Nucleic Acids Research, 2025, **53**, gkaf226
https://doi.org/10.1093/nar/gkaf226
Methods

Proxi-RIMS-seq2 applied to native microbiomes uncovers hundreds of known and novel ^{m5}C methyltransferase specificities

Weiwei Yang[1,†], Yvette Luyten[1,†], Emily Reister[2], Hayley Mangelson[2], Zach Sisson[2], Benjamin Auch[2], Ivan Liachko[2], Richard J. Roberts[1], Laurence Ettwiller[1,*]

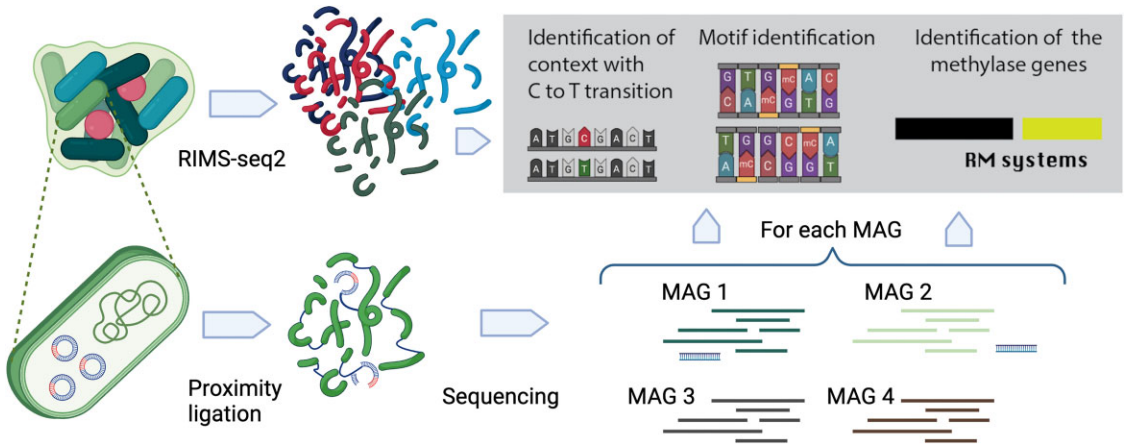
[1]New England Biolabs, Inc., 240 County Road, Ipswich, MA 01938, United States
[2]Phase Genomics, Inc., 1617 8th Ave N, Seattle, WA 98109, United States

[*]To whom correspondence should be addressed. Email: ettwiller@neb.com
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Methylation patterns in bacteria can be used to study restriction–modification or other defense systems with novel properties. While ^{m4}C and ^{m6}A methylation are well characterized mainly through PacBio sequencing, the landscape of ^{m5}C methylation is under-characterized. To bridge this gap, we performed RIMS-seq2 (rapid identification of methyltransferase specificity sequencing) on microbiomes composed of resolved assemblies of distinct genomes through proximity ligation. This high-throughput approach enables the identification of ^{m5}C methylated motifs and links them to cognate methyltransferases directly on native microbiomes without the need to isolate bacterial strains. Methylation patterns can also be identified on bacteriophage DNA and compared with host DNA, strengthening evidence for phage-host interactions. Applied to three different microbiomes, the method unveiled over 1900 motifs that were deposited in REBASE. The motifs include a novel eight-base recognition site (CAT^{m5}CGATG) that was experimentally validated by characterizing its cognate methyltransferase. Our findings suggest that microbiomes harbor arrays of untapped ^{m5}C methyltransferase specificities, providing insights into bacterial biology and biotechnological applications.

Graphical abstract



Introduction

The role of methylation, an important epigenetic mark, extends beyond the well-known restriction–modification (RM) systems in bacteria, playing roles in the orchestration of gene expression and other cellular mechanisms [1]. While ^{m4}C and ^{m6}A methylation patterns have been extensively identified, particularly through the use of Pacific Biosciences (PacBio) single-molecule sequencing, our understanding of the landscape of ^{m5}C methylation and methyltransferase specificities in bacteria lags behind. Efforts utilizing PacBio [2], Tet-assisted PacBio sequencing [3], and more recently, nanopore sequencing technology have begun to bridge this gap [4]. However, these methods require substantial amounts of native DNA and have limited throughput, posing significant hurdles.

Conversely, high-throughput short-read sequencing enables the rapid sequencing of entire microbiomes at very high

Received: October 12, 2024. **Revised:** December 13, 2024. **Editorial Decision:** February 24, 2025. **Accepted:** March 24, 2025

coverage, expanding the study of microbiomes to a broad range of starting materials and versatile selection protocols. In addition to sequencing genomes, methylome information can be obtained through short-read sequencing. However, the most common approach to overlay methylation to sequence information involves deamination of cytosine to uracil using bisulfite or enzymatic treatments. These treatments significantly alter the DNA sequence, making them incompatible with mixtures of nonmodel organisms typically observed in microbiomes for which no reference genomic sequences are available.

Recently, rapid identification of methyltransferase specificity sequencing (RIMS-seq [5]) has been developed to sequence nonmodel organisms, such as bacterial genomes, and simultaneously determine ^{m5}C methyltransferase specificities without requiring a reference genome. However, RIMS-seq is limited in its application to native microbiomes because short-read assemblies do not result in long enough contigs to assemble complete genomes from complex metagenomes.

To address this complexity, proximity ligation metagenomic sequencing has emerged as a strategy to deconvolute complex microbial communities [6–8]. This technique involves crosslinking DNA within intact cells prior to lysis, preserving the spatial information about sequences originating within the same cells. This information is recovered through ligating digested ends centered on a crosslink site, and the resulting chimeric molecules are analyzed via paired-end sequencing. When combined with a metagenomic assembly, these chimeric reads can be used to link contigs from the same consensus genome together into very high-quality bins (referred to throughout the manuscript as metagenome-assembled genomes or MAGs) and additionally to associate mobile elements like phages and plasmids with their microbial hosts without culturing.

In response to these challenges, we developed Proxi-RIMS-seq2, which combines an improved version of RIMS-seq and proximity ligation technologies to simultaneously assess the genetic and ^{m5}C epigenetic information on genome-resolved microbiomes. To exemplify the potential of this methodology, we applied Proxi-RIMS-seq2 to three highly diverse microbiomes and linked the methylated motifs to their cognate methyltransferases.

Materials and methods

Microbiome genomic DNA source

We performed RIMS-seq on four synthetic or native microbiome genomic DNA samples, including: (i) a mock gut microbiome genomic mix containing an equal mixture of 12 fully sequenced and authenticated bacterial species observed in the gut microbiome (ATCC, MSA-1006); (ii) human oral microbiome (Phase Genomics); (iii) human fecal microbiome (ZymoBIOMICS Fecal Reference, Zymo Cat #D6323); and (iv) vermiculture microbiome (Phase Genomics).

Genome-resolved microbiomes

A Hi-C library was created with the Phase Genomics ProxiMeta Hi-C version 4.0 Kit using the manufacturer-provided protocol [9]. Briefly, intact cells from two samples were crosslinked using a formaldehyde solution, simultaneously digested using the Sau3AI (^GATC) and MluCI (^AATT) restriction enzymes, and proximity ligated with biotinylated nucleotides to create chimeric molecules composed of fragments from different regions of genomes that were physically proximal *in vivo*. Proximity ligated DNA molecules were pulled down with streptavidin beads and processed into an Illumina-compatible sequencing library. Separately, using an aliquot of the original sample, DNA was extracted with a ZymoBIOMICS DNA Miniprep Kit (Zymo Research, #D4300), and a metagenomic shotgun library was prepared using ProxiMeta library preparation reagents. Sequencing was performed on an Illumina NovaSeq, generating PE150 read pairs for both Hi-C and shotgun libraries. Hi-C and shotgun metagenomic sequencing files were uploaded to the Phase Genomics cloud-based bioinformatics portal for subsequent analysis.

Shotgun reads were filtered and trimmed for quality and normalized using fastp [10] and then assembled with MEGAHIT [11, 12] using default options. Hi-C reads were then aligned to the assembly following the Hi-C kit manufacturer's recommendations (https://phasegenomics.github.io/2019/09/19/hic-alignment-and-qc.html). Briefly, reads were aligned using BWA-MEM [13] with the −5SP options specified and all other options default. SAMBLASTER [14] was used to flag polymerase chain reaction (PCR) duplicates that were later excluded from analysis. Alignments were then filtered with Samtools [15] using the −F 2304 filtering flag to remove nonprimary and secondary alignments. Metagenome deconvolution was performed with ProxiMeta [16], resulting in the creation of putative genome and genome fragment clusters. Clusters were assessed for quality using CheckM [17] and assigned preliminary taxonomic classifications with Mash [18].

Viral annotation, binning, and host association

Viral contigs were identified using VIBRANT (version 1.2.1) [19] with default settings. Putative viral contigs with bacterial and viral sequences present are annotated as prophages if 50% or more of the contig length is annotated as viral; otherwise, they are annotated as microbial.

ProxiPhage [14] viral binning utilizes a hybrid strategy that combines proximity ligation signal clustering with traditional metagenomic binning methods to address the limitations of each approach when used independently. An overlap network was then constructed where nodes represented contig clusters from each set, and edges indicated shared contigs between clusters. This network was processed using a proprietary greedy network collapse algorithm within the ProxiMeta platform to merge overlapping clusters into a single, refined set of vMAGs.

The quality of viral contigs and vMAGs was assessed using CheckV (version 1.0.1) [20]. Both viral and plasmid host assignments were carried out using the ProxiPhage host attribution tool. Long-range Hi-C linkages between viral contigs and their prokaryotic host genomes were analyzed to assign likely hosts. Viral- and plasmid-host linkages were filtered based on Hi-C linkage strength, connectivity ratios, and intra-MAG connectivity, using default settings for the ProxiPhage algorithm.

RIMS-seq library preparation and illumina sequencing

To prepare individual DNA libraries for RIMS-seq, we used genomic DNA (gDNA) input amounts ranging from 50 ng to 80 ng. The microbiome gDNAs were first sheared to 250 bp

with the Covaris S2 Focused Ultrasonicator. One reaction of the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB, #E7645) was used for each gDNA sample. Adaptor ligated libraries were purified with 1× volume of NEBNext Sample Purification Beads (NEB, #E7103) and eluted with 45 µl 0.1× Tris-EDTA buffer. NaOH was then added to the purified DNA libraries at 1 M with a 30 min incubation time at 60°C, followed by addition of an equimolar amount of acetic acid to neutralize the reaction. We included a uracil-specific excision reagent (USER, NEB, #M5505) treatment step to improve RIMS-seq library quality. The detailed protocol was published in our earlier work [21].

Prepared libraries were amplified and indexed with NEBNext Multiple Oligos for Illumina (NEB, #E6446). Short-read sequencing was processed on an Illumina NextSeq instrument with paired-end reads of 75 bp.

Reads are mapped to the MAGs reference sequences using BWA-MEM [22] and processed as described here [5].

Imbalance in known contexts

A total of 44 IUPAC motifs, known to be recognition sites for at least 10 distinct ^{m5}C methyltransferases cataloged in REBASE, were selected, and the positions of the methylated bases were recorded. The imbalance value was computed for all positions matching known motifs within a MAG. This value was calculated by subtracting the total number of C-to-T conversions in read 1 (R1) and G-to-A conversions in read 2 (R2) from the total number of C-to-T conversions in R2 and G-to-A conversions in R1, then normalizing to the total number of C-to-T conversions observed, using the following equation:

Imbalance value = [(C-to-T in R1 + G-to-A in R2) − (C-to-T in R2 + G-to-A in R1)] / [(C-to-T in R1 + G-to-A in R2) + (C-to-T in R2 + G-to-A in R1)]

Instances of C-to-T or G-to-A conversions were identified using mpileup output. High-quality conversions were considered when the base quality score was ≥35 for C-to-T or G-to-A when compared to the reference assembly within a MAG. To avoid considering positions that contain true genetic variants, any position where the percentage of C-to-T or G-to-A conversions exceeded 5% for at least five reads was ignored. Programs and a detailed manual for the *de novo* identification of motifs in Proxi-RIMS-seq2 are available on GitHub (https://github.com/Ettwiller/RIMS-seq).

Clustering of the motifs and MAGs based on the imbalance values was done using the pheatmap package (version 1.0.12) with the following options: clustering_distance_rows = "manhattan," clustering_distance_cols = "minkowski."

Proxi-RIMS-seq2 *de novo* motif identification

Using the mpileup files, ±7-bp flanking genomic regions (15 bp total) for which a high-quality (base quality score ≥ 35) C-to-T in R1 or G-to-A in R2 was found were extracted for the foreground. Positions and ±7-bp flanking genomic regions (15 bp total) for which a high-quality (base quality score ≥ 35) G-to-A in R1 or C-to-T in R2 was found were extracted for the background. C-to-T or G-to-A in the first position of reads was ignored. If the percentages of C-to-T or G-to-A were above 5% for at least five reads at any given position, the position was ignored (to avoid considering positions containing true variants).

De novo motif discovery in the mock microbiome was performed using mosdi-discovery using the following parameters: "mosdi-discovery -v discovery -q x -i -T 1e-100 -M 8,2,0,4 8 occ-count" using the foreground sequences with x being the output of the following command: "mosdi-utils count-qgrams -A dna using the background sequences. To identify additional motifs, the most significant motif found using mosdi-discovery is removed from the foreground and background sequences using the following parameter: "mosdi-utils cut-out-motif -M X," and the motif discovery process is repeated until no significantly enriched motif can be found. A motif is found significantly enriched in the foreground sequence if *P*-value < 1e−100.

De novo motif discovery in the microbiomes has been performed using DiNAMO. For each MAG, ±7-bp flanking genomic regions (15 bp total) for which a high-quality (base quality score ≥ 35) C-to-T in R1 or G-to-A in R2 was found were extracted for the foreground (file_foreground.fasta). Positions and their ±7-bp flanking genomic regions (15 bp total) for which a high-quality (base quality score ≥ 35) G-to-A in R1 or C-to-T in R2 was found were extracted for the background (file_background.fasta). DiNAMO was run using the following parameters: dinamo -pf file_foreground.fasta -nf file_background.fasta -l 8 -t 1

Identification, cloning, and expression of DNA methyltransferase genes

Putative DNA methyltransferase gene sequences were identified using HMMER [23] for the Pfam domain PF00145.20 (C-5 cytosine-specific DNA methylase) [24]. Methyltransferase target recognition domains (TRDs) were identified using the conserved signature motifs identified by Posfai *et al.* [25].

Methyltransferases sequences for MAGs 10 and MAGs 51:

>M_MAGs_10 VKSSNAKSSVVQSQSQIECVDLFCGIG GLTSGLAKGGIKVNAGIDVDEDCKFAYEKNNDAQFI LKDI SSLTGPQIRKFFGEGSISLLAGCAPC QPFSTYSRKS RKTKEDDKWSLVLHFGRLIKRAQPTLVTMEN V PQLI HHQVFDDFLKSLKGYSVWWKVVDCSELGIPQTRKR LVLLASKLGSIKLLEPDLSEPSTVRSA IEGLMPLEAGSC DPKDSLHAASDLSDINLRRIRASKPGGTWRDWNKSLLA KCHRKKSGATFPSVYGR MEWDKPSPTITTQCFGFGN GRFGHPEQDRAISLREAALIQTFPSDYCFLAPGEKLVFAK LGRLIGNAVPVKLGEIIANSLITHVDTYRT*

>M_MAGs_51 MIACIDFFCGVG GLTHGLVRGG VKVVAGVDVDPLCKYPYEANNEAKFIEQDVKQLS ATDLLPLWPKKSVTMLAGCAPC QPFSTYSRRGRM ERADAKWELAKEFGRLINQLQPKLVTMENV PQLA DHSVFKDFTKSLKGYNLWHGIVECSKYGVPQTRKR LVLLASKLGPISLAKPSIDESHPPTVRSAITGLLRLKAGE YDSSDYLHWACKLSPLNLKRIKASKPGGTWRDWDESLI SECHKKDTGETYPSVYGRMEWDKPSPTITTQCFGYGN GRFGHPEQNRAITLREAAILQTFPKGYKFLPKGELPTF AGMGRLIGNAVPVRLGEVIAQTVVAHVADHS*

Gene blocks for cloning M_MAGs_10 and M_MAGs_51 methyltransferase genes were synthesized by Integrated DNA Technology (Coralville, IA), and codons were optimized for expression in *Escherichia coli*. More specifically, the M_MAGs_10, a 1071-bp methyltransferase gene, was synthesized as two gene blocks (574 bp and 608 bp) containing overlaps designed for NEBuilder HiFi assembly. The M_MAGs_51, a 1080-bp methyltransferase gene, was synthesized as a single gene block containing overlaps designed for NEBuilder HiFi assembly.

The expression plasmid pACYC184 [26] was prepared by inverse PCR using Q5 Hot Start High Fidelity DNA Polymerase (NEB #M0493) according to the manufacturer's protocol. Forward and reverse primers (For: 5′-GCATGCACCATTCCTTGCGG-3′ and Rev: 5′-GGATCCACAGGACGGGTGTG-3′) were added at a final concentration of 400 nM. Cycling parameters were as follows: initial denaturation at 98°C for 30 s, followed by 25 cycles of 98°C for 10 s, 70°C for 20 s, and 72°C for 2.5 min, and a final extension step at 72°C for 2 min. PCR amplification was confirmed by visualization on a 0.8% agarose gel stained with ethidium bromide. PCR product was digested with 10 units of DpnI restriction enzyme (NEB #R0176) for 30 min at 37°C to digest methylated template DNA, followed by cleanup using the NEB Monarch nucleic acid purification kit (NEB #T1030) according to the manufacturer's instructions. Purified PCR product was quantified using the Qubit dsDNA broad range assay kit (Invitrogen, #Q32853).

The DNA methyltransferase genes were cloned in the pACYC184 expression vector under the control of a constitutive Tet promoter using NEBuilder HiFi DNA assembly master mix (NEB #E2621). Methyltransferase gene blocks were mixed in a 2:2:1 molar ratio (M_MAGs_10) or 2:1 molar ratio (M_MAGs_51) with PCR-amplified vector, followed by incubation at 50°C for 15 min as recommended by the manufacturer. Two microliters of NEBuilder HiFi reaction was transformed into NEB Express competent *E. coli* cells (#C2523; g*fhuA2 [lon] ompT gal sulA11 R (mcr-73::miniTn10*–Tet^S*)2 [dcm] R (zgb-210::Tn10*–Tet^S*) endA1 Δ (mcrC-mrr)114::IS10*). Individual colonies were selected and grown overnight in LB broth supplemented with chloramphenicol (25 μg/ml). Plasmid DNA and total DNA were isolated from overnight cultures using the Monarch® plasmid miniprep kit (NEB, #T1010) and the Monarch® genomic DNA purification kit (NEB, #T3010), respectively. Plasmid DNAs were sequence-verified by Oxford Nanopore Technology (ONT) sequencing using the EPI2ME clone validation workflow (https://github.com/epi2me-labs/wf-clone-validation) and total DNA was prepared for sequencing on PacBio Sequel II (Pacific Biosciences, Menlo Park, CA).

PacBio sequencing

In vivo modification activity and sequence specificity were analyzed by sequencing on the PacBio Sequel II. Prior to library preparation, input DNA was sheared to an average size of 5–10 kb using gTubes (Covaris, Woburn, MA) and concentrated using 0.6V Ampure beads (PacBio). Libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 (PacBio, #100-938-900) according to the manufacturer's protocol. Barcoded libraries were TET-treated following an abbreviated protocol for the NEBNext Enzymatic Methyl-Seq Conversion Module (NEB, #E7120) to enzymatically convert modified cytosines into 5-carboxylcytosine. Briefly, libraries were incubated for 1 h at 37°C with TET2 in the presence of TET2 reaction buffer, an oxidative supplement, DTT, and an oxidation enhancer, followed by the addition of stop reagent and incubation at 37°C for 30 min. TET2-treated libraries were cleaned up using 1V Ampure beads (PacBio).

TET-treated, barcoded libraries were prepared for sequencing on PacBio Sequel II using Sequel II Sequencing Kit 2.0 (PacBio, #101-820-200). PacBio SMRT Link web portal sample setup and run design calculators were used to determine polymerase binding and instrument loading conditions, respectively. Post-sequencing analyses, including the circular consensus sequencing and Microbial Genome Analysis programs, were performed using the SMRT Link (v11.1.0.166339), applying the default analysis parameters.

Result

RIMS-seq2 accurately identifies methylated context on a mock gut microbiome

Earlier versions of RIMS-seq and RIMS-seq2 were employed on mock microbiomes and human genomic DNA, respectively [5, 21]. We first tested RIMS-seq2 on the mock gut microbiome (ATCC® MSA-1006™) spiked with XP12 bacteriophage genomic DNA, where all cytosines have been replaced by ^{m5}C. RIMS-seq2 has a higher deamination rate compared to RIMS-seq and would require significantly less coverage per genome. Sequencing reads were downsampled to 1 million paired-end reads and were mapped to the mock community reference genomes. We calculated the imbalance of C-to-T transition between paired-end reads 1 and 2, established previously to be linearly correlated with methylation [5, 21]. Using XP12 as our deamination control, we achieved around a 1.42% deamination rate on ^{m5}C, which is consistent with the previous report of RIMS-seq2 [21] (Supplementary Fig. S1A). This deamination level is large enough to detect the ^{m5}C methylase specificity but too low to affect sequencing and assembly quality.

For each genome in the mock community, we applied mosdi-discovery [27], as described previously [5] to *de novo* identify motifs predicted to be methylated. Using this approach, we identified all the methylation motifs that were previously detected by RIMS-seq and validated by bisulfite sequencing (Supplementary Table S1). Additionally, motifs were discovered with significantly fewer reads than required by RIMS-seq. For instance, the GATC motif was identified with only 46 000 reads mapping to the *Bifidobacterium adolescentis* genome (Supplementary Table S1).

These results indicate that RIMS-seq2 can replace RIMS-seq for identifying methyltransferase specificities in mock microbiomes composed of 12 bacterial isolates at a fraction of the sequencing depth. However, it is important to note that the mock gut microbiome used in this study consists of an equal mixture by weight of a limited number of bacterial isolates with available high-quality reference genomes. Consequently, this setup is not an accurate representation of native microbiomes.

Proxi-RIMS-seq2 identifies contexts of elevated C-to-T transitions across phased-resolved microbiomes

Next, we performed RIMS-seq2 on two distinct native microbiomes: human oral and earthworm (vermicompost), for which shotgun sequencing and proximity ligation (ProximeMeta) were performed to resolve assemblies into genomes (Proxi-RIMS-seq2). Additionally, we evaluated a reference fecal microbiome (ZymoBIOMICS TruMatrix™ Fecal Reference) that was collected from healthy adult donors and homogenized in one large batch. This reference composite microbiome was sequenced using PacBio and proximity ligation to obtain a high-quality, genome-resolved reference fecal microbiome [28]. For all experiments, we aimed for an

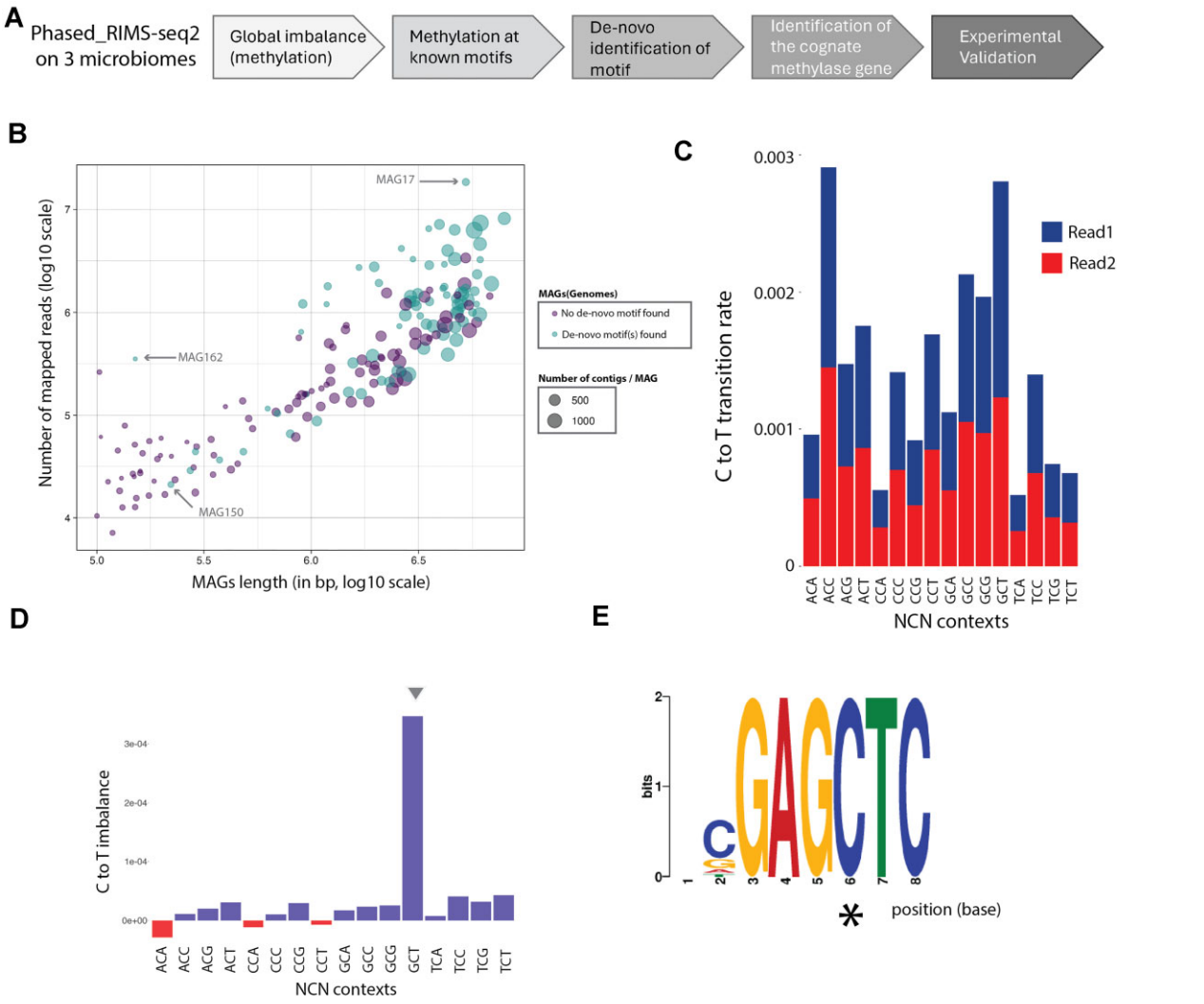


Figure 1 (**A**) Overview of the Proxi-RIMS-seq2 analytic steps: (i) Imbalances in MAGs indicate methylation. (ii) Imbalance at known motifs. (iii) Imbalance is used to *de novo* identify methylated motifs. (iv) Linking genotype (MAGs) with phenotype (methylation) associates methyltransferases with their predicted recognition motifs. (v) Methyltransferase activities can be cloned and expressed *in vivo* in strains lacking ^{m5}C methylation for validation. (**B**) Relationship between MAG length (in base pairs, *x*-axis) and read coverage (log_{10} scale, *y*-axis) within the genome-resolved vermicompost microbiome. Each data point represents an MAG, distinguished by the presence (green) or absence (purple) of predicted methylated motifs. Notably, MAG 17, belonging to the genus *Patulibacter*, exhibits the highest read coverage mapped to its consensus genome. (**C**) Bar plot representing the absolute C-to-T transition rate in R1 (blue) and R2 (red) at all 16 NCN contexts (with N = A,T,C, or G) in MAG 17. (**D**) Differential C-to-T transition rate between R1 and R2 (imbalance value) at all 16 NCN contexts in MAG 17. (**E**) PWM found to be most significantly associated with C-to-T imbalance in MAG 17 (* indicates the methylated cytosine).

estimated 1% deamination rate at methylated sites (Fig. 1A, see the "Material and methods" section). Proximity ligation libraries were produced using Phase Genomics' ProxiMeta kits, and the resulting libraries were sequenced on the Illumina NovaSeq X platform. Metagenomic assembly was performed using MEGAHIT, and MAG deconvolution and analysis was performed using the ProxiMeta platform, with phage and plasmid genomes being reconstructed as described [29]. Sequencing of RIMS-seq2 libraries yielded ~200 million read pairs for the vermicompost, 160 million read pairs for the dental microbiome, and 700 million read pairs for the fecal dataset. RIMS-seq2 reads were mapped to the genome-resolved reference sequences using BWA-MEM [22].

Using the vermicompost microbiome, we first assessed whether the imbalance between the C-to-T transition on R1 compared to R2 in RIMS-seq2 libraries could be observed despite the high heterogeneity of sequences typically found in population consensus genomes. Imbalance has been previously shown to represent damage rather than variants from the reference sequence [30]. In RIMS-seq, an imbalance of C-to-T represents a deamination of ^{m5}C. Given that a typical microbiome is highly complex, we focused on a representative of the *Patulibacter* genus, whose assembly is 92% complete and has the highest read coverage in the RIMS-seq2 sequencing dataset (Fig. 1B). Compared to mock microbiomes, we observed relatively higher C-to-T transition rates in all 16 three-base contexts (NCN with N = A,T,C, or G) for both R1 and R2 (Fig. 1C), presumably due to the heterogeneity captured in the population consensus genome. Despite this higher baseline variant rate, a prominent imbalance was observed in the GCT context (Fig. 1D), indicating methylation in this context or a related context.

Next, we searched the *Patulibacter* genome for genes containing an ^{m5}C methyltransferase domain using HMMER

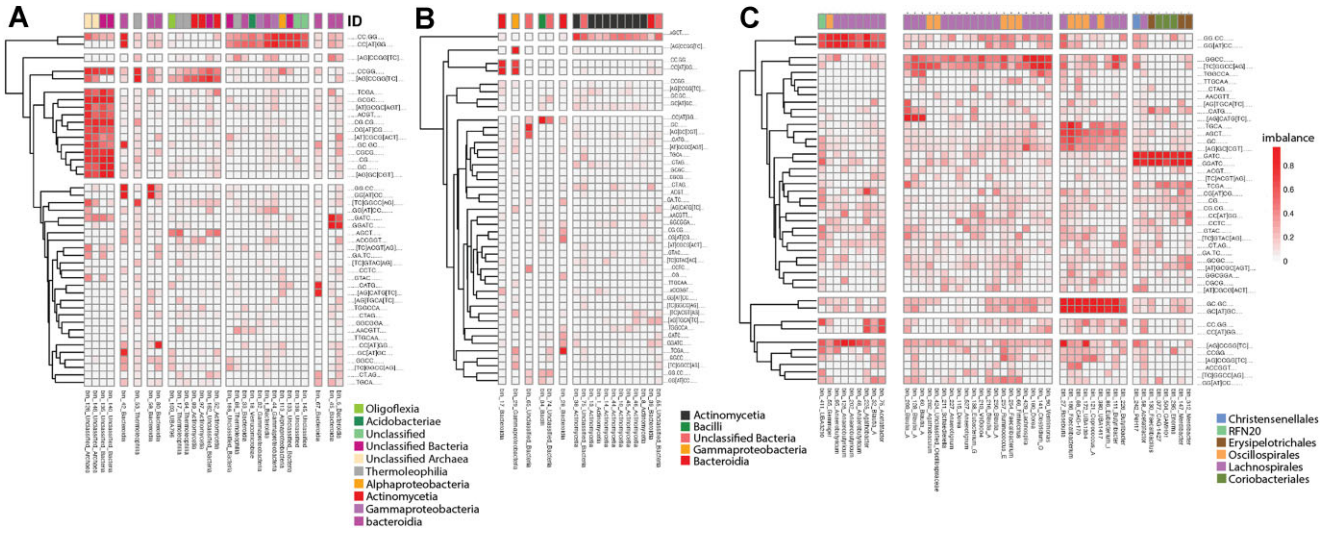


Figure 2. Imbalance profiles for selected clusters of MAGs (full dataset in Supplementary Fig. S2) in (**A**) vermicompost, (**B**) dental, and (**C**) fecal microbiome. At an imbalance of 1%, all cytosines at the underlined positions are predicted to be methylated. The *x*-axis corresponds to the annotated MAGs, *y*-axis corresponds to known recognition sites of ^{m5}C methyltransferases (as cataloged in REBASE), and *z*-axis (gray to red) indicates imbalance values, ranging from 0 to 1%. MAGs are color-coded according to their taxonomic order. Both MAGs and motifs are clustered based on their imbalance profiles. Rows are clustered using the Manhattan distance, and columns are clustered using the Minkowski distance, as implemented in Pheatmap.

[23]. A single hit containing an ^{m5}C methyltransferase domain was found in the assembly. A homology search in REBASE [31] identified the closest hits as M.Sgr13350I Type II methyltransferase with known activity at GAGCTC sites [31]. Knowing the expected motif content, we then sought to analyze the imbalance in all six-base NNNCNN contexts. The result reveals significant imbalance only in the GAGCTC context at ∼1% (Supplementary Fig. S1B and C). This result is consistent with the expected deamination rate for RIMS-seq2 and the predicted methyltransferase specificity in *Patulibacter*. Finally, to evaluate whether such motifs could have been identified *de novo* without prior knowledge of the predicted methylation motif, we used DiNAMO, a motif-searching algorithm that uses an exact discriminative method for discovering IUPAC motifs in DNA sequences [32]. DiNAMO was supplied with all 15-bp sequences flanking a C-to-T transition events in R1 (foreground) and R2 (background) and identifies a position weight matrix (PWM) containing the GAGCTC as the most significant motif associated with an excess of C-to-T transitions in R1 (Fig. 1E).

Taken together, these results indicate that C-to-T imbalance can be used to [i] assess whether a genome has ^{m5}C methylation, [ii] identify known methylated context(s), [iii] *de novo* identify methylated motif(s), and [iv] predict the association between the identified motif(s) and their cognate methyltransferase genes in individual genomes recovered from metagenomes. While this has already been demonstrated for individual genomes, we have now established that a similar strategy can be employed directly from complex genome-resolved microbiomes, despite the high genotypic and phenotypic variability typically observed in population consensus genomes.

Bacteria with similar methylation profiles tend to belong to the same order

Next, we examined the methylation profile of all three microbiome samples at 44 motifs known to be methylated by previously characterized methyltransferases from REBASE (see the "Materials and methods" section). For this, we computed for each MAG the imbalance value in all these known motifs and subsequently clustered both the MAGs and motifs according to these imbalance profiles.

We observed that genomes within a cluster often belong to bacteria of the same order (Fig. 2 and Supplementary Fig. S2). For instance, in the dental microbiome, methylation at AGCT motifs was predominantly found in the order Actinomycetales (Fig. 2B). In the fecal microbiome, methylation at GGNCC and GGWCC motifs was commonly found in the order Lachnospirales (Fig. 2C). While many methylation profiles are shared by related bacteria of the same order, other profiles were observed among evolutionarily divergent bacteria. For example, CCWGG and CCNGG contexts were methylated in a broad range of evolutionary diverse bacteria in the vermicompost (Fig. 2A).

Proxi-RIMS-seq2 *de novo* identifies ^{m5}C methylated motifs and their cognate methyltransferases directly on native microbiomes

Next, for each MAG, motifs associated with a C-to-T imbalance in the RIMS-seq2 paired-end libraries were identified using DiNAMO (see the "Materials and methods" section). Accordingly, these motifs are predicted to be methylated in their respective MAGs. Most of the MAGs composed of assemblies of at least 1×10^6 bp have at least one motif predicted to be methylated (Fig. 1B and Supplementary material). Nonetheless, methylated motifs can be predicted in MAGs with as low as ∼20 000 reads (i.e. vermicompost, MAGs 150) or MAGs as small as 150 000 bp of total sequences. For example, with only 150 843 bp and 5.49% completeness, we were still able to predict the two methylated nonoverlapping motifs GCCGGC and GAGCTC in vermicompost MAGs 162.

A total of 75 motifs, 166 motifs, and 1707 motifs were found in the 81 oral microbiome MAGs (0.92 motifs per MAG), the 176 vermicompost MAGs (0.94 motifs per MAG), and the 738 composite gut microbiome MAGs

(2.3 motifs per MAG), respectively (see Supplementary material). Motifs, methyltransferases, and MAGs were deposited into REBASE (http://rebase.neb.com/rebase/rebase.html). We conducted several manual inspections of the link between PWM and genotype data, often leading to successful matches even in complex cases involving multiple methyltransferase specificities. For example, we discovered five methylated motifs in MAG 4 with closest resemblance to *Algoriphagus terrigena* in the vermicompost microbiome (Supplementary Fig. S3). Six genes or gene fragments encoding ^{m5}C methyltransferase domains were identified in the *A. terrigena* genome assemblies. Of these, four genes could be confidently linked to their cognate methylated motifs based on sequence identity to previously characterized ^{m5}C methyltransferases with experimentally validated specificities [31]. The remaining motif (C<u>C</u>WGG) could not be assigned to a corresponding gene, likely due to insufficient sequence similarity to known methyltransferases or the presence of gaps in the assembled genome.

These results indicate that *de novo* discovery of methylated motifs can be performed directly on native genome-resolved microbiomes, even when only a small fraction of the genomic sequences is available. Crucially, the binning of genomes allows for the linkage of genotypes to their epigenetic states, thereby enabling the association of methyltransferases to their predicted activities.

Validation of newly identified methyltransferase specificities from Proxi-RIMS-seq2 data

The systematic association of novel methyltransferases with their sequence specificities at the microbiome level holds significant biotechnological potential, particularly in identifying enzymes with novel sequence specificities. To demonstrate the applicability of this strategy, we searched for novel motifs of interest that are predicted to be methylated.

One such newly identified methylation motif is CAT<u>C</u>GATG, which would correspond to a novel 8-bp recognition site. Methyltransferases recognizing such extended motifs are of significant biotechnological interest due to their potential association with restriction enzymes of similar specificity. This motif has been detected in two binned genomes from the vermicompost MAGs 10 and 51, corresponding to the genera *Oligoflexus* and *Anatilimnocola*, respectively. Using HMMER [23], we searched for genes containing a ^{m5}C methyltransferase domain and identified a single gene per binned genome. Pairwise comparison between *Oligoflexus* and *Anatilimnocola* methyltransferases shows 63.66% amino acid identity across the entire protein and 68% identity across the TRD, which would suggest similar specificity. The closest methyltransferase homolog with known sequence specificities to those in *Oligoflexus* and *Anatilimnocola* is M.Bbr28III from *Bifidobacterium breve* with 48% and 52% identity, respectively. M.Bbr28III has a known RT<u>C</u>GAY (with R = G or A and Y = C or T) motif specificity (partial of C**ATCGAT**G) that has been confirmed by bisulfite sequencing [31].

To experimentally validate the methylation specificities of these newly identified methyltransferases, we cloned and *in vivo* expressed them in an *E. coli* strain lacking endogenous ^{m5}C methylation while conserving ^{m6}A methylation at G<u>A</u>TC sites (NEB Express C2523 strain, Dcm−; Dam+, see the "Materials and methods" section). The *Oligoflexus* and *Anatilimnocola* methyltransferases were successfully cloned, and genomic DNA from the corresponding recombinant clone was sequenced using Tet-assisted PacBio SMRT-seq to assess methylation patterns [3]. No C methylation can be detected for the *Anatilimnocola* clone. In contrast, the *Oligoflexus* clone exhibited notable interpulse duration (IPD) ratio both at the G<u>A</u>TC motif, consistent with ^{m6}A methylation by the Dam methyltransferase, and at the CAT<u>C</u>GATG motif, indicative of ^{m5}C methylation (Fig. 3A and B). Using DiNAMO, we confirmed the presence of both the ^{m5}C modification at CAT<u>C</u>GATG and ^{m6}A methylation at G<u>A</u>TC motifs in the recombinant strain (Fig. 3C). These findings demonstrate the activity of the *Oligoflexus* m5C methyltransferase, which specifically targets the 8-bp recognition sequence CAT<u>C</u>GATG, as predicted by Proxi-RIMS-seq2.

Proxi-RIMS-seq2 on viral DNA strengthened phage-host association

Proximity ligation provides linkage information to enable the association of phages with their hosts. The physical interactions between phage DNA and the DNA of its microbial host can be captured *in vivo* using proximity ligation, and the ProxiPhage algorithm [29] is then used to reconstruct the vMAG and link it to its microbial host. We independently analyzed the predicted methylation profiles in both host and viral vermicompost MAGs and compared the identified motifs in each phage/host pair.

Out of 92 vMAGs, 13 have at least one motif predicted to be methylated in both the MAGs and their corresponding vMAGs. Out of these 13 vMAGs, 10 have a good match to at least one motif from their predicted host (Supplementary Fig. S4), and two have partial matches. These results indicate that the phages/hosts are sharing some of their methylation patterns and that the phasing of these interactions is reliable. In most of the cases, either the host or the bacteriophages have additional motifs, possibly indicating a more complex stratification in the phages/host association than population consensus genomes can resolve.

Conclusion

We developed Proxi-RIMS-seq2 to identify ^{m5}C motifs directly in microbiomes and have demonstrated the ability to *de novo* detect hundreds of such motifs across three distinct microbiomes. In some instances, we observed that bacteria within a microbiome and belonging to the same order share similar methylated motifs. This observation suggests the existence of a shared or stable epigenetic state that may facilitate genetic material exchange within the order. Further work is required to elucidate the extent of this phenomenon and the underlying mechanisms.

Two of these microbiomes were derived from unique samples, while the third, the fecal microbiome, is a microbial reference material composed of stool samples collected from numerous healthy donors. Application of Proxi-RIMS-seq2 to the fecal dataset revealed a higher number of methylation contexts per MAG compared with the other two microbiomes, suggesting a wide array of methyltransferase specificities. Given that the fecal dataset is a meta-metagenome sourced from multiple donors, this observed diversity likely re-

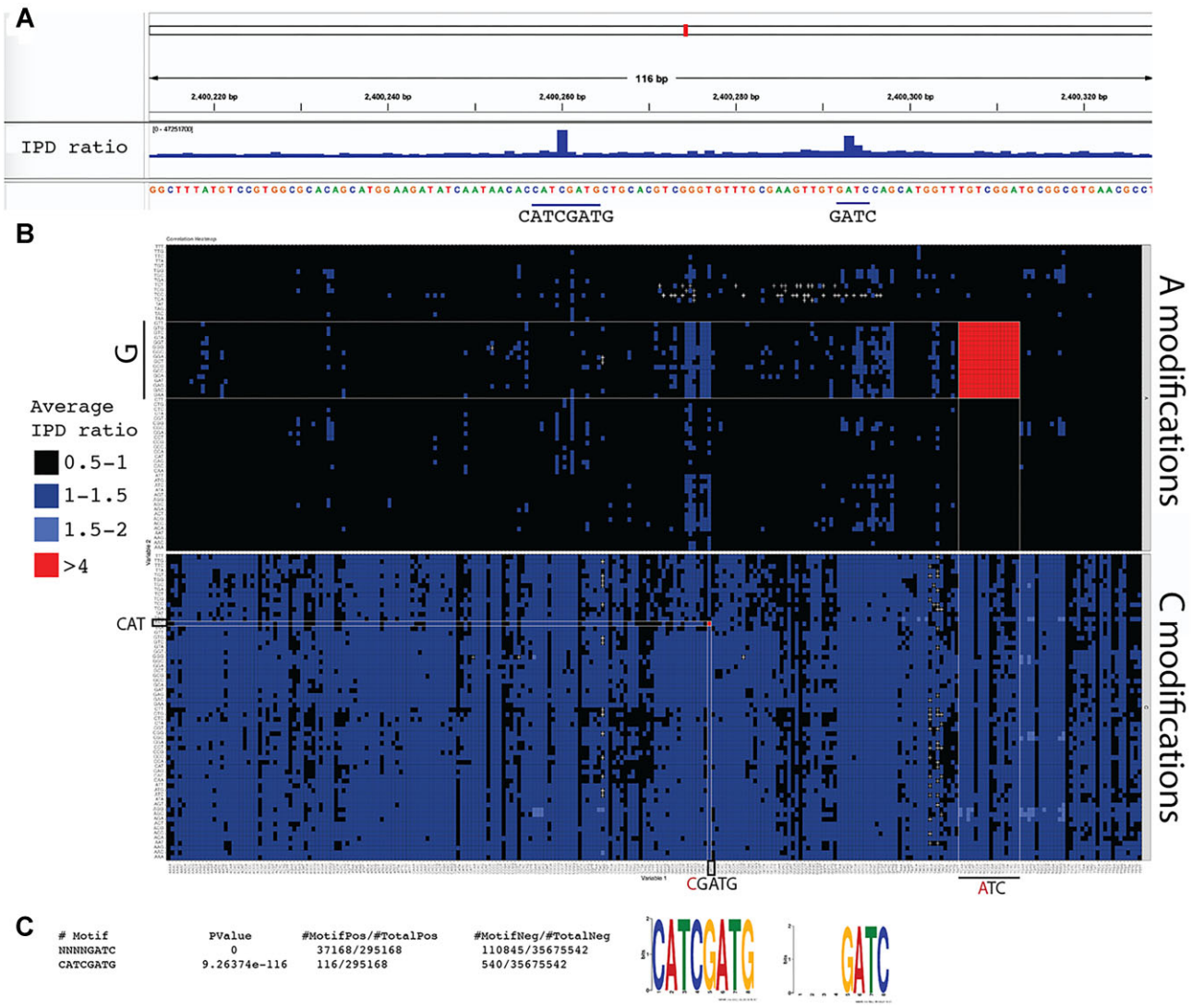


Figure 3. Validation of a new methyltransferase specificity using Tet-assisted PacBio sequencing. (**A**) IPD ratios at a specific *E.coli* locus containing both CATCGATG and GATC motifs. (**B**) Average IPD ratio for all possible 8 mers for A modifications (^{m6}A, upper panel) and C modifications (^{m4}C or ^{m5}C, lower panel). *Y*-axis corresponds to the first three nucleotides (oriented 3′–5′), *x*-axis corresponds to the last four nucleotides (orientated 5′–3′), and *z*-axis (color) corresponds to the average IPD ratio. (**C**) *De novo* identification of motifs (output of DiNAMO and PWM).

flects the natural variability of methyltransferase genes across different individuals rather than an increased number or specificity of methyltransferase genes within a single stool.

MAGs represent population consensus genomes, and the degree of similarity in the epigenetic states within such populations remains unknown. Consequently, the observed methylation motifs likely originate from methyltransferases that are common and stable among a substantial portion of individual bacteria within these populations. Thus, the identified motifs in the studies may be biased toward stable epigenetic marks, such as the well-studied Dcm methylation in *E. coli* [33]. Consistent with this statement, we often observed that methylated motif profiles are shared amongst related bacteria within the same order.

This study expands the potential applications of metagenomic data. For example, DNA methylation has been previously used as a complementary feature to enhance metagenomic binning [34]. Similarly, Proxi-RIMS-seq2 can be employed to validate genome binning from proximity ligation data. Beyond improving binning accuracy, Proxi-RIMS-seq2 can be used to investigate the dynamics of methylation directly within microbiomes and to identify novel methyltransferase specificities, as demonstrated in this study with the identification of a new CAT^{m5}CGATG recognition motif and its associated methyltransferases. This methyltransferase appears to function within an RM system, and we are currently characterizing the associated predicted restriction enzyme.

We have shown that it is possible to replace the conventional shotgun library with RIMS-seq while maintaining similar sequence accuracy and assembly statistics [5]. Similarly, RIMS-seq2 could substitute for the conventional shotgun library that is required for the proximity ligation protocol. In this case, we would achieve an efficient integration of methylation data with genome-resolved microbiome information, at minimal additional cost. Our study demonstrates that, despite decades spent searching for novel methyltransferase specificities for biotechnological applications, microbiomes still harbor hidden methyltransferase specificities yet to be discovered.

Acknowledgements

We are grateful for the NGS core sequencing group at NEB for the sequencing of RIMS-seq libraries, Charles Elfe for providing a convenient download of REBASE, Tamas Vincze for incorporating the data to REBASE, Brian Anton for useful comments, Shuang-Yong Xu for advice and analysis of the eight base-recognition restriction modification loci, Peter Weigele and Yian-Jiun Lee for providing Xp12 bacteriophage genomic DNA, Colleen Yancey for critical reading of the manuscript, and Alexey Fomenkov for advice on PacBio experiments.

Author contributions: Weiwei Yang (Data curation, Formal analysis, Investigation, Methodology, Validation, Writing—review & editing), Yvette Luyten (Investigation, Methodology, Writing—review & editing) Ivan Liachko (Conceptualization, Resources, Supervision, Funding acquisition, Writing—review & editing) Benjamin Auch (Conceptualization, Supervision, Project administration, Writing—review & editing) Emily Reister (Investigation, Methodology, Resources) Zach Sisson (Software, Formal analysis, Visualization) Hayley Mangelson (Software, Formal analysis, Visualization, Writing—review & editing) Richard Roberts (Methylation analysis), and Laurence Ettwiller (Conceptualization, Resources, Supervision, Formal analysis, Visualization, Funding acquisition, Writing – review & editing)

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

W.Y., Y.L., R.R., and L.E. are employees of New England Biolabs, Inc., a manufacturer of restriction enzymes and molecular biology reagents. E.R., H.M., Z.S., B.A., and I.L. are employees of Phase Genomics, the developer of metagenomic proximity ligation technology.

Funding

This work was supported in part by New England Biolabs, Inc., by NIH SBIR Grant R44AI172703 and a grant from the Bill & Melinda Gates Foundation to Phase Genomics. Funding to pay the Open Access publication charges for this article was provided by New England Biolabs, Inc.

Data availability

The raw data underlying this article are available in NCBI SRA (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1166287). MAGs, methyltransferase genes, and motifs were deposited in REBASE (http://rebase.neb.com/rebase/rebase.html). Code for C-to-T imbalance calculation can be found in https://github.com/Ettwiller/RIMS-seq or https://figshare.com/articles/software/RIMS-seq/27189405?file=49673466.

References

1. Anton BP, Roberts RJ. Beyond restriction modification: epigenomic roles of DNA methylation in prokaryotes. *Annu Rev Microbiol* 2021;**75**:129–49. https://doi.org/10.1146/annurev-micro-040521-035040
2. Seong HJ, Roux S, Hwang CY *et al.* Marine DNA methylation patterns are associated with microbial community composition and inform virus-host dynamics. *Microbiome* 2022;**10**:157. https://doi.org/10.1186/s40168-022-01340-w
3. Clark TA, Lu X, Luong K *et al.* Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biology* 2013;**11**:4. https://doi.org/10.1186/1741-7007-11-4
4. Tourancheau A, Mead EA, Zhang X-S *et al.* Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat Methods* 2021;**18**:491–8. https://doi.org/10.1038/s41592-021-01109-3
5. Baum C, Lin Y-C, Fomenkov A *et al.* Rapid identification of methylase specificity (RIMS-seq) jointly identifies methylated motifs and generates shotgun sequencing of bacterial genomes. *Nucleic Acids Res* 2021;**49**:e113. https://doi.org/10.1093/nar/gkab705
6. Marbouty M, Cournac A, Flot J-F *et al.* Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife* 2014;**3**:e03318. https://doi.org/10.7554/eLife.03318
7. Burton JN, Liachko I, Dunham MJ *et al.* Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3* 2014;**4**:1339–46. https://doi.org/10.1534/g3.114.011825
8. Bickhart DM, Kolmogorov M, Tseng E *et al.* Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nat Biotechnol* 2022;**40**:711–9. https://doi.org/10.1038/s41587-021-01130-z
9. Lieberman-Aiden E, van Berkum NL, Williams L *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**:289–93. https://doi.org/10.1126/science.1181369
10. Chen S, Zhou Y, Chen Y *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90. https://doi.org/10.1093/bioinformatics/bty560
11. Li D, Liu C-M, Luo R *et al.* MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**:1674–6. https://doi.org/10.1093/bioinformatics/btv033
12. Li D, Luo R, Liu C-M *et al.* MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 2016;**102**:3–11. https://doi.org/10.1016/j.ymeth.2016.02.020
13. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 2010;**26**:589–95. https://doi.org/10.1093/bioinformatics/btp698
14. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* 2014;**30**:2503–5. https://doi.org/10.1093/bioinformatics/btu314
15. Li H, Handsaker B, Wysoker A *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9. https://doi.org/10.1093/bioinformatics/btp352
16. Stewart RD, Auffret MD, Warr A *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun* 2018;**9**:870. https://doi.org/10.1038/s41467-018-03317-6
17. Parks DH, Imelfort M, Skennerton CT *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;**25**:1043–55. https://doi.org/10.1101/gr.186072.114
18. Ondov BD, Treangen TJ, Melsted P *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;**17**:132. https://doi.org/10.1186/s13059-016-0997-x
19. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020;**8**:90. https://doi.org/10.1186/s40168-020-00867-0

20. Nayfach S, Camargo AP, Schulz F *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 2021;**39**:578–85. https://doi.org/10.1038/s41587-020-00774-7
21. Yan B, Wang D, Ettwiller LM. Simultaneous assessment of human genome and methylome data in a single experiment using limited deamination of methylated cytosine. *Genome Res* 2024;**34**:904–13. https://doi.org/10.1101/gr.278294.123
22. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009;**25**:1754–60. https://doi.org/10.1093/bioinformatics/btp324
23. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;**14**:755–63. https://doi.org/10.1093/bioinformatics/14.9.755
24. Mistry J, Chuguransky S, Williams L *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* 2021;**49**:D412–9. https://doi.org/10.1093/nar/gkaa913
25. Pósfai J, Bhagwat AS, Pósfai G *et al.* Predictive motifs derived from cytosine methyltransferases. *Nucl Acids Res* 1989;**17**:2421–35. https://doi.org/10.1093/nar/17.7.2421
26. Chang AC, Cohen SN. Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. *J Bacteriol* 1978;**134**:1141–56. https://doi.org/10.1128/jb.134.3.1141-1156.1978
27. Marschall T, Rahmann S. Efficient exact motif discovery. *Bioinformatics* 2009;**25**:i356–64. https://doi.org/10.1093/bioinformatics/btp188
28. Portik DM, Feng X, Benoit G *et al.* Highly accurate metagenome-assembled genomes from human gut microbiota using long-read assembly, binning, and consolidation methods. bioRxiv, https://doi.org/10.1101/2024.05.10.593587, 11 May 2024, preprint: not peer reviewed.
29. Uritskiy G, Press M, Sun C *et al.* Accurate viral genome reconstruction and host assignment with proximity-ligation sequencing. bioRxiv, https://doi.org/10.1101/2021.06.14.448389, 14 June 2021, preprint: not peer reviewed.
30. Chen L, Liu P, Evans TC Jr *et al.* DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* 2017;**355**:752–6. https://doi.org/10.1126/science.aai8690
31. Roberts RJ, Vincze T, Posfai J *et al.* REBASE: A database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* 2023;**51**:D629–30. https://doi.org/10.1093/nar/gkac975
32. Saad C, Noé L, Richard H *et al.* DiNAMO: highly sensitive DNA motif discovery in high-throughput sequencing data. *BMC Bioinformatics* 2018;**19**:223. https://doi.org/10.1186/s12859-018-2215-1
33. Marinus MG, Morris NR. Isolation of deoxyribonucleic acid methylase mutants of *Escherichia coli* K-12. *J Bacteriol* 1973;**114**:1143–50. https://doi.org/10.1128/jb.114.3.1143-1150.1973
34. Beaulaurier J, Zhu S, Deikus G *et al.* Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol* 2018;**36**:61–9. https://doi.org/10.1038/nbt.4037

Received: October 12, 2024. **Revised:** December 13, 2024. **Editorial Decision:** February 24, 2025. **Accepted:** March 24, 2025