



Published in final edited form as:

Nature. 2017 September 14; 549(7671): 287–291. doi:10.1038/nature23881.

## Polycomb-like proteins link the PRC2 complex to CpG islands

Haojie Li<sup>1,#</sup>, Robert Liefke<sup>2,3,#</sup>, Junyi Jiang<sup>1,#</sup>, Jesse Vigoda Kurland<sup>4</sup>, Wei Tian<sup>1</sup>, Pujuan Deng<sup>1</sup>, Weidi Zhang<sup>1</sup>, Qian He<sup>1</sup>, Dinshaw J. Patel<sup>6</sup>, Martha L. Bulyk<sup>4,5</sup>, Yang Shi<sup>2,3</sup>, and Zhanxin Wang<sup>1,\*</sup>

<sup>1</sup>Key Laboratory of Cell Proliferation and Regulation Biology of Ministry of Education, College of Life Sciences, Beijing Normal University, 19 Xijiekouwai Avenue, Beijing 100875, People's Republic of China

<sup>2</sup>Division of Newborn Medicine and Epigenetics Program, Department of Medicine, Boston Children's Hospital, Boston, MA 02115, USA

<sup>3</sup>Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA

<sup>4</sup>Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

<sup>5</sup>Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

<sup>6</sup>Structural Biology Program, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA

### Abstract

The Polycomb repressive complex 2 (PRC2) mainly mediates transcriptional repression<sup>1,2</sup> and plays essential roles in various biological processes including the maintenance of cell identity and proper differentiation. Polycomb-like proteins (PCLs), including PHF1, MTF2 and PHF19, are PRC2 associated factors that form sub-complexes with PRC2 core components<sup>3</sup>, and have been

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: [http://www.nature.com/authors/editorial\\_policies/license.html#terms](http://www.nature.com/authors/editorial_policies/license.html#terms) Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).

\*Contact: wangz@bnu.edu.cn.

#These authors contributed equally to this work.

Correspondence and requests for materials should be addressed to Z.W. (wangz@bnu.edu.cn).

### Online Content

Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Supplementary Information is available in the online version of the paper.

### Author contributions

H.L. performed the protein expression, purification and the crystallographic studies. R.L. performed experiments in mESCs, genome-wide analyses and MTF2 complex experiments. J.J. did the ITC and EMSA assays. J.V.K performed protein binding microarray experiments. W.T., P.D., W.Z. and Q.H. assisted in cloning and protein purification. M.L.B. supervised the protein binding microarray research. All authors analyzed the data. Z.W. initialized the project, determined the crystal structures, designed the experiments with R.L. and Y.S., and wrote the paper with the help of R.L., D.J.P., J.V.K., M.L.B., and Y.S.

The authors declare competing financial interests: Y.S. is a co-founder of Constellation Pharmaceuticals and a member of its scientific advisory board, and a consultant for Active Motif, Inc.

The remaining authors declare no competing financial interests.

Readers are welcome to comment on the online version of the paper.

proposed to modulate PRC2's enzymatic activity or its recruitment to specific genomic loci<sup>4–13</sup>. Mammalian PRC2 binding sites are enriched in CG content, which correlate with CpG islands that display a low level of DNA methylation<sup>14</sup>. However, the mechanism of PRC2 recruitment to CpG islands is not fully understood. In this study, we solved the crystal structures of the N-terminal domains of PHF1 and MTF2 with bound CpG-containing DNAs in the presence of H3K36me3-containing histone peptides. We found that the extended homologous (EH) regions of both proteins fold into a winged-helix structure, which specifically binds to the unmethylated CpG motif but in a manner completely different from the canonical winged-helix motif-DNA recognition. We further showed that the PCL EH domains are required for efficient recruitment of PRC2 to CpG island-containing promoters in mouse embryonic cells. Our research provides the first direct evidence demonstrating that PCLs are critical for PRC2 recruitment to CpG islands, thereby further clarifying their roles in transcriptional regulation *in vivo*.

PHF1, MTF2 and PHF19 (also known as PCL1, PCL2 and PCL3, respectively) are mammalian Polycomb-like proteins that directly interact with PRC2<sup>4,5</sup>. They all possess a Tudor domain, two PHD fingers, an extended homologous region clustered at the N-terminus, and a chromo-like domain located at the C-terminus (Fig. 1a, Extended Data Fig. 1a). Currently, only the structures of the isolated Tudor domains of PCLs have been solved, which bind preferentially to histone H3 trimethylated at lysine 36 (H3K36me3)<sup>4,6,7,11,15,16</sup>. We solved the crystal structure of the PHF1 Tudor-PHD1-PHD2-EH cassette at 1.9 Å resolution (Extended Data Table 1). In the apo-form structure, these four domains organize into a compact upside-down triangle plus a handle architecture, with the Tudor, PHD1 and PHD2 domains forming the triangular head and the EH forming the handle (Fig. 1b). The Tudor and both PHDs have close contacts with one another, while the EH domain contacts only PHD2.

The PHF1 EH region folds into a domain containing three  $\alpha$ -helices and a curved three-stranded  $\beta$ -sheet. Structure-based homology search by the Dali server<sup>17</sup> demonstrated that it resembles a series of winged-helix motifs as proposed<sup>18</sup>. Comparison with the typical winged-helix motif of HNF-3 $\gamma$ <sup>19</sup> showed that the major structural elements are well superimposed, while large structural variations occur mainly at the wing-like loops (W1 and W2) and the loop between helix2 and helix3 (Fig. 1c).

Given that the winged-helix motif is the defining DNA binding domain of a family of forkhead transcription factors<sup>19</sup>, we speculated that PHF1 may also target specific DNA elements through its winged-helix motif in the EH region (EH<sub>WH</sub>). Through electrophoretic mobility shift assay (EMSA), we found that PHF1 neither binds DNA containing the consensus sequence (5'-GTAAACAA-3') recognized by all FOX-family members<sup>20</sup>, nor AT-rich DNA fragments (Fig. 1d). In contrast, the PHF1 cassette binds a 12-base pair CG-rich DNA with the palindromic sequence 5'-GGGCGGCCGCCC-3' containing 2 CpG motifs (referred to as 12mer-CpG, Fig. 1d). Isothermal titration calorimetry (ITC) based measurements demonstrated that PHF1 binds the 12mer-CpG DNA with a dissociation constant ( $K_d$ ) of around 1.2  $\mu$ M and with a molar ratio of around 2:1 (Fig. 1e and Extended Data Table 2). Changing the sequence to 5'-GGGGGGCCCCCC-3' that loses both CpG motifs but retains a GpC motif, abolishes the binding for PHF1 completely (Fig. 1d),

suggesting that it is the CpG motif, but not the GpC motif, that is required for binding. Consistently, all the DNAs tested without CpG motifs fail to bind the PHF1 cassette (Extended Data Fig. 2a and Extended Data Table 3). In vertebrates, the CpG motif is a frequent target of DNA methylation, resulting in hemi- or fully methylated substrates<sup>21</sup>. The PHF1 cassette shows reduced binding for the hemi-methylated 12mer-CpG DNA and a loss of binding for the fully methylated substrate (Fig. 1f). Taken together, we conclude that PHF1 EH<sub>WH</sub> preferentially binds unmethylated CpG-containing DNA substrates.

We solved the crystal structure of the binary complex of the PHF1 cassette bound to the 12mer-CpG DNA with a 3'-overhanging thymine (Fig. 2a and Extended Data Table 1). The DNA is recognized mainly through the W1 loop located on a positively charged surface of the EH<sub>WH</sub> (Extended Data Fig. 3a). The W1 loop penetrates into the CpG-containing major groove, with the Ile322-Lys323-Lys324 tripeptide forming extensive intermolecular contacts with both cytosines and guanines of a CpG duplex, thus contributing to the CpG selectivity (Fig. 2b, c). Bases C4 and C5', the symmetrically related cytosines of a CpG duplex, are anchored in place by forming a hydrogen bond each with the main chain carbonyl oxygens of Ile322 and Lys323, respectively. Their complementary guanines, G4' and G5 are each stabilized through a hydrogen bond with the side chains of Lys324 and Lys323, respectively. Methylation of either cytosine, or replacing the cytosines of the CpG segment with other bases, would disrupt these intermolecular hydrogen bonds, or cause steric clashes with the protein backbone. In addition, G3 and G6, the bases flanking the CpG dinucleotide, form additional hydrogen bonds with the side chains of Lys324 and Lys323, respectively, which further stabilizes the recognition and may account for the preference for flanking bases (Fig. 2b, c). Besides the above base-specific recognition, Lys326 interacts with the backbone phosphate from both G7' and C6' through hydrogen bonding; Lys269 and Tyr270, located on the  $\beta$ 1 strand of the EH<sub>WH</sub>, each interacts with the backbone phosphate of G2 through main chain hydrogen bonding. Overall, the EH<sub>WH</sub> targets the CpG-containing major groove over a 6-base pair footprint, while bases from the minor groove are not targeted (Fig. 2c). Due to the insertion of the W1 loop, the major groove of the bound DNA is distorted and 2.5 Å wider than that of a canonical B-form DNA (Extended Data Fig. 3b). Lys323 and Lys324 in the W1 loop play central roles in recognizing the CpG motif, as both the K323A and the K324A mutants show a complete loss of binding (Fig. 2d). By contrast, the I322A, R325A and K326A mutations do not or only modestly affect the binding affinity (Fig. 2d). The W1 loop-mediated DNA-recognizing mechanism of PHF1 EH<sub>WH</sub> is different from other known winged-helix motifs, among which the HNF-3 $\gamma$  winged-helix motif recognizes DNA mainly through the third  $\alpha$ -helix<sup>19</sup>, while the hRFX1 winged-helix motif makes sequence-specific contacts with the target DNA through both the third  $\alpha$ -helix and the W1 loop<sup>22</sup> (Extended Data Fig. 3c, d, e).

PCL proteins show high sequence similarities within their EH regions (Extended Data Fig. 1a), indicating that other PCL members may also recognize CpG-containing DNAs. Indeed, both MTF2 and PHF19 Tudor-PHD1-PHD2-EH cassettes bind the 12mer-CpG DNA, while mutating either of the first two lysines in their IKKKK motifs (IKKRK in PHF1) results in a complete loss of binding (Fig. 2e). Sequence alignments show that the CpG-recognizing IKK(R/K)K motif in the W1 loop is conserved in vertebrate PCL EH<sub>WH</sub> domains (less so in *Drosophila*), but is absent in other winged-helix motifs (Extended Data Fig. 1b), suggesting

that the CpG-recognition mechanism by the winged-helix motif is unique to the PCL proteins.

In the crystal structure, PHF1 makes sequence-specific interaction with a four-base segment of the bound DNA. To identify detailed CpG-containing motifs recognized by the PCL proteins, we used ITC and EMSA methods to measure the binding affinities of both the PHF1 and MTF2 PHD2-EH fragments for all 10 possible combinations of the NCpGN-containing DNA duplexes (N stands for any DNA base; Extended Data Fig. 2b, c and Extended Data Tables 2 and 3). Both PHF1 and MTF2 showed higher binding affinity for the (G/T)CpGG containing sequences. To further validate the DNA motifs recognized by PCL proteins, we performed unbiased protein binding microarray experiments using universal “all 10mer” arrays<sup>23</sup>, which confirmed that PHF1 and MTF2 preferentially bind to DNAs containing the (T/G)CpGG motifs, with guanines slightly preferred as the flanking bases on each side of the motifs (Fig. 2f).

Both the PHF1 and MTF2 Tudor-PHD1-PHD2-EH cassettes favor binding to the H3K36me3 peptide over the H3K27me3 peptide (Fig. 3a, b), similar to the results from isolated Tudor domains<sup>4,6,7,11,15,16</sup>, suggesting that the presence of the other domains does not interfere with the histone binding preference. In addition, we confirmed that the Tudor domains rather than the PHD1/2 fingers are responsible for the above recognition, as mutation of an aromatic-cage residue in the Tudor domain (Y47A for PHF1, Y62A for MTF2) led to a complete loss in binding affinity (Extended Data Table 2).

To further clarify the relationship of DNA and histone binding activities, we solved the crystal structures of the ternary complexes of both PHF1 and MTF2 Tudor-PHD1-PHD2-EH cassettes with bound 12mer-CpG DNA bearing a 3' overhang thymine in the presence of the H3(33-40)K36me3 peptide (Extended Data Table 1). The structures of both complexes superimpose well with each other except that their PHD1 domains display a small overall offset (Fig. 3c). The histone and DNA binding occur independently at the Tudor domain and the EH<sub>WH</sub> domain, respectively. Of note, the Lys36me3-engaging aromatic cage of PHF1 is composed of four aromatic residues (Fig. 3d), while in MTF2, the fourth aromatic residue is replaced by Ser86 (Fig. 3e). In addition, the PHF1-histone binding is further stabilized by sequence-specific interactions between Lys37 of H3 with Glu66 from the Tudor domain, and Arg40 of H3 with the residues located in the linker region between PHD1 and PHD2 (Fig. 3d). In contrast, MTF2 contacts only the backbone of the histone peptide (Fig. 3e). These differences may account for the relatively weaker binding affinity of MTF2 for the H3K36me3 peptide (Fig. 3b).

PCL proteins have been proposed to be involved in recruiting PRC2 to chromatin<sup>4,6,10,12,24</sup>. Analysis of publically available data<sup>10,12</sup> demonstrated that MTF2 and PHF19 colocalize with PRC2 at a subset of unmethylated CpG island-containing promoters in mouse embryonic stem cells (mESCs, Fig. 4a). Their binding locations show enrichment of CpG-rich DNA motifs (Fig. 4b), supporting a potential role of EH<sub>WH</sub> for the recruitment of PRC2 to these target genes. To investigate this hypothesis in more detail, we focused on MTF2, which is the dominant PCL protein in mESCs<sup>25</sup>. MTF2 is expressed in mESCs in three distinct isoforms due to alternative translational start sites<sup>24</sup> (Extended Data Fig. 4a, b). We

obtained MTF2 knockout (KO) mESCs by disrupting the *Mtf2* gene behind the third translational start site using CRISPR/Cas9 (Extended Data Fig. 4c–e). Consistent with a positive role of MTF2 for the function of PRC2, we observed in the KO cells a reduced chromatin association of SUZ12 and de-repression of PRC2 target genes (Fig. 4c, Extended Data Fig. 4e, f and Extended Data Table 4). Rescue experiments using either wild type MTF2 (isoform 2) or a CpG-binding deficient K339A-mutated MTF2 (Fig. 2e) demonstrated that the mutant has impaired chromatin binding ability (Fig. 4d). Consistently, the wild type but not the mutant MTF2 was able to partially rescue the gene expression levels and the chromatin association of SUZ12 (Fig. 4d and Extended Data Fig. 4g, h). To obtain a more comprehensive picture, we performed ChIP-Seq experiments for MTF2, SUZ12 and H3K27me3 in control, MTF2 KO, and rescued cells (Extended Data Fig. 5a). Comparison of MTF2 ChIP-Seq data in control and KO cells confirmed that MTF2 is strongly enriched at PRC2 target genes, and only subtly bound to CpG islands at active genes (Extended Data Fig. 5b). The lost chromatin association of MTF2 and SUZ12 in MTF2 KO cells could partially been restored when wild type MTF2 but not the K339A mutant was re-expressed (Fig. 4e, f), demonstrating a critical role of the EH<sub>WH</sub> domain for the chromatin binding of MTF2 and PRC2. In contrast, H3K27me3 was only mildly affected by the level of chromatin-bound MTF2 (Fig. 4e, f), which is similar to the previously observed minor consequences on H3K27me3 levels after MTF2 or PHF19 depletion *in vivo*<sup>9,10</sup> (Extended Data Fig. 5c). To further address the role of the MTF2 EH<sub>WH</sub> with respect to the function of PRC2, we purified human MTF2 containing PRC2 from HeLa-S cells (Extended Data Fig. 6a, b). EMSA experiments demonstrated that wild type but not mutant MTF2-PRC2 can bind to the 12mer-CpG DNA (Extended Data Fig. 6c), suggesting that besides the MTF2 EH<sub>WH</sub> domain, no other parts of MTF2-PRC2 can bind to CpG motifs. Consistently, the mutant MTF2-PRC2 possesses reduced methyltransferase activity on nucleosomes *in vitro* (Extended Data Fig. 6d). Together these data support a critical function of the MTF2 EH<sub>WH</sub> domain for the recruitment of PRC2 to chromatin.

Overall, the structural and biochemical analyses of both the PHF1 and the MTF2 N-terminal cassettes establish the PCL EH<sub>WH</sub> motifs as a new family of unmethylated CpG-containing DNA binding motifs, comparable to the canonical CpG-recognizing CXXC domains identified 17 years ago<sup>26</sup>. Unexpectedly, despite the structural divergence, PHF1/MTF2 EH<sub>WH</sub> and CFP1 CXXC<sup>27</sup> use similar principles underlying CpG DNA recognition (Extended Data Fig. 3f, g, h). PRC2 and its associated PCL proteins are commonly located at CpG islands<sup>14</sup>. Our finding that PCL proteins specifically recognize unmethylated CpG-motifs through their EH<sub>WH</sub> domains provides a direct link between CpG islands and PRC2 recruitment. Given that Polycomb-related gene regulation has been implicated in carcinogenesis<sup>1</sup>, our finding may provide a novel target for therapeutic intervention.

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

X-ray statistics are listed in Extended Data Table 1. ITC binding parameters are listed in Extended Data Table 2. DNA names and sequences are listed in Extended Data Table 3. Real-time PCR Primers are presented in Extended Data Table 4.

### Protein expression and purification

Constructs containing the PHF1 or MTF2 cassettes were made by inserting the corresponding cassettes into a hexahistidine-SUMO-tagged pRSFDuet-1 vector. The protein was expressed in *E. coli* Rosetta (DE3) cells at 37 °C until the OD<sub>600</sub> reached around 1.0, then the cells were cooled at 20 °C for around an hour before 0.2 mM IPTG and 0.1 mM ZnCl<sub>2</sub> were added to induce expression overnight. Cells were harvested by centrifugation at 4,500 g for 20 minutes. Cell pellets were re-suspended with the initial buffer containing 20 mM Tris-pH 7.0, 500 mM NaCl, 20 mM imidazole and sonicated for around 5 minutes. The soluble fraction of the cells was fractionated by centrifugation of the cell lysate at 25,000 g for an hour. Histidine-SUMO-tagged target protein was isolated through a nickel-charged HiTrap Chelating FF column from GE healthcare. The histidine-SUMO tag was then cleaved by incubating with histidine-tagged ULP1 protease and dialyzed with the initial buffer at 4 °C. The dialyzed solution was then reloaded onto a nickel-charged chelating column to remove both the histidine-tagged SUMO and ULP1. The flow through was diluted two-fold with 20 mM Tris-pH 7.0, 2 mM DTT, to yield a solution with half the initial salt concentration (250 mM NaCl), which was then loaded directly onto a heparin column to remove bound DNA. Target protein was separated by increasing the salt concentration of the low salt buffer (20 mM Tris-pH 7.0, 250 mM NaCl, 2 mM DTT) from 250 mM to 1 M NaCl through a linear gradient. The target protein was further purified by a hiload 200 16/600 gel-filtration column equilibrated with the low salt buffer, through which the resulting product was eluted as a monomer with high purity. Purified proteins were concentrated to around 20 mg/ml and stored in a -80 °C freezer.

PHF19 (31-377) was not stable in buffers with salt concentration lower than 500 mM NaCl. To enhance its stability for the EMSA analysis, PHF19 (31-377) fragment was cloned into a revised pRSFDuet-1 vector bearing a hexahistidine-MBP tag at the N-terminus and a GST tag at the C-terminus. The expression and purification procedure is similar as that of PHF1 and MTF2, except that both the histidine-MBP tag and the GST tag were not removed.

### Crystallization and structure resolution

Crystallization was carried out using the hanging-drop, vapor-diffusion method by mixing equal volume of protein and well solution. Crystals of both free forms of human PHF1 (26-340) were grown by mixing 1 µl protein at the concentration of 15 mg/ml with 1 µl crystallization buffer containing 0.1 M Tris-pH 8.0, 10% PEG 3,350, 22% ethylene glycerol at 4 °C. The crystals were picked and flash frozen directly in liquid nitrogen.

The binary complex of the human PHF1 (26-360) and DNA was prepared by mixing protein with the palindromic 12mer-CpG DNA duplex bearing a 3'-overhang thymine (5'-GGGCGGCCCGCCCT-3') at the molar ratio of 2:1:1. Crystals of the complex were grown under the condition of 0.1 M Tris-pH 8.5, 25% PEG 3,350, 0.2 M Li<sub>2</sub>SO<sub>4</sub>, 10 mM MgCl<sub>2</sub> at



4 °C. Crystals were flash frozen in the crystallization buffer containing 12% 2,3- butanediol as the cryoprotectant.

The ternary complex of the mouse PHF1 (26-360)/DNA/H3(29-41)K36me3 was prepared by mixing PHF1, DNA and the peptide at the molar ratio of 2:1.1:1.5. Complex crystals were grown at 20 °C in the crystallization buffer of 50 mM Bis-Tris pH 6.5, 50 mM ammonium sulfate, 30% pentaerythritol ethoxylate (15/4 EO/OH), which was also used as the cryoprotectant.

The ternary complex of the human MTF2 (42-358)/DNA/H3(33-40)K36me3 was prepared by mixing MTF2, DNA and the histone peptide at the molar ratio of 2:1.1:1.5. Crystals of the complex were grown at 20 °C in the crystallization buffer containing 0.1 M MES monohydrate-pH 6.5, 0.2 M ammonium sulfate, 25% PEG monomethyl ether 5,000, 10% glycerol. Crystallization buffer containing 20% glycerol was used as the cryoprotectant.

Data sets for the free form human PHF1 crystals were collected at Argonne National Laboratory (Argonne, USA) APS 19ID beamline at the wavelength of 0.97918 Å. The datasets were processed using the program HKL2000. Structure determination was carried out by PHENIX<sup>30</sup> through the SAD method using zinc anomalous signals. The initial partial model was auto-built by the ARP/wARP<sup>31</sup>, then manually rebuilt by Coot<sup>32</sup>, and further refined by PHENIX. There is one PHF1 molecule in one crystallographic asymmetric unit.

Data sets for the human PHF1/DNA binary complex crystals were collected at the Shanghai Synchrotron Radiation Facility (SSRF) beamline BL18U1 in China at the wavelength of 0.97791 Å. The structure of the binary complex was solved by molecular replacement method by PHENIX using the free form PHF1 (26-340) structure as the model. The structure of the binary complex was built and refined by the PHENIX program. There are three PHF1 molecules in one asymmetric unit, with one remaining in the free form, while the other two form a complex with a DNA duplex.

Data sets for the crystals of the mouse PHF1/DNA/histone ternary complex were collected at SSRF beamline BL19U1. The structure was solved by molecular replacement method using the free form PHF1 structure as the model. Model building and structure refinement are similar as that of the PHF1 binary complex structure.

Data sets for the human MTF2/DNA/histone ternary complex crystals were collected at SSRF beamline BL19U1 at the wavelength of 0.97853 Å. The structure of the ternary complex was solved by molecular replacement method using the free form PHF1 structure as the model. Model building and refinement were similar to that of the PHF1 binary complex structure.

### Electrophoretic Mobility-Shift Assay (EMSA)

Seventy-five picomoles of double-stranded DNA were mixed with increasing amount of recombinant PCL proteins in the buffer containing 20 mM Tris-pH 7.0, 200 mM NaCl and 2 mM DTT, and incubated at 4 °C for 20 minutes. The mixture was then loaded on a 1.2% agarose gel in the TAE buffer for electrophoresis and detected by ethidium bromide staining. Constructs containing PHF1 (26-360) and MTF2 (42-378) were used for the assay. To

enhance the solubility of PHF19, a construct containing PHF19 (31-377) plus an N-terminal hexahistidine-MBP tag and a C-terminal GST tag was used for the assay. All EMSA experiments were repeated at least three times.

### **Isothermal titration calorimetric measurement**

Calorimetric experiments were carried out at 10 °C with a MicroCal iTC200 instrument. To obtain better results, purified wild-type or mutant proteins or DNA duplexes were dialyzed overnight at 4 °C in the titration buffer containing 20 mM Tris-pH 7.0, 150 mM NaCl and 2 mM  $\beta$ -mercaptoethanol. Histone peptides were prepared by dissolving small aliquots of lyophilized peptides with the same buffer just before use. Titration was performed by injecting histone peptides or DNA fragments into protein samples. Calorimetric titration data were fitted with the Origin software under the algorithm of one binding-site model. All ITC measurements have been repeated at least twice.

### **Cell culture, Cellular fractionation ChIP and antibodies**

E14 mouse ES cells (E14TG2a) were obtained from ATCC and cultured in DMEM, 15% FCS, 1 x L-Glutamine (Invitrogen), 1 x Non-essential amino acids (Invitrogen), 1 x Sodium pyruvate, 1 x Penicillin/Streptomycin (Invitrogen), 0.15%  $\beta$ -mercaptoethanol and 100 Units/ml of LIF (Millipore) on gelatin-coated plates. The cells were tested for Mycoplasma contamination. Stable cell lines were obtained via infection with lentiviral vectors harboring the appropriate construct and selected via puromycin or blasticidin. MTF2 knockouts experiments were performed using LentiCRISPRv2<sup>33</sup> with the following gRNAs targets: (1: ATCACACTCGAGTCAATATG, 2: AGGGGTGGTGCCTTAAGAA, 3: ACTGTAACGGTAGACGTTTG, 4: AGAAGAAGAAGCATTGTGTTT). The gRNA target 4 was used to obtain MTF2 KO cells. Single cell clones were gained by limited dilution and validated by sequencing and Western. Rescue experiments were performed with lentiviral vectors expressing untagged mouse MTF2 (isoforms 2). The PAM sequence was synonymously mutated in rescue constructs.

Cellular fractionations were performed using “Subcellular Protein Fractionation Kit for Cultured Cells” (Thermo Scientific, #78840) according to manufacturer’s instructions, followed by Western blotting. ChIP experiments were performed by cross-linking ChIP as described<sup>34</sup>. In short, 100 million cells were crosslinked with 1% formaldehyde for 10 minutes. Subsequently, the cells were treated first with lysis buffer 1 (50 mM Tris-pH 8.0, 2 mM EGTA, 0.1% NP-40, 10% glycerol) for 10 minutes, homogenized and centrifuged. The obtained pellet was incubated with lysis buffer 2 (50 mM Tris-pH 8.0, 2 mM EGTA, 1% SDS) for 10 minutes and sonicated with a Biorupter to gain DNA fragments of 200–500 base pairs. After centrifugation, the supernatant was diluted in dilution buffer (50 mM Tris-pH 8.0, 5 mM EGTA, 200 mM NaCl, 0.5% NP-40) and pre-cleared for 1 hour using a protein A/G bead mix. Subsequently, 10–20  $\mu$ g antibody was added and the solution was incubated for 12 hours at 4 °C. The antibodies were bound using a protein A/G bead mix for 1 hour. The beads were washed twice with NaCl buffer (20 mM Tris-pH 8.0, 500 mM NaCl, 2 mM EGTA) and twice with LiCl buffer (20 mM Tris-pH 8.0, 500 mM LiCl, 2mM EGTA, 0.1% SDS, 1% NP-40). The precipitated DNA was eluted, de-crosslinked and purified through phenol/chloroform extraction. The obtained DNA was analyzed via qPCR or next



generation sequencing. Sequencing libraries were prepared using the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB, #E7370) with 10–20 ng DNA. For RNA-seq, whole RNA was prepared using Trizol and purified using Magnetic beads mRNA Isolation Kit (BioLabs, #S1550S). After mRNA fragmentation by heating the sample for 6 minutes at 95 °C, the mRNA was reverse transcribed using SuperScript III (Invitrogen, 18080-044), followed by Second Strand Synthesis (Invitrogen, 10812-014). RNA-seq libraries were constructed of 10–50 ng DNA using NEBNext DNA Library Prep Reagent Set (NEB, E6000). RNA-seq and ChIP-Seq libraries were analyzed using the Illumina HiSeq 2500 System.

Following antibodies were used: SUZ12 (Santa Cruz, sc-46264, Western), Suz12 (D39F6, Cell Signaling, ChIP), Actin (abcam, ab3280), Histone H3 (abcam, ab1791), H3K27me3 (Millipore, 07-449), H3K4me3 (Millipore, 04-745), MTF2 (Proteintech, 16208-1-AP).

### EMSA and HMTase reaction with human MTF2 complexes

HeLa-S cells were infected with Lentiviral constructs expressing human full-length Flag-HA-MTF2 or Flag-HA-MTF2 K339A. MTF2 complexes and empty vector Mock control were obtained in parallel from 5l HeLa-S cultures via single step purification using anti-Flag (M2) conjugated agarose beads (Sigma, A2220). Bound proteins were washed three times with TAP-buffer (50 mM Tris-pH 7.9, 100 mM KCl, 5 mM MgCl<sub>2</sub>, 0.2 mM EDTA, 10% Glycerol, 0.2 mM PMSF, 1 mM DTT, 0.1% NP-40) and subsequently eluted with 50 µl TAP-buffer containing 1 µg/ml Flag peptide. 1 µl of the Eluate was analyzed by Silver staining. EMSA were performed with equal volumes (0.5, 1, 2 and 3 µl) of the eluates using the 12mer-CpG sequence. For HMTase assay, mononucleosomes were incubated with 15 µl of the eluates for 2 hours at 25 °C using the following reaction buffer: 10 mM HEPES pH 7.4, 50 mM NaCl, 10 µM ZnCl<sub>2</sub>, 0.5 mM DTT, 2.5 mM MgCl<sub>2</sub>, 2 mM ATP, 5% Glycerol, 80 µM SAM<sup>35</sup>. The reaction products were analyzed by Western blotting.

### Bioinformatics analyses

RNA-Seq data were analyzing using TopHat and Cuffdiff<sup>36</sup>. ChIP-Seq data were aligned to mouse genome mm9 using Bowtie<sup>37</sup> with n = 1 and m =3 as parameter. Normalized Bigwig files were obtained using DeepTools<sup>38</sup>. Bioinformatics analyses were performed via the Cistrome platform<sup>39</sup> or Bioconductor<sup>40</sup>. Promoter reads were counted from -2000 to +2000 relative to the transcription start site and normalized to reads per million (rpm). Following public data sets were used: SUZ12 (GSM700554, GSM700553), PHF19 (GSM700556, GSM700555)<sup>10</sup>, MTF2 (GSM415050)<sup>12</sup>, MRE-Seq (GSM881347)<sup>29</sup>, H3K4me3 (GSM2027596)<sup>34</sup>. CpG island and promoter definitions were downloaded from the UCSC browser. Enriched motifs were identified by MEME-ChIP<sup>41</sup>.

### Protein binding microarray experiments and analysis

GST-fusion proteins for human PHF1 (165-360) and MTF2 (180-369) were expressed in BL21 (DE3) cells and affinity purified using Glutathione beads (Amersham). Subsequently, custom-designed “all-10mer” universal oligonucleotide arrays in 8 x 60K GSE array format (Agilent Technologies; AMADID #030236) were double-stranded and duplicate protein binding microarray experiments were performed essentially as described<sup>23,28</sup>. MTF2 was

assayed at a final concentration of either 500 nM or 900 nM, while PHF1 was assayed at a final concentration of 900 nM, in binding reactions containing 50  $\mu$ M zinc acetate, on either a fresh slide or a slide that had been stripped exactly once. Scans were acquired using a GenePix 4400A (Molecular Devices) microarray scanner. Microarray data quantification, normalization, and motif derivation were performed essentially as described previously using the Universal PBM Analysis Suite and the Seed-and-Wobble motif-derivation algorithm<sup>23,28</sup>.

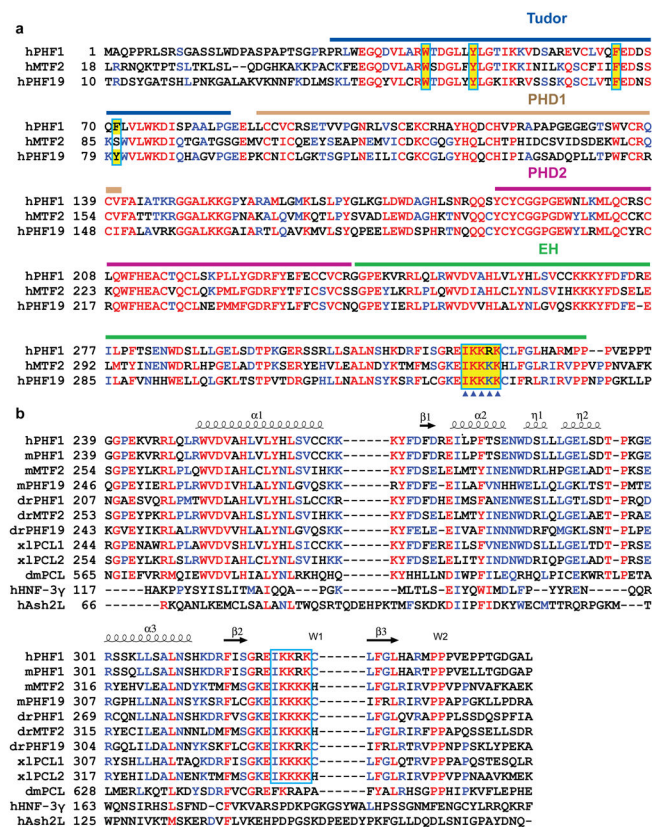
## Statistical Analysis

For statistical comparisons of two groups, one-way ANOVA followed by Tukey's post-hoc test was used.

## Data availability

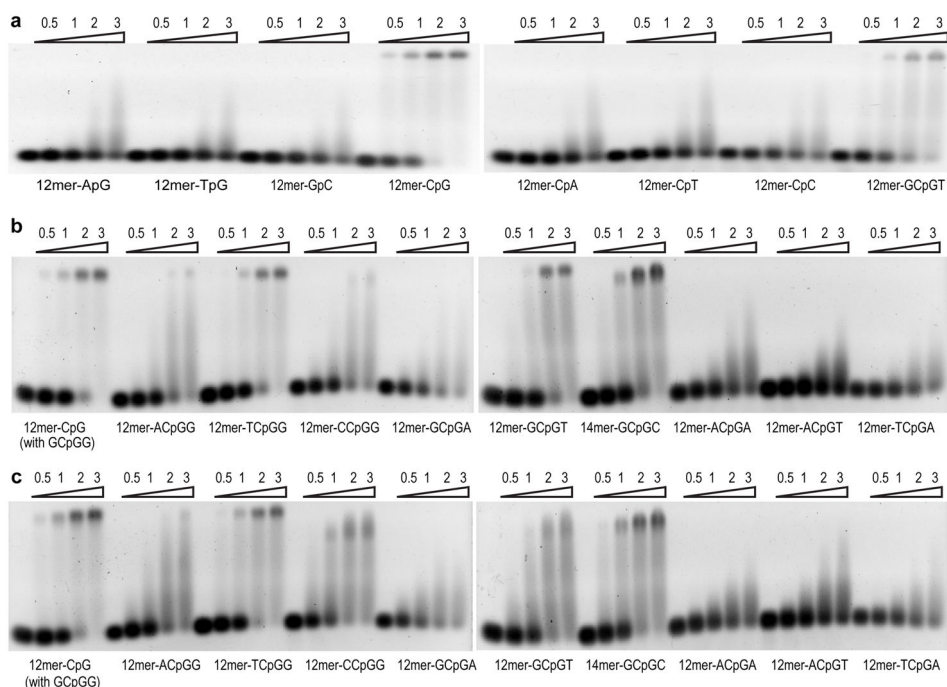
Atomic coordinates and structure factors for the apo-form PHF1 (two forms of crystals), binary complex of PHF1 with bound DNA, ternary complexes of PHF1 and MTF2 with bound DNA and histone peptide were deposited in the protein data bank with the accession codes of 5XFN, 5XFO, 5XFP, 5XFQ and 5XFR, respectively. ChIP-Seq and RNA-Seq data are available at the GEO repository: GSE97805. PBM data are available in the UniPROBE database (UniPROBE accession ID: KUR17A).

## Extended Data



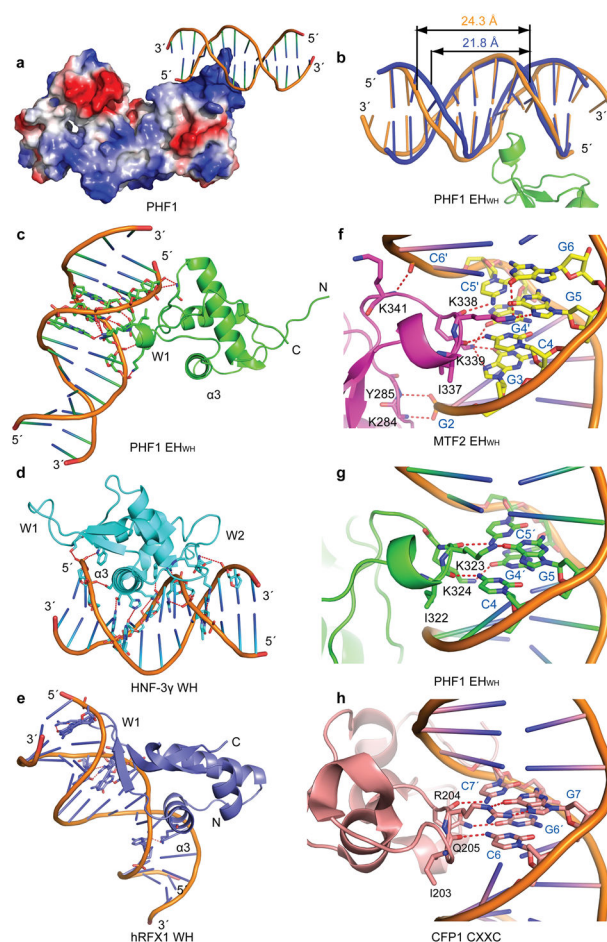
**Extended Data Figure 1. Sequence alignment of human PCL proteins, or the EH/WH regions from various species**

**a**, Sequence alignment of the N-terminal domains of human PCL proteins. Residues with high similarity are colored in red. Key residues mentioned in the text were highlighted yellow and indicated with blue triangles below. **b**, Sequence alignment of the EH domains from various species of PCL proteins and two typical winged-helix motifs. Conserved IKK(K/R)K motifs within the W1 loop of various PCLs were indicated in a blue box. Species abbreviations: h for *Homo sapiens*; m for *Mus musculus*; dr for *Danio rerio*; xl for *Xenopus laevis*; dm for *Drosophila melanogaster*.



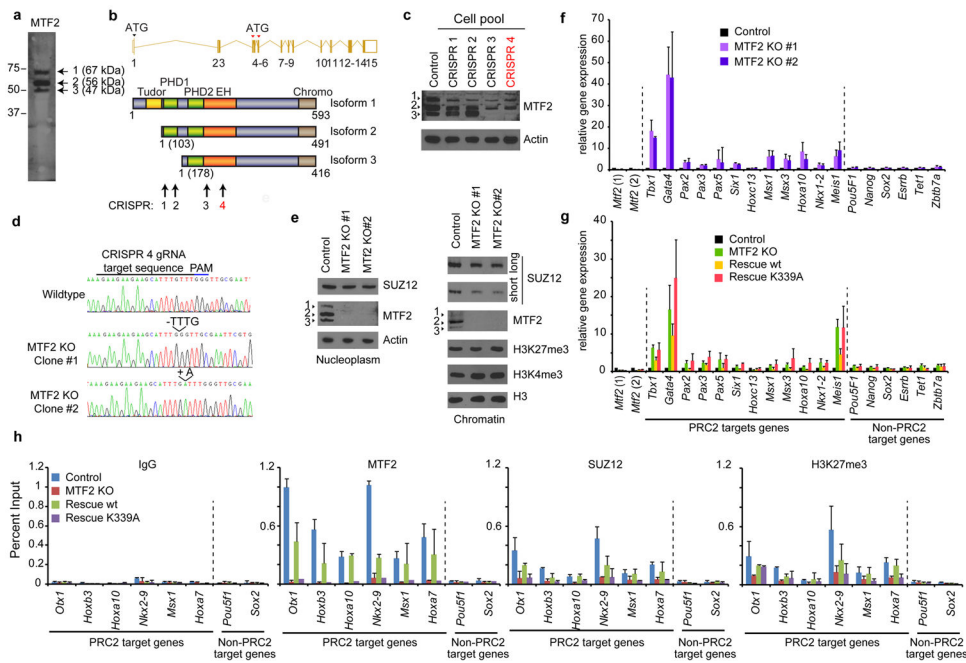
**Extended Data Figure 2. Binding analysis of PCLs with different CpG-motif substitutions or with CpG-containing DNAs varying in their flanking sequences**

**a**, EMSA results of the PHF1(26-360) fragment with different DNA duplexes bearing base substitutions in the CpG-motif. **b**, **c**, EMSA results of PHF1 (165-360) (panel b) or MTF2 (180-378) (panel c) with various NCpGN-containing DNA motifs, N stands for any DNA base. Protein to DNA molar ratio is shown above. Data shown are representative of at least three independent experiments. Uncropped Gels are shown in Supplementary Figure 1.



**Extended Data Figure 3. The comparisons of DNA-bound PHF1/MTF2 EH with two DNA bound winged-helix motifs and a CXXC domain**

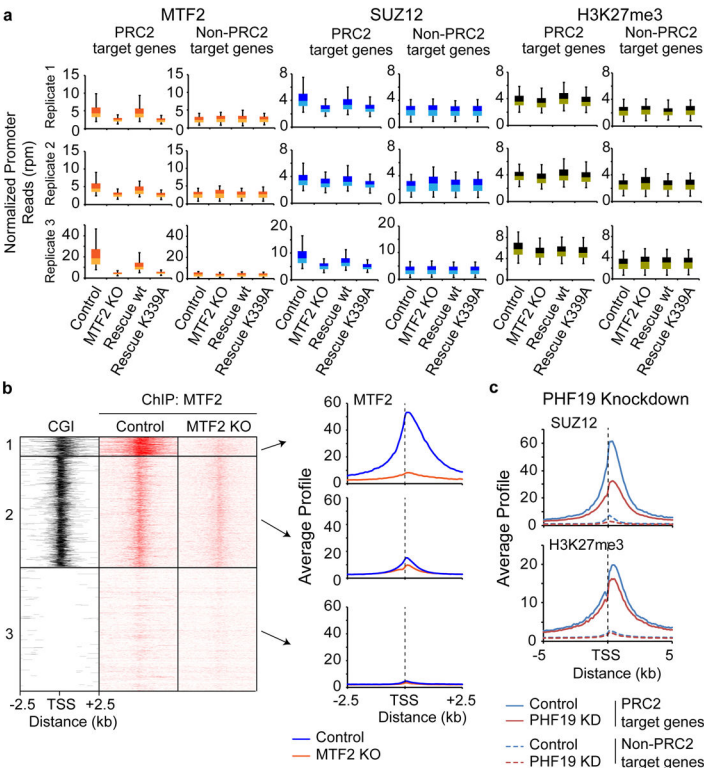
**a**, Electrostatic surface of the PHF1 cassette, with basic regions shown in blue and acidic regions in red. Bound DNA is shown in a cartoon representation. **b**, Superimposition of the PHF1 bound DNA (colored in orange) with a canonical B-form DNA (colored in blue, PDB: 1HQ7). **c, d, e**, Comparison of the DNA-recognizing details of the PHF1 EH (in **c**) with the winged-helix motifs of HNF-3 $\gamma$  (in **d**, PDB: 1VTN) and hRFX1 (in **e**, PDB: 1DP7) when all three domains were structurally aligned. **f, g, h**, Comparison of the CpG-recognition details of the MTF2 EH (in **f**) and the PHF1 EH (in **g**) with that of the CFP1 CXXC (in **h**, PDB: 3QMC). Of note, both cytosines of the CpG duplex form hydrogen bonds with the main chain carbonyl oxygens, while both guanines of the CpG duplex were also recognized by forming hydrogen bonds with the side chains.



#### Extended Data Figure 4. Creation of MTF2 KO mESCs and qPCR experiments

**a**, Western blot of endogenous MTF2 in mESCs. Three distinct isoforms are indicated. **b**, Schematic overview of the three MTF2 isoforms and their corresponding translational start sites. Positions of four test CRISPR gRNA targets are shown. **c**, Western blot of mESCs expressing a control of CRISPR construct or CRISPR constructs targeting the *Mtf2* gene as depicted in **b**. CRISPR 4 (in red) was used to obtain single cell clones. **d**, Sequence validation of two single cell clones. **e**, Western blotting of nucleoplasm and chromatin fractions from two MTF2 KO clones and control cells. **f, g**, RT-qPCR of control cells and two MTF2 KO clones (**f**) or control, KO, or MTF2 KO cells rescued with WT or K339A MTF2 (**g**). Data show mean  $\pm$  SD of three biological replicates. **h**, ChIP-qPCR experiments in control, MTF2 KO, and Rescued cells with the antibodies shown. Data show mean  $\pm$  SD of two biological replicates. Uncropped blots are shown in Supplementary Figure 1.

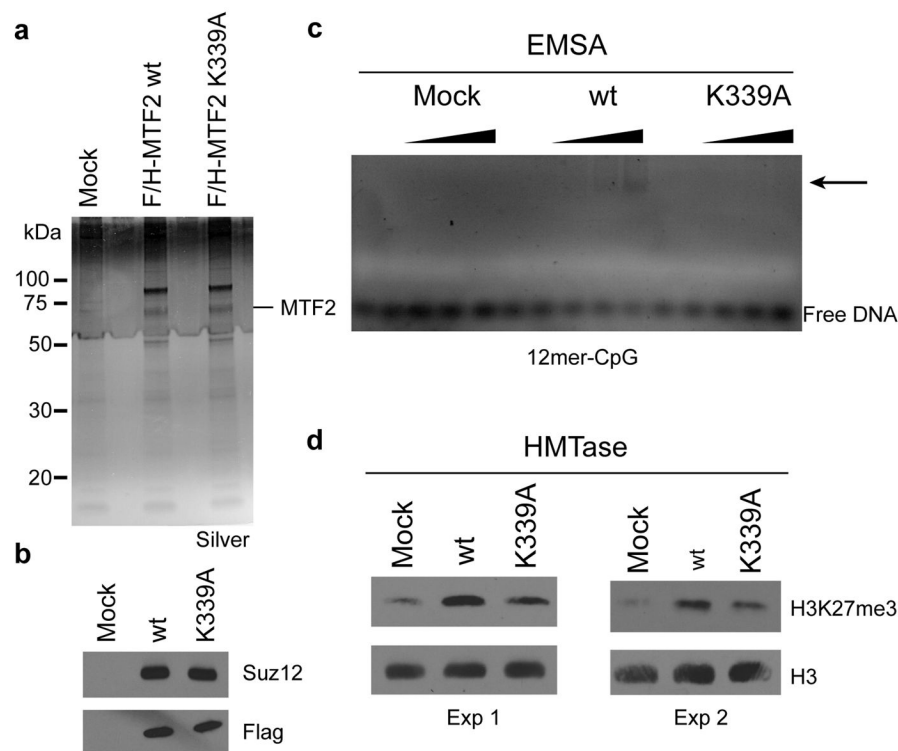




**Extended Data Figure 5. Analysis of the ChIP-seq experiments and PHF19 knockdown ChIP-seq data**

**a**, Comparison of normalized ChIP-Seq promoter reads (as in Fig. 4f) of three biological replicates for MTF2, SUZ12 and H3K27me3. The whisker-box plots represent the lower quartile, median and upper quartile of the data with 5 % and 95 % whiskers. **b**, Comparison of MTF2 ChIP-Seq data in Control and MTF2 KO cells (replicate 3) at the three promoter groups described in Fig. 4a. **c**, Promoter profiles of SUZ12 and H3K27me3 in control and PHF19 knockdown cells using publically available data<sup>10</sup>.





**Extended Data Figure 6. EMSA and HMTase experiments with purified MTF2-PRC2 complex**  
**a**, Silver staining of purified wildtype or K339A mutant human MTF2-PRC2 complexes (and Mock control) from HeLa-S cells. F/H = Flag-HA-tagged. **b**, Western blotting of the eluates from **a**. **c**, EMSA experiment with equal volume (0, 0.5, 1, 2, 3  $\mu$ l) of the eluates using the 12mer-CpG sequence. **d**, HMTase experiment using equal volume (15  $\mu$ l) of the eluates from **a**. Two technical replicates are shown. H3K27me3 levels were investigated by Western blotting. Uncropped blots are shown in Supplementary Figure 1.

**Extended Data Table 1**

X-ray statistics of the PHF1 and MTF2 Tudor-PHD1-PHD2-EH cassettes in the free or DNA and/or histone bound states.

Data collection and refinement statistics					
Crystal	Free human PHF1 (26-340) form 1	Free human PHF1 (26-340) form 2	Human PHF1 (26-360) and DNA complex	Human MTF2 (42-358) with DNA and H3(33-40)K36me3	Mouse PHF1 (26-360) with DNA and H3(29-41)K36me3
Beam line	APS-19ID	APS-19ID	SSRF-BL18U1	SSRF-BL19U1	SSRF-BL19U1
Wavelength	0.97918	0.97918	0.97791	0.97853	0.97852
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P3_12$	$C2$
Unit cell a, b, c ( $\text{\AA}$ )	40.0, 62.0, 135.4	61.5, 66.8, 76.2	109.6, 110.3, 118.9	137.7, 137.7, 101.2	141.2, 62.6, 97.3
Unit cell $\alpha$ , $\beta$ , $\gamma$ ( $^\circ$ )	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 120.0	90.0, 108.0, 90.0
Resolution ( $\text{\AA}$ )	36.49-1.90 (1.96-1.90) <sup>a</sup>	29.34-1.90 (1.96-1.90)	50.0-2.30 (2.34-2.30)	50.0-2.25 (2.30-2.25)	50.0-2.40 (2.46-2.40)

Data collection and refinement statistics					
Crystal	Free human PHF1 (26-340) form 1	Free human PHF1 (26-340) form 2	Human PHF1 (26-360) and DNA complex	Human MTF2 (42-358) with DNA and H3(33-40)K36me3	Mouse PHF1 (26-360) with DNA and H3(29-41)K36me3
R <sub>sym</sub>	0.135 (0.849)	0.132 (0.807)	0.123 (0.820)	0.124 (0.894)	0.086 (0.485)
I/σ (I)	16.6 (1.9)	18.8 (2.5)	22.4 (2.0)	19.6 (3.0)	17.9 (1.8)
Completeness (%)	97.3 (96.7)	100 (100)	99.9 (100)	100 (100)	99.1 (99.1)
Redundancy	4.7 (4.3)	6.6 (6.1)	9.2 (9.3)	16.9 (17.4)	3.6 (3.3)
Unique reflections	126328	171303	65188	52840	32456
R <sub>work</sub> /R <sub>free</sub> (%)	18.3/21.8	17.7/22.7	20.9/24.4	20.1/23.1	22.2/25.7
Number of non-H atoms					
Protein	2397	2509	7290	4979	5031
DNA	0	0	526	526	506
Water	183	203	190	331	56
ligands	4	4	28	8	8
Average B factors (Å <sup>2</sup> )					
Protein	27.9	24.7	62.1	38.9	57.8
DNA	no	no	58.6	41.0	72.1
Water	18.5	21.0	50.1	38.4	51.4
Other ligands	32.4	31.2	24.5	38.6	59.6
R.m.s. deviations					
Bond lengths (Å)	0.007	0.007	0.003	0.003	0.005
Bond angles (°)	0.998	1.04	0.641	0.634	0.793

<sup>a</sup>Highest resolution shell (in Å) shown in parentheses.

### Extended Data Table 2

ITC-based binding affinity measurements for the PCL cassettes or their mutants with DNAs or histones.

DNA or peptide	Protein Sample	K <sub>d</sub> (μM)	H (cal/mol)
12mer-CpG	PHF1 (26-360)	1.2 ± 0.3	−3608 ± 97
12mer-CpG	PHF1 (165-360)	0.5 ± 0.1	−5457 ± 88
12mer-ACpGG	PHF1 (165-360)	NB	
12mer-TCpGG	PHF1 (165-360)	0.8 ± 0.4	−1170 ± 80
12mer-CCpGG	PHF1 (165-360)	22 ± 4	4516 ± 467
12mer-GCpGA	PHF1 (165-360)	31 ± 2	4422 ± 207
12mer-GCpGT	PHF1 (165-360)	3.9 ± 0.5	−3053 ± 117
14mer-GCpGC	PHF1 (165-360)	11.3 ± 0.8	−4281 ± 125
12mer-ACpGA	PHF1 (165-360)	NB	
12mer-ACpGT	PHF1 (165-360)	NB	
12mer-TCpGA	PHF1 (165-360)	NB	
H3(29-43)K36me3	PHF1 (26-360)	2.0 ± 0.1	−9826 ± 62
H3(21-33)K27me3	PHF1 (26-360)	50 ± 7	−5970 ± 433

DNA or peptide	Protein Sample	$K_d$ ( $\mu$ M)	H (cal/mol)
H3(29-43)K36me3/R40A	PHF1 (26-360)	$5.2 \pm 0.3$	$-8169 \pm 70$
H3(1-15)K4me3	PHF1 (26-360)	$215 \pm 38$	$-8712 \pm 2195$
H3(29-43)	PHF1 (26-360)	NB	
H3(1-15)	PHF1 (26-360)	NB	
H3(29-43)K36me3	PHF1 (26-360)-Y47A	NB	
12mer-CpG DNA	MTF2 (180-378)	$2.1 \pm 0.3$	$-1767 \pm 44$
12mer-ACpGG	MTF2 (180-378)	$33 \pm 6$	$5836 \pm 1100$
12mer-TCpGG	MTF2 (180-378)	$6.4 \pm 1.0$	$2373 \pm 106$
12mer-CCpGG	MTF2 (180-378)	$12 \pm 1$	$5831 \pm 99$
12mer-GCpGA	MTF2 (180-378)	$22 \pm 7$	$3762 \pm 922$
12mer-GCpGT	MTF2 (180-378)	$25 \pm 4$	$3927 \pm 292$
14mer-GCpGC	MTF2 (180-378)	$9 \pm 2$	$2924 \pm 220$
12mer-ACpGA	MTF2 (180-378)	NB	
12mer-ACpGT	MTF2 (180-378)	NB	
12mer-TCpGA	MTF2 (180-378)	NB	
H3(29-43)K36me3	MTF2 (42-378)	$45 \pm 5$	$-3380 \pm 306$
H3(21-33)K27me3	MTF2 (42-378)	NB	
H3(29-43)	MTF2 (42-378)	NB	
H3(29-43)K36me3	MTF2 (42-378)-Y62A	NB	

NB, no detectable binding

### Extended Data Table 3

The names and sequences of the double-stranded DNAs used in the text. For each DNA duplex, only the sequence of one strand is listed in the table. Cytosine methylation is labeled as (m).

DNA name	DNA sequence
WH-motif	CTATGTAAACAAC
16mer-AT-rich	TTTTTATTAATAAAAA
12mer-CpG	GGGCGGCCGCC
12mer-GpC	GGGGGGCCCCC
12mer-ApG	GGGAGGCCTCCC
12mer-TpG	GGGTGGCCACCC
12mer-CpA	GGGCAGCTGCC
12mer-CpT	GGGCTGCAGCCC
12mer-CpC	GGGCCTAGGCC
12mer-ACpGG	GGACGGCCGTCC
12mer-TCpGG	GGTCGGCCGACC
12mer-CCpGG	GGCCGGCCGGCC
12mer-GCpGA	GGGCGATCGCCC

DNA name	DNA sequence
12mer-GCpGT	GGGCGTACGCCC
14mer-GCpGC	GGGCGCTAGCGCCC
12mer-ACpGA	GGACGATCGTCC
12mer-ACpGT	GGACGTACGTCC
12mer-TCpGA	GGTCGATCGACC
12mer-CpG-m1	GGGC(m)GGCCGCCC
12mer-CpG-m2	GGGC(m)GGCC(m)GCCC

Extended Data Table 4

Primers used for ChIP-qPCR and RT-qPCR.

Target genes	Forward primer sequences	Reverse primer sequences
<b>CHIP-qPCR</b>		
<i>Otx1</i>	AGTAGGCGTGCTCAGAGAGG	GGCCGGTCAAGAAGAAGTC
<i>Hoxb3</i>	CCGTGCGCATGAAGTACAAGA	CCTTAAGAGGGGGCTGGTAG
<i>Hoxa10</i>	CTTTTGGCGAGAACATCAAA	GTAGCCGGGTACTGGCACT
<i>Nkx2-9</i>	TGGCACCTTCCGGAATTG	AAGTGCGAGGCGCTCG
<i>Msx1</i>	ACAGAAAGAAATAGCACAGACCATAAGA	TTCTACCAAGTTCCAGAGGGACTTT
<i>Hoxa7</i>	GAGAGGTGGGCAAAGAGTGG	CCGACAACCTCATACCTATTCCTG
<i>Pou5f1</i>	GGCTCTCCAGAGGATGGCTGAG	TCGGATGCCCCATCGCA
<i>Sox2</i>	CCATCCACCCTTATGTATCCAAG	CGAAGGAAGTGGGTAAACAGCAC
<b>RT-qPCR</b>		
<i>Mtf2</i> (1)	ATGAGAGACTCTACAGGAGCAG	GCTAAGACATCTTGACCTCTTC
<i>Mtf2</i> (2)	CAGATGAAAAGTGGCTTTGTCTG	TGCATCCCATTCGAAGGTCAGC
<i>Tbx1</i>	CTGTGGGACGAGTTCAATCAG	TTGTCATCTACGGGCACAAAG
<i>Gata4</i>	CACAAGATGAACGGCATCAACC	CAGCGTGTGGTGGTAGTCTG
<i>Pax2</i>	AAGCCCGGAGTGATTGGTG	CAGGCGAACATAGTCGGGTT
<i>Pax3</i>	TCCCATGGTTGCGTCTCTAAG	CTCCACGTCAGGCGTTGTC
<i>Pax5</i>	CCATCAGGACAGGACATGGAG	GGCAAGTTCCACTATCCTTTGG
<i>Six1</i>	ATGCTGCCGTCGTTTGGTT	CCTTGAGCACGCTCTCGTT
<i>Hoxc13</i>	GCCGTCTACACGGACATCC	CCCCAAATGGGTAACCATAGC
<i>Msx1</i>	TGCTGCTATGACTTCTTTGCC	GCTTCCTGTGATCGGCCAT
<i>Msx3</i>	ACCTCCGCAAAACAAAAAC	CGCTCCGAATGGATAAGTAT
<i>Hoxa10</i>	CCTGCCGCGAACTCCTTTT	GGCGCTTCATTACGCTTGC
<i>Nkx1-2</i>	CGCTCTGCCCTATCAGACTTT	GGCCCAAGGAATGGAGTGA
<i>Meis1</i>	GCAAAGTATGCCAGGGGAGTA	TCCTGTGTTAAGAACCGAGGG
<i>Pou5f1</i>	AGAGGATCACCTTGGGGTACA	CGAAGCGACAGATGGTGGTC
<i>Nanog</i>	CACAGTTTGCTAGTTCTGAGG	GCAAGAATAGTTCTCGGGATGAA
<i>Sox2</i>	GCGGAGTGGAACCTTTTGTCC	GGGAAGCGTGTAATTATCCTTCT

Target genes	Forward primer sequences	Reverse primer sequences
<i>Esrrb</i>	GGACTCGCCGCCTATGTTC	CGTTAAGCATGTACTCGCATTG
<i>Tet1</i>	GCAGTGAACCCCGGAAAC	AGAGCCATTGTAAACCCGTG
<i>Zbtb7a</i>	CTTTGCGACGTGGTGATTCTT	CGTTCTGCTGGTCCACTACA

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the staff from BL17U1, BL18U1 and BL19U1 beamlines of National Facility for Protein Science in Shanghai (NFPS) at Shanghai Synchrotron Radiation Facility (SSRF) in China for assistance during data collection, and Caihong Yun at Peking University for help in data collection in U.S. We thank Guohua Jiang at Beijing Normal University for help in ITC studies. We thank Ashwini Jambhekar (Boston Children’s Hospital) for critical reading of the manuscript. This work was supported by the National Natural Science Foundation of China (31370719 and 31570729), Beijing Natural Science Foundation (5152015) and the Fundamental Research Funds for the Central Universities (2017EYT19) to Z.W. Early research on PCL protein expression undertaken by Z.W. in the lab of D.J.P. was supported by the Leukemia and Lymphoma Society and by the Memorial Sloan-Kettering Cancer Center Core Grant (P30 CA008748). Research on PCL proteins was supported by the German Research Foundation (DFG, LI 2057/1-1) to R.L., NIH/NHGRI R01 grant HG003985 to M.L.B., and the National Cancer Institute (R01 CA118487) to Y.S. Y.S. is an American Cancer Society Research Professor.

References

1. Comet I, Riising EM, Leblanc B, Helin K. Maintaining cell identity: PRC2-mediated regulation of transcription and cancer. *Nat Rev Cancer*. 2016; 16:803–810. DOI: 10.1038/nrc.2016.83 [PubMed: 27658528]

2. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. *Nature*. 2011; 469:343–349. DOI: 10.1038/nature09784 [PubMed: 21248841]

3. Hauri S, et al. A High-Density Map for Navigating the Human Polycomb Complexome. *Cell Rep*. 2016; 17:583–595. DOI: 10.1016/j.celrep.2016.08.096 [PubMed: 27705803]

4. Ballare C, et al. Phf19 links methylated Lys36 of histone H3 to regulation of Polycomb activity. *Nat Struct Mol Biol*. 2012; 19:1257–1265. DOI: 10.1038/nsmb.2434 [PubMed: 23104054]

5. Boulay G, Rosnoblet C, Guerardel C, Angrand PO, Leprince D. Functional characterization of human Polycomb-like 3 isoforms identifies them as components of distinct EZH2 protein complexes. *Biochem J*. 2011; 434:333–342. DOI: 10.1042/BJ20100944 [PubMed: 21143197]

6. Brien GL, et al. Polycomb PHF19 binds H3K36me3 and recruits PRC2 and demethylase NO66 to embryonic stem cell genes during differentiation. *Nat Struct Mol Biol*. 2012; 19:1273–1281. DOI: 10.1038/nsmb.2449 [PubMed: 23160351]

7. Cai L, et al. An H3K36 methylation-engaging Tudor motif of polycomb-like proteins mediates PRC2 complex targeting. *Mol Cell*. 2013; 49:571–582. DOI: 10.1016/j.molcel.2012.11.026 [PubMed: 23273982]

8. Cao R, et al. Role of hPHF1 in H3K27 methylation and Hox gene silencing. *Mol Cell Biol*. 2008; 28:1862–1872. DOI: 10.1128/MCB.01589-07 [PubMed: 18086877]

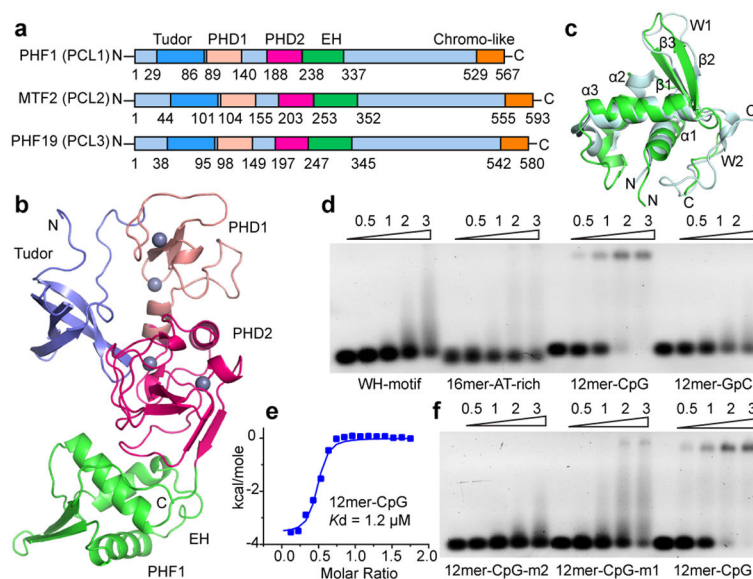
9. Casanova M, et al. Polycomblike 2 facilitates the recruitment of PRC2 Polycomb group complexes to the inactive X chromosome and to target loci in embryonic stem cells. *Development*. 2011; 138:1471–1482. DOI: 10.1242/dev.053652 [PubMed: 21367819]

10. Hunkapiller J, et al. Polycomb-like 3 promotes polycomb repressive complex 2 binding to CpG islands and embryonic stem cell self-renewal. *PLoS Genet*. 2012; 8:e1002576. [PubMed: 22438827]

11. Musselman CA, et al. Molecular basis for H3K36me3 recognition by the Tudor domain of PHF1. *Nat Struct Mol Biol.* 2012; 19:1266–1272. DOI: 10.1038/nsmb.2435 [PubMed: 23142980]
12. Walker E, et al. Polycomb-like 2 associates with PRC2 and regulates transcriptional networks during mouse embryonic stem cell self-renewal and differentiation. *Cell Stem Cell.* 2010; 6:153–166. DOI: 10.1016/j.stem.2009.12.014 [PubMed: 20144788]
13. Sarma K, Margueron R, Ivanov A, Pirrotta V, Reinberg D. Ezh2 requires PHF1 to efficiently catalyze H3 lysine 27 trimethylation in vivo. *Mol Cell Biol.* 2008; 28:2718–2731. DOI: 10.1128/MCB.02017-07 [PubMed: 18285464]
14. Ku M, et al. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* 2008; 4:e1000242. [PubMed: 18974828]
15. Kycia I, et al. The Tudor domain of the PHD finger protein 1 is a dual reader of lysine trimethylation at lysine 36 of histone H3 and lysine 27 of histone variant H3t. *J Mol Biol.* 2014; 426:1651–1660. DOI: 10.1016/j.jmb.2013.08.009 [PubMed: 23954330]
16. Qin S, et al. Tudor domains of the PRC2 components PHF1 and PHF19 selectively bind to histone H3K36me3. *Biochem Biophys Res Commun.* 2013; 430:547–553. DOI: 10.1016/j.bbrc.2012.11.116 [PubMed: 23228662]
17. Holm L, Rosenstrom P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* 2010; 38:W545–549. DOI: 10.1093/nar/gkq366 [PubMed: 20457744]
18. Callebaut I, Mornon JP. The PWAPA cassette: Intimate association of a PHD-like finger and a winged-helix domain in proteins included in histone-modifying complexes. *Biochimie.* 2012; 94:2006–2012. DOI: 10.1016/j.biochi.2012.05.025 [PubMed: 22664638]
19. Clark KL, Halay ED, Lai E, Burley SK. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature.* 1993; 364:412–420. DOI: 10.1038/364412a0 [PubMed: 8332212]
20. Biggs WH 3rd, Cavenee WK, Arden KC. Identification and characterization of members of the FKHR (FOX O) subclass of winged-helix transcription factors in the mouse. *Mamm Genome.* 2001; 12:416–425. DOI: 10.1007/s003350020002 [PubMed: 11353388]
21. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet.* 2012; 13:484–492. DOI: 10.1038/nrg3230 [PubMed: 22641018]
22. Gajiwala KS, et al. Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. *Nature.* 2000; 403:916–921. DOI: 10.1038/35002634 [PubMed: 10706293]
23. Berger MF, et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol.* 2006; 24:1429–1435. DOI: 10.1038/nbt1246 [PubMed: 16998473]
24. Li X, et al. Mammalian polycomb-like Pcl2/Mtf2 is a novel regulatory component of PRC2 that can differentially modulate polycomb activity both at the Hox gene cluster and at Cdkn2a genes. *Mol Cell Biol.* 2011; 31:351–364. DOI: 10.1128/MCB.00259-10 [PubMed: 21059868]
25. Kloet SL, et al. The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. *Nat Struct Mol Biol.* 2016; 23:682–690. DOI: 10.1038/nsmb.3248 [PubMed: 27294783]
26. Voo KS, Carlone DL, Jacobsen BM, Flodin A, Skalnik DG. Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Mol Cell Biol.* 2000; 20:2108–2121. [PubMed: 10688657]
27. Xu C, Bian C, Lam R, Dong A, Min J. The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. *Nat Commun.* 2011; 2:227. [PubMed: 21407193]
28. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc.* 2009; 4:393–411. DOI: 10.1038/nprot.2008.195 [PubMed: 19265799]
29. Xiao S, et al. Comparative epigenomic annotation of regulatory DNA. *Cell.* 2012; 149:1381–1392. DOI: 10.1016/j.cell.2012.04.029 [PubMed: 22682255]
30. Adams PD, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr.* 2010; 66:213–221. DOI: 10.1107/S0907444909052925 [PubMed: 20124702]

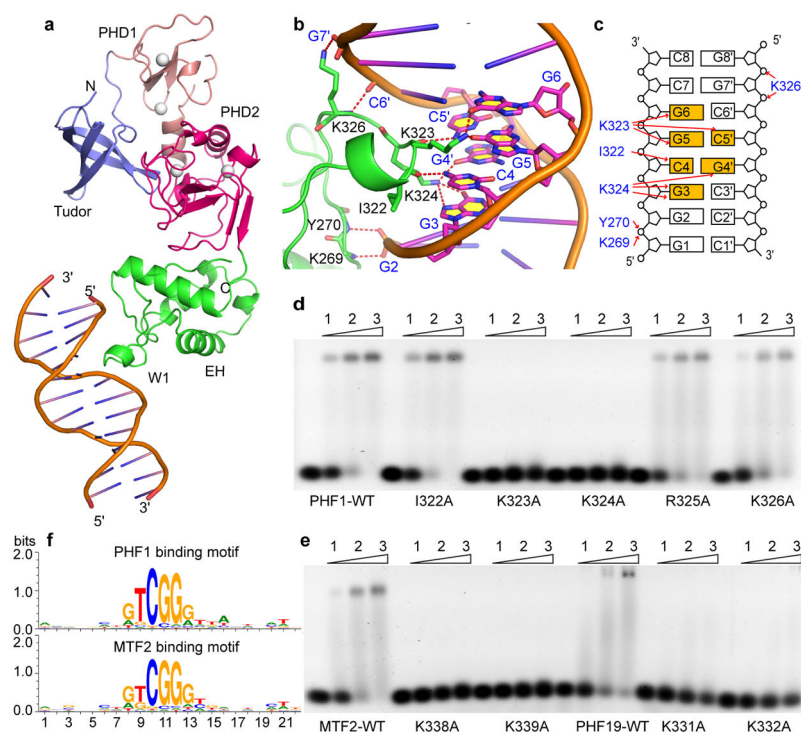


31. Langer G, Cohen SX, Lamzin VS, Perrakis A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc.* 2008; 3:1171–1179. DOI: 10.1038/nprot.2008.91 [PubMed: 18600222]
32. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr.* 2010; 66:486–501. DOI: 10.1107/S0907444910007493 [PubMed: 20383002]
33. Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods.* 2014; 11:783–784. DOI: 10.1038/nmeth.3047 [PubMed: 25075903]
34. Liefke R, Karwacki-Neisius V, Shi Y. EPOP Interacts with Elongin BC and USP7 to Modulate the Chromatin Landscape. *Mol Cell.* 2016; 64:659–672. DOI: 10.1016/j.molcel.2016.10.019 [PubMed: 27863226]
35. Kalb R, et al. Histone H2A monoubiquitination promotes histone H3 methylation in Polycomb repression. *Nat Struct Mol Biol.* 2014; 21:569–571. DOI: 10.1038/nsmb.2833 [PubMed: 24837194]
36. Trapnell C, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc.* 2012; 7:562–578. DOI: 10.1038/nprot.2012.016 [PubMed: 22383036]
37. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009; 10:R25. [PubMed: 19261174]
38. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* 2014; 42:W187–191. DOI: 10.1093/nar/gku365 [PubMed: 24799436]
39. Liu T, et al. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* 2011; 12:R83. [PubMed: 21859476]
40. Gentleman RC, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 2004; 5:R80. [PubMed: 15461798]
41. Bailey TL, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 2009; 37:W202–208. DOI: 10.1093/nar/gkp335 [PubMed: 19458158]



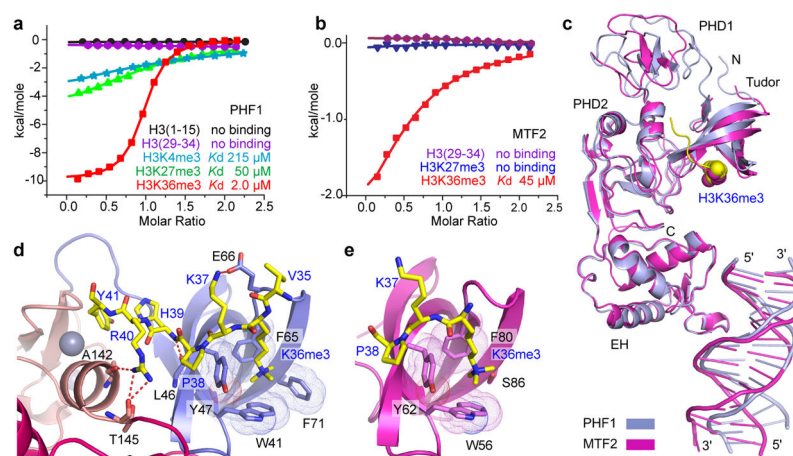
**Figure 1. PHF1 domain architecture, its free form structure and the binding analysis with various double-stranded DNAs**

**a**, Domain architecture of human PCL proteins. **b**, Free form structure of the PHF1 Tudor-PHD1-PHD2-EH cassette. The Tudor, PHD1, PHD2 and EH domains were colored in blue, salmon, magenta and green, respectively. Zinc ions were shown as grey balls. **c**, Overlapped structures of the PHF1 EH colored in green and the HNF-3γ winged-helix motif colored in cyan, with an r.m.s.d. of around 2.3 Å over 66 equivalent protein backbone atoms. **d**, EMSA results of the PHF1 cassette with different double-stranded DNAs. Protein to DNA molar ratios are shown above. **e**, ITC-based measurement of the PHF1 cassette with the 12mer-CpG DNA. **f**, EMSA analysis of the PHF1 cassette with hemi- or full- methylated 12mer-CpG DNAs. Protein to DNA molar ratios are indicated above. Data shown are representative of at least three independent experiments. Uncropped gels are shown in Supplementary Figure 1.



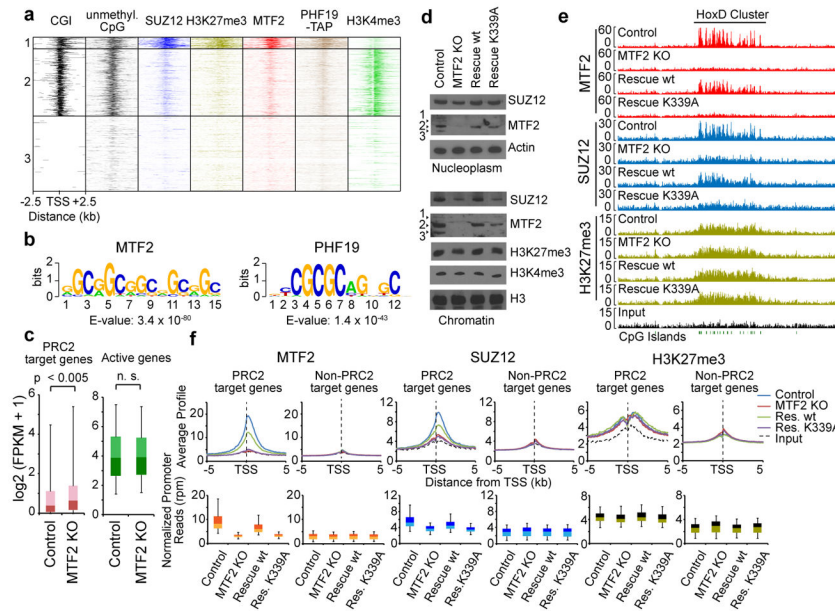
**Figure 2. Structural details of PHF1 with bound DNA, mutational analysis of the PCL cassettes, and identification of DNA motifs recognized by PHF1 and MTF2 through protein-binding microarrays**

**a**, Overall structure of the PHF1 cassette with bound DNA. **b**, Detailed interactions of the PHF1 EH domain with bound DNA. The PHF1 EH domain is colored in green. Hydrogen bonds are shown as red dotted lines. **c**, Schematic representation of PHF1-DNA interactions. **d, e**, EMSA results of the binding of 12mer-CpG DNA with wild type or mutant forms of PHF1 (in d), MTF2 and PHF19 (in e). Molar ratios of Protein to DNA are shown above. **f**, DNA binding specificity motifs recognized by the PHF1 and MTF2 PHD2-EH fragments identified from universal protein-binding microarrays using the Universal PBM Analysis Suite<sup>28</sup>. Information content (bits) on y-axis, position on x-axis. Data shown are representative of at least two independent experiments. Uncropped gels are shown in Supplementary Figure 1.



**Figure 3. Binding analysis of the PHF1 and MTF2 cassettes with various histone peptides and structural details of PHF1/MTF2 cassette-H3K36me3-DNA ternary complexes**

**a, b,** ITC-based measurements of the PHF1 (panel a) and the MTF2 (panel b) Tudor-PHD1-PHD2-EH cassettes with histone peptides. Data shown are representative of at least two independent experiments. **c,** Structural alignment of the PHF1-DNA-histone ternary complex (in blue) with that of the MTF2 ternary complex (in magenta). The PHF1-bound H3K36me3 peptide is colored in yellow, K36me3 was shown in a space-filling representation. **d, e,** Structural details of the interactions between the H3K36me3 peptide and the PHF1 cassette (panel d) or the MTF2 cassette (panel e) in their ternary complexes.



**Figure 4. The MTF2 EH domain is essential for PRC2 recruitment in mouse embryonic stem cells**

**a**, Heatmap of MTF2<sup>12</sup>, PHF19<sup>10</sup>, unmethylated CpGs<sup>29</sup> and SUZ12<sup>10</sup> at three promoter groups: CpG island (CGI)-containing promoters enriched for SUZ12 (Group 1,  $n = 2,008$ ), CGI-containing promoters with low SUZ12 (Group 2,  $n = 11,743$ ) or promoters without CGI (Group 3,  $n = 13,117$ ). **b**, Enriched DNA motifs at MTF2 and PHF19 bound locations. **c**, Gene expression (RNA-Seq) of control and MTF2 KO cells at PRC2 target genes and active non-PRC2 target genes (FPKM >1). The significance was estimated by one-way ANOVA with Tukey's post hoc test. n.s. = not significant. **d**, Western blotting of nucleoplasmic and chromatin fraction from mESCs that express endogenous MTF2 (Control), no MTF2 (MTF2 KO) or reintroduced wildtype (Rescue wt) and K339A mutant (Rescue K339A) MTF2 (isoform 2). **e**, Genome browser view of the HoxD cluster for ChIP-Seq data acquired from the four cell lines described. **f**, Promoter profiles of MTF2, SUZ12 and H3K27me3 at PRC2 target genes (Group 1 as in a) or non-PRC2 target genes (Group 2+3) in the four investigated cell lines. Normalized ChIP-seq promoter reads are presented as whisker blots. ChIP-Seq experiments were performed in three biological replicates, which were combined for the analysis (see also Extended Data Fig. 5a). The whisker-box plots represent the lower quartile, median and upper quartile of the data with 5 % and 95 % whiskers. Uncropped blots are shown in Supplementary Figure 1.