BBA - Gene Regulatory Mechanisms 1861 (2018) 1093–1101

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

Specificity of MLL1 and TET3 CXXC domains towards naturally occurring cytosine modifications

Anna Stroynowska-Czerwinska[a,1], Anna Piasecka[a,1], Matthias Bochtler[a,b,*]

[a] International Institute of Molecular and Cell Biology in Warsaw (IIMCB), Trojdena 4, 02-109 Warsaw, Poland
[b] Institute of Biochemistry and Biophysics PAS (IBB), Pawinskiego 5a, 02-106 Warsaw, Poland

ARTICLE INFO

Keywords:
CXXC domain
Mixed lineage leukemia (MLL)
K-specific methyltransferase (KMT2)
Ten-eleven-translocation (TET)
CpG island (CGI)
Cytosine modifications

ABSTRACT

CXXC domains have traditionally been considered as CpG specific DNA binding domains that are repelled by cytosine modifications. This view has recently been challenged by the demonstration that CXXC domain of TET3 has relaxed sequence specificity and binds with the highest affinity to symmetric DNA duplex containing 5caCpG. Here, we present a comparative analysis of the MLL1-CXXC and TET3-CXXC sequence specificity and tolerance to cytosine modifications (5-methyl, 5-hydroxymethyl, 5-formyl, 5-carboxyl) in CpG and non-CpG context. For the first time, we take into consideration possible interference from cytosine bases elsewhere in the sequence. We show that despite similar overall structure, MLL1-CXXC has greater sequence and modification specificity than TET3-CXXC. MLL1-CXXC is specific only for CpG and does not tolerate any cytosine modifications. In contrast, TET3-CXXC does not require the CpG context of cytosine bases. Methyl-, formyl- and carboxyl-modifications are tolerated by TET3-CXXC, but only preceding G. Based on our and other data we propose a parsimonious model of MLL1-CXXC and TET3-CXXC DNA binding. This model explains why the binding of modified DNA duplexes by TET3-CXXC requires in some cases a register shift and is therefore context-dependent.

1. Introduction

Cytosine methylation in vertebrates is mostly limited to the CpG context, in which up to 80% of cytosines are modified [1,2]. CpG-rich regions (termed the CpG islands, CGIs) are frequently associated with gene promoters that are methylated to varying extent in different cell types, causing transcriptional repression [3,4]. As all other enzymatically generated cytosine modifications in eukaryotic DNA are derived from 5-methylcytosine (5mC) [5–7], the CpG context is typical not only for 5mC, but also its oxidized derivatives: 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). In contrast to 5mC, its oxidized derivatives are associated rather with activation than repression of transcription [7,8].

CXXC domains are small DNA binding domains (~50 amino acids) that are built around two structural Zn^{2+} ions, which are chelated by eight cysteine residues. Most studies of CXXC domain containing proteins have suggested that CXXC domains are CpG dinucleotide specific and are repelled by methylation of at least one cytosine within CpG [9–13]. Recently, protein binding microarray (PBM) experiments have shown that the sequence specificity of CXXC domains is not as uniform as previously thought [14]. Based on the sequence logos of specifically bound DNA, CXXC domains were grouped into four different classes (I-IV) [14], which roughly correlate with phylogenetic groups [12]. Class I CXXC domains are represented by CFP1 that binds to DNA containing CpGpG trinucleotides [14]. Class II CXXC domains (present in MLL1 and MLL2, among others) are shown to be strictly CpG-specific, requiring no other DNA bases outside the CpG sequence [14]. Class III CXXC domains are present e.g. in the TET1 and TET3 proteins. According to the sequence logo, the sequence specificity of this subgroup is low, and there is at most a slight preference for CpG over CpA, CpT and CpC [14]. Finally, the class IV proteins (represented by the CXXC domains of DNMT1, among others) bind DNA weakly or not at all [14]. We decided to focus on the representative CXXC domains from group II and group III, MLL1 and TET3.

Abbreviations: MLL, mixed lineage leukemia; TET, ten-eleven translocation; MBD, methylated CpG binding domain; CGI, CpG island; C, cytosine; 5mC, 5-methylcytosine; 5hmC, 5-hydroxymethylcytosine; 5fC, 5-formylcytosine; 5caC, 5-carboxylcytosine; H3K4me1, H3K4 monomethylation; H3K4me3, H3K4 trimethylation; PC, pocket C; PG, pocket G; PG*, pocket G*; PC*, pocket C*; MLL1-CXXC, human MLL1 CXXC domain; TET3-CXXC, human TET3 CXXC domain; PBM, protein binding microarray; BER, base excision repair; EMSA, electrophoretic mobility shift assay

* Corresponding author at: IIMCB, Trojdena 4, 02-109 Warsaw, Poland.
E-mail address: mbochtler@iimcb.gov.pl (M. Bochtler).
[1] Contributed equally.

https://doi.org/10.1016/j.bbagrm.2018.10.009
Received 29 June 2018; Received in revised form 21 September 2018; Accepted 17 October 2018
Available online 22 October 2018
1874-9399/

The MLL1-4 (also called KMT2A-KMT2D) proteins have been named for the involvement of the first identified family member (MLL1) in mixed lineage leukemia (MLL) [15]. The MLL1-4 proteins are the catalytic cores of large multiprotein COMPASS complexes, which mediate positive genetic memory [16]. MLL1 (KMT2A) and MLL2 (KMT2B) catalyze H3K4 trimethylation (H3K4me3) at promoters, MLL3 (KMT2C) and MLL4 (KMT2D) proteins mediate H3K4 monomethylation (H3K4me1) at enhancers [17]. As CpG islands are typical for promoters, only promoter-specific MLL1 and MLL2 (but not MLL3 and MLL4) possess CpG-specific CXXC domains. Based on their biochemical properties, the MLL1 and MLL2 CXXC domains are thought to implement a positive feed-forward loop between activating histone marks and the absence of repressive DNA methylation marks at promoters [10].

Ten-eleven translocation (TET) proteins have been named for chromosomal rearrangements affecting the founding member of the family, TET1, one of three TET orthologues [18–20]. The TET proteins are α-ketoglutarate dependent dioxygenases that oxidize 5mC to 5hmC, 5fC, and finally 5caC [5,7]. The higher oxidation products of TET-mediated 5mC oxidation (5fC and 5caC) bear many hallmarks of DNA damage [21]. They are excised and replaced by base excision repair (BER), and possibly other DNA repair pathways [7,22]. In stark contrast to the widely held belief that CXXC domains require non-modified cytosine bases for binding, the TET3 CXXC domain has been reported to bind also 5mCpG in some sequence contexts [23]. Moreover, the TET3 CXXC domain has also been found to bind at least as well to 5caCpG (present on both strands) as to CpG [24].

Structures of CXXC domains have been extensively characterized using crystallography and NMR approaches. In particular, structures of the CXXC domain of human MLL1 in complex with non-methylated CpG containing DNA have been determined in solution using NMR (PDB ID: 2KKF) [10], and recently, also by X-ray crystallography (PDB ID: 4NW3) [14]. Even more structural information is available for the TET3 CXXC domain. Human and *Xenopus tropicalis* domains have been crystallized with non-modified CpG DNA (PDB IDs: 4Z3C, 4HP3) [14,23], and the *Xenopus tropicalis* domain has also been crystallized with methylated CpG (5mCpG) side on both DNA strands (PDB ID: 4HP1) [23]. In addition, there is a structure of the murine Tet3 CXXC domain crystallized in complex with 5-carboxylcytosine (5caC) replacing the cytosine of the CpG also on both DNA strands (PDB ID: 5EXH) [24].

The modification specificities of MLL1 and TET3 CXXC domains are partially known already. For the MLL1 CXXC domain, specificity has been determined with regard to DNA methylation [9,10], but the effect of other DNA modifications has not yet been investigated. In contrast, the modification specificity of the TET3 CXXC domain has been the subject of several studies [14,23,24]. Despite the demonstration by Xu and colleagues in 2012 that the TET3 CXXC domain requires only cytosine, but not the CpG context [23], almost all analyzed oligonucleotides contained several C bases. Moreover, most structural and biochemical data were collected using DNA duplexes that had symmetrically placed 5mC, 5hmC, 5fC and 5caC bases. Such arrangements of modified bases simplify interpretation of binding data, but are physiologically atypical.

Here, we report a comparison of the specificities of the human MLL1 and TET3 CXXC domains (MLL1-CXXC and TET3-CXXC) bound to dsDNA containing either cytosine, 5mC, or an enzymatically oxidized 5mC derivatives (5hmC, 5fC, 5caC), in CpG or non-CpG context. With respect to CXXC domain specificity towards non-modified DNA, our work extends a recent detailed analysis [14], but accounts more carefully for possible confounding influences of cytosine bases away from the site of interest. With respect to CXXC domain preferences for modified versus non-modified DNA bases, we focus on the physiologically relevant combinations of modifications in the two DNA strands (modified bases in dyads either with non-modified C or 5mC), rather than on the most straightforwardly analyzable, symmetric cases (such as 5caCpG/5caCpG dyads). We also pay much greater attention than previous authors to the influence of solubility and fluorescence anisotropy enhancing tags, which – when consistently used – do not alter the outcome of direct comparisons, but can significantly affect quantitative conclusions, especially regarding dissociation constants.

Based on our observations and the available crystal structures, we propose a model that summarizes observed MLL1-CXXC and TET3-CXXC domain sequence preferences and modification specificities. The model sheds light on CXXC molecular recognition of the DNA bases, explains why MLL1-CXXC recognizes only CpG in context independent way and justifies the register shifts observed in some TET3-CXXC/dsDNA co-crystal structures.

2. Results

2.1. CXXC domain constructs used in this work

CXXC domains were expressed in bacterial system as fusion proteins with various solubility enhancing tags (HisSUMO, SUMO, GST), and optionally cleaved from these tags. The affinity of these proteins to tested DNA oligonucleotides was then assessed using both electrophoretic mobility shift assays (EMSA) and fluorescence anisotropy experiments. During the course of this study, it became clear that the tags contributed clearly to keeping the CXXC domains soluble and functional (for DNA binding). By themselves, the tags exhibited at most marginal DNA binding (Fig. S1A). However, when the tags were fused to the CXXC domains, they had a significant and consistent effect not only on maximal fluorescence, but also on dissociation constants, both according to fluorescence anisotropy experiments (Fig. S1B) and the electrophoretic mobility shift assay (EMSA) (Fig. S1C). As perhaps expected, based on the positive charge of the His-tag, the HisSUMO tag enhanced affinity to DNA most strongly. Despite the more than twice smaller size compared to GST (≈11 kDa compared to ≈26 kDa), it also led to the highest observed fluorescence anisotropy (Fig. S1B). In EMSA gels, GST-MLL1-CXXC exhibited a preference for modified CCGT that was not seen with the non-tagged protein (Fig. S1D). Apart from GST-MLL1, comparisons between different CXXC domains for the same DNA, or one CXXC domain for different DNAs, yielded consistent results for different tags, so that qualitative conclusions were not affected (Fig. S2). In the following, we therefore show only representative results for the protein fused with HisSUMO-tag.

2.2. MLL1-CXXC is more CpG specific than TET3-CXXC

At the outset of our studies, we compared the sequence specificity of MLL1-CXXC and TET3-CXXC using 30-mer dsDNA that contained modified or non-modified cytosines in a central CpG dyad or other sequence context and several other cytosine bases elsewhere in the duplex (Table S1, 30mers). In the experiments with MLL1-CXXC, we observed a specific shift exclusively in the presence of non-modified CpG. In other lanes, only an unspecific smear was seen (observed above the specific band) (Fig. 1A). In contrast, multiple retarded bands were seen in the otherwise same experiment using the TET3-CXXC (note especially for non-CpG duplexes). Judging from band intensity and the degree of retardation, TET3-CXXC bound best to non-modified CpG duplexes (Fig. 1B, first lane). Otherwise, similar patterns of retardation were observed for all cytosine modifications (Fig. 1B).

To monitor retarded bands, we performed EMSA experiments with increasing concentration of MLL1-CXXC and TET3-CXXC with constant concentration of non-modified DNA duplexes (containing or lacking CpG dinucleotides) (Fig. 1C–F). We confirmed that MLL1-CXXC created only single specific band with CpG-containing duplex, even at high protein concentrations (Fig. 1C). In case of non-CpG DNA we detected only an unspecific smear (Fig. 1E). In contrast, TET3-CXXC generated several super-shifts, suggesting binding at multiple sites. Retarded bands appeared at lower protein concentration for CpG containing compared to non-CpG containing oligonucleotides, but eventually shifted bands were seen for all sequence contexts at high protein

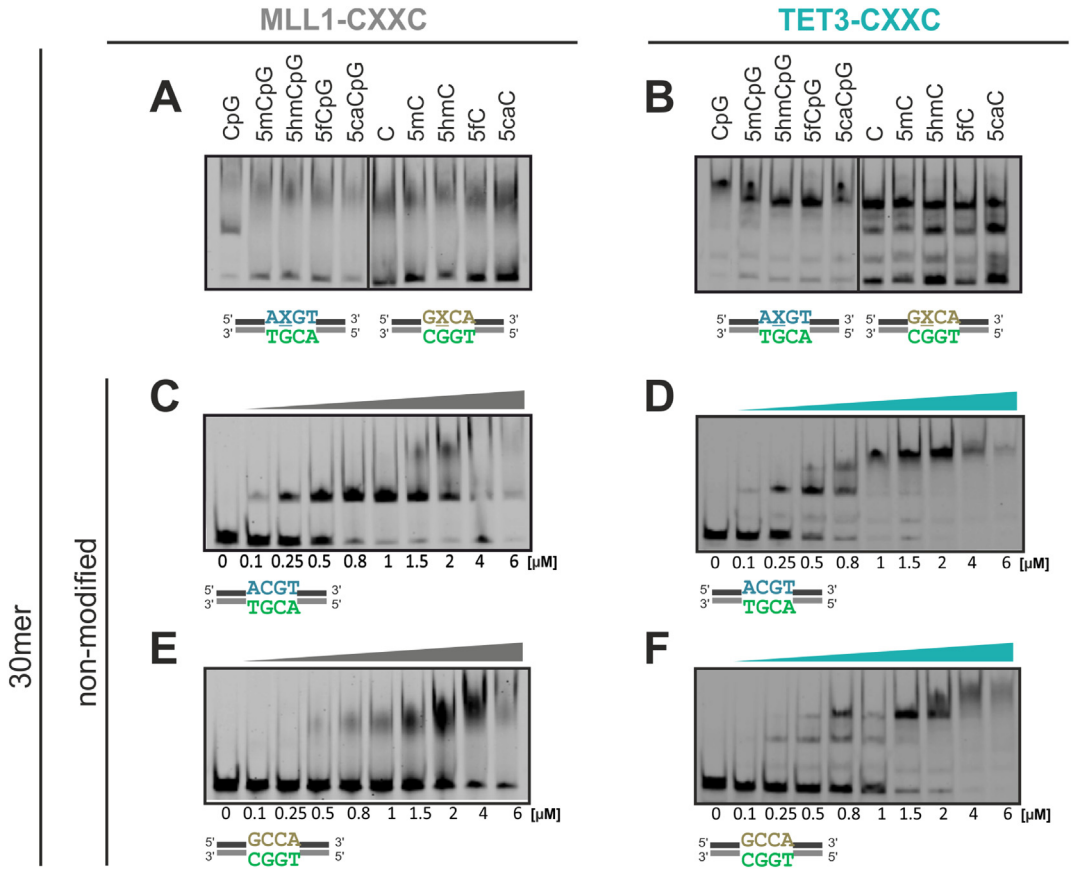

Fig. 1. Binding of MLL1-CXXC and TET3-CXXC to 30mer oligonucleotide (containing multiple C:G base pairs beside central non-modified or hemi-modified C in CpG or non-CpG context). EMSA experiments with A) MLL1-CXXC and B) TET3-CXXC with constant protein (2 μM) and dsDNA (100 nM) concentrations. C–F) EMSA experiments with increasing concentration of MLL1-CXXC (C, E) or TET3-CXX (D, F) and 100 nM non-modified dsDNA.

concentration (Fig. 1D and F). These results are consistent with literature data and confirm binding of TET3-CXXC to any cytosine, with moderate preference for the CpG context [23].

Fluorescence anisotropy experiments confirmed the preference for binding of MLL1-CXXC and TET3-CXXC to CpG over non-CpG containing DNA duplexes (Fig. S2A, S2D). They were also consistent with the EMSA result that the TET3-CXXC preference for CpG over non-CpG was very mild. Somewhat surprisingly the pronounced preference of MLL1-CXXC for CpG over non-CpG seen in the EMSA experiments was only clearly apparent in the fluorescence anisotropy experiments when the GST-tagged version of the protein fragment was used, whereas the difference appeared milder for the constructs with the HisSUMO and SUMO tags (Fig. S2A).

2.3. Design of improved, short (12mer) oligonucleotides without cytosine bases outside the site of interest

To avoid the confounding influence of C:G base pairs outside the CpG context in the case of TET3-CXXC (Fig. 1B, D, F), we designed a set of shorter (12mer) DNA duplexes that have only A:T base pairs in the flanking sequence (Table S1, 12mers). As a control we also included one non-CpG 12mer with multiple C:G base pairs (Table S1, ACAT (GC)).

In addition to DNA duplexes containing non-modified C in the complementary strand, we aimed to test the physiologically more relevant situation with 5mC rather than C in the complementary strand (possible only for CpG containing duplexes). Other modifications than 5mC were not investigated in the complementary strand, because combinations of 5hmC, 5fC and 5caC bases in the two DNA strands are unlikely *in vivo*, due to the scarcity of these bases [25]. Moreover, combinations of 5fC and 5caC are likely to be avoided *in vivo* because they create a double strand break hazard [7,26,27].

The set of oligonucleotides contained not only different modifications but also three sequence-contexts (CCGT, ACGT and TCGA; Table S1). These were chosen to minimize the number of C:G base pairs outside the CpG context. As a result of this design, the sequences differed from the CCGG sequence that is most frequently present in the TET3-CXXC and dsDNA co-crystal structures.

2.4. MLL1-CXXC binds only non-modified CpG

As in the experiments with 30mers (Fig. 1A), the EMSA experiment of MLL1-CXXC with 12mer oligonucleotides showed high specificity for DNA duplexes containing exclusively non-modified CpG (Fig. 2A, B). Any single CpG modification abolishes interaction between DNA and MLL1-CXXC. We also observed, that neither the bases upstream nor downstream of the CpG significantly influenced binding of MLL1-CXXC for the three tested sequence contexts (Figs. S2B, S3A). Moreover, MLL1-CXXC did not bind to any of CpA-containing oligonucleotides, even in the presence of additional C:G base pairs (Fig. 2A, compare C and C(GC) lanes of ACAT duplex). Binding experiments with short oligoduplexes did not exhibit the smear that was seen in the EMSA gels with 30mers. Perhaps it reflects lesser opportunities to bind to the short arms of the duplex flanking the site of interest. Fluorescence anisotropy experiments confirmed that MLL1-CXXC bound DNA containing an unmodified CpG dyad much better than DNA containing a modified dyad (Fig. S2C).

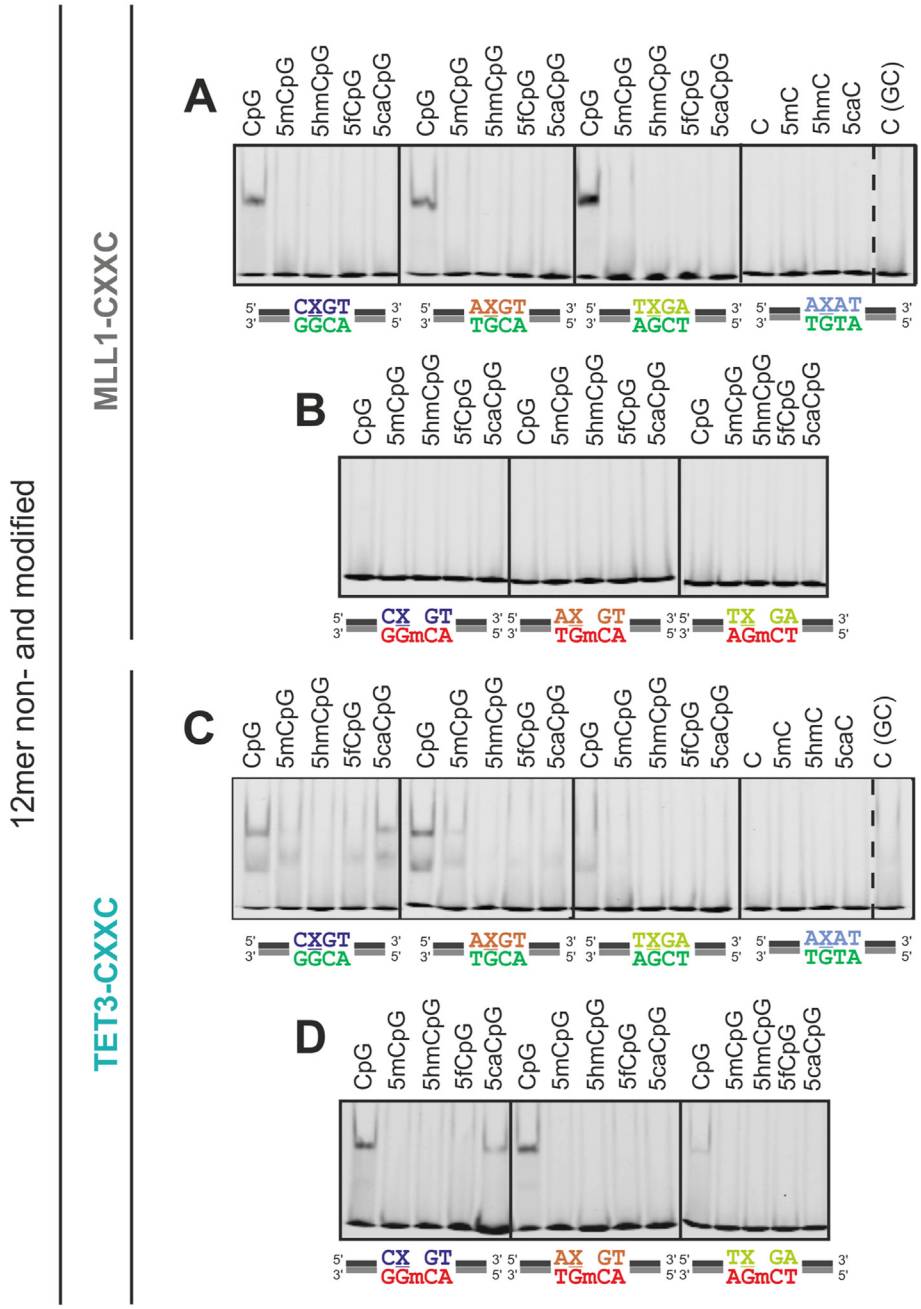

Fig. 2. Binding of MLL1-CXXC (A–B) and TET3-CXXC (C–D) to 12mer oligonucleotides lacking additional C:G base pair in the flanking sequence. EMSA of MLL1-CXXC with A) non-modified, hemi-modified or B) double-modified 12-mer DNA duplexes. Three different CpG sequence contexts (CCGT, ACGT, TCGA) and one non-CpG (ACAT) were analyzed. 12mer oligonucleotide with ACAT in the central position and additional C:G pairs was used as control with context similar to the 30mers. Similarly, EMSA of TET3-CXXC with C) non-modified and hemi-modified or D) double-modified 12-mer DNA duplexes.

2.5. TET3-CXXC has weaker specificity towards non-modified CpG than MLL1-CXXC

Like the MLL1-CXXC, the TET3-CXXC was bound to 12mer consistently when a non-modified CpG was present (Fig. 2C). According to the EMSA experiments, TET3-CXXC showed some preference for the sequence context outside the CpG dinucleotide (CCGT~ACGT > TCGA) (Fig. S3B), but this preference was not very clear from fluorescence anisotropy experiments (Fig. S2E). In the absence of additional cytosine bases, the non-CpG preference of TET3 was absent (compare ACAT and ACAT (GC) lanes) (Fig. 2C). Apparently, the TET3-CXXC requires multiple C:G base pairs to bind DNA lacking CpG. This observation stays in agreement with the binding of TET3-CXXC to 30mer containing several cytosine residues (Fig. 1B).

According to the EMSA experiments, TET3-CXXC bound duplexes that contained a modified C (only within CpG) also in a context-dependent manner (Fig. 2C). The binding occurred predominantly when the modified C of the CpG was 5mC and 5caC. Binding was tighter when the modified bases was preceded by another C (CCGT), weaker when preceded by an A (ACGT), and barely detectable for a T base (TCGA). Fluorescence anisotropy analysis confirmed binding with all modified CpG (5mC, 5hmC, 5fC, 5caC) in CCGT context (Fig. S2F). When the complementary strand contained 5mCpG instead of non-modified CpG, using EMSA we observed binding to non-modified top strand in all three sequence context as well as 5caCpG but only in CCGT context (Fig. 2D).

In the EMSA experiments with the 12mers and TET3-CXXC we

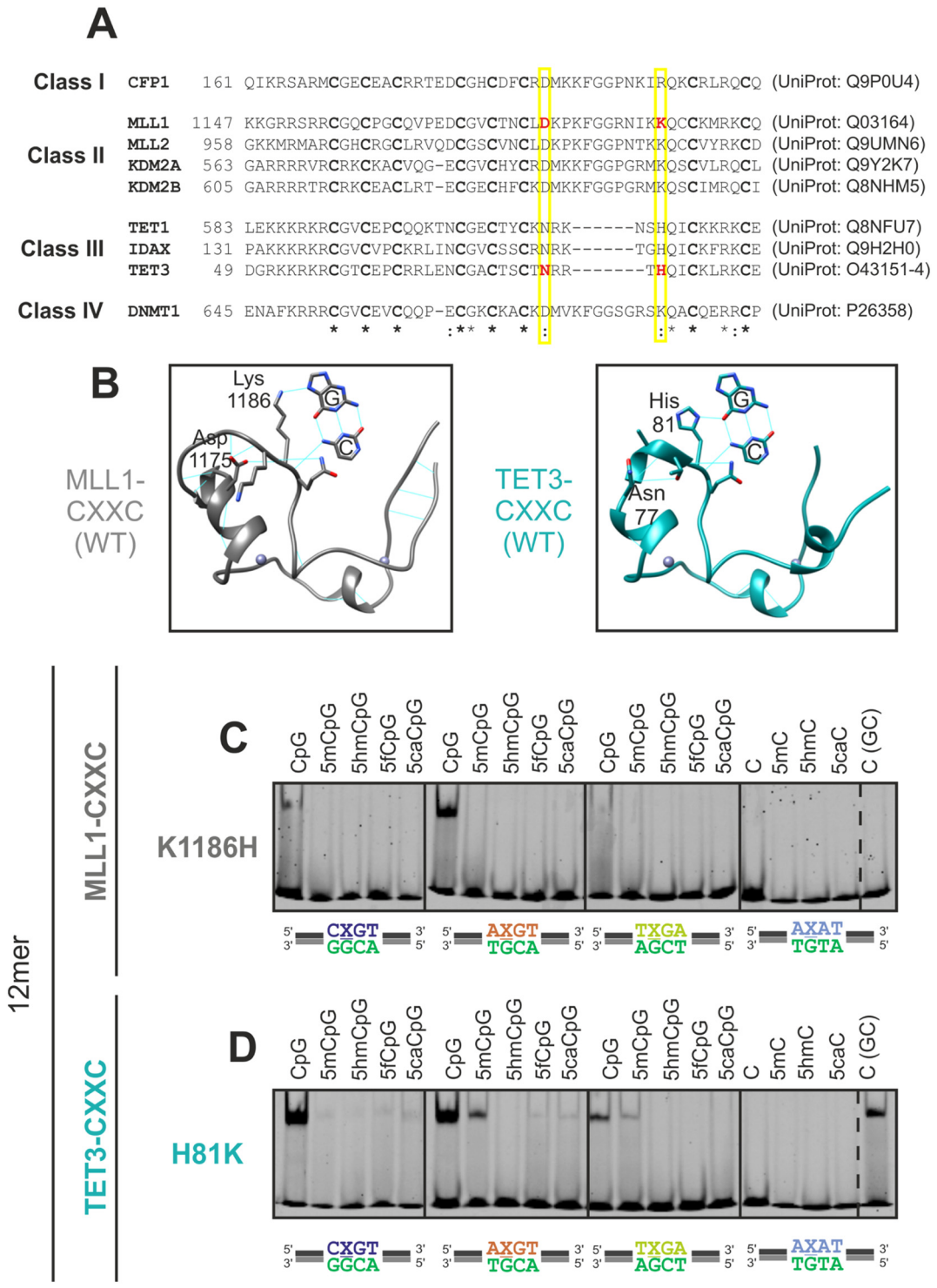

Fig. 3. Design of MLL1-like TET3-CXXC and TET3-like MLL1-CXXC chimeras by single point mutations and their binding properties. A) Protein sequence alignment of CXXC domains created using ClustalW [28,29], as in [14]. Cys residues involved in Zn^{2+} binding are bolded. Red labels identify sites of point-mutations. B) Crystal structures of wild-type: MLL1 (PDB ID: 4NW3) and TET3 (PDB ID: 4Z3C) [14]. EMSA results of 12mer DNA binding with C) MLL1-CXXC K1186H and D) TET3-CXXC H81K variants. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

observed two shifted bands (Fig. 2C). The upper band corresponded in mobility to the DNA complex of MLL1-CXXC, which has a similar mass and isoelectric point and is therefore expected to be similar. At higher protein concentration, a second, lower band gradually predominated (Fig. S3B). As this band represents a smaller or more charged species, it could not be interpreted as a conventional super-shift. The reasons for the appearance of the lower band are not clear. The occurrence of the lower band is also present with palindromic oligoduplexes (Fig. S3C) what excludes two equivalent binding modes as the cause of the shift. Moreover, we could also rule out single strand DNA binding as a possible cause (Fig. S3D). The lower and higher shifted bands predominated for TCGA (Fig. 2C) and 5caCpG/5mCpG in CCGT context (Fig. 2D), respectively. The reasons for this effect remain unclear.

2.6. Design of MLL1-like TET3-CXXC and TET3-like MLL1-CXXC chimeras

According to the sequence alignment (Fig. 3A), D1175/K1186 in MLL1-CXXC and N77/H81 in TET3-CXXC, among other amino acids, are strongly conserved within each CXXC class, but not between classes. D1175 in MLL1-CXXC (PDB ID: 4NW3) appears to be involved in rigidifying the structure in the DNA binding region (Fig. 3B) [9,10]. The "equivalent" residue N77 in TET3-CXXC (PDB ID: 4Z3C) has at most a minor structural role, but donates a hydrogen bond to the phosphodiester backbone of DNA [14]. K1186 in MLL1-CXXC and H81 in TET3-CXXC interact directly with the CpG target and could affect specificity. Prior tests of the D1175A and K1186A MLL1-CXXC variants have already shown that these residues are important for DNA binding, at least in MLL1 [9]. We wondered whether we could exchange the MLL1-CXXC and TET3-CXXC binding properties by single amino acid substitutions (MLL1-CXXC D1175, K1186 and TET3-CXXC N77, H81) (Fig. 3A, B).

2.7. Binding properties of MLL1-CXXC and TET3-CXXC chimeras

Surprisingly, we observed that both MLL1-CXXC D1175N (Fig. S4A) and TET3-CXXC N77D (Fig. S4B) variants lost affinity to DNA altogether. MLL1-CXXC K1186H variant was similar to the wild-type in its requirement for CpG context and repellence by cytosine modifications, but acquired a TET3-CXXC-like preference for flanking sequence (Fig. 3C). In contrast, the TET3-CXXC H81K variant changed specificity more drastically. The variant could still bind to non-modified and 5mCpG containing 12mer oligonucleotide but it lost almost completely the affinity towards other modifications (Fig. 3D). As the wild-type TET3-CXXC, the TET3-CXXC H81K variant still disfavored the TCGA context. In contrast to previous EMSA results for the wild-type TET3-CXXC (Fig. 2D, F), only a single shift predominated for non-modified CpG bound by TET3-CXXC H81K variant (Fig. 3D).

3. Discussion

3.1. A simple model of CXXC domains binding specificity

We analyzed available crystal structures of MLL1-CXXC (PDB ID: 4NM3) and TET3-CXXC (PDB ID: 4HP1, 4HP3, 4Z3C, 5EXH) together with the results obtained in this study. Based on this data we created a model for CXXC binding specificity that explains the available crystallographic data and is consistent with biochemical data reported in this work.

The CXXC domain model features PC:PG and PG*:PC* sites for adjacent C:G and G:C base pairs, respectively (where P stands for pocket) (Fig. 4A). The canonical binding mode places the CpG dinucleotides of the two DNA strands in the PC-PG* and PC*-PG sites (listed in 5′-3′-direction for the respective strands). For CXXC domains of relaxed sequence specificity, a register shift binding mode can occur, when the canonical binding mode is precluded by DNA modifications. MLL1-CXXC and TET3-CXXC differ in the selectivity of pockets and the propensity for register shifts. MLL1-CXXC is highly sequence selective and does not tolerate modifications. TET3-CXXC has more relaxed specificity. It may also accommodate 5mC:G pair in PC:PG, and a 5fC:G or 5ca:G pair in PG*:PC*. Placement of 5fC or 5caC in PG*, normally a guanine binding site, is facilitated by the formation of a favorable hydrogen bond or salt bridge to the 5-formyl or 5-carboxyl group, respectively.

3.2. Register shifts help to interpret crystallographic and biochemical data

The binding model presented in this work correctly predicts the register of binding of DNA in all MLL1-CXXC and TET3-CXXC crystal structures. All structures of CXXC domains in complex with DNA contain a CpG or modified CpG dinucleotide. The ones that contain a non-modified CpG (PDB IDs: 2KKF, 4NW3, 4Z3C, 4HP3) bind it in the expected register, with the C bases in PC and PC*, and the G bases in PG and PG*.

In the structures of the xtTet3-CXXC with symmetric 5mCpG (PDB ID: 4HP1) and mTet3-CXXC also with symmetric 5caCpG (PDB ID: 5EXH), the CpG dinucleotide is preceded by a C (CCGG sequence context). In both cases register shifts occur that place the Cp5mC or Cp5caC in the place normally taken by the CpG (Fig. 4A, top). In terms of the model, the PC:PG site occupied by C:G (preceding CpG) and PG*:PC* by 5mC/5caC:G base pairs. Our model assumes that 5fC and 5caC can only be accommodated when shifted. Single 5mC modification likely can fit into canonical register (5mC in PC site), but when modifications are present in both strands, a register shift is required (Fig. 4A, bottom). 5hmC bases cannot be accommodated either way, and therefore DNAs containing this base (and no other binding site) are not bound. In some cases, register shifted and non-shifted states may co-exist. Together with a PC preference (C > A > T), this may explain at least in part the context preference of TET3-CXXC (CCGT ~ ACGT > TCGA) seen in the EMSA experiments.

3.3. PC:PG and PG*:PC* properties in structural terms

Available structural data readily explain the shared properties of PC:PG and PG*:PC*, and the interactions with CpG containing DNA in the canonical binding mode. A specificity tripeptide (SPT, residues SPT1-3, 1185-KKQ-1187 in MLL1, 80-THQ-82 in TET3) contributes the sequence selective hydrogen bonding interactions.

PC:PG has similar, albeit not identical properties in MLL1-CXXC and TET3-CXXC, and is both sequence and modification specific (Fig. 4B). Sequence specificity appears to depend on hydrogen bonds in the major groove (from the cytosine N4 to the main chain carbonyl of the SPT2 residue, and from the SP2 side chain, to the N7 atom of the guanine). MLL1-CXXC PC:PG accepts well only a non-modified C:G base pair. TET3-CXXC may also accept 5mC:G pair, but not as well as a non-modified C:G pair. This difference may result from a slight shift of the cytosine base towards the sidewall, or a more rigid pocket due to stabilization by a neighboring helix [9], in MLL1-CXXC compared to TET3-CXXC. Rejection of 5hmC, 5fC and 5caC from the PC of TET3-CXXC is readily attributable to steric exclusion with the main chain ~5–6 Å away from the C5 atom of a bound cytosine. Thus, only the accommodation of a methyl group is marginally possible.

PG*:PC* differs between MLL1-CXXC and TET3-CXXC (Fig. 4C). In MLL1-CXXC, this pocket is also sequence specific and does not tolerate any other base than G in PG* site. In TET3-CXXC, the sequence specificity is relaxed and PG*:PC* can accommodate other base pairs. When a non-modified G:C pair is bound, the SPT3 glutamine donates a hydrogen bond to the guanine O6, the main chain carbonyl oxygen atom of SPT1 (K1185 in MLL1, T80 in TET3) accepts a hydrogen bond from the cytosine N4. When a register shift occurs, the SPT3 glutamine side chain leaves its canonical position and reaches out to the cytosine N4 of the complementary strand. This shift enlarges the PG* site and makes space for base modifications. In contrast, the PC* in both MLL1 and TET3 cannot accommodate any modifications. There is little space in the vicinity of the C5 atom in the crystal structures, and rejection of 5mC, 5hmC, 5fC or 5caC in PC* is readily explained by severe steric conflict (even addition of a methyl group leads to distances about 1.4 Å below the sum of van der Waals radii).

3.4. Biological implications of our data, showing high specificity of MLL1-CXXC for the non-modified CpG

At the outset of this project, we were aware of the high similarity between CpG binding regions (SPT), and of the very similar spaces for methyl groups in the MLL1-CXXC and TET3-CXXC complexes with DNA. We therefore initially expected that MLL1 may be recruited to sites of 5caC, as reported for TET3 [24]. Such recruitment would suggest a role in transcriptional activation, and appeared to be odds with

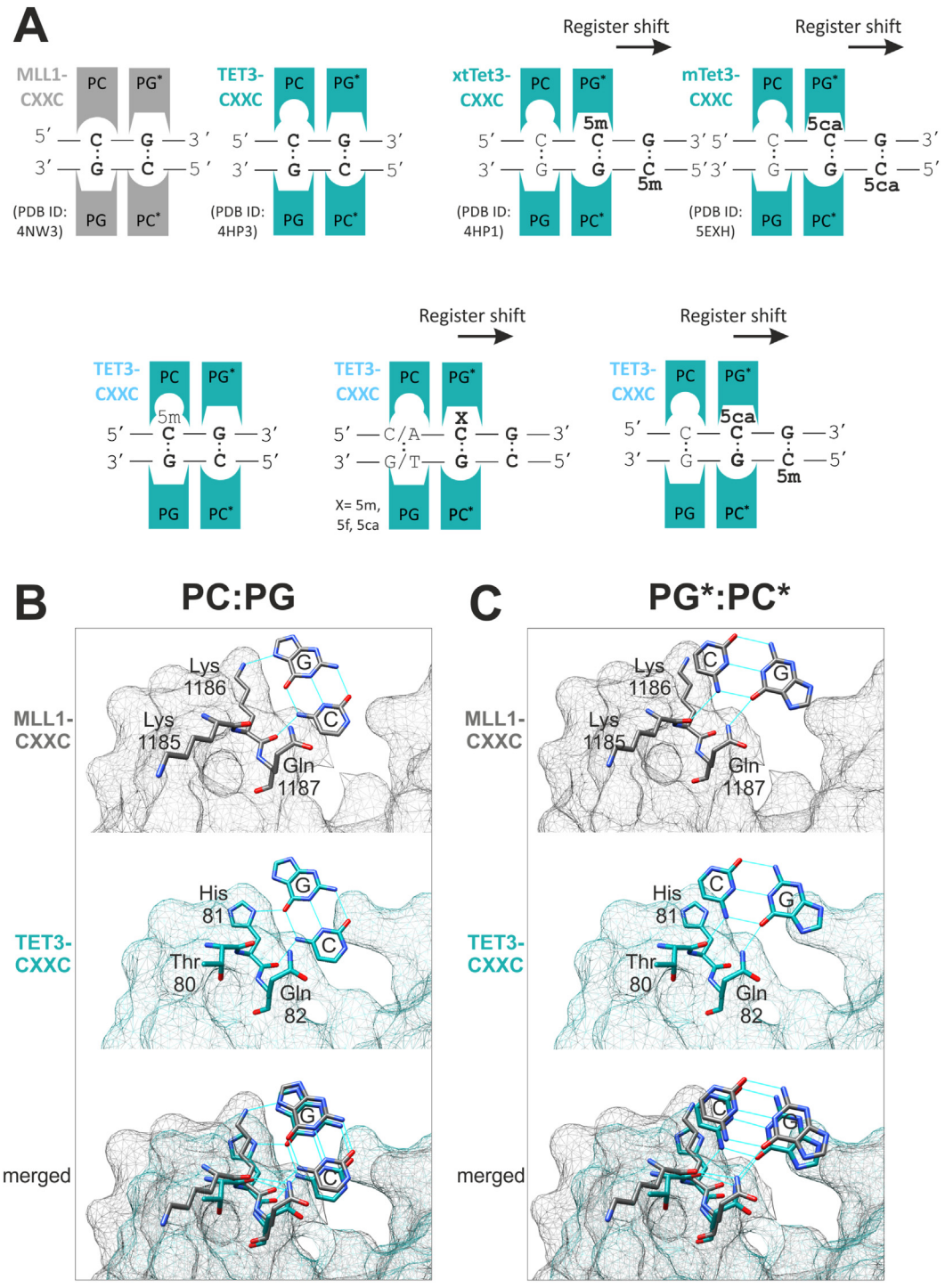

Fig. 4. MLL1-CXXC and TET3-CXXC binding with non- and modified dsDNA. A) Schematic model of MLL1-CXXC and TET3-CXXC binding to non- and modified CpG-containing dsDNA. Top – available crystal structures of MLL1 and TET3. Bottom – predicted registers. A single 5mC modification can be accommodated in PC, unlike larger modifications (5fC, 5caC) that require a shift of the binding register in 3′-direction of a top strand. Structures of B) PC:PG and C) PC*:PG* – MLL1 (PDB ID: 4NW3), TET3 (PDB ID: 4Z3C) and overlay of the two (top, middle, and bottom, respectively).

the accepted role of MLL1 in transcriptional memory, i.e. its action on chromatin that is already transcriptionally active. This work shows that MLL1-CXXC binds only to non-modified CpG. Therefore, there is no paradox, and the biochemical properties of MLL1-CXXC are consistent with the biological role of MLL1 in transcriptional memory.

Although our experiments have addressed only the properties of the MLL1-CXXC, it is very likely that the properties are general to the entire class II (as suggested also by Xu and colleagues [14]), and in particular to MLL2, with the same physiologically plausible implications. In addition to CXXC domain, we believe that for the precise recognition of the MLL1 and MLL2 biological targets, PHD domains also play an important role [30].

3.5. Biological importance of our data confirming TET3-CXXC binding to the modified CpG

Studies of the murine Tet proteins (including Tet3) have shown that the CXXC domains are not essential for Tet function [12]. Nevertheless,

it is likely that the CXXC domains are involved in protein targeting. Recruitment of TETs to non-methylated CpG regions suggests that TETs may be involved in the removal of "stray" methylation, introduced erroneously in otherwise non-methylated regions. On the other hand, recognizing the sites with 5caCpG and other modifications, suggests a model whereby the TETs demethylate a region in a concerted manner [24].

In this report, we confirmed previous studies [14,23] that TET3-CXXC binds also to DNA containing only multiple C:G base pairs. This observation may imply that the CXXC domain serves not only as a scanner for the CpG or modification but also as an anchor for C-rich sequences. Recruitment of TETs to 5caC sites, albeit limited to some sequence contexts (C preceding 5caC), has been interpreted as a mechanism to locally demethylate DNA in the vicinity of a "priming" 5caC base.

4. Materials and methods

4.1. Cloning, expression and purification of wild-type and variant CXXC domains

E. coli codon optimized synthetic genes (GeneArt, ThermoScientific) encoding the CXXC domain of human MLL1 (UniProt: Q03164, aa 1147-1203) and of human TET3 (UniProt: O43151-4, aa 47-101) were cloned into a pET28a-derived vector and pGEX-P6-2 to express the fragments with N-terminal His-SUMO-tag and N-terminal GST-tag, respectively. Variant constructs (MLL1-CXXC Asp1175Asn, Lys1186His as well as TET3-CXXC Asn77Asp, His81Lys) were obtained by site-directed mutagenesis on the wild type protein template of His-SUMO-tagged CXXC domains. Empty pGEX-P6-2 was used for GST overexpression. His-SUMO-CXXC and GST-CXXC proteins were overexpressed in *E. coli* BL21(DE3)RIL cells under kanamycin and chloramphenicol selection (50 μg/ml and 34 μg/ml, respectively), and in BL21(DE3) pLys cells under ampicillin selection (100 μg/ml), respectively. Cells were grown in LB medium to late logarithmic phase (OD_{600} of 0.5–0.8) at 37 °C, cooled for half an hour, and then induced (0.4 mM IPTG) for overnight protein expression at 24 °C. Cells were harvested the next day by centrifugation, and pellets from 1 l culture batches were frozen and stored for later use at -80 °C.

4.2. Protein purification

4.2.1. His-SUMO-CXXC proteins

Producer cells from 1 l of culture were resuspended in 40 ml ice-cold Lysis Buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl) and supplemented with PMSF and imidazole (final concentration 1 mM and 20 mM, respectively). After sonication, the lysate was cleared by ultracentrifugation. The supernatant was then applied on 5 ml Ni-NTA beads equilibrated with Lysis Buffer. After 1 h incubation, beads were extensively washed with High Salt Buffer (50 mM Tris-HCl pH 7.5, 1 M NaCl), and then Washing Buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 20 mM imidazole). Elution was performed with 5 ml Elution Buffer A (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 500 mM imidazole), immediately followed by 15 ml Elution Buffer B (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 300 mM imidazole). To avoid precipitation, protein was eluted to 20 ml solution containing 50 mM Tris-HCl, 300 μM $ZnSO_4$, 2 mM DTT. The eluate from the Ni-NTA column was then loaded on a Heparin column. After washing with Lysis Buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 150 μM $ZnSO_4$ and 1 mM DTT), protein was eluted in NaCl gradient (up to 1 M). The purity of eluate fractions was assessed by SDS-gel electrophoresis. Fractions were then pooled, concentrated using Amicon Ultra (Millipore, 3 kDa cut-off) and dialyzed overnight against storage buffer containing 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 150 μM $ZnSO_4$, 1 mM DTT, 50% glycerol. The next day protein concentration was measured, samples were aliquoted, frozen in liquid nitrogen, and stored at −80 °C until further use. All the purification steps except the dialysis were done on the same day.

4.2.2. GST-MLL1-CXXC, GST protein

Producer cells from 1 l of culture were re-suspended in 40 ml ice-cold Lysis Buffer (50 mM Tris-HCl pH 7.5, 400 mM NaCl, 5% glycerol, 150 μM $ZnSO_4$, 1 mM DTT) and supplemented with 0.25 mM PMSF. After sonication, the lysate was cleared by 30 min ultracentrifugation. Next, the supernatant was applied on 5 ml Glutathione Sepharose™ 4B (GE Healthcare) beads, equilibrated with Lysis buffer. After 2 h incubation, beads were extensively washed with Washing Buffer (50 mM Tris-HCl pH 8.0, 400 mM KCl, 5% glycerol, 150 μM $ZnSO_4$, 1 mM DTT). Protein was eluted by incubation of beads with 10 ml Elution Buffer (50 mM Tris-HCl pH 8.0, 150 mM KCl, 5% glycerol, 20 mM L-glutathione reduced, 150 μM $ZnSO_4$, 1 mM DTT) for overnight. The purified protein was concentrated on Amicon Ultra filters (Amicon, 10 kDa cut-off) and dialyzed overnight against storage buffer containing 1xPBS supplemented with 150 mM KCl, 150 μM $ZnSO_4$, 1 mM DTT, 10% glycerol. The next day protein concentration was measured and samples frozen until further use.

4.2.3. SUMO-CXXC, CXXC, HisSUMO and SUMO-tag proteins

His-SUMO tagged proteins were used to remove the His-tag by thrombin or HisSUMO-tag by Ulp1 cleavage, performed overnight in the cold room. Cleaved tags were removed by capture on Ni-NTA resin (2 h, 4 °C). Ni-NTA resins with bound HisSUMO-tag were further used for the incubation with Elution Buffer A (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 500 mM imidazole) to collect HisSUMO-tag. Next, the portion of the purified HisSUMO-tag was incubated with thrombin and incubated with Ni-NTA to collect purified SUMO-tag. MLL1-CXXC was alternatively also obtained by on-column, PreScission-catalyzed cleavage of the protein, instead of the standard glutathione elution step. The concentration of the CXXC domains without tag was assessed using Pierce™ BCA Protein Assay Kit (ThermoScientific).

4.3. Electrophoretic mobility shift assay (EMSA)

100 nM FAM-labelled double-stranded DNA oligonucleotides (Table S1) containing non- or modified central cytosine residue (5mC, 5hmC, 5fC, 5caC) were mixed either with MLL1-CXXC or TET3-CXXC (2 μM or increasing concentration 0.1–6 μM) in final 20 μl of reaction buffer (10 mM Tris-HCl, pH 7.5, 100 mM KCl, 1 mM dithiothreitol, 150 μM $ZnSO_4$). Reaction mixtures were incubated on ice for 30 min, after which were supplemented with 4 μl of 6× native loading solution (10 mM Tris-HCl, pH 7.6, 60% glycerol, 0.03% bromophenol blue) and immediately loaded on the 6% native PAGE. Electrophoresis was run in cold room using ice-cold 0.5× TBE as a running buffer and constant voltage (100 V) for 45 min. Resolved FAM-labelled DNA was detected using fluorescence readout by Typhoon Trio + (GE Healthcare).

4.4. Fluorescence anisotropy

The assay was performed in 96-well black, flat-bottom polystyrene NBS plates (Corning 3991) in 100 μl volume of reaction mixture each well. 100 nM fluorescently (FAM)-labelled double-stranded DNA (Table S1) were mixed with increasing concentration (0.5, 1, 2, 5, 8, 10, 15 μM) of MLL1-CXXC or TET3-CXXC proteins in reaction buffer (10 mM Tris-HCl, pH 7.5, 100 mM KCl, 1 mM dithiothreitol, 150 μM $ZnSO_4$). The mixtures were incubated at room temperature for 30 min and immediately afterwards the fluorescence anisotropy was measured by Tecan Infinite M1000 fluorescence microplate reader at 470 nm and 520 nm excitation and emission wavelength, respectively. The results were normalized to the sample without protein and binding curves were obtained using Prism software (GraphPad) with the one-site specific binding model with Hill slope.

Author contributions

All authors contributed to study design and data analysis. A.S-C. cloned plasmid, purified proteins, carried out fluorescence anisotropy as well as most of EMSA experiments, prepared the figures, and wrote parts of the manuscript. A.P. cloned plasmids, purified proteins, carried out some EMSA experiments, analyzed available crystal structures of CXXC domains, prepared structural pictures, and revised the text. M.B. initiated the project, analyzed available crystal structures of CXXC domains, wrote most of the manuscript, and supervised the study.

Conflict of interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Transparency document

The Transparency document associated with this article can be found, in online version.

Acknowledgement

We thank Dr. Asgar Abbas Kazrani for initial discussions about the study. This work was supported by a grant from the Polish National Science Centre (NCN) to M. B. (UMO-2014/14/M/NZ5/00558) and a special subsidy from the Polish Ministry of Science and Higher Education to A.P. (9124/E-529/M/2018).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagrm.2018.10.009.

References

[1] A.P. Bird, DNA methylation and the frequency of CpG in animal DNA, Nucleic Acids Res. 8 (1980) 1499–1504.
[2] A.M. Deaton, A. Bird, CpG islands and the regulation of transcription, Genes Dev. 25 (2011) 1010–1022.
[3] S. Saxonov, P. Berg, D.L. Brutlag, A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 1412–1417.
[4] M.G. Guenther, S.S. Levine, L.A. Boyer, R. Jaenisch, R.A. Young, A chromatin landmark and transcription initiation at most promoters in human cells, Cell 130 (2007) 77–88.
[5] M. Tahiliani, K.P. Koh, Y.H. Shen, W.A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L.M. Iyer, D.R. Liu, L. Aravind, A. Rao, Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1, Science 324 (2009) 930–935.
[6] S. Kriaucionis, N. Heintz, The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain, Science 324 (2009) 929–930.
[7] Y.F. He, B.Z. Li, Z. Li, P. Liu, Y. Wang, Q.Y. Tang, J.P. Ding, Y.Y. Jia, Z.C. Chen, L. Li, Y. Sun, X.X. Li, Q. Dai, C.X. Song, K.L. Zhang, C. He, G.L. Xu, Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA, Science 333 (2011) 1303–1307.
[8] S.C. Wu, Y. Zhang, Active DNA demethylation: many roads lead to Rome, Nat. Rev. Mol. Cell Biol. 11 (2010) 607–620.
[9] M.D. Allen, C.G. Grummitt, C. Hilcenko, S.Y. Min, L.M. Tonkin, C.M. Johnson, S.M. Freund, M. Bycroft, A.J. Warren, Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase, EMBO J. 25 (2006) 4503–4512.
[10] T. Cierpicki, L.E. Risner, J. Grembecka, S.M. Lukasik, R. Popovic, M. Omonkowska, D.D. Shultis, N.J. Zeleznik-Le, J.H. Bushweller, Structure of the MLL CXXC domain-DNA complex and its functional role in MLL-AF9 leukemia, Nat. Struct. Mol. Biol. 17 (2010) 62–68.
[11] C. Xu, C. Bian, R. Lam, A. Dong, J. Min, The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain, Nat. Commun. 2 (2011) 227.
[12] C. Frauer, A. Rottach, D. Meilinger, S. Bultmann, K. Fellinger, S. Hasenoder, M. Wang, W. Qin, J. Soding, F. Spada, H. Leonhardt, Different binding properties and function of CXXC zinc finger domains in Dnmt1 and Tet1, PLoS One 6 (2011) e16627.
[13] L.E. Risner, A. Kuntimaddi, A.A. Lokken, N.J. Achille, N.W. Birch, K. Schoenfelt, J.H. Bushweller, N.J. Zeleznik-Le, Functional specificity of CpG DNA-binding CXXC domains in mixed lineage leukemia, J. Biol. Chem. 288 (2013) 29901–29910.
[14] C. Xu, K. Liu, M. Lei, A. Yang, Y. Li, T.R. Hughes, J. Min, DNA sequence recognition of human CXXC domains and their structural determinants, Structure 26 (2018) 85–95.
[15] T. Clouaire, S. Webb, P. Skene, R. Illingworth, A. Kerr, R. Andrews, J.H. Lee, D. Skalnik, A. Bird, Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells, Genes Dev. 26 (2012) 1714–1728.
[16] T. Miller, N.J. Krogan, J. Dover, H. Erdjument-Bromage, P. Tempst, M. Johnston, J.F. Greenblatt, A. Shilatifard, COMPASS: a complex of proteins associated with a trithorax-related SET domain protein, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 12902–12907.
[17] D.Q. Hu, X. Gao, K.X. Cao, M.A. Morgan, G. Mas, E.R. Smith, A.G. Volk, E.T. Bartom, J.D. Crispino, L. Di Croce, A. Shilatifard, Not all H3K4 methylations are created equal: Mll2/COMPASS dependency in primordial germ cell specification, Mol. Cell 65 (2017) 460–475.
[18] R. Ono, T. Taki, T. Taketani, M. Taniwaki, H. Kobayashi, Y. Hayashi, LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23), Cancer Res. 62 (2002) 4075–4080.
[19] R.B. Lorsbach, J. Moore, S. Mathew, S.C. Raimondi, S.T. Mukatira, J.R. Downing, TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;q23), Leukemia 17 (2003) 637–641.
[20] W.A. Pastor, L. Aravind, A. Rao, TETonic shift: biological roles of TET proteins in DNA demethylation and transcription, Nat. Rev. Mol. Cell Biol. 14 (2013) 341–356.
[21] M. Bochtler, A. Kolano, G.L. Xu, DNA demethylation pathways: additional players and regulators, Bioessays 39 (2017).
[22] X.J. Wu, Y. Zhang, TET-mediated active DNA demethylation: mechanism, function and beyond, Nat. Rev. Genet. 18 (2017) 517–534.
[23] Y.F. Xu, C. Xu, A. Kato, W. Tempel, J.G. Abreu, C.B. Bian, Y.G. Hu, D. Hu, B. Zhao, T. Cerovina, J.B. Diao, F.Z. Wu, H.H. He, Q.Y. Cui, E. Clark, C. Ma, A. Barbara, G.J.C. Veenstra, G.L. Xu, U.B. Kaiser, X.S. Liu, S.P. Sugrue, X. He, J.R. Min, Y. Kato, Y.G. Shi, Tet3 CXXC domain and dioxygenase activity cooperatively regulate key genes for xenopus eye and neural development, Cell 151 (2012) 1200–1213.
[24] S.G. Jin, Z.M. Zhang, T.L. Dunwell, M.R. Harter, X. Wu, J. Johnson, Z. Li, J. Liu, P.E. Szabo, Q. Lu, G.L. Xu, J. Song, G.P. Pfeifer, Tet3 reads 5-carboxylcytosine through its CXXC domain and is a potential guardian against neurodegeneration, Cell Rep. 14 (2016) 493–505.
[25] L. Shen, H. Wu, D. Diep, S. Yamaguchi, A.C. D'Alessio, H.L. Fung, K. Zhang, Y. Zhang, Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics, Cell 153 (2013) 692–706.
[26] A. Maiti, A.C. Drohat, Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites, J. Biol. Chem. 286 (2011) 35334–35338.
[27] L. Zhang, X. Lu, J. Lu, H. Liang, Q. Dai, G.L. Xu, C. Luo, H. Jiang, C. He, Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA, Nat. Chem. Biol. 8 (2012) 328–330.
[28] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Clustal W and clustal X version 2.0, Bioinformatics 23 (2007) 2947–2948.
[29] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, R. Lopez, A new bioinformatics analysis tools framework at EMBL-EBI, Nucleic Acids Res. 38 (2010) W695–W699.
[30] Z. Wang, J. Song, T.A. Milne, G.G. Wang, H. Li, C.D. Allis, D.J. Patel, Pro isomerization in MLL1 PHD3-bromo cassette connects H3K4me readout to CyP33 and HDAC-mediated repression, Cell 141 (2010) 1183–1194.