

# A four-methylated mRNA signature-based risk score system predicts survival in patients with hepatocellular carcinoma

Yu Wang<sup>1</sup>, Zhiping Ruan<sup>1</sup>, Sizhe Yu<sup>1</sup>, Tao Tian<sup>1</sup>, Xuan Liang<sup>1</sup>, Li Jing<sup>1</sup>, Wenyuan Li<sup>1</sup>, Xiao Wang<sup>1</sup>, LCL Xiang<sup>1</sup>, F.X. Claret<sup>2</sup>, Kejun Nan<sup>1</sup>, Hui Guo<sup>1</sup>

<sup>1</sup>Department of Medical Oncology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, PR China

<sup>2</sup>Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

**Correspondence to:** Hui Guo; **email:** [guohuihappy97@163.com](mailto:guohuihappy97@163.com)

**Keywords:** score system, nomogram, prognosis, methylation, hepatocellular carcinoma

**Received:** September 16, 2018    **Accepted:** December 19, 2018    **Published:** January 10, 2019

**Copyright:** Wang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

## ABSTRACT

Evidence suggests that altered DNA methylation plays a causative role in the pathogenesis of various cancers, including hepatocellular carcinoma (HCC). Thus, methylated differently expressed genes (MDEGs) could potentially serve as biomarkers and therapeutic targets in HCC. In the present study, screening four genomics profiling datasets (GSE62232, GSE84402, GSE73003 and GSE57956) enabled us to identify a total of 148 MDEGs. A signature was then established based on the top four MDEGs (BRCA1, CAD, CDC20 and RBM8A). Taking clinical variables into consideration, we constructed a risk score system consisting of the four-MDEG signature and the patients' clinical features, which was predictive of prognosis in HCC. The prognostic value of the HCC risk score system was confirmed using TCGA HCC samples. The scores were then used to construct a nomogram, performance of which was evaluated using Harrel's concordance index (C-index) and a calibration curve. The signature-based nomogram for prediction of overall survival in HCC patients exhibited good performance and was superior to traditional staging systems (C-index: 0.676 vs 0.629,  $P < 0.05$ ). We have thus established a novel risk score system that is predictive of prognosis and is a potentially useful guide for personalized treatment of HCC patients.

## INTRODUCTION

Hepatocellular carcinoma (HCC) is the fifth most common cancer in China, where it is estimated to have killed 140 million people [1]. The leading cause of HCC is chronic infection with a hepatitis virus, alcohol abuse, exposure aflatoxin, tobacco smoking and diabetes [2]. Although there are a large number of studies examining HCC formation and progression, the precise mechanism underlying its pathogenesis remains unclear [3]. Moreover, the rate of early diagnosis of HCC is low; most patients are diagnosed with advanced disease. TNM stage at diagnosis is still regarded as the best predictor of survival [4]. However, because HCC is a

highly heterogeneous malignancy, the prognoses of patients with the same stage disease may differ due to inherent clinical and molecular diversities [5]. Therefore, new valid and reliable prognostic and predictive biomarkers are needed to improve risk prediction and offer better information for guiding personalized therapy.

Alterations in epigenetic modifications such as DNA methylation, histone acetylation and RNA interference are important heritable contributory factors in tumor development [6]. For example, altered DNA methylation is thought to contribute to the pathogenesis of a variety of cancers, including HCC [7]. However,

multiple studies indicate that a variety of genes are aberrantly hyper- or hypomethylated in HCC [8], but a comprehensive profile of the pathways within the interaction network remains to be elucidated. Several genes encoding epigenetic regulatory proteins, including EZH2 and HBV, have been shown to be involved in hepatocellular malignancy [9-10]. In addition, evidence now suggests that methylated mRNA may be a valid predictor of HCC [11]. But to the best of our knowledge, there are no prior studies examining methylated differentially expressed genes (MDEGs) on a genome-wide scale and focusing on predicting prognosis in HCC. In the present study, therefore, we comprehensively analyzed high-dimensional data from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) to build a novel MDEG-based risk score system that is predictive of prognosis and could potentially guide personalized therapy for HCC patients.

## RESULTS

### Identification and enrichment analysis of MDEGs in HCC

The flowchart for this study is shown in Figure 1. GSE62232, GSE84402, GSE73003 and GSE57956 comprised the training cohort downloaded from the GEO

database. The mRNA expression datasets GSE62232 and GSE84402 were calculated using the limma package in R (v 3.5.1). GSE62232 included 81 HCC and 10 normal liver samples, while, GSE84402 included 14 paired HCC and non-tumor samples (Affymetrix Human Genome U133 plus 2.0 platform). The GEO2R online analysis tool was used to calculate the datasets for the methylation difference profiles GSE73003 and GSE57956. The GSE73003 series consisted of 20 paired HCC and non-tumor samples, while GSE57956 consisted of 59 paired HCC and non-tumor samples (Illumina Human Methylation27 BeadChip). With cut-off criteria of  $P < 0.05$  and  $|\log_2FC| > 1$ , a total of 130 hypomethylation-high expression genes were detected by overlapping 3476 hypomethylated genes (4869 in GSE57956, 3748 in GSE73003) and 1945 upregulated genes (3972 in GSE62232, 3213 in GSE84402). Similarly, 18 hypermethylation-high expression genes were detected by overlapping 1689 hypermethylated genes (2651 in GSE67956, 1881 in GSE73003) and 338 downregulated genes (583 in GSE62232, 745 in GSE73003) (Figure 2A). To confirm that the  $P$  value and  $|\log_2FC|$  conform to logic using a different test, a representative volcano plot was constructed for GSE84402 (Figure 2B).

To obtain a deeper understanding of MDEGs, enrichment analysis with the Database for Annotation,

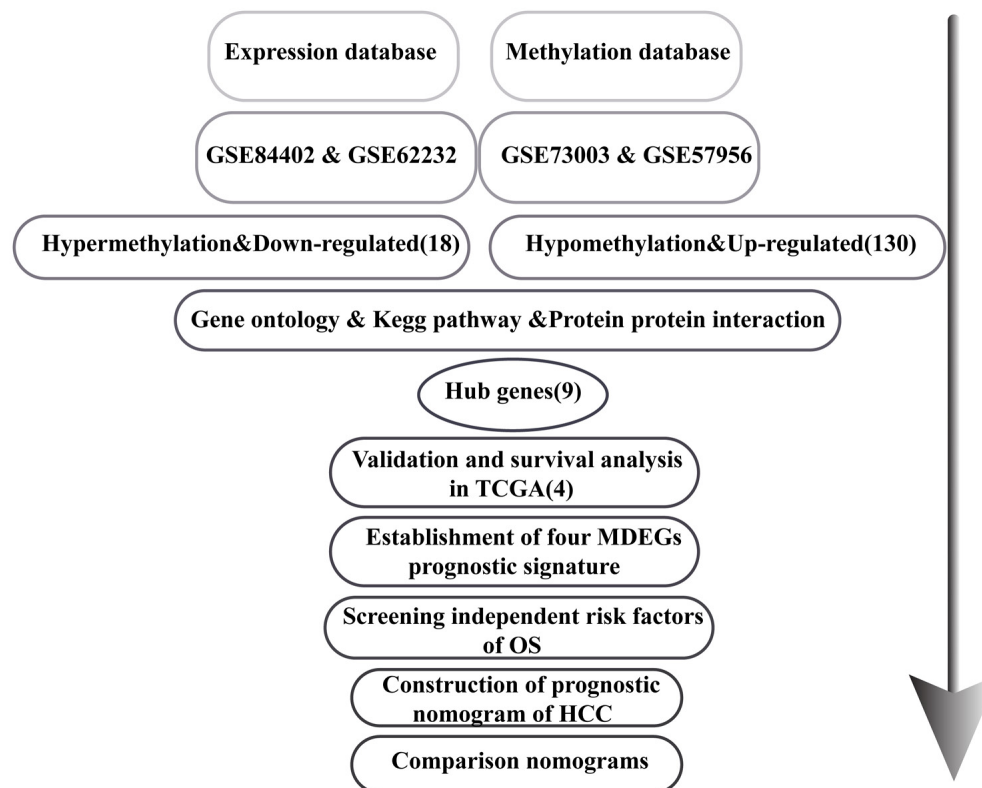
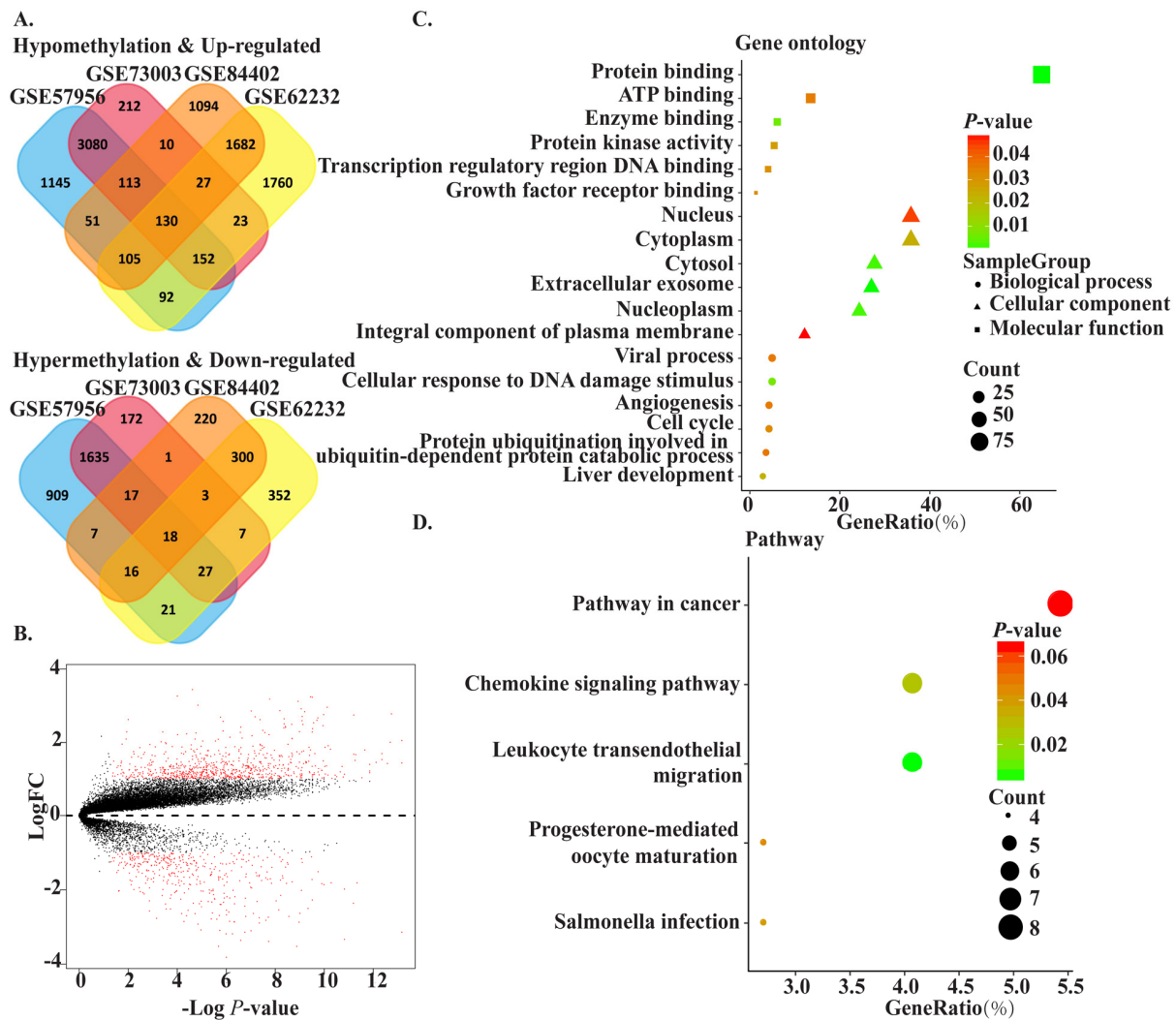


Figure 1. Flowchart of the study.

Visualization and Integrated Discovery (DAVID, <https://david.ncifcrf.gov/>) was used to elucidate biological function. The top significant terms emerging from the gene ontology enrichment analysis are shown in Figure 2C. MDEGs were enriched in “biological processes of cellular response to DNA damage stimulus,” “liver development,” “viral process,” “angiogenesis,” and “cell cycle.” Regarding molecular function, MDEGs showed enrichment in “protein binding,” “ATP binding,” “enzyme binding,” and “protein kinase activity.” Enrichment of cell components was mostly “nucleus region,” which suggests MDEGs may play an important role in transcription in HCC. Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis suggested that MDEGs were significantly enriched in pathways in “cancer,” “leukocyte transendothelial migration,” and “chemokine signaling pathway.” (Figure 2D).

### Identification of hub MDEGs and their clinical value in HCC

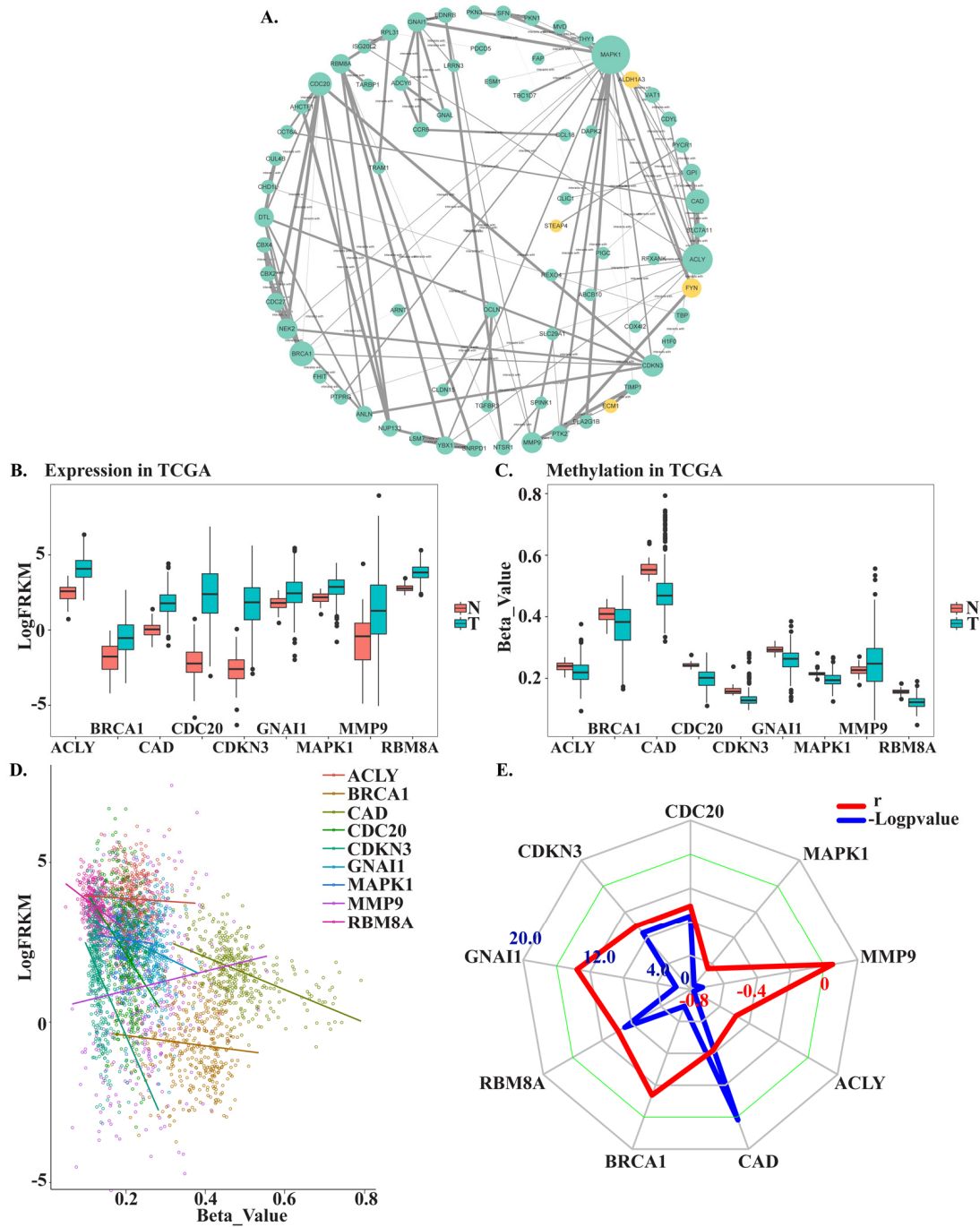
To identify the connections among MDEGs, a protein-protein interaction (PPI) network for MDEGs was constructed using STRING protein databases (Figure 3A). The top hub genes were CDC20, CDKN3, GNAI1, RBM8A, BRCA1, CAD, ACLY, MMP9, and MAPK1 based on a combined score >0.7 and connection numbers >8. To verify the hub genes, 371 HCC and 50 non-tumor samples were downloaded from TCGA as a validation cohort. Within this group, the expression and methylation values of most hub genes were consistent with the training group, with the exception of MMP9 (Figure 3B, C). We then further investigated the association between gene methylation and expression. The results showed a mild or moderate negative correlation,



**Figure 2. The methylated-differentially expressed genes identification and function.** (A) Venn of methylated-differentially expressed genes in gene expression datasets (GSE62232, GSE84402) and gene methylation datasets (GSE73003, GSE57956). (B) The volcano plot of GSE84402. Log<sub>2</sub> (FC) vs. -log<sub>10</sub> (p value) for differentially expressed mRNA. Red dot represents significant mRNA (log<sub>2</sub>|FC|>1, P<0.05). (C) The significant enriched gene ontology of MDEGs. (D) The significant enriched KEGG pathways of MDEGs.

which suggests methylation leads to decreased gene expression (Figure 3D-E). Expression of CDC20, CDKN3, GNAI1, RBM8A, BRCA1, and CAD showed a significant negative correlation with expression ( $p < 0.05$ ), whereas expression of ACLY, MMP9 and

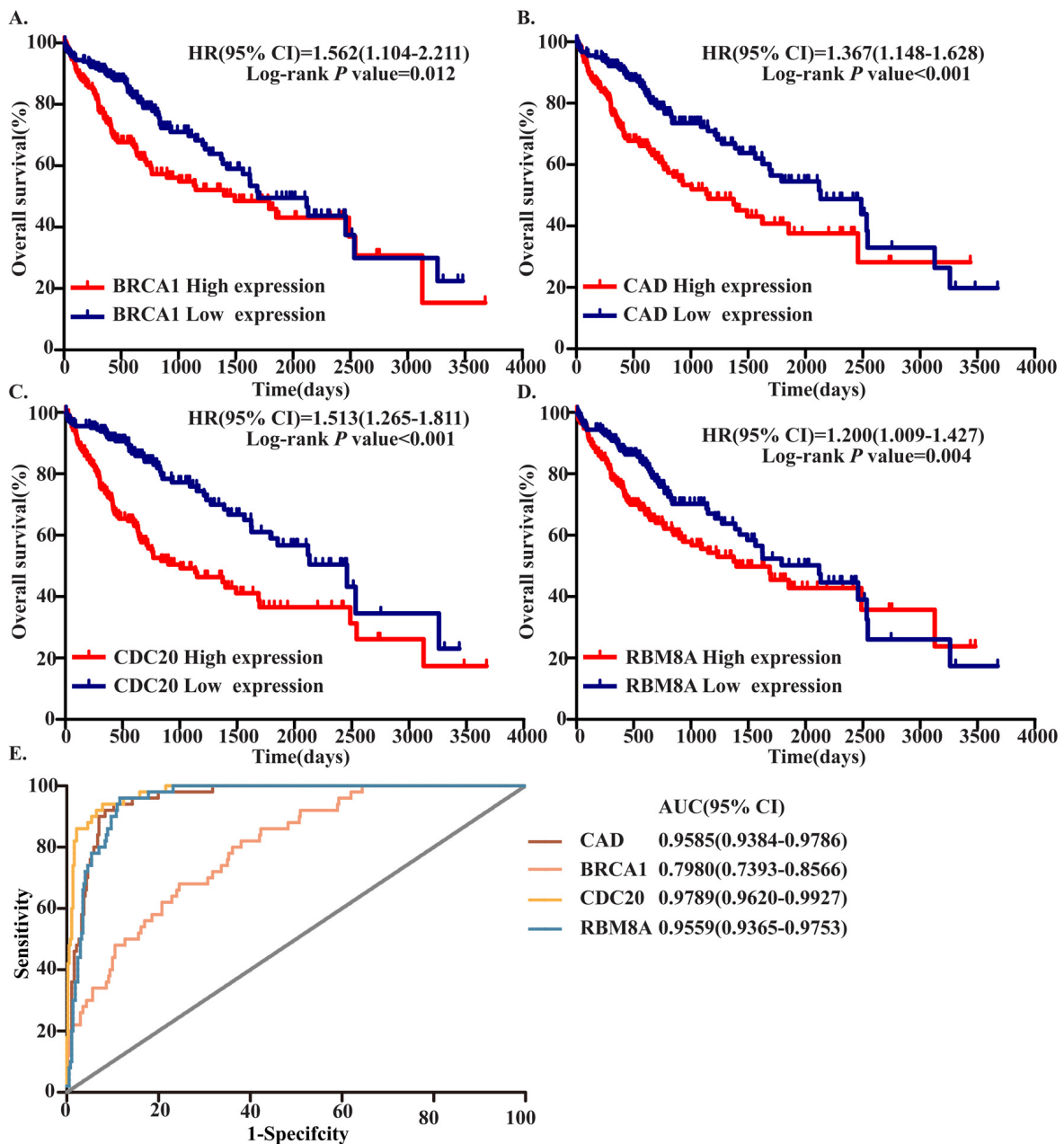
MAPK1 showed no correlation or was positively correlated. And results also were verified in MethHC, a database of DNA Methylation and gene expression in Human Cancer (<http://methhc.mbc.nctu.edu.tw/php/index.php>) (Supplementary Figure 1).



**Figure 3. Screening and verifying hub MDEGs.** (A) Protein-protein interaction network of MDEGs. Green dot represents hypo methylation-high expression gene. Yellow dot represents hyper methylation- low expression. The size of dot was decided by the connection degree of gene and the width of line between genes was decided by connectivity between two genes. (B) Expression of hub genes in TCGA. (C) Methylation of hub genes in TCGA. Beta-Value represents ratio of methylation. T represents tumor tissue, N represents normal tissue. (D) Correlation of expression and methylation of hub genes. (E) Radar map of hub genes correlation. Red line represents  $r$  and blue line represents  $-\text{Log}p\text{value}$ .

To identify hub MDEGs with potential prognostic value, we used the Kaplan-Meier method with the Log-rank test to evaluate the relation between expression of the aforementioned genes and the patients' overall survival (OS). Details of the clinical characteristics are presented in Supplementary Table 1. We found that OS was negatively related to expression of CDC20, RBM8A, BRCA1 and CAD, but had no relation with CDKN3 or GNAI1. Ultimately, the top four hub MDEGs were identified: CDC20, RBM8A, BRCA1 and

CAD. To further confirm the results, we verified the four hub genes in Gene Expression Profiling Interactive Analysis (GEPIA, <http://gepia.cancer-pku.cn/>) (Supplementary Figure 2). Receiver operating characteristic curve (ROC) analysis showed that all four of these genes have high sensitivity and specificity, which suggests high diagnostic value for distinguishing HCC patients from healthy individuals (Figure 4E). These four MDEGs may thus be useful as biomarkers for early diagnosis of HCC.



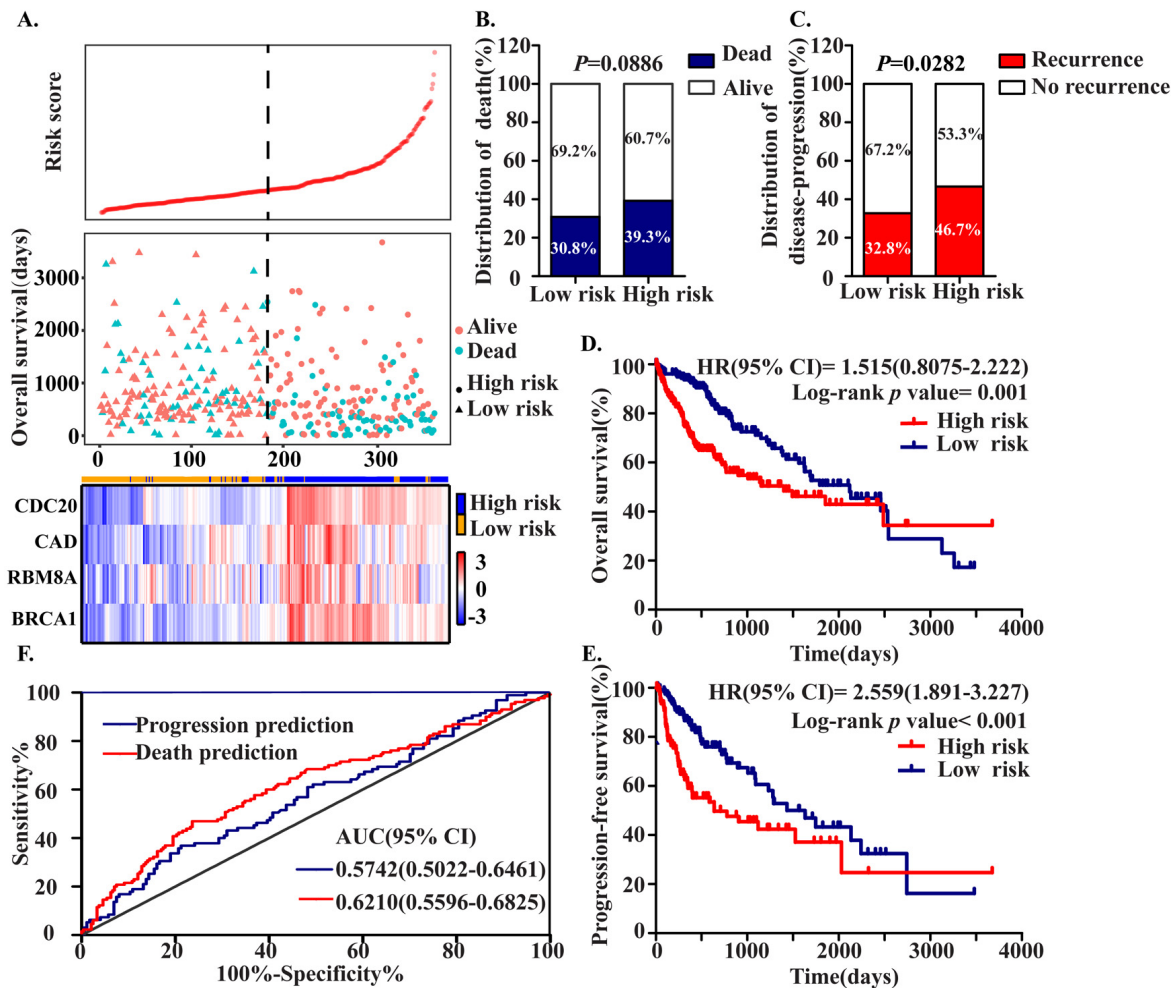
**Figure 4. Four hub MDEGs were associated with overall survival in HCC patients by using Kaplan-Meier curve and Log-rank test.** The patients were stratified into high expression group and low expression group according to median expression of each mRNA. (A) BRCA1; (B) CAD; (C) CDC20; (D) RBM8A. (E) ROC curves of the 4 hub MDEGs in HCC. The X axis shows false positive rate, presented as "1-Specificity". The Y axis indicates true positive rate, shown as "Sensitivity".

## Prognostic value of a four-MDEG signature risk score in HCC

To assess the prognostic value of CDC20, RBM8A, BRCA1 and CAD, we constructed a prognostic signature by integrating the expression of these four MDEGs using a regression coefficient. We then calculated a risk score for each patient and ranked them based on increasing score, after which patients were classified into a high-risk ( $n = 179$ ) or a low risk ( $n = 179$ ) group based on the median risk score. The risk score distribution, survival status, and expression profile of the four prognostic MDEGs are shown in Figure 5A. OS and progression-free survival (PFS) rates among patients were 60.7% and 53.3%, respectively, in the high-risk group, as compared to 69.2% and 67.2% in the low-risk group (Figure 5B, C). The hazard ratio (HR) of

high-risk group versus low-risk group was 1.515 for OS ( $P = 0.001$ , 95% confidence interval (CI) = 0.8075-2.222) and 2.559 for PFS ( $P < 0.001$ , 95%CI = 1.891-3.227). Thus, patients in the high-risk group had significantly poorer OS and PFS than patients in the low-risk group (Figure 5D, E).

A ROC analysis of the predictive efficiency of the four-MDEG signature suggested it had good performance with respect to both death and progression prediction (Figure 5F). Taking into consideration the patients' clinical features, including age, gender, clinical stage, T stage, grade, adjacent hepatic tissue inflammation, and HCC risk factors (virus infection, alcohol abuse, non-alcoholic fatty liver disease, hemochromatosis, alpha-1 antitrypsin deficiency), univariate and multivariate Cox regression analysis were used to assess the signature



**Figure 5. Construction of the Four MDEGs signature of HCC.** The patients were stratified into high risk group and low risk group based on median of risk score. (A) Risk score distribution of HCC patients, Survival status of each patient and Expression heatmap of the four hub MDEGs corresponding to each sample above. Red: high expression; Blue: low expression. (B, C) The distribution of death (B) and disease-progression (C) in high and low risk group. (D, E) Kaplan-Meier estimates of the overall survival (D) and progression-free survival (E) time of patients using the four MDEGs signature based risk score. (F) The ROC curve of the four MDEGs signature.

(high-risk vs. low-risk) with respect to OS. In the univariate analysis, clinical stage (HR = 2.229,  $P < 0.001$ ), T stage (HR = 2.534,  $P < 0.001$ ), HCC risk factors (HR = 0.631,  $P = 0.011$ ), fibrosis (HR = 0.542,  $P = 0.002$ ) and the four-MDEG signature (HR = 4.467,  $P < 0.001$ ) were all significantly associated with OS in HCC patients. To integrate all independent risk factors affecting OS for construction of a HCC prognostic nomogram, significant clinicopathological factors from the univariate analyses were entered into multivariate COX regression analyses. The results indicated that the four-MDEG signature (HR = 2.022,  $P < 0.001$ ) was a significant independent factor of OS, as were T stage (HR = 2.149,  $P < 0.001$ ) and HCC risk factors (HR = 0.651,  $P = 0.019$ ) (Table 1).

### Establishment of a nomogram for OS prediction in HCC

To provide a clinically associated quantitative method that could be used to predict the probabilities of 3- and 5-year OS in HCC, a prognostic nomogram was established in which the score integrated the three indepen-

dent prognostic factors, T stage, HCC risk factors and the four-MDEG signature (Figure 6A). Harrel's concordance index (C-index) for OS prediction was 0.676. The calibration curves for the nomogram for the 3- and 5- year OS rates showed good agreement between the prediction and the actual observation (Figure 6B). Each patient for whom there was complete clinical information about T stage, HCC risk factors, and the four-MDEG signature would obtain a Nomogram score reflecting total points. Using the Nomogram score, patients were divided into three risk groups based on the tertiles, which had cut-off values of 28.50 and 44.60. From KM analysis of the TCGA dataset, significant differences were observed between the high-, intermediate- and low-risk groups ( $P = 0.0003$ ) (Figure 6C).

### Comparison of predictive accuracy between the nomogram and a single independent factor

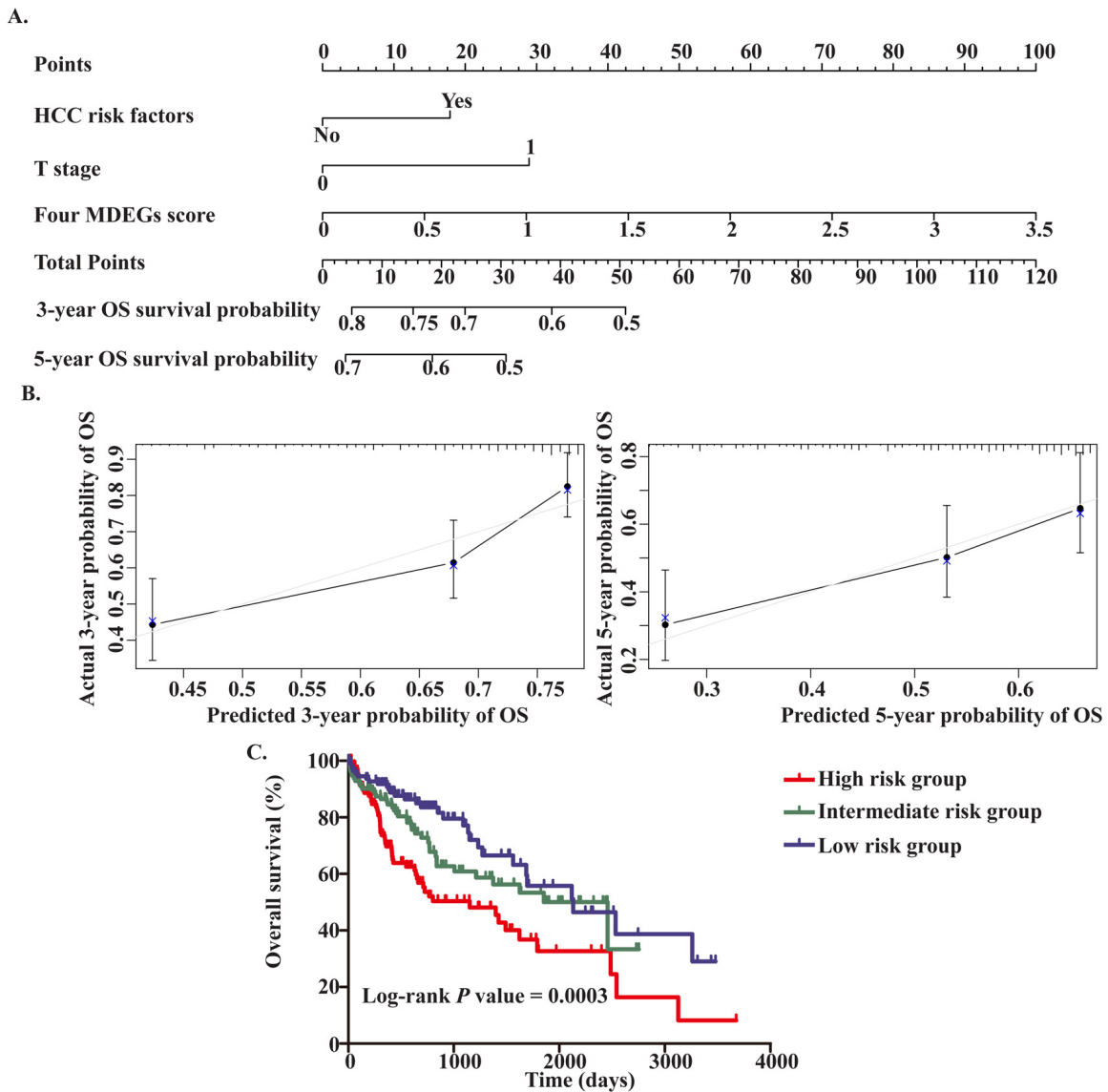
The TNM stage system is regarded as the best predictor of survival. Moreover, we found that T stage was an independent prognostic factor for OS in HCC. The predictive power of the nomogram for HCC prognosis

**Table 1. Univariate/multivariate COX regression analyses of clinicopathologic factors associated with OS.**

Variables	Univariate analysis		Multivariate analysis	
	HR(95%CI)	P	HR(95%CI)	P
Age( $\geq 65$ vs. $< 65$ )	1.265(0.893-1.791)	0.235		
Gender(Male vs. Female)	0.817(0.573-1.164)	0.262		
Clinical stage( III +IV vs. I+II)	2.229(1.559-3.188)	$<0.001^*$		
Grade(G3+G4 vs. G1+G2)	1.113(0.774-1.601)	0.564		
T stage(T3+T4 vs. T1+T2)	2.534(1.783-3.601)	$<0.001^*$	2.149(1.499-3.081)	$<0.001^*$
AFP( $<25$ ng/ml vs. $\geq 25$ ng/ml)	1.002(0.697-1.442)	0.991		
Adjacent hepatic tissue inflammation( Yes vs. No)	0.699(0.468-1.044)	0.699		
Fibrosis(Yes vs. No)	0.542(0.366-0.803)	0.002*		
Child-Pugh(A vs. B+C)	1.141(0.578-2.251)	0.703		
BMI( $\geq 25$ vs $< 25$ )	0.733(0.515-1.043)	0.084		
Family history(Yes vs. No)	1.225(0.858-1.748)	0.264		
HCC risk factors(Yes vs. No)	0.631(0.443-0.898)	0.011*	0.651(0.454-0.933)	0.019*
Four MDEGs signature (high risk vs. low risk)	4.467(1.995-10.002)	$<0.001^*$	2.022(1.486-2.753)	$<0.001^*$

Abbreviations: OS, overall survival; HR, hazard ratio; 95% CI, 95% confidence interval.

\*Statistically significant; AFP, Alpha-fetoprotein; BMI, body mass index.



**Figure 6. Establishment of the OS nomogram for HCC patients.** (A) Nomogram for predicting OS of HCC. There are three components in this nomogram: the four MDEGs score, HCC risk factor and T stage. Each of them generates points according to the line drawn upward. And the total points of the three components of an individual patient lie on "Total Points" axis which corresponds to the probability of 3-year and 5-year OS rate plotted on the two axes below. (B) Calibration plots of the nomogram for predicting OS rate at 3 year (Left) and 5 years (Right). The predicted and the actual probabilities of OS were plotted on the x- and y-axis, respectively. (C) Kaplan-Meier curves of three risk subgroups stratified by the total points the nomogram gives.

was compared with that of T stage. The C-index for OS prediction based on T stage was 0.629, which was significantly lower than the C-index for the nomogram (0.676,  $P < 0.05$ ). This suggests our nomogram is a more accurate predictor of OS in HCC than conventional staging systems and is potentially valuable for predicting survival of HCC patients.

## DISCUSSION

A valid and accurate molecule-based method for identifying patients who have a poor prognosis is ur-

gently needed to optimize their individual therapy. Therefore, effective and credible biomarkers and genetic signatures that can serve as prognostic predictors and treatment targets are critically needed for HCC. In the last decade, methylation has come to be recognized as an important epigenetic regulator of gene expression in eukaryotes, and it is now well established that methylation, especially DNA methylation, is crucially involved in multiple cancers, including HCC [12, 13]. Nishida showed that alterations in DNA methylation are a common feature of hepatocarcinogenesis [14]. Although some studies have identified



MDEGs in HCC [15, 16], their predictive value for HCC patients has not been systematically investigated until now. To our knowledge, this is the first study to develop a MDEG-based risk score that is predictive of prognosis in HCC.

We used methylation and expression microarrays with GEO databases to screen for MDEGs and were able to obtain a set of MDEGs through *in silico* analysis. Enrichment analysis of the MDEGs suggested they were involved in key biological processes, including DNA damage, viral processes, angiogenesis, and cell cycling. Given that hepatitis virus is a main cause of HCC, the presence of DNA damage due to integration of the virus genome into the host DNA is reasonable [17]. Also reasonable is the involvement of angiogenesis, since HCC is a highly vascular tumor. In addition, KEGG enrichments suggested that significantly enriched pathways include chemokine signaling pathways, which suggests that inflammation and immunity are critical factors in the pathogenesis, progression and metastasis of HCC [18]. Chemokine signaling reportedly influences HCC invasion and/or metastasis through effects on the tumor microenvironment [19, 20]. Evidence indicates, for example, that chemokines such as CCL5, CCL7, CXCL8 act via CCRs on myeloid-derived suppressor cells (MDSCs) to form an inhibitory tumor microenvironment that promotes tumor pathogenesis, progression and resistance [21-23].

Based on PPI analysis, we identified nine hub genes, which were verified and analyzed in a validation cohort from TCGA. Four of these hub genes, which were significantly associated with OS of HCC patients, were ultimately selected. Earlier research has consistently demonstrated that upregulation of BRCA1, CDC20, RBM8A and CAD promotes progression, invasion, metastasis and chemoresistance in HCC [24-27]. While the efficacy of any single marker is limited, a multiple-marker signature could have greater diagnostic and prognostic value. We therefore constructed a four-MDEG signature that was an independent prognostic factor for HCC patients. This signature was predictive of both OS and PFS. If applied, relatively minor examination using our risk score system could help identify high- and low-risk HCC patients and provide useful information that could aid in selecting a therapeutic strategy.

To increase the accuracy of the prediction of prognosis, both genetic and clinically-related variables were integrated into the nomogram. Ultimately, the OS nomogram included the four-MDEG signature, T stage and HCC risk factors. The nomogram for HCC performed well when used to predict OS, and its pre-

dictive ability was verified using a C-index and a calibration curve. Indeed, the nomogram provides greater predictive accuracy for OS than traditional systems. As regards the prognostic signature or nomogram, if we are able to put it into clinical practice in the future, we anticipate being able to identify patients at high-risk of cancer-related death before treatment, and recommend a more aggressive therapeutic strategies with dynamic surveillance. However, there are limitations to our study. First, whether the prognostic signature or nomogram can be applied to patients must be confirmed in larger groups of HCC patients. Second, the molecular mechanism of the four MDEGs in HCC remains to be explored further.

In summary, we have developed a novel four-MDEG expression-based risk score system for objectively and accurately predicting survival and prognosis in HCC patients. In addition, the MDEG signature could also shed new light on the role of methylation in the pathogenesis and progression of HCC, which may provide information helpful for selection of therapeutic strategies. The four MDEGs could potentially serve as biomarkers and therapeutic targets for dynamic surveillance and treatment of HCC patients.

## MATERIALS AND METHODS

### Data processing

The raw data and clinical information were download from the GEO (<https://www.ncbi.nlm.nih.gov/geo/>) and TCGA (<https://cancergenome.nih.gov/>). Gene methylation profiling of the GSE73003 and GSE57956 datasets was conducted using the GPL8490 platform (Illumina Human Methylation27 BeadChip), which included 27,578 highly informative CpG sites and more than 14476 genes (<http://www.illumina.com/pages.ilmn?ID=243>). Gene expression profiling of the GSE84402 and GSE62232 datasets was conducted using the GPL570 platform (Affymetrix Human Genome U133 plus 2.0 Array), which included 54675 unique probes and tested more than 23517 genes (<http://www.affymetrix.com/support/technical/byproduct.affx?product=hg-u133-plus>). The GSE73003 series consisted of 20 paired HCC and non-tumor samples. The GSE57956 series consisted of 59 paired HCC and non-tumor samples. The GSE84402 series included 14 paired HCC and non-tumor samples. And the GSE62232 series included 81 HCC and 10 normal liver samples. From TCGA, we downloaded 371 HCC and 50 non-tumor samples. The mRNA-seq data were preprocessed and submitted for analysis as the upper quantile normalized FPKM values. GEO2R was used to screen for genes differentially methylated between tumor and non-tumor samples. The differentially

expressed genes were identified using the limma package in R. Values of  $P < 0.05$  and  $|FC| \geq 1$  were considered significant. Finally, hypomethylation-high expression genes were detected by overlapping hypomethylated and upregulated genes; similarly, hypermethylation-low expression genes were detected by overlapping hypermethylated and downregulated genes.

### Functional and pathway enrichment analysis

Functional annotations in MDEGs were done using The Database for Annotation, Visualization and Integrated Discovery (DAVID; <https://david.ncifcrf.gov/>), which enriched gene oncology and pathways. Gene oncology involved three categories: cellular components, molecular function, and biological processes. Pathway enrichment was carried out using the Kyoto Encyclopedia of Genes and Genomes (KEGG, <https://www.kegg.jp/>), which contains information about genomes, biological pathways, diseases, and chemical substances. The criterion for significant enrichment of biological processes and pathways was  $P = 0.05$ .

### Hub MDEG screening and verification

STRING protein databases (<https://string-db.org/>) were used to evaluate interactive relationships among the MDEGs. We used Cytoscape software to construct a network based on the STRING results. Combined scores  $> 0.7$  and connection numbers  $> 8$  were deemed to indicate hub genes. To confirm the results, the hub MDEGs were validated in TCGA. The Pearson correlation test was used to assess the relationship between hub gene methylation and expression in HCC.

### Formulation of MDEG signatures and association of signatures and clinical features

ROC curve analysis was used to evaluate the diagnostic effectiveness of hub MDEGs. The prognostic value of hub MDEGs were evaluated using the Kaplan-Meier method with the Log-rank test. Hub MDEGs related to OS were considered to be prognostic. Using the combination of weighted MDEG expression values, independent hub MDEG biomarkers were integrated into a MDEG signature using a risk scoring method as shown in the following equation: Risk Score (patient) =  $\sum_{i=1}^n \frac{\text{expression}(mRNA_i)}{\text{median}} * \text{coefficient}(mRNA_i)$ . Here, Risk Score (patient) is a MDEG signature risk score for a HCC patient. In addition,  $mRNA_i$  represents the  $i$ th prognostic mRNA, while expression ( $mRNA_i$ ) is the expression value of  $mRNA_i$  for the patient. *Coefficient* ( $mRNA_i$ ) is the regression coefficient of  $mRNA_i$ , which represents the contribution of  $mRNA_i$  to the prognostic risk score. Based on the risk score, patients can be assigned to a high-risk or low-risk group. Subsequently,

a risk score system was constructed, and the median risk score was regarded as the cutoff point. HCC patients were then divided into high- and low-risk groups. Kaplan-Meier survival curves were calculated to compare survival and recurrence risk between the high- and low-risk groups.

### Statistical analysis

To identify independent predictors of OS in HCC, univariate Cox regression analysis was performed to evaluate the prognostic value of signatures with a threshold value of 0.2. Multivariate Cox regression analyses were conducted using Forward LR. Hazard ratios (HRs) and 95% confidence intervals (CIs) were computed based on the Cox regression analysis. A nomogram was constructed based on the results of the multivariate Cox regression analyses using rms version 3.5.1 (<http://www.r-project.org/>). The performance of the nomogram was assessed using Harrel's concordance index (C-index) and comparing the predicted and actual probabilities for OS. Bootstraps with 1,000 resamples were used for these activities. Comparisons between the nomogram and other staging systems were made using the rcorr.cens package in Hmisc and were evaluated using the C-index. Each patient received the total points from the nomogram (Nomo-score). KM curve analysis was performed to evaluate the performance of the nomogram by dividing patients into high-, intermediate- and low-risk groups using tertiles of the Nomo-scores as cut-off points. Values of  $P < 0.05$  were considered significant. Statistical analysis was performed using the IBM SPSS Statistics software program version 22.0 (IBM Corp., NY, USA).

### Abbreviations

MDEG: methylated differently expressed gene; HCC: hepatocellular carcinoma; OS: overall survival; PFS: progression-free survival; HR: hazard ratio; CI: confidence interval; KM curve: Kaplan-Meier curve.

### AUTHOR CONTRIBUTIONS

Hui Guo and Yu Wang designed the research and performed analysis; Zhiping Ruan, Sizhe Yu, Tao Tian and Xuan Liang collected and analyzed data and constructed figures; Li Jing, Wenyuan Li, Xiao Wang, LCL Xiang performed statistical analyses; F.X. Claret and Kejun Nan guided experiments; Hui Guo and Yu Wang drafted and revised the manuscript.

### ACKNOWLEDGEMENTS

This work benefited from the Gene Expression Omnibus (GEO) database and the Cancer Genome Atlas

(TCGA) database. We were grateful to the access to the resources and the efforts of the staff to expand and improve the two databases.

## CONFLICTS OF INTEREST

The authors declare that there are no financial or other relationships that might lead to a conflict of interest of the present article.

## FUNDING

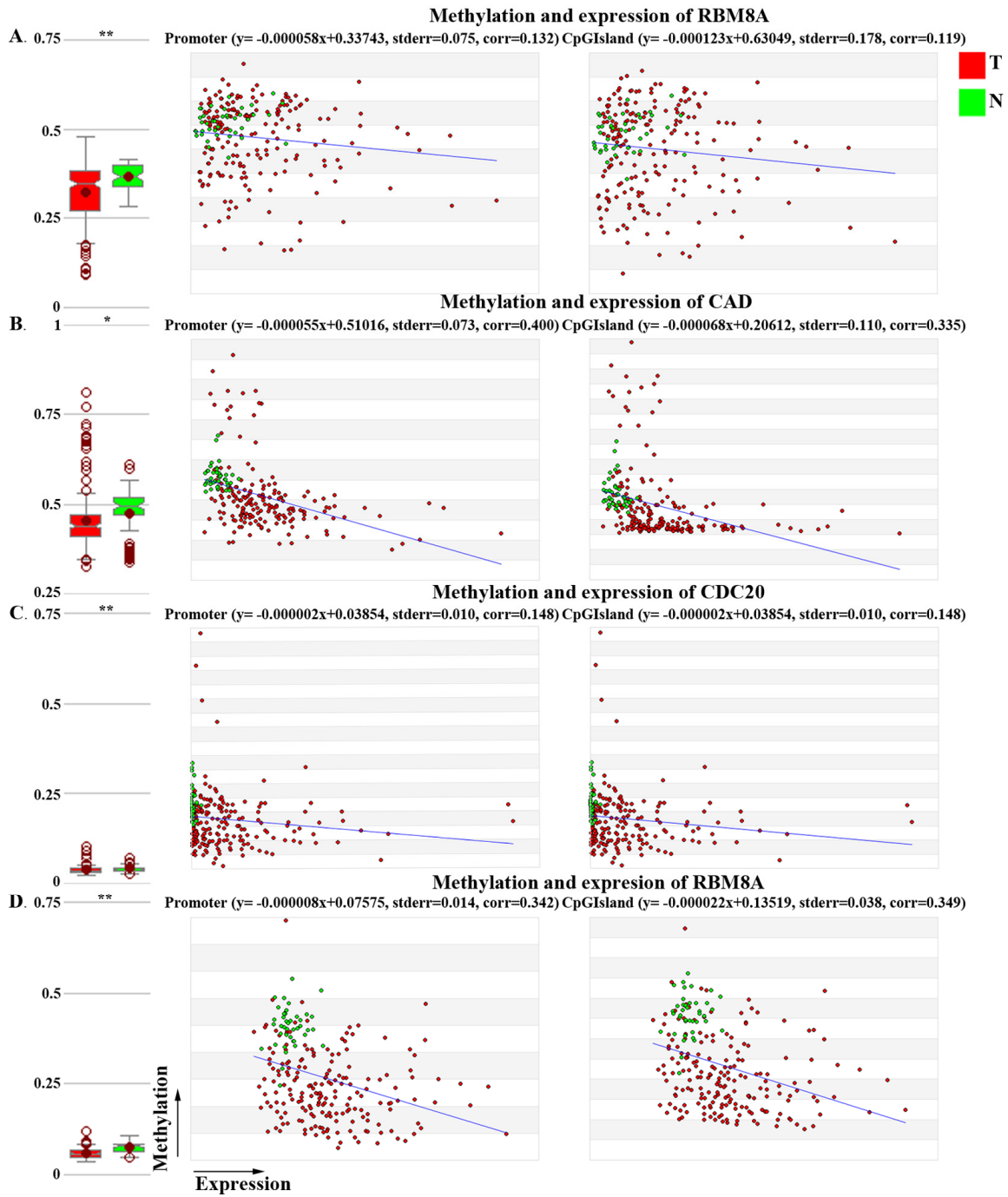
This study was supported in part by grant 81672432 from the National Natural Science Foundation of China (to H.G).

## REFERENCES

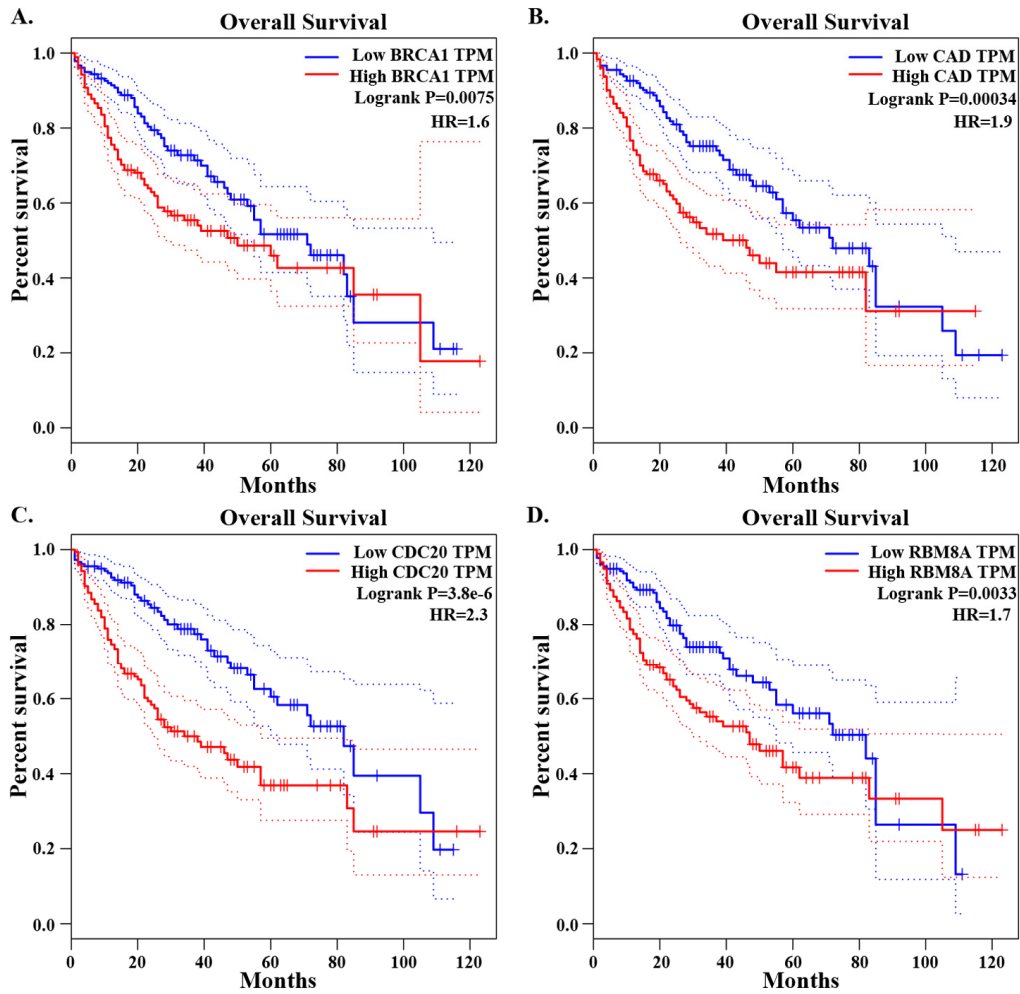
1. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. *CA Cancer J Clin.* 2017; 67:7–30. <https://doi.org/10.3322/caac.21387>
2. Forner A, Llovet JM, Bruix J. Hepatocellular carcinoma. *Lancet.* 2012; 379:1245–55. [https://doi.org/10.1016/S0140-6736\(11\)61347-0](https://doi.org/10.1016/S0140-6736(11)61347-0)
3. Guo H, Jing L, Cheng Y, Atsaves V, Lv Y, Wu T, Su R, Zhang Y, Zhang R, Liu W, Rassidakis GZ, Wei Y, Nan K, Claret FX. Down-regulation of the cyclin-dependent kinase inhibitor p57 is mediated by Jab1/Csn5 in hepatocarcinogenesis. *Hepatology.* 2016; 63:898–913. <https://doi.org/10.1002/hep.28372>
4. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, Meyer L, Gress DM, Byrd DR, Winchester DP. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin.* 2017; 67:93-99. <https://doi.org/10.3322/caac.21388>
5. Bruix J, Gores GJ, Mazzaferro V. Hepatocellular carcinoma: clinical frontiers and perspectives. *Gut.* 2014; 63:844–55. <https://doi.org/10.1136/gutjnl-2013-306627>
6. Chen Z, Li S, Subramaniam S, Shyy JY, Chien S. Epigenetic Regulation: A New Frontier for Biomedical Engineers. *Annu Rev Biomed Eng.* 2017; 19:195–219. <https://doi.org/10.1146/annurev-bioeng-071516-044720>
7. Nordlund J, Syvänen AC. Epigenetics in pediatric acute lymphoblastic leukemia. *Semin Cancer Biol.* 2018; 51:129–38. <https://doi.org/10.1016/j.semcancer.2017.09.001>
8. Sang L, Wang XM, Xu DY, Zhao WJ. Bioinformatics analysis of aberrantly methylated-differentially expressed genes and pathways in hepatocellular carcinoma. *World J Gastroenterol.* 2018; 24:2605–16. <https://doi.org/10.3748/wjg.v24.i24.2605>
9. Cai MY, Tong ZT, Zheng F, Liao YJ, Wang Y, Rao HL, Chen YC, Wu QL, Liu YH, Guan XY, Lin MC, Zeng YX, Kung HF, Xie D. EZH2 protein: a promising immunomarker for the detection of hepatocellular carcinomas in liver needle biopsies. *Gut.* 2011; 60:967–76. <https://doi.org/10.1136/gut.2010.231993>
10. Lee SM, Lee YG, Bae JB, Choi JK, Tayama C, Hata K, Yun Y, Seong JK, Kim YJ. HBx induces hypomethylation of distal intragenic CpG islands required for active expression of developmental regulators. *Proc Natl Acad Sci USA.* 2014; 111:9555–60. <https://doi.org/10.1073/pnas.1400604111>
11. Xu RH, Wei W, Krawczyk M, Wang W, Luo H, Flagg K, Yi S, Shi W, Quan Q, Li K, Zheng L, Zhang H, Caughey BA, et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat Mater.* 2017; 16:1155–61. <https://doi.org/10.1038/nmat4997>
12. Casadevall D, Kilian AY, Bellmunt J. The prognostic role of epigenetic dysregulation in bladder cancer: A systematic review. *Cancer Treat Rev.* 2017; 61:82–93. <https://doi.org/10.1016/j.ctrv.2017.10.004>
13. Hlady RA, Sathyanarayan A, Thompson JJ, Zhou D, Wu Q, Pham K, Lee JH, Liu C, Robertson KD. Integrating the Epigenome to Identify Drivers of Hepatocellular Carcinoma. *Hepatology.* 2019. <https://doi.org/10.1002/hep.30211>
14. Nishida N, Nishimura T, Nakai T, Chishina H, Arizumi T, Takita M, Kitai S, Yada N, Hagiwara S, Inoue T, Minami Y, Ueshima K, Sakurai T, Kudo M. Genome-wide profiling of DNA methylation and tumor progression in human hepatocellular carcinoma. *Dig Dis.* 2014; 32:658–63. <https://doi.org/10.1159/000367982>
15. Sun XJ, Wang MC, Zhang FH, Kong X. An integrated analysis of genome-wide DNA methylation and gene expression data in hepatocellular carcinoma. *FEBS Open Bio.* 2018; 8:1093–103. <https://doi.org/10.1002/2211-5463.12433>
16. Liu S, Miao C, Liu J, Wang CC, Lu XJ. Four differentially methylated gene pairs to predict the prognosis for early stage hepatocellular carcinoma patients. *J Cell Physiol.* 2018; 233:6583–90. <https://doi.org/10.1002/jcp.26256>
17. Lee WY, Bachtiar M, Choo CC, Lee CG. Comprehensive review of Hepatitis B Virus-associated hepatocellular carcinoma research through text mining and big data analytics. *Biol Rev Camb Philos Soc.* 2018. <https://doi.org/10.1111/brv.12457>
18. Yu S, Wang Y, Jing L, Claret FX, Li Q, Tian T, Liang X,

- Ruan Z, Jiang L, Yao Y, Nan K, Lv Y, Guo H. Autophagy in the “inflammation-carcinogenesis” pathway of liver and HCC immunotherapy. *Cancer Lett.* 2017; 411:82–89. <https://doi.org/10.1016/j.canlet.2017.09.049>
19. Zhu Y, Gao XM, Yang J, Xu D, Zhang Y, Lu M, Zhang Z, Sheng YY, Li JH, Yu XX, Zheng Y, Dong QZ, Qin LX. C-C chemokine receptor type 1 mediates osteopontin-promoted metastasis in hepatocellular carcinoma. *Cancer Sci.* 2018; 109:710–23. <https://doi.org/10.1111/cas.13487>
20. Liu J, Chen S, Wang W, Ning BF, Chen F, Shen W, Ding J, Chen W, Xie WF, Zhang X. Cancer-associated fibroblasts promote hepatocellular carcinoma metastasis through chemokine-activated hedgehog and TGF- $\beta$  pathways. *Cancer Lett.* 2016; 379:49–59. <https://doi.org/10.1016/j.canlet.2016.05.022>
21. Thakur A, Schalk D, Tomaszewski E, Kondadasula SV, Yano H, Sarkar FH, Lum LG. Microenvironment generated during EGFR targeted killing of pancreatic tumor cells by ATC inhibits myeloid-derived suppressor cells through COX2 and PGE2 dependent pathway. *J Transl Med.* 2013; 11:35. <https://doi.org/10.1186/1479-5876-11-35>
22. Qin Y, Vasilatos SN, Chen L, Wu H, Cao Z, Fu Y, Huang M, Vlad AM, Lu B, Oesterreich S, Davidson NE, Huang Y. Inhibition of histone lysine-specific demethylase 1 elicits breast tumor immunity and enhances antitumor efficacy of immune checkpoint blockade. *Oncogene.* 2018. <https://doi.org/10.1038/s41388-018-0451-5>. Epub ahead of print.
23. Kumar D, Xu ML. Microenvironment Cell Contribution to Lymphoma Immunity. *Front Oncol.* 2018; 8:288. <https://doi.org/10.3389/fonc.2018.00288>
24. Liao Y, Yuan S, Chen X, Zhu P, Li J, Qin L, Liao W. Up-regulation of BRCA1-associated RING Domain 1 Promotes Hepatocellular Carcinoma Progression by Targeting Akt Signaling. *Sci Rep.* 2017; 7:7649. <https://doi.org/10.1038/s41598-017-07962-7>
25. Cui X, Li X, Li J, Wang X, Sun W, Cheng Z, Ding J, Wang H. inFRank: a ranking-based identification of influential genes in biological networks. *Oncotarget.* 2017; 8:43810–21. <https://doi.org/10.18632/oncotarget.11878>
26. Liang R, Lin Y, Ye JZ, Yan XX, Liu ZH, Li YQ, Luo XL, Ye HH. High expression of RBM8A predicts poor patient prognosis and promotes tumor progression in hepatocellular carcinoma. *Oncol Rep.* 2017; 37:2167–76. <https://doi.org/10.3892/or.2017.5457>
27. Zhai X, Zhu H, Wang W, Zhang S, Zhang Y, Mao G. Abnormal expression of EMT-related proteins, S100A4, vimentin and E-cadherin, is correlated with clinicopathological features and prognosis in HCC. *Med Oncol.* 2014; 31:970. <https://doi.org/10.1007/s12032-014-0970-z>

SUPPLEMENTARY MATERIAL



**Supplementary Figure 1. Methylation and correlation with expression of the four hub genes. (A) BRCA1 (B) CAD (C) CDC20 (D) RBM8A** Red dot represents tumor sample; green dot represents normal sample. Horizontal axis is expression and vertical axis is methylation. \*:  $P < 0.05$ ; \*\*:  $P < 0.005$ .



**Supplementary Figure 2. Relation between expression of the four hub genes and OS. (A) BRCA1 (B) CAD (C) CDC20 (D) RBM8A.**

## SUPPLEMENTARY TABLE

Please browse the link in Full Text version to find the data of **Supplementary Table 1**. The **clinical characteristics** of TCGA.