**Long-Region Hypo-methylation in HBV Integration Regions Enhance HCC Non-invasive Surveillance by Low-pass Whole Genome-wide Bisulfite Sequencing.**

Haikun Zhang1,8, #, Peiling Dong2, #, Shicheng Guo3, #, Augusto Villanueva4, Chengcheng Tao1, Wenmin Zhao2, Jiakang Wang5, Ramsey Cheung6, Huiguo Ding2, Steven J. Schrodi3,7,\*, Dake Zhang1,\*, Changqing Zeng1,\*

1Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, 100101, China

2Department of Hepatology, Beijing You’an Hospital Affiliated with Capital Medical University, Beijing 100069, China

3Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI, USA

4Liver Cancer Research Program, Division of Liver Diseases, Tisch Cancer Institute, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

5Biology Department, Stonybrook University, Stonybrook, NY, USA

6

7Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, Madison, WI, USA

8University of Chinese Academy of Sciences, Beijing 100049, China

# These authors contributed equally to this work

\* Corresponding Author

## Abstract

Circulating cell-free DNA methylation has been demonstrated to be a promising strategy for non-invasive cancer diagnosis. However, the low-level of cell-free DNA in plasma and the high cost of whole genome bisulfite sequencing (WGBS) limit sequencing depth and subsequent biomarker identification of cell-free DNA in plasma. Here we demonstrate long-region hypo-methylation (LRM) in low-pass WGBS data (<5-million reads) provides high sensitivity and specificity surveillance to hepatocellular carcinoma (HCC). We applied a low-pass WGBS approach and demonstrated DNA methylation abnormalities in HCC occurred in HBV integration regions. These findings reflect the stage of liver disease progression thereby serving as a suitable surrogate for methylation level estimation in plasma cfDNA analysis of liver diseases.

## Introduction

Circulating cell-free DNA (cfDNA) are small double-stranded DNA fragments(1) found in plasma, urine, and other body fluids(2) originating from cell apoptosis and necrosis(3). In many settings, analyses of cfDNA can be regarded as a way to perform a “liquid biopsy”, which have previously demonstrated the potential not only for cancer diagnosis and prognosis (4, 5), but also for identifying patients with premalignant states, inflammatory conditions or traumas (1). Apoptotic and necrotic tumor cells can release cfDNA into the peripheral blood, which can reflect the tumor-related genetic features, including mutations, copy number aberrations and epigenetic changes(4). As such, cfDNA represents an important biomarker of oncogenesis. However, genome-wide methylation assays require large amounts of input DNA—conventional WGBS requires microgram input and reduced respresentation bisulfite sequencing (RRBS) requires 30ng of DNA input which is often approaching the maximum level of the cfDNA detected (or detectable) in a sample of human blood. In order to effectively investigate cfDNA, several novel methods have been proposed, such as scRRBS(6) and cfMeDIPseq(7).

Liver cancer is the fourth cause of cancer-related mortality worldwide. In the United States, liver cancer death rate increased 43% from 7.2 to 10.3 per 100,000 between 2000-2016(8). Hepatocellular carcinoma (HCC), the most frequent form of primary liver cancer, generally develops in patients with chronic liver disease due to hepatitis B (HBV), hepatitis C (HCV), alcohol abuse or non-alcoholic fatty liver disease(9, 10). Chronic inflammation, fibrosis, and aberrant hepatocyte regeneration favor a series of genetic and epigenetic events that culminate in hepatocyte malignant transformation. Hepatocarcinogenesis is a complex and poorly-understood multistep process that includes the histological transition from regenerative nodules in the context of cirrhosis, through dysplastic nodules and ultimately HCC (11-13). The high risk of HCC development in patients with cirrhosis (i.e., 2-4% annual risk) justifies the recommendation of biannual HCC surveillance with abdominal ultrasound (US) and serum alpha-fetoprotein (AFP) in patients at high-risk(14). Non-randomized studies suggest that early HCC detection increases the odds to receive a curative treatment and increase survival. However, the sensitivity of US and AFP is 63% to detect early stage HCC, which underscores the need for improved early detection tools. A number of studies have focused on cfDNA as a potential source of novel early detection biomarkers in HCC. This includes mutation profiling (15, 16), circulating tumor cells (CTCs)(17) and DNA methylation(18-23). As opposed to mutations and CTCs, DNA methylation analysis of cfDNA has the theoretical advantage of providing tissue of origin information, which is critical when cfDNA originates from a composite of cell types. Multiple studies have focused on the use of cfDNA methylation in cancer diagnosis in the areas of specific biomarkers (19, 23), pervasive hypo-methylation (18) and tissue of origin (20-22). Single cytosine measurement and high accuracy [facilitate](file:///E:\Program%20Files\youdao\Dict\7.5.2.0\resultui\dict\?keyword=facilitate) whole genome bisulfite sequencing (WGBS) to become the gold standard in DNA methylation analysis(24). One of the limitation of using WGBS for DNA methylation analyses on cfDNA is the need for deep sequencing (20, 22) which currently limits the wide-scale implementation in a clinical setting. Low depth sequencing in high sample numbers is a cost-effective strategy for cohort studies.(25) Utilizing reduced sequencing volume, low-pass sequencing and correspondingly low sequencing cost will be crucial to facilitate an easier clinical deployment of DNA methylation-based surveillance tools.

In this study, we evaluated the performance of low-pass whole genome bisulfite sequencing (WGBS) in cfDNA methylation profiling to identify the lowest sequencing depth for long-range methylation measurements. Applying the approach to a comparison of patients at different stages of liver diseases (hepatitis, cirrhosis and HCC), low-pass WGBS at 5-millilion sequencing reads was able to detect the hypo-methylation profiles of plasma cfDNA from patients with liver diseases. Additionally, we illustrate the efficacious diagnostic performance of low-pass WGBS assessing the methylation status of HBV integration regions.

## Results

### Efficacy of low pass sequencing strategy illustrated by re-sampling reads from cell-free WGBS data

In order to determine the impact of sequencing depth on methylation profiles in cell-free based WGBS data, we conducted a pilot study with 5 samples: one healthy individual (D1), one patient with hepatitis (D2), one patient with cirrhosis (D3) and 2 HCC patients (D4 and D5 of before and after surgery). The final read count equated to a mean of 58 million (M) reads per sample (**Supplementary Table 1**). The average DNA methylation across the genome was much lower in the HCC patient (D4; 53.56%) compared to healthy individual, hepatitis and cirrhosis (74.76%, 75.13 and 75.64%; **Supplementary Table 1**). To measure the methylation status of cfDNA in these samples we applied long range methylation (LRM). To identify the optimal region size of LRM, we divided the HCC genome (D4) into 500-Kb, 1-Mb, 1.5-Mb, 2-Mb and 2.5-Mb bins, respectively. For each region size, we calculated the average methylation level within each window for the genome. Then the percentage of regions with hypo-methylation (corresponding bin in D4 is less than -0.2 compared to healthy individual; see methods for details) was calculated in D4. The percentage of hypo-methylated regions was largest at the size of 2-Mb (**Figure S1; Supplementary Table 1**). So the LRM for all 1,382 autosomal 2-Mb regions were used for global methylation level calculation (**Materials and Methods**).

To determine the effective sequencing depth in low pass WGBS of cfDNA, we randomly sampled 1M to 10M mappable reads from each sequencing dataset (each composed of approximately 58M reads) and calculated the average methylation level for each 2-Mb region (MethylLRM). In each iteration, we calculated MethylLRM forall 2-Mb regions, and adopted correlation coefficient to show their consistency with those based on total sequencing reads. For each sequencing depth, we repeated the random extraction 100 times to examine the variation of the correlation coefficient, and the difference (coefficient of variation, CV) among 100 values of the correlation coefficient to assess sampling bias. We confirmed a high correlation between our low pass WGBS results as compared to all reads, with a CV below 4% in most of our samples (**Fig 1**). As predicted, when we increase the number of sequencing reads, MethylLRM was closer to the value calculated using total sequencing reads (**Fig 1**). The correlation coefficient between the methylation level from low-pass WGBS and the raw WGBS data saturates when using 5M or more reads. The correlation coefficient between MethylLRM at 5M reads and all sequencing reads was above 0.92 (Pearson’s correlation coefficient, P < 2.2x10-16, **Figure S2A-B**), and methylation level remained consistent after resampling 100-times (CV is 0.72%, 0.11%, 1.09%, 0.13%, 0.38% for D1, D2, D3, D4 and D5, respectively, **Fig 1**). In summary, we show how 5M mappable reads without redundancy in low pass WGBS is a reliable method to evaluate the methylation level of cfDNA samples in the long-range mode.

### Methylation level of plasma cfDNA from hepatitis and cirrhosis patients resembles healthy individuals

We next sought the ability of low pass WGBS of cfDNA to discriminate patients with different liver disease. Hence, we conducted low pass WGBS in plasma cfDNA of 54 individuals, including 17 HCC (3 early stage HCC, 5 advanced HCC and 9 HCC patients after surgery), 17 with hepatitis, 17 with cirrhosis and 3 healthy volunteers (**Supplementary Table 2**). On average, 10.2M mappable reads were obtained from each sample (IQR=6.3M, **Supplementary Table 3**). To evaluate the methylation levels in these samples, the LRM strategy was applied to define the hyper- or hypo-methylated LRM regions (**Materials and Methods**), using MethylLRM in healthy individuals as the baseline level. The percentage of hyper- or hypo-methylated LRM regions is shown for each patient (**Fig 2; Supplementary Table 3**). In hepatitis and cirrhosis patients, we found that hyper-long methylated regions (hyper-LMRs) accounted for <3% of total 1382 autosomal LMRs (**Fig 2A**), while hypo-long methylated regions (hypo-LMRs) accounted for 0.0-20.04% of the total LMRs, with only three patients exceeding 10% (**Fig 2B; Supplementary Table 3**), which showed that hepatitis and cirrhosis patients had similar cfDNA methylation levels with healthy individuals(**Fig 2**). Further, In early stage HCC patients, no hyper-LMR were identified, however hypo-LMRs accounted for 1.2% to 26.2% of the total LMRs. In advanced HCC patients, no hyper-LMR were identified, and hypo-LMR accounted for more than 65.7% of the total LMRs. (**Fig 2; Supplementary Table 3**). As expected, after surgery, most HCC patients (8/9) demonstrated similar cfDNA methylation levels to healthy individuals and patients with hepatitis or cirrhosis. Nevertheless, one (P45) out of the nine HCC patients exhibited a higher proportion of hypo-LMRs after surgery (69.9%, **Fig 2B**; **Supplementary Table 3**), and died two months later due to tumor recurrence, suggesting that tumor cells remained in that individual. Our results demonstrate that LMR could serve as a dynamic biomarker reflecting the genome-wide demethylation process from normal tissues to HCC and hence could be used as a measure of surgical efficacy.

### Differentially methylated CpGs (DMCs) and genes (DMGs) identified by low-pass cell-free WGBS

We identified DMCs and DMGs with low-pass cell-free WGBS data, and very limited CpGs were covered by our assay. On average, each cfDNA sample had 61,018 CpGs with sequencing depth over 5 reads (**Method, Supplementary Table 3**). In total, advanced HCC patients had 1,695 DMCsidentified (**Supplementary Table 4**), of which all the DMCs were hypo-methylated compared to healthy individuals. Among those, 23 DMCs were located in seven genes: *HFM1, PMF1, PMF1-BGLAP, DLG2, SENP5, SLCO5A1, REXO1L1P*. In the three early stage HCC patients, we identified 93 DMCs (**Supplementary Table 5**), of which 86 were in common with those observed in advanced HCC patients and five were located within *PMF1* and *PMF1-BGLAP* (**Fig 3A; Table 1**). Relatively high proportions of hypo-LMRs (>10%) were observed in one hepatitis and two cirrhosis patients (**Fig 2B**), possibly indicating their high HCC risk. In total, all four clinical groups had 74 DMCs in common (**Fig 3A**), which suggested that methylation changes may occur in the early stages of liver disease progression prior to HCC. **Table 1** displayed the genes with DMCs in the four comparisons. In particular, *SENP5* gene had seven significantly hypo-methylated DMCs with consistently high sequencing coverage across all individuals (149 reads, on average, **Supplementary Fig 3**, and **Fig 3B**). Intriguingly, all 7 DMCs that we found in intron 2 of *SENP5* were located near previously reported HBV integration sites in HCC (**Fig 3C**) (26).

### Overrepresentation of DMCs in repeat regions and surrounding HBV integration sites

Genome feature distribution of CpGs illustrated that they tended to be located at intergenic region and repeat regions (Methods, **Supplementary Fig 4A**), and CpGs in repeat regions had much higher sequencing depth in this low pass sequencing strategy compared to those in other regions (P < 2.2x10-16, Wilcoxon rank sum test; **Supplementary Fig 4B**). On average, 64% of all these CpGs were in the repeat regions (**Supplementary Fig 4C**), and this percentage varied from 49% to 87% across the samples. Differential methylation analysis required the CpG sites having over five sequencing reads in all samples (**Method**), and the resulting CpGs were over represented in repeat regions. Finally, 91% of DMCs of advanced HCC patients were located within repeat regions (**Fig 4A**). Considering that repeat regions are a known feature of HBV integration sites (27, 28), we subsequently analyzed the location of DMCs relative to reported HBV integration sites (26, 29-34). Among the 1,695 DMCs observed in advanced HCC patients, eighteen completely overlapped with the HBV integration sites, including two in *SENP5* (**Supplementary Table 8**). Additionally, 36.5% of the DMCs were located within a 100bp region either upstream or downstream of integration sites, and 95.8% of DMCs were within 5Kbp (**Fig 4A**). Overall, these DMCs were more enriched in HBV integration sites compared to promoter and gene coding regions (**Fig 4B**).

In order to evaluate whether methylation levels of CpGs near HBV integration sites could mirror the hypo-methylation profiles of cfDNA from HCC patients, CpGs with read depth exceeding 5 reads were analyzed in all 54 samples within 100bp flanking HBV integration sites and calculated the percentage of hypomethylated CpGs. These CpGs were found to be significantly hypo-methylated in advanced HCC patients, with 9.6% to 59.1% of CpGs being hypo-DMCs, while the proportion was generally reduced (2.6-10.2%) in early stage HCC patients. (**Fig 4C; Supplementary Table 3**). Then, All CpGs from each sample was used to calculate the average methylation level of the CpGs within the 100bp of the reported HBV integration sites (MethylHBV**)**. The advanced HCC patients still showed significantly hypo-methylation level compared to healthy individuals (<66.5%; P = 0.03, Wilcoxon rank sum test; **Fig 4D**; **Supplementary Table 3)**. However, for early stage HCC patients, this methylation level was relatively higher, ranging from 67.2% to 71%. Additionally, a strong negative correlation was observed between MethylHBV and alpha-fetoprotein (AFP) levels (R = -0.63, P = 8.4x10-7, Pearson’s correlation coefficient; **Fig 4D, Figure S5**).

To further assess the diagnostic accuracy of MethylHBV in HCC patients, a receiver operating characteristic (ROC) curve analysis based on generalized linear regression was conducted for 37 individuals without HCC (3 healthy individuals, 17 patients with hepatitis and 17 patients with cirrhosis) and 8 HCC patients (3 early stage HCC and 5 advanced HCC), and the area under the curve (AUC) was 0.93 (95% CI: 0.84-1.00; Fig S6). From the ROC curve, the sensitivity and specificity were 97.29% and 75% when MethylHBV was 67.28% as the optimal cutoff point for tumor detection. Moreover, four patients with hepatitis or cirrhosis (P2, P14, P18, P19) exhibited MethylHBV levels that approached the cutoff value (**Fig 4D, Table 2**). We also applied random forest (RF) to test the prediction performance of the low-pass WGBD data and we found the overall out-of-bag prediction accuracy was 92.16%. We noticed 1 of 43 non-HCC sample was predicted to HCC while 8 HCC was predicted to non-HCC and 3 mistake classified HCC samples were all belong to early HCC. We also applied five-fold cross-validate and with 100 random resampling in RF approach and find the average sensitivity, specificity and accuracy in test dataset were 62.5%, 97.6% and 91.1% respectively. When we check the samples side-by-side, we found one hepatitis patient, P14, had the average methylation level at 67.4% and abnormal AFP level (141.9 ng/ml; **Table 2**). Importantly, this patient developed clinically diagnosed HCC within six months. Patient P19 was diagnosed with alcoholic cirrhosis, with the transaminase indicator and presented with mild jaundice at each of the follow-up time points. For patient P2 (chronic hepatitis) and P18 (NASH-related cirrhosis) both had no observed abnormal measurements and have not had a detected tumor. Our results indicate hypo-methylation in HBV integration regions could be a potential biomarker to evaluate the risk of the transformation from hepatitis and cirrhosis to HCC in patients with chronic hepatitis B infection.

## Discussion

Chronic hepatitis B (CHB) is a major cause of HCC worldwide. Professional societies recommend HCC surveillance in patients with CHB with abdominal imaging (usually ultrasound) with or without alpha-fetoprotein (AFP) every 6 months to increase the likelihood of an early stage HCC diagnosis, thereby enabling more effective clinical interventions. Unfortunately alpha-fetoprotein is neither highly sensitive nor highly specific, and there is an unmet clinical need for new non-invasive diagnostic tests, such as liquid biopsy using circulating tumor cells (35). Although WGBS of cell-free DNA has been shown effective for cancer detection (21), the cost of cfDNA WGBS in cancer patients is one of challenges for wide application. We explored the cfDNA methylome of hepatitis, cirrhosis and HCC patients and examined the feasibility of HCC detection using low-pass WGBS. We demonstrated the measurement of long-range methylation could be applied in low-pass cell-free WGBS at 5-million reads to reflect liver disease status of hepatitis, cirrhosis and HCC. Moreover, DNA hypomethylation in HBV integration regions was shown to be potential biomarkers for cancer prognosis. Our result showed that DNA methylation levels in HBV integration regions were negatively correlated with AFP level, which corroborate the reliability of our result. Clinically, [add more discussion about this result and AFP?]

In a landmark paper, Chan and colleagues applied genome-wide pervasive hypomethylation in hepatocellular carcinoma detection and shown low sequencing depth of ~10 million reads was available for the cell-free detection for cancer (18). In our study, we required only 5M qualified reads for low-pass WGBS for 54 samples, and there were 2 samples only having 3.6M reads (**Supplementary Table 3**). In a 100-iteration resampling procedure, the average correlation coefficient was larger than 0.9 using 3M reads (**Fig 1**)—theoretically sufficient to evaluate methylation levels. This indicates that sequencing depth could be decreased to ~3 million reads with long-range DNA methylation measurements without substantially compromising accuracy.

One limit of cfDNA detection is the application of early stage cancer. Our work showed that the sensitivity of early stage HCC detection was much lower than advanced HCC. For early stage HCC patients, P35 and P36, both the proportion of hypo-LMRs (1.23% and 4.7%) and the average methylation level around HBV integration sites (70.48% and 71.48%) were similar to the healthy individuals and hepatitis patients. Both of these two patients had small tumor sizes (P35, 1.5cm; P36, less than 2cm, three lesions; **Supplementary Table 2**). Another application of cfDNA methylation analysis is to evaluate the residual tumor or risk of tumor recurrence after surgery. We found the methylation level of cfDNA should resemble with those from healthy individuals or patients with other chronic liver diseases after complete HCC resection. We found one HCC patient after interventional therapy, P45, showed significantly hypomethylation of the CpGs near the HBV integration sites (60.87%), indicating presence of hepatic micro-metastasis. The patient died two months later in follow-up due to multiple and recurrent lesions. The cutoff follow up of the five patients were 12 months without clinical evidence of HCC recurrence, and the other three patients had recurrence within 6 months.

Previous studies have been shown that the fragmentation process of cfDNA is not random (36, 37). Our results showed low-pass WGBS for cfDNA tended to capture fragments from repeat regions and HBV integration sites. More than 49% of CpGs located in the repeat regions and had a higher sequencing depth. When decreasing the sequencing volume, overrepresentation of genomic repeat regions was observed in our data. This suggests that the signal from repeat regions could remain given adequate sequencing depth in low pass WGBS. Since HBV integrations tend to localize at repeat regions, DMCs of advanced HCC patients were also enriched in previously reported HBV integration sites. Notably, CpGs near the HBV integration sites were likely to have methylation levels reflecting hypo-methyation status of tumor genome in HCC patients, which can discriminate the HCC patients from patients in other stages of liver diseases.

We adopted an approach focusing on 100bp upstream and downstream regions from HBV integration sites as surrogate regions for plasma hypomethylation analysis in HCC patients. Although we chose HBV integration sites as the indicator, it does not necessarily indicate that the analysis is only suitable for patients with HBV infection. In our sample set, we also included 3 patients without HBV infection (P1, P18 and P19; **Supplementary Table 2**). While HBV integration regions have molecular features suitable for HBV integrations, we also demonstrated that methylation changes in HBV integration regions may be common in HCC and independent of HBV infection. Interestingly, we found hypomethylation in HBV integration regions have higher sensitivity for HCC diagnosis. For example, P14 (chronic hepatitis) had an average value of 67.4% for the hypo-methylation HBV integration indicator which slightly exceeded the cutoff for HCC. We followed up this patient and found that he was subsequently diagnoses with HCC within 6 months. The sample from a chronic hepatitis patient, P2, showed that the proportion of LMRs was 17.8% and the average methylation level around HBV integration sites was 67.7%. Using the sample from a clinical visit 6 months following the initial sample collection, the proportion of LMRs dropped to 1.1%, whereas the average methylation around HBV integration sites slightly increased to 69%. This patient had no observed abnormal measurements and have not had a detected HCC in follow-up, showing that MethylHBV is stable than genome-wide LMR. As a predictor of HCC, the most challenging aspect is to determine appropriate cutoffs for disease status, which necessitates large sample sizes in future studies. Nevertheless, our study successfully illustrated that it is necessary to monitor the patients with suspicious methylation changes in cfDNA according to multiple indicators, combining their prognostic signals to improve accuracy.

Although we have found some stable methylation patterns using low-pass WGBS, we still need to validate these findings in larger studies. Such studies can be used to further develop these approaches and improve the accuracy of HCC diagnoses and surveillance. Larger studies will enable the determination of accurate cutoff values for disease stages, especially for those with small tumors. Furthermore, we anticipate that blood samples from HCC patients at multiple time points hold strong utility in tracking disease progression.

## Materials and Methods

### Sample collection

All the blood samples of patients were collected from Beijing You’an Hospital. Healthy individuals enrolled by Beijing Institute of Genomics were collected as controls. The diagnosis was made according to the guidelines for the prevention and treatment of chronic hepatitis B: a 2015 update (38). We collected age, gender, HBV-status, tumor size and Alanine aminotransferase (ALT) test, Aspartate aminotransferase (AST) test, bilirubin test, Alpha-fetoprotein (AFP) test and other related clinical information for related samples. Meanwhile, HCC patients was classified to early and late stage by BCLC system. The study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki and was approved by the Ethics Committee of Beijing You’an Hospital and Beijing Institute of Genomics. An informed consent was obtained from all patients and volunteers.

### Cell free DNA extraction

Ten microliters (ml) of whole blood was collected from each patient in Streck Cell-Free DNA BCT® tubes (Streck, Omaha, NE) and immediately transmitted to Beijing Institute of Genomics. Upon arrival, the blood collected in Streck BCT tubes were centrifuged at 3,000 × g for 15 minutes at 4°C within two hours. Subsequently, the plasma was transferred into a fresh microcentrifuge tube, followed by a 2nd centrifugation at 16,000 × g for 10 minutes at room temperature. Five ml of resultant plasma was used for cfDNA extraction using a QIAamp Circulating Nucleic Acid Kit (Qiagen, Valencia, CA). After extraction, total DNA was quantified using a Qubit dsDNAHS Assay kit (Life technologies, Grand Island, NY, USA). All DNA samples were stored at -80°C before sequencing library construction.

### Whole genome bisulfite sequencing and data processing

Using the TruSeq DNA Methylation Kit (Illumina Inc.) according to the manufacturers’ protocol. Total cfDNA (range from 0.5 ng to 88.7 ng) was used for sequencing library construction. Bisulfite conversion of cfDNA was performed using the EZ DNA Methylation-Gold Kit (Zymo Research) according to the instruction manual. During conversion, 0.5% methylated lambda DNA was included as a spike-in DNA control to estimate the conversion efficiency of unmodified cytosine. The sequencing libraries were then performed paired-end sequencing (2 × 100 bp) on an Illumina HiSeq 4000 (Illumina Inc., San Diego, CA, USA). The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive(39) in BIG Data Center(40), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession numbers CRA001537, CRA001537 that are publicly accessible at [http://bigd.big.ac.cn/gsa](http://bigd.big.ac.cn/gsa" \t "_blank).

After base calling, all paired-end fastq files were trimmed using cutadapt (v 1.8.3)(41) to removed adapter sequences and low quality bases with parameters ‘-q 15 --minimum-length 36’. HG19 reference genome was downloaded from ENSEMBL. Lambda genome was also included in the reference sequence for calculating bisulfite conversion rate. Filtered paired-end bisulfite sequencing data were mapped with Bismark (v0.14.5)(42) using with default parameters. After alignment, read duplicates were removed using the deduplicate\_bismark application included in the bismark software. Then the BAM files produced by Bismark were sorted using samtools (v 0.1.19) and overlapping paired-end reads were clipped using ClipOverlap function of bamUtil (<https://github.com/statgen/bamUtil>) to prevent counting twice from the same observation. For each CpG, the methylation level was combined from both DNA strands and estimated as mi/(mi + ui), where mi was defined as the number of methylated cytosines and ui was defined as the number of unmethylated cytosines. The number of methylated and unmethylated cytosines of long range regions were generated using R package methylKit. The average methylation level of each long range region (MethylLRM) was calculated as the total number of cytosines divided by the number of methylated cytosines.

### Identification of the optimal region size of long range methylation (LRM)

The HCC genome was divided into 500-Kb, 1-Mb, 1.5Mb, 2-Mb and 2.5-Mb, respectively. For each size, the average methylation level for each region from autosome were calculated. The hypo-methylated region were identified as methylation level difference larger than 0.2 compared to the corresponding region in heathy individual. Then the percentage of hypo-methylated regions across the genome was calculated. The largest percentage of hypo-methylated region size was selected as the optimal size of LRM

### Randomly re-sampling lower reads from medium WGBS data

A randomly sampling method was used to obtain low depth WGBS for 5 medium WGBS of cell-free DNA. (a) 1M to 10M read pairs (increasing by 1M step) was randomly extracted from each medium WGBS data set. (b) For each re-sampling, the average methylation level for each 2-Mb region (MethylLRM) from autosome were calculated and pearson correlation coefficient was used to show the correlation of all the autosomal MethylLRM between this re-sampling reads and total WGBS reads. This process was repeated for 100 times. (c) For each re-sampling, coefficient of variation (CV) for correlation coefficient was calculated across 100 randomly re-sampling to examine the variability of 100 extraction.

### Identification of hyper-LRMs and hypo-LRMs

We adopted the method of Chan et al.(18) to define the hyper- or hypo- MethylLRM compared to the healthy ref group. Only autosomes were included in this analysis. A 2-Mb region of a sample was defined as hyper- or hypo-methylated if its average methylation level was at least 3 SDs above or below the mean of the corresponding region of the healthy individuals. Then the number and percentage of hyper- or hypo- MethylLRM within the genome was calculated.

### Identification and annotation of the differentially methylated CpGs (DMCs) and genes (DMGs)

The identification of DMCs was generated using R package methylKit. The significance of the DMCs between case (hepatitis, cirrhosis, early stage HCC, advanced HCC) group and healthy group was performed by logistic regression test with at least 5-fold coverage. P-value was adjusted for multiple testing with the method of Hochberg and Benjamini. The CpG sites were considered differentially, if the Benjamini Hochberg corrected P value ≤ 0.05 and the methylation level difference was ≥ 0.2. Each DMCs was annotated for each RefSeq transcript obtained from ENSEMBL GRCh37. Promoters are defined as regions 2kb upstream from TSS for each RefSeq transcript. RepeatMasker annotations were obtained from UCSC.

### The enrichment score in each genomic region

The enrichment score for CpGs or DMCs was calculated by the following formula:The enrichment scorein the genomic element = log2 (# DMCsin the genomic element/# expected). # expected was computed as: # DMCsin the genome × # CpG sitesin the genomic element/# total CpG sitesin the genome. # means the number of sites.

### DNA methylation of CpGs near the HBV integration sites in hepatitis, cirrhosis and HCC

Identification of hypo-CpGs within the 100 bp upstream or downstream of HBV integration sites. The HBV integration sites were extracted from previous reports (26, 29-34).We extracted CpG within the 100 bp upstream or downstream of HBV integration sites. Only autosomal CpGs and CpGs with depth over 5 in all the 54 samples were included in the hypo-CpGs analysis. Similar to the identification of hypo-LRMs, a CpG of a sample was defined as hypo-methylated if its methylation level was 3 SDs or more below the mean of the corresponding CpGs of the healthy individuals. Then the percentage of hypo-CpGs was calculated. Average methylation level of the CpGs within the 100bp of the HBV integration sites. For each sample, the average methylation level of CpGs within the 100 bp upstream or downstream of HBV integration sites was included in all the CpGs. This value was calculated as the number of the total number of cytosines divided by the number of methylated cytosines within the 100bp of the HBV integration sites.

### Prediction analysis, Random Forest and ROC curves

We applied generalized linear regression to fit the prediction model with differential methylation regions as the predictor variables. In order to avoid over-fitting, we also applied five-fold cross-validation based random forest to make unbiased evaluation to the prediction. Random Forest approach was conducted with R package *randomForest*. Analysis of receiver operating characteristics (ROC) curves was constructed using R package pROC(43). The optimal cutoff was determined using the “coords" function. The optimal cut-off was the threshold that maximized the distance to the identity (diagonal) line.

**Acknowledgements**

This study is funded by Innovation Promotion Association CAS (2016098) and National Natural Science Foundation of China (81201700) to D.Z., Major State Basic Research Development Program (2014CB542006), the Key Research Program of the Chinese Academy of Sciences (KJZD-EW-L14) to C.Z., Capital's Funds for Health Improvement and Research (2018-1-1151) to P.D., and NLM training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM 5T15LM007359) to S.G.

**Authorship Contributions**

HZ and SG performed analyses, developed analysis methods and power calculations, interpreted results, and drafted the manuscript. PD enrolled patients and collected all the clinical information. CT and JK conducted sequencing experiments. ZW collected and prepared tissue samples for sequencing analysis and collected results of clinical assays. RC and AV interpreted results, provided liver cancer and hepatology clinical expertise, reviewed and edited the manuscript. HD aided in the analyses and reviewed the manuscript. HD provided clinical advice and reviewed the manuscript. XXX, DZ and CZ designed the study, supervised all experiments and analysis, provided molecular and cellular biology advice, reviewed and edited the manuscript.

**Disclosure of Conflicts of Interest**

The authors declare no conflict of interest.

**Abbreviations**

LRM Long-Region Methylation

HCC Hepatocellular Carcinoma

DMCs Differential Methylation CpGs

DMGs Differential Methylation Genes

HBV Hepatitis B virus

HCV Hepatitis C virus

GWBS Genome-wide Bisulfite Sequencing

RRBS [Reduced Representation Bisulfite Sequencing](https://en.wikipedia.org/wiki/Reduced_representation_bisulfite_sequencing)

cfDNA Circulating cell-free DNA

## Reference

1. Fleischhacker M, Schmidt B. Circulating nucleic acids (CNAs) and cancer - A survey. Biochimica Et Biophysica Acta-Reviews on Cancer 2007;1775:181-232.

2. Chan AK, Chiu RW, Lo YM, Clinical Sciences Reviews Committee of the Association of Clinical B. Cell-free nucleic acids in plasma, serum and urine: a new tool in molecular diagnosis. Ann Clin Biochem 2003;40:122-130.

3. Stroun M, Maurice P, Vasioukhin V, Lyautey J, Lederrey C, Lefort F, Rossier A, et al. The origin and mechanism of circulating DNA. Ann N Y Acad Sci 2000;906:161-168.

4. Schwarzenbach H, Hoon DSB, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nature Reviews Cancer 2011;11:426-437.

5. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, Pacey S, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer 2017;17:223-238.

6. Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat Genet 2017;49:635-642.

7. Shen SY, Singhania R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, Zuzarte PC, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature 2018;563:579-583.

8. JQ X. Trends in liver cancer mortality among adults aged 25 and over in the United States, 2000–2016. NCHS Data Brief, no 314 2018.

9. Chen CJ, Yu MW, Liaw YF. Epidemiological characteristics and risk factors of hepatocellular carcinoma. J Gastroenterol Hepatol 1997;12:S294-308.

10. Montesano R, Hainaut P, Wild CP. Hepatocellular carcinoma: from gene to public health. J Natl Cancer Inst 1997;89:1844-1851.

11. Stauffer JK, Scarzello AJ, Jiang Q, Wiltrout RH. Chronic inflammation, immune escape, and oncogenesis in the liver: a unique neighborhood for novel intersections. Hepatology 2012;56:1567-1574.

12. Aihara T, Noguchi S, Sasaki Y, Nakano H, Imaoka S. Clonal analysis of regenerative nodules in hepatitis C virus-induced liver cirrhosis. Gastroenterology 1994;107:1805-1811.

13. Schutte K, Bornschein J, Malfertheiner P. Hepatocellular carcinoma--epidemiological trends and risk factors. Dig Dis 2009;27:80-92.

14. European Association for the Study of the Liver. Electronic address eee, European Association for the Study of the L. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. J Hepatol 2018;69:182-236.

15. Labgaa I, Villacorta-Martin C, D'Avola D, Craig AJ, von Felden J, Martins-Filho SN, Sia D, et al. A pilot study of ultra-deep targeted sequencing of plasma DNA identifies driver mutations in hepatocellular carcinoma. Oncogene 2018;37:3740-3752.

16. Qu C, Wang Y, Wang P, Chen K, Wang M, Zeng H, Lu J, et al. Detection of early-stage hepatocellular carcinoma in asymptomatic HBsAg-seropositive individuals by liquid biopsy. Proc Natl Acad Sci U S A 2019;116:6308-6312.

17. Bhan I, Mosesso K, Goyal L, Philipp J, Kalinich M, Franses JW, Choz M, et al. Detection and Analysis of Circulating Epithelial Cells in Liquid Biopsies From Patients With Liver Disease. Gastroenterology 2018;155:2016-2018 e2011.

18. Chan KC, Jiang P, Chan CW, Sun K, Wong J, Hui EP, Chan SL, et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proc Natl Acad Sci U S A 2013;110:18761-18768.

19. Zhao Y, Xue F, Sun J, Guo S, Zhang H, Qiu B, Geng J, et al. Genome-wide methylation profiling of the different stages of hepatitis B virus-related hepatocellular carcinoma development in plasma cell-free DNA reveals potential biomarkers for early detection and high-risk monitoring of hepatocellular carcinoma. Clin Epigenetics 2014;6:30.

20. Sun K, Jiang P, Chan KC, Wong J, Cheng YK, Liang RH, Chan WK, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci U S A 2015;112:E5503-5512.

21. Lehmann-Werman R, Neiman D, Zemmour H, Moss J, Magenheim J, Vaknin-Dembinsky A, Rubertsson S, et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc Natl Acad Sci U S A 2016;113:E1826-1834.

22. Kang S, Li Q, Chen Q, Zhou Y, Park S, Lee G, Grimes B, et al. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome Biol 2017;18:53.

23. Xu RH, Wei W, Krawczyk M, Wang W, Luo H, Flagg K, Yi S, et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. Nat Mater 2017;16:1155-1161.

24. Li H, Jing C, Wu J, Ni J, Sha H, Xu X, Du Y, et al. Circulating tumor DNA detection: A potential tool for colorectal cancer management. Oncol Lett 2019;17:1409-1416.

25. Liu S, Huang S, Chen F, Zhao L, Yuan Y, Francis SS, Fang L, et al. Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. Cell 2018;175:347-359 e314.

26. Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, Lee NP, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet 2012;44:765-769.

27. Tu T, Budzinska MA, Shackel NA, Urban S. HBV DNA Integration: Molecular Mechanisms and Clinical Implications. Viruses 2017;9.

28. Yan H, Yang Y, Zhang L, Tang G, Wang Y, Xue G, Zhou W, et al. Characterization of the genotype and integration patterns of hepatitis B virus in early- and late-onset hepatocellular carcinoma. Hepatology 2015;61:1821-1831.

29. Jiang S, Yang Z, Li W, Li X, Wang Y, Zhang J, Xu C, et al. Re-evaluation of the carcinogenic significance of hepatitis B virus integration in hepatocarcinogenesis. PLoS One 2012;7:e40363.

30. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat Genet 2012;44:760-764.

31. Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, Kennemer MI, Guan Y, Lee W, et al. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Res 2012;22:593-601.

32. Ding D, Lou X, Hua D, Yu W, Li L, Wang J, Gao F, et al. Recurrent targeted genes of hepatitis B virus in the liver cancer genomes identified by a next-generation sequencing-based approach. PLoS Genet 2012;8:e1003065.

33. Li W, Zeng X, Lee NP, Liu X, Chen S, Guo B, Yi S, et al. HIVID: an efficient method to detect HBV integration using low coverage sequencing. Genomics 2013;102:338-344.

34. Toh ST, Jin Y, Liu L, Wang J, Babrzadeh F, Gharizadeh B, Ronaghi M, et al. Deep sequencing of the hepatitis B virus in hepatocellular carcinoma patients reveals enriched integration events, structural alterations and sequence variations. Carcinogenesis 2013;34:787-798.

35. Palmirotta R, Lovero D, Cafforio P, Felici C, Mannavola F, Pelle E, Quaresmini D, et al. Liquid biopsy of cancer: a multimodal diagnostic tool in clinical oncology. Ther Adv Med Oncol 2018;10:1758835918794630.

36. Jiang P, Sun K, Tong YK, Cheng SH, Cheng THT, Heung MMS, Wong J, et al. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. Proc Natl Acad Sci U S A 2018.

37. Chan KC, Jiang P, Sun K, Cheng YK, Tong YK, Cheng SH, Wong AI, et al. Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. Proc Natl Acad Sci U S A 2016;113:E8159-E8168.

38. Hou J, Wang G, Wang F, Cheng J, Ren H, Zhuang H, Sun J, et al. Guideline of Prevention and Treatment for Chronic Hepatitis B (2015 Update). J Clin Transl Hepatol 2017;5:297-318.

39. Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, Tang B, et al. GSA: Genome Sequence Archive<sup/>. Genomics Proteomics Bioinformatics 2017;15:14-18.

40. Members BIGDC. Database Resources of the BIG Data Center in 2019. Nucleic Acids Res 2019;47:D8-D14.

41. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 2011;17:3.

42. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 2011;27:1571-1572.

43. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77.

## Figure Legends

**Fig. 1. The efficiency of re-sampling sequencing depth for low pass WGBS.** Left of the figure showed the correlation coefficient between re-sampling low pass WGBS and total sequencing reads for 100 times from 1M to 10M. Right of the figure showed the coefficient of variation (CV) for 100 correlation coefficient between re-sampling low pass WGBS and total sequencing reads from 1M to 10M.

**Fig. 2 Whole genome-wide changed methylation of all the patients.** (A) The percentage of hyper-methylated long range regions (2-Mb) in chronic hepatitis, cirrhosis and HCC patients. (B) The percentage of hypo-methylated long range regions in chronic hepatitis, cirrhosis and HCC patients.

**Fig. 3. Differentially methylated CpGs (DMCs) identified in all the groups.** (A) Venn diagram showing the overlap of DMCs generated by 2 hypo-methylated chronic hepatitis patients, 1 hypo-methylated cirrhosis patient, 3 early stage HCC patients and 5 advanced HCC patients compared to healthy individuals. (B) Boxplot displays the methylation level of 7 DMCs of SENP5 in 3 healthy individuals, 17 hepatitis, 17 cirrhosis, 3 early stage HCC, 5 advanced HCC and 9 HCC patients after surgery. (D) The locus of 7 DMCs and 3 reported HBV integration sites in intron 2 of SENP5. The black dots represent the HBV integration sites and the orange vertical lines represent the 7 DMCs. The black bar labels in the bottom of the figure represent the locus of repeat marker in this region.

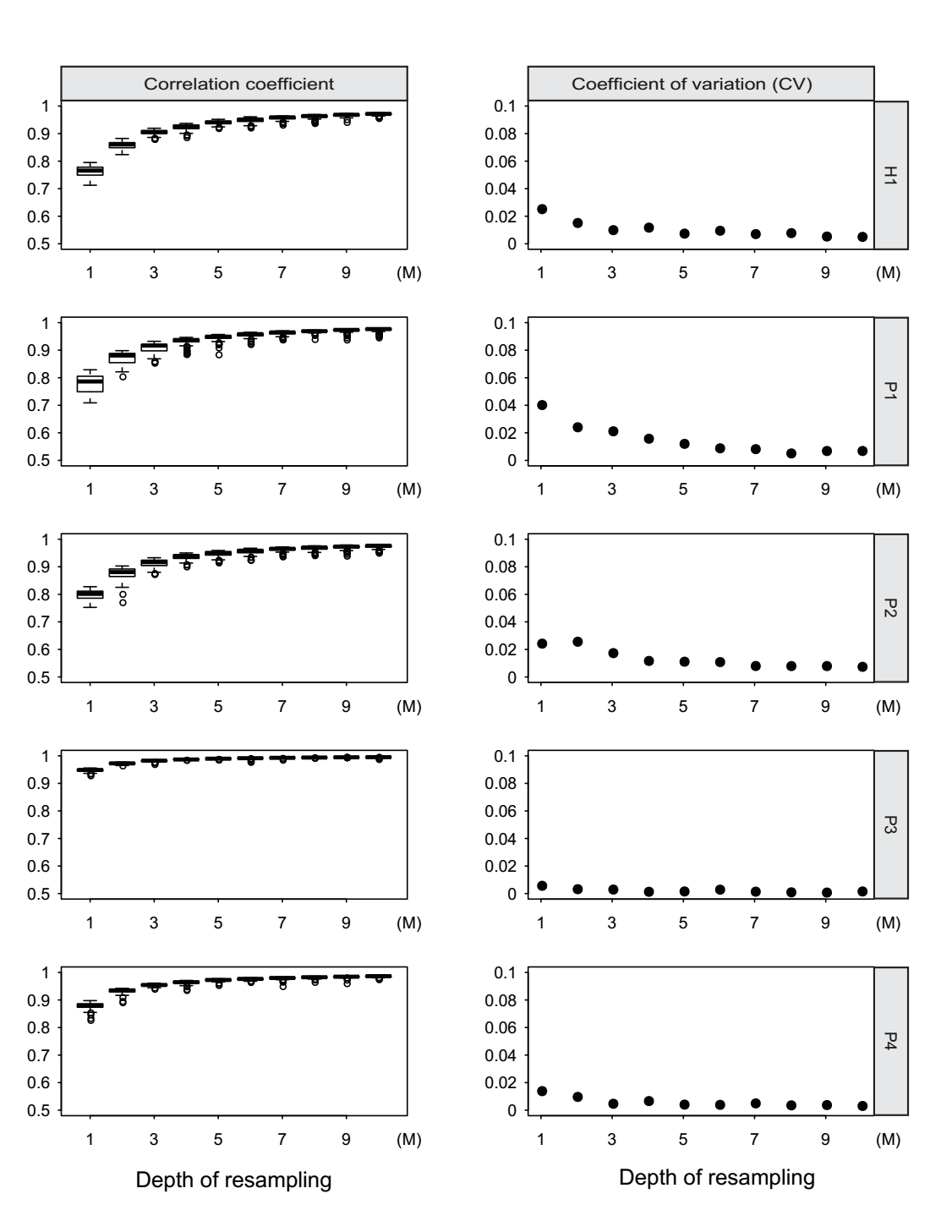
**Fig. 4. DMCs and CpGs are related to HBV integration sites.** (A) The percentage of DMCs located in different genomic elements and regions related to HBV integration sites. (B) The enrichment scores of DMCs in different genomic elements. (C) The heatmap display the methylation level of the CpGs located within 100 bp of the HBV integration sites in all the samples. (D) The average methylation level of all the CpGs located within 100 bp of the HBV integration sites in all the samples. The red arrows showed the examples of P14 patient.

**Table 1. Genes with DMCs between liver disease patients and healthy individuals.**

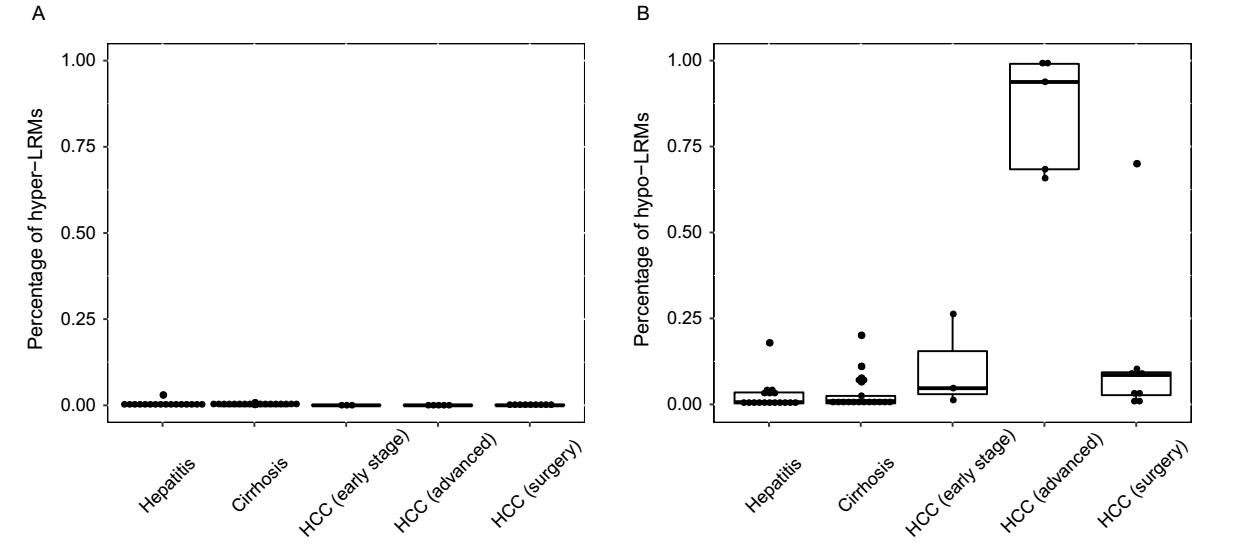
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chr** | **Start** | **End** | **CpG num** | **Position** | **Gene** |
| **Hepatitis vs. healthy** | | | | | |
| chr1 | 156186377 | 156186549 | 18 | gene body | PMF1 |
| chr1 | 156186377 | 156186549 | 18 | gene body | PMF1-BGLAP |
| chr1 | 91853073 | 91853096 | 3 | gene body | HFM1 |
| chr11 | 85195089 | 85195090 | 1 | gene body | DLG2 |
| **Cirrhosis vs. healthy** | | | | | |
| chr1 | 156186492 | 156186493 | 1 | gene body | PMF1 |
| chr1 | 156186492 | 156186493 | 1 | gene body | PMF1-BGLAP |
| chr8 | 70602451 | 70602487 | 2 | gene body | SLCO5A1 |
| **Early stage HCC vs. healthy** | | | | | |
| chr1 | 156186410 | 156186530 | 5 | gene body | PMF1 |
| chr1 | 156186410 | 156186530 | 5 | gene body | PMF1-BGLAP |
| **Advanced HCC vs. healthy** | | | | | |
| chr1 | 91852857 | 91852974 | 2 | gene body | HFM1 |
| chr1 | 156186392 | 156186540 | 10 | gene body | PMF1 |
| chr1 | 156186392 | 156186540 | 12 | gene body | PMF1-BGLAP |
| chr3 | 196625630 | 196625734 | 7 | gene body | SENP5 |
| chr8 | 70602486 | 70602487 | 1 | gene body | SLCO5A1 |
| chr8 | 86572360 | 86572383 | 2 | gene body | REXO1L1P |
| chr11 | 85195100 | 85195101 | 1 | gene body | DLG2 |

**Table 2. The information of hepatitis and cirrhosis patients with lower methylation levels of CpGs located in the 100 bp of HBV integration sites**

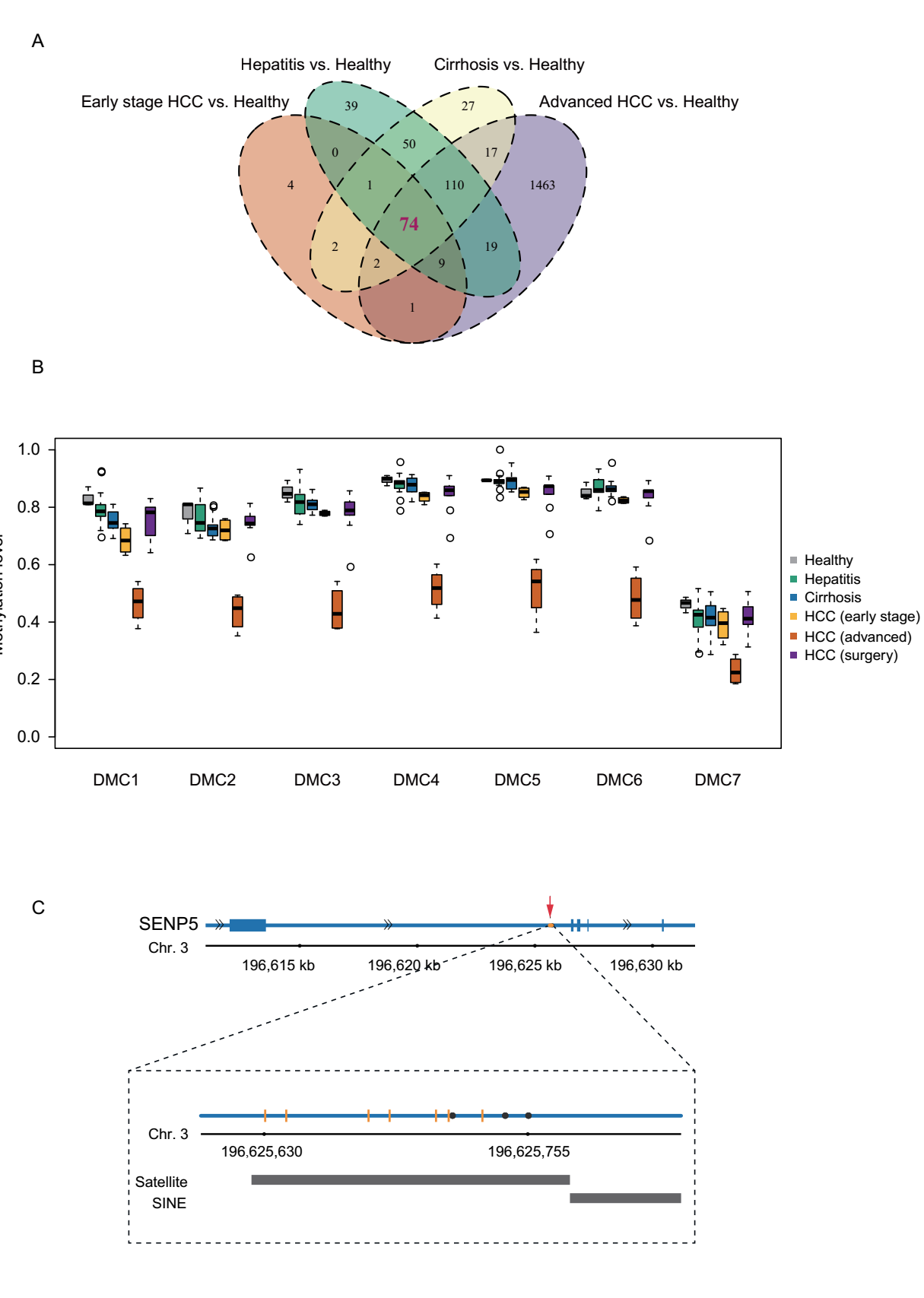
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient** | **Disease** | **Percentage of hypo LRMs** | **MethylHBV** | **AFP (ng/ml)** |
| P2 | chronic hepatitis | 17.87% | 67.69% | 2.2 |
| P14 | cirrhosis | 3.47% | 67.39% | 141.9 |
| P18 | nash-related cirrhosis | 20.04% | 66.04% | 2.07 |
| P19 | alcoholic cirrhosis | 11.00% | 67.96% | 8.46 |



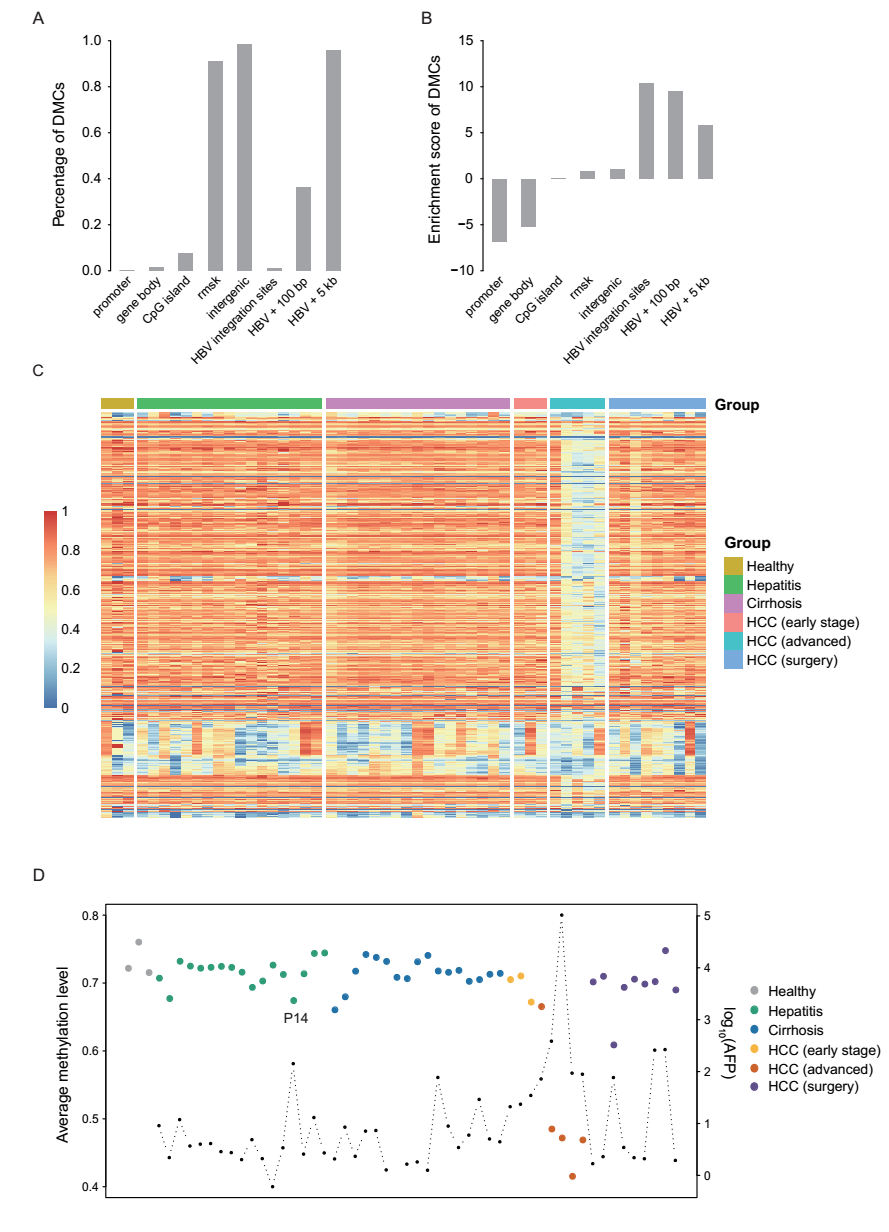
**Fig. 1. The efficiency of re-sampling sequencing depth for low pass WGBS.** Left of the figure showed the correlation coefficient between re-sampling low pass WGBS and total sequencing reads for 100 times from 1M to 10M. Right of the figure showed the coefficient of variation (CV) for 100 correlation coefficient between re-sampling low pass WGBS and total sequencing reads from 1M to 10M.



**Fig. 2 Whole genome-wide changed methylation of all the patients.** (A) The percentage of hyper-methylated long range regions (2-Mb) in chronic hepatitis, cirrhosis and HCC patients. (B) The percentage of hypo-methylated long range regions in chronic hepatitis, cirrhosis and HCC patients.



**Fig. 3. Differentially methylated CpGs (DMCs) identified in all the groups.** (A) Venn diagram showing the overlap of DMCs generated by 2 hypo-methylated chronic hepatitis patients, 1 hypo-methylated cirrhosis patient, 3 early stage HCC patients and 5 advanced HCC patients compared to healthy individuals. (B) Boxplot displays the methylation level of 7 DMCs of SENP5 in 3 healthy individuals, 17 hepatitis, 17 cirrhosis, 3 early stage HCC, 5 advanced HCC and 9 HCC patients after surgery. (D) The locus of 7 DMCs and 3 reported HBV integration sites in intron 2 of SENP5. The black dots represent the HBV integration sites and the orange vertical lines represent the 7 DMCs. The black bar labels in the bottom of the figure represent the locus of repeat marker in this region.



**Fig. 4. DMCs and CpGs are related to HBV integration sites.** (A) The percentage of DMCs located in different genomic elements and regions related to HBV integration sites. (B) The enrichment scores of DMCs in different genomic elements. (C) The heatmap display the methylation level of the CpGs located within 100 bp of the HBV integration sites in all the samples. (D) The average methylation level of all the CpGs located within 100 bp of the HBV integration sites in all the samples. The red arrows showed the examples of P14 patient.