bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# 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], Hideaki Yurino [1], Wei Qu [1], Jun Yoshimura [1], Yuji Takahashi [4], Jun Mitsui [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 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. Here, 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 when they are correlated. Both the sensitivity and precision of our algorithm were $>$84% for the genome of an inbred medaka (*Oryzias latipes*) strain within a practical read coverage of $<$18-fold. Using this method, we characterized the landscape of the methylation status of repetitive elements, such as LINEs, in the human genome, thereby revealing the strong correlation between CpG density and hypomethylation and detecting hypomethylation hot spots of LTRs and LINEs. We also comprehensively evaluated the methylation states for nearly identical ($>$ 99.8%) active transposons 4682 base pairs (bp) in length in the medaka genome, which were difficult to observe using bisulfite-treated short reads.**

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# 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 [1, 2, 3, 4, 5, 6, 7, 8, 9]. Many human diseases are associated with the disruption of DNA modifications. In particular, hypomethylation of repetitive elements, such as LINE-1 elements, has been related to some cancers [10, 11]. Although only a few LINE-1 elements exhibit activity in the human genome [12], transpositions of these elements have been reported in various cancer genomes [13, 14], and importantly, it has been reported that transpositions are correlated with hypomethylation of the promoter region of LINE-1 elements [15]. 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 [16, 17, 18, 19, 20]; 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 with multiple positions. Moreover, first/second-generation sequencing technology often fails to sequence DNA regions with a GC content $>$60% [21, 22] 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 [19, 20, 23, 24, 25]. 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 [26, 27], and this system is capable of sequencing reads of an average length of $\sim$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% [28], 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 [29].

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.

Consequently, the IPD ratio (IPDR), the ratio of the average IPD in DNA tem-

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

plates 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 [30], N4-methylcytosine [31], and N6-methylademine [30, 32, 33], as well as damaged DNA bases [34] 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 [30, 32, 35]. 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 [36]. Kinetic information for 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 (∼10%) in vertebrates and generally do not exist in isolation but often range over long hypomethylated regions [9, 24, 37, 38, 39, 40, 41]. Su *et al.* reported that the average length of unmethylated regions in five human cell types is ∼2 kb [42]. Thus, estimating regions of hypomethylated 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. Therefore, 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.

SMRT sequencing is unique in outputting long reads, which has been shown to be useful in a variety of applications, such as sequencing completion of bacterial genomes [43, 44], closing gaps in draft genomes [45], the *de novo* assembly of unknown genomes [46], sequencing of giant short tandem repeats (*e.g.*, CGG repeats) [28], and comprehensive characterization of mRNA isoforms [47]. 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 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 [39, 48, 49, 50, 51, 52, 53], because we may observe an intermediate DNA methylation level resulting from the mixture of different methylation states from two haplotypes [19, 54]. 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

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

medaka model system [55], because six medaka methylomes are available from early embryos, testes, and liver in two inbred strains [9] 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 [20]. 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 hypo- or hypermethylated, and methylation at non-CpG sites is very rare ($\sim$0.02%) [9], allowing us to focus on CpG sites only. We collected 25.87-fold coverage SMRT subreads from the testes of medaka Hd-rR (using 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. For sequencing two human samples, we employed the P6-C4 reagents and the P4-C2 or C2-C2 reagents, respectively. In total 2596378, 7279594, and 19115712 subreads were anchored to the medaka genome and the human genome, respectively. The mean mapped subread lengths were 7972 bases for medaka and 9254 and 2049 bases for human samples (Supplemental Table 1).

## 2.2 Prediction of the regional 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 hyper/hypomethylated 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: beta (orientation) and gamma (intercept). As hypomethylated regions are $\sim$2 kb in size (on average) and contain $\sim$50 CpG sites in vertebrate genomes [42], in the prediction, we assumed unmethylated regions to have at least 50 CpG sites and integrated the IPDR profiles to make predictions, which was effective in reducing noise in the IPD measurements. Finally, our method divides the genome into regions containing $\geq$50 CpG sites, such that each region is either hypomethylated or hypermethylated. An example of our prediction for the human genome is shown in Fig. 1b, in which gene promoter hypomethylation was captured correctly.

To determine whether our strategy is effective and its dependence on the amount of data available, we performed predictions using five medaka datasets with different read coverages, and determined the depth of coverage that would be sufficient to correctly detect unmethylated CpG sites. We calculated various accuracy measures, such as sensitivity (recall), specificity (1$-$false-positive rate), and precision by comparison our prediction on each CpG site with the methylation level determined in a bisulfite sequencing study [9]. As most CpG sites in the medaka genome are methylated con-

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sistently, 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 gamma (the intercept of the decision hyperplane) (Supplemental Fig. S2). Our prediction achieved 82.7% sensitivity and 86.6% precision with a 22.2-fold mapped read coverage (Fig. 1c).

## 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 hypomethylated or hypermethylated regions; however, such a dichotomous model is rather unrealistic, and more refined predictions involving multi-level 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 this concept for multi-class prediction. We made a classification using the linear discrimination process involving a separation (decision) hyperplane and determined the position of the hyperplane using the gamma parameter (Fig. 1d). Intuitively, the intermediately methylated CpGs are expected to be distributed more closely to the decision plane, and are therefore more ambiguous than the are CpGs with *bona fide* methylation states. Thus, to output the multi-class prediction, we perturbed gamma 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* hyper/hypomethylation should have extreme DML values, unlike intermediate or ambiguous predictions.

We empirically checked the accordance between the DMLs and intermediate or ambiguous methylation states. We compared our predictions made on our human sample (represented as DMLs) to the beta value (an indicator of methylation level expressed as a value ranging over [0,1]) obtained by Illumina BeadChip analysis (Fig. 1f). We divided DMLs and beta values into 10 bins and assigned all CpG sites with DMLs and beta values available into a pair of classes of corresponding bins. In Figures 1f-g, the width and height of each box are proportional to the relative number of CpG sites in the bin of DML or beta value, highlighting the strong correlation between DMLs and beta values. Similar analysis of the medaka genome also showed the expected correlation (Fig. 1g), although the number of CpGs with ambiguous methylation status was much lower than that in the human case, presumably because the medaka sample was collected from an inbred strain [9].

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## 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 hypomethylated if $\geq$50% of CpGs were predicted as hypomethylated. Fractions of hypomethylated repeat occurrences vary considerably among different classes of repetitive elements, from ~1% for L1 and Alu to ~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. S5).

We then examined the features for characterizing the differences between hypermethylated and hypomethylated repetitive elements. First, CpG density was significantly higher in the hypomethylated 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 hypomethylation because hypermethylation leads to depletion of CpG sites through deamination [56]. 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, hypomethylated occurrences were significantly more divergent than were hypermethylated 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. S4).

Next, we examined whether the hypomethylated 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 hypomethylated copies to all repetitive elements in individual non-overlapping bins 5 Mb in size (Fig. 2c-e). The non-random distribution patterns were more evident for LINE and LTR than for Alu. For example, we found hypomethylated LINEs to be enriched in the p-arm of chromosome 1 and in chromosomes 17 and 19. There were

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

hypomethylation 'hot spots' of LTR elements, *e.g.*, in chromosomes 6 and 9 (Supplemental Fig. S3). It is intriguing that some of these hypomethylation hot spots, such as those in the p-arms of chromosomes 6 and Y, seem to be shared among different classes of repetitive elements.

## 2.5 Analysis of the *Tol2* transposable element

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 [57]. 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 [58]. 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 [59].

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 hypermethylated 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 hypomethylated regions (number 14). From the bisulfite data, it appeared that the body of *Tol2*, from which data were missing, was hypomethylated. Nevertheless, our prediction estimated this region to be hypermethylated. These results demonstrate the ability of our method 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

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 hypomethylated, 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.

Although we presented an extension of our method to accommodate intermediate methylation states, the discrete methylation level, as defined in this article, is not identical to the actual proportion of methylated cytosines in the sample. That is, our prediction is inherently a qualitative classification. A method for truly quantitative observation of the methylation state using SMRT reads remains to be developed.

We explored the epigenetic landscape of repetitive elements within the human genome. An apparent limitation of our analysis is that we used the hg19 reference genome. By evaluating personal genomes instead of the reference genome, new insertions of these repetitive elements are often found, and such active occurrences should be of interest. 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 may 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.

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. 1f,g).

## 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 $\geq 1$ reads that mapped to C of either strand were considered covered. CpG sites that have a $\geq 1$ position within a 21 bp window with $\geq 3$ 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

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 [30, 34, 31, 32]. 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 [9]. 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 $\gamma$ using bisulfite sequencing data as the training data set to improve the prediction. 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 [30, 32, 35], 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 hypomethylated 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

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

single-base resolution. 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 and we 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 [42] 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 hypermethylation (hypomethylation, 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 hyper-/hypomethylated CpG blocks

Now, we must consider how to decompose $n$ CpG sites in the whole genome into hypermethylated regions $\{M_{\lambda \in \Lambda}\}$ and hypomethylated 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 [43, 44, 45, 46, 28] . 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:

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

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. We set $b$ to 50 in our experiments, but one can change

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 satisfy the constraint on the minimum number of CpG sites, we used the idea proposed by Csűrös [60].

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 [9]. In human cells, however, many partially methylated cytosines have been reported [19]. 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, $\boldsymbol{\beta}$, $\gamma$, and $b$, depend on the class to which the $i$-th CpG site belongs, we can redefine the new recurrences for three classes:

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

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

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

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.

For calculating more quantitative methylation level called discrete methylation levels, we performed prediction using the set of 10 perturbed gamma values (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

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

defined the methylation score at $i$ as $x/y$. We then defined the methylation status as ‘hypomethylated’ if the score was less than 0.5; otherwise, it was defined as ‘hypermethylated’.

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. To increase the number of CpG sites associated with methylation scores, we defined the methylation score when the CpG site was covered by one or more reads (*i.e.*, $y \geq 1$).

## 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 a hypo- or hypermethylated 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 actual status for each CpG site. CpG sites for which no bisulfite-treated reads were available were ignored. We treat a hypomethylated status as positive and a hypermethylated status as negative, because we are more interested in identifying rare hypomethylated regions accounting for $\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. 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, and CpG sites with a DML<0.4 were considered as hypomethylated. To further reduce the degree of hypomethylation 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 hypomethylated if more than half of the CpG sites were hypomethylated; otherwise, they were considered as hypermethylated.

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-e, we counted the numbers of hypomethylated and hypermethylated repeats in every 5 Mb bin and then used the Ideographica web server [61] 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.

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

### 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 [62] for analysis and visualization of the methylation patterns (Supplemental Fig. S5).

### 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 Figure 3 were produced using the UCSC Genome Browser (http://genome.ucsc.edu/) [63]. We used human bisulfite sequencing data and hypomethylated regions available in the GEO database [64].

## 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).

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## 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.

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# 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 hypo- and hypermethylated 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. 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, below the RefSeq gene track, black bars indicate hypomethylated 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 hypomethylated regions predicted from the bisulfite data. Green bars below indicate the alignability of short (100-bp) reads. The bottom row shows repeat masker tracks. **c.** Analysis of the sensitivity and precision (proportion of true-positives among the predicted positives) of our integration method. To estimate the read coverage sufficient to guarantee accuracy, we examined five datasets of different coverages ranging from $\sim$2-fold to $\sim$ 22-fold. **d.** IPDR profiles of CpGs are represented as points in the feature space. Predictions are made using a decision hyperplane (determined by gamma), and CpGs are classified as methylated (blue) or unmethylated (red). **e.** Multiple predictions using a set of different parameters define the discrete methylation level (DML) on each CpG site. **f.** DMLs (x-axis) correlated with the beta values of BeadChip (y-axis) for the CpG sites in our human sample. The beta values are color coded from 0 (red, hypomethylation) to 1 (blue, hypermethylation). The width is scaled to the relative number of CpG sites predicted as having that DML. The majority of CpG sites are hypomethylated, because most CpG sites on the BeadChip are designed on CpG islands. **g.** DMLs (x-axis) and methylation level monitored by bisulfite sequencing (y-axis) in our medaka sample according to the color coding and scaling shown in Fig. 1f. Most of the CpG sites were hypermethylated because we observed CpG methylation genome-wide.

### 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 hypomethylated (pink) repeat occurrences. The asterisks indicate statistical significance ($p < 1\%$) determined by the U test. **c-e.** Genome-wide distribution of hypomethylated repetitive elements. The ratio of hypomethylated repeat occurrences to all occurrences in each 5-Mb bin is indicated

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 hypomethylation and filled/blue indicates hypermethylation. 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, hypermethylation of *Tol2* was observed consistently by SMRT sequencing, while virtually no information was available on the *Tol2* region from bisulfite sequencing.

# REFERENCE

# References

[1] Weaver, I. C. *et al.* Epigenetic programming by maternal behavior. *Nat Neurosci* **7**, 847–54 (2004).

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

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

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

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

[6] Schmitz, R. J. *et al.* Transgenerational epigenetic instability is a source of novel methylation variants. *Science (New York, N.Y.)* **334**, 369–73 (2011).

[7] Molaro, A. *et al.* Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* **146**, 1029–41 (2011).

[8] Smith, Z. D. *et al.* A unique regulatory phase of dna methylation in the early mammalian embryo. *Nature* **484**, 339–344 (2012).

[9] Qu, W. *et al.* Genome-wide genetic variations are highly correlated with proximal dna methylation patterns. *Genome Res* **22**, 1419–25 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[10] Wilson, A. S., Power, B. E. & Molloy, P. L. Dna hypomethylation and human diseases. *Biochimica et biophysica acta* **1775**, 138–162 (2007).

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

[12] Beck, C. R. *et al.* LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170 (2010).

[13] Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012).

[14] Goodier, J. L. Retrotransposition in tumors and brains. *Mobile DNA* **5**, 11 (2014).

[15] Tubio, J. M. C. *et al.* Extensive transduction of nonrepetitive dna mediated by l1 retrotransposition in cancer genomes. *Science* **345** (2014).

[16] Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature* **452**, 215–219 (2008).

[17] Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell* **133**, 523–536 (2008).

[18] Meissner, A. *et al.* Genome-scale dna methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–70 (2008).

[19] Lister, R. *et al.* Human dna methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–22 (2009).

[20] Harris, R. A. *et al.* Comparison of sequencing-based methods to profile dna methylation and identification of monoallelic epigenetic modifications. (2010).

[21] Aird, D. *et al.* Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome biology* **12**, R18 (2011).

[22] Ross, M. *et al.* Characterizing and measuring bias in sequence data. *Genome Biology* **14**, R51 (2013).

[23] Bock, C. *et al.* Quantitative comparison of genome-wide dna methylation mapping technologies. *Nature biotechnology* **28**, 1106–1114 (2010).

[24] Gifford, C. A. *et al.* Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149 – 1163 (2013).

[25] Jiang, L. *et al.* Sperm, but not oocyte, {DNA} methylome is inherited by zebrafish early embryos. *Cell* **153**, 773 – 784 (2013).

[26] Korlach, J. *et al.* 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 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[27] Eid, J. *et al.* Real-time dna sequencing from single polymerase molecules. *Science (New York, N.Y.)* **323**, 133–8 (2009).

[28] Loomis, E. W. *et al.* Sequencing the unsequenceable: Expanded cgg-repeat alleles of the fragile x gene. *Genome Res* (2012).

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

[30] Flusberg, B. a. *et al.* Direct detection of dna methylation during single-molecule, real-time sequencing. *Nature methods* **7**, 461–5 (2010).

[31] Clark, T. a. *et al.* Characterization of dna methyltransferase specificities using single-molecule, real-time dna sequencing. *Nucleic acids research* **40**, e29 (2012).

[32] Fang, G. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic escherichia coli using single-molecule real-time sequencing. *Nature Biotechnology* (2012).

[33] Feng, Z. *et al.* Detecting dna modifications from smrt sequencing data by modeling sequence context dependence of polymerase kinetic. *PLoS Comput Biol* **9**, e1002935 (2013).

[34] Clark, T. a., Spittle, K. E., Turner, S. W. & Korlach, J. Direct detection and sequencing of damaged dna bases. *Genome integrity* **2**, 10 (2011).

[35] Schadt, E. E. *et al.* Modeling kinetic rate variation in third generation dna sequencing data to detect putative modifications to dna bases. *Genome Res* (2012).

[36] Clark, T. *et al.* Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via tet1 oxidation. *BMC Biology* **11**, 4 (2013).

[37] Eckhardt, F. *et al.* Dna methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* **38**, 1378–85 (2006).

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

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

[40] Nautiyal, S. *et al.* High-throughput method for analyzing methylation of cpgs in targeted genomic regions. *Proc Natl Acad Sci U S A* **107**, 12587–92 (2010). .

[41] Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–48 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[42] Su, J. *et al.* Cpg_mps: identification of cpg methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucleic Acids Res* (2012).

[43] Bashir, A. *et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology* **30**, 701–707 (2012).

[44] Zhang, X. *et al.* Improving genome assemblies by sequencing pcr products with pacbio. *BioTechniques* **53**, 61–2 (2012).

[45] English, A. C. *et al.* Mind the gap: Upgrading genomes with pacific biosciences rs long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).

[46] Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology* **30**, 693–700 (2012).

[47] Au, K. F. *et al.* Characterization of the human ESC transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E4821–30 (2013).

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

[49] Yamada, Y. *et al.* A comprehensive analysis of allelic methylation status of cpg islands on human chromosome 21q. *Genome Res* **14**, 247–66 (2004).

[50] Zhang, Y. *et al.* Dna methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS Genet* **5**, e1000438 (2009).

[51] Kerkel, K. *et al.* Genomic surveys by methylation-sensitive snp analysis identify sequence-dependent allele-specific dna methylation. *Nat Genet* **40**, 904–8 (2008).

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

[53] Hellman, A. & Chess, A. Extensive sequence-influenced dna methylation polymorphism in the human genome. *Epigenetics Chromatin* **3**, 11 (2010).

[54] Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in dna methylation associated with nuclear reprogramming. *Nat Biotechnol* **27**, 353–60 (2009).

[55] Kasahara, M. *et al.* The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714–719 (2007).

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

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

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[58] Iida, A. *et al.* Targeted reduction of the DNA methylation level with 5-azacytidine promotes excision of the medaka fish Tol2 transposable element. *Genetical research* **87**, 187–93 (2006).

[59] Koga, A. *et al.* Evidence for recent invasion of the medaka fish genome by the tol2 transposable element. *Genetics* **155**, 273 (2000).

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

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

[62] Kumaki, Y., Oda, M. & Okano, M. Quma: quantification tool for methylation analysis. *Nucleic acids research* **36**, W170–W175 (2008).

[63] Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic acids research* **42**, D764–70 (2014).

[64] Song, Q. *et al.* A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PloS one* **8**, e81148 (2013).

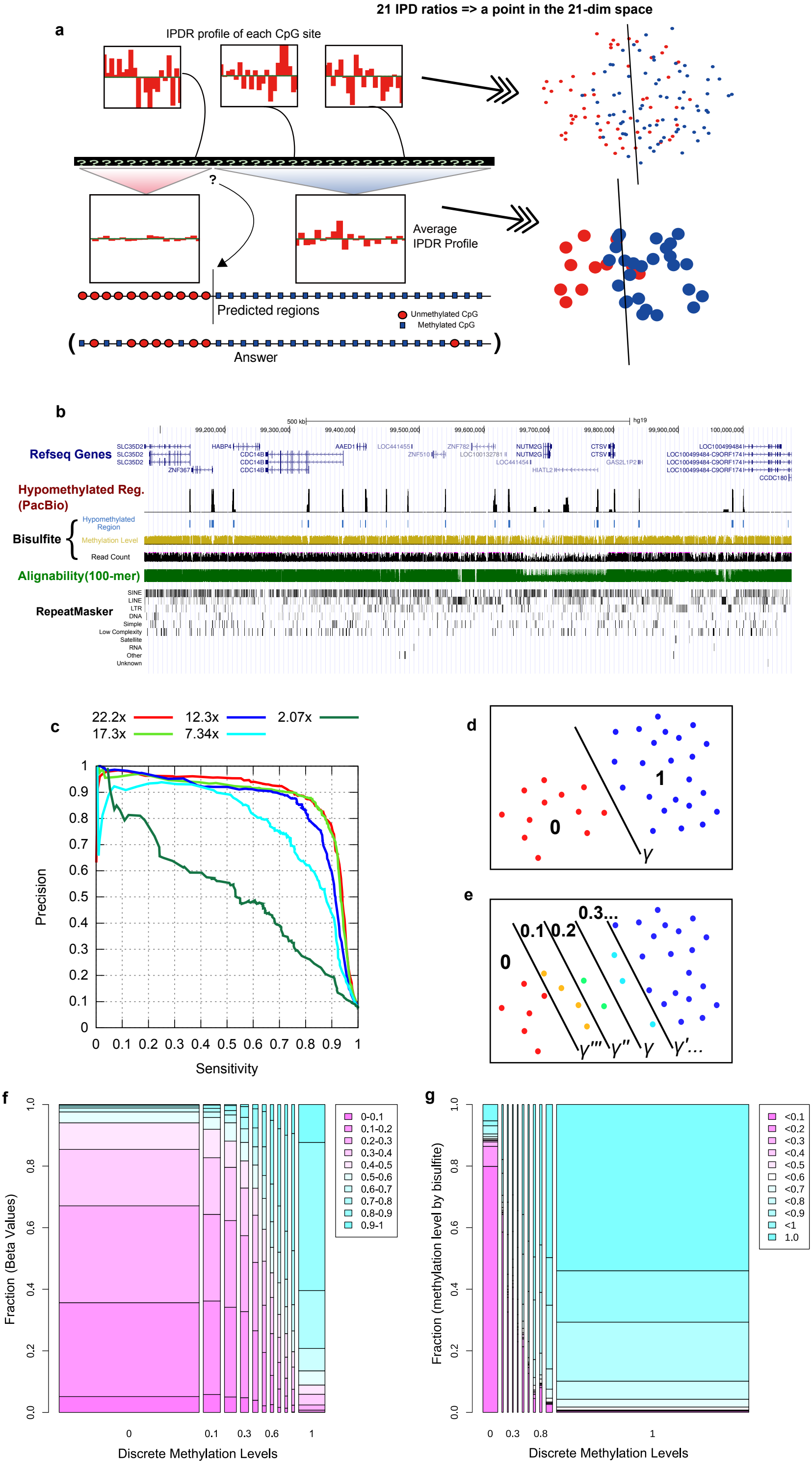
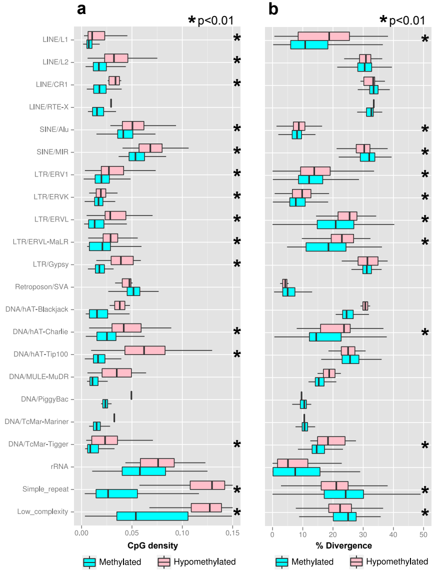
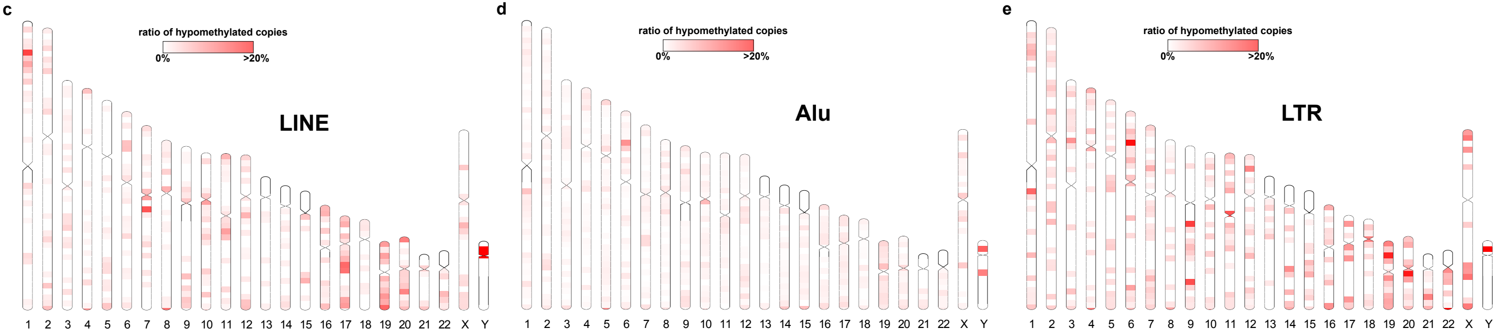
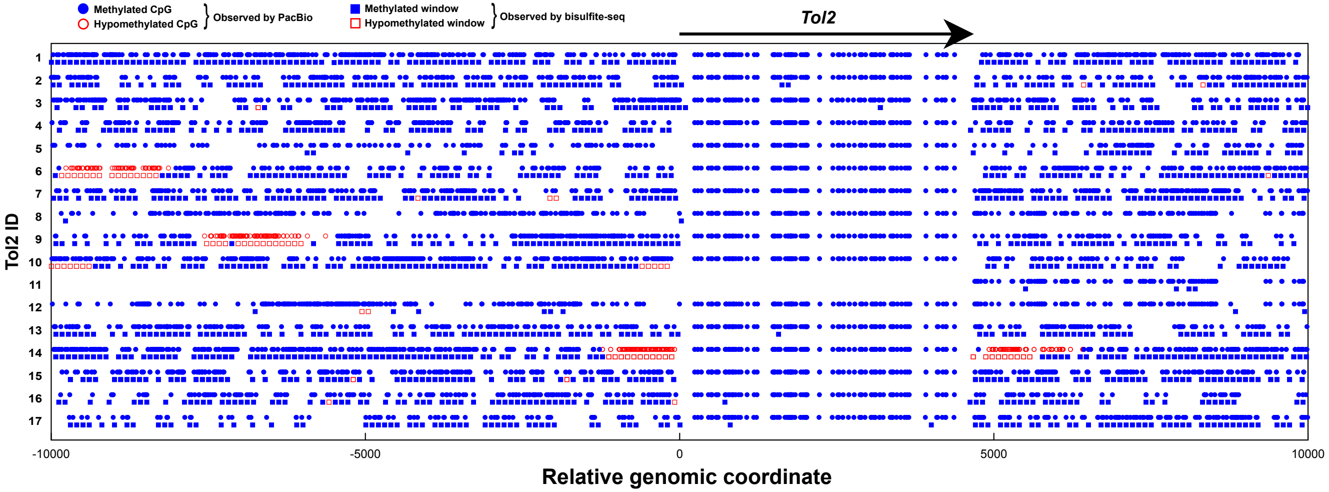

a
IPDR profile of each CpG site
21 IPD ratios => a point in the 21-dim space
?
Average IPDR Profile
Predicted regions
Unmethylated CpG
Methylated CpG
Answer
b
500 kb
hg19
99,200,000
99,300,000
99,400,000
99,500,000
99,600,000
99,700,000
99,800,000
99,900,000
100,000,000
Refseq Genes
SLC35D2
HABP4
AAED1
LOC441455
ZNF782
NUTM2G
CTSV
LOC100499484
ZNF367
CDC14B
ZNF510
LOC100132781
LOC441454
HIATL2
GAS2L1P2
LOC100499484-C9ORF174
CCDC180
Hypomethylated Reg. (PacBio)
Bisulfite
Hypomethylated Region
Methylation Level
Read Count
Alignability(100-mer)
RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
c
22.2x
17.3x
12.3x
7.34x
2.07x
Precision
Sensitivity
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
Fraction (Beta Values)
Discrete Methylation Levels
0.0
0.2
0.4
0.6
0.8
1.0
0
0.1
0.3
0.6
1
0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1
g
Fraction (methylation level by bisulfite)
Discrete Methylation Levels
0
0.3
0.8
1
<0.1
<0.2
<0.3
<0.4
<0.5
<0.6
<0.7
<0.8
<0.9
<1
1.0

**Table 1. Summary of methylation status on repetitive elements**

| Class | With >9 CpGs (A) | Covered (B) | B/A | Covered with >5x (C) | C/A | Hypomethylated (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
○ Hypomethylated CpG
Observed by PacBio
■ Methylated window
□ Hypomethylated 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

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# SUPPLEMENTAL FIGURE LEGENDS

### Supplemental Table S1. Statistics of SMRT sequencing data production

Summary statistics of SMRT sequencing data collected in this study.

### Supplemental Figure S1. The normal vector used for prediction

**a.** The normal vector $\beta$ used for prediction. We calculated $\beta$ as follows. Firstly, we classified the CpGs on the scaffold 1 in the medaka Hd-rR genome (version 1) into methylated CpGs and unmethylated CpGs according to bisulfite sequencing data. Next, for each CpG site, we calculate the IPD ratio profiles as the 21-dimensional vectors based on SMRT sequencing kinetics data. Then, using LDA (Linear Discriminant Analysis), we tried to find the best hyperplane that could separate these IPD ratio profiles into each class, namely, methylated or unmethylated. The normal vector of this hyperplane is denoted by $\beta$. **b.** The average IPDR profiles around unmethylated and methylated CpG sites. The x-axis shows the positions within 10 bp of the focal CpG site at the position represented by 0. The y-axis indicates IPDR values. The red- and blue-colored box plots at each position show the distributions of IPDR values around unmethylated and methylated CpG sites, respectively. The bottom, middle and top of each box plot indicate the first, second, and third quartiles, respectively, of the distribution.

### Supplemental Figure S2. Accuracy measures on the scaffold 1 and scaffold 2 of the medaka Hd-rR genome (version 1)

**a-c.** Matthew's correlation coefficient **(a)**, sensitivity **(b)**, and precision **(c)** as a function of the parameter $\gamma$, on the scaffold 1. **d-f.** Matthew's correlation coefficient **(d)**, sensitivity **(e)**, and precision **(f)** on the scaffold 2. The differently colored curves correspond to the different amount of sequencing data used for the prediction. Comparison

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of these statistics between scaffold 1 and scaffold 2 shows that $\gamma$ can be robustly optimized through maximization of MCC.

**Supplemental Figure S3. Examples of hypomethylated repeat occurrences in a hypomethylation 'hot spot'**

Three adjacent LTR1 elements were hypomethylated in this region (**a**), and a LTR12E element was located at a hypomethylated 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 hypomethylated 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 hypomethylated 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 S4. 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 hypomethylation and blue hypermethylation. The arrows show the hypomethylated occurrences that are clustered in terms of the sequence features.

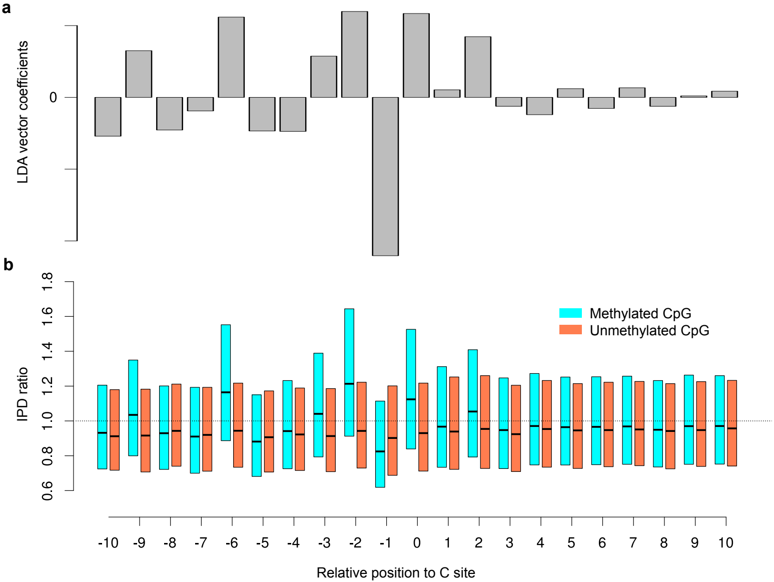
bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

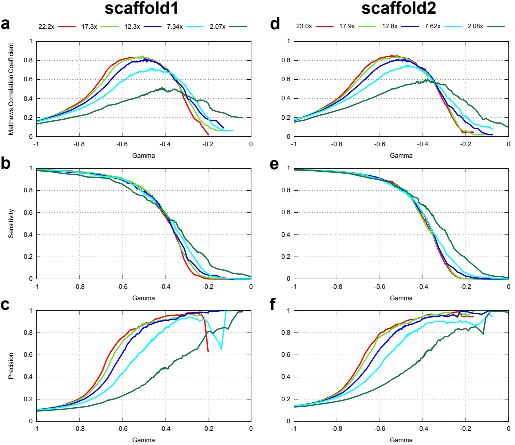
**Supplemental Table S2. The primers for nested PCR of the bisulfite treated blood DNA**

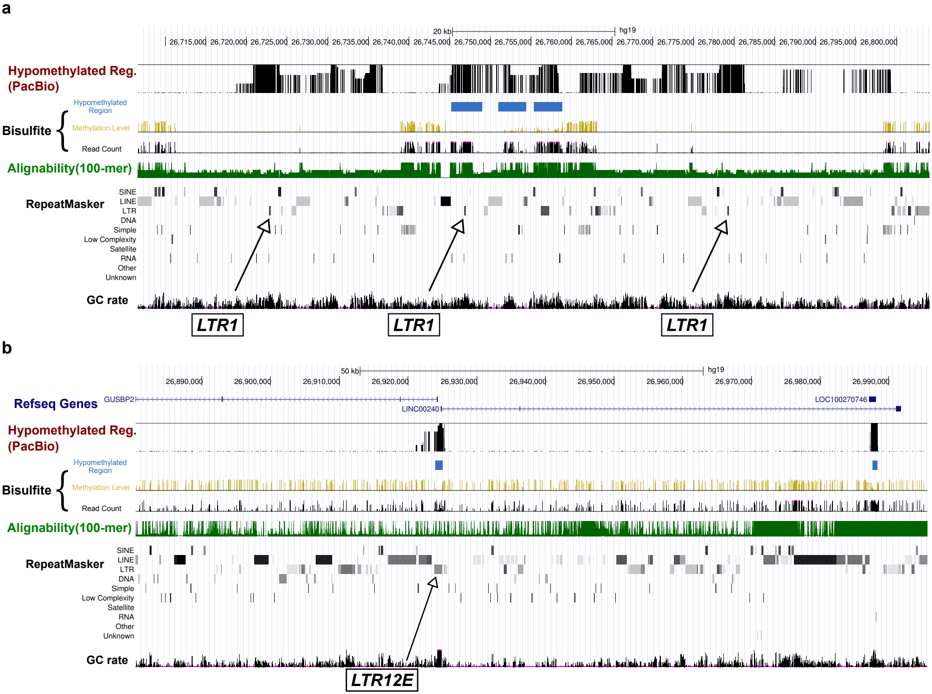
The primers for nested PCR are shown alongside the sequence IDs that correspond to those in Supplemental Figure S4, the sequence names, and the target genomic regions. For each entry, the forward primers appear in the top row, and the reverse primers appear in the second row.

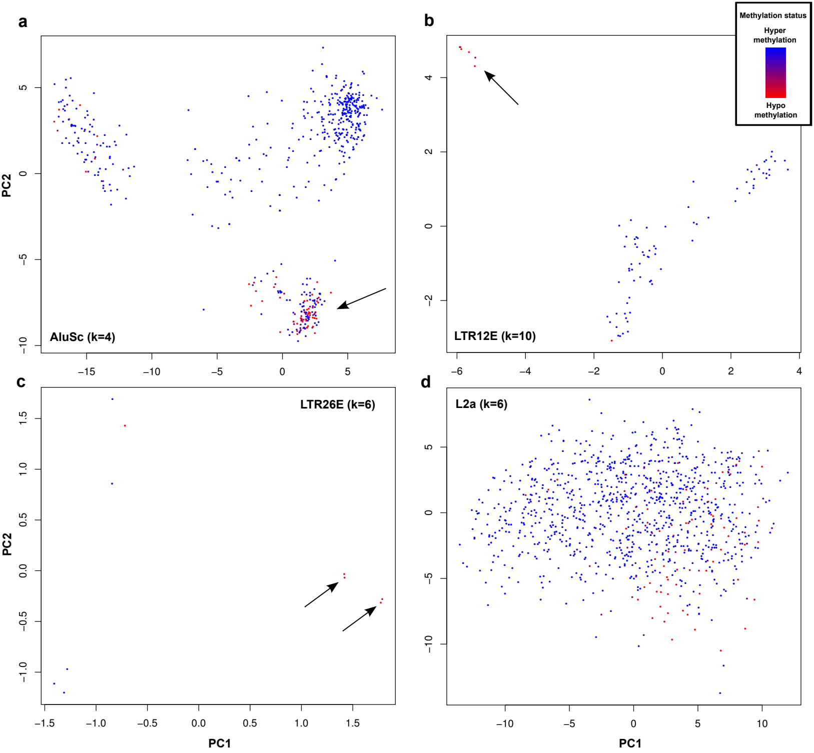
**Supplemental Figure S5. Methylation analysis of selected regions for validation of our prediction**

Of the 21 regions selected for validation of our method, 6 were amplified, and their Sanger sequencing reads were aligned to the target regions. In the alignments, the methylated (unconverted) CpGs are represented by the pink asterisks (*), and the unmethylated (converted) CpGs by the blue number sign (#). We can assess the efficiency of bisulfite conversion and the quality of the alignment by looking at non-CpG C sites (CpHs) because Cs in CpHs are usually unmethylated and should always be converted to Ts (represented by the colons (:)). Thus unconverted CpHs, which are highlighted by the brown exclamation marks (!), indicate low quality regions. The solid lines represent the other types of matches.


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


scaffold1
scaffold2
a
b
c
d
e
f
22.2x
17.3x
12.3x
7.34x
2.07x
23.0x
17.9x
12.8x
7.62x
2.08x
Matthews Correlation Coefficient
Sensitivity
Precision
Gamma
1
0.8
0.6
0.4
0.2
0
-1
-0.8
-0.6
-0.4
-0.2
0


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
Hypomethylated Reg. (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
Hypomethylated Reg. (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
b
c
d
Methylation status
Hyper methylation
Hypo methylation
AluSc (k=4)
LTR12E (k=10)
LTR26E (k=6)
L2a (k=6)
PC1
PC2
−15
−10
−5
0
5
−10
−5
0
5
−6
−4
−2
0
2
4
−2
0
2
4
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
−10
−5
0
5
10
−10
−5
0
5

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

C
* : Methylated CpG
C

C
# : Unmethylated CpG
T

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

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

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---
```

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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

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---------------------------------------------
```

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

| Statistics for medaka SMRT sequencing (P6) | |
|---|---|
| Number of SMRT cells | 38 |
| Number of mapped subreads | 2,596,378 |
| Mean mapped subread length (b) | 7,972 |
| Total bases of mapped subreads (b) | 20,698,432,471 |
| Coverage (medaka genome size = 800 Mb) | 25.87 |

| 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 |

bioRxiv preprint doi: https://doi.org/10.1101/018531; this version posted April 24, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

| 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 |  |