Molecular regulation of Polycomb complexes
in mouse embryonic stem cells

Matteo Perino
“Research is to see what everybody else has seen and think what nobody else has thought”

Albert Szent-Györgyi
Molecular regulation of Polycomb complexes
in mouse embryonic stem cells

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 27 Mei 2019
om 10.30 uur precies

door

Matteo Perino
geboren op 5 september 1988
te Rome, Italie
Promotor

Prof. dr. G.J.C. Veenstra

Manuscriptcommissie

Prof. dr. ir. H.G. Stunnenberg

Prof. dr. J.H. Jansen

Prof. dr. J. Gribnau (Erasmus MC)
Index of content

Chapter 1
General introduction .................................................................................................................................... 7

Chapter 2
Chromatin Control of Developmental Dynamics and Plasticity...............................................................17

Chapter 3
MTF2 recruits Polycomb Repressive Complex 2 by helical-shape- selective DNA binding ..................39

Chapter 4
Cooperative recruitment and interactions of PRC2 sub-complexes on chromatin ..............................67

Chapter 5
MTF2 controls meso-endodermal fate specification of mESC ............................................................89

Chapter 6
Summary and Discussion .........................................................................................................................109

Chapter 7
Epilogue ..................................................................................................................................................115
Chapter 1

General introduction
Multicellular organisms owe their success to the ability to specialize specific parts of their body to fulfil a subset of well-defined functions. On the one hand, this requires every cell to rely on every other to survive, on the other it allows for extreme specialization, resulting in increased efficiency and, ultimately, in improved evolutionary fitness. As an explanatory comparison, it is possible to consider colonial Hydrozoans such as the Portuguese man o’ war (Physalia phisalis, Fig 1) which, although resembling a single jellyfish-like organism, is a colony of independent individuals, each specialized in a critical function (hunting, digestion, floating, reproduction, etc.) and mutually depending on the others for survival (Mapstone, 2014). In multicellular organisms, both animals and plants, this specialization is represented by the wide diversity of cell types, tissues and organs that constitute their body. This is achieved through epigenetic modifications, an intricate system of regulatory mechanisms that integrates environmental and genetic cues, required to allow cells carrying virtually the same DNA to only express a well specified set of genes. This system not only drives genetically identical cells to acquire radically different identity and functions, but also does so in such a harmonised fashion to allow the development of sentient beings. Before identity specification, the first need of a eukaryotic cell is to package genomic DNA into the confined nuclear space, and to do so in a way that both avoid tangling of the extremely long DNA molecules and allows for a well organised structure, critical for an accurate control of gene expression. Chromatin is the macromolecular complex of proteins and DNA that composes such a compact and organised structure and, at the same time, allows for its regulation.

The basic element of this structure is the nucleosome, an octamer of histone proteins (H2A, H2B, H3, H4) around which DNA is wrapped (Luger et al., 1997). Biochemical modifications of these two players, DNA and histones, and their reciprocal interactions represent the basic elements of the epigenetic landscape, and set the stage for transcription factors and effector proteins. These can then read, erase, and modify single words or whole sentences, effectively executing the specified program or driving a switch when a change is required, due to either development progression or as response to environmental stimuli.

The rather mobile and unstructured N-terminal histone tails are the most studied, and different modifications— or marks — define the identity of specific regions. Trimethylation of lysine 4 (H3K4me3) is a mark of permissive/active promoters, often accompanied by acetylation of H3 lysine 9 (H3K9Ac). H3 lysine 4 monomethylation (H3K3me1) is present in broad region containing enhancers that, when active, will also acquire acetylation of H3 lysine 27 (H3K27Ac) and the presence of the cofactor responsible of its deposition, p300 (Kebede et al., 2015; Kouzarides, 2007; Lawrence et al., 2016; Perino and Veenstra, 2016). Cells not only have to acquire a specified
identity, but also, and maybe more importantly, they need to keep that identity and suppress alternative fates. Also, a number of parasitic repetitive DNA elements that have accumulated in eukaryotic genomes throughout evolution need to be kept silent to avoid aberrant expression and, in the worse cases, genomic instability. Hence, a number of repressive modifications have evolved. H3K9, as opposed to its “active” acetylated form, can be methylated, and its trimethylated form (H3K9me3) is a mark of stably repressed chromatin, along with trimethylation of H4 lysine 20 (H4K20me3), which play an important role in repressing repetitive elements (Bulut-Karslioglu et al., 2014; van Kruijsbergen et al., 2017; Varshney et al., 2015; Zylicz et al., 2015), together with the methylation of DNA itself, which also play an important role in regulating enhancer activity (Bogdanovic et al., 2016). These topics, together with how they influence, and are influenced by, higher-order chromatin organization will be further described and discussed in Chapter 2. The rest of this chapter will focus on other two repressive machineries, Polycomb Repressive Complex 1 and 2, and their interactions, which will be the focus of Chapter 3, Chapter 4 and Chapter 5.

**Polycomb regulation of development**

Embryonic development precision, robustness and reproducibility rely on waves of key transcription factor expression, often induced by specific concentration of signalling molecules, acting within narrow temporal windows and at specific locations of the developing embryo. These morphogens can induce a variety of different effects, making them a very ductile resource for the embryo. However this property could easily turn into a double edged sword, if the potential to exhibit pleiotropic effects is not controlled through tight regulation of the proteins bridging the signals to the transcription machinery, and the genes that are permissive for transcription. Additionally, during embryogenesis, developmental context and needs can change at an exceedingly fast rate, requiring a very flexible and dynamic regulatory landscape. In multicellular eukaryotes this system is orchestrated by the two Polycomb Repressive Complexes, PRC1 and PRC2, two extremely conserved epigenetic protein complexes that catalyse ubiquitination of H2A lysine 119 (H2AK119u) and methylation of H3 lysine 27 (H3K27me1/2/3) respectively.

**Polycomb repressive complexes**

PRC2 is a complex whose catalytic subunit, either EZH1 or EZH2, forms the core of the complex, along with EED and SUZ12, together with the histone-binding protein RBBP4 or RBBP7, present in a 1:1:1 ratio (Smits et al., 2013). These proteins can then associate with a number of substoichiometric units, giving rise to two PRC2 “flavours”: PRC2.1 can contain of one of the PCL homologs (PHF1, MTF2, PHF19, also named PCL1/2/3 respectively), and either EPOP (C17ORF96) or GM340 (C10ORF12), while PRC2.2 is defined by the presence of JARID2 and AEBP2 (Fig. 2a) (Holoch and Margueron, 2017; van Kruijsbergen et al., 2015). The composition of PRC1 is more complex, owing to the many paralogues that each of its core components has: two RING, five CBX, six PCGF, and three PHC (Di Croce and Helin, 2013). Additionally RYBP or its homologue YAF2 can be included in the core complex, although they are mutually exclusive with CBXs and PHCs, together with other substoichiometric units, leading to the classification of PRC1 into six different complexes (Fig 2b) (Gao et al., 2012).
The Polycomb paradigm

To perform their repressive activity, both complexes need to bind their target regions but, while in *Drosophila* some DNA sequences driving direct recruitment of Polycomb group (PcG) proteins have been described (Simon et al., 1993), in mammals such Polycomb Response Elements (PRES) have remained elusive. Instead, it has been known for long that *Drosophila* PRC1 can bind H3K27me3 via its Pc subunit, homolog of the mammalian CBXs (Cao et al., 2002). This led to a hierarchical model in which PRC2 would initiate the system by depositing H3K27me3 leading to reciprocal interaction and recruitment (Mozzetta et al., 2014; Xu et al., 2010). Furthermore the PRC2 core complex directly interacts with G9a/GLP, a methyltransferase of H3K9, indicating a feedback loop theoretically able to establish an H3K27me3 domain from a single modified histone.

**PRC2 recruitment**

**Chromatin-mediated recruitment**

In *Drosophila* PcG proteins are recruited to PRES by DNA binding proteins such as, among others, Pho and Gaga (Kassis and Brown, 2013). Distant homologues of these proteins do exist in vertebrates, such as YY1 (Shi et al., 1991), c-KROX (Matharu et al., 2010), ZBTB3 (Kumar, 2011), but...
did not retain the recruiting function (Matharu et al., 2010; Shi et al., 1991). PRC2 can also be recruited to the genome via both active and repressive chromatin modifications; EED has been shown to be able to bind H3K27me3 recruiting PRC2 to its own mark, generating a positive feedback loop theoretically able to establish an H3K27me3 domain from a single modified histone tail, since the binding also boosts PRC2 catalytic activity (Margueron et al., 2009; Poepsel et al., 2018). EED can also bind H3K9me3, although without stimulation of catalytic activity, and the PRC2 core complex directly interacts with G9a/GLP, a methyltransferase of H3K9, indicating reciprocal interaction and recruitment (Mozzetta et al., 2014; Xu et al., 2010). Furthermore the scaffolding protein HP1 has been shown to provide another mechanism to bridge H3K27me3 and H3K9me3 at heterochromatic regions, as HP1 dissociate from chromatin and is targeted for proteasomal degradation upon PRC2 knockdown (Boros et al., 2014). Finally JARID2/AEBP2-containing PRC2.2 can feed on the other Polycomb mark, the PRC1-deposited H2AK119u, as discussed above. Some histone modifications can also antagonize PRC2 binding, such as, among others, H3K27Ac, an active mark that competes for the same lysine, H3K4me3 which marks active promoters, and H3K36me3, which decorates the gene body of actively transcribed genes (van Kruijsbergen et al., 2015). H3K4me3 and H3K36me inhibit PRC2 trimethylation of K27 on the same histone tail (Schmitges et al., 2011), but H3K27me3 and H3K4me3 have been found on the same nucleosome in mESCs, on “bivalent” promoters, regulatory regions of genes mostly encoding for lineage-specific transcription factors (Harikumar and Meshorer, 2015; Voigt et al., 2013). This seems to be a configuration defining a primed promoter state, repressed in mESC but available a rapid activation upon exit of pluripotency (Voigt et al., 2012). However, although widely reported in literature for mESC, this primed state is not prevalent in vivo in neither Xenopus (Akkers et al., 2009), nor mouse embryos (Liu et al., 2016), although it can be detected in zebrafish (Vastenhouw et al., 2010) and at low levels in Xenopus (Akkers et al., 2009). H3K36me3, instead, could also function as a docking site for PRC2.1 via the tudor domain of PHF19 (Hunkapiller et al., 2012), which can also bind the H3K36me3 demethylase NO66 (Brien et al., 2012), potentially leading to repression of the targeted genes. More specifically, PHF19 is upregulated upon mESC differentiation (Kloet et al., 2016), and might be involved in dampening the pluripotency network upon initiation of differentiation.

Chromatin-based recruitment, however, could not explain initiation of H3K27me3 when it is most needed for cell fate decisions, since between fertilization and activation of the zygotic genome there is a wave of removal of several epigenetic modifications (Perino and Veenstra, 2016). Additionally, plasmids injected into developing embryos are marked with H3K27me3 when the constructs contain regions, called early nucleation sites, representing genomic locations where H3K27me3 is first deposited in Xenopus tropicalis development (van Heeringen et al., 2014). Remarkably, although not surprisingly given the conservation of Polycomb repression, the same regions are also targeted when transfected into mouse stem cells (mESCs) underscoring high conservation of this machinery across vertebrates at the finest molecular level (van Heeringen et al., 2014). Finally, it has also been shown that de novo, DNA-driven recruitment is required for H3K27me3 maintenance, not only initiation (Laprell et al., 2017). When PRE-containing constructs are introduced in Drosophila they are quickly decorated with H3K27me3 domains, however, upon targeted removal of the PRE from the construct, the modification is diluted, indicating that the self-sustaining PRC2-H3K27me3-EED(PCR2) loop, although highly efficient, is not sufficient for maintenance, and DNA-mediated recruitment is required (Laprell et al., 2017).
RNA-mediated recruitment

A means to provide sequence specificity to a non-DNA binding protein is via RNA intermediates, or scaffolds, and several PRC2 subunits have been shown to have affinity for both (long) non-coding RNA ((l)ncRNA) and nascent mRNA mostly in a specific and non-specific fashion (Brockdorff, 2013; Davidovich and Cech, 2015; Johnson and Straight, 2017). Among others, RNA-mediated PRC2 repression via Xist is required for inactivation of the X-chromosome in female mammalian cells (da Rocha et al., 2014; Zhao et al., 2008), HOTAIR in the repression of the Hox loci (Rinn et al., 2007; Tsai et al., 2010), and both EZH2 and JARID2 have been shown to bind scaffolding RNAs, stabilizing the PRC2 complex (Kaneko et al., 2014a). By contrast, binding to nascent mRNA seems to dampen PRC2 activity and correlates with H3K4me3 and H3K36me3 (Davidovich et al., 2015; Davidovich et al., 2013; Kaneko et al., 2014b; Kaneko et al., 2013). However, although potentially explaining PRC2 recruitment to a few exceptional cases, neither of these examples can explain initiation of Polycomb recruitment genome-wide.

Instructive recruitment to DNA

The open question about initiation of PRC2 recruitment in vertebrates has boosted a considerable amount of research aimed at the description, either bioinformatic or experimental, of a vertebrate PRE acting genome-wide, and of the proteins mediating this interaction (reviewed in (Bauer et al., 2016)). A clear sequence signature of PRC2 targets, especially in mammals, is richness of both CpG dinucleotide and G+C in the context of unmethylated DNA. In mammals CpGs mostly occur in clusters on regulatory regions called CpG islands (CGI), and their methylation, indeed, prevent both PRC1 and PRC2 binding, in the case of the former via inhibiting the recruitment of KDM2B, a CxxC-binding subunit of PRC1 (Bogdanovic et al., 2011; Lee et al., 2006; Long et al., 2013; Lynch et al., 2012). Although the role of CpG islands in PRC2 recruitment has been extensively studied (van Kruijsbergen et al., 2015), the connection between the presence of CGI and PRC2 recruitment is not intuitive. On the one hand, Polycomb targets are strongly enriched for CpGs, and some sequences from both the human and E.coli genome have been reported to recruit PRC2, once depleted of motif for transcriptional activators (Mendenhall et al., 2010). On the other hand, the simple enrichment for CpGs is not sufficient to explain PRC2 recruitment. Non-mammalian vertebrates do not show the strong enrichment of CpGs that are typical of mammalian CGI, yet Polycomb target are extremely conserved and characterized by the lack of DNA methylation. Additionally, in all vertebrates, CGI-containing promoters of housekeeping genes are also devoid of DNA methylation, and yet, never targeted by PRC2 (Bogdanovic and Gomez-Skarmeta, 2014; Perino and Veenstra, 2016; van Kruijsbergen et al., 2015), suggesting the presence of another layer of DNA-encoded information specifying the identity of Polycomb targets. This additional layer of information is underscored by the ability of machine learning algorithms to classify CGIs targeted by Polycomb from those that never receive the H3K27me3 modification (van Heeringen et al., 2014). PRC2-targeted CGIs also show enrichment of specific CpG-containing motifs, but so do untargeted sequences, again underscoring our incomplete understanding of the mechanism.

Contrarily, the search for a DNA-binding protein that could recruit PRC2 has recently seen a leap forward, with the description of the crystal structure of two PCL proteins, PHF1 and MTF2, in complex with a CpG-containing DNA bait (Choi et al., 2017; Li et al., 2017). These structures show that these proteins have a winged-helix domain whose second wing is able to enter the major
groove and directly contact the CpG and, in addition, the DNA phosphate backbone. These findings identify two candidate mediators of DNA-driven PRC2 recruitment and give a full molecular explanation for the requirement of CpGs, while at the same time still leave open the question of what makes the bound CpG different from the others.

We independently identified MTF2 as a DNA-binding PRC2 recruiter and described what determines its specificity for Polycomb targets, namely the specific 3D-shape of the DNA surrounding the bound CpG. This combination of CpG and shape-determining surrounding sequence indeed form a PRE which can be bound by MTF2 that is therefore able to initiate PRC2 recruitment at the vast majority of Polycomb target in mESC.

These findings (Chapter 3), together with a description of the interaction between PRC2.1, PRC2.2 and PRC1 in mESC (Chapter 4), and the role of MTF2 during cell differentiation (Chapter 5) are the main body of work discussed in this thesis.
References


Chapter 2

Chromatin Control of Developmental Dynamics and Plasticity

Matteo Perino and Gert Jan C. Veenstra


MP wrote the review, with help from GJCV
Abstract

Chromatin structure is intimately connected with gene expression and cell identity. Here we review recent advances in the field and discuss how establishment of cell identity during development is accompanied by large-scale remodeling of the epigenetic landscape and how this remodeling drives and supports lineage specification and maintenance. We discuss maternal control of the early embryonic epigenetic landscape, selective usage of enhancer clusters via 3D chromatin contacts leading to activation of transcription factor networks, and conserved regulation of developmental pathways by specific DNA demethylation of key regulatory regions. Together, these processes establish an epigenetic framework regulating different phases of embryonic development.

Introduction

Multicellular animals owe their complexity to their capacity to produce and maintain a multitude of different cell types that share virtually the same genomic DNA. Such complexity requires tight regulation of gene expression to unambiguously specify and constrain the developmental paths taken by cells in the embryo. Signaling molecules and networks of expressed transcription factors (TFs) cooperate in the context of genomic chromatin state to control cellular fate. Chromatin is the macromolecular complex of genomic DNA and protein that packages and condenses the DNA but also regulates its biochemical activities. Because of this regulation of genomic activity, chromatin can be considered a developmental cell type-specific filter of genomic sequence information that determines which genes are transcribed into RNA (Figure 1, left panel). Because of its capacity to act as a filter at various regulatory levels (Figure 1, middle panel), chromatin plays a critical role in cell identity and the embryonic developmental program.

In the classical description of chromosome architecture, chromatin is depicted as organized into active (A) and inactive (B) compartments, each with large topologically associating domains (TADs). Computational analysis applying hidden Markov models to large datasets has shown a more fine-grained distinction of chromatin states based on co-occurring epigenetic modifications (Ernst and Kellis, 2012). These different chromatin states demarcate distinct genomic elements such as promoters, enhancers, insulators, and gene bodies, and reflect their biochemical activities within the context of larger “genomic neighborhoods,” the TADs (Figure 1, middle panel). TADs are partitioned into smaller nested loops, whose function is to define the regulatory space within which genes can interact with their regulatory elements. Loop boundaries are determined by strong sequence-specific CTCF (CCCTC-binding factor) binding to DNA, generally providing transcriptional isolation from neighboring loops (for reviews on TADs and looping, see Ea et al., 2015; Pombo and Dillon, 2015). The relevance of TADs for gene regulation and development is exemplified by CRISPR/Cas9-mediated rearrangements of the TAD spanning the mouse Wnt6/Hhi/Epha4/Pax3 locus. Targeting the CTCF sites important for the TAD caused new promoter-enhancer interactions and pathogenic misexpression of genes, phenocopying human limb malformation syndromes (Lupianez et al., 2015).

Within TADs, chromatin structure is determined by the density, positioning, and composition of nucleosomes. Nucleosomes are composed of core histone proteins H2A, H2B, H3, and H4 (each present twice) and represent the basic unit of chromatin around which the DNA is wrapped.
Given their close interaction with DNA, nucleosomes, with their modifications and variant subunits, can control the accessibility of DNA to DNA-binding proteins (Figure 1, middle panel). For example, gene expression can be modulated by varying the density of nucleosomes positioned in close proximity to regulatory regions (Maehara and Ohkawa, 2016). Nucleosome composition also influences the tightness of DNA wrapping, therefore playing a role in chromatin accessibility. Histones in the canonical H2A-H2B/ H3-H4 nucleosome can be replaced by histone variants that locally alter chromatin compaction, influencing transcription. Histone variants have been shown to be involved in the regulation of both nucleosome stability at highly transcribed regions (H2A.Z, H3.3) and chromatin compaction, for example in X chromosome inactivation (macroH2A) and sperm cell DNA packing (H3.4). For detailed reviews of histone variants and their roles in chromatin structure and transcription, the reader is referred to Kamakaka and Biggins (2005) and Weber and Henikoff (2014).

In addition to the steric accessibility of DNA sequence, the state of chromatin is influenced by regulatory sequences that, along with TFs and coactivator complexes, can recruit chromatin-modifying enzymes. These enzymes deposit post-translational modifications (PTMs) on the histones (Figure 1, right panel). The most well-studied and relevant for transcriptional regulation are PTMs occurring on lysines of the histone N-terminal tails, particularly of histone H3 (Ausio et al., 1989). H3K4me3 (trimethylation of H3 lysine 4), for example, is present at active promoters, H3K4me1 (monomethylation of H3 lysine 4) and the p300-coactivator that mediates H3K27ac (acetylation of H3 lysine 27) decorate enhancers, while H3K27me3 (trimethylation of H3 lysine 27) is found at Polycomb target genes. A simplified overview of well-studied epigenetic modifications and chromosomal proteins is presented in the right-hand panel of Figure 1 (for a detailed review of histone modifications, see Lawrence et al., 2016). The relevance of these modifications is illustrated by the extensive epigenome maps of multilineage differentiation of human embryonic stem cells (ESCs) (Xie et al., 2013) and of early frog embryos (Hontelez et al., 2015), which both show that dynamic changes in histone methylation during determination of cell identities are tied to DNA methylation, another key epigenetic feature.

DNA methylation is a modification of the DNA itself, and can influence chromatin organization and transcription by recruiting methyl-CpG binding proteins which, in turn, can recruit histone deacetylase complexes, leading to repression (Bogdanovic and Gomez-Skarmeta, 2014). DNA methylation is a feature shared by all vertebrates and by many, but not all, non-vertebrate organisms. In vertebrates, up to 90% of all CG dinucleotides, which tend to be interspersed across the genome, are methylated, with the exception of unmethylated islands (Bogdanovic et al., 2016; Long et al., 2013). These regions are relatively rich in CG dinucleotides, especially in mammals (CpG islands), and often overlap with regulatory sequences (Bogdanovic and Gomez-Skarmeta, 2014).

Developing embryos provide a prime system to study how these systems interact to orchestrate lineage commitment and cell-identity specification in a physiological context. In this review, we focus on the roles of these epigenetic mechanisms in cell identity with some focus on embryonic development in vertebrates. The molecular mechanisms involving DNA (de) methylation and histone modifications have been reviewed elsewhere (Du et al., 2015; Pastor et al., 2013; Smith and Meissner, 2013; Speranzini et al., 2016).
Methylation of K27 of H3.3, a mark set by the Polycomb PRC2 complex, is also essential and be methylated at H3K4, ZGA is impaired in the paternal pronucleus (Aoshima et al., 2015). Assembly, and the first cell division (Inoue and Zhang, 2014; Lin et al., 2014). Also, if H3.3 cannot involves the chaperone Hira, which in the zygote is required for rRNA transcription, nuclear pore (Torres-Padilla et al., 2006; van der Heijden et al., 2005). Incorporation of this histone variant which are removed and replaced by nucleosomes containing the maternal histone variant H3.3 In mouse, ZGA occurs within a cell cycle after fertilization. Sperm DNA comes with protamines, 2015; Lee et al., 2014b; Li et al., 2013; Paranjpe and Veenstra, 2015; Tadros and Lipshitz, 2009). 

The earliest development following fertilization is under the full control of maternally deposited factors (proteins and RNA) in the egg or oocyte. This is particularly prominent in non-mammalian species developing outside the mother, but in all animals, including mammals, the zygotic genome is initially not transcribed until zygotic genome activation (ZGA), which marks the beginning of the transition to zygotic control of development (MZT, maternal-zygotic transition) (Biechele et al., 2015; Lee et al., 2014b; Li et al., 2013; Paranjpe and Veenstra, 2015; Tadros and Lipshitz, 2009). In mouse, ZGA occurs within a cell cycle after fertilization. Sperm DNA comes with protamines, which are removed and replaced by nucleosomes containing the maternal histone variant H3.3 (Torres-Padilla et al., 2006; van der Heijden et al., 2005). Incorporation of this histone variant involves the chaperone Hira, which in the zygote is required for rRNA transcription, nuclear pore assembly, and the first cell division (Inoue and Zhang, 2014; Lin et al., 2014). Also, if H3.3 cannot be methylated at H3K4, ZGA is impaired in the paternal pronucleus (Aoshima et al., 2015). Methylation of K27 of H3.3, a mark set by the Polycomb PRC2 complex, is also essential and

---

**Maternal-Zygotic Transition**

The earliest development following fertilization is under the full control of maternally deposited factors (proteins and RNA) in the egg or oocyte. This is particularly prominent in non-mammalian species developing outside the mother, but in all animals, including mammals, the zygotic genome is initially not transcribed until zygotic genome activation (ZGA), which marks the beginning of the transition to zygotic control of development (MZT, maternal-zygotic transition) (Biechele et al., 2015; Lee et al., 2014b; Li et al., 2013; Paranjpe and Veenstra, 2015; Tadros and Lipshitz, 2009).

In mouse, ZGA occurs within a cell cycle after fertilization. Sperm DNA comes with protamines, which are removed and replaced by nucleosomes containing the maternal histone variant H3.3 (Torres-Padilla et al., 2006; van der Heijden et al., 2005). Incorporation of this histone variant involves the chaperone Hira, which in the zygote is required for rRNA transcription, nuclear pore assembly, and the first cell division (Inoue and Zhang, 2014; Lin et al., 2014). Also, if H3.3 cannot be methylated at H3K4, ZGA is impaired in the paternal pronucleus (Aoshima et al., 2015). Methylation of K27 of H3.3, a mark set by the Polycomb PRC2 complex, is also essential and
required for heterochromatin formation (Santenard et al., 2010). However, some histone modifications are inherited from sperm, especially at developmental loci (Hammoud et al., 2009). Erasure of histone modifications is also important for preimplantation development, as mouse zygotes lacking maternally derived LSD1 (KDM1A), a histone demethylase for H3K4 and H3K9, exhibit changed histone methylation profiles and do not develop beyond the 2-cell stage (Ancelin et al., 2016). This shows that ZGA and early development require writing and erasing histone modifications that are associated with both activation and repression of transcription.

In the frog Xenopus, high-resolution time-course RNA sequencing (RNA-seq) has shown that transcription already starts during the early cleavage stages and before the midblastula stage, which is the stage traditionally associated with ZGA (Collart et al., 2014; Owens et al., 2016). The epigenetic landscape responsible for zygotic transcription is broadly determined by maternal factors, as identified by means of chromatin immunoprecipitation (ChIP) sequencing in transcriptionally inhibited embryos (Hontelez et al., 2015). In these embryos the deposition of H3K4me3 and H3K27me3 is maternally defined in the context of unmethylated DNA islands, and this control extends far beyond ZGA well into post-gastrulation development. Many of the promoters that become active during neurulation exhibit both a maternally dictated permissive chromatin state and a lack of DNA methylation (Figure 2). This shows that embryonic transcription is deeply influenced by maternal factors and that the MZT is a protracted process. On the other hand, recruitment of the p300 coactivator to distal enhancers is mostly governed by newly expressed factors, showing that, mechanistically, the MZT occurs at different times at different regulatory elements (Hontelez et al., 2015).

ZGA also involves changes in nucleosome occupancy at the transcription start site (TSS). It is commonly observed that regularly positioned nucleosomes extend into the gene from the promoter. In zebrafish, these arrays form during ZGA and involve the promoter of immediately transcribed genes but also those whose expression will occur later in development (Haberle et al., 2014; Zhang et al., 2014). The formation of these ordered arrays also correlates tightly with gene expression, and a high degree of promoter organization at ZGA is a better predictor of the future intensity of expression than the H3K4me3 level at the same stage (Zhang et al., 2014). The formation of arrays of nucleosomes at ZGA also marks a switch in the TSS, with maternal transcripts using different start sites than embryonic transcripts, as determined by base-pair resolution CAGE (cap analysis gene expression) (Figure 2). The underlying molecular logic of start-site selection is that promoters transcribed in oocytes feature an AT-rich sequence ("W box," resembling a TATA box) at >30 bp upstream of the TSS, while zygotic promoters have a C/G-rich sequence and a sharp C/G-A/T boundary ~50 bp downstream of the TSS, with the TSS defined by the position of the +1 nucleosome. These differing start-site “grammars” were confirmed by deletion analysis in oocytes and embryos (Haberle et al., 2014).

The onset of zygotic transcription naturally also requires TF activity. In Drosophila an important role is played by the zinc finger protein Zelda (Boija and Mannervik, 2015; Harrison and Eisen, 2015). Zelda is expressed prior to ZGA and is necessary for ZGA to proceed properly. Zelda-bound regions are kept nucleosome free during ZGA, whereas intense chromatinization is observed in Zelda knockout embryos (Sun et al., 2015), suggesting that Zelda acts as a “placeholder” for subsequent transcription (Harrison et al., 2011). No ortholog of Zelda is known in vertebrates, although the pluripotency TFs Oct4 and Sox2 may play an analogous crucial role in vertebrates.
zebrafish, Oct4, Nanog, and the Sox2 ortholog SoxB1 are the most highly translated TFs before ZGA and are required for activation of most zygotic genes, linking ZGA to pluripotency in this species (Lee et al., 2013; Leichsenring et al., 2013). Interestingly, among the transcripts activated by the pluripotency TFs is microRNA miR-430, which is important for the clearing of maternal mRNA. It is not known to what extent the pluripotency TFs can open up repressive chromatin at ZGA by themselves or whether they require auxiliary maternal factors. The pluripotency TFs, however, have been reported to bind to nucleosomal DNA both in vitro and in vivo, in line with a role as pioneer factors that can open up naive chromatin (Oldfield et al., 2014; Soufi et al., 2015) (see also below). As Oct4 is a maternal factor not only in fish but also in mammals, it could potentially play a conserved role in ZGA. The multisubunit CCAAT-binding NF-Y transcription factor can facilitate Oct4 binding (Oldfield et al., 2014; Soufi et al., 2015). Interestingly, early accessible promoters are enriched for the binding motif of NF-Y, and small interfering RNA-mediated NF-Y knockdown results in impaired generation of DNase I hypersensitive sites and reduced expression of many genes in 2-cell mouse embryos, suggesting a role for this TF complex in mouse ZGA (Lu et al., 2016).

The maternal-zygotic dichotomy is also reflected by the nature of the proteins regulated by the two mechanisms (Heyn et al., 2014). In fact, while the orthologs of maternally provided gene products are similar from Drosophila to mouse, this is much less the case for the first transcribed embryonic genes (Heyn et al., 2014). This is in line with the hourglass model of gene expression between species (Domazet-Loso and Tautz, 2010; Kalinka et al., 2010). Also, the genes encoding maternally provided proteins are larger and evolutionarily older compared with the first transcribed zygotic genes (Heyn et al., 2014). An evolutionarily conserved regulation is evident also at the level of developmental enhancers. A set of developmental enhancers that drive gene expression during the vertebrate phylotypic stage in fish, frogs, and mice are demethylated by TET enzymes during the phylotypic stage. Without Tet-mediated DNA demethylation, the chromatin accessibility of phylotypic enhancers is reduced and gene expression of zebrafish Tet1/2/3 triple-knockdown embryos is severely perturbed (Bogdanovic et al., 2016). Concordantly, Tet1/2/3 triple-knockout mouse ESCs differentiate poorly and do not support embryonic development (Dawlaty et al., 2014). These results identify a deeply conserved role for regulated DNA demethylation in zygotic body plan formation and embryonic development.

Overall, these findings show how the chromatin state is changed during the maternal-zygotic transition and subsequent development, affecting positioned nucleosomal arrays and the modification of histones. It is not always clear to what extent changes in chromatin state relate to transcriptional activity. However, in this case acquisition of the promoter mark H3K4me3 is required for ZGA (Aoshima et al., 2015) and is independent of transcription (Hontelez et al., 2015) or pre-initiation complex assembly (Haberle et al., 2014). This suggests a scenario where H3K4me3 forms a permissive landscape of marked promoters on which binding of the transcription initiation factor TFID is stabilized (Vermeulen et al., 2007). Together, these findings depict the activation of the zygotic genome as a continuous, gradual process, causing the MZT to span a range of developmental stages with distinct molecular transitions at developmental promoters and enhancers.
Figure 2. Chromatin state and the vertebrate maternal-zygotic transition. The transcription start site of maternal-zygotic genes is determined by a TATA-like sequence in oocytes and by a nucleosomal array in ZGA-stage embryos. H3K4 and H3K27 methylation is controlled by maternal factors in the absence of DNA methylation, and is temporally uncoupled from developmental stage-specific activation (at the stage of ZGA or later). On methylated promoters H3K4 methylation is associated with transcriptional activation. p300 co-activator recruitment to enhancers generally depends on zygotic transcription, whereas phylotypic enhancers also require demethylation at the phylotypic stage. Round spheres with four quadrants depict nucleosomes (green H3K4me3, red H3K27me3-modified); lollipops depict presence (black) or absence (white) of DNA methylation. Mammalian embryos are subject to a global wave of demethylation (not depicted) between fertilization and the blastocyst stage.

Pluripotency and Lineage Specification

Once the activation of the zygotic genome has occurred in the embryo, a cellular state of pluripotency needs to be established before the onset of gastrulation. In mammals, due to the need to produce extraembryonic tissues, an additional step is required. The first lineage choice takes place in the totipotent mammalian zygote and gives rise to the trophoectoderm (TE), which will produce the extraembryonic tissues, and the pluripotent inner cell mass (ICM), from which all the cells of the embryo originate. This first lineage commitment is marked by a strong reduction of histone mobility within the nucleus, as measured by quantification of fluorescence recovery after photobleaching (Boskovic et al., 2014). As the mouse embryo develops from the 2-cell stage (when ZGA occurs) to the 8-cell stage, the H2AZ, H3.1, and H3.2 turnover rate in the nucleosomes drops strikingly. Also, when looking at the derived lineages, ICM cells show higher chromatin mobility when compared with TE. This was further supported by electron microscopy experiments demonstrating the absence of heterochromatin foci and an overall looser chromatin structure in
2-cell embryos compared with the 8-cell stage (Boskovic et al., 2014), in line with the importance of chromatin compaction during development (Ahmed et al., 2010). Although not supporting an exclusive role of chromatin structure in inducing or maintaining cellular potency, these findings clearly underscore its involvement in early mammalian development.

**Chromatin Accessibility**

The compaction and resulting 3D structure of chromatin has two functions: fitting the genome in the cell nucleus and regulating gene expression. In a study comparing the histone sequence of 160 sequenced eukaryotic genomes ranging in size from 8 to 5,600 Mbp (Macadangdang et al., 2014), it was found that the canonical histone H2A has evolved to contain more arginines at specific positions of the N-terminal tail to allow a tighter packing of the DNA. A reduction of the nuclear diameter is observed when arginine-rich H2A is introduced in organisms with a smaller genome, and a chromatin relaxation when critical residues are mutated. Although the loss of these key arginine residues is not lethal, at least not in yeast, the wild-type yeast outcompeted the mutated counterpart, showing the potential evolutionary advantage provided by this compaction strategy (Macadangdang et al., 2014).

Because of this compaction, chromatin needs to open up at regulatory regions to facilitate transcription factor-DNA interactions. This can be observed in the gains and losses of DNase I and transposase hypersensitive sites during both early preimplantation development and ESC differentiation to different lineages (Lu et al., 2016; Stergachis et al., 2013; Wu et al., 2016). Profiling of chromatin accessibility of preimplantation mouse embryos by DNase I sequencing has shown that chromatin accessibility at regulatory regions increases progressively during development, with the proximal promoters being the earliest accessible locations, whereas distal regulatory regions such as enhancers become available later in development (Lu et al., 2016).

Upon lineage commitment of ESCs, differentiated cells share accessible regions with the ESC they are derived from, but this regulatory similarity decreases as cells mature. When cells commit to a specific lineage, the repertoire of accessible regulatory regions containing motifs for TFs specific to that lineage expands, whereas it shrinks for elements bound by TFs of other lineages (Stergachis et al., 2013). How then does chromatin accessibility change upon cell-lineage commitment? Some TFs, the pioneer factors, have the ability to bind their DNA motifs even in relatively compact chromatin (Adam and Fuchs, 2016; Zaret and Mango, 2016). This can facilitate the recruitment of chromatin-remodeling enzymes and the binding of other TFs, effectively opening up chromatin. This can be distinguished from factors that will only bind their motifs if they are accessible already (Sherwood et al., 2014). In hair follicle stem cells, for example, the expression of the SOX9 transcription factor can open up enhancer clusters (also referred to as super-enhancers) and protect them against Polycomb-mediated repression (Adam et al., 2015). These enhancer clusters in turn activate lineage-specific genes. The PHA-4/FoxA transcription factor can even recruit RNA polymerase II directly to the promoter in Caenorhabditis elegans embryos, creating a poised state prior to creating chromatin accessibility (Hsu et al., 2015). Even pioneer factors, however, bind only a fraction of their DNA motifs in the genome, suggesting that their binding benefits from cooperative interactions with other TFs, which is a feature commonly observed among TFs. Indeed, the CCAAT-binding NF-Y transcription factor complex facilitates enhanced binding of the core pluripotency TFs Oct4, Sox2, and Klf4 (Oldfield et al., 2014; Soufi et al., 2015). Also, the
pluripotency TFs themselves show a differential ability to target inactive loci. Based on single-cell single-molecule imaging in mouse ESCs, Sox2 finds its targets first, using a mixture of 3D diffusion and sliding on exposed DNA to find its motif. This binding is greatly stabilized by Oct4, resulting in a heterodimer stable enough to nucleate enhancerosome assembly (Chen et al., 2014). A similar hierarchy has been suggested in the transition from naive (ESC) to primed pluripotency (epiblastlike cells, EpiLCs) (Buecker et al., 2014). Both naive and primed stem cells express Sox2 and Oct4; however, Otx2 is required to recruit Oct4 to EpiLC-specific enhancers. In addition, binding of some factors is restricted by DNA methylation. For example, binding of NRF1 to regulatory elements in mouse ESCs is facilitated by a lack of methylation of two CG dinucleotides in its motif (Domccke et al., 2015). The hypomethylated state of the NRF1 motif was found to depend on surrounding TF motifs rather than on NRF1 binding itself. Interestingly, NRF1 has pioneer factor properties in the sense that it can establish DNase I hypersensitive sites in the unmethylated condition, although this is dependent on other TFs mediating a hypomethylated state (Domccke et al., 2015). These data provide fascinating insights into the mechanisms that control chromatin accessibility at regulatory regions, even if they call for more qualified statements as to what constitutes a pioneer factor. Moreover, they raise new questions regarding the relationships between the intrinsic properties of TFs to bind motifs in nucleosomal DNA, the repertoire and interactions of expressed TFs, and the gene-regulatory circuitry underlying the behavior and relative abundance of these TFs. All these aspects could contribute to what functionally defines pioneer behavior at dynamically accessible regions during lineage commitment.

**Chromatin Topology and Subnuclear Localization**

Compaction of large genomes not only necessitates local accessibility of relevant regulatory regions, it also requires robust regulation of chromosome topology to facilitate long-range promoter-enhancer interactions that are important for development. To understand the role of the 3D architecture of chromosomes in cell identity and gene expression, researchers have used chromatin conformation capture (3C)-related techniques such as 4C and Hi-C; these approaches crosslink chromosomal interactions in the nucleus followed by DNA fragment ligation and sequencing (Barutcu et al., 2016). Deep sequencing of Hi-C libraries of stem cells has been used to describe 3D chromatin architecture dynamics during differentiation in vitro (Dixon et al., 2015). Indeed, while the local contacts between regulatory regions are rearranged during development, the high-level TADs mostly maintain the same structure, despite an extensive switch between A (active) and B (not active) chromosomal compartments. This has been independently confirmed by Cohesin ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) and CTCF ChIP sequencing on both mouse and human ESCs (Dowen et al., 2014; Ji et al., 2016). Both of these studies show how a nested structure of Cohesin-bound CTCF dimers determines and limits the range of action of both active and Polycomb-repressed regulatory regions to the smallest nested loop; only within TADs do regulatory interactions change between cell types, as seen when comparing naive and primed states of pluripotency (Dowen et al., 2014; Ji et al., 2016).

Polycomb complexes play a role in 3D chromatin organization in several organisms. In Drosophila, super-resolution microscopy experiments showed that the Polycomb protein Ph is distributed in clusters of different sizes, and that these nuclear clusters are sites of chromatin interactions (Wani et al., 2016). 4C-sequencing analysis of the Bithorax Hox gene locus indeed shows that its 3D
structure is dependent on Ph binding. Also, Ph-mediated Polycomb clusters constitute functional units, and their disruption by dominant-negative Ph expression alters both chromatin structure and gene expression (Wani et al., 2016). In mammalian ESCs, the role of Polycomb in determining 3D chromatin structure goes beyond the local genomic neighborhood, as shown by long-range Polycomb-dependent interactions that involve the Hox genes and which extend across TADs and chromosomes. The biological significance and function of these Polycomb interactions is not fully understood, but they are characteristics of primed but not ground-state pluripotency (Joshi et al., 2015; Schoenfelder et al., 2015).

Although chromosome topology is generally conserved between species, a fascinating exception is provided by the HoxA and HoxD loci. In vertebrates, both of these gene clusters are split into two TADs that separate two distinct sets of enhancer elements, one on each side (referred to as anterior and posterior). In the chordate amphioxus, however, the Hox locus forms a single 3D structure under the almost exclusive control of enhancers in the anterior region (Acemel et al., 2016). This suggests that the vertebrate Hox clusters evolved their bipartite TAD structure by expanding and separating their regulatory inputs, allowing Hox genes to pattern distinct parts of the body, including limbs and digits.

Studies in C. elegans (Towbin et al., 2012) have described that the two worm H3K9 methyl transferases are required for targeting heterochromatic regions to the nuclear lamina. The proposed model involves a two-step process, in which Met-2 (worm homolog of mammalian SETDB1) deposits H3K9me1/2 on cytoplasmic histone H3, which is then trimethylated in the nucleus by SET-25 (worm homolog of mammalian Suv39h1/2) and targeted to the nuclear envelope (Towbin et al., 2012). H3K9 methylation, however, appears not to be the only repressive mechanism of these regions. Indeed, in both worm (Towbin et al., 2012) and ESC (Therizols et al., 2014), release of heterochromatic DNA from the nuclear lamina is not enough to induce transcription, which in ESCs requires the targeting of a strong transcriptional activator (Therizols et al., 2014).

Therefore it is clear that the 3D organization of the genome is highly regulated and critical for development. Limiting the contacts between genes and regulatory sequences by enclosing them in insulated genomic neighborhood contributes to the gene-regulatory specificity of developmental genes, and TADs and the enhancer landscapes they contain are therefore generally well conserved.

**Histone Modifications and Variants**

Within TADs, the regulatory landscape is highly dynamic during cell-lineage specification. This is linked to inductive signals if the cells are in a state of competence to respond. For example, Activin/Nodal signaling shapes the H3K4me3 landscape at gene promoters in human ESCs via the ML2/KTM2B COMPASS complex (Bertero et al., 2015). In this case the Activin/Nodal effectors SMAD2/3 cooperate with the core pluripotency factor NANOG to recruit DPY30-COMPASS to developmental genes. The Erk1/2 pathway is also involved in the regulation of pluripotency and lineage priming in mouse ESCs (Tee et al., 2014), being downstream of fibroblast growth factor (FGF)/Mek and LIF/ Stat3. Erk1-2 phosphorylation of RNA polymerase II (RNAPII) at serine 5 is indeed necessary for RNAPII binding to the bivalent promoters of developmental genes, marked by both H3K4 and H3K27 trimethylation (Tee et al., 2014). Interestingly, preimplantation- like
ground-state ESCs grown in the presence of Erk and GSK3b inhibitors (2i medium) show a lower intensity of H3K27me3 at loci of developmental genes when compared with primed ESCs grown in the presence of serum (Marks et al., 2012). These data illustrate how extracellular cues, expressed developmental TFs, and chromatin state can cooperate to change gene expression and induce cell-lineage commitment. These events, however, cannot happen at any point of the cell cycle. It has recently been shown that, at least in human ESCs, H3K4me3 deposition on bivalent genes, decorated with both permissive H3K4me3 and repressive H3K27me3 marks, is very dynamic. While H3K27me3 seems to be fairly stable during the ESC cell cycle, H3K4 residues at the same promoters undergo a massive demethylation during the G1 phase, a dynamic pattern that disappears when cells differentiate and the genes acquire a stable signature (Grandy et al., 2015). This H3K4me3 dynamic may prevent leaky expression of differentiation-inducing factors from permissive promoters, allowing the cell to undergo a self-renewal cycle (Grandy et al., 2015). A reshaping of the histone PTM landscape is also observed during primordial germ cell (PGC) specification. PGCs are a special example because they reactivate the pluripotency network, after having been specified as early mesoderm. The epigenome of these cells is subject to intense H3K9 demethylation, reduced levels of DNA methylation, and increased levels of the Polycomb mark H3K27me3, before it acquires the definitive oogenic/spermatogenic landscape (Kurimoto et al., 2015).

Histone H3 arginine methylation can affect the residency time of TFs on chromatin. Sox2 binds more stably to chromatin in two blastomers of the 4-cell stage mouse embryo, marking the cells that contribute most descendants to the pluripotent ICM of the blastocyst (White et al., 2016). Intriguingly, this differential Sox2 binding is induced by Carm1-dependent H3R26 methylation and subsequent chromatin relaxation, implying a regulatory role for this epigenetic modification in pluripotent lineage choice in the mouse embryo.

Along with histone PTMs, incorporation of replication-independent histone variants into chromatin also plays a role in defining developmental chromatin state. As discussed above, histone variant H3.3 plays a role in ZGA in mouse zygotes. Recent work has also shed light on the dual role of this histone variant in later mouse development and in stem cell differentiation in vitro. First, analysis of the embryonic lethal phenotype of H3.3 homozygous null mutants has uncovered the role of this variant in maintaining genome integrity at telomeres, centromeres, and pericentromeric heterochromatin (Jang et al., 2015). The absence of H3.3 has no major effect on gene expression in embryonic day 10.5 embryos, but leads to DNA damage, mitotic defects, and activation of the p53 pathway. Second, two studies confirmed the presence of H3.3 at both active and repressed loci, and the presence of H3.3 signifies nucleosome turnover at regulatory regions. The EP400 protein stimulated both transcription and deposition of H3.3 at hyperacetylated nucleosomes (Pradhan et al., 2016). H3.3 deposition by the Hira complex also facilitates PRC2 recruitment to developmental loci, showing an involvement in Polycomb-mediated repression (Banaszynski et al., 2013). Interestingly this last study also makes use of H3.3 KO stem cells, whose viability, together with the post-blastocyst lethality in vivo (Jang et al., 2015), suggest an essential role of H3.3 in post-blastocyst stages rather than stem cell maintenance.

Finally, nucleosome compaction can also affect gene expression. Linker histone H1 is known to regulate nucleosome compaction and repress nucleosome unwrapping, thereby limiting the accessibility of TF motifs within the nucleosome. This repressive potential is abolished by
acetylation of H3K56, which interferes with the H1-nucleosome interaction, re-establishing TF-binding site accessibility (Bernier et al., 2015). In addition, H1 mediates crosstalk with other repressive mechanisms such as Su(var)3–9-dependent H3K9 methylation in Drosophila (Boija and Mannervik, 2015). In ESCs it interferes with H3K4 methylation at imprinted promoters by SETD7 (SET7/9) and stimulates DNA methylation by DNMT1/3B (Yang et al., 2013).

**DNA Methylation**

Compared with the histone PTM landscape, chromosome topology is relatively stable and DNA methylation also is generally less dynamic. However, during mammalian development both the zygote and PGCs show global DNA demethylation waves during their development (Lee et al., 2014a). Bisulfite sequencing in human embryos has shown, as in mouse, an important role in repressing transcription from repetitive elements, especially the evolutionarily younger and more active transposons (Lee et al., 2014a). The DNA demethylation pattern, however, is somewhat different between the two species. Both human and mouse have a burst of DNA demethylation during the zygotic stage, but while humans only show subtle further demethylation until the morula stage (Guo et al., 2014), in mouse gradual demethylation lasts until the blastocyst is formed (Marks et al., 2012). There are also differences between human and mouse ESCs; whereas mouse ESCs are viable without DNA methylation, in human ESCs the complete loss of DNMT1 (DNA methyl transferase 1) activity causes rapid global DNA demethylation and cell death (Liao et al., 2015).

Locally, DNA methylation controls gene expression by regulating lineage-specific enhancer usage, as shown by enrichment of regulatory regions in differentially methylated regions (DMRs) in vertebrate embryos and tissues (Bogdanovic et al., 2016; Lee et al., 2015; Lowdon et al., 2014). Epigenetic profiling of different cell populations in human surface ectoderm shows how more closely related cell types share a more similar DMR pool, underscoring the role of DNA methylation in cell differentiation and the existence of lineage-specific methylation of regulatory regions (Lowdon et al., 2014). This is different from the phylotypic enhancer DMRs (discussed above), which undergo Tet-dependent DNA demethylation in adult mouse tissues from all of the three germ layers (Bogdanovic et al., 2016). DNA methylation plays a dual role during development: on one hand its stability is required to ensure constant repression of repetitive elements, and on the other hand local DNA methylation dynamics are required to ensure a permissive environment for developmental enhancers and regulatory regions to be able to exert their functions.

**Enhancer Usage**

Cell-type-specific gene activation via TF binding to enhancers is a key mechanism for establishing cellular identity. In ESCs a group of TFs, Nanog, Sox2, Oct4, Klf4, and Essrb, establish and maintain an autoregulatory transcriptional network in which the TFs mutually stimulate their expression via a specific set of enhancers (Whyte et al., 2013). These enhancers occur in close-proximity clusters (also referred to as super-enhancers), show a higher density of TF-binding sites, a relatively high level of H3K27 acetylation, and binding of the coactivator complex Mediator (Whyte et al., 2013). They are not fundamentally different from other enhancers in how they activate transcription (Pott and Lieb, 2015), but are relatively sensitive to changes in expression of TFs and coactivators; transcription activated by pluripotency super-enhancers show the earliest and most profound
downregulation upon small hairpin RNA-mediated knockdown of Oct4 and Mediator complex subunits (Whyte et al., 2013).

Changes in cell identity are reflected in changes in enhancer usage. Although half of the enhancers active in ground-state mouse ESCs lose activity during the differentiation to primed epiblast-like stem cells (EpiSCs), the transcription profiles of the two cell types are almost identical (Factor et al., 2014). Loss of enhancers that are only active in the ground state is compensated through the activity of distinct sets of regulatory regions, often called “poised” or “seed” enhancers. These seed enhancers are characterized by low levels of H3K27ac and/or H3K4me1 in ground-state mouse ESCs and only become active upon priming of ESCs to EpiSCs (Factor et al., 2014; Tsankov et al., 2015). This is accompanied with acquisition of additional enhancer activity in the close vicinity, turning the “seed” of an ESC poised enhancer into an EpiSC enhancer cluster. Indeed, more than 20% of poised enhancers present in mouse ESCs turn into super-enhancers during differentiation (Tsankov et al., 2015). These changes in enhancer usage require a topology that allows the new set of enhancers to interact with their target promoters (Figure 3). The role of enhancer clusters is not limited to early lineage commitment events, as they are also associated with the expression of key cell-identity genes in B cells, myotubes, T helper lymphocytes, and macrophages (Whyte et al., 2013).

A particularly intriguing question is how enhancer usage and chromatin state define cellular states of competence, specification, and determination, and how they affect the cell’s response to extracellular signaling. External stimuli such as morphogens and signaling molecules can shape the epigenetic landscape and transcriptional profile during the endoderm specification process, being able to homogeneously direct a pool of competent cells toward specific cell fates. Timely exposure to the bone morphogenetic protein, Wnt, and FGF signaling pathways is indeed sufficient to produce pure populations of fully differentiated pancreas and liver cells (Loh et al., 2014). For these results to be achieved in vitro, however, exposure to signaling molecules needs to occur at a specific time points, in the correct order, and at the right concentration, mirroring development in vivo. During development cells are guided through their differentiation program by signaling molecules, in a process involving first a loss of competence accompanying lineage specification, followed by final cell-fate determination. During this process chromatin modifications at enhancers reflect the loss of potency, and terminal differentiation is characterized by deposition of repressive heterochromatin marks at developmental genes (Loh et al., 2014). Such a loss of potency by lineage priming can occur very early, as shown in the case of myeloid progenitors (Paul et al., 2015). Several subpopulations of progenitors can be distinguished by single-cell RNA sequencing and differential usage of H3K4me2-positive regulatory regions. In their transcription profile these subpopulations are tied to one of seven hematopoietic fates, and in bone marrow transplantation experiments they can give rise to one or a few cell types (Paul et al., 2015). This suggests that their developmental potential is restricted and that they are (to some extent) refractory to switching identity. Indeed, as recently shown with reprogramming experiments, chromatin plays a key role in the maintenance of cellular identity as illustrated by knockdown of
Figure 3. Enhancer usage during embryonic stem cell differentiation. “Ground state” embryonic stem cells achieve expression (green) of pluripotency genes through specific enhancer clusters or super-enhancers (yellow), while differentiation genes are repressed (purple) or in a bivalent state (green and purple). In primed ESC expression of pluripotency-associated genes is often achieved via a switch in enhancer usage, which also may involve local loop rearrangement. Additional primed ESC genes may also start to be expressed. Upon cell fate determination lineage-specific genes are activated while pluripotency-associated and alternative-fate genes lose active marks and acquire a repressive chromatin environment.

the histone chaperone Caf-1, which enhanced cellular reprogramming toward different cell fates (Cheloufi et al., 2015), suggesting that chromatin itself confers cell-fate stability.

In summary, the expression of cell-identity-specifying genes is pivotal to embryonic development. The transcription of such master regulator genes tends to be controlled by enhancer clusters characterized by an exceptionally high level of active epigenetic marks and binding of TFs that form autoregulatory loops to sustain their own expression. Switching of regulatory networks is
facilitated by the relatively high sensitivity of enhancer clusters to the level of active epigenetic marks in their proximity. Chromatin likely provides a finely tuneable environment for enhancers mediating cellular competence, specification, and determination.

**Concluding Remarks**

Chromatin structure plays an essential role in embryonic development because it determines the environment within which the transcriptional machinery operates. Epigenetic modifications of chromatin are initially determined by maternally encoded factors, with maternal control extending far into development (Hontelez et al., 2015). In the context of permissive chromatin, proper nucleosome positioning at specific sequence-encoded signals is required to complete the switch to zygotic regulation (Haberle et al., 2014; Zhang et al., 2014). Once started, the transcriptional program needs to be dynamically orchestrated to specify different cell identities. In this process, signaling-mediated regulation of TF binding and enhancer usage acting within TADs play a critical role in shaping the transcriptional signature of the differentiating cells (Dixon et al., 2015; Dowen et al., 2014; Factor et al., 2014; Ji et al., 2016; Loh et al., 2014; Tsankov et al., 2015). While cell identities are being established, repression of alternative fates is required for stable cell differentiation and maintenance of complex tissues. DNA methylation and Polycomb proteins play a fundamental role in this process, but new players, such as histone chaperones, are emerging (Cheloufi et al., 2015).

It will be important to address a number of remaining questions. How are the influences of signaling and germ layer specification integrated with the cell-autonomous influences of maternal factors and specific signals embedded within the DNA sequence of regulatory regions to direct chromatin state and the developmental program? With what temporal dynamics and mechanistic hierarchy does this happen, and how are cellular competence and potency balanced with the need for commitment and epigenetic stability of cell fate? What does cell-fate specification and determination mean at the level of chromatin state of genes involved in different developmental programs? What determines whether a TF depends on the chromatin context of its target genes or whether it is able to initiate a remodeling process that will reshape the epigenetic landscape, making it permissive for gene expression? Is this the prerogative of a specific class of TFs, the “pioneer factors,” or can many TFs do this depending on expression level and interactions with other factors? How are chromatin-modifying enzymes, with PRC2 as a prime example, targeted to their genomic locations? Studies in recent years have made exciting inroads into these questions, and much progress is expected in addressing the outstanding questions in years to come.

To a large extent our increased understanding of chromatin and its role in embryonic development has been driven by new technologies. The emerging trends of “big data” analysis of embryonic development (integrating transcriptomics, proteomics, and epigenetics), the continuous development of new and ever more sensitive (epi)genomic techniques (ATAC sequencing [Buenrostro et al., 2013], ChIP/tagmentation [Schmidl et al., 2015], and single-cell techniques [Buenrostro et al., 2015; Jin et al., 2015]), and the combination of these approaches with powerful and fast genetic modifications (CRISPR/Cas9) should allow increasingly more powerful analyses of chromatin state, regulatory networks, and developmental plasticity in cellular and whole embryo model systems alike. Furthermore, integration of these powerful techniques, alongside synergistic approaches combining developmental and computational biology, will provide insight into the
profound questions associated with the multiple levels of complexity of the developing embryo, from egg to organism.

**Acknowledgments**

This work has been financially supported by the People Program (Marie Curie Actions) of the European Union’s Seventh Framework Program FP7 under REA grant agreement number 607142 (DevCom).

References

References


Chapter 3

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

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

* These authors contributed equally to the work


MP, SJvH and GJCV designed experiments and analysis. MP performed experiments and analysis. GvM and HM designed, performed and analyzed mESC whole-proteome experiments and contributed to ChIP. MP, GJCV, IDK and MV designed rescue experiments and Myc pulldowns. IDK performed and analyzed rescue experiments and Myc pulldowns. SvG produced and purified GST-MTF2. MV helped with DNA pulldown experimental design. MP and GJCV wrote the manuscript
Abstract

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 Mtfl2 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.

Introduction

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 RNA5,8,9; however, the DNA-targeting mechanism by which PRC2 is recruited to genes in vertebrates has remained elusive8,9. In Drosophila, PRC2 is recruited to Polycomb-response elements (PREs) by a variety of DNA-binding proteins9,10; 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 specificity11. 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 loci12,13. MTF2 was originally identified as a protein binding to the metal response element (MRE) of the mouse metallothionein (Mt1) promoter14, but was subsequently shown to modulate PRC2 activity at specific developmental genes, to regulate X chromosome inactivation and pluripotency15–17. 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 differentiation16,18–21. PHF19 binds to CpG islands, but this requires the histone-tail binding abilities of its Tudor domain as well22. 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 islands23. Recent studies have solved the crystal structures of PCL proteins and their interaction with
unmethylated DNA. 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. 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). 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). This algorithm successfully identifies the subset of DNA methylation-free islands that acquire H3K27me3 in human, frog and fish. 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 and mouse ESCs. 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 mESCs cultured in the 2i condition. These cells represent ground state pluripotency and closely resemble preimplantation embryos, wherein H3K27me3 is prevalent. 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, 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. Thus, these results suggest that our assay is enriching for an MTF2-containing PRC2 subcomplex, recently termed PRC2.1. The PRC2.1 complexes identified in our pulldown experiments...
Figure 3 PRC2 is recruited by short DNA sequences (previous page). a, Enrichment of k-mer occurrence in X. tropicalis H3K27me3 regions 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 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 Mtf2GT/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).
experiments also contain 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 GSTMTF2 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 Mtf2GT/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 Mtf2GT/GT mESCs. MTF2 was not detected in the Mtf2GT/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\(^ {24,25}\). Co-crystallization 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 ($\Delta$ C-term) all specifically bound to the wildtype 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, $\Delta$Apol, $\Delta$EcoRI/Blpl) 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$^{\Delta\Delta}$) 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$^{\Delta\Delta}$ mESCs 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.
Figure 4 | Functional domains of MTF2 (previous page). 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

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 (ChIPseq) for MTF2, EZH2, H3K27me3 and H3K4me3 in both wildtype and Mtf2GT/GT mESCs. MTF2 ChIP in Mtf2GT/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 for comparisons with unmethylated CpG islands. MTF2 ChIP-seq enrichment was found almost exclusively at a subset of unmethylated genomic locations, with good correspondance to EZH2 recruitment and H3K27me3 enrichment at all MTF2-binding sites (Fig. 3a,b). By contrast, in Mtf2GT/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. 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 complexes1, we investigated its relation with PRC2.2 and performed ChIP-seq for MTF2 in Jarid2−/− cells35 and for JARID2 in Mtf2GT/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. 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 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 Mtf2GT/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\textsuperscript{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\textsuperscript{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\textsuperscript{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 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 CpG24. 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\textsuperscript{37}. Furthermore, DNA helical shape is relevant for the prediction of bound vs. unbound transcription factor binding sites\textsuperscript{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\textsuperscript{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-
Figure 5 | 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.
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 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.

Figure 6: 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.
Figure 7 | 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). mC, 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.
alone (ROCAUCs > 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 co-crystal (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,
Figure 8 | 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.

PRC2.1 and PRC2.2 affect each other, as shown by reduced MTF2 binding in Jarid2−/− and JARID2 binding in Mtf2GT/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 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 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 a new angle on cancer biology, cellular reprogramming and stem cell biology, wherein Polycomb-mediated regulation is known to be important.


**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](https://doi.org/10.1038/s41588-018-0134-8).

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


Acknowledgements

We thank H. Koseki (RIKEN Research Center for Allergy and Immunology, Japan) for sharing the Mtf2GT/GT and Mtf2Δ/Δ 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.
**Supplementary Figures**

Supplementary Figure 1. TGCGAAA kmer recruits PRC2 and recombinant MTF2. a, DNA pulldown-mass spectrometry of 2i mESC nuclear extract using region 2 baits shows highly specific enrichment of PRC2 core subunits and associated proteins on the WT pulldown bait. Proteins on the right side of the plot are enriched in the wild type bait dataset. Highlighted proteins are enriched in both region 1 (Fig. 1b) and region 2 pulldowns. Each condition was measured in three independent experiments. FDR from two tailed t-test. b, DNA pulldown-western blot of 2i mESC nuclear extract with region 1 baits. Recruitment of PRC2 is inhibited by DNA methylation (mC) but not RNAse treatment. mC indicates methylated cytosines in the CG of the TGCGAAA kmer on both strands. Blots represent two independent pulldowns. c, DNA pulldown-western blots of nuclear extract from mESC using baits containing the same kmer but different flanking regions (Supplementary Table 1), representing different genomic locations. Blots represent at least two independent pulldowns. d, DNA pulldown-western blots of recombinant myc-tagged MTF2, myc-tagged C17ORF96 and both recombinant proteins. Recombinant MTF2 binding is specific to the WT DNA bait and mirrors the EZH2 recruitment pattern. C17ORF96 does not show DNA binding properties in either presence or absence of MTF2. Input lane represents 1% of the starting material. Blots represent three independent pulldowns. e, Coomassie staining of purified GST-MTF2 (left), and EMSA showing direct binding of GST-Mtf2 to region 1 probes. Blots represent at least three independent pulldowns. Uncropped gels available in supplementary information.
Supplementary Figure 2: MTF2 is required for PRC2 recruitment. a, Whole-proteome mass spectrometry quantification in 2i versus serum of PRC2 components enriched in 2i DNA pulldown. MTF2 and PRC2 core components show higher expression in 2i. Dotted lines represent two-fold change. Error bars represent SEM. Dashed lines represent log2 FC of 1. The vertical line between the dots indicates the mean FC. b, c, DNA pulldown-mass spectrometry of serum-grown wild type mESC nuclear extract using region 1 (b) and region 2 (c) baits showing specific enrichment of PRC2 core proteins and associated proteins on the WT pulldown bait. Proteins on the right side of the plot are enriched in the wild type bait dataset. Each condition was measured in three independent experiments. FDR from two tailed t-test. d, Stoichiometry of protein present in DNA-pulldown from serum grown mESC. MTF2 is the only non-core protein consistently enriched with a stoichiometric ratio to EZH2 in serum grown cells. Error bars represent standard deviation. Each condition was measured in three independent experiments. Dots represent means, error bars standard deviation. e, DNA pulldown-western blot of serum mESC nuclear extract from the indicated cell lines using region 1 baits, confirming loss of PRC2 recruitment in Mtf2^GT/GT mESC. Replicates of this experiment were detected three times by mass spectrometry (cfr panel b) and two by western blot. f, Whole-proteome mass spectrometry quantification of PRC2 components in (b, c) in wild type and Mtf2^GT/GT mESC. Absence of MTF2 does not affect PRC2 abundance, but only its recruitment to DNA. Dashed lines represent log2 FC of 1. The vertical line between the dots indicates the mean FC. N.D. = not detected in Mtf2^GT.
region 2 (c) baits showing specific enrichment of PRC2 core proteins and associated proteins on the WT pulldown bait. Proteins on the right side of the plot are enriched in the wild type bait dataset. Each condition was measured in three independent experiments. FDR from two tailed t-test. d, Stoichiometry of protein present in DNA-pulldown from serum grown mESC. MTF2 is the only non-core protein consistently enriched with a stoichiometric ratio to EZH2 in serum grown cells. Error bars represent standard deviation. Each condition was measured in three independent experiments. Dots represent means, error bars standard deviation. e, DNA pulldown-western blot of serum mESC nuclear extract from the indicated cell lines using region 1 baits, confirming loss of PRC2 recruitment in Mtf2<sup>GT/GT</sup> mESC. Replicates of this experiment were detected three times by mass spectrometry (cf panel b) and two by western blot. f, Whole-proteome mass spectrometry quantification of PRC2 components in (b, c) in wild type and Mtf2<sup>GT/GT</sup> mESC. Absence of MTF2 does not affect PRC2 abundance, but only its recruitment to DNA. Dashed lines represent log2 FC of 1. The vertical line between the dots indicates the mean FC. N.D. = not detected in Mtf2 GT

Supplementary Figure 3: MTF2 domains C-terminal to PHD2 are dispensable for PRC2 recruitment and not sufficient for MTF2-PRC2 interaction. a, DNA pulldown of E14 (wild type), Mtf2<sup>GT/GT</sup> and Mtf2<sup>Δ/Δ</sup> mESC with region1 bait. PRC2 recruitment is lost in Mtf2<sup>GT/GT</sup> cells while EZH2 from Mtf2<sup>Δ/Δ</sup> mESC show normal recruitment. Blots represent three independent pulldowns. Uncropped gels available as Supplementary Information. b, c, Interaction proteomics of rescue constructs in Mtf2<sup>GT/GT</sup> background. MTF2 isoform 2 (b) binds PRC2 while the construct only encoding the two PHD domains (c) does not. Each condition was measured in three independent experiments. FDR from two tailed t-test.
Supplementary Figure 4: Reliability of MTF2 antibody and reproducibility of ChIPseq experiments.  

**a, ChIP qPCR of MTF2 in wild type and Mtf2^GT/GT^ mESC showing MTF2 antibody specificity. This ChIP was confirmed in 2i grown cells.**

**b, Scatterplot of RPKM values for high-confidence peaks for ChIP-seq replicates, showing high reproducibility of ChIPseq samples.**

**Supplementary Figure 5: Extended characterization of ChIPseq data.**

- **a, b, Heatmap (left) and boxplot quantification (right) of RPKM-normalized ChIP-seq signal on EZH2 (a) or H3K27me3 (b) high-confidence peaks. MTF2 is required for PRC2 recruitment to the majority of its targets. n=7213 peaks for Ezh2 and n=6294 peaks for H3K27me3.**

- **c, RPKM quantification of ChIP-seq signal on repetitive elements, as defined by UCSC genome browser Repeat Masker track. (n=4048423 regions)**

- **d, Additional examples of ChIP-seq signal on peaks belonging to each of the clusters shown in Fig. 4a. Boxplots represent IQR, central bar the median, whiskers are 1.5 IQR. All shown ChIPseq were performed in duplicate excluding MTF2 in Eed-/-**
Supplementary Figure 5: Extended characterization of ChIPseq data. 

a, b, Heatmap (left) and boxplot quantification (right) of RPKM-normalized ChIP-seq signal on EZH2 (a) or H3K27me3 (b) high confidence peaks. MTF2 is required for PRC2 recruitment to the majority of its targets. n=7213 peaks for Ezh2 and n=6294 peaks for H3K27me3.

c, RPKM quantification of ChIP-seq signal on repetitive elements, as defined by UCSC genome browser Repeat Masker track. (n= 4048423 regions)

d, Additional examples of ChIP-seq signal on peaks belonging to each of the clusters shown in Fig. 4a. Boxplots represent IQR, central bar the median, whiskers are 1.5 IQR. All shown ChIPseq were performed in duplicate excluding MTF2 in Eed−/−.
Supplementary Figure 6: Comparison of ChIPseq signal from ChIP-seq of this study and from Li et al\textsuperscript{24}. Data were mapped in the same way and RPKM normalized. \textbf{a}, Heatmaps are centered on the high-confidence MTF2 peaks identified in this study. Li et al reported an analysis based on transcription start sites instead of peaks, and showed the Hoxd cluster as track example (cf. panel b). \textbf{b}, Genome browser screenshot showing an example of the tracks on the Hoxd locus.
Supplementary Figure 7: Identification of additional bound kmers and characterization of DNAsshape within MTF2 peaks (previous page). 

a, Performance of kmer-SVM algorithm for the identification of MTF2 peaks compared to H3K27me3-negative unmethylated DNA, quantified as Receiver Operator Characteristic Area Under the Curve (ROC-AUC, perfect prediction AUC = 1, random AUC = 0.5). G+C percentage and CpG richness (k=1, k=2) have decent predictive power. The algorithm performance, however, steeply increases with longer kmers, with trinucleotide-based prediction closely approaching the performance obtained with k=6 or k=7. n= 8092 regions.

b, Sequence logo of positive-scoring kmers (kmer-SVM of MTF2 peak summits, k=8, weight > 1.5) containing four bases to the left of the CpG dinucleotide, aligned on the CpG. There is a preference for G or C in front of the CpG.

c, Enrichment of high- and low-scoring kmers containing the GCG motif in MTF2 peak summits compared to unmethylated BioCap regions.

d, DNA pulldown of EZH2 and MTF2 with bait region 5 containing the CGCCCGG and TGCGCGCG kmers (cf. panel c) that are also found in the MRE3-4 element of the Mt1 gene. Blots represent three independent pulldowns. Uncropped gels available as Supplementary information.

e, DNA shape parameter distribution of MTF2 and BioCap regions aligned on the 5’ border of the peaks. Predicted DNA shape of MTF2 bound regions differs from the genomic context more than non Polycomb-targeted CpG islands. n=6357 peaks for Mtf2 and n=48247 regions for BioCap.

f, CpG density of MTF2 peak summits (100bp) and unmethylated (BioCap) or genomic regions n=24309 peak summits. Boxplots represent IQR, central bar the median, whiskers are 1.5 IQR.
Supplementary Figure 8: Prediction of bound GCGs using DNA shape information (previous page). a, Performance of shape based prediction of MTF2 binding, using Random Forest-based machine learning algorithm (Methods). The ROC-AUC values (>0.7 for all parameters) show that it is possible to identify bound GCGs using the predicted DNA shape only. n=58903 GCGs. b, DNA shape value distribution of all the GCG trinucleotides occurring in MTF2 peak summits (colored areas) compared with sequence composition matched control regions obtained by randomly shuffling the sequences in the positive set while keeping constant both number and position of GCGs (grey areas). Dark shades represent 10 percentiles around the median, lighter shade the IQR. The shape of MTF2-bound GCGs is highly specific and depends on flanking regions. n=58903 GCGs. c-d, Prediction of the DNA shape of the Region 6 containing TGAGCGCG kmer (c) and region 7 containing CCCGCGCG kmer (d) used for pulldown in Fig. 4d. Bandplots represent IQR of DNA shape parameters of top scoring GCG-containing kmers shown in Fig. 4a. Points represent DNA shape values of non-qualifying GCG contained in the kmers and of the qualifying GCG in the flanking regions. For clarity only the most relevant positions are shown. Number and position of GCG trinucleotides along the 30bp baits are shown above the panels. Only positions with GCGs are shown for clarity. n=2017 regions. e, Comparison of the predicted shape of region 1, region 6, region 7 with the original and reverse-complemented sequence used for the MTF2-DNA co-crystal24. Bandplots represent IQR of DNA shape parameters of top scoring GCG-containing kmers shown in Fig. 4a. Points represent DNA shape values of qualifying GCGs shown in panels c-d, contained in region 1 and in the DNA crystallized with MTF2. For clarity only the most relevant positions are shown. Number and position of GCG trinucleotides along the 30bp baits are shown above the panels. Only positions with GCGs are shown for clarity. n=2017 regions. f, GCG density of MTF2 peak summits (100bp) and unmethylated (BioCap) or genomic regions. n=8092 regions. Boxplots represent IQR, central bar the median, whiskers are 1.5 IQR. g, Quantification of shape qualifying GCGs for each shape parameter, showing a higher occurrence of properly shaped GCGs in MTF2-bound regions. Qualifying GCGs were defined as matching the IQR of the shape parameter from position -2 to position +1 of the GCG. n=8092 regions.
Chapter 4

Cooperative recruitment and interactions of PRC2 sub-complexes on chromatin

Matteo Perino*, Guido van Mierlo*, Sandra Wardle, Hendrik Marks# and Gert Jan C. Veenstra#

*, # These authors contributed equally to the work

MP and GvM designed the study with contributions from HM and GJCV. MP and GvM performed the experiments, and analysis with help from SW. MP and GvM wrote the manuscript, with help from HM and GJCV
Abstract

The Polycomb Repressive Complex 2 (PRC2) is essential for mammalian development as it catalyses the repressive histone mark H3K27me3. The core PRC2 complex consists of EED, EZH2 and SUZ12 and it is often associated with accessory subunits, resulting in the multimodal sub-complexes PRC2.1 and PRC2.2. MTF2 (in PRC2.1) and JARID2 (in PRC2.2) were shown to recruit PRC2 to target genes in ESCs. Currently, it remains unclear if and how these sub-complexes cooperate. Here, we combine a range of Polycomb mutant ESCs with chemical inhibition of PRC2 catalytic activity and ChIP-sequencing to address the complex interactions of the Polycomb system. Our results uncover that PRC2.1 and PRC2.2 are mutually dependent on each other for recruitment. In addition, the EED-H3K27me3 interaction is important for recruitment of both PRC2 sub-complexes and in part conveys their interdependence, but it is not the sole mediator. Strikingly, removal of either JARID2 or H3K27me3 has only minor effect on PRC2 recruitment, whilst combined ablation largely attenuates PRC2 recruitment, suggesting apparent redundancy between JARID2 and EED-H3K27me3 mediated recruitment of PRC2. Finally, we demonstrate that all core PRC2 recruitment occurs through the combined action of MTF2, the EED-H3K27me3 interaction and PRC1-mediated recruitment of JARID2-containing PRC2.2. The data presented here demonstrate interactions between PRC2 sub-complexes and highlight that these interactions are important to establish PRC2 at target sites.

Introduction

Cell fate specification and Embryonic Stem Cells (ESC) differentiation during embryonic development requires tightly controlled epigenetic programs. A key component safeguarding these processes is Polycomb Repressive Complex 2 (PRC2) that catalyses the histone modification H3K27me3, playing an essential role in the establishment of cellular identity by ensuring proper gene silencing$^1$. The critical role of PRC2 during developmental processes is underscored by the embryonic lethality observed in mice lacking a functional PRC2 complex$^{2-4}$. In recent years, our understanding of Polycomb regulation in terms of recruitment, enzymatic activity and sub-complex composition has significantly increased.

PRC2 consists of the core subunits EED, SUZ12 and the catalytic subunit EZH2. Next to these core subunits, PRC2 can contain multiple auxiliary subunits exerting functions such as guiding PRC2 to target genes and modulation of its enzymatic activity. These include PCL proteins (PHF1, MTF2 and PHF19), EPOP and GM340 (also called C10ORF12), which together with the core subunits constitute PRC2.1, and JARID2 and AEBP2, which are part of PRC2.2$^5$.

Within ESCs, a key model to study Polycomb biology, PRC2 core is mainly associated with MTF2 and EPOP, or AEBP2 and JARID2, whereas alternative PRC2.1 complexes containing PHF1, PHF19 and/or GM340 are very lowly abundant, due to the low expression of these subunits. Intriguingly, exit of the pluripotent state is accompanied by rapid downregulation of MTF2 and EPOP, and an upregulation of PHF1, PHF19 and GM340, suggesting that these factors are mainly important for lineage specification or during dissolution of the pluripotent state$^{6,7}$. In ESCs, it has been shown that both MTF2 and JARID2 are important for PRC2 recruitment, whereas ablation of either AEBP2 or EPOP does not affect PRC2 localization$^{8-14}$. Another important player in the recruitment of PRC2 is PRC1, which binds to non-methylated DNA via its subunit KDM2B. PRC1 deposits H2AK119ub
which can be bound by JARID2, resulting in PRC2 recruitment\textsuperscript{15–17}. Furthermore, after establishment of primary PRC2 recruitment, the complex can self-reinforce and spread from its target sites through an allosteric feedback loop by binding of EED WD40 domain to the PRC2 catalytic product, H3K27me3\textsuperscript{18,19}. Interestingly, this mechanism is not sufficient for H3K27me3 maintenance\textsuperscript{30}, indicating the constant requirement of \textit{de novo} recruitment of core PRC2 by its ancillary subunits. Conversely, the H3K27me3 mark can also be bound by canonical PRC1 complexes via the CBX7 subunit\textsuperscript{21} resulting in PRC2-mediated PRC1 recruitment and H2AK119ub deposition. Thus, PRC1 and PRC2 can be recruited independently, whilst simultaneously enhancing each other’s recruitment. Finally, both PRC1 and PRC2 recruitment can be fine-tuned by long non-coding RNAs, as well as through promiscuous binding to nascent RNA\textsuperscript{22–24}. Ample evidence has revealed that MTF2 and JARID2 together mediate primary PRC2 recruitment in ESCs, and ESCs lacking either or both MTF2 and JARID2 show severe PRC2 recruitment phenotypes\textsuperscript{25}, which depends to a large extent on MTF2-mediated DNA binding with a relatively minor contribution of JARID2\textsuperscript{9,14,26}. Interestingly, PRC2 recruitment is partially retained in both Mtf2 and Jarid2 mutant ESCs. As MTF2 and JARID2 are mutually exclusive in the PRC2 complex, this suggests that PRC2.1 and PRC2.2 synergize in establishing Polycomb at target genes\textsuperscript{27}, but how such a synergy materializes is currently unclear.

To investigate the interactions between PRC2.1 and PRC2.2, we combine a range of Polycomb mutant ESCs with chemical inhibition of PRC1 and PRC2 catalytic activity and ChIP-sequencing, to address the complex interactions of the Polycomb system. We demonstrate the individual attribution of primary recruitment mechanisms established by JARID2, MTF2 and H3K27me3. Our data strongly implicates that PRC2.1 and PRC2.2 act synergistically for PRC2 recruitment, which is in part mediated through H3K27me3. Furthermore, our data reveal that JARID2-mediated recruitment and the H3K27me3-EED interaction can substitute for each other. Finally, we show that core PRC2 recruitment occurs through the concerted action of MTF2, JARID2 and the EED-H3K27me3 interaction. These data indicate that PRC2 recruitment can be divided into two major axes, of which one relies more on MTF2-mediated DNA binding, and the other to a larger extent on JARID2-H3K27me3-PRC1 mediated recruitment. Moreover, these different recruitment axes seem to be differentially balanced at a different set of target sites. The data presented here demonstrate interactions between PRC2 sub-complexes and highlight that these interactions are important to establish PRC2 at target sites.

\textbf{Results}

\textit{PRC2 recruitment mainly depends on MTF2}

Recent advances have pinpointed three main recruitment mechanisms of PRC2: I) DNA-mediated recruitment via MTF2, II) recruitment via JARID2, and III) H3K27me3-mediated recruitment via EED (Fig 1a)\textsuperscript{14,17,18,25,26,34}. Currently, it remains unclear if and how these different recruiters cooperate in establishing PRC2 at target genes. To investigate this, we first evaluated whether these mechanisms drive recruitment to the same genomic sites by performing chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-seq) using antibodies against endogenous EZH2, H3K27me3, MTF2 and JARID2. We performed peak calling for EZH2 and determined the occupancy of H3K27me3, MTF2 and JARID2 on these peak sites, which revealed a
near-perfect overlap (Fig 1b). The same scenario was evident for peaks called in H3K27me3 or MTF2 data (Fig S1a-b). By contrast, we observed a large number of sharp JARID2 peaks with little or no occupancy of the other PRC2 subunits (Fig S1c). This could indicate that JARID2 exerts functions independent of the PRC2 complex as previously suggested in Drosophila\textsuperscript{35}, but, as these sites do not overlap with other PRC2 subunits, only peaks overlapping H3K27me3 regions (Fig S1d) were used for subsequent analysis.

Knowing that EZH2, H3K27me3, JARID2 and MTF2 largely co-localize on the genome, we aimed to understand how these factors are involved in recruitment of PRC2. First, we focused on MTF2 and JARID2 and used knockout ESCs for these subunits (Mtf2GT/GT and Jarid2-/-, respectively). These ESCs lack MTF2 or JARID2, respectively, but globally retain wildtype levels of core PRC2 subunits (Fig S1f). ChIP-sequencing revealed a major reduction for EZH2 and H3K27me3 at target sites in MTF2 mutants, whereas the reduction in Jarid2-/- ESCs is minor (Fig 1c, 1d). These observations are in line with previous reports attributing a prime role for MTF2 in PRC2 recruitment\textsuperscript{14,25,26}. To investigate whether PRC2.1 and PRC2.2 mediate each other's recruitment, we profiled the genomic locations bound by MTF2 and JARID2 in the knockout cells, which revealed that MTF2 and JARID2 affect each other’s recruitment, with the most profound effect of MTF2 knockout on JARID2 recruitment (Fig 1e, 1f). This indicates that the PRC2 sub-complexes directly or indirectly modulate their mutual recruitment. Next, to investigate the role of the allosteric EED feedback loop, we extended our analysis to wildtype ESCs treated with EED\textsuperscript{226}. By binding the EED WD40 domain, EED226 hampers the binding of EED to H3K27me3 while simultaneously inducing a conformational change that impedes stimulation of the EZH2 catalytic activity by EED\textsuperscript{31}. Importantly, EED226 does not disturb physical associations between core PRC2 subunits\textsuperscript{31}. After confirming the complete absence of H3K27me3 in EED226-treated ES cells (Fig S1g), we performed ChIP-seq for EZH2, MTF2 and JARID2. This revealed that in the absence of H3K27me3, EZH2 and MTF2 were retained on target sites at near-control levels (Fig 1c, e). JARID2 binding, instead, is >50% reduced by EED226 treatment (Fig 1f), suggesting JARID2 recruitment to target sites partly relies on H3K27me3. Typical examples of these observations above are visualized in Fig 1g. Taken together, these analyses are in line with previous observations that primary PRC2 recruitment in the absence of H3K27me3 occurs largely through MTF2\textsuperscript{14,26}

Figure 1. PRC2.1 and PRC2.2 binding and interdependency (next page). a) Schematic of the recruitment paths for PRC2.1 and PRC2.2. MTF2 binds to DNA and the EED subunit of core PRC2 (orange) binds to H3K27me3 as part of an allosteric feedback loop. The EZH2 subunit of core PRC2 catalyses H3K27 methylation. The PRC2.2 complex contains JARID2 but not MTF2. Both contain the core PRC2 subunits, however the interactions of the PRC2.1 and PRC2.2-specific subunits with chromatin are different. The arrow from JARID2 to DNA is dashed as DNA binding has been shown \textit{in vitro} but not \textit{in vivo}\textsuperscript{40} b) PRC2.1 (MTF2) and PRC2.2 (JARID2) co-localize to all EZH2 targets. c-f) Heatmap and rpkm quantification (boxplots) of PRC2 subunits and the catalytic product H3K27me3. EZH2 recruitment is heavily affected by the absence of MTF2, while JARID2 and H3K27me3 absence have minor effects (c). The effect of MTF2 and JARID2 on EZH2 recruitment is reflected on H3K27me3 deposition (d). MTF2 is marginally affected by H3K27me3 removal, but its binding is reduced to approximately half the WT level in the absence of JARID2 (e). JARID2 recruitment is strongly reduced in the absence of both H3K27me3 and MTF2 (f). g) Genome browser examples of PRC2 binding to classical Polycomb targets. Box plots represent median and interquartile range (IQR; whiskers, 1.5 IQR).
Stratification of Polycomb target sites reveals two major types of targets

We noticed that several of the clusters observed in Figure 1 showed distinct characteristics, such as the strength of binding or the width of the peaks (for example cluster 4 and 5 of the EZH2 ChIP-seq, Fig 1c). Moreover, the consequences of removal of MTF2 or JARID2 seems different per cluster (compare cluster 2 and 4 in Fig 1c, and cluster 3 and 5 in Fig 1f for the Mtf2GT/GT line, or cluster 2 and 5 in Fig 1e for the Jarid2-/- line). Furthermore, recent work implied that MTF2 recruitment relies on the physical presence of EED only in a subset of the EZH2 target26, which suggests distinct modes of recruitment to different genomic regions. To understand how PRC2 recruitment might be distinct depending on the genomic loci analysed, we included the public MTF2 ChIP-seq in ESCs lacking EED (and consequently the PRC2 core), BioCap data to identify regions free of DNA methylation36 (which is common for Polycomb targets and required for MTF2 binding to DNA26) and H3K4me3 ChIP-seq in wild type ESCs (to identify bivalent promoter elements, which represent the majority of Polycomb targets in ESCs37). We combined these data with those shown in Figure 1, confirmed them by ChIP-qPCR (Fig S2), and clustered them to identify dynamic changes among different conditions. Clustering of these data with the ChIP-seq tracks shown in Figure 1 on the common set of PRC2-bound regions revealed six major clusters (Fig 2a). Clusters 1-4 display strong BioCap and H3K4me3 signals and are likely bivalent promoters38, whereas clusters 5-6 show lower BioCap and H3K4me3 signals and could comprise silenced genes. We observed that the consequences of the perturbations for PRC2 recruitment varied per cluster (Fig 2b, S3a). A notable example includes the H3K27me3 signal, which is affected more in clusters 1-4 compared to cluster 5-6 in Mtf2GT/GT ESCs (Fig 2b, right top). Similar patterns hold true for EZH2 and JARID2 recruitment (Fig 2b, left top and right bottom). In addition, removal of EED results in very strong reduction of MTF2 recruitment in all clusters, although this seems slightly more profound in clusters 5-6 compared to cluster 1-4. As recent work revealed that MTF2 is recruited to CpGs in the context of a DNA shape with reduced helix twist26, we also analysed the DNA shape contents of each cluster. This revealed shape-matching GCG trinucleotides, characterized by a helix twist (HelT) and propeller twist (ProT) with values within the permissive ranges, are enriched in cluster 1-4 (Fig 2c), confirming previous findings26. Together, these analyses suggest that cluster 1-4 rely relatively more on MTF2-mediated recruitment compared to cluster 5-6.

Previous reports have suggested that Polycomb target sites contain distinct gene sets37. To attest whether cluster 1-4 and 5-6 also contain different sets of genes, we analysed the genomic elements underlying these clusters. We calculated enrichment (against all genes) of body structure whose development is influenced by the genes of every cluster and found that all clusters show strong enrichments for classical Polycomb targets, as expected from the strong binding of all PRC2 subunits to these regions (Fig S3b). Next, we stratified the clusters by calculating the enrichment only among PRC2 targeted genes, and found that cluster 5 and 6 contain genes prevalently related to body plan, limb, mesenchyme and branchial arches development, including all the Hox genes, while cluster 2 and 4 show a stronger enrichment of genes important for neural structures (Fig 2d). Collectively, these analyses identify at least two distinct Polycomb target regions, which could rely on different mechanisms of PRC2 recruitment.
Figure 3. H3K27me3 feedback loop and JARID2 are mutual backup for PRC2 recruitment. a) Heatmap showing the effect of H3K27me3 depletion on the binding of EZH2. WT and MTF2GT/GT shows mild reduction of EZH2 binding when treated with EED226 inhibitor, while the treatment is highly synergic with the depletion of JARID2. b) Bootstrapping-based RPKM quantification (methods) of the signal in (a). Each coloured dot represent the median of one round of bootstrapping, grey bar represent 99.9% confidence interval for the mean of bootstrapped values in each condition and cluster. c) Treatment with EED226 further affected MTF2 recruitment in Jarid2+/− and JARID2 recruitment in Mtf2GT/GT, with the former leading to recruitment patter closely resembling the Eed+/− line (cf. Fig2a). d) Bootstrapping-based RPKM quantification (methods) of the signal in (c) similar as in 3b. e) Genome browser view of example Polycomb targets. For each genotype two tracks are overlaid: the darker colour represent EED226 treated samples, the lighter colour untreated cells.
JARID2 and H3K27me3 are mutually redundant for PRC2 recruitment

Our analyses allowed us to investigate the individual contributions of MTF2, JARID2 and H3K27me3 to PRC2 recruitment. However, ablation of single subunits individually does not exclude the possibility of compensation by other factors. Thus, we combined knockouts of MTF2 and JARID2 with chemical removal of H3K27me3. We treated Mtf2<sup>GT/GT</sup> ESCs with EED226 to remove H3K27me3, which leaves only JARID2-mediated recruitment intact. Similarly, we combined removal of JARID2 knock-out with EED226 treatment, which would leave only the attribution of MTF2-mediated recruitment (cf. Fig 1a). In both situations, treatment with EED226 resulted in the complete removal of H3K27me3 (Fig S4). Next, we examined the effect on core attribution of MTF2-mediated recruitment (cf. Fig 1a). In both situations, treatment with EED226 simultaneously might phenocopy the effect of through H2AK119ub and JARID2, a scenario in which H3K27me3 and H2AK119ub are absent and especially in cluster 5-6 MTF2 recruitment was near-zero (Fig 3c-e). Interestingly, these JARID2 and H3K27me3 further reduced MTF2 recruitment compared to the removal of JARID2 alone, although the absence of H3K27me3 or JARID2 alone has only marginal effect on EZH2 recruitment, their combined ablation resulted in a major decrease of EZH2 recruitment, as shown both genome-wide (Fig 3a,b) and at specific locations (Fig 3e). These observations indicate that JARID2 and H3K27me3 are redundant for PRC2 recruitment. In addition, this demonstrates that MTF2-mediated recruitment by itself is not sufficient to establish a Polycomb domain but requires PRC2.2 and/or the EED feedback loop. We extended our analyses by performing ChIP-sequencing for JARID2 in Mtf2<sup>GT/GT</sup> +EED226 ESCs and for MTF2 in Jarid2<sup>−/−</sup>+EED226 ESCs. Removal of both JARID2 and H3K27me3 further reduced MTF2 recruitment compared to the removal of JARID2 alone, and especially in cluster 5-6 MTF2 recruitment was near-zero (Fig 3c-e). Interestingly, these are the same clusters in which Eed<sup>−/−</sup> ESCs display near-absence of MTF2 recruitment and where the GCG trinucleotides matching MTF2 DNA-shape requirements are enriched the least (Fig 2a, d). When focusing on JARID2 in Mtf2<sup>GT/GT</sup> +EED226 ESCs, we observed a reduction of recruitment in all clusters although the decrease was marginally stronger in cluster 5-6 (Fig 3b). Together, these data uncover an important contribution of the EED-H3K27me3 interaction to PRC2 recruitment, in particular for PRC2.2, and show that the relative importance of PRC2.1 and PRC2.2 differs across the genome.

JARID2 recruitment is largely dependent on PRC1

Our analyses indicate that both PRC2.1 and PRC2.2 are required to establish PRC2 recruitment to target genes. In this process, the allosteric feedback loop mediated by EED seems to buffer the absence of JARID2 and vice versa. We hypothesized this could be mediated through PRC1, as the catalytic subunit RING1B can deposit H2AK119ub which can in turn mediate JARID2 recruitment<sup>17,39</sup>. As PRC1-mediated PRC2 recruitment was shown to be mediated exclusively through H2AK119ub and JARID2, a scenario in which H3K27me3 and H2AK119ub are absent simultaneously might phenocopy the effect of Jarid2<sup>−/−</sup>+EED226 ESCs. To test this, we used RING1A/B double mutant ESCs treated with EED226 (Ring1a/b<sup>−/−</sup>+EED226) and performed ChIP-sequencing of EZH2, MTF2 and JARID2 in these ESCs. Interestingly, we observed that the obtained EZH2 and MTF2 profiles in Jarid2<sup>−/−</sup>+EED226 and Ring1a/b<sup>−/−</sup>+EED226 were nearly identical (Fig. 4a).
Figure 4. PRC1-PRC2 crosstalk in the absence of H3K27me3. a) Heatmap showing EZH2, MTF2 and JARID2 binding in the absence of H3K27me3 in PRC2 and PRC1 mutant lines in the absence of H3K27me3, JARID2 and RING1A/B mutant phenocopy each other with regard to EZH2 and MTF2 binding, suggesting and effect on the same recruitment mechanism. JARID2 recruitment is also strongly affected by the absence of RING1A/B, suggesting that JARID2 might be mediating PRC2 recruitment via binding to PRC1-deposited H2AK119ub. b) Average plot of the ChIP signal shown in (a), centred on called peaks. c) Heatmap showing Ring1b binding in several conditions. Ring1b is only mildly affected by removing H3K27me3 using EED226 (~40%). Binding is further attenuated in MTF2 and JARID2 mutant ESCs. d) Examples loci of the data shown in (c). e) Average plot of the ChIP signal shown in (c), centred on called peaks.

Also, JARID2 binding was affected in Ring1a/b−/−+EED226, especially in cluster 5 and 6, which are less dependent on MTF2 (Fig 4a-b). These observations indicate that JARID2 and PRC1 recruit PRC2 via the same recruitment mechanism. As JARID2 recruitment seems to be retained to some extent in the RING1a/b+EED226 condition, additional mechanisms might attract JARID2 to target sites, for example binding of JARID2 to DNA or RNA-mediated scaffolding. To extend our comprehension of PRC1/PRC2 interdependencies and disentangle the roles of H3K27me3 versus PRC2 subunits in PRC1 recruitment, we performed RING1B ChIP-seq in WT, MTF2^GT/GT, and Jarid2^−/−.
in the presence of EED226. Interestingly, removal of H3K27me3 in wildtype ESCs had only limited effect on RING1B recruitment (Fig 4c-e, Fig S5), which is in line with the observation that PRC1 can be recruited independent of PRC2\textsubscript{15}. Surprisingly, however, the combined absence of H3K27me3 and either MTF2 or JARID2 results in strong reduction of RING1B binding (Fig 4c-e). While the Polycomb dogma posits that PRC1 and PRC2 mutually affect each other only via their catalytic products (respectively H2AK119ub and H3K27me3), these data suggest that other H3K27me3-independent mechanisms exist that mediate PRC1 recruitment. As direct physical interaction between the two complexes has never been documented, this would require alternative explanations, such as scaffolding effects of RNA molecules. Alternatively, it is possible that the physical presence of PRC2 (which is strongly reduced in MTF2\textsuperscript{GT/GT}+EED226 and Jarid2\textsuperscript{-/-}+EED226 ESCs) stabilizes PRC1 binding to chromatin, which could shift the balance of a competition with additional factors or by building a generally more favourable chromatin context for PRC1 recruitment.

Figure 5. EZH2 recruitment depends on MTF2, H3K27me3 and H2AK119ub (previous page). a) Heatmap of ChIP-seq signal for EZH2 in multiple conditions including Mtf2\textsuperscript{GT/GT} cells double inhibited (d.i.) with EED226 (to remove H3K27me3) and MG132 (to remove H2AK119ub). b) Example loci of the data shown in (a). c) Average profiles of the ChIP signal shown in (a), centred on called peaks. Lower row is a zoom in of the area in the black box.
**PRC2 recruitment is mediated through a combined action of PRC2.1 and PRC2.2**

Finally, we investigated whether the residual EZH2 recruitment observed in *Ring1a/b−/−*+EED226 and *Jarid2−/−*+EED226 ESCs was mediated through MTF2. To do so, we used *Mtf2*{GT/GT} ESCs in which we removed H3K27me3 using EED226 and additionally H2AK119ub using MG132 (*Mtf2*{GT/GT} + d.i.), a drug that can be used to remove H2AK119ub\(^{15}\), and performed ChIP-seqencing for EZH2 in these ESCs. In this ‘triple ablation’ condition, the recruitment of EZH2 to target genes was completely abrogated (input levels, Fig 5, Figure S6). Collectively, these analyses demonstrate that the combined action of MTF2, the allosteric EED feedback loop and PRC1-mediated recruitment of JARID2-containing PRC2 sub-complexes are required for PRC2 recruitment in ESC.

**Discussion**

The mechanisms guiding PRC2 to target sites have long remained enigmatic. Clear evidence demonstrated that the allosteric feedback loop mediated by EED is important for spreading of PRC2 away from its initial nucleation site\(^{18}\). However, the mere presence of H3K27me3 is not sufficient to maintain PRC2 at its target genes, which indicated that continuous DNA-mediated sequence-specific recruitment is required to attract PRC2 to newly synthesized chromatin strands\(^{20}\). The discovery of facultative PRC2 subunits and the presence of functionally distinct sub-complexes has since greatly advanced our understanding of PRC2 recruitment\(^{27,42}\). In particular, individual ablation of all prime facultative subunits in ESCs revealed a major role for MTF2 in PRC2 recruitment, which together with JARID2 mediates the initial PRC2 recruitment to the initiation sites (‘nucleation sites’)\(^{14,25,26,40}\). However, MTF2 and JARID2 occupy distinct mutually-exclusive PRC2 sub-complexes and it remained unclear how these sub-complexes cooperate\(^{43}\). In this study, we aimed to unify the role of each prime constituent that mediates primary PRC2 recruitment. Our analyses further confirm previous observations that MTF2 is required for a significant proportion of PRC2 recruitment\(^{14,26}\) and extend on recent work highlighting that the remaining recruitment is mediated via JARID2\(^{25}\). We show that MTF2 and JARID2 modulate each other’s recruitment, which in part is mediated through the EED feedback loop and in part through PRC1. It remains a question how JARID2 affects MTF2 recruitment mechanistically, as absence of JARID2 does not affect EZH2 (and hence core PRC2\(^{42}\)) recruitment but does result in a ~50% reduction of MTF2 recruitment. A potential explanation would be that hybrid PRC2.1/2.2 complexes containing AEBP2 and MTF2 form that show altered function, such as for example JARID2-MTF2 hybrid complexes in AEBP2 mutant ESCs\(^{13}\). The formation of hybrid complexes could sequester the complex or inhibit MTF2-mediated recruitment. Alternatively, RNA-mediated scaffolding could be inhibited in the mutants, resulting in a reduced recruitment of the residual subunits.

Combining MTF2 or JARID2 absence with removal of H3K27me3 allowed us to gain additional insights into the establishment of Polycomb domains. A notable example includes the redundancy of JARID2 and H3K27me3 as observed in the current study, which indicates that these factors can substitute for each other, at least to a large extent. The observations made here concern a steady state situation and assessing such a redundancy mechanistically would require conditional removal of both factors followed by time-course ChIP-seq experiments. However, it is tempting to speculate that to initiate the formation of a Polycomb domain, only initial recruitment to target
sites is required to establish sufficient amount of PRC2 molecules that catalyse H3K27me3, which can kick-start an allosteric feedback loop. As primary recruitment is mainly mediated via MTF2, such a loop can still exist in the absence of JARID2 (when the allosteric loop can still occur through EED). In addition, when H3K27me3 is absent, we speculate that an alternative route could take over that requires JARID2 binding to H2AK119ub. As such, only combined ablation of both JARID2 and H3K27me3 would attenuate all possible feedback loops, leaving only DNA-mediated recruitment intact, as observed here. This interpretation is further substantiated by our analysis of ESCs lacking MTF2, H3K27me3 and H2AK119ub in which all core PRC2 recruitment is absent.

The observations in the current study further substantiate previous work that demonstrated that the role of PRC1 and PRC2 are large intertwined, as both complexes can be recruited independently but simultaneously modulate their mutual recruitment. Our analyses of EED226 treated ESCs reveals that ~40% of PRC1 recruitment depends on the presence of H3K27me3 (Fig 4e), which would be mediated through canonical complexes containing CBX7. The remainder of PRC2-independent PRC1 is likely recruited via KDM2B-mediated DNA binding, which is in line with previous observations showing that ~60% of RING1B recruitment is mediated by KDM2B. Surprisingly, our analyses reveal that MTF2 and JARID2 deficient ESCs treated with EED226 show a more profound decrease (than ~40%) of RING1B occupancy at target genes. This could indicate either a so far unknown link between PRC1 and PRC2, or alternatively, a stabilization of KDM2B-mediated recruitment to DNA by the physical presence of core PRC2, that would extend the residence time of PRC1 on chromatin. Finally, it is possible to imagine a scaffolding role of (l)ncRNAs, which, if simultaneously bound by PRC2 and PRC1, could explain the reduced RING1B recruitment in MTF2 and JARID2 mutants in absence of H3K27me3. Together, these observations further corroborate the dogma that PRC1 and PRC2 can be recruited autonomously, but are synergistic for each other’s function.

In addition to the insights into specific recruitment mechanisms, we provide evidence that Polycomb target regions can be sub-divided into (at least) two major groups. The largest group (in this study cluster 1-4) contains mainly bivalent genes which rely relatively more on PRC2.1-mediated recruitment. The smaller group (in this study cluster 5-6) seems to rely more on PRC1 and PRC2.2. Of note, cluster 5-6 contains very lowly expressed and developmentally relevant genes such as all the Hox genes (Table S1). As the two major target groups harbour functionally different genes, it is tempting to speculate that distinct recruitment mechanisms are required. This is supported by the observation that MTF2 is strongly downregulated upon differentiation while the paralogues PHF1 and PHF19 are upregulated and start occupying the PRC2.1 complex. Although these paralogues share many genomic targets in ESCs, they might have different genomic targets upon ESC differentiation as MTF2 is recruited by shape-specific DNA binding and PHF1 and PHF19 via binding to H3K36me3 and potentially sequence non-specific DNA binding. As such, it is possible that the bivalency of cluster 1-4 is resolved resulting in expression of these genes in early differentiation. Cluster 5-6, however, comprise the genes required for later differentiation (such as limb formation) and as PRC2.2 is stable during early differentiation, these genes will remain silenced.

Interestingly, inhibition of PRC2 catalytic activity using EED226, which results in the absence of H3K27me2/3, has only limited implications for core PRC2 recruitment, whereas mutating the WD40 domain of EED has major consequences. On the one hand, this could indicate that
inhibition by EED226, while preventing binding to H3K27me3 and allosteric activation of EZH2, still allows binding to potential other substrates for which EED has only limited affinity such as other tris(methylated histone lysines, unmodified histone tails, RNA scaffolds or DNA interactions. On the other hand, it is possible to conceive that mutations in the highly hydrophobic aromatic cage of the WD40 domain induce additional conformational changes on EED itself or other PRC2 subunits, resulting in molecular phenotypes more severe than the one attributable to the WD40 domain alone. Further structural studies will be required to clarify this question.

Collectively, the observations here provide novel insights into Polycomb recruitment in ESCs and provide a model in which PRC2 recruitment can be initiated solely through direct recruitment via DNA, after which PRC2.1/PRC2.2 and PRC2/PRC1 functional interactions are required to achieve the establishment of a complete Polycomb domain through self- and mutual reinforcement and spreading to neighbouring nucleosomes.

**Methods**

**Embryonic stem cell culture**

Wildtype E14 ESCs (129/Ola background) and knockout ESCs were maintained in Dulbecco’s Modified Eagle Medium (DMEM) containing 15% fetal bovine serum, 10 mM Sodium Pyruvate (Gibco), 5 μM beta mercaptoethanol (BME; Sigma) and Leukemia inhibitory factor (LIF: 1000U/ml; Millipore). Eed^-/- ESCs have been described by Schoeftner et al 28., Jarid2^-/- ESCs have been described in Landeira et al8. Mtf2 knockout (Mtf2GT/GT)29 and Ring1a^-/-/Ring1b+/- ESCs30 were a kind gift from Haruhiko Koseki. Ring1b ESCs are knockout for Ring1a and heterozygous for Ring1b. Full knockout of Ring1b was induced through treatment with Tamoxifen (OHT) for 2 days. To inhibit EED function, ESCs were treated with 10 μM EED22631 for 4 days. Complete removal of H3K27me3 was validated using western blot.

**Western blot**

Cell pellets were dissolved in RIPA buffer at a density of 10^4 cells per μl and briefly sonicated to ensure proper cell lysis. Proteins denatured in SDS-PAGE gels were transferred onto PVDF membranes. Primary antibodies used were rabbit anti-MTF2 (ProteinTech; 16208-1-AP), rabbit anti JARID2 (Novus Bio; NB100-2214), rabbit anti H3K27me3 (Millipore; 07-449), rabbit anti-H3 (Abcam; 1791). Secondary antibodies were HRP-conjugated anti-rabbit (Dako; P0217) and anti-mouse (Dako; P0161). Protein bands were visualized using Pierce ECL western blotting substrate (Thermo). Images were analysed using ImageJ.

**ChIP-sequencing**

Cells were crosslinked in 1% PFA at room temperature for 10 min. The crosslinking reaction was halted using 1.25M glycine and cells were harvested by scraping in buffer B (0.25% Triton X-100, 10 mM EDTA, 0.5 mM EGTA, 20 mM HEPES). The suspension was centrifuged for 5 min at 1600 rpm, 4 °C and the pellet was resuspended in 30 ml buffer C (150 mM, 1 mM EDTA, 0.5 mM EGTA, 50 mM HEPES) and rotated for 10 min at 4 °C. The nuclei were centrifuged 5 min at 1600 rpm, 4 °C and resuspended in incubation buffer (0.15% SDS, 1% Triton X-100, 150 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES) supplemented with Protease inhibitor. Nuclei were sonicated using a Bioruptor Pico to obtain chromatin with an average DNA length of 300 bp. The chromatin was
snap frozen and stored at -80 °C until further use. The sonicated chromatin was incubated overnight with the required antibody and pulled down using protein A/G magnetic beads. After washes, eluted chromatin was de-crosslinked overnight and purified with MinElute PCR Purification columns (Qiagen). After qPCR quality check for target enrichment, up to five ng/sample of ChIP were prepared for sequencing with the Kapa Hyper-prep Kit (Kapa Biosystems) using NEXTFlex adapters (Bio Scientific) and amplified with 8-11 cycles of PCR. After size-selection using E-gel (Invitrogen) to enrich for 300bp fragments, libraries were sequenced paired-end on an Illumina NextSeq500. qPCR analysis of ChIP DNA was performed with iQ SYBR Green Supermix (Bio-Rad) on a CFX96 Real-Time System C1000 Thermal Cycler (Bio-Rad). All the ChIPs in this study were performed at least in duplicate.

Antibodies

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), JARID2 (Novus Biologicals NB100-2214, Lot E2), RING1B (Abcam, AB3832 lot GR86503-25).

Bioinformatic analysis

Data from Perino et al.,26 were reprocessed in parallel with those of this study. To ensure maximum comparability (75bp single-end, 42bp paired-end) and accurate quantification, all fastq files were trimmed to 42bp using fastx_trimmer (version 0.0.13.2), and in case of paired-end sequencing only read_1 was used for mapping. All fastq files were mapped using bwa (version 0.7.10-r789), filtered to retain only uniquely mapping reads using mapping quality of 30 and samtools (version 1.7, flag -F 1024), and normalized for sequencing depth to produce bigwig. Peaks were called with MACS2-2.76 with qvalue 0.0001 using --call-summits for transcription factors and --broad for H3K27me3. Only peaks independently called in both replicates were used for downstream analysis. High-confidence peaks for each mark were obtained by merging peaks called in both replicates and overlapping by at least 50% of their length, and combined to obtain the list of all PRC2 peaks. Heatmaps of ChIPseq signal were generated using fluff v3.0.27, and clustered for dynamics using the “--g” option. ChIP metaplots were obtained with deeptools v 3.1.332. Anatomy tern enrichment was calculated using MouseMine33. RPKM bootstrapping analysis was performed using scipyr (v 1.1.0). RPKM from the two independent ChIPseq replicates were combined into a single pool. Values were drawed from this pool, recorded, and returned, such that every value could be drawed multiple times. For each bootstrapping round a number of values matching the total number of PRC2 peaks was drawed, and the median plotted as one dot in the swarmplot. Confidence intervals (99.9%) were calculated from 100 bootstrapping events.

Whole cell proteomes

Cell pellets were dissolved in RIPA buffer at a density of 10⁴ cells per µl and briefly sonicated to ensure proper cell lysis. Protein extracts (10 µg) were processed using Filter Aided Sample Preparation (FASP) and digested overnight with Trypsin. Peptide mixtures were desalted prior to LC-MS analysis.

References


**Supplementary Figure 1** (previous page). **a-d)** Heatmap of WT ChIP-seq signal on the indicated peak set. H3K27me3-negative JARID2 peaks were excluded from further analysis. **e)** Venn diagram showing the overlap of peaks called for each protein independently. **f)** Mass spectrometry quantification of PRC2 subunits in the different cell lines. Detection of JARID2 and MTF2 in the respective mutants (asterisks) are due to value imputation in Perseus. **g)** Western blot validation of EED226 depletion of H3K27me3, for the ChlP shown in Fig1-2. **h)** Scatterplot of peak RPKM showing high reproducibility of ChlP replicates.

Supplementary Figure 2. ChIP-qPCR validation of the deep sequencing data shown in Figure 2a. Each dot represents the enrichment value in one replicate.

**Supplementary Figure 3** (next page). **a)** Average plot of the ChIP signal shown in Fig 2a, centred on called peaks. **b)** Enrichment of anatomical terms in the genes associated with peaks in the six clusters shown in Fig 2a. Enrichment over all genes.
Supplementary Figure 4. Western blot validation of EED226 depletion of H3K27me3 for the ChIP shown in Fig 3-4.

Supplementary Figure 5. a) Examples loci of the data shown in Fig 4c, two for each cluster.
**Supplementary Figure 4.** Western blot validation of EED226 depletion of H3K27me3 for the ChIP show in Fig3-4

<table>
<thead>
<tr>
<th></th>
<th>Wildtype</th>
<th>Mtf2&lt;sup&gt;GT/GT&lt;/sup&gt;</th>
<th>Jarid2&lt;sup&gt;-/-&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctrl</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
<tr>
<td>+EED226</td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
</tbody>
</table>

### Supplementary Figure 5. a) Examples loci of the data shown in Fig 4c, two for each cluster.

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>chr5</th>
<th>118165000</th>
<th>118177500</th>
<th>chr11</th>
<th>95115000</th>
<th>95127500</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>WT + Eed226</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Mtf2&lt;sup&gt;GT/GT&lt;/sup&gt; + Eed226</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Jarid2&lt;sup&gt;-/-&lt;/sup&gt; + Eed226</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Input</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 2</th>
<th>chr1</th>
<th>128585000</th>
<th>128597500</th>
<th>chr10</th>
<th>67530000</th>
<th>67542500</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>WT + Eed226</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Mtf2&lt;sup&gt;GT/GT&lt;/sup&gt; + Eed226</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Jarid2&lt;sup&gt;-/-&lt;/sup&gt; + Eed226</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Input</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 3</th>
<th>chr4</th>
<th>109960000</th>
<th>109672500</th>
<th>chr6</th>
<th>28255000</th>
<th>28267500</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>WT + Eed226</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Mtf2&lt;sup&gt;GT/GT&lt;/sup&gt; + Eed226</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Jarid2&lt;sup&gt;-/-&lt;/sup&gt; + Eed226</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Input</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 4</th>
<th>chr2</th>
<th>147077500</th>
<th>147090000</th>
<th>chr8</th>
<th>12111000</th>
<th>12112500</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>WT + Eed226</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Mtf2&lt;sup&gt;GT/GT&lt;/sup&gt; + Eed226</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Jarid2&lt;sup&gt;-/-&lt;/sup&gt; + Eed226</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Input</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 5</th>
<th>chr5</th>
<th>139537500</th>
<th>139550000</th>
<th>chr5</th>
<th>35367500</th>
<th>35400000</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>WT + Eed226</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Mtf2&lt;sup&gt;GT/GT&lt;/sup&gt; + Eed226</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Jarid2&lt;sup&gt;-/-&lt;/sup&gt; + Eed226</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Input</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 6</th>
<th>chr18</th>
<th>53457500</th>
<th>53470000</th>
<th>chr18</th>
<th>53457500</th>
<th>53470000</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>WT + Eed226</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Mtf2&lt;sup&gt;GT/GT&lt;/sup&gt; + Eed226</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Jarid2&lt;sup&gt;-/-&lt;/sup&gt; + Eed226</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Input</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>
Supplementary Figure 6. a) Examples loci of the data shown in Fig 5, two for each cluster.
Chapter 5

MTF2 controls meso-endodermal fate specification of mESC

Matteo Perino, Chet Loh, Henk Stunnenberg and Gert Jan C. Veenstra

MP designed the study with contributions from GJCV. MP produced and collected the samples, performed and analysed bulk RNAseq. CL performed and analysed scRNAseq. HS provided support for scRNAseq. MP wrote the manuscript, with help from GJCV
**Abstract**

Polycomb Repressive Complex 2 (PRC2) is required for deposition of H3K27me3, a mark critical for acquisition and maintenance of cell identity. MTF2, a substoichiometric PRC2 component, has recently been shown to be of critical relevance for PRC2 recruitment in mESC, initiating the deposition of H3K27me3. Despite MTF2 requirement for mouse, Xenopus, and chicken embryogenesis, the molecular effect of MTF2 depletion in the early stages of embryonic development have not been addressed. Using bulk and single-cell RNA-seq we show that MTF2 is required to prevent cell identity specification toward somatic mesoderm, and to repress somatic mesoderm genes upon primordial germ cells specification and gain of pluripotency.

**Introduction**

Cell fate is acquired and maintained through the expression of a carefully balanced set of genes, whose functions define cell identity. These networks are established and maintained through epigenetic modifications that enforce a responsive, yet stable, regulatory landscape. A key player in this environment is Polycomb Repressive Complex 2 (PRC2) that, via the trimethylation of histone 3 lysine 27 (H3K27me3), ensures the repression of genes outside the appropriated network. The developmental relevance of this mechanism has prompted extensive research, leading to the identification and characterization of the components of this complex. Two core proteins, EED and SUZ12, associate with one of the two paralog catalytic subunits, EZH1 and EZH2, and one of the promiscuous histone-binding protein RBBP4/7, all required to assemble the complex. The absence of any of these proteins prevents the assembly of the core complex and leads to degradation of the other components (for reviews see references 5-9). Although these proteins are sufficient to assemble a catalytically functional PRC2 complex, numerous substoichiometric units can associate with the core, modifying its function, activity, and recruitment dynamics. The combination of these subunits defines two sub-complexes, recently named PRC2.1 and PRC2.2: PRC2.1 is defined by the presence of one of the three PCL homologs (PHF1, MTF2, and PHF19) and EPOP (C17ORF96) or GM340 (C10ORF12), while PRC2.2 is defined by the association with JARID2 and AEBP2.

PCL proteins in particular have recently been shown to have DNA binding properties, with MTF2 being fundamental for PRC2 recruitment in mouse embryonic stem cells (mESC). In mESC, MTF2 binds unmethylated CpG with a partially unwound DNA helix and mediates direct PRC2 recruitment to DNA, and is required for efficient mESC self-renewal. Upon exit of pluripotency, MTF2 is downregulated and replaced by increased expression of PHF1 and PHF19, which are not expressed in mESC. The role of MTF2, however, is not limited to mESC, and it is re-expressed in later development. Mtf2 mRNA is detected in the mouse brain from embryonic day 11.5 (E11.5) to adulthood, in the segmenting somites and tail (E11.5), developing limbs (E13.5), and thymus (E13.5, E15.5) (Allen mouse atlas, Fig S1). This expression pattern is in agreement with the multiple phenotypes observed in vivo upon Mtf2 expression perturbation in several organisms: I) in mice the absence of MTF2 leads to postnatal lethality and anterior-posterior patterning phenotypes, with cervical vertebrae assuming more posterior identity. MTF2 also regulates definitive erythropoiesis by repressing Wnt signalling in erythroid progenitors; II) in chicken, MTF2...
regulates left-right axis specification by repressing Shh signalling in the right side of the embryo, and its absence induces randomization of the direction of heart looping\(^{17}\), although this phenotype is not evident in mouse\(^{18}\); III) in Xenopus, mRNA injections induce malformation of neural structures and inhibit development of anterior structures in a dose-dependent fashion\(^{19}\).

Despite this extensive body of knowledge, little is known on how MTF2 regulates the transitions from mESC to the different lineages where it is expressed, especially at the molecular level. Thus we set out to characterize the role of MTF2 in mESC differentiation, using both bulk and single-cell RNAseq to detect transcriptome changes during undirected differentiation of wild type, Mtf2 (Mtf2\(^{GT/GT}\)) and Eed (Eed\(^{-/-}\)) mutant mESC to embryoid bodies.

**Results**

**MTF2 is required for pluripotent and differentiating mESC survival.**

In order to investigate the molecular role of MTF2 during the transition from pluripotency to lineage specification, we differentiated wild type (WT), Eed mutant (Eed\(^{-/-}\)), and Mtf2 gene trap mutant (Mtf2\(^{GT/GT}\)) mESC to embryoid bodies (EB) via withdrawal of the anti-differentiation factor LIF. We allowed the cells to differentiate, and performed time-course bulk RNA-seq and single cell RNA-seq experiments (Fig 1a). This undirected differentiation gives raise to all the cell lineages\(^{20}\), thus enabling detection of both enrichment and depletion of lineage-specific markers. We used mESC as reference, and compared Mtf2 mutant EB to both WT and Eed\(^{-/-}\), which represent the two extremes of the perturbation spectrum: the physiological mESC state, and a full PRC2 knock-out. Following differentiation we noticed that the Mtf2\(^{GT/GT}\) line had a slower growth rate (growth curve not shown), resulting in smaller EB already at day 4, a phenotype that was exacerbated at day 7 (Fig 1b). This is accompanied by an increased number of dead cells (not shown) and cells detaching from the EBs, resulting in an irregular shape profile, the latter being mildly present also in Eed\(^{-/-}\) (Fig 1b, day 4 and day 7). We then performed pairwise transcriptome comparison of the three mESC lines, and performed gene ontology analysis to capture functional enrichments of differentially expressed genes. We identified 374 and 753 upregulated genes, and 79 and 1918 downregulated genes when we compare Mtf2\(^{GT/GT}\) to WT and Eed\(^{-/-}\) respectively. As expected, genes upregulated in Eed\(^{-/-}\) are enriched for developmental terms, a classical feature of Polycomb mutants, partially present in Mtf2\(^{GT/GT}\) mESC as well (Fig 1c). Additionally, both mutants show enrichment for cell-substrate adhesion-related genes. Furthermore, when comparing the two PRC2 mutants, Mtf2\(^{GT/GT}\) shows a specific enrichment for multiple cell-death related terms, providing molecular support to the morphological phenotypes observed during EB differentiation. It is also worth noting that, although being broadly related due to classification as 'developmental' terms, upregulated genes in the two PRC2 mutants are enriched for distinct anatomical features, with those of Mtf2\(^{GT/GT}\) corresponding to the establishment of specific structures rather than development of general systems. As we previously showed by ChIP-sequencing in mESC\(^{11}\), these data suggest that MTF2 primarily regulates a subset of Polycomb targets in mESC, while also being required for mESC maintenance\(^{21}\).
and primordial germ cells

Our samples: a first vector, along the diagonal, follows the path of differentiation; a second one segregates the reproducibility of the experiments, and shows two main directions of gene expression variability in overall transcriptome change of the three mESC across all time points. We used principal component analysis (PCA) to reduce the space dimensionality and plotted the first two components, which together explain 82% of the detected variance (Fig 2a). PCA highlights the high reproducibility of the experiments, and shows two main directions of gene expression variability in our samples: a first vector, along the diagonal, follows the path of differentiation; a second one segregates the Eed−/− samples from the other two cell lines, indicating an overall different transcriptome already at the mESC stage, as suggested by the number of differentially expressed genes and their gene ontology enrichment.

In order to consider this initial difference in expression level when calling significant differences in gene expression, we applied Likelihood Ratio Test (LRT) statistics to analyse the time-course RNAseq (Methods). This approach evaluates the expression of each gene during the differentiation of each line and compares the respective patterns, testing for different temporal dynamic, rather than absolute difference in gene expression.

Figure 1: EB differentiation of mESC. a) Schematic of experimental design. b) Bright field pictures of mESC and differentiating EB. It is possible to notice the smaller size of $Mtf2^{GT/GT}$ EB, and the irregularly shaped edges of the EB from both mutant lines compared to the smooth, rounded WT EB. Scale bar 500μm. c) Biological processes gene ontology analysis of differentially expressed genes in pairwise comparison of mESC lines. p-values are corrected for multiple testing using Benjamini-Hochberg correction.

MTF2 controls the development of meso-endodermal tissues and primordial germ cells

Next, we investigated the role of MTF2 during undirected cell differentiation, and evaluated the overall transcriptome change of the three mESC across all time points. We used principal component analysis (PCA) to reduce the space dimensionality and plotted the first two components, which together explain 82% of the detected variance (Fig 2a). PCA highlights the high reproducibility of the experiments, and shows two main directions of gene expression variability in our samples: a first vector, along the diagonal, follows the path of differentiation; a second one segregates the Eed−/− samples from the other two cell lines, indicating an overall different transcriptome already at the mESC stage, as suggested by the number of differentially expressed genes and their gene ontology enrichment.

In order to consider this initial difference in expression level when calling significant differences in gene expression, we applied Likelihood Ratio Test (LRT) statistics to analyse the time-course RNAseq (Methods). This approach evaluates the expression of each gene during the differentiation of each line and compares the respective patterns, testing for different temporal dynamic, rather than absolute difference in gene expression.
Using this approach, we compared the expression of several lineage-specific marker genes to address the differentiation potential of the three mESC lines (Fig 2b, Fig S2). Compared to WT, both Polycomb mutants showed a reduced activation of neural and neuro-ectodermal genes, and an increased expression of key mesodermal markers such as T and Eomes. Mtf2\textsuperscript{GT/GT} EB also show stronger expression of endodermal and non-neural ectoderm markers, while Eed\textsuperscript{-/-} EB display a higher expression of genes associated with neural crest identity and retain higher expression of some pluripotency markers, such as Nanog and Prdm14, among those tested. To gain a more comprehensive overview of the different cell identity acquired during differentiation we calculated the anatomy term enrichment of the most significant and upregulated genes against the respective mESC line for each time-point (Fig 2c). While WT EB showed a very strong enrichment for neuro-ectodermal anatomy terms at both time points, mutant EB lack overrepresentation of genes expressed in these tissues, as suggested by marker genes expression. Both Polycomb mutants are instead characterized by the overrepresentation of mesodermal, mesenchymal and endodermal terms. When comparing Mtf2\textsuperscript{GT/GT} to Eed\textsuperscript{-/-} the former shows enrichment for specific mesodermal and endodermal terms, associated with limb and tail development, extraembryonic tissues, vascular system, and skeletal muscle development. This general pattern is mirrored in the anatomy category enriched in downregulated genes, with Eed\textsuperscript{-/-} showing downregulation of neural terms and Mtf2\textsuperscript{GT/GT} of nasal and sensory epithelia, while WT EB have lower expression of meso-endodermal terms (Fig 2d). Interestingly, WT cells are the only ones displaying significant downregulations of genes typical of early embryonic stage and connected to pluripotency. This downregulation, surprisingly, is only transient, and by day 7 only few terms retain significance, which is however largely reduced. This temporal dynamic is reflected by the expression of pluripotency marker genes (Fig 2b, Fig S2), that in are strongly re-expressed by day 7 in WT. A remarkable example is Prdm14, which is repressed to virtually no expression at day 4, but by day 7 WT EB show an expression level almost double the one detected in mESC. On the contrary, in Mtf2\textsuperscript{GT/GT} EB there is no reactivation, and Eed\textsuperscript{-/-} EB, which start form a higher expression in mESC, show a constant downregulation trend. As Prdm14 is highly expressed in both hematopoietic stem cells (HSC) and Primordial Germ Cells (PGC), two classes of multipotent proliferative cells, we further investigated whether we could identify a role for MTF2 in the acquisition of these cell identities. Few of the tested marker genes of the hematopoietic lineage\textsuperscript{22} show differential expression among the lines, and those whose expression is perturbed do not have a matching temporal dynamic with Prdm14 (Fig S3). Conversely, three other key characterising genes of PGC identity, Blimp1(Prdm1), Stella (Dppa3) and Tfap2c\textsuperscript{23}, show strong expression in WT EB at day 7 but are kept silent in Mtf2\textsuperscript{GT/GT} (Fig3a). Strikingly, in WT EB, Mtf2 mirrors the expression pattern of Prdm14, while another PCL paralog, Phf19 recapitulates the expression pattern of PGC marker genes, being strongly upregulated at day 7 in WT but not in Mtf2\textsuperscript{GT/GT} (Fig 3b). As a comparison, other PRC2 subunits do not show a comparable expression dynamic (Fig S4). As EB are a mixed population of cells with different identities, we asked whether this dynamic could represent the expression of PCL proteins in PGC. We therefore downloaded expression data of in vitro directed differentiation of mESC to PGC via an intermediate state of primed pluripotency represented by Epiblast-like cells (EpiLC)\textsuperscript{24}. Strikingly, we detect a significant downregulation of Mtf2 as mESC exit pluripotency and a subsequent significant upregulation to mESC-like levels upon acquisition of PGC identity, as defined by Blimp1/Stella double expression (Fig 3c). Similarly to EB, Phf19 expression is very low during the early stages of mESC differentiation, to then be strongly activated in Blimp1/Stella double-positive committed PGC.
Figure 2: LRT analysis of time-course bulk RNA-seq. a) PCA analysis, showing data reproducibility, the path of differentiation along the diagonal, and the starting difference of Eed\(^{-}\) cells from WT and Mtf2\(^{GT/GT}\). b) Temporal dynamic of gene expression for selected lineage markers, showing strong upregulation of meso-endodermal markers. Additional markers shown in Fig S2. c-d) Anatomy ontology enrichment among genes significantly upregulated (c) or downregulated (d) in LRT statistics. In (d) comparison not yielding significant results are not shown. p-values are corrected for multiple testing using Holm–Bonferroni correction.

Figure 3: PCLs expression in PGC differentiation (next page). a) Temporal dynamic of PGC marker genes, upregulated in WT EB at day 7. b) Temporal dynamic of Mtf2 and Phf19 expression, mirroring the expression pattern of Prdm14 (Mtf2) and the other PGC markers (Phf19). c) Expression of Mtf2 and Phf19 during directed mESC differentiation to PGC from 24. The expression pattern follows the one detected in EB. Mtf1 GT/GT embryoid bodies are enriched in mesoderm-committed cells.

To further dissect the heterogeneity of cell fates contained within our samples, we dissociated day 4 EB and performed single-cell RNA-seq, followed by unsupervised clustering of all the cells, and visualised them using t-SNE-based (t-Distributed Stochastic Neighbour Embedding) dimensionality reduction. We identified eight different clusters (Fig. 4a): one predominantly composed of WT cells (cluster 1), two of Eed\(^{-}\) cells (clusters 2 and 7) and two of Mtf2\(^{GT/GT}\) cells (clusters 3 and 5),
Figure 3: PCLs expression in PGC differentiation (next page). a) Temporal dynamic of PGC marker genes, upregulated in WT EB at day 7. b) Temporal dynamic of Mtf2 and Phf19 expression, mirroring the expression pattern of Prdm14 (Mtf2) and the other PGC markers (Phf19). c) Expression of Mtf2 and Phf19 during directed mESC differentiation to PGC from 24. The expression pattern follows the one detected in EB.

Mtf1<sup>GT/GT</sup> embryoid bodies are enriched in mesoderm-committed cells

To further dissect the heterogeneity of cell fates contained within our samples, we dissociated day 4 EB and performed single-cell RNA-seq, followed by unsupervised clustering of all the cells, and visualised them using t-SNE-based (t-Distributed Stochastic Neighbour Embedding) dimensionality reduction<sup>25</sup>. We identified eight different clusters (Fig. 4a): one predominantly composed of WT cells (cluster 1), two of Eed<sup>−/−</sup> cells (clusters 2 and 7) and two of Mtf2<sup>GT/GT</sup> cells (clusters 3 and 5),
with the remaining three showing variable contribution from the different genotypes (Fig 4b). Expression patterns of marker genes largely reflect the expression we identified in bulk RNA-seq, although not all genes were strongly detected (Fig 4c, Fig S5). It is however possible to identify a strong mesodermal (clusters 3 and 5) and endodermal (cluster 5) signature of the Mtf2GT/GT cells, a pattern also followed by the small, Eed⁻/⁻-specific cluster 7 (Fig 4c, Fig S5). These three clusters also show higher expression of keratin genes, in concordance with the bulk RNA-seq data. It is interesting to note that for multiple genes showing comparable expression in the two PRC2 mutant in the bulk RNAseq data, such as T, Eomes and Msx1, single cell analysis shows strong expression in few Eed⁻/⁻ cells while the majority of Mtf2GT/GT ones is affected (Fig 4c, Fig S5), suggesting that MTF2 presence is pivotal for repression of master regulators of the somatic mesoderm gene network. To more widely characterize the identity of these clusters we performed anatomy term enrichment analysis, and we indeed identified specific signatures for all the genotype-specific clusters (Fig 4d). On the contrary, those composed of mixed genotype populations (clusters 0, 4, and 6) had little to no anatomy term overrepresentation, and cluster 4 showed a trend toward higher expression of pro-apoptotic markers (Fig S6). Of note, this cluster has a large contribution from the Mtf2GT/GT population, resembling the finding in undifferentiated mESC, where the Mtf2GT/GT mESC showed enrichment of apoptosis-related biological processes (Fig 1c). Other clusters containing Mtf2GT/GT cells (cluster 3 and 5) show a strong and highly significant enrichment for meso-endodermal genes, together with one of the Eed⁻/⁻-specific clusters, number 7. The other Eed⁻/⁻ cluster, cluster 2, is instead mostly enriched for terms related to early embryogenesis, as is the only WT-specific cluster, cluster 1.

Finally, we asked whether we could detect differences in the specification of PGCs, at the single cell level. Despite the earlier time point, when compared to the strong pattern detected in day 7 bulk RNA-seq, we detect expression of all the four marker genes used to define PGC identity (Fig S7). It is interesting to note that it is possible to detect a strong downregulation of these genes at day 4 in all 3 clusters dominated by Mtf2GT/GT cells, with virtually no cells expressing Prdm14 and Stella. Despite the detectable expression of these markers in several clusters, the only two expressing all the four markers are cluster 1 and 2, exclusively composed of WT and Eed⁻/⁻ cells, respectively.

**Discussion**

The role of MTF2 in PRC2 recruitment and function has been extensively studied in mESC, and its relevance in more differentiated cell types has been shown in multiple systems⁵-⁹. We investigated the molecular requirement of MTF2 during undirected mESC early commitment, and identified a strong commitment of Mtf2GT/GT EB toward the mesodermal lineage. At the bulk level this pattern is also detectable in the Eed⁻/⁻ cells, although the enriched anatomy category are characterized by broader terms, which may be the result of the larger number of deregulated genes, already at the mESC stage. These two characteristics suggest that MTF2 is responsible for the repression of a specific subset of PRC2 targets, most likely those regulating the establishment of somatic mesoderm identity. This is supported by the larger fraction of single cells showing strong upregulation of key mesodermal markers in clusters enriched for Mtf2GT/GT cells when compared to Eed⁻/⁻-specific clusters, which instead show a broader spectrum of deregulations in bulk, and a necessary to confirm these observations, but, if confirmed, this trend could suggest that MTF2
Figure 4: scRNA-seq of day 4 EB. a-b) tSNE visualization of scRNA-seq data, coloured according to cluster number (a) or genotype (b). There is one WT-specific (1), two Eed<sup>-/-</sup>-specific (2, 7), two Mtf2<sup>GT/GT</sup>-specific (3-5), and three mixed (0, 4, 6) clusters. b) Expression of lineage marker genes in scRNA-seq, showing the strong signature of Mtf2<sup>GT/GT</sup> cells. Most cells in clusters 3 and 5 show high expression of mesodermal genes, accompanied by non-neural ectoderm (cluster 3, 5) and endodermal (cluster 5) marker genes. A similar but less strong pattern is visible in the Eed<sup>-/-</sup>-specific cluster 7. d) Anatomy enrichment analysis of cluster-defining genes, showing meso-endodermal signature for clusters 3, 5 and 7, and enrichment for early development terms for clusters 1 and 2.
removal of PRC2 repression would disrupt regulation of cell identity in a more chaotic way, dampening the overall differentiation ability of the cells without biasing it toward any specific lineage. This is also supported by the sustained expression of pluripotency markers and the earlier lethality of Eed knockout mice compared to Mtf2 mutants.

The high enrichment of mesodermal anatomy terms in differentiating Mtf2^{GT/GT} cells, together with the re-expression of Mtf2 during WT PGC specification, and the abolished expression of PGC marker genes in the Mtf2 mutant EB strongly suggest a role of Mtf2 as gate-keeper of the somatic mesoderm transcriptional program, in at least three occasions: I) during mESC self-renewal, to support pluripotency and avoid premature exit from pluripotency; II) upon early differentiation and fate restrictions, when MTF2 is required to prevent mesodermal specification and allow development of other tissues such as the nervous system; III) Finally, during PGC specification, when the somatic transcriptional programs need to be silenced to allow re-establishment of pluripotency as required for germ-cell lineage establishment. In light of this role of MTF2 in controlling mesodermal specification, it is interesting to consider the peculiar differentiation path of PGC, which after an initial fate commitment block the expression of somatic transcriptional network to re-acquire pluripotency. This pattern fits the expression of Mtf2, high in the two "pluripotent" states where mesodermal identity is actively repressed, and downregulated during mesoderm specification. Furthermore, given the drastic and extensive DNA de-methylation wave that accompanies PGC specification and the specific role of MTF2 as a PRC2 recruiter to unmethylated CpG islands, it is tempting to speculate that Mtf2 re-expression could be required to keep unwanted transcription of alternative developmental programs under control, thus acting as an ‘epigenetic parachute’ for the PGC demethylated genome. Finally, the strong upregulation of PHF19 upon PGC commitment will need to be further investigated. PHF19 is mutually exclusive with MTF2 in PRC2.1 composition, and thus competes with it for binding to the core PRC2 proteins. Its high affinity for H3K36me3, a marker of actively transcribed genes, however, might complement DNA-directed MTF2-mediated PRC2 recruitment, which is instead primarily targeted at unmethylated, CpG island-containing promoters.

Material and methods

Cell culture and differentiation

E14, Mtf2^{GT/GT} and Eed^{-/-} mESC were cultured as in. To induce differentiation 10^5 mESC per well were plated in Nunclon Sphera 6-well plates (Thermo Fisher Scientific) in mESC medium without LIF, and allowed to differentiate. After the first two days half of the culture medium was replace daily by pipetting and without removing the cells. mESC control samples were obtained by plating cells in the same condition in normal mESC medium with LIF, and harvested after two days.

RNA extraction and bulk RNA-seq

mESC/EBs were harvested by pipetting, and by washing the well with 5ml PBS to ensure complete collection. Cells were pelleted by gentle centrifugation (100g, 5 min) and directly dissolved in 1ml Trizol (Ambion). After isolation of the aqueous phase RNA was purified using RNeasy Kit (quiagen) following manufacturer protocol. Integrity of purified RNA was checked on an Agilent Bioanalyzer.
using the RNA 6000 Pico Kit. Intact RNA was depleted of rRNA and prepared for sequencing using the KAPA RNA HyperPrep Kit with RiboeRase (HMR) and the resulting library size-selected using e-gel (Thermo Fisher Scientific). Library were sequenced on an illumine NextSeq 500.

**Single-cell RNA library preparation**

The SORT-seq protocol used was adapted from the CEL-Seq2 method \(^{31}\), to facilitate FACs sorting of single-cells onto 384-well plates. Plates were spun down (1200g, 2mins, 4°C) and ERCC spike-in mix (1:50,000) was dispensed by the Nanodrop (BioNex Inc) into each well. 150nl of the Reverse Transcription (RT) mix was similarly dispensed into each well using the machine and the thermal cycling conditions were set at 4°C 5min; 25°C 10min; 42°C 1hr; 70°C 10min. Contents from the plates were pooled together and the cDNA was purified using AmpureXP (New England BioLabs) beads. In-vitro transcription (Ambion MEGA-Script) was then carried out overnight at 16°C, with the lid set at 70°C. An exonuclease digestion step was performed thereafter for 20mins at 37°C, followed by fragmentation of the RNA samples. After a beads cleanup, the samples were subjected to library RT and amplification to tag the RNA molecules with specific and unique sample indexes (Illumina), followed by a final beads cleanup (1:0.8, reaction mix: beads) and the sample cDNA libraries were all eluted in 20uL of DNase free water. Libraries were quantified using by a qPCR approach and sequenced on the NextSeq 500(Illumina) for 25 million reads per plate.

**Bioinformatic analysis**

Paired-end Illumina 75-bp sequencing files from Bulk RNAseq were mapped to the M16 release of GCRm38 annotation using STAR (version 2.5.3a) \(^{32}\), while also performing gene counts. Raw count tables were used as input for DEseq2 (version 1.18.1) \(^{33}\), using Wald statistics for mESC pairwise comparison and Likelihood Ratio Test statistics to identify statistically different temporal patterns. In both cases a significance cut off at 0.01 FDR was applied, and for Wald statistics a minimum of log2 fold change of 1 was requested. Gene Ontology enrichment analysis was performed with clusterProfiler (version 3.6.0) \(^{34}\). Anatomy ontology enrichment was performed MouseMine web interface \(^{35}\).

Raw reads from single-cell RNAseq were mapped and aligned to the mouse genome GRCm38/mm10 database using the Bowtie2 (version 2.3.4.2)36. Aligned reads were indexed and the final count table was derived using HTseq (version 0.9.1) 37. R package Scater (version 3.7) 38 was used for sample filtering and quality check, to remove dropouts (cells with < 5 reads) and cells with too few recovered genes (<500). The data was than analysed using R package Seurat (version 2.3.4) 39 for batch, read counts and gene counts normalization. Thereafter, the filtered and normalized reads were subject to a Principal Component Analysis using the package’s inbuilt functions. The cells are then clustered using a shared nearest neighbor (SNN) modularity optimization-based clustering algorithm (Seurat, version 2.3.4)40. First, k - nearest neighbors and SNN graphs were calculated, followed by an optimization step using the modularity function.
References


Supplementary Figure 1: *in situ* hybridization against *Mtf2* transcript. In all images it is possible to identify signal in the central nervous system. **a**) mid-sagittal section of E11.5 embryo. Magnification (**a'**) shows strong signal in segmenting tail. **b**) Parasagittal section of E13.5. Magnification (**b'**) shows details of *Mtf2* expression in developing limb. **c**) Mid-sagittal section of E13.5. (**c'**) shows details of expression in extraembryonic tissues, (**c'**) in the thymus. **d**) Mid-sagittal section of E15.5. Magnification (**d'**) shows strong signal in the thymus. Image credit: 2008 Allen Institute for Brain Science. Allen Developing Mouse Brain Atlas. Experiments No. 100056771, 100058493,100085137 (http://www.brain-map.org/search/index.html?query=mtf2&sp=Mouse&ag=Prenatal&tr=ISH&fa=false&e_sp=t&e_ag=t&e_tr=t&e_fa=t)
Supplementary Figure 2: Temporal expression dynamic of marker genes for different cell fates. Y-axis represents normalized counts. Asterisks (*) indicate significant differentially expressed gene under LRT statistics.
**Supplementary Figure 3:** Temporal expression dynamic of marker genes for HSC identity. Y-axis represents normalized counts. Asterisks (*) indicate significant differentially expressed gene under LRT statistics.

**Supplementary Figure 4:** Temporal expression dynamic of core PRC2 components, PRC.1 and PRC.2 specific subunits. Y-axis represents normalized counts.

**Supplementary Figure 5** (next page): Violin plot of scRNA-seq counts of marker genes for different cell fates. Y-axis represents normalized counts.
Supplementary Figure 6: Violin plot of scRNA-seq counts of pro-apoptotic marker genes. Y-axis represents normalized counts.

Supplementary Figure 7: Violin plot of scRNA-seq counts of PGC marker genes. Y-axis represents normalized counts.
**Supplementary Figure 6:** Violin plot of scRNA-seq counts of pro-apoptotic marker genes. Y-axis represents normalized counts.

**Supplementary Figure 7:** Violin plot of scRNA-seq counts of PGC marker genes. Y-axis represents normalized counts.
Chapter 6

Summary and Discussion
After an extended description of the role of the epigenetic environment in the determination of cell identity, and of the dynamics of chromatin during embryonic development (Chapter 1 and Chapter 2), the work of this thesis focuses on the dissection of molecular mechanisms involved in Polycomb recruitment in mouse embryonic stem cells (mESC, Chapter 3 and Chapter 4), and the requirement of the drivers of such mechanisms during the exit from pluripotency and early differentiation (Chapter 5).

Our work started with an effort aimed at the identification of the molecular mediator of Polycomb Repressive Complex 2 (PRC2) recruitment. PRC2-mediated deposition of H3K27me3 is an extremely effective, yet very flexible and precise mechanism to repress gene expression, employed across phyla, from plants to metazoans (Huang et al., 2017; Whitcomb et al., 2007). Despite a vast body of structural, biochemical, molecular, epigenetic, and developmental knowledge acquired on Polycomb biology during decades of research, a specific, yet fundamental, question had escaped finding an answer: How does PRC2 identifies its targets? What initiates this cascade of wonderfully balanced events, critical for the development of a single cell into a sentient being?

Several studies over the years have shed some light on this question. We know that PRC2 does not fulfil this task alone, but functionally interacts with several other chromatin modifiers, with PRC1 being the most studied. This field has also evolved with time, switching from a purely hierarchical view of PRC1 depending on PRC2 (this is still today called “canonical” PRC1 recruitment) to a more “interactive” model of cooperation between the complexes (Blackledge et al., 2014; Di Croce and Helin, 2013; Zilberman et al., 1995).

We have learned that facultative subunits can influence binding, catalytic activity, and interaction partners of the core PRC2 (Vizan et al., 2015). As a consequence, two different, mutually exclusive flavours of PRC2 have been identified, characterized by the presence of either one of the PCL proteins (PRC2.1, potentially with additional partners) or by binding with JARID2 and AEBP2 (PRC2.2) (Hauri et al., 2016; Holoch and Margueron, 2017).

In organisms such as Drosophila, the reduced complexity of the system has led to the identification of an handful of genomic regions, called Polycomb Response Elements (PRE), whose sequence is able to initiate Polycomb recruitment (Kassis and Brown, 2013; Muller and Kassis, 2006). These sequences, however, are hundreds of bases long, making difficult to pinpoint the molecular mechanism of recruitment and explain PRC2 specificity for these regions.

Drosophila PREs do not seem to have equivalent recruitment elements in vertebrates, with only the REST complex having shown some sequence-mediated recruitment at a minor fraction of PRC2 targets (Arnold et al., 2013; Dietrich et al., 2012; van Heeringen et al., 2014). Yet, the DNA sequence underlying PRC2 targets contains sufficient information to encode the positioning of H3K27me3 domain, and this information can be extracted by a machine learning algorithm and used to predict the location of PRC2 targets across vertebrates (van Heeringen et al., 2014).

**MTF2-mediated PRC2 recruitment**

We approached the PRC2 recruitment puzzle by combining the information coming from machine learning, the genome wide binding locations of PRC2, and their conservation across vertebrate species, and used this information to shortlist potential PRC2 recruiting sequences. When
experimentally tested, one of these sequences led to the identification of MTF2 as a DNA binding protein required for PRC2 recruitment in mESC (Chapter 3), as confirmed by independent studies (Li et al., 2017; Oksuz et al., 2018). Furthermore we identified the cues that drive MTF2 specificity, which is achieved through a hybrid readout of DNA sequence and shape. Crystallography data further supported our findings, showing that while MTF2 is able to establish direct contact with the bases of CG-containing sequences, this binding is complemented by interaction of additional residues with the DNA backbone (Li et al., 2017), a feature that we interpret as the DNA shape-reading component of MTF2 specificity. As MTF2 represents only one of the potential PRC2 recruitment mechanisms, and as it is present in only 40% of the PRC2 complexes in mESC (Kloet et al., 2016), we further investigated whether MTF2 is able to bind DNA independently, or depends on the association with PRC2 core or on potential interactions with PRC2. This resulted in the identification mutual interdependency of MTF2 and JARID2 and the classification of MTF2 targets in two categories: the primary targets, where MTF2 is able to bind independently, and the secondary ones, where it mostly depends on the presence of core PRC2 (Chapter 3).

**Cross-regulation of Polycomb complexes recruitment**

Intrigued by the unexpected finding about the reciprocal influence of MTF2 and JARID2 on binding, we followed up on the characterization of PRC2 interactions by extending our panel of knock-out lines and including the treatment with inhibitors of both PRC1 and PRC2. We reanalysed and generated new ChIP-seq datasets for five factors (EZH2, MTF2, JARID2, H3K27me3, H3K4me3, RING1B) in five lines (WT, Mtf2GT/GT, Eed−/−, Jarid2−/−, Ring1a/b−/−), either treated with inhibitors or not, for a total of 27 different conditions (Chapter 4). This extended analysis highlighted and clarified several important conceptual points. PRC2 recruitment mainly happens through two mechanisms, a DNA-MTF2 mediated branch and a PRC1-H2AK119ub-JARID2 mediated one, and simultaneous removal of both results in near-input level of EZH2 recruitment.

Both these pathways are enhanced by EED binding to H3K27me3, but the relevance of this positive feedback loop is much more important for the latter, where the combined removal of JARID2 and H3K27me3 results in a synergistic effect on PRC2 recruitment, suggesting that they can compensate for each other when only one of the two is absent.

PRC2.1 and PRC2.2 also mutually affect each other’s recruitment. This, though, is only partially relying on H3K27me3, a puzzling finding when put in the context of previous finding showing that MTF2 and JARID2 association with PRC2 core are mutually exclusive and that JARID2 removal minimally affect PRC2 core recruitment. This could be explained be the formation of non-functional hybrid complexes including MTF2 and AEBP2 in the Jarid2−/− line, similarly to what previously was shown in the AEBP2 mutant, the only case where MTF2 and JARID2 have been found in the same complex. It is also conceivable that the absence of either factor could be disrupting scaffolding effects of RNA molecules, or, alternatively, that the simultaneous presence of both sub-complexes might be required to keep the chromatin environment balanced toward PRC2 repression, for example increasing the resistance to competing activating factors.

Both the RNA and the chromatin environment hypothesis also helps to explain the last and probably most unexpected but exciting finding, namely the fact that PRC1 recruitment can be
affected by the absence of either MTF2 or JARID2 in cells already depleted of H3K27me3. Although reinforcing the concept that PRC2 and PRC1 can influence each other’s recruitment, this finding exposes a gap in our current understanding of the biochemistry of these interactions. Indeed, no attempt to identify a direct PRC1-PRC2 contact has so far been successful, leading to the widely accepted assumption that the interaction is exclusively mediated by binding to the histone tails modified by the other complex. While we do confirm this mechanism, we expose the presence of H3K27me3-independent mechanism(s) of RING1B recruitment to chromatin that rely on MTF2 and JARID2. Additional efforts, however, will be necessary to clarify the biochemical details of these mechanisms.

**MTF2 role in pluripotency and early differentiation**

Finally, we investigated the role of MTF2 in mESC during the exit from pluripotency and early differentiation by comparing gene expression of WT, Mtf2\textsuperscript{GT/GT}, and Eed\textsuperscript{-/-} mESC and embryoid bodies at different stages of differentiation, both in bulk and at the single cell level (Chapter 5). We identified a strong somatic mesoderm signature among the genes de-repressed in the absence of MTF2, while the overall absence of PRC2 repression in the EED mutant seems to induce a more chaotic perturbation, interfering with cell differentiation without a bias toward a specific lineage. Finally, we noticed a re-expression of MTF2 at later stages of differentiation of WT embryoid bodies. This pattern was followed by MTF2-dependent re-expression of the pluripotency-associated gene Prdm14, and accompanied by the increased transcription of marker genes of Primordial Germ Cell (PGC) identity (Saitou and Yamaji, 2012), and a steep rise in the level of Phf19, a PCL-family gene closely related to Mtf2. Strikingly, despite the highly heterogeneous nature of embryoid bodies, MTF2 and PHF19 expression pattern closely track the one detected during directed differentiation of mESC to PGC (Sasaki et al., 2015). Further validation of these findings will be required, but this scenario suggests that MTF2 functions as a lineage gate-keeper with multiple functions during embryogenesis: I) in mESC it may be required for self-renewal and prevention of premature exit from pluripotency; II) during early differentiation it appears to repress mesodermal fate, ensuring a balanced proportion of the three germ layers; III) during PGC specification it may be required for the downregulation of the somatic program and the re-expression of pluripotency factors, likely required by the highly proliferative nature of PGC.

**Concluding remarks and future perspectives**

The biochemical complexity and the biological relevance of the Polycomb system, shown by the lethality of knockout animal models, has for long time led to a strong focus on experiments performed in vitro or in extremely simplified systems, often with the exclusive use of purified recombinant proteins in very controlled conditions. While these studies have been instrumental to build the structural and biochemical knowledge on which more recent research stands, they have removed the object of their study from its physiological environment, native chromatin. This has gained us fundamental insights, but at the cost of losing biological complexity.

Recent technological advancement and increased throughput of genomics, transcriptomics and proteomics techniques has led to an explosion in the number of known sub-stoichiometric proteins, their association dynamics, and classes of biologically relevant sub-complexes for both PRC1 and PRC2. Integration of the power of high throughput approaches and advanced
bioinformatic techniques, coupled with the information about DNA shape-reading properties of some proteins coming from a different field of biology, has led to important discoveries that for the first time started answering questions that have remained open for decades, as shown in Chapter 3.

Advancements in targeted genome editing techniques have allowed the careful dissection of regulatory mechanisms, by assessing the effect of single and multiple perturbations on Polycomb complex function in their native environment, the chromatin template within a cell nucleus. This has allowed the identification of previously unknown relationships among Polycomb complexes, while at the same time pointing out areas of investigation where our understanding is far from complete, as discussed in Chapter 4.

Finally, Polycomb repression is pivotal because it controls transitions between cell identities, and fine tuning of Polycomb composition and activity plays an important role in this process. Indeed, in the case of PRC2.1, for example, different PCL proteins can associate with the core complex during differentiation, resulting in the redirection toward different target genes. We initiated to study the results of these modulations in the absence of MTF2 in chapter 5.

Beside the value of the particular biological insights provided by specific studies, what recent research in Polycomb biology has showcased are the power and potential of combining orthogonal approaches, that integrate large amounts of information from different sources. A major milestone of the research in molecular epigenetics in general, and Polycomb biology in particular, will be the detailed molecular description of all the regulatory steps driving lineage choices. This will help to gain a mechanistic understanding of how totipotent cells differentiate into the exact required cell type, at a specific point in time, in a specific location in the body, while synchronizing the process with neighbouring cells and tissues. No word can accurately explain the complexity of these processes, but this can only further underline the pivotal role of interdisciplinary approaches in future research aiming at better understanding the regulation of embryonic development.
References


Chapter 7

Epilogue
Samenvatting

Elk multicellulair organisme, zelfs de meest eenvoudige, bezit de mogelijkheid om cellen met hele verschillende en gespecialiseerde functies te genereren met behulp van dezelfde genetische informatie. Dit wordt gerealiseerd door te controleren hoe, wanneer en welk deel van deze ‘blueprint’ wordt gebruikt door middel van epigenetische regulatie. Via een complex systeem worden signalen uit de omgeving en van het genetisch materiaal geïntegreerd om vanaf hetzelfde DNA cellen te genereren met verschillenden identiteiten en functies. Dit alles gebeurt in een volledige harmonie zodat een volledig organisme kan worden gevormd. Hoofdstuk 1 en 2 zijn toegewijd aan een uitgebreide introductie in epigenetica, de regulatie hiervan en hoe deze dynamisch zijn gedurende de ontwikkeling van het embryo. De focus zal hier grotereels liggen op Polycomb Repressive Complex 2 (PRC2) en hoe dit gerekurtord wordt.

De plaatsing van histon 3 lysine 27 trimethylatie (H3K27me3) is een heel effectief en toch flexibel mechanisme om de repressie van genen te bewerkstelligen. Dit mechanisme is dan ook van planten tot gewervelde dieren en mensen geconserveerd (Huang et al., 2017; Whitcomb et al., 2007). Ondanks een enorme hoeveelheid van structurele, biochemische, moleculaire en epigenetische inzichten in Polycomb tijdens de ontwikkeling door tientallen jaren onderzoek is het lang onbekend gebleven hoe PRC2 zijn doelen herkent en vindt. De fundamentele vraag blijft hoe een samenwerking van verscheidene gebalanceerde gebeurtenissen, die essentieel zijn voor de ontwikkeling van een organisme, geïnitieerd worden.

In Hoofdstuk 3 beschrijven we hoe de combinatie van ivan computationele analyse, de binding van PRC2 op het genoom en de conservatie tussen verschillende organismen leidde tot enkele kandidaat DNA-sequenties die PRC2 kunnen rekruteren. Experimentele testen op deze sequenties resulteerde in de identificatie van MTF2 als een DNA-bindend eiwit dat nodig is om PRC2 te rekruteren in embryonale stamcellen (ESCs). Deze observatie is ook bevestigd in onaftankelijke onderzoeken (Li et al., 2017; Oksuz et al., 2018). Daarnaast hebben we de benodigdheden geïdentificeerd van de specifiteit van MTF2 voor DNA bewerkstelligen, namelijk een combinatie van DNA-sequentie en de DNA-vorm. Dit is verder onderbouwd door informatie van de kristalstructuur van MTF2 waaruit duidelijk werd dat MTF2 direct contact maakt met de basen van CG-bevattende sequenties, hetgeen versterkt wordt door interactie met extra residuen van de ‘ruggengraat’ van het DNA (Li et al., 2017). Dit geheel interpreteren wij als de DNA-vorm-specificiteit van MTF2. Aangezien MTF2 maar een van de meerdere mogelijke PRC2-rekruterende mechanismen is, en maar in 40% van de PRC2 complexen voor komt (Kloet et al., 2016), hebben we verder bekeken of MTF2 op zichzelf aan DNA kan binden of dat dit af hangt van interactie met de kern van PRC2 of potentiële interacties met PRC2.2. Dit leidde tot de identificatie van twee types MTF2 doel-locaties: degenen die wij aanduiden als ‘primair’, waar MTF2 op zichzelf kan binden, en de ‘secundaire’, waar de associatie van MTF2 met DNA af hangt van de aanwezigheid van de kern van PRC2.

In hoofdstuk 4 gaan wij hierop verder door middel van meerdere cellijnen waarin onderdelen van PRC1 en PRC2 zijn verwijderd. Dit completeren wij met chemische inhibitie van PRC1 en PRC2. We hebben hiervoor de locaties van zes verschillende factoren op het genoom geanalyseerd (EZH2, MTF2, JARID2, H3K27me3, H3K4me3, RING1B) in vijf cellijnen (controle, Mtf2GT/GT, Eed/-/-, Jarid2/-/-, Ring1a/b/-/-) al dan niet in combinatie met chemische inhibitie, resulterend in een totaal
van 27 verschillende condities. Deze uitgebreide analyse leidde tot nieuwe inzichten en verduidelijkte een aantal concepten van PRC2 reclutering. PRC2 reclutering gebeurt door middel van twee mechanismen: een DNA-MTF2 gemedieerde tak en een PRC1-H2AK119ub-JARID2 tak. Het tegelijk verstoren van deze twee takken resulteert in bijna volledige afwezigheid van EZH2 reclutering. Deze beide takken worden versterkt doordat EED bindt aan H3K27me3, maar dit is relevanter voor de laatstgenoemde. Hierin zorgt namelijk de gecombineerde afwezigheid van JARID2 en H3K27me3 in een synergistisch effect op PRC2 reclutering, hetgeen suggereert dat deze kunnen compenseren voor elkaar wanneer de ander weg is. PRC2.1 en PRC2.2 kunnen ook elkaars reclutering beïnvloeden. Dit is echter maar partieel gemedieerd door H3K27me3, wat ons verbaasde aangezien MTF2 en JARID2 niet tegelijk met de kern van PRC2 kunnen associëren en het weghalen van JARID2 nauwelijks effect heeft op PRC2 reclutering. Verder laten we zien dat de reclutering van PRC1 verstoord kan worden door de afwezigheid van MTF2 of JARID2 in cellen waar H3K27me3 al weg is. Al onderbouwt dit het concept dat PRC1 en PRC2 elkaars reclutering versterken, deze bevinding laten zien dat er een gat zit in onze huidige kennis van de biochemie onderliggend aan Polycomb interacties. Meer onderzoek zal nodig zijn om hier inzicht in te geven.

In Hoofdstuk 5 onderzoeken we de rol van MTF2 in embryonale stamcellen en tijdens de initiatie van differentiatie door het vergelijken van genexpressie profielen van controle, MTF2GT/GT en Eed/-/ ESCs en embryo-achtige lichamen op verschillende tijdstippen van differentiatie. Dit hebben we gedaan op populatieneiveau, maar ook op het niveau van enkele cellen. We hebben hiermee sterke mozaïek-mesoderme tekenen geïdentificeerd in de genen die gedereguleerd worden in de afwezigheid van MTF2. Dit is verschillend van wanneer de kern van PRC2 afwezig is, wat eerder leidt tot chaotische deregulatie. Dit uit zich door verstoorde differentiatie zonder dat er een duidelijke bias is richting een bepaalde type cellen. We observeerden dat MTF2 opnieuw tot expressie komt tijdens latere stadia van differentiatie. Dit ging vergezeld met re-expressie van de het pluripotentie-geassocieerde gen Prdm14, expressie van genen primitieve kiemcellen (PGC) kenmerken en een sterke op regulatie van Phf19, een PCL-familie eiwit dat sterk gerelateerd is aan Mtf2. Het is interessant om op te merken dat ondanks de sterke heterogeniteit van embryo-achtige lichamen, het expressiepatroon van MTF2 en PHF19 sterk het patroon van ESCs die worden gedifferentieerd naar PGCs (Sasaki et al., 2015). Verdere validatie van deze bevindingen is nodig, maar dit scenario suggereert dat MTF2 functioneert als een poortwachter met verschillende functies tijdens de vorming van het embryo: I) in ESCs is MTF2 nodig voor constante hernieuwing van ESCs en het voorkomen van premature differentiatie; II) tijdens vroege differentiatie onderdrukt MTF2 mesoderm programma’s, hetgeen zorgt voor een gebalanceerde verdeling van de drie kiem lagen; III) tijdens PGC specificatie en afzondering vanaf het mesoderm programma is MTF2 nodig voor het onderdrukken van het somatische mesoderm programma en re-expressie van pluripotentie factoren, wat waarschijnlijk nodig is voor de snel-delende PGCs.
References


**Curriculum vitae**

Matteo Perino was born in Rome, Italy, on September 5th, 1988. After completing scientifically oriented high school he studied Biotechnology at the University of Roma Tor Vergata, where he obtained his BSc with full marks and honours (110 e lode).

He next moved to Milan where he obtained his MSc in Medical, Molecular and Cellular Biotechnology from Università Vita-Salute S.Raffale with full marks and honours (110 e lode), with an experimental thesis on direct reprogramming of fibroblasts to dopaminergic neurons, which he further pursued with a scholarship from the Telethon Foundation after graduating.

In October 2013 he started his PhD in the DevCom consortium under the supervision of Gert Jan Veenstra where he studied the molecular details of Polycomb Repressive Complex 2 recruitment in mouse embryonic stem cells and *Xenopus Tropicalis* embryonic development.

Since April 2018 he is a postdoctoral fellow at EMBL where he studies regulation of zygotic genome activation in *Drosophila*.


**Publication list**


Acknowledgments

This section can only be started by thanking Gert Jan for organising DevCom and offering me a position in his group and within this wonderful consortium. You have been a great mentor and it is difficult to explain how much I learned during these years. We had our discussions, but I would choose the same position a thousand times again.

Speaking of learning, Simon I can’t forget that when I joined the lab I didn’t write a single line of code in my life. Yet. Your explanations, tutorial, workshops, suggestions, and examples are how I not only learned to code, but also to enjoy coding, eventually coming to consider myself a hybrid wet/dry biologist.

I wouldn’t have reached this point, without the influence and help of everyone in DevCom. A special mention goes to Saartje and Marion for managing everything and making it happen in real life. Ann, Ensieh, Giulia, Ines, Lindsay, Marta, Rita, Rosa, Bilal, Jamie, Matthew, Panos, it has been an amazing experience to do science while travelling Europe with you.

While I did travel a lot with DevCom, my daily lab life was of course in Nijmegen. Saartje, Ila, Siebe, Sarita, thank you so much for teaching me everything practical I know about working with frogs. Ila, Ann, the nights collecting embryos would have been much longer without you. Ron, thank you for taking such good care of our frogs. George, your patience while helping me debugging scripts would, alone, make you deserve a medal. Georgina, the sparks of Latin vibe you brought to the lab during the Dutch winters were wonderful. Chet, knowing the projects I left are in such good hands is really priceless. Hediche, Emil, Aiko, Madeleine, Sandra, it was great to meet and work with you. Supervising you taught me much more than you can imagine.

It has been a PhD of firsts, and mass spectrometry was one of them: Michiel it has been great to learn it from you and your group. Susan, Cristina, Ino, Matthew, Arne, Rick, Lisa, Pascal, Marijke, Nelleke, Raghu, thanks for the infinite number of advices, suggestions, tips and tricks you all have constantly provided. Hendrik, Guido, thank you for the countless brainstorming and discussions on Polycomb and mES. Guido, what I know about stem cells comes from you, it has been great to work together, and I hope the future will see us collaborate again.

Leonie, Bilge, Naomi, Karolina, Julien, Marco, Dei I really enjoyed our Friday meetings. Eva, Bowon, Kim, I couldn’t even have started thinking about writing a thesis without all the samples you sequenced, thank you so much. Maria, I’ve never been good with paperwork, but with you everything was easy. Thank you to everyone in the Mol(Dev)Bio departments for the general great atmosphere everyone contributed to.

Science was a large share of my PhD life, but it did not take all of my time in Nijmegen, were I found great friends. George, I can’t (and maybe is better not to :P) remember all the stories of these four years we could tell. From organising parties to daily housekeeping, from gossiping around to sharing personal situations, from cycling around carrying crutches to moving furniture up and down our steep and slippery stairs I couldn’t have imagined a better flatmate. Guido, Boris, Jakob, Roderick, George: nights out, brewing experiences, S.Patrick cross-country pub crawls, festivals, football games at Shamrock, if something fun was on schedule I could always count on you.

Jessie, Bowon, Sabine, Ian, Bilge, Naomi, Cristina, Matthew, Federica, the team was not complete without you and you never let us down. George, Guido, Ino, Lisa, Rick, it has been an honour (ad a lot of fun!) to share with you the duty on the borrel committee. Some of my most cherished memories in Nijmegen start with a successful borrel.

A special mention goes to the whole Fudoshin crew. Sharing a passion is a great way to make friends, and the nature of Judo can only help with it, but the atmosphere, enthusiasm, care,
Acknowledgments

This section can only be started by thanking Gert Jan for organising DevCom and offering me a position in his group and within this wonderful consortium. You have been a great mentor and it is difficult to explain how much I learned during these years. We had our discussions, but I would choose the same position a thousand times again.

Speaking of learning, Simon I can’t forget that when I joined the lab I didn’t write a single line of code in my life. Yet. Your explanations, tutorial, workshops, suggestions, and examples are how I not only learned to code, but also to enjoy coding, eventually coming to consider myself a hybrid wet/dry biologist.

I wouldn’t have reached this point, without the influence and help of everyone in DevCom. A special mention goes to Saartje and Marion for managing everything and making it happen in real life. Ann, Ensieh, Giulia, Ines, Lindsay, Marta, Rita, Rosa, Bilal, Jamie, Matthew, Panos, it has been an amazing experience to do science while travelling Europe with you.

While I did travel a lot with DevCom, my daily lab life was of course in Nijmegen. Saartje, Ila, Siebe, Sarita, thank you so much for teaching me everything practical I know about working with frogs. Ila, Ann, the nights collecting embryos would have been much longer without you. Ron, thank you for taking such good care of our frogs. George your patience while helping me debugging scripts would, alone, make you deserve a medal. Georgina, the sparks of Latin vibe you brought to the lab during the Dutch winters were wonderful. Chet, knowing the projects I left are in such good hands is really priceless. Hediche, Emil, Aiko, Madeleine, Sandra, it was great to meet and work with you. Supervising you taught me much more than you can imagine.

It has been a PhD of firsts, and mass spectrometry was one of them: Michiel it has been great to learn it from you and your group. Susan, Cristina, Ino, Matthew, Arne, Rick, Lisa, Pascal, Marijke, Nelleke, Raghu, thanks for the infinite number of advices, suggestions, tips and tricks you all have constantly provided. Hendrik, Guido, thank you for the countless brainstorming and discussions on Polycomb and mES. Guido, what I know about stem cells comes from you, it has been great to work together, and I hope the future will see us collaborate again. Leonie, Bilge, Naomi, Karolina, Julien, Marco, Dei, I really enjoyed our Friday meetings. Eva, Bowon, Kim, I couldn’t even have started thinking about writing a thesis without all the samples you sequenced, thank you so much. Maria, I’ve never been good with paperwork, but with you everything was easy. Thank you to everyone in the Mol(Dev)Bio departments for the general great atmosphere everyone contributed to.

Science was a large share of my PhD life, but it did not take all of my time in Nijmegen, were I found great friends. George, I can’t (and maybe is better not to :P) remember all the stories of these four years we could tell. From organising parties to daily housekeeping, from gossiping around to sharing personal situations, from cycling around carrying crutches to moving furniture up and down our steep and slippery stairs I couldn’t have imagined a better flatmate. Guido, Boris, Jakob, Roderick, George: nights out, brewing experiences, S.Patrick cross-country pub crawls, festivals, football games at Shamrock, if something fun was on schedule I could always count on you. Jessie, Bowon, Sabine, Ian, Bilge, Naomi, Ino, Cristina, Matthew, Federica, the team was not complete without you and you never let us down. George, Guido, Ino, Lisa, Rick, it has been an honour (ad a lot of fun!) to share with you the duty on the borrel committee. Some of my most cherished memories in Nijmegen start with a successful borrel.

A special mention goes to the whole Fudoshin crew. Sharing a passion is a great way to make friends, and the nature of Judo can only help with it, but the atmosphere, enthusiasm, care,
dedication and fun I found within this group simply left me speechless. Plus, most of the (little) Dutch I know I owe it to you. Thank you all very much!

Tutto questo non sarebbe stato possibile senza il supporto di chi conosco da una vita. **Lorenzo, Jacopo, Alessandro, Valerio, Luca, Goro, Manuel, Leonardo**, sapere di poter contare su di voi, di trovarvi ogni volta che fossi passato da Roma, mi ha dato la sicurezza di avere le spalle coperte di cui avevo bisogno per guardare così avanti.

**Mamma, Papà**, so che non deve essere stato facile vedere un figlio andare all’estero e perderne la quotidianità. So anche di avere la tendenza a farmi sentire poco, e me ne dispiaccio ogni volta che ci penso. Ricevere il vostro appoggio incondizionato e vedere l’orgoglio brillarvi negli occhi ad ogni successo che ricordi, dal più piccolo al più grande, dal più vecchio al più recente, è quello che mi ha dato il coraggio necessario per arrivare qui.

**Giulia**, quando sono partito stavamo insieme da poco più di tre anni, non sapevamo cosa ci riservasse il futuro, né se saremmo arrivati qui ancora insieme. Nonostante questo sei stata la mia prima e più convinta sostenitrice, spingendomi a partire ogni volta che hai notato la minima esitazione. Mi sei stata accanto in ogni momento di nostalgia, malinconia o sconforto. Mi hai dato idee e spunti scientifici nei momenti di difficoltà come la più brillante dei colleghi. Abbiamo festeggiato insieme ogni traguardo, comune o personale. Non è stato sempre tutto facile, accanto agli alti abbiamo avuto i nostri bassi, il dubbio, quando non la certezza a volte, che non avrebbe funzionato. Siamo cresciuti e cambiati, ma lo abbiamo fatto insieme, e dopo nove anni insieme posso dire che non c’è persona che avrei preferito accanto in questo viaggio. C’è un messaggio per te in questa tesi, semplice, forse banale, ma totalmente sincero. Buona ricerca ;)