博士論文

論文題目

Observing comprehensive DNA methylomes via single molecule real-time sequencing: application to diploid and centromeric methylation

（一分子 DNA シーケンサによる DNA メチル化情報の網羅的観測手法 ―
二倍体ゲノムとセントロメア領域への応用）

氏名　　鈴木　裕太

Observing comprehensive DNA methylomes via single molecule real-time sequencing: application to diploid and centromeric methylation

（一分子 DNA シーケンサによる DNA メチル化情報の網羅的観測手法 ― 二倍体ゲノムとセントロメア領域への応用）

By

SUZUKI, Yuta

鈴木　裕太

A Dissertation

博士論文

Submitted to
The Graduate School of Frontier Sciences
The University of Tokyo
On December 13, 2017

In Partial Fulfillment of the Requirements
For the Degree of Doctor of Science
Supervisor: Prof. Shin-ichi Morishita 森下 真一

Summary of the contents

Observing comprehensive DNA methylomes via single molecule real-time sequencing: application to diploid and centromeric methylation

(一分子 DNA シーケンサによる DNA メチル化情報の網羅的観測手法 ― 二倍体ゲノムとセントロメア領域への応用)

1. CpG methylation detection from kinetics information of SMRT sequencing data (Chapter 3)

SMRT (single molecule real-time) sequencing, or PacBio sequencing, has been adopted in hundreds of sequencing studies despite its relatively high cost and raw read error rate, because it can produce longer reads than conventional NGS (next generation sequencers) and its random error profile eventually enabled extremely accurate genome assembly. It is also useful in epigenetics studies as it can produce kinetics information that reflects the methylation status of DNA sample. However, methylation analysis using SMRT sequencing was not applied to vertebrate genomes including human genome, as no method had enough power to detect cytosine methylation accurately. I developed an algorithmic strategy, *AgIn*, to extract methylation information from SMRT reads of practical sequencing depth, and I reported the method can achieve good detection accuracy (~93% sensitivity and precision for detecting unmethylated CpGs) with reads of depth ~30x (Figure 1). The method was successfully applied to catalog methylation statuses of repetitive elements in human genome and extremely homologous (>99% similarity over 4.6kbpp) *Tol2* transposon in medaka genome [1]. The method was continuingly adapted to newer version of SMRT sequencing protocol and it currently works well with the latest P6-C4 chemistry for the PacBio RSII instrument.

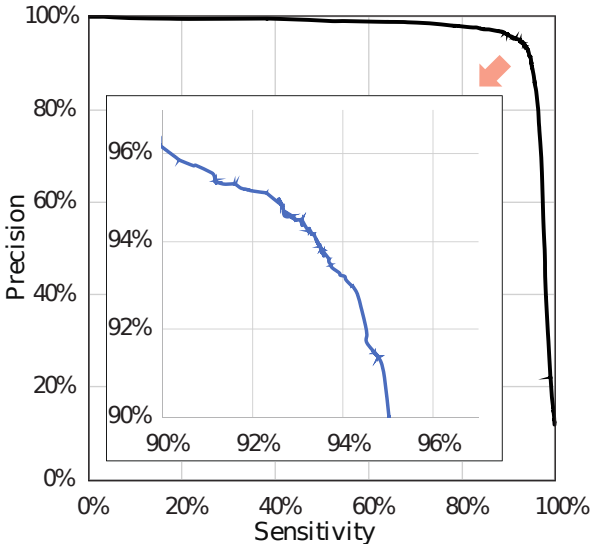


Figure 1. Prediction performance of AgIn method

2. Allele-specific methylation analysis using SMRT sequencing (Chapter 5)

This study is essentially an extension of AgIn to resolve another difficulty in the epigenetics studies. In diploid genomes, the methylation status of CpG sites in the same region can be different for two homologous chromosomes, and such situation is known as ***allele-specific methylation*** (ASM) events. ASM can regulate gene expression; for example, ***genomic imprinting***, in which the imprinted genes are expressed only from paternal or maternal chromosomes, are largely explained by the existence of ASM at the imprinting center. Also, disruption of ASM status is known to cause diseases. Several methods were developed to detect ASM events genome-widely, from use of methylation-sensitive restriction enzymes to model-

based estimation using short read bisulfite sequencing data, but these methods were unable to observe methylation status of the genome comprehensively for various reason. To overcome the situation, I developed a new strategy to observe directly genome-wide ASM events using SMRT sequencing reads, noting that the primary difference between two homologous chromosomes was nothing but heterozygous SNVs (single nucleotide variants), thus any method to detect ASM should relate heterozygous SNVs and methylation status around them. The proposed method assumes the availability of heterozygous SNV sites and, especially, their phasing information. This may sound demanding at first, but recent advent of linked-read technology (from 10x Genomics) lowered this hurdle. After aligning SMRT reads onto reference genome, the reads were separated according to alleles of heterozygous SNVs they contained, giving two sets of reads each represents single allele. Then, for each set, AgIn was applied to call CpG methylation status of the regions. By applying this strategy to two samples, AK1 (Asian Korean) and HG002 (Ashkenazim), I successfully identified thousands of CpG islands (CGIs) which shown ASM. The CGIs with strongest ASM signal were often located around the promoter regions of imprinted genes such as *TP73*, *ZNF597*, *ZNF331*, *HYMAI*, *MEST*, *PEG3*, *PEG13*. As a result, the list of CGIs with strong ASM signal was significantly (p=0.007, U test) populated with those were associated with imprinted genes. I also found that these ASM CGIs had unique distribution within genome in terms of chromatin state defined in ENCODE project; while most general CGIs were in TSS-like regions, ASM CGIs were rather found in actively transcribed regions or repressed regions. As an individual example, Figure 2 depicts the observed methylation statuses over the *GNAS* complex locus in AK1 genome. The locus is known to show a complex expression patterns regulated by ASM. I confirmed the ASM pattern of the locus was consistent with the known ASE (allele-specific expression) pattern. In collaboration with University of Iowa, I compared ASM calls with ASE data generated from long reads and short reads RNA-seq. As expected, I found that the expressing allele and the unmethylated allele were coincide for the ASM/ASE regions. Based on this observation, since

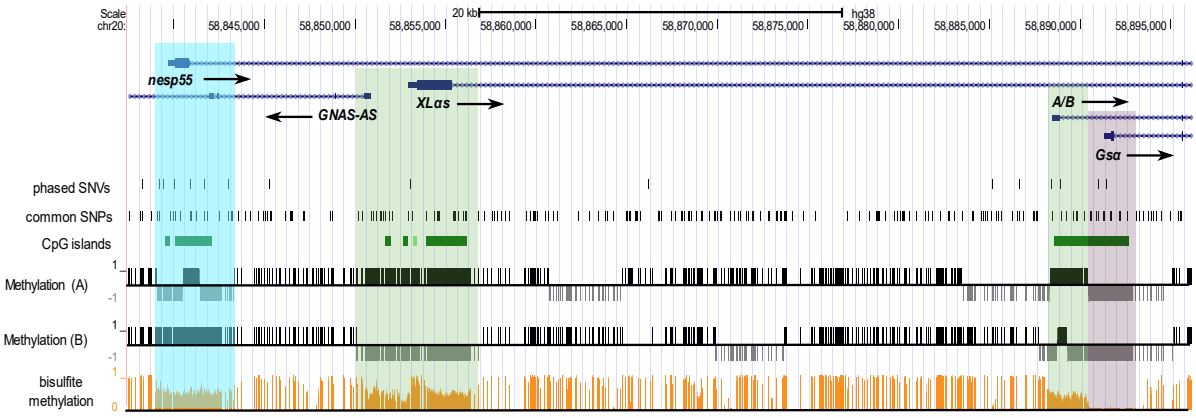


Figure 2. Allele-specific methylation over the GNAS locus. Genes in each of four shaded regions are known to be expressed (from the left) maternally, paternally, paternally, and biallelically.

ASE was much difficult to detect comprehensively due to scarcity of SNVs within exons, I claimed ASM can be a surrogate for ASE status of the gene. I also claimed these findings were possible only by using long reads because the majority of CpG sites in personal human genomes were located distant from any heterozygous SNV, which were sparsely distributed. This work is in preparation for publication [2].

3. CpG methylation analysis of centromeric repeat reagions in medaka genome (Chapter 4)

Centromeres were possibly the most difficult regions in any genome sequencing study, and epigenetic characterization of centromeres was largely indirect and descriptive as conventional method such as bisulfite sequencing could not observe methylation over its highly repetitive (Mbp-scale arrays of alpha-satellite) structure. I applied AgIn algorithm to medaka centromeres, which were assembled in contigs in the latest version of medaka genome [3], and identified the regions with unmethylated CpGs from a total of 11 chromosomes of two medaka inbred strains, Hd-rR and HSOK, which diverged ~2.5Mya (SNP rate ~ 2.5%) (Figure 3). By analyzing the sequence composition (k-mer distribution) of unmethylated and methylated repeats, I claimed these variations in methylation occurred recently, at least after the divergence of two strains (Figure 4). Therefore, it implied that the change in methylation status in each chromosome or strain could be independent, and I hypothesized that it may precede ultimate alteration in functionality of centromeres. I validated my methylation calls in centromeric regions by comparing them with calls from bisulfite sequencing data wherever they are available, although I found bisulfite sequencing was not sufficient to observe the overall picture of centromeric methylation patterns.

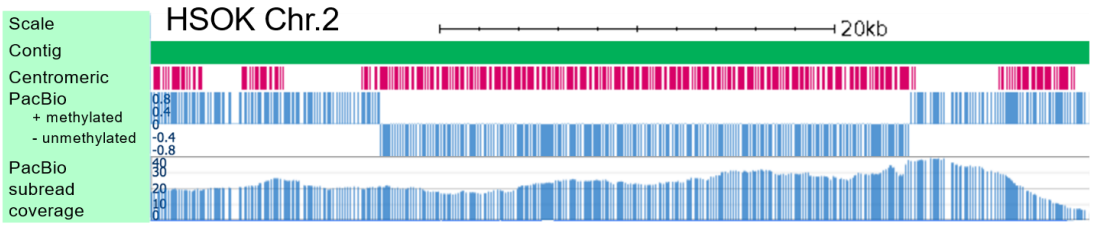


Figure 3. Methylation of centromeric repeats in medaka genome, HSOK chromosome 2.

Conclusion

In these works, I demonstrated the utility of SMRT sequencing for epigenetics study, especially for allele-specific methylation analysis and methylation analysis in complex region such as centromeres. The analyses of ASM in human genomes could retrieve imprinted genes and was consistent with the expression pattern of the ASE genes, which validated the accuracy of the method. By applying the method to centromeric repeat regions, I uncovered the cryptic methylation patterns of the regions, arriving at the hypothesis on an evolutionary drive of centromere sequences. This dissertation also contains two introductory chapters; Chapter 1 for general SMRT informatics and Chapter 2 for basics of kinetics information handling.

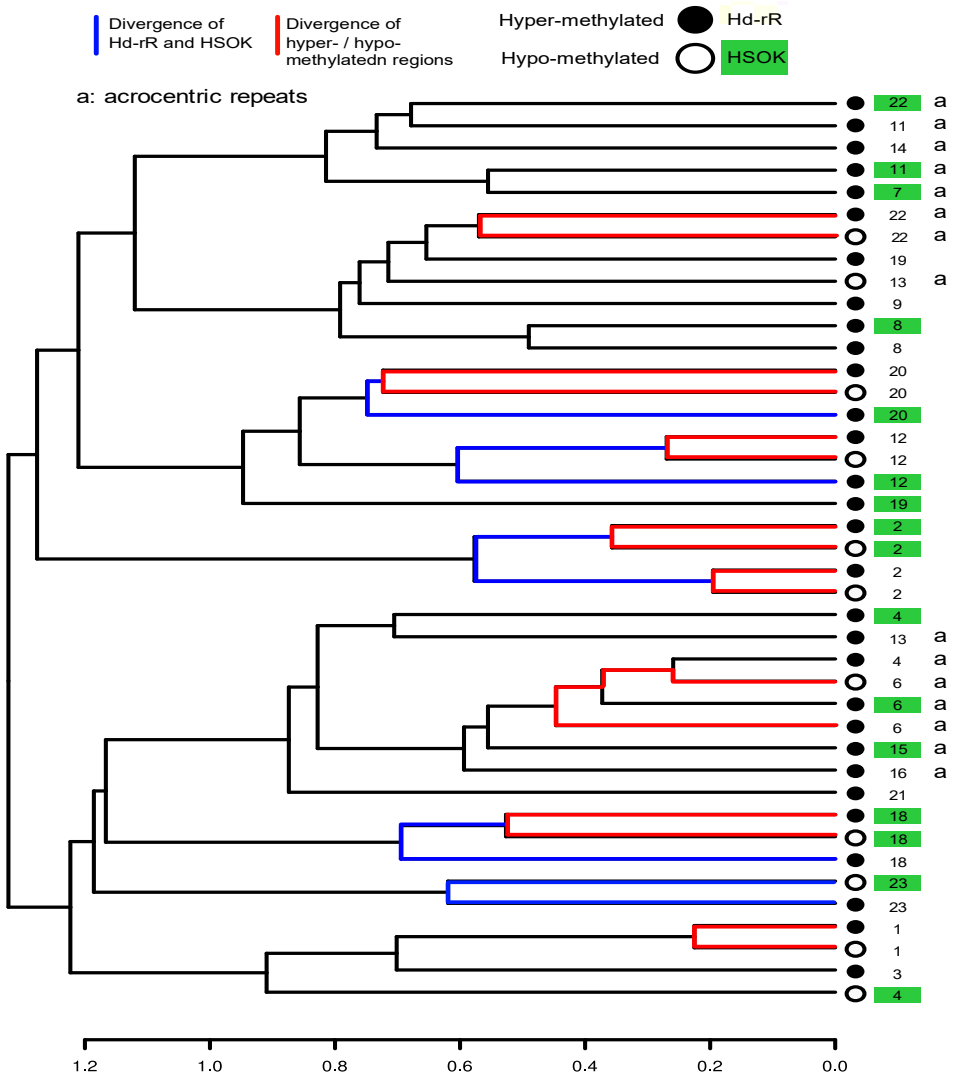


Figure 4. Clustering based on sequence similarity of centromeric repeats with unmethylated or methylated CpGs. Assuming the sequence similarity reflects the evolutionary distances, divergence of two strains precedes diversification of methylation pattern.

References

[1] Suzuki et al. "AgIn: measuring the landscape of CpG methylation of individual repetitive elements." *Bioinformatics*, 32, 19 **(2016)** 2911-2919.

[2] Suzuki et al. "Personal diploid methylomes and transcriptomes via phased heterozygous variants and single-molecule real-time sequencing." *in preparation*.

[3] Ichikawa, Tomioka, Suzuki, Nakamura, Doi et al. "Centromere evolution and CpG methylation during vertebrate speciation." *Nature Communications.* ", 8 **(2017)**: 1833.

Chapter 1

Informatics for PacBio long reads

Advances in SMRT biology and challenges in long read informatics

In 2011, advent of the PacBio RS sequencer and its SMRT (single molecule real-time) sequencing technology revolutionized the concept of DNA sequencing. Longer reads are promised to generate de novo assembly of much higher contiguity, and the claim was proved by several assembly projects[111, 89, 103]. The lack of sequencing bias was proved to be able to read regions which are extremely difficult for NGS (Next Generation Sequencers)[72].

None of these achievement, however, was just straightforward application of conventional informatics strategy developed for short read sequencers; the virtue of the long reads was not free at all. As many careful skeptics claimed in the early history of PacBio sequencing, the long reads seemed too noisy. Base accuracy was around $\sim$85% for single raw read, that is, $\sim$15% of bases were wrong calls, and indels consisted most of the errors. The higher error rate made it inappropriate to apply informatics tools designed for much accurate short read technologies.

Even the higher error rate is properly handled by sophisticated algorithms, the length of the reads itself can pose another problem. Computational burden of many algorithms

depends on the read length L. When only the short reads are assumed, it may be considered as constant, e.g., $L = 76, 150$, etc. The emergence of long read sequencer changed the situation drastically by improving the read length by orders of magnitudes, to thousands of bases, and to tens of thousands of bases by now. Besides the ongoing innovations for longer reads, there is a large variation in length of sequencing reads even in the same sequencing run. Therefore, the assumption that the read length is constant is not valid anymore, and one must have a strategy to handle (variably) long reads in reduced time (CPU hours) and space (memory footprint) requirement.

Availability of long read opened a door to the set of problems which were biologically existing in real but implicitly ignored by studies using short read sequencing. For example, we had to realize that a non-negligible fraction of reads could cover SVs (structural variants), requiring a new robust mapping strategy other than simply masking the known repetitive regions.

Consequently, many sophisticated algorithms had to be developed to resolve these issues; how to mitigate higher error rate, and how it can be done efficiently for long reads. The rest of this article covers some important innovations achieved and ongoing efforts in informatics area to make the most of long reads data.

1.1 Aligning noisy long reads with reference genome

When one aligns long reads against reference sequence, one must be aware that the variations between reads and reference stems from two conceptually separate causes. On one hand, there are sequencing errors in its simple sense, which is discrepancy between a read observed and actual sequence being sequenced. On the other hand, we expect a sample sequenced would have slightly different sequence than a reference sequence (otherwise there is no point in doing sequencing), and those difference are usually called variants. Though sequencing errors and sequence variants are conceptually different, however, they both appears just as

errors to us unless they have some criteria to distinguish them. The next two examples are for understanding why the distinction between two classes of error is relevant here.

Lets consider we have some noisy reads. Clearly, we cannot call sequence variants specific to the sample unless the frequency of sequencing errors is controlled to be sufficiently low compared to the frequency of variants. This is the reason why it is difficult for noisy reads to detect small nucleotide variants such as point mutations and indels.

Next, assume we have long reads. Then, there are more chances that the reads span the large variations such as structural variations (SVs) between a reference genome and the sample sequenced. This situation is problematic for aligners who considered any possible variation between reads and reference to be sequencing errors, for such aligners would fail to detect correct alignment as they need to introduce too much errors for aligning these sequences. Some aligners try to combat the situation by employing techniques such as chaining and split alignment. Some aligners (NGMLR, Minimap2) explicitly introduce an SV-aware scoring scheme such as a two-parts concave gap penalty, which reflects the two classes of variations between read and reference.

Sequence alignment is so fundamental in sequence analysis that it finds its application everywhere. For example, mapping sequencing reads to reference genome is the very first step of resequencing studies. Accuracy of mapping can directly be translated into the overall reliability of results. Also, mapping is often one of the most computationally intensive steps. Therefore, accurate and faster mapping software would benefit the whole area of resequencing studies. In the context of de novo assembly pipeline, it is used for detecting overlap among long reads. Of noted, desired balance of sensitivity and specificity of overlap detection is controlled differently than in mapping to reference, and could often be very subtle.

Though it is more or less subjective to make distinction between standalone aligners and aligners designed as a module of assembly pipeline or SV detection pipeline, we decided to cover some aligners in other sections. MHAP will be introdued in relation with Canu in the section devoted to assembly tools. Similarly, NGMLR will be detailed together with Sniffle

in the section for SV detection.

1.1.1 BWA-SW and BWA-MEM

Adopting the seed-and-extend approach, BWA-SW[68] builds FM-indices for both query and reference sequence. Then, DP (dynamic programming) is applied to these FM-indices to find all local matches, i.e., seeds, allowing mismatches and gaps between query and reference. Detected seeds are extended by Smith-Waterman algorithm. Some heuristics are explicitly introduced to speed up alignment of large-scale sequencing data and to mitigate the effect of repetitive sequences. BWA-MEM[65] inherits similar features implemented in BWA-SW such as split alignment, but is found on a different seeding strategy using SMEM (supermaximal exact matches) and reseeding technique to reduce mismapping caused by missing seed hits.

1.1.2 BLASR

BLASR[12] (Basic Local Alignment with Successive Refinement) is also one of the earliest mapping tools specifically developed for SMRT reads. Like BWA-MEM, it is probably the most widely used one to date. Bundled with official SMRT Analysis, it has been the default choice for the mapping (overlapping) step in all protocols such as resequencing, de novo assembly, transcriptome analysis, and methylation analysis. In the BLASRs paper, the authors explicitly stated it was designed to combine algorithmic devices developed in two separate lines of studies, namely, a coarse alignment method for whole genome alignment and a sophisticated data structure for fast short read mapping. Proven to be effective for handling noisy long read, the approach of successive refinement, or seed-chain-align paradigm, has become a standard principle.

BLASR first finds short exact matches (anchors) using either suffix array or FM index[33]. Then, the regions with clustered anchors aligned colinearly are identified as candidate mapping locations, by global chaining algorithm[1]. The anchors are further chained by sparse dynamic programming (SDP) within each candidate region[30]. Finally, it gives detailed

alignment using banded DP (dynamic programming) guided by the result of SDP. BLASR achieved 10-fold faster mapping of reads to human genome than BWA-SW algorithm at comparable mapping accuracy and memory footprint.

1.1.3 DALIGNER

DALIGNER[85] is specifically designed for finding overlaps between noisy long reads, though its concept can also be adopted for a generic long read aligner, as implemented in DAMAPPER (https://github.com/thegenemyers/DAMAPPER). Like in BLASR, DALIGNER also performs filter based on short exact matches. Instead of using BWT (FM index), it explicitly processes k-mers within reads by thread-able and cache coherent implementation of radix sort. Detected k-mers are then compared via block-wise merge sort, which reduces memory footprint to a constant depending only on the block size. To generate local alignment, it applies $O(ND)$ diff algorithm between two candidate reads[83]. DALIGNER achieved $22 \sim 39$-fold speedup over BLASR at higher sensitivity in detecting correct overlaps[85]. DALIGNER is supposed to be a component for read overlap (with DAMASKER for repeat masking, DASCRUBBER for cleaning up low quality regions, and a core module for assembly) of DAZZLER de novo assembler for long noisy reads, which will be released in future.

1.1.4 Minimap2

Minimap2[67] is one of the latest and state-of-the-art alignment program. Minimap2 is general-purpose aligner in that it can align short reads, noisy long reads, and reads from transcripts (cDNA) back to a reference genome. Minimap2 combines several algorithmic ideas developed in the field, such as locality-sensitive hashing as in Minimap and MHAP. For accounting possible SVs between reads and genome, it employs concave gap cost as in NGMLR, and it is efficiently computed using formulation proposed by Suzuki, Kasahara[113]. In addition to these features, the authors further optimized the algorithm, by transforming

the DP matrix from row-column coordinate to diagonal-antidiagonal coordinate for better concurrency in modern processors. According to the author of Minimap2, it is supposed to replace BWA-MEM, which is in turn a widely used extension of BWA-SW.

1.2 De novo assembly

As Lander-Waterman theory[59] would assert, the longer input reads are quite essential in achieving a high-quality genome assembly for repetitive genomes. Therefore, developing a de novo assembler for long read is naturally the most active area in the field of long read informatics.

To our knowledge, almost all assemblers published for long read take an overlap-layout-consensus (OLC) approach, where the overall task of assembly can be divided into the three steps. (1. Overlap) The overlaps between reads are identified as candidate pairs representing the same genomic regions, and the overlap graph is constructed to express these relations. (2. Layout) The graph is transformed to generate linear contigs. The step often starts by constructing the string graph[84], a string-labeled graph which encodes all the information in reads observed, and eliminates edges containing redundant information. (3. Consensus) The final assembly is polished. To eliminate errors in contigs, consensus is taken among reads making up the contigs.

Though we do not cover tools for the consensus step here, there are many of them released to date including official Quiver and Arrow bundled in SMRT Analysis[1], another official tool pbdagcon (https://github.com/PacificBiosciences/pbdagcon), Racon[118], and MECAT[123]. Of note, quality of a polished assembly can be much better than a short-read-based assembly due to the randomness of sequencing errors in long reads [14, 85].

[1]https://github.com/PacificBiosciences/GenomicConsensus

1.2.1 FALCON

FALCON[15] is designed as a diploid-aware de novo assembler for long read. It starts by carefully taking consensus among the reads to eliminate sequencing errors while retaining heterozygous variants which can distinguish two homologous chromosomes (FALCON-sense). For constructing a string graph, FALCON runs DALIGNER. The resulted graph contains haplotype-fused contigs and bubbles reflecting variations between two homologous chromosomes. Finally, FALCON-unzip tries to resolve such regions by phasing the associated long reads and local re-assembly. The contigs obtained are called haplotigs, which are supposed to be faithful representation of individual alleles in the diploid genome.

1.2.2 Canu and MHAP

MHAP[10] (Min-Hash Alignment Process) utilized MinHash for efficient dimensionality reduction of the read space. In MinHash, H hash functions are randomly selected, each of them maps k-mer into an integer. For a given read of length L, only the minimum values over the read are recorded for each of H hash functions. The k-mers at which the minimum is attained are called *min-mers*, and resulted representation is called *a sketch*. The sketch serves as a locality sensitive hashing of each read, for the similar sequences are expected share similar sketches. Because the sketch retains the data only on H min-mers, its size is fixed to H, independent of read length L. Built on top of MHAP, Canu[53] extends best overlap graph (BOG) algorithm[77] for generating contigs. A new bogart algorithm estimates an optimal overlap error rate instead of using predetermined one as in original BOG algorithm. This requires multiple rounds of read and overlap error correction, but eventually enables to separate repeats diverged only by 3%. Though BOG algorithm is greedy, the effect is mitigated in Canu by inspecting non-best overlaps as well to avoid potential misassemblies.

1.2.3 HINGE

While there is no doubt that obtaining more contiguous (i.e., higher contig N50) assembly is a major goal in genome assembly, the quest just for longer N50 may cause misassemblies if the strategy gets too greedy. Being aware that danger, HINGE[46] aims to perform the optimal resolution of repeats in assembly, in the sense that the repeats should be resolved if and only if it is supported by long read data available. To implement such a strategy is rather straightforward for de Bruijn graphs.In de Bruijn graph, its k-mers representing nodes are connected by edges when they co-occur next to each other in reads. In ideal situation, the genome assembly is realized as an Eulerian path, i.e., trail which visits every edge exactly once, in the de Bruijn graph. However, de Bruijn graphs are not robust for noisy long read, so overlap graphs are usually preferred for long read. One of the key motivations of HINGE is to give such a desirable property of de Bruijn graphs, to overlap graphs which is more error-resilient. To do so, HINGE enriches string graph with additional information called hinges based on the result of the read overlap step. Then, assembly graph with optimal repeat resolution can be constructed via a hinge-aided greedy algorithm.

1.2.4 Miniasm and Minimap

Minimap[66] adopts a similar idea as MHAP, for it uses minimizers to represents the reads compactly. For example, Minimap uses a concept of (w, k)-minimizer, which is the smallest (in the hashed value) k-mer in w consecutive k-mers. To perform mapping, Minimap searches for colinear sets of minimizers shared between sequences. Miniasm[66], an associated assembly module, generates assembly graph without error-correction. It firstly filters low-quality reads (chimeric or with untrimmed adapters), constructs graph greedily, and then cleans up the graph by several heuristics, such as popping small bubbles and removing shorter overlaps.

1.3 Detection of structural variants (SVs)

Sequence variants are called structural when they are explained by the mechanisms involving double-strand breaks, and are often defined to be variants larger than certain size (*e.g.*, 50 bp) for the sake of convenience. They are categorized into several classes such as insertions/deletions (including presence/absence of transposons), inversion, (segmental) duplication, tandem repeat expansion/contraction, etc. While some classes of SVs are notoriously difficult to detect via short reads (especially long inversions and insertions), long reads have promise to detect more of them by capturing entire structural events within sequencing reads.

1.3.1 PBHoney

PBHoney[29] implements combination of two methods for detecting SVs via read alignment to reference sequence. Firstly, PBHoney exploits the fact that the alignment of reads by BLASR should be interrupted (giving soft-clipped tails) at the breakpoints of SV events. PBHoney detects such interrupted alignments (piece-alignments) and clusters them to identify individual SV events. Secondly, PBHoney locates SVs by examining the genomic regions with anomalously high error rate. Such a large discordance can signal the presence of SVs because sequencing errors within PacBio reads are supposed to distribute rather randomly.

1.3.2 Sniffles and NGMLR

NGMLR[102] is a long-read aligner designed for SV detection, which uses two distinct gap extension penalties for different size range of gaps (*i.e.*, concave gap penalty) to align entire reads over the regions with SVs. Intuitively, the concave gap penalty is designed so that it can allow longer gaps in alignment while shorter gaps are penalized just as sequencing errors. Adopting such a complicated scoring scheme makes the alignment process computationally intensive[78], but NGMLR introduces heuristics to perform faster alignment. Then, an associated tool to detect SVs, Sniffles scans the read alignment to report putative SVs

which are then clustered to identify individual events and evaluated by various criteria. Optionally, Sniffle can infer genotypes (homozygous or heterozygous) of detected variants, and can associate nested SVs which are supported by the same group of long reads.

1.3.3 SMRT-SV

SMRT-SV[42] is a SV detection tool based on local assembly. It firstly maps long reads to reference genome, against which SVs are called. Then it searches signatures of SVs within alignment results, and 60 kbp regions around the detected signatures are extracted. The regions are to be assembled locally from those reads using Canu, then SVs are called by examining the alignment between assembled contigs and reference. Local assembly is performed for other regions (without SV signatures) as well to detect smaller variants.

1.4 Beyond DNA - Transcriptome analysis and methylation analysis.

SMRT sequencing has been found its applications outside DNA analysis as well. When it is applied to cDNA sequencing, long read would be expected to capture the entire structures of transcripts to elucidate expressing isoforms comprehensively. IDP (Isoform Detection and Prediction) [4] and IDP-ASE[24] are tools dedicated to analyze long read transcriptome data. To detect expressing isoforms from long read transcriptome data, IDP formulates it in the framework of integer programming. To estimate allele-specific expression both in gene-level and isoform-level, IDP-ASE then solves probabilistic model of observing each allele in short read RNA-seq. Both IDP and IDP-ASE effectively combines long read data for detection of overall structure of transcripts, and short read data for accurate base-pair level information.

In methylation analysis, official kineticsTools in SMRT Analysis has been widely used to detect base modification sites and to estimate sequence motives for DNA modification (see [34] for the principle of detection). Detecting 5-methyl-cytosines (5mC), which is by far

the dominant type of DNA modification in plants and animals, is challenging due to their subtle signal. Designed for detecting 5mC modifications in large genomes within practical sequencing depth, AgIn[114] exploits the observation that CpG methylation events in vertebrate genomes are correlated over neighboring CpG sites, and tries to assign the binary methylation states to CpG sites based on the kinetic signals under the constraint that a certain number of neighboring CpG sites should be in the same state. Making the most of high mappability of long read, AgIn has been applied to observe diversified CpG methylation statuses of centromeric repeat regions in fish genome[43], and to observe allele-specific methylation events in human genomes.

Concluding remarks

We have briefly described some innovative ideas in bioinformatics for an effective use of long read data. As concluding remarks, let me mention a few prospects for the future development in the field. By now, it is evident the quest for complete genome assembly is almost done, but the remaining is the most difficult part such as extremely huge repeats, centromeres, telomeres. While many state-of-the-art assemblers take the presence of such difficult regions into account and can carefully generate high quality assembly for the rest of genomes, it is remained open how to tackle these difficult part of the genome, how to resolve its sequence, not escaping from them.

Base modification analysis using PacBio sequencers may also have huge potential to distinguish several types of base modifications and to detect them simultaneously in the same sample[18], but only the limited number of modification types (6mA, 4mC, and 5mC) are considered for now. This is mainly due to the technical challenge to alleviate noise in kinetics data to distinguish each type of modifications and unmodified bases from each other.

That said, it will be no doubt that the field would be more attractive than ever, as the use of long read sequencer becomes a daily routine in every area of biological research, or

maybe even in clinical practice.

Chapter 2

Development of the model: from kinetics to methylation

In this chapter, we will detail the origin and the nature of kinetics information observed by PacBio sequencers. We will demonstrate how they look like in real sequencing data of a human genome which has CpG methylation as its canonical base modification. Then, some basic concepts of the model to predict (CpG) methylation status will be discussed. This chapter is intended to supplement the next chapter, where the model will be fully described and applied to answer some biological questions. Readers who are interested in the applications may read the next chapter first, and then come back to this chapter.

2.1 Kinetics information obtained with PacBio sequencing

Detection of base modifications in sample DNA using PacBio sequencing relies on the fact that kinetics of DNA polymerases is affected by the presence of base modifications in template DNA molecules they work on[34]. Of note, PacBio sequencing, i.e., single molecule real-time sequencing would be an appropriate sequencing method to observe such kinetics information.

By contrast, most Next Generation Sequencers adopt an opposite strategy to sequence DNA; In NGS, what is measured is essentially the synchronized behavior of the amplified DNA molecules, and is never a reflection of native molecular dynamics.

In processing of PacBio data, the raw measurements of fluorescent intensity are called the traces (Figure 2.1a). Then, these traces are interpreted via probabilistic model and transformed into a summarized data called the pulses (Figure 2.1b), which is the direct predecessor of the "basecalls". While these processing, which is called primary analysis, are done by internal code implemented in PacBio RS (I, II, and Sequel) system and is not fully configurable, most kinetics information are retained in basecalls data and in alignment results as well. The pulses are available on configuration, and traces can be outputted only for debugging purpose.

To infer the base modification status of DNA, inter-pulse duration (IPD) and pulse width (PW) are important measurement (Figure 2.1b). As their names stand, IPD and PW are metrics calculated for each pulse and the metrics are attached to each base associated with the pulse. Flusberg et al. demonstrated these IPD and PW are sufficient information to distinguish base modification statuses[34]. They shown that each of C (cytosines), mC (5-methylated cytosines), and hmC (5-hydroxymethylated cytosines) have apparently distinct distribution when they are projected onto the principal components. While both IPD and PW had strongly contributed to top two principal components, they focused only on IPD later in the study, and IPD are now widely considered as critical information for detecting methylation. Throughout this study, we consider only IPD though we here admit importance of PW in modification analysis remains open for future studies.

2.2 Determinants of IPD and IPD ratio

IPD has a dimension of time, and is in scale of seconds. They fluctuate so much in real data, which makes it extremely difficult to infer the base modification status from a single

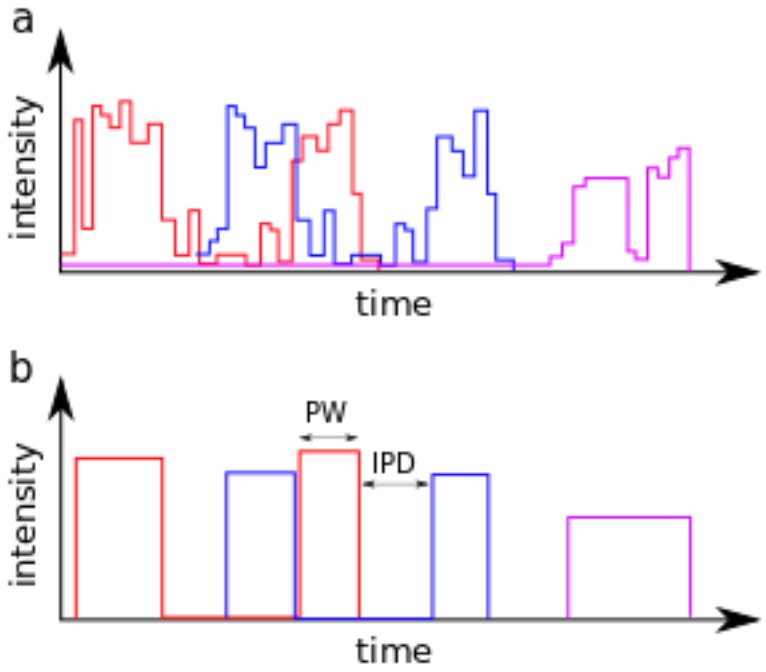


Figure 2.1: Kinetics information obtained from PacBio sequencing. **A.** Traces are raw measurements of fluorescent intensities. **B.** Traces are processed to give their compact representation, pulses. Both PW (pulse width) and IPD (inter-pulse duration) are both measured in the unit of time, and are supposed to contain information about the base modification in DNA template under replication.

measurement of IPD. Thus, a practical approach would be taking average over the IPD observations on the same genomic location to alleviate the effect of nonessential noise. To take an average over the different reads is essentially to observe the population of cells collectively for their modification status, thus the approach would be valid if we can assume the modification status is uniform within them in a sense.

Even after taking the average over the observations, IPD value for each genomic location would be expected to be greatly different each other. This is because IPDs essentially reflect the steric effect posed by a local configuration of bases, *i.e.*, a local sequence context and, of note, this is the very fact that enables the detection of base modification via examination of IPD, as we expect distinct IPDs are obtained for the context with base modifications and without them.

Based on this principle, several strategies would be possible to detect base modification from IPD data. One is to set up a complete model for the correspondence of each context (possibly with every type of modifications) and expected IPD values. With such a model, and with IPD measurements with enough precision, we can simply look up the model to infer the unknown sequence context. However, setting up such a model is difficult since it is unclear how to prepare the positive controls, that is, DNA molecules with modifications precisely at desired positions, with the only exception being the cases of bacterial genomes with known MTases. Therefore, another strategy, only viable option for non-bacterial study, is to prepare an unmodified DNA sample (a negative control) and compare IPD values between the control and the sample. Then, discrepancy between the control and the sample are assumed to be due to some base modifications in the sample. Thankfully, the negative controls need not to be prepared for every sample sequenced for modification analysis. Once enough amount of negative controls is available, we can set up the model that map sequence context to expected IPD value(*e.g., see [32]*). This model is known to be the in-silico IPD model and is now widely used.

By dividing the IPD value by the expected unmodified IPD (calculated via in-silico

model), we obtain IPD ratio. From the very definition, IPD ratio is expected to be 1.0 on bases without base modifications in their context, at least on average over many observations. On the other hand, large IPD ratio (> 2.0) often signals the presence of base modification(s) at, or around, the base associated to it[34, 31, 32]. In the next section, we will see how IPD ratio can reflect the base modification status of underlying sequence context, with the real example of CpG methylation in human genome.

2.3 Distinct IPD ratio profiles around methylated and unmethylated CpGs

Figure 2.2a shows the distribution of IPD ratio on each relative position around the CpG sites in a part of human genome. Only the focal CpG is fixed in this context; cytosine site is at position 0, and guanine site is at position 1. We call such a plot of IPD ratio distribution around some motifs as an IPD ratio profile. In the figure, we classified the CpG sites based on their methylation status inferred from bisulfite sequencing of the same DNA sample. While it was evident that methylated CpGs had distinct IPD ratio profile, the effect of the modification was not restricted to the modified cytosine base at position 0. For example, positions -1, -2, and -6 also exhibited a relatively large shift in IPD values. Interestingly, the position -1 shown a shift towards opposite direction than others; The incorporation of the corresponding bases tended to be faster when the context was modified.

Similarly, Figure 2.2b shows the IPD ratio profiles of CpG sites, but now we ignore the methylation status of the focal CpG sites, and they are classified according to the methylation status observed in the opposite strand. Strikingly, two profiles for methylated CpGs in Figure 2.2 are almost indistinguishable. This motivated us to use IPD ratio profiles for the opposite strand as well for predicting the CpG methylation status. In other words, we can effectively ignore the distinction between strands in predicting the CpG methylation status.

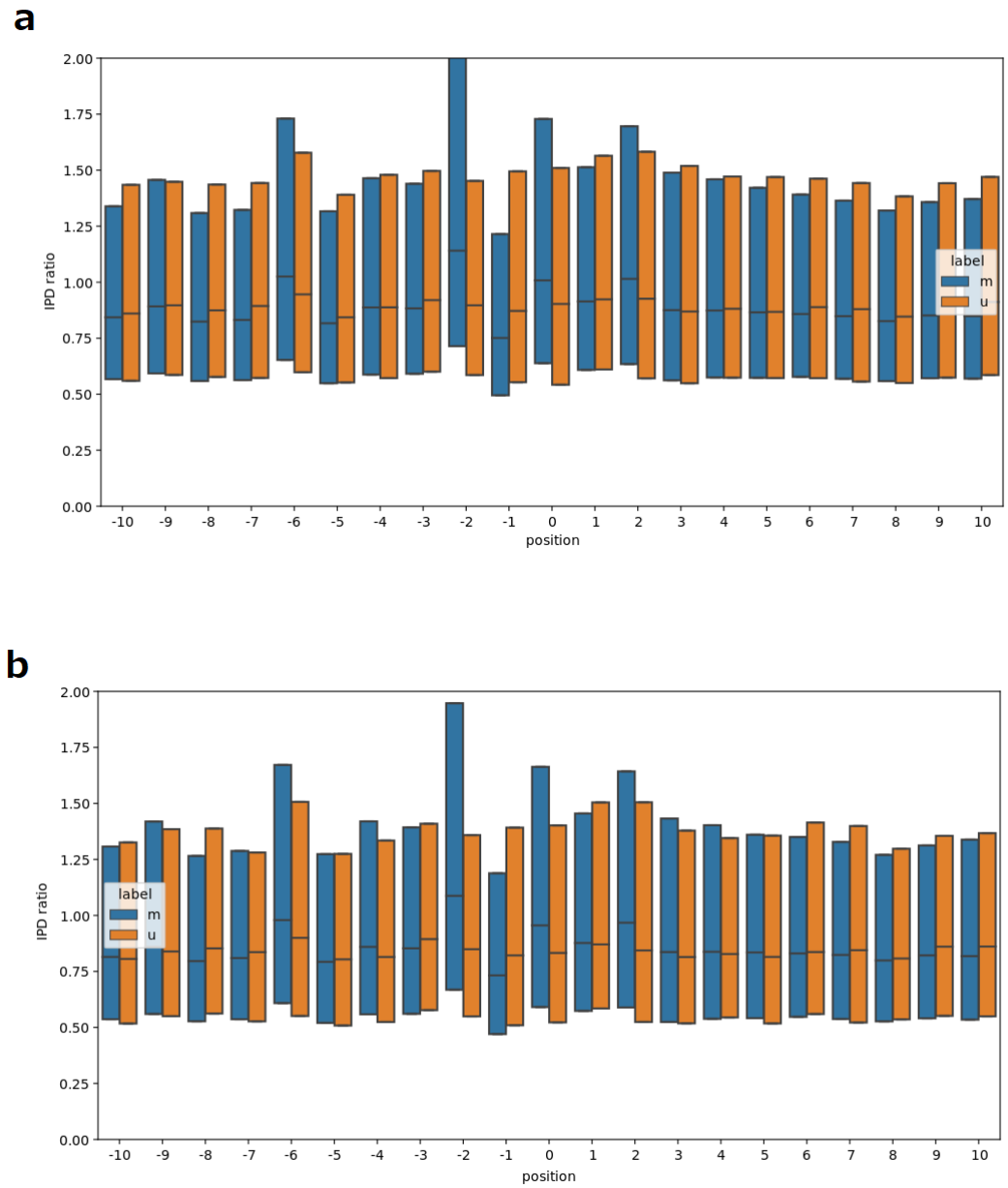


Figure 2.2: Boxplots for IPD ratio profiles around CpG sites. Mean values and .25/.75 quantiles are shown for each relative position to CpG sites. **A.** CpG sites were classified according to its methylation status. **B.** CpG sites were classified according to methylation status of the CpG sites in the opposite strand.

2.4 CpGs have a distinct IPD ratio profiles in CpG methylated samples

Most CpG sites in intact human genome are methylated with the notable exception of those in CpG islands. Consequently, CpGs have a distinct IPD ratio profiles compared to other 15 2-mers even before we classify them into methylated ones and unmethylated ones. Figure 2.3a-d show IPD ratio profiles for each 2-mer, and it is evident that CpGs have the most characteristic bumps and peaks in IPD ratio profile. Though we are not going to elaborate the possibility here, we may have a rough idea on the methylation status of the DNA sample, *e.g.*, whether it underwent amplification or not, which would have erased methylation. Therefore, to examine 2-mer IPD ratio profiles would be a harmless preliminary analysis before we apply more sophisticated methods.

2.5 The form of the regional prediction model

Let us turn to the specification of the model to predict methylated CpGs. Since the model is detailed in the chapter for the implementation and application of AgIn, we tried to pick a rather different perspective to avoid unnecessary repetition. The central idea behind AgIn is it focuses on predicting regional methylation status composed of a certain number of CpG sites, instead of methylation status of individual CpG sites. The motivation of this formulation is twofold. On one hand, methylation status of neighboring CpG sites are strongly correlated, thus such a formulation would enhance the prediction as it effectively models the biology of cytosine methylation[11, 26, 112]. On the other hand, it is costly to obtain enough sequencing data for accurate prediction of individual methylation status, so in a sense, we must reduce the resolution of prediction from individual CpG sites to region of CpG sites. The resulted model is described in the chapter for AgIn. We may simply summarize the form of the model as "linear separation of regional IPD ratio profiles via a

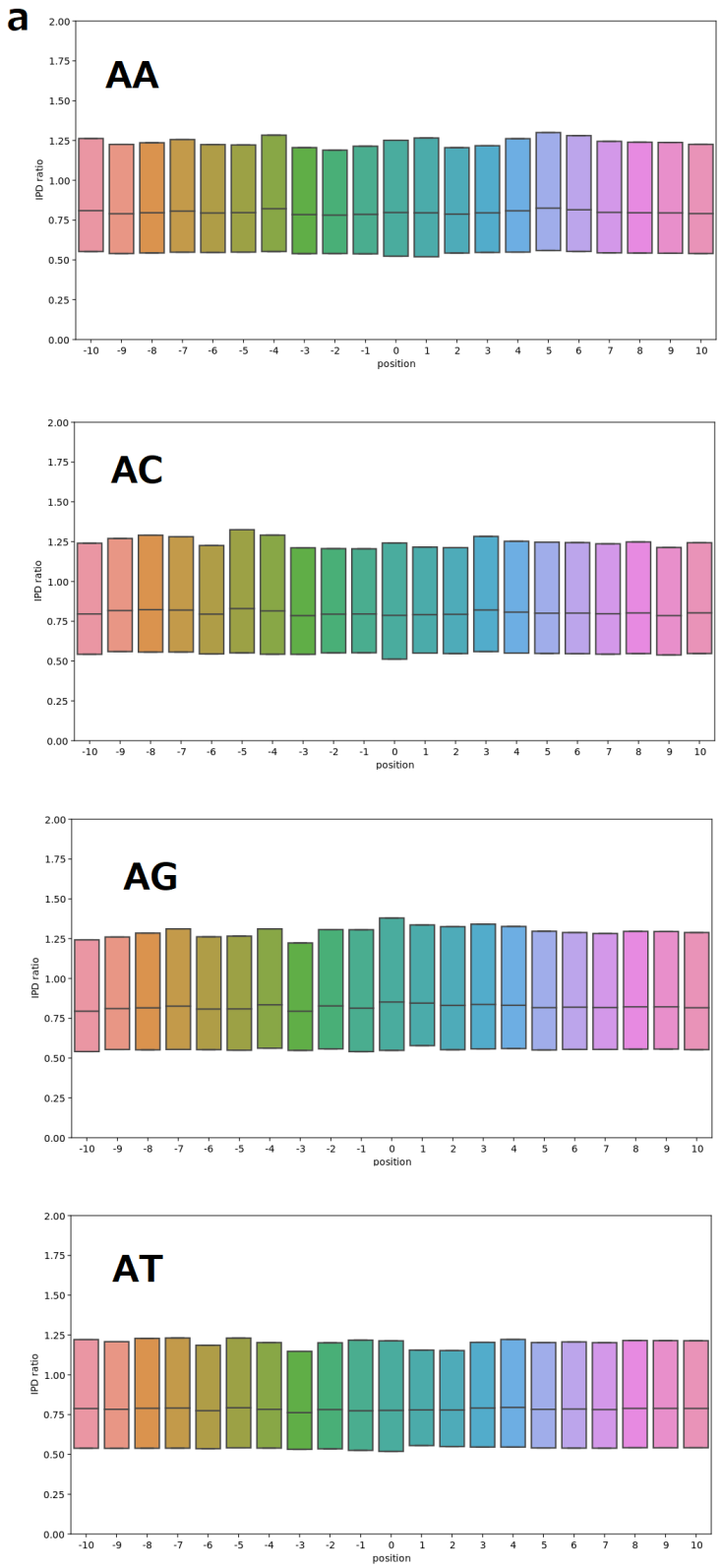
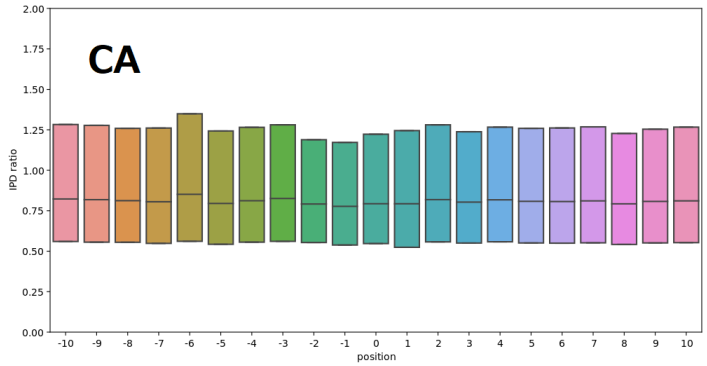
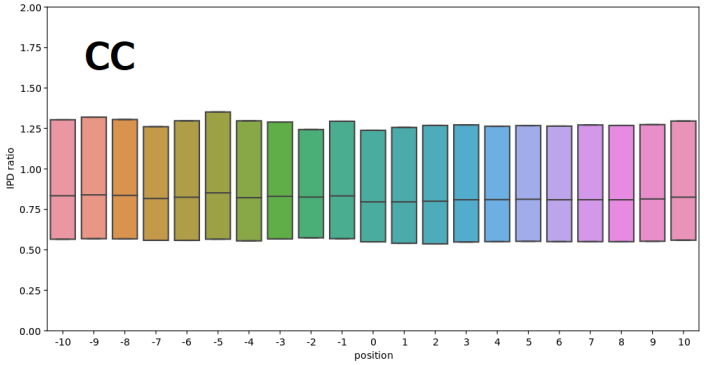
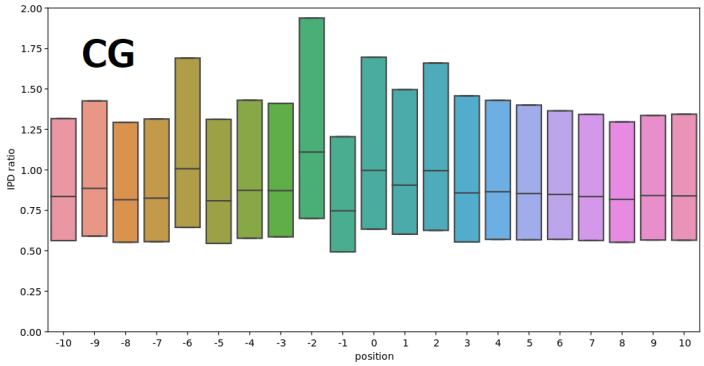


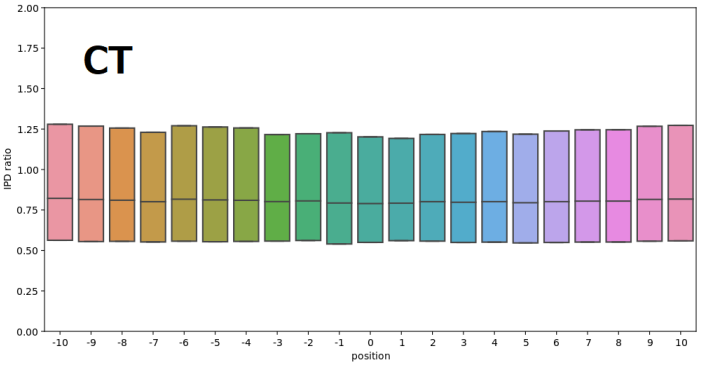
Figure 2.3: **A-D.** Mean values and .25/.75 quantiles are shown for each relative position to each 2-mer. CpG had the most characteristic IPD ratio profile among other 2-mers.

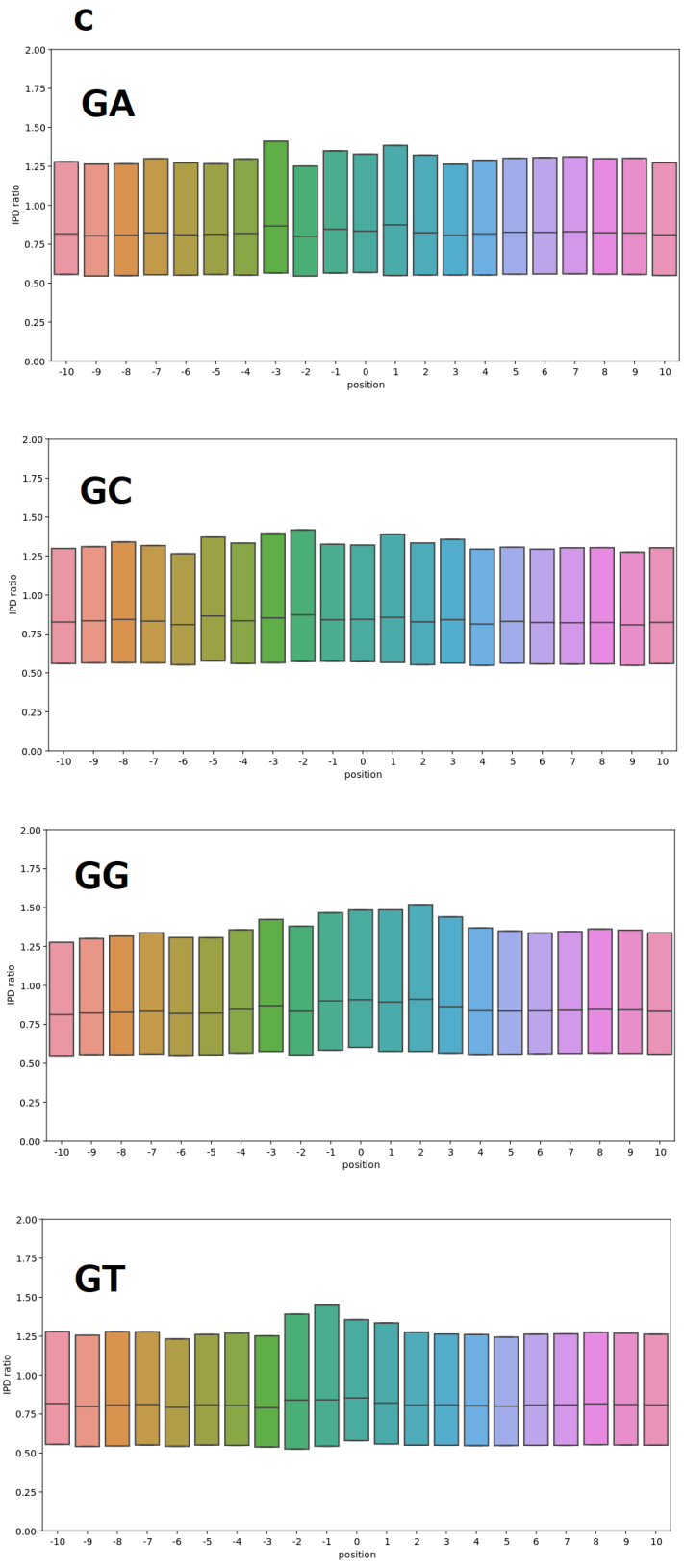
b

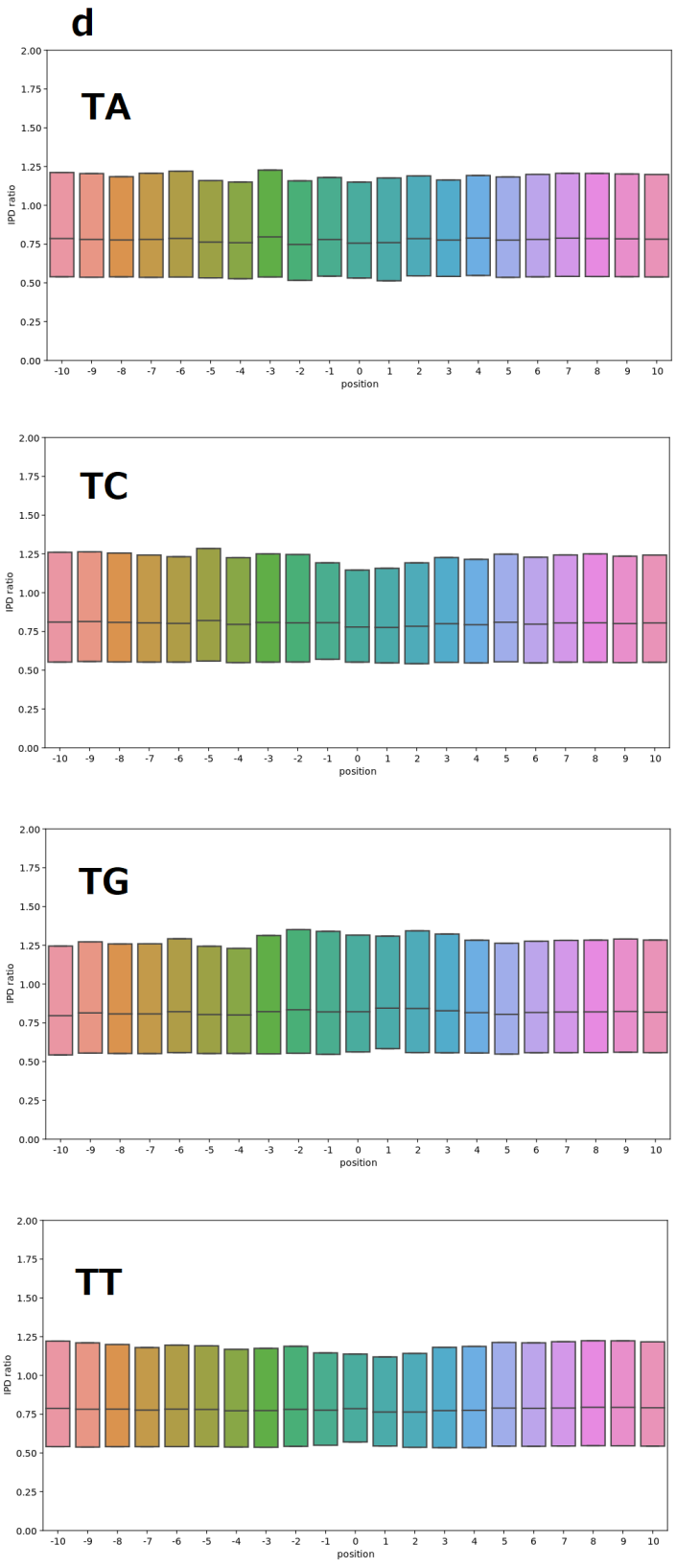

CA
IPD ratio
position


CC
IPD ratio
position


CG
IPD ratio
position


CT
IPD ratio
position


C
GA
GC
GG
GT
IPD ratio
position
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10


d
TA
TC
TG
TT
IPD ratio
position
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10

normal vector, β, and a bias term, γ". In the rest of this chapter, we will detail how each of these parameters were determined.

2.5.1 Optimization for beta using linear discriminant analysis

Beta (β) was optimized so that it can discriminate the IPD ratio profiles from methylated CpG sites and those from unmethylated ones. As we intend to use the vector for linear separation of regional IPD ratio profiles, it would be natural to optimize it so that it can work best for linear separation of individual IPD ratio. For that purpose, we applied linear discriminant analysis (LDA) for the set of individual IPD ratio profiles of CpG sites, each was labeled with its methylation status observed by bisulfite sequencing data. LDA will give us an optimal hyperplane (residing in the dimension of the data, 21 in our case). The normal vector which determine the optimal hyperplane was selected as our β.

2.5.2 Optimization for gamma (γ) and other parameters

Once the β was optimized, the rest of the parameters are optimized by examining the final prediction results of the model against the methylation status called by bisulfite sequencing. Other parameters to be optimized include, (1) γ (bias term) for controlling the sensitivity-specificity tradeoff, and (2) L for minimum number of CpG sites for each predicted block. Procedure will be detailed in the following chapter, which is on the specification of the fully developed model and its applications. So, let me move onto the good part of this study.

Chapter 3

Algorithm for CpG methylation detection - AgIn

3.1 Introduction

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 [120, 3, 76, 129, 99, 80, 109]. Some of human transposable elements (TEs), such as long interspersed nuclear elements (LINE), transpose actively within somatic cells along differentiation of neural tissues, and are partly regulated by DNA methylation [81, 82]. Each family of human TEs exhibits a variety of methylation statuses in different tissue types, which was found by looking at the mixture of methylation information on the consensus sequence of TEs in the same family [124]. Many human diseases are also associated with DNA methylation state of TEs. In particular, unmethylation of repetitive elements (REs), such as LINE-1 (L1) elements, has been related to some cancers [121, 94].

This work has been published in:
Suzuki, Yuta, et al. "AgIn: measuring the landscape of CpG methylation of individual repetitive elements." Bioinformatics 32.19 (2016): 2911-2919.

Although only a few L1 elements exhibit activity in the human genome [8], it has been reported in various cancer genomes [61, 38], and importantly, transposition is correlated with unmethylation in the promoter region of L1 elements [116]. Therefore, it is essential to develop an experimental framework that can characterize the methylation state of REs 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 [20, 70, 75, 71, 41]; 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. Especially, the younger and more active transposons retain higher fidelity and are therefore difficult to address using short reads.

The PacBio RS II sequencing system uses DNA polymerases to perform single-molecule real-time (SMRT) sequencing [54, 27], and is able to sequence reads of an average length of >10 kb. It is also able to sequence genomic regions with extremely high GC content. A striking example is the sequencing of a >2-kb region with GC content of 100% [72], indicating that SMRT sequencing is less vulnerable to sequence composition bias than first/second-generation sequencing is.

SMRT sequencing of bisulfite-treated DNA fragments may allow identification of DNA methylation; however, this approach is unlikely to process long, highly identical repeats because bisulfite treatment breaks DNA into fragments of <1500 bp [79, 128]. Instead, we explored another advantage of SMRT sequencing to detect DNA modifications directly.

3.2 Approach

In SMRT sequencing, we observe the base sequence in a single DNA molecule as the time course of the fluorescence pulses which reflect the incorporation processes of nucleotides. From this time course information, we define the inter-pulse duration (IPD), the time interval separating the two pulses of consecutive bases. Importantly, the IPD of the same genomic position varies and has a significant and predictable response to the presence of DNA modifications and damages [34].

Since the IPD tends to be perturbed systematically when DNA modifications are present, SMRT sequencing has been used to detect 5-hydroxymethylcytosine [34], N4-methylcytosine [19], N6-methylademine [34, 31, 32, 40], and damaged DNA bases [18] in bacteria and mitochondria. Though the sequence motifs with modifications can be detected with very low coverage [9], estimation of 5-methylcytosine (5-mC) residues using low-coverage reads is challanging. It requires extensive coverage (∼500x) at each position to clarify the base-wise 5-mC state and therefore becomes costly [34, 31, 96]. 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 [17]. 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 (∼10%) in vertebrates and generally do not exist in isolation but often range over long hypomethylated regions [92, 37, 26, 11, 106, 86, 124]. Su *et al.* reported that the average length of hypomethylated regions in five human cell types is ∼2 kb [112]. Thus, estimating regions with unmethylated CpG sites is informative in most cases. Moreover, integrating kinetic information over many CpG sites in a long region can increase the confidence in detecting methylation when the status of those sites is correlated. Therefore, it shows promise for predicting the methylation status in a block using low-coverage SMRT reads. In the rest of this article, we examine the feasibility of the approach and present a novel computational

algorithm that integrates SMRT sequencing kinetic data and determines the methylation status of CpG sites.

3.3 Methods

3.3.1 Outline of our method AgIn

Figure 3.1A shows a schematic representation of the basic concept of our method. To eliminate the context-dependent fluctuation of the IPD values, we calculated the IPD ratio (IPDR) on each genomic position as previously described [34]. This normalization is essential to compare the IPD values from different genomic positions with various sequence contexts. Then, we defined the IPDR profile of a CpG site as an array of IPDR measurements of 21 bp surrounding the CpG site because these neighboring positions have proven to be effective in predicting 5-hydroxymethylcytosine, N4-methylcytosine, and N6-methylademine in bacteria genomes in previous studies [34, 18, 19, 31]. With low coverage, the IPDR profiles at individual CpG sites are noisy and insufficient for determining whether each CpG site is methylated or not. However, if we could somehow identify the boundaries of hypomethylated/hypermethylated regions and take the average of the IPDR profiles for the CpGs within each region, then it 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. Averaging the IPDR profiles is also expected to alleviate the possible confounding effect from other types of modifications found in DNA. An example of our prediction for the human genome is shown in Figure 3.1B. Our method was able to estimate hypomethylation of long duplicated regions while the bisulfite sequencing provided little information. Supplementary Figure 3.S1C illustrates another example in which both methods were consistent in showing hypomethylation in the gene promoters.

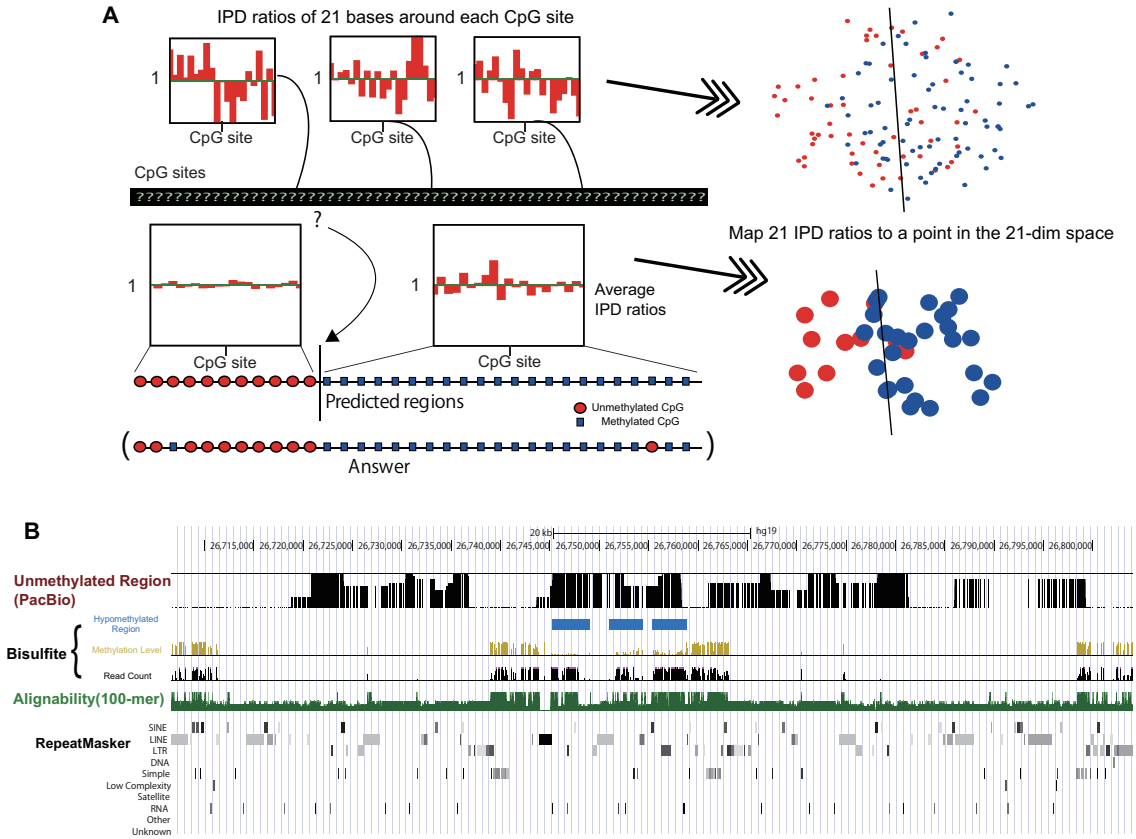


Figure 3.1: Outline of our integration method. **A.** The top three distributions show the typical inter-pulse duration ratio (IPDR) profiles within 10 bp of the CpG sites in the raw data. The IPDR profiles of individual CpG sites were treated as points in the 21-dimensional feature space. Red-colored unmethylated CpGs and blue-colored methylated CpGs are often 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 Supplementary Fig. 3.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 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. Both methods are consistent in showing hypomethylation 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 hypomethylation in these regions.

3.3.2 Estimating the methylation status at each CpG site

Suppose that the focal genome has n CpG sites. We denote the genomic position of C of the i-th CpG site by $p_i (i = 1, \ldots, n)$. For example, if the C of the second CpG site is at the 10th genomic position, "$p_2 = 10$." Our goal is to predict the methylation status, unmethylated or methylated, at p_i using information of the read coverage and the IPDRs at positions surrounding p_i. 21 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.

We used the SMRT Analysis pipeline to process raw kinetic data from SMRT sequencing to obtain the mean IPDR and the read coverage at each genomic position. Let r_i and r'_i denote the mean IPDR associated with position i of the forward and reverse strands, respectively, and let c_i and c'_i denote the read coverage at position i of the forward and reverse strands, respectively. To achieve a better prediction, we derive a modified IPDR vector from the raw read coverage and the IPDRs within 10 bases around p_i. For this purpose, we consider that 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 p_i to the position of C of the focal CpG in the plus strand, 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 $j = -10, \ldots, +10$. We attach more importance to the IPDR values associated with a higher read coverage and we quantify 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 take the sum of all the products and normalize it by dividing it by the total coverage. Finally, we obtain the 21-dimensional modified IPDR vector for 21 genomic positions around CpG site p_i:

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

We are now in a position to define a classifier that uses $\hat{X}(p_i)$ as explanatory variables and predicts the methylation status at p_i. We attempted to use linear discriminant analysis (LDA) with the discriminant function

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

where we optimized the values of coefficient vector β and variable γ using bisulfite sequencing data as the training data set to improve the prediction. Supplementary Figure 3.S1A and 3.S1D shows the optimized vector β that we used in this study. We do not claim this vector is the simplest one since excluding the low-contributing components from the parameter degraded the accuracy only by a little (Fig. 3.S3G). 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.' Our goal is to achieve a higher accuracy using a lower read coverage in order to reduce the cost.

3.3.3 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 motivates 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 the CpG sites in region A and its associated discriminant function:

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

Processing a longer region with sufficient CpG sites can improve the accuracy of the prediction, although it may overlook smaller regions. In our analysis, we imposed the constraint that each region contained at least b CpG sites. For example, we can set b to 50 because the average length of hypomethylated regions in five human cell types is approximately 2 kb [112] 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 coverage,

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

A with a larger magnitude of $w(A)\mathrm{F}(A)$ is better for prediction.

3.3.4 Decomposing the genome into hypomethylated/hypermethylated 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 regions occur alternatingly along the genome. We calculate 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 simplify this problem, we here mention one important characteristic of SMRT sequencing, that is, read coverage is not affected by the sequence composition [6, 130, 28, 52, 72]. Thus, the average coverage in A is constant at any position within 10bp relative to CpGs. Technically, we can assume that the average of coverages at the j-th position around all the

CpG sites in region A is a constant $\bar{c}$ that is dependent of A but is independent of j:

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

where $|A|$ denotes the number of CpG sites in A. 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 simplify the objective function:

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

where s_i denotes $21(\gamma\bar{c} + \sum_j \beta_j (c_{p_i+j} r_{p_i+j} + c'_{p_i+1-j} r'_{p_i+1-j}))$.

Finally, the objective function is a linear combination of s_i:

$$\begin{aligned}
& \sum_{\lambda \in \Lambda} w(M_\lambda)\mathrm{F}(M_\lambda) + \sum_{\mu \in M} -w(U_\mu)\mathrm{F}(U_\mu) \\
= \; & \sum_{\lambda \in \Lambda} \sum_{p_i \in M_\lambda} s_i + \sum_{\mu \in M} \sum_{p_i \in U_\mu} (-s_i)
\end{aligned}$$

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 motivates us to design an $O(n)$-time dynamic programming algorithm for calculating the optimal value efficiently. We consider the subproblem involving the first i CpG sites among all n sites, and let S_i^M and S_i^U be the maximum values of the objective function when the last i-th CpG site is methylated and unmethylated, respectively. S_i^M and S_i^U meet the following recurrences:

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

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

3.4 Results

3.4.1 SMRT sequencing and bisulfite data benchmark

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 (Supplementary Methods). 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 (Supplementary Methods). In total 2848641, 7279594, and 19115712 subreads mapped 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 (Supplementary Table S1).

As CpG methylation status reference data, we used the testes methylome of the medaka Hd-rR inbred strain by way of Illumina bisulfite sequencing [92]. 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%), allowing us to focus on CpG sites only. We evaluated the prediction accuracy of our integration method using the 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 then defined the methylation status as 'unmethylated' if the score x/y 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.*, $y \geq 10$) in order to make sure the scores were robust enough.

3.4.2 Computational performance

Our linear-time algorithm allows us to handle vertebrate-scale genomes with millions of CpG sites in a reasonable amount of time. It took 2.265 sec. on average to process 1 Mbp (1191 s to handle 525.7 Mb of medaka genome v.1) using a laptop PC (Intel i7-3612QM processor with a clock rate of 2.10 GHz and 7.8 GB of main memory).

3.4.3 Predicting the methylation state from kinetic data

We implemented our method using linear discrimination of the vectors of (average) IPDR profiles around the 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 (Fig. 3.1A). For better accuracy, we optimized two parameters of the decision hyperplane, the orientation and the intercept. Supplementary Figure 3.S1A (for P6-C4 reagents) and S1D (for P4-C2 reagents) show the optimized orientation. Our method divides the genome into regions containing $\geq b$ CpG sites, such that each region is either hypomethylated or hypermethylated. 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 when we could use a smaller value of b ($= 30, 35, 40, 45$) without degrading the accuracy of prediction.

We predicted the methylation status of each CpG site by checking whether the CpG site was located in an hypomethylated or hypermethylated region output by our 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 an unmethylated status as positive and a methylated status as negative because we are more interested in identifying rare hypomethylated regions accounting for a small portion (e.g., $\sim$10%) of CpG

sites.

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, and set the coverage to 0 otherwise, which slightly reduced the original average read coverage, 31.06-fold, to 29.9-fold on the chromosome 1. 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 [92]. 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 ($-8.0 \leq \gamma \leq 5.0$) (Fig. 3.2A and Supplementary Fig. 3.S2-3). When we used 100% of 29.9-fold subreads, setting b to 35 outperformed the other values (Fig. 3.2A). 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. To examine the coverage effect, we used five subread sets of coverage 20%, 40%, 60%, 80%, and 100% of 29.9-fold. For coverages of 20% and 40% of 29.9-fold, setting b to 50 performed best (Supplementary Fig. 3.S3). Both sensitivity and precision were $\sim 90\%$ for $b = 45$ even if the coverage is relatively small, 60% of 29.9-fold (Supplementary Fig. 3.S3C). In selecting b, it was suggested to use a larger value ($b = 50$) when the read coverage is small (15∼20-fold) so that the cumulative coverage (750∼1000-fold) is large enough. One can use a smaller value ($b = 35$) with sufficient read coverage (∼30-fold), and b can be decreased gradually with deeper coverage. Setting b to 1 corresponds to the case where the methylation state of each CpG is predicted independently, but it could not achieve a good accuracy, which confirmed the merit of our aggregating approach (Supplementary Fig. 3.S3F). The ROC curve, the tradeoff between false-positive rate and sensitivity, is also shown in Figure 3.2B. Overall, sensitivity and precision of our

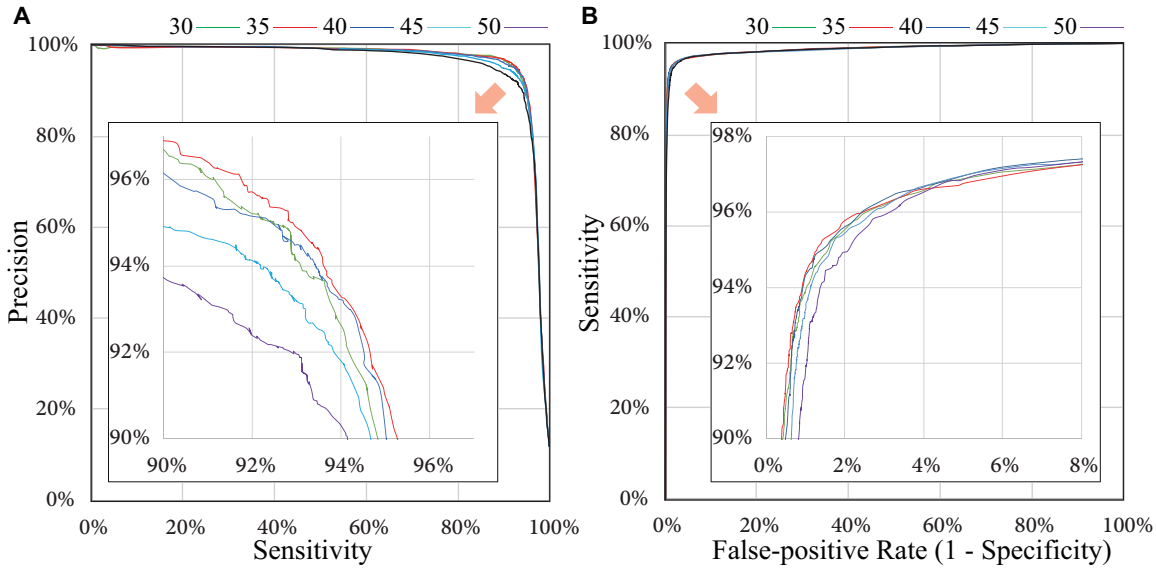


Figure 3.2: Accuracy of our method. **A.** The sensitivity and precision (proportion of true-positives among the predicted positives) are evaluated on individual CpG sites when we change the intercept of the hyperplane (between -8.0 and 5.0) and set the minimum number of CpG sites in a region, b, to 30, 35, 40, 45, and 50. **B.** The ROC curve of false-positive rate and sensitivity.

method are substantially high using a reasonable coverage of SMRT subreads.

3.4.4 Handling intermediate methylation states

We have introduced the 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 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 should improve the overall prediction accuracy.

Thus we extended our method in order to achieve more informative multi-class prediction and quantify the methylation level of each CpG, which we call *discrete methylation level* (DML, Supplementary Methods). Specifically, DML is calculated as the average prediction over the set of 10 parameters with different sensitivity-specificity combinations, thus it measures the robustness of the prediction. 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 (Supplementary Fig. 3.S4C,E), and we confirmed that measurements on 92.0% of CpG sites were in concordance within 0.25. 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 after normalizing the beta values (Supplementary Methods). We observed a weaker correlation ($R = 0.816$, Supplementary Fig. 3.S4D) 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 [119]. With the sequencing depth in our case, CpG sites with intermediate methylation were more difficult to predict than completely methylated/unmethylated cases (Supplementary Fig. 3.S4E). Therefore, excluding the prediction with intermediate levels improved the accuracy of the binary prediction (Supplementary Table S2). We concluded that DML serves to reflect the quantitative nature of methylation status in the samples to some extent, and is informative in achieving more accurate prediction as well.

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

We investigated how individual occurrences of repetitive elements (REs) were methylated in the human genome (Fig. 3.3A). Since some occurrences of REs contain no or very few CpG sites, we only consider those occurrences with at least 10 CpGs to exclude 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$50% of CpGs. We found that >96% were covered for every repeat type. To draw robust conclusions, we further applied a stringent quality control to each repeat occurrence so that the average read coverage be >5. Although this step reduced the number of repeat occurrences to be analysed by 3$-$18%, this could be mitigated simply by producing more data. Finally, we treated an occurrence as hypomethylated if $\geq$50% of CpGs were predicted as unmethylated. Similarly, we considered an occurence as methylated intermediately if $\geq$50% of CpGs were predicted as 0.3$\sim$0.7 in DML measurement. Fractions of hypomethylated repeat occurrences vary considerably among different classes of REs, from $\sim$1% for L1 and Alu to $\sim$50% for MIR and >70% for simple repeats and low-complexity regions. The fraction of intermediately methylated repeats was 1.4% among all the repeat classes.

To validate our prediction regarding the repeat occurrences, we selected 21 regions for

A

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%

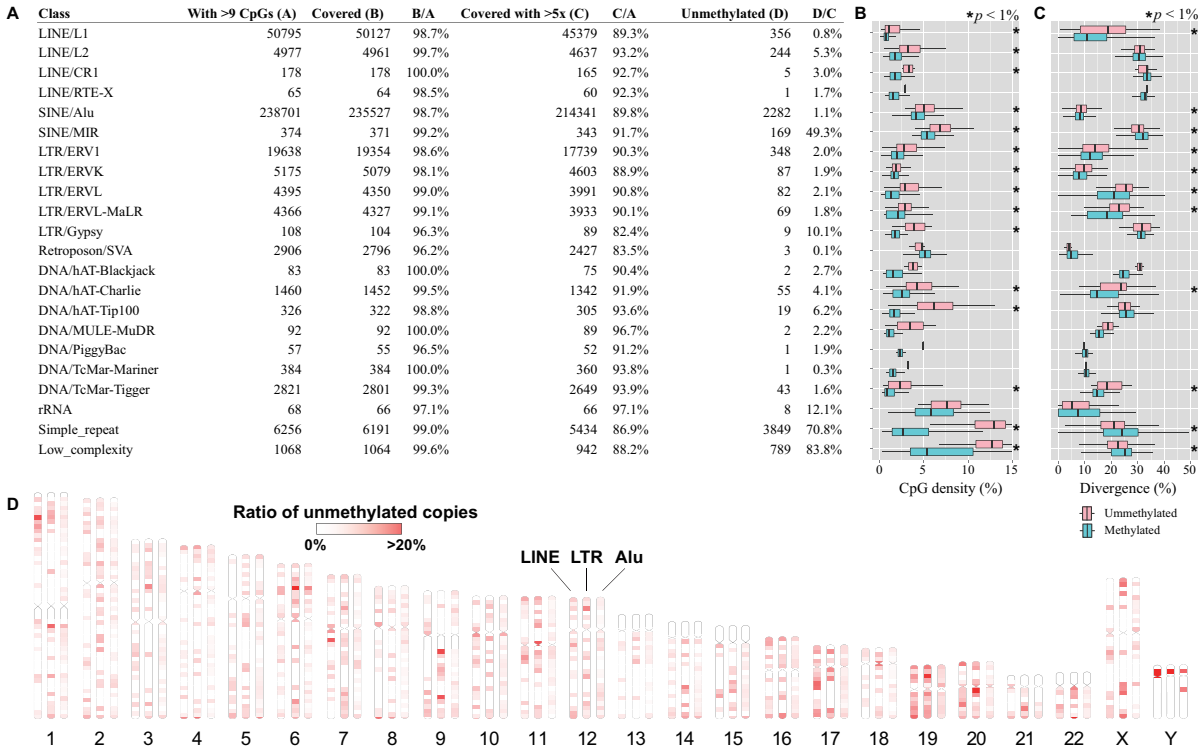


Figure 3.3: Epigenetic landscape of repetitive elements in the human genome. **A.** The table shows a summary of methylation status on repetitive elements (REs) that we select using the Repeat Library 20140131 (Smit, A., Hubley, R. and Green, P. Repeatmasker open-4.0 at http://www.repeatmasker.org). **B-C.** Distribution of CpG density (**B**) and sequence divergence from the representative in each repeat class (**C**) for methylated (cyan) and hypomethylated (pink) repeat occurrences. The asterisks indicate statistical significance ($p < 1\%$) determined by the U test. **D.** Genome-wide distribution of hypomethylated REs. The ratio of hypomethylated repeat occurrences to all occurrences in each 5-Mb bin is indicated by color shadings. We used the Ideographica web server to generate the image.

bisulfite Sanger sequencing, designed primers for nested PCR (Supplementary Table S3), and could amplify six regions (Supplementary Methods), indicating the difficulty in observing DNA methylation of REs using traditional bisulfite Sanger sequencing. In five (1 L1, 3 LTRs, 1 MIR) among the six amplified regions, we confirmed the consistency between our prediction and the methylation state observed by bisulfite Sanger sequencing (Supplementary Fig. 3.S5). The other one L1 element was predicted hypomethylated. In this region, however, 5 unmethylated CpG sites were followed by 5 methylated CpG sites, which showed our method was not reliable in determining the precise boundary and the individual calls should be interpreted carefully.

We then examined the features for characterizing the differences between hypermethylated and hypomethylated REs. First, CpG density was significantly higher in the hypomethylated occurrences in almost all classes of REs ($p < 1\%$, Fig. 3.3B). This observation was consistent with the known association between CpG-rich regions and unmethylation because methylation leads to depletion of CpG sites through deamination [22]. Second, sequence divergence from the representative in each repeat class showed a correlation with methylation status (Fig. 3.3C). 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. 3.3C), presumably because younger copies of a repeat element are less divergent and are likely to be targets of DNA methylation. Kernel principal component analysis (PCA) using spectrum kernel suggested, for some repeat types, that the methylation statuses were correlated partly with sequence features (Supplementary Fig. 3.S6).

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 REs for this analysis. We calculated the ratios of hypomethylated copies to all REs in individual non-overlapping bins 5 Mb in size (Fig. 3.3D). 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 hypomethylation 'hot spots' of LTR elements, *e.g.*, in chromosomes 6 and 9 (Supplementary Fig. 3.S7). 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 REs.

We further investigated the methylation states of LINE/L1 elements, the only known active autonomous retrotransposons in mammals [35]. 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) [91]; namely, 1) L1s with intact in the two ORFs that are likely to exhibit retro-transposition activity, 2) L1s with an intact ORF2 but a disrupted ORF1, and 3) non-intact L1s with two ORFs disrupted. We obtained the positions of these human LINE/L1 elements from L1Base [91] and analyzed their methylation stateses (Supplementary Table S4). Although 0.61% of non-intact L1s were hypomethylated, all of L1s with intact in two ORFs and L1s with an intact ORF2 were hypermethylated. We also checked the presence of LINE insertions that were novel to the hg19 reference genome. We assembled the SMRT reads using the FALCON assembler 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 [91]. 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 (their locations are in Supplementary Fig. 3.S8) were estimated to be methylated. These results confirmed putatively active LINE/L1 elements with intact ORFs were preferentially methylated.

3.4.6 *Tol2* transposable element in medaka

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 [48]. 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 [44]. Nevertheless, observing the methylation status of each *Tol2* copy using short reads is difficult, because *Tol2* is 4682 b in length, and ∼20 highly similar copies of *Tol2* exist in the genome [50].

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. BLAST search identified 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 the publicly available bisulfite-treated reads from the testes of the Hd-rR strain to these contigs and determined the methylation level on every 100-bp window using Bismark software.

The methylation status of these *Tol2*, observed by SMRT reads and bisulfite-sequencing, are shown in Figure 3.4. 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 using SMRT reads to clarify DNA methylation states of highly identical REs such as active transposons.

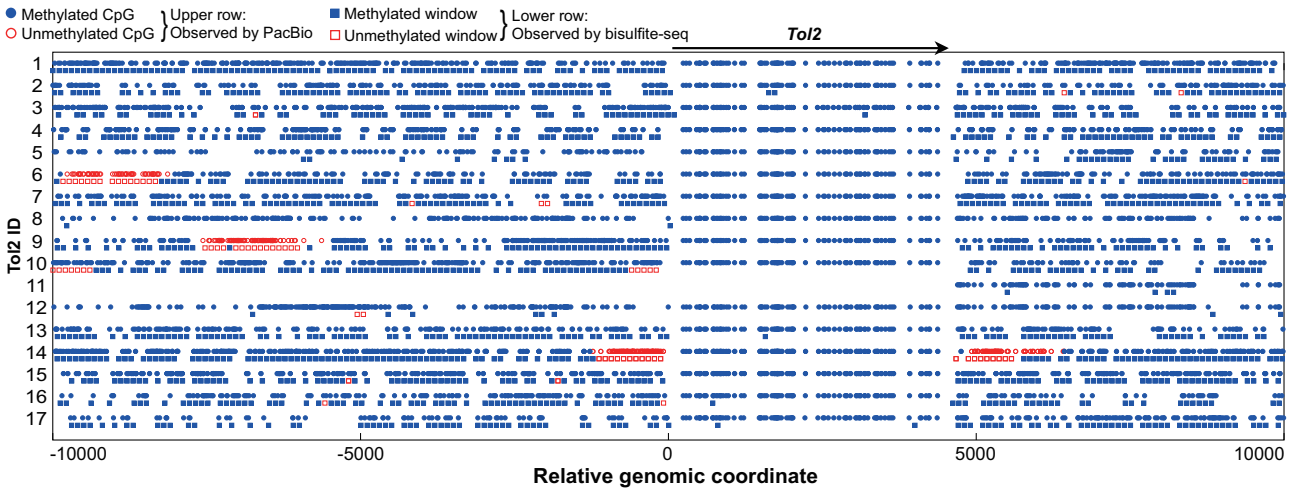


Figure 3.4: Methylation analysis of *Tol2*, a 4682-bp long autonomous transposon, in medaka. The new genome assembly of SMRT reads had 17 regions (contigs) that contained complete *Tol2* copies. The circles show our prediction of the methylation state of CpG sites, while the rectangles show the methylation 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.

3.5 Discussion

In this study, we addressed the problem of uncovering the landscape of DNA methylation of repetitive elements (REs). 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 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. 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 REs 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 REs 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 REs in personal human genomes.

We demonstrated the use of long SMRT reads can increase the potential comprehensiveness of the epigenetics study. In addition, our method can substantially reduce laboratory work. For example, in the projects of resequencing or *de novo* assembly using SMRT sequencing, you can call the methylation statuses of the sample as well, completely *in silico*, without any additional experiment. This is another important strength compared to conventional bisulfite sequencing or affinity-based assays.

3.6 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).

3.7 Supplementary Methods

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 $\sim$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$\times$ volume ratio of AMPure beads (Pacific Biosciences, Menlo

Park, CA, USA). SMRTbell^{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^{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^{TM} libraries were prepared with the DNA Template Preparation Kit 2.0 (3$\sim$10 kbp) (Pacific Biosciences, Menlo Park, CA, USA). Briefly, sheared DNA was end-repaired, and hairpin adapters were ligated using T4 DNA ligase. Incompletely formed SMRTbell^{TM} templates were degraded using a combination of exonucleases III and VII. The resulting DNA templates were purified using (0.45$\times$) 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^{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.

Handling intermediate or ambiguous methylation states

Suppelmentary Figure 3.S4 depicts 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 (Supplementary Fig. 3.S4A). 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 (Supplementary Fig. 3.S4B). Specifically, 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. We then defined the *discrete methylation level* (DML) ranging over $[0, 1]$ 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.

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, 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.

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 $\times$ 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 $\times$ 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 [57] for analysis and visualization of the methylation patterns (Supplemental Fig. 3.S5).

Other data sources and data visualization

Figure 3.1B and Supplemental Figures 3.S7, 3.S8 were produced using the UCSC Genome Browser (http://genome.ucsc.edu/) [47]. We used human bisulfite sequencing data and unmethylated regions available in the GEO database [110].

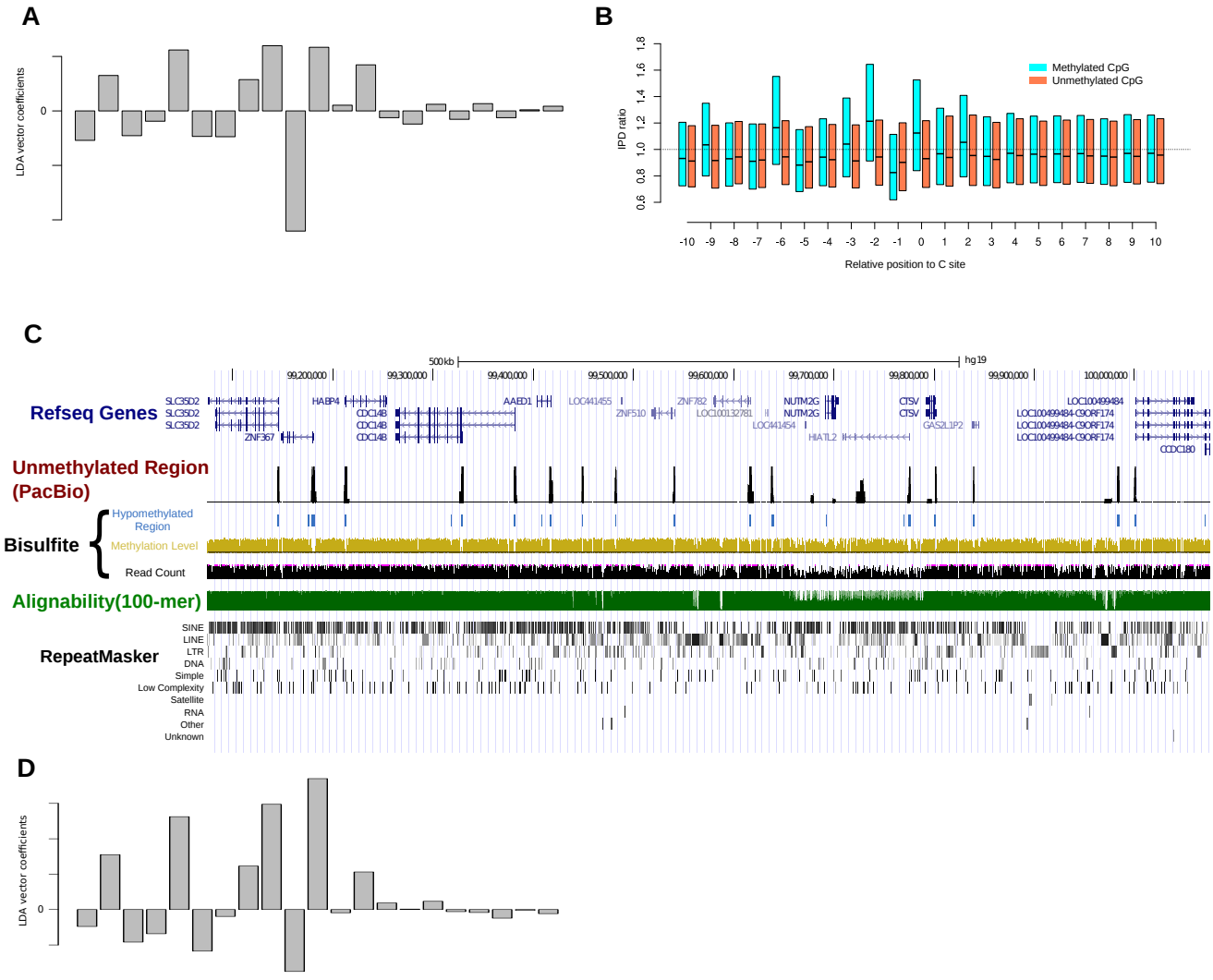


Figure 3.S1: **The normal vector used for prediction. A.** The normal vector $\boldsymbol{\beta}$ used for prediction with P6-C4 reagent. We calculated $\boldsymbol{\beta}$ as follows. Firstly, we classified the CpGs on the scaffold 1 in the medaka Hd-rR genome (version 1) into methylated CpGs and unmethylated CpGs according to bisulfite sequencing data. Next, for each CpG site, we calculate the IPD ratio profiles as the 21-dimensional vectors based on SMRT sequencing kinetics data. Then, using LDA (Linear Discriminant Analysis), we tried to find the best hyperplane that could separate these IPD ratio profiles into each class, namely, methylated or unmethylated. The normal vector of this hyperplane is denoted by $\boldsymbol{\beta}$. **B.** The average IPDR profiles around unmethylated and methylated CpG sites. The x-axis shows the positions within 10 bp of the focal CpG site at the position represented by 0. The y-axis indicates IPDR values. The red- and blue-colored box plots at each position show the distributions of IPDR values around unmethylated and methylated CpG sites, respectively. The bottom, middle and top of each box plot indicate the first, second, and third quartiles, respectively, of the distribution. **C.** An example in which both our method and bisulfite sequencing are consistent in showing unmethylation in gene promoters. The tracks are similar to those in Figure 3.1B. **D.** The normal vector $\boldsymbol{\beta}$ used for prediction with P4-C2/C2-C2 reagent.

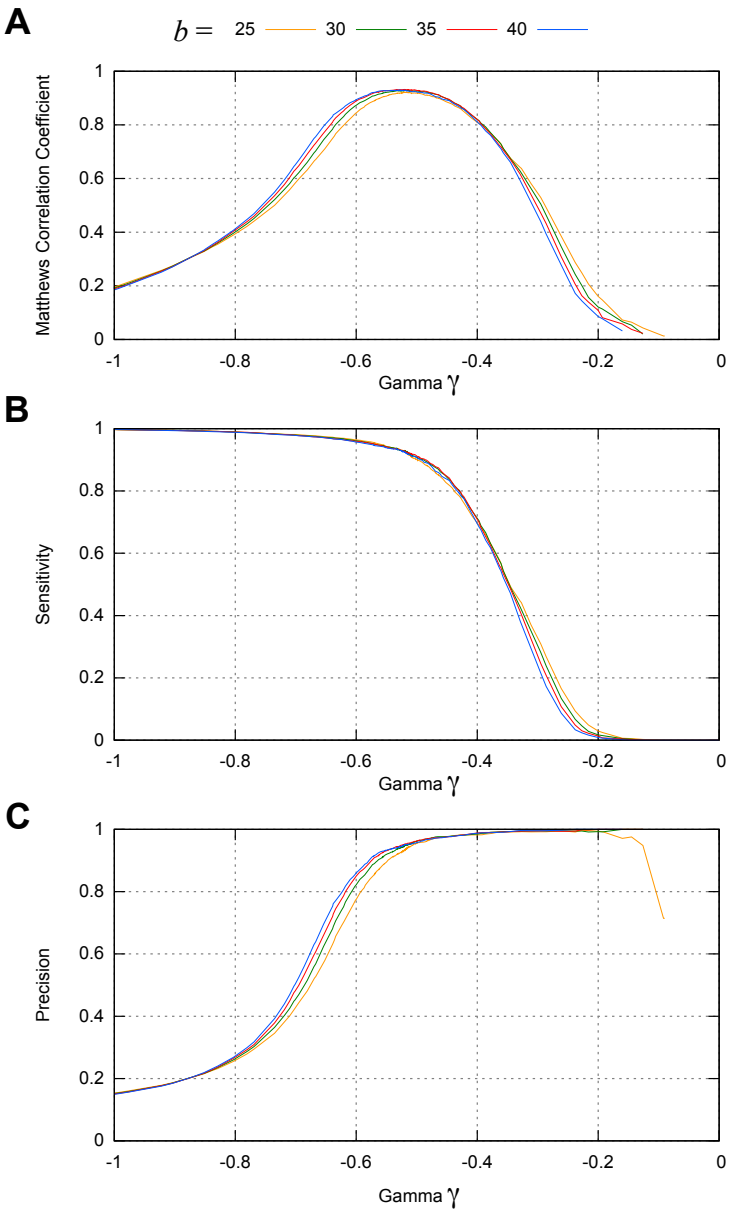
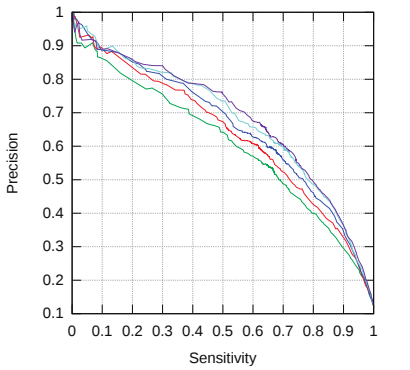
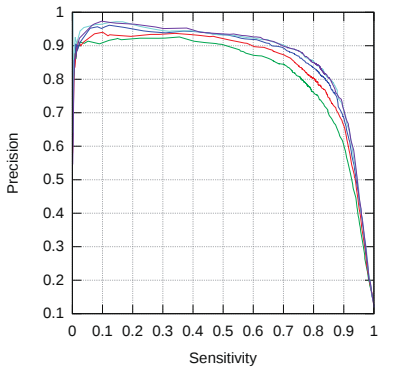


Figure 3.S3: **Accuracy metrics on the chromosome 1 of the medaka Hd-rR genome (version 2). A-C.** Matthew's correlation coefficient **(A)**, sensitivity **(B)**, and precision **(C)** as a function of the intercept of the hyperplane γ, on the chromosome 1 in the medaka genome (version 2) with a 29.9-fold mapped read coverage. Matthew's correlation coefficient represents an overall accuracy of our prediction. The differently colored curves correspond to the different lower bound of number of CpG sites, denoted by b, that was used for the prediction. Our prediction achieved 93.0% sensitivity and 94.9% precision at $b = 35$ and $\gamma = -0.526$. Or sensitivity (93.67%) and precision (93.88%) are close to each other when $b = 35$ and $\gamma = -0.540$.

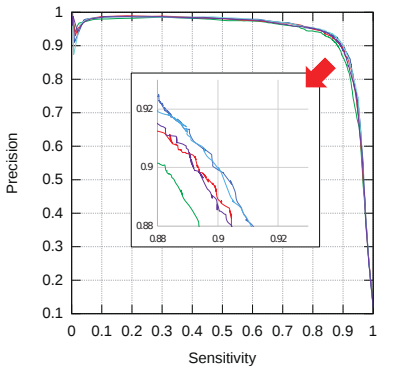
A (20% of 29.9x)



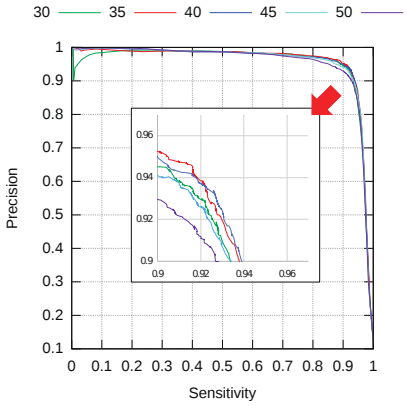
B (40% of 29.9x)



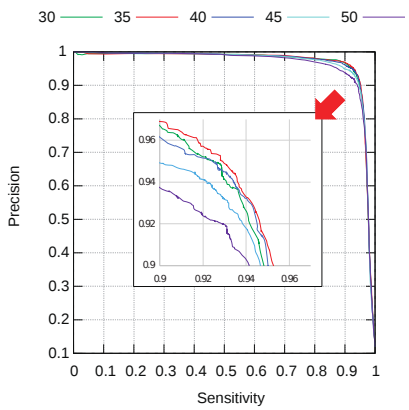
C (60% of 29.9x)



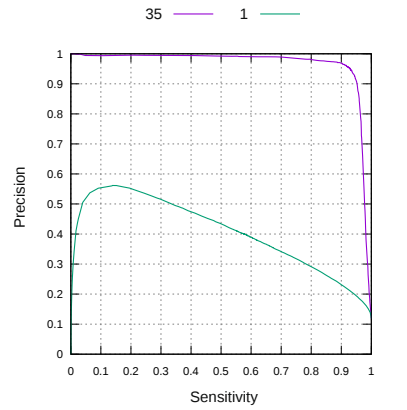
D (80% of 29.9x)



E (100% of 29.9x)



F (100% of 29.9x) b=35, b=1



G (100% of 29.9x) b=35 original or simplified parameter

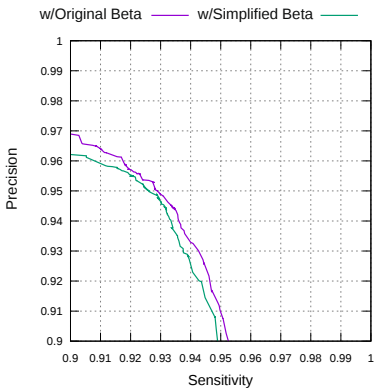


Figure 3.S3: **Sensitivity and precision of predicting unmethylated regions with $\geq b$ CpG sites for a variety of read coverages.** We continue to use b to denote a lower bound of the number of CpG sites in a region. For $b = 30, 35, 40, 45, 50$, we plot the sensitivity and precision curves when the read coverage is 20% of 29.9x (**A**), 40% of 29.9x (**B**), 60% of 29.9x (**C**), 80% of 29.9x (**D**), and 29.9x (**E**). The sensitivity and precision were evaluated on the chromosome 1 of the medaka Hd-rR genome (version 2). For better prediction with a smaller coverage, a wider window was favored. Precisely, setting b to 50 outperforms the other values for coverages, 20% and 40%, but it becomes inferior for 80% and 100%. In contrast, both sensitivity and precision increase for larger coverages, 80% and 100%, when b is set to smaller values, 35 and 40. In particular, Figure **E** shows that for coverage 100% (29.9x), setting b to 35 is better than other values of b. Figure **C** also highlights that even with a small coverage 60% of 29.9x, both sensitivity and precision are $\sim 90\%$ for $b = 45$. Figure **F** shows that the prediction is not accurate if each CpG site is treated independently (not as blocks). Figure **G** compares the performance with simplified beta (where the components for -7, +1, +3, +5$\sim$+10-th positions were truncated to 0) to that with the original full beta vector.

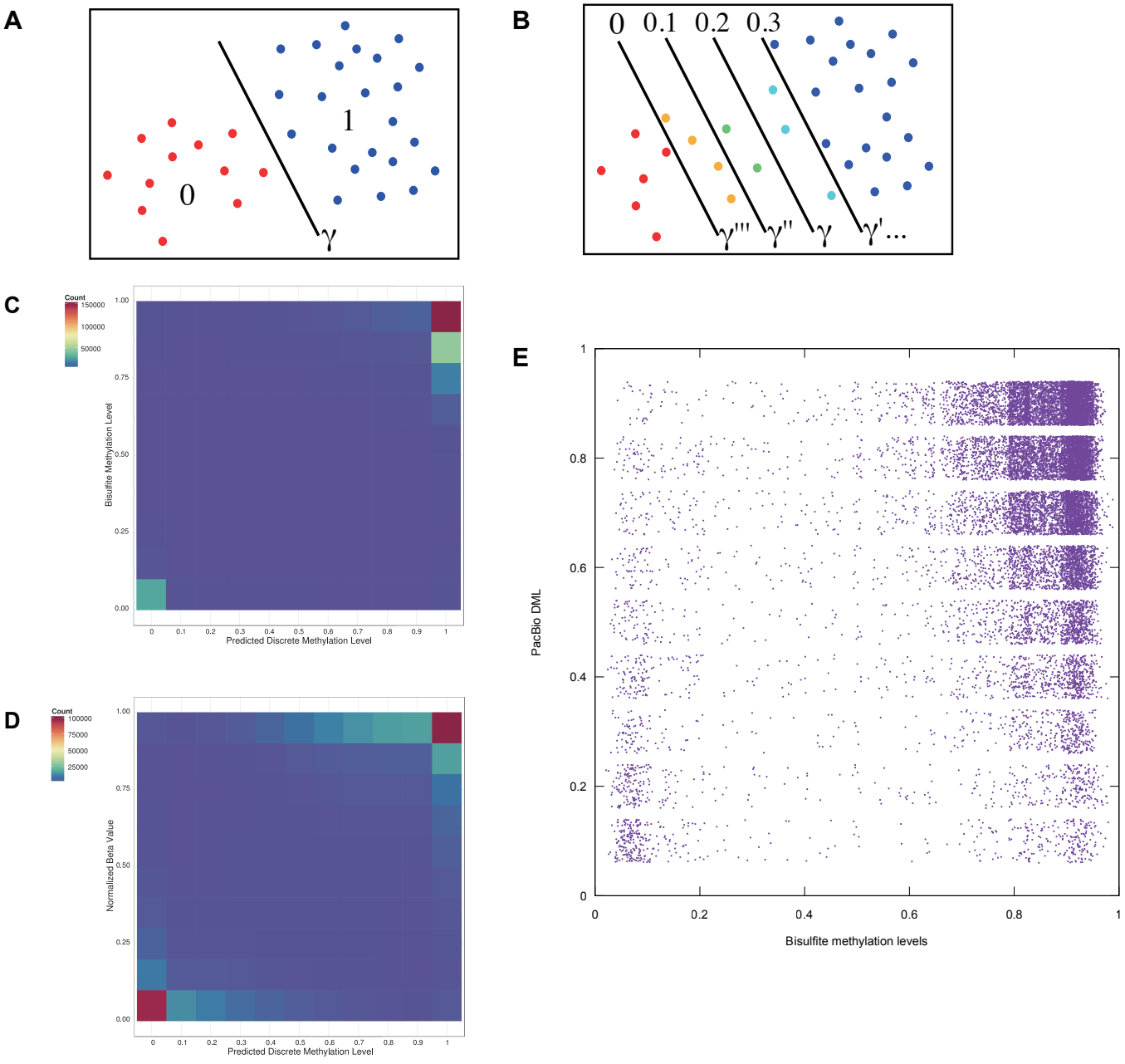
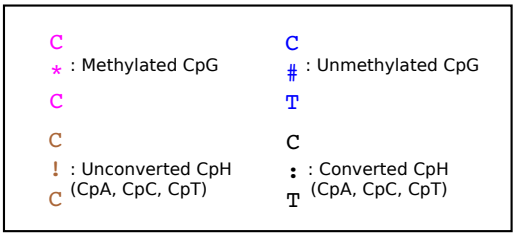

A
0
1
γ
B
0
0.1
0.2
0.3
γ'''
γ''
γ
γ'...
C
Count
150000
100000
50000
Bisulfite Methylation Level
Predicted Discrete Methylation Level
D
Count
100000
75000
50000
25000
Normalized Beta Value
Predicted Discrete Methylation Level
E
PacBio DML
Bisulfite methylation levels

Figure 3.S4: **Handling intermediate methylation states. A.** IPDR profiles of CpGs are represented as points in the feature space. Predictions are made using a decision hyperplane determined by its intercept γ, and individual CpGs are classified as methylated (blue) or unmethylated (red). **B.** Multiple predictions using a set of different intercept parameter values define the discrete methylation level (DML) on each CpG site. Specifically, after decomposing DNA into unmethylated and methylated regions for different intercept values of γ, we compute the ratio of methylated regions that cover each CpG site, and treat the ratio as the methylation level of the CpG site. **C.** DML (x-axis) and methylation level monitored by bisulfite sequencing (y-axis) in our medaka sample. The colors are based on the log of the number of CpG sites having corresponding DML value and bisulfite methylation level. These values were strongly correlated ($R = 0.884$) and the difference was within 0.25 for 92.0% of CpG sites. Most of the CpG sites were methylated because we observed CpG methylation in a genome-wide manner. **D.** DML (x-axis) correlated ($R = 0.732$) with the normalized beta values of BeadChip (y-axis) for the CpG sites in our human sample, and 75.4% of CpG sites are in concordance within 0.25. The majority of CpG sites are unmethylated, because most CpG sites on the BeadChip are designed on CpG islands. **E.** Scatterplot for methylation level monitored by bisulfite sequencing (x-axis) and DML (y-axis), on each CpG site, in the medaka sample.

Figure 3.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.



Sequence ID #5: LINE-1 (1 of 2)
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-----TCCSCCTCCT--------
```

.....

Sequence ID #7: LINE-1 (2 of 2)
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---------------------------------------------
```

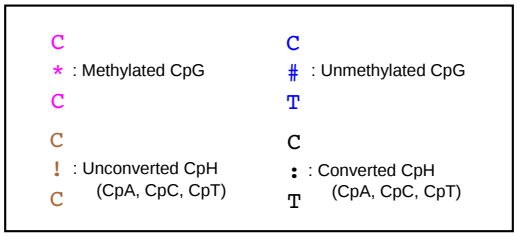
.....

Sequence ID #8: LTR (1 of 2)
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  TATTTTAGGTTTTTTTTTAGTTTTGATGGYGGGTTAGTATAAATTTTTGATKGGGATC---
```



Sequence ID #9: LTR (2 of 2)
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: LTR26C
Prediction: Hypomethylated
Region: chr19:11,848,508-11,850,380

```
Genome       661    ACCAGCGACCCCACACTCCAGCCGTCCCTGTCCACACCTCTAAACACCCCATCCCCAAAC
                           ! !:|:|! !!  !*|  !!                  :::: |::::|||:
Bisulfite      1    --------CYCTATAC-CCCCCGCYCC-----------------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: MIR2
Prediction: Hypermethylated
Region: chr19:47,905,568-47,906,031

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

Genome        61    CACTGGTTTCTGAGCCTCAGTTTTCCTAGTTACAAAGGACAGCCTCTGCCTGTGATGGGC
                    :|:|||||||:||||::|:|||||||::||||||:||||||:||::|:||::|||||||||
Bisulfite     43    TATTGGTTTTTGAGTTTTAGTTTTTTTAGTTATAAAGGATAGTTTTTGTTTGTGATGGGG

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

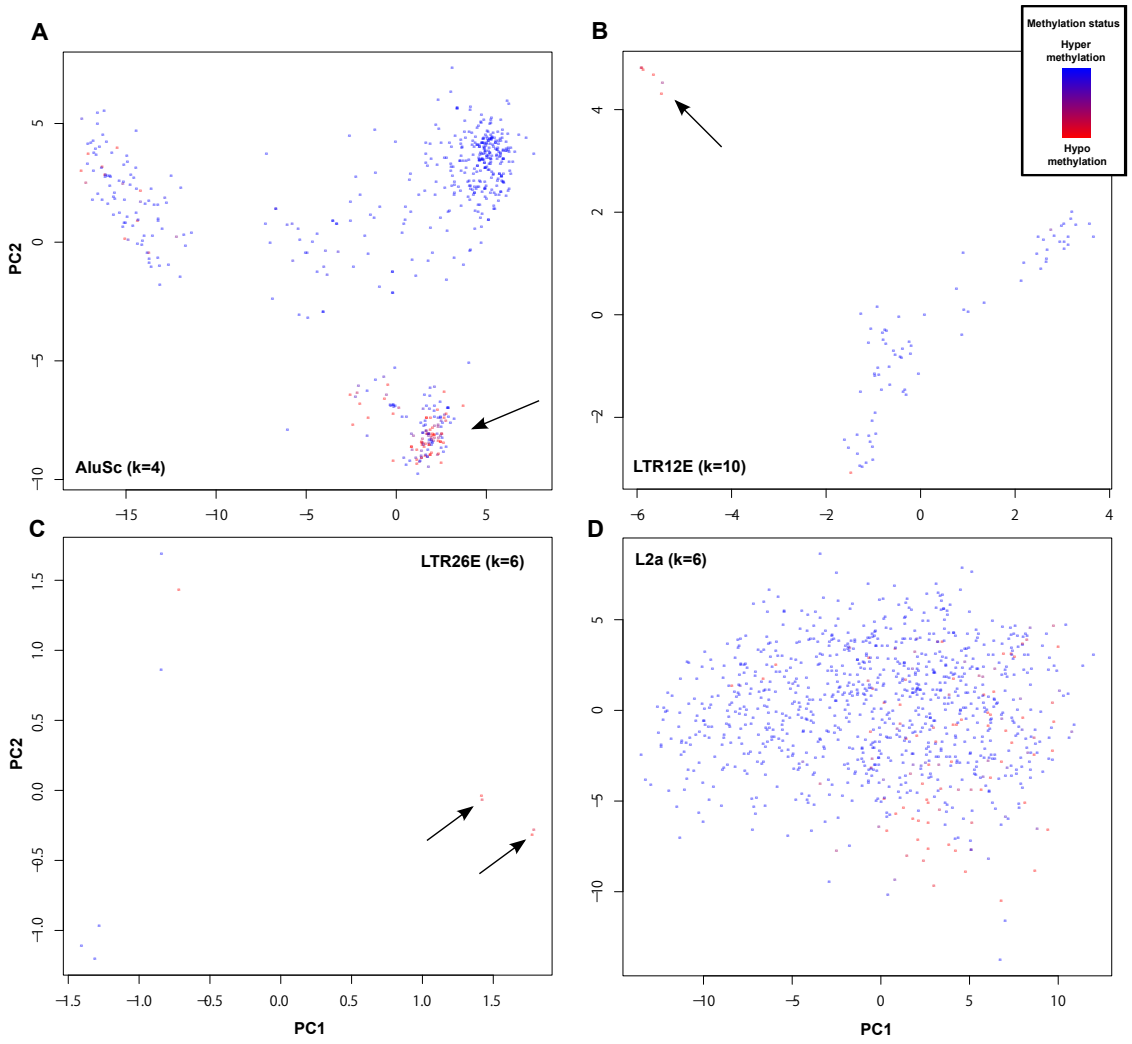


Figure 3.S6: **Kernel PCA analysis of sequence feature and methylation state.** The results of Kernel PCA analysis are shown for 4 selected classes of repetitive elements, AluSc **(A)**, LTR12E **(B)**, LTR26E **(C)**, and L2a **(D)**. We projected the repeat occurrences into the plane based on the distance metrics that we defined using the spectrum kernels and their top 2 principal components. The colors of the dots represent the methylation state of the repeat occurrences; namely, red indicates unmethylation and blue methylation. The arrows show the unmethylated occurrences that are clustered in terms of the sequence features.

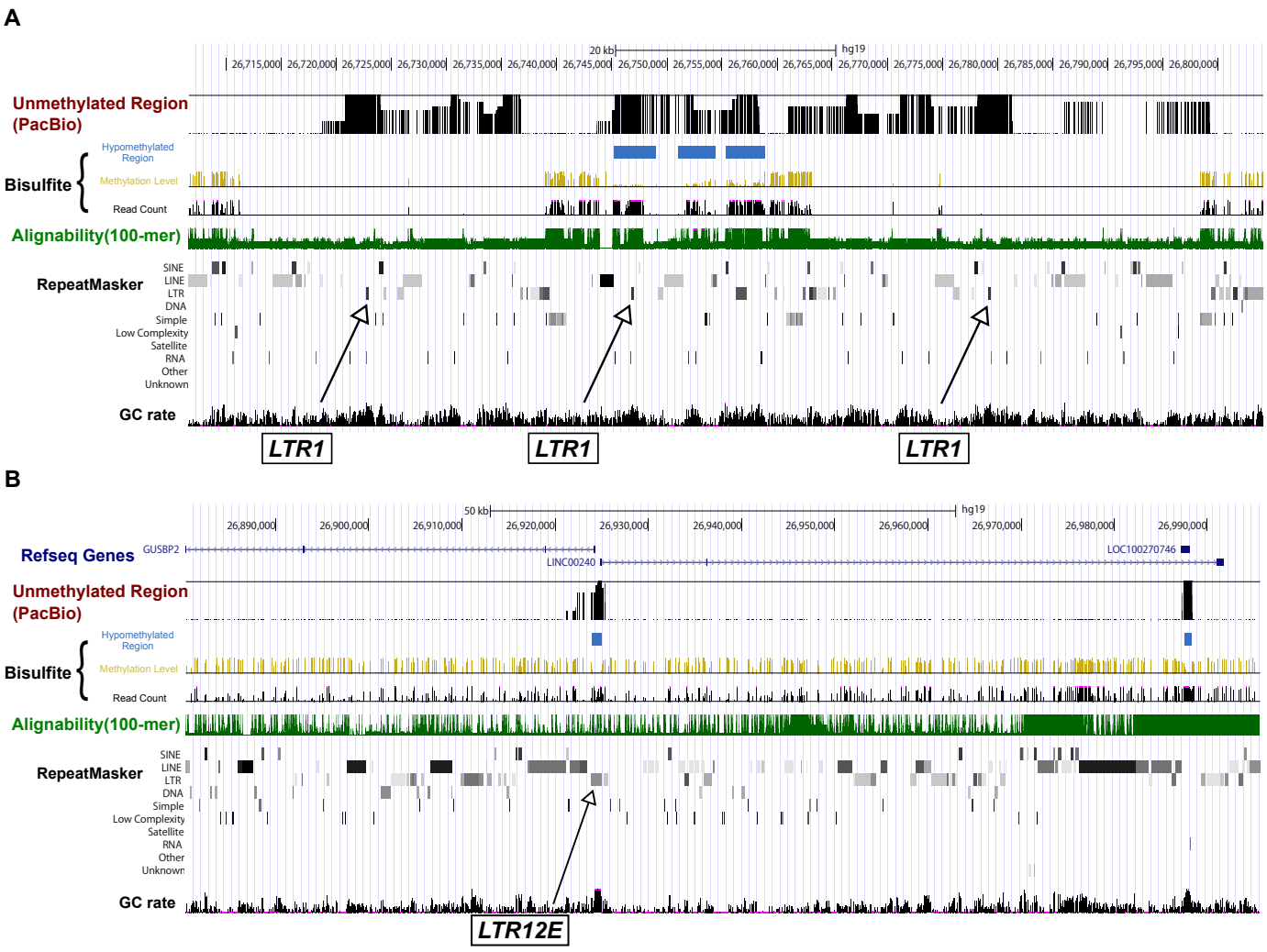


Figure 3.S7: **Examples of unmethylated repeat occurrences in a unmethylation 'hot spot'.** Three adjacent LTR1 elements were unmethylated in this region **(A)**, and a LTR12E element was located at a unmethylated bi-directional promoter region **(B)**. Both regions are on the p-arm of the chromosome 6. The arrows indicate the locations of LTR1 and LTR12E. From top to bottom, below the RefSeq gene track, black bars indicate unmethylated regions predicted from SMRT sequencing data using our method. Yellow and black bars show the methylation level and read coverage obtained from public bisulfite sequencing data, respectively, and blue boxes show unmethylated regions predicted from the bisulfite data. Green bars below indicate the alignability of short (100-bp) reads. The bottom rows shows repeat masker tracks and GC rate for every 5 bp window.

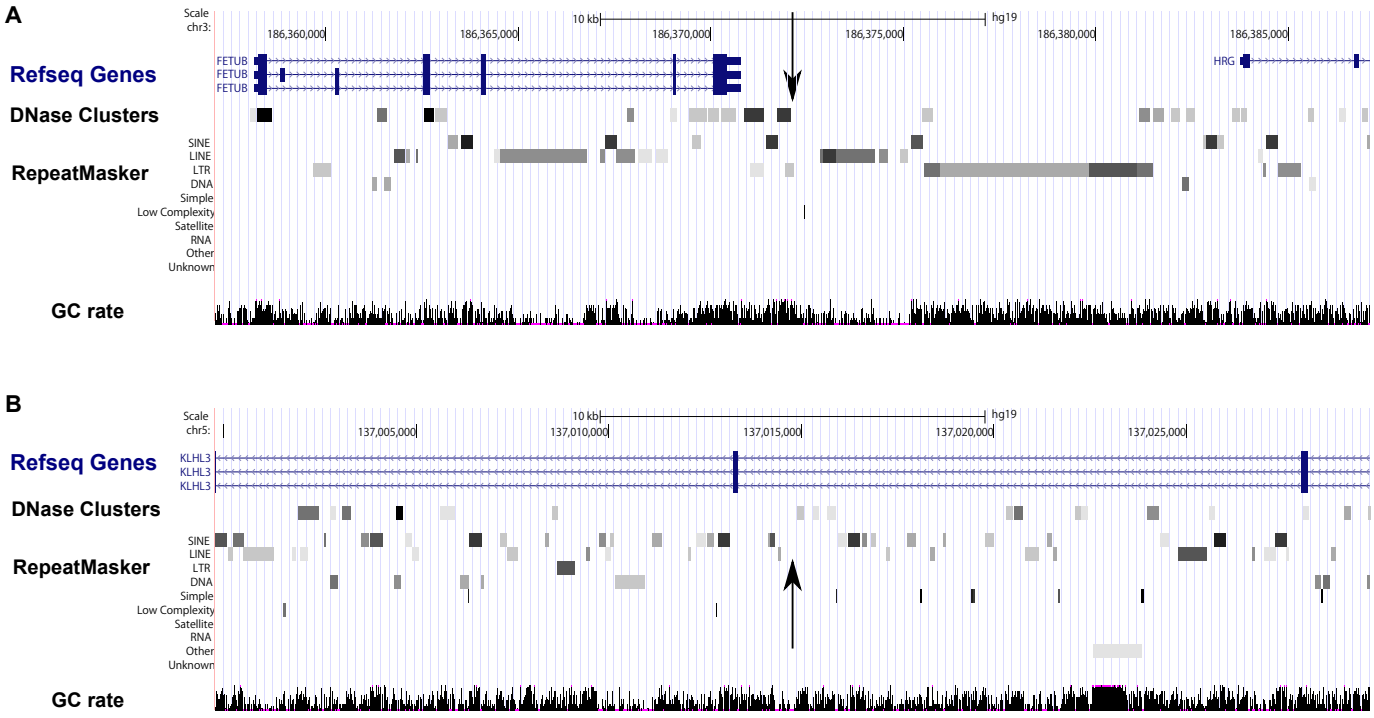


Figure 3.S8: **Two LINE insertions novel to hg19.** We identified two LINE insertions by comparing a new assembly obtained from SMRT reads and the hg19 reference genome. The vertical arrows indicate the locations of the identified novel insertions. Specifically, one is aligned at 186,372,132 in Chromosome 3 with identity 99.02%, and the other at 137,014,775 bp in Chromosome 5 with identity 98.71%. From top to bottom, the tracks shown are RefSeq genes, DNase clusters, repeat masker masked regions, and GC rate for every 5 bp window.

Chapter 4

Application of AgIn to centromeric repeats

4.1 CpG methylation in centromeric repeats

Epigenetic mechanisms are known to play a crucial role in the establishment and maintenance of centromeres[2]. The CpG methylation status in centromeres has been examined using methyl-sensitive restriction enzymes[13], fluorescence antibody labelling[127, 51], and bisulfite sequencing[126]. These studies showed that, on average, centromeric repeats were hypomethylated in core centromeres and were hypermethylated in pericentromeres in rice (Nipponbare)[127] and maize (Zea mays)[51]. Conversely, in mice (Mus musculus), the levels varied depending on tissue type, being higher for somatic cells, but intermediate and lower for sperm and oocytes, respectively[13, 126, 69]. However, these previous studies did not relate the methylation state of CpG sites with the structure of underlying centromeric repeats. We overcame this problem with our AgIn software to depict the global CpG methylation

This work has been published as a part of:
Ichikawa, Kazuki, et al. "Centromere evolution and CpG methylation during vertebrate speciation." Nature communications 8.1 (2017): 1833.

pattern over a broad range of medaka centromeric repeats at fine resolution, including the boundaries of centromeres. Indeed, we reconfirmed this in centromeric repeats where bisulfite sequencing data were available, AgIn predicted bisulfite results at an accuracy of 88.7% on unmethylated CpGs and 90.7% on methylated CpGs (Methods). In non-centromeric regions, AgIn is capable of estimating methylation states of CpG sites with a high accuracy (sensitivity and precision of ∼93.7%) from kinetic information of SMRT sequencing[114]; for example, Figure 4.1 shows typical examples of methylation states such that AgIn and bisulfite sequencing are concordant.

Adult medaka testes under reproductive laboratory conditions consist mainly of mature sperm and spermatogenic cells[58]. The centromeres obtained from these germ cells were found to be mostly hypermethylated (Fig. 4.2), which unexpectedly reflects the characteristics of somatic cells. We also reconfirmed this property by estimating the average methylation ratios of centromeres in testes and liver. Specifically, we aligned bisulfite-treated short reads from testes and liver[92] to the four representative centromeric monomers (given in Extended Data Fig. S3b in [43]). The average methylation ratio in testes was 72.9%, which was close to 65.3% in liver. However, we found that some centromeres contain hypomethylated domains. For example, Figure 4.2a and 4.2b show two syntenic centromeric repeat regions with unmethylated subregions in chromosome 2 of Hd-rR and HSOK (see [43]; the genetic marker correspondence in Figure 1b; dot plot in Extended Data Fig. 5b). Figure 4.2c shows that HSOK chromosome 4 contained two hypomethylated regions which exhibited sequence similarity to each other. Similarly, we observed hypomethylated centromeric repeats in four Hd-rR and three HSOK chromosomes (Fig. 4.2, 4.3). These examples showed diverged methylation patterns among centromeric repeats.

To understand this diversity, we analyzed underlying DNA sequences within centromeric repeats, and constructed a phylogeny tree of centromeric repeats with distinct methylation status in terms of the sequence similarity calculated with spectrum kernel (Fig. 4.2d; Methods)[62, 60]. Figure 4.1d shows the general tendency that the segregation of different

chromosomes occurred first, followed by the separation of Hd-rR and HSOK 25MYA[43]. Afterwards, hypo/hyper-methylated regions in individual chromosomes evolved independently and acquired unique sequence compositions that were not shared in common among different strains and chromosomes. This was confirmed by examining sufficiently large hypomethylated centromeric repeats in HSOK chromosomes, 2, 4, and 23 (Fig 4.2d, see also Extended Data Fig. 6 in [43]). We remark two deviations from this general tendency. Centromeric repeats in acrocentric chromosome 13 and 22 in Hd-rR are more similar in sequence than those in non-acrocentric chromosomes are, suggesting exchanges of repeats between acrocentric chromosomes. Hypomethylated repeats in HSOK chromosome 4 (orange repeats in Fig. 4.2c) are more similar to repeats in Hd-rR chromosome 1 than to repeats in chromosome 4, suggesting they might jump in HSOK chromosome 4 from another chromosome. Overall, DNA methylation in centromeres were correlated with centromere sequence phylogeny.

As a conclusion, this analysis is the first to reveal the specific pattern of hypomethylated and hypermethylated domains in centromeric repeats, which has been overlooked by traditional approaches. Analysis of underlying DNA sequence showed that the variation of non-acrocentric CpG methylation occurred after the divergence of two medaka strains (Hd-rR and HSOK), demonstrating that centromeres accumulated epigenetic diversity as well as the sequence diversity during speciation. Although centromere identity is known to be primarily defined by the epigenetic specification, in particular, by the presence of the histone H3 variant CenH3/CENP-A[73], a specific pattern of CpG methylation could play some roles in centromere evolution through meiotic centromere pairing.

4.2 Methods

4.2.1 Methylation calls using SMRT sequencing and bisulfite sequencing

Methylation call from SMRT long reads was performed as described[114]. For methylation analysis, we used SMRT reads sequenced with P6-C4 chemistry and avoided mixing reads from different polymerase and chemistry, which is not guaranteed to produce reliable result. Mapping and generation of modification summary (modifications.csv) were performed using SMRT Pipe with its default settings for the general resequencing protocol. The result was then processed by AgIn algorithm[114] to extract a set of hypomethylated regions. Specifically, we used the same parameters tuned for P6-C4 (β for P6-C4 and $\gamma = -0.55$), and set the minimum number of CpGs in each predicted region to 40. Bisulfite-treated short reads were downloaded from SRA (Accession No. SRX149585) and were processed by Bismark[55] to perform genome conversion, mapping of reads to converted genome, and production of methylation summary as bedGraph file. To align reads using bowtie2, we used the parameters: "-L 32 -N 0 ignore-quals". Each CpG site was classified as methylated if the strict majority of the mapped reads supported that it was methylated, otherwise as unmethylated. During the calculation of consistency between the results of AgIn and bisulfite sequencing, we considered CpG sites with bisulfite read coverage ranging from $2 \sim 9$, in order to exclude positions with an abnormally high coverage, which were likely to have identical copies in the genome. Among CpGs within the hypomethylated (hypermethylated, respectively) regions in centromeric repeats that we estimated from PacBio reads, 88.7 % (90.7%) were called as unmethylated (methylated) from bisulfite reads. Therefore, each technology supported the methylation calls from the other when methylation information is available from both. We also calculated the average methylation ratios in centromeres in testes and liver by using bisulfite-treated reads collected from testes and liver[92], and by aligning the reads to the four representative monomers (defined in Extended Data Figure 3b in [43]). The average methy-

lation ratio in testes was 72.9%, which was close to 65.3%, the average ratio in liver. Specifically, the respective numbers of methylated and unmethylated cytosines in liver were 20,245 and 10,827, which yielded the average 72.9% ($= 20,245/(20,245 + 10,827)$), while those in testes were 19,103 and 7,356, and the average was 65.3% ($= 19,103/(19,103 + 7,356)$).

4.2.2 Construction of a phylogenetic tree of hyper-/hypo-methylated centromeric regions

For the analysis of evolution of CpG methylation in centromeric repeats, we used all HdrR or HSOK chromosomes that had either hyper- or hypo-methylated centromeric repeat regions. Let $\mathbf{A}$ and $\mathbf{B}$ denote the normalized vector of k-mer frequencies in repeat regions, A and B, respectively such that $||\mathbf{A}||^2 = ||\mathbf{B}||^2 = 1$. To perform cluster analysis, we defined the distance between regions, A and B, by $D(A, B) = \sqrt{(||\mathbf{A} - \mathbf{B}||^2)}$. The formula is then transformed to $\sqrt{(||\mathbf{A}||^2 + ||\mathbf{B}||^2 - 2K(\mathbf{A}, \mathbf{B}))} = \sqrt{2 - 2K(A, B)}$, where $K(\mathbf{A}, \mathbf{B})$ denote the inner product of $\mathbf{A}$ and $\mathbf{B}$ that represents a sequence similarity between repeat regions, A and B, which is equivalent to the k-spectrum kernel[62], a widely used measure in sequence comparison. Based on these pairwise distance, we generated a hierarchical clustering of the regions with the UPGMA method[95]. In our analysis, we set k to 8 in Figure 4.2d because the setting could separate the segregation of chromosomes and the divergence of the medaka strains in the clustering. We calculated spectrum kernel, clustering, and final visualization were performed using R statistical environment (https://www.R-project.org/), and especially, the "kebabs" package for kernel-based analysis[88].

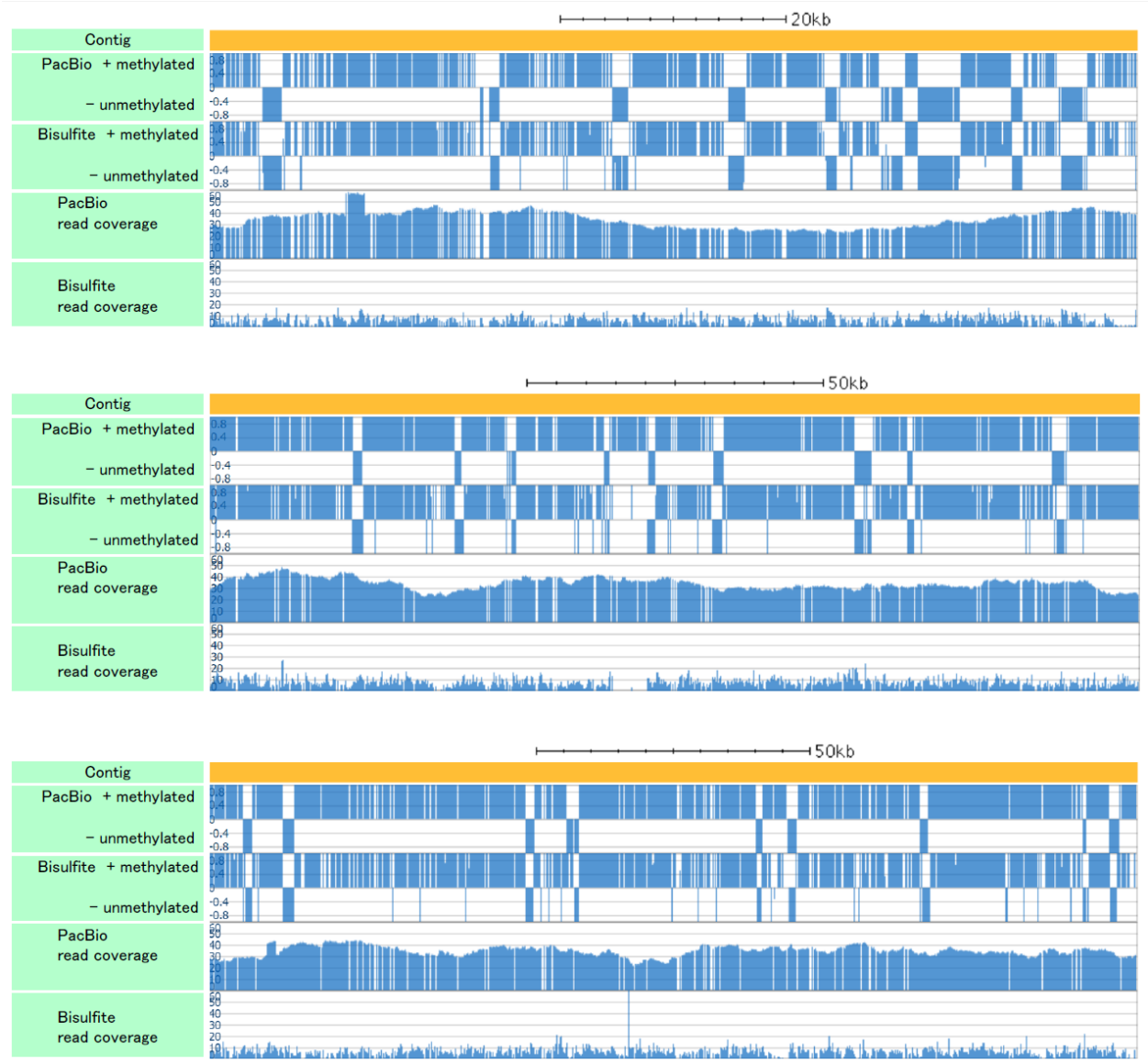


Figure 4.1: Three examples of genomic regions where CpG methylation states by PacBio sequencing and bisulfite sequencing are almost consistent. We display tracks for regional methylation prediction from PacBio reads (+, methylated; -, unmethylated), CpG-wise methylation from bisulfite reads, coverage of PacBio reads, and coverage of bisulfite reads in the Hd-rR genome. Three regions presented in the figure, Chr1:4,812K-4,894K, Chr1:5,586K-5,740K, and Chr2:8,679K-8,830K (from top to bottom) were selected from Hd-rR genome.

Figure 4.2: CpG methylation in centromeric repeats. **a.** The tracks shown here are, from the top, contigs layout as yellow bands, centromeric repeats as red bands, regional methylation prediction from PacBio reads (+, methylated; -, unmethylated), CpG-wise methylation from bisulfite reads, coverage of PacBio reads, coverage of bisulfite reads, and PacBio subreads alignments (red, forward; blue, reverse). The figure shows two centromeric repeat regions on Hd-rR chromosome 2 that were predicted as hypo-methylated from PacBio reads. Methylation calls by PacBio and bisulfite sequencing are inconsistent around the two unmethylated regions because most of bisulfite read coverages are very small (only 1) and are unreliable due to the repetitiveness of the centromeres. By contrast, PacBio reads achieved stable coverage over the repeat region. Some other chromosomes are presented in Figure 4.3. **b.** A part of HSOK chromosome 2 with an unmethylated region that is syntenic to the region in Figure 4.1a according to genetic markers. No bisulfite-treat short reads are available for the HSOK strain. The identity ratio between the representative monomers of the centromeric repeats in the Hd-rR and HSOK chromosome 2 was 85.7%.

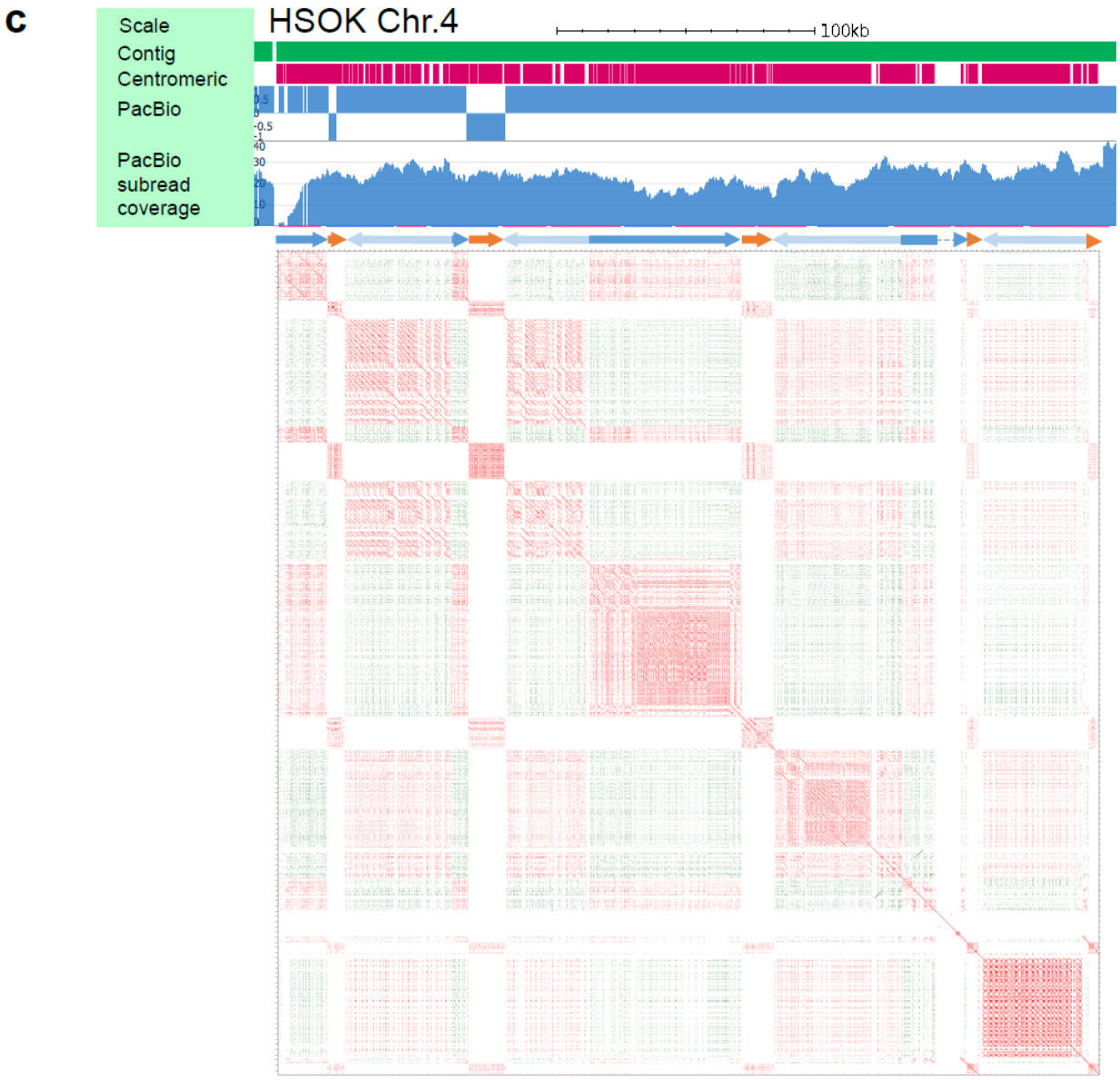


Figure 4.2: *(Cont.d)* **c.** A ~305 Kbp centromeric repeat region in HSOK chromosome 4. The lower portion shows a dot plot of the region. Forward and reverse matches are colored red and green, respectively. Each dot represents 40-mer sequence match. Blue and orange arrows displayed above the dot plot show two patterns of centromeric repeats that do not match. A light blue arrow is inverse orientation of a blue arrow. The left two regions represented by orange arrows are hypomethylated, though the other three orange arrows are not. This shows one illustrating example of gene conversion and non-allelic homologous recombination. Orange and blue repeats are respectively prevalent in acrocentric and non-acrocentric chromosomes. A possible scenario for this centromere evolution is that the orange repeat jumped into the blue repeat by gene conversion to create a basic pattern, and the pattern was duplicated multiple times by unequal crossover.

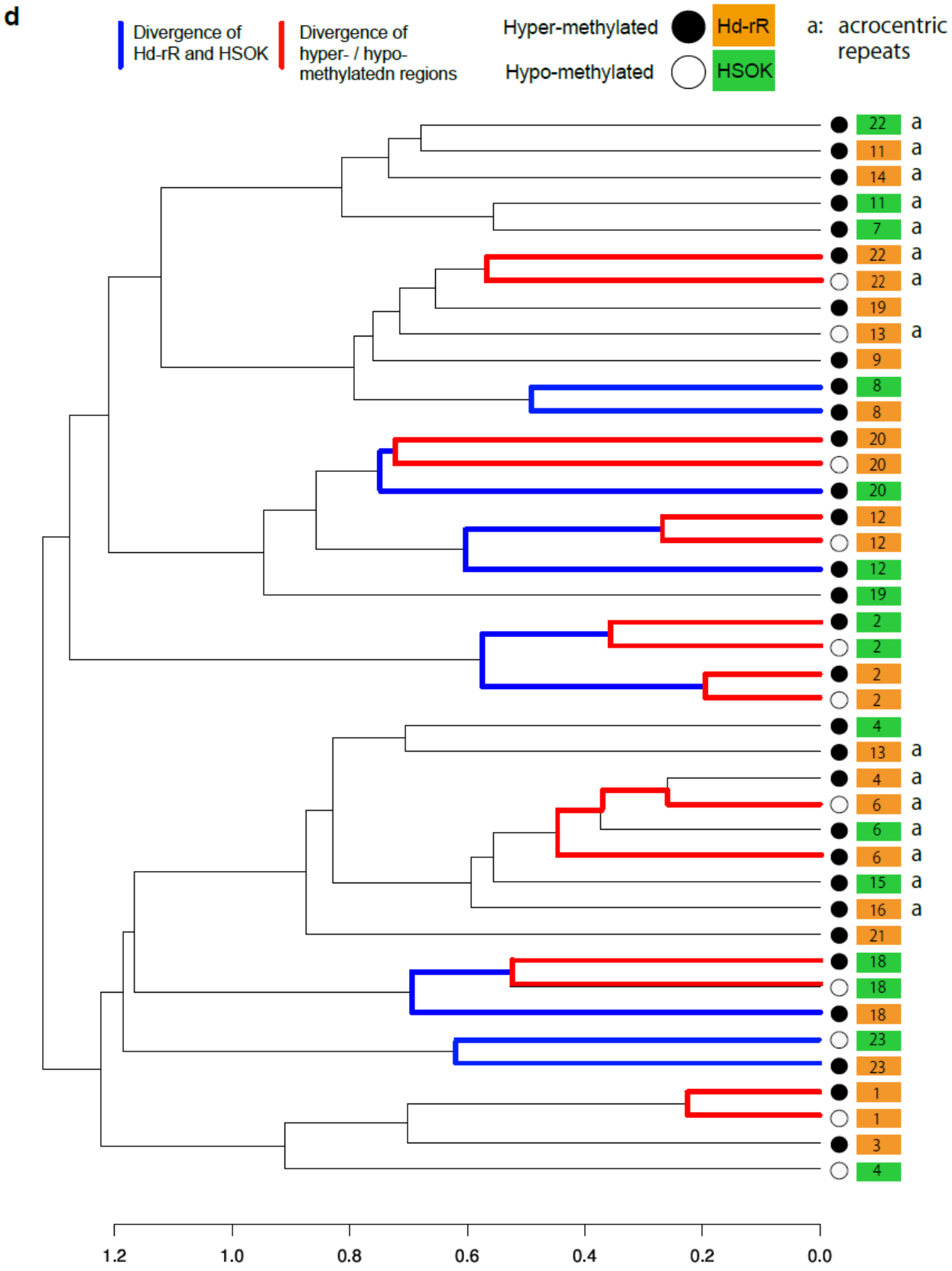

d
Divergence of Hd-rR and HSOK
Divergence of hyper- / hypo-methylatedn regions
Hyper-methylated
Hypo-methylated
Hd-rR
HSOK
a: acrocentric repeats
22 a
11 a
14 a
11 a
7 a
22 a
22 a
19
13 a
9
8
8
20
20
20
12
12
12
19
2
2
2
2
4
13 a
4 a
6 a
6 a
6 a
15 a
16 a
21
18
18
18
23
23
1
1
3
4
1.2
1.0
0.8
0.6
0.4
0.2
0.0

Figure 4.2: *(Cont.d)* **d.** We considered all centromeric repeat regions in Hd-rR and HSOK chromosomes. We clustered hyper- and hypo-methylated regions with at least 40 CpGs that we could reliably estimate from SMRT sequencing information (Method). We calculated the sequence similarities between all pairs of hyper/hypo-methylated regions using spectrum kernel, a robust method of characterizing sequence compositions of k-mers (strings of length k) in individual regions. From the similarities, we obtained a hierarchical clustering of hyper/hypo-methylated regions. The respective orange and green boxes represent the Hd-rR and HSOK strains. The black and white circles illustrate hyper- and hypo-methylated regions. Numbers indicate chromosome numbers. Black, blue, and red lines in the dendrogram respectively illustrated the timing of chromosome segregation, divergence of two strains (Hd-rR and HSOK), and divergence of hyper/hypo-methylated regions in an identical chromosome of the same strain. Seven pairs of hypomethylated and hypermethylated regions (from top to bottom: Hd-rR Chr. 22, 20, 12 HSOK Chr. 2, Hd-rR Chr. 2, HSOK Chr. 18, Hd-rR Chr. 1) are most similar to each other except for three exceptional cases (Hd-rR Chr. 13, Hd-rR Chr. 6, HSOK Chr. 4). The rightmost column labels acrocentric repeats, including Hd-rR Chr. 13 and 6, with a. Hd-rR Chr. 13 that might be exchanged from other acrocentric chromosomes. The hypomethylated Hd-rR Chr. 6 and the hypermethylated Hd-rR Chr. 4 were reciprocally most close, and they might be exchanged in Hd-rR. Orange repeats in Figure c might jump into HSOK Chr. 4.

Hd-rR chr.22

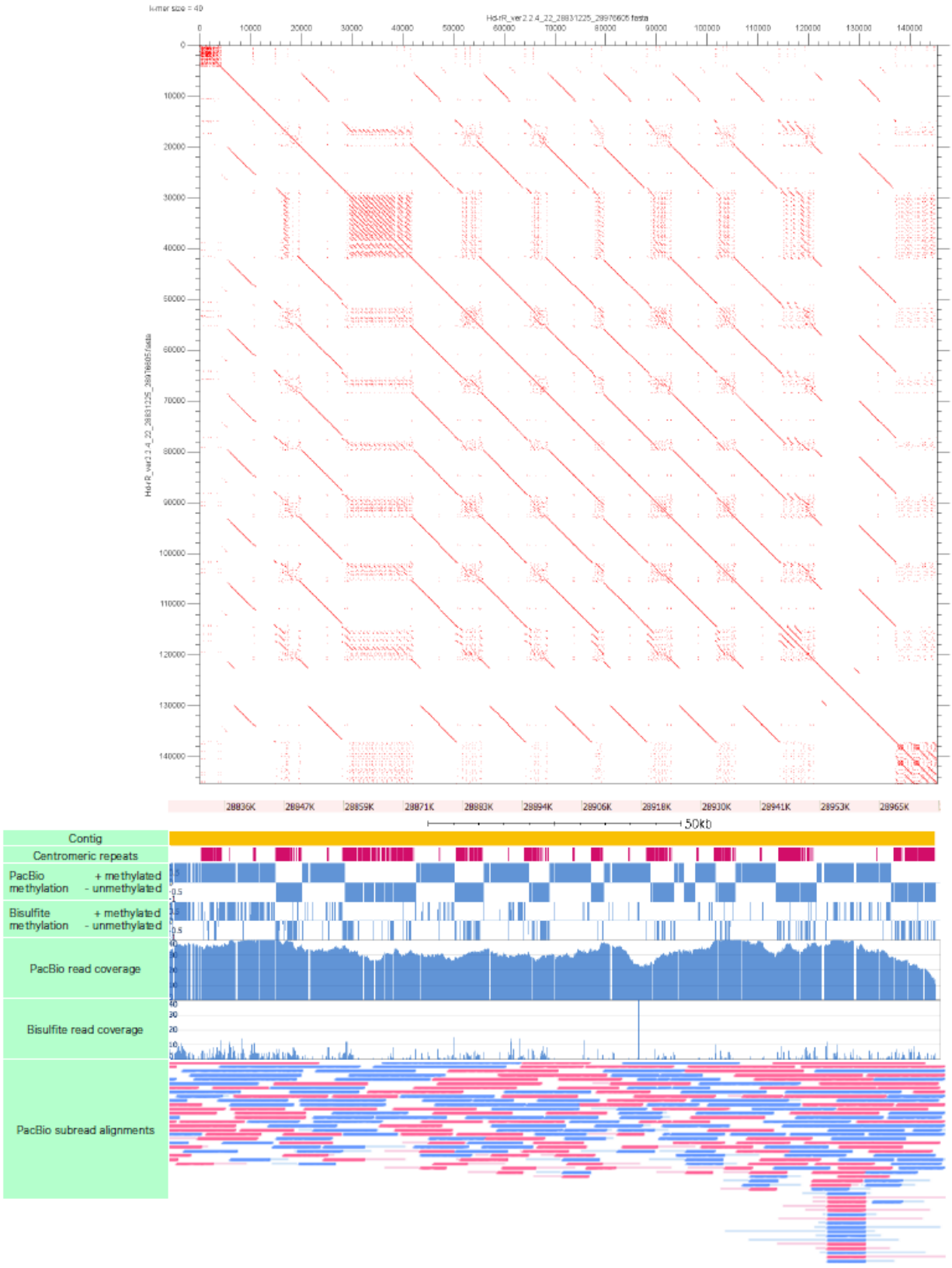


Figure 4.3: Validation of centromeric repeat regions and their CpG methylation states. Hd-rR Chromosome 22.

HSOK chr.23

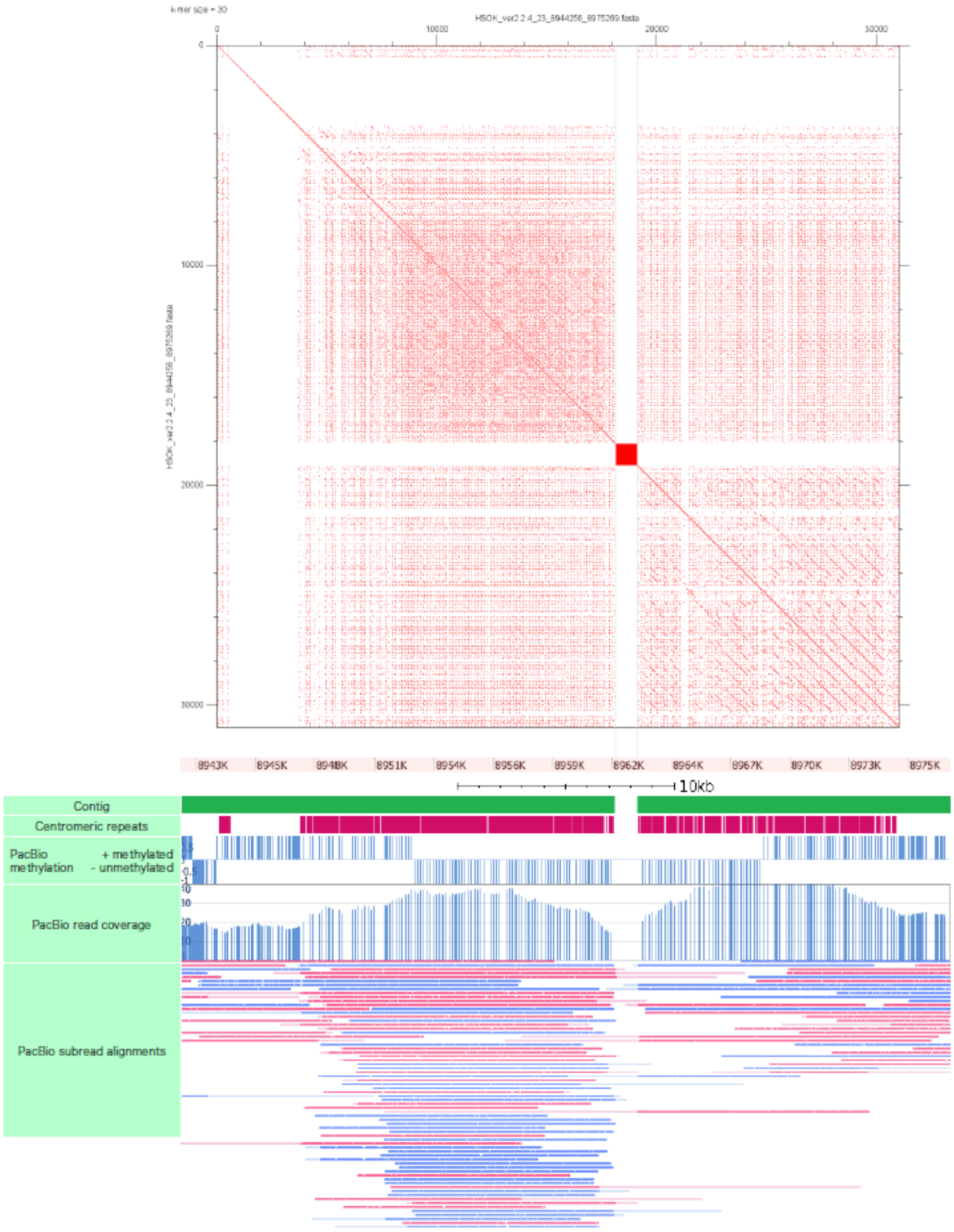


Figure 4.3: *(Cont.d)* Validation of centromeric repeat regions and their CpG methylation states. HNI Chromosome 23

Figure 4.3: We show two regions with centromeric repeats on chromosomes of the Hd-rR and HSOK genomes. (Top) Dot plot of centromeric regions, where each dot represents 30- or 40-mer sequence match (indicated at the top left in each figure). Red and green dots indicate forward and reverse matches, respectively. Red blocks indicate contigs gaps. We can observe multiple patterns of higher order repeats that are represented by lines parallel to the diagonal, uncovering broad divergence in higher order repeats. (Bottom) Snapshots of genome browser in centromeric regions. The yellow bars represent Hd-rR contigs, green bars HSOK contigs, and red bars centromeric repeats. Below the track for centromeric repeats, we display tracks for regional methylation prediction from PacBio reads (+, methylated; -, unmethylated), CpG-wise methylation from bisulfite reads, coverage of PacBio reads, coverage of bisulfite reads, and PacBio subreads alignments (red, forward; blue, reverse) by BLASR. As bisulfite data are unavailable for the HSOK genome, we generated two tracks for the methylation status calculated from PacBio subreads and for PacBio subread coverage at each CpG site. For information on other chromosomes, reader may refer to the original, Extended Data Figure. 5 in [43].

Chapter 5

Personal diploid methylomes and transcriptomes via phased heterozygous variants and single-molecule real-time sequencing

Abstract

Personal diploid methylomes, methylome pairs of individual homologous chromosomes, directly reflect allele-specific methylation (ASM) and control allele-specific expression (ASE) of genes, but are challenging to analyze. Phased heterozygous variants (PHVs) offer a unique opportunity to investigate personal diploid methylomes, if sequencing reads linking CpG methylation directly to neighboring PHVs are available. Analyzing two personal genomes (AK1 and HG002) with CpG-methylation sensitive single-molecule real-time (SMRT) reads, we found longer reads were more essential because only 11.3% $\sim$ 12.3% of CpG sites lay within 100 bp from their nearest PHVs whereas 72.2 % $\sim$ 81.3% within 8,000 bp. Bisulfite treatment is widely-used to observe CpG methylation but is not ideal because it breaks DNA into short fragments. We here propose a novel integration of PHVs and SMRT reads.

To study correlation of ASM with ASE on genes, we attempted to to build *personal diploid transcriptomes* by assigning RNA-seq (long and short) reads with informative PHVs to their alleles. Indeed, CpG islands showing ASM were often associated with known imprinted genes and resided more in the transcribed or repressed regions, which reflects ASE. We revealed complex ASM events controlling the differential ASE between alternative isoforms within genes (e.g., *ZNF331*) in the AK1 genome. We also note the scarcity of exons with PHVs (10.9 % of all exons) often hinders associating RNA-seq reads with their alleles, but correlation of ASM with ASE demonstrates the potential utility of ASM as complements for ASE. These findings highlight the need for long, CpG-methylation-sensitive SMRT reads in epigenetics study to construct comprehensive personal diploid methylomes and transcriptomes.

5.1 Introduction

DNA methylation plays important regulatory roles in a wide range of biological processes including differentiation, transposon repression, and cancer progression [45, 108, 100]. Several technological advances now enable us to study genome-wide DNA methylation [25], down to the resolution of a single base-pair [71]. Furthermore, single-cell biology can now be applied to epigenetics, allowing methylation to be measured at the single-cell level. This creates a unique research frontier [107, 39, 16]. Despite such advances in methodology, detection of allele-specific methylation (ASM) in which only one of the two homologous chromosomes is methylated in a specific region, remains challenging.

Conceptually, there are at least four known situations in which methylation may be intermediate (Supplementary Figure 5.1). First, mono-allelic methylation may be coupled to genomic imprinting, where genes in one of the homologous chromosomes escape methylation. Differences in methylation status are determined in a parent-of-origin-dependent manner, and are established during either gametogenesis or development. This situation might be the most extensively investigated form of ASM [64, 117, 74]. Second, methylation can be controlled by local *cis* variation. Thus, heterozygosity may trigger ASM. Unlike the case with imprinting, the methylation allele can be inherited from both parents. This general type of mono-allelic methylation has received much recent attention[125, 49, 98, 106, 97, 36, 63, 115]. Third, one of the two homologous loci may be methylated randomly, with no association evident with either the parent-of-origin or a *cis* variant. A well-known example of the above is X chromosome inactivation [93]. To determine whether mono-allelic methylation events are associated with genomic imprinting, *cis* effects, the presence of variants, or randomness, it is important to have information on the methylomes inherited from the parents. Finally, intermediate-type methylation may result from cell-to-cell variability within a sample population, even though the methylation status of all cells is in broad agreement [39]. This can be explored only via single-cell methylation analysis.

A number of methods are available to detect mono-allelic methylation, *i.e.*, ASM events.

Probabilistic models have been developed that allow ASM to be estimated from bisulfite sequencing data [31, 90, 122]. These models take advantage of the fact that ASM yields 50:50 mixtures of reads suggestive of methylation, or not, over specific regions. However, the converse does not hold in general. Such bi-modal observations can also be caused by cellular heterogeneity. Thus, it is not straightforward to conclude that an intermediate level of methylation reflects allelic differences.

In order to directly make distinction between two homologous chomosomes, several studies explicitly utilized heterozygous variants, as such variants define the differences between two homologous chromosomes. One approach involved a two-step experiment [49, 97]. In the first step, DNA fragments containing methylated alleles were enriched using a methylation-sensitive restriction enzyme (MSRE) or via methylated DNA immunoprecipitation (MeDIP). In the second step, sequence variants in the library were quantified using an SNP array or DNA sequencing. Variants associated with methylation might thus be over-represented when compared to an appropriate negative control. This approach is relatively cost-effective and comprehensive, but resolution is limited by the distribution of the relevant restriction enzyme cleavage sites, which are far sparser than CpG sites.

The other type of approach exploits heterozygous variants within bisulfite-treated sequencing reads [106, 36]. To assign a read to one of two alleles, the read must contain at least one informative (i.e., heterozygous) variant, in addition to the CpG site. However, we will show below that this condition is rarely satisfied when short bisulfite-treated reads are used; bisulfite breaks DNA into fragments of lengths that are typically less than 500 bp [79]; the reads are maximally 1,500 bp [128]. For example, Kuleshov et al. constructed a haplotyped genome using a read cloud containing long-range information and performed short-read bisulfite sequencing to survey ASM in a genome-wide manner [56].

As we will see later, it is difficult to observe comprehensive ASM for a given individual genome due to the lack of enough heterozygous variants available to short reads around the CpG sites. Consequently, current genome-wide overview of allele-specific methylation is com-

posed of observations for many individuals, as there are much more available heterozygous variants when a population is considered.

In the present work, we claim one needs long reads to directly observe genome-wide ASM on most CpG sites in individual genome. We developed an alternative method allowing us to analyze regions of intermediate methylation status; we used kinetic information obtained by PacBio sequencing to call regional CpG methylations, as reported previously [114]. We term the allele-specific methylome data obtained using phased long read as personal diploid methylome, in the sense it is comprehensive genome-wide ASM data obtained from single individual and it is based on the personal haplotype information.

Previous studies have revealed the prevalence of allele-specific expression (ASE) in human and demonstrated the link between ASM and ASE [101, 24]. We incorporated transcriptome data collected with long reads (as in "Iso-seq" studies using PacBio long reads[4]) and short reads to confirm that some of the ASM statuses we detected are consistent with their transcriptional activity, including their ASE statuses, *i.e.*, personal diploid transcriptomes.

5.2 Results

5.2.1 Generating the diploid methylomes and transcriptomes for AK1 and HG002 dataset

To demonstrate how our method to call ASM works, we used two independent datasets: AK1 (Asian Korean) [104] and HG002 (Ashkenazim Trio son) [132]. For both dataset, we followed the procedure outlined in Figure 5.1a to call methylation status independently for the two homologous chromosomes (*i.e.*, two haploids).

Intuitively, the goal of the procedure is to obtain two set of reads, each of which corresponds to one haploid. To achieve this, we focused on the phased heterozygous single nucleotide variants (hetSNVs) as they are the ultimate clues to distinguish two haploids. For the AK1 dataset, we identified the sites and their phasing of SNVs by aligning the scaf-

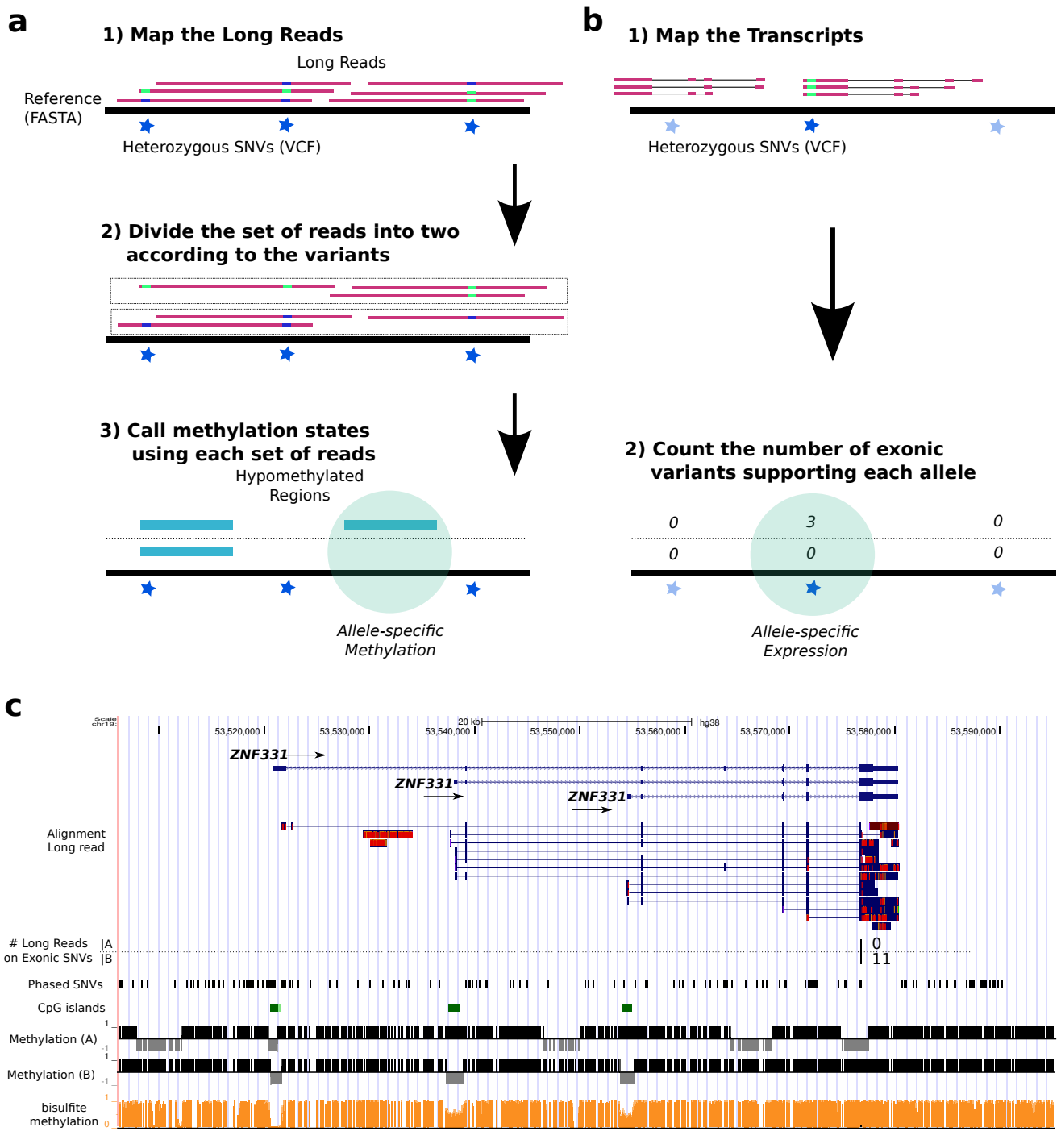

a
1) Map the Long Reads
Long Reads
Reference
(FASTA)
Heterozygous SNVs (VCF)
2) Divide the set of reads into two
according to the variants
3) Call methylation states
using each set of reads
Hypomethylated
Regions
Allele-specific
Methylation
b
1) Map the Transcripts
Heterozygous SNVs (VCF)
2) Count the number of exonic
variants supporting each allele
0
3
0
0
0
0
Allele-specific
Expression
c
Scale
chr13
20 kb
hg38
53,520,000
53,530,000
53,540,000
53,550,000
53,560,000
53,570,000
53,580,000
53,590,000
ZNF331
ZNF331
ZNF331
Alignment
Long read
Long Reads
on Exonic SNVs
A
B
0
11
Phased SNVs
CpG islands
Methylation (A)
Methylation (B)
bisulfite
methylation

Figure 5.1: **a.** Outline of the detection of allele-specific methylation (ASM). After mapping to the reference sequence, PacBio reads were assigned to one of two homologous chromosomes, depending on phased heterozygous variants (PHVs) evident on the reads. The methylation status of each diploid methylome was predicted using kinetics data from the PacBio sequencing process. **b.** Outline of the detection of allele-specific expression (ASE). Only exonic PHVs (the star with strong blue) can be utilized for distinguishing two homologous chromosomes. **c.** Example of a region exhibiting allele-specific methylation in diploid methylomes and transcriptomes. On the right two CpG islands (CGIs) shown in the middle, one allele (labeled A) is methylated and the other (B) is unmethylated. Each of the CGIs overlaps with promoter regions of distinct isoforms of known imprinted gene *ZNF331*. Bisulfite sequencing data in the bottom track exhibited intermediate-level methylation for the two CGIs showing ASM. From top to bottom, the panel shows following features: structures of genes, alignments of long RNA-seq (Iso-seq) reads, RNA-seq read counts of two alleles, which indicates ASE, sites of PHVs available in this personal genome (black marks), which were used to determine the allelic origins of the sequencing reads, annotated CpG islands (green rectangles), methylation levels of the CpG sites of two alleles that were predicted using SMRT reads (respective black and gray bars towards positive and negative indicate methylation and unmethylation), and publically available data on methylation levels via bisulfite sequencing (orange bars).

folds representing each haplotype into hg38 reference. For the HG002 dataset, the phased variants calculated using the linked-read technology were available [131]. We mapped the reads to the hg38 reference genome. Then, if the read contained matches to hetSNVs, we counted the number of hetSNVs supporting each allele. The assignment of allele for each read was determined by majority voting. Reads were excluded from further analysis if they contained none of the hetSNVs or the voting was tied. Then, we could study the SMRT read sets of both alleles separately and called the regional methylation status of genome-wide CpG sites using the kinetic information inherent in reads, as described previously [114]. The resulted set of methylation calls is *a personal diploid methylome*, as it comprises of two methylomes each representing one haploid.

For the AK1 dataset, RNA-seq data collected with long reads and short reads were available, thus we were able to use them to support that differential methylation between two alleles we detected was associated with differential transcription activity. The RNA-seq data was mapped to the genome and then the number of reads supporting transcription from each allele was recorded (Figure 5.1b), building a pair of transcriptomes in two homologous chromosomes, which we call *personal diploid transcriptomes*.

In total, 24,181,074 reads (210,782 Mb) from the AK1 dataset was aligned to hg38. The average length of the mapped reads was 8,717 bp. Of the reads, 13,857,752 (139,467 Mb) contained at least one match to a hetSNV, and a haplotype label was assigned. Notably, although these reads constituted 57.3% of all mapped reads in terms of read number, they contained 66.2% of the mapped bases. More bases were retained because longer reads were more likely to contain matches to the hetSNVs. In other words, reads with no matches were likely to be shorter, therefore affecting a relatively small number of bases. Consequently, the average length of reads assigned to a haplotype was 10,064 bp, 115% that of the average length of the original data set.

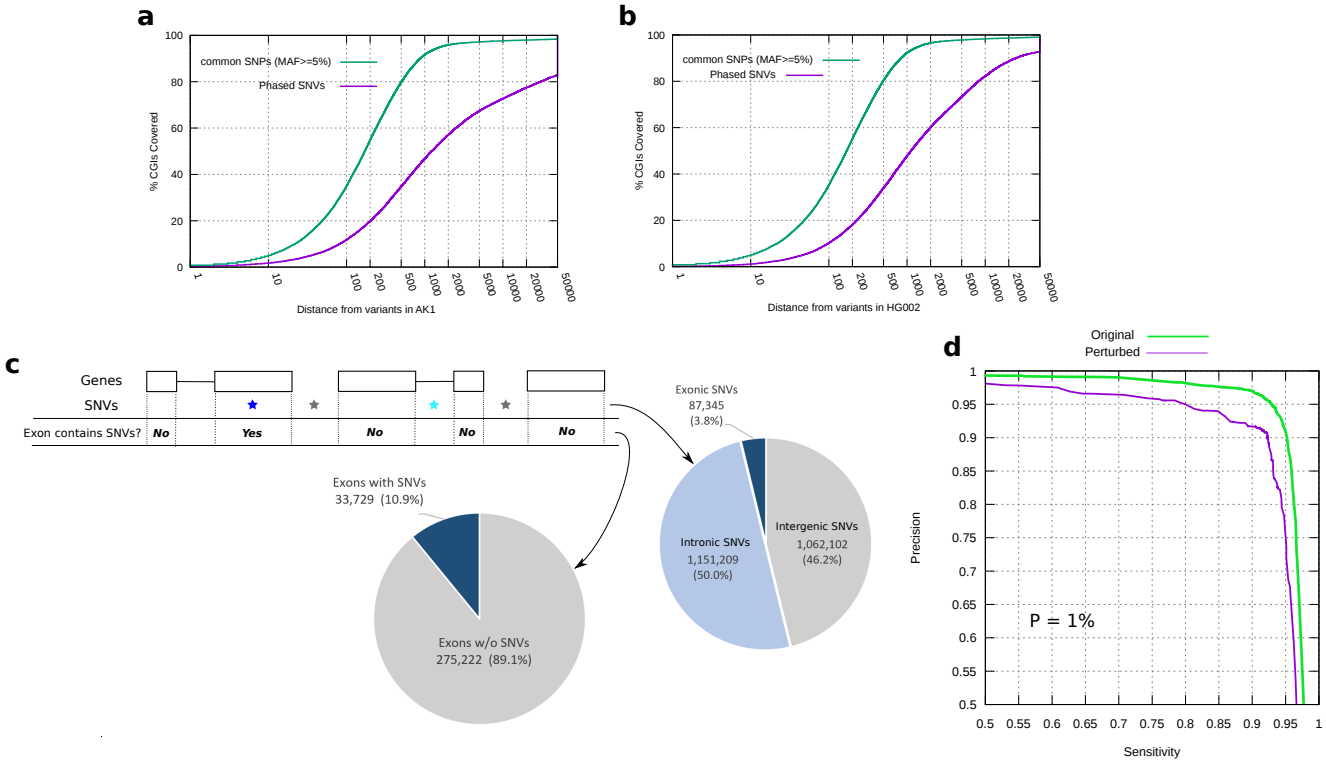
We obtained the similar statistics for the HG002 dataset as well. Starting from 23,031,407 reads (168,051 Mb) aligned to hg38, 13,676,974 reads (111,543 Mb) were assigned a haplotype

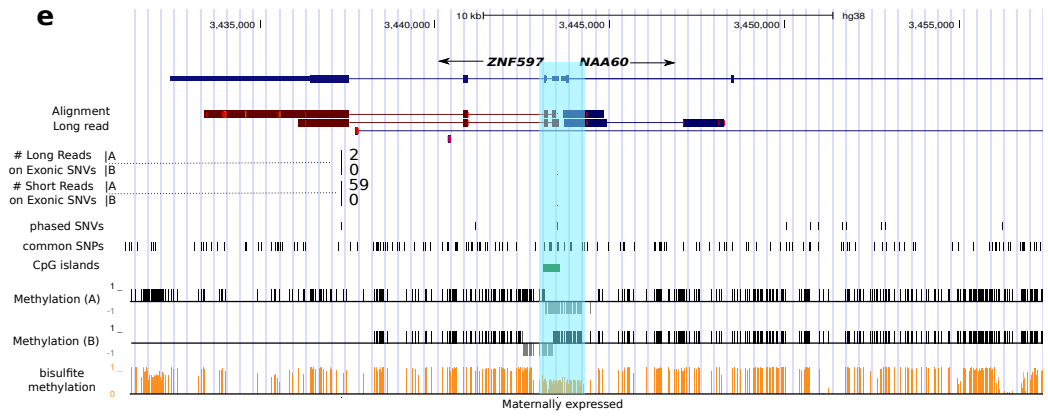
label. Thus we retained 59.4% of all mapped reads, and 66.4% of the mapped bases. Again, the average read length of reads assigned to a haplotype was 8,156 bp, 112% that of the average length of the original data set, 7,296 bp.

Figure 5.1c shows an example of an ASM detected using our method in the genomic region encoding a imprinted gene, *ZNF331*. There are 3 CpG islands (CGIs) in this region, and each CGI is corresponding to a promoter region of distinct isoforms of the *ZNF331* gene. While the CGI to the left in the panel was unmethylated for both alleles, the other two CGIs (in the middle and to the right) showed ASM, and our methylation calls informed us that the same allele (*i.e.*, allele B) was unmethylated. Of note, a publicly available methylation level annotation from a different sample by bisulfite sequencing suggested that these two CGIs are in intermediate methylation status. The alignment of long reads transcripts and the read counts at the exonic phased SNV supported that the corresponding two isoforms were transcribed exclusively from allele B. Thus it suggested the detected ASM was correlated with the transcriptional activity of the genes. We will cover other examples in the later sections to generalize this observation.

5.2.2 Distribution of phased heterozygous SNVs in two personal genomes

We next studied how the possibility of assigning the reads and CpGs into alleles would be limited by the distribution of phased heterozygous variants in personal genomes. Specifically, given a read of length l bp containing a CpG site, the allelic origin of the read can possibly be determined only when the nearest hetSNV is located within l bp from that site and both the CpG site and the hetSNV are covered by the same single read. Therefore, to assess the utility of long reads for determining the allelic origins of CpGs or CpG islands (CGIs), it is useful to calculate the proportions of CpGs or CGIs residing within specific distances from the nearest hetSNV (Figure 5.2a,b). These figures served as theoretical upper bounds for the proportions of CpGs or CGIs, for which the allelic methylation status could be determined


a
common SNPs (MAF>=5%)
Phased SNVs
% CGIs Covered
Distance from variants in AK1
b
common SNPs (MAF>=5%)
Phased SNVs
% CGIs Covered
Distance from variants in HG002
c
Genes
SNVs
Exon contains SNVs?
No
Yes
No
No
No
Exonic SNVs
87,345
(3.8%)
Intergenic SNVs
1,062,102
(46.2%)
Intronic SNVs
1,151,209
(50.0%)
Exons with SNVs
33,729 (10.9%)
Exons w/o SNVs
275,222 (89.1%)
d
Original
Perturbed
Precision
Sensitivity
P = 1%


e
ZNF597
NAA60
Alignment
Long read
Long Reads on Exonic SNVs |A
|B
Short Reads on Exonic SNVs |A
|B
2
0
59
0
phased SNVs
common SNPs
CpG islands
Methylation (A)
Methylation (B)
bisulfite methylation
Maternally expressed

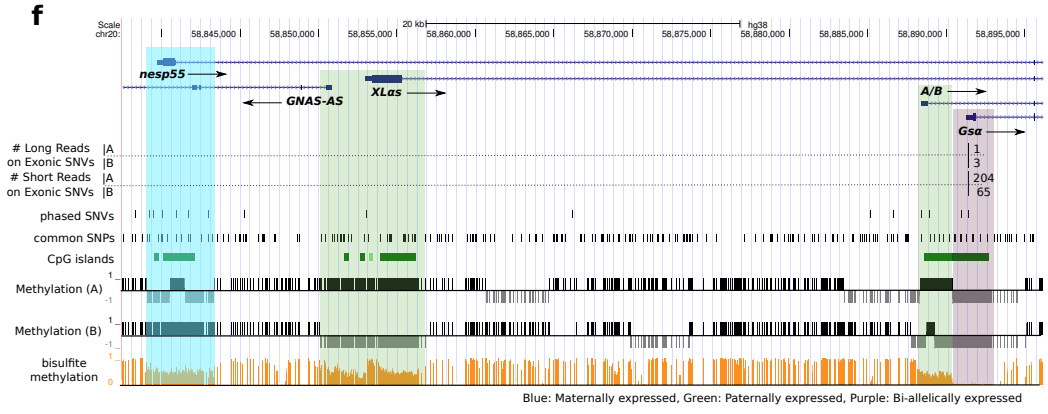

f
nesp55
GNAS-AS
XLas
A/B
Gsα
Long Reads on Exonic SNVs |A
|B
Short Reads on Exonic SNVs |A
|B
1
3
204
65
phased SNVs
common SNPs
CpG islands
Methylation (A)
Methylation (B)
bisulfite methylation
Blue: Maternally expressed, Green: Paternally expressed, Purple: Bi-allelically expressed

Figure 5.2: **a,b.** The proportions of CpG islands located within a distance in the x-axis from the nearest genomic features, common SNPs (green) and hetSNVs (purple), in each genome of (a) AK1 and (b) HG002. Common SNPs and hetSNVs distributed differently in both of the two personal genomes, and hetSNVs were essential in determining the proportions of CpG islands. **c.** Distribution of phased heterozygous variants (PHVs) with respect to exons. The left pie shows the proportion of exons containing PHVs, for which ASE status can be assessed directly. The right pie shows the ratios of PHVs in exonic (blue), intronic (pale blue), or intergenic (gray) regions, which classifies PHVs into three categories. **d.** Prediction performance (sensitivity and precision) of the method for perturbed IPD ratio (purple line). For comparison, typical performance statistics for original IPD are shown (green line). "$P = 1\%$" indicates that IPD was perturbed to similate 1% read assignment error as described in the text. **e.** Example showing personal diploid methylomes and transcriptomes in AK1 genome. The CGI at the bi-directional promoter region (area shaded in blue) for *ZNF597* and *NAA60* showed ASM. The RNA-seq reads (both long and short) supported that transcription was only from allele A, which is the allele unmethylated in the region. **f.** Personal diploid methylomes around *GNAS* complex locus in AK1 genome. The four regions are colored to show their known transcriptional pattern; maternally expressed (blue), paternally expressed (green), or expressed from both alleles (purple). Correspondingly, these regions shaded with different colors exhibited distinct methylation patterns to each others. Of note, allele-specific methylated regions were of intermediate methylation level according to bisulfite sequencing in the bottom. RNA-seq reads suggested the expression of the Gsα from both alleles.

using sequencing reads of given length of l. Of note, these depend on the distribution of phased hetSNVs available in a given sample, thus it can be quite different in individual samples.

One may attempt to approximate an upper bound using a set of common SNPs, typically the set of SNPs for which the minor allele frequencies (MAFs) are $\geq$5% in at least 1 of 26 major populations of dbSNPs[105]. As we observed that 97.3% of CpG sites lay within 500 bp from common SNPs (Supplemental Figure 5.S2), it would be possible for relatively short reads to determine the allelic origin of themselves, if these common SNPs are present in heterozygous in an individual genome. However, the conditions posed by the real distribution of phased hetSNVs was much severe.

Indeed in the AK1 dataset, at most 11.3%, 33.7% and 46.1% of CpGs could be assessed using reads of 100, 500 and 1,000 bp, respectively, whereas 72.2% of CpGs were apparent when the read length was 8,000 bp (Figure 5.2a, purple line). Similarly in the HG002 dataset, 12.3%, 37.5% and 51.1% of CpGs were within 100, 500 and 1,000 bp from the phased hetSNVs, respectively. Again, 81.3% of CpGs could be covered when the read length was 8,000 bp (Figure 5.2b). For the both samples, the sparseness of hetSNVs in individual genomes was emphasized. The common SNP sites have a much denser distribution, but of course most of the common SNPs are not heterozygous within any single individual genome (Supplemental Figure 5.S2). Therefore, longer reads are essential to detect ASMs in real-world situations.

When we try to detect allele-specific expression (ASE) using these phased hetSNVs, the only variants we can rely on are those appearing in the RNA-seq reads, *i.e.*, exonic variants. Thus the situation becomes even difficult for the analysis of ASE. For the AK1 dataset, 46.2% of the phased hetSNVs were found in the intergenic regions, and 50.0% were in the intronic regions (Figure 5.2c). Thus, only the remaining 3.8% of the variants were available in exons to call ASE in this individual. Similarly, 89.1% of $\sim$ 310k exons do not contain such variants, limiting the possibility of determination of expressing allele. While scarcity

of phased hetSNVs within exons was largely explained by the fact that exons constituted only a small fraction of the genome, the density of phased hetSNVs was also smaller over the exons, possibly due to presence of selective pressure on the coding sequences; on average, there were 0.68 SNVs within exons (0.79 SNVs within introns) per 1 kbp.

5.2.3 Simulation analysis of the method's accuracy

Due to the limited availability of ground-truth dataset for personal diploid methylomes, it is difficult to assess quantitatively the accuracy of ASM detection. Therefore, we approximated the accuracy by considering major potential sources of errors. As we described it in the previous sections, the method proceeds as firstly it assigns reads to one of the alleles, then it calls CpG methylation state for each allele. Thus, the overall accuracy are largely affected by errors in assignment of reads to alleles, and by errors in methylation detection itself. Once we assign reads to alleles, the methylation calls are generated just as described previously, and its accuracy was already assessed (both sensitivity and precision >93%)[114]. Therefore, in the followings, we investigated read assignment errors and their consequences in detail.

First, we noted that incorrect read assignment affects the final ASM calls by changing IPD (Inter-Pulse Duration, kinetics measurement reflecting the methylation status) statistics for each allele. Thus, if an assignment error occurs where the methylation statuses coincide for two alleles, it wont affect the IPD statistics, and such an error can be ignored as long as ASM is considered.By contrast, an assignment error within ASM regions alters the IPD statistics by mixing up IPDs from methylated CpGs on one allele and unmethylated CpGs on the other allele, which may miss the underlying ASM. Overall, read assignment errors deteriorate only sensitivity, not specificity, of ASM detection.

The read assignment errors can be due to several factors, namely, wrongly called/phased SNVs or sequencing errors in PacBio reads. Let us consider each case.

1. False-negative SNVs decrease the power to separate the reads, thereby decrease sensitivity of ASM detection. Incorrectly phased SNVs can decrease sensitivity by con-

founding other SNVs as well.

2. With false-positive SNVs, most reads around those SNVs would be assigned to the reference allele regardless of their true origin. The situation may be detected through imbalance of read depth between alleles.

3. Sequencing errors in long reads may seem the most problematic source of inaccuracy at first, but they are relevant only when they occur at SNV sites, and they affect IPD only when they support the wrong allele. Despite the high error rate in aggregate ($\sim$15%), it should be noted that mismatch errors are relatively rarer (<2%) than indels ($\sim$13%) in PacBio reads[87].

To approximate the effect of sequencing errors on accuracy of final ASM calls, we simulated perturbation in IPD and tested how it affects the methylation calls. Firstly, let us assume reads are assigned to an allele based on only 1 SNV site on them, sequencing errors occur at the SNV site for 2% of the reads, and half of them (1%) support the wrong allele. Then, expected frequency of read assignment error was calculated to be $\sim$1%. We added random perturbation, which was proportional to the frequency of read assignment errors, to IPD ratio of every position independently. With this setting, the predictive performance of the method was almost unchanged (both sensitivity and precision were >90%; Figure 5.2d), presumably because the random errors were averaged out in our prediction model. While our analysis simplifies the real situation, it conveys an intuition why sequencing errors would not severely affect the accuracy of the method contrary to the impression. Therefore, we concluded that the major source of inaccuracy of the method would be methylation detection itself, , which is guaranteed to be highly accurate (>93%)[114].

5.2.4 Allele-specific methylation on CGIs and allele-specific expression

Given the fact that exonic phased variants can be found only in small number of transcripts, genome-wide observation of ASM would provide alternative information about transcriptional activity of individual genomes. To prove this concept, we generated diploid methylomes for AK1 dataset and compared it with ASE analysis from RNA-seq data for the same dataset (Figure 5.2e,f).

At the middle of the first panel, a CGI was located at the promoter region of *ZNF597* (Figure 5.2e). We detected ASM around that CGI, and the allele A was unmethylated upstream of the transcription start site (TSS) of *ZNF597*, thus we could predict the gene was expressed exclusively from allele A. Consistent with the prediction, we found 2 long reads and 59 short reads support the transciption from the allele A while no reads supported the other allele, B. This identification of ASE was based on an exonic SNV within the last exon, and this was the only exonic SNV available in this region, highlighting the sparseness of the phased hetSNVs in exons.

The second example region is *GNAS* complex locus, where several isoforms of the *GNAS* gene are known to show allele-specific expression (Figure 5.2f) [7]. Specifically, while Gsα at the right end is expressed from both alleles, A/B transcripts, and *XL*α*s* are paternally (allele B in Figure 5.2f), and the *NESP55* is maternally expressed (allele A). In our result, that Gsα was expressed from both alleles was confirmed, as the exon specific to the isoform contained a phased variant, and both alternative alleles were observed in the RNA-seq reads. For the other isoforms, though we could not observe directly their allelic expression pattern due to the lack of phased variants in exonic regions, the CGIs located at the promoter regions of each isoforms showed an ASM pattern consistent with the expected expression pattern; the two CGI regions at the promoters of A/B transcripts, *XL*α*s*, and *GNAS-AS* were allele-specifically methylated on the same allele, B, and the CGI at the promoter of *NESP55* was methylated on the other allele, that is, allele A. Thus, we could predict the

expected expression pattern for this locus through their methylation pattern.

These examples demonstrated that the detected ASM status of CGIs can reflect the expression status of corresponding genes/isoforms. Therefore, such ASM on CGIs would be useful information especially when ASE is difficult to detect due to the absence of phased variant sites within exons.

5.2.5 The use of diploid methylomes to detect allele-specific methylation CGIs

Applying the same methodology to HG002 dataset, we determined the methylation status of genome-wide CGIs by summarizing the allelic methylation status of CpG sites contained in each CGI. Of the 26,866 CGIs in the entire genome, we studied 20,140 with at least 30 CpGs to focus on the more functional CGIs. Of these, 5,063 were not covered by long reads after allelic origin assignment, partly because they were relatively distant (4,016 were separated by $\geq$5000 bp) from their nearest hetSNVs. We required that all CGIs should be covered by a sufficient number ($\geq$16.0X for each haploid) of long reads, to reduce the false discovery level [114]. A total of 7,093 CGIs met these criteria. We calculated the methylation score for each CGI as the average of the methylation scores of all CpGs comprising the CGI (Figure 5.3a). Then, we selected the 70 CGIs with the top 1% of absolute differences ($\geq$0.68) in methylation scores between the "haploids" of the diploid methylomes (Supplemental Table 1).

For comparison, we analyzed the AK1 dataset to observe 7,322 (of 20,140) CGIs were not covered by any read, but 10,087 of remaining 12,818 had sufficient coverage ($\geq$16.0X), and 139 CGIs (1.3%) had methylation difference $\geq$0.68, which is almost consistent with the ratio in the HG002 dataset. Thus, in what follows, we continue our analysis using the HG002 data. We noted that the distances between these ASM CGIs and the hetSNVs were not necessarily small. Of the 70 ASM CGIs, 28 were separated by $\geq$500 bp, and 9 by $\geq$1,000 bp, from their nearest phased hetSNVs, which meant that the methylation status of alleles

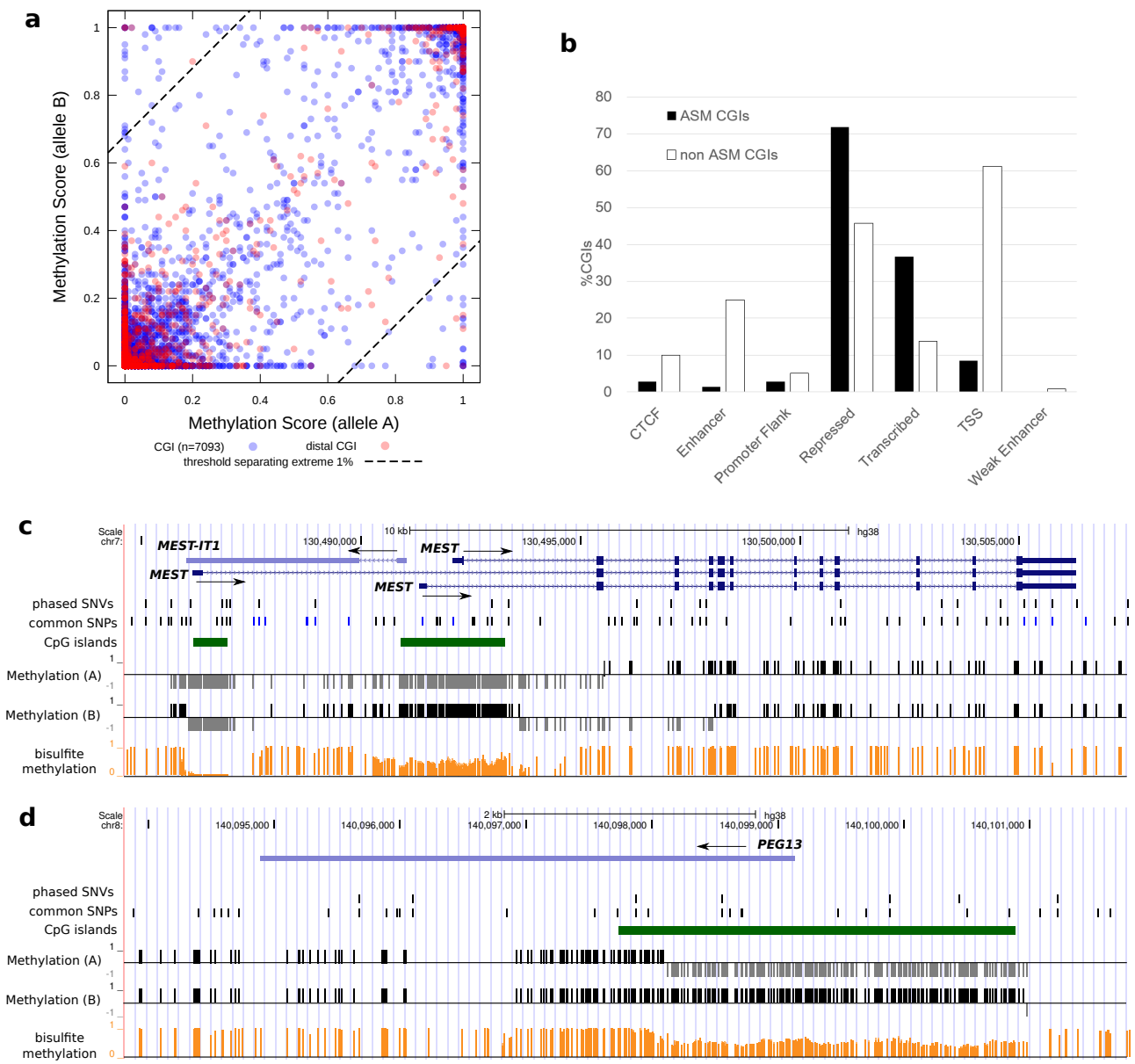

a
Methylation Score (allele B)
Methylation Score (allele A)
CGI (n=7093)
distal CGI
threshold separating extreme 1%
b
ASM CGIs
non ASM CGIs
%CGIs
CTCF
Enhancer
Promoter Flank
Repressed
Transcribed
TSS
Weak Enhancer
c
MEST-IT1
MEST
phased SNVs
common SNPs
CpG islands
Methylation (A)
Methylation (B)
bisulfite methylation
d
PEG13
phased SNVs
common SNPs
CpG islands
Methylation (A)
Methylation (B)
bisulfite methylation

Figure 5.3: **a.** Summary of the methylation scores (for each allele) of CGIs in personal diploid methylomes in HG002. Each CGI is shown as a circle. On the two opposite corners (top left and bottom right), CGIs with maximally 1% largest absolute differences in methylation levels between the two alleles were provisionally classified as allele-specifically methylated CGIs. Red circles: The corresponding CGIs were separated from the nearest hetSNVs by 1,000 bp or more. Blue circles: Separations $< 1,000$ bp. **b.** Distribution of each type of CGIs, ASM (black bar) or non-ASM (white bar), with respect to functional annotation of genomic regions. **c.** Example showing personal diploid methylomes in *MEST* gene-coding region of HG002 (Ashkenazim Trio Son) genome. Although the upstream CGI (with 66 CpG sites) was unmethylated in both alleles, the downstream larger CGI (with 184 CpG sites) exhibited ASM. The CGIs corresponded to the promoter regions of different isoforms of the genes. **d.** Another example of ASM around imprinted gene *PEG13*, *paternally expressed gene 13*.

of these CGIs could not be measured simultaneously when the reads were shorter than 500 or 1,000 bp.

In order to confirm that the detected ASM CGIs are functionally relevant, we compared the ASM CGIs with a genomic annotation from the combined segmentation by Segway and ChromHMM defined in the ENCODE project (Figure 5.3b) [21]. Of note, CGIs in general were significantly overlapping with TSS as expected. In contrast to this background, CGIs showing ASM were overlapping with segments annotated as "transcribed regions" or "repressed regions" more than TSS. This result may seem somewhat contradictory at first thought, since any single gene cannot be both transcribed and repressed at the same time, but it is still plausible that this would be a correct categorization for the regions with ASM genes because they can be, by definition, in two contrastive states in each of the alleles.

We also confirmed that our list of candidate ASM CGIs contained a number of CGIs overlapping with known imprinted genes. For example, we reproduced the expected ASM around known imprinted genes such as *MEST*(Figure 5.3c), *PEG13*(Figure 5.3d), *HYMAI*, *ZNF597*, *etc.*(Supplemental Table 1)[5]. Indeed, CGIs with larger difference of methylation between two alleles enriched with imprinted genes ($p = 0.007$, U test). That we successfully recovered the imprinted genes as ASM region confirmed the validity of our method again.

5.3 Discussion

In this work, we studied personal diploid methylomes to directly characterize the ASM status of genome-wide CGIs, based on a set of phased hetSNVs specific to each sample. Compared to the previous studies employing short read sequencing [106, 36, 56], one of the novelty of our approach is that we called methylation using kinetic information from long SMRT reads; we did not employ any chemical treatment such as bisulfite conversion, which breaks DNA into small fragments of length $< 1,500$ bp [128]. By this design, we could fully exploited the lengths of the PacBio reads ($> 8,000$ bp in our data). We determined the allelic origins of

more than half of the sequencing data, thus we were able to cover more CpGs in the genome. We previously reported that read coverage of ∼20x is required for detecting regional CpG methylation [114]. In order that sufficient read coverage is available for each allele after the separation, this number can include a margin since some reads will not contain any informative hetSNV and be filtered out. Therefore, 40∼50x of reads would be sufficient for the detection of ASM.

We considered several factors to be the main causes of read assignment errors, such as inaccuracy in the SNV set or erroneous phasing of them, and sequencing errors within raw PacBio reads. The simulation revealed, however, the accuracy would not be affected severely by sequencing errors, as they are random in its nature[87]. On the other hand, wrong SNV calls/phasing can be a source of biased errors, which should be alleviated by using a phased SNV set of better quality. Therefore, the overall accuracy of detected methylation statuses would essentially replicate the prediction performance of original methylation detection method, e.g., >90% for the regions with sufficient sequencing depth, say, 20-fold on each allele[114].

Other existing methods to detect genome-wide ASM had their own weaknesses. Use of methylation-sensitive restriction enzymes clearly requires that restriction sites be present, and the resolution and accessibility of the method is limited by the distribution of this additional genomic feature which may not be biologically relevant to ASM. Use of antibody for methylated cytosines followed by sequencing to detect heterozygous variants does not need additional features to anchor. However, it cannot conclude which CpG sites were really in ASM within the detected region, especially when ASM CGI was located close to other CGI. For example, Figure 5.2e illustrates that the two neighboring promoters for *Gsα* and *A/B* transcripts in the *GNAS* locus exhibit different methylation states. Another important advantage of our method is that long reads enables finding ASM associated with distal heterozygous variants, and we demonstrated that such cases were not necessarily rare as illustrated in Figure 5.2a-b.

The mechanism of establishment of ASM remains to be understood. Shoemaker *et al.* reported that disruption of CpG sites could lead to ASM around them [106]; however, distant mutations might be involved because, when we inspected diploid methylomes using long reads, we found that some methylation statuses can be associated only with distant SNVs from them. One future direction of the study would involve constructing the diploid methylomes for sufficient number of individuals, then it may delineate true causal relationship between each variant and ASM.

As we demonstrated, comprehensive information about genome-wide ASM status may complement ASE observation if we assume methylation status of promoter CGIs are expected to be correlated with its transcriptional activity. We touched a couple of examples to support this intuition with the help of RNA-seq data in AK1 dataset, and indicated the analysis of ASM could recapture some imprinted genes in HG002 dataset. While we cannot use epigenetic observation as a complete surrogate for expression data, it would complement ASE statuses of transcripts when they are more difficult to observe.

We also demonstrated that the long reads are essential for the study of ASM given a sparse distribution of heterozygous variants within individuals. In addition to that, the advent of linked-reads technology (such as 10x GemCode/Chromium) enabled us to extract long-range ($\sim$100 kbp) co-occurrences of DNA sequences. Such technology renders it easier to sequence individual genomes in a manner that the majority of variants are haplotype-phased [131]. The more accessible the haplotype-phased genomes becomes, the more reasonable it becomes to study epigenome being aware of the existence of two alleles. Therefore, to understand the full spectrum of ASM biology, it is essential to study diploid methylomes employing long reads.

5.4 Methods

5.4.1 Data source

DNA sequencing and RNA sequencing data for AK1 were obtained from a public repository (Accession No. PRJNA298944). DNA sequencing data and phased variants information for HG002 were obtained from a FTP repository of Genome in a Bottle consortium (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release). The original cell lines are available as well (Coriell GM24385). Both samples are lymphoblastoid cell lines (LCL).

5.4.2 Generating the diploid methylomes

For both dataset, we aligned raw PacBio reads to hg38 reference genome using standard mapping protocol of SMRT Analysis. Resulted `.cmp` files were first converted to SAM format to count the number of variants contained within each aligned read. We considered only SNVs as allele-distinguishing variants, then we were able to ignore the indels, which account majority of sequencing errors in PacBio reads. Consequently, read assignment to allele can be enough accurate for ASM detection, since the rate of mismatch error is relatively low ($< 2\%$). The assignment of allele for each read was determined by majority voting. If the read contains more than two PHVs within the reads, the evidence on each PHVs were just combined. Reads were excluded from further analysis if they did not contain any PHVs, or the voting was tied. Then `.cmp` files were partitioned accordingly by in-house scripts written in bash and scala. For each file, methylation calls were generated as previously described [114]. Read length statistics before/after the allelic assignment were calculated using `cmph5tools summary` command from SMRT Analysis.

To discuss the cause of read assignment errors in ASM detection pipeline and how they affect the accuracy of final ASM calls, we assumed IPD ratio statistics around the SNV are perturbed by:

$$\Delta IPD = random(-1, 1) * P * (IPD - 1.0), \tag{5.1}$$

where $random(-1, 1)$ is sampled from the uniform distribution over $[-1, 1]$ and $P = 1\%$.

5.4.3 Calculate the distribution of PHVs with respect to CGIs or exons

CpG islands annotation was retrieved from UCSC Genome Browser. For both PHVs and common SNPs, the distance from CGI was calculated as genomic distance from the center of CGI. To calculate the distribution of PHVs and common SNPs with respect to exons, a gene model (GENCODE ver.24) was intersected with each feature.

5.4.4 Identifying the CGIs with ASM

As AgIn reported methylation status of each CpG site, we calculated methylation level of CGI as the (unweighted) average of methylation status presented as 0 or 1 (unmethylation or methylation, resp.) within the CGI. Then filters were applied as described in the Result section. After identifying the ASM CGIs, each CGI was associated with gene(s) by manual inspection on a browser and the most closest gene (transcript) was recorded.

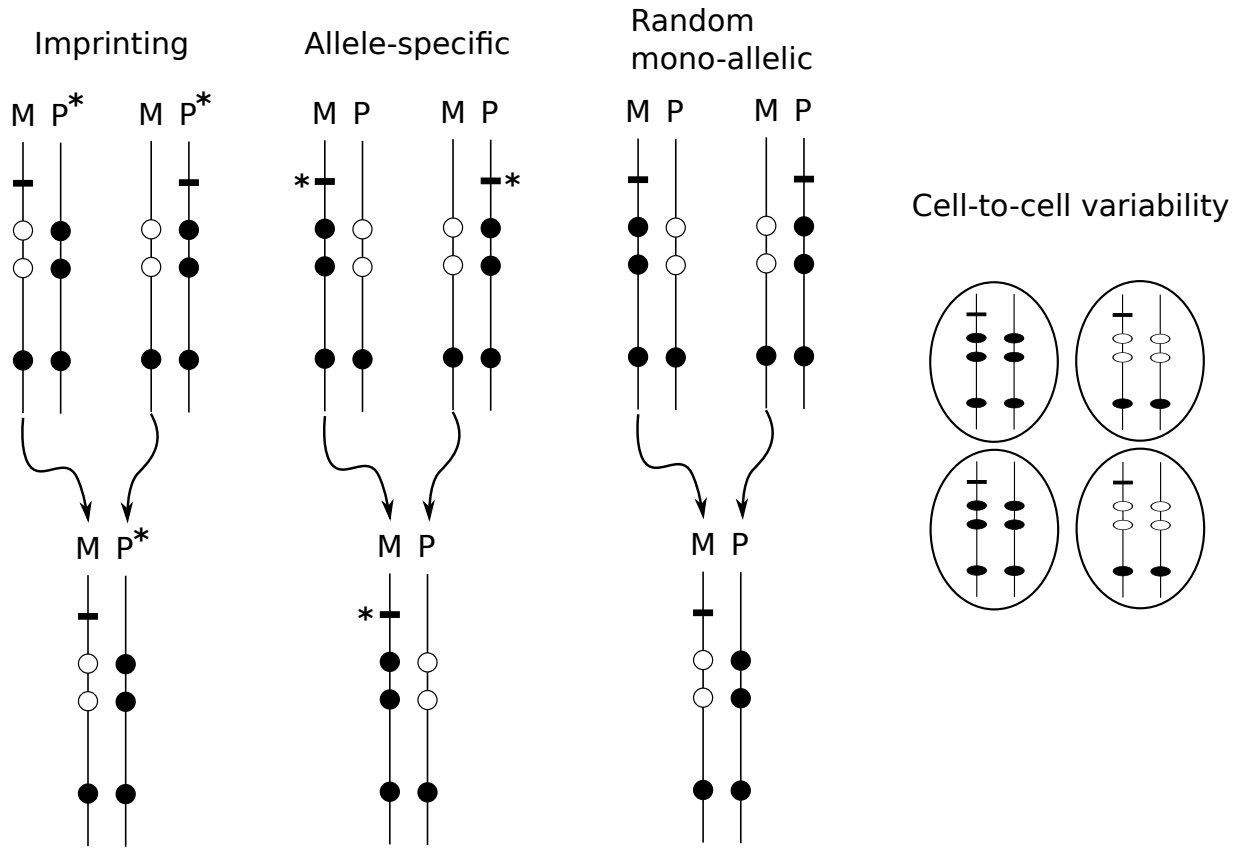


Figure 5.S1: Schematic illustration of the four types of diploid methylomes that may show intermediate levels of methylation. Each figure depicts the methylation status of CpG sites (circles) and sequence variants (horizontal bars) for maternal (M) and paternal (P) chromosomes of an imaginary trio. The asterisks indicate features correlated with methylation status. (1) Methylation associated with genomic imprinting. (2) Allele-specific methylation in the narrower sense. Methylation status is determined by the local genomic sequence. (3) Mono-allelic methylation where methylation is not correlated with the genome of the parent-of-origin or a local allele. (4) Cell-to-cell variability, or epigenomic heterogeneity.

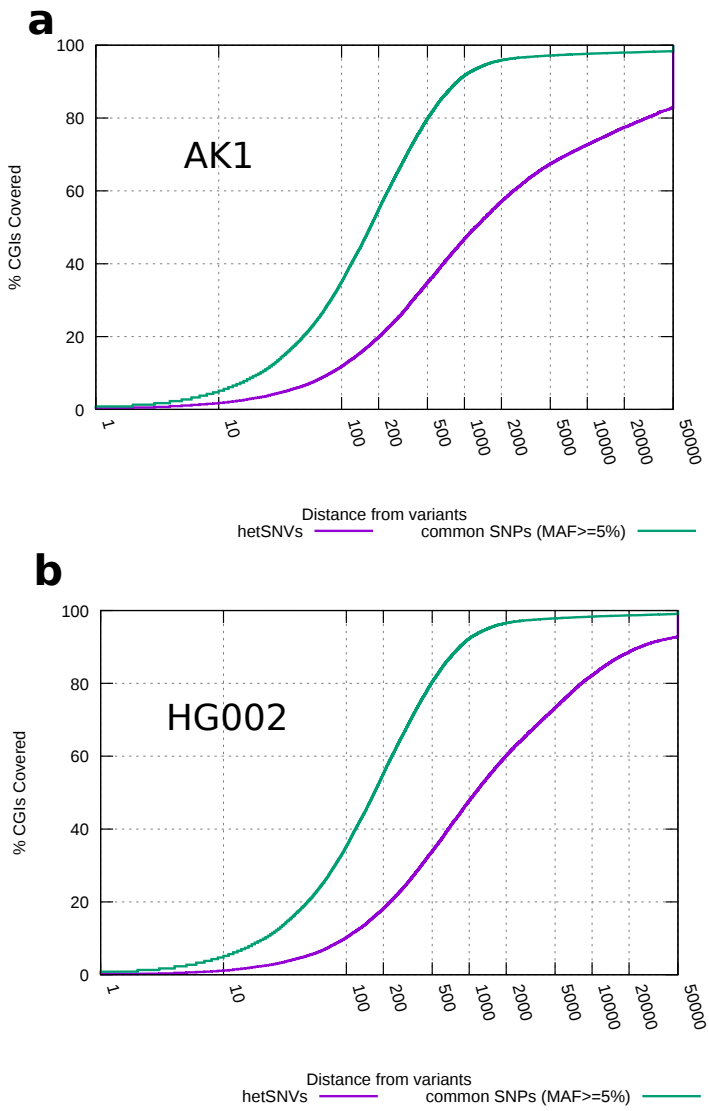


Figure 5.S2: Ratio of genomic CpG islands that were located within the given distance from the nearest genomic features, heterozygous SNVs (purple) or common SNPs with MAF≥5% (green). Each plot shows the comparison in **a** AK1 and **b** HG002.

Conclusion

Through the five chapters of the thesis, we demonstrated how one can extract useful epigenetic information from SMRT long reads, and how this method leads to comprehensive observation of CpG methylation status of large genomes. As we briefly reviewed in Chapter 1, a number of methods were devised to exploit long reads for high-quality genome assembly. While there are quite a variety of strategies, one of the most profound fact which benefits all such projects is that longer reads provide us an opportunity to resolve "difficult parts" of the genomes. This is mainly because identifying original location of reads is much easier for longer reads once one can handle the frequent errors inherent in long read technology.

Essentially, the project we demonstrated here is an application of this principle for epigenetic studies. Just as in studies of genome, short reads suffer from low mapability within low complexity regions. We tried to resolve this situation by using SMRT long reads which come with epigenetic information. Shortly after we set up our method, AgIn, we started with the study of known repeat elements in human and medaka genomes, demonstrating that observation of individual repeats is straightforward with long reads (Chapters 2,3). Then in Chapter 4, we moved on to the study of another cryptic part of genome: centromeres. While it was often the case that we had to give up saying something about centromeric regions with short read, we could observe by long read that there can be a variaty of methylation status within alpha-satellite arrays. Finally, in Chapter 5, we tackled the problem of investigating allele-specific methylation in human (diploid) genomes. To make the connection between this study and the previous chapters' clear, one could say that it involved the largest "repeats"

structures in human genome, that is, pairs of homologous chromosomes. By examining the distribution of heterozygous variants, we explicitly quantified why allele-specific methylation was difficult to study with short reads. With long reads, one could extract CpG methylation information independently for each allele. We termed this *a personal diploid methylome*, as it provided genome-wide landscape of allele specific methylation solely based on personal epigenetic data.

While we made our method grounded with a number of non-trivial applications, there remains many related problems open. For example, it would be of great interest if one could extend our idea to epigenetic studies of plants, which have a lot of non-canonical (non-CpG) methylation. Our study of centromeric regions was limited by availablity of genome assembly, and similar studies for human would require better assembly of human centromore, which remains an extremely challenging task.

Though its concept has been proven, we cannot predict exactly how far one can go with current (and with future single molecule sequencing) technologies. With ever-evolving sequencing technologies, it's on bioinformaticians' shoulder to invent appropriate theories and to implement them, to make the most of the state-of-the-art technologies of the time.

Acknowledgement

My special thank goes to my advisor Prof. Shinichi Morishita. Without his guidance the whole research project would have never be existed in the current form.

I'm very grateful to all my research collaborators and coauthors of original works, (honorific titles omitted; listed alphabetical order) Hideaki Yurino, Hiroyuki Ishiura, Hiroyuki Takeda, Jonas Korlach, Junko Taniguchi, Jun Mitsui, Jun Yoshimura, Kazuki Ichikawa, Kin Fai Au Koichiro Doi, Masahiko Kumagai, Naoki Irie, Ryohei Nakamura, Shingo Tomioka, Shinichi Morishita, Shoji Tsuji, Stephen W. Turner, Tatsuya Tsukahara, Wei Qu, Yui Uchida, Yuji Takahashi, Yunhao Wang and Yusuke Inoue, with whom the research was

always productive. They also generously granted permission to include our work as a part of this thesis.

I'd like to thank Yoshihiko Suzuki and Yuichi Motai as well as other members in Morishita lab for many insightful comments on the draft version of this thesis. Thank you all.

Bibliography

[1] Abouelhoda, M. I. and Ohlebusch, E. (2003). A local chaining algorithm and its applications in comparative genomics. In *International Workshop on Algorithms in Bioinformatics*, pages 1–16. Springer.

[2] Allshire, R. C. and Karpen, G. H. (2008). Epigenetic regulation of centromeric chromatin: old dogs, new tricks? *Nature Reviews Genetics*, **9**(12), 923.

[3] 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.

[4] Au, K. F., Sebastiano, V., Afshar, P. T., Durruthy, J. D., Lee, L., Williams, B. A., van Bakel, H., Schadt, E. E., Reijo-Pera, R. A., Underwood, J. G., *et al.* (2013). Characterization of the human esc transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences*, **110**(50), E4821–E4830.

[5] Baran, Y., Subramaniam, M., Biton, A., Tukiainen, T., Tsang, E. K., Rivas, M. A., Pirinen, M., Gutierrez-Arcelus, M., Smith, K. S., Kukurba, K. R., *et al.* (2015). The landscape of genomic imprinting across diverse adult human tissues. *Genome research*, **25**(7), 927–936.

[6] Bashir, A., Klammer, A. a., Robins, W. P., Chin, C.-S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P., Sebra, R., Sorenson, J., Bullard, J., Yen, J., Valdovino, M., Mollova, E., Luong, K., Lin, S., LaMay, B., Joshi, A., Rowe, L., Frace, M., Tarr, C. L., Turnsek, M., Davis, B. M., Kasarskis, A., Mekalanos, J. J., Waldor, M. K., and Schadt, E. E. (2012). A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology*, **30**, 701–707.

[7] Bastepe, M. (2007). The gnas locus: quintessential complex gene encoding gsα, xlαs, and other imprinted transcripts. *Current genomics*, **8**(6), 398–414.

[8] 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.

[9] Beckmann, N. D., Karri, S., Fang, G., and Bashir, A. (2014). Detecting epigenetic motifs in low coverage and metagenomics settings. *BMC bioinformatics*, **15**(Suppl 9), S16.

[10] Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., and Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology*, **33**(6), 623–630.

[11] 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.

[12] Chaisson, M. J. and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): application and theory. *BMC bioinformatics*, **13**(1), 238.

[13] Chen, F.-M. (1988). Effects of a: T base pairs on the b–z conformational transitions of dna. *Nucleic acids research*, **16**(5), 2269–2281.

[14] Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., *et al.* (2013). Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. *Nature methods*, **10**(6), 563–569.

[15] Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., *et al.* (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*, **13**(12), 1050–1054.

[16] Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G., and Reik, W. (2016). Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome biology*, **17**(1), 1.

[17] 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.

[18] Clark, T. A., Spittle, K. E., Turner, S. W., and Korlach, J. (2011). Direct detection and sequencing of damaged dna bases. *Genome integrity*, **2**(1), 10.

[19] 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.

[20] 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. (2008). Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, **452**(7184), 215–219.

[21] Consortium, T. E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

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

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

[24] Deonovic, B., Wang, Y., Weirather, J., Wang, X.-J., and Au, K. F. (2017). Idp-ase: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. *Nucleic acids research*, **45**(5), e32–e32.

[25] Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., Graef, S., Johnson, N., Herrero, J., Tomazou, E. M., *et al.* (2008). A bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature biotechnology*, **26**(7), 779–785.

[26] Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S. (2006). Dna methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, **38**(12), 1378–85.

[27] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-time dna sequencing from single polymerase molecules. *Science (New York, N.Y.)*, **323**, 133–8.

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

[29] English, A. C., Salerno, W. J., and Reid, J. G. (2014). Pbhoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC bioinformatics*, **15**(1), 180.

[30] Eppstein, D., Galil, Z., Giancarlo, R., and Italiano, G. F. (1992). Sparse dynamic programming i: linear cost functions. *Journal of the ACM (JACM)*, **39**(3), 519–545.

[31] Fang, G., Munera, D., Friedman, D. I., Mandlik, A., Chao, M. C., Banerjee, O., Feng, Z., Losic, B., Mahajan, M. C., Jabado, O. J., Deikus, G., Clark, T. a., Luong, K., Murray, I. a., Davis, B. M., Keren-Paz, A., Chess, A., Roberts, R. J., Korlach, J., Turner, S. W., Kumar, V., Waldor, M. K., and Schadt, E. E. (2012). Genome-wide mapping of methylated adenine residues in pathogenic escherichia coli using single-molecule real-time sequencing. *Nature Biotechnology*.

[32] 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.

[33] Ferragina, P. and Manzini, G. (2000). Opportunistic data structures with applications. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 390–398. IEEE.

[34] 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**(6), 461–465.

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

[36] Gertz, J., Varley, K. E., Reddy, T. E., Bowling, K. M., Pauli, F., Parker, S. L., Kucera, K. S., Willard, H. F., and Myers, R. M. (2011). Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genet*, **7**(8), e1002228.

[37] 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., Zhang, X., Coyne, M., Fostel, J. L., Holmes, L., Meldrim, J., Guttman, M., Epstein, C., Park, H., Kohlbacher, O., Rinn, J., Gnirke, A., Lander, E. S., Bernstein, B. E., and Meissner, A. (2013). Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell*, **153**(0), 1149 – 1163.

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

[39] Gravina, S., Dong, X., Yu, B., and Vijg, J. (2016). Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome. *Genome Biology*, **17**(1), 1–8.

[40] 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. (2015). Dna methylation on n 6-adenine in c. elegans. *Cell*, **161**(4), 868–878.

[41] 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., Olshen, A., Ballinger, T., Zhou, X., Forsberg, K. J., Gu, J., Echipare, L., O'Geen, H., Lister, R., Pelizzola, M., Xi, Y., Epstein, C. B., Bernstein, B. E., Hawkins, R. D., Ren, B., Chung, W.-Y., Gu, H., Bock, C., Gnirke, A., Zhang, M. Q., Haussler, D., Ecker, J. R., Li, W., Farnham, P. J., Waterland, R. a., Meissner, A., Marra, M. a., Hirst, M., Milosavljevic, A., and Costello, J. F. (2010). Comparison of sequencing-based methods to profile dna methylation and identification of monoallelic epigenetic modifications.

[42] Huddleston, J., Chaisson, M. J., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N., Vives, L., *et al.* (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*, **27**(5), 677–685.

[43] Ichikawa, K., Takeda, H., Yoshimura, J., Doi, K., Kumagai, M., Irie, N., Nakamura, R., Tomioka, S., Morishita, S., Uchida, Y., *et al.* (2017). Centromere evolution and cpg methylation during vertebrate speciation. *Nature Communications*, **8**(1), 1833.

[44] 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.

[45] Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, **13**(7), 484–492.

[46] Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A., and David, N. T. (2017). Hinge: long-read assembly achieves optimal repeat resolution. *Genome research*, **27**(5), 747–756.

[47] Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Dreszer, T. R., Fujita, P. a., Guruvadoo, L., Haeussler, M., Harte, R. a., Heitner, S., Hinrichs, A. S., Learned, K., Lee, B. T., Li, C. H., Raney, B. J., Rhead, B., Rosenbloom, K. R., Sloan, C. a., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M., and Kent, W. J. (2014). The UCSC Genome Browser database: 2014 update. *Nucleic acids research*, **42**(Database issue), D764–70.

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

[49] Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V. V., Schupf, N., Vilain, E., Morris, M., Haghighi, F., and Tycko, B. (2008). Genomic surveys by methylation-sensitive snp analysis identify sequence-dependent allele-specific dna methylation. *Nat Genet*, **40**(7), 904–8.

[50] 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.

[51] Koo, D.-H., Han, F., Birchler, J. A., and Jiang, J. (2011). Distinct dna methylation patterns associated with active and inactive centromeres of the maize b chromosome. *Genome research*, **21**(6), 908–914.

[52] 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., and Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, **30**, 693–700.

[53] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, **27**(5), 722–736.

[54] 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. (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.

[55] Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *bioinformatics*, **27**(11), 1571–1572.

[56] Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M., and Snyder, M. (2014). Whole-genome haplotyping using long reads and statistical methods. *Nature biotechnology*, **32**(3), 261.

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

[58] Kuwahara, Y., Shimada, A., Mitani, H., and Shima, A. (2003). γ-ray exposure accelerates spermatogenesis of medaka fish, oryzias latipes. *Molecular reproduction and development*, **65**(2), 204–211.

[59] Lander, E. S. and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**(3), 231–239.

[60] Lee, D., Karchin, R., and Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from dna sequence. *Genome research*, **21**(12), 2167–2180.

[61] Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., Lohr, J. G., Harris, C. C., Ding, L., Wilson, R. K., Wheeler, D. A., Gibbs, R. A., Kucherlapati, R., Lee, C., Kharchenko, P. V., Park, P. J., and Network, T. C. G. A. R. (2012). Landscape of somatic retrotransposition in human cancers. *Science*, **337**(6097), 967–971.

[62] Leslie, C., Eskin, E., and Noble, W. S. (2001). The spectrum kernel: A string kernel for svm protein classification. In *Biocomputing 2002*, pages 564–575. World Scientific.

[63] Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., and Yen, C.-a. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, **518**, 350–354.

[64] Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature*, **366**(6453), 362–365.

[65] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

[66] Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, **32**(14), 2103–2110.

[67] Li, H. (2017). Minimap2: versatile pairwise alignment for nucleotide sequences. *arXiv*, **1708**.

[68] Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, **26**(5), 589–595.

[69] Li, Y., Miyanari, Y., Shirane, K., Nitta, H., Kubota, T., Ohashi, H., Okamoto, A., and Sasaki, H. (2013). Sequence-specific microscopic visualization of dna methylation status at satellite repeats in individual cell nuclei and chromosomes. *Nucleic acids research*, **41**(19), e186–e186.

[70] 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.

[71] 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., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**(7271), 315–22.

[72] 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. (2013). Sequencing the unsequenceable: expanded cgg-repeat alleles of the fragile x gene. *Genome research*, **23**(1), 121–128.

[73] McKinley, K. L. and Cheeseman, I. M. (2016). The molecular basis for centromere identity and function. *Nature Reviews Molecular Cell Biology*, **17**(1), 16.

[74] Meaburn, E. L., Schalkwyk, L. C., and Mill, J. (2010). Allele-specific methylation in the human genome: implications for genetic studies of complex disease. *Epigenetics*, **5**(7), 578–582.

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

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

[77] Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, **24**(24), 2818–2824.

[78] Miller, W. and Myers, E. W. (1988). Sequence comparison with concave weighting functions. *Bulletin of mathematical biology*, **50**(2), 97–120.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] Myers, E. W. (1986). An o (nd) difference algorithm and its variations. *Algorithmica*, **1**(1), 251–266.

[84] Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*, **21**(suppl_2), ii79–ii85.

[85] Myers, G. (2014). Efficient local alignment discovery amongst noisy long reads. In *International Workshop on Algorithms in Bioinformatics*, pages 52–67. Springer.

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

[87] Ono, Y., Asai, K., and Hamada, M. (2012). Pbsim: Pacbio reads simulatortoward accurate genome assembly. *Bioinformatics*, **29**(1), 119–121.

[88] Palme, J., Hochreiter, S., and Bodenhofer, U. (2015). Kebabs: an r package for kernel-based analysis of biological sequences. *Bioinformatics*, **31**(15), 2574–2576.

[89] Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., *et al.* (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*, **12**(8), 780–786.

[90] Peng, Q. and Ecker, J. R. (2012). Detection of allele-specific methylation through a generalized heterogeneous epigenome model. *Bioinformatics*, **28**(12), i163–i171.

[91] 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.

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

[93] Riggs, A. D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, **14**(1), 9–25.

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

[95] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, **4**(4), 406–425.

[96] Schadt, E. E., Banerjee, O., Fang, G., Feng, Z., Wong, W. H., Zhang, X., Kislyuk, A., Clark, T. A., Luong, K., Keren-Paz, A., Chess, A., Kumar, V., Chen-Plotkin, A., Sondheimer, N., Korlach, J., and Kasarskis, A. (2012). Modeling kinetic rate variation in third generation dna sequencing data to detect putative modifications to dna bases. *Genome Res*.

[97] Schalkwyk, L. C., Meaburn, E. L., Smith, R., Dempster, E. L., Jeffries, A. R., Davies, M. N., Plomin, R., and Mill, J. (2010). Allelic skewing of dna methylation is widespread across the genome. *Am J Hum Genet*, **86**(2), 196–212.

[98] 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.

[99] 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.

[100] Schübeler, D. (2015). Function and information content of DNA methylation. *Nature*, **517**(7534), 321–326.

[101] Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., Rajagopal, N., Nery, J. R., Urich, M. A., Chen, H., *et al.* (2015). Human body epigenome maps reveal noncanonical dna methylation variation. *Nature*, **523**(7559), 212.

[102] Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., and Schatz, M. (2017). Accurate detection of complex structural variations using single molecule sequencing. *bioRxiv*, page 169557.

[103] Seo, J.-S., Rhie, A., Kim, J., Lee, S., Sohn, M.-H., Kim, C.-U., Hastie, A., Cao, H., Yun, J.-Y., Kim, J., *et al.* (2016a). De novo assembly and phasing of a korean human genome. *Nature*.

[104] Seo, J.-S., Rhie, A., Kim, J., Lee, S., Sohn, M.-H., Kim, C.-U., Hastie, A., Cao, H., Yun, J.-Y., Kim, J., *et al.* (2016b). De novo assembly and phasing of a korean human genome. *Nature*, **538**(7624), 243–247.

[105] Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, **29**(1), 308–311.

[106] 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.

[107] Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, **11**(8), 817–820.

[108] Smith, Z. D. and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, **14**(3), 204–220.

[109] 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.

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

[111] Steinberg, K. M., Schneider, V. A., Graves-Lindsay, T. A., Fulton, R. S., Agarwala, R., Huddleston, J., Shiryev, S. A., Morgulis, A., Surti, U., Warren, W. C., *et al.* (2014). Single haplotype assembly of the human genome from a hydatidiform mole. *Genome research*, **24**(12), 2066–2076.

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

[113] Suzuki, H. and Kasahara, M. (2017). Acceleration of nucleotide semi-global alignment with adaptive banded dynamic programming. *bioRxiv*, page 130633.

[114] Suzuki, Y., Korlach, J., Turner, S. W., Tsukahara, T., Taniguchi, J., Qu, W., Ichikawa, K., Yoshimura, J., Yurino, H., Takahashi, Y., *et al.* (2016). Agin: measuring the landscape of cpg methylation of individual repetitive elements. *Bioinformatics*, **32**(19), 2911–2919.

[115] Tang, A., Huang, Y., Li, Z., Wan, S., Mou, L., Yin, G., Li, N., Xie, J., Xia, Y., Li, X., *et al.* (2016). Analysis of a four generation family reveals the widespread sequence-dependent maintenance of allelic DNA methylation in somatic and germ cells. *Scientific Reports*, **6**, 19260.

[116] 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., Menzies, A., Roman-Garcia, P., Fullam, A., Gerstung, M., Shlien, A., Tarpey, P. S., Papaemmanuil, E., Knappskog, S., Van Loo, P., Ramakrishna, M., Davies, H. R., Marshall, J., Wedge, D. C., Teague, J. W., Butler, A. P., Nik-Zainal, S., Alexandrov, L., Behjati, S., Yates, L. R., Bolli, N., Mudie, L., Hardy, C., Martin, S., McLaren, S., O'Meara, S., Anderson, E., Maddison, M., Gamble, S., Group, I. B. C., Group, I. B. C., Group, I. P. C., Foster, C., Warren, A. Y., Whitaker, H., Brewer, D., Eeles, R., Cooper, C., Neal, D., Lynch, A. G., Visakorpi, T., Isaacs, W. B., van't Veer, L., Caldas, C., Desmedt, C., Sotiriou, C., Aparicio, S., Foekens, J. A., Eyfjörd, J. E., Lakhani, S. R., Thomas, G., Myklebost, O., Span, P. N., Børresen-Dale, A.-L., Richardson, A. L., Van de Vijver, M., Vincent-Salomon, A., Van den Eynden, G. G., Flanagan, A. M., Futreal, P. A., Janes, S. M., Bova, G. S., Stratton, M. R., McDermott, U., and Campbell, P. J. (2014). Extensive transduction of nonrepetitive dna mediated by l1 retrotransposition in cancer genomes. *Science*, **345**(6196).

[117] Tycko, B. (2010). Allele-specific DNA methylation: beyond imprinting. *Human Molecular Genetics*, **19**(R2), R210–R220.

[118] Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research*, **27**(5), 737–746.

[119] 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*, **10**(7), 662–669.

[120] 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.

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

[122] Wu, X., Sun, M.-a., Zhu, H., and Xie, H. (2015). Nonparametric Bayesian clustering to detect bipolar methylated genomic loci. *BMC bioinformatics*, **16**(1), 1.

[123] Xiao, C.-L., Chen, Y., Xie, S.-Q., Chen, K.-N., Wang, Y., Han, Y., Luo, F., and Xie, Z. (2017). Mecat: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *nature methods*, **14**(11), 1072–1074.

[124] Xie, W., Schultz, M. D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J. W., Tian, S., Hawkins, R. D., Leung, D., Yang, H., Wang, T., Lee, A. Y., Swanson, S. A., Zhang, J., Zhu, Y., Kim, A., Nery, J. R., Urich, M. A., Kuan, S., an Yen, C., Klugman, S., Yu, P., Suknuntha, K., Propson, N. E., Chen, H., Edsall, L. E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.-Y., Chi, N. C., Antosiewicz-Bourget, J. E., Slukvin, I., Stewart, R., Zhang, M. Q., Wang, W., Thomson, J. A., Ecker, J. R., and Ren, B. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, **153**(5), 1134–48.

[125] 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.

[126] Yamagata, K., Yamazaki, T., Miki, H., Ogonuki, N., Inoue, K., Ogura, A., and Baba, T. (2007). Centromeric dna hypomethylation as an epigenetic signature discriminates between germ and somatic cell lineages. *Developmental biology*, **312**(1), 419–426.

[127] Yan, H., Kikuchi, S., Neumann, P., Zhang, W., Wu, Y., Chen, F., and Jiang, J. (2010). Genome-wide mapping of cytosine methylation revealed dynamic dna methylation patterns associated with genes and centromeres in rice. *The Plant Journal*, **63**(3), 353–365.

[128] Yang, Y., Sebra, R., Pullman, B. S., Qiao, W., Peter, I., Desnick, R. J., Geyer, C. R., DeCoteau, J. F., and Scott, S. A. (2015). Quantitative and multiplexed dna methylation analysis using long-read single-molecule real-time bisulfite sequencing (smrt-bs). *BMC Genomics*, **16**, 350.

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

[130] 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.

[131] Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., *et al.* (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, **34**(3), 303–311.

[132] Zook, J., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C., Alexander, N., *et al.* (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data*, **3**, 160025.

Observing comprehensive DNA methylomes via single molecule real-time sequencing: application to diploid and centromeric methylation

一分子 DNA シーケンサによる DNA メチル化情報の網羅的観測手法 ― 二倍体ゲノムとセントロメア領域への応用

SUZUKI, Yuta
鈴木　裕太

Submitted on 2017/12/13