Integrative Analysis and Identification of an Excellent lncRNA Signature to Predict Prognosis in Patients with COAD

ZhiHua Chen
The first Affiliated Hospital of Fujian Medical University
YiLin Lin
The first Affiliated Hospital of Fujian Medical University
SuYong Lin
The first Affiliated Hospital of Fujian Medical University
Ji Gao
Fujian Medical University
Shao-Qin Chen (chenshaoqin1613@163.com)
The first Affiliated Hospital of Fujian Medical University

Research Article

Keywords: Long noncoding RNA, Colon cancer, Integrative analyses, Prognostic signature, Mechanism

Posted Date: July 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-641736/v1

Abstract

Backgroud: Tumour recurrence and metastasis lead to poor prognosis incolon cancer(COAD). Therefore We aimed to identify a lncRNA signature through an integrative analysis of copy number variation, mutation and transcriptome data to predict prognosis and explore its internal mechanism.

Methods: The lncRNA expression profile were collected fromThe Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO). TCGA data was randomly divided 3:1 intotraining andtesting cohort. In the training, weperformed integrated analyses of three candidate lncRNA sets that correlated with prognosis, copy number variations and mutations to establish a signature through Cox regression analysis. The robustness was determined in the testing and GEO.

Results: An 11-lncRNA signature that was significantly associated with prognosiswas constructed in the training (P<0.0001, HR=2.014) , And this signature was validated in the testing(P=0.0019, HR=3.374) and GSE17536(P=0.0076, HR=1.864). The signature is significantly related to MSI status and clinical prognostic factors. The prognostic-relatedrisk scores were significantly excellent than the other five models have been reported. Furthermore, GSEA suggested that the signature was involved in COAD development and metastasis-related pathways.

Conclusions: We identifiedansignature has strong robustness and can stably predict the prognosis of COAD in different platformsand may be implicated in COAD pathogenesis and metastasis and applied clinically as a prognostic marker.

Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer[1]. As the treatment technology for CRC continues to advance, its mortality rate has declined for several decades. However, CRC is still the second most common cause of cancer-related death worldwide because of its high rates of tumour recurrence and metastasis, which cause a poor prognosis[2]. Therefore, it is important to find a molecular model that can effectively identify and predict the risk of CRC recurrence and metastasis. Zhou Yiming [3] identified and structured a prognostic signature based on five candidate genes, REG1B, TGM6, NTF4, PNMA5, and HOXC13, that could recognize COAD patients at a high risk of metastasis.

Long noncoding RNA (lncRNA) is a major type of noncoding RNA (ncRNA) defined as an RNA transcript of more than 200 base pairs in length. Recently, several studies have shown that the aberrant expression of lncRNAs is closely related to the development of many human tumours, including CRC[4–6]. After the lncRNA CCAT1 (CARLo-5) was identified as an oncogene in COAD[7], a number of lncRNAs that are dysregulated in COAD have been identified[8], and the role of lncRNAs has recently received increasing attention. lncRNAs are characterized as oncogenes or tumour suppressor genes [9, 10] and play a role in different biological processes in COAD. lncRNAs participate in transcriptional and epigenetic regulation by interacting with genomic DNA, transcription factors, chromatin, spliceosomes, chromatin regulators and other nuclear proteins[11], and lncRNAs are always involved in posttranscriptional, translational and posttranslational regulatory processes[12]. Various lncRNAs are involved in carcinogenesis and progression by regulating COAD cell proliferation, migration and invasion[7, 13].

High-through put multi-omics sequencing data have laid a solid foundation for identifying genes associated with cancer prognosis. Multi-omics data analysis can reveal the mechanism of cancer development from multiple perspectives. To improve the predictive value and accuracy for COAD prognosis, we need to identify a robust lncRNA signature through an integrative analysis of prognosis-related lncRNA, copy number variation, mutation and transcriptome data with the help of multi-omics data analysis technology. In this study, We collected the data of the copy number variation, mutation and transcriptome of COAD tissues from The Cancer Genome Atlas (TCGA, n = 478) and the GEO(GSE17536 dataset, n = 177). A signature was established through integrated analyses, and validated in the diferent platforms of the testing cohort and the GSE17536 cohort. The microsatellite instability (MSI) status and clinical independence of the lncRNA signature were analysed. We also explored the pathways associated with the development and metastasis of COAD enriched in the

lncRNA signature (Fig. 1). Furthermore, Compared with other rognostic-related risk signatures, the signature was better of predicting prognosis in COAD patients.

Results

1 Comprehensive analysis of multi-omics data to obtain lncRNAs related to COAD prognosis

1.1 Lncrnas That Are Closely Related To Coad Prognosis

According to the univariate Cox regression analyse, we identified a total of 483 candidate prognostic lncRNAs from the training cohort, and information on the top 20 lncRNAs is shown in Table 2.

Table 2
Information on the top 20 candidate prognostic lncRNAs

lncRNA ID	*P*-value	HR	Low 95% CI	High 95% CI
ENSG00000274925	5.78E-06	1.156	1.086	1.231
ENSG00000272512	1.29E-05	1.219	1.115	1.333
ENSG00000275494	3.19E-05	1.051	1.027	1.077
ENSG00000258053	3.59E-05	1.114	1.059	1.173
ENSG00000260563	6.04E-05	1.086	1.043	1.131
ENSG00000245281	8.09E-05	0.362	0.219	0.600
ENSG00000247095	0.000138199	1.035	1.017	1.053
ENSG00000235245	0.000149251	1.104	1.049	1.162
ENSG00000272555	0.000187134	0.437	0.283	0.675
ENSG00000229380	0.000188209	1.309	1.136	1.507
ENSG00000273456	0.000197802	1.159	1.073	1.253
ENSG00000279148	0.000207384	0.422	0.268	0.666
ENSG00000228288	0.000246995	1.077	1.035	1.120
ENSG00000269680	0.000250415	1.253	1.110	1.413
ENSG00000227947	0.000294015	0.591	0.445	0.786
ENSG00000271781	0.000314732	1.068	1.031	1.108
ENSG00000238113	0.000381295	1.301	1.125	1.505
ENSG00000230641	0.000382509	0.476	0.316	0.717
ENSG00000226659	0.000402587	1.176	1.075	1.287
ENSG00000251637	0.000431394	0.495	0.335	0.732

1.2 lncRNAs that are closely related to gene copy number variation

We obtained lncRNAs that are closely related to gene copy number variation. A total of 137 lncRNAs that were significantly amplified on each fragment of the COAD genome (Fig. 2A), including LINC00392 on the 13q22.1 segment (q = 8.17E-12),

LINC01598 on the 20q11.21 segment (q = 5.75E-09) and LOC730183 on the 16p11.2 segment (q = 0.0014485), were identified. On the other hand, a total of 261 lncRNAs that were significantly deleted on each fragment in the COAD genome (Fig. 2B), including LINC00681 in the 8p22 segment (q = 2.74E-45), LOC101928728 in the 1p36.11 segment (q = 2.02E-09), and LINC00491 in the 5q22.2 segment (q = 1.48E-05), were identified.

1.3 lncRNAs that are closely related to gene mutations

Through MutSig2, we identified a total of 41 genes with significant mutation frequencies. The distribution of synonymous mutations, missense mutations, framework insertions or deletions, framework movements, nonsense mutations, cleavage sites and other nonsynonymous mutations in these 41 genes in TCGA COAD patient samples are shown in Fig. 3. We identified 41 genes, some of which have been reported to be closely related to the development of cancer, such as KRAS, TP53, APC, PIK3CA, and FBXW7. Among these 41 genes, we identified lncRNAs associated with gene mutations using each of the genes to mutate into a tag, and a rank-sum test was used to detect the difference between the expression of each lncRNA in the mutant and nonmutant groups. lncRNAs with a *P*-value < 0.01 were considered to be associated with a gene mutation; thus, we obtained 2712 lncRNAs related to gene mutations.

2 Identification Of An 11-lncrna Signature For Coad Survival

The comprehensive analysis revealed 147 lncRNAs associated with amplifications, deletions, and mutations from a total of 483 candidate prognostic lncRNAs. We analysed the change trajectory of each independent variable (Fig. 4A). It can be seen that with the gradual increase in lambda, the number of independent coefficients becomes closer to zero. We used three-fold cross-validation to build the model. The confidence interval under each lambda is shown in Fig. 4B. As shown in the figure, the model was optimal when lambda = 0.04078231. For this reason, we selected the lncRNAs obtained when lambda = 0.04078231 as the target lncRNAs to construct the model.

After Lasso Cox regression narrowed the scope, we obtained 21 target lncRNAs that were used to construct the model. A multivariate Cox survival analysis was performed on 21 lncRNAs, and the 11 lncRNAs with the lowest AIC values (AIC = 767.27) were used to construct the final model. Details of the 11 lncRNAs are shown in Table 3. The 11-lncRNA signature was then tested for its ability to predict survival in COAD patients.

Table 3
Eleven lncRNAs identified as significantly associated with OS in the training cohort

Ensembl Gene ID	Symbol	Coef	HR	Z-score	*P*-value	Low 95% CI	High 95% CI
ENSG00000269680	AC008760.1	2.2664	9.645	3.516	0.000438	2.727	34.115
ENSG00000215039	CD27-AS1	0.3273	1.387	2.03	0.042341	1.011	1.903
ENSG00000249550	LINC01234	0.4492	1.567	2.418	0.015587	1.089	2.255
ENSG00000247095	MIR210HG	0.3796	1.462	3.788	0.000152	1.201	1.779
ENSG00000180139	ACTA2-AS1	0.9405	2.561	3.111	0.001866	1.416	4.632
ENSG00000256546	AC156455.1	0.6567	1.928	2.736	0.006223	1.205	3.087
ENSG00000260805	AC092803.2	1.5048	4.503	2.02	0.043372	1.046	19.391
ENSG00000273576	AC009283.1	0.1432	1.154	2.15	0.031533	1.013	1.315
ENSG00000246627	CACNA1C-AS1	2.7294	15.323	4.146	3.39E-05	4.216	55.69
ENSG00000238113	LINC01410	2.0232	7.563	3.026	0.00248	2.039	28.045
ENSG00000235560	AC002310.1	1.3988	4.05	2.372	0.017675	1.275	12.863

3 Determination and analysis of the 11-lncRNA signature in the training cohort

The 11-lncRNA signature was then established using a multivariate Cox regression analysis with the following model:

$$RiskScore_{11} = 2.2664*exp^{AC008760.1}+0.3273*exp^{CD27-AS1}+0.4492*exp^{LINC01234}$$

$$+ 0.3796*exp^{MIR210HG}+0.9405*exp^{ACTA2-AS1}+0.6567*exp^{AC156455.1}$$

$$+ 1.5048*exp^{AC092803.2}+0.1432*exp^{AC009283.1}+2.7294*exp^{CACNA1C-AS1}$$

$$+ 2.0232*exp^{LINC01410}+1.3988*exp^{AC002310.1}$$

The risk score was calculated as the sum of the above gene expression values * the ordinal, and then we selected 0.9892846 as the cutoff (median risk score) and divided the samples into high-risk and low-risk groups. Finally, 247 patients were classified as low risk, and 110 patients were classified as high risk; significantly different OS rates were observed between the two groups in the training cohort (log-rank $P < 0.0001$, HR = 2.014) Fig. 5C). We acquired a five-year AUC of 0.83 according to the ROC curve for predicting survival in COAD patients (Fig. 5B). As the patient's risk score increased, the OS rate significantly decreased, and the number of deaths in the high-risk group increased significantly (Fig. 5A). According to the changes in the expression levels of the 11 different lncRNAs in the signature observed with increases in the risk score, the expression of ENSG00000246627 was correlated with a low risk of COAD, and the other 10 lncRNAs were identified as risk factors based on their high expression and correlation with a high risk of COAD.

4 Validation of the 11-lncRNA signature in the testing and GSE17536 cohorts

First, the 11-lncRNA signature was validated in the testing cohort; 88 patients were classified as low risk, and 31 patients were classified as high risk. There was a significant difference in OS between the two groups (log-rank $P = 0.0019$, HR = 3.374) (Fig. 6C). The five-year AUC was 0.66 according to the ROC curve (Fig. 6B). The results of the testing cohort were similar to those of the training cohort; as the patient's risk score increased, the OS time decreased significantly, and the number of deaths in the high-risk group increased significantly (Fig. 6A). Moreover, 10 lncRNAs (all lncRNAs except ENSG00000246627) were identified as risk factors.

Similarly, we used the same model and the same cut-off from the training cohort and verified the model's robustness using an external independent data cohort (GSE17536). Ultimately, 99 patients were classified as low risk, and 78 patients were classified as high risk. A significant difference in OS was observed between the two groups (log-rank $P = 0.0076$, HR = 1.864) (Fig. 7C). The five-year AUC was 0.71 according to the ROC curve (Fig. 7B). The GSE17536 cohort showed similar results to the TCGA training cohort. As the risk score increased, the survival time decreased significantly, and the number of deaths in the high-risk group increased. The expression of the 11 different signature lncRNAs also increased with the increase in the risk score, indicating that high expression of the 11 lncRNAs is associated with a high risk of COAD and could serve as a risk factor (Fig. 7A).

5 Independent predictive power of the 11-lncRNA signature according to the MSI status, tumour stage and clinicopathological characteristics

The patients were subdivided into a high-risk subgroup and a low-risk subgroup based on different MSI statues, and the 11-lncRNA signature was used to predict OS; the OS rate was significantly different between the high-risk and low-risk subgroups in the MSI-L and MSS groups (excluding MSH) (Fig. 8A-C). Based on their tumour stage, patients were subdivided into a high-risk group and a low-risk group in each stage, and the 11-lncRNA signature revealed no significant difference in OS at the II, III and IV stages (all stages except stage I) between the two groups (Fig. 8D-G). Furthermore, these results demonstrate that the 11-lncRNA signature model can better predict the OS of patients with different MSI statuses and tumour stages.

We systematically analysed the clinical information of the TCGA and GSE17536 patients, including age, sex, lymph node invasion status, pathology (T, N, and M classifications), tumour stage, and the 11-lncRNA signature grouping information (Table 4).

Table 4
Identification of the clinical factors and clinical independence associated with prognosis with univariate and multivariate Cox regression analyses

Variables	Univariate analysis			Multivariable analysis		
	HR	95% CI of HR	P-value	HR	95% CI of HR	P-value
TCGA training dataset						
11-lncRNA risk score						
Risk score (High/Low)	4.96	3.14–7.82	5.14E-12	4.39	2.58–7.46	4.26E-08
Age	1.02	0.99–1.04	0.07	1.03	1.01–1.05	0.002
Sex(Male/Female)	1.14	0.72–1.82	0.57	0.97	0.60–1.58	0.921
T3/T4 vs T1/T2	5.77	1.81–18.37	0.002	3.59	0.84-15.219	0.082
N1/N2 vs N0	3.05	1.85–5.01	1.04E-05	0.31	0.091–1.07	0.065
M1 vs M_0	2.58	1.61–4.16	8.72E-05	1.48	0.88–2.49	0.136
Stage □/□ vs Stage □/□	3.37	2.00-5.67	4.94E-06	7.56	1.94–29.47	0.003
TCGA validation dataset						
11-lncRNA risk score						
Risk score (High/Low)	3.06	1.36–6.89	0.0066	3.02	1.13–8.08	0.027
Age	1.02	0.98–1.05	0.28	1.04	0.99–1.08	0.083
Sex (Male/Female)	0.93	0.42–2.05	0.86	0.55	0.21–1.44	0.226
T3/T4 vs T1/T2	0.99	0.29–3.39	0.99	0.76	0.13–4.34	0.756
N1/N2 vs N0	2.52	1.13–5.58	0.02	0.54	0.059–4.87	0.583
M1 vs M_0	8.03	3.31–19.51	4.17E-06	5.85	1.88–18.15	0.002
Stage □/□ vs Stage □/□	2.86	1.24–6.56	0.01	5.05	0.41–61.95	0.205
GSE17536 validation dataset						
11-lncRNA risk score						
Risk score (High/Low)	1.86	1.17–2.97	0.0085	1.65	1.02–2.67	0.039
Age	1.006	0.98–1.02	0.49	1.02	1.002–1.04	0.029
Sex (Male/Female)	1.104	0.69–1.76	0.67	1.17	0.71–1.91	0.521
Stage □/□ vs Stage □/□	4.22	2.39–7.46	7.28E-07	4.226	2.36–7.56	1.21E-06

In the TCGA training cohort, we found significant survival-related correlations in the clinicopathological characteristics, with the exception of age and sex, according to the univariate Cox regression analysis, but we found that only the risk score (HR = 4.39, 95% CI = 2.58–7.46, *P* = 4.26E-08), age, and stage III/IV vs sage I/II were significantly related to survival according to the multivariate Cox regression analysis. The 11-lncRNA signature was verified to be clinically independent.

In the TCGA testing cohort, the risk score, N1/N2 vs N0, M1 vs M0, and stage III/IV vs stage I/II were significantly associated with survival according to the univariate Cox regression analysis, but only the risk score (HR = 3.02, 95% CI =

1.13–8.08, *P* = 0.027) and M1 vs M0 were significantly related to survival according to the multivariate Cox regression analysis. The 11-lncRNA signature was also verified to be clinically independent.

In the GSE17536 cohort, the risk score and stage III/IV vs stage I/II were significantly associated with survival according to the univariate Cox regression analysis, but only the risk score (HR = 1.65, 95% CI = 1.02–2.67, *P* = 0.039), age and stage III/IV vs stage I/II were significantly associated with survival according to the multivariate Cox regression analysis.

In conclusion, the 11-lncRNA signature is a prognostic indicator independent of other clinical factors and shows independent predictive performance with clinical application value.

6 Identification of the 11-lncRNA signature-associated biological pathways in the training cohort

The signalling pathways associated with the 11-lncRNA signature significantly enriched in the TCGA training cohort were detected by GSEA (Table 5). The signalling pathways that were significantly enriched in the high-risk and low-risk groups, were the Notch signalling pathway, the VEGF signalling pathway, the P53 signalling pathway and the cell cycle; all were significantly associated with the development and metastasis of COAD (Fig. 9).

Table 5
KEGG pathways significantly enriched in the high-risk and low-risk groups detected by GSEA

NAME	SIZE	ES	NES	NOM P-val	FDR q-val	FWER P-val
KEGG_NOTCH_SIGNALING_PATHWAY	47	-0.468	-1.633	0.028	1.000	0.726
KEGG_VEGF_SIGNALING_PATHWAY	75	-0.354	-1.486	0.048	1.000	0.916
KEGG_P53_SIGNALING_PATHWAY	67	0.434	1.660	0.015	0.936	0.632
KEGG_ALANINE_ASPARTATE_AND_GLUTAMATE_METABOLISM	32	0.464	1.601	0.025	0.552	0.752
KEGG_RIBOSOME	87	0.757	1.773	0.031	0.747	0.387
KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	133	0.418	1.658	0.035	0.632	0.638
KEGG_BASAL_TRANSCRIPTION_FACTORS	35	0.490	1.575	0.043	0.466	0.801
KEGG_CELL_CYCLE	124	0.459	1.609	0.050	0.657	0.745

7 Comparison of the 11-lncRNA signature with other COAD prognostic signatures

The ROC and OS KM curves of the five models are shown in Fig. 10. Significantly different OS rates were observed between the high-risk and low-risk groups using the six-lncRNA signature established by Zhao[17] (log-rank *P* = 0.0014,HR = 2.03) (Fig. 10B). We acquired a five-year AUC of 0.65 and a ten-year AUC of 0.67 according to the ROC (Fig. 10A). Significantly different OS rates were also observed between the two groups using the two-lncRNA signature established by Xue[18] (log-rank *P* = 0.018, HR = 1.73) (Fig. 10D), and we obtained a five-year AUC of 0.54 and a ten-year AUC of 0.47 (Fig. 10C). Significantly different OS rates were also observed with the 14-lncRNA signature reported by Xing[19](Fig. 10F), and we found a five-year AUC of 0.66 and a ten-year AUC of 0.53 (Fig. 10E). In addition, the six-lncRNA signature established by Fan[20](Fig. 10H) yielded a five-year AUC of 0.64 and a ten-year AUC of 0.41(Fig. 10G), and the 15-lncRNA signature obtained by Wang[21] (Fig. 10J) resulted in a five-year AUC of 0.78 and a ten-year AUC of 0.67 (Fig. 10I). The final comparison showed that our model was slightly better than the 15-lncRNA model and significantly better than the other four models.

Discussion

CRC is a common digestive tract tumour that is a serious threat to the health of patients. According to recent statistics, there are approximately 1.45 million new cases of CRC each year, which made it the third most prevalent cancer in 2018[1] and the second most prevalent cancer in American males in 2019[22]. Approximately 694,000 deaths have been reported every year, making the second most common cause of cancer-related death worldwide[2]. The recurrence and metastasis of CRC seriously affect the efficacy and prognosis of treatment. Identifying the risk of recurrence and metastasis can help us guide early intervention for the treatment of CRC, ultimately improving the prognosis. Therefore, it is very urgent to study and identify one or more efficient molecular models for predicting prognosis to guide treatment options and to improve the survival quality of CRC patients.

lncRNAs are a major class of ncRNAs with a length of more than 200 base pairs and are not translated into proteins[23, 24]. Over the past decade, it has become clear that certain lncRNAs have strong potential, and an in-depth study is needed to elucidate their mechanism of action. lncRNAs not only control the nuclear structure[25] but also regulate the expression of adjacent genes and act as amplifiers, with remarkable tissue specificity through various mechanisms[26]. lncRNAs have been shown to directly regulate gene expression at the transcriptional, posttranscriptional and epigenetic levels. lncRNAs have long nucleotide chains and intricate secondary structures and can interact with genomic DNA, chromatin, transcription factors, chromatin regulators, spliceosomes and other nuclear proteins[27]. lncRNAs are known to serve as tumour regulators and participate in complex networks of biological regulation[28, 29]. Therefore, the identification of lncRNAs closely related to tumour prognosis and the establishment of a signature that can predict the risk of tumour prognosis will be helpful for improving the prevention and treatment of tumours. Zhang GH[15] identified a novel four-lncRNA signature using Cox regression analysis to identify lncRNAs that correlated with the prognosis of 111 laryngeal cancer patients from the TCGA. The signature was shown to predict the prognosis of patients with laryngeal cancer and may influence the prognosis of laryngeal cancer through many pathways, such as regulating immunity and tumour apoptosis. Jie Li[30] also identified a five-lncRNA signature to predict the risk of tumour recurrence in breast cancer (BC) patients and found that it was independent of clinical prognostic factors, such as BC subtypes and adjuvant treatments.

Recently, an increasing number of researchers have focused on the role of lncRNAs in the development and progression of CRC and its significance to clinical prognosis[4, 6]. Many lncRNAs, such as MALAT1 and HOTAIR, have also been used as biomarkers in CRC[31, 32]. Several lncRNAs have not only been reported as markers in CRC diagnosis but also been shown to be correlated with patient prognosis (e.g., CCAL, PURPL and lnc-GNAT1-1)[33–35]. Therefore, this method was also used to identify a CRC-related lncRNA signature for predicting the prognosis of CRC[17–21] by analysing different datasets.

At the same time, as high-throughput multi-omics sequencing data have laid a solid foundation for identifying genes associated with cancer prognosis, multi-omics data analysis can reveal the mechanisms of cancer development from multiple perspectives[36]. We identified closely related genomics, gene mutations, epigenetics, and functions of lncRNAs that are inherently regulated by COAD[4, 6]. We show that dysregulated lncRNAs may be new prognostic and diagnostic biomarkers or therapeutic targets for clinical applications. Therefore, we identified a robust lncRNA signature through an integrative analysis of prognosis-related lncRNA, copy number variation, mutation and transcriptome data with the help of multi-omics data analysis technology. In this study, we fully integrated and analysed lncRNAs that are related not only to prognosis but also to copy number variations, mutations and transcriptome regulation in 359 COAD samples from the TCGA training cohort. We developed an 11-lncRNA signature that was validated in testing and GSE17536 cohorts. We also found that the signature had good robustness and good predictive ability of the prognosis risk in other independent datasets. Moreover, we compared the 11-lncRNA signature with other COAD prognostic signatures[17–21] to validate the good prediction of the prognosis risk of the 11-lncRNA signature. We found that the AUC of the 11-lncRNA signature was better than that of the other five signatures. Because the researchers did not confirm that the lncRNAs directly regulate gene expression at the transcriptional, post-transcriptional and epigenetic levels, we focused on the predictive role of lncRNAs in tumour prognosis and established lncRNA signatures to predict prognosis merely by analysing and identifying

lncRNAs associated with prognosis. However, the intrinsic interactive relationship between lncRNAs, genomics, gene mutations and epigenetics was not investigated. Ultimately, the studies failed to identify a good lncRNA signature to predict the risk of CRC prognosis. Significant differences in OS outcomes between the high-risk and low-risk groups were obtained using the described method[37].

Furthermore, we found that the 11-lncRNA signature had independent predictive power from the MSI status and tumour stage (e.g., MSI-L and MSS patients and patients in stages II, III and IV, excluding stage I), which may be due to a good survival rate in COAD patients with stage I [22]. This study also showed that the 11-lncRNA signature was a prognostic indicator independent of other clinical factors and had independent predictive performance with clinical application value. Additionally, we identified the 11-lncRNA signature-associated biological pathways that were significantly enriched in COAD patients as detected by GSEA. In summary, we determined that the Notch signalling pathway, the VEGF signalling pathway, the P53 signalling pathway and the cell cycle are significantly associated with the development and metastasis of COAD[38, 39].

In summary, we constructed a novel 11-lncRNA signature that can be used to predict the prognosis of patients with COAD and exploited the possible underlying mechanisms involved. The 11-lncRNA signature may indicate the potential roles of lncRNAs in COAD pathogenesis. The results will provide molecular diagnostic markers and therapeutic targets with clinical implications in COAD patients. Finally, we hope that all of the above results will be verified in basic experiments and clinical trials in further studies.

Conclusion

In our study, we integrated analysis of the lncRNAs were not only related to the prognosis, but also related to copy number variation, mutation and transcriptome data to identify and constructe a lncRNA signature to predict prognosis in COAD patients. And the signature was validated in the testing and GSE17536 cohorts and was independent of the MSI status and clinical prognostic factors. The signature was significantly better than the other five lncRNA signatures have been reported in the COAD.

Materials And Methods

1 Patients and data collection

We downloaded COAD RNA-Seq data, clinical follow-up information and copy number variation data from the SNP 6.0 chip in the TCGA database from the UCSC cancer browser (https://xenabrowser.net/datapages/), and we downloaded the mutation comment file (MAF) from the GDC client. We also downloaded the GSE17536 dataset, which included COAD expression profile data and clinical information, from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/).

We preprocessed the downloaded data, downloaded the fragments per kilobase of transcript per million mapped reads (FPKM) RNA-Seq data from the TCGA. The R package "caret" was used to randomly divide the samples into the training cohort (359 samples) and the testing cohort (119 samples). The lncRNA expression profile data is extracted according to the Ensemble ID of lncRNA in the GENCODE database. SeqMap was used to compare GSE17536 expression data (no mismatch was allowed). A total of 5,076 probes were annotated on the lncRNAs. All selected expression datasets in the training, testing and GSE17536 cohorts were $\log_2$ transformed for standardization.

Ultimately, a total of 478 samples from the TCGA were randomly divided at a ratio of 3:1 into a training cohort (n = 359) and a testing cohort (n = 119), and 177 samples were obtained from the GEO database (GSE17536). We obtained clinical pathology data, including the patient age, survival status, sex, lymph node metastasis status, T classification, N

classification, M classification and tumour stage (Table 1), from the three cohorts and found no significant differences among the groups.

Table 1
Clinical pathology data of the three cohorts

Characteristic		TCGA training cohort (n = 359)	TCGA testing cohort (n = 119)	*P*-value	GSE17536 (n = 177)
Age (years)	≤ 60	103	31	0.724	59
	> 60	225	76		118
Survival Status	Living	276	91	0.949	104
	Dead	81	28		73
Gender	female	157	45	0.349	81
	male	171	62		96
T	T 1	7	4	0.188	--
	T 2	63	12		--
	T 3	214	84		--
	T 4	43	7		--
N	N 0	184	71	0.172	--
	N 1	82	20		--
	N 2	62	16		--
M	M 0	234	87	0.027	--
	M 1	90	17		--
Tumour stage	Stage	60	13	0.131	24
	Stage	114	53		57
	Stage	95	29		57
	Stage	51	9		39

2 Identification of lncRNAs that are closely related to the prognosis of COAD

We performed a univariate Cox regression analysis to establish the correlation of lncRNA expression with overall survival (OS) in the TCGA training cohort, and the lncRNAs with significant P values ($P < 0.01$) were selected as candidates.

3 Identification of lncRNAs that are closely related to gene copy number variation

We used GISTIC 2.0 to identify genes with significant amplification or deletion from copy number variation data in the TCGA training cohort. We set a parameter threshold for fragments with amplification or deletion lengths greater than 0.1 and significant *P*-values ($P < 0.05$). The significantly amplified fragments in the genome and the genes amplified on each of the fragments were recorded and the significantly deleted fragments in the genome and the genes that were

significantly deleted on each fragment were recorded and incorporated to establish the lncRNAs associated with copy number variation.

4 Identification of lncRNAs that are closely related to gene mutations

We used MutSig2 to identify genes with significant mutations from the mutation annotation data of the TCGA training cohort, and with a threshold of $P < 0.05$. We identified lncRNAs associated with gene mutations using the rank-sum test to detect the difference in the expression of each lncRNA between the mutant and nonmutant groups, and each gene mutation was used as a label. lncRNAs with significant P values ($P < 0.01$) were considered to be associated with a gene mutation. Finally, the lncRNA dataset related to gene mutations was established.

5 Development and validation of the robustness of the lncRNA signature

By intersecting these three lncRNA sets (Prognosis-related lncRNAs, copy number variation-related lncRNAs and mutation-related lncRNAs), we obtained the targeted lncRNAs. The least absolute shrinkage and selection operator (Lasso) method, which is a compression estimate, was used to narrow the lncRNA range. We used the R package glmnet for the Lasso Cox regression analysis[14, 15]. Furthermore, we performed a multivariate Cox regression analysis, and stepwise regression was used to reduce the number of lncRNAs again. The lowest Akaike information criterion (AIC) value as the final model.

$$\text{RiskScore} = \text{coefficient}*\exp^{\text{lncRNA1}} + \text{oefficient}*\exp^{\text{lncRNA2}} + \text{oefficient}*\exp^{\text{lncRNA3}} + \ldots + \text{coefficient}*\exp^{\text{lncRNAn}}.$$

The risk score of each sample in the TCGA training cohort was then calculated. Based on the median risk score, the COAD patients were divided into two groups: a high-risk group and a low-risk group. A receiver operating characteristic (ROC) curve was used to test the accuracy of lncRNA signature to predict prognosis. Finally, the lncRNA signature was verified in the testing cohort. The GSE17536 cohort used the same model and the same cutoff as the TCGA training cohort.

6 Independent predictive power of the lncRNA signature based on different MSI statuses, tumour stages and clinicopathological characteristics

We divided the TCGA cohort into MSI-high (MSI-H), MSI-low (MSI-L) and microsatellite stable (MSS) groups according to the MSI phenotype information described by the TCGA network study[16]. We further analyze the relationship between this lncRNA signature and MIS status.

To identify the lncRNA signature model for clinical applications, we analysed the relationship between clinical information (including age, sex, lymph node invasion status, pathology (T, N, and M classifications), tumour stage) and lncRNA signature through univariate and multivariate Cox regression analyses.

7 Identification of the lncRNA signature-associated biological pathways with gene set enrichment analysis (GSEA)

We used GSEA and the “cluster profile R” package to conduct Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis. Significantly enriched pathways in the high-risk and low-risk groups in the TCGA training cohort were identified. The selected gene set was c2.cp.kegg.v6.0.symbols, which contained the KEGG pathways. In KEGG pathway analysis, $P < 0.05$ was considered to indicate statistical significance.

8 Comparison of the lncRNA signature with other COAD prognostic signatures

By reviewing the literature, we identified five prognostic-related risk models, a six-lncRNA signature (PMID: 30396175)[17], a two-lncRNA signature (PMID: 29254165)[18], a 14-lncRNA prognostic signature (PMID: 29565464)[19], a six-lncRNA signature (PMID: 29227531)[20] and a 15-lncRNA signature (PMID: 30510449)[21], for comparison with our lncRNA signature. To make the models comparable, we performed a multivariate Cox regression analysis to calculate the risk scores of the training set samples based on the corresponding genes in the three models. We evaluated the ROC of the five

models and then divided the samples into high-risk and low-risk groups according to the median risk score and analysed the difference in OS between the two groups.

Declarations

Acknowledgements

The authors are grateful to Ruiqing Chen and Lengxi Fu for technical assistance.

Disclosure statement

No potential conflict of interest was reported by the authors.

Authors' contributions

ZC, YL and SL conceived this experiment; ZC, YL and JG performed the experiments and dataanalysis; ZC, YL, JG and SC wrote and revisedthe manuscript. All authors approved the final version of themanuscript.

Funding

This work was supported by grants from the Fujian health youth research project (NO. 2019-2-20), the Fujian natural fund project (NO. 2019J01448) and the Fujian science and technology innovation joint fund Project (NO. 2019Y9133).

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author.

Ethics approval and consent to participate

The research used the published database data to conduct secondary mining research, and the research does not involve ethical approval.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Siegel Rebecca L,Miller Kimberly D,Jemal Ahmedin. Cancer statistics. 2019.CA Cancer J Clin. 2019; 69:7-34.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394-424.
3. Zhou Yiming,Zang Yiwen,Yang Yi, Xiang J, Chen Z. Candidate genes involved in metastasis of colon cancer identified by integrated analysis.Cancer Med. 2019;8: 2338-2347.
4. Kim T, Croce CM. Long noncoding RNAs: Undeciphered cellular codes encrypting keys of colorectal cancer pathogenesis. Cancer Lett. 2018;417:89-95.
5. Tang X, Qiao X, Chen C, Liu Y, Zhu J, Liu J. Regulation Mechanism of Long Noncoding RNAs in colon cancer Development and Progression. Yonsei Med J. 2019;60: 319-325.

6. Sun Z, Liu J, Chen C, Zhou Q, Yang S, Wang G, Song J, Li Z, Zhang Z, Xu J, Sun X. The Biological Effect and Clinical Application of Long Noncoding RNAs in Colorectal Cancer. Cell Physiol Biochem. 2018; 46: 431-441.
7. Nissan A, Stojadinovic A, Mitrani-Rosenbaum S, Halle D, Grinbaum R, Roistacher M, Bochem A. Colon cancer associated transcript-1: a novel RNA expressed in malignant and pre-malignant human tissues. Int J Cancer 2012;130:1598-606.
8. Han D, Wang M, Ma N, Xu Y, Jiang Y, Gao X. Long noncoding RNAs: novel players in colorectal cancer. Cancer Lett. 2015;361:13-21.
9. Xie X, Tang B, Xiao YF, Xie R, Li BS, Dong H, Zhou JY, Yang SM. Long noncoding RNAs in colorectal cancer. Oncotarget. 2016;7:5226-5239.
10. Shen P, Pichler M, Chen M, Calin GA, Ling H. To Wnt or Lose: The Missing NonCoding Linc in Colorectal Cancer. Int J Mol Sci. 2017; 18: e2003.
11. Schmitt AM, Chang HY. Chang, Long Noncoding RNAs: At the Intersection of Cancer and Chromatin Biology. Cold Spring Harb Perspect Med.2017;7:a026492.
12. Schmitt AM, Chang HY. Long Noncoding RNAs in Cancer Pathways. Cancer Cell. 2016;29:452-463.
13. Sun J, Ding C, Yang Z, Liu T, Zhang X, Zhao C, Wang J. The long noncoding RNA TUG1 indicates a poor prognosis for colorectal cancer and promotes metastasis by affecting epithelial-mesenchymal transition. J Transl Med 2016;14:42.
14. Meng J, Li P, Zhang Q, Yang Z, Fu S. A four-long non-coding RNA signature in predicting breast cancer survival.J Exp Clin Cancer Res. 2014; 33: 84.
15. Zhang G, Fan E, Zhong Q, Feng G, Shuai Y, Wu M, Chen Q, Gou X. Identification and potential mechanisms of a 4-lncRNA signature that predicts prognosis in patients with laryngeal cancer.Human genomics. 2019;13: 36.
16. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487: 330-7.
17. Zhao J, Xu J, Shang AQ, Zhang R. A Six-LncRNA Expression Signature Associated with Prognosis of Colorectal Cancer Patients.[J] .Cell. Physiol. Biochem., 2018, 50: 1882-1890.
18. Xue W, Li J, Wang F, Han P, Liu Y, Cui B. A long non-coding RNA expression signature to predict survival of patients with colon adenocarcinoma. Oncotarget. 2017; 8: 101298-101308.
19. Xing Y, Zhao Z, Zhu Y, Zhao L, Zhu A, Piao D. Comprehensive analysis of differential expression profiles of mRNAs and lncRNAs and identification of a 14-lncRNA prognostic signature for patients with colon adenocarcinoma. Oncol Rep. 2018;39: 2365-2375.
20. Fan Q,Liu B. Discovery of a novel six-long non-coding RNA signature predicting survival of colorectal cancer patients.J Cell Biochem. 2018;119: 3574-3585.
21. Wang X, Zhou J, Xu M, Yan Y, Huang L, Kuang Y, Liu Y, Li P, Zheng W, Liu H, Jia B. A 15-lncRNA signature predicts survival and functions as a ceRNA in patients with colorectal cancer.Cancer Manag Res. 2018;10: 5799-5806.
22. Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, Jemal A, Kramer JL. Cancer treatment and survivorship statistics, 2019.CA Cancer J Clin. 2019;69:363-385.
23. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145-66.
24. Cao J. The functional role of long non-coding RNAs and epigenetics. Biol Proced Online. 2014;16:11.
25. 25. Engreitz JM,Ollikainen N,Guttman M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression.Nat Rev Mol Cell Biol.2016;17:756-770.
26. Ransohoff JD,Wei Y,Khavari PA. The functions and unique features of long intergenic non-coding RNA.Nat Rev Mol Cell Biol.2018;19:143-157.
27. Kopp F,Mendell JT. Functional Classification and Experimental Dissection of Long Noncoding RNAs.Cell. 2018;172:393-407.

28. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. Towards a complete map of the human long non-coding RNA transcriptome.Nat RevGenet.2018;19:535-548.
29. Zhang H, Chen Z, Wang X, Huang Z, He Z, Chen Y. Long non-coding RNA: a new player in cancer. J Hematol Oncol. 2013;6:37.
30. Li J, Wang W, Xia P, Wan L, Zhang L, Yu L, Wang L, Chen X, Xiao Y, Xu C. Identification of a five-lncRNA signature for predicting the risk of tumor recurrence in patients with breast cancer.International journal of cancer.2018.143:2150-2160.
31. Wu S, Sun H, Wang Y, Yang X, Meng Q, Yang H, Zhu H, Tang W, Li X, Aschner M, Chen R. MALAT1 rs664589 polymorphism inhibits binding to miR-194-5p contributing to colorectal cancer risk, growth and metastasis.Cancer Res.2019;7.
32. Pan S, Liu Y, Liu Q, Xiao Y, Liu B, Ren X, Qi X, Zhou H, Zeng C, Jia L. HOTAIR/miR-326/FUT6 axis facilitates colorectal cancer progression through regulating fucosylation of CD44 via PI3K/AKT/mTOR pathway.Biochim Biophys Acta Mol Cell Res.2019;1866: 750-760.
33. Ma Y, Yang Y, Wang F, Moyer MP, Wei Q, Zhang P, Yang Z, Liu W, Zhang H, Chen N, Wang H. Long non-coding RNA CCAL regulates colorectal cancer progression by activating Wnt/β-catenin signalling pathway via suppression of activator protein 2α.Gut.2016;65: 1494-504.
34. Li XL, Subramanian M, Jones MF, Chaudhary R, Singh DK, Zong X, Gryder B, Sindri S, Mo M. Long Noncoding RNA PURPL Suppresses Basal p53 Levels and Promotes Tumorigenicity in Colorectal Cancer.Cell Rep. 2017; 20: 2408-2423.
35. Ye C, Shen Z, Wang B, Li Y, Li T, Yang Y, Jiang K, Ye Y, Wang S. A novel long non-coding RNA lnc-GNAT1-1 is low expressed in colorectal cancer and acts as a tumor suppressor through regulating RKIP-NF-κB-Snail circuit.J Exp Clin Cancer Res. 2016; 35: 187.
36. Archer TC, Ehrenberger T, Mundt F, Gold MP, Krug K, Mah CK, Mahoney EL, Daniel CJ, LeNail A. Proteomics, Post-translational Modifications, and Integrative Analyses Reveal Molecular Heterogeneity within Medulloblastoma Subgroups.Cancer Cell. 2018; 34: 396-410.
37. Luo W, Wang M, Liu J, Cui X, Wang H. Identification of a six lncRNAs signature as novel diagnostic biomarkers for cervical cancer.J Cell Physiol. 2019;7:1-8.
38. Sepulveda AR, Hamilton SR, Allegra CJ, Grody W, Cushman-Vokoun AM, Funkhouser WK. Molecular Biomarkers for the Evaluation of Colorectal Cancer: Guideline From the American Society for Clinical Pathology, College of American Pathologists, Association for Molecular Pathology, and the American Society of Clinical Oncology.J Clin Oncol.2017;35:1453-1486.
39. Fearon Eric R. Molecular genetics of colorectal cancer.Annu Rev Pathol.2011;6: 479-507.

Figures

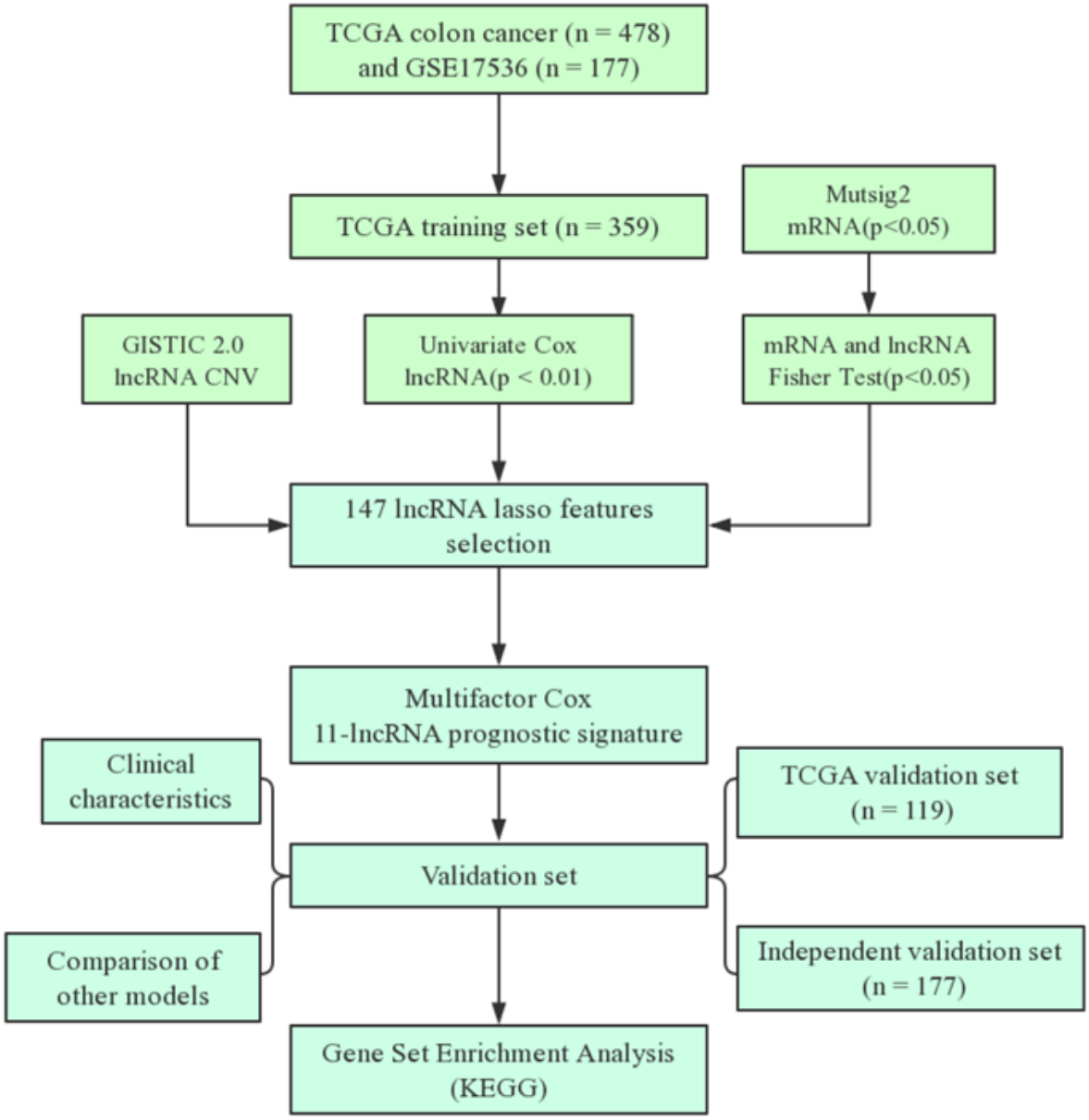

Figure 1

Workflow of identifying a COAD survival-related 11-lncRNA signature.

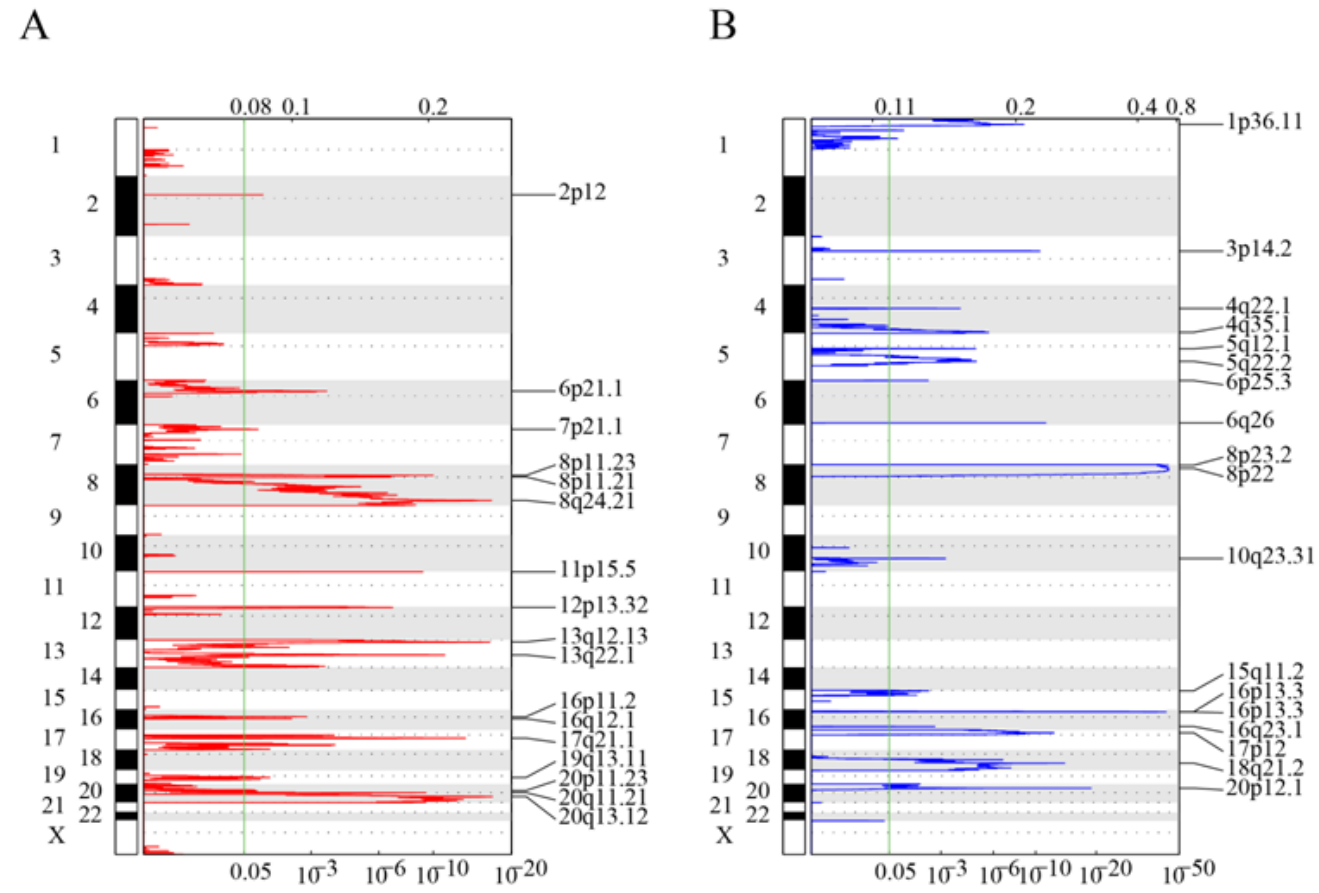

Figure 2

lncRNAs that are closely related to gene copy number variation. (A): A fragment that was significantly amplified in the COAD genome (p<0.05). (B): A fragment that was significantly deleted in the COAD genome (p<0.05).

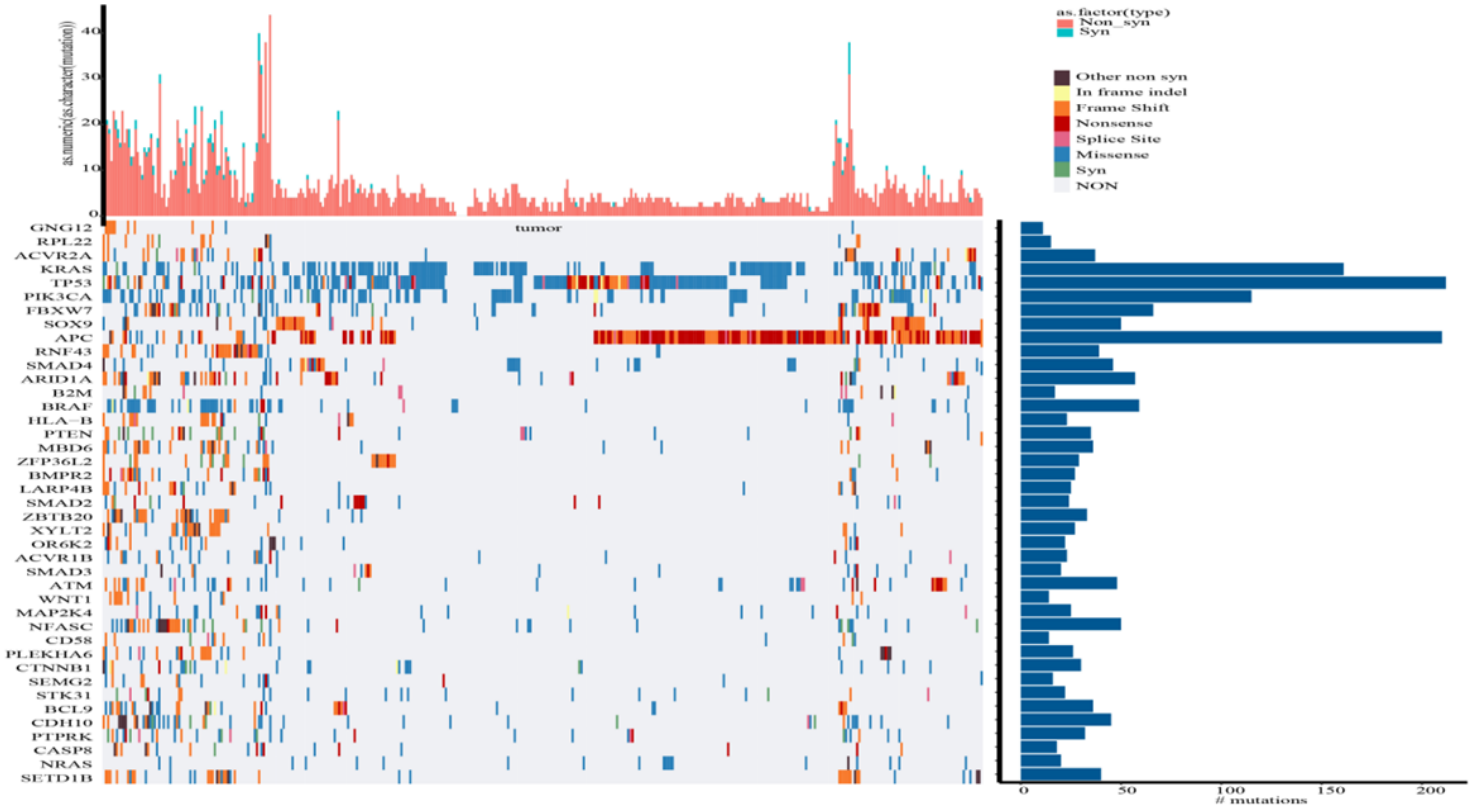

Figure 3

A total of 41 genes with significant mutation frequencies were identified through MutSig2. The upper histogram shows the total number of synonymous and nonsynonymous mutations in 41 genes per patient, and the right histogram shows the number of samples in which the 41 genes were mutated.

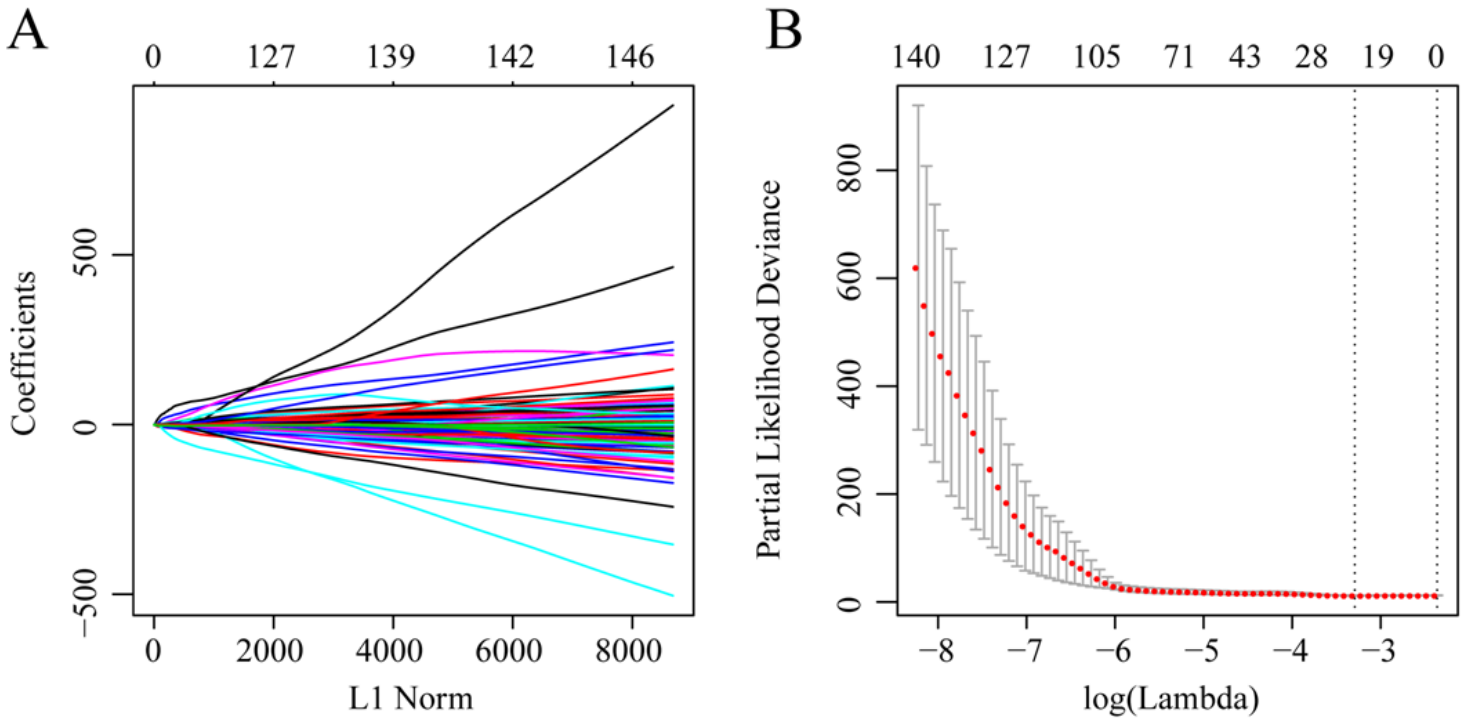

Figure 4

Target lncRNAs were identified and obtained by Lasso Cox regression. (A): The trajectory of each independent variable; the horizontal axis represents the log value of the independent lambda, and the vertical axis represents the coefficient of the independent variable; (B): The confidence interval under each lambda.

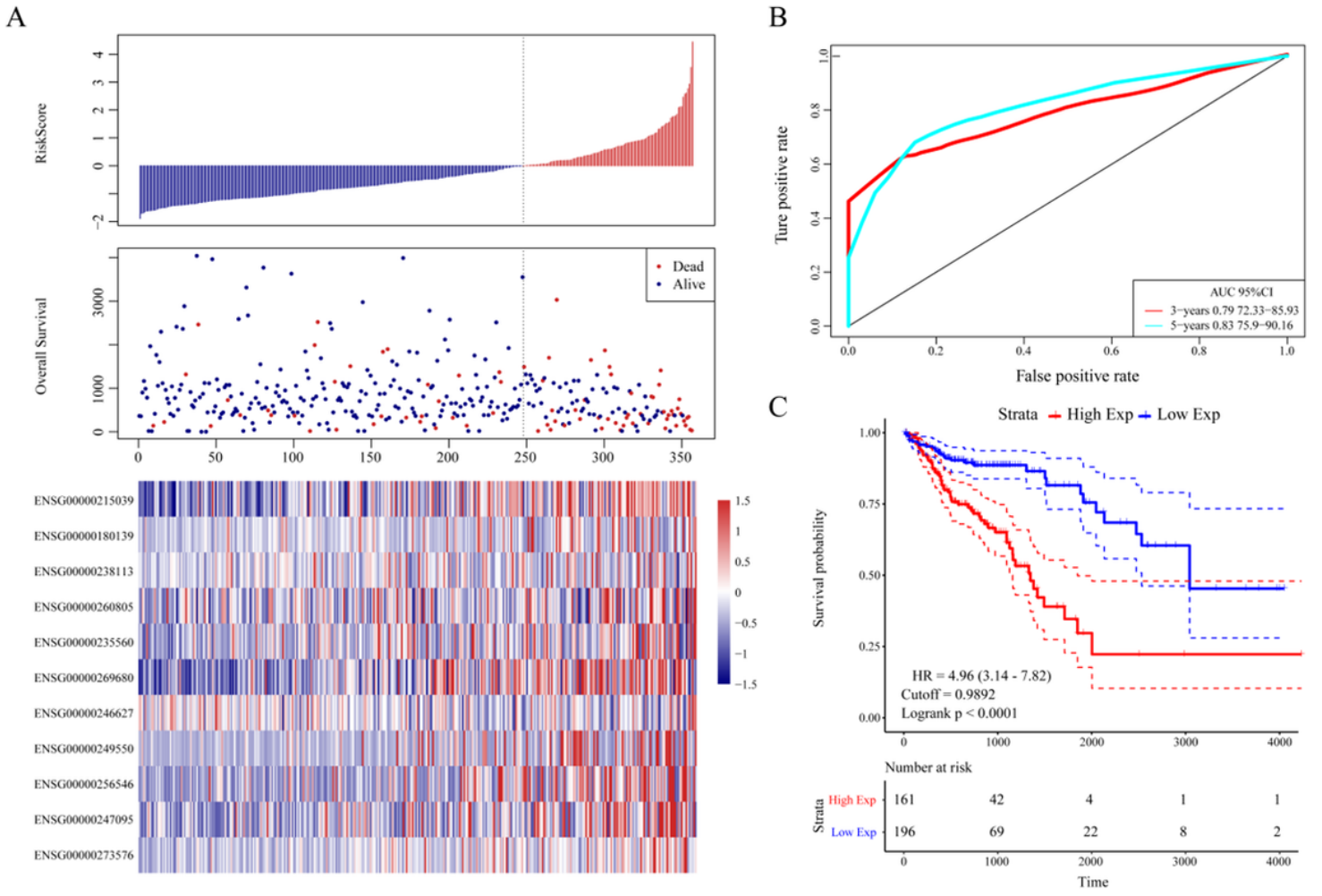

Figure 5

Determination and analysis of the 11-lncRNA signature in the training cohort. (A): Risk score, survival time, survival status and expression of the 11 lncRNAs in the TCGA training cohort. (B):11-lncRNA signature classification ROC curve and AUC. (C): 11-lncRNA signature Kaplan-Meier (KM) survival curve distribution in the TCGA training cohort.

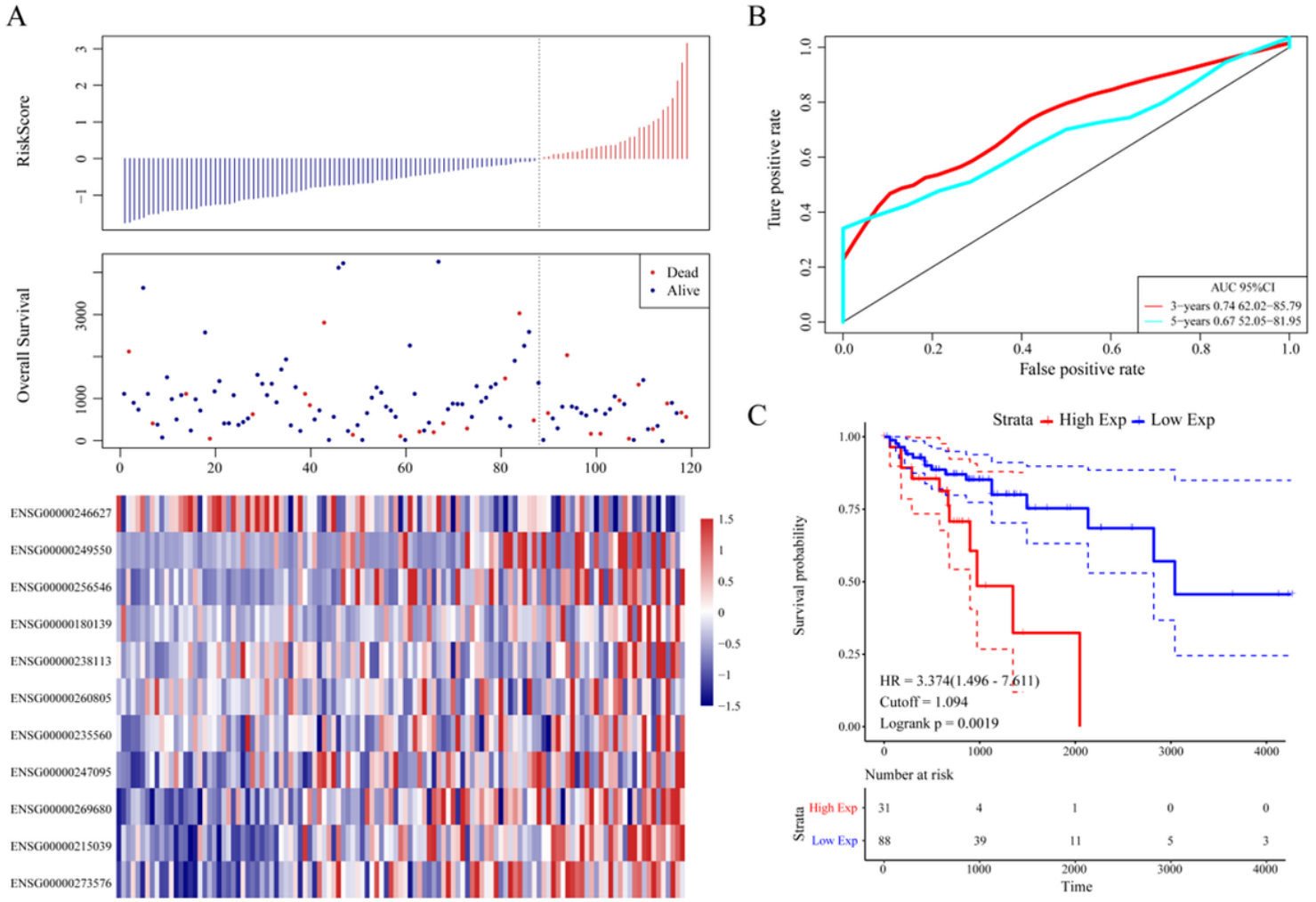

Figure 6

Validation of the 11-lncRNA signature in the testing cohort. (A): Risk score, survival time, survival status and expression of the 11 lncRNAs in the TCGA testing cohort. (B):11-lncRNA signature classification ROC curve and AUC. (C): 11-lncRNA signature KM survival curve distribution in the TCGA testing cohort.

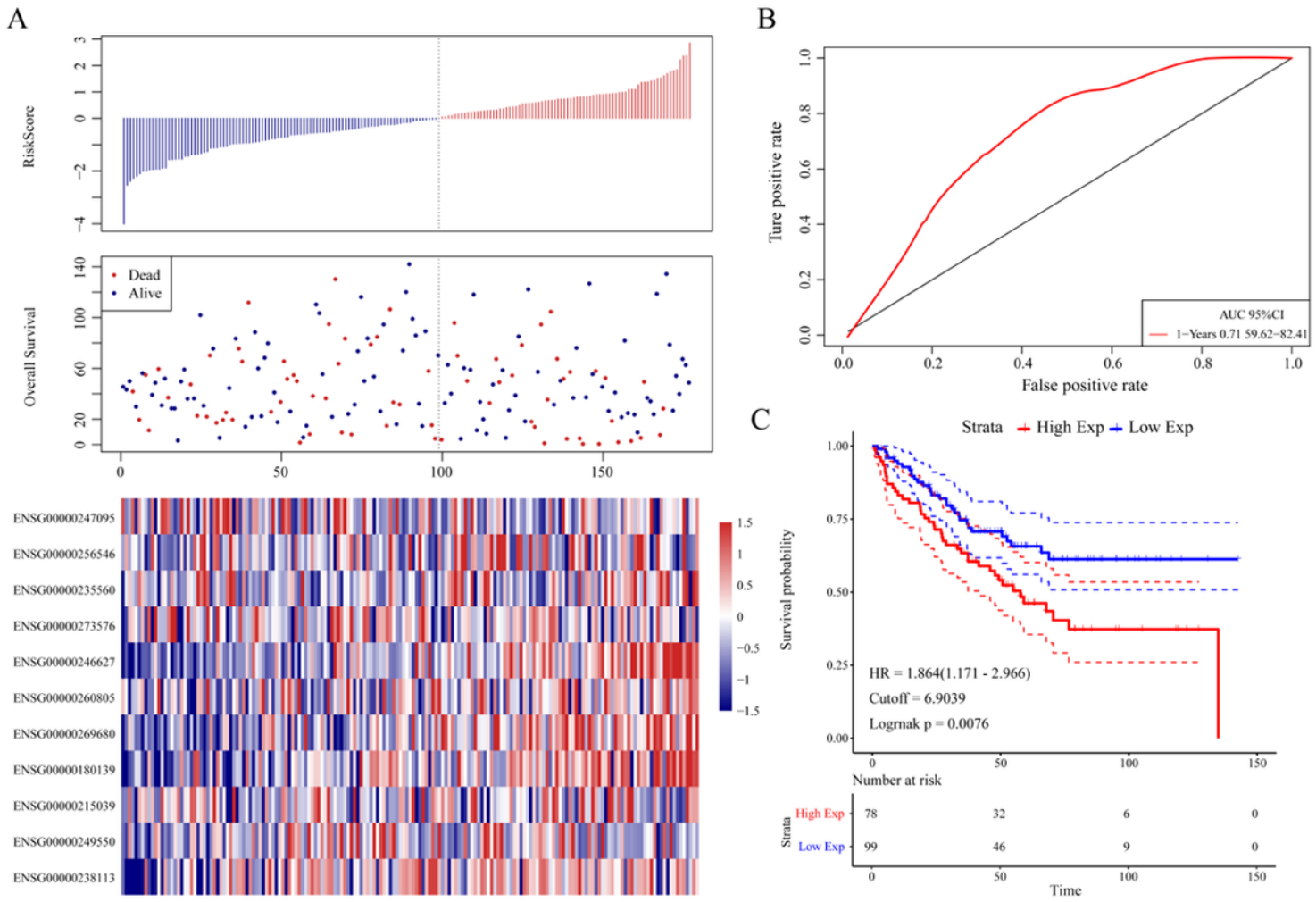

Figure 7

Validation of the 11-lncRNA signature in the GSE17536 cohort. (A): Risk score, survival time, survival status and expression of the 11 lncRNAs in the GSE17536 cohort. (B): The 11-lncRNA signature classification ROC curve and AUC in the GSE17536 cohort. (C): The 11-lncRNA signature KM survival curve distribution in the GSE17536 cohort.

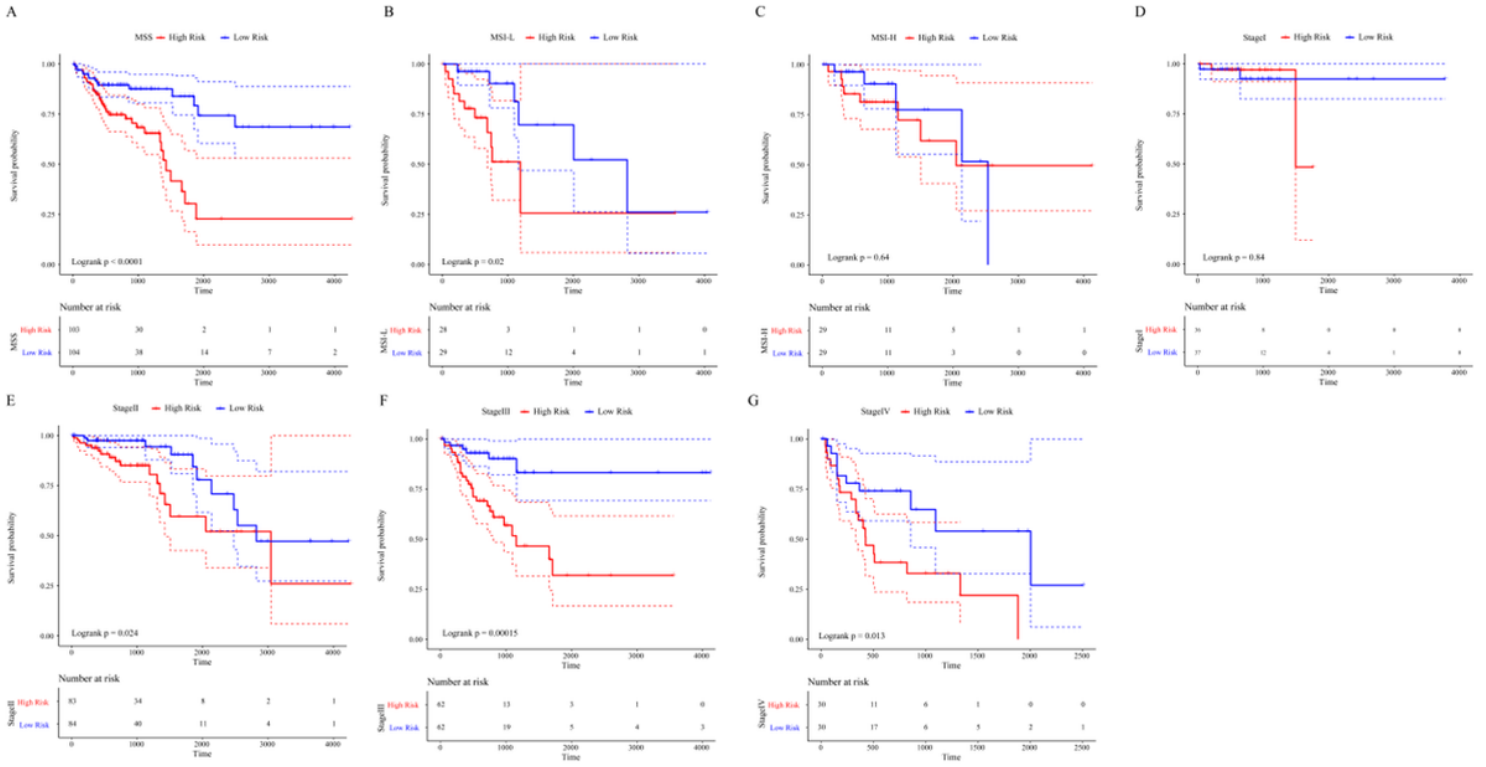

Figure 8

11-lncRNA signature KM survival curve distribution according to the MSI status and tumour stage. (A): 11-lncRNA signature KM survival curve distribution in the MSS group. (B): 11-lncRNA signature KM survival curve distribution in the MSI-L group. (C): 11-lncRNA signature KM survival curve distribution in the MSI-H group. (D): 11-lncRNA signature KM survival curve distribution in the stage I group. (E): 11-lncRNA signature KM survival curve distribution in the stage II group. (F): 11-lncRNA signature KM survival curve distribution in the stage III group. (G): 11-lncRNA signature KM survival curve distribution in the stage IV group.

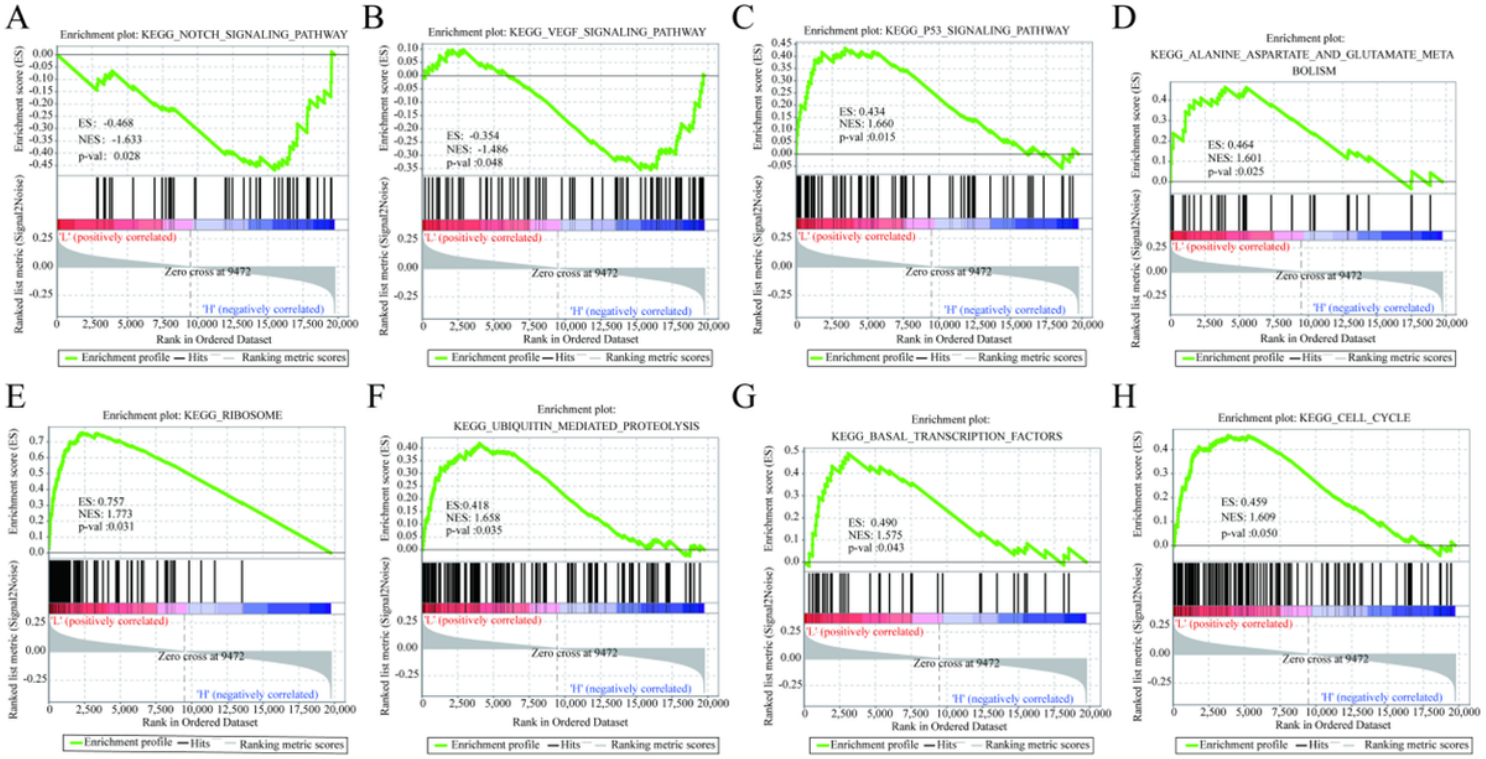

Figure 9

Signalling pathways associated with the 11-lncRNA signature were significantly enriched in the TCGA training cohort according to the GSEA. GSEA validated the enhanced activity of (A): “KEGG NOTCH SIGNALING PATHWAY”, (B): “ KEGG VEGF SIGNALING PATHWAY”, (C): “ KEGG P53 SIGNALING PATHWAY”, (D): “KEGG ALANINE ASPARTATE AND GLUTAMATE METABOLISM”, (E): “KEGG RIBOSOME”, (F): “KEGG UBIQUITIN MEDIATED PROTEOLYSIS”, (G): “KEGG BASAL RANSCRIPTION FACTORS”, and (H): “ KEGG CELL CYCLE”.

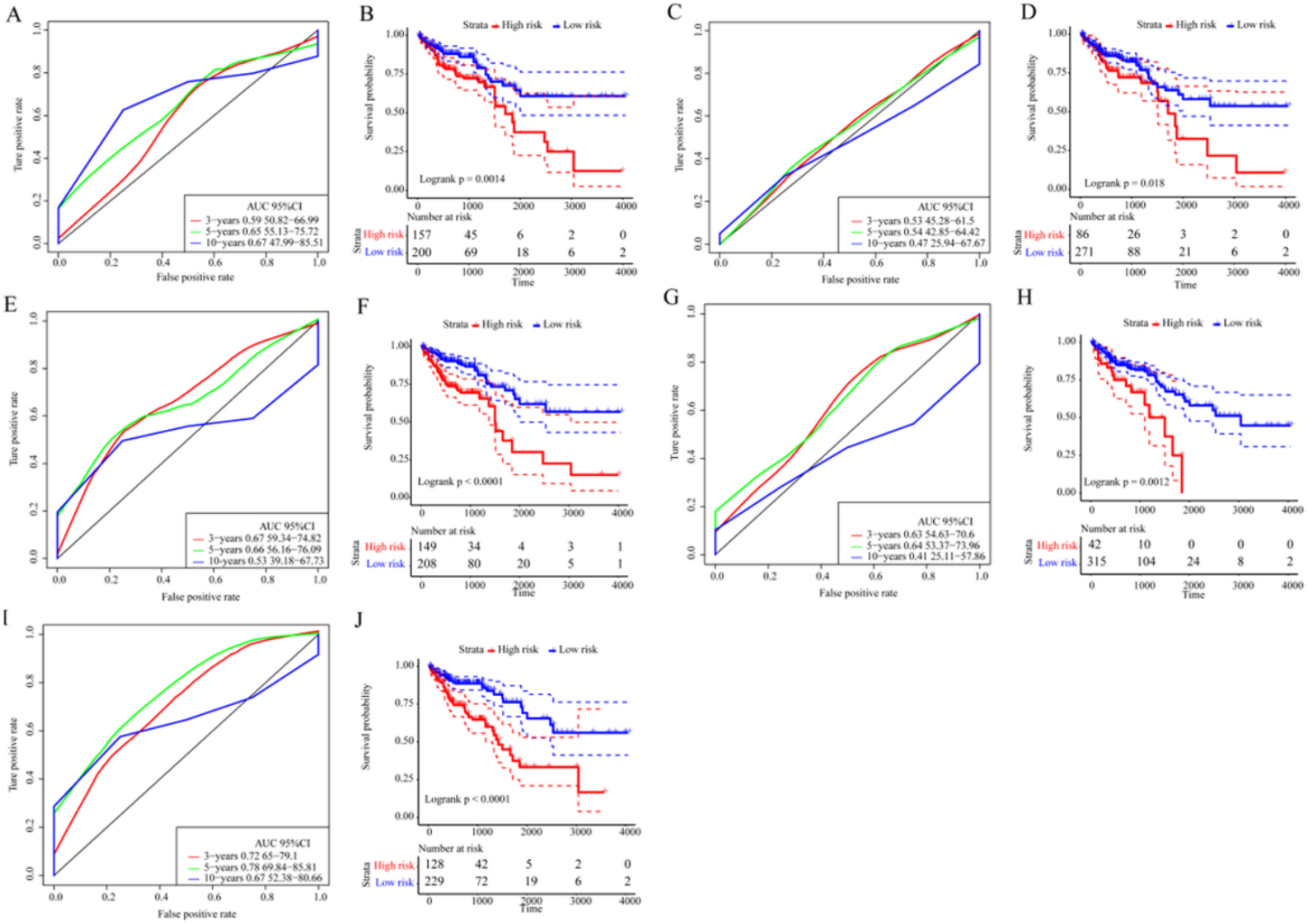

Figure 10

Comparison of the 11-lncRNA signature with other signatures. (A):The Zhao’s 6-lncRNA signature classification ROC curve and AUC. (B): The Zhao’s 6-lncRNA signature KM survival curve distribution.(C):The Xue’s 2-lncRNA signature classification ROC curve and AUC. (D): The Xue’s 2-lncRNA signature KM survival curve distribution.(E):The Xing’s 14-lncRNA signature classification ROC curve and AUC. (F): The Xing’s 14-lncRNA signature KM survival curve distribution.(G):The Fan’s 6-lncRNA signature classification ROC curve and AUC. (H): The Fan’s 6-lncRNA signature KM survival curve distribution. (I):The Wang’s 15-lncRNA signature classification ROC curve and AUC. (J): The Wang’s 15-lncRNA signature KM survival curve distribution.