Biochemical Society Annual Symposium No. 80

ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection

Hannah K. Long*†1, Neil P. Blackledge*†1 and Robert J. Klose*2

*Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, U.K., and †Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, U.K.

Abstract
Vertebrate DNA can be chemically modified by methylation of the 5 position of the cytosine base in the context of CpG dinucleotides. This modification creates a binding site for MBD (methyl-CpG-binding domain) proteins which target chromatin-modifying activities that are thought to contribute to transcriptional repression and maintain heterochromatic regions of the genome. In contrast with DNA methylation, which is found broadly across vertebrate genomes, non-methylated DNA is concentrated in regions known as CGIs (CpG islands). Recently, a family of proteins which encode a ZF-CxxC (zinc finger-CxxC) domain have been shown to specifically recognize non-methylated DNA and recruit chromatin-modifying activities to CGI elements. For example, CFP1 (CxxC finger protein 1), MLL (mixed lineage leukaemia protein), KDM (lysine demethylase) 2A and KDM2B regulate lysine methylation on histone tails, whereas TET (ten-eleven translocation) 1 and TET3 hydroxylate methylated cytosine bases. In the present review, we discuss the most recent advances in our understanding of how ZF-CxxC domain-containing proteins recognize non-methylated DNA and describe their role in chromatin modification at CGIs.

Background
The vast majority of cytosine methylation in vertebrates is found within the context of cytosine guanine dinucleotides (CpGs), occurring in up to 80% of CpGs in the genome [1,2]. Methylated CpGs are found broadly across the genome, covering both genic and intergenic regions and are specifically recognized by proteins that encode MBDs (methyl-CpG-binding domains) [3,4]. MBD proteins are generally found associated with co-repressor complexes and are thought to impose a repressive chromatin state through the activity of HDACs (histone deacetylases) [5]. In some instances, methylation of CpGs can also block access of transcription factors to their cognate binding sites to counteract transcription [5–7].

Despite the prevalence of CpG methylation, short (~1–2 kb) contiguous CpG-rich stretches of the genome exist which are generally refractory to DNA methylation [8,9]. These regions are known as CGIs (CpG islands) and are found in approximately 50–70% of vertebrate gene promoters suggesting they may play a role in gene regulation [2,10,11]. However, the precise mechanisms by which CGIs contribute to gene expression have remained largely enigmatic.

With the knowledge that methylated CpG dinucleotides are recognized by MBD proteins, it was proposed that non-methylated CpG dinucleotides may also act as a protein-binding site. To explore this possibility, Skalnik and colleagues conducted a phage-based ligand screen to discover protein factors that have the capacity to recognize non-methylated DNA and describe their role in chromatin modification at CGIs.
identified a non-methylated CGBP (CpG-binding protein) whose DNA-binding activity relied on a cysteine-rich ZF-CxxC (zinc finger-CxxC) domain [12]. The discovery of CGBP and the demonstration that the ZF-CxxC domain is responsible for non-methylated CpG-binding activity motivated bioinformatic analyses that led to the identification of an extended family of ZF-CxxC domain-containing proteins (Figure 1 and Table 1). To reflect its discovery as the first ZF-CxxC-domain containing protein, CGBP was later renamed CFP1 (CxxC finger protein 1).

The ZF-CxxC domain is characterized by two conserved cysteine-rich clusters which co-ordinate two Zn\(^{2+}\) ions intervened by a seemingly divergent sequence that effectively segregates the ZF-CxxC proteins into three distinct subtypes (Figure 2A). For the purposes of the present review, the three ZF-CxxC subtypes are referred to as type-1, -2 and -3. Proteins that encode type-1 ZF-CxxC domains Table 1 | ZF-CxxC domain-containing protein nomenclature

<table>
<thead>
<tr>
<th>Gene</th>
<th>CxxC nomenclature</th>
<th>Other names</th>
<th>NCBI Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kdm2a</td>
<td>Cxxc8</td>
<td>Fbx11, Jhdm1a, Ndy2</td>
<td>225876</td>
</tr>
<tr>
<td>Kdm2b</td>
<td>Cxxc2</td>
<td>Fbx10, Jhdm1b, Ndy1</td>
<td>30841</td>
</tr>
<tr>
<td>Cfp1</td>
<td>Cxxc1</td>
<td>Cgbp, Phf18</td>
<td>74322</td>
</tr>
<tr>
<td>Dnmt1</td>
<td>Cxxc9</td>
<td>Met11</td>
<td>13433</td>
</tr>
<tr>
<td>Mll1</td>
<td>Cxxc7</td>
<td>All1, Htx1, Kmt2a</td>
<td>214162</td>
</tr>
<tr>
<td>Mll2</td>
<td>Cxxc6</td>
<td>Wbp7, Kmt2b</td>
<td>75410</td>
</tr>
<tr>
<td>Mbd1</td>
<td>Cxxc3</td>
<td>Pcm1</td>
<td>17190</td>
</tr>
<tr>
<td>Tet1</td>
<td>Cxxc10</td>
<td>Lcx</td>
<td>52463</td>
</tr>
<tr>
<td>Tet3</td>
<td>Cxxc4</td>
<td>--</td>
<td>194388</td>
</tr>
<tr>
<td>Idax</td>
<td>Cxxc5</td>
<td>--</td>
<td>319478</td>
</tr>
<tr>
<td>Cxxc5</td>
<td>--</td>
<td>--</td>
<td>67393</td>
</tr>
</tbody>
</table>
include CFP1 and the histone H3 lysine 36 demethylases KDM2A and KDM2B. A recent series of studies have demonstrated that these proteins nucleate at CGIs in vivo, supporting the initial hypothesis that the ZF-CxxC domain may act as a CGI-targeting module [12–15]. However, the capacity of the ZF-CxxC domain to recognize CGIs in other family members, especially those in the type-2 and -3 subgroups, is less clear. In the present review, we examine our current understanding of ZF-CxxC domain structure followed by a more detailed discussion of the potential role that individual ZF-CxxC family members may play in CGI function.

Structure of the ZF-CxxC domain

The short (35–42 amino acids) primary sequence of the ZF-CxxC domain and its conspicuous arrangement of ion-co-ordinating cysteine residues suggested, even without...
atomic resolution information, that the ZF-CxxC domain would form a compact DNA-binding module (Figure 2A).

It was, however, by no means clear how this simple domain might provide such precise recognition of CpG dinucleotides and discriminate unmodified cytosine bases from the modified form which only differs by the presence of a single relatively inert methyl group. A recent succession of ZF-CxxC domain structures, in both unbound or DNA-associated states, has been instrumental in providing a detailed molecular and structural understanding of how this fascinating domain recognizes and interfaces with DNA [16–20]. These structures have also provided an important insight into why the type-1 and -3 ZF-CxxC domains possess unique DNA sequence-recognition properties.

Despite the sequence variation between type-1 and -3 ZF-CxxC domains, their overall domain architecture is highly similar. This is largely due to complete conservation of the two cysteine-rich clusters of CxxCxxxxxC/GxxCxxC and CxxRxxC motifs (Figure 2A). The eight cysteine residues within these clusters co-ordinate two Zn$^{2+}$ ions in a tetrahedral manner, stabilizing the ZF-CxxC domain in an extended crescent-shaped structure (Figure 2B). When bound to DNA, the ZF-CxxC domain lies perpendicular to the DNA axis and interrogates the major groove via a DNA-binding loop. Regions flanking the ZF-CxxC domain reach around to the opposite DNA face and interact with the minor groove (Figure 2B). By virtue of the fact that the ZF-CxxC domain essentially clamps around the DNA, it requires access to both the major and minor groove. This structural insight led to the realization that the ZF-CxxC domain must bind to linker regions of DNA between nucleosomes in vivo, as the physical association of DNA with histone octamers often prevents simultaneous access to the major and minor groove [21]. Therefore ZF-CxxC domain-mediated recognition of CGI DNA in vivo requires both the presence of non-methylated CpG dinucleotides and accessible internucleosomal DNA.

**Structural insights into DNA-binding specificity and capacity to discriminate between methylation states**

Despite the overall structural similarities within the ZF-CxxC domain fold, type-1 and type-3 ZF-CxxC domains exhibit divergence at the DNA-binding interface, which appears to define their DNA-binding specificity (Figure 2A). In the type-1 ZF-CxxC domains, an extended linker region located between the two cysteine-rich motifs contains a highly conserved KFGG (Lys-Phe-Gly-Gly) motif. The available structures for CFPI, MLL (mixed lineage leukaemia protein) 1, KDM2A and DNMT1 (DNA methyltransferase 1) suggest that the KFGG motif is not involved in sequence-specific DNA interactions, but may be required to provide rigidity to the ZF-CxxC domain fold (Figures 3A and 3C). This KFGG motif is followed by a hydrophilic positively charged DNA-binding loop which penetrates the DNA major groove in a wedge-like manner [17,18] (Figures 3A and 3C). The ZF-CxxC domain makes a number of base-specific and phosphodiester backbone (Figure 3C) interactions with the DNA substrate. Most significantly, the conserved KQ (Lys-Gln) motif [RQ (Arg-Gln) in the case of CFPI] from type-1 domains makes specific side-chain and backbone interactions with the double-stranded CpG dinucleotide-recognition sequence, forming hydrogen bonds with the cytosine bases from both DNA strands and a guanine from one of the two strands (Figures 3C and 3D). The remaining guanine in the double-stranded CpG is interrogated by the amino acid immediately N-terminal to the KQ or RQ motif via the carbonyl oxygen of the peptide backbone (Figures 3C and 3D). In type-1 ZF-CxxC domains, DNA binding is therefore mediated by a rigid tripeptide-recognition module (Figure 4A). Importantly, the close proximity of the DNA-binding loop to the CpG dinucleotide substrate is such that cytosine methylation would create a severe steric clash at the DNA-binding interface (Figure 4B). Tight packing of adjacent helices and the nearby Zn$^{2+}$ ion means that the DNA-binding tripeptide cannot undergo conformational change to accommodate the methyl moiety [18]. Consequently, in the presence of cytosine methylation, essential hydrogen bonds cannot form and DNA binding by the ZF-CxxC domain is prevented [17–19] (Figures 4A and 4B).

Interestingly, a recent structural study of the *Xenopus* TET (ten-eleven translocation) 3 type-3 ZF-CxxC domain revealed a more flexible mode of DNA binding that permits recognition of non-methylated cytosine bases in either a CpG or a non-CpG context. Similar to the type-1 domains described above, the type-3 ZF-CxxC domain of TET3 forms a crescent-like structure with a positively charged DNA-binding surface that wedges into the DNA major groove [20] (Figure 3B). However, the TET3 ZF-CxxC domain has a shortened linker before the DNA-binding loop that lacks the KFGG motif, whereas the DNA-binding interface contains an HQ (His-Gln) dipeptide corresponding to the KQ or RQ position of the type-1 domains (Figure 3C). Despite these differences, the TET3 ZF-CxxC domain bound a non-methylated CpG dinucleotide in an AC/GT context (Figures 3B, 3D and 4C). A second structure of the TET3 ZF-CxxC domain bound to a DNA molecule containing a non-methylated cytosine followed by a methylated CpG dinucleotide (CmCGG) revealed a unique capacity for the type-3 ZF-CxxC domain to interact with unmodified cytosine in a non-CpG context. In this sequence, the ZF-CxxC domain shifts one nucleotide along to interact with the non-methylated cytosine (Figure 4D). This shift leads to a steric clash between the methyl group and the Gln$^{91}$ side chain from the HQ motif, causing the Gln$^{91}$ and Ser$^{89}$ residues to become partially disordered and lose hydrogen-bonding with the DNA [20] (Figure 4D). Importantly, owing to the shortened linker region preceding the DNA-binding loop and loss of stabilizing hydrogen bonds (for example between Asp$^{189}$ and the DNA-binding loop in CFPI), the
Figure 3 | Structural insight into the DNA-binding properties of ZF-CxxC proteins

(A and B) Crystal structures of the (A) Homo sapiens DNMT1 (PDB code 3PTA) and (B) Xenopus tropicalis TET3 (PDB code 4HP3) ZF-CxxC domains in complex with DNA, viewed down the double-helix axis (left) and rotated 60° to the right (right). For both structures, the eight cysteine residues are highlighted in yellow and their interaction with the Zn^{2+} ions (represented as orange spheres) are shown by dashes. The KFGG motif is highlighted in pink and the DNA-binding tripeptide is highlighted in green, blue and red for serine, lysine/histidine and glutamine respectively. The right-hand panels highlight that the DNA-binding tripeptide loop interrogates the CpG dinucleotide via the major groove of DNA, and that the N- and C-terminal parts of the ZF-CxxC domain interact with the minor groove. Zn^{2+} ions are represented as spheres.

(C) A manually curated multiple sequence alignment of the amino acid sequence of four ZF-CxxC structures (PDB codes: CFP1, 3QMG; DNMT1, 3PTA; MLL, 2KLF; TET3, 4HP3). Residues reported to interact with the DNA backbone are marked with a grey box. Other residues are highlighted as in (A) and (B).

(D) Schematic representations of the base-specific hydrogen bond interactions between the DNA-binding tripeptide and a CpG dinucleotide. Side-chain interactions are shown by a continuous arrow and carbonyl oxygen interactions are shown by a broken arrow.

DNA-binding interface of the TET3 ZF-CxxC domain is not as rigid as those found in type-1 ZF-CxxC domains. The increased flexibility that this confers allows TET3 to seemingly recognize non-methylated cytosine bases in a broader range of sequence contexts, albeit with a slight preference for CpG [20].
Figure 4 | The effect of CpG methylation on DNA binding of type-1 and -3 ZF-CxxC domains
(A) The CFP1 DNA-binding tripeptide IRQ (Ile-Arg-Gln) forms both side-chain and backbone hydrogen bonds with the CpG dinucleotide (CpG from one DNA strand and C’pG’ from the other). Base-pairing hydrogen bonds are shown by black broken lines and ZF-CxxC tripeptide–DNA hydrogen bonds are shown as red broken lines. (B) As in (A), with methyl groups (me) at the 5 position of the cytosine rings. Cytosine methylation causes steric clash with the CFP1 tripeptide. (C) The TET1 DNA-binding tripeptide SHQ (Ser-His-Gln) forms both side-chain and backbone hydrogen bonds with the CpG dinucleotide as in (A). (D) DNA methylation of a CpCG-containing substrate causes TET1 to shift binding 1 bp along the DNA to interact with the non-methylated cytosine.

From binding non-methylated DNA in vitro to CpG island recognition and chromatin modification in vivo

In vitro binding analyses and structural studies have provided a molecular description of how the ZF-CxxC domain recognizes its DNA substrates. In most cases, these studies predict that ZF-CxxC domains should associate with non-methylated CGIs in vivo. Nevertheless, it has taken more than a decade since the discovery of CFP1 (CGBP) to convincingly demonstrate at the genome-scale that the ZF-CxxC domain can function as a CGI-targeting module [13,14]. In the following sections, we consider each of the individual ZF-CxxC domain-containing proteins and outline our current understanding of their DNA-binding properties and function in vivo.

KDM2A, KDM2B and FBXL19 (F-box and leucine-rich repeat protein 19)

KDM2A is a JmjC (Jumonji C) domain-containing histone lysine demethylase enzyme which catalyses removal of methylation from histone H3 Lys36 with a preference for the dimethyl modification state (H3K36me2) [22]. In addition to the JmjC domain, KDM2A also encodes a type-1 ZF-CxxC domain that binds specifically to DNA containing non-methylated CpGs in vitro [13]. KDM2A is significantly enriched at more than 90% of CGIs genome-wide in mouse ESCs (embryonic stem cells) [13] (Figure 5). Importantly, this includes CGI promoters of both expressed and non-expressed genes, suggesting that its nucleation on chromatin is dependent on recognition of non-methylated DNA as opposed to the transcriptional state of the associated gene.

H3K36me2, the substrate for KDM2A, is one of the most abundant histone modifications in mammalian cells, being found on 30–50% of total histone H3 and localizing to both inter- and intra-genic regions [23–25]. Importantly, KDM2A-bound CGIs are depleted of H3K36me2 and RNAi (RNA interference)-mediated knockdown of KDM2A results in increased H3K36me2 at these regions, suggesting that KDM2A plays an active role in removing H3K36me2 from CGIs [13]. Although the function of H3K36me2 remains poorly understood, in Saccharomyces cerevisiae H3K36me2 appears to be inhibitory to transcriptional initiation. This is in part thought to be mediated through binding of the EAF3 chromodomain-containing protein to H3K36me2 and recruitment of the HDAC-containing RPD3S corepressor complex [26,27]. Furthermore, it was demonstrated recently that H3K36 methylation can inhibit the interaction...
between histone chaperones and histone H3, effectively blocking histone exchange on chromatin and perhaps supressing further the capacity for non-regulatory regions to support transcriptional initiation [28]. Although it has yet to be unequivocally demonstrated that H3K36me2 leads to similar transcriptional repression in higher eukaryotes, pervasive H3K36me2 in the mammalian genome suggests that this modification may also contribute to the suppression of erroneous transcription initiation. Therefore it is tempting to speculate that targeting of KDM2A to CGIs, via its ZF-CxxC domain, leads to a specific depletion of H3K36me2 at CGIs, which could in turn help to create a favourable chromatin environment for initiation of transcription.

KDM2B, a paralogue of KDM2A, possesses an almost identical domain architecture including a type-1 ZF-CxxC domain (Figure 1). Similarly to KDM2A, KDM2B removes H3K36me2 [29] and can contribute to cellular immortalization, transformative capacity in cancer and reprogramming [29–33]. Recent ChIP-seq (chromatin immunoprecipitation sequencing)-based analysis indicates that KDM2B binds to CGIs genome-wide in a manner similar to that of KDM2A [13,15,34] (Figure 5). Intriguingly, detailed inspection of KDM2A- and KDM2B-binding profiles revealed a unique subset of CGIs that were preferentially enriched for KDM2B and depleted of KDM2A. These CGIs were generally associated with genes involved in embryo development, morphogenesis and cellular differentiation. In mouse ESCs, these type of genes are often bound by the PRCs (polycomb group repressive complexes) that function as transcriptional repressors [35], suggesting that KDM2B may contribute to polycomb-mediated transcriptional repression [15].

In mammals, the highly conserved polycomb system consists of two central PRCs called PRC1 and PRC2 [35,36]. Interestingly, PRCs appear to function almost exclusively at CGI elements, yet the mechanisms governing their recruitment to these sites remains poorly defined. The absence of an apparent sequence-specific DNA-binding domain within components of the canonical PRC1 and PRC2 complexes has led to the proposal that transient transcription factor or non-coding RNA-based interactions may provide a mechanism for targeting to CGIs [35]. Interestingly, experiments in cancer cells indicated that KDM2B associates with a variant PRC1 complex containing BCoR (Bcl-6-interacting co-repressor), PCGF1 [polycomb group RING (really interesting new gene) finger 1], RYBP [RING and YY1 (Yin and Yang 1)-binding protein], YAF2 (YY1-associated factor 2) and RING1B [37–39]. A similar complex was also purified from non-transformed mouse ESCs, suggesting that this variant PRC1 complex has a biological role in a non-malignant context [15]. On the basis of the ZF-CxxC-dependent capacity of KDM2B to recognize non-methylated CGI DNA and its enrichment at polycomb-occupied CGIs, it was hypothesized that KDM2B may contribute to recruitment of PRC1 to these sites (Figure 5). Indeed, knockdown of KDM2B using an shRNA (short hairpin RNA)-based approach caused a reduction in the levels of RING1B at polycomb target sites genome-wide, with a concomitant increase in expression of some polycomb-repressed genes [15,34,39a].

Although polycomb-repressed genes account for a relatively small subset of CGIs [36], KDM2B is present at virtually all CGIs through its ZF-CxxC-dependent recognition of non-methylated DNA. Interestingly, genome-resolution RING1B ChIP-seq analysis revealed that, in addition to the previously characterized CGIs known to be occupied by high levels of PRC1, the majority of
other CGIs in the genome also show low magnitude, yet appreciable enrichment of PRCI [15]. Binding of PRCI to these low-magnitude sites is dependent on KDM2B, suggesting that this targeting relies on recognition of non-methylated DNA. Therefore it appears that KDM2B recruits PRCI at low levels to CGIs genome-wide, possibly as a sampling mechanism for gene repression. It seems reasonable to hypothesize that, when this sampling module encounters the appropriate chromatin environment, possibly created by a lack of activating transcription factors, accumulation of PRCs can occur and transcriptional repression can be achieved.

The type-1 ZF-CxxC domain-containing protein FBXL19 is highly similar to KDM2A and KDM2B, with the exception that it lacks the N-terminal JmjC domain (Figure 1). Interestingly, both the KDM2A and KDM2B genes also have alternative transcription start sites downstream of their JmjC domain, giving rise to short forms of these proteins that closely resemble FBXL19 (Figure 1). The role of FBXL19 and the short forms of KDM2A and KDM2B remain poorly defined, but the presence of a presumably functional ZF-CxxC domain in each suggests that they probably recognize and affect CGI function.

**CFP1**

CFP1 encodes a type-1 ZF-CxxC domain and is essential for early mouse development [40]. The failure of CFP1-null ESCs to effectively differentiate in vitro is consistent with an important role for CFP1 in lineage commitment and perhaps relates to its capacity to bind CGIs and contribute to gene regulation [41]. The CFP1 protein is a component of the mammalian SETD1 (SET domain 1) H3K4 methyltransferase complex, which includes SETD1A or SETD1B, ASH2L (absent, small or homeotic 2-like), RbBP5 (retinoblastoma-binding protein 5), WDR (WD40 repeat) 5 and WDR82, and DPY-30 (dosage compensation protein 30) [42–44]. The SETD1 complex places H3K4 di- and tri-methylation (H3K4me2 and me3) [42,43]. These histone modifications are generally associated with the 5’ ends of genes [36,45–47], consistent with the localization of CFP1 [14,48]. Although the precise molecular function of H3K4me2/3 in vivo and its contribution to gene expression remain poorly defined, these marks are generally considered permissive to active transcription. This may be achieved by the recruitment of specific PHD (plant homeodomain) or tudor domain effector proteins [49–53].

Genome-wide binding studies in mouse brain tissue demonstrated that CFP1 associates with more than 80% of CGIs, and almost all CFP1-bound CGIs exhibit significant enrichment of H3K4me3 [14] (Figure 5). Similarly to KDM2A, localization of CFP1 to CGIs did not depend on the transcriptional state of the associated gene, suggesting that ZF-CxxC domain-mediated recognition of non-methylated DNA was primarily responsible for the chromatin-binding profiles of CFP1. Consistent with this observation, an exogenous CpG-rich DNA sequence lacking gene-regulatory features can recruit CFP1 and nucleate H3K4me3, apparently in the absence of transcription factors and RNAPII (RNA polymerase II) [14]. Interestingly, a subset of non-methylated CGIs associated with polycomb-mediated repression were not enriched for CFP1 [14], suggesting that, in some instances, the chromatin architecture at specific CGIs may restrict access of the ZF-CxxC domain.

In keeping with a role for CFP1 in targeting H3K4 methylation, mouse ESCs with constitutively deleted CFP1 exhibit a loss of H3K4me3 at up to half of CGIs in the mouse genome [54]. Somewhat surprisingly, however, loss of H3K4me3 was most prevalent at highly transcribed promoters and CFP1-null ESCs reconstituted with a mutant version of CFP1 lacking a functional ZF-CxxC domain restored normal H3K4me3 levels at affected genes [54]. This suggests that, in mouse ESCs, CFP1 can guide H3K4me3 to appropriate target sites in a manner that is independent of its DNA-binding activity, possibly through the activity of the CFP1 PHD domain which was shown recently to bind H3K4 methylation [55]. If the PHD domain is responsible for ZF-CxxC domain-independent targeting of CFP1 to CGIs, this would presumably require appropriate H3K4 methylation to be initiated at CGIs through alternative mechanisms. An intriguing possibility is that other H3K4 methyltransferases such as MLL1 or MLL2, which also encode ZF-CxxC domains, may fulfill this requirement.

In addition to H3K4me3 loss at CGI promoters, CFP1-null ESCs also appear to mistarget H3K4me3 methylation. The resulting ‘ectopic’ H3K4me3 peaks appear at numerous intergenic regions of the genome, and genes within the vicinity of these new H3K4me3 sites often displayed increased transcription [54]. Reintroduction of wild-type CFP1 abolished these ectopic H3K4me3 peaks, whereas a ZF-CxxC mutant did not. Therefore it appears that ZF-CxxC-independent mechanisms are capable of recruiting CFP1 to highly transcribed CGIs, whereas the ZF-CxxC domain of CFP1 is necessary for retention of the SETD1 complex at CGIs and to prevent its mis-localization to other regions of the genome.

**MLL1 and MLL2**

In addition to CFP1-containing SETD1 complexes, links between the mammalian H3K4 methylation systems and recognition of non-methylated DNA via ZF-CxxC domains extends to the MLL family. The MLL H3K4 methyltransferase family comprises four large proteins (MLL1–MLL4) that form independent multisubunit complexes that share a set of interaction partners with the SETD1 complexes, including ASH2L, WDR5, RbBP5 and DPY-30 [56]. MLL1 (also known as ALL-1, HRX, CXXC7 or KMT2A) and MLL2 (also known as MLL4, WBP7 or KMT2B) are closely related proteins that appear to have arisen through an evolutionary gene-duplication event [57,58]. They encode a type-1 ZF-CxxC domain (Figure 1), whereas MLL3 and MLL4 lack ZF-CxxC domains. The ZF-CxxC domains of MLL1 and MLL2 bind non-methylated DNA in vivo [59,60], but how they contribute to localization in vivo is not fully understood.
MLL1 plays an essential role in early mammalian development and in definitive haemopoiesis [61,62]. At a molecular level, MLL1 localizes to approximately 5000 gene promoters in human lymphoma cells, highly coincident with H3K4me3, RNAPII and active transcription [63]. MLL1 is also enriched across the HoxA cluster, a GC-rich genomic region exhibiting numerous CGIIs [63,64] (Figure 5). Therefore MLL1 localization exhibits hallmarks of ZF-CxxC-mediated recruitment, but, unlike KDM2 and CFP1 proteins [13–15], is restricted to a subset of CGI elements that are actively transcribed (Figure 5). Similarly menin, an N-terminal binding partner of MLL1 [65], associates with the 5′ end of approximately 2000 genes in a variety of cell types, frequently coinciding with MLL1-binding sites, H3K4me3 modification and high levels of gene expression [66]. The restriction of MLL1, and its binding partner menin, to a subset of CGIIs suggests that mechanisms independent of ZF-CxxC-mediated non-methylated targeting may play a role in MLL1 localization [64]. This more restricted binding pattern could be due to the activity of other chromatin-binding modules, including the N-terminal AT hooks of MLL1 which have been demonstrated to bind AT-rich regions of DNA [67] and a PHD finger (PHD3) which may recognize specific histone methylation marks [68–71]. Similarity, non-histone protein–protein interactions may also influence MLL1 localization. For example, members of PAF1C (RNA polymerase-associated factor 1 complex) interact with the CXXC domain of MLL1 [70,72]. Together, this complement of chromatin-binding activities probably shapes how MLL1 is recruited to appropriate target sites in vivo.

Chromosomal translocations that couple MLL1 to one of more than 60 known fusion partners have been implicated in driving aggressive adult and childhood leukaemias [73]. These translocation events result in the N-terminal portion of the MLL1 gene, including the AT-hooks and ZF-CxxC domain, being fused to the C-terminal portion of a translocation partner [65,74,75]. One of the most common MLL1 translocation events creates an MLL1–AF9 (ALL1–fused gene from chromosome 9 protein) fusion protein [76]. AF9 is a component of SEC (super-elongation complex) [77,78], which contributes to transcriptional elongation. The MLL1–AF9 fusion protein appears to result in aberrant targeting of SEC to normally silent MLL1 target genes, causing deleterious expression of these genes. Other MLL fusion proteins also affect target gene expression, but are thought to achieve this by recruitment of histone-modifying activities [77,79]. Interestingly, MLL–AF9 with a mutant ZF-CxxC domain exhibited severely reduced transforming potential [17,70,74], suggesting that the ZF-CxxC domain plays a crucial role in directing leukaemogenic fusion proteins to genomic targets.

MLL2 plays an essential role in early development, with MLL2 deletion causing embryonic lethality in mice at E10.5 (embryonic day 10.5) [80]. Despite having almost identical domain architecture and forming similar H3K4 methyltransferase complexes, MLL1 and MLL2 display some non-redundant functions [81,82]. For example, MLL2 is required for gametogenesis and also briefly in the zygote as a maternally derived factor [82,83]. Furthermore, MLL2 loss in macrophages causes gene-specific loss of H3K4me3 and loss of LPS (lipopolysaccharide)-triggered intracellular signalling [81]. Intriguingly, MLL2-fusion proteins have not been implicated in leukaemogenesis, which is perhaps surprising given that MLL1 and MLL2 have highly conserved ZF-CxxC domains and seemingly identical DNA-binding activities in vitro [16,60] (Figure 2A). This is exemplified by the observation that a synthetic MLL2–ENL (eleven-nineteen leukaemia) fusion protein was unable to transform haemopoietic cells, whereas a similar MLL1–ENL fusion is leukaemogenic [60]. Domain-swap experiments producing various MLL1 or MLL2 hybrid ENL fusions suggest that the ZF-CxxC domain and immediate flanking regions may be subtly different between MLL1 and MLL2, such that MLL2 fusions lack transforming potential [60].

DNMT1

DNMT1 is a large modular protein composed of a RFTS (replication foci-targeting sequence), a type-1 ZF-CxxC domain, a pair of BAH (bromo-adjacent homology) domains (BAH1 and BAH2), and a C-terminal catalytic domain (Figure 1). DNMT1 associates with PCNA (proliferating-cell nuclear antigen) at replication forks via its RFTS [84] where it copies pre-existing parental methylation patterns on to newly replicated daughter strands of DNA. During DNA replication, symmetrically methylated CpG dinucleotides become hemi-methylated as a result of semiconservative replication. Following replication, DNMT1 must recognize these sites and faithfully reinstate symmetrical methylation [85] (Figure 5). To achieve this, DNMT1 catalyses addition of a methyl group to hemi-methylated CpG dinucleotides with an efficiency 30–50-fold greater than for unmodified CpGs [84,86]. In part, its substrate specificity in vivo is dictated by a protein partner called UHRF1 (ubiquitin-like with PHD and RING finger domains 1) that recognizes hemi-methylated CpGs and is essential for correct targeting of DNMT1 [87–90].

The presence of a functional type-1 ZF-CxxC domain in DNMT1 [91] is perhaps somewhat surprising and counterintuitive given that the vast majority of CGIs are free of DNA methylation and the main substrate for DNMT1 is hemimethylated DNA. Nevertheless, a recent structural study provided a potentially interesting suggestion for how the ZF-CxxC domain of DNMT1 might function to limit DNMT1 to appropriate substrates [19]. By solving the crystal structure of a truncated form of DNMT1 in complex with DNA containing non-methylated CpGs [19], it became apparent that, when DNMT1 is bound to non-methylated CpG DNA, the ZF-CxxC domain occludes access of the DNMT1 catalytic site to the CpG dinucleotide. Furthermore, a highly acidic polypeptide loop which connects the ZF-CxxC domain to the BAH1 domain (termed the autoinhibitory linker) blocks the DNMT1 active-site cleft [19]. This led to the suggestion that,
when DNMT1 encounters an appropriate DNA substrate containing hemimethylated CpGs, the ZF-CxxC domain is unable to bind, causing the autoinhibitory loop to adopt an alternative conformation that renders the active site accessible. In support of this model, deletion of the ZF-CxxC domain and autoinhibitory linker increases the catalytic activity of DNMT1 specifically on non-methylated, but not hemimethylated, DNA substrates [19].

The ZF-CxxC-dependent autoinhibitory model was based on the study of a truncated form of DNMT1 that does not include the N-terminal RFTS domain. A subsequent analysis of full-length DNMT1 revealed that the ZF-CxxC domain did not influence its preference for hemimethylated over unmodified DNA substrates [92]. This observation was supported by structural studies using larger DNMT1 fragments that suggest that the RFTS domain can insert into the DNMT1 DNA-binding pocket and play an inhibitory role that prevails over the autoinhibitory linker implicated from the previous structural studies using smaller DNMT1 fragments [93,94]. Together, these studies suggest that DNMT1 has several in-built properties that help to limit its catalytic activity, with contributions from both ZF-CxxC domain-dependent and -independent mechanisms.

MBD1

The transcriptional repressor MBD1 encodes an MBD capable of recognizing methylated CpGs [95–97] and three ZF-CxxC domains (Figure 1). Two of the MBD1 ZF-CxxC domains (CxxC-1 and CxxC-2) are type-2 domains which lack a functional DNA-binding loop [95] (Figure 2A) and instead appear to function as protein–protein interaction modules [98,99]. The third ZF-CxxC domain (CxxC-3) is a type-1 domain capable of binding to non-methylated CpG dinucleotides in vitro [95,96]. The combination of both ZF-CxxC and MBDs in MBD1 suggests that it could potentially read non-methylated and methylated CpG dinucleotides individually or in combination [95]. However, point mutations in the CxxC-3 domain which disrupt DNA binding in vitro did not affect the recruitment of MBD1, suggesting that functional MBD1 targeting can be achieved in the absence of the ZF-CxxC domain [96]. Interestingly, in DNMT-null cells, where DNA methylation is lost, the DNA-binding capacity of the CxxC-3 domain results in targeting of MBD1 to non-methylated heterochromatin foci. It is therefore possible that the MBD1 CxxC-3 domain may act as a relevant targeting module in specific instances where DNA methylation levels are drastically reduced, for example, in pre-implantation embryos [96]. Nevertheless, in the majority of cases where the genome is pervasively methylated, the MBD appears to play a dominant role in guiding MBD1 to methylated DNA [96].

TET1 and TET3

Recently, it has become apparent that vertebrate genomes contain small yet significant levels of 5hmC (5-hydroxymethylcytosine) [100–103]. 5hmC is generated by oxidation of 5mC (5-methylcytosine) by the TET1–TET3 protein family [100,101] in an Fe(II)- and 2-oxoglutarate-(α-ketoglutarate) dependent manner. The capacity of TET proteins to convert 5mC into 5hmC prompted speculation that the TET proteins may form part of a mammalian DNA demethylation system [101,104]. In addition to the catalytic DSBH (double-stranded β-helix) domain, TET1 and TET3 encode an N-terminal ZF-CxxC domain (Figure 1). TET2 lacks a ZF-CxxC domain; however, the neighbouring IDAX (inhibition of the Dvl and axin complex protein) (CXXC4) protein has a ZF-CxxC domain which is very similar to those in TET1 and TET3 (Figure 2A), suggesting that TET2 and IDAX may have arisen from a duplication and partial inversion of either TET1 or TET3.

The three TET enzymes have distinct expression patterns and exhibit different phenotypes upon genetic perturbation, suggesting that they may have unique functions during development and in specific cell types. TET1 is highly expressed in mouse ESCs [101] where it maintains the pluripotent state by regulating the expression of pluripotency factors [105–107]. TET1 has also been implicated in the establishment of pluripotency during iPSC (induced pluripotent stem) cell reprogramming [108] and in the control of meiosis in female germ cells [109], again suggesting a role in cell fate decisions. In conflict with these reported roles for TET1 in pluripotency, other studies have failed to observe a loss of pluripotency upon knockdown of TET1, but did observe skewed differentiation [110,111]. Furthermore, TET1-null mouse ESCs remain undifferentiated and express pluripotency factors, but again display skewed differentiation [112]. These discrepancies may be explained by off-target effects of shRNAs [113], or that the phenotypes observed during acute TET1 loss are different from those seen during chronic loss of TET1 in the knockout mouse model [112]. Unlike TET1, TET3 expression is mostly restricted to the male pronucleus after fertilization [104] and ultimately TET3 lacks a ZF-CxxC domain; however, the neighbouring IDAX protein has a ZF-CxxC domain which is very similar to those in TET1 and TET3 (Figure 2A), suggesting that TET2 and IDAX may have arisen from a duplication and partial inversion of either TET1 or TET3.

The type-3 ZF-CxxC domains found in TET1 and TET3 differ from type-1 domains, as they exhibit a truncated linker region and a divergent DNA-binding loop (Figure 2A). The consequences of these differences are not fully understood, although a recent study suggests that TET3 can recognize non-methylated cytosine bases in any sequence context with a slight preference for CpG [20]. In contrast, it has been reported that the TET1 ZF-CxxC domain binds CpGs irrespective of methylation status [116,117] or that it lacks sequence-specific DNA-binding activity [118]. Despite these conflicting claims, in vivo evidence suggests that, at least in some instances, TET ZF-CxxC domains may constitute CGI-targeting modules. A number of independent studies have profiled TET1 localization genome-wide in mouse ESCs [111,117,119] and
all generally concluded that TET1 is preferentially enriched at gene promoters, with moderate enrichment in the exons of genes. Importantly, TET1 enrichment shows a strong positive correlation with CpG density, consistent with a potential ZF-CxxC-dependent CGI-targeting mechanism (Figure 5). Furthermore, mutation of the TET1 ZF-CxxC domain prevented interaction with CGI DNA in an in vitro pull-down assay [117].

On the basis of the enzymatic activity of TET proteins, there has been intense focus on determining where 5hmC is found in the genome. Genome-wide mapping studies using a variety of approaches have suggested that 5hmC is enriched at gene promoters with intermediate to high CpG density [111,117,119], bivalent promoters [111,117,119–121], within gene bodies [117] and promoters [106] of actively expressed, genes and at cis-regulatory elements [106,120,121]. Somewhat counterintuitively, CpG-rich promoters which exhibit the highest levels of TET1 appear to be largely devoid of 5hmC. This may be because the function of TET protein nucleation at CGIs is to ‘mop up’ aberrantly placed 5mC by conversion into 5hmC and perhaps subsequent reversal to the non-methylated state (Figure 5). In support of this contention, knockdown of TET1 results in acquisition of DNA methylation at specific CGIs [117]. Alternatively, it has also been reported that, for some genomic targets, TET1 has a repressive role that is independent of 5hmC enzymatic function particularly with respect to its potential ZF-CxxC-dependent CGI-targeting mechanism (Figure 5). Clearly, further study is required to fully understand the role of the ZF-CxxC domain in TET protein enzymatic function particularly with respect to its proposed role in counteracting DNA methylation.

Conclusions
In order to fully understand the contribution of CGIs to gene expression, an important future challenge is to elucidate the influence that ZF-CxxC proteins have on CGI function. Although there has been a significant amount of progress in this area over the last few years, clearly a more defined grasp of ZF-CxxC DNA-binding specificity and detailed understanding of ZF-CxxC domain-containing protein localization and function in vivo are essential in achieving this goal.

Acknowledgements
We thank Dr Nathan R. Rose for help with PyMOL and Dr Thomas A. Milne and Mr David A. Brown for a critical reading of the paper.

Funding
Work in the Klose laboratory is supported by the Wellcome Trust, the Lister Institute of Preventive Medicine, European Molecular Biology Organization (EMBO) and Cancer Research UK.

References
1 Bird, A.P. (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 8, 1499-1504


