Supplementary Information for

Epigenetic analysis of cell-free DNA by fragmentomic profiling

Qing Zhou[a,b,c,1], Guannan Kang[a,b,c,1], Peiyong Jiang[a,b,c,1], Rong Qiao[a,b,c], W.K. Jacky Lam[a,b,c], Stephanie C.Y. Yu[a,b,c], Mary-Jane L. Ma[a,b,c], Lu Ji[a,b,c], Suk Hang Cheng[a,b,c], Wanxia Gai[a,b,c], Wenlei Peng[a,b,c], Huimin Shang[a,b,c], Rebecca W.Y. Chan[a,b,c], Stephen L. Chan[d,e], Grace L.H. Wong[f], Linda T. Hiraki[g], Stefano Volpi[h,i], Vincent Wai-Sun Wong[f], John Wong[j], Rossa W. K. Chiu[a,b,c], K. C. Allen Chan[a,b,c], and Y. M. Dennis Lo[a,b,c,2]

Y. M. Dennis Lo

Email: loym@cuhk.edu.hk.

This PDF file includes:

1. Figure S1 to S5
2. Table S1 to S2
3. Materials and Methods
4. SI References

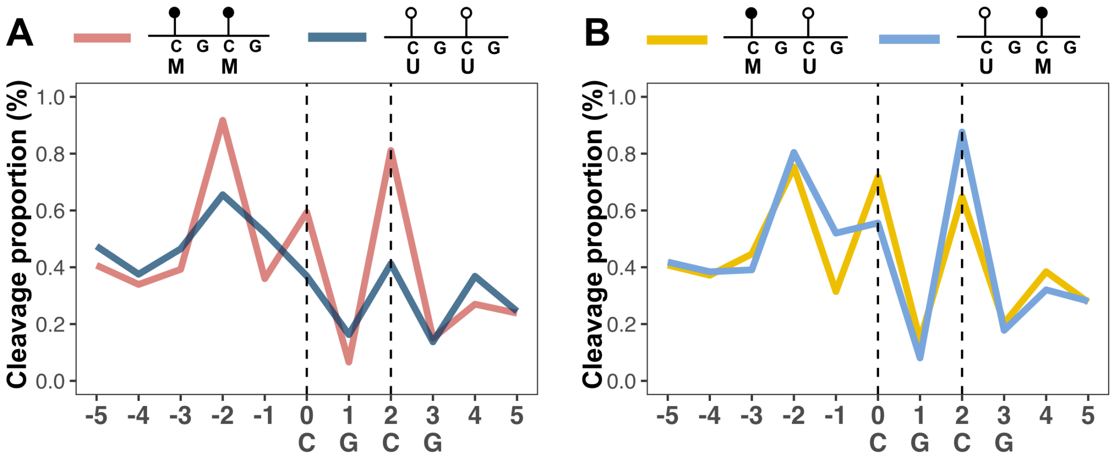

Fig. S1. Cleavage proportion of cfDNA molecules depending on CpG methylation status (whole genome sequencing data). (A) and (B) Cleavage profiles in windows each containing two tandem CpG dinucleotides spanning positions of 0, 1, 2, and 3 (i.e., CGCG subsequence) in the pool of cfDNA samples from 8 healthy controls. Red, dark blue, yellow, and light blue lines correspond to the cleavage profiles with different methylation configurations of two immediately adjacent CpG sites, namely MM, UU, MU, and UM where 'M' and 'U' represent the hypermethylated and hypomethylated state, respectively.

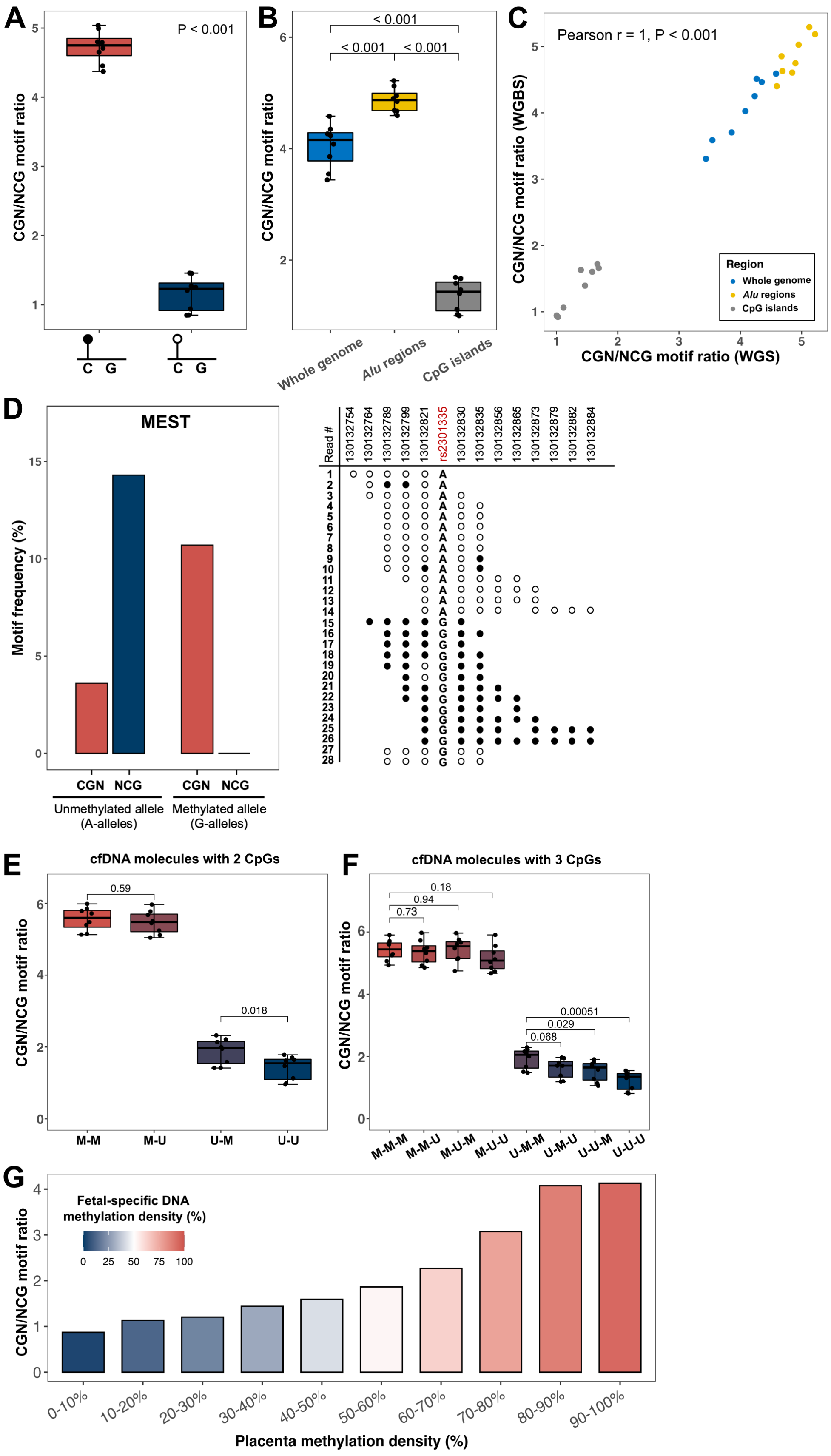
A
P < 0.001
CGN/NCG motif ratio
1
2
3
4
5
C G
C G
B
< 0.001
< 0.001
< 0.001
CGN/NCG motif ratio
2
4
6
Whole genome
Alu regions
CpG islands
C
Pearson r = 1, P < 0.001
CGN/NCG motif ratio (WGBS)
CGN/NCG motif ratio (WGS)
1
2
3
4
5
Region
Whole genome
Alu regions
CpG islands
D
MEST
Motif frequency (%)
0
5
10
15
CGN
NCG
CGN
NCG
Unmethylated allele
(A-alleles)
Methylated allele
(G-alleles)
Read #
130132754
130132764
130132789
130132799
130132821
rs2301335
130132830
130132835
130132856
130132865
130132873
130132879
130132882
130132884
E
cfDNA molecules with 2 CpGs
CGN/NCG motif ratio
0.59
0.018
0
2
4
6
M-M
M-U
U-M
U-U
F
cfDNA molecules with 3 CpGs
CGN/NCG motif ratio
0.18
0.94
0.73
0.00051
0.029
0.068
0
2
4
6
M-M-M
M-M-U
M-U-M
M-U-U
U-M-M
U-M-U
U-U-M
U-U-U
G
Fetal-specific DNA
methylation density (%)
0
25
50
75
100
CGN/NCG motif ratio
0
1
2
3
4
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
90-100%
Placenta methylation density (%)

Fig. S2. CGN/NCG motif ratio analysis. (A) Boxplot of CGN/NCG motif ratio between hypermethylated and hypomethylated CpGs from plasma DNA of 8 healthy control samples (whole genome non-bisulfite sequencing data). **(B)** CGN/NCG motif ratios of the whole genome, *Alu* regions, and CpG islands, respectively (whole genome sequencing data). **(C)** The correlation between the CGN/NCG motif ratio calculated based on the whole genome non-bisulfite sequencing data and the whole genome bisulfite sequencing data from 8 healthy control plasma DNA samples for the whole genome, Alu regions and CpG islands, respectively. **(D)** The methylation status of sequenced fragments mapped to an imprinting region (MEST gene, located at chr7:130,132,754-130,132,884). Each row with the black (methylated) and white (unmethylated) dots represents one plasma DNA molecule. Each dot represents one CpG site. Two groups of sequenced fragments carried A alleles and G alleles, respectively, at a SNP (rs2301335). The frequencies of CGN and NCG motifs related to the imprinting region are shown in the left panel. The right panel shows that cfDNA molecules carrying A alleles and G-alleles exhibit distinct methylation patterns. **(E) and (F)** Impact of methylation patterns with multiple adjacent CpGs on CGN/NCG motif ratio. Adjacent CpGs were defined as those CpG sites located within a range of 75 bp in size but not in tandem. CGN/NCG motif ratios across different combinations of methylation states for those molecules with 2 and 3 adjacent CpG sites were analyzed. For cfDNA molecules with 2 adjacent CpGs, there were a total of 4 combinations of methylation states. One combination could be the situation where the methylated CpG at 5' end was followed by a methylated CpG (denoted by "M-M", where 'M' represents the methylated CpG and '-' represents any one or more nucleotides). The other combinations could be "M-U", "U-M", "U-U", where 'U' represents the unmethylated CpG). For cfDNA molecules with 3 CpGs, there were a total of 8 combinations of methylation states, namely, "M-M-M", "M-M-U", "M-U-M", "M-U-U", "U-M-M", "U-M-U", "U-U-M", and "U-U-U". **(G)** The CGN/NCG motif ratios from fetal-specific cfDNA in maternal plasma DNA (3rd trimester) correlated with the methylation levels in the paired placenta tissue. CpGs were grouped into ten groups according to the methylation levels from paired placenta tissue. The y-axis represents the CGN/NCG motif ratio of fetal-specific cfDNA, and the graded colors in the bars represent the methylation density of fetal-specific cfDNA.

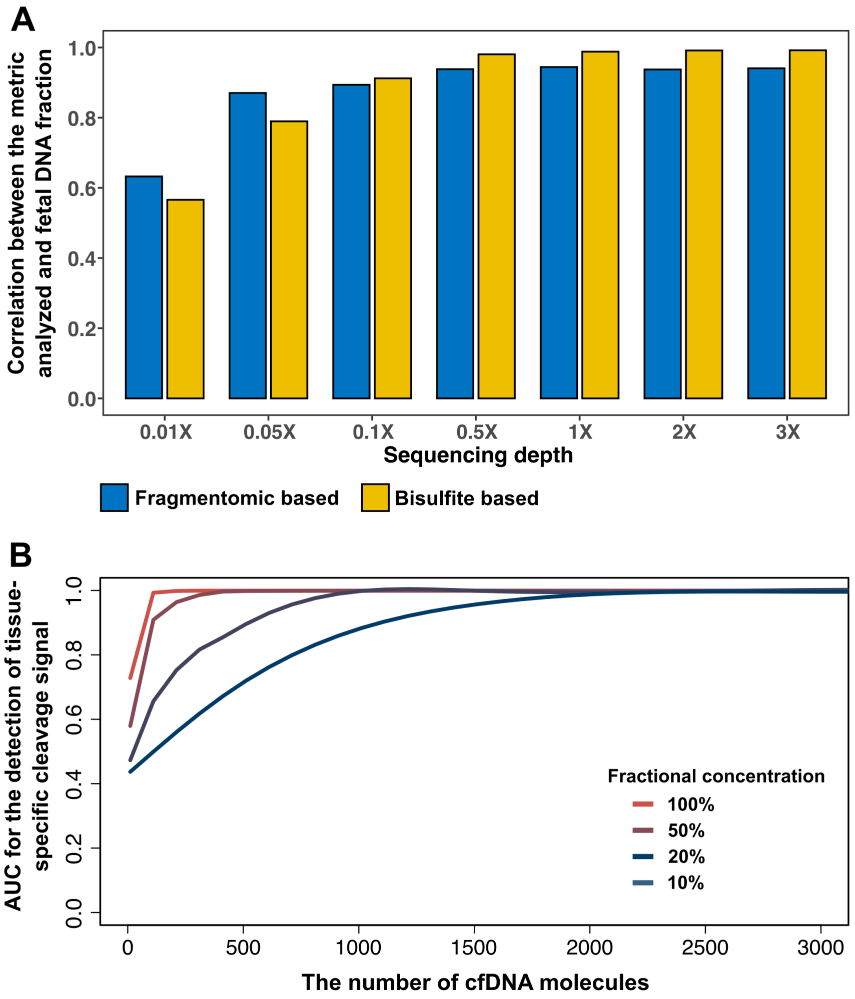

Fig. S3. (A) Comparison of methylation analyses based on cfDNA fragmentomics and bisulfite sequencing in pregnant women. The Y-axis represents the Pearson's correlation coefficient between the metric [CGN/NCG end motif (blue; the absolute difference between the CGN/NCG motif ratio from placenta-specific hypermethylated and hypomethylated CpGs)) or placental DNA contribution by methylation-based plasma DNA tissue mapping (yellow)] and fetal DNA fraction deduced by SNP-based approach. The X-axis represents different sequencing depths. **(B)** Impact of the number of sequenced cfDNA molecules and fractional concentration of target tissue DNA in plasma on the detection power of tissue-specific cleavage signal based on the computer simulation.

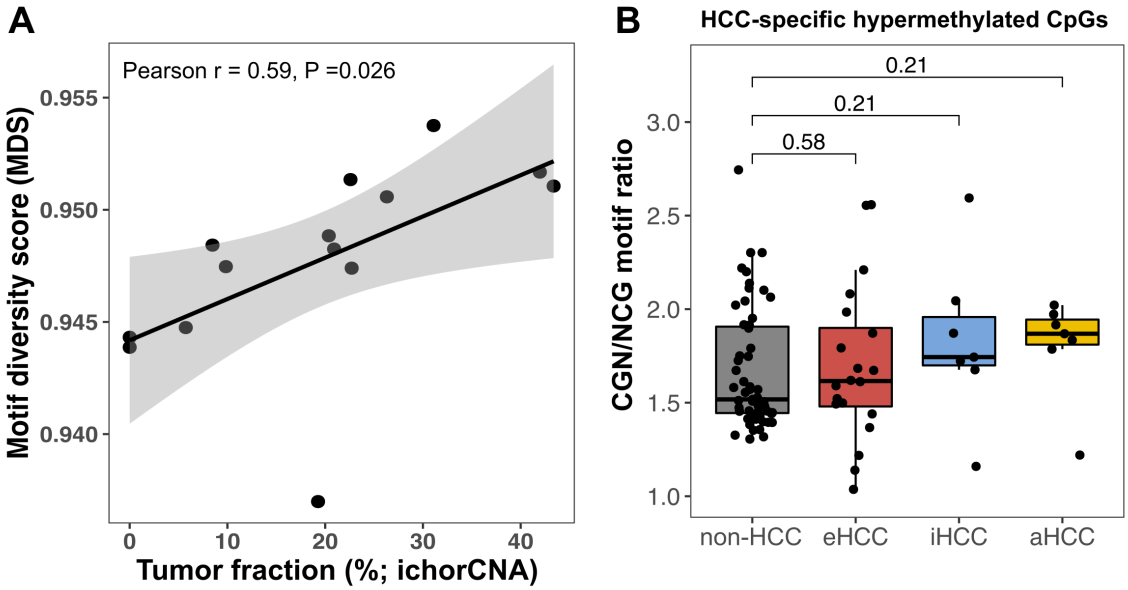

Fig. S4. (A) The correlation between the motif diversity score (MDS) and tumor DNA fraction calculated based on copy number aberrations in HCC cases. **(B)** The CGN/NCG motif ratio concerning HCC-specific hypermethylated CpGs in plasma DNA among non-HCC cases (healthy controls and HBV carriers), HCC cases with early (eHCC), intermediate (iHCC), and advanced (aHCC) stages.

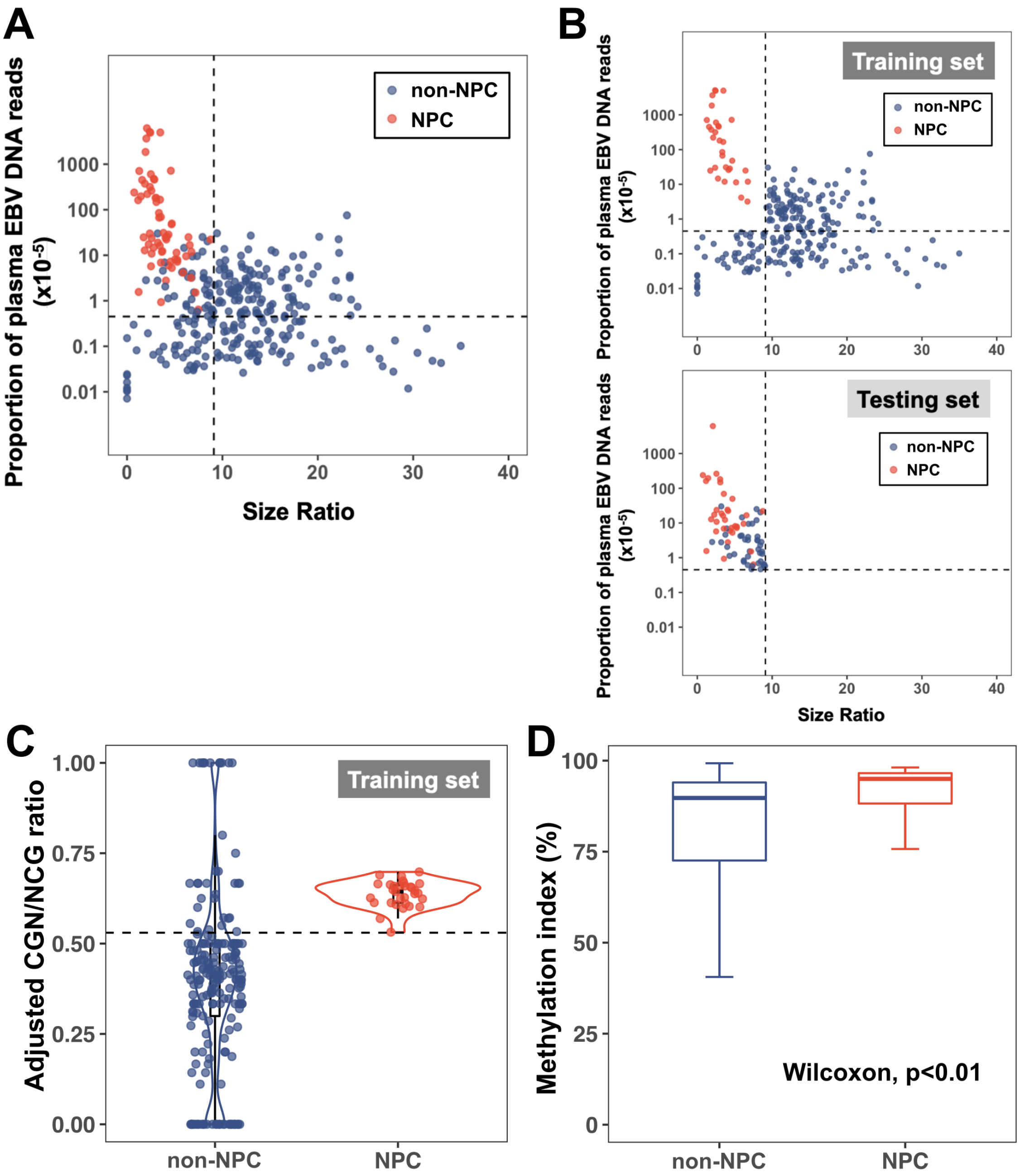

Fig. S5. CGN/NCG motif ratio used for NPC screening. (A) The classification cutoffs of EBV DNA proportion and EBV DNA size ratio used in a previous study (1). **(B)** Data points used in the training set (upper panel) and testing set (bottom panel). **(C)** The adjusted CGN/NCG motif ratios of informative CpGs in the EBV genome between non-NPC and NPC individuals (training set). **(D)** The methylation index of informative CpGs measured by bisulfite sequencing between non-NPC and NPC cases.

Sample names	Status	BCLC stage	Tumor DNA fraction estimated by ichorCNA (%)
TBR1015	eHCC	BCLC A	3.40
TBR1373	eHCC	BCLC 0	0.00
TBR1388	eHCC	BCLC A	4.58
TBR1586	eHCC	BCLC A	0.00
TBR1644	eHCC	BCLC A	13.85
TBR1662	eHCC	BCLC A	4.35
TBR1676	eHCC	BCLC 0	0.00
TBR1677	eHCC	BCLC 0	0.00
TBR1772	eHCC	BCLC A	0.00
TBR1774	eHCC	BCLC 0	0.00
TBR1828	eHCC	BCLC A	0.00
TBR1850	eHCC	BCLC A	6.45
TBR1886	eHCC	BCLC A	0.00
TBR1899	eHCC	BCLC A	6.44
TBR1916	eHCC	BCLC A	0.00
TBR1920	eHCC	BCLC 0	0.00
TBR1932	eHCC	BCLC A	0.00
TBR2038	eHCC	BCLC A	3.60
TBR2080	eHCC	BCLC A	5.90
TBR931	eHCC	BCLC 0	0.00
TBR1555	iHCC	BCLC B	5.74
TBR1723	iHCC	BCLC B	0.00
TBR846	iHCC	BCLC B	0.00
TBR852	iHCC	BCLC B	20.91
TBR858	iHCC	BCLC B	9.84
TBR874	iHCC	BCLC B	8.47
TBR964	iHCC	BCLC B	19.27
TBR1757	aHCC	BCLC C	22.59
TBR1838	aHCC	BCLC C	31.09
TBR1861	aHCC	BCLC C	43.37
TBR2000	aHCC	BCLC C	26.31
TBR2083	aHCC	BCLC C	22.71
TBR853	aHCC	BCLC C	20.37
TBR855	aHCC	BCLC C	41.96

Table S1. Cancer staging information and tumor DNA fraction estimated by ichorCNA for all HCC cases.

Plasma DNA	Sample types	Sequencing methods	sample sizes	Clinical information	Reference
	Healthy controls	WGBS	8	Healthy Control	(2)
		WGS	38		
	HBV carriers	WGS	13	HBV+, without cirrhosis	
			4	HBV+, with cirrhosis	
	HCC patients	WGS	20	HCC, early-stage	
			7	HCC, intermediate-stage	
			7	HCC, advanced-stage	
	Pregnant women	WGBS	10	1st trimester	
		WGBS	10	2nd trimester	
		WGBS	10	3rd trimester	
	Liver transplant samples	WGS	14	Liver Transplantation	(3)
	Pregnant women	WGBS	1	1st trimester	(4)
		WGBS	1	3rd trimester	
	Patients with DNASE1L3 deficiency	WGBS	4	Patients carrying DNASE1L3 homozygous mutation	(5)
	non-NPC individuals (n=272)	WGS	179	Transiently positive EBV	(1)
		WGS	93	Persistently positive EBV	
	NPC patients (n=65)	WGS	34	NPC, screening cohort	
		WGS	31	NPC, external cohort	
	non-NPC (n=160)	WGBS	110	Transiently positive EBV	(6)
		WGBS	50	Persistently positive EBV	
	NPC (n=47)	WGBS	33	NPC, screening cohort	
		WGBS	14	NPC, external cohort	
Tissue DNA	Placenta	WGBS	1	Pool of 4 samples	(4, 7)
	Buffy coat	WGBS	1	Pool of 6 samples	(4, 7)
	Liver	WGBS	1		(8)
	Lung	WGBS	1		
	Colon	WGBS	1		
	HCC	WGBS	1	Pool of 13 samples	(9)

Table S2. Summary for the datasets used in this study. WGBS: whole genome bisulfite sequencing; WGS: whole genome sequencing.

Materials and Methods

Identification of tissue-specific reads and calculation of tissue-specific DNA fraction

The genotypes regarding the maternal buffy coat and placenta/chorionic villus tissue samples were obtained using microarray-based genotyping technology (HumanOmni2.5 genotyping array Illumina), and informative SNPs were identified from sites where the mother was homozygous (denoted as AA genotype) and the fetus was heterozygous (denoted as AB genotype). Fetal-specific DNA fragments were identified according to the DNA fragments carrying fetal-specific alleles at informative SNP sites. In this scenario, the B allele was fetal-specific, and the DNA fragments carrying the B allele were deduced to be originated from the placenta. 'A' alleles were deemed shared alleles. The number of fetal-specific molecules (p) carrying the fetal-specific alleles (B) was determined. The number of molecules (q) carrying the shared alleles (A) was determined. The fetal DNA fraction across all cell-free DNA samples would be calculated by 2p/(p+q)*100%. To obtain maternal-specific DNA molecules, the other set of informative SNPs were identified from sites where the mother was heterozygous (denoted as AB genotype), and the fetus was homozygous (denoted as AA genotype). Maternal-specific DNA fragments were identified according to the DNA fragments carrying maternal-specific alleles at informative SNP sites. In this scenario, the B allele was maternal-specific, and the DNA fragments carrying the B allele were deduced to be originated from maternal alone. A similar data processing was applied to the identification of donor-derived DNA. Informative SNPs were identified from sites where the donor was heterozygous (denoted as AB genotype), and the recipient was homozygous (denoted as AA genotype). cfDNA fragments carrying the B allele were deduced to be originated from the donor. The number of DNA molecules carrying the donor-specific alleles (p) and carrying the shared alleles (q) was determined. The donor DNA fraction was calculated by 2p/(p+q)*100% for each sample with liver transplantation.

Identification of tissue-specific hypermethylated and hypomethylated CpGs

For tissue-specific DNA fraction estimation with CGN/NCG motif ratios, the tissue-specific hypermethylated CpGs referred to CpGs with a methylation density of over 70% in the target tissue and below 30% in the buffy coat, and hypomethylated CpGs referred to CpGs with a methylation density of below 30% in the target tissue and over 70% in the buffy coat. The same criterion was used to determine the HCC-specific hypermethylated and hypomethylated CpGs.

Computer simulation

To investigate the sensitivity of methylation detection by fragmentomics analysis, we performed computer simulation analysis to study how the sequencing depth would affect the detection of target molecules. We assumed that the plasma DNA contains the cfDNA molecules derived from the target tissue (e.g. the liver) with a fractional concentration $\boldsymbol{f}$ and the background DNA mainly of hematopoietic with a fractional concentration **(1-*f*)**. CGN and NCG end motif frequencies were assumed to be a linear combination of the target tissue and background DNA end motif distributions which were weighted by their corresponding fractional concentrations.

CGN and NCG end motif occurrences derived from the target tissue (i.e., $\boldsymbol{O(CGN)}_{\textbf{target tissue}}$ and $\boldsymbol{O(NCG)}_{\textbf{target tissue}}$) are assumed to follow the binomial distributions, being governed by a fractional concentration $\boldsymbol{f}$, the probability of methylation $\boldsymbol{p}$, the sequencing depth $\boldsymbol{d}$, as well as cleavage rate $\boldsymbol{r}$, as shown below:

$$\boldsymbol{O(CGN)_{target\ tissue} \sim Binom\big(d \times f \times p, r(CGN)\big),}\ \textbf{(1)}$$

$$\boldsymbol{O(NCG)_{target\ tissue} \sim Binom\big(d \times f \times p, r(NCG)\big),}\ \textbf{(2)}$$

where "Binom" represented the binomial distribution. Similarly, we could model CGN and NCG end motif occurrences derived from the background DNA (i.e., $\boldsymbol{O(CGN)}_{\textbf{background}}$ and $\boldsymbol{O(NCG)}_{\textbf{background}}$), based on the assumption of the binomial distributions.

We simulated $\boldsymbol{f}$ with the values of 10%, 20%, 50%, and 100%, respectively. For the methylated CpG sites, the probability of methylation $\boldsymbol{p}$ was assumed to be 0.95 whereas the unmethylated CpG sites were assumed to be 0.05. The cleavage rates ***(r)*** from methylated and unmethylated CpG sites were deduced from healthy control samples. In this simulation, the values of ***r(CGN)*** in methylated and unmethylated CpG sites were 0.067 and 0.029, respectively. The values of ***r(NCG)*** in methylated and unmethylated CpG sites were 0.0147 and 0.0269, respectively. We simulated CGN and NCG end motifs according to the binomial distributions mentioned above by varying the sequencing depth from 10X to 3000X. Hence, the CGN/NCG motif ratios could be calculated for those cfDNA molecules derived from the simulated tissue-specific methylated CpG sites and background unmethylated CpG sites, respectively. We further determined the area under the receiver operating characteristic curve (AUC) in the detection of tissue-specific cleavage signals related to the CpG methylation.

SVM model

The SVM model built with the frequency of each 5' CG-containing end motif (i.e., ACG, CCG, GCG, TCG, CGA, CGC, CGG, and CGT) was based on the leave-one-out strategy using R package (e1071). A 10-fold cross-validation was used to determine the best parameters used in the model.

NPC screening

272 and 65 EBV DNA positive individuals without NPC (non-NPC) and with NPC obtained from a previous report were divided into the training and testing sets (*SI Appendix,* Fig. S5A and B). In the training set, 31

NPC patients and 230 non-NPC individuals were classified by the previous method (1) based on EBV DNA proportion and EBV DNA size ratio (*SI Appendix, Fig*. S5B). In the testing set, 34 NPC cases and 42 non-NPC cases were not able to be differentiated according to the previous method based on EBV DNA proportion and EBV DNA size ratio (*SI Appendix,* Fig. S5B). In the training set, we pooled together all sequenced reads of EBV DNA in plasma DNA from non-NPC individuals and NPC patients into dataset A and dataset B, respectively. Based on these two datasets, the adjusted CGN/NCG motif ratio was calculated for each CpG site as follows:

$$\textbf{Adjusted CGN/NCG motif ratio} = \frac{\textbf{No. of 5}'\textbf{ CGN end motif}}{\textbf{No. of 5}'\textbf{ CGN end motif} + \textbf{No. of 5}'\textbf{ NCG end motif}}$$

1,425 CpG sites in the EBV genome were identified based on their adjusted CGN/NCG motif ratios fulfilling the criterion where the CGN/NCG motif ratios in dataset B were at least 20% higher than dataset A in the training set. The adjusted CGN/NCG motif ratio from those selected CpGs was calculated for each individual in the training dataset, and a cut-off was determined by using the lowest value of NPC cases in this dataset (*SI Appendix,* Fig. S5C). For the testing dataset comprising 42 non-NPC individuals and 34 NPC cases which were indistinguishable based on the previous method (1), the use of adjusted CGN/NCG motif ratio allowed additional 14 non-NPC cases to be excluded from the diagnostic conclusion of NPC, leading to an improved positive predictive value (i.e., 26.8%).

CNN model

The optimal parameters of the CNN model were determined when the overall prediction error between the output scores calculated by the sigmoid function and desired target output (binary values: 0 or 1) reached a minimum by iteratively updating model parameters. The overall prediction error was measured by the sigmoid cross-entropy loss function in the deep learning algorithm (https://pytorch.org/).

The CNN model used in this study made use of two one-dimensional (1D)-convolutional layers, each having 64 filters with a kernel size of 4. The activation function of the rectified linear unit (ReLU) was used for those convolutional layers. A batch normalization layer was applied subsequently, followed by a dropout layer with a dropout rate of 0.5. A flattened layer was further added, followed by a fully connected layer comprising 128 neurons with the use of the ReLU activation function. The output layer with one neuron was finally applied with a sigmoid activation function to yield the probabilistic score for a CpG site of being methylated (i.e., methylation score). The batch size was set to 64. The program for the CNN model was implemented based on the PyTorch machine learning framework (https://pytorch.org/). The model parameters learned from the training datasets were used to analyze the testing dataset to output a probabilistic score (i.e., the methylation score), indicating the likelihood of a CpG site being hypermethylated or hypomethylated.

The CpG sites used in the CNN modeling analysis were required to fulfill two criteria as below: (1) a CpG site needed to be covered by at least 50 sequenced molecules; (2) among DNA fragments mapped to the cleavage measurement window, there were at least 10 molecules ending with CGN or NCG motifs. Only the CpGs with their methylation index over 70% or below 30% were included for testing the feasibility of CNN model based methylation analysis at single CpG resolution. We randomly selected 91,212 CpGs, consisting of 45,606 hypermethylated and 45,606 hypomethylated CpGs, respectively. We used 70% and 15% of the prepared dataset to train and validate the model during the determination of model parameters. The remaining data (15%) which was not touched in the training process was used to test the model performance during generalization.

SI References

1. W. K. J. Lam *et al.*, Sequencing-based counting and size profiling of plasma Epstein-Barr virus DNA enhance population screening of nasopharyngeal carcinoma. *Proc Natl Acad Sci U S A* **115**, E5115-E5124 (2018).
2. P. Jiang *et al.*, Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov* **10**, 664-673 (2020).
3. W. Gai *et al.*, Liver- and colon-specific DNA methylation markers in plasma for investigation of colorectal cancers with or without liver metastases. *Clin Chem* **64**, 1239-1249 (2018).
4. F. M. Lun *et al.*, Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. *Clin Chem* **59**, 1583-1594 (2013).
5. S. C. Ding *et al.*, Jagged ends on multinucleosomal cell-free DNA serve as a biomarker for nuclease activity and systemic lupus erythematosus. *Clin Chem* 10.1093/clinchem/hvac050 (2022).
6. W. K. J. Lam *et al.*, Methylation analysis of plasma DNA informs etiologies of Epstein-Barr virus-associated diseases. *Nat Commun* **10**, 3256 (2019).
7. O. Y. O. Tse *et al.*, Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc Natl Acad Sci U S A* **118**, e2019768118 (2021).
8. K. Sun *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* **112**, E5503-5512 (2015).
9. K. C. Chan *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* **110**, 18761-18768 (2013).