ARTICLES
https://doi.org/10.1038/s41588-018-0134-8
nature genetics

MTF2 recruits Polycomb Repressive Complex 2 by helical-shape-selective DNA binding

Matteo Perino[1], Guido van Mierlo[2], Ino D. Karemaker[2], Siebe van Genesen[1], Michiel Vermeulen[2], Hendrik Marks[2], Simon J. van Heeringen[1*] and Gert Jan C. Veenstra[1*]

Polycomb-mediated repression of gene expression is essential for development, with a pivotal role played by trimethylation of histone H3 lysine 27 (H3K27me3), which is deposited by Polycomb Repressive Complex 2 (PRC2). The mechanism by which PRC2 is recruited to target genes has remained largely elusive, particularly in vertebrates. Here we demonstrate that MTF2, one of the three vertebrate homologs of *Drosophila melanogaster* Polycomblike, is a DNA-binding, methylation-sensitive PRC2 recruiter in mouse embryonic stem cells. MTF2 directly binds to DNA and is essential for recruitment of PRC2 both in vitro and in vivo. Genome-wide recruitment of the PRC2 catalytic subunit EZH2 is abrogated in *Mtf2* knockout cells, resulting in greatly reduced H3K27me3 deposition. MTF2 selectively binds regions with a high density of unmethylated CpGs in a context of reduced helix twist, which distinguishes target from non-target CpG islands. These results demonstrate instructive recruitment of PRC2 to genomic targets by MTF2.

Tight regulation of gene expression is essential for patterning, establishment of the body plan, and cell fate determination and maintenance during embryonic development. Transcription of many developmental master regulators is controlled by the highly conserved Polycomb group proteins via monomethylation, dimethylation and trimethylation of lysine 27 of histone H3 (H3K27me1, H3K27me2 and H3K27me3, respectively) and ubiquitination of lysine 119 of histone H2A (H2AK119ub) by Polycomb Repressive Complex 2 and 1 (PRC2 and PRC1) respectively. PRC2 is composed of a heterotrimeric core formed by EED, SUZ12 and one of two paralogs, EZH1 or EZH2. A number of associated proteins define two PRC2 subcomplexes, PRC2.1 (containing one of the Polycomblike proteins PHF1 (PCL1), MTF2 (PCL2), PHF19 (PCL3) and C17ORF96 (EPOP)) and PRC2.2 (containing AEBP2 and JARID2)[1]. Several of these PRC2-associated proteins modulate its catalytic activity and interact with chromatin (for reviews, see refs [2–6]). An extensive body of work has uncovered molecular interactions with other proteins and RNAs[2,3,7], but the DNA-targeting mechanism by which PRC2 is recruited to genes in vertebrates has remained elusive[8,9]. In *Drosophila*, PRC2 is recruited to Polycomb-response elements (PREs) by a variety of DNA-binding proteins[9,10]; however, these proteins are not functionally conserved in vertebrates. In mouse, JARID2 and AEBP2 bind DNA with a weak preference for GC-rich DNA but without apparent sequence specificity[11]. The three mammalian homologs of *Drosophila* Polycomblike, PHF1, MTF2 and PHF19, feature a Tudor domain and two PHD domains and have been implicated in PRC2 function. PHF1 is important for PRC2 catalytic activity in human and mouse cells, and perturbation of its function leads to deregulation of the Hox loci[12,13]. MTF2 was originally identified as a protein binding to the metal response element (MRE) of the mouse metallothionein (*Mt1*) promoter[14], but was subsequently shown to modulate PRC2 activity at specific developmental genes, to regulate X chromosome inactivation and pluripotency[15–17]. The Tudor domains of PHF1 and PHF19, but not MTF2, bind with high affinity to the transcriptional elongation mark H3K36me3, lending support to chromatin-driven recruitment of PRC2 to silence expressed genes upon differentiation[16,18–21]. PHF19 binds to CpG islands, but this requires the histone-tail-binding abilities of its Tudor domain as well[22]. In the highly methylated genomes of mammals, PRC2 and the H3K27me3 modification are almost exclusively present in a subset of DNA-methylation-free regions called CpG islands[23]. Recent studies have solved the crystal structures of PCL proteins and their interaction with unmethylated DNA[24,25]. However, how PRC2 discriminates between the CpG islands of developmental genes from those that are not targeted has remained unclear. Polycomb recruitment to unmethylated CpGs is conserved between mammals and anamniote vertebrates such as *Xenopus* and zebrafish, even though the CpG dinucleotide density of these regions in the latter is much lower, not forming islands as in mammals[26]. Previously, we showed that early H3K27me3 nucleation sites of the frog *Xenopus tropicalis* are able to induce de novo H3K27me3 deposition in mouse embryonic stem cells (mESCs)[26]. We identified a pan-vertebrate conserved sequence signature using a machine learning algorithm that can classify genomic regions on the basis of short sequences called *k*-mers (*k*-mer-based support vector machine algorithm, or *k*-mer-SVM)[27]. This algorithm successfully identifies the subset of DNA methylation-free islands that acquire H3K27me3 in human, frog and fish[26]. However, the mechanistic basis for conserved recruitment of PRC2 to a specific set of unmethylated CpG islands remained unexplained.

Results

PRC2 recruitment to unmethylated DNA. To identify DNA sequences involved in recruitment of PRC2 to DNA, we compared the enrichment of the *k*-mer-SVM motif sequences in Polycomb domains of *Xenopus* embryos[26] and mouse ESCs[28]. Among top-scoring motifs, we identified TGCGCAAA as the most strongly enriched motif in both vertebrate species (Fig. 1a). To test the functionality of this sequence in recruiting Polycomb proteins, we performed DNA pulldown coupled to mass spectrometry with nuclear extracts from

[1]Department of Molecular Developmental Biology, Radboud University, Faculty of Science, Radboud Institute for Molecular Life Sciences, Nijmegen, the Netherlands. [2]Department of Molecular Biology, Radboud University, Faculty of Science, Radboud Institute for Molecular Life Sciences, Nijmegen, the Netherlands. *e-mail: s.vanheeringen@science.ru.nl; g.veenstra@science.ru.nl

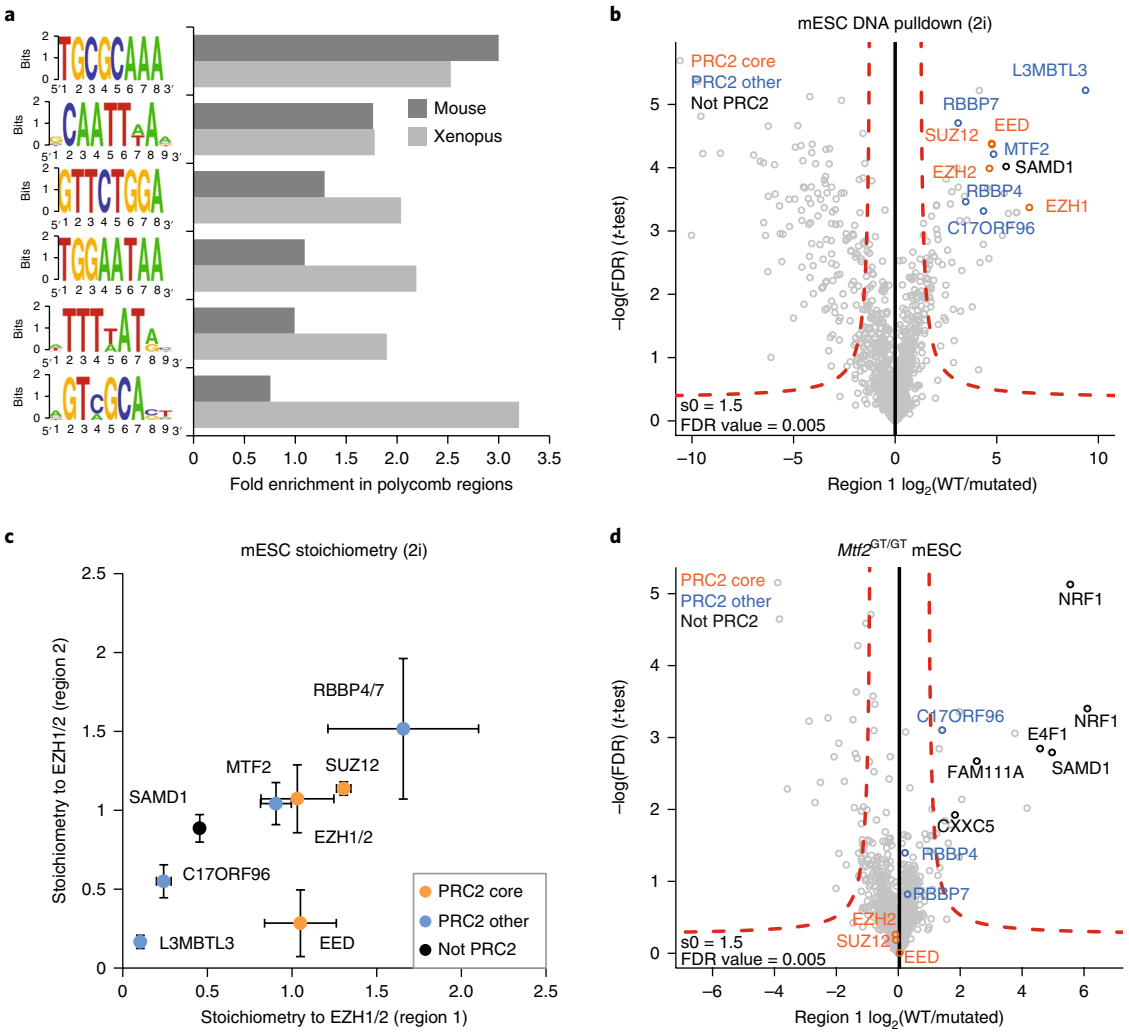

Fig. 1 | PRC2 is recruited by short DNA sequences. a, Enrichment of *k*-mer occurrence in *X. tropicalis* H3K27me3 regions[26] and EZH2 peaks in mESCs cultured in serum, showing the consistent enrichment of the TGCGCAAA motif in PRC2-targeted regions in both species when compared with untargeted CpG islands. **b**, DNA pulldown mass spectrometry of 2i mESC nuclear extract using region 1 baits. PRC2 core subunits and associated proteins show highly specific enrichment on the wild-type pulldown bait (right). Highlighted proteins are enriched in both region 1 and region 2 (Supplementary Fig. 1a) pulldowns. Each condition was measured in three independent experiments. False discovery rate (FDR) was calculated from a two-tailed *t* test. **c**, Stoichiometry of proteins highlighted in **b**. MTF2 is the only non-core protein consistently enriched in both baits with a stoichiometric ratio to EZH2. Each condition was measured in three independent experiments. Dots represent means, error bars s.d. **d**, DNA pulldown mass spectrometry of nuclear extract from serum-grown *Mtf2*$^{GT/GT}$ mESCs using region 1 baits. PRC2 core proteins fall in the background, indicating loss of recruitment. Each condition was measured in three independent experiments. FDR, false discovery rate from two-tailed *t* test; s0, artificial within-group variance (Perseus; see Methods).

mESCs cultured in the 2i condition[29]. These cells represent ground-state pluripotency and closely resemble preimplantation embryos, wherein H3K27me3 is prevalent[30]. Relative to a bait with four point mutations, the TGCGCAAA-containing 30-bp bait (region 1) showed highly specific binding of PRC2 core components EZH2/EZH1, EED and SUZ12, along with four PRC2-associated proteins, RBBP4 or RBBP7 (RBBP4/7), MTF2, C17ORF96 and L3MBTL3 (Fig. 1b). To rule out effects due to the sequence surrounding the motif or to the mutation itself, we repeated the experiment using baits with different flanking sequences and different point mutations in the motif (regions 2, 3 and 4, Supplementary Table 1). We assessed PRC2 recruitment by mass spectrometry and western blot and obtained similar results (Supplementary Fig. 1a–c). Notably, PRC2 recruitment was strongly reduced by DNA methylation of the central CpG of the *k*-mer (Supplementary Fig. 1b). RNase treatment of the samples did not affect PRC2 recruitment (Supplementary Fig. 1b), indicating the recruitment of PRC2 to DNA in this assay is not mediated by RNAs.

MTF2 is required for PRC2 binding to DNA. Since PRC2 core subunits do not bind DNA[3], we hypothesized that one of the proteins identified by mass spectrometry in our pulldown experiments would mediate binding of Polycomb to DNA. When calculating the stoichiometry of these proteins in the mass spectrometry data (Methods), MTF2 emerged as the only non-core protein that is stoichiometric relative to the catalytic subunit EZH2 (Fig. 1c) in all the replicates with both DNA baits. This compares favorably to the 0.4 stoichiometry of MTF2 to PRC2 core subunits observed in protein pulldowns with mESC nuclear extract[31]. Thus, these results suggest that our assay is enriching for an MTF2-containing PRC2 subcomplex, recently termed PRC2.1[11]. The PRC2.1 complexes identified in our pulldown experiments also contains C17ORF96 next to MTF2, albeit at substoichiometric levels. Therefore, we decided to test the DNA binding properties of Myc-tagged MTF2 and C17ORF96 produced in vitro. Myc-MTF2 was able to specifically bind to the wild-type bait, while Myc-C17ORF96

showed little if any binding regardless of the presence of MTF2 (Supplementary Fig. 1d). Moreover, purified recombinant GST-MTF2 delayed migration of a region 1 probe in an electrophoretic mobility shift assay (Supplementary Fig. 1e). Together with recent reports of the resolved crystal structure of MTF2 in complex with DNA[24], these results strongly support a role for MTF2 in DNA-mediated PRC2 recruitment.

To test whether MTF2 is necessary for recruitment of PRC2 to DNA, we used an *Mtf2* null line (*Mtf2*$^{GT/GT}$)[15] to assess the DNA binding ability of PRC2 in absence of MTF2. These cells did not grow well in the 2i medium, so we used serum-LIF (leukemia inhibitory factor) culture conditions instead. We then quantified PRC2 expression in the two systems with whole-proteome measurement by mass spectrometry and found a nearly double expression of MTF2 in 2i vs. serum conditions (Supplementary Fig. 2a), suggesting a higher requirement for MTF2 in 2i. Mass spectrometry confirmed specific PRC2 recruitment to both wild-type DNA baits in serum conditions, although with slightly lower enrichment (Supplementary Fig. 2b,c). In serum-grown wild-type cells we also found a similar MTF2 to EZH2 stoichiometry (Supplementary Fig. 2d), consistent in all the triplicates of each bait. By contrast, DNA pulldown with lysates of *Mtf2*$^{GT/GT}$ cells showed a complete disruption of PRC2 recruitment (Fig. 1d and Supplementary Fig. 2e), with none of the core PRC2 proteins being recruited to the bait. To assess whether loss of MTF2 affects the stability or abundance of PRC2 components rather than their recruitment, we performed whole-proteome analysis of wild-type and *Mtf2*$^{GT/GT}$ mESCs. MTF2 was not detected in the *Mtf2*$^{GT/GT}$ mutant mESCs, as expected, while the abundance of the PRC2 core components was unaffected (Supplementary Fig. 2f) and C17ORF96 expression was reduced. Notably, although JARID2 and AEBP2 were detected by mass spectrometry in the pulldown with wild-type cell extracts, they did not show sequence specificity for the wild-type bait and remained in the background cloud in both pulldowns, indicating that these proteins are not involved in this recruitment mechanism. The two other PCL homologs (PHF1, PHF19) were not detected, probably due to a very low abundance in mESCs[31]. Therefore, our pulldown results identify MTF2 as the protein required for PRC2 recruitment to DNA in vitro.

Functional domains of MTF2. The N-terminal part of all the mouse PCL paralogs is composed of a Tudor domain and two PHD domains. Multiple alignment of the mouse paralogs and the MTF2 proteins of different vertebrates shows the existence of a highly conserved region extending from the C terminus of the PHD2 domain (EH, extended homology domain; Fig. 2a and Supplementary Table 2)[32,33], followed by a lysine-rich region (Fig. 2b) that is more conserved among vertebrate MTF2 orthologs than among mouse PCL paralogs (Fig. 2a). Within the Polycomb complex, the PRC2 core component EZH1/2 is positioned in three-dimensional space in proximity to the Tudor domain (Lys68), the first PHD finger (Lys161) and the lysine-rich domain (Lys412)[31], while DNA binding mostly relies on the EH domain, which folds into a winged-helix structure[24,25]. Cocrystallization of either PHF1 or MTF2 with a short (12-bp) DNA sequence shows that the W1 loop of the EH domain enters into the major groove, directly contacting the CpG of the bait. As the PHD2 domain is required for PRC2 targeting[16], and given the potential for electrostatic interactions of the lysine-rich domain of MTF2 with DNA, we tested the DNA binding property of MTF2 constructs in DNA pulldown experiments (Fig. 2c). MTF2 isoform 2 (lacking the Tudor domain, hereafter referred to as MTF2) and the constructs also lacking the PHD1 domain (PHD2-stop) or the C-terminal domain (ΔC-term) all specifically bound to the wild-type DNA bait. Constructs encoding only the PHD domains, the C-terminal domain, or the PHD2 and EH domain but lacking the lysine-rich domain (PHD1+2, C-term, ΔApoI, ΔEcoRI/BlpI) lacked binding to DNA. Some nonspecific DNA interaction was detectable with constructs lacking the two PHD fingers (Pro256-stop) or encoding the lysine-rich domain plus C-terminal domain (Val353-stop), albeit with reduced affinity and minimal enrichment over background. This suggests that, as well as the DNA-binding EH domain, both the PHD2 and lysine-rich domains are necessary for binding. Since PHD2 interacts with the EH domain[24] it may be required to support EH binding to DNA, whereas the lysine-rich domain may further stabilize DNA binding by contacting our relatively long 30-bp bait.

Since the PHD1 domain was not required for DNA binding (Fig. 2c), we tested whether it is also dispensable for interactions with the PRC2 core complex. We used a mESC line (*Mtf2*$^{\Delta/\Delta}$)[15] that expresses only a shorter 46-kDa protein lacking the Tudor and PHD1 domains. DNA pulldown results (Supplementary Fig. 3a) with this line showed that the shorter MTF2 protein was still able to specifically recruit PRC2 to the wild-type bait, indicating that the PHD1 domain is not required for MTF2 interaction with PRC2. We tested this further in rescue experiments transfecting either Myc-MTF2 or Myc-PHD1+2 in *Mtf2*$^{GT/GT}$ cells and performing interaction proteomics analysis after pulldown for the Myc epitope tag. While Myc-MTF2 interacted with endogenous PRC2 (Supplementary Fig. 3b), the PHD1+2 protein did not interact with EZH2 (Supplementary Fig. 3c), indicating that PHD1 and PHD2 domains are not sufficient to mediate stable interaction with PRC2 in mESCs. In conclusion, PHD2 and the further C-terminal domains of MTF2 are sufficient for recruitment of PRC2 to DNA baits.

MTF2 recruits PRC2 genome-wide in mouse ES cells. To assess the role of MTF2 in PRC2 genome-wide recruitment in vivo, we performed chromatin immunoprecipitation and sequencing (ChIP-seq) for MTF2, EZH2, H3K27me3 and H3K4me3 in both wild-type and *Mtf2*$^{GT/GT}$ mESCs. MTF2 ChIP in *Mtf2*$^{GT/GT}$ cells resulted in virtually no chromatin recovery (Supplementary Fig. 4a), confirming the specificity of the anti-MTF2 antibody. ChIP-seq replicates were highly reproducible (Supplementary Fig. 4b) and were used to call high-confidence peaks used for downstream analysis (Methods). We used available BioCap data[34] for comparisons with unmethylated CpG islands. MTF2 ChIP-seq enrichment was found almost exclusively at a subset of unmethylated genomic locations, with good correspondence to EZH2 recruitment and H3K27me3 enrichment at all MTF2-binding sites (Fig. 3a,b). By contrast, in *Mtf2*$^{GT/GT}$ mESCs we found a striking genome-wide reduction in EZH2 recruitment and H3K27me3 deposition (Fig. 3a–c). Notably, the most intense reduction in H3K27me3 signal occurred in the central region of the peak (Fig. 3a,b), where most EZH2 (>80%) and H3K27me3 (>68%) enriched sites recruit MTF2 in wild-type mESCs. Virtually all EZH2 (96%) and most (>73%) of H3K27me3 peaks showed >50% reduction in *Mtf2* null cells (Supplementary Fig. 5a,b). Within this set, more than 65% of the total EZH2 peaks were even more strongly reduced (by >75%). By contrast, the active promoter mark H3K4me3 (Fig. 3a–c and Supplementary Fig. 5a,b) and the low levels of H3K27me3 at repetitive elements (Supplementary Fig. 5c) were largely unaffected. We note that a recent study reported no change in H3K27me3 upon *Mtf2* deletion[24]. We downloaded and mapped the raw data and found that the reduction in H3K27me3 may have escaped detection due to a lower sequencing depth and ChIP enrichment (Supplementary Fig. 6).

Since MTF2 is exclusively found in PRC2.1 complexes[1], we investigated its relation with PRC2.2 and performed ChIP-seq for MTF2 in *Jarid2*$^{-/-}$ cells[35] and for JARID2 in *Mtf2*$^{GT/GT}$ cells. Additionally, to assess the extent to which the presence of PRC2 stabilizes MTF2 binding, we also performed MTF2 ChIP-seq in *Eed*$^{-/-}$ cells, which completely lack functional PRC2[36]. We clustered the ChIP signals at EZH2-positive locations (Methods) and identified three clusters of peak regions (Fig. 4 and Supplementary Fig. 5d): sharp and medium-sized peaks (cluster 1 and cluster 2) where EZH2 binding was nearly

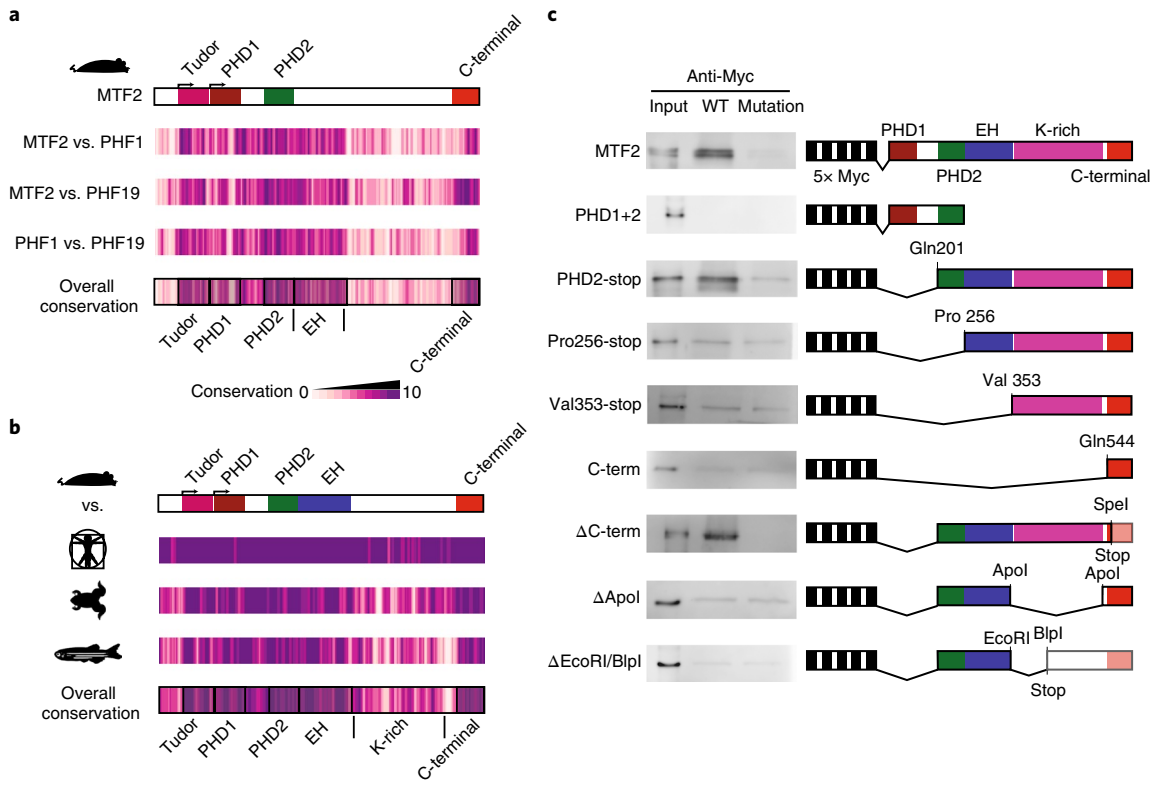

Fig. 2 | Functional domains of MTF2. a, Conservation score of pairwise and overall alignment of different mouse PCL homologs. Mouse PCL homologs show extended conservation C-terminal to PHD2 (Extended Homology, EH), which folds into a winged-helix structure, as shown using Praline conservation scores (0–10; Methods). **b**, Conservation score of human, *X. tropicalis* and zebrafish MTF2 in pairwise alignment with mouse MTF2 and overall vertebrate conservation. The lysine rich region (K-rich) of MTF2 is well conserved among vertebrates compared to the other PCL proteins in mouse (**a**), suggesting an evolutionary pressure against mutation in this area. **c**, DNA pulldown and western blot of the depicted MTF2 deletion constructs performed with region 1 baits, defining PHD2, EH and lysine-rich domains as the minimal region required for specific DNA binding. Input lane represents 1% of the starting material. Blots represent three independent pulldowns. Uncropped gels available in Supplementary Data 2.

abolished in the absence of MTF2, and broader peaks (cluster 3) where EZH2 binding was reduced but still present. Clusters 1 and 2 showed heavily reduced binding of JARID2 in the *Mtf2*$^{GT/GT}$ cells and reduced binding of MTF2 in *Jarid2*$^{-/-}$ cells. MTF2 binding was also decreased (but not abolished) in *Eed*$^{-/-}$ cells in cluster 1 and 2 regions, while MTF2 binding was abolished in cluster 3 regions in the absence of PRC2 (Fig. 4). The broad cluster 3 peaks also showed a more severe reduction of MTF2 binding in the absence of JARID2 and a relatively less severe reduction of JARID2 and EZH2 recruitment in the absence of MTF2. These observations suggest the presence of two different sets of PRC2 targets: a main one where both core PRC2 and JARID2 strongly depend on MTF2 for recruitment (MTF2 primary targets, clusters 1 and 2) and where the presence of PRC2 subsequently stabilizes MTF2 binding, and a smaller one, the secondary targets (cluster 3), where MTF2 cannot bind on its own without the PRC2 core complex and baseline binding is greatly enhanced by the presence of JARID2 (PRC2.2).

DNA sequence and helical shape dictate MTF2 binding. Given the role of MTF2 in PRC2 recruitment, we set out to predict vertebrate PREs on the basis of the sequences underlying MTF2 binding sites. We used the *k*-mer-SVM algorithm[26,27] to distinguish DNA methylation-free BioCap regions with and without MTF2 and found that MTF2-bound regions could be reliably classified on the basis of sequence alone using *k*-mers of different lengths. We then evaluated algorithm performance with the receiver operating characteristic area under the curve (ROC-AUC, Supplementary Fig. 7a). Nucleotide and dinucleotide content ($k=1$, $k=2$) showed reasonable classification power, reflecting known characteristics of Polycomb targets such as G+C richness and CpG dinucleotide density. Classification performance, however, improved substantially from $k=3$, with an optimal *k*-mer size of 6 or 7 base pairs (ROC-AUC 0.92, Supplementary Fig. 7a). This suggests a role for additional nucleotide positions in MTF2 binding site specification. The classification is based on multiple positive- and negative-scoring *k*-mers, including several with at least one CpG dinucleotide. However, we could not identify any strong consensus beyond the CpG dinucleotide itself, which is suggestive of sequence ambiguity among favored flanking sequences. The preferred sequence context, however, often contains a G just before the CpG (Supplementary Fig. 7b), as in the sequence of the original TGCGCAAA bait and the one used for the crystal structure[24]. We then calculated the enrichment of the highest and lowest scoring GCG-containing *k*-mers in MTF2 peak summits (Methods) and found that the ones with the highest SVM score were also the most enriched in MTF2 binding sites (Supplementary Fig. 7c). Strikingly, the two most enriched *k*-mers were both contained (with 1-bp permutation for one of them) in the MRE sequence previously shown to recruit MTF2 to the promoter of the *Mt1* gene[14]. We then tested the MRE sequence by DNA pulldown and confirmed its ability to recruit MTF2 and PRC2 (Supplementary Fig. 7d). These three *k*-mers, however, only share the presence of the GCG trinucleotide, which is not sufficient to explain the specificity for Polycomb recruitment, as negative-scoring *k*-mers and DNA baits not bound by MTF2 also contain GCG trinucleotides (Supplementary Fig. 7c and Supplementary Table 1). This led us to search for other

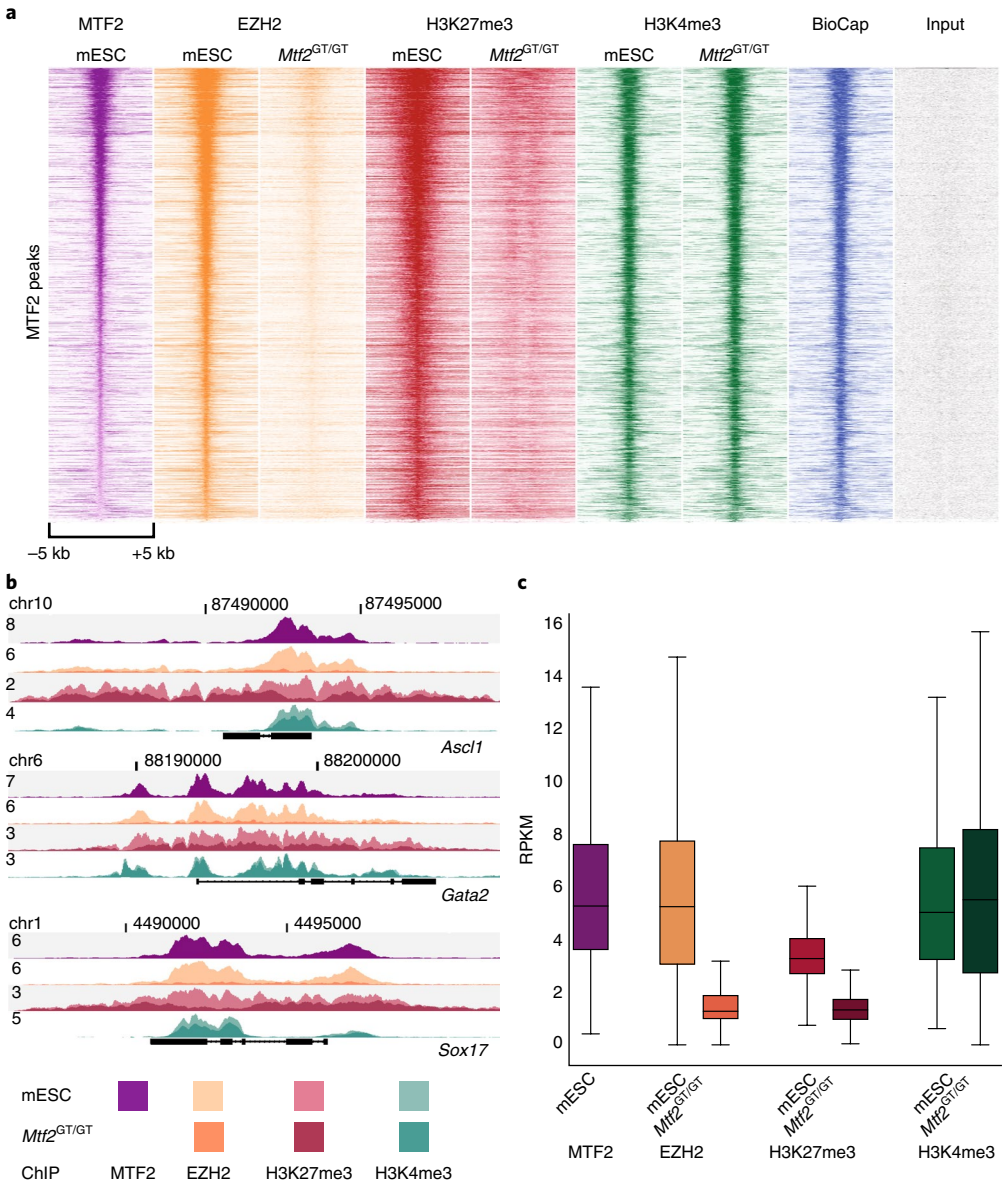

Fig. 3 | Mtf2 is required for PRC2 recruitment. a, Heat map of ChIP-seq signals on MTF2 high-confidence peaks, showing loss of genome-wide recruitment of PRC2 in *Mtf* $^{GT/GT}$ mESCs. **b**, Example of ChIP-seq signal (reads per million) on master regulators of embryonic development, known targets of PRC2 in vertebrates. **c**, Box plot quantification of signal shown in **a**, based on 6,357 MTF2 peaks. Box plots represent median and interquartile range (IQR; whiskers, 1.5 IQR). All ChIP-seq experiments were performed in duplicate.

sequence properties that could explain MTF2 binding specificity. The MTF2–DNA crystal structure shows a relatively unwound helix, and MTF2 interacts with the DNA backbone in addition to an unmethylated CpG[24]. Besides providing base-pair identity information, the DNA sequence is known to determine DNA helical structure, and in particular the GC dinucleotide has a strong effect widening the minor groove[37]. Furthermore, DNA helical shape is relevant for the prediction of bound vs. unbound transcription factor binding sites[38]. This is due to the stacking interactions between adjacent nucleotides that embed information about the three-dimensional shape of the DNA in the sequence of short *k*-mers[39]. However, multiple sequences can adopt the same shape, which could explain the lack of a classical consensus sequence flanking the GCG. Although DNA sequence and its associated helical shape are difficult to disentangle, we wondered to what extent differences in DNA helical shape can explain the differences in MTF2 binding to a variety of CpG-containing sequences. We therefore used DNA shape prediction tools (Methods) to investigate differences in the shape of the MTF2-recruiting *k*-mers and the unmethylated CpG islands they are found in. We found that unmethylated CpG islands showed an increased minor groove width, a decreased propeller twist and a decreased helix twist when compared to methylated flanking genomic regions, a difference that was even more pronounced in MTF2-bound regions and also correlated with CpG density (Supplementary Fig. 7e,f). At nucleotide resolution, while both positive- and negative-scoring SVM *k*-mers contained GCG trinucleotides, the helical structure of these *k*-mers showed opposite changes in propeller twist (respectively up-down and down-up at positions –1 and –2 relative to the GCG) and helix twist (respectively down and up at position –2; Fig. 5a), with additional differences in the minor groove width at the first G and a wider range of roll values in both flanking bases. The DNA pulldown baits we used for our experiments had a shape corresponding to that of the positive-scoring *k*-mers at the GCG and surrounding

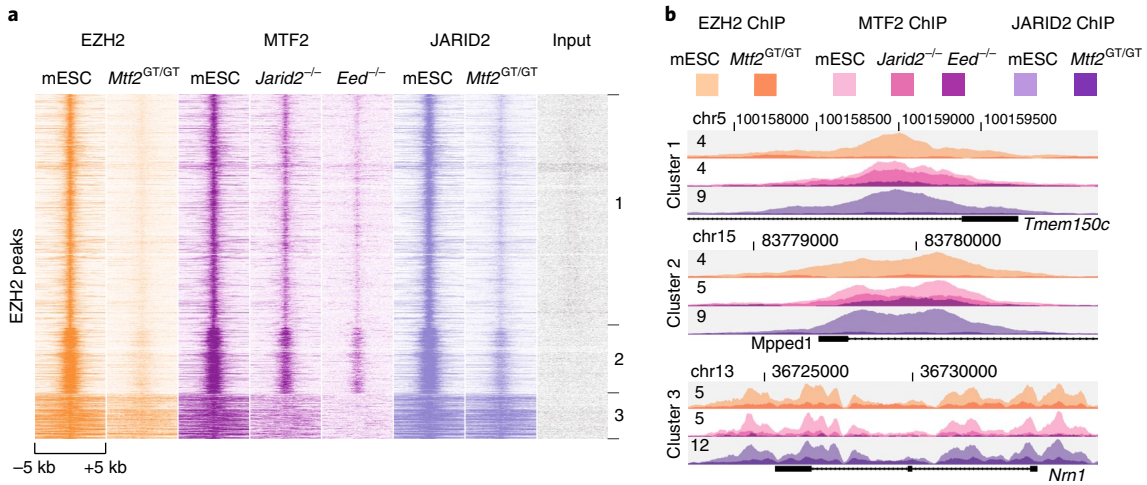

Fig. 4 | Primary and secondary MTF2 targets. a, Heat map of clustered ChIP-seq signal on EZH2 high-confidence peaks, showing different behavior in primary (clusters 1 and 2) and secondary (cluster 3) MTF2 targets. **b**, Example of ChIP-seq signal (reads per million) on peaks belonging to each of the clusters shown in **a**. With the exception of *Eed*$^{-/-}$, the ChIP-seq experiments were performed in duplicate.

positions, particularly at the 5′ side. We therefore decided to test the role of DNA shape in MTF2 binding by performing DNA pulldown with baits carrying single-base-pair mutations predicted to perturb DNA shape (Fig. 5b,c). DNA pulldowns performed with Myc-tagged MTF2 and with mESC nuclear extracts showed highly concordant mutation and DNA methylation sensitivities (Fig. 5c). Specifically, the central unmethylated CpG dinucleotide was critical but not sufficient for binding, as shown by the effect of flanking mutations that also affect the helical structure of the bait. Moreover, the mutations that most severely reduced MTF2 binding cause helical shape perturbations that lie outside the average shape profile of positive-scoring *k*-mers, while the least perturbing one almost perfectly mirrored the shape of the wild-type bait (Fig. 5b), lending further support to a role of DNA helical shape in MTF2 binding to DNA.

To further investigate the role of DNA shape in determining MTF2 binding sites, we tested whether we could predict MTF2 bound regions using only shape information. We predicted the DNA shape of all the GCG trinucleotides in MTF2 peak summits and used machine learning to classify them against nucleotide-composition-matched controls (Methods). The algorithm was able to identify differences between MTF2-bound vs. unbound unmethylated islands on the basis of helical shape alone (ROC-AUCs > 0.7) (Supplementary Fig. 8a). The distribution of DNA shape values also showed differences from control regions at several positions, including the central GCGs and the first neighboring bases (Supplementary Fig. 8b). To test the hypothesis that MTF2 binding relies on properly shaped GCG sequences, we tested more sequences for binding in the DNA pulldown assay. In particular, we tested GCG sequences from locations containing the top two enriched *k*-mers (regions 6 and 7) but in which the *k*-mers did not match the predicted ideal shape due to the flanking regions. Each of these regions, however, had at least one additional shape-matching GCG in the immediate vicinity (Supplementary Fig. 8c,d). The baits with the wild-type sequence efficiently recruited EZH2 and MTF2 (Fig. 5d), but this binding was not lost when mutating the GCG of the *k*-mer with disfavored shape properties, confirming our prediction. Instead, mutation of the GCGs with favorable shape flanking the *k*-mer abolished MTF2 binding and EZH2 recruitment (Fig. 5d), showing that DNA binding by MTF2 closely tracks DNA helical shape properties of qualifying GCG-containing sequences. The helical shape properties defined here are consistent with the helical shape of the DNA sequence used for the MTF2–DNA cocrystal (Supplementary Fig. 8e). Additionally, shape features might provide directionality to the binding site, thereby breaking the CpG palindrome, as shown by the reverse complement of the sequence in the crystal structure, which completely misses the acceptable feature range at critical positions (Supplementary Fig. 8e).

Next we quantified the occurrence of total and shape-matching GCGs and found that these preferred sequences were strongly enriched in Polycomb-targeted CpG islands but not in unbound CpG islands (Supplementary Fig. 8f,g), explaining the strong preference of MTF2 and PRC2 for a specific subset of unmethylated islands in the genome. We also tested whether a difference in the number of shape-qualifying GCGs could explain the differences of primary (clusters 1 and 2) and secondary (cluster 3) MTF2 targets discussed above (Fig. 4). The primary MTF2 targets showed a much higher enrichment of helical-shape-qualifying GCG compared to the broad, secondary MTF2 target peaks (Fig. 5e,f). Taken together, these analyses document the sequence and DNA helical shape properties of MTF2 binding and their role in PRC2 recruitment, defining a vertebrate analog of Polycomb response elements.

Discussion

Polycomb-mediated repression is critical for stem cell renewal and maintenance of cell identity. However, how PRC2 is targeted to DNA in vertebrates and how this relates to the well-known PREs present in *Drosophila* has been enigmatic. The experiments described here suggest a model for PRC2 recruitment (Fig. 6) that unifies a large body of observations: (i) instructive recruitment of PRC2 based on DNA sequence that is reminiscent of *Drosophila* PRE-based recruitment, (ii) a major role for unmethylated islands in PRC2 targeting, (iii) cooperation between PRC2.1 and PRC2.2, and (iv) DNA helical shape features distinguishing between Polycomb-recruiting and non-recruiting unmethylated islands. In vivo, we show that MTF2 is required for DNA-driven PRC2 recruitment to chromatin in mESCs. This is especially true for a large subset of EZH2 peaks, where primary MTF2 recruitment is necessary for both PRC2 core and JARID2 recruitment. A minority of EZH2 peaks instead show inhibition of MTF2 binding in the absence of JARID2 and EED, suggesting that MTF2 binding to these regions relies on the presence of PRC2.2. This difference in recruitment can be explained by the different enrichment for shape-qualifying MTF2 binding sites, which provides a potential general definition of vertebrate PREs. On both primary and secondary MTF2 targets, PRC2.1 and PRC2.2 affect each other, as

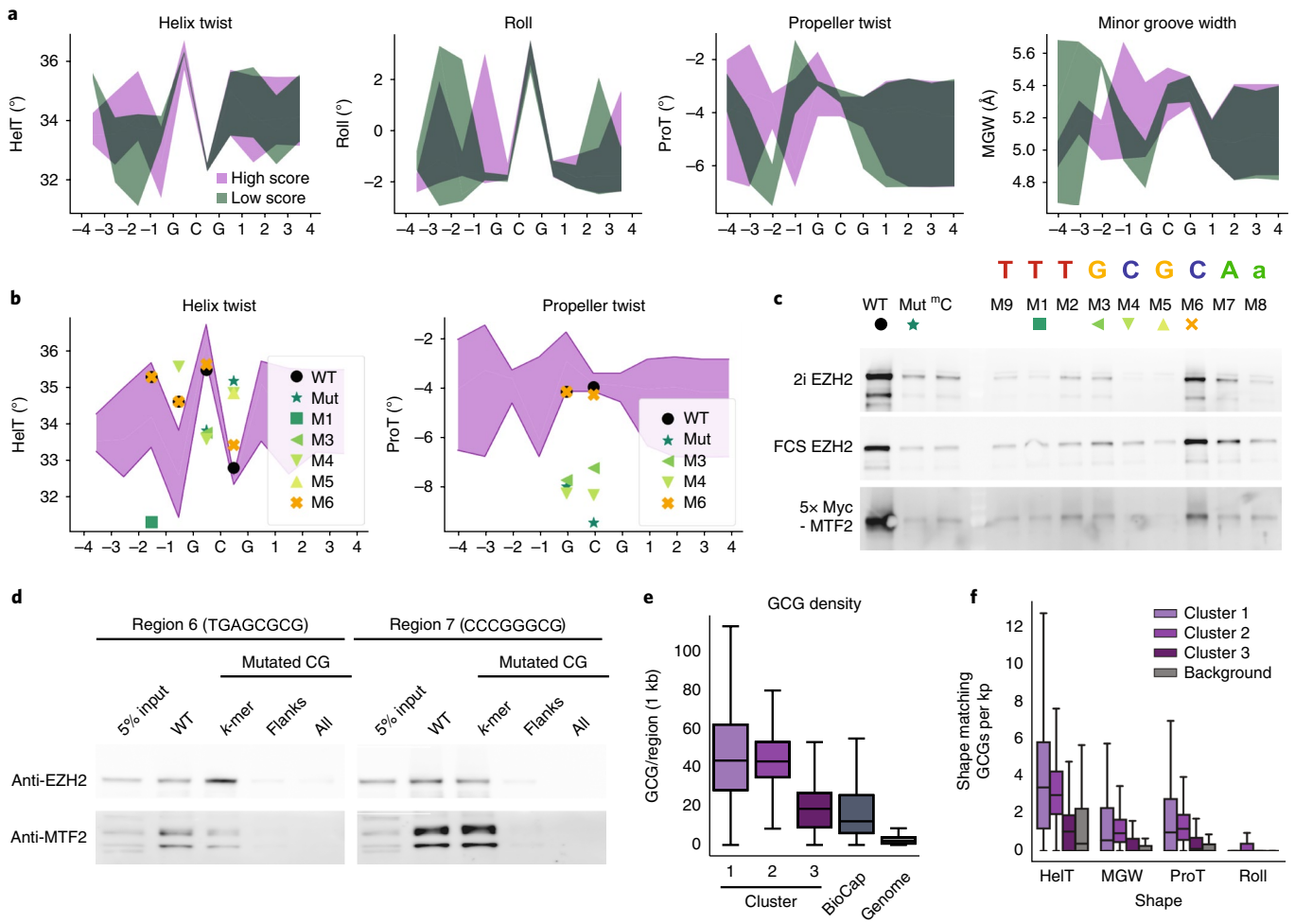

Fig. 5 | Role of DNA sequence and shape in MTF2 binding. a, Interquartile range (IQR) of DNA shape of GCG-containing *k*-mers (positive-scoring *k*-mers, purple, $n = 2{,}017$ regions; negative-scoring *k*-mers; green, $n = 872$ regions; see Methods). Helix twist (HelT), minor groove width (MGW), propeller twist (ProT) and roll show distinct profiles between the two groups. **b**, HelT and ProT of positive *k*-mers (IQR). Symbols represent the DNA shape values of baits with mutations (Mut, M1–M6, compare **c**; only relevant positions shown). **c**, DNA pulldown with mESCs (grown in serum (FCS) and 2i medium) and Myc-MTF2 using baits with mutations (Supplementary Table 1). M6, the mutation with the smallest effect on DNA shape (**b**), shows the weakest perturbation of binding (**c**). The *k*-mer forms an almost perfect palindrome (*k*-mer sequence indicated on top, flanking position mutated in M8), but the binding requirements appear asymmetric (M2 and M3 vs. M6 and M7). ^{m}C, methylated CG. **d**, DNA pulldown with mESC extract using baits with GCG-containing *k*-mers enriched in MTF2-bound regions. The *k*-mers do not match the preferred helical shape in the context of the bait, whereas each bait contains at least one GCG with preferred shape in the flanking region (Supplementary Fig. 7d,e and Supplementary Table 1). **e**, GCG density of EZH2 peaks stratified according to the clusters from Fig. 4a ($n = 4{,}894$, 1,365, 954, 7,213, 8,092 respectively in clusters 1–3, BioCap and genome). **f**, Quantification of shape-qualifying GCGs (Methods), showing higher enrichment of matching GCGs in primary MTF2 targets (clusters 1 and 2 from Fig. 4a). Background represents untargeted BioCap regions ($n = 4{,}812$, 1,356, 875, 7,004 respectively in clusters 1–3 and background regions). Blots are representative of three independent pulldowns. Uncropped gels available in Supplementary Data 2. Box plots: central bar, median; box, IQR; whiskers, 1.5 IQR.

shown by reduced MTF2 binding in *Jarid2*$^{-/-}$ and JARID2 binding in *Mtf2*$^{GT/GT}$ cells. Possible explanations of this reciprocal influence could be found in the presence of an intricate web of interactions among Polycomb complexes[2–6]: (i) the known binding of EED to H3K27me3, which would result in the indirect recruitment of both MTF2 and JARID2 to the chromatin; (ii) an indirect recruitment mediated by PRC1, which could bind PRC2-deposited H3K27me3 and catalyze H2AK119ub, which can in turn can be bound by JARID2; (iii) additional interplay of other PRC2 accessory proteins or (iv) interaction with RNAs. This scenario is also in line with the phenotype of *Mtf2* mutant mice, which show homeotic transformations but delayed lethality when compared to core PRC2 mutations[15]. In vitro, MTF2 sensitivity to DNA methylation and its role in recruitment is in line with its known association with PRC2[40] and modulation of its activity[16]. Additionally, while this study was in revision and as mentioned above, a crystal structure of MTF2 bound to DNA was published[24], confirming the DNA binding ability of MTF2. We note that crystallized MTF2 not only targets the bases of the CpG but also establishes direct contact with the backbone of the DNA. Besides the shape similarity of the DNA in the crystal to that of all our MTF2-bound baits, this strongly supports our prediction of DNA-shape-reading properties. Moreover, the *Drosophila* Polycomblike protein cooperates with Phol at the *Ubx* PRE to recruit PRC2[41], suggesting an important ancestral function of PCL proteins. Further effort will be required to explore the role of the PCL proteins in different cell types during differentiation and development. Also, how PRC2 is recruited to regions not relying on MTF2-bound DNA elements needs further investigation, as well as a careful dissection of the interaction web that orchestrates PRC2.1, PRC2.2 and PRC1 regulation. These findings open

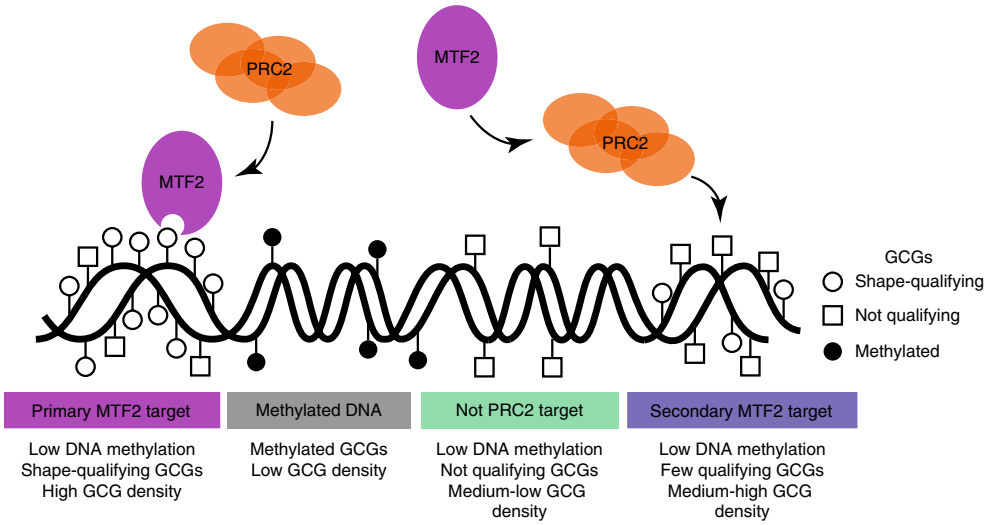

Fig. 6 | Model of MTF2-mediated PRC2 recruitment. At primary targets, MTF2 specifically binds unmethylated GCG trinucleotides that show specific features of DNA shape (white circles) in the context of CpG islands with a high density of CpGs, thereby recruiting PRC2 and nucleating the Polycomb domain. JARID2 and PRC2 core subunits support stronger binding and complete establishment of the domain (not depicted). At secondary targets GCGs do not have the preferred DNA shape (white squares), preventing direct MTF2 binding to DNA. Here PRC2 is nucleated via alternative mechanisms and MTF2 recruitment depends on JARID2 and PRC2 core subunits. MTF2 binding is absent at genes not targeted by PRC2 due to lack of shape-matching GCGs and alternative means of PRC2 recruitment. DNA methylation and the very low enrichment of GCGs prevent recruitment outside of CpG islands.

a new angle on cancer biology, cellular reprogramming and stem cell biology, wherein Polycomb-mediated regulation is known to be important.

URLs. Data track hub for UCSC Genome Browser, http://veenstra.science.ru.nl/trackhubm.htm and http://trackhub.science.ru.nl/hubs/mouse_veenstra/hub.txt; spp package, https://github.com/hms-dbmi/spp.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0134-8.

Received: 16 September 2017; Accepted: 6 April 2018;
Published online: 28 May 2018

References

1. Hauri, S. et al. A high-density map for navigating the human Polycomb complexome. *Cell Rep.* **17**, 583–595 (2016).
2. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. *Nat. Struct. Mol. Biol.* **20**, 1147–1155 (2013).
3. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343–349 (2011).
4. Comet, I., Riising, E. M., Leblanc, B. & Helin, K. Maintaining cell identity: PRC2-mediated regulation of transcription and cancer. *Nat. Rev. Cancer* **16**, 803–810 (2016).
5. Simon, J. A. & Kingston, R. E. Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. *Mol. Cell* **49**, 808–824 (2013).
6. Blackledge, N. P., Rose, N. R. & Klose, R. J. Targeting Polycomb systems to regulate gene expression: modifications to a complex story. *Nat. Rev. Mol. Cell Biol.* **16**, 643–649 (2015).
7. Brockdorff, N. Noncoding RNA and Polycomb recruitment. *RNA* **19**, 429–442 (2013).
8. Bauer, M., Trupke, J. & Ringrose, L. The quest for mammalian Polycomb response elements: are we there yet? *Chromosoma* **125**, 471–496 (2016).
9. Kassis, J. A. & Brown, J. L. Polycomb group response elements in Drosophila and vertebrates. *Adv. Genet.* **81**, 83–118 (2013).
10. Ringrose, L., Rehmsmeier, M., Dura, J. M. & Paro, R. Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster. *Dev. Cell* **5**, 759–771 (2003).
11. Grijzenhout, A. et al. Functional analysis of AEBP2, a PRC2 Polycomb protein, reveals a Trithorax phenotype in embryonic development and in ESCs. *Development* **143**, 2716–2723 (2016).
12. Cao, R. et al. Role of hPHF1 in H3K27 methylation and Hox gene silencing. *Mol. Cell. Biol.* **28**, 1862–1872 (2008).
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.* **28**, 2718–2731 (2008).
14. Inouye, C., Remondelli, P., Karin, M. & Elledge, S. Isolation of a cDNA encoding a metal response element binding protein using a novel expression cloning procedure: the one hybrid system. *DNA Cell Biol.* **13**, 731–742 (1994).
15. 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.* **31**, 351–364 (2011).
16. 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* **138**, 1471–1482 (2011).
17. Walker, E., Manias, J. L., Chang, W. Y. & Stanford, W. L. PCL2 modulates gene regulatory networks controlling self-renewal and commitment in embryonic stem cells. *Cell Cycle* **10**, 45–51 (2011).
18. Cai, L. et al. An H3K36 methylation-engaging Tudor motif of polycomb-like proteins mediates PRC2 complex targeting. *Mol. Cell* **49**, 571–582 (2013).
19. Musselman, C. A. et al. Molecular basis for H3K36me3 recognition by the Tudor domain of PHF1. *Nat. Struct. Mol. Biol.* **19**, 1266–1272 (2012).
20. Brien, G. L. et al. Polycomb PHF19 binds H3K36me3 and recruits PRC2 and demethylase NO66 to embryonic stem cell genes during differentiation. *Nat. Struct. Mol. Biol.* **19**, 1273–1281 (2012).
21. Ballaré, C. et al. Phf19 links methylated Lys36 of histone H3 to regulation of Polycomb activity. *Nat. Struct. Mol. Biol.* **19**, 1257–1265 (2012).
22. Hunkapiller, J. et al. Polycomb-like 3 promotes polycomb repressive complex 2 binding to CpG islands and embryonic stem cell self-renewal. *PLoS Genet.* **8**, e1002576 (2012).
23. Mendenhall, E. M. et al. GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet.* **6**, e1001244 (2010).
24. Li, H. et al. Polycomb-like proteins link the PRC2 complex to CpG islands. *Nature* **549**, 287–291 (2017).
25. Choi, J. et al. DNA binding by PHF1 prolongs PRC2 residence time on chromatin and thereby promotes H3K27 methylation. *Nat. Struct. Mol. Biol.* **24**, 1039–1047 (2017).
26. van Heeringen, S. J. et al. Principles of nucleation of H3K27 methylation during embryonic development. *Genome Res.* **24**, 401–410 (2014).
27. Lee, D., Karchin, R. & Beer, M. A. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* **21**, 2167–2180 (2011).

28. Marks, H. et al. The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**, 590–604 (2012).
29. Ying, Q. L. et al. The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).
30. Liu, X. et al. Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. *Nature* **537**, 558–562 (2016).
31. Kloet, S. L. et al. The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. *Nat. Struct. Mol. Biol.* **23**, 682–690 (2016).
32. Coulson, M., Robert, S., Eyre, H. J. & Saint, R. The identification and localization of a human gene with sequence similarity to Polycomblike of Drosophila melanogaster. *Genomics* **48**, 381–383 (1998).
33. O'Connell, S. et al. Polycomblike PHD fingers mediate conserved interaction with enhancer of zeste protein. *J. Biol. Chem.* **276**, 43065–43073 (2001).
34. Long, H. K. et al. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *Elife* **2**, e00348 (2013).
35. Landeira, D. et al. Jarid2 is a PRC2 component in embryonic stem cells required for multi-lineage differentiation and recruitment of PRC1 and RNA Polymerase II to developmental regulators. *Nat. Cell Biol.* **12**, 618–624 (2010).
36. Schoeftner, S. et al. Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. *EMBO J.* **25**, 3110–3122 (2006).
37. van der Heijden, T., van Vugt, J. J., Logie, C. & van Noort, J. Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. *Proc. Natl Acad. Sci. USA* **109**, E2514–E2522 (2012).
38. Mathelier, A. et al. DNA shape features improve transcription factor binding site predictions in vivo. *Cell Syst.* **3**, 278–286.e274 (2016).
39. Yang, L. et al. Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.* **13**, 910 (2017).
40. Vizán, P., Beringer, M., Ballaré, C. & Di Croce, L. Role of PRC2-associated factors in stem cells and disease. *FEBS J.* **282**, 1723–1735 (2015).
41. Savla, U., Benes, J., Zhang, J. & Jones, R. S. Recruitment of Drosophila Polycomb-group proteins by Polycomblike, a component of a novel protein complex in larvae. *Development* **135**, 813–817 (2008).

Acknowledgements

We thank H. Koseki (RIKEN Research Center for Allergy and Immunology, Japan) for sharing the *Mtf2*$^{GT/GT}$ and *Mtf2*$^{\Delta/\Delta}$ mESC lines and C. Fisher (MRC Clinical Sciences Centre, Imperial College School of Medicine, UK) for sharing the *Jarid2*$^{-/-}$ mESCs. We are grateful to M. Makowski for help and discussion. We thank P. Jansen, S. Kloet, A. H. Smits and L. N. Nguyen for advice and technical support with mass spectrometry, E. Janssen-Megens for help with Illumina sequencing, S. Wardle for help with ChIP and G. Georgiou for help with Python scripting. This work has been financially supported by the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program FP7 under grant agreement number 607142 (DevCom). Research in the group of H.M. is supported by a grant from the Netherlands Organization for Scientific Research (NWO-VIDI 864.12.007).

Author contributions

M.P., S.J.v.H. and G.J.C.V. designed experiments and analysis. M.P. performed experiments and analysis. G.v.M. and H.M. designed, performed and analyzed mESC whole-proteome experiments and contributed to ChIP. M.P., G.J.C.V., I.D.K. and M.V. designed rescue experiments and Myc pulldowns. I.D.K. performed and analyzed rescue experiments and Myc pulldowns. S.v.G. produced and purified GST-MTF2. M.V. helped with DNA pulldown experimental design. M.P. and G.J.C.V. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0134-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.J.H. or G.J.C.V.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Cell culture and mESC lines. E14[28], *Mtf2*$^{GT/GT15}$, *Mtf2*$^{\Delta/\Delta15}$, *Jarid2*$^{-/-35}$ and *Eed*$^{-/-36}$ mESC lines were cultured in serum or 2i medium as described in ref. [28]. Cultured cells were harvested for nuclear protein extraction as in ref. [31] and nuclear extract used for DNA pulldown. All cell lines were tested for mycoplasma contamination.

Protein production. *Mtf2* and *C17orf96* coding sequences were amplified from mESC cDNA, cloned in the pT7TS plasmid already containing the Myc tag, and Sanger sequenced to check for mutations. MTF2 deletion constructs were obtained either by restriction digestion deletion or by PCR amplification and cloning into the same vector. Linearized plasmids were used for in vitro transcription with the Amplicap-Max T& High Yield Message Maker kit (CellScript) and the purified mRNA translated with the Wheat Germ Extract kit (Promega). GST-MTF2 (pGex-5×1_mMtf2_iso2) was produced in *Escherichia coli* C3013 (NEB). 500 ml LB medium was inoculated with 5 ml of the overnight culture and incubated at 30 °C in a shaker at 200 r.p.m. At OD_{600} = 0.3 the culture was placed in a shaker at 15 °C at 225 r.p.m. for 2 h. Cells were induced by adding IPTG (0.1 mM), further cultured overnight at 15 °C, pelleted for 20 min at 5,000 r.p.m. in a GSA rotor (Sorvall) and resuspended in 30 ml ice-cold 10 mM Tris-HCl, pH 7.3, 140 mM NaCl, 2.7 mM KCl, 1 µM $ZnCl_2$ (TNKZ). The suspension was sonicated (6×30 s) and centrifuged at 20,000 r.p.m. in a SS34 rotor (Sorvall) at 4 °C. The supernatant was loaded on a - ml GST-trap-FF column (GE) that was pre-equilibrated with TNKZ at room temperature at 0.2 ml/min and then washed with 15 ml TNKZ. GST-mMTF2 was eluted with 5 ml 10 mM GSH (Sigma) in TNKZ in 0.5-ml fractions.

Gel shift. The indicated amount of protein was incubated with 20 pmol of biotinylated region 1 probe for 30 min on ice in 11% glycerol, 16 mM Tris-HCl, pH 8.0, 60 mM KCl, 10 ng/µl BSA, 10 ng/µl dIdC, 0.6 mM DTT, 1 mM $ZnCl_2$. 8% polyacrylamide (80:1) gels in 0.25×TBE supplemented with 0.5 mM $ZnCl_2$ were prerun at 4 °C for 1 h. Samples were loaded on the gel and run at 10 V/cm, transferred to nylon membranes and UV-cross-linked with 120 µJ/cm^2. Biotinylated probes were detected by chemiluminescence using the Chemiluminescent Nucleic Acid Detection Module Kit (ThermoFisher, 89880).

PCL protein conservation. PCL fasta sequences were obtained from Uniprot and aligned using PRALINE[42]. A five-amino-acid sliding window was used to calculate the average local conservation based on the PRALINE conservation score for each amino acid, and the data were color-coded using custom Python code.

Whole-cell proteome analysis. Cells were lysed in radioimmunoprecipitation assay (RIPA) buffer on ice followed by mild sonication to ensure efficient cell lysis. Protein mixtures were denatured using a standard filter-aided sample preparation (FASP) workflow[43] and digested overnight with trypsin. Tryptic peptides were desalted and purified with StageTips[44] before liquid chromatography and tandem mass spectrometry.

Myc pulldown and mass spectrometry. Nuclear extracts from *Mtf2*$^{GT/GT}$ mESCs transiently transfected using polyethylenimine were used to perform label-free Myc pulldowns in triplicate. Per pulldown, 15 µl of Myc-trap_A 50% bead slurry (Chromotek) was used. Beads were washed three times with buffer C (300 mM NaCl, 20 mM HEPES KOH pH 7.9, 20% (v/v) glycerol, 2 mM $MgCl_2$, 0.2 mM EDTA, 0.25% NP40, 0.5 mM DTT, complete protease inhibitor). Nuclear extract (450 µg) and ethidium bromide (final concentration 50 µg/ml) were diluted to a total volume of 400 µl with buffer C and rotated with the beads for 90 min at 4 °C. After beads were washed twice with buffer C with 0.5% NP40, twice with PBS plus 0.5% NP40, and twice with PBS, all supernatant was removed using a 30 G syringe. Beads were then resuspended in 50 µl elution buffer (2 M urea, 100 mM Tris, pH 8.5, 10 mM DTT) and incubated for 20 min in a thermoshaker at 1,400 r.p.m. at room temperature. After addition of 50 mM iodoacetamide (IAA), beads were incubated for 10 min at 1,400 r.p.m. at room temperature in the dark. Proteins were then on-bead digested into tryptic peptides by addition of 0.25 µg trypsin and subsequently incubated for 2 h at 1,400 r.p.m. at room temperature. The supernatant was transferred to new tubes and further digested overnight at room temperature with an additional 0.1 µg of trypsin. Tryptic peptides were acidified with 0.5% TFA and purified on C18-StageTips1. Tryptic peptides were eluted from StageTips and separated on an Easy-nLC 1000 (Thermo Scientific) connected online to an LTQ-Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific), using a 114-min gradient of acetonitrile (7–32%), followed by washes at 50% then 90% acetonitrile, for 140 min of total data collection. Scans were collected in data-dependent top-speed mode of a 3-s cycle with dynamic exclusion set at 60 s. Peptides were searched against the UniProt mouse proteome with MaxQuant (version 1.5.1.0), using default settings and match between runs enabled. Data were analyzed with Perseus (version 1.4.0.0) version and R (Supplementary Data 1).

DNA pulldown and mass spectrometry. DNA pulldown and mass spectrometry measurement were performed as in ref. [45] using oligonucleotides representing *X. tropicalis* genomic regions, centered on the *k*-mer in analysis. Briefly, biotinylated 30-bp oligonucleotide were bound to streptavidin–Sepharose beads (GE Healthcare) and incubated with either mESC nuclear extract or recombinant proteins. After extensive washes, bound proteins were either analyzed by western blot or digested with trypsin for mass spectrometry analysis. DNA pulldown samples were measured on a Q Exactive (Thermo Scientific) and whole proteomes on an Orbitrap Fusion Tribrid (Thermo Fisher Scientific) mass spectrometer. All experiments using mass spectrometry were performed in triplicate, with wild-type and mutant bait samples processed in parallel. Mass spectrometry raw data were aligned to Uniprot mouse proteome, using MaxQuant (version1.5.3.30) enabling options for 'match between runs', LFQ and iBAQ[46]. Output tables were analyzed with Perseus (version 1.5.0.15, MaxQuant package), stoichiometry ratios calculated from iBAQ values as in ref. [47] and plotting performed in R. Oligonucleotides used for DNA pulldown are listed in Supplementary Table 1.

ChIP and sequencing. Chromatin extraction and ChIP were performed as described in ref. [31]. Five nanograms per ChIP sample were prepared for sequencing with the Kapa Hyper-prep Kit (Kapa Biosystems) using NEXTflex adapters (Bio Scientific) and amplified with ten cycles of PCR. Libraries were size-selected to obtain 300-bp fragments using E-gel (Invitrogen) and sequenced on an Illumina NextSeq machine to obtain 75-bp reads. Quantitative PCR analysis of ChIP DNA was performed with iQ SYBR Green Supermix (Bio-Rad) on a CFX96 Real-Time System C1000 Thermal Cycler (Bio-Rad). Oligonucleotides used for ChIP-qPCR are listed in Supplementary Table 3.

Antibodies. ChIP was performed using 3 µl per sample of the following antibodies: MTF2 (Aviva System Biology ARP34292, lot QC49692-42166), H3K27me3 (Millipore 07-449, lot 2717675), EZH2 (Diagenode C15410039, lot 003), H3K4me3 (Ab858, lot GR240214-4), JARID2 (Novus Biologicals NB100-2214, Lot E2). Western blots were stained with EZH2 (Cell Signaling 52465, lot 7, 1:2,000), Myc (Santa Cruz sc-789, lot D1715, 1:1,000) or MTF2 (Protein Tech, 16208-1-AP, lot 88-478-4522, 1:2,000) primary antibodies, which were detected using Dako secondary antibodies (P0161, lot 20033538, P0217 20040441).

Bioinformatic analysis. Illumina 75-bp sequencing files were mapped using bwa (version 0.7.10-r789) and normalized for sequencing depth before loading in the UCSC Genome Browser track hub (see "Data availability"). Peaks were called with MACS2-2.7[48] using the –no-model option and manual shift provided with the –extsize parameter. The extent of shifting was calculated with spp R library (see URLs). A *q* value threshold of 0.001 was applied in all cases and either the –call-summits (MTF2 and EZH2) or the –broad (H3K27me3 and H3K4me3) parameter was used. High-confidence conserved peaks were identified with MAnorm[49] allowing a maximum of 1.5-fold change between replicates. Peak summits were defined as the non-overlapping 100-bp region around the summits called by MACS in the high-confidence peaks. Heat maps of ChIP-seq signal were generated using fluff v2.1.0[50] and clustered for dynamics using the "–g" option. The same analysis pipeline was applied for analysis of published data. Motif search and *k*-mer analysis was performed with GimmeMotifs v0.8.6[51] and *k*-mer-SVM v1.0[27]. *k*-mers for enrichment analysis were selected among those ending with a GCG trinucleotide to maximize the number of base pairs (and therefore the information content) on the 5′ end of the sequence. The resulting pool of *k*-mers was filtered by SVM weight, and *k*-mers scoring higher than 1.5 or lower than –1.5 were used for enrichment analyses. DNA shape predictions were performed on sequences from MTF2 peak summits containing the *k*-mers (in natural sequence context) using the sliding pentamer model and R package described in refs [52,53]. Random forest classification of DNA shape features was performed using the scikit-learn Python package[54]. Control regions were generated shuffling the sequence of each entry of the positive set while preserving the number of pre-existing GCGs and their position in the sequence. Shape-qualifying GCGs of BioCap regions with and without MTF2 were defined as those falling within the IQR of the shape prediction for the high-scoring SVM *k*-mers for the shape parameter in analysis. Only positions from –2 to + 1 of the GCG were considered to define shape match.

Statistical test and reproducibility. Significance of mass spectrometry data was determined by two-tailed *t* test in Perseus (see above). For each condition, three samples were processed separately. ChIP-seq experiments were performed in duplicate as indicated in the legends, and reproducibility was verified by correlation analysis and visual inspection in the genome browser.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The R script used to plot mass spectrometry data is available in Supplementary Data 1.

Data availability. ChIP-seq reads, coverage as genome browser tracks, and peak files have been deposited in the GEO repository under accession code GSE94300. Proteomic data have been deposited in the ProteomeXchange via the PRIDE[55] partner repository under identifier PXD005821. Processed data can be visualized with the UCSC Genome Browser using the Trackhub link indicated in the URLs section. Figures 2c and 5c,d and Supplementary Figs. 1b–e, 2e, 3a and 7d have associated source data in Supplementary Data 2.

References

42. Simossis, V. A. & Heringa, J. The PRALINE online server: optimising progressive multiple alignment on the web. *Comput. Biol. Chem.* **27**, 511–519 (2003).
43. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).
44. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
45. Bogdanović, O. et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat. Genet.* **48**, 417–426 (2016).
46. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
47. Smits, A. H., Jansen, P. W., Poser, I., Hyman, A. A. & Vermeulen, M. Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Res.* **41**, e28 (2013).
48. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
49. Shao, Z., Zhang, Y., Yuan, G. C., Orkin, S. H. & Waxman, D. J. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.* **13**, R16 (2012).
50. Georgiou, G. & van Heeringen, S. J. fluff: exploratory analysis and visualization of high-throughput sequencing data. *PeerJ* **4**, e2209 (2016).
51. van Heeringen, S. J. & Veenstra, G. J. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* **27**, 270–271 (2011).
52. Chiu, T. P. et al. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* **32**, 1211–1213 (2016).
53. Zhou, T. et al. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.* **41**, W56–W62 (2013).
54. Fabian Pedregosa, G.V. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
55. Vizcaíno, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44 D1**, D447–D456 (2016).

Corresponding Author: GJC Veenstra, SJ van Heeringen

Date: Mar 27, 2018

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

No statistical methods were performed to determine sample size. ChIP-seq was performed in two independent replicates, as recommended by the ENCODE ChIP-seq guidelines. For the mass spectrometry enrichment analyses, triplicates of test and control pulldowns are required for a two-tailed t-test and FDR calculation. This is a standardized workflow.

2. Data exclusions

Describe any data exclusions.

replicate 2 of MTF2 ChIPseq in EED KO did not pass QC and was not included in analysis

3. Replication

Describe whether the experimental findings were reliably reproduced.

All ChIPseq experiment were performed in duplicate all DNA-pulldown/ Mass spectrometry in triplicate. Replicates passing QC were considered successful replication of the experiments. Individual replicates were consistent for the effects reported.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Randomized design not possible for ChIPseq and DNApulldown experiments

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blind design not possible for ChIPseq and DNApulldown experiments

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☒	☐	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

ChIPseq and DNAshape analysis:
Base-calling was performed by the Illumina CASAVA software.
Reads were aligned to the mouse genome (mm10) using bwa 0.7.10-r789 with default settings.
Peaks were called with MACS2-2.7(Zhang et al., 2008) using the --no-model option and manual shift provided with the –extsize parameter. The extent of shifting was calculated with spp R library (https://github.com/hms-dbmi/spp). A q value threshold of 0.001 was applied in all cases and either the --call-summits (MTF2 and EZH2) or the --broad (H3K27me3 and H3K4me3) parameters was used. High-confidence conserved peaks were identified with MAnorm(Shao et al., 2012) allowing a maximum of 1.5 fold change between replicates. Peak summits were defined as the non-overlapping 100bp region around the summits called by MACS in the high confidence peaks. Heatmaps of ChIPseq signal were generated using fluff v2.1.0(Georgiou and van Heeringen, 2016). The same analysis pipeline was applied for analysis of published data. Motif search and kmer analysis was performed with GimmeMotif v0.8.6(van Heeringen and Veenstra, 2011) and kmer-SVM v1.0(Lee et al., 2011). Kmers for enrichment analysis were selected among those ending with a GCG trinucleotide to maximize the number of base pairs (and therefore the information content) on the 5' end of the sequence. The resulting pool of kmers was filtered by SVM weight and kmer scoring higher that 1.5 or lower that -1.5 were used for enrichment. DNA shape predictions were performed on sequences from MTF2 peak summits containing the kmers in analysis using the sliding pentamer model and R package described in(Chiu et al., 2016; Zhou et al., 2013). Random forest classification of DNA shape features was performed using the scikit-learn python package(Fabian Pedregosa, 2011). Control regions were generated shuffling the sequence of each entry of the positive set while preserving the number of pre-existing GCG and their position in the sequence. Shape-qualifying GCGs of BioCap regions with and without MTF2 were defined as those falling within the IQR of the shape prediction for the high-scoring SVM kmers for the parameter in analysis. Only positions from -2 to +1 of the GCG were considered to define shape match.
DNA pulldown Mass Spectrometry:
Mass spectrometry raw data were aligned to Uniprot mouse proteome, using MaxQuant (version1.5.3.30) enabling options for 'match between runs', LFQ, and iBAQ(Cox J. and Mann M., 2008). Output tables were analyzed with Perseus (version 1.5.0.15, MaxQuant package), stoichiometry ratios calculated from iBAQ values as in47 and plotting performed in R.
Interaction Proteomics:
Peptides were searched against the UniProt mouse proteome with MaxQuant (version 1.5.1.0), using default settings and match between runs enabled. Data were analyzed with Perseus (version 1.4.0.0)version and in-house R scripts available as Supplementary Information.
Protein conservation:
PCL fasta sequences were obtained from Uniprot and aligned using PRALINE(Simossis, V. A. & Heringa, J. 2003). A five-amino acid sliding window was used to calculate the average local conservation based on the PRALINE conservation score for each amino acid, and color coded using custom python code.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No restrictions

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	ChIP was performed using 3ul/sample of the following antibodies: MTF2 (Aviva System Biology ARP34292, lot QC49692-42166), H3K27me3 (Millipore 07-449, lot 2717675), EZH2 (Diagenode C15410039, lot 003), H3K4me3 (Ab8580, lot GR240214-4), JARID2 (Novus Biologicals NB100-2214, Lot E2). Mtf2 Ab was validated for ChIP in house (Fig S4a). All other antibodies are validate by the manufacturer and ChIP/IP grade. Western blots were stained with EZH2 (Cell Signaling 52465, lot 7, 1:2000), Myc (Santa Cruz sc-789, lot D1715, 1:000) or MTF2 (Protein Tech, 16208-1-AP, lot 88-478-4522) primary antibodies. All antibodies are validated for WB by the manucacturer.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	E14 ref 28, Mtf2GT/GT ref15, Mtf2Δ/Δ ref15, Jarid2-/- ref35 and Eed-/- ref36
b. Describe the method of cell line authentication used.	Mtf2GT/GT were genotyped by PCR as in ref 15, Mtf2Δ/Δ , Jarid2-/- and Eed-/- by WB
c. Report whether the cell lines were tested for mycoplasma contamination.	All lines are routinely tested for mycoplasma contamination
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	no commonly misidentified cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	no animals were used in the study

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	the study did not involve human research participants