SUPPLEMENTARY FIGURES

supplementary figure 1 Technical robustness of the 5hmC-Seal technique. (A) The 5hmC-Seal count data across cfDNA replicates from the same individual across different starting quantities (i.e., 1 ng, 2 ng, 5 ng, and 10 ng) and two batches (i.e., different technicians and reagent lots) show high correlation (Pearson's $r > 0.99$). (B) The principal components analysis (PCA) shows that there are no systematic biases in the processing of these samples. CHB: chronic hepatitis B virus infection; Control: healthy individuals and patients with benign liver lesions; HCC: hepatocellular carcinoma.

supplementary figure 2 The wd-scores distinguish HCC from controls. The wd-scores show a high diagnostic performance for distinguishing (A) HCC from controls (i.e., patients with benign liver lesions and healthy individuals); (B) small HCC (≤ 2.0 cm) from controls; and (C) early stage patients (stage 0/A) from controls. AUC: area under curve; CI: confidence interval; HCC: hepatocellular carcinoma.

supplementary figure 3 Performance of the wd-scores for HCC patients with low AFP levels. The wd-scores show a high diagnostic performance for distinguishing HCC from controls (i.e., patients with benign liver lesions and healthy individuals) at different cutoffs: (A) AFP < 20 ng/mL; and (B) AFP < 400 ng/mL. Note, at 20 ng/mL, 160 and 311 HCC patients would have been misclassified by AFP alone in the training set and validation set 1, respectively; at 400 ng/mL, 261 and 529 HCC patients would have been misclassified by AFP alone in the training set and validation set 1, respectively. AFP: α-fetoprotein; AUC: area under curve; CI: confidence interval; HCC: hepatocellular carcinoma.

supplementary figure 4 Performance of the wd-scores for HCC patients with unknown stage information. The wd-scores show a high diagnostic performance for distinguishing a small set of HCC with missing stage information (n = 147) from (A) non-HCC (i.e., patients with CHB/LC and controls); and (B) patients with CHB/LC. AUC: area under curve; CHB, chronic hepatitis B virus infection; CI: confidence interval; HCC: hepatocellular carcinoma; LC: liver cirrhosis.

supplementary figure 5 The wd-scores distinguish HCC from non-HCC regardless of CHB/LC history. The diagnostic performance of the wd-scores for HCC by CHB/LC history is shown in (A) the training set and (B) validation set 1. The HCC patients are grouped into: +CHB/+LC: HCC with both CHB and LC history; -CHB /-LC: HCC with neither CHB nor LC history; -CHB /+LC: HCC with no CHB history, but with LC history; and +CHB/-LC: HCC with CHB history, but no LC history. AUC: area under curve; CHB, chronic hepatitis B virus infection; CI: confidence interval; HCC: hepatocellular carcinoma; LC: liver cirrhosis.

supplementary figure 6 Performance of the wd-scores for HCC by BCLC stage. The diagnostic performance of the wd-scores by BCLC stage is shown: (A) stage 0; (B) stage A; (C) stage B; (D) stage C; (E) stage 0/A; and (F) stage 0/A/B. Note, only early stage patients (stage 0/A) are included in the training set. AUC: area under curve; BCLC: Barcelona Clinic Liver Cancer staging system; CI: confidence interval; HCC: hepatocellular carcinoma.

supplementary figure 7 Performance of the wd-scores for HCC by tumor size. The diagnostic performance of the wd-scores by tumor size is shown: (A) tumor size ≤ 2.0 cm; (B) 2.0 cm < tumor size ≤4.0 cm; (C) 4.0 cm < tumor size ≤ 6.0 cm; (D) 6.0 cm < tumor size ≤ 8.0 cm; (E) 8.0 cm < tumor size ≤ 10.0 cm; and (F) tumor size >10.0 cm. AUC: area under curve; CI: confidence interval; HCC: hepatocellular carcinoma.

supplementary figure 8 The wd-scores distinguish small HCC from CHB/LC. The wd-scores show a high diagnostic performance for distinguishing small HCC from CHB/LC in

the training set and validation set 1. AUC: area under curve; CHB, chronic hepatitis B virus infection; CI: confidence interval; HCC: hepatocellular carcinoma; LC: liver cirrhosis.

supplementary figure 9 Potential cancer specificity of the diagnostic model for early HCC. The boxplots show that the wd-scores of a set of 89 Chinese patients (stage I/II, n = 50) with primary pancreatic ductal adenocarcinoma (PDAC) are significantly lower than the scores of HCC patients. The cfDNA samples from PDAC patients were processed and profiled using the 5hmC-Seal approach. The same bioinformatic processing pipeline was applied to the PDAC dataset.

supplementary figure 10 Differentially modified genes in cfDNA are enriched with differentially expressed genes in TCGA. Differentially modified genes in terms of 5hmC between early HCC and non-HCC in the training set shows enrichment with differentially expressed genes in TCGA samples (HCC vs. normal tissues) relative to random sampling (e.g., hypergeometric test $p < 0.0001$ for the top 200 modified genes in cfDNA). HCC: hepatocellular carcinoma; TCGA: The Cancer Genome Atlas.

SUPPLEMENTARY METHODS

Clinical definitions and study subjects

Tumor size was defined as the largest diameter of the primary tumor recorded in the pathology reports or radiographic imaging (mainly MRI). Small HCC was defined as a single HCC nodule with a diameter ≤ 2.0 cm. Healthy controls were collected from individuals who underwent routine physical examinations and had normal liver biochemistry, no liver mass under ultrasound, and no history of viral hepatitis or malignant diseases. Samples for benign liver lesions were derived from patients diagnosed with various benign hepatic diseases, including cysts, cavernous hemangiomas, hepatocellular adenomas, and focal nodular hyperplasia. Patients with a history of CHB but no LC were those with the presence of HBsAg for the past 6 months and without development of malignant tumors, according to the American Association for the Study of Liver Diseases guidelines.[1] Patients with LC were identified based on clinical manifestations and no evidence of hepatic mass in 3 months prior to recruitment, according to the guideline of the Japanese Society of Gastroenterology.[2] Diagnosis was confirmed based on tissue pathology for those patients (~83%) whose tumor biopsies were available, mostly through surgical removal. For those late stage patients whose tumors could not be removed, the diagnosis was based on the Liver Imaging Reporting AND Data System (LI-RADS),[3] needle biopsy, serum markers etc. Tumor stages were determined according to the Barcelona Clinic Liver Cancer (BCLC) staging system.[4] The majority of the HCC patients (~81%) included in this study were subjected to surgical removal, ~11% under transarterial chemoembolization (TACE), ~5% under radiofrequency ablation (RFA), and ~3% under other treatments (e.g., transplant).

For the 1,204 HCC patients from Zhongshan Hospital of Fudan University and The Eastern Hepatobiliary Surgery Hospital, the average age at diagnosis was 56.9±10.9 years (range, 21 to 88 years), median BMI was 23.4 (range 15.59, 38.19), 84.2% (n = 1,014) were males, and 91.5% (n = 1,102) reported CHB or LC. Of these 1,204 HCC patients, 48.1 % (n = 579) were stage 0/A, which were randomly grouped into the training set and the main validation set (i.e., "validation set 1"). For the 392 non-cancer patients with CHB/LC history, the average age at collection of blood was 40.6±13.3 years, and 67.6% (n = 265) were males. An independent validation set of 60 HCC patients from other participating hospitals were

included for exploring external validity (i.e., "validation set 2"). In addition, the majority of the LC-related HCC in this study were related to CHB (~95%), a major risk factor for LC and HCC in China. Supplementary table 1-2 provide more details about the study subjects.

Preparation of DNA samples

Peripheral blood was used to prepare cfDNA for the 5hmC-Seal profiling. Briefly, 5-10 mL of peripheral blood was collected from each subject using EDTA anticoagulant tubes, and the plasma was processed within 6 h by centrifuging twice at 1,350 × g for 12 min, and then centrifuging at 13,500 × g for 12 min. The processed plasma samples (about 2 mL/subject) were immediately stored at -80°C. The QIAamp Circulating Nucleic Acid Kit (Qiagen, Germany) was used to isolate cfDNA. For 26 HCC patients with available cfDNA samples, we also collected their tumors and adjacent tissue samples for comparison. The tissue samples were stored at -80°C after surgical removal. About 10-25 mg of tissue was collected using a scalpel after sample thawing. Genomic DNA (gDNA) was then isolated using the ZR Genomic DNA Tissue Kit (Zymo Research, California, USA).

Preparation of the 5hmC-Seal libraries and sequencing

Detailed information about 5hmC-Seal library construction and sequencing has been reported previously.[5, 6] All samples were randomized with respect to phenotype/diagnosis status in the 5hmC-Seal library construction and sequencing steps. The 5hmC-Seal combines a series of routine sample preparation and sequencing steps with a unique pull-down step based on covalent chemistry. In the first routine step, the Illumina-compatible adaptors were installed on cfDNA or fragmented gDNA. Next, T4 bacteriophage β-glucosyltransferase was used to transfer an engineered glucose moiety containing an azide group onto the hydroxyl group of 5hmC across the human genome. The azide group was then chemically modified with biotin, and streptavidin beads were used for affinity enrichment of 5hmC-containing DNA fragments. Subsequently, routine PCR amplification and sequencing (PE38) were conducted using the Illumina NextSeq500 platform (Illumina, California, USA) at Shanghai Epican Genetech Co. Ltd. (Shanghai, China). Out of the 2,574 individuals we recruited for this study, 20 cfDNA samples were excluded due to technical reasons such as low quality of the DNA.

There are several scientifically and clinically relevant characteristics of the 5hmC-Seal profiling method that distinguish it from other existing epigenomic technologies. First, it is an unbiased genome-wide method, thus it can be used to generate data-driven *de novo* diagnostic models that do not require pathology-related assumptions to inform sampling/analytical/probe panel strategies for particular diseases, as would be required before the use of for example reduced representative bisulfite sequencing, probe-based methods, or PCR-based methods. A routine 5hmC-Seal profiling analysis can examine the entire human genome, covering ~22,000 known genes as well as various regulatory elements,[5] offering potentially cost-effective data and flexibility compared with single gene/locus-based approaches.

Processing of the 5hmC-Seal data

The raw sequencing reads from the 5hmC-Seal were first processed to remove adapter sequences using Trimmomatic (v0.36).[7] Low quality bases at the 5' and 3' were trimmed based on phred score to a minimum length of 30 bp. The sequencing reads were aligned to the human genome reference (hg19) using Bowtie2 (v2.2.6) with the end-to-end alignment mode.[8] Read pairs were concordantly aligned with fragment length ≤ 500 bp and with average ≤ 1 ambiguous base and up to four mismatched bases per 100 bp length. Alignments with Mapping Quality Score ≥ 10 were counted for overlap with gene bodies (annotated by the GENCODE Project),[9] using featureCounts (v1.5.0-p3),[10] without strand information. The

raw fragment counts were then normalized using DESeq2,[11] which performs an internal normalization that corrects for library size. Our previous study in cfDNA samples from patients with colon and stomach cancers suggested that the gene body is a robust genomic feature for summarizing 5hmC modifications using the 5hmC-Seal.[5] To visualize the distinguishing capacity of the detected 5hmC marker genes, the heatmaps (figure 4A-B) were plotted using the hierarchical clustering[12] of samples (columns) and genes (rows) based on the Euclidean distance in terms of normalized read counts.

Exploration of gene regulatory elements

The histone modification data, specifically H3K4me1 and H3K27ac that mark enhancers, derived from various adult tissues from the Roadmap Epigenomics Project[13] were used to explore potential gene regulatory relevance of the cfDNA samples in a random set of HCC patients (n = 50) and healthy individuals (n = 50). Tissue-derived H3K4me1 and H3K27ac data were downloaded from the Roadmap Epigenomics Project, including liver (E066), primary T cells (E034), brain (E098), colon (E075), stomach (E094), heart (E095), lung (E096), and pancreas (E098).[13] Specifically, the Gapped Peak data (BED tracks) were downloaded (accessed on February 10, 2019) from the Roadmap. The 5hmC profiles from this study were summarized by counting overlaps with the Roadmap H3K4me1/H3K27ac peaks, as well as their flanking regions divided into 2.5K bp bins, e.g., +5K bp to +2.5K bp, +2.5K bp to the start sites of H3K4me1 peaks. For each cfDNA sample, the average read counts across all H3K4me1 or H3K27ac peaks were calculated, and were normalized to the 2.5K bp bins. The average length of the liver-derived H3K4me1 and H3K27ac peaks were 3,596 bp and 4,145bp respectively. The 5hmC-Seal data were found to be enriched in liver-derived H3K4me1 and H3K27ac regions, while depleted in their flanking regions (+/-5K bp and +/-2.5K bp; figure 2B, C). The average fold changes between HCC patients and healthy individuals were calculated for each set of tissue-derived histone modification peaks. The fold changes were tested against the null hypothesis (fold change = 1) using the two-sided t-test for the ratios of two means, and the 95% CIs for the fold changes were calculated using the Fieller's method[14] (figure 2D, E). In figure 2B, C, the x-axis is the relative position related to the start/end of respective histone modification peak, and the y-axis is the average count across all peaks in all samples. In figure 2D, E, the x-axis represents the ratio between the average counts of 50 HCC and 50 healthy individuals. In addition, to further demonstrate the potential underlying biology for the detected 5hmC marker genes, the predicted enhancer regions were obtained from the EnhancerAtlas[15] database which was based on the ENCODE (Encyclopedia of DNA Elements) Project (GM12878).[16]

Comparison between plasma and tissue samples

For the 26 HCC patients with available plasma cfDNA, tumor, and adjacent tissue samples, we evaluated the correlation relationships across different sources of DNA and different individuals. All genes were ranked by their variability (i.e., variance) in 5hmC modification levels in cfDNA samples, tumors, or tumor-adjacent tissue samples, separately. The top ranked genes with the highest variability in cfDNA samples were evaluated for overlapping with those top ranked genes in tumors or tumor-adjacent tissue samples. The observed number of overlapped top ranked genes between cfDNA and tumor/adjacent tissue samples was compared with the number of overlapped genes from random sampling using the hypergeometric test (figure 3A). For the top 30 most variable genes in cfDNA, the Pearson's correlation was calculated across the 26 pairs of cfDNA/tumor (TU), and cfDNA/adjacent tissue (TI) samples (figure 3B-C).

Comparison within controls

The controls were comprised of healthy individuals and patients with benign liver lesions. There were no significant differences in terms of differential 5hmC-Seal count data between healthy individuals and patients with benign liver lesions in the training set and validation set 1 at 5% false discovery rate.

Evaluation of functional relevance of the diagnostic model

The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis using the NIH/DAVID tool[17, 18] was used to explore the underlying biological connections of the 917 candidate genes that were included for further feature selection and model building (see supplementary table 8). To provide further biological insights into the marker genes, we compared the 6,691 genes showing potentially differential 5hmC modification in gene bodies ($p<0.01$) between early HCC and controls in the training set (see online supplementary table 3, top100 genes shown as examples) with the differentially expressed genes in HCC from The Cancer Genome Atlas (TCGA).[19] In total, RNA expression data from 374 primary HCC tumors from US patients and 50 normal liver tissues were obtained from TCGA for the 6,691 genes using the XENA browser (http://xena.uscs.edu/). The $\log_2$ fold changes of gene expression were calculated between tumors and normal tissues. The top differentially modified genes in cfDNA samples, ranked by p-value from logistic regression for early HCC vs. non-HCC in the training set, were evaluated for overlapping with the top differentially expressed genes obtained from TCGA, and compared with the distribution of overlapped genes from random sampling using the hypergeometric test (e.g., hypergeometric test $p < 0.0001$ for the top 200 modified genes in cfDNA; supplementary figure 10). For each given number of top ranked genes (e.g., top 200), the random sampling process was performed for 1,000 times, and the mean value (e.g., overlapping between a random gene set and the top 200 differentially modified genes in cfDNA) was used for the hypergeometric test and figure illustration.

To explore potential molecular mechanisms underlying the marker genes, we evaluated the relationships between the 5hmC levels of the marker genes and the 5hmC levels in liver-derived histone modification marks or predicted enhancer regions.The 5hmC profiles in cfDNA were summarized by counting overlaps with these histone modification marks or the predicted enhancer regions. Overall, the Pearson's correlation tests showed significant correlation between the 5hmC profiles in gene bodies and the 5hmC profiles in the combined H3K4me1/H3K27ac/enhancer regions for the majority of marker genes (see online supplementary table 6). The 5hmC coefficients for the comparison between early HCC and non-HCC in individual histone marks or enhancer regions were calculated using the logistic regression adjusted for age and gender in the training set. Our findings suggested a possible underlying mechanism for at least some marker genes (e.g., *SOX9*), e.g., driven by the differential modification at individual histone modification marks/enhancers in these genes (see online supplementary table 6-7).

REFERENCES

1 Lok AS, McMahon BJ. Chronic hepatitis B: update 2009. *Hepatology* 2009;50:661-2.

2 Fukui H, Saito H, Ueno Y*, et al.* Evidence-based clinical practice guidelines for liver cirrhosis 2015. *J Gastroenterol* 2016;51:629-50.

3 Chernyak V, Fowler KJ, Kamaya A*, et al.* Liver Imaging Reporting and Data System (LI-RADS) Version 2018: Imaging of Hepatocellular Carcinoma in At-Risk Patients. *Radiology* 2018;289:816-30.

4 Pons F VM, Llovet JM. Staging systems in hepatocellular carcinoma. HPB (Oxford) 2005.
5 Li W, Zhang X, Lu X*, et al.* 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Res* 2017;27:1243-57.
6 Han D, Lu X, Shih AH*, et al.* A Highly Sensitive and Robust Method for Genome-wide 5hmC Profiling of Rare Cell Populations. *Mol Cell* 2016;63:711-9.
7 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114-20.
8 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357-9.
9 Harrow J, Frankish A, Gonzalez JM*, et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012;22:1760-74.
10 Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;30:923-30.
11 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15:550.
12 Eisen MB, Spellman PT, Brown PO*, et al.* Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998;95:14863-8.
13 Roadmap Epigenomics C, Kundaje A, Meuleman W*, et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317-30.
14 Zerbe GO. On Fieller's Theorem and the General Linear Model. *The American Statistician* 1978;32:103-5.
15 Gao T, He B, Liu S*, et al.* EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics* 2016;32:3543-51.
16 An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57-74.
17 Kanehisa M, Sato Y, Kawashima M*, et al.* KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;44:D457-62.
18 Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009;37:1-13.
19 Cancer Genome Atlas Research N, Weinstein JN, Collisson EA*, et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 2013;45:1113-20.