



CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains

Hui Huang^{1,2}, Quan Zhu³, Adam Jussila^{1,4}, Yuanyuan Han³, Bogdan Bintu⁵, Colin Kern³, Mattia Conte⁶, Yanxiao Zhang¹, Simona Bianco⁶, Andrea M. Chiariello⁶, Miao Yu¹, Rong Hu¹, Melodi Tastemel⁷, Ivan Juric⁸, Ming Hu⁸, Mario Nicodemi^{6,9,10}, Xiaowei Zhuang¹⁰ and Bing Ren^{1,3,7,11} ✉

Insulators play a critical role in spatiotemporal gene regulation in animals. The evolutionarily conserved CCCTC-binding factor (CTCF) is required for insulator function in mammals, but not all of its binding sites act as insulators. Here we explore the sequence requirements of CTCF-mediated transcriptional insulation using a sensitive insulator reporter in mouse embryonic stem cells. We find that insulation potency depends on the number of CTCF-binding sites in tandem. Furthermore, CTCF-mediated insulation is dependent on upstream flanking sequences at its binding sites. CTCF-binding sites at topologically associating domain boundaries are more likely to function as insulators than those outside topologically associating domain boundaries, independently of binding strength. We demonstrate that insulators form local chromatin domain boundaries and weaken enhancer-promoter contacts. Taken together, our results provide genetic, molecular and structural evidence connecting chromatin topology to the action of insulators in the mammalian genome.

The spatial and temporal patterns of gene expression are encoded in the genome in the form of *cis*-regulatory elements, which are categorized into promoters, enhancers, insulators and other less studied regulatory sequences, including repressive/silencing elements^{1–3}. In animals, insulators play an essential role in cell-type-specific gene expression by protecting genes from improper regulatory signals from the neighboring chromatin environment⁴. Enhancer-blocking insulators act in a position-dependent manner in that they prevent enhancer-dependent gene activation only when placed between the enhancer and target gene^{5–7}. Insulators were initially identified in *Drosophila*, where the molecular machinery for insulation was first elucidated^{4,5,8}. The first identified enhancer-blocking insulator in vertebrates is the 5'-HS4 element of the chicken β -globin locus⁹. Detailed analysis of this insulator led to the finding that the evolutionarily conserved zinc-finger-family transcription factor CTCF, first identified as a DNA-binding protein at the chicken *c-Myc* gene promoter¹⁰, was essential for its enhancer-blocking activity¹¹. Alterations in the CTCF protein or its binding sites at insulators have since been implicated in a broad spectrum of human diseases^{12–14}. In addition to its function at insulators, CTCF has also been demonstrated to play roles in transcriptional repression, gene activation, alternative splicing and class switch recombination depending on the context of genomic locus^{10,15–19}. There are reports that CTCF binding at gene promoters could promote, instead of block, enhancer-promoter

interactions^{20,21}. To date, exactly how and where CTCF mediates insulator function remains unclear.

CTCF has long been postulated to function as an organizer of three-dimensional chromosome architecture^{1,22,23}. Genome-wide chromosome conformation capture (3C) analyses showed that the interphase chromosomes in mammalian cells are partitioned into megabase-sized topologically associating domains (TADs)^{24,25}, and CTCF-binding sites (CBSs) were found at over 75% of TAD boundaries²⁴, suggesting a probable link between TAD boundaries and CTCF-mediated transcriptional insulation. Supporting this connection, disruption of TAD boundaries has been shown to permit ectopic enhancer-promoter contacts and aberrant gene expression, thereby leading to developmental abnormalities and cancer^{16,26}. Additionally, depletion of CTCF can lead to the weakening or disappearance of TADs^{27–29}. CTCF drives TAD formation by working together with the cohesin complex to establish dynamic chromatin loops between distant CBSs, likely through a loop-extrusion process^{29–39} or other mechanisms such as phase separation^{40–45}. However, it is still debated whether TAD boundaries are sufficient to provide transcriptional insulation. Rapidly dissolving the global TAD structure by acute depletion of CTCF or cohesin subunits altered transcription of only a small number of genes in many different cellular contexts^{27,29,33,35,37,46}. Moreover, deletion of CTCF sites at the TAD boundary of the developmental locus *Sox9-Kcnj2* did not cause discernible phenotypes⁴⁷. Furthermore, a majority of

¹Ludwig Institute for Cancer Research, La Jolla, CA, USA. ²Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, CA, USA. ³Center for Epigenomics, Department of Cellular and Molecular Medicine, School of Medicine, University of California, San Diego, La Jolla, CA, USA. ⁴Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA, USA. ⁵Howard Hughes Medical Institute, Department of Chemistry and Chemical Biology and Department of Physics, Harvard University, Cambridge, MA, USA. ⁶Dipartimento di Fisica, Università di Napoli Federico II, and INFN Napoli, Complesso di Monte Sant'Angelo, Naples, Italy. ⁷Department of Cellular and Molecular Medicine, School of Medicine, University of California, San Diego, La Jolla, CA, USA. ⁸Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH, USA. ⁹Berlin Institute for Medical Systems Biology, Max Delbrück Centre (MDC) for Molecular Medicine, Berlin, Germany. ¹⁰Berlin Institute of Health (BIH), Berlin, Germany. ¹¹Institute of Genomic Medicine and Moores Cancer Center, School of Medicine, University of California, San Diego, La Jolla, CA, USA. ✉e-mail: biren@health.ucsd.edu

CBSs are not located at TAD boundaries, and whether these CTCF sites may function as insulators is unclear. These observations warrant an in-depth investigation of the role that CTCF and TADs play in transcriptional insulation.

To better understand where and how CTCF may mediate transcriptional insulation in the genome, we have developed an insulator reporter assay to evaluate the function of any DNA fragments in blocking enhancer-dependent transcriptional activation in mouse embryonic stem cells (mESCs). Using this system, we demonstrated that isolated single CTCF sites have weak or no insulator activity, regardless of its DNA-binding strength. Instead, multiple copies of CTCF sites placed in tandem can provide a potent insulation effect. We also observed that CBSs at TAD boundaries could function as potent insulators, while the CTCF sites not located at TAD boundaries were incapable of insulating transcription. We attributed this difference in insulation activity to a sequence located 10–20 base pairs (bp) upstream of the CTCF core motifs, which promotes optimal insulation likely through contacts with CTCF's zinc-fingers 9–11. We further discovered that insulators act by forming local TAD boundaries to reduce productive enhancer–promoter contacts, using both 3C-based assays and high-throughput multiplexed DNA fluorescence in situ hybridization (FISH) techniques. These results, taken together, shed light on how CTCF mediates transcriptional insulation in mammalian cells and establish a direct link between TAD boundaries and insulators.

Results

An insulator reporter assay in mESCs. To quantitatively assay insulator activities in the context of native chromatin in cells, we engineered the *Sox2* gene locus in the F123 hybrid mESC line (*Mus musculus castaneus* × S129/SvJae)⁴⁸. We and others previously showed that a super-enhancer located ~110 kilobases (kb) downstream of the *Sox2* gene was responsible for more than 90% of its expression in the mESCs^{49,50}. We reasoned that the insulator activity of DNA elements could be measured by the reduction in *Sox2* gene expression when inserted between the *Sox2* gene and the downstream super-enhancer. To create the insulator reporter, we first tagged the two copies of the *Sox2* gene with *egfp* (CAST allele) and *mcherry* (129 allele) to quantify allelic *Sox2* expression by live-cell fluorescence-activated cell sorting (FACS; Fig. 1a and Extended Data Fig. 1a). Subsequently, we inserted a negative-selection fusion gene pCAG-*HyTK* flanked by a pair of heterotypic flippase recognition sites (FRT/F3) between the *Sox2* gene and its downstream super-enhancer on the CAST allele (Fig. 1a and Extended Data Fig. 1b). As enhancer-blocking insulation is position dependent, we created a control clone with the same replaceable cassette placed further downstream of the *Sox2* super-enhancer at equal distance on the CAST allele (Fig. 1a and Extended Data Fig. 1c). The pCAG-*HyTK* marker gene can be replaced by a donor sequence using the recombinase-mediated cassette exchange (RMCE) strategy (Fig. 1b and Supplementary Fig. 1a). By killing off unmodified mESCs with ganciclovir, we could achieve nearly 100% efficiency of marker-free insertion (Supplementary Fig. 1b).

As the insertion was specifically on the CAST allele, we used the 129 allele as the internal control to correct clone-to-clone variations in *Sox2* expression (Fig. 1b and Supplementary Fig. 2a,b), which allowed quantitative comparisons of insulator activities of different CBSs. We tested the insulation activity of a total of 11 different CBSs selected from several known TAD boundaries and chromatin loop anchors (Supplementary Table 1). Each CBS insert was amplified from mouse or human genomic DNA by PCR and was 1–4 kb in length. Surprisingly, isolated single CBSs tested in both the forward and reverse orientations generally exhibited little or no insulator effect (Fig. 1c). Only two of the probed CBSs in the reverse orientation and four of the probed CBSs in the forward orientation showed significant yet modest insulator effects (Fig. 1c). The CBS

of a canonical insulator, the HS5 sequence of the human β -globin locus, reduced *Sox2* expression by $11.0\% \pm 1.9\%$ when inserted in the forward orientation but had no effect in the reverse orientation (Fig. 1c and Supplementary Fig. 2c,d). On average, individual isolated CBSs in the forward and reverse orientations reduced *Sox2* expression to $93.0\% (\pm 6.5\%)$ and $97.0\% (\pm 6.0\%)$ of parental cells with no insertion, respectively (Fig. 1c).

Tandem CTCF sites enable strong transcriptional insulation.

We hypothesized that multiple CBSs collectively may provide more robust insulation, as TAD boundaries are enriched for clustered CBSs^{24,51}. To test this possibility, we constructed a series of insertion clones harboring multiple CBSs from the *Sox9–Kcnj2* TAD boundary (Extended Data Fig. 2a). Two or more CBSs were PCR amplified from mouse genomic DNA, ligated together and inserted between the *Sox2* gene and super-enhancer on the CAST allele by RMCE as described above. We found that two CBSs, in forward tandem, reverse tandem or divergent orientations, all had a significantly stronger insulation effect than individual CBSs alone (Fig. 2a). Notably, combining a weak CBS insulator with one that had a negligible insulator activity gave rise to stronger insulation than the summed effects of the two individual sites (Fig. 2a), suggesting that CBSs could have synergistic insulation effects. Nevertheless, a weak CBS insulator did not enhance the insulator activity of a stronger CBS insulator if placed in convergent orientation (Fig. 2a). Next, we measured the insulator activity of CBS clusters consisting of up to all four CBSs from the *Sox9–Kcnj2* TAD boundary. Analyses using chromatin immunoprecipitation followed by sequencing (ChIP-seq) indicated that CTCF was recruited to the extra copy of the boundary sequence inserted in the *Sox2* domain (Extended Data Fig. 2b). We found that the insulation effect became stronger as the number of CBSs increased, regardless of the orientation of the CTCF motifs (Fig. 2b and Supplementary Table 2). Interestingly, the enhancement of insulation conferred by each additional CBS became smaller when the number of CBSs exceeds two (Extended Data Fig. 2c). Consistent with the requirement for CTCF in transcriptional insulation, removal of the binding motifs of CTCF within the inserts completely abolished the insulation effects of CBSs (Fig. 2c). Furthermore, introducing CTCF sites downstream of the *Sox2* super-enhancer did not reduce but rather slightly increased *Sox2* expression (Fig. 2b), likely due to the insulation of interactions between the super-enhancer and further downstream chromatin. Taken together, these results suggest that multiple CBSs arranged in tandem can function as a potent insulator due to synergistic or additive effects from individual sites.

Surprisingly, we observed that the insulator containing four CBSs was able to reduce *Sox2* expression by $38.47 \pm 3.16\%$, rather than completely blocking the *Sox2* super-enhancer activity (Fig. 2b). The reduction of *Sox2* expression from the CAST allele was further confirmed by allele-sensitive RNA-seq analysis (Extended Data Fig. 2d,e). Interestingly, this insulator substantially increased cell-to-cell variations in *Sox2* expression, evidenced by the accumulation of cells with extremely low *Sox2*–eGFP signals (Extended Data Fig. 2f). Moreover, the subpopulation of cells expressing ultralow *Sox2*–eGFP could revert to the state of higher expression level after extended culturing, suggesting that the cell-to-cell variation of *Sox2* gene expression was a metastable state (Extended Data Fig. 2g). Furthermore, CTCF insulation did not change the active chromatin state on either the *Sox2* promoter or its enhancer (Extended Data Fig. 2h,i). Collectively, these results suggest that CBS-mediated insulation is permissive and highly dynamic.

CTCF-mediated insulation depends on sequence context. To better understand the sequence requirements for CTCF-mediated insulation, we synthesized insulators by concatenating multiple 139-bp genomic DNA sequences, each containing a 19-bp CTCF motif

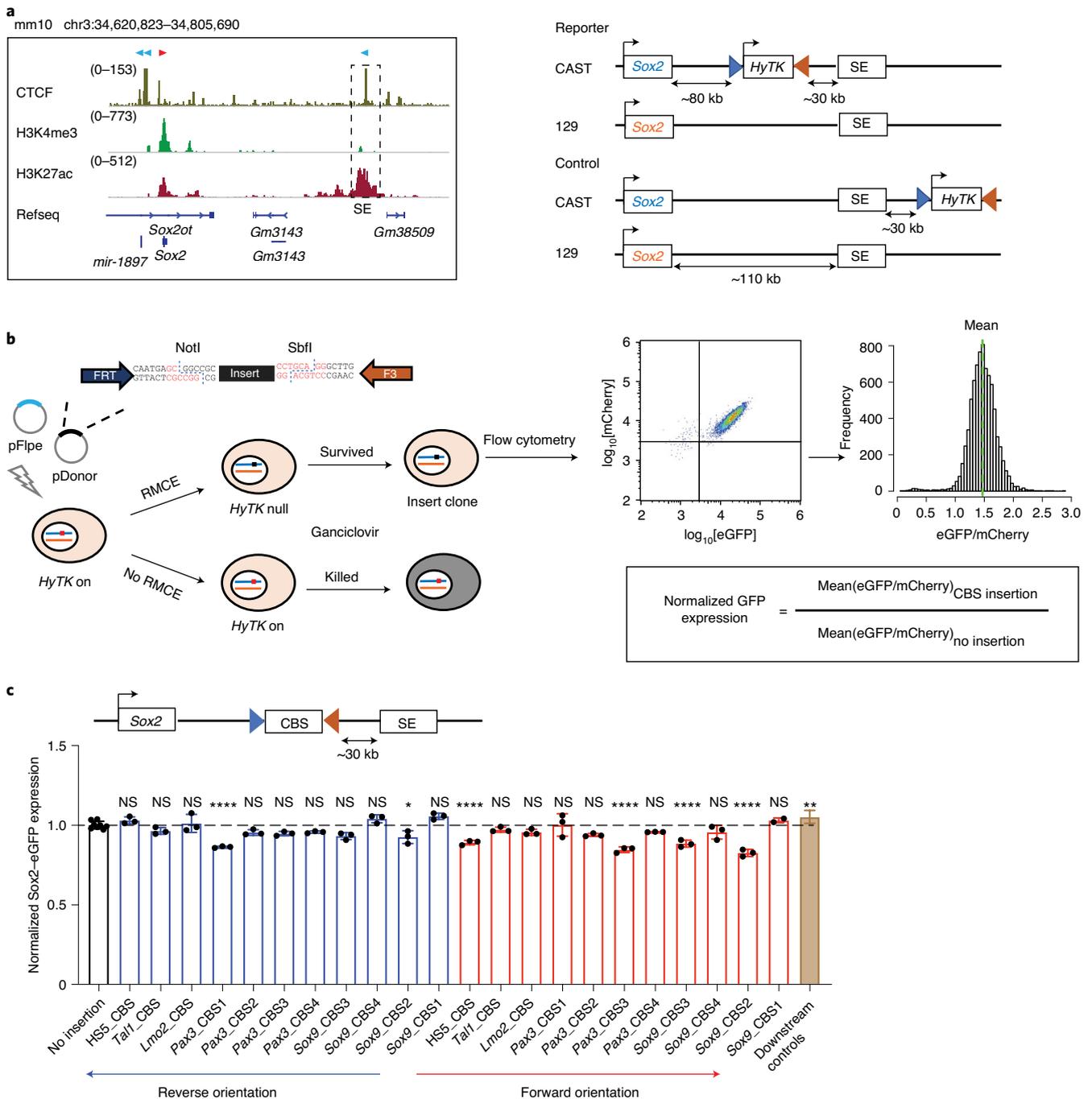


Fig. 1 | A sensitive insulator reporter assay measures the insulation activity of different CBSs at the *Sox2* locus in mESCs. **a, Left, the regulatory landscape of the *Sox2* locus in mESCs. The orientations of the CTCF sites are indicated on the top of the signal tracks (normalized read counts). Right, genetic constructs of mESC lines. Boxed *Sox2* in blue represents the *Sox2-p2a-egfp* in situ fusion gene; boxed *Sox2* in orange represents the *Sox2-p2a-mcherry* in situ fusion gene. The hygromycin phosphotransferase-thymidine kinase fusion gene *HyTK* is flanked by the flippase recognition sites FRT and F3. SE, super-enhancer. **b**, Experimental scheme to insert a test sequence into the *Sox2* locus by RMCE. The flippase expression plasmid and donor plasmid containing the test sequence were coelectroporated into cells. The orientation of the insert was controlled by the positions of the NotI and SbfI restriction enzyme sites. mESC clones containing the insert were picked and genotyped, and allelic *Sox2* expression was measured by FACS. **c**, A bar graph showing the normalized *Sox2*-eGFP expression of the no insertion clone ($n=8$), different CBS insertion clones ($n=3$; for *Sox9_CBS1* in the forward orientation, $n=2$) and downstream insertion controls ($n=27$). Each dot represents an independently picked colony. The P values were determined by one-way analysis of variance with Bonferroni's multiple comparisons test. Data are mean \pm s.d. The exact P values for each comparison are listed in Supplementary Table 9. NS, $P > 0.05$; * $P \leq 0.05$; ** $P \leq 0.01$; **** $P \leq 0.0001$.**

and two 60-bp flanking sequences. Each site was selected from the aforementioned CBSs (Supplementary Tables 3 and 4). Consistent with the observations described above, the synthetic DNA sequences

showed additive effects in transcriptional insulation (Extended Data Fig. 3a). Additionally, ChIP-seq analyses confirmed the recruitment of CTCF and the cohesin complex to the synthetic insulators

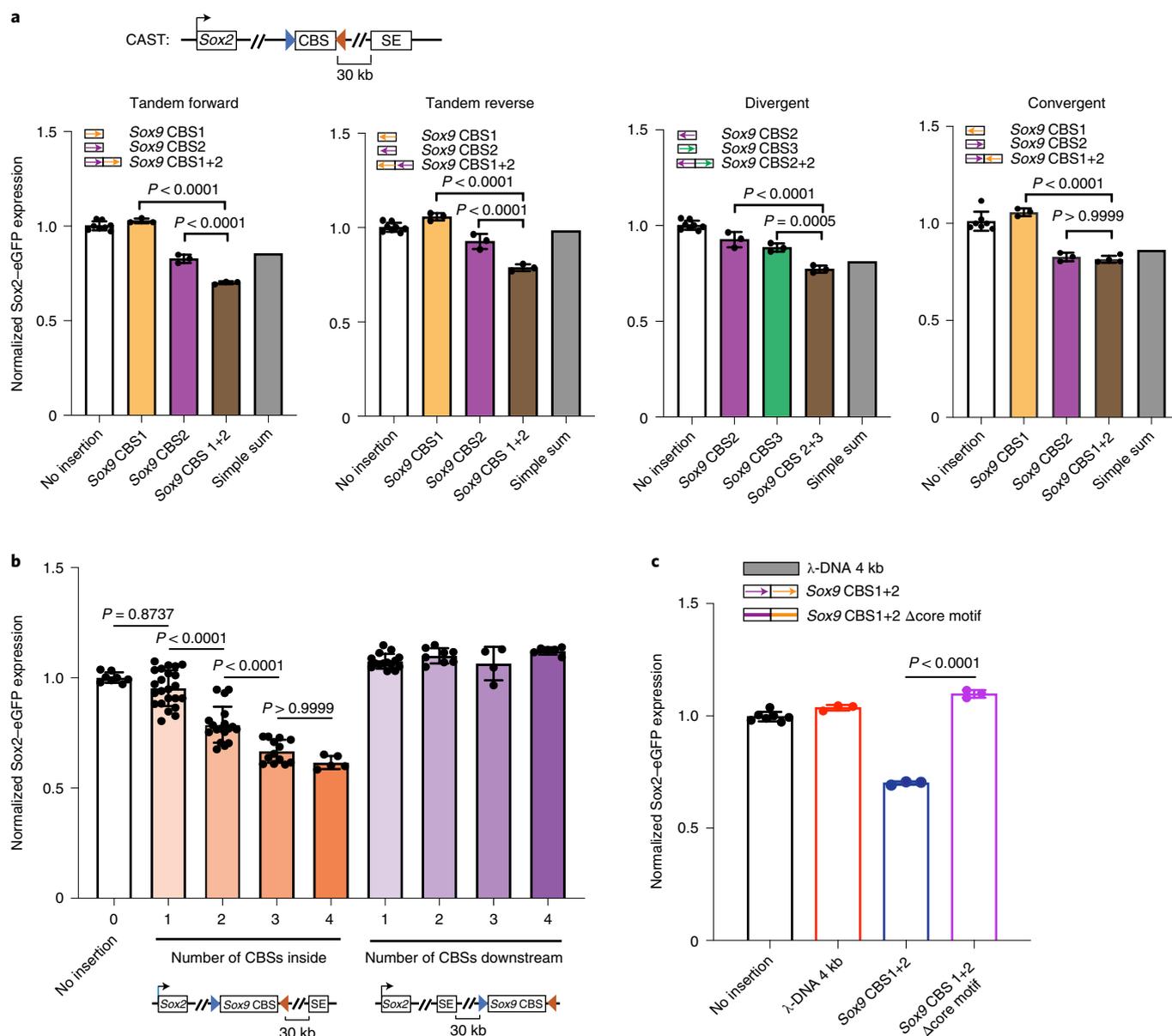


Fig. 2 | Multiple CTCF sites in tandem enable strong transcriptional insulation. a, Bar graphs showing insulation effects of two combined CBSs from the *Sox9-Kcnj2* TAD boundary (no insertion $n=8$, for convergent group, $n=7$; insertion clones $n=3$, for Sox9 CBS1+2 in convergent group $n=4$). Individual CBS sequences were combined by PCR to create two-CBS insertions. Arrows indicate the motif orientation of each individual CBS. Every insertion construct was created by an independent RMCE experiment. **b**, A bar graph showing insulation effects of multiple CBSs from the *Sox9-Kcnj2* TAD boundary. Individual or combined CBS sequences were PCR cloned from mouse genomic DNA. Every insertion construct was created by an independent RMCE experiment (0 CBS, $n=8$; 1 CBS inside, $n=23$; 2 CBSs inside, $n=18$; 3 CBSs inside, $n=13$; 4 CBSs inside, $n=5$; 1 CBS downstream, $n=15$; 2 CBSs downstream, $n=8$; 3 CBSs downstream, $n=4$; 4 CBSs downstream, $n=6$). **c**, A bar graph showing insulation effects of λ -DNA ($n=3$), a combined two-CBS sequence, Sox9 CBS1+2 ($n=3$), and Sox9 CBS1+2 Δ core motifs, which is the same two-CBS sequence but with the two 19-bp CTCF core motifs deleted ($n=3$). Inserts were comparable in length (~ 4 kb). Data are mean \pm s.d. The P values were determined by one-way analysis of variance with Bonferroni's multiple comparisons test.

(Fig. 3a). Interestingly, we observed that CBSs with longer flanking sequences (1 kb or longer) had stronger insulation effects than the shorter 139-bp CBSs, suggesting the existence of additional elements that could facilitate insulation (Extended Data Fig. 3b).

Using the same approach, we also tested whether CBSs from outside TAD boundaries could function as insulators. We selected multiple CBSs from non-TAD boundary regions in the genome, concatenated multiple 139-bp genomic sequences containing CTCF-binding motifs together and tested their insulation ability in

our insulator reporter assay (Supplementary Table 4). Surprisingly, although these non-TAD boundary CBSs displayed stronger CTCF binding than those from TAD boundaries at their original loci (Extended Data Fig. 3c), the synthetic DNA sequences made up of 6 or 15 tandemly arrayed 139-bp CBSs from non-boundary regions were unable to function as insulators, despite the presence of strong CTCF ChIP-seq signals at the insertion site (Fig. 3b and Extended Data Fig. 3d), indicating that CTCF binding alone is insufficient to result in transcriptional insulation.

To further dissect the sequence dependence of CTCF-mediated insulation, we exchanged the core motifs of 139-bp boundary CBSs with those of the synthetic CBSs from non-boundary regions (Supplementary Table 4). Combining boundary CBS core motifs with non-boundary adjacent sequences resulted in a much weaker insulation effect than with their original neighboring sequences of equal lengths (Fig. 3c). In contrast, replacing adjacent sequences of non-boundary CBSs with those from boundary sites significantly strengthened their insulation effect (Fig. 3c). However, when the adjacent sequences were scrambled or kept the same for boundary and non-boundary core motifs, their effects in insulating *Sox2* expression were comparable (Fig. 3c). Together, these results suggest that transcriptional insulation by CTCF is sequence-context dependent, requiring DNA elements flanking the CTCF-binding motif. It should be noted that ChIP-seq analysis showed that differential insulation activity of the synthetic insulators is not strictly correlated with CTCF occupancy (Extended Data Fig. 4a–d).

To further delineate the key element within the flanking sequences that promote transcriptional insulation, we tested the insulator activity of a series of synthetic CBSs with gradually decreasing flanking sequences from each side. Interestingly, strong insulation was retained at a synthetic insulator with just 20-bp flanking sequences on both sides of the core CTCF-binding motifs, but significantly reduced when the flanking sequences were shortened to 10 bp (Fig. 3d), suggesting a critical role for the 10–20-bp flanking sequences of the core CTCF-binding motif in insulation. We used the GLAM2 tool⁵², a multiple sequence aligner that allows gaps and deletions among motifs, to identify a composite element in the six boundary CBSs (Extended Data Fig. 4e). We found a central motif that matches the CTCF core motif and an upstream motif at the same location as a previously reported element recognized by CTCF zinc-fingers 9–11^{53–55} (Fig. 3e). To test whether CTCF zinc-fingers 9–11 indeed contribute to transcriptional insulation, we deleted the DNA segment coding for zinc-fingers 9–11 from both copies of the endogenous CTCF gene using CRISPR editing tools as previously described³⁷ (Extended Data Fig. 5a–c). Deletion of CTCF zinc-fingers 9–11 significantly weakened insulation of the boundary CBSs but did not further reduce the insulation strength of the synthetic insulator with just 10-bp flanking sequences (Fig. 3f and Extended Data Fig. 5d,e). Together, these results suggest that flanking sequences of the boundary CBSs promote CTCF-mediated transcriptional insulation likely through contacts with CTCF zinc-fingers 9–11. Further, ChIP-seq analysis showed that CTCF binding to CBSs with just 10-bp flanking sequences did not decrease notably (Extended Data Fig. 4a,b).

Insulators form TADs and weaken enhancer-promoter contacts. Previous studies suggest that the *Sox2* super-enhancer forms long-range chromatin contacts with the *Sox2* promoter^{50,56}. We

hypothesized that insulators may change chromosome topology to limit enhancer–promoter communication. To test this hypothesis, we performed proximity ligation ChIP-seq (PLAC-seq⁵⁷; also known as HiChIP⁵⁸) experiments using mESC clones with various insulators inserted at the *Sox2* locus to detect promoter-centered chromatin contacts. Contact frequencies between the *Sox2* promoter and downstream super-enhancer were similar between the CAST and 129 alleles in mESCs with no insertion (Fig. 4a). Inserting two CBSs from the *Sox9–Kcnj2* TAD boundary between the *Sox2* promoter and super-enhancer reduced the enhancer–promoter contacts significantly (Fig. 4a). Consistent with the observed dosage-dependent insulation effects, the *Sox2* enhancer–promoter contacts on the CAST allele were further reduced in cells with the insertion of four CBSs (Fig. 4a). By contrast, placing two or four CBSs downstream of the *Sox2* super-enhancer did not reduce the *Sox2* enhancer–promoter contacts (Fig. 4a). These results support the model that insulators act by reducing the enhancer–promoter contacts.

To further understand the effect of the insulators on local chromatin structure, we performed in situ Hi-C experiments⁵⁹ with mESC clones containing either two or four CBSs inserted between the *Sox2* gene and its super-enhancer on the CAST allele (Fig. 4b,c). On the 129 allele, the *Sox2* promoter and downstream super-enhancer were found to be in a single TAD (Fig. 4b). By contrast, insertion of two CBSs between the *Sox2* gene and super-enhancer on the CAST allele created a new TAD boundary that separated the *Sox2* locus into two local chromatin domains (Fig. 4b). Introducing four CBSs in the same location created an even stronger TAD boundary, and contacts across the new local domains were further reduced (Fig. 4c). Additionally, we found that the inserted CBSs showed elevated levels of chromatin contacts with the CBSs located on the *Sox2* promoter and super-enhancer, following the convergent rule⁵⁹ (Extended Data Fig. 6a–d). Collectively, these results suggest that CTCF-dependent insulators create local TAD domains by forming chromatin loops between convergent CBSs.

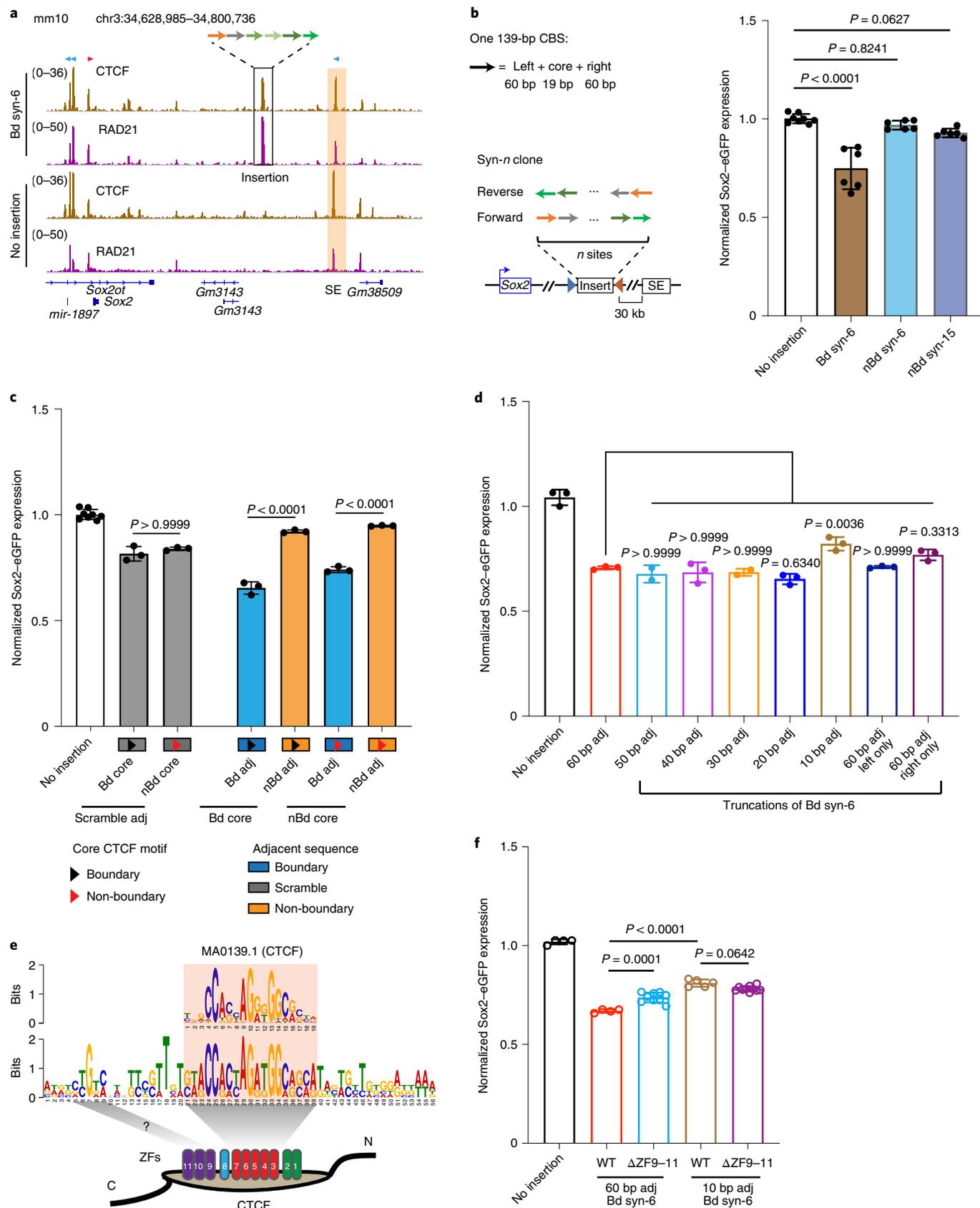
Visualizing *Sox2* locus by multiplexed FISH for DNA and RNA.

To directly visualize the impacts of insulators on chromatin architecture, we used the recently developed multiplexed DNA FISH imaging method to trace the chromatin conformation^{60–62}. We traced the three-dimensional structure of the 210-kb genomic region (chr3:34,601,078–34,811,078) containing the *Sox2* and super-enhancer loci across thousands of individual chromosomes at 5-kb intervals. We partitioned the 210-kb region into forty-two 5-kb segments and sequentially labeled and imaged each segment using 14 rounds of hybridization of readout probes with a three-color imaging scheme (Fig. 5a, Extended Data Fig. 7a–c and Supplementary Tables 5 and 6). The identity of the CAST allele was determined within each nucleus based on the presence of FISH

Fig. 3 | Synthetic insulators reveal sequence requirements for CTCF-mediated enhancer-blocking activity. **a**, ChIP-seq signal tracks of CTCF and RAD21 (normalized read counts). The Bd syn-6 mESC clone contains the insertion of six 139-bp boundary (Bd) CBSs (*Sox9_CBS1-4*, *Pax3_CBS3* and *HS5_CBS*; Supplementary Tables 3 and 4) between *Sox2* and its super-enhancer. Sequencing reads from the insertion clone were aligned to a customized mm10 genome that included the inserted sequence at the target location. Motif orientations of nearby and inserted CBSs are indicated on the top of the signal tracks. The *Sox2* super-enhancer is highlighted in the orange box. **b**, A bar plot showing insulation effects of synthetic sequences containing tandemly arrayed 139-bp CBSs from boundary and non-boundary (nBd) regions. For each synthetic sequence, six insertion clones were picked with three of them in the forward orientation and the other three in the reverse orientation. **c**, A bar plot showing insulation effects of recombined tandemly arrayed 139-bp CBSs. Core motifs of boundary and non-boundary CBSs were combined with their native adjacent sequences, scrambled adjacent sequences or exchanged adjacent sequences with each other ($n=3$). Each test sequence contains six tandemly arrayed 139-bp CBSs. The order of the six CBS core motifs was kept the same. **d**, A bar plot showing the insulation effect of the Bd syn-6 sequence with truncated adjacent sequences. All insertions were between the *Sox2* promoter and super-enhancer ($n=3$; for 50 bp adj, $n=2$). **e**, A composite motif discovered in the six boundary CBSs tested. Each CBS consists of a 19-bp core motif and 20-bp adjacent sequences on both sides. The motif was searched by the GLAM2 program of the MEME suite. ZFs, zinc-fingers. **f**, A bar plot showing the impact of CTCF zinc-finger 9–11 deletion on the insulation effects of boundary CBSs containing 60-bp adjacent sequences and 10-bp adjacent sequences (no insertion, $n=4$; for WT with 60 bp adj, $n=4$; for $\Delta ZF9-11$ with 60 bp adj, $n=9$; for WT with 10 bp adj, $n=5$; for $\Delta ZF9-11$ with 10 bp adj, $n=10$). The P values were determined by one-way analysis of variance with Bonferroni's multiple comparisons test. Data are mean \pm s.d.

signal corresponding to the 7.5-kb 4CBS insulator sequence inserted into the CAST allele that was absent in the 129 allele (Fig. 5a and Extended Data Fig. 7d).

We first carried out chromatin tracing experiments with the mESC clone containing an insertion of the 4CBS insulator between the *Sox2* gene and the downstream super-enhancer on the CAST



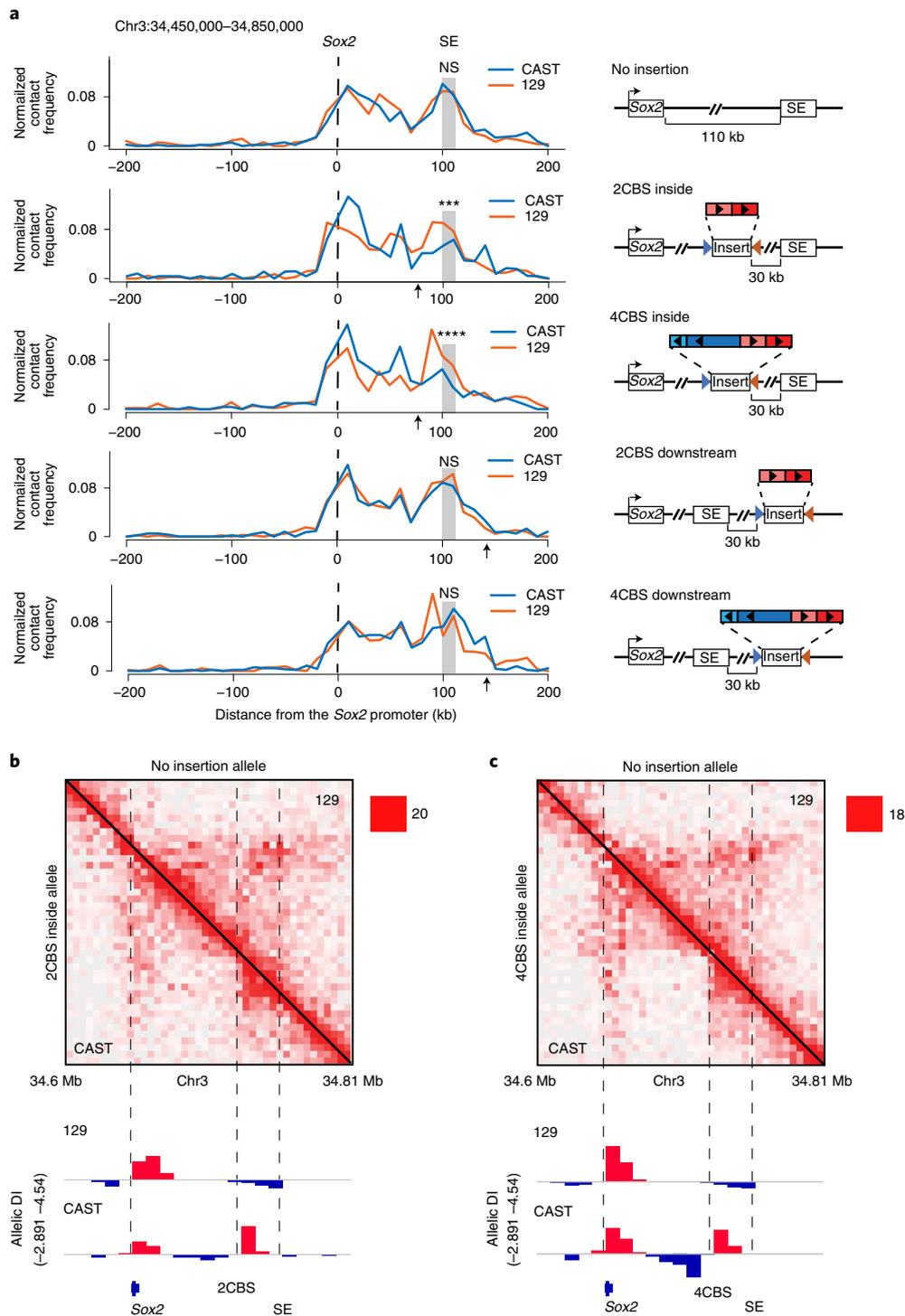


Fig. 4 | Enhancer-blocking insulator forms local chromatin domains and reduces *Sox2* enhancer–promoter chromatin contacts. **a**, Allelic chromatin contacts from PLAC-seq data are shown at the viewpoint of the *Sox2* promoter ($n=2$, replicates were merged). PLAC-seq experiments were carried out using a monoclonal antibody (Millipore, O4-745) to H3K4me3. Sequencing reads were mapped to the mm10 reference genome and split into CAST and 129 alleles based on the haplotypes of the parental strains. DNA fragments connecting the promoter and each of the surrounding 10-kb bins were counted. Contact frequency was normalized by the total *cis* contacts of the *Sox2* promoter for each allele; interactions within the 10-kb *Sox2* promoter bin are not shown. Arrows indicate the insertion location of CBSs. Fisher exact tests of *Sox2* enhancer–promoter contacts of the two alleles were performed (two-sided tests; NS, $P > 0.05$; *** $P = 4.91 \times 10^{-4}$; **** $P = 5.34 \times 10^{-5}$). Right, insertion construct matching each clone on the left. The CBS clusters were obtained from the *Sox9–Kcnj2* TAD boundary by PCR. **b,c**, Allelic Hi-C contact map at the *Sox2* locus. mESCs with the insertion of two CBSs or four CBSs from the *Sox9–Kcnj2* TAD boundary in the CAST allele were used for the experiments. Hi-C reads were mapped to the mm10 reference genome and split into CAST and 129 alleles based on the haplotypes of the parental strains. The allele-specific contact matrix was normalized by Knight–Ruiz matrix balancing. Top right, no insertion allele (129). Bottom left, insertion allele from the same cells (CAST). Bottom, allelic directionality index (DI) score of Hi-C interaction frequency ($n=2$, replicates were merged).

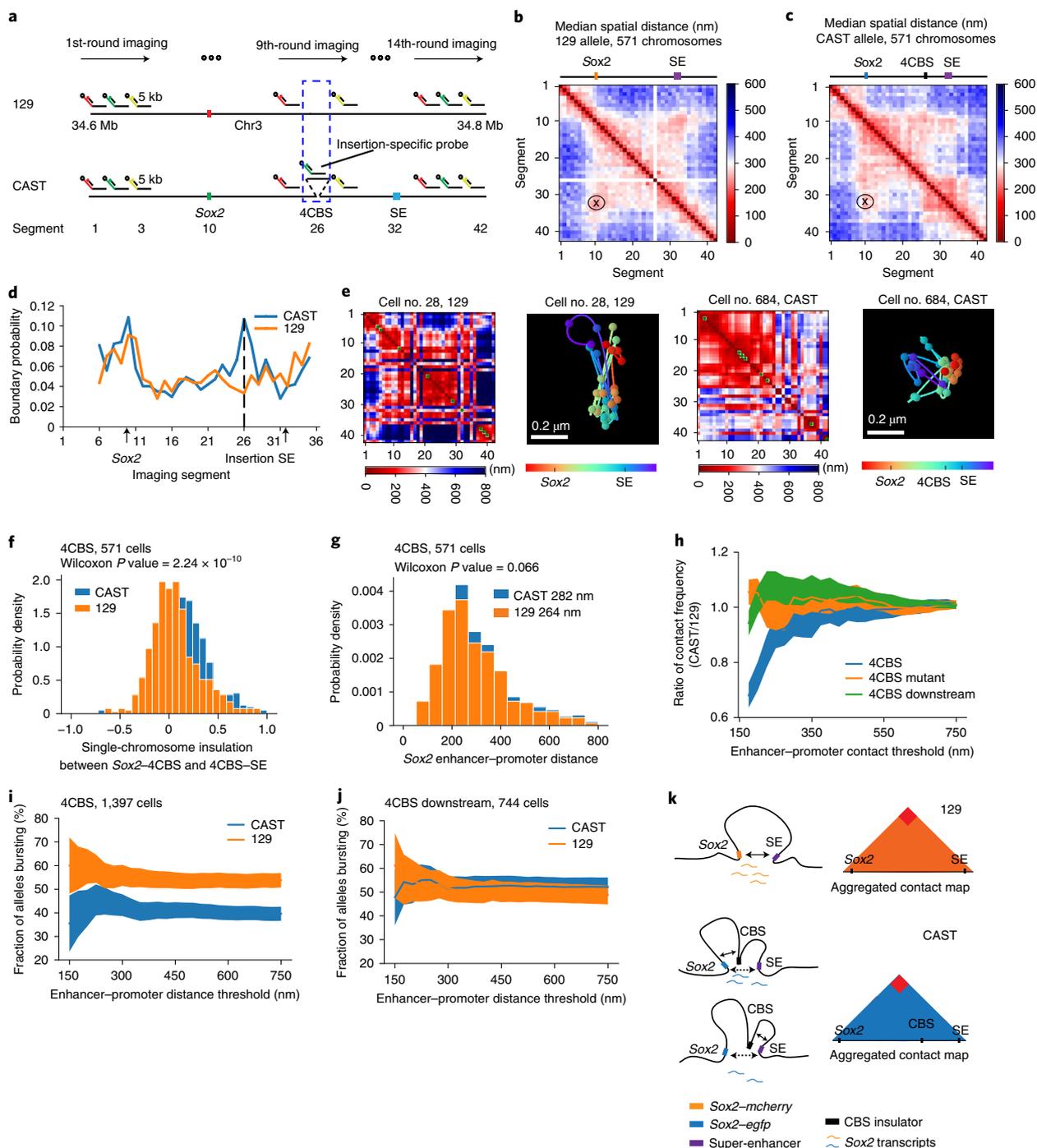


Fig. 5 | Effects of an enhancer-blocking insulator on chromatin topology and transcription revealed by multiplexed FISH. **a**, Scheme of the chromatin tracing experiments targeting the 210-kb *Sox2* region (chr3:34,601,078–34,811,078). **b,c**, Median-spatial-distance (nm) between imaged segments in **(a)** for the 129 **(b)** and CAST **(c)** chromosomes. **d**, The probability of each segment to be a single-chromosome domain boundary for the two alleles in **b,c**. The 26th segment on the CAST allele is the 4CBS insertion. **e**, Exemplary single-chromosome structures of the imaged *Sox2* locus (segments are pseudo-colored) of the CAST and 129 alleles. Green pixels on the interpolated matrices indicate missing values in the displayed examples of chromatin traces. **f**, The distribution of single-chromosome insulation scores for each of the alleles between the *Sox2* promoter–4CBS insertion (segments 10–25) and 4CBS insertion–*Sox2* enhancer (segments 26–33) regions. A two-sided Wilcoxon rank-sum test was performed. **g**, The distribution of *Sox2* enhancer–promoter distance for the CAST and 129 chromosomes in **b,c**. A two-sided Wilcoxon rank-sum test was performed. **h**, The ratio of *Sox2* enhancer–promoter contact frequency of CAST chromosomes to that of 129 chromosomes. The distribution of contact frequency ratio (CAST/129) of the 4CBS ($n=571$ cells) clone is significantly different from that of the 4CBS mutant ($n=659$ cells) and 4CBS downstream ($n=784$ cells) clones, with P values of two-sided Kolmogorov–Smirnov tests equal to 6.38×10^{-5} and 1.09×10^{-9} , respectively. Shadow indicates the 95% confidence interval based on binomial distribution. **i,j**, The bursting frequency of the *Sox2* gene on the CAST and 129 chromosomes. **i**, The 4CBS clone ($n=1,397$ cells); **j**, the control clone with 4CBS inserted downstream of the *Sox2* super-enhancer ($n=744$ cells). Shadow indicates the 95% confidence interval based on binomial distribution. **k**, A model of the *Sox2* locus on the two alleles. On the 129 allele, the super-enhancer interacts with the *Sox2* promoter and activates transcription of the *Sox2* gene. On the CAST allele, the CBS insulators can interact with both the *Sox2* promoter and the super-enhancer, resulting in fewer productive enhancer–promoter contacts.

allele. We obtained chromatin tracing data from 571 cells where both CAST and 129 alleles were robustly discerned (Methods). Consistent with results from Hi-C (Fig. 4c), the median-spatial-distance matrix for the 129 allele showed a single TAD harboring both the *Sox2* and super-enhancer loci, whereas the spatial distance matrix for the CAST allele showed two TADs with a new boundary formed at the insertion site separating the *Sox2* and super-enhancer loci (Fig. 5b,c and Extended Data Fig. 8a–c). Accordingly, individual CAST chromosomes were more likely to form a boundary at the 4CBS insertion (Fig. 5d,e). Moreover, the level of insulation between the two sub-regions to either side of the inserted 4CBS, containing the *Sox2* promoter and the super-enhancer, was statistically significantly enhanced on the CAST alleles (Fig. 5f).

As controls, we also performed chromatin tracing experiments with one mESC line where all CTCF-binding motifs of the insertion were removed, and another cell line where the insertion was at an equal distance further downstream of the *Sox2* super-enhancer. We obtained chromatin tracing data on both CAST and 129 alleles from 659 and 784 cells of the two cell lines, respectively. Based on FACS analyses, neither control insert reduced *Sox2* expression on the CAST allele (Extended Data Fig. 8d). Consistently, no local chromatin domain boundary was visible between the *Sox2* and super-enhancer loci, and spatial insulation between the *Sox2* gene and the super-enhancer was indistinguishable between the CAST and 129 alleles (Extended Data Fig. 8e–j). Interestingly, the distances between regions across the insulator were increased on the CAST allele compared to the 129 allele, whereas the mutant CBS inserted at the same location did not increase the distance between regions across the insertion (Extended Data Fig. 9a,b). In contrast, the 4CBS insulator inserted downstream of the *Sox2* super-enhancer appeared to promote segregation of the *Sox2* domain from downstream chromatin, which may explain the slightly increased *Sox2* expression in this clone (Extended Data Fig. 9c).

Surprisingly, although the 4CBS insulator substantially reduced *Sox2* expression and the contact frequency between *Sox2* and its super-enhancer, the median spatial distance between the *Sox2* super-enhancer and promoter only mildly increased on the CAST alleles (282 nm) compared to the 129 alleles (264 nm; Wilcoxon rank-sum test, $P=0.066$; Fig. 5g). We hypothesized that only on a small fraction of chromosomes the *Sox2* super-enhancer was in physical proximity with the *Sox2* promoter to engage in productive transcription, and insertion of an insulator on the CAST allele could reduce this fraction of engaged *Sox2* enhancer–promoter configuration selectively on the CAST allele. To test this hypothesis, we quantified the fraction of CAST alleles that showed a spatial distance between the *Sox2* promoter and the super-enhancer shorter than a particular threshold and compared it to that of the 129 alleles in the same cells. Indeed, in the mESCs where the 4CBS insulator was inserted between the *Sox2* gene and super-enhancer on the CAST allele, the ratio between the fraction of CAST alleles with spatially proximal enhancer–promoter pairs and the fraction of 129 alleles with spatially proximal enhancer–promoter pairs was much smaller than 1, at a spatial distance threshold of 150 nm, and the ratio increased gradually to 1 at a spatial distance threshold of ~300 nm (Fig. 5h). By contrast, no reduction of this ratio was observed at a shorter spatial threshold in mESC clones where CTCF motifs were deleted from the insulator, or when the insulator sequence was inserted downstream of the *Sox2* super-enhancer (Fig. 5h).

To further study how insulators affect enhancer–promoter spatial proximity and enhancer-dependent transcriptional activation at single-cell resolution, we simultaneously probed the chromatin structure with multiplexed DNA FISH and the transcripts at the *Sox2* locus with single-molecule RNA FISH^{61,63}. We first hybridized three sets of RNA-FISH probes targeting *Sox2*, *egfp* and *mcherry*, each with a unique readout sequence to distinguish the transcripts made from the two *Sox2* chromosome copies in each cell

(Supplementary Table 7). We then performed multiplexed DNA FISH with the same cells to trace the local chromatin configuration. The *Sox2* chromatin loci that spatially overlapped with nascent *Sox2* transcripts were designated as transcriptionally bursting loci, and the remaining *Sox2* loci without a coincident transcript were regarded to be in the resting state (Extended Data Fig. 10a). Consistent with the RNA-seq analysis described above (Extended Data Fig. 2d,e), the frequency of detecting the nascent *Sox2* transcripts on the CAST allele was substantially lower than that of the 129 allele in the 4CBS clone (Extended Data Fig. 10b). By contrast, the frequency of detecting nascent *Sox2* transcripts on the CAST allele was slightly higher than that of the 129 allele in the control cells in which the CBS insulator was inserted downstream of the *Sox2* enhancer (Extended Data Fig. 10b). Consistent with the findings of previous studies^{61,64}, we found that nascent *Sox2* transcripts were detected across a wide range of spatial distances between the *Sox2* enhancer and promoter, although the median enhancer–promoter distances at the *Sox2* gene with coincident nascent transcripts were slightly but significantly shorter than those on the resting loci (Extended Data Fig. 10c,d). However, the fraction of the *Sox2* genes with coincident nascent transcripts on the CAST allele in the 4CBS clone was consistently lower than that on the 129 allele even though the spatial distances between the enhancer and promoter are similar (Fig. 5i). By contrast, the fraction of the *Sox2* genes with coincident nascent transcripts was comparable between the two alleles when the 4CBS was inserted downstream of the *Sox2* enhancer (Fig. 5j). These results, taken together, suggest that enhancer proximity is positively correlated to transcriptional activity at the target gene in general; however, alone it is not sensitive enough to differentiate transcriptional states. In summary, our results suggest that CTCF insulators decrease the frequency of transcription bursting at *Sox2* when inserted between the enhancer and promoter, likely by establishing local chromatin domain boundaries that weaken productive communications between a spatially close enhancer and promoter (Fig. 5k).

Discussion

The sequence-specific DNA-binding protein CTCF plays a role in both chromatin organization and transcriptional insulation, but exactly how chromatin topology is related to transcriptional insulation remains to be understood. In this study, we developed an experimental system using mESCs to quantify the enhancer-blocking activity of insulators in the native chromatin context at the *Sox2* locus. The well-defined distal enhancer of *Sox2* gene activation afforded an excellent opportunity to quantify the effects of insulator insertions on local chromatin structure and transcription *in cis*. We determined the insulator activity of a number of CBSs either alone or in various combinations, and demonstrated that potent insulation was rendered by two or more clustered CBSs. Importantly, we found that CTCF binding alone was insufficient to confer insulation activity; rather, sequences immediately adjacent to CTCF-binding motifs were required for potent insulator function. Consistent with this observation, CBSs within TAD boundaries are more likely to function as insulators than those not located at TAD boundaries, regardless of the strength of their binding by CTCF. Finally, using two complementary approaches to profile chromatin architecture, we showed that CTCF likely mediates transcriptional insulation by creating local chromatin domain boundaries and reducing the frequency of productive enhancer–promoter contacts. Our results, therefore, provide mechanistic insights into the link between TAD boundaries that are enriched for CBSs and CTCF-mediated transcriptional insulation.

We demonstrated that several factors may be involved in CTCF-mediated transcriptional insulation in mammalian cells. First, a single CBS has weak insulation effects, which vary depending on the orientation of the CTCF motif. The orientation bias is

likely due to a pair of convergent CBSs located on the *Sox2* promoter and enhancer. The CBS insertion is predicted to loop with the enhancer CBS in a forward orientation⁵⁹. A loop formed with the enhancer may block enhancer activity more efficiently than one formed with the promoter. Given that the inserted insulator is closer to the *Sox2* enhancer, where CTCF binding is stronger, it is also possible that looping with CBS on the *Sox2* super-enhancer is more efficient, thereby, favoring insulation by forward-orientated CBSs.

Second, we found that multiple CBSs taken from TAD boundaries exert potent transcriptional insulation activities. Our finding is consistent with a recent study of mouse *Pcdh* clusters reporting that insertion of tandem CTCF sites could block enhancers from activating proximal genes⁶⁵. These observations with CTCF insulators are different from those with the *Drosophila* gypsy insulator, which was ineffective in blocking enhancer activity when two tandem copies were tested^{66,67}.

Third and more importantly, through sequence swapping experiments, we showed that sequences immediately adjacent to CTCF-binding motifs were necessary for enhancer-blocking function. We further found an upstream element in the flanking sequences of CTCF-binding motifs to be crucial for transcriptional insulation. Previous studies reported an upstream motif that stabilizes CTCF binding via interactions with zinc-fingers 9–11 of CTCF^{53–55,68}. We speculate that CTCF zinc-fingers 9–11 may promote transcriptional insulation by inducing a tertiary structure on insulators that stabilizes CTCF–cohesin interactions, thereby blocking the loop-extrusion process that facilitates long-range enhancer–promoter contacts. It is noteworthy that deleting zinc-fingers 9–11 did not fully abolish insulation of the boundary CBSs, suggesting the involvement of additional factors in transcriptional insulation.

Our study also relates the chromatin structure involving enhancer–promoter contacts, as revealed by various 3C-based and microscopy-based experiments, to enhancer-dependent transcription. From both the 3C-based and imaging experiments, we found that the insertion of multiple CBSs in tandem, with the appropriate flanking sequences, induced the formation of a TAD boundary at the insertion site and reduced interactions between the enhancer and the promoter. Spatial proximity between an enhancer and a promoter has been thought to be positively correlated with enhancer-dependent activation in general. However, recent studies have also shown that spatial proximity is not strictly correlated with transcriptional activation⁶⁹, and is a poor predictor of transcriptional activity in live cells⁶⁴. We showed that the transcriptional activities of the *Sox2* promoter, measured by the frequency of nascent transcripts detected at the gene locus in a population, could be reduced by insulators with only modest changes to the enhancer–promoter proximity. These studies, together, highlight that enhancer–promoter proximity is just one of the many elements regulating transcriptional activity in mammalian cells. The point-to-point spatial distance between the enhancer and promoter does not fully reflect the chromatin structure of the entire locus. Simultaneous imaging of chromatin and transcripts indicated that *Sox2* transcription could take place in chromosomes showing a broad range of spatial distances between the enhancer and *Sox2* promoter⁶¹. One possibility is that the *Sox2* super-enhancer forms a phase-separated environment⁷⁰, where the *Sox2* gene needs not be very close to its enhancer to be activated. Another possibility is that the temporal duration of *Sox2* enhancer–promoter interaction is relatively short compared to a transcriptional bursting cycle, which would make it difficult to capture the two events simultaneously in fixed cells using FISH. Finally, transcription is not likely to happen immediately after enhancer–promoter contacts⁷¹. The lagging between these two events could also explain the lack of strict correlation between enhancer–promoter proximity and transcriptional bursting in live cells⁶⁴.

In summary, our results suggest that CTCF sites in the genome are not all equivalent to each other, and CTCF-mediated insulation depends on both dosage and upstream flanking sequences. Our findings explain why CBSs at TAD boundaries are more likely to act as transcriptional insulators than those outside TAD boundaries. One potential limitation of the current study is that the insulation effects of CBSs were tested only in the *Sox2* locus. Future experiments will be needed to demonstrate whether observations made from the *Sox2* locus can be generalized to other gene loci in the mammalian genome.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-021-00863-6>.

Received: 23 July 2020; Accepted: 2 April 2021;

Published online: 17 May 2021

References

- Hnisz, D., Day, D. S. & Young, R. A. Insulated neighborhoods: structural and functional units of mammalian gene control. *Cell* **167**, 1188–1200 (2016).
- Kellis, M. et al. Defining functional DNA elements in the human genome. *Proc. Natl Acad. Sci. USA* **111**, 6131–6138 (2014).
- Levine, M., Cattoglio, C. & Tjian, R. Looping back to leap forward: transcription enters a new era. *Cell* **157**, 13–25 (2014).
- West, A. G., Gaszner, M. & Felsenfeld, G. Insulators: many functions, many mechanisms. *Genes Dev.* **16**, 271–288 (2002).
- Geyer, P. K. & Corces, V. G. DNA position-specific repression of transcription by a *Drosophila* zinc finger protein. *Genes Dev.* **6**, 1865–1873 (1992).
- Recillas-Targa, F., Bell, A. C. & Felsenfeld, G. Positional enhancer-blocking activity of the chicken β -globin insulator in transiently transfected cells. *Proc. Natl Acad. Sci. USA* **96**, 14354–14359 (1999).
- Stief, A., Winter, D. M., Stratling, W. H. & Sippel, A. E. A nuclear DNA attachment element mediates elevated and position-independent gene activity. *Nature* **341**, 343–345 (1989).
- Gurudatta, B. V. & Corces, V. G. Chromatin insulators: lessons from the fly. *Brief. Funct. Genomic Proteomic* **8**, 276–282 (2009).
- Chung, J. H., Bell, A. C. & Felsenfeld, G. Characterization of the chicken β -globin insulator. *Proc. Natl Acad. Sci. USA* **94**, 575–580 (1997).
- Lobanenkov, V. V. et al. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken *c-myc* gene. *Oncogene* **5**, 1743–1753 (1990).
- Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405**, 482–485 (2000).
- Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in *IDH* mutant gliomas. *Nature* **529**, 110–114 (2016).
- Katainen, R. et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015).
- Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.* **17**, 520–527 (2001).
- Filippova, G. N. et al. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian *c-myc* oncogenes. *Mol. Cell Biol.* **16**, 2802–2813 (1996).
- Lupianez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
- Shukla, S. et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79 (2011).
- Vostrov, A. A. & Quitschke, W. W. The zinc finger protein CTCF binds to the APB β domain of the amyloid β -protein precursor promoter. Evidence for a role in transcriptional activation. *J. Biol. Chem.* **272**, 33353–33359 (1997).
- Zhang, X. et al. Fundamental roles of chromatin loop extrusion in antibody class switching. *Nature* **575**, 385–389 (2019).
- Guo, Y. et al. CTCF/cohesin-mediated DNA looping is required for protocadherin alpha promoter choice. *Proc. Natl Acad. Sci. USA* **109**, 21081–21086 (2012).
- Guo, Y. et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* **162**, 900–910 (2015).

22. Ghirlando, R. & Felsenfeld, G. CTCF: making the right connections. *Genes Dev.* **30**, 881–891 (2016).
23. Phillips-Cremins, J. E. & Corces, V. G. Chromatin insulators: linking genome organization to cellular function. *Mol. Cell* **50**, 461–474 (2013).
24. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
25. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
26. Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).
27. Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944 (2017).
28. Luppino, J. M. et al. Cohesin promotes stochastic domain intermingling to ensure proper regulation of boundary-proximal genes. *Nat. Genet.* **52**, 840–848 (2020).
29. Wutz, G. et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* **36**, 3573–3599 (2017).
30. Alipour, E. & Marko, J. F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res.* **40**, 11202–11212 (2012).
31. Davidson, I. F. et al. DNA loop extrusion by human cohesin. *Science* **366**, 1338–1345 (2019).
32. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
33. Haarhuis, J. H. I. et al. The cohesin release factor WAPL restricts chromatin loop extension. *Cell* **169**, 693–707 (2017).
34. Kim, Y., Shi, Z., Zhang, H., Finkelstein, I. J. & Yu, H. Human cohesin compacts DNA by loop extrusion. *Science* **366**, 1345–1349 (2019).
35. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320 (2017).
36. Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl Acad. Sci. USA* **112**, E6456–E6465 (2015).
37. Vian, L. et al. The energetics and physiological impact of cohesin extrusion. *Cell* **173**, 1165–1178 (2018).
38. Wutz, G. et al. ESCO1 and CTCF enable formation of long chromatin loops by protecting cohesin^{STAG1} from WAPL. *eLife* **9**, e52091 (2020).
39. Brackley, C. A. et al. Nonequilibrium chromosome looping via molecular slip links. *Phys. Rev. Lett.* **119**, 138101 (2017).
40. Barbieri, M. et al. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl Acad. Sci. USA* **109**, 16173–16178 (2012).
41. Bianco, S. et al. Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat. Genet.* **50**, 662–667 (2018).
42. Brackley, C. A., Taylor, S., Papantonis, A., Cook, P. R. & Marenduzzo, D. Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. *Proc. Natl Acad. Sci. USA* **110**, E3605–E3611 (2013).
43. Buckle, A., Brackley, C. A., Boyle, S., Marenduzzo, D. & Gilbert, N. Polymer simulations of heteromorphic chromatin predict the 3D folding of complex genomic loci. *Mol. Cell* **72**, 786–797 (2018).
44. Conte, M. et al. Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation. *Nat. Commun.* **11**, 3289 (2020).
45. Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G. & Onuchic, J. N. Transferable model for chromosome architecture. *Proc. Natl Acad. Sci. USA* **113**, 12168–12173 (2016).
46. Schwarzer, W. et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017).
47. Despang, A. et al. Functional dissection of the *Sox9-Kcnj2* locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).
48. 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**, 759–773 (2003).
49. Li, Y. et al. CRISPR reveals a distal super-enhancer required for *Sox2* expression in mouse embryonic stem cells. *PLoS ONE* **9**, e114485 (2014).
50. Zhou, H. Y. et al. A *Sox2* distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev.* **28**, 2699–2711 (2014).
51. Kentepozidou, E. et al. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol.* **21**, 5 (2020).
52. Frith, M. C., Saunders, N. F., Kobe, B. & Bailey, T. L. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput. Biol.* **4**, e1000071 (2008).
53. Nakahashi, H. et al. A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.* **3**, 1678–1689 (2013).
54. Xu, D. et al. Dynamic nature of CTCF tandem 11 zinc fingers in multivalent recognition of DNA as revealed by NMR spectroscopy. *J. Phys. Chem. Lett.* **9**, 4020–4028 (2018).
55. Yin, M. et al. Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. *Cell Res.* **27**, 1365–1377 (2017).
56. Yan, J. et al. Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell Res.* **28**, 204–220 (2018).
57. Fang, R. et al. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.* **26**, 1345–1348 (2016).
58. Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
59. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
60. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362** (2018).
61. Mateo, L. J. et al. Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* **568**, 49–54 (2019).
62. Wang, S. et al. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598–602 (2016).
63. Su, J. H., Zheng, P., Kinrot, S. S., Bintu, B. & Zhuang, X. Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell* **182**, 1641–1659 (2020).
64. Alexander, J. M. et al. Live-cell imaging reveals enhancer-dependent *Sox2* transcription in the absence of enhancer proximity. *eLife* **8**, e41769 (2019).
65. Jia, Z. et al. Tandem CTCF sites function as insulators to balance spatial chromatin contacts and topological enhancer-promoter selection. *Genome Biol.* **21**, 75 (2020).
66. Cai, H. N. & Shen, P. Effects of cis arrangement of chromatin insulators on enhancer-blocking activity. *Science* **291**, 493–495 (2001).
67. Muravyova, E. et al. Loss of insulator activity by paired Su(Hw) chromatin insulators. *Science* **291**, 495–498 (2001).
68. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).
69. Benabdallah, N. S. et al. Decreased enhancer-promoter proximity accompanying enhancer activation. *Mol. Cell* **76**, 473–484 (2019).
70. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A phase separation model for transcriptional control. *Cell* **169**, 13–23 (2017).
71. Chen, H. et al. Dynamic interplay between enhancer-promoter topology and gene activity. *Nat. Genet.* **50**, 1296–1303 (2018).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Cell culture. The hybrid F123 mESC line (F₁ *Mus musculus castaneus* × S129/SvJae, maternal 129/Sv, paternal CAST) was from Rudolf Jaenisch's laboratory at the Whitehead Institute at MIT. The wild-type F123 mESC line and engineered clones were maintained in feeder-free, serum-free 2i conditions (1 μM PD03259010, 3 μM CHIR99021, 2 mM glutamine, 0.15 μM monothio glycerol, 1,000 U ml⁻¹ LIF). The growth medium was changed every day. Cells were dissociated by Accutase (AT104) and passaged onto 0.2% gelatin-coated plates every 2–3 days.

Genetic engineering of the *Sox2* locus. Tagging of the *Sox2* gene with a fluorescent reporter was performed by CRISPR–Cas9-mediated homologous recombination. Specifically, a guide RNA expression plasmid (pX330, Addgene no. 42230) targeting the 3' UTR of the *Sox2* gene, with *egfp* and *mcherry* donor plasmids, was coelectroporated into wild-type F123 cells by the Neon transfection system (MPK1096). Cells were recovered for 2 days; then eGFP⁺ mCherry⁺ cells were sorted by FACS and seeded onto a new 0.2% gelatin-coated 60-mm dish. A second round of FACS was performed 5 days later to enrich eGFP⁺ mCherry⁺ cells. A total of 500–1,000 double-positive single cells were seeded onto a new 60-mm dish and single colonies were picked manually a further 5 days later. Allele-specific genotyping of *Sox2* was performed with primers spanning CAST/129 SNPs. A clone with the CAST allele *Sox2* gene fused with *egfp* and the 129 allele *Sox2* gene fused with *mcherry* was selected as the parental clone. Subsequently, the *HyTK* fusion gene was integrated into the CAST allele of the parental clone by CRISPR–Cas9 editing. Specifically, electroporated cells were recovered for 2 days and then cultured in growth medium containing 200 μg ml⁻¹ hygromycin for 7 days. Surviving cells were dissociated into single cells and seeded at a density of 500–1,000 cells per 60-mm dish. After 5 days, colonies were manually picked and genotyped with primers spanning CAST/129 SNPs. Genotyping primers were synthesized by Integrated DNA Technologies (Supplementary Table 8).

Donor plasmids cloning for RMCE. The donor vector was adapted from the pUC19 plasmid. Two heterotypic flippase recognition sites, FRT/F3, as well as NotI and SbfI restriction enzyme recognition sites, were added into the pUC19 plasmid by PCR. The donor vector was then digested with the enzyme cocktail of NotI–HF (NEB, R3642S), SbfI–HF (NEB, R3189S) and rSAP (NEB, M0371S) for 4 h at 37°C. Individual CBSs were PCR amplified from mouse or human genomic DNA. The PCR primers contain overhang sequences of NotI and SbfI sites to specify CTCF motif orientation. PCR products were purified by gel electrophoresis, digested and ligated into the donor vector. Ligation products were transformed into Stbl3 chemically competent cells. Positive clones were screened by PCR and plasmids were extracted using the Qiagen Plasmid Plus Midi Kit (catalog no. 12943) and validated by Sanger sequencing.

Marker-free insertion in mESCs by RMCE. A flippase expression plasmid (pFlpe; Addgene no. 13787) and a donor plasmid (pDonor) were coelectroporated into 0.1 million insulator reporter or control cells at a ratio of 1:4 (pFlpe/pDonor = 1 μg:4 μg). Cells were recovered for 2 days and cultured in growth medium containing 2 μM ganciclovir for another 5 days. Surviving cells were dissociated into a single-cell suspension and seeded at a density of 500–1,000 cells per 60-mm dish. After 5 days, six colonies were picked for PCR genotyping. Genomic DNA was then extracted by Qiagen DNeasy Blood & Tissue Kits (no. 69506 and no. 69581). For each insert, three independent clones were randomly picked for FACS analysis and subsequent studies. Individual CBSs were combined by PCR to create CBS clusters. Specifically, the 4CBS cluster from the *Sox9–Kcnj2* TAD boundary was composed of genomic sequences from chr11:111,523,291–111,524,273, chr11:111,531,104–111,533,964 and chr11:111,535,307–111,538,959. The PCR primers were synthesized by Integrated DNA Technologies (Supplementary Table 8).

Deleting zinc-fingers 9–11 of CTCF in mESCs. Deletion of CTCF zinc-fingers 9–11 was achieved by CRISPR–Cas9-mediated homologous recombination as previously described³⁷. Briefly, coding sequences of exons 10–12 of the *Ctcf* gene, together with an SV40 polyA signal, were inserted into exon 9 of the *Ctcf* gene in situ, resulting in only zinc-fingers 1–8 of the CTCF protein being functional. About 0.15 million cells were transfected with a mixture of guide RNA expression plasmid (pX330, 1 μg), homologous recombination repair plasmid (4 μg) and a coelectroporation marker (0.1 μg, puromycin resistant). After 2 days' recovery, cells were treated with 1 μg ml⁻¹ puromycin for another 3 days. Surviving cells were suspended into single cells and seeded at a density of 500–1,500 cells per 10-cm Petri dish. After 5 days, single colonies were manually picked and genotyped by PCR.

FACS data acquisition and analysis. Cells were treated with Accutase (no. AT104) at 37°C for 5–7 min and resuspended into single cells with 2 ml warm 2i/LIF medium. Cells were then spun down at 1,000 r.p.m. for 4 min and washed twice with 5 ml PBS. Cell pellets were resuspended into single cells with 1 ml PBS and filtered through the 35-μm strainer cap of a FACS tube (FSC-9005). Then, cells were sorted by a Sony sorter SH800 (Cell Sorter Software 2.1.5) in analysis mode using a 130-μm chip. For each insertion clone, both GFP and mCherry signals were

recorded for 10,000 cells. Cells were first gated by SSCA–FSCA for live cells, and then by FSA–FSH for singlets using FlowJo 10.0.7r2. Fluorescence signals of cells that passed gating were exported in csv files and analyzed in R 3.6.0. Specifically, the GFP signal is normalized by the mCherry signal from the same cell. For each insertion clone, the normalized *Sox2*–eGFP expression was calculated as:

$$\text{Mean} \left(\frac{\text{eGFP}}{\text{mCherry}} \right)_{\text{insertion}} / \text{mean} \left(\frac{\text{eGFP}}{\text{mCherry}} \right)_{\text{no insertion}}$$

To better estimate instrument variability in FACS, we used replicates of the no insertion clone in all experiments as controls when testing the significance of the insulation effects of the inserted DNA elements.

ChIP-seq. Cells were dissociated into single cells and crosslinked by 1% formaldehyde in PBS for 15 min at room temperature. Crosslinking was then quenched by 0.125 M glycine and cells were washed twice with 5 ml cold PBS. Permeabilized nuclei were prepared with the Covaris truChIP Chromatin Shearing Kit (PN520154) following the manufacturer's instructions. A total of 1–3 million nuclei were sonicated in a 130-μl microtube by a Covaris M220 instrument (power, 75 W; duty factor, 10%; cycle per burst, 200; time, 10 min; temperature, 7°C). Sonicated chromatin was diluted with 1× shearing buffer into a total volume of 1 ml and spun down at 15,000 r.p.m. at 4°C to remove cell debris. A 5 μg quantity of antibodies were added to the supernatant and incubated overnight at 4°C with gentle rotation (CTCF, ab70303, lot GR3281212-6.7.8; RAD21, ab992, lot GR3253930-8, GR3310168-11; H3K4me3, Millipore, 04-745, lot 3243412; H3K27ac, Active Motif, 39685, lot 33417016). Chromatin was pulled down by protein G Sepharose beads (GE, 17061801) and washed three times with RIPA buffer (10 mM Tris (pH 8.0), 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate), twice with high-salt RIPA buffer (10 mM Tris (pH 8.0), 300 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate), once with LiCl buffer (10 mM Tris (pH 8.0), 250 mM LiCl, 1 mM EDTA, 0.5% IGEAL CA-630, 0.1% sodium deoxycholate) and twice with TE buffer (10 mM Tris (pH 8.0), 0.1 mM EDTA). Washed chromatin was reverse-crosslinked overnight with 2 μl proteinase K (P8107S, NEB) at 65°C (1% SDS, 10 mM Tris (pH 8.0), 0.1 mM EDTA), column purified and subjected to end repair, A-tailing, adapter ligation and PCR amplification. Final libraries were purified by SPRI beads (0.8:1) and quantified with the Qubit HS dsDNA kit (Q32854).

RNA-seq. Total RNA from cells was extracted using the TRIzol Plus RNA purification kit (Thermo Fisher Scientific, catalog no.: 12183555). RNA-seq libraries were prepared from 4 μg total RNA using Illumina TruSeq Stranded mRNA Library Prep Kit Set A (RS-122-2101; Illumina) or Set B (RS-122-2102; Illumina). RNA-seq libraries were sequenced on Illumina NextSeq 550 and HiSeq 4000 platforms (75-bp paired ends).

PLAC-seq. PLAC-seq libraries were prepared as previously described^{57,58} with minor modifications. In brief, 2–3 million cells were crosslinked for 15 min at room temperature with 1% methanol-free formaldehyde and quenched for 5 min at room temperature with 0.2 M glycine. The crosslinked cells were lysed in 300 μl Hi-C lysis buffer (10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 0.2% IPEGAL CA-630) for 15 min on ice and then washed once with 500 μl lysis buffer (2,500g for 5 min). Subsequently, cells were resuspended in 50 μl 0.5% SDS and incubated for 10 min at 62°C and then quenched by 160 μl 1.56% Triton X-100 for 15 min at 37°C. Then, 25 μl 10× NEBuffer 2 and 100 U MboI were added to digest chromatin for 2 h at 37°C with shaking (1,000 r.p.m.). Digested fragments were biotin-labeled and subsequently ligated by T4 DNA ligase buffer (NEB) for 2 h at 23°C with 300 r.p.m. gentle rotation. Chromatin was sheared and washed as described for ChIP-seq. Dynabeads (M-280 sheep anti-rabbit IgG, catalog: 11203D) coated with 5 μg H3K4me3 antibodies (Millipore, 04-745, lot 3243412) were used for immunoprecipitation. Pulled down chromatin was treated with 10 μg RNase A for 1 h at 37°C, reverse-crosslinked by 20 μg proteinase K at 65°C for 2 h, and then purified with the Zymo DNA Clean & Concentrator-5 kit. Ligation junctions were enriched by 25 μl myOne T1 streptavidin Dynabeads. Libraries were prepared using the QIAseq Ultraflow Input Library Kit (Qiagen, no. 180492). Final libraries were size-selected with SPRI beads (0.5:1 and 1:1), quantified and submitted for paired-end sequencing.

Hi-C. Cells were processed in the same way as for PLAC-seq before the chromatin shearing steps. Briefly, after the ligation step, nuclei were digested by 50 μl proteinase K (20 mg ml⁻¹) for 30 min at 55°C. DNA was then purified by ethanol precipitation and resuspended in 130 μl 10 mM Tris-HCl (pH 8.0). Purified DNA was sonicated by a Covaris M220 instrument with the following parameters: duty cycle, 10%; power, 50; cycles per burst, 200; time, 70 s. DNA fragments smaller than 300 bp were removed by Ampure XP bead-based dual size selection (0.55:1 and 0.75:1). Biotin-labeled free DNA ends were cleaned up by end-repair reaction and ligation junctions were enriched by streptavidin Dynabeads as described for PLAC-seq. Ligation junctions were then purified and subjected to A-tailing, adapter ligation and PCR amplification. Final libraries were purified by 0.75× Ampure XP beads, quantified and submitted for pair-end sequencing.

Multiplexed FISH imaging for chromatin tracing. Glass coverslips were treated with poly-L-lysine for 30 min at 37 °C. Then, glass coverslips were washed twice with 5 ml PBS and treated with 0.2% gelatin for another 20 min at 37 °C. A total of 2.5 million mESCs were seeded in a 6-cm plastic dish containing the treated glass coverslip. After 20 h, cells were crosslinked using 4% paraformaldehyde, followed by chromatin tracing experiments as described in a previous publication⁶⁰. Briefly, the entire 210-kb *Sox2* region was labeled by a library of primary Oligopaint probes^{60,61}. Each primary probe consists of a unique 42-nucleotide readout sequence that is specific for each 5-kb DNA segment. Next, secondary readout probes complementary to the readout sequences on the primary probes were added to the cells. Last, fluorophore-labeled common imaging probes complementary to the secondary probes were added to the cells to allow three-dimensional diffraction-limited imaging of individual DNA segments. After each round of imaging, the fluorescence signal was extinguished by using both tris(2-carboxyethyl) phosphine cleavage at a concentration of 50 μM in 2× SSC and high-power photobleaching. The process was repeated until all DNA segments were labeled and imaged. We performed three-color imaging by using three secondary readout imaging probes that were conjugated with Cy3, Cy5 and Alexa 750, respectively. In this case, three consecutive 5-kb chromatin segments were labeled by each round of imaging. A pool of 42 oligonucleotide probe sets was designed to scan the 210-kb *Sox2* locus with each set covering a 5-kb DNA region.

Multiplexed RNA and DNA FISH imaging at the *Sox2* locus. The dual-modality FISH imaging was performed as recently described⁶³. Briefly, the sample was prepared as for the multiplexed FISH imaging for chromatin tracing except that after cells were crosslinked by 4% paraformaldehyde, the sample was hybridized with oligonucleotide probes (Supplementary Table 7) targeting the *Sox2*, *egfp* and *mcherry* transcripts followed by imaging (final concentration 100 nM). Immediately after the RNA FISH imaging was completed, the sample was washed with 50% formamide to remove residual fluorescence readout probes and crosslinked again with 4% paraformaldehyde before multiplexed FISH imaging for chromatin tracing.

Data analysis. ChIP-seq. Sequenced reads were aligned to the reference mouse genome mm10 using bowtie2 (version 2.2.9). Unmapped reads and PCR duplicates were removed. For clones with the insertion of synthetic CBSs, reads were aligned to a customized mm10 reference genome that includes the inserted sequence. The mapping pipeline is available at <http://renlab.sdsc.edu/huh025/chipseq-PE/>. Signal tracks were generated with the command `bamCoverage (version 3.3.1) --normalizingRPKM -bs 50 --smoothLength 150`. Peaks were called by `macs2 (version 2.1.1.20160309)` with default parameters.

RNA-seq. The RNA-seq alignment and quantification pipeline is available at <https://github.com/ren-lab/rnaseq-pipeline>. Briefly, reads were aligned to mm10 (GRCm38) and GENCODE GTF version M25 with rnaSTAR⁷² (version 020201). In particular, we created two extra chromosomes for the two tagged *Sox2* alleles. PCR duplicates were removed using Picard. Reads uniquely mapped to *egfp* and *mcherry* sequences were counted using samtools. *Sox2* expression from the CAST and 129 alleles was quantified by reads per kilobase of transcript per million mapped reads (RPKM) values of the *egfp* and *mcherry* gene, respectively.

PLAC-seq. To resolve allele-specific interactions, we created VCF files containing SNPs with respect to the mm10 reference genome for the parental strains CAST/EIJ and 129SV/Jae. Specifically, whole-genome sequencing reads from the two strains were mapped to mm10 and deduplicated, and SNPs were called using bcftools. We removed heterozygous SNP calls and those with sequencing depth less than 5 and quality less than 30 and further removed SNPs that were present in both strains. We used a modified mapping procedure from the WASP⁷³ pipeline (version 0.3.4) to detect allele-specific contacts. Since the WASP pipeline ignores indels, we further removed all reads that map to within 50 bp from the nearest indel. We modified the original WASP mapping procedure by replacing the bowtie2 alignment tool with BWA-MEM and integrated the MAPS⁷⁴ feather post-filtering pipeline to resolve the chimeric reads. The analysis pipeline is available at <https://github.com/ijuric/Sox2AllelicAnalysis>.

Hi-C. To process Hi-C data, we used our in-house pipeline available at <https://github.com/ren-lab/hic-pipeline>. Briefly, Hi-C reads were aligned to mm10 using BWA-MEM (version 0.7.12-r1039) for each read separately and then paired. For chimeric reads, only 5' end-mapped locations were kept. Duplicated read pairs mapped to the same location were removed to leave only one unique read pair. The output bam files were transformed into juicer file format for visualization in Juicebox 1.11.08. Contact matrices were normalized using the Knight–Ruiz matrix balancing method⁷⁵. The directionality index score for each sample was generated at 50-kb resolution with a 2-Mb window (40 bins) as described in previous work⁷⁴. Haplotype phasing was performed using the obtained CAST/129 VCF file. This created two contact matrices corresponding to the CAST allele and the 129 allele for each Hi-C library. For each phased haplotype of chromosome 3, the directionality index score was generated at 10-kb resolution with a 50-kb window (5 bins).

Chromatin tracing data processing. Custom software was used to obtain images of chromatin architecture as described previously⁶⁰ with minor modifications. The software identifies the centroid position of each 5-kb chromatin segment using diffraction-limited z-stack images acquired by epifluorescence microscopy. Chromosome locations were first identified via the segmentation of the nuclei in each field of view using a convolutional neural network. The segmentation masks were then applied to limit the chromosome candidates to the two most likely clusters of fluorescent spots present in each nucleus. We then selected the two spots that showed the strongest averaged fluorescence signal over all imaging rounds as the two alleles for each nucleus. To avoid selecting the same chromosome, we also required the two spots to be separated by at least 10 pixels (1.08 μm). The algorithm then utilized the identified chromosome locations to select candidate spots of the imaged 5-kb chromatin segments in every round of imaging. A Gaussian fitting algorithm was then used to fit both the signal of each of the candidate segments and the signals of the fiducial beads. The chromatic aberration, flat-field and drift correction algorithms were adopted from a published work⁶⁰.

The candidate spot of each segment was then further evaluated for its likelihood to be accepted or rejected as estimated by an expectation–maximization algorithm. The expectation–maximization algorithm computes a score based on the product of three terms—the brightness of the spot, the proximity of the spot to the estimated chromosome centroid position and the proximity of the spot to a moving average localization of the candidates selected in the previous five rounds of imaging—for each candidate spot of a segment. The expectation–maximization algorithm selected the highest-scoring candidate spot for each chromosome segment in each round of imaging, while all remaining candidate spots were not considered in subsequent analyses.

The misidentification rate was computed as the percentage of fluorescent spots among the top discarded candidate spots that had scores above the expectation–maximization score threshold that we chose. Finally, only chromosomes that contained accepted segments with a score above the selected threshold across at least ~50% of imaging rounds (22/42 rounds) were kept for further analysis. The detection efficiency of each segment for each experiment was computed as the fraction of segments with accepted candidate spots based on the above procedure. We kept only cells in which one and only one chromosome was detected as being positive for the insertion. In addition, we required the signal of the insertion to be greater than 1/2 of the median value of all segments and at least two times stronger than the signal from the other allele. In this way, the misclassification of the two alleles is estimated to be less than 5%. The insulation score was calculated for each chromosome as the natural log of the ratio of the median distance between loci across domains and the median distance between loci within domains. The *Sox2* enhancer–promoter distance was calculated using the median pairwise Euclidean distance between the genomic locations of the *Sox2* gene (9th–11th region) and its enhancer (30th–32nd region) for every chromosome.

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

Data availability

All next-generation sequencing data are available under GEO accession [GSE153403](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE153403). Raw images of multiplexed FISH experiments and raw FACS data for specific mESC colonies are available on request. Source data are provided with this paper.

Code availability

Code for analyses of multiplexed FISH data can be found on Github at <https://github.com/epigen-UCSD/huang-natgen2021>.

References

- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
- 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).
- Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).

Acknowledgements

We are grateful for comments from members of the Ren laboratory. This study was supported by funding from the Ludwig Institute for Cancer Research and NIH (U54 DK107977 and UM1HG011585, to B.R., X.Z., M.H. and M.N., and 3U54DK107977-05S1 to B.R.). Y.Z. was supported by 1K99CA252020-01. X.Z. is a Howard Hughes Medical Institute Investigator.

Author contributions

B.R. and H.H. conceived the study. B.R. supervised the study. H.H. performed insulator assays and related analysis. R.H. and M.Y. performed PLAC-seq and Hi-C experiments.

I.J. and M.H. analyzed PLAC-seq data. M.T. performed western blot experiments. Y.Z. performed Hi-C analysis. Q.Z. and Y.H. performed chromatin tracing experiments with help from B.B. and X.Z. A.J., B.B., C.K., M.C., S.B., A.M.C. and M.N. analyzed chromatin tracing data. The manuscript was written by H.H. and B.R. with input from all coauthors.

Competing interests

B.R. is a cofounder and consultant for Arima Genomics, Inc. and cofounder of Epigenome Technologies. X.Z. is a cofounder and consultant for Vizgen, Inc. The remaining authors declare no competing interests.

Additional information

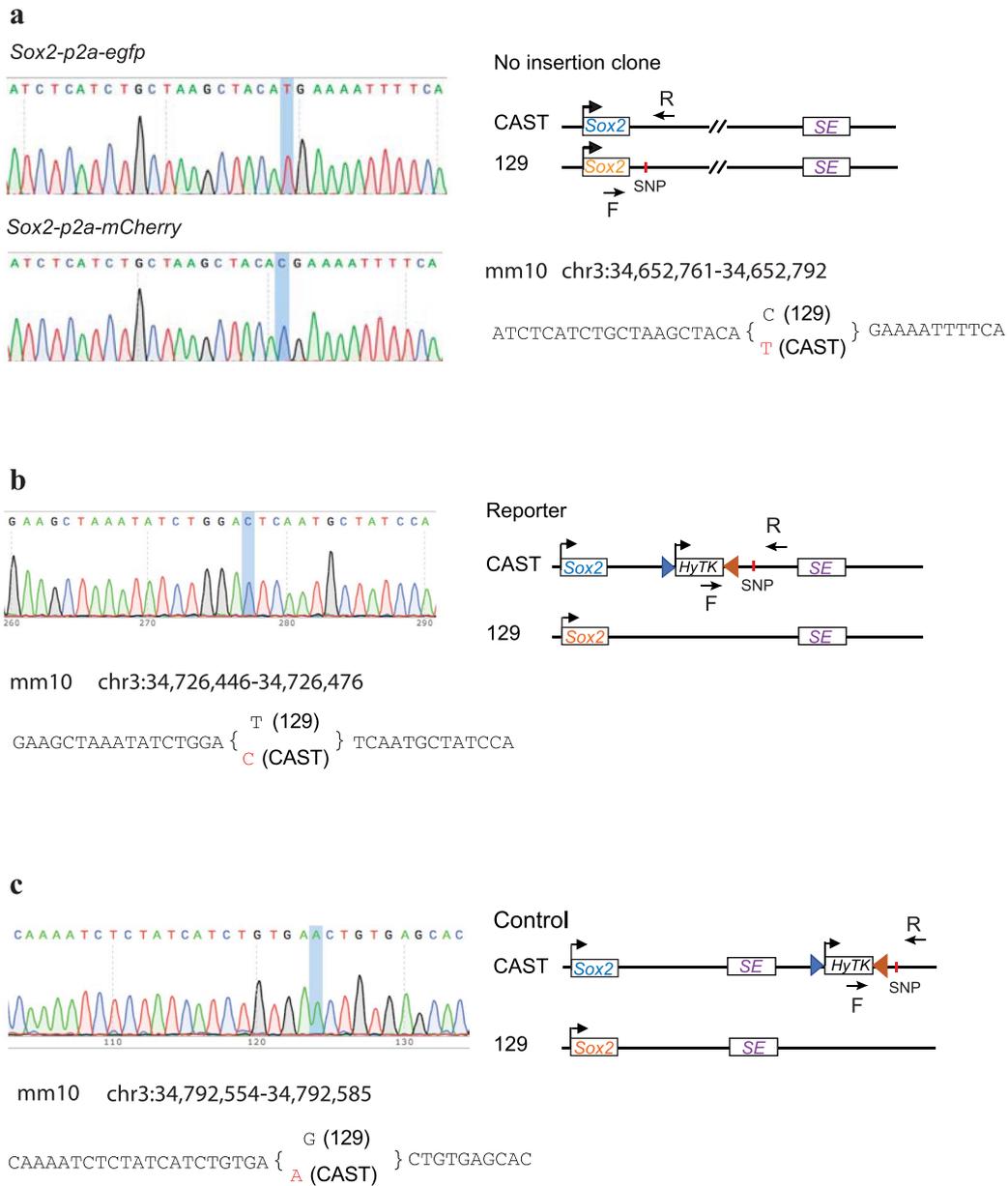
Extended data is available for this paper at <https://doi.org/10.1038/s41588-021-00863-6>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-021-00863-6>.

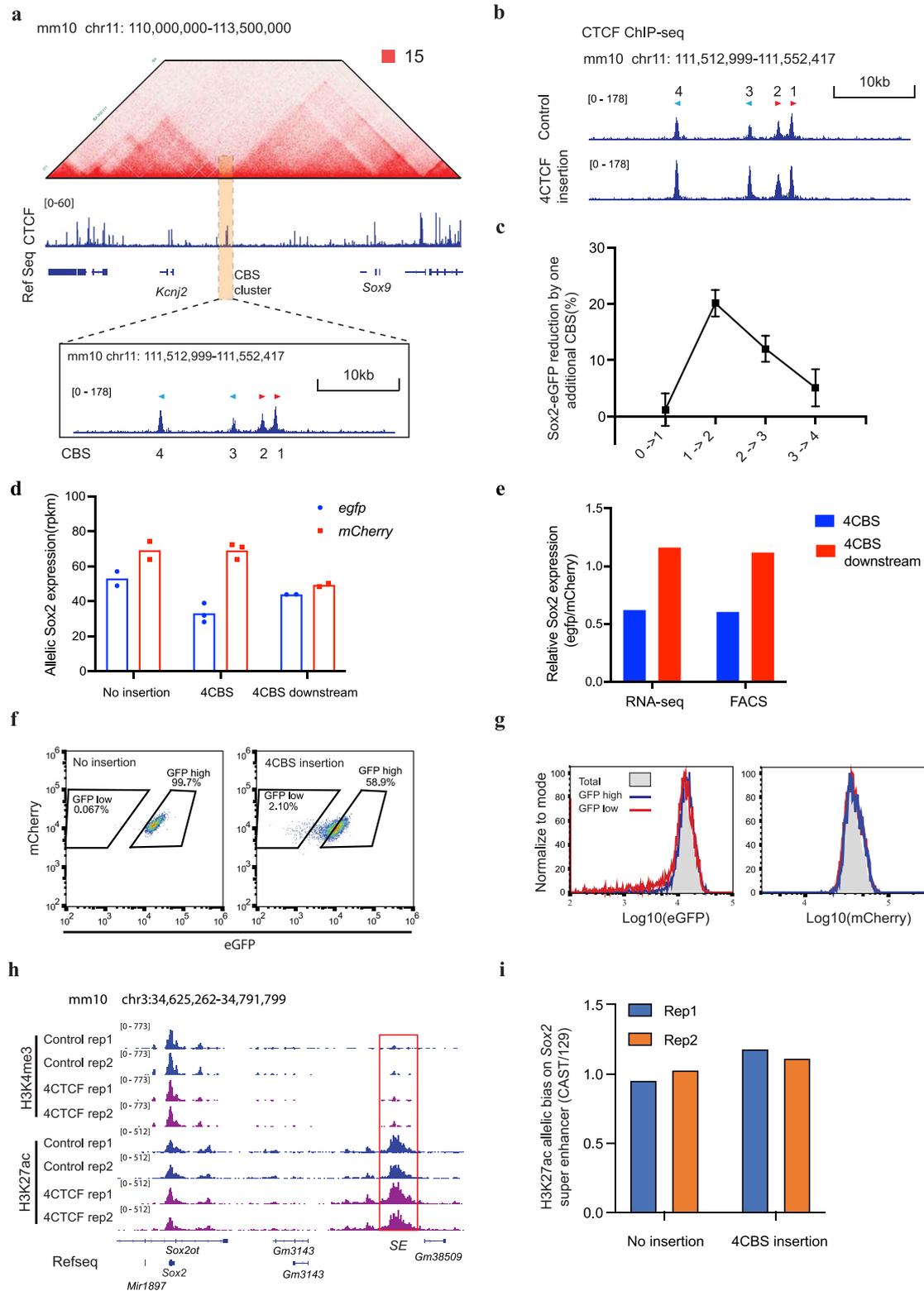
Correspondence and requests for materials should be addressed to B.R.

Peer review information *Nature Genetics* thanks Qiang Wu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

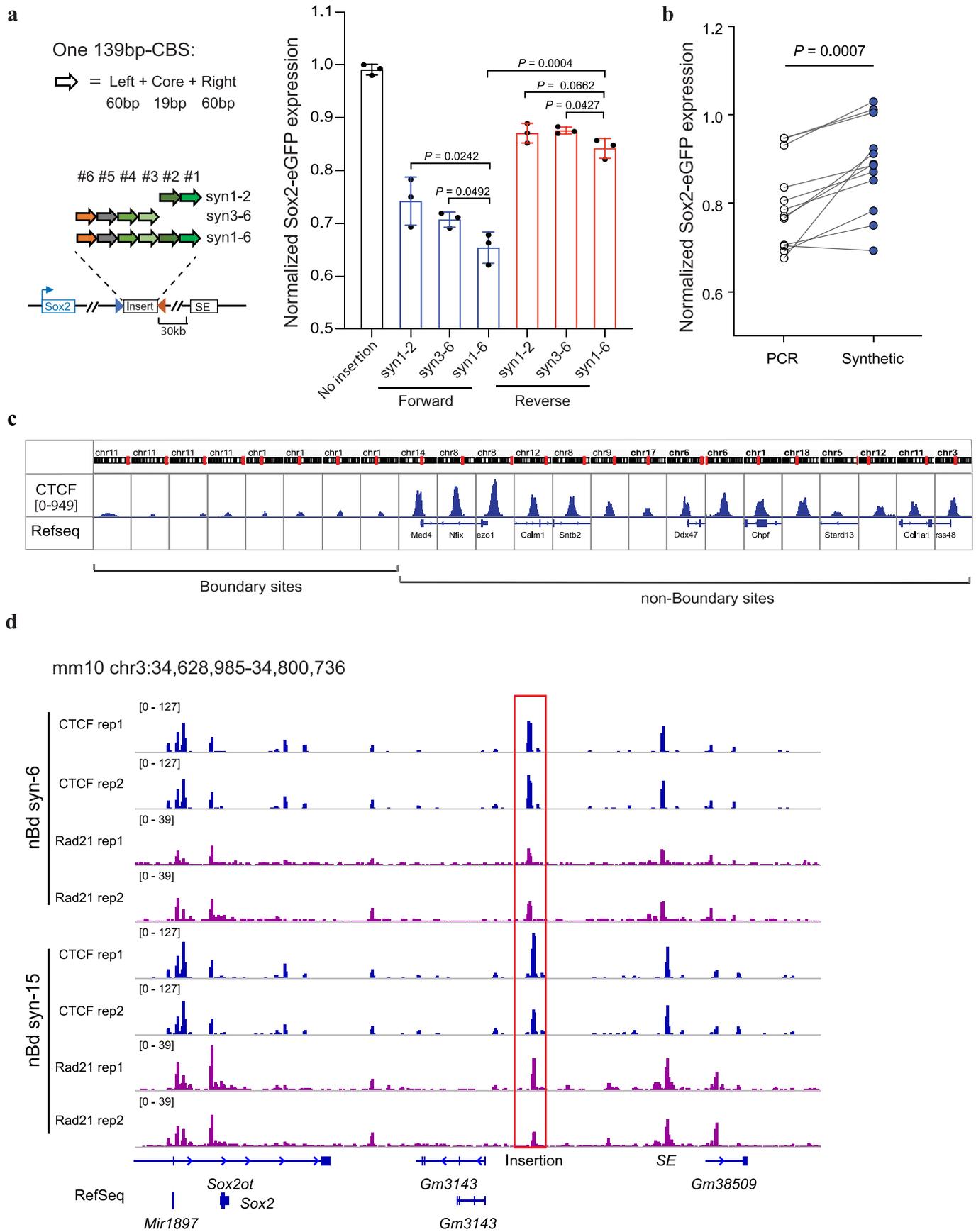


Extended Data Fig. 1 | Genotyping mESC reporter cell lines. a, Genotyping *egfp* and *mcherry* labeled *Sox2* gene. Left, Sanger sequencing results for allele-specific PCR products. Allele-specific SNP is highlighted. Right, the construct of the clone and the SNP information used to distinguish the two alleles. The reverse primer was common, while the forward primer was allele-specific, matching with *egfp* and *mcherry* sequence, respectively. **b-c**, Genotyping the Insulator reporter and control cell lines. Left, Sanger sequencing and SNP information. Right, Construct of the clone and positions of PCR primers. The forward primer is specific to the inserted *HyTK* gene. **b**, insulator reporter cell line. **c**, Insulator control cell line.



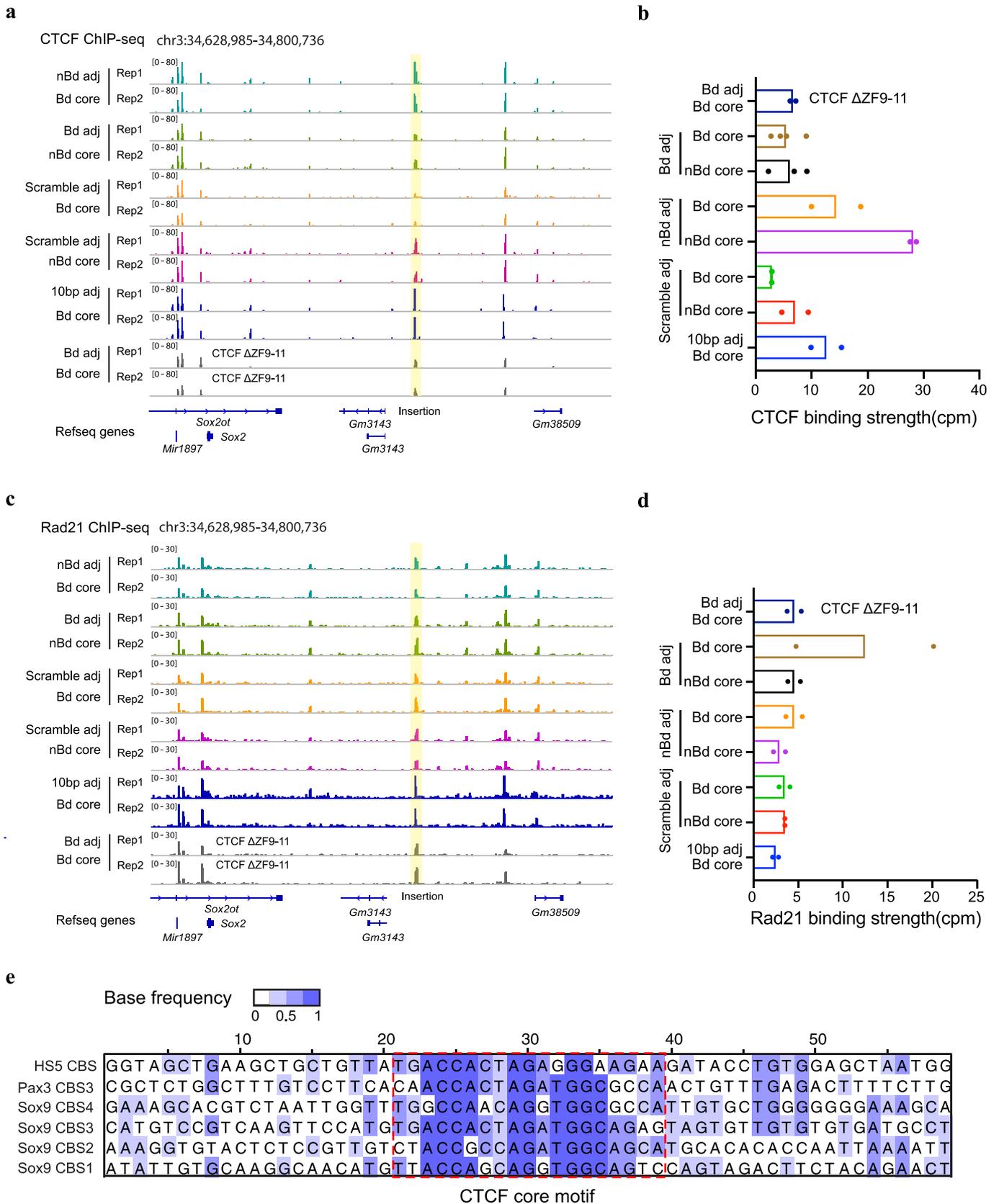
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Insulation features of CBSs from the *Sox9-Kcnj2* TAD boundary. **a**, Hi-C contact map of the *Sox9-Kcnj2* locus in mouse ES cells. Zoom in view shows the four CTCF binding sites cloned for insulator activity test. **b**, ChIP-seq of CTCF in the no insertion clone and the clone with an extra copy of the four *Sox9-Kcnj2* TAD boundary CBS inserted inside the *Sox2* domain. **c**, Reduction in *Sox2*-eGFP expression by one additional CBS. The comparison was between the clones presented in Fig. 2b. (0 CBS, n = 8; 1 CBS inside, n = 23; 2 CBS inside, n = 18; 3 CBS inside, n = 13; 4 CBS inside, n = 5; Data are mean \pm sd). **d**, Allele-specific *Sox2* expression in the no insertion clone (n = 2), the 4CBS clone (n = 3), and the 4CBS downstream clone (n = 2) as measured by RNA-seq. *Sox2* expression from the CAST and 129 allele was represented by normalized read counts (rpkm) of the tagged *egfp* and *mcherry* gene, respectively. **e**, Relative *Sox2* expression in the 4CBS and the 4CBS downstream clone in **d** measured by RNA-seq and FACS. The *Sox2* expression from the *egfp* allele was first normalized to the *mcherry* allele, then compared to the no insertion clone. **f**, FACS profiling of the no insertion clone and the 4CBS clone. **g**, FACS profiling of GFP^{low}, GFP^{high} sub-populations, and the unsort total population of the 4CBS insertion clone in **f** after extended culturing for 8 days. Left, GFP signal, right, mCherry signal from the same cells. **h**, ChIP-seq of H3K4me3 and H3K27ac in the no insertion clone and the 4CBS clone (n = 2). **i**, Allelic quantification of H3K27ac signal on the *Sox2* super-enhancer of clones in **h**. H3K27ac ChIP-seq reads on the *Sox2* super-enhancer were normalized by the total reads mapped to chromosome 3 for each allele.



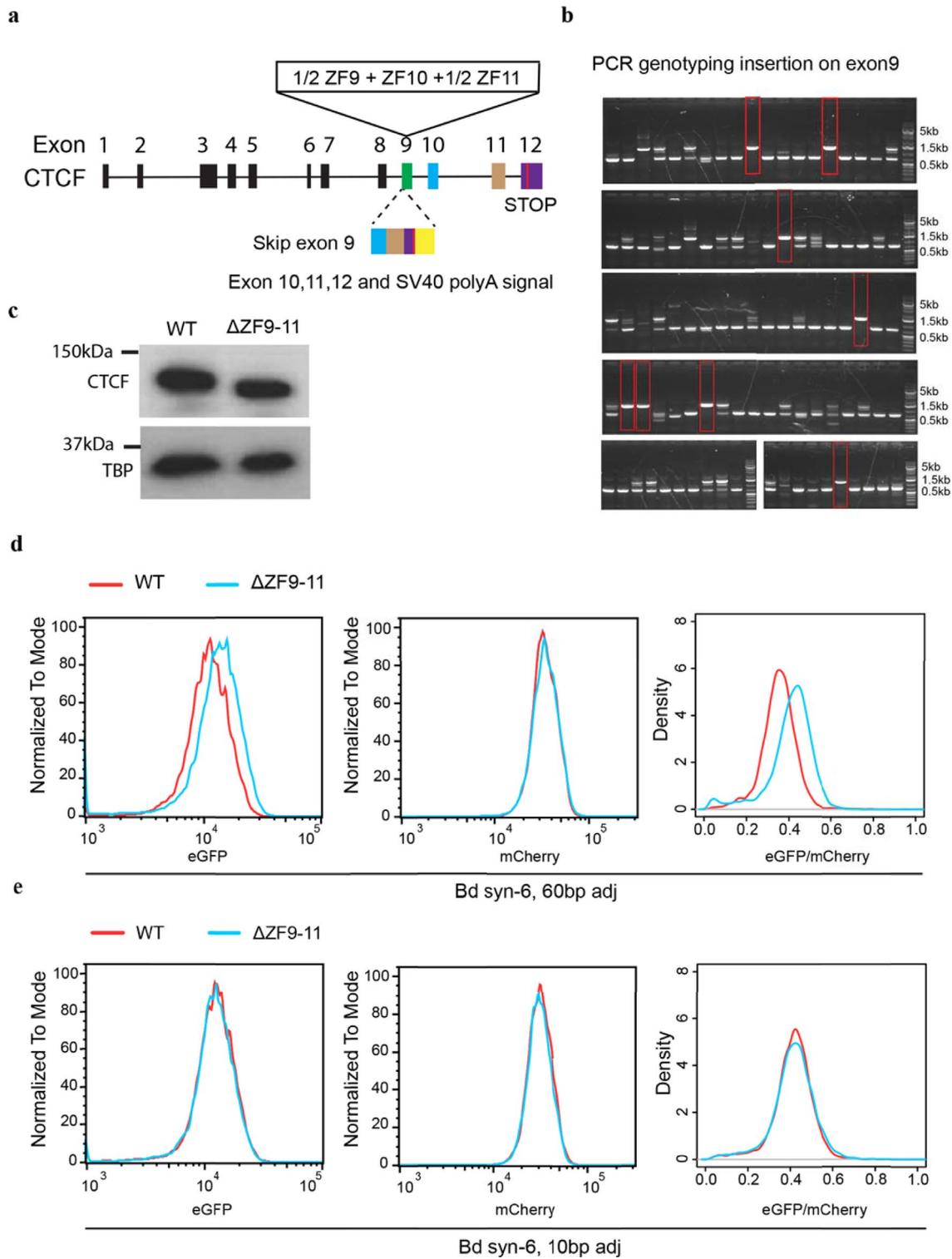
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Insulation effects of synthetic CTCF binding sites. **a**, Additive insulation by synthetic CBS from boundary regions. Left top, compositions of one 139bp-CBS that was synthesized; Left bottom, tandemly arrayed 139bp-CBSs tested for insulator activity. Right, normalized Sox2-eGFP expression of clones with the tandemly arrayed 139bp-CBSs inserted between the Sox2 gene and its super-enhancer. Blue, CBS core motifs were in forward orientation; Red, CBS core motifs were in reverse orientation. Insertions were on the CAST allele only. $n=3$, unpaired t-test, two-tailed. Data are mean \pm sd. **b**, Insulation effects of PCR cloned large size CBSs (1–4 kb) and the synthesized 139bp-CBSs that contain the same CTCF motifs. ($n=12$, paired t-test, two-tailed, $***P=0.0007$.). **c**, CTCF binding strength at selected boundary sites and non-boundary sites in mouse ES cells. CHIP-seq signals of CTCF are shown in 2-kb window. **d**, ChIP-seq of CTCF and Rad21 in clones with the insertion of six (nBd-syn6) or fifteen (nBd-syn15) 139-bp CBSs obtained from non-boundary regions. ChIP-seq reads were mapped to a customized mm10 genome that included the inserted sequence at the target site. Insertion position is highlighted in the red box.



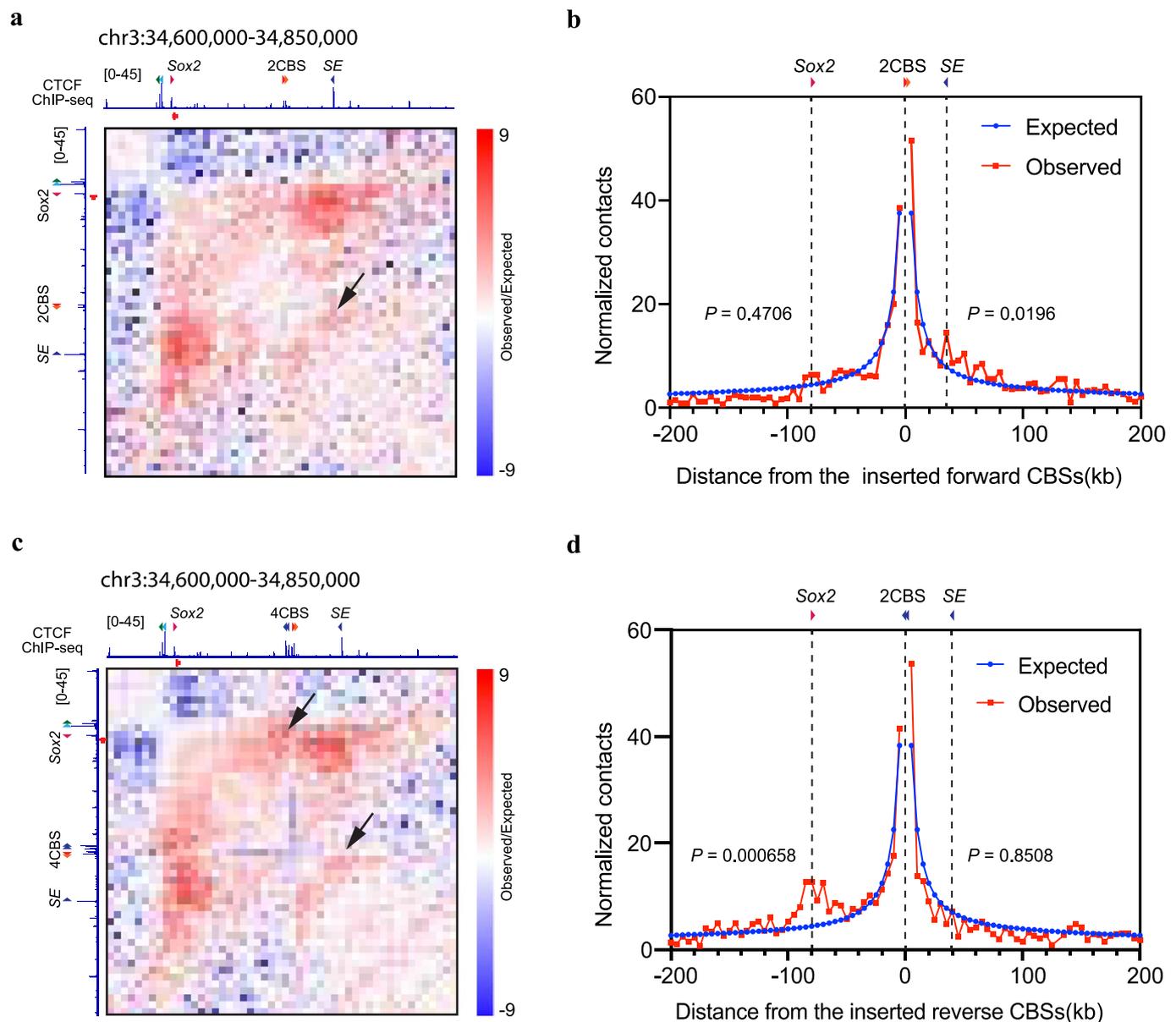
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | ChIP-seq analysis of CTCF and cohesin binding at the synthetic insulators in various insulator reporter clones. **a**, ChIP-seq signal tracks of CTCF in clones with the insertion of different synthetic CBS variants ($n=2$). Each insertion consists of six CBSs that were tandemly arrayed in forward orientation (Supplementary Table 4). The insertion location is highlighted in the yellow box. **b**, CTCF binding strength (counts per million uniquely mapped reads) at the insertion location in the clones in **(a)**. For each clone, ChIP-seq reads were mapped to a specific customized genome that contains the corresponding insertion in the *Sox2* locus ($n=2$; for bd core with bd adj, $n=4$; for nbd core with bd adj $n=3$.). **c**, ChIP-seq signal tracks of Rad21 in the same clones in **(a)** ($n=2$). The insertion location is highlighted in the yellow box. **d**, Rad21 binding strength (counts per million uniquely mapped reads) at the insertion location in the clones in **(c)**. For each clone, ChIP-seq reads were mapped to a specific customized genome that contains the corresponding insertion in the *Sox2* locus ($n=2$). **e**, Sequence alignment of the six boundary CBSs. Each CBS consists of 19-bp core motif plus 20-bp adjacent sequences on both sides. The color indicates the base frequency at each position. The CTCF motifs are highlighted in the red box.

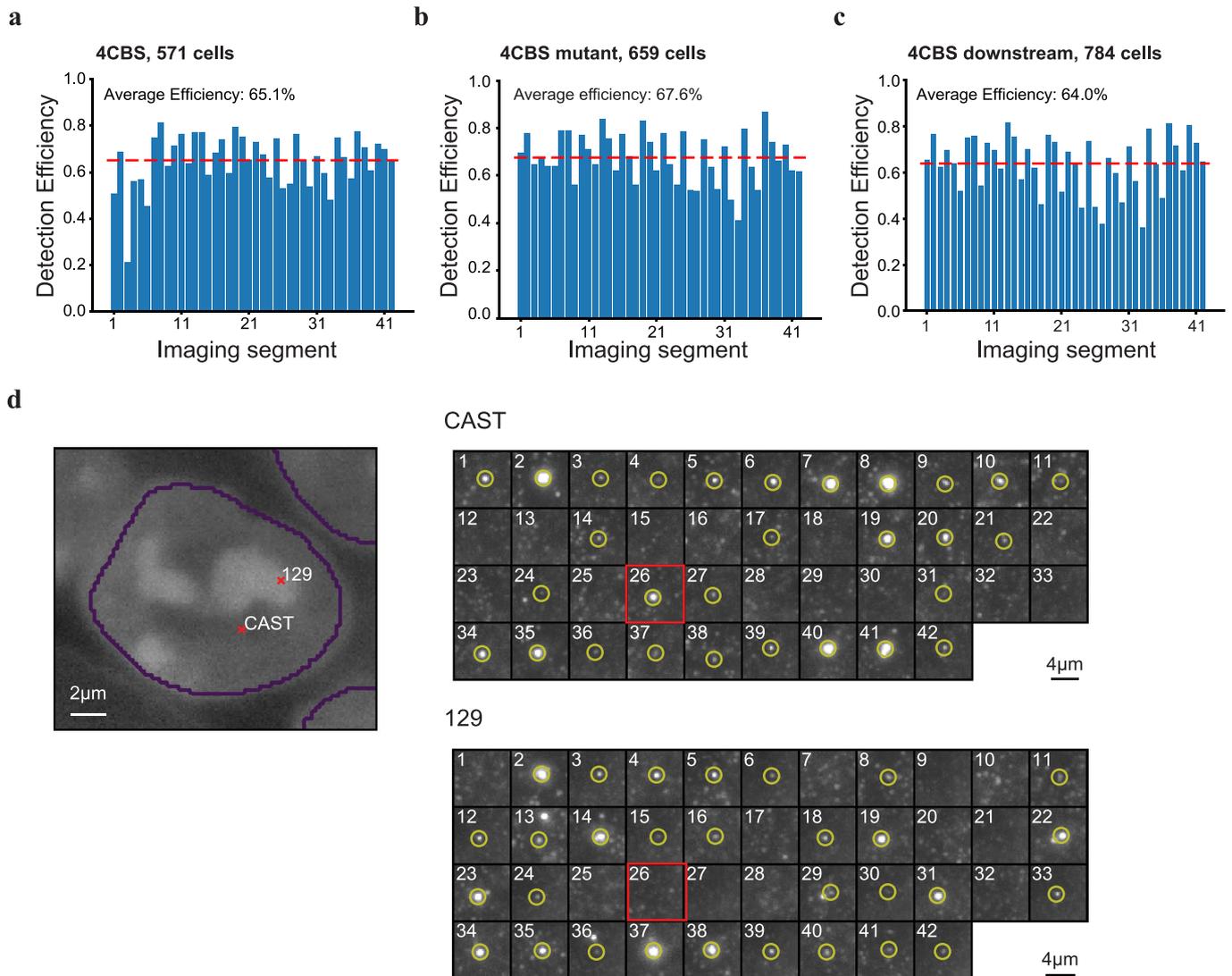


Extended Data Fig. 5 | See next page for caption.

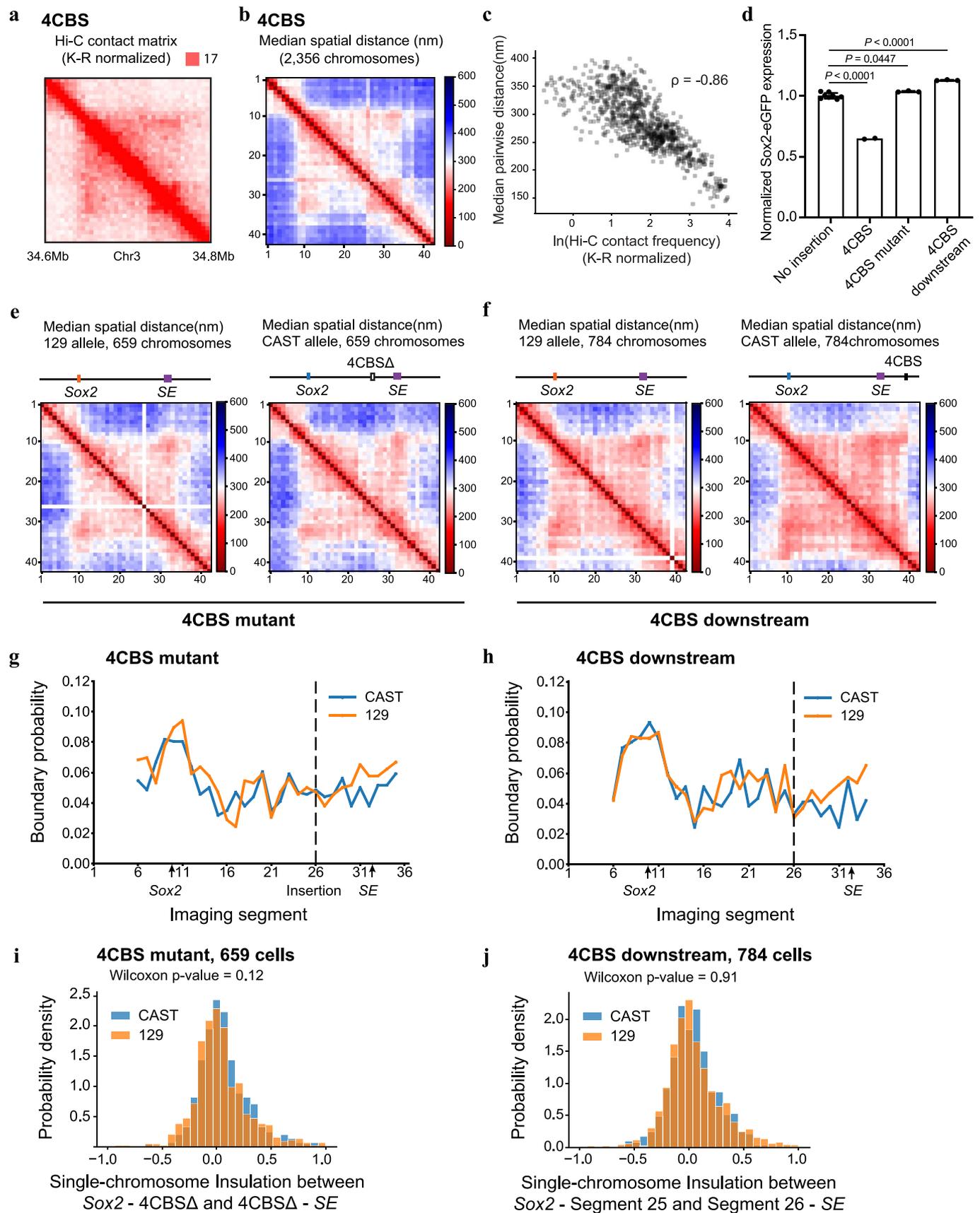
Extended Data Fig. 5 | Impact of CTCF ZF-9-11 deletion on transcriptional insulation by a synthetic insulator. **a**, A schematic shows the experimental design to delete ZF9-11 of CTCF in mESCs. The exon 10, 11, partial of exon12, and an SV40 polyA signal were inserted into exon 9, resulting in the skip of exon9 in mRNA of the CTCF gene. **b**, PCR genotyping CTCF Δ ZF9-11 mutant colonies. Genotyping primers spanned the insertion in exon 9. Highlighted in red boxes were homozygous mutant colonies evidenced by a single large-sized PCR fragment. PCR products from the homozygous mutant clones were further confirmed by Sanger sequencing. Genotyping of the homozygous mutants was repeated once with similar results. **c**, Western blot of CTCF in wild type and an exemplary CTCF Δ ZF9-11 mutant clone. The primary antibody was the same one used for CHIP-seq (catalog: ab70303, lot GR3281212-7). Bottom, TBP loading control (primary antibody: sc-421, lot #B0304). Western blot was repeated once with similar results. **d**, Impact of CTCF zinc fingers 9-11 deletion on insulation effects of boundary CBSs with sixty-base-pair adjacent sequences. Left, eGFP profile of exemplary clones expressing wild-type and mutant CTCF protein; middle, mCherry profile of the same cells; right, normalized eGFP signal (eGFP/mCherry) of the wild-type and mutant clones. **e**, Impact of CTCF zinc fingers 9-11 deletion on insulation effects of boundary CBSs with ten-base-pair adjacent sequences. Left, eGFP profile of exemplary clones expressing wild-type and mutant CTCF protein; middle, mCherry profile of the same cells; right, normalized eGFP signal (eGFP/mCherry) of the wild-type and mutant clones.



Extended Data Fig. 6 | Chromatin contacts at inserted CBSs. **a**, K-R normalized Hi-C matrix (Observed/Expected) in the clone with the two forward *Sox9-Kcnj2* TAD boundary CBSs inserted between the *Sox2* promoter and super-enhancer ($n=2$, replicates were merged). Hi-C reads were mapped to a customized chromosome 3 containing the insertion of the two forward CBSs. ChIP-seq signal of CTCF and orientations of the inserted CBSs and CBSs around the *Sox2* promoter and super-enhancer were shown. The black arrow indicates the interactions between the inserted CBSs and the CBS on the *Sox2* super-enhancer. **b**, Virtual 4C derived from Hi-C contacts in (**a**) at the viewpoint of the two inserted CBSs. Contacts were counted in each 5kb-bin. Contacts between the inserted CBSs and the *Sox2* promoter or super-enhancer were compared to expected values, two-sided Poisson test. **c**, K-R normalized Hi-C matrix (Observed/Expected) in the clone with the four *Sox9-Kcnj2* TAD boundary CBSs inserted between the *Sox2* promoter and super-enhancer ($n=2$, replicates were merged). Hi-C reads were mapped to a customized chromosome 3 containing the insertion of the four CBSs. ChIP-seq signal of CTCF and orientations of the inserted CBSs and CBSs around the *Sox2* promoter and super-enhancer were shown. The black arrows indicate the interactions between the inserted CBSs and the CBS on the *Sox2* promoter and super-enhancer. **d**, Virtual 4C derived from Hi-C contacts in (**c**) at the viewpoint of the two reverse-orientated CBSs inserted between the *Sox2* promoter and super-enhancer. Contacts were counted in each 5kb-bin. Contacts between the two reverse CBSs and the *Sox2* promoter or super-enhancer were compared to expected values, two-sided Poisson test.

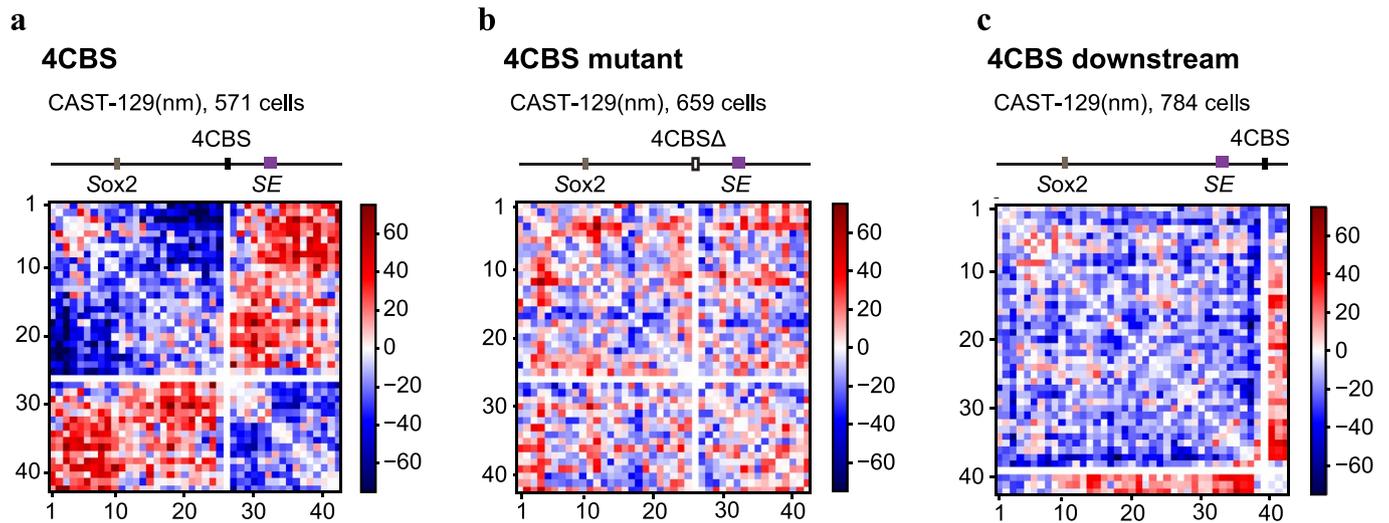


Extended Data Fig. 7 | Allele classification by multiplexed DNA FISH. **a-c**, Bar plots showing detect efficiency of the 42 segments of chromatin tracing experiments in the ‘4CBS’ clone (**a**), the ‘4CBS mutant’ clone (**b**), and the ‘4CBS downstream’ clone (**c**). Detect efficiency of each segment was calculated as the fraction of chromosomes that showed a positive fluorescence signal at the specific imaging round. **d**, Exemplary images of allele classification. Left, nuclei segmentation and the positions of CAST and 129 allele in the nucleus. Right, images of the forty-two 5-kb segments (chr3:34,601,078-34,811,078) of the CAST and 129 allele. The hybridization probes of the 26th segment (highlighted in the red box) specifically targeted the 4CBS sequence. The chromosome positive for the 26th segment (inserted 4CBS) was classified as CAST allele, the negative chromosome in the same cell was classified as 129 allele. Cells with both chromosomes positive or both chromosomes negative for the 26th segment were discarded.

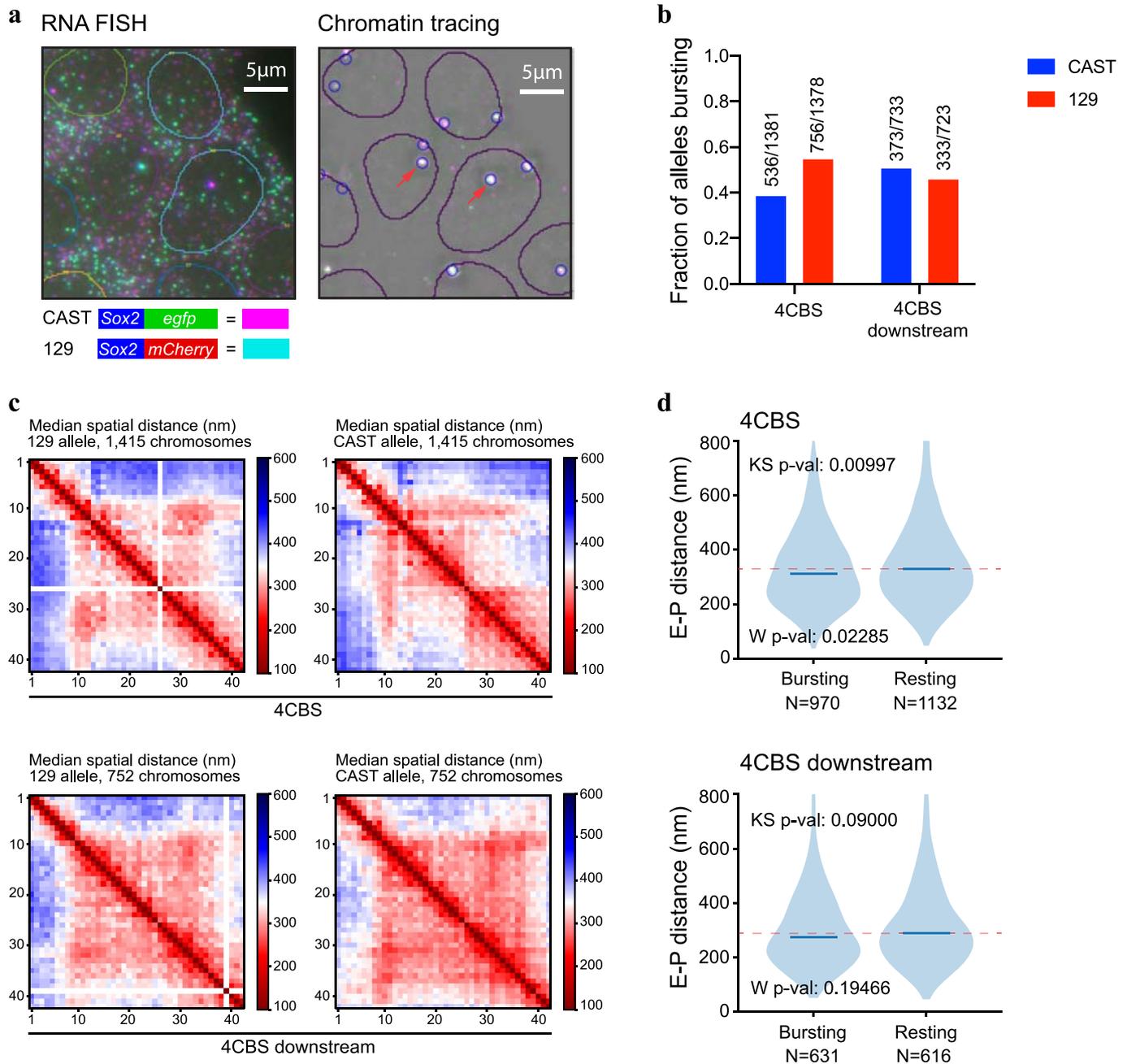


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Spatial organization of the Sox2 locus in engineered mESCs. **a**, Bulk Hi-C contact matrix (K-R normalized) of the Sox2 locus in the 4CBS clone. **b**, Median pairwise distance of the same Sox2 region measured by chromatin tracing experiment in the same clone in **a**. CAST and 129 chromosomes were combined. **c**, Correlation between the Hi-C contact frequency matrix (**a**) and median distance matrix(**b**). **d**, Normalized Sox2-eGFP expression in the no insertion clone(n=8), the '4CBS' clone (same cells in **a-b**, n=2), and two insertion controls, '4CBS mutant' (n=3) and '4CBS downstream' (n=3). One-way analysis of variance with Bonferroni's multiple comparisons test. Data are mean \pm sd. **e-f**, Median spatial-distance matrix for the 210 kb Sox2 region (chr3: 34601078-34811078) of 129 (left) and CAST (right) chromosomes of the '4CBS mutant' clone(**e**) and the '4CBS downstream clone'(**f**). The 26th segment was imaged by 4CBS specific probes in **e**. Similarly, the 38th segment was imaged by 4CBS specific probes in **f**. **g-h**, The probability of forming single-chromosome domain boundaries at each segment for the two alleles of the '4CBS mutant' clone (**g**), and the '4CBS downstream' clone (**h**). **i**, The distribution of single-chromosome insulation scores for each of the alleles between Sox2 promoter - 4CBS Δ insertion (segments 10-25) and 4CBS Δ insertion - Sox2 super-enhancer (segments 26-33), respectively. Two-sided Wilcoxon rank-sum test was performed. **j**, The distribution of single-chromosome insulation scores for each of the alleles between the same two domains (segment 10-25 and segment 26-33) in (**i**) for the '4CBS downstream' clone. Insulation score was calculated in the same way as in (**i**). Two-sided Wilcoxon rank-sum test was performed.



Extended Data Fig. 9 | Allele differences in median spatial distance. a-c, Difference of the median distance matrices between the CAST and 129 allele of the '4CBS' clone (**a**), the '4CBS mutant' clone (**b**) and the '4CBS downstream' clone (**c**).



Extended Data Fig. 10 | Imaging of both nascent transcripts and chromatin structure at the *Sox2* locus. **a, Example images of RNA FISH(left) and chromatin tracing(right) in the same cells. Circles indicate individual nuclei. RNA probes targeting *Sox2*, *egfp*, and *mCherry* are color-coded. Transcripts from the CAST allele are indicated by dots in purple pseudo color. Transcripts from the 129 allele are indicated by dots in cyan pseudo color. Arrows highlight examples of bursting alleles. **b**, Bursting frequencies of the CAST and 129 allele in the 4CBS clone and the control clone with 4CBS inserted downstream of the *Sox2* SE. The numbers of bursting chromosomes and total chromosomes are indicated on the top of each bar. **c**, Median spatial distance matrices of the CAST and 129 allele in the 4CBS and the control clone (4CBS inserted downstream of the *Sox2* super-enhancer). Multiplexed DNA FISH experiments were performed in the same cells following the RNA FISH experiments. **d**, Enhancer-promoter distances of the bursting and resting chromosomes in the 4CBS and the 4CBS downstream clone. A two-sided KS test between the distributions and a two-sided Wilcoxon test were performed.**