**Long-region hypo-methylation in HBV integration regions aids non-invasive surveillance to hepatocellular carcinoma by low-pass genome-wide bisulfite sequencing.**

Haikun Zhang1,8, #, Peiling Dong2, #, Shicheng Guo3, #, Chengcheng Tao1, Wei Chen1,4, Wenmin Zhao2, Jiakang Wang5, Ramsey Cheung6, Augusto Villanueva7, Jian Fan1,9, Huiguo Ding2, Steven J. Schrodi3,8,\*, Dake Zhang1,4\*, 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

4Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, Beihang University, Beijing 100083 China

5Biology Department, Stonybrook University, Stonybrook, NY, US

6 Department of Gastroenterology and Hepatology, VA Palo Alto Health Care System and Stanford University, Palo Alto, CA, USA

7Liver Cancer Research Program, Division of Liver Diseases, Tisch Cancer Institute, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

8Department of Medical Genetics, University of Wisconsin-Madison, Madison, WI, USA

9University of Chinese Academy of Sciences, Beijing 100049, China

# These authors contributed equally to this work; \* Corresponding Author

Steven J. Schrodi, Ph.D.

Department of Medical Genetics

School of Medicine and Public Health

University of Wisconsin-Madison

Madison, WI, 53706

Tel: (608) 262-1069

Email: [schrodi@wisc.edu](mailto:schrodi@wisc.edu)

Dake Zhang, Ph.D.

Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics

Chinese Academy of Sciences, Beijing, 100101,

Tel: (010) 8409-7566

Email: [dakezhang@gmail.com](mailto:dakezhang@gmail.com)

Changqing Zeng, Ph.D.

Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics

Chinese Academy of Sciences, Beijing, 100101,

Tel: (010) 8409-7818

Email: [czeng@big.ac.cn](mailto:czeng@big.ac.cn)

## Abstract

Circulating cell-free DNA (cfDNA) methylation has been demonstrated to be a promising approach for non-invasive cancer diagnosis. However, the low-level of cfDNA and high cost of whole genome bisulfite sequencing (WGBS) significantly hinders the clinical implementation of a methylation-based cfDNA early detection biomarker. Here we proposed a novel method in which we utilized long-region methylation around HBV integration regions in low-pass WGBS data (~5 million reads) generated from cfDNA to detect methylation changes. We found a significant enrichment of CpGs and differential methylation loci in intergenic and repeat regions, especially in HBV integration sites using low-pass WGBS of cfDNA. Moreover, methylation profiles nearby HBV integration sites (MethylHBV) were found to enhance the prediction performance. We applied the method to investigate dynamic methylation changes in cfDNA from blood samples of patients with hepatitis, cirrhosis, early and advanced hepatocellular carcinoma (HCC). Models based on DNA methylation within 5kb flanking HBV integration sites and specifically hypomethylated long range regions demonstrated that low-pass cfDNA methylation data exhibited powerful discrimination ability as quantified by receiver operating characteristic (ROC) curves. The results demonstrate that low-pass cfDNA WGBS could be used as a low-cost and minimaly invasive approach for early HCC detection in the context of surveillance programs.

Overall, these DMCs were more enriched in HBV integration sites compared to promoter and gene coding regions

which supports DNA methylation around HBV integration sites as a more sensitive indicator to detect HCC compared to average methylation level across the genome.

Strikingly, compared to tumor tissue and buffy coat, cfDNA samples were less enriched functional elements (CpG island, promoter and exon) and more enriched in intergenic, repeat regions and HBV integration regions in both randomly 10M reads and deep sequencing. (**3C**; **Fig S5C**), suggesting this enrichment is a feature of cfDNA and the bias of cfDNA release.

## Introduction

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 (1, 2). Hepatocellular carcinoma (HCC), the most frequent form of primary liver cancer, generally develops in patients with chronic liver disease due to hepatitis B virus (HBV), hepatitis C virus (HCV), alcohol abuse or non-alcoholic fatty liver disease (3, 4). 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 (5-7). The high risk of HCC development in patients with cirrhosis (i.e., 2-7% annual risk) justifies the recommendation of biannual HCC surveillance with abdominal ultrasound (US) with or without serum alpha-fetoprotein (AFP) in patients at high-risk (8). 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 (9), which underscores the need for improved early detection tools.

Circulating cell-free DNA (cfDNA) are small double-stranded DNA fragments (10) found in plasma, urine, saliva, cerebrospinal fluid (CSF) and other body fluids (11) originating of cell apoptosis and necrosis (12). In many settings, analyses of cfDNA can be regarded as a way to perform a “liquid biopsy”, which have been produced promising results for genetic testing (13, 14), early cancer detection (15, 16) and prognosis prediction (17, 18). Apoptotic and necrotic tumor cells can release cfDNA into the peripheral blood, which reflects tumor-related genetic features, including cfDNA fragment size (cfDNAsize) (19), mutations, copy number aberrations and epigenetic changes (17). Meanwhile, cfDNA also carries tissues-specific information which provides promising abilities for tissue-of-origin mapping (20-24). As such, cfDNA could be used as an important biomarker in clinical settings. There are different technologies to investigate methylation changes in cfDNA, including scRRBS (20) and cfMeDIPseq (23). A number of studies have focused on cfDNA as a potential source of novel early detection biomarkers in HCC based on DNA methylation (25-30). Multiple studies have focused on the use of cfDNA methylation in cancer diagnosis in the areas of specific biomarkers (26, 30), hypo-methylation (25) and tissue of origin (27-29). Single cytosine measurement and high accuracy have enabled whole genome bisulfite sequencing (WGBS) to become the gold standard in DNA methylation analysis (31). One challenge in detecting circulating tumor DNA (ctDNA) in plasma is the minor fraction of ctDNA amidst the background of total cfDNA. This is particularly true in patients with early stage cancers and in the minimal residual disease setting, which benefits from deep sequencing producing a more sensitive indicator for early cancer detection and surveillance (27, 29, 32). That said, low depth sequencing in high sample sizes is a cost-effective strategy for cohort studies (33). 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 investigated the performance of low-pass WGBS in cfDNA methylation profiling and HCC prediction. We explored the DNA methylation state around HBV integration regions of HCC patients. We evaluated the minimum sequencing depth for long-range methylation measurements flanking 5kb of reported HBV integration sites (MethylHBV) and provided the landscapes of low-pass WGBS in healthy individuals, cirrhosis, hepatitis and HCC patients. Finally, we proposed methylation around HBV-integration regions (MethylHBV) and specifically hypo-methylated regions with length larger than 1Mb carry utility to predict HCC from non-HCC samples.

## Results

### DNA methylation around HBV integration sites mirrors the hypo-methylation of HCC patients

In order to explore methylation profiles in cell-free based WGBS data, we conducted a pilot study with 5 cfDNA samples using high-depth WGBS: one healthy individual (D1), one patient with chronic 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 (**Table S1**; **Fig S1A**). The average DNA methylation across the genome was much lower in the HCC patient (D4; 53.56%) compared to healthy individual, chronic hepatitis and cirrhosis (74.76%, 75.64% and 75.13%; **Table S1**). We found the genome distribution of CpGs in WGBS data tended to be located at intronic, intergenic and repeat regions (**Fig 1A**).

Next, we identified differentially methylated CpGs (DMCs) and differentially methylated gene (DMGs) with cell-free WGBS data between HCC patient and healthy individual. On average, each cfDNA sample had 7,274,674 CpGs with sequencing depth over 5 reads (**Table S1**). In total, we identified 2,670 DMCs in HCC patient compared to healthy individual (D4 vs. D1; **Table S2**), of which 99.8% were hypo-methylated in HCC patients. These hypo-DMCs clearly separated HCC patient from healthy individual, patient with chronic hepatitis, cirrhosis and HCC patient after surgery (**Figure 1B**). Among hypo-DMCs, 174 DMCs (6.5% of 2,670 DMC) were located in promoter or gene body of 77 genes. In particular, *SENP5* gene had six significantly hypo-methylated DMCs with consistently high sequencing coverage across all the five individuals (an average of 295 reads; **Fig 1B** and **Fig S1B**). Intriguingly, all six DMCs that we found in intron 2 of *SENP5* were located near previously reported HBV integration sites in HCC (**Fig 1C**) (34).

We found that 80% of DMCs of HCC patients were located within repeat regions (**Fig 1D**). Considering that repeat regions are a known target for HBV integration (35, 36), we analyzed the location of DMCs relative to reported HBV integration sites (34, 37-42). Totally, we collected 6072 HBV integration sites from published researches (Table S3). Among the 2,670 DMCs observed in HCC patient, 21 completely overlapped with the HBV integration sites, including one in *SENP5*. Additionally, 26.8% of the DMCs were located within a 100bp region either upstream or downstream of integration sites, and 73.9% of DMCs were within 5Kbp (**Fig 1D**). Overall, these DMCs were more enriched in HBV integration sites compared to promoter and gene coding regions (**Fig 1E**, P < 2.2x10-16, Fisher’s exact test).

Although CpGs (depth ≥ 5) of cell-free DNA were likely more to locate at HBV integration sites (**Fig 1A**, Fisher’s exact test), DMCs were more enriched in HBV integration sites compared to CpGs (**Fig 1A**; **Fig 1E**). With above findings, we further exmamined whether DNA methylation levels around HBV integration regions could represent the hypo-methylation of HCC genome and be used in optimization of prediction model for HCC. In HCC tumor tissues and paired buffy coat samples in a previous study (25), the hypomethalytion near the HBV intergration sites were observed in both tumor and buffy coat, and the closer to integration sites, the lower methylation levels. Methylation levels were further reduced in tumor tissue, especially within 100bp region near these sites (**Fig 1F and Fig S2**). We calculated the average methylation level of the CpGs within a region near HBV integration sites we collected (tMethylins) in each tissue sample, as the indicatior for methylation level **(**region size 100bp for tissue samples, **Methods)**. Although tMethylins was lower than average methylation level across the genome (Methylgenome) in both buffy coat and tumor tissue, and tumor samples had a significantly smaller tMethylins compared with buffy coat (P = 8.8x10-5, t test; **Fig S3**). Particlularly, tMethylins was significantly lower than Methylgenome in tumor tissue samples (P = 8.8x10-5, t test; **Fig 1G**), which supports DNA methylation around HBV integration sites as a more sensitive indicator to detect HCC compared to average methylation level across the genome.

### Hypomethalytion of regions near HBV integration sites effectively detected by a low pass sequencing strategy in cell-free WGBS data

Considering the dispersive and limited genomic regions represented by cfDNA fragments, particularly in patients with early stage HCC, long-range methylation around HBV integaration sites (cfMethylins, region size 5Kb) was applied to measure the methylation status of cfDNA in the five cfDNA samples at medium sequencing volume (each composed of approximately 58M reads). As expected,cfMethylins was much lower in the HCC patient (49.85%) compared to healthy individual, chronic hepatitis and cirrhosis (72.72%, 71.58% and 71.92%; **Table S1**; **Fig S1A**). To determine the effective small sequencing depth, we randomly sampled 1M to 10M mappable reads from each sequencing dataset and caculated cfMethylins respectively (**Methods**).As predicted, when we used more sequencing reads, cfMethylins was closer to the value calculated using total sequencing reads. The correlation coefficient between the methylation level from low-pass WGBS and total WGBS data saturates when using 5M or more reads (**Fig 2A, Fig S4**). The correlation coefficient at permulated regions between 5M resampling reads and all sequencing reads was above 0.77 (Pearson’s correlation coefficient, **Fig 2, Fig S4**), and methylation level remained consistent after resampling 100-times (CV is 3.8%, 4.5%, 2.4%, 3.0%, 5.1% for D1, D2, D3, D4 and D5, respectively, **Fig S4**). In summary, we demonstrate that 5M mappable reads without redundancy in low-pass WGBS is a reliable approach to evaluate the methylation level of cfDNA samples in the long-range mode.

We next sought to evaluate the ability of low-pass WGBS of cfDNA to discriminate the patients with different liver diseases. We conducted low-pass WGBS to the circulating cfDNA which are from 54 individuals, including 17 HCC (3 early stage HCC, 5 advanced HCC and 9 HCC patients after surgery; 16 were HBsAg positive and 1 was anti-HBs positive) , 17 with cirrhosis (14 from HBV, 1 from NASH, 1 from alcohol and 1 cryptogenic cirrhosis), 17 with hepatitis B and 3 healthy volunteers (**Table S4**). On average, 10.2M mappable reads were obtained (IQR=6.3M, **Table S5**). The cfDNA fragment size (cfDNAsize)in HCC samples were significantly shorter than non-HCC samples (P=0.003, Wilcoxon rank sum test), consistent with recent observation(19). Particularly, cfDNAsize in advanced HCC group were much shorter than those in healthy individuals (P < 2.2x10-16, Wilcoxon rank sum test; **Fig 3A)**, and the size seemed to decrease along with liver disease progression (**Fig 3A)**. As expected, the distribution of CpGs capatured by low-pass WGBS also tended to be located at intergenic and repeat regions. Moreover, higher in low-pass WGBS had much higher enrichment score of regions around reported HBV integration sites than high-depth WGBS datasets (**Fig 3B, Fig 1A**). To figure out the enrichment at repeat regions is a feature of cfDNA or artifacts of WGBS, we randonmly extracted 10M single reads from published high-depth cfDNA WGBS datasets (25), including 58 cfDNA samples and 30 tissue sampels as well as analyzed region enrichment score. Overrepresntation of regions around reported HBV integration sites was also observed in these datasets (**Fig 3C, Fig S5**). Strikingly, compared to tumor tissue and buffy coat, cfDNA samples were less enriched functional elements (CpG island, promoter and exon) and more enriched in intergenic, repeat regions and HBV integration regions in both randomly 10M reads and high-depth data. (randomly 10M reads in **Fig 3C**; high-depth reads in **Fig S5C**), suggesting this enrichment is a feature of cfDNA and the bias of cfDNA release.

Using our low-pass WGBS datasetes, we explored whether DNA methylation in HBV integration regions could mirror the hypo-methylation profiles of cfDNA from HCC patients and the potential for early HCC detection. Accodrding to cfMethylins , the advanced HCC patients showed significantly hypo-methylation level compared to healthy individuals (<66.1%; P = 0.03, Wilcoxon rank sum test; **Fig 4A**; **Table S4)**. However, for early stage HCC patients, this methylation level was relatively higher, ranging from 68.5% to 72.3%. As expected, after surgery, most HCC patients (8/9) demonstrated similar cfDNA methylation levels to healthy individuals and patients with chronic hepatitis or cirrhosis. Nevertheless, one (P45) out of the nine HCC patients exhibited a lower methylation after surgery (68.49%, **Fig S5A**; **Table S4**), and died two months later due to tumor recurrence, suggesting that there were micro-metastasis with tumor cells in that individual. Additionally, a negative correlation was observed between cfMethylins and alpha-fetoprotein (AFP) levels (Pearson’s correlation coefficient = -0.59, P = 5.9 x10-6; **Fig 4A-B**). Besides, cfMethylins seemed to have no difference among healthy individuals, patients with chronic hepatitis and cirrhosis((P >0.1, Wilcoxon rank sum test; **Fig 4A**). We also included one patient with acute hepatitis B in the hepatitis group and found that MethylHBVLR from this patient was similar to patients with chronic hepatitis (**Table S4**; **Fig 4A**).

### DNA methylation around HBV integration regions enhance HCC prediction

We evaluated cfMethylins by their differentiation ability to HCC from non-HCC saples using Receiver operating characteristic (ROC) curves based on a logistic regression model. MethylHBVLR showed the distinguish ability of HCC from non-HCC with AUC=0.85. To solve the minor release of cfDNA and the lower sensitity in early stage HCC, we screened for candidate hypo-methylated regions around HBV integration sites. These regions should be long enough to be constantly detected at low-pass WGBS and could be suitable as markers of early stage HCC. For 6072 reported HBV integration sites, we first obtained 3083 one with sequening reads more than 10 within 5K flanking regions in all the 54 samples, which were stable detected at low-pass sequencing. Then we merged two neighbor sites with their flanking regions into one larger region if their distance was less than 1Mb. At last, We obtained 144 candidate merged regions larger than 1Mb. They were used for further random forest based feature selection in healthy individuls and early stage HCC patients. Top 5 regions were identified in distinguishing patients from healthy individuals (chr13: 19442162-20713822; chr1: 10121993-12279387; chr10: 11149668-13266296; chr10: 38027603-39151628; chr10: 84035111-85772043) **(Methods)**. All our cfDNA samples had these regions well sequenced, with the minimum amount of sequencing reads at 1,991 (**Table S5**). Their methylation levels were significantly lower in either early stage or advanced HCC patients than in healthy individuals, and demonstrated obvious decreasing tendency along with disease progression (**Fig 4C; Table S6**). We found the prediction model using region 1, 2 and 5 provided improved prediction for HCC patients (AUC>0.85; **Fig 4D**). They all exhibited improved discrimination performance compared to clinical variables (ALT, AST, Tbil, AFP) (Figure S6B).

To validate our finding, we applied this method in an independent cohort in a previous study (25). This cfDNA cohort was comprised of 32 healthy individuals and 26 HCC patients with early stage (BCLC stage is A or B) based on single-end bisulfite sequencing. To achieve similar sequencing depth, we randomly sampled 10M reads from each plasma sample. All the 5 regions showed significantly decreasing methylation in early stage HCC compared to healthy individuals (Wilcoxon rank sum test; **Table S7**; **Fig 4E**). Particlularly, our approach demonstrated better performance of region5 in HCC detection than genome-wide hypomethylation analysis in the paper (AUC=0.93; **Fig 4F**). With all the 58 cfDNA samples, region5 was still the best solo HCC indicator (AUC=0.918, **Fig S6C**). Moreover, the combination of multiple regions provided improved prediction for HCC. When combing region1, region2, region3 region5 all together in the predicition model, it provided the best prediction performance with AUC=0.954 (**Fig 4G**).

## Discussion

Patients with chronic liver disease are at risk of HCC development, highest among those with cirrhosis. Professional societies recommend HCC surveillance in those patients at high risk who will benefit from early diagnosis so they might receive curative therapies. The recommended strategy for surveillance includes abdominal ultrasound with or without alpha-fetoprotein (AFP) every six months. However, image examination required special equipment (the ultrasound machine) and trained personnel to perform and interpret the study, potential barriers especially considering the large population of patients with HBV infection in China. Ultrasound is also operator dependent. Therefore, there is an unmet clinical need for new non-invasive diagnostic tests that is not operator dependent, such as liquid biopsy using circulating tumor cells (43). Unfortunately, The European Association for the Study of the Liver did not recommend the use of any existing tumor markers such as AFP and L3 fraction for HCC surveillance due to their suboptimal performance for early detection, and in the prior version of the American Association for the Liver Diseases, AFP was felt to lack both sensitivity or specificity for early detection of HCC. Subjects at highest risk for HCC are those with chronic hepatitis and advanced fibrosis; hepatic inflammation can result in elevation of AFP and up to 30% of HCC was non-AFP producing. Current study found a strong negative correlation between MethylHBVLR and AFP levels. However, unlike AFP, the MethylHBVLR level was not affected by the presence of inflammation, hence making it a more specific tumor marker. Currently new blood-based measurements are commonly compared with AFP, which had already been shown to have inadequate sensitivity and specificity, hence we believe future comparison should be between new biomarkers and ultrasound for early detection of HCC. Although WGBS of cfDNA has been shown effective for cancer detection (28), the cost of cfDNA WGBS in cancer patients is one of challenges for wide application. In this paper, 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 DNA methylation around HBV integration regions could be applied in low-pass cell free WGBS at 5-million reads to reflect liver disease status of chronic hepatitis, cirrhosis and HCC. Moreover, DNA hypomethylation in HBV integration regions was shown promising results as a potential biomarker for early HCC detection.

Previous studies have been shown that the fragmentation process of cfDNA is not random (44, 45). Our results show low-pass WGBS for cfDNA tended to capture fragments from repeat regions and HBV integration sites. Due to open chromation regions are easily degradated, fragments from open chromation regions (promoter and gene coding regions) were less likely to be detected in cfDNA. When decreasing the sequencing volume, overrepresentation of genomic repeat regions and HBV integration regions were observed in cfDNA. This suggests that the signal from these regions could remain given adequate sequencing depth in low-pass WGBS. Since HBV integrations tend to localize at repeat regions, DMCs of advanced HCC patient were also enriched in previously reported HBV integration sites.

We adopted an approach based on regions around HBV integration sites as surrogate regions for plasma hypomethylation analysis in HCC patients. We screened for hypomethylated HBV regions with length larger than 1M for early stage HCC detection. 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 three patients without HBV infection (P1, P18 and P19; **Table S2**). While HBV integrations carried by dominant tumor clones are likely to have some specific DNA molecular features (26, 46-48), 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, one chronic hepatitis patient, P14, had the MethylHBV5k at 69.5%, the methylation level of region5 at 72.4% and an abnormal AFP level (141.9 ng/ml). The corresponding P14 blood sample was initially labeled as chronic hepatitis since he was a follow-up patient with chronic HBV infection; however, he was diagnosed with HCC in this examination and died 8 months later. Therefore, it is plausible that the patient had significant circulating tumor cells at the time of sample collection since his AFP was also significantly elevated. Except P14 (chronic hepatitis), the sample from a chronic hepatitis patient, P2, showed that the methylation level of region5 was 70.7% and the MethylHBV5k was 68.5%. Using the sample from a clinical visit 6 months following the initial sample collection, the methylation level of region5 increased to 73.92%, whereas the MethylHBV5k increased to 71.34%. This patient had no detected HCC in follow-up. 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. We compared our strategy with genome-wide hypomethylation analysis in a public dataset, and our strategy had improved classification performance compared to the genome-wide hypomethylation analysis used in the original publication (25). Moreover, the calculation of methylation in theses regionsdoes not rely on a reference panel of healthy individuals, and is thus independent of either sequencing quality and inclusion criteria of the reference panel.

Target sequencing have already achieved certain progress in tumor detection, but genome wide characterization of methylation profiles is the promising direction to overcome the false negative errors due to tumor heterogeneity and optimize the genomic regions used for surrogating the methlyation level changes specific to tumor patients, such as previously reported HBV integration sites in our observation. We believe low-pass WGBS will faciliate efforts using large sample size for novel solutions and finally improve the clinical implementation of methylation evaluation. Although we have found some stable methylation patterns using low-pass WGBS, these findings still need to be validated in larger studies. The low-coverage caused by the low-pass WGBS sequencing introduced analysis challenges, however, it may still have clinical utility in augmenting early detection of HCC. This study can serve as a platform to motivate further development of low-pass DNA methylation approaches to improve the accuracy of HCC diagnoses and surveillance. Subsequent larger studies will aid in 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.

In summary, we demonstrate that DNA methylation around HBV integration regions could serve as HCC detection biomarkers. We also demonstrated DNA methylation around HBV integration regions reflects genome-wide demethylation changes from non-tumoral tissues to HCC and could be used as a low-cost approach detect minimal tumoral residual disease after surgical resection. In summary, our study provided a novel low-cost HCC cancer diagnosis strategy in which HBV integration and multiple hypomethylated long range regions were employed.

## 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 of chronic hepatitis B was made according to the guidelines for the prevention and treatment of chronic hepatitis B: a 2015 update (49). 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 were classified as early and late stage according to the Barcelona Clinic Liver Cancer staging system, considering A as early stage, C and D as late stage. 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 (IRB number 2016H005). An informed written 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 shipped to Beijing Institute of Genomics. Upon arrival, the blood was 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 (50) in BIG Data Center (51), 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>.

After base calling, all paired-end fastq files were trimmed using cutadapt (v 1.8.3) (52) 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) (53) 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 (v0.1.19) 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 m/(m + u), where m was defined as the number of methylated cytosines and u was defined as the number of unmethylated cytosines. The number of methylated and unmethylated cytosines of 1 Mb regions were generated using R package methylKit. The average methylation level of each long range region was calculated as the total number of cytosines divided by the number of methylated cytosines.

### cfDNA fragment size determination and distribution

Unique reads with well alignments to human genome (hg19) were applied for cfDNA fragment size evaluation. The end positions and start positions were extracted to calculate the cfDNA size and the distribution were prepared for different samples.

### The enrichment score in each genomic region

CpGs with depth over 5 were used for calculation of enrichment score. The enrichment score is defined as follows: is the number of DMC sites in the genomic element; where the expected value , is the number of DMC sites in the genome, is the number of CpG sites in the genomic element, and is the total number of CpG sites in the genome.

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

The identification of DMCs was generated using the R package methylKit (54). The significance of the DMCs departure between two groups was calculated using a logistic regression test with at least 5-fold coverage. P-value was adjusted for multiple testing with the method of Benjamini and Hochberg (55). The CpG sites were considered different between cases and controls 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 Genome Browser (56). The HBV integration sites were extracted from previous reports (34, 37-42).

### Calculation of average methylation level within the 5kb of HBV integaration sites (MethylHBV5k)

Average methylation level of the CpGs within the 5kb of the HBV integration sites (MethylHBV5k) was determined. For each sample, all the CpGs with depth over 1 read were extracted. The average methylation level within the 5 kbp upstream or downstream of HBV integration sites (MethylHBV5k) was included in all the CpGs with depth over 1 read. This value was calculated as the number of the total number of methylated cytosines divided by the number of total cytosines within the 5kb of the HBV integration sites.

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

Regions within 5kb of reported HBV integration sites were applied to measure the methylation status. Overlapping regions were merged to form a single region. A random sampling method was used to obtain low depth WGBS for 5 pilot WGBS of cell-free DNA. 1M to 10M read pairs (increasing by 1M step) was randomly extracted from each WGBS data set. In each iteration, we randomly permuted genomic regions of 5Kb around the reported HBV integration sites using BEDTools shuffle (57). The average methylation level of permuted regions of this randomly sampled low-pass reads and the average methylation level of permuted regions of total sequencing reads were calculated. The permutation was repeated 100 times and a correlation coefficient was adopted to measure the consistency between low-pass resampling reads and those based on total sequencing reads. For each sequencing depth, we repeated the random extraction 10 times to examine the variation of the correlation coefficient, and the difference (coefficient of variation, CV) among 10 values of the correlation coefficient was used to assess dispersion in the sampling process.

### Feature selection based on HBV integration regions

Regions within 5kb of HBV integration sites with depth over 10 reads were selected and the adjacent regions less than 1Mb were extended to a single region. Merged regions with length larger than 1Mb were used for the feature selection procedure. Feature selection was conducted using the R package caret based on a random forest algorithm.

### Prediction analysis and ROC curves

The AUCs measure the discrimination between HCC and non-HCC samples. AUC values calculated in our dataset were averaged AUC calculated across the the five-fold cross validation runs on the overall test dataset. The procedure is that the data including selected hypo-methylated regions, MethylHBV5k, AFP, cfDNAsize were divided into 5 equal parts and each of them was set as the test dataset while the remaining as the training dataset. In the training stage, a logistic regression-based prediction model was used. Analysis of receiver operating characteristics (ROC) curves was constructed using R package PredictABEL.

**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. The authors greatly acknowledge Dr Yuk Ming Dennis Lo and his team in the Chinese University of Hong Kong for his cfDNA data.

**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. SJS provided analysis advice, aided in coordinating and supervised all the scientific activities, reviewed and edited the manuscript. 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.

**Data Availability**

The raw sequence data were deposited in the Genome Sequence Archive in BIG Data Center, 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>.

**Code Availability**

All the related software and script were used in the manuscript are available through GitHub at <https://github.com/Shicheng-Guo/low-pass-WGBS/blob/master/readme.md>

**Abbreviations**

HCC Hepatocellular Carcinoma

cfDNA Circulating cell-free DNA

cfDNAsize Fragment size of circulating cell-free DNA

MethyHBV5k Methylation level in HBV intergration regions

DMCs Differential Methylation CpGs

DMGs Differential Methylation Genes

HBV Hepatitis B virus

WGBS Whole Genome Bisulfite Sequencing

## Reference

1. 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.

2. Villanueva A. Hepatocellular Carcinoma. N Engl J Med 2019;380:1450-1462.

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

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

5. 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.

6. 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.

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

8. 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.

9. Tzartzeva K, Obi J, Rich NE, Parikh ND, Marrero JA, Yopp A, Waljee AK, et al. Surveillance Imaging and Alpha Fetoprotein for Early Detection of Hepatocellular Carcinoma in Patients With Cirrhosis: A Meta-analysis. Gastroenterology 2018;154:1706-1718 e1701.

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

11. 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.

12. 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.

13. Waldron D. Cancer genomics: A nucleosome footprint reveals the source of cfDNA. Nat Rev Genet 2016;17:125.

14. Bahcall OG. Genetic testing: cfDNA screening for trisomy 21 tested in unselected pregnancies. Nat Rev Genet 2015;16:316-317.

15. Corcoran RB, Chabner BA. Cell-free DNA Analysis in Cancer. N Engl J Med 2019;380:501-502.

16. Fiala C, Diamandis EP. Cell-free DNA Analysis in Cancer. N Engl J Med 2019;380:501.

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

18. 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.

19. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, Jensen SO, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 2019;570:385-389.

20. 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.

21. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, Samet Y, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun 2018;9:5068.

22. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, Jensen SO, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 2019.

23. 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.

24. Li W, Zhang X, Lu X, You L, Song Y, Luo Z, Zhang J, et al. 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. Cell Res 2017;27:1243-1257.

25. 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.

26. 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.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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.

32. Fece de la Cruz F, Corcoran RB. Methylation in cell-free DNA for early cancer detection. Ann Oncol 2018;29:1351-1353.

33. 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.

34. 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.

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

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

46. Zhao ZH, Fan YC, Zhao Q, Dou CY, Ji XF, Zhao J, Gao S, et al. Promoter methylation status and expression of PPAR-gamma gene are associated with prognosis of acute-on-chronic hepatitis B liver failure. Clin Epigenetics 2015;7:115.

47. Kuramoto J, Arai E, Tian Y, Funahashi N, Hiramoto M, Nammo T, Nozaki Y, et al. Genome-wide DNA methylation analysis during non-alcoholic steatohepatitis-related multistage hepatocarcinogenesis: comparison with hepatitis virus-related carcinogenesis. Carcinogenesis 2017;38:261-270.

48. Zhang X, Hu Y, Justice AC, Li B, Wang Z, Zhao H, Krystal JH, et al. DNA methylation signatures of illicit drug injection and hepatitis C are associated with HIV frailty. Nat Commun 2017;8:2243.

49. 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.

50. 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.

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

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

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

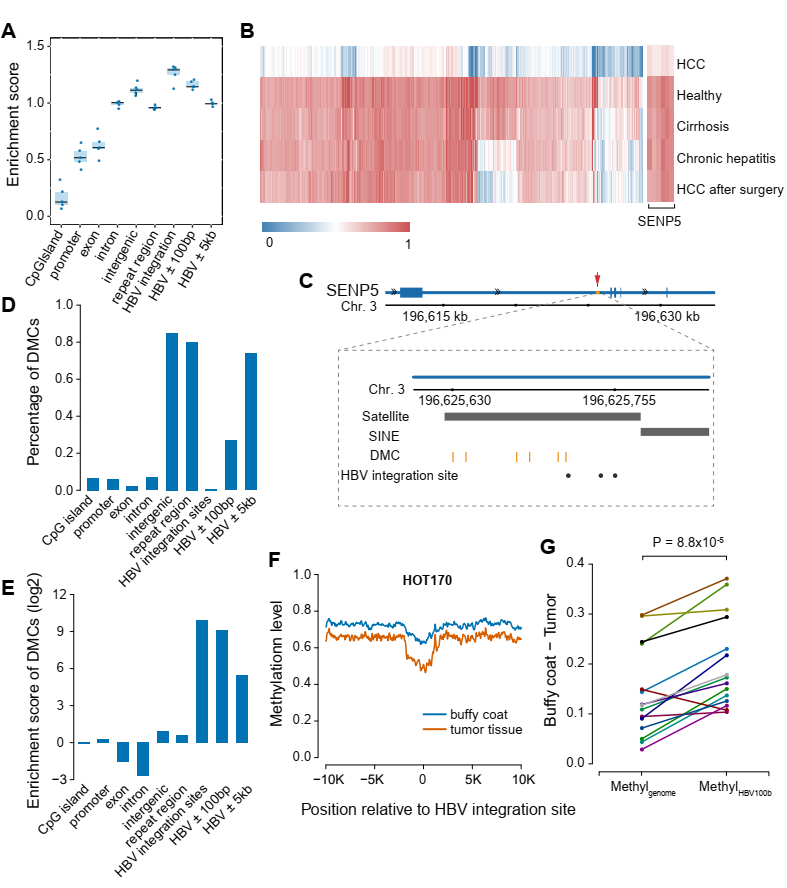
54. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 2012;13:R87.

55. Green GH, Diggle PJ. On the operational characteristics of the Benjamini and Hochberg False Discovery Rate procedure. Stat Appl Genet Mol Biol 2007;6:Article27.

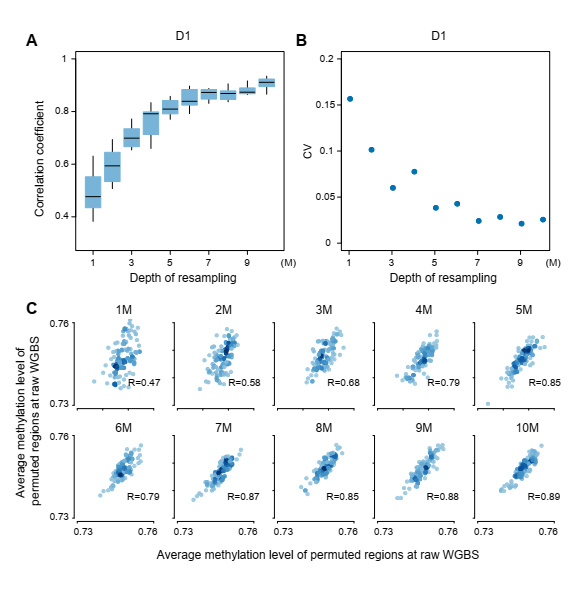
56. Hung JH, Weng Z. Visualizing Genomic Annotations with the UCSC Genome Browser. Cold Spring Harb Protoc 2016;2016.

57. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841-842.

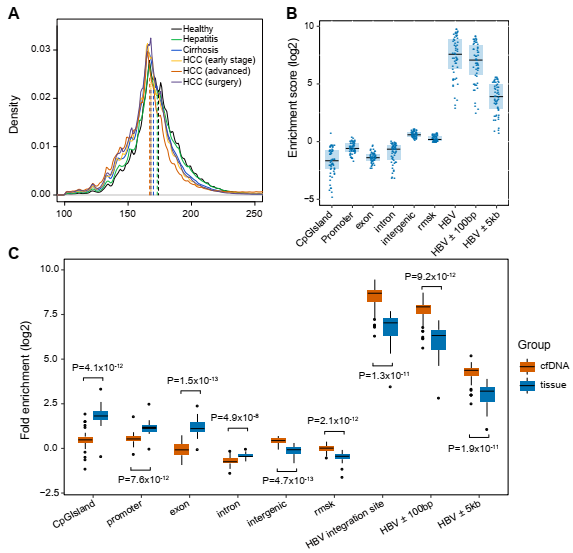
## Figure Legends



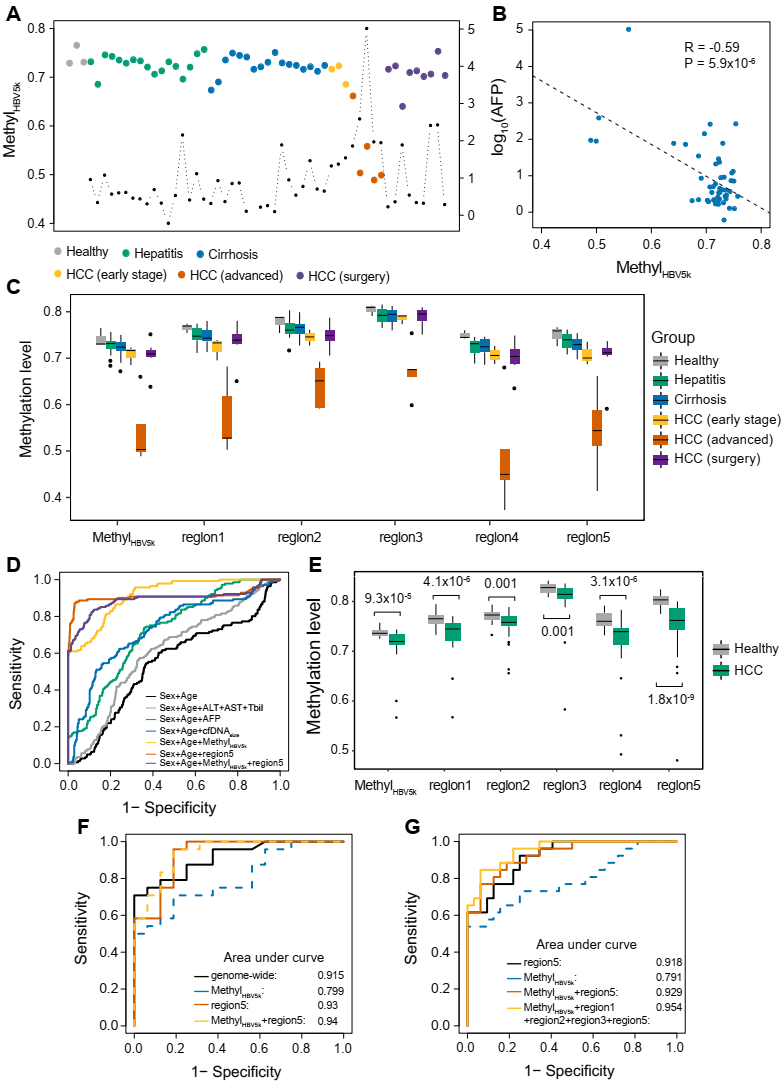
**Fig. 1. DNA methylation around HBV integration regions.** (A) The enrichment score of CpGs in different genomic elements in 5 pilot WGBS. (B) The heatmap displays the methylation level of DMCs between HCC patient and healthy individual in all the 5 individuals. (C) The locus of 6 DMCs and 3 reported HBV integration sites in intron 2 of SENP5. The black dots represent the HBV integration sites and the orange vertical lines represent the 6 DMCs. The black bar labels represent the locus of repeat marker in this region. (D) The percentage of DMCs located at different genomic elements and regions surrounding HBV integration sites. (E) The enrichment scores of DMCs at different genomic elements. (F) The average DNA methylation level profiles along 10 kb upstream and downstream of the HBV integration sites in buffy coat and tumor tissue of patient HOT170. (G) The difference of average methylation level across the genome (Methylgenome) and average methylation level of the CpGs near the HBV integration sites (MethylHBV) between Buffy coat and tumor tissue.



**Fig. 2. The efficiency of re-sampling sequencing reads for low pass WGBS in healthy individual.** (A) The boxplot shows the correlation coefficient between re-sampling low-pass WGBS and total sequencing reads for 10 times from 1M to 10M . (B) The coefficient of variation (CV) for 10 correlation coefficient between re-sampling low-pass WGBS and total sequencing reads from 1M to 10M. (C) The correlation of average methylation level of permutated regions between different re-sampling reads and total sequencing reads at one re-sampling.



**Fig. 3. Landscape of plasma cfDNA in healthy individuals, hepatitis, cirrhosis and HCC patients.** (A) The distribution of cfDNA fragment size in the group of healthy, hepatitis, cirrhosis, early stage HCC advanced HCC and HCC after surgery. The vertical dashed lines indicate the median values in all distributions. (B) The enrichment scores of CpGs at different genomic elements and regions surrounding HBV integration sites of all the 54 cfDNA samples at low-pass WGBS. (C) The enrichment scores of CpGs at different genomic elements of cfDNA and tissue samples by randomly re-sampling 10M reads from published dataset.



**Fig. 4. Hypomethylation around HBV integration regions enhance HCC prediction.** (A) The average methylation level of CpGs located within 5 kb of the HBV integration sites (MethylHBV5k) in all the samples. The black dot represents for AFP level for the corresponding individual. (B) The correlation between AFP (log10) and average methylation level of the CpGs within the 5 kb of the reported HBV integration sites (MethylHBV5k). (C) Boxplot displays MethylHBV5k and the methylation level of top 5 selected regions in all the samples. (D) Receiver operating characteristics (ROC) curve based on five-fold cross-validation for HCC patientd detection by different indicators in discriminating HCC patients from individuals without HCC (healthy individuals, patients with hepatitis and cirrhosis). (E) The comparisons between healthy individuals and patients with early stage HCC using MethylHBV5k and the methylation level of top 5 selected regions in the validation dataset. (F) ROC curves for 16 healthy individuals and 24 HCC patients using genome-wide hypomethylation analysis, MethylHBV5k and region5. (G) ROC curves for HCC patient detection using all the healthy individuals and HCC patients.