Science Advances
AAAS

Manuscript Template

FRONT MATTER

Title

- Circulating cell-free DNA based low-pass genome-wide bisulfite sequencing aids non-invasive surveillance to Hepatocellular carcinoma.
- Low-pass WGBS of cfDNA aids early detection to HCC.

Authors

Haikun Zhang[1,8,#], Peiling Dong[2,#], Shicheng Guo[3,#], Chengcheng Tao[1], Wenmin Zhao[2], Jiakang Wang[4], Ramsey Cheung[5], Augusto Villanueva[6], Huiguo Ding[2], Steven J. Schrodi[3,7,*], Dake Zhang[1,*], Changqing Zeng[1,*]

Affiliations

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

[2]Department of Hepatology, Beijing You'an Hospital Affiliated with Capital Medical University, Beijing 100069, China.

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

[4]Biology Department, Stonybrook University, Stonybrook, NY, USA.

[5] Department of Gastroenterology and Hepatology, VA Palo Alto Health Care System and Stanford University, Palo Alto, CA, USA.

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

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

[8]University of Chinese Academy of Sciences, Beijing 100049, China.

* Corresponding author. Email: czeng@big.ac.cn (C.Z.); dakezhang@gmail.com (D.Z.); schrodi.steven@mcrf.mfldclin.edu (S.J.S.)

# These authors contributed equally to this work

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 ($Methyl_{LRM}$) in low-pass WGBS data (~5 million reads) generated from cfDNA to detect methylation changes in patients with hepatocellular carcinoma (HCC). We found a significant enrichment of differential methylation loci in intergenic and repeat regions, especially in HBV integration sites. Moreover, methylation profiles nearby HBV integration sites ($Methyl_{HBV}$) were found to enhance the prediction performance. Multiple machine learning models based on $Methyl_{LRM}$, $Methyl_{HBV}$, cfDNA fragment size demonstrated low-pass cfDNA methylation data provided powerful discriminating ability, and could be used as a low-cost approach for early HCC detection in the context of surveillance programs.

MAIN TEXT

Introduction

Circulating cell-free DNA (cfDNA) are small double-stranded DNA fragments found in plasma, urine, and other body fluids originating from cell apoptosis and necrosis (*1*). 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 (*2*), early cancer detection

and prognosis prediction (*3, 4*). Apoptotic and necrotic tumor cells can release cfDNA into the peripheral blood, which reflects tumor-related genetic features, including cfDNA fragment size ($cfDNA_{size}$) (*5*), mutations, copy number aberrations and epigenetic changes(*3*). Meanwhile, cfDNA also carries tissues-specific information which provides promising abilities for tissue-of-origin mapping (*6, 7*). 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 (*6*) and cfMeDIPseq. However, genome-wide methylation assays require large amounts of input DNA—conventional WGBS requires microgram input and reduced representation bisulfite sequencing (RRBS) requires 30ng of DNA input which is often approaching the maximum level of the cfDNA detected (or detectable) in a human blood sample.

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 virus (HBV), hepatitis C virus (HCV), alcohol abuse or non-alcoholic fatty liver disease (*9*). 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 (*10, 11*). 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 (*12*). 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 (*13*), circulating tumor cells (CTCs) and DNA methylation (*15-19*). 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 mixture of cell types. Multiple studies have focused on the use of cfDNA methylation in cancer diagnosis in the areas of specific biomarkers (*16, 19*), hypo-methylation (*15*) and tissue of origin (*17, 18, 20*). Single cytosine measurement and high accuracy have enabled whole genome bisulfite sequencing (WGBS) to become the gold standard in DNA methylation analysis (*21*). One of the limitations of using WGBS for DNA methylation analyses on cfDNA is the need for deep sequencing (*17, 18*) 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 (*22*). 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 cancer prediction. We evaluated the minimum sequencing depth for long-range methylation measurements and provided the landscapes of low-pass WGBS in healthy individuals, cirrhosis, hepatitis and HCC patients. Finally, we proposed long-region DNA methylation within 2-Mbp ($Methyl_{LRM}$), methylation around HBV-integration region ($Methyl_{HBV}$) and cfDNA fragment size ($cfDNA_{size}$) to predict HCC from non-HCC samples.

Results

Efficacy of a 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 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). The average DNA methylation across the genome was much lower in the HCC patient (D4; 53.56%) compared to healthy individual, cirrhosis and chronic hepatitis (74.76%, 75.64% and 75.13%; Table S1). Long range methylation ($Methyl_{LRM}$) was applied to measure the methylation status of cfDNA in these samples. To identify the optimal region size of $Methyl_{LRM}$, we divided the HCC genome (D4) into 500-Kb, 1-Mb, 1.5-Mb, 2-Mb and 2.5-Mb bins. For each region size, the average methylation level was calculated within each window for the genome. Then, the percentage of regions with hypo-methylation was determined in the HCC sample (D4). With the 2-Mb as the window size, the percentage of hypo-methylated regions in D4 was found to have the maximum ratio (Figure S1A; Table S1). Finally, the $Methyl_{LRM}$ for all 1,382 autosomal 2-Mb regions (LRM_{2M}) 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 ($Methyl_{LRM}$). In each iteration, we calculated $Methyl_{LRM}$ for all 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, $Methyl_{LRM}$ 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 $Methyl_{LRM}$ at 5M reads and all sequencing reads was above 0.92 (Pearson's correlation coefficient, $P < 2.2x10^{-16}$, Figure S1B-C), 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.

Landscape of plasma cfDNA in HCC, chronic hepatitis, cirrhosis patients and healthy individuals

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 S2). On average, 10.2M mappable reads were obtained (IQR=6.3M, Table S3). We estimated the

cfDNA fragment size ($cfDNA_{size}$) cross all the samples and found the median of $cfDNA_{size}$ in HCC samples are significantly shorter than non-HCC samples (P=0.009, logistic regression), consistent with recent observation (*5*).The $cfDNA_{size}$ in different groups showed that the order from short to long was patients with advanced HCC, patients with cirrhosis, early stage HCC and HCC after surgery, and patients with hepatitis and healthy control (Figure 2A**)**. Particularly, $cfDNA_{size}$ in advanced HCC group were much shorter than those in healthy individuals ($P < 2.2x10^{-16}$, Wilcoxon rank sum test; Figure 2A**)**. We apply principal components analysis (PCA) based on 2-Mb regions to investigate the data structure of low-pass WGBS data across all of the samples. We found a obvious separation between advanced HCC and the remaining samples (Figure S2).

To evaluate the methylation levels in these samples, we applied the $Methyl_{LRM}$ strategy to define the hyper- or hypo-methylated regions, using $Methyl_{LRM}$ in healthy individuals as the baseline level. The percentage of hyper- or hypo-methylated regions is shown for each patient (Fig 2B-C; Table S3). In chronic hepatitis and cirrhosis patients, we found that hyper-long methylated regions (hyper-LRM) accounted for <3% of total 1382 autosomal LRMs (Fig 2B), while hypo-long methylated regions (hypo-LRM) accounted for 0.0-20.04% of the total LRMs, with only three patients exceeding 10% (Fig 2C; Table S3). This data suggest that patients with chronic hepatitis and cirrhosis have similar cfDNA methylation levels with healthy individuals. We also included one patient with acute hepatitis B in the hepatitis group and found that the profile from this patient was similar to patients with chronic hepatitis (percentage of hypo-LRMs is 0.36%; Table S3). Further, in early stage HCC patients, no hyper-LRM were identified, however hypo-LRMs accounted for 1.2% to 26.2% of the total LRMs. In advanced HCC patients, no hyper-LRM were identified, and hypo-LRM accounted for more than 65.7% of the total LRMs (Fig 2B-C; Table S3). 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 higher proportion of hypo-LRMs after surgery (69.9%, Fig 2C; Table S3), and died two months later due to tumor recurrence, suggesting that there were micro-metastasis with tumor cells in that individual. We find a significant positive relationship between AFP and the percentage of hypo-LRMs ($R=0.6$, $P=3.9\times10^{-6}$, Pearson's correlation coefficient). We also evaluated the diagnostic potential of low-pass WGBS data to distinguish HCC from non-HCC samples and we found the percentage of hypo-LRMs showed better diagnosis performance than AFP (AUC= 0.885 vs 0.711).

Differentially methylated CpGs and genes identified by low-pass circulating cell-free WGBS

With traditional analysis pipeline (*23*), we identified differentially methylated CpGs (DMCs) and differentially methylated gene (DMGs) with low-pass cell-free WGBS data. On average, each cfDNA sample had 61,018 CpGs with sequencing depth over 5 reads (Table S3). In total, we identified 1,695 DMCs in advanced HCC patients (Table S4), 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, SENP5, SLCO5A1, REXO1L1P, DLG2*. In the one early stage HCC patients (percentage of hypo-LRMs=26.27%), we identified 249 DMCs (Table S4), of which 207 were in common with those observed in advanced HCC patients and nine were located within *PMF1* and *PMF1-BGLAP*. Relatively high proportions of hypo-LRMs (>10%) were observed in one chronic hepatitis and two cirrhosis patients (Fig 2C), possibly indicating their high HCC risk. In total, all four clinical groups had 165 DMCs in common (Fig 3A), which suggested that these DNA methylation loci might be considered as early biomarkers for HCC diagnosis. Fig 3A displayed the genes with DMCs in four comparisons. Moreover, 31 DMCs were

identified between early stage HCC and cirrhosis patients and 760 DMCs were identified between advanced HCC and early stage HCC patients, with no overlap detected between the two comparisions (Table S4). In particular, *SENP5* gene had seven significantly hypo-methylated DMCs with consistently high sequencing coverage across all individuals (149 reads, on average, Figure S3, 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) (*24*).

DNA methylation around HBV integration and cfDNA fragment size enhance HCC prediction

We found the distribution of CpGs capatured by low-pass WGBS tended to be located at intergenic and repeat regions (Figure S4A). Also, 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; Figure S4B). On average, 64% of all these CpGs were in the repeat regions (Figure S4C), and this percentage varied from 49% to 87% across the samples. Differential methylation analysis required the CpG sites having at least five sequencing reads in all samples, 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 target for HBV integration (*25*), we analyzed the location of DMCs relative to reported HBV integration sites (*24, 26-31*). Among the 1,695 DMCs observed in advanced HCC patients, eighteen completely overlapped with the HBV integration sites, including two in *SENP5* (Table S5). Though *TERT* promoter and *KMT2B* are known HBV integration sites, due to the low-pass approach, no adequate reads were detected in these regions. 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).

With above findings, we are interested in whether DNA methylation in HBV integration regions could mirror the hypo-methylation profiles of cfDNA from HCC patients and then provide extra improvement for methylation prediction model. We collected all the CpGs with read depth exceeding 5 reads 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 (Fig 4C**)**, with 9.6% to 59.1% of CpGs being hypo-CpGs, while the proportion was generally reduced (2.6-10.2%) in early stage HCC patients (Fig 4C; Table S3). Then, the average methylation level of the CpGs within the 100bp of the reported HBV integration sites ($Methyl_{HBV}$) was calculated in each sample **(**Methods**)**. 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; Table S3**)**. 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 $Methyl_{HBV}$ and alpha-fetoprotein (AFP) levels (R = -0.63, P = $8.4x10^{-7}$, Pearson's correlation coefficient; Fig 4D-E).

Finally, multiple machine learning and deep learning models were applied to evaluate the performance of identifying patients with HCC compared to those without HCC with different features which included $Methyl_{LRM}$, $Methyl_{HBV}$ and $cfDNA_{size}$. The percentage of hypo-LRMs based logistic regression model showed the distinguish ability of HCC from non-HCC with AUC=0.885. Moreover, $Methyl_{HBV}$ provided improved prediction for HCC patients (AUC=0.92). We found the prediction model using $Methyl_{HBV}$ and $cfDNA_{size}$ provide the best prediction performance with AUC=0.937 (Figure 4F; Table 1). Applying

five-fold cross-validation and using 100 random resampling iterations on the RF model, the average sensitivity, specificity and accuracy were found to be 62.5%, 97.6% and 91.1%, respectively. These data were also subjected to a neural network classifier using the top 10 features selected by the RF approach using the training set. The neural network prediction in the test set attained an AUC=0.90 (Figure S5).

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 (*32*). 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 $Methyl_{HBV}$ and AFP levels. However, unlike AFP, the $Methyl_{HBV}$ 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 (*20*), 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 long-range methylation 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 reports applying genome-wide hypomethylation in HCC detection and shown low sequencing depth of ~10 million reads was available for the cell-free detection for cancer (*15*). In our study, we required only 5M qualified reads for low-pass WGBS for 54 samples, and there were two samples only having 3.6M reads (Table S3). 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. In our analysis based on limited sample size, all five advanced HCC patients were detected according to this measurement. But for patients with early stage HCC, the sensitivity of our DNA methylation approach in plasma is lower. Specifically, P35 and P36, both the proportion of hypo-LRMs (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 chronic hepatitis patients. Both of these two patients had small tumor sizes (P35, 1.5cm;

P36, less than 2cm, three lesions; Table S2). Multiple indicators, including $Methyl_{HBV}$, cell-free DNA fragment size ($cfDNA_{size}$) and AFP, need to be combined for early stage HCC.

Previous studies have been shown that the fragmentation process of cfDNA is not random (*33*). Our results show low-pass WGBS for cfDNA tended to capture fragments from repeat regions and HBV integration sites. More than 49% of CpGs were 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.

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 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 (*34*, *35*), 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 average methylation level at 67.4%, the proportion of hypo-LRMs at 3.47% and abnormal AFP level (141.9 ng/ml; Table 1). Its blood sample was initially labeled as chronic hepatitis since he was a follow-up patient with chronic HBV infection; however, he was

diagnosed as HCC in this examination and died 8 month later. Therefore, he was likely to has circulating tumor cell at the time since his AFP was significantly elevated. Except P14 (chronic hepatitis), the sample from a chronic hepatitis patient, P2, showed that the proportion of hypo-LRMs 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 LRMs dropped to 1.1%, whereas the average methylation around HBV integration sites slightly increased to 69%. This patient had no detected HCC in follow-up, showing that $Methyl_{HBV}$ is more stable than genome-wide LRM. 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, 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 $Methyl_{LRM}$, $Methyl_{HBV}$ and $cfDNA_{size}$ could serve as HCC detection biomarkers. We also demonstrated LRM 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, DNA methylation and cfDNA fragment size 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 (*36*). 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).

After base calling, all paired-end fastq files were trimmed using cutadapt (v 1.8.3) (*37*) 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) (*38*) 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 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 long range regions were generated using R package methylKit. The average methylation level of each long range region ($Methyl_{LRM}$) 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 segments. For each size, the average methylation level for each region from autosome was 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

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. Logistic regression was applied to test the association between the median of $cfDNA_{size}$ in HCC and nonHCC samples.

Randomly re-sampling lower reads from medium WGBS data

A random 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 resampling, the average methylation level for each autosomal 2-Mb region ($Methyl_{LRM}$) was calculated and a Pearson correlation coefficient was used to show the correlation for all the autosomal $Methyl_{LRM}$ between the

resampled reads and the total WGBS reads. This process was repeated 100 times. (c) For each resampling, a coefficient of variation (CV) for the correlation coefficient was calculated across the 100 random resamples to examine the variability of the 100 extractions.

Identification of hyper-LRMs and hypo-LRMs

We adopted the method of Chan et al. (*15*) to define the hyper- or hypo-$Methyl_{LRM}$ compared to the healthy reference group. Only autosomes were included in this analysis. A 2-Mb region from 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 within the healthy individuals. Lastly, the number and percentage of hyper- or hypo-$Methyl_{LRM}$ within the genome were calculated.

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

The identification of DMCs was generated using the R package methylKit (*39*). 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. 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.

The enrichment score in each genomic region

The enrichment score for CpGs or DMCs was calculated by the following formula:The enrichment score$_{\text{in the genomic element}}$ = $\log_2$ (# DMCs$_{\text{in the genomic element}}$/# expected). # expected was computed as: # DMCs$_{\text{in the genome}}$ × # CpG sites$_{\text{in the genomic element}}$/# total CpG sites$_{\text{in the genome}}$. # means the number of sites.

Identification of hypo-CpGs within the 100bp of HBV integration sites

The HBV integration sites were extracted from previous reports (*24, 26-31*). We extracted CpG within the 100bp upstream or downstream of HBV integration sites. Only autosomal CpGs and CpGs with depth over 5 reads 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. Next, the percentage of hypo-CpGs was calculated.

Calculation of average methylation level within the 100bp of HBV integaration sites (Methyl$_{\text{HBV}}$)

Average methylation level of the CpGs within the 100bp of the HBV integration sites (Methyl$_{\text{HBV}}$) was determined. For each sample, all the CpGs with depth over 1 read were extracted. The average methylation level within the 100 bp upstream or downstream of HBV integration sites (Methyl$_{\text{HBV}}$) 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 100bp of the HBV integration sites.

Prediction analysis, logistic regression, Random Forest and ROC curves

Classification performance for the low-pass WGBS data was calculated using five-fold cross-validation combined wrapped logistic regression (LR), random forest (RF) and neural network (NN). The AUCs measure the discrimination between HCC and non-HCC

samples. All the AUC values calculated in the manuscript were averaged AUC calculated across the the five-fold cross validation runs on the overall test dataset. The detailed procedure is that the data including $Methyl_{LRM}$, $Methyl_{HBV}$, AFP, $cfDNA_{size}$ were divided into 5 equal parts and each of them was set as test dataset while the remaining as the training dataset. In the training stage, logistic regression based prediction model was fitted with feature selection by the Akaike information criterion (AIC) criteria with forward and backward selection. The detailed procedure is that we first starts with the full model and eliminates one predictor at a time, at each step considering whether AIC shows significant decrease by adding back in the variable removed at the previous step. Finally, we make the prediction with the prediction model built in training stage to test dataset and summarize the prediction sensitivity, specificity and accuracy. We also applied five-fold cross-validation based random forest to reduce bias of the prediction. Random Forest approach was conducted with R package randomForest. The neural network algorithm was based on the R package neuralnet. Feature selection was conducted in the training set under 10-fold cross validation with the top features ranked using the MeanDecreaseGini function. Model performance was then evaluated separately in the training and test sets. Analysis of receiver operating characteristics (ROC) curves was constructed using R package PredictABEL.

H2: Supplementary Materials

Fig. S1. Determination of optimal region size and effective sequencing depth of low pass WGBS.

Fig. S2. PCA based on average methylation level of 2-Mb region of all the samples.

Fig. S3. The depth of 7 DMCs of SENP5 in all the samples.

Fig. S4. The genome feature distribution of CpGs at the low-pass WGBS.

Fig. S5. Neural network prediction using the top 10 features selected by RF in training dataset.

Table S1. The statistical information of 5 pilot WGBS samples

Table S2. Clinic information of all the individuals.

Table S3. The statistical information of low pass WGBS.

Table S4. DMCs between different groups.

Table S5. 18 DMCs overlap with HBV integration sites.

References and Notes

1. M. Stroun, P. Maurice, V. Vasioukhin, J. Lyautey, C. Lederrey, F. Lefort, A. Rossier, X. Q. Chen, P. Anker, The origin and mechanism of circulating DNA. *Ann N Y Acad Sci* **906**, 161-168 (2000).
2. D. Waldron, Cancer genomics: A nucleosome footprint reveals the source of cfDNA. *Nat Rev Genet* **17**, 125 (2016).
3. H. Schwarzenbach, D. S. B. Hoon, K. Pantel, Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer* **11**, 426-437 (2011).
4. J. C. M. Wan, C. Massie, J. Garcia-Corbacho, F. Mouliere, J. D. Brenton, C. Caldas, S. Pacey, R. Baird, N. Rosenfeld, Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer* **17**, 223-238 (2017).
5. S. Cristiano, A. Leal, J. Phallen, J. Fiksel, V. Adleff, D. C. Bruhm, S. O. Jensen, J. E. Medina, C. Hruban, J. R. White, D. N. Palsgrove, N. Niknafs, V. Anagnostou, P. Forde, J. Naidoo, K. Marrone, J. Brahmer, B. D. Woodward, H. Husain, K. L. van Rooijen, M. W. Orntoft, A. H. Madsen, C. J. H. van de Velde, M. Verheij, A. Cats, C. J. A. Punt, G. R. Vink, N. C. T. van Grieken, M. Koopman, R. J. A. Fijneman, J. S. Johansen, H. J. Nielsen, G. A. Meijer, C. L. Andersen, R. B. Scharpf, V. E. Velculescu, Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385-389 (2019).

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

7. S. Cristiano, A. Leal, J. Phallen, J. Fiksel, V. Adleff, D. C. Bruhm, S. O. Jensen, J. E. Medina, C. Hruban, J. R. White, D. N. Palsgrove, N. Niknafs, V. Anagnostou, P. Forde, J. Naidoo, K. Marrone, J. Brahmer, B. D. Woodward, H. Husain, K. L. van Rooijen, M. W. Orntoft, A. H. Madsen, C. J. H. van de Velde, M. Verheij, A. Cats, C. J. A. Punt, G. R. Vink, N. C. T. van Grieken, M. Koopman, R. J. A. Fijneman, J. S. Johansen, H. J. Nielsen, G. A. Meijer, C. L. Andersen, R. B. Scharpf, V. E. Velculescu, Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, (2019).

8. A. Villanueva, Hepatocellular Carcinoma. *N Engl J Med* **380**, 1450-1462 (2019).

9. C. J. Chen, M. W. Yu, Y. F. Liaw, Epidemiological characteristics and risk factors of hepatocellular carcinoma. *J Gastroenterol Hepatol* **12**, S294-308 (1997).

10. J. K. Stauffer, A. J. Scarzello, Q. Jiang, R. H. Wiltrout, Chronic inflammation, immune escape, and oncogenesis in the liver: a unique neighborhood for novel intersections. *Hepatology* **56**, 1567-1574 (2012).

11. K. Schutte, J. Bornschein, P. Malfertheiner, Hepatocellular carcinoma--epidemiological trends and risk factors. *Dig Dis* **27**, 80-92 (2009).

12. e. e. e. European Association for the Study of the Liver. Electronic address, L. European Association for the Study of the, EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. *J Hepatol* **69**, 182-236 (2018).

13. C. Qu, Y. Wang, P. Wang, K. Chen, M. Wang, H. Zeng, J. Lu, Q. Song, B. H. Diplas, D. Tan, C. Fan, Q. Guo, Z. Zhu, H. Yin, L. Jiang, X. Chen, H. Zhao, H. He, Y. Wang, G. Li, X. Bi, X. Zhao, T. Chen, H. Tang, C. Lv, D. Wang, W. Chen, J. Zhou, H. Zhao, J. Cai, X. Wang, S. Wang, H. Yan, Y. X. Zeng, W. K. Cavenee, Y. Jiao, Detection of early-stage

hepatocellular carcinoma in asymptomatic HBsAg-seropositive individuals by liquid biopsy. *Proc Natl Acad Sci U S A* **116**, 6308-6312 (2019).

14. I. Bhan, K. Mosesso, L. Goyal, J. Philipp, M. Kalinich, J. W. Franses, M. Choz, R. Oklu, M. Toner, S. Maheswaran, D. A. Haber, A. X. Zhu, R. T. Chung, M. Aryee, D. T. Ting, Detection and Analysis of Circulating Epithelial Cells in Liquid Biopsies From Patients With Liver Disease. *Gastroenterology* **155**, 2016-2018 e2011 (2018).

15. K. C. Chan, P. Jiang, C. W. Chan, K. Sun, J. Wong, E. P. Hui, S. L. Chan, W. C. Chan, D. S. Hui, S. S. Ng, H. L. Chan, C. S. Wong, B. B. Ma, A. T. Chan, P. B. Lai, H. Sun, R. W. Chiu, Y. M. Lo, Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc Natl Acad Sci U S A* **110**, 18761-18768 (2013).

16. Y. Zhao, F. Xue, J. Sun, S. Guo, H. Zhang, B. Qiu, J. Geng, J. Gu, X. Zhou, W. Wang, Z. Zhang, N. Tang, Y. He, J. Yu, Q. Xia, 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* **6**, 30 (2014).

17. K. Sun, P. Jiang, K. C. Chan, J. Wong, Y. K. Cheng, R. H. Liang, W. K. Chan, E. S. Ma, S. L. Chan, S. H. Cheng, R. W. Chan, Y. K. Tong, S. S. Ng, R. S. Wong, D. S. Hui, T. N. Leung, T. Y. Leung, P. B. Lai, R. W. Chiu, Y. M. Lo, Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci U S A* **112**, E5503-5512 (2015).

18. S. Kang, Q. Li, Q. Chen, Y. Zhou, S. Park, G. Lee, B. Grimes, K. Krysan, M. Yu, W. Wang, F. Alber, F. Sun, S. M. Dubinett, W. Li, X. J. Zhou, CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol* **18**, 53 (2017).

19. R. H. Xu, W. Wei, M. Krawczyk, W. Wang, H. Luo, K. Flagg, S. Yi, W. Shi, Q. Quan, K. Li, L. Zheng, H. Zhang, B. A. Caughey, Q. Zhao, J. Hou, R. Zhang, Y. Xu, H. Cai, G. Li, R. Hou, Z. Zhong, D. Lin, X. Fu, J. Zhu, Y. Duan, M. Yu, B. Ying, W. Zhang, J. Wang, E. Zhang, C. Zhang, O. Li, R. Guo, H. Carter, J. K. Zhu, X. Hao, K. Zhang, Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat Mater* **16**, 1155-1161 (2017).

20. R. Lehmann-Werman, D. Neiman, H. Zemmour, J. Moss, J. Magenheim, A. Vaknin-Dembinsky, S. Rubertsson, B. Nellgard, K. Blennow, H. Zetterberg, K. Spalding, M. J. Haller, C. H. Wasserfall, D. A. Schatz, C. J. Greenbaum, C. Dorrell, M. Grompe, A. Zick, A. Hubert, M. Maoz, V. Fendrich, D. K. Bartsch, T. Golan, S. A. Ben Sasson, G. Zamir, A. Razin, H. Cedar, A. M. Shapiro, B. Glaser, R. Shemer, Y. Dor, Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci U S A* **113**, E1826-1834 (2016).

21. H. Li, C. Jing, J. Wu, J. Ni, H. Sha, X. Xu, Y. Du, R. Lou, S. Dong, J. Feng, Circulating tumor DNA detection: A potential tool for colorectal cancer management. *Oncol Lett* **17**, 1409-1416 (2019).

22. S. Liu, S. Huang, F. Chen, L. Zhao, Y. Yuan, S. S. Francis, L. Fang, Z. Li, L. Lin, R. Liu, Y. Zhang, H. Xu, S. Li, Y. Zhou, R. W. Davies, Q. Liu, R. G. Walters, K. Lin, J. Ju, T. Korneliussen, M. A. Yang, Q. Fu, J. Wang, L. Zhou, A. Krogh, H. Zhang, W. Wang, Z. Chen, Z. Cai, Y. Yin, H. Yang, M. Mao, J. Shendure, J. Wang, A. Albrechtsen, X. Jin, R. Nielsen, X. Xu, Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell* **175**, 347-359 e314 (2018).

23. Y. Li, J. Zhu, G. Tian, N. Li, Q. Li, M. Ye, H. Zheng, J. Yu, H. Wu, J. Sun, H. Zhang, Q. Chen, R. Luo, M. Chen, Y. He, X. Jin, Q. Zhang, C. Yu, G. Zhou, J. Sun, Y. Huang, H.

Zheng, H. Cao, X. Zhou, S. Guo, X. Hu, X. Li, K. Kristiansen, L. Bolund, J. Xu, W. Wang, H. Yang, J. Wang, R. Li, S. Beck, J. Wang, X. Zhang, The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol* **8**, e1000533 (2010).

24. W. K. Sung, H. Zheng, S. Li, R. Chen, X. Liu, Y. Li, N. P. Lee, W. H. Lee, P. N. Ariyaratne, C. Tennakoon, F. H. Mulawadi, K. F. Wong, A. M. Liu, R. T. Poon, S. T. Fan, K. L. Chan, Z. Gong, Y. Hu, Z. Lin, G. Wang, Q. Zhang, T. D. Barber, W. C. Chou, A. Aggarwal, K. Hao, W. Zhou, C. Zhang, J. Hardwick, C. Buser, J. Xu, Z. Kan, H. Dai, M. Mao, C. Reinhard, J. Wang, J. M. Luk, Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet* **44**, 765-769 (2012).

25. H. Yan, Y. Yang, L. Zhang, G. Tang, Y. Wang, G. Xue, W. Zhou, S. Sun, Characterization of the genotype and integration patterns of hepatitis B virus in early- and late-onset hepatocellular carcinoma. *Hepatology* **61**, 1821-1831 (2015).

26. S. Jiang, Z. Yang, W. Li, X. Li, Y. Wang, J. Zhang, C. Xu, P. J. Chen, J. Hou, M. A. McCrae, X. Chen, H. Zhuang, F. Lu, Re-evaluation of the carcinogenic significance of hepatitis B virus integration in hepatocarcinogenesis. *PLoS One* **7**, e40363 (2012).

27. A. Fujimoto, Y. Totoki, T. Abe, K. A. Boroevich, F. Hosoda, H. H. Nguyen, M. Aoki, N. Hosono, M. Kubo, F. Miya, Y. Arai, H. Takahashi, T. Shirakihara, M. Nagasaki, T. Shibuya, K. Nakano, K. Watanabe-Makino, H. Tanaka, H. Nakamura, J. Kusuda, H. Ojima, K. Shimada, T. Okusaka, M. Ueno, Y. Shigekawa, Y. Kawakami, K. Arihiro, H. Ohdan, K. Gotoh, O. Ishikawa, S. Ariizumi, M. Yamamoto, T. Yamada, K. Chayama, T. Kosuge, H. Yamaue, N. Kamatani, S. Miyano, H. Nakagama, Y. Nakamura, T. Tsunoda, T. Shibata, H. Nakagawa, Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat Genet* **44**, 760-764 (2012).

28. Z. Jiang, S. Jhunjhunwala, J. Liu, P. M. Haverty, M. I. Kennemer, Y. Guan, W. Lee, P. Carnevali, J. Stinson, S. Johnson, J. Diao, S. Yeung, A. Jubb, W. Ye, T. D. Wu, S. B. Kapadia, F. J. de Sauvage, R. C. Gentleman, H. M. Stern, S. Seshagiri, K. P. Pant, Z. Modrusan, D. G. Ballinger, Z. Zhang, The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res* **22**, 593-601 (2012).

29. D. Ding, X. Lou, D. Hua, W. Yu, L. Li, J. Wang, F. Gao, N. Zhao, G. Ren, L. Li, B. Lin, Recurrent targeted genes of hepatitis B virus in the liver cancer genomes identified by a next-generation sequencing-based approach. *PLoS Genet* **8**, e1003065 (2012).

30. W. Li, X. Zeng, N. P. Lee, X. Liu, S. Chen, B. Guo, S. Yi, X. Zhuang, F. Chen, G. Wang, R. T. Poon, S. T. Fan, M. Mao, Y. Li, S. Li, J. Wang, Jianwang, X. Xu, H. Jiang, X. Zhang, HIVID: an efficient method to detect HBV integration using low coverage sequencing. *Genomics* **102**, 338-344 (2013).

31. S. T. Toh, Y. Jin, L. Liu, J. Wang, F. Babrzadeh, B. Gharizadeh, M. Ronaghi, H. C. Toh, P. K. Chow, A. Y. Chung, L. L. Ooi, C. G. Lee, Deep sequencing of the hepatitis B virus in hepatocellular carcinoma patients reveals enriched integration events, structural alterations and sequence variations. *Carcinogenesis* **34**, 787-798 (2013).

32. R. Palmirotta, D. Lovero, P. Cafforio, C. Felici, F. Mannavola, E. Pelle, D. Quaresmini, M. Tucci, F. Silvestris, Liquid biopsy of cancer: a multimodal diagnostic tool in clinical oncology. *Ther Adv Med Oncol* **10**, 1758835918794630 (2018).

33. P. Jiang, K. Sun, Y. K. Tong, S. H. Cheng, T. H. T. Cheng, M. M. S. Heung, J. Wong, V. W. S. Wong, H. L. Y. Chan, K. C. A. Chan, Y. M. D. Lo, R. W. K. Chiu, Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *Proc Natl Acad Sci U S A*, (2018).

34. J. Kuramoto, E. Arai, Y. Tian, N. Funahashi, M. Hiramoto, T. Nammo, Y. Nozaki, Y. Takahashi, N. Ito, A. Shibuya, H. Ojima, A. Sukeda, Y. Seki, K. Kasama, K. Yasuda, Y.

Kanai, Genome-wide DNA methylation analysis during non-alcoholic steatohepatitis-related multistage hepatocarcinogenesis: comparison with hepatitis virus-related carcinogenesis. *Carcinogenesis* **38**, 261-270 (2017).

35. X. Zhang, Y. Hu, A. C. Justice, B. Li, Z. Wang, H. Zhao, J. H. Krystal, K. Xu, DNA methylation signatures of illicit drug injection and hepatitis C are associated with HIV frailty. *Nature communications* **8**, 2243 (2017).

36. J. Hou, G. Wang, F. Wang, J. Cheng, H. Ren, H. Zhuang, J. Sun, L. Li, J. Li, Q. Meng, J. Zhao, Z. Duan, J. Jia, H. Tang, J. Sheng, J. Peng, F. Lu, Q. Xie, L. Wei, C. M. A. Chinese Society of Hepatology, C. M. A. Chinese Society of Infectious Diseases, Guideline of Prevention and Treatment for Chronic Hepatitis B (2015 Update). *J Clin Transl Hepatol* **5**, 297-318 (2017).

37. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3 (2011).

38. F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).

39. A. Akalin, M. Kormaksson, S. Li, F. E. Garrett-Bakelman, M. E. Figueroa, A. Melnick, C. E. Mason, methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology* **13**, R87 (2012).

Acknowledgments

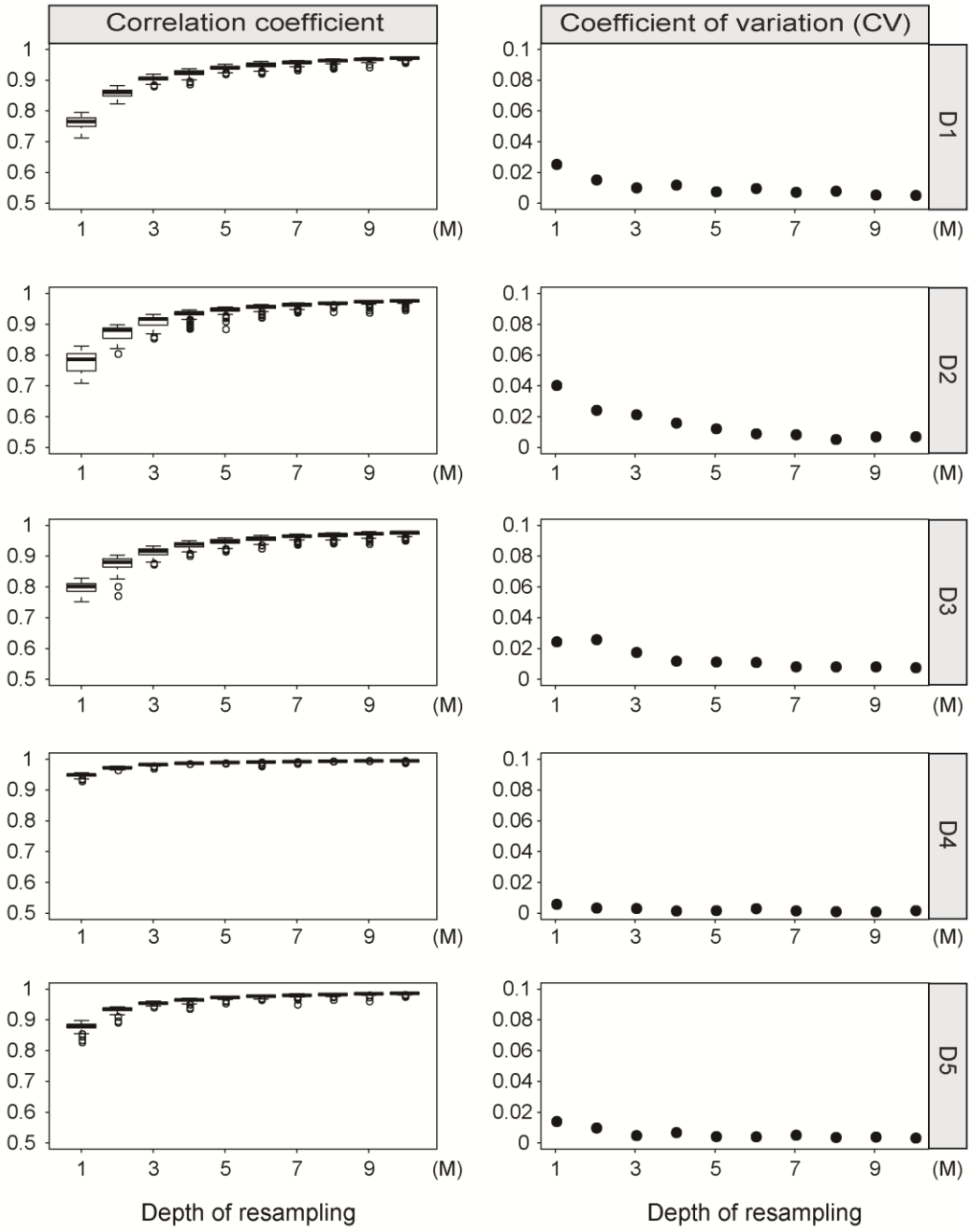
Funding: This study is funded by Innovation Promotion Association CAS (2016098) and National Natural Science Foundation of China (81201700) to D.Z., Major State Basic Research Development Program (2014CB542006), the Key Research Program of the Chinese Academy of Sciences (KJZD-EW-L14) to C.Z., Capital's Funds for Health Improvement and Research (2018-1-1151) to P.D.

Author 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.

Competing interests: The authors declare no conflict of interest.

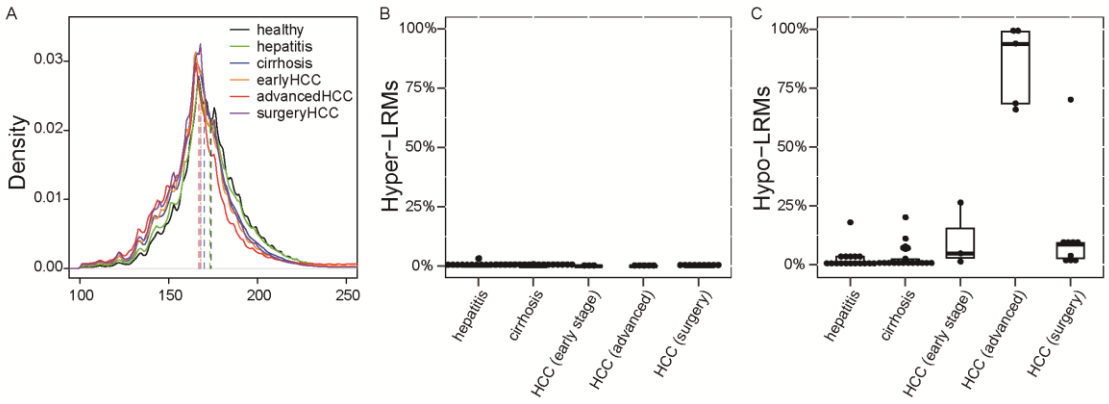
Data and materials availability: The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive in BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession numbers CRA001537, CRA001537 that are publicly accessible at http://bigd.big.ac.cn/gsa. 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

Figures and Tables



679

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



685

686 **Fig. 2. Landscape of plasma cfDNA in healthy individuals, hepatitis, cirrhosis and**

687 **HCC patients.** (A) The distribution of cfDNA fragment size in the group of

688 healthy, hepatitis, cirrhosis, early stage HCC advanced HCC and HCC after

689 surgery. The vertical dashed lines indicate the median values in all distributions.

690 (B) The percentage of hyper-methylated long range regions (2-Mb) in hepatitis,

691 cirrhosis and HCC patients. (C) The percentage of hypo-methylated long range

692 regions in hepatitis, cirrhosis and HCC patients.

693

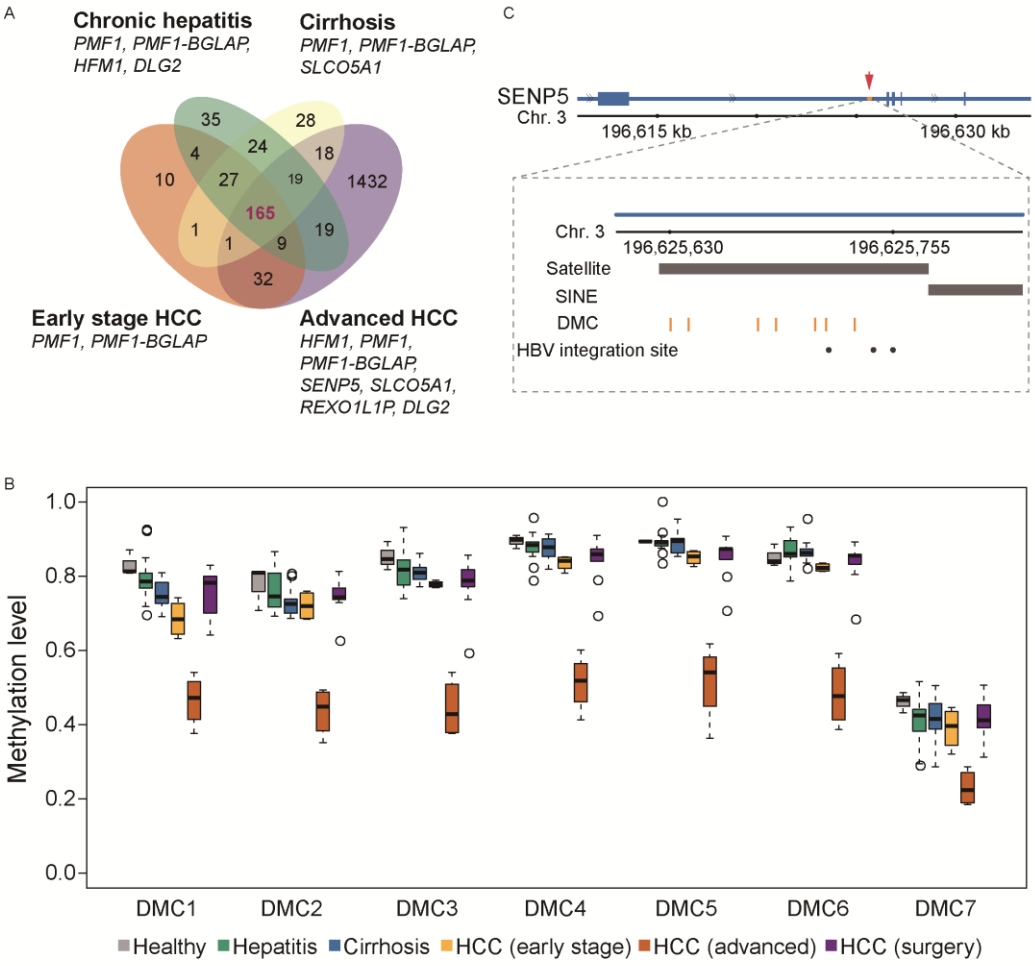


Fig. 3. Differentially methylated CpGs (DMCs) identified by low-pass cell-free WGBS. (A) Left venn diagram showing the overlap of DMCs generated by 1 hypo-methylated chronic hepatitis patients, 2 hypo-methylated cirrhosis patient, 1 hypo-methylated early stage HCC patients and 5 advanced HCC patients compared to healthy individuals. Genes represent the genes annotated with DMCs in each comparison. (B) Boxplot displays the methylation level of 7 DMCs of SENP5 in all the individuals. (C) 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 represent the locus of repeat marker in this region.

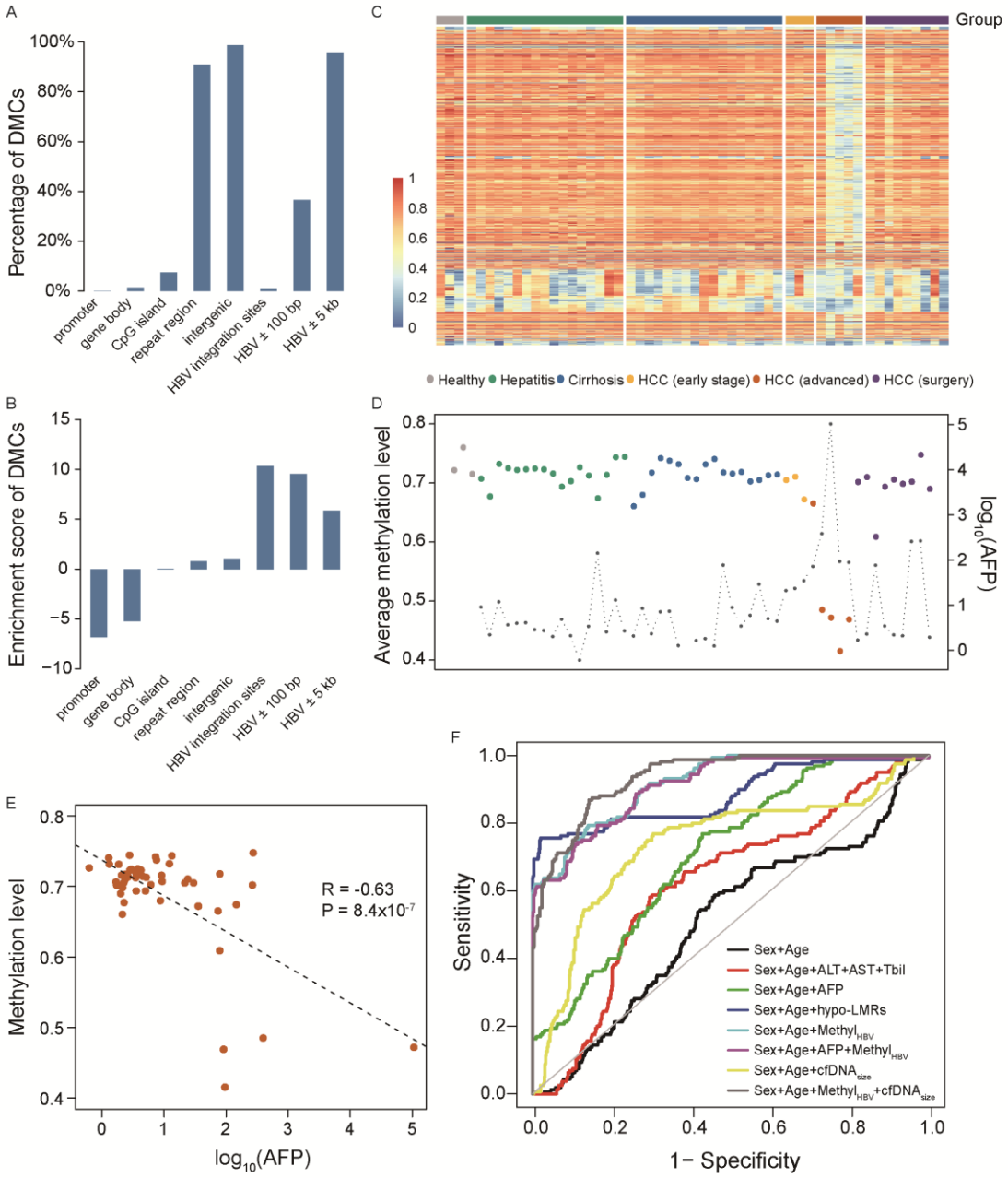


Fig. 4. Overrepresentation of DMCs and CpGs surrounding HBV integration sites. (A) The percentage of DMCs located at different genomic elements and regions surrounding HBV integration sites. (B) The enrichment scores of DMCs at different genomic elements. (C) The heatmap displays the methylation level of the CpGs (>5 reads) located within 100bp of the HBV integration sites in all the samples. (D) The average methylation level of CpGs located within 100 bp of the

712 HBV integration sites ($Methyl_{HBV}$) in all the samples. The black dot represents for
713 AFP level for the corresponding individual. (E) The correlation between AFP
714 (log_{10}) and average methylation level of the CpGs within the 100bp of the reported
715 HBV integration sites ($Methyl_{HBV}$).(F) Receiver operating characteristics (ROC)
716 curve based on five-fold cross-validation for HCC patient detection by different
717 indicators in discriminating HCC from individuals without HCC.

718

719 **Table 1. Area under curve (AUC) for HCC patients detection by different indicators**
720 **and prediction models**

indicator	AUC (95%CI)
Sex+Age	0.51 (0.46-0.56)
Sex+Age+ALT+AST+Tbil	0.62 (0.57-0.67)
Sex+Age+AFP	0.71 (0.67-0.75)
Sex+Age+$cfDNA_{size}$	0.74 (0.69-0.79)
Sex+Age+hypoLMRs	0.89 (0.85-0.92)
Sex+Age+$Methyl_{HBV}$	0.92 (0.90-0.94)
Sex+Age+AFP+$Methyl_{HBV}$	0.91 (0.89-0.94)
Sex+Age+$Methyl_{HBV}$+$cfDNA_{size}$	0.94 (0.92-0.96)

Supplementary Materials

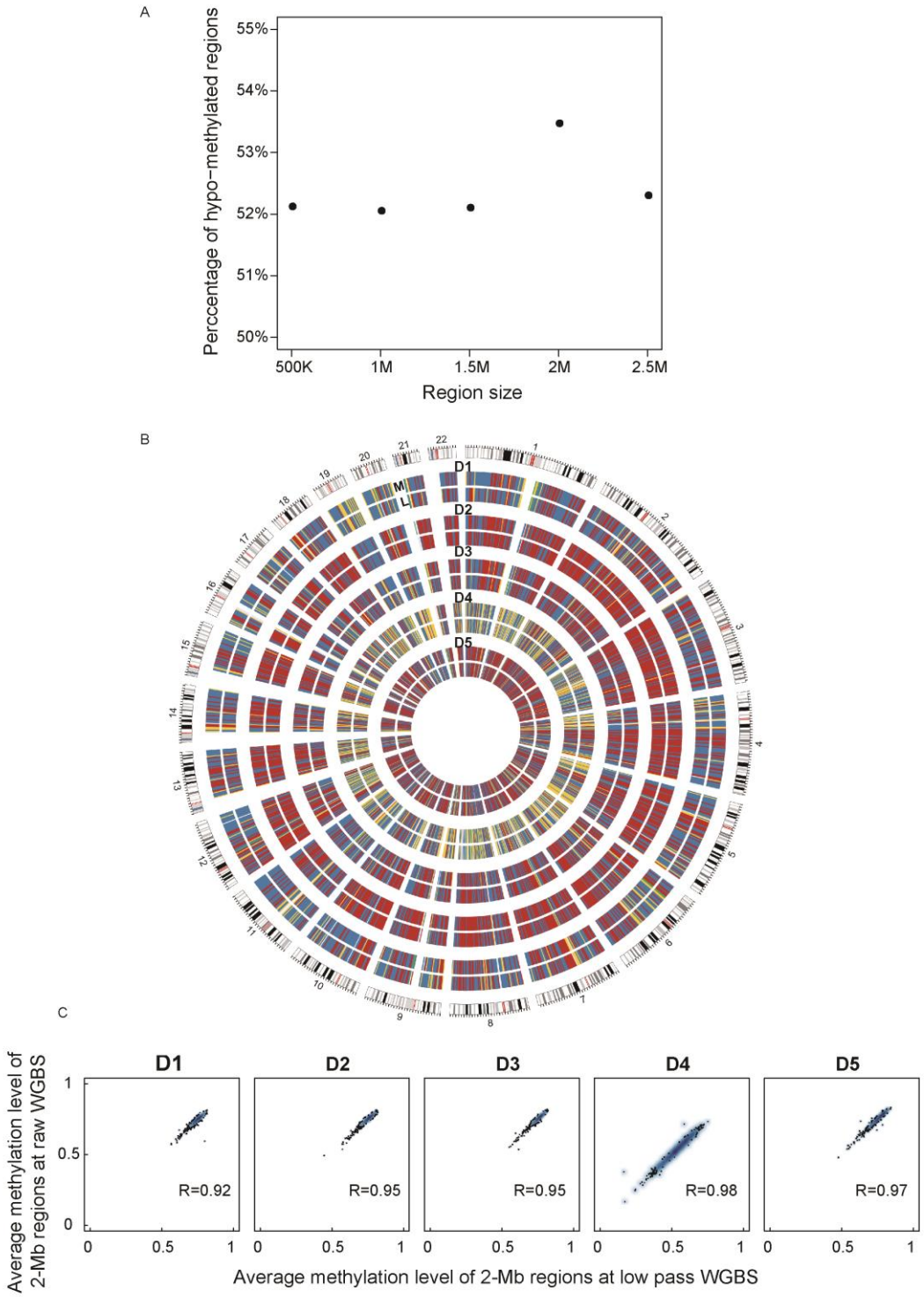


Fig. S1. Determination of optimal region size and effective sequencing depth of low pass WGBS. (A) Percentage of hypo-methylated regions at 500-Kb, 1-Mb, 1.5-Mb, 2-Mb and 2.5-Mb size in the HCC patient (D4). (B) Comparison of average

methylation level of 2-Mb regions between 5M re-sampling reads and total sequencing reads from 5 individuals. Genome-wide DNA methylation level of 2-Mb regions for each comparison are shown in circos. The data represent the average methylation levels for 2-Mb regions. “M” represents the total WGBS and “L” represents the 5M re-sampling reads from total WGBS. Colors represent (from green, purple, yellow, blue and red) the methylation level from low to high. (C). The correlation of average methylation level of 2-Mb regions between 5M re-sampling reads and total sequencing reads. The Pearson’s correlation coefficient is large than 0.92 in all the 5 samples.

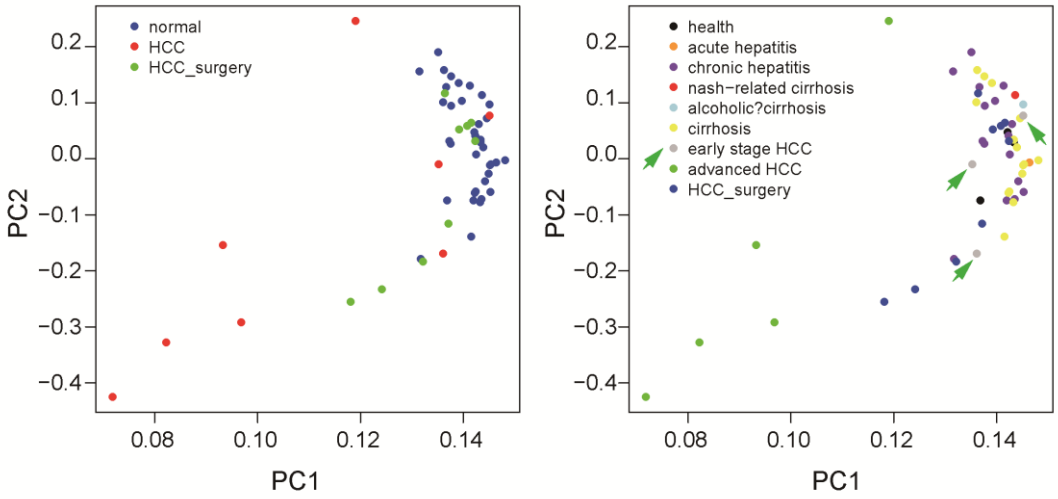
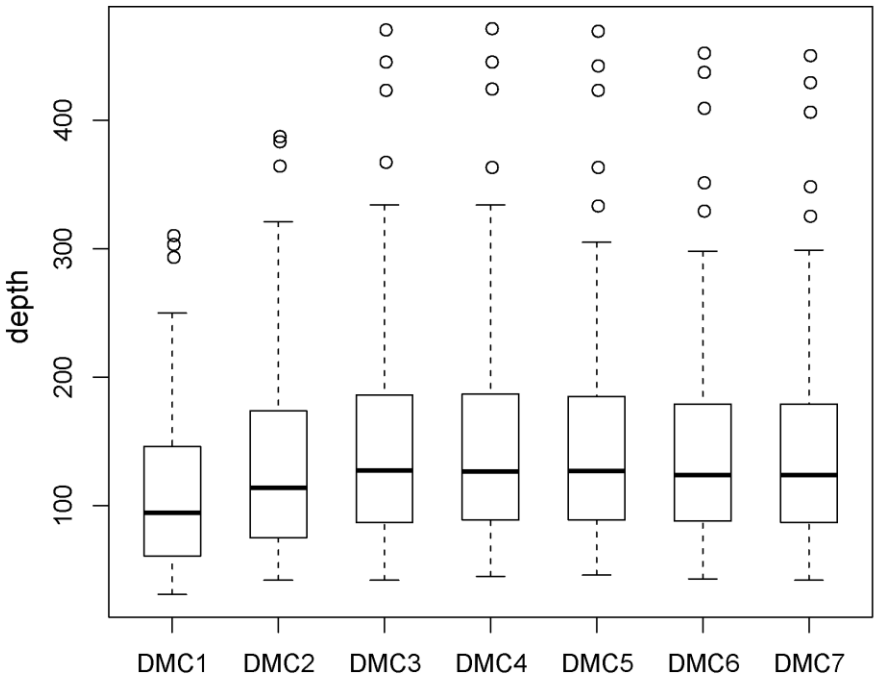
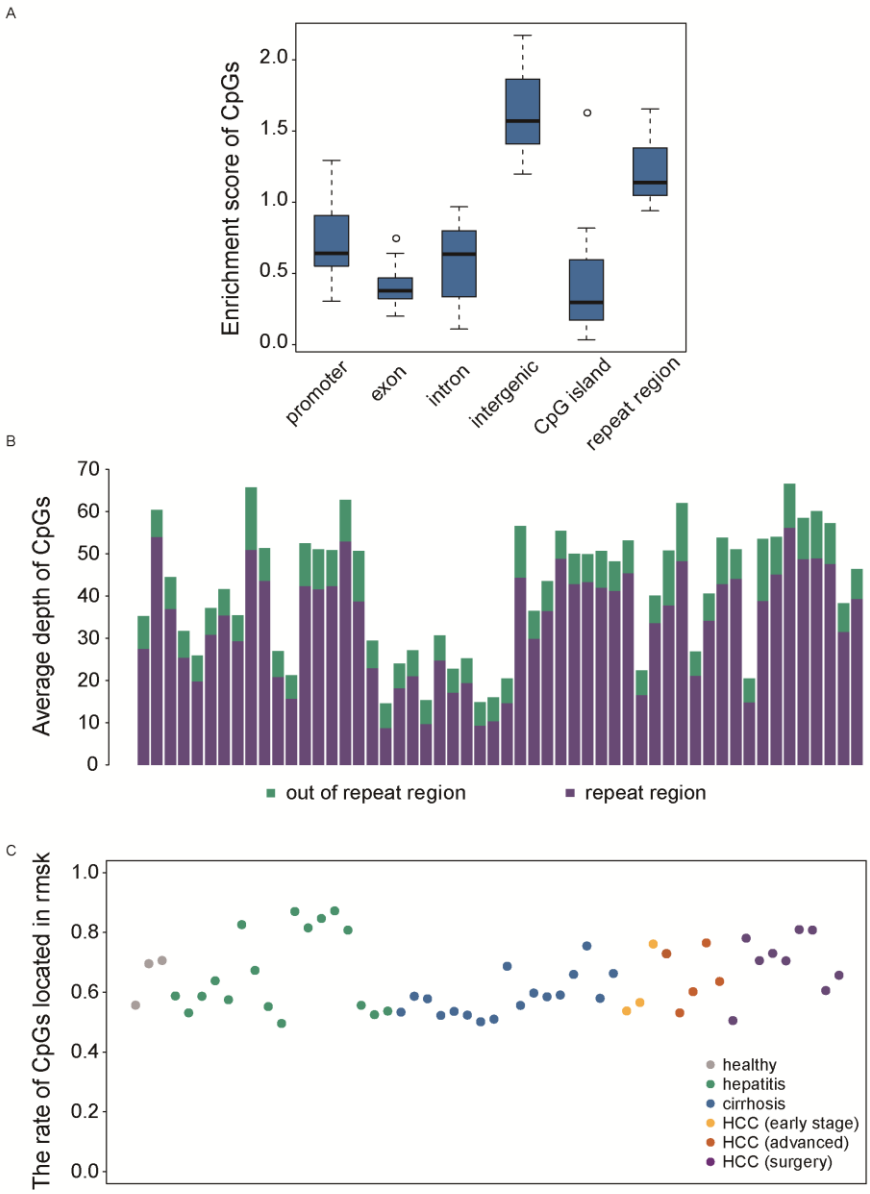


Fig. S2. PCA based on average methylation level of 2-Mb region of all the samples.



739

740 **Fig. S3. The depth of 7 DMCs of SENP5 in all the samples.**



741

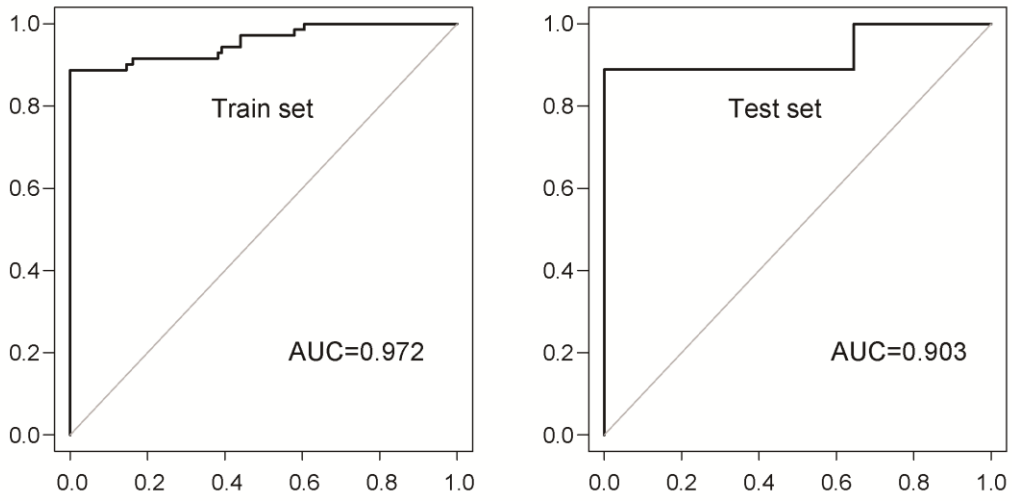
742 **Fig. S4. The genome feature distribution of CpGs at the low-pass WGBS.** (A) The

743 enrichment scores of CpGs in promoter, exon, intron, intergenic, CpG island and

744 repeat regions of all the samples. (B) The average depth of CpGs located in repeat

745 regions and CpGs located outside of repeat regions. (C) The percentage of CpGs

746 located in repeat regions in all the individuals.



747

748 **Fig. S5. Neural network prediction using the top 10 features selected by RF in**

749 **training dataset.**

750

751 **Data files S1: An Excel file including Table S1 to Table S5**

752