Title

Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation

Journal

Nature Structural & Molecular Biology, 28(2)

ISSN

1545-9993

Authors

Kubo, Naoki
Ishii, Haruhiko
Xiong, Xiong
et al.

Publication Date

2021-02-01

DOI

10.1038/s41594-020-00539-5

Peer reviewed

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 February ; 28(2): 152–161. doi:10.1038/s41594-020-00539-5.

Promoter-proximal CTCF-binding promotes long-range-enhancer dependent gene activation

Naoki Kubo[1], **Haruhiko Ishii**[1], **Xiong Xiong**[2], **Simona Bianco**[3], **Franz Meitinger**[1], **Rong Hu**[1], **James D. Hocker**[1], **Mattia Conte**[3], **David Gorkin**[4], **Miao Yu**[1], **Bin Li**[1], **Jesse R. Dixon**[5], **Ming Hu**[6], **Mario Nicodemi**[3], **Huimin Zhao**[2,7], **Bing Ren**[1,4,8,*]

[1]Department of Cellular and Molecular Medicine, University of California San Diego School of Medicine, La Jolla, CA, USA

[2]Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[3]Department of Physics, University of Naples Federico II, and INFN Complesso di Monte Sant'Angelo, Naples, Italy

[4]Department of Cellular and Molecular Medicine, Center for Epigenomics, University of California San Diego School of Medicine, La Jolla, CA, USA

[5]Salk Institute for Biological Studies, La Jolla, CA, USA

[6]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH, USA.

[7]Departments of Chemistry, Biochemistry, and Bioengineering, and Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[8]Department of Cellular and Molecular Medicine, Moores Cancer Center and Institute of Genome Medicine, University of California San Diego School of Medicine, La Jolla, CA, USA

*Correspondence to: biren@ucsd.edu.

Author contributions:

N.K., H.I., and B.R. conceived the project. N.K., H.I., X.X., and H.Z. engineered cell lines. N.K., R.H., J.D.H., and Z.Y. carried out library preparation. N.K. and F.M. performed cell cycle analysis. N.K., S.B., M.C., M.N., D.G., and B.L. performed data analysis. M.Y., M.H., and J.D. contributed to computational analysis and experimental design. N.K. and B.R. wrote the manuscript. All authors edited the manuscript.

Competing interests:

B.R. is a co-founder of Arima Genomics, Inc. and Epigenome Technologies, Inc..

Reporting Summary statement:

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

Code Availability Statement:

PLAC-seq and the other analyses in this study were performed by combining public software as described in Methods.

Data Availability statement:

All datasets generated in this study have been deposited to Gene Expression Omnibus (GEO), with accession number GSE94452. Hi-C dataset analyzed in Extended Data Fig. 3g–i was provided from Dr. Benoit Bruneau (GSE98671). Accession codes for the mouse tissue datasets used in Fig. 5 and Extended Data Fig. 10 are listed in Supplementary Table 6; data sets are available from the ENCODE portal (https://www.encodeproject.org/).

Source data are available with the paper online.

Abstract

The CCCTC-binding factor (CTCF) works together with the cohesin complex to drives the formation of chromatin loops and topologically associating domains, but its role in gene regulation has not been fully defined. Here, we investigated the effects of acute CTCF loss on chromatin architecture and transcriptional programs in mouse embryonic stem cells undergoing differentiation to neural precursor cells. We identified CTCF-dependent enhancer-promoter contacts genome-wide and found that they disproportionally affect genes that are bound by CTCF at promoter and dependent on long-distance enhancers. Disruption of promoter-proximal CTCF binding reduced both long-range enhancer-promoter contacts and transcription, which are restored by artificial tethering of CTCF to the promoter. Promoter-proximal CTCF binding is correlated with transcription of over 2,000 genes, across a diverse set of adult tissues. Taken together, our study shows that CTCF binding to promoters may promote long-distance-enhancer dependent transcription at specific genes in diverse cell types.

Introduction:

Transcriptional regulation in mammalian cells is orchestrated by cis-regulatory elements that include promoters, enhancers, insulators and other less well characterized sequences[1,2]. Large-scale projects such as ENCODE have annotated millions of candidate cis-regulatory elements in the human genome and genomes of other mammalian species[3–5]. A majority of these candidate regulatory elements are located far from transcription start sites (i.e. promoters), display tissue and cell-type specific chromatin accessibility, and likely act as enhancers to regulate cell-type specific gene expression. Enhancers can activate genes at great genomic distances, making it difficult to predict their target genes from sequence information alone. Increasingly, maps of the chromatin topology are used to infer target genes of enhancers, based on the observations that enhancers are frequently positioned close to their target gene promoters in 3D space at the time of gene activation[6]. However, the exact role of chromatin topology in enhancer-dependent gene regulation remains to be clearly defined[7,8].

The CCCTC-binding factor (CTCF) plays a critical role in chromatin architecture[9–13]. It works together with the cohesin complex to establish chromatin domains genome-wide, and forms long-range chromatin loops between CTCF binding sites (CBSs) via a mechanism involving loop extrusion [14,15]. CTCF has also been shown to be necessary for enhancer-promoter (E-P) contacts for specific genes, such as the proto-cadherin gene clusters, and for class switch recombination in B lymphocytes[16–18]. On the other hand, acute depletion of CTCF has been shown to result in only moderate change of gene expression profiles despite the global loss of chromatin loops anchored at CBSs and weakening of chromatin domain boundaries[13,19]. In addition, although CTCF is essential for embryonic development in multiple types of tissues[20], a recent study reported a dispensable role for CTCF in immune cell differentiation[21]. To better understand the apparent discrepancy in the functional role of CTCF in dynamic gene regulation in different cell types, comprehensive analysis of CTCF-dependent E-P contacts during cell differentiation and exploration of the role of CTCF binding in establishment of specific E-P contacts are needed.

Here we use two different approaches to perturb chromatin topology at CBSs in mouse embryonic stem cells (mESCs), in order to define the role of CTCF-driven chromatin organization in gene regulation and cellular differentiation. First, we used auxin-inducible degron[22,23] to acutely deplete CTCF protein levels in a genetically engineered mouse ES cell line, and study the changes in chromatin topology genome-wide in both undifferentiated ES cells and in neural precursor cells (NPCs) derived from the CTCF-depleted ES cells. To identify promoter-anchored contacts at high-resolution that cannot be precisely revealed by conventional *in situ* Hi-C and might be in the similar resolution level as Micro-C[24], we also performed promoter-centric chromatin conformation capture assays, PLAC-seq (also known as HiChIP)[25,26]. We observed hundreds of lost and newly formed E-P and promoter-promoter (P-P) contacts at dysregulated genes, and found that removal of CTCF binding at the promoter reduces E-P and P-P contacts and gene expression, suggesting that CTCF binding at promoters plays an active role in establishment of promoter-anchored contacts. In the second approach, we used CRISPR technology to artificially tether CTCF to a promoter. We demonstrated that targeted recruitment of CTCF to a promoter is required to establish long-range chromatin contacts between the promoter and distal elements and to activate gene expression. Furthermore, we characterized the features of CTCF-dependent genes and found that the impact of CTCF loss on gene regulation is determined not only by CTCF binding at promoters but also the distribution of nearby enhancers. The role of promoter-proximal CTCF binding in transcriptional regulation is further supported by the observation that over 2,300 mouse genes display a significant correlation between CTCF occupancy at the promoter and tissue-specific gene expression patterns. Our findings therefore uncover the mechanisms of CTCF-dependency in gene regulation and provide direct evidence for a role of CTCF-binding at promoters in activation of genes located in regions sparse in active enhancers, which is distinct from its function at insulator sequences.

Results:

CTCF depletion impedes differentiation of mESC towards neural precursor cells

To investigate the functional role of CTCF-driven chromatin loops in gene regulation, we utilized an auxin-inducible degron system to acutely deplete CTCF protein in mESC and examined the impact of CTCF loss on gene expression and chromatin architecture during mESC differentiation to NPCs (Fig. 1a, Supplementary Table 1). The depletion of CTCF was verified by Western blotting and by ChIP-seq analysis showing that chromatin occupancy of CTCF was nearly completely lost in both ESCs and NPCs, along with loss of cohesin accumulation (Extended Data Fig. 1 and Supplementary Table 2). The CTCF-depleted cells exhibited a delay in the formation of neuronal axons during neural differentiation treatment with cell colonies remaining in ESC-like round-shape and reverted to normal NPC morphology after washing out of auxin (Fig. 1b).

We next investigated the impact of CTCF loss on gene expression using RNA-seq. Consistent with previous reports[13], transcription from most genes was unaffected in CTCF-depleted ESC. Additionally, the gene expression profiles were largely uninterrupted during cell differentiation (Fig. 1c). However, transcription of several hundreds of genes (186 genes, 1.4% in ESCs, 353 genes, 2.7% in NPCs) decreased significantly upon CTCF loss

(FDR < 0.05, fold change > 2) (Fig. 1d, Extended Data Fig. 2a, b, and Supplementary Table 3). Genes related to neural differentiation (e.g. *Pcdh* cluster genes, *Pax6*, *Tubb3*) were highly enriched in these CTCF-dependent genes (Extended Data Fig. 2c), consistent with the observation that CTCF-depleted ESCs underwent abnormal neural differentiation as described above.

CTCF loss affects hundreds of E-P and P-P contacts at dysregulated genes in mESC and NPC

Previous studies have shown that CTCF plays an important role in establishment of chromatin loops and domains in the mammalian genome[9–13,27]. To better understand how CTCF-dependent chromatin topology contributes to gene regulation, we first defined the changes of chromatin architecture as a result of CTCF depletion. To this end, we performed in situ Hi-C experiments with ESC and NPC, each before and after auxin-induced depletion of CTCF. Topologically associating domains (TADs) are DNA segments characterized by strong intra-domain interactions and relatively weak inter-domain interactions in Hi-C, and the strength of the TAD boundaries can be quantified by the insulation score, a ratio between the number of cross-border interactions and the sum of intra-domain interactions within the two adjacent TADs[28]. As shown in Extended Data Fig. 3, CTCF depletion resulted in substantial loss of chromatin contacts between convergent CBSs (genomic distance > 100 kb), supporting an essential role for CTCF in the formation of these chromatin organizational features (Extended Data Fig. 3a). We also observed significant weakening of TAD boundaries in both ESC and NPC (Extended Data Fig. 3b–e and Supplementary Table 4). These results are consistent with previous findings indicating CTCF's role in the formation of most, if not all, TADs in mammalian cells[13] (Extended Data Fig. 3g–i). On the other hand, we also observed relatively well-preserved TAD boundaries in CTCF-depleted cells. As transcription has been suggested to play a role in boundary formation[9,29], insulation of TAD boundaries was generally stronger when they overlapped with housekeeping genes. Consistent with the previous hypothesis, insulation scores at these TAD boundaries remained low even after CTCF depletion, indicative strong insulation. This observation suggests that CTCF is not the only factor to modulate domain insulation at TAD-boundaries with highly activated transcription (Extended Data Fig. 3d).

To more precisely define the changes in chromatin topology due to CTCF loss, we performed PLAC-seq (also known as HiChIP)[25,26], which interrogates chromatin contacts from active or poised gene promoters at high resolution by combining Hi-C and chromatin immunoprecipitation. We used antibodies against the histone modification H3K4me3, which marks active or poised promoters, to detect chromatin contacts centered on these genomic regions. We obtained between 300 and 400 million paired-end reads for each replicate (Supplementary Table 1). To determine the differential chromatin contacts in ESCs and NPCs, we analyzed 11,900 gene promoters with similar levels of H3K4me3 ChIP-seq signal using a negative binomial model for each distance-stratified 10-kb interval (Supplementary Figure 1a, b, Methods). In total, we found 5,913 chromatin contacts between the promoters of 4,573 genes and distal elements (active enhancers or promoters) to be significantly induced during the neural differentiation (FDR < 0.05), and 1,594 contacts centered on 1,294 genes significantly decreased, most of which could not be identified in deeply

sequenced Hi-C data[29] (Fig. 2a, Supplementary Figure 1c, d). We observed a higher number and longer-range of E-P and P-P contacts induced during neural differentiation than in mESC (Fig. 2a, Supplementary Figure 1d, e, and Extended Data Fig. 4a). As expected, these dynamic changes of E-P and P-P contacts were positively correlated with the changes of active histone modifications such as H3K27ac and H3K4me1 (Extended Data Fig. 4b). Our datasets also confirmed previously reported dynamic E-P contacts during ESC differentiation (e.g. *Sox2*, *Dnmt3b*)[10,30] (Extended Data Fig. 4c). Analysis of the relationship between dynamic chromatin contacts at promoters and transcription is complicated by the fact that many genes had multiple E-P contacts, and the changes of each individual chromatin contact were not always positively correlated with differential gene expression (Extended Data Fig. 4c–e). We therefore devised an Active-Inactive-Contact (AIC) value, in which we used solely contact count by combining multiple E-P and P-P contacts (Extended Data Fig. 4f, Methods). This value showed a positive correlation with gene expression changes (Extended Data Fig. 4g, h). Interestingly, we also observed a large number of promoter-anchored contacts even at inactive genes (Extended Data Fig. 4i) and identified over a thousand chromatin contacts with distal elements displaying repressive histone mark H3K27me3 (Extended Data Fig. 4j, k). The above results, taken together, support the potential of our datasets for analyzing individual E-P and P-P contacts[29].

Using the same approach, we determined the chromatin contacts dependent on CTCF in ESCs and NPCs by comparing the PLAC-seq data collected from cells before and after CTCF depletion. The chromatin contacts between convergent CBSs were severely reduced upon CTCF loss (Extended Data Fig. 5a, b), consistent with the results from Hi-C assays. Surprisingly, the majority of chromatin contacts between enhancers and promoters remained unchanged despite the global weakening of TADs and disruption of chromatin loops. Chromatin contacts between just 394 and 806 enhancer-promoter (E-P) or promoter-promoter (P-P) pairs in ESCs and NPCs, respectively, decreased significantly upon CTCF loss (FDR < 0.05), while chromatin contacts between 44 and 109 pairs in ESCs and NPCs, respectively, increased upon CTCF loss (Fig. 2b, Supplementary Table 5, Supplementary Figure 1c). Consistent with potential for some promoters to act as enhancers of other genes[31,32] (Extended Data Fig. 4h), disruption of P-P contacts upon CTCF loss was accompanied by down-regulation of gene expression (Extended Data Fig. 5d, e), although the expression of the genes in each pair did not always move in the same direction upon loss of contacts (Extended Data Fig. 5e). Regarding the genomic distance of chromatin contacts, CTCF-dependent E-P and P-P contacts in differentiated NPCs generally span longer genomic distances than those in undifferentiated ESCs (Extended Data Fig. 5c), which is consistent with the increase in long-range promoter-anchored contacts in differentiated NPCs (Fig. 2a). Most importantly, only 283 pairs of E-P and P-P contacts, out of 5,913 contacts that are normally induced during differentiation, failed to be induced in the absence of CTCF (Fig. 2c). This observation provides an explanation for the mild change of gene expression profiles upon CTCF depletion, and suggests CTCF/cohesin-independent mechanisms of enhancer-promoter contacts, as recently reported[33].

CTCF binding at gene promoters drives CTCF-dependent E-P and P-P contacts

We next investigated the features of CTCF-dependent E-P and P-P contacts and genes. Since ChIP-seq levels of histone modifications (H3K27ac, H3K4me1, and H3K4me3) were virtually unaffected by CTCF depletion, the observed changes in chromatin contacts in ESCs and NPCs were likely a direct consequence of CTCF loss (Extended Data Fig. 6). We first categorized CTCF-dependent reduced chromatin contacts based on the presence of CBSs on anchor sites. The majority of CTCF-dependent E-P and P-P contacts were anchored by CBSs at promoter regions (81% in ESC, 64% in NPC), although they were not always anchored by convergent CTCF on their both anchor sites (Fig. 2d, Extended Data Fig. 7a). Importantly, most of these CTCF-dependent E-P and P-P contacts were located within CTCF-CTCF loops or at the loop anchors, suggesting that CTCF-dependent E-P and P-P contacts might be physically supported by surrounding CTCF-CTCF loops. In line with these results, CTCF-dependent down-regulated genes were strongly enriched for CBSs at their promoters, not at their distal elements (Fig. 2e, Extended Data Fig. 7c–f). Furthermore, the degree of enrichment of CTCF-dependent E-P and P-P contacts increased with the number of CBSs around the anchors (Extended Data Fig. 7b) and the larger number of CBSs was also observed around promoters of CTCF-dependent genes (Fig. 2f). By contrast, the anchors of E-P and P-P contacts induced upon CTCF loss were not enriched for CBS (Extended Data Fig. 7a, b), and a higher fraction of induced chromatin contacts were crossing over multiple CTCF sites such as TAD boundary regions, suggesting that CTCF depletion leads to newly formed chromatin contacts between enhancers and promoters that were formerly insulated by CTCF binding (Extended Data Fig. 7g–i). Taken together, CTCF dependency in gene regulation was highly impacted by enriched CBSs around promoters, indicating that CTCF promotes long-range E-P and P-P contacts by binding directly to their promoters and controls transcription of a select number of genes.

Artificial tethering of CTCF to a gene promoter facilitates distal element dependent transcription

To delineate the causal relationship between CTCF-mediated long-range chromatin contacts anchored at a promoter and distal element dependent transcription, we next focused on the *Vcan* gene, which encodes a protein that plays an important role in axonal outgrowth[34] and neural differentiation[35]. *Vcan* is induced during NPC differentiation, and the induction is lost upon CTCF depletion along with the loss of a long-range P-P contacts (350 kb range) anchored by a CBS only on the *Vcan* promoter side, suggesting that the *Vcan* gene may be regulated by a CTCF-dependent P-P contact. We used CRISPR-mediated genome editing to delete a 118-bp sequence containing the CTCF binding motif at the promoter of *Vcan* gene. Upon removal of the CTCF binding sequence, *Vcan* expression was significantly reduced. This reduction in *Vcan* expression could be restored partially by tethering the CTCF protein to the mutated *Vcan* promoter using a dCas9-CTCF fusion and a guide RNA (gRNA) targeting a sequence adjacent to the deleted CTCF binding sequence, in two different experiments using distinct gRNAs (Fig. 3a, b, Extended Data Fig. 8a–c). The rescue of the *Vcan* expression by the artificially tethered CTCF was also dependent on the presence of the promoter of *Xrcc4*/*Tmem167* gene located at 350 kb downstream of *Vcan* gene (Fig. 3a, b, the cell line depicted on second from the bottom). PLAC-seq experiments showed that the *Vcan* promoter-anchored contacts within the TAD were significantly reduced upon removal

of the promoter-proximal CTCF binding site (WT vs CTCF motif del in Fig. 3c, d), though the degree of the change was not as severe as those observed after the global loss of CTCF (auxin - vs auxin + in Fig. 3c, d). Similarly, the artificial tethering of CTCF to *Vcan* promoter restored the intra-TAD contacts from *Vcan* promoter (dCas9-CTCF vs dCas9 alone in Fig. 3c, d), and importantly, the long-range chromatin contacts between *Vcan* promoter and the 350 kb downstream distal *Xrcc4*/*Tmem167* promoter was also re-established (Fig. 3e, f, Extended Data Fig. 8d).

Taken together, our results demonstrated that promoter-proximal CTCF binding can promote long-range promoter-anchored chromatin contacts and facilitates gene activation driven by distal enhancers. Polymer modelling based on the strings and binders switch (SBS) model also supports such changes of chromatin contacts upon CTCF depletion[36,37], providing at the same time information of the underlying 3D spatial organization of chromatin around the *Vcan* gene at the single-molecule level (Extended Data Fig. 8e–g). Finally, it is interesting to note that the distal element driving *Vcan* transcription was also itself a promoter of gene, but we confirmed that this distal promoter is responsible for the activation of *Vcan* gene transcription by deleting its sequences (Fig. 3a, b, the cell line depicted on the bottom).

Promoter-proximal CTCF-regulated genes tend to reside in enhancer sparse regions

While we found that the genes affected by CTCF loss tend to be occupied by CTCF at the promoters, we next addressed the distribution of active enhancers around CTCF-dependent/-independent genes in ESC and NPC. CTCF-independent genes (Fig. 1d) were generally close to active enhancers especially in NPCs (Extended Data Fig. 9a, b) and appeared to be regulated by short range interactions (≤ 50 kb, PLAC-seq peak signal p-value < 0.01) (Fig. 4a), implying that they are regulated by short-range E-P and P-P contacts formed independently of CTCF (Extended Data Fig. 9c, d). By contrast, CTCF-dependent genes were generally regulated by long-range E-P and P-P contacts (≥ 100 kb, PLAC-seq peak signal p-value < 0.01) especially in NPCs (Fig. 4b, Extended Data Fig. 9e). Similarly, genes up-regulated upon CTCF depletion differed from those down-regulated in whether they had multiple active enhancers around them or not. While the down-regulated genes tended to be located at enhancer sparse regions defined as having two enhancers or less around transcription start site (TSS) within 200 kb (Fig. 4b), the up-regulated genes were close to multiple enhancers (Fig. 4b, Extended Data Fig. 9f, and Supplementary Figure 2 for their examples). These results suggest that CTCF-dependent genes are generally dependent on distal elements. To support this, we further examined the overlap with genes whose regulation is dependent on Mll3 and Mll4, major H3K4 monomethyltransferases on distal enhancers[38,39]. As expected, the overlap between Mll3/4 dependent genes that were differentially expressed in Mll3/4 double knockout cells[40] and the CTCF-dependent genes in NPCs were highly significant (p-value = 3.6e-42, odds ratio = 3.8) (Extended Data Fig. 9g). Lastly, we addressed the variable impacts of CTCF loss at gene promoters, because there were still hundreds of genes that were not affected by CTCF loss despite having CTCF binding at gene promoters in ESCs and NPCs. (Fig. 2e). We classified these genes occupied by CTCF at promoters (TSS ±1 kb) into two groups based on the number of enhancers around the gene and the distance of their E-P contacts (Fig. 4c, left panel). When genes were located at enhancer dense regions (7 enhancers or more around TSS ± 200 kb) and displayed

short chromatin contacts with distal elements (< 50 kb, PLAC-seq peak signal p-value < 0.01), the reduction in gene expression upon CTCF loss was moderate despite CTCF binding at their promoters. Taken together, CTCF dependency in gene regulation was determined not only by the distribution of CBSs but also the distribution of active elements around promoters, and importantly, these results suggest that presence of multiple active enhancers nearby genes might compensate for the loss of CTCF at promoters.

Promoter occupancy by CTCF correlates with tissue-restricted gene expression at over 2,000 of genes

The above findings suggest a previously under-appreciated mechanism for CTCF in gene regulation. In contrast to its well-established role in forming chromatin loops, TAD boundaries and insulators, we demonstrated that CTCF also directly binds to gene promoters to promote long-range E-P(P-P) contacts and enable enhancer-dependent gene expression in mammalian cells. In mouse ESCs and NPCs, several hundred genes are subject to regulation by this mechanism. These include the proto-cadherin gene clusters that were previously reported to be regulated by CTCF binding at the promoters and the distal enhancer[16] (Supplementary Figure 2a). To further explore the extent of genes subject to this CTCF-dependent mechanism, we examined public ChIP-seq datasets of CTCF binding and RNA-Seq across multiple mouse tissues (9 tissue samples from ENCODE[4,41], Supplementary Table 6). Consistent with this postulated mechanism, multiple CBSs are enriched around promoters with relatively conserved binding motif sequences (Fig. 5a, Extended Data Fig. 10a–c) and ChIP-seq signals around promoters (TSS ±10 kb) show positive correlation with gene expression in over 2,300 mouse genes in these tissues ($r > 0.6$, 2,332 genes), many of which could not be explained by DNA methylation levels at the promoter-proximal CBSs (Fig. 5b, Extended Data Fig. 10d). Interestingly, high lineage-specificity in transcription as measured by Shannon entropy that define transcriptome diversity based on its distribution[42] was predominantly found in the forebrain-specific genes and the most enriched gene ontology (GO) term in this gene group was related to “synapse assembly”. On the other hand, GO terms related to “signaling pathway” were enriched in the other tissue-specific genes (Fig. 5c, Extended Data Fig. 10e, f). Many forebrain-specific and heart-specific genes were down-regulated in CTCF-depleted NPCs and CTCF knockout heart tissue[43], respectively (Extended Data Fig. 10g, h), supporting that many of these genes are indeed regulated by CTCF binding to the gene promoters in a lineage-specific manner.

Discussion:

CTCF- and cohesin-mediated chromatin structures such as TADs and CTCF loops have been postulated to play a role in constraining enhancer-promoter communications[9–12,27]. However, the vast majority of genes are expressed normally in the absence of CTCF or Cohesin[13,19], raising questions about the role of chromatin architecture, especially E-P contacts, in gene regulation. Here, we unveiled the mechanisms of CTCF-dependent/-independent gene regulation in different cell identities, and provided multiple layers of evidence that CTCF not only actively forms TADs and CTCF loops, but also directly promotes E-P and P-P contacts and potentially contributes to activation of thousands of lineage-specific genes. CTCF binding to the promoter of such genes is necessary for

establishing their E–P and P-P contacts, and we demonstrated that artificial tethering of CTCF to gene promoter could promote P-P contacts and gene activation. The active function of promoter-proximal CTCF in promoting long-range E-P(P-P) contacts may affect different subsets of genes in different cell types due to the variable distribution of enhancers and CBSs. Supporting this, CTCF loss leads to variable phenotypes in different cell types. For example, CTCF knockout leads to severe developmental defects in many tissues[20] and our study shows abnormal cell differentiation from ESCs to NPCs in the absence of CTCF, while CTCF is dispensable for immune cell transdifferentiation with minor effects on transcription that is comparable to our findings[21]. It is also reasonable to assume that some specific long-range E-P or P-P contacts that potentially determine lineage-specificity require the strong structure of long-range CTCF loops including TAD structures[44]. In other words, CTCF-mediated structures might help promoter-proximal CTCF to reel distal enhancers to the target gene promoter, possibly via loop extrusion, predominantly from one side, as we observed significant change of chromatin contacts frequency inside the TAD in the rescue experiments at the *Vcan* gene. Therefore, the orientation of CTCF binding motif at promoters may be important to determine the predominant direction for searching distal enhancers (Extended Data Fig. 10c). In our dCas9-CTCF tethering experiments, we designed gRNAs to target top and bottom strands separately, but both gRNAs could restore the chromatin loops and gene expression. It might due to presumably flexible orientation of dCas9-CTCF protein that does not bind to DNA directly. Detailed analysis of CTCF protein structure and its binding orientation will illuminate this mechanism.

Our study also revealed that the majority of individual E-P and P-P contacts are not affected by CTCF loss, which could explain the modest change of gene expression profiles upon CTCF loss. While CTCF occupancy might not contribute to forming of the majority of E-P contact[45], their dynamics during cell differentiation were associated with the enhancer activities[29]. In addition to the enhancer activity itself, it can be assumed that these CTCF-independent E-P contacts are mediated by other factors. In NPCs, neuronal transcription factors such as Pax6 contribute to chromatin folding, which is compatible with highly variable changes of E-P contacts upon neural differentiation[29]. Besides these lineage-specific expressed factors, there are several chromatin folding proteins that are more broadly expressed across tissues. For example, Yin Yang 1 (YY1) has been shown to regulate E-P contacts in mouse embryonic stem cells[46,47]. Another genomic interaction mediator, LIM-domain-binding protein 1 (LDB1) is also known to control long-range and trans interactions[48,49] that regulate specific gene sets such as olfactory receptor genes[49] and genes for cardiogenesis[48]. Besides E-P contacts, P-P contacts are also considered as a key factor for gene regulation network because many promoters may have enhancer-like activity[31,32] and promote cooperative regulation between genes[31,50]. However, the chromatin folding factors between promoters and how CTCF binding affects such P-P contacts and gene regulation have not been elucidated. Our study identified hundreds of CTCF-dependent P-P contacts in ESCs and NPCs, and demonstrated that CTCF binding at promoter can promotes E-P and P-P contacts and activates gene expression. The synchronous activity between interacting promoters is also broadly consistent with the phase separation model[51], and further work should elucidate the relationship between CTCF-mediated genome structure and phase separation transcription machinery. It should be noted that PLAC-seq datasets in

this study have H3K4m3 antibody bias between E-P and P-P contacts, therefore it cannot determine which type of contacts is predominantly affected by CTCF loss.

Lastly, we showed a link between CTCF binding and lineage-specific gene expression, yet the factor(s) regulating tissue-specific CTCF binding at promoters remains unclear. DNA methylation is the most well-studied determinants of CTCF binding[52,53]. CTCF occupancy is inhibited by DNA methylation and tissue-specific DNA methylation dynamics might affect the tissue-specific CTCF binding at promoters. For example, as reported in a recent study[47] and our datasets, global loss of CTCF binding occurred during differentiation from ESCs to NPCs (Extended Data Fig. 1d), and these observations are consistent with the reports that pluripotency in ESCs is associated with global DNA demethylation[54,55]. On the other hand, a recent study showed that the vast majority of CTCF binding was not affected upon a drastic reduction of DNA methylation by double knockout of the major DNA methyltransferases DNMT1 and DNMT3B[56]. Our analysis also showed lack of correlation between the majority of tissue-specific promoter-proximal CTCF binding and DNA methylation. It is possible that 5-carboxylcytosine[57] or cooperative binding with undetermined factors play a role in tissue-specific CTCF binding.

Our study only surveys the impact of CTCF loss after a certain period of CTCF depletion and some genes might be affected by secondary effects of the abnormal differentiation. Nevertheless, this study uncovers the genome-wide alteration of E-P and P-P contacts upon CTCF loss during neural differentiation and provides new insight into the functional role of CTCF in gene regulation. CTCF has been implicated in a variety of human diseases. It has been previously reported that CBSs are highly mutated in several cancer types[58,59] and somatic CTCF mutations also occur in about one-quarter of endometrial carcinoma[60]. Thus, further study of the mechanism of CTCF-mediated gene regulation in various cell types could help to elucidate the causes of these diseases.

Methods:

Cell lines

The F1 Mus musculus castaneus × S129/SvJae mouse ES cells (XY; F123 cells)[61] (a gift from Rudolf Jaenisch) were cultured in KnockOut Serum Replacement containing mouse ES cell media: DMEM 85%, 15% KnockOut Serum Replacement (Gibco), penicillin/ streptomycin (Gibco), 1× non-essential amino acids (Gibco), 1× GlutaMax (Gibco), 1000 U/ml LIF (Millipore), 0.4 mM β-mercaptoethanol. The cells were typically grown on 0.2% gelatin-coated plates with irradiated mouse embryonic fibroblasts (MEFs) (GlobalStem). Cells were maintained by passaging using Accutase (Innovative Cell Technologies) on 0.2% gelatin-coated dishes (GENTAUR) at 37°C and 5% CO2. Medium was changed daily when cells were not passaged. Cells were checked for mycoplasma infection and tested negative.

Construction of the plasmids

The CRISPR/Cas9 plasmid (CTCF-mouse-3sgRNA-CRISPRexp-AID) was assembled using the Multiplex CRISPR/Cas9 Assembly System kit (a gift from Takashi Yamamoto, Addgene kit #1000000055). Oligonucleotides for three gRNA templates were synthesized, annealed

and introduced into the corresponding intermediate vectors. The first gRNA matches the genome sequence 23 bp upstream of the stop codon of mouse CTCF. The oligonucleotides with sequences (5'-CACCGTGATCCTCAGCATGATGGAC-3') and (5'-AAACGTCCATCATGCTGAGGATCAC-3') were annealed. The other two gRNAs direct *in vivo* linearization of the donor vector: the first pair of oligonucleotides are (5'-CACCGCTGAGGATCATCTCAGGGGC −3') and (5'-AAACGCCCCTGAGATGATCCTCAGC −3'); the second pair is (5'-CACCGATGCTGGGGCCTTGCTGGC-3') and (5'-AAACGCCAGCAAGGCCCCAGCATC-3'). The three gRNA-expressing cassettes were incorporated into one single plasmid using Golden Gate assembly. The donor vector (mCTCF24-AID-donor-Neo) was constructed using PCR and Gibson Assembly Cloning kit (New England Biolabs). The insert cassette includes sequences that codes for a 5GA linker, the auxin-induced degron (AID), a T2A peptide and the neomycin resistant marker, and is flanked by 24-bp homology arms to integrate into the CTCF locus. The left and right arms have sequences CCTGAGATGATCCTCAGCATGATG and GACCGGTGATGCTGGGGCCTTGCT, respectively. The AID coding sequence was amplified from pcDNA5-H2B-AID-EYFP (a gift from Don Cleveland, Addgene plasmid #47329) and the T2A-Neo^R was amplified from pAC95-pmax-dCas9VP160–2A-neo (a gift from Rudolf Jaenisch, Addgene plasmid #48227). The sequence for the 5GA linker was included in one of the primers. The original donor backbone was a gift from Dr. Ken-ichi T. Suzuki from Hiroshima University, Hiroshima, Japan.

The donor vector encodes the following amino acid sequence that corresponds to the 24-bp left homology arm of CTCF, a 5GA linker, AID, T2A, and Neo^R:
PEMILSMMGAGAGAGAGAGSVELNLRETELCLGLPGGDTVAPVTGNKRGFSETVDL
KLNLNNEPANKEGSTTHDVVTFDSKEKSACPKDPAKPPAKAQVVGWPPVRSYRKNV
MVSCQKSSGGPEAAAFVKVSMDGAPYLRKIDLRMYKSYDELSNALSNMFSSFTMG
KHGGEEGMIDFMNERKLMDLVNSWDYVPSYEDKDGDWMLVGDVPWPMFVDTCK
RLRLMKGSDAIGLAPRAMEKCKSRAGSGEGRGSLLTCGDVEENPGPRLETRMGSAI
EQDGLHAGSPAAWVERLFGYDWAQQTIGCSDAAVFRLSAQGRPVLFVKTDLSGALN
ELQDEAARLSWLATTGVPCAAVLDVVTEAGRDWLLLGEVPGQDLLSSHLAPAEKVS
IMADAMRRLHTLDPATCPFDHQAKHRIERARTRMEAGLVDQDDLDEEHQGLAPAEL
FARLKARMPDGEDLVVTHGDACLPNIMVENGRFSGFIDCGRLGVADRYQDIALATR
DIAEELGGEWADRFLVLYGIAAPDSQRIAFYRLLDEFF*.

The lentiviral vector for expressing TIR1 (Lentiv4-EFsp-Puro-2A-TIR1–9Myc) was constructed using PCR and Gibson Assembly Cloning kit (New England Biolabs). The backbone was modified from lentiCRISPR v2 (a gift from Feng Zhang, Addgene plasmid #52961) and the TIR1–9myc fragment was amplified from pBabe TIR1–9myc (a gift from Don Cleveland, Addgene plasmid #47328). The expressing cassette includes a puromycin resistant marker followed by sequences that code for P2A peptide and TIR1–9myc protein. The gene expression is driven by EFS promoter in the original lentiCRISPR v2. The maps and the sequences of the plasmids are available at the following URLs. CTCF-mouse-3sgRNA-CRISPRexp-AID (https://benchling.com/s/seq-1R4nJ8quYptUqerRWSdX), mCTCF24-AID-donor-Neo (https://benchling.com/s/seq-LtJu9OTscKJNCEMOk8ok),

Lentiv4-EFsp-Puro-2A-TIR1–9Myc (https://benchling.com/s/seq-6wSCsW3Kr9S1igXZ8H9K).

Transfection and establishment of CTCF-AID knock-in mESC clones

The cells were passaged once on 0.2% gelatin-coated feeder-free plates before transfection. The cells were transfected using the Mouse ES Cell Nucleofector Kit (Lonza) and Amaxa Nucleofector (Lonza) with 10 μg of the CRISPR plasmid and 5 μg of the donor plasmid following the manufacturer's instructions. After transfection, the cells were plated on drug-resistant MEFs (GlobalStem). Two days after transfection, drug selection was started by addition of 160 μg/ml G418 (Geneticin, Gibco) to the medium. Drug-resistant colonies were isolated and the clones with AID knock-in on both alleles were found by performing PCR of the genomic DNA using primers specific to sequences flanking the 3' end of the CTCF coding sequence (AAATGTTAAAGTGGAGGCCTGTGAG and AAGATTTGGGCCGTTTAAACACAGC). The sequence at the CTCF-AID junction on both alleles were checked by sequencing of allele-specific PCR products, which were generated by using either a CTCF-129-specific (CTGACTTGGGCATCACTGCTG) or a CTCF-Cast-specific (GTTTTGTTTCTGTTGACTTAGGCATCACTGTTA) forward primer and a reverse primer in the AID coding sequence (GAGGTTTGGCTGGATCTTTAGGACA). The expression of CTCF-AID fusion protein was confirmed by observing the difference in the molecular weight compared to the control cells by Western blot with anti-CTCF antibody (Millipore, 07–729).

Lentivirus production and infection

We produced the lentivirus for expressing TIR1–9myc using Lenti-X Packaging Single Shots system (Clontech) and infected the CTCF-AID knock-in mESCs following the manufacturer's instructions. After infection, the cells were selected by culturing with 1 μg/ml puromycin. Drug-resistant colonies were isolated and expression of TIR1–9myc was confirmed by Western blot using anti-Myc antibody (Santa Cruz, sc-40). Clones expressing high level of TIR1–9myc were used for the subsequent experiments.

Preparation of CTCF-depleted cells and neural progenitor cell differentiation

The CTCF-AID knock-in mESCs expressing TIR1–9myc were passaged on 0.2% gelatin-coated plates without MEFs. We added 1 ul 500 mM auxin (Abcam, ab146403) per 1 ml medium to deplete CTCF, and changed medium with auxin every 24 hours. Cells were harvested 24, 48 or 96 hours after starting auxin treatment. For NPC differentiation[62,63], the CTCF-AID knock-in mESCs were grown on MEFs and passaged on 0.2% gelatin-coated plates without MEFs one day before starting differentiation treatment. The cells were plated sparsely to avoid passaging to new plates during neural differentiation because most of the cells failed to attach to new plates after auxin treatment. On day 0, auxin was added to the CTCF-depleted cell samples, and LIF was deprived from the culture medium 6 hours after adding auxin. From day 1, 5 uM retinoic acid (Sigma, R2625) was added with LIF-deprived medium and auxin was also added continuously to the CTCF-depleted cell samples. Cells were harvested on day 2, day 4 and day 6. To harvest auxin-washout samples, auxin treatment was stopped on day 4 or day 6 and differentiation treatment was continued for

another 2 days. Alkaline phosphatase staining was performed on each time point using the AP Staining kit II (Stemgent, 00–0055).

Antibodies

Antibodies used in this study were rabbit anti-CTCF (Millipore, 07–729, for western blotting), rabbit anti-Histone H3 (abcam, ab1791, for western blotting), rabbit anti-CTCF (Active Motif, 61311, for microChIP-seq), rabbit anti-Rad21 (Santa Cruz, sc-98784, for microChIP-seq), rabbit anti-H3K4me1 (abcam, ab8895, for ChIP-seq), rabbit anti-H3K4me3 (Millipore, 04–745, for ChIP-seq), rabbit anti-H3K27ac (Active Motif, 39133, for ChIP-seq), mouse anti-H3K27me3 (Active Motif, 61017, for ChIP-seq), mouse anti-Myc antibody (Santa Cruz, sc-40, for western blotting), and mouse anti-Cas9 (Cell Signaling, 14697, for western blotting). Goat anti-Rabbit IgG (H+L)-HRP (Bio Rad, 1706515) and Goat anti-Mouse IgG (H+L)-HRP (Invitrogen, 31430) were used as secondary antibody for western blotting.

Western blotting

Cells were washed with PBS and scraped in cold PBS, and pelleted to be stored at −80°C. Two million cells were resuspended in 100 μL lysis buffer (20 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA, 1mM EGTA, 1% Triton X-100, 1x complete protease inhibitor (Roche)), and sonicated for 10 minutes total ON time with pulses of 15 second ON and OFF, and 40% amplitude using QSONICA 800R (Qsonica). Protein concentration was measured using Pierce BCA Protein Assay Kit (Thermo Fisher). Laemmli Sample Buffer (Bio-Rad) with 355 mM 2-Mercaptoethanol was mixed with 15 μg of each sample and incubated for 5 minutes at 95°C. The samples were run on 4–15% Mini-PROTEAN® TGX™ Precast Gels (Bio-Rad), and transferred onto nitrocellulose membranes at 100 V for 1 hour. The membranes were rinsed with 1x TBST and blocked with 5% dry milk at room temperature for 45 minutes. After washing with TBST, the membranes were incubated with diluted antibody in the blocking buffer overnight at 4°C. After overnight incubating, membranes were washed 4 times 5minutes in 1x TBST at room temperature, and incubated with secondary antibody in blocking buffer at room temperature for 45 minutes. After washing 4 times with TBST, the substrates were detected using Pierce ECL Western Blotting Substrate (Thermo Fisher).

Cell cycle analysis

Cells were grown in 6-well plates. After dissociation with Accutase (Innovative Cell Technologies), 2–5 million cells were washed with PBS and re-suspended in 300 μl ice-cold PBS. Cells were fixed for a minimum of 24h at 4°C after drop-wise addition of 800 μl ice-cold ethanol. After fixation, cells were pelleted and re-suspended in PBS containing 0.1% Triton X-100, 20 μg/mL Propidium iodide and 50 μg/ml RNase A. Cells were incubated for 30 min at 37°C before subjected to flow cytometry analysis.

MicroChIP-seq library preparation

MicroChIP-seq experiments for CTCF and Rad21 were performed as described in ENCODE experiments protocols (“Ren Lab ENCODE Chromatin Immunoprecipitation Protocol for

MicroChIP" in https://www.encodeproject.org/documents/) with minor modifications. Libraries were sequenced on HiSeq2500 or HiSeq4000 single end for 50 bp. Two biological replicates were prepared for each sample. See Supplementary Note for more detailed information.

ChIP-seq library preparation

ChIP-seq experiments for each histone mark were performed as described in ENCODE experiment protocols ("Ren Lab ENCODE Chromatin Immunoprecipitation Protocol" in https://www.encodeproject.org/documents/) with minor modifications. Libraries were sequenced on HiSeq4000 single end for 50 bp. Two biological replicates were prepared for each sample. See Supplementary Note for more detailed information.

RNA-seq library preparation

Total RNA was extracted from 1–2 million cells using the AllPrep Mini kit (QIAGEN) according to the manufacturer's instructions and 1 μg of total RNA was used to prepare each RNA-seq library. The libraries were prepared using TruSeq Stranded mRNA Library Prep Kit (Illumina). Library quality and quantity were estimated with TapeStation (Agilent Technologies) and Qubit (Thermo Fisher Scientific) assays. Libraries were sequenced on HiSeq4000 using 50 bp paired-end. Two biological replicates were prepared for each sample.

Hi-C library preparation

In situ Hi-C experiments were performed as previously described using the MboI restriction enzyme[4]. Libraries were sequenced on Illumina HiSeq 4000. Two biological replicates were prepared for each sample. See Supplementary Note for more detailed information.

PLAC-seq library preparation

PLAC-seq experiments were performed as previously described[25]. Libraries were sequenced on Illumina HiSeq 4000. Two biological replicates were prepared for each sample. See Supplementary Note for more detailed information.

ChIP-seq data analysis

Each fastq file was mapped to mouse genome (mm10) with BWA[64] -aln with "-q 5 -l 32 -k 2" options. PCR duplicates were removed using Picard MarkDuplicates (https://github.com/broadinstitute/picard) and the bigWig files were created using deepTools [65] with following parameters: bamCompare --binSize 10 --normalizeUsing RPKM --ratio subtract (or ratio). The deepTools was also used for generating heatmaps. Peaks were called with input control using MACS2 [66] with regular peak calling for narrow peaks (e.g. CTCF) and broad peak calling for broad peaks (e.g. H3K27me3, K3K4me1). Enhancer regions were characterized by the presence of both H3K4me1 and H3K27ac peaks, but not H3K4me3 peak. Promoter regions that are potentially activate genes like enhancers were characterized by the presence of H3K4me1, H3K27ac, and H3K4me3 peaks. Repressive distal elements that were analyzed in Extended Data Fig. 4j were characterized by the presence of H3K4me1 and

H3K27me3 peaks, but not H3K27ac peak. We used DEseq2[67] to calculate differences in peak levels between samples.

RNA-seq data analysis

RNA-seq reads (paired-end, 100 bases) were aligned against the mouse mm10 genome assembly using STAR[68]. The mapped reads were counted using HTSeq[69] and the output files from two replicates were subsequently analyzed by edgeR[70] to estimate the transcript abundance and to detect the differentially expressed genes. Only genes that had H3K4me3 ChIP-seq peaks on TSS were used for downstream analysis (Fig. 1d, e). Differentially expressed genes were called by FDR < 0.05 and fold change > 2 thresholds. RPKM was calculated using an in-house pipeline.

Hi-C data analysis

Hi-C reads (paired-end, 50 or 100 bases) were aligned against the mm10 genome using BWA[64] -mem. Reads mapped to the same fragment were removed and PCR duplicate reads were removed using Picard MarkDuplicates. Raw contact matrices were constructed using in-house scripts with 10 or 40 kb resolution, and then normalized using HiCNorm[71]. We used juicebox pre[72] to create hic file with -q 30 -f options. To visualize Hi-C data, we used Juicebox[72] and 3D Genome Browser (http://www.3dgenome.org). Topological domain boundaries were identified at 40-kb or 10-kb resolution based on the directionality index (DI) score and a Hidden Markov Model as previously described[1], and they were also identified based on insulation scores using peakdet (Billauer E, 2012. http://billauer.co.il/peakdet.html). The insulation score analysis was performed as previously described[73] and insulation scores on TAD boundaries were calculated by taking the average value of scores that overlapped with TAD boundaries. The stripe calling was performed using a homemade pipeline (shared from Feng Yue lab, Penn State University) that is based on the algorithm proposed in a previous study[74]. We used HiCCUPS[72] with options “-r 10000 -k KR -f 0.001 -p 2 -i 5 -d 50000” to identify Hi-C peaks as chromatin loops, and then we chose CTCF associated loops among them that were overlapped with convergently oriented CTCF ChIP-seq peaks in control cells. The aggregate analysis of CTCF associated loops were performed using APA[72] with default parameters.

To assess global changes in TAD boundary strength between samples, we performed a comparison of each samples’ aggregated boundary contact profile (Extended Data Fig. 2c). First, to generate a consensus set of TAD boundaries we performed a simple merge between boundaries from clone 1 before auxin treatment (Clone 1, 0 hr) and boundaries from clone 2 before auxin treatment (Clone 2, 0 hr). Two filtering steps were used to generate the final set of consensus boundaries: 1) We discarded boundaries there were within 3.04 Mb of a chromosome start or end, because we would not be able to extract a submatrix of the correct size for the aggregate analysis; 2) We discarded boundaries > 200kb, because these often represent regions of disorganized chromatin between TADs, rather than true TAD boundaries. Next, we extracted a Hi-C sub-matrix for each boundary in each sample. Each sub-matrix consists of a window of 3.04 Mb centered on the midpoint of the boundary in question. These boundary sub-matrices were then averaged to generate one 3.04 Mb matrix representing the average boundary contact profile in a given sample. To facilitate

comparison between samples, average boundary contact profiles were then normalized across samples using standard quantile normalization. We then made pairwise comparisons between samples by subtracting the average boundary contact profile of sample 1 from the average boundary contact profile of sample 2. The list of consensus TAD boundaries used here is the same as that described for the aggregate boundary analysis above.

PLAC-seq data analysis

PLAC-seq reads (paired-end, 50 bases) were aligned against the mm10 genome using BWA[64] -mem. Reads mapped to the same fragment were removed and PCR duplicate reads were removed using Picard MarkDuplicates. Filtered reads were binned at 10 kb size to generate the contact matrix. Individual bins that were overlapped with H3K4me3 peaks on TSSs were used for downstream differential contact analysis. For the peak calling, we used MAPS[75] with default settings and FitHiChIP[76] with coverage bias correction with default settings in 10 kb resolution.

For differential contact analysis, the raw contact counts in 10 kb resolution bins that have the same genomic distance were used as inputs for comparison. To minimize the bias from genomic distance, we stratified the inputs into every 10-kb genomic distance from 10 kb to 150 kb, and the other input bins with longer distances were stratified to have uniform size of input bins that were equal to that of 140–150 kb distance bins. Since each input showed negative binomial distribution, we used edgeR[70] to get the initial set of differential interactions. We only used bins that have more than 20 contact counts in each sample of two replicates for downstream analysis. The significances of these differential interactions are either due to the difference in their H3K4me3 ChIP coverage or 3D contacts coverage. Therefore, the chromatin contacts overlapping with differential ChIP-seq peaks (FDR < 0.05, logFC < 0.5) were removed and only the chromatin contacts with the same level of H3K4me3 ChIP-seq peaks were processed. In this differential analysis, we used all bins for inputs that included non-significant interactions that were not identified by MAPS or FitHiChIP peak caller, because the majority of short-range interactions were not identified as significant peaks due to their high background and the changes in the short-range interactions might be also critical for gene regulation. We identified a large number of differentially changed short-range interactions even though many of them were not identified as significant peaks, and we observed a clear correlation between these differentially changed interactions and the changes of active enhancer levels on their anchor regions during neural differentiation, suggesting these interaction changes might reflect the biological changes. We used significance level with change direction ($-/+ \log_{10}$(p-value)) instead of fold change to show the changes of interactions, because fold change tends to be small value especially in short-range interactions even though the change is actually significant for biological aspects.

Active/inactive contact (AIC) ratio/value

The change of chromatin contacts on enhancers and promoters are affected by the alteration of enhancer activities such as H3K27ac and H3K4me1 levels (Extended Data Fig. 4b), and it is also well known that gene expression levels have positive correlation with these active marks around their TSS. These findings suggest that information of contact counts itself

should involve the information of enhancer activity. Furthermore, the majority of genes have multiple E-P contacts with variable changes of contact frequencies. Therefore, we designed AIC value to represent quantitative activity of multiple E-P contacts and aimed to show the relationship between gene regulation change and E-P contacts change without using any quantitative values of histone marks. First, we summed total contact counts on active elements and promoters in each gene. As for promoter-promoter (P-P) contacts, they have similar function as E-P contacts[32]. However, it is still unclear that the same contact frequency of P-P contacts has the same effect as that of E-P contacts. Moreover, in H3K4me3 PLAC-seq datasets, P-P contacts correspond to peak-to-peak interactions that have generally higher contact counts than that of peak-to-non-peak interactions. Therefore, we divided the P-P contact counts by a certain integer that showed the highest correlation coefficient between gene expression changes and AIC value changes before summing total active contact counts. We tested integers from 0 to 10 to divide the P-P contact counts, and dividing by 3 and 8 showed the highest correlation coefficient between gene expression changes and AIC value changes in ESCs and NPCs, respectively. The simply summing of active contact counts is still not proper for comparison between different samples because they are affected by the difference of H3K4me3 peak levels on TSS in different samples. Therefore, in order to cancel the bias from the H3K4me3 peak levels in different samples, we also calculated total contact counts on inactive (non-active) regions and computed active/inactive contact (AIC) ratio on each gene by following formula.

AIC ratio on gene A = Sum of Active Contact counts (SAC) on gene A / Sum of Inactive Contact counts (SIC) on gene A

Next, we calculated the average of SICs from the comparing two samples on each gene, and multiplied them by AIC ratios to calculate AIC values. AIC values are computed as pseudo contact counts to perform differential analysis by edgeR[70] after rounding them to their nearest integer. The bias from different H3K4me3 levels on TSS in different samples does not occur by multiplying the common average value of the SICs.

AIC value on gene A = AIC ratio on gene A × Average of SICs of compared two samples on gene A

We also computed the changes of AIC values using Hi-C datasets in the same way, and we could observe comparable correlations with gene expression changes. In Extended Data Fig. 4h, we also calculated AIC values of E-P contacts and P-P contacts separately. In this case, SAC was calculated using only E-P contact counts or P-P contact counts, and SIC was calculated in the same way as described above.

Odds ratio calculation for CTCF-dependent E-P contacts enrichment

For Fig. 3a, b all PLAC-seq contacts (10 kb resolution) on promoter and enhancer were classified based on the distance from anchor sites (enhancer side or promoter side) to the nearest CTCF binding sites (Fig. 3a, categorized into 4×4 bins). They are also classified based on the number of CTCF motif sites around each anchor site (10 kb bin ±5 kb) (Fig. 3b, categorized into 5×5 bins). Then, we generated 2×2 tables based on whether they are

CTCF-dependent contacts or not (FDR < 0.05) and whether they were categorized into the bin or not. Odds ratios and p-values (Fisher's test) on each 2×2 tables were calculated.

For Fig. 3c, d all PLAC-seq contacts (10 kb resolution) on promoter and enhancer were classified based on the distance from anchor sites (enhancer side or promoter side) to the nearest CTCF binding sites (Fig. 3c, categorized into 4×4 bins). They are also classified based on the number of CTCF motif sites around each anchor site (10 kb bin ±5 kb) (Fig. 3d, categorized into 5×5 bins). Then, we generated 2×2 tables based on whether they are on differentially down-regulated genes or not (FDR < 0.05) and whether they were categorized into the bin or not. For the E-P contacts on differentially down-regulated genes, chromatin contacts that were identified as significant by peak calling were counted (p value < 0.01). Odds ratios and p-values (Fisher's test) on each 2×2 tables were calculated.

For Fig. 4a, all genes were classified based on the distance to the nearest interacting enhancer and the number of enhancers around TSS (< 200 kb) (categorized into 3×3 bins). The distance to the nearest interacting enhancer is represented by the shortest genomic distance of significant PLAC-seq peaks on enhancers and promoters (p-value < 0.01). Then, we generated 2×2 tables based on whether they are CTCF-independent stably-regulated genes or not (FDR < 0.05) and whether they were categorized into the bin or not. Odds ratios and p-values (Fisher's test) on each 2×2 tables were calculated. In Fig. 4b and Extended Data Fig. 9e, the same analysis as Fig. 4b was performed in CTCF-dependent down-regulated genes and CTCF-dependent up-regulated genes.

3D modelling

We used the Strings & Binders Switch (SBS) polymer model[36] to dissect the 3D spatial organization of the *Vcan* gene region in wild type and CTCF depleted NPC cells. In the SBS view, a chromatin filament is modelled as a Self-Avoiding Walk (SAW) chain of beads, comprising different specific binding sites for diffusing cognate molecular factors, called binders. Different types of binding sites are visually represented by distinct colors. Beads and binders of the same color interact with an attractive potential, so driving the folding of the chain. All binders also interact unspecifically with all the beads of the polymer by a weaker energy affinity (see below). We estimated the optimal number of distinct binding site types describing the locus and their arrangement along the polymer chain by using the PRISMR algorithm, a previously described machine learning based procedure[37]. In brief, PRISMR takes as input a pairwise experimental contact map (e.g. Hi-C) of the studied genomic region and, via a standard Simulated Annealing Monte Carlo optimization, returns the minimal number of different binding site types and their arrangement along the SBS polymer chain, which best reproduce the input contact map. Next, we ran Molecular Dynamics (MD) simulations of the inferred SBS polymers so to produce a thermodynamics ensemble of single molecule 3D conformations.

We focused on the genomic region chr13:89,200,000–92,000,000 (mm10) encompassing the mouse *Vcan* gene, in wild type and CTCF depleted NPC cells. Applied to our Hi-C contacts data of the region, at 10kb resolution, the PRISMR algorithm[37] returned in both cases polymer models made of 6 different types of binding sites. In our simulations, beads and binders interact via standard potentials of classical polymer physics studies[77] and the system

Brownian dynamics is defined by the Langevin equation. By using the LAMMPS software[78], we ran massive parallel MD simulations so producing an ensemble of, at least, 10^2 independent conformations. We started our MD simulations from initial SAW configurations and let the polymers evolve up to 10^8 MD time steps when the equilibrium globule phase is reached. We explored a range of specific and non-specific binding energies in the weak biochemical energy range, respectively from $3.0K_BT$ to $8.0K_BT$ and from $0K_BT$ to $2.7K_BT$, where K_B is the Boltzmann constant and T is the system temperature. For the sake of simplicity, those affinity strengths are the same for all the different types. All details about the model and MD simulations are described in [36,37]. To better highlight the locations of the *Vcan* gene and its regulatory elements in the two different cases, we produced a coarse-grained version of the polymers. We interpolated the coordinates of the beads with a smooth third-order polynomial curve and used the POV-ray software (Persistence of Vision Raytracer Pty. Ltd) to produce the 3D images.

For the model derived contact maps, we computed the average contact frequencies from our MD derived ensemble of 3D polymer model conformations for each cell type. We followed a standard approach that considers a pair of polymer sites in contact if their physical distance is lower than a threshold distance[37]. To compare model contact maps with corresponding Hi-C data in each cell type, we used the HiCRep stratum adjusted correlation coefficient (SCC)[79], a bias-corrected correlation designed for Hi-C comparison, with a smoothing parameter h=5 and an upper bound of interaction distance equal to 1.5Mb. To compute the model frequencies of multiple contacts, we proceeded similarly. Specifically, fixed a viewpoint site k, we accounted for a triple contact (i,j,k) between k and any pair of sites i,j along the locus if their relative physical distances were all lower than the threshold distance.

CTCF motif deletion and tethering dCas9-CTCF

The CRISPR/Cas9 system was used to delete CTCF motif nearby *Vcan* promoter. The sequences of the DNA targeted are listed below (the protospacer adjacent motif is underlined). The guide RNAs were generated using GeneArt Precision gRNA Synthesis Kit (Invitrogen).

5’-TTCAGCACAAGCGGAAAATAGGG-3’,

5’-CTGCTTGCAGTTGGGTGTTTCGG-3’

Transfection of gRNA and Cas9 ribonucleoprotein (EnGen SpyCas9, New England Biolabs) into mESCs was performed using Neon Transfection System, 10 ul tip kit (Life Technologies). The cells were grown for approximately one week, and individual colonies were picked into a 96-well plate. After expanding cells, genotyping by PCR and Sanger sequencing were performed to confirm the motif deletion.

For the generation of dCas9-CTCF tethered cell lines, plasmids containing sequences for dCas9 and CTCF were generated by modifying lenti-dCas-VP64-Blast (a gift from Feng Zhang, Addgene #61425). The VP64 cassette was replaced by CTCF sequences to generate dCas9-CTCF and neomycin resistant marker that was taken from pAC95-pmax-dCas9VP160–2A-neo (a gift from Rudolf Jaenisch, Addgene 48227) was inserted. To

generate gRNA plasmids to recruit dCas9, the gRNA oligos were inserted into the backbone vector (pSPgRNA, a gift from Charles Gersbach, Addgene Plasmid #47108). The gRNA was designed to target the top and bottom strand of *Vcan* promoter-proximal region which is close to the deleted CTCF motif locus. The sequences of the DNA targeted are listed below (the protospacer adjacent motif is underlined).

5'-CCTGCCTCCTTGGACAGAGACGG-3' (for top strand)

5'-GTCCCTTCCGTCTCTGTCCAAGG-3' (for bottom strand)

The plasmids for dCas9-CTCF and gRNA were extracted using PureLink HiPure Plasmid Midiprep Kit (Invitrogen). For the electroporation, 350 ng of dCas9-CTCF plasmids and 600 ng of gRNA plasmids (1 ul) were added to 0.1–0.2 million mESCs resuspended in 10 ul Buffer R (Invitrogen), and electric pulse was delivered with the setting of 1200 V, 20 ms, and 2 pulses. After culturing approximately 10 days, individual colonies were picked and genotyping and western blotting were performed to confirm the sequences from the transfection plasmids and their protein expression.

For deletion of enhancer region that is interacting with *Vcan* promoter. The sequences of the DNA targeted are listed below (the protospacer adjacent motif is underlined).

5'- AGGAACGGCCCATTCCCGAGGGG-3',

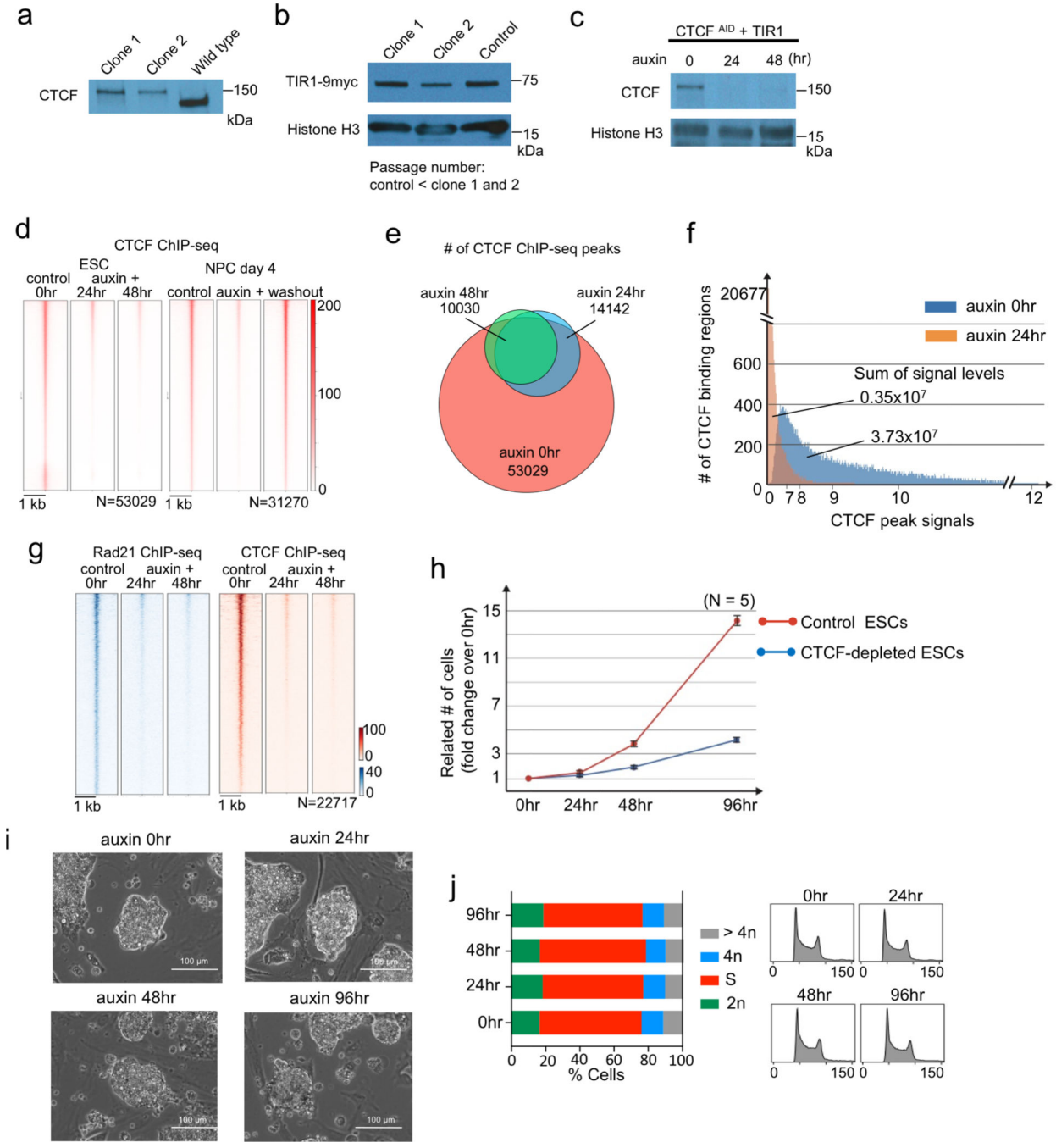
5'- CAATCAATAATAACACGCATAGG-3'

Generating gRNA and transfection of gRNA and Cas9 ribonucleoprotein into mESCs were performed in the same way as the deletion of CTCF motif was done. Genotyping by PCR and Sanger sequencing were performed to confirm the deletion.

Analysis of CTCF-occupied promoter genes in multiple mouse tissues

To analyze the CTCF ChIP-seq signals around promoters, we calculated fold changes of sample RPKM over input RPKM in each 50-bp bin and summed them in each promoter region (TSS ±10 kb) when the 50-bp bins were located at the regions of optimal IDR thresholded ChIP-seq peaks. Then correlation coefficient between these summed CTCF ChIP-seq signals and RNA-seq RPKM values across 9 mouse tissues was computed in each gene. The random datasets used as control in Fig. 5b was generated by randomly assigning the CTCF ChIP-seq signal levels at each promoter to each gene expression level (RNA-seq). Heatmap was generated for genes with high correlation coefficient (> 0.6). The values in the heatmap were calculated by $\log_2$(summed ChIP-seq signals / average value of all tissues) for promoter-proximal CTCF signal and $\log_2$(RPKM / average RPKM of all tissues) for gene expression. Lineage specificity of transcription was measured by Shannon entropy[42]. For DNA methylation levels around promoters, DNA methylation rates at CBSs (motif sequences ±100 bp) were calculated and averaged in each promoter region (TSS ±10 kb). PhastCons score[80] was used for conservation analysis. The highest pahstCons score at each CTCF motif locus was represented as the conservation score of each CBS.

Extended Data



Extended Data Fig. 1. Depletion of CTCF and characterization of CTCF-depleted cells.

a, b, Western blot showing AID-tagged CTCF and wild type CTCF (a) and the expression of TIR1 protein in two mESC clones. TIR1 expression in these clones that went through multiple passages was comparable to that in control cells with lower passage number. Uncropped images are available as source data online.

c, Western blot showing acute depletion of CTCF protein after 24 and 48 hours of auxin treatment. Uncropped images are available as source data online.

d, Heatmaps showing CTCF ChIP-seq signals centered at all regions of CTCF peaks identified in the control cells and CTCF occupancy at the same regions in CTCF-depleted cells at each time point of differentiation.

e, Venn-diagram comparing the number of CTCF ChIP-seq peaks identified in control and CTCF-depleted ESCs at each time point.

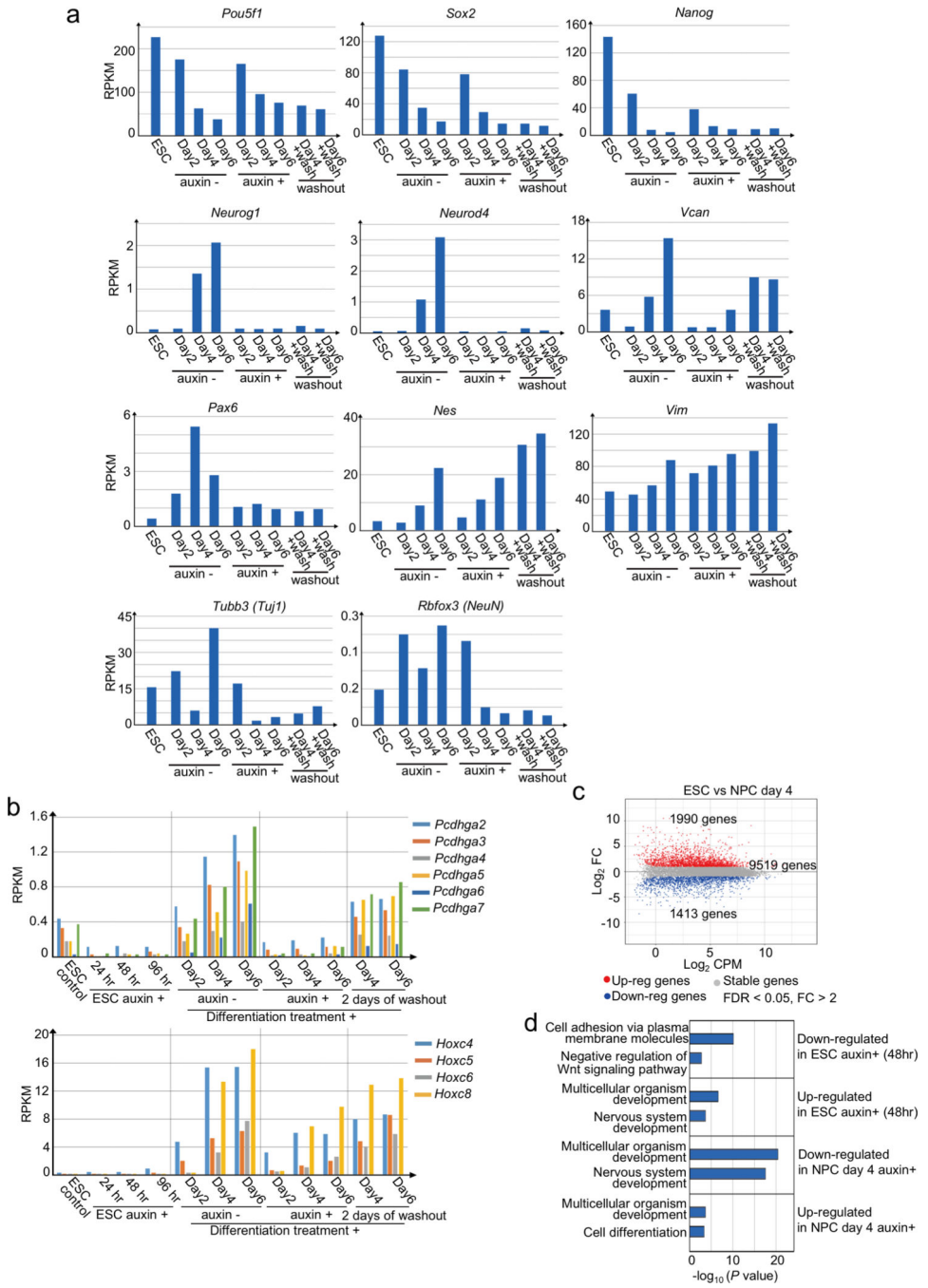
f, Histogram showing the number of CTCF binding regions in *y*-axis and the associated CTCF ChIP-seq signal level in *x*-axis. The CTCF signal levels in control cells and auxin treated cells were calculated for CTCF peak regions identified in the control cells.

g, Heatmaps comparing Rad21 ChIP-seq signals centered at all regions of Rad21 peaks identified in control and CTCF-depleted ESCs at each time point (left, blue heat map). CTCF occupancy are also shown (right, red heat map).

h, Growth curves of mouse ESCs with or without auxin treatment. Date are plotted as averages +/− standard deviation (n=5 independent experiments).

I, Bright-field microscopy images of mouse ESC colonies before and after auxin treatment.

j, Cell cycle analysis by flow cytometry using propidium iodide staining in control ESCs and after 24, 48, and 96 hours of auxin treatment.



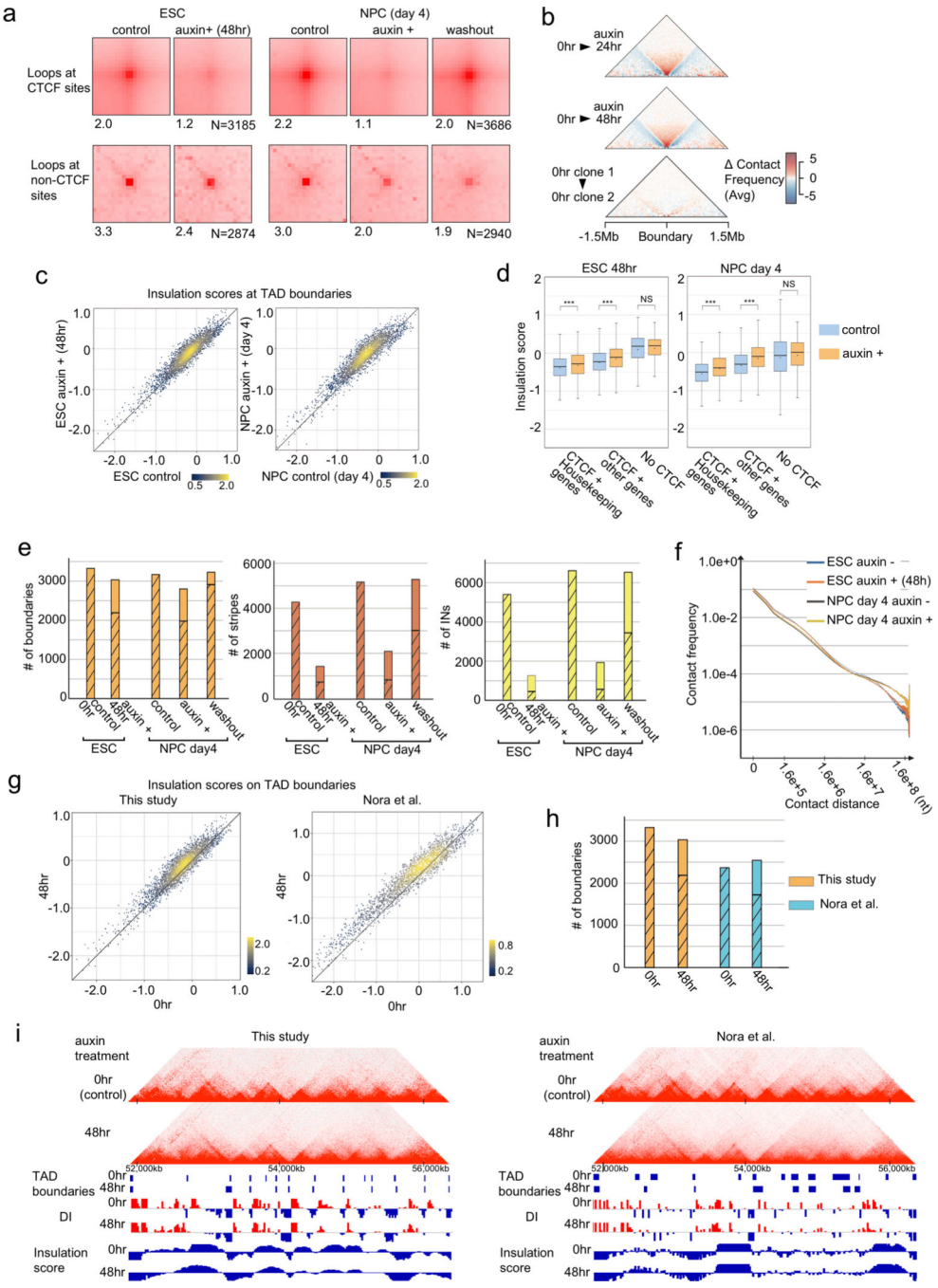
Extended Data Fig. 2. Transcriptional changes during neural differentiation of control and CTCF-depleted mES cells.

a, Gene expression profiles of pluripotent marker genes (*Pou5f1*, *Sox2*, *Nanog*) and examples of induction failure gene upon CTCF loss that is important for nervous system development (*Neurog1, Neurod4, Vcan, Pax6, Tubb3* (*Tuj1*), *Rbfox3* (*NeuN*)) in control and CTCF-depleted cells during differentiation from ESC to NPC and 2 days after washing out of auxin in NPCs.

b, Gene expression profiles of *Pcdhga* and *Hoxc* gene clusters during multiple days of auxin treatment in ESCs and during differentiation from ESC to NPC in control and CTCF-depleted cells followed by washing out auxin in NPCs.

c, Transcriptional changes between control ESCs and NPCs (day 4). Differentially up-regulated and down-regulated genes are plotted in red and blue, respectively (fold change > 2, FDR < 0.05).

d, Top two enriched GO terms of the sets of differentially expressed genes upon CTCF loss are shown along with p values (Fisher's exact test).



Extended Data Fig. 3. Global disruption of chromatin architecture upon CTCF loss.

a, APA on Hi-C peak loci (> 100-kb looping range) on convergent CTCF binding sites identified in control ESCs (n=3185) and NPCs (n=3686) and on Hi-C peak loci that have no CBSs (n=2874 (ESCs), n=2940 (NPCs). Scores on the bottom represent focal enrichment of peak pixel against pixels in its lower left.

b, Aggregate boundary analysis showing average change in boundary strength between samples. Each triangle is a contact map of the difference in the average contact profile at

TAD boundaries between two time points. The bottom column shows difference in the average boundary profile between the two control samples.

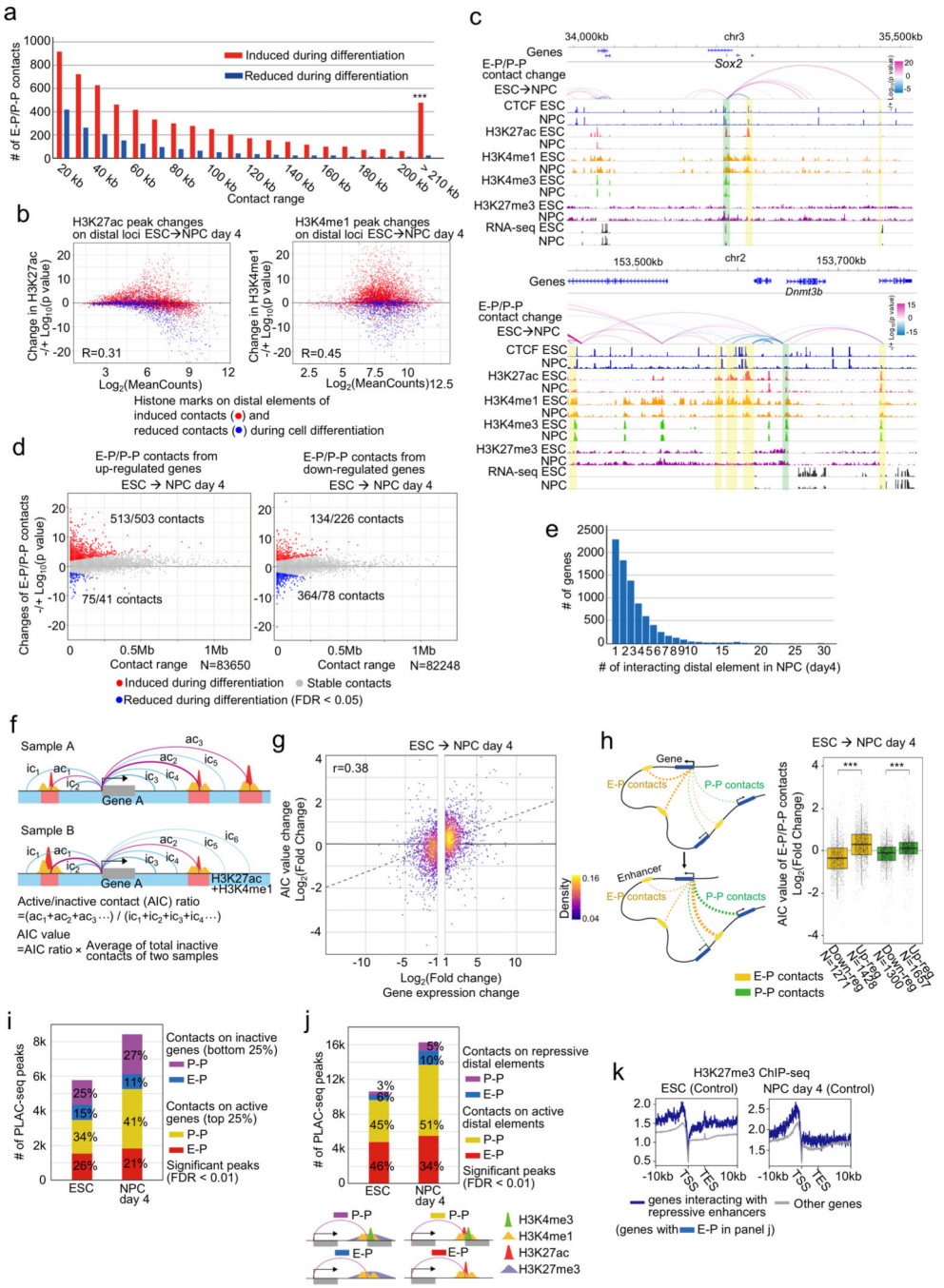
c, Scatter plots of insulation scores at TAD boundaries in control and auxin treated ESCs (left) and NPCs (right). A higher score denotes lower insulation.

d, Boxplots showing insulation scores at TAD boundaries that overlapped with housekeeping genes and CBSs, with other genes and CBSs, and TAD boundaries without CBSs in control and auxin treated ESCs and NPCs. All boxplots hereafter are defined as: Central bar, median; lower and upper box limits, 25th and 75th percentiles, respectively; whiskers, minimum and maximum value within the rage of (1st quartile-1.5*(3rd quartile-1st quartile)) to (3rd quartile+1.5*(3rd quartile- 1st quartile)). *** p value < 0.001, two-tailed t-test.

e, The number of TAD boundaries (left), stripes (middle), and insulated neighborhoods (INs) in control, CTCF-depleted, and auxin washout cells. Hatched bars indicate overlap with control cells.

f, Hi-C contact frequencies at each genomic distance.

g–i, Comparison of Hi-C datasets generated in this study and by Nora et al.[13]. Scatter plots of insulation scores at all TAD boundaries (g). Number of TAD boundaries in control and CTCF-depleted cells from both studies. Hatched bars indicate the overlap with control cells (h). Genome browser snapshots showing Hi-C contact heatmaps, TAD boundaries, directionality indices (DIs), and insulation scores analyzed in the two independent studies at the same genomic region in control and CTCF-depleted cells (i).



Extended Data Fig. 4. Features of E-P and P-P contacts that change during neural differentiation in control cells.

a, Histogram showing the number of significantly induced (red) and reduced (blue) E-P contacts between ESCs and NPCs and their genomic distances. *** p value < 0.001, Pearson's Chi-squared test.

b, Scatter plots showing changes of H3K27ac and H3K4me1 ChIP-seq signals at distal elements that display significantly induced (red) or reduced (blue) E-P or P-P contacts during neural differentiation.

c, Genome browser snapshots of *Sox2* (top) and *Dnmt3b* (bottom) loci. Arcs show changes of H3K4me3 PLAC-seq contacts on active elements and promoters between ESCs and NPCs (see Methods for details). The colors of arcs represent degrees of interaction change between samples (blue to red, $-/+\log_{10}$(p-value)) (Fisher's exact test). Promoter regions of *Sox2* and *Dnmt3b* and interacting enhancer regions are shown in green and yellow shadows, respectively. CTCF, H3K4me1, H3K27ac, H3K4me3, H3K27me3 ChIP-seq and RNA-seq in ESCs and NPCs (day 4) are also shown.

d, Scatter plots showing changes of E-P or P-P contacts anchored on up-regulated (left) and down-regulated (right) genes between ESCs and NPCs. Genomic distances between their two loop anchor sites are plotted on *x*-axis. Significantly induced and reduced chromatin contacts are shown as red and blue dots, respectively (FDR < 0.05).

e, Histogram showing the number of genes and the number of their interacting distal elements in NPCs. Genes without significant chromatin contacts were removed in this analysis.

f, Schematic representation of the AIC model to compute the correlation between changes of multiple E-P contacts and gene expression levels. H3K27ac and H3K4me1 peaks are shown as red and yellow peaks, respectively, and regions where these two types of peaks overlap are defined as active elements (red colored regions). Promoter-centered chromatin contacts on these active elements are shown as red arcs (active contacts) and other chromatin contacts are shown as blue arcs (inactive contacts). AIC ratio and value was calculated as indicated on the bottom (see Methods for details).

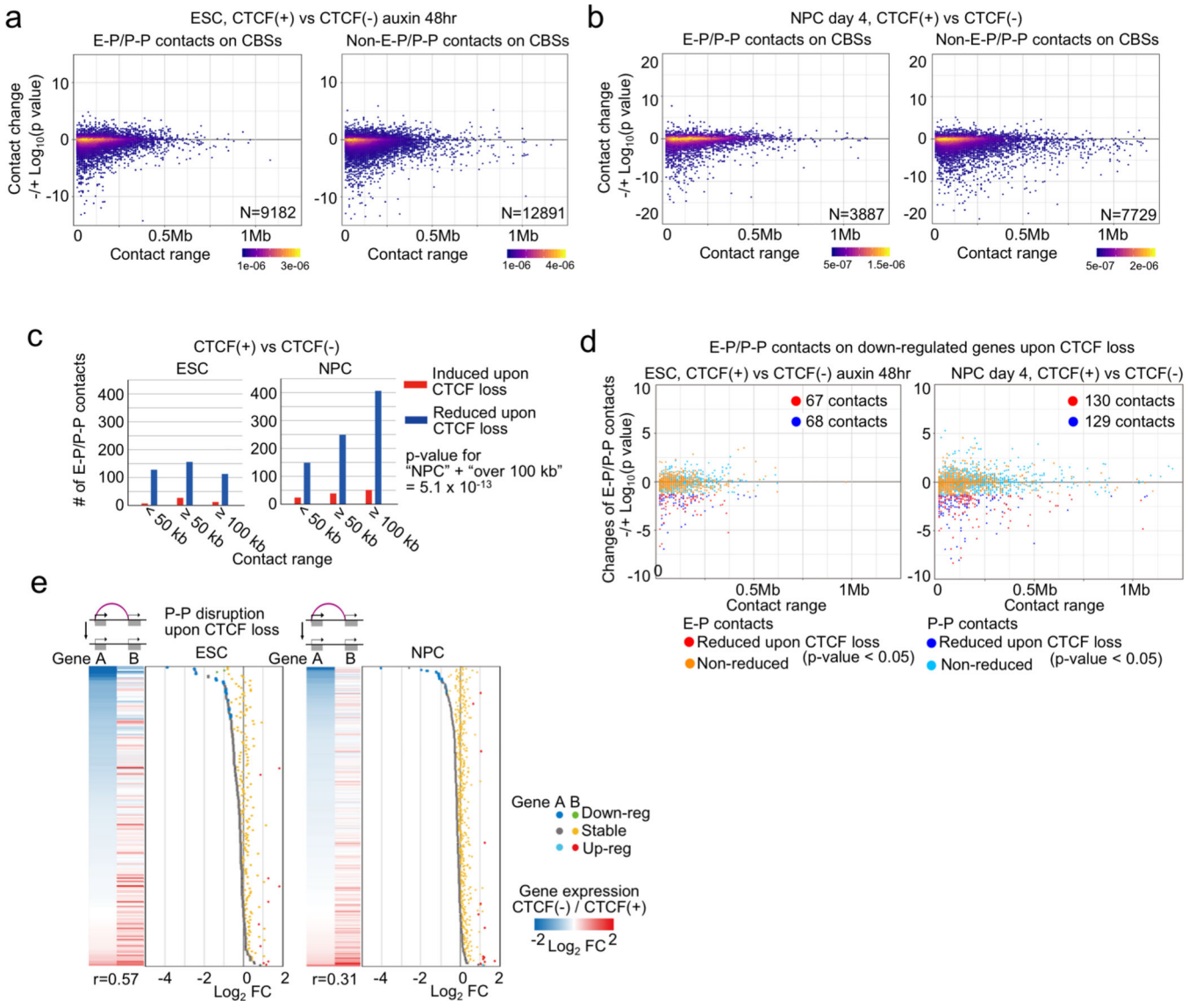
g, Scatter plots showing changes of AIC values and gene expression levels in differentially expressed genes during neural differentiation with linear approximation.

h, (Left) Schematic representation of a model to calculate AIC values using only P-P or E-P contacts. Promoter-centered chromatin contacts on active enhancers are shown as yellow arcs and chromatin contacts on other promoters are shown as green arcs. AIC ratios and values of P-P contacts and E-P contacts to other inactive contacts were calculated as shown in panel f. (Right) Box plots showing changes of the AIC values of P-P and E-P contacts in differentially expressed genes. The number of data points is indicated on the bottom. *** p value < 0.001, two-tailed t-test.

i, Histogram of the number of significant PLAC-seq peaks (FDR < 0.01) on P-P and E-P pairs anchored on active and inactive genes (top and bottom 25% of gene expression) in ESCs and NPCs.

j, Histogram showing the number of significant PLAC-seq peaks (FDR < 0.01) on P-P and E-P pairs anchored on active distal elements (presence of H3K4me1 and H3K27ac) and repressive distal elements (presence of H3K4me1 and H3K27me3, but not H3K27ac peaks) in ESCs and NPCs. Schematic representation of each type of chromatin contact is shown on the bottom.

k, Average enrichments of H3K27me3 ChIP-seq signals on TSSs and TESs of genes that interact with repressive distal enhancers identified in panel (j). H3K27me3 ChIP-seq signals on other genes are shown as control.



Extended Data Fig. 5. Changes of chromatin contacts upon acute CTCF loss illuminate relationship between CTCF-dependent P-P contacts and gene regulation.

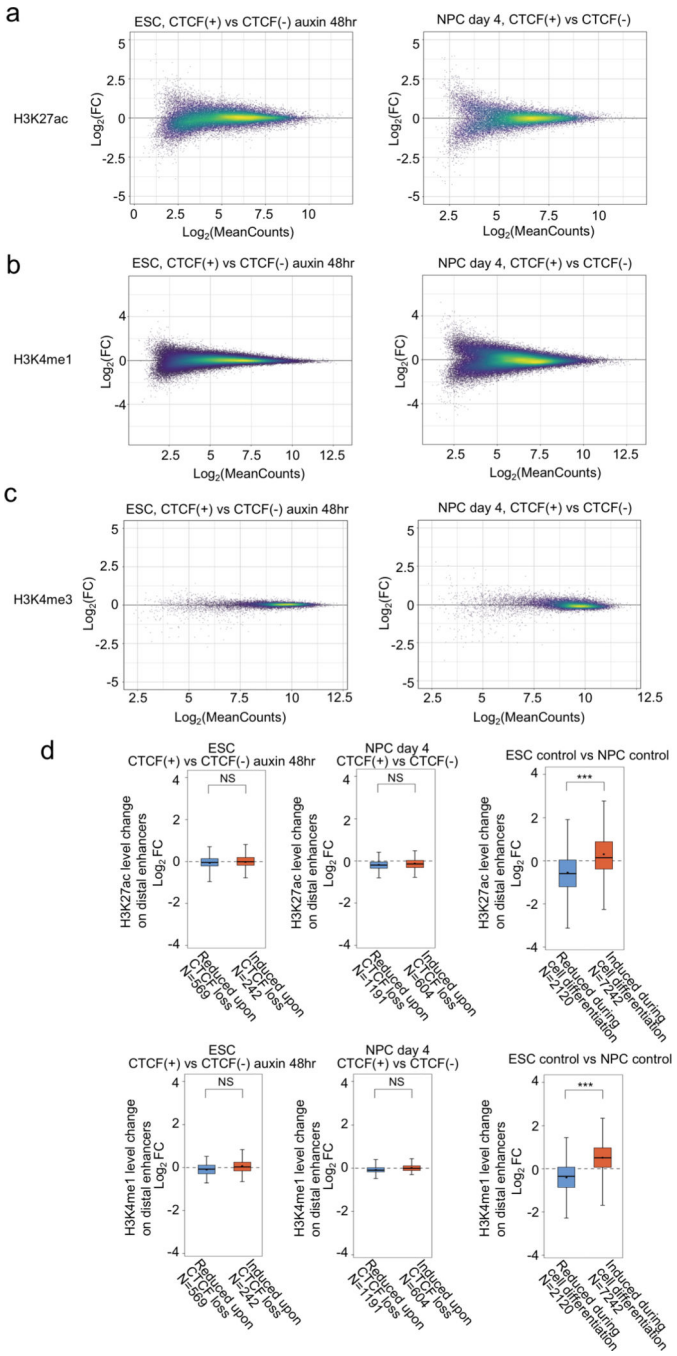
a, b, Scatter plots showing changes of H3K4me3 PLAC-seq contacts (*y*-axis) on convergently oriented CBSs and their loop ranges (*x*-axis). Chromatin contacts in CTCF-depleted cells were compared to the chromatin contacts in control cells in ESC (a) and NPC stage (day 4) (b). The plots were classified based on whether they are on promoters and enhancers (E-P and P-P) (left) or not (right).

c, Histograms showing the number of significantly changed E-P(P-P) contacts upon CTCF loss and their genomic distances in ESC (left) and NPC (right) stages. Significantly induced and reduced contacts are shown as red and blue bars, respectively. P value: Pearson's Chi-squared test for the comparison of the number of chromatin contacts that were long-range (≥ 100 kb) or not between the ESCs and NPCs.

d, Scatter plots showing changes of E-P(P-P) contacts anchored on CTCF-dependent down-regulated genes in ESC (left) and NPC stage (right). Chromatin contacts were classified

based on whether they were E-P or P-P contacts (red vs blue dots). Their genomic ranges are plotted in *x*-axis. The number of reduced E-P and P-P contacts are also shown, respectively (p value < 0.05).

e, Heatmaps and dotplots showing gene expression changes (fold change, FC) of genes that lost P-P contacts upon CTCF loss in ESC (left) and NPC (right) stages. Each gene pair interacting through CTCF-dependent P-P contacts is shown as either Gene A or Gene B. Gene A have a lower log2-FC than Gene B. Pearson Correlation coefficients (r) between the gene expression changes of the two paired genes are shown on the bottom of the heatmaps. Blue and green dots; down-regulated gene A and B (FDR < 0.05), gray and yellow dots; stably regulated gene A and B, light blue and red dots; up-regulated gene A and B (FDR < 0.05).

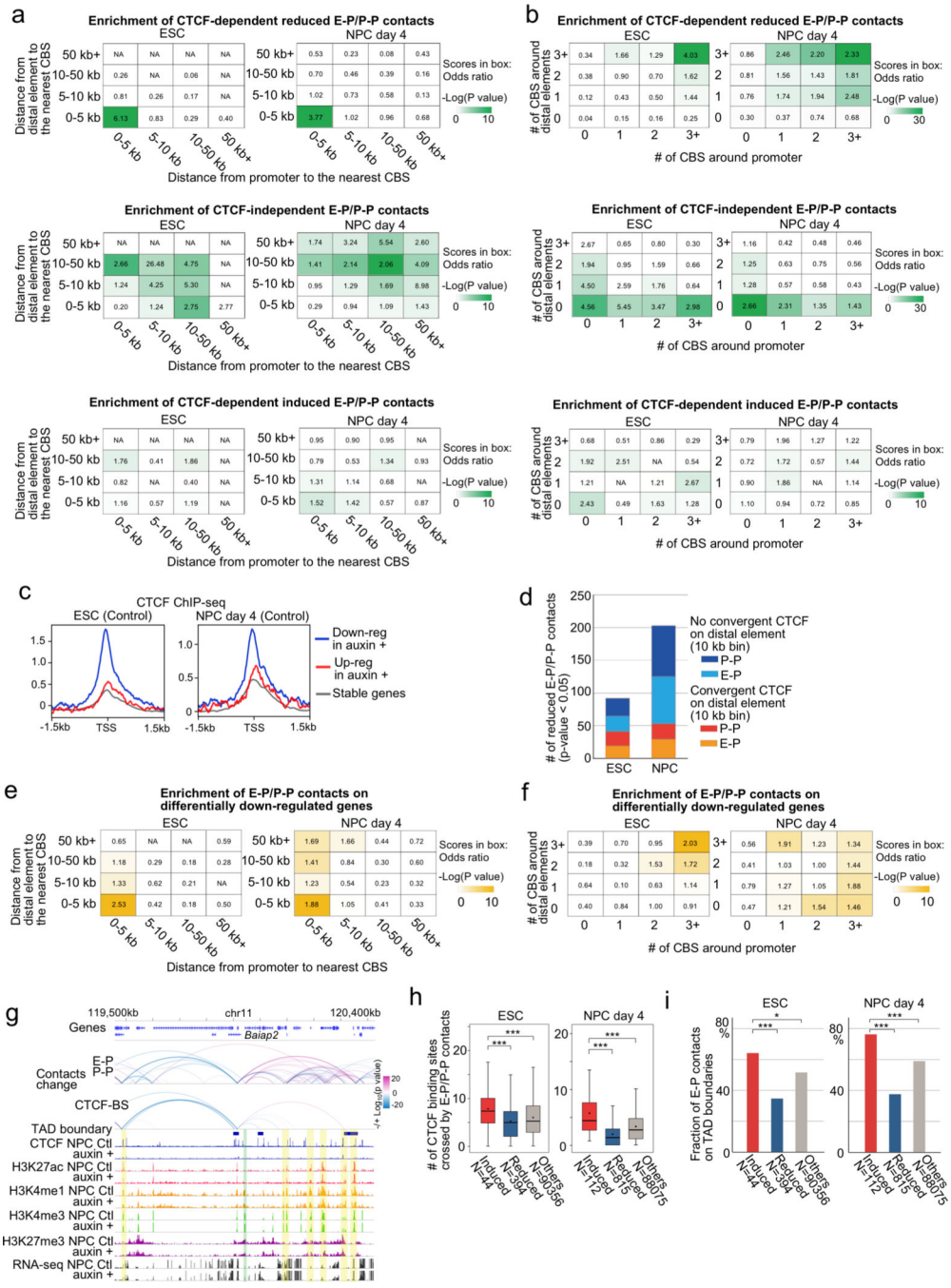


Extended Data Fig. 6. CTCF loss does not measurably alter histone modification at promoters and enhancers.

a–c, Scatter plots showing the changes of H3K27ac (a) and H3K4me1 (b) ChIP-seq signal levels upon CTCF loss on all significant peak regions in ESCs (left) and NPCs (right). The changes of H3K4me3 ChIP-seq signal levels on all peak regions on transcription start sites (TSSs) are also shown (c).

d, Boxplots showing the changes of H3K27ac (top) and H3K4me1 (bottom) ChIP-seq signal levels on distal element loci of all analyzed E-P contacts. The changes upon CTCF depletion in ESC (left) and NPC stage (middle) and the changes during neural differentiation (right)

are shown. The numbers of data points are also indicated on the bottom. NS not significant, *** p value < 0.001, two-tailed t-test.



Extended Data Fig. 7. Features of CTCF-dependent/-independent E-P and P-P contacts.

a, **b,** Enrichment analysis of CTCF-dependent reduced E-P and P-P contacts (top), CTCF-independent E-P and P-P contacts (middle), and CTCF-dependent induced E-P and P-P contacts (bottom). Chromatin contacts were categorized based on the distance from the loop anchor sites on the distal element side (vertical columns) or promoter side (horizontal

columns) to the nearest CBS (a) or based on the number of CBSs around loop anchor sites (10 kb bin ±5kb) on the distal element side (vertical columns) or promoter side (horizontal columns) (b). Enrichment values are shown by odds ratio (scores in boxes) and p-values (color) in ESCs (left) and NPCs (right) (see Methods).

c, Average enrichment of CTCF ChIP-seq signals on TSSs of CTCF-dependent up-regulated (red) or down-regulated (blue) and CTCF-independent genes (gray) in ESCs (left) and NPCs (right).

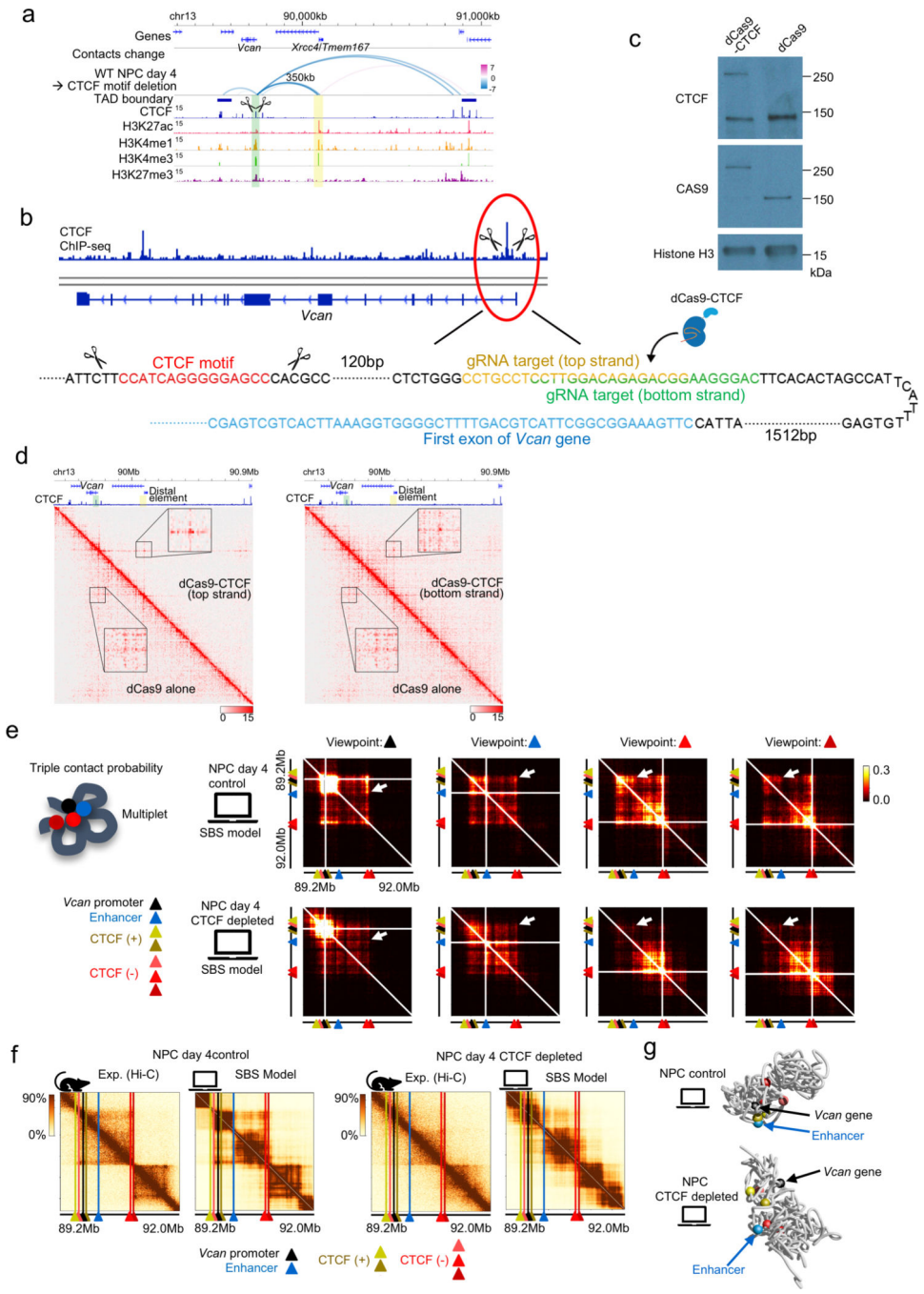
d, Histograms showing the number of reduced CTCF-dependent E-P and P-P contacts (p value < 0.05) anchored on CTCF-dependent down-regulated gene promoters with CBSs (TSS ± 5 kb) in ESCs (left) and NPCs (right). Chromatin contacts were classified based on whether their interacting distal elements were anchored on convergent CTCF or not (within 10 kb bin).

e, f, Enrichment analysis of E-P and P-P contacts anchored on CTCF-dependent down-regulated genes categorized based on the distance from the loop anchor sites on the distal element side (vertical columns) or promoter side (horizontal columns) to the nearest CBS (e). The same enrichment analysis categorized based on the number of CBSs around loop anchor sites (10 kb bin ±5kb) on distal element side (vertical columns) or promoter side (horizontal columns) (f). Enrichment values shown by odds ratio (scores in boxes) and p-values (color) in ESCs (left) and NPCs (right) (see Methods).

g, Genome browser snapshots of the *Baiap2* locus. Arcs show changes of chromatin contacts on E-P and on CBSs. The colors of arcs represent change from control cells to CTCF-depleted cells (blue to red, $-/+\log_{10}$(p-value)). Promoter regions of *Baiap2* and interacting enhancer regions are shown in green and yellow shadows, respectively. CTCF, H3K4me1, H3K27ac, H3K4me3, H3K27me3 ChIP-seq, and RNA-seq in control and CTCF-depleted NPCs, and TAD boundaries in control cells are also shown.

h, Boxplots showing the number of CBSs located between two anchor sites of significantly induced (red) or reduced (blue) E-P contacts upon CTCF loss, and CTCF-independent E-P contacts (gray). The numbers of data points are indicated on the bottom. *** p value < 0.001, two-tailed t-test.

i, The fraction of E-P and P-P contacts that overlapped with TAD boundaries in ESCs (left) and NPCs (left). The numbers of data points are indicated on the bottom. * p value < 0.05, *** p value < 0.001, Pearson’s Chi-squared test.



Extended Data Fig. 8. SBS polymer model and rescue experiments using dCas9-CTCF.

a, Genome browser snapshots of the *Vcan* locus. Arcs show changes of chromatin contacts anchored on the *Vcan* promoter, distal enhancer, and CBSs identified s between wild type NPCs and NPCs in which promoter-proximal CTCF motif sequences were deleted. The colors of arcs represent degrees of interaction change upon the deletion of CTCF motif sequences (blue to red, −/+$\log_{10}$(p-value)). The promoter region and interacting enhancer region are shown in green and yellow shadows, respectively. CTCF, H3K27ac, H3K4me1,

H3K4me3, and H3K27me3 ChIP-seq, and TAD boundaries in wild type NPCs are also shown.

b, Schematic representation of the dCas9-CTCF rescue experiments.

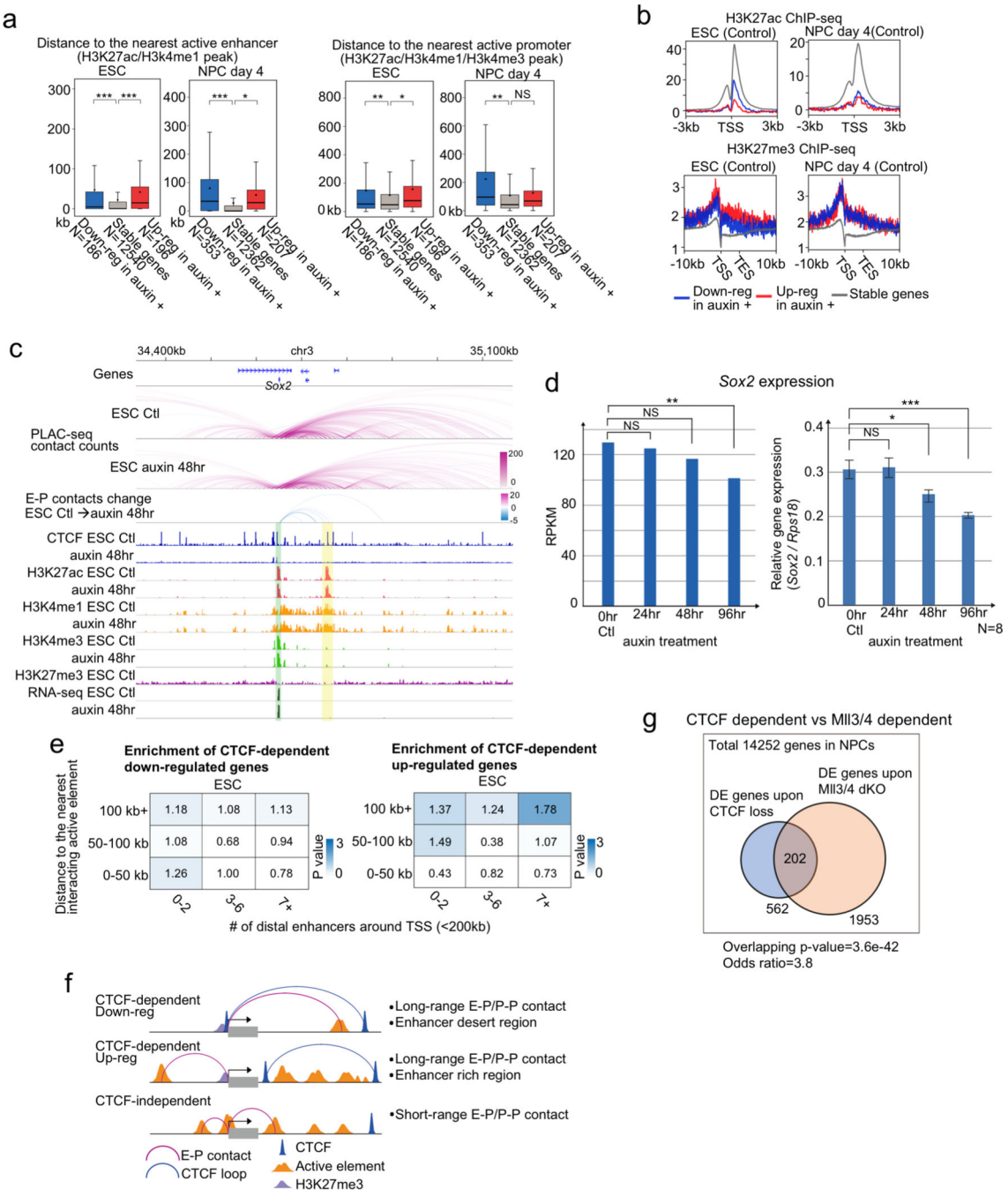
c, Western blot of cells lysates expressing dCas9-CTCF or dCas9 control plasmids. Uncropped images are available as source data online.

d, Snapshots of heatmaps around *Vcan* showing mapped reads of PLAC-seq in dCas9-CTCF (top and bottom strand) and dCas9 control cell lines. Peaks of chromatin contacts between the *Vcan* gene promoter and the downstream distal element are shown in zoom-in.

e, SBS model showing triplet interactions between the *Vcan* promoter (black), distal enhancer (blue) and CBSsthat weaken upon CTCF depletion (white arrows). Heatmaps from each viewpoint in control and CTCF-depleted NPCs are shown. CTCFs (+) (browns) and CTCFs (−) (reds) are convergently oriented.

f, Hi-C contact maps (left) of the *Vcan* locus in control and auxin treated NPCs and the SBS polymer model (right) (HiCRep stratum adjusted correlation SCC = 0.76 and SCC = 0.62 respectively). Genomic positions of *Vcan* promoter (black), distal enhancer (blue) and relevant motif-oriented CBSs (brown and red) are shown by colored triangles.

g, SBS derived 3D structures of the *Vcan* locus in control and CTCF-depleted NPCs, with relevant elements indicated by colored beads (color as in (b)).



Extended Data Fig. 9. Mechanisms of CTCF-dependent/-independent gene regulation.

a, Boxplots showing the distance from TSS to the nearest enhancer (left) and promoter (right) region in ESCs and NPCs. Red: CTCF-dependent up-regulated genes, blue: CTCF-dependent down-regulated genes, gray: CTCF-independent stably regulated genes. The numbers of genes analyzed in each group are indicated on the bottom. *** p value < 0.001, ** p value < 0.01, * p value < 0.05, two-tailed t-test.

b, Average enrichment of H3K27ac ChIP-seq signals on TSSs of CTCF-dependent up-regulated (red), and down-regulated (blue) genesand CTCF-independent genes (gray) in

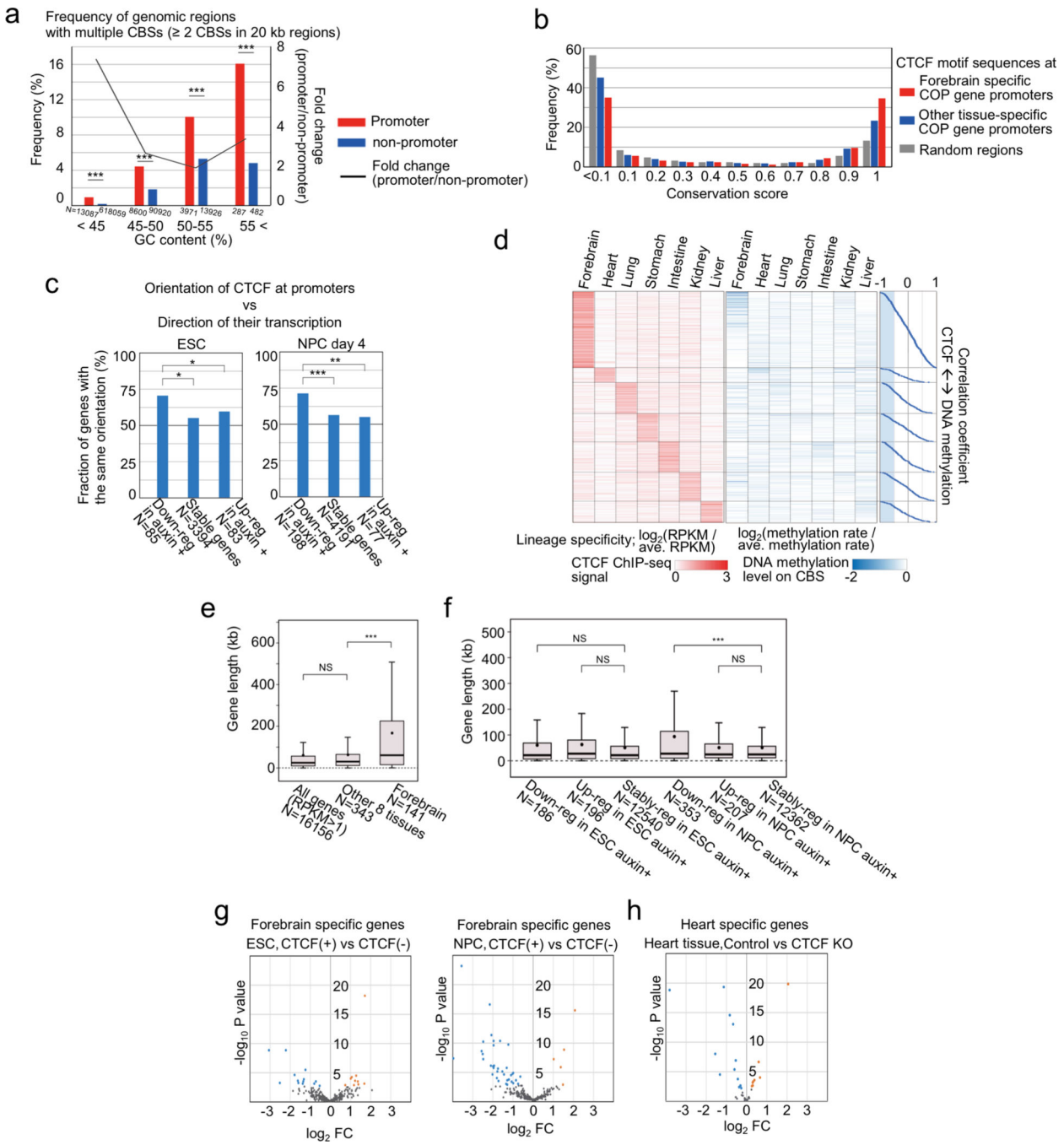
ESCs (left) and NPCs (right). Average enrichment of H3K27me3 ChIP-seq signals on TSSs and TESs are also shown (bottom).

c, d, Genome browser snapshot of the *Sox2* locus (c) whose reduction of expression level was moderate 24 or 48 hours after CTCF depletion in ESCs in RNA-seq and qPCR (d). The arcs show PLAC-seq contact counts in control (top) and CTCF-depleted ESCs (middle) at every 10-kb bin. Changes of chromatin contacts on enhancers and the *Sox2* promoter are also shown (bottom). *Sox2* gene promoter and interacting super enhancer are shown in green and yellow shadows, respectively. CTCF, H3K4me1, H3K27ac, H3K4me3, and H3K27me3 ChIP-seq, RNA-seq in control and CTCF-depleted ESCs are shown. The error bars in the right panel of (d) indicate standard deviation of 8 independent experiments. RPKM values were calculated from two RNA-seq replicates. NS not significant, * p value < 0.05, ** p value < 0.01, *** p value < 0.001, two-tailed t-test.

e, Enrichment analysis of CTCF-dependent down-regulated (left) and up-regulated (right) genes categorized based on the distance to the nearest interacting enhancer (vertical columns) and the number of enhancers around TSS (< 200 kb) (horizontal columns) in ESCs. Enrichment values are shown by odds ratio (scores in boxes) and p-values (color). The distance to the nearest interacting enhancer is represented by the shortest genomic distance of significant PLAC-seq peaks on enhancers and promoters (p-value < 0.01). (see Fig. 4b for the same analysis in NPCs).

f, Model for the general features of CTCF-dependent down-regulated (top), up-regulated genes (middle), and CTCF-independent genes (bottom).

g, Venn-diagram showing overlapping between CTCF dependent genes and Mll3/4 dependent genes in NPCs. Statistical significance based on Fisher’s exact test. Odds ratio represents the strength of association.



Extended Data Fig. 10. Features of tissue-specific CTCF occupied promoter genes.

a, Histogram showing frequencies of genomic regions with 2 or more CBSs in all analyzed 9 tissues, classified based on GC content levels. Black line shows fold change between the two groups. Total numbers of genomic regions analyzed in each group are indicated on the bottom. *** p value < 0.001, two-tailed t-test.

b, Histogram showing frequencies of CTCF motif sequences and their PhastCons conservation scores.

c, Histogram showing the fractions of genes whose promoter CTCF binding motifs were the same direction with the orientation of transcription. The fractions in CTCF-dependent down- and up-regulated genes and CTCF-independent genes in ESCs and NPCs are shown. The numbers of genes analyzed in each group are indicated on the bottom. * p value < 0.05, ** p value < 0.01, *** p value < 0.001, Pearson's Chi-squared test.

d, Heatmap showing lineage-specific DNA methylation levels at CBSs (motif sequences ±100 bp) in promoter regions of genes shown in Fig. 5c. The DNA methylation levels at multiple CBSs in the same promoter region (TSS ±10 kb) were averaged. Lineage-specificity of DNA methylation levels shown in the heatmap are calculated by $\log_2$(DNA methylation level / average methylation level of all tissues). The heatmap was sorted by correlation coefficient between CTCF ChIP-seq signal levels and DNA methylation levels across multiple tissues in each group. Each correlation coefficient is shown in the scatter plots (right) ($r < -0.5$, highlighted in blue).

e, Boxplots showing length of lineage-specific genes with CTCF occupied promoter that had high correlation coefficient (> 0.6) in Fig 5b, c. Forebrain-specific genes and other lineage-specific genes are shown at right and middle, respectively. All genes whose RNA-seq RPKM value is more than 1 in at least one tissue sample were used as control (left). The numbers of genes analyzed in each group are indicated on the bottom. NS not significant, *** p value < 0.001, two-tailed t-test.

f, Boxplots showing gene length of CTCF-dependent down-regulated, up-regulated and CTCF-independent genes in ESCs and NPCs. The numbers of genes analyzed in each group are indicated on the bottom. NS not significant, *** p value < 0.001, two-tailed t-test.

g, Volcano plots showing the gene expression changes of the forebrain-specific CTCF-occupied genes between control cells and CTCF-depleted cells in ESCs (left) and NPCs (right).

h, Volcano plots showing gene expression changes of heart-tissue-specific CTCF-occupied genes between control heart tissue and CTCF knockout heart tissue.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

We thank Drs. Elphrege Nora and Benoit Bruneau for exchanging datasets and reagents. We would like to give special thanks to Samantha Kuan for operating the sequencing instruments and Tristin Liu and Zhen Ye for helping with experiments. We would like to acknowledge the help of Drs. Victor Lobanenkov and Arshad Desai for giving helpful advice and the help of Drs. Feng Yue, Xiaotao Wang, Ivan Juric, and Armen Abnousi for sharing computational pipelines. We would also like to give special thanks to Drs. Ramya Raviram, Rongxin Fang, Yanxiao Zhang, Anthony Schmitt, and Sora Chee for sharing helpful information and protocols, as well as all the other members of the Ren laboratory. This work was supported by the Ludwig Institute for Cancer Research (B.R.), NIH (1U54DK107977–01) (B.R.), NIH (1U54DK107965) (H.Z.), a Ruth L. Kirschstein Institutional National Research Award from the National Institute for General Medical Sciences (T32 GM008666) (J.D.H.), and a Postdoc fellowship from the TOYOBO Biotechnology Foundation (N.K.).

References:

1. Heintzman ND et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459, 108–12 (2009). [PubMed: 19295514]

2. Long HK, Prescott SL & Wysocka J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell 167, 1170–1187 (2016). [PubMed: 27863239]

3. Shen Y. et al. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116–20 (2012). [PubMed: 22763441]

4. Andersson R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014). [PubMed: 24670763]

5. Consortium EP An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]

6. Yu M. & Ren B. The Three-Dimensional Organization of Mammalian Genomes. Annu Rev Cell Dev Biol 33, 265–289 (2017). [PubMed: 28783961]

7. Benabdallah NS et al. Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. Mol Cell 76, 473–484 e7 (2019).

8. Alexander JM et al. Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. Elife 8(2019).

9. Dixon JR et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–80 (2012). [PubMed: 22495300]

10. Phillips-Cremins JE et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell 153, 1281–95 (2013). [PubMed: 23706625]

11. Nora EP et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381–5 (2012). [PubMed: 22495304]

12. Rao SS et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–80 (2014). [PubMed: 25497547]

13. Nora EP et al. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell 169, 930–944.e22 (2017).

14. Sanborn AL et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A 112, E6456–65 (2015).

15. Fudenberg G. et al. Formation of Chromosomal Domains by Loop Extrusion. Cell Rep 15, 2038–49 (2016). [PubMed: 27210764]

16. Monahan K. et al. Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-alpha gene expression. Proc Natl Acad Sci U S A 109, 9125–30 (2012). [PubMed: 22550178]

17. Zhang X. et al. Fundamental roles of chromatin loop extrusion in antibody class switching. Nature 575, 385–389 (2019). [PubMed: 31666703]

18. Lee J, Krivega I, Dale RK & Dean A. The LDB1 Complex Co-opts CTCF for Erythroid Lineage-Specific Long-Range Enhancer Interactions. Cell Rep 19, 2490–2502 (2017). [PubMed: 28636938]

19. Rao SSP et al. Cohesin Loss Eliminates All Loop Domains. Cell 171, 305–320 e24 (2017).

20. Arzate-Mejia RG, Recillas-Targa F. & Corces VG Developing in 3D: the role of CTCF in cell differentiation. Development 145(2018).

21. Stik G. et al. CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response. Nat Genet (2020).

22. Nishimura K, Fukagawa T, Takisawa H, Kakimoto T. & Kanemaki M. An auxin-based degron system for the rapid depletion of proteins in nonplant cells. Nat Methods 6, 917–22 (2009). [PubMed: 19915560]

23. Holland AJ, Fachinetti D, Han JS & Cleveland DW Inducible, reversible system for the rapid and complete degradation of proteins in mammalian cells. Proc Natl Acad Sci U S A 109, E3350–7 (2012).

24. Krietenstein N. et al. Ultrastructural Details of Mammalian Chromosome Architecture. Mol Cell 78, 554–565 e7 (2020).

25. Fang R. et al. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. Cell Res 26, 1345–1348 (2016). [PubMed: 27886167]

26. Mumbach MR et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods 13, 919–922 (2016). [PubMed: 27643841]

27. Dowen JM et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell 159, 374–387 (2014). [PubMed: 25303531]
28. Krijger PH et al. Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming. Cell Stem Cell 18, 597–610 (2016). [PubMed: 26971819]
29. Bonev B. et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell 171, 557–572 e24 (2017).
30. Li Y. et al. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. PLoS One 9, e114485 (2014).
31. Li G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84–98 (2012). [PubMed: 22265404]
32. Diao Y. et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat Methods 14, 629–635 (2017). [PubMed: 28417999]
33. Thiecke MJ et al. Cohesin-Dependent and -Independent Mechanisms Mediate Chromosomal Contacts between Promoters and Enhancers. Cell Rep 32, 107929 (2020).
34. Landolt RM, Vaughan L, Winterhalter KH & Zimmermann DR Versican is selectively expressed in embryonic tissues that act as barriers to neural crest cell migration and axon outgrowth. Development 121, 2303–12 (1995). [PubMed: 7671797]
35. Wu Y. et al. Versican V1 isoform induces neuronal differentiation and promotes neurite outgrowth. Mol Biol Cell 15, 2093–104 (2004). [PubMed: 14978219]
36. Chiariello AM, Annunziatella C, Bianco S, Esposito A. & Nicodemi M. Polymer physics of chromosome large-scale 3D organisation. Sci Rep 6, 29775 (2016).
37. Bianco S. et al. Polymer physics predicts the effects of structural variants on chromatin architecture. Nat Genet 50, 662–667 (2018). [PubMed: 29662163]
38. Herz HM et al. Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. Genes Dev 26, 2604–20 (2012). [PubMed: 23166019]
39. Hu D. et al. The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. Mol Cell Biol 33, 4745–54 (2013). [PubMed: 24081332]
40. Yan J. et al. Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. Cell Res 28, 387 (2018). [PubMed: 29497152]
41. He Y. et al. Spatiotemporal DNA methylome dynamics of the developing mouse fetus. Nature 583, 752–759 (2020). [PubMed: 32728242]
42. Martinez O. & Reyes-Valdes MH Defining diversity, specialization, and gene specificity in transcriptomes through information theory. Proc Natl Acad Sci U S A 105, 9709–14 (2008). [PubMed: 18606989]
43. Lee DP et al. Robust CTCF-Based Chromatin Architecture Underpins Epigenetic Changes in the Heart Failure Stress-Gene Response. Circulation 139, 1937–1956 (2019). [PubMed: 30717603]
44. Wutz G. et al. ESCO1 and CTCF enable formation of long chromatin loops by protecting cohesin(STAG1) from WAPL. Elife 9(2020).
45. Hsieh TS et al. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Mol Cell 78, 539–553 e8 (2020).
46. Weintraub AS et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell 171, 1573–1588 e28 (2017).
47. Beagan JA et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res 27, 1139–1152 (2017). [PubMed: 28536180]
48. Caputo L. et al. The Isl1/Ldb1 Complex Orchestrates Genome-wide Chromatin Organization to Instruct Differentiation of Multipotent Cardiac Progenitors. Cell Stem Cell 17, 287–99 (2015). [PubMed: 26321200]
49. Monahan K, Horta A. & Lomvardas S. LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. Nature 565, 448–453 (2019). [PubMed: 30626972]
50. Schoenfelder S. et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res 25, 582–97 (2015). [PubMed: 25752748]
51. Hnisz D, Shrinivas K, Young RA, Chakraborty AK & Sharp PA A Phase Separation Model for Transcriptional Control. Cell 169, 13–23 (2017). [PubMed: 28340338]

52. Wang H. et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res 22, 1680–8 (2012). [PubMed: 22955980]
53. Renda M. et al. Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. J Biol Chem 282, 33336–45 (2007).
54. Leitch HG et al. Naive pluripotency is associated with global DNA hypomethylation. Nat Struct Mol Biol 20, 311–6 (2013). [PubMed: 23416945]
55. Ficz G. et al. FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. Cell Stem Cell 13, 351–9 (2013). [PubMed: 23850245]
56. Maurano MT et al. Role of DNA Methylation in Modulating Transcription Factor Occupancy. Cell Rep 12, 1184–95 (2015). [PubMed: 26257180]
57. Nanan KK et al. TET-Catalyzed 5-Carboxylcytosine Promotes CTCF Binding to Suboptimal Sequences Genome-wide. iScience 19, 326–339 (2019). [PubMed: 31404833]
58. Katainen R. et al. CTCF/cohesin-binding sites are frequently mutated in cancer. Nat Genet 47, 818–21 (2015). [PubMed: 26053496]
59. Kaiser VB, Taylor MS & Semple CA Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. PLoS Genet 12, e1006207 (2016).
60. Cancer Genome Atlas Research, N. et al. Integrated genomic characterization of endometrial carcinoma. Nature 497, 67–73 (2013). [PubMed: 23636398]

References:

61. Gribnau J, Hochedlinger K, Hata K, Li E. & Jaenisch R. Asynchronous replication timing of imprinted loci is independent of DNA methylation, but consistent with differential subnuclear localization. Genes Dev 17, 759–73 (2003). [PubMed: 12651894]
62. Strubing C. et al. Differentiation of pluripotent embryonic stem cells into the neuronal lineage in vitro gives rise to mature inhibitory and excitatory neurons. Mech Dev 53, 275–87 (1995). [PubMed: 8562428]
63. Bain G, Kitchens D, Yao M, Huettner JE & Gottlieb DI Embryonic stem cells express neuronal properties in vitro. Dev Biol 168, 342–57 (1995). [PubMed: 7729574]
64. Li H. & Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–60 (2009). [PubMed: 19451168]
65. Ramirez F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 44, W160–5 (2016). [PubMed: 27079975]
66. Zhang Y. et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137 (2008). [PubMed: 18798982]
67. Love MI, Huber W. & Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014). [PubMed: 25516281]
68. Dobin A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013). [PubMed: 23104886]
69. Anders S, Pyl PT & Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–9 (2015). [PubMed: 25260700]
70. Robinson MD, McCarthy DJ & Smyth GK edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–40 (2010). [PubMed: 19910308]
71. Hu M. et al. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics 28, 3131–3 (2012). [PubMed: 23023982]
72. Durand NC et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst 3, 99–101 (2016). [PubMed: 27467250]
73. Crane E. et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 523, 240–4 (2015). [PubMed: 26030525]
74. Vian L. et al. The Energetics and Physiological Impact of Cohesin Extrusion. Cell 175, 292–294 (2018). [PubMed: 30241609]

75. Juric I. et al. MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. PLoS Comput Biol 15, e1006982 (2019).

76. Bhattacharyya S, Chandra V, Vijayanand P. & Ay F. Identification of significant chromatin contacts from HiChIP data by FitHiChIP. Nat Commun 10, 4221 (2019). [PubMed: 31530818]

77. Kremer K. & Grest GS Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. The Journal of Chemical Physics 92, 5057–5086 (1990).

78. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. Journal of computational physics 117, 1–19 (1995).

79. Yang T. et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res 27, 1939–1949 (2017). [PubMed: 28855260]

80. Siepel A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15, 1034–50 (2005). [PubMed: 16024819]

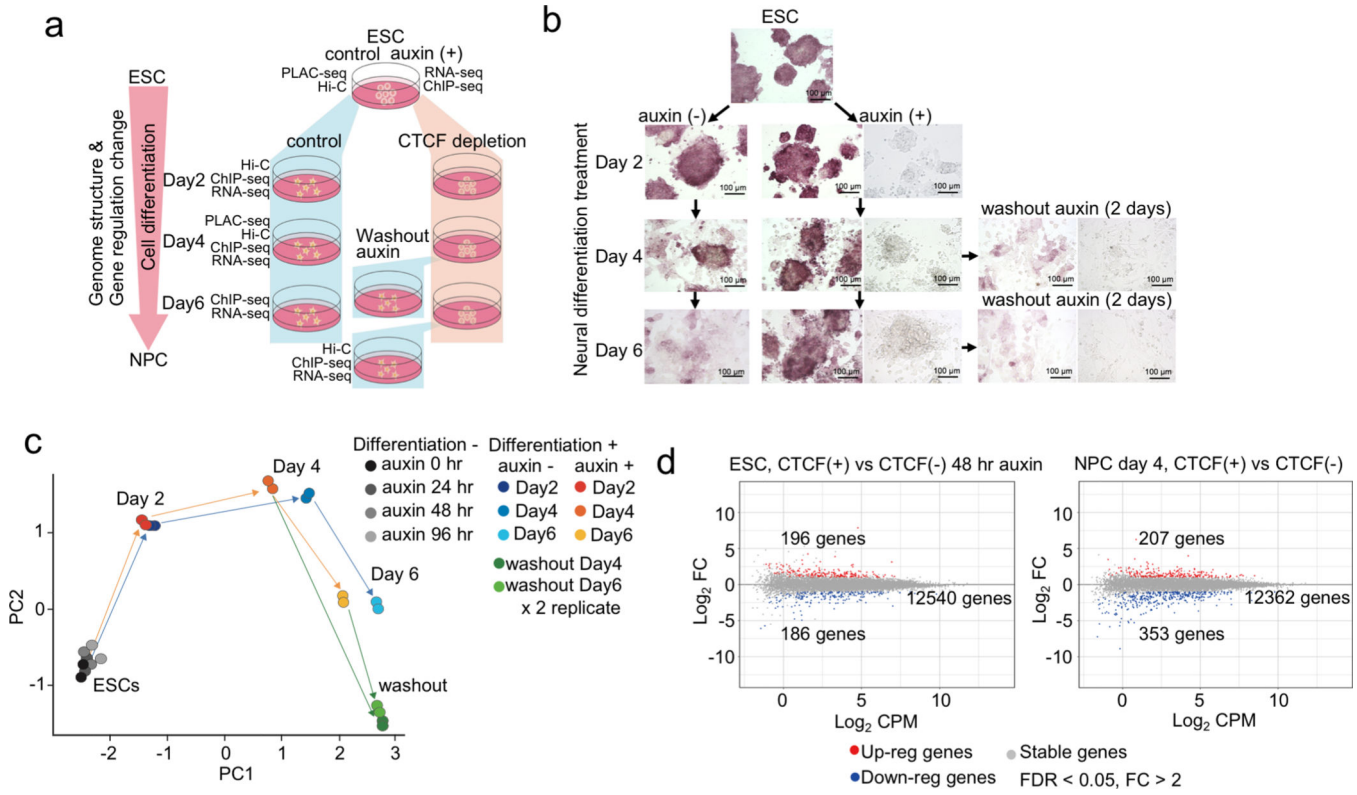


Fig. 1 |. CTCF loss impedes cell differentiation from ESC to NPC.

a, Schematic representation of experimental design and sample preparation. The auxin-inducible degron system was used to deplete CTCF during cell differentiation from ESCs to NPCs (day 2, 4, 6). An additional 2 days of neural differentiation was performed after washing out auxin for day 4 and day 6 differentiated cells. Types of experiments performed at each time point are indicated.

b, Microscopic images of cells under the treatment regime described in (a). Alkaline phosphatase staining was performed at every time point. Non-stained bright-field images of each auxin treated sample and auxin washout sample are also shown on the right.

c, Principal component analysis of gene expression profiles of control and CTCF-depleted cells at each time point of cell differentiation and 2 days after washing out auxin. Gene expression profiles in ESCs with multiple days of auxin treatment (24, 48, and 96 hours) were also analyzed. Two replicates of each sample are shown.

d, Gene expression changes upon CTCF depletion in ESCs (left, 48 hours with or without auxin) and in differentiated cells (right, differentiation day 4 with or without auxin). Differentially up-regulated and down-regulated genes are plotted in red and blue, respectively (fold change > 2, FDR < 0.05).

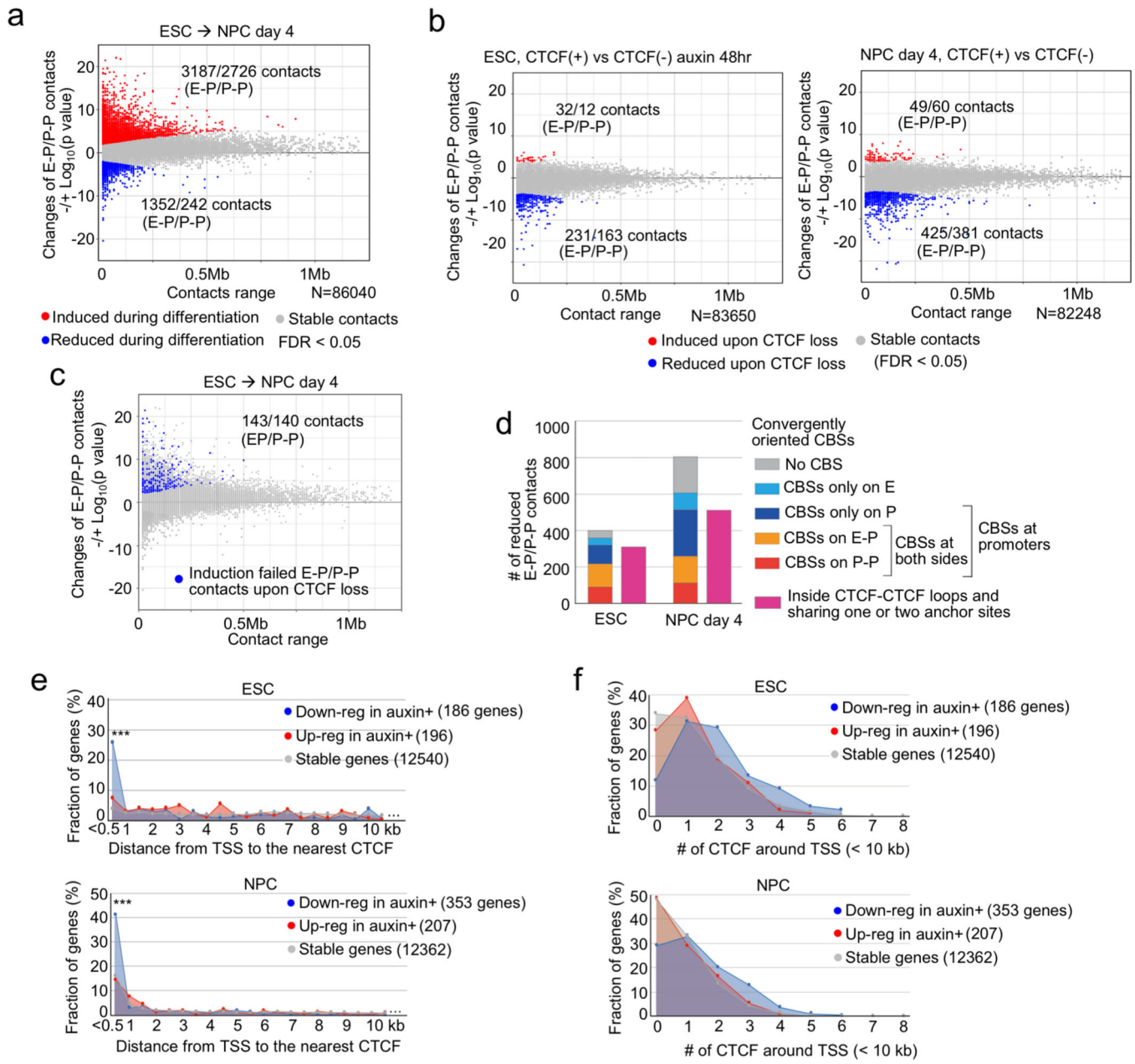


Fig. 2 |. CTCF loss reduces promoter-anchored contacts at a modest number of genes.

a, b, Scatter plots showing genome-wide changes of chromatin contacts anchored on promoters and enhancers (y-axis) identified in differential interaction analysis between ESCs and NPCs (a) and between control and CTCF-depleted cells in ESC (b, left) and NPC stage (day 4) (b, right). Genomic distances between their two loop anchor sites are plotted in *x*-axis. Significantly induced and reduced chromatin contacts are shown as red and blue dots, respectively (FDR < 0.05). The interaction changes are shown by significance value (−/+$\log_{10}$(p-value)) (Fisher's exact test, n=2 independent experiments). The numbers of significantly changed E-P and P-P contacts are indicated.

c, E-P and P-P contacts that were significantly induced during neural differentiation but significantly reduced upon CTCF loss were plotted as blue dots on the scatter plots displayed in (a). The numbers of significantly changed E-P and P-P contacts are indicated.
d, Histogram showing the numbers of significantly reduced E-P and P-P contacts upon CTCF loss that were categorized based on whether their anchor sites (10-kb bin) on promoter and enhancer regions have convergently oriented CTCF binding sites (CBSs) or not. Purple bars show the number of E-P and P-P contacts that were located inside CTCF-CTCF loops and whose anchor sites overlapped with at least one anchor site of CTCF-CTCF loops.
e, f, Fraction of genes classified based on genomic distance from TSS to the nearest CTCF ChIP-seq peak (e), and fraction of genes classified based on the number of CTCF ChIP-seq peaks around TSS (< 10 kb) (f) are shown. CTCF-depletion induced down-regulated genes (blue), up-regulated genes (red), and CTCF-independent genes (gray) in ESCs (top) and NPCs (bottom) are shown. *** p value < 0.001, Pearson's Chi-squared test for the comparison of the number of genes that had CTCF at TSS (< 500 bp) or not between down-regulated genes and the other types of genes.

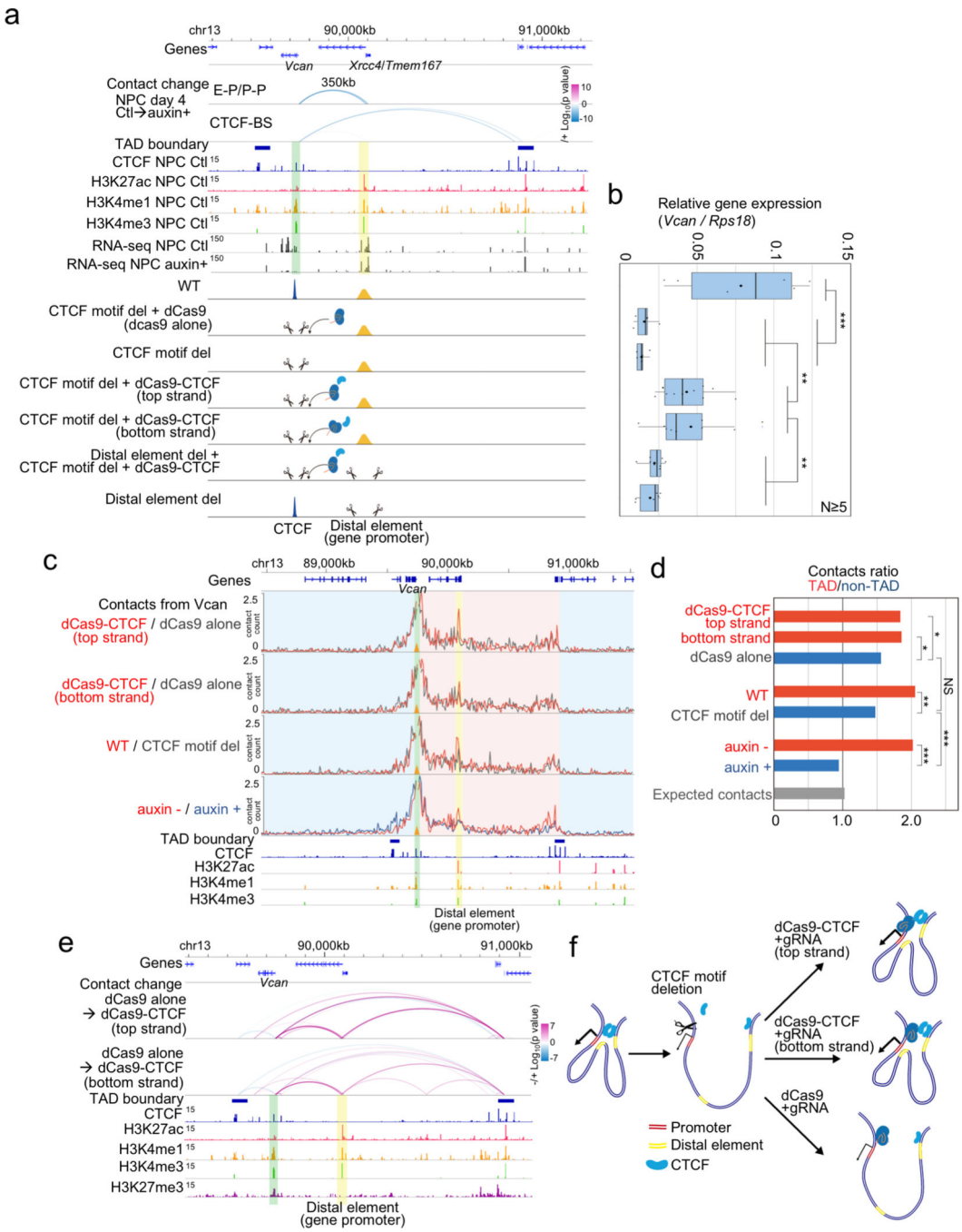


Fig. 3 |. Rescue of CTCF-dependent chromatin contacts and gene activation by artificial tethering of CTCF to the promoter.

a, Genome browser snapshots of a region around the *Vcan* gene. Arcs show the changes of chromatin contacts on enhancers and promoters (E-P/P-P) and CTCF binding sites (CTCF-BS); colors represent degrees of interaction change (blue to red, $-/+\log_{10}$(p-value)) (Fisher's exact test, n=2 independent experiments). The *Vcan* promoter and the 350 kb downstream distal active element (corresponding to the *Xrcc4/Tmem167* gene promoter) are shown in green and yellow shadows, respectively. ChIP-seq of H3K27ac, H3K4me3, H3K4me1,

RNA-seq, and TAD boundaries in NPCs are also shown. Schematic description of each cell line is shown at the bottom.

b, Boxplots of *Vcan* expression levels for each cell line indicated in (a) after differentiation (NPC). RT-qPCR assay was performed 5 times for each cell line. Central bar, median; lower and upper box limits, 25th and 75th percentiles, respectively; whiskers, minimum and maximum value within the rage of (1st quartile-1.5*(3rd quartile- 1st quartile)) to (3rd quartile+1.5*(3rd quartile- 1st quartile)). ** p value < 0.01 and *** p value < 0.001, two-tailed t-test.

c, The effects of artificial tethering of dCas9-CTCF (top and bottom strand) on promoter-anchored contacts at *Vcan* are compared to dCas9 alone control (red vs gray line). Normalized contact counts of H3K4me3 PLAC-seq (lines in 10-kb resolution) originating from the *Vcan* promoter (shadowed in green) are shown. Yellow shadow: promoter-promoter contacts. Red shadow: contacts with regions inside TAD. Blue shadow: contacts with regions outside TAD. Wild type cells, *Vcan* promoter-proximal CBS deleted cells, and auxin treated/untreated cells are shown for comparison. Data merged from two independent experiments are shown in the figure.

d, Ratio of total *Vcan* promoter-anchored contact counts throughout intra-TAD (shadowed in red in (c)) to other total contact counts outside the TAD (shadowed in blue in (c)). (TAD/non-TAD ratio) were computed in each sample. The ratio calculated by expected contact counts in wild type NPCs is shown at the bottom. *** p value < 0.001, ** p value < 0.01, * p value < 0.05, Pearson's Chi-squared test for the comparison of the total contact counts inside the TAD or outside the TAD between the compared two replicates samples.

e, Changes of chromatin contacts upon artificial tethering of CTCF at the promoter on top and bottom strand are shown. The colors of arcs represent degrees of interaction change from control NPCs (dCas9 alone) to dCas9-CTCF tethered NPCs (blue to red, $-/+\log_{10}$(p-value)) (Fisher's exact test, two independent experiments).

f, Schematic representation of observations of the rescue experiments.

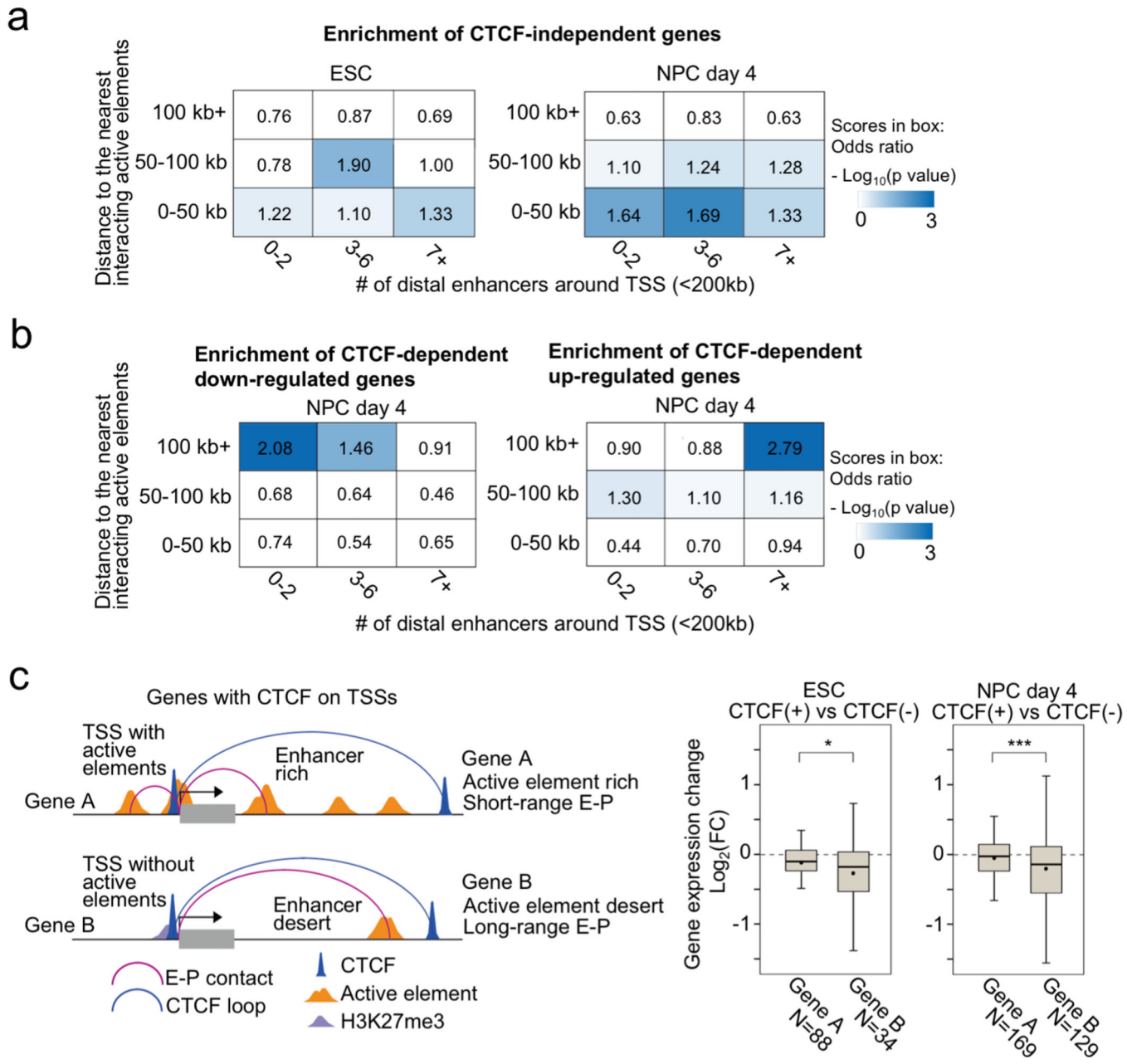


Fig. 4 |. General features of CTCF-dependent/-independent genes.

a, Enrichment analysis of CTCF-independent genes categorized based on the distance to the nearest interacting enhancer (vertical columns) and the number of enhancers around TSS (< 200 kb) (horizontal columns) in ESCs (left) and NPCs (right). Enrichment values are shown by odds ratio (scores in boxes) and p-values (color). The distance to the nearest interacting enhancer is represented by the shortest genomic distance of significant PLAC-seq peaks on enhancers and promoters (p-value < 0.01) For details on the odds ratio calculation and statistical analysis, see Methods.

b, Enrichment analysis of CTCF-dependent down-regulated genes (left) and up-regulated genes (right) categorized based on the distance to the nearest interacting enhancer (vertical

columns) and the number of enhancers around TSS (< 200 kb) (horizontal columns) in NPCs. Enrichment values are shown by odds ratio (scores in boxes) and p-values (color). The distance to the nearest interacting enhancer is represented by the shortest genomic distance of significant PLAC-seq peaks on enhancers and promoters (p-value < 0.01) For details on the odds ratio calculation and statistical analysis, see Methods. Extended Data Fig. 9e shows the same analysis in ESCs.

c, Schematic representation of two types of genes with CTCF binding peaks on their TSSs (< 1 kb). For Gene A, the shortest E-P contact (PLAC-seq peak signal p-value < 0.01) is shorter than 50 kb genomic distance and there are 7 enhancers or more around the TSS (< 200 kb). For Gene B, the shortest E-P contact is longer than 50 kb and there are 2 enhancers or less around TSS. Boxplots of gene expression changes upon CTCF loss in Gene A and Gene B classes in ESCs (left) and NPCs (right). Central bar, median; lower and upper box limits, 25^{th} and 75^{th} percentiles, respectively; whiskers, minimum and maximum value within the rage of (1^{st} quartile-1.5*(3^{rd} quartile- 1^{st} quartile)) to (3^{rd} quartile+1.5*(3^{rd} quartile- 1^{st} quartile)). * p value < 0.05 and *** p value < 0.001, two-tailed t-test.

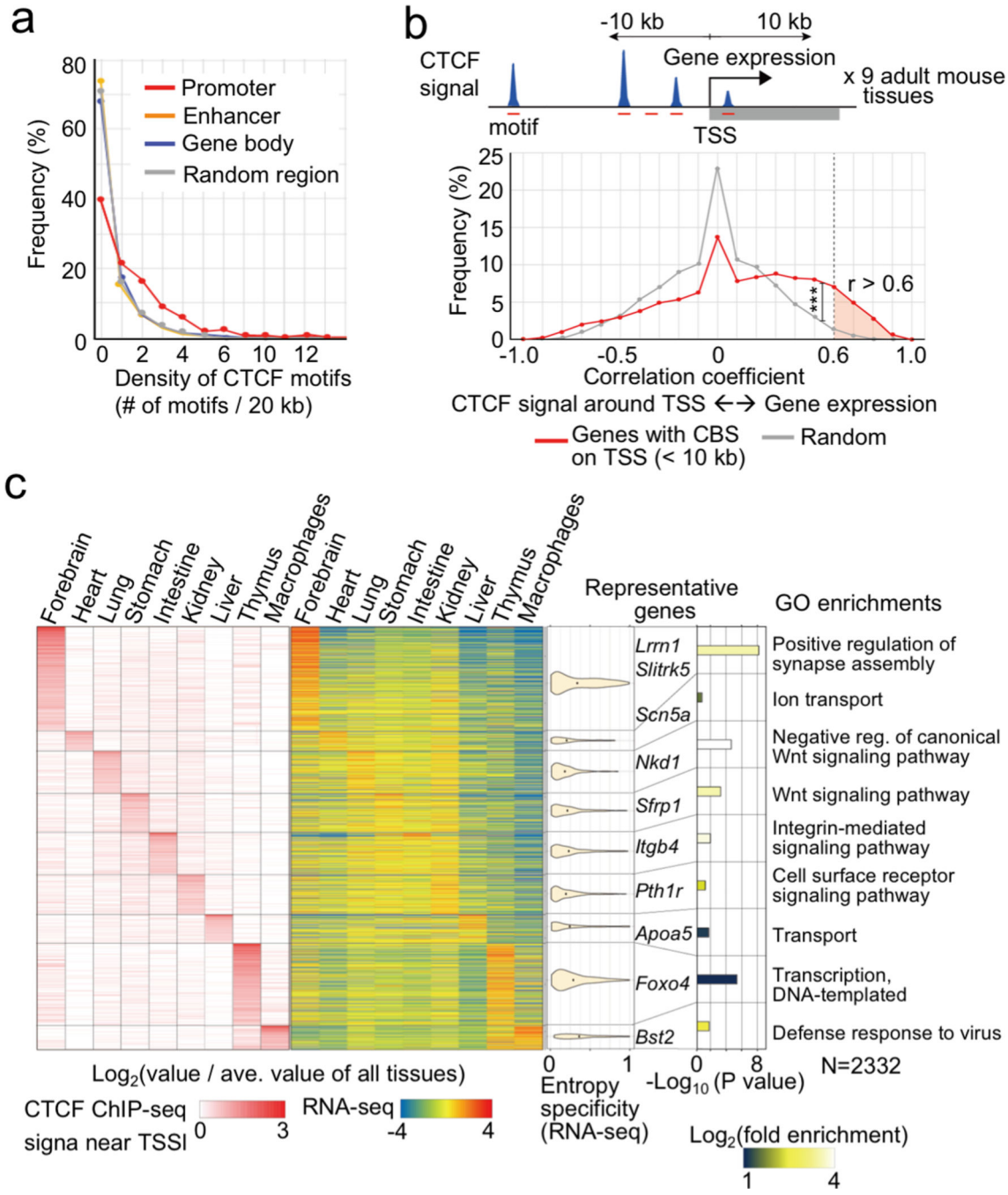


Fig. 5 |. Promoter-proximal CTCF binding correlates with transcription at thousands of mouse genes.

a, Frequency of genomic regions and their density of CTCF motifs bound by CTCF in at least one adult mouse tissue. The genomic regions were classified into promoters, enhancers (identified in ESCs and NPCs), gene bodies, and random regions.

b, Schematics of steps to compute the correlation between CTCF occupancy around TSS and transcripts levels across multiple mouse tissues (top, for details see Methods). Frequencies of genes are plotted based on the Pearson Correlation Coefficient between the

CTCF ChIP-seq signals around TSS (< 10 kb) and transcripts levels across multiple mouse tissues (red line). The same plots analyzed using randomly shuffled CTCF ChIP-seq datasets are shown in gray as control. *** p value < 0.001, Pearson's Chi-squared test for the comparison of fraction of genes with positive correlation coefficient ($r \geq 0.6$) or others ($r < 0.6$) between the two groups.

c, Heatmaps of lineage-specificity of CTCF ChIP-seq signals around promoters and gene expression levels. Genes with high correlation coefficients (> 0.6) in panel (b) are shown (2,332 genes). Lineage-specificity was calculated as $\log_2$(value / average value of all tissues). The violin plots also show the lineage-specificity of transcription measured by Shannon entropy in each gene group. The width is proportional to the sample size. Top enriched GO terms of each group genes are shown with fold enrichment, p-value (Fisher's exact test), and their representative genes.