

## RESEARCH ARTICLE

# Structural and functional diversity among Type III restriction-modification systems that confer host DNA protection via methylation of the N4 atom of cytosine

Iain A. Murray<sup>1\*</sup>, Yvette A. Luyten<sup>1</sup>, Alexey Fomenkov<sup>1</sup>, Nan Dai<sup>1</sup>, Ivan R. Corrêa, Jr.<sup>1</sup>, William G. Farmerie<sup>2</sup>, Tyson A. Clark<sup>3</sup>, Jonas Korlach<sup>3</sup>, Richard D. Morgan<sup>1</sup>, Richard J. Roberts<sup>1</sup>

**1** New England Biolabs, Ipswich, Massachusetts, United States of America, **2** Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, Florida, United States of America, **3** Pacific Biosciences Inc., Menlo Park, California, United States of America

\* [murray@neb.com](mailto:murray@neb.com)



## OPEN ACCESS

**Citation:** Murray IA, Luyten YA, Fomenkov A, Dai N, Corrêa IR, Jr., Farmerie WG, et al. (2021) Structural and functional diversity among Type III restriction-modification systems that confer host DNA protection via methylation of the N4 atom of cytosine. PLoS ONE 16(7): e0253267. <https://doi.org/10.1371/journal.pone.0253267>

**Editor:** Albert Jeltsch, Universität Stuttgart, GERMANY

**Received:** March 1, 2021

**Accepted:** June 1, 2021

**Published:** July 6, 2021

**Copyright:** © 2021 Murray et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the manuscript and its [Supporting information](#) files. SRA sequence read archives for four of the six strains described: *A. species H* (SRX10618618), *P. lemoignei* (SRX10565630 and SRX10565631), *E. coli* NCTC86 (SRX2568521) and *F. nodosum* Rt17-B1 (SRX1058899). Two strains were sequenced by others and not deposited so we cannot supply raw read archives.

## Abstract

We report a new subgroup of Type III Restriction-Modification systems that use m4C methylation for host protection. Recognition specificities for six such systems, each recognizing a novel motif, have been determined using single molecule real-time DNA sequencing. In contrast to all previously characterized Type III systems which modify adenine to m6A, protective methylation of the host genome in these new systems is achieved by the N4-methylation of a cytosine base in one strand of an asymmetric 4 to 6 base pair recognition motif. Type III systems are heterotrimeric enzyme complexes containing a single copy of an ATP-dependent restriction endonuclease-helicase (Res) and a dimeric DNA methyltransferase (Mod). The Type III Mods are beta-class amino-methyltransferases, examples of which form either N6-methyl adenine or N4-methyl cytosine in Type II RM systems. The Type III m4C Mod and Res proteins are diverged, suggesting ancient origin or that m4C modification has arisen from m6A MTases multiple times in diverged lineages. Two of the systems, from thermophilic organisms, required expression of both Mod and Res to efficiently methylate an *E. coli* host, unlike previous findings that Mod alone is proficient at modification, suggesting that the division of labor between protective methylation and restriction activities is atypical in these systems. Two of the characterized systems, and many homologous putative systems, appear to include a third protein; a conserved putative helicase/ATPase subunit of unknown function and located 5' of the *mod* gene. The function of this additional ATPase is not yet known, but close homologs co-localize with the typical Mod and Res genes in hundreds of putative Type III systems. Our findings demonstrate a rich diversity within Type III RM systems.

**Funding:** New England Biolabs provided funding in the form of salary to Iain A. Murray, Richard D. Morgan, Yvette A. Luyten, Alexey Fomenkov, Ivan R. Corrêa Jr, Nan Dai, and Richard J. Roberts, and in the form of supporting publication fees. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section. No additional external funding was received for this study.

**Competing interests:** Iain A. Murray, Richard D. Morgan, Yvette A. Luyten, Alexey Fomenkov, Ivan R. Corrêa Jr, Nan Dai, and Richard J. Roberts are full-time employees of New England Biolabs. There are no patents, products in development or marketed products to declare. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

## Introduction

The Type III restriction enzymes EcoPI and EcoP15I—originally isolated from coliphage P1 and *E. coli* plasmid P15, respectively—are among the earliest examples of RM systems to be identified and characterized [1, 2]. Type III systems are typically composed of two functionally distinct proteins, Mod and Res. The Mod subunit is responsible for specific DNA sequence recognition and host-protective DNA methylation. The Res subunit contains the endonuclease domain for DNA cleavage and an ATP hydrolysis domain related to SF2 helicases. Restriction requires both mod and res subunits, whereas methylation generally requires only the mod enzyme. In all previously known systems, the mod subunit confers host DNA protection via N6-methylation of a deoxyadenosine base located within just one strand of a non-palindromic target DNA sequence. The protein sequences of the EcoPI and EcoP15I Res subunits are almost identical and the corresponding Mod subunits are also closely related. However, the amino acid sequences encoding the specificity-defining target recognition domains (TRD) within each of the Mod subunits are divergent. M.EcoPI methylates the adenine base at the third position of 5' -AGACC-3' sequences and M.EcoP15I methylates the adenine base at the fifth position of 5' -CAGCAG-3' sequences [3, 4]. In the absence of cognate methylation, substrate DNAs are restricted by a Mod-Res complex in a reaction that requires ATP hydrolysis and the presence of two or more recognition sites arranged in an inverse orientation—either head-to-head or tail-to-tail [5–8]. Initial models of Type III enzymes—invoking a heterotetrameric structure comprised of two Mod and two Res subunits per active complex [9]—were subsequently revised in favor of a heterotrimeric Mod-2/Res-1 architecture [10, 11]. Structural analyses of EcoP15I—using small-angle X-ray scattering [12] and X-ray crystallography [13]—are also consistent with a single Res subunit associated with a Mod homodimer. In the crystal structure one subunit of the Mod dimer appears to be primarily responsible for DNA binding while the second catalyzes methylation of the adenine at the 5th position in the EcoP15I target sequence. Res-catalyzed phosphodiester bond hydrolysis of unmethylated substrates typically yield products with two base-pair (bp) 5'-extensions at a position 25 and 27 base-pairs 3' of the Mod motif [3, 4]. Several additional Type III RM systems have been characterized, including HinfIII which methylates the second A in the sequence 5' -CGAAT-3', PstII which methylates the A at the fourth position in the sequence 5' -CTGATG-3', and BceSI that methylates the A at the fourth position of the sequence 5' -CGAAG-3' [14–16]. However, in contrast to the Type II RM systems that provided the critical tools for the development of recombinant DNA technologies, relatively few Type III RM systems have been characterized in detail. This reflects the difficulty of characterizing cleavage sites by the Res component, which often gives incomplete cleavage. In addition, the analyses typically required to define the sequence motifs recognized by Mod subunits—and the identity of the specific A that was methylated—were technically challenging. However, with the introduction of Single Molecule Real Time (SMRT) sequencing technology it is now possible to identify N6-methyldeoxyadenosine (m6A) and N4-methyldeoxycytosine (m4C) modifications on a genome-wide scale thereby providing a direct readout of Mod specificities [17–19].

Based on SMRT sequencing observations of asymmetric, single strand m4C modification motifs in organisms that harbor putative Type III RM systems, we characterized the recognition specificity for six Type III systems that use m4C methylation for host protection and confirmed endonuclease cleavage typical of Type III enzymes for one of these systems. Additionally, we identify putative recognition motifs for 24 additional Type III m4C systems from methylome analyses of SMRT sequenced genomes and identify hundreds of Type III systems highly similar to the characterized MTases that likely also depend upon m4C modification. These findings demonstrate Type III RM systems can employ either m6A or m4C for

host protection, allowing even greater diversity of their recognition target as they evolve to defend against phage predation, along with an expanded potential for phase-variable transcription regulation.

## Materials and methods

### Materials

Restriction endonucleases, T4-DNA ligase, Phusion-HF and Q5 DNA polymerases, Gibson and HiFi assembly reagents, S-adenosylmethionine, Nucleoside Digestion Mix and competent *E. coli* cells were from New England Biolabs (Ipswich, MA, USA). Plasmid DNAs and PCR products were purified using spin-column reagents from New England Biolabs. DNA oligonucleotides (for use in PCR or DNA sequencing applications) and a codon-optimized synthetic gene were purchased from Integrated DNA Technologies (Coralville, IA, USA). Genomic DNAs of *D. thermolithotrophum* and *F. nodosum* Rt17-B1 were purchased from the DSMZ-German Collection of Microorganisms and Cell Cultures GmbH.

**Identification of candidate Type III systems.** A group of putative Type III RM systems that use m4C modification for host DNA protection were identified using Pacific Biosciences single-molecule real-time (SMRT) sequencing of bacterial genomic DNAs to identify asymmetric motifs having m4C methylation in one DNA strand [17]. SMRT sequencing was performed on the RSII platform and m4C methylation was characterized using the "RS\_Modification\_and\_Motif\_Analysis.1" program from the SMRT Analysis 2.3.0 software package. Raw sequence data is available from the SRA sequence read archives for four of the six strains described: *A. species* H (SRX10618618), *P. lemoignei* (SRX10565630 and SRX10565631), *E. coli* NCTC86 (SRX2568521) and *F. nodosum* Rt17-B1 (SRX1058899). Putative Type III RM systems were predicted using the SeqWare software in REBASE, based on sequence similarity between the predicted protein sequences and known Type III Mod and Res proteins.

**Isolation of genomic DNAs.** *A. species* H, *C. inhibens* subspecies *gilichinskyi* [20] and *P. lemoignei* genomic DNA samples were isolated using phenol-chloroform extraction followed by isopropanol precipitation. Genomic DNA of *E. coli* NCTC86 was a gift from Dr Swaine L. Chen.

**Cloning and recombinant expression of *mod* genes.** Candidate *mod* genes from each of the six bacterial strains were amplified from their respective genomic DNAs using Q5 high-fidelity DNA polymerase—and introduced into the high-copy pRRS plasmid vector using the HiFi *in vitro* DNA assembly method (NEB). DNA sequencing of the cloned genes and flanking vector sequences was used to confirm production of the *bona fide* recombinant plasmids. In the case of the *P. lemoignei mod* gene a non-canonical TTG initiation codon in the native strain was replaced with ATG for expression in *E. coli*. Cloned *mod* genes were expressed using either *E. coli* ER2683 (K strain, *dam*+/*dcm*+), *E. coli* C2523 (B strain, *dam*+/*dcm*-) or *E. coli* ER2796, a K strain which lacks all host-encoded methyltransferase activities [21].

**Confirmation of methylation status of DNAs isolated from *E. coli* cells expressing Type III *mod* genes.** Pacific Biosciences SMRT sequencing of plasmid and/or genomic DNA samples was used to confirm sites of incorporation of m4C modifications catalyzed by each of the Mod enzymes.

**Quantitative analyses of *Mod* activities using an LC-MS assay.** Relative abundances of unmethylated (dC) and methylated (m4C) were determined using liquid chromatography and mass spectrometry (LC-MS). In this case DNA samples were isolated from a *dcm*- *E. coli* host strain to eliminate m5C contamination which would otherwise confound the analyses. DNA samples were hydrolyzed to nucleosides using the Nucleoside Digestion Mix (NEB). LC-MS/

MS analysis was performed in duplicate by injecting digested polynucleotide samples on an Agilent 1290 UHPLC equipped with a G4212A diode array detector and a 6490A Triple Quadrupole Mass Detector operating in the positive electrospray ionization mode. UHPLC was carried out using a Waters XSelect HSS T3 XP column (2.1 × 100 mm, 2.5 μm) with the gradient mobile phase consisting of methanol and 10 mM aqueous ammonium formate (pH 4.4). Data acquisition was performed in the dynamic multiple reaction monitoring (DMRM) mode. Each nucleoside was identified in the extracted chromatogram associated with its specific MS/MS transition: dC [M+H]<sup>+</sup> at m/z 228 → 112 and m4C [M+H]<sup>+</sup> at m/z 242 → 126. External calibration curves with known amounts of the nucleosides were used to calculate their ratios within the samples analyzed.

**Restriction analyses.** Two grams of *Escherichia coli* NCTC86 cells were suspended in 20 ml sonication buffer (20 mM Tris pH8.0, 50 mM NaCl, 0.1 mM EDTA), sonicated to create a cell extract and clarified by centrifugation at 30,000g for 15 minutes at 4°C. The extract was applied to a 5 ml Heparin HiTrap HP column equilibrated in sonication buffer and 3 ml fractions were collected during a 100 ml gradient from 50 mM to 1M NaCl in the same buffer. Fractions were assayed for endonuclease activity in 1X CutSmart buffer (NEB) supplemented with 1 mM ATP. A fraction containing Eco86II was identified and used to map cut sites on pUC19 and pBR322 plasmids. The Eco86II fraction was incubated with the closed circular plasmid DNA for 30 minutes, the reaction mixture was divided and one of several REases that cut each plasmid once at different positions around the plasmid was added and the reaction continued for 30 minutes more. DNA fragments were analyzed by electrophoresis on 1% agarose gels.

**Sequence alignments.** Protein sequence alignments of Type III restriction enzymes described in this study were constructed using PROMALS3D: a tool for multiple protein sequence and structure alignments [22].

## Results

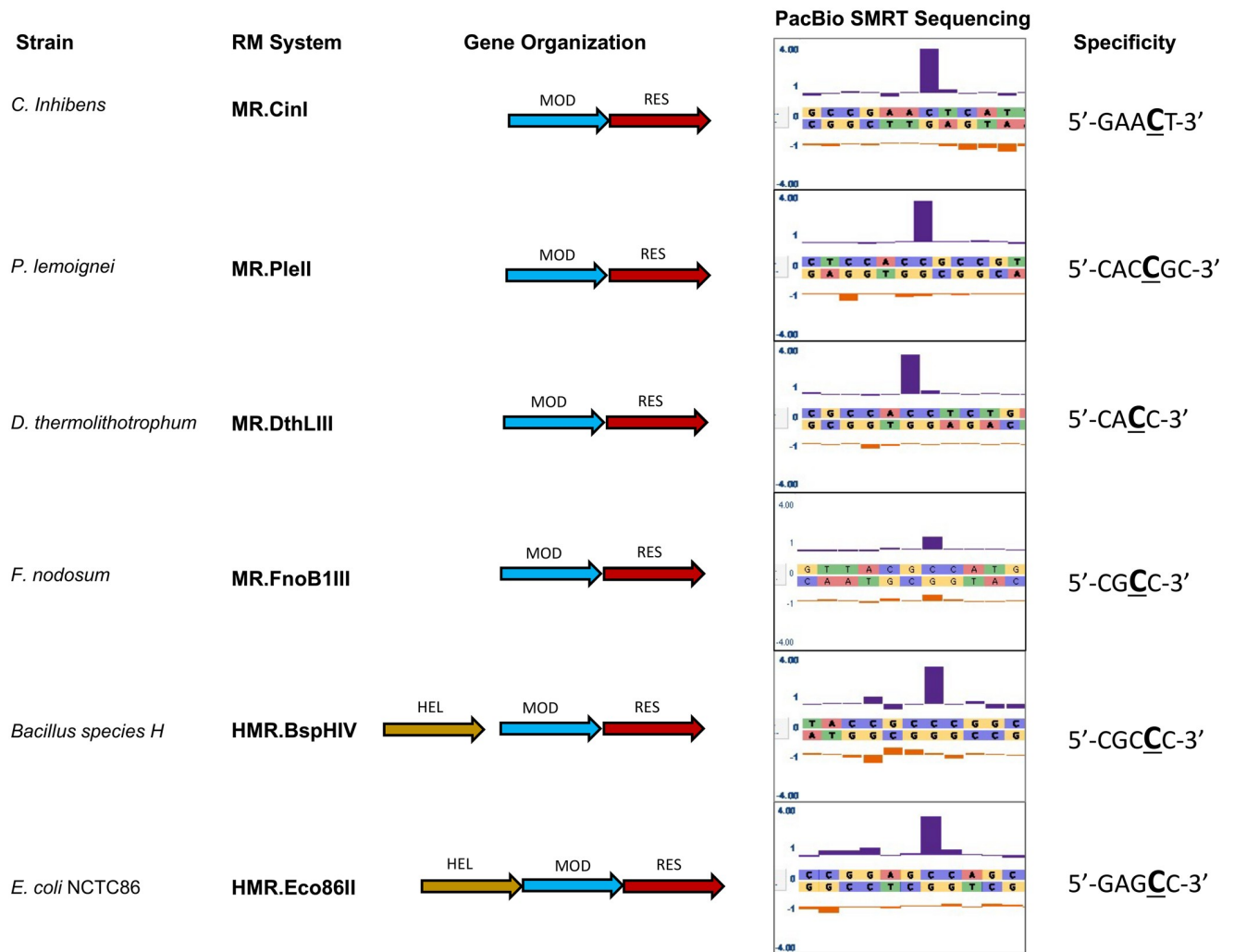
We characterized the recognition sequence motif, and the base modified to m4C within that motif, for six m4C Type III restriction-modification (RM) systems, each recognizing a novel target motif (Table 1). These systems were identified based upon observing an asymmetric, single-strand-modified motif typical of Type III RM systems, but having m4C methylation rather than m6A, in the SMRT sequencing methylome analyses of six diverged bacteria: *Carnobacterium inhibens* subspecies *gilichinskyi*, *Pseudomonas lemoignei*, *Desulfurobacterium thermolithotrophum*, *Fervidobacterium nodosum* Rt17-B1, *Escherichia coli* NCTC86 and *Acinetobacter species* H (formerly *Bacillus species* H, thus systems are named BspHI, BspHII, etc.) (Fig 1). Putative Type III RM system genes were identified within the genome sequences of

**Table 1. Characterized m4C Type III R-M systems.**

<u>Enzyme</u>	<sup>a</sup> <u>Specificity</u>	<u>Host Organism</u>	<u>gDNA source</u>
BspHIV	CG <u>CC</u>	<i>Acinetobacter species</i> H.	NEB 394
CinI	GA <u>ACT</u>	<i>Carnobacterium inhibens</i>	DSM 13024
DthLIII	CA <u>CC</u>	<i>Desulfurobacterium thermolithotrophum</i>	DSM 11699
Eco86II	GAG <u>CC</u>	<i>Escherichia coli</i> NCTC86	Dr. Swaine L. Chen
FnoB1III	CG <u>CC</u>	<i>Fervidobacterium nodosum</i> Rt17-B1	DSM 5306
PleII	CAC <u>CGC</u>	<i>Pseudomonas lemoignei</i>	NEB 418

<sup>a</sup>The cytosine base methylated in the recognition motif is underlined.

<https://doi.org/10.1371/journal.pone.0253267.t001>



**Fig 1. Identification of six Type III restriction-modification (RM) systems that methylate the N4 atoms of deoxycytosine bases.** Four systems share a common architecture with previously-identified Type III systems comprised of Res and Mod subunits, whereas two additional systems appear to encode a third protein subunit located 5' of their respective *mod* genes. Note that in the case of the BspHIV system there is approximately 140bp of non-coding sequence between the 3'-end of the helicase-like ATPase subunit and the 5'-end of the *mod* gene. SMRT-sequencing was employed to identify derived consensus methylation motifs for each of the six Type III RM systems. The tall blue bars indicate the Inter-Pulse Duration (IPD) values associated with positions of N4C methylation.

<https://doi.org/10.1371/journal.pone.0253267.g001>

these strains by a sequence similarity search comparing putative protein sequences to known Type III RM systems, using SeqWare software from the REBASE database [23]. Five of these six strains contained just one putative Type III system, while the *Acinetobacter species H* strain contained two putative Type III systems.

Each of these RM systems contains a Mod subunit with conserved amino-methyltransferase motifs, including the AdoMet binding motif I (FGG) and catalytic motif IV (DPPY/F) (Fig 2). Similarly, the Res subunit of each system contains conserved Walker A (GSGKT) and Walker B (DEXH) motifs typical of ATP-dependent helicase domains, and a canonical PD-(D/E)XK endonuclease catalytic motif in the carboxy terminal portion of the protein (Fig 3).

To match the identified specificity motif having m4C methylation to the putative Type III systems, and to identify which BspH system is responsible for the m4C-modified motif, each

Colored PROMALS3D alignment of m4C-Type III Methyltransferase proteins (sequences in aligned order)



Mtase catalytic motif IV "DPPY/F" and AdoMet binding motif I "F-G-G" highlighted in yellow. red text indicates predicted helices, blue text indicates predicted beta strands.

Fig 2. Alignment of the six characterized Type III Methyltransferases that methylate the N4 atom of cytosine. Each of the six Mod proteins are beta-class methyltransferases and contain conserved sequence motifs corresponding to the methyltransferase catalytic motif IV (DPPY) and the AdoMet binding motif I (FGG) described by Malone et al., [24].

https://doi.org/10.1371/journal.pone.0253267.g002

Mod gene, or both Mod plus Res gene, was PCR amplified, ligated into a suitable vector and transformed into an E. coli host. The transformed cells were grown to allow methylase expression and host modification, then total genomic and plasmid DNA was isolated and methylation motifs were identified using SMRT sequencing and methylome analyses.

SMRT sequencing and LC-MS identification of m4C modification and recognition motifs

In contrast to all previously characterized Type III RM systems which form m6A, each of the newly identified systems methylate deoxycytosine bases yielding m4C. The methyltransferase of C. inhibens subspecies gilchinskyi (M.CinI) methylates the C at the fourth position of 5' - GAAC T - 3' when expressed in E. coli, though the SMRT methylome analysis indicates methylation of this particular construct is partial (Table 2). The methyltransferase of P. lemoignei (M. PleII) methylates the C at the fourth position of 5' - CACCG C - 3'. In this case expression of the Mod gene from the high copy plasmid pRRS resulted in full methylation of the natural target motif, CACCGC, but also additional methylation at near cognate sites representing relaxed specificity or 'star' activity. The methylome analysis indicated significant methylation at CACCGC sites (Table 2), and lesser modification at CACCGT and CAACGC. This increased modification is likely a result of over-expression in a non-native context, as no modification of the extra star sites was observed in the native host. The methyltransferase of E. coli NCTC86

Colored PROMLAS3D alignment of m4C-Type III Endonuclease proteins (sequences in aligned order)



Helicase Walker A motif **GSGKT** and Walker B motif **DEAH** highlighted in yellow. PD-ExK endonuclease motifs **PD** and **DPK** (or **EPK**/**EPH**) highlighted in magenta. red text indicates predicted helices, blue text indicates predicted beta strands.

**Fig 3. Alignment of the six characterized Type III endonucleases.** The Res subunit of each system contains a conserved Walker A (GSGKT) and Walker B DEAD-BOX/DEXH motif typically associated with ATP-dependent DNA helicase proteins, and a canonical PD-(D/E)XK endonuclease catalytic motif located toward the carboxy terminus.

<https://doi.org/10.1371/journal.pone.0253267.g003>

(M.Eco86II) methylates the C at the fourth position of 5' -GAGCC-3'. Perhaps unsurprisingly, recombinant expression of this *E. coli* gene in our laboratory *E. coli* strain gave full methylation that was specific for the native site, even when expressed from the high copy vector, Table 2. We expressed the methyltransferase of both of the putative Type III systems found in *A. species* H and performed methylome analysis. M.BspHIV, locus tag HUK62\_04085, was found to efficiently methylate the C at the fourth position of 5' -CGCCC-3' when expressed in *E. coli*. The other putative Type III system from this strain was found to modify adenine to m6A in the recognition motif AGCACC and was designated BspHIII (MTase gene locus tag is HUK62\_00355) Table 2. The *D. thermolithotrophum* and *F. nodosum* R17-1B1 systems (M. DthLIII and M.FnoB1III) were found to methylate the Cs at the third position of 5' -CACC-3' and 5' -CGCC-3', respectively, the latter motif having been previously reported by Blow *et al.*, [25].

Recombinant expression of the *mod* gene alone for M.BspHIV, M.CinI, M.Eco86II and M. PleII was sufficient to generate cytosine methylation in a laboratory *E. coli* strain lacking any

Table 2. Modification detection by SMRT sequencing.

(dam GATC m6A reported for <i>E. coli</i> clones expressing dam MTase)				
Enzyme	MTase construct	<sup>a</sup> m4C % detected	Coverage	dam m6A %
BspHIV		CGCCC		
	Native organism	98.5	181x	na
	M.BspHIV-pRRS	97.6	90x	99.9
CinI		GA <u>ACT</u>		
	Native organism	98.4	146x	na
	M.CinI-pRRS	25.3	82x	99.9
DthLIII		CA <u>CC</u>		
	Native organism	85.5	550x	na
	M.DthLIII-pRRS	0	93x	99.9
	M.DthLIII+R.DthLIII-pRRS	90.6	2954x	100
Eco86II		GAG <u>CC</u>		
	Native organism	92.3	77x	99.6
	M.Eco86II-pRRS	98.9	76x	99.9
FnoB1III		CG <u>CC</u>		
	Native organism	47	not reported	na
	M.FnoB1III-pRRS	0	81x	99.9
	M.FnoB1III+R.FnoB1III-pRRS	63.7	163x	99.9
PleII		CAC <u>CGC</u>		
	Native organism	99.9	158x	na
	M.PleII-pRRS	73.9	86x	99.9
	M.PleII-pRRS CACCNC "star"	69.3	86x	99.9

<sup>a</sup>m4C percent detected" is percent of m4C motifs called as modified using the standard PacBio methylome analysis with default settings. Coverage is per strand for the motif. dam m6A methylation calls are shown as a control for systems expressed in *E. coli* carrying the dam methyltransferase.

<https://doi.org/10.1371/journal.pone.0253267.t002>

endogenous methylation (ER2796). In addition to the SMRT sequencing methylome analyses, the m4C modified cytosine was detected via hydrolysis of the plasmid DNA to individual nucleosides, followed by quantitation using liquid chromatography and mass spectrometry (LC-MS), Table 3. Methylated cytosine residues were detected for these four systems expressing Mod alone. The percentage of modified C bases detected closely matched the expected value for full modification at the specific recognition motif for M.BspHIV and M.Eco86II (Table 3), in agreement with the SMRT methylome analysis detection of 97.6% and 98.9% of their sites as methylated (Table 2). An overabundance of modified cytosine relative to the number of target sites in the M.PleII expressing strain. Table 3 reflects the additional modification at near cognate 'star' sites for this clone that is observed in the SMRT methylome analysis, Table 2. For M.CinI the LC-MS data showed only half the expected percentage of modified cytosine, which again agrees well with the SMRT methylome analysis that detected only partial modification.

### Modification requirement for both MTase and ENase

In the case of the two systems isolated from thermophilic species, *D. thermolithotrophum* and *F. nodosum Rt17-B1*, we were unable to detect activity of the respective *mod* genes in *E. coli* using either SMRT methylome analyses or the LC-MS assay. Initially we assumed this was most likely a consequence of expressing the genes at a temperature much lower than optimal



Table 3. LC-MS analyses of m4C modification for Type III systems.

Enzyme	Genes	Specificity	m4C%	Target	Plasmid	Expected m4C%	
			(m4C/m4C+C)	sites	size(kb)	sites/total "C"	
BspHIV	mod	CG <u>CC</u>	0.35%	9	5.9	0.36%	
			0.35%				
CinI	mod	GA <u>ACT</u>	0.21%	9	4.7	0.43%	
			0.21%				
DthLIII	mod	CA <u>CC</u>	0.0%	28	5.3	1.34%	
			0.0%				
	mod+res			1.00%	32	8.3	1.11%
				1.00%			
FnoB1III	mod	CG <u>CC</u>	0.0%	31	5.3	1.48%	
			0.0%				
	mod+res			1.31%	34	8.3	1.10%
				1.35%			
Eco86II	mod	GAG <u>CC</u>	0.23%	6	5.9	0.24%	
			0.23%				
PleII	mod	CA <u>CCGC</u>	1.24%	5	5.3	0.18%	
			1.25%				

Plasmid DNAs expressing Type III *mod* genes were isolated from stationary phase cultures of *E. coli* C2523 (*dcm*-/*dam*+). Two independent isolates of each plasmid DNA were hydrolyzed to nucleosides using the Nucleoside Digestion Mix and resolved and quantified using LC-MS. m4C percentage methylation data indicate the mean values for each pair of independent plasmid replicates. Expected m4C percentage indicates the percentage of recognition motif cytosine targets in the substrate plasmid. The methylated bases from each isolate are highlighted in bold text. Note that in the case of the two thermophilic isolates co-expression of both *mod* and *res* genes is required for expression of their cognate MTase activities.

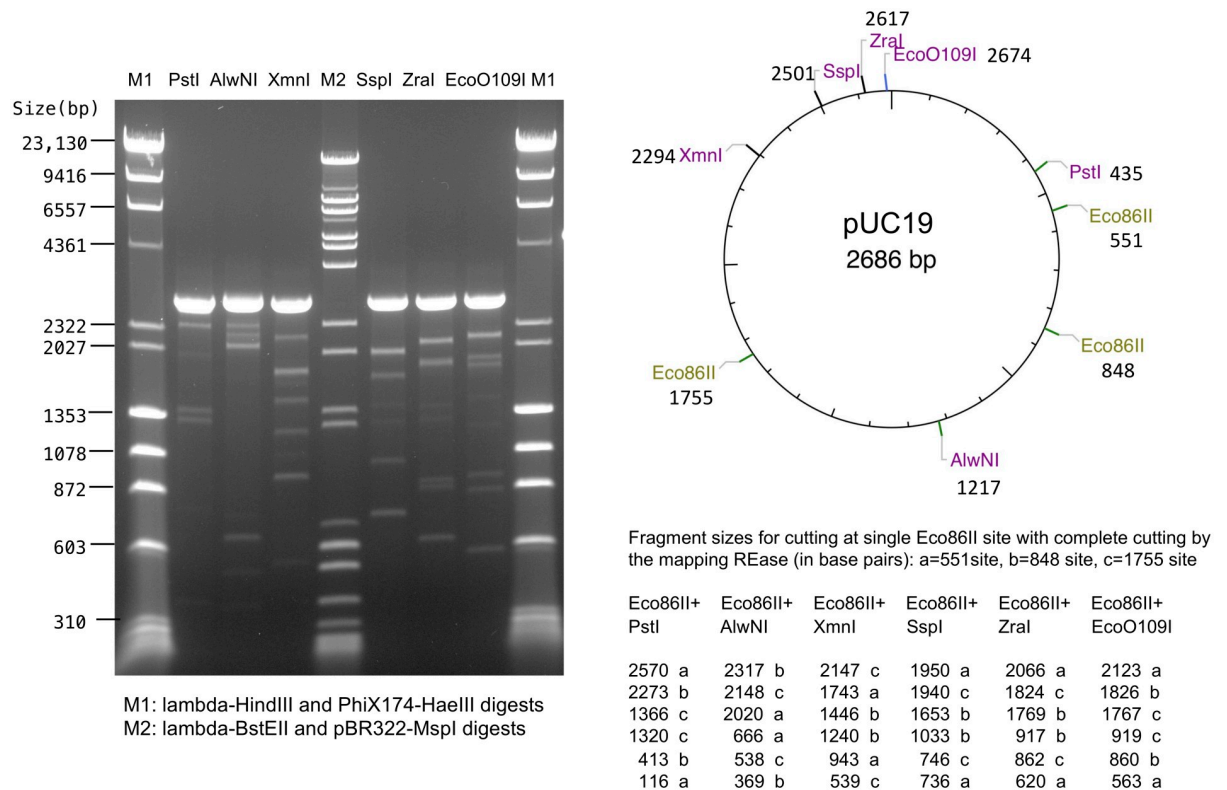
<https://doi.org/10.1371/journal.pone.0253267.t003>

for the native strains, or alternatively, that the codon preferences of the thermophile genes were suboptimal for *E. coli* expression. In an attempt to address the latter, we obtained a synthetic *Desulfurobacterium mod* gene with a sequence that was codon optimized for expression in the heterologous *E. coli* host. However, when the synthetic gene was expressed in *E. coli*, we were still unable to detect any evidence of m4C modification by LC-MS, Table 3. This suggested that expression of the isolated *mod* gene alone was insufficient to confer methylation competency.

Although we are unaware of any precedent among previously characterized Type III systems, we decided to address the possibility that expression of the cognate *res* gene might be required to support methyltransferase activity. We therefore co-expressed the *mod* and *res* genes of both the *Desulfurobacterium* and *Fervidobacterium* RM systems in *E. coli* and in each case the previously absent m4C methyltransferase activity was detected upon co-expression of *mod* with the cognate *res* genes, Table 3.

### Restriction endonuclease activity

To test if these systems have Type III endonuclease activity, we performed *in vitro* assays using a protein extract prepared from the Eco86II system native strain. We prepared a crude protein extract and partially purified this by ion exchange/affinity chromatography over a Heparin column. Heparin fractions were assayed for endonuclease activity by cutting either pUC19 or pBR322 plasmid DNA, both of which have head to head (HtH) oriented 5' -GAGCC-3' Eco86II sites. After allowing the Type III system to act on the circular plasmid DNA, which



**Fig 4. Eco86II endonuclease activity.** Partially purified Eco86II enzyme was used to digest pUC19 plasmid substrate containing 3 sites: one in the forward direction and two in reverse orientation. After cutting with Eco86II, the reaction was divided and cut with six REases that cleave pUC19 once at known positions. Although the Eco86II digestion is only partial, fragments having lengths corresponding to the distance between the known REases and the Eco86II cleavage positions are observed, demonstrating specific Eco86II cleavage downstream from its GAGCC recognition motif. The original gel photograph is available as [S1 Raw image](#).

<https://doi.org/10.1371/journal.pone.0253267.g004>

enables 1D communication between HtH sites, the reaction was split and REases that cut once at various locations were added to linearize the plasmid and drop out diagnostic fragments generated by complete cutting at the linearizing REase site and any cutting by Eco86II. Although only partial Eco86II cutting was observed, the sites of cutting mapped to the expected positions for Eco86II (approximately 25 bp 3' to the GAGCC Eco86II recognition sites in these DNAs), confirming that the Eco86II system has restriction as well as methylation function (Fig 4).

### Additional putative ATPase gene adjacent to MTase and ENase in some systems

The genomic organization for four of the new systems is similar to that of previously characterized Type III systems insofar as each contains a prototypical arrangement of *mod* and *res* genes. The remaining two systems differ in that they each have an additional, highly conserved gene located immediately 5' of the *mod* gene and which has amino acid motifs typical of ATP-dependent SF2 DNA helicases. This potential extra helicase-like ATPase gene was not required for methyltransferase activity. We performed MS-MS peptide spectroscopy on the partially purified Eco86II restriction enzyme isolated from native cells (Fig 4) which detected both the *mod* and *res* proteins but no peptides from this putative additional helicase-like ATPase.

Table 4. Putative m4C Type III R-M systems.

Enzyme	Specificity	Host Organism
Aha10306IV	CCACG	Acinetobacter haemolyticus NCTC10306
Bbr10542II	TCACCG	Bordetella bronchiseptica NCTC10542
Bca78520I	GAACCC	Brevibacterium casei 3012STDY7078520
Bhi13200II	CACCGC	Bordetella hinzii NCTC13200
Bpe564I	AGCCGCC	Bordetella pertussis H564
EcoF9792I	GAGCC	Escherichia coli F9792
EcoB4103II	GAGCC	Escherichia coli O27:H7 B4103-1
Gel1640III	GGACCG	Geopsychrobacter electrophilus DSM 15401
Kaq16071II	GGACT	Kangiella aquimarina DSM 16071
MspCY2I	AGCGCC	Microbacterium sp KROCY2
MspMM1I	CCACCC	Microbulbifer species CCB-MM1
Pmi29II	GAACT	Proteus mirabilis AR_0029
Sen6016III	GAGCC	Salmonella enterica NCTC6016
Sen10536IV	GAGCC	Salmonella enterica NCTC10436
TmeBIV	CGCC	Thermosiphon melanesiensis BI429 DSM 12029
VflA21I	CARCC	Vibrio fluvialis AK 1296-A2-1
Vna16374II	CCACCG	Vibrio natriegens CCUG 16374
Vsp20II	GCACCC	Vogesella species NEB_P20

Additional Type III m4C systems predicted from SMRT methylome data and bioinformatics as reported in the REBASE database. These additional Type III m4C systems are predicted from bacteria that have a single Type III system based on bioinformatic analysis of their predicted genes, and which have a single Type III-like single strand m4C modified motif detected in SMRT methylome analysis. While these systems have not been verified by biochemical assays or by expressing the system in a heterologous host and performing SMRT methylome analyses, the prediction of the recognition motif and function for these systems is of high confidence.

<https://doi.org/10.1371/journal.pone.0253267.t004>

Type III systems that use m4C methylation are widespread. There are now 24 examples of Type III RM systems in REBASE [23] that can confidently be predicted to match a SMRT methylome m4C specificity motif. These enzymes recognize 17 different motifs and 5 different patterns of bases that are specifically recognized 5' (2, 3 or 4 base pairs) and 3' (1, 2 or 3 base pairs) to the cytosine that is methylated, Table 4.

## Discussion

In this work we describe the first examples of Type III RM systems which exploit methylation of the N4 atoms of deoxycytosine for host DNA protection. Using SMRT sequencing of genomic DNAs we identified six candidate m4C Type III RM systems and defined the sequence specificities of their respective *mod* genes.

The m4C Type III systems characterized have rather diverse Mod protein sequences (Fig 2). This suggests either m4C methylation arose early in the evolution of Type III systems, or that it has arisen multiple times within diverged families of Type III systems. Since m6A and m4C MTases have identical chemistry in acting on an exocyclic nitrogen it seems likely that m4C modification could readily evolve from m6A MTases. The Res proteins are also quite diverse, also suggesting ancient lineage (Fig 3).

These six Type III RM systems characterized appear to fall into three groups. The *C. inhibens* and *P. lemoignei* group have a typical Type III architecture consisting of two genes, *mod* and *res*. In each case, recombinant expression of the respective *mod* gene alone was sufficient to confer robust m4C methyltransferase activity.

The second group is comprised of two systems isolated from the thermophilic bacteria *D. thermolithotrophum* and *F. nodosum Rt17-B1*. Robust methyltransferase activity was only detected in these systems when both the *mod* and *res* genes were co-expressed, suggesting that unlike other Type III systems *mod* alone is insufficient for methylation.

The third group of Type III systems, identified in the genomes of *A. species H* and *E. coli* NCTC86, differ from the other systems by virtue of an additional gene located 5' of the *mod* and *res* genes. This additional gene contains sequence motifs typically associated with ATP-dependent helicase enzymes but does not contain a PD-ExK or another endonuclease motif. This putative ATPase is not homologous to the RES subunit of the adjacent Type III system. The biological role of this additional putative helicase-like ATPase gene is unclear although it appears to be intimately associated with its neighboring Type III system through the logic of guilt by association based on their consistently close proximity in more than 500 genomes. The Res subunits of Type III systems function by using the energy of ATP hydrolysis to rearrange the Mod2Res complex into a translocation competent complex that diffuses along the DNA to contact another Mod2Res complex bound at an inverted site. Whether this potential additional helicase-like ATPase supplements the translocase activities of the canonical Res subunit remains to be seen, but the intimate genomic association with the Mod and Res in these systems suggests this additional helicase-like ATPase plays some role in these systems. We examined the partially purified Eco86II enzyme preparation that gave partial cutting *in vitro* (Fig 4) using Mass Spec and did not observe any peptides from this additional helicase-like ATPase gene, indicating this extra protein is not required for Eco86II cleavage. Any role it may have will require further investigation.

Type III m6A methylation has been shown important in regulating gene expression and thus virulence through phase variable on-off expression in a number of bacteria [26]. The m4C Type III Mod systems extend this capability, but to date no examples have been described.

With these findings, m4C modification is now found across Type I, Type II and Type III RM systems. The Type I m4C systems have gamma-class MTases and all known such systems have one m6A and one m4C MTase protein and modify one strand to m6A and one to m4C. The Type I MTase proteins and their corresponding S recognition domains are highly similar, suggesting recent evolution of the m4C MTase from an m6A MTase ancestor. In contrast, the Type III m4C systems reported here are highly diverged and thus likely of ancient origin. Type II beta-class MTases are also quite diverged and likewise include enzymes that methylate either m6A or m4C. SMRT methylome analyses strongly suggest m4C modification also functions in further bacterial defense systems such as BREX systems.

## Supporting information

### S1 File. Raw sequence data accession numbers.

(DOCX)

### S1 Raw image.

(PDF)

## Acknowledgments

The authors wish to express their gratitude to Dr Swaine L. Chen of the National University of Singapore, for his generous gift of genomic DNA of *E. coli* NCTC86, the original *Escherichia coli* isolate of Theodor Escherich and source of the Eco86II system [27].

## Author Contributions

**Investigation:** Iain A. Murray, Yvette A. Luyten, Alexey Fomenkov, Nan Dai, Ivan R. Corrêa, Jr., William G. Farmerie, Tyson A. Clark, Jonas Korlach, Richard D. Morgan.

**Project administration:** Iain A. Murray, Richard J. Roberts.

**Supervision:** Richard J. Roberts.

**Writing – original draft:** Iain A. Murray, Richard D. Morgan.

**Writing – review & editing:** Iain A. Murray, Richard D. Morgan, Richard J. Roberts.

## References

1. Haberman A, Heywood J, Meselson M. DNA modification methylase activity of *Escherichia coli* restriction endonucleases K and P. *Proc Natl Acad Sci*. 1972; 69(11):3138–41. <https://doi.org/10.1073/pnas.69.11.3138> PMID: 4564204
2. Reiser J, Yuan R. Purification and properties of the P15 specific restriction endonuclease from *Escherichia coli*. *J Biol Chem*. 1977; 252(2):451–6. PMID: 137900
3. Bächli B, Reiser J, Pirrotta V. Methylation and cleavage sequences of the EcoP1 restriction-modification enzyme. *J Mol Biol*. 1979; 128(2):143–63. [https://doi.org/10.1016/0022-2836\(79\)90123-2](https://doi.org/10.1016/0022-2836(79)90123-2) PMID: 219202
4. Hadi SM, Bächli B, Shepherd JCW, Yuan R, Ineichen K, Bickle TA. DNA recognition and cleavage by the EcoP15 restriction endonuclease. *J Mol Biol*. 1979; 134(3):655–66. [https://doi.org/10.1016/0022-2836\(79\)90372-3](https://doi.org/10.1016/0022-2836(79)90372-3) PMID: 231671
5. Meisel A, Bickle TA, Krieger DH, Schroeder C. Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. *Nature*. 1992; 355(6359):467–9. <https://doi.org/10.1038/355467a0> PMID: 1734285
6. Meisel A, Mackeldanz P, Bickle TA, Krüger DH, Schroeder C. Type III restriction endonucleases translocate DNA in a reaction driven by recognition site-specific ATP hydrolysis. *EMBO J*. 1995; 14(12):2958–66. PMID: 7796821
7. Saha S, Rao DN. ATP hydrolysis is required for DNA cleavage by EcoPI restriction enzyme. *J Mol Biol*. 1995; 247(4):559–67. [https://doi.org/10.1016/s0022-2836\(05\)80137-8](https://doi.org/10.1016/s0022-2836(05)80137-8) PMID: 7723013
8. Janscak P, Sandmeier U, Szczelkun MD, Bickle TA. Subunit assembly and mode of DNA cleavage of the type III restriction endonucleases EcoP1I and EcoP15I1 Edited by J. Karn. *J Mol Biol*. 2001; 306(3):417–31. <https://doi.org/10.1006/jmbi.2000.4411> PMID: 11178902
9. van Aelst K, Tóth J, Ramanathan SP, Schwarz FW, Seidel R, Szczelkun MD. Type III restriction enzymes cleave DNA by long-range interaction between sites in both head-to-head and tail-to-tail inverted repeat. *Proc Natl Acad Sci*. 2010; 107(20):9123–8. <https://doi.org/10.1073/pnas.1001637107> PMID: 20435912
10. Wyszomirski KH, Curth U, Alves J, Mackeldanz P, Möncke-Buchner E, Schutkowski M, et al. Type III restriction endonuclease EcoP15I is a heterotrimeric complex containing one Res subunit with several DNA-binding regions and ATPase activity. *Nucleic Acids Res* 2012; 40(8):3610–22. <https://doi.org/10.1093/nar/gkr1239> PMID: 22199260
11. Butterer A, Pernstich C, Smith RM, Sobott F, Szczelkun MD, Tóth J. Type III restriction endonucleases are heterotrimeric: comprising one helicase-nuclease subunit and a dimeric methyltransferase that binds only one specific DNA. *Nucleic Acids Res* 2014; 42(8):5139–50. <https://doi.org/10.1093/nar/gku122> PMID: 24510100
12. Gupta YK, Yang L, Chan S-H, Samuelson JC, Xu S-y, Aggarwal AK. Structural insights into the assembly and shape of Type III restriction-modification (R-M) EcoP15I complex by small-angle X-ray scattering. *J Mol Biol*. 2012; 420(4–5):261–8. <https://doi.org/10.1016/j.jmb.2012.04.026> PMID: 22560991
13. Gupta YK, Chan S-H, Xu S-Y, Aggarwal AK. Structural basis of asymmetric DNA methylation and ATP-triggered long-range diffusion by EcoP15I. *Nat Commun*. 2015; 6:7363-. <https://doi.org/10.1038/ncomms8363> PMID: 26067164
14. Piekarowicz A, Bickle TA, Shepherd JCW, Ineichen K. The DNA sequence recognised by the HinIII restriction endonuclease. *J Mol Biol*. 1981; 146(1):167–72. [https://doi.org/10.1016/0022-2836\(81\)90372-7](https://doi.org/10.1016/0022-2836(81)90372-7) PMID: 6267295
15. Sears A, Peakman LJ, Wilson GG, Szczelkun MD. Characterization of the Type III restriction endonuclease PstII from *Providencia stuartii*. *Nucleic Acids Res* 2005; 33(15):4775–87. <https://doi.org/10.1093/nar/gki787> PMID: 16120967

16. Xu S-y, Nugent RL, Kasamkattil J, Fomenkov A, Gupta Y, Aggarwal A, et al. Characterization of Type II and III Restriction-Modification Systems from *Bacillus cereus* Strains ATCC 10987 and ATCC 14579. *J Bacteriol.* 2012; 194(1):49–60. <https://doi.org/10.1128/JB.06248-11> PMID: 22037402
17. Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, et al. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res* 2011; 40(4):e29–e. <https://doi.org/10.1093/nar/gkr1146> PMID: 22156058
18. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, et al. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat Biotechnol.* 2012; 30:1232. <https://doi.org/10.1038/nbt.2432> PMID: 23138224
19. Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, et al. The methylomes of six bacteria. *Nucleic Acids Res* 2012; 40(22):11450–62. <https://doi.org/10.1093/nar/gks891> PMID: 23034806
20. Leonard MT, Panayotova N, Farmerie WG, Triplett EW, Nicholson WL. Complete Genome Sequence of *Carnobacterium gilichinskyi* Strain WN1359T (DSM 27470T). *Genome Announc.* 2013; 1(6).
21. Anton BP, Mongodin EF, Agrawal S, Fomenkov A, Byrd DR, Roberts RJ, et al. Complete Genome Sequence of ER2796, a DNA Methyltransferase-Deficient Strain of *Escherichia coli* K-12. *PLoS One.* 2015; 10(5):e0127446–e. <https://doi.org/10.1371/journal.pone.0127446> PMID: 26010885
22. Pei J, Kim BH, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* 2008; 36(7):2295–300. <https://doi.org/10.1093/nar/gkn072> PMID: 18287115
23. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* 2015; 43(Database issue):D298–9. <https://doi.org/10.1093/nar/gku1046> PMID: 25378308
24. Malone T, Blumenthal RM, Cheng X. Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes. *J Mol Biol.* 1995; 253(4):618–32. <https://doi.org/10.1006/jmbi.1995.0577> PMID: 7473738
25. Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, et al. The Epigenomic Landscape of Prokaryotes. *PLoS Genet.* 2016; 12(2):e1005854. <https://doi.org/10.1371/journal.pgen.1005854> PMID: 26870957
26. Seib KL, Srihanta YN, Atack JM, Jennings MP. Epigenetic Regulation of Virulence and Immuno-evasion by Phase-Variable Restriction-Modification Systems in Bacterial Pathogens. *Annu Rev Microbiol.* 2020; 74:655–71. <https://doi.org/10.1146/annurev-micro-090817-062346> PMID: 32689914
27. Escherich T. Die Darmbakterien des Säuglings und ihre Beziehungen zur Physiologie der Verdauung. 1886.