Miura *et al. BMC Biotechnology* 2022, **22**(1):33
https://doi.org/10.1186/s12896-022-00765-3

BMC Biotechnology

RESEARCH ARTICLE Open Access

Identification, expression, and purification of DNA cytosine 5-methyltransferases with short recognition sequences

Fumihito Miura[1*], Miki Miura[1], Yukiko Shibata[1], Yoshikazu Furuta[2], Keisuke Miyamura[1], Yuki Ino[1], Asmaa M. A. Bayoumi[3], Utako Oba[1,4] and Takashi Ito[1]

Abstract

Background: DNA methyltransferases (MTases) are enzymes that induce methylation, one of the representative epigenetic modifications of DNA, and are also useful tools for analyzing epigenomes. However, regarding DNA cytosine 5-methylation, MTases identified so far have drawbacks in that their recognition sequences overlap with those for intrinsic DNA methylation in mammalian cells and/or that the recognition sequence is too long for fine epigenetic mapping. To identify MTases with short recognition sequences that never overlap with the CG dinucleotide, we systematically investigated the 25 candidate enzymes identified using a database search, which showed high similarity to known cytosine 5-MTases recognizing short sequences.

Results: We identified MTases with six new recognition sequences, including TCTG, CC, CNG, TCG, GCY, and GGCA. Because the recognition sequence never overlapped with the CG dinucleotide, MTases recognizing the CC dinucleotide were promising.

Conclusions: In the current study, we established a procedure for producing active CC-methylating MTases and applied it to nucleosome occupancy and methylome sequencing to prove the usefulness of the enzyme for fine epigenetic mapping. MTases that never overlap with CG dinucleotides would allow us to profile multiple epigenomes simultaneously.

Keywords: DNA methyltransferase, Recognition sequence, Methylome sequencing

Background

DNA methylation at position 5 of cytosine (5mC) is widely observed in the genomic DNA of living organisms and is one of the major epigenetic modifications. Classically, in mammalian cells, increased 5mC level was thought to lead to gene silencing, whereas the levels of gene body methylation positively correlate with the expression levels of genes [1–4]. Accordingly, the distribution patterns of 5mC are cell type-specific, and the genome-wide distribution of 5mC or methylome has always attracted much attention.

The deposition of 5mC in mammalian cells depends on three DNA methyltransferases: DNMT1, DNMT3A, and DNMT3B [2, 5, 6]. DNMT1 is a maintenance DNA methyltransferase that introduces methylation into hemimethylated CG dinucleotides. In contrast, DNMT3A and DNMT3B are de novo methyltransferases that introduce DNA methylation into non-methylated CG sites. Although some exceptions are known for pluripotent stem cells, oocytes, neurons, and glial cells, DNA methylation on non-CG sites rarely occurs in normal somatic

*Correspondence: fumihito@med.kyushu-u.ac.jp

[1] Department of Biochemistry, Kyushu University Graduate School of Medical Sciences, 3-1-1 Maidashi, Higashi-Ku, Fukuoka 812-8582, Japan
Full list of author information is available at the end of the article

tissues [7, 8]. Therefore, it is commonly accepted that 5mC is applied almost exclusively to CG dinucleotides, and the cytosines in contexts other than CG dinucleotides are unmethylated in most mammalian cells.

The deposition of 5mC is closely related to other epigenetic modifications. For example, 5mC is depleted on the promoters of actively transcribed genes, and trimethylation of lysine 4 of histone H3 (H3K4me3) is enriched [9]. Another example is the di- and tri-methylation of histone H3 lysine 36 (H3K36me2 and H3K36me3), which are positively related to transcriptional activity. H3K36me2 and H3K36me3 are enriched in gene bodies and recruit DNMT3A/B to introduce 5mC. Therefore, these modifications support gene body methylation [10, 11]. Because many epigenetic modifications are mutually related, simultaneous detection of DNA methylation with other epigenetic markers on the same DNA molecule would be useful. Establishing such a measurement would help adapt epigenetic analysis to more complex biological phenomena. However, most current methods for detecting epigenomes are unsuitable for simultaneous detection.

For example, chromatin immunoprecipitation followed by sequencing (ChIP-Seq) [12, 13] is the most commonly used technique for detecting epigenomes. In ChIP-Seq, the chromatin is first fragmented, the target epigenome on the fragmented chromatin is then immunoprecipitated, and DNA bound to the epigenome is sequenced. Because ChIP-Seq is based on chromatin fragmentation, the proximally located epigenomes on different nucleosomes should be separated in this first fragmentation step. Therefore, measuring the relationships among epigenetic modifications on a single DNA molecule is not practical with ChIP-Seq. Accordingly, fragmentation-based procedures are not suitable for measuring multiple epigenetic modifications simultaneously on a single DNA molecule.

There are some fragmentation-free procedures for epigenomic detection. For example, DNA methyltransferase (MTase) can be used as a probe to measure chromatin accessibility. Nucleosome occupancy and methylome sequencing (NOMe-Seq) is one of the first methods to measure chromatin accessibility with an MTase [14], and it has recently been combined with a nanopore sequencer [15]. The same principle was recently applied to Fiber-seq [16], a combination of non-specific adenine N-6 MTase M.EcoGII and a nanopore sequencer. Although chromatin accessibility requires the measurement of DNA–protein interactions of unspecified proteins, interactions between DNA and specific proteins can also be detected with MTase. A representative technology is DamID [17]. In DamID, dam methyltransferase fused with a protein of interest is expressed in target cells, and specific interactions of the fused protein with DNA can be measured by detecting DNA methylation introduced at the N-6 position of adenosine. Although methylation signals were detected using a methylation-sensitive restriction enzyme in the original DamID report [17], this is not an absolute requirement for detecting DNA methylation signals. Indeed, DNA methylation introduced with DamID was detected using single-molecule sequencers. Therefore, MTases enable fragmentation-free and simultaneous detection of multiple epigenomes on a single DNA molecule.

For epigenetic analysis based on MTases, the resolution is primarily determined by the frequency of the recognized sequences in the genome, and the length of the recognized sequence determines the frequency of sequences. As one of the most basic epigenome units, nucleosomes, are wrapped with approximately 150 base pairs (bp) of genomic DNA, the resolution required for epigenome mapping should be higher than 150 bp. Therefore, the recognition sequence must be 3 bp (64-bp resolution) or less to achieve sufficient resolution for epigenetic measurement. Many MTases have been identified from bacterial and viral sources, and the use of non-specific adenine N6 methyltransferases, such as M. EcoGII, can help achieve fine mapping of chromatin accessibility [16]. However, the technologies that use adenine N6-methylation (m6A) cannot be combined with PCR amplification because DNA replication causes the loss of methylation marks. Both sequencers from Oxford Nanopore Technologies and Pacific Biosciences for m6A detection require microgram quantities of DNA, whereas DNA recovered from target samples for epigenetic analysis is usually limited to less than a few tens of nanograms. Therefore, epigenetic mapping technologies using 5mC have a practical advantage of being combined with bisulfite sequencing or its enzymatic alternative to enable PCR amplification of the sequencing library. Accordingly, we focused on DNA cytosine C-5 MTases. However, the available MTases that can be used as probes for fine epigenome mapping are limited to only three in REBASE [18, 19]: M. SssI (CG), M. CviPI (GC), and M. CviPII (CCD) (the target cytosines are underlined). In addition, because of the intrinsic DNA methylation of the CG dinucleotide in mammalian cells, sequences fully (M.SssI) or partially (M.CviPI) overlapping with CG dinucleotides are problematic. Therefore, we launched a project to identify cytosine C-5 MTases with short recognition sequences that never overlap with CG dinucleotides.

The current study presents a systematic strategy for identifying cytosine C-5 MTases with short recognition sequences. This strategy enabled us to identify the MTases of six novel recognition sequences. In addition, of the identified MTases, we focused on the CC-recognizing

MTase because the non-overlapping specificity with CG dinucleotide is beneficial for epigenome mapping of mammalian cells. Further, we used the CC-recognizing MTase as a probe for NOMe-Seq to show its great potential for use for fine epigenetic mapping.

Results

Identification of cytosine C-5 MTases structurally homologous to M.CviPII and M.CviQIX

We identified two MTases that recognize C-5 cytosine, rather than the CG dinucleotide, in short sequences using REBASE [18, 19] (Fig. 1A). The first, M.CviPI (REBASE: 3772, GenBank: AAC64006.1) [20], which could recognize GC dinucleotides, has been commercialized by NEB. The other was M.CviQIX (REBASE: 3175) [21], annotated to methylate the first C of CCD trinucleotide (D is a degenerated expression of A, G, or C).

M.CviPI and M.CviPII resembled each other [22] but were identified to recognize different short sequences GC and CCD, respectively. Thus, we hypothesized that amino acid sequences structurally homologous to both M. CviPI and M. CviPII (and M. CviQIX) with a similar degree of difference between the proteins might recognize different short sequences. The BLASTP searches [23] using M.CviPI and M.CviPII as queries on non-redundant protein sequences (nr) [24] on the NCBI website identified hundreds of high-scoring hits for both. M.CviPI was identified as a homolog of M.CviPII with a score of 844 and vice versa. Thus, we roughly set the similarity cutoff score as 500. Using these criteria, 25 amino acid sequences homologous to the entire structure of the queries were identified (Fig. 1B and Additional file 1: Fig. S1), and the results of two blastp searches shared all genes (not shown). These genes were structurally close, as expected from the procedure used for identification (Fig. 1B and Additional file 1: Fig. S1).

Several differences were observed in the multiple alignments of the amino acid sequences (Fig. 1B). For example, N-terminal amino acid extension was evident in the CCD-recognizing enzymes M.CviPII and M.CviQIX (Fig. 1B and Additional file 1: Fig. S1). Amino acids 217 to 229 of the GC-recognizing enzyme M.CviPI seemed to be insertions (Fig. 1B and Additional file 1: Fig. S1). Using a phylogenetic tree based on multiple sequence alignments, we recognized a good diversity of the 25 genes; M.CviPI and M.CviQIX were reasonably separated into different clades (see below).

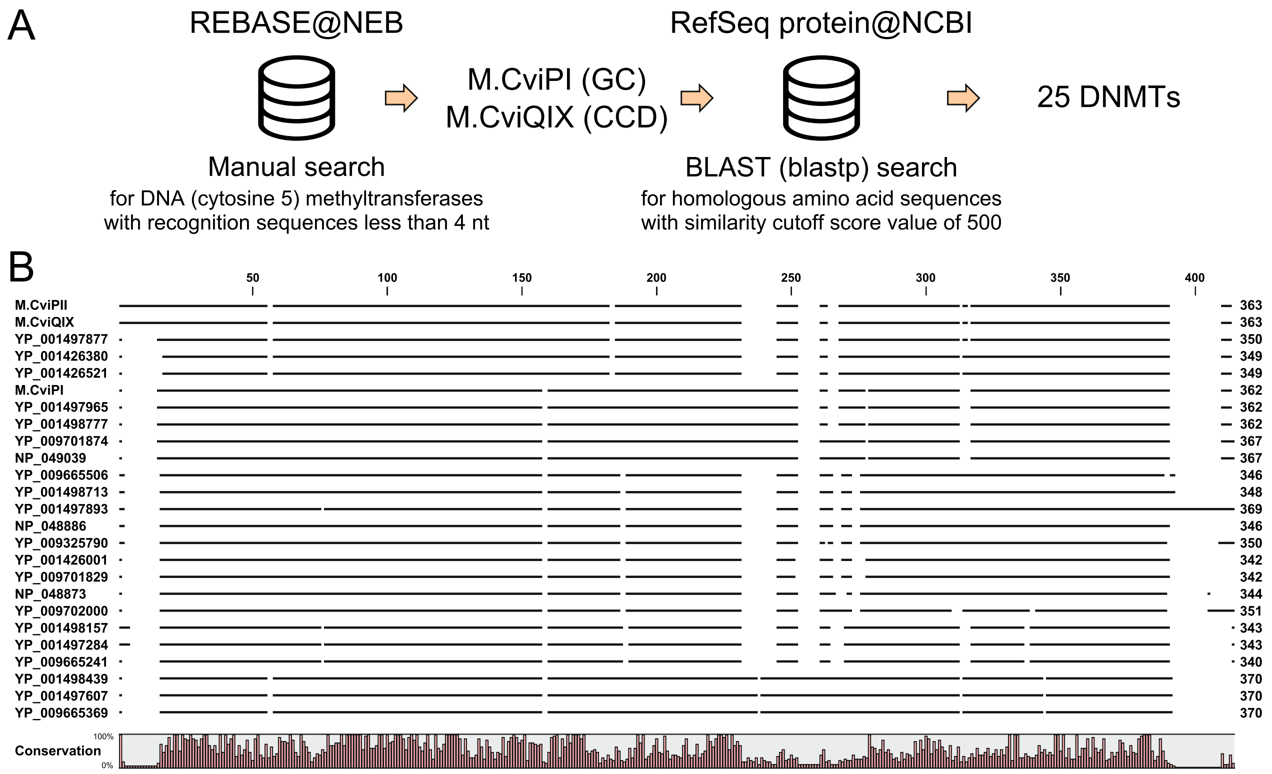

Fig. 1 Strategy to identify the MTase genes recognizing short sequence motifs **A** Scheme used to search the candidate MTases in the databases. **B** Multiple alignments of the 25 genes. Homologous regions are expressed with thick bars. The same alignment with the single amino acid resolution is provided in Additional file 1: Fig. S2

A dual-affinity strategy for the expression and purification of active MTases

Before expressing all 25 genes, we investigated *Escherichia coli* protein expression systems using M.CviPI as a model because the successful expression of M.CviPI was previously reported by another group [20]. At first, a commercially available cold-shock promoter system (pColdI) was used. Some growth defects were observed after subcloning M.CviPI into pColdI, even when we used the 5mC-tolerant bacterial strains NEB10β and T7Express that lacked *mcrA*, *mcrBC*, and *mrr*. The expression of recombinant M.CviPI in bacterial cells was good (Additional file 1: Fig. S2A), and the genomic DNA extracted from the cells was resistant to the methylation-sensitive restriction enzyme HaeIII, which recognized 5′-GGCC-3′, after the induction of gene expression (Additional file 1: Fig. S2B). The pColdI vector possesses an N-terminal-tagged His6 tag, so the recombinant protein could be purified via immobilized-metal affinity chromatography (IMAC). Because we wanted to combine M.CviPI with antibodies in future studies, fusion proteins of M.CviPI with the antibody-interacting domain of protein A, called ZZ [25], were also constructed. We found that M.CviPI fused with the ZZ domain showed higher expression levels in bacterial cells and higher purity after His6 tag-based purification, especially when ZZ was fused at the C-terminal of M.CviPI (Additional file 1: Fig. S2C), which could be partially explained by the protein solubilization effect of the ZZ domain [26]. Although MTase activity was strong enough with only His-tag-based affinity purification (Additional file 1: Fig. S2C), M.CviPI purified using only the IMAC column with Ni Sepharose resin (HisTrap HP column) contained non-negligible impurities (Additional file 1: Fig. S2D). To further purify the recombinant M.CviPI, we used a dual-affinity purification strategy via fusing the target protein with the Strep-II tag at its C-terminus (Fig. 2A). Additional affinity purification of M.CviPI with Strep-Tactin Sepharose (StrepTrap column) produced proteins with over 95% homogeneity (Fig. 2B) with strong DNA methylation activities (Fig. 2C). Therefore, we used the cold-shock promoter-based bacterial expression system with the N-terminal His6 tag, ZZ, and Strep-II tag at the C-terminus of the target protein in the following experiments (pCZS vector Additional file 1: Fig. S3A).

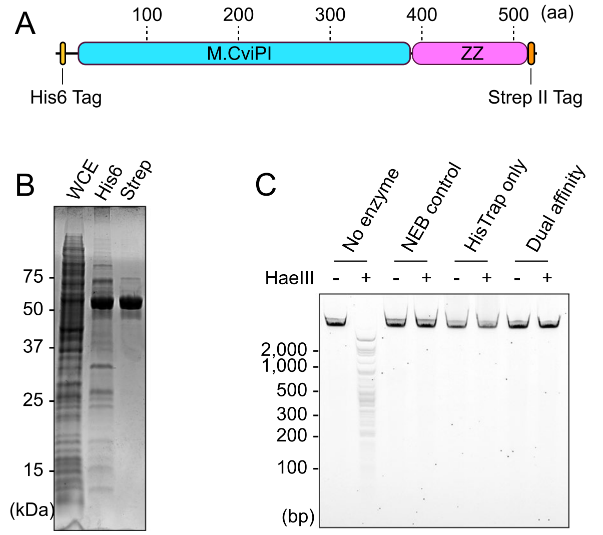

Fig. 2 Dual-affinity purification strategy used to purify the candidate MTases. **A** Structure of the expressed protein. **B** M.CviPI proteins purified only via HisTrap purification and dual-affinity purification combined with HisTrap and StrepTrap purifications. **C** DNA methylation activities of purified M.CviPI were assayed using the methylation-sensitive restriction enzyme HaeIII. Unmethylated lambda DNA (Promega) was used as the substrate DNA. As a control, M.CviPI purchased from NEB was included

N-terminal deletion enhances the MTase activity of M.CviQIX and M.CviPII

We also tried to produce an active M.CviQIX protein annotated to have MTase activity at the first C of the CCD trinucleotide. If the proteins expressed in the bacterial cells were active, the genomic DNA of the cells would become resistant to both HaeIII and HpaII digestion. Using methylation-sensitive restriction digestion assays, we screened suitable expression systems for M.CviQIX (Fig. 3A–C and Additional file 1: Fig. S4A and B). However, we could not detect strong DNA methylation activities in most bacterial cells with genes subcloned into several expression vectors (Fig. 3A–C). Strong MTase activity was observed in cells expressing M.CviQIX subcloned in the expression vectors pCTF and pCNS, which produced M.CviQIX fused with a trigger factor (TF) and a SUMO motif at the N-terminal, respectively (Fig. 3A). Therefore, we tried to purify the products of these N-terminal TF- and SUMO-tagged constructs. However, both proteins showed only weak MTase activity after affinity purification using the N-terminal His6 tag (not shown).

Although SUMO is said to have a protein solubilization effects [27], genomic DNA extracted from the bacterial cells with SUMO-tagged M.CviQIX at the C-terminal showed less resistance to methylation-sensitive restriction enzymes than N-terminally tagged ones (Fig. 3C), which indicated that the N-terminal tagging of M.CviQIX contributed to the production of active M.CviQIX in the bacterial cells. Additionally, the lack of any MTase activity for the affinity purified N-terminally tagged proteins indicated that the complete N-terminal structure inhibited MTase activity. Furthermore, we observed exceptional amino acid extensions at the N-terminal of both

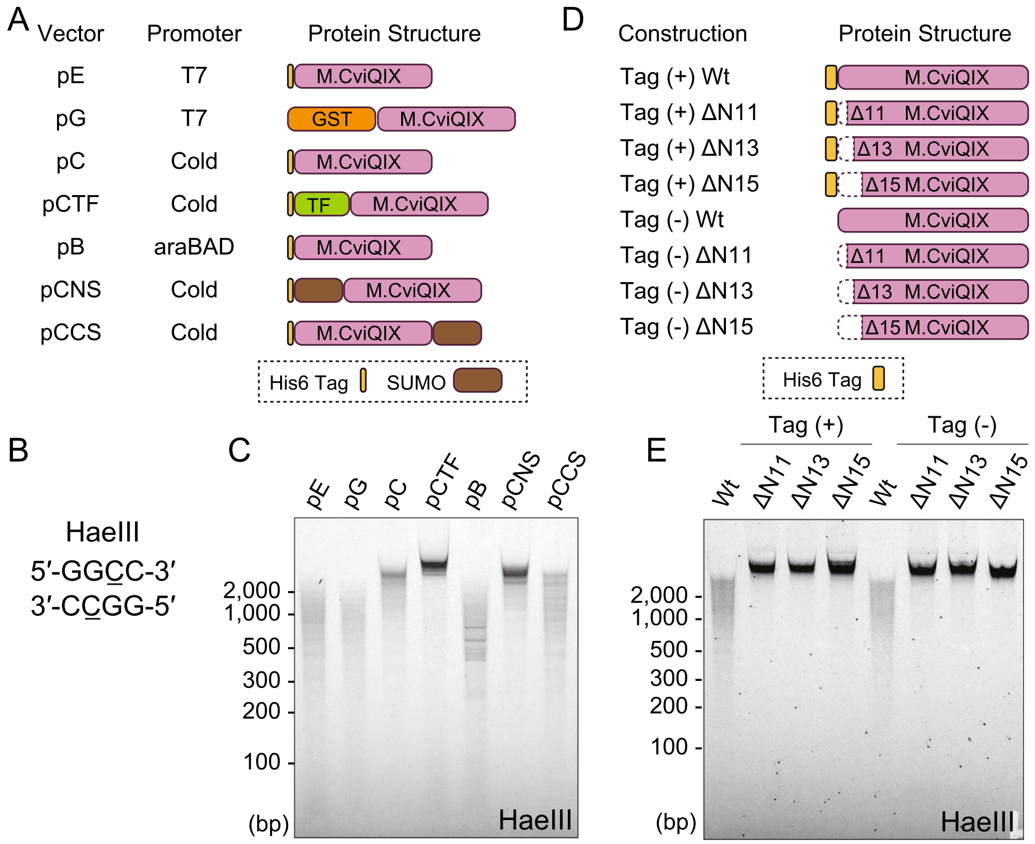

Fig. 3 N-terminal extension of M.CviQIX has an inhibitory effect on DNA methylation. **A** Schemes showing recombinant protein structures of M.CviQIX. **B** Recognition sequence of the restriction enzyme HaeIII. The recognition sequence contains the CC dinucleotide. If the first C is 5-methylated, HaeIII cannot cut the substrate DNA. **C** Genomic DNA digestion with HaeIII. DNAs extracted from bacterial cells that express the M.CviQIX proteins shown in A were digested with HaeIII and analyzed on E-Gel Ex. **D** Schemes showing M.CviQIX proteins with or without N-terminal deletions and vector-derived sequences. Tag denotes the vector-derived sequence. **E** Genomic DNA digestion of bacterial cells using the constructions in D

M.CviQIX and M.CviPII (Fig. 1B). Therefore, we hypothesized that the N-terminal extensions might cause a low methylation activity of purified M.CviQIX and M.CviPII.

As expected, deletion of the M.CviQIX extension enhanced the methylation of the host genome (Fig. 3D and E and Additional file 1: Fig. S4C). Because there were three methionine residues in the N-terminal extension, we speculated that translation initiation might be achieved through one of these residues. Thus, we prepared three N-terminal deletion mutants of M.CviQIX that lacked 11, 13, or 15 amino acids and compared the methylation activities of these proteins (Fig. 3D). Genomic DNA extracted from all three deletion mutants showed similar levels of resistance to digestion with the methylation-sensitive restriction enzymes HaeIII and HpaII (Fig. 3E and Additional file 1: Fig. S4E). Therefore, any of these three methionine residues could be used as an initiation codon to produce an active M.CviQIX. Note that the N-terminal sequence derived from the expression vector would have a minimal effect on this conclusion because we observed the same result with the constructs possessing no amino acid sequences belonging to the expression vector used (see no tag constructs, Fig. 3D and E and Additional file 1: Fig. S4D and E). We confirmed the same conclusion with the N-terminal deletion mutants of M.CviPII (Additional file 1: Fig. S4F and G). Therefore, the N-terminal amino acid sequences of M.CviQIX and M.CviPII had an inhibitory effect on DNA methylation activity.

Whole-genome bisulfite sequencing (WGBS)-based determination of the MTase recognition sequences

Next, we synthesized 22 additional MTase genes and aimed to subclone all 25 genes into the expression vector pCZS (Additional file 1: Fig. S3A). We used the 15 amino acid deletion (NDel15) mutants of M.CviQIX and M.CviPII. Fifteen out of the 25 MTase genes were successfully subcloned into pCZS (Additional file 2:

Table S1). In contrast, 10 genes could not be subcloned into the pCZS vector. In addition, we observed that some genes caused severe growth defects after transformation, suggesting leaky expression of the cold-shock promoter of pCZS under uninduced conditions. Therefore, to enhance the subcloning efficiency, we replaced the cold-shock promoter of pCZS with that of the araBAD system to obtain pBZS (Additional file 1: Fig. S3B) and successfully subcloned all MTase genes into pBZS (Additional file 2: Table S1).

We induced the expression of these MTases in *E. coli* cells, extracted the bacterial genomic DNA, and performed WGBS using our recent post-bisulfite adapter tagging (PBAT) protocol, tPBAT [28, 29] (Fig. 4A). Methylation levels were determined for each cytosine residue. The enrichment of specific nucleobases surrounding the methylated Cs was evaluated by calculating the position weight matrix weighted with methylation levels (PWM-M) and methylation logo (M-logo). For example, as shown in Fig. 4B, PWM-M and M-logo for M.CviPI indicated strongly enriched G at the -1 position of C, which indicated that M.CviPI recognized the GC dinucleotide. Using the same strategy, we successfully defined the recognition sequences for 24 out of the 25 MTases (Fig. 4C, Table 1, Additional file 2: Table S1, and Additional file 1: Fig. S5). Excluding overlaps, these recognition sequences comprised eight unique sequences (TCTG, CG, CC, CNG, TCG, GCY, GC, and GGCA; the methylated cytosines are underlined). Of these, six (TCTG, CC, CNG, TCG, GCY, and GGCA) were novel recognition sequences of MTases (Table 1).

Purification of MTases in vitro

Next, we attempted to purify the MTases. After cultivating bacterial cells and inducing protein expression, the target MTases were purified using the dual-affinity strategy employing His- and Strep-II-tag (Fig. 2). All purifications were performed under the same conditions to systematically evaluate the feasibility of active MTase purification, and the condition used was the one we empirically established to purify the active M.CviPI. Proteins were stored in the presence of 100 µg/mL bovine serum albumin (BSA), 10 mM DTT, and 50% glycerol to preserve enzymatic activity during storage at − 20 °C.

Most purified proteins appeared highly homogenous upon SDS-PAGE, as shown in Additional file 1: Fig. S6A. However, it was difficult to obtain sufficient yields of several proteins. We found that some of these MTases showed strong methylation activities when assayed with methylation-sensitive restriction enzymes and/or WGBS using lambda phage DNA as a substrate (Additional file 2: Table S1 and Additional file 1: Fig. S6B). As a result,

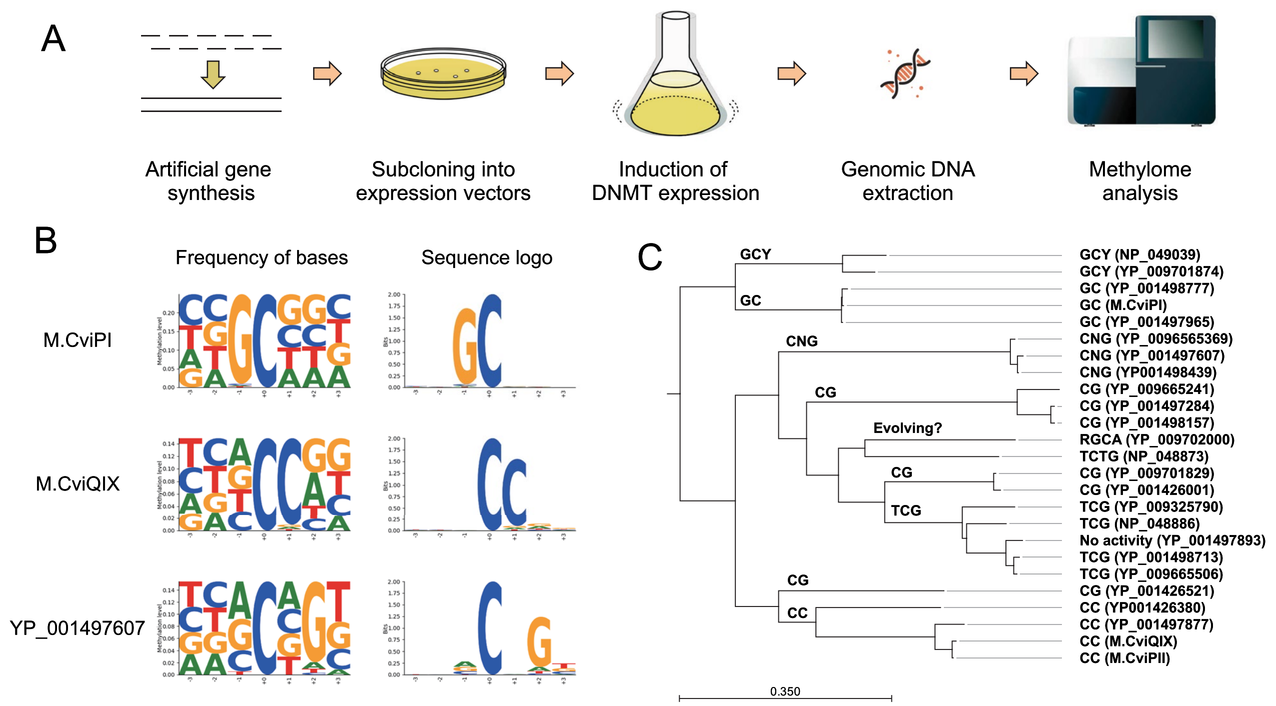

Fig. 4 Experimental strategy to define the recognition sequences of the candidate MTase. **A** Scheme used. **B** Examples of identified sequence motifs. **C** Phylogenetic tree was drawn based on the amino acid sequence similarities. The determined recognition sequences were arranged according to the gene name. The common recognition sequence of each clade is annotated at the top of the clade

Table 1 Summary of the analysis of candidate DNMTs

Entry	Gene	Recognition Sequence	pCZS		pBZS	
			Methylation motif	**In vitro activity**	**Methylation motif**	**In vitro activity**
1	M.CviPI	GC	GC	+	GC	+
2	M.CviPII (NDel15)	CC	CC	+	CC	−
3	M.CviQIX (NDel15)	CC	CC	+	CC	+
4	NP_048873	TCTG	TCTG	+	–	−
5	NP_048886	TCG	TCG	−	–	−
6	NP_049039	GCY	GCY	−	–	−
7	YP_001426001	CG	CG	+	CG	+
8	YP_001426380	CC	CC	−	–	−
9	YP_001426521	CG	CG	+	–	−
10	YP_001497284	CG	–	−	CG	−
11	YP_001497607	CNG	–	−	CNG	+
12	YP_001497877	CC	–	−	CC	−
13	YP_001497893	–	–	−	–	−
14	YP_001497965	GC	–	−	GC	+
15	YP_001498157	CG	–	−	CG	−
16	YP_001498439	CNG	–	−	CNG	+
17	YP_001498713	TCG	TCG	−	–	−
18	YP_001498777	GC	–	−	GC	−
19	YP_009325790	TCG	TCG	+	–	−
20	YP_009665241	CG	–	−	CG	−
21	YP_009665369	CNG	–	−	CNG	+
22	YP_009665506	TCG	TCG	−	–	−
23	YP_009701829	CG	–	−	CG	+
24	YP_009701874	GCY	GCY	+	GCY	−
25	YP_009702000	RGCA	–	−	RGCA	+

16 active MTases out of 19 were purified (Additional file 2: Table S1). Some of these proteins lost their activity during purification and storage. We also observed batch-to-batch differences in the activities and purities of these proteins. Therefore, to use these MTases in experiments, further optimization was required for their expression, purification, and storage.

NOMe-Seq with a CC-recognizing MTase produces similar methylation patterns as GCMT

Of the six newly identified recognition sequences (TCTG, CC, CNG, TCG, GCY, and GGCA), the CC dinucleotide was promising because the short recognition sequence enabled fine epigenome mapping at approximately 10-bp resolution. Therefore, we aimed to establish an MTase that recognizes CC dinucleotide as a probe for epigenome mapping. We purified the four CC-methylating MTases, M.CviPII, M.CviQIX, YP_001426380, and YP_001497877 (Table 1 and Additional file 2: Table S1), and found that the N-terminal deletion mutant of M.CviQIX expressed under the control of the cold-shock promoter showed the highest yield and activity (not shown). The purified mutant M.CviQIX is hereafter referred to as CCMT. We investigated the salt concentration in the CCMT reaction and found that 50 mM sodium chloride was the best condition for stringent DNA methylation (Additional file 1: Fig. S7). Notably, M.CviQIX recognizes CC and CCD [21] and REBASE [18, 19], respectively. Although CCMT recognized every trinucleotide (CCA, CCC, CCG, and CCT) and introduced a methyl moiety into the first C of all (Additional file 1: Fig. S5), we observed slightly weak methylation of CCC trinucleotides in some reaction conditions (Additional file 1: Fig. S7), which might cause a discrepancy between the results reported in the literature and databases.

Next, we used CCMT for NOMe-Seq of formaldehyde-crosslinked budding yeast (*S. cerevisiae*) nuclei. Formalin-fixed yeast spheroplasts were permeabilized with an NP-40-containing buffer and treated with CCMT. Genomic DNA was recovered using our recently developed procedure for formalin-fixed samples (Oba U et al., unpublished). Finally, the purified DNA was used for

WGBS based on the tPBAT protocol [28]. Because the budding yeast cell did not have any intrinsic DNA methylation, all DNA methylation signals were assumed to be introduced by the MTase treatment. As a control, we performed NOMe-Seq using M.CviPI purchased from NEB (GCMT).

We calculated the methylation levels and compared GCMT- and CCMT-treated nuclei; these two datasets showed similar methylation patterns. (Fig. 5A and B, Additional file 1: Table S2). The aggregation plots of CCMT and GCMT signals centered at several genomic features showed a similar appearance (Fig. 5C and D and Additional file 1: Fig. S8A). We observed sharp DNA methylation peaks in the promoter regions of the genes (Fig. 5B). These strong DNA methylation signals colocalized well with the signals detected using ATAC-Seq; therefore, these signals were thought to indicate open chromatin regions. The peaks detected upon CCMT- and GCMT-based NOMe-Seq and ATAC-Seq were well-colocalized (Additional file 1: Fig. S8B).

The DNA methylation levels at the CC and GC dinucleotides aggregated well at the promoter-proximal regions (Fig. 5C and D). Undoubtedly, these signals indicated that nucleosome-free regions were frequently observed at the gene promoters. The promoter-proximal DNA methylation signal at the − 100 to − 250 region of the transcription start site appeared to be inversely correlated with the low signal region detected via MNase-Seq (Fig. 5E). The aggregated DNA methylation signals at the transcription start sites showed a periodicity of 160 to 170 bp, and were inversely correlated with those observed using MNase-Seq (Fig. 5B–E). These observations recapitulated the known properties of NOMe-Seq and MNase-seq [30].

CCMT-based NOMe-Seq achieves fine mapping of the mammalian epigenome

Next, we applied CCMT-based NOMe-Seq to analyze IMR-90 mammalian cells. We obtained high-quality methylome data with 24.5 × and 25.7 × coverage for CCMT-treated and untreated cells, respectively (Additional file 1: Table S3). As expected, intrinsic DNA methylation levels at the CG sites were almost identical between the datasets (Fig. 6A and B and Additional file 1: Fig. S9A and B). In contrast, DNA methylation levels at the CC sites increased only in the CCMT-treated nuclei (Fig. 6C and D). These results indicated that intrinsic CG methylation and CCMT-induced methylation at the CC dinucleotide could be explicitly separated.

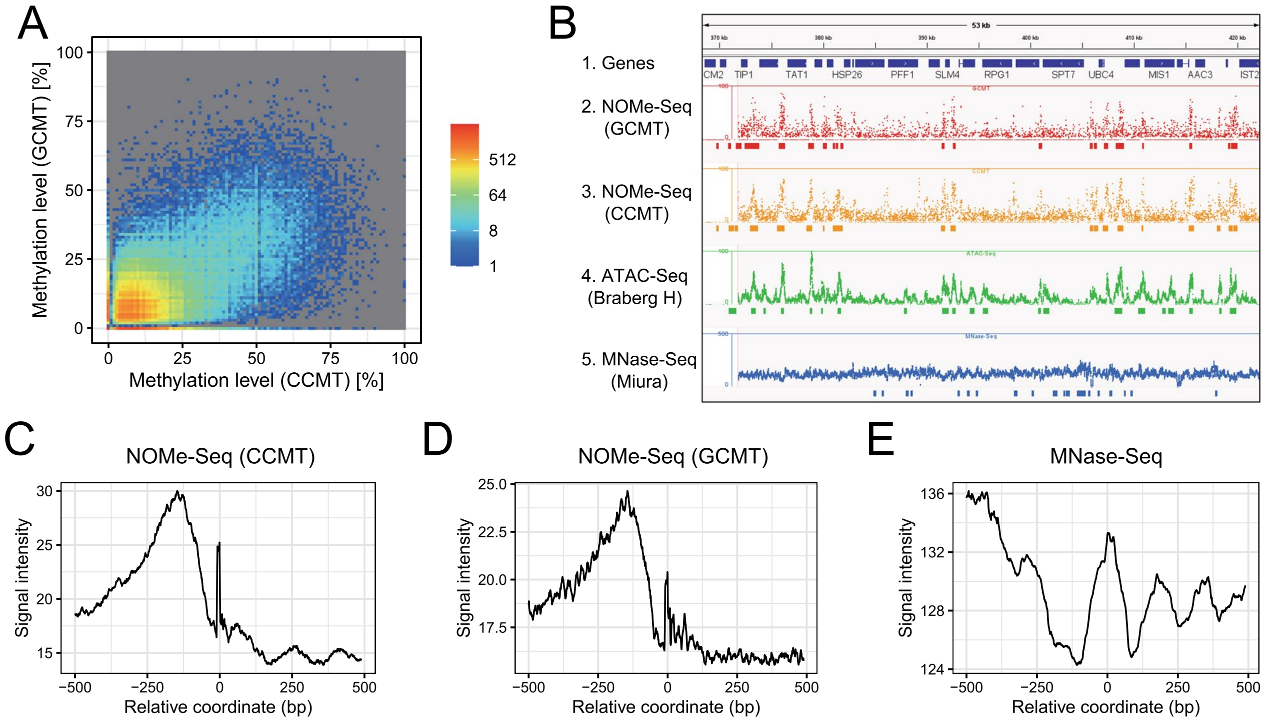

Fig. 5 NOMe-Seq of fixed yeast nuclei. **A** Comparison of methylation levels between CCMT- and GCMT-treated nuclei. Bin size is 10 bp. **B** Genome browser screenshot showing annotated yeast genes (Track 1), DNA methylation levels at the GC sites (Track 2, NOMe-Seq using GCMT) and CC sites (Track 3, NOMe-Seq using CCMT), read coverage of ATAC-Seq (Track 4, SRR12926697) [36], and MNase-Seq (Track 5, SRR6729489) [37]. **C–E** Aggregation plots around the initiation codons of yeast genes. DNA methylation signals at CC (**C**) and GC (**D**) sites, as well as read coverage of MNase-Seq (**E**) (SRR6729489), are shown

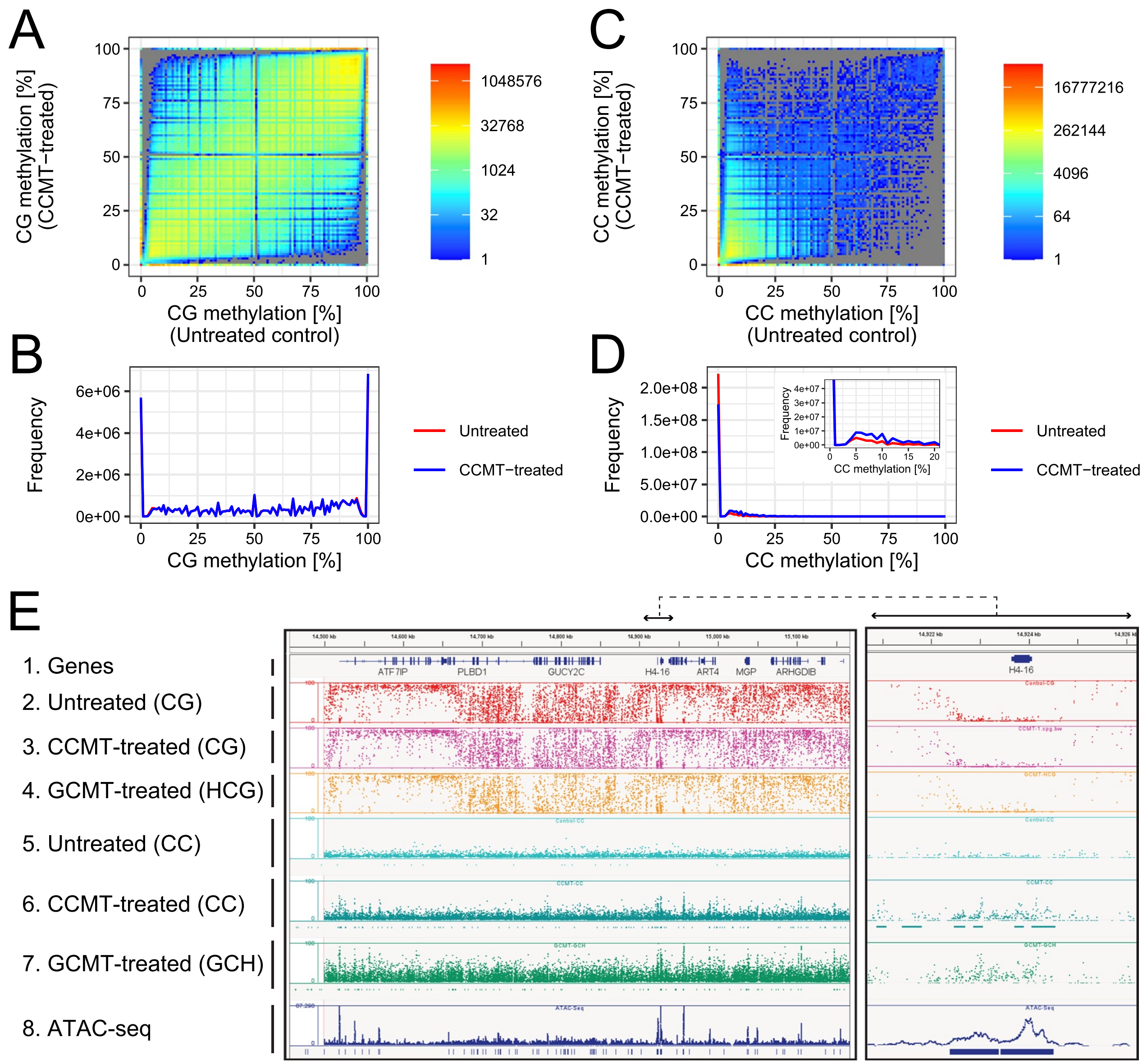

Fig. 6 NOMe-Seq of IMR-90 cells. **A**–**D** Comparisons of DNA methylation levels between CCMT-treated and untreated nuclei. Intrinsic DNA methylation at the CG sites (A and B) and methylation at CC sites introduced via CCMT in vitro (C and **D**) are shown. **E** Genome browser screenshots showing annotated human genes (Track 1), intrinsic methylation levels (Track 2, 3, and 4 for untreated, CCMT-treated, and GCMT-treated ones, respectively), MTase-induced methylations (Track 5, 6, and 7 for untreated, CCMT-treated, and GCMT-treated, respectively), and read coverage of ATAC-Seq (SRR7765313) are shown (Track 8). The right panel is the enlarged view of the indicated location of the left panel. For the tracks showing MTase-induced methylation patterns (Track 5, 6, and 7), peaks detected with macs2 are indicated at the bottom of each track

Then, we visually inspected the two datasets using a genome browser. We observed several large domains with different intrinsic DNA methylation levels at CG sites in both datasets (Fig. 6E). These domains were undoubtedly highly methylated regions and partially methylated domains, the existence of which has been described for IMR-90 cells [31]. The comparison between CCMT-based NOMe-Seq and untreated control indicated that these two datasets almost had the same qualities in terms of the intrinsic DNA methylome. In contrast, we observed strong and sharp methylation peaks for CC methylation in the treated samples, whereas none was observed in the untreated control (Fig. 6E). These peaks largely colocalized with the peaks identified using ATAC-Seq. Therefore, the peaks observed in the CCMT-induced methylation signals indicated highly accessible chromatin regions (Fig. 6E). These results indicated that CCMT-based NOMe-Seq simultaneously measures both the intrinsic methylome and chromatin accessibility of mammalian cells.

We compared the CCMT-based NOMe-Seq with the previously established GCMT-based procedure. We treated IMR-90 cells with GCMT and obtained methylome data with exactly similar procedure conducted for CCMT. We obtained high-quality methylome data with 22.7 × coverage (Additional file 1: Table S3). Note that, to exclude ambiguous GCG sites, only DNA methylation levels in HCG and GCH residues were examined for intrinsic (CG) and artificially induced (GC) DNA methylation, as described by Kelly et al. [14]. The patterns of intrinsic DNA methylation levels were indistinguishable between CCMT- and GCMT-based NOMe-Seq (Fig. 6E). In addition, the artificially induced DNA methylation levels at CC and GC sites showed almost similar patterns between the datasets (Fig. 6E). Although the methylation levels were higher for GCMT-based NOMe-Seq than for CCMT-based NOMe-Seq, the methylation levels of NOMe-Seq could be adjusted by fine-tuning of the MTase amount used, as previously indicated (15). These results indicated that both CCMT- and GCMT-based NOMe-Seq produce almost similar qualities of data.

To determine whether CCMT-based NOMe-Seq could also detect characteristic aggregation patterns of GCMT-induced DNA methylation around several genomic features [14], we performed the same analysis using the CCMT-based NOMe-Seq data. Methylation patterns around CTCF binding sites were observed in both CCMT- and GCMT-based NOMe-Seq but not in control MTase-untreated WGBS (Additional file 1: Fig. S9C–E). We also found a perk signal in both aggregation plots for GCMT and CCMT around the transcription start sites (Additional file 1: Fig. S9F–H) and histone H3 lysine 4 trimethylation peaks (Additional file 1: Fig. S9I–K). These results also supported the usefulness of CCMT in epigenome mapping.

Ideally, while 233,643,193 GCH sites were available for GCMT-based NOMe-Seq, 302,651,847 sites were usable for epigenome mapping of the human genome with CCMT (Table 2). When Cs covered with at least 10 reads were included in the analysis, 215,834,149 sites were available for CCMT in the WGBS data (see untreated control, Table 2), whereas 165,985,668 sites could be used for GCMT (Table 2). These results indicated that about 30% more sites were available with CCMT than with GCMT. The difference in the number of data points directly affected the resolution; the theoretical resolution of the CCMT-based NOMe-Seq (10 bp) was better than that of the GCMT-based NOMe-Seq (13 bp) (Table 2). Although the differences in the analytical resolution between the CCMT- and GCMT-based NOMe-Seq seemed to have little effect on the identification of accessible chromatin regions (Fig. 6E), higher analytical resolution would be beneficial for the fine mapping of the epigenetic state of chromatin.

Discussion

Despite the increasing demand for MTases as tools for mapping the epigenome, the available enzymes are limited, especially for cytosine C5 MTases of short recognition sequences. However, genome sequences have been determined for many bacterial and viral strains. Therefore, we hypothesized that MTases of such short recognition sequences already exist in databases, although they are still experimentally unidentified.

In this study, we conducted a screening to identify the DNA cytosine C5 MTases with short recognition sequences. We devised a strategy that combines database search, artificial gene synthesis, and WGBS. We successfully determined the recognition sequences of most MTases and purified several MTases with DNA methylation activities. Of the MTases identified, we chose one CC-recognizing enzyme and established it as a new probe for epigenome mapping.

With advanced DNA sequencing and artificial gene synthesis technologies, we have been able to

Table 2 Resolution of the NOMe-Seq analysis

Sequence motif	Genome size (hg19)	Number of motifs in hg19	Analytical resolution (bp)	Number of motifs with ten or more read coverage					
				Control (No MTase)		NOMe-Seq (CCMT)		NOMe-Seq (GCMT)	
CG	3,137,161,264	57,400,172	55	37,593,561	65.5%	35,968,639	62.7%	32,224,228	56.1%
HCG		43,525,732	72	28,851,898	66.3%	27,661,596	63.6%	24,764,078	56.9%
GC		247,517,634	13	174,727,329	70.6%	169,559,103	68.5%	151,895,283	61.4%
GCH		233,643,193	13	165,985,668	71.0%	161,252,061	69.0%	144,435,133	61.8%
CC		302,651,847	10	215,834,149	71.3%	208,676,370	68.9%	187,023,053	61.8%

systematically evaluate the recognition sequences of candidate MTases. Similar screenings can be easily set up and applied to any amino acid sequence in the database. During the preparation of this manuscript, Ilinsky et al. showed a similar strategy to determine the recognition sequences of MTases [32]. They combined the expression of artificially synthesized genes in *E. coli* cells and sequenced the bacterial genomic DNA using a nanopore sequencer [32]. In the current study, in contrast, we used WGBS because we thought that the reliability of 5mC detection with a nanopore sequencer is currently limited, especially for residues other than the CG dinucleotide. These two successful determinations of recognition sequences using modern DNA sequencing technology would accelerate the identification of unique MTases.

To identify MTases of short recognition sequences, we focused on the mutual sequence similarity of the two identified MTases of different short recognition sequences. We selected structurally similar amino acid sequences in the public database and revealed that these proteins recognize and methylate short sequences. Therefore, the similarity-based approach was effective. Although we have focused on the 25 amino acid sequences registered in RefSeq, many amino acid sequences in the database remain to be identified. Therefore, similar screening should be conducted in the future.

In mammalian cells, cytosine in the CG dinucleotide is methylated by endogenous MTases. In contrast, GCMT methylates cytosine in GC. Based on the difference in the recognition sequences, GCMT has been used for the simultaneous detection of endogenous DNA methylation and chromatin accessibility in NOMe-Seq [14]. However, at the GCG trinucleotide, one cannot discriminate the causative MTases for methylation and, because of this, it is difficult to exploit the full resolution of GCMT for epigenome mapping. Therefore, MTases with non-overlapping recognition sequences with CG dinucleotides would be beneficial. Thus, CCMT is promising because its recognition sequence does not overlap with CG.

The current study has shown the benefits of non-CG overlapping MTases in epigenetic mapping for the simultaneous detection of intrinsic DNA methylation. Similarly, additional MTases that never overlap with CG dinucleotides would expand our ability to profile multiple epigenomes simultaneously. Thus, MTases that recognize CA and CT would be important because these two sequences never overlap with CG and CC. Therefore, we are interested in the two MTases NP_048873 and YP_009702000 that recognize T$\underline{\text{C}}$TG and GG$\underline{\text{C}}$A, respectively. Additional screening for genes, particularly homologous to these two MTases, could help identify other MTases that recognize CA and CT residues, which cannot be identified using previous methods.

Conclusions

In this study, we have screened MTases similar to both M. CviPI and M. CviQIX to find several new MTases of short recognition sequences. We focused on MTases recognizing the CC dinucleotide or CCMT and established a method for producing active CCMT. Using budding yeast and human cell lines, we showed the practical usefulness of CCMT for fine epigenetic mapping.

Methods

Sequence analysis

The amino acid sequences of M. CviPI and M. CviQIX were downloaded from REBASE (http://rebase.neb.com/rebase/rebase.html) [18, 19]. Sequence analyses, including the blastp search, the calculation of multiple alignments, and phylogenetic tree drawing, were performed using the CLC Main Workbench (Qiagen, Hilden, Germany).

Note that there were some ambiguities in relation to M. CviQIX. The name M.CviQIX was assigned to an unidentified enzyme that might methylate the first C of the CC dinucleotide [21]. Although M.CviQIX was originally isolated from the Chlorella virus strain NY-2A [21], the NY-2A genome sequence (NC_009898) did not contain M.CviQIX. The amino acid sequence of M. CviQIX was unavailable in REBASE, but the sequence we downloaded from REBASE 2 years ago (May 17, 2019) was identical to an MTase found in the genome of a different virus strain AR158 (DQ491003, Gene ID C619L or RefSeq: YP_001498700). M.CviQIX was almost identical to M.CviPII (GenBank: AAV84097) [22]; only 3 out of 363 amino acids were different between M.CviQIX and M.CviPII. As expected from the high similarity, M. CviPII was identified to methylate the first C of CCD trinucleotides [22].

Artificial gene synthesis and subcloning into expression vectors

The genes encoding the MTases were synthesized by Eurofin Genomics Inc. (Tokyo, Japan, for M.CviPI, M.CviPII, and NP_048886) and GenScrypt (Tokyo, Japan, for the other genes) with codon optimization for *E. coli* (Additional file 1: Supplementary Information). First, the genes were synthesized with no internal BamHI and EcoRI recognition sequences and with sequences at the 5’ and 3’ terminals, respectively. Then, the gene fragments were subcloned into the BamHI-EcoRI site of pCZS and pBZS (Additional file 1: Fig. S1).

Small-scale induction of MTases and bacterial genomic DNA extraction

The bacterial expression vectors pCZS and pBZS (Additional file 1: Fig. S3) carrying a gene encoding methyltransferase were introduced into T7Express (New England Biolabs, Ipswich, MA, USA). The cells were inoculated in 3 mL of LB medium (1% [w/v] tryptone, 0.5% [w/v] yeast extract, and 0.5% [w/v] sodium chloride) with 100 µg/mL carbenicillin (Nacalai Tesque, Kyoto, Japan) and grown at 37 °C overnight with shaking at 200 rpm. Fifty microliters of each cell suspension was diluted in 3 mL of LB medium with carbenicillin, and the cells were cultivated at 37 °C for 3 h with shaking. For pCZS, the cell suspension was cooled in ice-cold water for 30 min, and 3 µL of 1 M isopropyl β-D-1-thiogalactopyranoside (IPTG, Nacalai Tesque) was added to induce protein expression. For pBZS, 3 µL of 30% (w/v) L-arabinose was added to induce protein expression. Cells were cultivated with shaking at 200 rpm and 16 °C overnight (pCZS) or 37 °C for 3 h (pBZS). The cells were collected via centrifugation at 2500 × *g* for 15 min and used for DNA purification using the DNeasy Blood & Tissue Kit (Qiagen), following the optional protocol for gram-negative bacteria provided by the manufacturer. The genomic DNA concentration was determined using the Qubit dsDNA BR Assay Kit and Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA, USA).

Analysis of DNA methylation with restriction enzyme digestion

Genomic DNA (100 ng) was added to a 10-µL reaction containing 1 × CutSmart Buffer and an appropriate restriction enzyme shown in Table 1. The reaction mixture was incubated at 37 °C for 1 h, and the enzyme was heat-inactivated at 70 °C for 10 min. After restriction enzyme treatment, DNA was analyzed using an E-Gel Power SNAP system on a 2% E-Gel Ex gel (Thermo Fisher Scientific).

Purification of recombinant MTases

The bacterial cells transformed with pCZS or pBZS expressing an MTase-encoding gene were inoculated in 30 mL of LB medium with 100 µg/mL carbenicillin and cultured at 37 °C overnight with shaking at 150 rpm. The seed culture was transferred into a 2-L Erlenmeyer flask with baffles containing 1 L of LB medium with carbenicillin and further cultivated at 37 °C for 4 h, shaking at 100 rpm. For pCZS, the flask was cooled in ice-cold water for 30 min, and protein expression was induced by adding 230 mg of IPTG per 1 L of medium. For pBZS, 3 g of L-arabinose per 1 L of medium was added to induce protein expression. The flask was incubated at 16 °C overnight (pCZS) or 37 °C for 4 h (pBZS) with shaking at 100 rpm. The cells were collected via centrifugation at 2500 × *g* for 15 min, and collected cells were frozen at − 80 °C until use.

The bacterial cells were resuspended in 20 mL of HisTrap buffer A (20 mM sodium phosphate, pH 7.5, 800 mM NaCl, 20 mM imidazole, 1 mM DTT, and 10% glycerol) containing a protease inhibitor cocktail (Nacalai Tesque) and disrupted via sonication. The lysate was cleared via centrifugation at 15,000 × *g* for 15 min and filtered using a 0.45-µm syringe filter. The first-round affinity purification was performed using a 5-mL HisTrap HP column (Cytiva, Marlborough, MA, USA) with HisTrap buffer A and HisTrap buffer B (20 mM sodium phosphate, pH 7.5, 800 mM NaCl, 200 mM imidazole, 1 mM DTT, and 10% glycerol). The fraction containing the target protein was further purified using 1-mL StrepTrap HP (Cytiva) with StrepTrap buffer A (100 mM Tris–HCl, pH 8.0; 150 mM NaCl; 1 mM EDTA; and 1 mM DTT) and StrepTrap buffer B (100 mM Tris–HCl, pH 8.0; 150 mM NaCl; 1 mM EDTA; 1 mM DTT; and 2.5 mM d-desthiobiotin). Chromatographic purification was performed using the AKTA Start System (Cytiva). The dual-affinity purified protein was concentrated via ultrafiltration using an Amicon Ultra-4 30 K device (Millipore), supplemented with BSA and glycerol to obtain final concentrations of 100 µg/mL and 50% (v/v), respectively, and stored at − 20 °C until use.

WGBS

WGBS was performed following the PBAT strategy [29] with recent improvements using a highly efficient single-stranded DNA ligation technique [28]. The molar concentration of the sequencing library was determined using the library quantitation kit from Takara Bio Inc. (Shiga, Japan). Small-scale sequencing was performed using Illumina MiSeq with the MiSeq v3 Reagent Kit in the paired-end mode for 2 × 75 cycles (for bacterial and budding yeast cells) or with the MiSeq v2 Reagent Kit nano on the paired-end mode for 2 × 150 cycles (for lambda DNA). For large-scale sequencing of mammalian cells, paired-end sequencing for 2 × 150 cycles using HiSeq X Ten was performed by Macrogen Japan Corp. (Tokyo, Japan). The reads were mapped to a reference genome (J02459 for lambda phage, NC_000913 for bacterial cells, sacCer3 for yeast cells, and hg19 for human cells) with BMap (https://github.com/FumihitoMiura/Project-2). The mapped alignments were summarized using in-house developed pipelines (https://github.com/FumihitoMiura/Project-2).

Calculation of the PWM-M and M-logo

From the given genome sequence, a set of every subsequence of length $2\,l+1$ with a C nucleobase at the center of the subsequence was taken, and PWM-M elements were calculated using the following equation:

$$M_{ij} = \frac{\prod_{k=1}^{N}\left\{m_k \times I_i\left(S_{(K+j)}\right)\right\}}{N}$$

where i is nucleobase A, C, G, or T in the matrix, and j denotes the relative coordinate from the center of the subsequence. N is the total number of C nucleobases in the given genome sequence, and K is the genomic coordinate of kth C. S denotes the nucleobases A, C, G, or T in the genomic sequence, and $S_{(K+j)}$ indicates the nucleobase at coordinate $K+j$. $I_i(S)$ is an indicator function where $I_i(S)$ is 1 if $S=i$, and 0 otherwise. m_k denotes the methylation level at kth C.

From PWM-M, the normalized methylation-weighted frequency f_{ij} was calculated as follows:

$$f_{ij} = \frac{M_{ij}}{\sum_i M_{ij}}$$

Then, the Shannon entropy H_{ij} and the information content $R_{ij}R_j$ of position j were calculated, respectively [33, 34]:

$$H_j = -\sum_{i\in\{A,C,G,T\}} f_{ij} \times \log_2 f_{ij}$$

$$R_j = \log_2(4) - H_j - 0.541$$

Finally, the height of the logo for each nucleobase i at position j was calculated as follows:

$$height_{ij} = f_{ij} \times R_j$$

The calculated PWM-M and the M-logo were visualized using the Logomaker [35].

Budding yeast NOMe-Seq

S288C cells (Biological Resource Center at the National Institute of Technology and Evaluation, NBRC1136) were inoculated into 50 mL of YPD media (1% yeast extract, 2% Bacto peptone, and 2% dextrose) and cultivated with vigorous shaking at 30 °C overnight. Cell suspension was diluted with 1 L of YPD and further cultured at 30 °C for 4 h with shaking. Cells were collected via centrifugation at 2000×g for 15 min, washed twice with 50 mL of water, and resuspended in 20 mL of 1 M sorbitol. To the cell suspension, 20 μL of β-mercaptoethanol and 20 mg of Zymolyase 100 T (Seikagaku Kogyo Co., Tokyo, Japan) were added and incubated at room temperature for 15 min to degrade the cell wall. Then, the spheroplasts were washed twice with 20 mL of 1 M sorbitol and resuspended in 20 mL of 1 M sorbitol. Fixation was performed by adding 240 μL of 38% formaldehyde (Wako Chemicals, Osaka, Japan) to the resuspended spheroplasts and incubating at room temperature for 5 min. The fixed spheroplasts were washed twice with 20 mL of 1 M sorbitol and pelleted via centrifugation at 5000×g for 5 min. The spheroplasts were then resuspended in 10 mL of nuclei preparation solution (10 mM HEPES–KOH, pH 7.5; 10 mM NaCl; 3 mM $MgCl_2$; and 0.5% [v/v] Nonidet P 40 substitute) and incubated at room temperature for 5 min. The nuclei were collected via centrifugation at 5000×g for 5 min and washed twice with 10 mL of nuclei wash solution (10 mM HEPES–KOH, pH 7.5; 10 mM NaCl, and 3 mM $MgCl_2$). The nuclei were resuspended in 10 mL of nuclei wash solution, of which 500 μL was dispensed into 20 tubes and stored at − 80 °C until use.

A frozen nuclei suspension was thawed and resuspended in 500 μL of 1 × methylation buffer (50 mM Tris–HCl, pH 8.5, 50 mM NaCl, 1 mM DTT, 1.6 mM SAM). Then, an appropriate amount of either M.CviPI (NEB) or N-terminal 15 amino acid deletion mutant of M.CviQIX (termed CCMT, see RESULTS section), prepared in our laboratory, was added. After incubation at 37 °C for 1 h, the nuclei were pelleted via centrifugation at 5000×g for 1 min and resuspended in 40 μL of 10 mM Tris–HCl, pH 8.0. The nuclei suspension was supplemented with 5 μL of 20 mg/mL proteinase K and 10% (w/v) SDS and incubated at 50 °C for 1 h. The reaction was supplemented with 150 μL of 1 M Tris–HCl, pH 8.0, overlayed with white mineral oil, and incubated at 80 °C for 24 h. The aqueous phase was transferred to a new tube and mixed with 200 μL of Buffer AL and isopropanol. The mixture was loaded onto the DNeasy column by centrifuging at 10,000×g for 1 min, and the column was washed sequentially with 500 μL of Buffer W1 and Buffer W2. Finally, the purified DNA was eluted with 200 μL of Buffer AE and used for the WGBS analysis using the tPBAT protocol. Note that Buffer AL, Buffer W1, Buffer W2, and Buffer AE are DNeasy Blood & Tissue Kit components.

NOMe-Seq of mammalian cells

The IMR-90 cells (JCRB cell bank #JCRB9054) were cultivated in four T75 flasks containing 30 mL of Minimum Essential Media (Thermo Fisher Scientific) supplemented with 10% fetal bovine serum (Thermo Fisher Scientific) and 50 U/mL penicillin–streptomycin (Thermo Fisher Scientific) at 37 °C in an atmosphere of 100% humidity and 5% CO_2. Cells grown were treated with trypsin and collected via centrifugation at 2000×g for 5 min. Cells were washed twice with 20 mL of D-PBS (Nacalai Tesque) and resuspended in CELLBANKER 1 (Takara

Bio) at 1×10^7 cells/mL. The cell suspension was dispensed in 500 μL and stored at − 80 °C until use.

Before use, the frozen cell suspension was thawed and washed twice with PBS. Cells were then resuspended in nuclei preparation buffer (10 mM Tris–HCl, pH 8.0, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA) and 0.5% (w/v) NP-40, and incubated on ice for 10 min. The nuclei were collected via centrifugation at 2000×*g* for 5 min, resuspended in with 500 μL of 1 × methylation buffer with an optimized amount of CCMT, and incubated at 37 °C for 30 min. Then, the nuclei were collected via centrifugation at 2000×*g* for 5 min, and the supernatant was removed. The nuclei were dissolved in 20 μL of 20 mg/mL proteinase K and 200 μL of Buffer ATL was added. The reaction was incubated at 56 °C for 1 h and supplemented with 200 μL of Buffer AL and 200 μL of isopropanol. The mixture was loaded onto the DNeasy column and washed with 500 μL of buffer W1 or Buffer W2. The purified DNA was eluted with 200 μL of Buffer AE and used for WGBS.

Abbreviations

BSA: Bovine serum albumin; IMAC: Immobilized-metal affinity chromatography; PBAT: Post-bisulfite adapter tagging; TF: Trigger factor; WGBS: Whole-genome bisulfite sequencing.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12896-022-00765-3.

Additional file 1. Figs. S1–S14, Tables S2–S3, and the nucleotide sequences of genes.

Additional file 2. **Table S1.** Summary of the analysis of candicate DNMTs.

Acknowledgements

We would like to thank Editage (www.editage.com) for English language editing.

Author contributions

FM conceived the study, performed experiments, analyzed data, and wrote the paper. MM, YS, KM, YI, AB performed experiments. YF and UO contributed to experimental design. TI contributed to experimental design and wrote the manusript. All authors have read and approved the manuscript.

Funding

This work was supported by the Platform Project for Supporting Drug Discovery and Life Science Research, Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS) and Basis for Supporting Innovative Drug Discovery and Life Science Research from AMED (grant number JP20am0101103 to TI and JP22ama121022j0001 to FM), JSPS KAKENHI (grant number JP17H06305 to TI, and grant number JP20H03243 to FM], and National Cancer Center Research and Development Fund (2020-A-7). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data, or in writing the manuscript.

Availability of data and materials

All sequencing data produced in this study were deposited to NCBI GEO and NCBI SRA with accession numbers GSE192699, GSE192949, GSE202353, GSE202354, and SRR17348378-SRR17348416. The publicly available data used in the current study were downloaded from NCBI SRA with accession numbers SRR12926697, SRR6729489, and SRR7765313.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

FM has a patent pending for CCMT and confirms that the patent does not restrict research applications of the method. The remaining authors have no conflicts of interest to declare.

Author details

[1] Department of Biochemistry, Kyushu University Graduate School of Medical Sciences, 3-1-1 Maidashi, Higashi-Ku, Fukuoka 812-8582, Japan. [2] Division of Infection and Immunity, Research Center for Zoonosis Control, Hokkaido University, Sapporo 001-0020, Japan. [3] Department of Biochemistry, Faculty of Pharmacy, Minia University, El-Minia 61511, Egypt. [4] Department of Pediatrics, Kyushu University Graduate School of Medical Sciences, 3-1-1 Maidashi, Higashi-Ku, Fukuoka 812-8582, Japan.

Received: 10 May 2022 Accepted: 25 October 2022
Published: 4 November 2022

References

1. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol. 2019;20:590–607.
2. Edwards JR, Yarychkivska O, Boulard M, Bestor TH. DNA methylation and DNA methyltransferases. Epigenetics Chromatin. 2017;10:23.
3. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484–92.
4. Petryk N, Bultmann S, Bartke T, Defossez P-A. Staying true to yourself: mechanisms of DNA methylation maintenance in mammals. Nucleic Acids Res. 2021;49:3020–32.
5. Jones PA, Liang G. Rethinking how DNA methylation patterns are maintained. Nat Rev Genet. 2009;10:805–11.
6. Tajima S, Suetake I, Takeshita K, Nakagawa A, Kimura H. Domain structure of the Dnmt1, Dnmt3a, and Dnmt3b DNA methyltransferases. Adv Exp Med Biol. 2016;945:63–86.
7. Jang HS, Shin WJ, Lee JE, Do JT. CpG and non-CpG methylation in epigenetic gene regulation and brain function. Genes. 2017;8:148.
8. Patil V, Ward RL, Hesson LB. The evidence for functional non-CpG methylation in mammalian cells. Epigenetics. 2014;9:823–8.
9. van de Lagemaat LN, Flenley M, Lynch MD, Garrick D, Tomlinson SR, Kranc KR, et al. CpG binding protein (CFP1) occupies open chromatin regions of active genes, including enhancers and non-CpG islands. Epigenetics Chromatin. 2018;11:59.
10. Teissandier A, Bourc'his D. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. EMBO J. 2017;36:1471–3.
11. Dhayalan A, Rajavelu A, Rathert P, Tamas R, Jurkowska RZ, Ragozin S, et al. The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. J Biol Chem. 2010;285:26114–20.
12. Solomon MJ, Larsen PL, Varshavsky A. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell. 1988;53:937–47.
13. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316:1497–502.
14. Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome Res. 2012;22:2497–506.
15. Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat Methods. 2020;17:1191–9.

16. Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science. 2020;368:1449–54.
17. Steensel BV, Henikoff S. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. Nat Biotechnol. 2000;18:424–8.
18. Roberts RJ, Macelis D. REBASE-restriction enzymes and methylases. Nucleic Acids Res. 1997;25:248–62.
19. 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:D298–9.
20. Xu M, Kladde MP, Van Etten JL, Simpson RT. Cloning, characterization and expression of the gene coding for a cytosine-5-DNA methyltransferase recognizing GpC. Nucleic Acids Res. 1998;26:3961–6.
21. Zhang Y, Nelson M, Nietfeldt J, Xia Y, Burbank D, Ropp S, et al. Chlorella virus NY-2A encodes at least 12 DNA endonuclease/methyltransferase genes. Virology. 1998;240:366–75.
22. Chan SH, Zhu Z, Van Etten JL, Xu SY. Cloning of CviPII nicking and modification system from chlorella virus NYs-1 and application of Nt.CviPII in random DNA amplification. Nucleic Acids Res. 2004;32:6187–99.
23. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
24. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.
25. Nilsson B, Moks T, Jansson B, Abrahmsén L, Elmblad A, Holmgren E, et al. A synthetic IgG-binding domain based on staphylococcal protein A. Protein Eng Des Sel. 1987;1:107–13.
26. Inouye S, Sahara Y. Soluble protein expression in *E. coli* cells using IgG-binding domain of protein A as a solubilizing partner in the cold induced system. Biochem Biophys Res Commun. 2008;376:448–53.
27. Malakhov MP, Mattern MR, Malakhova OA, Drinker M, Weeks SD, Butt TR. SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. J Struct Funct Genom. 2004;5:75–86.
28. Miura F, Shibata Y, Miura M, Sangatsuda Y, Hisano O, Araki H, et al. Highly efficient single-stranded DNA ligation technique improves low-input whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res. 2019;47: e85.
29. Miura F, Enomoto Y, Dairiki R, Ito T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Res. 2012;40: e136.
30. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. 2019;20:207–20.
31. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.
32. Ilinsky VV, Kozlova EM, Degtyarev SK, Yankovsky NK, Makeev VJ. The Efficiency of 5-methylcytosine identification in DNA of *Escherichia coli* cells that carry bacterial DNA methyltransferase genes using an Oxford nanopore device. Biophysics. 2020;65:889–93.
33. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. J Mol Biol. 1986;188:415–31.
34. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18:6097–100.
35. Tareen A, Kinney JB. Logomaker: beautiful sequence logos in Python. Bioinformatics. 2019;36:2272–4.
36. Braberg H, Echeverria I, Bohn S, Cimermancic P, Shiver A, Alexander R, et al. Genetic interaction mapping informs integrative structure determination of protein complexes. Science. 2020. https://doi.org/10.1126/science.aaz4910.
37. Miura F, Fujino T, Kogashi K, Shibata Y, Miura M, Isobe H, et al. Triazole linking for preparation of a next-generation sequencing library from single-stranded DNA. Nucleic Acids Res. 2018;46: e95.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.