Research Paper

A novel molecular-clinicopathologic nomogram to improve prognosis prediction of hepatocellular carcinoma

Zhongjing Zhang[1,*], Wanqing Weng[1,*], Weiguo Huang[1], Boda Wu[1], Yi Zhou[1], Jie Zhang[1], Tuo Deng[1], Wen Ye[1], Jiecheng Zhang[1], Jianyang Ao[1], Qiyu Zhang[1], Keqing Shi[2]

[1]Department of Hepatopancreatobiliary Surgery, The First Affiliated Hospital, Wenzhou Medical University, Wenzhou 325015, Zhejiang Province, PR China
[2]Precision Medical Center Laboratory, The First Affiliated Hospital, Wenzhou Medical University, Wenzhou 325015, Zhejiang Province, PR China
*Co-first author

Correspondence to: Keqing Shi, Qiyu Zhang; **email:** skochilly@163.com, qiyuz@126.com
Keywords: hepatocellular carcinoma, the cancer genome atlas, long non-coding RNA, nomogram, time-dependent receiver operating characteristic
Received: October 4, 2019 **Accepted:** May 20, 2020 **Published:** June 30, 2020

ABSTRACT

Background: Emerging evidence suggests that long non-coding RNA (lncRNA) plays a crucial part in the development and progress of hepatocellular carcinoma (HCC). The objective was to develop novel molecular-clinicopathological prediction methods for overall survival (OS) and recurrence of HCC.
Results: An 8-lncRNA-based classifier for OS and a 14-lncRNA-based classifier for recurrence were developed by LASSO COX regression analysis, both of which had high accuracy. The tdROC of OS-nomogram and recurrence-nomogram indicates the satisfactory accuracy and predictive power. The classifiers and nomograms for predicting OS and recurrence of HCC were validated in the Test and GEO cohorts.
Conclusions: These two lncRNA-based classifiers could be independent prognostic factors for OS and recurrence. The molecule-clinicopathological nomograms based on the classifiers could increase the prognostic value.
Methods: HCC lncRNA expression profiles from the cancer genome atlas (TCGA) were randomly divided into 1:1 training and test cohorts. Based on least absolute shrinkage and selection operator method (LASSO) COX regression model, lncRNA-based classifiers were established to predict OS and recurrence, respectively. OS-nomogram and recurrence-nomogram were developed by combining lncRNA-based classifiers and clinicopathological characterization to predict OS and recurrence, respectively. The prognostic value was accessed by the time-dependent receiver operating characteristic (tdROC) and the concordance index (C-index).

INTRODUCTION

Hepatocellular carcinoma (HCC) is one of the leading causes of cancer-related mortality worldwide [1]. Considerable progress has been achieved in the prevention, monitoring, early screening, diagnosis and treatment of HCC over the past few decades. However, in many countries, the incidence and specific mortality of HCC continue to rise [2]. There are a number of reasons for the high mortality rate of HCC; most importantly, in many parts of the world, patients are diagnosed at an advanced stage [1]. Thence, it is of great clinical implication to identify effective tumor markers and explore their role in the occurrence and development of HCC.

Next-generation sequencing (NGS) is a powerful platform for high-throughput sequencing of different

genetic factors, which helps researchers to obtain more accurate and comprehensive data of gene variation [3]. The rich and standardized clinical data and abundant samples for different types of cancer generated by the Cancer Genome Atlas (TCGA) enabled a joint analysis of multiple influencing factors associated with tumor oncogenesis [4, 5].

Long non-coding RNAs (lncRNAs) which contain more than 200 nucleotides is a type of the non-coding RNAs (ncRNAs) [6]. For a long time, lncRNA is considered as a kind of non-functional RNA, but emerging research has proved that these RNAs were important regulators of gene expression networks [7, 8]. Their functions contain controlling nuclear architecture and mRNA stability, participating in the transcription, translation and post-translational modifications, which involve all aspects of cellular gene expression [9, 10]. In recent researches, many lncRNAs have seemed as biomarkers of early detection and prognosis of HCC, but these studies only involved minority lncRNA and lack a large number of clinical samples for analysis [11, 12].

In current study, we collected a large cohort of HCC patients who contained clinical information and complete sequencing results in the TCGA database. Thereafter, we performed least absolute shrinkage and selection operator method (LASSO) COX select model, a method that could be applied to high dimensional regression prediction, to establish and validate two multi-lncRNA-based classifiers which have high veracity of predicting overall survival (OS) and recurrence in HCC patients.

RESULTS

Data processing

The workflow of this article is shown in Figure 1. In the expression profiles of HCC tumors compared with the samples from normal tissues, we identified 669 differentially expressed lncRNAs (DElncNRAs) of $|\log \text{Fold Change}| \geq 2$ and $p < 0.05$(Supplementary Table 1 and Figure 2A). Of which, 595 lncRNAs were down-regulated and 74 lncRNAs were up-regulated. As shown in Figure 2B, significant differential expression was detected between the tumor and the adjacent normal groups. Subsequently, the DElncRNAs with $P < 0.05$ were selected by univariate COX regression analysis. Therefore, a total of 191 OS-related lncRNAs and 86 recurrence-related lncRNAs were reserved for further study (Figure 2C). After taking the intersection with GSE76427 and GSE116174, 21 recurrence-related lncRNAs and 85 OS-related lncRNAs were finally obtained for classifier development. A total of 312 patients with OS data were randomized 1:1 into two groups, the training cohort (n=156) and the test cohort (n=156). The GSE116174 (n=64) was reserved as a validation cohort for predicting OS. Meanwhile, a total of 269 patients with recurrence data were randomly divided equally into two groups, the training cohort (n= 130) and the test cohort (n=139). GSE76427 (n= 81) was used as a validation cohort to validate recurrence-related models. LASSO COX selection method was applied to training cohort to develop a prediction model (OS: Figure 3A, 3B; recurrence: 3F, 3G). As shown in Supplementary Tables 2, 8 OS-related DElncRNAs and 14 recurrence-related DElncRNAs were identified by the LASSO COX selected model.

Multi-lncRNAs-based classifier

In order to contrive multi-lncRNAs-based classifiers for predicting OS and recurrence in HCC, LASSO COX selection method was performed with the 85 OS related lncRNAs and 21 recurrence related lncRNAs expression data. An 8-lncRNA-based classifier for OS (Figure 3A and 3B) and a 14-lncRNAs-based classifier for recurrence were constructed by training cohort (Figure 3E, 3F). All those lncRNAs are listed in Supplementary Table 2. 8-lncRNAs-based classifier = 0.0299*EXP (AC090921.1) + 0.0125*EXP (AC096637.2) + 0.1838* EXP(AP002478.1) + 0.2221*EXP (C10orf91)+ 0.0437* EXP (LINC01116) + 0.0251*EXP (LINC01224) + 0.0137*EXP (MAFG-DT) -0.1168*EXP(SERTAD4-AS1); 14-lncRNAs-based classifier = -0.0255*EXP (AC004477.1)+0.1647*EXP(AC010307.4)+0.0416*EX P(AC034229.4)+0.1580*EXP (AC209154.1)+0.3958* EXP(C10orf91)+0.0233*EXP(CDKN2A-DT)+0.0037* EXP(CDKN2B-AS1)+0.00057*EXP(FIRRE)-0.1140* EXP(LINC01549)+0.1813*EXP (LINC01572)+ 0.0958*EXP(MAFA-AS1)+ 0.1348*EXP(MAFG-DT)- 0.365*EXP (MIR9-3HG)-0.0761*EXP (SNHG25). All patients were divided into low and high risk groups according to the optimal cut-off value calculated by X-TILE. The optimal cutoff value for the OS-related classifier was 0.2, and for the recurrence-related classifier was 0.1. The Kaplan-Meier log rank test illustrated that there were significant differences in OS and recurrence in the training cohort (Supplementary Figure 1A, 1E), the test cohort (Supplementary Figure 1B, 1F), the TCGA cohort (Supplementary Figure 1C, 1G), and the GEO cohort (Supplementary Figure 1D, 1H).

Patient characteristics

Since the training cohort and test cohort were equally randomly grouped, there was no significant difference or deviation between them. (Supplementary Tables 9–11, Table 1).

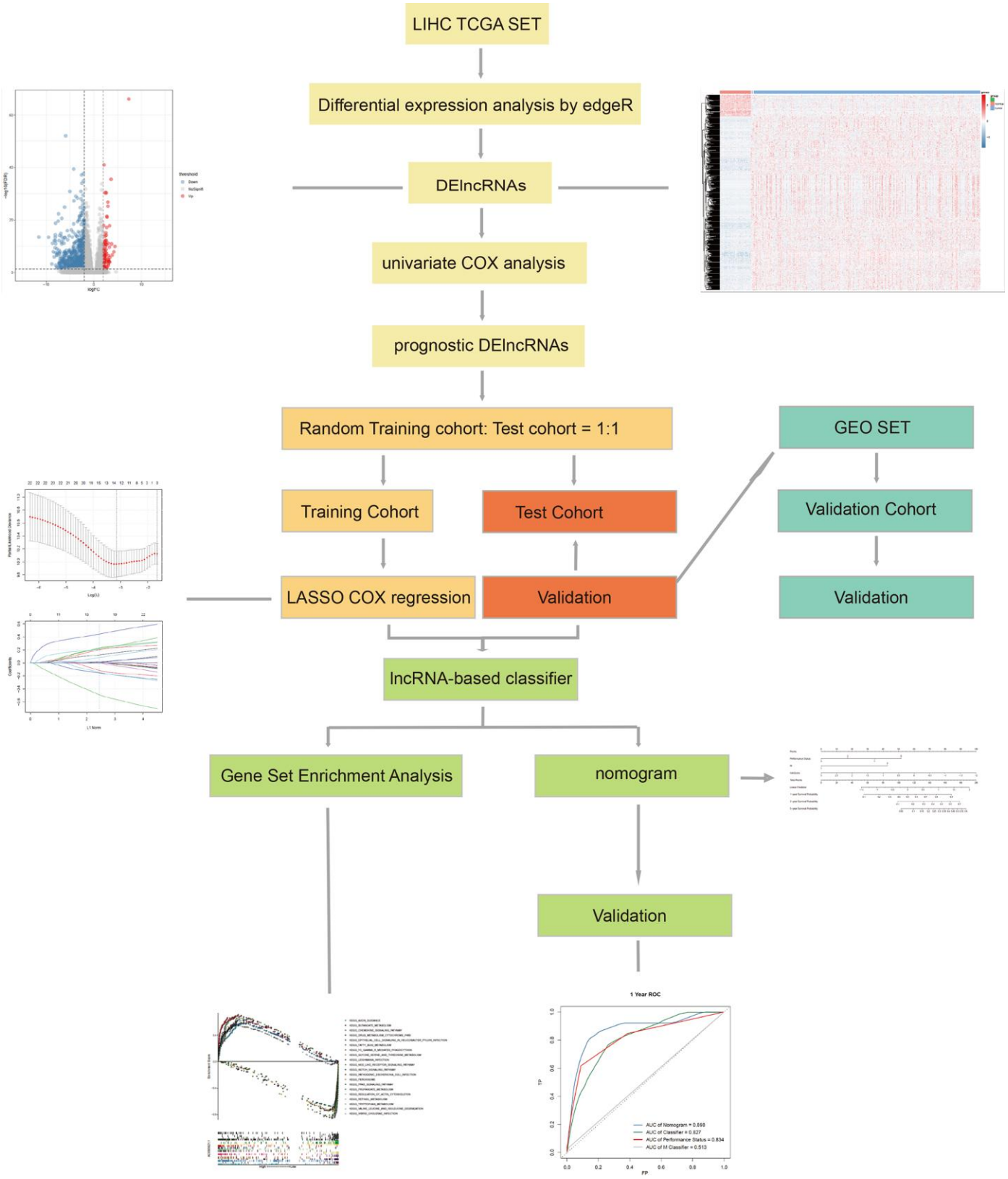


Figure 1. The workflow of this work.

Overall survival

In the training cohort, 156 patients were enrolled. As shown in Supplementary Table 9, there were no significant differences were detected in the distribution of age (P = 0.598), neoplasm histologic grade (P = 0.179), vascular invasion (P = 0.872), performance status (P = 0.155), TNM T stage (P=0.523), TNM M stage (P = 0.298), adjacent hepatic tissue inflammation (P = 0.656), liver fibrosis Ishak score category (P = 0.923), family history (P = 0.922), race category (P = 0.968), HBV infection (P = 0.080), HCV infection (P = 0.139), alcohol consumption (P = 0.287), Child-Pugh classification (P = 0.068), AJCC pathological stage (P = 0.521) and gender (P = 0.849).

In the test cohort, the distribution of data was similar to the training cohort. The proportion of patients with performance status (2 + 3) (P = 0.018), TNM N stage (N1) (P = 0.031) and neoplasm histologic Grade (G3+G4) (P = 0.027) in the high-risk group was higher than the low risk group.

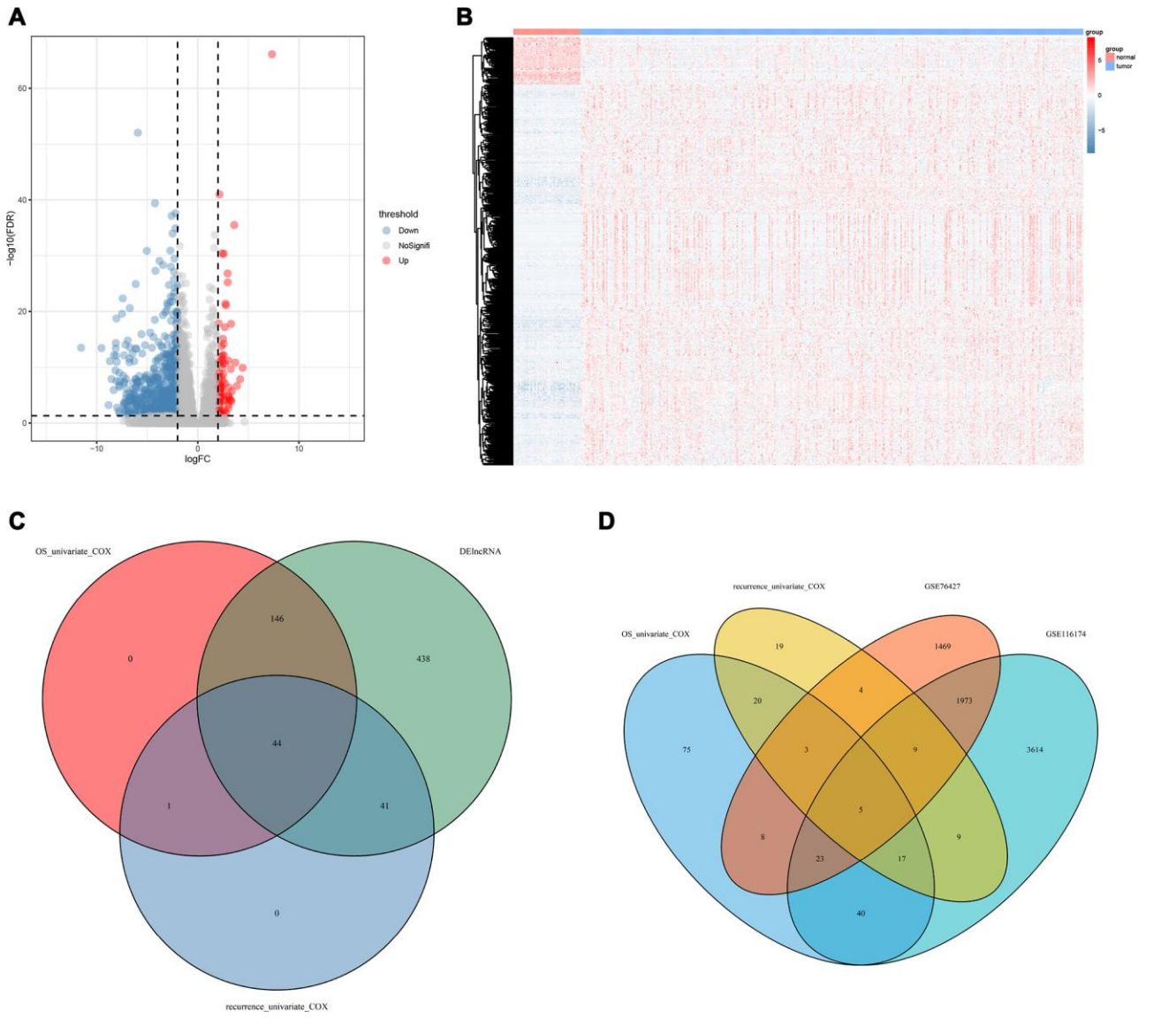


Figure 2. Prognostic DElncRNAs identification process. (**A**) Volcano plot of differentially expressed lncRNAs in TCGA-LICH dataset; (**B**) Hierarchical clustering of HCC with or without tumor using 669 differentially expressed lncRNAs using Euclidean distance and average linkage clustering; (**C**) Venn diagram of prognostic DElncRNAs in prognostic lncRNAs (OS/recurrence multivariate cox p < 0.05) and DElncRNAs(|logFC| >2 and padj < 0.05); (**D**) Venn diagram of lncRNAs related to OS/recurrence. TCGA, The Cancer Genome Atlas; LICH, Liver hepatocellular carcinoma; HCC, hepatocellular carcinoma; DElncRNA, differentially expressed long non-coding RNA; OS, overall survival; LASSO, least absolute shrinkage and selection operator method.

In the TCGA cohort, 312 patients were included for further study. The proportion of patients with performance status (2 + 3) (P = 0.018), tumor grade (G3 + G4) (P = 0.012), and HBV infection (P = 0.043) in the high-risk group was higher than the low risk group.

In the GSE116174 cohort, 64 patients were enrolled. As shown in Supplementary Table 9, there were no significant differences were detected in the distribution of age (P = 0.516), gender (P = 0.418), vascular invasion (P = 0.612), HBV infection (P = 0.849), HCV infection (P = 0.139), Alcohol consumption (P = 0.167), and AJCC pathological stage (P = 0.754).

As shown in Supplementary Figure 2A–2D), the AJCC pathological stage, performance status, HBV infection and neoplasm histologic grade are significantly correlated with the 8-lncRNAs-based classifier. The 8-lncRNAs-based classifier scores for the performance status (2 & 3 & 4), stage (III & IV), HBV positive, and tumor grade (G3 & G4) groups were higher than those of the performance status (0 & 1), stage (I & II), HBV negative, and tumor grade (G1 & G2) groups.

Recurrence

In the training cohort, 130 patients were enrolled. As shown in Supplementary Table 10, the proportion of patients with performance status (2 + 3) (P=0.026), Child-Pugh classification (C&D) (P=0.030), TNM T stage (T3 + T4) (P=0.006), and AJCC pathological stage (III & IV) (P=0.023) in the high-risk group was higher than the low risk group.

In the test cohort, the proportion of patients with tumor grade (G3 + G4) (P=0.006) in the high-risk group was higher than the low risk group. The remaining risk factors were not significantly different in distribution compared to the training cohort (Supplementary Table 10).

In the TCGA cohort, a total 269 patients were enrolled. As shown in Supplementary Table 10, The proportion of patients with tumor grade (G3 + G4) (P = 0.005), HBV infection (P = 0.001), TNM T stage (T3 + T4) (P = 0.029), TNM N stage (N1) (P = 0.023), and AJCC pathological stage (III & IV) (P = 0.007) in the high-risk group was higher than the low risk group.

In the GSE76427 cohort, 81 patients were enrolled. As shown in Supplementary Table 10, there were no significant differences were detected in the distribution of age (P = 0.960), AJCC pathological stage (P = 0.303) and gender (P = 0.117).

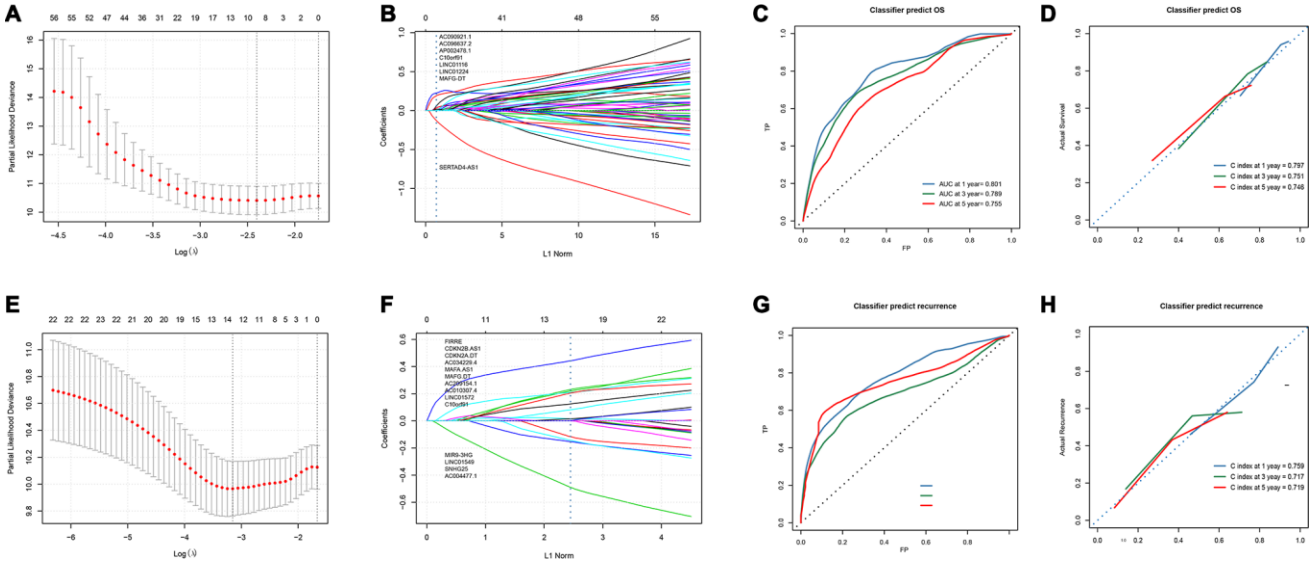


Figure 3. Development and verification of 8-lncRNAs-based and 14-lncRNAs-based classifiers. (**A**) LASSO coefficient profiles of the 86 Significant difference lncRNAs in OS set. A vertical line is drawn at the value chosen by 10-fold cross-validation; (**B**) Ten-time cross-validation for tuning parameter selection in the LASSO model; (**C**, **D**) Time-dependent ROC curves and calibration curves of 8-lncRNAs-based classifier; (**E**) Time-dependent ROC curves of Liao's biomarker for overall survival; (**F**) LASSO coefficient profiles of the 21 Significant difference lncRNAs in recurrence set, A vertical line is drawn at the value chosen by 10-fold cross-validation; (**G**) Ten-time cross-validation for tuning parameter selection in the LASSO model; (**H**) Time-dependent ROC curves and calibration curves of 14-lncRNAs-based classifier. LASSO, least absolute shrinkage and selection operator method; lncRNA, long non-coding RNA; OS, overall survival; ROC, receiver operating characteristic.

Table 1. Univariate and multivariate COX analyses of the lncRNA-based classifier for recurrence.

Prognostic parameter	Univariate analysis			Multivariate analysis		
	HR	95% CI	P value	HR	95% CI	P value
Training Cohort						
RiskScore	4.434	2.860-6.874	**0.001**	6.210	3.092-12.474	**0.001**
Age	0.980	0.613-1.566	0.932			
M	5.661	1.325-24.179	**0.019**	1.957	0.388-9.871	0.416
N	6.040	0.794-45.943	0.082			
Stage	2.392	1.364-4.196	**0.002**	4.872	2.023-11.732	**0.001**
T classification	2.486	1.469-4.207	**0.001**			
Bilirubin	1.140	0.908-1.430	0.259			
Child-Pugh classification	0.744	0.262-2.111	0.579			
Performance Status	1.766	1.276-2.445	**0.001**	1.119	0.740-1.692	0.595
Family History	0.851	0.508-1.428	0.542			
Fraction Genome Altered	2.187	0.671-7.133	0.194			
Grade	0.650	0.383-1.104	0.111			
Adjacent hepatic tissue inflammation	1.098	0.704-1.712	0.681			
HBV	0.403	0.230-0.705	**0.001**	0.782	0.402-1.984	0.782
HCV	1.463	0.845-2.530	0.173			
Alcohol	1.063	0.639-1.769	0.814			
Liver fibrosis Ishak score category	1.077	0.887-1.308	0.453			
Mutation Count	1.001	1.000-1.002	0.085			
Platelet count	1.000	1.000-1.000	0.452			
Race Category	1.123	0.893-1.413	0.320			
Albumin	1.067	1.014-1.123	**0.013**	1.008	0.953-1.065	0.792
Gender	1.252	0.745-2.103	0.386			
Vascular Invasion	0.785	0.509-1.211	0.273	0.536	0.291-0.987	**0.045**
BMI	0.979	0.936-1.024	0.363			
AFP	0.992	0.938-1.050	0.792			
Test Cohort						
RiskScore	1.448	1.060-1.977	**0.020**	1.448	0.903-2.324	0.125
Age	0.829	0.520-1.320	0.430			
M	1.000		1.000			
N	0.845	0.117-6.134	0.868			
Stage	0.001	1.530-3.945	**0.001**	0.658	0.051-8.531	0.749
T classification	2.232	1.405-3.547	**0.001**	2.311	0.196-27.305	0.506
Bilirubin	1.045	0.948-1.152	0.379			
Child-Pugh classification	3.187	1.218-8.338	**0.018**	2.516	0.636-9.949	0.188
Performance Status	1.922	1.471-2.513	**0.001**	1.494	0.802-2.784	0.206
Family History	0.929	0.566-1.526	0.772			
Fraction Genome Altered	2.598	0.808-8.350	0.109			
Grade	1.509	0.960-2.373	0.075			
Adjacent hepatic tissue inflammation	1.386	0.880-2.181	0.159			
HBV	0.537	0.314-0.918	**0.023**	0.793	0.395-1.592	0.514
HCV	1.578	0.864-2.882	0.138			
Alcohol	1.104	0.700-1.742	0.670			
Liver fibrosis Ishak score category	0.989	0.842-1.161	0.891			
Mutation Count	1.002	1.000-1.004	0.118			
Platelet count	1.000	1.000-1.000	0.260			

Race Category	0.995	0.785-1.259	0.964			
Albumin	0.986	0.937-1.038	0.592			
Gender	0.962	0.612-1.512	0.866			
Vascular Invasion	1.101	0.739-1.640	0.637	0.953	0.516-1.761	0.878
BMI	1.004	0.982-1.027	0.712			
AFP	1.042	0.988-1.097	0.127			
TCGA Cohort						
RiskScore	2.065	1.625-2.626	**0.001**	2.043	1.458-2.863	**<0.001**
Age	0.903	0.650-1.255	0.543			
M	7.067	2.197-22.730	**0.001**	7.520	2.250-25.14	**0.001**
N	1.648	0.405-6.701	0.485			
Stage	2.405	1.683-3.435	**0.001**			
T classification	2.320	1.643-3.277	**0.001**	1.646	0.374-7.247	0.510
Bilirubin	1.060	0.974-1.153	0.180			
Child-Pugh classification	1.323	0.658-2.661	0.433			
Performance Status	1.863	1.520-2.283	**0.001**	1.770	1. 376-2.276	**<0.001**
Family History	0.901	0.630-1.288	0.568			
Fraction Genome Altered	2.356	1.026-5.411	**0.043**	1.177	0.368-3.765	0.784
Grade	1.034	0.740-1.445	0.845			
Adjacent hepatic tissue inflammation	1.222	0.891-1.676	0.213			
HBV	0.457	0.311-0.671	**0.001**	0.849	0.510-1.414	0.530
HCV	1.453	0.971-2.172	0.069			
Alcohol	1.074	0.768-1.504	0.676			
Liver fibrosis Ishak score category	1.028	0.908-1.164	0.660			
Mutation Count	1.001	1.000-1.002	**0.031**			
Platelet count	1.000	1.000-1.000	0.189			
Race Category	1.054	0.895-1.241	0.530			
Albumin	0.999	0.994-1.004	0.700			
Gender	1.078	0.769-1.512	0.663			
Vascular Invasion	0.908	0.677-1.217	0.517	0.951	0.656-1.397	0.790
BMI	0.998	0.975-1.021	0.853			
AFP	1.016	0.978-1.056	0.403			
GSE76427 Cohort						
RiskScore	1.433	0.920-2.232	0.112			
Age	1.069	0.586-1.951	0.828			
Gender	0.613	0.269-1.394	0.243			
Stage	1.279	0.608-2.693	0.517			

HR, Hazard ratio; CI, confidence interval; lncRNA, long non-coding RNA.

As shown in Supplementary Figure 2A–2D), the AJCC pathological stage, performance status, HBV infection and neoplasm histologic grade were significantly correlated with the 14-lncRNAs-based classifier. The 14-lncRNAs-based classifier scores for the performance status (2 & 3), stage (III & IV), HBV positive, and tumor grade (G3 & G4) groups were higher than those of the performance status (0 & 1), stage (I & II), HBV negative, and tumor grade (G1 & G2) groups. In addition, we also investigated the relationship between a total of 20 lncRNAs. The results are shown in Supplementary Table 9.

Prognosis value of the lncRNA-based classifiers

Additionally, we assessed the prognostic value of lncRNA-based classifiers.

Overall survival

Cox univariate analysis showed that the Performance Status, the tumor stage, TNM T classification, HBV infection, and the 8-lncRNA-based classifier were correlated with OS, whether in the training cohort, test cohort, or the TCGA cohort. After multivariable

adjustment by these variables, Performance Status (HR: 2.589, 95% CI: 1.355-4.947; P = 0.004), TNM M stage (HR: 7.703, 95% CI: 1.603-37.021; P = 0.011), and the lncRNA-based classifier (HR: 15.483, 95% CI: 6.149-38.989; P < 0.001) remained to be powerful and independent factors for OS in the TCGA Cohort (Supplementary Table 11). In addition, multiple lncRNAs-based classifier was still an independent risk factor in the validation cohort (GSE116174).

In the time-dependent ROC curve, the 8-lncRNAs-based classifiers can effectively predict the 1-year, 3-year, and 5-year survival rates, and their AUC is 0.801, 0.789 and 0.755, respectively (Figure 3C). The average predicted probability (predicted survival rate) and Kaplan-Meier estimated (observed survival rate) were plotted, and the dotted line indicated the ideal reference line corresponding to the predicted survival rate and the actual survival rate. The calibration curve of 1-, 3- and 5-year survival probability based on 8-lncRNAs-classifier were in good agreement with the actual observed values. The C-index of 1-year, 3-year, and 5-year were 0.797, 0.751 and 0.746 respectively, indicating that the prediction model had high accuracy (Figure 3D). Compared with tdROC of liao et al., [13] a larger AUC indicated that our model had a good prediction ability (Supplementary Figure 4A–4C).

Recurrence

Cox univariate analysis showed that Performance Status, the tumor stage, TNM T stage, TNM M stage, HBV infection, and the 14-lncRNA-based classifier were correlated with recurrence, whether in the training cohort, test cohort, or the TCGA cohort. After multivariable adjustment by these variables, Performance Status (HR: 1.608, 95% CI: 1.213-2.131; P < 0.001), TNM M stage (HR: 5.782, 95% CI: 1.631-20.501; P = 0.007) and the lncRNA-based classifier (HR: 2.076, 95% CI: 1.457-2.957; P < 0.001) remained to be powerful and independent factors for recurrence in the TCGA cohort (Table 1).

In the time-dependent ROC curve, the 14-lncRNAs-based classifiers can effectively predict the 1-year, 3-year, and 5-year survival rates with AUC of 0.800, 0.686 and 0.789, respectively (Figure 3G). The average predicted probability (predicted survival rate) and Kaplan-Meier estimated (observed survival rate) were plotted, and the dotted line indicated the ideal reference line corresponding to the predicted survival rate and the actual survival rate. The calibration curve of 1-, 3- and 5-year survival probability based on 14-lncRNAs-classifier are in good agreement with the actual observed values. The C-index of 1-year, 3-year, and 5-year were 0.759, 0.717 and 0.719 respectively, indicating that the prediction model had good performance (Figure 3H). Compared with tdROC of liao et al., [13] a larger AUC indicated that our model had a good prediction ability. (Supplementary Figure 4D–4F)

Construction and evaluation of the nomogram

Subsequently, we constructed a gene-clinical nomogram (Figure 4A, 4B) by multivariate cox regression analysis (Supplementary Table 11, Table 1), combined with clinical characterization and lncRNAs-based classifier. TNM M stage, Performance Status and an 8-lncRNAs-based classifier were included in the gene-clinical nomogram of OS, while TNM M stage, Performance Status and 14-lncRNAs-based classifier were included in the gene-clinical nomogram of recurrence. Nomograms can visually predict the prognosis of patients according to their genes and clinical information, and accurately predict the survival and recurrence of patients at 1-, 3-, and 5 years. Moreover, the score of the nomogram was retained for development and validation of the performance of the nomogram risk score (Supplementary Tables 3 and 4). The risk score of OS-nomogram = 1/(-45.629*Classifier - 47*M - 16*Performance Status + 134.689). The risk score of recurrence-nomogram = 1/(-20*Classifier - 43*M -17.3*Performance Status + 94.7). And the accuracy of the nomogram in 1, 3 and 5 years was analyzed by tdROC, and the corresponding calibration curve was drawn (Figure 4).

Overall survival

The OS-nomogram based on an 8-lncRNAs-based classifier combined with the TNM M stage and Performance Status has an AUC of 0.898, 0.834, and 0.814 for predicting 1, 3, and 5 years of OS, respectively. The C-index of 1, 3, and 5 years was 0.878, 0.838, and 0.828, respectively (Figure 4C–4F). The results indicated that the combination of the lncRNA-based classifier models, TNM M stage and Performance Status could enhance the capability to predict the prognosis of survival. Kaplan-Meier curve analysis showed that the two groups divided by cutoff value (0.006953) calculated by X-tile were still statistically significant in OS (Supplementary Figure 3A–3C).

Kaplan-Meier curve showed that patients in the training cohort, the test cohort and the TCGA cohort distributed by lncRNA-based classifiers with Performance Status had significantly different prognosis (p < 0.0001, Figure 5A–5C). As shown in Supplementary Table 5, the tests were performed by log rank test between groups.

Recurrence

The td-ROC showed that the recurrence-nomogram based on classifier, TNM M stage and Performance Status has an AUC of 0.786, 0.711, and 0.752 for predicting 1, 3, and 5 years of recurrence, respectively. The C-index of 1, 3, and 5 years was 0.719, 0.692, and 0.692, respectively (Figure 4G–4J). The Kaplan-Meier curve analysis also indicated that the prognosis of patients stratified by cutoff value (0.01470) calculated by X-tile was significantly different (Supplementary Figure 3D–3F). The lncRNA-based classifiers with TNM stage could distinguish patients in the training cohort, the test cohort and the TCGA cohort into the different risk of recurrence (p< 0.0001, Figure 5D–5F). As shown in Supplementary Table 6, the tests were performed by log rank test between groups.

GSEA identifies KEGG signaling pathway

In order to investigate different activated KEGG signaling pathways in HCC, GSEA was performed on 8 OS-related lncRNAs and 14 recurrence-related lncRNAs expression datasets. We considered the difference as statistically significant when | NES | ≥ 1, NOM p-value <0.01 and FDR q-val <0.25. All significant enrichment pathways were listed in Supplementary Tables 7 and 8. Figure 6 showed the most significant KEGG pathway function enrichment of 8-OS-related lncRNAs. Figure 7 showed the most significant KEGG pathway function enrichment of 14-recurrence-related lncRNAs. The results showed that Aminoacyl TRNA biosynthesis, Arginine and proline metabolism, Basal transcription factor, Base excision repair, Bladder cancer, Cytoplasmic DNA sensing

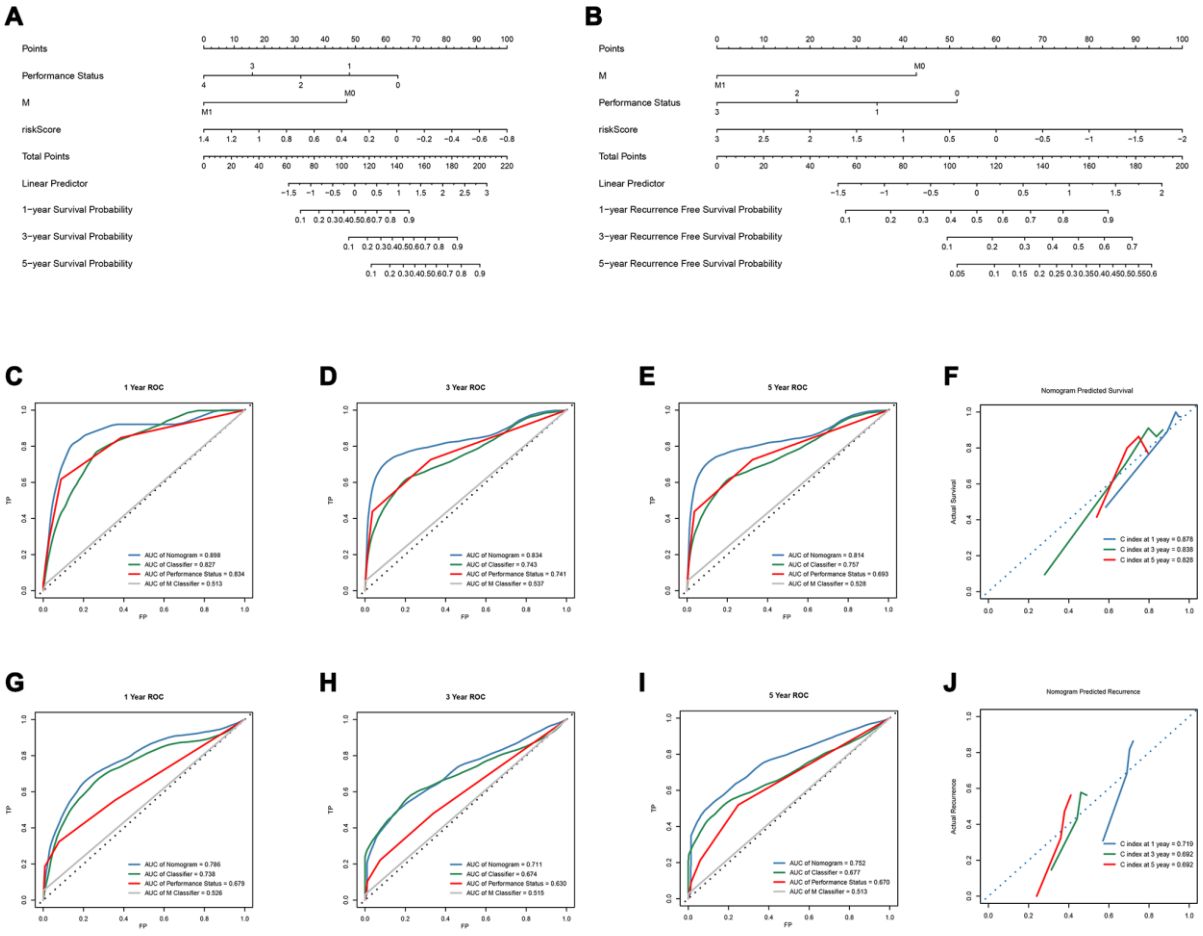


Figure 4. Development and verification of OS-nomogram and recurrence-nomogram. (**A**) OS-nomogram based on 8-lncRNAs-based classifier, TNM M classifier and Performance Status; (**B**) recurrence-nomogram based on 14-lncRNAs-based classifier, TNM M classifier and Performance Status; (**C**–**E**) The 1, 3, and 5-year Time-dependent ROC curves compare the prognostic accuracy of the OS-nomogram; (**F**) 1, 3, and 5 year calibration curve and C-index of the OS-nomogram; (**G**–**I**) The 1, 3, and 5-year Time-dependent ROC curves compare the prognostic accuracy of the recurrence-nomogram; (**J**) 1, 3, and 5 year calibration curve and C-index of the recurrence-nomogram. OS, overall survival; lncRNA, long non-coding RNA; ROC, receiver operating characteristic; C-index, concordance index.

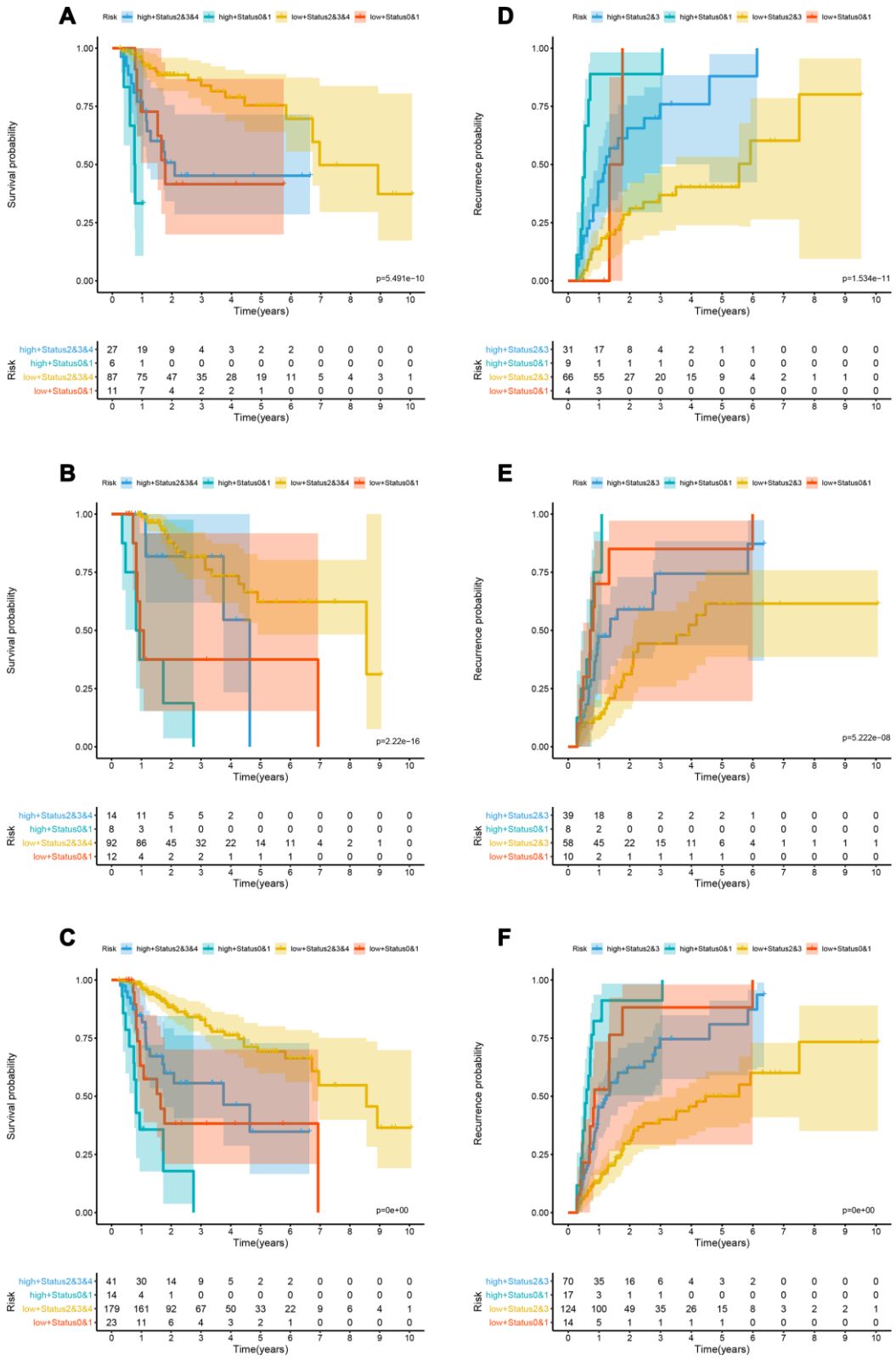


Figure 5. Kaplan-Meier analysis in the training, validation and whole cohorts according to the lncRNA-based classifiers stratified by clinicopathological risk factors. (**A**–**C**) Kaplan-Meier survival curves of LIHC patients with combinations of lncRNA-classifier and TNM T classifier in the training, test and TCGA cohorts for OS; (**D**–**F**) Kaplan-Meier survival curves of LIHC patients with combinations of lncRNA-classifier and TNM stage in the training, test and TCGA cohorts for OS. lncRNA, long non-coding RNA; OS, overall survival; LIHC, Liver hepatocellular carcinoma.

pathway, DNA replication, Epithelial Signaling in Helicobacter Pylori Infection, Focal adhesion, Gap junction, homologous recombination, leukocyte transendothelial migration, mapk signaling pathway, Nod like receptor signaling pathway, P53 signaling pathway, pathways in cancer, Nucleotide excision repair, RNA degradation, cell cycle, spliceosome, VEGF signaling pathway, and WNT signaling pathways showed consistent enrichment in the up-regulated phenotypes of 8-OS-related lncRNAs (Figure 6). The results showed that Basal transcription factor, Base excision repair, DNA replication, mismatch repair, Lysosome, proteasome, homologous recombination, leukocyte transendothelial migration, P53 signaling pathway, pathways in cancer, Nucleotide excision repair, RNA degradation, cell cycle, and spliceosome signaling pathways showed consistent enrichment in the up-regulated phenotypes of 14-recurrence-related lncRNAs (Figure 7).

DISCUSSION

HCC is a malignant tumor with high heterogeneity, which adds to the difficulty of prognosis and treatment [14]. The progression of hepatocellular carcinoma involves genetic and epigenetic changes, which are closely related to the poor prognosis of HCC [15]. Previous studies have shown that commonly used clinicopathological parameters (such as age, TNM staging, sex, viral infection, and AFP levels) were not sufficient to accurately predict prognosis of patients. Emerging evidence illustrates that lncRNA plays a critical role in regulating the progression of hepatocellular carcinoma (HCC) [16, 17]. Some studies have been conducted from the genetic perspective to screen lncRNA as a biomarker of HCC in the past few decades [18, 19]. To date, however, studies have attempted to predict the prognosis of patients by gene signature, but limited by sample size, biological heterogeneity, inappropriate data processing methods, and validation methods, there was usually not a good prediction ability.

In this study, we developed two lncRNA-based classifiers to predict survival and recurrence, respectively. Compared with previous studies, our research has the following advantages. First, we included 312 patients in the OS group and 269 patients in the recurrence group to reduce the deviation caused by insufficient sample size. Since only a small number of lncRNA were identified in previous studies, 15113 lncRNAs were isolated from the gene expression profile of LICH data set in this study. Furthermore, in order to identify useful lncRNA markers in high-dimensional data sets, an appropriate approach is required. The LASSO-Cox regression model was a popular tool for regression using high-dimensional predictors, which can more effectively perform dimensionality analysis on high-throughput sequencing data to construct more accurate gene signatures. The experience expansion of LASSO punishment can reduce the error discovery rate in the high-dimensional Cox regression model [20]. Finally, in addition to factors such as age, gender, TNM stage, and tumor stage, we also analyze AFP, HBV, HCV, Alcohol, Family history, Fibrosis Ishak score/Liver cirrhosis, BMI, Platelet result, Performance status, Child-Pugh grade, ALB, Region/Race, Adjacent tissue inflammation, etc. Among them, AFP, HCV, Alcohol, Family history, Fibrosis Ishak score/Liver cirrhosis, BMI, Platelet result, ALB, Region/Race,

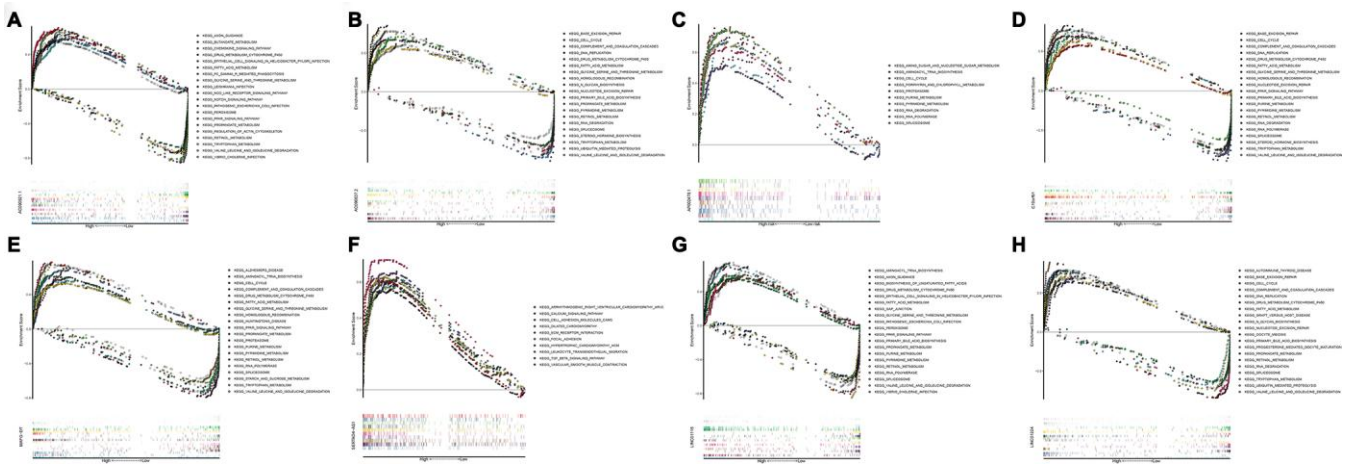


Figure 6. Gene set enrichment analysis of lncRNAs of overall survival related classifier. GSEA results showed in (**A**) AC090921.1, (**B**) AC096637.2, (**C**) AP002478.1, (**D**) C10orf91, (**E**) MAFG−DT, (**F**) SERTAD4−AS1, (**G**) LINC01116, and (**H**) LINC01224. GSEA, Gene set enrichment analysis.

Adjacent tissue inflammation are not independent factors with significant statistical differences. The lack of data in some of these samples may affect the results, and larger sample studies are still needed. In conclusion, the prognostic ability of classifiers in this study is more reliable and accurate than previous studies. In addition, the nomogram risk score based on classifier and clinical characterization as a method to predict prognosis provides a visual method for predicting OS and recurrence in HCC patients. The nomogram based on 8-lncRNAs-based classifier combined with TNM M stage and performance status can visually predict OS, and a 14-lncRNAs-based classifier combined with TNM M stage and performance status can be used to visually predict recurrence, both having excellent predictive power and accuracy.

Some of the lncRNAs involved in this study have been investigated in past studies. AP002478.1 can predict hepatitis virus positive HCC as prognostic targets [21]. Research by Lou et al suggests that C10orf91 is one of five lncRNAs expression as competing endogenous RNAs in regulating hepatoma carcinoma [22]. LINC01116 was significantly associated with HCC patients' poor outcomes [23]. CDKN2B-AS1 has been reported promotes tumor growth and metastasis of HCC by targeting let-7c-5p/NAP1L1 axis [24]. FIRRE has been reported in the literature to activate the Wnt/β-catenin signaling pathway to promote the growth of diffuse large B lymphoma cells by regulating the nuclear translocation of β-catenin [25]. Chen et al. found that LINC01572 can distinguish between early and advanced lung squamous cell carcinoma [26]. MIR9-3HG was considered to be related to the survival time of Head and neck squamous cell carcinoma in the study by Hu et al. [27]. These lncRNAs have been studied in a variety of cancers, including HCC. In the classifier of our study, AC090921.1, AC096637.2,

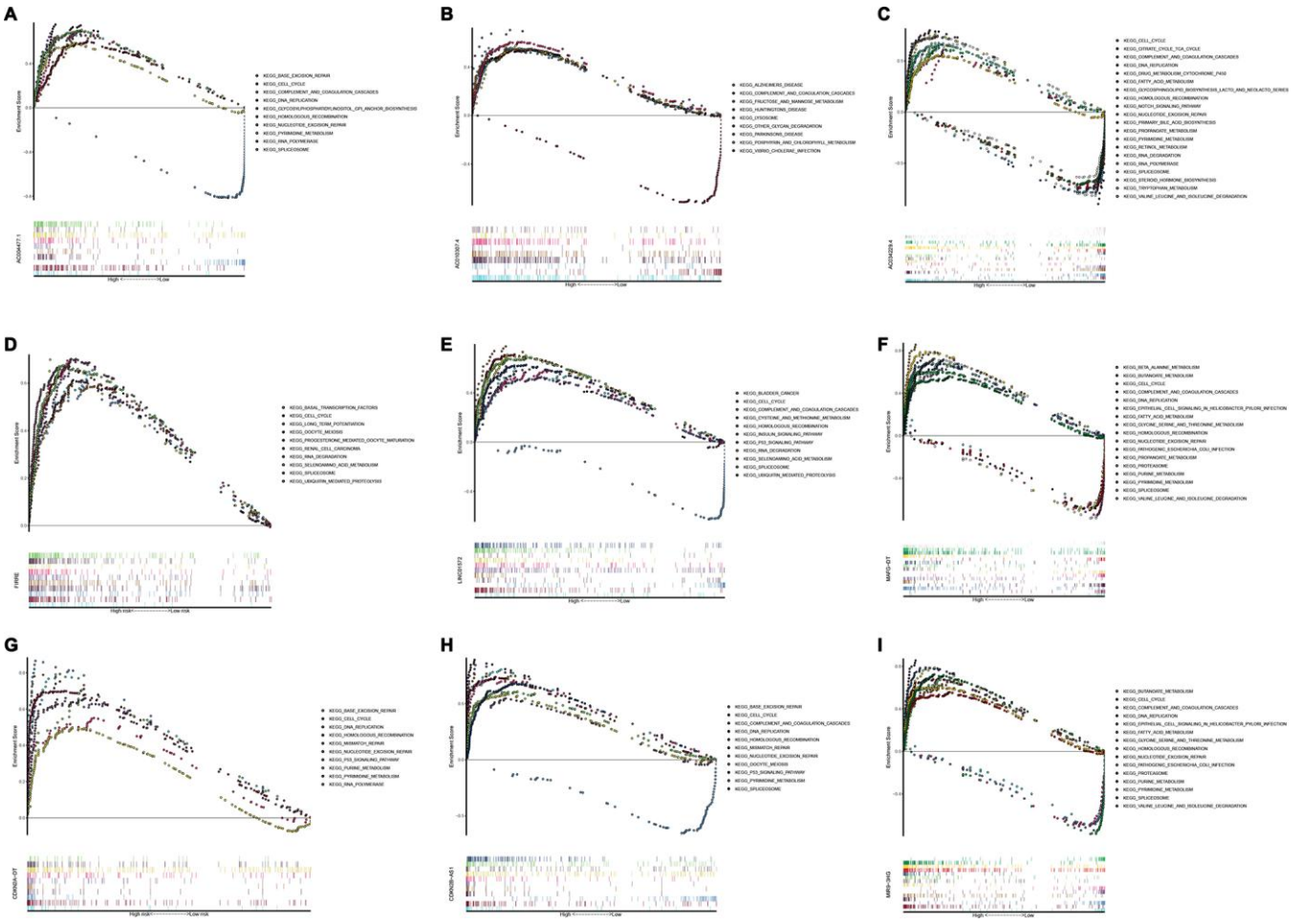


Figure 7. Gene set enrichment analysis of lncRNAs of recurrence related classifier. GSEA results showed in (**A**) AC004477.1, (**B**) AC010307.4, (**C**) AC034229.4, (**D**) FIRRE, (**E**) LINC01572, (**F**) MAFG−DT, (**G**) CDKN2A−DT, (**H**) CDKN2B−AS1, and (**I**) MIR9−3HG. GSEA, Gene set enrichment analysis.

LINC01224, SERTAD4-AS1, AC004477.1, AC034229.4, AC209154.1, CDKN2A-DT, CDKN2B-AS1, LINC01549, MAFA-AS1, MAFA-AS1, MAFG-DT, MIR9-3HG and SNHG25 has not been reported to related to HCC biology, the functions and mechanisms of these lncRNAs in HCC need to be further investigated.

To further explore the function of the 20 lncRNAs in this study. GSEA was used to detect its genetic enrichment. KEGG pathway analysis showed that these genes are associated with rich metabolic pathways. Enrichment with phenotypic consistency was also found in pathways such as Aminoacyl TRNA biosynthesis, Arginine and proline metabolism, Basal transcription factor, Base excision repair, Bladder cancer, Cytoplasmic DNA sensing pathway, DNA replication, Epithelial Signaling in Helicobacter Pylori Infection, Focal adhesion, Gap junction, homologous recombination, leukocyte transendothelial migration, Mapk signaling pathway, Nod like receptor signaling pathway, p53 signaling pathway, pathways in cancer, nucleotide excision repair, RNA degradation, cell cycle, spliceosome, VEGF signaling pathway, and WNT signaling pathways. These results indicate that these 20 lncRNAs may participate in the occurrence and development of HCC through these pathways. As an important process of cell division and growth, the active DNA replication pathway promotes tumor growth and proliferation. Studies have shown that N7-alkyl-dG can block DNA replication, suggesting that these lncRNAs can be potential targets for tumor drug treatment [28]. In this study, multiple lncRNAs were enriched in the cell cycle pathway, and drugs acting on the cell cycle may benefit patients [29, 30]. However, further research is needed to investigate and verify the function of these 22 lncRNAs.

Current research inevitably has some limitations that can be explored in the future. First, we developed a lncRNAs-based classifier based on half of the LIHC data and used another part for verification, but the limited number of validation sets required a larger sample to further validate our model. Secondly, this study was based on a study of the TCGA database that determines a retrospective study, and a larger sample of more regional prospective studies was still needed. Third, in this study, the significance of lncRNAs in the development of HCC is unquestionable, but the mechanism behind it was not yet clear and further researches were needed. Moreover, the RNA sequencing data of this study were based on clinical specimens, which increase the difficulty of clinical application. Finally, whether it was TCGA's RNA-seq or GEO's Array chip, its expensive price was also an obstacle to clinical practice. If we could extract the lncRNA we need a more accessible blood sample, it would become a more potentially valuable method.

In conclusion, we proved that the lncRNA-based classifier devised by LASSO cox method can accurately predict survival and recurrence, and divide HCC patients into low- and high-risk groups. Furthermore, the novel nomogram constructed based on this classifier combined with clinical characterization can not only visually predict HCC survival and recurrence, but also increase its prognostic value, making it a potentially valuable biomarker signature in clinical practice.

MATERIALS AND METHODS

Patient data

Liver Hepatocellular Carcinoma (LIHC) read counts data was downloaded from TCGA, a publicly available portal (up to May 10, 2019, https://tcga-data.nci.nih.gov/tcga/). Forty-nine adjacent non-tumor samples and 368 HCC samples were obtained after the removal of non-HCC patients and patients who lost critical data. Clinical characteristics of patients were obtained from the cBioportal platform (http://cbioportal.org/) [31], A web resource for visual exploration and analysis multidimensional cancer genome data. The exclusion criteria were as follows: 1) not HCC samples; 2) samples with clinical data but without lncRNA sequence data; 3) samples missing important clinical or biological data; and 4) Patients were followed up for less than three months. After the removal of non-HCC patients and patients lacking critical information, 312 patients were finally reserved for further study. The expression matrices for GSE76427 and GSE116174 were downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). The expression matrices for GSE76427 and GSE116174 were downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). GSE116174 contains 64 patients with OS > 90 days. In GSE76427, 81 patients with recurrence follow-up longer than 90 days were used for further analysis.

Data processing

LncRNAs of LIHC set was re-annotated by the gene annotation file " gencode.v30.long_noncoding_RNAs ", which is Downloaded from the gencode website. 15113 lncRNAs were identified from LIHC set. The expression value of each lncRNA was normalized with the TMM function of the 'limma' and 'edgeR' package for further analysis [32]. We used the 'edgeR' package to test all the data to identify lncRNAs that were differentially expressed by $|logFC| > 2$ and $padj < 0.05$ in the tumor compared with normal samples. Then, the

differentially expressed lncRNAs (DElncRNAs) were subjected to univariate Cox regression, and lncRNAs with $p < 0.05$ were identified as prognostic DElncRNAs for further research. GEO's lncRNAs were identified by the gpl platform annotation file and the Fasta file "gencode.v32.lncRNA_transcripts.fa".3494 lncRNAs were identified from GSE76427 and 5690 lncRNAs were identified from GSE116174. The OS-related prognostic DElncRNAs that intersect with GSE116174 are used for the development of OS-classifiers. Prognostic DElncRNAs associated with recurrences that intersect with GSE76427 are used to construct recurrence-classifiers. Because the detection methods used by the TCGA database and the GEO database are different, the background noise is also different. So we log2 (x + 1) the TCGA data set and normalize it using the zscore method. zscore normalization is also performed in the GEO dataset. The "sva" package is used to remove batch effects between TCGA and GEO datasets.

Construction of lncRNAs classifier

LASSO is a commonly used high-dimensional predictive regression method. The method is a compression estimation. By constructing a penalty function, it can get a more refined model, and make it compress some regression coefficients, that is, the sum of the absolute values of the forcing coefficients is less than a certain fixed value. Also, set some regression coefficients to be zero. Therefore, it retains the advantage of subset contraction and is a biased estimation for processing data with complex collinearity [33]. The lncRNAs related to OS and recurrence were identified by LASSO COX regression model [34]. The regression coefficients (β) of each related lncRNAs are reserved for the development of lncRNAs-based classifier. The lncRNAs based classifier = $\sum$ EXP(lncRNA) * β. Based on the optimal cut-off value calculated by x-tile software version 3.6.1 (Yale University School of Medicine, New Haven, CT, USA), the LIHC set was divided into high and low risk groups [35]. The time-dependent receiver operating characteristic (tdROC) curve analysis, calibration curve analysis, Kaplan-Meier survival analysis were used to evaluate predictive ability of the models in training cohort, test cohort, TCGA cohort, and the GEO cohort. After that, we began to construct genomic-clinical nomograms to predict the prognosis and mortality of each HCC patient individually [36].

Data analysis

The Chi-square test, COX survival analysis, and other data processing were completed by SPSS 19.0. Kaplan-Meier log rank test was calculated by medcalc (Version 19.0). Time-dependent ROC (tdROC) was used to assess the performance of lncRNA-based classifier with "time ROC" package in R software(Version 3.6.1). And area under ROC (AUC) was used to assess the accuracy of the prediction. Calibration curve and C-index are performed by 'rms' package. The larger C-index indicates that the prediction model has better accuracy [37]. 'Hmisc', 'rms', and 'survival' were used to develop a nomogram. When all the hypotheses are $P < 0.05$, the difference is statistically significant.

Gene set enrichment analysis

To identify the activation of different KEGG signaling pathways in HCC, we conducted GSEA between down-regulated and up-regulated phenotypes. The lncRNAs of the classifier were divided into up- and down-regulated groups by median. Gene Set Enrichment Analysis (GSEA) was performed by JAVA program GSEA 4.0.2 with the MSigDB Collection (c2.cp.kegg.v7.0.symbol). Normalized enrichment score (NES), nominal p-value and false discovery rate (FDR) were used to quantify enrichment magnitude and statistical significance, respectively [38]. When $|NES| \geq 1$, FDR q-val <0.25 and NOM p-value <0.01 were considered significant.

Abbreviations

AUC: area under receiver operating characteristics; CI: confidence interval; C-index: concordance index; HCC: hepatocellular carcinoma; HR: hazard ratio; lncRNA: long non-coding RNA; LASSO: least absolute shrinkage and selection operator method; OS: overall survival; ROC: receiver operating characteristic; TCGA: The Cancer Genome Atlas; tdAUC: time-dependent area under receiver operating characteristic; td-ROC: time-dependent receiver operating characteristic.

AUTHOR CONTRIBUTIONS

Zhongjing Zhang: study design, data collection, and analysis, interpreted data, drafted the manuscript. Wanqing Weng, Weiguo Huang and Tuo Deng, data analysis and helped to draft the manuscript. Boda Wu, Yi Zhou, Jie Zhang, Wen Ye, Jiecheng Zhang, Jianyang Ao: data collection and analysis, interpreted data, prepared figures. KeQing Shi and Qiyu Zhang: study design, study supervision, obtained funding and helped to draft the manuscript. All authors saw and approved the final version of the paper.

CONFLICTS OF INTEREST

These authors declare no conflicts of interest.

FUNDING

This work was supported by grants from the Natural Science Foundation of Zhejiang Province (LY17H160057).

REFERENCES

1. Yang JD, Hainaut P, Gores GJ, Amadou A, Plymoth A, Roberts LR. A global view of hepatocellular carcinoma: trends, risk, prevention and management. Nat Rev Gastroenterol Hepatol. 2019; 16:589–604. https://doi.org/10.1038/s41575-019-0186-y PMID:31439937

2. Forner A, Reig M, Bruix J. Hepatocellular carcinoma. Lancet. 2018; 391:1301–14. https://doi.org/10.1016/S0140-6736(18)30010-2 PMID:29307467

3. Carbonell D, Suárez-González J, Chicano M, Andrés-Zayas C, Triviño JC, Rodríguez-Macías G, Bastos-Oreiro M, Font P, Ballesteros M, Muñiz P, Balsalobre P, Kwon M, Anguita J, et al. Next-generation sequencing improves diagnosis, prognosis and clinical management of myeloid neoplasms. Cancers (Basel). 2019; 11:1364. https://doi.org/10.3390/cancers11091364 PMID:31540291

4. Alifrangis CC, McDermott U. Reading between the lines; understanding drug response in the post genomic era. Mol Oncol. 2014; 8:1112–19. https://doi.org/10.1016/j.molonc.2014.05.014 PMID:24957465

5. Diederichs S, Bartsch L, Berkmann JC, Fröse K, Heitmann J, Hoppe C, Iggena D, Jazmati D, Karschnia P, Linsenmeier M, Maulhardt T, Möhrmann L, Morstein J, et al. The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. EMBO Mol Med. 2016; 8:442–57. https://doi.org/10.15252/emmm.201506055 PMID:26992833

6. Peschansky VJ, Wahlestedt C. Non-coding RNAs as direct and indirect modulators of epigenetic regulation. Epigenetics. 2014; 9:3–12. https://doi.org/10.4161/epi.27473 PMID:24739571

7. Khandelwal A, Bacolla A, Vasquez KM, Jain A. Long non-coding RNA: a new paradigm for lung cancer. Mol Carcinog. 2015; 54:1235–51. https://doi.org/10.1002/mc.22362 PMID:26332907

8. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, Nagano T, Mancini-Dinardo D, Kanduri C. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell. 2008; 32:232–46. https://doi.org/10.1016/j.molcel.2008.08.022 PMID:18951091

9. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, et al. Landscape of transcription in human cells. Nature. 2012; 489:101–08. https://doi.org/10.1038/nature11233 PMID:22955620

10. Yao RW, Wang Y, Chen LL. Cellular functions of long noncoding RNAs. Nat Cell Biol. 2019; 21:542–51. https://doi.org/10.1038/s41556-019-0311-8 PMID:31048766

11. Yang X, Sun L, Wang L, Yao B, Mo H, Yang W. LncRNA SNHG7 accelerates the proliferation, migration and invasion of hepatocellular carcinoma cells via regulating miR-122-5p and RPL4. Biomed Pharmacother. 2019; 118:109386. https://doi.org/10.1016/j.biopha.2019.109386 PMID:31545291

12. Zhou JF, Shi YT, Wang HG, Yang XZ, Wu SN. Overexpression of long noncoding RNA HOXC13-AS and its prognostic significance in hepatocellular carcinoma. Eur Rev Med Pharmacol Sci. 2019; 23:7369–74. https://doi.org/10.26355/eurrev_201909_18843 PMID:31539123

13. Liao X, Yu T, Yang C, Huang K, Wang X, Han C, Huang R, Liu X, Yu L, Zhu G, Su H, Qin W, Deng J, et al. Comprehensive investigation of key biomarkers and pathways in hepatitis B virus-related hepatocellular carcinoma. J Cancer. 2019; 10:5689–704. https://doi.org/10.7150/jca.31287 PMID:31737106

14. Lin DC, Mayakonda A, Dinh HQ, Huang P, Lin L, Liu X, Ding LW, Wang J, Berman BP, Song EW, Yin D, Koeffler HP. Genomic and epigenomic heterogeneity of hepatocellular carcinoma. Cancer Res. 2017; 77:2255–65. https://doi.org/10.1158/0008-5472.CAN-16-2822 PMID:28302680

15. Kou Y, Koag MC, Lee S. N7 methylation alters hydrogen-bonding patterns of guanine in duplex DNA. J Am Chem Soc. 2015; 137:14067–70. https://doi.org/10.1021/jacs.5b10172 PMID:26517568

16. Wang B, Tang D, Zhang Z, Wang Z. Identification of aberrantly expressed lncRNA and the associated TF-

mRNA network in hepatocellular carcinoma. J Cell Biochem. 2020; 121:1491–503.
https://doi.org/10.1002/jcb.29384
PMID:31498488

17. Shen JY, Li C, Wen TF, Yan LN, Li B, Wang WT, Yang JY, Xu MQ. Alpha fetoprotein changes predict hepatocellular carcinoma survival beyond the milan criteria after hepatectomy. J Surg Res. 2017; 209:102–11.
https://doi.org/10.1016/j.jss.2016.10.005
PMID:28032546

18. Long J, Zhang L, Wan X, Lin J, Bai Y, Xu W, Xiong J, Zhao H. A four-gene-based prognostic model predicts overall survival in patients with hepatocellular carcinoma. J Cell Mol Med. 2018; 22:5928–38.
https://doi.org/10.1111/jcmm.13863
PMID:30247807

19. Zhang Z, Ouyang Y, Huang Y, Wang P, Li J, He T, Liu Q. Comprehensive bioinformatics analysis reveals potential lncRNA biomarkers for overall survival in patients with hepatocellular carcinoma: an on-line individual risk calculator based on TCGA cohort. Cancer Cell Int. 2019; 19:174.
https://doi.org/10.1186/s12935-019-0890-2
PMID:31312112

20. Ternès N, Rotolo F, Michiels S. Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional cox regression models. Stat Med. 2016; 35:2561–73.
https://doi.org/10.1002/sim.6927 PMID:26970107

21. Zhou RS, Zhang EX, Sun QF, Ye ZJ, Liu JW, Zhou DH, Tang Y. Integrated analysis of lncRNA-miRNA-mRNA ceRNA network in squamous cell carcinoma of tongue. BMC Cancer. 2019; 19:779.
https://doi.org/10.1186/s12885-019-5983-8
PMID:31391008

22. Lou X, Li J, Yu D, Wei YQ, Feng S, Sun JJ. Comprehensive analysis of five long noncoding RNAs expression as competing endogenous RNAs in regulating hepatoma carcinoma. Cancer Med. 2019; 8:5735–49.
https://doi.org/10.1002/cam4.2468
PMID:31392826

23. Jiang H, Shi X, Ye G, Xu Y, Xu J, Lu J, Lu W. Up-regulated long non-coding RNA DUXAP8 promotes cell growth through repressing krüppel-like factor 2 expression in human hepatocellular carcinoma. Onco Targets Ther. 2019; 12:7429–36.
https://doi.org/10.2147/OTT.S214336
PMID:31571902

24. Huang Y, Xiang B, Liu Y, Wang Y, Kan H. LncRNA CDKN2B-AS1 promotes tumor growth and metastasis of human hepatocellular carcinoma by targeting let-7c-5p/NAP1L1 axis. Cancer Lett. 2018; 437:56–66.
https://doi.org/10.1016/j.canlet.2018.08.024
PMID:30165194

25. Shi X, Cui Z, Liu X, Wu S, Wu Y, Fang F, Zhao H. LncRNA FIRRE is activated by MYC and promotes the development of diffuse large b-cell lymphoma via Wnt/β-catenin signaling pathway. Biochem Biophys Res Commun. 2019; 510:594–600.
https://doi.org/10.1016/j.bbrc.2019.01.105
PMID:30739786

26. Chen WJ, Tang RX, He RQ, Li DY, Liang L, Zeng JH, Hu XH, Ma J, Li SK, Chen G. Clinical roles of the aberrantly expressed lncRNAs in lung squamous cell carcinoma: a study based on RNA-sequencing and microarray data mining. Oncotarget. 2017; 8:61282–304.
https://doi.org/10.18632/oncotarget.18058
PMID:28977863

27. Hu Y, Guo G, Li J, Chen J, Tan P. Screening key lncRNAs with diagnostic and prognostic value for head and neck squamous cell carcinoma based on machine learning and mRNA-lncRNA co-expression network analysis. Cancer Biomark. 2020; 27:195–206.
https://doi.org/10.3233/CBM-190694
PMID:31815689

28. Koag MC, Kou Y, Ouzon-Shubeita H, Lee S. Transition-state destabilization reveals how human DNA polymerase β proceeds across the chemically unstable lesion N7-methylguanine. Nucleic Acids Res. 2014; 42:8755–66.
https://doi.org/10.1093/nar/gku554
PMID:24966350

29. Fekry MI, Ezzat SM, Salama MM, Alshehri OY, Al-Abd AM. Bioactive glycoalkaloides isolated from solanum melongena fruit peels with potential anticancer properties against hepatocellular carcinoma cells. Sci Rep. 2019; 9:1746.
https://doi.org/10.1038/s41598-018-36089-6
PMID:30741973

30. Kou Y, Koag MC, Cheun Y, Shin A, Lee S. Application of hypoiodite-mediated aminyl radical cyclization to synthesis of solasodine acetate. Steroids. 2012; 77:1069–74.
https://doi.org/10.1016/j.steroids.2012.05.002
PMID:22583912

31. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013; 6:pl1.
https://doi.org/10.1126/scisignal.2004088
PMID:23550210

32. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014; 15:R29. https://doi.org/10.1186/gb-2014-15-2-r29 PMID:24485249

33. Duan J, Soussen C, Brie D, Idier J, Wan M, Wang YP. Generalized LASSO with under-determined regularization matrices. Signal Processing. 2016; 127:239–46. https://doi.org/10.1016/j.sigpro.2016.03.001 PMID:27346902

34. Tibshirani R. The lasso method for variable selection in the cox model. Stat Med. 1997; 16:385–95. https://doi.org/10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3 PMID:9044528

35. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. Clin Cancer Res. 2004; 10:7252–59. https://doi.org/10.1158/1078-0432.CCR-04-0713 PMID:15534099

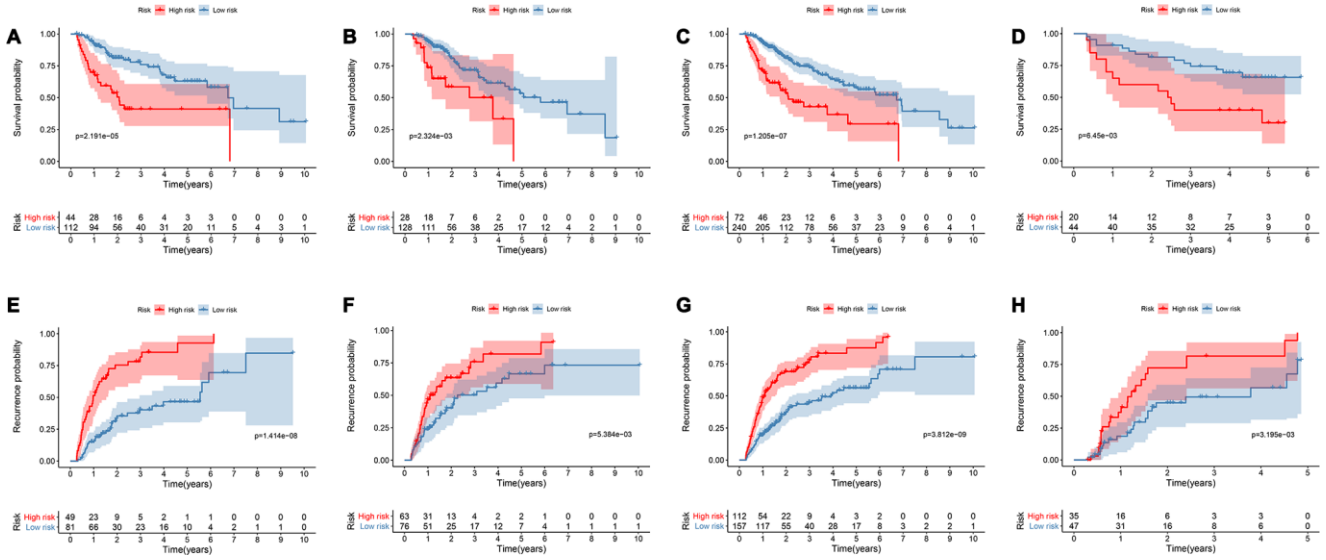
36. Tang H, Wu Z, Zhang Y, Xia T, Liu D, Cai J, Ye Q. Identification and function analysis of a five-long noncoding RNA prognostic signature for endometrial cancer patients. DNA Cell Biol. 2019; 38:1480–98. https://doi.org/10.1089/dna.2019.4944 PMID:31539276

37. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med. 2013; 32:5381–97. https://doi.org/10.1002/sim.5958 PMID:24027076

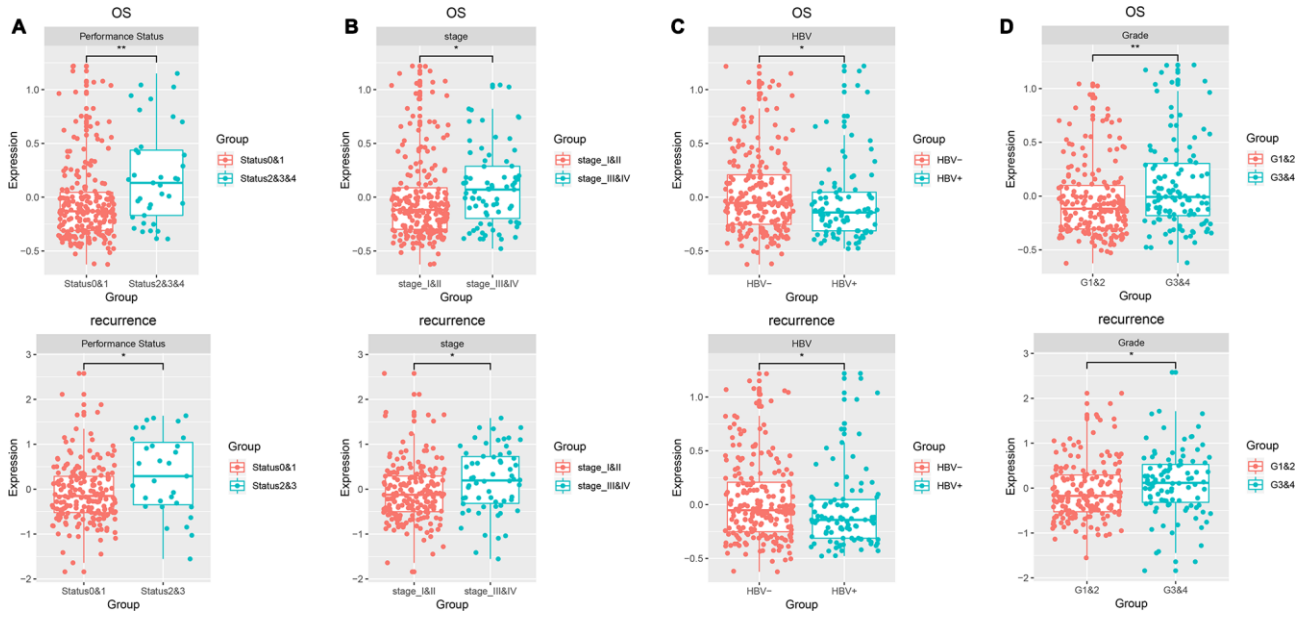
38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102:15545–50. https://doi.org/10.1073/pnas.0506580102 PMID:16199517

SUPPLEMENTARY MATERIALS

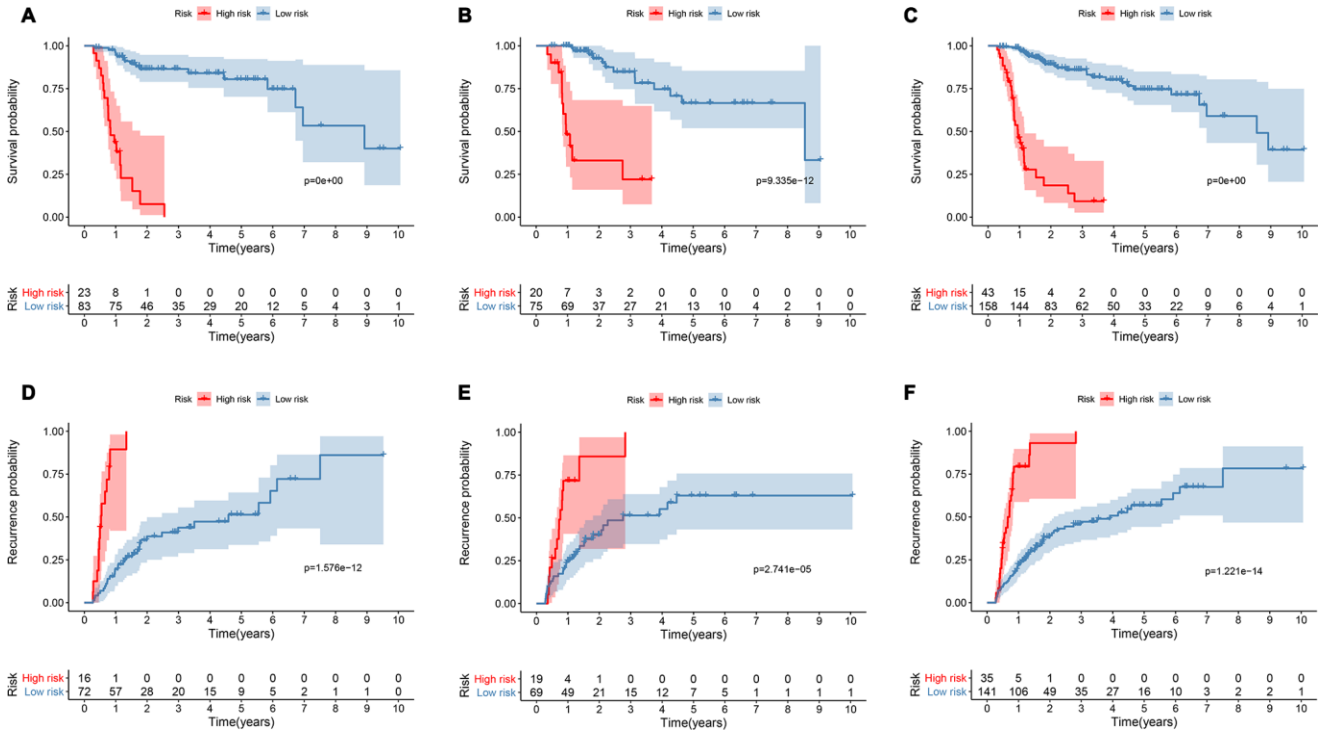
Supplementary Figures



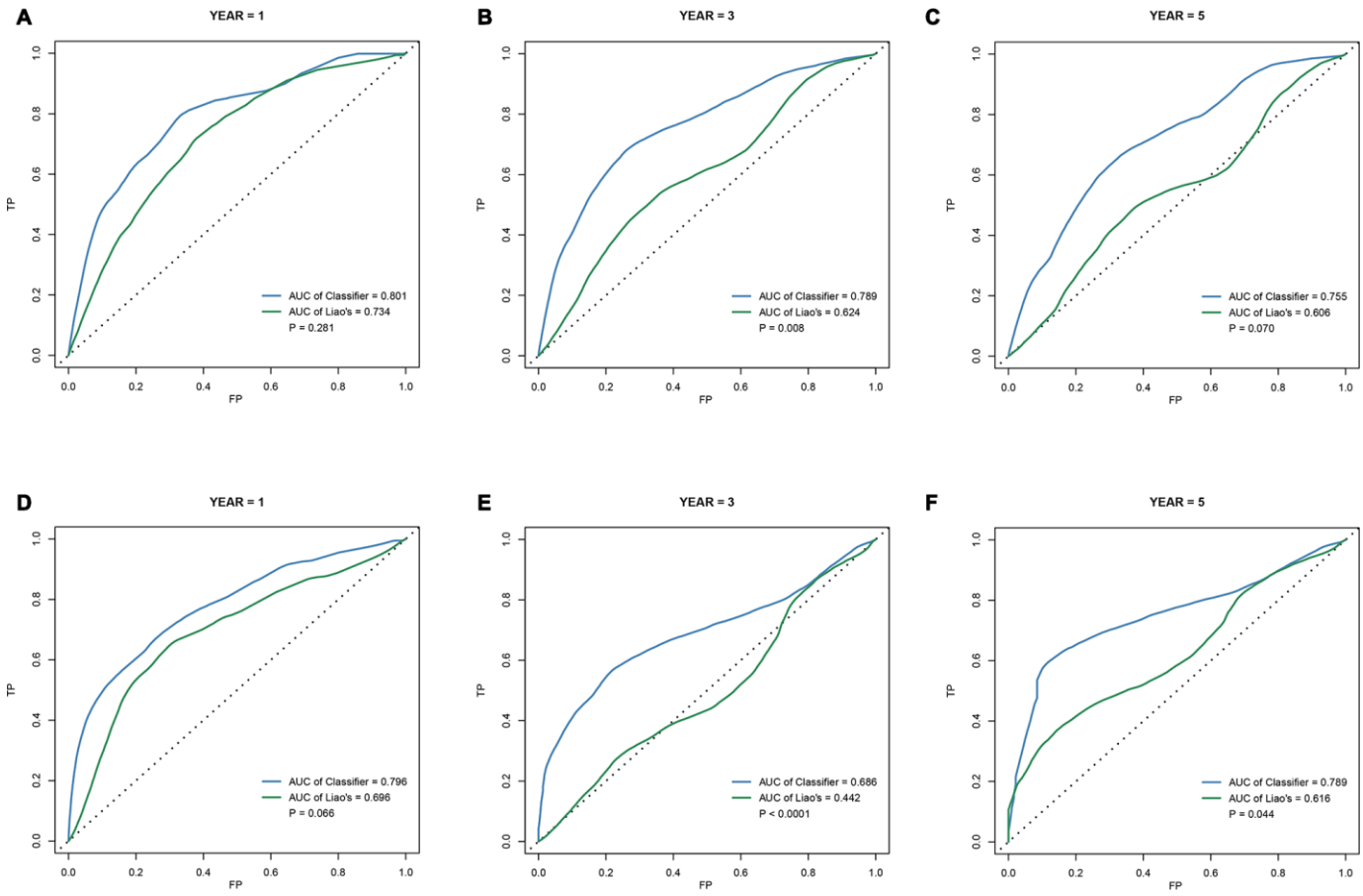
Supplementary Figure 1. Kaplan-Meier analysis in the training, validation and whole cohorts according to the lncRNA-based classifiers. Kaplan-meier survival analysis was performed to predict overall survival in the (**A**) training cohort, (**B**) test cohort, (**C**) TCGA cohort and (**D**) GEO cohort, according to the high-risk and low-risk groups stratified by the 8-lncRNAs-based classifier. Kaplan-meier survival analysis was performed to predict recurrence in the (**E**) training cohort, (**F**) test cohort, (**G**) TCGA cohort and (**H**) GEO cohort,, according to the high-risk and low-risk groups stratified by the 14-lncRNAs-based classifier.



Supplementary Figure 2. Boxplot of lncRNA-based classifier score in patients with clinicopathological risk factors. Boxplot of 8-lncRNAs-based classifier score and 14-lncRNAs-based classifier score in patients with (**A**) Performance Status, (**B**) TNM stage, (**C**) HBV, and (**D**) grade.



Supplementary Figure 3. Kaplan-Meier analysis in the training, validation and whole cohorts according to the molecular-clinicopathological nomograms. Kaplan-meier survival analysis was performed to predict overall survival in the (**A**) training cohort, (**B**) test cohort, and (**C**) TCGA cohort, according to the high-risk and low-risk groups stratified by the OS-nomogram. Kaplan-meier survival analysis was performed to predict recurrence in the (**D**) training cohort, (**E**) test cohort, and (**F**) TCGA cohort, according to the high-risk and low-risk groups stratified by the recurrence-nomogram. OS, overall survival.



Supplementary Figure 4. Comparison of Classifiers and Liao's Biomarkers. (**A**–**C**) The 1, 3, and 5-year Time-dependent ROC curves compare the prognostic accuracy of the OS-related Classifier and Liao's Biomarkers (**D**–**F**) The 1, 3, and 5-year Time-dependent ROC curves compare the prognostic accuracy of the recurrence -related Classifier and Liao's Biomarkers; OS, overall survival; lncRNA, long non-coding RNA; ROC, receiver operating characteristic.

Supplementary Tables

Please browse Full Text version to see the data of Supplementary Table 1.

Supplementary Table 1. lncRNA differentially expressed in tumor vs. normal.

Supplementary Table 2. The detailed information of lncRNAs for constructing the prognostic signature.

Gene name	ENSG_ ID	Gene_type	bp	Chromosome	β	Cutoff Value
8-lncRNA-based classifier for OS						**0.2**
AC090921.1	ENSG00000214803	lincRNA	1195	Chromosome 8: 124,192,671-124,247,398	0.0299	0.7
AC096637.2	ENSG00000265415	antisense	1846	Chromosome 17: 59,202,677-59,203,829	0.0125	0.8
AP002478.1	ENSG00000266401	antisense	1455	Chromosome 18: 3,653,030-3,656,282	0.1838	1.5
C10orf91	ENSG00000180066	lincRNA	1846	Chromosome10: 132,444,327-132,449,408	0.2221	1.3
LINC01116	ENSG00000163364	lincRNA	1502	Chromosome 21: 44,477,850-44,478,493	0.0437	1.4
LINC01224	ENSG00000269416	lincRNA	2766	Chromosome 19: 23,399,233-23,416,075	0.0251	1.2
MAFG-DT	ENSG00000265688	bidirectional_promoter_ lncRNA	1895	Chromosome 17: 81,927,829-81,930,753	0.0137	1.3
SERTAD4-AS1	ENSG00000203706	antisense	726	Chromosome 1: 210,231,456-210,234,047	-0.1168	-0.9
14-lncRNA-based classifier for recurrence						**0.1**
AC004477.1	ENSG00000263412	processed_transcript	2910	Chromosome 17: 48,045,141-48,048,073	-0.0255	0.6
AC010307.4	ENSG00000250244	antisense	389	Chromosome 5: 133,256,492-133,275,977	0.1647	1.5
AC034229.4	ENSG00000272417	lincRNA	441	Chromosome 5: 10,203,600-10,204,040	0.0416	1.0
AC209154.1	ENSG00000276399	lincRNA	3801	Chromosome 17: 22,406,019-22,413,744	0.1580	0.3
C10orf91	ENSG00000180066	lincRNA	1846	Chromosome10: 132,444,327-132,449,408	0.3958	0.5
CDKN2A-DT	ENSG00000224854	antisense	823	Chromosome 9: 21,966,929-21,967,751	0.0233	0.7
CDKN2B-AS1	ENSG00000240498	antisense	7173	Chromosome 9: 21,994,139-22,128,103	0.0037	0.7
FIRRE	ENSG00000213468	processed_transcript	5506	Chromosome X: 131,688,779-131,830,862	0.00057	1.1
LINC01549	ENSG00000232560	lincRNA	1702	Chromosome 21: 17,438,821-17,450,104	-0.1140	-1.1
LINC01572	ENSG00000261008	lincRNA	3298	Chromosome 16: 72,236,281-72,665,014	0.1813	0.5
MAFA-AS1	ENSG00000254338	antisense	417	Chromosome 8: 143,417,679-143,419,150	0.0958	1.3
MAFG-DT	ENSG00000265688	bidirectional_promoter_ln cRNA	1895	Chromosome 17: 81,927,829-81,930,753	0.1348	-0.8
MIR9-3HG	ENSG00000255571	lincRNA	5607	Chromosome 15: 89,361,579-89,398,487	-0.365	-0.7
SNHG25	ENSG00000266402	lincRNA	278	Chromosome 17: 64,145,970-64,146,476	-0.0761	0.4
LINC02499	ENSG00000250436.1	lincRNA	763	Chromosome 4: 73,508,803-73,534,128	-0.19279	

Supplementary Table 3. Correlation points about nomogram prediction of overall survival.

Performance Status	Points	riskScore	Points	Total Points	1-year Survival Probability	Total Points	3-year Survival Probability	Total Points	5-year Survival Probability
0	64	-0.8	100	70	0.1	105	0.1	122	0.1
1	48	-0.6	91	84	0.2	119	0.2	135	0.2
2	32	-0.4	82	93	0.3	128	0.3	145	0.3
3	16	-0.2	73	102	0.4	137	0.4	153	0.4
4	0	0	64	110	0.5	145	0.5	161	0.5
M	**Points**	0.2	55	117	0.6	152	0.6	169	0.6
0	47	0.4	45	126	0.7	161	0.7	177	0.7
1	0	0.6	36	135	0.8	170	0.8	187	0.8
		0.8	27	149	0.9	184	0.9	200	0.9
		1	18						
		1.2	9						
		1.4	0						

Supplementary Table 4. Correlation points of nomogram prediction of recurrence.

Performance Status	Points	riskScore	Points	Total Points	1-year Survival Probability	Total Points	3-year Survival Probability	Total Points	5-year Survival Probability
0	52	-2	100	55	0.1	99	0.1	103	0.05
1	34	-1.5	90	75	0.2	118	0.2	119	0.1
2	17	-1	80	89	0.3	132	0.3	130	0.15
3	0	-0.5	70	101	0.4	144	0.4	139	0.2
M	**Points**	0	60	112	0.5	155	0.5	146	0.25
0	43	0.5	50	123	0.6	167	0.6	153	0.3
1	0	1	40	135	0.7	178	0.7	159	0.35
		1.5	30	149	0.8			165	0.4
		2	20	168	0.9			170	0.45
		2.5	10					176	0.5
		3	0					181	0.55
								187	0.6

Supplementary Table 5. Log rank test of 8-lncRNA-based classifier combined with performance status

Log Rank (Mantel-Cox)	group	low+Status0&1		low+Status2&3&4		high+Status0&1		high+Status2&3&4	
		chi-square	P values	chi-square	P values	chi-square	P values	chi-square	P values
Training cohort	low+Status0&1			11.3	0.001	15.511	0	34.829	0
	low+Status2&3&4	11.3	0.001			0.003	0.959	4.066	0.044
	high+Status0&1	15.511	0	0.003	0.959			4.208	0.04
	high+Status2&3&4	34.829	0	4.066	0.044	4.208	0.04		
Test cohort	low+Status0&1			19.896	0	2.102	0.147	65.568	0
	low+Status2&3&4	19.896	0			2.489	0.115	1.502	0.22
	high+Status0&1	2.102	0.147	2.489	0.115			12.862	0
	high+Status2&3&4	65.568	0	1.502	0.22	12.862	0		
TCGA cohort	low+Status0&1			30.199	0	18.938	0	108.295	0
	low+Status2&3&4	30.199	0			0.858	0.354	5.167	0.023
	high+Status0&1	18.938	0	0.858	0.354			14.022	0
	high+Status2&3&4	108.295	0	5.167	0.023	14.022	0		

Supplementary Table 6. Log rank test of 14-lncRNA-based classifier combined with performance status.

Log Rank (Mantel-Cox)	group	low+Status0&1		low+Status2&3		high+Status0&1		high+Status2&3	
		chi-square	P values	chi-square	P values	chi-square	P values	chi-square	P values
Training cohort	low+Status0&1			1.934	0.164	16.623	0	42.66	0
	low+Status2&3	1.934	0.164			0.079	0.778	2.609	0.106
	high+Status0&1	16.623	0	0.079	0.778			7.997	0.005
	high+Status2&3	42.66	0	2.609	0.106	7.997	0.005		
Test cohort	low+Status0&1			15.487	0	8.36	0.004	27.763	0
	low+Status2&3	15.487	0			2.423	0.12	0.334	0.563
	high+Status0&1	8.36	0.004	2.423	0.12			5.571	0.018
	high+Status2&3	27.763	0	0.334	0.563	5.571	0.018		
TCGA cohort	low+Status0&1			19.964	0	24.73	0	65.312	0
	low+Status2&3	19.964	0			1.395	0.238	3.5	0.061
	high+Status0&1	24.73	0	1.395	0.238			14.741	0
	high+Status2&3	65.312	0	3.5	0.061	14.741	0		

Please browse Full Text version to see the data of Supplementary Tables 7, 8.

Supplementary Table 7. Gene sets enriched in overall survival related classifier.

Supplementary Table 8. Gene sets enriched in recurrence related classifier.

Supplementary Table 9. Correlation between overall survival-classifier-related lncRNAs and recurrence-classifier-related lncRNAs.

lncRNA-based classifier for OS	lncRNA-based classifier for recurrence	cor	p value
AC090921.1	AC034229.4	0.208874156	0.000578559
AC090921.1	AC209154.1	0.134326063	0.027898715
AC090921.1	C10orf91	0.21696736	0.000346205
AC090921.1	LINC01549	-0.167598529	0.005952943
AC090921.1	MAFA-AS1	0.145372058	0.017247764
AC090921.1	MAFG-DT	0.180382687	0.003041263
AC096637.2	AC034229.4	0.285522048	2.02E-06
AC096637.2	C10orf91	0.230304099	0.000142449
AC096637.2	CDKN2A-DT	0.143301482	0.018919805
AC096637.2	LINC01572	0.12455441	0.041604634
AC096637.2	MAFA-AS1	0.131372909	0.031559886
AC096637.2	MAFG-DT	0.28949342	1.43E-06
AC096637.2	SNHG25	0.198976	0.001056771
AP002478.1	AC004477.1	0.259711142	1.66E-05
AP002478.1	AC010307.4	0.259539077	1.69E-05
AP002478.1	AC209154.1	0.244094535	5.38E-05
AP002478.1	C10orf91	0.140072512	0.021808497
AP002478.1	CDKN2A-DT	0.172834139	0.004545744
AP002478.1	CDKN2B-AS1	0.163611629	0.007274098
AP002478.1	FIRRE	0.146847125	0.016136627
AP002478.1	LINC01572	0.238742496	7.90E-05
AP002478.1	MAFG-DT	0.293178891	1.03E-06
C10orf91	AC004477.1	0.169090332	0.005516772
C10orf91	AC010307.4	0.223319151	0.000228295
C10orf91	AC034229.4	0.224312505	0.000213672
C10orf91	CDKN2B-AS1	0.130183888	0.033145673
C10orf91	FIRRE	0.170799934	0.005052355
C10orf91	LINC01549	-0.218043958	0.000322882
C10orf91	LINC01572	0.149601672	0.014228252
C10orf91	MAFA-AS1	0.169959418	0.005276101
C10orf91	MAFG-DT	0.241248479	6.61E-05
C10orf91	MIR9-3HG	0.189558585	0.001826946
LINC01116	C10orf91	0.271001376	6.79E-06
LINC01116	MAFG-DT	0.131794777	0.031013017
LINC01224	AC034229.4	0.29166297	1.18E-06
LINC01224	AC209154.1	0.168876724	0.005577414
LINC01224	C10orf91	0.241786227	6.36E-05
LINC01224	FIRRE	0.243081896	5.79E-05
LINC01224	LINC01572	0.149374266	0.014377938
LINC01224	MAFA-AS1	0.270744672	6.93E-06
LINC01224	MAFG-DT	0.158255714	0.009458505
LINC01224	MIR9-3HG	0.257762954	1.93E-05
LINC01224	SNHG25	0.145502969	0.017146522
MAFG-DT	AC004477.1	0.217947287	0.000324915
MAFG-DT	AC010307.4	0.249828932	3.53E-05
MAFG-DT	AC034229.4	0.235960325	9.62E-05

MAFG-DT	C10orf91	0.241248479	6.61E-05
MAFG-DT	CDKN2A-DT	0.149773026	0.014116366
MAFG-DT	CDKN2B-AS1	0.120368864	0.049013373
MAFG-DT	LINC01572	0.197606609	0.001146105
MAFG-DT	MAFA-AS1	0.170178478	0.005216945
MAFG-DT	MIR9-3HG	0.19080438	0.001701746
MAFG-DT	SNHG25	0.23188722	0.000127749
SERTAD4-AS1	AC034229.4	-0.156188541	0.010446167
SERTAD4-AS1	C10orf91	-0.145826459	0.016898567
SERTAD4-AS1	CDKN2B-AS1	-0.133468183	0.02892273
SERTAD4-AS1	LINC01549	0.259481918	1.69E-05
SERTAD4-AS1	MAFA-AS1	-0.268111694	8.57E-06
SERTAD4-AS1	SNHG25	-0.140105715	0.021776952

Please browse Full Text version to see the data of Supplementary Tables 10–12.

Supplementary Table 10. Correlations between risk score of the 8-lncRNA-based classifier with overall survival and clinicopathological characteristics in training cohort, test cohort, TCGA cohort and GEO cohort.

Supplementary Table 11. Correlations between risk score of the 14-lncRNA-based classifier with recurrence and clinicopathological characteristics in training cohort, test cohort, TCGA cohort and GEO cohort.

Supplementary Table 12. Univariate and multivariate COX analyses of the lncRNA-based classifier for OS.