Distinctive Klf4 mutants determine preference for DNA methylation status.

Xing Zhang, Emory University
Hideharu Hashimoto, Emory University
Dongxue Wang, Emory University
Alyse N. Steves, Emory University
Peng Jin, Emory University
Robert M. Blumenthal, University of Toledo
Xiaodong Cheng, Emory University

Journal Title: Nucleic Acids Research
Volume: Volume 44, Number 21
Publisher: Oxford University Press | 2016-12-01, Pages 10177-10185
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1093/nar/gkw774
Permanent URL: https://pid.emory.edu/ark:/25593/rtn1h

Final published version: http://dx.doi.org/10.1093/nar/gkw774

Copyright information:
© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic Acids Research. For commercial re-use, please contact journals.permissions@oup.com.
This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/).

Accessed April 25, 2019 2:59 AM EDT
Distinctive Klf4 mutants determine preference for DNA methylation status

Hideharu Hashimoto¹, Dongxue Wang¹, Alyse N. Steves², Peng Jin³, Robert M. Blumenthal⁴, Xing Zhang¹ and Xiaodong Cheng¹,²,*

¹Department of Biochemistry, Emory University, Atlanta, GA 30322, USA, ²Genetics and Molecular Biology Graduate Program, Laney Graduate School, Emory University, Atlanta, GA 30322, USA, ³Department of Human Genetics, Emory University, Atlanta, GA 30322, USA and ⁴Department 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 orTpG within a specific sequence. Mouse Klf4 DNA binding domain has roughly equal affinity for methylated CpG or TpG, and slightly lower affinity for unmethylated 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 (NH2) 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 genomewide 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
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-3′] (15) or ChIP-seq [5′-GGGYG(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 3α or 3β (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 (Stratagen) as described (20). Bacterial cells
were cultured at 37°C in Luria–Bertani medium, the temperature was shifted to 16°C at OD600nm = 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 GTG TGG C-3′ and 5′-TTG CCA CGC CTC-3′ (where X = C or 5mC) or 5′-GAG GTG TGG C-3′ and 5′-TTG CCA CGC CTC-3′. Curves were fit individually using PRISM 5.0 software. $K_D$ values were calculated as $[mP] = [maximal mp] \times [(C)/K_D + [C]] + [baseline mp]$, 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 (22-1D 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/Nanog 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 Cy 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 (Egrl/Zif268) (31,32,33). Interestingly, in an early phage display study of Egrl/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 co-crystal 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 GCG 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 compared with 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 E446P, but much lower affinities for E446D (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 C-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 Ado 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 recognizes three 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.

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).
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

<table>
<thead>
<tr>
<th>Klf4 (ZnF)</th>
<th>Wild-Type</th>
<th>Wild-Type</th>
<th>E446P</th>
<th>E446D</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA (M=5mC)</td>
<td>5'-GGAGGTGTGGC-3'</td>
<td>5'-GGAGGTGTGGC-3'</td>
<td>5'-GGAGGTGTGGC-3'</td>
<td>5'-GGAGGTGTGGC-3'</td>
</tr>
<tr>
<td>PDB</td>
<td>5KE6</td>
<td>5KE7</td>
<td>5KE9</td>
<td>5KEB</td>
</tr>
<tr>
<td>Beamline (SIR-CAT)</td>
<td>22-BM</td>
<td>22-BM</td>
<td>22-ID</td>
<td>22-ID</td>
</tr>
<tr>
<td>Unit cell (Å)</td>
<td>a=b=50.9, c=130.5</td>
<td>a=b=50.8, c=130.4</td>
<td>a=b=51.4, c=130.6</td>
<td>a=b=51.1, c=131.6</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>34.68-1.99 (2.04-1.99)</td>
<td>34.60-2.06 (2.11-2.06)</td>
<td>25.24-2.34 (2.42-2.34)</td>
<td>47.64-2.45 (2.52-2.45)</td>
</tr>
<tr>
<td>Rmerge</td>
<td>0.073 (0.982)</td>
<td>0.085 (0.821)</td>
<td>0.099 (0.887)</td>
<td>0.054 (0.390)</td>
</tr>
<tr>
<td>R̃oobs</td>
<td>(0.814)</td>
<td>(0.797)</td>
<td>(0.919)</td>
<td>(0.812)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.9 (100.0)</td>
<td>99.9 (100.0)</td>
<td>99.5 (100.0)</td>
<td>98.0 (73.2)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>7.0 (7.2)</td>
<td>7.0 (7.1)</td>
<td>10.7 (10.0)</td>
<td>7.1 (2.5)</td>
</tr>
<tr>
<td>Reflections (observed)</td>
<td>86,505</td>
<td>77,806</td>
<td>85,360</td>
<td>48,373</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>12,435 (891)</td>
<td>11,193 (819)</td>
<td>7,979 (763)</td>
<td>6,808 (349)</td>
</tr>
<tr>
<td>Refinement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>1.99</td>
<td>2.06</td>
<td>2.34</td>
<td>2.45</td>
</tr>
<tr>
<td>No. Reflections</td>
<td>12,409</td>
<td>11,161</td>
<td>7,915</td>
<td>6,710</td>
</tr>
<tr>
<td>Rmerge</td>
<td>0.202 / 0.232</td>
<td>0.194 / 0.216</td>
<td>0.214 / 0.272</td>
<td>0.248 / 0.286</td>
</tr>
<tr>
<td>No. Atoms (without H)</td>
<td>Protein</td>
<td>707</td>
<td>719</td>
<td>728</td>
</tr>
<tr>
<td></td>
<td>DNA</td>
<td>404</td>
<td>405</td>
<td>404</td>
</tr>
<tr>
<td></td>
<td>Zn</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Solvent</td>
<td>89</td>
<td>114</td>
<td>14</td>
</tr>
<tr>
<td>B-factors (Å²)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>45.1</td>
<td>37.3</td>
<td>73.4</td>
<td>68.9</td>
</tr>
<tr>
<td>DNA</td>
<td>36.6</td>
<td>29.7</td>
<td>54.3</td>
<td>53.6</td>
</tr>
<tr>
<td>Zn</td>
<td>38.2</td>
<td>32.2</td>
<td>68.0</td>
<td>66.1</td>
</tr>
<tr>
<td>Solvent</td>
<td>42.0</td>
<td>38.2</td>
<td>49.1</td>
<td>50.0</td>
</tr>
<tr>
<td>R.m.s. deviations</td>
<td>Bond length (Å)</td>
<td>0.006</td>
<td>0.003</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>Bond angles (°)</td>
<td>0.7</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>All atom clashscore</td>
<td>0.5</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Ramachandran (%)</td>
<td>Favored</td>
<td>98.8</td>
<td>100.0</td>
<td>98.9</td>
</tr>
<tr>
<td></td>
<td>Allowed</td>
<td>1.2</td>
<td>0.0</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>Rotamer outliers (%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cα deviation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Values in parentheses correspond to highest resolution shell; Wavelength (1.0 Å); Space group (P4₁2₁2₁ with a=b=γ=90°)
* Rmerge = Σ | I(I) - <I(I)> | / Σ I(I), where I is the observed intensity and <I(I)> is the averaged intensity from multiple observations;
* <I(q)> = averaged ratio of the intensity (I) to the error of the intensity (σI);
* R̃oobs = Σ | Fcal - Fobs | / Σ | Fcal | , where Fobs and Fcal are the observed and calculated structure factors, respectively;
* Rmerge was calculated using a randomly chosen subset (5%) of the reflections not used in refinement.
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 can be methylated, or TpG. MeCP2 recognizes consensus-binding elements, containing either CpA/CpG, which can be methylated, or TpG/CpA, which can be methylated, or TpG. MeCP2 recognizes consensus-binding elements, containing either CpA/CpG, which can be methylated, or TpG/CpA, which can be methylated, or TpG. MeCP2 recognizes consensus-binding elements, containing either CpA/CpG, which can be methylated, or TpG/CpA, which can be methylated, or TpG. MeCP2 recognizes consensus-binding elements, containing either CpA/CpG, which can be methylated, or TpG/CpA, which can be methylated, or TpG. We note that it 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-derivered 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxycytosine (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 21). 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 of Klf4-DNA have been submitted to PDB under accession number 5KE10 (E446D version of Klf4).

The X-ray structures (coordinates and structure factor files) and the E446D version of Klf4.

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 edited 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.]; 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


