Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an environmental prokaryotic community.

Satoshi Hiraoka[1,†,*], Yusuke Okazaki[2], Mizue Anda[3], Atsushi Toyoda[4], Shin-ichi Nakano[2], Wataru Iwasaki[1,3,5,*]

1 Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan

2 Center for Ecological Research, Kyoto University, Otsu, Japan

3 Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan

4 National Institute of Genetics, Mishima, Japan

5 Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Japan

† Present Address: Research and Development Center for Marine Biosciences, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka, Japan

Email: hiraokas@jamstec.go.jp, iwasaki@bs.s.u-tokyo.ac.jp

Abstract

DNA methylation plays important roles in prokaryotes, such as in defense mechanisms against phage infection, and the corresponding genomic landscapes—prokaryotic epigenomes—have recently begun to be disclosed. However, our knowledge of prokaryote methylation systems has been severely limited to those of culturable prokaryotes, whereas environmental communities are in fact dominated by uncultured members that must harbor much more diverse DNA methyltransferases. Here, using single-molecule real-time and circular consensus sequencing techniques, we revealed the 'metaepigenomes' of an environmental prokaryotic community in the largest lake in Japan, Lake Biwa. A total of 19 draft genomes from phylogenetically diverse groups, most of which are yet to be cultured, were successfully reconstructed. The analysis of DNA chemical modifications identified 29 methylated motifs in those genomes, among which 14 motifs were novel. Furthermore, we searched for the methyltransferase genes responsible for the methylation of the detected novel motifs and confirmed their catalytic specificities via transformation experiments involving artificially synthesized genes. Finally, we found that genomes without DNA methylation tended to exhibit higher phage infection levels than those with methylation. In summary, this study proves that metaepigenomics is a powerful approach for revealing the vast unexplored variety of prokaryotic DNA methylation systems in nature.

Introduction

DNA methylation is a major class of epigenetic modification that is found in diverse prokaryotes, in addition to eukaryotes[1]. For example, prokaryotic DNA methylation by sequence-specific restriction-modification (RM) systems that protect host cells from invasion by phages or extracellular DNA has been well characterized and is utilized as a key tool in biotechnology[2,3,4]. In addition, recent studies have revealed that prokaryotic DNA methylation plays additional roles, performing various biological functions, including regulation of gene expression, mismatch DNA repair, and cell cycle functions[5–9]. Research interest in the diversity of prokaryotic methylation systems is therefore growing due to their importance in microbial physiology, genetics, evolution, and disease pathogenicity[7,10]. However, our knowledge of the diversity of prokaryotic methylation systems has been severely limited thus far because most studies must focus only on the rare prokaryotes that are cultivable in laboratories.

The recent development of single-molecule real-time (SMRT) sequencing technology provides us with another tool for observing DNA methylation. An array of DNA methylomes of cultivable prokaryotic strains, including N6-methyladenine (m6A), 5-methylcytosine (m5C), and N4-methylcytosine (m4C) modifications, have been revealed by this technology[11–14]. Despite its high rates of base-calling and methylation-detection errors per raw read[15,16], SMRT sequencing technology can produce ultralong reads of up to 60 kb with few context-specific biases (*e.g.*, GC bias)[17]. This characteristic enables SMRT sequencing to achieve high accuracy by merging data from many erroneous raw reads originating from clonal DNA molecules, typically from cultivated prokaryotic populations[18]. Alternatively, in an approach referred to as circular consensus sequencing (CCS), a circular DNA library is prepared as a sequence template to allow the generation of a single ultralong raw read containing multiple sequences ('subreads') that correspond to the same stretch on the template[19,20]; therefore, a cultivated clonal population is not required to achieve high accuracy[21]. However, CCS has thus far been applied in only a few shotgun metagenomics studies[22] and, to the best of our knowledge, has not yet been applied to 'metaepigenomics' or direct methylome analysis of environmental microbial communities, which are usually constituted by uncultured prokaryotes.

Here, we applied CCS to shotgun metagenomic and metaepigenomic analyses of freshwater microbial communities in Lake Biwa, the largest lake in Japan, to reveal the genomic and epigenomic characteristics of the environmental microbial communities using the PacBio Sequel platform. Freshwater habitats are rich in phage-prokaryote interactions[23–26], which are known to be closely related to prokaryotic DNA methylation. CCS analyses of the environmental microbial samples allowed reconstruction of draft genomes and the identification of their methylated motifs, at least 14 of which were novel. Furthermore, we computationally predicted and experimentally confirmed four methyltransferases (MTases) responsible for the detected methylated motifs. Importantly, two of the four MTases were revealed to recognize novel motif sequences.

Materials and methods

Sample collection

Water samples were collected at a pelagic site (35°13′09.5″N 135°59′44.7″E) in Lake Biwa, Japan (Fig. S1a) on December 26, 2016. The sampling site was located approximately 3 km from the nearest shore and had a depth of 73 m. The lake has a permanently oxygenated hypolimnion and was thermally stratified during sampling (Fig. S1b). Water sampling into prewashed 5-L Niskin bottles was conducted at depths of 5 m and 65 m, above and below the thermally stratified layer, respectively. The vertical profiles of temperature, dissolved oxygen concentrations, and chlorophyll *a* concentrations were measured using a conductivity, temperature, and depth probe *in situ*. Equipment that could come into direct contact with the water samples in the following steps was either sterilized by autoclaving or disinfected with a hypochlorous acid solution. The water samples were transferred to sterile bottles, kept cool in the dark, and immediately transported to the laboratory. Water samples with a total volume of approximately 30 L were prefiltered through 5-μm membrane PC filters (Whatman). Microbial cells were collected using 0.22-μm Sterivex filters (Millipore) and immediately stored at −20°C in a refrigerator until analysis.

DNA extraction and SMRT sequencing

The microbial DNA captured on the Sterivex filters was retrieved using a PowerSoil DNA Isolation Kit (QIAGEN) according to the supplier’s protocol with slight modifications. The filters were removed from the container, cut into 3-mm fragments, and directly suspended in the extraction solution from the kit for cell lysis. The bead-beating time was extended to 20 minutes to yield sufficient quantities of DNA for SMRT sequencing, with reference to Albertsen *et al.*[27] SMRT sequencing was conducted using a PacBio Sequel system (Pacific Biosciences) in two independent runs according to the manufacturer’s standard protocols. SMRT libraries for CCS were prepared with a 4-kb insertion length, and two SMRT cells were used for each sample as technical replicates.

Bioinformatic analysis of CCS reads

Reads that contained at least three full-pass subreads were retained to generate consensus sequences (CCS reads) using the standard PacBio SMRT software package with the default settings. Only CCS reads with >97% average base-call accuracy were retained. For taxonomic assignment of the CCS reads, Kaiju[28] in *Greedy-5* mode with the NCBI NR database[29] and Kraken[30] with the default parameters and complete prokaryotic genomes from RefSeq[31] were used. CCS reads that potentially encoded 16S ribosomal RNA (rRNA) genes were extracted using SortMeRNA[32] with the default settings, and the 16S rRNA sequences were predicted by RNAmmer[33] with the default settings. The 16S rRNA sequences were taxonomically assigned using BLASTN[34] searches against the SILVA database release 128[35], where the top-hit sequences with e-values ≤1E-15 were retrieved.

CCS reads were *de novo* assembled using Canu[18] with the *-pacbio-corrected* setting and Mira[36] with the settings for PacBio CCS reads, according to the provided instructions. After removal of the assembled contigs that were suggested to contain repeats, the contigs were binned into genomes using MetaBAT[37] based on genome coverage and tetra-nucleotide frequencies as genomic signatures, where the genome coverage was calculated by mapping the CCS reads to the binned genomes using BLASR[38] with the settings for PacBio CCS reads. The quality of all genomes was assessed using CheckM[39], which estimates completeness and contaminations based on taxonomic collocation of prokaryotic marker genes with the default settings. Sequence extraction and taxonomic assignment of 16S rRNA genes in each genome bin were conducted using RNAmmer[33] with the default settings. Taxonomic assignment of the genome bins was based on the 16S rRNA genes if found or on the taxonomic groups most frequently estimated by CAT[40] otherwise (and Kaiju[28] if CAT did not provide an estimation).

Coding sequences (CDSs) in each genome bin were predicted using Prodigal[41] with the default settings. Functional annotations were achieved through GHOSTZ[42] searches against the eggNOG[43] and Swiss-Prot[44] databases, with a cut-off e-value ≤1E-5, and HMMER[45] searches against the Pfam database[46], with a cut-off e-value ≤1E-5. A maximum-likelihood (ML) tree of the genome bins was constructed on the basis of the set of 400 conserved prokaryotic marker genes using PhyloPhlAn[47] with the default settings. Prophages were predicted using PHASTER[48] with the default settings, and their sequence alignment was conducted using LAST[49] with the default settings. CRISPR arrays were predicted using the CRISPR

Recognition Tool[50] with the default settings, and *cas* genes were annotated by querying 101 known CRISPR-associated genes in TIGRFAM[51] using HMMER[45] with a threshold of e-value ≤1E-5.

Metaepigenomic and RM system analyses

DNA methylation detection and motif analysis were performed according to BaseMod (https://github.com/ben-lerch/BaseMod-3.0). Briefly, the subreads were mapped to the assembled contigs using BLASR,[38] and interpulse duration ratios were calculated. Candidate motifs with scores higher than the default threshold value were retrieved as methylated motifs. Those with infrequent occurrences (<50) or very low methylation fractions (<1%) in each genome bin were excluded from further analysis.

Genes encoding MTases, restriction endonucleases (REases), and DNA sequence-recognition proteins were detected by BLASTP[34] searches against an experimentally confirmed gold-standard dataset from the Restriction Enzyme Database (REBASE)[52], with a cut-off e-value of ≤ 1E-15. Sequence specificity information for each hit MTase gene was also retrieved from REBASE.

Experimental verification of MTase activities

Four estimated MTase genes (EMGBS3_12600, EMGBS15_03820, EMGBS10_10070, and EMGBD2_08790) were artificially synthesized with codon optimization and cloned into the pUC57 cloning vector by Genewiz (Table S1). The genes were subcloned into the pCold III expression vector (Takara Bio) using an In-FusionHD Cloning Kit (Takara Bio). The gene-specific oligonucleotide primers used for polymerase chain reaction and recombination are described in Table S2. For verification of the EMGBS10_10070 gene function, the 5’-ACGAGTC-3’ sequence was inserted downstream of the termination codon for the sake of the methylation assay (the first five-base ACG**A**G sequence was the estimated methylated motif, and the last five-base GAGTC is recognized by the restriction enzyme PleI) (Table S1).

The constructs were transformed into *Escherichia coli* HST04 *dam*$^-$/*dcm*$^-$ (Takara Bio), which lacks endogenous MTases. The *E. coli* strains were cultured in LB broth medium supplemented with ampicillin. MTase expression was induced according to the supplier’s protocol. Plasmid DNAs were isolated using the FastGene Xpress Plasmid PLUS Kit (Nippon Genetics). SalI was employed to linearize the plasmid DNAs encoding EMGBS3_12600 and EMGBS15_03820 and then inactivated by heat. Methylation statuses were assayed by enzymatic digestion using the following restriction enzymes: BceAI and TseI for EMGBS3_12600, DpnII and XmnI for EMGBS15_03820, PleI for EMGBS10_10070, and FokI for EMGBD2_08790. All restriction enzymes were purchased from New England BioLabs. All digestion reactions were performed at 37°C for 1 h, except for those involving TseI (8 h) and FokI (20 min). Notably, although TseI digestion is conducted at 65°C in the manufacturer's protocol, we adopted a temperature of 37°C to avoid cleavage of methylated DNA.

We further verified the methylated motifs that were newly estimated in this study, *i.e.*, those of EMGBS10_10070 and EMGBD2_08790. Chromosomal DNA was extracted from cultures of the transformed *E. coli* strains using a PowerSoil DNA Isolation Kit (QIAGEN) according to the supplier's protocol. SMRT sequencing was conducted using PacBio RSII (Pacific Biosciences), and methylated motifs were detected via the same method described above.

Data deposition

The raw sequencing data and assembled genomes were deposited in the DDBJ Sequence Read Archive and DDBJ/ENA/GenBank, respectively (Table S3). All data were registered under BioProject ID PRJDB6656.

Results and discussion

Water sampling, SMRT sequencing, and circular consensus analysis

Water samples were collected at a pelagic site in Lake Biwa, Japan, at 5 m (biwa_5m) and 65 m depths (biwa_65m), from which PacBio Sequel produced a total of 2.6 million (9.6 Gbp) and 2.0 million (6.4 Gbp) subreads, respectively (Table 1). The circular consensus analysis produced 168,599 and 117,802 CCS reads, with lengths of 4,474 ± 931 and 4,394 ± 587 bp, respectively (Table 1 and Fig. S2). In the shallow sample data, at least 90% of the CCS reads showed high quality (Phred quality scores >20) at each base position, except for the 5′-terminal five bases and 3′-terminal bases after the 5,638th base. In the deep sample data, the same was true, except for the 5′-terminal four bases and 3′-terminal bases after the 5,356th base (Fig. S3).

Taxonomic analysis

Taxonomic assignment of the CCS reads was performed using Kaiju[28] and the NCBI NR database[29] (Fig. 1). The assignment ratios were >88% and >56% at the phylum and genus levels, respectively, which were higher than those for the Illumina-based shotgun metagenomic analysis of lake freshwater and other environments using the same computational method[28]. Kraken[30] with complete prokaryotic and viral genomes in RefSeq[31] (Fig. S4a-c) provided similar results but resulted in much lower assignment ratios (30% and 27%, respectively), likely due to the lack of genomic data for freshwater microbes in RefSeq. 16S rRNA sequence-based taxonomic assignment via BLASTN searches against the SILVA database[53] also provided consistent results (Fig. S4d-f). It should be noted that 16S rRNA-based and CDS-based taxonomic assignments can be affected by 16S rRNA gene copy numbers and genome sizes, respectively.

At the phylum level, Proteobacteria dominated both samples, followed by Actinobacteria, Verrucomicrobia, and Bacteroidetes (Fig. 1). Chloroflexi and Thaumarchaeota were especially abundant in the deep water sample, consistent with previous findings[54,55]. The ratio of Archaea was particularly low in the shallow sample (0.6 and 6.9% in biwa_5m and biwa_65m, respectively). Although the filter pore-size range (5–0.2 μm) was not suitable for most viruses and eukaryotic cells, non-negligible ratios corresponding to their existence were observed in the shallow sample. The dominant eukaryotic phylum was Opisthokonta (2.68 and 0.92%), followed by Alveolata (1.67 and 0.45%) and Stramenopiles (1.45 and 0.15%). Among viruses, Caudovirales and Phycodnaviridae were the most abundant families in both samples. Caudovirales are known to act as bacteriophages, while Phycodnaviridae primarily infect eukaryotic algae. The third most abundant viral family was Mimiviridae, whose members are also known as 'Megavirales' due to their large genome size (0.6–1.3 Mbp)[56,57]. Viruses without double-stranded DNA (*i.e.*, single-stranded DNA and RNA viruses) were not observed because of the experimental method employed. Overall, the taxonomic composition was consistent with those obtained in previous studies on microbial communities in freshwater lake environments, reflecting the fact that SMRT sequencing provides taxonomic compositions consistent with those obtained using short-read technologies, such as the Illumina MiSeq and HiSeq platforms[58,59].

Metagenomic assembly and genome binning

The CCS reads from the shallow and deep samples were assembled into 554 and 345 contigs, respectively, using Canu[18] (Table S4). The corresponding N50 values were 83 and 76 kbp, and the longest contigs had lengths of 481 and 740 kbp, respectively. Notably, the contigs were much longer than those obtained in a previous study that applied CCS for shotgun metagenomics analysis of an active sludge microbial community[22]. We also used Mira[36] for metagenomic assembly, but this resulted in shorter longest contigs (148 and 151 kbp, respectively) and N50 values (19 and 18 kbp, respectively).

The contigs were binned to genomes using MetaBAT[37], which is a reference-independent binning tool, based on CCS-read coverage and tetranucleotide frequency (Fig. 2 and Table 2). Among a total of 899 contigs, 390 (43.3%) were assigned to fifteen and four bins from the shallow and deep samples, respectively. We obtained a draft genome for each bin, where the completeness of the genome ranged from 17–99% (67% on average). Estimated contamination levels were low (<3% in each bin). Based on the total contig size and estimated genome completeness of each bin, the genome sizes were estimated to range from 1.0–5.6 Mbp. The GC content ranged from 29–68%, and the average N50 was 24 kbp, with a maximum of 1.67 Mbp.

The nineteen genome bins belonged to seven phyla (Table 2 and Fig. S5). Among these genome bins, ten contained 16S rRNA genes, and many of them showed top hits to uncultured clades; thus, our CCS-based approach was estimated to have truly targeted multiple uncultured prokaryotes. Seven genome bins were predicted to belong to the phylum Actinobacteria, including *Candidatus* Planktophila (BS7), one of the most dominant bacterioplankton lineages in freshwater systems[60,61]. Metagenomic bins affiliated with other dominant freshwater lineages were also recovered, including *Candidatus* Methylopumilus (BS12)[62], the freshwater lineage (LD12) of Pelagibacterales (BS14)[63,64], and Nitrospirae (BD2) and *Candidatus* Nitrosoarchaeum (BD3), the predominant nitrifying bacteria and archaea in the hypolimnion, respectively[54,55]. Four bins were affiliated with the phylum Verrucomicrobia (BS6, BS8, BS10, and BD4), in line with a previous study[65]. The BS3 and BD1 genome bins likely represent members of the CL500-11 group (class Anaerolineae) of the Chloroflexi phylum, where BD1 presented the highest coverage of >45×. This group is a dominant group in the hypolimnion of Lake Biwa and is frequently found in deep oligotrophic freshwater environments worldwide[66]. Overall, the phylogeny of the reconstructed genomes likely reflects the major dominant lineages present in the water of Lake Biwa.

Metaepigenomic analysis

A total of 29 methylated motifs were detected in ten genome bins (Table 3). Their methylation ratios ranged from 19–99%, which can be affected by modification detection power, *i.e.*, these ratios are likely lower than the true methylation levels. Three motifs from the BS12 genome bin contained overlapping sequences (HCAG**C**TKC, BGMAG**C**TGD, and GMAG**C**TKC, where B: G/T/C, D: G/A/T, H: A/C/T, K: G/T, and M: A/C, where the underlined bold face indicates methylation sites) that were likely due to incomplete detection of a single methylated motif or heterogeneous motif sequences between closely related lineages contained within that genome bin. A palindromic motif and five complementary motif pairs that likely reflect double-strand methylation were observed in the BS15 bin (*e.g.*, a pair of **A**GCNNNNNNCAT and **A**TGNNNNNNGCT). It may also be notable that three genome bins from the Chloroflexi phylum (BS1, BS3, and BD1) shared the same motif sequence set (G**A**NTC, TTA**A**, and G**C**WGC, where W: A/T), likely due to evolutionarily shared methylation systems.

Overall, even if such overlapping, complementary, and shared motif sequences are considered, at least 14 motifs still presented no match to existing recognition sequences in the REBASE repository. This result demonstrates the existence of unexplored diversity of DNA methylation systems in environmental prokaryotes, which include many uncultured strains.

Known MTases that correspond to detected methylated motifs

To identify MTases that can catalyze the methylation reactions of the detected methylated motifs, systematic annotation of MTase genes was performed. Sequence similarity searches against known genes identified 20 MTase genes in nine genome bins (sequence identities ranged from 23–71%) (Table 4). The most abundant group was Type II MTases, followed by Type I and Type III MTases, a trend that is consistent with the general MTase distribution[13,67]. Several genes encoding REases and DNA sequence-recognition proteins were also detected (Table 4). The known motifs of seven of the 20 MTases were matched to those identified in our metaepigenomic analysis (Table 3). For example, the genome bin BD3 contained two MTases, whose recognition motif sequences were AG**C**T and G**A**TC according to the sequence homology-based prediction, which were perfectly congruent with the two motifs detected in our metaepigenomic analysis. It may be notable that these two motifs were also reported in an enrichment-culture study of the closely related genus *Candidatus* Nitrosomarinus catalina[68] and are therefore likely evolutionarily conserved within their group. In the BS14 bin, a similar one-to-one perfect match was also observed. The two Chloroflexi genome bins BS3 and BD1 were characterized by the same set of three methylated motifs, each of which contained three MTases. No MTase gene was found in the other Chloroflexi bin BS1, likely due to its low estimated genome completeness of 31% (Table 2). Among these MTases, two were predicted to show methylation specificities that were congruent with two of the detected motifs, G**A**NTC and TTA**A** (the other MTase and motif will be discussed in the next section). Collectively, these observations suggest that metaepigenomic analysis is an effective tool for identifying the methylation systems of environmental prokaryotes.

Unexplored diversity of prokaryotic methylation systems

Among the 20 detected MTases, 13 MTases did not present known recognition motifs that matched those identified in our metaepigenomic analysis (Tables 3 and 4). Although homology search-based MTase identification and recognition motif estimation are frequently conducted in genomic and metagenomic studies, this result suggests that these approaches are not sufficient, and direct observation of DNA methylation is needed to reveal the methylation systems of diverse environmental prokaryotes.

As noted earlier, each of the BS3 and BD1 bins had three MTase genes, two of which were congruent to two of the detected motifs. The other MTase from each bin (EMGBS3_12600 and EMGBD1_09320 in BS3 and BD1, respectively) showed the highest sequence similarity to an MTase that was reported to recognize A**C**GGC; however, the other methylated motif detected in the BS3 and BD1 bins was G**C**WGC.

In the BS15 genome bin, six MTases and eleven methylated motifs were detected, but none of the MTases and motifs matched each other. At the methylation type level, five MTases and all of the methylated motifs were of the m6A type. We predicted that the EMGBS15_03820 MTase, which is estimated to exhibit non-specific m6A methylation activity, is actually a sequence-specific enzyme that recognizes a G**A**ANNNNTTC motif that was detected through metaepigenomic analysis, because the adjacent gene EMGBS15_03830 encodes an REase that targets the same GAANNNNTTC sequence.

In the BS8 genome bin, one MTase and one methylated motif were detected; however, the estimated motif of this MTase was incongruent with the detected motif (the estimated and detected motifs were ACG**A**NNNNNNGRTC and **A**GGNNNNNRTTT, respectively, where R: G/A). This MTase is predicted to function in an RM system because of the existence of the neighboring REase and DNA-sequence recognition protein genes.

In the BS10 genome bin, one MTase and one methylated motif were detected, and their motifs were also incongruent (GCA**A**GG and ACG**A**G, respectively).

In the BD2 genome bin, two MTases and one methylated motif were detected. The two MTases were predicted to display m6A and m5C methylation activities, while the detected motif contained an m6A site. Thus, the former MTase was predicted to catalyze the methylation reaction, although their motifs were again incongruent (GRGGA**A**G and TANGG**A**B, respectively). It should also be noted that these MTases appear to constitute a recently proposed system known as the Defense Island System Associated with Restriction-Modification (DISARM), which is a phage-infection defense system composed of MTase, helicase, phospholipase D, and DUF1998 genes[69]. To our knowledge, this is the first DISARM system identified in the phylum Nitrospirae.

In the BS6 genome bin, one MTase gene was found, but we could not detect any methylated motif, and we therefore anticipate that this MTase gene does not exhibit methylation activity or the corresponding methylation motif was undetected due to the low sensitivity of SMTR sequencing to m5C modification as described previously [13,14]. However, in the BS12 genome bin, we detected methylated motifs but no MTase genes. We assume that the MTase genes corresponding to this bin were missed due to insufficient genome completeness (although the estimated completeness was 81%), or because these MTase genes have diverged considerably from MTase genes found in cultivable strains, or because thee MTases belong to a new group.

Experimental verification of MTases with new methylated motifs

Among the MTases whose estimated methylated motifs were not congruent with our metaepigenomic results, we experimentally verified the methylation specificities of the four MTases: EMGBS3_12600 in BS3 (and EMGBD1_09320 in BD1, which has exactly the same amino acid sequence), EMGBS15_03820 in BS15, EMGBS10_10070 in BS10, and EMGBD2_08790 in BD2 (Table 4). We constructed plasmids that each carried one of the artificially synthesized MTase genes, which we then transformed *E. coli* cells that lacked endogenous MTases, forced their expression, and observed the methylation status of the isolated plasmid DNA by REase digestion.

Although the estimated methylated motif sequence of EMGBS3_12600 was A**C**GGC, the unaccounted-for motif sequence observed in BS3 was G**C**WGC. Thus, we hypothesized that the true recognition sequence of EMGBS3_12600 is G**C**WGC. The REase digestion assay showed that TseI (GCWGC specificity) did not cleave the plasmids when EMGBS3_12600 was expressed in the cells, which clearly supports our hypothesis (Fig. 3a). Furthermore, we confirmed that BceAI (ACGGC specificity) cleaved plasmids regardless of whether EMGBS3_12600 was expressed, indicating that the EMGBS3_12600 protein does not show ACGGC sequence specificity (Fig. 3a). Accordingly, we named this protein M.AspBS3I, as a novel MTase that possesses G**C**WGC specificity (Table 4).

While the homology-based analysis predicted EMGBS15_03820 as a non-sequence specific MTase, its adjacency to an REase and the results of the metaepigenomic analysis suggested that this MTase presents G**A**ANNNNTTC sequence specificity. The REase digestion assay showed that XmnI (GAANNNNTTC specificity) did not cleave the plasmids only when EMGBS15_03820 was expressed in the cells, which also supports our hypothesis (Fig. 3b). Furthermore, we confirmed that DpnII (GATC specificity) cleaved the plasmids regardless of whether EMGBS15_03820 was expressed, indicating that EMGBS15_03820 is not a nonspecific MTase. We named this protein M.FspBS15I, as a novel MTase that possesses G**A**ANNNNTTC methylation specificity (Table 4).

For EMGBS10_10070 in BS10 and EMGBD2_08790 in BD2, we also conducted REase digestion assays to confirm the recognition motif sequences. Based on the results of the metaepigenomic analysis, their motifs were predicted to be ACG**A**G and TANGG**A**B, respectively. Expression of each gene altered the electrophoresis patterns of the digested plasmids to contain fragments that resulted from inhibition of REase cleavage at the estimated methylation sites (Fig. S6). Furthermore, we additionally conducted SMRT sequencing analysis using the PacBio RSII platform to examine the methylation status of the chromosomal DNA of the *E. coli* transformed with each of the two MTase genes. The results were basically consistent (Table S5): ACG**A**G was actually detected as the methylated motif in *E. coli* transformed with EMGBS10_10070, and we named the protein M.OspBS10I. In the case of EMGBD2_08790, the detected TAHGG**A**B motif was almost the same, but a subset of the estimated TANGG**A**B motif (*i.e.*, TAGGG**A**B was excluded), and this difference could be due to *E. coli*-specific conditions (*e.g.*, cofactors and sequence biases), insufficient data, or inaccuracy of the methylated motif detection method. Regardless of this minor difference, we concluded that EMGBD2_08790 is a novel MTase gene responsible for methylation of the TAHGG**A**B motif and we named the protein M.NspBD2I accordingly.

Genome bins that lack methylation systems and phage infection

Among the nineteen genome bins, no methylated motifs were detected in nine genome bins (MTase genes were also not detected, except in the BS6 genome bin). This high ratio of methylation-lacking organisms contrasts remarkably with a previous report in which prokaryotic genomes were found to rarely lack DNA methylation systems (<7%)[13]. Notably, those nine genome bins contained seven Actinobacteria bins, indicating that the dominant Actinobacteria in Lake Biwa lack methylation systems, although a number of methylated motifs and corresponding MTases have been reported in Actinobacteria[13].

Because DNA methylation is known to play a role in opposing phage infection[2–4], we conducted *in silico* prophage detection to evaluate whether prokaryotes in Lake Biwa tend to be infected by phages. Within the nineteen genome bins, more than one prophage was found in ten genome bins (Table 2 and S6). Among these ten bins, six overlapped with the nine genome bins in which no methylated motifs were identified. The prophages showed little sequence similarity to each other except for two pairs and likely resulted from independent and repetitive infections (Fig. S7). Thus, phage infection and prophage integration appear to frequently occur in prokaryotes that lack DNA methylation systems. We also investigated the presence of CRISPR/Cas systems as another major prokaryotic mechanism against phage infections[70–73]. We identified possible CRISPR arrays in three genome bins, BS3, BS8, and BD3, which exhibit methylation systems but no prophages, although the first two genome bins contained no associated *Cas* genes.

Based on these results, we assume that the possession of prophages is tolerable in lake freshwater environments, and thus, the evolutionary pressure to develop or retain methylation systems is low. These results also suggest that uncultured and cultivable strains may be under different selection pressures regarding DNA methylation systems, and the true diversity of microbial methylation systems must be examined in the future using metaepigenomic approaches.

Conclusion

The present study demonstrated the effectiveness of the metaepigenomic approach powered by SMRT sequencing and CCS, showing obvious advantages over sequence similarity-based and culture-based methylation system analyses and short-read metagenomics. The CCS reads facilitated metagenomic assembly, binning, and protein sequence-based taxonomic assignment from an environmental sample that contained dominant uncultured prokaryotes. Most importantly, this approach revealed several methylated motifs, including novel ones in environmental prokaryotes, and subsequent experiments identified four MTases responsible for those reactions. The anti-correlation pattern between the presence of prophages and methylation was consistent with past observations that methylation systems inhibit phage infection and phage-mediated genetic exchange, although the underlying ecological background and mechanisms must be examined in the future.

The current throughput of SMRT sequencing may be still insufficient to apply the metaepigenomic approach to more diverse and complex samples. Because deep sequencing coverage (>25× subreads for each DNA strand) is required for the reliable detection of DNA methylation, it is still difficult to obtain sufficient sequencing reads to recover long contigs and detect methylated motifs for ‘rare’ species (typically those with <1% relative abundance). In addition to rapid and ongoing technological advances in SMRT sequencing, the emergence of Oxford Nanopore Technology may provide as another long-read, single-molecule, and methylation-detectable technology[74,75]. Another problem is that the detectable types of DNA modifications are limited (*i.e.*, m4C, m5C, and m6A) with the currently available SMRT sequencing technology, while many other DNA chemical modifications occur in nature[76]. In addition to advances in sequencing methods, novel bioinformatic tools will be critical for metaepigenomic analyses of environmental prokaryotes.

A recent study showed that sets of methylated motifs and MTases can vary widely, even between closely related strains[77], where metaepigenomics is expected to enable differential methylation analyses between populations. In addition, genus-level conservation of MTases that are not associated with REases is sometimes observed, which suggests that MTases play unexplored adaptive roles, in addition to their functions in combating phages[13,78]. Novel MTases may be adopted for biotechnological uses, such as DNA recombination and methylation analyses[79]. It is envisioned that metaepigenomics of environmental prokaryotes under different sampling conditions and environments will significantly deepen our understanding of the enigmatic evolution of prokaryotic methylation systems and broaden their application potential.

Author Contributions

SH conceived the study, performed the bioinformatics analyses and experiments, and wrote the manuscript. YO and SN performed the water sampling. AM performed the experiments. AT performed the genomic and metagenomic sequencing. WI conceived the study, wrote the manuscript, and supervised the project. All authors read and approved the final manuscript.

Funding

This work was supported by the Japan Science and Technology Agency (CREST), the Japan Society for the Promotion of Science (Grant Numbers 15J00971, 15J08604, 15H01725, 16H06154, and 17H05834), the Ministry of Education, Culture, Sports, Science, and Technology in Japan (221S0002 and 16H06279), and Leave a Nest Grant.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

The SMRT sequencing was supported by National Institute of Genetic, Research organization of information and systems, Mishima, Japan. We thank Yoshinori Nii, Masashi Yoshino, and Satoko Fukuda for their helpful suggestions and experimental supports. We are grateful to Yukiko Goda and Tetsuji Akatsuka for their assistance in the field sampling. We also thank Metabologenomics, Inc. for financial support.

References

1. Kumar, R. & Rao, D. N. Role of DNA Methyltransferases in Epigenetic Regulation in Bacteria. in *Subcellular Biochemistry* (ed. Kundu, T. K.) **61,** 81–102 (Springer Netherlands, 2013).
2. Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **8,** 317 (2010).
3. Kobayashi, I. Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.* **29,** 3742–3756 (2001).
4. Makarova, K. S., Wolf, Y. I., Snir, S. & Koonin, E. V. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.* **193,** 6039–6056 (2011).

5. Wion, D. & Casadesús, J. N^6-methyl-adenine: An epigenetic signal for DNA–protein interactions. *Nat. Rev. Microbiol.* **4,** 183–192 (2006).

6. Low, D. A. & Casadesús, J. Clocks and switches: Bacterial gene regulation by DNA adenine methylation. *Curr. Opin. Microbiol.* **11,** 106–112 (2008).

7. Casadesus, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* **70,** 830–856 (2006).

8. Vasu, K. & Nagaraja, V. Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.* **77,** 53–72 (2013).

9. Kozdon, J. B. *et al.* Global methylation state at base-pair resolution of the *Caulobacter* genome throughout the cell cycle. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E4658–E4667 (2013).

10. Srikhanta, Y. N., Fox, K. L. & Jennings, M. P. The phasevarion: Phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. *Nat. Rev. Microbiol.* **8,** 196 (2010).

11. Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7,** 461–465 (2010).

12. Clark, T. A. *et al.* Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.* **40,** e29 (2012).

13. Blow, M. J. *et al.* The Epigenomic Landscape of Prokaryotes. *PLoS Genet.* **12,** e1005854 (2016).

14. Murray, I. A. *et al.* The methylomes of six bacteria. *Nucleic Acids Res.* **40,** 11450–11462 (2012).

15. Vinet, L. & Zhedanov, A. A 'missing' family of classical orthogonal polynomials. *Science (80-.).* **323,** 133–138 (2010).

16. Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30,** 693–700 (2012).

17. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics. Proteomics Bioinformatics* **13,** 278–289 (2015).

18. Koren, S. *et al.* Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27,** 722–736 (2017).

19. Fichot, E. B. & Norman, R. S. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome* **1,** 10 (2013).

20. Gao, S. *et al.* PacBio full-length transcriptome profiling of insect mitochondrial gene expression. *RNA Biol.* **13,** 820–825 (2016).

21. Hiraoka, S., Yang, C. & Iwasaki, W. Metagenomics and bioinformatics in microbial ecology: Current status and beyond. *Microbes Environ.* **31,** 204–212 (2016).

22. Frank, J. A. *et al.* Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci. Rep.* **6,** 25373 (2016).

23. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536,** 425 (2016).

24. Moon, K., Kang, I., Kim, S., Kim, S.-J. & Cho, J.-C. Genomic and ecological study of two distinctive freshwater bacteriophages infecting a Comamonadaceae bacterium. *Sci. Rep.* **8,** 7989 (2018).

25. Moon, K., Kang, I., Kim, S., Kim, S.-J. & Cho, J.-C. Genome characteristics and environmental distribution of the first phage that infects the LD28 clade, a freshwater methylotrophic bacterial group. *Environ. Microbiol.* **19,** 4714–4727 (2017).

26. Ghai, R., Mehrshad, M., Megumi Mizuno, C. & Rodriguez-Valera, F. Metagenomic recovery of phage genomes of uncultured freshwater actinobacteria. *ISME J.* **11,** 304–308 (2017).

27. Albertsen, M., Karst, S. M., Ziegler, A. S., Kirkegaard, R. H. & Nielsen, P. H. Back to basics—The influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. *PLoS One* **10,** e0132783 (2015).

28. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7,** 11257 (2016).

29. Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **45,** D12–D17 (2017).

30. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).

31. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. RefSeq microbial genomes database: New representation and annotation strategy. *Nucleic Acids Res.* **42,** D553–D559 (2014).

32. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28,** 3211–3217 (2012).

33. Lagesen, K. *et al.* RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35,** 3100–3108 (2007).

34. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10,** 1–9 (2009).

35. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41,** D590–D596 (2013).

36. Chevreux B & Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. in *German conference on bioinformatics* **99,** 45–56 (Hanover, Germany, 1999).

37. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3,** e1165 (2015).

38. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. *BMC Bioinformatics* **13,** 238 (2012).

39. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (2015).

40. Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Contig annotation tool CAT robustly classifies assembled metagenomic contigs and long sequences. *bioRxiv* 072868 (2016). doi:10.1101/072868

41. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).

42. Suzuki, S., Kakuta, M., Ishida, T. & Akiyama, Y. Faster sequence homology searches by clustering subsequences. *Bioinformatics* **31,** 1183–1190 (2015).

43. Powell, S. *et al.* EggNOG v4.0: Nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42,** D231–D239 (2014).

44. UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* **41,** D43–D47 (2013).

45. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41,** e121 (2013).

46. Finn, R. D. *et al.* The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–D285 (2016).

47. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4,** 2304 (2013).

48. Arndt, D. *et al.* PHASTER: A better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44,** W16–W21 (2016).

49. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21,** 487–493 (2011).

50. Bland, C. *et al.* CRISPR recognition tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8,** 209 (2007).

51. Haft, D. H. *et al.* TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.* **41,** D387–D395 (2013).

52. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE—a database for DNA restriction and modification: Enzymes, genes and genomes. *Nucleic Acids Res.* **38,** D234–D236 (2010).

53. Yilmaz, P. *et al.* The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **42,** D643-8 (2014).

54. Okazaki, Y. & Nakano, S.-I. Vertical partitioning of freshwater bacterioplankton community in a deep mesotrophic lake with a fully oxygenated hypolimnion (Lake Biwa, Japan). *Environ. Microbiol. Rep.* **8,** 780–788 (2016).

55. Okazaki, Y. *et al.* Ubiquity and quantitative significance of bacterioplankton lineages inhabiting the oxygenated hypolimnion of deep freshwater lakes. *ISME J.* **11,** 2279–2293 (2017).

56. Colson, P. *et al.* "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Arch. Virol.* **158,** 2517–2521 (2013).

57. Claverie, J.-M. *et al.* Mimivirus and Mimiviridae: Giant viruses with an increasing number of potential hosts, including corals and sponges. *J. Invertebr. Pathol.* **101,** 172–180 (2009).

58. Tsai, Y.-C. *et al.* Resolving the complexity of human skin metagenomes using single-molecule sequencing. *MBio* **7,** e01948-15 (2016).

59. Singer, E. *et al.* Next generation sequencing data of a defined microbial mock community. *Sci. Data* **3,** 160081 (2016).

60. Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A guide to the natural history of freshwater lake bacteria. *Microbiol. Mol. Biol. Rev.* **75,** 14–49 (2011).

61. Neuenschwander, S. M., Ghai, R., Pernthaler, J. & Salcher, M. M. Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. *ISME J.* **12,** 185–198 (2018).

62. Salcher, M. M., Neuenschwander, S. M., Posch, T. & Pernthaler, J. The ecology of pelagic freshwater methylotrophs assessed by a high-resolution monitoring and isolation campaign. *ISME J.* **9,** 2442 (2015).

63. Salcher, M. M., Pernthaler, J. & Posch, T. Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria that rule the waves (LD12). *ISME J.* **5,** 1242–1252 (2011).

64. Henson, M. W., Lanclos, V. C., Faircloth, B. C. & Thrash, J. C. Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. *ISME J.* **12,** 1846–1860 (2018).

65. Cabello-Yeves, P. J. *et al.* Reconstruction of diverse verrucomicrobial genomes from metagenome datasets of freshwater reservoirs. *Frontiers in Microbiology* **8,** 2131 (2017).

66. Okazaki, Y., Hodoki, Y. & Nakano, S. Seasonal dominance of CL500-11 bacterioplankton (phylum Chloroflexi) in the oxygenated hypolimnion of Lake Biwa, Japan. *FEMS Microbiol. Ecol.* **83,** 82–92 (2013).

67. Oliveira, P. H., Touchon, M. & Rocha, E. P. C. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42,** 10618–10631 (2014).

68. Ahlgren, N. A. *et al.* Genome and epigenome of a novel marine Thaumarchaeota strain suggest viral infection, phosphorothioation DNA modification and multiple restriction systems. *Environ. Microbiol.* **19,** 2434–2452 (2017).

69. Ofir, G. *et al.* DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.* **3,** 90–98 (2018).

70. Rath, D., Amlinger, L., Rath, A. & Lundgren, M. The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie* **117,** 119–128 (2015).

71. Jore, M. M., Brouns, S. J. J. & van der Oost, J. RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. *Cold Spring Harb. Perspect. Biol.* **4,** a003657 (2012).

72. Seed, K. D. Battling phages: How bacteria defend against viral attack. *PLoS Pathog.* **11,** e1004847 (2015).

73. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **13,** 722 (2015).

74. Rand, A. C. *et al.* Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14,** 411–413 (2017).

75. Stoiber, M. H. *et al. De novo* identification of DNA modifications enabled by genome-guided nanopore signal processing. *bioRxiv* 094672 (2016). doi:10.1101/094672

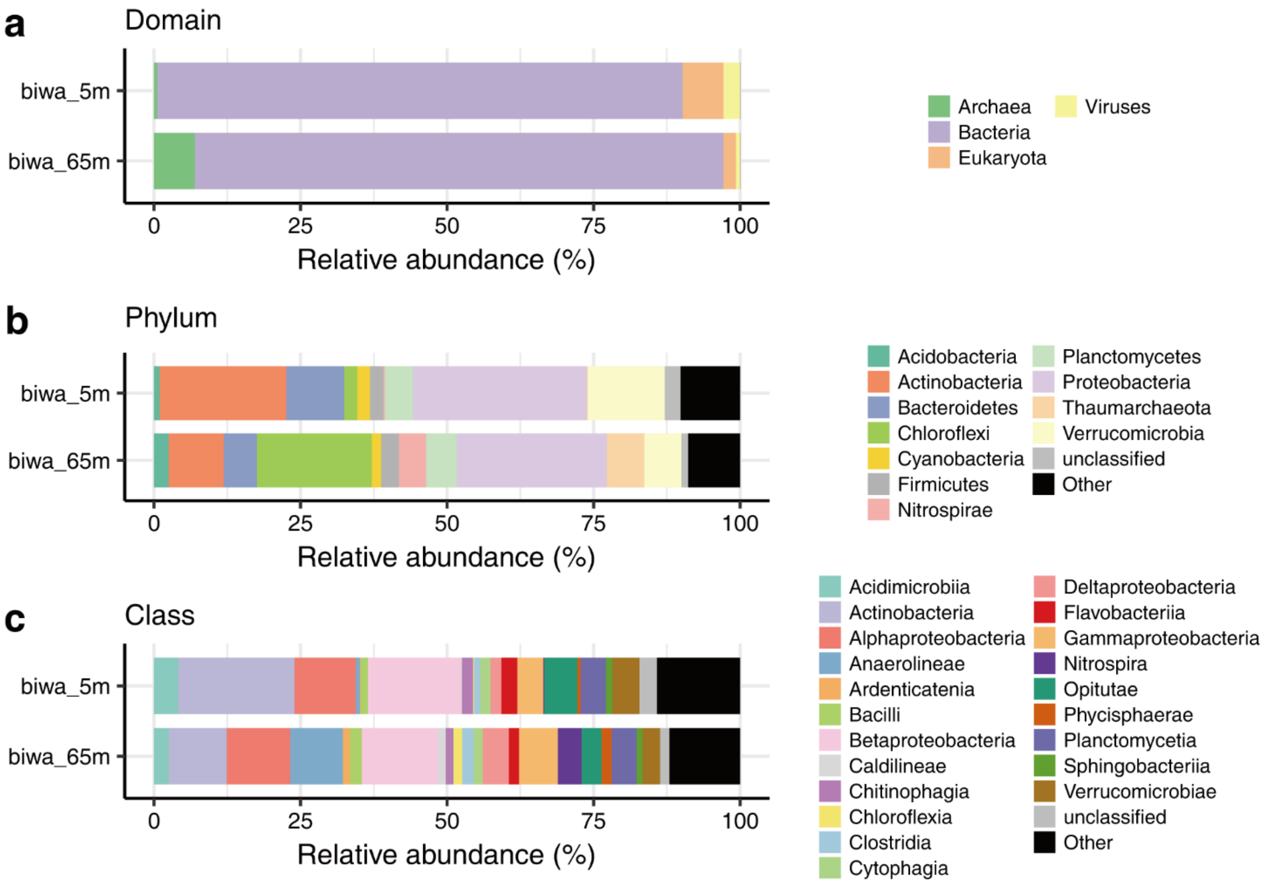
76. Davis, B. M., Chao, M. C. & Waldor, M. K. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr. Opin. Microbiol.* **16,** 192–198 (2013).

77. Kojima, K. K. *et al.* Population evolution of *Helicobacter pylori* through diversification in DNA methylation and interstrain sequence homogenization. *Mol. Biol. Evol.* **33,** 2848–2859 (2016).

78. Seshasayee, A. S. N., Singh, P. & Krishna, S. Context-dependent conservation of DNA methyltransferases in bacteria. *Nucleic Acids Res.* **40,** 7066–7073 (2012).

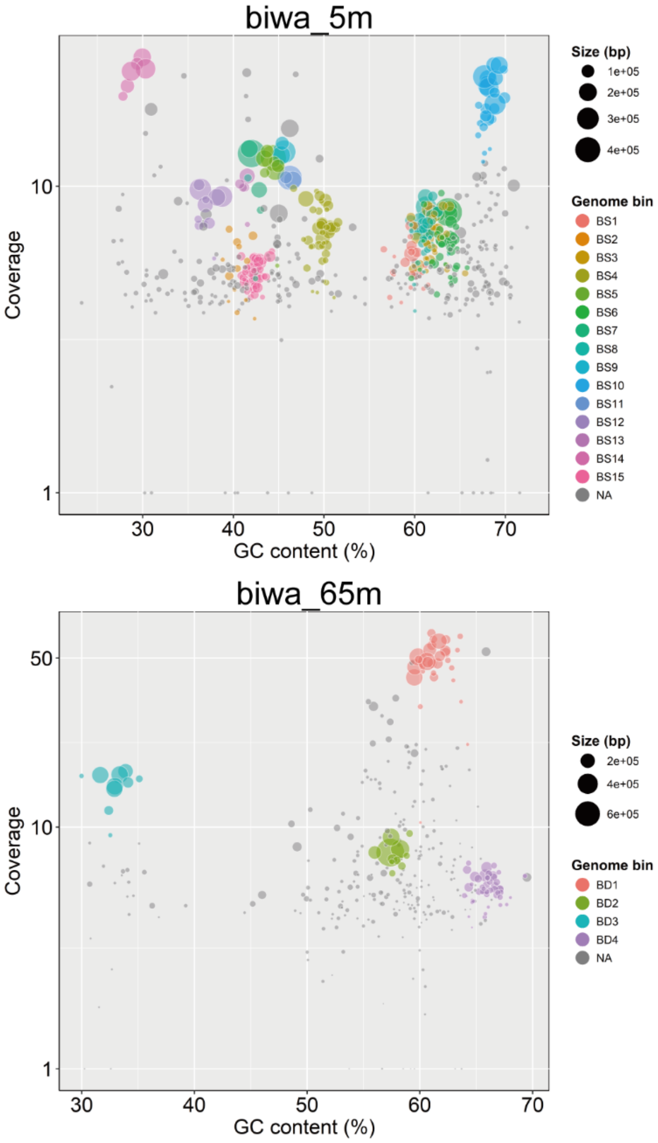
79. Buryanov, Y. & Shevchuk, T. The use of prokaryotic DNA methyltransferases as experimental and analytical tools in modern biology. *Anal. Biochem.* **338,** 1–11 (2005).

589 **Figures**



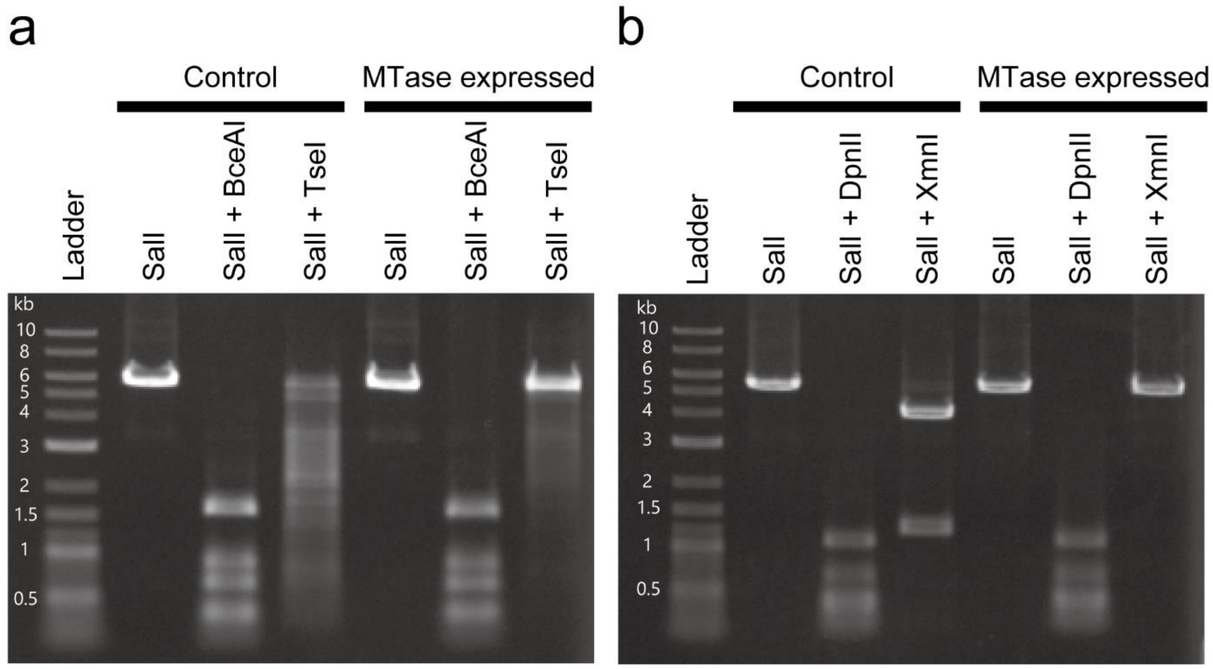
590

591 **Figure 1.** Phylogenetic distribution of CCS reads. Estimated relative abundances at the (**a**) domain, (**b**)
592 phylum, and (**c**) class levels are shown. Eukaryotic and viral reads are ignored, and groups with <1%
593 abundance are grouped as ‘Others’ in **b** and **c**.



594

595 **Figure 2.** Genome binning of the assembled contigs. Each circle represents a contig, where the color and size
596 represent its assigned bin and total sequence length, respectively. Contigs not assigned to any bin are indicated
597 in gray (named 'NA'). The x-axis and y-axis represent GC% and genome coverage, respectively.



598

599 **Figure 3.** REase digestion assays. **a** Assay of the EMGBS3_12600 gene (and EMGBD1_09320, which has 600 the same amino-acid sequence). BceAI and TseI were used, where the plasmid contained 12 (ACGGC) and 21 601 (GCWGC) target sites, respectively. Plasmid DNAs were linearized using SalI before the assay. An NEB 602 2-log DNA ladder was employed as a size marker. **b** Assay of the EMGBS15_03820 gene. DpnII and XmnI 603 were used, where the plasmid contained 27 (GATC) and two (GAANNNNTTC) target sites, respectively.

604

Tables

Table 1. Statistics of SMRT sequencing and CCS-read analysis.

Sample	biwa_5m	biwa_65m
Sequenced reads	850,494	688,436
Total base pairs (bp)	9,570,723,004	6,419,717,083
CCS reads	168,599	117,802
Read length (bp)	4,474 ± 931	4,394 ± 587
Total base (bp)	754,416,328	517,663,806
16S rRNA	170	106
Length (bp)	1,491 ± 64	1,468 ± 104

Table 2. Statistics for genome bins.

Genome bin	Lineage	Estimated genome size (Mb)	Contigs	N50 (bp)	GC content (%)	Completeness (%)	Contamination (%)	16S rRNA	CDSs	Coverage	Methylated motifs	MTases	Prophage
BS1	Bacteria; Chloroflexi[1]	2.24	21	64,528	59.5	30.6	0.0	0	751	5.79	3	0	0
BS2	Bacteria; Actinobacteria[1]	1.57	13	28,617	40.6	16.9	0.0	0	363	5.13	0	0	1
BS3	Bacteria; Chloroflexi; Anaerolineae; Anaerolineales; Anaerolineaceae; uncultured; uncultured Crater Lake bacterium CL500-11	3.35	36	58,996	61.8	49.1	0.0	1	1,646	6.91	3	3	0
BS4	Bacteria; Actinobacteria; Acidimicrobiia; Acidimicrobiales; Acidimicrobiaceae; CL500-29 marine group	2.31	40	61,750	49.8	76.8	1.3	1	2,066	6.67	0	0	2
BS5	Bacteria; Actinobacteria; Actinobacteria; Frankiales; Sporichthyaceae; hgcI clade; uncultured *Clavibacter* sp.	1.51	8	190,417	44.2	71.6	0.0	1	1,209	10.02	0	0	2
BS6	Bacteria; Verrucomicrobia; Opitutae; Opitutae vadinHA64; uncultured bacterium	2.27	37	100,045	63.4	89.2	0.7	1	1889	6.85	0	1	1
BS7	Bacteria; Actinobacteria; Actinobacteria; Frankiales; Sporichthyaceae; hgcI clade; uncultured *Candidatus* Planktophila sp.	1.49	6	470,028	42.1	58.4	0.6	1	948	9.26	0	0	0
BS8	Bacteria; Verrucomicrobia[2]	2.71	34	102,020	61.2	82.5	2.0	0	2,121	7.34	1	1	0
BS9	Bacteria; Actinobacteria[2]	1.65	3	315,861	45.5	37.6	0.0	0	677	12.09	0	0	3
BS10	Bacteria; Verrucomicrobia; Opitutae; Opitutae vadinHA64; uncultured bacterium	2.55	24	1,672,582	68.4	95.9	2.7	1	2,165	17.93	1	1	2
BS11	Bacteria; Actinobacteria; Actinobacteria; Frankiales; Sporichthyaceae; hgcI clade; uncultured actinobacterium	1.03	3	365,154	46.3	62.1	0.0	1	675	10.28	0	0	1
BS12	Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; *Candidatus* Methylopumilus; uncultured bacterium	1.40	10	169,468	37.3	80.7	0.4	1	1,289	8.37	3	0	1
BS13	Bacteria; Actinobacteria; Actinobacteria[1]	1.49	5	47,968	41.3	19.0	0.0	0	351	7.56	0	0	0
BS14	Proteobacteria; Alphaproteobacteria; Pelagibacterales[1]	1.02	6	222,441	29.4	88.6	0.0	0	1,075	20.45	1	1	1
BS15	Bacteria; Bacteroidetes; Sphingobacteriia; Sphingobacteriales; Chitinophagaceae; Filimonas; uncultured bacterium	4.08	44	45,979	42.4	43.1	0.1	1	1,908	5.57	11	6	0
BD1	Bacteria; Chloroflexi[1]	2.89	30	157,947	60.9	90.9	0.9	0	2,429	45.74	3	3	0
BD2	Bacteria; Nitrospirae[1]	1.92	11	313,929	57.6	93.9	0.9	0	1,890	8.01	1	2	2
BD3	Archaea; Thaumarchaeota; Marine Group I; Unknown Order; Unknown Family; *Candidatus* Nitrosoarchaeum	1.48	10	250,506	33.0	98.5	1.9	1	1,869	13.93	2	2	0
BD4	Bacteria; Verrucomicrobia[2]	2.09	49	46,663	65.9	81.5	0.7	0	1,705	5.98	0	0	0

[1] Estimated using CAT

[2] Estimated using Kaiju

Table 3. Detected methylated motifs.

Genome bin	Detected methylated motif	Modification Type	Motif in REBASE	Number of methylated sites	Number of motif sequences	Methylation ratio (%)	Mean modification QV	Mean motif coverage
BS1	G**A**NTC	m6A	Yes	1,813	2,070	87.6%	58.0	35.2
	TTA**A**	m6A	Yes	1,264	1,522	83.0%	55.5	34.1
	G**C**WGC	m4C	Yes	3,026	15,948	19.0%	38.4	40.6
BS3	**G**ANTC	m6A	Yes	3,724	4,014	92.8%	66.1	41.3
	TTA**A**	m6A	Yes	3,036	3,338	91.0%	62.4	40.4
	G**C**WGC	m4C	Yes	13,821	54,026	25.6%	39.5	46.4
BS8	**A**GGNNNNNRTTT	m6A	No	80	276	29.0%	39.6	65.8
BS10	ACG**A**G	m6A	No	1,986	7,185	27.6%	45.0	171.4
BS12	GMAG**C**TKC	m4C	No	169	220	76.8%	50.9	83.5
	HCAG**C**TKC	m4C	No	124	293	42.3%	46.8	79.0
	BGMAG**C**TGD	m4C	No	78	185	42.2%	46.3	76.3
BS14	G**A**NTC	m6A	Yes	2,856	2,880	99.2%	190.6	166.9
BS15	G**A**ANNNNTTC	m6A	Yes	1,309	1,472	88.9%	55.6	30.9
	AGCNNNNNNCAT	m6A	No	642	726	88.4%	56.0	29.4
	ATGNNNNNNGCT	m6A	No	619	726	85.3%	52.0	29.8
	AGCNNNNNNGTG	m6A	No	311	349	89.1%	56.9	30.4
	C**A**CNNNNNNGCT	m6A	No	293	349	84.0%	53.3	30.9
	CA**A**NNNNNNNNCTTG	m6A	No	205	256	80.1%	49.4	29.1
	CA**A**GNNNNNNNDTTG	m6A	No	164	214	76.6%	48.7	28.7
	TT**A**GNNNNNCCT	m6A	No	87	99	87.9%	51.3	29.8
	AGGNNNNNCTAA	m6A	No	77	99	77.8%	49.4	29.7
	GYT**A**NNNNNNNTTRG	m6A	No	76	89	85.4%	56.0	31.3
	CYA**A**NNNNNNNTAVCH	m6A	No	59	127	46.5%	53.5	32.6
BD1	G**C**WGC	m4C	Yes	72,730	77,932	93.3%	140.2	297.3
	G**A**NTC	m6A	Yes	6,754	6,844	98.7%	346.3	281.7
	TTA**A**	m6A	Yes	5,475	5,564	98.4%	325.3	270.9
BD2	TANGG**A**B	m6A	No	1,276	1,367	93.3%	64.4	48.5
BD3	G**A**TC	m6A	Yes	9,446	9,618	98.2%	122.1	93.7
	AG**C**T	m4C	Yes	5,974	6,224	96.0%	84.0	92.1

R= G/A, Y= T/C, M= A/C, K= G/T, S= G/C, W= A/T, H= A/C/T, B= G/T/C, V= G/C/A, D= G/A/T, N= G/A/T/C

Table 4. Detected MTases, REases, and specificity subunit genes.

		Bioinformatic prediction							Experimental verification
Genome bin	CDS ID	Gene type[1]	Top-hit protein in REBASE	Identity (%)	Predicted recognition motif	Modification type	RM type	MTase name	Confirmed recognition motif
BS3	EMGBS3_04270	M	M.SstE37II	58.9	G**A**NTC	m6A	II		
	EMGBS3_09240	M	M.Sth20745I	71.4	TTA**A**	m6A	II		
	EMGBS3_12600	M	M1.BceSIII	22.9	A**C**GGC	m4C	II	M.AspBS3I	G**C**WGC
BS6	EMGBS6_08960	M	M.SinI	57.0	GGW**C**C	m5C	II		
BS8	EMGBS8_10720	R	DvuI	36.3	?	-	I		
	EMGBS8_10740	S	S.PveNS15I	32.4	?	-	I		
	EMGBS8_10750	M	M.RbaNRL2II	55.6	ACG**A**NNNNNNNGRTC	m6A	I		
BS10	EMGBS10_10070	RM	CjeFIII	23.7	GCA**A**GG	m6A	II	M.OspBS10I	ACG**A**G
BS14	EMGBS14_10020	M	M.Bsp460I	56.7	G**A**NTC	m6A	II		
BS15	EMGBS15_02830	M	M.Bli37I	56.6	G**A**YNNNNNRTC	m6A	I		
	EMGBS15_02840	M	M.EcoNIH1III	59.2	G**A**TGNNNNNNTAC	m6A	I		
	EMGBS15_02870	S	S.PveNS15I	47.2	?	-	I		
	EMGBS15_02930	R	DvuI	38.4	?	-	I		
	EMGBS15_03820	M	M.EcoGI	25.8	non-specific	m6A	II	M.FspBS15I	G**A**ANNNNTTC
	EMGBS15_03830	R	XmnI	34.0	GAANNNNTTC	-	II		
	EMGBS15_04560	R	GmeII	33.8	TCCAGG	-	III		
	EMGBS15_04600	M	M.FpsJII	53.4	CGC**A**G	m6A	III		
	EMGBS15_05670	M	M.FnuDI	59.8	GGCC[2]	m4C	II		
	EMGBS15_05690	R	BhaII	45.6	GGCC	-	II		
	EMGBS15_12460	M	M.Mva1261III	37.1	CT**A**NNNNNNNRTTC	m6A	I		
BD1	EMGBD1_08400	M	M.Sth20745I	71.0	TTA**A**	m6A	II		
	EMGBD1_09320	M	M1.BceSIII	22.9	A**C**GGC	m4C	II	M.AspBS3I	G**C**WGC
	EMGBD1_19510	M	M.SstE37II	58.9	G**A**NTC	m6A	II		
BD2	EMGBD2_08760	M	M.HgiDII	55.0	GTCGAC[1]	m5C	II		
	EMGBD2_08790	RM	AquIV	28.5	GRGGA**A**G	m6A	II	M.NspBD2I	TAHGG**A**B
	EMGBD2_08800	R	LpnPI	56.3	CCDG	-	II		
BD3	EMGBD3_00670	M	M.Mma5219II	45.9	AG**C**T	m4C	II		
	EMGBD3_01960	M	M.AvaVI	50.3	G**A**TC	m6A	II		

[1] M: Methyltransferase, R: Restriction endonuclease, S: specificity subunit

[2] Modified base undetermined