World Journal of Gastroenterology

Submit a Manuscript: https://www.f6publishing.com

World J Gastroenterol 2019 January 14; 25(2): 220-232

DOI: 10.3748/wjg.v25.i2.220

ISSN 1007-9327 (print) ISSN 2219-2840 (online)

ORIGINAL ARTICLE

Basic Study

Six-long non-coding RNA signature predicts recurrence-free survival in hepatocellular carcinoma

Jing-Xian Gu, Xing Zhang, Run-Chen Miao, Xiao-Hong Xiang, Yu-Nong Fu, Jing-Yao Zhang, Chang Liu, Kai Qu

ORCID number: Jing-Xian Gu (0000-0002-6183-4004); Xing Zhang (0000-0001-9568-8124); Run-Chen Miao (0000-0003-1651-3970); Xiao-Hong Xiang (0000-0002-9020-9187); Yu-Nong Fu (0000-0001-8324-0613); Jing-Yao Zhang (0000-0002-1177-0401); Chang Liu (0000-0001-7916-4670); Kai Qu (0000-0002-1138-3727).

Author contributions: Gu JX and Qu K designed the research; Gu JX, Zhang X, and Miao RC collected and analyzed data; Xiang XH and Fu YN constructed the figures; Gu JX, Zhang JY, Liu C, and Qu K drafted and revised the manuscript.

Supported by The National Natural Science Foundation of China, No. 81773128 and No. 81871998; the Natural Science Basic Research Plan in Shaanxi Province of China, No. 2017JM8039; China Postdoctoral Science Foundation, No. 2018m641000; and Research Fund for Young Star of Science and Technology in Shaanxi Province, No. 2018KJXX-022.

Conflict-of-interest statement: None.

Data sharing statement: The data used in this manuscript are accessible through https://www.ncbi.nlm.nih.gov/geo/ and https://portal.gdc.cancer.gov/.

Jing-Xian Gu, Xing Zhang, Run-Chen Miao, Xiao-Hong Xiang, Yu-Nong Fu, Jing-Yao Zhang, Chang Liu, Kai Qu, Department of Hepatobiliary Surgery, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, Shaanxi Province, China

Corresponding author: Kai Qu, PhD, MD, Associated Professor, Department of Hepatobiliary Surgery, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, Shaanxi Province, China. qukai@xjtu.edu.cn
Telephone: +86-13609117104
Fax: +86-29-85323900

Abstract

BACKGROUND

Recent evidence shows that long non-coding RNAs (lncRNAs) are closely related to hepatogenesis and a few aggressive features of hepatocellular carcinoma (HCC). Increasing studies demonstrate that lncRNAs are potential prognostic factors for HCC. Moreover, several studies reported the combination of lncRNAs for predicting the overall survival (OS) of HCC, but the results varied. Thus, more effort including more accurate statistical approaches is needed for exploring the prognostic value of lncRNAs in HCC.

AIM

To develop a robust lncRNA signature associated with HCC recurrence to improve prognosis prediction of HCC.

METHODS

Univariate COX regression analysis was performed to screen the lncRNAs significantly associated with recurrence-free survival (RFS) of HCC in GSE76427 for the least absolute shrinkage and selection operator (LASSO) modelling. The established lncRNA signature was validated and developed in The Cancer Genome Atlas (TCGA) series using Kaplan-Meier curves. The expression values of the identified lncRNAs were compared between the tumor and non-tumor tissues. Pathway enrichment of these lncRNAs was conducted based on the significantly co-expressed genes. A prognostic nomogram combining the lncRNA signature and clinical characteristics was constructed.

RESULTS

The lncRNA signature consisted of six lncRNAs: *MSC-AS1*, *POLR2J4*, *EIF3J-AS1*, *SERHL*, *RMST*, and *PVT1*. This risk model was significantly associated with the RFS of HCC in the TCGA cohort with a hazard ratio (HR) being 1.807 (95%CI [confidence interval]: 1.329-2.457) and log-rank *P*-value being less than 0.001. The

Manuscript source: Unsolicited manuscript

Received: October 10, 2018
Peer-review started: October 12, 2018
First decision: November 15, 2018
Revised: December 5, 2018
Accepted: December 19, 2018
Article in press: December 20, 2018
Published online: January 14, 2019

best candidates of the six-lncRNA signature were younger male patients with HBV infection in relatively early tumor-stage and better physical condition but with higher preoperative alpha-fetoprotein. All the lncRNAs were significantly upregulated in tumor samples compared to non-tumor samples ($P < 0.05$). The most significantly enriched pathways of the lncRNAs were TGF-β signaling pathway, cellular apoptosis-associated pathways, *etc*. The nomogram showed great utility of the lncRNA signature in HCC recurrence risk stratification.

CONCLUSION

We have constructed a six-lncRNA signature for prognosis prediction of HCC. This risk model provides new clinical evidence for the accurate diagnosis and targeted treatment of HCC.

Key words: Long non-coding RNAs; Hepatocellular carcinoma; Prognostic signature; Recurrence-free survival; Least absolute shrinkage and selection operator

Core tip: In this study, we constructed a six-lncRNA signature that could predict the recurrence-free survival (RFS) of hepatocellular carcinoma (HCC) *via* a novel formulation method - the least absolute shrinkage and selection operator (LASSO). LASSO, based on penalized regression, can handle all the independent variables simultaneously, which has been demonstrated to be more accurate than stepwise regression. This lncRNA-signature was further validated and developed in another independent dataset from The Cancer Genome Atlas project. Our risk score system showed great utility in predicting the RFS of HCC and provided extra evidence for guiding targeted treatment of HCC.

Citation: Gu JX, Zhang X, Miao RC, Xiang XH, Fu YN, Zhang JY, Liu C, Qu K. Six-long non-coding RNA signature predicts recurrence-free survival in hepatocellular carcinoma. *World J Gastroenterol* 2019; 25(2): 220-232
URL: https://www.wjgnet.com/1007-9327/full/v25/i2/220.htm
DOI: https://dx.doi.org/10.3748/wjg.v25.i2.220

INTRODUCTION

Hepatocellular carcinoma (HCC) is the fifth leading cause of cancer-related death worldwide and its incidence rate per year remains increasing rapidly[1]. HCC is an extremely heterogenous tumor from either clinical or molecular aspect, which is mainly due to the unique somatic genomic alteration patterns of each tumor[2]. In recent years, more mutation genes have been revealed to be involved in the tumorigenesis and progression of HCC, such as *TP53*, *CTNNB1*, and *mTOR*[3]. These molecular markers will help identify high-risk patients and provide guidance for therapeutic strategies directed at the individual patient. Due to the high heterogeneity of HCC, although therapeutic modalities have largely improved over the past decades, the prognosis remains unsatisfactory[4,5]. Therefore, more effective biomarkers for early diagnosis and precise prognostic prediction are in urgent need.

Long non-coding RNAs (lncRNAs) belong to the family of non-coding RNAs and measures longer than 200 nucleotides in length[6]. Accumulating evidence has shown that lncRNAs play an important part in a series of cellular biological processes and are associated with the initiation, progression, and migration of a wide range of malignancies including HCC[7]. However, so far, only a few lncRNAs like *HOTAIR*, *HULC*, and *TERC* have been well described about their oncogenic roles in HCC[8-10]. Apart from them, more lncRNAs have been proposed as the diagnostic or prognostic biomarkers of HCC recently[11]. Yet by far, only a few prognostic models of HCC based on lncRNAs have been developed. For all we know, a total of three studies have reported the prognostic models based on lncRNAs of HCC[12-14]. Although all three previous studies used the overall group or part of The Cancer Genome Atlas (TCGA) cohort as the discovery dataset, the results varied. Therefore, to derive a more convincing result and find more potentially functional lncRNAs in HCC, in the present study, we selected GSE76427 from Gene Expression Omnibus (GEO) database as the discovery dataset and another independent dataset from TCGA database as the validation series. Moreover, the risk score system of HCC was formulated by a contemporary clinico-practical statistical method, the least absolute shrinkage and selection operator (LASSO) algorithm which was more accurate than multivariate COX model used by the previous three studies[15]. We here aimed to construct a robust lncRNA expression-based signature to improve prognostic prediction of HCC *via* comprehensive genomic data analysis.

MATERIALS AND METHODS

Microarray datasets

Microarray data of GSE76427 were downloaded from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) database. GSE764267 was conducted by GPL10558 (Illumina HumanHT-12 V4.0 expression beadchip). It contained 115 HCC tissue samples and 52 adjacent non-tumor tissue samples. Out of 115 tumor samples, 108 with complete follow-up information (recurrence status and recurrence-free survival [RFS]) were included in the discovery dataset. All 108 participants underwent curative resection for HCC. Among the 108 participants, 22 were female and the other 86 were male. The average age of all the participants in the discovery dataset was 63.4 years old. With respect to HCC stage, 86 of 108 were in stage I or II, 21 in stage III/IV, and only 1 patient had no record of cancer stage. The median follow-up was 1.17 years. HCC recurrence was diagnosed according to the established criteria reached by International Working Party[16]. Each array from GPL10558 consisted of more than 47,000 probes corresponding to more than 31,000 annotated genes including coding and non-coding genes, microRNAs, rRNAs, and other short RNAs. We extracted all the long non-coding RNAs from GPL10558 for the preliminary screening of prognostic lncRNAs. Three hundred and thirty-seven HCC patients with recurrence information from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) constituted the validation dataset. The lncRNA expression profiles, clinical characteristics, follow-up data, and genetic mutation information of the TCGA cohort were downloaded. Besides, the RNA-Seq data of the 49 paired non-tumor tissue samples from the TCGA database were also obtained. The median RFS of GSE76427 and TCGA series was 8.4 and 13.0 months, respectively.

Construction and confirmation of an lncRNA signature

Univariate COX regression analysis was performed to screen the prognostic lncRNAs. Then, LASSO was applied to the construction of an HCC prognostic signature with the screened lncRNAs[15]. LASSO statistical algorithm was conducted using “glmnet” package in the R software (version 3.4.0, https://www.r-project.org/)[17]. Based on the expression levels of each sample, LASSO identified the eligible lncRNAs for the risk system and generated the corresponding coefficients for each of them.

The risk scores of each sample from the discovery and validation groups were calculated according to the risk model. The respective medians of two groups were used as the cut-off value to divide patients into high-risk and low-risk groups. Kaplan-Meier curves were plotted to compare the RFS of high-risk and low-risk patients. Meanwhile, *P*-values and hazard ratio (HR) with 95% confidence interval (CI) were generated by log-rank tests. Stratified survival analysis was carried out to identify the best candidates for the prognostic signature. The overall group was divided into subgroups by their clinical characteristics. Kaplan-Meier analysis was performed in the subgroups using the same cut-off value as the overall group. Kaplan-Meier curves were plotted using GraphPad Prism software (version 7.0).

Expression levels of the identified lncRNAs in tumor and non-tumor tissues were compared using TCGA RNA-seq data. The receiver operating characteristic (ROC) curve was plotted with GraphPad Prism (version 7.0). The area under the ROC curve (AUC) for evaluating discriminatory ability was calculated as well. Besides, the distribution of high-risk or low-risk patients in early- and late-stage subgroups were also compared *via* Chi-square test.

Function prediction of the prognostic lncRNAs

Pearson correlation analyses were conducted between the identified lncRNAs and the protein-coding genes in TCGA dataset based on their expression levels. The correlation coefficient > 0.4 and *P* < 0.001 were considered significantly correlated. The significantly co-expressed mRNAs were thrown into a publicly available web tool, Enrichr, for BioCarta (http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) pathway enrichment[18]. The enriched BioCarta terms were sorted by rank based ranking, an algorithm assessing the deviation from the expected rank to the mean rank[19].

Statistical analysis

Univariate and multivariate COX regression analyses were carried out in TCGA dataset to identify the independent risk factors for the RFS of HCC patients. A composite nomogram predicting the RFS of HCC was established based on the independent factors using the "rms" package of R statistical software. The concordance index (C-index) was calculated to evaluate the discriminatory ability of the nomogram. And calibration curves were plotted to compare the predicted and actual probabilities of RFS. Each component of the nomogram gives points and the sum of them represents the total points a patient receives. All the participants were divided into different risk groups according to their total points. Kaplan-Meier analysis was utilized to compare the RFS of different risk groups. Statistical analyses were performed with SPSS 23.0 (SPSS, Chicago, IL), unless otherwise indicated. A P-value < 0.05 was considered statistically significant.

RESULTS

Construction of a risk score system associated with RFS in HCC

The flow chart of the study procedure is presented in Figure 1. All the lncRNAs in the discovery dataset (GES76427) were subjected to univariate COX analysis and those significantly associated with RFS (P < 0.05) were considered as prognostic ones for LASSO modelling. The risk score formula for RFS was calculated as follows: risk score = 0.021355462 × (expression value of *MSC-AS1*) + 0.018051929 × (expression value of *POLR2J4*) + 0.016385849 × (expression value of *EIF3J-AS1*) + 0.01340867 × (expression value of *SERHL*) + 0.012263937 × (expression value of *RMST*) + 0.007303891 × (expression value of *PVT1*). From the formula, it is seen that these lncRNAs were all risk factors for HCC recurrence (coefficient > 0). And the value of their respective coefficients represented how much impact they had on the RFS prediction. It is obvious that *MSC-AS1* had the most while *PVT1* had the least impact. The risk model generates a risk score for each participant. Using the median of the risk scores of the whole discovery group, 9.100, as the cut-off value, 108 patients were classified as high-risk or low-risk ones (Figure 2A). The recurrence status, RFS period, and six lncRNAs' expression value of each patient are presented in Figure 2A as well. Kaplan-Meier curves showed that the low-risk group had significantly longer RFS than the high-risk group with a P-value from log-rank test of 0.024 (HR = 1.842, 95%CI: 1.026-3.309) (Figure 2B).

Validation and development of a prognostic signature in TCGA cohort

To confirm the predictive ability of the lncRNA-signature, validation analysis was carried out in a group of 337 HCC patients from the TCGA project. The whole validation group was divided into high-risk and low-risk groups accordingly in the discovery dataset. The median score (0.0357) of the whole validation group was adopted as the cut-off value. Survival analysis showed great performance of the risk model in stratifying high-risk and low-risk patients with a log-rank P-value being less than 0.001 (HR = 1.807, 95%CI: 1.329-2.457) (Figure 2C).

Stratified survival analysis in the validation series was conducted to further investigate the suitable patient group of the six-lncRNA signature. The cut-off value in the subgroups was consistent with that in the overall group (0.0357). The result of stratification analysis was shown in Table 1. Our risk score system was more applicable to the patients possessing the following characteristics: TNM stage I/II, male gender, younger than 60 years, Asian, with family history, hepatitis B virus (HBV) infection, alcohol consumption, ECOG = 0, and higher levels of preoperative serum albumin (ALB, >3,5g/dl) and alpha-fetoprotein (AFP, >20 ng/ml).

Differential expression of the identified lncRNAs in tumor and non-tumor tissues

To investigate the expression profile of the identified lncRNAs, we compared the expression values of the six lncRNAs between tumor and non-tumor samples in TCGA dataset. The results showed that all six lncRNAs were significantly upregulated in HCC presenting with remarkably significant P-values ($P < 1.0 \times 10^{-10}$) (Figure 3A-F). In addition, the ROC curve showed the great utility of the combined six-lncRNA signature in discriminating tumor from non-tumor tissues. The AUC was 0.932 (95%CI: 0.898-0.966) (Figure 3G). Furthermore, the distribution of high-risk and low-risk patients in different stages was also examined. Patients in late-stage (TNM stage III/IV) had a higher likelihood of being high-risk patients than those in early-stage (TNM stage I/II) (P < 0.05), implying that higher risk score was associated with relatively late HCC stage (Figure 4H).

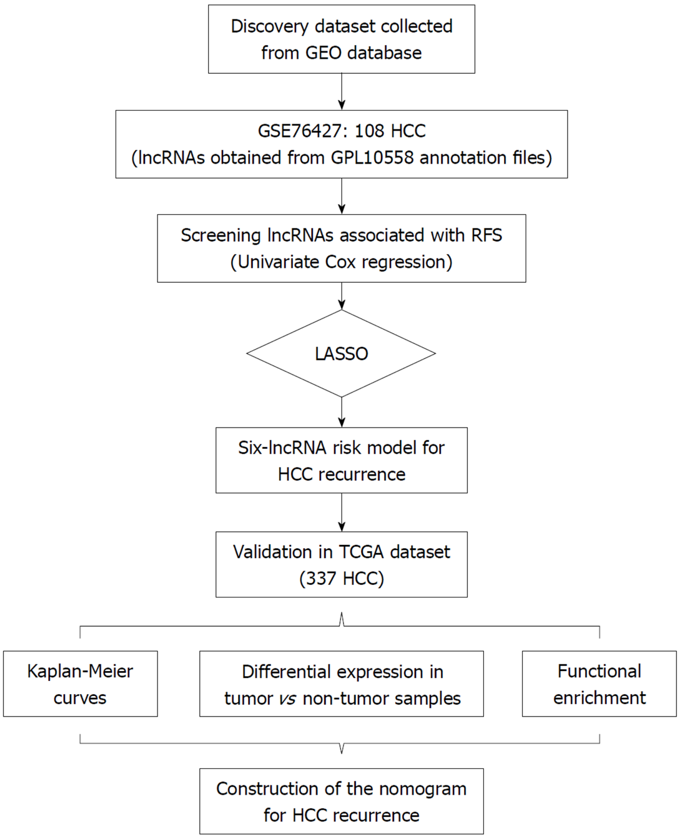
Discovery dataset collected from GEO database
GSE76427: 108 HCC (lncRNAs obtained from GPL10558 annotation files)
Screening lncRNAs associated with RFS (Univariate Cox regression)
LASSO
Six-lncRNA risk model for HCC recurrence
Validation in TCGA dataset (337 HCC)
Kaplan-Meier curves
Differential expression in tumor *vs* non-tumor samples
Functional enrichment
Construction of the nomogram for HCC recurrence

Figure 1 Overall design of the present study. HCC: Hepatocellular carcinoma; RFS: Recurrence-free survival.

Functional enrichment analysis of the six lncRNAs

To further investigate the potential biological roles of the identified six lncRNAs, BioCarta pathways were enriched using the co-expressed protein-coding genes of these lncRNAs. A gene significantly correlating with at least one of the six lncRNAs (Pearson coefficient > 0.4 and P <0.001) was considered eligible for pathway enrichment. Top ten highly enriched pathways are shown in Figure 4. These co-expressed genes of the lncRNAs clustered most significantly in TGF-β signaling pathway, internal ribosome entry pathway, granzyme A mediated apoptosis, FAS signaling pathway, calcium signaling by HBx, p38/MAPK signaling pathway, *etc*. Most of them are classical and vital pathways involved in HCC initiation and progression.

Establishment of a nomogram predicting RFS in HCC patients

To develop a composite predictor for the RFS of HCC patients, we combined the six-lncRNA signature, clinicopathological characteristics, and *TP53* mutation status together for the screening of the independent factors for RFS. The results from univariate and multivariate COX regression analyses showed that the identified independent risk factors for RFS included the six-lncRNA score, TNM stage, and ECOG [Eastern Cooperative Oncology Group] score (P < 0.05) (Table 2). The nomogram for RFS prediction was comprised of the three factors (Figure 5A). The C-index of the nomogram was 0.684 (95%CI: 0.635-0.733). The calibration curves for the probability of recurrence at 1 year and 3 years showed good agreement between the prediction from the nomogram and the actual observations (Figure 5B). Each patient got the total points according to the scoring of the nomogram. The tertiles of all the total points were used as the cut-off value (6.800 and 3.200) to divide the patients into high-, intermediate- and low-risk groups. The Kaplan-Meier analysis of the three risk subgroups indicated the great utility of the composite nomogram in discriminating HCC patients with good, intermediate, and poor prognosis (Figure 5C).

DISCUSSION

Although extensive research efforts have been made on proposing the diagnostic and prognostic indicators of HCC in the past few decades, there is still a long way to go to

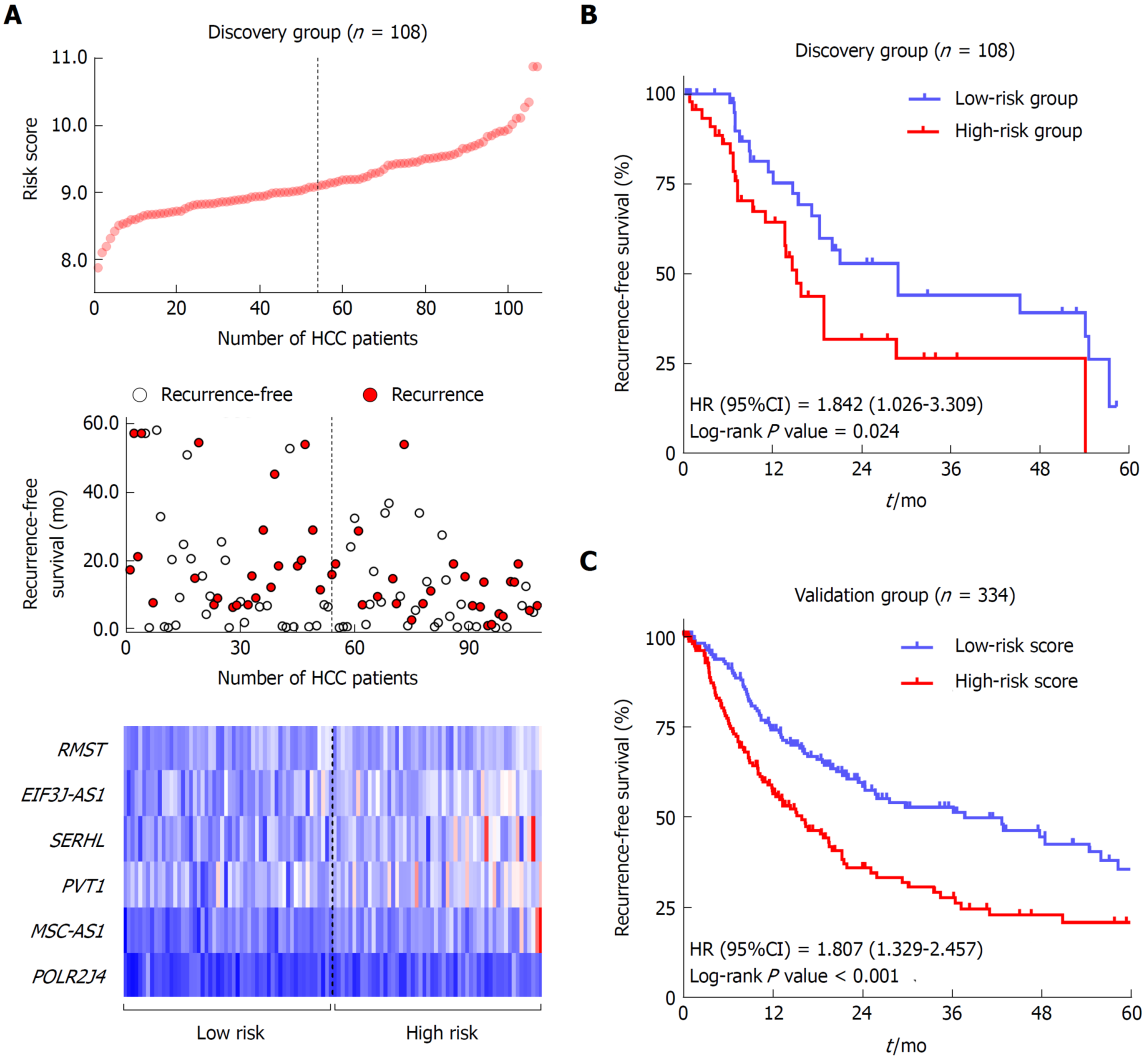

Figure 2 Construction and validation of a prognostic lncRNA signature for hepatocellular carcinoma. A: LncRNA risk score distribution (*Upper*), the recurrence status and recurrence-free survival (RFS) period (*Middle*), and expression profiles of the six lncRNAs (*Lower*) of the 108 patients in the discovery dataset. B: The Kaplan-Meier curve of the RFS between the high-risk and low-risk groups stratified by the median risk score in GSE76427 series. C: The Kaplan-Meier curve of the RFS between the high-risk and low-risk groups stratified by the median risk score in The Cancer Genome Atlas cohort.

construct a system of the molecular classifications of HCC[20]. A vast majority of the published studies investigating HCC predictors or predictive signatures were focused on protein-coding genes or microRNAs[21,22]. Currently, a group of non-coding RNAs, lncRNAs, which were overlooked previously, have attracted much attention. A growing number of studies have demonstrated that dysregulated lncRNAs in HCC are closely involved in the hepatocarcinogenesis, progression, and migration, and might be potential biomarkers for early detection, targeted therapies, and prognosis evaluation of HCC[7]. Thus, to help clarify the prognostic value of lncRNAs and refine prediction in HCC, we here carried out a comprehensive screening of the lncRNAs significantly associated with HCC recurrence in the public database (GSE76427) and used the significant lncRNAs to construct a prognostic model. We also presented the validation and development analysis of the established risk score system which showed good performance in predicting the RFS of HCC.

Compared to the lncRNA signatures of the three previous studies[12-14], our model was totally different. The reasons are probably as follows: First, our risk score system was constituted by LASSO penalized regression. Unlike traditional stepwise regression the previous studies used, LASSO algorithm can analyze all the independent variables simultaneously and tend to pick the most influential variables[17]. The coefficients of less influential variables will become zero after introduced to penalty following a regularization path[15]. Therefore, this formulation method was far more accurate than the stepwise regression of multivariate COX model, especially when dealing with very large datasets, like genomics[23]. Second, the three published studies all employed TCGA cohort as the discovery dataset while we used the GSE76427 for discovery and TCGA for validation. It was mainly the

Table 1 Stratified analysis of recurrence-free survival in The Cancer Genome Atlas cohort

Characteristic	Recurrence-free survival		
	High-risk/low-risk	HR (95%CI)	*P*-value
Overall	165/169	1.807 (1.329-2.457)	< 0.001[a]
TNM stage			
I/II	104/130	1.595 (1.071-2.375)	0.018[a]
III/IV	48/29	1.295 (0.759-2.208)	0.340
Gender			
Male	111/116	2.150 (1.476-3.133)	< 0.001[a]
Female	53/54	0.940 (0.552-1.598)	0.818
Age (yr)			
≤ 60	81/80	2.244 (1.435-3.510)	< 0.001[a]
> 60	83/90	1.155 (0.758-1.761)	0.499
Hepatitis B			
With HBV	39/60	2.013 (1.056-3.837)	0.022[a]
Without HBV	46/32	1.371 (0.685-2.743)	0.371
Hepatitis C			
With HCV	26/25	0.964 (0.466-1.997)	0.922
Without HCV	46/54	1.434 (0.792-2.596)	0.224
Alcohol consumption			
Yes	55/52	2.209 (1.301-3.752)	0.002[a]
No	107/110	1.404 (0.954-2.066)	0.080
Liver cirrhosis			
Yes	27/48	1.353 (0.697-2.626)	0.340
No	62/63	1.263 (0.770-2.070)	0.354
Albumin (g/dL)			
≤ 3.5	36/41	0.908 (0.464-1.776)	0.776
> 3.5	88/110	1.682 (1.132-2.501)	0.007[a]
Creatinine (mg/dL)			
< 1.1	85/101	1.419 (0.945-2.130)	0.084
≥ 1.1	39/50	1.686 (0.891-3.189)	0.088
Alpha-fetoprotein (ng/mL)			
≤ 20	55/84	0.874 (0.535-1.428)	0.590
> 20	58/58	1.780 (1.066-2.972)	0.021[a]
Platelet(× 10^9/L)			
≤ 250	80/102	1.517 (0.983-2.340)	0.052
> 250	45/52	1.431 (0.818-2.502)	0.181
Race			
Asian	73/75	2.160 (1.329-3.510)	0.001[a]
White	78/82	1.275 (0.847-1.919)	0.238
BMI			
< 24	75/71	1.597 (1.007-2.533)	0.039[a]
≥ 24	75/86	1.692 (1.085-2.638)	0.017[a]
Family history			
Yes	45/55	1.883 (1.074-3.304)	0.022[a]
No	102/89	1.367 (0.919-2.034)	0.120
ECOG			
0	64/88	2.106 (1.279-3.468)	0.002[a]
> 0	61/53	1.334 (0.837-2.125)	0.223

[a]Statistically significant. BMI: Body mass index; ECOG: Eastern Cooperative Oncology Group; HBV: Hepatitis B virus; HCV: Hepatitis C virus.

discovery dataset that determined the key lncRNAs for further validation and

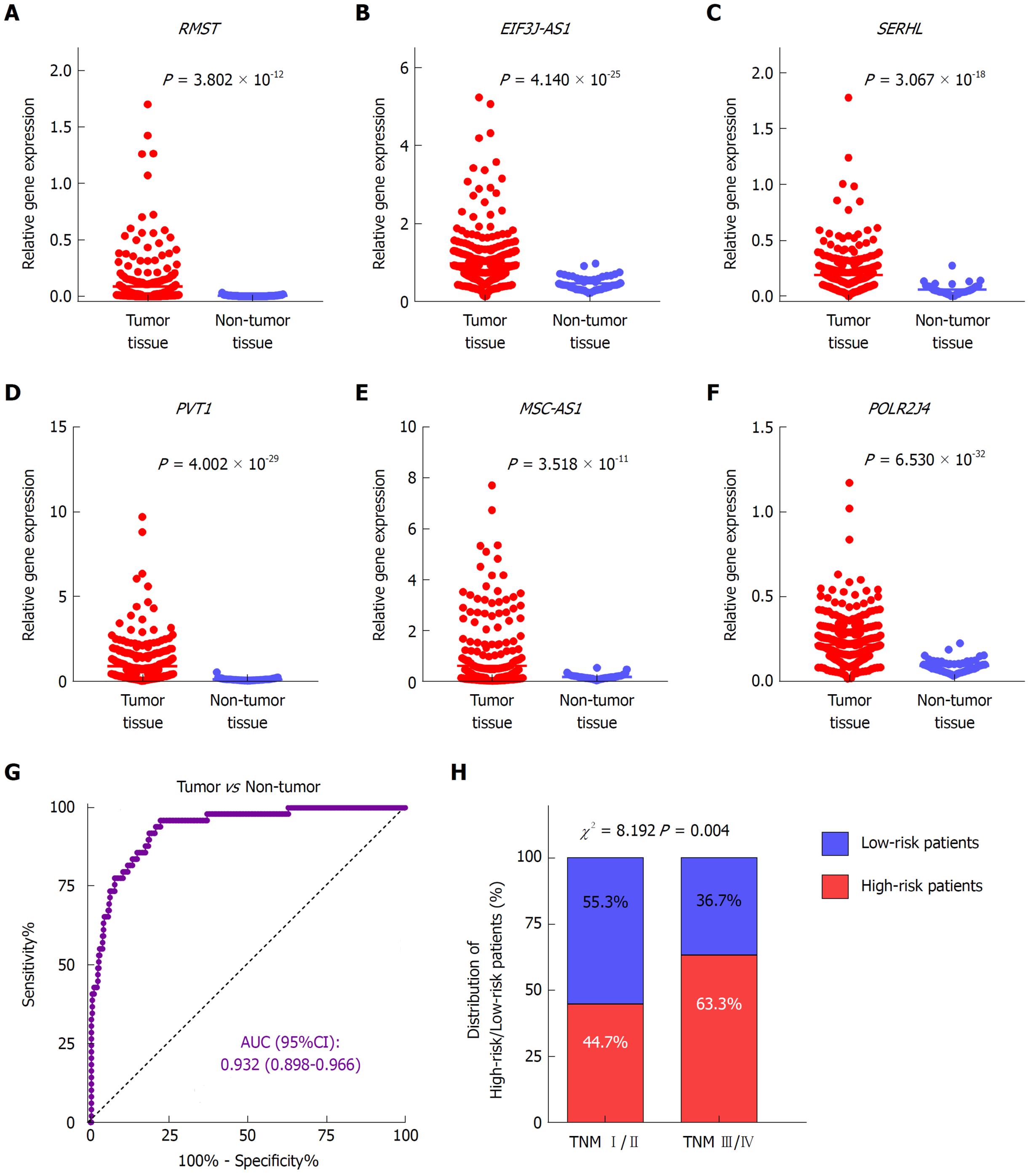

Figure 3 Expression patterns of the six lncRNAs in tumor and non-tumor tissues. Differential expression of *RMST* (A), *EIF3J-AS1* (B), *SERHL* (C), *PVT1* (D), *MSC-AS1* (E), and *POLR2J4* (F) between hepatocellular carcinoma and non-cancerous samples. G: The ROC curve of tumor tissue *vs* non-tumor tissue discriminated by the six-lncRNA signature. (H) Comparisons of the distribution of high-risk and low-risk patients in early stage (TNM I/II) and late stage (TNM III/IV) by the chi-square test.

investigation. Moreover, all the three studies adopted overall survival (OS) as an evaluating indicator of prognosis while our study used RFS for prognostic prediction, which implied that our signature might be more suitable for recurrence assessment.

The expression values of the lncRNAs constituting the risk model in patients' tumor tissue can be tested *via* liver biopsy or from the surgical specimen. Out of the six lncRNAs, *PVT1* (plasmacytoma variant translocation 1) has been demonstrated to be the activator of myelocytomatosis (*MYC*), a well-described oncogene[24]. *PVT1* is upregulated in a wide variety of malignancies, particularly in digestive cancers, and is associated with a poor clinical outcome[25]. In HCC, studies showed that *PVT1* could promote cell proliferation and stem-cell like potential by upregulating *NOP2*[26]. Recent studies also revealed that *PVT1* regulated the iron metabolism and cell apoptosis in HCC and promoted tumor progression and metastasis[27]. All these

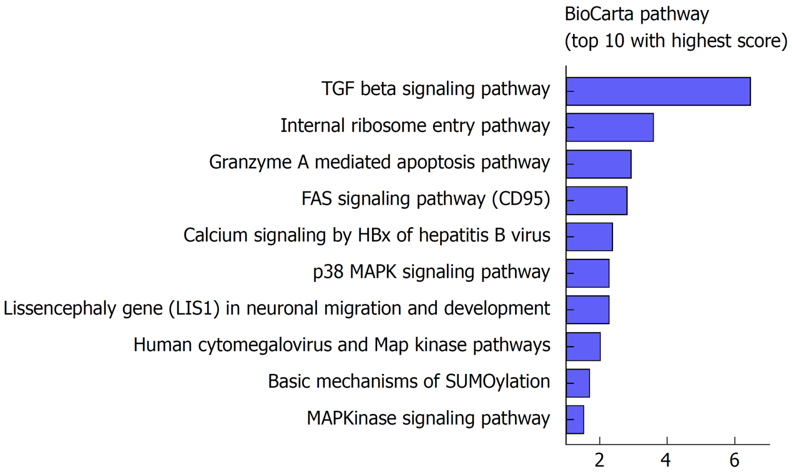

Figure 4 Functional annotation of the six lncRNAs. Top 10 significantly enriched BioCarta pathways using the co-expressed mRNAs of the six lncRNAs in The Cancer Genome Atlas database.

results suggested that *PVT1* might be a powerful biomarker for HCC survival. Our prognostic signature enrolled *PVT1*, favoring the prognostic value of *PVT1* in HCC as well. In case of *RMST* (rhabdomyosarcoma 2-associated transcript), it was reported to be involved in stem cell differentiation and neurogenesis[28]. However, to the best of our knowledge, there has not been any research into the role of *RMST* in HCC by now, either clinically or from the standpoint of molecular mechanism. Instead, our study provided novel evidence that *RMST*, as well as the other four lncRNAs (*MSC-AS1*, *POLR2J4*, *EIF3J-AS1*, and *SERHL*), with no published studies reporting their biological functions so far as we know, might be potential predictors of HCC prognosis, and further studies are needed to validate these results and investigate its molecular characteristics.

In this study, we identified a novel and robust lncRNA signature for prognostic prediction of HCC. Stratified survival analysis showed that this six-lncRNA signature was more suitable for the recurrence prediction of relatively younger (aged ≤ 60 years) Asian male patients with HBV infection, family history, and history of alcohol consumption who are in TNM stage I or II and better physical condition (ECOG = 0 and preoperative ALB > 3.5 g/dL) but with higher preoperative AFP. To further refine prediction, a nomogram combining the molecular signature and clinical markers was constructed. Although the biological functions of the identified lncRNAs in HCC have not been researched or reported except *PVT1*, pathway enrichment revealed that these lncRNAs might exert influence on tumorigenesis and progression of HCC through the TGF-β pathway and cellular apoptosis-associated pathways. In a word, our study highlighted the prognostic value of the six-lncRNA signature and suggested practical applications in prognostic prediction and targeted therapy of HCC.

Table 2 Univariate and multivariate Cox regression analyses of clinicopathologic characteristics associated with recurrence in The Cancer Genome Atlas samples

Variable	Univariate analysis		Multivariate analysis	
	HR (95%CI)	*P*-value	HR (95%CI)	*P*-value
Six-lncRNA risk score	1.341 (1.110-1.620)	0.002[a]	1.270 (1.018-1.585)	0.034[a]
TNM stage (IV/III/II/I)	1.727 (1.441-2.070)	< 0.001[a]	1.705 (1.393-2.089)	< 0.001[a]
Gender (male/female)	0.982 (0.711-1.355)	0.911	-	-
Age (> 60/≤ 60)	0.956 (0.707-1.294)	0.773	-	-
HBV (yes/no)	0.804 (0.505-1.281)	0.359	-	-
HCV (yes/no)	1.271 (0.798-2.207)	0.313	-	-
Alcohol consumption (yes/no)	1.062 (0.769-1.467)	0.717	-	-
Liver cirrhosis (yes/no)	1.271 (0.861-1.877)	0.228	-	-
Albumin (> 3.5/≤ 3.5 g/dL)	0.968 (0.659-1.424)	0.870	-	-
Creatinine (≥ 1.1/< 1.1 mg/dA)	0.739 (0.511-1.069)	0.109	-	-
AFP (> 20/≤ 20 ng/mL)	1.380 (0.972-1.959)	0.072		
Platelet (> 250/≤ 250 × 10^9/L)	1.314 (0.932-1.854)	0.119	-	-
Race (Asian/White)	1.270 (0.928-1.739)	0.136	-	-
BMI (≥ 24/< 24)	0.860 (0.697-1.062)	0.161	-	-
Family history (yes/no)	0.920 (0.655-1.292)	0.630	-	-
ECOG (> 0/0)	1.858 (1.329-2.598)	< 0.001[a]	1.486 (1.045-2.114)	0.028[a]
TP53 mutation (yes/no)	1.389 (1.004-1.922)	0.047[a]	1.326 (0.894-1.967)	0.161

[a] Statistically significant. AFP: Alpha-fetoprotein; BMI: Body mass index; ECOG: Eastern Cooperative Oncology Group; HR: Hazard ratio; HBV: Hepatitis B virus; HCV: Hepatitis C virus.

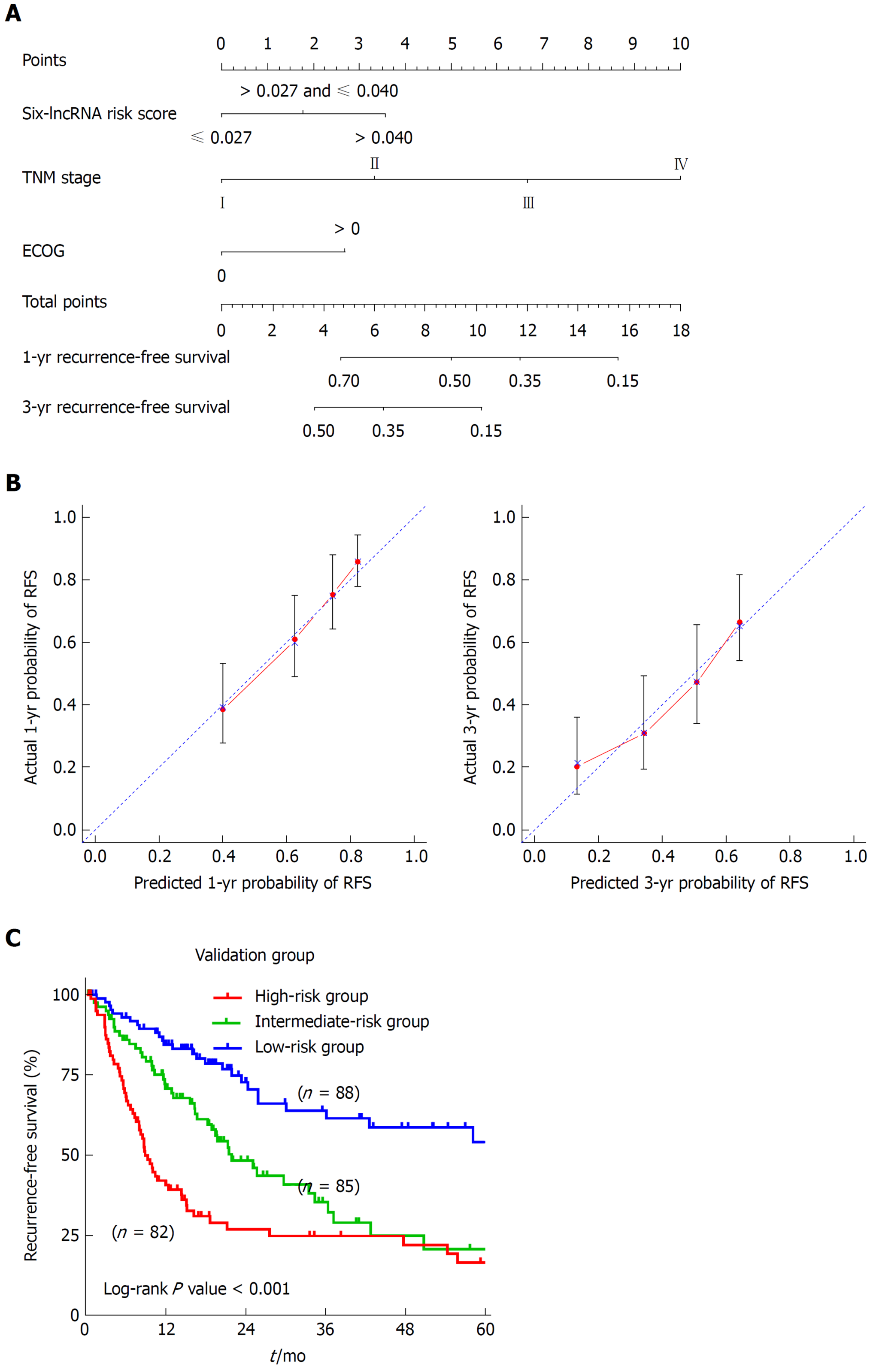

Figure 5 Construction of a nomogram for recurrence-free survival prediction in hepatocellular carcinoma. A: The composite nomogram consists of the six-lncRNA score, TNM stage, and ECOG score. Each component generates their respective points according to the “Points” line drawn above. Add the points from three variables together and find the location of the total points on “Total Points” line. Then draw a vertical line from “Total Points” axis to the two lower lines which corresponds to the predicted 1-year and 3-yr recurrence-free survival (RFS) rates by the nomogram. B: Calibration curves of the nomogram for the estimation of RFS rates at 1-year (Left) and 3-years (Right). The predicted and actual 1-yr and 3-yr RFS-probabilities were drawn on the x and y axis, respectively. C: The Kaplan-Meier curve of three risk subgroups stratified by the tertiles of total points derived from the nomogram.

ARTICLE HIGHLIGHTS

Research background

Hepatocellular carcinoma (HCC) is the most common type of liver cancer which remains a severe health issue worldwide. In recent years, genetic markers and predictive models have been put forward for improving the management of HCC. Meanwhile, many statistical techniques have been used for data mining in a series of large public databases involving the high-throughput genetic data of cancers. With the help of the most advanced clinic-practical methods, more accurate and robust prognostic models can be constructed for HCC.

Research motivation

Researchers have tried to constitute a prognostic model based on molecular biomarkers for HCC over these years. Long non-coding RNAs (lncRNAs) are novel predictive indicators. Although a few attempts have been made to construct lncRNA-based models for HCC, more are needed for further really significant findings.

Research objectives

By analyzing data from two databases, Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA), we wanted to identify a prognostic signature for HCC which is comprised of the potential functional lncRNAs.

Research methods

The latest statistical algorithm, the least absolute shrinkage and selection operator (LASSO), was utilized to constitute our predictive model. This method was performed based on the significant lncRNAs screened based on the lncRNA expression profiles from the GEO database. The expression values of the candidate lncRNAs were also examined in the HCC and normal liver tissues. The robustness of this model was validated using TCGA dataset. The suitable patients and other clinical applicability of the lncRNA-signature were explored as well.

Research results

The risk score system for predicting the recurrence of HCC was constructed based on the six lncRNAs (*MSC-AS1*, *POLR2J4*, *EIF3J-AS1*, *SERHL*, *RMST*, and *PVT1*) using LASSO. All six lncRNAs were aberrantly expressed in HCC and non-tumor tissue and they were significantly enriched in TGF-β signaling pathway and cellular apoptosis-related pathways. The best candidates we identified were younger early-staged male patients with HBV infection and family history in better physical condition but with higher preoperative AFP. To broaden the application scope of the model, a nomogram involving the lncRNA signature and other clinicopathological characteristics was formulated.

Research conclusions

The six-lncRNA signature showed great predictive ability in prognostic evaluation of HCC patients. This tool may help perform risk stratification and provide more individualized clinical advice for each patient.

Research perspectives

Our study offered extra evidence that lncRNAs are potential functional regulators in HCC progression. Finding effective molecular biomarkers and predictive signatures of HCC prognosis are future direction calling urgently for groundbreaking attempts.

REFERENCES

1 **Siegel RL**, Miller KD, Jemal A. Cancer Statistics, 2017. *CA Cancer J Clin* 2017; **67**: 7-30 [PMID: 28055103 DOI: 10.3322/caac.21387]

2 **Schulze K**, Nault JC, Villanueva A. Genetic profiling of hepatocellular carcinoma using next-generation sequencing. *J Hepatol* 2016; **65**: 1031-1042 [PMID: 27262756 DOI: 10.1016/j.jhep.2016.05.035]

3 **Llovet JM**, Zucman-Rossi J, Pikarsky E, Sangro B, Schwartz M, Sherman M, Gores G. Hepatocellular carcinoma. *Nat Rev Dis Primers* 2016; **2**: 16018 [PMID: 27158749 DOI: 10.1038/nrdp.2016.18]

4 **Villanueva A**, Hernandez-Gea V, Llovet JM. Medical therapies for hepatocellular carcinoma: a critical view of the evidence. *Nat Rev Gastroenterol Hepatol* 2013; **10**: 34-42 [PMID: 23147664 DOI: 10.1038/nrgastro.2012.199]

5 **Wallace MC**, Preen D, Jeffrey GP, Adams LA. The evolving epidemiology of hepatocellular carcinoma: a global perspective. *Expert Rev Gastroenterol Hepatol* 2015; **9**: 765-779 [PMID: 25827821 DOI: 10.1586/17474124.2015.1028363]

6 **Ulitsky I**, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. *Cell* 2013; **154**: 26-46 [PMID: 23827673 DOI: 10.1016/j.cell.2013.06.020]

7 **Klingenberg M**, Matsuda A, Diederichs S, Patel T. Non-coding RNA in hepatocellular carcinoma: Mechanisms, biomarkers and therapeutic targets. *J Hepatol* 2017; **67**: 603-618 [PMID: 28438689 DOI: 10.1016/j.jhep.2017.04.009]

8 **Gupta RA**, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010; **464**: 1071-1076 [PMID: 20393566 DOI: 10.1038/nature08975]

9 **Wang J**, Liu X, Wu H, Ni P, Gu Z, Qiao Y, Chen N, Sun F, Fan Q. CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic*

Acids Res 2010; **38**: 5366-5383 [PMID: 20423907 DOI: 10.1093/nar/gkq285]

10 **Mitchell M**, Gillis A, Futahashi M, Fujiwara H, Skordalakes E. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat Struct Mol Biol* 2010; **17**: 513-518 [PMID: 20357774 DOI: 10.1038/nsmb.1777]

11 **Zheng C**, Liu X, Chen L, Xu Z, Shao J. lncRNAs as prognostic molecular biomarkers in hepatocellular carcinoma: a systematic review and meta-analysis. *Oncotarget* 2017; **8**: 59638-59647 [PMID: 28938667 DOI: 10.18632/oncotarget.19559]

12 **Zhao QJ**, Zhang J, Xu L, Liu FF. Identification of a five-long non-coding RNA signature to improve the prognosis prediction for patients with hepatocellular carcinoma. *World J Gastroenterol* 2018; **24**: 3426-3439 [PMID: 30122881 DOI: 10.3748/wjg.v24.i30.3426]

13 **Wang Z**, Wu Q, Feng S, Zhao Y, Tao C. Identification of four prognostic LncRNAs for survival prediction of patients with hepatocellular carcinoma. *PeerJ* 2017; **5**: e3575 [PMID: 28729955 DOI: 10.7717/peerj.3575]

14 **Liao X**, Yang C, Huang R, Han C, Yu T, Huang K, Liu X, Yu L, Zhu G, Su H, Wang X, Qin W, Deng J, Zeng X, Ye X, Peng T. Identification of Potential Prognostic Long Non-Coding RNA Biomarkers for Predicting Survival in Patients with Hepatocellular Carcinoma. *Cell Physiol Biochem* 2018; **48**: 1854-1869 [PMID: 30092592 DOI: 10.1159/000492507]

15 **Gao J**, Kwan PW, Shi D. Sparse kernel learning with LASSO and Bayesian inference algorithm. *Neural Netw* 2010; **23**: 257-264 [PMID: 19604671 DOI: 10.1016/j.neunet.2009.07.001]

16 **International Working Party**. Terminology of nodular hepatocellular lesions. *Hepatology* 1995; **22**: 983-993 [PMID: 7657307]

17 **Friedman J**, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010; **33**: 1-22 [PMID: 20808728]

18 **Kuleshov MV**, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016; **44**: W90-W97 [PMID: 27141961 DOI: 10.1093/nar/gkw377]

19 **Chen EY**, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013; **14**: 128 [PMID: 23586463 DOI: 10.1186/1471-2105-14-128]

20 **Forner A**, Reig M, Bruix J. Hepatocellular carcinoma. *Lancet* 2018; **391**: 1301-1314 [PMID: 29307467 DOI: 10.1016/S0140-6736(18)30010-2]

21 **Pinyol R**, Montal R, Bassaganyas L, Sia D, Takayama T, Chau GY, Mazzaferro V, Roayaie S, Lee HC, Kokudo N, Zhang Z, Torrecilla S, Moeini A, Rodriguez-Carunchio L, Gane E, Verslype C, Croitoru AE, Cillo U, de la Mata M, Lupo L, Strasser S, Park JW, Camps J, Solé M, Thung SN, Villanueva A, Pena C, Meinhardt G, Bruix J, Llovet JM. Molecular predictors of prevention of recurrence in HCC with sorafenib as adjuvant treatment and prognostic factors in the phase 3 STORM trial. *Gut* 2018 [PMID: 30108162 DOI: 10.1136/gutjnl-2018-316408]

22 **Fu Q**, Yang F, Xiang T, Huai G, Yang X, Wei L, Yang H, Deng S. A novel microRNA signature predicts survival in liver hepatocellular carcinoma after hepatectomy. *Sci Rep* 2018; **8**: 7933 [PMID: 29785036 DOI: 10.1038/s41598-018-26374-9]

23 **McNeish DM**. Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. *Multivariate Behav Res* 2015; **50**: 471-484 [PMID: 26610247 DOI: 10.1080/00273171.2015.1036965]

24 **Tseng YY**, Moriarity BS, Gong W, Akiyama R, Tiwari A, Kawakami H, Ronning P, Reuland B, Guenther K, Beadnell TC, Essig J, Otto GM, O'Sullivan MG, Largaespada DA, Schwertfeger KL, Marahrens Y, Kawakami Y, Bagchi A. PVT1 dependence in cancer with MYC copy-number increase. *Nature* 2014; **512**: 82-86 [PMID: 25043044 DOI: 10.1038/nature13311]

25 **Zhou DD**, Liu XF, Lu CW, Pant OP, Liu XD. Long non-coding RNA PVT1: Emerging biomarker in digestive system cancer. *Cell Prolif* 2017; **50** [PMID: 29027279 DOI: 10.1111/cpr.12398]

26 **Wang F**, Yuan JH, Wang SB, Yang F, Yuan SX, Ye C, Yang N, Zhou WP, Li WL, Li W, Sun SH. Oncofetal long noncoding RNA PVT1 promotes proliferation and stem cell-like property of hepatocellular carcinoma cells by stabilizing NOP2. *Hepatology* 2014; **60**: 1278-1290 [PMID: 25043274 DOI: 10.1002/hep.27239]

27 **Xu Y**, Luo X, He W, Chen G, Li Y, Li W, Wang X, Lai Y, Ye Y. Long Non-Coding RNA PVT1/miR-150/ HIG2 Axis Regulates the Proliferation, Invasion and the Balance of Iron Metabolism of Hepatocellular Carcinoma. *Cell Physiol Biochem* 2018; **49**: 1403-1419 [PMID: 30205391 DOI: 10.1159/000493445]

28 **Ng SY**, Bogu GK, Soh BS, Stanton LW. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. *Mol Cell* 2013; **51**: 349-359 [PMID: 23932716 DOI: 10.1016/j.molcel.2013.07.017]

P- Reviewer: Fava G, Kornberg A, Sung WW

S- Editor: Gong ZM **L- Editor:** Wang TQ **E- Editor:** Yin SY

Baishideng®

Published By Baishideng Publishing Group Inc
7901 Stoneridge Drive, Suite 501, Pleasanton, CA 94588, USA
Telephone: +1-925-2238242
Fax: +1-925-2238243
E-mail: bpgoffice@wjgnet.com
Help Desk:http://www.f6publishing.com/helpdesk
http://www.wjgnet.com