Nanomotif: Identification and Exploitation of DNA Methylation Motifs in Metagenomes using Oxford Nanopore Sequencing

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 vital for understanding microbial biology, but 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 - a tool for identification of methylated motifs in metagenomic contigs. We demonstrate how this enables MAG contamination detection, association of mobile genetic elements, and linking of motifs with the responsible methyltransferase directly from Nanopore data.

Main

In all domains of life, genomes are subjected to epigenetic modifications, which directly influences gene expression, replication, and repair processes. 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 the motif and cleaves the DNA if it lacks the specific methylation. All DNA in the host must therefore have the correct methylation pattern for it to persist, including mobile genetic elements[2,3]. 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. 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[4], but currently has a low sensitivity for 5mC which requires a high sequencing coverage (250x)[5,6]. In 2023, ONT introduced all context methylation detection models making 5mC and 6mA methylation calls readily available with high sensitivity (https://github.com/nanoporetech/dorado). Despite this, only one effort has been made to utilize ONT methylation calls for methylation motif discovery in bacteria[7], but none which 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]. However, these methodologies suffer from the low PacBio sensitivity for 5mC[2,8] or require whole genome amplification for detection of motifs using ONT[7].

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, bin association of unbinned contigs, and linking of MTase genes 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 (>80%) positions. Motif candidates are then built iteratively by considering enriched bases around the methylated position. Afterwards, windows that constitute the specific motif are removed and the process repeated to identify additional motifs in the contig (supplementary note 1). Motifs *de novo* identified in the contig are referred to as 'direct detected'. Afterwards, all direct detected motifs are scored across all contigs to identify missed motifs and referred to as 'indirect detected'.

We benchmarked Nanomotif's motif finder on three monocultures by segmenting their genomes to a varying number of motif occurrences and coverages to simulate metagenomic conditions (Fig. 1c and Supplementary Fig. 1-3) and compared Nanomotif to MicrobeMod[9], the only other tool for performing motif discovery using ONT methylation calls[9]. Nanomotif achieved a high recall rate at low coverage and occurrences across all benchmarks, vastly outperforming MicrobeMod. Nanomotif detected G**6mA**TC with high sensitivity at genome coverage of 10x and motif occurrences of 10. Furthermore, Nanomotif maintained a high recall rate for more complex motifs such as GGC**6mA**$(N)_6$TGG at low coverage and motif occurrence. The *de novo* search algorithm can sporadically miss complex bipartite motifs like GGC**6mA**$(N)_6$TGG, but only one direct motif identification on a single contig is required for subsequent indirect detection of the motif across all contigs (Fig. 1d).

We applied the Nanomotif motif finder to identify putative methylated motifs in ten monocultures. A total of 25 unique motifs were identified with 19 highly methylated (>95%) in at least one species, which is consistent with previous observations[1]. Motifs observed with reduced degree of methylation, may result from involvement in regulatory functions[1]. All plasmids exhibited methylated motifs consistent with their corresponding genomes, highlighting methylation as a potential feature for plasmid host association - a difficult task with conventional metagenomic binning features (Fig. 1b). A unique feature of Nanomotif is that motifs can be identified in complex metagenomic samples. We therefore used Nanomotif on four increasingly complex metagenomic samples (Fig. 2e). In all metagenomic samples, except soil, the average number of motifs pr. metagenome-assembled genome (MAG) range between 1-2 and at least one motif was identified in >75% of high-quality (HQ) MAGs. In soil, at least one motif was identified in 35% of HQ-MAGs. This is in the same range as previous small-scale meta-epigenomic studies, which identified methylation motifs in approximately 50% of MAGs using PacBio[10,11].

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 the responsible methyltransferases.
Current MAG contamination evaluation tools rely on lineage-specific markers derived from genome databases[12–14], however, as the databases are far from complete, and exceptions exist even within closely related organisms, it is a difficult task. Using methylation patterns, contamination in MAGs can be directly detected as the methylation patterns must match across all contigs in a bin. Using the Nanomotif contamination detection module, we highlight two HQ MAGs from the anaerobic digester in Fig. 2a, which in both cases include contigs with inconsistent methylation patterns. In bin.1.257, contig 3819 (151 kbp) and 28180 (39 kbp), both completely lack G**Am6**TC methylation, despite the remaining bin being methylated at 46% of GATC positions. Another example is contig 77426 (69 kbp) in bin.1.84, which shares no methylated motifs with the bin. In a few cases, the methylation degree for a motif varies heavily within a bin. For example, in contig 75285 of bin.1.84, the methylation degree for TTCGA**Am6** deviated from the bin consensus, leading to its identification as putative contamination. The cause of such varying methylation degrees are not fully understood, but may be related to unknown biological factors rather than the contig actually being a contaminant. Overall hundreds of contigs were flagged as putative contamination across the complex metagenomic samples, including in HQ MAGs (Fig. 2e). In three cases, decontamination changed the MAG quality from MQ to HQ (Supplementary Fig. 4-7 and Supplementary data 3). 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 comparing the methylation pattern of unbinned contigs to bins in the sample. The contig must have a perfect unique match to a bin for it to be associated. Using Nanomotif contig inclusion module we highlight contigs 600, 609, and 1929, classified as two plasmids and a virus, which were assigned to bin.1.1 with a perfect and unique methylation profile match (Fig. 2d). The plasmids were likely missed in the binning as they have a 2-3x higher coverage compared to the chromosomal contigs of bin.1.1 (Fig. 2c). Associating mobile genetic elements with MAGs is of major importance as these can carry vital functionality[15]. For instance, geNomad identified three antimicrobial resistance genes (Supplementary data 4) in contig 600 that would have been missed using traditional binning features.

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. 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[16,17]. Therefore, we developed the Nanomotif MTase-linker module, which links methylation motifs to their corresponding MTase and, when present, their entire RM system (Supplementary data 1 & 2). We were able to confidently link 24 out of 31 detected motifs to an MTase in the monocultures (Fig. 1b). Of these, ten were associated with a complete RM system. In the metagenomic samples, nanomotif successfully linked MTase genes to 12-32% of identified motifs (Fig. 2e), and found that 57-72% these genes were part of a complete RM system. 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 metagenomes. Furthermore, we provide simple implementations that utilize these motifs for robust identification of putative contamination in MAGs, association of mobile genetic elements to hosts, and linkage of motifs to restriction-modification systems. As Nanopore sequencing becomes better at detecting modifications, the value of Nanomotif will increase further. Currently, more than 40 and 150 covalent modification types are known for DNA and RNA, respectively[5,18,19]. As the detection of these becomes reliable, they can readily be integrated into Nanomotif.

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

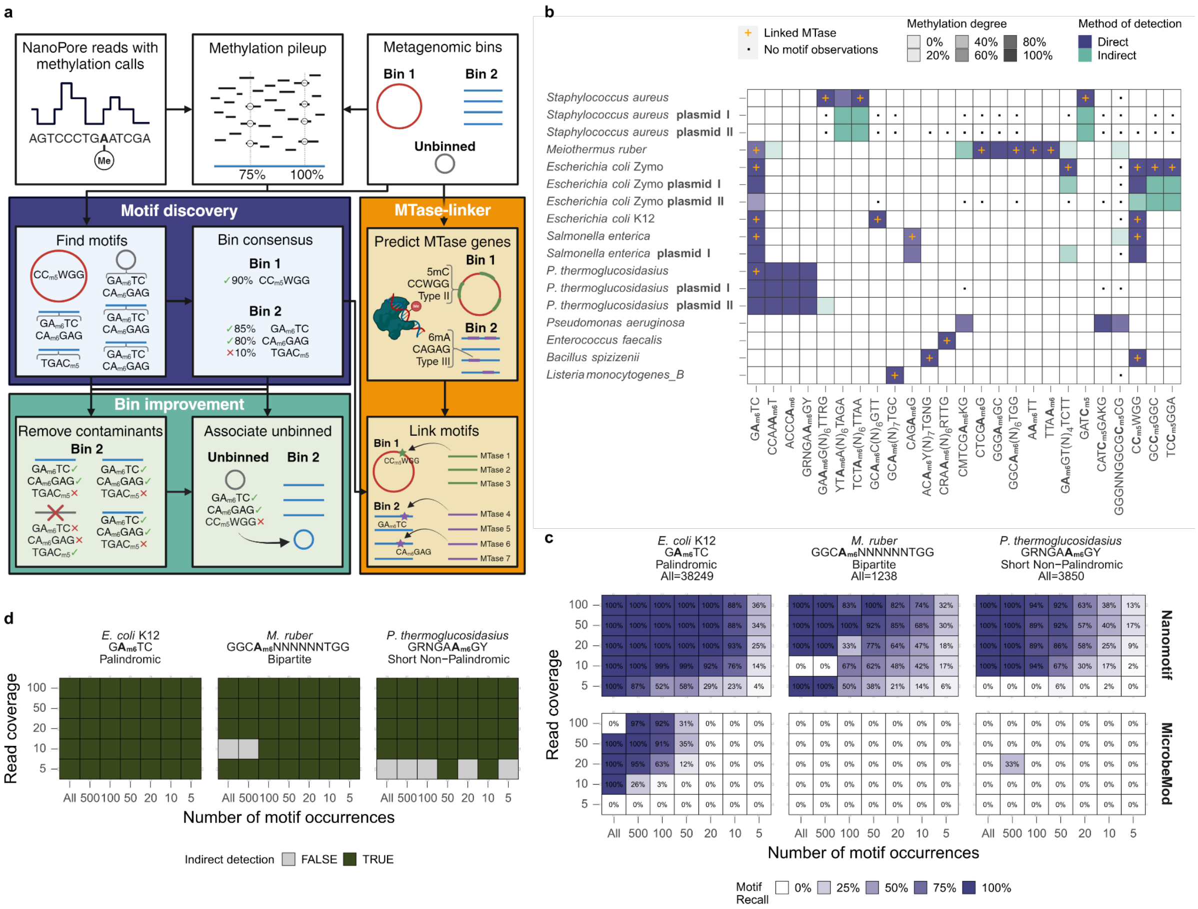
Nanomotif is available at https://github.com/MicrobialDarkMatter/nanomotif. Code for reproducing figures and supplementary resources can be found at https://github.com/SorenHeidelbach/nanomotif-article.

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).

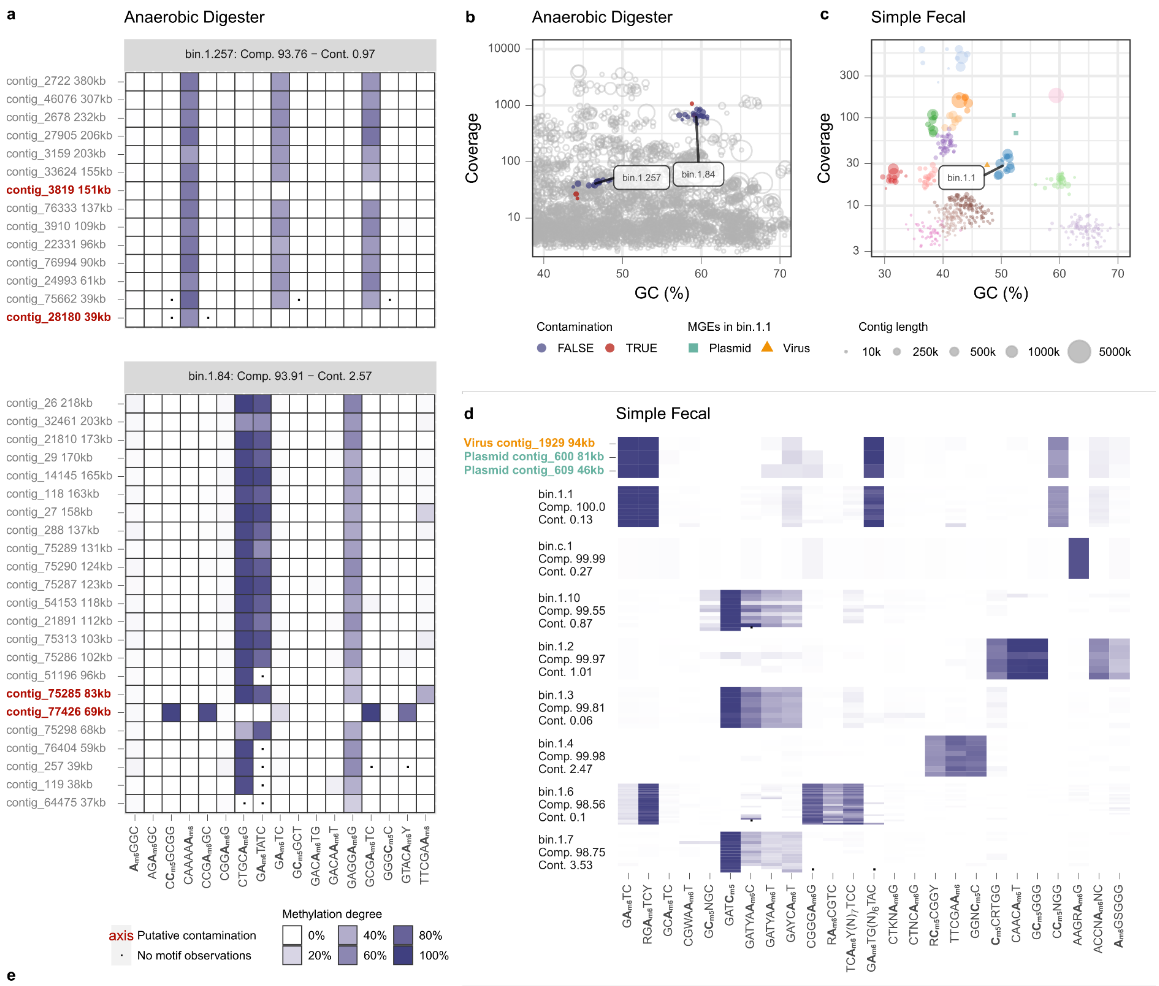
Ethics

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.



Read coverage	E. coli K12 All	500	100	50	20	10	5	M. ruber All	500	100	50	20	10	5	P. thermoglucosidasius All	500	100	50	20	10	5
Nanomotif 100	100%	100%	100%	100%	100%	88%	36%	100%	100%	83%	100%	82%	74%	32%	100%	100%	94%	92%	63%	38%	13%
50	100%	100%	100%	100%	100%	88%	34%	100%	100%	100%	92%	85%	68%	30%	100%	100%	89%	92%	57%	40%	17%
20	100%	100%	100%	100%	100%	93%	25%	100%	100%	33%	77%	64%	47%	18%	100%	100%	89%	86%	58%	25%	9%
10	100%	100%	99%	99%	92%	76%	14%	0%	0%	67%	62%	48%	42%	17%	100%	100%	94%	67%	30%	17%	2%
5	100%	87%	52%	58%	29%	23%	4%	100%	100%	50%	38%	21%	14%	6%	0%	0%	0%	6%	0%	2%	0%
MicrobeMod 100	0%	97%	92%	31%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
50	100%	100%	91%	35%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
20	100%	95%	63%	12%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	33%	0%	0%	0%	0%	0%
10	100%	26%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
5	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

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,** Heatmap of *de novo* identified motifs and their methylation degree in the monocultures. **c,** Benchmarking of a palindrome, bipartite, and short non-palindromic motif with Nanomotif and MicrobeMod[9]. The low motif recall of MicrobeMod, at high coverage and high motif occurrence settings, primarily stems from identification of similar motifs that are not identical to the benchmarking motif, e.g. SNG**Am6**TC instead of G**Am6**TC. **d,** For each condition in the top panel of c, green indicates when a motif was included for indirect detection, and therefore included in downstream processes.



e

	# Sequenced bases (Gbp)	# Contigs	Contig N50 (Kbp)	# MAGs (HQ)	# Motifs (HQ)	Motifs/MAG (HQ)	% MAGs w/ motif (HQ)	# MTases*	# MTases in RM−systems*	Linked motifs* (%motifs)	Contamination (#contigs (HQ))	Contigs included (#contigs (MGEs))
Monocultures	–	–	–	10 (10)	31 (31)	3.1 (3.1)	100% (100%)	52	21	24 (77%)	–	–
Simple fecal	4.2	2034	184.9	14 (8)	25 (19)	1.8 (2.4)	85% (100%)	32	9	6 (32%)	11 (0)	135 (27)
Complex fecal	49.1	11625	202.4	190 (93)	245 (192)	1.3 (2.1)	56% (75%)	591	247	51 (27%)	179 (20)	712 (518)
Anaerobic digestor	176.0	65077	98.4	423 (230)	667 (486)	1.6 (2.1)	60% (77%)	1391	583	57 (12%)	771 (121)	3498 (744)
Soil	206.6	374780	51.7	552 (66)	173 (26)	0.3 (0.4)	30% (35%)	105	26	8 (31%)	435 (7)	12472 (3399)

**in HQ−MAGs*

Fig. 2: Nanomotif MAG contamination detection and association of mobile genetic elements. a, Methylation profile of two HQ bins recovered from the Anaerobic digester sample. Contigs highlighted in red are putative contamination identified by Nanomotif. **b,** GC% and coverage of the anaerobic digester sample. **c,** GC% and coverage of the simple fecal sample. Contigs are colored according to the assigned bin. **d,** Methylation profile of the HQ bins in the simple fecal sample and highlighted plasmid & viral contigs. **e,** Sample stats from binning and the Nanomotif modules.

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. ZymoBIOMICS HMW DNA Standard D6322 was used for the remaining monoculture organisms.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). The complex fecal sample was collected at Aalborg University with consent from the provider. Sampling of the anaerobic digester sludge has been described elsewhere[20].

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 the simple fecal sample was extracted with the DNeasy PowerSoil Pro kit as described previously[21]. 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[20].

Sequencing

All samples were sequenced on the Promethion24 using the R10.4.1 nanopore. Libraries were prepared with SQK-LSK114 for the anaerobic digester and the complex fecal sample, whereas the other samples were prepared with the SQK-NBD114-24 ligation kit. Samples were basecalled with Dorado v0.3.2+d8660a3 using the dna_r10.4.1_e8.2_400bps_sup@v4.2.0 model and DNA methylation was called with the respective methylation models for 5mC and 6mA.

Assembly and binning

All samples were assembled and binned using the mmlong2-lite v1.0.2 pipeline available at https://github.com/Serka-M/mmlong2-lite. Briefly, flye[22] is used for assembly and polished using medaka (https://github.com/nanoporetech/medaka). Eukaryotic contigs are removed with tiara[23] before assembly coverage is calculated with minimap2[24]. Binning is performed as an ensemble using SemiBin[25], MetaBat2[26], and GraphMB[27], whereafter the best bin is chosen with DAS tool[28]. Recovered MAGs were evaluated with CheckM2[14].

Methylation pileup

Reads with methylation calls were mapped to the assembly using minimap2 v2.24[24] using default settings. Nanopore's modkit v0.2.4 (https://github.com/nanoporetech/modkit) was used to generate the methylation pileup from mapped reads using default settings.

Motif identification

Nanomotif was developed using python 3.9. Nanomotif motif discovery algorithm has three submodules, “find-motifs”, “score-motifs” and “bin-consensus”. “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. “score-motifs” takes the complete set of motifs and calculates a Beta-Bernoulli model for all motifs in all contigs. “bin-consensus” evaluates which motifs are considered highly methylated motifs within bins. All subcommands are gathered in a parent command “complete-workflow”, 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. For details about the algorithm see supplementary note 1.

Benchmark

Direct motif identification was benchmarked using motifs identified in the monocultures, whose validity was manually verified. Benchmarking was performed across two parameters; read coverage and number of motif occurrences. Lower coverage was achieved using rasusa[29] by subsetting the total length of reads to a multiple of the assembly length of the respective benchmarking organisms. Motif occurrences is the number of times a motif sequence occurs on the reference. For each benchmarking setup, the reference was split into chunks, containing exactly the number of motif occurrences being benchmarked; if the final chunk does not satisfy the number of motif occurrences, it is dropped from the benchmark. If the number of chunks, resulting from splitting the reference, exceeded 100, 100 chunks are randomly sampled and used for benchmarking. The methylation pileup is generated during the MicrobeMod execution. For a fair comparison, the same methylation pileup was also used for Nanomotif for direct motif identification. Then motif identification was performed with Nanomotif using the “find-motifs” command (version 0.1.19) and MicrobeMod using the “call_methylation” command (version 1.0.3). We calculate the recall rate for each benchmark condition as the number of chunks, where the motif was identified with the correct motif sequence, correct methylation position, and correct methylation type, divided by the number of benchmarking chunks. Benchmarking of indirect motifs identification was conducted on the Nanomotif output from the comparison above, where all chunks from the reference were treated collectively as a single bin. The motif was only reported as being identified if it was reported exactly as the benchmarking motif.

Benchmarks in supplementary figures S1-3 were performed on pileups generated as described in the “methylation pileup” section. This benchmark was performed on all chunks resulting from the splitting of the reference.

Contamination detection

Contamination is evaluated using “nanomotif detect_contamination” which defines a methylation pattern for each bin, and compares the methylation pattern of each contig in the bin against the bin consensus pattern. If a mismatch is observed between the contig and the bin consensus, the contig is reported as contamination.

Firstly, motifs not exceeding 25% mean methylation (--mean_methylation_cutoff) or observed less than 500 times in the bin consensus (--n_motif_bin_cutoff) are removed. The remaining motifs are used to create an extended bin consensus using the methylation detected for all contigs in a bin. Ambiguous motifs, defined as motifs where more than 40% of the mean methylation values (--ambiguous_motif_percentage_cutoff) in a bin are between 5% and 40%, are then removed. After removing ambiguous motifs, a motif is considered methylated in the bin if the mean bin methylation is at least 25% (--mean_methylation_cutoff). This creates a binary index for each motif as either methylated or not methylated. For methylated motifs in a bin, the standard deviation of the mean methylation values for each motif is calculated. To be included in the calculation of the standard deviation, the contig must have at least 10 motif occurrences and the motif must be at least 10 % methylated. Each motif is then scored for a given contig in a given bin. If the motif is methylated in the bin consensus, the motif in the contig is deemed methylated if the mean methylation is higher than the bin consensus mean methylation minus four standard deviations or if the contig mean value is above 40%. If the bin consensus mean minus four standard deviations is lower than 10% then the threshold is set to 10%. If the bin consensus is not methylated for a given motif, then the contig is deemed methylated if the mean methylation degree exceeds 25%. Given these criteria, a methylation mismatch score is calculated between the bin consensus and each contig. If one mismatch is found the contig is reported as contamination.

Include contigs

The “nanomotif include_contigs” scores all unbinned contigs and contigs reported as contamination similar to the “detect_contamination” module. Contigs are hereafter compared to each bin consensus pattern. If a perfect unique match with at least 5 comparisons (--min_motif_comparisons) is found between a contig and a bin, the contig is assigned to that bin. Only contigs and bins with at least one positive methylation are considered. Mobile genetic elements were identified using geNomad 1.7.4[29].

MTase-Motif-Linker

The Nanomotif MTase-linker module initially uses Prodigal[30] for protein-coding gene prediction (default settings) followed by DefenseFinder[31] 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 score is selected. The output file defense_finder_systems.tsv is used to determine whether the identified MTase is part of a complete RM system.

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[32], 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 against a custom database of methyltransferases with known target recognition motif from REbase[33]. We employ a threshold of 80% sequence identity and 80% query coverage to confidently predict the target recognition motif. Lastly, the 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.

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. Clark, T. A. *et al.* Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.* **40**, e29–e29 (2012).
5. Tue Kjærgaard Nielsen *et al.* Detection of nucleotide modifications in bacteria and bacteriophages; strengths and limitations of current technologies and software. *Mol. Ecol.* (2022) doi:10.1111/mec.16679.
6. Liu, J. *et al.* Bacmethy: A novel and convenient tool for investigating bacterial DNA methylation pattern and their transcriptional regulation effects. *iMeta* e186 (2024) doi:10.1002/imt2.186.
7. 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).
8. Wilbanks, E. G. *et al.* Metagenomic methylation patterns resolve bacterial genomes of unusual size and structural complexity. *ISME J.* **16**, 1921–1931 (2022).
9. 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*. http://biorxiv.org/lookup/doi/10.1101/2023.11.13.566931 (2023) doi:10.1101/2023.11.13.566931.
10. Hiraoka, S. *et al.* Diverse DNA modification in marine prokaryotic and viral communities. *Nucleic Acids Res.* **50**, 1531–1550 (2022).
11. Hiraoka, S. *et al.* Metaepigenomic analysis reveals the unexplored diversity of DNA

methylation in an environmental prokaryotic community. *Nat. Commun.* **10**, 159 (2019).

12. 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).
13. Orakov, A. *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 178 (2021).
14. 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).
15. 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).
16. 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).
17. Yasui, K. *et al.* Improvement of bacterial transformation efficiency using plasmid artificial modification. *Nucleic Acids Res.* **37**, e3 (2009).
18. Sood, A. J., Viner, C. & Hoffman, M. M. DNAmod: the DNA modification database. *J. Cheminformatics* **11**, 30 (2019).
19. Boccaletto, P. *et al.* MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Res.* **50**, D231–D235 (2022).
20. 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).
21. 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).
22. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
23. Karlicki, M., Antonowicz, S. & Karnkowska, A. Tiara: deep learning-based classification

system for eukaryotic sequences. *Bioinformatics* **38**, 344–350 (2022).

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

25. Pan, S., Zhu, C., Zhao, X.-M. & Coelho, L. P. A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat. Commun.* **13**, 2326 (2022).

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

27. Lamurias, A., Sereika, M., Albertsen, M., Hose, K. & Nielsen, T. D. Metagenomic binning with assembly graph embeddings. *Bioinformatics* **38**, 4481–4487 (2022).

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

29. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* 1–10 (2023) doi:10.1038/s41587-023-01953-y.

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

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

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

33. 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).

Supplementary Figures

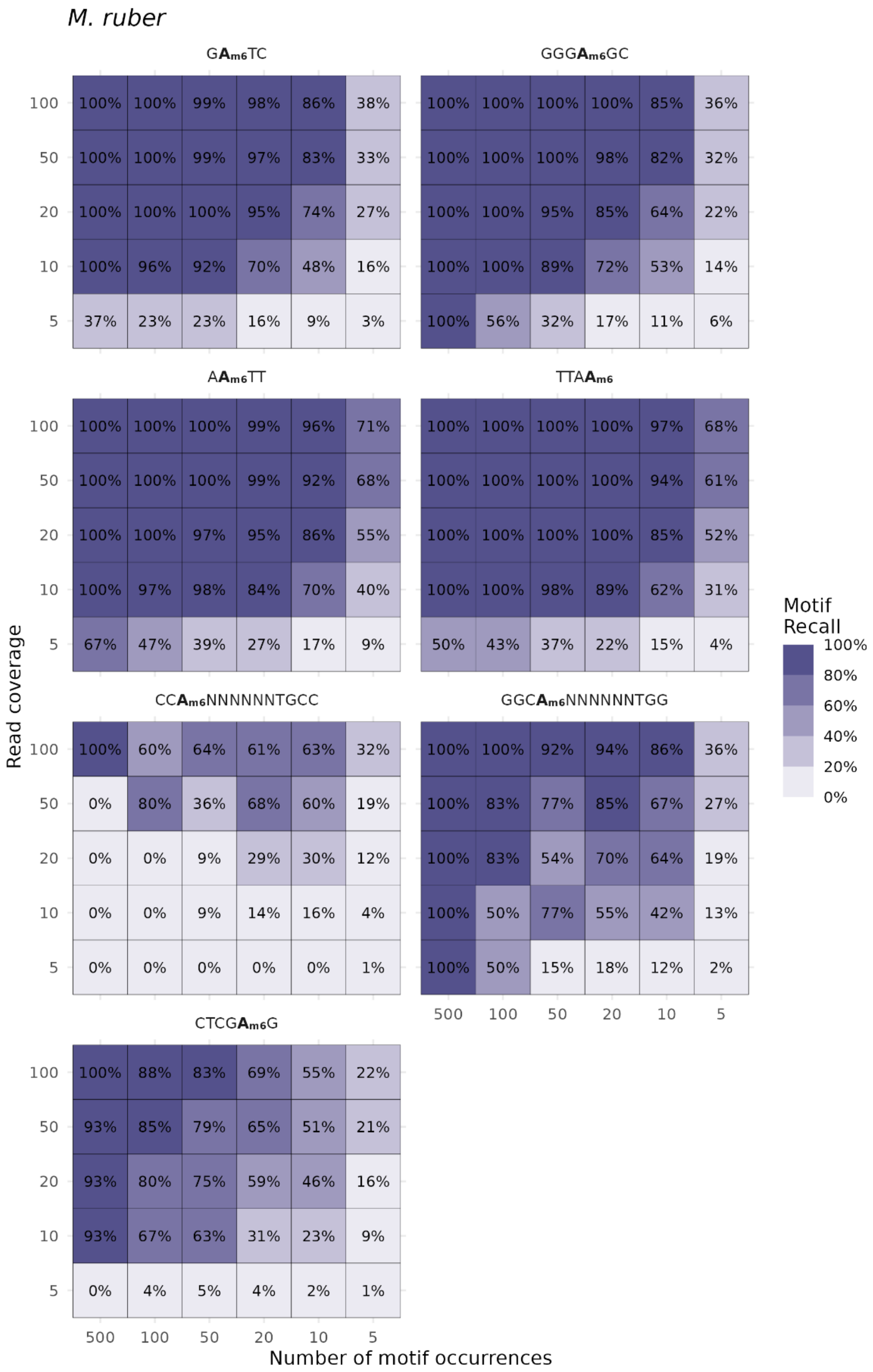


Fig. S1: Benchmarking of motif identification in *Meiothermus ruber* was conducted using Nanomotif for direct motif identification. In each benchmark, the reference sequence was divided into chunks, each containing the specified number of motif occurrences. For the purpose of recall calculation, 'true positives' are defined as the number of chunks in which the exact same motif as the benchmarking motif was identified.

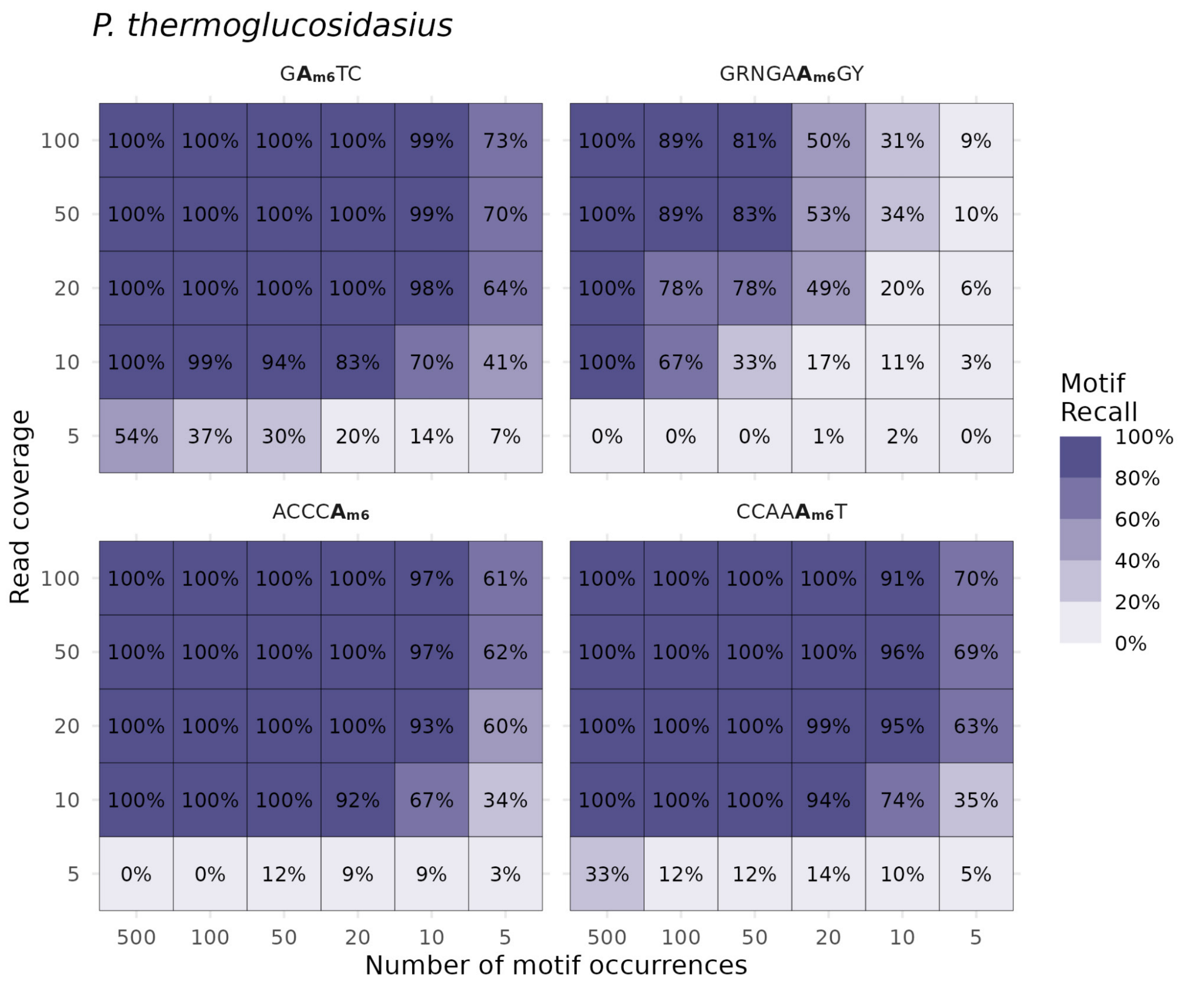


Fig. S2: Benchmarking of motif identification in *Parageobacillus thermoglucosidasius* conducted using Nanomotif for direct motif identification. In each benchmark, the reference sequence was divided into chunks, each containing the specified number of motif occurrences. For the purpose of recall calculation, 'true positives' are defined as the number of chunks in which the exact same motif as the benchmarking motif was identified.

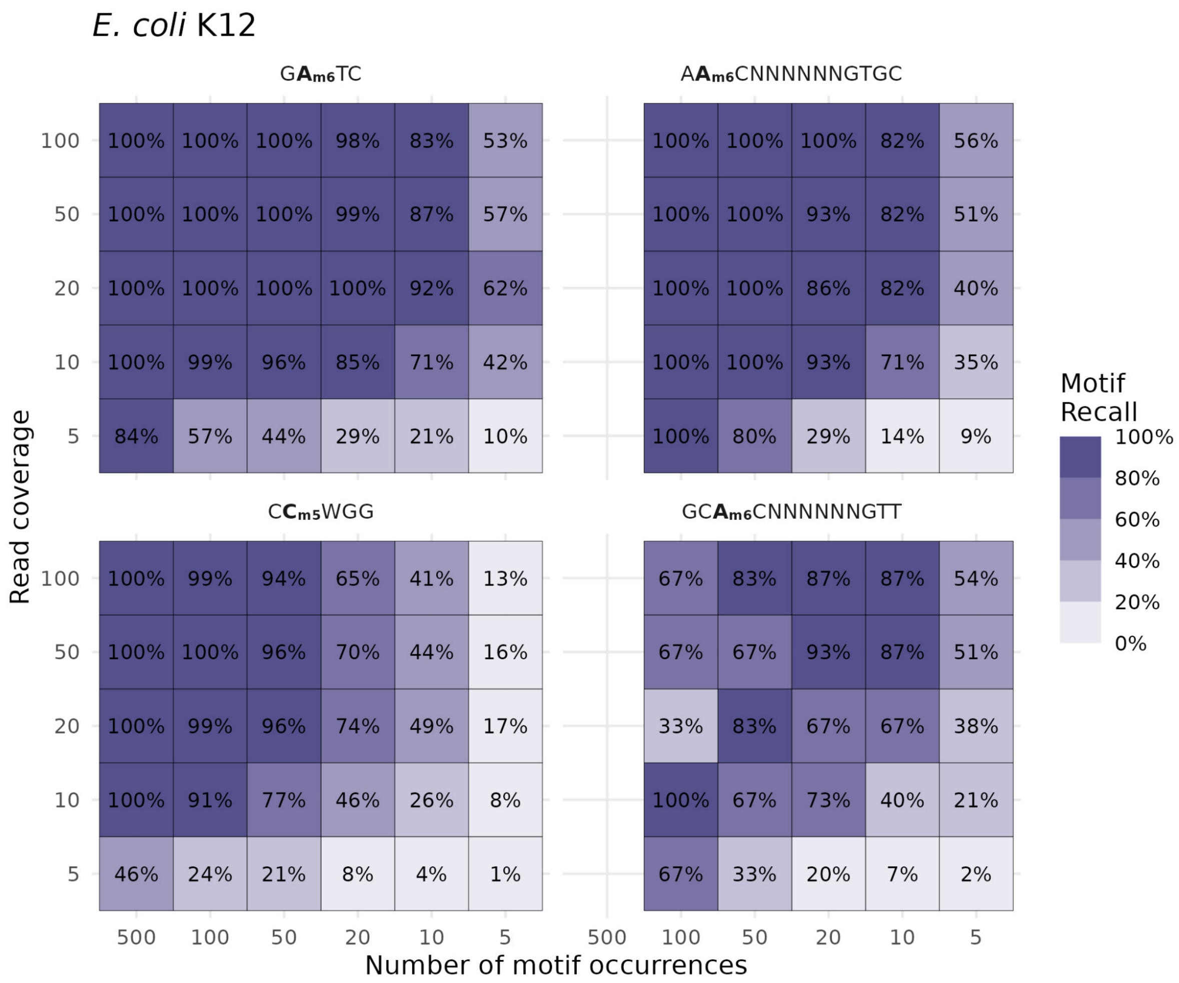


Fig. S3: Benchmarking of motif identification in *Escherichia coli* conducted using Nanomotif for direct motif identification. In each benchmark, the reference sequence was divided into chunks, each containing the specified number of motif occurrences. For the purpose of recall calculation, 'true positives' are defined as the number of chunks in which the exact same motif as the benchmarking motif was identified. The bipartite motifs were not benchmarked at 500 motif occurrences, as the complete genome of E. coli did not contain this many motif occurrences.

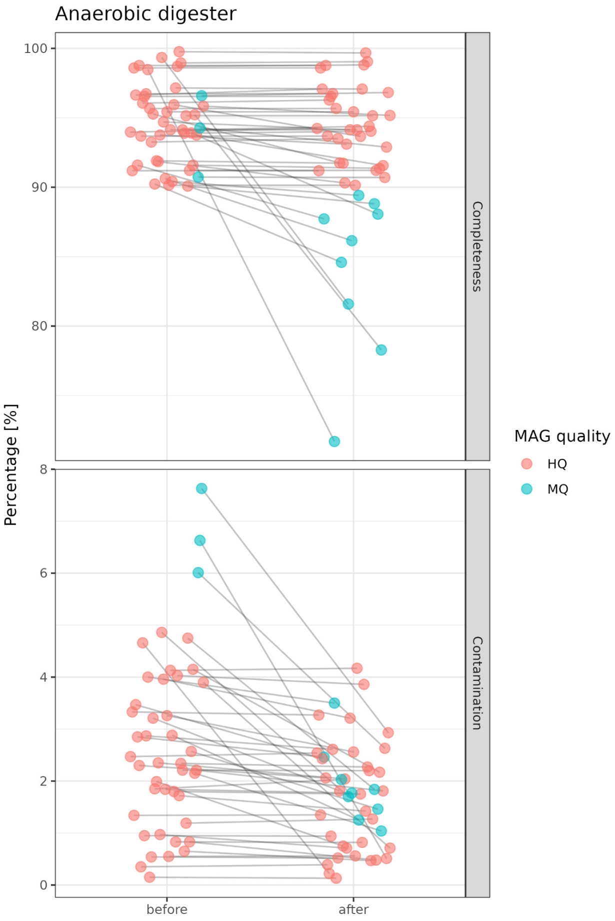


Fig. S4: Completeness and contamination before and after removal of putative contamination. Only MAGs that are either HQ before and/or after decontamination are shown.

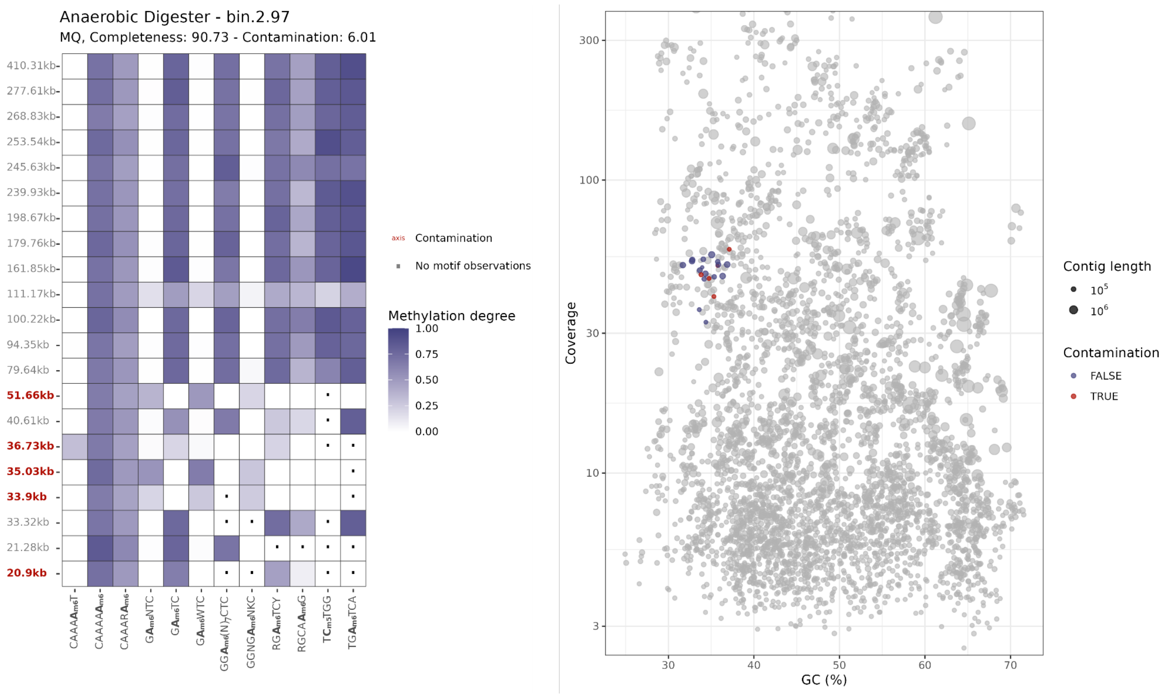


Fig. S5: Methylation pattern and GC% - coverage plot of bin.2.97 in the Anaerobic Digester. bin.2.97 is of medium quality before decontamination and high quality after.

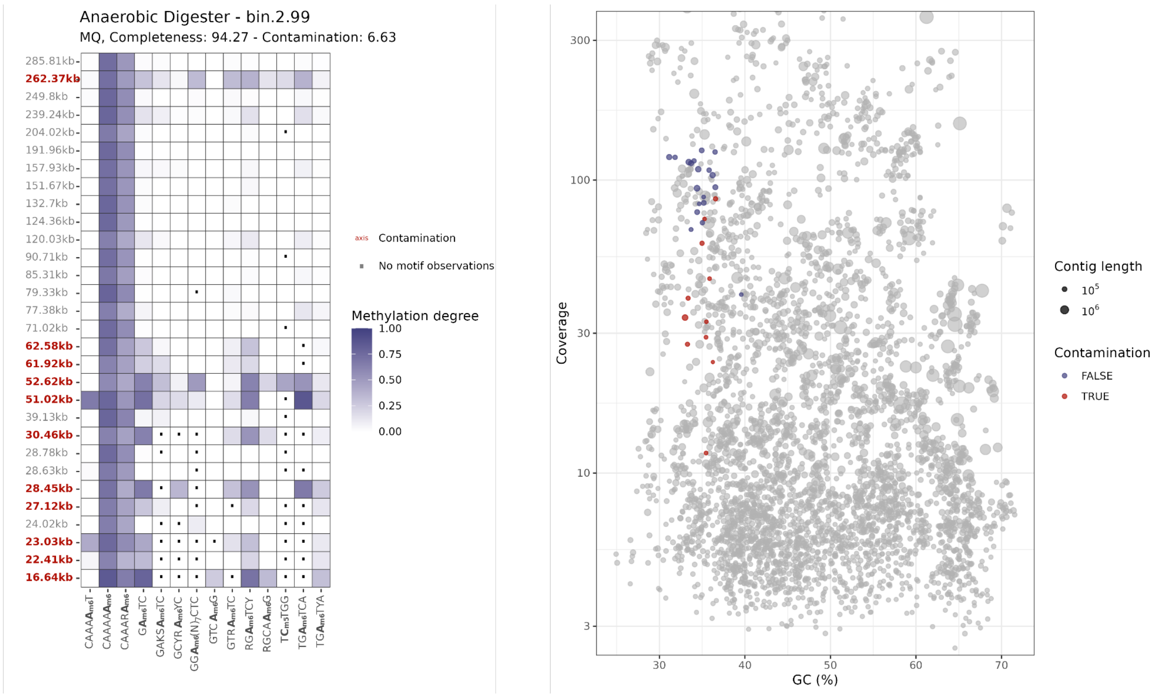


Fig. S6: Methylation pattern and GC% - coverage plot of bin.2.99 in the Anaerobic Digester. bin.2.99 is of medium quality before decontamination and high quality after.

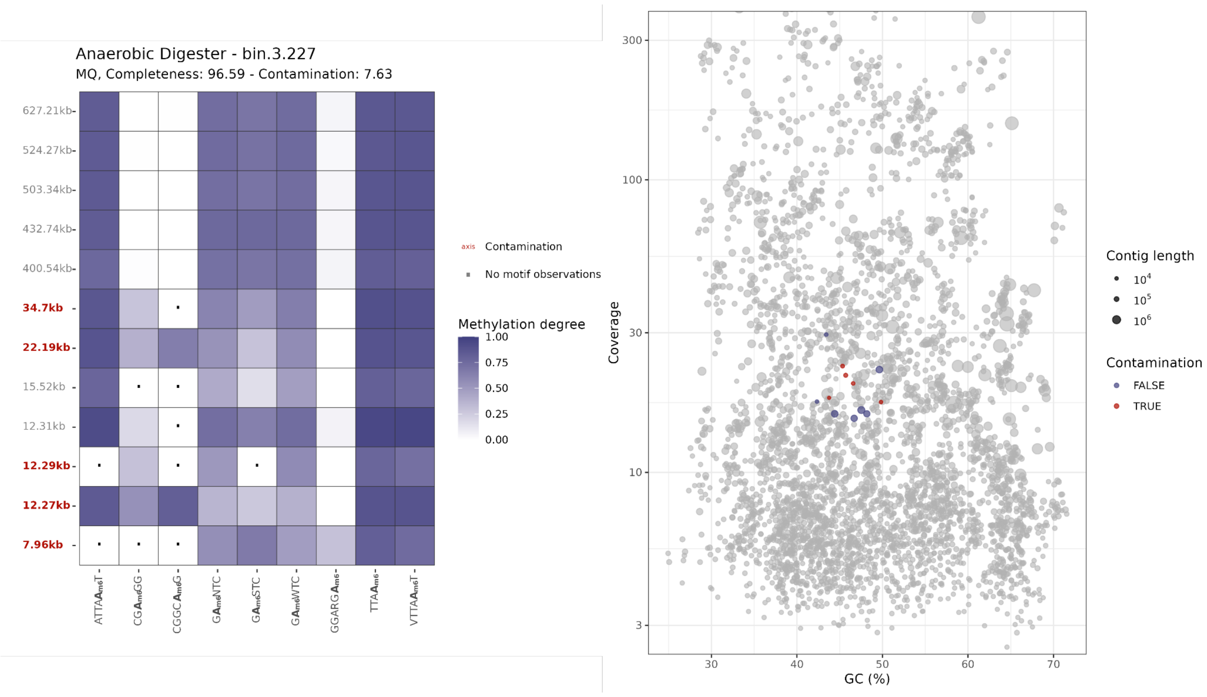


Fig. S7: Methylation pattern and GC% - coverage plot of bin.3.227 in the Anaerobic Digester. bin.3.227 is of medium quality before decontamination and high quality after.

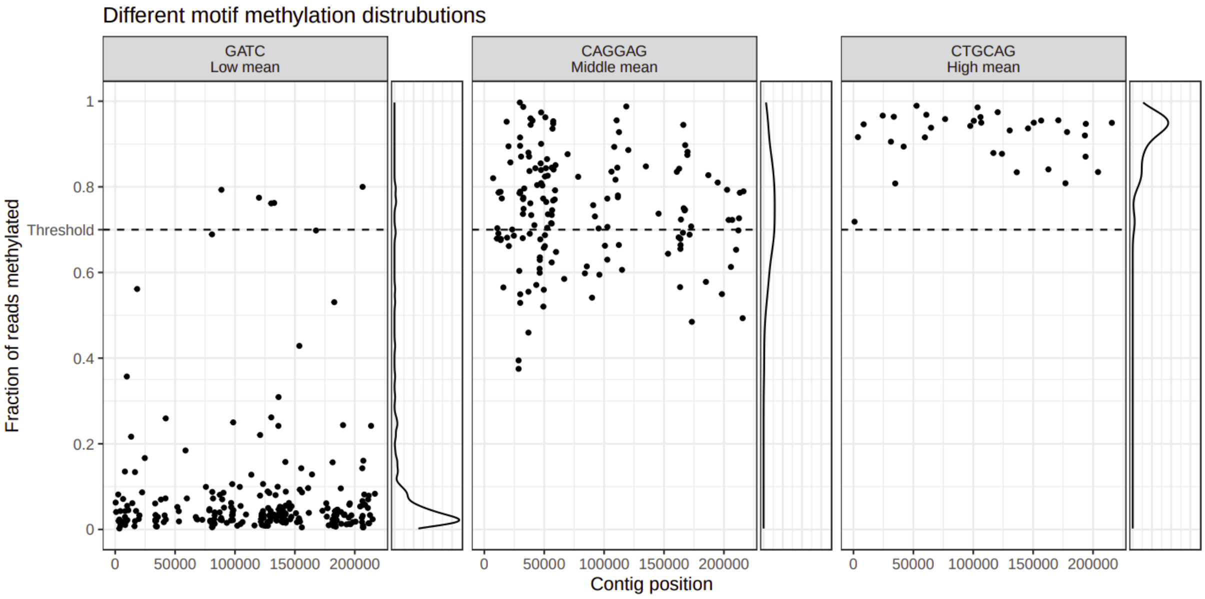
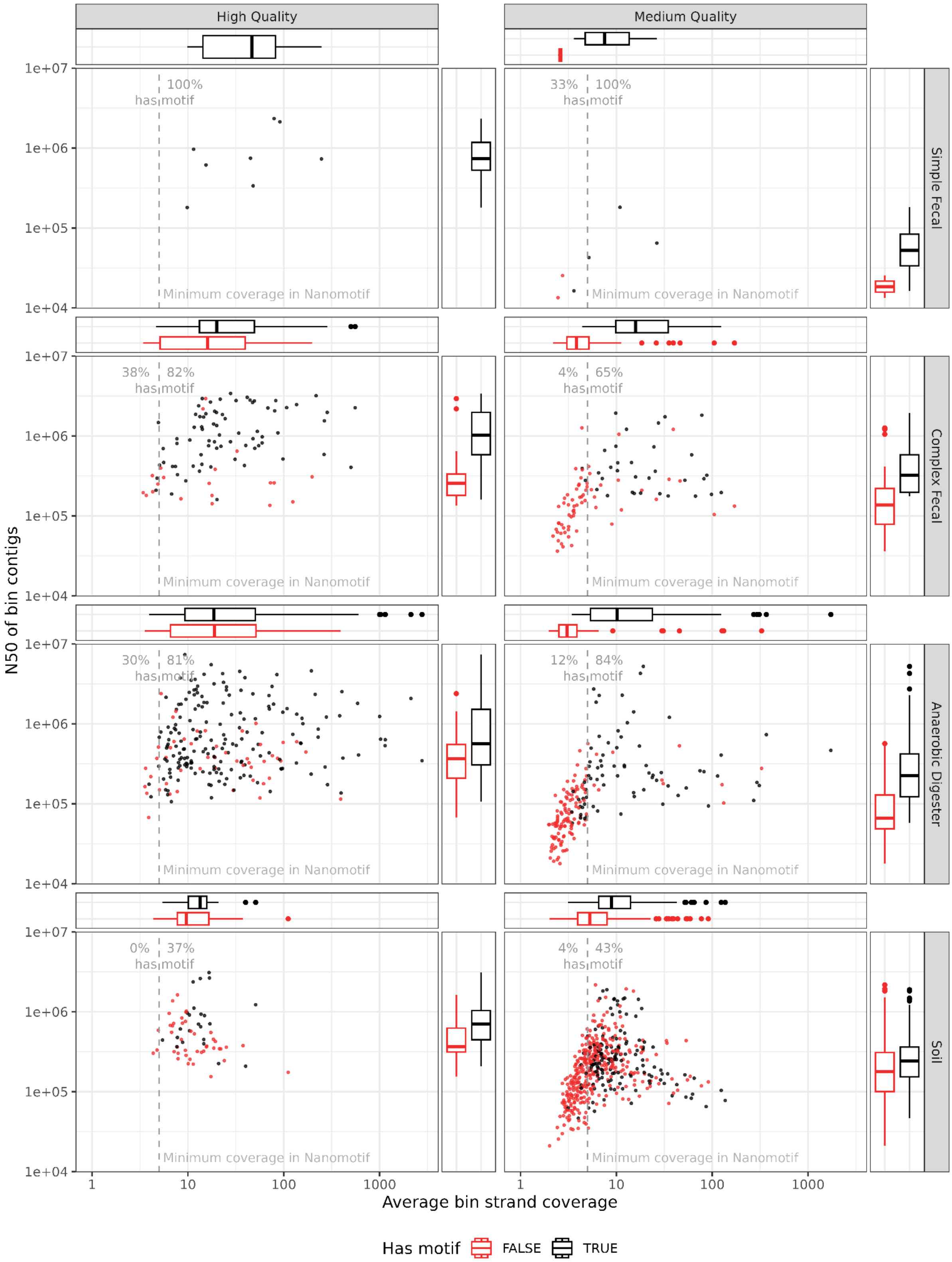


Fig. S8: Three different methylation degrees of motifs in contig 26 of the anaerobic digester sample bin.1.84. Each dot is an occurrence of the motif on the contig, with its respective fraction methylated reads. The threshold line indicates the general methylation threshold, above which a position is considered methylated and below which a position is considered non-methylated, when calculating mean methylation of the motif. The density of the fraction of reads methylated is shown to the right of the plotting frame. The CAGGAG motif, which has a density distribution around the threshold, gives rise to middle value means, despite all positions having a fraction of methylated reads >0.35.



451

452 **Fig. S9:** Motif identification in bins across N50 and coverage for metagenomic samples. Demonstration of drop off in percentage of bins with identified motif when the strand coverage drops below the threshold in nanomotif. The percentage of bins with at least one bin-consensus motif, has been indicated on both sides of the nanomotif coverage threshold.

456

Supplementary Note 1

Direct motif identification in contigs

The assembly sequence and the methylation pileup from modkit 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 (C for 5mC and A for 6mA). Firstly, positions with less than 5 valid bases at a position are removed. We then define two ways in which a position can be methylated; generally methylated positions, which is used when evaluating the degree of methylation of a motif and confidently methylated position, which is used for extracting sequences for the search algorithm. The fraction modified threshold for these are by default 0.70 and 0.80, respectively.

Motif search is initiated at a seed motif (the default is the respective base to the evaluated methylation type, C for 5mC and A for 6mA). To determine which position to expand 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 position and a positional nucleotide frequency table is calculated. 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 from the contig and a positional nucleotide frequency table of the same dimensions is calculated. For each relative position, 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. 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; the frequency of a base in the methylation sequence frequency table must be above 35% 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 it is not a generally methylated position. Positions filtered away from insufficient valid bases are not counted. 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 difference in mean from the preceding motif. The mean represents the degree of motif methylation, a value expected generally to tend towards 1 in fully methylated organisms. 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.

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

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 >0.1, 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 with 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

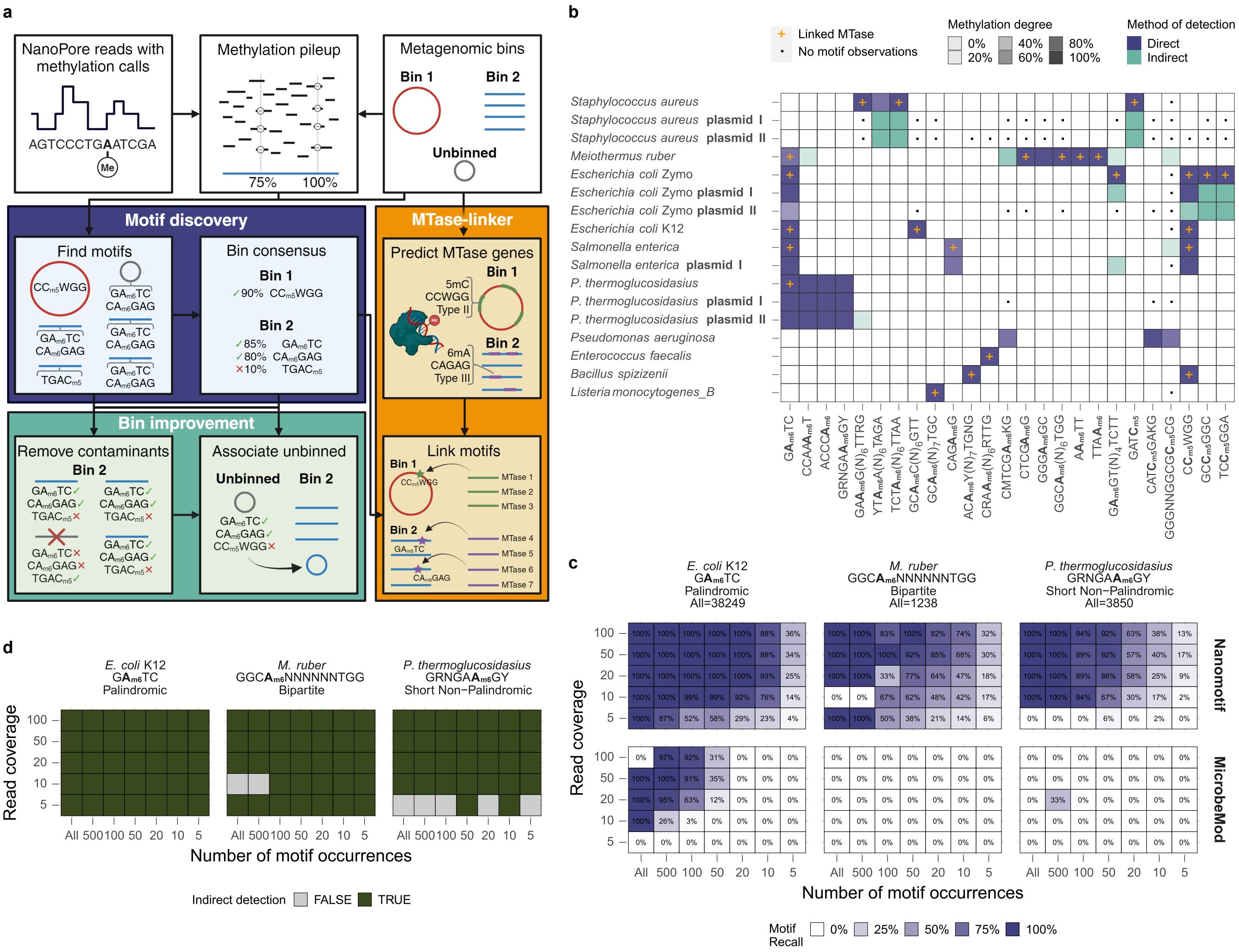
Direct motif identification is performed on one contig without any information from other contigs in an assembly. To detect potentially missed motifs in contigs, we perform what we term indirect detection of motifs in contigs, so called as they are only detected because the motif was directly detected 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.

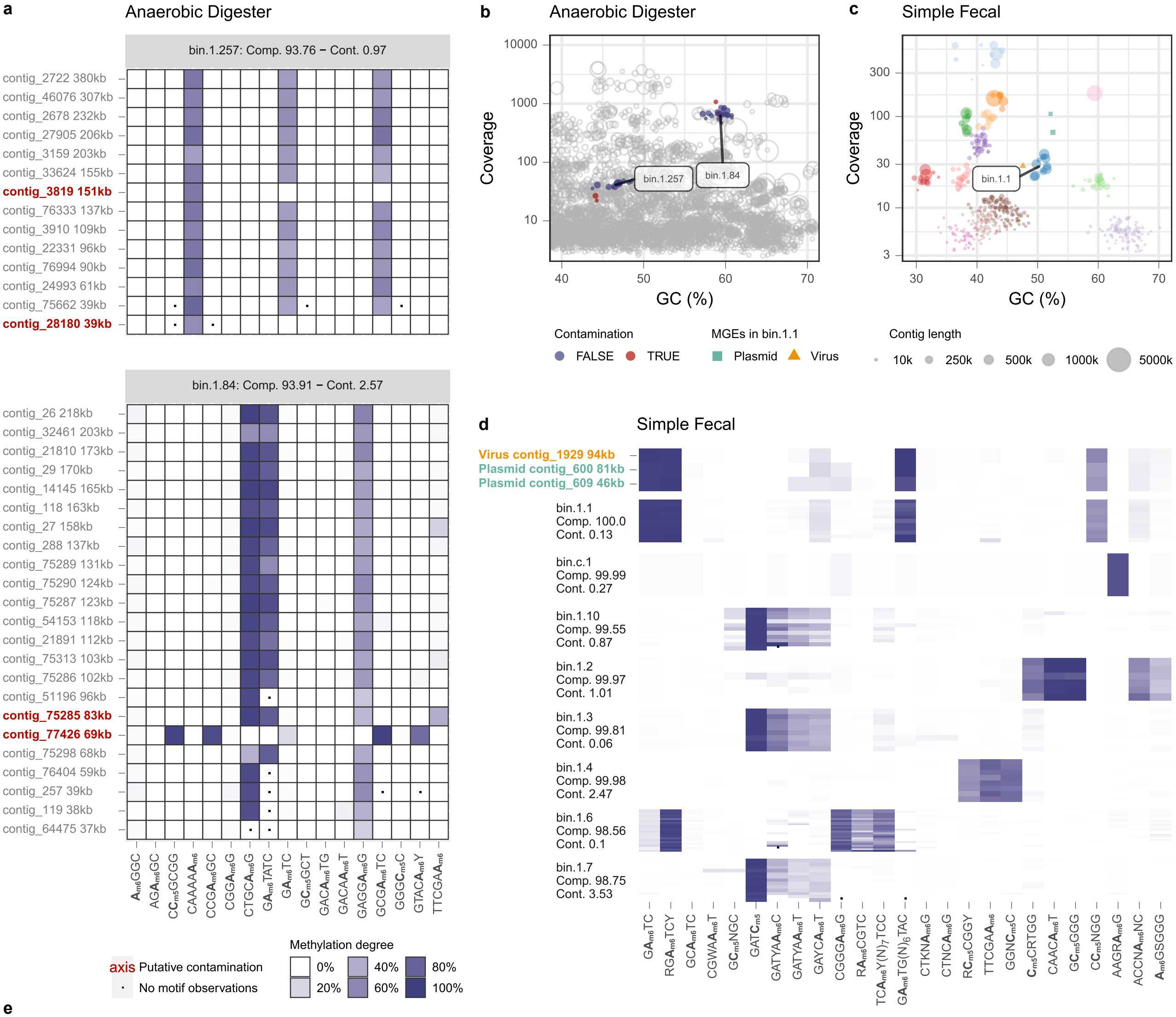
Bin consensus

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 Data

(See supplementary_data folder)


a
NanoPore reads with methylation calls
AGTCCCTGAATCGA
Me
Methylation pileup
75%
100%
Metagenomic bins
Bin 1
Bin 2
Unbinned
Motif discovery
Find motifs
CC$_{m5}$WGG
GA$_{m6}$TC
CA$_{m6}$GAG
TGAC$_{m5}$
Bin consensus
Bin 1
✓90% CC$_{m5}$WGG
Bin 2
✓85% GA$_{m6}$TC
✓80% CA$_{m6}$GAG
✗10% TGAC$_{m5}$
MTase-linker
Predict MTase genes
5mC CCWGG Type II
6mA CAGAG Type III
Me
Link motifs
MTase 1
MTase 2
MTase 3
MTase 4
MTase 5
MTase 6
MTase 7
Bin improvement
Remove contaminants
Associate unbinned
Unbinned
b
Linked MTase
No motif observations
Methylation degree
0%
20%
40%
60%
80%
100%
Method of detection
Direct
Indirect
Staphylococcus aureus
Staphylococcus aureus plasmid I
Staphylococcus aureus plasmid II
Meiothermus ruber
Escherichia coli Zymo
Escherichia coli Zymo plasmid I
Escherichia coli Zymo plasmid II
Escherichia coli K12
Salmonella enterica
Salmonella enterica plasmid I
P. thermoglucosidasius
P. thermoglucosidasius plasmid I
P. thermoglucosidasius plasmid II
Pseudomonas aeruginosa
Enterococcus faecalis
Bacillus spizizenii
Listeria monocytogenes_B
GA$_{m6}$TC
CCAAA$_{m6}$T
ACCCA$_{m6}$
GRNGAA$_{m6}$GY
GAA$_{m6}$G(N)$_6$TTRG
YTA$_{m6}$A(N)$_6$TAGA
TCTA$_{m6}$(N)$_6$TTAA
GCA$_{m6}$C(N)$_6$GTT
GCA$_{m6}$(N)$_7$TGC
CAGA$_{m6}$G
ACA$_{m6}$Y(N)$_7$TGNG
CRAA$_{m6}$(N)$_6$RTTG
CMTCGA$_{m6}$KG
CTCGA$_{m6}$G
GGGA$_{m6}$GC
GGCA$_{m6}$(N)$_6$TGG
AA$_{m6}$TT
TTAA$_{m6}$
GA$_{m6}$GT(N)$_4$TCTT
GATC$_{m5}$
CATC$_{m5}$GAKG
GGGNNGGCGC$_{m5}$CG
CC$_{m5}$WGG
GCC$_{m5}$GGC
TCC$_{m5}$GGA
c
E. coli K12
GA$_{m6}$TC
Palindromic
All=38249
M. ruber
GGCA$_{m6}$NNNNNNTGG
Bipartite
All=1238
P. thermoglucosidasius
GRNGAA$_{m6}$GY
Short Non−Palindromic
All=3850
Nanomotif
100 | 100% 100% 100% 100% 100% 88% 36% | 100% 100% 83% 100% 82% 74% 32% | 100% 100% 94% 92% 63% 38% 13%
50 | 100% 100% 100% 100% 100% 88% 34% | 100% 100% 100% 92% 85% 68% 30% | 100% 100% 89% 92% 57% 40% 17%
20 | 100% 100% 100% 100% 100% 93% 25% | 100% 100% 33% 77% 64% 47% 18% | 100% 100% 89% 86% 58% 25% 9%
10 | 100% 100% 99% 99% 92% 76% 14% | 0% 0% 67% 62% 48% 42% 17% | 100% 100% 94% 67% 30% 17% 2%
5 | 100% 87% 52% 58% 29% 23% 4% | 100% 100% 50% 38% 21% 14% 6% | 0% 0% 0% 6% 0% 2% 0%
MicrobeMod
100 | 0% 97% 92% 31% 0% 0% 0% | 0% 0% 0% 0% 0% 0% 0% | 0% 0% 0% 0% 0% 0% 0%
50 | 100% 100% 91% 35% 0% 0% 0% | 0% 0% 0% 0% 0% 0% 0% | 0% 0% 0% 0% 0% 0% 0%
20 | 100% 95% 63% 12% 0% 0% 0% | 0% 0% 0% 0% 0% 0% 0% | 0% 33% 0% 0% 0% 0% 0%
10 | 100% 26% 3% 0% 0% 0% 0% | 0% 0% 0% 0% 0% 0% 0% | 0% 0% 0% 0% 0% 0% 0%
5 | 0% 0% 0% 0% 0% 0% 0% | 0% 0% 0% 0% 0% 0% 0% | 0% 0% 0% 0% 0% 0% 0%
Read coverage
All 500 100 50 20 10 5
Number of motif occurrences
Motif Recall
0%
25%
50%
75%
100%
d
E. coli K12
GA$_{m6}$TC
Palindromic
M. ruber
GGCA$_{m6}$NNNNNNTGG
Bipartite
P. thermoglucosidasius
GRNGAA$_{m6}$GY
Short Non−Palindromic
Read coverage
100
50
20
10
5
All 500 100 50 20 10 5
Number of motif occurrences
Indirect detection
FALSE
TRUE



e

	# Sequenced bases (Gbp)	# Contigs	Contig N50 (Kbp)	# MAGs (HQ)	# Motifs (HQ)	Motifs/MAG (HQ)	% MAGs w/ motif (HQ)	# MTases*	# MTases in RM−systems*	Linked motifs* (%motifs)	Contamination (#contigs (HQ))	Contigs included (#contigs (MGEs))
Monocultures	–	–	–	10 (10)	31 (31)	3.1 (3.1)	100% (100%)	52	21	24 (77%)	–	–
Simple fecal	4.2	2034	184.9	14 (8)	25 (19)	1.8 (2.4)	85% (100%)	32	9	6 (32%)	11 (0)	135 (27)
Complex fecal	49.1	11625	202.4	190 (93)	245 (192)	1.3 (2.1)	56% (75%)	591	247	51 (27%)	179 (20)	712 (518)
Anaerobic digestor	176.0	65077	98.4	423 (230)	667 (486)	1.6 (2.1)	60% (77%)	1391	583	57 (12%)	771 (121)	3498 (744)
Soil	206.6	374780	51.7	552 (66)	173 (26)	0.3 (0.4)	30% (35%)	105	26	8 (31%)	435 (7)	12472 (3399)

**in HQ−MAGs*