PNAS
www.pnas.org

Supplementary Information for

Phage-encoded Ten-eleven Translocation Dioxygenase (TET) Is Active in C5-Cytosine Hypermodification in DNA

Evan J. Burke[a,1], Samuel S. Rodda [a,1], Sean R. Lund [a,1], Zhiyi Sun[a], Malcolm R. Zeroka[a], Katherine H. O'Toole[a], Mackenzie J. Parker[a], Dharit S. Doshi[a], Chudi Guan[a], Yan-Jiun Lee[a], Nan Dai[a], David M. Hough[a], Daria A. Shnider[a], Ivan R. Corrêa[a], Jr.[a], Peter R. Weigele[a,2], and Lana Saleh[a,2]

[a] Research Department, Molecular Enzymology Division, New England Biolabs, Ipswich, MA 01938

[1]E.J.B., S.S.R., and S.R.L. contributed equally to this work.

[2]To whom correspondence should be addressed: weigele@neb.com or saleh@neb.com

This PDF file includes:

Supplementary Materials and Methods
Figures S1 to S18
Tables S1 to S3
SI References

SI Materials and Methods

Data Sources. The metavirome contigs were obtained from the GOV 2.0 (1) and IMG/VR (2) databases (SI Appendix, Table S1). IMG/VR contig names consists of three alphanumeric strings: the first number corresponds to the IMG Genome ID, the second corresponds to the Gold Analysis Project ID, and the third identifies the contig from within that project's data set. The GOV 2.0 contig IDs have two types of assemblies. The first uses all the reads (*i.e.* ALL_assembly) for population genetics, auxiliary metabolic gene discovery, and other genome-based discoveries, and the second uses the smallest number of reads that exist in any library (i.e. subsampled_assembly) for diversity metrics comparisons across different ocean regions, water depths, and other factors. The "NODE_#" is a unique number given to each contig generated from the assembly per sample. The contigs are arranged according to sequence length with the longer contig given the smaller "NODE_" number.

Bioinformatic Analysis of C5-MT/TET-encoding Contigs. A dataset of predicted protein sequences was examined with BLAST to find matches to C5-MT and TET queries. TET searches were performed using the TET/JBP family sequences from the following organisms: *Mycobacterium* phage Cooper (YP_654899.1), *Mycobacterium* phage Acadian (AER48916.1), *Persicivirga* phage P12024L (AFM54746.1), mouse TET 1, 2, and 3, JBP1 and 2 from *Leishmania major*, and SAR234 cluster bacterium (WP_010154308.1). The C5-MT queries were derived from an initial dataset of 4158 DNA cytosine-C5-methyltransferases obtained from REBASE. These were passed through CD-HIT with a 30 % cutoff to make a smaller, non-redundant dataset and then representatives from each cluster were selected by hand for use in the queries.

Unannotated metagenomic contigs (401 contigs) were then obtained from the databases by text matching the protein ID to the contig ID. The region containing the pairs, as well as several kilobases up and downstream, were extracted for analysis. Open reading frames were annotated using the KBase Knowledge Discovery Environment (3). A more thorough annotation process was also applied using the HMMER hmmscan program (http://hmmer.org; HmmerWeb version 2.41.1). Subsequently, filters were implemented for only high-confidence annotations by applying an *E* value cutoff of 1 x 10^{-4} in BLAST searches for either C5-MT or TET. This dataset (130 contigs) was then examined for the proximity and direction of the genes of interest. Contigs were manually inspected and removed if the C5-MT and TET coding frames were antiparallel or were separated by more than 5 other genes. This refinement limited the dataset to only 77 contigs. In several cases, metagenomic contigs from distinct geographical locations exhibited identical gene sequence and organization. Therefore, a manual filtering process was applied to ensure that only unique gene contigs would be considered for experimental testing. Of the 401 starting contigs, 32 met all the aforementioned criteria (SI Appendix, Fig. S2 and Table S1). On this high-confidence dataset, structural modeling was applied using the Phyre2 protein fold recognition server (4). High confidence (> 60 %) domain analysis was manually applied as annotations to the clusters in Geneious Prime software (version 2020.0.5) to allow previously undefined open reading frames to be annotated and already annotated genes to be confirmed.

Sequence Alignment and Phylogenetic Tree Classification of C5-MT and TET Proteins. Sequence alignment was performed by MAFFT Multiple Aligner v7.388 (JTT200 Scoring Matrix, 1.53 Gap Open Penalty, 0.123 Offset Value) Biomatters Ltd.

within Geneious. Phylogenetic tree classification was done using Geneious Tree Builder, Jukes-Cantor Distance Model, UPGMA Method.

Chemicals, Media, Enzymes, and Microbes. Unless otherwise specified, all chemicals were obtained from Sigma-Aldrich and used without additional purification. Media and media components were sourced from Beckton Dickinson-Difco. Plasmids were obtained from Genscript. Enzymes, DNA purification kits, as well as competent *E. coli* cells were obtained from NEB.

***In Vivo* C5-MT and TET Characterization of Activities.** Two plasmids with compatible origins of replication and antibiotic resistance were used in our *in vivo* studies. The C5-MT genes were cloned into pACYC184 plasmid (NEB) that has a p15A origin of replication and a *chloramphenicol* resistance gene. TET genes were cloned into the isopropyl β-D-1-thiogalactopyranoside (IPTG)-inducible pJS119K replicon with ColE1 origin of replication and *kanamycin* resistance (5). 10-100 ng of each plasmid was transformed into T7 Express Competent *E. coli* cells following the protocol provided by the supplier. Cells were grown at 37 ºC on rich media containing 10 g tryptone, 5 g NaCl, and 5 g yeast extract per 1 L and 50 μg/mL of each of the appropriate antibiotics. TET genes were induced with 250 μM IPTG and the cultures were grown at 18 ºC for 16 h before harvest.

After collection of cell pellets by centrifugation, the Monarch® gDNA Purification Kit was used to extract genomic DNA. The Monarch® PCR and DNA Cleanup Kit was used to clean up the DNA to a suitable purity. gDNA was digested to nucleosides by a 1 h treatment at 37 ºC with Nucleoside Digestion Mix in a 20 μL reaction volume following the supplier's protocol. Presence of base modifications on the isolated *E. coli* genome as a result of the *in vivo* expressed enzyme activities was investigated using LC/MS or LC-MS/MS.

EM-seq Analysis:

Mapping sites methylated by phage C5-MT (Fig. 4 and SI Appendix Fig. S8). NEBNext® Enzymatic Methyl-seq (EM-seq) kit was used as per instructions for Standard Insert Libraries with few modifications. 100 ng of gDNA isolated from *E. coli* cells with the C5-MT gene expressed overnight was combined with 2 ng of unmethylated Lambda DNA (Promega) and 3 ng of methylated Xp12 DNA (all cytosine sites are methylated) prior to shearing acoustically to 300 bp using a COVARIS S2 instrument equipped with a microTUBE-50. After end-repair and 5mC-adaptor ligation, three enzymatic treatments were applied to discriminate 5mC sites, which are generated by phage C5-MT, from unmethylated Cs (SI Appendix, Fig. S7 (light brown box)) (6). These steps include in a single-pot reaction treatment of the DNA with mouse TET2 and T4 5hmC-DNA β-glucosyltransferase (BGT), which affect the conversion of 5mC to 5-carboxycytosine (5caC) and 5-GlcβmC. The resulting DNA is denatured using formamide and reacted with APOBEC3A, which can only deaminate unprotected cytosine resulting in its conversion to uracil. Following the completion of the protocol, library qualities were assessed on an Agilent Tapestation and quantified using the High Sensitivity D1000 kit. In the event that the library contained primer dimers, the library was processed using 5 equivalents of binding buffer using Monarch® PCR and DNA Cleanup Kit. Libraries were concentrated and pooled at 4 nM for running on an Illumina NextSeq. 5caC and 5-GlcβmC sites are read as C, while unmethylated C is sequenced as T (SI Appendix, Fig. S7 (light brown box)) (6).

Mapping sites hydroxymethylated by phage TET (Fig. 5 and SI Appendix Fig. S8 through S11). E. coli gDNA obtained from an overnight expression of phage C5-MT and TET genes was used to map methylcytosine sites hydroxylated by phage TET. The same protocol described above was followed with the exception that the internal control DNA was comprised of 2 ng of unmethylated Lambda DNA and 3 ng of T4 *gt -/-* DNA. The latter DNA was used to monitor 5hmC protection by BGT. Other modifications to the protocol included omission of mouse TET2 oxidation step and treatment with only BGT (SI Appendix Fig. S7 (blue box)). Also, 5hmC-containing adaptors were used in the adaptor ligation step (7). Following the completion of the protocol, library qualities were assessed on an Agilent Tapestation and quantified using the High Sensitivity D1000 kit. In the event that the library contained primer dimers, the library was processed using 5 equivalents of binding buffer using Monarch® PCR and DNA Cleanup Kit. Libraries were concentrated and pooled at 4 nM for running on an Illumina NextSeq. 5-GlcβmC sites are read as C. 5mC sites that were not hydroxylated by phage TET were deaminated by APOBEC3A to T and sequenced as such (SI Appendix, Fig. S7 (blue box)) (7).

Mapping hydroxymethylation of Xp12 methylated cytosine sites by TET43 (Fig. 6B). 100 ng of Xp12 DNA hydroxymethylated by TET43 *in vitro* was used as input DNA. The rest of the protocol was identical to that described above for hydroxymethylated *E. coli* gDNA.

EM-seq Data Processing and (Hydroxy)Methylation Calling. Raw Illumina sequencing reads were trimmed to remove adapter sequences and low-quality bases from the 3´-end using the Trim Galore software (https://github.com/FelixKrueger/TrimGalore). This process also removed reads that became unpaired or too short (< 20 nt) after adapter/quality trimming. The remaining read pairs were then C to T converted and mapped to the converted reference genomes of *E. coli* (B strain C2566) plus the two transformation vectors (pJS119k and pACYC184) by the Bismark program with default Bowtie2 setting (8). The numbers of converted and non-converted copies of each cytosine site in the *E. coli* genome were summarized from the aligned reads using Bismark methylation extractor. The raw modification level of each covered cytosine site is calculated as: number of unconverted Cs/ (number of unconverted Cs + number of converted Cs). The Fisher's exact test was then applied to call differentially methylated bases in C5-MT-expressed *E. coli* samples in comparison to no-enzyme control with an adjusted q-value of < 0.01 and percent methylation difference > 25% using the methylKit package of R (9). For 5hmC calling, the same analysis was conducted to compare C5-MT and TET co-expressed samples to the no-enzyme control. Flanking sequences of called 5mC or 5hmC sites were extracted from the *E. coli* genome using custom Perl script and were used for motif analysis and plotting by the WebLogo program (Fig. 4 and Fig. 5) (10). Statistical analysis and plotting of sequence-specific modification level were conducted in R (version 3.6.3) (R Core Team (2020); http://www.r-project.org/index.html) using all the cytosine sites in the *E. coli* genome that are in certain sequence context (e.g., GCN, NGC and NGCN) (SI Appendix, Fig. S8, Fig. S9, Fig. S10, Fig. S11, and Fig. S12).

Protein Isolation:

Solubility Screen: Solubility of C5-MT43, GT43-I, GT43-II, GT14-I, and GT14-II were tested using a selection of buffers to determine the optimal purification condition for each of these enzymes. Buffers included sodium citrate pH 5.0, 2-(*N*-morpholino)ethanesulfonic acid (MES) pH 6.0, sodium acetate pH 6.0, potassium phosphate pH 7.0, *N*-[*tris*(hydroxymethyl)methyl]-2-aminoethanesulfonic acid (TES) pH 7.5, 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) pH 7.5, *tris*(hydroxymethyl)aminomethane hydrochloride (Tris-HCl) pH 8.0, and [*tris*(hydroxymethyl)methylamino]propanesulfonic acid (TAPS) pH 9.0. For each buffer type, lysis/wash and elution buffers were prepared. The lysis/wash buffer consisted of 20 mM buffer, 500 mM NaCl, 20 mM imidazole, 1 mM tris(2-carboxyethyl)phosphine (TCEP), and 0.1 % triton X-100. Elution buffer consisted of 20 mM buffer, 500 mM NaCl, 500 mM imidazole, and 1 mM TCEP. Small aliquots of roughly 20 mg wet cell pellets were used to screen protein solubility. Cells were lysed using QSonica Q500 Sonicator with four-prong attachment with the following settings: 30 s (2 s on, 2 s off, pulse) per round, 4 rounds, 15 s rest between rounds. Cells were kept cold in a QSonica freezer block during sonication. The lysates were then centrifuged at 16,0000 x g for 20 min to separate insoluble cell debris. The resulting supernatants containing soluble His-tagged protein were purified using NEBExpress® Ni Spin Columns following the standard protocol. Samples of the lysate soluble fraction, column flow-through, and column elution were analyzed by sodium dodecyl sulfate and polyacrylamide gel electrophoresis (SDS-PAGE). Conditions resulting in the highest protein yield and purity were adopted for large-scale purification. TES pH 7.5 and Tris-HCl pH 8.0 were used in the purification of C5-MT43 and GT43-I, respectively (SI Appendix, Fig. S14A and S16C). Details of the ÄKTA purification are in the next section. MES pH 6.0 was used in the purifications of GT14-I and GT14-II (SI Appendix, Fig. S16C). These enzymes were purified using the NEBExpress® Ni Spin Columns following the manufacturer's protocol.

ÄKTA Protein Purification: C5-MT43, TET43, and GT43-I were purified on a large-scale using an ÄKTA purification system. Roughly 20 g wet cell pellets were resuspended in 50 mL of the appropriate lysis buffer and lysed using a DyHydromatics Shear Jet PL300 Processor with the chiller set to 4 ºC and a pressure of 18,000 PSI with the cell slurry kept on ice. The slurry was passed through the instrument twice to ensure complete lysis. Purification was conducted using a GE Pharmacia ÄKTA Purifier 100 UPC system or an ÄKTA Go system (GE Healthcare). The lysate was loaded onto a HisTrap HP 5mL column at a flow rate of 5 mL/min in lysate/wash buffer. The column was washed with 30 column volume (CV) of the same buffer and then the protein was eluted with a 20 CV linear gradient of 20-500 mM imidazole. The gradient was held at 500 mM imidazole for another 1 CV and 5 mL fractions were collected throughout the procedure.

Fractions containing the protein of interest were identified by SDS-PAGE, pooled, and loaded on a HiTrap Heparin HP 5 mL column. The protein pool was diluted in the appropriate buffer in the presence of 1 mM TCEP to bring the salt concentration down to 50 mM. The Heparin column purification specifics used were as follows: 5 mL/min flow rate, 10 CV wash with buffer A (50 mM buffer, 50 mM NaCl, and 1 mM TCEP), elution gradient 50-1000 mM NaCl over 40 CVs with buffer B (50 mM buffer, 1000 mM NaCl, and 1 mM TCEP), and 5 mL fraction size. The protein of interest was identified by SDS-

PAGE, pooled, and dialyzed against 2 L of storage buffer (50 mM buffer, 250 mM NaCl, 1 mM TCEP, and 50 % (v/v) glycerol). TES pH 7.5 was used in the purification of C5-MT43 (SI Appendix, Fig. S14A). Tris-HCl pH 8.0 was used in the purification of C5-TET43 (SI Appendix, Fig. S13) and GT43-I (SI Appendix, Fig. S16C).

Protein Identities and Concentrations. Protein identities were confirmed by Peptide Sequencing using an Orbitrap ESI with LC-MS/MS. Protein concentrations were determined spectrophotometrically by using molar absorption coefficients at 280 nm (ε_{280}) calculated according to the method of Gill and von Hippel (11).

DNA Substrate Preparation:

Biotinylated Lambda DNA. Lambda DNA was digested to completion with restriction endonuclease AseI according to the manufacturer's guidelines. After heat inactivation at 80 °C for 20 min, the digest was purified using a Monarch® PCR and DNA Cleanup Kit. Fragmented lambda DNA at a final concentration of 300 ng/µL was biotinlylated by incubating 5 U/µL of Klenow DNA polymerase containing 50 µM biotin-16-dUTP and 50 µM alpha-thio-dATP at 37 °C overnight in 1× NEBuffer 2. The DNA was then purified with the Monarch® PCR and DNA Cleanup Kit and stored in nuclease free water until usage as substrate in C5-MT43 *in vitro* activity assays (SI Appendix, Fig. S14B).

Xp12 and T4 Phage gDNA: Xp12 and T4 phage DNA were isolated from source and then purified according to the previously published procedure (12).

DNA Shearing. Lambda, Xp12, or T4gt DNA was sheared roughly to 1500 base pair length in 10 mM Tris buffer pH 8.0 with 1 mM EDTA using the COVARIS S2 instrument with the following settings: 5 % duty cycle, intensity 3 and 200 cycles/burst for 40 s. The sheared DNA was then purified using a Monarch® PCR and DNA Cleanup Kit and stored in nuclease free water.

***In Vitro* Enzyme Activity.**

C5-MT43 Activity Assay: Reactions were prepared in 20 µL volume containing 300 ng of biotinylated-lambda DNA substrate, 2.5 µM of purified C5-MT43, 100 mM buffer (as indicated in SI Appendix, Fig. S14B), 50 mM NaCl, 2.5 mM TCEP, and 160 µM SAM. Reactions were maintained at 37 ºC for ~ 17 h and quenched with 0.8 U of Proteinase K for 1 h at 37 ºC. In the meantime, streptavidin magnetic beads (NEB) were pre-treated by washing with 20 mM Tris pH 7.5, 500 mM NaCl, 1 mM EDTA, 6 % PEG and resuspended in 2 x the concentrations mentioned above. An equal volume of the resuspended streptavidin beads is added to the proteinase K-quenched enzyme reaction volume and the solution was gently shaken for 20 min to ensure DNA binding. The beads were then extracted using a magnetic rack (~ 5 min), washed 3 times with 20 mM Tris pH 7.5, 500 mM NaCl, 1 mM EDTA, and briefly allowed to dry (~ 2 min) before addition of the Nucleoside Digestion Mix.

TET43 Activity Assay: TET43 activity was tested in a reaction volume of 50 µL containing 50 mM MES pH 6.0, 70 mM NaCl, 5 mM 2OG, 80 µM Fe(II), ascorbate as indicated in Fig. 6 legend, 300 ng sheared Xp12 DNA, and 20 µM of purified enzyme. The reaction was maintained at 37 ºC for ~ 17 h and quenched with 0.8 U of Proteinase K for 1 h at 37 ºC. An equal volume of SPRI beads (Beckman Coulter) were added to the quenched reaction volume and the solution was gently shaken for 20 min to ensure DNA binding. The beads were then extracted using a magnetic rack (~ 5 min), washed 3 times with 70 % ethanol, and briefly allowed to dry (~ 2 min) before addition of the Nucleoside Digestion Mix.

In this reaction, the final concentration of Xp12 gDNA is 6 ng/μL, which corresponds to 144 μM (Xp12 is 64,272 bp in length). The GC content of this phage is 68.2 %, which translates to 6.3 μM 5mC. Among these 5mCs, 33.7 % is in the Gp5mC context, which results in 2.1 μM Gp5mC.

GT Activity Assay: Reactions were prepared in 50 μL volume containing 300 ng of sheared T4 phage DNA substrate, ~ 15 μM of purified GT in storage buffer, and 400 μM of UDP-glucose or derivatives. Salt and buffer conditions were as indicated in SI Appendix, Fig. S17. Reactions were maintained at 37 ºC for ~ 17 h and quenched with 0.8 U of Proteinase K for 1 h at 37 ºC. An equal volume of SPRI beads (Beckman Coulter) were added to the quenched reaction volume and the solution was gently shaken for 20 min to ensure DNA binding. The beads were then extracted using a magnetic rack (~ 5 min), washed 3 times with 70 % ethanol, and briefly allowed to dry (~ 2 min) before addition of the Nucleoside Digestion Mix.

LC-MS Analysis: LC-MS analysis of DNA digested into nucleosides using the Nucleoside Digestion Mix was performed on an Agilent 1200 Series LC-MS System equipped with a G1315D diode array detector and a 6120 single quadrupole mass detector in both positive (+ESI) and negative (-ESI) electrospray ionization modes. LC was performed on a Waters Atlantis T3 column (4.6 × 150 mm, 3 μm) with a gradient mobile phase consisting of aqueous ammonium acetate (10 mM, pH 4.5) and methanol. The relative abundance of each nucleoside was determined by dividing the UV absorbance by the corresponding extinction coefficient. The relative abundance of 5-

GlcαmC, 5-GlcβmC, 419 u (+Glc), and 460 u (+GlcNAc) species were estimated using the extinction coefficient of 5hmC at 273 nm.

LC-MS/MS analysis was performed by injecting digested DNAs 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 (+ESI). Ultra-high-performance liquid chromatography (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). MS 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: C $[M+H]^+$ at m/z 228 →112, 5mC $[M+H]^+$ at m/z 242→126, 5hmC $[M+H]^+$ at m/z 258→142, 5fC $[M+H]^+$ at m/z 256→140, and 5caC $[M+Na]^+$ at m/z 294→178. T $[M+Na]^+$ at m/z 265→149, 5hmU $[M+Na]^+$ at m/z 281→165, 5fU $[M+Na]^+$ at m/z 279→163, and 5caU $[M+Na]^+$ at m/z 295→179. External calibration curves with known amounts of the canonical and non-canonical nucleosides were used to calculate the ratios of individual nucleosides within the samples analyzed.

SI References

1. A. C. Gregory *et al.*, Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109-1123 e1114 (2019).
2. D. Páez-Espino *et al.*, IMG/VR: a database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res.* **45**, D457-D465 (2017).
3. A. P. Arkin *et al.*, KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* **36**, 566-569 (2018).
4. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845-858 (2015).
5. J. P. Fürste *et al.*, Molecular cloning of the plasmid RP4 primase region in a multi-host-range tacP expression vector. *Gene* **48**, 119-131 (1986).
6. R. Vaisvila *et al.*, EM-seq: Detection of DNA Methylation at Single Base Resolution from Picograms of DNA. *BioRxiv* https://doi.org/10.1101/2019.12.20.884692 (2019).
7. Z. Sun *et al.*, Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome Res.* 10.1101/gr.265306.120 (2021).
8. F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).
9. A. Akalin *et al.*, MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87 (2012).
10. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188-1190 (2004).
11. S. C. Gill, P. H. von Hippel, Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **182**, 319-326 (1989).
12. Y. J. Lee, P. R. Weigele, Detection of modified bases in bacteriophage genomic DNA. *Methods Mol. Biol.* **2198**, 53-66 (2021).

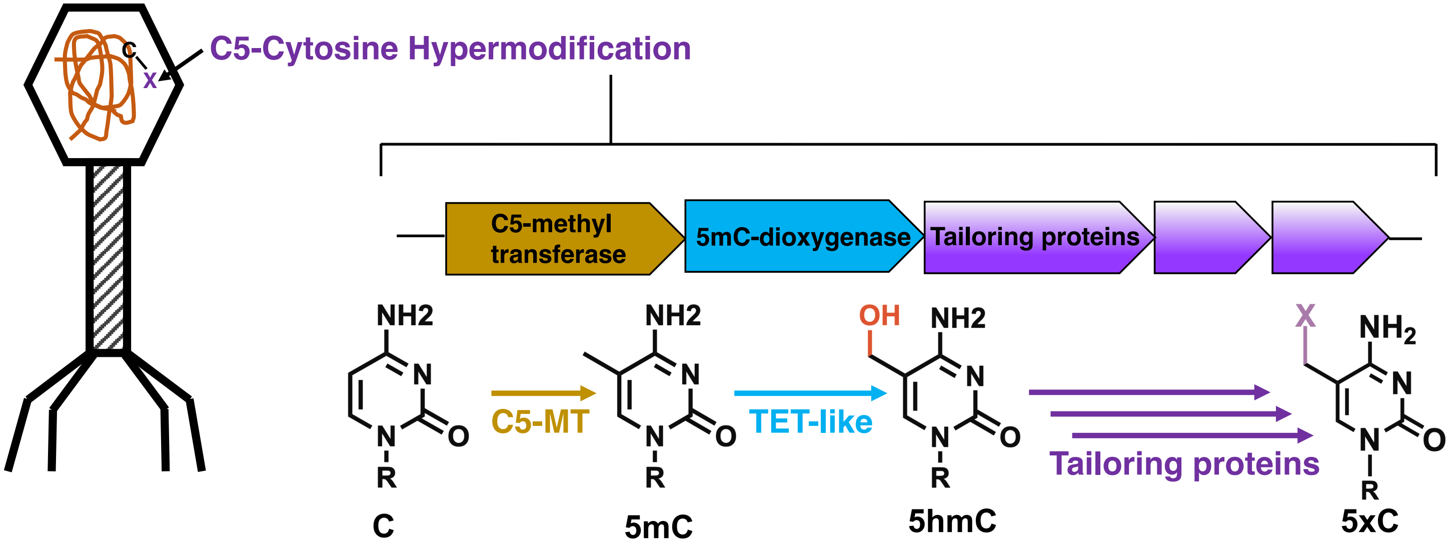

Fig. S1. A schematic representation of TET-encoding biosynthetic gene cluster specifying complex cytosine modifications on phage DNA.

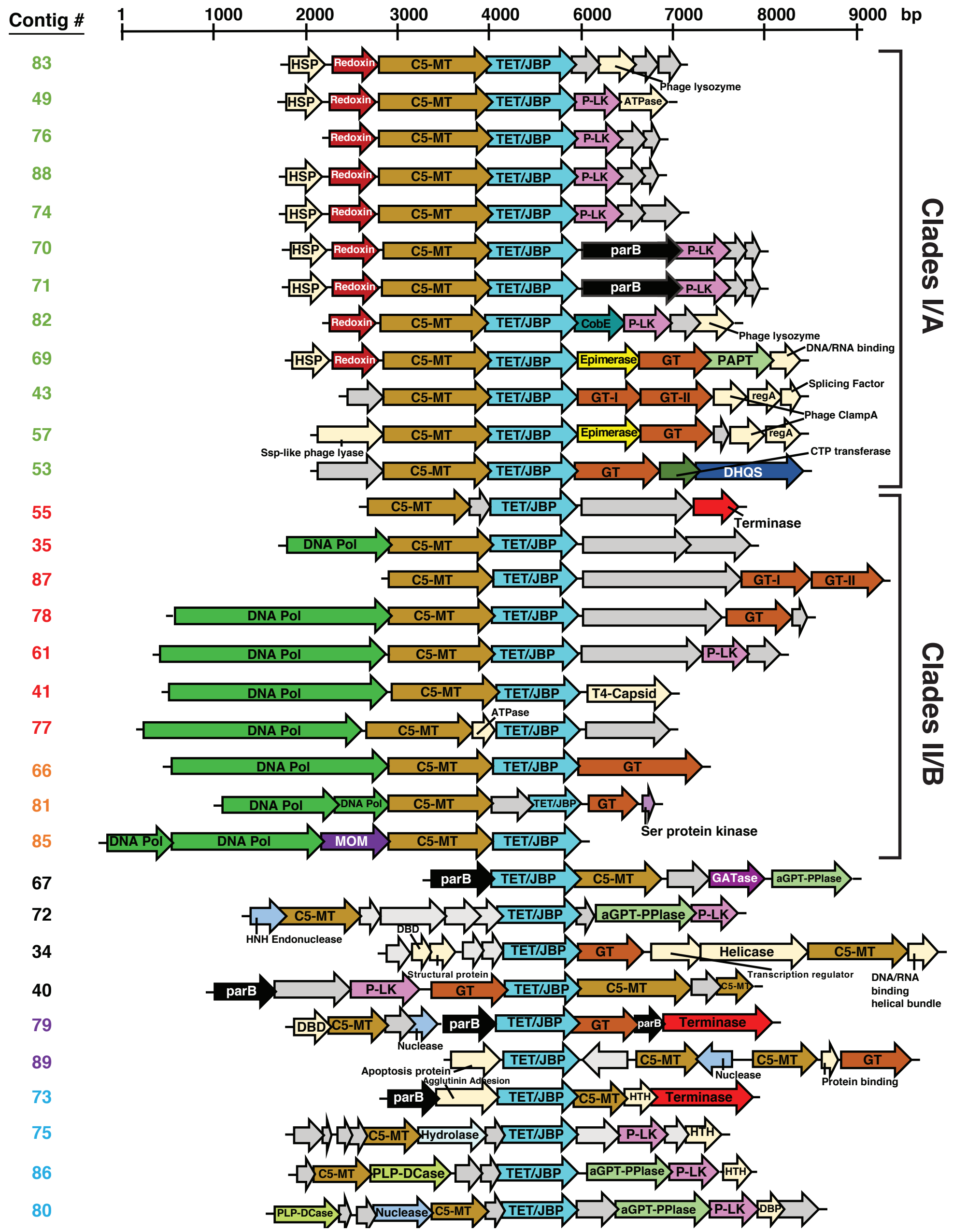
Contig #
1 1000 2000 3000 4000 6000 7000 8000 9000 bp
83 HSP Redoxin C5-MT TET/JBP Phage lysozyme
49 HSP Redoxin C5-MT TET/JBP P-LK ATPase
76 Redoxin C5-MT TET/JBP P-LK
88 HSP Redoxin C5-MT TET/JBP P-LK
74 HSP Redoxin C5-MT TET/JBP P-LK
70 HSP Redoxin C5-MT TET/JBP parB P-LK
71 HSP Redoxin C5-MT TET/JBP parB P-LK
82 Redoxin C5-MT TET/JBP CobE P-LK Phage lysozyme
69 HSP Redoxin C5-MT TET/JBP Epimerase GT PAPT DNA/RNA binding
43 C5-MT TET/JBP GT-I GT-II regA Splicing Factor Phage ClampA
57 C5-MT TET/JBP Epimerase GT regA
Ssp-like phage lyase
53 C5-MT TET/JBP GT DHQS CTP transferase
Clades I/A
55 C5-MT TET/JBP Terminase
35 DNA Pol C5-MT TET/JBP
87 C5-MT TET/JBP GT-I GT-II
78 DNA Pol C5-MT TET/JBP GT
61 DNA Pol C5-MT TET/JBP P-LK
41 DNA Pol C5-MT TET/JBP T4-Capsid
77 DNA Pol C5-MT ATPase TET/JBP
66 DNA Pol C5-MT TET/JBP GT
81 DNA Pol DNA Pol C5-MT TET/JBP GT Ser protein kinase
85 DNA Pol DNA Pol MOM C5-MT TET/JBP
Clades II/B
67 parB TET/JBP C5-MT GATase aGPT-PPlase
72 C5-MT TET/JBP aGPT-PPlase P-LK HNH Endonuclease DBD
34 TET/JBP GT Helicase C5-MT Structural protein Transcription regulator DNA/RNA binding helical bundle
40 parB P-LK GT TET/JBP C5-MT C5-MT
79 DBD C5-MT parB TET/JBP GT parB Terminase Nuclease
89 TET/JBP C5-MT C5-MT GT Apoptosis protein Nuclease Protein binding
73 parB TET/JBP C5-MT HTH Terminase Agglutinin Adhesion
75 C5-MT Hydrolase TET/JBP P-LK HTH
86 C5-MT PLP-DCase TET/JBP aGPT-PPlase P-LK HTH
80 PLP-DCase Nuclease C5-MT TET/JBP aGPT-PPlase P-LK DBP

Fig. S2. C5-MT/TET- encoding contigs from viral metagenomic origin. A number assignment is shown to the left of each contig and a more detailed description is in Table S1 and *SI Materials and Methods*. Genes with a beige color assignment are phage regulatory proteins. Genes in grey correspond to proteins of undefined function. The predicted enzymatic functionalities of the rest of the colored genes are as labeled. Abbreviations are as following: HSP = Heat shock protein; regA = phage endoribonuclease translational repressor; DBD = DNA binding domain; DBP = DNA binding protein; PLP Dcase = PLP-dependent decarboxylase, GATase = glutamine amidotransferase; HTH= helix-turn-helix domain; MOM = Adenine modification enzyme MOM; CobE = CobE/GbiG C-terminal domain-like. Other abbreviations have been defined in the main text.

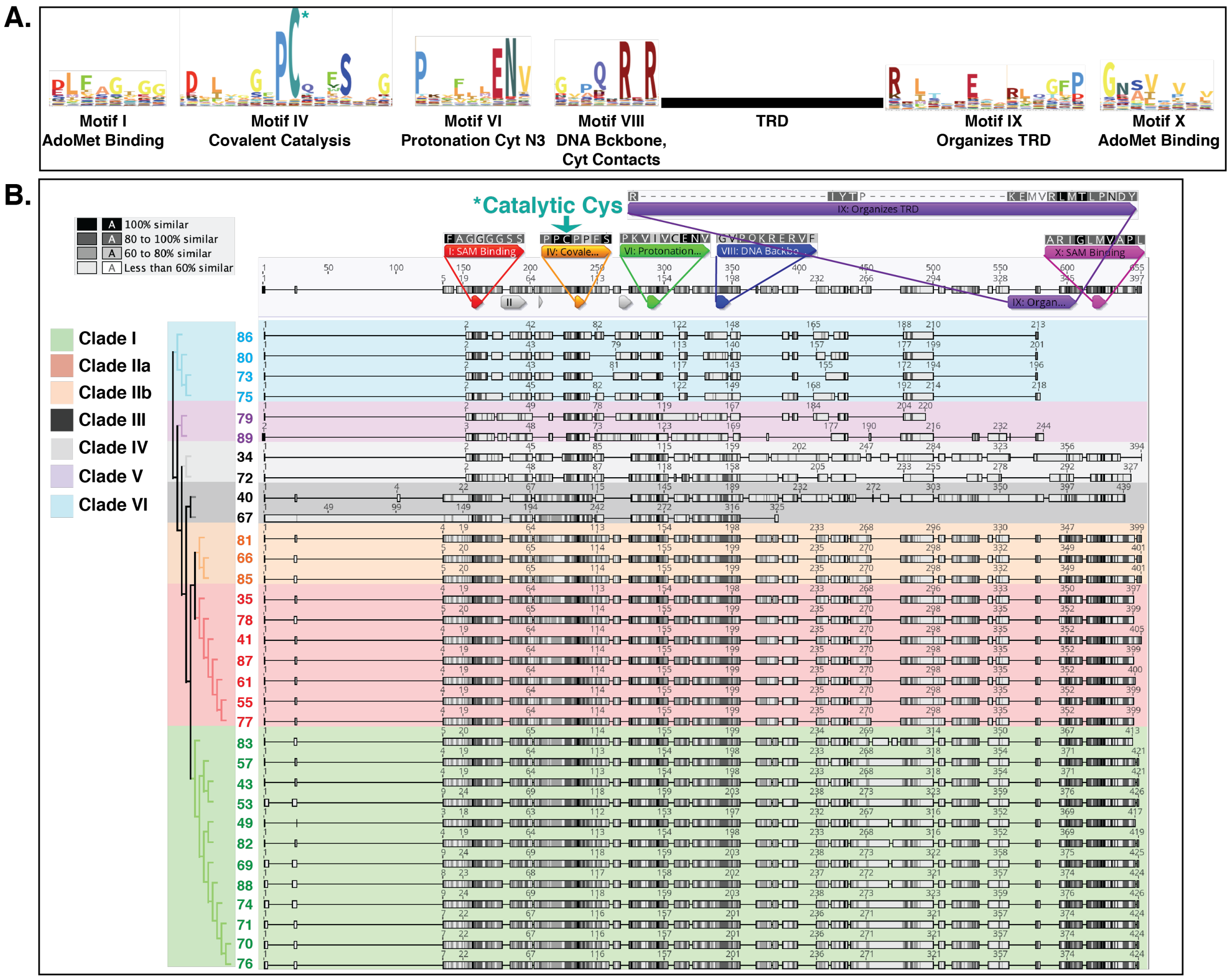

Fig. S3. *(A)* Sequence logos of conserved catalytic motifs in C5-MTs as obtained from EMBL-EBI Pfam HMM logo tool (DNA_methylase (PF00145)). The name and the proposed function of each of the motifs are indicated below the logo. Abbreviations are as following: TRD = target recognition domain; Cyt = cytosine. *(B)* Sequence alignment of the studied phage C5-MTs highlighting domain similarity between the various enzymes.

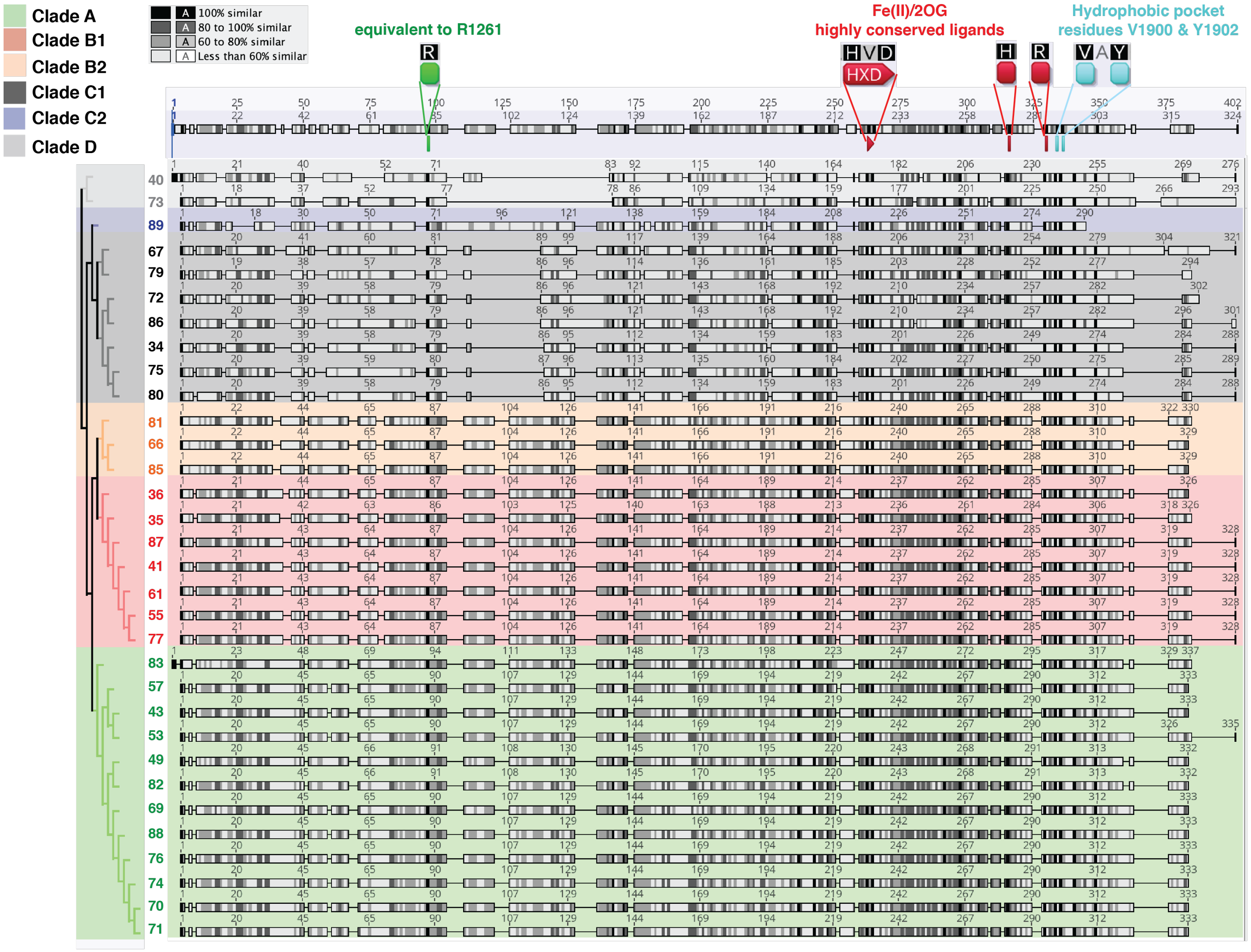

Fig. S4. Sequence alignment of the studied phage TETs highlighting domain similarity between the various enzymes. On top, we show highly conserved amino acids equivalent to human TET2 residues, which have been shown to be important in the catalytic function of the enzyme.

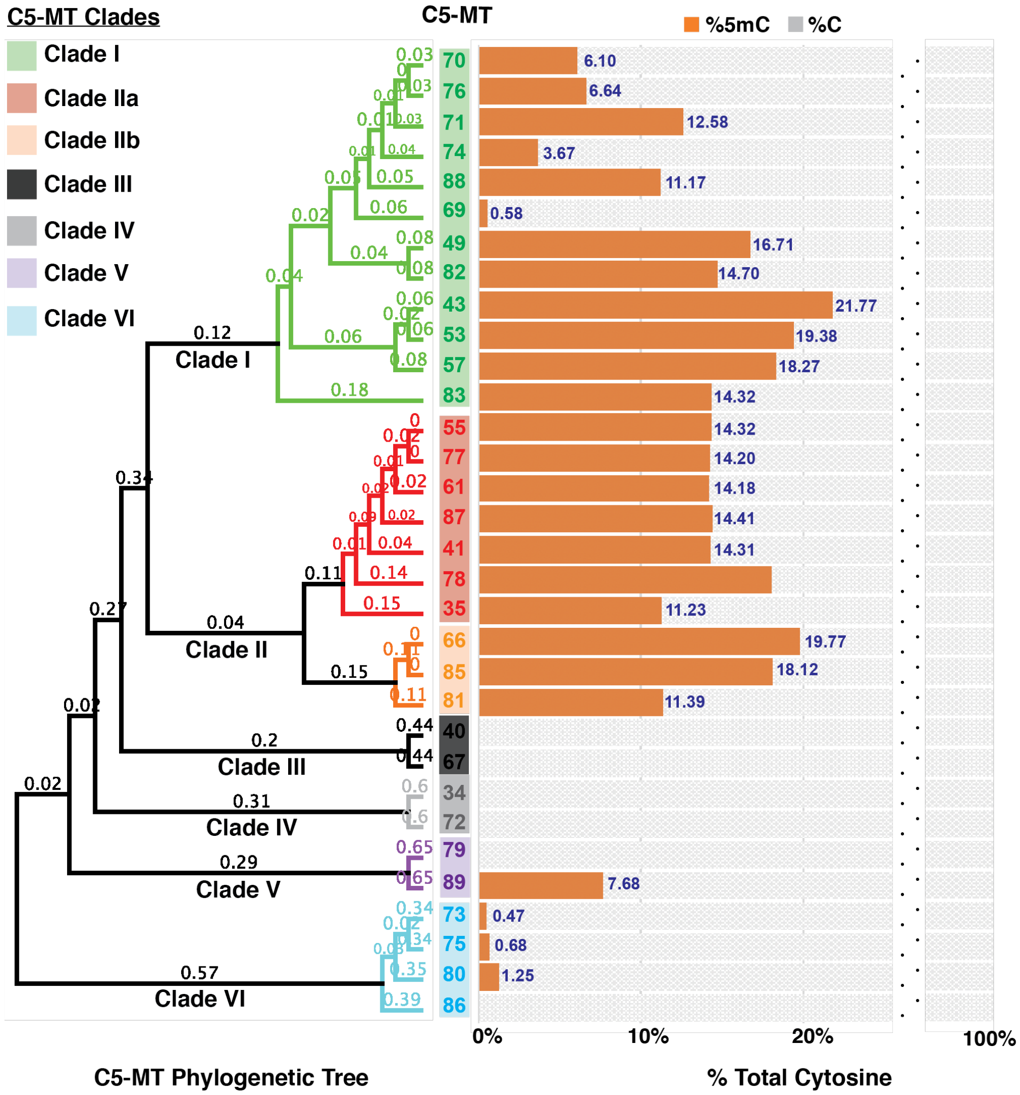

Fig. S5. Phylogenetic analysis of C5-MTs by their amino acid sequences (left) and LC-MS/MS analysis of % 5mC formed *in vivo* on *E. coli* gDNA per total cytosine upon expression of C5-MT from a specific contig (right).

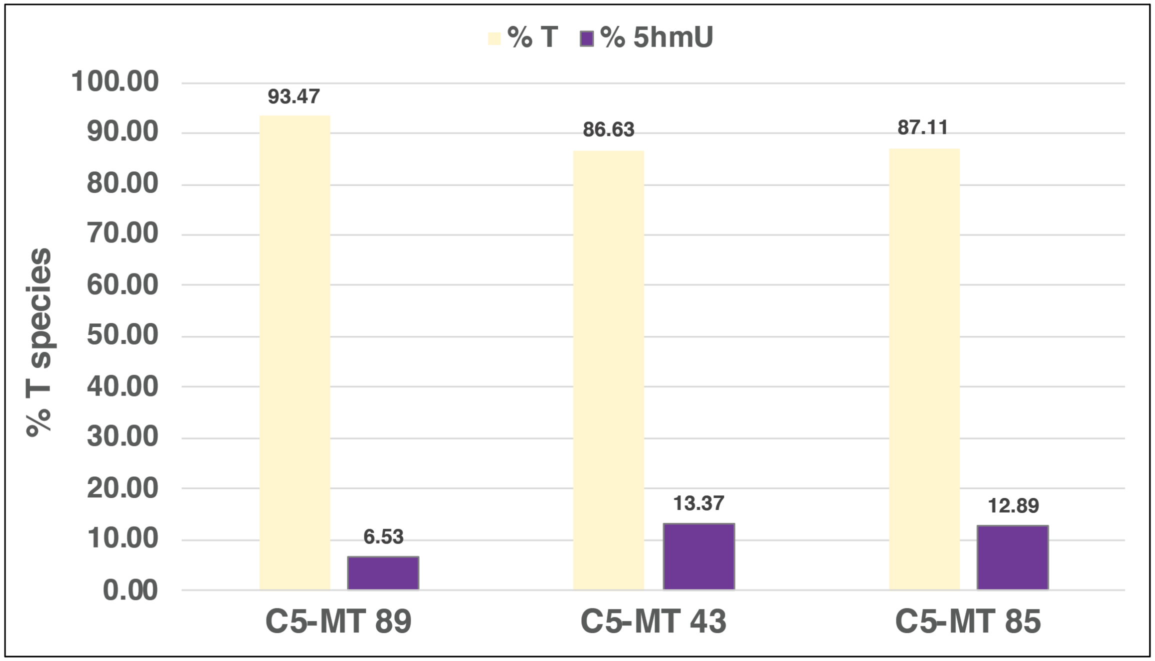

Fig. S6. LC-MS/MS data showing T hydroxylation (formation of 5hmU) *in vivo* on *E. coli* gDNA upon expression of TET89 (clade C2) with the indicated C5-MTs.

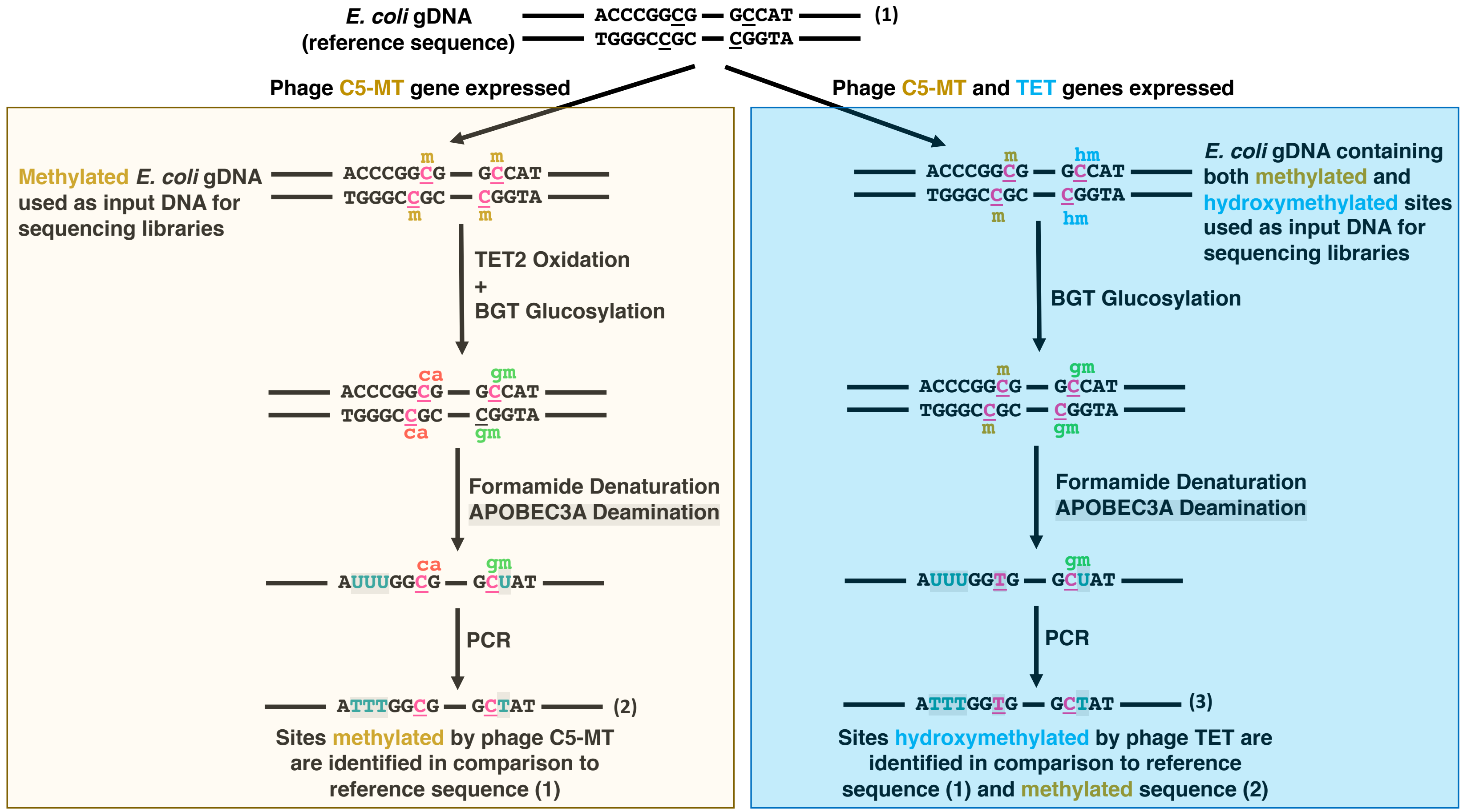

Fig. S7. Scheme describing methods used for mapping of sites methylated by phage C5-MT (light brown box on the left) or sites hydroxymethylated by TET (blue box on the right). For identification of methylation sites (light brown box), three enzymes are used: TET2, BGT, and APOBEC3A, as shown in the first two enzymatic conversion steps. Following PCR amplification and Illumina-based sequencing, 5caC and 5-GlcβmC (labeled as gmC in the scheme) are read as C, while unmethylated C is sequenced as T (sequence **(2)**). Comparison of sequence **(2)** to **(1)** allows identification of sites methylated by phage C5-MT. For the identification of sites hydroxymethylated by phage TET, only BGT and APOBEC3A are used in this method. 5mC sites not converted by phage TET are ultimately deaminated by APOBEC3A into T while gmC is read as C during sequencing. Comparison of sequence **(3)** to **(2)** and **(1)** allows identification of sites hydroxylated by phage TET.

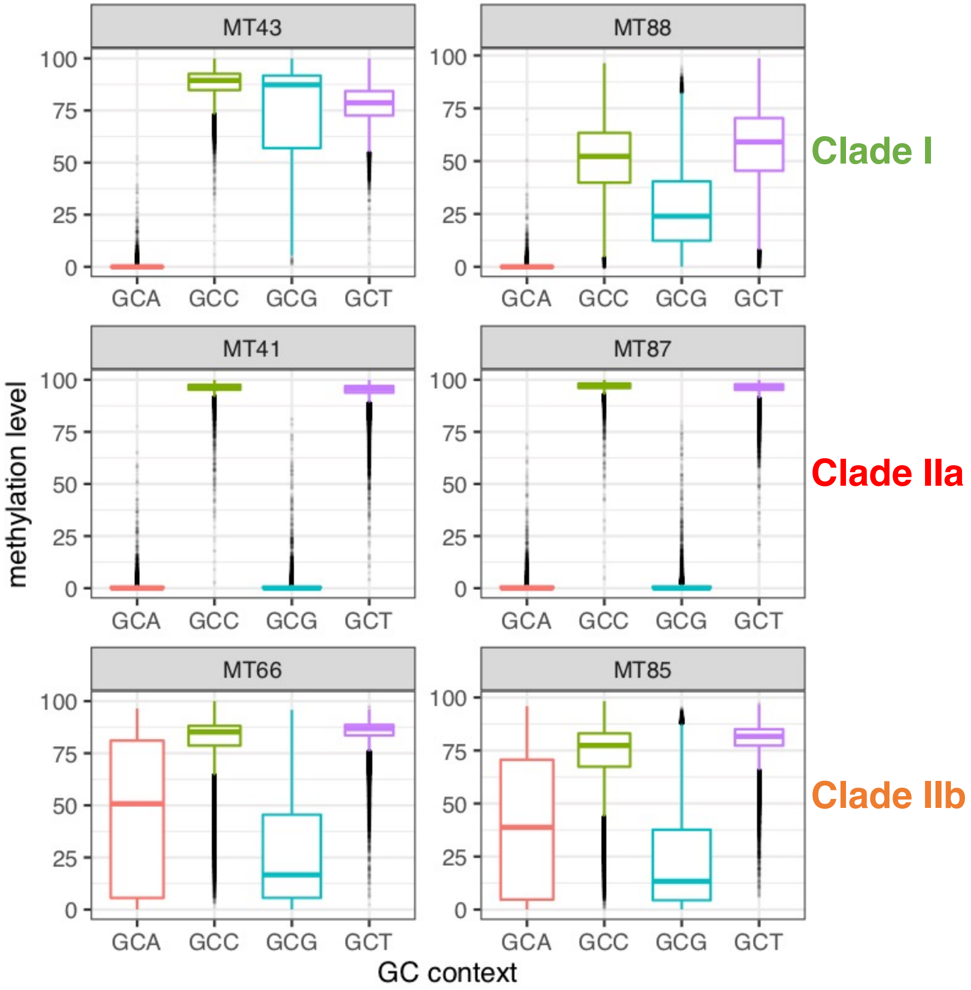

Fig. S8. Box plots showing 5mC levels (percentage of methylated copies among all the sequenced copies) per GCN sites on *E. coli* gDNA by C5-MT. Experimental details pertaining sequencing and generation of box plots are described in SI Appendix Materials and Methods.

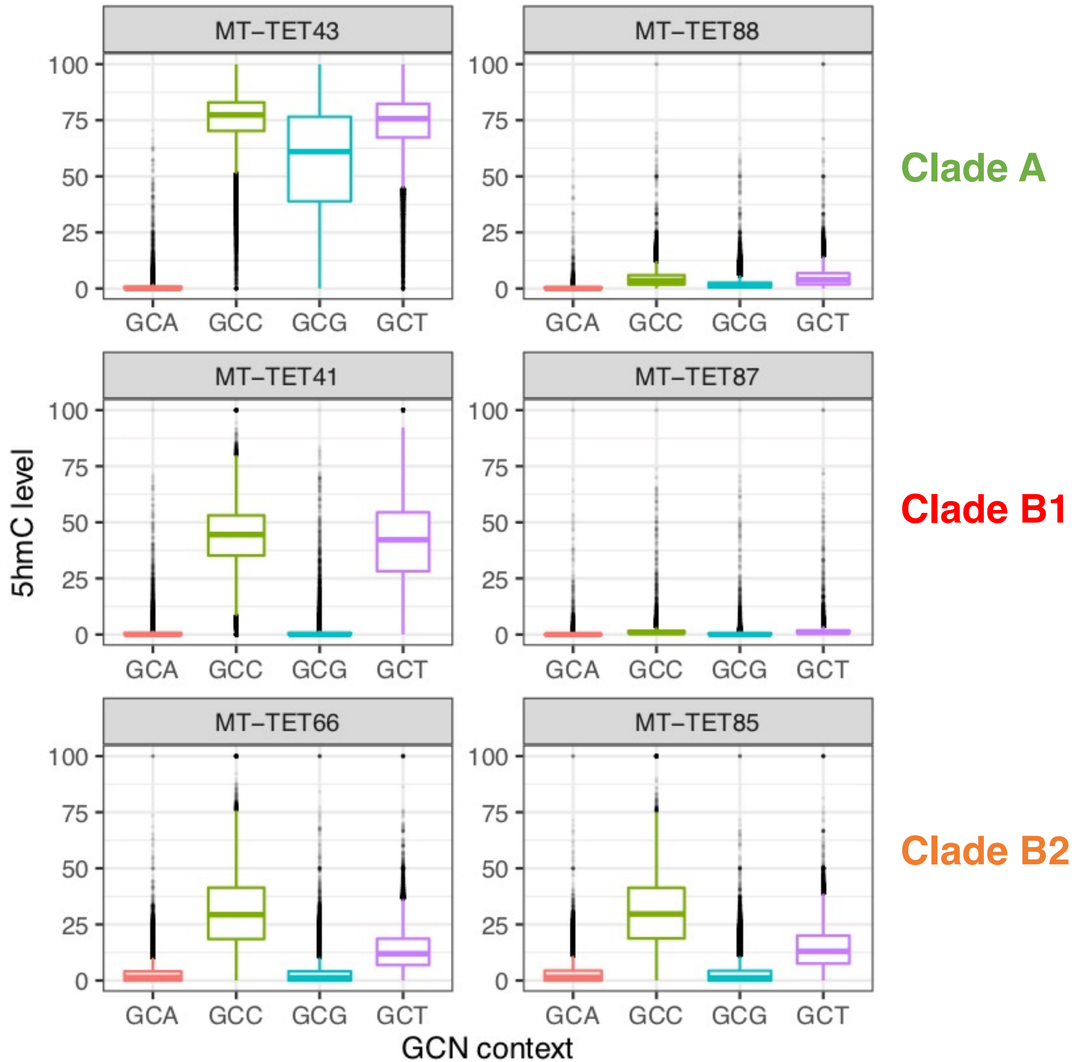

Fig. S9. Box plots showing 5hmC levels (percentage of hydroxymethylated copies among all the sequenced copies) per GCN sites on *E. coli* gDNA by TET. Experimental details related to sequencing and generation of box plots are described in SI Appendix Materials and Methods.

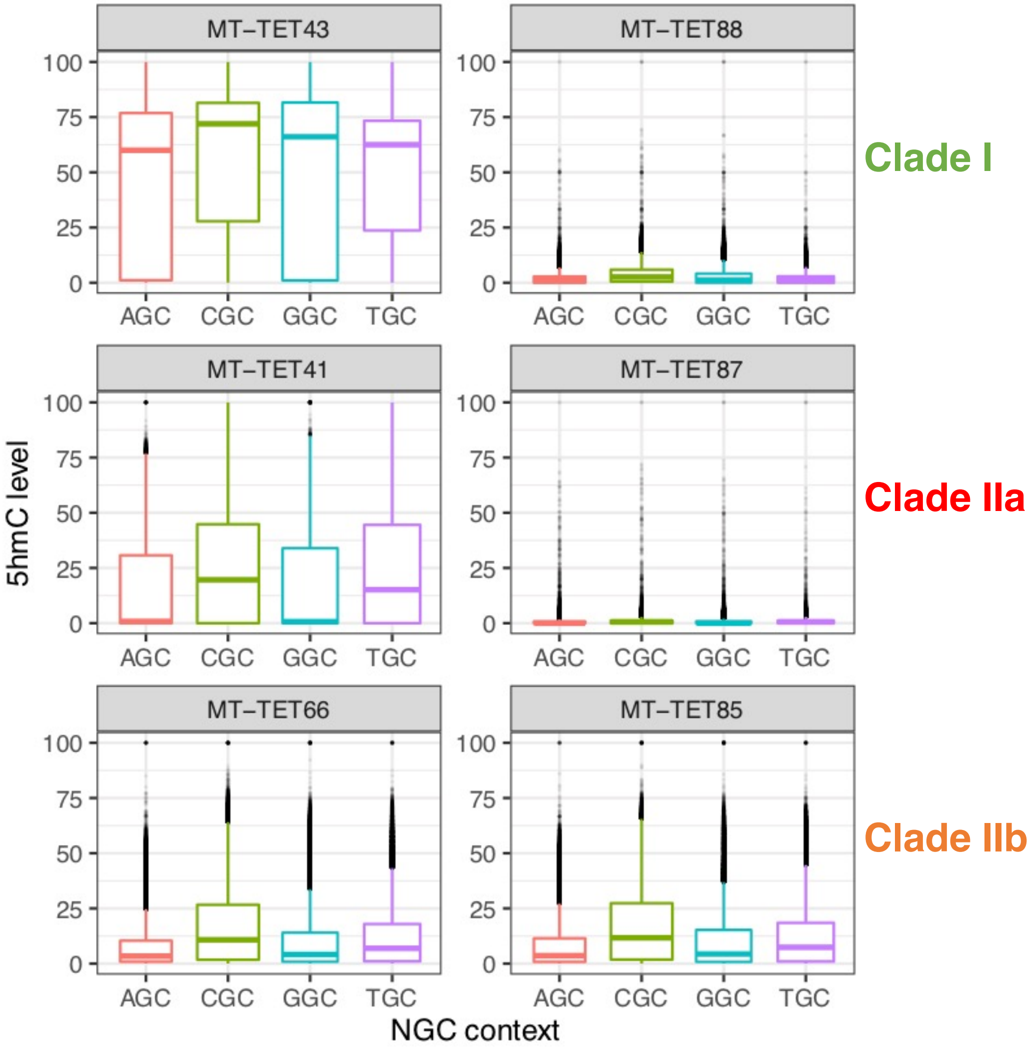

Fig. S10. Box plots showing 5hmC levels (percentage of hydroxymethylated copies among all the sequenced copies) per NGC sites on *E. coli* gDNA by TET. Experimental details related to sequencing and generation of box plots are described in SI Appendix Materials and Methods.

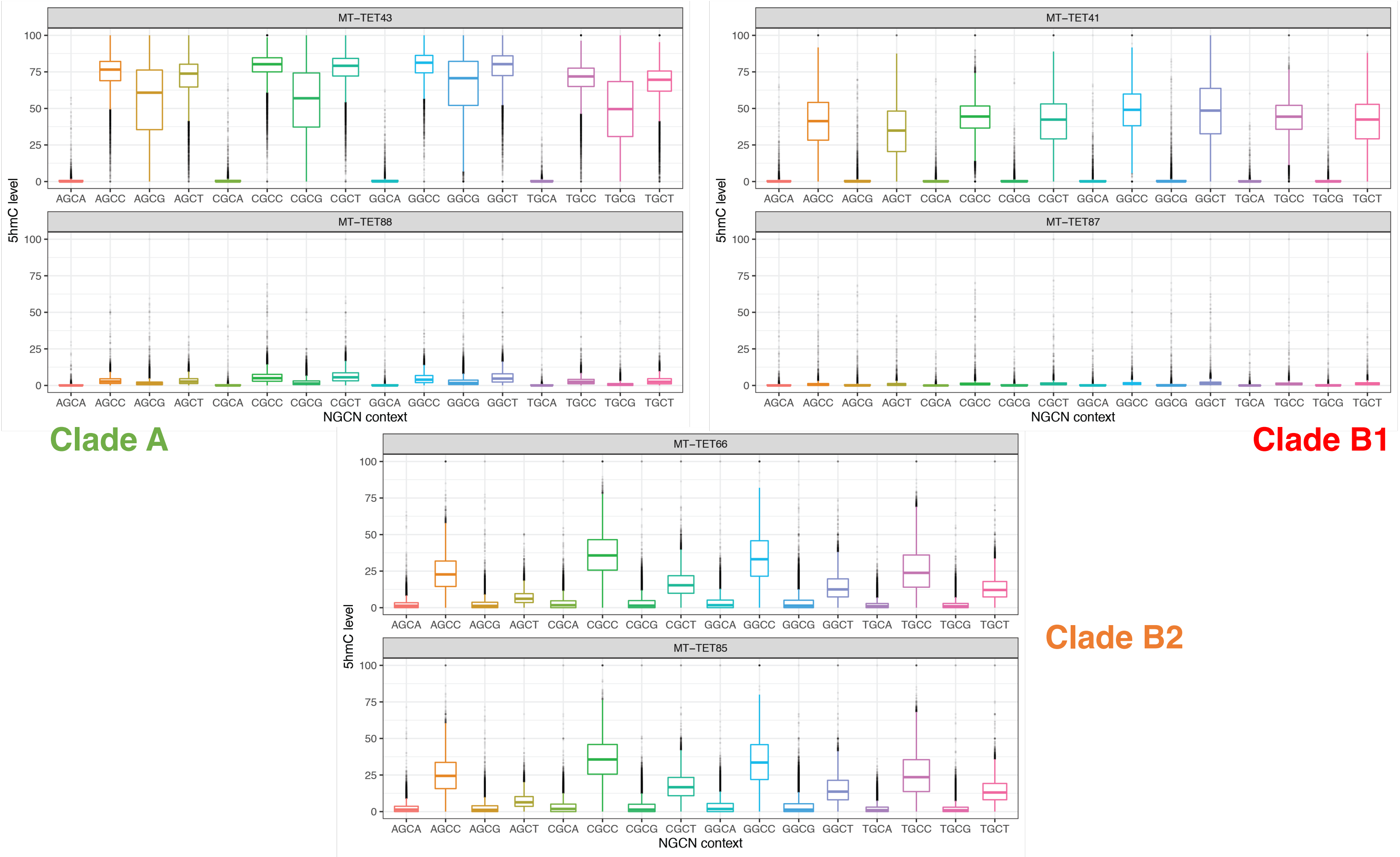

Fig. S11. Box plots showing 5hmC levels (percentage of hydroxymethylated copies among all the sequenced copies) per NGCN sites on *E. coli* gDNA by TET. Experimental details related to sequencing and generation of box plots are described in SI Appendix Materials and Methods.

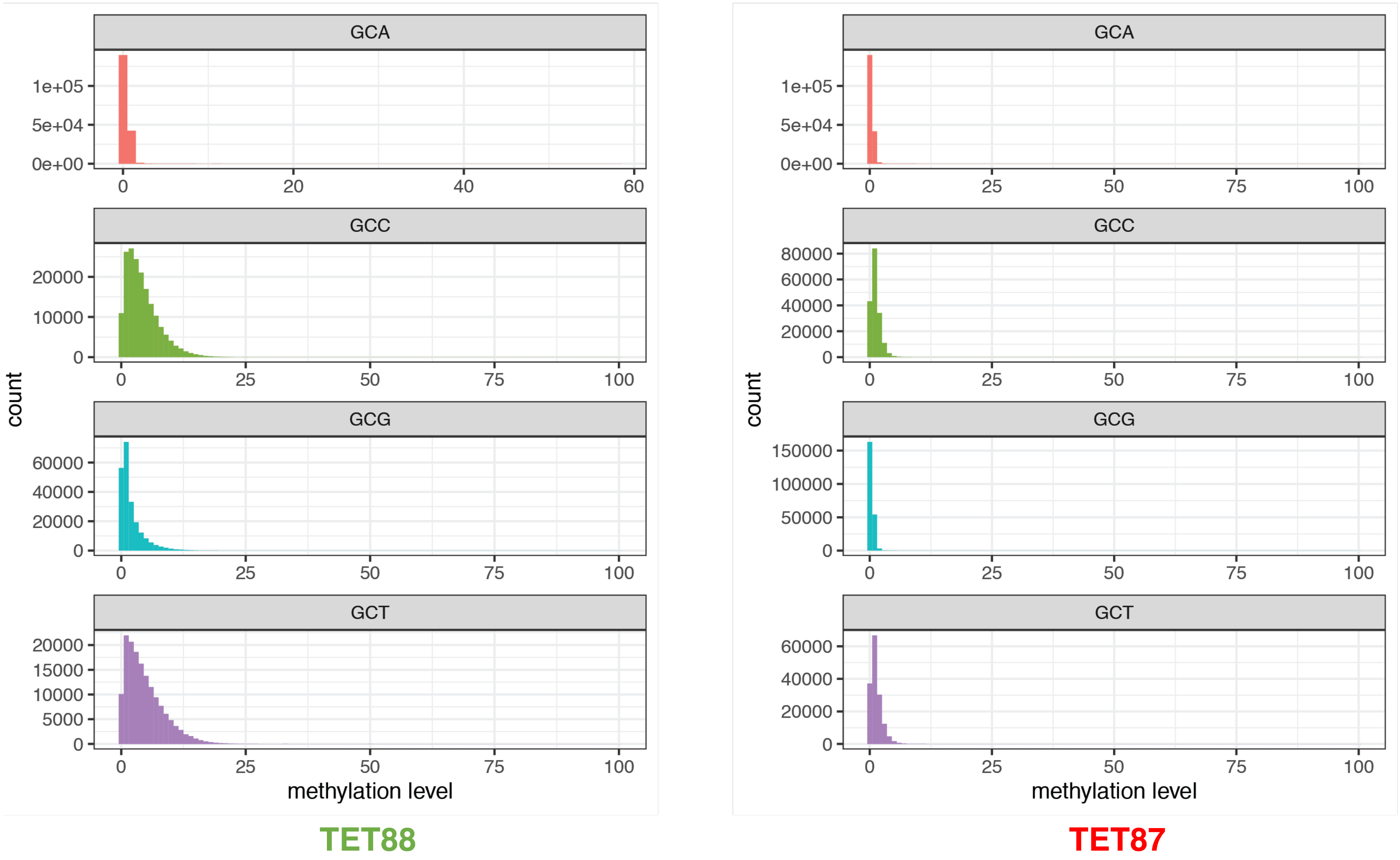

Fig. S12 A histogram plot showing the number of GCN sites converted to GhmCN at different hydroxymethylation levels by TET88 (left) and TET87 (right).

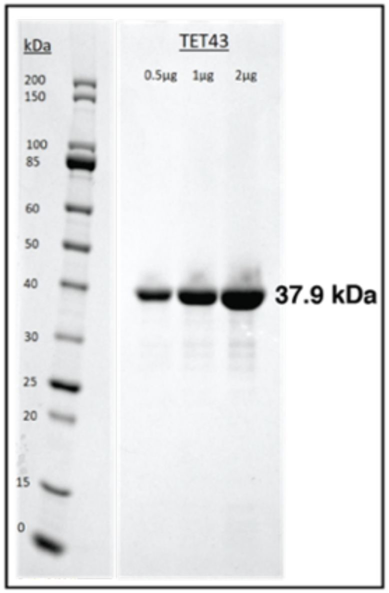

Fig. S13. Coomassie-stained SDS-PAGE gel showing TET43 purity. The protein identity was confirmed by Peptide Sequencing using an Orbitrap ESI with LC-MS/MS.

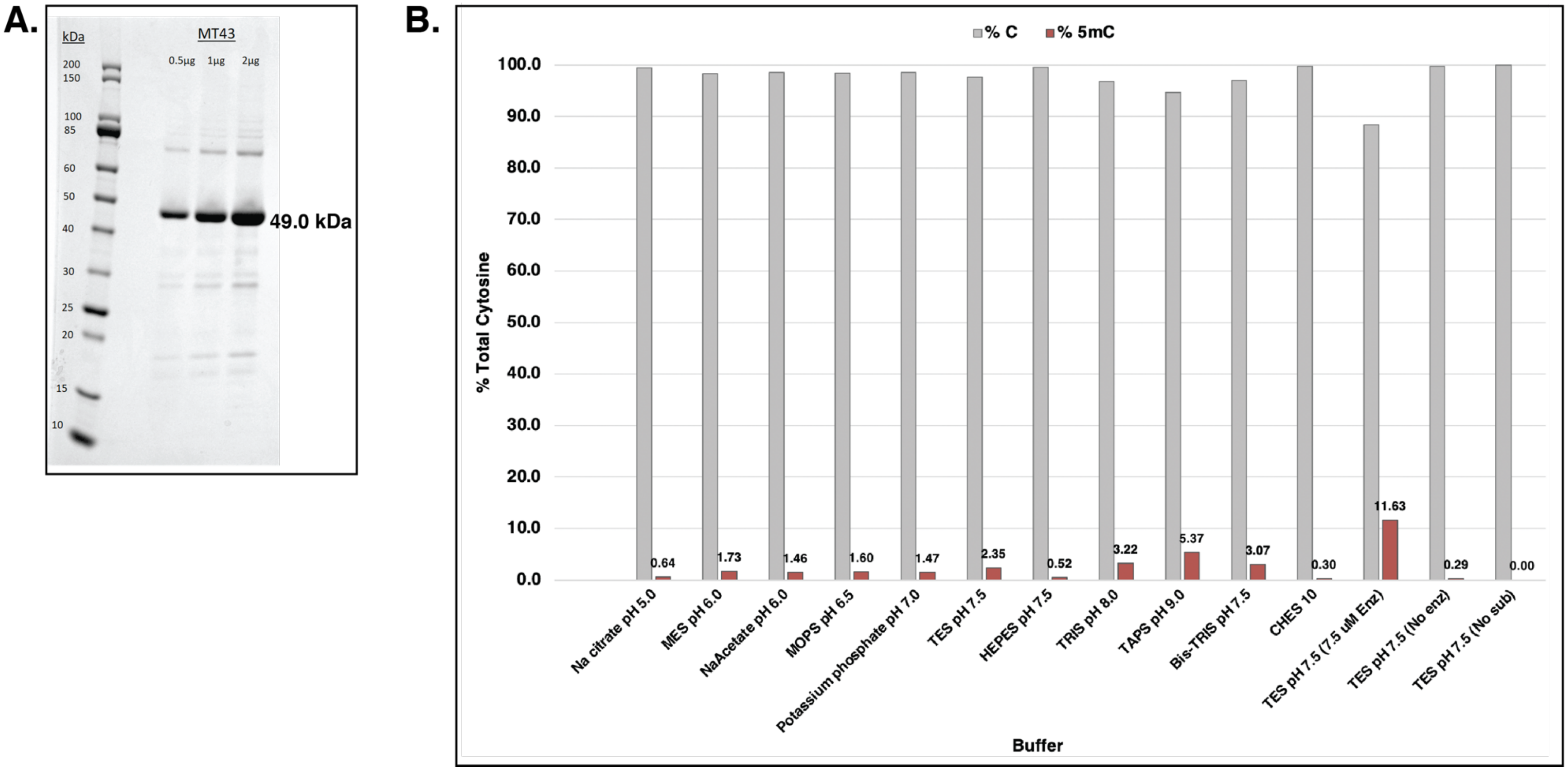

Fig. S14. *(A)* Coomassie-stained SDS-PAGE gel showing C5-MT43 purity. The protein identity was confirmed by Peptide Sequencing using an Orbitrap ESI with LC-MS/MS. *(B)* % total cytosines on lambda DNA (15 ng/µL) (0.77 µM cytosine) methylated by C5-MT43 (2.5 µM) as detected by LC-MS/MS under different *in vitro* conditions detailed in the Methods section. The reactions were incubated for ~ 17 h at 37 ºC.

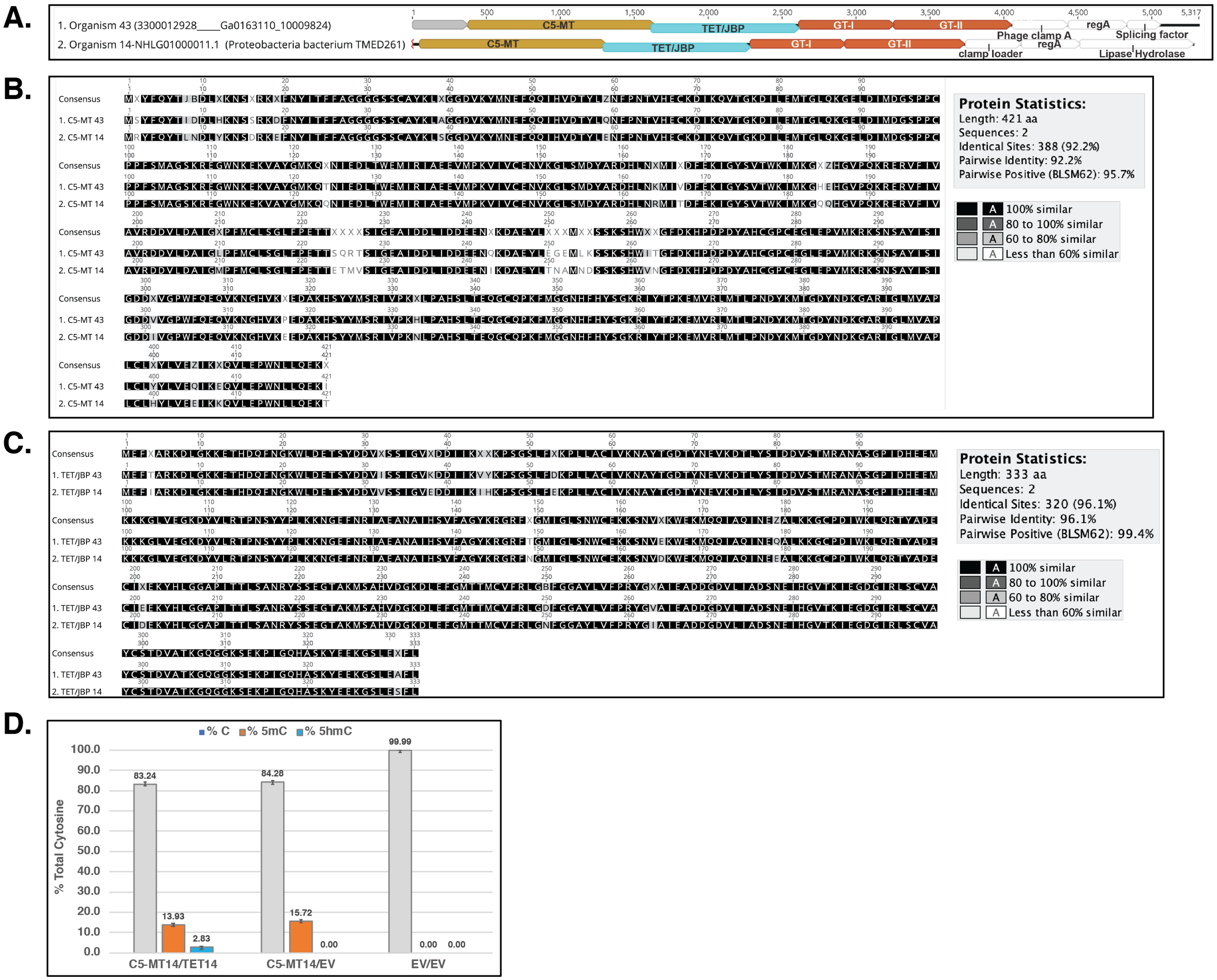

Fig. S15. *(A)* Gene architecture of contigs 43 and 14. *(B)* Sequence alignment of C5-MT43 and 14. *(C)* Sequence alignment of TET43 and 14. and *(D)* LC-MS/MS analysis of the *in vivo* activities of the co-expressed C5-MT and TET14 on *E. coli* gDNA.

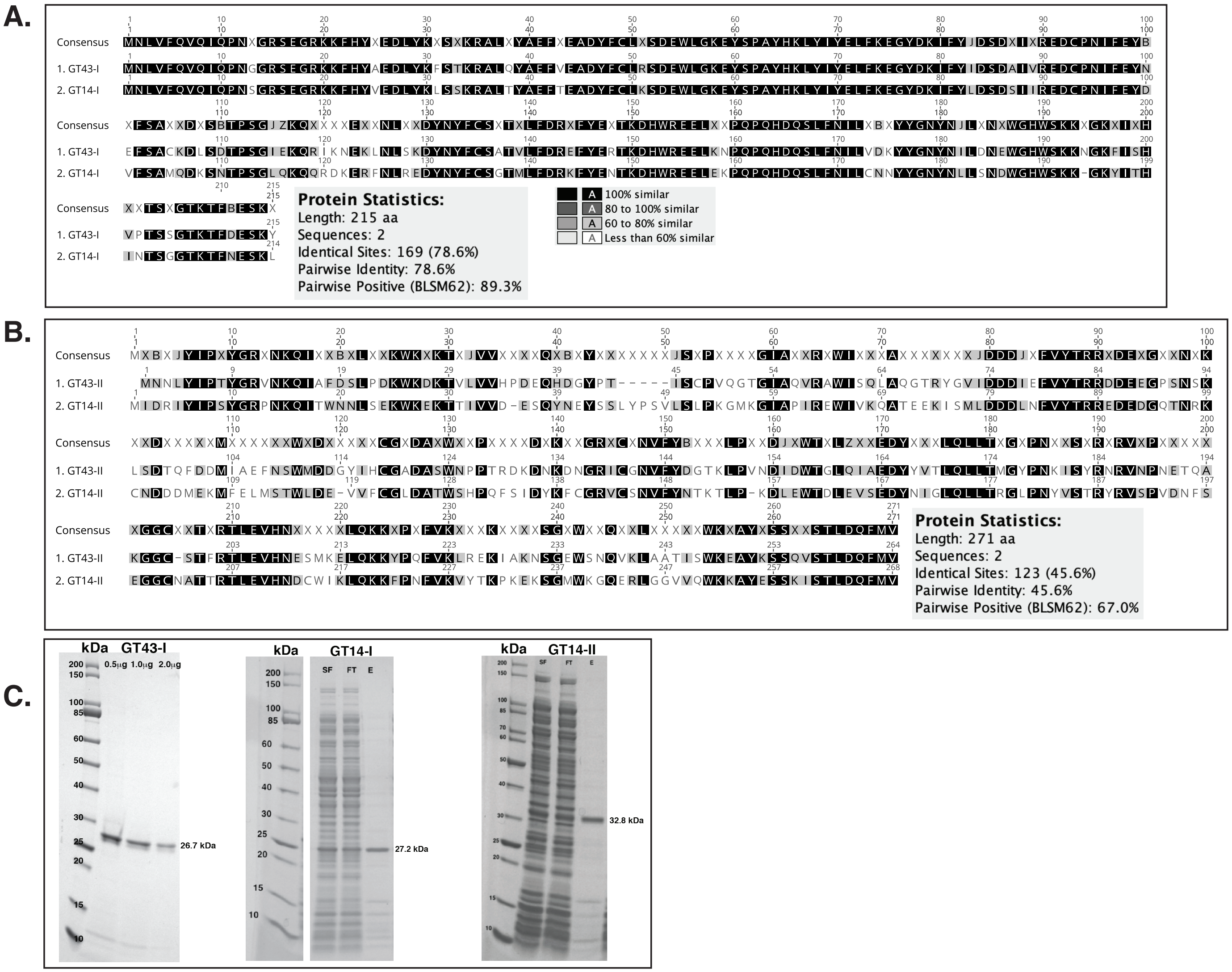
A.
Consensus MNLVFQVQIQPNXGRSEGRKKFHYXEDLYKXSXKRALXYAEFXEADYFCLXSDEWLGKEYSPAYHKLYIYELFKEGYDKIFYJDSDXIXREDCPNIFEYB
1. GT43-I MNLVFQVQIQPNGGRSEGRKKFHYAEDLYKFSTKRALQYAEFVEADYFCLRSDEWLGKEYSPAYHKLYIYELFKEGYDKIFYIDSDAIVREDCPNIFEYN
2. GT14-I MNLVFQVQIQPNSGRSEGRKKFHYVEDLYKLSSKRALTYAEFTEADYFCLKSDEWLGKEYSPAYHKLYIYELFKEGYDKIFYLDSDSIIREDCPNIFEYD
Consensus XFSAXXDXSBTPSGJZKQXXXXEXXNLXXDYNYFCSXTXLFDRXFYEXTKDHWREELXXPQPQHDQSLFNILXBXYYGNYNJLXNXWGHWSKKXGKXIXH
1. GT43-I EFSACKDLSDTPSGIEKQRIKNEKLNLSKDYNYFCSATVLFDREFYERTKDHWREELKNPQPQHDQSLFNILVDKYYGNYNILDNEWGHWSKKNGKFISH
2. GT14-I VFSAMQDKSNTPSGLQKQQRDKERFNLREDYNYFCSGTMLFDRKFYENTKDHWREELEKPQPQHDQSLFNILCNNYYGNYNLLSNDWGHWSKK-GKYITH
Consensus XXTSXGTKTFBESKX
1. GT43-I VPTSSGTKTFDESKY
2. GT14-I INTSGGTKTFNESKL
Protein Statistics:
Length: 215 aa
Sequences: 2
Identical Sites: 169 (78.6%)
Pairwise Identity: 78.6%
Pairwise Positive (BLSM62): 89.3%
A 100% similar
A 80 to 100% similar
A 60 to 80% similar
A Less than 60% similar
B.
Consensus MXBXJYIPXYGRXNKQIXXBXLXXKWKXKTXJVVXXXXQXBXYXXXXXXXJSXPXXXXGIAXXRXWIXXXAXXXXXXXJDDDJXFVYTRRXDEXGXXNXK
1. GT43-II MNNLYIPTYGRVNKQIAFDSLPDKWKDKTVLVVHPDEQHDGYPT-----ISCPVQGTGIAQVRAWISQLAQGTRYGVIDDDIEFVYTRRDDEEGPSNSK
2. GT14-II MIDRIYIPSYGRPNKQITWNNLSEKWKEKTTIVVD-ESQYNEYSSLYPSVLSLPKGMKGIAPIREWIVKQATEEKISMLDDDLNFVYTRREDEDGQTNRK
Consensus XXDXXXXXMXXXXXXWXDXXXXXCGXDAXWXXPXXXXDXKXXGRXCXNVFYBXXXLPXXDJXWTXLZXXEDYXXXLQLLTXGXPNXXSXRXRVXPXXXXX
1. GT43-II LSDTQFDDMIAEFNSWMDDGYIHCGADASWNPPTRDKDNKDNGRICGNVFYDGTKLPVNDIDWTGLQIAEDYYVTLQLLTMGYPNKISYRNRVNPNETQA
2. GT14-II CNDDDMEKMFELMSTWLDE-VVFCGLDATWSHPQFSIDYKFCGRVCSNVFYNTKTLP-KDLEWTDLEVSEDYNIGLQLLTRGLPNYVSTRYRVSPVDNFS
Consensus XGGCXXTXRTLEVHNXXXXXLQKKXPXFVKXXXKXXXXSGXWXXQXXLXXXXXWKXAYXSSXXSTLDQFMV
1. GT43-II KGGC-STFRTLEVHNESMKELQKKYPQFVKLREKIAKNSGEWSNQVKLAATISWKEAYKSSQVSTLDQFMV
2. GT14-II EGGCNATTRTLEVHNDCWIKLQKKFPNFVKVYTKPKEKSGMWKGQERLGGVVQWKKAYESSKISTLDQFMV
Protein Statistics:
Length: 271 aa
Sequences: 2
Identical Sites: 123 (45.6%)
Pairwise Identity: 45.6%
Pairwise Positive (BLSM62): 67.0%
C.
kDa
GT43-I
0.5μg 1.0μg 2.0μg
200
150
100
85
60
50
40
30
25
20
15
10
26.7 kDa
kDa
GT14-I
SF FT E
200
150
100
85
60
50
40
30
25
20
15
10
27.2 kDa
kDa
GT14-II
SF FT E
200
150
100
85
70
60
50
40
30
25
20
15
10
32.8 kDa

Fig. S16. Sequence alignments of *(A)* GT-I 43 and 14 and *(B)* GT-II 43 and 14. *(C)* Coomassie-stained SDS-PAGE gel showing GT43-I, GT14-I, and GT14-II purity. GT43-I was ÄKTA-purified while GT14-I and II were purified using NEBExpress® Ni Spin Columns. SF = soluble fraction; FT = flow through; E = elution. Protein identities were confirmed by Peptide Sequencing using an Orbitrap ESI with LC-MS/MS.

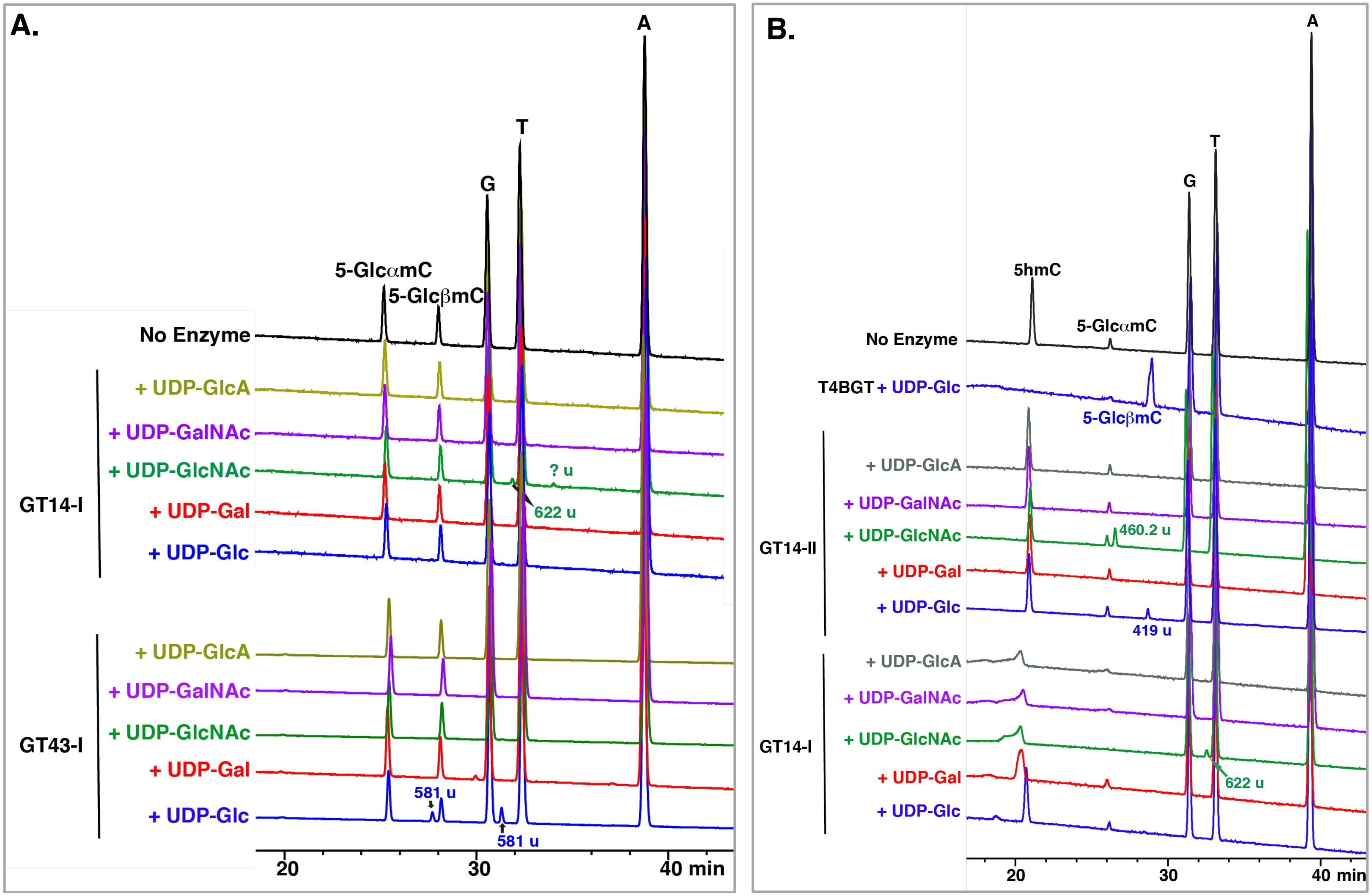

Fig. S17. LC-MS analysis showing the *in vitro* activity of *(A)* purified GT14-I and GT43-I with T4 phage *wt* DNA. *(B)* purified GT14-I and II with T4 phage *gt-/-* DNA. In *(A)* and *(B)*, [GT] = ~ 15 μM and [DNA] = 6 ng/μL (3.3 μM modified cytosines). The reactions were incubated at 37 ºC for ~ 17 h.

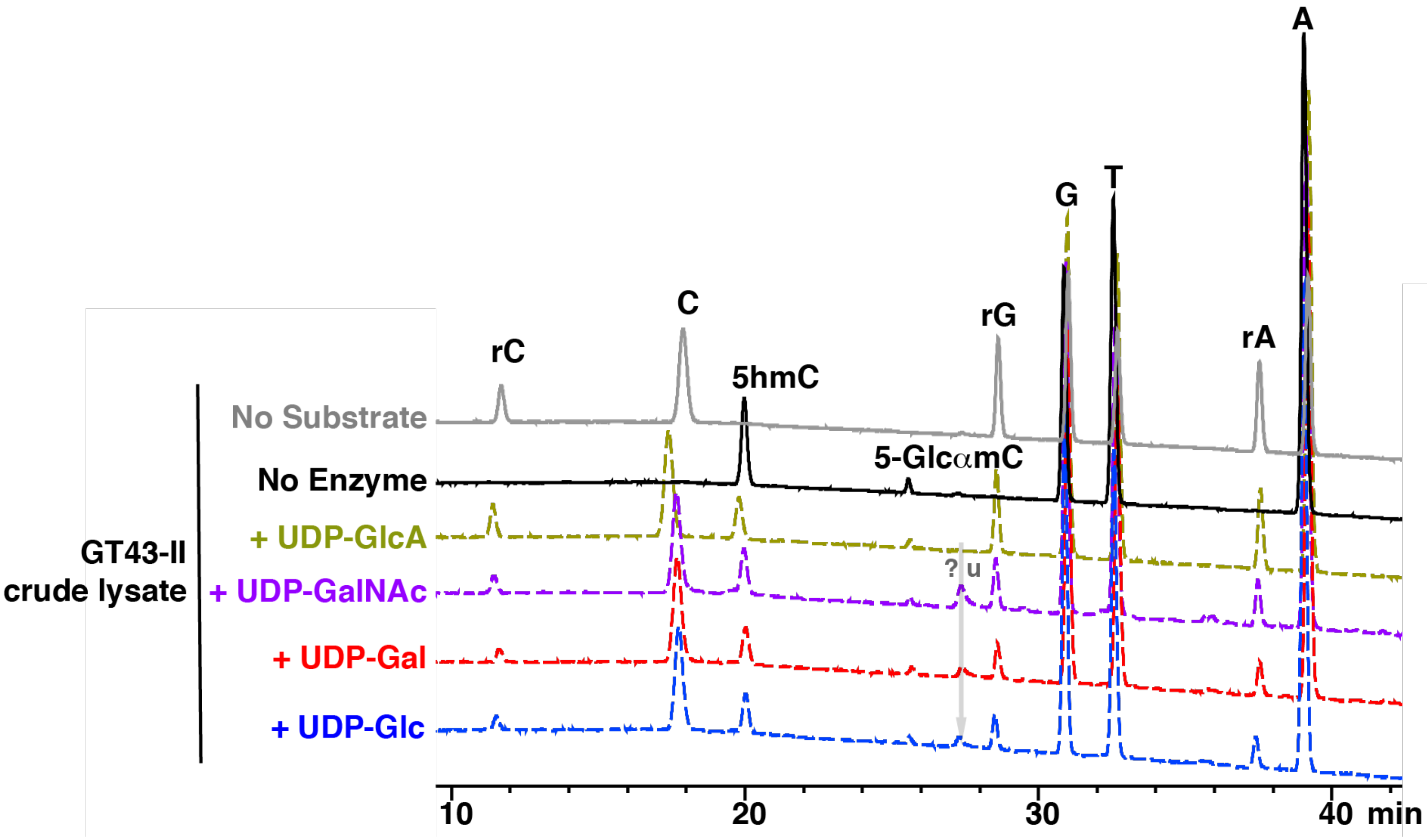

Fig. S18. LC-MS analysis showing the *in vitro* activity of 10 µL GT43-II crude lysate with 50 ng/µL T4 phage *gt-/-* DNA (27.5 µM modified cytosines) in the presence of different UDP-sugars (400 µM final concentration). The reactions were incubated at 37 ºC for ~ 17 h. r = ribo form of the nucleoside. The peak labeled with a "? u" has a mass signal that is below the sensitivity of detection of the instrument.

Contig #	Accession Number	Sequence Length	Database
83	3300009508_____Ga0115567_10014341	7,093	IMG/VR
49	3300009550_____Ga0115013_10000019	16,324	IMG/VR
76	3300005941_____Ga0070743_10000793	12,218	IMG/VR
88	Station206_SUR_ALL_assembly_NODE_16923_length_5542_cov_88.571168	5,542	GOV2.0
74	Station188_DCM_ALL_assembly_NODE_6288_length_12953_cov_7.669561	12,953	GOV2.0
70	3300000168_____LPjun09P1210mDRAFT_c1000008	19,050	IMG/VR
71	3300009507_____Ga0115572_10001606	18,298	IMG/VR
82	Station32_SUR_ALL_assembly_NODE_4389_length_7226_cov_2.950774	7,226	GOV2.0
69	Station168_IZZ_ALL_assembly_NODE_4375_length_22934_cov_10.789938	22,934	GOV2.1
43	3300012928_____Ga0163110_10009824	5,317	IMG/VR
57	3300009593_____Ga0115011_10000038	20,162	IMG/VR
53	3300009790_____Ga0115012_10003245	9,607	IMG/VR
55	3300000929_____NpDRAFT_10020496	7,886	GOV2.0
35	3300017697_____Ga0180120_10003723	7,582	GOV2.0
87	Station191_SUR_ALL_assembly_NODE_11004_length_5776_cov_11.142458	5,776	GOV2.0
78	3300006413_____Ga0099963_1015931	11,262	IMG/VR
61	3300006990_____Ga0098046_1000722	11,710	IMG/VR
41	3300006789_____Ga0098054_1000597	21,132	IMG/VR
77	3300000224_____SI34jun09_10mDRAFT_1000741	11,430	IMG/VR
66	Station34_DCM_ALL_assembly_NODE_476_length_29640_cov_4.296941	29,640	GOV2.0
81	Station30_DCM_ALL_assembly_NODE_2814_length_8891_cov_2.853554	8,891	GOV2.0
85	Station122_DCM_ALL_assembly_NODE_4607_length_6363_cov_4.648066	6,363	GOV2.0
67	3300008007_____Ga0100386_10001213	25,382	IMG/VR
72	Station76_MES_COMBINED_FINAL_NODE_3609_length_17389_cov_5.877120	17,389	GOV2.0
34	3300009079_____Ga0102814_10006475	7,255	IMG/VR
40	3300013130_____Ga0172363_10003614	12,573	IMG/VR
79	Station85_MES_COMBINED_FINAL_NODE_953_length_32906_cov_4.548964	11,050	GOV2.0
89	Station189_MES_ALL_assembly_NODE_24926_length_5457_cov_5.283784	5,457	GOV2.0
73	Station52_SUR_ALL_assembly_NODE_1027_length_14824_cov_73.851581	14,824	GOV2.0
75	Station191_SUR_ALL_assembly_NODE_3166_length_12853_cov_84.951321	12,853	GOV2.0
86	3300009409_____Ga0114993_10017275	6,095	IMG/VR
80	Station208_SUR_ALL_assembly_NODE_6354_length_9043_cov_12.757232	9,043	GOV2.0

Table S1. Description of the viral metagenomic contigs used in this study. Details on accession numbers and databases used are found in *SI Materials and Methods.*

TET Clade	pJS119k-TET	pACYC-C5-MT	% C	% 5mC	% 5hmC
C1	34	34	99.98	0.01	0.00
C1	34	43	78.13	21.86	0.00
C1	34	41	83.18	16.82	0.00
C1	34	85	80.17	19.82	0.00
D	40	40	99.98	0.01	0.00
D	40	43	78.08	21.92	0.00
D	40	41	85.42	13.64	0.93
D	40	85	-	-	-
C1	67	67	99.97	0.02	0.00
C1	67	43	77.59	22.41	0.00
C1	67	41	85.72	14.26	0.01
C1	67	85	81.70	18.30	0.00
C1	72	72	99.99	0.00	0.00
C1	72	43	77.66	22.34	0.00
C1	72	41	85.30	14.69	0.00
C1	72	85	78.93	21.07	0.00
C1	75	75	98.87	1.12	0.00
C1	75	43	77.19	22.81	0.00
C1	75	41	85.66	14.33	0.00
C1	75	85	86.57	13.43	0.00
C1	80	80	99.37	0.63	0.00
C1	80	43	77.76	22.24	0.00
C1	80	41	85.54	14.45	0.00
C1	80	85	82.08	17.92	0.00
C1	86	86	99.99	0.00	0.00
C1	86	43	77.58	22.41	0.00
C1	86	41	85.57	14.43	0.00
C1	86	85	77.44	22.56	0.00

TET Clade	pJS119k-TET	pACYC-C5-MT	% C	% 5mC	% 5hmC
D	73	73	99.88	0.10	0.00
D	73	43	85.58	14.42	0.00
D	73	41	79.42	20.58	0.00
D	73	85	77.53	22.47	0.00
C1	79	79	99.98	0.01	0.00
C1	79	43	77.27	22.72	0.00
C1	79	41	85.33	14.66	0.00
C1	79	85	80.57	19.43	0.00
C2	89	89	93.07	6.93	0.00
C2	89	43	78.98	20.92	0.09
C2	89	41	89.38	10.59	0.03
C2	89	85	82.92	17.02	0.05
A	69	69	99.55	0.44	0.00
A	69	43	78.24	21.38	0.37
A	69	41	85.32	14.68	0.00
A	69	85	79.95	17.47	2.58
A	70	70	94.36	5.60	0.03
A	70	43	77.01	21.08	1.90
A	70	41	81.73	17.26	1.00
A	70	85	73.41	23.47	3.11
A	74	74	97.05	2.94	0.01
A	74	43	77.58	21.52	0.89
A	74	85	79.14	20.24	0.62
B2	81	81	82.60	17.39	0.00
B2	81	43	77.25	22.74	0.00
B2	81	41	85.75	14.24	0.00
B2	81	85	80.39	19.61	0.00

Table S2. LC-MS/MS analysis of the *in vivo* activities of select TETs with C5-MT 43 (clade I), 41 (clade IIa), and 85 (clade IIb) on *E. coli* gDNA. Detectable methyl hydroxylation activity is highlighted in fluorescent yellow.

Sample Definition	Trace #	5hmC	5-GlcαmC	460.2 u	419 u	G	T	A
GT14-II + UDP-Glc+T4*gt -/-* DNA	2	0.486	0.059	0	0.054	0.550	1.083	1.000
GT14-II + UDP-GlcNAc+T4g*t -/-* DNA	3	0.419	0.059	0.111	0	0.548	1.086	1.000
T4*gt -/-* DNA (No enzyme control)	1	0.549	0.058	0	0	0.546	1.087	1.000

Sample Definition	Trace #	5-GlcαmC	5-GlcβmC	G	T	A
GT14-I + UDP-Glc+T4gt *wt* DNA	6	0.326	0.198	0.544	1.081	1.000
T4gt *wt* DNA (No enzyme control)	5	0.364	0.235	0.543	1.079	1.000

Sample Definition	Trace #	5-GlcαmC	5-GlcβmC	G	T	A
GT43-I + UDP-Glc+T4gt *wt* DNA	7	0.310	0.131	0.543	1.083	1.000
T4gt *wt* DNA (No enzyme control)	not shown	0.357	0.217	0.543	1.079	1.000

Table S3. Relative abundance of each nucleoside normalized to adenosine (set as 1). The relative abundance of 5-GlcαmC, 5-GlcβmC, 419 u species, and 460 u species were estimated using the extinction coefficient of 5hmC at 273 nm.