Jia et al. Genome Biology (2020) 21:75
https://doi.org/10.1186/s13059-020-01984-7

Genome Biology

RESEARCH Open Access

Tandem CTCF sites function as insulators to balance spatial chromatin contacts and topological enhancer-promoter selection

Zhilian Jia[1†], Jingwei Li[1†], Xiao Ge[1†], Yonghu Wu[1], Ya Guo[1] and Qiang Wu[1,2*]

Abstract

Background: CTCF is a key insulator-binding protein, and mammalian genomes contain numerous CTCF sites, many of which are organized in tandem.

Results: Using CRISPR DNA-fragment editing, in conjunction with chromosome conformation capture, we find that CTCF sites, if located between enhancers and promoters in the protocadherin (*Pcdh*) and *β-globin* clusters, function as an enhancer-blocking insulator by forming distinct directional chromatin loops, regardless whether enhancers contain CTCF sites or not. Moreover, computational simulation in silico and genetic deletions in vivo as well as dCas9 blocking in vitro revealed balanced promoter usage in cell populations and stochastic monoallelic expression in single cells by large arrays of tandem CTCF sites in the *Pcdh* and immunoglobulin heavy chain (*Igh*) clusters. Furthermore, CTCF insulators promote, counter-intuitively, long-range chromatin interactions with distal directional CTCF sites, consistent with the cohesin "loop extrusion" model. Finally, gene expression levels are negatively correlated with CTCF insulators located between enhancers and promoters on a genome-wide scale. Thus, single CTCF insulators ensure proper enhancer insulation and promoter activation while tandem CTCF topological insulators determine balanced spatial contacts and promoter choice.

Conclusions: These findings have interesting implications on the role of topological chromatin insulators in 3D genome folding and developmental gene regulation.

Keywords: CTCF, Insulator, Promoter/enhancer selection, 3D genome, Gene regulation, Loop extrusion, Cohesin, Chromatin polymer simulation, Bayesian networks, Topological spatial contacts

Background

Genetic studies have long described the phenomenon of position effect variegation (PEV) [1], suggesting that the spatial organization of chromatin domains has an important influence on gene expression [2–4]. Early studies revealed that boundary elements, also known as insulators, restrict promoter activity from the position effects of its chromatin contexts [5, 6]. In particular, through a series of transgenic experiments, Grosveld and colleagues have identified dominant boundary elements flanking the human *β-globin* locus, which determine its position-independent expression in transgenic mice [5]. It has since been established that insulators play an essential role in shielding the position effects of chromatin conformation and in blocking enhancers or silencers from improperly activating or repressing non-cognate promoters, respectively [2, 3, 6–8].

* Correspondence: qiangwu@sjtu.edu.cn; qwu123@gmail.com
†Zhilian Jia, Jingwei Li and Xiao Ge contributed equally to this work.
[1]MOE Key Lab of Systems Biomedicine, Center for Comparative Biomedicine, State Key Lab of Oncogenes and Related Genes, Shanghai Cancer Institute, Joint International Research Laboratory of Metabolic & Developmental Sciences, Institute of Systems Biomedicine, Xin Hua Hospital, Shanghai Jiao Tong University, Shanghai 200240, China
[2]The Third Affiliated Hospital of Guangzhou Medical University, Guangzhou 510150, China

The mammalian CTCF is the best characterized genome architectural protein that binds to insulator elements [2, 8]. CTCF directionally and dynamically binds to tens of thousands of CTCF-binding sites (CBS elements) in mammalian genomes through the combinatorial usage of its 11 zinc fingers [9, 10]. CTCF, together with the associated cohesin, a ring-shaped complex embracing DNA, mediates genome-wide long-range chromatin interactions [2]. Interestingly, these interactions are preferentially formed between forward-reverse convergent CBS pairs [11–14]. The CBS elements in the boundaries between neighboring chromatin domains are configured in a reverse-forward divergent orientation, which are thought to restrict enhancer activity to promoters within each insulated neighborhood [12, 15, 16]. Thus, the boundary CBS elements may function as insulators to block cohesin loop extrusion [11, 12, 17–20]. However, whether and how internal CBS elements function as insulators remain incompletely understood.

Recent topologically associated domain (TAD) perturbations by targeted degradation of CTCF or cohesin revealed that loss of chromatin loops genome-wide differentially affect gene expression [21, 22]. Numerous studies have shown that CTCF/cohesin-mediated chromatin loop domains or TADs are important for gene regulation in specific loci [12, 15, 16, 23]. Insertion, mutation, deletion, inversion, or duplication of CBS elements alters chromatin topology and gene expression [12, 14–16, 18, 23–25]. Emerging evidence suggests that spatial control of genome topology by CTCF/cohesin regulates gene expression; however, how numerous CBS elements in mammalian genomes function as insulators to control proper promoter activation and its balanced usage remains obscure.

Similar to the enormous diversity of DSCAM1 proteins in *Drosophila*, combinatorial *cis*- and *trans*-interactions between clustered cell surface protocadherin (Pcdh) proteins in mammals, encoded by the three closely linked *α*, *β*, and *γ* gene clusters (Fig. 1a in mice), endow individual neurons with a unique identity code and specific self-recognition module, which are required for neuronal migration and connectivity, dendrite self-avoidance and tiling, and axon outgrowth and even spacing in the brain [26–32]. The *Pcdh α* and *γ* clusters contain more than a dozen highly similar, tandem-arrayed, unusually large "alternate" variable exons and 2 or 3 divergent C-type variable exons, respectively (Fig. 1a). These variable exons are followed by 3 downstream small constant exons, reminiscent of the variable and constant genome organizations of immunoglobulin (*Ig*), T cell receptor (*Tcr*), and UDP-glucuronosyltransferase (*Ugt*) clusters [26, 28, 33]. Each of the *Pcdhα* "alternate" variable exons (*α1-α12* in mice) carries its own promoter, which is flanked by two forward-oriented CBS elements (Fig. 1a). However, the *αc1* promoter carries only one forward-oriented CBS, and the *αc2* promoter has no CBS element (Fig. 1a). Two distal *Pcdhα* enhancers, *HS7* and *HS5-1*, are located downstream, and one of which, *HS5-1*, is flanked by two reverse-oriented CBS (*HS5-1a* and *HS5-1b*) elements [34, 35]. Multiple long-distance chromatin interactions between these remote enhancers and *Pcdhα* target promoters form a transcription hub and determine the promoter choice, but the underlying mechanisms are unknown [35, 36].

Here, by a combination of CRISPR DNA-fragment editing [37, 38] and chromosome conformation capture [3] experiments as well as Bayesian modeling, we show that ectopic and endogenous CTCF sites function as topological insulators in an orientation-independent manner through CTCF-mediated directional chromatin looping throughout the mammalian genome. In addition, genetic experiments, in conjunction with computational polymer simulation of cohesin loop extrusion, demonstrate that tandem-arrayed CTCF sites ensure stochastic spatial accessibility of repertoires of promoters and their balanced usage.

Results

Exogenous directional CTCF sites function as protocadherin insulators in cellular model in vitro

To investigate the mechanisms of cell-specific *Pcdh* gene expression in the brain, we performed single-cell RNA-seq of mouse cortical neurons and found members of the *Pcdhα* cluster are expressed in single neurons in a combinatorial and stochastic manner (Fig. 1b), similar to the stochastic monoallelic expression patterns of *Pcdhα* in single Purkinje cells in the cerebellum [28, 39]. In addition, maximum likelihood modeling confirms the stochastic monoallelic expression patterns in single cells of the mouse neocortex (Additional file 1: Figure S1a,b) [40].

iWe next made use of the HEC-1-B cell line, which monoallelically expresses *α6* and *α12* (Additional file 1: Figure S1c-f; note that humans have 13 alternate variable exons), as a single-cell model system to investigate mechanisms of gene regulation [12]. We performed CBS insertions by DNA-fragment editing and screened for single-cell CRISPR clones (Fig. 1c, d) [37, 38]. We first nserted single ("F") or tandem ("FF") forward-oriented CBS elements into the location between the *Pcdhα* cluster and its *HS5-1* enhancer (Fig. 1d–f) and carried out quantitative high-resolution chromosome conformation capture copy followed by next-generation sequencing (QHR-4C) experiments (Additional file 1: Figure S2). QHR-4C revealed prominent long-distance chromatin interactions between *HS5-1* and the inserted CBS elements, and a concurrent decrease of chromatin interactions between *HS5-1* and the *Pcdhα* promoters (Fig. 1g–j and Additional file 1: Figure S3a,b). In addition, CBS mutations abolish these effects (Fig. 1g–j and Additional file 1: Figure S3a,b). Consistent

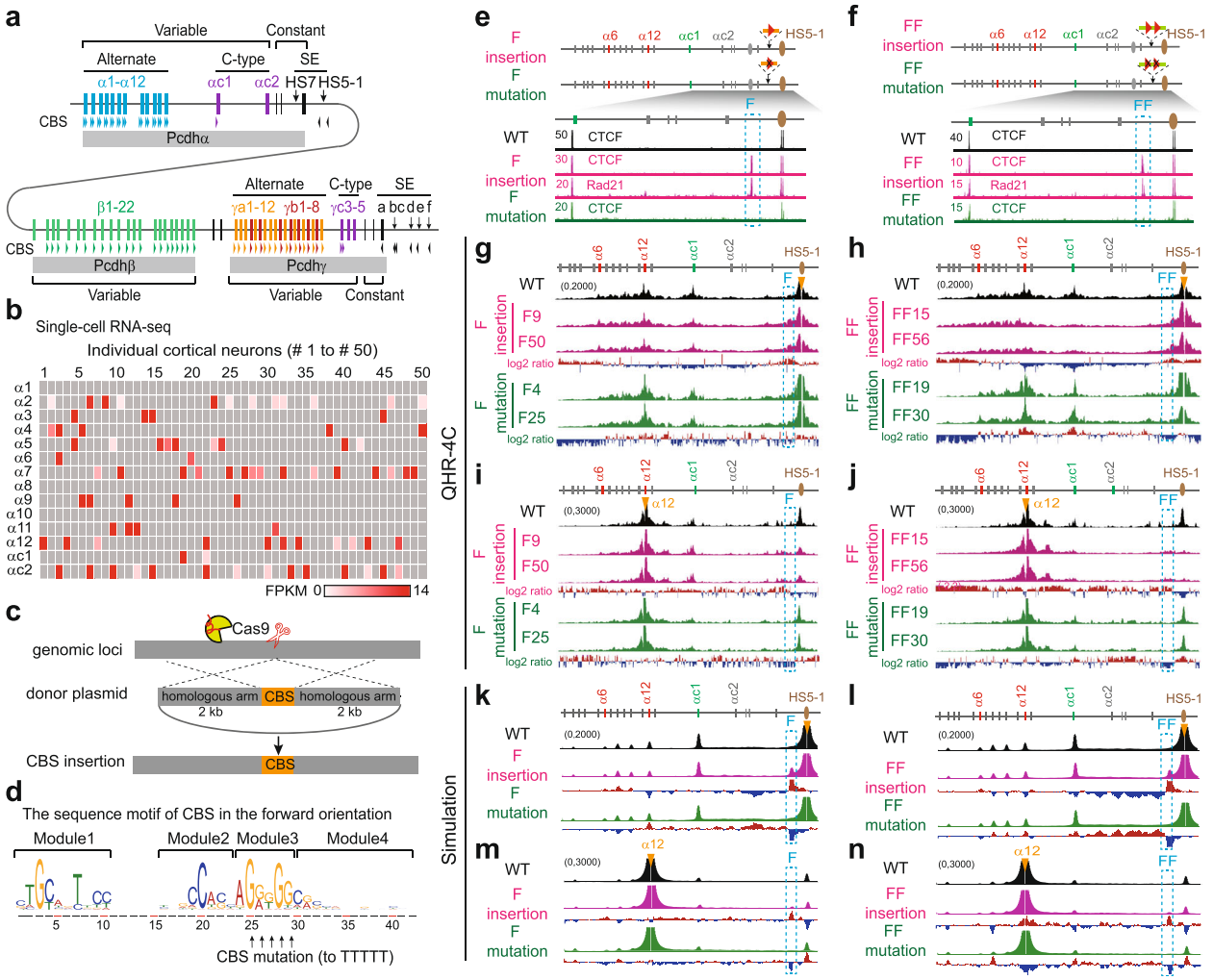

Fig. 1 CTCF sites as insulators in *Pcdha* through directional chromatin looping. **a** Schematic of the three mouse *Pcdh* clusters. Genomic organizations of the *Pcdh α* and *γ* clusters are similar. Both contain variable exons, each of which is preceded by its own promoter and spliced to respective downstream constant exons. The *Pcdhβ* cluster contains only variable exons. Each of the variable exons (except *αc2*, *β1*, *γc4*, and *γc5*) carries one or two forward-oriented CTCF sites (indicated by tandem arrowheads). Two distal enhancers, *HS7* (with no CBS) and *HS5-1* (flanked by two reverse-oriented CBS elements of *HS5-1a* and *HS5-1b*), are located between the *Pcdh α* and *β* clusters. The super-enhancer of the *Pcdh β* and *γ* clusters, which contains a tandem array of reverse-oriented CBS elements, is located downstream of *Pcdhγ*. CBS, CTCF binding site; HS, hypersensitive site; SE, super-enhancer. **b** Single-cell expression patterns of the *Pcdha* genes in mouse cortical cells. **c** CRISPR insertion of CBS elements by homologous recombination. **d** Schematic of the sequence motif (CTCF binding site contains four modules in the order of modules 1–4 for the forward orientation) and its mutation. **e**, **f** CTCF and Rad21 ChIP-seq of single-cell CRISPR clones with one or two forward-oriented CBS elements inserted into the location between the *Pcdha* cluster and its downstream *HS5-1* enhancer. **g**–**j** QHR-4C interaction profiles of four CRISPR clones (F9 and F50 for one-CBS insertion; FF15 and FF56 for two-CBS insertion) as well as their CBS mutations (F4 and F25; FF19 and FF30) with *HS5-1* (**g**, **h**) or *α12* (**i**, **j**) as a viewpoint (VP), represented by arrowheads. **k**–**n** Chromatin interaction profiles from an ensemble of 350,000 simulated conformations with *HS5-1* (**k**, **l**) or *α12* (**m**, **n**) as a viewpoint, corresponding to **g**–**j**, respectively. Log2 ratios (insertion vs wild-type or mutation vs insertion) are shown under the 4C profiles

with the decrease of enhancer-promoter interactions, RNA-seq revealed a significant decrease of *α6* and *α12* expression levels, and CBS mutations rescue their expression (Additional file 1: Figure S3c,d). In summary, the inserted forward-oriented CBS elements block the long-distance chromatin spatial contacts between the *HS5-1* enhancer and its target promoters and thus function as chromatin insulators by competing with the target *Pcdhα* promoters.

We next inserted three different reverse-oriented CBS elements each into distinct locations in the *Pcdhα* cluster (Additional file 1: Figures S3e-j and S4). We found that each competes with the *HS5-1* enhancer to form long-distance chromatin interactions with target promoters and thus functions as an insulator (Additional file 1: Figures S3e-j and S4). Finally, we inserted reverse-forward CBS pairs ("RF" or "RRFF") into the

location between the *Pcdhα* cluster and the *HS5-1* enhancer. We found that they also function as insulators (Additional file 1: Figures S5 and S6).

Forward-reverse CTCF sites do not compromise insulation activity

Previous studies demonstrated that *Drosophila* paired insulators compromise the insulation activity of each other [41, 42]. To test the orientation of mammalian insulators, we inserted four tandem CBS elements in a forward-reverse configuration between the *Pcdhα* cluster and its *HS5-1* enhancer (Additional file 1: Figure S7a). We found, surprisingly, these inward forward-reverse CBS elements still function as insulators. Specifically, QHR-4C and RNA-seq revealed a significant decrease of chromatin interactions between *HS5-1* and the *Pcdhα* promoters as well as their decreased expression (Additional file 1: Figure S7b-f). This suggests that, different from fly insulators, the mammalian forward-reverse tandem CTCF sites do not compromise their insulation activities. As a control, the inserted outward reverse-forward boundary CBS elements function as insulators as expected (Additional file 1: Figures S5 and S6).

We conclude that both forward and reverse ectopic CBS elements function as insulators for the *Pcdhα* genes through CTCF-mediated directional looping (Fig. 1 and Additional file 1: Figures S3-S7), namely, CTCF insulators function in an orientation-independent manner. However, their insulation mechanisms are distinct. The forward or reverse CBS elements form long-distance chromatin interactions with the *Pcdhα* enhancers or promoters (presumably by cohesin sliding through the oncoming convergent CTCF sites, Additional file 1: Figure S7b,c), respectively, in an orientation-dependent manner. Thus, the relative locations and orientations of inserted CBS elements determine their insulation specificity through directional looping to distinct CTCF sites in the *Pcdhα* cluster.

CTCF insulators enhance distal promoter usage

Interestingly, the inserted CTCF insulators mainly block enhancer contacts with the proximal *Pcdhα* promoters (Fig. 1g, h and Additional file 1: Figures S3f, S5b, S6b, and S7b). Surprisingly, the insertion of CTCF insulators augments long-distance chromatin interactions between the *HS5-1* enhancer and the distal *Pcdhα* promoters (Fig. 1g, h and Additional file 1: Figures S3f, S5b, S6b, and S7b). To understand this puzzling phenomenon, we simulated polymer conformation dynamics of the *Pcdhα* cluster by "two-headed" cohesin loop extrusion on a coarse-grained chromatin fiber (Additional file 1: Figure S7g), based on the locations and relative orientations of the CBS elements that are dynamically bound by CTCF proteins (Additional file 1: Figure S8a-c) [9, 10, 12, 18–20].

We assume that cohesin slides along the *Pcdhα* chromatin fiber until it encounters an opposite CBS element or another sliding cohesin (Additional file 1: Figure S7g) [18, 19, 43]. Remarkably, computational 3D polymer simulations revealed that, in addition to proximal *Pcdhα* promoter insulation, continuous cohesin extrusion of chromatin loops results in a significant increase of chromatin interactions between the *HS5-1* enhancer and the distal *Pcdhα* promoters upon insertions of various CTCF insulators (Fig. 1k, l and Additional file 1: Figures S3i, S5f, S6f, and S7e), consistent with the observed data from the QHR-4C experiments (Fig. 1g, h and Additional file 1: Figures S3f, S5b, S6b, and S7b). Finally, by applying the relative maximum entropy approach with independent Gaussian errors, we optimized our polymer simulations and obtained strong evidence that CTCF insulators promote distal chromatin interactions (Additional file 1: Figure S8d).

We next simulated chromosome conformation of the *Igh* cluster which also contains a large repertoire of tandem variable CTCF sites (Additional file 1: Figure S8e,f) [33] and found that, similar to that in the *Pcdhα* cluster, insertion of various CTCF insulators in different orientations also augments distal variable gene segment (V_H) utilization (Additional file 1: Figure S8f,g). Thus, CTCF-mediated directional looping of tandem-arrayed CBS elements determines the promoter balance of both *Pcdhα* and *Igh* gene clusters.

Topological looping of distal-to-distal CTCF sites in the *Pcdh β/γ* clusters

Similar to the *Pcdhα* cluster, the promoter of each member of the *Pcdh β* and *γ* clusters (except *β1*, *γc4*, and *γc5*) carries a forward CBS, and the downstream super-enhancer contains a tandem array of reverse-oriented CBS elements (Fig. 1a) [12, 44]. It is not clear how members of the *Pcdh β* and *γ* clusters are regulated by these tandem reverse CBS elements. Single-cell RNA-seq and maximum likelihood modeling demonstrated that single cortical neurons express random combinations of roughly up to 4 isoforms of the *Pcdhβ* family and 4 isoforms of the *Pcdhγ* family in the mouse brain (Fig. 2a and Additional file 1: Figure S9a). However, the deletion of CTCF sites *b-e* in the super-enhancer mainly impairs the expression of members of the *Pcdhβ* cluster in single cells in the mouse cortex (Fig. 2b compared with Fig. 2a).

To investigate whether tandem CTCF sites in the *Pcdh β* and *γ* clusters and their downstream super-enhancer also balance spatial chromatin contacts and promoter choice, we performed QHR-4C experiments with a repertoire of the *Pcdh β* and *γ* promoters as a viewpoint using mouse cortical tissues (Fig. 2c, d). Remarkably, the

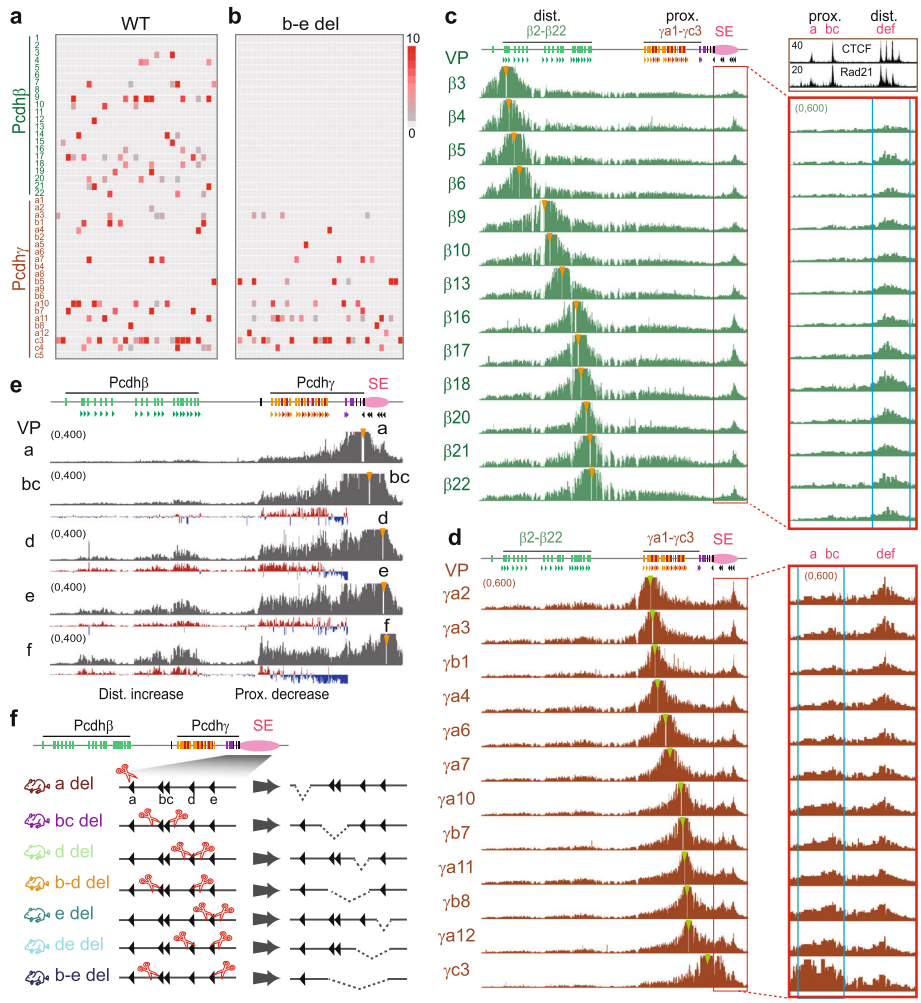

Fig. 2 Tandem CTCF sites balance the usage of *Pcdh β* and *γ* promoters. **a**, **b** Single-cell RNA-seq of cortical neurons of the WT (**a**) and *CBS b-e* (**b**) deletion mice. Note that the absence of *Pcdhβ* expression in single cortical neurons of the CTCF sites *b-e* deletion mice. **c** QHR-4C profiles with a repertoire of the *Pcdhβ* promoters as a viewpoint show that they form spatial chromatin contacts with the distal CTCF sites *d-f*, but not proximal CTCF sites *a-c*, within the super-enhancer in the mouse cortical tissues. Inset in the upper right corner, ChIP-seq with a specific antibody against CTCF or Rad21. **d** QHR-4C profiles with a repertoire of the *Pcdhγ* promoters as a viewpoint show that, in addition to distal CTCF sites, they form gradually increased spatial chromatin contacts with the proximal CTCF sites *a-c* in the super-enhancer in the mouse cortical tissues. **e** QHR-4C interaction profiles with a repertoire of increasingly distal CTCF sites in the super-enhancer as a viewpoint show increased spatial chromatin contacts with the *Pcdhβ* cluster. **f** Schematic of the deletions of individual CTCF sites or their combinations in the *Pcdh β* and *γ* super-enhancer in mice. SE, super-enhancer; del, deletion

regulation of the *Pcdh β* and *γ* promoters appears topological. Namely, there are specific long-distance chromatin interactions between members of the *Pcdhβ* cluster and the distal CTCF sites *d-f*, but not proximal CTCF sites *a-c* (despite that all six CTCF sites *a-f* are bound by CTCF and cohesin, inset in the upper right corner of Fig. 2c), in the downstream super-enhancer (Fig. 2c). However, when using a repertoire of the *Pcdhγ* promoters as a viewpoint, in addition to the distal CTCF sites *d-f*, there appear increased spatial chromatin contacts with the proximal CTCF sites *a-c* of the downstream super-enhancer (Fig. 2d). Finally, to confirm this spatial regulation of the *Pcdh β* and *γ* promoters, we performed QHR-4C experiments with each of the super-enhancer CBS repertoire as a viewpoint and found increased long-range chromatin interactions between distal forward CTCF sites of the *Pcdh* variable promoters and distal reverse CTCF sites of the super-enhancer (Fig. 2e). Therefore, members of the *Pcdh β* and *γ* clusters are regulated topologically by the distal and proximal CTCF sites, respectively, within the downstream super-enhancer.

Tandem CTCF sites balance usage of *Pcdh β* and *γ* promoters

To further investigate the mechanism of tandem-arrayed CBS function in the super-enhancer, we generated a

series of deletions of individual CTCF sites or their combinations in mice (Fig. 2f and Additional file 1: Figure S9b). QHR-4C experiments revealed that deletions of these CTCF sites result in a significant increase of long-distance chromatin interactions between the *Pcdhγ* promoters and the super-enhancer, as well as a significant decrease of long-distance chromatin interactions between the *Pcdhβ* promoters and the super-enhancer (Fig. 3a and Additional file 1: Figures S10 and S11).

To pinpoint these topological effects to CTCF sites but not enhancers, we used catalytically inactive Cas9 (dCas9 for dead Cas9) CRISPR systems to specifically block each CBS within deletions without perturbing enhancers. QHR-4C experiments confirmed a significant increase with proximal *Pcdhγ* and a significant decrease with *Pcdhβ* (Fig. 3b). Finally, we confirmed this topological regulation in deletion mice and dCas9-blocking system by QHR-4C with the *Pcdhβ17* promoter as a viewpoint (Fig. 3c, d). We conclude that, similar to the *Pcdhα* and *Igh* clusters, endogenous tandem CTCF sites function as topological insulators to balance spatial enhancer contacts and promoter choice of the *Pcdh β* and *γ* clusters.

Endogenous CTCF sites function as protocadherin insulators

We next tested whether each of the endogenous tandem arrays of the forward-oriented *Pcdh* CBS elements functions as an insulator. We found that the deletion of the *αc1* CBS element results in a significant increase of long-distance chromatin interactions between *HS5-1* and the *Pcdhα* genes upstream of *αc1* (Fig. 4a, b). In addition, this deletion results in a significant increase of *α6* and *α12* expression levels (Fig. 4c). Moreover, the deletion of the *α12* CBS element also results in a significant increase of chromatin interactions between *HS5-1* and the upstream *Pcdhα* genes (Fig. 4d, e) as well as of the *α6* expression levels (Fig. 4f). Together, these data suggest that each endogenous CBS element functions as an insulator for its respective upstream *Pcdhα* genes.

To investigate whether each of the two reverse-oriented CBS elements (*HS5-1a* and *HS5-1b*) flanking the *HS5-1* enhancer also functions as an insulator, we deleted each of them in mice in vivo and performed 5C, QHR-4C, and RNA-seq experiments using mouse cortical tissues (Fig. 4g–k). Deletion of the *HS5-1b* CBS (Additional file 1: Figure S12a,b), which is at the boundary between the *Pcdhα* and *Pcdhβγ* subTADs [12], results in an aberrant increase of long-distance chromatin interactions between *HS5-1* and the 5′ isoforms of the *Pcdhβ* cluster (Fig. 4g, h) as well as an aberrant activation of their promoters (Fig. 4j). Remarkably, even for the *Pcdhβ1* promoter, which does not carry CBS, the long-distance chromatin interactions with *HS5-1* is still aberrantly increased, suggesting that *HS5-1b* CBS functions as an insulator to block the *HS5-1* enhancer from the improper activation of the *Pcdhβ1* promoter (Fig. 4h). By contrast, both the chromatin interactions of *HS5-1* with the proximal alternate *Pcdhα* genes as well as their expression levels are significantly decreased (Fig. 4g, i, j). This suggests that the boundary *HS5-1b* CBS element is an insulator that restricts the *HS5-1* enhancer activity from the aberrant activation of the *Pcdhβ* promoters. As a control, homozygous deletion of the internal *HS5-1a* CBS element (Additional file 1: Figure S12a,b) results in no expression alteration of the 5′ isoforms of the *Pcdhβ* cluster (Fig. 4k). Therefore, although both *HS5-1a* and *HS5-1b* CBS elements are required for bridging the *HS5-1* enhancer to the *Pcdhα* promoters (Fig. 4g, i–k), only the boundary *HS5-1b* CBS element functions as an insulator blocking the *HS5-1* enhancer activity from aberrantly activating the *Pcdhβ* genes.

Finally, to further investigate whether the insulation activity of the CBS *HS5-1b* is orientation-dependent, we generated a mouse line with the CBS *HS5-1b* inverted (Additional file 1: Figure S12a,b). Strikingly, neither the expression levels of 5′ isoforms of the *Pcdhβ* cluster nor their long-distance chromatin interactions with the *HS5-1* enhancer are significantly increased (Additional file 1: Figure S12c-e). By contrast, both expression levels of the proximal alternate *Pcdhα* genes and their long-distance chromatin interactions with the *HS5-1* enhancer are significantly decreased (Additional file 1: Figure S12c,f). Thus, the inverted CBS *HS5-1b* still functions as an insulator to block the *HS5-1* enhancer from improperly activating the *Pcdhβ* cluster but no longer is able to bridge the *HS5-1* enhancer with the proximal alternate *Pcdhα* genes. This again demonstrates that the insulation activity of CTCF insulators is orientation-independent, but the directional looping of CTCF sites is orientation-dependent. We conclude that both endogenous CTCF sites in the native genomic locations and inserted exogenous CTCF sites in ectopic locations function as insulators in an orientation-independent manner.

Insulators for *Pcdh* and *β-globin* enhancers with no CTCF site

We next prepared mice with a deletion of the entire *HS5-1* fragment including the two flanking CBS elements of *HS5-1a* and *HS5-1b* (Fig. 5a–e and Additional file 1: Figure S12a,b). We found that the long-distance chromatin interactions between the *HS7* enhancer and the 5′ isoforms of the *Pcdhβ* cluster are significantly increased upon the *HS5-1* deletion (Fig. 5a–c). In addition, the expression levels of the 5′ isoforms of the *Pcdhβ* cluster are also significantly increased (Fig. 5e). This suggests that the two *HS5-1* CBS elements function as an insulator to block the activity of the *HS7* enhancer, which contains no CBS, from aberrantly activating the *Pcdhβ* promoters. As a control, we

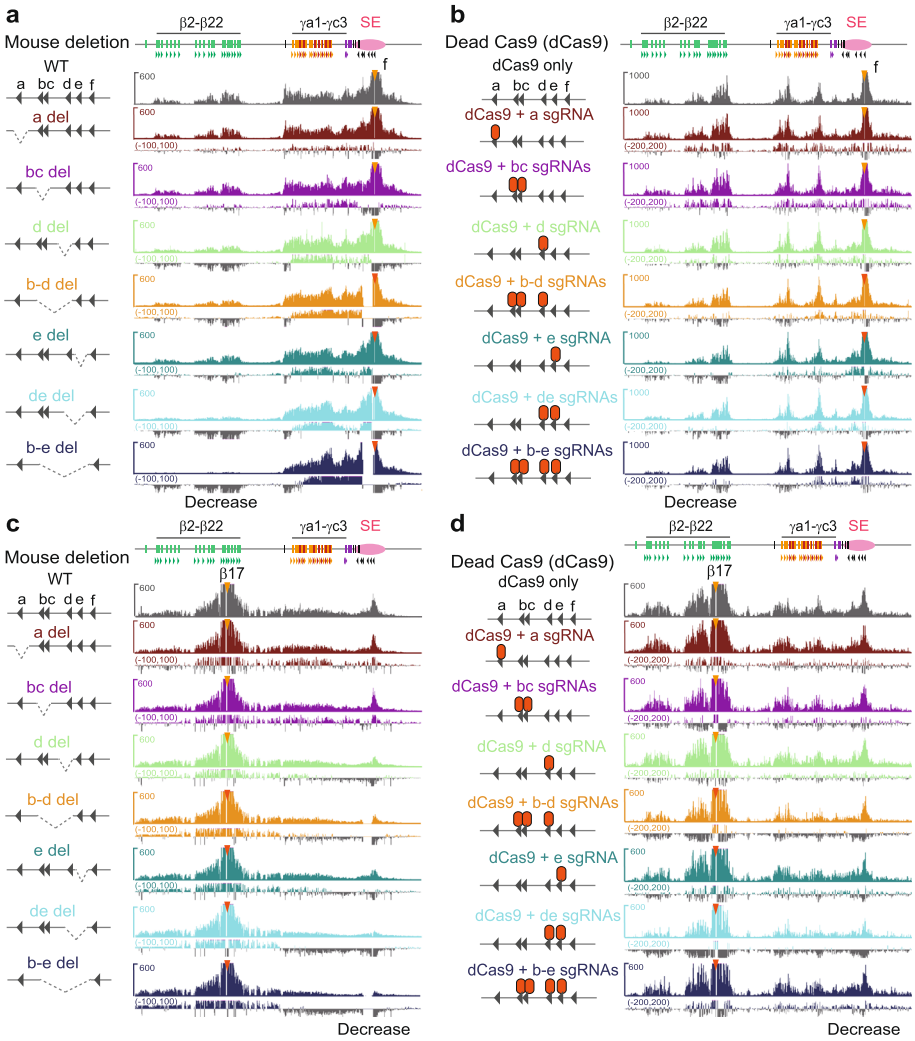

Fig. 3 The topology of spatial chromatin contacts of super-enhancer with the *Pcdh β* and *γ* promoters. **a** QHR-4C interaction profiles with the CBS *f* of the downstream super-enhancer as a viewpoint in cortical tissues of mice with a series of deletions of individual CTCF sites or their combinations. **b** QHR-4C interaction profiles with the CBS *f* of the downstream super-enhancer as a viewpoint using cells transfected with dCas9 only or dCas9 with sgRNAs targeting individual CTCF sites or their combinations to pinpoint the effects to CTCF sites. **c**, **d** QHR-4C interaction profiles with the upstream *Pcdhβ17* promoter CBS as a viewpoint in mice with CTCF site deletions or in cells with dCas9-blocked CTCF sites confirm the decreased interactions with the downstream super-enhancer. WT, wild-type; del, deletion

inverted in situ the same *HS5-1* fragment including the two reverse-oriented CTCF sites in mice in vivo (Fig. 5a and Additional file 1: Figure S12a,b). In contrast to the *HS5-1* deletion, neither *HS7* chromatin looping interactions with nor expression levels of the 5′ isoforms of the *Pcdhβ* cluster are significantly increased (Fig. 5b, d, f). These remarkable differences between deletion and inversion of *HS5-1* clearly show that the two endogenous *HS5-1* CBS elements function as an insulator to block the *HS7* enhancer from aberrantly activating the *Pcdhβ* gene expression, and its insulation activity is orientation-independent in vivo, consistent with the insertions of exogenous CBS elements of either orientation in cell lines in vitro (Fig. 1 and Additional file 1: Figures S3-S7).

To further investigate whether this is true for the *β-globin* cluster, we next inserted a pair of reverse-forward CBS elements (designated "RF2" to be distinguished from the first "RF" in Additional file 1: Figure S5) into the location between the five globin promoters and the *HS2* enhancer, which also contains no CBS (Fig. 5g). ChIP-seq confirmed the binding of CTCF/cohesin to the inserted CBS pair but not its mutant sites (Fig. 5g). QHR-4C experiments with either the *HS2* enhancer or the *HBG2* promoter as a viewpoint demonstrated a significant decrease of the *β-globin* enhancer-promoter interactions (Fig. 5h and Additional file 1: Figure S13a). Consistently, the expression levels of all *β-globin* genes are significantly decreased, and the decrease is rescued by CBS mutations (Fig. 5i).

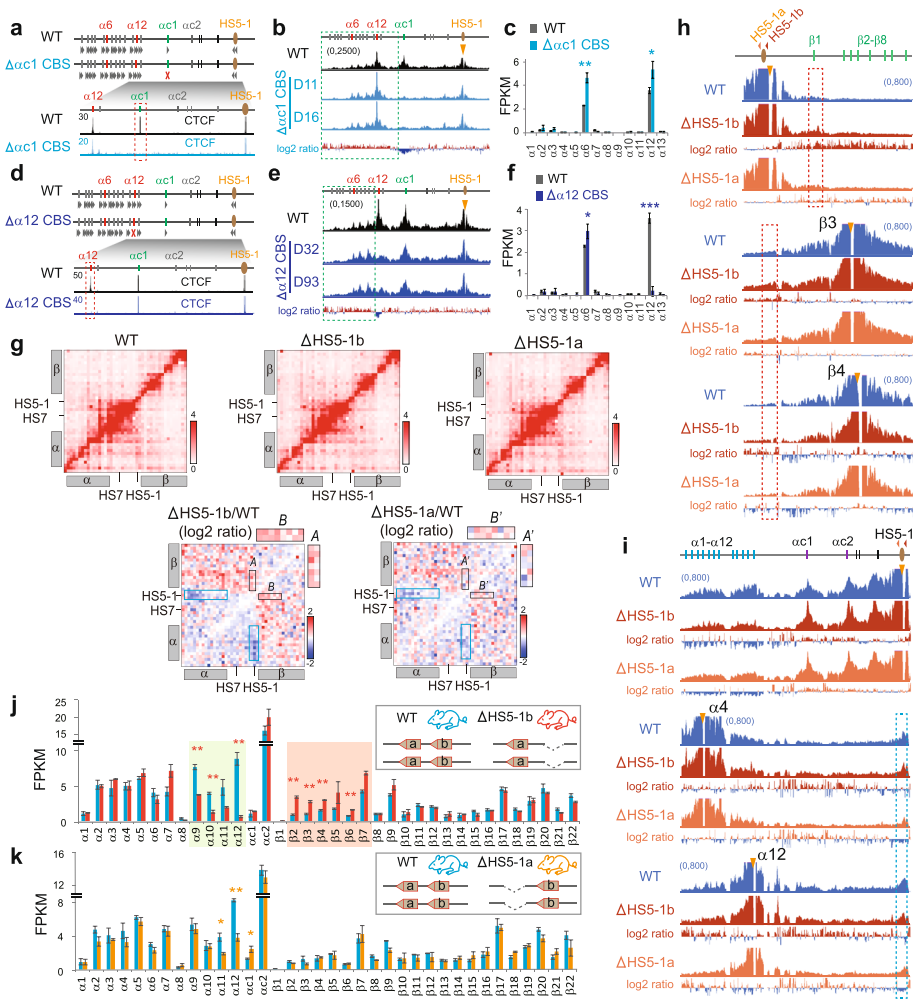

Fig. 4 Endogenous CTCF sites as *Pcdh* insulators. **a** CTCF ChIP-seq of the *Pcdhac1* CBS deletion CRISPR HEC-1-B cell clones. **b** QHR-4C profiles of the long-range chromatin contacts with *HS5-1* as a viewpoint in two single-cell CRISPR clones (D11, D16) with the deletion of the endogenous *ac1* CBS. Log2 ratios (deletion vs wild-type) are also shown. **c** RNA-seq of the WT and *ac1* CBS-deleted CRISPR clones. **d**–**f** Corresponding to **a**–**c**, respectively, but with *a12* CBS deletion in two single-cell CRISPR clones (D32, D93). **g** 5C interaction profiles of the *Pcdh a* and *β* clusters in cortical tissues of the *HS5-1b* or *HS5-1a* CBS deletion mice. The log2 ratios of chromatin interactions of *HS5-1* with *Pcdha* or 5′ isoforms of *β* gene repertoire are highlighted by blue or black rectangles, respectively. Note the significant increase of chromatin interactions between *HS5-1* with 5′ isoforms of the *Pcdhβ* cluster upon the *HS5-1b* deletion as indicated by enlargement of Insets *A* and *B*, compared with no alteration upon the *HS5-1a* deletion as indicated by enlargement of Insets *A*′ and *B*′. **h**, **i** QHR-4C confirmed the increased interactions with 5′ isoforms of the *Pcdhβ* cluster and the decreased interactions with the *Pcdha* cluster. **j**, **k** RNA-seq revealed increased expression levels of the 5′ isoforms of the *Pcdhβ* cluster in the homozygous CBS *HS5-1b* deletion (**j**) mice in comparison with the *HS5-1a* deletion (**k**) mice as controls. Data as mean ± SD, **p* < 0.05, ***p* < 0.01, ****p* < 0.001. One-tailed Student's *t* test

QHR-4C with *5′HS5* or *3′HS1* as a viewpoint, which contains a CBS element located outside of and beyond the *β-globin* enhancer and promoter regions, respectively, revealed opposite chromatin looping interactions with the inserted reverse-forward CBS pair (Additional file 1: Figure S13b-d). Thus, CBS elements, if inserted between enhancers and promoters with no CBS, also function as insulators by forming long-distance chromatin looping interactions with CBS elements located in the endogenous genome outside of and beyond respective enhancer and promoter regions. Finally, we inserted various combinations of CBS elements upstream and/or downstream of the *HS7* enhancer, which contains no CTCF site, of the *Pcdhα* cluster and found that the inserted CBS elements block long-distance chromatin interactions of the *HS7* enhancer (Additional file 1: Figure S13e-g). In conjunction with the data of the endogenous *Pcdh* CBS deletion, we conclude that CBS elements function as insulators for enhancers with no CBS.

Genome-wide CTCF sites function as insulators

To see whether genome-wide CTCF-bound CBS elements function as insulators for enhancers, we analyzed

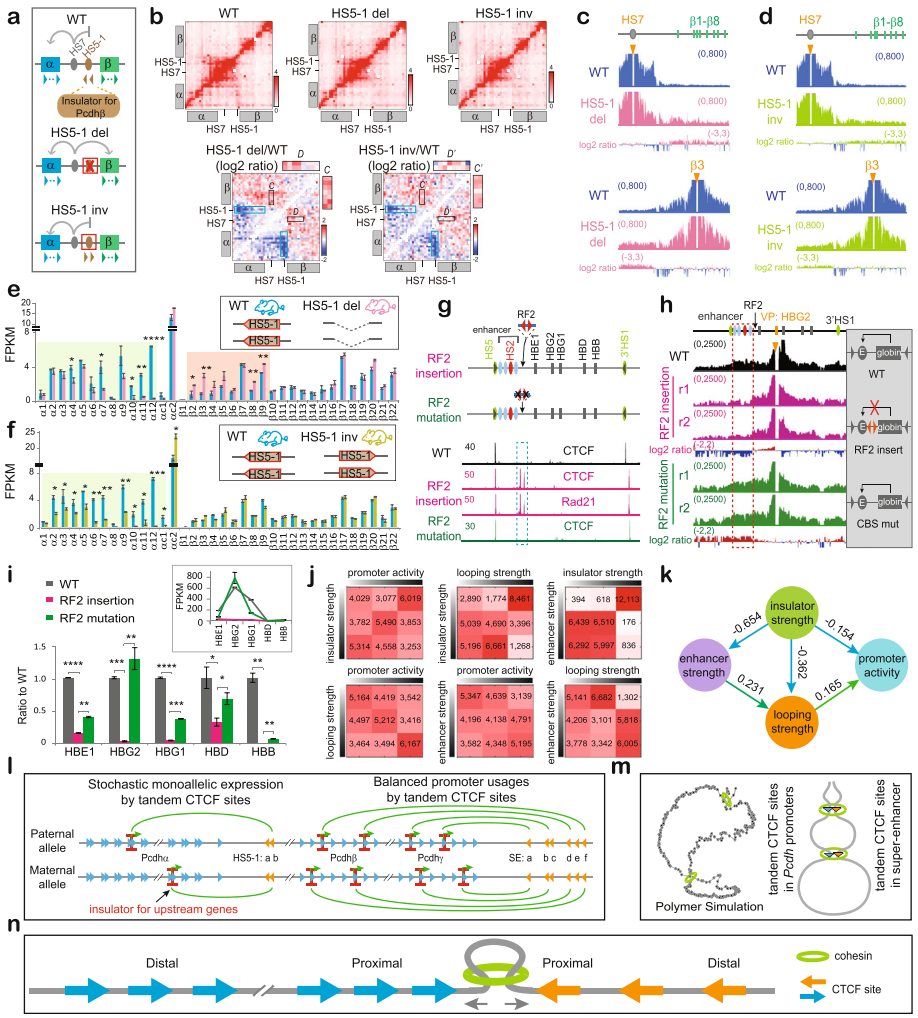

Fig. 5 Insulators for enhancers with no CTCF site. **a** Schematic of the *HS5-1* CBS elements as an insulator of the *HS7* enhancer for the *Pcdhβ* genes. **b** 5C interaction profiles of the *Pcdh α* and *β* clusters in the *HS5-1* deletion (del) or inversion (inv) mice in vivo. The log2 ratios of chromatin interactions of *HS5-1* with *Pcdhα* and of *HS7* with 5′ isoforms of the *Pcdhβ* cluster are highlighted by blue or black rectangles, respectively. Note the significant increase of chromatin interactions between *HS7* with 5′ isoforms of the *Pcdhβ* cluster upon *HS5-1* deletion as indicated by the enlargement of insets *C* and *D*. **c** QHR-4C with *HS7* or *β3* as a viewpoint confirms the increased interactions between *HS7* and 5′ isoforms of the *Pcdhβ* cluster in homozygous *HS5-1* deletion mice. **d** QHR-4C with *HS7* or *β3* as a viewpoint confirms no significant alteration of interactions between *HS7* and 5′ isoforms of the *Pcdhβ* cluster in homozygous *HS5-1* inversion mice. **e** RNA-seq of cortical tissues of the WT and *HS5-1* deletion mice. **f** RNA-seq of cortical tissues of the WT and *HS5-1* inversion mice. **g** CTCF and Rad21 ChIP-seq of human single-cell *β-globin* CRISPR clones with insertion of a pair of reverse-forward CBS elements ("RF2"). **h** QHR-4C profiles with the human *β-globin HBG2* promoter as a viewpoint. **i** RNA-seq reveal decreased expression levels (normalized to WT) of the human *β-globin* repertoire. The actual expression levels are shown in the inset. Data as mean ± SD, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, $****p < 0.0001$. One-tailed Student's *t* test. **j** The pairwise contingency tables showing interrelationships of each pair of variables among genome-wide insulator strength, promoter activity, and enhancer and looping strength. **k** The optimal structure of relationships among genome-wide insulator strength, promoter activity, enhancer and looping strength during human epidermal differentiation learned by Bayesian networks. **l** Directional CTCF looping underlies stochastic monoallelic *Pcdh α* and *βγ* gene expression and balanced promoter usage. In particular, stochastic and monoallelic CTCF-mediated directional chromatin looping underlies activation of one and only one variable promoter in each chromosome in the *Pcdha* cluster, while up to 4 promoters are activated in each chromosome in the *Pcdhβ/γ* clusters. **m** A polymer simulated 3D Hulu (gourd) model of tandem CTCF sites topologically balancing spatial chromatin contacts and enhancer-promoter selection. **n** Mechanistic interpretation of the 3D Hulu model in the context of bidirectional cohesin "loop extrusion" through tandem CTCF sites

developmental plasticity of insulators, promoters, and enhancers during human epidermal differentiation by Bayesian networks learned by the max-min hill-climbing algorithm [45] using categorized factors (Fig. 5j) inferred from previously published capture Hi-C data [46]. Remarkably, we find a direct inverse relationship between insulator strength and promoter activity (Fig. 5k). Moreover, insulators also regulate promoter activity indirectly through enhancers by perturbing the looping strength of their spatial chromatin contacts (Fig. 5k). Thus, these

Bayesian network analyses, in conjunction with our deletion and inversion experiments in mice in vivo (Figs. 2, 3, 4, and 5), suggest that CTCF-bound CBS elements function as insulators through directional chromatin looping across the human genome.

Discussion

Considerable progress has been made in understanding the stochastic expression of large repertoires of gene clusters by spatial regulation of chromatin contacts [12, 28, 33, 47]. In particular, the allelic insulation (Fig. 5l) by CTCF-mediated directional looping may be epigenetically regulated by methylation of CBS elements [35, 48]. Some CBS elements, such as the boundary *Pcdh HS5-1b* site, contain no CpG dinucleotides [35]. Consequently, the *HS5-1b* site has constitutive and cell-invariant CTCF/cohesin occupancy and functions as a chromatin insulator for the downstream *Pcdhβ* genes (Fig. 5l). Other CTCF sites are regulated by DNA methylation and have a cell-specific pattern of the CTCF/cohesin occupancy in single neurons [28]. For example, each CBS within the alternate variable promoter of members of the *Pcdh α*, *β*, and *γ* clusters contains a CpG dinucleotide that is methylated in specific subpopulations of neurons in the brain [28, 35, 49]. Therefore, the 5′ boundary CBS element of each *Pcdh* loop domain is cell-specific and distinct for single neurons, thus functions as a chromatin insulator for its respective upstream genes (Fig. 5l). Consistently, our computational modeling suggests that members of the three *Pcdh* families are expressed monoallelically in individual neurons (Additional file 1: Figures S1 and S9) [40].

This explains the long-standing puzzle of stochastic *Pcdhα* monoallelic expression in single cells (Additional file 1: Figure S1) [39]. Specifically, for any unmethylated promoter CBS element, it forms long-distance chromatin contacts with downstream enhancers and therefore is activated through chromatin looping. These chromatin looping interactions function as an insulator for all of its upstream *Pcdhα* promoters, resulting in inactivation or silencing of them in each chromosomal allele (Fig. 5l). In addition, all of its downstream *Pcdhα* promoters are not activated because by mathematical definition, if any downstream promoter is activated by enhancer looping, none of its upstream promoters could be activated (Fig. 5l). Similarly, for the more complex regulation of the *Pcdh β* and *γ* clusters, our genetic experiments demonstrate that they are topologically regulated by the tandem CTCF sites within the downstream super-enhancer (Figs. 2 and 3), as explained in the Hulu model of topological gene regulation (Fig. 5m). Therefore, only one isoform of the *Pcdhα* cluster (Additional file 1: Figure S1a,b) and up to 4 isoforms of the *Pcdh β* and *γ* clusters (Additional file 1: Figure S9a) are expressed from each chromosomal allele in individual neurons in the brain (Fig. 5l).

We posit a general mechanism for the Hulu model by which tandem directional CTCF sites function as topological insulators in the context of the cohesin "loop extrusion" (Fig. 5m, n). Because each CBS element has permeability for cohesin sliding [20], continuous active chromatin loop extrusion by cohesin in an ATP-dependent manner bridges inner convergent CTCF sites first (Fig. 5n). After sliding through the inner proximal CTCF sites, cohesin will then stall at the intermediate tandem CTCF sites (Fig. 5n). Finally, cohesin will reach the outer distal CTCF sites (Fig. 5n). Based on our experimental observation and mathematical prediction in the clustered *Pcdh* and *Ig* genes, the two "heads" of cohesin are stalled or anchored at CTCF sites in the two arrays of convergent tandem CTCF sites, resulting in long-range chromatin interactions between proximal-proximal CBS elements as well as between distal-distal CBS elements (Fig. 5m, n). In other words, two "heads" of cohesin complex anchor proximal-proximal or distal-distal CBSs through continuous active loop extrusion of chromatin fibers which are asymmetrically blocked by permeable CTCF insulators. The functional consequences of these interactions caused by tandem CTCF insulators are the decreased proximal chromatin interactions and increased distal chromatin interactions. Thus, tandem directional CTCF sites function as topological insulators to balance higher-order chromatin contacts and promoter choice, eliminating bias of spatial chromatin accessibilities between proximal and distal promoters by remote enhancers.

Overwhelming evidence suggests that the function of insulators is orientation-independent, but the chromatin looping of CBS elements is directional [7, 11, 12, 14, 24, 33, 50]. CTCF mediates specific directional loop formation through asymmetric anchoring of the ring-shaped cohesin complex, which slides along chromatin fibers to actively extrude loops [3, 12, 18, 19, 33, 51]. Our data are consistent with the predominant chromatin interactions between forward-reverse CBS pairs [11, 12]. In addition, there are numerous cases of tandem CTCF sites across mammalian genomes [12, 16, 33, 52]. Since the binding of CTCF to genome-wide CBS elements is not static but rather dynamic [9, 10] and there is variable permeability of CTCF extrusion barriers [20], this suggests that cohesin slides through the proximal CTCF sites within tandem CBS arrays to more distal sites (Fig. 5n). Curiously, our computational simulation in silico and genetic deletion in vivo revealed that tandem-arrayed CBS elements ensure balanced usage of associated promoters in specific and equal spatial chromatin contacts in general. Thus, our data on *Pcdh*, *β-globin*, and *Igh* clusters suggest that directional CTCF chromatin looping between convergent CBS

elements underlies insulator function and that tandem CTCF sites ensure balanced promoter spatial accessibility in the 3D genome folding and regulation. However, since there are numerous gene clusters and hundreds of thousands CTCF sites in mammalian genomes, whether all tandem CTCF sites function in a similar manner in vivo waits further studies.

Conclusion

In the present study, we show by CRISPR DNA-fragment editing, in conjunction with mathematic modeling and chromosome conformation capturing, that tandem directional CTCF sites function as topological insulators to enhance long-distance chromatin interactions with distal CTCF sites and to balance promoter-enhancer selections. Specifically, ectopic and endogenous CTCF sites function as insulators in an orientation-independent manner through CTCF-mediated directional chromatin looping. In addition, in combination with computational simulations of cohesin "two-headed" chromatin loop extrusion, we demonstrate that tandem CTCF sites ensure proper spatial accessibility of distal promoters by remote enhancers and balanced usage of target promoters. Finally, we report that tandem CTCF sites regulate long-distance chromatin looping in the mammalian genome in a topological manner.

Methods

Cell culture

Human endometrial HEC-1-B cells (ATCC) were cultured in MEM medium (Hyclone), supplemented with 10% (v/v) FBS (Gibco), 2 mM glutamine (Gibco), 1 mM sodium pyruvate (Sigma), and 1× penicillin-streptomycin (Gibco). Human K562 and mouse Neuro-2A cells (ATCC) were cultured in DMEM medium (Hyclone) supplemented with 10% (v/v) FBS and 1× penicillin-streptomycin. Cells were maintained at 37 °C in a humidified incubator containing 5% (v/v) of CO_2 and were passaged every 3 days.

In vitro transcription of sgRNA pairs and Cas9 mRNA for microinjection

The preparation of sgRNA pairs and Cas9 mRNA was recently described [38]. Briefly, to obtain sgRNAs for microinjection of zygotes, we performed in vitro transcription using DNA templates generated by PCR with a forward primer containing a T7 promoter followed by targeting sequences and a common reverse primer. In vitro transcription was performed with the MEGAshortscript Kit (Life Technologies) using T7 polymerase by incubating at 37 °C for 5 h. The template DNA was removed by digestion with DNaseI. The transcribed sgRNAs were purified with the MEGAclear Kit (Life Technologies) and eluted in TE buffer (0.2 mM EDTA). The sequences of primers used for preparing sgRNAs were listed in Additional file 2: Table S1.

To obtain Cas9 mRNA for the microinjection of zygotes, the Cas9 coding sequence was cloned into pcDNA3.1 plasmid under the control of the T7 promoter. The plasmid was then linearized by XbaI and used for in vitro transcription with the mRNA transcription system according to the manufacturer's instructions (Life Technologies). After digestion of the DNA template, the transcribed Cas9 mRNA was purified with the MEGAclear Kit (Life Technologies).

Generation of the CBS deletion and inversion mice by CRISPR DNA-fragment editing

Mice were maintained at 23 °C in a 12-h (7:00–19:00) light and 12-h (19:00–7:00) dark schedule in an SPF mouse facility. For each CRISPR deletion or inversion of CBS elements, Cas9 mRNA (100 ng/μl) and a pair of sgRNAs (50 ng/μl each) targeting the region flanking the CBS elements were injected into the cytoplasm of one-cell embryos of the C57BL/6 mice. After recovering for 2 h at 37 °C incubator, the embryos were then implanted into the oviducts of the pseudo-pregnant ICR mice. The newborn F0 mice (Additional file 2: Table S2) were then screened for targeted deletions or inversions by PCR using specific primer pairs (Additional file 2: Table S1). The amplified PCR products were then cloned and confirmed by Sanger sequencing. The F0 mice with targeted deletions or inversions were maintained and crossed to obtain F1 mice. F1 mice were genotyped again for heterozygous deletion or inversion. Heterozygous F1 mice were then crossed to obtain homozygous F2 mice. For all of the 5C and RNA-seq experiments, only the wild-type littermates were used as controls.

Single-cell RNA-seq

Single-cell RNA-seq experiments were performed as previously described [40]. Briefly, for neurons, the P0 mouse brain was dissected, and the tissue from the cerebral cortex was digested with 0.013% of collagenase in Neurobasal Medium (Gibco) at 37 °C for 3 min. The collagenase was neutralized by adding an excess amount of Neurobasal Medium. A single-cell suspension was made by gentle pipetting and then filtered through 100-μm cell strainers (BD Biosciences). For HEC-1-B cells, trypsin was added to the culture dish, and the single cells were suspended in the culture medium. Single cells were then picked under the microscope by using a microcapillary pipette into the thin-walled PCR tube containing 2 μl of cell lysis buffer, 1 μl of oligo-dT primer, and 1 μl of dNTP mix. After reverse transcription, the cDNA was pre-amplified by PCR. The cDNA library was then purified, tagmented, and ligated with adapters using the Nextera XT DNA Library Preparation kit (Illumina FC-

131-1096). Finally, the adapter-ligated fragments were further amplified by PCR and purified with AMPure XP beads (Beckman). The single-cell RNA-seq libraries were pooled and sequenced using an Illumina Hiseq 2500 platform.

Plasmid construction

The plasmids of sgRNAs for cell transfection experiments were constructed as previously described [37, 38]. Briefly, pairs of complementary oligonucleotides for generating sgRNAs (Additional file 2: Table S1) were annealed with 5′ overhangs of "ACCG" and "AAAC," and cloned into a BsaI-linearized pGL3 vector under the control of the U6 promoter. To insert CBS elements into distinct genomic regions, circular donor plasmids with about 2-kb homologous arms flanking the inserted sequence were used as donors for CRISPR-based homologous recombination. To construct donor plasmids, we amplified the CBS elements, as well as the genomic sequences flanking the insertion site by PCR. CBS elements and the two homologous arms with 20 bp of overlapping sequences were jointed together with the EcoRI and HindIII digested Puc19 vector using the multi-fragment recombination system (Vazyme). All of the plasmids constructed were confirmed by Sanger sequencing. The primer sequences used for the construction of sgRNAs and the donor plasmids were shown in Additional file 2: Table S1.

Screening CBS insertion and deletion single-cell clones by CRISPR DNA-fragment editing

Generation of the CRISPR single-cell clones with CBS element insertions and deletions was performed as previously described [12, 38]. Briefly, cells were transfected with a plasmid mix using Lipofectamine 3000 reagents (Thermo) in a 12-well plate. For CBS insertions and mutations, Cas9 (0.3 μg) and donor plasmids (0.5 μg) were co-transfected with one sgRNA construct (0.2 μg) targeting the insertion site. For CBS deletions, Cas9 plasmids (0.4 μg) were co-transfected with two sgRNA constructs (0.3 μg each) targeting the two ends of the deletion fragments. The sgRNA constructs contained a puromycin-resistant gene which can be used for selection. Forty-eight hours after transfection, puromycin (Sigma) was added to the culture medium at a final concentration of 2 μg/ml. The culture medium was replaced every day with puromycin for a total of 3 days. Puromycin was then removed, and cells were cultured in normal culture medium for 2 days. The cells were then suspended into a single-cell solution and plated into 96-well plates at the concentration of about one cell per well. Two weeks after plating, single-cell clones were marked manually under a microscope and replaced with fresh culture medium. Four weeks after plating, the single-cell clones were screened for insertion, mutation, or deletion by PCR. At least two individual clones for each insertion, mutation, or deletion were obtained and analyzed. We screened for a total of 1948 single-cell clones, and 80 homozygous clones were obtained and analyzed (Additional file 2: Table S3). Single-cell clones for each editing were confirmed by Sanger sequencing. The primers used for genotyping were listed in Additional file 2: Table S1.

Targeted blocking of CTCF sites by dCas9

We used a well-established method to block CBS functions by dCas9 [53–55]. We first mutated sequences encoding the RuvC and HNH domains of Cas9 to generate a pcDNA3.1 plasmid encoding a catalytically dead Cas9 (dCas9) which lacks the endonuclease activity. The plasmid backbone contains a puromycin-resistance gene which is suitable for puromycin selection. In addition, we chose the sgRNA sequence to target module 2 and module 3 of CTCF sites according to the molecular structures of CTCF-DNA complexes [56, 57]. The sgRNA expression plasmids were constructed by annealing two overlapping primers and inserting the annealed dsDNA into the plasmid backbone as previously described [37]. The primers used for generating sgRNA plasmids were listed in Additional file 2: Table S1.

We make use of the mouse neuroblastoma cell line Neuro-2A as an established model system to investigate the role of CBS in the regulation of the clustered *Pcdh* genes [12]. The Neuro-2A cells cultured to 70% confluency in 6-well plates were transiently transfected with 1.25 μg dCas9 and 1.25 μg sgRNA plasmids using Lipofectamine 3000 transfection reagent (Invitrogen) with the protocol recommended by the manufacturers. We transfected dCas9 with the plasmid targeting Gal4 for the control group. Forty-eight hours after transfection, cells were selected with 2 μg/ml of puromycin diluted in culture medium for 4 days. The survival cells were cultured for another 3 days in normal culture medium without puromycin and harvested for QHR-4C experiments.

ChIP-seq experiments

ChIP experiments were performed as previously described [35] with modifications. Briefly, 4×10^6 of cells were cross-linked by 1% formaldehyde in 10% FBS/PBS for 10 min at room temperature. Cells were then lysed twice with ice-cold lysis buffer (20 mM Tris-HCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate, and 1× protease inhibitors, pH 7.5) for 10 min with slow rotations. The lysed cells were then sonicated to obtain DNA fragments of about 200–500 bp using the Bioruptor system (high energy, with working time of 30 s and resting time of 30 s, 30 cycles). After removal of the insoluble debris, the lysate was incubated with

specific antibodies against CTCF (07-729; Millipore), RAD21 (ab992; Abcam), or NIPBL (A301-779A; Bethyl Laboratories) and purified by protein A-agarose beads (16-157; Millipore). NIPBL and CTCF ChIP-seq for the *Igh* locus were recently published [58]. ChIP DNA was extracted and prepared for high-throughput sequencing using a DNA library preparation kit for Illumina (NEB). ChIP-seq libraries were sequenced on a HiSeq X Ten platform (Illumina).

Quantitative high-resolution chromosome conformation capture copy (QHR-4C)

We developed a QHR-4C method to detect genomic elements that are close to any viewpoint of interest with high efficiency and specificity. This method is conceptually similar to UMI-4C and HTGTS [59]. We used this method to study chromatin conformation of the clustered *Pcdh* and *β-globin* loci from as few as 50,000 cells. After the cells were harvested and crosslinked, chromatins within the nuclei were digested in situ by a restriction enzyme. The chromosome conformation is then captured by proximal ligation. After fragmentation by sonication, a linearized amplification step is applied to enrich ligation events associated with a specific viewpoint using a single primer tagged with biotin. The amplified single-stranded biotin-tagged DNA fragments were purified with streptavidin beads and ligated with a staggered adapter. Finally, QHR-4C libraries were generated by PCR.

Compare to the regular 4C, QHR-4C has several advantages. First, the chosen viewpoint is much more flexible in QHR-4C. In the regular 4C, the size of the viewpoint fragments should be at least 200 bp to allow for efficient self-circulating in the second ligation step. In addition, there must be at least one restriction enzyme cutting site within the viewpoint fragment to allow for self-circulation. By contrast, the only requirement for viewpoint selections in QHR-4C is the matching of a linearized amplification primer. Second, the regular 4C could not detect chromatin interactions of the fragments that do not contain the second restriction enzyme cutting site. However, QHR-4C, which does not require the second digestion step, is able to detect these chromatin interactions and allows for better coverage of genomic regions of interests. Third, since the ends of the captured DNA fragments are generated by sonication, the captured dsDNA ends are random and unique, and thus can be used as an identifier for quantifying the long-range chromatin interactions. Finally, multiplexing QHR-4C is much easier than the regular 4C experiments.

Briefly, single cells from various CRISPR single-cell clones and mouse cortical tissues were centrifuged at 500*g* for 5 min, and the pellets were used for QHR-4C experiments. The cell pellets were suspended for crosslinking in 900 μl 2% formaldehyde at room temperature for 10 min. The crosslinking reaction was stopped by adding and mixing with 100 μl of 2 M glycine for a final concentration of 200 mM. The fixed cells were spun down at 800*g* at 4 °C for 5 min and washed twice by suspending briefly in 1 ml ice-cold PBS. Cells were then permeabilized twice with 200 μl ice-cold 4C permeabilization buffer each for 10 min (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 5 mM EDTA, 0.5% NP-40, 1% Triton X-100, and 1× protease inhibitors). After centrifugation, the pellet was resuspended in 73 μl water, 10 μl of 10× DpnII buffer (we used DpnII enzyme as an example, using the recommended buffer for other enzymes), and 2.5 μl of 10% SDS. The reaction was performed at 37 °C for 1 h with constant shaking at 900 rpm. 12.5 μl of 20% Triton X-100 was added into the reaction to quench SDS and incubated at 37 °C for 1 h with shaking at 900 rpm. The cells were then digested in situ overnight at 37 °C with 2 μl of DpnII (10 U/μl) while shaking at 900 rpm. After the inactivation of DpnII at 65 °C for 20 min, the pellets of the nuclei were collected by centrifuging at 1000*g* for 1 min, and the supernatant was removed completely, which ensures the subsequent ligation reaction can be performed in a small volume. Proximity ligation was carried out for 24 h at 16 °C with 1 μl T4 DNA ligase (400 unit/μl) in 100 μl 1× T4 ligation buffer. The ligated product was then reverse crosslinked by heating to 65 °C for 4 h in the presence of 1 μl proteinase K (10 mg/ml) to digest proteins. The DNA was then extracted using phenol-chloroform. One microliter glycogen (20 mg/ml) was added to facilitate DNA precipitation. The precipitated DNA was dissolved in 50 μl water. We sonicated the ligated DNA using the Bioruptor system (with low energy setting at a train of 30-s sonication with 30-s interval for 12 cycles) to obtain DNA fragments ranging from 200 to 600 bp.

After fragmentation, a linearized amplification step is applied to enrich the ligation events associated with a specific viewpoint, using a 5′ biotin-tagged primer (Additional file 2: Table S1) complementary to the viewpoint fragment in 100 μl of PCR system for a total of 60 cycles. This primer should be neither too close to the DpnII site to facilitate the nested PCR at the final amplification step nor too far away from the DpnII site to maximize the product amount. The amplification products were denatured by incubating at 95 °C for 5 min and immediately chilled on ice to obtain ssDNA. The ssDNA was then enriched and purified with Streptavidin Magnetic Beads (Invitrogen) according to the manufacturer's instructions.

The ssDNA on beads was then ligated in 15 μl ligation buffer with 0.1 μM of adapters (Additional file 2: Table S1) at 16 °C for 24 h. We chose the adapter sequence that matched the 3′ end of the Illumina P7 sequence so that

one PCR step can produce sequencing libraries. The adapters were generated by annealing two complementary primers in annealing buffer (25 mM NaCl, 10 mM Tris-HCl pH 7.5, 0.5 mM EDTA). After ligation, free adapters were removed by washing the beads twice with the B/W buffer (5 mM Tris-HCl, 1 M NaCl, 0.5 mM EDTA, pH 7.5). The DNA on beads was resuspended in 10 μl water. Finally, the QHR-4C libraries were generated by one-step PCR amplification (94 °C, 2 min; 94 °C, 10 s; 60 °C, 15 s; 72 °C, 1 min for 19 cycles; and a final extension at 72 °C, 5 min) with captured DNA on beads as the template and a pair of PCR primers. The forward primer matches the Illumina P5 and the viewpoint sequence adjacent to the DpnII site with barcodes, and the reverse primer matches Illumina P7 with indexes (primer sequences are listed at Additional file 2: Table S1). The PCR products were purified with a PCR purification kit (Qiagen). About 100 QHR-4C libraries with different combinations of barcodes and indexes were pooled and sequenced on an Illumina HiSeq X Ten platform. All of the QHR-4C experiments for each CRISPR clone and CRISPR mouse lines were performed with two biological replicates.

Circularized chromosome conformation capture

The circularized chromosome conformation capture (4C) experiments were performed as previously described [12, 14]. Briefly, cells were counted, and about 2×10^6 cells were used for each 4C experiment. After cross-linking with 2% formaldehyde, cells were lysed twice with cold lysis buffer, digested with DpnII, and ligated with T4 DNA ligase. The ligated samples were purified using the High-Pure PCR Product Purification kit (Roche). The 4C-seq libraries were generated by PCR using a high-fidelity DNA polymerase (Vazyme). All of the 4C experiments were performed with biological replicates. 4C-seq libraries were sequenced on the HiSeq X Ten platform. 4C primers used were listed in Additional file 2: Table S1.

Chromosome conformation capture carbon copy

Chromosome conformation capture carbon copy (5C) experiments were performed as previously described [60, 61]. Briefly, a total of 46 forward and 46 reverse primers covering the mouse *Pcdh* α and β clusters were designed by My5C tools (http://my5c.umassmed.edu) [62]. These primers are a subset of the 5C primer set covering all three *Pcdh* gene clusters [63]. All forward primers contain a 5′ end T7 universal primer sequence (CGGTA ATACG ACTCA CTATA GCC) preceding a unique sequence which is followed by AAG at the 3′ end. All reverse primers contain CTT at 5′ end followed by a unique sequence and a complementary T3 universal sequence (TCCCT TTAGT GAGGG TTAAT A). All reverse primers were 5′-phosphorylated.

Generation of 5C libraries for sequencing

The P0 mouse cortical tissues were dissociated to obtain single-cell suspension as described above in the single-cell RNA-seq experiments. A total of 10^7 cells were cross-linked and digested with HindIII (NEB). After inactivating HindIII, the digested DNA was ligated with T4 DNA ligase and purified. As a control, DNA of six bacterial artificial chromosomes (BACs) covering the three *Pcdh* clusters was also digested, ligated, and purified. The purified mouse cortical DNA was mixed with 1 μg of salmon sperm DNA (Sigma). The control BAC DNA (5 ng) was mixed with 1.5 μg of salmon sperm DNA. These samples were then each mixed with 1.7 fmol of each 5C primer and 1 μl of 10 × 5C annealing buffer (20 mM Tris-acetate pH 7.9, 50 mM potassium acetate, 10 mM magnesium acetate, 1 mM DTT) in a total volume of 10 μl and denatured at 95 °C for 5 min. Annealing was performed by incubation at 48 °C for 16 h. The annealed DNA was ligated by adding Taq DNA ligase (NEB) in the 5C ligation buffer (25 mM Tris-HCl pH 7.6, 31.25 mM potassium acetate, 12.5 mM magnesium acetate, 1.25 mM NAD, 12.5 mM DTT and 0.125% Triton X-100). The ligation reaction was performed for 1 h at 48 °C followed by incubation for 10 min at 65 °C to stop the ligation reaction. The ligated products were amplified by PCR with Illumina primer pairs. The amplified libraries were purified with a PCR purification kit (QIAGEN) for high-throughput sequencing.

5C reads mapping

The 5C libraries were sequenced with the 90-bp pair-end mode by the Hi-seq 2500 platform of Illumina. All 5C experiments were performed with two biological replicates. The read depth of each sample was equal to about 2 million (Additional file 2: Table S4). Pearson correlation coefficients between the two biological replicates range from 0.967081 to 0.99251 (Additional file 2: Table S5). We used 56-bp reads for mapping. Each of the paired-end reads was independently mapped using the local mapping mode of Bowtie2 with default parameters. Only both of the paired-end reads uniquely mapped to a single 5C interaction were used for downstream analyses. We found that about 96% of paired-end reads can be uniquely mapped (Additional file 2: Table S4). The read count was then normalized to 1 million for each sample to correct the difference in sequencing depth.

5C bias correction

Bias may be introduced in many steps in 5C experiments including, but not limited to, differences in the

crosslinking efficiency, differences in restriction enzyme digestion efficiency, differences in ligation efficiency, differences in 5C primer and PCR amplification efficiency, and differences in DNA sequencing efficiency. All of these potential biases are shared by all experimental groups as we used the same sets of primers and investigated the same genomic region. As a result, the bias can be partially neutralized as we focused on the differences between each sample. In addition, we performed BAC control experiments to reduce 5C primer and PCR amplification bias. Finally, we filtered primers by a statistical method known as Loess.

Locally estimated scatterplot smoothing

Locally estimated scatterplot smoothing (Loess) locally fits the response y_i (5C interaction frequency) to the predictor x_i (genomic distance) for $i \in [1, n]$ by a function from a specific parametric class, say polynomials of degree 1 or 2, which provide an estimate $\hat{g}(x)$. A function $w_{\hat{x}}(x)$ with local support is used to weight the predictors around $\hat{x}$.

$$w_{\hat{x}}(x) = \begin{cases} \left(1-\left(\frac{|x-\hat{x}|}{d}\right)^3\right)^3, |x-\hat{x}| \le d, \\ 0, |x-\hat{x}| > d, \end{cases}$$

where d is the distance from $\hat{x}$ to the $\lceil \alpha n \rceil$th closest predictor in $\{x_1, x_2, \cdots, x_n\}$, and α is the percentage of data points used to calculate the response for $\hat{x}$. Under the assumption that the errors $\epsilon_i \coloneqq y_i - \hat{g}(x_i)$ are independent Gaussian random variables with 0 means and constant variances σ^2, Loess does weighted least squares, i.e., $\hat{g}(\boldsymbol{x}_{\hat{x}}) = X(X^T W X)^{-1} X^T W \boldsymbol{y}_{\hat{x}}$, where $\boldsymbol{x}_{\hat{x}} = (x_{i_1}, x_{i_2}, \cdots, x_{i_m})^T$ such that $\{x_{i_1}, x_{i_2}, \cdots, x_{i_m}\} = \{x_i | w_{\hat{x}}(x_i) > 0, 1 \le i \le n\}$, $\boldsymbol{y}_{\hat{x}} = (y_{i_1}, y_{i_2}, \cdots, y_{i_m})$, $W = \mathrm{diag}(w_{\hat{x}}(x_{i_1}), w_{\hat{x}}(x_{i_2}), \cdots, w_{\hat{x}}(x_{i_m}))$, and X depends on the parametric class used for the local regression. In the case of a polynomial of degree 2:

$$X = \begin{pmatrix} 1 & x_{i_1} & x_{i_1}^2 \\ 1 & x_{i_2} & x_{i_2}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{i_m} & x_{i_m}^2 \end{pmatrix}.$$

Denote $L \coloneqq X(X^T W X)^{-1} X^T W$. Then, the covariance matrix of the errors $\boldsymbol{\epsilon}_{\hat{x}} \coloneqq \boldsymbol{y}_{\hat{x}} - \hat{g}(\boldsymbol{x}_{\hat{x}})$ is $\sigma^2 (I-L)(I-L)^T \approx \frac{\boldsymbol{\epsilon}_{\hat{x}}^T \boldsymbol{\epsilon}_{\hat{x}} (I-L)(I-L)^T}{\mathrm{tr}(I-L)(I-L)^T}$, which gives the standard deviation SD_i for each data point.

Primer filtering

We performed data correction using locally estimated scatterplot smoothing (Loess) to calculate Z scores (a measurement of the number of standard deviations a data point is from the average value) of each 5C chromatin interaction. First, we calculated the global average relationship $\hat{g}$ between the interaction frequency and genomic distance via Loess smoothing for each sample. We used Loess [64] implemented in R to calculate the Z score $Z_i \coloneqq (y_i - \hat{g}(x_i))/\mathrm{SD}_i$ with default setting and the span of $\alpha = 0.01$. In this equation, y_i and x_i are the interaction frequency and genome distance of pair i, respectively. In addition, SD_i is the standard deviation of $y_i - \hat{g}(x_i)$. The overall interaction profile of each primer is then compared to the global average. If the individual Loess of a primer is higher or lower than 0.85 of the global average, it is flagged as problematic. If a primer is flagged in more than 40% of the datasets from all samples, it is removed from the downstream analyses from all datasets [65–67]. Using this threshold, we removed 7 primers (mpcdh-for-2, mpcdh-for-8, mpcdh-for-21, mpcdh-rev-6, mpcdh-rev-12, mpcdh-rev-17, mpcdh-rev-25) from the downstream analyses.

Singleton removal

In 5C experimental data, there are instances that 5C interactions resulting from aberrant PCR amplifications were much higher than neighboring interactions by more than an order of magnitude. These abnormal interactions may be caused by PCR over-amplification, the so-called PCR "blowouts" or abnormal singletons. To remove these singletons, we calculated the Z score for each 5C interaction. If the Z score of a 5C singleton is larger than 12, the singleton is removed [65–67]. In total, three singletons (mpcdh_for_14 - mpcdh_rev_30, mpcdh_for_25 - mpcdh_rev_39, and mpcdh_for_28 - mpcdh_rev_22) have been removed.

After data correction, we normalized 5C interactions by dividing the BAC sample. The mean ratio of two biological replicates is shown as heatmaps. To compare the interaction profiles, the log2 ratio between mutant and wild-type groups is calculated and shown as heatmaps.

RNA-seq experiments

RNA-seq experiments were performed as previously described [12] with modifications. Briefly, total RNA from mouse cortical tissues or cultured cells was extracted using TRIzol reagents (Life Technologies) following the manufacturer's instructions. Total mRNA was prepared from 1 μg total RNA using poly(A) mRNA magnetic isolation reagents (NEB) and fragmented at 94 °C for 15 min. RNA was then reverse-transcribed into cDNA with random primers. After end repairing and A-tailing, cDNA was ligated with adapters and amplified by PCR with Illumina sequencing primers. All RNA-seq experiments were performed with biological replicates. RNA-seq libraries were sequenced on a HiSeq X Ten platform.

High-throughput sequencing and data analyses

High-throughput analyzing pipelines were the same as previously described [12, 35] with some modifications. Briefly, reads that passed the Illumina quality filter were considered for alignments. For 4C-seq data, reads were aligned to the reference human (GRCh37/hg19) or mouse (NCBI37/mm9) genome using the Bowtie2 program. The reads per million (RPM) value was calculated using the r3Cseq program (version 1.20) in the R package (version 3.3.3). For QHR-4C data, duplicated paired-end reads were removed by FastUniq (version 1.1) program, and only the unique reads were used for analyses using the Bowtie and r3Cseq program. For ChIP-seq data analyses, reads were mapped to the reference genome (human GRCh37/hg19 or mouse NCBI37/mm9) or the modified genome with insertions using the Bowtie2 program. Peaks were called by the MACS program [68] (version 1.4.2) with a cutoff p value of 10^{-5}. For RNA-seq and single-cell RNA-seq data, reads were aligned using Hisat2 (version 2.0.4) to the human genome (GRCh38/hg38) or mouse genome (GRCm38/mm10), and the FPKM value was calculated using the Cufflinks program (version2.1.1).

Maximum likelihood modeling of *Pcdh* stochastic expression

Since single-cell RNA-seq data of each neuron are resulted from the combined expression of two sets of paternal and maternal chromosomes, upon the assumption that two chromosomal sets express independently, we first decomposed the RNA-seq data of single cells from the anterior lateral motor and primary visual cortices [40] into the expression of each chromosomal sets.

Let G be the total number of considered genes (for example, $G = 12$ in the mouse *Pcdhα* cluster). Define the whole gene set as $\mathcal{G} = \{g | 1 \le g \le G\}$. Because there are 81.56% and 86.64% single cells from anterior lateral motor and primary visual cortices express no more than 2 *Pcdhα* isoforms (Additional file 1: Figure S1), respectively, we assume that the *Pcdhα* cluster on a single chromosomal allele expresses at most H genes ($H = 2$ here in the *Pcdhα* cluster). Define $\mathbb{G}_H \coloneqq \{\mathcal{S} | \mathcal{S} \subset \mathcal{G}, |\mathcal{S}| \le H\}$, where $|\mathcal{S}|$ is the total number of elements in the set $\mathcal{S}$, as the set of all the subsets of $\mathcal{G}$ that contain less than H elements. In other words, $\mathbb{G}_H$ gives all possible gene sets that can be expressed from a single chromosomal allele.

Define the Cartesian product $\mathbb{G}_H^2 \coloneqq \{(\mathcal{S}, \mathcal{T}) | \mathcal{S}, \mathcal{T} \in \mathbb{G}_H\}$ as all possible combinatorial expression sets from both chromosomal alleles. For $\mathcal{R} \subset \mathcal{G}$, let $N_\mathcal{R}$ be the number of single cells that express the gene set $\mathcal{R}$. Define $\mathbb{G}_{H,\mathcal{R}}^2 \coloneqq \{(\mathcal{S}, \mathcal{T}) | (\mathcal{S}, \mathcal{T}) \in \mathbb{G}_H^2, \mathcal{S} \cup \mathcal{T} = \mathcal{R}\}$. $\mathbb{G}_{H,\mathcal{R}}^2 = \varnothing$ if and only if $|\mathcal{R}| > 2H$. $|\mathbb{G}_{H,\mathcal{R}}^2| > 1$ means that there are more than one way of the *Pcdhα* isoforms to be expressed from both chromosomal alleles to achieve the total expressed gene set $\mathcal{R}$. Define $N_{\mathcal{S},\mathcal{T}}$ as the number of single cells that the first chromosomal allele expresses gene set $\mathcal{S}$ and the second chromosomal allele expresses gene set $\mathcal{T}$. $N_{\mathcal{S},\mathcal{T}}$ is hidden. By definition, $N_\mathcal{R} = \sum_{(\mathcal{S},\mathcal{T}) \in \mathbb{G}_{H,\mathcal{R}}^2} N_{\mathcal{S},\mathcal{T}}$. Define $(N_{\mathcal{S},\mathcal{T}})_{\mathbb{G}_H^2} \coloneqq \{N_{\mathcal{S},\mathcal{T}} | (\mathcal{S}, \mathcal{T}) \in \mathbb{G}_H^2\}$ and $(N_\mathcal{R})_\mathcal{G} \coloneqq \{N_\mathcal{R} | \mathcal{R} \subset \mathcal{G}\}$ under the independent assumption $P_{\mathbb{G}_H^2}(\mathcal{S}, \mathcal{T}) = P_{\mathbb{G}_H}(\mathcal{S}) P_{\mathbb{G}_H}(\mathcal{T})$, where $P_{\mathbb{G}_H^2}$ and $P_{\mathbb{G}_H}$ are distributions (probability measures) on $\mathbb{G}_H^2$ and $\mathbb{G}_H$. We choose $P_{\mathbb{G}_H}$ which maximizes the likelihood $P[(N_\mathcal{R})_\mathcal{G} | P_{\mathbb{G}_H}]$. This is achieved by alternately maximizing the complete likelihood $P[(N_{\mathcal{S},\mathcal{T}})_{\mathbb{G}_H^2} | P_{\mathbb{G}_H}]$ over $P_{\mathbb{G}_H}$, and calculating the conditional expectation [69] as $E_{(N_{\mathcal{S},\mathcal{T}})_{\mathbb{G}_H^2} | (N_\mathcal{R})_\mathcal{G}, P_{\mathbb{G}_H^2}}(N_{\mathcal{S},\mathcal{T}})_{\mathbb{G}_H^2}$. To be exact, do $P_{\mathbb{G}_H}(\mathcal{S}) \propto \sum_{\mathcal{T} \in \mathbb{G}_H} N_{\mathcal{S},\mathcal{T}}$ and $N_{\mathcal{S},\mathcal{T}} \propto P_{\mathbb{G}_H}(\mathcal{S}) P_{\mathbb{G}_H}(\mathcal{T})$ until convergence. Note that the second equation is done under the constraint $N_\mathcal{R} = \sum_{(\mathcal{S},\mathcal{T}) \in \mathbb{G}_{H,\mathcal{R}}^2} N_{\mathcal{S},\mathcal{T}}$. We initially assume that $N_{\mathcal{S},\mathcal{T}} = N_\mathcal{R} / |\mathbb{G}_{H,\mathcal{R}}^2|$ for $(\mathcal{S}, \mathcal{T}) \in \mathbb{G}_{H,\mathcal{R}}^2$.

Polymer simulation of tandem-arrayed CTCF sites

We used a method to simulate long-distance chromatin interactions based on cohesin loop extrusion on a coarse-grained DNA fragment [19]. The modeled DNA fragment is divided into roughly equal bins. Long-distance chromatin interactions between one bin and all other bins are determined by their 3D distances according to the polymer simulation of cohesin loop extrusion.

"Two-headed" cohesin loop extrusion

Cohesin complex may extrude chromatin fiber individually [70] and asymmetrically [71, 72], or may even use the "inchworm" model [72, 73]. For clarity, we assume that cohesin loop extrusion with "two heads" as previously proposed [19]. Cohesin can be loaded stochastically on any location or in a specific position by NIPBL and start to extrude chromatin fibers in opposite directions. The extrusion process is continuous until blocked by oriented CBS which bound CTCF protein in an anti-parallel manner [10].

Coarse-grained polymer simulations

Based on the loop extrusion model, we simulate QHR-4C long-distance chromatin interactions according to the previous polymer modeling method, which is pioneered by the Mirny and Dekker laboratories, and assume the chromatin fiber as a polymer of 10-nm monomers each contains roughly three nucleosomes (about 600 bp) with excluded volume interactions and without topological constraints [19, 74]. We first divide the human *Pcdh* locus (chr5:140160700-140920300 of the GRCh37/hg19 assembly) into $L = 1266$ bins (monomers) each of about 600 bp in length for coarse-grained

simulations [19, 75]. Thus, the entire *Pcdh* locus is considered as a polymer containing 1266 monomers. The simulation consists of both 1D (one dimensional) lattice loop extrusion processes and 3D (three dimensional) polymer simulations with molecular dynamics.

1D lattice loop extrusion

In the 1D lattice loop extrusion, "two heads" of the cohesin (loop-extrusion factor) independently extrude a DNA loop in opposite directions in an ATP-dependent manner until blocked by CTCF insulators asymmetrically or dropping off from the coarse-grained chromatin fiber (Additional file 1: Figure S7g) [19, 76]. In addition, the cohesin ring cannot pass through each other during extrusion. Finally, CBS can block cohesin sliding in an orientation-dependent manner [11, 18, 19].

The concepts of cohesin separation and processivity are introduced to characterize loop extrusion [19, 20]. Accordingly, cohesin separation is the mean distance between consecutive sliding cohesin complexes on a chromatin fiber, and cohesin processivity λ is the mean size of the extruded loops. Specifically, for L bins and separation d, the number of cohesins on the *Pcdh* locus is calculated as $\lfloor L/d \rfloor$. The initial locations of these cohesins are determined according to the loading probabilities inferred from the NIPBL ChIP-seq data. Both heads of a cohesin either occupy the same bin or two adjacent bins with a probability of 0.5 for each. Different cohesins cannot occupy the same bin. At each step, a cohesin may drop off from the chromatin fiber or polymer with the probability $2/\lambda$, where λ is the processivity. If one cohesin drops off, a new one will be immediately loaded to the polymer according to the loading probabilities from the NIPBL ChIP-seq data but avoiding existing ones. This keeps the number of cohesin complexes unchanged for the *Pcdh* locus. Finally, both "heads" of a cohesin complex can extrude through a bin if it is unoccupied by CTCF or another cohesin.

We determine cohesin loading by calculating the coverage of the NIPBL ChIP-seq for each bin. Eighty percent of cohesins load to the chromatin fiber according to the probabilities proportional to NIPBL coverages of bins, and 20% load randomly [77]. We design the following two methods to calculate CBS permeability or CTCF occupancy for cohesin loop extrusion. The first one is based on ChIP-seq experimental data for CTCF occupancy. The second one is based on dynamic interactions between CTCF and its genomic target sites [9, 10, 56, 57].

Estimation of permeability of bins based on CTCF ChIP-seq data

Since cohesin accumulates at CBS only when it is occupied by CTCF proteins [78], it has been established that CTCF binding strength of a site can be translated into cohesin permeability of that site [19]. The orientations of bins are determined by CTCF sites within the bins. The CTCF sites are called by the FIMO program [79] from the experimental CTCF ChIP-seq data in the *Pcdh* locus. We first map CTCF ChIP-seq reads to the *Pcdh* cluster by Bowtie2 [80]. The CTCF occupancies (cohesin stalling probability) are called by MACS2 [68]. Each has a fold enrichment value x.

If a bin contains CTCF sites in only one orientation, it stalls opposite cohesins with the probability $\mathcal{T} = \frac{1}{1+\exp(-\frac{x}{\zeta}-\mu)}$ for $x > 0$ and 0 for $x = 0$, where $\zeta = 40$, $\mu = 4$, and x is the CTCF enrichment [19]. If a bin contains CTCF sites in both orientations, it stalls cohesins in both directions with stalling probabilities calculated separately. The CTCF occupancy and cohesin permeability of the *Pcdh* locus in the CBS-inserted clones are estimated similarly according to their ChIP-seq data.

Estimation permeability of cohesin sliding through oriented CTCF array with no ChIP-seq data available

It was recently reported that CTCF binding to dsDNA is much more dynamic than cohesin and that the residence time of cohesin on DNA fiber is at least 10-fold more than CTCF [9]. The dynamic binding of CTCF to oriented CTCF sites provides hindrance for cohesin sliding [9, 10, 56, 57]. In this scenario, the permeability is calculated as follows. If there are n consecutive CTCF sites c_1, c_2, $\cdots$, c_n, from distal to proximal, with a permeability of p_1, p_2, $\cdots$, p_n, respectively, we want to know the mean attempting times x_n for cohesin ring to slide through the entire CBS array from proximal to distal. For the first attempt, the proximal CBS has a probability p_n to allow cohesin ring to pass through. Thus, the cohesin needs x_{n-1} attempting times on average to slide through the remaining CBS array c_1, c_2, $\cdots$, c_{n-1}. Otherwise, the proximal CBS c_n has the probability $(1 - p_n)$ to block cohesin ring passing through. Thus, one attempting time has been used and cohesin still needs x_n attempting times on average to slide through the entire CBS array c_1, c_2, $\cdots$, c_n. In summary:

$$x_n = p_n x_{n-1} + (1-p_n)(1 + x_n).$$

Since $x_0 = 1$, by mathematical induction, $x_n = \sum_{i=1}^{n} 1/p_i - n + 1$. Then, one obtains the overall permeability $1/x_n$ of CTCF sites c_1, c_2, $\cdots$, c_n.

3D polymer simulations

Lennard-Jones (LJ) reduced units

Bins are considered as monomers with diameter σ and mass m. The Langevin equation:

$$m\frac{d^2r}{dt^2} = -\nabla U - \gamma\frac{dr}{dt} + \sqrt{2k_BT\gamma}\eta(t)$$

is rescaled to:

$$\frac{d^2r}{dt^2} = -\nabla U - \alpha\gamma\frac{dr}{dt} + \sqrt{2\alpha\gamma}\eta(t)$$

by LJ reduced units [81] that m, σ, k_BT, and $(\sigma^2 m/k_BT)^{1/2}$ are units of mass, distance, energy, and time, respectively, where $\alpha = \frac{\sigma}{(mk_BT)^{1/2}}$. We set $m = 100$ Da, $\sigma = 1$ nm, $T = 300$ K, and $\gamma = 0.01$ ps^{-1}m according to previous reports [19].

Bonds in the reduced units

The repulsive potential is defined as previously described [19].

$$U_{\mathrm{REP}} = \mathrm{REPe}\left\{1 + \left(\frac{r\mathrm{REPrmin}}{\mathrm{REPsigma}}\right)^{12}\left[\left(\frac{r\mathrm{REPrmin}}{\mathrm{REPsigma}}\right)^{2} - 1\right]/\mathrm{REPemin}\right\},$$

where REPe = 1.5, $\mathrm{REPrmin} = \sqrt{6/7}$, REPsigma = 1.05, and $\mathrm{REPemin} = \frac{46656}{823543}$. Harmonic bond $U_{\mathrm{HAR}} = k(r-d)^2$ is used between adjacent monomers with $k = 100$ and $d = 1$, and cohesin-bounded monomers with $k = 25$ and $d = 0.5$. The polymer stiffness is described by $U_{\mathrm{STI}} = 2(1 - \cos\theta)$.

Langevin velocity Verlet algorithm

The time step $\Delta t = 80$ ts [19]. The velocities v, forces f, and positions r of monomers are updated by the Langevin velocity Verlet algorithm [82].

$$v = v + \frac{\Delta t}{2f} + b\Delta w,$$

$$r = r + cv,$$

$$f = f(r),$$

$$v = av + b\Delta w + \frac{\Delta t}{2f},$$

where $a \coloneqq \frac{2-\alpha\gamma\Delta t}{2+\alpha\gamma\Delta t}$, $b \coloneqq \sqrt{\alpha\gamma\Delta t/2}$, and $c \coloneqq \frac{2\Delta t}{2+\alpha\gamma\Delta t}$.

Simulation of QHR-4C data process

We simulated long-distance chromatin interaction profiles between a viewpoint of interest and its target genomic regions by coarse-grained modeling. We first transform the experimental contact frequencies from restriction fragments to coarse-grained bins of 600 bp.

Unlike Hi-C and 5C data, 4C data with different viewpoints, even for the same cell types, cannot be compared directly because of their inconsistent scales. Assume viewpoint $i \in [1, I]$ forms J_i valid pairs (i, j) for $j \in [1, J_i]$. Let u_{ij} be the contact frequency of pair (i, j). We choose k_i for $i \in [1, I]$ and α minimizing the geometric standard deviation of $\frac{u_{ij}}{k_i s_{ij}^{-\alpha}}$ (the contact frequency decreases with the 1D distance roughly in power law [19])

$$\mathrm{GSD} \coloneqq \exp\left\{\sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}\left[\log\left(\frac{u_{ij}}{k_i s_{ij}^{-\alpha}}\right) - \beta\right]^2}{J}}\right\},$$

where β is the mean of $\log(\frac{u_{ij}}{k_i s_{ij}^{-\alpha}})$ and $J \coloneqq \sum_{i=1}^{I} J_i$. α and $\log k_i$ solve the linear algebra:

$$\begin{aligned}\frac{\partial\,\mathrm{logGSD}}{\partial\alpha} &= \sum_{i=1}^{I}\sum_{j=1}^{J_i} 2\,\log u_{ij}\left(\log s_{ij} - \beta\right) \\ &+ \alpha\sum_{i=1}^{I}\sum_{j=1}^{J_i} 2\,\log s_{ij}\left(\log s_{ij} - \beta\right) \\ &+ \sum_{i=1}^{I}\log k_i\sum_{j=1}^{J_i}(-2)\left(\log s_{ij} - \beta\right) = 0,\end{aligned}$$

$$\begin{aligned}\frac{\partial\,\mathrm{logGSD}}{\partial\,\log k_w} &= \sum_{i=1}^{I}\sum_{j=1}^{J_i} 2\,\log u_{ij}\left(-\delta_{i,w} + J_w/J\right) \\ &+ \alpha\sum_{i=1}^{I}\sum_{j=1}^{J_i} 2\,\log s_{ij}\left(-\delta_{i,w} + J_w/J\right) \\ &+ \sum_{i=1}^{I}\log k_i\sum_{j=1}^{J_i}(-2)\left(-\delta_{i,w} + J_w/J\right) = 0.\end{aligned}$$

Without loss of generality, fix $k_1 = 1$ to remove the redundancy among the equations of $\log k_i$ for $i \in [1, I]$. Finally, divide u_{ij} by k_i to obtain comparable 4C contact frequencies.

Since there are data of two biological replicates available, both mean and variance of contact frequencies are calculated for each pair. Finally, pairs with 0 mean contact frequency are excluded from the fitting of 4C simulation by the approach of relative maximum entropy.

Relative maximum entropy approach to correct polymer simulations by rescaled QHR-4C data

In statistical mechanics, the 3D conformations of the polymer are microstates, which cannot be observed directly in experiments. In single-cell experiments, some macroscopic variables, such as contact strength between monomers, can be observed for single microstates. In multiple-cell experiments, only the mean of macroscopic variables over an ensemble of microstates can be observed. As an inverse problem, inferring the distribution of microstates from the macroscopic variables can be achieved in two different ways. The first is the maximum entropy approach. One searches for the best in all microstate distributions which coincide with the observed macroscopic variables, and choose the distribution with

the maximum entropy. The justification for this is that one should introduce as little information as possible other than that from the direct observation. The second is the model-based simulation. One sets up a computation model, such as the cohesin loop extrusion, and simulates many microstates. The difficulty comes from the parameter choice. Generally, novel methods are used to optimize parameters by minimizing the differences between the macroscopic variables calculated from the simulated microstates and those observed from experiments. Depending on the problem, the optimization process can be extremely hard and achieve very limited improvements.

The advantage of the maximum entropy approach is that the predicted distribution of microstates resulting in the same macroscopic variables as the observations. The disadvantage is that it abandons all known central mechanisms, such as the cohesin loop extrusion. On the contrary, the model-based simulation includes known mechanisms to set up the model but predicts the macroscopic variables usually deviating from the observations. We use the relative maximum entropy approach [83] that combines the advantages of both the maximum entropy approach and the model-based simulations. The basic idea is quite similar to the maximum entropy approach. One searches the best in all microstate distributions which coincide with the observed macroscopic variables. The difference is that the distribution with the maximum entropy relative to that determined by the underlying model, namely the relative entropy instead of the entropy, is chosen. The entropy is actually the relative entropy to the uniform distribution, which is the least informative distribution. If an underlying model is set up based on new information, then the least informative distribution will be determined by the underlying model. The relative entropy, or the negative Kullback-Leibler divergence, is a measurement of the similarity between two distributions. The relative maximum entropy approach actually selects a distribution closest to the least informative one among those satisfying the experimental observations.

We apply the relative maximum entropy approach [83] to correct polymer simulations by the rescaled QHR-4C data. Pairs with 0 contact frequencies in both replicates are excluded. In simulations, long-distance chromatin interactions between the bin of viewpoint and all other bins are determined by their spatial distances in the 3D conformation and the capture radius $\mathbb{C}$ (the distance at which two monomers are determined to be in contact). Let $k \coloneqq \arg\min_{k'>0} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \left(\overline{u}_{ij} - \frac{\max(0, p_{ij} - \hat{p})}{k'} \right)^2$ be a multiplier transforming the contact frequencies to the contact probabilities, where $\overline{u}_{ij}$ is the mean contact frequencies over the rescaled 4C replicates, p_{ij} is the contact probabilities of simulations, and $\hat{p}$ is the median contact probabilities over all pairs in the *Pcdh* locus. We subtract $\hat{p}$ to remove the abnormally high background contact probabilities due to the small period box for simulations. Specifically, we initialize the polymer of length $L = 1266$ by a compact conformation [84] (a cubic lattice) in a period box of size $\left(\frac{L}{\rho}\right)^{\frac{1}{3}} \approx 18.4984$ with the monomer density $\rho = 0.2$. Even for the relatively short capture radius of 2, the background contact probabilities in such a small box are much higher than those observed in 4C data.

Let $P_0(q)$ be the distribution of the 3D conformation q of the *Pcdh* locus determined by the underlying loop extrusion model. Let $c_{ij}(q) = 1$ if monomers i and j are within the capture radius, and $c_{ij}(q) = 0$ otherwise. To prevent overfitting, we assume independent Gaussian errors $\epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij}; 0, \sigma_{ij}^2)$ with variance $\sigma_{ij}^2 \coloneqq \max[\sigma_{\min}^2, \tilde{\sigma}_{ij}^2]$ for the contact probabilities of (i, j), where $\tilde{\sigma}_{ij}^2$ is the variance of ku_{ij} over experimental replicates, and $\sigma_{\min}^2$ is the minimally allowed variance. Then the union distribution $Q_0(q, \epsilon)$ is:

$$Q_0(q, \epsilon) = P_0(q) \prod_{j=1}^{J_i} \mathcal{N}\left(\epsilon_{ij}; 0, \sigma_{ij}^2\right).$$

Force $Q(q, \epsilon)$ to reproduce the experimentally observed mean contact probabilities, i.e.:

$$\int \left(c_{ij}(q) + \epsilon_{ij}\right) Q(q, \epsilon) dq d\epsilon = \xi_{ij} \coloneqq \min\left(1, k\overline{u}_{ij} + \min\left(p_{ij}, \hat{p}\right)\right),$$

while maximizing the relative entropy:

$$\mathcal{S}[Q][Q_0] \coloneqq -\int Q(q, \epsilon) \log[Q(q, \epsilon)/Q_0(q, \epsilon)] dq d\epsilon. \tag{1}$$

By the variational methods [83]:

$$Q(q, \epsilon) \propto Q(q, \epsilon; \lambda) \coloneqq \exp\left(-\sum_{i=1}^{I} \sum_{j=1}^{J_i} \lambda_{ij} \left[c_{ij}(q) + \epsilon_{ij}\right] \right) Q_0(q, \epsilon), \tag{2}$$

where λ is determined by Eq. (1) and the normalization restraint $\int Q(q, \epsilon) dq d\epsilon = 1$. This λ must minimize [83]:

$$\Gamma(\lambda) = \log\left[\int Q(q,\epsilon;\lambda)dqd\epsilon\right] + \sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}\epsilon_{ij}$$
$$= \log\left[\int \exp\left(-\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}c_{ij}(q)\right)P_0(q)dq\right]$$
$$+\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}\epsilon_{ij} + \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}\sigma_{ij}^2.$$

The gradient and Hessian of $\Gamma(\lambda)$ are:

$$\frac{\partial\Gamma}{\partial\lambda_{ij}} = \epsilon_{ij} - \langle c_{ij}(q)\rangle + \lambda_{ij}\sigma_{ij}^2,$$

$$\frac{\partial^2\Gamma}{\partial\lambda_{(i_1j_1)}\partial\lambda_{i_2j_2}} = \langle c_{i_1j_1}(q)c_{i_2j_2}(q)\rangle - \langle c_{i_1j_1}(q)\rangle\langle c_{i_2j_2}(q)\rangle + \delta_{i_1i_2}\delta_{j_1j_2}\sigma_{i_1j_1}^2,$$

where for arbitrary function f(q), define:

$$\langle f(q)\rangle \coloneqq \frac{\int f(q)\,\exp\left(-\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}c_{ij}(q)\right)P_0(q)dq}{\int \exp\left(-\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}c_{ij}(q)\right)P_0(q)dq}.$$

Sample conformations q_1, q_2, q_3, ⋯, q_N from the distribution $P_0(q)$ determined by the underlying model by simulations. Then:

$$\langle f(q)\rangle \approx \frac{\sum_{n=1}^{N} f(q_n)\,\exp\left(-\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}c_{ij}(q_n)\right)}{\sum_{n=1}^{N}\exp\left(-\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}c_{ij}(q_n)\right)}. \quad (3)$$

$\sigma_{\min}^2 > 0$ promises the strictly positive definition of Hessian, thereby the optimization is strictly convex. Increasing $\sigma_{\min}^2$ not only speeds up the convergence, but also keeps $|\lambda_{ij}|$ small, thereby avoiding overfitting. However, it also extracts less information from the experiments. Therefore, we set $\sigma_{\min}^2 = 0.01$ and solve λ by the trust region algorithm.

Optimization of processivity, separation, and capture radius

We set both processivity and separation to 100, 200, or 400 [19]. For each pair of processivity and separation, we do the following simulations. First, we anneal the loop extrusion dynamics by 1,000,000 1D simulation time steps. We then anneal the 3D dynamics by 2000 blocks, each of which contains one 1D simulation time step and 1250 3D simulation time steps. Finally, we simulate 50,000 blocks and obtain 50,000 conformations.

The above process is repeated twice to obtain 100,000 conformations for each pair of processivity and separation. We then use the relative maximum entropy approach to calculate $\min_{\lambda}\Gamma(\lambda)$ for each pair of processivity and separation and each capture radius of 2, 3, or 4. The pair of processivity 400 and separation 200, which maximizes the average $\min_{\lambda}\Gamma(\lambda)$ for capture radius of 2, 3, or 4, is considered as optimal because by Eq. (2):

$$\mathcal{S}[Q][Q_0] = -\int Q(q,\epsilon)\left\{\log\left[\frac{Q(q,\epsilon;\lambda)}{Q_0(q,\epsilon)}\right] - \log\left[\int Q\left(q',\epsilon';\lambda\right)dq'd\epsilon'\right]\right\}dqd\epsilon$$
$$= \int\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}\left[c_{ij}(q)+\epsilon_{ij}\right]Q(q,\epsilon)dqd\epsilon + \log\left[\int Q(q,\epsilon;\lambda)dqd\epsilon\right]$$
$$= \log\left[\int \exp\left(-\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}c_{ij}(q)\right)P_0(q)dq\right]$$
$$+\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}\left(\langle c_{ij}(q)\rangle - \lambda_{ij}\sigma_{ij}^2\right)$$
$$+\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}^2\sigma_{ij}^2 \approx \log\left[\sum_{n=1}^{N}\exp\left(-\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}c_{ij}(q_n)\right)\right] - \log N$$
$$+\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}\left(\langle c_{ij}(q)\rangle - \lambda_{ij}\sigma_{ij}^2\right) + \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\lambda_{ij}^2\sigma_{ij}^2.$$

As convergence, $\frac{\partial\Gamma}{\partial\lambda_{ij}} = \xi_{ij} - \langle c_{ij}(q)\rangle + \lambda_{ij}\sigma_{ij}^2 = 0$, thereby $\mathcal{S}[Q][Q_0] \approx \Gamma(\lambda)$. The optimal parameter set is repeated 26 times for each sample to obtain 1,300,000 conformations. $\hat{\lambda} \coloneqq \mathrm{argmin}_{\lambda}\,\Gamma(\lambda)$ obtained from the wild-type sample by the relative maximum entropy is then used to weight conformations of the mutant samples by Eq. (3) to obtain final entropy-corrected contact probabilities $\tilde{p}_{ij}$.

Hulu model

Consider a simple genomic region containing four convergent CBS elements (two forward CBS elements followed by two reverse ones) with spaces of 100 bins and stalling probabilities $\mathcal{T} = 0.97$. We assume that cohesins mainly load between the inner convergent CBS pair (30 times faster than other locations), two heads of a cohesin advance in the same speed, and the processivity is large enough. We simulate 210,000 conformations for this model region with processivity 400 and separation 400. The contact map $\mathbb{H}$ is generated by capture radius 2. To give an intuitive expression of the Hulu structure, we transformed $\mathbb{H}$ to a distance matrix $\mathbb{D}$ by $d_{ij} = h_{ij}^{-1}$ and applied the non-metric multidimensional scaling with Kruskal's normalized stress-1 criterion.

Genome-wide insulator analyses by Bayesian networks

Bayesian networks are a powerful and widely used probabilistic model to infer the underlying conditional dependency of factors shared by a group of instances. In our case, each instance is a promoter, which has four factors: the enhancer strength, the insulator strength, the loop strength, and the promoter activity. Bayesian networks are a non-cyclic directed graph which uses nodes to represent factors and arrows to

connect them. The networks are learned by maximizing the posterior likelihood. An arrow from the insulator strength to the promoter activity means it is a direct dependence. We analyzed 207,663 enhancer-promoter contacts of the capture Hi-C data for the genome-wide relationship between insulators, enhancers, and promoters [46]. For each bait promoter fragment, containing a promoter whose activity is represented by the expression level f_i, with starting chromosomal coordinate s_i and terminating coordinate t_i, we denote it by $[s_i, t_i]$. It forms long-distance chromatin contacts, measured as loop counts l_{ij} in the capture Hi-C experiments, with a putative enhancer fragment $[s_j, t_j]$. The enhancer strength e_j of the fragment $[s_j, t_j]$ is defined as its total H3K27ac signals from ChIP-seq experiments. The insulator strength u_{ij} of the loop is defined as the total CTCF ChIP signals in the interval $[\min(t_i, t_j) + a, \max(s_i, s_j) - a]$ (a = 500 bp to exclude the rare cases that promoters or enhancers themselves contain CTCF binding sites) if $\min(t_i, t_j) + a \le \max(s_i, s_j) - a$, and zero otherwise. Let $\mathcal{T}_i$ be the set of enhancer fragments which have chromatin contacts with the bait promoter fragment $[s_i, t_i]$. The total enhancer strength for the bait promoter $[s_i, t_i]$ is defined as $E_i \coloneqq \sum_{j\in\mathcal{T}_i} e_j$. The mean chromatin looping strength is defined as $L_i \coloneqq \frac{\sum_{j\in\mathcal{T}_i} e_j l_{ij}}{E_i}$. The mean insulator strength is defined as $U_i \coloneqq \frac{\sum_{j\in\mathcal{T}_i} e_j u_{ij}}{E_i}$. Finally, we use the ranking of the above variables on day 0, day 3, and day 6 to discrete them. For example, let f_i^0, f_i^3, and f_i^6 be the expression levels of the promoter $[s_i, t_i]$ in days 0, 3, and 6, respectively, with $f_i^3 < f_i^0 < f_i^6$. Then, we set $f_i^0 = 2$, $f_i^3 = 1$, and $f_i^6 = 3$ to learn the structure of the Bayesian network by the following method.

Let $\mathbf{X} \coloneqq \{X_1, X_2, \cdots, X_n\}$ be the set of n discrete random variables. $\mathbf{x} \coloneqq \{x_1, x_2, \cdots, x_n\}$ is the specific value of $\mathbf{X}$. x_i^k for $1 \le k \le r_i$ are the r_i possible values of X_i. Given a network structure S, $\mathbf{Pa}_i^S \subset \mathbf{X}$ are the parents of X_i, and $\mathbf{pa}_i^S \subset \mathbf{x}$ are the corresponding specific value. $\mathbf{pa}_i^{S,j}$ for $1 \le j \le q_i^S$ are the q_i^S possible values of $\mathbf{Pa}_i^S$. Define:

$$\boldsymbol{\theta}^S \coloneqq \cup_{i=1}^n \boldsymbol{\theta}_i^S \coloneqq \cup_{i=1}^n \cup_{j=1}^{q_i^S} \boldsymbol{\theta}_{ij}^S \coloneqq \cup_{i=1}^n \cup_{j=1}^{q_i^S} \cup_{k=1}^{r_i} \left\{\theta_{ijk}^S\right\},$$

$$\boldsymbol{\alpha}^S \coloneqq \bigcup_{i=1}^{n} \boldsymbol{\alpha}_i^S \coloneqq \bigcup_{i=1}^{n} \bigcup_{j=1}^{q_i^S} \boldsymbol{\alpha}_{ij}^S \coloneqq \bigcup_{i=1}^{n} \bigcup_{j=1}^{q_i^S} \bigcup_{k=1}^{r_i} \left\{\alpha_{ijk}^S\right\},$$

where $\theta_{ijk}^S > 0$, $\alpha_{ijk}^S > 0$, $\sum_{k=1}^{r_i} \theta_{ijk}^S = 1$. Introduce the independence assumption.

$$p(\boldsymbol{\theta}^S|\boldsymbol{\alpha}^S, S) = \prod_{i=1}^{n}\prod_{j=1}^{q_i^S} p\left(\boldsymbol{\theta}_{ij}^S|\boldsymbol{\alpha}_{ij}^S, S\right).$$

Assume $p(x_i^k|\mathbf{pa}_i^{S,j}, \boldsymbol{\theta}_i^S, S) = \theta_{ijk}^S$ and $p(\boldsymbol{\theta}_{ij}^S|\boldsymbol{\alpha}_{ij}^S, S) = \mathcal{D}(\boldsymbol{\theta}_{ij}^S|\boldsymbol{\alpha}_{ij}^S)$, where $\mathcal{D}(\boldsymbol{\theta}_{ij}^S|\boldsymbol{\alpha}_{ij}^S)$ is the Dirichlet distribution of $\boldsymbol{\theta}_{ij}^S$ with parameter $\boldsymbol{\alpha}_{ij}^S$.

$D \coloneqq \{d_l | 1 \le l \le m\}$ are m samples. d_{li} and $\mathbf{pa}_{li}^S$ are the values of variable i and its parents in sample l, respectively. Define:

$$\delta_{ijk}^{S,l} \coloneqq \begin{cases} 1, \mathbf{pa}_{li}^S = \mathbf{pa}_i^{S,j}, d_{li} = x_i^k, \\ 0, \text{otherwise}, \end{cases} \quad N_{ijk}^{S,l} \coloneqq \sum_{l'=1}^{l-1} \delta_{ijk}^{S,l'},$$

$$\mathbf{N}_{ij}^{S,l} = \left\{N_{ijk}^{S,l} | 1 \le k \le r_i\right\}, D_l \coloneqq \left\{d_{l'} | 1 \le l' < l\right\}.$$

$N_{ijk}^{S,l}$ is the number of samples in D_l with variable i taking the kth value x_i^k and its parents taking the jth value $\mathbf{pa}_i^{S,j}$. Then, it is well known that $p(\boldsymbol{\theta}_{ij}^S|D_l, \boldsymbol{\alpha}_{ij}^S, S) = \mathcal{D}(\boldsymbol{\theta}_{ij}^S|\boldsymbol{\alpha}_{ij}^S + \mathbf{N}_{ij}^{S,l})$. Also:

$$\begin{aligned} p(\boldsymbol{\theta}_i^S|D_l, \boldsymbol{\alpha}^S, S) &= \frac{p(D_l|\boldsymbol{\theta}_i^S, S)p(\boldsymbol{\theta}_i^S|\boldsymbol{\alpha}^S, S)}{p(D_l|\boldsymbol{\alpha}^S, S)} \\ &= \frac{\left[\prod_{j=1}^{q_i^S}\prod_{k=1}^{r_i}\left(\theta_{ijk}^S\right)^{N_{ijk}^{S,l}}\right]\left[\prod_{j=1}^{q_i^S} p\left(\boldsymbol{\theta}_{ij}^S|\boldsymbol{\alpha}^S, S\right)\right]}{p(D_l|\boldsymbol{\alpha}^S, S)} \\ &= \frac{\left[\prod_{j=1}^{q_i^S} p\left(D_l|\boldsymbol{\theta}_{ij}^S, S\right)\right]\left[\prod_{j=1}^{q_i^S} p\left(\boldsymbol{\theta}_{ij}^S|\boldsymbol{\alpha}^S, S\right)\right]}{p(D_l|\boldsymbol{\alpha}^S, S)} \\ &= \frac{\left[\prod_{j=1}^{q_i^S} p\left(D_l, \boldsymbol{\theta}_{ij}^S|\boldsymbol{\alpha}^S, S\right)\right]}{p(D_l|\boldsymbol{\alpha}^S, S)} = \prod_{j=1}^{q_i^S} \mathcal{D}\left(\boldsymbol{\theta}_{ij}^S|\boldsymbol{\alpha}_{ij}^S + \mathbf{N}_{ij}^{S,l}\right). \end{aligned}$$

Thus:

$$\begin{aligned} p(D|\boldsymbol{\alpha}^S, S) &= \prod_{l=1}^{m}\prod_{i=1}^{n}\int d\boldsymbol{\theta}_i^S p(d_{li}|\mathbf{pa}_{li}^S, \boldsymbol{\theta}_i^S, S)p(\boldsymbol{\theta}_i^S|D_l, \boldsymbol{\alpha}^S, S) \\ &= \prod_{l=1}^{m}\prod_{i=1}^{n}\int d\boldsymbol{\theta}_i^S \left[\prod_{j=1}^{q_i^S}\prod_{k=1}^{r_i}\left(\theta_{ijk}^S\right)^{\delta_{ijk}^{S,l}}\right]\left[\prod_{j=1}^{q_i^S}\mathcal{D}\left(\boldsymbol{\theta}_{ij}^S|\boldsymbol{\alpha}_{ij}^S + \mathbf{N}_{ij}^{S,l}\right)\right] \\ &= \prod_{l=1}^{m}\prod_{i=1}^{n}\int\left[\prod_{j=1}^{q_i^S} d\boldsymbol{\theta}_{ij}^S\right]\left[\prod_{j=1}^{q_i^S}\prod_{k=1}^{r_i}\left(\theta_{ijk}^S\right)^{\delta_{ijk}^{S,l}}\right]\left[\prod_{j=1}^{q_i^S}\frac{\Gamma\left(\left|\boldsymbol{\alpha}_{ij}^S + \mathbf{N}_{ij}^{S,l}\right|\right)}{\prod_{k=1}^{r_i}\Gamma\left(\alpha_{ijk}^S + N_{ijk}^{S,l}\right)}\prod_{k=1}^{r_i}\left(\theta_{ijk}^S\right)^{\alpha_{ijk}^S + N_{ijk}^{S,l}-1}\right] \\ &= \prod_{l=1}^{m}\prod_{i=1}^{n}\left[\prod_{j=1}^{q_i^S}\frac{\Gamma\left(\left|\boldsymbol{\alpha}_{ij}^S + \mathbf{N}_{ij}^{S,l}\right|\right)}{\prod_{k=1}^{r_i}\Gamma\left(\alpha_{ijk}^S + N_{ijk}^{S,l}\right)}\right]\prod_{j=1}^{q_i^S}\int d\boldsymbol{\theta}_{ij}^S\prod_{k=1}^{r_i}\left(\theta_{ijk}^S\right)^{\delta_{ijk}^{S,1}+\alpha_{ijk}^S+N_{ijk}^{S,l}-1} \\ &= \prod_{l=1}^{m}\prod_{i=1}^{n}\prod_{j=1}^{q_i^S}\frac{\Gamma\left(\left|\boldsymbol{\alpha}_{ij}^S + \mathbf{N}_{ij}^{S,l}\right|\right)}{\Gamma\left(\left|\boldsymbol{\alpha}_{ij}^S + \mathbf{N}_{ij}^{S,l+1}\right|\right)}\prod_{k=1}^{r_i}\frac{\Gamma\left(\alpha_{ijk}^S + N_{ijk}^{S,l+1}\right)}{\Gamma\left(\alpha_{ijk}^S + N_{ijk}^{S,l}\right)} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{q_i^S}\frac{\Gamma\left(\left|\boldsymbol{\alpha}_{ij}^S\right|\right)}{\Gamma\left(\left|\boldsymbol{\alpha}_{ij}^S + \mathbf{N}_{ij}^{S}\right|\right)}\prod_{k=1}^{r_i}\frac{\Gamma\left(\alpha_{ijk}^S + N_{ijk}^{S}\right)}{\Gamma\left(\alpha_{ijk}^S\right)}, \end{aligned}$$

where $N_{ijk}^S \coloneqq \sum_{l=1}^m \delta_{ijk}^{S,l}$ and $\mathbf{N}_{ij}^S = \{N_{ijk}^S | 1 \le k \le r_i\}$.

The independence assumption is valid by assuming prior modularity, marginal likelihood equivalence, and Dirichlet More importantly, there exists $\boldsymbol{\alpha} \coloneqq \{\alpha_{\mathbf{x}} | \mathbf{x} \in \mathcal{X}\}$ independent of S, such that $\alpha_{ijk}^S = \sum_{\mathbf{x}\in\mathcal{X}_{ijk}^S} \alpha_{\mathbf{x}}$ $\forall S$, where $\mathcal{X} \coloneqq \otimes_{i=1}^n \mathcal{X}_i$, $\mathcal{X}_i \coloneqq$

$\{x_i^k | 1 \le k \le r_i\}$ and $\mathcal{X}_{ijk}^S \coloneqq \{\mathbf{x} \in \mathcal{X} | x_i = x_i^k, \mathbf{pa}_i^S = \mathbf{pa}_i^{S,j}\}$. So, $p(D|\boldsymbol{\alpha}^S, S) = p(D|\boldsymbol{\alpha}, S)$. For simplicity, we assume $\alpha_{\mathbf{x}} \equiv \alpha$ (uniform priors) and $\boldsymbol{\alpha} = 1$ (limited prior information). The best structure is defined as $S^* \coloneqq \mathrm{argmax}_S p(S|D, \boldsymbol{\alpha}) = \mathrm{argmax}_S p(D|\boldsymbol{\alpha}, S)p(S)/p(D|\boldsymbol{\alpha})$. For simplicity, assume that $p(S)$ is uniformly distributed over all possible structures. To find S^*, we first transform the data into an all-dimensions tree and then apply the max-min hill-climbing (MMHC) algorithm [45].

Statistics and reproducibility

All statistical tests used were performed using R 3.5 and Microsoft Excel. All of the statistical tests used are described in the relevant text. p values are provided as exact values where possible and otherwise are reported as a range. All of QHR-4C, 5C, and RNA-seq experiments were performed with at least two biological replicates. Single-cell CRISPR CBS insertion clones and their corresponding mutant clones were screened for at least two clones for each genotype.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-01984-7.

Additional file 1: Figure S1. Stochastic and monoallelic expression of the *Pcdha* genes in single cells. **Figure S2.** A sensitive QHR-4C method for one-to-all capture of chromosome conformations. **Figure S3.** Both forward and reverse CBS elements inserted between the *Pcdha* cluster and its downstream *HS5-1* enhancer function as insulators. **Figure S4.** Reverse CBS elements inserted between *Pcdh a13* and *ac1* function as an insulator for the upstream genes. **Figure S5.** Reverse-forward CBS pair as an insulator for the *Pcdha* genes. **Figure S6.** Reverse-forward tandem CBS pairs as an insulator for the *Pcdha* genes. **Figure S7.** Forward-reverse convergent CTCF sites do not compromise their insulation activity. **Figure S8.** Polymer simulations of the chromatin looping interaction profiles upon CBS insertions or their mutations in the *Pcdh* and *Igh* clusters. **Figure S9.** Tandem CTCF sites ensure stochastic and balanced *Pcdh* gene expression. **Figure S10.** Topology of spatial chromatin contacts between the *Pcdh* β and γ clusters and the downstream super-enhancer. **Figure S11.** Topology of spatial chromatin contacts between the *Pcdh*γ clusters and the downstream super-enhancer. **Figure S12.** Genotyping of the mouse lines of various *HS5-1* CBS deletions and inversions. **Figure S13.** Tandem CTCF sites function as insulators for enhancers with no CBS.

Additional file 2: Table S1. Oligonucleotides in this study. **Table S2.** CRISPR deletion and inversion mice. **Table S3.** CRISPR single-cell clones. **Table S4.** Mapping statistics of the 5C data. **Table S5.** Pearson correlations between 5C replicates.

Additional file 3: Review history.

Acknowledgements

We thank Drs M. Capecchi, D. Czajkowsky, C. Hou, and T. Maniatis for the critical reading of the manuscript.

Review history

This manuscript was previously reviewed in another journal. The review history can be found as Additional file 3.

Additional information

Barbara Cheifet was the primary editor of this manuscript and managed its editorial process and peer review with collaboration with the rest of the editorial team.

Authors' contributions

QW conceived the research. ZJ and XG, assisted by YW and YG, did the experimental work. JL, assisted by ZJ and YW, performed the mathematical simulation and computational modeling. ZJ, JL, and QW wrote the manuscript with inputs from all authors. ZJ and QW would like to dedicate this paper to the memory of Wuhan victims of COVID-19 during this difficult time. The authors read and approved the final manuscript.

Funding

This work was supported by grants from the National Natural Science Foundation of China (31630039 and 31700666), the Ministry of Science and Technology of China (2017YFA0504203 and 2018YFC1004504), and the Science and Technology Commission of Shanghai Municipality (19JC1412500).

Availability of data and materials

High-throughput sequencing files (QHR-4C, RNA-seq, and ChIP-seq) have been deposited into the NCBI Gene Expression Omnibus (GEO) database with the accession number GSE138646 [85]. 5C data are available from the Sequence Read Archive (SRA) under the accession number PRJNA576991 [86]. The codes for 1D lattice and 3D polymer simulations of tandem-arrayed CTCF sites, maximum-likelihood modeling of *Pcdh* stochastic and monoallelic expression, and genome-wide insulator analyses by Bayesian networks are available at GitHub (https://github.com/ljw20180420/balance_codes) [87].

Ethics approval and consent to participate

All animal experiments were approved by the Institutional Animal Care and Use Committee (IACUC) of Shanghai Jiao Tong University (protocol#: 1602029).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 4 February 2020 Accepted: 4 March 2020
Published online: 23 March 2020

References

1. Müller HJ. Types of visible variations induced by X-rays in *Drosophila*. J Genet. 1930;22:299–334.
2. Phillips-Cremins JE, Corces VG. Chromatin insulators: linking genome organization to cellular function. Mol Cell. 2013;50:461–74.
3. Dekker J, Mirny L. The 3D genome as moderator of chromosomal communication. Cell. 2016;164:1110–21.
4. Furlong EEM, Levine M. Developmental enhancers and chromosome topology. Science. 2018;361:1341–5.
5. Grosveld F, van Assendelft GB, Greaves DR, Kollias G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. Cell. 1987;51:975–85.
6. Chung JH, Whiteley M, Felsenfeld G. A 5′ element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila*. Cell. 1993;74:505–14.
7. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell. 1999;98:387–96.
8. Ghirlando R, Felsenfeld G. CTCF: making the right connections. Genes Dev. 2016;30:881–91.
9. Hansen AS, Pustova I, Cattoglio C, Tjian R, Darzacq X. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. Elife. 2017;6:25776.
10. Xu D, Ma R, Zhang J, Liu Z, Wu B, Peng J, et al. Dynamic nature of CTCF tandem 11 zinc fingers in multivalent recognition of DNA as revealed by NMR spectroscopy. J Phys Chem Lett. 2018;9:4020–8.
11. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.

12. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell. 2015;162:900–10.
13. Rudan MV, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 2015;10:1297–309.
14. de Wit E, Vos ESM, Holwerda SJB, Valdes-Quezada C, Verstegen MJAM, Teunissen H, et al. CTCF binding polarity determines chromatin looping. Mol Cell. 2015;60:676–84.
15. Hnisz D, Weintraub AS, Day DS, Valton AL, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016;351:1454–8.
16. Narendra V, Bulajic M, Dekker J, Mazzoni EO, Reinberg D. CTCF-mediated topological boundaries during development foster appropriate gene regulation. Genes Dev. 2016;30:2657–62.
17. Merkenschlager M, Nora EP. CTCF and cohesin in genome folding and transcriptional gene regulation. Annu Rev Genomics Hum Genet. 2016;17:17–43.
18. Sanborn AL, Rao SSP, Huang SC, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A. 2015;112:E6456–E65.
19. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of chromosomal domains by loop extrusion. Cell Rep. 2016;15:2038–49.
20. Nuebler J, Fudenberg G, Imakaev M, Abdennur N, Mirny LA. Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc Natl Acad Sci U S A. 2018;115:E6697–E706.
21. Nora EP, Goloborodko A, Valton AL, Gibcus JH, Uebersohn A, Abdennur N, et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. Cell. 2017;169:930–44 e22.
22. Rao SSP, Huang SC, St Hilaire BG, Engreitz JM, Perez EM, Kieffer-Kwon KR, et al. Cohesin loss eliminates all loop domains. Cell. 2017;171:305–20.
23. Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25.
24. Hou C, Zhao H, Tanimoto K, Dean A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. Proc Natl Acad Sci U S A. 2008;105:20398–403.
25. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature. 2016;529:110–4.
26. Wu Q, Maniatis T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell. 1999;97:779–90.
27. Lefebvre JL, Kostadinov D, Chen WV, Maniatis T, Sanes JR. Protocadherins mediate dendritic self-avoidance in the mammalian nervous system. Nature. 2012;488:517–21.
28. Toyoda S, Kawaguchi M, Kobayashi T, Tarusawa E, Toyama T, Okano M, et al. Developmental epigenetic modification regulates stochastic expression of clustered protocadherin genes, generating single neuron diversity. Neuron. 2014;82:94–108.
29. Schreiner D, Weiner JA. Combinatorial homophilic interaction between gamma-protocadherin multimers greatly expands the molecular diversity of cell adhesion. Proc Natl Acad Sci U S A. 2010;107:14893–8.
30. Chen WV, Nwakeze CL, Denny CA, O'Keeffe S, Rieger MA, Mountoufaris G, et al. Pcdhalphac2 is required for axonal tiling and assembly of serotonergic circuitries in mice. Science. 2017;356:406–11.
31. Fan L, Lu YC, Shen XL, Shao H, Suo L, Wu Q. Alpha protocadherins and Pyk2 kinase regulate cortical neuron migration and cytoskeletal dynamics via Rac1 GTPase and WAVE complex in mice. Elife. 2018;7:35242.
32. Mountoufaris G, Canzio D, Nwakeze CL, Chen WV, Maniatis T. Writing, reading, and translating the clustered protocadherin cell surface recognition code for neural circuit assembly. Annu Rev Cell Dev Biol. 2018;34:471–93.
33. Jain S, Ba Z, Zhang Y, Dai HQ, Alt FW. CTCF-binding elements mediate accessibility of RAG substrates during chromatin scanning. Cell. 2018;174:102–16 e14.
34. Kehayova P, Monahan K, Chen W, Maniatis T. Regulatory elements required for the activation and repression of the protocadherin-alpha gene cluster. Proc Natl Acad Sci U S A. 2011;108:17195–200.
35. Guo Y, Monahan K, Wu H, Gertz J, Varley KE, Li W, et al. CTCF/cohesin-mediated DNA looping is required for protocadherin alpha promoter choice. Proc Natl Acad Sci U S A. 2012;109:21081–6.
36. Allahyar A, Vermeulen C, Bouwman BAM, Krijger PHL, Verstegen M, Geeven G, et al. Enhancer hubs and loop collisions identified from single-allele topologies. Nat Genet. 2018;50:1151–60.
37. Li J, Shou J, Guo Y, Tang Y, Wu Y, Jia Z, et al. Efficient inversions and duplications of mammalian regulatory DNA elements and gene clusters by CRISPR/Cas9. J Mol Cell Biol. 2015;7:284–98.
38. Shou J, Li J, Liu Y, Wu Q. Precise and predictable CRISPR chromosomal rearrangements reveal principles of Cas9-mediated nucleotide insertion. Mol Cell. 2018;71:498–509.
39. Esumi S, Kakazu N, Taguchi Y, Hirayama T, Sasaki A, Hirabayashi T, et al. Monoallelic yet combinatorial expression of variable exons of the protocadherin-alpha gene cluster in single neurons. Nat Genet. 2005;37:171–6.
40. Tasic B, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature. 2018;563:72–8.
41. Cai HN, Shen P. Effects of cis arrangement of chromatin insulators on enhancer-blocking activity. Science. 2001;291:493–5.
42. Muravyova E, Golovnin A, Gracheva E, Parshikov A, Belenkaya T, Pirrotta V, et al. Loss of insulator activity by paired Su(Hw) chromatin insulators. Science. 2001;291:495–8.
43. Srinivasan M, Scheinost JC, Petela NJ, Gligoris TG, Wissler M, Ogushi S, et al. The cohesin ring uses its hinge to organize DNA using non-topological as well as topological mechanisms. Cell. 2018;173:1508–19 e18.
44. Yokota S, Hirayama T, Hirano K, Kaneko R, Toyoda S, Kawamura Y, et al. Identification of the cluster control region for the protocadherin-beta genes located beyond the protocadherin-gamma cluster. J Biol Chem. 2011;286:31885–95.
45. Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn. 2006;65:31–78.
46. Rubin AJ, Barajas BC, Furlan-Magaril M, Lopez-Pajares V, Mumbach MR, Howard I, et al. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. Nat Genet. 2017;49:1522–8.
47. Monahan K, Horta A, Lomvardas S. LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. Nature. 2019;565:448–53.
48. Canzio D, Nwakeze CL, Horta A, Rajkumar SM, Coffey EL, Duffy EE, et al. Antisense lncRNA transcription mediates DNA demethylation to drive stochastic protocadherin alpha promoter choice. Cell. 2019;177:639–53 e15.
49. Wu Q, Zhang T, Cheng JF, Kim Y, Grimwood J, Schmutz J, et al. Comparative DNA sequence analysis of mouse and human protocadherin gene clusters. Genome Res. 2001;11:389–404.
50. Tanimoto K, Liu Q, Bungert J, Engel JD. Effects of altered gene order or orientation of the locus control region on human beta-globin gene expression in mice. Nature. 1999;398:344–8.
51. Busslinger GA, Stocsits RR, van der Lelij P, Axelsson E, Tedeschi A, Galjart N, et al. Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. Nature. 2017;544:503–7.
52. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell. 2015;163:1611–27.
53. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell. 2013;152:1173–83.
54. Zhang Y, Zhang X, Ba Z, Liang Z, Dring EW, Hu H, et al. The fundamental role of chromatin loop extrusion in physiological V(D)J recombination. Nature. 2019;573:600–4.
55. Tarjan DR, Flavahan WA, Bernstein BE. Epigenome editing strategies for the functional annotation of CTCF insulators. Nat Commun. 2019;10:4258.
56. Hashimoto H, Wang D, Horton JR, Zhang X, Corces VG, Cheng X. Structural basis for the versatile and methylation-dependent binding of CTCF to DNA. Mol Cell. 2017;66:711–20 e3.
57. Yin ML, Wang JY, Wang M, Li XM, Zhang M, Wu Q, et al. Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. Cell Res. 2017;27:1365–77.
58. Vian L, Pekowska A, Rao SSP, Kieffer-Kwon KR, Jung S, Baranello L, et al. The energetics and physiological impact of cohesin extrusion. Cell. 2018;173:1165–78.
59. Schwartzman O, Mukamel Z, Oded-Elkayam N, Olivares-Chauvet P, Lubling Y, Landan G, et al. UMI-4C for quantitative and targeted chromosomal contact profiling. Nat Methods. 2016;13:685–91.
60. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome conformation capture carbon copy (5C): a massively parallel

solution for mapping interactions between genomic elements. Genome Res. 2006;16:1299–309.
61. Dostie J, Dekker J. Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc. 2007;2:988–1002.
62. Lajoie BR, van Berkum NL, Sanyal A, Dekker J. My5C: web tools for chromosome conformation capture studies. Nat Methods. 2009;6:690–1.
63. Lu Y, Shou J, Jia Z, Wu Y, Li J, Guo Y, et al. Genetic evidence for asymmetric blocking of higher-order chromatin structure by CTCF/cohesin. Protein Cell. 2019;10:914–20.
64. Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc. 1988;83:596–610.
65. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109–13.
66. Smith EM, Lajoie BR, Jain G, Dekker J. Invariant TAD boundaries constrain cell-type-specific looping interactions between promoters and distal elements around the CFTR locus. Am J Hum Genet. 2016;98:185–201.
67. Kundu S, Ji F, Sunwoo H, Jain G, Lee JT, Sadreyev RI, et al. Polycomb repressive complex 1 generates discrete compacted domains that change during differentiation. Mol Cell. 2017;65:432–46 e5.
68. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
69. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Methodol. 1977;39:1–38.
70. Hansen AS, Cattoglio C, Darzacq X, Tjian R. Recent evidence that TADs and chromatin loops are dynamic structures. Nucleus. 2018;9:20–32.
71. Nichols MH, Corces VG. A CTCF code for 3D genome architecture. Cell. 2015;162:703–5.
72. Yatskevich S, Rhodes J, Nasmyth K. Organization of chromosomal DNA by SMC complexes. Annu Rev Genet. 2019;53:445–82.
73. Nichols MH, Corces VG. A tethered-inchworm model of SMC DNA translocation. Nat Struct Mol Biol. 2018;25:906–10.
74. Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny LA, et al. Organization of the mitotic chromosome. Science. 2013;342:948–53.
75. Marko JF, Siggia ED. Polymer models of meiotic and mitotic chromosomes. Mol Biol Cell. 1997;8:2217–31.
76. Alipour E, Marko JF. Self-organization of domain structures by DNA-loop-extruding enzymes. Nucleic Acids Res. 2012;40:11202–12.
77. Schwarzer W, Abdennur N, Goloborodko A, Pekowska A, Fudenberg G, Loe-Mie Y, et al. Two independent modes of chromatin organization revealed by cohesin removal. Nature. 2017;551:51–6.
78. Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, et al. Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell. 2008;132:422–33.
79. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017–8.
80. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
81. Rapaport DC. The art of molecular dynamics simulation. 2nd ed. Cambridge: Cambridge University Press; 2004.
82. Kröger M. Models for polymeric and anisotropic liquids. 2005th edn. Berlin: Springer; 2005.
83. Cesari A, Reißer S, Bussi G. Using the maximum entropy principle to combine simulations and solution experiments. Computation. 2018;6:15–39.
84. Imakaev MV, Tchourine KM, Nechaev SK, Mirny LA. Effects of topological constraints on globular polymers. Soft Matter. 2015;11:665–71.
85. Jia Z, Li J, Ge X, Wu Y, Guo Y, Wu Q. Tandem CTCF sites function as insulators to balance spatial contacts and topological enhancer-promoter selection. Datasets Gene Expression Omnibus. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138646.
86. Jia Z, Li J, Ge X, Wu Y, Guo Y, Wu Q. Tandem CTCF sites function as insulators to balance spatial contacts and topological enhancer-promoter selection. Datasets Sequence Read Archive. 2020. https://www.ncbi.nlm.nih.gov/sra/PRJNA576991.
87. Jia Z, Li J, Ge X, Wu Y, Guo Y, Wu Q. Tandem CTCF sites function as insulators to balance spatial contacts and topological enhancer-promoter selection. Computational Codes. 2020. https://github.com/ljw20180420/balance_codes.

Publisher's Note