CpG underrepresentation and the bacterial CpG-specific DNA methyltransferase M.MpeI

Marek Wojciechowski[a], Honorata Czapinska[a], and Matthias Bochtler[a,b,c,1]

[a]Laboratory of Structural Biology, International Institute of Molecular and Cell Biology, 02-109, Warsaw, Poland; [b]Institute of Biochemistry and Biophysics of the Polish Academy of Sciences, 02-106, Warsaw, Poland; and [c]Laboratory of Structural Biology, Schools of Chemistry and Biosciences, Cardiff University, Cardiff CF10 3AT, United Kingdom

Edited* by Robert Huber, Max Planck Institute of Chemistry, Planegg-Martinsried, Germany, and approved November 19, 2012 (received for review May 11, 2012)

Cytosine methylation promotes deamination. In eukaryotes, CpG methylation is thought to account for CpG underrepresentation. Whether scarcity of CpGs in prokaryotic genomes is diagnostic for methylation is not clear. Here, we report that Mycoplasms tend to be CpG depleted and to harbor a family of constitutively expressed or phase variable CpG-specific DNA methyltransferases. The very CpG poor *Mycoplasma penetrans* and its constitutively active CpG-specific methyltransferase M.MpeI were chosen for further characterization. Genome-wide sequencing of bisulfite-converted DNA indicated that M.MpeI methylated CpG target sites both in vivo and in vitro in a locus-nonselective manner. A crystal structure of M.MpeI with DNA at 2.15-Å resolution showed that the substrate base was flipped and that its place in the DNA stack was taken by a glutamine residue. A phenylalanine residue was intercalated into the "weak" CpG step of the nonsubstrate strand, indicating mechanistic similarities in the recognition of the short CpG target sequence by prokaryotic and eukaryotic DNA methyltransferases.

bisulfite sequencing | cytosine deamination | genome evolution | microbiology | structural biology

Cytosine methylation is essential in many eukaryotic organisms (1). However, it is also mutagenic, because it drives deamination (2, 3). Methyl transfer promotes hydrolysis (2), and once formed, 5-methylcytosine (5mC) deaminates twice as fast as cytosine (3). Moreover, 5-methylation of cytosines subverts DNA repair, because deamination of 5mC leads to thymine (a legitimate DNA base) and not uracil (an illegitimate DNA base).

In many eukaryotes, particularly in vertebrates, cytosine methylation is predominantly found in the CpG (5′-CG-3′) sequence context (4, 5). Consistent with methylation-driven loss, the CpG dinucleotide is strongly underrepresented in vertebrate genomes (6). CpG depletion is most pronounced in intergenic regions, less so in promoter regions and introns, and least in exons, which are subject to the strongest selective pressure (*SI Appendix*, Table S1). It is not clear whether CpG loss due to methylation in mammals is ongoing (6) or whether equilibrium between loss and restoration of cytosines in the CpG context has been reached (7).

In prokaryotes, C5 methylation of cytosines occurs in diverse sequence contexts. It is predominantly associated with restriction-modification systems, targeting short palindromic or nearly palindromic sites (8). Whether sequence-specific cytosine methylation in prokaryotes is correlated with depletion of target sites is unclear. Slight underrepresentation of palindromes in the DNA of phages and some bacteria has been reported (9, 10), but the effects are very weak. The (near) lack of a genomic methylation footprint could be due to transient association of restriction-modification systems with their hosts, higher proportion of functionally constrained DNA in prokaryotes, or the scarcity of methylation sites (most target sequences of prokaryotic DNA methyltransferases are at least four nucleotides long). The latter explanation leaves open the possibility that prokaryotic cytosine methyltransferases with dinucleotide targets could still shape dinucleotide frequencies in host genomes. *Spiroplasma monobiae* MQ1, which hosts the well-known bacterial CpG-specific DNA methyltransferase (M.SssI), is a proprietary strain and has not had its entire genome sequenced. The M.SssI coding gene itself is CpG depleted (5 CpGs, 23 expected on the basis of the GC content of the gene) (11), but some ribosomal RNA coding DNA sequences are not.

Results

CpG Depletion in Prokaryotes. There are no reports on CpG depletion in prokaryotes. Therefore, we scanned more than 3,000 bacterial genomes available from the National Center for Biotechnology Information (NCBI) for CpG underrepresentation (Fig. 1 and *SI Appendix*, Tables S2 and S3), and also as a control for underrepresentation of other palindromic dinucleotides (*SI Appendix*, Fig. S1). For each species we determined CpG underrepresentation in the whole genome and in nonprotein coding ("intergenic") regions to avoid bias from codon preferences. Although both overall and intergenic CpG depletion were not smoothly distributed across the "tree of life", a good correlation between the two measures of CpG depletion was found (Pearson coefficient 0.67, after removal of near duplicates 0.62). The overall very CpG poor genomes of the hyperthermophilic *Methanocaldococcus infernus* (strain ME, NC_014122), intracellular *Mycoplasma penetrans* (strain HF-2, NC_004432) (12), and dental pathogen *Fusobacterium nucleatum* (ssp. *nucleatum*, NC_003454) (13) were analyzed in more detail. The context of remaining CpGs in the three genomes was largely dependent on the GC content of the genome but showed only slight preferences for flanking bases in the case of the *M. penetrans* genome (*SI Appendix*, Fig. S2). In all three outlier genomes, only CpGs were strongly underrepresented, but other dinucleotide frequencies did not deviate much from expectation (*SI Appendix*, Fig. S3). In contrast to the situation in eukaryotes (*SI Appendix*, Table S1), CpGs were more underrepresented in coding than in noncoding regions (Fig. 1). This observation and the dependence of CpG underrepresentation on reading frame (*SI Appendix*, Table S4) could be explained by the near absence of CpG-containing codons in all three cases (*SI Appendix*, Fig. S4). Codon bias could account for the small number of CpGs in *M. infernus* and *F. nucleatum* but could not explain CpG underrepresentation in noncoding regions of *M. penetrans*.

Candidate Constitutive and Phase Variable CpG Methyltransferases in Prokaryotes. To correlate CpG depletion with methylation, we used BLAST and the known amino acid sequence of M.SssI

Author contributions: M.B. designed research; M.W., H.C., and M.B. performed research; M.W., H.C., and M.B. analyzed data; and M.B. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: The atomic coordinates and structure factors have been deposited in the Protein Data Bank, www.pdb.org (PDB ID code 4DKJ). The sequencing data have been deposited in the GEO archive (accession no. GSE41022).

[1]To whom correspondence should be addressed. E-mail: mbochtler@iimcb.gov.pl.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1207986110/-/DCSupplemental.

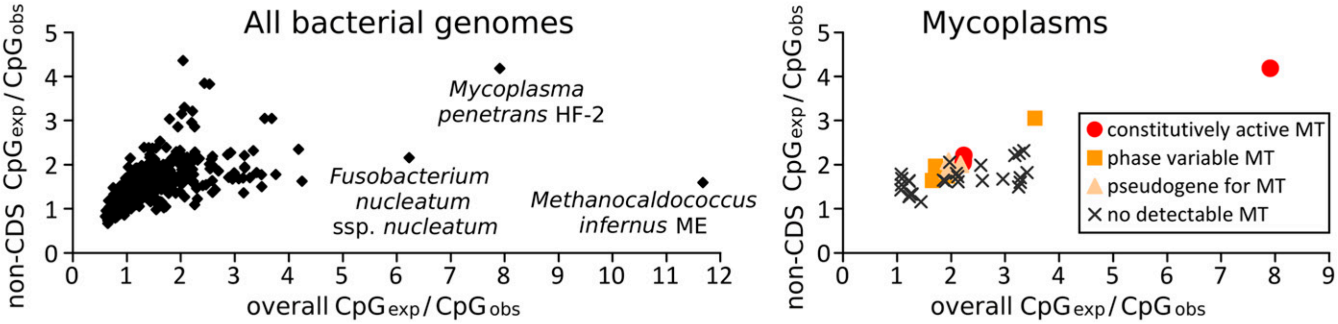

Fig. 1. CpG underrepresentation in bacterial genomes. The degree of CpG underrepresentation (expressed as a ratio of expected and observed CpGs) was determined for all completed genomes in NCBI (excluding nearly duplicate entries for closely related strains). Calculations were carried out either for entire genomes or only for regions not annotated as protein coding. To exclude plasmids from the analysis only genomes above 500 kb were considered. (*Right*) Mycoplasms are coded according to the predicted properties of M.MpeI orthologs in these genomes.

methyltransferase to screen for orthologs in bacterial genomes. None are found in most bacteria (including the outlier genomes of *M. infernus* and *F. nucleatum*). However, with search parameters optimized for distant sequence similarity, a family of predicted C5 methyltransferases was identified. Seemingly intact methyltransferase genes were found in a few Mycoplasms, including *M. penetrans* (*mpeORF4940*), which was noted for overall and intergenic CpG depletion. In other Mycoplasms, such as *Mycoplasma pulmonis* (14) or *Mycoplasma crocodyli* (15), frameshifts in the genomic sequence suggested recent inactivation of a previously functional methyltransferase gene. Interestingly, resequencing of the *M. crocodyli* methyltransferase gene (*mcrMORF235*) did not confirm the frameshift. The discrepancy between our and the published sequence mapped to a dinucleotide repeat region, strongly suggesting a polymerase slippage event during bacterial propagation rather than a sequencing error (*SI Appendix*, Fig. S5). Most but not all predicted CpG methyltransferases contain such dinucleotide repeat tracts and are therefore probably phase variable (*SI Appendix*, Table S5). We note that slipped strand mispairing at repetitive DNA is common in bacteria (16, 17) and has been described in Mycoplasms in the context of variations of the repertoire of potentially immunogenic surface proteins (18).

CpG Underrepresentation and CpG Methyltransferase Expression in Mycoplasms. *M. penetrans* stood out in the screens for CpG underrepresentation and harbors a predicted constitutively active CpG methyltransferase. We further noticed that on average CpGs tend to be more underrepresented in Mycoplasms than in other bacteria (Fig. 1). Moreover, Mycoplasms that harbor predicted constitutively active or phase variable methyltransferases seemed to be more CpG depleted on average than Mycoplasms without such a methyltransferase, but there were also clear exceptions to this trend (Fig. 1 and *SI Appendix*, Table S5).

CpG Methylation in Mycoplasms and Other Selected Bacteria. To check our predictions about DNA methylation, we subjected selected bacterial genomes to bisulfite sequencing. Samples were initially analyzed by conventional Sanger sequencing (*SI Appendix*, Tables S6 and S7). The relatively sparse data obtained in this manner indicated methylation of all tested CpG sites for *M. penetrans* and *M. crocodyli* and no CpG methylation for the other tested genomes (*M. infernus*, *F. nucleatum*, and also *M. agalactiae* PG2, *M. pulmonis*, and *M. mycoides* ssp. *capri*). For further experimental studies we selected the CpG depleted *M. penetrans* with a constitutively expressed DNA methyltransferase.

Bisulfite Sequencing of the *M. penetrans* Genome. In eukaryotes, CpG methylation is locus-dependent. To distinguish between locus-dependent and global CpG methylation, we bisulfite converted *M. penetrans* genomic DNA. The deaminated DNA was amplified (to overcome problems with conversion damage) and subjected to Illumina MiSeq reversible terminator sequencing. Because sequence coverage was very uneven (*SI Appendix*, Fig. S6), the data were analyzed in two ways, giving equal weight either to reads or to cytosines in the genome (*SI Appendix*, Table S8). The latter way of analyzing the data showed that most cytosines in the genome were either unmethylated (fraction of reads indicating methylation ≤0.1) or highly methylated (fraction of reads indicating methylation >0.9) (*SI Appendix*, Fig. S7). Both ways of analyzing the data indicated nearly complete CpG methylation (Table 1). For most CpGs, independent methylation information was available for both DNA strands (*SI Appendix*, Fig. S8). The few sites that were weakly or not methylated in one strand were typically well methylated in the other strand. Manual inspection of the few CpGs that were flagged as unmethylated indicated no obvious patterns. Therefore we attribute these sites to imperfect bisulfite conversion and conclude that CpG methylation in *M. penetrans* is not locus-specific. We also note that an unbiased analysis of methylation sites readily identified CpG and RGCY as methyltransferase targets (Fig. 2 and *SI Appendix*, Table S8). The latter is consistent with the description of an AGCT-specific restriction-modification system in *M. penetrans* but suggests that the methyltransferase has a more relaxed specificity than currently indicated in the database (8). Together the data show that whole genome bisulfite sequencing can be a very useful tool to assign specificities of C5 methyltransferases with excellent statistical support.

In Vitro Characterization of the *M. penetrans* CpG-Specific Methyltransferase. To confirm the assignment of M.MpeORF4940P as the CpG-specific DNA methyltransferase, we overexpressed a histidine-tagged version of the protein in *Escherichia coli*. The protein was purified (*SI Appendix*, Fig. S9) and tested for methyltransferase activity in vitro (using phage λ DNA as the substrate) (*SI Appendix*, Fig. S10). As predicted, the in vitro methylated DNA was protected against the CpG methylation sensitive HpaII (CCGG target sequence) but remained susceptible to the CpG methylation insensitive isoschizomeric MspI endonuclease (*SI Appendix*, Fig. S10). To determine the exact in vitro specificity of M.MpeORF4940P, we used the enzyme to methylate genomic DNA from Dcm-negative *E. coli* strain ER2566 and subjected this DNA or unmethylated control DNA to bisulfite sequencing (Table 1 and *SI Appendix*, Table S8). Similar analysis as for the *M. penetrans* DNA showed that almost all CpG sites in M.MpeORF4940P-treated DNA, but not control DNA, were highly methylated (Table 1 and *SI Appendix*, Fig. S11). Together the functional data about M.MpeORF4940P justify renaming this protein to M.MpeI in accordance with nomenclature guidelines (the first described DNA methyltransferase from *M. penetrans*) (19).

Table 1. High-throughput analysis of CpG methylation

Genome	*M. penetrans* First	*M. penetrans* Second	*E. coli* (control) First	*E. coli* (control) Second	*E. coli* M.MpeI First	*E. coli* M.MpeI Second
Read-based analysis						
Unconverted (methylated) CpGs	340,263	311,814	23,648	23,843	855,154	718,669
Converted (unmethylated) CpGs	7,913	7,575	2,962,773	2,616,758	57,677	49,008
Unconverted (methylated) CpGs (%)	97.7	97.6	0.8	0.9	93.7	93.6
Site-based analysis						
CpGs	5,680		699,086		699,086	
With methylation info	5,289		382,161		364,082	
Unmethylated (%)	15 (0.3)		371,726 (97.3)		12,717 (3.5)	
Intermediately methylated (%)	403 (7.6)		8,131 (2.1)		33,081 (9.1)	
Highly methylated (%)	4,871 (92.1)		2,304 (0.6)		318,284 (87.4)	

Methylation analysis was done by bisulfite conversion and reversible terminator sequencing. Complete statistics are provided in *SI Appendix*, Table S8. In the read-based analysis, the first and second sequences from paired end sequencing were analyzed independently. In the site-based analysis, all reads from paired end sequencing were pooled. Sites were classified as unmethylated if the fraction of reads indicating methylation was ≤0.1 and as highly methylated if it was >0.9.

Crystallization of the M.MpeI–DNA Complex. To better understand the CpG specificity of M.MpeI and the tolerance of the related DNA methyltransferases to sequence insertions, we aimed for a crystal structure of the protein in complex with target DNA. The physiological substrates of DNA methyltransferases, particularly those with symmetric target sequences, are hemimethylated (after DNA replication). Therefore, we chose a hemimethylated oligoduplex for cocrystallization. Moreover, we placed a 5-fluorocytosine (5FC) in the position of the substrate base. We expected that this design would interfere with the last step of catalysis, the regeneration of the active enzyme, by preventing the elimination of the active site cysteine (20). As a cosubstrate we used S-adenosylmethionine (SAM). The best crystal was obtained in a drop that contained a mixture of SAM and its degradation product 5′-methylthioadenosine (MTA) (from scission of SAM to MTA and homoserine lactone) (21), and perhaps also the methyltransferase coproduct S-adenosylhomocysteine (SAH). Diffraction data extended to 2.15-Å resolution and could be interpreted by molecular replacement using the M.HhaI–DNA complex as a search model [Protein Data Bank (PDB) code 2C7P] (22). The refined structure has been deposited in the PDB with the accession code 4DKJ (*SI Appendix*, Table S9).

Overall Structure and Active Site. The M.MpeI DNA cocrystal structure displays "classic" features of a C5-methyltransferase–DNA complex (Figs. 3–5). The 5FC (replacing the substrate base in the structure) is extruded from the DNA, where its place is taken by an intercalating amino acid (Gln141). The flipped base is accommodated in a pocket of the methyltransferase with the Watson–Crick edge facing Glu184. Residues Arg228 and Arg230 are also close to the flipped base (approximately 3.0 Å) and donate one or two hydrogen bonds to its O2 (Fig. 4*A*). We see no evidence for methyl group transfer to the substrate base but also no clear methyl group on the cofactor. The SAM binding site is probably occupied by a mixture of SAM, MTA, and SAH. Details are difficult to discern, because the density is well defined only for the base, much weaker for the sugar, and poor for the sulfonium or thioether region (Fig. 4*B*). The catalytic cysteine residue of M.MpeI (Cys135) is well positioned for nucleophilic attack on the C6 atom. However, the Sγ–C6 distance (2.9 Å) is intermediate between a covalent bond (1.8 Å) and noncovalent interaction (3.6 Å). A similar distance (2.6–2.8 Å) is seen in complexes of other C5 methyltransferases with DNA in the presence of SAH. This has been interpreted as evidence for the formation of a partial covalent bond (23, 24), which should be accompanied by a slight loss of substrate base planarity. However, our resolution is insufficient to detect such detail. In solution, neither MTA nor SAH promote irreversible covalent bond formation, even at neutral pH (*SI Appendix*, Fig. S12).

Structural Basis of M.MpeI Sequence Specificity. M.MpeI not only interacts with the flipped substrate base but also engages in

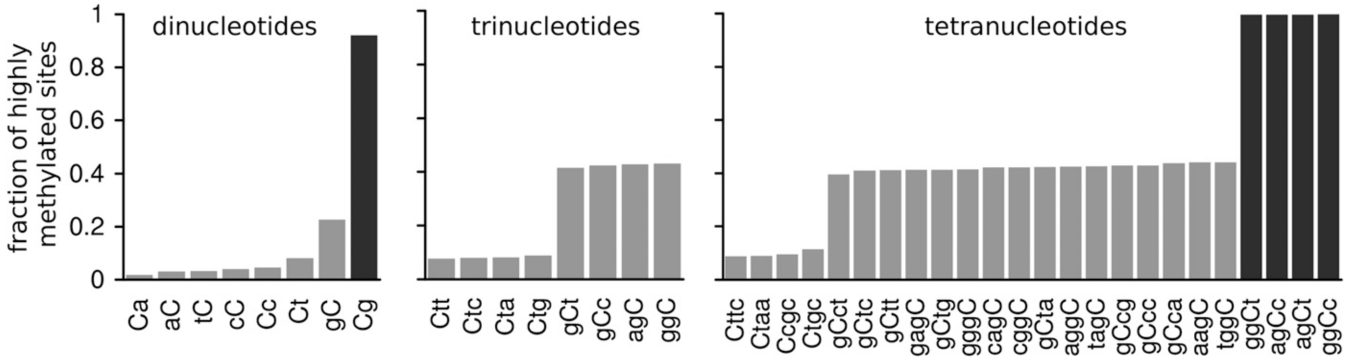

Fig. 2. Determination of C5 methyltransferase target sequences. For every possible dinucleotide (*Left*), trinucleotide (*Center*), or tetranucleotide (*Right*) context, the fraction of highly methylated cytosines (>90% of reads indicating methylation) was determined. Candidate target sequences were ordered according to the fraction of highly methylated sites, and all (*Left*) or the highest-scoring (*Center* and *Right*) sites were plotted. Sites that were already identified as methylation targets by a shorter subsequence were omitted for clarity. The cytosine that was analyzed for methylation is marked by a capital "C"; all other bases in the putative recognition sequences are given in small letters.

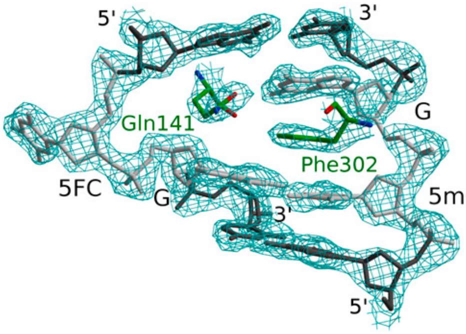

Fig. 3. Zoom into the M.MpeI–DNA complex. The M.MpeI target sequence (light gray) together with one flanking base pair on each side (dark gray) and the intercalating protein residues (green) are shown. The composite omit map is contoured at 1.5 σ (cyan). The displacement of the flipped out substrate cytosine (here 5FC) by an amino acid side chain (Gln141) is typical for C5 methyltransferases. The intercalation of Phe302 into the easily unstackable 5mCpG step of the nonsubstrate strand is characteristic for the M.MpeI–DNA complex and probably contributes to DNA recognition.

extensive interactions with the other three bases of the specifically recognized sequence (*SI Appendix*, Table S10). The guanine opposite the flipped base (the "estranged" guanine) is recognized by a few hydrogen bonds: Ser304 Oγ and N donate hydrogen bonds to its O6, and Asn303 N to its N7 atom. In addition, Gln141, which displaces the substrate C (or 5FC), accepts a bifurcated hydrogen bond from the N1 and N2 atoms of G (Fig. 4*C*). The most striking feature of the sequence recognition by M.MpeI is the insertion of Phe302 from the major groove side into the 5mCpG step of the nonsubstrate DNA strand. The height of this step is almost twofold larger than usual, which results in a substantial DNA bend or kink (Fig. 3). Despite the severe distortion, the second base pair of the recognition sequence remains Watson–Crick hydrogen bonded. The C of this base pair donates a hydrogen bond from its N4 atom to the carboxylate group of Glu305. The G faces solvent with its minor groove edge and accepts a hydrogen bond from Ala323 N to its O6 atom (Fig. 4*D*).

Discussion

Recognition of the CpG Target Sequence. The recognition of a short sequence such as CpG requires multiple interactions. The insertion of Gln141 was expected, but intercalation of Phe302 in the 5mCpG step came as a surprise. Because 5′-pyrimidine-purine-3′ dinucleotides can be unstacked more easily than other dinucleotides (25), the intercalation might contribute to sequence specificity. Interestingly, a similar role of methionine intercalation has recently been reported for the CpG sequence recognition by ThaI restriction endonuclease (26).

M.MpeI vs. Dnmt1 and M.HhaI. It is instructive to compare the DNA complexes of M.MpeI, Dnmt1 (the only eukaryotic DNA methyltransferase cocrystallized with DNA) (27), and M.HhaI (the prototypical bacterial methyltransferase, which methylates CpG in the GCGC context) (28) (Fig. 5 and *SI Appendix*, Fig. S13). The catalytic cysteine (Cys135 in M.MpeI, Cys1229 in Dnmt1, Cys81 in M.HhaI) and glutamate (Glu184, Glu1269, and Glu119, respectively) as well as two arginine residues that flank the pocket for the flipped base are conserved. In all three protein–DNA complexes the flipped substrate base is displaced by insertion of amino acid side chains from the minor (Gln141 in M.MpeI, Met1235 in Dnmt1) or major (Lys1537 in Dnmt1, Gln237 in M.HhaI) groove side. Nevertheless, recognition of the CpG sequence differs between the three complexes. M.MpeI and Dnmt1, but not M.HhaI, unstack the CpG step of the nonsubstrate strand by intercalation of an amino acid side chain from the major groove side (Phe302 in M.MpeI, Trp1512 in Dnmt1). There are also differences in CpG recognition between the M.MpeI and Dnmt1 cocrystal structures. In the case of M.MpeI, the estranged guanine interacts with the intercalating Gln141. This mode of interaction makes CpG recognition context independent. In the Dnmt1 DNA complex, the equivalent guanine interacts with a guanine on the 5′-side of the substrate strand CpG, at least in the crystal. However, further experiments show that the guanine–guanine interaction in the Dnmt1 DNA cocrystal is not important in solution (29). Thus, CpG recognition by M.MpeI and Dnmt1 might be more similar than suggested by the comparison of the two crystal structures.

Link Between CpG Methylation and Depletion. Because the reasoning linking CpG methylation and depletion is universal, it should also apply to prokaryotes. The correlative data in this study suggest that this is the case, but other factors are also involved and tend to obscure clear-cut correlations. Codon use could have a particularly strong influence, because 80–90% of bacterial genome sequences are protein coding. Whether extreme codon frequencies are a consequence of DNA methylation is currently unclear. The case of *M. penetrans* suggests so, but then *M. infernus* and *F. nucleatum* should have also harbored CpG-specific DNA methyltransferases, which is currently speculative. Interestingly, the *M. penetrans* CpG methyltransferase

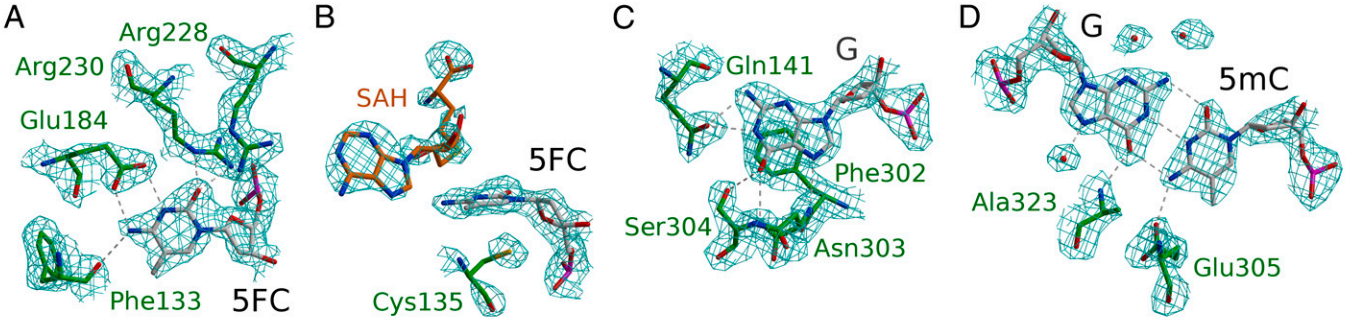

Fig. 4. M.MpeI–DNA interactions. (*A*) Region of the flipped 5FC base. For the Michaelis complex, a single hydrogen bond is formed between the Watson–Crick edge of the substrate base and Glu184 at neutral pH. At acidic pH, Glu should be protonated and a second hydrogen bond may be formed. The same, but with swapped roles of proton donor and acceptor, may happen during the reaction when the addition of the catalytic cysteine to the base converts the N3 atom from a hydrogen bond acceptor to donor (24). (*B*) Alternative view of the flipped base with the substrate 5FC residue, catalytic cysteine, and coproduct SAH molecule. (*C*) Recognition of the "estranged" guanine base. (*D*) Recognition of the G:5mC pair adjacent to the flipped base. The hydrogen bonds to the water molecules on the sugar edge of the guanine are ambiguous and are therefore not displayed. In all panels the composite omit electron density map was contoured at 1.5 σ.

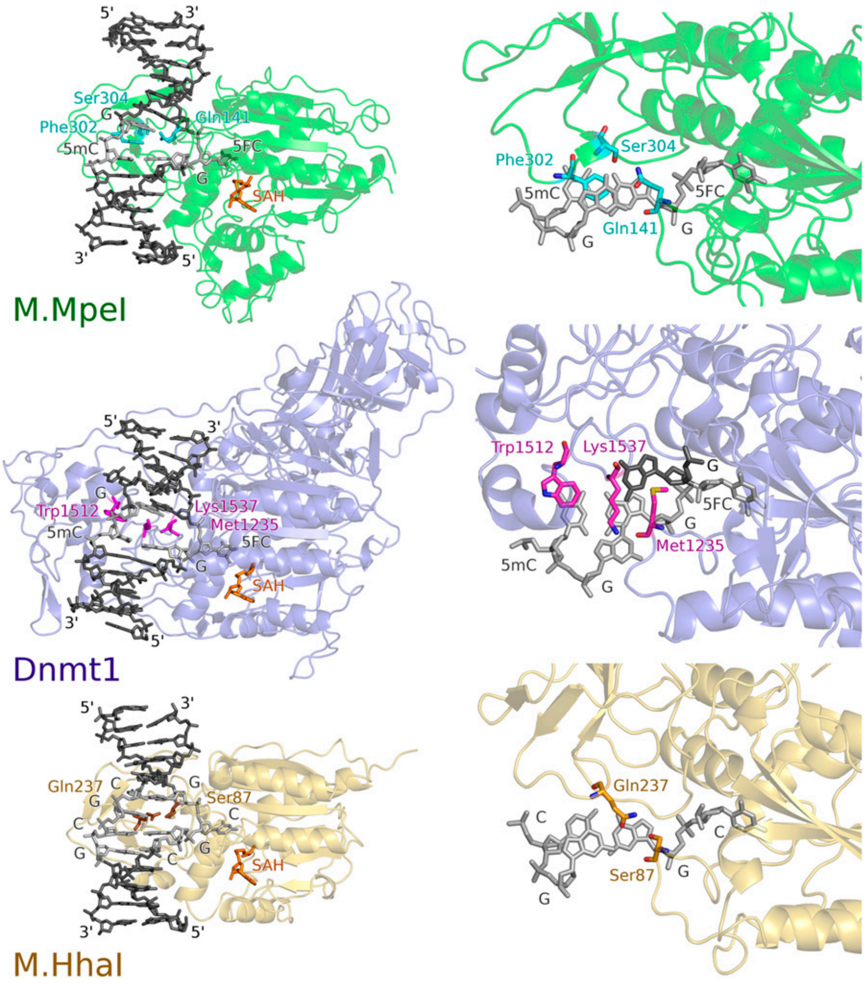

Fig. 5. Structural comparison of methyltransferase–DNA complexes. M.MpeI–, Dnmt1–, and M.HhaI–DNA complexes (PDB codes 4DKJ, 4DA4, and 3MHT) were superposed according to the position of the 5FC substrate base or its analog. The DNA intercalating residues (Gln141 and Phe302 in M.MpeI; Met1235, Trp1512, and Lys1537 in Dnmt1; Ser87 and Gln237 in M.HhaI) are shown in stick representation. The target CpG sequences are indicated with faint gray color and labeled. (*Left*) Overall structure of the three enzymes together with bound DNA and coproduct S-adenosylhomocysteine. (*Right*) The structures are 90° rotated and zoomed on the target CpG sequences. A single flanking G nucleotide in Dnmt1 complex that forms a nonstandard base pair with the specifically recognized G is indicated in black.

operon contains an ORF for a predicted nuclease. The latter might act as a restrictase in a restriction-modification system or as a Vsr-like protein to repair methylation-induced deamination damage. At present, we cannot exclude that CpG methylation and depletion in some Mycoplasms represent independent adaptations to some external pressure either. The host immune systems might well provide such selection. DNA with unmethylated CpGs is very immunogenic, more so than "host like" DNA (30–32). The pathogenic Mycoplasms, which are already less "visible" to the host immune system than other bacteria owing to the absence of LPS, might thus profit from both CpG methylation and depletion. Whether CpG methylation contributes to *Mycoplasma* pathogenicity and whether CpG depletion is caused by, rather than correlated with, CpG methylation, remains to be addressed in future studies.

Materials and Methods

Genomic DNA Preparation. Live cultures of *F. nucleatum* obtained from DSMZ were propagated in modified peptone-yeast extract-glucose (PYG) medium (www.dsmz.de) in Anaerobe Atmosphere Generating Bags (Oxoid) for 48 h and then harvested by centrifugation. *E. coli* ER2566 was cultured in LB broth. Genomic DNA was isolated using the Genomic Mini DNA isolation kit (A&A Biotechnology). Genomic DNA of *M. crocodyli*, *M. agalactiae* PG2, *M. penetrans* HF-2, *M. pulmonis* UAB CTIP, *M. mycoides* ssp. *capri*, and *M. infernus* ME was obtained from D. Brown (University of Florida, Gainesville, FL), C. Citti (INRA ENVT, Toulouse, France), Y. Sasaki (National Institute of Infectious Diseases, Tokyo, Japan), P. Sirand-Pugnet (Université Bordeaux, Bordeaux, France), F. Thiaucourt (Centre International de Recherche en Agronomie pour le Développement, Montpellier, France), and W. Whitman (University of Georgia, Athens, GA), respectively.

DNA Methylation in Vitro. Two micrograms of phage λ or *E. coli* genomic DNA was suspended in 49 μL of 10 mM Tris·HCl (pH 8.0), 50 mM NaCl, 1 mM DTT, and 160 μM SAM; and methylated with 1 μL of M.MpeI (1 mg/mL), 1 μL of M.SssI (4 U/μL; NEB) for 4 h at 37 °C, or left unmethylated. Product was purified by chloroform extraction, precipitated with LiCl/ethanol, and suspended in water.

Bisulfite Conversion and Sample Workup. Bisulfite conversion was carried out using the EZ Methylation Kit Gold (Zymo Research). For the analysis of methylation at randomly selected sites by Sanger sequencing, we used MethPrimer (33) designed primers (*SI Appendix*, Table S6) and DreamTaq DNA polymerase (Fermentas) for PCR amplification. For the analysis of methylation at the genome level by reversible terminator sequencing, we amplified bisulfite converted genomic DNA using the GenomePlex Whole Genome Amplification Kit (Sigma). The preparation of Illumina MiSeq libraries and the actual sequencing were carried out by Genomed as a commercial service.

Analysis of Genome Sequencing Data. Paired end DNA reads were obtained from Illumina MiSeq reversible terminator sequencing. Raw sequences included primers or primer concatamers from the genome amplification step. Undisclosed primer sequences were guessed with TagCleaner (34) and removed with the cutadapt program (35) (requiring a remaining minimum length of 25 bp), followed by additional cutting of 5 bp from either side. Reads were mapped to the sequenced *M. penetrans* and *E. coli* genomes using the Bismark program (36), which relies on bowtie (37) for efficient alignment of reads. Because of the genome amplification step, sequence coverage was very uneven. Hence we did not attempt to remove paired end read redundancy and instead used in-house software to map reads to sites. Methylation information for different sites was given equal weight independent of the number of reads for the analysis of methylation target sequences. For unbiased determination of methyltransferase target sequences, these were enumerated and ordered according to the fraction of highly methylated sites.

Cloning. The genomic DNA of *M. penetrans* strain HF-2 (12) was obtained from Yuko Sasaki (National Institute of Infectious Diseases, Tokyo, Japan). The gene encoding M.MpeI was amplified by PCR and cloned into the NcoI and XhoI sites of pET-28a (Novagen, Km^R), yielding a full-length ORF with a C-terminal LEHHHHHH tag. *M. penetrans* has a nonstandard genetic code, therefore four TGAs in the cloned sequence were mutated to TGG with Pfu Plus! DNA polymerase (Eurx) using conventional point mutagenesis. The obtained pET-28a::MMpe construct coded for a protein with the unintended mutations Q68R and K71R that might result from the natural variation between strains. An S295P mutation was introduced deliberately to prevent proteolysis.

Protein Expression and Purification. *E. coli* ER 2566 strain (NEB) was transformed with pET-28a::MMpe and grown overnight on LB-agar supplemented with 1% glucose. Selected colonies were used to inoculate a preculture grown in LB medium with 0.1% glucose at 37 °C. After 8 h, 20 mL of the preculture was added to 1 L of Terrific broth supplemented with 1 mM MgCl$_2$, 0.05% glucose, and 0.2% lactose. The expression culture was grown for 18 h at 28 °C and harvested by centrifugation (20 min; 4,000 × *g*). The pellet was suspended in 20 mL of disruption buffer [20 mM Tris·HCl (pH 7.5), 100 mM NaCl, 10 mg/mL lysozyme, 1 mg/mL DNase I, and protease inhibitor mixture from Sigma], incubated for 15 min at 37 °C and disrupted by sonication. The cell lysate was cleared by ultracentrifugation (30 min; 190,000 × *g*) and precipitated with an equal volume of a saturated (NH$_4$)$_2$SO$_4$ solution. After 1 h incubation on ice the salted-out proteins were removed by centrifugation (30 min; 10,000 × *g*). The soluble fraction was titrated to pH 8.0 using 1 M Tris·HCl (pH 8.5) and loaded onto a Nickel-NTA column (Qiagen) preequilibrated with 20 mM Tris·HCl (pH 7.5), 300 mM NaCl, 20 mM imidazole, and 0.01% nonaethylene glycol monododecyl ether. The column was extensively washed with the same buffer and then with 20 mM Tris·HCl (pH 7.5), 300 mM NaCl, 40 mM imidazole, and 0.01% nonaethylene glycol monododecyl ether. M.MpeI was eluted with 20 mM Tris·HCl (pH 7.5), 300 mM NaCl, and 100 mM imidazole, and subjected to gel filtration on a HiLoad 16/60 Superdex 75 column (GE Healthcare) in 10 mM Tris·HCl (pH 7.5), 50 mM NaCl, and 1 mM DTT.

Crystal Structure Determination. M.MpeI (5 mg/mL) in 10 mM NaOH·Hepes (pH 8), 120 mM NaCl, 5% glycerol (wt/vol), and dsDNA (5′-GTTCAG5mCG-CATGTG-3′ and 5′-CCACATG5FCGCTGAA-3′) in 10 mM Tris·HCl (pH 8.0) and SAM were mixed 1:1:1. Crystals were obtained from 10% PEG 3350, 150 mM NaCl, and 50 mM sodium citrate (final pH, 5.6) and cryoprotected with glycerol (25% vol/vol final concentration). Diffraction data to 2.15 Å were collected at 100 K and 0.9795 Å (IO2 beamline, Diamond Light Source) and processed with MOSFLM and SCALA (38). DIBER (39) confirmed the presence of protein and DNA in the crystals. The structure was solved using MOLREP (38) and the CHAINSAW (38) modified M.HhaI model (PDB code 2C7P). ARP/wARP (40) was used for model building, COOT (38), REFMAC (38), and CNS (41) for refinement. Of the final model, 97.9% protein residues were in favored Ramachandran plot regions [MolProbity (42)]. *SI Appendix*, Table S9 summarizes data collection and refinement statistics. Atomic coordinates and structure factors for the M.MpeI-DNA complex structure have been deposited in the PDB under accession code 4DKJ.

ACKNOWLEDGMENTS. We thank D. Brown, C. Citti, Y. Sasaki, P. Sirand-Pugnet, F. Thiaucourt, and W. Whitman for providing genomic DNA of *M. crocodyli*, *M. agalactiae* PG2, *M. penetrans* HF-2, *M. pulmonis* UAB CTIP, *M. mycoides* ssp. *capri*, and *M. infernus* ME, respectively; S. Klimasauskas for identifying the site of proteolytic cleavage in the original M.MpeI protein; and T. Sorensen and J. Sandy for granting beam time and assisting during data collection at the IO2 beamline of Diamond Light Source. M.W. was supported by a scholarship from the European Social Fund, Human Capital Operational Programme (UDA-POKL.08.02.01-14-041/09). H.C. is funded by the TEAM program cofinanced by the Foundation for Polish Science and the European Union European Regional Development Fund [Grant TEAM/2010-6/1 (to M.B.)]. Funding was also provided by European Community Seventh Framework Programme "Proteins in Health and Disease" Grant GA 229676 and Polish Ministry of Science Grants N N302 654640 (to M.B.), N N301 425038 (to H.C.), and IP2011 0609 71 (to M.W.).

1. Li E, Bestor TH, Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* 69(6):915–926.
2. Shen JC, Rideout WM, 3rd, Jones PA (1992) High frequency mutagenesis by a DNA methyltransferase. *Cell* 71(7):1073–1080.
3. Shen JC, Rideout WM, 3rd, Jones PA (1994) The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res* 22(6):972–976.
4. Jeltsch A (2010) Molecular biology. Phylogeny of methylomes. *Science* 328(5980):837–838.
5. Lister R, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315–322.
6. Cooper DN, Krawczak M (1989) Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum Genet* 83(2):181–188.
7. Sved J, Bird A (1990) The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci USA* 87(12):4692–4696.
8. Roberts RJ, Vincze T, Posfai J, Macelis D (2010) REBASE—a database for DNA restriction and modification: Enzymes, genes and genomes. *Nucleic Acids Res* 38(Database issue):D234–D236.
9. Karlin S, Burge C, Campbell AM (1992) Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res* 20(6):1363–1370.
10. Gelfand MS, Koonin EV (1997) Avoidance of palindromic words in bacterial and archaeal genomes: A close connection with restriction enzymes. *Nucleic Acids Res* 25(12):2430–2439.
11. Renbaum P, et al. (1990) Cloning, characterization, and expression in *Escherichia coli* of the gene coding for the CpG DNA methylase from *Spiroplasma* sp. strain MQ1(M.SssI). *Nucleic Acids Res* 18(5):1145–1152.
12. Sasaki Y, et al. (2002) The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. *Nucleic Acids Res* 30(23):5293–5300.
13. Kapatral V, et al. (2002) Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586. *J Bacteriol* 184(7):2005–2018.
14. Chambaud I, et al. (2001) The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res* 29(10):2145–2153.
15. Brown DR, et al. (2011) Genome sequences of *Mycoplasma alligatoris* A21JP2T and *Mycoplasma crocodyli* MP145T. *J Bacteriol* 193(11):2892–2893.
16. Levinson G, Gutman GA (1987) Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Mol Biol Evol* 4(3):203–221.
17. van Belkum A, Scherer S, van Alphen L, Verbrugh H (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* 62(2):275–293.
18. Shaw BM, Simmons WL, Dybvig K (2012) The Vsa shield of *Mycoplasma pulmonis* is antiphagocytic. *Infect Immun* 80(2):704–709.
19. Roberts RJ, et al. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res* 31(7):1805–1812.
20. Osterman DG, DePillis GD, Wu JC, Matsuda A, Santi DV (1988) 5-Fluorocytosine in DNA is a mechanism-based inhibitor of HhaI methylase. *Biochemistry* 27(14):5204–5210.
21. Morana A, et al. (2002) Stabilization of S-adenosyl-L-methionine promoted by trehalose. *Biochim Biophys Acta* 1573(2):105–108.
22. Neely RK, et al. (2005) Time-resolved fluorescence of 2-aminopurine as a probe of base flipping in M.HhaI-DNA complexes. *Nucleic Acids Res* 33(22):6953–6960.
23. Kumar S, et al. (1997) DNA containing 4′-thio-2′-deoxycytidine inhibits methylation by HhaI methyltransferase. *Nucleic Acids Res* 25(14):2773–2783.
24. Gerasimaitė R, Merkienė E, Klimašauskas S (2011) Direct observation of cytosine flipping and covalent catalysis in a DNA methyltransferase. *Nucleic Acids Res* 39(9):3771–3780.
25. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci USA* 95(19):11163–11168.
26. Firczuk M, Wojciechowski M, Czapinska H, Bochtler M (2011) DNA intercalation without flipping in the specific ThaI-DNA complex. *Nucleic Acids Res* 39(2):744–754.
27. Song J, Teplova M, Ishibe-Murakami S, Patel DJ (2012) Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. *Science* 335(6069):709–712.
28. Klimasauskas S, Kumar S, Roberts RJ, Cheng X (1994) HhaI methyltransferase flips its target base out of the DNA helix. *Cell* 76(2):357–369.
29. Bashtrykov P, Ragozin S, Jeltsch A (2012) Mechanistic details of the DNA recognition by the Dnmt1 DNA methyltransferase. *FEBS Lett* 586(13):1821–1823.
30. Krieg AM, et al. (1995) CpG motifs in bacterial DNA trigger direct B-cell activation. *Nature* 374(6522):546–549.
31. Moldoveanu Z, Love-Homan L, Huang WQ, Krieg AM (1998) CpG DNA, a novel immune enhancer for systemic and mucosal immunization with influenza virus. *Vaccine* 16(11-12):1216–1224.
32. Hemmi H, et al. (2000) A Toll-like receptor recognizes bacterial DNA. *Nature* 408(6813):740–745.
33. Li LC, Dahiya R (2002) MethPrimer: Designing primers for methylation PCRs. *Bioinformatics* 18(11):1427–1431.
34. Schmieder R, Lim YW, Rohwer F, Edwards R (2010) TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* 11:341.
35. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17(1):10–12.
36. Krueger F, Andrews SR (2011) Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27(11):1571–1572.
37. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
38. Collaborative Computational Project, Number 4 (1994) The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* 50(Pt 5):760–763.
39. Chojnowski G, Bochtler M (2010) DIBER: Protein, DNA or both? *Acta Crystallogr D Biol Crystallogr* 66(Pt 6):643–653.
40. Langer G, Cohen SX, Lamzin VS, Perrakis A (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* 3(7):1171–1179.
41. Brünger AT, et al. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* 54(Pt 5):905–921.
42. Lovell SC, et al. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* 50(3):437–450.