Title: Community-wide epigenetics provides novel perspectives on the ecology
and evolution of marine microbiome

Authors: Hoon Je Seong¹, Simon Roux², Chung Yeon Hwang³, Woo Jun Sul^{1*}

Affiliations:

¹*Department of Systems Biotechnology, Chung-Ang University; Anseong, Republic of Korea.*

²*DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.*

³*School of Earth and Environmental Sciences and Research Institute of Oceanography, Seoul National University; Seoul, Republic of Korea.*

*Corresponding author. Email: sulwj@cau.ac.kr

Abstract: DNA methylation in prokaryotes is involved in many different cellular processes including cell cycle regulation and defense against viruses. To date, most prokaryotic methylation systems have been studied in culturable microorganisms, resulting in a limited understanding of DNA methylation from a microbial ecology perspective. Here, we analyze the distribution patterns of several microbial epigenetics marks in the ocean microbiome through genome-centric metagenomics across all domains of life. We show that overall, DNA methylation can readily be detected across dominant oceanic bacterial, archaeal, and viral populations, and microbial epigenetic changes correlate with population differentiation. Furthermore, our genome-wide epigenetic analysis of *Pelagibacter* suggests that GANTC, a DNA methyltransferase target motif, is related to the cell cycle and is affected by environmental conditions. Yet, the presence of this motif also partitions the phylogeny of the *Pelagibacter* phages, possibly hinting at a competitive co-evolutionary history and multiple effects of a single methylation mark.

One-Sentence Summary: DNA methylation patterns are associated with ecological changes and virus-host dynamics in the marine microbiome.

Main text: DNA methylation is an important epigenetic modification in which methyl groups are added to nucleotides by methyltransferases (MTases). Although much of the literature has focused on DNA methylation in eukaryotes, the last decade has seen increased research on DNA methylation in prokaryotes due to the emergence of third-generation sequencing technology. In prokaryotes, DNA methylation has been largely associated with restriction–modification (RM) systems, which protect host cells against invasion by viruses or against horizontal gene transfer of extracellular DNA by distinguishing host DNA from sequence-specific DNA methylation. DNA methylation also has roles in gene expression regulation, virulence, DNA mismatch repair, and cell-cycle regulation in prokaryotes. DNA MTases that methylate nucleotides without cognate restriction enzymes (REases) are referred to as orphan MTases. Some well-studied orphan MTases play important physiological roles beyond RM systems, including transcriptional regulation and cell phenotype variations (*1-4*). Thus, there has been increasing interest in the role of prokaryotic methylation systems in bacterial genetics, phenotypic changes, and pathogenesis. However, the implications for the diversity and meaning of prokaryotic methylation systems in environmental conditions are unclear. The primary reason is that previous research has focused extensively on culturable prokaryotes, whereas the majority of marine bacteria are currently not culturable under laboratory conditions.

The advent of long-read single-molecule real-time (SMRT) sequencing has opened a new era in methylation research. While SMRT sequencing cannot currently identify all known base modifications, it can readily detect the two major DNA modifications found in prokaryotes: N^6 -methyladenosine (m^6A) and N^4 -methylcytosine (m^4C), including from metagenomes. This methylation information has been used in the past to improve prokaryotic genome binning (*5, 6*). However, outside these few examples, few attempts have been made to apply long-read metagenomic sequencing to detect DNA methylation directly in the environmental microbiome (*7*), and link methylation patterns to ongoing eco-evolutionary processes.

Here, we applied meta-epigenomic analysis to genome-centric metagenomics of the ocean microbiome in the northwest Pacific Ocean to reveal the role of DNA methylation in environmental microbial communities. We report that the DNA methylome is differentiated by taxonomic lineage

and is affected by the complexity of the community, i.e., the co-existence of multiple closely related strains. We further link methylation patterns to cell cycle regulation and phage defense for *Pelagibacter* genomes, highlighting the multiple roles played by DNA methylation in one of the dominant bacteria of the marine environment.

Results and Discussion. Novel microbial genomes from the northwest Pacific Ocean metagenome. In the 2015 Shipborne Pole-to-Pole Observations (SHIPPO) project of the Korea Polar Research Institute, we conducted shotgun metagenomic sequencing using ocean surface samples from 10 stations (referred to as St2–St11) by traveling about 4,000 km from the Pacific Northwest to the Bering Sea during July 22–29, 2015 (fig. S1, table S1). □ To capture free-living organisms, we extracted genomic DNA after 0.22–1.6-μm size filtering. □ Ten seawater samples were sequenced with 154.4 Gb (read average length: 151 bp) using the Illumina HiSeq 4000 platform and 32.2 Gb (read average length: 797.9 bp) using the Pacific Biosciences (PacBio) RSII (3.2 cells per sample) platform, respectively. Extensive computational analysis was performed on all samples to reconstruct the genomes across the kingdom using a combination of individual, co-, and hybrid assembly, binning, and refinement methods (Fig. 1). This strategy allowed the recovery of a total of 15,056 viral, 252 prokaryotic, 56 giant viral, and 6 eukaryotic metagenome-assembled genomes (MAGs, specifically referred to here as SHIPPO vMAGs, proMAGs, gvMAGs, and eukMAGs, respectively). A total of 252 dereplicated proMAGs (99% average nucleotide identity; ANI) with ≥50% completeness and <10% contamination remained (average completeness: 78.80±14.09, and contamination: 2.76±2.37), 105 originated from the individual binning and 147 from the co-assembly binning. Forty-seven had >90% completeness and <5% contamination (near-complete), of which only three were high-quality MAGs that fit the Minimum Information about a Metagenome-assembled Genome (MIMAG) criteria (8) including rRNA and tRNA. These proMAGs mainly consist of the bacterial phyla of Proteobacteria ($n = 120$), Bacteroidota ($n = 88$), Actinobacteriota ($n = 11$), and the archaeal phyla of Thermoplasmatota ($n = 15$) (Fig. 2A).

This assembly strategy substantially improved the fraction of mapped metagenomic reads. Overall, the average mappability of all samples was 38.03% (std. 2.88) (fig. S2). Most of the reads

mapped to vMAGs and proMAGs, with relatively smaller representation of gvMAGs and eukMAGs. We compared our proMAGs with the TARA (9) and Global Ocean Reference Genomes Tropics (GORG-Tropics) (10) datasets to evaluate the novelty of our recovered genomes (Fig. 2B). Although these proMAGs and single-cell amplified genomes datasets came from a global-scale study (9, 10), only three proMAGs overlapped at the species level ($\geq 95\%$ ANI) with our proMAGs. Furthermore, 95% of the proMAGs obtained here could not be classified at the species level, even though the Genome Taxonomy Database (GTDB) (11) includes genomes of uncultured organisms derived from shotgun metagenomics and single-cell genomics. Thus, despite previous extensive ocean metagenomic binning efforts such as those undertaken on data from mega-surveys like the *Tara* Ocean Expedition (TARA) and the Global Ocean Survey (9, 10), the northwest Pacific Ocean datasets from this study provide substantial novel genomic information on ocean microbiomes.

DNA MTases in marine microorganisms are rarely associated with an RM system. To characterize the role of MTases in ocean microbial communities, we first identified the type of MTases and their cognate REases distribution from the genome catalog established here and derived from previous ocean microbiome surveys (GORG, TARA, and SHIPPO). Of the total 5,713 medium-quality proMAGs, we found 67.18% (3,838) and 19.45% (1,111) of proMAGs encoded one or more MTases and REases, respectively (table S2). Among the four MTase types—I, II, III, and IV—type II MTase was found most frequently (94.71%) in proMAGs with MTases, followed by type I (14.02%) and III (2.16%). Of all the proMAGs, only 14.77% had a complete RM system; most consisted of type I and III MTases: 76.39% and 74.70% of type I and III MTases constituted a complete RM system, respectively, whereas only 3.19% of type II MTases constituted an RM system. By contrast, most MTases (86.09%) belonged to orphan MTases, thus lacking counterpart REases, and consisted of type II MTases. To compare the genome size of ocean prokaryotes and the number of genes related to RM systems, 424 near-complete proMAGs were used. Although the number of MTase and REase correlated with genome size, the genome of proMAG with one or more RM system was significantly larger than that of proMAG with orphan MTase (Wilcoxon's rank-sum P value $< 2.2 \times 10^{-16}$). In addition, in the case of proMAG with RM system, the correlation between MTase and genome size (r^2 :

0.29) was higher than for REase (r^2 : 0.16) (Fig. 2C). In the ocean microbial community, loss of the type II RM system may be caused by the selective pressures of genome streamlining in the pelagic environment (12), which contributes to retaining essential cellular mechanisms rather than defense systems. By contrast, type I and III MTases and their cognate REases typically serve as a defense mechanism through the RM system and thus are harbored in genomes of relatively large size.

Beyond MTases detected in proMAGs, a total of 959 MTases were found in the SHIPPO MAGs catalog, including all ($n = 6$) of the eukMAGs, 62.5% ($n = 35$) of the gvMAGs, 36.90% ($n = 93$) of the proMAGs, and 4.28% ($n = 645$) of the vMAGs (Fig. 2D and table S3). Consistent with the abovementioned results, type II MTases were the most frequently detected for all domains (Fig. 2D). All but two MTases were solitary or orphan MTases that have no counterpart REases. Eukaryotes and giant viruses had an average of 5.0 and 2.2 MTases per genome, whereas fewer MTases were found in the prokaryotic and viral genomes (1.58 and 1.09 per genome, respectively).

DNA methylome of SHIPPO MAGs. We next studied the DNA methylation patterns of the ocean microbiome and compared DNA methylation profiles of each SHIPPO MAG across samples. We first performed a principal coordinate analysis (PCoA) based on the Kulczynski dissimilarity of 5-mer DNA methylation profiles for individual MAG-sample pairs (requiring 10× coverage with 20% genome breadth for proMAGs and gvMAGs, 10% for eukMAGs, and 60% for vMAGs). DNA methylation profiles were grouped clearly by domain, i.e., separating eukaryotes, prokaryotes, and virus MAGs (Fig. 3A). In particular, Alphaproteobacteria harbored distinct methylation profiles compared to all other microbial organisms, and Alphaproteobacteria proMAGs were partitioned from each other down to the family level. However, the methylation profile of 5-mers could not be distinguished at the species level, as in the example of the *Pelagibacteraceae* cluster, which consisted of SHIPP_PRO_33, SHIPP_PRO_245, and SHIPP_PRO_247. Furthermore, proMAGs belonging to Flavobacteriia, Actinobacteria, Gammaproteobacteria, and Bacteroidia were also difficult to distinguish by their methylation profile.

To identify the exact DNA methylated motif, the methylated motif information was collected via the Restriction Enzyme database (REBASE) (13). Although additional motifs were discovered from

the *de novo* motif finding algorithm MultiMotifMaker (14) across SHIPPO MAGs, most motifs were discovered in previous studies (GATC, GANTC, CGCG, VATB, underlining indicates methylation position) (15, 16). A total of 1,357 motifs were searched along the mapped region of each SHIPPO MAG. When <20% of each motif was methylated in each genome, it was considered noise and excluded from the candidate methylation motifs. Ninety-five candidate methylated motifs were detected, of which 17 and 76 represented m⁴C and m⁶A modifications, respectively. The other two were non-palindromic motifs (CGTCTC/GAGACG, GAAGA/TCTTC) with both m⁴C and m⁶A methylation. Among the methylated motifs, 13 motifs were shared across domains. GANTC was found most frequently in several families belonging to Alphaproteobacteria and in *Caudovirales* (Fig. 3B). CGCG and GGTAG were detected in both archaeal proMAGs affiliated to MGIIA and *Caudovirales* vMAGs (Fig. 3B). The GATC motif was found in gvMAGs (*Phycodnaviridae*), *Caudovirales*, and unclassified vMAGs (Fig. 3B). In addition, there were methylated motifs unique to vMAGs, such as CCWGG and GGCC, and unique motifs were found in each family, such as TTAA (*Microbacteriaceae*), and TGCGA (MGIIA) (Fig. 3B). The high diversity of methylated motifs in vMAG suggests the existence of many unknown MTases encoded on prokaryotic and/or viral genomes, and is consistent with a role of viral genome methylation in virus-host arms race.

To match the methylated motifs with MTases of each SHIPPO MAG, candidate MTases were searched against REBASE (13) reference MTases with recognition sequence information. Only the recognition sequences of 12 MTases were known, and except for the recognition sequences of one viral MTase, it was confirmed that the methylated motif sequences were identical to the recognition sequence information (table S3). For example, in the seven Alphaproteobacteria proMAGs, we could identify methylation signals from several motifs, including GANTC, GASTC, GAGTC, and CGANNNNNNAATC, but these were represented by GATNC, for which the congruent recognition sequences of the best similarity reference MTase could be found (table S3). There were four methylation motifs in an archaeal proMAG (SHIPPO_PRO_101), and after deduplication, two different motifs (CGCG and GGTAG) were represented (table S3). One congruent MTase could be found (CGCG) among these two methylation motifs, but no MTase matched the other methylation

motif (GGTAG). Although the GGTAG motif was novel in that it is previously unreported, this result limited the confirmation of the archaeal novel methylation system, likely due to the fact that this MAGs does not represent a complete genome. Furthermore, except for 12 of the 1,124 MTases, it was either difficult to identify the methylation profile due to the long-read sequencing depth or the lack of MTase recognition sites in the previous database made it difficult to compare between MTases and motif sequences.

DNA methylation patterns are linked to population genomic structure. Several studies of bacterial DNA methylome suggested that different bacterial strains have different methylome patterns, even within species (6, 17, 18). These changes can be caused by the presence of MTases or by phase-variable MTases that respond to changes in the environment (19). However, microbial DNA methylation changes in complex environments have not yet been measured directly; therefore, we analyzed intra-species DNA methylation variation at the sampling station. The fixation index (F_{ST}) was used to compare the similarity in the population differentiation between samples. proMAGs with low base-pair coverage were excluded because the F_{ST} had to be calculated by the allele frequencies within the species. F_{ST} was calculated for dominant proMAGs using a mapping region that overlaps at least 40% breadth with 10× depth in all samples with short reads. The dissimilarity of methylomes was calculated by the methylation frequency of 5-mer nucleotides in a genome. Therefore, to compare the DNA methylation changes for different sampling stations, only six species-level proMAGs were fulfilled by the sequencing coverage of long- and short reads. In five of the six proMAGs, DNA methylome differences and population differentiation across samples correlated significantly, regardless of the distance between sampling stations (Pearson correlation, P value < 0.05; Fig. 3C). A significant correlation was also found in Gammaproteobacteria proMAGs that showed no specific methylated motif (weak methylated motifs ratio <10%). These results may indicate that when multiple strains with different methylation profiles of motifs are present in the environment, they affect the methylation pattern at the species level.

Genome-wide DNA methylation analysis of *Pelagibacter* in environmental samples. Next, we analyzed the methylation pattern of dominant proMAGs at the single-nucleotide level to investigate in

detail the environmental-dependent changes in methylation. Identifying the single-nucleotide-level DNA methylation from a genome-wide perspective required a relatively deeper and wider read coverage. Four proMAGs (SHIPPO_PRO_33, SHIPPO_PRO_246, SHIPPO_PRO_64, and SHIPPO_PRO_101) were covered over >40% of the breadth of the genome at 20× depth per strand. Of these, only a SHIPPO_PRO_33 proMAG affiliated with *Pelagibacter* overlapped at 65.61% of the genome breadth in all 10 samples. The average breadth of the genome coverage with 20× per strand of SHIPPO_PRO_33 was 90.93% (std. 7.40). The in-depth overlapped coverage of long-reads enabled a comparison of methylation patterns between samples for this specific *Pelagibacter* proMAG.

As mentioned above, the MTases of most marine microorganisms were found without counterpart REases, and conserved MTases were observed without counterpart REases in the four *Pelagibacter* proMAGs. In particular, the *Pelagibacter* proMAGs showed that the methylated motif (GANTC) and the recognition motif of their MTase were consistent. The GANTC motif in the *Pelagibacter* proMAGs was globally uniformly distributed throughout the genome (Shapiro–Wilk test: genic >0.05, intergenic >0.05, and regulatory >0.05). For the bacterial MTases involved in gene regulation, methylated motifs were frequently located upstream of regulated genes (20). In this study, we found that GANTC was enriched in the intergenic region of *Pelagibacter*, indicating that GANTC favors an epigenetic role (P value = 0.05, fig. S3).

To compensate for the lack of sequencing depth in the genome between samples, we focused on the genome-wide epigenetic analysis of the *Pelagibacter* proMAG (SHIPPO_PRO_33; 12 contigs; 87.30 completeness, 3.79 contamination). Only the nucleotide positions of the overlapped regions that could measure methylation in all samples were compared. A total of 2,494 GANTC motifs were detected on both strands of the SHIPPO_PRO_33 genome, and only 1,719 GANTC (68.93%) were included in the overlapped region (the average depth of reads was 112.85×, std. 60.56) for measuring the methylation at 10 sampling stations. Although most motifs are known to be methylated throughout the prokaryotic genome (16), some motif sites have been found that remain unmethylated (20–22). These unmethylated sites may be due to competitive binding between MTases and regulatory proteins

due to epigenetic regulation (23, 24). The methylation rate of GANTC varied in the range 71.31–94.77% depending on the sample, of which 2.44% (42 sites) remain unmethylated in all samples. In most cases, unmethylated sites were frequently observed in intergenic regions (66.67%), including regulatory regions (16.70%) (table S3). The GANTC position of one of these seven unmethylated positions of the regulatory region contains the *sufE* (K04488) regulatory region of the Suf (sulfur-forming) system. This gene is part of the SufSE complex with SufS and is involved in the transport of sulfur to the sulfur mobilization protein. This gene was also significantly more highly expressed in the absence of dimethylsulfoniopropionate (DMSP), sulfur-limited in the laboratory, along with *sufS* (25). In addition, we observed unmethylated GANTC in the regulatory region of *groEL* (chaperones gene, K04077), which is also more expressed in sulfur-restricted conditions (25). However, we found that most GANTC motifs (99.86%, 3477/3482) were fully methylated (both strands methylated) under the DMSP-rich laboratory culture condition of *Candidatus Pelagibacter Giovannoni* NP1 (26), which is most closely related to SHIPPO_PRO_33 (ANI $\geq$ 92.36). All but five GANTC sites were methylated throughout the chromosome. Two of the five unmethylated sites were fully methylated in both strands of the tRNA (Phe) region, and the rest were methylated in the intergenic region. However, the SHIPPO_PRO_33 proMAGs from environmental samples lacked genomic integrity and could not be compared with the five unmethylated GANTC positions of *Cand. P. Giovannoni* NP1. The differences in unmethylated GANTCs between environmental and laboratory settings suggest that they are the result of epigenetic regulation in competition for nutrients, such as sulfur.

The DNA methylation signal at the nucleotide resolution is calculated from the pooled interpulse duration (IPD) ratio in separate molecules for each genomic locus. Due to the often-found epigenetic heterogeneity (27, 28), these aggregated methylation signals indicate the methylation of cell fractions at the nucleotide level (hereafter referred to as methylation fraction). The methylation fraction of GANTCs was observed in heterogeneity across samples, particularly in St6, St8, and St9. The differences in the methylation fractions on nucleotide sites were referred to as single nucleotide methylation variation (SNMV). Compared to other samples, the number of unmethylated motifs was higher in St6, St8, and St9 and showed different SNMV patterns at each nucleotide position (Fig. 4A).

By contrast, all samples, except for St6, St8, and St9, had similar SNMVs to each other and were more closely clustered, regardless of latitude (Fig. 4B). For example, pairs St2 and St3, and St10 and St11, distanced geographically by about 20° latitude and 3,000 km distance, were grouped closely through PCoA (Fig. 4B). In particular, differences in strain-level composition (single nucleotide variants; SNVs) were found between these two groups, but in SNMVs, these groups were clustered together (Fig. 4B). These results indicated that DNA methylation at the single nucleotide level differed under environmental conditions regardless of strain composition, which suggests that dynamic cellular events occur among various *Pelagibacter* in northwest Pacific Ocean surface waters.

Cell-cycle regulation of *Pelagibacter* by MTase activity. A gradual decrease was observed in the exponential phase of the genome-wide methylation fraction of GANTC from the replication origin (*ori*) to replication terminus (*ter*) of the *Cand. P. Giovannoni* NP1 chromosome (fig. S4). In addition, the same pattern was observed with the *Pelagibacter* proMAG (SHIPPO_PRO_33) across several samples under real environmental conditions (Fig. 4D). The MTase CcrM, which methylates the GANTC motif, is a representative example of an MTase involved in the cell-cycle regulation in Alphaproteobacteria (29). As chromosomal replication proceeds, the methylation status of the parental strand is maintained, whereas GANTC remains unmethylated when a new daughter strand is generated through a replication fork (hemimethylated) (Fig. 4C). *CcrM* is only expressed at the end of chromosome replication, so the chromosome remains hemimethylated until the end of replication. When DNA methylation signals are pooled from separate molecules for each genomic locus, the methylation fraction may result in relatively lower values near the *ori* due to the ratio of hemimethylation to full methylation following different DNA replication instances for different cells in the exponential phase (fig. S4). Based on the above hypothesis, a relatively lower methylation fraction in the *ori* of the *Pelagibacter* chromosome suggests that GANTC methylation may also be involved in the cell cycle of *Pelagibacter*. As in the nutrient-rich laboratory environment, the same pattern of methylation fractions in the *Pelagibacter* chromosome was observed in marine environments (Fig. 4E); this suggests that *Pelagibacter* is also in the exponential phase in real marine environments.

DNA methylation of the viral genome and the possibility of legacy from the host. The RM system is ubiquitous in about 90% of the bacterial genome (30) and acts as a defense mechanism to distinguish between self and non-self DNA. However, due to the lack of cognate REases in most genomes in this study, DNA methylation is thought to be used to regulate the cellular mechanisms rather than defense mechanisms of exogenous invasive DNA, such as the RM system, the Bacteriophage EXclusin (BREX) system (31), and the Defense Island System Associated with restriction–modification (DISARM) (32) in the ocean microbial communities. We only observed methylation patterns in 83 vMAGs associated with 67 vOTUs (viral operational taxonomic units) in a total of 15,056 putative viral genomes. The 83 vMAGs grouped into five major lineages based on pairwise genome similarity (Fig. 5A). Most of the clades were composed of vMAGs belonging to the *Caudovirales* order, except for clade V which showed genomic similarity to high-quality gvMAGs, and is likely composed of genome fragments related to *Phycodnaviridae*. Within the individual clades, there were typically two or three distinct subclades with outgroups, and some methylation patterns were shared between subclades. For instance in clade-III, specific adenine methylation motifs were found within subclades, which were not shared with other clades. Although these subgroups are phylogenetically related, the two different methylated adenine motifs (GRRGA/None and CCATC/GATGG) did not overlap and represent entirely different motif patterns. Cytosine methylation was dominant throughout clade-II, and methylation was observed in CGCG, GAGCTC, and CYCGRG motifs according to subgroups. On the other hand, both adenine and cytosine methylation were found in clade-IV, and specific methylation was indicated in the GANTC and CCWGG motifs, respectively, depending on the clade. In addition, the methylation in the GATC motif was consistent in clade-I and, together with the motifs found in clades-IV, suggests that it may be associated with bacterial methylation system as a methylation motif frequently found in Proteobacteria. For clade-V, adenine methylation was detected, as previously reported for *Phycodnaviridae* (33); in our study, these were found on GATC and CATG motifs. These two methylated motifs were observed spanning genomes belonging to *Phycodnaviridae*, and importantly the same motif has been previously reported in some of their green algae hosts (15).

To investigate how viral methylation patterns are associated with MTase genes, we next evaluated the distribution of MTases according to the phylogenetic distribution of vMAG clades. Of 83 vMAGs, we found only 14 vMAGs encoded MTase genes, all of which belonged to the Type II MTase. Most MTases showed phylogenetic consistency with the vMAGs (Fig. 5B), along with consistencies in the methylated motifs. For instance, 4 vMAGs in the clade-IIIb, all methylated to the CCATC/GATGG motif were found to encode closely related MTases, yet these MTases had less than 50% similarity to compare within the REBASE. This lack of similarity to characterized MTases was a common observation for vMAG-encoded MTases, and there was thus only one instance of methylated motifs consistent with the predicted recognition sites based on the closest existing Mtases, for vOTU_N_648 in Clade-I. This suggests that most viral MTase systems are still poorly characterized, and we expect that methylation motifs for these novel MTases could be predicted based on the methylation patterns and phylogenetic distributions of their corresponding vMAGs. Viruses primarily use DNA methylation to counter host defense systems, such as the RM system, although they also exploit methylation to signal the transition in their status from the latent to lytic state (34). The methylation of viral DNA is also inherently transient as it is either derived from viral-encoded MTase or the host's MTase and can be demethylated via passage through an MTase-free host. However, as mentioned above, bacterial DNA methylation in the marine environment does not appear to be a defense mechanism primarily. Viruses use their own MTases as a counter-defense and as a signal to initiate their lysis state and DNA packaging (35, 36), however incongruences were observed in our study between the methylome of the viral genome and the recognition motif of the predicted MTase. Therefore, DNA methylation of the viral genome in the marine environment is most likely a marker left behind from a previous host. To support the hypothesis, we further analyzed the DNA methylomes of the viral genomes detected at multiple sampling stations. Only four genomes had >60% of the average breadth of the genome coverage with 10x depth spanning three or more sampling stations. Notably, the DNA methylation profile differed according to the sampling station. For example, in the genome of vOTU sg 2962, most motifs associated with GANTC (GAGTC, GAWTC, and GASTC) were methylated in St6 and St8, whereas DNA methylation was only detected in GAGAG in St7 (Fig.

5B). Although further studies are needed to understand the biological and ecological significance of DNA methylation in viruses in the real world, an explanation for our results is that different hosts altered the virus methylome with MTases that recognize different motifs.

Possible implication of methylation in SAR11 phage-host interactions. Most GANTC positions were methylated throughout the *Cand. P. Giovannoni* NP1 genome (fig. S4), and this methylation mark is thought to be related to cell cycle regulation (see above). However, uneven methylation patterns were observed in specific regions of the genome due to the lack of GANTC motifs in the proviral region (Fig. 6A). This raised questions about the potential use of the same methylation mark as a defense system, and the evasion of this defense system by pelagiphages (phages infecting SAR11 bacteria) at point in the evolutionary history of the phage-host pairs.

In the eight proMAGs of the SHIPPO catalog spanning the order Pelagibacterales, their genomes lacked all genes associated with typical defense mechanisms including CRISPR and RM systems. While SAR11 is exposed to an environment where phage predation is frequent, it is currently thought that by reducing its genome size, it can gain an edge in the environment because of its superior nutrient uptake competitiveness (37). We thus interpret the lack of RM systems in SHIPPO *Pelagibacterales* proMAGs as a sign of a strong selection pressure towards genome reduction, suggesting that DNA methylation would not be currently used as an anti-phage defense by these bacterial populations.

We next reconstructed the phylogeny of 33 publicly available pelagiphage genomes based on a head–tail connector protein and compiled the frequency of GANTC motifs in the same genomes (Fig. 6B). Surprisingly, about half of the pelagiphage genomes showed a clear depletion in GANTC motif, and the calculated phylogenetic trees clearly showed that the phages clustered according to the density of GANTC motifs in their genome (Fig. 6B and 6C). This difference was particularly observed in the subfamily *Autographivirinae* of the family *Podoviridae*. This raises the possibility that methylation marks may have been used as anti-phage defense in the past, and some pelagiphages would have retained a genome composition bias since. Alternatively, it is possible that some strains within the SAR11 clade do use RM systems as phage defense, and the difference in GANTC frequency would be

related to differences in host range between phage clades.

Conclusion. Our results describe the biological meaning of DNA methylation in marine microorganisms that have not been previously interpreted from an environmental perspective. Diverse DNA methylation patterns across the kingdoms and changes in marine microbial communities have contributed to a wide range of implications, from the role of DNA methylation in marine microbes to the co-evolution history between phage-hosts. In this meta-epigenomic study, we describe how genome-wide epigenetic analysis and phage-host-related methylation could be implemented as novel interpretations of the microbiome, together with the construction of future metagenome-methylation databases.

References

1. K. Vasu, V. Nagaraja, Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.* **77**, 53-72 (2013).
2. M. De Ste Croix, I. Vacca, M. J. Kwun, J. D. Ralph, S. D. Bentley, R. Haigh, N. J. Croucher, M. R. Oggioni, Phase-variable methylation and epigenetic regulation by type I restriction–modification systems. *FEMS Microbiol. Rev.* **41**, S3-S15 (2017).
3. A. Tan, J. M. Attack, M. P. Jennings, K. L. Seib, The capricious nature of bacterial pathogens: phasevarions and vaccine development. *Front. Immunol.* **7**, 586 (2016).
4. M. A. Sánchez-Romero, J. Casadesús, The bacterial epigenome. *Nat. Rev. Microbiol.* **18**, 7-20 (2020).
5. J. Beaulaurier, S. Zhu, G. Deikus, I. Mogno, X. S. Zhang, A. Davis-Richardson, R. Canepa, E. W. Triplett, J. J. Faith, R. Sebra, E. E. Schadt, G. Fang, Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.* **36**, 61-69 (2018).
6. A. Tourancheau, E. A. Mead, X.-S. Zhang, G. Fang, Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat. Methods*, **18**, 491-498 (2021).

7. S. Hiraoka, Y. Okazaki, M. Anda, A. Toyoda, S. I. Nakano, W. Iwasaki, Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an environmental prokaryotic community. *Nat. Commun.* **10**, 159 (2019).
8. R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Elie-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, The Genome Standards Consortium, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. M. Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, T. Woyke, Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725-731 (2017).
9. T. O. Delmont, C. Quince, A. Shaiber, Ö. C. Esen, S. T. M. Lee, M. S. Rappé, S. L. McLellan, S. Lückner, A. M. Eren, Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* **3**, 804-813 (2018).
10. M. G. Pachiadaki, J. M. Brown, J. Brown, O. Bezuidt, P. M. Berube, S. J. Biller, N. J. Poulton, M. D. Burkart, J. J. La Clair, S. W. Chisholm, R. Stepanauskas, Charting the complexity of the marine microbiome through single-cell genomics. *Cell* **179**, 1623-1635 (2019).
11. D. H. Parks, M. Chuvochina, P.-A. Chaumeil, C. Rinke, A. J. Mussig, P. Hugenholtz, A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079-1086 (2020).
12. S. J. Giovannoni, J. C. Thrash, B. Temperton, Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553-1565 (2014).
13. R. J. Roberts, T. Vincze, J. Posfai, D. Macelis, REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **43**, D298-D299 (2015).

14. T. Li, X. Zhang, F. Luo, F.-X. Wu, J. Wang, MultiMotifMaker: a multi-thread tool for identifying DNA methylation motifs from Pacbio reads. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**, 220-225 (2020).
15. S. Zhu, J. Beaulaurier, G. Deikus, T. P. Wu, M. Strahl, Z. Hao, G. Luo, J. A. Gregory, A. Chess, C. He, A. Xiao, R. Sebra, E. C. Schadt, G. Fang, Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. *Genome Res.* **28**, 1067-1078 (2018).
16. D. Wion, J. Casadesús, N⁶-methyl-adenine: an epigenetic signal for DNA–protein interactions. *Nat. Rev. Microbiol.* **4**, 183-192 (2006).
17. P. H. Oliveira, J. W. Ribis, E. M. Garrett, D. Trzilova, A. Kim, O. Sekulovic, E. A. Mead, T. Pak, S. Zhu, G. Deikus, M. Touchon, M. Lewis-Sandari, C. Beckford, N. E. Zeitouni, D. R. Altman, E. Webster, I. Oussenko, S. Bunyavanich, A. K. Aggarwal, A. Bashir, G. Patel, F. Wallach, C. Hamula, S. Huprikar, E. E. Schadt, R. Sebra, H. van Bakel, A. Kasarskis, R. Tamayo, A. Shen, G. Fang, Epigenomic characterization of *Clostridioides difficile* finds a conserved DNA methyltransferase that mediates sporulation and pathogenesis. *Nat. Microbiol.* **5**, 166-180 (2020).
18. G. Fang, D. Munera, D. I. Friedman, A. Mandlik, M. C. Chao, O. Banerjee, Z. Feng, B. Losic, M. C. Mahajan, O. J. Jabado, G. Deikus, T. A. Clark, K. Luong, I. A. Murray, B. M. Davis, A. Keren-Paz, A. Chess, R. J. Roberts, J. Korlach, S. W. Turner, V. Kumar, M. K. Waldor, E. E. Schadt, Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* **30**, 1232-1239 (2012).
19. Y. N. Srikhanta, R. J. Gorrell, J. A. Steen, J. A. Gawthorne, T. Kwok, S. M. Grimmond, R. M. Robins-Browne, M. P. Jennings, Phasevarion mediated epigenetic gene regulation in *Helicobacter pylori*. *PLoS ONE* **6**, e27569 (2011).

20. M. J. Blow, T. A. Clark, C. G. Daum, A. M. Deutschbauer, A. Fomenkov, R. Fries, J. Froula, D. D. Kang, R. R. Malmstrom, R. D. Morgan, J. Posfai, K. Singh, A. Visel, K. Wetmore, Z. Zhao, E. M. Rubin, J. Korlach, L. A. Pennacchio, R. J. Roberts, The epigenomic landscape of prokaryotes. *PLoS Genet.* **12**, e1005854 (2016).
21. W. B. Hale, M. van der Woude, D. A. Low, Analysis of nonmethylated GATC sites in the *Escherichia coli* chromosome and identification of sites that are differentially methylated in response to environmental stimuli. *J. Bacteriol.* **176**, 3438-3441 (1994).
22. S. Tavazoie, G. M. Church, Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in *E. coli*. *Nat. Biotechnol.* **16**, 566-571 (1998).
23. H. N. Lim, A. van Oudenaarden, A multistep epigenetic switch enables the stable inheritance of DNA methylation states. *Nat. Genet.* **39**, 269-275 (2007).
24. S. Ardisson, P. Redder, G. Russo, A. Frandi, C. Fumeaux, A. Patrignani, R. Schlapbach, L. Falquet, P. H. Viollier, Cell cycle constraints and environmental control of local DNA hypomethylation in α -proteobacteria. *PLoS Genet.* **12**, e1006499 (2016).
25. D. P. Smith, C. D. Nicora, P. Carini, M. S. Lipton, A. D. Norbeck, R. D. Smith, S. J. Giovannoni, Proteome remodeling in response to sulfur limitation in "*Candidatus Pelagibacter ubique*". *mSystems* **1**, e00068-16 (2016).
26. R. M. Morris, K. R. Cain, K. L. Hvorecny, J. M. Kollman, Lysogenic host–virus interactions in SAR11 marine bacteria. *Nat. Microbiol.* **5**, 1011-1015 (2020).
27. J. Casadesús, D. A. Low, Programmed heterogeneity: epigenetic mechanisms in bacteria. *J. Biol. Chem.* **288**, 13929-13935 (2013).
28. A. S. Manso, M. H. Chai, J. M. Attack, L. Furi, M. De Ste Croix, R. Haigh, C. Trappetti, A. D. Ogunniyi, L. K. Shewell, M. Boitano, T. A. Clark, J. Korlach, M. Blades, E. Mirkes, A. N. Gorban, J. C. Paton, M. P. Jennings, M. R. Oggioni, A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat. Commun.* **5**, 5055 (2014).
29. G. Zweiger, G. Marczynski, L. Shapiro, A *Caulobacter* DNA methyltransferase that functions

- only in the predivisional cell. *J. Mol. Biol.* **235**, 472-485 (1994).
30. P. H. Oliveira, M. Touchon, E. P. C. Rocha, The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**, 10618-10631 (2014).
 31. T. Goldfarb, H. Sberro, E. Weinstock, O. Cohen, S. Doron, Y. Charpak-Amikam, S. Afik, G. Ofir, R. Sorek, BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* **34**, 169-183 (2015).
 32. G. Ofir, S. Melamed, H. Sberro, Z. Mukamel, S. Silverman, G. Yaakov, S. Doron, R. Sorek, DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.* **3**, 90-98 (2018).
 33. W. H. Wilson, J. L. Van Etten, M. J. Allen, "The *Phycodnaviridae*: the story of how tiny giants rule the world. *Curr. Top. Microbiol. Immunol.* **328**, 1-42 (2009).
 34. K. Hoelzer, L. A. Shackelton, C. R. Parrish, Presence and role of cytosine methylation in DNA viruses of animals. *Nucleic Acids Res.* **36**, 2825-2837 (2008).
 35. N. Sternberg, J. Coulby, Cleavage of the bacteriophage P1 packaging site (pac) is regulated by adenine methylation. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 8070-8074 (1990).
 36. J. Murphy, J. Mahony, S. Ainsworth, A. Nauta, D. van Sinderen, Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. *Appl. Environ. Microbiol.* **79**, 7547-7555 (2013).
 37. S. J. Giovannoni, H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, M. S. Rappé, J. M. Short, J. C. Carrington, E. J. Mathur, Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242-1245 (2005).
 38. C.-C. Lo, P. S. Chain, Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinform.* **15**, 1-8 (2014).
 39. H. Xu, X. Luo, J. Qian, X. Pang, J. Song, G. Qian, J. Chen, S. Chen, FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS ONE* **7**, e52249 (2012).

40. D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676 (2015).
41. J. R. Wang, J. Holt, L. McMillan, C. D. Jones, FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinform.* **19**, 1-11 (2018).
42. M. Kolmogorov, D. M. Bickhart, B. Behsaz, A. Gurevich, M. Rayko, S. B. Shin, K. Kuhn, J. Yuan, E. Polevikov, T. P. Smith, P. A. Pevzner, metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103-1110 (2020).
43. J. Neauport, L. Lamagnere, H. Bercegol, F. Pilon, J.-C. Birolleau, Polishing-induced contamination of fused silica optics and laser induced damage density at 351 nm. *Opt. Express* **13**, 10163-10171 (2005).
44. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, (2013).
45. J. Alneberg, B. S. Bjarnason, I. De Bruijn, M. Schimer, J. Quick, U. Z. Lahti, N. J. Loman, A. F. Andersson, C. Quince, Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144-1146 (2014).
46. D. D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, Z. Wang, MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
47. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043-1055 (2015).
48. M. R. Olm, C. T. Brown, B. Brooks, J. F. Banfield, dRep: a tool for fast and accurate genomic

- comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864-2868 (2017).
49. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359 (2012).
50. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
51. R. R. Wick, L. M. Judd, C. L. Gorrie, K. E. Holt, Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).
52. F. Asnicar, A. M. Thomas, F. Beghini, C. Mengoni, S. Manara, P. Manghi, Q. Zhu, M. Bolzan, F. Cumbo, U. May, J. G. Sanders, Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 1-10 (2020).
53. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293-W296 (2021).
54. P. T. West, A. J. Probst, I. V. Grigoriev, B. C. Thomas, J. F. Banfield, Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* **28**, 569-580 (2018).
55. P. Saary, A. L. Mitchell, R. D. Finn, Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol.* **21**, 244 (2020).
56. T. O. Delmont, M. Gaia, D. D. Hinsinger, P. Fremont, C. Vanni, A. F. Guerra, A. M. Eren, A. Kourlaiev, L. d'Agata, Q. Clayssen, E. Villar, Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *bioRxiv*, 2020.2010.2015.341214 (2021).

57. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).
58. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
59. L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268-274 (2015).
60. B. Q. Minh, M. A. T. Nguyen, A. von Haeseler, Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188-1195 (2013).
61. S. Roux, F. Enault, B. L. Hurwitz, M. B. Sullivan, VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
62. D. Paez-Espino, G. A. Pavlopoulos, N. N. Ivanova, N. C. Kyrpides, Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat. Protoc.* **12**, 1673-1682 (2017).
63. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
64. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
65. D. Paez-Espino, S. Roux, I.-M. A. Chen, K. Palaniappan, A. Ratner, K. Chu, M. Huntemann, T. B. K. Reddy, J. C. Pons, M. Llabrés, E. A. Elie-Fadrosh, N. N. Ivanova, N. C. Kyrpides, IMG/VR v. 2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.* **47**, D678-D686 (2019).

66. A. C. Gregory, A. A. Zayed, N. Conceição-Neto, B. Temperton, B. Bolduc, A. Alberti, M. Ardyna, K. Arkhipova, M. Carmichael, C. Cruaud, C. Dimier, G. Domínguez-Huerta, J. Ferland, S. Kandels, Y. Liu, C. Marec, S. Pesant, M. Picheral, S. Pisarev, J. Poulain, J.-É. Tremblay, D. Vik, Tara Oceans Coordinators, M. Babin, C. Bowler, A. I. Culley, C. de Vargas, B. E. Dutilh, D. Iudicone, L. Karp-Boss, S. Roux, S. Sunagawa, P. Wincker, M. B. Sullivan, Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109-1123 (2019).
67. F. Schulz, S. Roux, D. Paez-Espino, S. Jungbluth, D. A. Walsh, V. J. Denef, K. D. McMahon, K. T. Konstantinidis, E. A. Elie-Fadrosh, N. C. Kyrpides, T. Woyke, Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432-436 (2020).
68. S. Nayfach, A. P. Camargo, F. Schulz, E. Elie-Fadrosh, S. Roux, N. C. Kyrpides, CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578-585 (2021)
69. S. Schloissnig, M. Arumugam, S. Sunagawa, M. Mitreva, J. Tap, A. Zhu, A. Waller, D. R. Mende, J. R. Kultima, J. Martin, K. Kota, Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45-50 (2013).
70. A. Lomsadze, K. Gemayel, S. Tang, M. Borodovsky, Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res.* **28**, 1079-1089 (2018).
71. V. Ter-Hovhannisyan, A. Lomsadze, Y. O. Chernoff, M. Borodovsky, Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979-1990 (2008).
72. D. Hyatt, G. L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 1-11 (2010).
73. M. Kanehisa, Y. Sato, K. Morishima, BlastKOALA and GhostKOALA: KEGG tools for

- functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726-731 (2016).
74. C. Bland, T. L. Ramsey, F. Sabree, M. Lowe, K. Brown, N. C. Kyrpides, P. Hugenholtz, CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinform.* **8**, 1-8 (2007).
 75. R. C. Edgar, PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinform.* **8**, 1-6 (2007).
 76. M. Huntemann, N. N. Ivanova, K. Mavromatis, H. J. Tripp, D. Paez-Espino, K. Palaniappan, E. Szeto, M. Pillay, I. M. A. Chen, A. Pati, T. Nielsen, The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v. 4). *Stand. Genom. Sci.* **10**, 1-6 (2015).
 77. D. Laslett, B. Canback, ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11-16 (2004).
 78. A. I. Rissman, B. Mau, B. S. Biehl, A. E. Darling, J. D. Glasner, N. T. Perna, Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* **25**, 2071-2073 (2009).
 79. T. A. Knijnenburg, S. A. Ramsey, B. P. Berman, K. A. Kennedy, A. F. Smit, L. F. Wessels, P. W. Laird, A. Aderem, I. Shmulevich, Multiscale representation of genomic signals. *Nat. Methods* **11**, 689-694 (2014).

Acknowledgments:

The work conducted by the U.S. Department of Energy Joint Genome Institute (SR) is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231.

Funding:

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A2A16055779), and by Ministry of Science and ICT (NRF-2017R1A2B4008968, NRF-2019R1A2C1090861).

Competing interests:

Authors declare no competing interests.

Data and materials availability:

The raw Illumina and SMRT sequence files used in this study were deposited in the NCBI BioProject under the accession PRJNA784005. Analysis codes of the meta-epigenomics are available at the repository (<https://github.com/hoonjeseong/Meta-epigenomics>) under a MIT license.

Supplementary Materials

Materials and Methods

Supplementary Text

fig. S1 to S5

table S1 to S5

Fig. 1. Meta-epigenome analysis scheme of ocean surface samples. A schematic overview of meta-epigenomics. Meta-epigenomics using genome-centric metagenomics from the binning approach of short- and long-read assemblies, followed by identifying the epigenetic signals of genomes from long-read mapping.

Fig. 2. Phylogenetic tree of MAGs obtained from SHIPPO. (A) A phylogenetic tree of prokaryotic SHIPPO MAGs using core genes from PhyloPhlan2. A total of 252 strain-level MAGs were obtained; each bar outside the tree represents the number of methyltransferase (MTase) genes present in each MAG. (B) The distribution of restriction enzyme and MTase types from SHIPPO MAGs across kingdoms. (C) Prokaryotic SHIPPO MAGs were compared against genomes from *Tara* Oceans (TARA) and Global Ocean Reference Genomes Tropics (GORG-Tropics) datasets using FastANI. (D) The number of genes associated with the restriction–modification (RM) system is plotted against the genome size for each ocean microbiome MAG (SHIPPO MAG, TARA, and GORG-Tropics). Points are shaped depending on the type of the complete and orphan RM system. MAG: metagenome-assembled genome; SHIPPO: Shipborne Pole-to-Pole Observations.

Fig. 3. Meta-epigenomic profile of MAGs across all sampling stations. (A) Principal coordinate analysis (PCoA) clustering by the 5-mer methylation features of Shipborne Pole-to-Pole Observations (SHIPPO) MAGs based on Kulczynski dissimilarity. Each point represents each species-level MAG in each sample. The black-dashed circles represent family-level clusters of MAGs across samples. The colored-solid lines represent species-level clusters belonging to *Pelagibacteraceae* across samples. (B) The maximum methylation ratios of motifs are represented in each MAG at the family level; the highest methylation value among all sample sites is colored. (C) Population differentiation versus methylome across sampling stations. For the most prevalent Shipborne Pole-to-Pole Observations (SHIPPO) MAGs, scatterplots show the relationship of 5-mer methylome dissimilarity based on Bray–Curtis and population differentiation by sampling distance. MAG: metagenome-assembled genome (eukaryotic: eukMAG, prokaryotic: proMAG, viral MAG: vMAG, giant viral

MAG: gvMAG); F_{ST} : fixation index.

Fig. 4. Genome-wide epigenetic analysis of *Pelagibacter* MAG across samples. (A) The UpSet plot compares GANTC methylation at each genome position on a *Pelagibacter* MAG (SHIPPO_PRO_33) across samples. More than 0.5 of the methylation fractions were considered methylated at each genome position; the color of each bar depends on the genic (G: green), intergenic (I: navy), and regulatory region (R: orange). The column bar indicates the intersection of the number of methylation positions on the MAG across samples. The left bar represents the total number of methylated positions on the MAG for each sample. (B) Principal coordinate analysis (PCoA) clustering by SNMV and SNV based on the Bray–Curtis distance on overlapped regions for all samples. (C) A model of the methylation pattern according to the cell-cycle progression in *Alphaproteobacteria*. (D) The methylation fraction comparisons of the GANTC motif between genomic regions of *ori* (replication origin), *ter* (replication terminus), and other regions. (E) The genome-wide distribution of methylated fractions for the GANTC motif indicates the trend of cell-cycle progress throughout the genome. MAG: metagenome-assembled genome; SHIPPO: Shipborne Pole-to-Pole Observations; SNMV: single nucleotide methylation variation; SNV: single nucleotide variant.

Fig. 5. DNA methylation of the viral genome. (A) The methylome of 83 prevalent viral Shipborne Pole-to-Pole Observations (SHIPPO) of metagenome-assembled genomes (vMAGs) is indicated by the heatmap with their phylogeny from genome similarities. The star represents vMAGs in (C). (B) Phylogenetic comparison of the 14 vMAGs harboring MTase and its MTase genes. (C) Changes in vMAG methylation profiles were measured at three sampling stations. Circles represent the methylation ratio of each motif, and bars represent the read mapping breadth of viral genomes with 10× depth.

Fig. 6. Genomic characterization between *Pelagibacter* and pelagiphage according to the GANTC motif. (A) Distribution of GANTC motif methylation in both strands of *Cand. Pelagibacter*

Giovannoni NP1 genome. The inner blue and red bars indicate the coding sequence region of the genome. The methylation fraction bar for each strand is red for values ≥ 0.9 , orange for ≥ 0.6 , and bright yellow for ≥ 0.3 . The prophage region of the genome is highlighted by a purple bar from Pleška *et al.* (39). The GATC motif depletion region is marked through a multiscale signal representation (MSR) analysis. **(B)** Phylogenetic tree of pelagiphages of the head–tail connector protein. Red boxes represent the GATC depletion subgroups of pelagiphage. **(C)** The GATC motif density comparisons between genomes of *Pelagibacter* and pelagiphage.

Metagenomics

Genome-centric metagenomics

Meta-epigenomic analysis

A**Phylum** **MTase type**

Tree scale: 0.1

B **C** **D**

A **B** **C** **D** **E**