Measuring the Landscape of CpG Methylation of Individual Repetitive Elements

Yuta Suzuki[1,5], Jonas Korlach[2], Stephen W. Turner[2], Tatsuya Tsukahara[3], Junko Taniguchi[1], Wei Qu[1], Kazuki Ichikawa[1], Jun Yoshimura[1], Hideaki Yurino[1], Yuji Takahashi[4], Jun Mitsui[4], Hiroyuki Ishiura[4], Shoji Tsuji[4], Hiroyuki Takeda[3], and Shinichi Morishita[1,5]

1. Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8583, Japan

2. Pacific Biosciences, 1380 Willow Rd. Menlo Park, CA 94025, USA

3. Department of Biological Sciences, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

4. Department of Neurology, Graduate School of Medicine, The University of Tokyo, Tokyo 113-8655, Japan

5. Corresponding authors (Yuta Suzuki ysuzuki@cb.k.u-tokyo.ac.jp, Shinichi Morishita moris@cb.k.u-tokyo.ac.jp)

ABSTRACT

Determining the methylation state of regions with high copy numbers is challenging for second-generation sequencing, because the read length is insufficient to map reads uniquely, especially when repetitive regions are long and nearly identical to each other. Single-molecule real-time (SMRT) sequencing is a promising method for observing such regions, because it is not vulnerable to GC bias, it performs long read lengths, and its kinetic information is sensitive to DNA modifications. We propose a novel algorithm that combines the kinetic information for neighboring CpG sites and increases the confidence in identifying the methylation states of those sites. Both the sensitivity and precision of our algorithm were $\sim$93.7% on CpG site basis for the genome of an inbred medaka (*Oryzias latipes*) strain within a practical read coverage of $\sim$30-fold. The method is quantitatively accurate because we observed a high correlation coefficient ($R = 0.884$) between our method and bisulfite sequencing, and 92.0% of CpG sites were in concordance within 0.25. Using this method, we characterized the landscape of the methylation status of repetitive elements, such as LINEs, in the human genome, thereby revealing the strong correlation between CpG density and unmethylation and detecting unmethylation hot spots of LTRs and LINEs. We could uncover the methylation states for nearly identical active transposons, two novel LINE insertions of identity $\sim$99% and length 6050 base pairs (bp) in the human genome, and sixteen *Tol2* elements of identity $>$99.8% and length 4682 bp in the medaka genome.

1 BACKGROUNDS

There has been a great deal of interest in identification of genome-wide epigenetic DNA modifications in recent years, because DNA modifications play an essential role in cellular and developmental processes (Weaver et al., 2004; Anway et al., 2005; Jirtle and Skinner, 2007; Miller, 2010; Zemach et al., 2010; Schmitz et al., 2011; Molaro et al., 2011; Smith et al., 2012; Qu et al., 2012). Some of human transposable elements, such as long interspersed nuclear elements (LINE), are reported to transpose actively within somatic cells along differentiation of neural tissues, and to be partly regulated by DNA methylation (Muotri et al., 2005, 2010). Each family of human transposable elements is in a variety of methylation statuses according to tissue type by looking at the mixture of methylation information on the consensus sequence of TEs in the same family (Xie et al., 2013). Many human diseases are also associated with DNA methylation states of transposable elements. In particular, unmethylation of repetitive elements, such as LINE-1 elements, has been related to some cancers (Wilson et al., 2007; Ross et al., 2010). Although only a few LINE-1 elements exhibit activity in the human genome (Beck et al., 2010), transpositions of these elements have been reported in various cancer genomes (Lee et al., 2012; Goodier, 2014), and importantly, it has been reported that transpositions are correlated with unmethylation of the promoter region of LINE-1 elements (Tubio et al., 2014). Therefore, it is essential to develop an experimental framework that can characterize the methylation state of repetitive elements in a genome-wide manner.

The advent of second-generation sequencing technology has increased the efficiency of the generation of precise genome-wide methylation maps at a single-base resolution using bisulfite treatment (Cokus et al., 2008; Lister et al., 2008; Meissner et al., 2008; Lister et al., 2009; Harris et al., 2010); however, these sequencing-based technologies have difficulty in characterizing the methylation status of CpGs in regions that are highly similar to other regions. Bisulfite-treated short reads from these regions often fail to map uniquely to their original positions; instead, they are likely to be aligned ambiguously to multiple genomic positions. Moreover, first/second-generation sequencing technology often fails to sequence DNA regions with a GC content >60% (Aird et al., 2011; Ross et al., 2013) and may exhibit bias against GC-rich regions. These inherent problems of second-generation sequencing may result in underrepresentation of methylation information on specific DNA regions, such as transposable elements and low-complexity repeat sequences (Lister et al., 2009; Harris et al., 2010; Bock et al., 2010; Gifford et al., 2013; Jiang et al., 2013). Especially, the younger and more active transposons are thought to retain higher fidelity and are therefore difficult to address using short reads.

In the PacBio RS sequencing system, DNA polymerase is used to perform single-molecule real-time (SMRT) sequencing (Korlach et al., 2008; Eid et al., 2009), and this system is capable of sequencing reads of an average length of >10 kb. SMRT sequencing is also able to sequence genomic regions with extremely high GC contents. A striking example is a previous report of the sequencing of a >2-kb region with a GC content of 100% (Loomis et al., 2012), indicating that SMRT sequencing is less vulnerable to sequence composition bias than is first/second-generation sequencing. SMRT sequencing of bisulfite-treated DNA fragments may allow identification of DNA methylation within long regions; however, this approach is not promising because bisulfite treatment divides DNA into short fragments $<$1000 bp (Miura et al., 2012).

Instead, we explored another advantage of SMRT sequencing to detect DNA modifications. In SMRT sequencing, we can observe the base sequence in a single DNA molecule as each corresponding nucleotide is incorporated using the time course of the fluorescence pulses. From this time course information, we can determine the inter-pulse duration (IPD), which is defined as the time interval separating the pulses of two neighboring bases. Importantly, the IPD of the same genomic position varies and has a significant and predictable response to DNA modifications due to the sensitivity of DNA polymerase kinetics to DNA modifications and damage (Flusberg et al., 2010).

Consequently, the IPD ratio (IPDR), the ratio of the average IPD in DNA templates with modifications to that in control templates, tends to be perturbed systematically, allowing identification of DNA modifications (Fig. 1A). Indeed, SMRT sequencing methods have been used to detect changes in 5-hydroxymethylcytosine (Flusberg et al., 2010), N4-methylcytosine (Clark et al., 2012), and N6-methylademine (Flusberg et al., 2010; Fang et al., 2012; Feng et al., 2013; Greer et al., 2015), as well as damaged DNA bases (Clark et al., 2011) in bacteria and mitochondria; however, estimation of 5-methylcytosine (5-mC) residues using low-coverage reads is prone to errors and requires extensive coverage at each position to clarify the base-wise 5-mC state and therefore becomes costly (Flusberg et al., 2010; Fang et al., 2012; Schadt et al., 2012). Clark *et al.* attempted to improve the detection of microbial 5-mC in the *Escherichia coli* and *Bacillus halodurans* genomes using Tet1-mediated oxidation to convert 5-mC into 5caC in SMRT reads of ~150x coverage per DNA strand (Clark et al., 2013). Therefore, kinetic information from low-coverage SMRT reads at a single CpG site is not reliable for predicting the methylation status.

In this study, we exploited the facts that unmethylated CpG dinucleotides are rare ($\sim$10%) in vertebrates and generally do not exist in isolation but often range over long unmethylated regions (Qu et al., 2012; Gifford et al., 2013; Eckhardt et al., 2006; Bock et al., 2008; Shoemaker et al., 2010; Nautiyal et al., 2010; Xie et al., 2013). Su *et al.* reported that the average length of unmethylated regions in five human cell types is $\sim$2 kb (Su et al., 2012b). Thus, estimating regions of unmethylated CpG sites is informative in most cases. Similarly, integrating kinetic information for many CpG sites in a long region can increase the confidence in detecting methylation when the status of those sites is correlated and shows promise for predicting the methylation status in a block using low-coverage SMRT reads. We examined the feasibility of the approach and present a novel computational algorithm that integrates SMRT sequencing kinetic data and determines the methylation statuses of CpG sites.

To demonstrate a possible application of our method, we investigated methylation status of individual occurrences of transposable elements in the human genome. Only a few studies have addressed this. In one study, the authors designed a microarray specifically to characterize changes in transposon methylation of cancer cells (Szpakowski et al., 2009), and other study utilized a bisulfite sequencing dataset (Su et al., 2012a). These methods, however, cannot observe the repetitive elements which have not diverged enough to be distinguished by short reads, or those which are novel insertions and simply absent from the reference genome. Therefore, we examined the possibility of determining the methylation statuses of highly similar occurrences of transposable elements in human and medaka fish (*Oryzias latipes*), which could be investigated only by using long reads.

2 RESULTS

2.1 Bisulfite data benchmark and SMRT sequencing

It is necessary to take into account allele-specific DNA methylation in the analysis of the methylomes of diploid genomes (Shoemaker et al., 2010; Chandler et al., 1987; Yamada et al., 2004; Zhang et al., 2009; Kerkel et al., 2008; Schilling et al., 2009; Hellman and Chess, 2010), because we may observe an intermediate DNA methylation level resulting from the mixture of different methylation states from two haplotypes (Lister et al., 2009; Deng et al., 2009). To assess the ability of SMRT sequencing to monitor the DNA methylation status, DNA extracted from a haploid cell line would serve as an ideal template, avoiding situations in which two alleles are differentially methylated. In nonhuman model organisms, inbred strains also provide a clean resource, because the two haplotypes are almost identical in sequence, suggesting that the methylation statuses of the two haplotypes may also match. Therefore, we used the medaka model system (Kasahara et al., 2007), because six medaka methylomes are available from early embryos, testes, and liver in two inbred strains (Qu et al., 2012) by way of Illumina bisulfite sequencing, which outperforms three other frequently used sequence-based methods in terms of the genome-wide percentage of CpGs covered (Harris et al., 2010). As CpG methylation status reference data, we used the testes methylome of the medaka Hd-rR inbred strain. In this dataset, most of the CpG sites in the medaka genome are either unmethylated or methylated, and methylation at non-CpG sites is very rare ($\sim$0.02%) (Qu et al., 2012), allowing us to focus on CpG sites only. We collected 31.06-fold coverage SMRT subreads from the testes of medaka Hd-rR (assuming an estimated genome size of 800 Mb) using P6-C4 reagents. We also collected

22.45-fold and 13.06-fold coverage SMRT reads from human peripheral blood of two Japanese individuals. Thus, we have 3 datasets in total, 1 for medaka and 2 for human. For sequencing two human samples, we employed the P6-C4 reagents and the P4-C2 or C2-C2 reagents, respectively. In total 2848641, 7279594, and 19115712 subreads were anchored to the medaka genome and the human genome, respectively. The mean mapped subread lengths were 8722 bases for medaka and 9254 and 2049 bases for 2 human samples (Supplemental Table 1).

2.2 Prediction of the methylation state from kinetic data

Figure 1A shows a schematic representation of the basic concept of our method. First, as a raw ingredient for prediction, we defined the IPDR profile of a CpG site as an array of IPDR measurements of 21 bp surrounding the CpG site. With low coverage, the IPDR profiles at individual CpG sites are noisy and insufficient for determining whether the focal CpG site is unmethylated or methylated. However, if we could somehow identify the boundaries of unmethylated/methylated regions, it would be possible to take the average of the IPDR profile for the CpGs within each region and would allow better prediction of the methylation state of each region from its average IPDR profile, which has less noise than the profile of a single CpG site.

We implemented our method using linear discrimination of the vectors of (average) IPDR profiles around the focal CpG sites. We represented the vectors as points residing in the Euclidean space of the appropriate dimension and attempted to separate the points by a decision hyperplane. For better accuracy, we optimized two parameters of the decision hyperplane, the orientation and intercept. Supplemental Fig. S1A shows the optimized orientation. As unmethylated regions are

~2 kb in size (on average) and contain ~50 CpG sites in vertebrate genomes (Su et al., 2012b), in the prediction, we first assume unmethylated regions to have at least 50 CpG sites and integrate the IPDR profiles to make predictions, which is effective in reducing noise in the IPD measurements. We will later examine whether we can reduce the number of CpG sites, denoted by b, while preserving the accuracy. Our method divides the genome into regions containing $\geq b$ CpG sites, such that each region is either unmethylated or methylated. An example of our prediction for the human genome is shown in Fig. 1B, in which our method is able to estimate unmethylation of long duplicated regions while the bisulfite sequencing provides little information. Supplemental Fig. S1C illustrates another example in which both methods are consistent in showing unmethylation in gene promoters.

While setting lower bound b to 50 is supported by the plausible heuristics with biological grounds, a looser bound ($b < 50$) allows us to detect shorter regions. We therefore examined whether we could use a smaller value of b $(= 30, 35, 40, 45)$ without degrading the accuracy of prediction. To evaluate the accuracy of our method, we used the chromosome 1 of length 34,959,811 bp in the medaka genome (version 2) that we assembled from SMRT subreads. For predicting CpG methylation accurately, we guaranteed that each CpG site was covered by at least three subreads, which slightly reduced the original average read coverage, 31.06-fold, to 29.9-fold on the chromosome 1. To examine the coverage effect, we used five subread sets of coverage 20%, 40%, 60%, 80%, and 100% of 29.9-fold. We calculated various accuracy measures, such as sensitivity (recall), specificity (1−false-positive rate), and precision by comparing our prediction on each CpG site with the methylation level determined in a bisulfite sequencing study (Qu et al., 2012). As most CpG sites in the medaka genome are methylated consistently, there are only a small number of positive examples of unmethylated CpGs, and therefore, precision is more informative than specificity in evaluation. We made the trade-off between sensitivity and precision through the selection of the intercept of the decision hyperplane (Supplemental Fig. S2-3). When we used 100% of 29.9-fold subreads, setting b to 35 outperformed the other values (Supplemental Fig. S3), and for better readability, Figure 1C illustrates the sensitivity and precision curve for $b = 35$ only. Our prediction achieved 93.7% sensitivity and 93.9% precision, or 93.0% sensitivity and 94.9% precision, depending on the selection of the intercept. In contrast, for a smaller coverage, a wider window (a larger value of b) was favored; for example, for coverages of 20% and 40% of 29.9-fold, setting b to 50 performed best (Supplemental Fig. S3). Both sensitivity and precision were $\sim 90\%$ for $b = 45$ even if the coverage is relatively small, 60% of 29.9-fold (Supplemental Fig. S3C). Overall, sensitivity and precision of our method are substantially high using a reasonable coverage of SMRT subreads.

2.3 Handling intermediate or ambiguous methylation states

We have introduced a two-class model of our prediction that assigns all of the CpG sites into either unmethylated or methylated regions; however, such a dichotomous model is rather unrealistic, and more refined predictions involving multilevel methylation states or even continuous methylation levels are desirable. For example, an intermediate level of CpG methylation could result from the distinct methylation states of two DNA molecules of diploid cells, although each cytosine must be either methylated or unmethylated in a single DNA molecule. More generally, a sampled cell population can be epigenetically heterogeneous, which would possibly show a spectrum of methylation levels according to its composition. Finally, prediction allowing intermediate states can represent the ambiguity of the prediction, and exclusion of such ambiguous predictions is expected to improve the overall prediction accuracy.

Taking these points into consideration, we extended our method to achieve more complex and informative multi-class predictions. Figures 1D-E depict the concept for multi-class prediction using hypothetical data points. We made a classification using the linear discrimination process involving a separation (decision) hyperplane and determined the position of the hyperplane using the intercept parameter denoted by γ (Fig. 1D). Intuitively, the intermediately methylated CpGs are expected to be distributed more closely to the decision plane, and are therefore more ambiguous than CpGs with *bona fide* methylation states are. Thus, to output the multi-class prediction, we perturbed the intercept γ around its optimal value to produce multiple predictions on each CpG site, which is illustrated by the parallel displaced hyperplanes (Fig. 1E). We then defined the discrete methylation level (DML) as the fraction of predictions that favored 'methylation'. The robust predictions on the *bona fide* methylation states should have extreme DML values, unlike intermediate or ambiguous predictions.

We checked the accordance between our DML and intermediate or ambiguous methylation level captured by two other quantitative methods, bisulfite sequencing and Illumina BeadChip. On the medaka sample, we observed a strong correlation (R = 0.884) between our DML and methylation level calculated from bisulfite sequencing, and we confirmed that 92.0% of CpG sites were in concordance within 0.25 (Fig. 1F). We also compared our DML on the human sample to the beta value (an indicator of methylation level expressed as a value ranging over [0,1]) obtained from Illumina BeadChip (Fig. 1G). We observed a weaker correlation ($R = 0.816$)

and a smaller fraction (75.4%) of CpG sites in concordance within 0.25 presumably because the beta value is less quantitative than the methylation level calculated from bisulfite sequencing (Wang et al., 2015). Overall, DML serves to reflect the quantitative nature of methylation status in the samples.

2.4 Genome-wide methylation pattern of repetitive elements in the human genome

We investigated how individual occurrences of repetitive elements were methylated in the human genome, as summarized in Table 1. Of note, some occurrences of repetitive elements contain no or very few CpG sites, and thus we only consider those occurrences with at least 10 CpGs to exclude other less informative cases. First, we checked whether SMRT reads could address the repetitive regions in a useful manner for methylation analysis. Specifically, we considered a repeat occurrence to be covered by uniquely mapped SMRT reads if the IPD ratio was available on $\geq$50% of CpGs, and found that $>$96% were covered for every repeat type. To draw robust conclusions, we further applied a stringent quality control process to each repeat occurrence such that the read coverage was $>$5. Although this step reduced the number of repeat occurrences under consideration by $3-18$%, this reduction could be mitigated simply by producing more data. Finally, we treated an occurrence as unmethylated if $\geq$50% of CpGs were predicted as unmethylated. Fractions of unmethylated repeat occurrences vary considerably among different classes of repetitive elements, from $\sim$1% for L1 and Alu to $\sim$50% for MIR and $>$70% for simple repeats and low-complexity regions. To validate our prediction regarding the repeat occurrences, we selected 21 regions for bisulfite Sanger sequencing, designed primers for nested PCR (Supplemental Table S2), and could amplify six regions, indicating the difficulty in observing DNA methylation of repetitive elements using traditional bisulfite Sanger sequencing. In the six amplified regions, we confirmed the consistency between our prediction and the methylation state observed by bisulfite Sanger sequencing (Supplemental Fig. S4).

We then examined the features for characterizing the differences between methylated and unmethylated repetitive elements. First, CpG density was significantly higher in the unmethylated occurrences in almost all classes of repetitive elements ($p < 1\%$, Fig. 2A). This observation was consistent with the known association between CpG-rich regions and unmethylation because methylation leads to depletion of CpG sites through deamination (Cooper and Krawczak, 1989). Second, sequence divergence from the representative in each repeat class also showed a correlation with methylation status (Fig. 2B). For most classes, with the apparent exception of simple repeats, low-complexity regions, and MIR elements, unmethylated occurrences were significantly more divergent than were methylated occurrences ($p < 1\%$, Fig. 2B), presumably because younger copies of a repeat element are less divergent and are likely to be targets of DNA methylation. We also examined whether the methylation status of repetitive elements could be correlated partly with sequence features. Kernel principal component analysis (PCA) using spectrum kernel suggested positive answers for some repeat types (Supplemental Fig. S5).

Next, we examined whether the unmethylated repeat occurrences were distributed uniformly or non-uniformly throughout the entire genome. We selected three major classes (LINE, Alu, and LTR) of repetitive elements for this analysis. We calculated the ratios of unmethylated copies to all repetitive elements in individual non-overlapping bins 5 Mb in size (Fig. 2C). The non-random distribution patterns were more evident for LINE and LTR than for Alu. For example, we found unmethylated LINEs to be enriched in the p-arm of chromosome 1 and in chromosomes 17 and 19. There were unmethylation 'hot spots' of LTR elements, *e.g.*, in chromosomes 6 and 9 (Supplemental Fig. S6). It is intriguing that some of these unmethylation hot spots, such as those in the p-arms of chromosomes 6 and Y, seem to be shared among different classes of repetitive elements.

We further investigated the methylation states of LINE/L1 elements, the only active autonomous retrotransposons in mammals (Furano, 2000). Although most of LINE/L1 insertions contain many mutations, Penzkofer *et al.* categorize full-length L1 elements into three classes according to the conservation of two open reading frames (ORFs) (Penzkofer et al., 2005); namely, L1s with intact in the two ORFs that are likely to exhibit retro-transposition activity, L1s with an intact ORF2 but a disrupted ORF1, and non-intact L1s. We obtained the positions of these human LINE/L1 elements from L1Base (Penzkofer et al., 2005) and analyzed their methylation states (Supplemental Table 3). Although 0.61% of non-intact L1s were unmethylated, all of L1s with intact in two ORFs and L1s with an intact ORF2 were methylated. We also checked the presence of LINE insertions that were novel to the hg19 reference genome. We assembled the SMRT reads using FALCON and searched the assembly for novel LINE insertions that matched a hot L1 element (GenBank: M80343.1) of size 6050 bp with identity $> 98.5\%$. The hot L1 element was used as the representative according to the procedure of L1Base (Penzkofer et al., 2005). We identified two novel instances covered by sufficient depth of SMRT reads that allowed us to call their methylation statuses confidently. Both of the two LINE insertions were estimated to be methylated, and the positions of these two occurrences are found in Supplemental Fig. S7.

These results confirmed putatively active LINE/L1 elements with intact ORFs were targets of DNA methylation.

2.5 Analysis of the *Tol2* transposable element in medaka

The medaka has an innate autonomous transposon known as *Tol2*, which is one of the first examples of autonomous transposons in vertebrate genomes and a useful tool for genetic engineering of vertebrates, such as zebrafish and mice (Kawakami, 2007). The excision activities of *Tol2* are promoted when DNA methylation is reduced by 5-azacytidine treatment, which suggests that DNA methylation is one of the mechanisms regulating the *Tol2* transposition (Iida et al., 2006). Nevertheless, observing the methylation status of each *Tol2* copy using short reads is difficult, because *Tol2* is 4682 b in length, and $\sim$20 highly similar copies of *Tol2* exist in the genome (Koga et al., 2000).

To elucidate the methylation status of each *Tol2* copy, we applied our method to a new assembly of the Hd-rR genome obtained exclusively from SMRT reads. We found 17 copies of *Tol2* contained entirely within this assembly, all of which were essentially identical ($>$99.8% sequence identity). We then called the methylation status of these *Tol2*. For comparison, we mapped bisulfite-treated short reads to these contigs and determined the methylation level. The methylation status of these *Tol2*, observed by SMRT reads and bisulfite-sequencing, are shown in Fig. 3. While virtually no *Tol2* copies were mapped by bisulfite reads, as expected from their extremely high fidelity, 16 of 17 copies were anchored by SMRT reads, and all were predicted to be methylated by our method. For the regions examined by both SMRT reads and bisulfite-treated short reads, our prediction was consistent with the methylation level calculated from the bisulfite-treated reads. For example, one *Tol2* copy was surrounded by unmethylated regions (number 14). From the bisulfite data, it appeared that the body of *Tol2*, from which data were missing, was unmethylated. Nevertheless, our prediction estimated this region to be methylated. These results demonstrate the ability of our method using SMRT reads to clarify DNA methylation states of highly identical repetitive elements such as active transposons.

3 DISCUSSION

In this study, we addressed the problem of uncovering the landscape of DNA methylation of repetitive elements. To this end, we developed a unique application of SMRT sequencing to epigenetics. This direction had been already explored in the research community for bacterial and viral species. However, this application in large vertebrate genomes has been largely unexplored because of the subtle cytosine methylation signals in the kinetic information. Therefore, we proposed a new method to utilize relatively small amounts of kinetic information by incorporating a model reflecting our prior knowledge on the regional patterns of CpG methylation of vertebrate genomes. We confirmed the validity of our strategy by comparing the prediction to bisulfite sequencing data on medaka and to BeadChip analysis on human samples. These two datasets had very different characteristics, which seemed to be partly because of the methods used (*i.e.*, BeadChip was designed to observe mainly CpG islands that are often unmethylated, while bisulfite sequencing is used for genome-wide methylation analysis) and partly because of the nature of the samples used (*i.e.*, the medaka samples were derived from an inbred strain, while the human samples were from diploid cells). Despite such differences in characteristics, our method using the same parameters performed almost equally well for both datasets (Fig. 1F,G). These observations suggested that the choice of parameters is robust for a wide variety of samples, which is a desirable feature for any method. We also presented an extension of our method to accommodate intermediate methylation states, the discrete methylation level (DML), and confirmed a high correlation ($R = 0.884$) between DML and bisulfite methylation level.

We explored the epigenetic landscape of repetitive elements within the human genome. Using the hg19 reference genome is an apparent limitation. By assembling individual personal genomes instead of the reference genome, new insertions of these repetitive elements are expected to be found, and such active occurrences should be of interest. Indeed, we detected two novel LINE insertions that were estimated to be methylated. Importantly, the more recent the insertion event, the less divergent it would be from the original copy, and therefore, there would be less likelihood of it being anchored by short reads. In such cases, long SMRT reads shed new light on the ecosystem of active repetitive elements in personal human genomes.

Finally, our method had important strengths compared with conventional tools for epigenetic studies, such as bisulfite sequencing or affinity-based assays, with not only an expected increase in comprehensiveness by virtue of long SMRT reads, but also in the remarkable reduction of laboratory work. If an epigenetic study is conducted alongside a resequencing study or a *de novo* assembly study using SMRT sequencing, the methylation status could be called solely *in silico*, and no additional experiments would be necessary.

METHODS

Software availability

Our software program AgIn (Aggregate on Intervals) is available at: https://github.com/hacone/AgIn

Preparation of genomic DNA and SMRT sequencing

DNA was extracted from the testes of Hd-rR medaka with the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany), following the tissue protocol. Genomic DNA was isolated from peripheral blood leukocytes of two Japanese patients using standard procedures after informed consent. The DNA featured A280/260 values of ∼1.8 and formed a clear, sharp band on agarose gel electrophoresis.

For the medaka sample and one human sample, genomic DNA was sheared using g-Tube devices (Covaris Inc., Woburn, MA, USA), targeting 20 kb fragments at 4300 rpm, 150 ng/µl and purified using 0.45× volume ratio of AMPure beads (Pacific Biosciences, Menlo Park, CA, USA). SMRTbell$^{\mathrm{TM}}$ libraries were prepared with the DNA Template Preparation Kit 1.0 (Pacific Biosciences, Menlo Park, CA, USA) using the “20-kb Template Preparation using BluePippin Size Selection System (15 kb Size Cutoff)” protocol. Sequencing primer was annealed to the template at 0.833 nM concentration. SMRT bell$^{\mathrm{TM}}$ templates were sequenced using magnetic bead loading, C4 chemistry, and polymerase version P6. Sequence data were collected on the magnetic bead collection protocol, 20 kb insert size, stage start, and 240 min movies in PacBio RS Remote.

For the other human sample, sequencing was performed as follows. Genomic DNA was sheared with using g-TUBE devices, targeting 10 kb fragments. SMRTbell$^{\mathrm{TM}}$

libraries were prepared with the DNA Template Preparation Kit 2.0 (3∼10 kbp) (Pacific Biosciences, Menlo Park, CA, USA). Briefly, sheared DNA was end-repaired, and hairpin adapters were ligated using T4 DNA ligase. Incompletely formed SMRTbell$^{\mathrm{TM}}$ templates were degraded using a combination of exonucleases III and VII. The resulting DNA templates were purified using (0.45×) SPRI magnetic beads (AMPure; Agencourt Bioscience, Beverly, MA, USA). Sequencing primers were annealed to the templates at a final concentration of 5 nM template DNA. SMRTbell$^{\mathrm{TM}}$ library was sequenced using Magbead loading, C2 chemistry, and Polymerase version C2 or P4. Sequence data were collected on the PacBio RS for 120 min.

Regarding two human samples, the latter sample matches the one used for Illumina BeadChip analysis. We used the sequencing data and methylation state prediction from this sample solely for the analysis of intermediate methylation state prediction (Fig. 1G).

Normalization of beta values of Illumina BeadChip

The respective beta values of an unmethylated CpG and a methylated CpG are not always equal to 0 and 1. Indeed, in our data, the distribution of raw beta values of Illumina BeadChip had bimodal peaks at 0.04 and 0.89. To compare beta values with our DML data (Fig. 1G), we treated 0.04 and 0.89 as unmethylated and methylated states respectively and normalized raw beta values by setting $x \leq 0.04$ to 0, $0.89 \leq x$ to 1, and $0.04 < x < 0.89$ to $(x - 0.04)/(0.89 - 0.04)$, proportionally.

Raw IPDR and read coverage

We used the PacBio RS SMRT pipeline to process raw kinetic data from SMRT sequencing to obtain the mean IPDR and read coverage at each genomic position. r_i and r'_i denote the mean IPDR associated with position i of the forward and reverse strands, respectively, and c_i and c'_i denote the read coverage at position i of the forward and reverse strands, respectively. To remove outlier noise inherent in raw data, mean IPDRs $>$10 were Winsorized to 10 and positions with less than three reads were excluded from the data (the latter was handled by SMRT Pipe). In bisulfite sequencing, CpG sites with $\geq$ 10 reads that mapped to C of either strand were considered covered. CpG sites that have a ≥ 1 position within a 21 bp window with ≥ 3 SMRT reads were counted as covered.

Estimating the methylation status at individual CpG sites

Suppose that the focal genome has n CpG sites. We can assign identifiers ranging from 1 to n to individual n CpG sites and denote the genomic position of C of the i-th CpG site by p_i. For example, the second CpG site at the 10th genomic position is denoted by "$p_2 = 10$." Our goal was to predict the methylation status, unmethylated or methylated, at p_i using information on read coverage and IPDR at positions surrounding p_i. We used positions within 10 bases around p_i because these neighboring positions have proven to be effective in predicting 5-hydroxymethylcytosine, N4-methylcytosine, and N6-methylademine in bacteria genomes in previous studies (Flusberg et al., 2010; Clark et al., 2011, 2012; Fang et al., 2012). Neighboring positions are denoted by $p_i + j$ for $j = -10, \ldots, +10$ in the plus strand. For example, the positions 5 bases upstream and downstream of

p_i are $p_i - 5$ and $p_i + 5$, respectively.

To achieve a better prediction, we derived a modified IPDR vector from raw read coverage and IPDR within 10 bases around p_i. For this purpose, we took into account the property that any CpG site in one strand is reverse complementary to the CpG in the other strand, and the methylation status of Cs at a pair of CpG sites in both strands is consistent in most cases, making it meaningful to combine IPDR information for both strands to predict the methylation status. To represent positions in the minus strand, we note that since we set the position of C of the focal CpG in the plus strand to p_i, the position of C of the CpG in the minus strand is $p_i + 1$, and the surrounding positions are $p_i + 1 - j$ for $j = -10, \ldots, +10$. In addition, we attached more importance to IPDR values associated with a higher read coverage and we quantified this as $c_{p_i+j} \times r_{p_i+j}$ in the plus strand ($c'_{p_i+1-j} \times r'_{p_i+1-j}$ in the minus strand). We then took the sum of all the products and normalized it by dividing it by the total number of reads. Finally, we obtain the 21-dimensional modified IPDR vector for 21 genomic positions around CpG site p_i:

$$\hat{\boldsymbol{X}}(p_i)_j = \frac{c_{p_i+j} r_{p_i+j} + c'_{p_i+1-j} r'_{p_i+1-j}}{c_{p_i+j} + c'_{p_i+1-j}} \qquad (j = -10, \ldots, +10).$$

We are now in a position to define a classifier that uses $\hat{\boldsymbol{X}}(p_i)$ as explanatory variables and predicts the methylation status at p_i, which is also estimated independently by bisulfite sequencing (Qu et al., 2012). We attempted to use linear discriminant analysis (LDA) with the discriminant function

$$\mathrm{F}(p_i) = \boldsymbol{\beta} \cdot \hat{\boldsymbol{X}}(p_i) + \gamma,$$

where we optimized values of the coefficient vector $\boldsymbol{\beta}$ and variable γ using bisulfite sequencing data as the training data set to improve the prediction. Supplemental Fig. S1A shows the optimized vector $\boldsymbol{\beta}$ that we used in this study. If the sign of the discriminant function, $\mathrm{F}(p_i)$, is positive, the methylation status at p_i is defined as 'methylated'; otherwise, it is defined as 'unmethylated.' We note that according to previous studies (Flusberg et al., 2010; Fang et al., 2012; Schadt et al., 2012), estimating 5-methylcytosine residues with low read coverage, for example $c_{p_i+j} + c'_{p_i+1-j} < 100$, is prone to errors, demanding hundreds of reads, which is extremely costly to achieve.

Predicting the methylation status of CpG blocks

In vertebrates, unmethylated CpG dinucleotides are rare ($\sim$10%) and do not always exist in isolation, but they are likely to range over long unmethylated regions. This motivated us to integrate low-coverage reads around CpGs in a region to yield high-coverage for estimating the methylation status in the entire region, rather than at a single-base resolution. Let A denote a region. The following formula expresses the average IPDR vector for 21 genomic positions around all of the CpG sites in region A and its associated discriminant function:

$$\begin{aligned} \hat{\boldsymbol{X}}(A)_j &= \frac{\sum_{p_i \in A}(c_{p_i+j} r_{p_i+j} + c'_{p_i+1-j} r'_{p_i+1-j})}{\sum_{p_i \in A}(c_{p_i+j} + c'_{p_i+1-j})} \qquad (j = -10, \ldots, +10). \\ \mathrm{F}(A) &= \boldsymbol{\beta} \cdot \hat{\boldsymbol{X}}(A) + \gamma \end{aligned}$$

In processing a longer region with more CpG sites, the accuracy of methylation status prediction can improve, although smaller regions may be overlooked. In our analysis, we impose the constraint that each region has at least b CpG sites. For example, we can set b to 50 because Su *et al.* report that the average length of unmethylated regions in five human cell types is approximately 2 kb (Su et al., 2012b) and the average distance between neighboring CpG sites in the medaka genome is 53.5 bases, although this constraint should be adjusted according to each individual situation.

The possibility of the methylation (unmethylation, respectively) of A increases with a larger positive (negative) value of $\mathrm{F}(A)$, as well as for a larger total number of reads

$$w(A) = \sum_{p_i \in A, j=-10,\ldots,+10} (c_{p_i+j} + c'_{p_i+1-j}).$$

Thus, region A associated with a larger value of $w(A)\mathrm{F}(A)$ is better.

Decomposing the genome into unmethylated/methylated CpG blocks

Now, we must consider how to decompose n CpG sites in the whole genome into methylated regions $\{M_{\lambda \in \Lambda}\}$ and unmethylated regions $\{U_{\mu \in M}\}$ such that all regions are disjoint from each other, their union covers all CpG sites, and the two types of region occur alternatingly along the genome. To obtain better regions, we calculated the optimal decomposition of regions that maximizes the following objective function:

$$\sum_{\lambda \in \Lambda} w(M_\lambda)\mathrm{F}(M_\lambda) + \sum_{\mu \in M} -w(U_\mu)\mathrm{F}(U_\mu).$$

To solve this problem, we here mention one important characteristic of SMRT sequencing. Read coverage is not affected by the sequence composition in SMRT sequencing (Bashir et al., 2012; Zhang et al., 2012; English et al., 2012; Koren et al., 2012; Loomis et al., 2012). Thus, we assume that the sum of reads at the j-th position around all CpG sites in region A is a constant $\bar{c}(A)$ independent of j:

$$\sum_{p_i \in A} (c_{p_i+j} + c'_{p_i+1-j}) = \bar{c}(A) \qquad \text{for } \ j = -10, \ldots, 10$$

This allows us to transform $w(A)$ into a simpler form:

$$w(A) \ = \sum_{p_i \in A, j=-10,\ldots,+10} (c_{p_i+j} + c'_{p_i+1-j}) = 21\bar{c}(A)$$

Subsequently, we can also simplify the objective function:

$$\begin{aligned}
& w(A)F(A) \\
&= w(A)(\boldsymbol{\beta} \cdot \hat{\boldsymbol{X}}(A) + \gamma) \\
&= 21\bar{c}(A) \left(\sum_{j=-10,\ldots,+10} \boldsymbol{\beta}_j \frac{\sum_{p_i \in A}(c_{p_i+j} r_{p_i+j} + c'_{p_i+1-j} r'_{p_i+1-j})}{\bar{c}(A)} \right) + \gamma w(A) \\
&= 42 \left(\sum_j \boldsymbol{\beta}_j \sum_{p_i \in A} \frac{(c_{p_i+j} r_{p_i+j} + c'_{p_i+1-j} r'_{p_i+1-j})}{2} \right) \\
&\quad + \sum_{p_i \in A, j} \gamma (c_{p_i+j} + c'_{p_i+1-j}) \\
&= 42 \sum_{p_i \in A} \sum_j \left(\boldsymbol{\beta}_j \frac{(c_{p_i+j} r_{p_i+j} + c'_{p_i+1-j} r'_{p_i+1-j})}{2} + \gamma \frac{c_{p_i+j} + c'_{p_i+1-j}}{42} \right) \\
&= 42 \sum_{p_i \in A} s_i,
\end{aligned}$$

where s_i denotes $\sum_j \left(\boldsymbol{\beta}_j \frac{(c_{p_i+j} r_{p_i+j} + c'_{p_i+1-j} r'_{p_i+1-j})}{2} + \gamma \frac{c_{p_i+j} + c'_{p_i+1-j}}{42} \right)$ in the second last formula because the value only depends on read coverage and IPDR values at 21 genomic positions surrounding p_i. Consequently, our objective function to optimize became a linear combination of s_i:

$$\sum_{\lambda\in\Lambda} w(M_\lambda)\mathrm{F}(M_\lambda) + \sum_{\mu\in M} -w(U_\mu)\mathrm{F}(U_\mu) \quad = \quad \sum_{\lambda\in\Lambda}\sum_{p_i\in M_\lambda} s_i + \sum_{\mu\in M}\sum_{p_i\in U_\mu}(-s_i)$$

Although we set s_i to a score calculated from weighted IPDR information, we can set s_i to a log-likelihood function of the form - log Q_i for some likelihood function Q_i.

This simple form motivated us to design a dynamic programming algorithm for calculating the optimal value efficiently. We considered the subproblem involving the first i CpG sites among all n sites, and let S_i^M and S_i^U be the maximum value of the objective function when the last i-th CpG site was methylated and unmethylated, respectively. S_i^M and S_i^U meet the following recurrences:

$$\begin{aligned} S_{i+1}^M &= \max\{S_i^M + s_{i+1},\ S_{i-b+1}^U + \sum_{k=i-b+2}^{i+1} s_k\} \\ S_{i+1}^U &= \max\{S_i^U - s_{i+1},\ S_{i-b+1}^M + \sum_{k=i-b+2}^{i+1} (-s_k)\} \end{aligned}$$

The first max term implies extension of the running region by one CpG site, while the second term means a switch from the previous methylation status and the initiation of a new region with $\geq b$ CpG sites. For example, we can set b to 50, but one can change the requirement for the minimum number of CpG sites in a region by making an appropriate adjustment to the second term. Of S_n^M and S_n^U, the larger value gives the maximum value, and tracing back the optimal path from the maximum value provides all the boundaries between neighboring methylated and unmethylated regions. To calculate regions satisfying the constraint on the minimum number of CpG sites, we generalized the dynamic programming idea proposed by Csűrös (Csűrös, 2004).

Thus far, we have assumed two possible methylation statuses, methylated or unmethylated, because this situation is true in most cases in our inbred strain sample (Qu et al., 2012). In human cells, however, many partially methylated cytosines have been reported (Lister et al., 2009). To consider such situations, we need to extend our algorithm to involve scores for three methylation statuses, methylated, unmethylated, and partially methylated. One can redesign the score function and the recurrence for each class. For example, making the parameters, $\boldsymbol{\beta}$, γ, and b, depend on the class to which the i-th CpG site belongs, we can redefine the new recurrences for three classes:

$$\begin{aligned}
S^M_{i+1} &= \max\{S^M_i + s^M_{i+1},\ \max\{S^U_{i-b^M+1}, S^P_{i-b^M+1}\} + \sum_{k=i-b^M+2}^{i+1} s^M_k\} \\
S^U_{i+1} &= \max\{S^U_i + s^U_{i+1},\ \max\{S^P_{i-b^U+1}, S^M_{i-b^U+1}\} + \sum_{k=i-b^U+2}^{i+1} s^U_k\} \\
S^P_{i+1} &= \max\{S^P_i + s^P_{i+1},\ \max\{S^M_{i-b^P+1}, S^U_{i-b^P+1}\} + \sum_{k=i-b^P+2}^{i+1} s^P_k\},
\end{aligned}$$

where P denotes “partially methylated,” b^C indicates the minimum region length for each class $C \in \{M, U, P\}$, and

$$s^C_i = \sum_j \left(\beta^C_j \frac{(c_{p_i+j} r_{p_i+j} + c'_{p_i+1-j} r'_{p_i+1-j})}{2} + \gamma^C \frac{c_{p_i+j} + c'_{p_i+1-j}}{42} \right).$$

One might wonder if the hidden Markov Model (HMM) can be used for computing unmethylated and methylated regions; however, it is not obvious that using HMM guarantees the requirement that each range has $\geq b$ CpG sites.

Computing discrete methylation levels

For calculating more quantitative methylation level named *discrete methylation levels*, we performed prediction using the set of 10 perturbed intercept values (γ ranging from -12% to +24% by 4%) so we obtain 10 predictions on each CpG site. Then, on each CpG site, the number of predictions that favored methylation were divided by 10, yielding the discrete methylation level ranging over $[0, 1]$.

Methylation status calculated from bisulfite sequencing

We evaluated the prediction accuracy of our integration method using methylation scores calculated from bisulfite-treated Illumina reads as the answer set. Let S be the set of bisulfite-treated Illumina reads covering the i-th CpG site, x be the number of methylated CpGs in S at i, and y be the coverage of S at i (the size of S). We defined the methylation score at i as x/y. We then defined the methylation status as 'unmethylated' if the score was less than 0.5; otherwise, it was defined as 'methylated'.

We need to carefully constrain the value of the coverage y. Allowing a lower value of y is likely to produce more erroneous methylation scores, while using y greater than a higher threshold would reduce the number of CpGs associated with their methylation scores. The average coverage was 9.40-fold in our bisulfite-treated reads collected from testes of the Hd-rR medaka inbred strain; however, the coverage at individual CpG sites varied to some extent. We defined the methylation score only when the CpG site was covered by 10 or more reads (*i.e.*, $y \geq 10$) so that make sure the scores are robust enough.

Prediction accuracy of our method at individual CpG sites

We predicted the methylation status of each CpG site by checking whether the CpG site was located in an unmethylated or methylated region according to the output of our integration method. We measured the accuracy of the prediction by checking the consistency between the prediction and the methylation score associated with each CpG site. CpG sites without methylation score (due to the lack of bisulfite-treated reads) were ignored. We treat a unmethylated status as positive and a methylated status as negative, because we are more interested in identifying rare unmethylated regions accounting for a small portion (e.g., $\sim$10%) of CpG sites.

Methylation analysis of human repetitive elements

We started the analysis by listing repetitive elements using the Repeat Library "20140131" (Smit, A., Hubley, R. & Green, P. Repeatmasker open-4.0 at http://www.repeatmasker.org). Only repetitive elements containing at least 10 CpG sites were considered. We calculated the methylation levels of CpG sites as discrete methylation levels (DML), and CpG sites with a DML<0.4 were considered as unmethylated. To further reduce the degree of unmethylation assigned false-positively, we filtered out repetitive elements with an average read coverage on CpG sites of <5.0. Finally, we treated repetitive elements as unmethylated if more than half of the CpG sites were unmethylated; otherwise, they were considered as methylated.

The relationships between methylation state and CpG density or divergence were tested for statistical significance using the Mann-Whitney U test. To draw ideograms in Figure 2C, we counted the numbers of unmethylated and methylated repeats in every 5 Mb bin and then used the Ideographica web server (Kin and Ono, 2007) to generate the images. In Supplemental Figure S4, Kernel PCA analysis was performed using spectrum kernel. As the magnitude of sequence divergence among occurrences was markedly variable for different types of repetitive elements, it was necessary to optimize the k-mer size for each type of repetitive element to achieve better visualization.

Validation of our prediction by bisulfite Sanger sequencing

Bisulfite conversion of genomic DNA was performed using a commercially available kit (MethlEasy Xceed Rapid DNA Bisulphite Modification Kit; Human Genetic Signatures, NSW, Australia). Briefly, 5 μg of DNA was denatured by 0.3 M NaOH for 15 minutes at 37°C. Subsequently, the samples were incubated with bisulfite solution for 45 minutes at 80°C. After purification, the eluted samples were incubated for 20 minutes at 95°C. The converted DNA was stored at -20°C for PCR amplification.

To perform targeted PCR on the 21 regions selected for validation, we designed primers for nested PCR to amplify 111∼622bp fragments of bisulfite-converted DNA (Supplemental Table S2). Primer pairs were purchased from Life Technologies (Supplementary Information). PCR was performed in a volume of 50 μL containing 1 × EpiTaq PCR Buffer, 2.5 mM MgCl2, 0.3 mM dNTP mix, 20 pmol primers, 1.25 units TakaraEpiTaq HS polymerase (Shiga, Japan), and 50 ng bisulfite-converted DNA. PCR conditions were 40 cycles of 98°C for 10 seconds, 55°C for 30 seconds, and 72°C for 1 minute. To check the quality of the PCR products, 2% agarose gel electrophoresis was used in 1 × TAE buffer at 50 volts for 15 minutes. The amplified products were visualized using a LED transilluminator, and the product bands were purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel GmbH & Co. KG, Dueren, Germany). Targeted PCR products were sequenced directly using ABI3730 sequencers with BigDye v3.1 chemistry (Applied Biosystems, Foster City, CA, USA).

Finally, we processed the obtained sequencing data using the QUMA online tool (Kumaki et al., 2008) for analysis and visualization of the methylation patterns (Supplemental Fig. S3).

Methylation analysis of medaka *Tol2* elements

In Figure 3, we applied our method to observe the methylation state of a new medaka assembly. For comparison, we also called the methylation state on every 100-bp window using Bismark software and the publicly available bisulfite-treated reads from the testes of the Hd-rR strain. Among the assembly, we identified 17 contigs containing *Tol2* elements by BLAST search.

Other data sources and data visualization

Figure 1B and Supplemental Figures S5-S6 were produced using the UCSC Genome Browser (http://genome.ucsc.edu/) (Karolchik et al., 2014). We used human bisulfite sequencing data and unmethylated regions available in the GEO database (Song et al., 2013).

DATA ACCESS

The sequence data (SMRT reads) from the medaka sample are deposited at the NCBI Archive (Accession No. SRP020483). Sequence data from a Japanese individual are available under controlled access through the National Bioscience Database Center (NBDC, accession number JGAS00000000003).

ACKNOWLEDGMENTS

This work was supported in part by MEXT (Grants-in-Aid for Scientific Research 22129008 and 23241058 [SM]). We would like to thank Ryohei Nakamura for stimulating discussions.

FIGURE LEGENDS

Figure 1. Outline of our integration method

A. The top three distributions show the typical Inter-Pulse Duration Ratio (IPDR) profiles within 10 bp of the CpG sites in the raw data. The IPDR profiles of individual CpG sites were treated as points in the 21-dimensional feature space. Red-colored unmethylated CpGs and blue-colored methylated CpGs are difficult to separate using a hyperplane. Therefore, initially, we had little knowledge about the methylation status of each CpG site from the raw data, as illustrated by the question marks at the CpG sites. Our algorithm predicts the boundary of unmethylated and methylated CpG sites. The average IPDR profiles of the two regions, which have clearly distinct IPDR profiles, are shown below the two regions separated by the boundary (see the detailed IPDR profiles in Supplemental Fig. S1B). Red circles and blue boxes represent unmethylated and methylated CpGs, respectively, predicted by our algorithm (annotated as 'predicted regions') and were observed by bisulfite sequencing (labeled 'answer'). In the feature space, red and blue disks represent the IPDR profiles of predicted regions. **B.** Comparison of our prediction with the available human genome methylome data. From top to bottom, 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 row shows repeat masker tracks. Both methods are consistent in showing unmethylation on the three blue-colored regions. No read counts of the bisulfite data are available in long duplicated regions where the alignability is quite low, but our method can estimate unmethylation in these regions. **C.** The sensitivity and precision (proportion of true-positives among the predicted positives) of our integration method are evaluated on individual CpG sites when we change the intercept of the hyperplane and set the minimum number of CpG sites in a region, b, to 35. See the curves for $b =$ 30,40,45, and 50 in Supplemental Fig S3. **D.** 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). **E.** 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. **F.** 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. **G.** 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.

Figure 2. Epigenetic landscape of repetitive elements in the human genome

A-B. Distribution of CpG density (**a**) and sequence divergence from the representative in each repeat class (**b**) for methylated (cyan) and unmethylated (pink) repeat occurrences. The asterisks indicate statistical significance ($p < 1\%$) determined by the U test. **C.** Genome-wide distribution of unmethylated repetitive elements. The ratio of unmethylated repeat occurrences to all occurrences in each 5-Mb bin is indicated by color shadings. Prediction of the methylation state was performed after quality control as described in the text.

Figure 3. Methylation analysis of *Tol2*, a 4682-bp long autonomous transposon, in medaka

The new genome assembly of SMRT reads had 17 regions (contigs) that contained complete *Tol2* copies. The circles show our prediction of the methylation state of CpG sites, while the rectangles show the methylation states within each 100 bp window obtained from bisulfite sequencing. For both tracks, open/red indicates unmethylation and filled/blue indicates methylation. The arrow above indicates the region of *Tol2* insertions. As the eleventh region was located at the extreme of the contig, *Tol2* was not observed successfully by either SMRT sequencing or bisulfite sequencing. For the other 16 regions, methylation of *Tol2* was observed consistently by SMRT sequencing, while virtually no information was available on the *Tol2* region from bisulfite sequencing.

References

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A., 2011. Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome biology*, **12**(2):R18.

Anway, M. D., Cupp, A. S., Uzumcu, M., and Skinner, M. K., 2005. Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science*, **308**(5727):1466–9.

Bashir, A., Klammer, A. a., Robins, W. P., Chin, C.-S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P., *et al.*, 2012. A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology*, **30**:701–707.

Beck, C. R., Collier, P., Macfarlane, C., Malig, M., Kidd, J. M., Eichler, E. E., Badge, R. M., and Moran, J. V., 2010. LINE-1 retrotransposition activity in human genomes. *Cell*, **141**(7):1159–1170.

Bock, C., Tomazou, E. M., Brinkman, A. B., Müller, F., Simmer, F., Gu, H., Jäger, N., Gnirke, A., Stunnenberg, H. G., and Meissner, A., *et al.*, 2010. Quantitative comparison of genome-wide dna methylation mapping technologies. *Nature biotechnology*, **28**(10):1106–1114.

Bock, C., Walter, J., Paulsen, M., and Lengauer, T., 2008. Inter-individual variation of dna methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Res*, **36**(10):e55.

Chandler, L. A., Ghazi, H., Jones, P. A., Boukamp, P., and Fusenig, N. E., 1987. Allele-specific methylation of the human c-ha-ras-1 gene. *Cell*, **50**(5):711–7.

Clark, T., Lu, X., Luong, K., Dai, Q., Boitano, M., Turner, S., He, C., and Korlach, J., 2013. Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via tet1 oxidation. *BMC Biology*, **11**(1):4.

Clark, T. a., Murray, I. a., Morgan, R. D., Kislyuk, A. O., Spittle, K. E., Boitano, M., Fomenkov, A., Roberts, R. J., and Korlach, J., 2012. Characterization of dna methyltransferase specificities using single-molecule, real-time dna sequencing. *Nucleic acids research*, **40**:e29.

Clark, T. a., Spittle, K. E., Turner, S. W., and Korlach, J., 2011. Direct detection and sequencing of damaged dna bases. *Genome integrity*, **2**:10.

Cokus, S. J., Feng, S. H., Zhang, X. Y., Chen, Z. G., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E., *et al.*, 2008. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, **452**(7184):215–219.

Cooper, D. and Krawczak, M., 1989. Cytosine methylation and the fate of cpg dinucleotides in vertebrate genomes. *Human genetics*, **83**(2):181.

Csűrös, M., 2004. Maximum-scoring segment sets. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **1**(4):139–150.

Deng, J., Shoemaker, R., Xie, B., Gore, A., LeProust, E. M., Antosiewicz-Bourget, J., Egli, D., Maherali, N., Park, I. H., Yu, J., *et al.*, 2009. Targeted bisulfite sequencing reveals changes in dna methylation associated with nuclear reprogramming. *Nat Biotechnol*, **27**(4):353–60.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., *et al.*, 2006. Dna methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, **38**(12):1378–85.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al.*, 2009. Real-time dna sequencing from single polymerase molecules. *Science (New York, N.Y.)*, **323**:133–8.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., *et al.*, 2012. Mind the gap: Upgrading genomes with pacific biosciences rs long-read sequencing technology. *PLoS ONE*, **7**:e47768.

Fang, G., Munera, D., Friedman, D. I., Mandlik, A., Chao, M. C., Banerjee, O., Feng, Z., Losic, B., Mahajan, M. C., Jabado, O. J., *et al.*, 2012. Genome-wide mapping of methylated adenine residues in pathogenic escherichia coli using single-molecule real-time sequencing. *Nature Biotechnology*, .

Feng, Z., Fang, G., Korlach, J., Clark, T., Luong, K., Zhang, X., Wong, W., and Schadt, E., 2013. Detecting dna modifications from smrt sequencing data by modeling sequence context dependence of polymerase kinetic. *PLoS Comput Biol*, **9**(3):e1002935.

Flusberg, B. a., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. a., Korlach, J., and Turner, S. W., 2010. Direct detection of dna methylation during single-molecule, real-time sequencing. *Nature methods*, **7**:461–5.

Furano, A., 2000. The biological properties and evolutionary dynamics of mammalian line-1 retrotransposons. *Prog. Nucleic Acid Res. Mol. Biol.*, **64**:255–294.

Gifford, C. A., Ziller, M. J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A. K., Kelley, D. R., Shishkin, A. A., Issner, R., *et al.*, 2013. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell*, **153**(0):1149 – 1163.

Goodier, J. L., 2014. Retrotransposition in tumors and brains. *Mobile DNA*, **5**(1):11.

Greer, E. L., Blanco, M. A., Gu, L., Sendinc, E., Liu, J., Aristizábal-Corrales, D., Hsu, C.-H., Aravind, L., He, C., and Shi, Y., *et al.*, 2015. Dna methylation on n 6-adenine in c. elegans. *Cell*, **161**(4):868–878.

Harris, R. A., Wang, T., Coarfa, C., Nagarajan, R. P., Hong, C., Downey, S. L., Johnson, B. E., Fouse, S. D., Delaney, A., Zhao, Y., *et al.*, 2010. Comparison of sequencing-based methods to profile dna methylation and identification of monoallelic epigenetic modifications.

Hellman, A. and Chess, A., 2010. Extensive sequence-influenced dna methylation polymorphism in the human genome. *Epigenetics Chromatin*, **3**(1):11.

Iida, A., Shimada, A., Shima, A., Takamatsu, N., Hori, H., Takeuchi, K., and Koga, A., 2006. Targeted reduction of the DNA methylation level with 5-azacytidine promotes excision of the medaka fish Tol2 transposable element. *Genetical research*, **87**(3):187–93.

Jiang, L., Zhang, J., Wang, J.-J., Wang, L., Zhang, L., Li, G., Yang, X., Ma, X., Sun, X., Cai, J., *et al.*, 2013. Sperm, but not oocyte, {DNA} methylome is inherited by zebrafish early embryos. *Cell*, **153**(4):773 – 784.

Jirtle, R. L. and Skinner, M. K., 2007. Environmental epigenomics and disease susceptibility. *Nat Rev Genet*, **8**(4):253–62.

Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Dreszer, T. R., Fujita, P. a., Guruvadoo, L., Haeussler, M., *et al.*, 2014. The UCSC Genome Browser database: 2014 update. *Nucleic acids research*, **42**(Database issue):D764–70.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., *et al.*, 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, **447**(7145):714–719.

Kawakami, K., 2007. Tol2: a versatile gene transfer vector in vertebrates. *Genome Biol*, **8**(Suppl 1):1–10.

Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V. V., Schupf, N., Vilain, E., *et al.*, 2008. Genomic surveys by methylation-sensitive snp analysis identify sequence-dependent allele-specific dna methylation. *Nat Genet*, **40**(7):904–8.

Kin, T. and Ono, Y., 2007. Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat. *Bioinformatics*, **23**(21):2945–6.

Koga, A., Shimada, A., Shima, A., Sakaizumi, M., Tachida, H., and Hori, H., 2000. Evidence for recent invasion of the medaka fish genome by the tol2 transposable element. *Genetics*, **155**(1):273.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. a., McCombie, W. R., Jarvis, E. D., *et al.*, 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, **30**:693–700.

Korlach, J., Marks, P. J., Cicero, R. L., Gray, J. J., Murphy, D. L., Roitman, D. B., Pham, T. T., Otto, G. a., Foquet, M., and Turner, S. W., *et al.*, 2008. Selective aluminum passivation for targeted immobilization of single dna polymerase molecules in zero-mode waveguide nanostructures. *Proceedings of the National Academy of Sciences of the United States of America*, **105**:1176–81.

Kumaki, Y., Oda, M., and Okano, M., 2008. Quma: quantification tool for methylation analysis. *Nucleic acids research*, **36**(suppl 2):W170–W175.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., Lohr, J. G., Harris, C. C., Ding, L., Wilson, R. K., *et al.*, 2012. Landscape of somatic retrotransposition in human cancers. *Science*, **337**(6097):967–971.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R., 2008. Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, **133**(3):523–536.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., *et al.*, 2009. Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**(7271):315–22.

Loomis, E. W., Eid, J. S., Peluso, P., Yin, J., Hickey, L., Rank, D., McCalmon, S., Hagerman, R. J., Tassone, F., and Hagerman, P. J., *et al.*, 2012. Sequencing the unsequenceable: Expanded cgg-repeat alleles of the fragile x gene. *Genome Res*, .

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., *et al.*, 2008. Genome-scale dna methylation maps of pluripotent and differentiated cells. *Nature*, **454**(7205):766–70.

Miller, G., 2010. Epigenetics. the seductive allure of behavioral epigenetics. *Science*, **329**(5987):24–7.

Miura, F., Enomoto, Y., Dairiki, R., and Ito, T., 2012. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic acids research*, **40**(17):e136–e136.

Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W. R., Hannon, G. J., and Smith, A. D., 2011. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell*, **146**(6):1029–41.

Muotri, A. R., Chu, V. T., Marchetto, M. C., Deng, W., Moran, J. V., and Gage, F. H., 2005. Somatic mosaicism in neuronal precursor cells mediated by l1 retrotransposition. *Nature*, **435**(7044):903–910.

Muotri, A. R., Marchetto, M. C., Coufal, N. G., Oefner, R., Yeo, G., Nakashima, K., and Gage, F. H., 2010. L1 retrotransposition in neurons is modulated by mecp2. *Nature*, **468**(7322):443–446.

Nautiyal, S., Carlton, V. E., Lu, Y., Ireland, J. S., Flaucher, D., Moorhead, M., Gray, J. W., Spellman, P., Mindrinos, M., Berg, P., *et al.*, 2010. High-throughput method for analyzing methylation of cpgs in targeted genomic regions. *Proc Natl Acad Sci U S A*, **107**(28):12587–92. .

Penzkofer, T., Dandekar, T., and Zemojtel, T., 2005. L1base: from functional annotation to prediction of active line-1 elements. *Nucl. Acids Res.*, **33**:D498–D500.

Qu, W., Hashimoto, S., Shimada, A., Nakatani, Y., Ichikawa, K., Saito, T. L., Ogoshi, K., Matsushima, K., Suzuki, Y., Sugano, S., *et al.*, 2012. Genome-wide genetic variations are highly correlated with proximal dna methylation patterns. *Genome Res*, **22**(8):1419–25.

Ross, J. P., Rand, K. N., and Molloy, P. L., 2010. Hypomethylation of repeated dna sequences in cancer. *Epigenomics*, **2**(2):245–269.

Ross, M., Russ, C., Costello, M., Hollinger, A., Lennon, N., Hegarty, R., Nusbaum, C., and Jaffe, D., 2013. Characterizing and measuring bias in sequence data. *Genome Biology*, **14**(5):R51.

Schadt, E. E., Banerjee, O., Fang, G., Feng, Z., Wong, W. H., Zhang, X., Kislyuk, A., Clark, T. A., Luong, K., Keren-Paz, A., *et al.*, 2012. Modeling kinetic rate variation in third generation dna sequencing data to detect putative modifications to dna bases. *Genome Res*, .

Schilling, E., El Chartouni, C., and Rehli, M., 2009. Allele-specific dna methylation in mouse strains is mainly determined by cis-acting sequences. *Genome Research*, **19**(11):2028–2035.

Schmitz, R. J., Schultz, M. D., Lewsey, M. G., O'Malley, R. C., Urich, M. a., Libiger, O., Schork, N. J., and Ecker, J. R., 2011. Transgenerational epigenetic instability is a source of novel methylation variants. *Science (New York, N.Y.)*, **334**:369–73.

Shoemaker, R., Deng, J., Wang, W., and Zhang, K., 2010. Allele-specific methylation is prevalent and is contributed by cpg-snps in the human genome. *Genome Res*, **20**(7):883–9.

Smith, Z. D., Chan, M. M., Mikkelsen, T. S., Gu, H., Gnirke, A., Regev, A., and Meissner, A., 2012. A unique regulatory phase of dna methylation in the early mammalian embryo. *Nature*, **484**:339–344.

Song, Q., Decato, B., Hong, E. E., Zhou, M., Fang, F., Qu, J., Garvin, T., Kessler, M., Zhou, J., and Smith, A. D., *et al.*, 2013. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PloS one*, **8**(12):e81148.

Su, J., Shao, X., Liu, H., Liu, S., Wu, Q., and Zhang, Y., 2012a. Genome-wide dynamic changes of DNA methylation of repetitive elements in human embryonic stem cells and fetal fibroblasts. *Genomics*, **99**(1):10–17.

Su, J., Yan, H., Wei, Y., Liu, H., Wang, F., Lv, J., Wu, Q., and Zhang, Y., 2012b. Cpg_mps: identification of cpg methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucleic Acids Res*, .

Szpakowski, S., Sun, X., Lage, J. M., Dyer, A., Rubinstein, J., Kowalski, D., Sasaki, C., Costa, J., and Lizardi, P. M., 2009. Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. *Gene*, **448**(2):151–167.

Tubio, J. M. C., Li, Y., Ju, Y. S., Martincorena, I., Cooke, S. L., Tojo, M., Gundem, G., Pipinikas, C. P., Zamora, J., Raine, K., *et al.*, 2014. Extensive transduction of nonrepetitive dna mediated by l1 retrotransposition in cancer genomes. *Science*, **345**(6196).

Wang, T., Guan, W., Lin, J., Boutaoui, N., Canino, G., Luo, J., Celedón, J. C., and Chen, W., 2015. A systematic study of normalization methods for Infinium 450 K methylation data using whole-genome bisulfite sequencing data. *Epigenetics*, (September):00–00.

Weaver, I. C., Cervoni, N., Champagne, F. A., D'Alessio, A. C., Sharma, S., Seckl, J. R., Dymov, S., Szyf, M., and Meaney, M. J., 2004. Epigenetic programming by maternal behavior. *Nat Neurosci*, **7**(8):847–54.

Wilson, A. S., Power, B. E., and Molloy, P. L., 2007. Dna hypomethylation and human diseases. *Biochimica et biophysica acta*, **1775**:138–162.

Xie, W., Schultz, M., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J., Tian, S., Hawkins, R., Leung, D., *et al.*, 2013. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, **153**(5):1134–48.

Yamada, Y., Watanabe, H., Miura, F., Soejima, H., Uchiyama, M., Iwasaka, T., Mukai, T., Sakaki, Y., and Ito, T., 2004. A comprehensive analysis of allelic methylation status of cpg islands on human chromosome 21q. *Genome Res*, **14**(2):247–66.

Zemach, A., McDaniel, I. E., Silva, P., and Zilberman, D., 2010. Genome-wide evolutionary analysis of eukaryotic dna methylation. *Science*, **328**(5980):916–9.

Zhang, X., Davenport, K. W., Gu, W., Daligault, H. E., Munk, a. C., Tashima, H., Reitenga, K., Green, L. D., and Han, C. S., 2012. Improving genome assemblies by sequencing pcr products with pacbio. *BioTechniques*, **53**:61–2.

Zhang, Y., Rohde, C., Tierling, S., Jurkowski, T. P., Bock, C., Santacruz, D., Ragozin, S., Reinhardt, R., Groth, M., Walter, J., *et al.*, 2009. Dna methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS Genet*, **5**(3):e1000438.

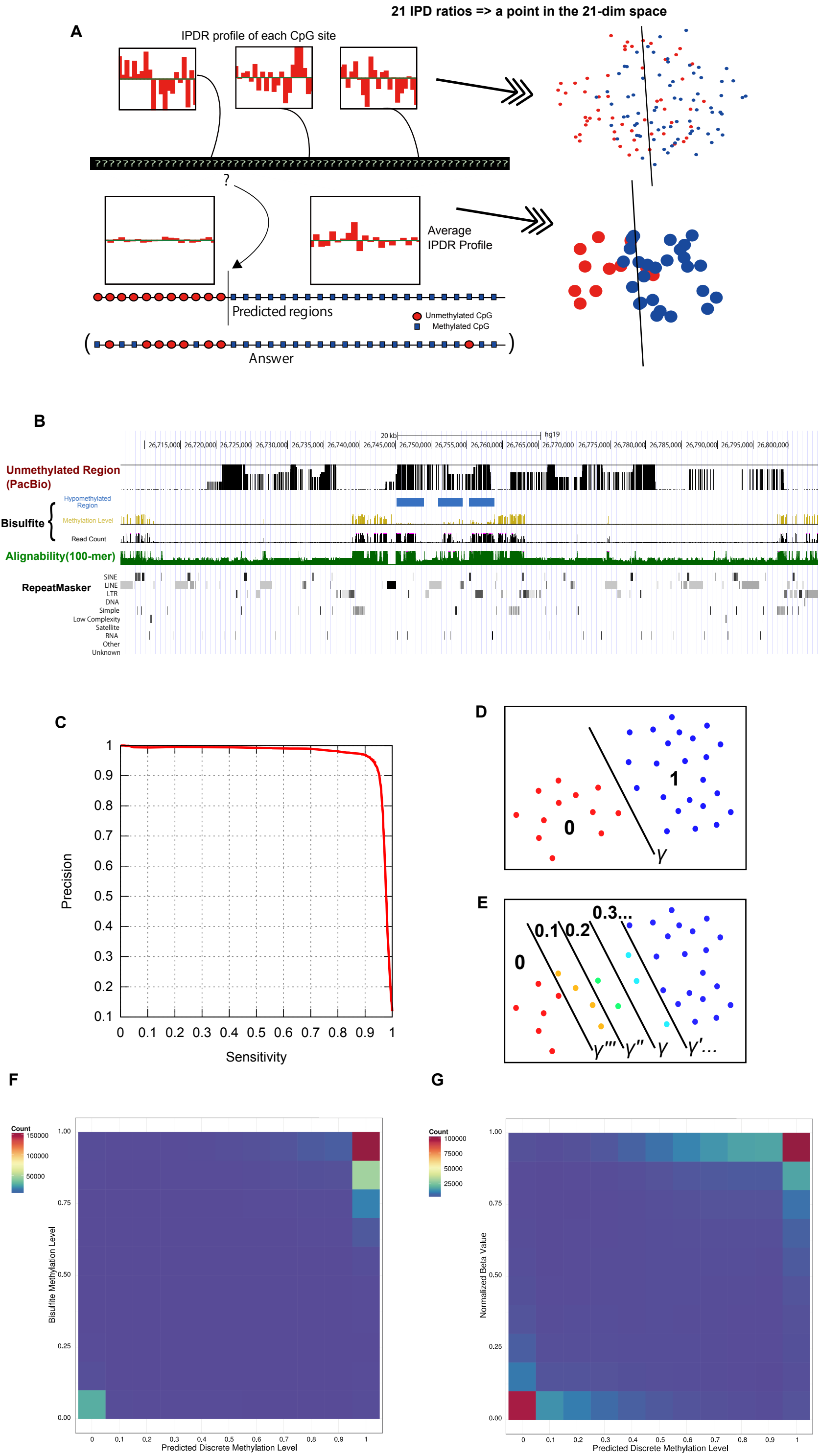
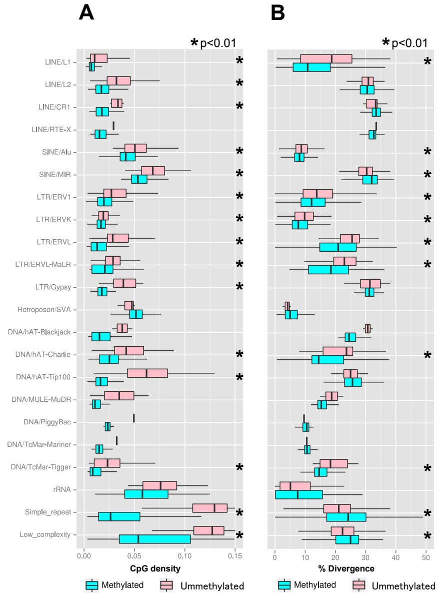
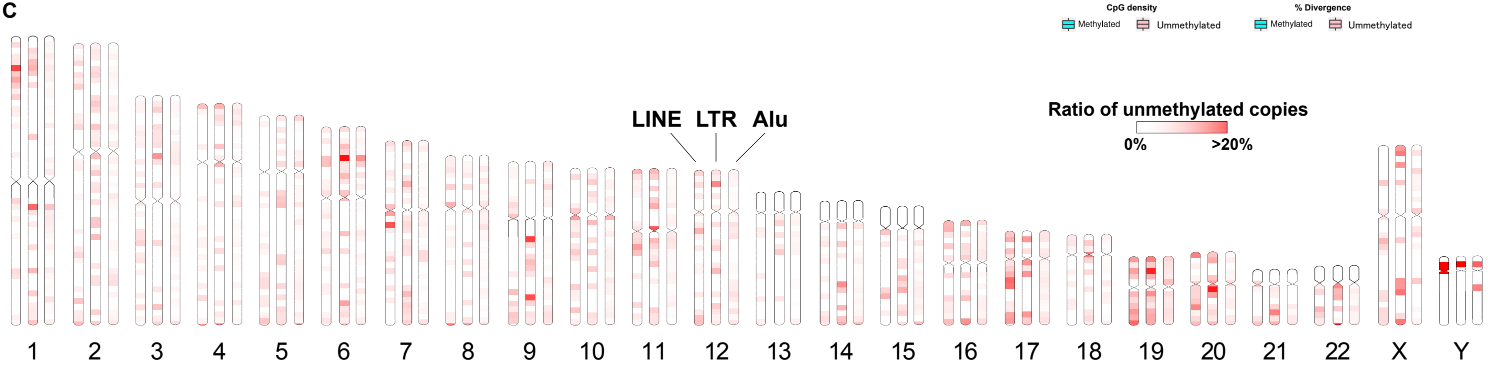
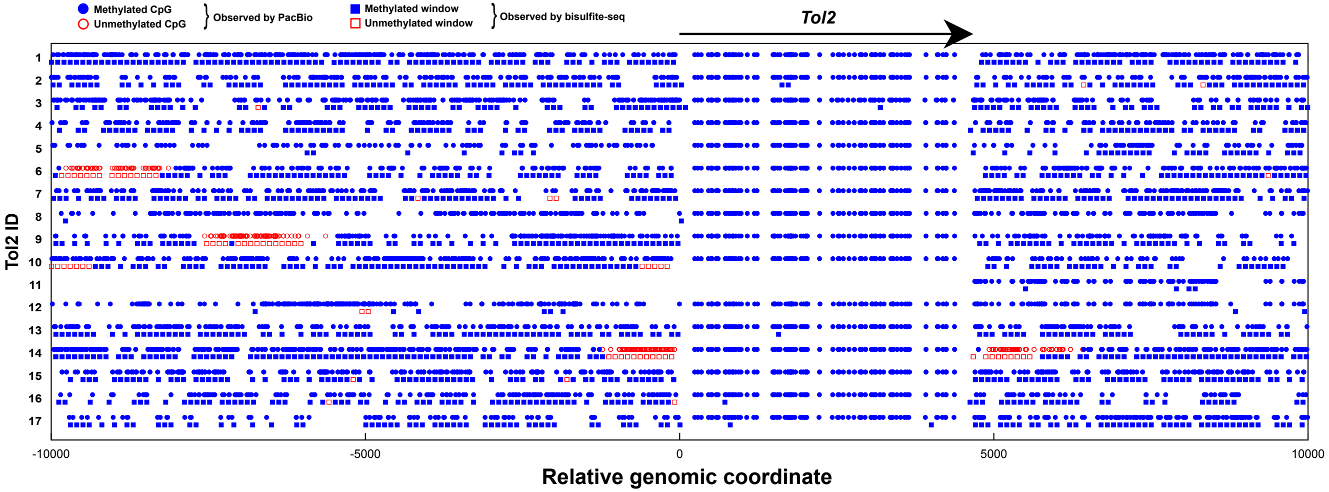

A
21 IPD ratios => a point in the 21-dim space
IPDR profile of each CpG site
??
?
Average
IPDR Profile
Predicted regions
Unmethylated CpG
Methylated CpG
Answer
B
20 kb
hg19
26,715,000
26,720,000
26,725,000
26,730,000
26,735,000
26,740,000
26,745,000
26,750,000
26,755,000
26,760,000
26,765,000
26,770,000
26,775,000
26,780,000
26,785,000
26,790,000
26,795,000
26,800,000
Unmethylated Region
(PacBio)
Hypomethylated
Region
Bisulfite
Methylation Level
Read Count
Alignability(100-mer)
RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
C
Precision
Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
D
0
1
γ
E
0
0.1
0.2
0.3...
γ'''
γ''
γ
γ'...
F
Count
150000
100000
50000
Bisulfite Methylation Level
1.00
0.75
0.50
0.25
0.00
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Predicted Discrete Methylation Level
G
Count
100000
75000
50000
25000
Normalized Beta Value
1.00
0.75
0.50
0.25
0.00
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Predicted Discrete Methylation Level

Table 1. Summary of methylation status on repetitive elements

Class	With >9 CpGs (A)	Covered (B)	B/A	Covered with >5x (C)	C/A	Unmethylated (D)	D/C
LINE/L1	50795	50127	98.7%	45379	89.3%	356	0.8%
LINE/L2	4977	4961	99.7%	4637	93.2%	244	5.3%
LINE/CR1	178	178	100.0%	165	92.7%	5	3.0%
LINE/RTE-X	65	64	98.5%	60	92.3%	1	1.7%
SINE/Alu	238701	235527	98.7%	214341	89.8%	2282	1.1%
SINE/MIR	374	371	99.2%	343	91.7%	169	49.3%
LTR/ERV1	19638	19354	98.6%	17739	90.3%	348	2.0%
LTR/ERVK	5175	5079	98.1%	4603	88.9%	87	1.9%
LTR/ERVL	4395	4350	99.0%	3991	90.8%	82	2.1%
LTR/ERVL-MaLR	4366	4327	99.1%	3933	90.1%	69	1.8%
LTR/Gypsy	108	104	96.3%	89	82.4%	9	10.1%
Retroposon/SVA	2906	2796	96.2%	2427	83.5%	3	0.1%
DNA/hAT-Blackjack	83	83	100.0%	75	90.4%	2	2.7%
DNA/hAT-Charlie	1460	1452	99.5%	1342	91.9%	55	4.1%
DNA/hAT-Tip100	326	322	98.8%	305	93.6%	19	6.2%
DNA/MULE-MuDR	92	92	100.0%	89	96.7%	2	2.2%
DNA/PiggyBac	57	55	96.5%	52	91.2%	1	1.9%
DNA/TcMar-Mariner	384	384	100.0%	360	93.8%	1	0.3%
DNA/TcMar-Tigger	2821	2801	99.3%	2649	93.9%	43	1.6%
rRNA	68	66	97.1%	66	97.1%	8	12.1%
Simple_repeat	6256	6191	99.0%	5434	86.9%	3849	70.8%
Low_complexity	1068	1064	99.6%	942	88.2%	789	83.8%







● Methylated CpG
○ Unmethylated CpG
} Observed by PacBio
■ Methylated window
□ Unmethylated window
} Observed by bisulfite-seq
Tol2
Tol2 ID
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
-10000
-5000
0
5000
10000
Relative genomic coordinate

Legends for supplemental tables and figures

Supplemental Table S1. Statistics of SMRT sequencing data production

Summary statistics of SMRT sequencing data collected in this study.

Supplemental Table S2. 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 Supplemental Figure S4, 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.

Supplemental Table S3. 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.

Supplemental Figure S1. The normal vector used for prediction

A. The normal vector β used for prediction. We calculated β 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 β. **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.

Supplemental Figure S2. Accuracy measures 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 $\gamma = -0.540$.

Supplemental 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 $b = 45$.

Supplemental Figure S4. 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.

Supplemental Figure S5. Kernel PCA analysis of sequence feature and methylation state

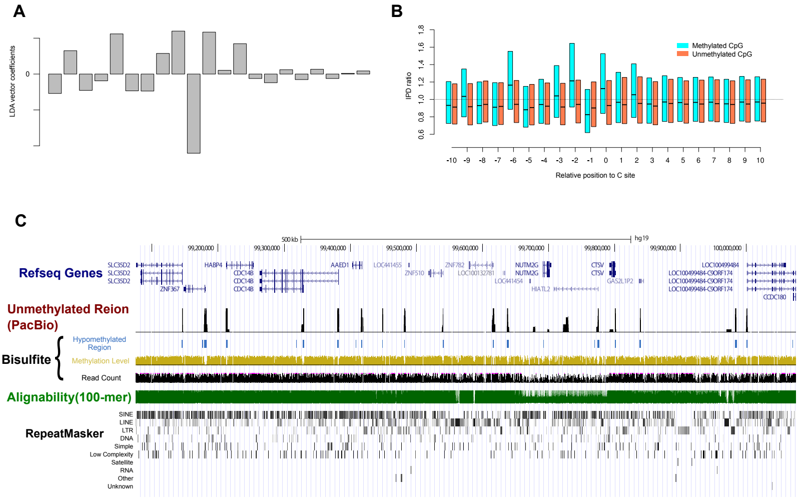
The results of Kernel PCA analysis are shown for 4 selected classes of repetitive elements, AluSc **(A)**, LTR12E **(B)**, LTR26E **(C)**, and L2a **(D)**. We projected the repeat occurrences into the plane based on the distance metrics that we defined using the spectrum kernels and their top 2 principal 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.

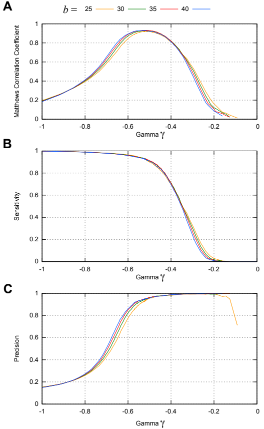
Supplemental Figure S6. Examples of unmethylated repeat occurrences in a unmethylation 'hot spot'

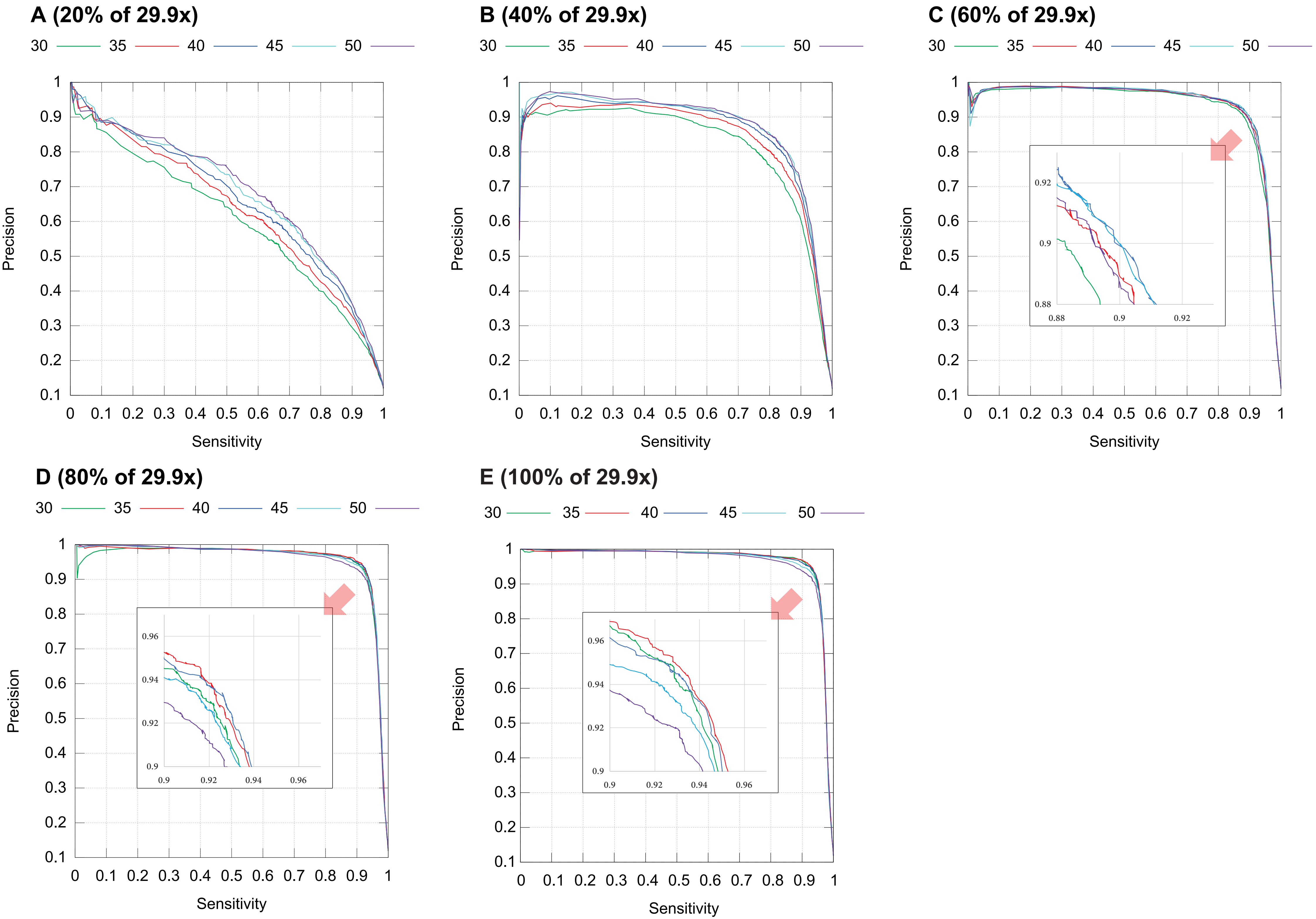
Three adjacent LTR1 elements were unmethylated in this region **(A)**, and a LTR12E element was located at a unmethylated bi-directional promoter region **(B)**. Both regions are on the p-arm of the chromosome 6. The arrows indicate the locations of LTR1 and LTR12E. From top to bottom, below the RefSeq 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.

Supplemental Figure S7. 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.


A
LDA vector coefficients
0
B
IPD ratio
0.6
0.8
1.0
1.2
1.4
1.6
1.8
Methylated CpG
Unmethylated CpG
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
Relative position to C site
C
500 kb
hg19
99,200,000
99,300,000
99,400,000
99,500,000
99,600,000
99,700,000
99,800,000
99,900,000
100,000,000
Refseq Genes
SLC35D2
SLC35D2
SLC35D2
ZNF367
HABP4
CDC14B
CDC14B
CDC14B
AAED1
LOC441455
ZNF510
ZNF782
LOC100132781
LOC441454
NUTM2G
NUTM2G
HIATL2
CTSV
CTSV
GAS2L1P2
LOC100499484
LOC100499484-C9ORF174
LOC100499484-C9ORF174
LOC100499484-C9ORF174
CCDC180
Unmethylated Reion (PacBio)
Bisulfite
Hypomethylated Region
Methylation Level
Read Count
Alignability(100-mer)
RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown


A
$b =$ 25 30 35 40
Matthews Correlation Coefficient
1
0.8
0.6
0.4
0.2
0
-1
-0.8
-0.6
-0.4
-0.2
0
Gamma γ
B
Sensitivity
1
0.8
0.6
0.4
0.2
0
-1
-0.8
-0.6
-0.4
-0.2
0
Gamma γ
C
Precision
1
0.8
0.6
0.4
0.2
0
-1
-0.8
-0.6
-0.4
-0.2
0
Gamma γ


A (20% of 29.9x)
B (40% of 29.9x)
C (60% of 29.9x)
D (80% of 29.9x)
E (100% of 29.9x)
30
35
40
45
50
Precision
Sensitivity
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.88
0.9
0.92
0.94
0.96

```
C                                  C
* : Methylated CpG                 # : Unmethylated CpG
C                                  T

C                                  C
! : Unconverted CpH                : : Converted CpH
C     (CpA, CpC, CpT)              T     (CpA, CpC, CpT)
```

Sequence ID #5
Prediction: Hypomethylated
Region: chrX:17,366,059-17,366,763

```
Genome      301  GGGAGTGACCCAATTTTCCAGGTGCCGTCCATCACCCCTTTCTTTGACTAGGAAAGGGAA
                                                                    | ||||||
Bisulfite     1  ---------------------------------------------------GTAAGGGAC

Genome      361  CTCCCTGACCCCTTGCGCTTCCCGAGTGAGGCAATGCCTCGCCCTGCTTCGGCTCGCGCA
                 :|:::|| ::::|  #|:||: #||||||||:||||::  |:::||:| #||: #|#|:|
Bisulfite    10  TTTTTTGMTTTTTYSTGTTTTYTGAGTGAGGTAATGTTCYGTTTTGTTCTGGTCTGTGTA

Genome      421  CGGTGCGTGCACCCACTGACCTGCGCCCACTGTCTGGCACTCCCTAGTGAGATGAACCTG
                 *||||*|||:|:::|:|||::||*|:::|:|||:|||:|:|:::||||||||||||::||
Bisulfite    70  CGGTGCGTGTATTTATTGATTTGCGTTTATTGTTTGGTATTTTTTAGTGAGATGAATTTG

Genome      481  GTA-CCTTAGATGGAAATGCAGAAATCACCGGTCTTCTGCGTCGCTCACGCTGGGAGCTA
                 ||| ::|||||||||||||:||||||:|!*|| :||     |* ! ! # !|
Bisulfite   130  GTATTTTTAGATGGAAATGTAGAAATTACCGGYTTT-----TCCCSCCTCCT--------
```

Sequence ID #7
Prediction: Hypermethylated
Region: chr6:123,793,104-123,793,890

```
Genome      301  CCTCAGTCGGGAAGTGCAAGGGGTCAGGGAGTTCCCCTTCCGAGTCAAAGAAAGGGGTGA
                                     || |:||||||||::::||: ||||:||||||| ||| ||
Bisulfite     1  -------------------AGGKTTAGGGAGTTTTTTTTTYGAGTTAAAGAAA-GGGCGA

Genome      361  CGGACAGCACCTGGAAAATCGGGTCACTCCCACCCGAATACTGCGCTTTTCCGACAGGCT
                 *|||:  :|::||||||||* |||:|:|::::|::*|||||:||*|:||||:*||:||| |
Bisulfite    41  CGGATM-TATTTGGAAAATCAGGTTATTTTTATTCGAATATTGCGTTTTTTCGATAGG-T

Genome      421  TAAAAAACGGCGCACCACAAGATTATATCCCACACCTGGCTCGGAGGGTCCTACGCCCAC
                 |||||||*||*|:|::|:||||||||||:::|:| :|||:|*|||||||::||*| ::|*
Bisulfite    99  TAAAAAACGGCGTATTATAAGATTATATTTTATA-TTGGTTCGGAGGGTTTTACG-TTAC

Genome      481  GGAATCTCGCTGATTGCTAGCACAGCAGTCTGAGATCAAACTGTAAGGCGGCAGC-AAGG
                 |||||:|*|:||||||:|||:|:||:|||:||||||:|||:|||||||*||:||: || |
Bisulfite   157  GGAATTTCGTTGATTGTTAGTATAGTAGTTTGAGATTAAATTGTAAGGCGGTAGTAAASG

Genome      540  CTGGGGGAGGGGCGCCCGCCATTGCCCAGGCTTTCTTAGGAAAACAAAGCAGCCGGGAAG
                 :|||||  ||| * !
Bisulfite   217  TTGGGGAMGGGKCCC---------------------------------------------
```

Sequence ID #8
Prediction: Hypomethylated
Region: chr11:5,829,621-5,830,339

```
Genome      181  CTATCCTTCACTGGAATCGTAACTGAGGCT--CAATTCGCCTATCCTTTAGCCCCACCT-
                                            |:|  !|                 :!! |:!|
Bisulfite     1  ---------------------------GTTGGCA-----------------TCCMATCTA

Genome      238  --GCTGGAGGCTCTTTGCATCCTTTCGCTTTGTCCACTCTGGCCGCTTCCCTCGTGGGAA
                    :||||||:| ||||:||::|||#|:|||||::|:|:|||:#|:||:::|#|||||||
Bisulfite    17  TWRTTGGAGGTTYTTTGTATTTTTTTGTTTTGTTTATTTTGGTTGTTTTTTTTGTGGGAA

Genome      296  TATTTCAGGTTCCTCTTAGCCTTGATGGCGGGTCAGCATAAACCCCTGAT-GGGACCCCC
                 |||||:|||||::|:||||::||||||| ||||:||:|||||::::|||| ||||:!
Bisulfite    77  TATTTTAGGTTTTTTTTAGTTTTGATGGYGGGTTAGTATAAATTTTTGATKGGGATC---
```

C
* : Methylated CpG
C

C
: Unmethylated CpG
T

C
! : Unconverted CpH (CpA, CpC, CpT)
C

C
: : Converted CpH (CpA, CpC, CpT)
T

Sequence ID #9

Prediction: Hypomethylated

Region: chr1:89,663,480-89,664,077

```
                  GTTACAGGAAAGTAAACAGTACTAGGTGCAGGGGCTTTAATTCTATCA-CAAGGTGATAG
                                     ||!||            ||   |:|| :|  ||||||||||
Bisulfite    129  -------------------TACTA------------TTWMWTTTAWTAWWAAGGTGATAG

Genome       352  AAGCGGGGCTTTGGGCTTTATCAACCAGACACAAACGCGGGGGGCTCTGGGTGCTGTTAA
                  ||| ||||:||||||:|||||:||::|||:|:|||#| ||||||:|:||||||:|| |||
Bisulfite    158  AAGWGGGGTTTTGGGTTTTATTAATTAGATATAAATG-GGGGGGTTTTGGGTGTTGCTAA

Genome       412  CCGGGCGAAT-TCCTGGGAACTGCGGGTATGGCTTGCCACAGTACCTTATCAGTTAATTG
                  :#||| |||  |!! |
Bisulfite    217  TTGGGYGAAYCTCCCG--------------------------------------------
```

Sequence ID #11

Prediction: Hypomethylated

Region: chr19:11,848,508-11,850,380

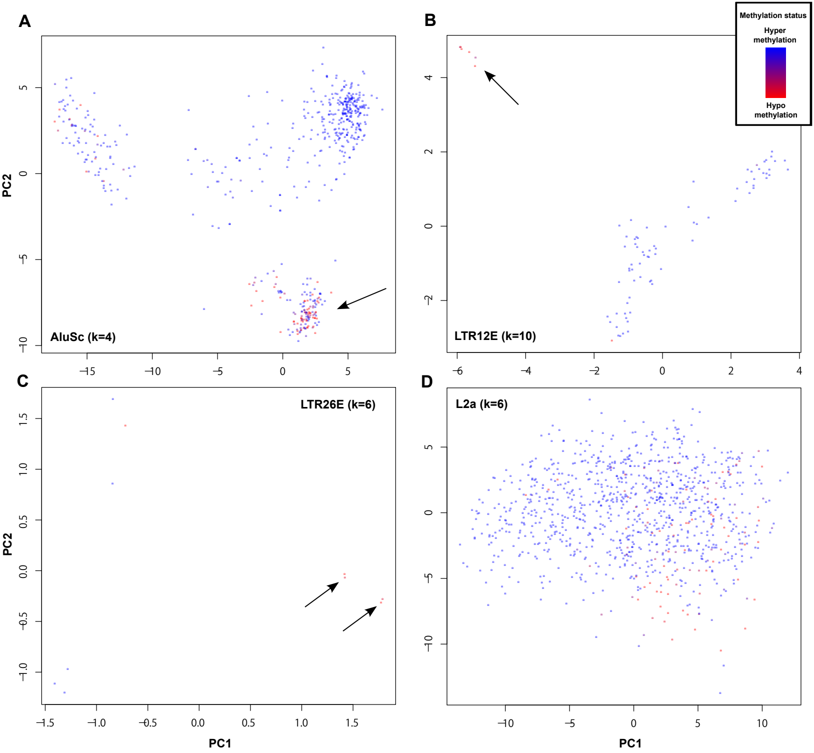
```
Genome       661  ACCAGCGACCCCACACTCCAGCCGTCCCTGTCCACACCTCTAAACACCCCATCCCCAAAC
                          ! !:|:|! !!  !*|  !!                  :::: |::::|||:
Bisulfite      1  --------CYCTATAC-CCCCCCGCYCC------------------TTTTTTTTTTAAAT

Genome       721  CTCTCAGGGAGGCGGATCTGGGGTGTCCTCCCCTCTCCCCCATTAAACTGTTTCTGCTGC
                  :|:|:   |  | ||  :||||| | ::|::::|:|:::::|| |||:|||||:||:||:
Bisulfite     34  TTTTTTKKGRRGKGGGGTTGGGGGGKTTTTTTTTTTTTTTTATCAAATTGTTTTTGTTGT

Genome       781  AGCCTTCGGCGTCTCGGTGCAGTGACTCGGGCCGTGAACCTGTGCCGGTTACAACTGCAC
                    ::||#||#||:|#||||:|||||:|#|||:#|||||::||||:#|||||:||:||:|:
Bisulfite     94  AGTTTTTGGTGTTTTGGTGTAGTGATTTGGGTTGTGAATTTGTGTTGGTTATAATTGTAT

Genome       841  AATCTGGGGAGACGCGGAGCTGCGGGCGCGGAGCTGCCCAGAGAGGGCGCCGGGGCCGGG
                  |||:||||||||#|#||||:||#|||#|#||||:||:::||||||||#|:#||||:#|||
Bisulfite    154  AATTTGGGGAGATGTGGAGTTGTGGGTGTGGAGTTGTTTAGAGAGGGTGTTGGGGTTGGG

Genome       901  GCCGCAGCGGCCGAGCAGGGACGGGACAGGACGCCCGGGGTCCCGGCTGCCGCCCCAGCC
                  |:#|:||#||:#|||:|||||#||||:||||#|::#|||||::#||:||:#|::::|| :
Bisulfite    214  GTTGTAGTGGTTGAGTAGGGATGGGATAGGATGTTTGGGGTTTTGGTTGTTGTTTTAG-T

Genome       961  CCATCTTGCGGCCCA-GGGGACCAAGGGCAGAGCTGCGCCAGGGGCACTGGGATTTGCAG
                  ::||:| |  |:::| |  | !!||   !       *
Bisulfite    273  TTATTTYGYSGTTTACGCCGMCCAAMCSC----YYMC-----------------------
```

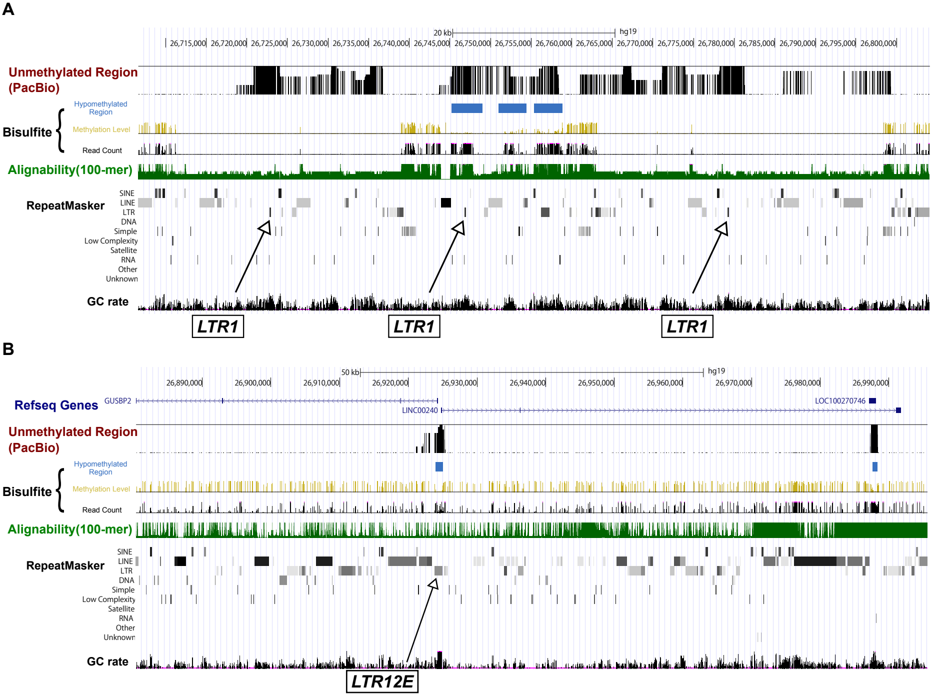
Sequence ID #15

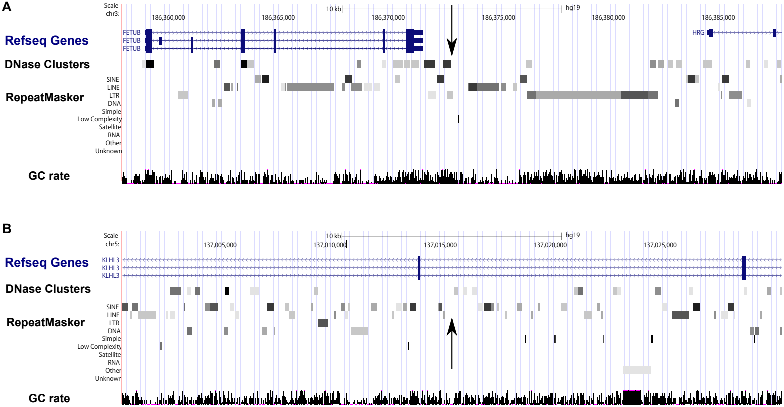
Prediction: Hypermethylated

Region: chr19:47,905,568-47,906,031

```
Genome         1  TCTCTCTCTGGGGGGTGGAGGGGACAGAGATCTGGAAAACTGAGAACCCCAAGGGACTCA
                                 |||| | ||! |     |||||||:||||||::::||||||:|:|
Bisulfite      1  -------------CCTGGACGAGACCG-----TGGAAAATTGAGAATTTTAAGGGATTTA

Genome        61  CACTGGTTTCTGAGCCTCAGTTTTCCTAGTTACAAAGGACAGCCTCTGCCTGTGATGGGC
                  :|:||||||:||||::|:||||||::||||||:||||||:||::|:||::|||||||||
Bisulfite     43  TATTGGTTTTTGAGTTTAGTTTTTTTAGTTATAAAGGATAGTTTTTGTTTGTGATGGGG

Genome       121  GCTGACACACGTGGCACAGTTCCCCATGTGTCCCTCGAAATACCTCCACCATCAGCACAA
                  |:|||:|:|*||||:|:||||::::||||||:::|*||||||::|::|::||:||:|:||
Bisulfite    103  GTTGATATACGTGGTATAGTTTTTTATGTGTTTTTCGAAATATTTTTATTATTAGTATAA

Genome       181  TCATCCTACGAGACAGGCACGGCCGCTCTCCCCATTCTCCAGATGTGGAAACCGGGGCCC
                  |:||::||*||||:|||:|*||:*|:|:|::::|||:|::|||||||||||:*||||:::
Bisulfite    163  TTATTTTACGAGATAGGTACGGTCGTTTTTTTTATTTTTTAGATGTGGAAATCGGGGTTT

Genome       241  AGCCAGGTGAAGTCGTAA-CCCGAGGTGCCA-TAGCTGTTGCGTTCCAGAGGCGAGA-TT
                  ||::|||||||||*|||| ::*||||||::| |||:|||||*|||::|||||*|||| ||
Bisulfite    223  AGTTAGGTGAAGTCGTAATTTCGAGGTGTTATTAGTTGTTGCGTTTTAGAGGCGAGATTT

Genome       298  CAAACCC--ACGTCCGTCCGGAAGCCTTGGAAGTGAGGGTGTGCCTGCCTAACCTGCTCA
                  :||| ::     |:*
Bisulfite    283  TAAAWTTWAWYSTTC---------------------------------------------
```


A
B
C
D
AluSc (k=4)
LTR12E (k=10)
LTR26E (k=6)
L2a (k=6)
PC1
PC2
Methylation status
Hyper methylation
Hypo methylation
−15
−10
−5
0
5
−6
−4
−2
0
2
4
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
−10
−5
0
5
10
−10
−5
0
5
−2
0
2
4
−1.0
−0.5
0.0
0.5
1.0
1.5
−10
−5
0
5


A
20 kb
hg19
26,715,000
26,720,000
26,725,000
26,730,000
26,735,000
26,740,000
26,745,000
26,750,000
26,755,000
26,760,000
26,765,000
26,770,000
26,775,000
26,780,000
26,785,000
26,790,000
26,795,000
26,800,000
Unmethylated Region (PacBio)
Bisulfite
Hypomethylated Region
Methylation Level
Read Count
Alignability(100-mer)
RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
GC rate
LTR1
LTR1
LTR1
B
50 kb
hg19
26,890,000
26,900,000
26,910,000
26,920,000
26,930,000
26,940,000
26,950,000
26,960,000
26,970,000
26,980,000
26,990,000
Refseq Genes
GUSBP2
LINC00240
LOC100270746
Unmethylated Region (PacBio)
Bisulfite
Hypomethylated Region
Methylation Level
Read Count
Alignability(100-mer)
RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
GC rate
LTR12E


A
Scale
chr3:
10 kb
hg19
186,360,000
186,365,000
186,370,000
186,375,000
186,380,000
186,385,000
Refseq Genes
FETUB
FETUB
FETUB
HRG
DNase Clusters
RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
GC rate
B
Scale
chr5:
10 kb
hg19
137,005,000
137,010,000
137,015,000
137,020,000
137,025,000
Refseq Genes
KLHL3
KLHL3
KLHL3
DNase Clusters
RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
GC rate

Statistics for medaka SMRT sequencing (P6)	
Number of SMRT cells	38
Number of mapped subreads	2,848,641
Mean mapped subread length (b)	8,722
Total bases of mapped subreads (b)	24,846,441,723
Coverage (medaka genome size = 800 Mb)	31.06

Statistics for human SMRT sequencing (P6)	
Number of SMRT cells	111
Number of mapped subreads	7,279,594
Mean mapped subread length (b)	9,254
Total bases of mapped subreads (b)	67,364,373,129
Coverage (human genome size = 3 Gb)	22.45

Statistics for human SMRT sequencing (P4)	
Number of SMRT cells	377
Number of mapped subreads	19,115,712
Mean mapped subread length (b)	2,049
Total bases of mapped subreads (b)	39,177,531,604
Coverage (human genome size = 3 Gb)	13.06

Sequence_Id	Sequence_Name	Region	Primer_Sequence (5' to 3')	Primer Sequence (for nested PCR) (5' to 3')
1	FLI_CHR11_F	chr11:92,869,695-92,870,491	TGTATGAGTATGTTTAGTGT	TATATGGGGGAGGAGTTAAGATGGT
	FLI_CHR11_R		CTACTATCTTTTTATTTATCTATACCC	CACCCCTTTCTTTAACTCAAAAAAAA
2	FL_CHR7_F	chr7:34,945,686-34,946,552	GAATATATGAGTAAATGAAGGATGT	TTTTATTAGGGAGTGTTAGATAGTGGG
	FL_CHR7_R		TACCCCCAAAAATAAAAACT	TATTAAATACCCCTCCCCCAACCTC
3	FL_CHR8_F	chr8:98,307,662-98,308,329	TTGGTATTTGTAAGAAATTAGGGA	TGATTTTTGTATTTTTATTTGAGGTAT
	FL_CHR8_R		CTTACACTTCCCAAATAAAACAA	ATAAAACAATACCTCACCCTACTTC
4	FL_CHR19_F	chr19:35,351,062-35,351,674	AAGATATTTATTTAAGGAGGAG	GGTGATTTTTGTATTTTTAGTTGAGGTAT
	FL_CHR19_R		ACCTAATCAAACCTAAACAATAAC	AAAAAAAACTCCCTAACCCCTTAC
5	FL_CHRX_F	chrX:17,366,059-17,366,763	GGTGGGAGTGATTTAATTTTTTA	GAAAGGGAATTTTTTGATTTTTTG
	FL_CHRX_R		ACTTTATTTATACAACTTCTATTC	AAAACAACTCTAATCTATAACTCCCAAC
6	FL_CHR1_F	chr1:90,218,123-90,218,684	TTTGGTTGTTTTGTTTATTTAAGT	TTAAGTAAGTTTGGGTAATGGTGGG
	FL_CHR1_R		ATCTCTTAAATACCTTAACC	AAAAAATCAAAAAATTCCCTTTCC
7	FL_CHR6_F	chr6:123,793,104-123,793,890	GGAAGGATAAATAGTTTAATAAAGG	TTATTAGGGAATGTTAGATAGTGGG
	FL_CHR6_R		CCCAAAAATAAAACCTACAAAA	TTTCCTAAAAAAACCTAAACAATAAC
8	LTR_CHR11_F	chr11:5,829,621-5,830,339	TTGTAATATTTTTTATATTGGG	TTGTTGGAGGTTTTTTGTATTTTTT
	LTR_CHR11_R		AAAAAATCTTCAATCATCCT	ATCCAATCTATAATTCTATAATCACCTCAT
9	LTR_CHR1_F	chr1:89,663,480-89,664,077	AAATTTTTGTTTTTTGGAGTTTTA	TTTTAATTTTATTATAAGGTGATAGAAG
	LTR_CHR1_R		TAACTCTCCCTTAACTAAAA	TAAAATACTATAACAAACCATACCC
10	LTR_CHR6_F	chr6:26,924,100-26,924,635	AAGTTTTTTAAAGTTTTTATTAG	
	LTR_CHR6_R		ACCTACCATACTAAAACCCT	
11	LTR26C_CHR19_F	chr19:11,848,508-11,850,380	AGGTTGAAGATTTTATAAGGGAA	TTTTATTTTTAAATTTTTTAGGGAGG
	LTR26C_CHR19_R		TAAAACCCACACTAACTTTT	ATCTACAAATCCCAATACCCCTAAC
12	AluSc_F	chr15:80,352,853-80,353,500	AGTTGTAATTAGTTGTGAGGAAGT	TTGTAATTTTAGTATTTTGGGAGGT
	AluSc_R		CCCCTAAAACTCTAAAAAAA	AAAAAAAATTTAATCCTATTTCTC
13	AluSc2_F	chr12:11,667,503-11,668,057	TTTTGAGTGTTTTTGGTTTTGGA	TTTTTATGTTAAGAATAGTTTTGGT
	AluSc2_R		TTTCTCTATTTTTCAACTATTCACC	TAAAATAAAATCTCTCTCTATCACC
14	MIR_F	chr8:145,157,891-145,158,610	GAGGAGTAAAGAAATATAAG	AGTTTTGTAAAGTAGGTTTAGGTAGGTTTT
	MIR_R		RAAACCCAAAATTAAACCCCT	CCACCTAAATACCCTTAAACAATTATATTT
15	MIR2_F	chr19:47,905,568-47,906,031	GGATAGAGATTTGGAAAATTGA	TGGAAAATTGAGAATTTTAAGGGATTTATA
	MIR2_R		AAACAAATTAAACAAACACACCC	AACACACCCTCACTTCCAAAACTTC
16	LTR26E_CHR19_F	chr19:12,510,921-12,511,712	TTTTGGAAAGAAAGAAGGGAT	
	LTR26E_CHR19_R		CATTCTACTAAATAAACTCC	AATTCAAAATCTAAAAACTCCRAC
17	LTR26E_CHR1_F	chr1:245,287,681-245,288,390	AGGATTAAGAAGAATTTTGGA	AGAAATATAGTTGTAAGTAAG
	LTR26E_CHR1_R		CTATAAAACACAACAAACTTAACC	AAATCCATCTCCCTAAAAAA
18	L2b_CHR9_F	chr9:137,028,161-137,028,803	ATATTTTGTAGTTATTTTTGA	ATTTTTTTTAGAATTTAGGG
	L2b_CHR9_R		ACACCCAAATCCAAATCCAAA	CCCAAAAACCTATACAAAAA
19	L2b_CHR8_F	chr8:141,097,083-141,098,123	GGAGTATAGATGGAATATTAATAG	TTGTTTAGGTTAAAATTTTAAAAGATATTT
	L2b_CHR8_R		CACAACTAATACAAAAACCCAAA	TACAAAAACCCAAAAATAATACCAC
20	AluY_CHR1_F	chr1:202,975,549-202,976,334	TATTTTGTAAGTTTAGGGGTGTT	GTGTTTTTTTTGTTTGTGGT
	AluY_CHR1_R		CCCACACCTATATTAATTAAAA	TTCTTCTTAATAACTCCTCT
21	LSUrRNAHsa_F	chr9:79,186,495-79,187,160	AGTTTTTTATTGGGATGGTATGT	TTTTTTAATAATAATGAGATGGGG
	LSUrRNAHsa_R		CCCTTTCTTTTTTTCTTTTTTC	

Class	With >9 CpGs (A)	Covered with SMRT reads (B)	B/A	Covered with >5x (C)	C/A	Unmethylated (D)	D/C
non-intact L1s due to multiple mutations	11646	11497	98.72%	10465	89.86%	64	0.61%
L1s with an intact ORF2 but a disrupted ORF1	103	100	97.09%	91	88.35%	0	0.00%
L1s with intact in the two ORFs	145	133	91.72%	119	82.07%	0	0.00%