Mapping DNA Methylation to Methyltransferases

in Microbial Communities

Jeppe Støtt Bøjer

Aalborg University

Department of Chemistry and Bioscience

Frederik Bajers Vej 7H, 9220 Aalborg Ø, Denmark

Abstract

Across all domains of life, the genetic code is overlaid with epigenetic modifications that extend beyond the primary nucleotide sequence. The most common and nearly universal mechanism of epigenetic signaling is DNA methylation. In bacteria, it modulates a range of biological processes, including host defense mechanisms, cell cycle regulation, gene expression, and virulence. This modification is facilitated by DNA methyltransferases, which dictate the methylation patterns of bacterial genomes in a motif-specific manner, often differing among species and strains. Recent technological advances in Nanopore sequencing now enable the direct detection of DNA methylation from a standard sequencing run. Despite this, only a few efforts have been made to utilize ONT methylation calls for methylation motif discovery in bacteria, but none which scales or extends motif discovery to metagenome sequencing of microbial communities. To address this, we developed Nanomotif, a fast, scalable, bioinformatic tool for identification and utilization of methylation motifs in metagenomic samples. The MTase-linker submodule of Nanomotif replaces existing manual and non-scalable methods with a modern, user-friendly bioinformatics tool that pairs methylation motifs to their cognate DNA methyltransferases. In the era of metagenomics, tools like this are essential for faster epigenetic profiling across entire microbial communities. Motif-methyltransferase pairs not only help circumvent restriction-modification barriers but also open new avenues to explore the functional roles of methylation and its implications for microbial physiology and ecology.

Title: Mapping DNA Methylation to Methyltransferases in Microbial Communities

Semester: 9th & 10th

Project period: September 1st, 2023, to August 1st, 2025

ETCS: 60

Supervisor: Mads Albertsen

Total Pages: 55

Supplementary Pages: 0

Jeppe Støtt Bøjer

Jeppe Støtt Bøjer

Preface

This thesis is based on the following study:

Nanomotif: Leveraging DNA Methylation Motifs for Genome Recovery and Host Association of Plasmids in Metagenomes from Complex Microbial Communities

Søren Heidelbach, Sebastian Mølvang Dall, Jeppe Støtt Bøjer, Jacob Nissen, Lucas N.L. van der Maas, Mantas Sereika, Rasmus H. Kirkegaard, Sheila I. Jensen, Sabrina Just Kousgaard, Ole Thorlacius-Ussing, Katja Hose, Thomas Dyhre Nielsen, Mads Albertsen. *bioRxiv 2024.04.29.591623; doi: https://doi.org/10.1101/2024.04.29.591623*

The latest version of the manuscript is included in the last chapter of this thesis. Most of the work was carried out between September 2023 and June 2024. My main contribution was the development and implementation of the MTase-Linker module, which is the focus of the third chapter. I further participated in data analysis and interpretation, and was actively involved in manuscript preparation, critical revision, and final submission.

To provide context for the study, the first chapter introduces the broader field of prokaryotic epigenomics.

References are notated according to Cite Them Right 12th edition – Harvard.

I am extremely grateful to have worked under the guidance and encouragement of my supervisor, Professor Mads Albertsen. I would also like to thank Søren Heidelbach and Sebastian Mølvang Dall for their invaluable support in the conceptual design of the MTase-Linker module, assistance with analyzing Nanopore sequencing data, and their advice on scientific writing and coding practices. Finally, I sincerely thank the entire Albertsen Lab for their support, enthusiasm, and thoughtful scientific discussions throughout this project.

Abbreviations

Abbreviation	Full description	Abbreviation	Full description
A	Adenine	*Mod*	MTase gene Type III
Agn43	Antigen 43 gene	*Modtype*	DNA modification type
bp	Base pairs	MQ	Medium quality
C	Cytosine	ONT	Oxford Nanopore Technology
Dam	DNA adenine methyltransferase	OxyR	Oxidative stress regulator/
DNA	Deoxyribonucleic acid	PacBio	Pacific Biosciences
G	Guanine	PPM	Positional probability matrix
gtr	Glycosyltransferase operon	R/REase	Restriction enzyme
HMM	Hidden Markov model	*Res*	Restriction gene Type III
hsdM	MTase gene Type I	RM system	Restriction-modification system
hsdR	Restriction gene Type I	T	Thymine
hsdS	Sequence recognition gene Type I	S	Sequence recognition subunit
HQ	High quality	SAM	*S*-adenosyl-L-methionine
IPD	Inter-pulse duration	4mC	N4-methylcytosine
KL-divergence	Kullback–Leibler divergence	5mC	C5-methylcytosine
MAG	metagenome-assembled genome	6mA	N6-methyladenine
M/MTase	DNA methyltransferase		

Table of Contents

Abstract ..... **i**

Preface ..... **ii**

Abbreviations ..... **iii**

Table of Contents ..... **iv**

Introduction **1**

The Prokaryotic Epigenome 2

RM systems: Guardians of the Genome 3

Gene Regulation by DNA Methylation 5

Deciphering the Epigenetic Code 6

Aim and Objectives **8**

MTase-Linker **9**

Annotating MTase Genes 10

Linking MTase Genes with Motifs 11

Homology-based Recognition Motif Prediction 12

RM System Type Predicts Motif Type 13

Modification Type Prediction 13

Conclusion **14**

References **15**

Manuscript **20**

Nanomotif: Leveraging DNA Methylation Motifs for Genome Recovery and Host Association of Plasmids in Metagenomes from Complex Microbial Communities 21

Abstract 21

Main 21

Data availability 24

Code availability 24

Acknowledgements 24

Ethics 25

Materials And Methods 28

References 33

Supplementary Figures 36

Supplementary Note 1 45

Supplementary Note 2 48
Supplementary Note 3 48

Introduction

Invisible to the naked eye, microbes make up the vast majority of Earth's life forms (Hug *et al.*, 2016; Timmis *et al.*, 2017). They inhabit every environment where macroscopic organisms exist, and they are the sole life forms in extreme environments like the deep trenches and acidic hot springs (Brock, 1985; Cavicchioli *et al.*, 2019). Dating back at least 3.8 billion years to the origin of life on Earth, microbes have become essential to many vital processes (Cavicchioli *et al.*, 2019). They are central to carbon and nutrient cycling, play essential role in soil structure and fertility, act as key producers and sinks of greenhouse gases, and influence various physiological activities in humans, animals and plants (Timmis *et al.*, 2017; Cavicchioli *et al.*, 2019). Without microbes, life as we know it would rapidly cease to exist on Earth (Gilbert and Neufeld, 2014).

Today, both academic and industrial sectors are investing significant effort into remediating the environmental damage caused by human activity. Yet, despite their fundamental role in sustaining ecosystems across the biosphere, microbes remain largely overlooked in these efforts (Cavicchioli *et al.*, 2019; Crowther *et al.*, 2024). As the earliest forms of life on Earth, microbes have evolved remarkable evolutionary, functional, and metabolic diversity - offering a vast and largely untapped toolkit for tackling both current and future environmental challenges (Timmis *et al.*, 2017). This immense microbial potential is not entirely unfamiliar to humanity. In fact, microbes have been harnessed since the dawn of civilization for everyday processes such as brewing beer, fermenting cheese and wine and baking bread (Buchholz and Collins, 2013; Timmis *et al.*, 2017). With the discovery of DNA as the blueprint of life in the 1960s (Watson and Crick, 1953), microbial technology and product development have surged, paving the way for industrial and pharmaceutical advancement including drug and enzyme production (Buchholz and Collins, 2013). Despite the discovery of microbes more than 300 years ago, the vast majority of microbial diversity and functionality still remains unexplored (Gest, 2004; Albertsen, 2023; Singleton *et al.*, 2024).

Historically, the study of microbes has relied heavily on isolating and studying them in pure culture within laboratory settings. However, access to microbes by cultivation methods is still quite limited, and the vast majority of microbial diversity continues to elude cultivation (Rinke *et al.*, 2013). As a result, modern microbial research increasingly depends on genomic data, particularly whole genomes and marker genes, which are readily obtained through DNA extraction and sequencing. These genomic elements serve as the fundamental units for exploring microbial diversity, evolution, and function (Pérez-Cobas, Gomez-Valero and Buchrieser, 2020; Albertsen, 2023).

To date, most genomic investigations have focused on the primary sequence of DNA - the arrangement of the four nucleotide bases: guanine (G), cytosine (C), adenine (A), and thymine (T). Technologies for sequencing the primary DNA code have advanced rapidly over the past two decades, enabling the decoding of genetic information for a vast diversity of macro- and microorganisms (Hug *et al.*, 2016; Satam *et al.*, 2023). Though, still far from the estimates of millions (Louca *et al.*, 2019) to billions (Larsen *et al.*, 2017) of microbial species expected to constitute the Earths biosphere, ~144,000 prokaryotic species now have a genomic representation in the public databases (Parks *et al.*, 2022). This growing genomic catalog continues to expand our understanding of microbial life and holds tremendous promises for unlocking novel biotechnological innovations and applications.

While the primary DNA sequence play a well-established role in biology, other genomic features that are equally vital to microbial physiology have received comparatively less attention and remain poorly understood (Hofer, Liu and Balasubramanian, 2019; Sánchez-Romero and Casadesús, 2020). Although DNA modifications were discovered in bacteria more than half a century ago, limited methodological advancements long hindered comprehensive exploration in this area (Beaulaurier, Schadt and Fang, 2019). Recent breakthroughs in third-generation sequencing technologies, however, have overcome many of these limitations - enabling genome-wide detection of multiple types of DNA modifications at single-nucleotide resolution (Nielsen *et al.*, 2023). Researchers are now better equipped than ever before to uncover the previously hidden epigenetic mechanisms, which govern the physiology and functional capabilities of microbes.

The Prokaryotic Epigenome

The most common and nearly universal mechanism of epigenetic signaling is DNA methylation (Sánchez-Romero and Casadesús, 2020). In prokaryotes, DNA methylation occurs in three forms: C5 and N4 cytosine methylation (5mC and 4mC) and N6 adenine methylation (6mA), where 6mA is the most common form (**Figure 1**) (Sánchez-Romero and Casadesús, 2020).

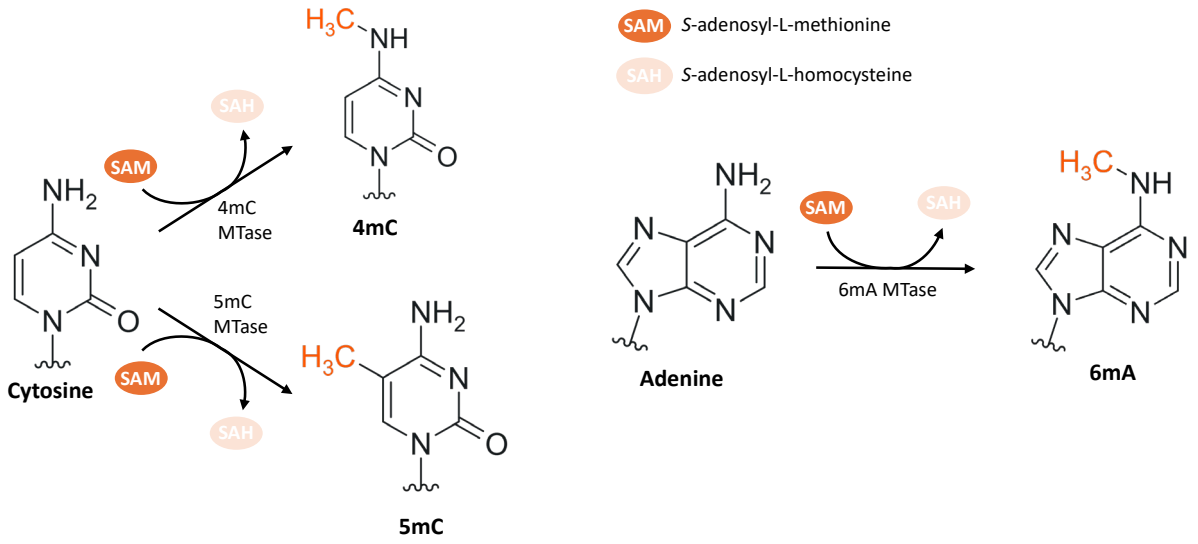


Figure 1. Primary types of DNA methylation in prokaryotes. Chemical structures of the three primary types of DNA methylation found in prokaryotes: C5-methylcytosine (5mC), N4-methylcytosine (4mC), and N6-methyladenine (6mA). These modifications are catalyzed by DNA methyltransferases (MTases), which transfer a methyl group from *S*-adenosyl-L-methionine (SAM) to the appropriate position on the unmodified target base. Adapted from (Beaulaurier, Schadt and Fang, 2019).

DNA is methylated by DNA methyltransferases (MTases), which transfer a methyl group from *S*-adenosyl-L-methionine to the appropriate position on the target base. Based on the position to which the methyl group is transferred, MTases can be divided into two classes, exocyclic amino MTases (4mC and 6mA) and endocyclic MTases (5mC) (Gao *et al.*, 2023). These enzymes methylate specific DNA sequence contexts, called motifs, and create unique methylation patterns on prokaryotic genomes. For example, the *Escherichia coli* K-12 strain encodes three active MTases, recognizing 5'-GATC-3', 5'-CCWGG-3', 5'-AACNNNNNNGTGC-3'

motifs (Adhikari and Curtis, 2016). Nearly every occurrence of these target motifs is methylated, and typically, if a DNA motif is recognized by an MTase, more than 95% of its occurrences are modified (Beaulaurier, Schadt and Fang, 2019). DNA methylation is both abundant and widespread across prokaryotic taxa. The specificity domain of MTases, which determines the recognized motif, varies greatly among species, contributing to a remarkable diversity of methylation patterns throughout the prokaryotic kingdom (Blow *et al.*, 2016; Beaulaurier, Schadt and Fang, 2019). A landmark study investigating a diverse set of 230 prokaryotic genomes observed motifs in 93% of genomes with an average of 3 methylated motifs per genome (Blow *et al.*, 2016).

RM systems: Guardians of the Genome

Historically, prokaryotic DNA methylation and MTases have been associated with restriction-modification systems (RM systems), consisting of MTases and cognate restriction enzymes (REases) (Roberts, 2003; Seong, Han and Sul, 2021). The RM systems are the most abundant and widespread antiphage system, present in more than 80% of prokaryotic genomes, with an average of more than two RM systems per genome (Oliveira, Touchon and Rocha, 2014; Tesson *et al.*, 2022). It protects the host against foreign DNA elements by distinguishing non-methylated foreign DNA from its own methylated DNA. The former being recognized and cleaved by the cognate REase (**Figure 2**) (Ershova *et al.*, 2015).

RM systems have been classified into four main types, I, II, III, and IV based on their subunit composition, cleavage site, sequence recognition motif and cofactor requirements (Ershova *et al.*, 2015). In **Figure 3**, genes and subunit compositions involved in the first three RM system types are illustrated.

Type I RM systems are encoded by three genes: *hsdM* (MTase gene), *hsdS* (sequence recognition gene), and *hsdR* (restriction gene). The Type I MTase is a complex composed of two MTase subunits and one S subunit (M_2S_1, **Figure 3**), whereas the type I REase is composed of two MTase subunits and two restriction subunits with one S subunit to form a complex ($R_2M_2S_1$, **Figure 3**). The methylation of Type I RM systems occurs on both strands of a bipartite motif, for example 5'-AACNNNNNNGTGC-3'. Cleavage by the REase complex occurs up to several kilobases away from the bipartite recognition motif (Ershova *et al.*, 2015).

Type II RM systems are usually produced by two genes (*M* and *R*, **Figure 3**) that encode an REase and MTase, respectively. MTases are active as monomers, but REases are composed of various complexes ranging from monomers to tetramers. They mostly bind to short (4-8 base pair (bp)) palindromic sequences, like 5'-GATC-3'. Methylation occurs inside the motif on both strands, and cleavage occurs either inside or nearby the recognition motif (Ershova *et al.*, 2015). A specific subtype of Type II system, Type IIG, is produced by a single gene that encodes a protein with both restriction and methyltransferase activity. These typically recognize non-palindromic motifs either as a short, continuous sequence or a bipartite, discontinuous sequence (Roberts *et al.*, 2023).

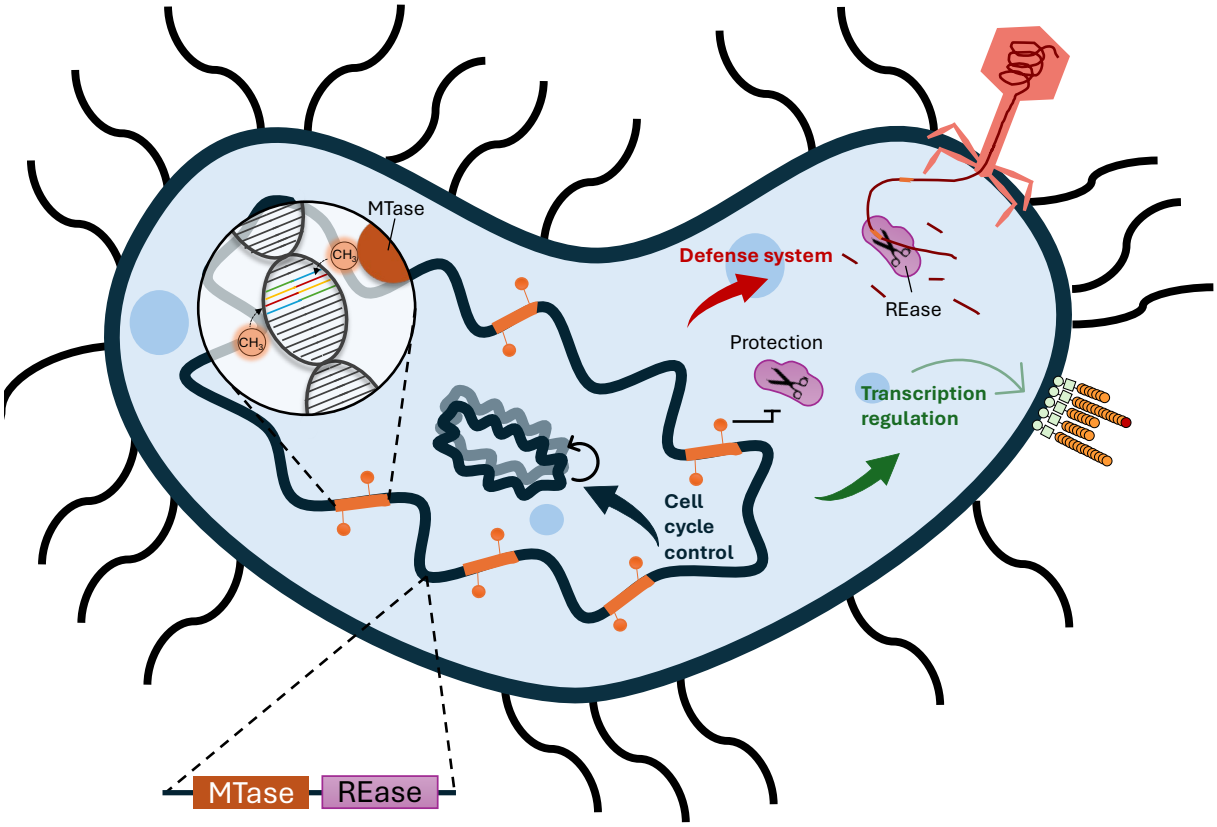


Figure 2. **Roles of Prokaryotic DNA Methylation.** In prokaryotes, restriction-modification (RM) systems serve as widespread defense mechanisms against foreign DNA. These systems distinguish self from non-self DNA through methylation patterns. The Type II RM system, illustrated here, includes a DNA methyltransferase (MTase, orange) and a restriction enzyme (REase, purple). Both enzymes recognize short, palindromic sequence motifs in the genome (thick orange lines). Unmethylated recognition sites, such as those found in an infecting phage genome, are cleaved by the REase. Motif specific methylation by the MTase protects the host genome from cleavage by the REase. In addition to their role in genome defense, several MTases regulate important cellular process such as transcription and cell cycle. Figure inspired by (Seong, Han and Sul, 2021; Wilbanks et al., 2022).

Type III RM systems consist of two *mod* and *res* genes encoding proteins that recognize, modify, and cleave specific DNA sequence motifs. Type III MTases are composed of two Mod-subunits (M_2, **Figure 3**), and because only Mod-subunits contain the DNA-binding specific domain, a type III REase consists of a complex with a Res and two Mod-subunits (R_1M_2, **Figure 3**). Type III RM systems bind to short (4-6 bp) non-palindromic motifs, e.g. 5'-CGAAT-3', and methylate only one DNA strand after binding (Ershova *et al.*, 2015).

Finally, unlike other types of RM systems, the Type IV RM systems comprise only the REase, which hydrolyze methylated DNA. The Type IV RM system has evolved to have low sequence specificity (unlike other RM systems) to protect host cells from a wide range of foreign DNA with various methylation patterns (Ershova *et al.*, 2015).

In addition to the RM systems, many bacterial and archaeal genomes harbor at least one Type II MTase, not associated with any REase (Oliveira, Touchon and Rocha, 2014; Blow *et al.*, 2016). These MTases are designated orphan MTases. Unlike RM MTases, which are poorly conserved, many orphan DNA methyltransferases are conserved at the genus level (Oliveira and Fang, 2021; Gao *et al.*, 2023). An example of an conserved, orphan MTase, is the *E.coli* Dam enzyme, methylating 5'-GATC-3', which have homologs widespread in γ-Proteobacteria (Oliveira and Fang, 2021).

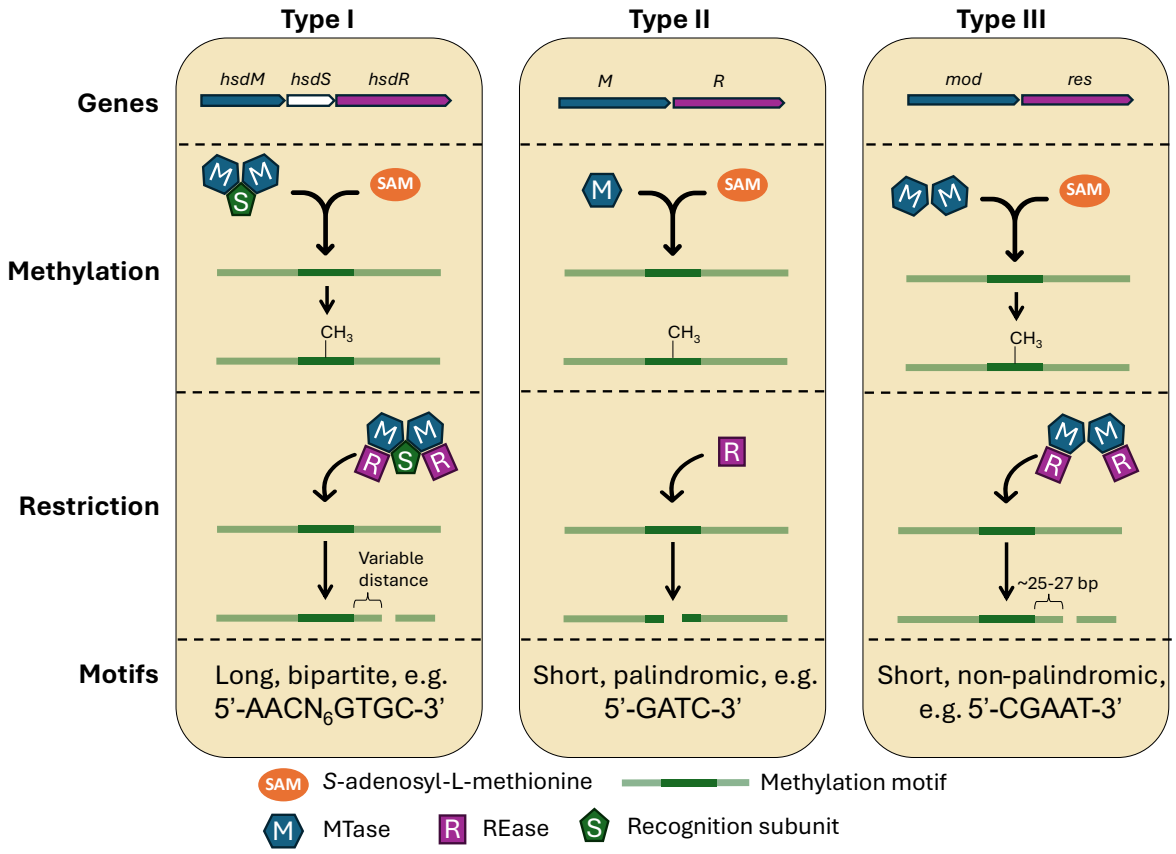


Figure 3. Structural and Functional Organization of Type I-III RM systems. Genes: Gene interlocation scheme. *hsdM, M, mod* - DNA methyltransferase genes; *hsdR, R, res* – restriction enzyme genes; *hsdS* - gene of Type I sequence recognition protein. **Methylation:** Subunit composition of DNA methylation complex. **Restriction**: Subunit composition of DNA restriction complex. **Motifs**: Examples of characteristic target motifs. Figure adapted from (Beaulaurier, Schadt and Fang, 2019).

Gene Regulation by DNA Methylation

The precise target motifs and biological functions of most MTases remain largely unknown (Beaulaurier, Schadt and Fang, 2019; Won and Yim, 2024). In addition to their role in genome defense, several MTases have been shown to induce significant changes in gene expression, contributing to processes such as biofilm formation and pathogenicity (**Figure 2**) (Kwiatek *et al.*, 2015; Kumar *et al.*, 2018). Traditionally, MTases within RM systems were believed to function solely in genome protection, while epigenetic regulation was attributed exclusively to orphan MTases. However, this distinction has blurred, as both RM system and orphan MTases now have been implicated in transcriptional regulation (Vasu and Nagaraja, 2013; Sánchez-Romero and Casadesús, 2020).

One well-characterized mechanism of epigenetic regulation involves DNA methylation in promoter or regulatory regions, where it can directly influence transcription by modulating the binding of regulatory proteins. In many cases, DNA-binding proteins and MTases compete for access to the same sequences: methylation can block repressor binding, promoting gene expression, while protein binding can prevent methylation and maintain a repressed state (Sánchez-Romero and Casadesús, 2020). This regulatory interplay is exemplified by the orphan Dam MTase in *Escherichia coli* and *Salmonella enterica* (**Figure 4**). In *E. coli*, three

GATC motifs located upstream of the *agn43* gene, which encodes the outer membrane protein Antigen 43 involved in biofilm formation, are subject to dynamic regulation by Dam and the repressor protein OxyR. When the GATC sites are methylated by Dam, OxyR cannot bind, allowing gene expression. Conversely, when OxyR binds, it blocks Dam methylation and represses transcription (Adhikari and Curtis, 2016).

A similar methylation-dependent switch regulates the *gtr* operon in *S. enterica*, which encodes enzymes involved in O-antigen modification. Here, four GATC sites upstream of the operon influence transcription in a position-dependent manner: methylation of the sites closest to the transcription start site promotes gene expression, while methylation of the more distal sites represses it (Broadbent, Davies and Van Der Woude, 2010).

These examples highlight phase variation, a form of epigenetic regulation that enables the generation of phenotypically distinct yet genetically identical subpopulations. This variability enhances bacterial adaptability to fluctuating environmental conditions. Beyond phase variation, DNA methylation also contributes to other cellular processes such as cell cycle control and DNA repair (Zweiger, Marczynski and Shapiro, 1994; Sánchez-Romero and Casadesús, 2020). As such, DNA methylation functions as a versatile and dynamic regulator of prokaryotic physiology.

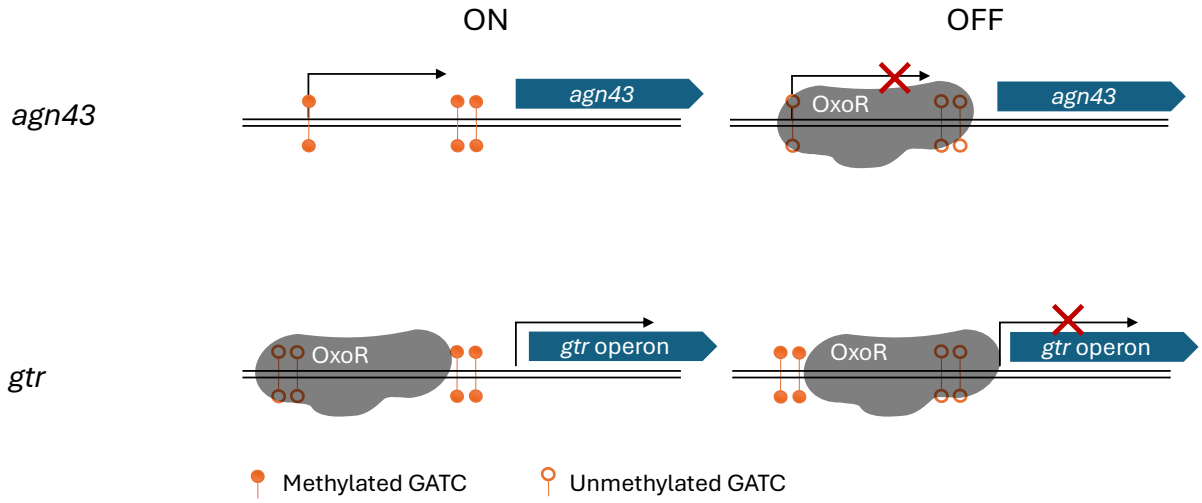


Figure 4. Models of *agn43* and *gtr* Dam-dependent Phase Variation. For the *agn43* gene, three GATC motifs are located within the promotor region. In the OFF phase, the transcriptional repressor OxyR binds to the unmethylated GATC sites, preventing RNA polymerase access and repressing transcription. In the ON phase, Dam methylates the GATC sites, which blocks OxyR binding and permits transcription. In case of the *gtr* operon, four GATC motifs are situated immediately upstream of the transcription start site. In the OFF phase, the two distal GATC sites are methylated, while OxyR binds to the two unmethylated GATC sites closest to the transcription start site, blocking RNA polymerase access and inhibiting transcription. In the ON phase, the methylation pattern is reversed: the proximal sites are methylated allowing the RNA polymerase to access the transcription start site. Cartoon not to scale. Figure adapted from (Broadbent, Davies and Van Der Woude, 2010).

Deciphering the Epigenetic Code

Historically, methodological development for DNA methylation detection has been devoted towards characterizing 5mC modifications, as this modification is the most widespread methylation type in eukaryotes. Bisulphite-based short-read sequencing has been the gold standard method for detection of 5mC modifications for many years (Beaulaurier, Schadt and Fang,

2019). In bisulfite sequencing, genomic DNA is treated with bisulfite, which deaminates unmethylated cytosine residues to uracil, while 5mC residues remain unchanged. During downstream sequencing analysis, methylated cytosines are detected as cytosines, whereas unmethylated cytosines, converted to uracil, are detected as thymine (Frommer *et al.*, 1992). Bisulfite sequencing provides high sensitivity and accuracy for 5mC detection. Yet, this method's application in prokaryotes is limited because it fails to detect the more prevalent 6mA modification and is less effective at resolving 4mC modifications (Beaulaurier, Schadt and Fang, 2019).

In addition to bisulfite sequencing, other methods for mapping specific modified bases exist, utilizing chemical or enzymatic treatments before second-generation sequencing (Hofer, Liu and Balasubramanian, 2019). As previously mentioned, prokaryotic MTases typically target defined sequence motifs. To assess the methylation status of these motifs across the genome, genomic DNA can be digested using one or more methylation-sensitive restriction enzymes with known recognition sites. The resulting pattern of cut and uncut restriction sites reflects the underlying methylation landscape (Zweiger, Marczynski and Shapiro, 1994). This approach is robust, reliable, and accurate, but it is limited to studying methylation motifs that match the specificities of the available restriction enzymes. Therefore, while it is effective for assessing methylation within known sequence motifs, it is generally not suitable for discovering new motifs (Beaulaurier, Schadt and Fang, 2019).

Today, all approaches based on second-generation sequencing require specialized treatment prior to sequencing, which is often laborious and may lead to general nucleotide deterioration (Nielsen *et al.*, 2023). Advancements in long-read, third-generation sequencing technologies, specifically, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have enabled direct detection of DNA methylations and other modifications without the need for pretreatment (Nielsen *et al.*, 2023). Both methods rely on disturbances in the inherent sequencing signals caused by modified bases. PacBio detects changes in polymerase kinetics when incorporating modified bases compared to canonical bases (Clark *et al.*, 2012), while Nanopore detects changes in the current signal as modified bases pass through the pore compared to their canonical counterparts (Nielsen *et al.*, 2023). The current state of the PacBio technology (August 2024) requires a minimum coverage of 25x for 4mC and 6mA modifications, and 250x coverage for 5mC (Nielsen *et al.*, 2023). In metagenomes, with low-abundant species and bacterial populations exhibiting heterogeneous methylation, a 250x coverage for 5mC is simply unfeasible. The last 5-10 years, multiple research efforts have developed software tools for detection of modified bases in nanopore sequencing data (Stoiber *et al.*, 2016; Liu *et al.*, 2019; Ni *et al.*, 2019; Tourancheau *et al.*, 2021; Bonet *et al.*, 2022). However, many of these tools fail from being confined to limited sequence contexts, limited training data, or requiring whole genome amplified DNA, and since the release of the software tool Dorado [1] these tools have become legacy tools. Oxford Nanopore technology released their new basecaller, namely, Dorado, which beyond the traditional A, T, C, and G basecalling can detect multiple modified base types in all sequence contexts including 4mC, 5mC, and 6mA.

[1] https://github.com/nanoporetech/dorado

Aim and Objectives

It is evident that DNA methylation plays a crucial role in prokaryotes, modulating various biological processes, including host defense mechanisms, cell cycle regulation, gene expression, and virulence (Sánchez-Romero and Casadesús, 2020). Despite this fact, research on microbial methylation systems has so far been focused on a limited selection of culturable bacteria. This small sample size confines our knowledge of microbial methylation systems particularly in terms of diversity, distribution and functionality (Blow *et al.*, 2016; Hiraoka *et al.*, 2022). To fully understand the biological significance of this modification, it is essential to expand current research efforts and map the prokaryotic methylome across a broader portion of the tree of life. Recently, third-generation sequencing technologies have paved the way for this expansion by enabling direct, high-resolution detection of DNA methylations, making methylation data readily accessible for, in principle, any genome in the tree of life. Nevertheless, comprehensive mapping of the prokaryotic methylome goes beyond just detecting methylated nucleotides. It requires identification of the related methylation motifs and association of the MTase enzymes responsible for these modifications (Beaulaurier, Schadt and Fang, 2019).

While fast and scalable methylation motif discovery tools have been developed for metagenomes (Heidelbach *et al.*, 2024), methods used to assign methylation motifs to cognate MTase genes in metagenomes remain underdeveloped. Current methods involves either querying a database of MTases with known target motif followed by manual expert assignment (Blow *et al.*, 2016; Hiraoka *et al.*, 2022; Seong *et al.*, 2022) or employing experimental means (Jensen *et al.*, 2019; Hiraoka *et al.*, 2022; Zhang *et al.*, 2023). Both approaches are labor-intensive and difficult to scale for high-throughput applications. In the era of meta(epi)genomics, a state-of-the-art, high-throughput tool, which can make these modification-enzyme pairs readily available in novel, unculturable species and on the metagenomic scale, is needed.

The overall aim of this project is to develop a scalable, bioinformatic approach for linking DNA methylation motifs to their cognate MTase genes in prokaryotic genomes, including those found in metagenomic datasets. To achieve this, the project will pursue the following specific objectives:

1. Examine existing methods for MTase gene annotation and motif-MTase assignment in prokaryotic genomes and metagenomes, and evaluate their strengths, limitations, and suitability for metagenomic applications.
2. Design a computational pipeline capable of:
    - Identifying putative MTase genes and associated defense system genes in complete genomes or metagenome-assembled genomes (MAGs).
    - Shortlisting putative MTase genes responsible for a specific, observed DNA methylation motif.
3. Integrate the pipeline as a submodule into the Nanomotif framework, ensuring compatibility with high-throughput sequencing datasets as well as user-friendliness.
4. Evaluate the performance of the pipeline by applying it to both complete genomes and complex metagenomic datasets.

MTase-Linker

To address the aim outlined in the previous chapter, we developed MTase-Linker, a submodule of Nanomotif (see (Heidelbach *et al.*, 2024) or the **Manuscript** in the final chapter of this thesis). MTase-Linker is designed to provide scalable and high-throughput insights into bacterial methylation systems - not only in culturable prokaryotic genomes, but also in complex metagenomes.

MTase-Linker is a user-friendly, command-line pipeline that:

- Identifies putative MTase genes and associated defense system genes in complete genomes and MAGs;
- Shortlists putative MTase genes responsible for specific, observed DNA methylation motifs.

An overview of the MTase-Linker workflow is illustrated in **Figure 5** and a complete description of the pipeline can be found in the methods section of the **Manuscript**.

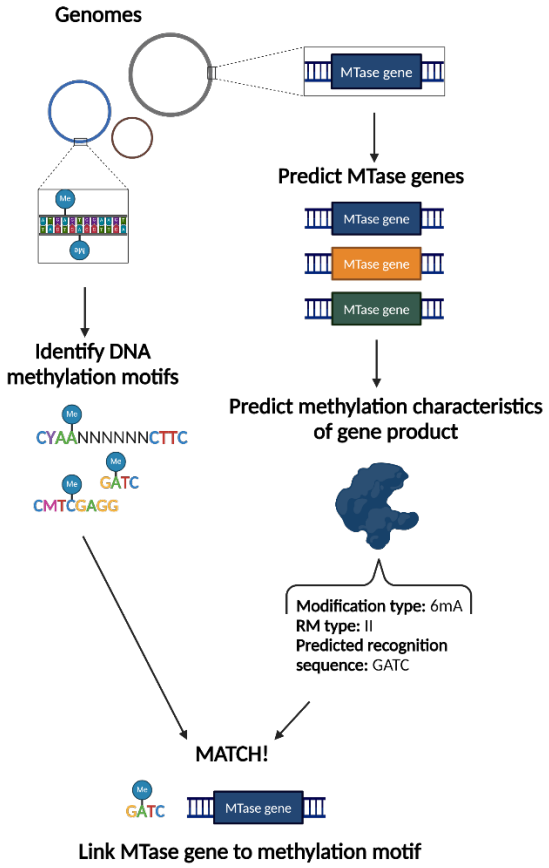


Figure 5. Workflow of MTase-Linker. Given a genome or MAG as input, the pipeline initially annotates MTase and related defense system genes. Subsequently, the methylation characteristics are predicted for each MTase gene product. This includes predicting the expected modification type (4mC, 5mC, 6mA) and RM system type (Type I, Type II, Type III), and inferring the predicted target recognition motif based on sequence similarity to a database of MTases with known target motif. For each genome, this information is compared with the methylation motifs identified by the motif discovery module of Nanomotif. Methylation motifs are associated with MTase genes by leveraging the predicted gene products' methylation characteristics to narrow down responsible MTases for each detected motif. See methods in **Manuscript.**

Annotating MTase Genes

To establish active RM MTases or orphan MTases within a prokaryotic organism, it is essential first to identify the genes encoding these enzymes within their genome. Annotating MTase genes within a prokaryotic genome can, however, pose significant challenges for at least two reasons:

(1) The sequence space of MTase domains is vast and diverse (Samokhina and Alexeevski, 2023; Tisza *et al.*, 2023). Consequently, local alignment annotation approaches, such as BLASTP, are insufficient for annotating MTases in microbial environments like soil, where high novelty is expected.
(2) Many MTase incorporate additional domains such as DNA helicase domains, and some MTase domains have close homology to other genes, particularly RNA methyltransferases. Standard annotation tools like Bakta may mislabel these genes as “Helicase” or provide ambiguous labels like “methyltransferase” (Samokhina and Alexeevski, 2023; Tisza *et al.*, 2023).

To overcome these challenges, manually curated profile hidden Markov models (HMMs) targeting MTases and RM-system genes have been generated by multiple research groups. HMMs are probabilistic models used to annotate protein or gene sequence families within a genome. They encompass the variability and conserved patterns within a family, making them particularly effective at capturing a broader range of sequences while simultaneously distinguishing closely related gene or protein families. Multiple state-of-the-art, computational tools apply HMMs to annotate MTases. For example, DNA Methylase Finder (Tisza *et al.*, 2023) and MicrobeMod (Crits-Christoph *et al.*, 2023) specifically annotates all potential DNA methyltransferases and neighboring RM-systems genes in a genome. rmsFinder (Shaw, Rocha and MacLean, 2023) only targets Type II RM-systems while DefenseFinder (Tesson *et al.*, 2022) and PADLOC (Payne *et al.*, 2021) systematically annotates all known antiphage systems including RM-systems. Some of these tools use the same profile HMMs originally generated by (Oliveira, Touchon and Rocha, 2014), and later updated by the tools’ developers. In the MTase-Linker pipeline, DefenseFinder is used for this purpose. The authors of this tool have manually curated all their models and report a sensitivity above 91% for all MTase profiles when searching against REbase, a comprehensive and extensively curated database of RM system genes (Roberts *et al.*, 2023). The false positive rate of these profiles is difficult to determine without experimental validation. However, no systematic false positive pattern was noticeable, when 11 prokaryotic strains were manually compared to literature and previous annotations in REbase (see supplementary note 2 in **Manuscript**), and when MTase sequences identified by MTase-Linker were annotated using Bakta (Schwengers *et al.*, 2021); the state-of-the-art tool for microbial genome annotation (**Figure *6***). Furthermore, annotating MTase genes as part of complete RM systems strengthens the predictions, as co-localization of related genes supports each other's identification.

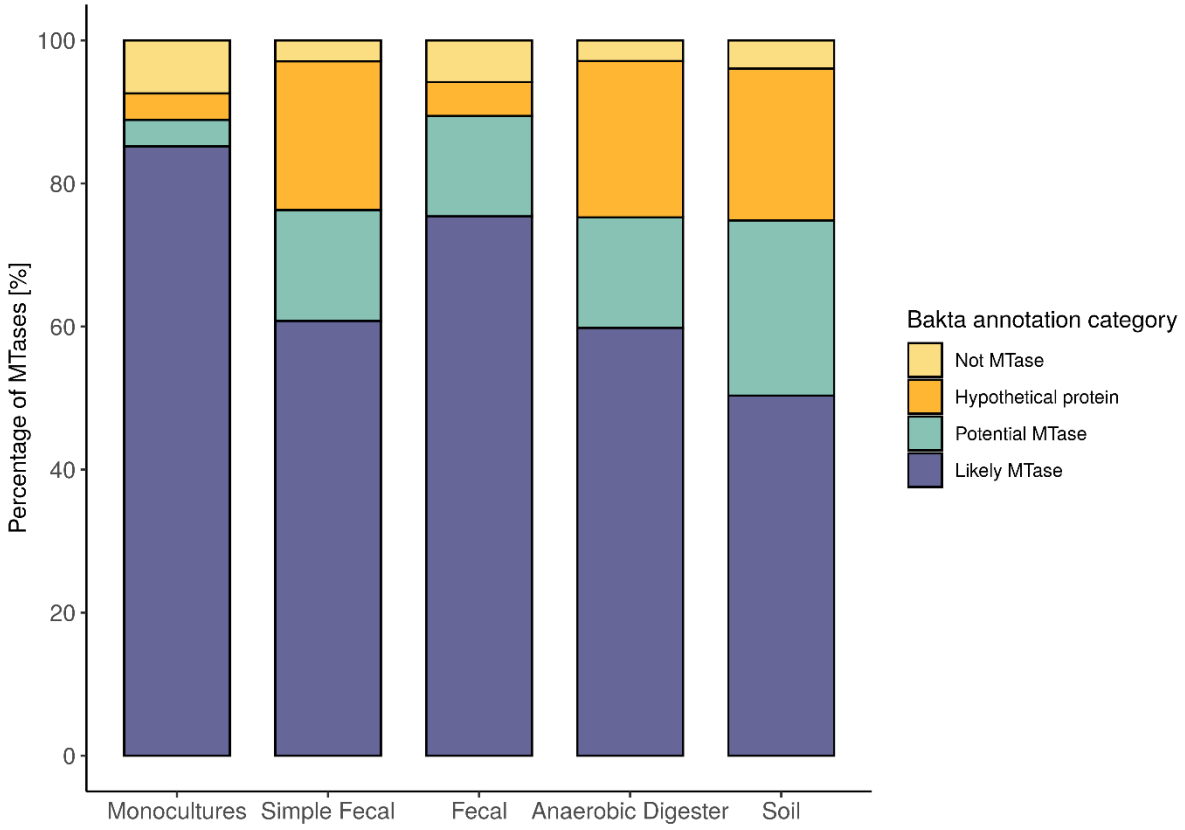


Figure 6. Bakta Annotation of MTase Sequences. MTase amino acid sequences identified by MTase-Linker in ten monocultures and four metagenomes were annotated using Bakta. These annotations were manually grouped into four categories: **Likely MTase**, sequences with high-confidence annotations as DNA methyltransferases; **Possible MTase**, sequences exhibiting some characteristic annotation of MTases but lacking definitive evidence; **Not MTase**, sequences likely misannotated as MTases, showing evidence of other functions; and **hypothetical proteins**.

Linking MTase Genes with Motifs

Linking specific DNA methylation motifs to their cognate MTases through genomic analysis alone is not an easy feat. The regulatory roles of many MTases within the cell often render them functionally silent under most conditions. As a result, prokaryotic genomes often encode more MTases genes than the number of distinct motifs detected on the genome, complicating efforts to accurately assign target motifs (Blow *et al.*, 2016; Tisza *et al.*, 2023). A commonly employed strategy involves sequence homology-based prediction, which relies on querying a database of MTases with known recognitions motifs to infer target motifs of unknown MTases. Nonetheless, multiple studies have demonstrated that this approach frequently lacks accuracy and fails to predict the recognition motifs for a substantial proportion of MTases (Hiraoka *et al.*, 2019, 2022; Jensen *et al.*, 2019; Seong *et al.*, 2022). Additionally, the presence of most RM systems and orphan MTases in prokaryotic genomes is connected to horizontal gene transfer. Most MTases are located within the shell or cloud compartments of the pangenome, and many are associated with mobile genetic elements, such as plasmids (Oliveira, Touchon and Rocha, 2014). This leads to significant variation in the set of MTases even among closely related species or strains, which ultimately limits the effectiveness of taxonomic approaches for establishing linkages between MTases and their recognition motifs. To overcome these challenges, MTase-Linker uses three different approaches to shortlist MTase genes responsible for observed DNA methylation motifs.

Homology-based Recognition Motif Prediction

In MTase-Linker, identified MTase sequences are first queried against REbase using BLASTP to estimate recognition motifs based on sequence similarity to MTase genes with known recognition motif. Predicted target motifs are used to directly establish a link to observed methylation motifs in the genome. The default threshold for motif inference is set at 80% identity and 80% query coverage, aligning with thresholds used in a previous study for motif estimation via sequence similarity (Tisza *et al.*, 2023). An analysis of the gold standard proteins in REbase was conducted to evaluate the robustness of these thresholds (**Figure 7**). This analysis revealed that only nearly identical homologs of Type I MTases and Type IS subunits recognized the same motif, while homologs of Type II and III MTases exhibited identical target motifs at identity levels above 50% and 70%, respectively, provided the query coverage was above 80%. These findings are consistent with previous studies (Oliveira, Touchon and Rocha, 2016).

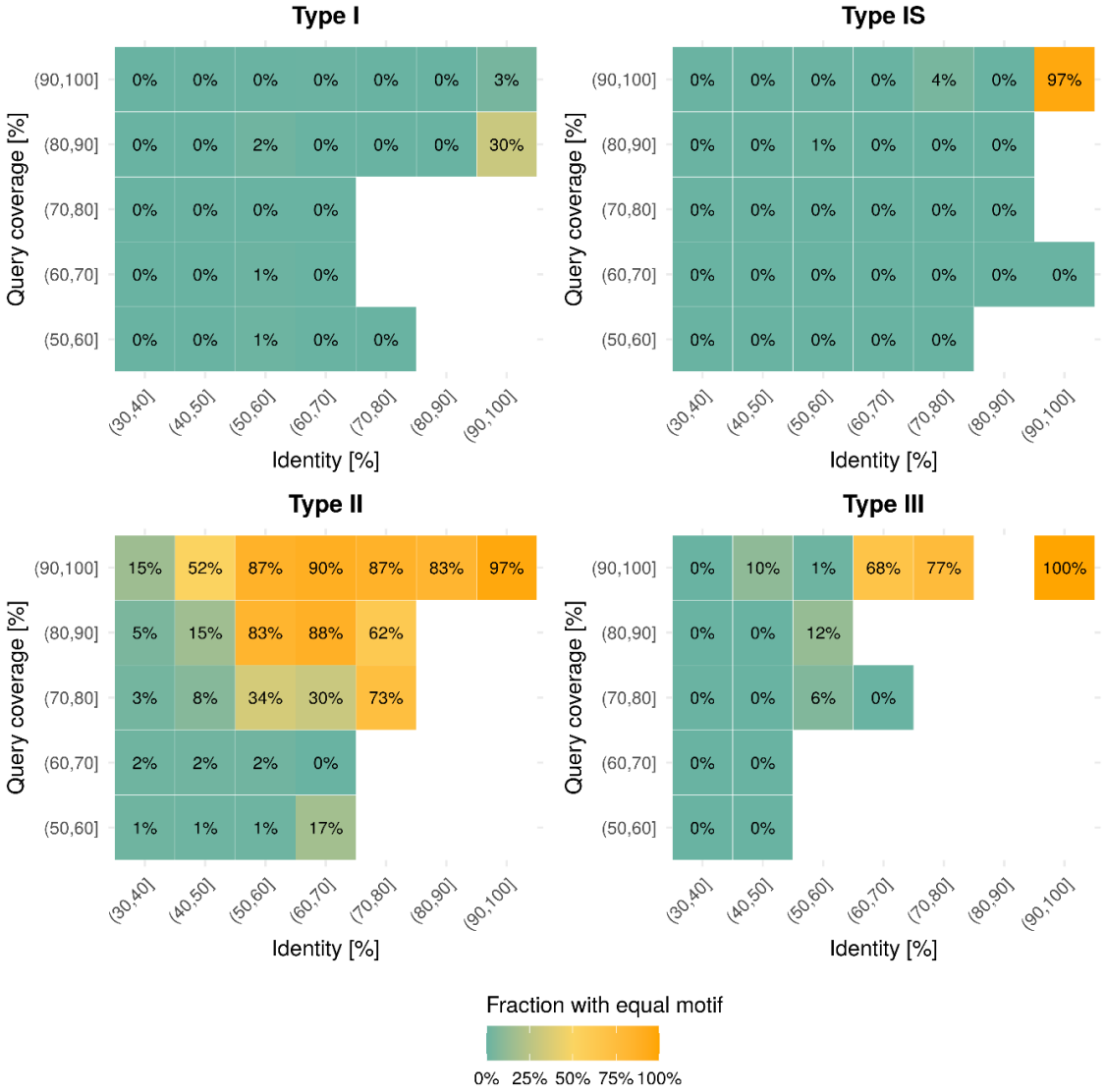


Figure 7. Relationship between target recognition motif and protein sequence similarity in MTases. Gold standard MTase proteins from REbase were subjected to an all-vs-all BLASTP analysis. The heatmap displays the frequency at which MTases of a given type recognize the same motif across different identity and query coverage intervals. The analysis includes 729 Type I, 745 Type II, and 99 Type III MTases and 712 Type IS subunits.

RM System Type Predicts Motif Type

In MTase-Linker, each MTase gene product is classified according to its RM system type using profile HMMs specific to each type. This classification helps predict the recognition motif type of the MTases, which in turn is used to narrow down the number of observed methylation motifs that the specific MTases could be responsible for (**Figure 3**).

Modification Type Prediction

Similar to the classification of RM system types, profile HMMs specific to 5mC and 6mA/4mC modifications are used to predict the modification type of MTase gene products. This information further narrows down the list of observed methylation motifs that the specific MTases could be responsible for. The profile HMMs used to predict the modification types has been retrieved from the Interpro database (See methods section of the **Manuscript**) and a preliminary analysis of the gold standard proteins in REbase was conducted to evaluate the performance of these models in modification type prediction (**Figure 8**). The analysis shows a robust prediction with an accuracy above 97% for both 5mC and 6mA/4mC prediction models.

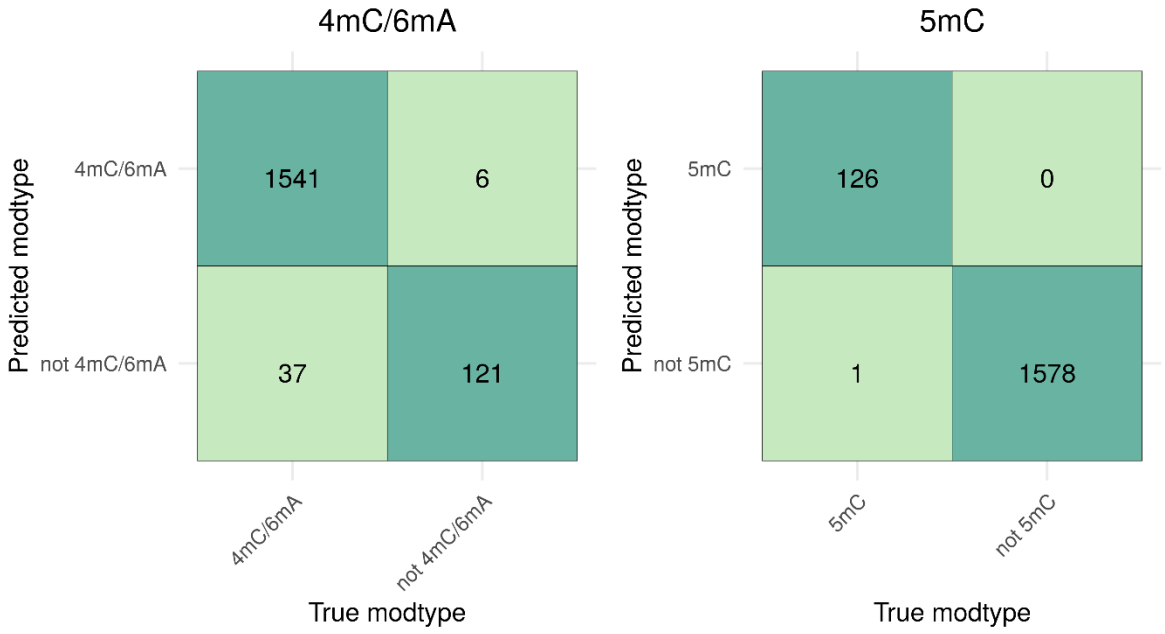


Figure 8. Modification Type Prediction Performance. Confusion matrix showcasing the performance of the profile hidden Markov models used to predict the modification types (modtype) on the gold standard proteins of Rebase. The set of hidden Markov models used is PF01555.22, PF02384.20, PF12161.12, PF05869.15, PF02086.19, PF07669.15, PF13651.10, PF00145.21 from the Interpro database.

Conclusion

With Nanomotif comprehensive mapping of the bacterial methylome is now broadly accessible using standard Nanopore sequencing - even in complex metagenomic datasets and without the need for additional methods. The MTase-linker submodule modernizes the motif-MTase pairing process by replacing existing manual approaches with a scalable, user-friendly bioinformatics tool that links methylation motifs to their cognate MTase and defense system genes. This streamlines analysis and lowers the barrier for researchers to explore DNA methylation in uncultured or poorly characterized microbes. These motif-MTase pairs can not only help circumvent genetic transformation barriers but can also open avenues to explore the functional roles of methylations and their implications for microbial physiology across a broad spectrum of prokaryotic taxa.

As the REBASE database - on which the MTase-linker relies - continues to expand, the tool's capacity to generate high-confidence motif-MTase pairs will only strengthen, facilitating discovery in increasingly complex metagenomic samples. One possible direction for future development is the integration of advanced protein structure prediction and alignment tools, which might improve MTase gene annotation and motif assignment (Jumper *et al.*, 2021; Heinzinger *et al.*, 2024; Van Kempen *et al.*, 2024).

The open-source codebase, freely available on GitHub, empowers the community to adapt and extend Nanomotif for diverse applications. For example, it can serve as a foundation for developing methods to predict genetic flux between bacteria (Oliveira, Touchon and Rocha, 2016), monitor epigenetic changes in microbial communities under stress (D'Aquila *et al.*, 2023), or design synthetic biology systems with epigenetic control mechanisms (Komera *et al.*, 2024).

References

Adhikari, S. and Curtis, P.D. (2016) 'DNA methyltransferases and epigenetic regulation in bacteria', *FEMS Microbiology Reviews*, 40(5), pp. 575–591. Available at: https://doi.org/10.1093/femsre/fuw023.

Albertsen, M. (2023) 'Long-read metagenomics paves the way toward a complete microbial tree of life', *Nature Methods*, 20(1), pp. 30–31. Available at: https://doi.org/10.1038/s41592-022-01726-6.

Beaulaurier, J., Schadt, E.E. and Fang, G. (2019) 'Deciphering bacterial epigenomes using modern sequencing technologies', *Nature Reviews Genetics*, 20(3), pp. 157–172. Available at: https://doi.org/10.1038/s41576-018-0081-3.

Blow, M.J. *et al.* (2016) 'The Epigenomic Landscape of Prokaryotes', *PLOS Genetics*, 12(2), pp. 1–28. Available at: https://doi.org/10.1371/journal.pgen.1005854.

Bonet, J. *et al.* (2022) 'DeepMP: a deep learning tool to detect DNA base modifications on Nanopore sequencing data', *Bioinformatics*, 38(5), pp. 1235–1243. Available at: https://doi.org/10.1093/bioinformatics/btab745.

Broadbent, S.E., Davies, M.R. and Van Der Woude, M.W. (2010) 'Phase variation controls expression of *Salmonella* lipopolysaccharide modification genes by a DNA methylation-dependent mechanism', *Molecular Microbiology*, 77(2), pp. 337–353. Available at: https://doi.org/10.1111/j.1365-2958.2010.07203.x.

Brock, T.D. (1985) 'Life at High Temperatures', *Science*, 230(4722), pp. 132–138. Available at: https://doi.org/10.1126/science.230.4722.132.

Buchholz, K. and Collins, J. (2013) 'The roots—a short history of industrial microbiology and biotechnology', *Applied Microbiology and Biotechnology*, 97(9), pp. 3747–3762. Available at: https://doi.org/10.1007/s00253-013-4768-2.

Cavicchioli, R. *et al.* (2019) 'Scientists' warning to humanity: microorganisms and climate change', *Nature Reviews Microbiology*, 17(9), pp. 569–586. Available at: https://doi.org/10.1038/s41579-019-0222-5.

Clark, T.A. *et al.* (2012) 'Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing', *Nucleic Acids Research*, 40(4), pp. 1–12. Available at: https://doi.org/10.1093/nar/gkr1146.

Crits-Christoph, A. *et al.* (2023) 'MicrobeMod: A computational toolkit for identifying prokaryotic methylation and restriction-modification with nanopore sequencing', *bioRxiv* [Preprint]. Available at: https://doi.org/10.1101/2023.11.13.566931.

Crowther, T.W. *et al.* (2024) 'Scientists' call to action: Microbes, planetary health, and the Sustainable Development Goals', *Cell*, 187(19), pp. 5195–5216. Available at: https://doi.org/10.1016/j.cell.2024.07.051.

D'Aquila, P. *et al.* (2023) 'Epigenetic-Based Regulation of Transcriptome in Escherichia coli Adaptive Antibiotic Resistance', *Microbiology Spectrum*, 11(3), pp. 1–16. Available at: https://doi.org/10.1128/spectrum.04583-22.

Ershova, A.S. *et al.* (2015) 'Role of restriction-modification systems in prokaryotic evolution and ecology', *Biochemistry (Moscow)*, 80(10), pp. 1373–1386. Available at: https://doi.org/10.1134/S0006297915100193.

Frommer, M. *et al.* (1992) 'A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.', *Proceedings of the National Academy of Sciences*, 89(5), pp. 1827–1831. Available at: https://doi.org/10.1073/pnas.89.5.1827.

Gao, Q. *et al.* (2023) 'Bacterial DNA methyltransferase: A key to the epigenetic world with lessons learned from proteobacteria', *Frontiers in Microbiology*, 14:1129437, pp. 1–19. Available at: https://doi.org/10.3389/fmicb.2023.1129437.

Gest, H. (2004) 'The discovery of microorganisms by Robert Hooke and Antoni van Leeuwenhoek, Fellows of The Royal Society', *Fellows of The Royal Society. Notes Rec. R. Soc. Lond.*, 58(2), pp. 187–201. Available at: https://doi.org/10.1098/rsnr.2004.0055.

Gilbert, J.A. and Neufeld, J.D. (2014) 'Life in a World without Microbes', *PLoS Biology*, 12(12), pp. 1–3. Available at: https://doi.org/10.1371/journal.pbio.1002020.

Heidelbach, S. *et al.* (2024) 'Nanomotif: Identification and Exploitation of DNA Methylation Motifs in Metagenomes using Oxford Nanopore Sequencing', *bioRxiv* [Preprint]. Available at: https://doi.org/10.1101/2024.04.29.591623.

Heinzinger, M. *et al.* (2024) 'Bilingual language model for protein sequence and structure', *NAR Genomics and Bioinformatics*, 6(4), pp. 1–15. Available at: https://doi.org/10.1093/nargab/lqae150.

Hiraoka, S. *et al.* (2019) 'Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an environmental prokaryotic community', *Nature Communications*, 10, 159, pp. 1–10. Available at: https://doi.org/10.1038/s41467-018-08103-y.

Hiraoka, S. *et al.* (2022) 'Diverse DNA modification in marine prokaryotic and viral communities', *Nucleic Acids Research*, 50(3), pp. 1531–1550. Available at: https://doi.org/10.1093/nar/gkab1292.

Hofer, A., Liu, Z.J. and Balasubramanian, S. (2019) 'Detection, Structure and Function of Modified DNA Bases', *Journal of the American Chemical Society*, 141(16), pp. 6420–6429. Available at: https://doi.org/10.1021/jacs.9b01915.

Hug, L.A. *et al.* (2016) 'A new view of the tree of life', *Nature Microbiology*, 1(5), pp. 1–6. Available at: https://doi.org/10.1038/nmicrobiol.2016.48.

Jensen, T.Ø. *et al.* (2019) 'Genome-wide systematic identification of methyltransferase recognition and modification patterns', *Nature Communications*, 10, 3311, pp. 1–9. Available at: https://doi.org/10.1038/s41467-019-11179-9.

Jumper, J. *et al.* (2021) 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596(7873), pp. 583–589. Available at: https://doi.org/10.1038/s41586-021-03819-2.

Komera, I. *et al.* (2024) 'Microbial Synthetic Epigenetic Tools Design and Applications', *ACS Synthetic Biology*, 13(6), pp. 1621–1632. Available at: https://doi.org/10.1021/acssynbio.4c00125.

Kumar, S. *et al.* (2018) 'N4-cytosine DNA methylation regulates transcription and pathogenesis in Helicobacter pylori', *Nucleic Acids Research*, 46(7), pp. 3429–3445. Available at: https://doi.org/10.1093/nar/gky126.

Kwiatek, A. *et al.* (2015) 'Type III Methyltransferase M.NgoAX from Neisseria gonorrhoeae FA1090 Regulates Biofilm Formation and Interactions with Human Cells', *Frontiers in Microbiology*, 6, 1426, pp. 1–15. Available at: https://doi.org/10.3389/fmicb.2015.01426.

Larsen, B.B. *et al.* (2017) 'Inordinate Fondness Multiplied and Redistributed: the Number of Species on Earth and the New Pie of Life', *The Quarterly Review of Biology*, 92(3), pp. 229–265. Available at: https://doi.org/10.1086/693564.

Liu, Q. *et al.* (2019) 'Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data', *Nature Communications*, 10, 2449, pp. 1–11. Available at: https://doi.org/10.1038/s41467-019-10168-2.

Louca, S. *et al.* (2019) 'A census-based estimate of Earth's bacterial and archaeal diversity', *PLOS Biology*. Edited by J.K. Jansson, 17(2), pp. 1–30. Available at: https://doi.org/10.1371/journal.pbio.3000106.

Ni, P. *et al.* (2019) 'DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning', *Bioinformatics*, 35(22), pp. 4586–4595. Available at: https://doi.org/10.1093/bioinformatics/btz276.

Nielsen, T.K. *et al.* (2023) 'Detection of nucleotide modifications in bacteria and bacteriophages: Strengths and limitations of current technologies and software', *Molecular Ecology*, 32(6), pp. 1236–1247. Available at: https://doi.org/10.1111/mec.16679.

Oliveira, P.H. and Fang, G. (2021) 'Conserved DNA Methyltransferases: A Window into Fundamental Mechanisms of Epigenetic Regulation in Bacteria', *Trends in Microbiology*, 29(1), pp. 28–40. Available at: https://doi.org/10.1016/j.tim.2020.04.007.

Oliveira, P.H., Touchon, M. and Rocha, E.P.C. (2014) 'The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts', *Nucleic Acids Research*, 42(16), pp. 10618–10631. Available at: https://doi.org/10.1093/nar/gku734.

Oliveira, P.H., Touchon, M. and Rocha, E.P.C. (2016) 'Regulation of genetic flux between bacteria by restriction–modification systems', *Proceedings of the National Academy of Sciences*, 113(20), pp. 5658–5663. Available at: https://doi.org/10.1073/pnas.1603257113.

Parks, D.H. *et al.* (2022) 'GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy', *Nucleic Acids Research*, 50(D1), pp. D785–D794. Available at: https://doi.org/10.1093/nar/gkab776.

Payne, L.J. *et al.* (2021) 'Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types', *Nucleic Acids Research*, 49(19), pp. 10868–10878. Available at: https://doi.org/10.1093/nar/gkab883.

Pérez-Cobas, A.E., Gomez-Valero, L. and Buchrieser, C. (2020) 'Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses', *Microbial Genomics*, 6(8), pp. 1–22. Available at: https://doi.org/10.1099/mgen.0.000409.

Rinke, C. *et al.* (2013) 'Insights into the phylogeny and coding potential of microbial dark matter', *Nature*, 499(7459), pp. 431–437. Available at: https://doi.org/10.1038/nature12352.

Roberts, R.J. (2003) 'A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes', *Nucleic Acids Research*, 31(7), pp. 1805–1812. Available at: https://doi.org/10.1093/nar/gkg274.

Roberts, R.J. *et al.* (2023) 'REBASE: a database for DNA restriction and modification: enzymes, genes and genomes', *Nucleic Acids Research*, 51(D1), pp. D629–D630. Available at: https://doi.org/10.1093/nar/gkac975.

Samokhina, M. and Alexeevski, A. (2023) 'A Novel Classification System for Prokaryotic DNA Methyltransferases Based on 3D Catalytic Domain Topology', *bioRxiv* [Preprint]. Available at: https://doi.org/10.1101/2023.12.13.571470.

Sánchez-Romero, M.A. and Casadesús, J. (2020) 'The bacterial epigenome', *Nature Reviews Microbiology*, 18(1), pp. 7–20. Available at: https://doi.org/10.1038/s41579-019-0286-2.

Satam, H. *et al.* (2023) 'Next-Generation Sequencing Technology: Current Trends and Advancements', *Biology*, 12, 997, pp. 1–25. Available at: https://doi.org/10.3390/biology12070997.

Schwengers, O. *et al.* (2021) 'Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification: Find out more about Bakta, the motivation, challenges and applications, here.', *Microbial Genomics*, 7(11), pp. 1–13. Available at: https://doi.org/10.1099/mgen.0.000685.

Seong, H.J. *et al.* (2022) 'Marine DNA methylation patterns are associated with microbial community composition and inform virus-host dynamics', *Microbiome*, 10(1), pp. 1–19. Available at: https://doi.org/10.1186/s40168-022-01340-w.

Seong, H.J., Han, S.-W. and Sul, W.J. (2021) 'Prokaryotic DNA methylation and its functional roles', *Journal of Microbiology*, 59(3), pp. 242–248. Available at: https://doi.org/10.1007/s12275-021-0674-y.

Shaw, L.P., Rocha, E.P.C. and MacLean, R.C. (2023) 'Restriction-modification systems have shaped the evolution and distribution of plasmids across bacteria', *Nucleic Acids Research*, 51(13), pp. 6806–6818. Available at: https://doi.org/10.1093/nar/gkad452.

Singleton, C. *et al.* (2024) 'Microflora Danica: the atlas of Danish environmental microbiomes', *bioRxiv* [Preprint]. Available at: https://doi.org/10.1101/2024.06.27.600767.

Stoiber, M. *et al.* (2016) '*De novo* Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing', *bioRxiv* [Preprint]. Available at: https://doi.org/10.1101/094672.

Tesson, F. *et al.* (2022) 'Systematic and quantitative view of the antiviral arsenal of prokaryotes', *Nature Communications*, 13, 2561, pp. 1–10. Available at: https://doi.org/10.1038/s41467-022-30269-9.

Timmis, K. *et al.* (2017) 'The contribution of microbial biotechnology to sustainable development goals', *Microbial Biotechnology*, 10(5), pp. 984–987. Available at: https://doi.org/10.1111/1751-7915.12818.

Tisza, M.J. *et al.* (2023) 'Roving methyltransferases generate a mosaic epigenetic landscape and influence evolution in Bacteroides fragilis group', *Nature Communications*, 14, 4082, pp. 1–14. Available at: https://doi.org/10.1038/s41467-023-39892-6.

Tourancheau, A. *et al.* (2021) 'Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing', *Nature Methods*, 18(5), pp. 491–498. Available at: https://doi.org/10.1038/s41592-021-01109-3.

Van Kempen, M. *et al.* (2024) 'Fast and accurate protein structure search with Foldseek', *Nature Biotechnology*, 42(2), pp. 243–246. Available at: https://doi.org/10.1038/s41587-023-01773-0.

Vasu, K. and Nagaraja, V. (2013) 'Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense', *Microbiology and Molecular Biology Reviews*, 77(1), pp. 53–72. Available at: https://doi.org/10.1128/MMBR.00044-12.

Watson, J.D. and Crick, F.H.C. (1953) 'Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid', *Nature*, 171(4356), pp. 737–738. Available at: https://doi.org/10.1038/171737a0.

Wilbanks, E.G. *et al.* (2022) 'Metagenomic methylation patterns resolve bacterial genomes of unusual size and structural complexity', *The ISME Journal*, 16(8), pp. 1921–1931. Available at: https://doi.org/10.1038/s41396-022-01242-7.

Won, C. and Yim, S.S. (2024) 'Emerging methylation-based approaches in microbiome engineering', *Biotechnology for Biofuels and Bioproducts*, 17(1), p. 96. Available at: https://doi.org/10.1186/s13068-024-02529-x.

Zhang, Y. *et al.* (2023) 'A sequential one-pot approach for rapid and convenient characterization of putative restriction-modification systems', *mSystems*, 8(6), pp. 1–18. Available at: https://doi.org/10.1128/msystems.00817-23.

Zweiger, G., Marczynski, G. and Shapiro, L. (1994) 'A Caulobacter DNA Methyltransferase that Functions only in the Predivisional Cell', *Journal of Molecular Biology*, 235(2), pp. 472–485. Available at: https://doi.org/10.1006/jmbi.1994.1007.

Manuscript

MTase-linker is developed as a submodule of Nanomotif. Nanomotif is a Python package designed to explore and utilize methylation in prokaryotic genomes using Nanopore sequencing. In the following manuscript, we demonstrate how Nanomotif can offer valuable insights into the bacterial epigenome. These insights can be applied to associate plasmids with the genome and to evade restriction-modification (RM) systems during genetic transformation.

This article is formatted as a brief communication, with all components - including the main text, methods, figures, references, and supplementary information - annotated with line number to clearly distinguish the elements that constitute the paper and the broader project. Figures and literature will be referenced independently of the broader project.

Nanomotif: Leveraging DNA Methylation Motifs for Genome Recovery and Host Association of Plasmids in Metagenomes from Complex Microbial Communities

Søren Heidelbach[1], Sebastian Mølvang Dall[1], Jeppe Støtt Bøjer[1], Jacob Nissen[2], Lucas N.L. van der Maas[3], Mantas Sereika[1], Rasmus H. Kirkegaard[1], Sheila I. Jensen[3], Sabrina Just Kousgaard[4,5], Ole Thorlacius-Ussing[4,5], Katja Hose[6], Thomas Dyhre Nielsen[2], Mads Albertsen[1]*

[1]Center for Microbial Communities, Aalborg University, Denmark

[2]Department of Computer Science, Aalborg University, Denmark

[3]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Denmark

[4]Department of Gastrointestinal Surgery, Aalborg University Hospital, Denmark

[5]Department of Clinical Medicine, Aalborg University, Denmark

[6]Institute of Logic and Computation, TU Wien, Austria

*corresponding author.

Abstract

DNA methylation is found across all domains of life but is a rarely used feature in recovery of metagenome-assembled genomes (MAGs). Recently, Oxford Nanopore introduced all context methylation detection models. We leveraged this to develop Nanomotif, which identifies and exploits methylation motifs for enhanced MAG recovery. We demonstrate how Nanomotif enables database-independent contamination removal from high-quality MAGs and host association of plasmids directly from Nanopore sequencing data in complex metagenomes.

Main

In all domains of life, genomes are subjected to epigenetic modifications, which directly influences gene expression, replication, and repair processes[1]. In bacteria, the most common epigenetic modification is DNA methylation, which primarily acts as a host-defense mechanism against phages[1]. DNA methylation is facilitated by DNA methyltransferases (MTases), which recognizes specific DNA sequences, called motifs, and adds a methyl group to the DNA[1,2]. MTases often appear in restriction-modification systems, where a restriction enzyme recognizes a specific motif and cleaves the DNA if it lacks methylation. All DNA in the host must therefore have the correct methylation pattern for it to persist, including mobile genetic

elements[2,3]. This evolutionary arms race has given rise to a great diversity of MTase recognition sequences[4]. Historically, DNA methylations have been identified using bisulfite conversions followed by short-read sequencing[1]. In recent years, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have enabled direct detection of DNA methylations without the need for pre-treatment[5]. The most common methylations in bacteria are 5-methylcytosine (5mC), N6-methyladenine (6mA), and N4-methylcytosine (4mC). PacBio was first to demonstrate *de novo* detection of DNA methylation[5], but currently has a low sensitivity for 5mC which requires a high sequencing coverage (250x)[6,7]. During 2023-24, ONT introduced all context methylation detection models making 4mC, 5mC and 6mA methylation calls readily available with high sensitivity (https://github.com/nanoporetech/dorado). Despite this, only few efforts have been made to utilize ONT methylation calls for methylation motif discovery in bacteria[8–10], but none which scales or extends motif discovery to metagenome sequencing of microbial communities.

In metagenomics, DNA methylation motifs are directly applicable in binning by clustering contigs, assess contamination in bins, and associate mobile genetic elements to specific microbial hosts. Previous studies have utilized methylation motif information for metagenomic binning and association of plasmids[2,11]. However, these methodologies suffer from the low PacBio sensitivity for 5mC[2,11] or require whole genome amplification for detection of motifs using ONT[12]. Building on the recent methylation calling capabilities of ONT sequencing, we developed Nanomotif, a fast, scalable, and sensitive tool for identification and utilization of methylation motifs in metagenomic samples. Nanomotif offers *de novo* methylated motif identification, metagenomic bin contamination detection, association of unbinned contigs to existing bins, and linking of restriction-modification systems to methylation motifs (Fig. 1A)

Nanomotif finds methylated motifs in individual contigs by first extracting windows of 20 bases upstream and downstream of highly methylated positions (>80% methylated). Motif candidates are then built iteratively using a beta-Bernoulli model, which evaluates whether the new candidate is more methylated relative to its originating candidate motif. The motif candidate search is directed using the KL-divergence from a non-methylated background, which rapidly guides Nanomotif through the motif search space, greatly decreasing search time compared to other algorithms (Supplementary note 1, Tab. S1).

We investigated a total of 28 monocultures, including 11 REBASE gold standard strains with known methylation motifs. The 11 monocultures with 75 expected methylation motifs were further split into 6 strains (29 motifs) for training and 5 strains (46 motifs) for testing. We benchmarked Nanomotif against Modkit[10], MicrobeMod[8] and MotifMaker[13] on both training and test monocultures. Only Nanomotif and Modkit performed satisfactory, and were therefore included in further benchmarks (Fig. S1 & S2). Nanomotif achieved a high recall rate and precision across all monocultures, identifying 68 out of the 75 expected motifs and 15 other motifs (Fig. 1B). Nine of the other motifs were closely related to a non-identified expected motif (Fig. 1B-#1-4). In *M. ruber*, RGAT**4mC**Y was missed, as it is a sub-motif of G**6mA**TC (Fig. 1B-#5). In *A. variabilis*, **4mC**YCGRG and ATGC**6mA**T were missed due to only 36 and 74 occurrences, respectively (Fig. 1B-#6). Lastly, four rare motifs were identified in *M. ruber* (57-474 counts) that likely represent noise due to increased 5mC false positive rate in high GC% organisms (Fig. 1B-#7 and Fig. S3).

To simulate metagenomic conditions, we further benchmarked motif identification by segmenting the test monoculture genomes to a varying number of fragment sizes and coverages (Fig.

1.C). Nanomotif and Modkit perform similarly on palindromic motifs, which can be confidently identified in 25 kbp fragments even at 25x coverage. Palindromic motifs are generally shorter and therefore easier to detect because of their higher frequency and simplicity. For non-palindromic and bipartite motifs Nanomotif and Modkit have similar performance on 10-100 kbp, however on 1000 kbp fragments Modkit finds more false-positives leading to a drop in precision and hence F1-score (Fig. S4 & S5). Lastly, we benchmarked scalability, where Modkit used 23-40 times more total time compared to Nanomotif (Tab. S1).

A unique feature of Nanomotif is the scalability to complex metagenomic samples. We therefore used Nanomotif on five increasingly complex metagenomic samples (Fig. 2A). Across all metagenomic samples, at least one motif was identified within 87% of metagenome-assembled genomes (MAGs) above 10x coverage, and within these the median number of identified motifs was 3. This is more than previously reported in small-scale meta-epigenomic studies, which only identified methylation motifs in approximately 50% of MAGs using PacBio[14,15].

Building on the motif discovery algorithm, we developed three modules for Nanomotif, which uses the motif methylation pattern; MAG contamination detection, inclusion of unbinned contigs, and linking of motifs to cognate methyltransferases.

Current MAG contamination evaluation tools rely on lineage-specific markers derived from genome databases[16–18], however, it is a difficult task as the databases are far from complete, and exceptions exist even within closely related organisms. To enable *de novo* contamination detection in MAGs, we leveraged Nanomotif to identify methylation motifs and then used ensemble clustering on the methylation pattern of bins (see methods). The 28 monocultures were used to benchmark the contamination module by fragmenting the monocultures into one 1600 kbp fragment and several 20 kbp fragments and then randomly assigning 20 kbp fragments to other bins (Fig. 2B, see methods). Nanomotif was able to achieve high sensitivity and precision with a mean of 89% and 91%, respectively. Most monocultures had near perfect contamination removal across all benchmarks (Fig. S6). We then applied the contamination detection module to the five real metagenomes of increasing complexity. The median number of contaminants in MAGs where at least one contaminant was detected was 1-2 for HQ MAGs. For example, bin 1.169 (HQ MAG) from the anaerobic digester (Fig. 2C) included contig_6001 (80 kbp) that completely lacked CAAAA**6mA** and G**6mA**TC methylation, despite the remaining bin being methylated at >75% in these motifs. In total, 196 contaminants were removed across 90 MAGs from the complex communities (Fig. S7). Each putative contaminant was manually reviewed, and in 84 out of 90 MAGs, the removal appeared accurate based on the methylation pattern, which matches the precision observed in the benchmark. This indicates a high potential for methylation to serve as a powerful post-binning cleanup, especially as this information is directly available for all new Nanopore sequencing projects.

The Nanomotif contig inclusion module assigns unbinned contigs to existing bins by training a linear discriminant analysis model, random forest, and k-neighbors classifier on the decontaminated bins (see methods). In case all three classifiers agree with a joint mean probability >0.8, the contig is assigned to the bin. Nanomotif achieved a high precision of 96% and moderate recall of 66% across the 28 fragmented monocultures described above (Fig. S8). In the five real complex metagenomic samples, the include module added a median of 1–4 contigs per bin for HQ MAGs. Associating mobile genetic elements with MAGs is of major importance as these can carry vital functionality[19]. This can be very difficult for traditional binners due to large variation in coverage or GC-content from the host, but should be possible if a unique

methylation motif signal is present. For example, using methylation motifs three contigs from the Simple Fecal sample were assigned to bin 1.7 and classified as two plasmids and a virus (Fig. 2d-e). It should be noted that the inclusion module is not a binner and assignments should be considered putative as methylation motif patterns can be shared across MAGs, which is also reflected in our efforts to prioritize precision over recall.

Restriction-modification (RM) systems are often substantial obstacles to genetic transformation, which pose a significant barrier for the implementation of novel bacteria as cell factories[20]. Circumventing these systems through RM system evasion or through heterologous expression of the methyltransferases in the cloning host (RM system mimicking) has shown to increase transformation efficiency significantly[20,21]. Therefore, we developed the Nanomotif MTase-linker module, which links methylation motifs to their corresponding MTase and, when present, their entire RM system (Fig. S9, supplementary note 2 & 3, and supplementary data 2). Across 11 monocultures, 52 putative orphan MTases and 29 RM-systems (exclusive type IV) were identified. 19 RM systems were associated with an active methylation motif, and a total of 42 out of 71 detected motifs could be linked to a single MTase gene or RM system with high confidence. Across 549 recovered HQ MAGs from five metagenomic samples, Nanomotif identified 3,123 putative MTase genes, of which 1,297 were associated with RM systems. For 76% of the detected motifs, at least one candidate MTase with matching methylation characteristics was identified within the same genome. Additionally, Nanomotif, successfully generated a high-confidence set of target motif annotations for 232 MTases. Hence, Nanomotif has the potential to drastically increase the number of putative links between motifs and MTase genes, thereby vastly improving the molecular toolbox and the RM system databases.

With Nanomotif, *de novo* motif discovery is now seamlessly possible with standard Nanopore sequencing, even for short and low coverage contigs from complex metagenomes. We provide simple implementations that utilize these motifs for robust identification of putative contamination in MAGs, association of mobile genetic elements to hosts, and linking of motifs to restriction-modification systems. This greatly enhances the resolution of metagenomic investigations, opening the possibility of linking extrachromosomal DNA elements to the host. This capability, which previously required additional, laborious methods, is now readily available with Nanomotif using standard Nanopore Sequencing.

Data availability

Sequencing data generated during the current study is available in the European Nucleotide Archive (ENA) repository, under the accession number PRJEB74343. Assemblies, bins, and output from Nanomotif are available at https://doi.org/10.5281/zenodo.10964193.

Code availability

Nanomotif is available at https://github.com/MicrobialDarkMatter/nanomotif. Code for reproducing figures and supplementary resources can be found at https://github.com/SorenHeidelbach/nanomotif-article.

Acknowledgements

The study was funded by grants from VILLUM FONDEN (130690, 50093), the Poul Due Jensen Foundation (Microflora Danica) and the European Research Council (101078234). We further acknowledge the Novo Nordisk Foundation within the framework of the Fermentation-

based Biomanufacturing Initiative (FBM), (Grant no. NNF17SA0031362), and the Novo Nordisk Foundation (Grant no. NNF20CC0035580).

Ethics

The simple fecal sample was collected as part of a study registered at ClinicalTrials.gov (Trial number NCT04100291). The study adhered to the Good Clinical Practice requirements and the Revised Declaration of Helsinki. The participant provided signed written informed consent to participate and allowed for the sample to be used in scientific research. Consent could be withdrawn at any time during the study period. Conduction of the study was approved by the Regional Research Ethics Committee of Northern Jutland, Denmark (project number N-20150021). The complex fecal sample was collected at Aalborg University with consent from the provider to be used in this study.

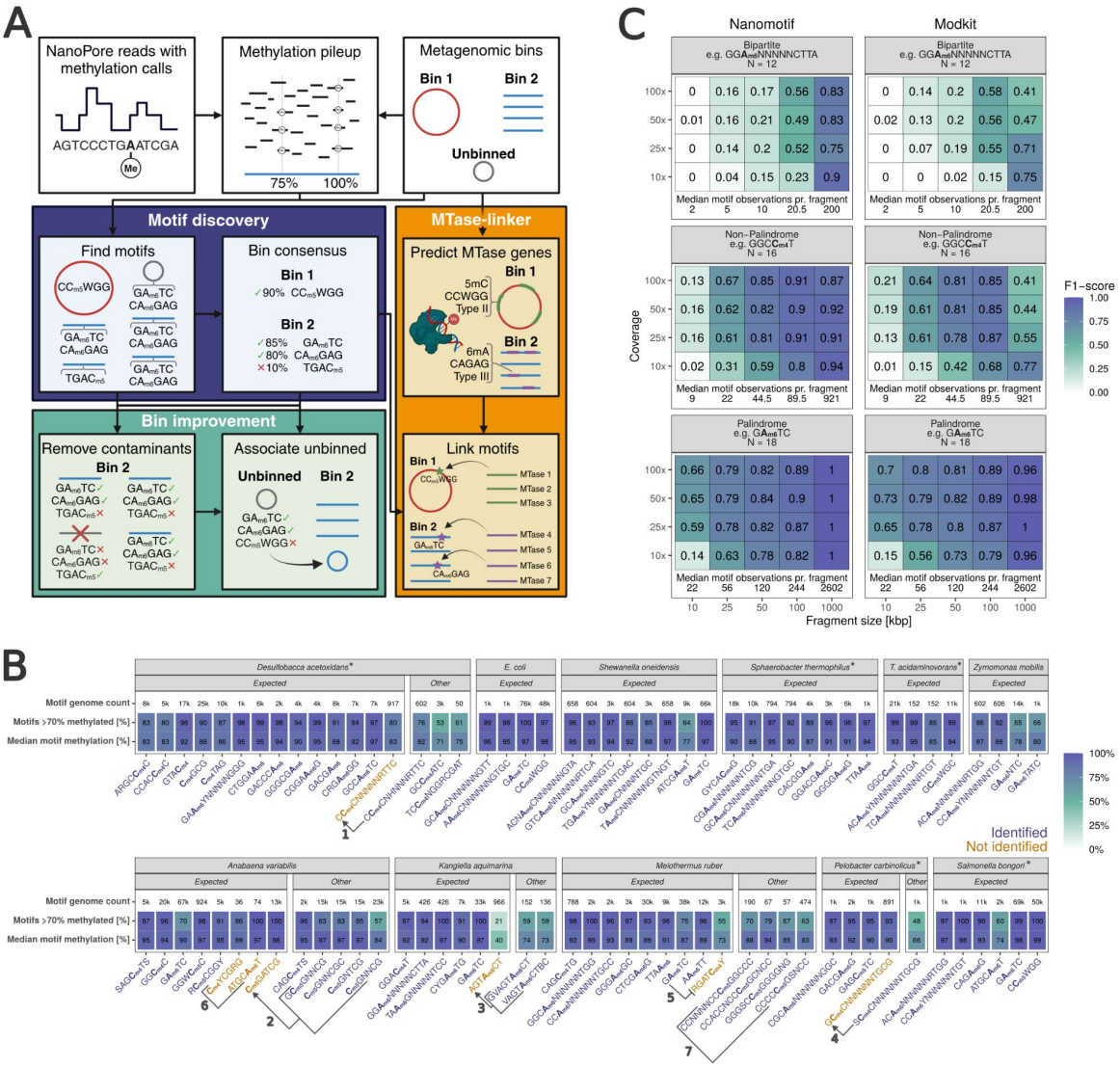


Fig. 1: Nanomotif overview and benchmark. a, Overview of Nanomotif functionality. White boxes on the top row are required inputs for Nanomotif, colored boxes are Nanomotif modules. **b,** Motif identification using Nanomotif in 11 monocultures with 75 known motifs, the 5 test monocltures are marked with *. For each motif three values are shown; the number of occurrences in the genome, the percent of motif positions with methylation >70% and the median methylation of motif positions. Motifs are grouped by whether or not they were expected to be found or not (See supplementary data 1.a). Seven cases of faulty motif identification have been annotated 1-7. **c,** Benchmarking of palindrome, bipartite, and non-palindromic motif identification with Nanomotif and Modkit[7] across different coverages and reference lengths. The number of expected motifs for each motif type is given in the header.

A

Sample	# Gbp (%binned)	# Contigs (%binned)	N50 [kbp]	# Bins MQ | HQ	% Bin w. motif MQ | HQ	Median motifs pr. Bin MQ | HQ*	Median contaminants MQ | HQ**	Median contigs included MQ | HQ**
Fecal, simple	4.2 (91%)	1584 (61%)	231	1 | 10	100% | 100%	1 | 3	1 | 2	19 | 4
ZymoFecal	25.5 (55%)	20562 (21%)	81	19 | 53	54% | 73%	1 | 3	2 | 1	8 | 3
Fecal, complex	131.1 (62%)	16315 (26%)	181	38 | 117	62% | 89%	2 | 4	2 | 1	2 | 2
Anaerobic digester	164.4 (69%)	51041 (21%)	102	41 | 205	85% | 88%	4 | 4	5 | 1	2 | 1
Soil	206.0 (22%)	114545 (8%)	91	159 | 30	63% | 59%	2 | 2	1 |NA	3 | 2

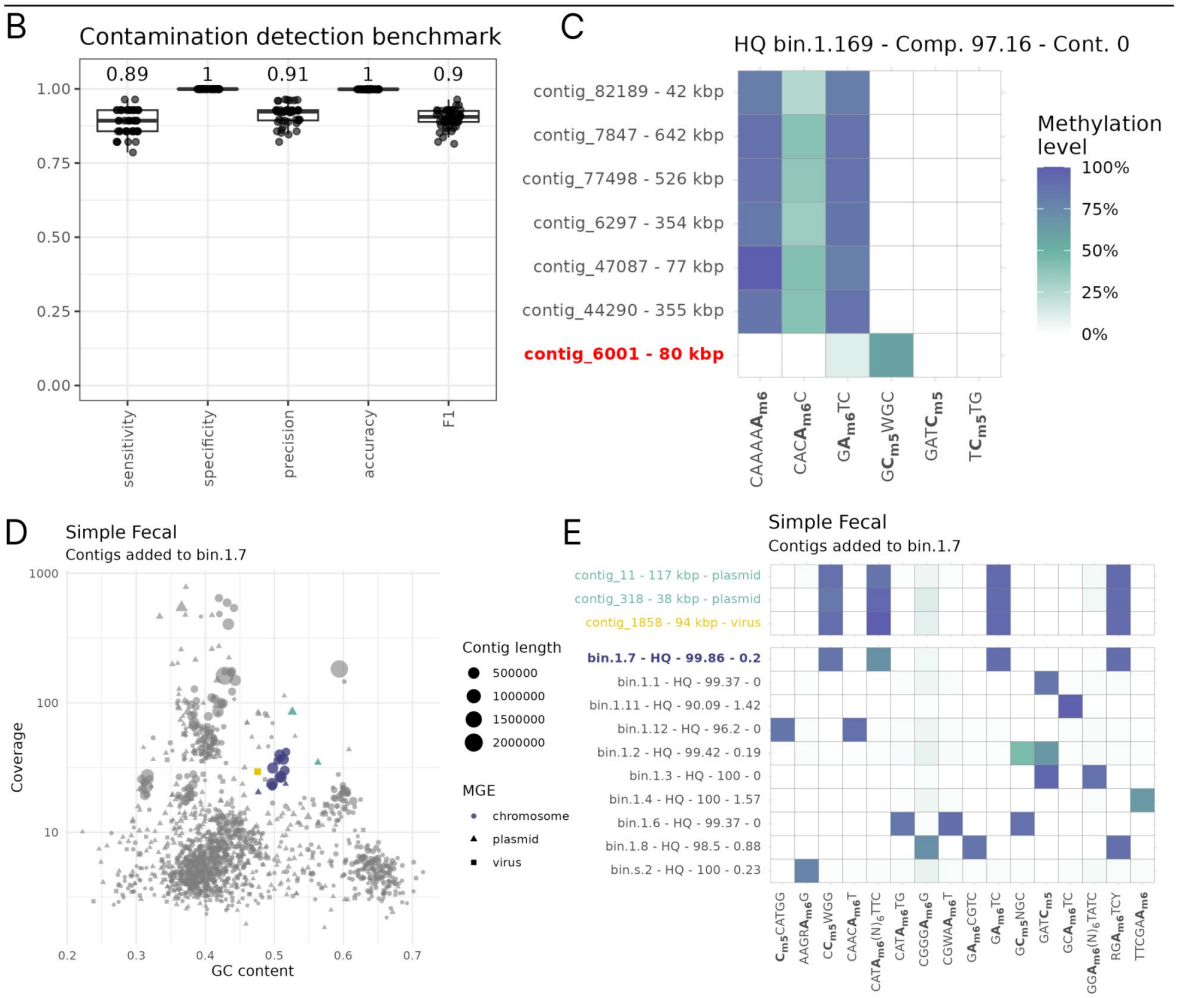


Fig. 2: Nanomotif MAG contamination detection and association of mobile genetic elements. a, Sample stats from binning and the Nanomotif modules. Only values for bins above 10x coverage are reported. *The median is reported for bins with at least one motif identified. **The median is reported for bins where contaminants were removed or contigs included. **b,** Nanomotif detect_contamination benchmark metrics from 50 benchmark datasets. 28 monocultures at 40x coverage were fragmented into one 1600 kbp fragment and several 20 kbp fragments. One randomly selected fragment was randomly assigned to another bin in each benchmark dataset. **c,** Example of contamination removal (red) from a HQ MAG from an anaerobic digester. **d,** GC% and coverage of bin.1.7 in the simple fecal sample (blue). Three contigs, predicted as two plasmids and one virus, are assigned to the bin with the Nanomotif include_contigs module. **e,** Methylation profile of the HQ bins in the simple fecal sample and highlighted plasmid & viral contigs assigned to bin.1.7.

Materials And Methods

Sampling

Escherichia coli K-12 MG1655 (labcollection), *Meiothermus ruber* 21 (DSM 1279), and *Parageobacillus thermoglucosidasius* DSMc 2542 were grown overnight in LB, DSMZ 256 Thermus ruber medium, and SPY medium, respectively. Genomic DNA for Desulfobacca acetoxidans ASRB2 (DSM 11109), Sphaerobacter thermophilus S 6022 (DSM 20745), Thermanaerovibrio acidaminovorans Su883 (DSM 6589), Kangiella aquimarina SW-154 (DSM 16071), Anabaena variabilis PCC 7120 (DSM 107007), Pelobacter carbinolicus Bd1 (DSM 2380) and Salmonella bongori 1224.72 (DSM 13772) was ordered from Leibniz-Institute DSMZ. Raw pod5 files for *Shewanella oneidensis MR-1, Cellulophaga lytica Cy I20, LIM -21 (DSM 7489), Kangiella aquimarina SW-154 (DSM 16071), Zymomonas mobilis subsp. pomaceae Barker I* and the remaining 15 monocultures used for contamination and inclusion benchmarks were acquired from s3://cultivarium-sequencing/MICROBEMOD-DATA-NOV2023/pod5/[8].

ZymoHMW, ZymoOral (D6332), ZymoGut (D6331) and ZymoFecal (D6323) were ordered from ZymoBIOMICS.

The simple fecal sample was collected at Aalborg University Hospital at the Department of Gastrointestinal Surgery as part of a clinical trial (ClinicalTrials.gov NCT04100291)[22]. The complex fecal sample was collected at Aalborg University with consent from the provider.

Sampling of the anaerobic digester sludge has been described elsewhere[23]. Sampling of soil has been described elsewhere[24].

DNA Extraction

DNA from cell pellets of overnight grown cultures of *E. coli* K-12 MG1655 and *M. ruber* 21 was extracted with the PureLink Genomic DNA mini kit (Invitrogen, Thermo Fisher Scientific, USA) following manufacturer's instructions with final elution in DNAse/RNAse free water. DNA from cell pellets of *P. thermoglucosidasius (*DSM 2542) was extracted with the MasterPure Gram positive DNA purification kit (Biosearch Technologies (Lucigen)), according to manufacturer's instructions with a 60 min incubation step and final elution in DNAse/RNAse free water. DNA from ZymoOral (D6332), ZymoGut (D6331) and ZymoFecal (D6323) was extracted with the DNeasy PowerSoil Pro kit according to manufacturer's instructions and suppliers suggestions. DNA from the simple fecal sample was extracted with the DNeasy PowerSoil Pro kit as described previously[25]. DNA from Complex fecal sample was extracted using DNeasy PowerSoil Pro kit according to manufacturer's instructions. DNA was extracted from the anaerobic digester as described previously[23].

Sequencing

All samples were sequenced on the Promethion24 using the R10.4.1 nanopore. Anaerobic digester, complex fecal and ZymoFecal (D6323) were prepared with SQK-LSK114. Monocultures, ZymoOral (D6332), ZymoGut (D6331) were prepared with SQK-RBK114.24. ZymoHMW was prepared with the SQK-NBD114-24 ligation kit. Sequencing of simple fecal is described elsewhere[22]. Sequencing of soil is described elsewhere[24]. All samples were basecalled with Dorado v0.8.1 using the dna_r10.4.1_e8.2_400bps_sup@v5.0.0 model and DNA methylation was called with the respective v2 methylation models for 4mC_5mC and 6mA.

Assembly and binning

All samples were assembled and binned using the mmlong2-lite v1.1.0 pipeline available at [26]. Briefly, metaFlye (v2.9.4)[27] is used for assembly and eukaryotic contigs are removed with Tiara (v1.0.3)[28] before assembly coverage is calculated with read mapping via minimap2 (v2.28)[29]. Binning is performed iteratively as an ensemble using SemiBin2 (v2.1.0)[30], MetaBat2 (v2.15)[31], VAMB (v3.0.3)[32], and COMEbin (v1.0.4)[33] whereafter the best bin is chosen with DAS tool (v1.1.3)[34]. MAG quality was classified according to the MIMAG definition (Bowers et al., 2017). Completeness and contamination were evaluated with CheckM2 while rRNA and tRNA genes were found with barrnap (v0.9, https://github.com/tseemann/barrnap) and tRNAscan-SE (v2.0.16)[35], respectively.

Methylation pileup

Reads with methylation calls were mapped to the assembly using minimap2 v2.24[29] using default settings. Nanopore's modkit v0.4.0 (https://github.com/nanoporetech/modkit) was used to generate the methylation pileup from mapped reads using default settings.

Motif identification

MicrobeMod v1.0.3 with default settings was used for all motif identification experiments. motifMaker (smartlink v13.1.0) with default settings was used for all motif identification experiments. Modkit pileup is not directly compatible with motifMaker and had to be converted to the same format as the output of ipdSummary. As the goal was a comparison of the motif identification algorithm, we extracted generally methylated positions (>70% methylated) and generated a GFF formatted file similar to the output of ipdSummary, marking all extracted positions with high Q-score and IPD Ratio. Modkit v0.4.0 was used for all motif identification experiments. Default parameters were used for full genomes and scalability experiment. The setting --min-sites 20 was used for benchmark with lowered coverage and fragmentation of reference. This was done for fair comparison to Nanomotif minimum motif count of 20.

Nanomotif v0.4.16 was used for all experiments. Nanomotif motif discovery algorithm has two main submodules, "find-motifs" and "bin-consensus". All subcommands are gathered in a parent command "motif_discovery", which was executed with the following arguments for all samples: threshold_methylation_confident=0.8, threshold_methylation_general=0.7, search_frame_size=41, threshold_valid_coverage=5, minimum_kl_divergence=0.05, min_motif_score=0.2. "find-motifs" identifies motifs in contigs, referred to as directly identified motifs. This is done using a greedy search and candidates are selected based on a Beta-Bernoulli model, where each motif occurrence is Bernoulli trial, being a success if the fraction of methylation of reads at the position is above a predefined threshold (default 0.70).

The Beta-Bernoulli was chosen in order to include uncertainty in the motif scoring process, instead of a point estimate. The exact steps performed for motif identification is outlined in the pseudo code below. For full details see supplementary note 1.

```
INPUT:
   assembly - List of contigs sequences
   modkitPileup - Methylation pileup data
OUTPUT:
   Identified methylated motifs
BEGIN
| Initialize identifiedMotifs as an empty list
| FOR sequence IN assembly DO
| | FOR sequencePosition IN sequence DO
| | | IF coverage(sequencePosition) < coverageThreshold (default: 5) THEN
| | | | Mark position as NA
| | | ELSE IF fractionModified(sequencePosition) < methylatedThreshold (default: 0.7) THEN
| | | | Mark position as not methylated
| | | ELSE
| | | | Mark position as methylated
| | | | IF fractionModified(position) >= confidentlyMethylatedThreshold (default: 0.8) THEN
| | | | | Mark position as confidently methylated
| | Initialize seedMotif based on methylation type (C for 5mC, C for 4mC, A for 6mA)
| | currentMotif = seedMotif
| | sampledSeqs = Extract n sequences randomly from sequence (default n: 10,000)
| | sampledPPM = Positional nucleotide probability matrix of sampledSeqs
| | WHILE stopping criteria NOT met DO
| | | methylatedSeqs = Extract sequences at confidently methylated positions (default window: 41)
| | | Remove from methylatedSeqs any sequences that match any motif in identified motifs
| | | WHILE stopping criteria NOT met DO
| | | | Initialize motifCandidates as an empty list
| | | | methylatedPPM = Calculate positional nucleotide probability matrix of methylatedSeqs
| | | | FOR position IN methylatedPPM DO
| | | | | Compute KL-divergence from methylatedPPM[, position]  to sampledPPM[, position]
| | | | | IF KL-divergence > threshold (default: 0.05) THEN
| | | | | | Identify valid bases WHERE
| | | | | | | - methylatedPPM[, position] > 25%
| | | | | | | - methylatedPPM[, position] > sampledPPM[, position]
| | | | | | FOR baseCombination IN validBases DO
| | | | | | | newMotif = expand currentMotif with baseCombination at position
| | | | | | | add newMotif to motifCandidates
| | | | FOR motifCandidate IN motifCandidates DO
| | | | | Compute Beta-Bernoulli posterior parameters
| | | | | Compute score and priority
| | | | Update currentMotif to lowest-priority motifCandidate
| | | | IF highestScoringMotif.score < currentMotif.score THEN
| | | | | highestScoringMotif = currentMotif
| | | | Subset methylatedSeqs to sequences matching new currentMotif
| | | | Stopping criteria:
| | | | | highestScoringMotif.score not improved for n round (default: 10)
| | | IF highestScoringMotif.score > 0.2 THEN
| | | | Add highestScoringMotif to identifiedMotifs
| | | | Exclude sequences containing identified motifs
| | Stopping criteria:
| | | - 25 low-scoring motifs dropped
| | | - <1% methylation sequences remain
| | Remove motifs contained within more generic motifs
| | Remove motifs with isolated bases
| | mergeableMotifs  = motifs with similar sequences (Hamming distance ≤ 2):
| | FOR motifPair IN mergeableMotifs DO
| | | IF methylation of merged motif ≥ methylation pre-merge motifs THEN
| | | | Accept merged motif
| | | ELSE
```

```
| | | | Retain pre-merge motifs
| | Identify complementary motifs
| | Add processed motifs for contig to identifiedMotifs
| RETURN identifiedMotifs
END
```

After identification of motifs in each contig “bin-consensus” evaluates the full set of identified motifs in the contigs belonging to the bin. It performs motif merging like the post processing steps in the motif identification algorithm, but for the bin motif set. Then motifs which are not methylated more than 25% methylated in 75% of the bin are removed.

Motif Identification Benchmark

Motif identification was benchmarked using motifs identified in 11 monocultures with known methylated motifs. Full set of expected motif and motif evidence is available in supplementary data 1.a. Six monocultures were used for parameter justification of the “find-motifs” algorithm; *Shewanella oneidensis MR-1, Kangiella aquimarina DSM 16071, Anabaena variabilis ATCC 27893, Escherichia coli K-12 substr. MG1655, Meiothermus ruber DSM 1279, Zymomonas mobilis subsp. pomaceae.* A grid search was performed over the three most important parameters, to justify the final parameter settings (Fig. S10). The algorithm performance is stable across the selected parameters, and the final set parameters were chosen to increase sensitivity in metagenomic settings. Five monocultures were used for testing; *Desulfobacca acetoxidans DSM 11109, Salmonella bongori NCTC 12419, Sphaerobacter thermophilus DSM 20745, Pelobacter carbinolicus, Thermanaerovibrio acidaminovorans DSM 6589*. Testing monocultures were not seen during tuning.

Benchmark metrics were calculated by comparing identified motif with expected motif in monocultures. If an identified motif matches an expected motif exactly, it counts as a true positive. False positives are counted as motifs not in the expected motif set. False negatives are counted as motifs in the expected set which are not identified. Both forward and reverse complement are counted as a motif for motifs which are not palindromic. For benchmarks, precision, recall and F1-score are reported.

Reduced information benchmarking was performed across two parameters; read coverage (10, 25, 50, 100x) and contig size (10, 25, 50, 100, 1000 kbp). Read coverage affects false positive and false negative in calling of generally methylated positions, as lower coverage is more sensitive to non-systematic false positive and false negative calls at the reads level. Lower coverage was achieved using Rasusa[28] by subsetting the total length of reads to a multiple of the assembly length of the respective benchmarking organisms. As contig size is proportional with motif occurrences, smaller contigs will have fewer motif observations, thereby less information for motif identification. Differing contig sizes were created by chunking the reference genome of the monoculture to fix sized windows using SeqKit2 (v2.5.1)[36] (windows were not allowed to overlap). Up to 20 chunks from the procedure for each monoculture were included in the benchmark. To fairly compare across fragmentation lengths, we reduced the minimum required motif observations to 20, the same requirements Nanomotif utilises for motif identification in contigs.

To benchmark execution time of Modkit on metagenomic samples, we split the pileup file into separate files, each containing the information related to the contigs of a single bin. Then the assembly was split into fasta files, each containing the contig sequences related to a single bin. Execution time for preprocessing of files was not included in the reported run time or CPU

hours. To identify motifs in a bin, we executed Modkit on the files corresponding to a single bin. Modkit was executed with a single thread, and parallelly executed in 144 instances.

Contamination detection

Contamination is evaluated using “nanomotif detect_contamination” which uses ensemble clustering. In case all four clustering methods, HDBSCAN, Gaussian Mixture Model, Agglomerative Clustering, and spectral clustering marks a contig as an outlier, the contig is marked as contamination.

Firstly, motif methylation is calculated as follows: The mean read methylation for each motif observation is calculated for motif observations with at least 3 read mappings. The median value of these is then reported as the motif methylation. This methylation value was more robust for smaller contigs compared to the mean of means. Before clustering, motif methylation is filtered for each contig if the product of number of motif observations and the mean read coverage is less than 24. This way the methylation value of a contig with a one motif observation is trusted if the read coverage is at least 24 or the contig has at least 8 motif observations. After filtering, contigs with missing motif values are imputed with the bin mean and PCA is performed to reduce dimensions while retaining 90% of variance. Contigs are then clustered with HDBSCAN (min_samples=3, min_cluster_size=2), Gaussian mixture model (n_components = n_bins, covariance_type=”full”), Agglomerative clustering (n_clusters = n_bins, affinity = “nearest_neighbors”) and spectral clustering (n_clusters = n_bins). For each clustering method, the bin cluster is the cluster with the largest fraction of the bin length which must constitute at least 85% of the total bin length. In case all methods agree a contig does not belong to the bin, the contig is flagged as a contaminant. Contamination contigs are then assigned to the “unbinned” pool.

Include contigs

The “nanomotif include_contigs” module will attempt to assign unbinned contigs to the bin after decontamination. The include_contigs module uses an ensemble of supervised machine learning techniques; random forest (n_estimators=100), linear discriminant analysis (solver = ”svd”), and k-nearest neighbors classifier (n_neighbors = 3) classifier to assign unbinned contigs to bins. Firstly, the three classifiers are trained on the binned contigs after dimensionality reduction (see contamination detection). Missing motif observations in unbinned contigs are then imputed with a pseudo value randomly chosen between 0 and 0.15, whereafter features are z-score normalized and projected using the binned conversion matrices. A contig is assigned to a bin if all three classifiers agree and the mean probability of the three classifiers is above 0.80.

Contamination and inclusion benchmark

A synthetic benchmark dataset was constructed for developing and evaluating the contamination and inclusion module. 28 monocultures (see Fig S6) were sampled to 40x coverage and fragmented into 20 kbp fragments. For each fragmented monoculture, a single long 1600 kbp contig was reconstructed by stitching together 80 consecutive 20 kbp fragments, while the remaining fragments were retained unaltered. Nanomotif were then used to find motifs anew and contamination and inclusion were evaluated using the found motifs.

To evaluate the contamination module 50 datasets were created where one randomly chosen contig from each monoculture was randomly added to another. For the inclusion benchmark 50 datasets were created where a random contig was removed from each monoculture. The 1600 kbp contig was not shuffled or removed.

MTase-Linker

The Nanomotif MTase-linker module initially uses Prodigal v.2.6.3[37] for protein-coding gene prediction (default settings) followed by DefenseFinder v1.2.0[38] to predict MTases and related RM-system genes. The output file defense_finder_hmmer.tsv is filtered for all RM-related MTase hits. When a single gene has several model hits, the model that yields the highest w. The output file defense_finder_systems.tsv is used to determine whether the identified MTase hit is part of a complete defense system. MTase hits associated with non-methylation-mediated defense systems are excluded. Additionally, RM type IIG MTase hits not identified by DefenseFinder as part of a complete RM system are also removed.

Using hmmer (with parameter –cut_ga) the predicted MTase protein sequences are queried against a set of hidden markov models (PF01555.22, PF02384.20, PF12161.12, PF05869.15, PF02086.19, PF07669.15, PF13651.10, PF00145.21) from the PFAM database[39], to predict the modification type (5mC or 6mA/4mC). Furthermore, to infer the probable target recognition motif, the MTase protein sequences are queried using BLASTP (Blast v.2.14.1) against a custom database of methyltransferases with known target recognition motif from REbase[40]. We employ a threshold of 80% sequence identity and 80% query coverage to confidently predict the target recognition motif. Lastly, the RM system, RM sub-type, mod-type, and predicted motif information for each methyltransferase gene are used to link methylation motifs to the genes. The pipeline identifies high confidence MTase-motif matches, labeled as “linked”, through either a precise match between the predicted motif and the detected motif or when a single gene and a single motif share a similar combination of methylation features, which are unique within a MAG. When a high confidence match cannot be elucidated, the MTase-Motif-linker assigns feasible candidate genes, with the corresponding motif type and modification type, for each motif.

MGE classification

Contig were labeled as Mobile genetic elements (MGE) when classified as viral or plasmidal by GeNomad (v1.7.0) with a score >0.75, had a mean coverage above 10x and a minimum length of 10kbp.

References

1. Seong, H. J., Han, S.-W. & Sul, W. J. Prokaryotic DNA methylation and its functional roles. *J. Microbiol.* **59**, 242–248 (2021).

2. Beaulaurier, J. *et al.* Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.* **36**, 61–69 (2018).

3. Seong, H. J., Roux, S., Hwang, C. Y. & Sul, W. J. Marine DNA methylation patterns are associated with microbial community composition and inform virus-host dynamics. *Microbiome* **10**, 157 (2022).

4. Oliveira, P. H. & Fang, G. Conserved DNA Methyltransferases: A Window into Fundamental Mechanisms of Epigenetic Regulation in Bacteria. *Trends Microbiol.* **29**, 28–40 (2021).

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

6. Liu, J.-H. *et al.* Bacmethy: A novel and convenient tool for investigating bacterial DNA methylation pattern and their transcriptional regulation effects. *Imeta* **3**, e186 (2024).

7. Tue Kjærgaard Nielsen *et al.* Detection of nucleotide modifications in bacteria and bacteriophages; strengths and limitations of current technologies and software. *Molecular Ecology. 2023;32:1236–1247.* (2022) doi:10.1111/mec.16679.

8. Crits-Christoph, A., Kang, S. C., Lee, H. H. & Ostrov, N. MicrobeMod: A computational toolkit for identifying prokaryotic methylation and restriction-modification with nanopore sequencing. *bioRxiv* (2023) doi:10.1101/2023.11.13.566931.

9. Tidwell, A. K., Faust, E., Eckert, C. A., Guss, A. M. & Alexander, W. Discovering methylated DNA motifs in bacterial nanopore sequencing data with MIJAMP. *bioRxiv* 2024.08.14.607972 (2024) doi:10.1101/2024.08.14.607972.

10. *Modkit: A Bioinformatics Tool for Working with Modified Bases*. (Github).

11. Wilbanks, E. G. *et al.* Metagenomic methylation patterns resolve bacterial genomes of unusual size and structural complexity. *ISME J.* **16**, 1921–1931 (2022).

12. Tourancheau, A., Mead, E. A., Zhang, X.-S. & Fang, G. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat. Methods* **18**, 491–498 (2021).

13. Li, T., Zhang, X., Luo, F., Wu, F.-X. & Wang, J. MultiMotifMaker: A Multi-Thread Tool for Identifying DNA Methylation Motifs from Pacbio Reads. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**, 220–225 (2020).

14. Hiraoka, S. *et al.* Diverse DNA modification in marine prokaryotic and viral communities. *Nucleic Acids Res.* **50**, 1531–1550 (2022).

15. Hiraoka, S. *et al.* Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an environmental prokaryotic community. *Nat. Commun.* **10**, 159 (2019).

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

17. Orakov, A. *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 178 (2021).

18. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).

19. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732 (2005).

20. Johnston, C. D. *et al.* Systematic evasion of the restriction-modification barrier in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 11454–11459 (2019).

21. Yasui, K. *et al.* Improvement of bacterial transformation efficiency using plasmid artificial modification. *Nucleic Acids Res.* **37**, e3 (2009).

22. Kousgaard, S. J. *et al.* The effect of non-pooled multi-donor faecal microbiota

transplantation for inducing clinical remission in patients with chronic pouchitis: Results from a multicentre randomised double-blinded placebo-controlled trial (MicroPouch). *J. Crohns. Colitis* (2024) doi:10.1093/ecco-jcc/jjae0

23. Sereika, M. *et al.* Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* **19**, 823–826 (2022).

24. Singleton, C. M. *et al.* Microflora Danica: the atlas of Danish environmental microbiomes. *bioRxiv* (2024) doi:10.1101/2024.06.27.600767.

25. Jensen, T. B. N., Dall, S. M., Knutsson, S., Karst, S. M. & Albertsen, M. High-throughput DNA extraction and cost-effective miniaturized metagenome and amplicon library preparation of soil samples for DNA sequencing. *PLoS One* **19**, e0301446 (2024).

26. Sereika, M. Recovery of novel microbial genomes. (Aalborg University, 2024). doi:10.54337/aau718059137.

27. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

28. Karlicki, M., Antonowicz, S. & Karnkowska, A. Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics* **38**, 344–350 (2022).

29. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

30. Pan, S., Zhao, X.-M. & Coelho, L. P. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics* **39**, i21–i29 (2023).

31. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

32. Nissen, J. N. *et al. Binning Microbial Genomes Using Deep Learning*. http://biorxiv.org/lookup/doi/10.1101/490078 (2018).

33. Wang, Z. *et al.* Effective binning of metagenomic contigs using contrastive multi-view representation learning. *Nat. Commun.* **15**, 585 (2024).

34. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).

35. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).

36. Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss army knife for sequence and alignment processing. *Imeta* **3**, e191 (2024).

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

38. Tesson, F. *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* **13**, 2561 (2022).

39. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

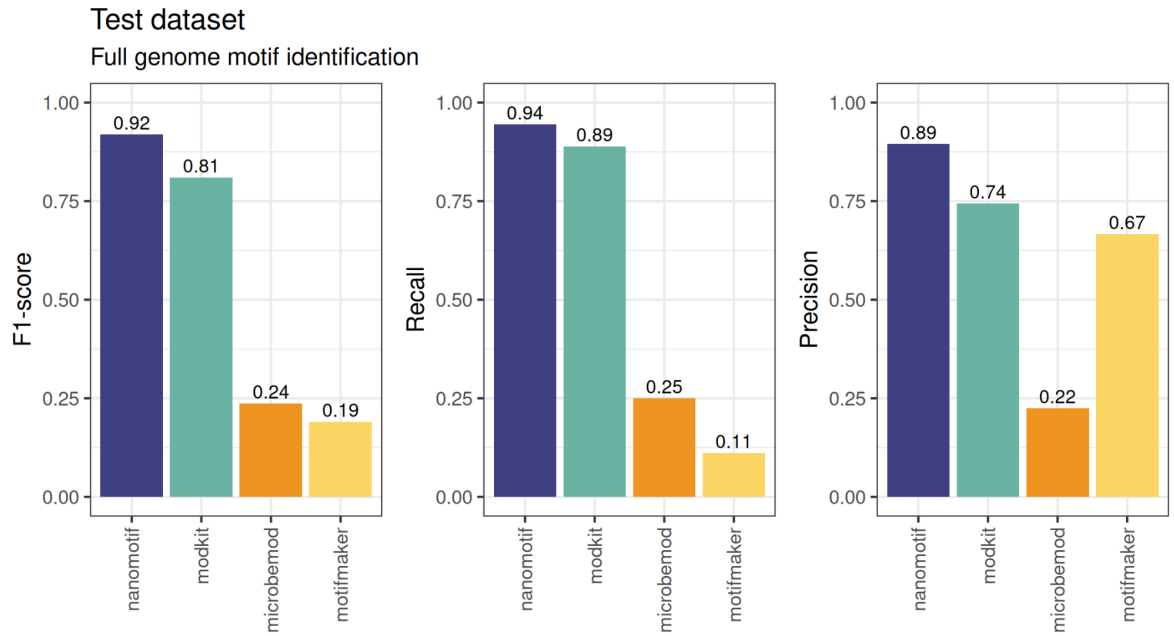
40. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE: a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **51**, D629–D630 (2023).

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

Supplementary Figures

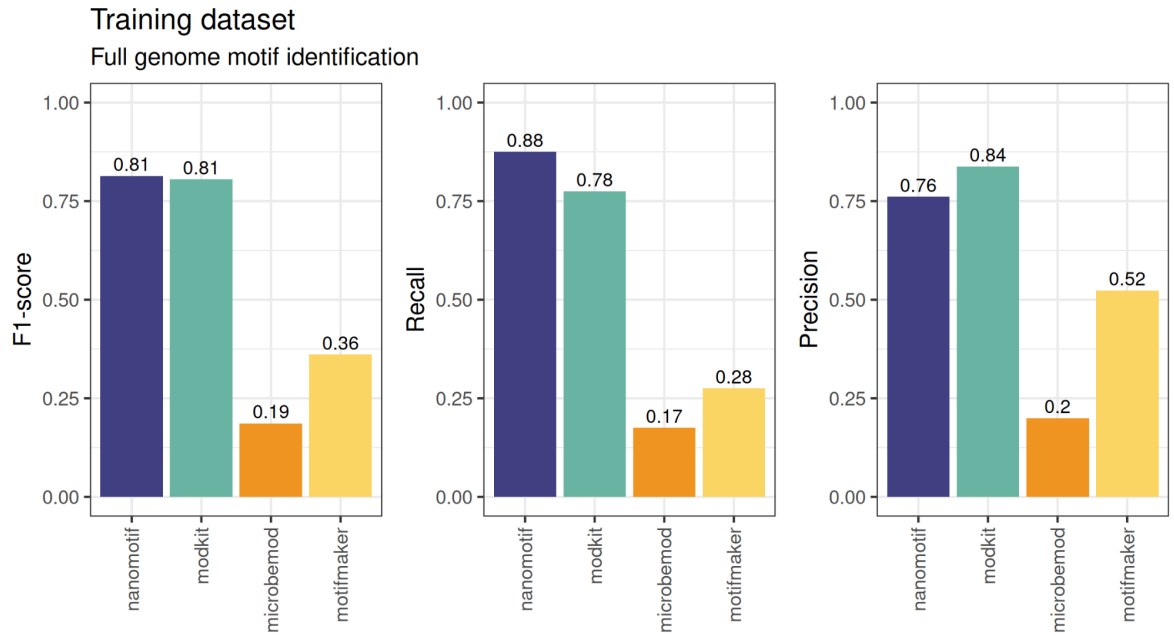
Dataset	***Tool***	***CPU Time [h]***	***Total Time [h:m:s]***	***Time compared to Nanomotif***
Fecal, simple	modkit	6.7	6:32:03	120x
	nanomotif	0.8	0:03:14	1x
Fecal, complex	modkit	18.0	1 day, 2:11:43	35x
	nanomotif	17.0	0:44:15	1x
ZymoFecal	modkit	14.0	14:28:34	38x
	nanomotif	7.6	0:22:17	1x
Anaerobic Digester	modkit	86.0	1 day, 23:51:30	23x
	nanomotif	65.0	2:02:50	1x

Tab. S1: Benchmark of Nanomotif and Modkit motif identification performance. Running modkit was infeasible for the soil sample and is hence not included. Total time benchmark was performed using 144 AMD EPYC 7H12 CPUs.



565

566 **Fig. S1:** Full genome motif identification benchmark on 46 expected motifs from the five test
567 monocultures.

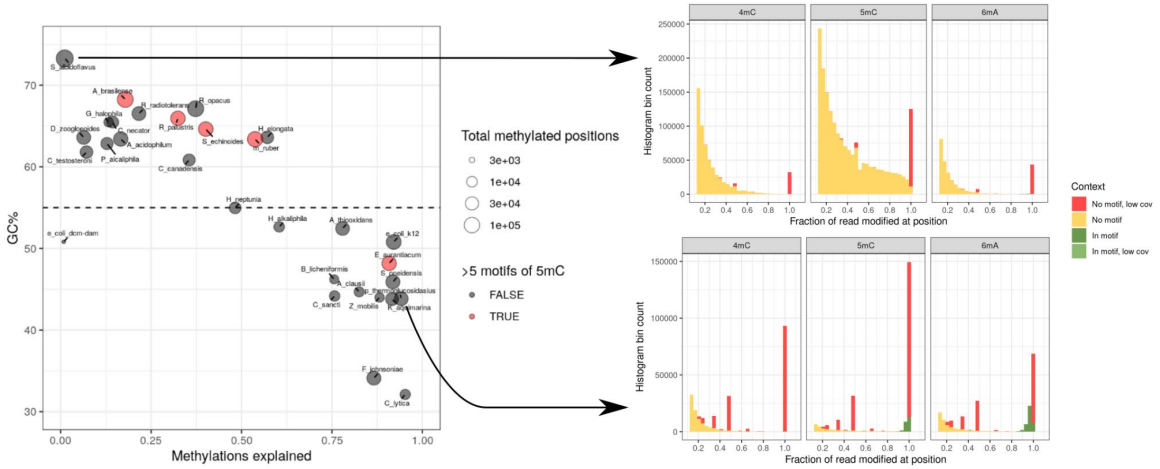


568

569

570 **Fig. S2:** Full genome motif identification benchmark on 29 expected motifs from the six training
571 monocultures.

572



573

574 **Fig. S3:** High 5mC false positives in high GC% species. Left panel; On the x-axis is the fraction
575 of methylated positions occurring explained by a motif and the GC% of the organism genome.
576 As the GC% increases, the degree of explained methylated positions decreases. Additionally,
577 large degrees of false positive 5mC motif are identified in the high GC% organisms. Right
578 panel shows two organisms: 1) *S. albidoflavus*, with a GC% of 74%, where 5mC has a con-
579 tinuous drop off in fraction of read modified at C positions, while none of these positions are
580 explained by a 5mC motif. 2) *P. thermoglucosidasius*, with a GC% of 44%, where 5mC distri-
581 bution is split into two groups; a low fraction group with no motif associated and a high fraction
582 group, all explained by a motif.

Precision

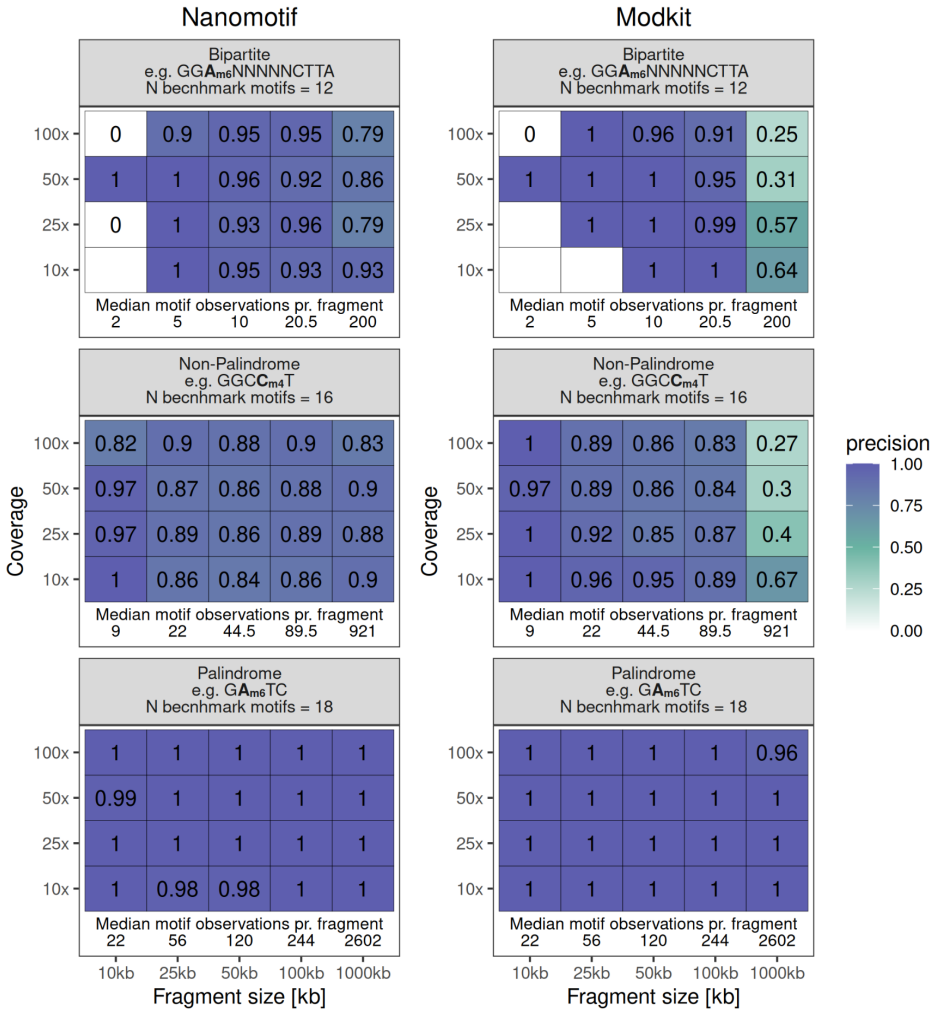
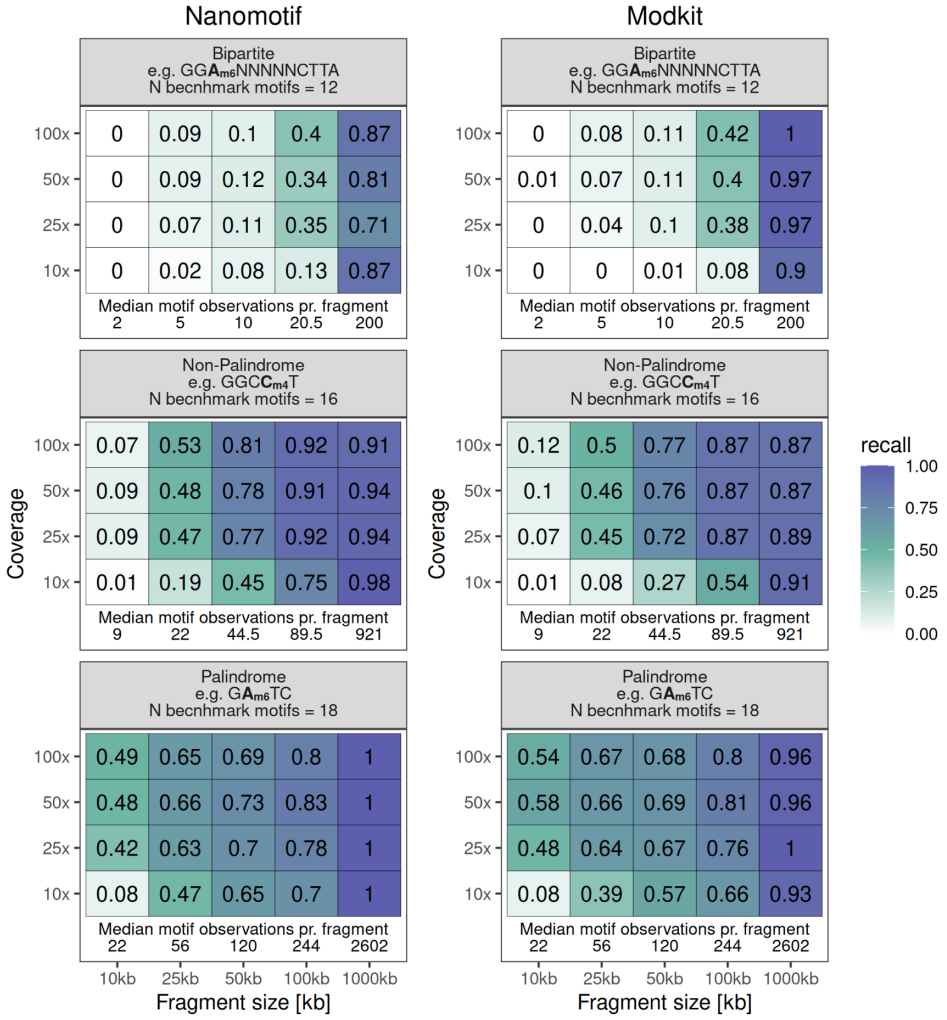


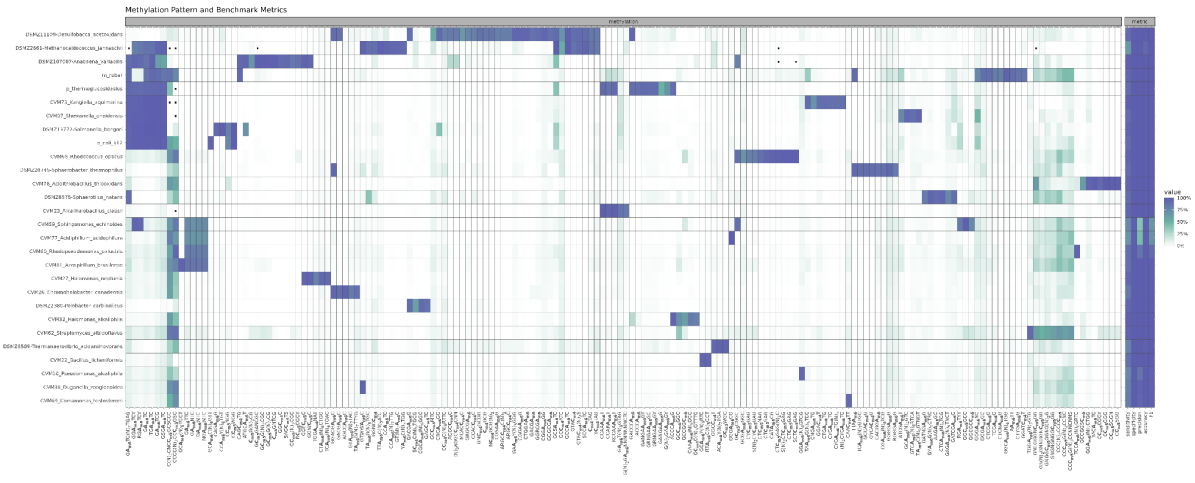
Fig. S4: Precision of benchmark presented in figure 1.C.

Recall



585

586 **Fig. S5:** Recall of benchmark presented in figure 1.C.



587

588 **Fig. S6:** Methylation pattern of monocultures included in the benchmark dataset along with
589 mean sensitivity, specificity, precision, accuracy, and F1 score across the 50 datasets where
590 a contig from each bin has been assigned to another.

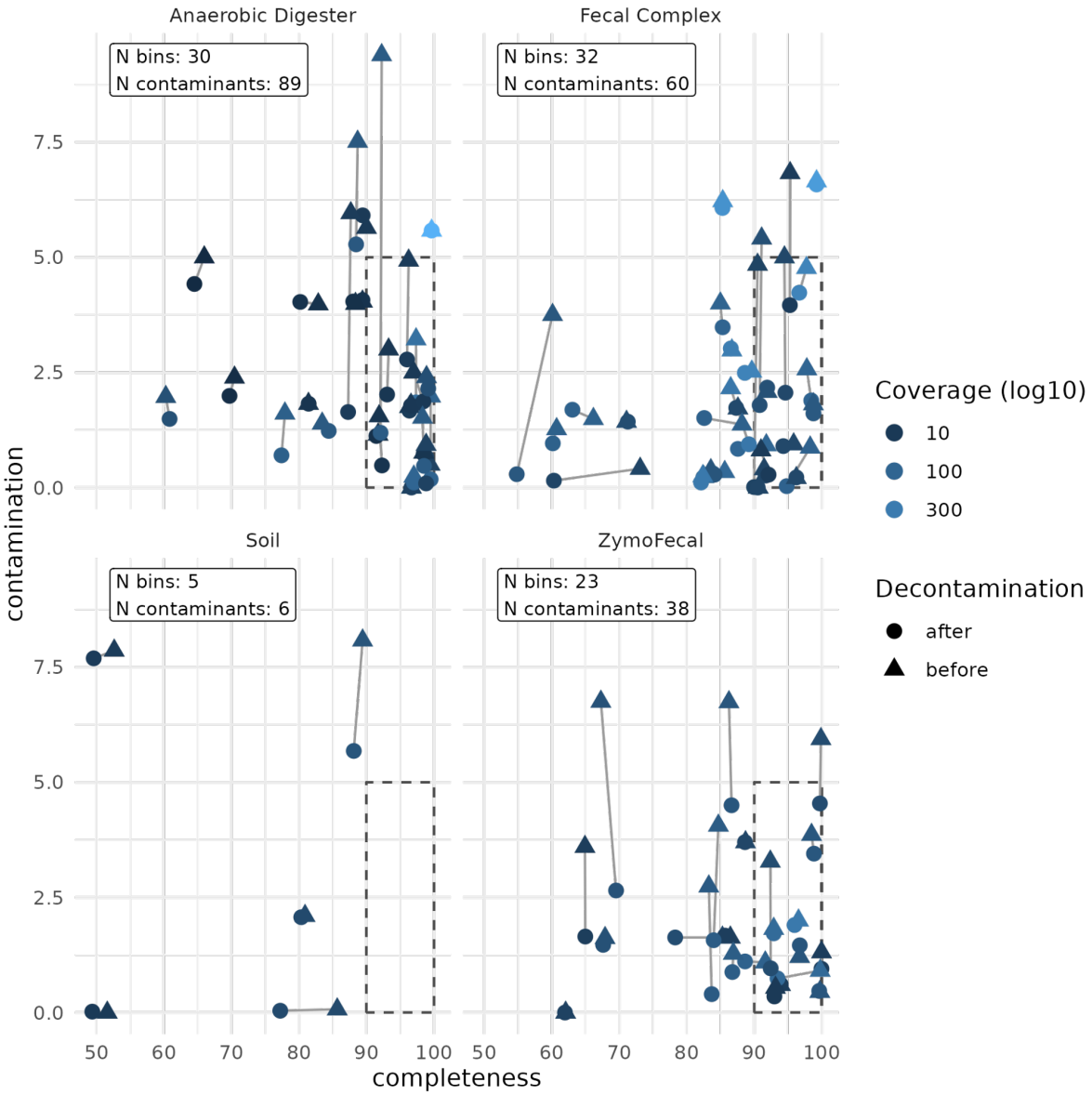


Fig. S7: Completeness and contamination of MQ and HQ bins before and after removal of putative contamination. Dashed boxes mark completeness >= 90%, contamination <5%. One contaminant was removed in Fecal Simple but is not shown.

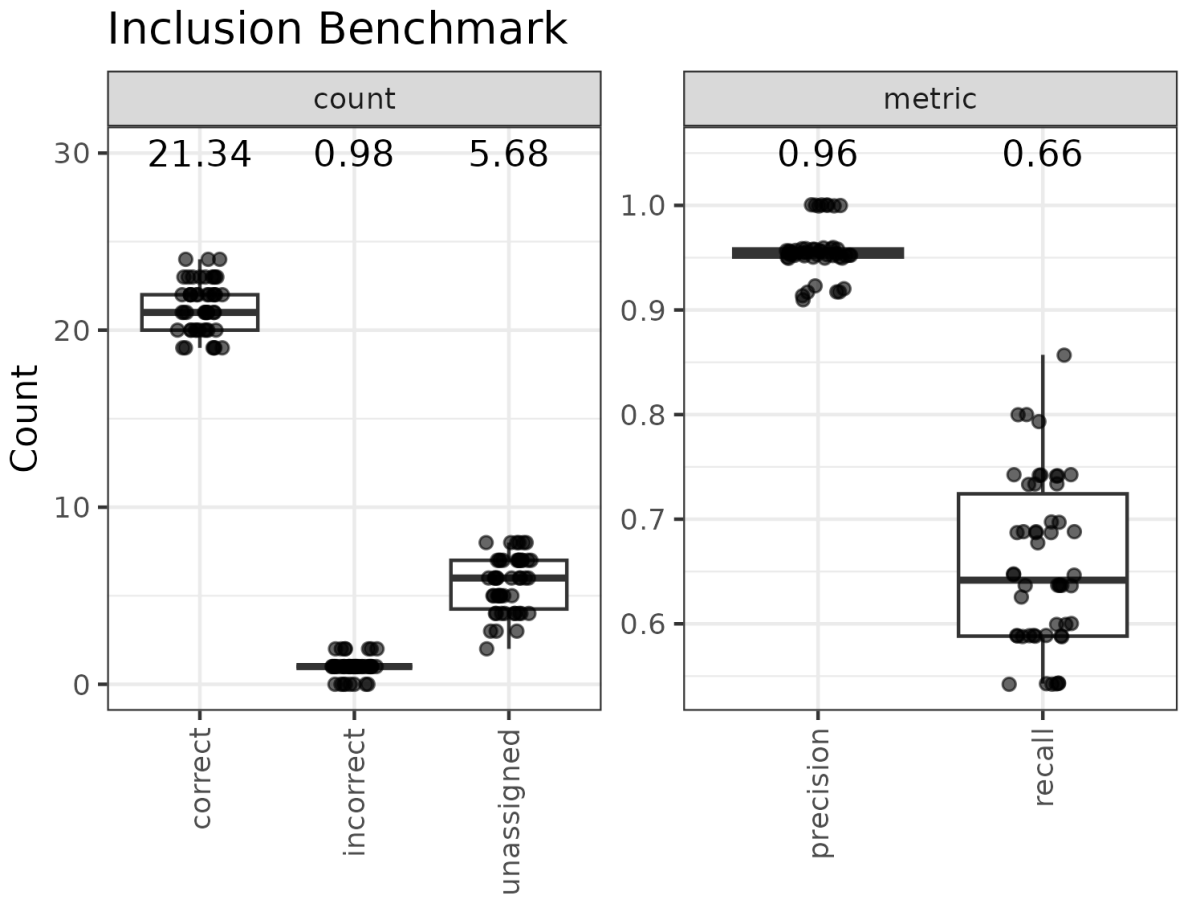


Fig. S8: Number of correct, incorrect and unassigned classification from nanomotif include module using the monoculture benchmark.

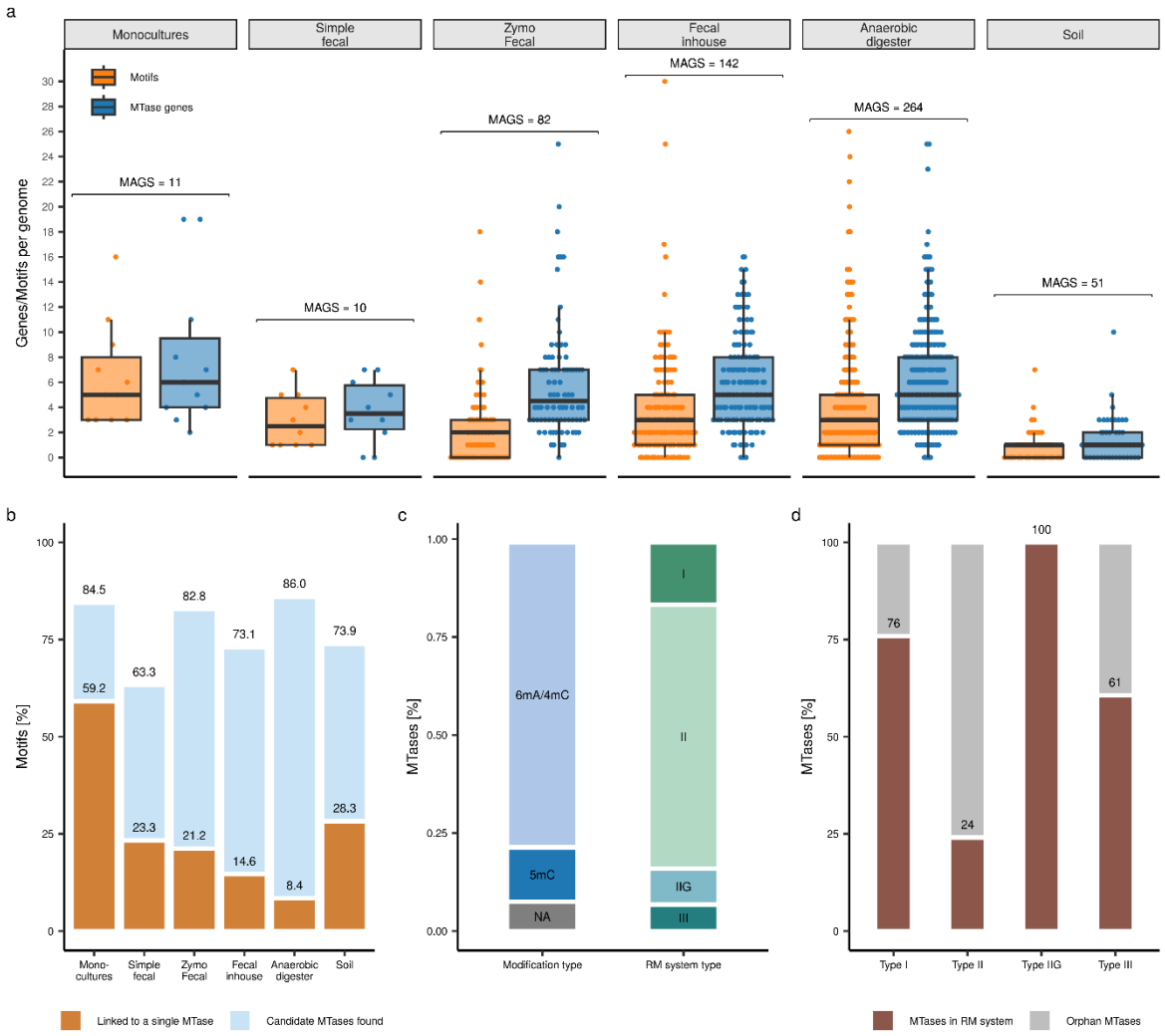


Fig. S9: Summary of putative MTase annotations and motif assignments in 11 monocultures and recovered HQ MAGs from five metagenomes. **a**, Distribution of detected motifs and MTase genes per. genome. **b**, Percentage of motifs with a high confidence link to a MTase gene (orange), or motifs for which one or multiple candidate genes have been found (blue). **c**, Breakdown of MTases by modification type and RM-system type. NA indicates unidentified modification type. **d**, Proportion of MTases involved in RM-systems.

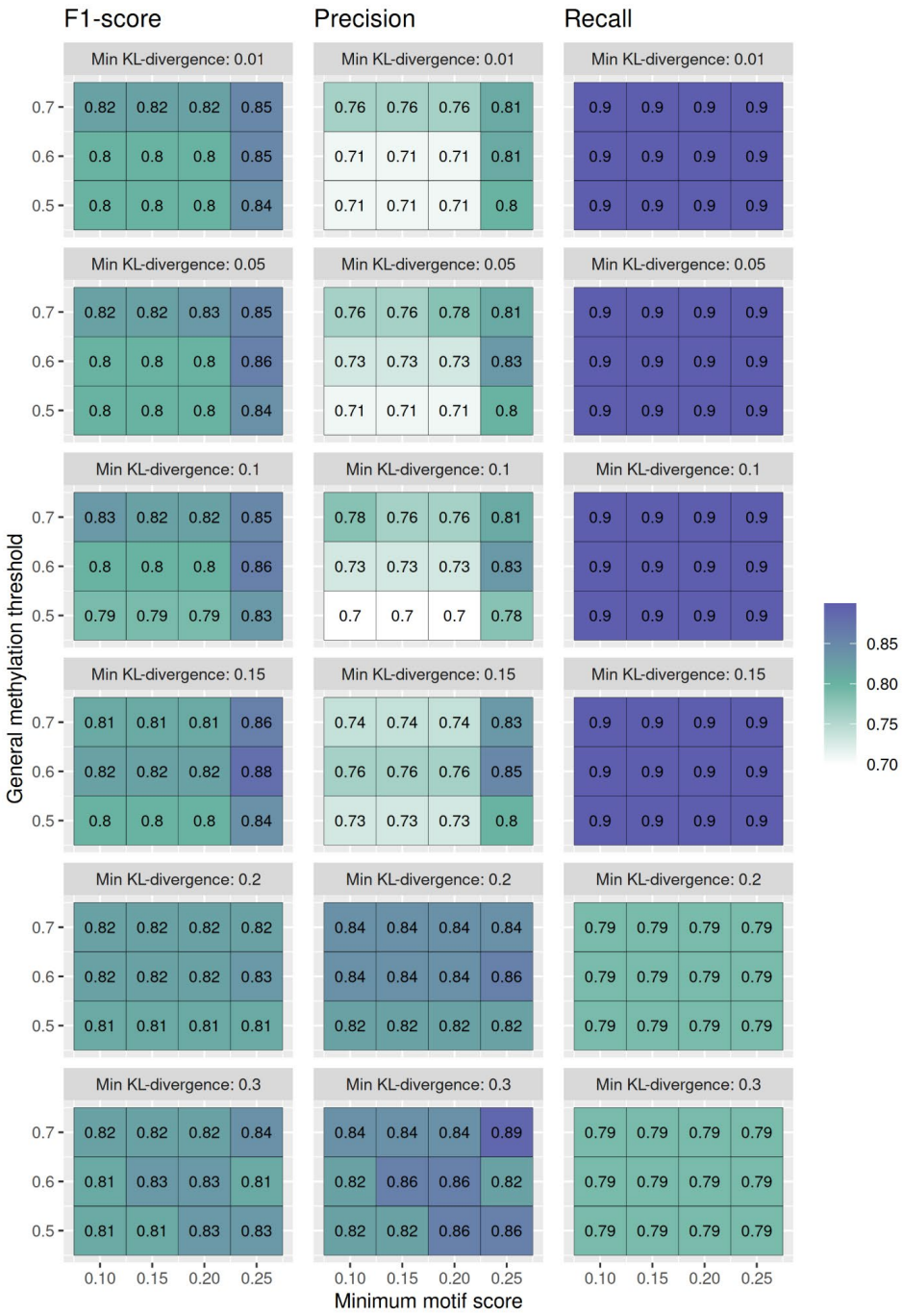


Fig. S10: Parameter sweep across three of Nanomotif motif identification most important parameters. The sweep was performed with expected motifs of the six training monocultures. Motif recall is mostly affected by the minimum KL-divergence parameters, whereas precision is reversely affected by the minimum KL-divergence. Precision is greatest at higher general methylation threshold and higher minimum motif score. Motif identification is generally stable across parameters. Parameters min_motif_score: 0.20, general_methylation_threshold: 0.7 and min_kl_divergence:0.5 was used. This was selected for higher sensitivity in low information references.

Supplementary Note 1

Direct motif identification in contigs

The assembly sequence and the methylation pileup from “modkit pileup” are used to identify methylated motifs.

Motifs are identified in each contig sequence separately from other contigs in an assembly. We use the “fraction modified” value in the modkit pileup output to determine if a position on the contig is methylated. “Fraction modified” corresponds to the number of mapped reads modified at the position divided by the number of valid bases at the position, which is the number of reads with the same canonical base as the respective modification type (A for 6mA and C for 5mC/4mC).

Each position with a valid base is classified as; NA, if the coverage below the threshold (default 5), not methylated, if the fraction of read methylation is below threshold (default 0.7) or methylated. We further define two ways in which a position can be methylated; generally methylated positions, where fraction of read methylation is above the threshold (default 0.7) and confidently methylated position, where fraction of read methylation is above the threshold (default 0.8).

Motif search is initiated at a seed motif (the default is the respective base to the evaluated methylation type, C for 5mC & 4mC and A for 6mA). To determine which position to expand from the motif, we extract sequences in a window around all confidently methylated positions (default window size is 41, 20 bases upstream and 20 bases downstream of the methylated position). These sequences are aligned with respect to the methylation, generating a methylated sequence set, S. A positional probability matrix (PPM) is then calculated from the set S.

$$p_{ij} = \frac{1}{S} \sum_{s \in S} \begin{cases} 1 \text{ if } s[i] = j \\ 0 \text{ if } s[i] \neq j \end{cases} \quad , j \in \{\mathrm{A,T,G,C}\}$$

This generates a 4x41 table, where the 41 columns correspond to the relative position with respect to the methylation and the 4 rows correspond to the nucleotide. Next, 10,000 sequences of the same window size are sampled with replacement from the contig, S_{sample}, and a positional nucleotide frequency table of the same dimensions is calculated. The 10,000 sampled windows serve only to generate a background positional frequency table. Resampling only reinforces the true nucleotide frequency of the background, hence resampling is non problematic.

$$q_{ij} = \frac{1}{S} \sum_{s \in S_{sample}} \begin{cases} 1 \text{ if } s[i] = j \\ 0 \text{ if } s[i] \neq j \end{cases} \quad , j \in \{\mathrm{A,T,G,C}\}$$

For each relative position, i, the KL-divergence is calculated from the four frequencies of the methylated sequence frequency table to the four frequencies of the sampled sequence frequency table.

$$D_{KL}^{(i)}(\mathbf{P} \parallel \mathbf{Q})) = \sum_{j \in \{\mathrm{A,\ T,\ G,\ C}\}} p_{ij} \log(\frac{p_{ij}}{q_{ij}})$$

This generates a vector of size 41, where each entry corresponds to a KL-divergence value. Positions are, per default, only considered for expansion if the KL-divergence is greater than 0.05. After selecting which position to expand, we select which bases to incorporate at each of these positions by two criteria; 1. the frequency of a base in the methylation sequence frequency table must be above 25% and the frequency of a base must be above the frequency in the sampled sequence frequency table. If more than one base at a position meets this criteria, we keep both of them and combinations of them a, e.g. accepting A and G at relative position 2 with seed NNANN would give rise to NNANA, NNANG and NNANR.

Each new motif candidate after the expansion is evaluated using a beta-Bernoulli model, treating each motif occurrence as a Bernoulli trial, being a success if it is a generally methylated position and a failure if not a methylated position. We use a Beta(α=0, β=1) as a prior, which means the posterior is also a Beta distribution with the parameters:

$$\alpha = \alpha_{prior} + n_{methylated}, \beta = \beta_{prior} + n_{non-methylated}$$

The posterior distribution is used to score each motif using the mean, standard deviation, and mean difference from the preceding motif. The mean represents the degree of motif methylation, a value expected to increase as the motif is refined. The standard deviation is used to penalize when few observations are present. Mean difference is expected to be high, when a desirable nucleotide addition is made, as it keeps the N highly methylated motif variants and disregards 4-N non-methylated motif variants, and is approximately zero for nucleotide insertion which contributes nothing to the recognition sequence.

$$score = mean_{diff} \cdot mean \cdot -log_{10}(standard\ deviation)$$

After scoring each of the new motifs, the highest scoring motif is stored. Next, one of the motifs is selected for propagation to the new set of motifs. The objective of the search is to converge on the motif candidate contributing the most positive methylation sites. The search heuristic is therefore formulated to minimize the proportion of generally methylated positions removed and maximize the proportion of non-methylated positions removed with respect to the seed motif. Concretely, the heuristic is calculated using the α and β parameters of the beta-Bernoulli posterior of the current motif and the seed motif, as they represent the number of methylated and non-methylated motif sites.

$$priority = \left(1 - \left(\frac{\alpha_{current}}{\alpha_{seed}}\right)\right) \cdot (\beta_{current}/\beta_{seed})$$

The motif with the lowest priority is then chosen for the next iteration. For the next iteration, the methylation sequences extracted initially are subsetted to those only containing the motif picked for expansion. After this the positional frequency table and KL-divergence is recalculated and the same procedure as before follows. The algorithm expands and scores following the steps described above, until the maximum score of a motif has not increased for 10 rounds or no more motif candidates are left to explore. The best scoring motif is then kept and saved to candidate motifs if its score is >0.2, otherwise dropped. The whole procedure is then repeated from the same seed, but removing sequences containing previously identified

candidate motifs from methylated sequences. This is continued until 25 candidate motifs with insufficient score have been dropped or only 1% of methylation sequences remain.

After all candidate motifs have been identified in a contig, they are subjected to a series of post-processing steps to improve final motifs. First, motifs which are a sub motif of other motifs are removed, which is the case if the sequence of any other motif is contained within the sequence of the current motif, e.g. C**5mC**WGG would give rise to removal of **6mA**CCWGG, as CCWGG is contained within ACCWGG. This step was added to mitigate false positive motifs resulting from 5mC methylations in close proximity to adenine can result in 6mA methylation calls, which subsequently produce a sufficiently strong signal to "detect" 6mA motifs. In this case we accept the possibility of removing similar motifs with different methylation types. Next we remove motifs which have isolated bases, defined as a non N position with at least 2 N's on both sides. Next we merge motifs whose sequences are similar, which can be the case for more generic motifs such as CCWGG, where CCAGG and CCTGG were found as separate motifs, but should constitute one motif. Motif merging is done by constructing a distance graph between all motifs, where motifs are only connected if the hamming distance is 2 or less. Motifs are then defined to be part of the same cluster in the graph if they are mutually reachable. All motifs within the same cluster are merged into a single motif, representing all motifs contained within the cluster. The merged motif is only accepted if the mean degree of methylation is not less than 0.2 of the mean methylation of the pre merge motifs, otherwise the premerge motifs are kept as is. Finally, motifs are queried for motif complements. If another motif is the complementary sequence of the motif, it gets removed and added as a complementary motif instead. Palindromic motifs are always considered as the complementary of itself.

Indirect motif detection

Direct motif identification is performed on one contig without any information from other contigs in an assembly. To detect potentially missed motifs in contigs, we perform what we term indirect detection of motifs in contigs, so called as they are only detected because the motif was directly detected with high confidence in another contig. To get indirectly identified motifs, we take the complete set of all motifs identified in all contigs and calculate α and β of the Beta posterior of the beta-Bernoulli model for all contigs. We report the α and beta parameters as the number of motif methylations and non-methylations, respectively.

Bin consensus

Bin consensus is evaluated by taking the complete set of motifs for a bin and checking if a motif meets a set of criteria. Firstly, a motif has to have been directly detected in at least one of the contigs in the bin. Next, we remove motifs that are not methylated in at least 75% of the contigs in the bin. We estimate this by counting the number of motif occurrences in contigs with a mean methylation of a motif above 25% and dividing by the total number of motif occurrences in the bin; if the fraction of motif occurrences present in methylated contigs is above 0.75, they are kept. Lastly, of the kept motifs, sub-motifs are removed as described in the post-processing step in the direct motif identification section. The remaining motifs are considered bin consensus motifs.

Supplementary Note 2

Annotation of gold standard proteins and their methylation motifs

We initially analyzed 11 prokaryotic strains known to encode gold-standard (GS) RM system proteins using Nanomotif (Fig. S9a, supplementary data 1 & 2). GS proteins are those whose biochemical functions have been experimentally validated with their exact coding DNA sequences identified[40]. The MTase-linker module of Nanomotif successfully annotated 43 out of 44 gold-standard MTase enzymes and linked the associated motifs, if active, across the 11 monocultural genomes. The single missing annotation is likely a false negative, as DefenseFinder had multiple HMM-profile hits for this gene, but ultimately assigned the corresponding gene as part of a non-methylation mediated defense system. Among the 44 GS MTases, 40 were found to be active. Notably, the motifs identified for all active GS MTases precisely match the specificities reported for the GS MTases in REBASE. For the non-GS epigenetic systems in these 11 prokaryotic strains, our gene annotations and motif assignments are generally aligned with existing REBASE entries. However, in *S. oneidensis*, *K.aquimarina,* and *D. acetoxidans*, the MTase-linker module annotated additional type II MTases beyond those previously reported in REBASE. While some may represent false positives, for example, *contig_2_3607* in *S. oneidensis*, others are supported by active motifs without any alternative gene assignments, for example *contig_1_1589* in *K. aquimarina* (supplementary data 2).

In total, 42 out of 71 detected motifs were confidently linked to annotated RM systems or orphan MTases. For 18 additional motifs, candidate genes with matching methylation features were identified. Notably, the remaining motifs include several from *Anabaena variabilis* that may represent variants of fewer distinct motifs, and *M.ruber,* representing noise motifs attributed to the increased false positive rate of ONT's methylation models in high GC contexts (Fig. S3). Apart from these unassigned motifs and two motifs in *D. acetoxidans*, all detected motifs were either supported by REBASE entries or corroborated by previous PacBio sequencing data[41], further validating the accuracy and reliability of the detected motifs.

Supplementary Note 3

Discovery of epigenetics systems in diverse metagenomes

We next analyzed a diverse set of prokaryotic communities using nanomotif (Fig. S9). Using the MTase-linker, 3123 MTase genes were detected across 549 HQ MAGs, resulting in a median of 6 MTases genes per genome and 3 RM systems encompassing an MTase per genome (Fig. S9a). Type II MTases were the most abundant (67%). This was followed by type I (16%), IIG (9%), and III (7%) (Fig. S9c). Type II MTases were also the most prevalent of all types not associated with an RM-system (Fig. S9d). Only 24% of type II MTases were co-located with a cognate restriction enzyme. Previous studies have also reported the frequent presence of orphan MTases in a wide range of bacterial genomes[41]. However, it is important to acknowledge that some associated REase genes may have gone undetected due to sequence divergence, especially in the complex samples with high novelty. In such cases, there may be an overestimation of orphan type II MTase genes. Despite this, many examples of orphan MTases are indeed genuine, and they represent a large group of MTases with non-RM functions. Similarly, 24% and 39% of Type I and Type III MTases, respectively, were unexpectedly identified without

775 corresponding restriction enzymes or Type IS subunits. In the soil sample, a significantly lower
776 number of MTase genes per genome compared to the other metagenomes was observed as
777 well. This discrepancy is likely due to the limited sensitivity of the HMM models in complex
778 samples.

779 For 76% of detected motifs, at least one candidate MTase with similar methylation character-
780 istics was found within the same genome (Fig. S9b). In 232 cases, a single motif could be
781 confidently linked to a specific MTase gene or RM system, resulting in a high-confidence set
782 of MTase target motif annotations. Notably, 65 of these motifs involved 5mC modifications,
783 which are notoriously difficult to detect with SMRT sequencing. This highlights the potential of
784 Nanomotif to accurately annotate MTase target recognition motifs in metagenomes, including
785 those with 5mC modifications.