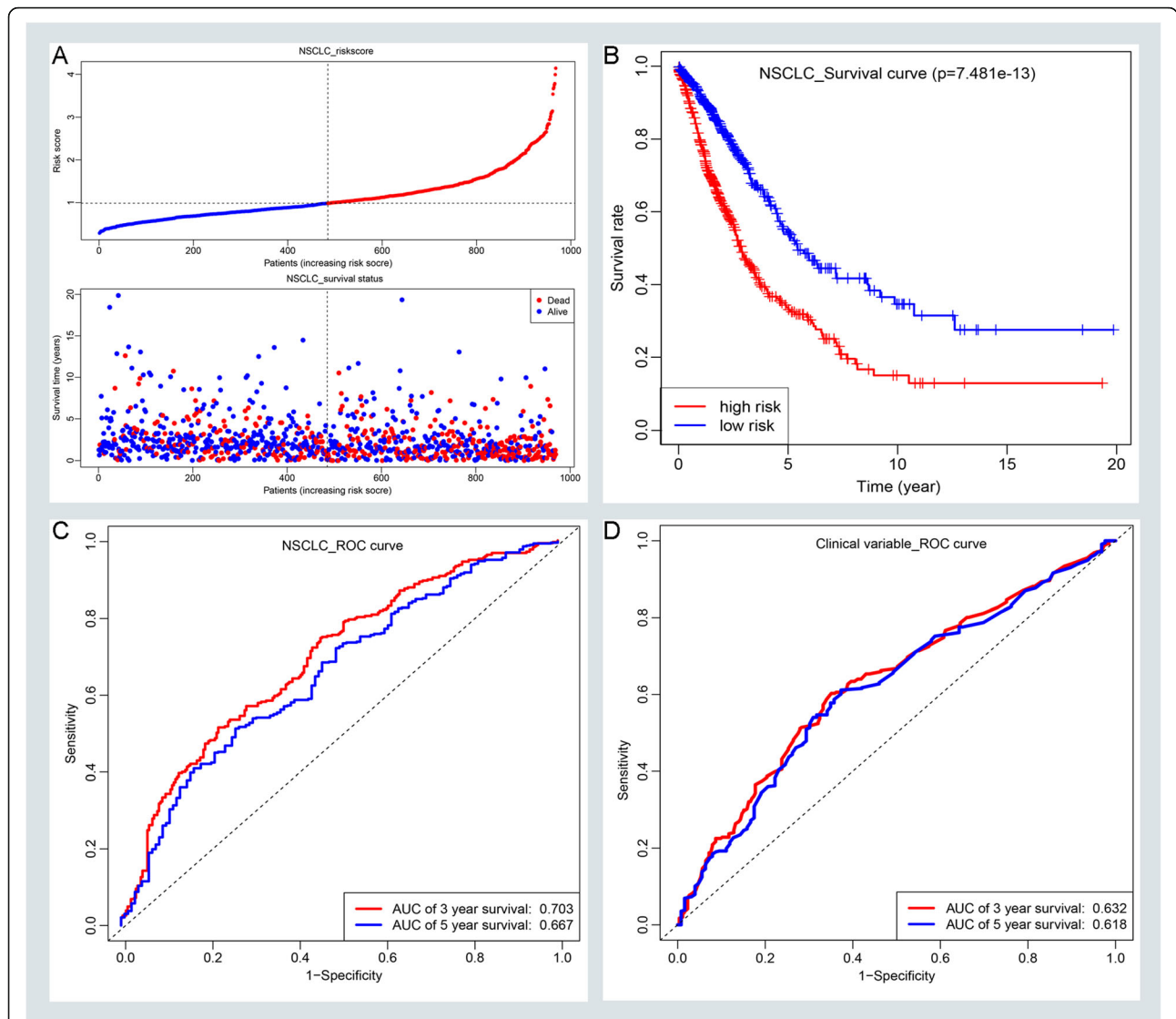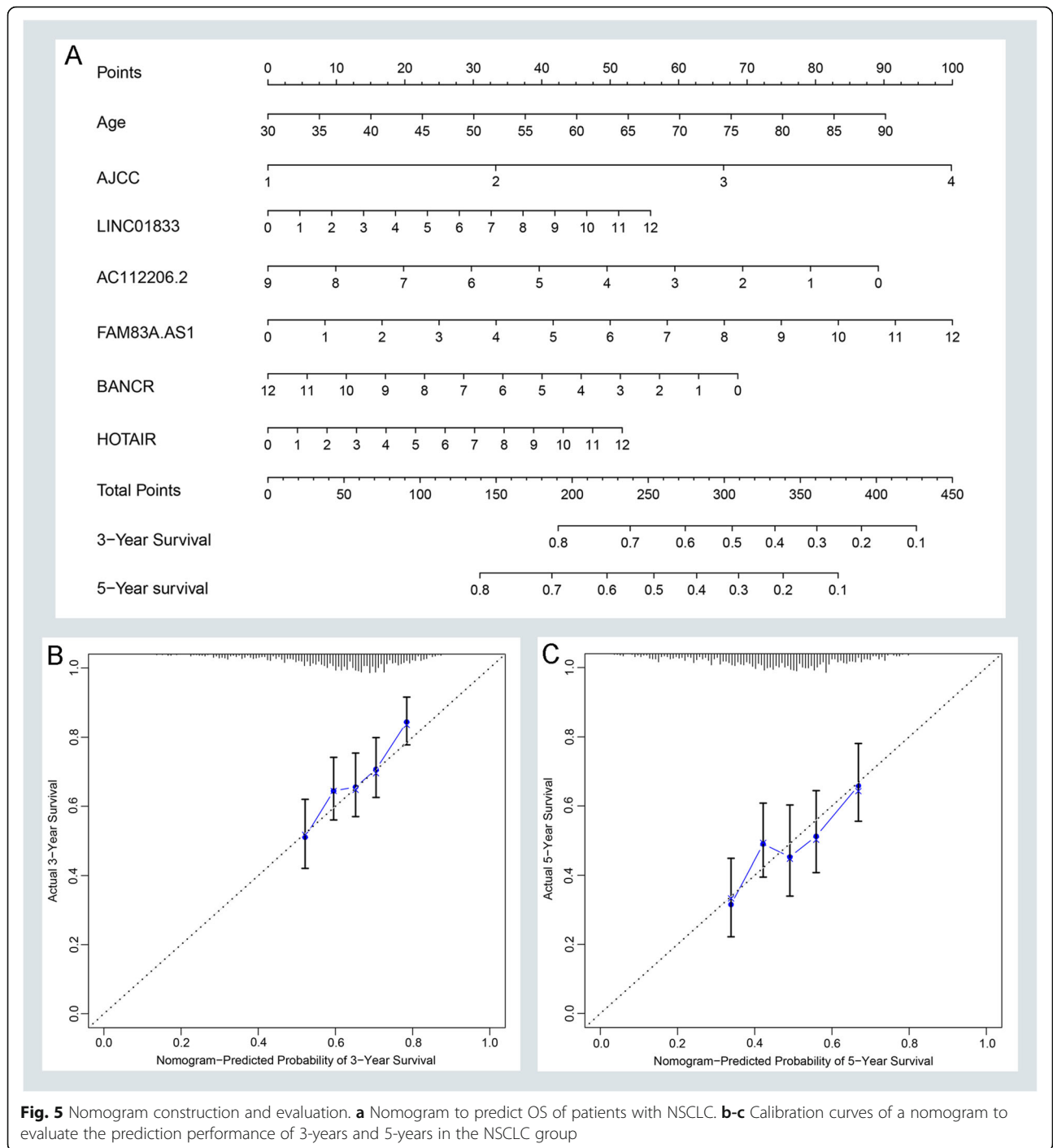
Fig. 4 Assessing the prognostic performance of the risk score formula in the NSCLC group. **a** The risk score distribution and OS status of the formula in the NSCLC group. **b** Kaplan-Meier curves for OS based on the formula in the NSCLC group. The tick-marks on the curve represent the censored patients. **c** ROC curve analysis of the formula for predicting OS in the NSCLC group. **d** ROC curve analysis of the age and AJCC staging for predicting OS in the NSCLC group

operation [34]. Low expression of GAS6-AS1 was connected with the occurrence of lymph node metastasis and an independent biomarker for the prognostic outcome of NSCLC [35]. Therefore, it is urgent to set up a reliable prognostic model for NSCLC.

In this research, we successfully identified a survival nomogram based on lncRNAs and clinical variables of NSCLC. We combined age and AJCC staging with 5 lncRNAs into a risk formula and weighted each parameter to detect their relationship with overall survival. And a nomogram was established to vividly quantify the OS probability of each variable. Notably, we weighted each lncRNA into a nomogram instead of integrating 5 lncRNAs as a whole [36, 37]. Because it is difficult to test

all variables at a time in clinical practice. In our nomogram, we weighted a risk point to every variable and calculated patient survival at 3 or 5 years based on the total risk point. And the predictive performance of our nomogram was superior to the traditional age and AJCC stage. Moreover, our nomogram is easy to understand. Its simplicity will allow clinicians to quickly evaluate survival outcomes and make decisions about individual NSCLC patients. Even individuals without a medical background can easily understand the meaning of our nomogram. Those features will make our nomogram an accurate and effective biomarker for clinical applications.

Additionally, the functional analysis indicated that 5 lncRNAs were also involved in several cancer pathways,

Fig. 5 Nomogram construction and evaluation. **a** Nomogram to predict OS of patients with NSCLC. **b-c** Calibration curves of a nomogram to evaluate the prediction performance of 3-years and 5-years in the NSCLC group

for example, p53 and MicroRNAs in cancer pathways. Moreover, some studies have found their role in the progression of NSCLC. The high expression of HOTAIR is involved in many types of biological processes in NSCLC. For instance, HOTAIR contributes to the down-regulating expression of p21WAF1/CIP1, thereby inducing the cisplatin resistance of A549 cells [38]. Wang et al. demonstrated that HOTAIR was reported to

facilitate the proliferation and migration of A549 and H838 cells through sponging with miR-326, thus control the expression of phox2a [39]. The high expression of HOTAIR can be controlled by Col-1, thereby promoting the formation of microenvironment and progression of NSCLC [40]. Several studies have recently revealed that the lower expression of BANCR was related to the initiation and progression of NSCLC. Sun et al. found that

Fig. 6 Schematic representations of 5 lncRNA-related ceRNA regulatory network in the progression of NSCLC
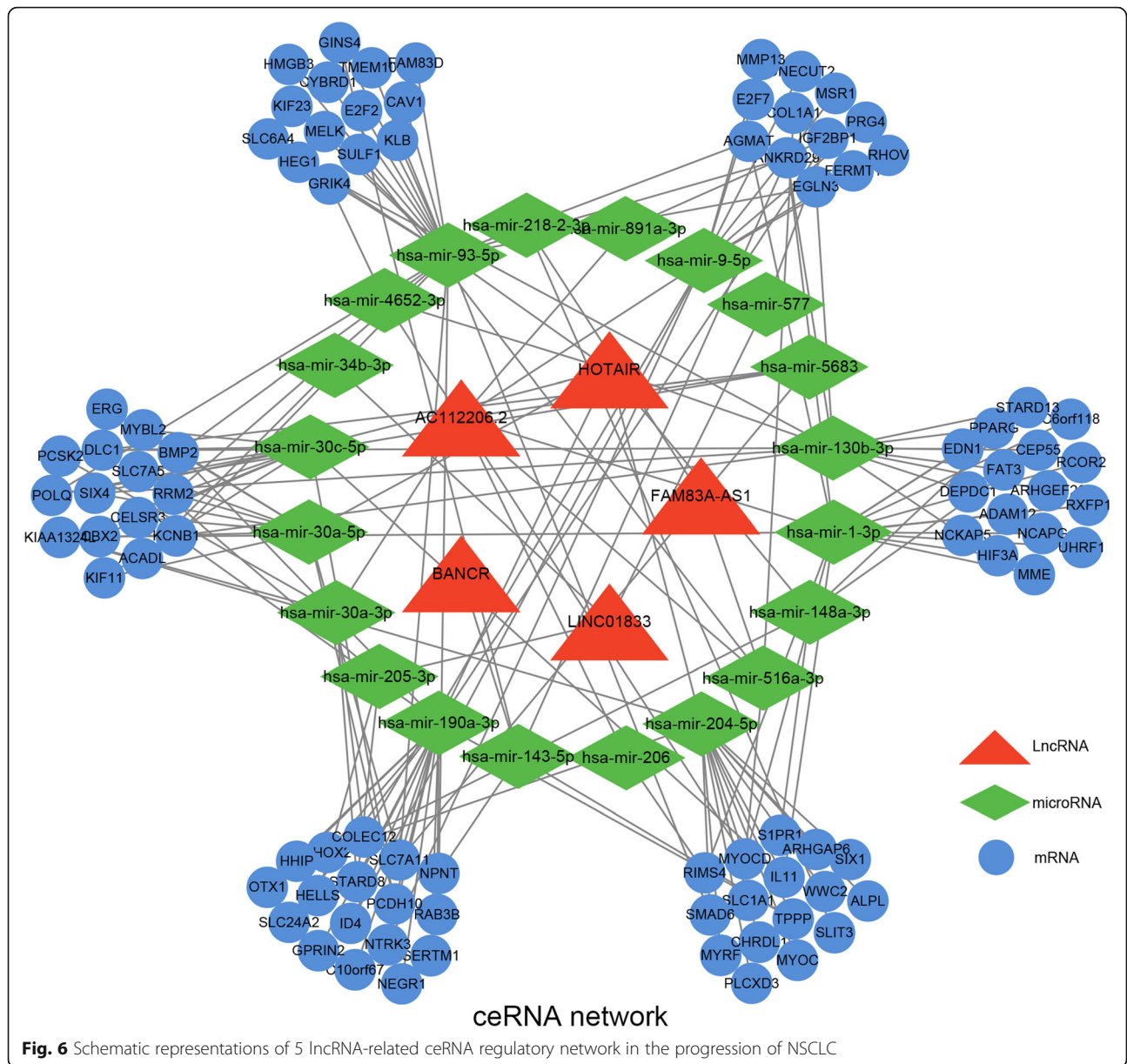
lower expression of BANCR can encourage the epithelial-mesenchymal transition of A549 and SPC-A1 cells and improve their ability in invasion and metastasis [41]. Up-regulating BANCR was tightly linked with radiotherapy for lung cancer [42]. Jiang et al. discovered that BANCR was able to moderate the ability of invasion and metastasis in lung cancer through the p38 MAPK and JNK pathway rather than the ERK MAPK pathway [43]. LINC01833 was demonstrated to promote the infiltration and metastasis of LUAD by adsorbing miR-519e-3p through a sponge and regulate S100A4 expression [44]. FAM83A-AS1 was proved to enhance cell migration, invasion and EMT by modulating the miR-150-5p/MMP14 pathway [45]. So, the lncRNAs in the

nomogram can not only serve as a prognostic biomarker but also function as a regulator in the occurrence and progression of tumors. Notably, those lncRNAs can act as a ceRNA network, thereby participating in cancer progression. For instance, HOTAIR/miR-149-5p/HNRNPA1 axis promotes the cell growth, migration, and invasion in NSCLC [46]; FAM83A-AS1/miR-150-5p/MMP14 regulates LUAD progression and invasion [45]; BANCR/miR-338-3p/IGF1R network regulated Raf/MEK/ERK pathway, thereby encouraging the proliferation, migration, invasion and epithelial-mesenchymal transition (EMT) of esophageal cancer [47]. Those studies indicated that the lncRNAs in the nomogram can function as a therapeutic target for NSCLC.

Although our lncRNAs-related nomogram showed good performance in survival prediction of NSCLC, some limitations can be detected in our research. First, our study is a retrospective study. But 970 patients is a large sample. So our results are reliable. Additionally, our data lacked information such as chemo-radiotherapy history, smoking history, and patients' disease history. This may result from the limitations of our data. So, further clinical studies are needed to verify our results when applied to clinical practice. Last, we identified survival-related lncRNAs to construct the nomogram, which might overlook some valuable information. All in all, despite these limitations in our study, we believe that our persistent efforts will eventually establish an ideal prognostic model in clinical practice.

Conclusions

In summary, by a large sample analysis, we successfully constructed a nomogram based on lncRNAs and clinical variables that predicts the survival of NSCLC patients. And the predictive performance of our prognostic nomogram was better than the traditional AJCC stage and age. In addition to the survival prediction of our nomogram, functional analysis and ceRNA network also indicate that they might involve in cancer progression and potentially serve as a therapeutic target for NSCLC.

Abbreviations
NSCLC: Non-small cell lung cancer; LUAD: Lung adenocarcinoma; LUSC: Lung squamous carcinoma; lncRNAs: Long non-coding RNAs; CPHR: Cox proportional hazards regression; TCGA: The Cancer Genome Atlas; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Gene and Genomes; AJCC: American Joint Committee on Cancer; OS: Overall survival; HR: Hazard ratio; CI: Confidence interval; ROC: Receiver operating curves; AUC: Area Under Curve; ceRNA: Competing endogenous RNA

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12885-021-08207-7.

Additional file 1 : **Supplementary Figure 1.** Screening differentially expressed mRNAs (DEMs) in three groups. (A–C) The volcano plots of DEMs in the TCGA_NSCLC group, TCGA_LUAD group, and TCGA_LUSC group with thresholds of |log2FC| > 2, average expression > 2, and adjust *P*-value < 0.05, respectively. The red dots and blue dots represent the up-regulated and down-regulated DEMs, separately. (D) The intersection of DELs in three groups.

Additional file 2 : **Supplementary Figure 2.** Assessing the prognostic performance of the risk score formula in LUAD and LUSC group. (A-B) The risk score distribution and OS status of the formula in LUAD and LUSC group, respectively. (C-D) Kaplan-Meier curves for OS based on the formula in LUAD and LUSC group, separately. The tick-marks on the curve represent the censored patients. (E-F) ROC curve analysis of the formula for predicting OS in LUAD and LUSC group, respectively.

Additional file 3 : **Supplementary Figure 3.** Evaluating the prediction performance of the nomogram. (A-B) Calibration curves of a nomogram to evaluate the prediction performance of 3-years and 5-years in the LUAD group. (C-D) Calibration curves of a nomogram to evaluate the prediction performance of 3-years and 5-years in the LUSC group.

Additional file 4 : **Supplementary Figure 4.** Functional enrichment analysis for the differentially expressed mRNAs (DEMs) of 5 lncRNAs. (A-C) The top ten enriched GO terms of qualified DEMs with a correlation coefficient greater than 0.2 in biological processes (BP), cellular components (CC), and molecular functions (MF), respectively. (D) The top ten enriched KEGG pathways of qualified DEMs in the NSCLC database.

Additional file 5 : **Supplementary Figure 5.** Screening differentially expressed miRNA (DEMIs) in three groups. (A–C) The volcano plots of DEMIs in the TCGA_NSCLC group, TCGA_LUAD group, and TCGA_LUSC group with thresholds of |log2FC| > 1, average expression > 1, and adjust *P*-value < 0.05, respectively. The red dots and blue dots represent the up-regulated and down-regulated DEMIs, separately. (D) The intersection of DEMIs in three groups and the LncBase database.

Additional file 6 : **Supplementary Figure 6.** The relationship between risk models and clinical variables. (A-E) High-risk score of our risk model was related to age (*P* < 0.001), male (*P* < 0.001), smoking status (*P* = 0.0165), AJCC stage (*P* < 0.001), and EGFR mutation (*P* = 0.0063) with a statistical difference. (F) But there was no statistically significant difference in KRAS mutation (*P* = 0.3895).

Additional file 7 : **Supplementary Table 1**. Clinical characteristics of NSCLC patients in this study.

Additional file 8 : **Supplementary Table 2.** Univariate and multivariate Cox regression analysis in NSCLC.

Additional file 9 : **Supplementary Table 3.** The lncRNA-miRNA pairs were predicted by the LncBase database.

Additional file 10 : **Supplementary Table 4.** The qualified lncRNA-miRNA and miRNA-mRNA pairs were constructed in the ceRNA network.

Additional file 11 : **Supplementary Table 5.** Differentially expressed RANs in our analysis.

Authors' contributions
YY, RKM designed this project, performed data analysis and wrote the manuscript. All authors have read and approved the final manuscript.

Availability of data and materials
The datasets generated during the current study are available in The Cancer Genome Atlas (TCGA) program (https://portal.gdc.cancer.gov/repository).

Declarations

Ethics approval and consent to participate
We declare that all the clinical data used in this article were obtained from The Cancer Genome Atlas Program (TCGA) database, and all the data recorded by TCGA have already met strict scientific, technical, and ethical requirements: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/policies.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Shengjing Hospital of China Medical University, Shenyang 110004, Liaoning, China. [2]Department of Thoracic Surgery, Shengjing Hospital of China Medical University, Shenyang 110004, Liaoning, China.

References

1. Siegel R, Miller K, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019; 69(1):7–34. https://doi.org/10.3322/caac.21551.
2. New M, Keith R. Early detection and chemoprevention of lung cancer. F1000Research. 2018;7:61.
3. Tang Y, Qiao G, Xu E, Xuan Y, Liao M, Yin G. Biomarkers for early diagnosis, prognosis, prediction, and recurrence monitoring of non-small cell lung cancer. Onco Targets Ther. 2017;10:4527–34. https://doi.org/10.2147/OTT.S142149.
4. Morris K, Mattick J. The rise of regulatory RNA. Nat Rev Genet. 2014;15(6): 423–37. https://doi.org/10.1038/nrg3722.
5. Engreitz J, Haines J, Perez E, Munson G, Chen J, Kane M, et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature. 2016;539(7629):452–5. https://doi.org/10.1038/nature20149.
6. Bhan A, Soleimani M, Mandal S. Long noncoding RNA and cancer: a new paradigm. Cancer Res. 2017;77(15):3965–81. https://doi.org/10.1158/0008-5472.CAN-16-2634.
7. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell. 2011;146(3):353–8. https://doi.org/10.1016/j.cell.2011.07.014.
8. Xiong G, Liu C, Yang G, Feng M, Xu J, Zhao F, et al. Long noncoding RNA GSTM3TV2 upregulates LAT2 and OLR1 by competitively sponging let-7 to promote gemcitabine resistance in pancreatic cancer. J Hematol Oncol. 2019;12(1):97. https://doi.org/10.1186/s13045-019-0777-7.
9. Zu F, Han H, Sheng W, Sun J, Zang H, Liang Y, et al. Identification of a competing endogenous RNA axis related to gastric cancer. Aging (Albany NY). 2020;12:20540–60.
10. Chen X, Wang Z, Tong F, Dong X, Wu G, Zhang R. lncRNA UCA1 promotes gefitinib resistance as a ceRNA to target FOSL2 by sponging miR-143 in non-small cell lung cancer. Mol Ther Nucleic Acids. 2020;19:643–53. https://doi.org/10.1016/j.omtn.2019.10.047.
11. Li L, Zhang L, Zhang Y, Zhou F. Increased expression of LncRNA BANCR is associated with clinical progression and poor prognosis in gastric cancer. Biomed Pharmacother. 2015;72:109–12.
12. Shen X, Bai Y, Luo B, Zhou X. Upregulation of lncRNA BANCR associated with the lymph node metastasis and poor prognosis in colorectal cancer. Biol Res. 2017;50(1):32. https://doi.org/10.1186/s40659-017-0136-5.
13. Zhang R, Xia Y, Wang Z, Zheng J, Chen Y, Li X, et al. Serum long non coding RNA MALAT-1 protected by exosomes is up-regulated and promotes cell proliferation and migration in non-small cell lung cancer. Biochem Biophys Res Commun. 2017;490(2):406–14. https://doi.org/10.1016/j.bbrc.2017.06.055.
14. Weinstein J, Collisson E, Mills G, Shaw K, Ozenberger B, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10): 1113–20. https://doi.org/10.1038/ng.2764.
15. Yates A, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020;48:D682–8.
16. Ritchie M, Phipson B, Wu D, Hu Y, Law C, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
17. Li R, Liu X, Zhou XJ, Chen X, Li JP, Yin YH, et al. Identification of a prognostic model based on immune-related genes of lung squamous cell carcinoma. Front Oncol. 2020;10:1588. https://doi.org/10.3389/fonc.2020.01588.
18. Zhong G, Lou W, Yao M, Du C, Wei H, Fu P. Identification of novel mRNA-miRNA-lncRNA competing endogenous RNA network associated with prognosis of breast cancer. Epigenomics. 2019;11(13):1501–18. https://doi.org/10.2217/epi-2019-0209.
19. Huang DW, Sherman B, Lempicki R. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1): 44–57. https://doi.org/10.1038/nprot.2008.211.
20. Kuleshov M, Jones M, Rouillard A, Fernandez N, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–7. https://doi.org/10.1093/nar/gkw377.
21. Karagkouni D, Paraskevopoulou MD, Tastsoglou S, Skoufos G, Karavangeli A, Pierros V, et al. DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts. Nucleic Acids Res. 2020;48(D1): D101–10. https://doi.org/10.1093/nar/gkz1036.
22. Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. 2018;46(D1):D296–302. https://doi.org/10.1093/nar/gkx1067.
23. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. 2020;48(D1):D127–31. https://doi.org/10.1093/nar/gkz757.
24. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504. https://doi.org/10.1101/gr.1239303.
25. Ravani P, Parfrey P, MacRae J, James M, Quinn R, Malberti F, et al. Modeling survival of arteriovenous accesses for hemodialysis: semiparametric versus parametric methods. Clin J Am Soc Nephrol. 2010;5(7):1243–8. https://doi.org/10.2215/CJN.06190809.
26. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw. 2011;39(5): 1–13. https://doi.org/10.18637/jss.v039.i05.
27. Núñez E, Steyerberg EW, Núñez J. Regression modeling strategies. Rev Esp Cardiol. 2011;64(6):501–7. https://doi.org/10.1016/j.recesp.2011.01.019.
28. Li J, Ma S. Time-dependent ROC analysis under diverse censoring patterns. Stat Med. 2011;30(11):1266–77. https://doi.org/10.1002/sim.4178.
29. Rami-Porta R, Asamura H, Travis W, Rusch V. Lung cancer - major changes in the American Joint Committee on cancer eighth edition cancer staging manual. CA Cancer J Clin. 2017;67(2):138–55.
30. Jain D, Roy-Chowdhuri S. Molecular pathology of lung cancer cytology specimens: a concise review. Arch Pathol Lab Med. 2018;142(9):1127–33. https://doi.org/10.5858/arpa.2017-0444-RA.
31. Rooney C, Sethi T. Advances in molecular biology of lung disease: aiming for precision therapy in non-small cell lung cancer. Chest. 2015;148(4):1063–72. https://doi.org/10.1378/chest.14-2663.
32. Radkiewicz C, Dickman P, Johansson A, Wagenius G, Edgren G, Lambe M. Sex and survival in non-small cell lung cancer: a nationwide cohort study. PLoS One. 2019;14(6):e0219206. https://doi.org/10.1371/journal.pone.0219206.
33. Ahmed T, Lycan T, Dothard A, Ehrlichman P, Ruiz J, Farris M, et al. Performance status and age as predictors of immunotherapy outcomes in advanced non-small-cell lung cancer. Clin Lung Cancer. 2020;21(4):e286–93. https://doi.org/10.1016/j.cllc.2020.01.001.
34. Liu X, Liu Z, Sun M, Liu J, Wang Z, De W. The long non-coding RNA HOTAIR indicates a poor prognosis and promotes metastasis in non-small cell lung cancer. BMC Cancer. 2013;13(1):464. https://doi.org/10.1186/1471-2407-13-464.
35. Han L, Kong R, Yin D, Zhang E, Xu T, De W, et al. Low expression of long noncoding RNA GAS6-AS1 predicts a poor prognosis in patients with NSCLC. Med Oncol (Northwood, London, England). 2013;30(4):694.
36. Wang Y, Du L, Yang X, Li J, Li P, Zhao Y, et al. A nomogram combining long non-coding RNA expression profiles and clinical factors predicts survival in patients with bladder cancer. Aging (Albany NY). 2020;12(3):2857–79. https://doi.org/10.18632/aging.102782.
37. Liu Z, Sun M, Lu K, Liu J, Zhang M, Wu W, et al. The long noncoding RNA HOTAIR contributes to cisplatin resistance of human lung adenocarcinoma cells via downregualtion of p21(WAF1/CIP1) expression. PLoS One. 2013; 8(10):e77293. https://doi.org/10.1371/journal.pone.0077293.
38. Wang R, Chen X, Xu T, Xia R, Han L, Chen W, et al. MiR-326 regulates cell proliferation and migration in lung cancer by targeting phox2a and is regulated by HOTAIR. Am J Cancer Res. 2016;6(2):173–86.
39. Zhuang Y, Wang X, Nguyen H, Zhuo Y, Cui X, Fewell C, et al. Induction of long intergenic non-coding RNA HOTAIR in lung cancer cells by type I collagen. J Hematol Oncol. 2013;6(1):35. https://doi.org/10.1186/1756-8722-6-35.
40. Lin T, Fu Y, Zhang X, Gu J, Ma X, Miao R, et al. A seven-long noncoding RNA signature predicts overall survival for patients with early stage non-small cell lung cancer. Aging (Albany NY). 2018;10(9):2356–66. https://doi.org/10.18632/aging.101550.
41. Sun M, Liu X, Wang K, Nie F, Kong R, Yang J, et al. Downregulation of BRAF activated non-coding RNA is associated with poor prognosis for non-small cell lung cancer and promotes metastasis by affecting epithelial-mesenchymal transition. Mol Cancer. 2014;13(1):68. https://doi.org/10.1186/1476-4598-13-68.
42. Chen J, Chen M, Zheng Y, Wang S, Shen Z. Up-regulation of BRAF activated non-coding RNA is associated with radiation therapy for lung cancer. Biomed Pharmacother. 2015;71:79–83.
43. Jiang W, Zhang D, Xu B, Wu Z, Liu S, Zhang L, et al. Long non-coding RNA BANCR promotes proliferation and migration of lung carcinoma via MAPK pathways. Biomed Pharmacother. 2015;69:90–5.

44. Zhang Y, Li W, Lin Z, Hu J, Wang J, Ren Y, et al. The long noncoding RNA Linc01833 enhances lung adenocarcinoma progression via MiR-519e-3p/ S100A4 Axis. Cancer Manag Res. 2020;12:11157–67. https://doi.org/10.2147/ CMAR.S279623.

45. Xiao G, Wang P, Zheng X, Liu D, Sun X. FAM83A-AS1 promotes lung adenocarcinoma cell migration and invasion by targeting miR-150-5p and modifying MMP14. Cell Cycle. 2019;18(21):2972–85. https://doi.org/10.1080/1 5384101.2019.1664225.

46. Li H, Cui Z, Lv X, Li J, Gao M, Yang Z, et al. Long non-coding RNA HOTAIR function as a competing endogenous RNA for miR-149-5p to promote the cell growth, migration, and invasion in non-small cell lung cancer. Front Oncol. 2020;10:528520. https://doi.org/10.3389/fonc.2020.528520.

47. Song W, Wang K, Yang X, Dai W, Fan Z. Long non-coding RNA BANCR mediates esophageal squamous cell carcinoma progression by regulating the IGF1R/Raf/MEK/ERK pathway via miR-338-3p. Int J Mol Med. 2020;46(4): 1377–88. https://doi.org/10.3892/ijmm.2020.4687.