HHS Public Access

Author manuscript

Nat Struct Mol Biol. Author manuscript; available in PMC 2021 July 04.

Published in final edited form as:

Nat Struct Mol Biol. 2021 January ; 28(1): 103–117. doi:10.1038/s41594-020-00535-9.

Motif-driven interactions between RNA and PRC2 are rheostats that regulate transcription elongation

Michael Rosenberg[†,1], **Roy Blum**[†,1], **Barry Kesner**[1], **Eric Aeby**[1], **Jean-Michel Garant**[2,3], **Attila Szanto**[1], **Jeannie T. Lee**[1,*]

[1]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA USA; Department of Genetics, Harvard Medical School, Boston, MA 02114, USA.

[2]Canada's Michael Smith Genome Sciences Centre, Vancouver, BC, V5Z 4S6, Canada.

[3]RNA Group/Groupe ARN, Département de Biochimie, Faculté de Médecine et des Sciences de la Santé, Pavillon de Recherche Appliquée au Cancer, Université de Sherbrooke, 3201 rue Jean-Mignault, Sherbrooke, Québec, J1E 4K8, Canada.

Abstract

Although Polycomb repressive complex 2 (PRC2) is now recognized as an RNA-binding complex, the full range of binding motifs and why PRC2-RNA complexes often associate with active genes have not been elucidated. Here we identify high-affinity RNA motifs whose mutations weaken PRC2 binding and attenuate its repressive function in mouse embryonic stem cells. Interactions occur at promoter-proximal regions and frequently coincide with pausing of RNA Polymerase II (POL-II). Surprisingly, while PRC2-associated nascent transcripts are highly expressed, ablating PRC2 further upregulates expression via loss of pausing and enhanced transcription elongation. Thus, PRC2-nascent RNA complexes operate as rheostats to fine-tune transcription by regulating transitions between pausing and elongation, explaining why PRC2-RNA complexes frequently occur within active genes. Nascent RNA also targets PRC2 in cis and downregulates neighboring genes. We propose a unifying model in which RNA specifically recruits PRC2 to repress genes through POL-II pausing and, more classically, H3K27-trimethylation.

[*]Corresponding author: lee@molbio.mgh.harvard.edu.
[†]Equal contribution

AUTHOR CONTRIBUTIONS

M.R. and J.T.L. conceived of the PRC2-RNA interactome project. M.R. designed the dCLIP method and conducted the dCLIP, EMSA, Western, and cellular functional assays. R.B. devised the bioinformatics pipeline and performed the computational analysis for identification of consensus motifs. B.K. aligned sequencing data, and called dCLIP-seq peaks. E.A performed PRO-seq experiments and generated the data. J-M.G contributed the machine learning algorithm for rG4 analysis and for comparison between EZH2 and SUZ12 dCLIP replicates. A.S. generated RNA-seq data for 16.7 cells and performed initial alignment. M.R., R.B., and J.T.L. analyzed data, determined project direction, drafted figures, and wrote the manuscript.

POTENTIAL CONFLICTS OF INTEREST STATEMENT

J.T.L. is a cofounder of Translate Bio and Fulcrum Therapeutics, and is also a scientific advisor to Skyhawk Therapeutics.

INTRODUCTION

Polycomb repressive complex 2 (PRC2) is a multi-subunit complex that regulates stem cell pluripotency, cell differentiation, and embryonic development [1]. PRC2 possesses three core subunits: Suppressor of Zeste 12 Protein Homolog (SUZ12), Embryonic Ectoderm Development (EED), and the catalytic subunit, Enhancer of Zeste 2 (EZH2), that trimethylates histone H3-lysine 27 (H3K27me3) [2]. PRC2 binds and regulates expression of thousands of mammalian genes. An important unresolved question is how PRC2 is recruited to target genes. In mammals, various chromatin features have been proposed to recruit PRC2, but there is presently no evidence for specific DNA sequences as a general recruiting mechanism [3]. Because PRC2 interacts with a large ensemble of RNAs, RNA has also been proposed as sequence-specific targeting mechanisms[4–9]. The original model that cis-acting transcripts lend specificity to PRC2 recruitment [6] now finds support in many systems, including X-inactivation, genomic imprinting, cardiac differentiation, apoptosis, and telomere regulation [5,6,10–13]. Despite this, several paradoxical observations have seemingly complicated the simple model. First, RNA binding reduces the enzymatic activity of PRC2 [14–17]. Second, the RNA-binding activity can also be observed at active genes [5,7,8,18]. Both findings appear — at face value — to be inconsistent with PRC2's repressive function. These observations led to additional models in which (i) PRC2 binds RNA and DNA mutually exclusively to regulate an ordered assembly on chromatin [7,14,15,19], (ii) RNA holds PRC2 poised in check [8,14,17], and (iii) PRC2 surveys "junk mail" to prevent unnecessary transcription [18]. Although these proposals are not mutually exclusive, a unifying model has yet to be developed. Further confounding the field are CLIP studies reporting that PRC2 makes nonspecific contacts with many transcripts *in vivo* [7,8,18], thereby calling into question the specificity and regulatory function of PRC2-RNA interactions. These reports have led to characterization of PRC2 as "promiscuous" or nonspecific, implying a poor ability to discriminate RNA sequences [7,8,18–20].

Yet, PRC2 is known to have high affinity for some transcripts, including Xist RNA (K_d 20–80 nM)[14,21]. PRC2 is also known to prefer G-rich sequences [22] and to contact RNA via defined surfaces along EZH2, SUZ12, and EED [16,23]. Contact with EZH2's catalytic domain is especially notable [16], given that RNA inhibits the methyltransferase activity of EZH2 [14,17]. Nonetheless, proof of specificity ultimately rests on discrete binding motifs in the RNA, either in the primary RNA sequence or in its folded structure. Revealing such motifs would require a high-fidelity method of generating RNA-binding footprints. A major limitation of earlier CLIP methodologies has been the reliance on antibodies for pulling down protein-RNA complexes, as the low nanomolar affinities of antibody-antigen interactions (K_d 10^{-8} -10^{-9} M) preclude stringent washes during RNA purification. This limitation may explain why existing experiments all recovered large transcriptomes, making it difficult to identify footprints and underlying motifs for PRC2 binding [5,7,8,24].

To circumvent this problem, we previously developed "dCLIP" (denaturing CLIP) [25], which took advantage of *in vivo* biotin tagging of proteins, and generated proof-of-concept using the CBX7 subunit of Polycomb repressive complex 1 (PRC1), a biochemically distinct Polycomb complex that ubiquitylates histone H2A at lysine 119 [25]. Because biotin-streptavidin interactions have among the highest affinities and greatest specificity of any

non-covalent biological interactions ($K_d=10^{-15}M$), this approach enables purification of RNA-protein interactions under denaturing conditions, thereby allowing separation of true binding transcripts from non-specific ones. Here, we employ dCLIP to identify motifs for PRC2's two RNA-binding subunits, EZH2 and SUZ12, and use the resulting discoveries to derive a unified model.

RESULTS

dCLIP identifies subunit-specific RNA footprints for PRC2

We bio-tagged EZH2, SUZ12, and EED in mouse embryonic stem (mES) cells stably expressing BirA biotinylase (Fig. 1a,b) and performed dCLIP in day 7 (D7) mES cells differentiated into embryoid bodies. We UV-crosslinked cells, performed an RNAse protection step to degrade exposed RNA while preserving the bound RNA footprint, and then subjected resulting complexes to a stringent denaturing purification on streptavidin beads in the presence of 8M urea, 2% SDS, and 1M NaCl, thereby eliminating RNAs not covalently photo-crosslinked to PRC2 prior to deep sequencing. Although EED makes contact with RNA within PRC2 [16,23], our EED dCLIP yielded no convincing targets in multiple biological replicates regardless of whether we tagged the N- or C-terminus (Extended Data Fig. 1a, b). These data suggest that EED may not stably contact RNA, in agreement with an earlier study [14].

By contrast, EZH2 and SUZ12 (Extended Data Fig. 1c,d) yielded robust libraries in two independent replicates. Peak-calling using *PeakRanger* revealed >10,000 statistically significant peaks for each protein (Extended Data Table 1). For EZH2, 10,468 and 13,267 peaks were observed in two replicates. For SUZ12, 11,580 and 13,951 were observed in two replicates. By employing *deepTools*, we averaged the significance values (−log(p-value)) of strand-specific peaks enriched in at least one out of two replicates per bin. Scatter plots were generated by applying 1-kb bin size, and Pairwise-Pearson correlation (PPC) analysis was performed for each of two biological replicates (Extended Data Fig. 2a), yielding positive correlation coefficient values for both EHZ2 and SUZ12. There was less overlap between EZH2 and SUZ12 peaks (Extended Data Fig. 2b), but there was excellent concordance between transcripts hit by EZH2 and SUZ12 (Extended Data Fig. 2c), suggesting that the two subunits bind different regions of the same transcript. The unrelated CBX7 protein showed low concordance with EZH2 and SUZ12 (Extended Data Fig. 2d). For EZH2, the median footprint was 183 nt). For SUZ12, the median was 146 nt (Fig. 1c).

As a complementary approach, we employed an artificial neural network (ANN) used previously for predicting RNA G-quadruplexes [26]. For each replicate and PRC2 subunit, we introduced dCLIP FASTA sequences, trained the ANN, and performed accuracy testing using sequences obtained from the same interactome. A good predictive power was shown by the average area under the receiver operating characteristic (ROC) curve (AUC) between 0.712±0.012 and 0.783±0.011 (Extended Data Fig. 3, Extended Data Table 2). Second, we employed ANN for classifying the interactome of a corresponding biological replicate to predict the accuracy. Notably, there was compelling sequence similarity between the two biological replicates of EZH2 and SUZ12, with comparable average ANN accuracy values ranging between 0.7±0.004 and 0.773±0.008 (Extended Data Fig. 3, Extended Data Table

2). In contrast, the AUC scores of two CTCF dCLIP interactomes [27] (negative controls) were considerably lower, revealing poor concordance between interactomes of PRC2 and CTCF. Thus, PRC2-RNA interactomes were specific and highly reproducible among biological replicates.

We asked where the binding peaks occurred with respect to genes. Using the Genomic Association Tester (GAT) computational tool [28], we found that EZH2 and SUZ12 predominantly bound intronic sequences (Fig. 1d), agreeing with previous studies [8,9,17] and hinting that PRC2 favors binding nascent transcripts. When normalized to their general occurrence in the transcriptome (RNA input, Fig. 1d), intron enrichment was dwarfed by over-representation of 5'UTR, coding exons (CDS), and 3'UTR (Fig. 1e). Metagene analysis showed that, in addition to enrichment in the gene body, there were spikes around the transcription start site (TSS) and the transcription termination site (TTS) (Fig. 1f). The TTS spike did not occur in CBX7 [25], consistent with a low correlation between EZH2-SUZ12 versus CBX7 targets (Extended Data Fig. 2d).

Our analysis revealed a total of 3,839 significant PRC2-interacting RNAs in D7 mES cells, called on the basis of their having peaks in at least half (2 out of 4) of PRC2 libraries (EZH2 and SUZ12). We refer to the top 383 hits (10%) that bind both subunits as "high binders" (Fig. 1g, green dots), with highly enriched dCLIP signals (Extended Data Table 3). While high binders tend to have higher expression levels, there were 5,507 transcripts possessing similar expression levels (RPKM) without any reproducible dCLIP signal (Fig. 1g, red versus green dots). Thus, having high level expression alone is insufficient to ensure PRC2 binding. The dCLIP hits showed good correlation with the original EZH2 RNA interactome [5] and an EZH2 PAR-CLIP dataset [8] (Extended Data Fig. 4a,4b), but showed poorer correlation with an SUZ12 iCLIP dataset [7] (Extended Data Fig. 4c).

PRC2's interaction with nascent RNA fine-tunes gene expression states

To study the relationship between PRC2's RNA-binding sites versus chromatin binding sites, we performed a heatmap analysis plotting each gene's dCLIP signal against the EZH2 and H3K27me3 ChIP-seq intensities around the TSS (±3 kb) in mES cells of the same genetic background [29] (16.7 cells and $Tsix^{TST}/+$ cells have a similar transcriptomic profile on d7 (Extended Data Fig. 4d)). The enrichment signals were concordant between EZH2 and H3K27me3 (Fig. 2a), but dCLIP enrichment was strikingly different. There was an inverse correlation between dCLIP and ChIP signals overall (Extended Data Fig. 4e). EZH2 and SUZ12 RNA binding was most enriched over genes with lowest promoter-proximal EZH2 and H3K27me3 ChIP signals (Fig. 2a, b), concordant with the idea that PRC2 is frequently found at active genes [7,8,15,17,18].

To investigate mechanisms underlying the differences in PRC2's RNA binding relative to PRC2's function on chromatin, we created a three-tier regulatory database of genome-wide datasets obtained from differentiating mouse mES cells: (i) PRC2 dCLIP interactome; (ii) H3K27me3 ChIP-seq [29]; and (iii) RNA-seq of WT mES cells and their EED knockout (EED-KO) counterparts [30] (Extended Data Fig. 4f,g). We then categorized three gene groups based on the nature of their interactions with PRC2 (Extended Data Table 4). "Canonical" genes (n=603) include those that are repressed by high-level PRC2 and H3K27me3

enrichment (Fig. 2c, top). “CLIP” genes (n=414) have high PRC2-RNA (dCLIP) signals over the genes but paradoxically low PRC2 and H3K27me3 enrichment (Fig. 2c, middle). Because there is an uncoupling of RNA-mediated PRC2 recruitment and PRC2 catalysis on chromatin, we consider this group to exhibit non-canonical behavior. “No CLIP” genes (n=467) have low dCLIP, low PRC2/H3K27me3 ChIP signals, and are highly expressed (Fig. 2c, bottom). No CLIP genes reinforce the idea (Fig. 1g) that high-level expression is insufficient for PRC2-RNA binding and further argue for PRC2’s discriminating potential.

We asked how loss of PRC2 activity affected each category. Given that EED is a required subunit for EZH2’s methyltransferase activity, we analyzed RNA-seq data in EED-KO cells. Cumulative distribution plots (CDPs) for fold-changes between EED-KO and control cells showed that Canonical genes — the well-established Polycomb target genes — increased in expression, as expected (FC=3.3, Fig. 3a). No CLIP genes — which were previously highly expressed — showed a bimodal distribution in Eed-KO cells, with 30% being increased, 70% being decreased, and a net FC=0.67. These effects were likely indirect, considering that No CLIP genes have neither PRC2 nor RNA binding. No CLIP genes remained highly expressed when EED was ablated.

By contrast, CLIP genes — expressed genes with PRC2 bound to the nascent transcripts — showed a much different behavior. Because CLIP genes are among the most highly expressed in mES cells, one might not expect further increases in gene expression following EED-KO. However, they did increase in expression (Fig. 3a). CLIP genes also exhibited a bimodal distribution, with ~60% increasing, ~40% decreasing, and a net FC=1.3. The right-shift relative to the No CLIP profile was highly significant (P = 4e-18), as was the shift relative to Canonical genes (P = 2e-104). Analysis of absolute RPKM values corroborated the findings (Fig. 3b) and highlighted the fact that, even though the CLIP gene set showed the greatest RPKM values at baseline, they significantly increased in expression after PRC2 was ablated. To exclude the null hypothesis, we tested a randomized gene set with matching RPKM values in 10,000 randomized iterations but found no significant differences between the CLIP- and No CLIP-matched randomized groups after EED ablation (Extended Data Fig. 5a, b). Thus, CLIP genes represent a unique class of Polycomb targets: Already highly expressed, they are further upregulated when PRC2 is ablated. Thus, PRC2 controls gene expression in a non-binary fashion [30–32]. Instead, in the case of CLIP genes, PRC2 may function more like a rheostat to fine-tune expression using an RNA-mediated mechanism.

Interaction between PRC2 and nascent RNA regulates POL-II pausing

To explore the rheostat model, we first asked if CLIP genes fall into specific functional categories. Using *PantherDB* Gene Ontology (GO) [33], we observed that the CLIP group showed enrichment for genes in the cell differentiation pathway and general gene regulation (e.g. *Aebp2, Mybl2, Zfp57*; Fig. 3c, Extended Data Table 5), consistent with PRC2’s well-established role in regulating gene silencing during cell differentiation. The rheostat model is intriguing, given recent reports linking PRC2 to transcriptional elongation of by RNA Polymerase II (POL-II) [30–32]. Mammalian transcription is now understood to be regulated primarily at the level of transcription elongation, rather than initiation [34], via the release of promoter-proximal POL-II pausing[35]. Notably, paused POL-II is always associated with a

short nascent transcript [36]. Furthermore, PRC2 is now known to methylate Elongin and downregulate POL-II elongation [30]. PRC2 can also be antagonized by elongation factor, SPT6, to maintain stem cell pluripotency [32]. Given the relationship between PRC2 and POL-II elongation, we asked if nascent RNA could be a missing link in the regulation of pause-release.

The tall enrichment peaks around the TSS in EZH2 and SUZ12 dCLIP metagene profiles distinctly contrasted with CBX7's (Fig. 1f), but are consistent with PRC2 binding to promoter-associated transcripts [8,37]. This observation suggests a tantalizing relationship to promoter-proximal pausing. We therefore performed PRO-seq (Precision nuclear run-on sequencing), an epigenomic method that maps POL-II active sites with base-pair resolution and tracks sites of nascent transcription across the genome [38]. To determine likelihood of pausing in CLIP versus No CLIP genes, we calculated the "Pausing Index" (PI) based on POL-II's "Traveling Ratio" [39,40], which compares POL-II densities at −30 to +300 relative to TSS versus the rest of the gene body. In two PRO-seq biological replicates, the PI of CLIP genes was significantly higher than that of No CLIP genes (Fig. 3d). To exclude the possibility that this difference was due to a difference in expression levels per se of the two gene sets, we performed a statistical analysis to examine the difference in median PI in 10,000 iterations of randomized RPKM-matched gene sets and observed highly significant difference from the observed CLIP/NoCLIP values (Fig. 3e). We also utilized PRO-seq data to empirically map true pausing sites based on local signal maxima of POL-II at promoter-proximal regions (Extended Data Fig. 5c, d). Indeed, a metagene analysis of the 414 CLIP genes showed EZH2 and SUZ12 dCLIP peaks just downstream of the POL-II pause site (Fig. 3f).

To test the pausing model, we generated PRC2-ablated 16.7 mES cells by degron-tagging the SUZ12 subunit at the C-terminus using the recently developed FKBP-Cereblon dTAG system [41]. The degron tag resulted in a nearly complete abrogation of SUZ12 protein and a concomitant abrogation of H3K27me3 levels in homozygous clones (Fig. 4a). There was no significant change in levels of pluripotency factors (Extended Data Fig. 6a), suggesting no major effect on mES pluripotency [42]. When subjected to differentiation conditions, SUZ12-degron cells showed higher levels of pluripotency factors and lower levels of differentiation markers as expected (Extended Data Fig. 6b) [42]. However, Xist RNA still increased in the female mES cells lacking SUZ12 (Extended Data Fig. 6c), suggesting that the differentiation defect is relatively mild. These results indicate that SUZ12-degron cells are a suitable experimental system to explore a link between PRC2-RNA interactions and transcriptional elongation in differentiating mES cells.

We performed PRO-seq in the SUZ12-degron cells and found a dramatic effect on transcriptional elongation of CLIP genes. The PI for CLIP genes was significantly increased relative to no CLIP genes, indicating reduced pausing and increased elongation in PRC2-ablated cells (Fig. 4c). Concordantly, *edgeR* differential analysis of altered PRO-seq signal revealed a significantly increased response in CLIP genes (Fig. 4d). In wildtype cells, the representative CLIP genes, *L1td1*, *Mybl2*, *Nodal*, and *Ogdhl*, exhibited strong POL-II pausing at the promoter-proximal end just downstream of the TSS (red asterisk, Fig. 4e, Extended Data Fig. 6d). We note that, even with strong pausing, elongation still occurred,

albeit at lower levels, explaining why CLIP genes have robust expression in wildtype cells (Fig. 3b). In SUZ12-degron cells, pausing was visibly reduced and elongation strikingly enhanced, as evidenced by dramatically increased in POL-II density in the gene body (Fig. 4e). In contrast, No CLIP genes, *Hspb11*, *E2f5*, and *Ctsb*, continued to show promoter-proximal pausing (red asterisk, Fig. 4f; Extended Data Fig. 6d).

Together, these findings explain the upregulation of CLIP genes and lack of effect on no CLIP genes in EED-KO cells (Fig. 3a,b). Thus, active genes subject to PRC2-RNA rheostat regulation can further increase expression in a PRC2-sensitive manner. On the other hand, highly expressed genes without an RNA-associated PRC2 (i.e., No CLIP genes) do not respond to PRC2 dissociation. These data therefore point to a distinct mechanism in which POL-II pause-release is regulated by a complex containing nascent RNA and PRC2.

Identification of RNA motifs for PRC2 binding

We asked if PRC2-RNA motifs underlie pausing. The dCLIP method was previously developed to produce footprints small enough to enable motif identification [25]. For EZH2 and SUZ12, we employed a multi-layered pipeline [25] and identified four PRC2-binding motifs in the CLIP set (Fig. 5a). Interestingly, EZH2 and SUZ12 both favored these motifs, suggesting that the two subunits bind to similar RNA sequences and/or to motifs located in proximity to each other. Two of them (P7, P14) were G-rich, consistent with a previously noted in vitro preference for G-rich RNA [19,22]. Two others were not G-rich, but were instead C- and U-rich (P1, P10; Fig. 5a), contrasting with the prior study [22]. This difference in in vivo preferences may reflect accessory factors inside cells [43].

To validate these motifs, we sampled various CLIP transcripts and investigated their affinities for PRC2. Using P14 as a test case, we carried out in vitro binding assays for two P14-containing CLIP transcripts, Adipocyte-Enhancer Binding Protein 2 (Aebp2) and Heat Shock Protein Family A Member 5 (HspA5). In electrophoretic mobility shift assays (EMSA), a robust upward shift of both Aebp2 and HspA5 probes, with a measured affinity (K_d, dissociation constant) of 93.31±4.79 nM and 11.31±0.32 nM, respectively (Fig. 5b,c). Analysis of four other P14-containing CLIP targets also revealed high affinities (Fig. 5d–g) — Fbxo5 with a K_d of 28.52±1.43 nM, Cbx7 with 35.33±4.74 nM, Mcm2 with 77.29±2.75 nM, and Erlec1 with 12.49±0.64 nM. These affinities were comparable that for Xist Repeat A (20–80 nM) [14,21,23]. When G-nucleotides were mutated to C's, all 6 binding sites exhibited dramatically lowered affinities — e.g., Fbxo5 from K_d of 28.52±1.43 nM to 2058±1274 nM, *Mcm2* from 77.29±2.75 nM to 3258±2133 nM. Thus, PRC2 binds the new motifs with high specificity.

PRC2 motifs linked to promoter-proximal pausing of POL-II

To test a functional linkage between PRC2 motifs and POL-II pausing, we performed UV-crosslink RNA immunoprecipitation (UV-RIP) with RT-qPCR and observed enriched pulldown of the RNA by EZH2 and SUZ12 in multiple biological replicates (Fig. 6a), confirming the physical interactions in vivo. We then asked if there is a general enrichment for PRC2 motifs downstream of pause sites. All 6 tested target genes (Fig. 5) harbored P14 motifs at promoter-proximal regions and, significantly, EZH2 and SUZ12 dCLIP peaks

occurred just downstream of pause sites defined by PRO-seq (Fig. 6b, Extended Data Fig. S7). Metagene analysis also unveiled a general tendency for PRC2 motifs to cluster just downstream of pause sites (Fig. 6c). Furthermore, when CLIP genes were segmented into 3 bins based on proximity of motifs to pause sites, motifs within 2 kb were most sensitive to SUZ12 perturbation, as demonstrated by a larger ΔPI (Fig. 6d). Altogether, these data link the EZH2/SUZ12 motifs to promoter-proximal pausing.

Potential to form RNA G-quadruplex

G-rich motifs can potentially fold into four-stranded RNA structures termed RNA G-quadruplexes (rG4) [44,45]. As one in vitro study proposed that PRC2 has specificity for rG4 motifs [22], we asked whether G-rich motifs identified in our study have rG4 potential using two *in-silico* strategies: i) *PQSfinder* [46], an algorithm that prioritizes presence of four consecutive G's segmented by loops of semi-arbitrary lengths; and ii) *G4RNA Screener* [26], an artificial neural network trained with sequences of experimentally validated rG4s. To validate the tools, we identified bioinformatically all putative rG4 sites in an arbitrary set of 18 genes and plotted their rG4 content against gene expression response to an rG4-stabilizing compound, carboxypyridostatin (cPDS) [47,48], calculated the linear correlation between the two variables, and observed that rG4 content is positively correlated with elevated transcript levels using either algorithm (Fig. 7a, Extended Data Fig. 8a).

For PRC2 dCLIP libraries, we calculated an rG4 ratio (% of dCLIP summit regions with rG4 potential) and tested rG4 enrichment relative to two CLIP controls — CTCF [27] and CBX7 [25] — over a randomized RNA sequences from differentiating mES transcriptome. Both *G4RNA Screener* and *PQSfinder* revealed a significant rG4 enrichment in EZH2 and SUZ12 dCLIP replicates relative to randomized controls (Fig. 7b, Extended Data Fig. 8b). While CBX7 also showed enrichment, it was less enriched than PRC2. In contrast, CTCF showed a strong depletion relative to randomized controls (Fig. 7b, Extended Data Fig. 8b).

We then asked if CLIP transcripts have a higher rG4 probability density compared to No CLIP transcripts, using *G4RNA Screener* and *PQSfinder* to screen for rG4 in CLIP summits versus simulated summits generated by random sampling of No CLIP transcripts. Density profiles showed significantly higher rG4 content in CLIP summits (left - Fig. 7c, Extended Data Fig. 8c). When analyzed across the entire gene, there was also a significantly higher likelihood of rG4 in CLIP transcripts (right - Fig. 7c, Extended Data Fig. 8d). rG4 was also markedly enriched at empirically defined POL-II pause sites (Fig. 7d), thereby coinciding with EZH2 and SUZ12 RNA binding (Fig. 3f). Thus, multiple strands of research align on the conclusion that EZH2 and SUZ12 bind rG4 and this feature of nascent transcripts mediates PRC2's role in promoter-proximal pausing of POL-II.

PRC2-P14 motif binding reduces transcriptional activity

Next, we tested the effect of PRC2-P14 interactions on transcription inside cells. Motif mutations in multiple P14-containing transcripts, including Aebp2, HspA5, and Fbxo5, abrogated PRC2 binding in vitro (Fig. 5b,c). We recreated the same P14 mutations for a reporter assay by cloning the promoters and 5'UTR with their respective pause sites and P14 motifs into a promoterless FireFly luciferase (Luc) vector [49] (Fig. 7e). By comparing Luc

expression in mES cells carrying wildtype (P14-WT) versus mutant (P14-Mut) constructs, we found ~2-fold higher expression for mutants relative to WT, for both *Aebp2* and *HspA5* (Fig. 7f). If this effect were mediated by P14-PRC2 interactions, abrogating SUZ12 might result in one of two outcomes. If PRC2-P14 solely induced pausing, depleting PRC2 might not further increase expression, as PRC2 and P14 would be in the same pathway. However, if PRC2's trimethylation of H3-K27 or other functions also contributed to transcription regulation, ablating PRC2 could cause a further increase in Luc expression.

To test these possibilities, we used the SUZ12-degron system (Fig. 4a,b) to abrogate PRC2 function in clone 14H. For Aebp2 and HspA5, no further increase in Luc expression was observed, suggesting that PRC2-P14 interactions operated predominantly at the level of POL-II pause-release. For *Fbxo5*, on the other hand, mutating P14 and abrogating SUZ12 caused a further increase in Luc expression, suggesting that *Fbxo5* is regulated by both pause-release and additional PRC2's functions. Because all three CLIP genes are already highly expressed, their further upregulation through the loss of PRC2-P14 interaction is especially notable. Collectively, our data argue for PRC2-P14 interactions as the basis for transcriptional suppression inside cells. For CLIP genes, the mechanism is mediated predominantly by POL-II pausing, but parallel pathways such as the classic H3K27-trimethylation may facilitate suppression.

RNA also targets PRC2 in cis and promotes POL-II pausing at neighboring genes

Heretofore, we have only considered PRC2 bound to nascent RNAs produced from the same genes. We now return to the classical model that RNA targets PRC2 to neighboring genes in cis [20,50,51] by asking CLIP transcripts impact expression of neighboring genes (Fig. 8a). For each CLIP gene, we identified a pair of nearest neighbor genes (upstream, downstream) — so-called "CLIP neighboring genes" (CLIP-NG, n=218). We compared effects of ablating PRC2 on CLIP-NG versus neighboring genes of No CLIP genes (No CLIP-NG, n=238) and of Canonical genes (Canonical-NG, n=279). CDP showed a marked right-shift (upregulation) of CLIP-NG, compared to No CLIP-NG ($P < 8e-04$, Wilcoxon test) (Fig. 8a), with an overall fold-change of +1.5 for CLIP-NG (Fig. 8b). Thus, PRC2-RNA interactions can also target or spread silencing to neighboring genes.

We asked if CLIP transcripts suppress neighboring genes also by facilitating POL-II pausing. First, we excluded other underlying factors. In WT cells, CLIP-NG and No CLIP-NG both showed low expression (Fig. 8b), indicating that transcription is not a distinguishing feature between neighbors that do or do not respond to PRC2-RNA disruptions. Pausing indices were also similar for CLIP-NG and No CLIP-NG ($P < 0.72$, Wilcoxon test; Fig. 8c). Furthermore, while canonical genes have larger distances to their NN, the NN distances were not different for CLIP-NG and No CLIP-NG (Extended Data–Fig. 9a, $p<0.11$). Because the ability of noncoding RNAs to spread PRC2's H3K27me3 activity has been correlated with transcript abundance and stability [51], we performed an RNA stability assay on CLIP versus No CLIP transcripts using 5,6-Dichlorobenzimidazole 1-b-D-ribofuranoside (DRB) to block transcription and measure RNA decay rates over 8-hours. CLIP transcripts were only slightly more stable than No CLIP (Extended Data–Fig.

9b). Lastly, analysis of H3K27me3 ChIP-seq coverage also showed no differences between CLIP-NG versus No CLIP-NG (Extended Data–Fig. 9c).

However, analysis of ChIP-seq coverage of Serine-5-phosphorylated POL-II (POL-II-S5P) — the isoform associated with transcription initiation or "poised" POL-II — revealed a significant difference. POL-II-S5P coverage in the promoter regions of Canonical, CLIP, and No CLIP genes all showed good correlation with expression levels, as expected (Fig. 8b,d), but there was strikingly stronger enrichment of POL-II-S5P over promoters of CLIP-NG relative to No CLIP-NG and Canonical-NG (Fig. 8d). We therefore conclude that, in the classical model whereby nascent RNA targets PRC2 to a neighboring gene, the repressive mechanism also at least partially involves an effect on POL-II pausing to block the transition to productive elongation.

DISCUSSION

PRC2's interaction with RNA has been intensively debated and disparately interpreted. The classical model proposed that RNA plays a central role in targeting PRC2 to specific loci in cis to silence genes [52]. However, some considered the model incompatible with PRC2-RNA interactions occurring at active genes and with RNA's inhibitory effect on PRC2's catalytic activity [8,18,20]. Our current work reconciles these apparent contradictions. Indeed, unique cis-acting transcripts lend specificity to targeting of chromatin complexes, a function that no other type of factor can assume [4]. Xist RNA is a classic example [6], but there is now an extensive list of such transcripts, including Kcnq1ot1 and TERRA [5,6,10–13]. Our current work shows that nascent transcripts bound to PRC2 exerts a repressive effect on either their own transcriptional elongation and/or on that of other genes in cis (Fig. 8e,f). Although PRC2-RNA interactions could occur on processed transcripts, our study predominantly implicates nascent transcripts because the majority of dCLIP peaks hit introns (Fig. 1d,e) and the effect on pause-release inherently implicates nascent RNA.

Some confusion arose in the aftermath of early studies because PRC2-RNA interactions localize within active genes [8,18], the argument being that if RNA's role were to recruit PRC2, the genes to which PRC2 is recruited should be silent [7,8,18,20]. We reconcile these opposing views with the understanding that gene expression is not binary decision and nascent transcripts associated with PRC2 act as rheostats to fine-tune expression levels (Fig. 8e,f,). The RNA/PRC2 complex serves as a juggernaut for POL-II's ability to advance transcription. At so-called CLIP genes, PRC2 interacts with nascent transcripts via specific motifs at promoter-proximal regions to stall POL-II. Importantly, PRC2-RNA binding reduces, but does not preclude, transcription elongation — explaining how PRC2 can be found within active genes. However, when PRC2 is ablated, pausing is lost and POL-II elongates with enhanced efficiency — accounting for the massive gene upregulation on top of existing expression (Fig. 3,4). By contrast, at No CLIP genes (nascent transcripts not bound by PRC2), pausing is not affected by PRC2 loss, consistent with PRC2 ablation having no effect on No CLIP genes (Fig. 3a,b). Our findings could also explain the paradoxical profiles of so-called "bivalent genes" [53], which express in mES cells despite being marked by repressive H3K27me3 and H2AK119ub1 Polycomb modifications [54,55]. Because RNA from bivalent genes bind PRC2 at their 5' ends [5,37], we suggest that this

interaction might also temper expression via enhanced pausing. Given PRC2's crosstalk with transcription elongation factors, such as Elongin A [30] and SPT6 [32], we postulate that RNA is a missing link in the control of pause-release.

Further confusion had arisen because RNA also blunts PRC2 catalysis [7,14,17]. Yet, as originally proposed [14], the mutually exclusive binding of PRC2 to RNA versus DNA is essential for the regulatory mechanism, as it permits an ordered assembly of PRC2 first on RNA before it is transferred to chromatin [7,14–16,19]. For instance, Xist RNA recruits PRC2 and holds its activity in check until it comes into contact with JARID2 [14,56], which positively regulates the histone methyltransferase activity [14,56]. Altogether, our study demonstrates two parallel functions for PRC2-nascent RNA interactions: (i) Recruitment of PRC2 for the marking of chromatin with H3K27me3, and (ii) the control of POL-II pause-release (Fig. 8e,f).

Thus, PRC2's interaction with RNA is neither promiscuous nor non-specific, and is instead predicated on specific motifs, with affinities in the 10–90 nM range (Fig. 5) on par with Xist's and discrete RNA-binding pockets within the multi-subunit PRC2 complex [14,21,57]. Although this study focused on POL-II pausing and H3K27me3 methylation, PRC2-RNA interactions clearly have other functions. For one, PRC2 can carry out non-histone methylation [30,58]. For another, during the stress response, the EZH2 subunit induces the ribozyme activity of SINE B2 RNA [57,59]. Finally, we note that the 3' UTR is also enriched for PRC2 motifs (Fig. 1f). These 3'UTR-PRC2 interactions could have distinct functions worth future investigation as well.

METHODS

Contact for Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the corresponding author, Dr. Jeannie T. Lee (lee@molbio.mgh.harvard.edu).

Experimental Model and Subject Details

16.7 mouse female embryonic stem cells were described previously [61]. Stem cell lines were routinely maintained in 500 U/ml LIF, DME, and 15% FCS on gamma-irradiated mouse embryonic fibroblasts feeder layer. For differentiation, 7×10^5 cells were plated on pre-gelatinized 150mm TC plates and grown in monolayer for 7 days in DME + 15% FBS without LIF. All RNA-seq and ChIP-seq data were derived from the 16.7 line. In some instances, we use the 16.7 derivative, *Tsix*$^{TST/+}$, which has an X-inactivation choice mutation that affects only which X chromosome will be selected for inactivation and does not affect overall transcriptomics. This point is demonstrated by the high correlation shown in Extended Data Fig. 4d. Both are in the 16.7 background. The one exception is the RNA-seq data for the EED-KO, which we used to determine transcriptomic changes in the absence of PRC2 (Fig. 3a–c, Fig. 9a–c). These data came from a study by Ardehali et al [30], in which the knockout was performed in another ES cell line. Ardehali's ES line from which the KO was created and our 16.7 have very similar transcriptomic profiles. This similarity is shown in Extended Data–Fig. 4f,g.

Stable Transfection

The following plasmid vectors were used for stable transfection into 16.7 mES cells:

pBrCAG – Avi - GFP-mouse EZH2 and pBrCAG – Avi - GFP-mouse SUZ12 plasmids were used for stable expression of Avi-GFP-tagged EZH2 and SUZ12 in dCLIP-seq experiments.

pEF1aBirAV5His plasmid was utilized for stable expression of V5-His-tagged BirA bacterial biotinylase in 16.7 mES cells.

pEF1a-Flag-biotag-PGKpuro-mouse Eed and pBrCAG – mouse Eed – Avi-HA plasmids were employed for stable transfection of mouse EED carrying biotinylation tag in 16.7 cells expressing BirA biotinylase.

pEF1aBirAV5His and pEF1-Flag-Biotag plasmid vectors were a kind gift from Dr. Stuart Orkin (Harvard Medical School) and have been described previously by Kim et al [62]

pBrCAG-Avi-GFP plasmid was a kind gift from Dr. Mitinori Saitou, Department of Anatomy and Cell Biology, Graduate School of Medicine, Kyoto University.

To create mouse ES cells with stable expression of recombinant proteins, 16.7 mouse ES cells were grown to 70% confluence on embryonic feeder layer in T75 flasks. Cells were trypsinized and 2×10^7 cells electroporated with 30μg of linearized vector in PBS using GenePulser II (Bio-Rad). Positive cells were selected using growth media supplemented with 1μg/ml Puromycin (Gibco) alone or in combination with 300μg/ml G418. Stable transfection and expression of recombinant proteins was confirmed by PCR genotyping and Western blotting with specific antibodies.

Denaturing CLIP method – small scale, large scale and library prep

Denaturing CLIP including library prep was conducted essentially as described previously[25]. Detailed protocol is provided in Supplementary Note 1.

Gene Expression

RNA was extracted from cells with Trizol reagent (Thermo Fisher Scientific) according to manufacturer instructions, DNAse - treated, and cDNA libraries were constructed using Superscript III reverse-transcriptase (Thermo Fisher Scientific) and qPCR was performed with specific primers Primer sequences are given in Extended Data Table 6.

Electrophoretic Mobility Shift Assays

RNA-EMSA was conducted essentially as described previously [14,25]. Detailed protocol is provided in Supplementary Note 1.

Western Blotting

Western Blotting was conducted essentially as described previously[25]. Detailed protocol is provided in Supplementary Note 1.

Quantification and Statistical Analysis of qPCR data

Data represents the average ± standard deviation for at least 3 biological replicates as stated in the Fig. legends. P values were determined by unpaired two-tailed student t-test unless otherwise stated.

Analysis of dCLIP-Seq data, RNA-seq data and RNA-seq vs CLIP-seq Analysis

Data analysis was conducted essentially as described previously[25]. Detailed pipeline is provided in Supplementary Note 1.

Gene-specific correlation between EZH2 and SUZ12 dCLIP-seq

We utilized the matrix of total read counts per gene normalized to gene length (see 'RNA-seq vs CLIP-seq Analysis') in order to perform correlation of enriched signals at gene level between EZH2 and SUZ12 dCLIP-seq. For this analysis we considered genes that had enriched dCLIP-seq signals in at least one out of two biological replicates in both PRC2 subunits ('homo-binders'), and genes that had dCLIP-seq signal in two biological replicates of only one subunit and had shown no signal in two biological replicates of the other subunit ('hetero-binders') (Extended Data Fig. 2c). In parallel, by employing the corresponding matrix of normalized read counts obtained for CBX7 dCLIP-seq [25] we performed correlation analysis between CBX7 and each of PRC2's subunits (Extended Data Fig. 2d).

Heat maps Assembly

For depicting the antagonistic interplay of PRC2 binding to either ChIP or RNA transcripts we first called *MACS2* peaks (qvalue cutoff = 0.05, -f BAMPE, and -bdg) for EZH2 and H3K27me3 (narrow, and broad, respectively), by contrasting each sample (uniquely aligned reads) against its corresponding control sample. Fold-enrichment signal (*FE*; IP sample's signal over its corresponding Input sample's signal) was calculated per each sample by executing *MACS2 bdgcmp* using *--method FE*. By applying *Homer* [63] suit's *makeTagDirectory* and *annotatePeaks* algorithms we obtained a fold enrichment (FE) matrix of H3K27me3 and EZH2 ChIP-seq over the proximate promoter region [±3kb from the transcription start site (TSS)] of a set of ~28k known canonical gene transcripts (mm9). We sorted the gene matrix by H3K27me3 occupancy and by employing *deeptools* [64] generated a heat map showing the densities of H3K27me3 and EZH2 ChIP-seq signals. By applying *Homer* [63] we summarized uniquely aligned reads overlapping the composite (merged) track of all strand-specific enriched *PeakRanger* peaks raised from four PRC2 dCLIP-seq libraries, and obtained a read counts matrix over gene bodies normalized to gene length, of four PRC2 dCLIP-seq libraries over gene bodies of all known canonical genes. The distribution of dCLIP signal over gene bodies within this heat map is indicated by a density plot (top, black line) and compared to their density over a randomly permutated heat map (top, dashed gray line) – Fig. 2a.

For mapping signals of PRC2 dCLIP-seq (EZH2 and SUZ12 interactomes) and ChIP-seq (EZH2 or H3K27me3 (GSM905445 and GSM905446, respectively)), we employed seqMINER [65] by using the gene bodies of genes displaying either PRC2-dCLIP or PRC2-ChIP signals as reference coordinates (Fig. 2b). For depicting profiles of PRC2 dCLIP-seq signals, we used uniquely aligned reads of PRC2 dCLIP-seq libraries overlapping the

composite (merged) track of all strand-specific enriched *PeakRanger* peaks raised from four PRC2 dCLIP-seq libraries (two EZH2 replicates, and two SUZ12 replicates) (see "*PeakRanger*" peak calling under STAR Methods' "Analysis of dCLIP-Seq data" section). For depicting profiles of ChIP-seq libraries, we first employed MACS2 (v2.1.1.20160309) with parameters (qvalue cutoff = 0.05, -f BAMPE) to call peaks for EZH2 and H3K27me3 (narrow, and broad, respectively), and depicted with seqMINER uniquely aligned reads overlapping the enriched MACS2 peaks. We generated the heatmap depicted in Fig. 2b by utilizing seqMINER for clustering genes based on their gene body dCLIP-seq signals and identified a gene cluster ("dCLIP" cluster) containing genes exhibiting a reproducible pattern of PRC2 dCLIP-seq signal in all four dCLIP libraries (n=2,038). We identified in parallel a second larger cluster ("ChIP" cluster) containing genes that display a strong PRC2-ChIP (H3K27me3 and EZH2) signal (n=4,783). To assess reproducibility and specificity within a pair of PRC2 dCLIP-seq (or ChIP-seq) libraries we calculated Spearman correlations for each of the genes (rows) depicted in the "dCLIP cluster" by comparing a 200-bins density matrix between the two paired libraries (Extended Data–Fig. 2). Altogether, we performed 14 rounds of pairwise analysis comparing each dCLIP-seq library to any one of the other dCLIP-seq libraries (homo-pairs), or ChIP-seq libraries (hetero-pairs). The distribution of Spearman correlation coefficients of each of the 14 pairwise comparisons (6 homo-pairs and 8 hetero-pairs) was presented as a boxplots group. Box boundaries represent 25th and 75th percentiles; the center line represents the median; whiskers indicate ±1.5 times the interquartile range (IQR).

Reproducibility of biological replicates of dCLIP-seq

To determine reproducibility among dCLIP peaks, we utilized *deepTools* [64] analysis. We applied 1-kb bin size and per each PRC2 subunit compared the significance values (−log(p-value)) of strand-specific peaks enriched in at least one out of two replicates, per bin (bins with enriched signal in a single replicate, who had no enrichment in both replicates of the complement PRC2 subunit, were discarded). Pairwise-Pearson correlation (PPC) analysis was undertaken for evaluating reproducibility between two biological replicates, and scatter plots were generated (Extended Data Fig. 2a). Overall positive correlation was observed with PCC ranging from 0.3 to 0.35 (Extended Data Fig. 2a). In parallel, we also carried out same correlative analysis between the merged strand-specific enriched dCLIP peaks of EZH2 and SUZ12 (Extended Data Fig. 2b). The obtained PCC values were positive albeit of much lower magnitude compared to these obtained between biological replicates of the same PRC2 subunit (Extended Data Fig. 2a).

For verifying concordance between biological replicates at sequence similarity level we employed an artificial network generating platform previously developed by Garant et. al. [26]. The platform, written in python and implementing the PyBrain library, supports the training of a custom artificial neural network (ANN), and is available at http://gitlabscottgroup.med.usherbrooke.ca/J-Michel/g4rna_screener_dev along with its manual and documentation. We first tested the platform ability to generate accurate machine learning models by employing the platform's script 'train_from_fasta.py' for training an ANN based on a previously published G-quadruplex (G4RNA) dataset [26]. We then utilized the platform's script "screen.py" by employing the ANN model and a testing set for

generating prediction scores. The predictive power of the generated model, determined by calculating the area under the receiver operating characteristic (ROC) curve (AUC), was high (AUC > 0.95) (data not shown) and comparable to the predictive power (AUC > 0.92) of the original model (G4RNA_2016–11-07.pkl) previously generated by Garant et al. [26], thus attesting for the platform's capability to generate highly accurate ANN models. Next, per each of the PRC2 dCLIP-seq libraries, we generated a training set composed of 50% positive ('True') and 50% negative ('False') cases. For generating a 'True' subset we randomly selected half of the library's summit regions (100 bp in size) and fetched their transcriptome sequences in FASTA format. For creating a 'False' subset we randomly picked an equal number of simulated summit regions from the expressed transcriptome of 16.7 D7 cells. Per each of the libraries, we employed its training set for generating an ANN, and followed by performing accuracy testing (validation) using a testing set that was generated in the same manner as the training set. To this end we utilized the platform's script "screen.py" by employing the ANN model that we generated for estimating prediction scores for the testing set. We found that the predictive power scores (AUC) of all four generated ANNs were fair, ranging between 0.71 and 0.78 (Extended Data Table 2). While these accuracy values attested of a fairly good prediction power, they also indicated that the models had a relatively low classifier efficiency in comparison to the model we trained on G-quadruplex sequences (see above). These lower accuracy scores were however expected given the complex nature of PRC2 subunits interactomes that are furthermore diversified than a well-defined sequence dataset such as G-quadruplex. In order to evaluate the sequence concordance between two biological replicates, we created per each library an additional testing set composed of summit region sequences randomly picked from its replicate library. In addition, we also generated two testing set controls by using the summit region sequences of two CTCF CLIP-seq libraries previously generated by our lab [27]. We measured the prediction accuracy scores of these three testing sets by using the library's ANN, calculated AUC scores and plotted per each of the libraries four ROC curves corresponding to (1) testing sequences of the library itself; (2) testing sequences of the corresponding biological replicate library; (3+4) testing sequences of two control samples (CTCF CLIP-seq). To obtain statistical margins of for the accuracy measurements performed for each library, we employed the analysis described above five times. ROC curves depicted in Extended Data Fig. 3 are representatives of one out of five repeated measurements. Extended Data Table 2 presents all mean accuracy values and their margins.

Metagene, Genomic Features and Biological Process Analysis

Data analysis was conducted essentially as described previously[25]. Detailed pipeline is provided in Supplementary Note 1.

Gene Classification and CDP plotting

To investigate whether the RNA-binding activity of PRC2 may be involved in modulation of gene expression, we composed a three-tier database consisted of (1) PRC2 interactome, (2) H3K27me3 chromatin-binding mapping, and (3) PRC2-dependent transcriptome, all generated in WT embryoid bodies during the mid-stage of differentiation (~D7). We compiled our PRC2 subunits interactomes (two biological replicates per each, EZH2 and SUZ12 dCLIP-seq) by employing *Homer* [63 for generating read count matrix summarized

over gene bodies and normalized to gene length. Similarly, we generated for H3K27me3 ChIP-seq (GSM905446), a fold enrichment (FE) matrix summarized over gene bodies, including minimal promoter region (1 kb upstream of the TSS) and normalized by gene +promoter length [29]. We took advantage of a recently produced whole genome RNA-seq dataset (GSE104657) of WT (control) and Eed-null embryoid bodies (EB), containing two biological replicates per condition [30]. We employed sRAP [66] for performing data normalization and calculation of log2 ratios (based on averaged signals of two replicates per each condition), and statistical significance (p-value). We characterized three gene groups based on their direct physical interactions with PRC2 (Extended Data Table 4): (1) Canonical genes: genes enriched by H3K27me3 (FPKM>1), with significantly increased ($p < 0.05$) transcript levels in EED-KO mES cells relative to their control counterparts; (2) CLIP genes: genes whose RNA transcripts were highly enriched in at least 2 out of 4 PRC2 dCLIP-seq interactomes, H3K27me3 deposition levels were low or absent (FPKM<1), and expression levels were co-regulated in both biological replicates upon EED-KO ($p < 0.1$). (3) No CLIP genes: genes whose RNA transcripts were depleted of all four PRC2 dCLIP-seq interactomes, H3K27me3 deposition levels were low or absent (FPKM<1), and expression levels were co-regulated in both biological replicates upon EED-KO ($p < 0.1$). To avoid bias introduced by lowly expressed genes we ignored in both, CLIP and No CLIP gene groups, lowly expressed genes that their expression levels were ranked in the lowest 10^{th} percentile. Per each of the gene groups, we plotted the fold changes in gene expression between EED-KO cells versus their control counterparts, as a cumulative distribution plot (CDP). CDP plots, boxplots, and scatter plots were constructed with R software (www.R-project.org).

Randomized expression-matched controlled analysis: Due to the inherent differences in basal expression levels of CLIP genes and No CLIP genes we performed randomized controlled analysis in order to rule out the possibility that the strong right shift of CLIP genes was an outcome of their higher gene expression levels, or of a random event. We created a gene pool consists of all expressed genes that were not considered as either CLIP, No CLIP or Canonical genes (as before, we ignored lowly expressed genes with expression levels ranked below the lowest ranked 10^{th} percentile of CLIP genes). We then computed per each of the gene groups (CLIP and No CLIP) its kernel density estimator for the group's gene expression levels (RNA-seq). Based on the kernel's fitted density probabilities, we then randomly sampled per each of the gene groups (CLIP and No CLIP) its corresponding number of genes from the gene pool. We repeated the randomization process 10,000 times thus creating two control gene groups: CLIP-matched genes, and No CLIP-matched genes, mirroring expression levels of CLIP and No CLIP genes, respectively. As shown by boxplots depicted in Extended Data Fig. 5a, the distribution of expression levels of randomly selected genes indeed nicely mirrored the distribution of expression levels of CLIP and No CLIP genes (Fig. 3b). We further calculated the CDP plots of fold change responses to EED-KO (Extended Data Fig. 5b) and observed a very minor difference (=0.07) between the median fold changes of "CLIP-matched" genes versus "No CLIP-matched" genes. We further employed our randomization expression-matched analysis in order to rule out the chance that the significant higher pausing index scores (majored by PRO-seq analysis) of CLIP genes compared to No CLIP genes were a mere consequence of

their relative higher gene expression levels. The randomization procedure performed 10,000 times, subtracted each time between the median pausing indices obtained for randomly selected expression-matched CLIP genes and the median pausing indices of No CLIP genes. The differences in median scores were then plotted as density plot (Fig. 3e) and empirical significance (p-value) was estimated relative to the actual median difference between CLIP gene No CLIP gene groups (5.73).

Motif Analysis of CLIP genes

Motif analysis was conducted essentially as described previously[25]. Detailed pipeline is provided in Supplementary Note 1.

Neighboring Genes

Per each of the gene groups (CLIP, No CLIP, Canonical), we identified for each gene, its nearest pair of neighbor genes residing upstream and downstream relative to its gene body, and co-regulated in both biological replicated upon EED-KO ($p < 0.1$). For generating boxplots presenting ChIP-seq levels at CLIP, No CLIP, and Canonical genes, and their corresponding neighbor genes ("CLIP-NG", "No CLIP-NG" and "Canonical-NG"), we generated per H3K27me3 ChIP-seq (GSM905446), and POL-II-S5 (GSM905457) a fold enrichment (FE) matrix summarized over gene bodies, including minimal promoter region (1 kb upstream of the TSS) and normalized by gene+promoter length [29]. Similarly, and as described under STAR Methods' "Gene Classification and CDP plotting" section, we also plotted RNA-seq signals (RPKM) as boxplots (Fig. 9b). Note that for the purpose of these analyses we focused on all genes (CLIP, No CLIP, or Canonica genesl) that had at least one matched neighboring gene. Genes without neighboring genes were left out of these analyses.

Luciferase assays

Aebp2, HspA5, Fbxo5, fragments containing partial promoter + 5'UTR sequences were harvested by genomic PCR using Q5 PCR master-mix (NEB) and cloned into pCR-Blunt-II-Topo vector (Thermo Fisher Scientific). P14 motif mutation was introduced using QuikChange Site-Directed Mutagenesis Kit (Agilent). WT and mutated sequences were cloned into pGL4.14 vector (Promega) [49] via Acc65I + EcoRV restriction (NEB), to enable expression of FireFly luciferase ORF fused to promoter + 5'UTR sequences derived from *Aebp2, HspA5, Fbxo5* genes. Promoterless vector plasmid pGL4.14 served as control. For transfecting cells we used 2.25 μg pGL4.14 – derived plasmid DNA into 1×10^6 16.7 mES cells in suspension, in total volume of 250 μl Opti-MEM medium, with 7.5 μl Lipofectamine 3000 reagent and 5 μl P3000 reagent, according to manufacturer's instructions (Thermo Fisher Scientific). Per each reaction, 0.25 μg Renilla luciferase – expressing pGL4.74 plasmid was introduced as transfection control. After 20min incubation on room temperature, transfected cells were seeded on gelatin-coated surface inside a 6-well tissue culture plates in normal mES cells growth media. After 24hrs incubation on 37^0C in humidified conditions, cells were washed once with 2 ml PBS and harvested in 1 ml Trizol reagent (Thermo Fisher Scientific). RNA was extracted following manufacturer's instructions, with 1.7 μl 15 mg/ ml GlycoBlue reagent (Thermo Fisher Scientific). RNA was eluted in 100 μl nuclease-free DDW and treated with 1 μl (2 u) TurboDNAse (Thermo Fisher Scientific) per 10 μg RNA for 30min on 37^0C in final concentration of 200 μg/ ml RNA.

Then, the entire volume was transferred to Phase-Lock Gel Heavy 2 ml tubes (5 Prime GmbH) and extracted twice with equal volume of acid phenol:chlorophorm solution (AM9722, Thermo Fisher Scientific) according to manufacturer's protocol. The remaining phenol was removed by adding half-volume of 24:1 chloroform:isoamyl alcohol solution (Sigma). The aqueous phase was collected, and ethanol precipitated by adding $1/10^{th}$ volume of 3M NaAcetate and 3 volumes of 100% ethanol. The mixture was incubated 1hr on -80^0C or overnight on -20^0C. RNA was precipitated 15min 21,130xG 4^0C. Sup removed, and pellets washed 1×1 ml 75% ethanol, 5min 21,130xG. Elution was performed with 50μl nuclease-free water. qPCR assays were performed on CFX96 real-time PCR system (Bio-Rad). Specific primers are listed in Extended Data Table 6. Threshold cycle values were utilized to calculate FireFly luciferase transcript expression relative to promoterless construct, with Renilla luciferase transcript serving as a reference gene.

rG4 evaluation

2×10^6 Day 6 16.7 mouse mES cells were seeded on gelatin-coated 6-well dishes. After 24hrs, media was replaced with a 2 ml aliquot containing various concentrations of carboxypyridostatin (cPDS). After 24hrs, cells were washed with 2 ml PBS and harvested in 1 ml Trizol reagent (Thermo Fisher Scientific). RNA was extracted following manufacturer's instructions, with 1.7 μl 15 mg/ ml GlycoBlue reagent (Thermo Fisher Scientific). RNA was eluted in 100 μl nuclease-free DDW and treated with 1μl (2u) TurboDNAse (Thermo Fisher Scientific) per 10 μg RNA for 30 min on 37^0C in final concentration of 200 μg/ml RNA. Then, the entire volume was transferred to Phase-Lock Gel Heavy 2 ml tubes (5 Prime GmbH) and extracted twice with equal volume of acid phenol:chloroform solution (AM9722, Thermo Fisher Scientific) according to manufacturer's protocol. The remaining phenol was removed by adding half-volume of 24:1 chloroform:isoamyl alcohol solution (Sigma). The aqueous phase was collected, and ethanol precipitated by adding $1/10^{th}$ volume of 3 M NaAcetate and 3 volumes of 100% ethanol. The mixture was incubated 1 hr on -80^0C or overnight on -20^0C. RNA was precipitated 15 min 21,130xG 4^0C. Sup removed, and pellets washed 1×1 ml 75% ethanol, 5 min 21,130xG. Elution was performed with 50μl nuclease-free water. qPCR assays were performed on CFX96 real-time PCR system (Bio-Rad). Specific primers are listed in Extended Data Table 6. Threshold cycle values were utilized to calculate transcript expression relative to untreated control with β-actin RNA serving as a reference gene.

rG4 *in silico* analysis

For quantifying the content of RNA G-quadruplexes (rG4) within our CLIP and No CLIP genes we employed two bioinformatics methods: (1) *PQSfinder*[46]: an algorithm-based tool identifying putative rG4 elements according to presence of consensus motif composed of four consecutive guanine runs separated by semi-arbitrary loops; and (2) *G4RNA Screener*[26,67]: artificial neural network tool that was trained by sequences of experimentally validated rG4 elements for predicting probable rG4 structures in a candidate sequence. For identifying rG4 elements using the *PQSfinder* tool, we applied default parameters, (i.e. regular expression constraints of G[1,10].[0,9]G[1,10]), and minimal PQS score of 40 (as demonstrated by the user guide). Per each of the four PRC2 dCLIP-seq libraries we scanned the FASTA sequences of 100bp summit regions in a strand-specific manner for identifying

putative rG4 elements and calculated the fraction of summit regions harboring at least one rG4 element ("rG4-ratio"). As a comparison, we screened in parallel, the summit regions of two CTCF CLIP libraries obtained from differentiating cells (D3) (GSM1540988 and GSM1540990) [27]. In order to assess the statistical significance of rG4-ratio obtained for each CLIP library, we employed empirical test approach by creating per each sample 2,000 random sets, each simulating an equal number of summit regions as in the tested library. For generating each set of random regions we utilized the entire expressed transcriptome of 16.7 D7 mES cells and used the Bioconductor suite "regioneR" for randomly selecting (without overlap) random sets of 100bp-RNA regions (simulating 100bp summit regions of each of the original dCLIP libraries). We then employed *PQSfinder*, using the same settings described above for identifying putative rG4s within each of the random RNA sets, for calculating per each set its rG4-ratio. Finally, we plotted the rG4-ratio of each of the original CLIP libraries (illustrated as circular dots) along with the rG4-ratios of its corresponding series of 2000 random control sets as boxplot (Fig. 7b).

In order to sift through RNA transcripts of dCLIP transcriptomes and score their similitude to the G4 folding sequences of G4RNA database via artificial neural network, we employed the machine learning algorithm *G4RNA Screener* [26,67]. The python source code ("screen.py"), as well as its artificial neural network's pickle file: "G4RNA_2016–11-07.pkl" (http://gitlabscottgroup.med.usherbrooke.ca/J-Michel/g4rna_screener), were installed on our slurm server, and executed serially by the server's high-performance processor. FASTA sequences of summit regions (100bp) from each of four PRC2 dCLIP libraries were screened using the composite of all three default threshold parameters implemented by the web interface version (cGcC ≥ 4.5 & G4H ≥ 0.9 & G4NN ≥ 0.5) as a minimal reporting criteria. We scanned each summit region thoroughly, by applying a gradient of 9 different window sizes (ranging from 15 bp to 55bp) and a step length between windows (-s) of 1 bp. Next, per each summit region we combined (merged) all reported rG4 sequences into one rG4 track. We finally calculated per each of the libraries the fraction of summit regions harboring at least one rG4 element ("rG4-ratio"). As in our *PQSfinder* screening procedure, for comparison, we analyzed in parallel the summit region sequences of two CTCF CLIP libraries [27], and evaluated the statistical significance of rG4-ratios by carrying out empirical test approach using 2,000 random sets (per each sample) picked from the entire expressed transcriptome of 16.7 D7 mES cells, while applying the same screening *G4RNA Screener* parameters used to screen dCLIP libraries (see a more detailed description above, related to the *PQSfinder* protocol) (Fig. 7a).

To determine if the abundance per gene of rG4 elements detected within the summit regions of CLIP genes ("empirical" rG4 content) was higher relative to that detected in equivalent transcriptome content of No CLIP genes, we did the following analysis: First, we created two simulated sets of RNA sequences randomly selected within the gene bodies of No CLIP genes, and which their lengths mirrored the lengths of merged summit regions of CLIP genes. We then employed *G4RNA Screener* and *PQSfinder* tools for screening putative rG4s within these summit regions (and simulated regions) and calculated per each gene the accumulated rG4 coverage. We divided this accumulated rG4 coverage by the gene length to obtain the "empirical" percentage of putative rG4 content per gene (Fig. 7b).

To determine if the nascent RNA of CLIP genes had in general a higher potential to generate rG4 elements compared to No CLIP genes (“potential” rG4 content), we performed a complementary analysis by screening rG4 across the entire gene body of both CLIP genes and No CLIP genes. The accumulative rG4 coverage was divided by the gene length to obtain the “potential” percentage of putative rG4 content per gene (Fig. 7c).

RNA stability

2×10^6 Day 6 16.7 mouse ES cells were seeded on gelatin-coated 6-well dishes. After 24hrs, media was replaced with a 2 ml aliquot containing 75μM 5,6-Dichlorobenzimidazole 1-b-D-ribofuranoside (DRB) transcriptional inhibitor. At appropriate time points, cells were washed with 2 ml PBS and harvested in 1 ml Trizol reagent (Thermo Fisher Scientific). RNA was extracted following manufacturer’s instructions, with 1.7 μl 15 mg/ml GlycoBlue reagent (Thermo Fisher Scientific). RNA was eluted in 100 μl nuclease-free DDW and treated with 1 μl (2 u) TurboDNAse (Thermo Fisher Scientific) per 10 μg RNA for 30 min on 37^0C in final concentration of 200 μg/ ml RNA. Then, the entire volume was transferred to Phase-Lock Gel Heavy 2 ml tubes (5 Prime GmbH) and extracted twice with equal volume of acid phenol:chloroform solution (AM9722, Thermo Fisher Scientific) according to manufacturer’s protocol. The remaining phenol was removed by adding half-volume of 24:1 chloroform:isoamyl alcohol solution (Sigma). The aqueous phase was collected, and ethanol precipitated by adding $1/10^{th}$ volume of 3M NaAcetate and 3 volumes of 100% ethanol. The mixture was incubated 1hr on -80^0C or overnight on -20^0C. RNA was precipitated 15 min 21,130xG 4^0C. Sup removed, and pellets washed 1×1 ml 75% ethanol, 5min 21,130xG. Elution was performed with 50μl nuclease-free water. qPCR assays were performed on CFX96 real-time PCR system (Bio-Rad). Specific primers are listed in Extended Data Table 6. Threshold cycle values were utilized to calculate transcript expression relative to time 0 control with 18S ribosomal RNA serving as a reference gene.

Generation of *SUZ12-degron* cell lines.

FKBP-tag (degron tag) [41] was introduced into the C-terminus of endogenous Suz12 gene, in frame with SUZ12 protein ORF, by CRISPR-assisted homologous recombination. Suz12-C guide RNA (CAGTGTCTGTTCAAAACATG) was designed using the https://zlab.bio/guide-design-resources website and cloned into pSpCas9(BB)-2A-GFP vector [68]. Donor vector was prepared by Gibson Assembly cloning kit (NEB) using EcoRI-digested pCR4-Topo vector and overlapping PCR products containing left homology arm, FKBP-tag and right homology arm, which were generated by Q5 PCR master-mix (NEB). For CRISPR-reporter vector [69], 3.6 kb fragment containing CRISPR – target sequence was prepared by genomic PCR and cloned into pCR-Blunt-II-Topo vector (Thermo Fisher Scientific). Then, 2,271 kb ScaI – XmnI fragment of cloned sequence was introduced into Eco53KI – digested pMB1610-pRR-Puro CRISPR reporter plasmid. 20 μg of transfection mix (12.4 μg pSpCas9-Suz12-C gRNA, 6.2 μg AhdI – linearized donor vector and 1.4 μg CRISPR – reporter vector) were nucleofected into 2×10^6 16.7 Day 0 mES cells (Lonza Nucleofector II, Mouse ES Cell Nucleofector Solution) as per manufacturer’s instructions, using A-030 program. After nucleofection, cells were seeded on 10 cm feeder plates. After 24 hrs incubation on 37^0C, cells were selected with 2 μg/ ml Puromycin for 48 hrs and transferred to normal growth conditions until well-sized colonies developed. Colonies were lifted into

96-well plates. Homologous recombination and expression of recombinant SUZ12 was confirmed by PCR genotyping and Western blotting with specific antibodies.

PRO-seq library construction and data analysis

PRO-seq library construction and dana analysis were performed essentially as described previously was performed as described previously [38] . The detailed protocol is provided in Supplementary Note 2.

Pausing index analysis

For quantifying differences in the retention of promoter-paused POL-II between CLIP and No CLIP genes, we employed a previously published analytic method denoted as “pausing index (PI) analysis” (also known as “traveling ratio (TR) analysis”) [39,40], which compares the ratio between POL-II density in the promoter-proximal region and the gene body region. The detailed pipeline is provided in Supplementary note 2.

Empiric Mapping of POL-II Pausing Sites

To closely determine enrichment of dCLIP-seq signal (and their motifs/rG4s) at POL-II pausing sites we utilized the PRO-seq data of two biological replicates for mapping empirical pausing sites. First, we called narrow peaks by using macs2 and intersected them against the read depth tracks of PRO-seq in order to accurately identify the exact location of peak summit. Second, by using the transcription start site (TSS) coordinates of all mm9 RefSeq genes from the UCSC we annotated to each TSS its most enriched PRO-seq summit within the promoter-proximal region (−30 to +300 relative to TSS). Approximately 95% of all genes had been annotated with a corresponding PRO-seq summit. For the remaining genes we further performed an additional annotation round using an extended promoter-proximal region (−300 to +2000 relative to TSS). For further analysis we considered all genes that their promoter-proximal summits (maxima signal) in two PRO-seq biological replicates were agreeable and distanced no more than 100bp from each other (6,536 and 6,359 genes on the positive and the negative strands, respectively). To assess our method performance, we plotted strand-specific PRO-seq and RNA-seq signals around empirically defined pausing sites and confirmed that the predominant transcriptional signals emanate from the genomic regions located downstream to the defined pausing sites (Extended Data Fig. 5c). We further profiled our empirically defined pausing sites relative to the TSS positions of their annotated Refseq genes and observed that as expected, the vast majority of empirically defined POL-II pausing sites were located approximately −30 to +300 bp relative to TSS (Extended Data Fig. 5d). By employing *deepTools* [64] we profiled the enrichment distribution of dCLIP-seq signal (Fig. 3f), enriched dCLIP motifs (Fig.6c,d), and dCLIP rG4 elements (Fig. 7d) at the vicinity of empirically defined POL-II pausing sites.

UV-RIP for Nascent Transcripts

UV-RIP was conducted essentially as described previously[25]. Detailed protocol is provided in Supplementary Note 1.

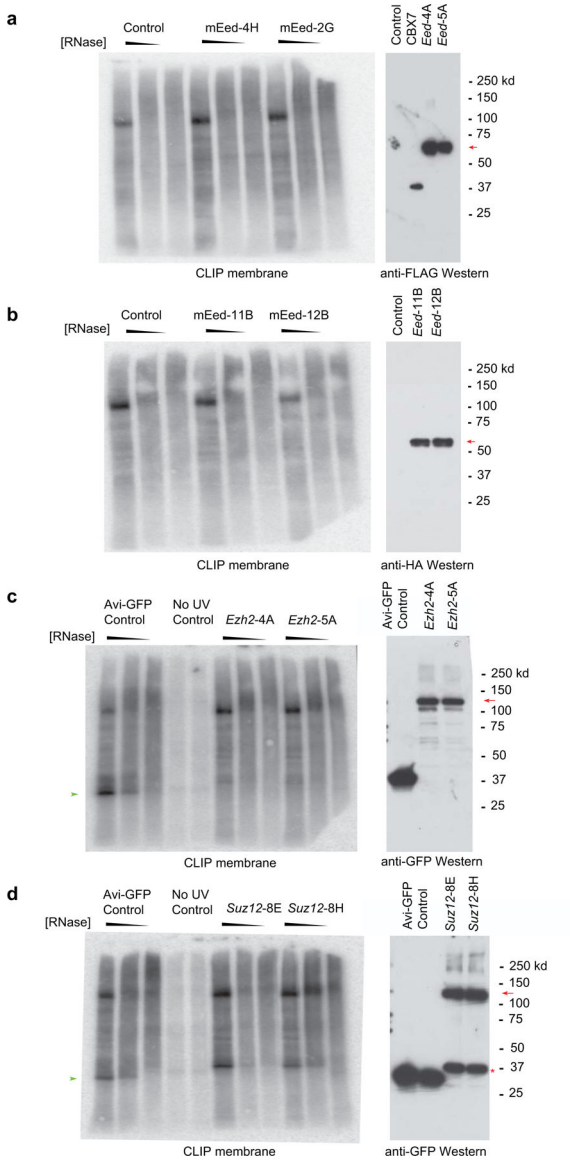
Reporting Summary Statement

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

Data and Software Availability

All sequencing data have been deposited in GEO under the accession number GSE141700.

Extended Data



Extended Data Fig. 1. Denaturing CLIP of EED, EZH2, and SUZ12 in 16.7 mES cells.

(a) Representative dCLIP experiment with N-terminally – Flag-Biotagged EED protein. Left panel, autoradiography of dCLIP experiment. Right panel, Western blot with anti-FLAG

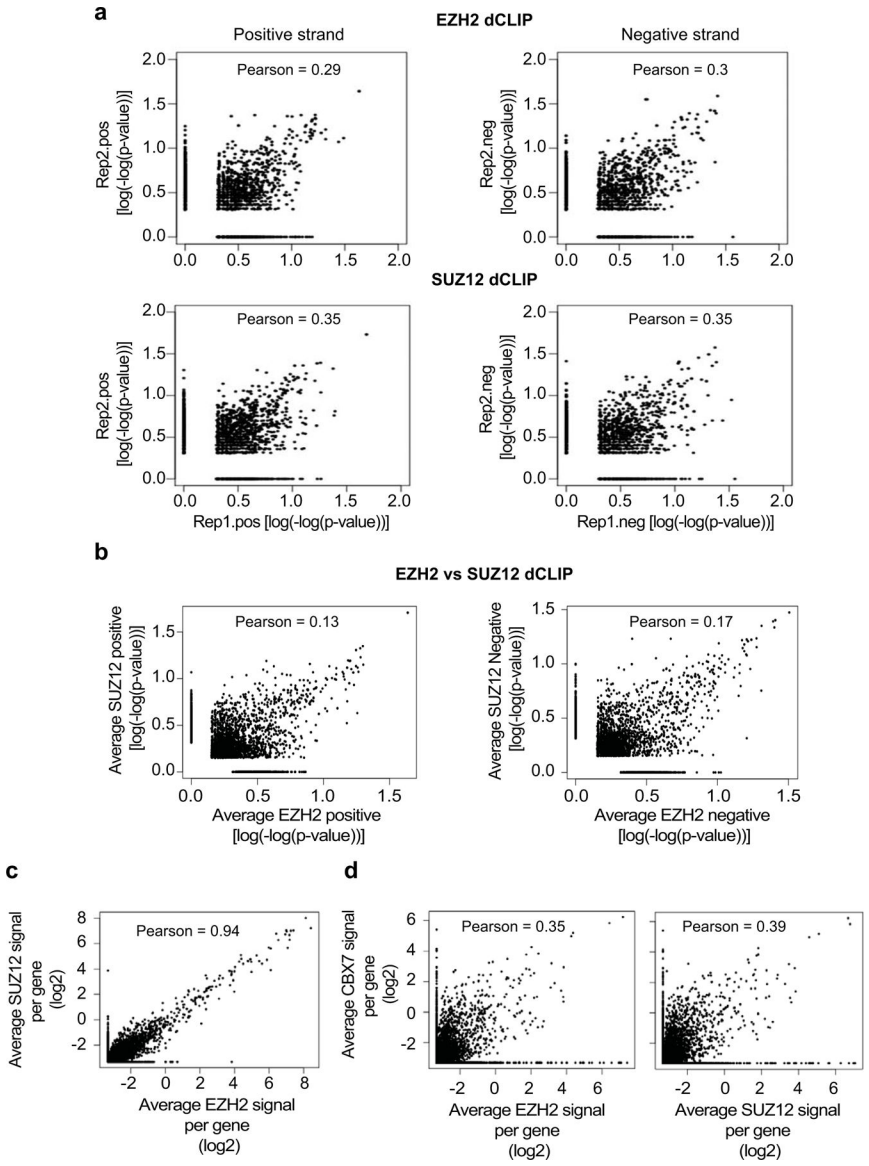
antibody. CBX7-Flag was used as control for FLAG Western. Red arrow, FLAG-Biotagged-EED signal.

(b) Representative dCLIP experiment with C-terminally – HA-Biotagged EED protein. Left panel, autoradiography of dCLIP experiment. Right panel, Western blot with the anti-HA antibody. Red arrow, HAG-Biotagged-EED signal. *mEed-4H* and *mEed-2G* are two clonal cell lines expressing physiological levels of FLAG-Biotagged-EED. *mEed-11B* and *mEed-12B* are two clonal cell lines expressing physiological levels of HA-Biotagged-EED. Note the lack of EED-specific dCLIP signal in both panels.

(c) Representative dCLIP experiments for EZH2. Left panel, autoradiography of dCLIP experiment. Right panel, Western blot with the anti-GFP antibody. Red arrows, GFP-Biotagged-EZH2 / SUZ12 signal. *Ezh2*-4A and *Ezh2*-5A are two clonal 16.7 mES cell lines expressing physiological levels of GFP-Biotagged-EZH2.

(d) Representative dCLIP experiments for SUZ12. *Suz12*-8E and *Suz12*-8H are two clonal cell lines expressing physiological levels of GFP-Biotagged-SUZ12. Red asterisk in (b) depicts short N-terminal truncated fragment characteristic for N-terminally-tagged SUZ12 protein [76].

n=3 for all the representative experiments. Unprocessed blots are provided in Source data Extended Data Fig.1



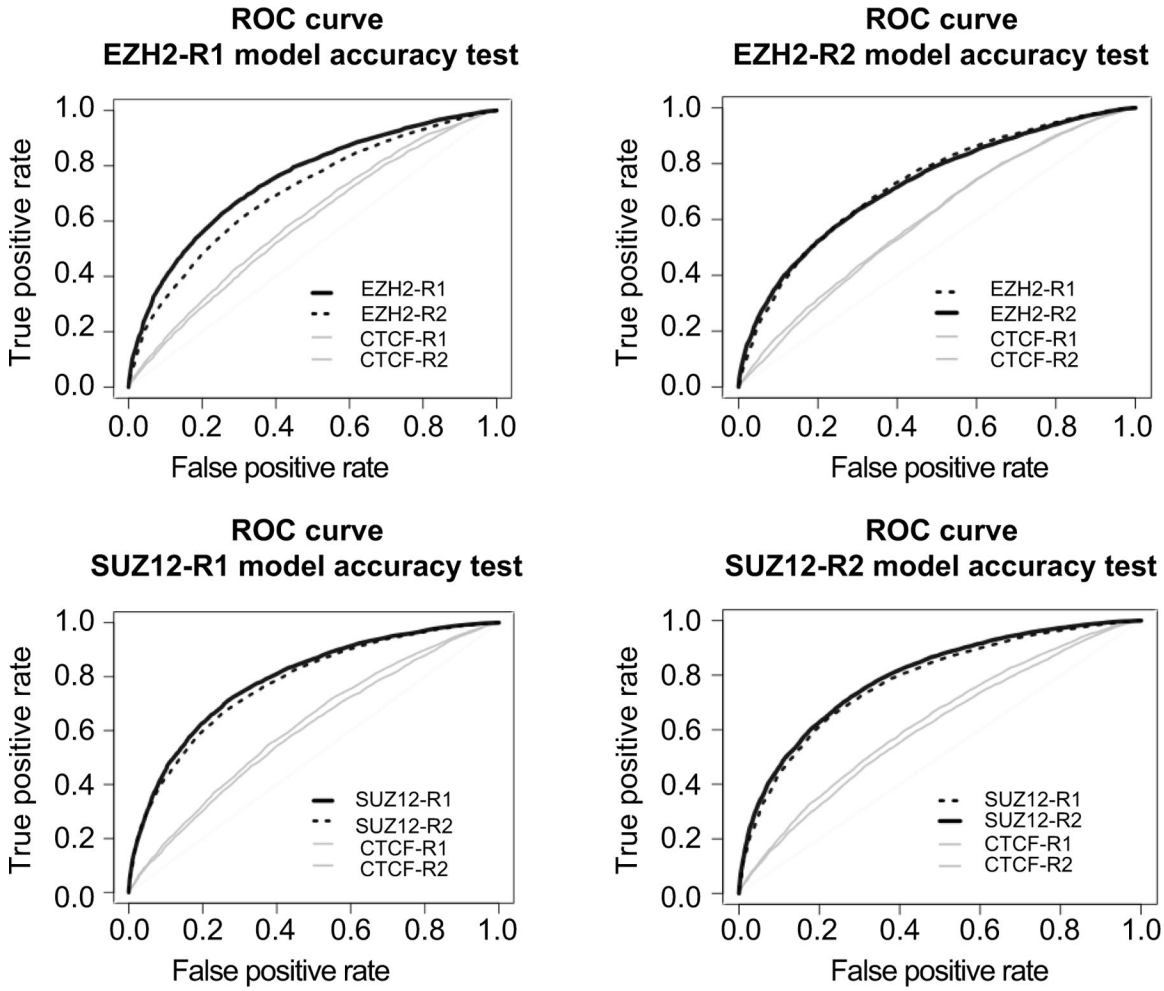
Extended Data Fig. 2. Correlation between biological dCLIP replicates and EZH2 versus SUZ12 dCLIP.

(a) A genome-wide pairwise comparisons of enriched dCLIP peaks over 1kb bins per two biological replicates of EZH2 and SUZ12 dCLIP-seq samples (see Methods for details). Pearson's correlation coefficients (R^2) are shown.

(b) A genome-wide pairwise comparisons of enriched dCLIP peaks over 1kb bins between EZH2 and SUZ12 dCLIP-seq. Pearson's correlation coefficients (R^2) are shown.

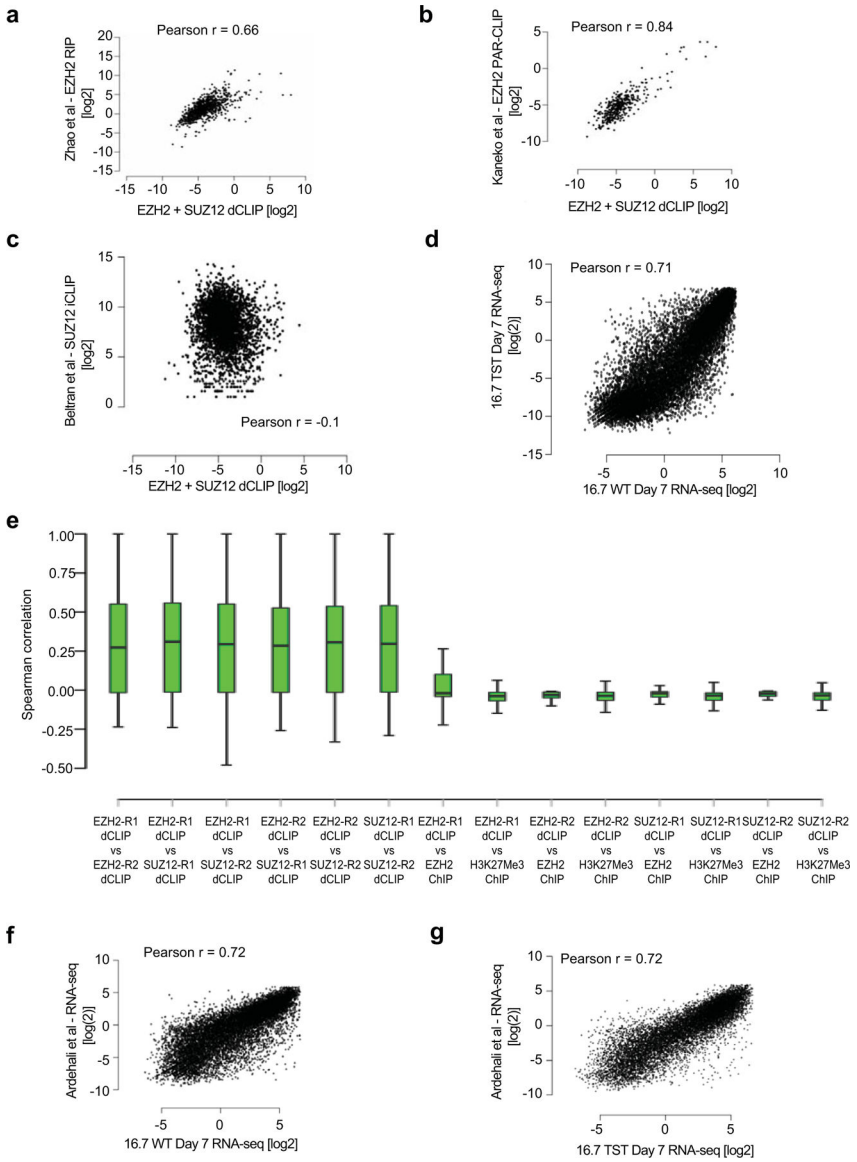
(c) Gene-based pairwise comparisons of enriched dCLIP peaks between EZH2 and SUZ12 dCLIP-seq. Pearson's correlation coefficients (R^2) are shown.

(d) Gene-based pairwise comparisons of enriched dCLIP peaks between EZH2 / SUZ12 dCLIP-seq and CBX7 dCLIP [29]. Pearson's correlation coefficients (R^2) are shown. Note a much lower correlation between EZH2/SUZ12 dCLIP enriched genes and CBX7 dCLIP enriched genes, despite the same method being applied in both studies.



Extended Data Fig. 3. Machine learning modeling.

Receiver Operating Characteristic (ROC) curves for artificial neural network of each of dCLIP-seq libraries, along with ROC curves for testing sets generated from a corresponding biological replicate of dCLIP-seq, and two control samples of CTCF CLIP. For each of PRC2 subunits we introduced the interactome sequences obtained from one biological replicate, trained an ANN model and performed accuracy testing (validation) using sequences obtained from the same interactome as well as the corresponding biological replicate. Two samples of CTCF interactome [31] were employed as a control. The predictive power of each of the four ANNs generated based on the RNA sequences of dCLIP-seq libraries was determined by the average area under the ROC curve (AUC). See also Extended Data Table 2 and text for further details.



Extended Data Fig. 4. Correlation between transcriptomic datasets.

(a) Comparison of enriched peaks signal per gene to EZH2-RIP data from Zhao et al [13].

(b) Comparison of enriched peaks signal per gene to EZH2 PAR-CLIP data from Kaneko et al [10].

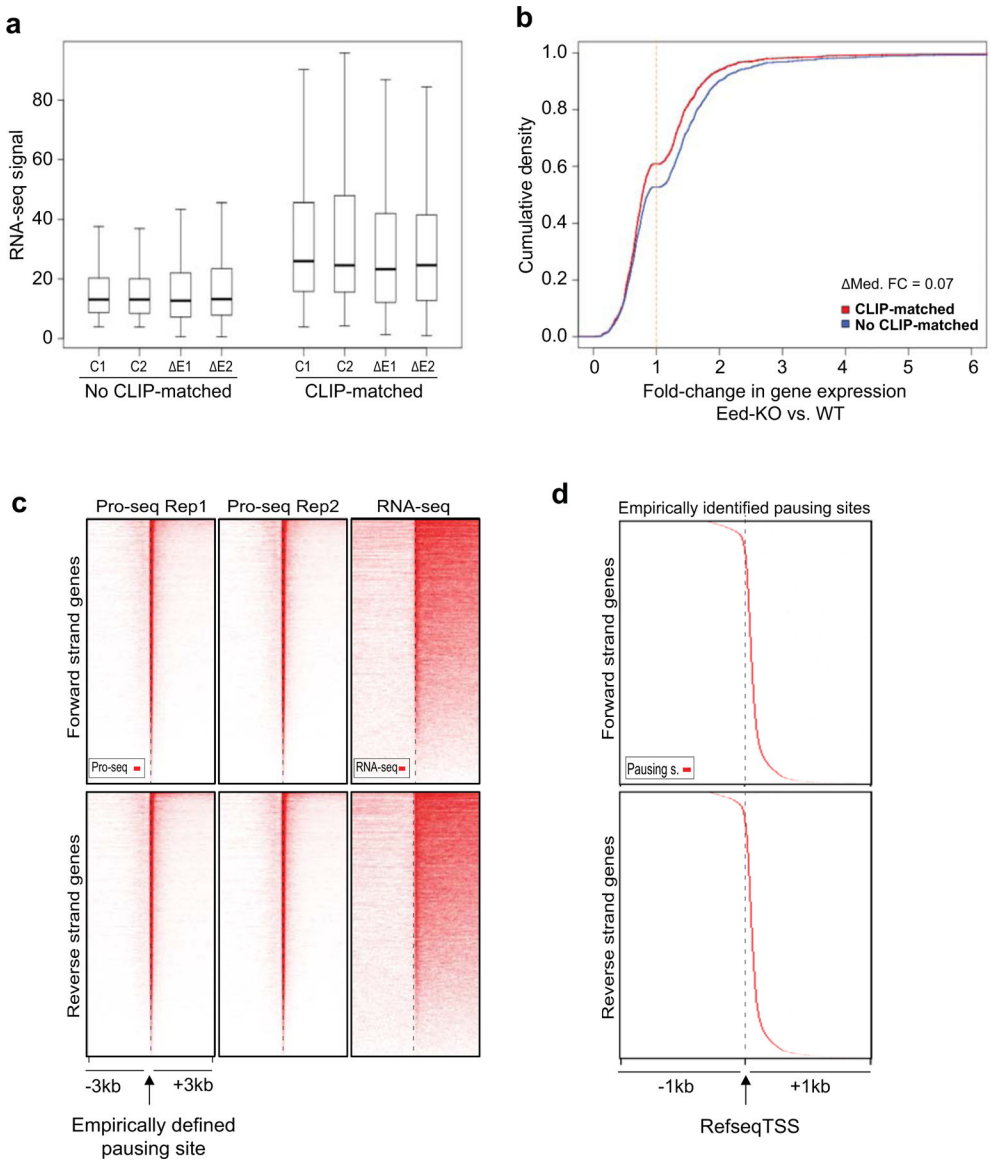
(c) Comparison of enriched peaks signal per gene to SUZ12 iCLIP data from Beltran et al [6]. (enrichment values per-gene obtained from GSE120696 were used).

(d) Wildtype 16.7 cells used in this study and *Tsix*$^{TST/+}$ 16.7 cells [33] have a similar transcriptomic profile on d7, as the *Tsix*^{TST} allele affects only the choice of which X chromosome will be inactivated, but does not affect the general transcriptome. Shown is a comparison of RPKM values per gene for the two cell lines.

(e) Concordance between homo-pairs of dCLIP-seq libraries and hetero-pairs of dCLIP-seq libraries and ChIP-seq libraries. Spearman correlation was calculated within each library pair for each of the regions (rows) depicted in the heatmap by comparing between the two

matrices of normalized gene body density signal. The distribution of Spearman correlation coefficients per gene (n=2038) of each of the library pairs was presented as a boxplots group. Box boundaries represent 25th and 75th percentiles; the center line represents the median; whiskers indicate ±1.5 times the interquartile range (IQR). EZH2 and H3K27Me3 ChIP-seq datasets were from Pinter et al [33] and originated from 16.7 cells at Day 7 of differentiation.

(f,g) Comparison of RNA-seq signal per gene in 16.7 WT Day 7 mES cells (f) or *Tsix*^{TST}/+ 16.7 mES cells [33] (g) versus Day 5 WT CJ7 mouse ES cells from Ardehali et al [34] (RPKM values per-gene were obtained from GSE104657 – sub-datasets GSM2805147 and GSM2805148). *16.7 and Tsix*^{TST}*/+ transcriptomic profiles showed good correlation with profile from* Ardehali et al [34]. Note high Pearson correlation values in (d), (e) and (f).



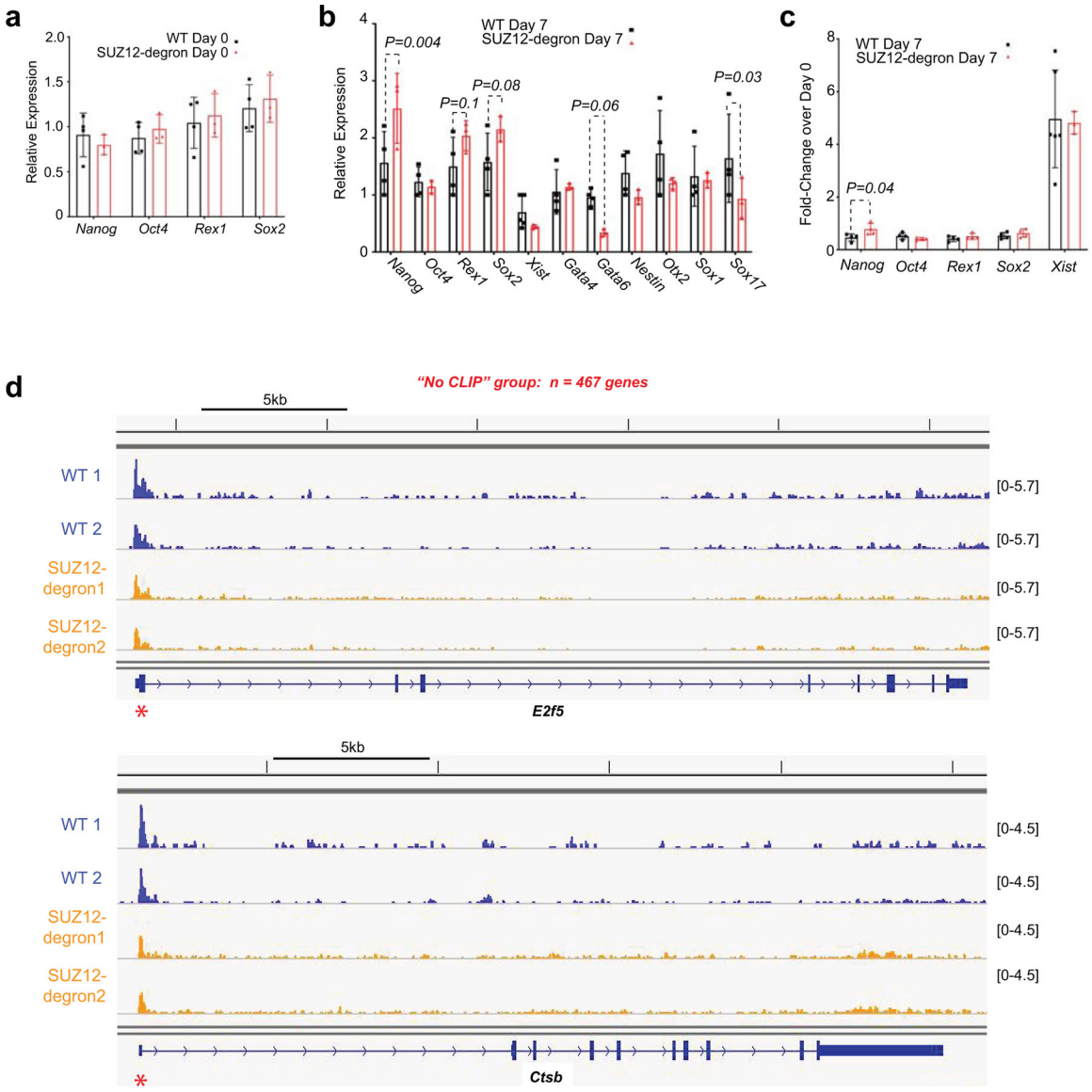
Extended Data Fig. 5. Randomized controls and Distribution of Pausing Indices for the data presented in Figure 3e.

(a) To rule out the possibility that the strong right shift observed in CLIP genes (Fig. 3a) was due to higher gene expression levels, we generated randomized models for CLIP and No CLIP categories matched for expression levels of CLIP and No CLIP gene groups, respectively (see Methods for more details). Boxplots for each of the two gene groups (No CLIP-matched (n=467), and CLIP-matched (n=414)) showing the distribution of gene expression measured by RNA-seq (RPKM) in two WT samples and two EED-KO samples. Box boundaries represent 25th and 75th percentiles; center line represents the median; whiskers indicate ± 1.5× IQR).

(b) A comparison of gene expression changes following Eed knockout (EED-KO) in the randomized model: CLIP-matched control group compared to No CLIP-matched control group. We plotted CDP curves for CLIP-matched controls (n=414), and No CLIP-matched controls (n=467), depicting fold-change alterations in expression levels (RNA-seq) upon EED-KO. Statistical significance of differential gene expression response to Eed ablation was calculated by Wilcoxon test (unpaired, one-sided) between CLIP-matched gene group (red) versus No CLIP-mathced gene group (blue).

(c) Heatmap depicting PRO-seq (two biological replicates) and RNA-seq signal distribution in the vicinity of empirically defined POL-II pause sites. These pause sites were utilized for the analysis in Fig. 3f.

(d) Localization distribution of empirically defined POL-II pause sites relative to Refseq-annotated TSS. See Methods for further details.

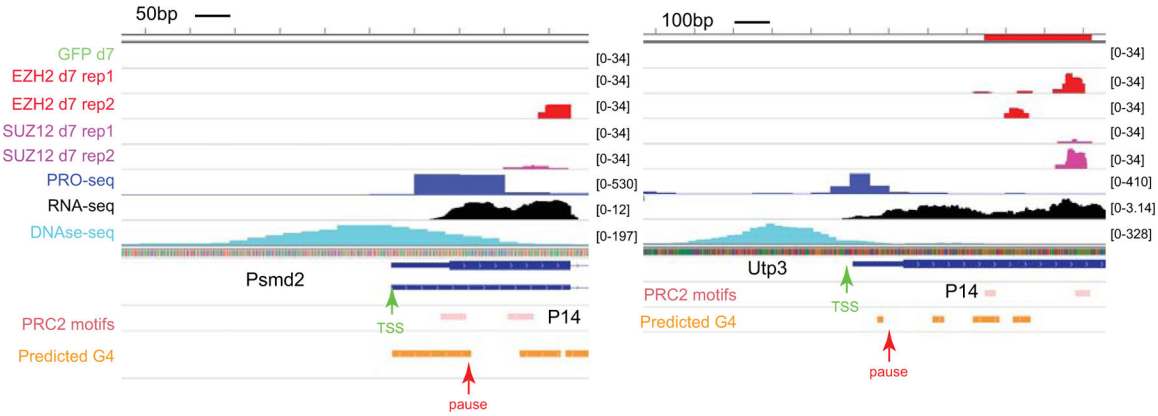


Extended Data Fig. 6. Effect of SUZ12-degradation on pluripotency factors and differentiation markers.

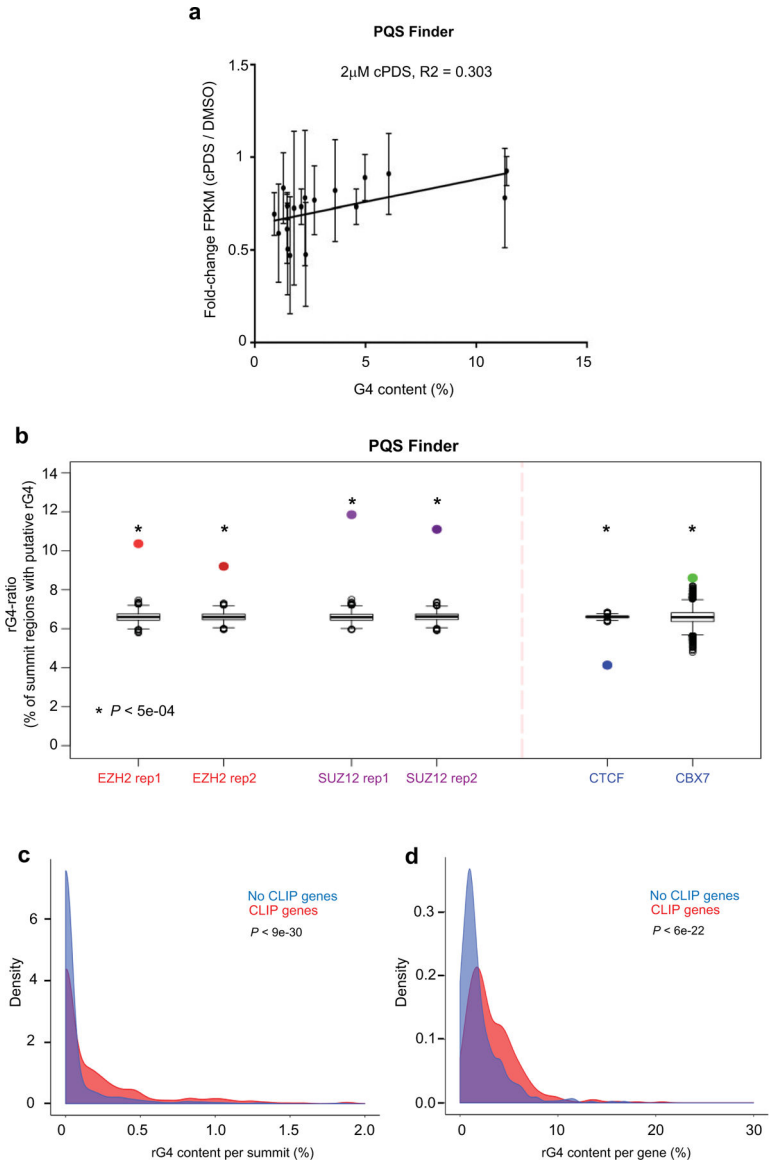
(a) Relative expression of pluripotency factors in undifferentiated 16.7 cells vs 16.7 SUZ12-degron cells, normalized to b-actin. Four biological replicates of 16.7 WT cells and 3 biological replicates of 16.7 SUZ12-degron cells were analyzed. One replicate of 16.7 WT cells was set as a reference (Expression = 1) and other replicates of WT and all values for SUZ12-degron cells were normalized to this replicate.

(b) Relative expression of pluripotency factors (Nanog, Oct4, Rex1 and Sox2), X chromosome inactivation marker (Xist) and differentiation markers (Gata4, Gata6, Nestin, Sox1 and Sox17) in Day 7 differentiated WT cells vs SUZ12-degron cells. Data normalized to B-actin. 4 biological replicates of 16.7 WT cells and 3 biological replicates of 16.7 SUZ12-degron cells were analyzed. One replicate of 16.7 WT cells was set as a reference (Expression = 1). The significance was determined by unpaired Student t-test.

(c) Fold-change in the expression of pluripotency factors (Nanog, Oct4, Rex1 and Sox2) and X chromosome inactivation marker (Xist) in Day 7 WT and SUZ12-degron cells.
(d) Additional examples for No CLIP genes shown in Figure 4f.
Statistical source data for (a,b,c) is provided in Source Data Extended Data Fig.6.



Extended Data Fig. 7. Additional examples of representative genes shown in Figure 6b.
Additional examples of genes that manifest PRC2 dCLIP signals at their 5' regions. Note proximity of G-rich RNA-binding motif (P14) within PRC2 dCLIP peaks. G4RNA Screener tool indicated potential rG4-forming structures (orange bars). Pausing sites were defined empirically by PRO-seq signals (summits) for active POL-II. DNAse-hypersensitivity sites were from Vierstra et al [67]. See also Figure 6b.



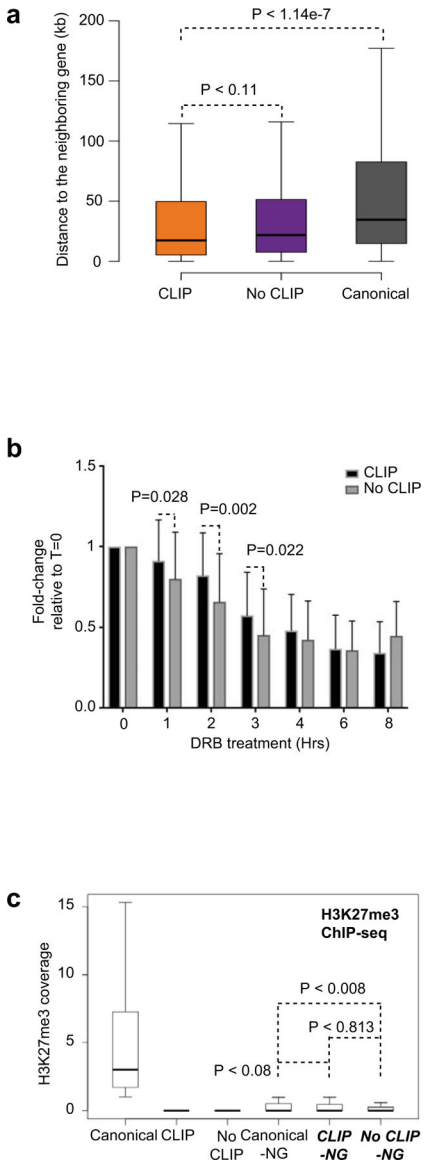
Extended Data Fig. 8. G-quadruplex motifs are enriched in interacting transcripts and POL-II pause sites.

(a) Positive correlation between RNA stability measured following treatment with rG4 stabilizer cPDS (2μM) [relative to DMSO controls] and putative rG4 gene content screened by artificial enrual network algorithm, *PQSfinder*. RNA stability data presented is Mean +/− SD of 4 biological replicates.

(b) *PQSfinder* demonstrates that rG4 is highly enriched in CLIP targets relative to No CLIP transcripts. rG4 ratio defined as % of summit regions with putative rG4. rG4-ratio of PRC2 dCLIP summit-regions (100 nt around the most significant binding site ±50 nt) harboring putative rG4 element was plotted for two EZH2 and two SUZ12 replicates, with CTCF and CBX7 as controls. Black box plots: rG4-ratios of 2,000 random sets, each simulating an equal number of summit regions as in the tested library.

(c) Left: rG4 abundance per gene (“Empirical rG4 content”) is higher in CLIP versus No CLIP transcripts. rG4 elements (per *PQSfinder*) in CLIP summit regions and in simulated

No CLIP summit regions. Density plots of rG4 content per gene are presented for CLIP and No CLIP genes, respectively. Right: Nascent RNAs of CLIP genes have a higher potential to generate rG4 structures compared to nascent transcripts of No CLIP genes. *p* values by Wilcoxon test (unpaired, one-sided).
Statistical source data for (a) is provided in Source Data Extended Data Fig.8.



Extended Data Fig. 9. RNA also targets PRC2 to neighboring genes to control POL-II pausing.
(a) Boxplots depicting the distribution of distances between Canonical-NG (n=279), CLIP-NG (n=218), or No CLIP-NG (n=238) and the linked TSS of nearest corresponding neighboring gene. Box boundaries represent 25th and 75th percentiles; center line represents the median; whiskers indicate ± 1.5× IQR). P values were determined using One-sided Wilcoxon test. Note the insignificant differences in distances between CLIP genes and No CLIP genes, whereas the distances of Canonical genes were significantly longer.

(b) Comparison of RNA stability between selected CLIP genes (n=13) and No CLIP genes (n=11) in Day 7 16.7 cells. Data presented is a Mean +/− S.D of 3 biological replicates. Significance determined by unpaired student t-test.

(c) Boxplots depicting H3K27me3 ChIP-seq enrichment levels (−1 kb to TTS) in the WT mES cells for CLIP-NG (n=218), No CLIP-NG (n=238) and Canonical-NG (n=279) and corresponding CLIP (n=414), No CLIP (n=467) and Canonical (n=603) genes. ChIP-seq datasets for Day 7 16.7 cells were from Pinter et al [33]. Box boundaries represent 25th and 75th percentiles; center line represents the median; whiskers indicate ± 1.5× IQR). P values were determined using One-sided Wilcoxon test.

Statistical source data for (b) is provided in Source Data Extended Data Fig.9.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank R. Aguilar, H. Lee and H. Sunwoo for sharing reagents and providing support on initial stages of the project. We also thank T. Jégu, A. Kriz, and all the Lee lab members for helpful advice and numerous constructive discussions. We give special thanks to the MGH Next Generation Sequencing Core and Harvard University Nascent Transcriptomics Core for excellent technical assistance. This work was funded by grants from the NIH (R01-HD097665) and HHMI to J.T.L.

REFERENCES

1. Simon JA & Kingston RE Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. Molecular cell 49, 808–24 (2013). [PubMed: 23473600]
2. Margueron R & Reinberg D The Polycomb complex PRC2 and its mark in life. Nature 469, 343–9 (2011). [PubMed: 21248841]
3. Laugesen A, Hojfeldt JW & Helin K Molecular Mechanisms Directing PRC2 Recruitment and H3K27 Methylation. Mol Cell 74, 8–18 (2019). [PubMed: 30951652]
4. Lee JT Epigenetic regulation by long noncoding RNAs. Science 338, 1435–9 (2012). [PubMed: 23239728]
5. Zhao J et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Molecular cell 40, 939–53 (2010). [PubMed: 21172659]
6. Zhao J, Sun BK, Erwin JA, Song JJ & Lee JT Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322, 750–6 (2008). [PubMed: 18974356]
7. Beltran M et al. The interaction of PRC2 with RNA or chromatin is mutually antagonistic. Genome Res 26, 896–907 (2016). [PubMed: 27197219]
8. Kaneko S, Son J, Shen SS, Reinberg D & Bonasio R PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. Nature structural & molecular biology (2013).
9. Guil S et al. Intronic RNAs mediate EZH2 regulation of epigenetic targets. Nature structural & molecular biology 19, 664–70 (2012).
10. Montero JJ et al. TERRA recruitment of polycomb to telomeres is essential for histone trymethylation marks at telomeric heterochromatin. Nat Commun 9, 1548 (2018). [PubMed: 29670078]
11. Pandey RR et al. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell 32, 232–46 (2008). [PubMed: 18951091]
12. Kotzin JJ et al. The long non-coding RNA Morrbid regulates Bim and short-lived myeloid cell lifespan. Nature 537, 239–243 (2016). [PubMed: 27525555]

13. Klattenhoff CA et al. Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. Cell 152, 570–83 (2013). [PubMed: 23352431]
14. Cifuentes-Rojas C, Hernandez AJ, Sarma K & Lee JT Regulatory interactions between RNA and polycomb repressive complex 2. Mol Cell 55, 171–85 (2014). [PubMed: 24882207]
15. Wang X et al. Molecular analysis of PRC2 recruitment to DNA in chromatin and its inhibition by RNA. Nat Struct Mol Biol 24, 1028–1038 (2017). [PubMed: 29058709]
16. Zhang Q et al. RNA exploits an exposed regulatory site to inhibit the enzymatic activity of PRC2. Nat Struct Mol Biol 26, 237–247 (2019). [PubMed: 30833789]
17. Kaneko S, Son J, Bonasio R, Shen SS & Reinberg D Nascent RNA interaction keeps PRC2 activity poised and in check. Genes Dev 28, 1983–8 (2014). [PubMed: 25170018]
18. Davidovich C, Zheng L, Goodrich KJ & Cech TR Promiscuous RNA binding by Polycomb repressive complex 2. Nat Struct Mol Biol 20, 1250–7 (2013). [PubMed: 24077223]
19. Beltran M et al. G-tract RNA removes Polycomb repressive complex 2 from genes. Nat Struct Mol Biol 26, 899–909 (2019). [PubMed: 31548724]
20. Davidovich C & Cech TR The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. RNA 21, 2007–22 (2015). [PubMed: 26574518]
21. Davidovich C et al. Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA. Mol Cell 57, 552–8 (2015). [PubMed: 25601759]
22. Wang X et al. Targeting of Polycomb Repressive Complex 2 to RNA by Short Repeats of Consecutive Guanines. Mol Cell 65, 1056–1067 e5 (2017). [PubMed: 28306504]
23. Long Y et al. Conserved RNA-binding specificity of polycomb repressive complex 2 is achieved by dispersed amino acid patches in EZH2. Elife 6(2017).
24. Khalil AM et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proceedings of the National Academy of Sciences of the United States of America 106, 11667–72 (2009). [PubMed: 19571010]
25. Rosenberg M et al. Denaturing CLIP, dCLIP, Pipeline Identifies Discrete RNA Footprints on Chromatin-Associated Proteins and Reveals that CBX7 Targets 3' UTRs to Regulate mRNA Expression. Cell Syst 5, 368–385 e15 (2017). [PubMed: 29073373]
26. Garant JM, Perreault JP & Scott MS Motif independent identification of potential RNA G-quadruplexes by G4RNA screener. Bioinformatics 33, 3532–3537 (2017). [PubMed: 29036425]
27. Kung JT et al. Locus-specific targeting to the X chromosome revealed by the RNA interactome of CTCF. Mol Cell 57, 361–75 (2015). [PubMed: 25578877]
28. Heger A, Webber C, Goodson M, Ponting CP & Lunter G GAT: a simulation framework for testing the association of genomic intervals. Bioinformatics 29, 2046–8 (2013). [PubMed: 23782611]
29. Pinter SF et al. Spreading of X chromosome inactivation via a hierarchy of defined Polycomb stations. Genome Res 22, 1864–76 (2012). [PubMed: 22948768]
30. Ardehali MB et al. Polycomb Repressive Complex 2 Methylates Elongin A to Regulate Transcription. Mol Cell 68, 872–884 e6 (2017). [PubMed: 29153392]
31. Mousavi K, Zare H, Wang AH & Sartorelli V Polycomb protein Ezh1 promotes RNA polymerase II elongation. Mol Cell 45, 255–62 (2012). [PubMed: 22196887]
32. Wang AH et al. The Elongation Factor Spt6 Maintains ESC Pluripotency by Controlling Super-Enhancers and Counteracting Polycomb Proteins. Mol Cell 68, 398–413 e6 (2017). [PubMed: 29033324]
33. Mi H et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc 14, 703–721 (2019). [PubMed: 30804569]
34. Jonkers I & Lis JT Getting up to speed with transcription elongation by RNA polymerase II. Nat Rev Mol Cell Biol 16, 167–77 (2015). [PubMed: 25693130]
35. Min IM et al. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes Dev 25, 742–54 (2011). [PubMed: 21460038]
36. Adelman K & Lis JT Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet 13, 720–31 (2012). [PubMed: 22986266]
37. Kanhere A et al. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. Mol Cell 38, 675–88 (2010). [PubMed: 20542000]

38. Mahat DB et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc 11, 1455–76 (2016). [PubMed: 27442863]
39. Danko CG et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Mol Cell 50, 212–22 (2013). [PubMed: 23523369]
40. Rahl PB et al. c-Myc regulates transcriptional pause release. Cell 141, 432–45 (2010). [PubMed: 20434984]
41. Nabet B et al. The dTAG system for immediate and target-specific protein degradation. Nat Chem Biol 14, 431–441 (2018). [PubMed: 29581585]
42. Pasini D, Bracken AP, Hansen JB, Capillo M & Helin K The polycomb group protein Suz12 is required for embryonic stem cell differentiation. Molecular and cellular biology 27, 3769–79 (2007). [PubMed: 17339329]
43. Sarma K et al. ATRX Directs Binding of PRC2 to Xist RNA and Polycomb Targets. Cell 159, 1228 (2014). [PubMed: 28898627]
44. Fay MM, Lyons SM & Ivanov P RNA G-Quadruplexes in Biology: Principles and Molecular Mechanisms. J Mol Biol 429, 2127–2147 (2017). [PubMed: 28554731]
45. Kwok CK & Merrick CJ G-Quadruplexes: Prediction, Characterization, and Biological Application. Trends Biotechnol 35, 997–1013 (2017). [PubMed: 28755976]
46. Hon J, Martinek T, Zendulka J & Lexa M pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. Bioinformatics 33, 3373–3379 (2017). [PubMed: 29077807]
47. Biffi G, Di Antonio M, Tannahill D & Balasubramanian S Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. Nat Chem 6, 75–80 (2014). [PubMed: 24345950]
48. Rocca R et al. Molecular recognition of a carboxy pyridostatin toward G-quadruplex structures: Why does it prefer RNA? Chem Biol Drug Des 90, 919–925 (2017). [PubMed: 28459507]
49. Sun S et al. Jpx RNA activates Xist by evicting CTCF. Cell 153, 1537–51 (2013). [PubMed: 23791181]
50. Lee JT & Bartolomei MS X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell 152, 1308–23 (2013). [PubMed: 23498939]
51. Schertzer MD et al. lncRNA-Induced Spread of Polycomb Controlled by Genome Architecture, RNA Abundance, and CpG Island DNA. Mol Cell (2019).
52. Lee JT Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Genes & development 23, 1831–42 (2009). [PubMed: 19684108]
53. Bernstein E et al. Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. Molecular and cellular biology 26, 2560–9 (2006). [PubMed: 16537902]
54. Stock JK et al. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nat Cell Biol 9, 1428–35 (2007). [PubMed: 18037880]
55. Brookes E et al. Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. Cell Stem Cell 10, 157–70 (2012). [PubMed: 22305566]
56. Kaneko S et al. Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. Mol Cell 53, 290–300 (2014). [PubMed: 24374312]
57. Zovoilis A, Cifuentes-Rojas C, Chu HP, Hernandez AJ & Lee JT Destabilization of B2 RNA by EZH2 Activates the Stress Response. Cell 167, 1788–1802 e13 (2016). [PubMed: 27984727]
58. Dobenecker MW et al. Coupling of T cell receptor specificity to natural killer T cell development by bivalent histone H3 methylation. J Exp Med 212, 297–306 (2015). [PubMed: 25687282]
59. Hernandez AJ et al. B2 and ALU retrotransposons are self-cleaving ribozymes whose activity is enhanced by EZH2. Proc Natl Acad Sci U S A 117, 415–425 (2020). [PubMed: 31871160]
60. Vierstra J et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science 346, 1007–12 (2014). [PubMed: 25411453]

Methods References

61. Lee JT & Lu N Targeted mutagenesis of Tsix leads to nonrandom X inactivation. Cell 99, 47–57 (1999). [PubMed: 10520993]
62. Kim J, Cantor AB, Orkin SH & Wang J Use of in vivo biotinylation to study protein-protein and protein-DNA interactions in mouse embryonic stem cells. Nat Protoc 4, 506–17 (2009). [PubMed: 19325547]
63. Heinz S et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576–89 (2010). [PubMed: 20513432]
64. Ramirez F, Dundar F, Diehl S, Gruning BA & Manke T deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res 42, W187–91 (2014). [PubMed: 24799436]
65. Ye T et al. seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res 39, e35 (2011). [PubMed: 21177645]
66. Warden CD, Yuan Y & Wu X Optimal Calculation of RNA-Seq Fold-Change Values. International Journal of Computational Bioinformatics and In Silico Modeling 2, 285–292 (2013).
67. Garant JM, Perreault JP & Scott MS G4RNA screener web server: User focused interface for RNA G-quadruplex prediction. Biochimie 151, 115–118 (2018). [PubMed: 29885355]
68. Ran FA et al. Genome engineering using the CRISPR-Cas9 system. Nat Protoc 8, 2281–2308 (2013). [PubMed: 24157548]
69. Flemr M & Buhler M Single-Step Generation of Conditional Knockout Mouse Embryonic Stem Cells. Cell Rep 12, 709–16 (2015). [PubMed: 26190102]

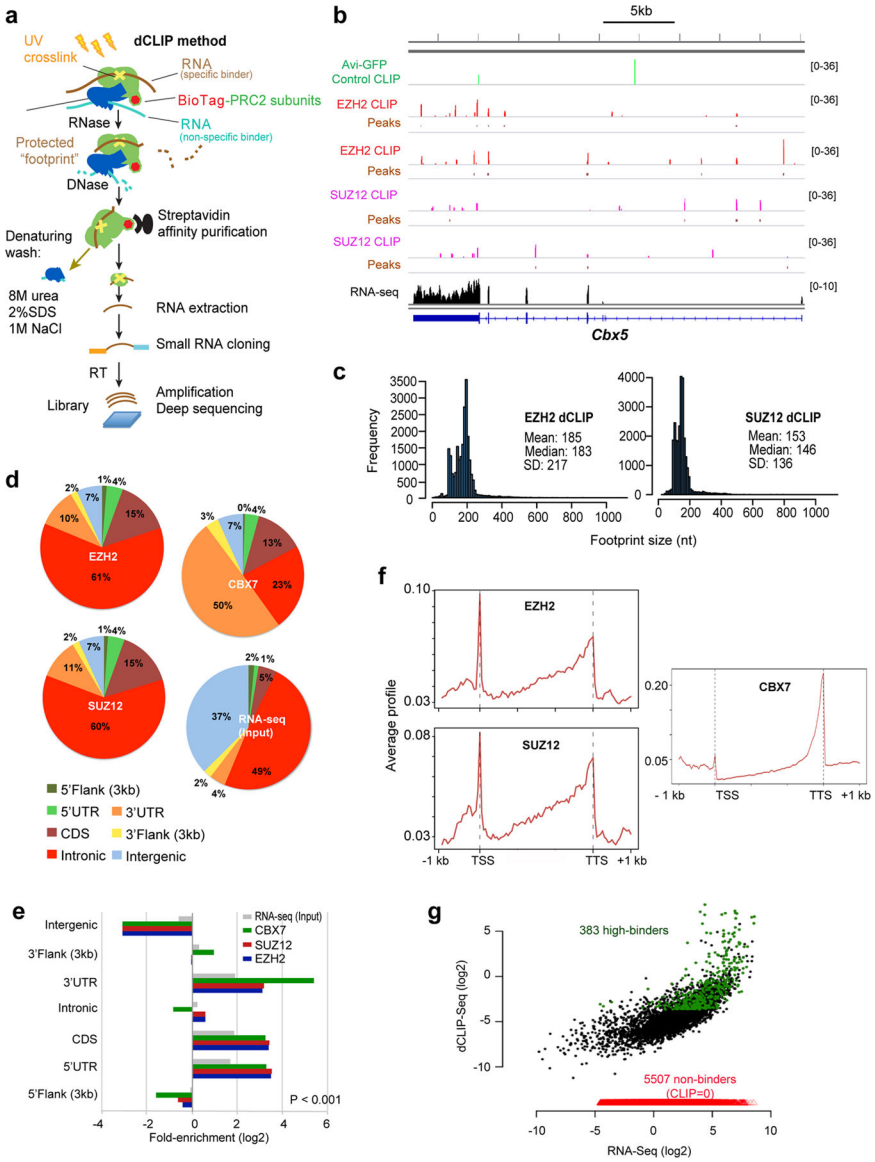


Figure 1. Denaturing CLIP of EZH2 and SUZ12 in Differentiating 16.7 mES cells.

(a) Schematic workflow for dCLIP assay.

(b) Representative EZH2 and SUZ12 dCLIP profiles for selected gene, *Cbx5*.

(c) Strand-specific enriched peaks (called by "*PeakRanger*") from two replicate dCLIP libraries for EZH2 and SUZ12, each were pooled, and overlapped peaks were merged into long enrichment regions in a strand-specific manner. Length distribution frequency of the enriched dCLIP peaks, as well as mean, median, and standard deviation were calculated.

(d) Genomic Association Test (GAT) analysis of dCLIP-seq for EZH2 (upper pie), SUZ12 (lower pie), and CBX7 (right pie) interactomes showing distribution of enriched signals in various genomic features. Input RNA-seq from same time point differentiating 16.7 mES cells (left pie) emphasizes differences in genomic feature distribution between dCLIP-seq signal and cellular transcriptome.

(e) GAT enrichment analysis for EZH2, SUZ12, control CBX7 dCLIP and cellular transcriptome (input RNA-seq). All fold enrichment scores were significantly enriched ($p < 0.001$, an empirical P value using 10,000 simulations).
(f) Cis-regulatory Element Annotation System (CEAS) metagene analysis of EZH2 and SUZ12 dCLIP interactomes, compared to control CBX7 interactome. TSS, transcriptional start site. TTS, transcriptional termination site.
(g) Correlation between gene expression levels and dCLIP signal. Black, expressed RefSeq genes with reproducible dCLIP signal. Green, genes with the highest dCLIP signals consistently detected in both subunits. Red, expressed genes with no reproducible dCLIP signals.

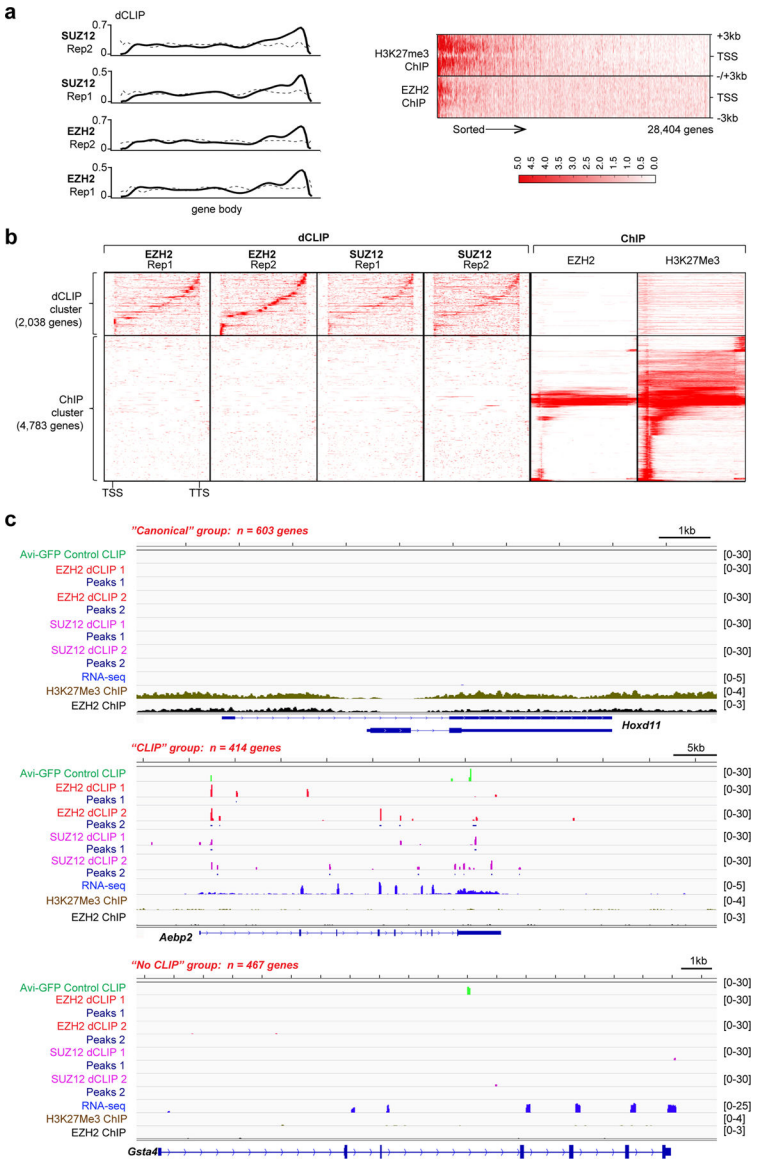


Figure 2. Integrative dCLIP-seq and ChIP-seq analyses of PRC2 subunits and H3K27me3 reveal a new class of PRC2-interacting nascent transcripts.

(a) Heatmap (bottom) showing H3K27me3 and EZH2 densities at all unique RefSeq TSSs ± 3 kb, sorted by H3K27me3 occupancy. The distribution of dCLIP signals produced by gene bodies of genes within this heatmap is indicated by a density plot (top, black line) and compared to their density over a randomly permutated heat map (top, dashed gray line). Data are from two biological replicates of each, EZH2, and SUZ12 dCLIP samples.

(b) Gene clustering based on dCLIP-seq signal of PRC2 subuntis and ChIP-seq signals of EZH2 and H3K27Me3. The “dCLIP” cluster contains genes exhibiting a highly reproducible pattern of PRC2 dCLIP-seq signal in all four dCLIP libraries. These genes are largely devoid of EZH2 and H3K27me3 ChIP-seq signals. The “ChIP” cluster contains genes with strong PRC2-ChIP (H3K27me3 and EZH2) signals. Note that dCLIP signal over these genes is considerably low or even undetectable.

(c) Representative gene classes. Top: Canonical gene, Hoxd11. Canonical genes exhibit strong EZH2 (black) and H3K27me3 (khaki) ChIP-seq signals throughout their gene bodies, with no expression (no RNA signals in the RNA-seq and dCLIP tracks). Middle: CLIP gene, *Aebp2* (middle). Note reproducible peaks in both replicates of EZH2 (red) and SUZ12 (pink) dCLIP. By contrast, there is weak or no ChIP-seq signal for EZH2 and H3K27Me3. Bottom: No CLIP gene, *Gsta4* (bottom). Despite high level expression (RNA-seq track) and absent EZH2 and H3K27me3 ChIP-seq signal, these No CLIP genes have weak to nonexistent dCLIP signal in both replicates of EZH2 and SUZ12 dCLIP.

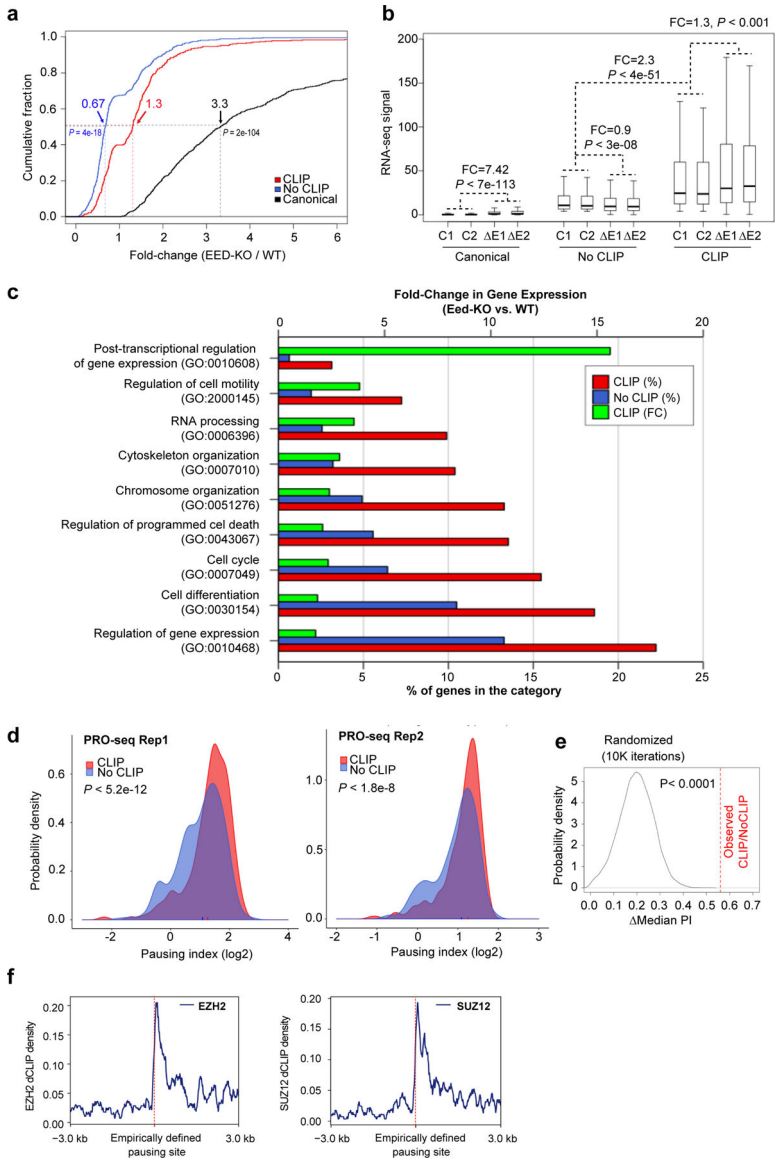


Figure 3. PRC2's interaction with nascent RNA fine-tunes gene expression states and regulates POL-II pausing.

(a) A comparison of gene expression alterations following Eed knockout (EED-KO) [30] in CLIP gene group compared to No CLIP and Canonical gene groups. We generated Cumulative Distribution Plots (CDPs) for CLIP genes (n=414), No CLIP genes (n=467), and Canonical genes (n=603), depicting fold-change in mRNA levels of genes upon EED-KO. Statistical significance of differential gene expression after Eed ablation was calculated by Wilcoxon test (unpaired, one-sided) between CLIP genes (red) versus No CLIP genes (blue), and between CLIP genes versus Canonical genes (black).

(b) Boxplots of gene expression distribution (RPKM) in CLIP (n=414), No CLIP (n=467), and Canonical (n=603) groups. Box boundaries represent 25th and 75th percentiles; center line represents the median; whiskers indicate ± 1.5× IQR). P values were determined using a one-sided non-parametric Wilcoxon test and fold change (FC) were calculated between the median values of two EED-KO samples versus the median values of two WT samples.

(c) Biological process enrichment of CLIP genes relative to No CLIP genes. Analysis was performed online (http://www.pantherdb.org/) using Fisher test and FDR correction ($p < 0.05$, FDR corrected).

(d) Probability density plots of "pausing index" (PI). PRO-seq data (n=2; Day 7 16.7 mES cells) was analyzed to compute statistical significance using Wilcoxon test (one-sided). Red and blue ticks on the X-axis depict median (log2) pausing index values for CLIP and No CLIP datasets respectively.

(e) Distribution of differences in median fold-change of pausing index values in 10,000 iterations of randomized FPKM-matched control groups (density plot; maxima <1) versus the actual median difference between CLIP and No CLIP gene groups (vertical dashed red line = 0.56).

(f) Metagene analysis of EZH2 and SUZ12 dCLIP interactomes in the vicinity of empiricially defined POL-II pausing site. Pausing sites were determined as regions of the highest density of active POL-II within the promoter-proximal region.

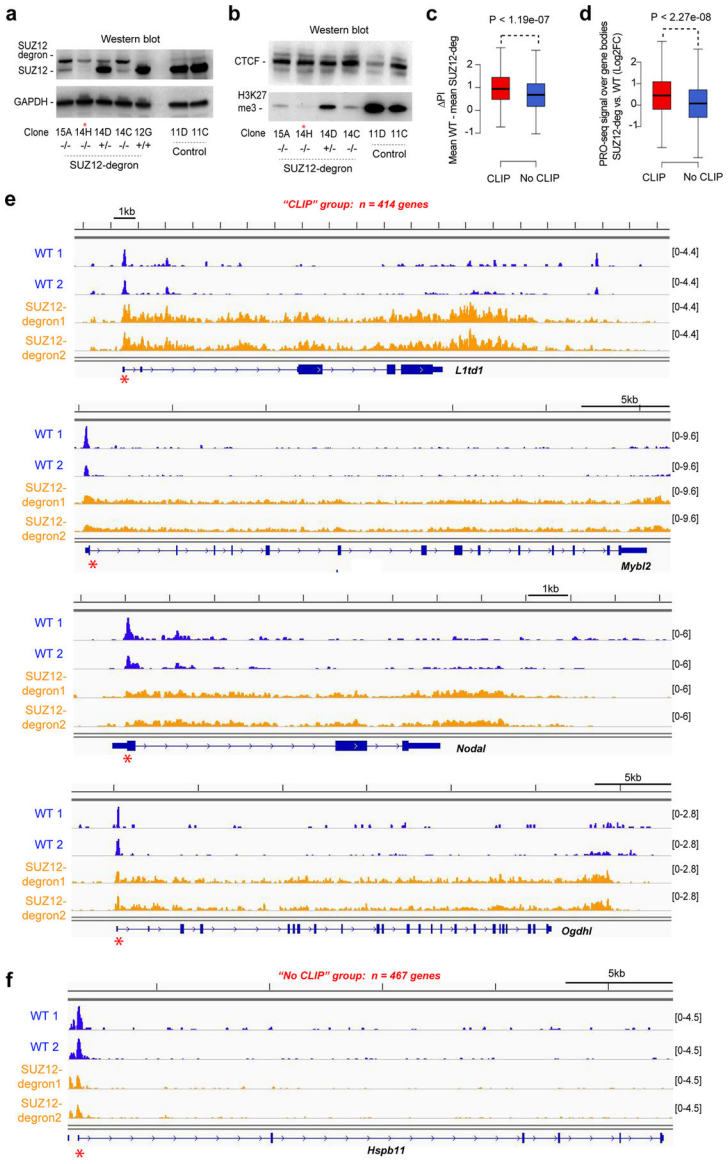


Figure 4. SUZ12 ablation abolishes POL-II pausing at promoter-proximal regions of CLIP genes.

(a) Derivation of SUZ12-degron mES cells (16.7 background). FKBP tag was introduced in the C-terminus of endogenous *Suz12* gene by CRISPR. Western Blot for SUZ12, with GAPDH protein as a loading control. Clones 14D and 14C are heterozygous clones, whereas clones 14C, 14H and 15A are homozygous. Clone 14H (asterisk) used for all subsequent studies.

(b) Western blot for H3K27me3 confirms SUZ12 ablation. CTCF, loading control. Clone 14H (asterisk) used for all subsequent studies.

(c) Boxplot showing change in pausing index (ΔPI) following SUZ12 degradation for CLIP (n=414) versus No CLIP (n=467) genes. Average pausing index (PI) scores of two PRO-seq samples from WT mES cells were substracted from the average PI of two PRO-seq replicates from SUZ12-degraded mES cells. Median shown, with 25th and 75th percentiles

delineated by boxes and whiskers indicating ± 1.5× IQR. P values were determined using a one-sided non-parametric Wilcoxon test.

(d) Boxplots showing distribution of log2 fold-change (log_2FC) in mean PRO-seq signals between SUZ12-degraded and WT cells. Average fold-change values between each cell line (two replicates per each) were calculated by performing *edgeR* differential analysis of PRO-seq signal amenanating within gene bodies (+300 bp to TTS) of CLIP and No CLIP genes. Median shown, with 25th and 75th percentiles delineated by boxes and whiskers indicating ± 1.5× IQR. P values were determined using a one-sided non-parametric Wilcoxon test.

(e) Changes in PRO-seq profile for representative CLIP genes (high ΔPI) after SUZ12 elimination. Red asterisk, pause site. Two biological replicates (1,2) shown.

(f) Changes in PRO-seq profile for a representative No CLIP gene (low ΔPI) after SUZ12 elimination. Two biological replicates (1,2) shown. Red asterisk, pause site. See Extended Data Figure 6 for more examples.

Unprocessed blots for Fig. 4 are provided as source data

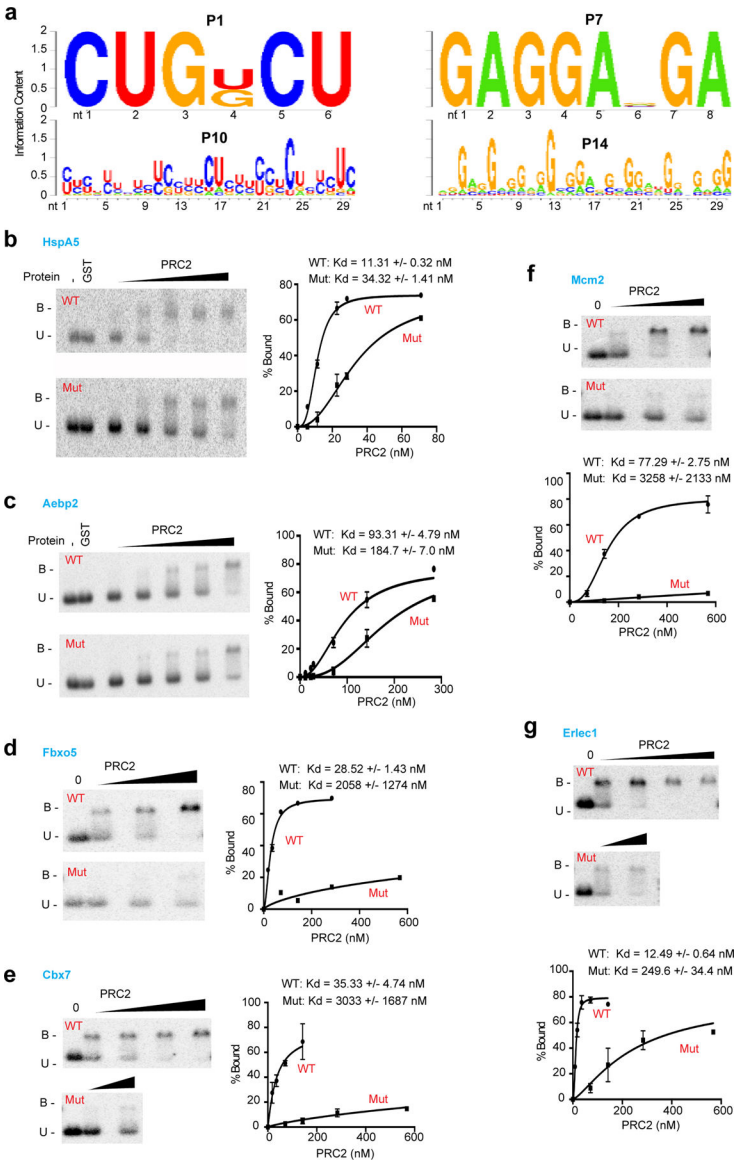


Figure 5. Identification of RNA-Binding Motifs for PRC2.

(a) RNA-Binding motifs identified by MEME / CLOVER analysis in CLIP gene group using EZH2 and SUZ12 dCLIP data.

(b,c) Representative RNA-EMSA with purified PRC2 complex and 5 nM *in vitro* transcribed RNA carrying P14 RNA-binding motif in WT or mutated (Mut) form for HspA5 (b) or Aebp2 (c). Lane 1, no protein. Lane 2, 71 nM GST. Other lanes with increasing PRC2 concentrations: 6, 11, 23, 28, 71 nM. Right: K_d quantitation of binding for WT and Mut probes. B, bound; U, unbound probe. n=2 for Aebp2, n=3 for HspA5-WT and n=4 for HspA5-Mut.

(d) RNA-EMSA with PRC2 and 5 nM Fbxo5 RNA carrying P14 RNA-binding motif in WT or Mut form. WT probes tested at 0, 17.75, 35.5, 142 nM PRC2. Mut probes tested at 0, 71, 142, 284 nM PRC2. Right: K_d quantitation, n=3 for Fbxo5-WT and n=4 for Fbxo5-Mut

(e) RNA-EMSA with PRC2 and 5 nM Cbx7 RNA carrying P14 RNA-binding motif in WT or Mut form. WT probes tested at 0, 17.75, 35.5, 71, and 142 nM PRC2. Mut probes tested at 0, 142, 284 nM PRC2. Right: K_d quantitation, n=3 for Cbx7-WT and n=4 for Cbx7-Mut.
(f) RNA-EMSA with PRC2 and 5 nM Mcm2 RNA carrying P14 RNA-binding motif in WT or Mut form. WT probes tested at 0, 35.5, 142, 284 nM PRC2. Mut probes tested at 0, 71, 142, 284 nM PRC2. Right: K_d quantitation,n=3 independent assays.
(g) RNA-EMSA with PRC2 and 5 nM Erlect1 RNA carrying P14 RNA-binding motif in WT or Mut form. WT probes tested at 0, 17.75, 35.5, 71, and 142 nM PRC2. Mut probes tested at 0, 142, 284 nM PRC2. Right: K_d quantitation,n=4 independent assays.
Unprocessed images and statisitical data for Fig.5 are provided as source data.

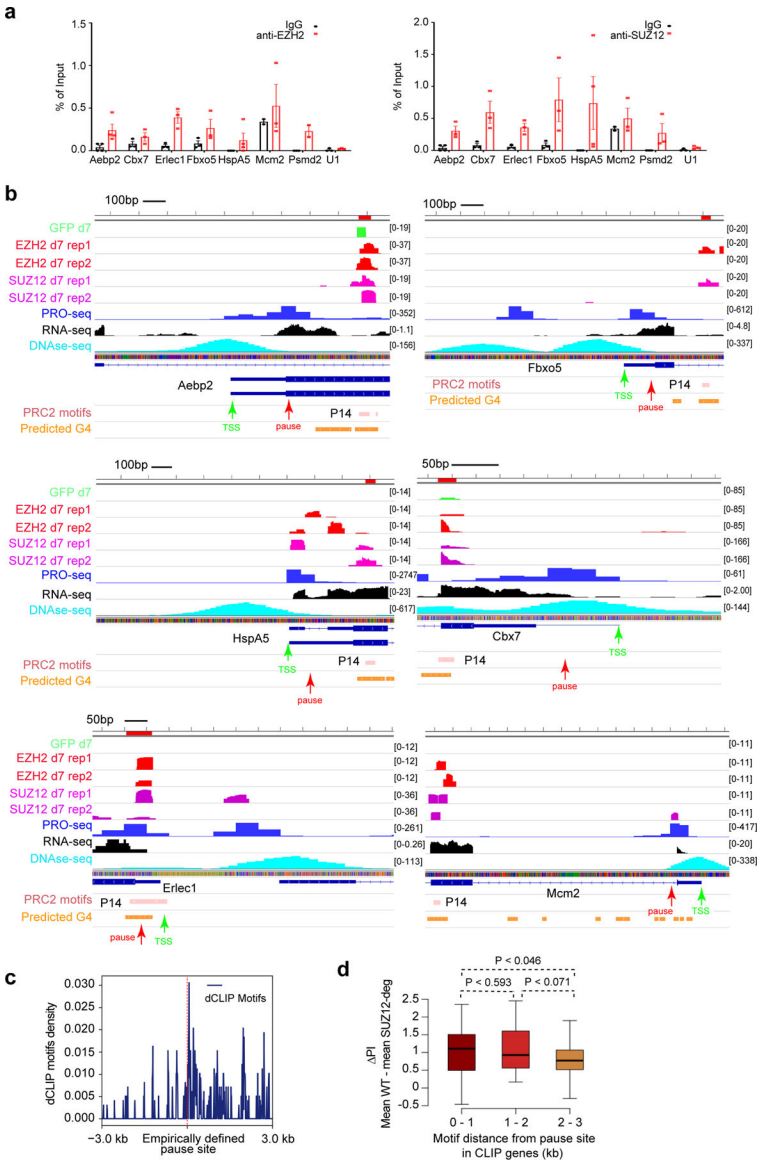


Figure 6. PRC2 motifs are linked to promoter-proximal POL-II pausing.

(a) UV-RIP qPCR validates CLIP targets as bona fide PRC2-binding nascent transcripts. Subunits EZH2 and SUZ12 were both tested, with similar overall trends. Mean +/− SD of two biological replicates with duplicates for every antibody is shown. Significance determined by unpaired student t-test. Statistical data is provided as source data.

(b) Six representative genes manifesting PRC2 dCLIP signals at their 5’ regions. Note proximity of G-rich RNA-binding motif (P14) within PRC2 dCLIP peaks. G4RNA Screener tool indicated potential rG4-forming structures (orange bars). Pausing sites were defined empirically by PRO-seq signals (summits) for active POL-II. DNAse-hypersensitivity sites were from Vierstra et al [60]. Additional examples shown in Extended Data Fig. 7.

(c) Metagene analysis of RNA-binding motifs enriched in PRC2 dCLIP interactomes in the vicinity of empiricially defined POL-II pause site. Enrichment of dCLIP motifs just downstream of pause sites.

(d) Boxplots depicting distribution of ΔPI values (Average PI score of WT cells substracted from average PI score of SUZ12-degron cells from 2 biological replicates) for CLIP genes as a function of distance from PRC2 motif to pause site. Bin #1 (0–1 kb; n=58), Bin #2 (1–2 kb; n=20), Bin #3 (2–3; n=23). Median shown, with 25th and 75th percentiles delineated by boxes and whiskers indicating ± 1.5× IQR. P values determined using one-sided Wilcoxon test.

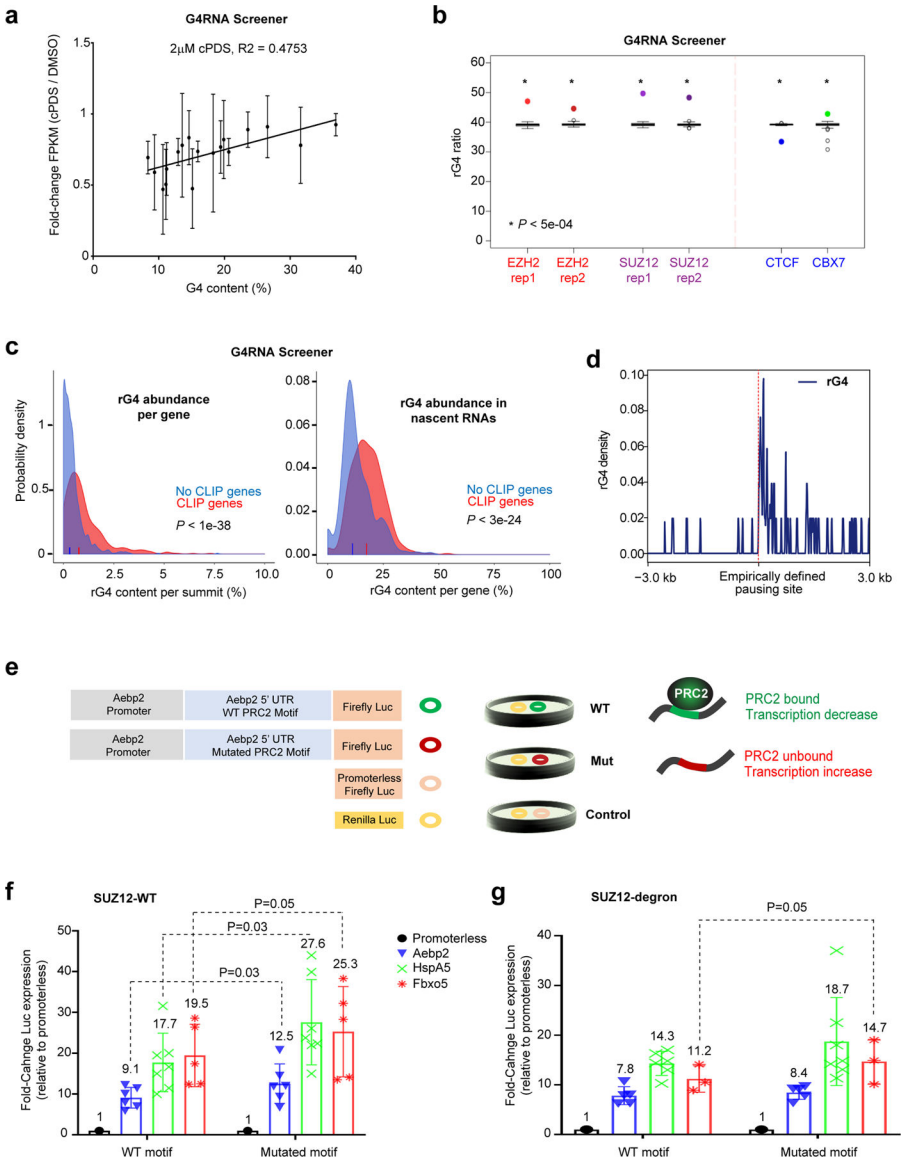


Figure 7. G-quadruplex motifs are enriched at POL-II pause sites and their ablation results in transcription upregulation.

(a) Positive correlation between RNA stability measured following treatment with rG4 stabilizer cPDS (2μM) [relative to DMSO controls] and putative rG4 gene content screened by artificial neural network algorithm, *G4RNA Screener*. RNA stability data presented is Mean +/− SD. n=4 independent assays.

(b) *G4RNA Screener* demonstrates that rG4 is highly enriched in CLIP targets relative to No CLIP transcripts. rG4 ratio defined as % of summit regions with putative rG4. rG4-ratio of PRC2 dCLIP summit-regions (100 nt around the most significant binding site ±50 nt) harboring putative rG4 element was plotted for two EZH2 and two SUZ12 replicates, with CTCF and CBX7 as controls. Black box plots: rG4-ratios from n=2,000 random sets each simulating an equal number of summit regions as in the tested library.

(c) Left: rG4 abundance per gene (“Empirical rG4 content”) is higher in CLIP versus No CLIP transcripts. rG4 elements (per *G4RNA Screener*) in CLIP summit regions and in

simulated No CLIP summit regions. Density plots of rG4 content per gene are presented for CLIP and No CLIP genes, respectively. Right: Nascent RNAs of CLIP genes have a higher potential to generate rG4 structures compared to nascent transcripts of No CLIP genes. *p* values by Wilcoxon test (unpaired, one-sided).

(d) *Deeptools* metagene analysis of rG4 elements (per PQSFinder) reveals accumulation dCLIP rG4s around POL-II pause site (x=0).

(e) rG4 motifs from *Aebp2*, *HspA5*, or *Fbxo5* were cloned into promoterless plasmid upstream of FireFly luciferase (Luc) ORF and transiently transfected into Day 0 16.7 mES cells.

(f,g). Dotplots showing RT-qPCR levels of the resulting fused transcript, using primers specific to FireFly luciferase and Renilla luciferase as reference. Fold-change in Luc expression relative to promoterless construct is plotted. Data presented is a Mean +/− SD of at least three biological replicates. Significance determined by a paired student t-test for *Aebp2* and *Fbxo5* and unpaired student t-test for *HspA5*.

Statistical data for panels a,f,gis provided as source data Fig.7.

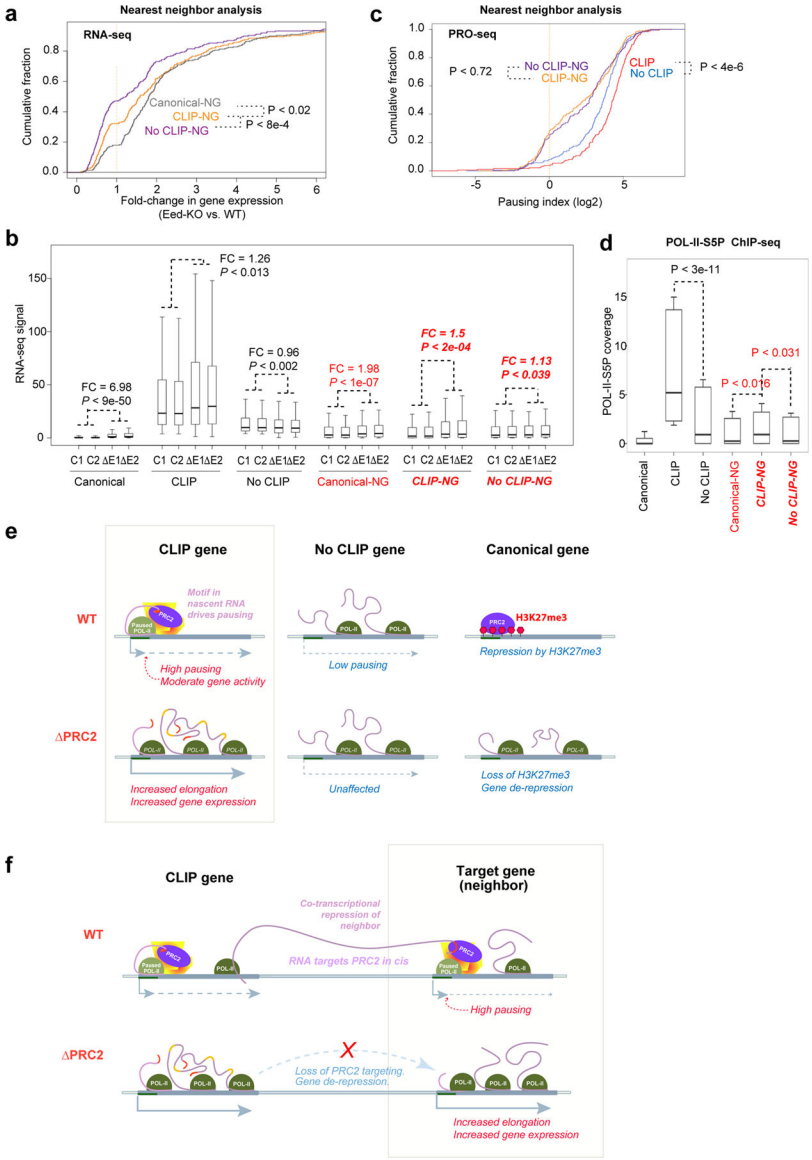


Figure 8. Unified Model: RNA recruits PRC2 for rheostat control of transcription elongation in cis and also at neighboring genes.

(a) CDPs for fold-change (FC) in gene expression following Eed knockout (EED-KO) in CLIP-NG (n=218) versus no CLIP-NG (n=238) and canonical-NG (n= n=279). Gene expression datasets were from Ardehali et al [30]. FC calculated between the median values of two EED-KO samples versus two WT samples. P values by one-sided non-parametric Wilcoxon test.

(b) Boxplots for indicated gene groups showing the distribution of RNA-seq [30] RPKM values in two WT samples (C1, C2) and two EED-KO (ΔE1, ΔE2) samples. Box boundaries, 25th and 75th percentiles. Center line, median. Whiskers ± 1.5× IQR. P values by one-sided non-parametric Wilcoxon test. Fold-change (FC) calculated between the median values of two EED-KO samples versus two WT samples.

(c) CDP of pausing index based on PRO-seq experiments, comparing CLIP versus No CLIP transcripts and CLIP-NG versus No CLIP-NG. Note significant difference for CLIP versus

No CLIP group, whereas no significant difference was see for CLIP-NG versus No CLIP-NG. P values by one-sided non-parametric Wilcoxon test.

(d) Boxplots depicting POL-II-Ser5 ChIP-seq enrichment levels (−1 kb to TTS) in the WT mES cells for CLIP-NG (n=218), No CLIP-NG (n=238) and Canonical-NG (n=279) gene groups and their corresponding CLIP (n=414), No CLIP (n=467) and Canonical (n=603) neighbors. ChIP-seq datasets for Day 7 16.7 cells were from Pinter et al [29]. Box boundaries represent 60th and 90th percentiles; center line represents the 75th percentile; whiskers indicate ± 1.5× IQR).

(e) Regulation of POL-II pausing by PRC2-nascent RNA interactions. Binding of PRC2 to nascent transcripts controls pause-release and transcription elongation by POL-II. Loss of PRC2 affects both POL-II pausing in CLIP genes, and H3K27me3 in Canonical genes.

(f) Integration of the pause-release model with the classical model in which RNA targets PRC2 to neighboring genes in cis. Loss of PRC2 leads to increased expression of two linked genes.