

RESEARCH

Open Access



Screening the genome for HCC-specific CpG methylation signatures as biomarkers for diagnosis and prognosis evaluation

Rui-kun Zhang¹ and Jia-lin Liu^{2*}

Abstract

Background: Hepatocellular carcinoma (HCC) is one of the most common and invasive malignant tumors in the world. The change in DNA methylation is a key event in HCC.

Methods: Methylation datasets for HCC and 17 other types of cancer were downloaded from The Cancer Genome Atlas (TCGA). The CpG sites with large differences in methylation between tumor tissues and paracancerous tissues were identified. We used the HCC methylation dataset downloaded from the TCGA as the training set and removed the overlapping sites among all cancer datasets to ensure that only CpG sites specific to HCC remained. Logistic regression analysis was performed to select specific biomarkers that can be used to diagnose HCC, and two datasets—GSE157341 and GSE54503—downloaded from GEO as validation sets were used to validate our model. We also used a Cox regression model to select CpG sites related to patient prognosis.

Results: We identified 6 HCC-specific methylated CpG sites as biomarkers for HCC diagnosis. In the training set, the area under the receiver operating characteristic (ROC) curve (AUC) for the model containing all these sites was 0.971. The AUCs were 0.8802 and 0.9711 for the two validation sets from the GEO database. In addition, 3 other CpG sites were analyzed and used to create a risk scoring model for patient prognosis and survival prediction.

Conclusions: Through the analysis of HCC methylation datasets from the TCGA and Gene Expression Omnibus (GEO) databases, potential biomarkers for HCC diagnosis and prognosis evaluation were ascertained.

Background

Hepatocellular carcinoma (HCC) is one of the most common and invasive malignant tumors in the world. It is the most common primary malignant tumor of the liver and the third most common cause of cancer-related death worldwide [1–3]. In China, liver cancer is the fourth most common type of cancer and the third most common cause of cancer-related death [4]. The incidence rate of HCC is increasing steadily [5]. According to the

Surveillance, Epidemiology, and End Results (SEER) program, the HCC incidence rate increased by 3.1% annually from 2008 to 2012. The incidence rates for males and females were 11.5 and 3.9 per 100,000. Likewise, the mortality rates for males and females with HCC increased by 2.8% and 3.4%, respectively, occupying the top spot for the deadliest cancer during the period [6]. In 2012 alone, there were 782,000 new HCC cases and 745,000 HCC-related deaths, accounting for 9.1% of all cancer fatalities. It is estimated that by 2025, more than a million cases of HCC will be diagnosed annually [7, 8].

According to estimations, more than 50% of newly diagnosed HCC cases occur in China [8]. More than half of HCC-related fatalities occur in China [9]. The primary risk factors for HCC include hepatitis B virus (HBV)

*Correspondence: szliujialin@126.com

² Department of Hepatobiliary and Pancreatic Surgery, Shenzhen Traditional Chinese Medicine Hospital, No.1 Fuhua Road, Shenzhen 518000, Guangdong, China
Full list of author information is available at the end of the article



infection, hepatitis C virus (HCV) infection, non-alcohol-related steatohepatitis, exposure to aflatoxin and chronic alcohol poisoning [10, 11]. HBV and HCV infection are the primary reasons for chronic liver disease and HCC in China. The incidence rates of HCV-related chronic liver cancer and HCC are also increasing [12, 13].

Most HCC patients do not display significant symptoms in early disease stages, so they are often diagnosed after the disease reaches stage III or IV. Liver resection, liver transplantation, radiofrequency ablation, transarterial chemoembolization and sorafenib are the primary clinical treatments for HCC [14]. While progress has been made in HCC clinical treatment and medical management, since most patients undergoing these treatments have late-stage cancer, the overall prognosis for HCC patients remains unsatisfactory, with a 17% 5-year survival rate [15, 16].

Alpha-fetoprotein (AFP) is a commonly used cancer marker for early screening, diagnosis and treatment evaluation. However, due to its limited sensitivity and specificity, this serum biomarker is of limited usefulness in the treatment of early-stage cancer patients [17–20]. Therefore, the American Association for the Study of Liver Diseases (AASLD) and European Association for the Study of the Liver (EASL) do not mandate using AFP for HCC diagnosis. Thus, finding a better biomarker for HCC diagnosis is an extremely important research direction.

Changes in DNA methylation are important epigenetic events for cancer. The term “epigenetic” refers to changes in gene expression mediated by mechanisms other than alterations in the primary nucleotide sequence of a gene [21, 22]. DNA methylation refers to reactions using *S*-adenosyl-methionine as a methyl donor that are catalyzed by enzymes called DNA methyltransferases (DNMTs), which add methyl groups to the cytosine ring to form methyl cytosines [23]. In mammalian gene sets, methylation only occurs at the 5' position of cytosines adjacent to guanines in CpG sites [21]. CpG islands in the gene promoter regions are the most common sites. High levels of methylation at these promoter regions represent the most evident epigenetic changes. Every type of human tumor is related to the methylation of genes [24, 25]. A high level of methylation in the CpG island in the promoter region stops the combination of DNA polymerase and transcription factors. Thus, the transcription of the target gene is inhibited [26]. For example, DNA methylation can deactivate tumor suppressor genes and DNA repair genes by reducing its transcriptional activity and can reduce the expression of E-cadherin. These effects allow the development of tumors [27]. Manel Esteller et al. [28] summarized the relationship between gene methylation and various types of cancer, such as

bladder cancer, cervical cancer, melanoma and glioma, with p16 methylation as a trait; high levels of p14 and APC methylation are common in gastrointestinal cancer (such as colon cancer and stomach cancer), while GSTP1 demonstrates high methylation levels in breast cancer and prostate cancer.

In HCC, O⁶-methylguanine-DNA-methyltransferase (MGMT) is another important DNA repair gene, and it is most active in the liver. The methylation of promoter region CpG islands reduces or eliminates MGMT expression in HCC cases [29, 30]. A total of 67.86% of HCC cases exhibit a high level of p73 gene methylation. The inactivation of p16INK4a caused by the methylation of promoters is also one of the main causes of HCC [27].

The methylation of DNA can be detected in phlegm, bronchial irrigation fluid, urine, blood, catheter fluid, lymph nodes and cancer tissues [22]. Thus, DNA methylation has great potential as a biomarker for the early diagnosis of cancer. A high level of SEPT9 gene methylation can be used to diagnose colorectal cancer [31]. A prognostic model of 7 CpG sites has already been applied in oral squamous cell carcinoma (OSCC) [32]. There have been reports of six methylation biomarkers that can be used to differentiate HCC patients and healthy humans [33]. Xu et al. [34] discovered 10 types of methylation biomarkers that can be used to diagnose HCC, but previous research lacked specificity. These biomarkers cannot be used to completely differentiate HCC and other types of cancer. In addition, the causal relationship between DNA methylation status and HCC outcome is still not clear. Thus, we integrated The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) Illumina 450 K DNA methylation datasets for HCC for this research to identify potential biomarkers for HCC diagnosis and prognosis.

Methods

Data preparation

We downloaded data for 377 cases (with paired tumor and paracancerous tissues) of HCC. The data included methylation data based on Illumina Human Methylation 450 BeadChip's tertiary methylation datasets, genetic expression data and relevant clinical data from <https://gdc-portal.nci.nih.gov/>. The paired tumor and paracancerous tissues were used to analyze different methylation sites. These data were collectively used as the training set.

Two DNA methylation datasets (GSE157341 and GSE54503) and the relevant clinical data were downloaded as the validation set. We downloaded methylation data for 17 other cancers from the following TCGA datasets: BLCA, BRCA, CESC, CHOL, COAD, ESCA, KIRC, KIRP, LUAD, LUSC, PAAD, PCPG, PRAD, READ,

SKCM, THCA and UCEC. The methylation datasets for 15 other types of cancer were not chosen due to the lack of paired paracancerous tissue samples. The beta (β) value of the methylation level is expressed as $M/(M+U+100)$. M is the number of Infinium II probes (methylated), while U is the number of Infinium I probes (nonmethylated).

Data preprocessing

R (v 4.0.3) was used to import the various methylation datasets, genetic data, and clinical data. The Impute package was used to fill in the missing values for methylation data. The Champ package [35] and minfi package [36] were used to filter (probes with detection p values > 0.01 , probes with < 3 beads in at least 5% of samples per probe, non-CpG probes contained in the current dataset, SNP-related probes, multihit probes, the cross-reactive probes identified by Chen et al. [37], and probes located on the X and Y chromosomes were discarded) and normalize that data. Low-quality probe signal sites were removed. Principal component analysis was applied to remove outliers. The Limma package [38] was used to log₂-transform and normalized the gene expression data. Samples without survival time information were excluded.

DNA differential methylation analysis

In the HCC dataset (referred to as the LIHC dataset) downloaded from the TCGA, 59 patients had matching cancer tissues and paracancerous tissues. After filtering the data and setting the 59 pairs of matching cancer and paracancerous tissues as the discovery set, the differentially methylated sites were identified. The difference between the methylation levels of tumor samples and normal tissues was calculated as $\beta_{\text{tumor}} - \beta_{\text{normal}}$. Student's t test was applied to compare the difference between the two groups. If the absolute value of the β difference level of a differential methylation point (DMP)'s ($|\beta_{\text{tumor}} - \beta_{\text{normal}}| > 0.4$ and the p value < 0.001 , it was considered meaningful. As a high methylation level is commonly associated with inappropriate transcriptional silencing, DMPs with a β difference > 0.4 and a p value < 0.05 were used for further research. The same method was applied to assess the methylation data for 17 other types of cancer.

Results

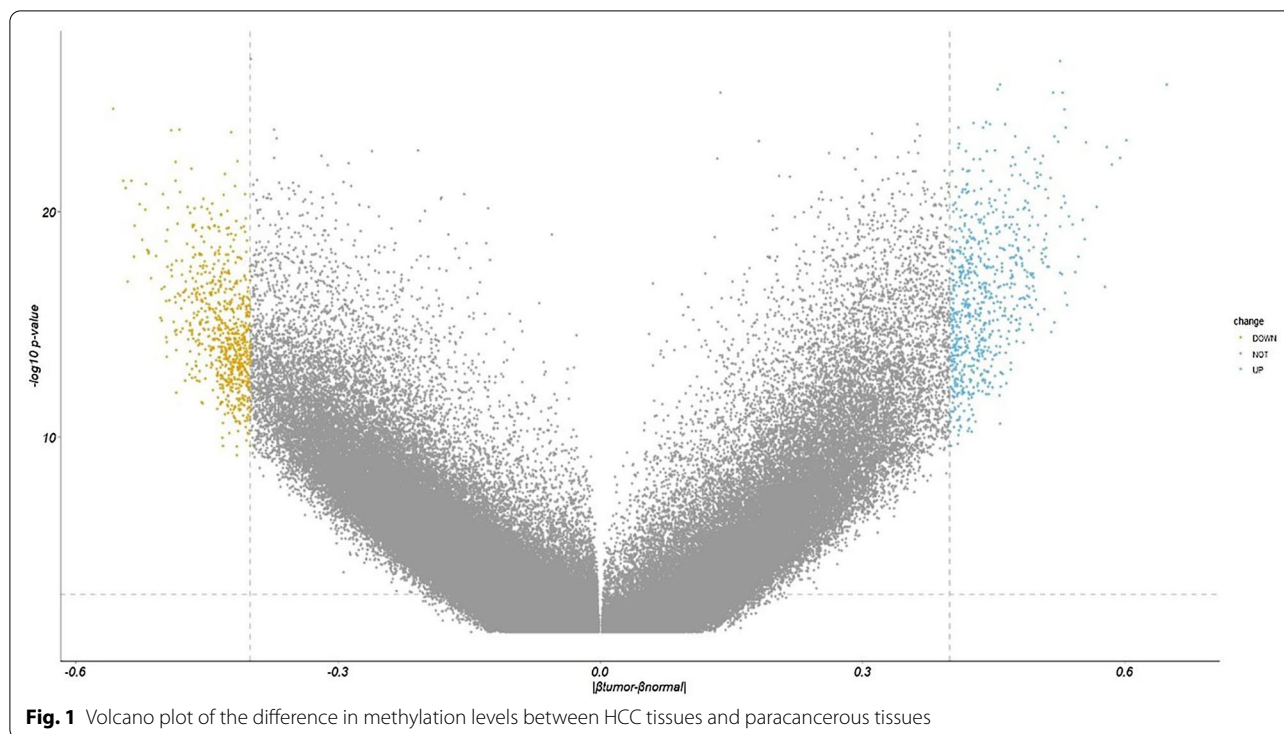
Results of the differential methylation analysis

In the discovery set of the LIHC dataset (59 pairs of cancer tissues and paracancerous tissues), based on the selection standard of $|\beta_{\text{tumor}} - \beta_{\text{normal}}| > 0.4$ and p value < 0.001 , a total of 1374 DMPs were selected. A total of 678 DMPs had increased methylation, while 696 DMPs

had decreased methylation (Fig. 1). It should be specified that the CpG distribution of different array platforms is different and that different individuals have differences in their methylation patterns [39]. Therefore, the CpG distribution displayed in this research is referenced from the Illumina HumanMethylation450 BeadChip dataset downloaded from GEO (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gpl13534>). The DMPs of the 17 other cancer types can be found in Additional file 1 (Figure S0). Among the DMPs with decreased methylation (Fig. 2), 3.03% were located within the CpG island region, and 21.9% were located within the promoter region. Moreover, 71.8% of the DMPs with increased methylation (Fig. 2) were located within the CpG island region, and 36.4% were located within the promoter region. It is worth noting that the Infinium HumanMethylation450 BeadChip array includes probes designed to target CpG islands, as well as shores, shelves and "open sea" regions. The results shown in Fig. 2 are strongly influenced by the CpGs localization and by the array design. This means that is predictable to find highly methylated CpG in the "Island" and less methylated CpGs in "open sea" [40]. In the promoter region, 69.3% of the CpG sites are located within the CpG island, meaning that high levels of gene methylation tend to occur in the promoter region, especially within the CpG islands. High methylation levels in gene promoter regions are usually the cause of transcriptional silencing, especially the silencing of cancer suppressor genes.

Selecting the diagnostic HCC-specific sites

From the LIHC methylation data of the discovery set, 678 DMPs with increased methylation were selected. The same standard was applied to the 17 other cancer datasets. BLCA yielded 1525 DMPs with increased methylation, BRCA yielded 1006 DMPs with increased methylation, CESC yielded 62 DMPs with increased methylation, CHOL yielded 1396 DMPs with increased methylation, COAD yielded 3074 DMPs with increased methylation, ESCA yielded 9465 DMPs with increased methylation, KIRC yielded 146 DMPs with increased methylation, KIRP yielded 577 DMPs with increased methylation, LUAD yielded 125 DMPs with increased methylation, LUSC yielded 2337 DMPs with increased methylation, PAAD yielded 157 DMPs with increased methylation, PCPG yielded 295 DMPs with increased methylation, PRAD yielded 439 DMPs with increased methylation, READ yielded 2278 DMPs with increased methylation, SKCM yielded 70 DMPs with increased methylation, THCA yielded 26 DMPs with increased methylation, and UCEC yielded 8527 DMPs with increased methylation. The UpSetR package was used to find the intersection of the 678 DMPs with increased



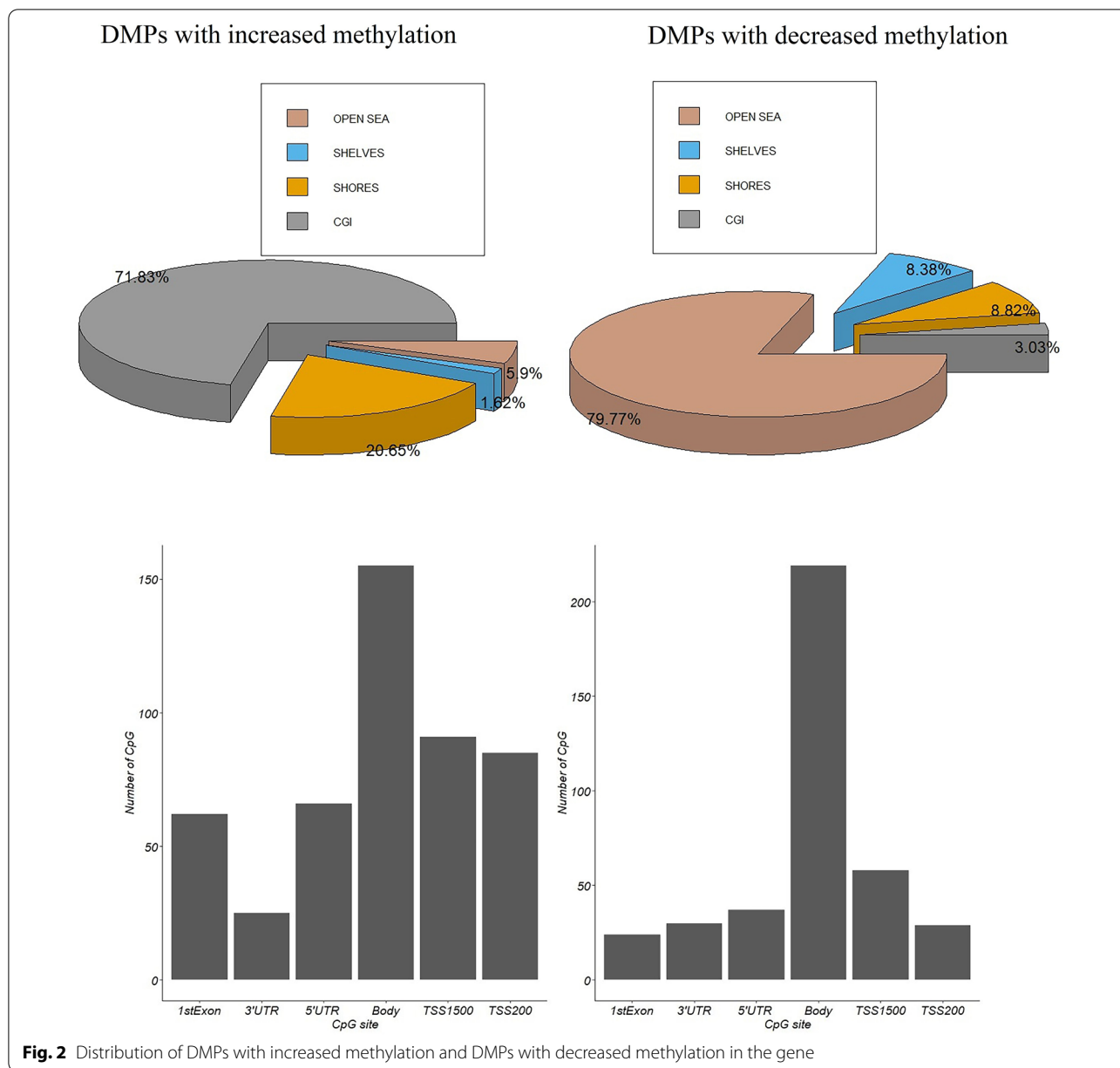
methylation in the LIHC dataset and those for the other 17 types of cancer, as demonstrated in Fig. 3. We removed intersecting DMPs and 168 CpG sites that were specific to HCC. Of the remaining DMPs, 176 with increased methylation were in the promoter region (ranging from transcription start site (TSS200 to TSS1500)). The Pearson coefficient was calculated to assess the correlation between the methylation of these 176 CpG sites and their corresponding gene expression values. If a single CpG site corresponded to multiple genes, the first one served as a reference. Sites that had a Pearson coefficient < -0.2 and a p value < 0.05 were considered significantly related. There were a total of 47 sites significantly related to gene expression. When these sites were overlapped with the 168 CpG sites specific to HCC, we identified 20 HCC-specific CpG sites that are in the promoter region and are significantly and negatively related to the expression of their corresponding genes.

To search for relevant CpG sites related to HCC, we downloaded clinical information for 377 HCC patients. The methylation data from the 20 CpG sites and the status information of the patients were assessed by logistic regression analysis model. According to the stepwise regression method, sites with p values lower than 0.05 were selected. Six CpG sites matched this criterion, indicating that they can be used as diagnostic HCC-specific CpG sites (Table 1, Additional file 1: Figures S1 and S2). The relationship between the 6 CpG

sites and the expression levels of their corresponding genes can be seen in the Additional file 1 (Additional file 1: Figure S3). We used a receiver operating characteristic (ROC) curve to evaluate the model, and the area under the ROC curve (AUC) was 0.97. The six diagnostic HCC-specific CpG sites had a relatively high accuracy (Fig. 4).

Accuracy evaluation of the model with the GEO dataset

To further evaluate the accuracy of the six CpG sites in the TCGA dataset, we downloaded two other datasets from the GEO database that are also based on the Illumina HumanMethylation450 BeadChip (GSE54503 and GSE157341). The GSE54503 dataset includes information on 66 pairs of HCC tissues and corresponding paracancerous tissues. The GSE157341 dataset includes information on 239 cancer tissues and 35 normal liver tissues. We applied logistic regression models in the GSE54503 dataset to predict the accuracy of the six-CpG-based signature; it had relatively high accuracy within the dataset with an AUC of 0.9711. In the GSE157341 dataset, the six-CpG-based signature yielded an AUC of 0.8802. The comparisons performed with the three datasets indicate that the six-CpG-based signature has high utility (Fig. 5). Therefore, the six-CpG-based signature shows great potential in differentiating normal and cancerous tissue.



Selecting the prognostic HCC-specific sites

We screened information in the LIHC dataset downloaded from the TCGA to find sites that are specific to HCC, are in the promoter region (ranging from TSS200 to TSS1500) and are negatively correlated with the expression of their corresponding genes. The Champ package was used to filter and normalize the data of 377 samples; 70 samples were filtered out. The remaining 307 samples were randomly divided into two sets: the training set (155 samples) and the validation set (152 samples). Finally, the entire dataset was used as another validation set.

In the training set, the methylation data were combined with survival status and survival time data. A univariate Cox regression model was constructed to assess the correlation of 47 CpG sites with the overall survival status. Finally, three CpG sites closely related to patient survival were chosen: cg08167706, cg03757145 and cg09626894 (refer to Table 2 for detailed information). The methylation level of the three sites and the coefficient from the Cox regression model were linearly combined. The formula for patients' prognosis risk score is as follows: $cg08167706 \times -0.935 + cg03757145 \times 1.300 + cg09626894 \times 1.082$. Generally, the higher the risk score is, the worse

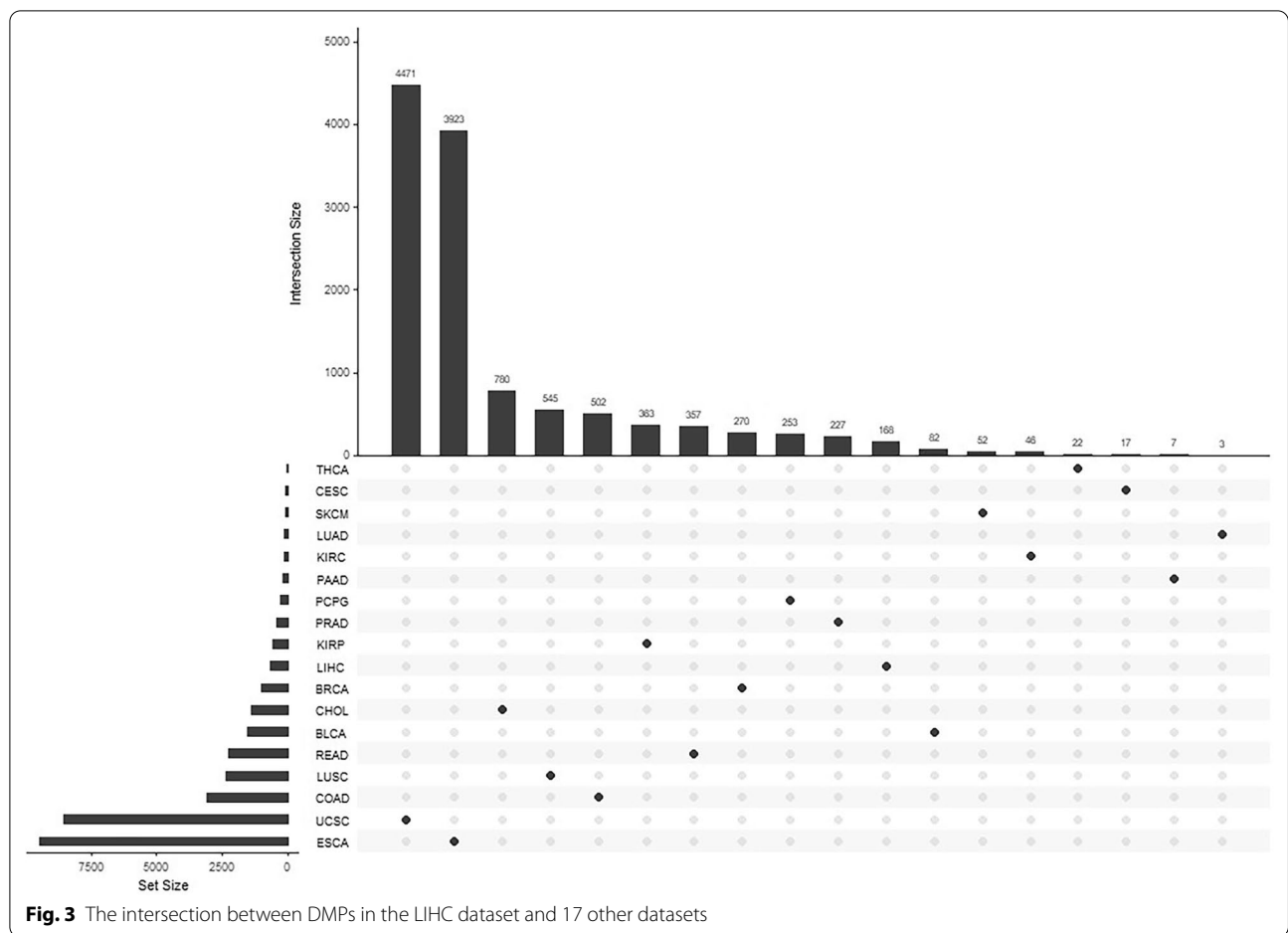


Fig. 3 The intersection between DMPs in the LIHC dataset and 17 other datasets

Table 1 Detailed information on HCC-specific CpG sites

ID	Gene Symbol	HR	Coefficient	95% CI for the Coefficient	p value
cg26581504	BCO2	3507.98	8.162797	5.40 to 11.67	2.31×10^{-7}
cg05106294	DKK3	689.79	6.536393	2.76 to 11.30	2.42×10^{-3}
cg20342184	GRHL2	60.14	4.096723	0.29 to 8.81	5.59×10^{-2}
cg23623667	KCNQ1	149.58	5.00	1.03 to 9.79	2.22×10^{-2}
cg14250130	PFKP	0.001	-6.51	(- 11.20) to (- 2.15)	3.87×10^{-3}
cg13564825	PPP1R14A	141.75	4.95	1.64 to 8.99	7.09×10^{-3}

the prognosis outcome is. The relationship between the three CpG sites and the expression of their corresponding genes can be seen in Additional file 1: Figure S4.

CpG sites and patient prognosis

The median of scores for the three CpG sites for each of the patients (0.582) was set as the threshold to separate the patients into high-risk (n=77) and low-risk (n=78) groups. The prognosis scores and survival status data are shown in Fig. 6. The death rate in the high-risk group was

higher than that in the low-risk group. Kaplan–Meier survival analysis and the sum of log-rank tests were used to examine and compare the difference in survival between the groups. The Kaplan–Meier survival curve (Fig. 7a) showed that the overall survival rate of the high-risk group was significantly lower than that of the low-risk group ($p < 0.0001$). The three-year survival rate and five-year survival rate for the high-risk group were 38% and 16.1%, respectively, while the three-year survival rate and five-year survival rate for the low-risk group were 78.9% and 62.4%,

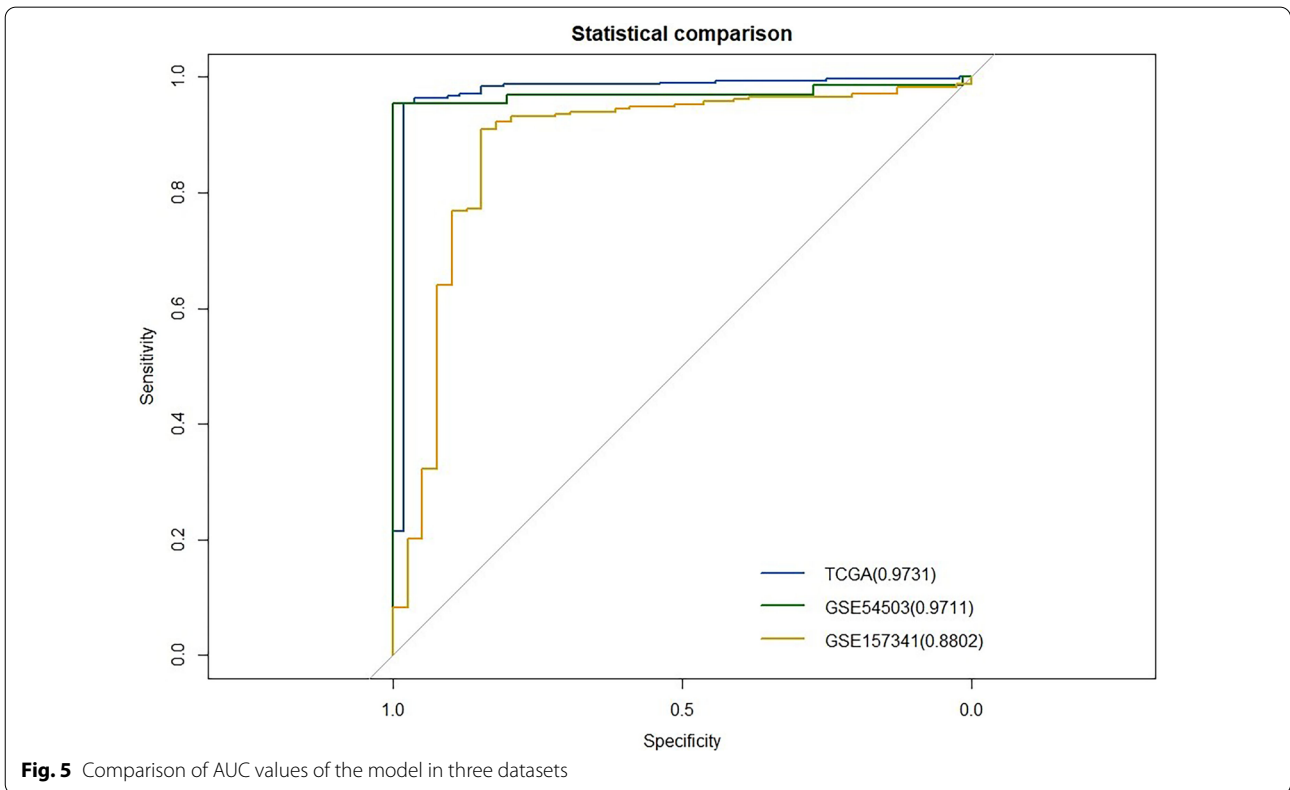
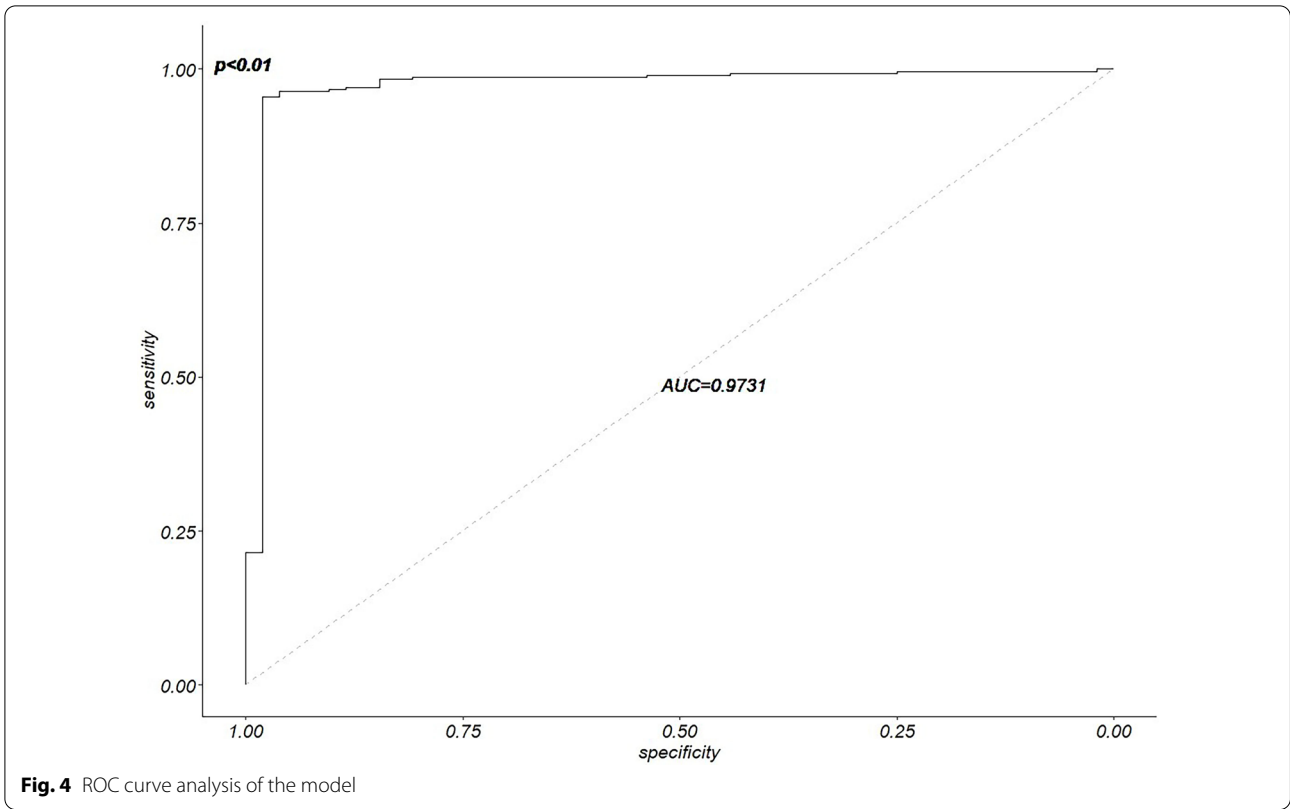


Table 2 CpG sites related to patient survival

ID	Gene symbol	HR	Coefficient	95% CI for the HR	p value
cg08167706	AKR1B1	0.392	-0.935	0.156–0.981	0.045
cg03757145	CDKL2	3.672	1.300	1.051–12.832	0.041
cg09626894	CFTR	2.952	1.082	1.051–8.291	0.039

respectively. We applied a ROC curve to evaluate the three-CpG-based signature, and the AUC was 0.678 (Fig. 7b). This means the model has good performance. When conducting a univariate Cox regression model with the three-CpG-based signature risk score and factors such as age, sex, alcoholism, tumor grade, recurrence, clinical stage and HBV infection status, the three-CpG-based signature risk score was significantly related to the survival of patients (hazard ratio [HR]=4.286, $p=0.00002$, see Additional file 1: Table S1).

Validation of the three-CpG-based signature based on the validation set and the entire dataset

To validate the utility of the prognosis model, the formula $\text{cg08167706} \times -0.935 + \text{cg03757145} \times 1.300 + \text{cg09626894} \times 1.082$ was applied to evaluate the prognosis of all patients. The threshold of 0.582 was used to evaluate the three-CpG-based signature. Similar to the previous approach, the patients were separated into two groups: a high-risk group ($n=80$) and a low-risk group ($n=72$). Kaplan–Meier survival analysis was used to compare the difference in survival of the two patient groups, and the results were similar to those of the training set analysis. The overall survival rate of the high-risk group patients was significantly lower than that of the low-risk group patients (Fig. 8a). The AUC was 0.621 (Fig. 8b). Univariate Cox regression analysis of clinical information and the three CpG sites revealed that the three-CpG-based signature was significantly related to the survival of patients (HR=2.293, $p=0.026$, refer to Additional file 1: Table S1). We extended the validation to the entire dataset ($n=307$) and separated the patients into a high-risk patient group ($n=157$) and a low-risk patient group ($n=150$), and the overall survival rate of the high-risk group patients was significantly lower than that of the low-risk group patients (Fig. 9a). The AUC was 0.65 (Fig. 9b). The univariate Cox regression model indicated that the three-CpG-based signature was significantly related to patient survival (HR=2.532, $p=0.0001$, Additional file 1: Table S1).

Independent predictive effect of the three-CpG-based signature

To examine whether the three-CpG-based signature can predict survival independent of clinical factors, such as age, sex, alcoholism, tumor grade, recurrence, clinical

stage and HBV infection, in the training set, validation set, and entire dataset, these variables were incorporated into a multivariate Cox regression model with the three-CpG-based signature. The results suggested that after adjusting for clinical factors, the three-CpG-based signature was an independent factor in the training set (HR=3.971, $p<0.01$), the validation set (HR=2.405, $p<0.01$) and the entire dataset (HR=2.205, $p<0.01$). In addition, it was discovered that the clinical stage was also significantly related to the survival of patients. In the layered analysis for the three-CpG-based signature and clinical staging, the patients were separated into a high-risk group and a low-risk group. Regardless of whether the patient belonged to the high-risk group or the low-risk group, the survival rate for patients who were in stage III and stage IV was significantly lower than that for those in stages I and II (Fig. 10, $p<0.01$). Finally, a nomogram was developed to evaluate the influence of each of the variables on prognosis (Additional file 1: Figure S5). As shown in the figure, the three-CpG-based signature contributed the most to the prognosis. This means that the three-CpG-based signature was the best indicator of patient survival. Finally, the flowchart of the search process for CpG sites used for diagnosis and prognosis is shown in the Additional file 1: Figure S6.

Discussion

Gene methylation is a symbol of tumor occurrence. The occurrence of every type of tumor is related to high gene methylation levels. To identify biomarkers that are specific to HCC and can aid in its diagnosis, methylation data for HCC and 17 other types of cancer based on the Illumina 450 K DNA array were downloaded from the TCGA. Based on the criteria of absolute β difference >0.4 and p value <0.001 , we filtered out DMPs with increased methylation that were not specific to HCC. We applied logistic regression analysis and identified 6 CpG sites that are unique to HCC and can be used to diagnose HCC: cg26581504 (BCO2), cg05106294 (DKK3), cg20342184 (GRHL2), cg23623667 (KCNQ1), cg14250130 (PFKP), and cg13564825 (PPP1R14A). These CpG sites can be used to effectively distinguish between cancer tissues and normal liver tissues. The AUC for the ROC curve of the model incorporating these sites is 0.973. We also used two independent HCC datasets from the GEO to

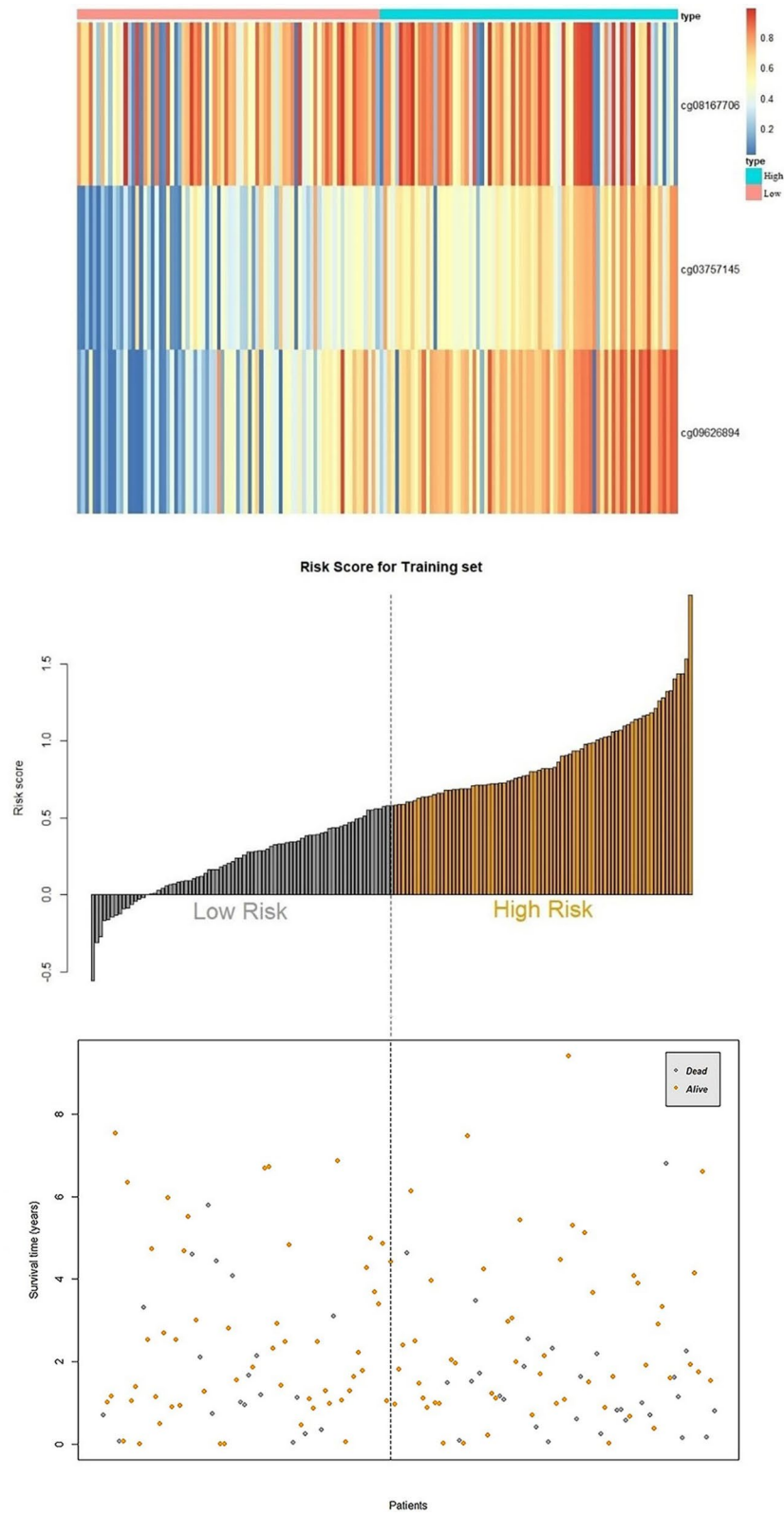
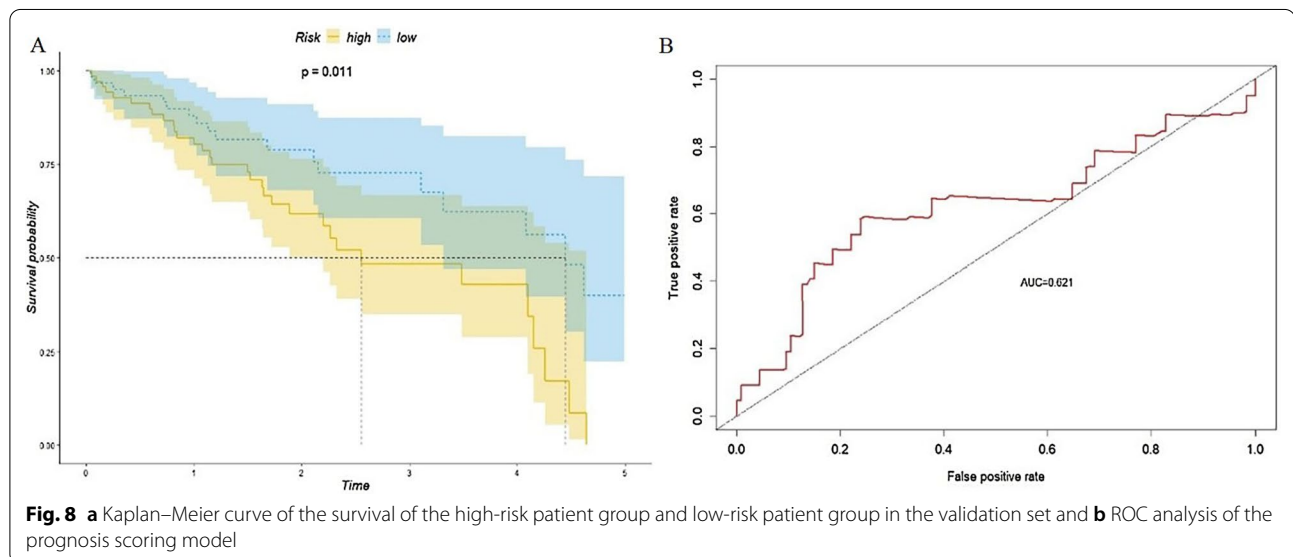
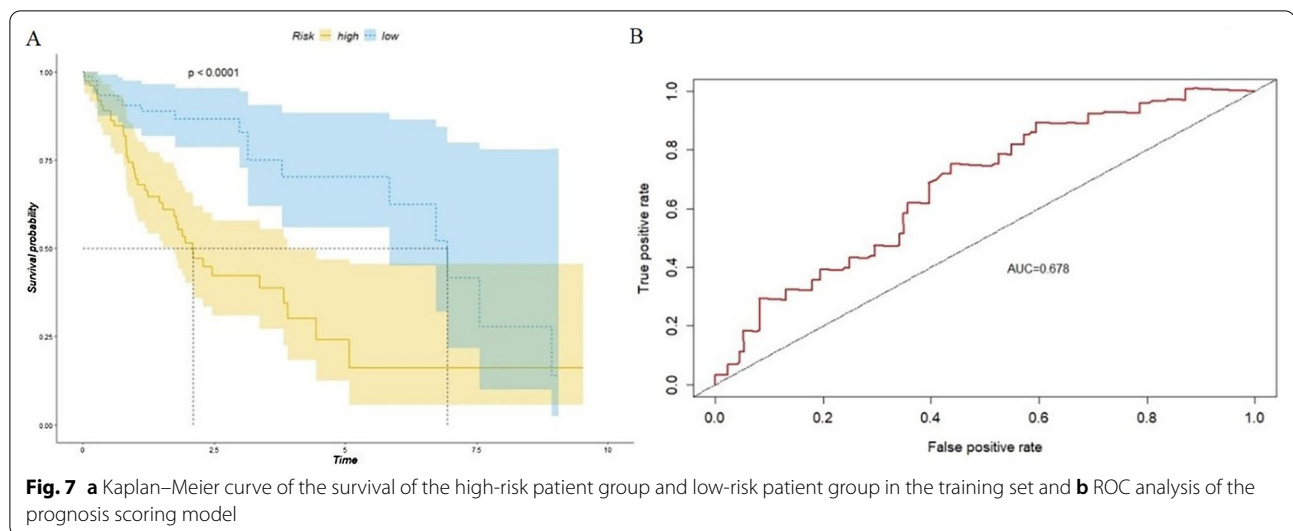


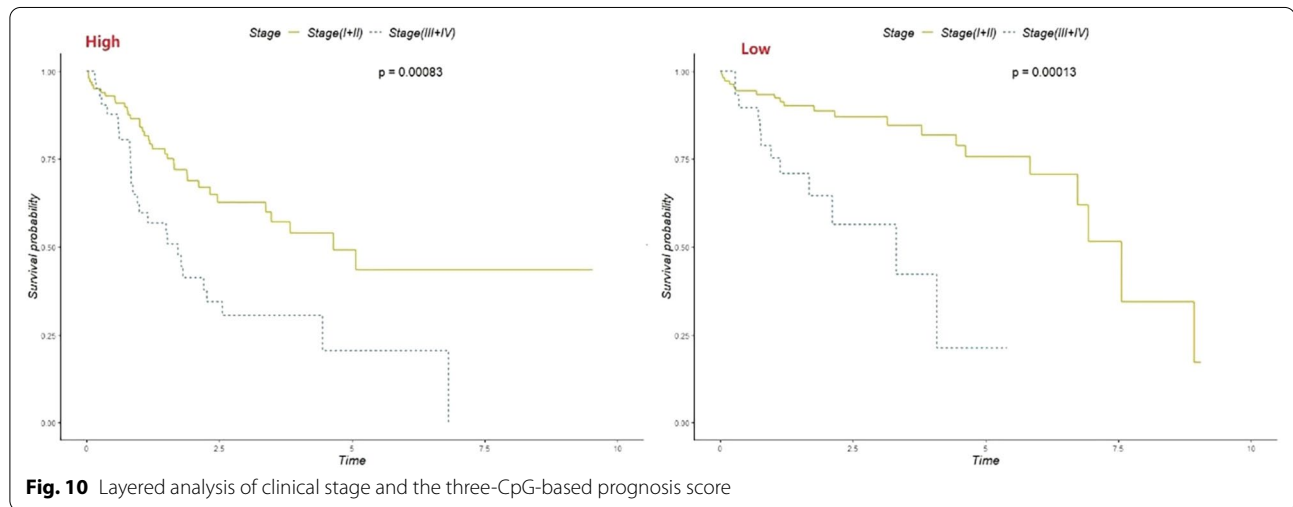
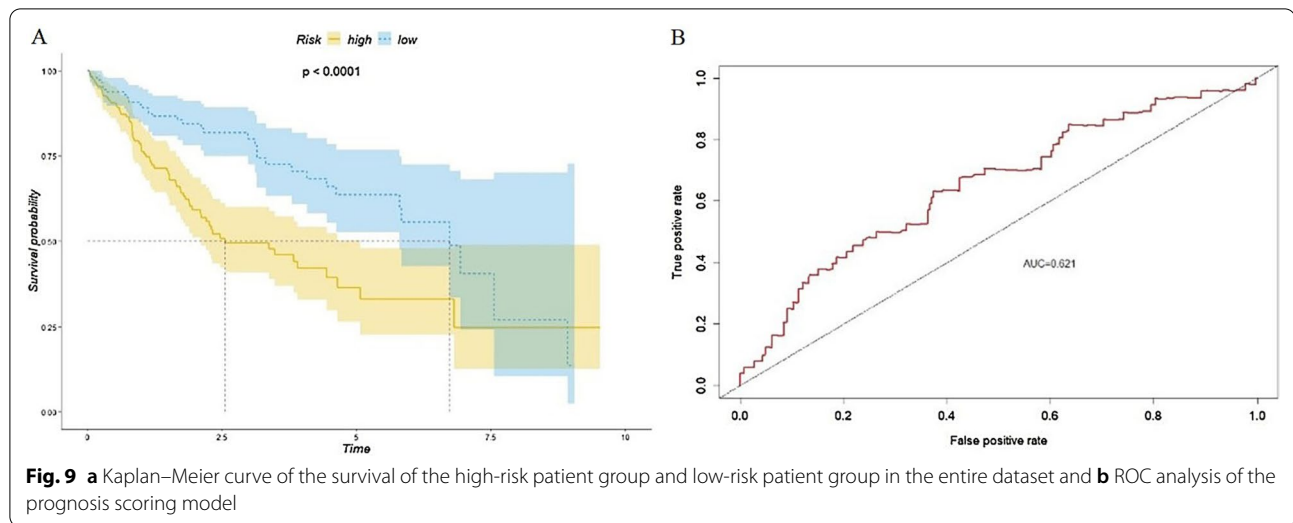
Fig. 6 Methylation level, prognosis score and survival status for the three CpG sites



validate the accuracy of these biomarkers. The AUCs for the ROC curves were 0.971 and 0.880. This means that these 6 biomarkers have high utility for the diagnosis of HCC. What separates this research from a previous study on diagnostic markers is that this study is the first to use the intersection method to filter highly altered DMPs between different TCGA datasets. The results are different between the two studies [33, 34].

In the diagnosis of HCC, this six-CpG-based signature has relatively high sensitivity and specificity. For patients who are already diagnosed with HCC, we also hope to develop a CpG-based signature that can evaluate the prognosis of patients for early intervention and disease management. Due to the lack of survival information and

clinical data that can be used for validation, this research separated samples in the LIHC dataset from TCGA into two sets: a training set and a validation set. Analysis of the training set revealed three CpG sites as candidate biomarkers related to prognosis: cg08167706 (AKR1B1), cg03757145 (CDKL2) and cg09626894 (CFTR). These three sites were used as the basis for prognosis scoring. According to the prognosis scores, the patients were divided into a high-risk group and a low-risk group, and the overall survival rate of the high-risk group was significantly lower than that of the low-risk group ($p < 0.001$). The model was then validated with the validation set and the whole dataset, and ROC analysis was performed. The results showed that the three-CpG-based prognosis



evaluation model can be used to effectively evaluate the prognosis of patients. Furthermore, clinical variables and the three-CpG-based signature were put into univariate and multivariate Cox regression models. Nomograms were also constructed for these clinical variables and the three-CpG-based signature. These results indicate that the signature is an important independent factor for predicting patient survival.

Conclusions

In summary, this study identified a six-CpG-based signature that may be used for diagnosing HCC and a three-CpG-based signature for predicting the survival of patients with HCC through analyzing the methylation profile and clinical data from TCGA and GEO.

In the diagnosis of HCC, taking tissues for methylation sequencing is certainly invasive. Plasma ctDNA assessment is both minimally invasive and can provide more dynamic monitoring of cancer. Clinically, peripheral plasma ctDNA is already applied for the early screening and detection of cancer. In HCC, the methylation levels of tumor DNA and matched plasma ctDNA are highly correlated [34]. Research indicates that CpG assessment of HCC tissue samples is no more effective than that of ctDNA samples [41]. In future research, we will gather our own samples and apply the six-CpG-based signature and the three-CpG-based signature for the assessment of HCC tissues and peripheral plasma ctDNA to find the best diagnosis and prognosis plan for HCC patients.

Abbreviations

LIHC: Liver hepatocellular carcinoma; TCGA: The Cancer Genome Atlas; GEO: Gene Expression Omnibus; DNMTs: DNA methyltransferases; MGMT: O⁶-methylguanine-DNA-methyltransferase; DMP: Differential methylation point; TSS: Transcription start site.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12920-021-01015-9>.

Additional file 1. Table S1. Univariate and multivariate Cox regression analysis for the whole dataset. **Figure S0.** Volcano plot of the difference in methylation levels for the cancer tissues and paracancerous tissues of 17 other cancer types. **Figure S1.** Methylation levels of the six HCC-specific CpG sites. **Figure S2.** Coefficients for the six HCC-specific CpG sites based on the univariate Cox regression model. **Figure S3.** Six HCC-specific CpG sites used for diagnosis and the expression levels of their corresponding genes. **Figure S4.** Three CpG sites related to patient survival and the expression levels of their corresponding genes. **Figure S5.** Nomogram of clinical factors and the three-CpG-based prognosis score. **Figure S6.** The flowchart of searching process of CpG sites used for diagnosis and prognosis

Acknowledgements

Yi-jin Yang at Shenzhen University and AJE provided English editing assistance, and we thank the patients and investigators who participated in the TCGA and GEO for providing data.

Authors' contributions

RZ and JL conceived of the study design and determined appropriate surveys and survey items for the study. RZ collected the data and performed the data analysis. RZ and JL interpreted analysis results and participated in drafts and revisions of this manuscript. All authors reviewed and provided approval for the final version of the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

TCGA: <https://tcga-data.nci.nih.gov>, now hosted at GDC: <https://portal.gdc.cancer.gov/>. GSE157341: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157341>. GSE54503: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54503>

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Author details

¹Health Science Center, Shenzhen University, Shenzhen, China. ²Department of Hepatobiliary and Pancreatic Surgery, Shenzhen Traditional Chinese Medicine Hospital, No.1 Fuhua Road, Shenzhen 518000, Guangdong, China.

Received: 27 February 2021 Accepted: 11 June 2021

Published online: 19 June 2021

References

- Qiu X, Huang Y, Zhou Y, Zheng F. Aberrant methylation of TRIM58 in hepatocellular carcinoma and its potential clinical implication. *Oncol Rep*. 2016;36(2):811–8.
- Bosch FX, Ribes J, Borràs J. Epidemiology of primary liver cancer. *Semin Liver Dis*. 1999;19(3):271–85.
- Dufour JF, Johnson P. Liver cancer: from molecular pathogenesis to new therapies: summary of the EASL single topic conference. *J Hepatol*. 2010;52(2):296–304.
- Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. *CA: Cancer J Clin*. 2016;66(2):115–32.
- El-Serag HB, Rudolph KL. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology*. 2007;132(7):2557–76.
- Ryerson AB, Ehemann CR, Altekruse SF, Ward JW, Jemal A, Sherman RL, et al. Annual report to the nation on the status of cancer, 1975–2012, featuring the increasing incidence of liver cancer. *Cancer*. 2016;122(9):1312–37.
- Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136(5):E359–86.
- Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA: Cancer J Clin*. 2015;65(2):87–108.
- Chen WQ, Zheng RS, Zhang SW, Li N, Zhao P, Li GL, et al. Report of incidence and mortality in china cancer registries, 2008. *Chin J Cancer Res*. 2012;24(3):171–80.
- El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma. *Gastroenterology*. 2012;142(6):1264–73.e1.
- Forner A, Llovet JM, Bruix J. Hepatocellular carcinoma. *Lancet (London, England)*. 2012;379(9822):1245–55.
- Cui Y, Jia J. Update on epidemiology of hepatitis B and C in China. *J Gastroenterol Hepatol*. 2013;28(Suppl 1):7–10.
- Duan Z, Jia JD, Hou J, Lou L, Tobias H, Xu XY, et al. Current challenges and the management of chronic hepatitis C in mainland China. *J Clin Gastroenterol*. 2014;48(8):679–86.
- Llovet JM, Brú C, Bruix J. Prognosis of hepatocellular carcinoma: the BCLC staging classification. *Semin Liver Dis*. 1999;19(3):329–38.
- El-Serag HB. Hepatocellular carcinoma. *N Engl J Med*. 2011;365(12):1118–27.
- Njei B, Rotman Y, Ditch I, Lim JK. Emerging trends in hepatocellular carcinoma incidence and mortality. *Hepatology (Baltimore, MD)*. 2015;61(1):191–9.
- Bruix J, Sherman M. Management of hepatocellular carcinoma: an update. *Hepatology (Baltimore, MD)*. 2011;53(3):1020–2.
- Sherman M. Serological surveillance for hepatocellular carcinoma: time to quit. *J Hepatol*. 2010;52(4):614–5.
- Bolondi L. Screening for hepatocellular carcinoma in cirrhosis. *J Hepatol*. 2003;39(6):1076–84.
- Ward J, Robinson PJ. How to detect hepatocellular carcinoma in cirrhosis. *Eur Radiol*. 2002;12(9):2258–72.
- Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev*. 2002;16(1):6–21.
- D'Urso A, Brickner JH. Mechanisms of epigenetic memory. *Trends Genetics*. 2014;30(6):230–6.
- Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med*. 2003;349(21):2042–54.
- Baylin SB, Herman JG. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet*. 2000;16(4):168–74.
- Jones PA, Laird PW. Cancer epigenetics comes of age. *Nat Genet*. 1999;21(2):163–7.
- Liu M, Jiang L, Guan XY. The genetic and epigenetic alterations in human hepatocellular carcinoma: a recent update. *Protein Cell*. 2014;5(9):673–91.
- Tischoff I, Tannapfe A. DNA methylation in hepatocellular carcinoma. *World J Gastroenterol*. 2008;14(11):1741–8.
- Esteller M, Corn PG, Baylin SB, Herman JG. A gene hypermethylation profile of human cancer. *Can Res*. 2001;61(8):3225–9.
- Esteller M, Hamilton SR, Burger PC, Baylin SB, Herman JG. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. *Can Res*. 1999;59(4):793–7.

30. Gerson SL, Trey JE, Miller K, Berger NA. Comparison of O6-alkylguanine-DNA alkyltransferase activity based on cellular DNA content in human, rat and mouse tissues. *Carcinogenesis*. 1986;7(5):745–9.
31. Warren JD, Xiong W, Bunker AM, Vaughn CP, Furtado LV, Roberts WL, et al. Septin 9 methylated DNA is a sensitive and specific blood test for colorectal cancer. *BMC Med*. 2011;9:133.
32. Shen S, Wang G, Shi Q, Zhang R, Zhao Y, Wei Y, et al. Seven-CpG-based prognostic signature coupled with gene expression predicts survival of oral squamous cell carcinoma. *Clin Epigenetics*. 2017;9:88.
33. Cheng J, Wei D, Ji Y, Chen L, Yang L, Li G, et al. Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. *Genome Med*. 2018;10(1):42.
34. Xu RH, Wei W, Krawczyk M, Wang W, Luo H, Flagg K, et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat Mater*. 2017;16(11):1155–61.
35. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, et al. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics (Oxford, England)*. 2017;33(24):3982–4.
36. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics (Oxford, England)*. 2014;30(10):1363–9.
37. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*. 2013;8(2):203–9.
38. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucl Acids Res*. 2015;43(7):e47.
39. Palumbo D, Affinito O, Monticelli A, Cocozza S. DNA Methylation variability among individuals is related to CpGs cluster density and evolutionary signatures. *BMC Genomics*. 2018;19(1):229.
40. Price ME, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin*. 2013;6(1):4.
41. Hlady RA, Zhao X, Pan X, Yang JD, Ahmed F, Antwi SO, et al. Genome-wide discovery and validation of diagnostic DNA methylation-based biomarkers for hepatocellular cancer detection in circulating cell free DNA. *Theranostics*. 2019;9(24):7239–50.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

