Legends

Supplementary Table S1.

Statistics of SMRT sequencing data production. Summary statistics of SMRT sequencing data collected in this study.

Supplementary Table S2.

Accuracy metrics improvement by excluding intermediate predictions. By excluding CpGs with intermediate prediction, the accuracy of binary prediction was improved. For example, our method achieved $>95\%$ sensitivity and precision when 7 % of CpGs excluded.

Supplementary Table S3.

The primers for nested PCR of the bisulfite treated blood DNA. The primers for nested PCR are shown alongside the sequence IDs that correspond to those in Supplementary Figure S5, the sequence names, and the target genomic regions. For each entry, the forward primers appear in the top row, and the reverse primers appear in the second row. The primers with circled ids (5, 7, 8, 9, 11, and 15) were able to amplified the regions.

Supplementary Table S4.

DNA methylation states of full-length LINE/L1 elements. According to the three classes of full-length LINE/L1 elements in L1Base, we examined DNA methylation states of LINE/L1 elements in each class.

Supplementary Figure S1.

The normal vector used for prediction. A. The normal vector $\boldsymbol{\beta}$ used for prediction with P6-C4 reagent. We calculated $\boldsymbol{\beta}$ as follows. Firstly, we classified the CpGs on the scaffold 1 in the medaka Hd-rR genome (version 1) into methylated CpGs and unmethylated CpGs according to bisulfite sequencing data. Next, for each CpG site, we calculate the IPD ratio profiles as the 21-dimensional vectors based on SMRT sequencing kinetics data. Then, using LDA (Linear Discriminant Analysis), we tried to find the best hyperplane that could separate these IPD ratio profiles into each class, namely, methylated or unmethylated. The normal vector of this hyperplane is denoted by $\boldsymbol{\beta}$. **B.** The average IPDR profiles around unmethylated and methylated CpG sites. The x-axis shows the positions within 10 bp of the focal CpG site at the position represented by 0. The y-axis indicates IPDR values. The red- and blue-colored box plots at each position show the distributions of IPDR values around unmethylated and methylated CpG sites, respectively. The bottom, middle and top of each box plot indicate the first, second, and third quartiles, respectively, of the distribution. **C.** An example in which both our method and bisulfite sequencing are consistent in showing unmethylation in gene promoters. The tracks are similar to those in Figure 1B. **D.** The normal vector $\boldsymbol{\beta}$ used for prediction with P4-C2/C2-C2 reagent.

Supplementary Figure S2.

Accuracy metrics on the chromosome 1 of the medaka Hd-rR genome (version 2). A-C. Matthew's correlation coefficient **(A)**, sensitivity **(B)**, and precision **(C)** as a function of the intercept of the hyperplane γ, on the chromosome 1 in the medaka genome (version 2) with a 29.9-fold mapped read coverage. Matthew's correlation coefficient represents an overall accuracy of our prediction. The differently colored curves correspond to the different lower bound of number of CpG sites, denoted by b, that was used for the prediction. Our prediction achieved 93.0% sensitivity and 94.9% precision at $b = 35$ and $\gamma = -0.526$. Or sensitivity (93.67%) and precision (93.88%) are close to each other when $b = 35$ and $\gamma = -0.540$.

Supplementary Figure S3.

Sensitivity and precision of predicting unmethylated regions with $\geq b$ CpG sites for a variety of read coverages. We continue to use b to denote a lower bound of the number of CpG sites in a region. For $b = 30, 35, 40, 45, 50$, we plot the sensitivity and precision curves when the read coverage is 20% of 29.9x (**A**), 40% of 29.9x (**B**), 60% of 29.9x (**C**), 80% of 29.9x (**D**), and 29.9x (**E**). The sensitivity and precision were evaluated on the chromosome 1 of the medaka Hd-rR genome (version 2). For better prediction with a smaller coverage, a wider window was favored. Precisely, setting b to 50 outperforms the other values for coverages, 20% and 40%, but it becomes inferior for 80% and 100%. In contrast, both sensitivity and precision increase for larger coverages, 80% and 100%, when b is set to smaller values, 35 and 40. In particular, Figure **E** shows that for coverage 100% (29.9x), setting b to 35 is better than other values of b. Figure **C** also highlights that even with a small coverage 60% of 29.9x, both sensitivity and precision are $\sim 90\%$ for $b = 45$. Figure **F** shows that the prediction is not accurate if each CpG site is treated independently (not as blocks). Figure **G** compares the performance with simplified beta (where the components for -7, +1, +3, +5$\sim$+10-th positions were truncated to 0) to that with the original full beta vector.

Supplementary Figure S4.

Handling intermediate methylation states. A. IPDR profiles of CpGs are represented as points in the feature space. Predictions are made using a decision hyperplane determined by its intercept γ, and individual CpGs are classified as methylated (blue) or unmethylated (red). **B.** Multiple predictions using a set of different intercept parameter values define the discrete methylation level (DML) on each CpG site. Specifically, after decomposing DNA into unmethylated and methylated regions for different intercept values of γ, we compute the ratio of methylated regions that cover each CpG site, and treat the ratio as the methylation level of the CpG site. **C.** DML (x-axis) and methylation level monitored by bisulfite sequencing (y-axis) in our medaka sample. The colors are based on the log of the number of CpG sites having corresponding DML value and bisulfite methylation level. These values were strongly correlated ($R = 0.884$) and the difference was within 0.25 for 92.0% of CpG sites. Most of the CpG sites were methylated because we observed CpG methylation in a genome-wide manner. **D.** DML (x-axis) correlated ($R = 0.732$) with the normalized beta values of BeadChip (y-axis) for the CpG sites in our human sample, and 75.4% of CpG sites are in concordance within 0.25. The majority of CpG sites are unmethylated, because most CpG sites on the BeadChip are designed on CpG islands. **E.** Scatterplot for methylation level monitored by bisulfite sequencing (x-axis) and DML (y-axis), on each CpG site, in the medaka sample.

Supplementary Figure S5.

Methylation analysis of selected regions for validation of our prediction. Of the 21 regions selected for validation of our method, 6 were amplified, and their Sanger sequencing reads were aligned to the target regions. In the alignments, the methylated (unconverted) CpGs are represented by the pink asterisks (*), and the unmethylated (converted)

CpGs by the blue number sign (#). We can assess the efficiency of bisulfite conversion and the quality of the alignment by looking at non-CpG C sites (CpHs) because Cs in CpHs are usually unmethylated and should always be converted to Ts (represented by the colons (:)). Thus unconverted CpHs, which are highlighted by the brown exclamation marks (!), indicate low quality regions. The solid lines represent the other types of matches.

Supplementary Figure S6.

Kernel PCA analysis of sequence feature and methylation state. The results of Kernel PCA analysis are shown for 4 selected classes of repetitive elements, AluSc **(A)**, LTR12E **(B)**, LTR26E **(C)**, and L2a **(D)**. We projected the repeat occurrences into the plane based on the distance metrics that we defined using the spectrum kernels and their top 2 principal components. The colors of the dots represent the methylation state of the repeat occurrences; namely, red indicates unmethylation and blue methylation. The arrows show the unmethylated occurrences that are clustered in terms of the sequence features.

Supplementary Figure S7.

Examples of unmethylated repeat occurrences in a unmethylation 'hot spot'. Three adjacent LTR1 elements were unmethylated in this region **(A)**, and a LTR12E element was located at a unmethylated bi-directional promoter region **(B)**. Both regions are on the p-arm of the chromosome 6. The arrows indicate the locations of LTR1 and LTR12E. From top to bottom, below the RefSeq gene track, black bars indicate unmethylated regions predicted from SMRT sequencing data using our method. Yellow and black bars show the methylation level and read coverage obtained from public bisulfite sequencing data, respectively, and blue boxes show unmethylated regions predicted from the bisulfite data. Green bars below indicate the alignability of short (100-bp) reads. The bottom rows shows repeat masker tracks and GC rate for every 5 bp window.

Supplementary Figure S8.

Two LINE insertions novel to hg19. We identified two LINE insertions by comparing a new assembly obtained from SMRT reads and the hg19 reference genome. The vertical arrows indicate the locations of the identified novel insertions. Specifically, one is aligned at 186,372,132 in Chromosome 3 with identity 99.02%, and the other at 137,014,775 bp in Chromosome 5 with identity 98.71%. From top to bottom, the tracks shown are RefSeq genes, DNase clusters, repeat masker masked regions, and GC rate for every 5 bp window.