Published online 4 September 2016 *Nucleic Acids Research, 2016, Vol. 44, No. 21* ***10177–10185***
doi: 10.1093/nar/gkw774

Distinctive Klf4 mutants determine preference for DNA methylation status

Hideharu Hashimoto[1], Dongxue Wang[1], Alyse N. Steves[2], Peng Jin[3], Robert M. Blumenthal[4], Xing Zhang[1] and Xiaodong Cheng[1,2,*]

[1]Department of Biochemistry, Emory University, Atlanta, GA 30322, USA, [2]Genetics and Molecular Biology Graduate Program, Laney Graduate School, Emory University, Atlanta, GA 30322, USA, [3]Department of Human Genetics, Emory University, Atlanta, GA 30322, USA and [4]Department of Medical Microbiology & Immunology and Program in Bioinformatics, The University of Toledo College of Medicine & Life Sciences, Toledo, Ohio 43614, USA

Received July 08, 2016; Revised August 22, 2016; Accepted August 23, 2016

ABSTRACT

Reprogramming of mammalian genome methylation is critically important but poorly understood. Klf4, a transcription factor directing reprogramming, contains a DNA binding domain with three consecutive C2H2 zinc fingers. Klf4 recognizes CpG or TpG within a specific sequence. Mouse Klf4 DNA binding domain has roughly equal affinity for methylated CpG or TpG, and slightly lower affinity for unmodified CpG. The structural basis for this key preference is unclear, though the side chain of Glu446 is known to contact the methyl group of 5-methylcytosine (5mC) or thymine (5-methyluracil). We examined the role of Glu446 by mutagenesis. Substituting Glu446 with aspartate (E446D) resulted in preference for unmodified cytosine, due to decreased affinity for 5mC. In contrast, substituting Glu446 with proline (E446P) increased affinity for 5mC by two orders of magnitude. Structural analysis revealed hydrophobic interaction between the proline's aliphatic cyclic structure and the 5-methyl group of the pyrimidine (5mC or T). As in wild-type Klf4 (E446), the proline at position 446 does not interact directly with either the 5mC N4 nitrogen or the thymine O4 oxygen. In contrast, the unmethylated cytosine's exocyclic N4 amino group (NH_2) and its ring carbon C5 atom hydrogen bond directly with the aspartate carboxylate of the E446D variant. Both of these interactions would provide a preference for cytosine over thymine, and the latter one could explain the E446D preference for unmethylated cytosine. Finally, we evaluated the ability of these Klf4 mutants to regulate transcription of methylated and unmethylated promoters in a luciferase reporter assay.

INTRODUCTION

The control of gene expression in mammals relies substantially on the methylation status of genomic DNA. Mammalian DNA methyltransferases methylate cytosines at the ring carbon 5 position, generating 5-methylcytosine (5mC), usually within the dinucleotide sequence context of CpG (1–3) or CpA (4–9). As CpG is symmetrical with the same sequence on both DNA strands, methylation yields a symmetric modification pattern (Figure 1A) that would be transiently hemimethylated (methylated on one strand only) following replication. In contrast, CpA/TpG is intrinsically hemimethylated, meaning that the normal 5-carbon methylation of thymine (5mU) is always present, while the paired CpA may or may not be methylated (Figure 1A).

A critical role in modulating DNA methylation is played by proteins called 'reprogramming factors'. The genome-wide levels of 5mCpA/TpG (and not of 5mCpG/5mCpG) undergo dynamic changes during germ line differentiation (6), during brain development from fetus to young adult (7) and in the neonatal prospermatogonia-to-spermatogonia transition (8,10,11). In embryonic stem (ES) cells, around 25% of the cytosine methylations occur in non-CpG contexts, mainly CpA (6). This CpA methylation disappeared upon induced differentiation of ES cells, and was restored in induced pluripotent stem (iPS) cells by the four Yamanaka reprogramming factors (Oct3/4, Sox2, c-Myc, and Klf4) (6). The Yamanaka reprogramming factors (12), as well as self-renewal regulators such as the homeobox protein Nonag, recognize sequences containing CpA/TpG (13).

A key role played by the reprogramming factors might be the ability to recognize DNA methylation status. Klf4 is one of 26 members of the specificity protein/Krüppel-like factor (Sp/Klf) family of zinc finger (ZnF) transcription factors (14–16). Depending on tissue context, it can act as a tumor suppressor, oncogene, or both (17). Klf4 protein has an N-terminal domain showing no similarity to any known structures (XC, personal observation via threading analysis) and a C-terminal DNA-binding domain composed of

*To whom correspondence should be addressed. Tel: +1 404 727 8491; Fax: +1 404 727 3746; Email: xcheng@emory.edu

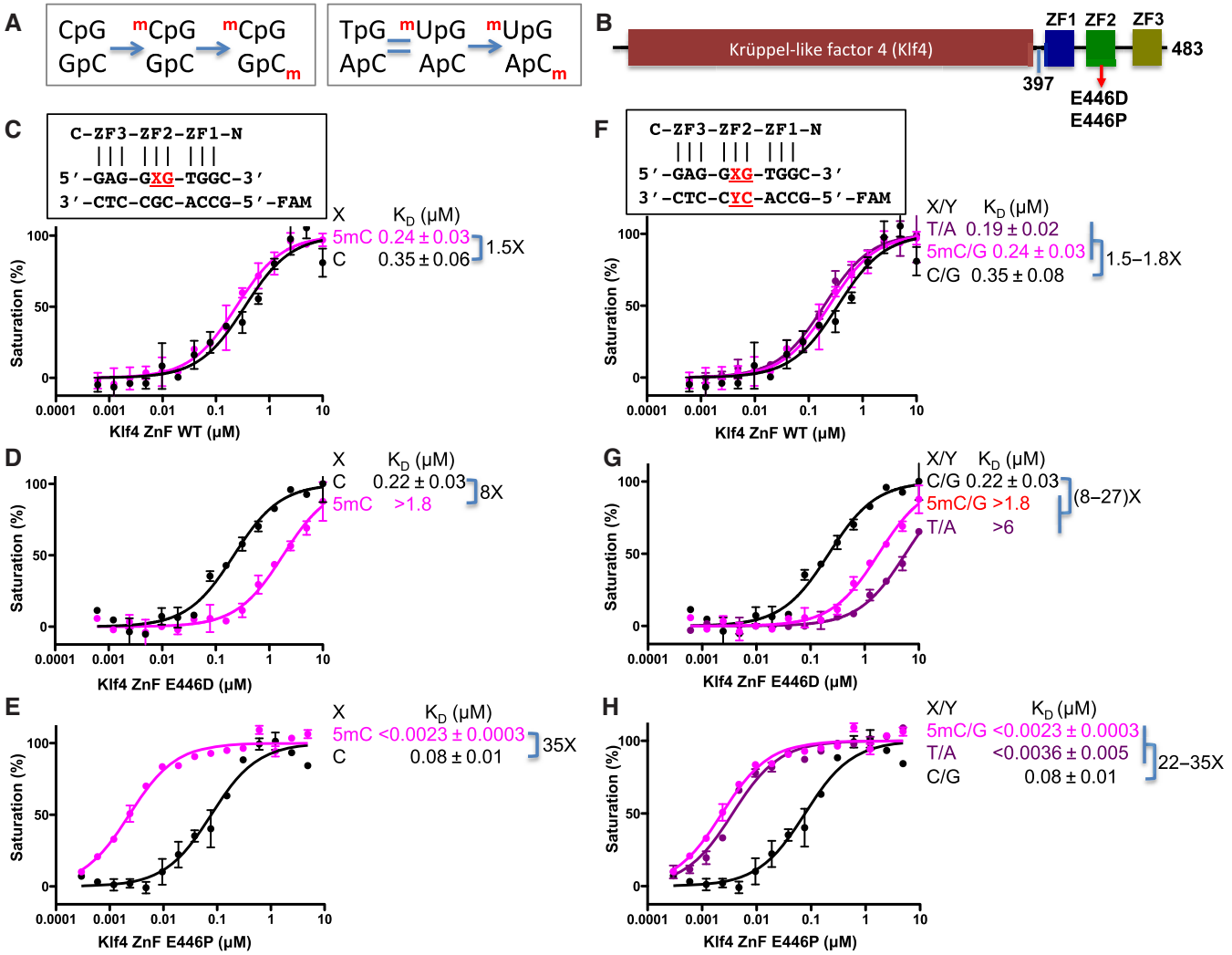

Figure 1. 5mCpG, TpG and unmodified CpG binding by three variants of Klf4. (**A**) Similarity and difference between CpG and TpG dinucleotides. Bases in red have a methyl group on the 5-carbon. (**B**) Schematic representation of mouse Klf4, containing a C-terminal Zinc Finger (ZnF) DNA binding domain comprising three fingers in tandem. Residues 397–483 were used in the DNA binding assays. (**C**–**E**) Comparison by fluorescence polarization of Klf4 DNA binding domain, WT (E446), D446 and P446 variants on oligos containing unmodified C or 5mC (hemi-methylated). (**F**–**H**) Comparison by fluorescence polarization of Klf4 WT (E446), D446 and P446 variants on oligos containing methylated CpG, TpG and unmodified CpG. We note that P466 variant binds too tightly against methylated DNA (K_D being lower than probe concentration of 5 nM) (panels E and H).

three standard Krüppel-like zinc fingers (Figure 1B). Recent studies from us and others indicate that Klf4 binds methylated DNA (18–20). The consensus binding elements for Klf4, determined by either classic base-specific mutagenesis [5′-(A/G)(G/A)**GGYG**Y-3′] (15) or ChIP-seq [5′-G**GGYG**(T/G)GG-3′] (13), share a central **GGYG**, where Y is pyrimidine (C or T). The consensus contains either CpG, which can be methylated, or TpG, which is intrinsically methylated on one strand and can be methylated on the other strand (CpA) by DNA methyltransferase 3a or 3b (4,5,21) (Figure 1A).

Previously, we showed that the *in vitro* binding affinity of the mouse Klf4 DNA binding domain for methylated DNA oligonucleotide is only slightly stronger (∼1.5X) than that for the corresponding unmodified oligonucleotide (20). In an attempt to better understand discrimination between methylated 5mCpG (or TpG) and unmodified CpG, we designed two Klf4 mutants affecting the residue that contacts the methylated base (Glu446) (20). We analyzed their interactions with methylated and unmethylated DNA both biochemically and structurally, and evaluated the transcription potentials of these Klf4 mutants in a luciferase reporter assay.

MATERIALS AND METHODS

Mutagenesis, protein expression and purification

Glutathione S-transferase (GST)-tagged mouse Klf4 ZnF1-3 fragment (Uniprot Q60793; residues 396–483; pXC1248) and its mutants Glu446-to-Pro (E446P; pXC1328), Glu446-to-Asp (E446D; pXC1411) were cloned into the pGEX6P-1 vector and expressed in *Escherichia coli* BL21(DE3)-RIL codon plus (Stratagene) as described (20). Bacterial cells

were cultured at 37°C in Luria–Bertani medium, the temperature was shifted to 16°C at OD_{600nm} = 0.5, adding $ZnCl_2$ to 25 μM. Supplying 0.2 mM isopropyl β-D-1-thiogalactopyranoside for 16 h induced the Klf4 proteins. The bacteria were harvested and lysed by sonication in 20 mM Tris-HCl (pH 7.5), 250 mM NaCl, 5% (v/v) glycerol and 0.5 mM tris(2-carboxyethyl)phosphine (TCEP), followed by centrifugation for 60 min at 16 000 rpm. After purification on Glutathione Sepharose 4B (GE Healthcare), the GST tag on the recombinant protein was removed by PreScission protease (purified in-house), resulting in the additional N-terminal residues Gly-Pro-Leu-Gly-Ser (GPLGS) relative to the native sequence. Protein was further purified on tandem HiTrap-Q-SP columns and Superdex-200 (16/60) (GE Healthcare) and concentrated in 20 mM Tris–HCl (pH 7.5), 200 mM NaCl, 5% (v/v) glycerol and 0.5 mM TCEP. The yields of the mutant proteins were similar to that of the wild-type protein, but E446P alone was difficult to concentrate to more than ∼1 mg/ml under these conditions. Instead, E446P mutant and double strand oligonucleotides were mixed and then concentrated together to ∼8 mg/ml.

DNA binding assay by fluorescence polarization

Fluorescence polarization assays for DNA binding were performed in 20 mM Tris–HCl (pH 7.5), 150 mM NaCl, 5% (v/v) glycerol and 0.5 mM TCEP at room temperature (∼22°C) using a Synergy 4 Microplate Reader (BioTek) as described (20). Fluorescently labeled double-stranded DNA probe (5 nM) and various amounts of Klf4 proteins, with a final volume of 40 μl, were incubated in a 384-well plate for 0.5 h before measurement. The sequences of 6-carboxy-fluorescein (FAM)-labeled double-stranded oligonucleotides were 5′-GAG GXG TGG C-3′ and FAM-5′-TTG CCA CGC CTC-3′ (where X = C or 5mC) or 5′-GAG GTG TGG C-3′ and FAM-5′-TTG CCA CAC CTC-3′. Curves were fit individually using PRISM 5.0 software. K_D values were calculated as [mP] = [maximum mP] × [C]/(K_D + [C]) + [baseline mP], where [mP] is millipolarization and [C] is protein concentration. Averaged K_D and its standard error were reported. We have found that the absolute magnitude of binding affinity by Klf4 is sensitive to the percentage of glycerol and concentration of NaCl used in the buffer, though the ratios of K_D values on the different DNA substrates are only minimally affected. We also take extra precaution to use the same bottle of buffer for most assays.

Crystallography

The concentrated wild-type Klf4 and E446D mutant proteins (∼8 mg/ml) were incubated with annealed oligonucleotides at an equimolar ratio for 0.5 h on ice before crystallization. E446P variant and double strand oligonucleotides were mixed and concentrated together to ∼8 mg/ml. All of the final solutions contained 0.8 mM protein–DNA complex. Crystals were obtained by the sitting-drop method; the mother liquor contained 100 mM Tris-HCl (pH 8.5), 250 mM NaCl and 20% (w/v) polyethylene glycol 8000. Crystals grew within 3 days at 16°C. The crystals were flash frozen by plunging into liquid nitrogen. X-ray diffraction data were collected at the SER-CAT beamline at the Advanced Photon Source (22-ID and 22-BM), Argonne National Laboratory. HKL2000 (22) or XDS (23) and anisotropic server (24) were used for the data processing. The structures were solved by molecular replacement with the coordinates of 4M9E as an initial searching model using the Phaser (25). Model refinement (including hydrogen atoms) was performed with COOT (26) and PHENIX (27). Molecular graphics were generated with the Pymol program (DeLano Scientific LLC).

Luciferase reporter assay

HEK293T cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% heat-inactivated fetal bovine serum (FBS), antibiotics (100 U/ml penicillin and 100 μg/ml streptomycin) and non-essential amino acids. HEK293T cells were transfected in a 96-well plate with various expression plasmids together with the pGL4.2-Basic-2XCR4 (2C or 2T) (0.1 μg/well) and renilla (0.1 μg/well) using Lipofectamine 2000. After 36 h post-transfection, cells were lysed for measurement of luciferase activity using Dual-Glo® Luciferase Assay System (Promega Corporation). Luminescence was read on a Synergy 4 microplate reader (BioTek). Firefly luciferase activities were corrected for the renilla activity and normalized to the negative control as 1.0 (n = 3). The statistics significance was analyzed by one-way analysis of variance (ANOVA) with post-hoc Tukey HSD (Honestly Significant Difference) test.

pMXs-full length Klf4 (WT, E446P and E446D) (28), and pcDNA3.1-Flag-full length Klf4 (WT, E446P and E446D), pGL4.2-Basic-2XCR4 (2C) and pGL4.2-Basic-2XCR4 (2T) were generated. Two copies of CR4, which contains binding sites of mouse Klf4, Oct4, and Sox2/Nanag were subcloned into pGL4.2-[Luc/Puro] vector (generating 2C), and the Klf4 binding site mutated Cyt-to-Thy by PCR (generating 2T). The pRL-SV40 vector was used as a transfection control. Anti-Klf4 polyclonal antibody (Abcam ab129473, lot# GR147393-1; 1:5000), and anti-rabbit IgG (SouthernBiotech, cat. No. 4050-05; 1:2000) were used for western blotting.

RESULTS

Development of two Klf4 mutants with increased selectivity between methylated and unmethylated DNA

In an earlier study, we found that Glu446 of Klf4 exhibits one of the largest conformational differences when bound to methylated versus unmethylated CpG DNA. In the structure of the Klf4 bound to methylated DNA (20), the methyl group of 5mC in the recognition strand makes a van der Waals contact with the aliphatic carbon Cγ and forms a weak C-H…O type of hydrogen bond with one of the carboxylate oxygen atoms of Glu446. Similar interaction between glutamate and methylated cytosine has been observed in Kaiso (29), Zfp57 (30), Wilms tumor protein (WT1) (31) and early growth response factor (Egr1/Zif268) (31,32,33). Interestingly, in an early phage display study of Egr1/Zif268, an aspartate (D) residue (rather than E in the

wild type) shows a distinct preference for binding (unmodified) cytosine (34). This observation led Choo and Klug to comment that 'The physical basis for the interaction of aspartate/glutamate and cytosine is not yet clear, since hydrogen bonding contacts between these groups have yet to be observed in zinc finger cocrystal structures' (35). In more recent bacterial one-hybrid experiments where only unmodified bases were present, aspartate was again found to preferentially juxtapose to cytosine (36). However, as noted by Choo and Klug, the aspartate/glutamate interaction with cytosine requires more study; and particularly in the context of important, 5mC-responsive regulatory proteins. We thus generated a Glu446-to-Asp (E446D) mutant in Klf4, to test whether the substitution would reverse the order of binding preference and result in a mutant transcription factor with higher affinity for unmethylated DNA.

To complement the E446D study, and potentially generate a useful research tool, we were also interested in generating Klf4 variants that strongly prefer 5mC. The mismatch repair endonuclease MutH (37) uses a proline juxtaposed to a methyl group in its hemimethylated recognition sequence (though the methyl occurs on adenine rather than on C in that case). In addition, proline was found to preferentially contact thymine, presumably through interaction with the 5-methyl group (36). Previously we replaced the corresponding glutamate in WT1 with proline, resulting in a WT1 variant that highly prefers 5mC over unmodified C (31). For these reasons, and to further test the potential role of proline in recognizing methylated DNA, we generated a Glu446-to-Pro (E446P) mutant of Klf4.

Fluorescence polarization was used to measure the dissociation constants (K_D) of Klf4 mutants and double-stranded oligonucleotides containing a single CpG dinucleotide (5′-GAG G**C**G TGG C-3′) with and without methylation at the bolded C. First, we repeated our previous observation for WT Klf4 that the binding affinity for methylated DNA is only slightly stronger (1.5X) than that of unmodified DNA under the assay conditions (Figure 1C). We kept the unmodified C on the bottom strand, because with ZnF proteins, only one DNA strand is involved in base-specific contacts (the "top" strand, depicted as containing the recognition sequence), while the bottom strand interacts mainly with water molecules (20). As predicted, the mutant E446D displayed a preference for unmodified C compared with 5mC, under our conditions by a factor of eight (Figure 1D). The selectivity of the E446D variant for C compared with 5mC stems from a 7.5-fold decrease in affinity for 5mC, together with a slight (1.6X) increase in affinity for C (compare Figure 1C and 1D). Also as predicted, the E446P variant of Klf4 strongly distinguished sequences containing 5mC from the C-containing oligonucleotide, by a factor of ∼35 (Figure 1E). Compared to the WT, the E446P affinity for 5mC increased dramatically by two orders of magnitude, while the increase for C was modest (∼4-fold). As mentioned above, the consensus binding elements for Klf4 can contain either CpG or TpG. We repeated the binding assays for oligonucleotides containing TpG, and found that they had similar affinities to those of the 5mCpG duplexes for WT and E466P, but much lower affinities for E466D (Figure 1F–H).

Structural basis for recognition of 5mC/T versus unmodified C

While the relative binding preferences of the mutant Klf4 proteins agreed with our predictions, which were based on analogy to other proteins, the structural basis for such preferences had not been determined. To understand why E446P and E446D respond so differently to 5mC and unmodified C, we determined the co-crystal structures of each variant. We used 10-bp duplexes containing 5mCpG or TpG within the consensus sequence in complex with E446P, or unmodified CpG in complex with E446D (Table 1). In addition, we also determined the structure of WT in complex with TpG containing oligo (Table 1), to be compared with the previously-solved structure of WT with methylated CpG (20). The structures were determined to the resolution range of 2.0–2.5 Å. Except for the side chain of residue 446 (see below) and a rotation of ZnF3 (Supplementary Figure S1), the overall structure of the Klf4 is essentially unchanged among these complexes. The three zinc fingers of Klf4 bind in the major groove of the DNA (Figure 2A). ZnF3 interacts with the 5′ GAG sequence, ZnF2 interacts with the central GTG or GCG (methylated or unmodified) and ZnF1 interacts with the 3′ TGG. We focus here on the methylation-responsive ZnF2.

As shown above, Klf4 has very similar affinities for sequences having G-5mC-G or G-T-G as the central triplet (Figure 1F). Just as in Klf4 WT-5mC interactions (20), we found that the methyl group of thymine in the top strand makes van der Waals contacts with the guanidine group of Arg443, which in turn forms bifurcated hydrogen bonds with the 3′ guanine (Figure 2B), forming a methyl-Arg-Gua triad (33). The thymine methyl group also interacts with the aliphatic Cγ atom and carboxylate group of Glu446, forming a weak (3.6 Å) C-H…O type of hydrogen bond (Figure 2C) – a common but underappreciated interaction in biomolecules and molecular recognition (38). The Ade base-paired with this Thy does not exhibit any side chain specific interactions, but is engaged with a layer of water molecules in both the major and minor grooves (Figure 2C). Together, these observations can account for the nearly-identical Klf4 binding affinities for the G-5mC-G and G-T-G sequences.

Considering next the E446P variant, it was entirely possible that inserting a proline in the first turn (the third residue) of the helix would destabilize it (39) and abolish binding, but obviously this did not occur (Figure 2D). Neither the WT E446 side chain nor the corresponding proline directly contact the 5mC N4 or thymine O4, consistent with lack of discrimination between C and T. E446P preserves the methyl-Arg-Gua triad, but the methyl group of thymine (or 5mC) also makes an extensive van der Waals contact with the proline residue (Figure 2E and F), which could explain the significantly enhanced binding affinity (Figure 1H).

Lastly, we consider the E446D variant of Klf4, which shows preference for unmethylated DNA (Figure 1D and G). Unlike the side chain carboxylate group of Glu446 of WT Klf4, the aspartate carboxylate group of the E446D variant forms a hydrogen bond with the cytosine N4 (Figure 2H and I) and would thus exclude thymine. Significantly, the aspartate forms a C-H…O type of hydrogen bond with the

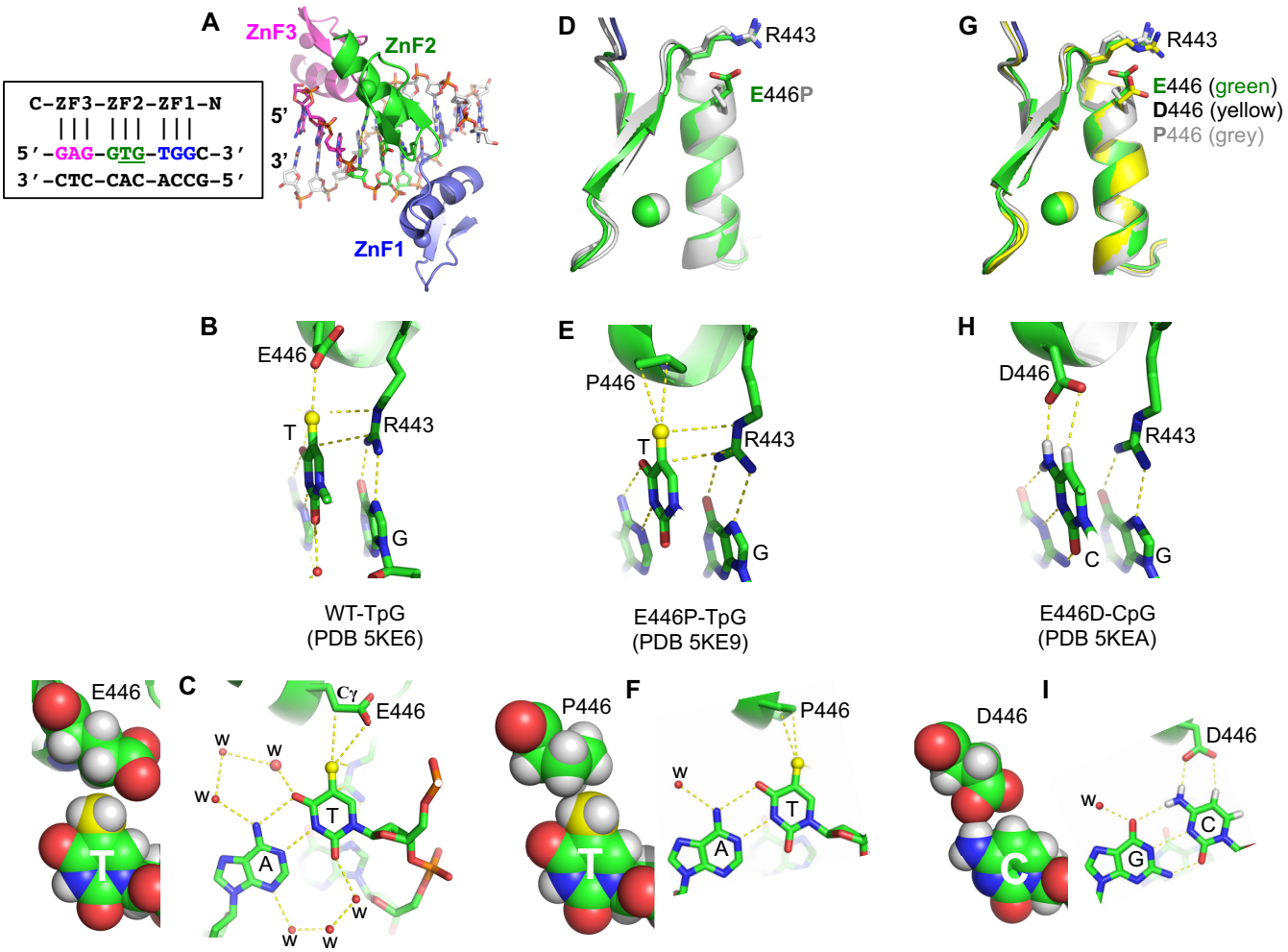

Figure 2. Structures of Klf4 DNA binding domains, WT (E446), D446 and P446, in complex with TpG and CpG duplex oligonucleotides. (**A**) The mouse Klf4 ZnF DNA binding domain binds in the major groove of DNA with ZnF1 (blue), ZnF2 (green) and ZnF3 (magenta). Each ZnF recognizesthree adjacent DNA base pairs (boxed). (**B** and **C**) Klf4 WT (E446) interactions with TpG/CpA dinucleotide. Space filling model includes hydrogen (grey), carbon (green), oxygen (red) and nitrogen atoms (blue). The methyl carbon atom attached to the ring C5 atom is colored yellow. (**D**) Superimposition of ZnF2 in WT (E446 in green) and E446P mutant (in grey). The sphere is the Zn atom. (**E** and **F**) P446 variant interactions with TpG/CpA dinucleotide. (**G**) Superimposition of ZnF2 in WT (E446 in green) and E446D mutant (in yellow), and E446P mutant (in grey). (**H** and **I**) D446 variant interacts with unmodified CpG dinucleotide. Note hydrogen atoms on the cytosine ring were shown to illustrate the C-H…O type of hydrogen bond. Modeling a methyl group onto unmodified C potentially results in repulsion with D446 in the C-specific conformation.

ring C5 atom (Figure 2H and I), such that a methyl group on C5 would sterically obstruct Asp446 from adopting the unmodified-Cyt-specific conformation, perhaps explaining the diminished binding to the methylated oligo by E446D variant (Figure 1G).

Evaluation of transcriptional regulatory activity of Klf4 mutants

Having generated Klf4 variants with increased binding selectivity for or against methylated (5mCpG or TpG) versus unmodified DNA (CpG), we next asked whether the substitutions affected the ability to activate transcription. We first modified a pGL4.2-Basic-6XCR4 reporter plasmid (28) to have two copies of CR4, which contains at different positions the binding sites of Klf4, Oct4 and Sox2/Nanog (Figure 3A). We introduced either TpG ('2T' reporter plasmid, representing methylated DNA) or CpG ('2C' reporter plasmid, representing unmodified DNA) within the Klf4 binding sites upstream of the *lux* luciferase gene (Figure 3A). The Klf4 E446D variant exhibited the highest transactivation activity following transfection of either LTR or CMV into HEK293 cells (Figure 3B). E446D exhibited 2–2.5X greater activities in transactivation on the 2C reporter than with 2T. The same trend is observed when Klf4 was transfected together with native Oct4, Sox2 and Nanog (OSN in Figure 3C; the red lines between panels B, C and D indicate the changed scales of the luciferase activity along y axes), and with modified factors fused to the murine Yap transcription activation domain (28) (OySyNy in Figure 3D and Supplementary Figure S2). While overall transactivation activities are lower than those of the D446 variant, both E446 (WT) and the P446 variant seem to have higher activity with 2T than with 2C under the CMV driven expression (Figure 3C and D).

Table 1. Summary of X-ray diffraction and structural refinement statistics (*)

Klf4 (ZnF)	Wild-Type	Wild-Type	E446P	E446D
DNA (M=5mC)	5'-GAGG**TG**TGGC-3' 3'-CTCC**AC**ACCG-5'	5'-GAGG**TG**TGGC-3' 3'-CTCC**AM**ACCG-5'	5'-GAGG**TG**TGGC-3' 3'-CTCC**AC**ACCG-5'	5'-GAGG**CG**TGGC-3' 3'-CTCC**GC**ACCG-5'
PDB	5KE6	5KE7	5KE9	5KEB
Beamline (SER-CAT)	22-BM	22-BM	22-ID	22-ID
Unit cell (Å)	a=b=50.9, c=130.5	a=b=50.8, c=130.4	a=b=51.4, c=130.6	a=b=51.1, c=131.6
Resolution (Å)	34.68-1.99 (2.04-1.99)	34.60-2.06 (2.11-2.06)	25.24-2.34 (2.42-2.34)	47.64-2.45 (2.52-2.45)
[a] R_{merge}	0.073 (0.982)	0.085 (0.821)	0.099 (0.887)	0.054 (0.390)
CC1/2*	(0.814)	(0.797)	(0.919)	(0.812)
[b] <I/σI>	16.6 (2.3)	16.1 (2.9)	16.6 (4.1)	20.4 (2.2)
Completeness (%)	99.9 (100.0)	99.9 (100.0)	99.5 (100.0)	98.0 (73.2)
Redundancy	7.0 (7.2)	7.0 (7.1)	10.7 (10.0)	7.1 (2.5)
Reflections (observed)	86,505	77,806	85,360	48,373
Unique reflections	12,435 (891)	11,193 (819)	7,979 (763)	6,808 (349)
Refinement				
Resolution (Å)	1.99	2.06	2.34	2.45
No. Reflections	12,409	11,161	7,915	6,710
[c] R_{work}/ [d] R_{free}	0.202 / 0.232	0.194 / 0.216	0.214 / 0.272	0.248 / 0.286
No. Atoms (without H)				
Protein	707	719	728	699
DNA	404	405	404	404
Zn	3	3	3	3
Solvent	89	114	14	5
B-factors (Å^2)				
Protein	45.1	37.3	73.4	68.9
DNA	36.6	29.7	54.3	53.6
Zn	36.2	32.2	68.0	66.1
Solvent	42.0	38.2	49.1	50.0
R.m.s. deviations				
Bond length (Å)	0.006	0.003	0.005	0.004
Bond angles (°)	0.7	0.7	0.6	0.5
All atom clashscore	0.5	0.0	0.5	1.0
Ramachandran (%)				
Favored	98.8	100.0	98.9	97.6
Allowed	1.2	0.0	1.2	2.4
Rotamer outliers (%)	0	0	0	0
C_β deviation	0	0	0	0

*Values in parenthesis correspond to highest resolution shell; Wavelength (1.0 Å); Space Group ($P4_32_12$ with α=β=γ=90°)
[a] $R_{merge} = \Sigma \mid I - \langle I \rangle \mid / \Sigma I$, where I is the observed intensity and <I> is the averaged intensity from multiple observations;
[b] <I/σI> = averaged ratio of the intensity (I) to the error of the intensity (σI);
[c] $R_{work} = \Sigma \mid F_{obs} - F_{cal} \mid / \Sigma \mid F_{obs} \mid$, where F_{obs} and F_{cal} are the observed and calculated structure factors, respectively;
[d] R_{free} was calculated using a randomly chosen subset (5%) of the reflections not used in refinement.

The relative activities on 2T and 2C reporters seem to be correlated to the in vitro binding affinity we measured for each variant. We were surprised to find a large difference in protein expression level among the variant proteins, particularly under CMV promoter. However, there does not appear to be a correlation between protein level and luciferase activity: when the amount of Klf4 protein produced (as reflected by western blot density) was plotted against luciferase activity (see Supplementary Figure S3), it clearly shows that E446D mutant proteins have roughly the same higher transactivation activity than WT, regardless of vastly different protein expression level. In addition, E446D showed higher luciferase levels on unmethylated (2C) reporter than the methylated (2T) reporter, as expected.

Surprisingly, when the methylated (2T) reporter fusion was combined with the CMV driven expression, that combination gave expected levels of luciferase expression (compared to the LTR driven expression with 2T reporters – compare gray circles and squares in Supplementary Figure S3), but gave much higher levels of Klf4 accumulation as judged by western blot density. While this elevated level of protein is unexpected, it further demonstrates that elevated Klf4 levels did not result in elevated luciferase expression, suggesting that the amounts of Klf4 in these assays were not a limiting factor. One caveat is that we do not know whether or how rapidly the 2C reporter becomes methylated in the HEK293 cells (40,41); any such methylation would tend to reduce the differences between 2C and 2T reporters.

DISCUSSION

The role of DNA methylation is critical in mammalian development, but we are still working to understand its mechanics. This includes the generation, maintenance and erasure of methyl marks, and also—as in this study—how those

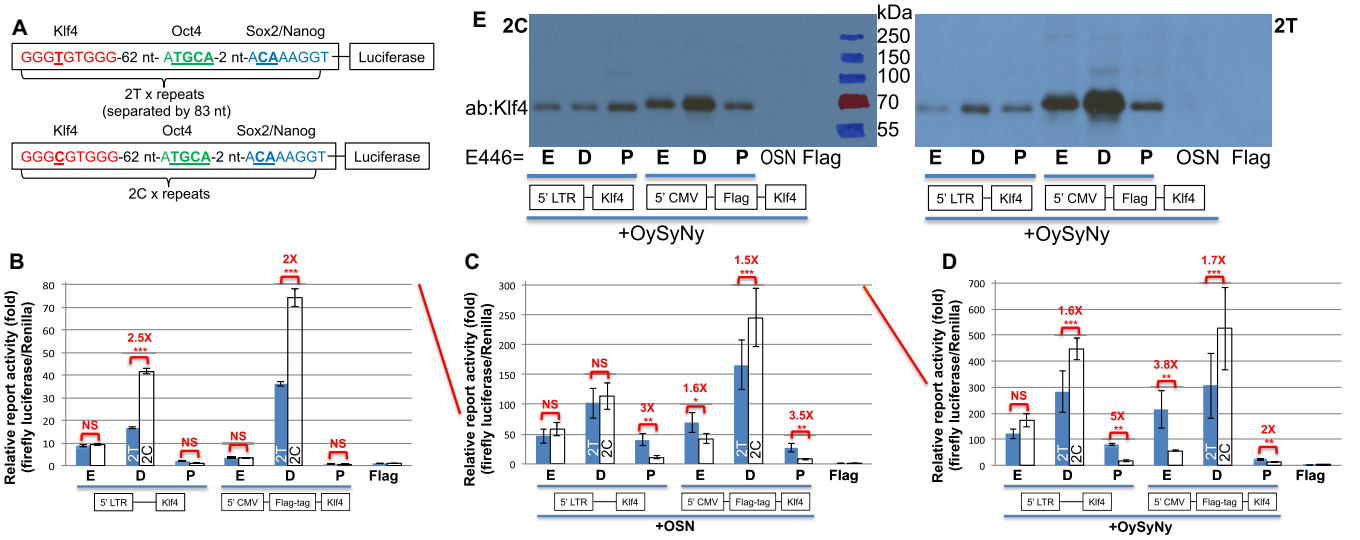

Figure 3. Comparison of transcriptional effects of full-length Klf4 WT versus D446 and P446 variants. (**A**) Schematic definition of 2T (permanently methylated due to 5-methyl on Thy) and 2C (unmethylated Cyt) reporter plasmids. (**B**) HEK293 cells were transfected in a 96-well plate with Klf4 expression plasmids along with the 2T or 2C vector and Renilla control vector. Thirty-six hours posttransfection, cells were lysed for the measurement of luciferase activity. Firefly luciferase activities were normalized based on the Renilla activity. Error bars represent SEM from three (n = 3) duplicates (***$P < 0.001$; NS, not significant). The red lines between panels B, C and D indicate the changed scales of the luciferase activity along y axes. (**C** and **D**) The different luciferase activities were measured following induction of (C) OSNK or (D) OySyNyK. Error bars represent SEM from three (n = 3) duplicates (***$P < 0.001$; **$P < 0.01$; *$P < 0.05$; NS, not significant). (**E**) Western blot analysis of HEK293 cells (~8000 cells) treated with 2C (left) or 2T (right) DNA for 36 h from the same samples shown in panel D, and probed with anti-Klf4 antibody.

methyl marks are detected. DNA 5mC is a major epigenetic signal that acts to regulate chromatin structure and ultimately gene expression. These modifications protrude into the major groove of DNA, the primary recognition surface for proteins, and change its atomic shape and pattern of electrostatic charge. In principle, such changes can alter the way in which proteins bind to their recognition sequences in DNA by strengthening the interactions, weakening them or abolishing them altogether (42). This, in turn, can modulate gene expression and control cellular metabolism and is believed to be one of the principal mechanisms underlying epigenetic processes such as differentiation, development, aging and disease.

Many transcription factors (e.g. Klf4 and MeCP2) recognize consensus-binding elements, containing either CpG/CpG, which can be methylated, or TpG/CpA, which is intrinsically methylated on one strand and can be methylated on the other strand (Figure 1A). Recent work suggests that MeCP2 binds methylated CpA sites with similar affinity to that of fully methylated CpG (43–45), similar to what was found with Klf4. It seems quite plausible that transcription factors are responsive to different states of cytosine modification [including Tet-derived 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine (46–49), as well as Tet-dependent thymine modification to 5-hydroxymethyluracil (50,51)]. This responsiveness likely makes gene activity controllable by these modifications on a much finer scale than a simple 'on' or 'off.'

Klf4 plays critical roles in many biological processes including nuclear reprogramming (12) and in regulation of immunosuppressive myeloid-derived fibrocytes in tumor metastases (52). As it is a C2H2 ZnF, protein Klf4 has a well-defined region (the middle finger) responsible for detecting DNA methylation status, making it a good focus for studies on how methyl marks are read. The E446-to-D change behaved as predicted for preferred recognition of unmodified cytosine, and the structural analysis provided a credible explanation (Figure 2I). The E446-to-P change also behaved as predicted, based on analogies to other proteins (MutH (37) and WT1 (31)) for preferential recognition of methylated cytosine, and was also consistent with the structural analysis that revealed van der Waals contacts between the proline and 5-position methyl group (Figure 2F). Interestingly, we could find no instances of naturally-occurring variation at the equivalent of E446. Using NCBI BLINK, with human Klf4 as the search seed, even proteins having just 41% identity (the lowest level we saw where the full ORF still aligns) are fully conserved at and around the E446 residue of the DNA binding domain (Supplementary Figure S4).

The set of three Klf4 variants could be useful tools in better understanding the *in vivo* roles of methylation-responsive transcription factors. We suggest that the E446D variant of Klf4 might also be useful in improving the extremely low efficiency for somatic cell reprogramming using the original Yamanaka factors (53). The same approach could also applied to Oct4 and Myc, for selective mutants of recognition of DNA modifications (or lack thereof), whereas Sox2 is insensitive to DNA modification located in the major groove (54) because the Sox2 DNA binding domain binds in the minor groove (55). We note that it may be possible to greatly improve the efficiency of generating pluripotent stem cells by employing reprogramming factors (Klf4, Oct4 and Myc) that have been engineered to

exhibit stronger preference for unmethylated DNA, such as the E446D version of Klf4.

ACCESSION NUMBERS

The X-ray structures (coordinates and structure factor files) of Klf4-DNA have been submitted to PDB under accession number 5KE6 and 5KE7 (WT-TpG), 5KE8 (E446P-5mCpG) and 5KE9 (E446P-TpG), and 5KEA and 5KEB (E446D-CpG).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Brenda Baker at the organic synthesis unit of New England Biolabs for synthesizing the oligonucleotides. The Emory University School of Medicine supported the use of Southeast Regional Collaborative Access Team (SER-CAT) 22-BM beamline at the Advanced Photon Source, Argonne National Laboratory; Use of the Advanced Photon Source was supported by the U. S. Department of Energy, Office of Science.
Author Information: H.H. performed crystallographic and DNA binding experiments, and generated E446P mutant; D.W. performed luciferase reporter assay; A.N.S. generated E446D mutant; P.J. provided retroviral vectors for the modified factors; R.M.B. performed data analysis and assisted in preparing the manuscript; X.Z. and X.C. organized and designed the scope of the study.

FUNDING

National Institutes of Health (NIH) [GM049245-23 to X.C]; National Science Foundation Graduate Research Fellowship Program [DGE-1444932 to A.N.S.]; Georgia Research Alliance Eminent Scholar [to X.C.]. Funding for open access charge: NIH.
Conflict of interest statement. None declared.

REFERENCES

1. Bestor,T., Laudano,A., Mattaliano,R. and Ingram,V. (1988) Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J. Mol. Biol.*, **203**, 971–983.
2. Okano,M., Xie,S. and Li,E. (1998) Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat. Genet.*, **19**, 219–220.
3. Okano,M., Bell,D.W., Haber,D.A. and Li,E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.
4. Ramsahoye,B.H., Biniszkiewicz,D., Lyko,F., Clark,V., Bird,A.P. and Jaenisch,R. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 5237–5242.
5. Gowher,H. and Jeltsch,A. (2001) Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: The enzyme modifies DNA in a non-processive manner and also methylates non-CpG [correction of non-CpA] sites. *J. Mol. Biol.*, **309**, 1201–1208.
6. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
7. Lister,R., Mukamel,E.A., Nery,J.R., Urich,M., Puddifoot,C.A., Johnson,N.D., Lucero,J., Huang,Y., Dwork,A.J., Schultz,M.D. *et al.* (2013) Global epigenomic reconfiguration during mammalian brain development. *Science*, **341**, 1237905.
8. Kubo,N., Toh,H., Shirane,K., Shirakawa,T., Kobayashi,H., Sato,T., Sone,H., Sato,Y., Tomizawa,S., Tsurusaki,Y. *et al.* (2015) DNA methylation and gene expression dynamics during spermatogonial stem cell differentiation in the early postnatal mouse testis. *BMC Genomics*, **16**, 624.
9. Vlachogiannis,G., Niederhuth,C.E., Tuna,S., Stathopoulou,A., Viiri,K., de Rooij,D.G., Jenner,R.G., Schmitz,R.J. and Ooi,S.K. (2015) The Dnmt3L ADD domain controls cytosine methylation establishment during spermatogenesis. *Cell Rep.*, **10**, 944–956.
10. He,Y. and Ecker,J.R. (2015) Non-CG Methylation in the human genome. *Annu. Rev. Genomics Hum. Genet.*, **16**, 55–77.
11. Hashimoto,H., Zhang,X., Vertino,P.M. and Cheng,X. (2015) The mechanisms of generation, recognition, and erasure of DNA 5-methylcytosine and thymine oxidations. *J. Biol. Chem.*, **290**, 20723–20733.
12. Takahashi,K. and Yamanaka,S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663–676.
13. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
14. Zhang,W., Shields,J.M., Sogawa,K., Fujii-Kuriyama,Y. and Yang,V.W. (1998) The gut-enriched Kruppel-like factor suppresses the activity of the CYP1A1 promoter in an Sp1-dependent fashion. *J. Biol. Chem.*, **273**, 17917–17925.
15. Shields,J.M. and Yang,V.W. (1998) Identification of the DNA sequence that interacts with the gut-enriched Kruppel-like factor. *Nucleic Acids Res.*, **26**, 796–802.
16. Nandan,M.O. and Yang,V.W. (2009) The role of Kruppel-like factors in the reprogramming of somatic cells to induced pluripotent stem cells. *Histol. Histopathol.*, **24**, 1343–1355.
17. Park,C.S., Shen,Y., Lewis,A. and Lacorazza,H.D. (2016) Role of the reprogramming factor KLF4 in blood formation. *J. Leukoc. Biol.*, **99**, 673–685.
18. Spruijt,C.G., Gnerlich,F., Smits,A.H., Pfaffeneder,T., Jansen,P.W., Bauer,C., Munzel,M., Wagner,M., Muller,M., Khan,F. *et al.* (2013) Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell*, **152**, 1146–1159.
19. Hu,S., Wan,J., Su,Y., Song,Q., Zeng,Y., Nguyen,H.N., Shin,J., Cox,E., Rho,H.S., Woodard,C. *et al.* (2013) DNA methylation presents distinct binding sites for human transcription factors. *eLife*, **2**, e00726.
20. Liu,Y., Olanrewaju,Y.O., Zheng,Y., Hashimoto,H., Blumenthal,R.M., Zhang,X. and Cheng,X. (2014) Structural basis for Klf4 recognition of methylated DNA. *Nucleic Acids Res.*, **42**, 4859–4867.
21. Baubec,T., Colombo,D.F., Wirbelauer,C., Schmidt,J., Burger,L., Krebs,A.R., Akalin,A. and Schubeler,D. (2015) Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature*, **520**, 243–247.
22. Otwinowski,Z., Borek,D., Majewski,W. and Minor,W. (2003) Multiparametric scaling of diffraction intensities. *Acta Crystallogr. A*, **59**, 228–234.
23. Kabsch,W. (2010) Xds. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 125–132.
24. Strong,M., Sawaya,M.R., Wang,S., Phillips,M., Cascio,D. and Eisenberg,D. (2006) Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 8060–8065.
25. McCoy,A.J., Grosse-Kunstleve,R.W., Adams,P.D., Winn,M.D., Storoni,L.C. and Read,R.J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.*, **40**, 658–674.
26. Emsley,P., Lohkamp,B., Scott,W.G. and Cowtan,K. (2010) Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 486–501.

27. Adams,P.D., Mustyakimov,M., Afonine,P.V. and Langan,P. (2009) Generalized X-ray and neutron crystallographic analysis: more accurate and complete structures for biological macromolecules. *Acta Crystallogr. D Biol. Crystallogr.*, **65**, 567–573.
28. Zhu,G., Li,Y., Zhu,F., Wang,T., Jin,W., Mu,W., Lin,W., Tan,W., Li,W., Street,R.C. *et al.* (2014) Coordination of engineered factors with TET1/2 promotes early-stage epigenetic modification during somatic cell reprogramming. *Stem Cell Rep.*, **2**, 253–261.
29. Buck-Koehntop,B.A., Stanfield,R.L., Ekiert,D.C., Martinez-Yamout,M.A., Dyson,H.J., Wilson,I.A. and Wright,P.E. (2012) Molecular basis for recognition of methylated and specific DNA sequences by the zinc finger protein Kaiso. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 15229–15234.
30. Liu,Y., Toh,H., Sasaki,H., Zhang,X. and Cheng,X. (2012) An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. *Genes Dev.*, **26**, 2374–2379.
31. Hashimoto,H., Olanrewaju,Y.O., Zheng,Y., Wilson,G.G., Zhang,X. and Cheng,X. (2014) Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. *Genes Dev.*, **28**, 2304–2313.
32. Zandarashvili,L., White,M.A., Esadze,A. and Iwahara,J. (2015) Structural impact of complete CpG methylation within target DNA on specific complex formation of the inducible transcription factor Egr-1. *FEBS Lett.*, **589**, 1748–1753.
33. Liu,Y., Zhang,X., Blumenthal,R.M. and Cheng,X. (2013) A common mode of recognition for methylated CpG. *Trends Biochem. Sci.*, **38**, 177–183.
34. Choo,Y. and Klug,A. (1994) Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 11168–11172.
35. Choo,Y. and Klug,A. (1997) Physical basis of a protein-DNA recognition code. *Curr. Opin. Struct. Biol.*, **7**, 117–125.
36. Gupta,A., Christensen,R.G., Bell,H.A., Goodwin,M., Patel,R.Y., Pandey,M., Enuameh,M.S., Rayla,A.L., Zhu,C., Thibodeau-Beganny,S. *et al.* (2014) An improved predictive recognition model for Cys2-His2 zinc finger proteins. *Nucleic Acids Res.*, **42**, 4800–4812.
37. Lee,J.Y., Chang,J., Joseph,N., Ghirlando,R., Rao,D.N. and Yang,W. (2005) MutH complexed with hemi- and unmethylated DNAs: Coupling base recognition and DNA cleavage. *Mol. Cell*, **20**, 155–166.
38. Horowitz,S. and Trievel,R.C. (2012) Carbon-oxygen hydrogen bonding in biological structure and function. *J. Biol. Chem.*, **287**, 41576–41582.
39. Stoll,R., Lee,B.M., Debler,E.W., Laity,J.H., Wilson,I.A., Dyson,H.J. and Wright,P.E. (2007) Structure of the Wilms tumor suppressor protein zinc finger domain bound to DNA. *J. Mol. Biol.*, **372**, 1227–1245.
40. Zang,L., Nishikawa,M., Ando,M., Takahashi,Y. and Takakura,Y. (2015) Contribution of Epigenetic Modifications to the Decline in Transgene Expression from Plasmid DNA in Mouse Liver. *Pharmaceutics*, **7**, 199–212.
41. Mitsui,M., Nishikawa,M., Zang,L., Ando,M., Hattori,K., Takahashi,Y., Watanabe,Y. and Takakura,Y. (2009) Effect of the content of unmethylated CpG dinucleotides in plasmid DNA on the sustainability of transgene expression. *J. Gene Med.*, **11**, 435–443.
42. Dantas Machado,A.C., Zhou,T., Rao,S., Goel,P., Rastogi,C., Lazarovici,A., Bussemaker,H.J. and Rohs,R. (2015) Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief. Funct. Genomics*, **14**, 61–73.
43. Guo,J.U., Su,Y., Shin,J.H., Shin,J., Li,H., Xie,B., Zhong,C., Hu,S., Le,T., Fan,G. *et al.* (2014) Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.*, **17**, 215–222.
44. Gabel,H.W., Kinde,B., Stroud,H., Gilbert,C.S., Harmin,D.A., Kastan,N.R., Hemberg,M., Ebert,D.H. and Greenberg,M.E. (2015) Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, **522**, 89–93.
45. Kinde,B., Gabel,H.W., Gilbert,C.S., Griffith,E.C. and Greenberg,M.E. (2015) Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 6800–6806.
46. Tahiliani,M., Koh,K.P., Shen,Y., Pastor,W.A., Bandukwala,H., Brudno,Y., Agarwal,S., Iyer,L.M., Liu,D.R., Aravind,L. *et al.* (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930–935.
47. Ito,S., Shen,L., Dai,Q., Wu,S.C., Collins,L.B., Swenberg,J.A., He,C. and Zhang,Y. (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, **333**, 1300–1303.
48. He,Y.F., Li,B.Z., Li,Z., Liu,P., Wang,Y., Tang,Q., Ding,J., Jia,Y., Chen,Z., Li,L. *et al.* (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, **333**, 1303–1307.
49. Pfaffeneder,T., Hackner,B., Truss,M., Munzel,M., Muller,M., Deiml,C.A., Hagemeier,C. and Carell,T. (2011) The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Ed. Engl.*, **50**, 7008–7012.
50. Pfaffeneder,T., Spada,F., Wagner,M., Brandmayr,C., Laube,S.K., Eisen,D., Truss,M., Steinbacher,J., Hackner,B., Kotljarova,O. *et al.* (2014) Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat. Chem. Biol.*, **10**, 574–581.
51. Pais,J.E., Dai,N., Tamanaha,E., Vaisvila,R., Fomenkov,A.I., Bitinaite,J., Sun,Z., Guan,S., Correa,I.R. Jr, Noren,C.J. *et al.* (2015) Biochemical characterization of a Naegleria TET-like oxygenase and its application in single molecule sequencing of 5-methylcytosine. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 4316–4321.
52. Shi,Y., Ou,L., Han,S., Li,M., Pena,M.M., Pena,E.A., Liu,C., Nagarkatti,M., Fan,D. and Ai,W. (2014) Deficiency of Kruppel-like factor KLF4 in myeloid-derived suppressor cells inhibits tumor pulmonary metastasis in mice accompanied by decreased fibrocytes. *Oncogenesis*, **3**, e129.
53. Wang,Y., Chen,J., Hu,J.L., Wei,X.X., Qin,D., Gao,J., Zhang,L., Jiang,J., Li,J.S., Liu,J. *et al.* (2011) Reprogramming of mouse and human somatic cells by high-performance engineered factors. *EMBO Rep.*, **12**, 373–378.
54. Sun,Z., Terragni,J., Borgaro,J.G., Liu,Y., Yu,L., Guan,S., Wang,H., Sun,D., Cheng,X., Zhu,Z. *et al.* (2013) High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells. *Cell Rep.*, **3**, 567–576.
55. Remenyi,A., Lins,K., Nissen,L.J., Reinbold,R., Scholer,H.R. and Wilmanns,M. (2003) Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.*, **17**, 2048–2059.