A CpGs Methylation Signature as Potential Marker for Early Diagnosis of Hepatocellular Carcinoma from HBV-Related Liver Disease using Multiplex Bisulfite Sequencing

Kang Li

Beijing Youan Hospital <https://orcid.org/0000-0003-0843-0382>

Yi Song

Capital Medical University Affiliated Beijing Friendship Hospital

Ling Qin

Capital Medical University Youan Hospital

Ang Li

Capital Medical University Youan Hospital

Sanjie Jiang

Cambridge University

Lei Ren

Air force medical center, PLA

Chaoran Zang

Capital Medical University Youan Hospital

Jianping Sun

Capital Medical University Youan Hospital

Yan Zhao

Capital Medical University Youan Hospital

Yonghong Zhang (✉ zhangyh@ccmu.edu.cn)

Capital Medical University

Research

Keywords: Multiplex bisulfite sequencing, CpGsmethylation, early HCC, diagnostic model, enhanced bootstrap validation

Posted Date: July 29th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-743735/v1>

Abstract

Background: Aberrant methylation of CpG sites served as epigenetic marker for building diagnostic, prognostic, and recurrence models for hepatocellular carcinoma (HCC).

Methods: Using Illumina 450K and EPIC Beadchip, we identified 34 CpG sites in peripheral blood mononuclear cell (PBMC) DNA that were differentially methylated in early HCC versus HBV-related liver diseases (HBVLD). We employed multiplex bisulfite sequencing (MBS) based on next-generation sequencing (NGS) to measure methylation of 34 CpG sites in PBMC DNA from 654 patients that were divided into a training set ($n = 442$), a test set ($n = 212$). Using training set, we selected and built a six-CpG-scorer (including cg14171514, cg07721852, cg05166871, cg18087306, cg05213896, and cg18772205), applying least absolute shrinkage and selection operator (LASSO) regression. We performed multivariable analyses of four candidate risk predictors (including six-CpG-scorer, age, sex, AFP level), using 20 times imputation of missing data, non-linearly transformed and backwards feature selection with logistic regression. The final model's regression coefficients were calculated according to "Rubin's Rules". The diagnostic accuracy of model was internally validated with 10000 bootstrap validation dataset, and then applied to the test set for validation.

Results: The area under the receiver operating characteristic curve (AUROC) of the model was 0.81 (95% CI, 0.77-0.85) and it showed good calibration, decision curve analysis. Using enhanced bootstrap validation, adjusted C-statistics and adjusted brier score was 0.809 and 0.199, respectively. The model also showed AUROC value of 0.84 (95% CI 0.79-0.88) of diagnosis for early HCC in test set.

Conclusions: Our model basing on six-CpG-scorer was a reliable diagnosis tool for early HCC from HBVLD. The usage of MBS method can realize large-scale detection of CpG sites in clinical diagnosis of early HCC and benefit the majority of patients.

Background

Hepatocellular carcinoma (HCC) is the most common primary liver tumor with high morbidity and mortality. It has been estimated that hepatitis B virus (HBV) infection is cause for 50–80% of HCC cases worldwide [1]. The progression of multi-stage hepatocarcinogenesis from chronic HBV infection (CHB) to HBV-related liver cirrhosis (LC), finally to HBV-related HCC are complex. There is 2%-7% incidence of LC per year in CHB patients [2], and 2–4% annual incidence of HCC in LC patients [3]. Meanwhile, CHB patients can also develop HCC without cirrhosis. The new data shows that annual HCC rates ranged 0.03–1.57% among non-LC Asian CHB male patients [4]. In involving 34952 patients study suggested that 2.29% CHB patients would develop HCC despite hepatitis B surface antigen (HBsAg) seroclearance [5]. There are several mechanisms related with HBV-related hepatocarcinogenesis progression including viral regulatory HBV x protein interrupting liver cell proliferation and increasing HBV replication [6]; integration of HBV DNA into the host cell genome provoking host cell chromosomal alterations and

insertional mutagenesis of cancer genes [7]; host genomic and epigenetic aberrant variation inducing by inflammation [8].

The high mortality rate of HCC is mostly due to its discovery at advanced stages. There is an urgent unmet need for biomarkers that can detect early HCC development in CHB and LC liver background. DNA methylation profiles for early stage of HCC in peripheral blood mononuclear cells (PBMC) are different from healthy controls as well as CHB. HCC has a specific DNA methylation signature in easily accessible PBMC and can serve as “noninvasive” biomarkers for detection of early HCC as well as HCC progression [9]. Several studies have demonstrated CpG methylation prepared for constructing diagnostic, prognostic, and recurrence models for HCC [10–12].

However, in these models, measurement methylation of CpGs was performed by pyrosequencing or methylation specific PCR. These methods were laborious and fussy work [13] for detection methylation of dozens of CpGs and inevitable increase of inter batch differences in big size samples. That could cause adverse effect on diagnostic efficacy of the model.

Therefore, in this study we applied multiplex bisulfite sequencing (MBS), which based on next-generation sequencing (NGS) method to assess methylation status of 34 CpGs in 654 samples at the same time. Multivariable methods identified a minimal set of CpGs to achieve optimal prediction. Meanwhile we handled with multiple imputation to complete missing data and bootstrapped training datasets for backward feature selection. The model incorporated both the six-CpG-scorer methylation panel and demographic and clinical characteristics risk factors for early diagnosis onset of HCC from HBV-related liver disease (HBVLD) including CHB and LC.

Material And Methods

Data processing for CpGs selection

The raw IDAT files of Illumina 450K Beadchip data including CHB patients ($n = 10$) and Barcelona clinic liver cancer staging system (BCLC) 0 ($n = 10$), BCLC-A ($n = 10$), BCLC-B ($n = 10$) and BCLC-C ($n = 9$) was obtained from Gene Expression Omnibus (accession number: GSE67170) [9], which was prepared for getting differentially methylated CpGs between CHB and each HCC stage. Forty-four paired PBMC sample were collected from 22 patients from LC stage to HCC BCLC-0 and -A. The genome-wide DNA methylation were analyzed using Illumina EPIC Beadchip (data did not published). The R package ChAMP was applied for methylation raw data processing and the paired t tests was used to analyze the differentially methylated CpGs between LC and HCC BCLC-0 and -A. The differentially methylated CpGs with $|\Delta \text{beta}| \geq 0.2$ and remaining significant after Bonferroni-corrected P value < 0.05 were selected.

Patients study population.

A total of 654 patients from Beijing You' An Hospital were included in this study and 442 patients were randomly assigned to the training set while the remaining participants ($n = 212$) were included in the test

set. The demographic and clinical characteristics data of patients who were admitted from August 8, 2010, to July 27, 2019 at the Beijing You'An Hospital were collected from hospital electronic medical records. The diagnosis of CHB was based on the Asian-Pacific clinical practice guidelines on the management of hepatitis B: a 2015 update [14], diagnosis of hepatitis B related LC was based on the guideline of prevention and treatment for chronic hepatitis B (2010 Version) [15], diagnosis of HCC according to 2012 EASL clinical practice guidelines: Management of chronic hepatitis B virus infection, Barcelona clinic liver cancer staging system (BCLC) [16]. The study was approved by the Institutional Ethics Committee of the Beijing You'An Hospital. All participants signed written informed consents on enrolment.

Pbmc Dna Bisulfite Converted

After extraction, PBMC DNA bisulfite was converted (ZYMO Research, Irvine, USA). We measured the concentration of the single strand DNA using a Qubit 2.0 (Thermo, USA, catalog Q10212) to ensure that adequate amounts of DNA had been converted.

Multiplex Bisulfite Sequencing

A panel contained 34 'CpG-free' primer pairs was designed for targeting to all candidate CpGs. Library preparation was performed by nest-PCR. First round PCR reaction was set up as follows: bisulfite converted DNA 5µl; forward primer mix (10µM) 1µl; reverse primer mix (10µM) 1µl; 2×PCR Ready Mix 15µl (total 25µl) (KAPA HiFi HotStart ReadyMix PCR). The plate was sealed and PCR performed in a thermal instrument (BIO-RAD, T100TM) using the following program: 1 cycle of denaturing at 98°C for 3min, then 27 cycles of denaturing at 98°C for 20 s, annealing at 60°C for 4 min, final elongation at 72°C for 2 min. Finally hold at 10°C. The PCR products were checked using electrophoresis in 1 % (w/v) agarose gels in TBE buffer (Tris, boric acid, EDTA) stained with ethidium bromide (EB) and visualized under UV light. Then we used AMPure XP beads to purify the amplicon product. After that, the second round PCR was performed. PCR reaction was set up as follows: first round PCR amplicon product (10ng/µl) 2µl; universal P7 primer with barcode (10µM) 1µl; universal P5 primer with barcode (10µM) 1µl; 2×PCR Ready Mix 15µl (total 30µl) (Kapa HiFi Ready Mix). The plate was sealed and PCR performed in a thermal instrument (BIO-RAD, T100TM) using the following program: 1 cycle of denaturing at 98°C for 1 min, then 8 cycles of denaturing at 98°C for 20 s, annealing at 60°C for 20 s, elongation at 72°C for 30s, and a final extension at 72°C for 2 min. Then we used AMPure XP beads to purify the second round PCR amplicon product. The libraries were then quantified and pooled. Paired-end sequencing of the library was performed on the HiSeq XTen sequencers (Illumina, San Diego, CA).

Data Qc And Snp Calling

MBS raw data was processed by Sangon Biotech. Briefly, raw reads were filtered according to two steps: (1) removing adaptor sequence if reads contains by cutadapt (v 1.2.1); (2) removing low quality bases from reads 3' to 5' ($Q < 20$) by PRINSEQ-lite (v 0.20.3); (3) Bismark (version v0.22.1) (www.bioinformatics.babraham.ac.uk/projects/) was used for CpGs detect with default parameters.

CpGs Selection And Assessment

A multi-CpG-based scorer was constructed for diagnosis the early HCC patients from HBVLD in the training dataset. The least absolute shrinkage and selection operator (LASSO) with cross-validation method was applied to select the most significant CpGs from 34 ones described in Supplementary Table 1. The process was mainly performed “glmnet” package [17] based the R software (Version 4.0.3) and depicted in Fig. 1A.

Statistical Analysis

Firstly, the restricted cubic splines (RCS) was applied to detect the possible nonlinear dependency of the relationship between early HCC and 17 continuous variables, using 3 knots at the 25th, 50th, and 75th, and percentiles of corresponding variable. There were potential threshold associations between early HCC and five indices (age, direct bilirubin (DBil), total protein (TP), albumin (ALB) and hemoglobin) (p value for nonlinear < 0.001). We have categorized the age to < 43 years and ≥ 43 years groups, DBil to $< 7 \mu\text{mol/L}$ and $\geq 7 \mu\text{mol/L}$ groups, TP to $< 65 \text{ g/L}$ and $\geq 65 \text{ g/L}$ groups, ALB to $< 42 \text{ g/L}$ and $\geq 42 \text{ g/L}$ groups, hemoglobin to $< 140 \text{ g/L}$ and $\geq 140 \text{ g/L}$ groups based on the RCS curve. The other 4 indices (total bilirubin (TBil), γ -glutamyltranspeptidase (γ -GT), platelet count (PLT) and lymphocyte count (LYM)) were categorized to normal and abnormal groups based on the medical reference value, respectively. The other seven indices (alanine aminotransferase (ALT), aspartate aminotransferase (AST) alkaline phosphatase (ALP), white blood cell count (WBC), monocyte count (MONO), neutrophil count (NUET) and alpha-fetoprotein (AFP)) were log-transformed, respectively (Supplementary table 1).

Missing values in original dataset were imputed 20 times using imputation with predictive mean matching (“mice” package) [18]. For each of 20 complete dataset, 500 bootstrapping datasets were generated, and backwards feature selection with the Akaike Information Criterion (AIC) was performed on the 10000 datasets (“rms”) [19]. Second-order interaction terms were also tested for inclusion. The selected predictors constructed the final model, which regression coefficients were calculated using “psfmi” package according to “Rubin's Rules” [20, 21]. The enhanced bootstrap method was to evaluate the stability of the diagnosis model. Discrimination ability was assessed by the area under the receiver operating characteristic curve (AUROC) and C-statistics. Calibration was applied for assessment of agreement between predicted and observed risks across of the population. The nomogram presented the results of the modeling process. Decision curve analysis (DCA) was performed the model clinical usefulness. The construction process referred to this report [22] and the model analysis conforms to the reporting standards of STARD [23] and was depicted in Fig. 1B.

Result

Demographic and Characteristics of the patients

After a strict pathological diagnosis and exclusion process, 168 patients with CHB, 173 patients with liver cirrhosis, 148 patients with HCC BCLC-0 stage, and 165 patients with HCC BCLC-A stage were prospectively collected in Beijing You'An hospital and included into this study (Fig. 1A). A total of 654 patients were randomly assigned to the training set ($n = 442$) and test set ($n = 212$). Demographic and clinical characteristics of patients were shown in Table 1.

Table 1
Clinical characteristics of the enrolled participants in this study (n = 654)

	Training set (442)		Test set (212)	
	HBVLD (n = 230)	early HCC (n = 212)	HBVLD (n = 111)	early HCC (n = 101)
Age(year,mean ± SD)	45.37 ± 11.9	46.22 ± 11.7	45.38 ± 12.08	52.24 ± 10.36
Sex (M/F)	170/60	174/38	80/31	90/11
ALT (U/L)	108 (47.0%)	128 (60.4%)	51(45.9%)	60(59.4%)
< 40	121(52.6%)	83(39.2%)	60(54.1%)	41(40.6%)
≥ 40				
AST (U/L)	125(54.3%)	140(66.0%)	59(53.2%)	55(54.5%)
< 40	104(45.2%)	71(30.9%)	52(46.8%)	46(44.5%)
≥ 40				
Total bilirubin (μmol/L)	144(62.6%)	147(69.3%)	79(71.1%)	63(62.4%)
< 21	85(37.0%)	64 (30.2%)	32(18.9%)	38(37.6%)
≥ 21				
Direct bilirubin (μmol/L)	169(73.5%)	175(82.5%)	88(79.3%)	72(71.3%)
< 7	60(26.1%)	36(17.0%)	23(20.7%)	29(28.7%)
≥ 7				
Total protein (g/L)	168(73.4%)	135(59.0%)	85(76.6%)	65(64.4%)
< 65	61(26.5%)	76(35.8%)	26(23.4%)	36(35.6%)
≥ 65				
Albumin (g/L)	136(59.1%)	105(49.5%)	72(64.9%)	45(44.6%)
< 40	93(40.4%)	106(50.0%)	39(35.1%)	56(55.4%)
≥ 40				
γ-GT (U/L)	121(52.6%)	95(44.8%)	52(46.8%)	40(39.6%)
< 45	96(41.7%)	95(44.8%)	52(46.8%)	61(60.4%)
≥ 45				

Abbreviations: HBVLD, HBV-related liver disease; CHB, chronic hepatitis B; LC, HBV-related liver cirrhosis; HCC, Hepatocellular carcinoma; BCLC, Barcelona clinic liver cancer staging system; γ-GT, γ-glutamyltranspeptidase; WBC, white blood cell. ALT, Alanine aminotransferase; AST, Aspartate aminotransferase.

	Training set (442)		Test set (212)	
	HBVLD (n = 230)	early HCC (n = 212)	HBVLD (n = 111)	early HCC (n = 101)
Alkaline phosphatase (U/L)	172(74.8%)	142(67.0%)	83(74.8%)	48(47.5%)
≤ 100	46(20.0%)	48(22.6%)	22(19.8%)	37(36.6%)
> 100				
WBC count × 10 ⁹ /L	174(75.7%)	157(74.1%)	79(71.2%)	74(73.3%)
< 3.5	46(20.0%)	42(19.8%)	24(21.6%)	14(13.9%)
3.5–9.5	10(4.3%)	12 (5.7%)	8(7.2%)	13(12.9%)
> 9.5				
Hemoglobin (g/L)	143(62.2%)	138(65.1%)	77(69.4%)	60(59.4%)
< 130	87(37.8%)	73(34.4%)	34(30.6%)	41(40.6%)
≥ 130				
Platelet count × 10 ⁹ /L	129(56.1%)	108(50.9%)	70(63.1%)	58(57.4%)
< 125	101(43.9%)	104(49.1%)	41(36.9%)	43(42.6%)
≥ 125				
Lymphocyte count×10 ⁹ /L	172(74.8%)	149(70.3%)	85(76.6%)	58(57.4%)
< 1.1	58(25.2%)	62(29.2%)	26(23.4%)	43(42.6%)
≥ 1.1				
Monocyte count× 10 ⁹ /L	214(93.1%)	182(85.8%)	107(96.4%)	81(80.2%)
< 0.6	16(6.9%)	29(13.7%)	4(3.6%)	20(19.8%)
≥ 0.6				
Neutrophil count× 10 ⁹ /L	166(72.2%)	167(78.8%)	80(72.1%)	71(70.3%)
< 1.8	53(23.0%)	17(8.0%)	26(23.4%)	12(11.9%)
1.8–6.3	11(4.8%)	17(8.0%)	5(4.5%)	18(17.8%)
> 6.3				
Abbreviations: HBVLD, HBV-related liver disease; CHB, chronic hepatitis B; LC, HBV-related liver cirrhosis; HCC, Hepatocellular carcinoma; BCLC, Barcelona clinic liver cancer staging system; γ-GT, γ-glutamyltranspeptidase; WBC, white blood cell. ALT, Alanine aminotransferase; AST, Aspartate aminotransferase.				

	Training set (442)		Test set (212)	
	HBVLD (n = 230)	early HCC (n = 212)	HBVLD (n = 111)	early HCC (n = 101)
Alpha-fetoprotein (ng/mL)	178(77.4%)	110(51.9%)	86(77.5%)	41(40.6%)
< 20	50(21.7%)	96(45.3%)	25(22.5%)	58(57.4%)
≥ 20				
Abbreviations: HBVLD, HBV-related liver disease; CHB, chronic hepatitis B; LC, HBV-related liver cirrhosis; HCC, Hepatocellular carcinoma; BCLC, Barcelona clinic liver cancer staging system; γ-GT, γ-glutamyltranspeptidase; WBC, white blood cell. ALT, Alanine aminotransferase; AST, Aspartate aminotransferase.				

CpGs selection and panel signature building for early HCC diagnosis

On the basis of our previous studies [9], we portrayed differentially methylated CpGs between CHB and each of the HCC phases utilizing Limma package with Bonferroni-correction. The number of specifically methylated CpGs between CHB and each of the HCC phases included 2285 for BCLC-0; 2233 for BCLC-A; 3345 for BCLC-B; and 23596 for BCLC-C. There were 326 differentially methylated CpGs could specifically distinguish early HCC (BCLC-0 and BCLC-A stages) from CHB (Fig. 2A). Moreover, 20 robust CpG differentially methylated sites were used in this study from 33 CpGs ($|\Delta\beta| \geq 0.2$). Meanwhile, 36 differentially methylated CpGs ($|\Delta\beta| \geq 0.2$) could specifically distinguish early HCC from LC using paired *t* tests with Bonferroni-correction, and 17 CpGs were selected for this study (unpublished work). In all, methylation ration of 34 CpGs (three CpGs overlap) were investigated using MBS in the HBVLD and early HCC smaples (Supplementary Table 2 and Supplementary Table 3).

Thirty-four CpGs were reduced to six potential predictors using the LASSO regression model. The cross-validated error plot and a coefficient profile plot of the LASSO regression model were produced (Fig. 2B and 2C). The logistic regression was used for building the six-CpG-scorer: where risk score = $-0.87 - 3.73 \times \text{cg14171514} + 2.58 \times \text{cg07721852} + 6.91 \times \text{cg05166871} - 9.85 \times \text{cg18087306} + 4.50 \times \text{cg05213896} + 4.39 \times \text{cg18772205}$. We then applied this formula to calculate the risk score for early HCC of each patient based on their individual six CpGs methylation ration. The risk score was significantly increased in the early HCC samples versus the HBVLD samples ($p < 2.22 \times 10^{-16}$) (Fig. 2D). The six CpGs and their combination six-CpG-scorer also showed diagnostic accuracy (Fig. 2E, Supplementary Table 4). According to determining of maximum Youden index, 0 severed as the optimal cutoff point of risk score. Therefore we classified those patients with risk score < 0 as low-risk group, and those with risk score ≥ 0 as high-risk group. The distribution of demographic and clinical characteristics did not vary significantly between the high-risk and low-risk group (Supplementary Table 5).

Six-cpg-scorer Signature Was Independent Of Clinical Factors

The univariate logistic analysis was performed in ten thousand bootstrap datasets. Six of the 18 candidate variables indicated a higher early HCC risk; among these were higher age (≥ 43 years), male, individuals with higher AFP, higher six-CpG-scorer, lower TP ($< 65\text{g/L}$), and lower TBil (Table 2).

Four variables were selected by the backward feature selection procedure in ten thousand bootstrap datasets (Table 2). Age, sex, AFP and six-CpG-scorer signature were independent risk factors for early HCC. The six-CpG-scorer also showed significantly higher predictive accuracy than demographic and clinical risk factors.

Table 2. All 18 variables included in the backwards feature selection analysis.				
	Univariable		Multivariable	
Variable	OR (95% CI)	P Value	OR (95% CI)	P Value
Age (years), ≥ 43 vs <43	5.07 (3.07-8.56)	1.75e-07 ***	5.23 (3.24-8.63)	2.56e-08 ***
Sex (Female vs Male)	0.32(0.17-0.56)	0.0011 **	0.32(0.19-0.53)	0.000248 ***
Alpha-fetoprotein (AFP) (ng/ml), $\log_{10}$ a	1.50 (1.35-1.69)	1.85e-09 ***	1.48(1.34-1.64)	3.05e-10 ***
Six-CpG-Scorer	2.37(1.83-3.12)	1.10e-07 ***	2.58 (2.01-3.36)	1.08e-09 ***
Total protein(g/L), ≥ 65 vs <65	0.38 (0.19-0.71)	0.01209 *	0.32(0.20-0.51)	0.56
Total bilirubin(μ mol/L), ≥ 21 vs <21	0.32 (0.14-0.69)	0.01588 *	0.45(0.32-0.64)	0.78
AST (U/L), $\log_{10}$	0.77 (0.45-1.28)	0.40551		
ALT (U/L) , $\log_{10}$	0.51(0.43-1.06)	0.15320		
Direct bilirubin (μ mol/L), ≥ 7 vs <7	0.55 (0.77-2.35)	0.38832		
γ -GT (U/L), ≥ 45 vs <45	1.15 (0.83-1.61)	0.47649		
Albumin (g/L), ≥ 40 vs <40	0.74 (0.41-1.36)	0.42147		
Alkaline phosphatase (U/L), $\log_{10}$	0.77 (0.44-1.35)	0.44623		
Hemoglobin (g/L), ≥ 115 vs <115	1.68 (0.91-3.14)	0.16827		
White blood cell count $\times 10^9$ /L, $\log_{10}$	0.84 (0.17-4.24)	0.85430		
Monocyte count $\times 10^9$ /L, truncate_99 + $\log_{10}$ b	2.29 (1.14-4.67)	0.05410 .		
Platelet count $\times 10^9$ /L, ≥ 125 vs <125	0.84(0.50-1.42)	0.58796		

Lymphocyte count $\times 10^9/L$, ≥ 1.1 vs <1.1	0.82 (0.46-1.49)	0.58026
Neutrophil count $\times 10^9/L, \log_{10}$	1.08(0.37-2.86)	0.89871
a: nonlinear transformation; b: Monocyte count variable was truncated with 1th and 99 th percentiles and then performed with nonlinear transformation; Abbreviations: OR, odds ratio; γ -GT, γ -glutamyltranspeptidase; ALT, Alanine aminotransferase; AST, Aspartate aminotransferase. Signif. codes: '***' 0.001, '**' 0.01, '*' 0.05.		

Development Of An Individualized Early Hcc Diagnosis Nomogram

The early HCC diagnosis model incorporated the four risk predictors and estimated on the 20 complete datasets according to Rubin's Rule. The prognostic index X (based on logistic regression model coefficients) was: $X = -1.0944708 - 0.7183741 \times \text{Sex (Male = 1, Female = 2)} + 1.7286974 \times \text{Age } [(< 43 \text{ years}) = 1, (\geq 43 \text{ years}) = 2] + 0.2761166 \times \log(\text{AFP}) + 0.7902764 \times \text{six-CpG-scorer}$. And the calculation of the predicted risk of early HCC from HBVLD: $\text{risk of eHCC} = \frac{1}{1 + \exp(-X)}$. It presented as the early HCC nomogram (eHCC nomogram) (Fig. 3A).

Estimation for the C-Statistics and Brier score by bootstrap Validation

Internal validation was performed using the 500 times resampling enhanced bootstrap method from each 20 complete datasets. The result showed negligible model optimism. The apparent C-statistics and apparent Brier score was 0.805 and 0.200, respectively. The optimism of the C-statistics and Brier score was -0.0042 and 0.00164, respectively. The adjusted C-statistics and Brier score was 0.809 and 0.199, respectively.

Diagnostic Performance And Clinical Usefulness Of Ehcc Nomogram

The AUROC of the model was 0.81 (95% CI, 0.77–0.85) in training set. The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) when used in differentiating the early HCC

from HBVLD were 70.0%, 77.8%, 74.4%, and 73.6%, respectively.

The calibration curve of the eHCC nomogram for the probability of early HCC demonstrated good agreement between prediction and observation in training set (Fig. 3B). The decision curve analysis of the eHCC nomogram and that for the model without six-CpG-scorer was presented in (Fig. 3C). The DCA showed that if the threshold probability of a patient or doctor was > 10%, using eHCC nomogram to predict early HCC adds more benefit than either the treat-all-patients scheme or the treat-none scheme (Fig. 3D).

Diagnostic Performance Of The Ehcc Nomogram In Test Set

We further enrolled 212 patients including 111 HBVLD and 101 early HCC to serve as test set for validation of the diagnostic potential. The risk score was significantly increased in the early HCC versus the HBVLD group in test set ($p = 2.7 \times 10^{-7}$) (Fig. 4A). The eHCC nomogram achieved an AUROC value of 0.84 (95% CI 0.79–0.88) between the early HCC and HBVLD (Fig. 4B). The calibration curve demonstrated good agreement between prediction and observation in early HCC (Fig. 4C). The sensitivity, specificity, PPV, NPV were used in differentiating the early HCC from HBVLD were 68.9%, 82.9%, 80.0%, and 71.8%, respectively. The nomogram also indicated good clinical benefits both in DCA, which suggested an obvious diagnosis efficacy for early HCC from HBVLD (Fig. 4D).

Discussion

The identification of differentially methylated CpGs in genome-wide from Methylation Chip (450K and EPIC) often required a test and/or replication in extra cohorts. The pyrosequencing, bisulfite-conversion-based methylation PCR, PCR cloning and Sanger sequencing method were laborious and fussy work [13] for detection methylation of each CpG site and inevitable increase of inter batch differences in big size samples. The deep sequencing of specific target CpGs were often called for MBS, which based on NGS method to assessed methylation status at target DNA regions and with very high coverage [13]. The important and recommendable characteristics of MBS were low cost, low DNA input, high repeatability and scalability. In addition, the MBS were compatible with common NGS platforms using standard equipment, making it available to most laboratories [24]. Construction of amplicon libraries according to target CpGs panel was key step of MBS. In multiplex PCR using bisulfite converted DNA as template, mixed-base primers, which pyrimidine base Y contains a 'mixture' of cytosines and thymines at the cytosine site, and purine base R contains a mixture of guanines and adenines at the guanine site, greatly affected the amplification efficiency and led to a decrease in library quality. In the design of this study, we applied 34 'CpG-free' primer pairs (i.e. primers that do not bind to a region containing CpG dinucleotides) targeting each region cover candidate CpGs, which had been considered to produce 'non-preferential' amplification of bisulfite-converted DNA genome [25]. We successfully establish a version of 'CpG-free' primer pairs panel applied in MBS to simultaneously investigate the methylation status of 34 candidate CpGs across a large sample size ($n = 654$). The complete and accurate methylation data of

34 candidate CpGs allowed us to screen out and integrated into a six-CpG-scorer with application of LASSO regression.

CpGs methylation served as an apparatus focused on non-invasive biomarkers for chronic diseases (age-related diabetes [26], diabetic embryopathy [27], coronary heart disease [28], nonalcoholic fatty liver disease [29], or cancer (bladder cancer [30], rectal cancer [31], HCC[32]) had been set up by compelling studies. Our recent study looked at CpGs methylation alterations of HBV related liver disease and developed a model discriminating compensated cirrhosis patients from CHB and decompensated cirrhosis ones based on two CpGs biomarkers [33]. Our previous study also supported the feasibility of using 350 CpGs for detection of each stages of HCC from healthy control [9]. In this study, we successfully established and validated a novel diagnostic nomogram based on six-CpG-scorer to distinguish early HCC patients from CHB and LC ones. Furthermore, this proposed six-CpG-scorer can diagnose the early HCC better than other demographic and characteristics risk factors. Meanwhile, eHCC nomogram validated considerable diagnostic potential to early HCC and indicated good clinical benefits in DCA.

The CpGs severing as independent risk factors of HCC from mining analysis of Methylation Chip data might point out a novel trace for the exploration of HCC progress mechanism. Basing on Illumina Human Beadchip 27K array, sphingomyelin phosphodiesterase 3 (SMPD3) and neurofilament, heavy polypeptide (NEFH) were found to behave as tumor suppressor genes in HCC after validation in vitro and in vivo [34]. Gentilini *et al*/ applied Methylation 450K BeadChip array and showed that four epigenetically regulated candidate genes (adherens junctions associated protein 1 (AJAP1), adenosine deaminase RNA specific B2 (ADARB2), protein tyrosine phosphatase receptor type N2 (PTPRN2) and sidekick cell adhesion molecule 1 (SDK1) potentially involved in the pathogenesis of HCC [35]. The finding of AJAP1 was consistent with clinical observation of low AJAP1 expression as an independent factor for predicting disease-free survival (DFS) [36]. Mechanically, AJAP1 could block epithelial to mesenchymal transition (EMT) via suppressing β -catenin/zinc finger E-box binding homeobox 1 (ZEB1) signaling.

In this study, cg14171514 located within 5'UTR of neuroblast differentiation-associated protein (AHNAK) gene, and significantly hypomethylated both in the PBMC DNA of early HCC patients comparing with HBVLD ones in this study. Consistently, the aberrant hypomethylation status in AHNAK gene promoter region from HCC patients PBMC DNA was verified in methylation-specific PCR (MSP) results of our previous studies [37]. Aberrant AHNAK methylation level in PBMC DNA was the associated with HCC. In addition, AHNAK mRNA had been reported overexpression in liver cancer tissues by qPCR [37] and TGCA data (Supplementary Fig. 1). Our novel results from immunoprecipitation in combination with mass spectrometry (IP-MS) analysis strongly indicated that AHNAK protein involved in and promoted HCC progress (data did not published). The cg18087306 was located within body of Lamin B2 (LMNB2) gene, whose expression level in HCC tissue was higher relative to peritumor tissue. LMNB2 was a promising HCC prognostic and diagnostic biomarker [38].

Conclusions

We had showed characteristic changes of 34 CpGs in HBVLD and early HCC across a clinical cohort, identified six-CpG-scorer and validated their diagnostic efficacy. Thus, we proposed that six-CpG-scorer basing nomogram become potential non-invasive diagnostic tool for early HCC from HBVLD, and performed internal validation using the 500 times resampling enhanced bootstrap method from each 20 complete datasets to evaluate the stability and then applied to the test set for validation. The external validation of the monogram will be needed in much larger cohorts from different centers or regions to promote the efficacy and stability for further benefit HCC populations.

Abbreviations

HCC: hepatocellular carcinoma

CHB: chronic HBV infection

LC: HBV-related liver cirrhosis

HBVLD: HBV-related liver diseases

MBS: multiplex bisulfite sequencing

NGS: next-generation sequencing

PBMC: Peripheral blood mononuclear cell

BCLC: Barcelona clinic liver cancer staging system

AUROC: area under the receiver operating characteristic curve

RCS: restricted cubic splines

DCA: Decision curve Analysis

AIC: Akaike Information Criterion

LASSO: least absolute shrinkage and selection operator

Declarations

Availability of data and materials

The raw data of illumineEPIC and multiplex bisulfite sequencing can be made available upon request. Patient and clinicopathological data in the study cohorts cannot be made publicly available due to their content of identifiable human data. Requests to access the datasets should be directed to corresponding author.

Ethics approval and consent to participate

The study was approved by the Institutional Ethics Committee of the Beijing You'An Hospital, Approval No. of Ethics Committee: Jing you kelunzi[2017]26# . All participants signed written informed consents on enrolment. This study followed the recommendations described by the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

All the authors declare that there are no conflicts of interest to disclose.

Funding

This project was supported by grants National Key R&D Program of China (2020YFE0202400), Beijing Natural Science Foundation (7202069, 7191004), Capital's Funds of Health Improvement and Research (CFH2020-1-2182, CFH2020-2-1153), Beijing Municipal Science & Technology Commission (Z171100001017078) and Beijing Key Laboratory (BZ0373).

Author contributions

YHZ and KL conceived and designed the experiments. KL and YS, LQ and AL performed all of the experiments. YS designed multiplex bisulfite sequencing and acquired raw data. KL, LR and SJJ conducted the statistical analysis. KL, LQ, CRZ, JPS, and YZ contributed collection and processing of samples. All authors read and approved the final manuscript submitted for publication.

Acknowledgements

Not applicable.

References

1. Venook AP, Papandreou C, Furuse J, Ladrón de Guevara L. The Incidence and Epidemiology of Hepatocellular Carcinoma: A Global and Regional Perspective. *The Oncologist*. 2010;15(S4):5-13.
2. Hsu YS, Chien RN, Yeh CT, Sheen IS, Chiou HY, Chu CM, et al. Long-term outcome after spontaneous HBeAg seroconversion in patients with chronic hepatitis B. *Hepatology*. 2002;35(6):1522-7.
3. El-Serag HB. Hepatocellular carcinoma. *N Engl J Med*. 2011;365(12):1118-27.
4. Liu M, Tseng T-C, Jun DW, Yeh M-L, Trinh H, Wong GLH, et al. Transition rates to cirrhosis and liver cancer by age, gender, disease and treatment status in Asian chronic hepatitis B patients. *Hepatol Int*. 2021; 15:71–81.

5. Liu F, Wang X-W, Chen L, Hu P, Ren H, Hu H-D. Systematic review with meta-analysis: development of hepatocellular carcinoma in chronic hepatitis B patients with hepatitis B surface antigen seroclearance. *Aliment Pharmacol Ther.* 2016;43(12):1253-61.
6. Guerrieri F, Belloni L, D'Andrea D, Pediconi N, Le Pera L, Testoni B, et al. Genome-wide identification of direct HBx genomic targets. *BMC Genomics.* 2017;18(1):017-3561.
7. Buendia MA, Neuveut C. Hepatocellular carcinoma. *Cold Spring Harb Perspect Med.* 2015;5(2):a021444. doi: 10.1101/cshperspect.a021444.
8. Gehring AJ, Ho ZZ, Tan AT, Aung MO, Lee KH, Tan KC, et al. Profile of tumor antigen-specific CD8 T cells in patients with hepatitis B virus-related hepatocellular carcinoma. *Gastroenterology.* 2009;137(2):682-90.
9. Zhang Y, Petropoulos S, Liu J, Cheishvili D, Zhou R, Dymov S, et al. The signature of liver cancer in immune cells DNA methylation. *Clin Epigenetics.* 2018;10(1):8. doi: 10.1186/s13148-017-0436-1. eCollection 2018.
10. Long J, Chen P, Lin J, Bai Y, Yang X, Bian J, et al. DNA methylation-driven genes for constructing diagnostic, prognostic, and recurrence models for hepatocellular carcinoma. *Theranostics.* 2019;9(24):7251-67.
11. Sceusi EL, Loose DS, Wray CJ. Clinical implications of DNA methylation in hepatocellular carcinoma. *HPB.* 2011;13(6):369-76.
12. Jiang HY, Ning G, Wang YS, Lv WB. 14-CpG-Based Signature Improves the Prognosis Prediction of Hepatocellular Carcinoma Patients. *Biomed Res Int.* 2020;8:9762067.
13. Rainbow DB, Yang X, Burren O, Pekalski ML, Smyth DJ, Klarqvist MD, et al. Epigenetic analysis of regulatory T cells using multiplex bisulfite sequencing. *Eur J Immunol.* 2015;45(11):3200-3.
14. Sarin SK, Kumar M, Lau GK, Abbas Z, Chan HL, Chen CJ, et al. Asian-Pacific clinical practice guidelines on the management of hepatitis B: a 2015 update. *Hepatol Int.* 2016;10(1):1-98.
15. Chinese Society of Hepatology and Chinese Society of Infectious Diseases CMA. [The guideline of prevention and treatment for chronic hepatitis B (2010 version)]. *Zhonghua Gan Zang Bing Za Zhi.* 2011;19(1):13-24.
16. Liver EAFTSOT. EASL clinical practice guidelines: Management of chronic hepatitis B virus infection. *J Hepatol.* 2012;57(1):167-85.
17. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw.* 2010;33(1):1-22.
18. Buuren Sv, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in R. *J Stat Softw.* 2011;45(3), 1–67.
19. Harrell FE. Regression Modeling Strategies: with applications, to linear models, logistic and ordinal regression, and survival analysis, 2nd ed. Heidelberg: Springer. Biometrics 2016; 72: 1006.
20. Heymans MW, van Buuren S, Knol DL, van Mechelen W, de Vet HC. Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Med Res Methodol.* 2007;7(33):1471-

21. Eekhout I, van de Wiel MA, Heymans MW. Methods for significance testing of categorical covariates in logistic regression models after multiple imputation: power and applicability analysis. *BMC Med Res Methodol.* 2017;17(1):129.
22. Markaki M, Tsamardinos I, Langhammer A, Lagani V, Hveem K, Røe OD. A Validated Clinical Risk Prediction Model for Lung Cancer in Smokers of All Ages and Exposure Types: A HUNT Study. *EBioMedicine.* 2018;31:36-46.
23. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ.* 2015;28(351) :h5527. doi: 10.1136/bmj.h5527.
24. Korbie D, Lin E, Wall D, Nair SS, Stirzaker C, Clark SJ, et al. Multiplex bisulfite PCR resequencing of clinical FFPE DNA. *Clin Epigenetics.* 2015;7(1):15-67.
25. Clark SJ, Harrison J, Paul CL, Frommer M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.* 1994;22(15):2990-7.
26. Grant CD, Jafari N, Hou L, Li Y, Stewart JD, Zhang G, et al. A longitudinal study of DNA methylation as a potential mediator of age-related diabetes risk. *Geroscience.* 2017;39(5-6):475-89.
27. Schulze KV, Bhatt A, Azamian MS, Sundgren NC, Zapata GE, Hernandez P, et al. Aberrant DNA methylation as a diagnostic biomarker of diabetic embryopathy. *Genet Med.* 2019;21(11):2453-61.
28. Horvath S, Gurven M, Levine ME, Trumble BC, Kaplan H, Allayee H, et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biol.* 2016;17(1):016-1030.
29. Hyun J, Jung Y. DNA Methylation in Nonalcoholic Fatty Liver Disease. *Int J Mol Sci.* 2020;21:8138. doi: 10.3390/ijms21218138.
30. Kitchen MO, Bryan RT, Emes RD, Luscombe CJ, Cheng KK, Zeegers MP, et al. HumanMethylation450K Array-Identified Biomarkers Predict Tumour Recurrence/Progression at Initial Diagnosis of High-risk Non-muscle Invasive Bladder Cancer. *Biomark Cancer.* 2018;8(10) :1179299–17751920.
31. Exner R, Pulverer W, Diem M, Spaller L, Woltering L, Schreiber M, et al. Potential of DNA methylation in rectal cancer as diagnostic and prognostic biomarkers. *Br J Cancer.* 2015;113(7):1035-45.
32. Villanueva A, Portela A, Sayols S, Battiston C, Hoshida Y, Méndez-González J, et al. DNA methylation-based prognosis and epidrivers in hepatocellular carcinoma. *Hepatology.* 2015;61(6):1945-56.
33. Li K, Qin L, Jiang S, Li A, Zhang C, Liu G, et al. The signature of HBV-related liver disease in peripheral blood mononuclear cell DNA methylation. *Clin Epigenetics.* 2020;12(1):81.
34. Revill K, Wang T, Lachenmayer A, Kojima K, Harrington A, Li J, et al. Genome-wide methylation analysis and epigenetic unmasking identify tumor suppressor genes in hepatocellular carcinoma. *Gastroenterology.* 2013;145(6):1424-35.
35. Gentilini D, Scala S, Gaudenzi G, Garagnani P, Capri M, Cescon M, et al. Epigenome-wide association study in hepatocellular carcinoma: Identification of stochastic epigenetic mutations through an innovative statistical approach. *Oncotarget.* 2017;8(26):41890-902.

36. Ezaka K, Kanda M, Sugimoto H, Shimizu D, Oya H, Nomoto S, et al. Reduced Expression of Adherens Junctions Associated Protein 1 Predicts Recurrence of Hepatocellular Carcinoma After Curative Hepatectomy. *Ann SurgOncol*. 2015;22(3):015-4695.
37. Sun L, Li K, Liu G, Xu Y, Zhang A, Lin D, et al. Distinctive pattern of AHNAK methylation level in peripheral blood mononuclear cells and the association with HBV-related liver diseases. *Cancer Med*. 2018;7(10):5178-86.
38. Kong W, Wu Z, Yang M, Zuo X, Yin G, Chen W. LMNB2 is a prognostic biomarker and correlated with immune infiltrates in hepatocellular carcinoma. *IUBMB Life*. 2020;72(12):2672-85.

Figures

Figure 1

Study design and analysis protocol. A: Study design and flow diagram. A total of 654 PBMC samples were prospectively collected. Thirty-four CpGs were investigated using multiplex bisulfite sequencing in 341 HBVLD (CHB: 168, LC: 173) and 313 early HCC group (BCLC-0:148, BCLC-A: 165). The least absolute shrinkage and selection operator method (LASSO) cross validation was introduced to selected 6 CpGs and built six-CpG-scorer in the training dataset (n=442, 212 early HCC and 230 HBVLD). B: Schematic

representation of statistical analysis. The analyses were performed on 18 selected predictors measured on 442 parents involved in the training set. Twenty complete datasets were created after 20 times missing imputation, and 17 candidate variables were detected the possible nonlinear dependency of the relationship with early HCC and seven variables of them were transformed in a non-linear fashion. Resampling 500 times were from each of the 20 complete datasets, leading to a total number of 10000 bootstrap datasets. A backward feature selection with AIC was repeated on each of 10000 bootstrap datasets for selecting the most relevant risk variables for early HCC. The factors chosen at least once during the feature selection procedure constituted the final mode, whose coefficients were estimated using Rubin's Rule from the 20 complete datasets. In internal validation, the model predictiveness and correcting overfit were assessed using 10000 separate enhanced bootstrapped datasets. The nomogram presented the final mode. Decision curve analysis and clinical impact curves were performed the final model clinical usefulness. The final mode also showed an obvious diagnosis potential in test set (n=212, 101 early HCC and 111 HBVLD). HBVLD, HBV-related liver disease; HCC, hepatocellular carcinoma; BCLC, Barcelona Clinic Liver Cancer staging system.

Figure 2

Construction of the six-CpG-scorer in training set A: Venn diagram of the overlap between differentially methylated CpGs between different HCC BCLC stages (0-C) and CHB. B: Five-fold CV LASSO coefficients of 34 candidate CpGs. The first black vertical dotted line corresponds to the $\lambda_{\min}$ that minimized binomial deviance during CV. The second black vertical dotted line corresponds to λ_{1se} , used for the selection of six biomarker CpGs. C: LASSO coefficient profile of the 34 early HCC-associated CpGs. A

vertical line was drawn at the value chosen by five-fold cross-validation. D: The six-CpG-scorer of each participant from the liver disease, early HCC group in training dataset was calculated and the score was significantly increased in the early HCC versus the HBV-related liver disease (HBVLD) samples (Wilcoxon rank sum test was used to compare six-CpG risk score between two groups). E: Area under the receiver operating characteristic curve (AUROC) of six-CpG-scorer and its component CpGs. Data were AUROC (95% CI). HCC, hepatocellular carcinoma; BCLC, Barcelona Clinic Liver Cancer staging system; LASSO, least absolute shrinkage and selection operator method.

Figure 3

Performance evaluation and clinical usefulness of eHCC nomogram. A: Nomogram was based on age, sex, alpha-fetoprotein (AFP) and six-CpG-scorer. B: Calibration curves of the eHCC nomogram in training dataset. The x-axis represented the predicted probability of early HCC risk from nomogram. The y-axis represented the actual early HCC rate. The blue diagonal dotted line represented a perfect performance by an ideal nomogram. The red dot line represented the performance of the eHCC nomogram; the green solid line represented bootstrap-corrected performance of eHCC nomogram. C: Area under the receiver operating characteristic curve (AUROC) of eHCC nomogram in training dataset. Data were AUROC (95%

CI). D: Decision curve analysis (DCA) of the nomogram predicted early HCC in training set. The x-axis indicated the threshold probability and Y-axis indicated the net benefit. The blue curve represented the eHCC nomogram. The red line indicated the assumption that all patients were early HCC. The horizontal green line indicated the assumption that there were early HCC. eHCC, early hepatocellular carcinoma.

Figure 4

Diagnostic performance of the eHCCnomogram in test set A: The risk score was significantly increased in the early HCC samples versus the HBVLD group ($p=2.7\times10^{-7}$). B:Area under the receiver operating characteristic curve (AUROC) of eHCCnomogram in test set. Data were AUROC (95% CI). C: Calibration curves of the eHCCnomogram in test set. The x-axis represented the predicted probability of HCC risk from nomogram. The y-axis represented the actual HCC rate. The blue diagonal dotted line represented a perfect performance by an ideal nomogram. The red dot line represented the performance of the eHCCnomogram, the green solid line represented bootstrap-corrected performance of eHCCnomogram. D: Decision curve analysis (DCA) of the nomogram predicted early HCC in test set. The x-axis indicated the threshold probability, and Y-axis indicated the net benefit. The blue curve represented the

eHCCnomogram. The red line indicated the assumption that all patients were HCC. The horizontal green line indicated the assumption that there were early HCC.eHCC, early hepatocellular carcinom.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigure1.pptx](#)
- [SupplementaryTable1.doc](#)
- [SupplementaryTable2.xlsx](#)
- [SupplementaryTable3.docx](#)
- [SupplementaryTable4.doc](#)
- [SupplementaryTable5.doc](#)