Mapping Protein Occupancy on DNA with an Unnatural Cytosine Modification in Bio-orthogonal Contexts

Ruiyao Zhu[1], Christian E. Loo[2], Christina M. Hurley[2],

Jared B. Parker[1,3], Rahul M. Kohli[1,3*]

[1]Department of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

[2]Graduate Group in Biochemistry, Biophysics, and Chemical Biology, University of Pennsylvania, Philadelphia, PA, 19104, USA

[3]Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA, 19104, USA

[*] Corresponding Author

Abstract

The epigenome provides a dynamic layer of gene regulatory control above the static genetic sequence. Epigenetic DNA base modifications are a key regulator of gene expression, which in mammalian genomes predominantly occur in CpG contexts and are disproportionately distributed in regulatory regions like promoters. Chromatin-associated proteins and transcription factors work in tandem with these modifications to further control gene regulation. Given the interplay of these factors, mapping DNA base modifications concurrently with protein-DNA occupancy is therefore critical to interpreting the epigenome. Current multi-modal mapping methods employ DNA methyltransferases that mark accessible protein-unbound DNA in non-CpG contexts. However, the overlap of this exogenous DNA methylation with naturally occurring modifications can confound readouts and significantly limit compatibility with methods to simultaneously read epigenetic states. To circumvent these limitations, we explored the possibility of introducing an unnatural DNA base modification as an alternative label for protein occupancy. Here, we report our efforts to rationally engineer non-CpG-specific DNA carboxymethyltransferases, characterize their neomorphic activity, and assess DNA carboxymethylation as a reporter of protein occupancy on DNA. We find that DNA carboxymethylation of cytosines in GpC contexts shows broad compatibility with the most widely used epigenetic detection methods and reliably reports on protein occupancy state. Our results demonstrate that unnatural DNA modifications in bio-orthogonal sequence contexts, coupled with either chemical or enzymatic deamination, can potentiate new approaches to multimodal epigenetic profiling.

Introduction

Chemical modifications of DNA bases enable dynamic control over gene regulation and cell fate.(Nabel et al., 2012) In mammalian genomes, DNA base modifications primarily occur at cytosines within cytosine–guanine (CpG) dinucleotide contexts, with 5-methylcytosine (5mC) being the most common and well-characterized.(Schubeler, 2015) 5mC is generated by DNA methyltransferases (MTases) using S-adenosyl-L-methionine (SAM) as a methyl donor.(Iyer et al., 2011) 5mC is typically regarded as a repressive epigenetic mark with diverse roles that include transcriptional silencing, imprinting, and X-chromosome inactivation.(Schubeler, 2015) 5mC dynamics are regulated in part by further modification, with ten–eleven translocation (TET) dioxygenases oxidizing 5mC to 5-hydroxymethylcytosine (5hmC) and subsequent higher oxidation states, which can relieve the gene repression imposed by 5mC marks.(Kohli and Zhang, 2013) Importantly, the functional consequences of these DNA base modifications depend critically on their context, whether buried within nucleosome-wrapped DNA or exposed and available to interact with transcription factors.(Ambrosi et al., 2017; Kreibich et al., 2023). Direct DNA modifications combined with DNA-interacting proteins thus constitute a highly complex epigenetic landscape.

Given the interconnected nature of these regulatory processes, jointly profiling modification and protein occupancy of DNA is essential for understanding how gene expression is controlled. Numerous methods have been established to profile this regulatory network as individual layers in isolation, including either DNA modification mapping or protein occupancy.(Dey et al., 2012; Minnoye et al., 2021; Wang, T. et al., 2021) Conventional chemical approaches include methods that discriminate among cytosine modification states, such as selective deamination with bisulfite sequencing (BS-Seq), as well as approaches that profile histone-bound regions, such as hydroxyl radical footprinting or dimethyl sulfate.(Frommer et al., 1992; Jain and Tullius, 2008; Umeyama and Ito, 2018) However, chemical discrimination approaches can be destructive, requiring large DNA inputs and causing biased or selective fragmentation of DNA.(Nielsen, 1990; Tanaka and Okamoto, 2007) To overcome these limitations, alternative detection methods have recently been developed that harness DNA modifying 'writer' enzymes as 'readers' of the epigenome. For example, APOBEC3A (A3A), a DNA cytosine deaminase involved in viral defense, has been repurposed to profile DNA modifications. Leveraging its ability to selectively deaminate unmodified cytosines and 5mC, but not protected 5hmC, A3A can be used to localize 5hmC in APOBEC-coupled epigenetic sequencing (ACE-Seq), or partnered with TET enzymes to distinguish unmodified C from both 5mC and 5hmC (EM-Seq).(Schutsky et al., 2018; Vaisvila et al., 2021) Similarly, enzymatic approaches have been employed in methods such as DNase-Seq or ATAC-Seq to fragment and thereby map open chromatin. (Buenrostro et al., 2015; Song and Crawford, 2010) In contrast to chemical approaches, these enzymatic approaches are non-destructive, enabling the analysis of minute samples, such as rare cell populations or even single cells.

Nevertheless, DNA fragmentation can still introduce cleavage bias and limit resolution across long reads. Consequently, enzymatic approaches that selectively label DNA have been developed to profile protein occupancy. For instance, methods like SMAC-Seq, SAMOSA, and Fiber-Seq use an adenine DNA MTase, EcoGII, to mark accessible regions with 6-methyladenine (m6A), a natural modification that occurs predominantly in plants and prokaryotes but is otherwise absent in mammalian genomes.(Abdulhay et al., 2020; Shipony et al., 2020; Stergachis et al., 2020) Methods including NOMe-Seq and its variants instead leverages a GpC-specific MTase M.CviPI to selectively install 5mC in free DNA regions unbound by proteins like histone or transcription factors. Endogenous and the resulting exogenous base modifications can be partially distinguished based on CpG or GpC context, allowing for simultaneous profiling of DNA methylation and accessibility.(Kelly et al., 2012; Lee et al., 2020; Pott, 2017; Wasney and Pott, 2023) Despite these advances, the use of natural DNA base

modifications as protein occupancy labels presents inherent limitations. Most notably, the identical chemical nature of these marks with natural modifications limits their compatibility with detection methods and subsequent readout. Specifically, NOMe-Seq is unable to resolve 5mC from 5hmC and requires disambiguation of 5mC in overlapping CpG and GpC sequence contexts. On the other hand, the requirement for third-generation sequencing to read m6A significantly increases sample input requirements for Fiber-Seq and limits the use of more common and reliable methods for studying cytosine modifications, such as BS-Seq or EM-Seq.(Kong et al., 2024; Searle et al., 2023)

Here, we considered the possibility that the enzymatic introduction of unnatural base modifications could complement current methods for multimodal mapping. In support of the potential utility of unnatural DNA modifications in epigenetic analysis, we have recently shown that a CpG-specific DNA MTase (M.MpeI) can be engineered to take on the neomorphic ability to accept carboxy-S-adenosyl-L-methionine (CxSAM), a sparse natural metabolite.(Wang, T. and Kohli, 2021) The engineered M.MpeI N374K converts unmodified CpGs into ones with 5-carboxymethylcytosine (5cxmC), an unnatural base modification that is resistant to both bisulfite and enzymatic deamination. The CpG-specific DNA carboxymethyltransferase (CxMTase) enabled the development of Direct Methylation sequencing (DM-Seq), a fully enzymatic method for profiling 5mC directly at single-base resolution.(Wang, Tong et al., 2023)

The precedent established by the CpG-specific DNA CxMTase suggests the possibility that unnatural DNA modifications could also be used as orthogonal labels in non-CpG contexts to map DNA-protein interactions. In considering options, we posited that an ideal method for simultaneous detection of DNA base modification and protein occupancy would introduce marks that are distinct from native modifications and compatible with both chemical and enzymatic methods to detect DNA modifications (Figure 1). To potentiate such an approach, here, we engineer two novel CxMTases with differing non-CpG context specificities, characterize their activities through biochemical and sequencing-based assays, and demonstrate their ability to act as sensors of protein occupancy. We thus show that engineered CxMTases can support a fully enzymatic, non-destructive sequencing pipeline for mapping of DNA-protein interactions, extending the utility of DNA modifying enzymes in resolving heterogeneous epigenetic states.

Results and Discussion

Construction and Screening of non-CpG MTase Candidates

In exploring candidates to introduce unnatural labels into DNA, we considered that the ideal engineered DNA MTase would avoid overlap with endogenous cytosine modifications in CpGs. Additionally, as the frequency of the MTase recognition sequence determines the resolution of protein occupancy footprinting of DNA, shorter recognition sequences could offer higher resolution. With these considerations in mind, we chose to examine two previously identified cytosine DNA MTases that have a specific dinucleotide recognition sequence: M.CviPI (GpC) and M.CviQIX (CpC) (target cytosine underlined).(Miura et al., 2022)

Given the conserved nature of SAM-dependent cytosine MTases, we hypothesized that mutations at residues homologous to M.MpeI N374 could potentially confer carboxymethylation activity in other cytosine MTases.(Bheemanaik et al., 2006) As the homologous residues were not immediately obvious from a primary sequence alignment (Figure S1A), we generated AlphaFold3-predicted structures of M.CviPI and M.CviQIX bound to DNA substrates containing either a methylated GpC or a methylated CpC, respectively.(Abramson et al., 2024; Thompson et al., 1994) We aligned the predicted structures with the crystal structure of M.MpeI (PDB 4DKJ) bound to a CpG-containing DNA substrate and the S-adenosyl-L-homocysteine (SAH) byproduct. Both predicted structures showed high agreement with M.MpeI around the active site, exhibiting features of a Rossmann fold with alternating α helices and β strands and the base-flipping mechanism common for DNA MTases (Figure S1B). (Chouhan et al., 2019; Ren et al., 2022) Using these models, we identified M.CviPI Y205 and M.CviQIX N218 as the residues closest to the Cα of the N374 site in M.MpeI and therefore a target for potential mutagenesis (Figure 2). In our prior saturation mutagenesis exploration at N374 of M.MpeI, we found that the introduction of a cationic side chain best supported CxSAM activity.(Wang, T. and Kohli, 2021) This finding could be rationalized in molecular modeling of M.MpeI N374K with CxSAM, which highlighted the contribution of a newly established salt-bridge interaction.(Loo et al., 2023) Building on these precedents, we therefore chose to focus our efforts on K and R mutants of M.CviPI Y205 and M.CviQIX N218.

Given that high enzyme activity is essential for sequencing method development, we first evaluated the impact of potential CxMTase conferring mutations on canonical MTase activity. We assayed the two MTases using endogenous SAM, building on our prior work with M.MpeI, which showed that the gain in CxMTase activity was accompanied by maintained MTase activity (Wang et Kohli, 2021). To this end, we individually expressed the wild-type and variant enzymes in a methylation-tolerant strain of *E. coli*, using WT M.MpeI as a comparison. We then assessed MTase activity on host DNA *in vivo* by extracting gDNA from the *E. coli* and digesting it with HaeIII, a GGCC-recognizing restriction enzyme blocked by modification at the internal cytosine (Figure 3A). As expected, WT M.MpeI did not protect against HaeIII digestion, consistent with its specific methylation of CpG sites. In contrast, both wild-type M.CviPI and M.CviQIX conferred significant protection, indicating that most GGCC sites in the gDNA were modified. With both K and R mutants for each enzyme, we observed a pattern consistent with partial protection, suggesting reduced but detectable activity. Notably, the mutants of M.CviPI exhibited higher levels of partial protection relative to that of M.CviQIX (Figure 3B), suggesting that M.CviPI variants are more active MTases than M.CviQIX mutants when overexpressed in *E. coli*. We therefore proceeded with the GpC-specific M.CviPI for more detailed biochemical investigation.

Engineered GpC MTases Gain Neomorphic Carboxymethylation Activity

Given that CxSAM is 400-fold less abundant than SAM in *E. coli*, we attributed the activity detected upon enzyme overexpression in *E. coli* predominantly to SAM-mediated generation of 5mC. (Kim et al., 2015) To examine enzymatic activity with SAM and CxSAM *in vitro*, we next expressed and purified M.CviPI WT, Y205K, and Y205R (Figure S2). We then incubated an unmodified pUC19 plasmid DNA substrate with decreasing concentrations of each enzyme, in the presence of a fixed concentration of either excess SAM or CxSAM. Modification levels were again assessed by digestion with HaeIII (Figure 4A). As anticipated, pUC19 plasmids incubated with WT M.CviPI were efficiently digested by HaeIII in the absence of SAM, but were protected in its presence, consistent with efficient GpC methylation. In line with the *in vivo* screen, where partial protection of gDNA was observed, both Y205K and Y205R mutants were also capable of protecting pUC19 from digestion in the presence of SAM. However, unlike the *in vivo* screen examining gDNA protection, the methylation efficiency appeared similar across WT and mutants *in vitro* (Figure 4B). This discrepancy likely reflects limiting SAM concentrations *in vivo*, which disproportionately affected the mutants. When these mutants were incubated with CxSAM, WT M.CviPI failed to protect pUC19 from HaeIII digestion, consistent with its constrained active site and inability to utilize CxSAM. By contrast, both Y205R and Y205K incubated with CxSAM protected pUC19 from HaeIII digestion, supporting their efficient use of CxSAM as a cofactor, consistent with our hypothesis.

To further characterize the potential CxMTase activity, we next performed an orthogonal assay using a duplexed oligonucleotide rather than plasmid DNA as the substrate. In this assay, we used a fluorescently labeled 27-bp oligonucleotide (FAM-S27, Table S1) with a single GpC site within an HaeIII (GGCC) recognition sequence. In the unlabeled complement strand, we included a 5mC at the GpC site (S27-Comp-5mC, Table S1), allowing us to focus on activity at a single unmodified GpC in the labeled strand without confounding. Following incubation with the M.CviPI variants and either SAM or CxSAM, S27-Comp-5mC was exchanged with an excess of the complementary strand lacking 5mC (S27-Comp-C) to allow us to interrogate with HaeIII and determine whether the GpC in FAM-S27 is unmodified (digested) or modified (protected from digestion) (Figure 4C). In this assay, we observed that both WT and mutant enzymes incubated with SAM protected FAM-S27 from HaeIII digestion, suggesting retention of MTase activity, consistent with the pUC19-based assay. By contrast, when incubated with CxSAM only the Y205K and Y205R mutants showed significant protection of the duplex, consistent with a gain of CxMTase activity with the mutants (Figure 4D). To definitively establish the identity of the modification, we performed intact oligonucleotide LC/MS on the initial duplex after incubation with M.CviPI Y205K and CxSAM. In addition to the two expected peaks corresponding to the masses of FAM-S27 and S27-Comp-5mC, we observed a unique peak with a mass shift of 57.8 Da relative to FAM-S27, consistent with the presence of a single 5-carboxymethylcytosine (5cxmC) in the modified oligonucleotide (Figure 4E).

Bio-orthogonal Carboxymethylation Offers Compatibility with Common Deamination Methods

Having validated the successful engineering of GpC CxMTases, we next aimed to quantitatively characterize their activities on a more complex DNA substrate and additionally test their compatibility with different epigenetic sequencing methods. To that end, we performed transfer reactions on an unmodified l phage genome (48.5 kb) with either SAM or CxSAM, and then resolved the resulting DNA modifications at base resolution (Figure 5A). In these sequencing pipelines, the modified DNA was sheared and sequenced to a total depth of greater than 120-fold using both BS-Seq and ACE-Seq. The use of both chemical and enzymatic sequencing approaches allows for resolution of C, 5mC, and 5cxmC. Based on established precedent, both 5mC and 5cxmC are protected from C-to-T conversion

in BS-Seq, whereas only 5cxmC is protected from C-to-T conversion by A3A in the ACE-Seq pipeline.(Wang, Tong et al., 2023) As a positive control, we also included the engineered M.MpeI N374K CpG CxMTase to allow for comparison of CpG and GpC modifying enzyme activities.

First focusing on BS-Seq to compare enzymatic activity with SAM and CxSAM as substrates, we observed the efficient use of SAM as a substrate across M.CviPI WT and mutants as measured by the average level of protection of GpC sites from C-to-T conversion (Figure 5B). This result suggests efficient MTase activity in GpC contexts, consistent with our previous *in vitro* assays that only interrogated GGCC sites using sensitivity to HaeIII cleavage. At the SAM concentration examined, there was no significant difference in modification level by M.CviPI WT and mutants. Using CxSAM with the WT enzyme, we observed no significant protection of GpC sites from C-to-T conversion relative to background levels of C-to-T non-conversion observed in the absence of SAM. By contrast, when the Y205K and Y205R mutants were used with CxSAM, average protection levels of 64.5% and 59.0% were observed, respectively, for GpCs from C-to-T conversion in BS-Seq. We observed no significant differences in protection between the two variants, although M.CviPI Y205K showed less variance in replicates relative to Y205R. Notably, all M.CviPI variants showed no significant modification in WCG contexts with C-to-T non-conversion relative to background as expected (Figure S3A).

Unlike 5mC generated from SAM, 5cxmC generated from CxSAM is known to resist enzymatic deamination in the ACE-Seq pipeline.(Wang, Tong et al., 2023) Indeed, when we focused our analysis on individual GpC sites in the genome, when SAM was used as a substrate, these sites remained unconverted in BS-Seq but underwent significant C-to-T conversion in ACE-Seq, leaving no readily detectable signal in the latter pipeline. By contrast, samples treated with the M.CviPI mutants and CxSAM yielded unconverted GpC sites in both BS-Seq and ACE-Seq. Moreover, these signals from ACE-Seq correlate well with signals from BS-Seq (Figure 5C), which indicates that M.CviPI mutants incubated with CxSAM are efficiently protecting the GpC sites with a modification that resists both chemical and enzymatic conversion and thereby compatible with multiple detection methods.

To further characterize the substrate context specificity of these enzymes, we generated a position weight matrix of the modified sites for each enzyme (Figure 5D; Figure S3B; Figure S3C). As expected, all variants show a preference for GpC contexts. WT M.CviPI is known to have some off-target activity at CpC site, which can confound footprinting studies in NOMe-Seq.(Kelly et al., 2012) Interestingly, while our analysis by BS-Seq revealed this expected off-target CpC activity with WT M.CviPI and SAM, we observed that both mutants exhibit improved context specificity with SAM, while retaining similar levels of methyl transfer. These observations suggest that M.CviPI Y205K and Y205R, even when used in conjunction with SAM, may offer useful advantages to WT M.CviPI for use in epigenetic mapping. With the mutant enzymes, we find that this selectivity pattern is further enhanced with CxSAM as a substrate, with minimal activity detectable outside of a GpC context. We also examined the sequence context flanking target sites (Figure S4) to look for bias in GpC modification. We observed that the CpG CxMTase shows higher bias, with discrimination against -1 A or +2 R (A or G) relative to the CpG site, while the GpC CxMTase shows lower overall bias, with only minor discrimination against +1 A relative to the GpC site.

In prior work, we found that carboxymethylation activity of the CpG CxMTase M.MpeI N374K favors the generation of hemi-carboxymethylated CpG sites, likely due to the steric bulk of the first carboxymethyl group hindering efficient modification of the opposite strand CpG.(Wang, Tong et al., 2023) This asymmetric activity posed a challenge in the development of DM-Seq that was resolved by introducing 5mC in a copy-strand to help direct the CxMTase to the target strand. To examine if the

GpC CxMTases also favor asymmetric modification, we correlated the modification on the Watson strand with those in the Crick strand for GpC dyads across the l phage genome (Figure S3D; Figure S3E). With the M.MpeI N374K enzyme, asymmetric modification is readily apparent, with lower levels of modification on one strand when the opposite strand CpG is more heavily modified (top left or bottom right corners of plot). Compared to the CpG CxMTase, the GpC CxMTases show significantly less asymmetry, as evidenced by a limited number of sites where modification on one strand is associated with the absence of modification in the other strand. Additionally, while CpG CxMTase asymmetry limits global modification to approximately 50%, GpC CxMTases do not exhibit the same constraint. Indeed, we found that increasing CxSAM concentration from 300 μM to 800 μM enhances carboxymethylation to 82.7% of sites without notable asymmetry or off-target activity (Figure S3C). Thus, the use of GpC CxMTases could offer some distinctive advantages over CpG CxMTase by supporting the potential symmetric introduction of 5cxmC on GpC sequences in both DNA strands.

GpC Carboxymethylation Distinguishes Protein-Bound States

Having engineered CxMTases, characterized their activities, and assessed their compatibility with different modes of deamination, we last aimed to apply the neomorphic enzyme as a potential reporter of protein-bound states of DNA. In established NOMe-Seq approaches, WT M.CviPI is used to selectively introduce 5mC modifications into DNA that is unencumbered by proteins, as bound histones or other DNA-binding proteins block MTase activity. To model whether CxMTase activity and the resulting bio-orthogonal 5cxmC mark could similarly be used in footprinting approaches, we devised an *in vitro* assay to assess functionality.

Our reporter system employs the key regulatory DNA binding protein LexA from the bacterial DNA damage response, also known as the SOS response.(Culyba et al., 2018; Maslowska et al., 2019) In this system, LexA is a transcriptional regulator that binds to and represses a large collection of DNA damage response genes through specific interactions with an SOS box DNA binding motif in the gene promoters. In the presence of DNA damage, LexA is induced to undergo self-proteolysis, resulting in induction of the SOS response. In our assay design, we employed a catalytically inactive LexA S119A capable of binding to DNA but not proteolysis. We designed an oligonucleotide containing an embedded central SOS box, which included GpC dinucleotides in a HhaI cleavage site (GCGC), along with a matched control where the SOS binding motif was scrambled (Table S1). We amplified the duplexes with a 5'-fluorescently labeled primer and then used a gel shift assay to validate that LexA S119A bound to the SOS-box–containing duplex, but not the scrambled duplex (Figure S5A).

We next moved to evaluating whether the M.CviPI variants could act as reporters of protein-bound states. Given lower variance in modification activity of M.CviPI Y205K in the previous assay on the lambda genome, we chose to treat the SOS-box–containing and control duplex with WT M.CviPI or M.CviPI Y205K, using SAM or CxSAM as the respective co-substrate, following incubation in the presence or absence of LexA S119A. The reaction was then quenched with Proteinase K, and the resulting duplexes were purified and subjected to HhaI digestion to interrogate the modification state of the target GCGC site within the SOS box (Figure S5B). In the absence of M.CviPI variants, both duplexes can be efficiently digested by HhaI. In the absence of LexA, the SAM/WT M.CviPI and CxSAM/M.CviPI Y205K pairs both result in protection of duplexes from digestion, consistent with efficient modification of the HhaI site. In the presence of LexA S119A, however, while enzymatic protection of the scrambled duplex is unaffected, the SOS box containing duplex is no longer efficiently protected from digestion by HaeIII, suggesting that the LexA-bound DNA is not accessible to the GpC WT MTase or CxMTase (Figure S5C).

To more rigorously examine whether we could footprint LexA, we next aimed to resolve its positioning at single-base resolution via BS-Seq (Figure 6; Figure S6). With the scrambled duplex, we observed that either the SAM/WT M.CviPI pair or the CxSAM/M.CviPI Y205K pair result in modifications across GpC sites in the duplex, independent of the presence or absence of LexA S119A. In contrast, with the oligonucleotide containing the SOS box binding motif, we observed a significantly reduced GpC modification in the presence of LexA S119A. Importantly, this reduction in modification was localized to the GpCs within the LexA binding sites, while GpCs outside of the binding window maintained similar modification patterns. Upon closer examination of the base-resolution patterns, we also observe CpC modification with the M.CviPI WT/SAM pair but not with the M.CviPI Y205K/CxSAM pair, consistent with our earlier evidence for increased selectivity with the variant enzyme. The selectivity and modification footprint are thus sufficient to indicate the presence of the 20 bp LexA binding site. Taken together, these results indicate that the GpC CxMTase can be blocked by the presence of protein bound to DNA, while reliably reporting on neighboring free DNA, offering orthogonal information on the epigenetic state of underlying DNA.

Conclusion

DNA modifying enzymes, traditionally viewed as 'writers', are increasingly being harnessed as 'readers' of the epigenetic state of DNA. Beyond resolving DNA modification states, enzymatic 'writers' have been employed to discriminate between regions of DNA that are protein-bound vs. protein-free, revealing insights into the multi-modal epigenetic code. While such approaches have typically relied upon the use of natural DNA MTases and SAM, we sought to develop a bio-orthogonal label to report on protein occupancy, enabling broad compatibility with common epigenetic sequencing methods and more versatile readouts.

Here, we used structure-guided mutagenesis to engineer two MTases that act in sequence contexts distinct from CpG, where natural modifications most commonly occur in mammalian genomes. We then demonstrate their ability to use the synthetic cofactor CxSAM to introduce the unnatural base 5cxmC in a non-CpG sequence context, which is compatible with BS and A3A-based sequencing.

This work builds on our prior success engineering the CpG-specific CxMTase M.MpeI N374K, showcasing that similar mutations at structurally homologous residues within M.CviPI can also result in CxMTase activity. Our findings support a generalizable strategy for converting MTases to CxMTases and suggest that this residue, originally thought to participate in the hydrogen bonding network that promotes MTase activity, is permissive to variation, offering a means to expand cofactor selectivity.(Jurkowski and Jeltsch, 2011; Zhang, X. and Bruice, 2006) Complementing prior studies that have focused on 'space opening' mutations, allowing for the use of bulky SAM analogs, we show that introducing cationic residues can enhance catalysis with CxSAM, demonstrating the potential of charge-to-charge interactions in designing additional targeted partnerships between engineered MTases and SAM analogs.(Dalhoff et al., 2006; Lukinavicius et al., 2012)

While the process of generating the neomorphic CxMTase highlights some shared plasticity in MTases, the specific GpC CxMTase variants that we developed and characterized have some unique attributes that make them useful tools. For one, while asymmetry of CpG CxMTase activity presented a major challenge in the development of DM-Seq, we found that the GpC CxMTase activity was not affected by carboxymethylation on the opposite strand. Further study could elucidate the mechanistic determinants of the symmetric activity of CxMTases, which could potentiate a more efficient DM-Seq workflow and fewer limitations in the technological application of CxMTases. Second, while the WT M.CviPI is known to have some off-target activity at non-GpC sites, we found that the M.CviPI Y205K showed higher fidelity for acting on GpC sites with either SAM or CxSAM as substrates. Thus, the engineered MTases may also offer distinct advantages in NOMe-Seq and other pipelines involving the use of M.CviPI WT. (Aughey et al., 2018; Krebs et al., 2017; Pardo et al., 2011; Wang, Yunhao et al., 2019)

Most notably, we envision that these CxMTases can generate unnatural modifications in distinctive contexts to serve as reporters of the protein-DNA occupancy by selectively labeling unbound DNA. Through an *in vitro* assay, we demonstrated that the GpC CxMTase can footprint the binding of the transcriptional repressor LexA to a 20-bp SOS box DNA binding motif. Interestingly, in the LexA-DNA bound structure, the central region of the SOS box is not directly engaged with LexA.(Zhang, A. P. et al., 2010) Nonetheless, the GpCs in this region appear resistant to the CxMTase, likely because the region is insufficiently accessible to M.CviPI. While further work is needed to benchmark the performance of CxMTase in more complex workflows with mammalian genomic DNA samples, our work presents proof of concept for enzymatic installation of unnatural DNA base modifications as a novel, non-destructive approach to footprint protein occupancy on DNA without fragmentation or interference with native DNA base modifications.

Using unnatural DNA modifications as protein occupancy markers brings unique advantages over previous approaches that rely on natural modifications. For instance, 5cmxC resists deamination and can therefore overcome the inability of exogenous methylation labels to be retained in next generation-based workflows. As both 5hmC and 5cxmC are resistant to enzymatic deamination with A3A, the GpC CxMTase should be useful for multi-modal profiling of 5hmC together with open chromatin, analysis not readily feasible with current approaches. Moreover, the GpC CxMTase opens new possibilities for long-read sequencing platforms. We posit that 5cxmC, given its distinct chemical character, will likely generate a unique signal from natural base modifications. While further efforts are needed to characterize their long-read sequencing signatures, we propose that unnatural modifications like carboxymethylation present a novel, versatile solution for jointly mapping DNA modification and protein occupancy at the single-molecule level, facilitating a more complete understanding of how the functional interactions between the chemical marks on DNA bases with protein regulators shape gene expression.

Author Information

Corresponding Author

Rahul M. Kohli - *Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*; *Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*; https://orcid.org/0000-0002-7689-5678; Email: rkohli@pennmedicine.upenn.edu

Authors

Ruiyao Zhu – *Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Christian E. Loo - *Graduate Group in Biochemistry, Biophysics, and Chemical Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Christina M. Hurley - *Graduate Group in Biochemistry, Biophysics, and Chemical Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Jared B. Parker – *Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States; Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Notes

The University of Pennsylvania has patents pending for CxMTase enzymes with R.M.K as an inventor. C.E.L. currently serves as a scientific advisor to biomodal.

Acknowledgements

R.Z. was supported by the Vagelos Molecular Life Sciences program at the University of Pennsylvania. This work was funded by the National Institutes of Health, R01-HG010646 to R.M.K. and F31-HG012892 to C.E.L. We gratefully acknowledge Hao Wu, Julia Leu, Ying Li, and Aleksia Barka for guidance on enzymatic footprinting, Noa Erlitzki for guidance on computational analysis, and all Kohli lab members for helpful discussions. We thank Eleanor T. Daniels and Jacob M. Seiple for reagent preparation and technical assistance.

References

Abdulhay, N.J., McNally, C.P., Hsieh, L.J., Kasinathan, S., Keith, A., Estes, L.S., Karimzadeh, M., Underwood, J.G., Goodarzi, H., Narlikar, G.J., and Ramani, V. (2020). Massively multiplex single-molecule oligonucleosome footprinting. Elife *9,* 10.7554/eLife.59404.

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J.*, et al.* (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature *630,* 493–500.

Ambrosi, C., Manzo, M., and Baubec, T. (2017). Dynamics and Context-Dependent Roles of DNA Methylation. J. Mol. Biol. *429,* 1459–1475.

Aughey, G.N., Estacio Gomez, A., Thomson, J., Yin, H., and Southall, T.D. (2018). CATaDa reveals global remodelling of chromatin accessibility during stem cell differentiation in vivo. Elife *7,* 10.7554/eLife.32341.

Bheemanaik, S., Reddy, Y.V.R., and Rao, D.N. (2006). Structure, function and mechanism of exocyclic DNA methyltransferases. Biochem. J. *399,* 177–190.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. *109,* 21.29.1–21.29.9.

Chouhan, B.P.S., Maimaiti, S., Gade, M., and Laurino, P. (2019). Rossmann-Fold Methyltransferases: Taking a "β-Turn" around Their Cofactor, S-Adenosylmethionine. Biochemistry *58,* 166–170.

Culyba, M.J., Kubiak, J.M., Mo, C.Y., Goulian, M., and Kohli, R.M. (2018). Non-equilibrium repressor binding kinetics link DNA damage dose to transcriptional timing within the SOS gene network. PLoS Genet. *14,* e1007405.

Dalhoff, C., Lukinavicius, G., Klimasauskas, S., and Weinhold, E. (2006). Direct transfer of extended groups from synthetic cofactors by DNA methyltransferases. Nat. Chem. Biol. *2,* 31–32.

Dey, B., Thukral, S., Krishnan, S., Chakrobarty, M., Gupta, S., Manghani, C., and Rani, V. (2012). DNA-protein interactions: methods for detection and analysis. Mol. Cell. Biochem. *365,* 279–299.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci. U. S. A. *89,* 1827–1831.

Iyer, L.M., Abhiman, S., and Aravind, L. (2011). Natural history of eukaryotic DNA methylation systems. Prog. Mol. Biol. Transl. Sci. *101,* 25–104.

Jain, S.S., and Tullius, T.D. (2008). Footprinting protein-DNA complexes using the hydroxyl radical. Nat. Protoc. *3,* 1092–1100.

Jurkowski, T.P., and Jeltsch, A. (2011). On the Evolutionary Origin of Eukaryotic DNA Methyltransferases and Dnmt2. PloS One *6,* e28104.

Kelly, T.K., Liu, Y., Lay, F.D., Liang, G., Berman, B.P., and Jones, P.A. (2012). Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome Res. *22,* 2497–2506.

Kim, J., Xiao, H., Koh, J., Wang, Y., Bonanno, J.B., Thomas, K., Babbitt, P.C., Brown, S., Lee, Y., and Almo, S.C. (2015). Determinants of the CmoB carboxymethyl transferase utilized for selective tRNA wobble modification. Nucleic Acids Research *43,* 4602–4613.

Kohli, R.M., and Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. Nature *502,* 472–479.

Kong, Y., Zhang, Y., Mead, E.A., Chen, H., Loo, C.E., Fan, Y., Ni, M., Zhang, X., Kohli, R.M., and Fang, G. (2024). Critical assessment of nanopore sequencing for the detection of multiple forms of DNA modifications. bioRxiv

Krebs, A.R., Imanci, D., Hoerner, L., Gaidatzis, D., Burger, L., and Schübeler, D. (2017). Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. Mol. Cell *67,* 411–422.e4.

Kreibich, E., Kleinendorst, R., Barzaghi, G., Kaspar, S., and Krebs, A.R. (2023). Single-molecule footprinting identifies context-dependent regulation of enhancers by DNA methylation. Mol. Cell *83,* 787–802.e9.

Lee, I., Razaghi, R., Gilpatrick, T., Molnar, M., Gershman, A., Sadowski, N., Sedlazeck, F.J., Hansen, K.D., Simpson, J.T., and Timp, W. (2020). Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat. Methods *17,* 1191–1199.

Loo, C.E., Hix, M.A., Wang, T., Cisneros, G.A., and Kohli, R.M. (2023). Revealing Drivers for Carboxy-S-adenosyl-l-methionine Use by Neomorphic Variants of a DNA Methyltransferase. ACS Chem. Biol. *18,* 2224–2232.

Lukinavicius, G., Lapinaite, A., Urbanaviciute, G., Gerasimaite, R., and Klimasauskas, S. (2012). Engineering the DNA cytosine-5 methyltransferase reaction for sequence-specific labeling of DNA. Nucleic Acids Res. *40,* 11594–11602.

Maslowska, K.H., Makiela-Dzbenska, K., and Fijalkowska, I.J. (2019). The SOS system: A complex and tightly regulated response to DNA damage. Environ. Mol. Mutagen. *60,* 368–384.

Minnoye, L., Marinov, G.K., Krausgruber, T., Pan, L., Marand, A.P., Secchia, S., Greenleaf, W.J., Furlong, E.E.M., Zhao, K., Schmitz, R.J., Bock, C., and Aerts, S. (2021). Chromatin accessibility profiling methods. Nat. Rev. Methods Primers *1,* 10.1038/s43586–9. Epub 2021 Jan 21.

Miura, F., Miura, M., Shibata, Y., Furuta, Y., Miyamura, K., Ino, Y., Bayoumi, A.M.A., Oba, U., and Ito, T. (2022). Identification, expression, and purification of DNA cytosine 5-methyltransferases with short recognition sequences. BMC Biotechnol. *22,* 33–3.

Nabel, C.S., Manning, S.A., and Kohli, R.M. (2012). The curious chemical biology of cytosine: deamination, methylation, and oxidation as modulators of genomic potential. ACS Chem. Biol. *7,* 20–30.

Nielsen, P.E. (1990). Chemical and photochemical probing of DNA complexes. J. Mol. Recognit. *3,* 1–25.

Pardo, C.E., Darst, R.P., Nabilsi, N.H., Delmas, A.L., and Kladde, M.P. (2011). Simultaneous single-molecule mapping of protein-DNA interactions and DNA methylation by MAPit. Curr. Protoc. Mol. Biol. *Chapter 21,* Unit 21.22.

Pott, S. (2017). Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. Elife *6,* 10.7554/eLife.23203.

Ren, R., Horton, J.R., Hong, S., and Cheng, X. (2022). Recent Advances on DNA Base Flipping: A General Mechanism for Writing, Reading, and Erasing DNA Modifications. Adv. Exp. Med. Biol. *1389,* 295–315.

Schubeler, D. (2015). Function and information content of DNA methylation. Nature *517,* 321–326.

Schutsky, E.K., DeNizio, J.E., Hu, P., Liu, M.Y., Nabel, C.S., Fabyanic, E.B., Hwang, Y., Bushman, F.D., Wu, H., and Kohli, R.M. (2018). Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. Nat. Biotech. *36,* 1083–1090.

Searle, B., Müller, M., Carell, T., and Kellett, A. (2023). Third-Generation Sequencing of Epigenetic DNA. Angew. Chem. Int. Ed Engl. *62,* e202215704.

Shipony, Z., Marinov, G.K., Swaffer, M.P., Sinnott-Armstrong, N.A., Skotheim, J.M., Kundaje, A., and Greenleaf, W.J. (2020). Long-range single-molecule mapping of chromatin accessibility in eukaryotes. Nat. Methods *17,* 319–327.

Song, L., and Crawford, G.E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. *2010,* pdb.prot5384.

Stergachis, A.B., Debo, B.M., Haugen, E., Churchman, L.S., and Stamatoyannopoulos, J.A. (2020). Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science *368,* 1449–1454.

Tanaka, K., and Okamoto, A. (2007). Degradation of DNA by bisulfite treatment. Bioorg. Med. Chem. Lett. *17,* 1912–1915.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22,* 4673–4680.

Umeyama, T., and Ito, T. (2018). DMS-seq for In Vivo Genome-Wide Mapping of Protein-DNA Interactions and Nucleosome Centers. Curr. Protoc. Mol. Biol. *123,* e60.

Vaisvila, R., Ponnaluri, V.K.C., Sun, Z., Langhorst, B.W., Saleh, L., Guan, S., Dai, N., Campbell, M.A., Sexton, B.S., Marks, K.*, et al.* (2021). Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome Res. *31,* 1280–1289.

Wang, T., and Kohli, R.M. (2021). Discovery of an Unnatural DNA Modification Derived from a Natural Secondary Metabolite. Cell. Chem. Biol. *28,* 97–104.e4.

Wang, T., Loo, C.E., and Kohli, R.M. (2021). Enzymatic approaches for profiling cytosine methylation and hydroxymethylation. Mol. Metab. 101314.

Wang, T., Fowler, J.M., Liu, L., Loo, C.E., Luo, M., Schutsky, E.K., Berríos, K.N., DeNizio, J.E., Dvorak, A., Downey, N.*, et al.* (2023). Direct enzymatic sequencing of 5-methylcytosine at single-base resolution. Nature Chem Bio *19,* 1004–1012.

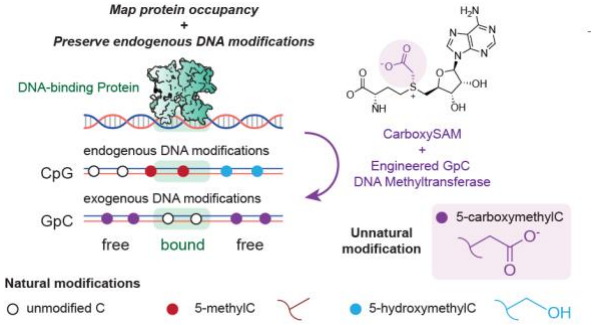
Wang, Y., Wang, A., Liu, Z., Thurman, A.L., Powers, L.S., Zou, M., Zhao, Y., Hefel, A., Li, Y., Zabner, J., and Au, K.F. (2019). Single-molecule long-read sequencing reveals the chromatin basis of gene expression. Genome Res. *29,* 1329–1342.

Wasney, M., and Pott, S. (2023). Simultaneous Measurement of DNA Methylation and Nucleosome Occupancy in Single Cells Using scNOMe-Seq. Methods Mol. Biol. *2611,* 231–247.

Zhang, A.P., Pigli, Y.Z., and Rice, P.A. (2010). Structure of the LexA-DNA complex and implications for SOS box measurement. Nature *466,* 883–886.

Zhang, X., and Bruice, T.C. (2006). The mechanism of M.HhaI DNA C5 cytosine methyltransferase enzyme: A quantum mechanics/molecular mechanics approach. Proceedings of the National Academy of Sciences of the United States of America *103,* 6148–6153.

Abstract Graphic



Figures

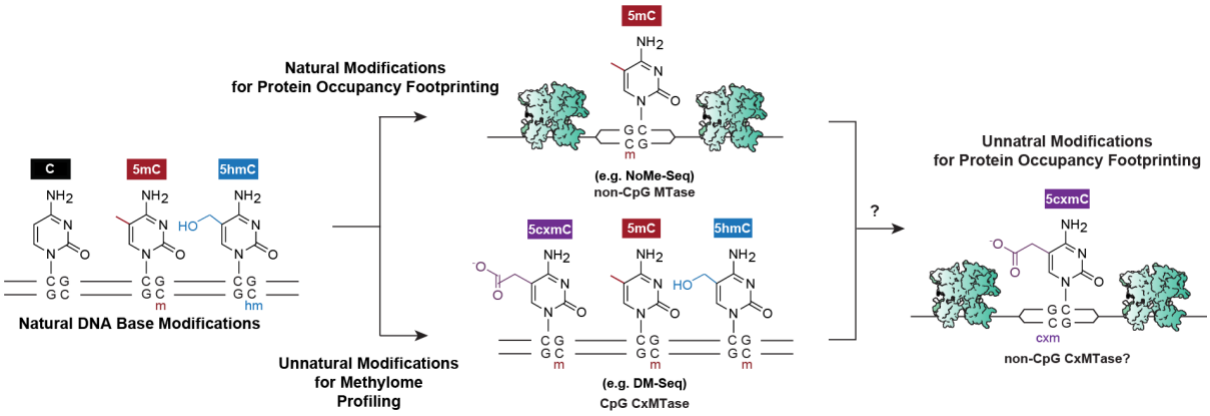


Figure 1. Unnatural modifications in non-native contexts for joint profiling of epigenetic bases and protein occupancy. Natural DNA base modifications in mammalian genomes include 5mC and 5hmC (left). 5mC has been used as to label protein-unbound DNA in methods like NOMe-Seq (top middle). 5cxmC is an unnatural DNA base modification resistant to enzymatic conversion, with installation of 5xcmC in CpG contexts having enabling the developement of DM-Seq (bottom middle). Explored here is the concept that 5cxmC installed in non-CpG contexts (right) could be used for protein occupancy mapping in a mode compatible with chemical and enzymatic sequecning approaches.

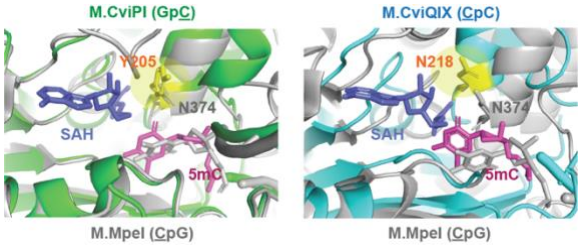


Figure 2. Homologous residue targets for mutagenesis to confer carboxymethylation. AlphaFold3-predicted structures of M.CviPI (left, green) and M.CviQIX (right, cyan) aligned with the crystal structure of M.MpeI (grey, PDB 4DKJ) bound with DNA with a 5mC (pink) and the side product SAH (dark blue).

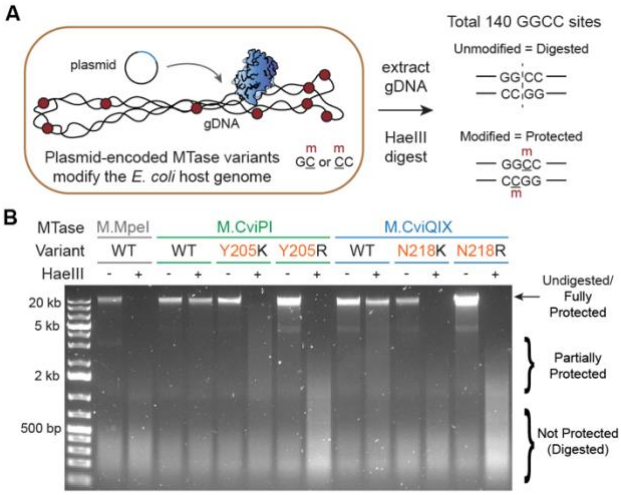


Figure 3. *In vivo* activity screen of MTase candidates. (A) Scheme of activity screen in *E. coli*. Genomic DNA (gDNA) is extracted from *E. coli* expressing the plasmid-encoded MTase variants. GGCC sites in the genome can be modified by active M.CviPI (GpC) or M.CviQIX (CpC) using intracellular SAM as a co-factor, while M.MpeI (CpG) is not expected to modify this site. The modified DNA is protected from restriction digest by HaeIII. (B) Agarose gel showing undigested or HaeIII digested gDNA recovered from cells expressing different MTase variants.

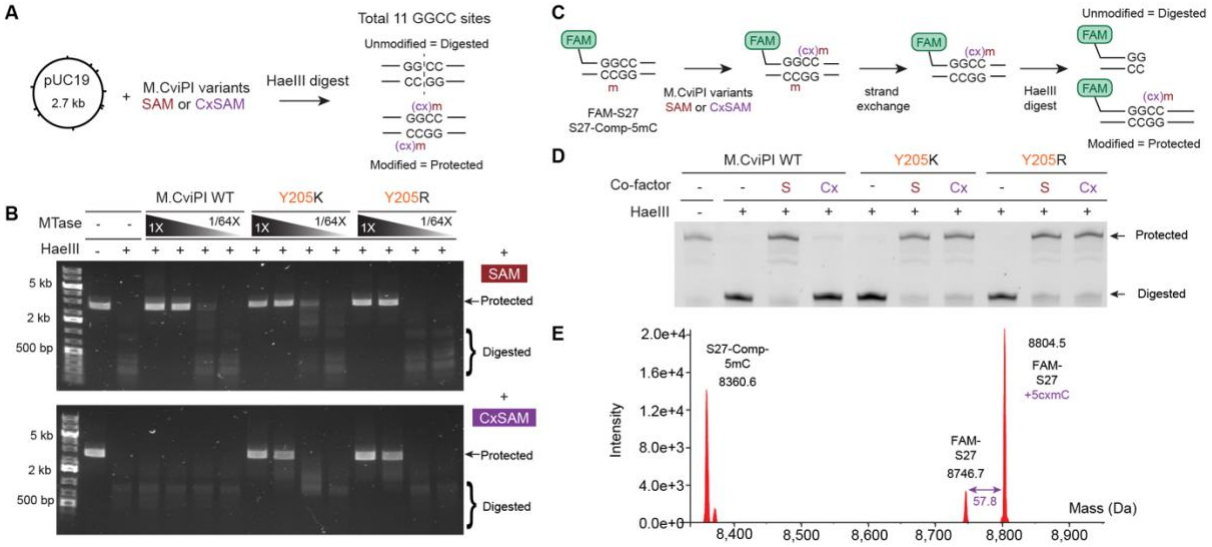


Figure 4. *In vitro* activity assays of MTase candidates. (A) Scheme of pUC19-based activity assay. The unmodified pUC19 substrate is incubated with M.CviPI variants and either SAM or CxSAM. The substrate is then purified and digested by HaeIII to report on modification status. (B) Agarose gel of restriction digest results. Modification is detected as protection against digestion. The starting concentration of all three variants were normalized and the enzyme was titrated with 4-fold serial dilutions from 1X to 1/64X while SAM/CxSAM concentration is fixed at 300 µM. (C) Scheme of oligonucleotide-based activity assay. The FAM labeled top strand (FAM-S27) containing a single GGCC site is annealed with a methylated complement. The duplex was incubated with M.CviPI variants and SAM or CxSAM. The methylated bottom strand is exchanged for an unmodified strand to permit evaluation of protection with HaeIII digestion. (D) Shown is the DNA Denaturing PAGE of duplexes analyzed by the activity assay where each M.CviPI variant was incubated with no (-), SAM (S), or CxSAM (Cx). (E) Intact LC/MS trace of oligonucleotide duplex treated with M.CviPI Y205K and CxSAM that displays the expected mass shift for the carboxymethylated oligonucleotide.

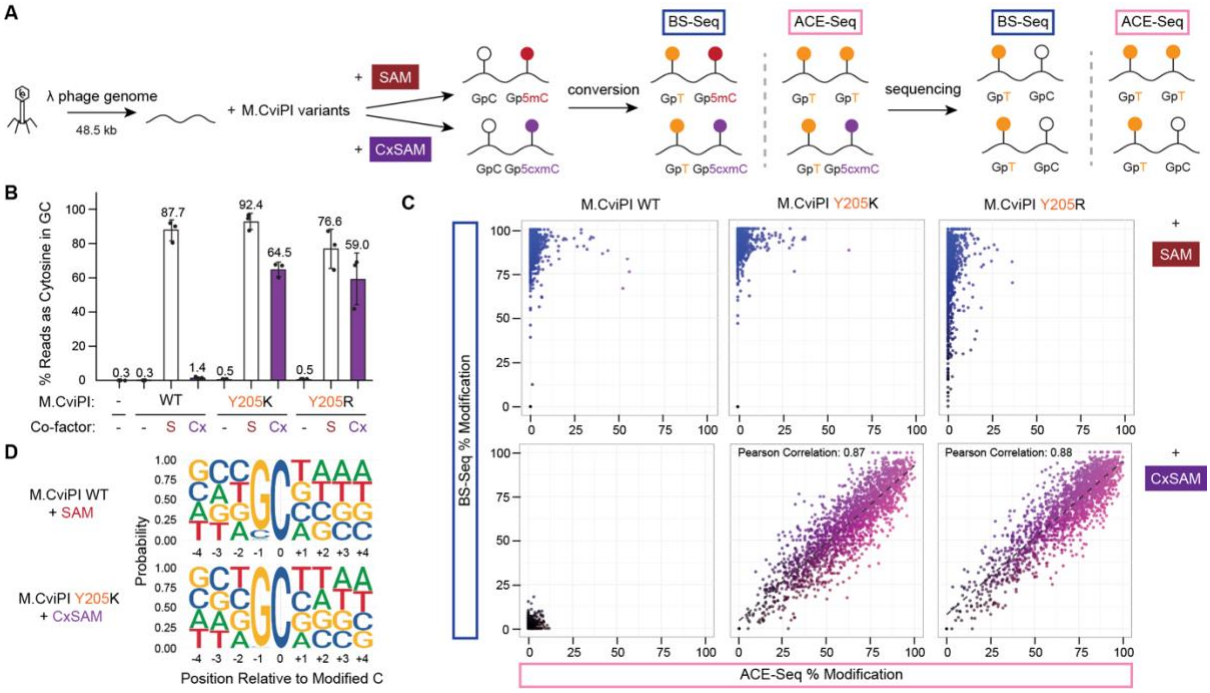


Figure 5. Base-resolution characterization of modification patterns on l phage genomic DNA. (A) Scheme of sequencing-based characterization. (B) Cytosine modification is measured by protection of GpC sites from C-to-T conversion (% reads as C) in GpC contexts. Shown are individual data points from independent replicates (n = 3), with +/- 1 standard deviation plotted and mean values shown above bars. (C) Correlation plots between BS-Seq and ACE-Seq after reaction of l phage gDNA with different M.CviPI and co-factor pairs. Each dot represents a single GpC site. Data is filtered for GpC sites with at least 15 total reads and randomly downsampled to 2000 sites to avoid overplotting. A strong Pearson correlation is indicated by a dashed line in the last two subplots. (D) Modification logos of different enzyme and co-factor pairs.

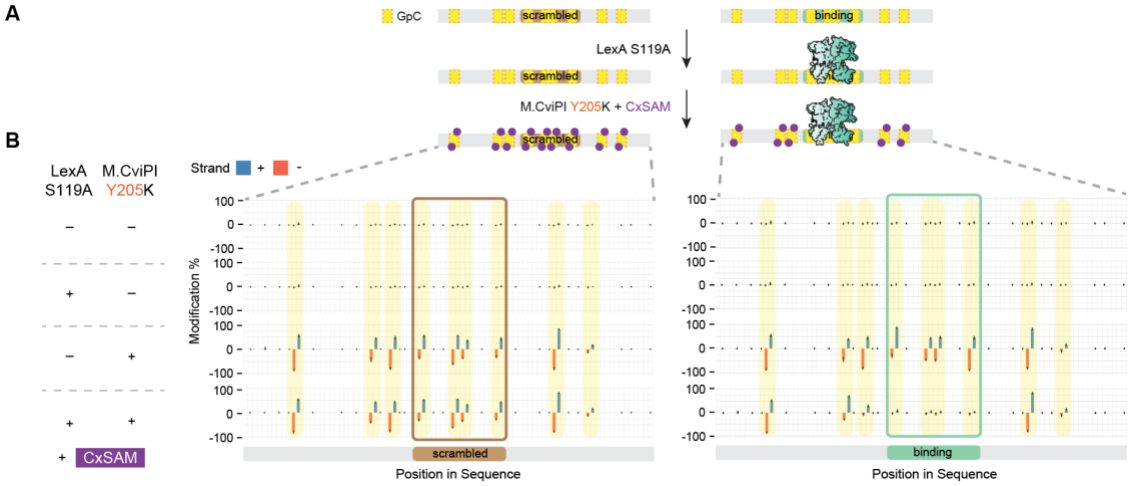


Figure 6. GpC footprinting assay for LexA-operator DNA binding. (A) Scheme of LexA-operator footprinting assay. (B) Modification level on + strand (blue, upward) and – strand (orange, downward) assessed by BS-Seq. Shown are individual data points from independent replicates (n = 3), with +/- 1 standard deviation plotted. The scrambled 20-bp LexA binding site is indicated by the brown colored box. The 20-bp LexA binding site is indicated by the green colored box.