nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0218-5

Promoter bivalency favors an open chromatin architecture in embryonic stem cells

Glòria Mas[1,2], Enrique Blanco[1], Cecilia Ballaré[1], Miriam Sansó[1,3], Yannick G. Spill[1,4], Deqing Hu[5,6], Yuki Aoi[5,6], François Le Dily[1], Ali Shilatifard[5,6], Marc A. Marti-Renom[1,4,7,8*] and Luciano Di Croce[1,7,8*]

In embryonic stem cells (ESCs), developmental gene promoters are characterized by their bivalent chromatin state, with simultaneous modification by MLL2 and Polycomb complexes. Although essential for embryogenesis, bivalency is functionally not well understood. Here, we show that MLL2 plays a central role in ESC genome organization. We generate a catalog of bona fide bivalent genes in ESCs and demonstrate that loss of MLL2 leads to increased Polycomb occupancy. Consequently, promoters lose accessibility, long-range interactions are redistributed, and ESCs fail to differentiate. We pose that bivalency balances accessibility and long-range connectivity of promoters, allowing developmental gene expression to be properly modulated.

Embryonic stem cells (ESCs) can differentiate into all the cell types of the organism, a process that relies on the complex interplay between transcription factors, epigenetic regulators, and signaling molecules. The coordinated actions of such mechanisms regulate with exquisite temporal and spatial precision the developmental gene expression programs that dictate cell fate. The exceptional transcriptional plasticity of ESCs is achieved by a unique chromatin landscape and structure[1]. One feature that characterizes (but is not restricted to) the chromatin landscape of ESCs is the presence of 'bivalent promoters'. Bivalent promoters are regions of epigenetically modified nucleosomes that carry histone H3 trimethylated at lysine 4 (H3K4me3) and lysine 27 (H3K27me3), which are associated with gene activation and repression, respectively[2]. Genome-wide mapping of histone modifications in ESCs showed that bivalently marked genes are essential for developmental processes and lineage specification[2–5]. The histone methyltransferases responsible for catalyzing the bivalent marks H3K27me3 and H3K4me3 are the Polycomb repressive complex 2 (PRC2) and the Trithorax MLL2/COMPASS (KMT2B; WBP7), respectively[6–9]. The majority of bivalent promoters are also co-occupied by the Polycomb repressive complex 1 (PRC1)[10].

Bivalent chromatin is essential for embryogenesis, as loss or mutation of Polycomb and MLL2 complexes leads to developmental defects or embryonic lethality[6,11–13]. Despite its importance, the role of bivalency at promoters of developmental genes is still unclear. It has been suggested that the coexistence of active and repressive marks maintains bivalent genes in a poised transcriptional state, prepared for activation or repression depending on the developmental stimuli[2,6,9]. However, this model is challenged by the fact that depletion of PRC2 or MLL2 complex subunits in ESCs causes minimal defects in gene expression[6]. Moreover, cells lacking MLL2 remain pluripotent and are able to induce developmental gene expression upon retinoic acid differentiation[6,7].

Genome architecture is emerging as a fundamental mechanism for genome regulation. It is known that distal regulatory regions and promoters are in physical contact to fine-tune developmental gene expression[14,15]. Despite being transcriptionally poised, bivalent promoters in ESCs are highly interconnected[16,17]. PRC1 and PRC2 are required for maintaining long-range interactions between bivalent promoters[14,17–21]. PRC2 is necessary for the formation of intra- and inter-chromosomal interactions between a subset of bivalent promoters in the transition from ground state pluripotent ESCs toward a developmentally more advanced, primed-like state[19]. Moreover, PRC2 mediates interactions between poised enhancers and developmental genes to facilitate their transcriptional induction upon differentiation[21]. PRC1 seems to play a more global role in spatially organizing interactions between bivalent genes, particularly in the *Hox* gene network[17,22]. Further, the PRC1-dependent spatial organization of the genome is thought to maintain proper gene silencing[17].

Here, we have comprehensively interrogated the changes in the epigenetic landscape, genome architecture, genome accessibility, and gene expression in ESCs that have been rapidly and acutely depleted of MLL2. MLL2 loss results in increased occupancy of Polycomb complexes, rewiring of long-range interactions and decreased nascent transcription around bivalent promoters. Together, our results delineate the balancing functions of bivalency in maintaining promoter regions interconnected and accessible, to properly and efficiently orchestrate developmental gene regulatory programs.

Results

reChIP-seq experiments provide a catalog of bona fide bivalent genes. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has traditionally been the method of choice to evaluate the genomic occupancy of histone modifications. However, ChIP-seq cannot unequivocally distinguish co-occupancy of histone modifications at the same nucleosomes, at different alleles, or in specific cells within the cell population. Sequential ChIP-seq (reChIP-seq) overcomes this limitation and has recently been applied to map

[1]Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. [2]Present address: Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL, USA. [3]Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain. [4]CNAG-CRG, Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [5]Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA. [6]Simpson Querrey Center for Epigenetics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [7]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [8]ICREA, Barcelona, Spain. *e-mail: martirenom@cnag.crg.eu; luciano.dicroce@crg.eu

co-occupancy of H3K4me3 and H3K27me3 in CD4$^+$ memory T cells[23]. To define bona fide bivalent regions in mouse ESCs (mESCs), we performed individual ChIP-seq and reChIP-seq experiments for H3K4me3 and H3K27me3 using mESCs isolated from blastocysts carrying conditional knockout alleles for *Mll2*/*Kmt2b* (refs.[12,24,25]). Overlap between individual H3K4me3 and H3K27me3 ChIP-seq identified 6,817 potential bivalent genes (Fig. 1a and Supplementary Table 1). We next performed reChIP-seq, with either H3K4me3 followed by H3K27me3, or H3K27me3 followed by H3K4me3. We verified that the chromatin used for reChIP-seq was sheared to mono-nucleosome size (~150 bp on average), to maximize the probability that both histone modifications were found in the same nucleosome (Supplementary Fig. 1a). We removed possible carry-over peaks from the first antibody by subtracting the peaks obtained from the corresponding H3K4me3–immunoglobulin-G (IgG) or H3K27me3–IgG reChIP-seq experiments. In total, we identified 4,778 and 5,582 loci containing H3K4me3–H3K27me3 and H3K27me3–H3K4me3, respectively (Fig. 1a). The overlap between ChIP-seq and reChIP-seq experiments identified a total of 3,868 bona fide bivalent genes (Fig. 1a and Supplementary Table 1). As expected, the 3,868 bona fide bivalent genes showed relatively low levels of H3K27ac at the promoter, and low H3K36me3 levels at the coding region, compared with a subset of 9,948 non-bivalent genes (decorated by H3K4me3 alone; Supplementary Fig. 1b and Supplementary Table 1). Gene ontology (GO) analysis confirmed that the 3,868 bona fide bivalent genes encode for proteins involved in morphogenesis and embryonic development (Supplementary Fig. 1c). To validate our bioinformatic analyses, we performed reChIP-qPCR (Supplementary Fig. 1d and Supplementary Table 2).

Next, we mapped the genome-wide occupancy of MLL2 and the PRC2 and PRC1 subunits SUZ12 and RING1B, respectively. As expected, our 3,868 bona fide bivalent genes were enriched for MLL2, SUZ12, and RING1B (Fig. 1b,c). We found MLL2 peaks at the transcription start site (TSS) of 62% of H3K4me3-marked promoters[7,8]. Importantly, the majority of the bona fide bivalent genes are MLL2 targets (Supplementary Fig. 1e).

We also found that approximately 20% of bona fide bivalent peaks mapped to non-TSS regions (524 non-TSS peaks from a total of 3,039 peaks; Fig. 1d). The majority (66.5%) of bivalent non-TSS peaks were located in intragenic regions (including coding sequence, introns, and UTRs), while the rest (33.5%) were found in intergenic regions (that is, further than ±2.5 kb from any annotated gene; Supplementary Fig. 1f). Both bivalent TSS and non-TSS peaks were enriched in MLL2, H3K4me3, and H3K27me3 compared with an IgG control (Fig. 1e). Additional validation would be required to determine whether these bivalently marked non-TSS regions act as transcriptional elements.

MLL2 and H3K4me3 restrain local Polycomb occupancy at bivalent genes. In ESCs, MLL2/COMPASS catalyzes the deposition of H3K4me3 at developmental gene promoters[26]. To characterize the role of MLL2 in bivalency, we employed tamoxifen-inducible conditional *Mll2* knockout (KO) mESCs to acutely eliminate MLL2 protein[8,24] and performed ChIP-seq to evaluate changes in H3K4me3 and H3K27me3. In agreement with previous reports, knockout of *Mll2* led to a reduction in H3K4me3 at bivalent promoters (Fig. 2a). Unexpectedly, this was accompanied with an increase of H3K27me3 levels at bivalent genes (Fig. 2a). To further examine this effect, we focused on a subset of 1,669 genes, within the 3,868 bona fide bivalent genes, that are MLL2 targets and have greater than twofold decrease in H3K4me3 signal following MLL2 depletion (Supplementary Fig. 2a). Throughout the manuscript, we will refer to this gene subset as 'HD' bivalent genes, for 'highly dependent' on MLL2. The 1,669 HD bivalent genes showed a marked increase in H3K27me3 levels upon *Mll2* KO (Fig. 2b, Supplementary Fig. 2a, and Supplementary Table 1). The HD gene set also exhibited the highest levels of H3K27me3 in wild type (WT) cells, compared with the complete 3,868 bivalent gene catalog (Supplementary Fig. 2b). From the 9,948 non-bivalent gene set, we extracted a group of 4,213 genes with average expression comparable to bivalent genes (less than 5 total RNA RPKMs). In this collection of 4,213 non-bivalent genes, no significant changes were observed in H3K4me3 and H3K27me3 levels after MLL2 depletion (Fig. 2b and Supplementary Fig.2b). In addition, *Mll2* KO resulted in increased occupancy of SUZ12 and RING1B at bivalent genes (Fig. 2c–e). To confirm these observations, we performed ChIP-seq for H3K4me3, H3K27me3, SUZ12, and RING1B using *Drosophila* spike-in as an exogenous reference genome[27]. Indeed, the spike-in normalized ChIP-seq datasets confirmed the decrease in H3K4me3, concomitant to the increase in H3K27me3, SUZ12, and RING1B, at bivalent genes (Supplementary Fig. 2c). Importantly, *Mll2* KO did not significantly alter either the SUZ12 and RING1B global protein levels or their subcellular localization, but did lead to moderate decreases in H3K4me3 and H3K4me2 levels and a slight increase in H3K27me3 levels (Supplementary Fig. 2d–f).

Previous studies have reported that ESCs carrying constitutive MLL2 KO can be continually expanded in vitro with limited changes in gene expression[7,12,24]. We sought to re-evaluate whether transcriptional alterations accompanied the increased Polycomb occupancy observed in tamoxifen-inducible conditional *Mll2* KO mESCs. Total RNA-seq identified 564 genes significantly upregulated, and 744 genes significantly downregulated, after MLL2 depletion (Supplementary Fig. 2g and Supplementary Table 3). As expected, the pluripotency gene network had no major alterations in expression (Supplementary Table 3). Most downregulated genes were MLL2 targets associated with meiosis and neuronal development (Supplementary Fig. 2h,i). When assessing the transcriptional levels of our catalog of bivalent genes, we observed a subtle but significant decrease in transcription of bivalent genes, whereas non-bivalent genes remained unchanged (Supplementary Fig. 2j). Interestingly, while the occupancy of RPB1 (the largest subunit of the RNA polymerase II complex) was not significantly altered, the occupancy of the Ser5-phosphorylated form of RNA polymerase II (RNAPII-Ser5P) significantly decreased at the TSS of our HD gene subset in *MII2* KO cells (Supplementary Fig. 2k–m).

H3K4me3 inhibits the PRC2 methyltransferase activity toward H3K27me3 in vitro[28]. To test whether the presence of H3K4me3 reduces the levels of H3K27me3 in bivalent promoters, we performed ChIP-qPCR to evaluate the changes in H3K27me3 in clustered regularly interspaced short palindromic repeat (CRISPR) knock-in mESCs carrying a catalytically inactive MLL2[29]. Indeed, catalytically impaired MLL2 cells had reduced H3K4me3 levels, concomitant with increased H3K27me3, at bivalent genes compared with WT cells (Fig. 2f). Together, these data indicate that MLL2 and its catalytic activity restrict the occupancy of Polycomb complexes, thus balancing the levels of H3K27me3 and RNAPII-Ser5P and thereby maintaining a low but active transcriptional state at bivalent promoters.

Depletion of MLL2 causes A-to-B compartment rearrangements. Polycomb complexes are required to establish and maintain regulatory long-range interactions between bivalent regions in mESCs[14,17–19,21,22]. Our results indicated that MLL2 restricts occupancy of Polycomb at bivalent promoters. We hypothesized that MLL2 might also modulate long-range contacts between bivalent genes. To assess this, we performed in situ Hi-C experiments from *Mll2* WT and acutely depleted *Mll2* KO mESCs[30]. Comparing the contact matrices and correlation coefficient matrices showed that MLL2 depletion did not induce major rearrangements in high-order genome organization (Fig. 3a,b). At the megabase scale, we did not find significant changes in topologically associating domain (TAD) border strength or size

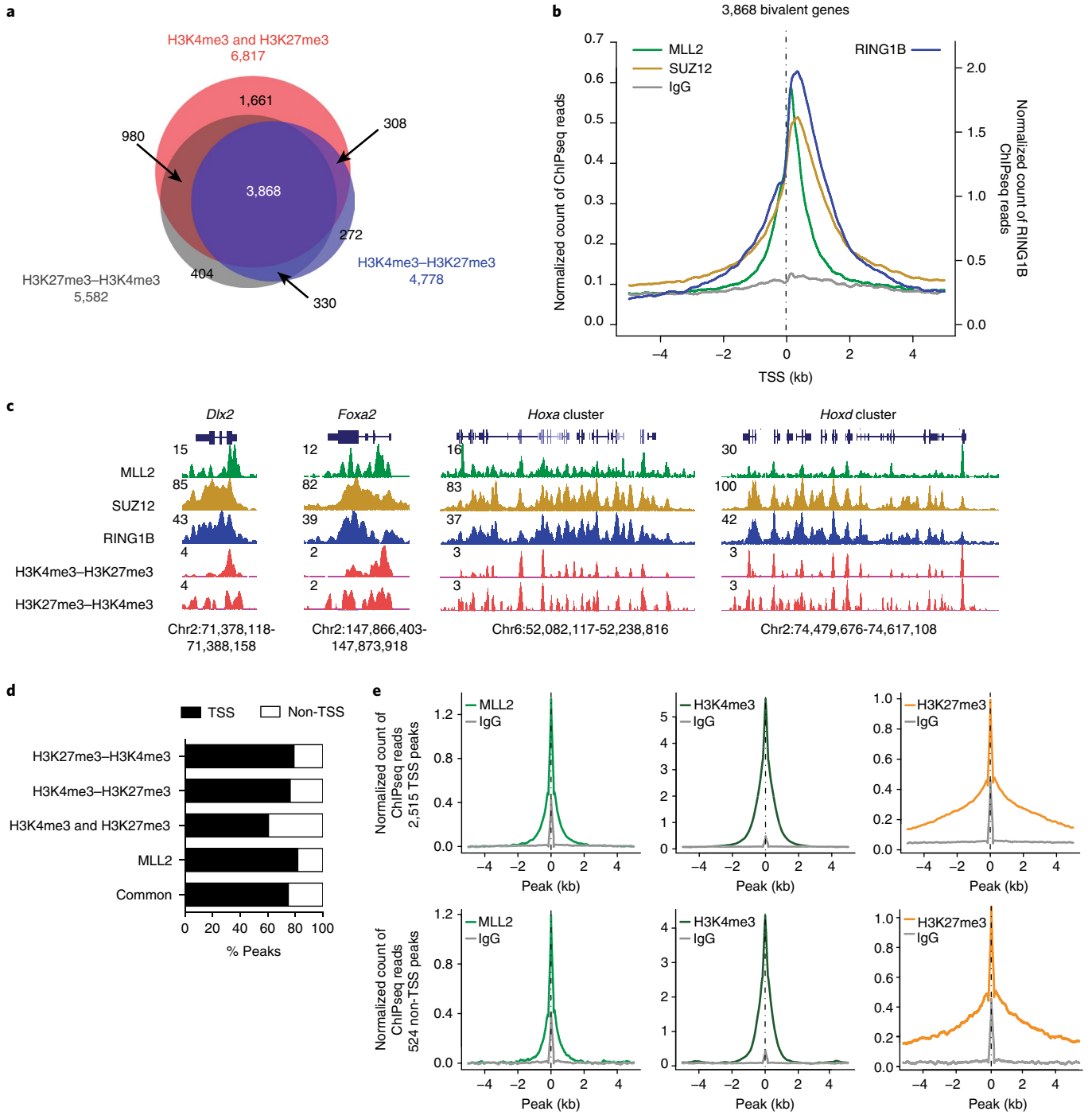

Fig. 1 | reChIP-seq experiments identified a set of 3,868 bona fide bivalent genes in mESCs. a, Venn diagram showing the overlap between the 6,817 co-target genes from H3K4me3 and H3K27me3 individual ChIP-seq experiments (red), 4,778 target genes from H3K4me3–H3K27me3 reChIP-seq (blue), and 5,582 target genes from H3K27me3–H3K4me3 reChIP-seq (gray). All experiments were performed in mESCs carrying knockout alleles for *Mll2/Kmt2b*, before inducing the knockout. The overlap identifies 3,868 bona fide bivalent genes. **b**, TSS (±5 kb) enrichment plot of the indicated ChIP-seq at 3,868 bivalent genes in WT mESCs. Enrichment levels are normalized for the total number of reads of each sample. **c**, UCSC genome browser snapshots of MLL2, SUZ12, RING1B, and reChIP-seq experiments (H3K4me3–H3K27me3; H3K27me3–H3K4me3) in the indicated genes, *Hoxa* and *Hoxd* clusters. Chromosomal coordinates of the sequencing tracks are indicated. $N = 2$ independent cell culture replicates were performed and showed similar results. **d**, Percentage of peaks of the indicated ChIP-seq and reChIP-seq experiments located in TSS (±2.5 kb around the TSS) and non-TSS (peaks outside ±2.5 kb around the TSS). Common peaks were identified by overlapping results from individual H3K4me3, H3K27me3, and MLL2 ChIP-seq as well as those from H3K4me3–H3K27me3 and H3K27me3–H3K4me3 reChIP-seq experiments (giving 2,515 common TSS peaks and 524 common non-TSS peaks). **e**, Peak (±5 kb) enrichment plots of the indicated ChIP-seq experiments at 2,515 bivalent TSS peaks (upper panels) and 524 bivalent non-TSS peaks (lower panels).

(Supplementary Fig. 3a)[31,32]. Correspondingly, MLL2 loss did not significantly alter CTCF genome distribution or occupancy (Supplementary Fig. 3b).

The cross-correlation interaction matrices at 100 kb identified 11,086 bins in compartment A (open, actively transcribed chromatin) and 14,213 bins in compartment B (closed, transcriptionally

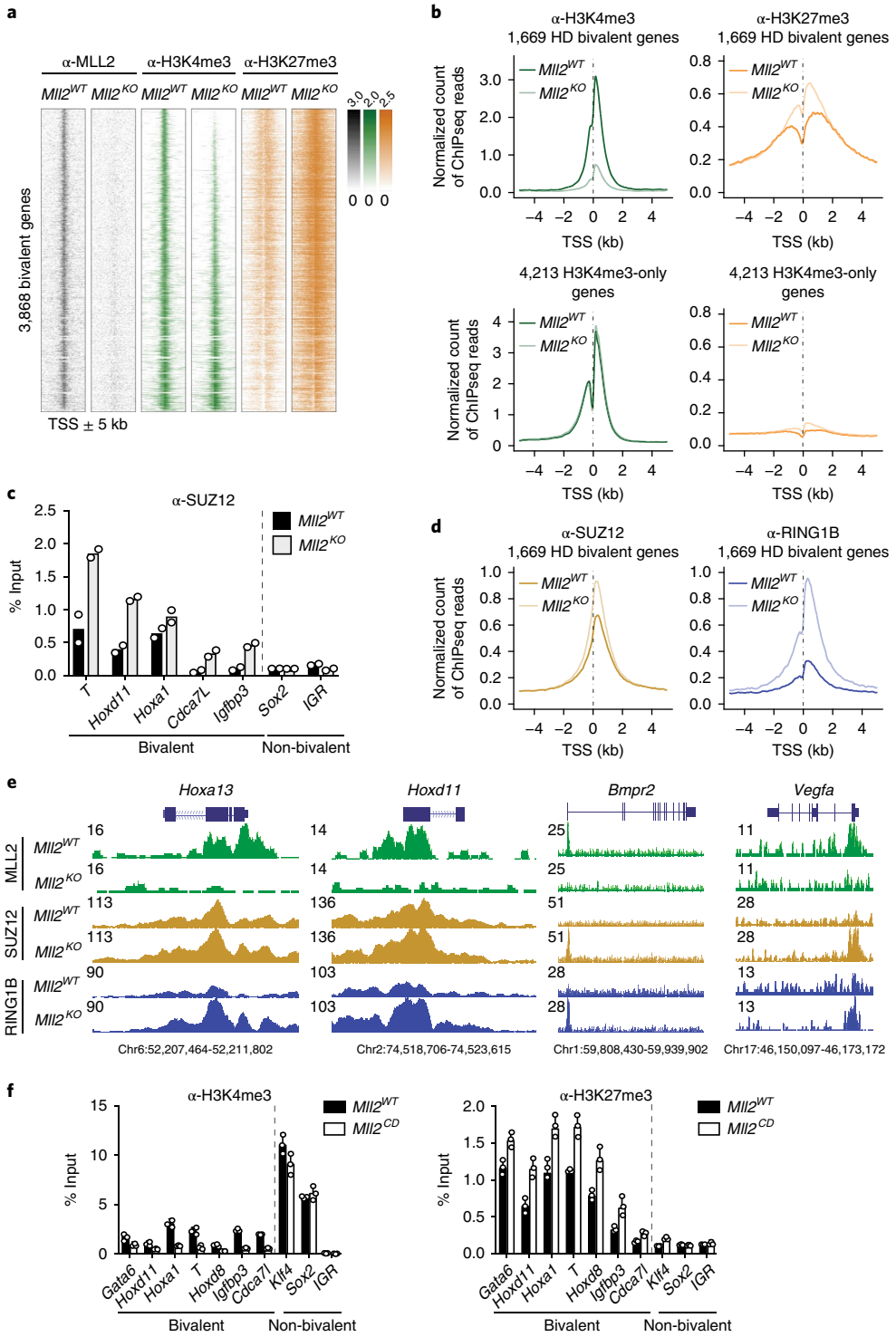

Fig. 2 | Depletion or catalytic impairment of MLL2 results in local increase in Polycomb occupancy at bivalent genes. a, Heatmaps of MLL2, H3K4me3, and H3K27me3 centered at the TSS (±5 kb) of 3,868 bona fide bivalent genes, showing the levels in *Mll2* wild-type (*Mll2*^{WT}) and *Mll2* knockout (*Mll2*^{KO}) mESCs. Genes are ranked by the degree of loss of H3K4me3 after *Mll2* depletion. Numbers in color scale show the correspondence between values and colors. **b**, TSS enrichment plots (±5 kb) of H3K4me3 (left) and H3K27me3 (right) ChIP-seq signal at 1,669 HD bivalent genes (top) and at 4,213 non-bivalent genes marked only by H3K4me3 (bottom), in *Mll2*^{WT} and *Mll2*^{KO} ESCs. **c**, ChIP-qPCR of SUZ12 in *Mll2*^{WT} and *Mll2*^{KO} cells. Examples of bivalent genes (*T* (brachyury), *Hoxd11*, *Hoxa1*, *Cdca7L*, *Igfbp3*) and non-bivalent genes (*Sox2*, *IGR* (intergenic region)) are shown. Values of $n = 2$ independent cell cultures are shown. **d**, TSS enrichment plots (±5 kb) at 1,669 HD bivalent genes from *Mll2*^{WT} and *Mll2*^{KO} cells for SUZ12 and RING1B. **e**, UCSC genome browser snapshots of MLL2, SUZ12, and RING1B in the indicated genes. Chromosomal coordinates of the sequencing tracks are indicated. $N = 2$ independent cell culture replicates were performed and showed similar results. **f**, ChIP-qPCR of H3K4me3 (left) and H3K27me3 (right) in *Mll2*^{WT} and catalytically dead *Mll2* (*Mll2*^{CD}) mESCs. Values represent the mean, and error bars correspond to the s.e.m. of $n = 3$ independent cell cultures.

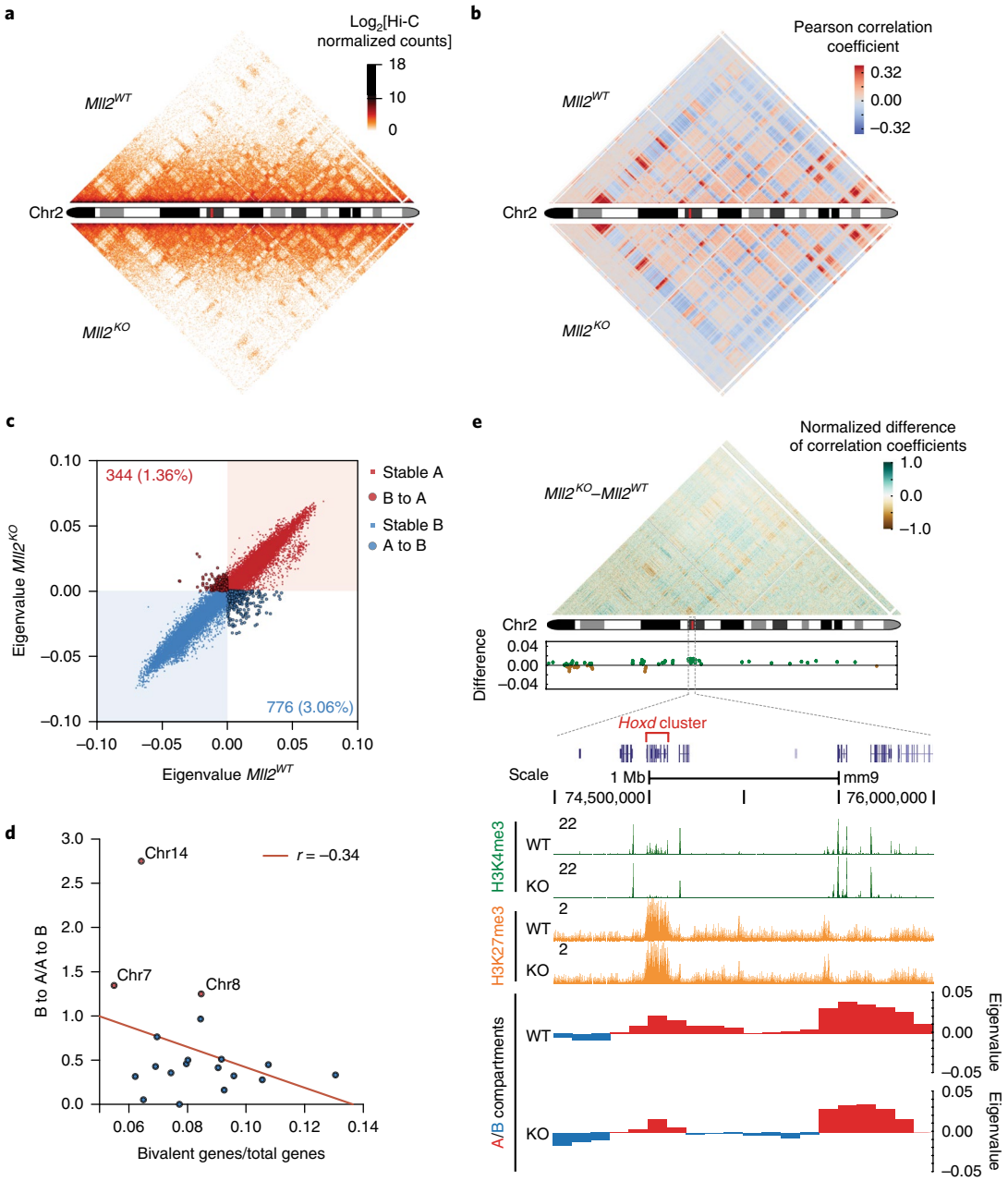

Fig. 3 | Depletion of MLL2 results in an A-to-B compartment switch of bivalent regions. a, Hi-C normalized interaction matrix for *Mll2*^{WT} (above diagonal) and *Mll2*^{KO} (below diagonal) mESCs at 100-kb resolution for chromosome 2 (~182 Mb). **b**, Pearson correlation coefficient matrix of normalized interaction ratios in *Mll2*^{WT} and *Mll2*^{KO} ESCs across chromosome 2 at a 100-kb resolution. Positive correlation is shown in red and negative correlation in blue. **c**, Scatter plot of eigenvector in *Mll2*^{WT} versus *Mll2*^{KO} ESCs. **d**, Correlation between the proportion of bivalent genes with respect to the total number of genes in the genome, with the proportion of A-to-B with respect to B-to-A switchers. Red dots represent mouse chromosomes with ratio (B to A/A to B) higher than 1.0. $N = 4$ independent cell culture experiments were analyzed. Pearson correlation coefficient (r) is indicated. **e**, Matrix of normalized differences in correlation coefficients between *Mll2*^{WT} and *Mll2*^{KO} in chromosome 2. Middle panel, bins that switched compartments (change of sign in correlation coefficient). Green dots, switches from A to B; orange dots, switches from B to A. Lower panels zoom in chromosome 2 region including the *Hoxd* cluster (red), with ChIP-seq tracks for H3K4me3 and H3K27me3, and A/B compartment changes for the *Hoxd* cluster and nearby bins. The genome is binned in 100 kb bins, and the region covers chromosome 2 from 73 Mb to 76 Mb. All eigenvalues corresponding to this region decreased in *Mll2*^{KO}.

repressed chromatin)[33]. We verified that compartment A regions contained genes with higher expression compared with compartment B regions (Supplementary Fig. 3c). MLL2 loss caused approximately 3% of the genome bins to switch from the compartments A to B, while 1.4% of the bins switched from B to A (Fig. 3c). To investigate which regions of the genome switch between compartments, we correlated the proportion of bivalent genes per chromosome with the proportion of changes in the A and B compartments. Interestingly, the higher proportion of bivalent genes present in a chromosome, the higher frequency of A-to-B compartment switches that occurred on MLL2 depletion (Fig. 3d). Only three chromosomes (Chr7, Chr8, and Chr14) resulted in more B-to-A changes compared with A-to-B. In total, 148 genes out of our 1,669 HD bivalent gene set were located within regions that switched

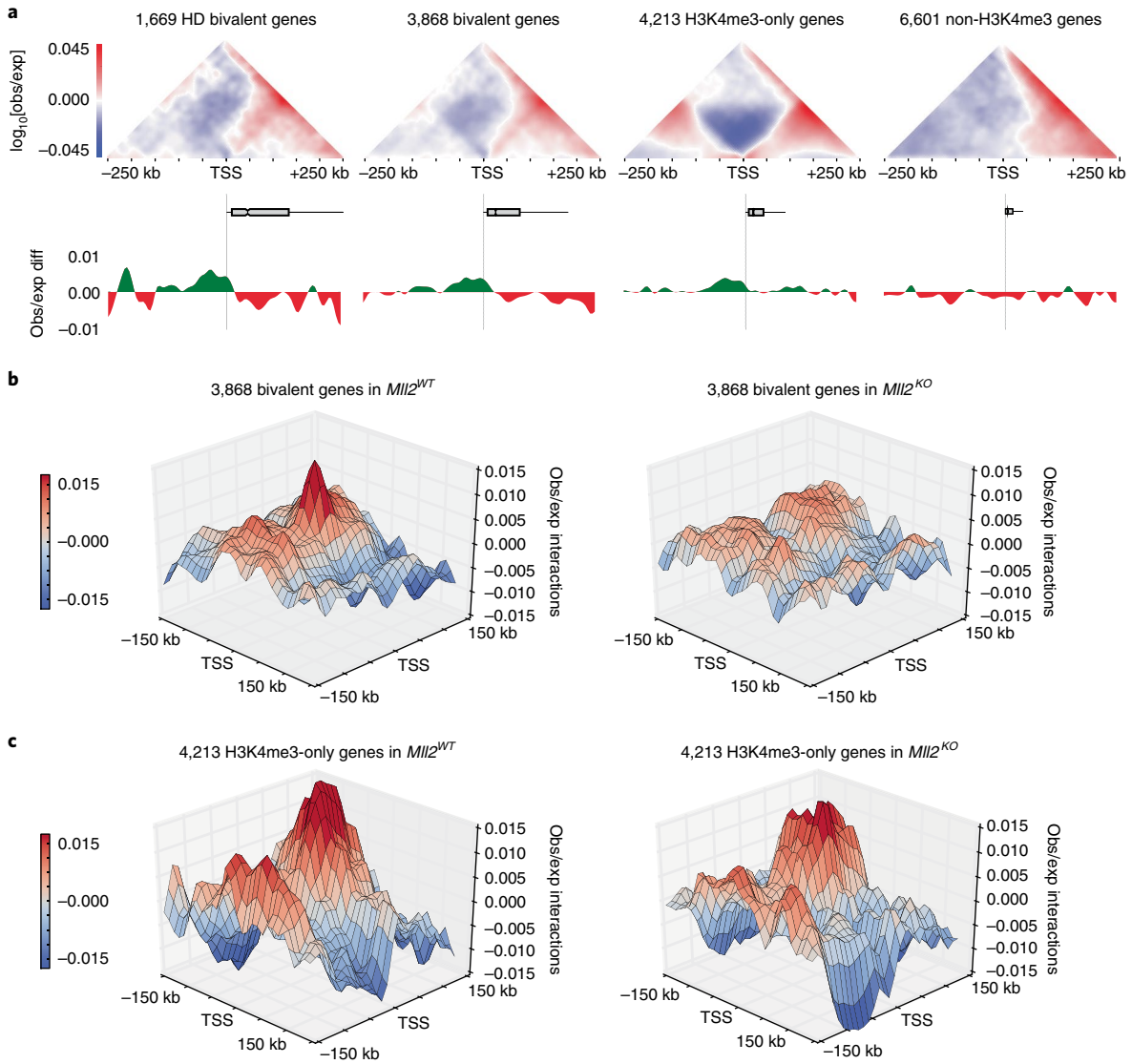

Fig. 4 | MLL2 loss leads to scattering of distal interactions at bivalent TSS. a, Upper panels: average $\log_{10}$ of observed versus expected interaction meta-matrices for ±250 kb from the TSS of the indicated subsets of genes. Direction of TSS was corrected by strand and direction of transcription, and interaction submatrices were normalized by decay. The distribution of gene size from the TSS is shown in grey. Lower panels: differential interactions in *Mll2*^{KO} mESCs compared with WT cells. **b**,**c**, Three-dimensional interaction meta-plots (10-kb resolution) depicting observed versus expected interaction frequencies between TSS of bivalent genes (**b**) and non-bivalent genes (**c**). Hubs visualize intra-TAD (300 kb–1 Mb) cross-talk between TSS sites. Areas shown are centered on the respective TSS (±150 kb).

from compartment A to compartment B on MLL2 loss, including important developmental regulators such as *Hand2*, *Neurog2*, and *Prdm6*. Although the difference was not statistically significant, the expression of these 148 genes trended toward downregulation (Supplementary Fig. 3d). We also observed that A-to-B compartment switching occurred near well-known bivalent genes such as the *Hoxd* cluster (Fig. 3e). Our results indicated that MLL2 loss and increased occupancy of Polycomb at bivalent promoters did not affect the overall organization of the ESC genome, but may induce a switch from the A to the B compartment in regions proximal to bivalent genes.

Interactions between bivalent promoters redistribute upon MLL2 loss. To further explore the role of MLL2 in the spatial organization of bivalent genes, we next examined promoter contacts at high resolution (5 kb) using a meta-loop analysis similar to the paired-end spatial chromatin analysis (PE-SCAn; see Online Methods)[34]. We interrogated the average of observed over expected interaction frequencies ±250 kb around the TSS of specific gene subsets, centering the interaction submatrices at the TSS taking into account the transcription directionality (Fig. 4a). Notably, TSS of bivalent genes interacted more frequently downstream of the TSS (toward gene bodies) than TSS of 4,213 H3K4me3-only (non-bivalent) genes, which interacted both upstream and downstream from the TSS (Fig. 4a, upper panels). As previously reported, bivalent genes showed larger size compared with non-bivalent genes[35]. Remarkably, in *Mll2* KO mESCs, bivalent genes displayed reduced connectivity with the gene body (downstream of the TSS) and increased interactions within the TSS and upstream from the TSS (Fig. 4a, lower panels, and Supplementary Fig. 3e). Importantly, interactions of non-bivalent genes downstream from the TSS were not affected. Our results indicate that MLL2 depletion affected the short-range (within 250 kb) contacts of bivalent TSS, by reducing their interaction with the gene body and increasing the interaction frequency with elements upstream from the gene body.

To identify the elements with increased interactions with bivalent TSS in *Mll2* KO cells, we analyzed the frequency of mid-range interactions (>300 kb up to 1 Mb) between the TSS of our 3,868 bivalent genes—that is, interactions between bivalent genes located within the same TAD (Fig. 4b). We observed that interactions were focused between TSS sites in WT cells (Fig. 4b, left panel). Strikingly, *Mll2* KO mESCs showed a clear change in the pattern of interactions around the TSS, from a strictly focal pattern to more diffused (Fig. 4b, right panel). In contrast, non-bivalent genes displayed a similar focused pattern of interactions at TSS in *Mll2* WT and *Mll2* KO mESCs (Fig. 4c, left and right panels, respectively). Finally, no changes were observed in the frequency of long-range inter-TAD interactions (from 2 to 10 Mb) between bivalent genes in WT and *Mll2* KO mESCs (Supplementary Fig. 3f). Together, these data indicate that bivalency focalizes short- and mid-range interactions between bivalent genes mostly at the TSS, and disruption of the MLL2-Polycomb balance results in a redistribution of interactions around the TSS region.

MLL2 catalytic activity contributes to focalize interactions. Given that loss of MLL2 catalytic activity resulted in increased levels of H3K27me3 at bivalent genes, we next sought to determine whether H3K4me3 is required for the establishment of the focalized pattern of interactions at bivalent promoters. To this end, we performed in situ Hi-C in WT and catalytically dead MLL2 ESCs ($MLL2^{CD}$). Similar to *Mll2* KO, $MLL2^{CD}$ ESCs showed minor changes in high-order genome organization and limited A/B compartment changes: 2.53% of the genome bins switched from B to A, while 2.99% of genome bins switched from A to B (Supplementary Fig. 4a). The proportion of bivalent genes in a chromosome again correlated with the frequency of A-to-B compartment switching (Supplementary Fig. 4b). The chromosomal region near the *Hoxd* cluster also presented an A-to-B compartment switching, resembling *Mll2* KO conditions (Supplementary Fig. 4c).

Our analyses therefore suggested that MLL2 deletion or abrogation of its catalytic activity induce minor A-to-B compartment switches. When examining the interaction frequencies flanking bivalent TSS, we found that in $MLL2^{CD}$ cells, bivalent promoters displayed a reduction in interactions with the gene body, but increased interactions upstream the TSS (Supplementary Fig. 4d, lower panels). Conversely, non-bivalent promoters presented fewer changes in interaction frequency flanking the TSS (Supplementary Fig. 4e). Furthermore, the TSS-focalized pattern of interactions between bivalent TSS in WT cells also became redistributed in $MLL2^{CD}$ cells (Supplementary Fig. 4f). This effect was less pronounced in non-bivalent genes (Supplementary Fig. 4g). Overall, our results indicate that MLL2 catalytic activity contributes to centralize interactions between bivalent genes at the TSS.

MLL2 controls promoter accessibility and nascent transcription. Given that MLL2 loss caused a redistribution of bivalent TSS interactions, we next examined whether this would be accompanied by alterations in chromatin accessibility, by performing ATAC-seq experiments in *Mll2* WT and *Mll2* KO ESCs[36]. A total of 59,008 accessible peaks were identified in *Mll2* WT cells, while a reduced number (51,530 peaks) were found in *Mll2* KO cells (Fig. 5a). The genomic distributions of the accessible peaks in both cell types were similar (Fig. 5a). Global comparison of the ATAC-seq profiles between *Mll2* WT and *Mll2* KO mESCs indicated that ~7,500 peaks lost accessibility to various degrees after MLL2 depletion, with 3,147 peaks in *Mll2* WT cells at least twofold less accessible in *Mll2* KO cells (Fig. 5b, left panel). In turn, only 601 peaks gained accessibility by at least twofold in *Mll2* KO cells (Fig. 5b, right panel). Interestingly, 86% of the 3,147 peaks that robustly lost accessibility in *Mll2* KO cells were located in intergenic and intragenic regions, while 14% were found at promoters (Fig. 5c). Furthermore, these 3,147 peaks showed a marked decrease in H3K4me3 levels after MLL2 depletion (Fig. 5d, left panel), with minor changes in H3K27me3 (Fig. 5d, right panel). We identified 1,258 genes associated with the 3,147 peaks with lost accessibility in *Mll2* KO cells (peaks located 2.5 kb upstream from the TSS to the transcription end site). Notably, the expression of these 1,258 genes was significantly downregulated in *Mll2* KO cells (Supplementary Fig. 5a). GO term analysis of these genes showed enrichment in neuronal development and function (Supplementary Fig. 5b,c). Thus, loss of H3K4me3 in *Mll2* KO cells correlates with reduced chromatin accessibility and downregulated transcription, potentially at bivalently marked regions. Direct inspection of bivalent promoters showed a decrease in accessibility after MLL2 depletion (Fig. 5e,f, left panels), while non-bivalent genes were not affected (Fig. 5e,f, right panels).

To further correlate the changes in accessibility and global genome architecture induced by MLL2 loss, we examined the ATAC-seq levels of regions that underwent A/B compartment switch. Regions switching from A to B in *Mll2* KO cells had a significant decrease in chromatin accessibility (Supplementary Fig. 5d, left panel), compared with regions switching from B to A (Supplementary Fig. 5d, right panel). Thus, our Hi-C and ATAC-seq data collectively suggest that MLL2 restrains Polycomb binding to, and chromatin compaction of, bivalent regions of the stem cell genome.

These results prompted us to investigate in more detail the effects of MLL2 loss in transcription. We performed Precision nuclear Run-On sequencing (PRO-seq), to map the distribution of nascent transcription by tracing transcriptionally engaged RNA polymerases in vivo[37]. We analyzed differential nascent transcription emerging from known exons, and found a total of 1,136 genes significantly differentially expressed, of which 626 genes were downregulated in *Mll2* KO conditions. Interestingly, almost half of them are included within our bona fide bivalent gene set (Supplementary Fig. 5e,f). Moreover, GO term analyses showed functional links to cell proliferation, morphogenesis, and development of different tissues (Supplementary Fig. 5g). Notably, bivalent genes exhibited a significant downregulation in nascent transcription around the TSS after MLL2 depletion (Fig. 5g). Both sense and antisense nascent transcripts (±2 kb around the TSS) were downregulated around bivalent TSS (Fig. 5h and Supplementary Fig. 5h). Such changes were relatively modest in non-bivalent genes (Fig. 5i,j). Hence, MLL2 plays a role in maintaining a basal level of accessibility and nascent bidirectional transcription around bivalent TSS, which may have important biological significance in the context of development.

MLL2 KO impairs embryoid body formation and transcriptional programs. PRC2 and MLL2 are dispensable for maintaining ESC self-renewal but essential during differentiation[7,8,25,38]. *Mll2* mutant mice are embryonic lethal around E10.5, highlighting its fundamental functions in early embryogenesis. In the course of lineage differentiation, the genome undergoes extensive rewiring of long-range interactions, which are associated with changes in gene expression[39,40]. To gain insight into the consequences of the MLL2-dependent reorganization of bivalent genes during differentiation, we differentiated *Mll2* WT and *Mll2* KO ESCs into embryoid bodies (EBs) and performed RNA-seq. EBs mimic early development of the embryo. The morphology and size of *Mll2* KO EBs at days 6 and 9 were remarkably different than *Mll2* WT EBs (Fig. 6a and Supplementary Fig. 6a), as reported in ref. [25]. *Mll2* WT 9-day-old EBs induced the expression of 2,697 genes markedly involved in embryonic development, validating our differentiation protocol[41] (Fig. 6b, Supplementary Fig. 6b, and Supplementary Table 3). EBs generated from *Mll2* KO mESCs were also able to induce developmental gene expression programs (Fig. 6c and Supplementary Fig. 6c). Interestingly, however, up to 745 genes failed to induce in *Mll2* KO EBs at day 9 (Supplementary Fig. 6d), including bona fide

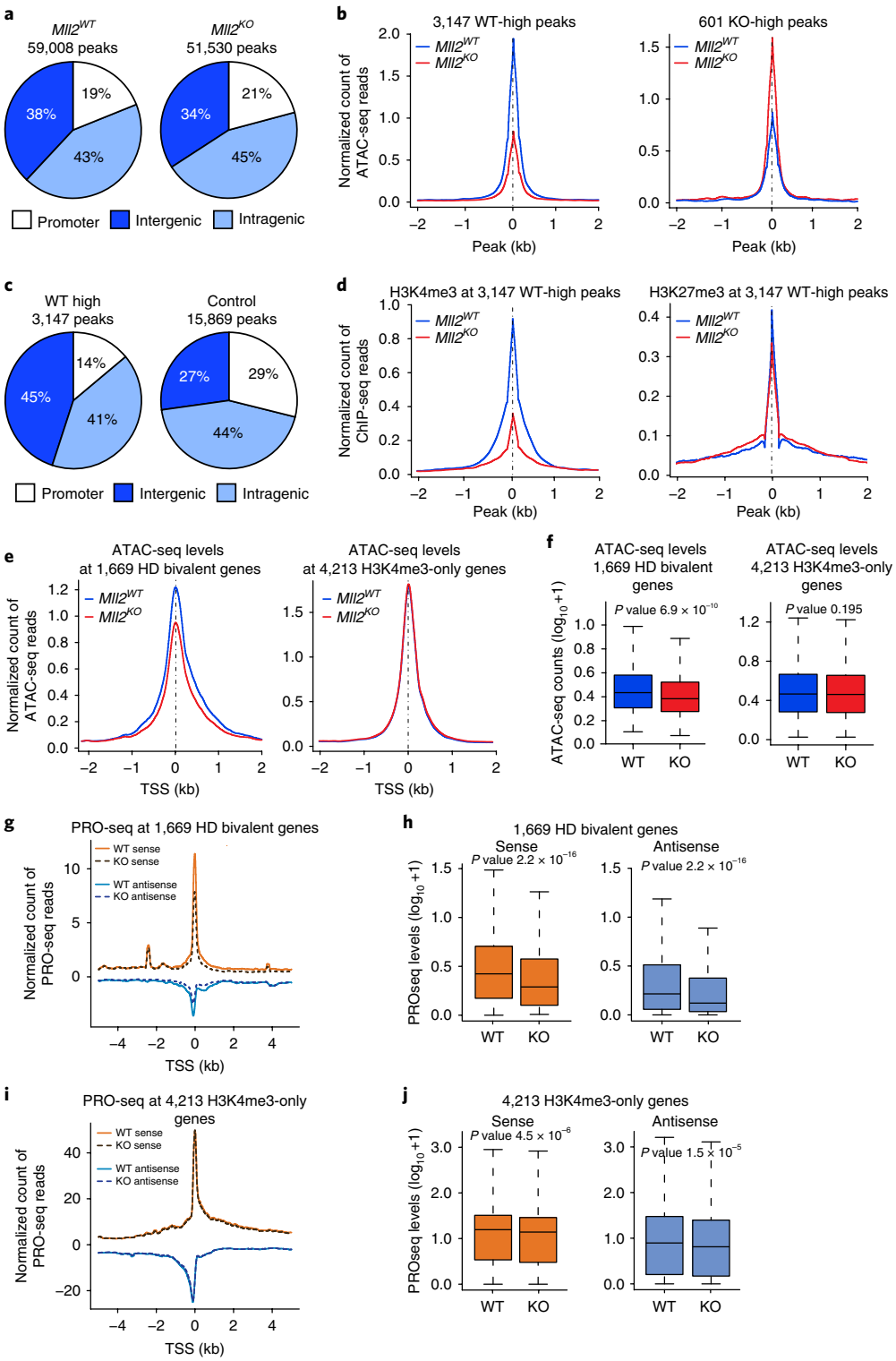

Fig. 5 | Bivalent gene promoters lose local chromatin accessibility after MLL2 depletion. a, Genome-wide distribution of ATAC-seq peaks in *Mll2*^{WT} and *Mll2*^{KO} cells. **b**, ATAC-seq levels at 3,147 peaks decreasing more than twofold after *Mll2*^{KO} (left), and at 601 peaks increasing more than twofold after *Mll2*^{KO} (right). **c**, Distribution of 3,147 ATAC-seq peaks with more than twofold decrease after *Mll2*^{KO} (left), and 15,869 ATAC-seq peaks not changing (<0.5 normalized units; right). **d**, ChIP-seq of H3K4me3 (left) and H3K27me3 (right) in *Mll2*^{WT} and *Mll2*^{KO} cells, at 3,147 ATAC-seq peaks decreasing more than twofold in *Mll2*^{KO}. **e**, ATAC-seq peak plots (±2.5 kb) in *Mll2*^{WT} and *Mll2*^{KO} cells at 1,669 HD bivalent genes (left), and 4,213 non-bivalent genes (right). **f**, Boxplots of ATAC-seq levels for bivalent (left) and non-bivalent (right) genes in *Mll2*^{WT} and *Mll2*^{KO} ESCs. Bottom and top of boxes are first and third quartiles, and line represents the median ($n = 3$ independent cell cultures). Whiskers represent 1.5 times the interquartile range from the median. *P* values were computed using Wilcoxon test (two-sided). **g**, PRO-seq peak plots (±5 kb from TSS) of sense and antisense transcripts in *Mll2*^{WT} and *Mll2*^{KO} cells at 1,669 HD bivalent genes. Signals were coverage and spike-in normalized. **h**, Boxplots of PRO-seq levels for 1,669 HD bivalent genes in *Mll2*^{WT} and *Mll2*^{KO} cells ($n = 2$ independent cell cultures). Sense and antisense regions correspond to ±2 kb respectively, from TSS in the direction of transcription. Boxes and whiskers as in **f**. *P* values were computed using Wilcoxon test (two-sided). **i**,**j**, Same as **g** and **h**, respectively, but for 4,213 non-bivalent genes.

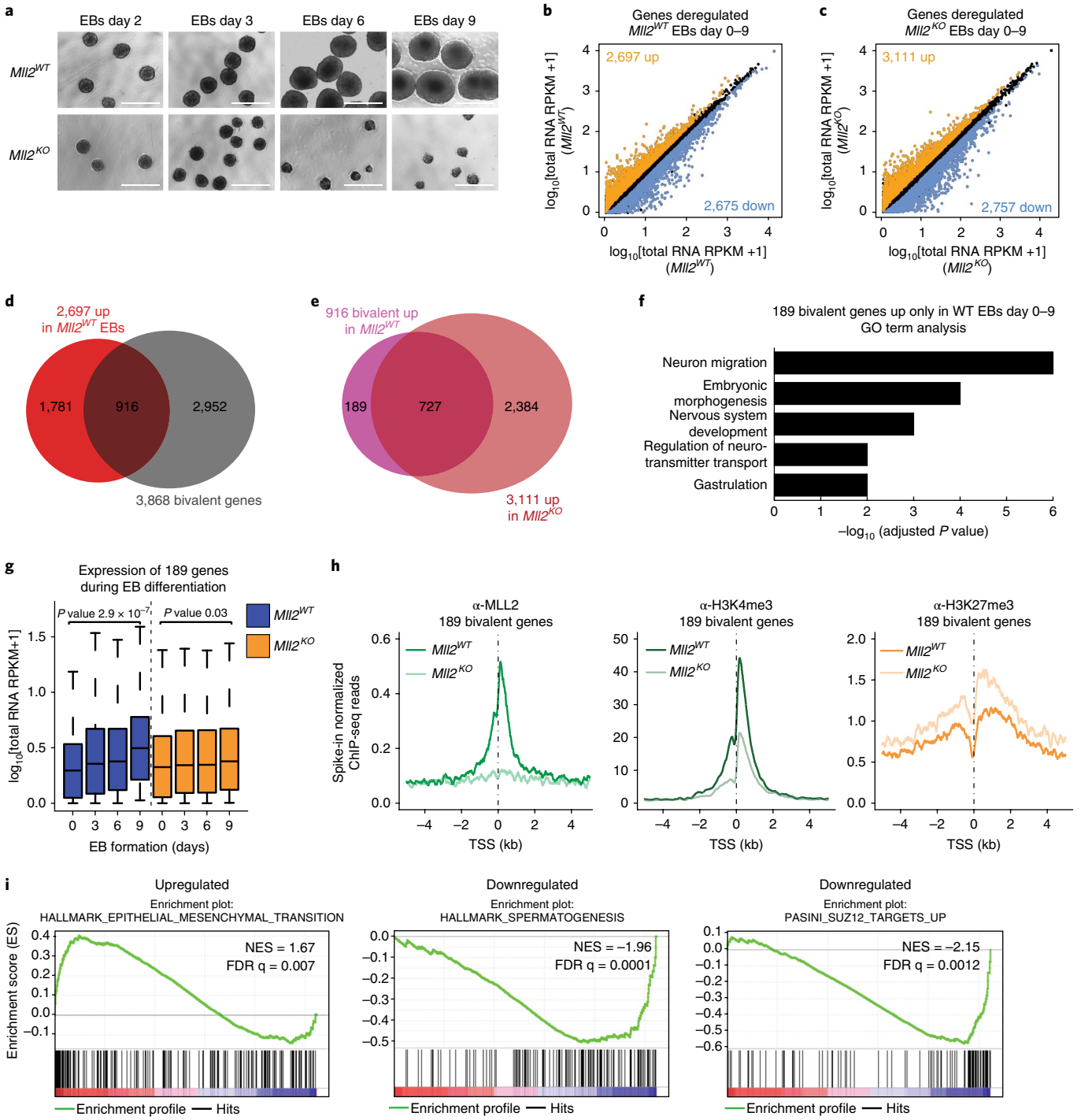

Fig. 6 | Loss of MLL2 leads to impaired embryoid body formation and dysregulation of developmental gene expression. a, EBs derived from *Mll2*^{WT} and *Mll2*^{KO} ESCs (×20 magnification; scale bar, 100 µM) at days 2, 3, 6, and 9 during EB differentiation. $N = 3$ independent cell cultures showed similar results. **b**, Differentially expressed genes ($n = 2$; adjusted P value < 0.1) from day 0 to 9-day-old *Mll2*^{WT} EBs. Genes significantly upregulated (orange) and downregulated (blue) are shown. **c**, Same as **b**, but for *Mll2*^{KO} EBs. **d**, Overlap between 2,697 genes upregulated in *Mll2*^{WT} EBs from day 0 to day 9 (red), with 3,868 bona fide bivalent genes (gray). **e**, Overlap between 916 bivalent genes upregulated in *Mll2*^{WT} EBs from day 0 to day 9 (pink), and 3,111 genes upregulated in *Mll2*^{KO} EBs from day 0 to day 9 (orange). **f**, GO analysis of 189 bivalent genes upregulated only in *Mll2*^{WT} EBs from day 0 to day 9. **g**, Boxplot of expression levels for 189 bivalent genes during differentiation of *Mll2*^{WT}- and *Mll2*^{KO}-derived EBs. Day 0 shows expression levels in ESCs. Boxes represent first and third quartiles, and median is the line within. Whiskers correspond to 1.5 times the interquartile range from the median ($n = 2$ independent cell cultures). P values were calculated with Wilcoxon test (two-sided). **h**, Spike-in normalized ChIP-seq TSS enrichment plots (±5 kb) of indicated proteins or histone modifications at 189 bivalent genes, in *Mll2*^{WT} and *Mll2*^{KO} ESCs. **i**, GSEA of genes deregulated between *Mll2*^{WT} and *Mll2*^{KO} EBs at day 9 (normalized enrichment score (NES); false discovery rate-adjusted P value (FDR q)).

bivalent genes with key functions in embryonic development and differentiation (*Sema3a*, *Hand2*, *Gata3*, *Mesp1*, *Wnt7b*) and important epigenetic regulators (*Kdm3a*, *Setdb1*, *Arid4b*, *Ash1L*, *Ezh1*, *Kdm6a*; Supplementary Table 3). Intriguingly, a significant fraction of genes (1,159 of the 3,111 genes; 37%) were aberrantly upregulated in *Mll2* KO EBs at day 9 (Supplementary Fig. 6d). GO term

analysis showed a functional enrichment in genes related to embryonic development as well as extracellular matrix organization, an essential process in the structural organization of the embryonic germ layers during development[42] (Supplementary Fig. 6e). Moreover, MGI mammalian phenotype ontology significantly associated these 1,159 aberrantly upregulated genes with neonatal and postnatal lethality (Supplementary Fig. 6f).

To further investigate the correlation between the macroscopic features of the EBs and gene expression, we next compared the expression profiles of bivalent genes in WT and KO EBs. Approximately 34% of the genes induced in *Mll2* WT conditions at day 9 are bona fide bivalent genes (916 of the 2,697 genes; Fig. 6d). Although most of the 916 genes were also induced in *Mll2* KO EBs, 189 bivalent genes showed a delayed activation in *Mll2* KO EBs (Fig. 6e,g). Out of the 189 genes, 64 (34%) were part of the 1,669 HD gene subset. The 189 genes displayed the characteristic loss of H3K4me3 and gain of H3K27me3 on MLL2 depletion (Fig. 6h), were functionally linked to embryo morphogenesis and neuronal development (Fig. 6f), and included master regulators of embryonic development, such as *Gata3*, *Foxf1*, *Hand2*, *Mixl1*, *Mesp1*, and *Meis2*. Gene set enrichment analysis (GSEA) showed that the epithelial–mesenchymal-transition pathway was among the most upregulated gene signatures in *Mll2* KO EBs, while genes associated with spermatogenesis and PRC2/SUZ12 targets[43] were downregulated (Fig. 6i).

Collectively, our results suggest that MLL2 targets bivalent promoters to maintain a low transcriptional state, accessible and physically interconnected with other bivalent genes. ESCs lacking MLL2 have impaired ability to generate embryoid bodies, and the kinetics of activation of developmental gene expression is greatly defective. These data put forward the interesting possibility that MLL2/H3K4me3 may not maintain a developmentally poised transcriptional state, but rather specify the robust and timely induction of bivalent promoters.

Discussion

Here, we comprehensively characterized the consequences of MLL2 depletion on the architecture of the embryonic stem cell genome. Abrogation of MLL2 or its catalytic activity led to increased levels of H3K27me3 at bivalent promoters. Mechanistically, loss of H3K4me3 and the presence of H3K27me3 in asymmetrically modified bivalent nucleosomes might stimulate PRC2 activity[4,44,45]. We note, however, that our catalog of bivalent genes is defined by ChIP and reChIP-seq experiments, which do not distinguish symmetrical (in the same histone H3 tail) or asymmetrical localization of the bivalent modifications. Interestingly, one of the best-characterized MLL2 target genes, *Magoh2*, was shown to have increased levels of H3K27me3, Polycomb subunits, and DNA methylation in ESCs lacking MLL2[24,29,46]. At bivalent promoters, we speculate that MLL2 depletion should not alter the DNA methylation state for two reasons: first, bivalent promoters overlap with hypomethylated CpG islands in ESCs[9], and second, our gene expression data in EBs suggest that many bivalent promoters that gain Polycomb in *Mll2* KO cells are able to induce transcription upon differentiation and thus might not become silenced by DNA methylation.

Previous studies did not find alterations of H3K27me3 global levels in *Mll2* KO mESCs[7,8]. This apparent discrepancy could be explained by distinct growth conditions and/or the relatively short time in which inducible KO cells were MLL2 depleted in our study. Indeed, constitutive *Mll2* KO cells (*Mll2*$^{-/-}$) had a 2.5-fold increase in *Mll1* messenger RNA levels compared with either *Mll2* WT ESCs or inducible *Mll2* KO cells. This suggests that MLL1 compensates for the long-term loss of MLL2 in constitutive *Mll2*$^{-/-}$ cells, as previously reported[8,12].

A growing body of research demonstrates that Polycomb complexes stabilize long-range interactions between developmental gene promoters[14]. The local increase in PRC2 and PRC1 occupancy around bivalent TSS in *Mll2* KO mESCs might explain the redistribution and formation of new interactions. Our RNA-seq and PRO-seq profiling also highlighted a more profound dysregulation in sense and antisense transcription around bivalent promoters than previously anticipated. The degree of transcriptional change between *Mll2* WT and *Mll2* KO increased during EB formation (Supplementary Table 3). Recently, MLL2 was shown to stabilize enhancer–promoter interactions at the *Gata6* and *Sox17* promoters[47], and to activate enhancers of the germ cell specification genes *Prdm14* and *Prdm1*[29]. We have now established an active role of MLL2 as a genome-wide orchestrator of long-range interactions, while maintaining the genes in a transcriptionally accessible state. Other COMPASS family members, such as MLL3 or MLL4, mark enhancer regions in ESCs and mediate enhancer–promoter interactions to ensure gene activation during lineage specification[48–50].

Our data showed dramatic macroscopic changes for six-day-old EBs derived from *Mll2* KO ESCs compared with WT, indicating that global differentiation is severely compromised in cells lacking MLL2. *Mll2* KO EBs aberrantly induced expression of genes functionally related to embryo morphogenesis and extracellular matrix organization, and failed to activate a subset of genes encoding for transcription factors and epigenetic regulators with key roles in differentiation and development (Fig. 6e,f and Supplementary Fig. 6d–f). These transcriptional abnormalities might contribute to the morphologic defects observed in *Mll2* KO EBs. Notably, genes encoding the Polycomb targeting factors Jarid2 and Phf19 were among the significantly downregulated genes in *Mll2* KO EBs (Fig. 6i). Jarid2 is essential during embryogenesis, and Phf19-depleted EBs display a differentiation delay[51,52]. The decrease in these Polycomb targeting factors may lead to a failure in lineage specification, as developmental genes of different lineages would be allowed to increase their expression within the same cells, thereby losing cell identity.

Our study provides a comprehensive view on how MLL2/COMPASS and Polycomb complexes balance their activities to regulate accessibility, nascent bidirectional transcription, and long-range interactivity of bivalent genes, to ensure a robust and timely induction of developmental gene expression, which may allow for proper lineage specification during embryogenesis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0218-5.

Received: 4 June 2017; Accepted: 1 August 2018;
Published online: 17 September 2018

References

1. Tee, W. W. & Reinberg, D. Chromatin features and the epigenetic regulation of pluripotency states in ESCs. *Development* **141**, 2376–2390 (2014).
2. Bernstein, B. E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
3. Azuara, V. et al. Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.* **8**, 532–538 (2006).
4. Voigt, P. et al. Asymmetrically modified nucleosomes. *Cell* **151**, 181–193 (2012).
5. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
6. Piunti, A. & Shilatifard, A. Epigenetic balance of gene expression by Polycomb and COMPASS families. *Science* **352**, aad9780 (2016).
7. Hu, D. et al. The Mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1093–1097 (2013).
8. Denissov, S. et al. Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant. *Development* **141**, 526–537 (2014).

9. Voigt, P., Tee, W. W. & Reinberg, D. A double take on bivalent promoters. *Genes Dev.* **27**, 1318–1338 (2013).
10. Ku, M. et al. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* **4**, e1000242 (2008).
11. Aloia, L., Di Stefano, B. & Di Croce, L. Polycomb complexes in stem cells and embryonic development. *Development* **140**, 2525–2534 (2013).
12. Glaser, S. et al. Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development. *Development* **133**, 1423–1432 (2006).
13. Laugesen, A. & Helin, K. Chromatin repressive complexes in stem cells, development, and cancer. *Cell. Stem. Cell.* **14**, 735–751 (2014).
14. Mas, G. & Di Croce, L. The role of Polycomb in stem cell genome architecture. *Curr. Opin. Cell Biol.* **43**, 87–95 (2016).
15. Sexton, T. & Cavalli, G. The role of chromosome domains in shaping the functional genome. *Cell* **160**, 1049–1059 (2015).
16. Schoenfelder, S. et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
17. Schoenfelder, S. et al. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat. Genet.* **47**, 1179–1186 (2015).
18. Entrevan, M., Schuettengruber, B. & Cavalli, G. Regulation of genome architecture and function by Polycomb proteins. *Trends. Cell Biol.* **26**, 511–525 (2016).
19. Joshi, O. et al. Dynamic reorganization of extremely long-range promoter–promoter interactions between two states of pluripotency. *Cell. Stem. Cell.* **17**, 748–757 (2015).
20. Denholtz, M. et al. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. *Cell. Stem. Cell.* **13**, 602–616 (2013).
21. Cruz-Molina, S. et al. PRC2 Facilitates the regulatory topology required for poised enhancer function during pluripotent stem cell differentiation. *Cell. Stem. Cell.* **20**, 689–705.e9 (2017).
22. Kundu, S. et al. Polycomb repressive complex 1 generates discrete compacted domains that change during differentiation. *Mol. Cell* **65**, 432–446.e5 (2017).
23. Kinkley, S. et al. reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in $CD4^+$ memory T cells. *Nat. Commun.* **7**, 12514 (2016).
24. Glaser, S. et al. The histone 3 lysine 4 methyltransferase, Mll2, is only required briefly in development and spermatogenesis. *Epigenetics Chromatin* **2**, 5 (2009).
25. Lubitz, S., Glaser, S., Schaft, J., Stewart, A. F. & Anastassiadis, K. Increased apoptosis and skewed differentiation in mouse embryonic stem cells lacking the histone methyltransferase Mll2. *Mol. Biol. Cell.* **18**, 2356–2366 (2007).
26. Andreu-Vieyra, C. V. et al. MLL2 is required in oocytes for bulk histone 3 lysine 4 trimethylation and transcriptional silencing. *PLoS Biol.* **8**, e1000453 (2010).
27. Orlando, D. A. et al. Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep.* **9**, 1163–1170 (2014).
28. Schmitges, F. W. et al. Histone methylation by PRC2 is inhibited by active chromatin marks. *Mol. Cell* **42**, 330–341 (2011).
29. Hu, D. et al. Not All H3K4 methylations are created equal: Mll2/COMPASS dependency in primordial germ cell specification. *Mol. Cell* **65**, 460–475.e6 (2017).
30. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
31. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
32. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
33. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**, 390–403 (2013).
34. de Wit, E. et al. The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **501**, 227–231 (2013).
35. Smith, E. & Shilatifard, A. Enhancer biology and enhanceropathies. *Nat. Struct. Mol. Biol.* **21**, 210–219 (2014).
36. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9 (2015).
37. Mahat, D. B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016).
38. Chamberlain, S. J., Yee, D. & Magnuson, T. Polycomb repressive complex 2 is dispensable for maintenance of embryonic stem cell pluripotency. *Stem Cells* **26**, 1496–1505 (2008).
39. Freire-Pritchett, P. et al. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *eLife* **6**, e21926 (2017).
40. Dixon, J. R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
41. Morey, L. et al. Polycomb regulates mesoderm cell fate-specification in embryonic stem cells through activation and repression mechanisms. *Cell Stem Cel.* **17**, 300–315 (2015).
42. Poh, Y. C. et al. Generation of organized germ layers from a single mouse embryonic stem cell. *Nat. Commun.* **5**, 4000 (2014).
43. Pasini, D., Bracken, A. P., Hansen, J. B., Capillo, M. & Helin, K. The Polycomb group protein Suz12 is required for embryonic stem cell differentiation. *Mol. Cell. Biol.* **27**, 3769–3779 (2007).
44. Margueron, R. et al. Role of the Polycomb protein EED in the propagation of repressive histone marks. *Nature* **461**, 762–767 (2009).
45. Rickels, R. et al. An evolutionary conserved epigenetic mark of Polycomb response elements implemented by Trx/MLL/COMPASS. *Mol. Cell* **63**, 318–328 (2016).
46. Ladopoulos, V. et al. The histone methyltransferase KMT2B is required for RNA polymerase II association and protection from DNA methylation at the MagohB CpG island promoter. *Mol. Cell. Biol.* **33**, 1383–1393 (2013).
47. Singh, A. M. et al. Cell-cycle control of bivalent epigenetic domains regulates the exit from pluripotency. *Stem Cell Reports* **5**, 323–336 (2015).
48. Herz, H. M. et al. Enhancer-associated H3K4 monomethylation by Trithorax-related, the *Drosophila* homolog of mammalian Mll3/Mll4. *Genes Dev.* **26**, 2604–2620 (2012).
49. Hu, D. et al. The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. *Mol. Cell. Biol.* **33**, 4745–4754 (2013).
50. Yan, J. et al. Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell Res.* **28**, 204–220 (2018).
51. Morey, L., Santanach, A. & Di Croce, L. Pluripotency and epigenetic factors in mouse embryonic stem cell fate regulation. *Mol. Cell. Biol.* **35**, 2716–2728 (2015).
52. Ballare, C. et al. Phf19 links methylated Lys36 of histone H3 to regulation of Polycomb activity. *Nat. Struct. Mol. Biol.* **19**, 1257–1265 (2012).

Acknowledgements

This work would not have been possible without F. Stewart and M. Obst (Biotechnology Center, TU Dresden, Germany), who generously provided the C57B1/6 *Mll2*$^{F/F}$ mouse ESCs. We are indebted to L. Morey and members of the Di Croce laboratory for insightful discussions and critical reading of the manuscript. We thank V. A. Raker for scientific editing, and the CRG Genomics Unit for their help in genomic experiments. This work was partially supported by the European Research Council under the 7th Framework Program FP7/2007–2013 (ERC grant agreement 609989) and the European Union's Horizon 2020 research and innovation programme (grant agreement 676556) to M.A.M.-R. We also acknowledge support of the Spanish Ministry of Economy and Competitiveness (BFU2016-75008-P), Centro de Excelencia Severo Ochoa 2013–2017 (SEV-2012-0208), AGAUR, and Fundació 'La Marató de TV3' to L.D.C. We also acknowledge a JSPS Research Fellowship for Young Scientists to support Y.A. of US National Cancer Institute grant R35CA197569 to A.S.H.

Author contributions

G.M. and L.D.C. designed the study. G.M., C.B. and M.S. carried out ChIP-seq experiments; G.M. and C.B. conducted RNA-seq experiments. G.M. performed western blot, in situ Hi-C, and ATAC-seq experiments. E.B. performed bioinformatic analyses for ChIP-seq, RNA-seq, ATAC-seq, and PRO-seq data. Y.S. and M.A.M.-R. performed Hi-C data processing and analyses. D.H. and A.S. generated the catalytically inactive MLL2 mouse ESCs. Y.A. performed PRO-seq experiments. F.L.D. assisted with in situ Hi-C experiments and helped to interpret Hi-C analyses. All authors contributed to the discussion and interpretation of the results. L.D.C. supervised the experiments and provided intellectual support in the design and interpretation of the data. G.M. wrote the manuscript with input from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0218-5.

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

Correspondence and requests for materials should be addressed to M.A.M. or L.D.

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

Methods

ESC lines and culture conditions. Conditional KO mESCs for the H3K4 methyltransferase *Kmt2b*/*Mll2*/*Wbp7* (reference sequence file identifier NC_000073.6; NM_001290573.1; NP_001277502.1. for *Mus Musculus* build GRCm38.p4), *Mll2*$^{Fl/Fl}$; *Rosa26::CreERT2*, were isolated from blastocyst outgrowths and generously provided by F. Stewart's laboratory[12]. Cells were cultured in feeder-free 0.1% gelatin-coated dishes, in DMEM (SIGMA G5154) supplemented with 20% FBS (HyClone), with 0.2% β-mercaptoethanol (Gibco), 1% Pen/Strep (Gibco), 1% sodium pyruvate (Gibco), 1% non-essential amino acids (MEM NEAA, Gibco), 1% GlutaMAX (Gibco), and 1,000 U ml^{-1} recombinant LIF (Millipore). Induction of Cre-recombination was achieved by treating cells with 10^{-6} M 4-hydroxytamoxifen during 96 h as previously described[12]. For PRO-seq experiments, 20% of Stasis Stem Cell Qualified US origin FBS (GEMINI, catalog#100–525) was used instead of Hyclone FBS. MLL2^{CD} (Y2602A mutant) ESCs were generated by CRISPR-mediated knock-in of mouse V6.5 cells in A.S.'s laboratory[29]. WT and catalytically dead MLL2 cells were cultured in dishes coated with 0.1% gelatin and irradiated mouse embryonic fibroblasts feeders, in 20% FBS DMEM media containing the abovementioned supplements and LIF. To remove contaminating feeder cells, ESCs cells where trypsinized and pre-plated for two rounds for 30 min each plating. Supernatant containing feeder-cleared ESCs was then collected and pelleted.

Whole-cell extracts, histone extracts, and chromatin fractionation. One p150 cm^2 confluent plate of *Mll2* WT and one plate of *Mll2* KO mESCs (treated 96 h with 10^{-6} M 4-hydroxytamoxifen) were trypsinized and collected, washed with PBS, and resuspended in 200 µl of 1× Laemmli buffer (2% SDS, 10% glycerol, 5% 2-mercaptoethanol, 62.5 mM Tris-HCl pH 6.8). Extracts were incubated at 98 °C for 10 min and sonicated for 30 s at high potency using a Bioruptor (Diagenode) until they clarified. Protein was quantified with a BCA Protein Assay Kit (Pierce) and 10–20 µg were loaded on SDS-PAGE gels.

Histone extraction was performed starting with approximately five million ESCs, harvested by centrifugation and washed with cold PBS. Pellets were resuspended by vortexing in 500 µl of cold lysis buffer (10 mM Tris-HCl pH 6.5, 50 mM sodium bisulfite, 1% Triton X-100, 10 mM $MgCl_2$, 8.6% sucrose, 10 mM sodium butyrate). Lysate was centrifuged at full speed for 15 s and the resulting supernatant discarded. Cold lysis buffer (1 ml) was added to the pellet, vortexed, and centrifuged at full speed for 15 s. Pellets were washed again with 1 ml cold lysis buffer, then with 1 ml of wash buffer (10 mM Tris pH 7.4, 13 mM EDTA). Samples were centrifuged at full speed for 15 s, and the supernatant was discarded. Extraction of histones was performed by adding 100 µl of 0.4 M H_2SO_4 to the remaining pellets, vortexing vigorously, and incubating on ice for 1 h. Samples were centrifuged at full speed for 5 min, and the supernatant was transferred to a new tube. Histones were precipitated by adding 900 µl of 80% cold acetone, vortexing, and incubating overnight at −20 °C. Acid-extracted histones were pelleted by centrifuging at full speed for 10 min. Pellets were air-dried and resuspended in 30 µl of 100 mM Tris-HCl pH 8. Histones were quantified using Bradford and 1.5 ng of protein was loaded in SDS-PAGE gels. Relative quantification of western blot bands was performed with ImageJ. Briefly, high-resolution scanned images were converted to JPEG in 'grayscale' mode. Mean gray values were obtained from bands of the same size, background was deducted, and the ratio of the value of the *Mll2*^{KO} band relative to the *Mll2*^{WT} band was calculated.

Chromatin fractionation experiments were performed as previously described[53] following a modified version of the Mendez and Stillman protocol[54]. Immunoblots were blocked in 5% milk in TBS-T (TBS pH 8, 0.1% Tween-20), antibodies incubated overnight at 4 °C in the same buffer, and washes were performed in TBS-T.

Full scans of gels and western blot images have been included in the Supplementary Information.

Antibodies used in this study. Several histone and non-histone antibodies were used in this study. Anti-H3K4me1, anti-H3K4me2, anti-H3K4me3, and anti-Mll2 antibodies were generated in A.S.'s laboratory and described elsewhere[7]. We used commercially available anti-H3K27me3 (Millipore, 07–449), anti-H3K4me3 (Diagenode, C15410003), anti-H3K36me3 (Abcam, ab9050), anti-H3K27ac (Millipore 07–360), anti-H3 (Abcam, ab1791), anti-SUZ12 (Abcam, ab12073), anti-EZH2 (Cell Signaling, 5246), anti-TUBULIN (Abcam, ab7291), anti-GAPDH (Santa Cruz, SC-32233) anti-IgG (Abcam, 172730), anti-CTCF (Millipore, 07–729), anti-RPB1-NTD (Cell Signaling Technologies, Rabbit mAb #14958 (D8L4Y)), and anti-RNAPolII-Ser5P antibody (clone CTD4H8; Mouse Monoclonal, Covance, MMS-128P). Anti-RING1B is a homemade antibody[55].

ChIP-seq, reChIP-seq, and ChIP-qPCR. For ChIP and reChIP experiments, ESC plates were cross-linked for 10 min at room temperature with 1% formaldehyde DMEM (without serum), incubated 5 min at room temperature with 125 mM glycine, followed by 15 min incubation on ice. Cells were washed twice in cold PBS and harvested by scraping in PBS supplemented with protease inhibitors (cOmplete EDTA-free, Sigma) and 1 mM PMSF, pelleted, and frozen at −80 °C. ChIP experiments were carried out following a standard protocol in the laboratory[55]. Briefly, cross-linked cell pellets were lysed in ice-cold ChIP buffer (1× volume SDS buffer (50 mM Tris-HCl pH 8.1; 100 mM NaCl; 5 mM EDTA pH 8; 0.5% SDS) plus 0.5 volume Triton dilution buffer (100 mM Tris-HCl pH 8.6; 100 mM NaCl; 5 mM EDTA pH 8; 5% Triton X-100) with protease inhibitors). Samples were sonicated for 16 cycles (30 s on/30 s off, high potency, Diagenode Bioruptor) and insoluble material was pelleted. To check chromatin concentration, 25 µl of the supernatant were mixed with 175 µl PBS, plus 2.5 µl proteinase K (20 mg ml^{-1}), de-cross-linked for 5 h at 65 °C in a shaker followed by DNA purification using a PCR purification kit (Qiagen). DNA was quantified using NanoDrop and 800 ng was loaded in a 1.4% agarose gel to test sonication efficiency. ChIP reactions were performed overnight using chromatin corresponding to 50 µg of DNA (or 20 µg for histone modifications) plus 5 µg specific antibody in a final volume of 500 µl ChIP buffer plus protease inhibitors. Immunocomplexes were recovered with 40 µl of protein A or G-agarose bead slurry incubated for 2 h at 4 °C. Immunoprecipitated material was washed three times with 1 ml low salt buffer (50 mM HEPES pH 7.5; 140 mM NaCl; 1% Triton X-100), one time with 1 ml high salt buffer (50 mM HEPES pH 7.5; 500 mM NaCl; 1% Triton X-100), and eluted in 200 µl freshly prepared elution buffer (1% SDS, 100 mM $NaHCO_3$). DNA complexes were de-cross-linked at 65 °C for 5 h, treated with proteinase K and the DNA purified using a PCR purification kit (Qiagen).

For ChIP experiments with spike-in control, an equal amount of *Drosophila melanogaster* S2 cell chromatin was added to each chip reaction (2.5% of the ESCs chromatin for histone modification ChIPs, and 0.08% for the rest of the ChIPs), together with 1 µg of an antibody against *Drosophila*-specific histone variant, H2Av (Active Motif, catalog no. 61686).

For ChIP-seq experiments using anti-MLL2, anti-RING1B, RPB1-NTD, and RNAPolII-Ser5P antibodies, we followed the protocol described in Beringer et al.[53].

reChIP-seq experiments were performed using reChIP-IT kit (Active Motif, 53016) following the manufacturer's instructions.

ChIP-seq and reChIP-seq libraries were prepared with 2–10 ng of ChIP or input DNA material using the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB, E7370L) as per the manufacturer's instructions. ChIP-seq libraries were size selected to remove fragments below 100 bp and amplified for 10 PCR cycles. Libraries were sequenced on a HiSeq 2000 platform (Illumina) to obtain approximately 30 million reads per library (50 bp, single end).

Primers utilized in ChIP-qPCR validations are listed in Supplementary Table 2.

In situ Hi-C. To generate in situ Hi-C libraries, we followed the protocol described in Rao et al.[30] with minor modifications. Briefly, frozen pellets from 20 million cross-linked ESCs were thawed and intact nuclei were permeabilized and digested with 600 U of MboI restriction enzyme (NEB, R0147M) in 1×NEB2 buffer overnight. We next filled the 5′-overhangs with biotin-14-dATP (Life Technologies, 19524–016), and performed in nuclei ligation overnight with 10,000 U of T4 DNA ligase (NEB, M0202M). Samples were treated with 10 µl of 10 mg ml^{-1} RNase A for 15 min at 37 °C and incubated overnight at 65 °C with 20 µl of 10 mg ml^{-1} proteinase K to reverse cross-links. DNA was purified using the phenol:chloroform:isoamyl alcohol method. To prepare in situ Hi-C libraries, 3 µg of Hi-C DNA was sheared with a Bioruptor Pico (Diagenode) to obtain fragments ranging 300–500 bp (8 cycles total of 20 s on, 60 s off each cycle). Sonicated DNA fragments were then subjected to biotin pull-down for 30 min at room temperature using magnetic streptavidin T1 beads (Dynabeads MyOne Streptavidin T1, 65601), at a ratio of 1.5 µg of sonicated DNA:50 µl beads slurry. Biotin pull-down washes and the enzymatic reactions to prepare Illumina-compatible libraries were performed on the beads, following the Rao et al.[30] indications. For the final amplification of the Hi-C libraries, we performed 10 PCR reactions per condition using 5 µl of beads per PCR reaction, Illumina primer mix, and 12.5 µl of 2× NEBNext High Fidelity PCR Master Mix (NEB, M0541S) in a 25 µl final volume per PCR. Libraries were amplified for 8 cycles. At the end of the reaction, supernatants of all the PCRs were pooled and beads were discarded, bringing the total volume of the library to 250 µl. Libraries were purified using 250 µl of AMPure XP beads (1× volume) and following the manufacturer's instructions. Final in situ Hi-C libraries were resuspended in 30 µl of 10 mM Tris-HCl pH 8.5 and quantified. Each library was sequenced in at least two lanes on the HiSeq 2500 platform (75 bp, paired-end reads, Illumina). We sequenced four biological replicates and one technical replicate per condition (*Mll2*^{WT} and *Mll2*^{KO}), reaching approximately 1.2 billion raw reads per condition (Supplementary Table 4). For in situ Hi-C experiments using WT V6.5 and *Mll2* Y2602A catalytically dead embryonic stem cells, 2 biological replicates per condition were generated, reaching 0.75 billion raw reads (Supplementary Table 4).

ATAC-seq. ATAC-seq libraries for *Mll2*^{WT} and *Mll2*^{KO} (treated 96 h with 10^{-6} M 4-hydroxytamoxifen) cells were generated following the Buenrostro et al. protocol[36]. Briefly, 50,000 cells per condition were washed in cold PBS and resuspended in 50 µl of cold lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% (v/v) Igepal CA-630). Samples were centrifuged for 10 min at 500*g*, 4 °C and the cell pellet resuspended in the transposition reaction mix (25 µl 2× transposition reaction buffer from Nextera kit, 2.5 µl Nextera Tn5 transposase from Nextera kit, 22.5 µl nuclease-free water) and incubated at 37 °C for 30 min. Samples were purified using the Qiagen MinElute PCR Purification Kit. Transposed DNA was eluted in 10 µl of elution buffer, and subjected to PCR

amplification using barcoded primers and NEBNext High Fidelity PCR Master Mix as described previously[36]. We determined that the optimal total number of PCR cycles for the ATAC-seq libraries was 8. ATAC-seq libraries were purified using 1.8× volumes of AMPure XP beads to remove fragments below 100 bp. Library quality was assessed using a Bioanalyser High Sensitivity DNA Analysis Kit (Agilent). Three biological replicates per condition were sequenced (50 bp, paired-end reads) on a HiSeq 2500 platform (Illumina), to obtain approximately 170 million raw reads per condition.

RNA-seq and embryoid bodies (EBs) experiments. Approximately 5 million mESCs from two independent cell cultures each of *Mll2* WT and *Mll2* KO (treated with carrier or 10^{-6} M 4-hydroxytamoxifen for 96 h, respectively), corresponding to day 0, were collected for RNA extraction. *Mll2* WT and *Mll2* KO mESCs were differentiated into EBs with the 'hanging drops' method as previously described[55,56]. Briefly, a suspension of 2×10^6 cells in 40 ml of 10% FBS GMEM without LIF was plated in 20 µl drops and incubated for 2 d. Drops containing EBs were then collected and transferred to non-attachment plates in fresh 10% FBS GMEM without LIF. Culture medium was changed every 2 d and pellets for RNA extraction were collected at Day 9. RNA was extracted using RNEasy Kit (Qiagen).

For RNA-seq, RNA samples were quantified and quality was evaluated using Bioanalyser (RIN > 9.9). One µg of total RNA was used as starting material to prepare RNA-seq libraries with ribosomal RNA depletion using TruSeq Stranded Total RNA Library Prep Kit (Illumina, RS-122-2201). Two biological replicates per condition were sequenced in a HiSeq 2000 (50 bp, single-end reads) to obtain about 40 million raw reads per sample.

PRO-seq experiments. We performed PRO-seq following a previously described protocol[37], with some modifications. After 4-hydroxytamoxifen treatment for 96 h, nuclei were isolated by Dounce homogenizer with loose pestle. Approximately 10^7 nuclei were subjected to nuclear run-on (30 °C, 3 min) in the presence of 25 µM biotin-11-ATP/GTP/CTP/UTP (PerkinElmer) and 6×10^5 *Drosophila* S2 spike-in nuclei. Total RNA were extracted and hydrolyzed in 0.2 M NaOH (on ice, 10 min). Biotinylated nascent RNA was purified by Dynabeads M-280 streptavidin (Invitrogen). After 3′ VRA3 adapter ligation and the second purification by Dynabeads, 5′ cap and 5′ hydroxyl RNA were converted to 5′ monophosphorylated RNA by RppH (NEB) and PNK (NEB), respectively. After 5′ VRA5 adapter ligation and the third purification by Dynabeads, complementary DNA was generated by reverse transcription with SuperScript III (Invitrogen) and RP1 primer. PCR by Phusion Hot Start (Thermo) and RP1/RPIx primer sets in 10 cycles amplified indexed DNA libraries. DNA libraries of 140–350 bp were size-selected by Pippin HT with 2% gel cassette 20B (Sage Science). Two independent cell cultures per condition were then sequenced by NextSeq 500 system (Illumina) with single-read runs.

ChIP-seq, reChIP-seq, RNA-seq, ATAC-seq, and PRO-seq data processing. ChIP-seq and reChIP-seq samples were mapped against the mm9 mouse genome assembly using Bowtie[57] with the option -m 1 to discard reads that did not map uniquely to one region. MACS[58] was run with the default parameters but adjusting the shift size to 100 bp to perform the peak calling of ChIP-seq experiments. For the reChIP-seq samples, the corresponding IgG control was subtracted from each experiment and only regions above one normalized unit were considered as peaks. Distribution of ChIP-seq peaks across different regions of the genome was calculated by counting the number of peaks fitting on each class of region accordingly to RefSeq annotations[59]. Each set of target genes was retrieved by matching those ChIP-seq peaks in the region from 2.5 kb upstream of the TSS until the end of the transcripts annotated in RefSeq[59]. Functional analysis of the target gene sets using GO and other genomic libraries was performed with Enrichr[60]. The aggregated plots showing the average distribution of ChIP-seq reads around the TSS of each target gene set were generated by counting the number of reads on this region for each gene and then averaging this value for the total number of mapped reads of each sample and the number of targets of the gene list. To generate the same plots but showing the average distribution of ChIP-seq reads along the body of a meta-gene derived from a target gene set, each gene was converted into a uniform gene of 100 positions to count the number of reads along this region, calculating the mean at each point of the resulting meta-gene profile afterwards. The aggregated plots showing the average distribution of ChIP-seq reads of a collection of peaks were generated by counting the number of reads around the center of each peak and normalizing for the total number of peaks in the set. Heat maps displaying the density of ChIP-seq reads around the TSS of each target gene set were generated by counting the number of reads on bins of 100 bp around ±5 kb of the TSS for each individual gene, and normalizing this value with the total number of mapped reads of the sample (without spike-in correction). The scale values were kept constant between WT and KO rows. No additional normalization was applied on the values within the same row. In all the cases, we used the release 68 of the RefSeq annotations[59] as provided by the University of California Santa Cruz (UCSC) Genome Browser on the refGene.txt file[61]. This RefSeq version contains 34,904 transcripts corresponding to 24,338 mouse genes. No preprocessing filtering steps were performed on this file. All the isoforms of each gene were considered to calculate the distribution of peaks along the genome features and to match peaks to target genes. For the aggregated plots and the heatmaps, instead, only one isoform that was arbitrarily selected per gene was included. ChIP-seq samples normalized by spike-in were mapped against a synthetic genome constituted by the mouse and the fruit fly chromosomes (mm9 + dm3) using Bowtie with the option -m 1 to discard reads that did not map uniquely to one region. The aggregated plots of ChIP-seq samples containing spike-in were generated by counting the number of reads mapped in mouse around the TSS for each gene and then averaging these values for the total number of reads mapped on the fruit fly genome and the number of targets of the gene list, as previously described[27]. The UCSC Genome Browser was used to generate screenshots of the genomic landscape of selected genes[61].

ATAC-seq triplicate samples were mapped against the mm9 mouse genome assembly using Bowtie with the option –m 1 to discard those reads that could not be uniquely mapped in just one region and the option -X 2000 to define the maximum insert size for paired-end alignment[57]. Mitochondrial reads were removed from each resulting map and down-sampling was applied to obtain the same number of mapped fragments per sample. Correlation between biological replicates was assessed to ensure high reproducibility before pooling each set of replicates. MACS was run with the default parameters but adjusting the shift size to 100 bp to perform the peak calling[58]. The peaks reported in both conditions were gathered on a single set of peaks and the maximum number of normalized reads within each peak in *Mll2* WT and *Mll2* KO conditions was determined. Peaks with at least twofold-enrichment in one of the conditions were considered as differential peaks for further analyses.

RNA-seq replicate samples were mapped against the mm9 mouse genome assembly using TopHat[62] with the option -g 1 to discard reads that could not be uniquely mapped to just one region. DESeq2[63] was run to quantify the expression of every annotated transcript using the RefSeq catalog of exons and to identify each set of differentially expressed genes. Expression values shown in the boxplots correspond to the average RPKMs across the two replicates in each condition. GSEA of the pre-ranked lists of genes by DESeq2 stat value was performed with the GSEA software[64]. Mouse genes were ranked by the ratio between KO and WT RNA-seq expression on EBs at day 9.

PRO-seq raw data were trimmed by cutadapt 1.14[65] and Trimmomatic 0.33[66]. Each sample was next mapped to the mouse and fruit fly genome assemblies (mm9 and dm3, respectively) using Bowtie. The number of fly reads on each experiment was used to normalize the corresponding genome-wide strand-specific profiles. The aggregated plots showing the average distribution of PRO-seq reads around the TSS of each target gene set were generated by counting the number of reads on this region for each gene in the sense and antisense orientations, and then averaging this value for the total number of fly mapped reads of each sample and the number of targets of the gene list. To quantify the PRO-seq level in sense or in antisense orientation for a group of genes, we defined a region of 2 kb upstream of the TSS (antisense strand, opposite to the direction of transcription) and another region of 2 kb downstream of the TSS of each gene (sense strand, in the direction of transcription). For each experiment, the total number of PRO-seq reads was normalized to the number of *Drosophila* spike-in reads. Then, we counted the number of PRO-seq normalized reads found in the 2 kb upstream or 2 kb downstream of each TSS, taking into account the direction of transcription. DESeq2[63] was run to systematically identify the set of genes showing significant differences in PRO-seq transcription.

Supplementary Table 4 includes sequencing statistics for ChIP-seq, RNA-seq, ATAC-seq, and PRO-seq experiments performed in this study.

Hi-C data analysis. Hi-C data were processed using TADbit[67] for read quality control, read mapping, interaction detection, interaction filtering, and matrix normalization. First, the reads were checked using an implemented FastQC protocol in TADbit. This allowed discarding problematic samples and detecting systematic artefacts. Next, we used a fragment-based strategy in TADbit for mapping the remaining reads to the reference mouse genome (mm9). The mapping strategy resulted in >80% of reads mapped uniquely to the genome (Supplementary Table 4). Next, we filtered non-informative contacts between two reads, including self-circles, dangling-ends, errors, random breaks, or duplicates as previously described[67]. The final interaction matrices resulted in 426 and 363 million of valid interactions for the $Mll2^{WT}$ and $Mll2^{KO}$ samples, respectively (Supplementary Table 4). These valid interactions were then used to generate genome-wide interaction maps at 100 kb, 40 kb, and 10 kb to segment the genome into the so-called A/B compartments, TADs, and perform a meta-analysis, respectively[32,68]. A/B compartments were calculated using normalized and decay corrected matrices[69] by calculating the first component of a principal component analysis of chromosome-wide matrices of the Pearson product-moment internal correlation as implemented in TADbit. TADs were next identified using 40-kb resolution normalized and decay corrected matrices with the TAD border detection method implemented in TADbit, which also provided a measure of the strength of each detected TAD border. Finally, to assess whether particular parts of the Hi-C interaction matrices had common structural features, we performed a meta-analysis of the region using the TADbit package[67] by merging individual local sub-matrices at 5 kb resolution in similar fashion as previously published (PE-SCAn)[34]. Contact maps for selected loci (Supplementary Figure 3e) were visualized at 2-kb resolution using the distance *Z*-scored adjusted interactions[70].

Supplementary Table 4 includes sequencing statistics for ChIP-seq, RNA-seq, ATAC-seq, PRO-seq and in situ Hi-C experiments performed in this study.

Construction of main gene sets. We defined the 3,868 bona fide bivalent gene set as the overlap between the 6,817 genes that were decorated by H3K4me3 and H3K27me3 in our ChIP-seq experiments, and the 4,778 and 5,582 loci containing H3K4me3–H3K27me3 and H3K27me3–H3K4me3, respectively, using reChIP-seq. The 1,669 HD bivalent genes (highly dependent on MLL2) correspond genes within the 3,868 bona fide gene set that are MLL2 targets and exhibit greater than twofold decrease in the number of H3K4me3 normalized reads upon $Mll2^{KO}$. The 9,948 non-bivalent gene set corresponds to genes decorated by H3K4me3 only (and not included in the list of H3K27me3 target genes). From the 9,948 non-bivalent gene set, we extracted the 4,213 non-bivalent gene set, which encompasses non-bivalent genes with average expression levels similar to bivalent genes (less than 5 RPKMs). Finally, the 6,601 non-H3K4me3 gene set corresponds to genes not included in the other non-bivalent groups, and thus are not decorated by H3K27me3 or H3K4me3.

Supplementary Table 1 contains the list of genes belonging to each subset.

Statistics and reproducibility. ChIP-seq and reChIP-seq experiments were performed independently at least twice with similar results. ChIP and reChIP products from these experiments were validated by ChIP-qPCR using known target genes and negative controls before and after sequencing library preparation. ChIP-seq experiments requiring accurate evaluations of differences between conditions (H3K27me3, H3K4me3, SUZ12, RING1B) were additionally validated with *Drosophila* spike-in antibody and chromatin controls (Active Motif). Two biological replicates (that is, independent cell line cultures) for MLL2 ChIP-seq were performed in WT and KO mESCs, and the results confirmed specificity and reproducibility. We performed an additional two or three biological replicates for ChIP-qPCR experiments, obtained similar results, and are shown in the figures. Three biological replicates per condition were performed for ATAC-seq, and two biological replicates for RNA-seq experiments. High statistical correlation ($r > 0.95$) between independent cell culture replicates was verified for every sequencing dataset, and figures show the averaged results from the multiple replicates. Western blot to assess histone modifications levels between *Mll2* WT and *Mll2* KO mESCs were performed at least 3 times ($n = 3$ independent cell cultures), and we used two different commercially available antibodies for H3K4me3 and H3K27me3. All experiments showed similar results, and representative images were selected for the final figures. To ensure reproducibility and robustness of ChIP-seq, Hi-C, ATAC-seq, and RNA-seq results, our datasets were cross-validated with published WT mESCs datasets. PRO-seq experiments were performed twice, with similar results, using two independent cell line cultures each in $Mll2^{WT}$ and $Mll2^{KO}$ conditions. One of the replicates was used to generate the figure panels. Four independent biological replicates and one technical replicate, reaching a total of 1.2 billion raw reads per condition, were analyzed for in situ Hi-C experiments. Two biological replicates were generated for in situ Hi-C experiments on MLL2 catalytic dead mutant and corresponding WT ESCs, reaching a total of 0.75 billion raw reads. Embryoid body experiments were performed 3 times ($n = 3$ independent cell cultures) and representative images were selected for final figures and images.

Reporting Summary. Further information on research design can be found in the Nature Research Reporting Summary linked to this article.

Data availability

Raw data and processed information for the ChIP-seq, in situ Hi-C, ATAC-seq, PRO-seq, and RNA-seq sequencing experiments generated in this study have been deposited in the Gene Expression Omnibus (GEO) under accession GSE99530.

References

53. Beringer, M. et al. EPOP functionally links elongin and Polycomb in pluripotent stem cells. *Mol. Cell* **64**, 645–658 (2016).
54. Mendez, J. & Stillman, B. Chromatin association of human origin recognition complex, cdc6, and minichromosome maintenance proteins during the cell cycle: assembly of prereplication complexes in late mitosis. *Mol. Cell. Biol.* **20**, 8602–8612 (2000).
55. Morey, L. et al. Nonoverlapping functions of the Polycomb group Cbx family of proteins in embryonic stem cells. *Cell. Stem. Cell.* **10**, 47–62 (2012).
56. Aloia, L. et al. Zrf1 is required to establish and maintain neural progenitor identity. *Genes Dev.* **28**, 182–197 (2014).
57. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome. Biol.* **10**, R25 (2009).
58. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome. Biol.* **9**, R137 (2008).
59. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
60. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
61. Tyner, C. et al. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.* **45**, D626–D634 (2017).
62. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
63. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome. Biol.* **15**, 550 (2014).
64. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
65. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
66. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
67. Serra, F. et al. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.* **13**, e1005665 (2017).
68. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
69. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**, 999–1003 (2012).
70. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhikeras guide to Hi-C analysis: practical guidelines. *Methods* **72**, 65–75 (2015).

nature research

Corresponding author(s): Dr Luciano Di Croce, Dr Marti-Renom

☐ Initial submission ☒ Revised version ☐ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

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

Experimental design

1. Sample size

Describe how sample size was determined.

No statistical methods were used to determine sample sizes. We followed ENCODE recommendations to decide the number of replicate experiments: ChIP-seq was performed in two independent replicates, with additional 2-3 independent replicates to validate by ChIP-qPCR; ATAC-seq and PRO-seq were performed in 3 and 2 independent replicates, respectively; RNA-seq were performed in 2 independent replicates.

2. Data exclusions

Describe any data exclusions.

TAD borders not called in independent biological replicates were excluded in subsequent analyses. Methods section includes further details on read and bin filtering strategies used for Hi-C data analyses, as well as processing of other sequencing datasets.

3. Replication

Describe whether the experimental findings were reliably reproduced.

No attempts to reproduce experimental findings failed.

4. Randomization

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

Randomization is not relevant to this study because no comparisons between experimental groups were made.

5. Blinding

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

Blinding was not relevant to this study because all metrics were derived from absolute quantitative methods without human subjectivity.

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

6. Statistical parameters

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

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

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

Software

Policy information about availability of computer code

7. Software

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

ChIP-seq, reChIP-seq and ATAC-seq datasets were mapped using BOWTIE (Langmead and Salzberg, 2012), peak calling using MACS (Zhang, 2008). RNA-seq data was mapped using TopHat (Trapnell, 2009) and differential gene expression was analyzed with DESeq2 (Love et al., 2014). GO term analyses were performed with ENRICHR (Kuleshov, 2016) and GSEA (Subramanian, 2005). Hi-C data was processed using TADbit (Serra et al., 2017). PRO-seq data was processed with CUTADAPT (Martin, 2011) and TRIMMOMATIC (Bolger, 2014). Wilcox test was performed using the wilcox.test() R function.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No restrictions. With the exception of the home-made anti-RING1B antibody used for ChIP-seq, which is available upon request to the corresponding authors, no other unique materials were used in this manuscript.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

ChIP experiments were performed using 4 ug antibody/sample. Anti-H3K4me1, anti-H3K4me2, anti-H3K4me3 and anti-Mll2 antibodies were generated in A. Shilatifard Lab and validated in Hu et al., 2013. We used commercially available anti-H3K27me3 (Millipore, 07-449; Lot. 2736613), anti-H3K4me3 (Diagenode, C15410003; Lot. A1062D), anti-H3K36me3 (Abcam, ab9050; Lot. GR220922-1), anti-H3K27ac (Millipore 07-360; Lot. 2686927), anti-H3 (Abcam, ab1791; Lot. GR293197-1), anti-Suz12 (Abcam, ab12073; Lot. GR777123-1), anti-Ezh2 (Cell signaling, 5246; Lot. 9), anti-Tubulin (Abcam, ab7291; Lot. GR138941-1), anti-Gapdh (Santa Cruz, SC-32233; Lot. G0110), anti-IgG (Abcam, 172730; Lot. GR262233-9), anti-CTCF (Millipore, 07-729; Lot. 2517762), anti-Rpb1-NTD (D8L4Y) (Cell Signaling Technologies, Rabbit mAb #14958; Lot.1) and anti-RNAPolII-Ser5P (clone CTD4H8) (Covance, Mouse Monoclonal, MMS-128P; Lot. #E10407LF). All commercial antibodies were validated by the manufacturer and ChIP/IP grade. Anti-Ring1b is a homemade antibody, validated in Morey et al 2012.
For Western Blot, antibodies were used at the concentrations recommended by the manufacturer.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

Conditional knock-out Mll2 (KMT2b, Gene ID: 75410; MGI:109565) mouse embryonic stem cells were isolated from blastocyst outgrowths by Dr. Stewart Laboratory (UT Dresden, Germany).
Mouse embryonic stem cells carrying catalytic inactive Mll2 mutations were generated by CRISPR-mediated knock-in in Dr Shilatifard Laboratory and described in Hu et al., Molecular Cell 2017. Irradiated MEF feeders were isolated from C57Bl6 blastocysts.

b. Describe the method of cell line authentication used.

Cre-recombination of knock-out alleles were performed for every experiment using PCR utilizing the primers described in Glaser et al, Development 2006 plus additional control primers (to detect a fragment between exons 2 and 3 in wild-type but not in knock-out Mll2 cells, and primers to detect beta-actin as loading control). Further validation of knock-down was done by western blot to detect Mll2 protein levels.

c. Report whether the cell lines were tested for mycoplasma contamination.

Cell lines were tested for mycoplasma contamination in two independent laboratories (Dr Stewart Lab, UT Dresden, Germany and Dr Di Croce Lab, Barcelona). Routinary checkups for mycoplasma contamination were performed in every experiment and were always negative.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

Animals and human research participants

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

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The study did not involve human research participants.