PNAS
www.pnas.org

Supplementary Information for

Genomewide detection of cytosine methylation by single molecule real-time sequencing

O.Y. Olivia Tse [a,b,1], Peiyong Jiang[a,b,1], Suk Hang Cheng[a,b,1], Wenlei Peng[a,b], Huimin Shang[a,b], John Wong[c], Stephen L. Chan[d,e], Liona C.Y. Poon[f], Tak Y. Leung[f], K.C. Allen Chan[a,b,e], Rossa W.K. Chiu[a,b], Y.M. Dennis Lo[a,b,e,2]

[2]To whom correspondence may be addressed: E-mail: loym@cuhk.edu.hk.

This PDF file includes:

1. **Figures S1 to S13**
2. **Tables S1 and S2**
3. **Methods and Materials**
    a. Implementation of Hidden Markov Model (HMM)
    b. Sample Recruitment and Processing
    c. SMRTbell™ Template Library Preparation and Sequencing
    d. Procedures for Training and Testing the HK Model

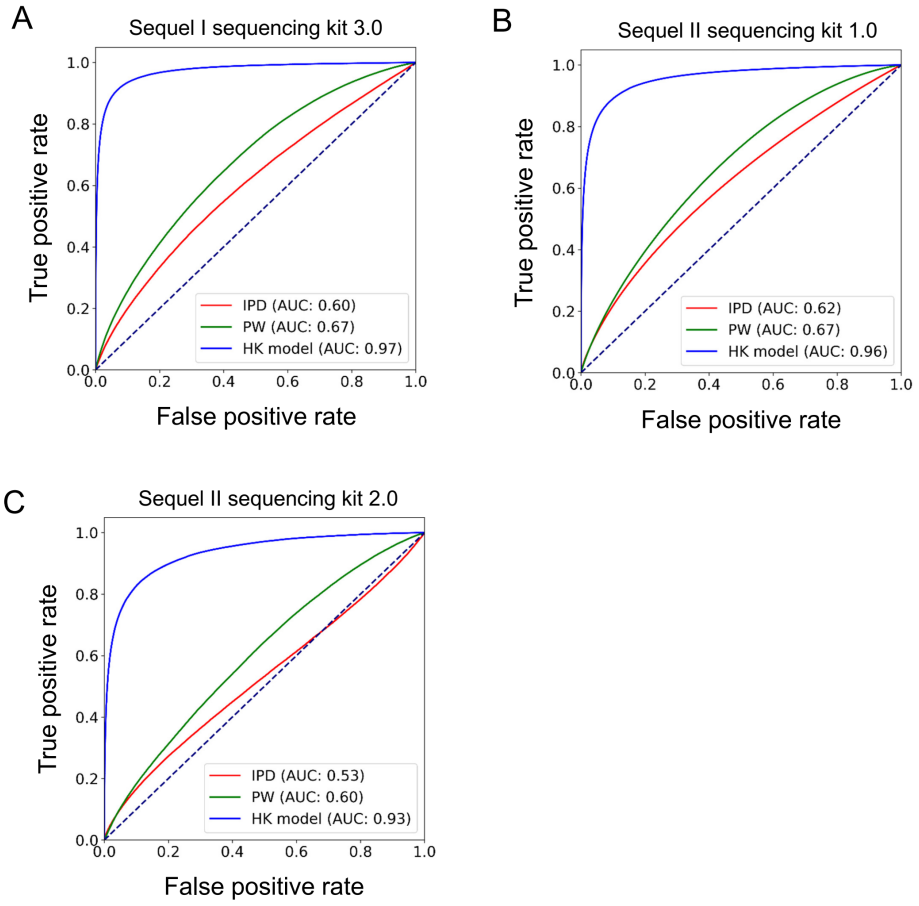

Fig. S1. Performance comparison for 5mC detection using HK model, IPD and PW at cytosines within CpG sites. (A) Sequel I sequencing kit 3.0. (B) Sequel II sequencing kit 1.0. (C) Sequel sequencing kit 2.0.

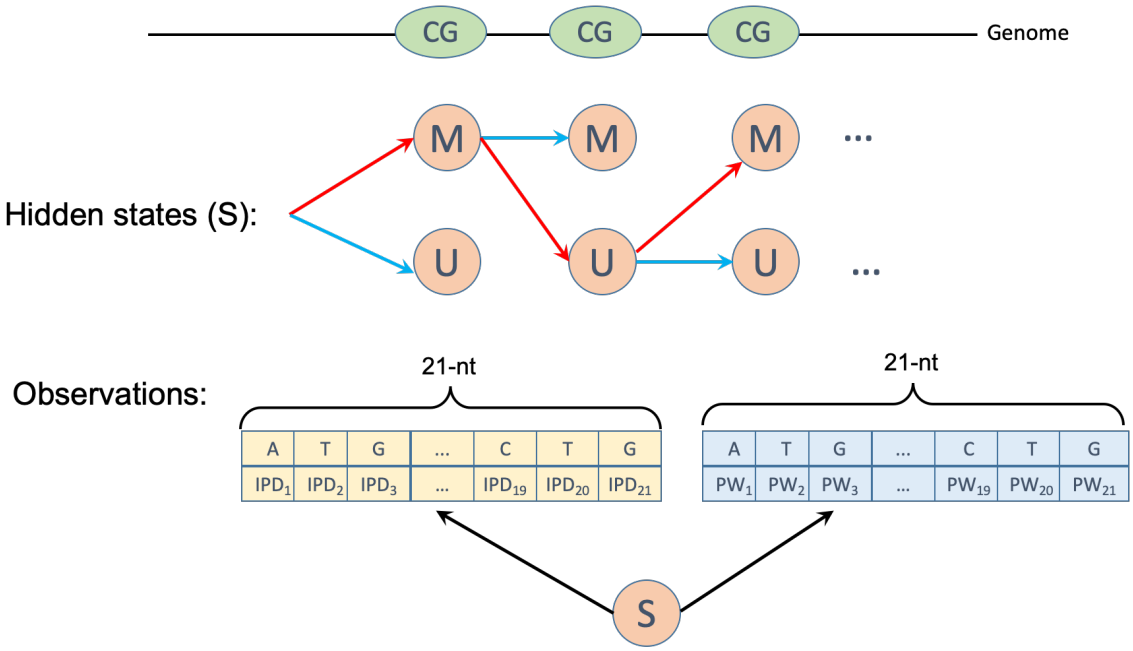

Fig. S2. Conceptual illustration for the HMM model for methylation analysis. Red arrows indicate the most likely chain of hidden states for explaining the observations (i.e. IPDs and PWs).

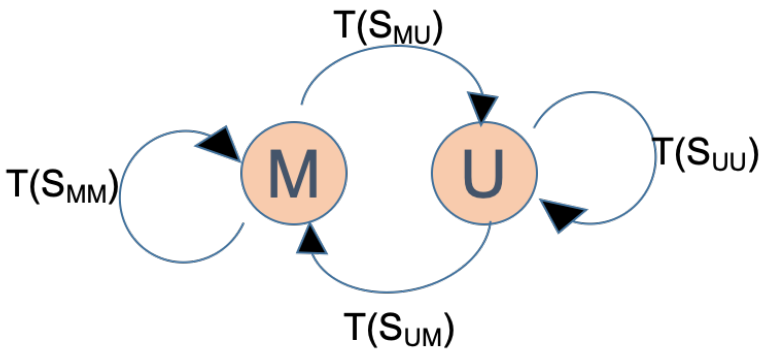

Fig. S3. The illustration for transition probabilities of the hidden states in the HMM.

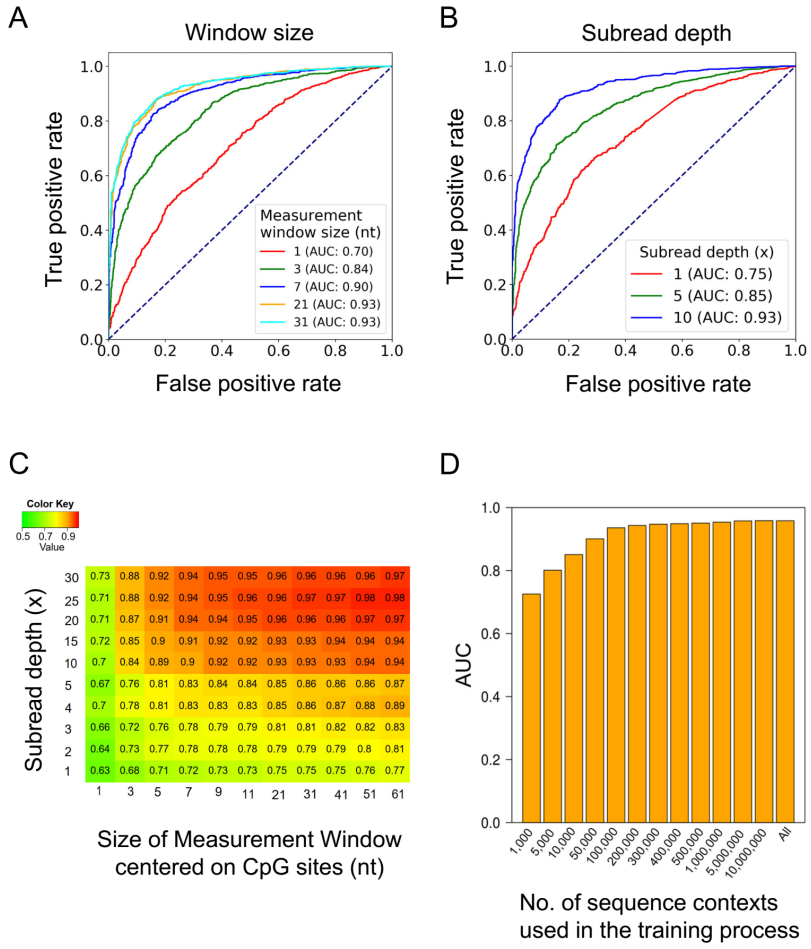

Fig. S4. Effect of window size and subread depth on the performance of 5mC detection. (A) ROC curves for HK models based on various measurement window sizes including 1, 3, 7, 21 and 31 nt, when the subread depth was at 10x. (B) ROC curves for HK models based on various subread depths with 1, 5, and 10x, when the measurement window was at 21 nt in size. (C) Heatmap for AUC values varied according to the different subread depths and measurement window sizes.

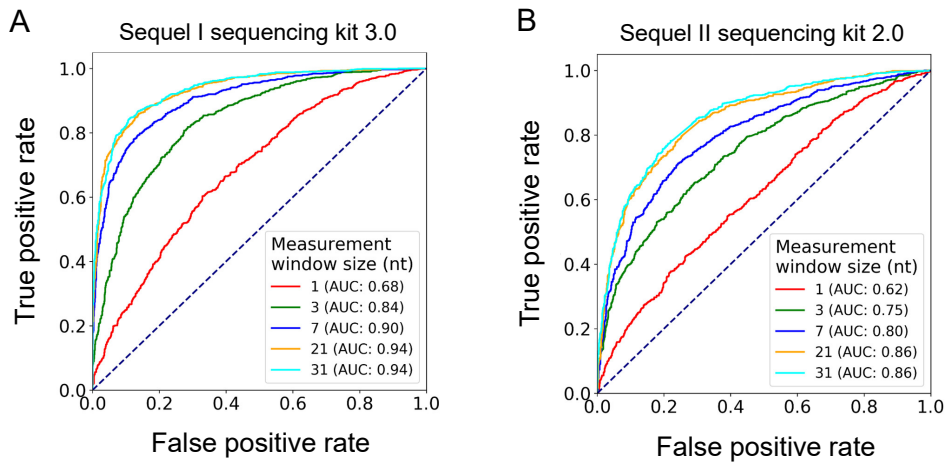

Fig. S5. Effect of window size on the performance of 5mC detection using Sequel I sequencing kit 3.0 (A) and Sequel II sequencing kit 2.0 (B).

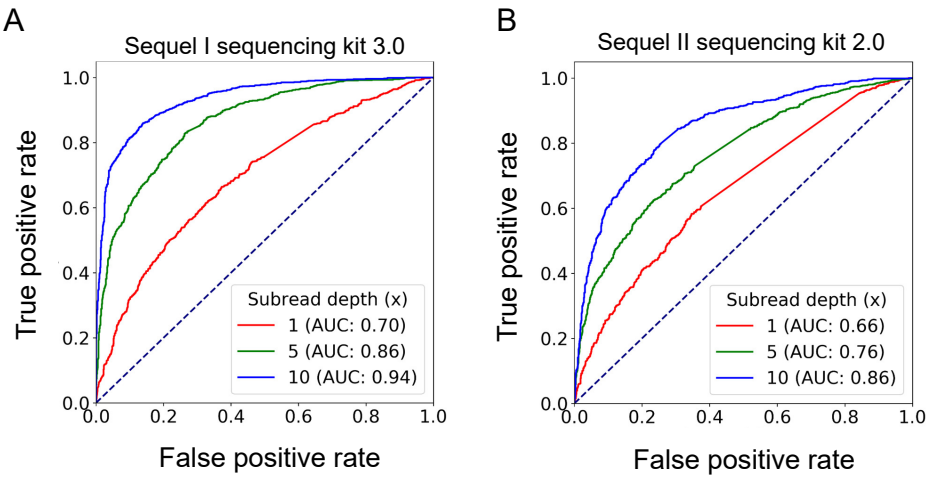

Fig. S6. Effect of subread depth on the performance of 5mC detection using Sequel I sequencing kit 3.0 (A) and Sequel II sequencing kit 2.0 (B).

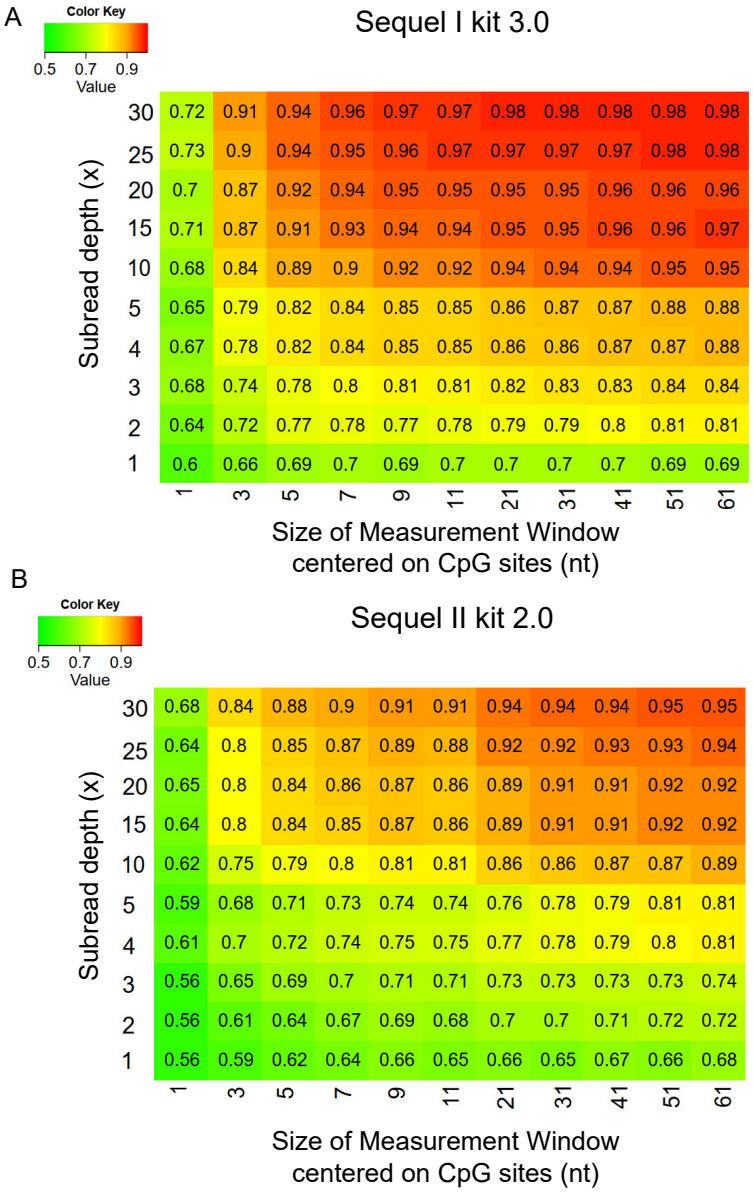

Fig. S7. Effect of measurement window size and subread depth on the performance of 5mC detection using the Sequel I sequencing kit 3.0 (A) and Sequel II sequencing kit 2.0 (B).

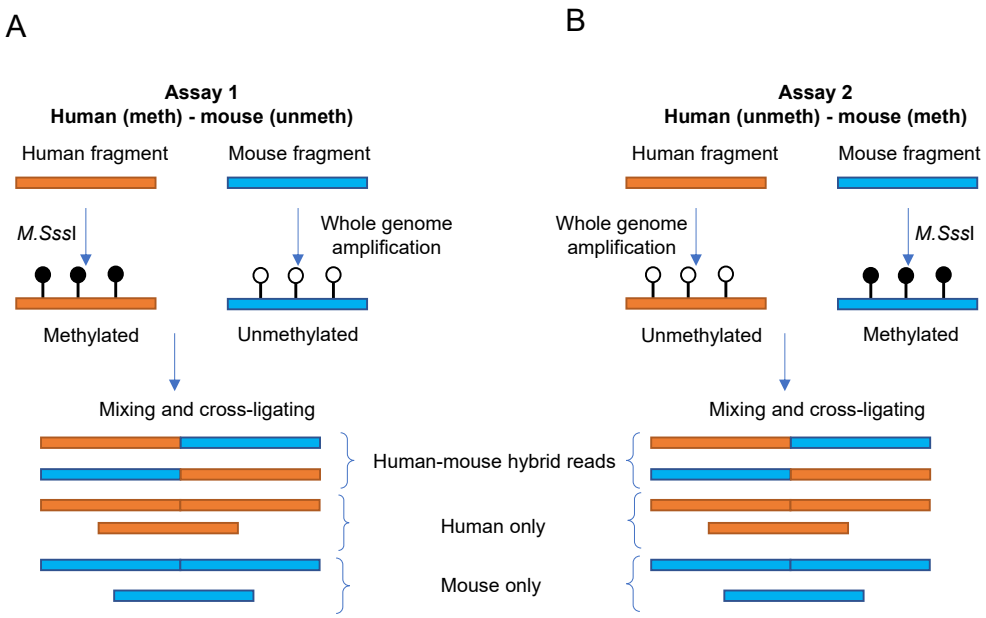

Fig. S8. Design for human-mouse hybrid fragments. (A) Assay for generating Human (meth) – mouse (unmeth) dataset. (B) Assay for generating the Human (unmeth) – mouse (meth) dataset. 'meth': methylated. 'unmeth': unmethylated.

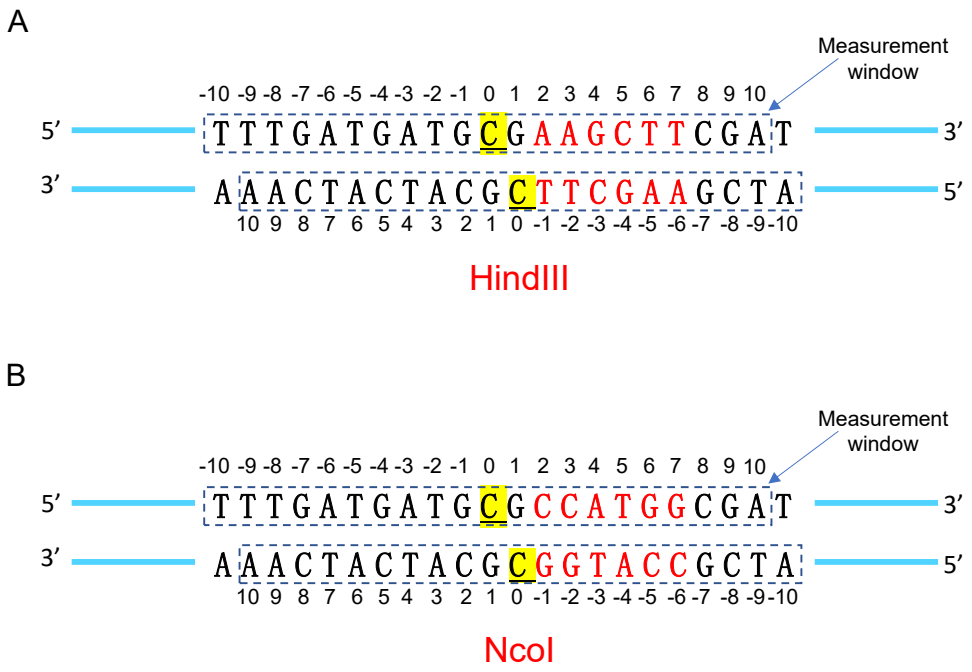

Fig. S9. Illustration for the nucleotide composition surrounding the restriction enzymes for construction of the human-mouse hybrid DNA fragments. (A) An example of base composition surrounding the restriction enzyme HindIII. (B) An example of base composition surrounding the restriction enzyme NcoI. The bases shown in red correspond to the restriction recognition sites. The underscored cytosine is the cytosine within the CpG site that is subjected to the interrogation of methylation status. The position of the underscored cytosine is annotated as 0. The other downstream and upstream positions relative to the interrogated cytosine are annotated as positive (1, 2, 3, …, 10) and negative integers (-1, -2, -3, …, -10).

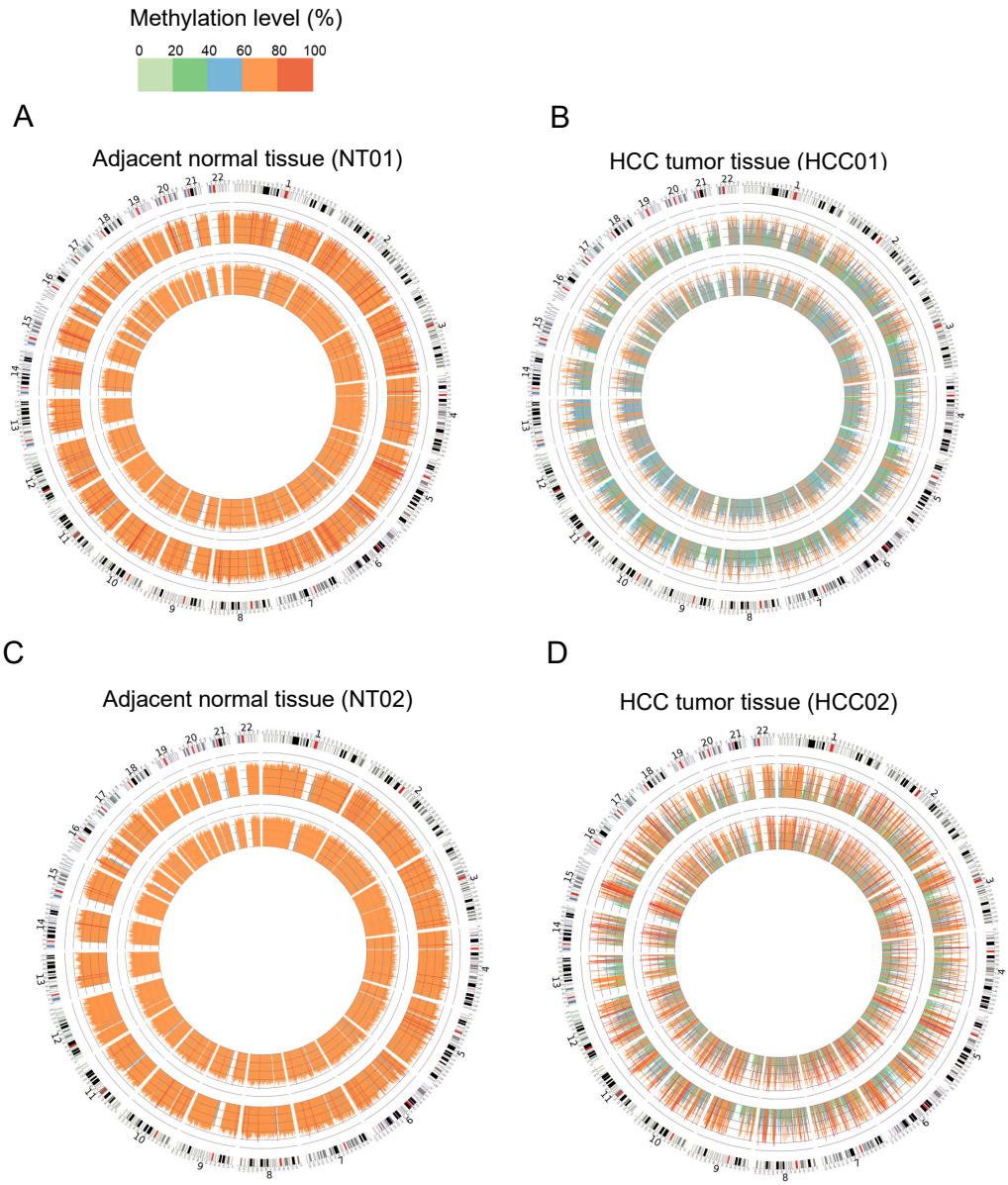

Fig. S10. Circos plots show methylation levels determined by the HK model (inner ring) and BS-seq (outer ring) across different 1-Mb regions of the human genome for adjacent nontumoral tissue (NT01) (A), HCC tumor tissue (HCC01) (B), adjacent nontumoral tissue (NT02) (C), HCC tumor tissue (HCC02) (D).

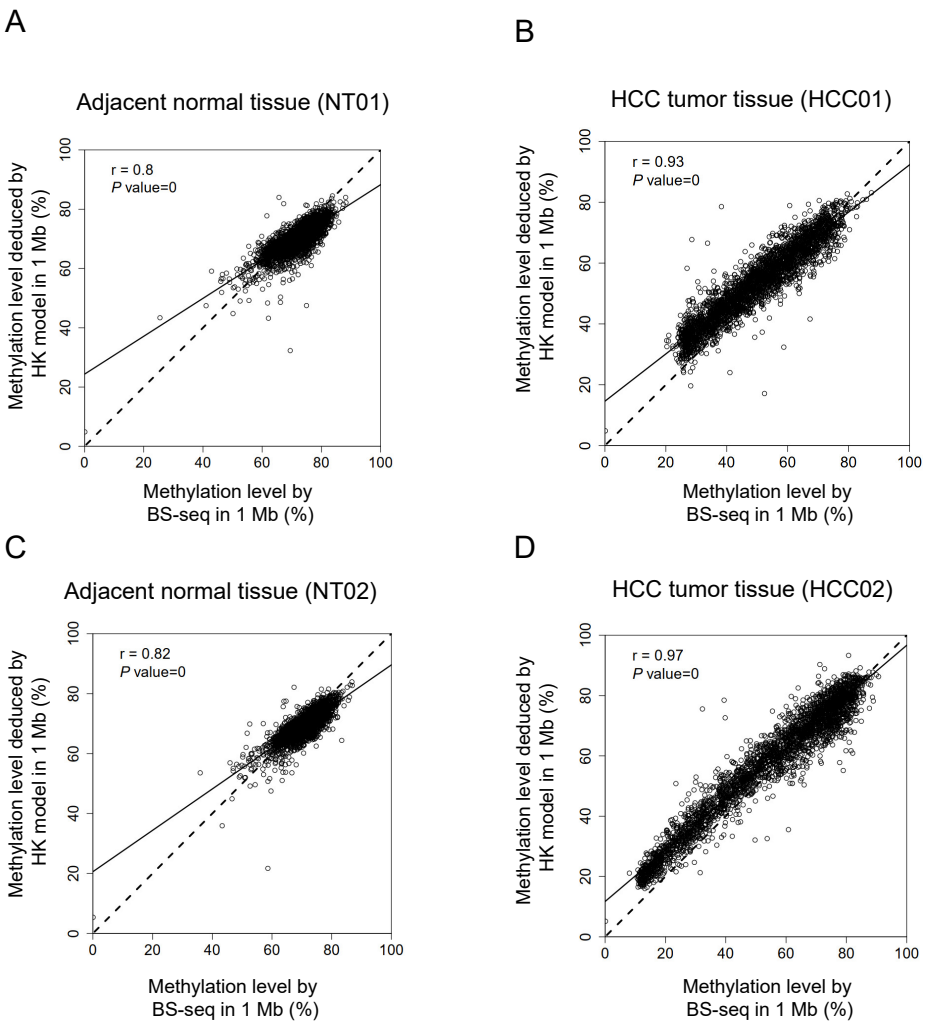

Fig. S11. Scatter plots show correlations of methylation level in each 1-Mb genomic region determined by the HK model and BS-seq for adjacent nontumoral tissue (NT01) (A), HCC tumor tissue (HCC01) (B), adjacent nontumoral tissue (NT02) (C), HCC tumor tissue (HCC02) (D).

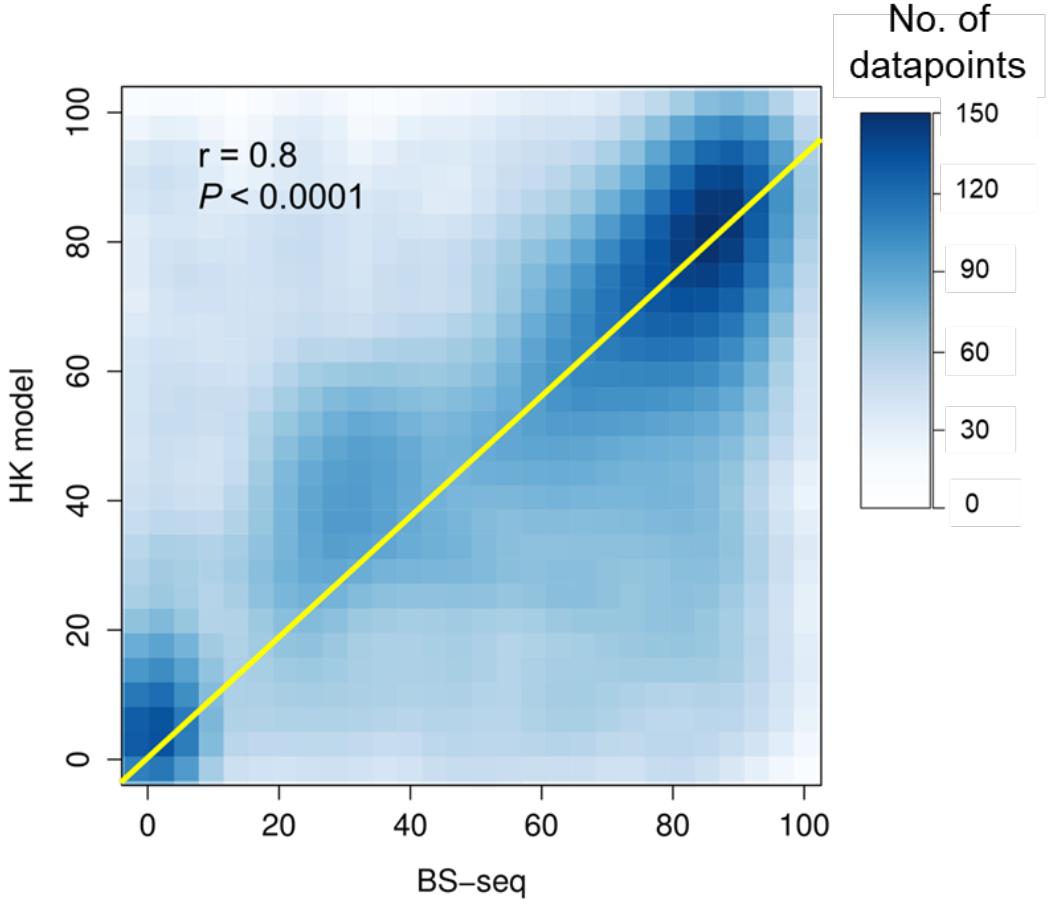

Fig. S12. Correlation of methylation levels between the HK model and BS-seq at single-base resolution.

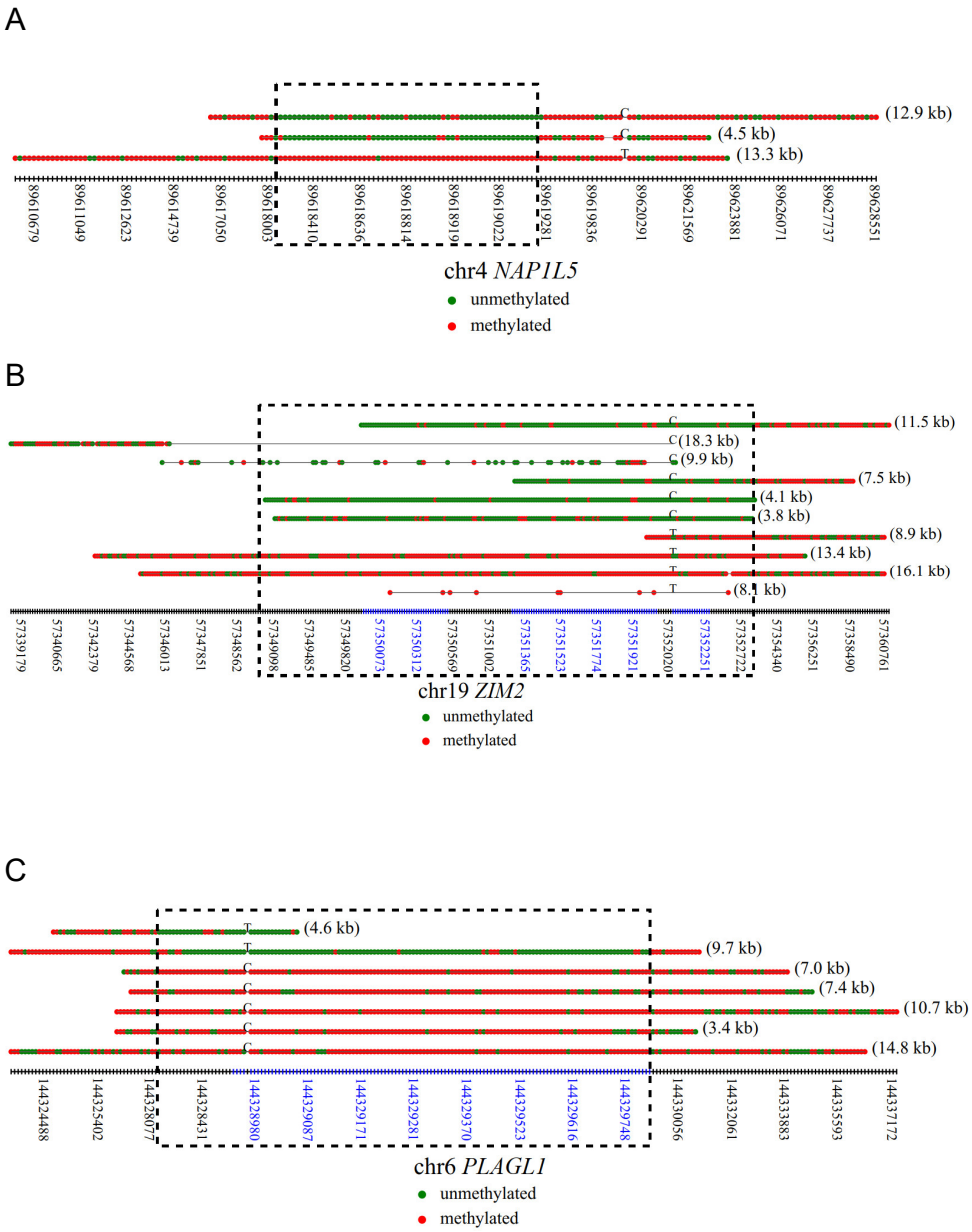

Fig. S13. Methylation patterns in a single molecule derived from imprinted regions including *NAP1L5* (A), *ZIM2* (B), and *PLAGL1* (C).

Table S1. Statistics for the numbers of sequence contexts used in different training and testing datasets.

Kit types	Dataset types	No. of CpG sites			No. of sequence context surrounding CpG sites			% of total no. of sequence context surrounding CpG sites in the human genome			AUC
		Total	WGA	M.SssI	Total	WGA	M.SssI	Total	WGA	M.SssI	
Sequel I kit v3	Training	656,466	328,233	328,233	561,855	287,390	288,831	2.71	1.38	1.39	0.97
	Testing	656,466	328,233	328,233	559,329	285,986	287,680	2.69	1.38	1.39	0.97
Sequel II kit v1	Training	11,272,552	5,636,276	5,636,276	6,788,277	3,858,728	3,844,729	32.70	18.59	18.52	0.96
	Testing	11,272,554	5,636,277	5,636,277	6,780,477	3,856,099	3,833,107	32.66	18.57	18.46	0.96
Sequel II kit v2	Training	325,780	162,890	162,890	270,780	136,379	141,160	1.30	0.66	0.68	0.94
	Testing	325,782	162,891	162,891	271,769	137,120	141,367	1.31	0.66	0.68	0.93

Table S2. Summary of sequencing kits and throughput used for different samples.

Sample names	Sample descriptions	No. of circular consensus sequences (CCSs)	Sequencing kits
M02	Placenta sample treated by CpG methyltransferase M.SssI	1,408,345	Sequel II sequencing kit 1.0
W02	Placenta sample with whole genome amplification	1,918,619	
BC01	Buffy coat	11,563,500	
BC02	Buffy coat	1,180,644	
BC03	Buffy coat	3,823,329	
BC04	Buffy coat	3,800,392	
BC05	Buffy coat	2,295,002	
HepG2	HCC cell line	5,139,689	
NT01	nontumoral tissue adjacent to HCC01 tumor tissue	4,702,130	
NT02	nontumoral tissue adjacent to HCC02 tumor tissue	4,471,370	
HCC01	HCC tumor tissue sample	1,147,985	
HCC02	HCC tumor tissue sample	2,157,196	
PL01	Placenta	1,573,540	
H01	Human methylated ligated with mouse unmethylated	3,476,786	
H02	Human unmethylated ligated with mouse methylated	3,901,995	
M01	Placenta sample treated by CpG methyltransferase M.SssI	419,713	Sequel I sequencing kit 3.0
W01	Placenta sample with whole genome amplification	444,079	
M03	Placenta sample treated by CpG methyltransferase M.SssI	92,673	Sequel II sequencing kit 2.0
W03	Placenta sample with whole genome amplification	155,163	

Methods and Materials

a. Implementation of Hidden Markov Model (HMM)

We attempted to use a Hidden Markov Model (HMM) for classifying the methylated and unmethylated cytosines at CpG sites. Briefly, we assumed that the methylation status of CpG sites along the genome followed a Markov chain. In this model, there were two hidden states for each CpG dinucleotide, namely methylation (M) and unmethylation (U). We used the first-order HMM, assuming that the hidden methylation state at each CpG depended only on the methylation state of the most preceding CpG. The possible observations at a given CpG site included interpulse durations (IPDs) and pulse widths (PWs) across individual nucleotides (i.e. sequence context) within a measurement window (21 nt). The empirical distributions of IPD and PW at each nucleotide surrounding a CpG site in question were determined using datasets generated by M.SssI-treated DNA and amplified DNA.

Figure S10 showed that conceptual model structure of HMM. The hidden state (S) had two possible values for each CpG site (i.e. $\boldsymbol{M}$ and $\boldsymbol{U}$). The observations of kinetic values including IPD and PW within a 21-nt measurement window were assumed to be associated with the hidden state ($\boldsymbol{S}$) at a given sequence context ($\boldsymbol{C}$).

To implement the HMM analysis, we need to set initial probabilities, $\boldsymbol{A_k}$, where $\boldsymbol{k} \in \boldsymbol{(M, U)}$. $\boldsymbol{A_k}$ has the below values in the present HMM.

$$A_M = 0.5,$$
$$A_U = 0.5.$$

The transition probabilities, $\boldsymbol{T}(\boldsymbol{S_{ij}})$, were illustrated in Figure E, where $\boldsymbol{i, j} \in \boldsymbol{(M, U)}$.

As the transition probability $\boldsymbol{T}$ would depend on the nucleotide distance between two CpG sites, we let $\boldsymbol{T}(\boldsymbol{S_{ij}})$ follow the below distribution:

$$T(S_{ij} \mid d) \begin{cases} f(S_{ij} \mid d) & d \le 200 \text{ nt} \\ 0.5 & d > 200 \text{ nt} \end{cases}$$

where $\boldsymbol{d}$ is the nucleotide distance between two consecutive CpG sites; $\boldsymbol{f}$ is the frequency of the combinations of methylation states between two CpG sites, which was empirically deduced from tissue samples with a high sequenced depth using BS-seq (75x2 paired-end reads).

For the emission probabilities, there are two types of observations, IPD and PW metrics.

The distribution of IPD is denoted by $\boldsymbol{\phi}$:

$$\boldsymbol{\phi}\ (\mathbf{IPD} \mid \boldsymbol{S},\ \boldsymbol{i},\ \boldsymbol{C}),$$

where *'$\boldsymbol{S}$'* is the hidden state having two possible values (i.e. $\boldsymbol{M}$, $\boldsymbol{U}$);

'$\boldsymbol{i}$' is a relative position in a measurement window ranging from -10 to +10;

'$\boldsymbol{C}$' is the sequence context at the position, where $\boldsymbol{C} \in (\boldsymbol{A},\ \boldsymbol{C},\ \boldsymbol{G},\ \boldsymbol{T})$.

The distribution of PW is denoted by $\boldsymbol{\Psi}$:

$$\boldsymbol{\Psi}\ (\boldsymbol{PW} \mid \boldsymbol{S}, \boldsymbol{i}, \boldsymbol{c}),$$

where *'$\boldsymbol{S}$'* is the hidden state having two possible values (i.e. $\boldsymbol{M}$, $\boldsymbol{U}$);

'i' is a relative position in the measurement window ranging from -10 to +10;

'$\boldsymbol{c}$' is the sequence context at the position, where $\boldsymbol{C} \in (\boldsymbol{A},\ \boldsymbol{C},\ \boldsymbol{G},\ \boldsymbol{T})$.

The most likely chain of hidden states (i.e. methylation patterns) was determined by the Viterbi algorithm (1) as below:

$$\boldsymbol{V_{1,k}} = \boldsymbol{P(O_1|k)} \cdot \boldsymbol{A_k}$$

$$\boldsymbol{V_{i,k}} = \mathbf{max}_{\in(\boldsymbol{M},\boldsymbol{U})}\ (\boldsymbol{P(O_i|k)} \cdot \boldsymbol{T_{j,k}} \cdot \boldsymbol{V_{i-1,j}})$$

Where $\boldsymbol{V_{1,k}}$ is the probability of methylation state at the first CpG site with $\boldsymbol{k}$ as its initial state; $\boldsymbol{V_{i,k}}$ *is the probablity of the most probable* methylation *state* chain (i.e. methylation patterns along a DNA sequence) *for the first* $\boldsymbol{i}$ CpG positions with $\boldsymbol{k}$ *as its final* methylation *state*; $\boldsymbol{O_i}$ is the observations regarding IPD and PW metrics within a measurement window related to the CpG $\boldsymbol{i}$.

$$P(O_i|k) = \prod_{m=-10}^{m=+10} \phi\ (\text{IPD} \mid k,\ m,\ C) \cdot \Psi\ (\text{PW} \mid k, \text{m, c})$$

Thus, $\log(V_{i,k}) = \max_{\in(M,\,U)} \sum_{m=-10}^{m=+10} \log \phi\ (\text{IPD} \mid k,\ m,\ C) + \log \Psi\ (\text{PW} \mid k, \text{m, c}) + \log T_{j,k} + \log V_{i-1,j}$

i.e. $\log(V_{i,k}) = \max_{\in(M,\,U)} \sum_{m=-10}^{m=+10} \log \phi\ (\text{IPD} \mid k,\ m,\ C) + \log \Psi\ (\text{PW} \mid k, \text{m, c}) + \log \text{T}(S_{j,k} \mid \text{d}) + \log V_{i-1,j}$

b. Sample Recruitment and Processing

Peripheral blood samples from all participants were collected into EDTA-containing tubes, and buffy coat was isolated after centrifugation. The tumor tissues and their paired adjacent nontumoral tissues of the HCC patients were obtained during their cancer resection surgery. Placenta tissue samples were obtained from pregnant women after delivery. HepG2 hepatocellular carcinoma cell line was purchased from the American Tissue Culture Collection (ATCC; Catalog no. HB-8065; Manassas, VA). Genomic DNA was extracted from buffy coat samples according to the protocol of the QIAamp DNA Blood Mini Kit. Tissue and cell line DNA were extracted with the QIAamp DNA Mini Kit (Qiagen).

c. SMRTbell™ Template Library Preparation and Sequencing

The sheared DNA molecules were subjected to single molecule real-time (SMRT) sequencing template construction using a SMRTbell Template Prep Kit 1.0 and Expression Prep Kit 2.0 (PacBio). Sequencing primer annealing and polymerase binding conditions were calculated with the SMRT Link v8.0 software (PacBio). Briefly, sequencing primer was annealed to the sequencing template, and then a polymerase was bound to templates using a Sequel I Binding Kit 3.0 or Sequel II Binding and Internal Control Kit 2.0 (PacBio). Sequencing was performed on a Sequel SMRT Cell 1M or Sequel II SMRT Cell 8M. Sequencing movies were collected on the Sequel systems for 20 hours with a Sequel Sequencing Kit 3.0 (PacBio), or for 30 hours with a Sequel II Sequencing Kit 1.0 or 2.0.

Different versions of the reagent kit used for SMRT-seq would contain different DNA polymerases with different kinetic properties. The kinetic properties of a DNA polymerase would play an important role in the performance of 5mC detection. The detailed information about reagent kits used in this study were listed below according to the manufacturer's instructions (Pacific Biosciences).

(1) The Sequel Binding Kit 3.0 consists of reagents, enabling the prepared DNA template libraries to be bound to the Sequel Polymerase 3.0. The resultant DNA polymerase/template complex will be suited for sequencing on the Sequel System. The Sequel Polymerase 3.0 should be used only with Sequel Sequencing Kit 3.0.

(2) The Sequel II Binding Kit 1.0 consists of reagents, enabling the prepared DNA template libraries to be bound to the Sequel II Polymerase 1.0 for sequencing on the Sequel II System. Sequel II Polymerase 1.0 should be used only with Sequel II Sequencing Kit 1.0.

(3) The Sequel II Binding Kit 2.0 consists of reagents, enabling the prepared DNA template libraries to be bound to the Sequel II Polymerase 2.0 on the Sequel II System. Sequel II Binding Kit 2.0 is to be used for all samples except for samples with short inserts < 3 kb such as amplicons. Sequel II Polymerase 2.0 should be used only with Sequel II Sequencing Kit 2.0.

d. Procedures for training and testing the HK model

The optimal parameters of the HK model were obtained when the overall prediction error between the output scores calculated by the sigmoid function and desired target outputs (binary values: 0 or 1) reached a minimum by iteratively adjusting model parameters. The overall prediction error was measured by the sigmoid cross-entropy loss function in deep learning algorithms (https://keras.io/). The model parameters learned from the training datasets were used for analyzing the testing dataset to output a probabilistic score (referred to as the methylation score in this study) which would indicate the likelihood of a CpG site being methylated. In practice, certain signal patterns might overlap between the methylated or unmethylated CpG datasets during the training process. Furthermore, the signal pattern in a test measurement window might have variations from those present in the training datasets. Hence, the methylation score output by the HK model was a continuous probabilistic score ranging from 0 to 1, instead of discrete binary values. For example, if the methylation score of a CpG site was 0.9, then applying a methylation score threshold of 0.5 would result in that CpG site being classified as methylated. In contrast, if the methylation score was 0.1, applying the same methylation score threshold of 0.5 would result in that site being classified as unmethylated.

Reference

1. Viterbi A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inf Theory* 13(2):260–269.