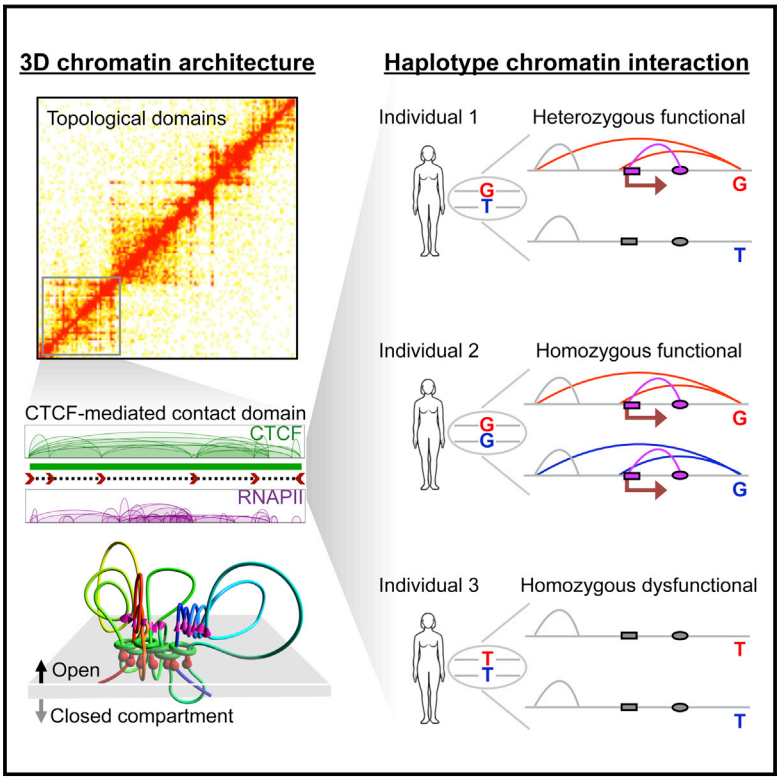
Article

Cell

CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription

Graphical Abstract



Authors

Zhonghui Tang, Oscar Junhong Luo, Xingwang Li, ..., Dariusz Plewczynski, Guoliang Li, Yijun Ruan

Correspondence

yijun.ruan@jax.org

In Brief

Advanced ChIA-PET shows that CTCF/cohesin and RNA polymerase II arrange spatial organization for coordinated transcription. Haplotype variants exhibit allelic effects on chromatin topology and transcription that link disease susceptibility.

Highlights

- ChIA-PET is inclusive in mapping 3D genome at multi-scale and nucleotide resolution
- CTCF foci spatially arrange RNAPII transcription concordant in CTCF-motif direction
- SNPs alter haplotype chromatin topology and function that link to disease risks
- 3D genome models elucidate topological framework for transcriptional regulation

Accession Numbers

GSE72816

Tang et al., 2015, Cell *163*, 1611–1627
December 17, 2015
http://dx.doi.org/10.1016/j.cell.2015.11.024

CellPress

Cell

CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription

Zhonghui Tang,[1,12] Oscar Junhong Luo,[1,12] Xingwang Li,[1,2,12] Meizhen Zheng,[1] Jacqueline Jufen Zhu,[1,3] Przemyslaw Szalaj,[4,5,6] Pawel Trzaskoma,[7] Adriana Magalska,[7] Jakub Wlodarczyk,[7] Blazej Ruszczycki,[7] Paul Michalski,[1] Emaly Piecuch,[1,3] Ping Wang,[1] Danjuan Wang,[1] Simon Zhongyuan Tian,[1] May Penrad-Mobayed,[8] Laurent M. Sachs,[9] Xiaoan Ruan,[1] Chia-Lin Wei,[10] Edison T. Liu,[1] Grzegorz M. Wilczynski,[7] Dariusz Plewczynski,[6] Guoliang Li,[2,11] and Yijun Ruan[1,2,3,*]

[1]The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06030, USA
[2]National Key Laboratory of Crop Genetic Improvement, College of Life Sciences and Technology, Huazhong Agricultural University, Wuhan, Hubei 430070, China
[3]Department of Genetics and Genome Sciences, University of Connecticut Health Center, 400 Farmington Avenue, Farmington, CT 06030, USA
[4]Center for Bioinformatics and Data Analysis, Medical University of Bialystok, ul. Jana Kilinskiego 1, 15-089 Bialystok, Poland
[5]I-BioStat, Hasselt University, Agoralaan Building D, 3590 Diepenbeek, Belgium
[6]Centre of New Technologies, University of Warsaw, S. Banacha 2c, 02-097 Warsaw, Poland
[7]Nencki Institute of Experimental Biology, 3 Pasteur Street, 02-093 Warsaw, Poland
[8]Université Paris-Diderot-Paris 7, Centre National de la Recherche Scientifique and Institut Jacques Monod, 15 rue Hélène Brion, 75205 Paris Cedex, France
[9]Centre National de la Recherche Scientifique and Muséum National d'Histoire Naturelle, 57 Rue Cuvier, 75231 Paris Cedex, France
[10]Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA
[11]College of Informatics, Huazhong Agricultural University, Wuhan, Hubei 430070, China
[12]Co-first author
*Correspondence: yijun.ruan@jax.org
http://dx.doi.org/10.1016/j.cell.2015.11.024

SUMMARY

Spatial genome organization and its effect on transcription remains a fundamental question. We applied an advanced chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) strategy to comprehensively map higher-order chromosome folding and specific chromatin interactions mediated by CCCTC-binding factor (CTCF) and RNA polymerase II (RNAPII) with haplotype specificity and nucleotide resolution in different human cell lineages. We find that CTCF/cohesin-mediated interaction anchors serve as structural foci for spatial organization of constitutive genes concordant with CTCF-motif orientation, whereas RNAPII interacts within these structures by selectively drawing cell-type-specific genes toward CTCF foci for coordinated transcription. Furthermore, we show that haplotype variants and allelic interactions have differential effects on chromosome configuration, influencing gene expression, and may provide mechanistic insights into functions associated with disease susceptibility. 3D genome simulation suggests a model of chromatin folding around chromosomal axes, where CTCF is involved in defining the interface between condensed and open compartments for structural regulation. Our 3D genome strategy thus provides unique insights in the topological mechanism of human variations and diseases.

INTRODUCTION

The human genome consists of more than 3 billion nucleotides, spanning over 2 m in length. Packaging this genomic material within the micrometer-sized nuclear space requires extensive folding (Bickmore, 2013). Such folding is presumed to be both specific and functional (Ong and Corces, 2014). However, details regarding general folding principles, distinct topologies, and/or relationships to gene activity are still largely unknown.

Current technologies in studying 3D structures of the human genome include 3D fluorescence in situ hybridization (3D-FISH) nuclear imaging and 3D genome mapping. 3D-FISH can visualize realistic chromosome conformation and individual contacts within nucleus (Cremer et al., 2008). However, it lacks sufficient genomic detail and accuracy. The core strategy in 3D genome mapping is nuclear proximity ligation (Cullen et al., 1993), which allows detection of distant genomic segments residing in close spatial proximity to one another, yet are linearly far away. Using this strategy, a number of high-throughput methods have been developed for genome-wide chromatin interaction mapping, including chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) and high-throughput chromosome conformation capture (Hi-C) (Fullwood et al., 2009; Lieberman-Aiden et al., 2009). ChIA-PET was designed to detect genome-wide chromatin interactions mediated by specific protein factors, whereas Hi-C was developed to capture all chromatin contacts. Hi-C has been proven effective for mapping

large-scale structures, such as topologically associated domains (TAD) (Dixon et al., 2012); however, it lacks the resolution to detect precise functional interactions mediated by proteins. In contrast, by inclusion of chromatin immunoprecipitation (ChIP), ChIA-PET is unique in detecting protein factor-mediated chromatin interactions and is capable of generating high-resolution (~100 bp) genome-wide chromatin interaction maps with binding-site specificity among functional elements in human and mouse (Li et al., 2012; Zhang et al., 2013).

To comprehensively characterize the 3D topology of chromatin interactions between functional elements and higher-order organization in the human genome, we applied ChIA-PET, targeting on two protein factors, CCCTC-binding zinc finger protein (CTCF) and RNA polymerase II (RNAPII) in a number of human cell lines. CTCF is the master weaver of genome organization (Ong and Corces, 2014), and Hi-C studies further correlated CTCF binding at TAD boundaries (Dixon et al., 2012; Rao et al., 2014). RNAPII is involved in transcription of all protein-coding and many non-coding genes (Sims et al., 2004). Therefore, comprehensive analyses of chromatin interactions mediated by these two factors have the potential to reveal the overall relationship between organizational structure and transcriptional function. Here, we demonstrate that ChIA-PET is inclusive for mapping both ChIP-enriched and non-enriched chromatin contacts with haplotype specificity and nucleotide resolution, and we uncover detailed chromatin topology that provides the framework for regulating transcriptional activity.

RESULTS

ChIA-PET Is Multifaceted for Chromatin Interaction Mapping

In addition to the original ChIA-PET data deposited in the ENCODE project (ENCODE Project Consortium, 2012), we have generated new CTCF- and RNAPII-mediated chromatin interaction datasets using an improved ChIA-PET protocol (Figure S1A) for longer reads (2 × 150 bp). Altogether, we collected 364 million uniquely mapped ChIA-PET reads in 12 ChIA-PET libraries from four human cell lines: GM12878, HeLa, K562, and MCF7 for analysis (Table S1).

A ChIA-PET experiment delivers paired-end-tag (PET) sequencing data from self-ligation and inter-ligation products (Figures 1A and S1B). The self-ligation PET data identify ChIP-enriched protein-binding sites. The clustered inter-ligation PET data detect enriched interactions mediated by the ChIP targeted protein factor, whereas the singleton inter-ligation data reflect higher-order topological proximity, similar to Hi-C data (Figures S1B–S1E). Therefore, in theory, the multifaceted ChIA-PET data are ideal for comprehensive 3D genome mapping.

Recently, a study using in situ Hi-C generated 4.9 billion contact reads in GM12878 cells and found the majority of chromatin interaction loops to be associated with CTCF-binding sites (Rao et al., 2014). We compared this dataset with our CTCF ChIA-PET and demonstrated that the two datasets displayed very similar contact patterns (Figure 1B). In addition, ChIA-PET identified specific CTCF loops with binding-site resolution at 100s bp level. Importantly, only a single HiSeq 2500 rapid mode run or even a MiSeq run was sufficient for ChIA-PET to reach the same output as in situ Hi-C, plus 10-fold higher resolution. Thus, ChIA-PET is cost-effective, inclusive, and reproducible (Data S1, I; Supplemental Experimental Procedures) of generating multi-layer mapping data, capturing protein binding sites and enriched functional chromatin interactions, as well as non-enriched chromatin contacts for higher-order chromosomal conformations.

CTCF Organizes Chromatin Contact Domain into CTCF Foci

CTCF-mediated chromatin interactions are pervasive across the entire genome in the four tested human cell lines (Figure S2A). Based on the CTCF interaction clustering scheme, we identified 53,741 high quality CTCF-mediated interactions in GM12878 cells (Figures S2B–S2E; Supplemental Experimental Procedures). Considering that cohesin protein complex is highly associated with CTCF in chromatin biology (Ong and Corces, 2014), we examined the CTCF-anchor sites for co-occupancy by cohesin using ChIP sequencing (ChIP-seq) data of subunits RAD21 and SMC3. The vast majority (99%) of the CTCF interactions had cohesin co-occupancy in either one (n = 10,952) or both anchors (n = 42,297). Moreover, interactions with cohesin support on both anchors had significantly higher contact frequency than those with cohesin only on one anchor (Figure 2A, left). Since the sites with co-occupancy of CTCF and cohesin represent the most biologically relevant regions in our study, we focused further analyses on this subset.

We investigated the CTCF motifs in anchors of CTCF loops identified in this study. Of the 42,297 interactions, ~83% (n = 35,230) had CTCF motifs in both anchors with unique orientation. Among these, 33.1% were in tandem (i.e., tandem loop) and 64.5% (n = 22,709) in convergent orientation (i.e., convergent loop) (Figures 2A, right, and S2F; Table S2; Supplemental Experimental Procedures). Thus, CTCF-mediated chromatin loops have a clear orientation preference (convergent) that represents strong interactions with high contact frequency (Figure 2A, right), in agreement with in situ Hi-C data (Rao et al., 2014). In addition, tandem loops were present in significant numbers with intermediate interacting strength and span (Figures 2A and S2G), and most of them (82%) were positioned within convergent loops, perhaps representing a subgroup with possible supplemental function. A possible reason for the tandem loops discovered in our study but not in Hi-C is likely due to the power of specific enrichment in CTCF ChIA-PET experiments (Table S3; Supplemental Experimental Procedures).

To further understand how cohesin cooperates with CTCF in chromatin biology, we analyzed ChIP-seq data of RAD21 and SMC3 and suggest that cohesin may surround the CTCF occupancy but have a preference toward the 3′ side of CTCF motif (Data S1, II). To more precisely define CTCF/cohesin co-occupancy, we performed ChIP-nexus (He et al., 2015) to identify the specific borders of DNA footprints bound by cohesin in relation to CTCF. The RAD21 ChIP-nexus data detected one 5′ border that is very close to the 5′ border of CTCF and two 3′ border sites with the weak one matching exactly to the CTCF 3′ border and the stronger one being 40 bp downstream (Figure 2B; Data S1, II). Similar patterns were observed for SMC3

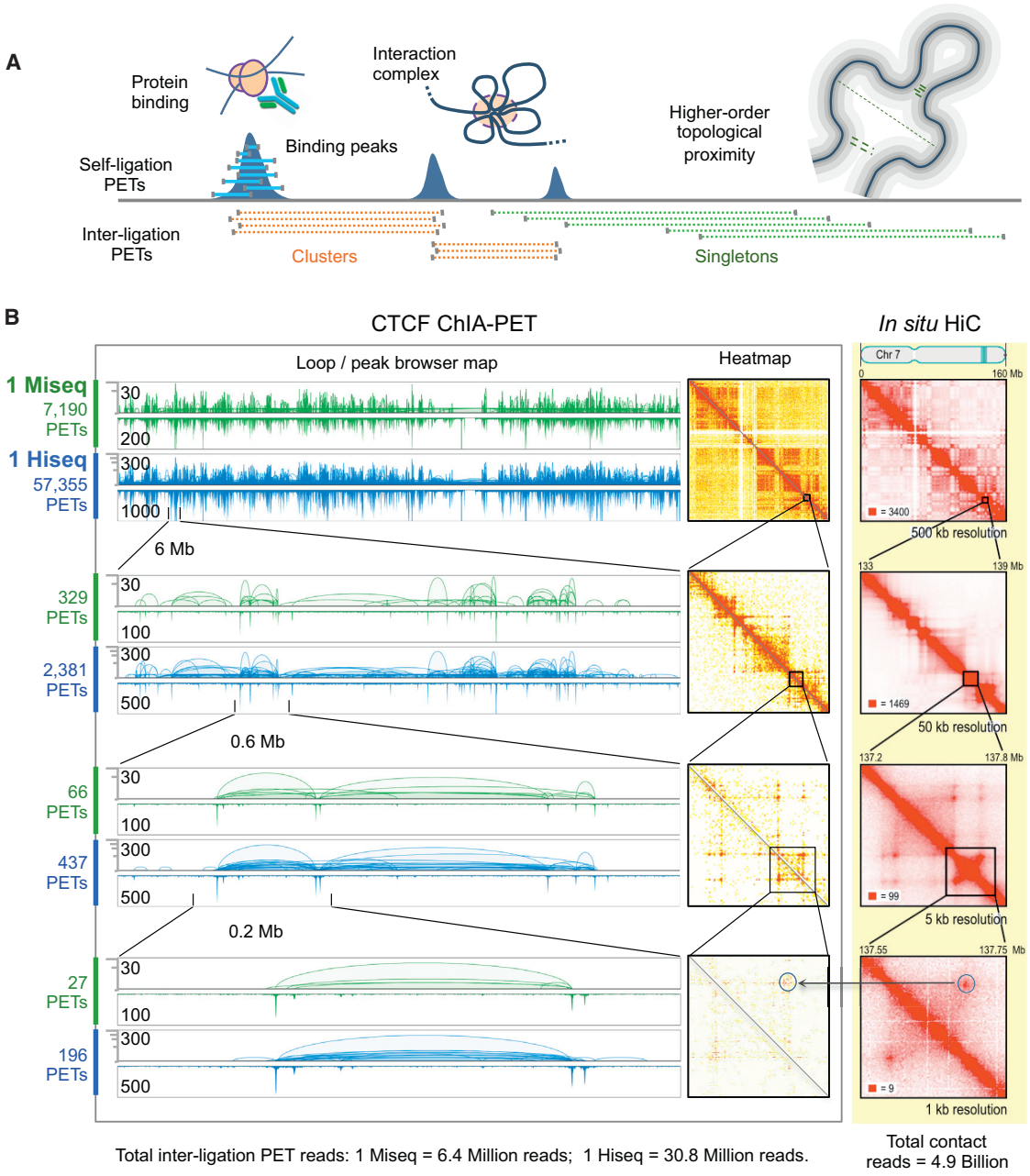


Figure 1. Characteristics of ChIA-PET Data for 3D Genome Mapping

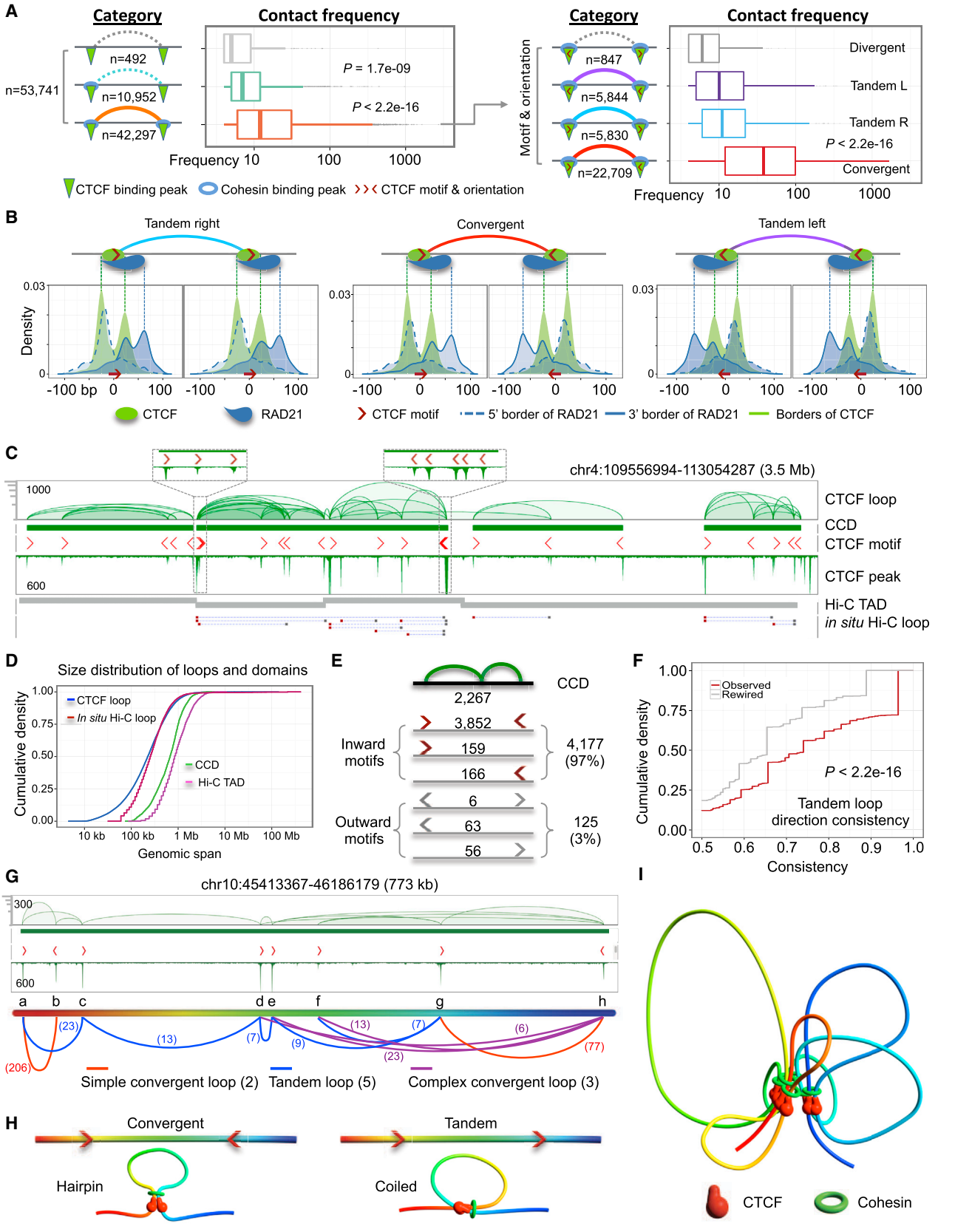
(A) Graphic of ChIA-PET mapping properties including binding peaks, enriched chromatin interactions, and non-enriched singleton PETs inferring topological neighborhood proximity.

(B) Comparison between CTCF ChIA-PET and in situ Hi-C data (GM12878). Left: loop/peak map views of CTCF ChIA-PET data at different zoom-in scopes. For each data track, loop view is at top, peak view at bottom; y axis indicates the contact frequency of loops (log10 scale) and intensity of binding peaks (linear scale). The maximum frequency and intensity are given in each data track. PET counts on the left side of each track show the numbers of interaction PETs detected in the given region. Middle and right: CTCF ChIA-PET contact heatmap and matched zoom-in regions to the in situ Hi-C contact heatmap (Rao et al., 2014). Total numbers of sequence reads generated for the in situ Hi-C data, and the CTCF ChIA-PET data are given at the bottom.

See also Figure S1, Table S1, and Data S1.

(Data S1, II). Collectively, these results suggest part of the cohesin ring complex directly overlap with the CTCF binding around the motif and embrace additional space downstream in the 3′ direction.

Many CTCF/cohesin-mediated loops often interconnect and continuously cover large chromatin segments (Figure 2C). Based on connectivity and contact frequency (Figure S3A; Supplemental Experimental Procedures), we identified 2,267 CTCF-mediated



(legend on next page)

chromatin contact domains (CCDs) in GM12878 cells (Figures 2D and S3B). Comparison to Hi-C data showed that CTCF ChIA-PET and non-selective Hi-C were highly concordant in detecting chromatin domain structures (Figures 2C, 2D, S3C, and S3D), indicating that CTCF-associated chromatin interactions are abundant in human 3D nucleome.

At the 4,534 boundaries of the 2,267 CCDs, the majority contained inward-facing CTCF-binding motifs (Figures 2E and S3E). Many CCD boundaries contained multiple CTCF-binding peaks with inward-facing motifs (Figures 2C, insets, and S3F), suggesting a possible double-knot-tie mechanics for tightening domain end structures (Figures S3G–S3I; Supplemental Experimental Procedure). In contrast, tandem loops were evenly distributed within the CCD space (Figure S3I) and showed high consistency in motif orientation (Figures 2F and S3J).

Interconnected CTCF binding and looping with convergent and tandem motif orientations constitute the finer details in chromatin topology. Figure 2G illustrates a CCD composed of multiple CTCF loops. Theoretically, CTCF dimerization with motifs could be either symmetric or asymmetric. Since the cohesin ring complex is most likely bound toward the 3′ downstream of CTCF motif (Figure 2B), we speculate that CTCF dimerization with two interacting motifs favors symmetric conformation. Therefore, a pair of convergent motifs would form a "hairpin loop," while a pair of tandem motifs form a "coiled loop" (Figure 2H). This principle may have critical topological implications for the 3D structure of CCD and the overall chromosomal topology and genome organization. The example CCD in Figure 2G shows all eight CTCF anchors (a–h) to be interconnected, thus, suggesting anchors in this CCD form an aggregated scaffold of "CTCF/cohesin focus" (Figure 2I). Consequently, the overall properties of genome organization would be collectively shaped by the 2,267 CTCF/cohesin foci (2,267 CCDs) in GM12878 cells.

RNAPII Transcription Factories Are Spatially Associated with CTCF/Cohesin Foci

To functionally characterize CTCF/cohesin-mediated chromatin topology, we overlaid functional genomics data (RNAPII ChIA-PET, chromatin state, and RNA-sequencing [RNA-seq]) with CCD footprint on the genome landscape (Figure 3A). The results indicated that most transcription activity occurs within CTCF-looped chromatin structures (Figure S4A). Most of the RNAPII-associated loops are smaller than CTCF loops (Figure S4B), and the vast majority of RNAPII-looping structures are included within CCD-defined genomic space (Figures S4B and S4C; Table S2).

To dissect the associations with transcription, we divided CTCF/cohesin-bound chromatin loop structures into "anchor" and "loop" regions and then examined their epigenomic and transcriptional features (Supplemental Experimental Procedures). Unlike the loop regions, the anchors were enriched with active epigenomic markers, RNAPII occupancy and the presence of transcription start site (TSS) (Figure 3B), suggesting that CTCF-anchor regions are the foci for transcriptional activity. Unexpectedly, further detailed analyses focusing on anchors uncovered structure-function features related to transcription activity and directionality at three distinct levels. First, signals of active epigenomic markers tend to be higher toward 3′ direction of CTCF motif for both convergent and tandem loops (Figure 3C). More strikingly, TSS at anchor regions showed clear strand-specificity and directional enrichment along with CTCF motif orientation, indicating that a sub-group of genes are pre-positioned within the CTCF-defined anchor regions with their promoters in harmony with CTCF motif orientation, thus dictating the directionality of transcription. Second, active epigenomic markers, RNAPII, and TSS densities were highly enriched around the anchors of tandem loops compared to the convergent loops (Figure 3C). The same trend was also observed for B cell-specific transcription factors (TFs), e.g., ELF1 and ZEB1, but not for chromatin structural proteins CTCF, RAD21, and ZNF143 (Figure 3D). Third, the paired anchors involved in tandem loops also exhibited directionality: the "head" anchor tends to have higher signal intensities of active epigenomic markers, RNAPII, and B cell-specific TF binding than the "tail" anchor (Figures 3C, 3D, and S4D), while there was no such notable difference for the anchors of convergent loops. It implies that the head anchor in

Figure 2. CTCF-Defined Chromatin Looping Topology

(A) Characterization of CTCF-mediated loops in relation to cohesin binding and CTCF motif orientation. See also Figures S2E–S2G.

(B) CTCF and RAD21 binding patterns centered on CTCF motif. ChIP-Exo data of CTCF (Rhee and Pugh, 2011) and ChIP-nexus data of RAD21 were plotted as density curve around CTCF-motif sites. Borders of DNA footprints were identified by occupancy peaks. The green peaks depict the two borders of CTCF occupancy. The dashed blue line shows the peak position depicting 5′ border of RAD21 footprint. The solid blue line shows bimodal peaks from the 3′ border of RAD21 occupancy. See also Data S1, II.

(C) A mapping browser screenshot shows the CTCF-defined chromatin interactions and contact domains. Hi-C determined TADs (Dixon et al., 2012) and in situ Hi-C identified loops (Rao et al., 2014) are also shown. CTCF-motif position and orientation at the corresponding interaction anchors are shown as red arrows. Inset: zoom-in regions highlight CCD boundaries having multiple CTCF-binding peaks and motifs with inward-facing orientation.

(D) Cumulative density plot shows the genomic span distribution of individual CTCF loops, CCDs, in situ Hi-C loops and Hi-C TADs. See also Figures S3C and S3D.

(E) Statistics of CTCF-motif orientation at CCD boundaries. See also Figure S3E.

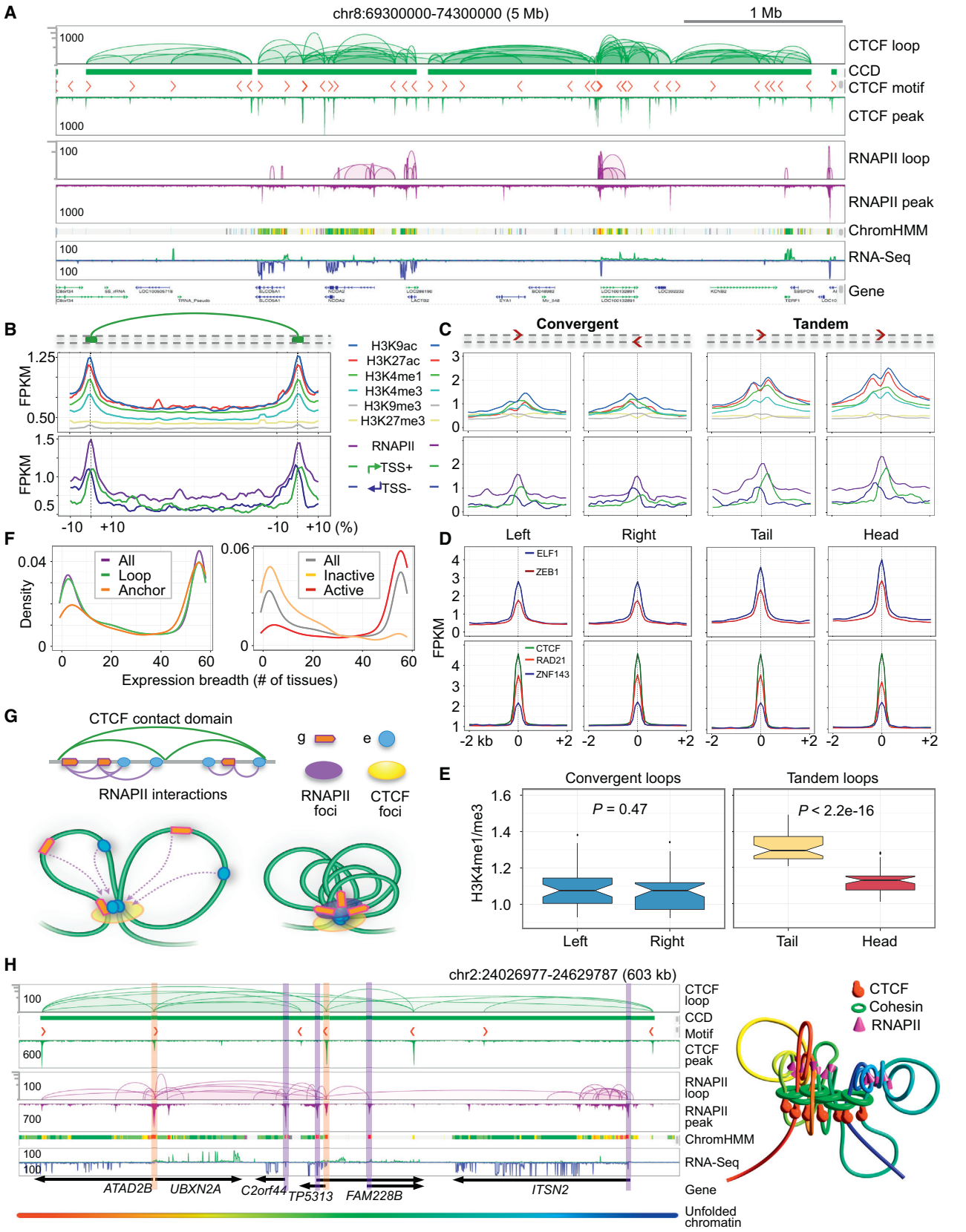
(F) Cumulative density of the consistency of motif orientation for tandem loops resided within the same CCD unit (red). The rewired data (gray) refers to the tandem looping directions randomly assigned (either left or right). It showed that the observed tandem loops within a given CCD have significantly higher directional consistency than random chance. P value is calculated by Wilcoxon test. See also Figure S3J.

(G) An example CCD of two simple-convergent (not cross other loop anchor, red), five tandem (blue), and three complex-convergent (cross other loop anchor, purple) loops from eight anchors (a–h). The motifs in the five tandem loops are all in the rightward direction. Numbers in brackets depict the contact frequency.

(H) Proposed models, hairpin for convergent and coiled for tandem loops.

(I) Simulated 3D model of chromatin looping structure using the contact frequency and genomic span in G based on the folding principles proposed in (H). This model is an average representation of data derived from millions of cells. The simulation is detailed in Supplemental Experimental Procedures.

See also Figures S2 and S3, Tables S2 and S3, and Data S1.



(legend on next page)

tandem loops may have more promoter property, whereas the tail anchor could possess more enhancer potential. To test this, H3K4me1 (an enhancer marker) and H3K4me3 (a promoter marker) ChIP-seq data were used to assess the relative strength of promoter and enhancer. The ratio of H3K4me1/H3K4me3 at the tail anchor of tandem loops was significantly higher than the head anchor (Figure 3E, right), indicating that the tail anchor is more likely involved in enhancer function, whereas the head anchor is more related to promoter. In contrast, no difference was observed for the paired anchors of convergent loops (Figure 3E, left). Collectively, these data suggest that tandem loops possess distinctive features important for organizing transcription activity.

In GM12878 cells, thousands of genes and enhancers were found proximal to CTCF/cohesin anchors (i.e., anchor-genes/enhancers), and the rest were scattered within the CTCF/cohesin loop regions (i.e., loop-genes/enhancers) (Supplemental Experimental Procedures). We examined the anchor- and loop-genes based on their expression profiles in 56 different human tissues. Gene expression breadth analysis (Supplemental Experimental Procedures) indicated that anchor-genes were significantly less tissue-specific than loop-genes (Figure 3F, left). Further analysis revealed that active anchor-genes were almost exclusively housekeeping (Figures 3F, right, S4E, and S4F), emphasizing the notion that CTCF interaction anchors selectively enrich for constitutively expressed genes.

To investigate how active anchor-genes relate to active loop genes and enhancers, we examined their connectivity by RNAPII ChIA-PET, as described previously (Li et al., 2012). Using the newly generated RNAPII ChIA-PET data (Table S1), most active loop-genes were found connected to anchor-genes and/or anchor-enhancers (Table S2; Supplemental Experimental Procedures) through RNAPII-mediated interactions. Therefore, most active genes are connected, either directly or indirectly, to the anchors of CTCF loops (Figure 3G), suggesting that anchor-genes and anchor-enhancers could serve as nucleation points to aggregate related loop-genes toward corresponding CTCF/cohesin anchors for coordinated transcription. Figure 3H illustrates an example CCD with seven interconnecting CTCF anchors and a number of CTCF anchor-genes/enhancers and loop-genes/enhancers. RNAPII interactions indicated that these genes and enhancers were interconnected, which could be viewed (spatially) as a transcription factory docked to the chromatin structural base of CTCF focus. More examples are shown in Figures S4G and S4H.

Together, CTCF and RNAPII ChIA-PET analyses, along with chromatin state and RNA readout data, revealed that the basic topological units of chromatin looping structures and transcriptional function are highly cooperative for housekeeping functions and cell-type specificity.

Haplotype Variants Alter CTCF-Mediated Chromatin Structure and Function

Allelic differences between two homologous chromosomes can influence inheritance characteristics in the human genome (McDaniell et al., 2010). The means by which allele-specific genetic variation affects chromatin organization has become an intriguing question (Leung et al., 2015). Hi-C analyses have demonstrated that nuclear proximity ligation is an efficient genome-wide approach for haplotype mapping of chromatin interactions (Selvaraj et al., 2013). Here, we sought to use ChIA-PET to investigate haplotype-specific chromatin interactions and subsequent structural and functional consequences.

We identified 65,718 phased PET reads mapped intra-chromosomally in the phased GM12878 genome, of which the vast majority (n = 64,805, 98.6%) were *cis*-interacting PETs (Figures 4A, S5A and S5B; Supplemental Experimental Procedures). To investigate haplotype-specific chromatin interactions mediated by CTCF and RNAPII, we focused on phased PETs that were clustered to represent interactions enriched by these factors. Using this criterion, we identified 350 haplotype-biased anchors and 1,728 interactions mediated by CTCF, and 299 haplotype-biased anchors and 1,322 interactions mediated by RNAPII (Figures 4B and 4C; Supplemental Experimental Procedures).

Figure 3. Relationship between CTCF/Cohesin-Mediated Chromatin Structure and RNAPII-Associated Transcriptional Function

(A) Browser view of a 5 Mb genomic segment with four CCDs showing overlapped CTCF, RNAPII ChIA-PET data along with chromatin state (ChromHMM) and RNA-seq data. In the ChromHMM track, red for active promoter, yellow for enhancer and green for transcribed region. See Supplemental Experimental Procedures for the detailed color code.

(B) Aggregation density plots showing histone modification (top), RNAPII binding, and TSS (bottom) distribution profiles around the CTCF anchors and the loop regions. x axis, CTCF-anchors were taken from the anchor center with ±10% extension proportional to the enclosed loop regions. y axis, intensity (FPKM).

(C) Similar to (B), but only around the anchor centers (±2 kb). Anchors of convergent and tandem loops are analyzed separately.

(D) Similar to (C), but ChIP-Seq data of selected TFs are plotted. Upper: ELF1 and ZEB1. Lower: CTCF, RAD21, and ZNF143.

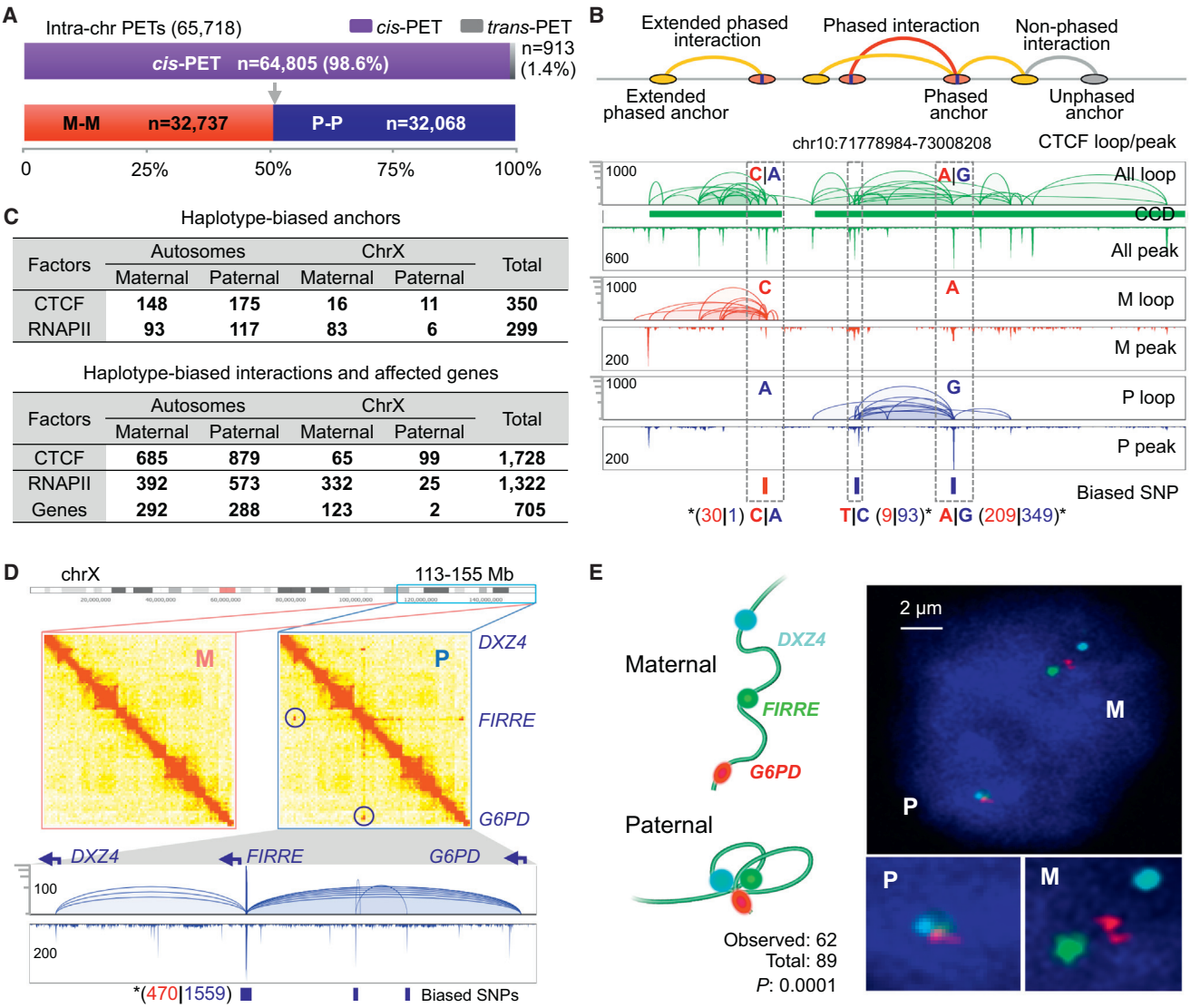
(E) Boxplots for the ratio of H3K3me1/H3K4me3 ChIP-seq at the anchors of convergent and tandem loops. High ratio suggests enhancer potential, low value indicates promoter function.

(F) Expression breadth (number of tissues a gene is expressed in) of CTCF anchor-genes and loop-genes in GM12878 (left). Anchor-genes (yellow) are significantly less represented as tissue-specific than loop-genes (green) ($p < 2.2e-16$, nonparametric Kolmogorov-Smirnov test). Anchor-genes are further divided as active (red) and inactive (yellow) for analysis (right). The expression breadth of all genes (gray) is included as reference.

(G) Proposed chromatin model. Top: a schematic CCD with anchor-gene/enhancer and loop-gene/enhancer associated with RNAPII interactions. g, gene; e, enhancer. Bottom left: CTCF-mediated loop model shows relative anchor and loop positions. Dotted arrow lines indicate the connectivity brought by RNAPII. Bottom right: RNAPII-participated model shows that RNAPII draws loop-genes/enhancers toward the CTCF anchors, docking the RNAPII foci onto the CTCF foci.

(H) Browser view of a CCD with complex sub-domain structures. It involves a numbers of anchor-genes/enhancers and loop-gene/enhancers, which are also connected by RNAPII-mediated loops. Orange and purple vertical bars highlight the promoters of anchor-gene and loop-gene, respectively. Right: a simulated 3D model for this topological domain mediated by CTCF/cohesin and the embedded transcriptional complex. This model is an average representation of data obtained from millions of cells.

See also Figure S4, Table S2, and Supplemental Experimental Procedures.



Factors	Autosomes		ChrX		Total
	Maternal	Paternal	Maternal	Paternal	
CTCF	**148**	**175**	**16**	**11**	**350**
RNAPII	**93**	**117**	**83**	**6**	**299**

Factors	Autosomes		ChrX		Total
	Maternal	Paternal	Maternal	Paternal	
CTCF	**685**	**879**	**65**	**99**	**1,728**
RNAPII	**392**	**573**	**332**	**25**	**1,322**
Genes	**292**	**288**	**123**	**2**	**705**

Figure 4. Haplotype Mapping of Chromatin Interaction

(A) Statistics of phased PETs in GM12878. Intra-chromosomal PETs were distinguished as *cis*-PETs and *trans*-PETs. A *cis*-PET has the two tags mapped to phased SNPs with the same haplotype (M-M or P-P); a *trans*-PET has the two tags mapped to phased SNPs in the opposite haplotypes (M-P or P-M).

(B) Identification of haplotype chromatin interactions. Top: schematic of the haplotype phasing of ChIA-PET mapping. Phased SNPs with CTCF or RNAPII binding were first identified. Interaction anchors overlapping with phased SNPs are referred as "Phased anchors" (vertical bar indicates the phased SNP). Interaction loops originating from paired phased anchors are "Phased interactions" (red). Interactions with only one side originating from phased anchors are "Extended phased interactions" (yellow). All other interactions that cannot be reliably determined are "Unphased interactions" (gray). Bottom: an example CCD, where three phased SNPs are identified with significant haplotype bias in CTCF-binding. The SNP nucleotides are color coded for their haplotypes (red, maternal; blue, paternal). Allele-specific binding frequencies in PET counts are given in parentheses with corresponding color code. *p < < 0.05, binomial test.

(C) Statistics of haplotype-biased anchors and interactions. Allele-specific genes associated with haplotype-biased RNAPII loops are also shown.

(D) Haplotype-specific super-long interactions mediated by CTCF connecting three loci: *DXZ4*, *FIRRE*, and *G6PD* in ChrX. Upper: contact heatmaps of the maternal (M) and paternal (P) homologs showing the contacts of the three loci only identified in paternal. Lower: loop/peak view of interactions mediated by CTCF in paternal ChrX. Phased allele frequencies of the SNPs at the *FIRRE* locus are shown in aggregate. The simulated 3D models for ChrX and the *DXZ4-FIRRE-G6PD* segment are presented in Figures S7F and S7G.

(E) DNA-FISH validation of the *DXZ4-FIRRE-G6PD* interactions. Left: expected conformations and probe design. Right: microscopic image in a nucleus with two clusters of the three testing probes. The numbers of total examined nuclei and nuclei with the expected probe pattern are shown. P value calculated by binomial test.

See also Figures S5 and S7.

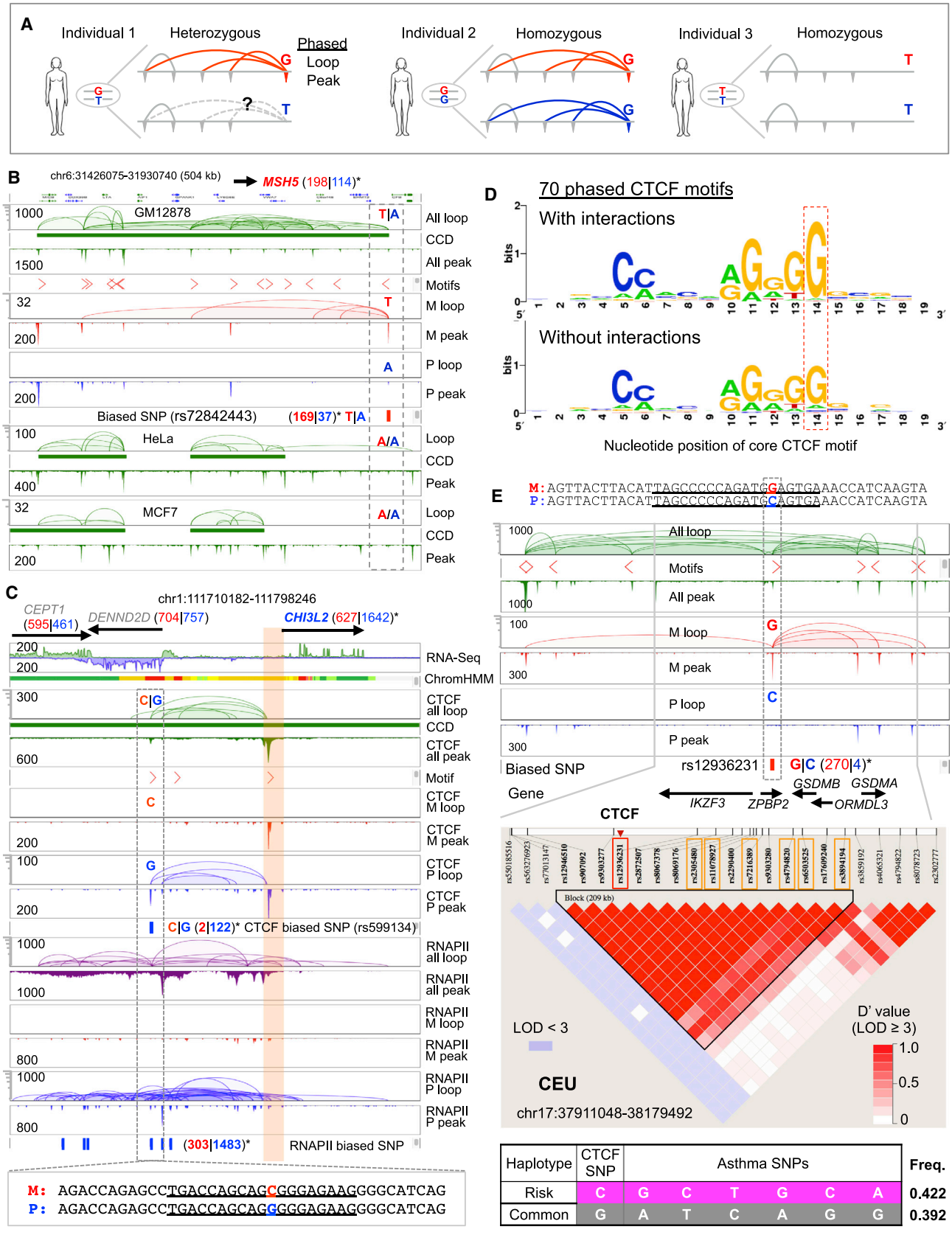
Among the identified haplotype-biased chromatin interactions mediated by CTCF was the well-studied *H19-IGF2* locus (Nativio et al., 2011) involved in genomic imprinting of autosomes (Figure S5C, left, and additional example in Figure S5C, right) thus demonstrating the robustness of our approach. We then explored whether allelic variations alter chromatin 3D structure between homologous chromosomes. Indeed, we found paternally biased super-long interactions (13 Mb) mediated by CTCF on ChrX connecting the *DXZ4* and *FIRRE* loci (Figure 4D), which has been reported previously (Horakova et al., 2012; Rao et al., 2014). In addition, we identified another super long-range interaction between *FIRRE* and *G6PD* (23 Mb) exclusively on the same haplotype as indicated by both contact heatmaps and CTCF-enriched interactions. This interaction was further validated by DNA-FISH (Figure 4E).

Next, we investigated whether SNPs could directly alter chromatin topology and function at a finer scale (e.g., domain structure and individual loop). Others have demonstrated deletion and inversion of DNA fragments containing CTCF/cohesin binding sites could disrupt the nearby TAD structure and alter associated gene transcription (Dowen et al., 2014; Guo et al., 2015). Despite the success, CRISPR/Cas9-engineered regions in these studies involved from 100s to 1,000s bp, which could contain other sequence elements of unknown function. To avoid potential "collateral damage," we exploited the SNP "genotype" as single nucleotide "perturbation" and evaluated the corresponding CTCF binding/looping property as the "phenotype" in different human individuals (cell lines) with either heterozygous or homozygous allele composition (Figure 5A). We first focused on the 39 allele-biased CTCF interaction anchors (i.e., 39 loci with phased SNPs as "naturally occurred" single nucleotide "perturbation") located at CCD boundary regions. As shown in Figure 5B, a heterozygous SNP (maternal "T" and paternal "A") was located on the 3′ boundary of a CCD. The maternal "T" exhibited strong CTCF binding (i.e., functional allele) while the paternal "A" allele showed weak binding (i.e., dysfunctional allele). We hypothesize that the loss of CTCF binding at the "A" allele in this locus would cause loss of CTCF-mediated looping and, in turn, alter CCD structure. To test this, we examined other individuals (cell lines) with homozygous "A/A" genotype at this locus. Indeed, in HeLa and MCF7 cells, no CTCF-mediated interactions originated from this locus and the corresponding CCD structures in these two cell lines were drastically different from GM12878 (see another case in Figure S6A). Together, these analyses validated that functional CTCF sites are necessary to maintain the proper CCD boundaries and suggest that single nucleotide variations in CTCF binding sites could alter chromatin topology.

We then explored if "naturally occurred SNP perturbation" could alter CTCF tandem loops and consequently impact transcription inside CCDs. We identified 50 loci where allele-specific tandem loops were associated with genes harboring phased SNPs in the gene body, thus, testable for allelic expression bias. From them, 22 (44%) displayed significant allele-specific expression ($p < 0.05$; see examples in Figure S6B). In a particular example, the promoters of *DENND2D* and *CHI3L2* reside at the opposite anchors of a CTCF tandem loop that is paternal-specific (Figure 5C). Consistently, the RNAPII occupancy and associated chromatin loops in this region are also paternal-biased, indicating this CTCF tandem loop is functionally involved in a paternal-specific transcriptional regulation. However, only *CHI3L2* exhibited paternal-biased gene expression (see also Figure S6C) but not *DENND2D*. This example supports that, in an allele-specific manner, the enhancer at the tail anchor of a CTCF tandem loop could interact with gene promoters proximal to the head anchor of the loop in concordance with the motif orientation for transcription regulation (Figures 3C–3E). Additionally, the phased SNP with allele-biased CTCF-binding coverage was located inside the CTCF-binding motif, and this SNP was the only nucleotide difference between the two homologs in kilobase-wide span. This observation impelled us to systematically examine how SNPs in the CTCF motif would subsequently influence CTCF binding and looping.

Of the SNPs mapped within the 350 allelic-specific anchors bound by CTCF (Figure 4C), 70 reside in the core motif (Figure S6D). The alignment of each of the heterozygous SNPs within the CTCF motif showed that alleles having strong CTCF-binding possess canonical motif consensus, whereas alleles with weak or no binding had deviated motif sequences, especially with position 14 ("G") being the most affected (Figures 5D and S6E), suggesting that this "G" is critical for CTCF-binding affinity. For example, at position 14 of a CTCF motif in Chr17q21, the maternal and paternal alleles are G|C, respectively (Figure 5E). Despite this SNP being the only variable site in this locus and the surrounding kilobase region, the CTCF binding and interaction in maternal allele is 68-fold stronger than paternal, indicating that nucleotide "C" in the motif is disruptive for CTCF to bind to the paternal allele in this locus. Since SNP variation in CTCF motif could lead to such drastic alteration in CTCF binding and looping, we sought to determine whether changes in chromatin topology would link to human diseases.

We systematically assessed disease association of the disrupted CTCF-mediated interactions by examining the linkage disequilibrium (LD) between the 70 SNPs residing in CTCF motif (i.e., CTCF-SNPs) and GWAS identified disease-associated SNPs (Supplemental Experimental Procedures). We hypothesized that disease-associated SNPs and dysfunctional CTCF-SNPs would be linked if they reside in strong LD blocks. In this setting, 32 of the 70 CTCF-SNPs were documented in dbSNP database, and eight showed LD with disease-associated SNPs in the tested populations (Data S1, IV; Table S4). Since GM12878 originated from an individual of European ancestry, we further focused on the CTCF-SNPs with disease association by LD in CEU population. One of the CTCF-SNPs found to be associated with asthma is SNP rs12936231 (Figure 5E). Of the two alleles (G|C), the dysfunctional "C" has been documented as a high-risk allele for asthma and autoimmune diseases and suggested to alter chromatin remodeling and domain-wide transcription of certain genes (e.g., *ZPBP2*, *GSDMB*, and *ORMDL3*) (Verlaan et al., 2009). Separate expression quantitative trait loci (eQTL) data also suggested that this allele alters the expression level of *GSDMB* and *ORMDL3* (Montgomery et al., 2010; Stranger et al., 2007). Interestingly, six additional asthma-associated SNPs were found in the same LD block (209 kb) with this CTCF-SNP (rs12936231) in CEU having high pairwise correlation ($D' > 0.7$), and the haplotype comprising these seven risk alleles was frequently (0.422) observed in the CEU population



Haplotype	CTCF SNP	Asthma SNPs						Freq.
Risk	C	G	C	T	G	C	A	0.422
Common	G	A	T	C	A	G	G	0.392

(legend on next page)

(Figure 5E). Collectively, these results suggest that the disruption of CTCF motif by this "C" allele, which abrogates CTCF binding, looping, and chromatin topology, may be the primary molecular event leading to disease susceptibility, while the other six asthma-SNPs were likely non-causative "bystanders." Moreover, these findings highlight the potential of allelic chromatin topology analyses to infer mechanisms by which SNPs are associated with disease and traits.

RNAPII-Mediated Chromatin Interactions Regulate Allele-Specific Transcription

To investigate whether allelic variations affect RNAPII-mediated chromatin interaction and/or result in functional consequences, we identified 1,322 haplotype-specific RNAPII interactions that involved 299 haplotype-specific RNAPII interaction anchors in GM12878 (Figure 4C; Supplemental Experimental Procedures). Our haplotype analyses showed significant ChrX maternal-specificity of RNAPII interactions and gene expression (Figure 4C), consistent with the fact that paternal-X in GM12878 is imprinted (Rozowsky et al., 2011). Thus, at chromosomal scale, the connection between allele-specific RNAPII interaction and gene transcription is broadly established. To further validate that haplotype-specific chromatin interaction could lead to allele-specific transcription regulation (Figure 6A), we examined 40 TFs for their allelic binding specificity and found that the vast majority (95%) of the TF binding allele-biases were consistent with the RNAPII allele-specificity (Figure 6B; Supplemental Experimental Procedures). Furthermore, we identified 89 genes with allele-biased expression involved in allele-specific RNAPII interactions, and the majority (n = 79) displayed the same allele-specificity in transcription as the allele-biased RNAPII interactions (Figures 6C and 6D; Table S5). For example, RNAPII binding and interactions over the SNPs at the *XIST* (X inactive specific transcript) locus showed highly consistent allele bias with haplotype-specific expression of *XIST* (Figure S5D). Such observations at the *XIST* locus demonstrate the accuracy of our haplotype chromatin interaction analysis.

Our data also reveals the haplotype effect of long-range enhancer-promoter interactions. For example, the promoters of *LOC374443*, *CLEC2D*, and *CLECL1* were in contact with an RNAPII-associated multi-gene complex (Figure 6E). It is observed RNAPII occupancy in the paternal allele at the upstream enhancer and promoter sites (300 kb apart) were ~3-fold and ~2.5-fold higher, respectively, than the matched maternal allele. In addition, the distal enhancer exhibited >3-fold higher paternal-biased binding by three B cell-specific TFs: BCL3, EBF1, and TCF12. We also observed 3.5-fold higher paternally biased expression in these three genes (Figure 6E). In contrast, nearby genes not directly involved in the RNAPII-mediated interactions showed balanced expression between homologs. This example supports the notion that allele-specificity at distant enhancers is also effective in regulating allele-specific RNAPII activity at the target genes with high specificity.

3D Genome Models Elucidate the Human Genome Structure and Function

Using an integrated 3D NucleOme Modeling Engine (3D-NOME) that builds a hierarchical tree structure to represent the 3D genome with increasing resolution (Figures S7A–S7C) (unpublished data), we simulated the 3D genome models from the combined CTCF and RNAPII ChIA-PET data derived from GM12878. In these models, several known chromosomal features were captured, e.g., at the whole nucleome level, large chromosomes were preferentially projected in the periphery of nucleus, and small chromosomes were positioned in the inner nuclear space (Figures S7D and S7E; Supplemental Experimental Procedures). To gain further specificity at the level of individual chromosomes, we modeled Chr1 at different resolutions and observed a

Figure 5. SNPs Altering Allelic CTCF Chromatin Interaction and the Functional Implication

(A) Schematic of using SNP as single nucleotide "perturbation" for validation of CTCF-mediated chromatin interactions. In individual 1, phased SNP and allele-specific CTCF binding are used to determine the functional and dysfunctional alleles for CTCF interaction. However, it is not immediate ready to extrapolate "no binding = no looping." In individual 2 and 3, homozygous alleles at the corresponding SNP location, possessing either the functional or the dysfunctional CTCF interaction allele, were analyzed for the presence or absence of CTCF binding and looping, respectively, thus, validating the function of CTCF in mediating chromatin interaction.

(B) An example using data from GM12878, HeLa and MCF7 shows CCD structures perturbed by SNP. A phased SNP (maternal "T," paternal "A") is identified at the right boundary of a CCD in GM12878. Differential strength of CTCF binding (169:37) was detected and the CTCF loops were extrapolated based on the biased binding. At this SNP locus, both HeLa and MCF7 were of homozygous "A/A" (dysfunctional CTCF allele), had no CTCF binding, and no chromatin contact originated from. *$p < < 0.05$; binomial test. See also Figure S6A.

(C) An example in GM12878 illustrating CTCF tandem loop with allele-specificity and consequent impact on allele-biased transcription. A phased SNP (rs599134) is located in the CTCF motif (dashed box highlighted) of a "tail" anchor of a tandem loop with the "head" anchor and CTCF motif (highlighted in orange) proximal to the promoter of *CHI3L2*. The CTCF binding and looping in this region are paternal-specific, and the RNAPII binding and interactions are significantly paternal-biased as indicated by multiple heterozygous SNPs in this region. The expression of *CHI3L2* also exhibited significant paternal-bias. In contrast, the genes (*CEPT1* and *DENND2D*) immediately upstream of the tandem loop showed balanced expression. Nucleotide sequences of the highlighted CTCF-binding site are shown at the bottom with the motif underlined. *$p < < 0.05$, binomial test.

(D) Logos from 70 CTCF motifs with allelic SNP disruption on CTCF interaction. Haplotype motifs with strong CTCF bindings had canonical consensus (top), motifs with weak CTCF binging displayed deviated consensus (down), especially at position 14. Examples of SNPs in CTCF motif disrupting CTCF binding and looping patterns are shown in Data S1, III.

(E) CTCF motif disrupted by SNP is linked to disease susceptibility. Top: an example of allele-specific disruption on CTCF interaction by having SNP within a CTCF motif. SNP (rs12936231) resides at motif position 14 of a CTCF interaction site. Middle: linkage disequilibrium between this CTCF-SNP (rs12936231, in red box) and the other six asthma-associated SNPs (in orange boxes) in the CEU population. These seven SNPs are identified in a significant LD block (D' value > 0.5 and LOD $\geq$ 3) as highlighted in black triangle. Bottom: haplotypes of these seven SNPs associated with asthma in the CEU population. The dysfunctional "C" allele of the CTCF-SNP (rs12936231) is frequently (0.422) associated with the risk alleles of the other 6 SNPs in CEU.

See also Figure S6, Table S4, and Data S1.

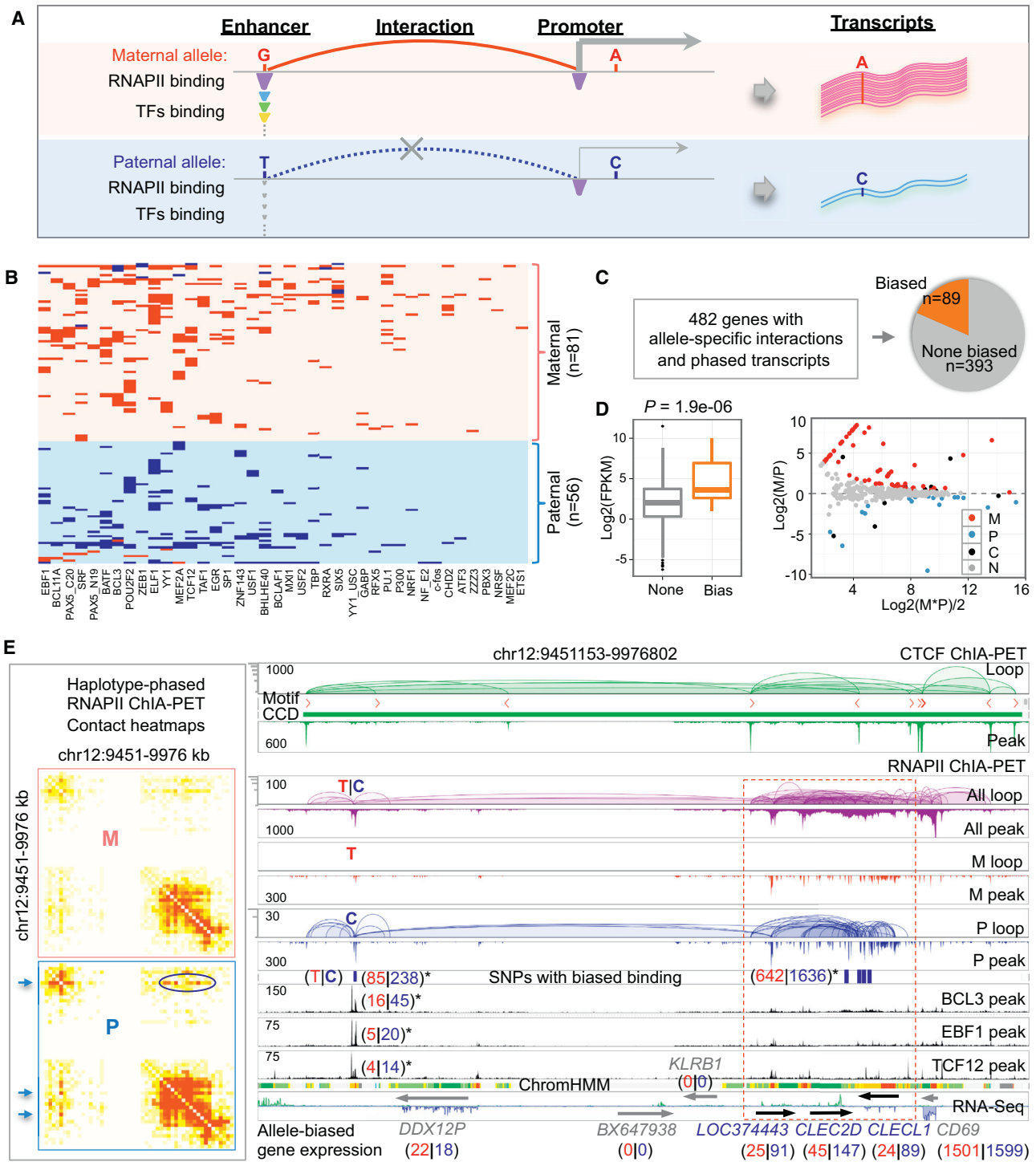


Figure 6. Allele-Biased Chromatin Interactions Mediated by RNAPII

(A) Schematic of using SNPs to investigate allelic-effects of transcription regulation via haplotype-biased occupancy and interaction mediated by RNAPII and TFs.

(B) Profile of allele-biased binding by 40 TFs at the allele-biased anchors (maternal 81, paternal 56) of RNAPII interactions. Each row represents an allele-biased RNAPII anchor with allele-biased binding by at least one TF. Each column represents one of the 40 tested TF. Each colored tile indicates TF binding bias: maternal, red; paternal, blue.

(C) Genes involved in allele-specific RNAPII-mediated interactions with phased transcripts. Eighty-nine genes showed significant allele-biased gene expression.

(legend continued on next page)

putative conformation, whereby its two chromosomal arms bend and extend in the same direction (Figure 7A). Since our mapping data were derived from millions of cells, the predicted 3D model would reflect an average representation. It is possible that the body of Chr1 is fluidly changing between "open" and "closed" conformations (Figure 7B). Our 3D DNA-FISH results supported such speculation (Figure 7C; Supplemental Experimental Procedures).

From 3D genome simulation and DNA-FISH nuclear imaging, as well as the growing knowledge of chromatin topology (Boyle et al., 2011; Kalhor et al., 2012), a general feature of chromosome topology starts to emerge. Although much less condensed, chromosomes in interphase still maintain their core axes, comprised of mostly condensed heterochromatin segments (probably other matrix-filling material as well) (Figure 7D). The loosely organized "open" chromatin segments could extend laterally outward (as chromatin loops) around the chromosome axis, similar to the morphology of lampbrush chromosome (Morgan, 2002). We envision CTCF and cohesin (possibly with other factors e.g., topoisomerase II) (Liang et al., 2015) localizing on the surface of the core chromosome axis and defining the interface between the inner condensed (inside the chromosome axis core) and the outer loose domains. The condensed chromosome axis core could help to maintain the desired shape and physical properties for the overall chromosome territories, which are also impermeable as repressive or inactive compartments. On the contrary, the loose domains are open and permissive to molecules mediating nuclear functions.

A key component in our model is the spatial overlap of RNAPII-associated transcription factories with the CTCF/cohesin contributed structural foci. To test this, we conducted super-resolution structured illumination microscopic (SIM) analysis of immunostaining using antibodies against CTCF and RNAPII in GM12878 cells (Figure 7E; Supplemental Experimental Procedures). To further increase the detection resolution, we performed Förster resonance energy transfer (FRET) assay using fluorescence lifetime imaging microscopy (FLIM) to detect the co-localization of CTCF and RNAPII (Figure 7F; Supplemental Experimental Procedures). Together, the SIM and FRET-FLIM analyses validated that the CTCF and RNAPII foci are indeed in close distance in the human nucleome.

Last, to test whether CTCF prevalently locates along the chromosomal axes, we exploited lampbrush chromosomes. Although not ideal, lampbrush chromosome is still considered the classic model for chromatin looping morphology. We examined the lampbrush chromosomes isolated from newt nuclei using immunostaining to CTCF, and showed the CTCF signals predominantly along chromosome axes instead of the laterally extended chromatin loops (Figure 7G; Supplemental Experimental Procedures).

DISCUSSION

The inclusiveness of ChIA-PET for enriched specific chromatin interactions and non-enriched higher-order chromatin contacts and the high degree of data correlation between ChIA-PET and Hi-C demonstrated here are extremely encouraging to the field of 3D genome biology. In addition, the high accuracy of haplotype-resolved chromatin contact mapping by ChIA-PET and Hi-C, in agreement with the concept of chromosome territory established by DNA-FISH, further collectively validates the primary principles of our strategies in characterizing genuine events of chromatin interaction and 3D genome topology. The immediate technical challenges are to further improve the efficiency, accuracy, specificity, and throughput of our technologies for 3D genome mapping in mixed and individual cells.

TAD is an important concept in chromatin biology established recently. However, the detailed structures and associated functions are still to be uncovered. Our analytic strategy focusing on CTCF and RNAPII represents a highly practical, efficient, and comprehensive solution to provide mechanistic insights on sub-TAD structures and the embedded functions for transcription regulation. The high degree alignment of CCDs identified by CTCF ChIA-PET to TADs identified by Hi-C further support the idea that CTCF, along with cohesin, is a major contributor that shape chromatin topology in the human nucleome. As indicated in this study, most of the convergent CTCF loops (that probably require more energy and are more stable once established) are engaged to define the CCD/TAD structures and boundaries, whereas the tandem loops (that possibly require less energy and are more dynamic) are involved inside CCD/TAD for transcription and regulatory functions. The discovery of orientational alignment between CTCF motifs and large portion of genes proximal to chromatin interaction anchors elucidates an interesting directional framework of chromatin topology for coordinated transcription regulation between the interplays of CTCF foci and RNAPII machinery. Although we provided initial nuclear imaging evidence to support the proximity of CTCF and RNAPII foci within the nuclear space, additional validations are expected in the near future.

Given the structural importance of CTCF to chromatin configuration, we anticipate that strong CTCF binding sites would be

(D) Left: boxplot of the expression levels of genes with (red box) and without (gray box) allele-bias. Genes with allele-bias (n = 89) are of significantly higher expression than the none-biased (n = 393) (p = 1.9e-06). Right: MA-plot of the allele-biased gene expression levels. x axis measures expression abundance, y axis indicates differential expression between the two haplotypes. M, maternal-biased, red, n = 61; P, paternal-biased, blue, n = 18; C, contradictively biased with the corresponding haplotype-biased RNAPII interaction, black, n = 10; N, no bias, gray, n = 393. See also Table S5.

(E) An example shows allele-biased RNAPII binding/looping and the regulatory effect on the associated genes. Left: haplotype contact heatmaps (M, maternal; P, paternal) of the genomic segment indicated paternal haplotype-specific long-range chromatin interactions (blue arrows). Right: loop/peak browser view. On the left side, an enhancer was identified. This enhancer overlaps with a phased SNP (maternal "T," paternal "C") and connects downstream to an RNAPII-mediated interaction complex involving 3 genes (*LOC374443*, *CLEC2D*, *CLECL1*). There are 11 phased SNPs in the multi-gene complex. Both of the enhancer and the gene complex exhibited paternal-biased RNAPII binding and interactions. The expression of the three genes is also paternal-biased. In addition, the enhancer also showed paternal-biased binding by 3 B cell-specific TFs BCL3, EBF1, and TCF12. All allele-specific sequence reads coverage by RNAPII and TF binding and transcripts are shown in aggregate numbers in the parentheses. *$p < < 0.05$, binomial test.
See also Table S5.

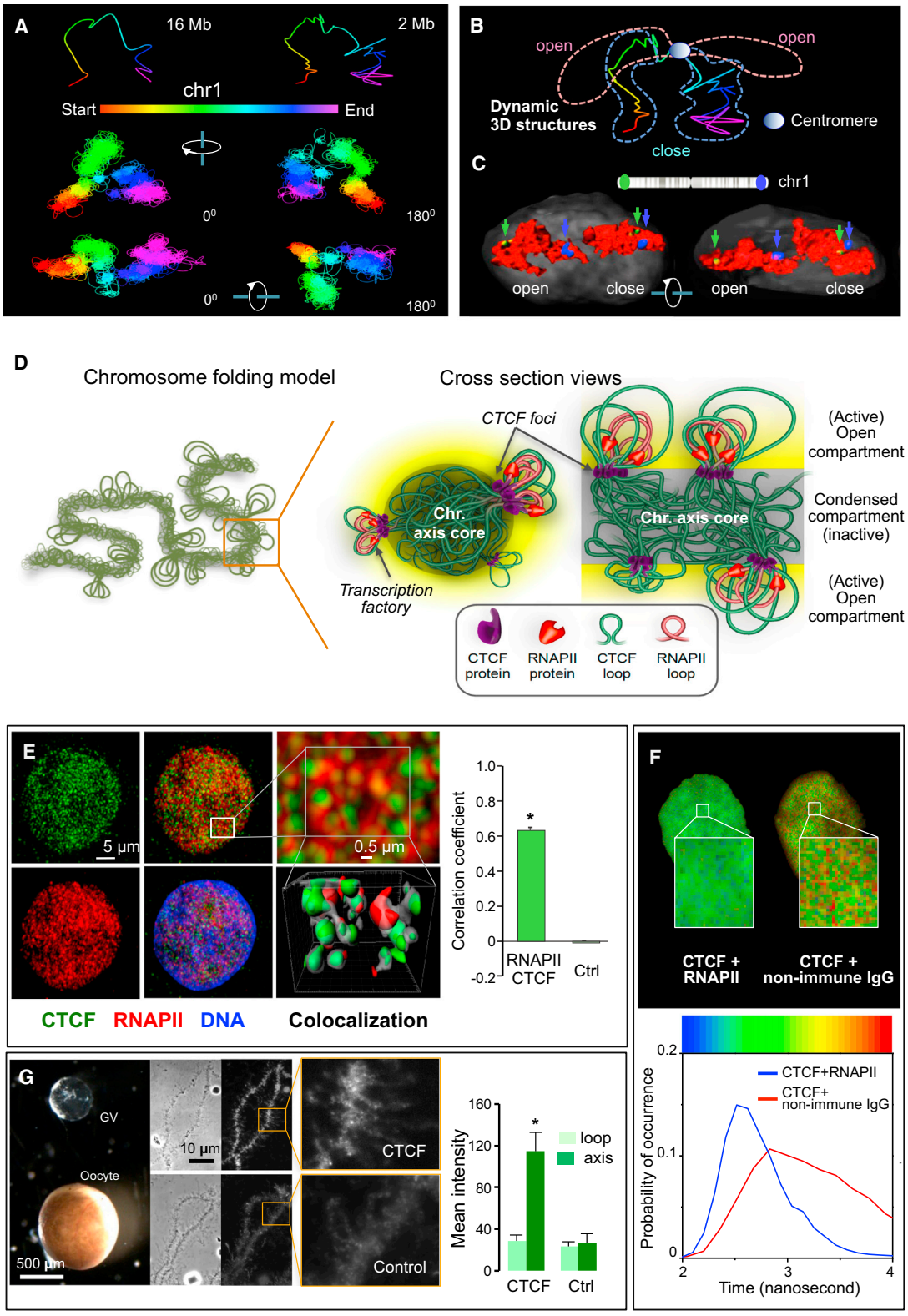


Figure 7. Chromatin Model of CTCF Foci and RNAPII Transcription Factories

(A) 3D models of Chr1 at 3 resolutions (16 Mb, 2 Mb, 100 bp) with views from different angles. The color bar indicates the proportional genomic coordinates of 3D models.

(legend continued on next page)

the candidate mutational targets, where single nucleotide changes may have dramatic consequences. Conveniently, our strategy using SNP-based haplotype mapping of chromatin interactions and allelic-biased chromatin structure features enabled us to use "naturally occurred point mutations" as effective means for "naturally occurred single nucleotide perturbation" in human individuals to study genetic effects on chromatin structures and consequent gene expression possibly linked to disease susceptibility. Together, several lines of evidence provided in this study suggest that genetic variations that affect CTCF/cohesin-mediated chromatin topology could lead to changes in gene expression, thus, laying the molecular foundation for genome topology change, disease susceptibility and evolutionary adaptation. With further detailed comprehension of the chromatin-organizing role played by CTCF, RNAPII, and other protein factors, a clearer view of 3D genome topology and associated nuclear function will soon emerge.

EXPERIMENTAL PROCEDURES

More details are available in Supplemental Experimental Procedures.

Long Read ChIA-PET

Instead of using MmeI digestion as in the original ChIA-PET protocol (Fullwood et al., 2009), long read ChIA-PET uses Tn5 tagmentation to generate random size of PET templates for long tag sequencing reads (2 × 150 bp) by HiSeq2500 (Supplemental Experimental Procedures).

ChIA-PET Data Processing

Short-read ChIA-PET data was previously generated by us and deposited in ENCODE data repository (Table S5) and processed using the original ChIA-PET Tool (Li et al., 2010). Long-read ChIA-PET data was processed by a customized ChIA-PET data processing pipeline.

ChIP-Nexus

ChIP-nexus on RAD21 and SMC3 in GM12878 cells was performed as described in He et al. (2015).

3D DNA-FISH

DNA-FISH of GM12878 cells was performed according to Cremer et al. (2008) with minor modifications. The 3D images of all-chromosome painting were acquired with the Zeiss LSM 780 confocal microscope. The 3D Chr1 territory was obtained using Imaris (Bitplane) and Amira (FEI).

Immunostaining and SIM Super-Resolution Microscopy

The CTCF and RNAPII immunofluorescence staining in GM12878 was performed according to routine procedures (Hall et al., 2015). Specimens were analyzed using a Zeiss ELYRA PS.1 super-resolution system. The super-resolved images were generated using Zeiss Zen 2012 black edition software.

Co-localization Detection of CTCF and RNAPII by FRET-FLIM

The FRET-FLIM analysis of the same specimen was performed following the same immunocytochemistry protocol described above. The measurement of fluorescence lifetime of the donor was performed on a Picoquant PicoHarp 300 time-correlated single photon counting (TCSPC) system attached to Leica Sp8 confocal microscope, using 63× oil immersion objective (NA 1.4).

Lampbrush Chromosome

Ovarian biopsies were performed on adult female newts. Germinal vesicles (nuclei) from stage V-VI oocytes were manually isolated. Lampbrush chromosomes were prepared as previously described (Penrad-Mobayed et al., 2010) and then immunostained with CTCF antibody subject for standard light transmitted and florescence microscopy. The fluorescence signals were measured on the chromosome axes and lateral loops.

ACCESSION NUMBERS

The accession number for the data reported in this paper is GEO: GSE72816.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, five tables, and two data files and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.11.024.

AUTHOR CONTRIBUTIONS

G.L. and Y.R. conceptualized the study. X.L. improved ChIA-PET. M.Z. adopted ChIP-nexus. X.L., M.Z., P.W., D.W., and X.R. generated data. Z.T. and

(B) An ensemble model of Chr1 folding dynamics in GM12878.
(C) 3D images of DNA-FISH for the two copies of Chr1 (red) in a nucleus from different angles. The positional patterns of the two probes (green and blue) indicate two chromosomal conformations, "open" and "close."
(D) An overall model of chromosomal folding involving CTCF and RNAPII. Left: chromosome in interphase is loosely organized with chromatin loops extended from the condensed chromosome axis core that maintains the overall conformation of chromosome territory. Right: zoom-in transverse and longitudinal cross section views. CTCF locate on the surface of chromosome axis core, defining the interphase of the inner condensed (inactive) and the outer open (active) compartment for transcription.
(E) Super-resolution SIM microscopic images of CTCF and RNAPII immunostains. Left: CTCF (green) and RNAPII (red) foci in GM12878 nucleus. Middle: merged images from CTCF and RNAPII without (middle top) and with (middle bottom) DNA stain (Hoechst 33342, blue). Top right: zoom-in merged image. Bottom right: 3D reconstruction of co-localization with the depth of the scanned volume. Bar chart: statistics of Spearman's correlation values between CTCF and RNAPII signals from 21 cell nuclei. Control (Ctrl) is from random sampling of 100 nm-sized CTCF and RNAPII immunostained images. Data are shown as mean with SEM. *$p < 0.001$.
(F) FLIM of GM12878 nuclei subjected to FRET. Nuclei were stained immunofluorescently. Left: CTCF + RNAPII co-immunostained. Right: CTCF + non-immune IgG as negative control. Alexa488-labeled CTCF served as a donor for FRET, while Cy3 labeled RNAPII as an acceptor. Color-coded pixels correspond to values of mean fluorescent lifetimes as indicated by color bar below. Bottom: distribution curve of fluorescence lifetime in the experiments, CTCF + RNAPII (blue) and CTCF + non-immune IgG (red). The occurrence of FRET between the donor and acceptor (co-localization of CTCF/RNAPII with the inter-molecular distances between the fluorophores $\leq$10 nm) is revealed by the shortening of the lifetime (ns) of the donor fluorescence.
(G) CTCF immunostain of lampbrush chromosome. Left: light microscopy of oocyte, the germinal vesicle (GV, nucleus), and lampbrush chromosomes isolated from *Pleurodeles waltl*, with zoom-in confocal microscopy of lampbrush chromosomes stained by CTCF and IgG antibodies. The confocal microscopic images show that the CTCF signals are mostly concentrated along chromosome axis, but the control IgG signal are scattered evenly. Right: bar chart of immunostaining measurements on chromosome axis and laterally extended chromatin loops. The CTCF signals are significantly higher on chromosome axis than on chromatin loops (*$p < 0.001$). Data are presented as mean with SEM.
See also Figure S7, Data S2, and Supplemental Experimental Procedures.

O.J.L. performed data analysis and interpretation. M.P. and L.M.S. analyzed lampbrush chromosome. J.J.Z., P.T., A.M., J.W., B.R., and G.M.W. performed microscopic analyses. P.S., P.M., E.P., and D.P. developed software for 3D simulation and visualization. Z.T., O.J.L., and Y.R. wrote the manuscript with input from X.L., E.T.L., C.W., and G.L.

ACKNOWLEDGMENTS

Y.R. is supported by the Director Innovation Fund of The Jackson Laboratory, NCI R01 CA186714, NHGRI R25HG007631, NIDDK U54DK107967 (4DN), and the Roux family as the Florine Roux Endowed Chair in Genomics and Computational Biology. X.L. is supported in part by China "111 project" (B07041). Polish National Science Centre supports G.M.W. [UMO-2012/05/E/NZ4/02997]; D.P. and P.S. [2014/15/B/ST6/05082; UMO-2013/09/B/NZ2/00121]; and J.W. [DEC-2012/06/M/NZ3/00163]. D.P. and P.S. are also supported by National Leading Research Centre in Bialystok and the European Union under the European Social Fund. The authors thank C.Z. Zhang for initial DNA-FISH, Agnieszka Walczak and Katarzyna Krawczyk for FISH discussion, Rafael Casellas, Michael Stitzel, and Duygu Ucar for manuscript discussion, and Gosia Popiel for help on preparing Figure S7. Some of the genome sequences described in this research were derived from a HeLa cell line. Henrietta Lacks, and the HeLa cell line that was established from her tumor cells without her knowledge or consent in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. The request to use HeLa data for this research was approved by the NIH Director based on the recommendations of the Advisory Committee to the Director and the evaluation by its HeLa Genome Data Access Working Group (http://acd.od.nih.gov/hlgda.htm). The HeLa genomic datasets used for analysis described in this manuscript were obtained from the database of Genotypes and Phenotypes (dbGaP) through dbGaP: phs000640.

Received: May 30, 2015
Revised: September 12, 2015
Accepted: November 10, 2015
Published: December 10, 2015

REFERENCES

Bickmore, W.A. (2013). The spatial organization of the human genome. Annu. Rev. Genomics Hum. Genet. *14*, 67–84.

Boyle, S., Rodesch, M.J., Halvensleben, H.A., Jeddeloh, J.A., and Bickmore, W.A. (2011). Fluorescence *in situ* hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosome Res. *19*, 901–909.

Cremer, M., Grasser, F., Lanctôt, C., Müller, S., Neusser, M., Zinner, R., Solovei, I., and Cremer, T. (2008). Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes. Methods Mol. Biol. *463*, 205–239.

Cullen, K.E., Kladde, M.P., and Seyfred, M.A. (1993). Interaction between transcription regulatory regions of prolactin chromatin. Science *261*, 203–206.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., and Young, R.A. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell *162*, 900–910.

Hall, M.H., Magalska, A., Malinowska, M., Ruszczycki, B., Czaban, I., Patel, S., Ambrożek-Latecka, M., Zołocińska, E., Broszkiewicz, H., Parobczak, K., et al. (2015). Localization and regulation of PML bodies in the adult mouse brain. Brain Struct. Funct. http://dx.doi.org/10.1007/s00429-015-1053-4.

He, Q., Johnston, J., and Zeitlinger, J. (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat. Biotechnol. *33*, 395–401.

Horakova, A.H., Moseley, S.C., McLaughlin, C.R., Tremblay, D.C., and Chadwick, B.P. (2012). The macrosatellite DXZ4 mediates CTCF-dependent long-range intrachromosomal interactions on the human inactive X chromosome. Hum. Mol. Genet. *21*, 4367–4377.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90–98.

Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A.Y., Yen, C.A., Lin, S., Lin, Y., Qiu, Y., et al. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature *518*, 350–354.

Li, G., Fullwood, M.J., Xu, H., Mulawadi, F.H., Velkov, S., Vega, V., Ariyaratne, P.N., Mohamed, Y.B., Ooi, H.S., Tennakoon, C., et al. (2010). ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol. *11*, R22.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84–98.

Liang, Z., Zickler, D., Prentiss, M., Chang, F.S., Witz, G., Maeshima, K., and Kleckner, N. (2015). Chromosomes progress to metaphase in multiple discrete steps via global compaction/expansion cycles. Cell *161*, 1124–1137.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

McDaniell, R., Lee, B.K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., et al. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. Science *328*, 235–239.

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.

Morgan, G.T. (2002). Lampbrush chromosomes and associated bodies: new insights into principles of nuclear structure and function. Chromosome Res. *10*, 177–200.

Nativio, R., Sparago, A., Ito, Y., Weksberg, R., Riccio, A., and Murrell, A. (2011). Disruption of genomic neighbourhood at the imprinted IGF2-H19 locus in Beckwith-Wiedemann syndrome and Silver-Russell syndrome. Hum. Mol. Genet. *20*, 1363–1374.

Ong, C.T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.

Penrad-Mobayed, M., Kanhoush, R., and Perrin, C. (2010). Tips and tricks for preparing lampbrush chromosome spreads from Xenopus tropicalis oocytes. Methods *51*, 37–44.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Rhee, H.S., and Pugh, B.F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell *147*, 1408–1419.

Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol. *7*, 522.

Selvaraj, S., R Dixon, J., Bansal, V., and Ren, B. (2013). Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat. Biotechnol. *31*, 1111–1118.

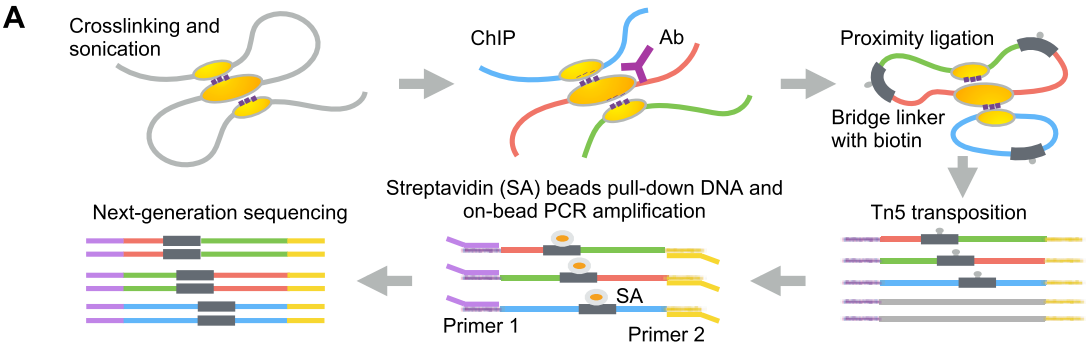
Sims, R.J., 3rd, Mandal, S.S., and Reinberg, D. (2004). Recent highlights of RNA-polymerase-II-mediated transcription. Curr. Opin. Cell Biol. *16*, 263–271.

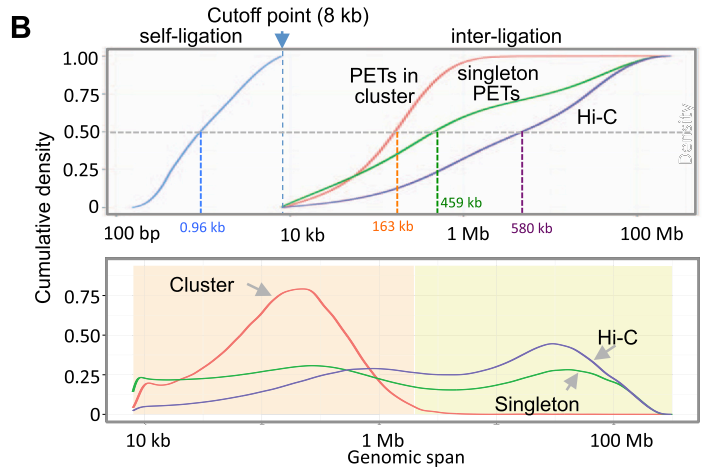
Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat. Genet. *39*, 1217–1224.

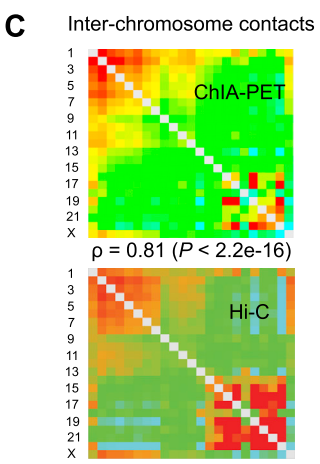
Verlaan, D.J., Berlivet, S., Hunninghake, G.M., Madore, A.M., Larivière, M., Moussette, S., Grundberg, E., Kwan, T., Ouimet, M., Ge, B., et al. (2009). Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. Am. J. Hum. Genet. *85*, 377–393.

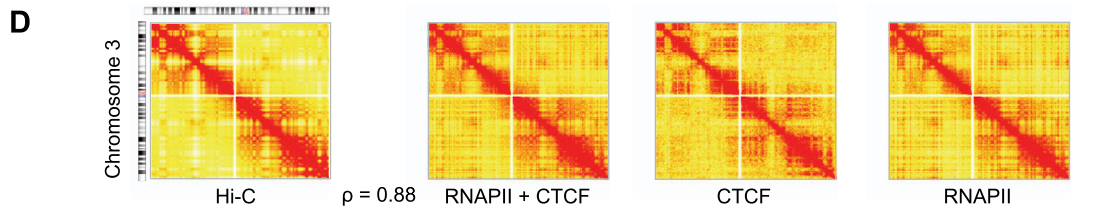
Zhang, Y., Wong, C.H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y., Lim, J., Tai, E., Poh, H.M., Wong, E., et al. (2013). Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature *504*, 306–310.

Cell



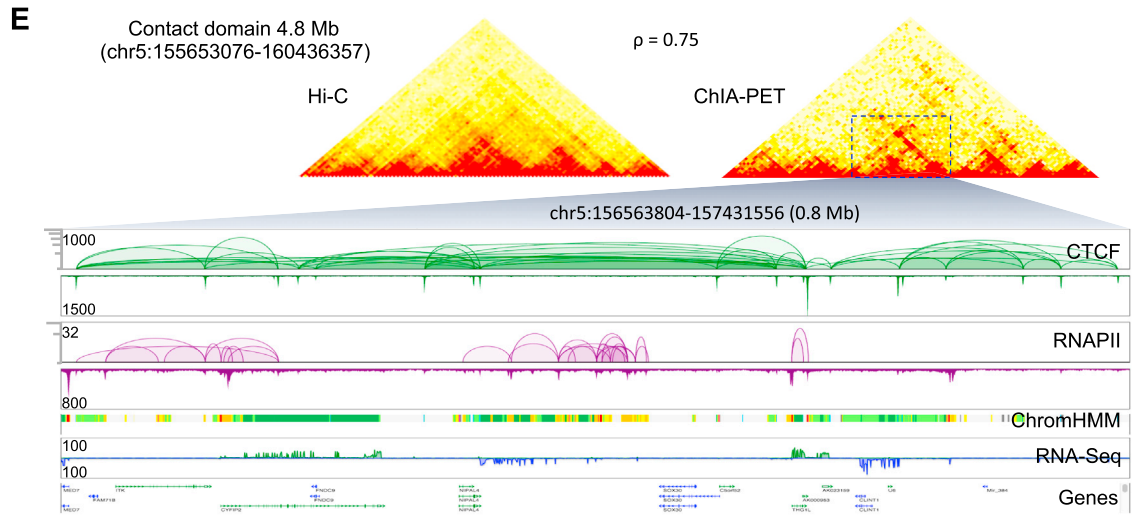






Spearman's correlation coefficient between GM12878 Hi-C and ChIA-PET per chromosome

Chr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X
Cor.	0.83	0.87	0.88	0.90	0.86	0.81	0.87	0.88	0.84	0.87	0.77	0.83	0.91	0.75	0.82	0.76	0.80	0.89	0.68	0.81	0.82	0.78	0.86



(legend on next page)

Figure S1. Characterization of ChIA-PET Data, Related to Figure 1

(A) Schematic workflow of long-read ChIA-PET. After crosslinking, sonication and antibody pull-down, the ChIP DNA fragments in chromatin complex are proximity ligated by a bridge linker with biotin. Then the chromatin complexes are de-crosslinked and Tn5 transposase is used to cut the DNA fragments and add adaptor simultaneously. Streptavidin beads are used to pull-down the ChIP enriched DNA fragments, and then the fragments are PCR amplified and are subject to paired-end sequencing.

(B) Genomic span distribution comparison between ChIA-PET and Hi-C data in GM12878 cells. ChIA-PET data used for analysis here consists of combined CTCF and RNAPII datasets. Upper: Cumulative density plots of the genomic span of ChIA-PET self-ligation, inter-ligation and Hi-C paired end reads. ChIA-PET inter-ligation PETs are segregated into clustered and singleton PETs. The genomic span size at the point of 50% for each data category is shown. PET singletons and Hi-C mapping reads generally have similar distribution patterns covering large size range, and are larger than PETs in clusters. Lower: Density plots of ChIA-PET and Hi-C data over genomic span. The enriched ChIA-PET clusters are mostly in sub-megabase range having peak size around 200 kb, reflecting protein factor specific interactions, whereas the singleton inter-ligation reads from ChIA-PET resembles Hi-C data in terms of genomic span. Hi-C is more enriched with data having larger genomic span.

(C) Comparison between normalized ChIA-PET and Hi-C contact frequency heatmaps of GM12878 data at inter-chromosome level. Spearman's correlation coefficient is shown.

(D) Comparison of intra-chromosomal contact frequency using heatmaps between Hi-C and ChIA-PET data of Chr3 in GM12878. The general heatmap patterns are very similar across different datasets. Spearman's correlation coefficients of contract frequency between Hi-C and combined CTCF and RNAPII ChIA-PET data for each test chromosome in GM12878 are included. See Supplemental Experimental Procedures.

(E) Comparison of intra-chromosomal contact frequency using heatmaps between Hi-C and ChIA-PET data for a large segment in Chr5. Detailed ChIA-PET detected chromatin interaction loops and binding peaks mediated by CTCF and RNAPII in a large contact domain (0.8Mb) is also shown. Chromatin state (ChromHMM), strand-specific RNA-Seq (green: positive strand; blue: negative strand) and gene tracks are also shown. For each data track, the numbers at the left side indicate the maximum interaction frequency (log10 scale) and the highest tag count of binding peaks or RNA-Seq readout (linear scale). For the ChromHMM track, red is for active promoter, yellow is for enhancer and green is for transcribed region. See Supplemental Experimental Procedures for detailed color code.

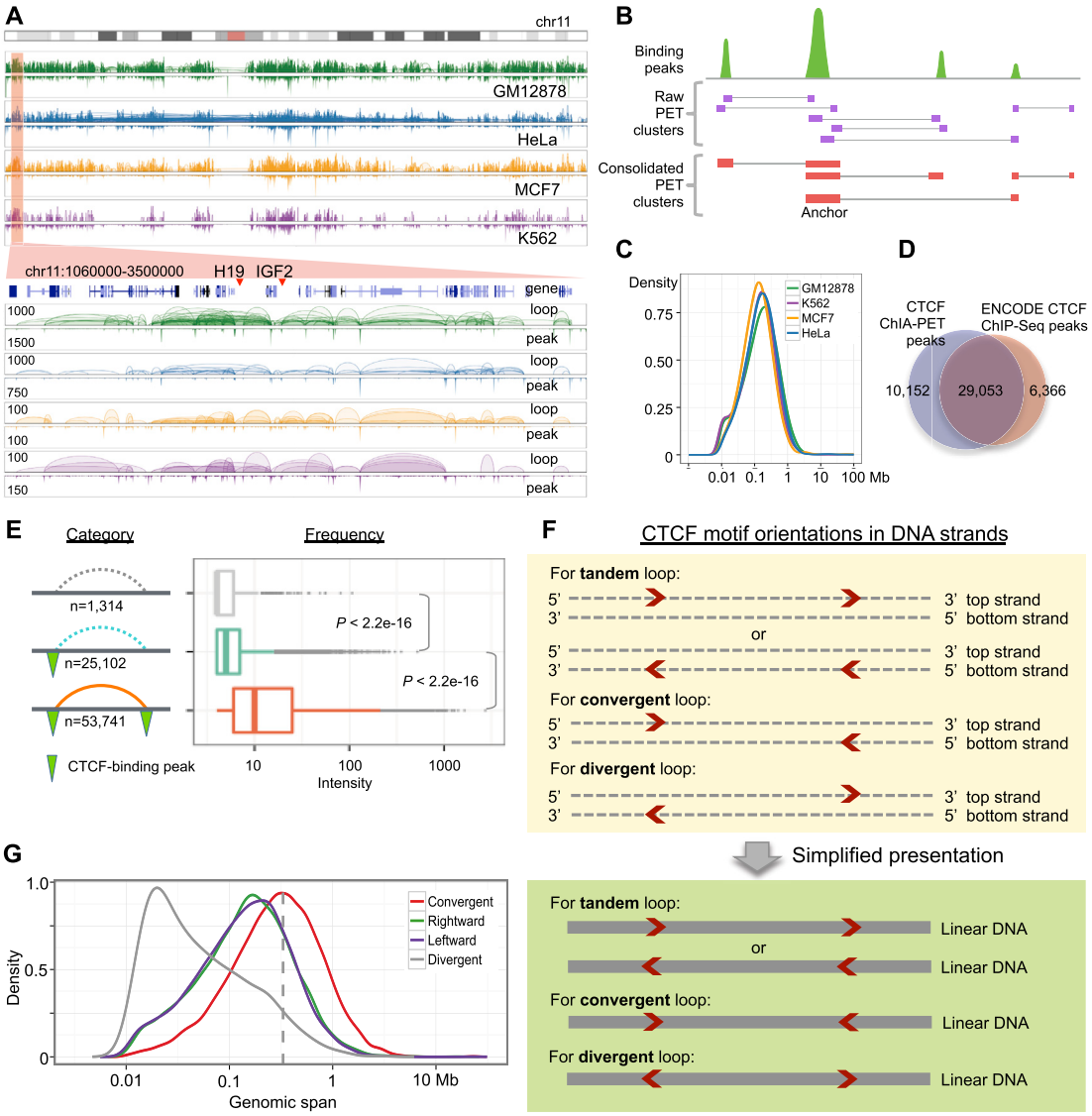


Figure S2. Characterization of CTCF-Mediated Chromatin Interactions, Related to Figure 2

(A) A chromosome-wide view and zoom-in section of CTCF-mediated chromatin interaction loops and binding peaks in GM12878, HeLa, MCF7 and K562 cells. CTCF interaction loops are ubiquitously identified across the entire chromosome in all 4 cell-types. The zoom-in view shows that CTCF-defined chromatin interactions are well conserved in these cell lines.

(B) Schematic illustration of PET clustering. There are two steps to generate merged PET clusters. The individual interaction PET reads (2x150 bp) were clustered based on overlapping with 500 bp extensions. The raw PET clusters were further collapsed to generate the merged PET clusters based on if tag clusters are overlapped with the same protein factor binding sites. The binding sites were identified by merging the raw PET cluster anchors (see extended experimental procedures). Such process streamlines the PET cluster data structure. The bars in purple denote the anchors of raw PET clusters. The bars in red denote the final merged anchors.

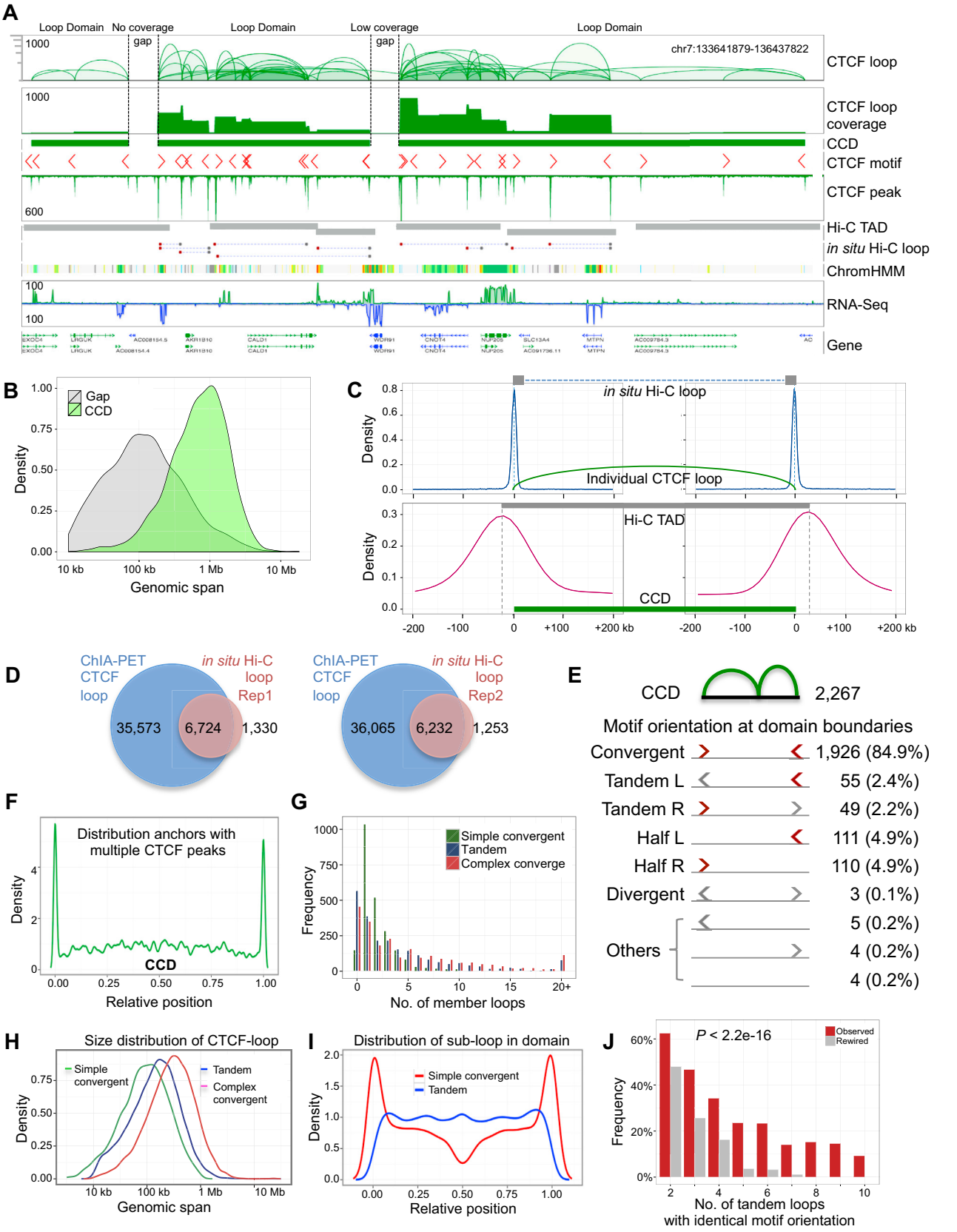
(C) Genomic span distribution of CTCF interaction loops in GM12878, HeLa, K562 and MCF7. The loop span size distributions across different cell types are almost identical. The peak loop size is in 100-200 kb range.

(D) Comparison between CTCF-binding peaks identified by ChIA-PET and ENCODE CTCF ChIP-Seq data in GM12878. Venn diagram shows 74% of the ChIA-PET identified peaks are in directly overlap with the ENCODE CTCF peaks.

(E) Dissection of CTCF interaction loops according to relationship with ENCODE identified CTCF-binding peaks in GM12878. Of the 80,157 CTCF loops, 53,741 have the anchors directly overlapping with CTCF-binding peaks, and such loops are of significantly higher interaction intensity measured by PET counts.

(F) Schematic of CTCF-binding motif orientation at the CTCF-mediated interaction achors.

(G) Genomic span distribution of CTCF loops with convergent, tandem (leftward and rightward) and divergent motif patterns. The peak size of the convergent loops at approximately 250 kb is indicated by the gray vertical dashed line.



(legend on next page)

Figure S3. Definition and Characterization of CTCF/Cohesin-Mediated Chromatin Contact Domain, Related to Figure 2

(A) An example locus from GM12878 illustrating the connectivity between individual CTCF loops, contact frequency, loop coverage and the derived CTCF-mediated contact domains. The connectivity of the CTCF loops provides mostly continuous coverage along the chromosome within a given segment, which defines the contact domains. The regions with very low or no loop coverage are considered as domain gaps (see extended experimental procedures). Hi-C identified TADs (IMR90) and in situ Hi-C (GM12878) identified chromatin loops are included for comparison. CTCF motifs at the interaction anchors are shown as red arrows representing the motif strand orientation. Chromatin states and strand-specific RNA-Seq data (green: forward strand; blue: reverse strand) are also included. Numbers on the left side in each data track represent the maximum level of frequency for each corresponding track.

(B) Genomic span distributions of the CTCF contact domains and gap regions. The CTCF contact domains are of much coherent in genomic spans with peak size around 1 Mb. In comparison, the domain gaps are generally smaller with a broader size distribution range (from ~10 kb to 10 Mb).

(C) Comparison of CCDs with Hi-C TADs. Upper: The distribution of the distance from the individual CTCF loop boundaries to the closest boundaries of the matched in situ Hi-C loops. Left and right boundary distances are calculated separately. Lower: For overlapped TADs and CCDs, the differences in distance of boundaries were measured. Using the boundaries of CCDs as the center point, the distribution of the distance from the CCD left and right boundaries to the nearest left and right boundaries of Hi-C defined TADs (IMR90) are in wide range, respectively. In comparison, the TAD boundaries spread wide range when considering the CCD boundaries as reference, whereas the in situ Hi-C loop boundaries are much closer to the CTCF loop anchors.

(D) Venn diagrams showing the relationships between CTCF ChIA-PET loops and the in situ Hi-C detected loops in replicate 1 and 2, respectively. Vast majority of the in situ Hi-C detected chromatin loops (83.5% and 83.3% of replicate 1 and 2, respectively) are recapitulated by CTCF ChIA-PET loops (See Extended Experimental Procedures for details). Meanwhile, the CTCF ChIA-PET identified CTCF loops are more than 4-fold of the in situ Hi-C detected loops, which suggests that the CTCF loops identified by ChIA-PET represent a much large body of detailed and complex CTCF-mediated chromatin looping structures.

(E) Summary statistics and illustration of the pair-wise CTCF motif orientation at the boundaries of CCDs. Majority (84.9%) of the CCDs are of convergent CTCF motif pattern at the outmost boundary anchors.

(F) Distribution of the positions of the CTCF loop anchors having multiple CTCF-binding peaks/motifs relative to the genomic span of CCDs. Such type of interacting anchors (Figure 2C, insets) is strongly enriched at the boundaries of CCDs.

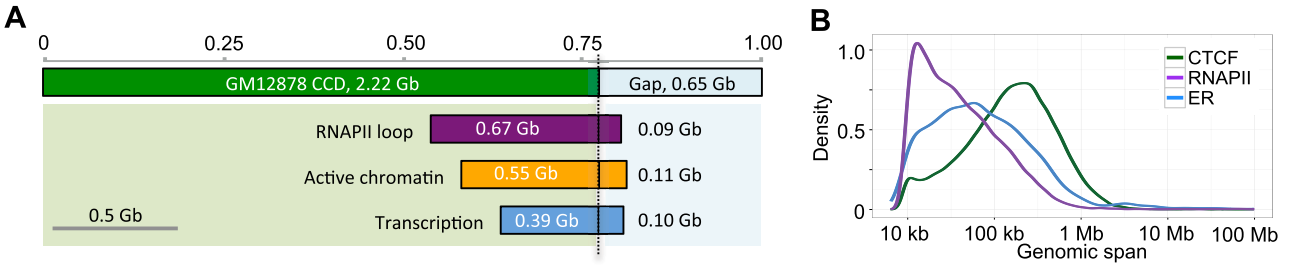
(G) Histogram of the numbers of simple-convergent, tandem and complex-convergent loops located in a CCD. Most of the CCDs contain 1 to 2 simple-convergent loops, and on average, 4.3 tandem (median: 3) and 6.4 (median: 5) complex-convergent loops.

(H) Distribution of the genomic spans of simple-convergent, tandem and complex-convergent CTCF loops. Generally, complex-convergent loops are of the largest genomic span; tandem loops are smaller than complex-convergent, but larger than the simple-convergent loops.

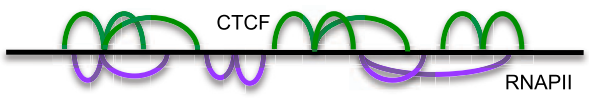
(I) Distribution of the relative positions of the simple-convergent and tandem loops in the CTCF contact domain defined spaces. The center points of any individual simple-convergent and tandem loops are used for calculation. Simple-convergent loops are enriched near the boundaries of the CCDs, whereas tandem loops are rather evenly distributed in the genome space defined by CCDs.

(J) Distribution of the frequency of CCDs with a certain number of tandem loops having identical CTCF motif orientation. For example, for CCDs containing 2 tandem loops, > 60% of them have all of these 2 member tandem loops with identical CTCF motif orientation. This is significantly higher than rewired data. For the rewired data, the motif orientations of the tandem loops within a contact domain are randomly assigned. The P-value is calculated by Wilcoxon test.

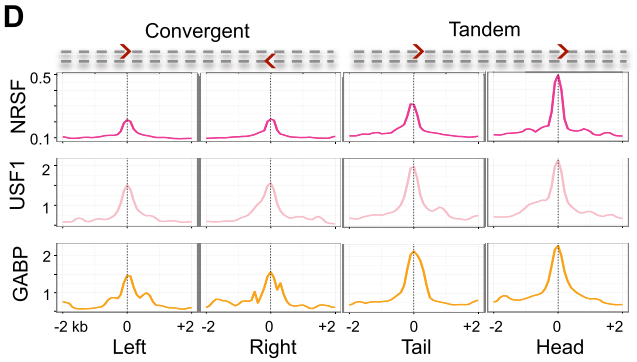
See also Supplemental Experimental Procedures.

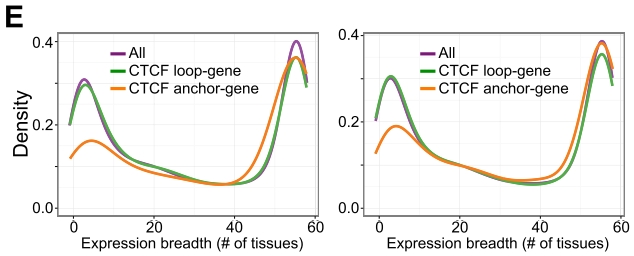


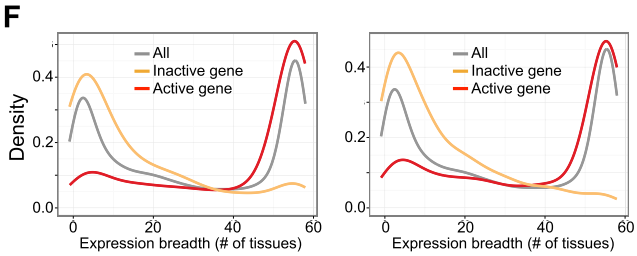
C

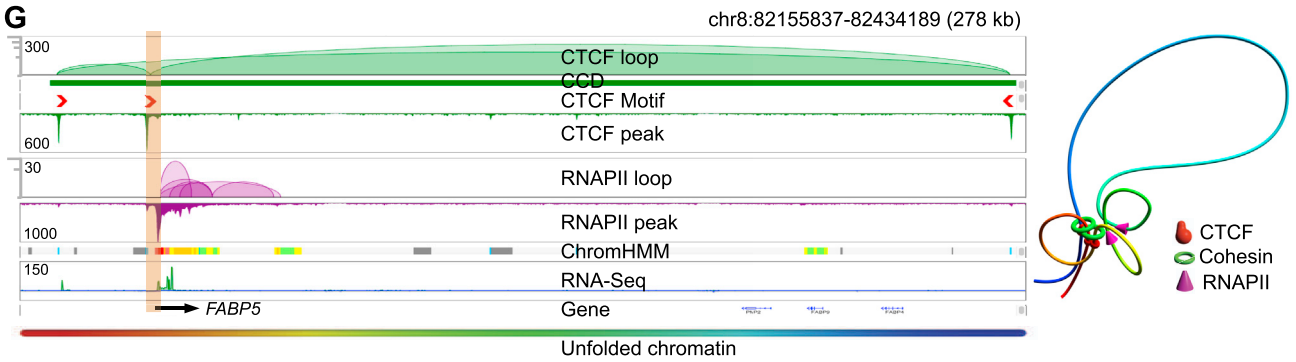


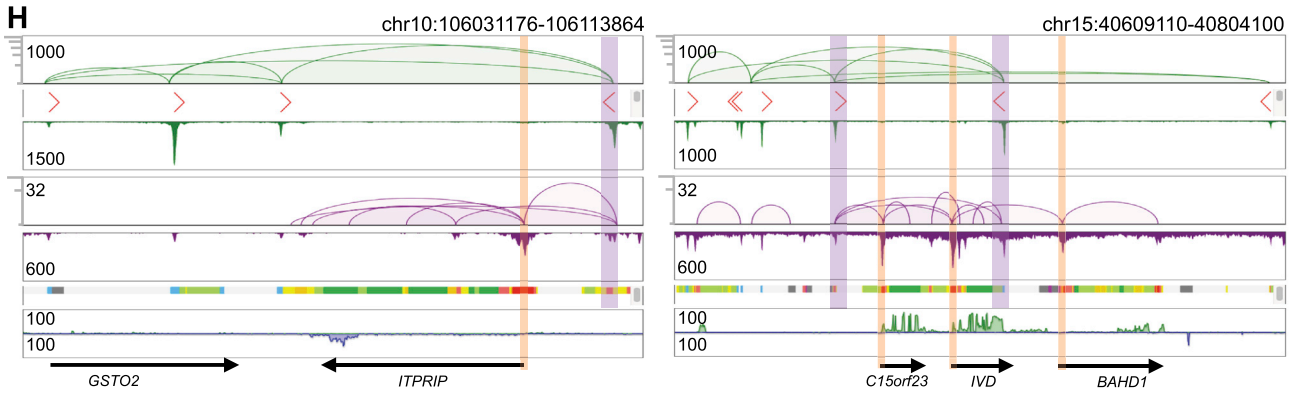
RNAPII interaction	In CCD	In CCD gap	Connect CCDs
Individual loop (50,045)	92.0%	6.9%	1.1%
SG complex (1,141)	86.6%	12.3%	1.1%
MG complex (1,592)	77.1%	7.0%	15.9%











(legend on next page)

Figure S4. Summary and Characterization of CTCF and RNAPII Interaction Structures, Related to Figure 3

(A) Bar chart of genomic coverage by CCDs. The majority of the genome (hg19 assembly gaps excluded) in GM12878 cells is covered by CTCF loops (2.22 Gb, 77%), whereas a minor portion is the gap region (0.65 Gb, 23%). Within the CTCF-charted genomic landscape, most of transcriptional activities represented by RNAPII-associated chromatin loops (0.67 Gb), active chromatin markers (0.55 Gb), and expressed transcripts (0.39 Gb) are found in the CCDs, whereas only small parts are in the gap regions, respectively.

(B) Genomic span distribution of CTCF, RNAPII and ER mediated loops. Generally, CTCF loops are of larger size, and RNAPII loops are much smaller (mostly less than 100 kb). ER loop (from MCF7 cells) sizes are more uniformly distributed from 10 kb to 1 Mb.

(C) Intersection relationship between CCDs and RNAPII mediated chromatin interaction loops in GM12878 cells. Individual RNAPII-mediated loops are often clustered into complex structures, some complex involved with only single gene (SG), others with multiple genes (MG). Most of the RNAPII looping structures are contained within the CCDs, and some take place in CTCF domain gaps. Very few of RNAPII mediated gene complexes go across the CTCF contact domain boundaries or connect multiple domains.

(D) Aggregation density plots showing the NRSF, USF1 and GABP binding signal distribution profiles around the centers (±2kb) of CTCF interaction anchors. CTCF loops with paired motifs in convergent and tandem orientations in anchor regions are analyzed separately. Loops with paired motifs in tandem leftward and rightward orientations are combined for analysis, and for data plotting, the x axis coordinates of tandem left data are flipped to fit the tandem right orientation shown.

(E) Expression breadth (number of tissues a gene is expressed in) of CTCF anchor-genes and CTCF loops-genes identified in K562 (left) and MCF7 (right). The expression breadth of all genes included in the expression tissue panel is shown as control. CTCF anchor-genes in both K562 and MCF7 are significantly less enriched with tissue-specific genes ($p < 2.2e-16$, nonparametric Kolmogorov-Smirnov test).

(F) Expression breadth of active and inactive CTCF anchor-genes in K562 (left) and MCF7 (right). The expression breadth of all genes included in the expression tissue panel is shown as control (see more details in Supplemental Experimental Procedures).

(G) An example illustrating a CCD containing a single active gene with its promoter in close proximity to CTCF anchor. Orange vertical bar highlights promoter located at CTCF anchor. Right: Proposed 3D chromatin model surrounding this gene. This model is an average representation derived from 3D mapping data obtained from millions of cells.

(H) Examples of CTCF anchor-gene/enhancer and CTCF loop-gene/enhancer in relationship to RNAPII associated chromatin interactions within CCDs. Left: In this CTCF contact domain in Chr10, the 4 CTCF interaction anchors (with motifs) are connecting together to form a CTCF-focus. This CTCF-defined structure includes an active gene that is located in the intervening loop region, with two anchor-enhancers and a few loop-enhancers as well. Right: A CTCF domain in Chr15 contains 5 CTCF anchors (with motifs) that would form a CTCF-focus. There are 3 actively expressed genes located within this CTCF structure. All 3 promoters of the genes are located in loop regions (i.e., loop-genes). However, 2 of the connecting enhancers are proximal to CTCF anchors, consequently linking the genes to the CTCF anchor regions. Structurally, the transcription factory involving these 3 genes is docked on top of this CTCF-focus. Based on these examples, our proposed model is that RNAPII associated interactions between the gene promoters and the enhancers would draw genes in to proximity of CTCF anchors, or CTCF-focus. The resulting RNAPII complex would form a transcription factory for the expression of genes. The vertical bars in orange indicate promoters, and bars in purple highlight enhancers that are in CTCF anchor regions.

See also Supplemental Experimental Procedures.

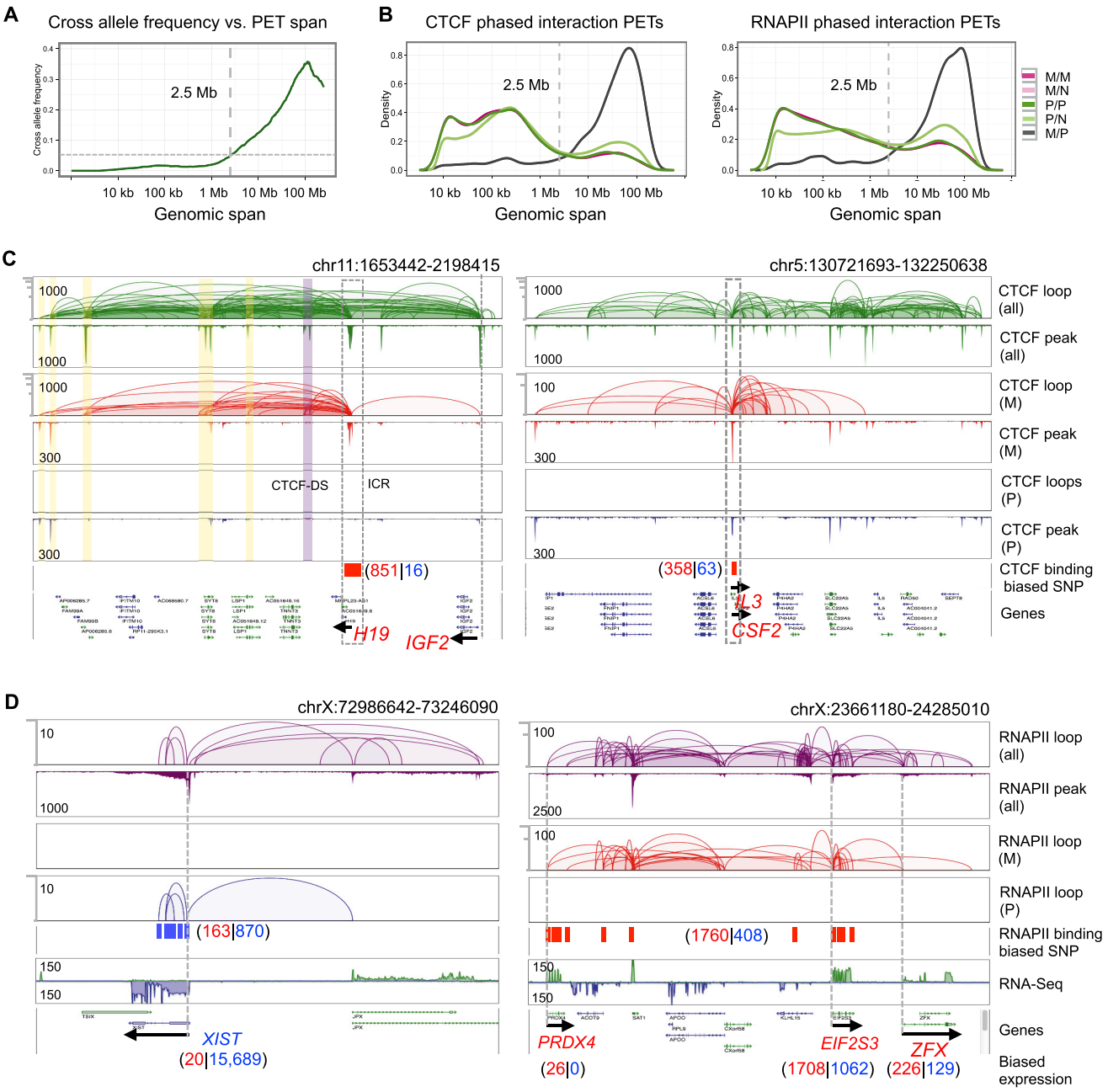


Figure S5. Haplotype Mapping of Chromatin Interaction by ChIA-PET, Related to Figures 4, 5, and 6

(A) Cumulative frequency of cross-haplotype interaction PETs as a function of genomic span. Intra-chromosomal ChIA-PET PETs with less than 2.5 Mb of genomic span are overwhelmingly *cis*-interacting.

(B) Genomic span distribution of phased CTCF (left) and RNAPII (right) ChIA-PET PETs. Cross-haplotype interaction PETs are of abnormally large genomic span than the other classes. Such observation is consistent with our prior knowledge that chimeric PETs in ChIA-PET data are of large genomic span (Li et al., 2012). M/M: maternal cis interaction; P/P: paternal cis-interaction; M/N and P/N: extended maternal and paternal cis interaction; M/P: cross-allele interaction.

(C) Examples illustrating haplotype biased CTCF interactions. Left: The classic *H19-IGF2* haplotype-specific CTCF interaction locus. CTCF-binding peak at the ICR (imprinting control region) upstream of *H19* is exclusively detected on the maternal allele. The maternal tag count (851) in the CTCF binding peak region (boxed by gray dotted line) is in red; the paternal tag count (16) in blue is likely reflecting basal level detection noise. Beside the previously reported maternal homolog exclusive ICR to CTCF-DS (Downstream; purple highlight) CTCF-mediated interaction, 5 additional distal loci (yellow highlight) are identified to interact with ICR exclusively on the maternal homolog. Right: The *IL3* and *CSF2* locus is exhibiting maternal biased CTCF-binding and interactions. Numbers on the left side in each data tack are for the maximum frequency of peaks and loops.

(legend continued on next page)

(D) Left: *XIST* responsible for ChrX inactivation is involved in paternal allele exclusive RNAPII interactions and shows haplotype specific expression from the paternal homolog. Right: *PRDX4*, *EIF2S3* and *ZFX* are involved in maternal allele specific RNAPII interactions and are of biased expression in the maternal homolog.

In (C) and (D), phased SNPs coverage frequencies by CTCF and RNAPII binding and RNA-Seq data from the maternal (red) and paternal (blue) haplotypes are shown in aggregate if multiple SNPs were involved.

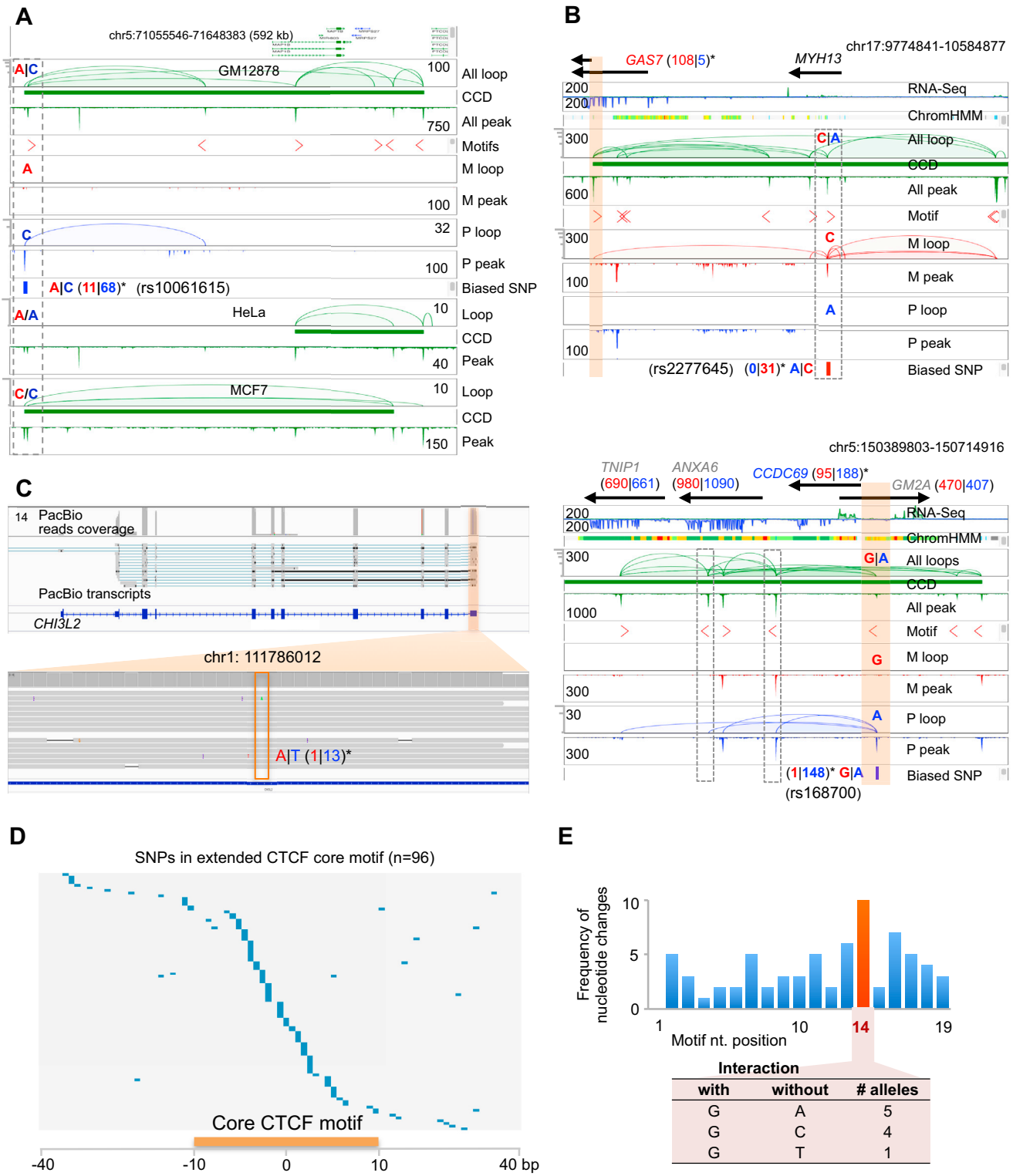


Figure S6. Allele-Specific CTCF Interaction and Disease Association Analysis, Related to Figure 5

(A) One example using data from GM12878, HeLa and MCF7 illustrating CCD structures perturbed by SNP variation. A phased SNP (maternal "A" and paternal "C") at the left boundary of a CCD in GM12878 exhibited differential CTCF interaction (PET counts: maternal "A" = 11, paternal "C" = 68). In this case, the paternal allele "C" is the functional one for CTCF interaction, while the maternal "A" allele is the dysfunctional one. In HeLa and MCF7, distinctive CTCF binding, looping and CCD structures were observed corresponding to the two heterozygous genotypes in GM12878. In HeLa, it has the homozygous dysfunctional allele "A/A" at

(legend continued on next page)

the SNP site. The CTCF binding is weak and no chromatin interactions initiated from this site. In contrast, MCF7 has homozygous functional genotype ("C/C") at this site, and exhibits similar CTCF binding and CCD structure as in GM12878. The genotype information for MCF7 cells were imputed from CTCF and histone modification mark ChIP-Seq data.

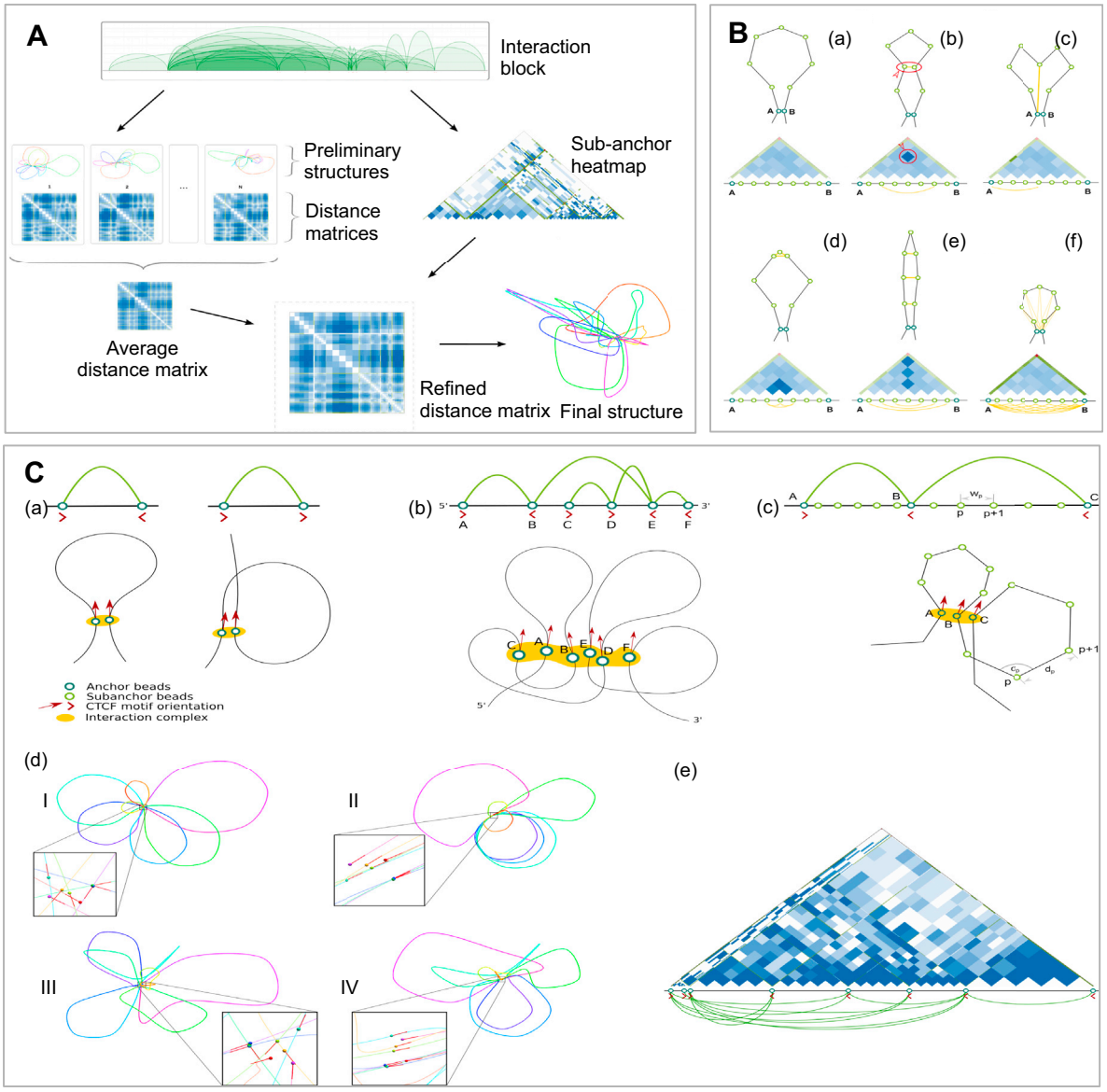
(B) Two examples in GM12878 illustrating allele-specific CTCF-mediated loops in tandem motif orientation and the consequent impact to transcription. Top: A phased SNP (rs2277645) located in a CTCF binding site (dashed box highlighted region) involved in maternal-allele specific CTCF binding and looping (PET count: maternal "C" = 31, paternal "A" = 0). A tandem CTCF loop connecting this and another site harboring gene *GAS7*, whose promoter (highlighted with orange box) is proximal to left anchor of the loop. Allele-specific RNA-Seq data also show that the expression of *GAS7* is almost exclusively from maternal allele (RNA-Seq read count: maternal = 108, paternal = 5). Down: A phased SNP (rs168700) locus exhibited haplotype-specific CTCF binding exclusively to the paternal allele (PET count: maternal "G" = 1, paternal "A" = 148), and is proximal to the *CCDC69* promoter (highlighted with orange box). This CTCF anchor is connected to two upstream anchors (highlighted in dashed boxes) forming two paternal-allele specific interactions with motifs in tandem orientation. The expression of *CCDC69* is significantly biased (2-fold) to the paternal allele (RNA-Seq read count: maternal = 95, paternal = 188), but not the other 3 genes nearby.

(C) Genome browser view of PacBio ISO-Seq readout of *CHI3L2* in GM12878. The red box highlighted base shows the base pair coverage of the phased SNP, which has a strong paternal allele bias.

(D) Heatmap showing the position of phased SNPs relative to the nearest CTCF-binding motif in GM12878. In total, 96 phased SNPs are located in the CTCF-binding core motif (20 bp, 70 SNPs) and the extended region (±40 bp, 26 SNPs). The positions of the SNPs are shown in blue.

(E) Frequency of nucleotide changes (heterozygous SNP) in positions of CTCF motif. Top: A bar chart showing the positional distribution of the 70 SNPs identified in CTCF motifs, nucleotide position 14 has the highest perturbation frequency (n = 10) by heterozygous SNP. Bottom: All 10 alleles with the conserved nucleotide "G" at this position retain CTCF interaction, and the opposite allele with the other genotypes ("A" = 5, "C" = 4, "T" = 1) at this position showed no CTCF interaction.

(A–C) *$p < 0.05$. See Supplemental Experimental Procedures.



(legend on next page)

Figure S7. Strategy and Method for Modeling 3D Genome Using ChIA-PET Data and Modeling Results, Related to Figure 7

(A) Loop structure reconstruction at sub-anchor level. During loop structure reconstruction, each interaction blocks is considered separately. A serial of preliminary structures is generated based on the simplified consideration of energy of the shape of loop, as well as the corresponding distance matrices. Those distance matrices are aggregated to obtain an average distance matrix. Using the sub-anchor heatmap, the average distance matrix is refined. Based on the refined distance matrix, the final structure with properly positioned loops is generated.

(B) Strategy to determine the shapes of loops. Between two strong interacting anchors, the intervening chromatin space will form a loop structure. However, the specific shape of the particular loop will be determined by weak interactions from outside and inside of the loop. For simplicity, we only consider the weak force of interactions within the loop structure. Sub-anchor weak interactions can be detected using sub-anchor heatmap. In the heatmap that absents significant contact signal (a), the loop is shaped as a regular and circular form. In the heatmap with only one significant contact signal (b, c), the contact sites in the loops are being brought close together. With multiple significant contact signals in the heatmap, the loop is configured with different shape (d, e). The strong, but uniform contact signals in the heatmap result the loop to be smaller without modification of its regular and circular properties (f).

(C) Simulation of the 3D structure with consideration of CTCF sequence motif orientation. Two principles are proposed for folding of loops with CTCF sequence motif. The loop with convergent motifs forms hairpin structure, and the loop with tandem motif forms coiled loop structure (a). In (b), schematics show the chromatin folding structure of a CTCF interaction block at 2D level according to the proposed folding principles. During the simulation at sub-anchor level (c), the genomic distance (w_p) is translated to physical distance (d_p) between two corresponding nodes. The angle α_i between loop segments is also considered to constrain the structure. The CTCF motif orientations force the loops to form different shapes. In (d), an example to show the simulated 3D structures on a selected CTCF interaction block (Chr17:63226521-64565170) with different modeling constraints as shown. In the basic model, only the loop segment lengths and angles are considered, resulting a regular chromatin configuration with freely distributed loops (I). Including the CTCF motif orientations in the model introduces more constraints in positioning loops and results in limited flexibility of loops, which made the loops aligned side-by-side (II). Considering the contact signals at sub-anchor level (e) introduces more constraints on the loop conformation and results in irregular loops (III in (d)). Including both CTCF motif orientations and contact signal at sub-anchor heatmap introduces more constraints for loop conformation and yields irregular but aligned loops (IV in (d)). A zoom-in view on the interaction complex is shown for each structure. In the zoom-in view, the spheres denote the interaction anchors, and the red bars indicated the motif alignment.

(D) The 3D structure of nucleome derived from ChIA-PET data at "chromosome" (Left), "segment" (Middle) and "sub-anchor" (Right) resolutions. The color codes at right side of panel denote each chromosome.

(E) 3D DNA-FISH representation of chromatin territory of all chromosomes with false colors. The index of each chromosome is marked on its chromatin territory.

(F) 3D modeling of entire maternal ChrX and *DXZ4-FIRRE-G6PD* loci (zoom-in section).

G. 3D modeling of entire paternal ChrX and *DXZ4-FIRRE-G6PD* loci (zoom-in section).

See also Supplemental Experimental Procedures.

Cell

Supplemental Information

CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription

Zhonghui Tang, Oscar Junhong Luo, Xingwang Li, Meizhen Zheng, Jacqueline Jufen Zhu, Przemyslaw Szalaj, Pawel Trzaskoma, Adriana Magalska, Jakub Wlodarczyk, Blazej Ruszczycki, Paul Michalski, Emaly Piecuch, Ping Wang, Danjuan Wang, Simon Zhongyuan Tian, May Penrad-Mobayed, Laurent M. Sachs, Xiaoan Ruan, Chia-Lin Wei, Edison T. Liu, Grzegorz M. Wilczynski, Dariusz Plewczynski, Guoliang Li, and Yijun Ruan

Supplemental Information

Supplemental Experimental Procedures

Experimental methods

Long-read ChIA-PET library construction and sequencing

GM12878 and HeLa CTCF and RNAPII long-read ChIA-PET (ChIA-PET v2) libraries were performed similar to the short-read ChIA-PET (ChIA-PET v1) protocol (Fullwood et al., 2009) with modifications. Briefly, approximately 100-200 million cells were harvested and fixed by 30 ml of 1.5 mM EGS (ethylene glycol bis[succinimidylsuccinate]) in PBS buffer for 45 min at room temperature. Next, formaldehyde was added to final concentration of 1% to cross-link the cells for another 20 min at room temperature and then neutralized with 0.125 M glycine. The cross-linked cells were lysed by cell lysis buffer and nuclear lysis buffer. Chromatin was obtained and subjected to fragmentation with an average length of 300 bp by sonication. The anti-PolII monoclonal antibody 8WG16 (Covance, MMS-126R) and anti-CTCF antibody (Abcam, ab70303) were used to enrich RNAPII and CTCF-bound chromatin fragments, respectively. ChIP DNA on beads was used for ChIA-PET library preparation. After performing the end-repair and A-tailing using T4 DNA polymerase (NEB) and Klenow enzyme, the ChIP DNA ends were proximity-ligated by the single biotinylated bridge-linker:
Forward strand: 5'- [5Phos]CGCGATATC/iBIOdT/TATCTGACT-3',
Reverse strand: 5'- [5Phos]GTCAGATAAGATATCGCGT-3',
with the 3' nucleotide T over-hanging on both strands. Proximity ligation DNA was reverse cross-linked and fragmented and added sequencing adaptors simultaneously by using Tn5 transposase (Nextera kit, Illumina). DNA fragments contained the bridge-linker at ligation junctions were captured by Streptividin beads, and used as templates for PCR amplification. These DNA products were then subjected to size-selection and paired-end sequencing (2×150 bp) using Illumina Hi-Seq 2500. This modified ChIA-PET method is referred as “long-read ChIA-PET” due to the longer sequence tag produced, which facilitates higher mapping confidence and base pair coverage.

ChIP-nexus library construction and sequencing

Human GM12878 cells were harvested and treated with 1% formaldehyde for 10 min, then quenched with 0.125 M of Glycine. Approximately, eighty million cells were lysed with 10 ml of 0.1% FA buffer (50mM Tris.HCl, pH 7.5, 150mM Nacl, 1mM EDTA, 0.1% SDS), rotating for 6 min at room temperature, then added 900 µL of 10% SDS, and rotated for 2 min at 37 °C to obtain nuclei. Nuclei were suspended in 0.1% FA buffer and subjected for sonication to get chromatin fragments from 100 bp to 500 bp.The fragmented chromatin was aliquot to two tubes for Rad21 (Abcam, ab992) and SMC3 (Abcam, ab9263) antibody immunoprecipitation as previously described (Goh et al., 2012). In brief, 100 µL of protein G/A dynabeads (Life technologies) bounded 10 µg of antibody were incubated with the aliquot chromatin respectively. ChIP chromatin was processed through the ChIP-nexus assay as described in He et al. (2015). The three ChIP-nexus DNA samples with different barcode were sequenced on an Illumina Nextseq 500 platform with single-end sequencing primer. See later section and Figure 2B and Data S1, II.

PacBio Iso-Seq

Total RNA of GM12878 was extracted by using RNeasy Mini Kit (Qiagen). A cDNA library was prepared and separated into three size fractions (0.5-2.0 kb, 2.0-3.0 kb, 3.0-6.0 kb) according to the PacBio Iso-Seq protocol “Using Clontech cDNA Prep and the

SageELF™ Size Selection System". The prepared Iso-Seq library was sequenced on PacBio RS II system with 4 SMRT Cells. PacBio Iso-Seq data was processed by the PacBio SMRT Portal pipeline. Results are presented in Figure S6C.

DNA-FISH validation of *DXZ4-FIRRE-G6PD* co-localization

DNA-FISH was performed to validate the allele-specific co-localization of *DXZ4-FIRRE-G6PD* detect by CTCF ChIA-PET in GM12878 cells. GM12878 cells were harvested and fixed with methanol: acetic acid (3:1) and dropped onto slides for FISH. Fixed cells were then dehydrated through 70%, 85%, 100% ethanol series. Labeled probes were denatured at 76 °C for 2 min and hybridized to pretreated slides at 37 °C overnight. Slides were washed with 0.4X SSC for 2 min at 72 °C, followed by 1 min 0.1% Triton/PBS wash and 3 min 2X SSC wash at room temperature. After washing, slides were mounted with Prolong Diamond Antifade Mountant with DAPI (Life Technologies) and observed under Leica TCS SP8 confocal microscope with 63X objective. Images were analyzed with Image J. The DNA-FISH probe for *FIRRE* is from BAC clone RP11-754H22 labeled with FITC (green) (EmpireGenomics). The probes for *G6PD* and *DXZ4* are customized by Agilent, labeled with Cy3 (orange) and Dyomics 415 (aqua), respectively, covering ChrX:153,686,069-153,783,639 and ChrX:114,955,098-115,027,743 regions. The results are presented in Figure 4E.

3D DNA-FISH for whole nucleome and chromosome 1

DNA-FISH was performed essentially according to Cremer et al. (2008) with some modifications. The cells were fixed in suspension, with 4% paraformaldehyde in PBS, and spun down onto a glass slide using a cytocentrifuge. Then the cells were permeabilized with 0.5% Triton X-100 in PBS (RT, 20 min). To augment permeabilization, the cells were immersed in a cryoprotectant solution (20% glycerol in PBS, RT, 2h) followed by their repeated freezing-thawing above the surface of the liquid nitrogen (4 times for 30 sec). Subsequently the cells were washed in PBS containing 0.05% Triton X-100 (RT, 3 times for 5 min), treated with 0.1 N HCl (RT, 5 min), washed again in 2xSSC at 37° C and incubated in 50% formamide in 2xSSC (4° C, overnight). For visualization of chromosome 1 plus its p and q arms termini, the cells and the probe were simultaneously subjected to DNA denaturation at 80ºC for 5 min. The hybridization was performed at 37°C in humid dark chamber for 2 days. We used the following probes (all from Kreatech): 1) sub-telomere 1pter/ D1S2217/ 170/ 800/green /KBI-40201, 2) sub-telomere 1qter/ D1S555/ 170/ 350/ KBI-40202, and chromosome 1 painting/ blue/ KBI-30001B). After hybridization the specimen was washed, counterstained with TO-PRO-3 DNA-binding dye, mounted, and examined under the Zeiss LSM 780 confocal microscope, using the 405 nm diode laser, 488 nm line of argon laser, 561 nm DPSS diode laser, and 633 nm HeNe laser, at 70 nm/pixel lateral resolution, and 210 nm z-spacing. The image-stacks were deconvolved using Huygens software (SVI) with the maximum-likelihood estimation (MLE) algorithm, and subjected to automatic segmentation of the nuclei using our proprietary software "Segmentation magick" (Walczak et al., 2013).

For all-chromosome territory painting we used 24XCyte MetaSystems kit, and performed the FISH essentially as described above, except that the probe and the specimen were denatured separately. The probe was treated according to the manufacturer's protocol, but the specimen was denatured at 70°C for 5 min. The images of all-chromosome painting were acquired with the Zeiss LSM 780 confocal microscope using the 355 nm

UV laser, 458 nm line of argon laser, 561 nm DPSS diode laser, and 633 nm HeNe laser, with objective and the spatial resolution as stated above. DNA was counterstained using Hoechst 33342. To separate signals of considerably spectrally overlapping fluorophores, lambda-scanning using 32-channels detector was performed followed by previously described nonnegative linear unmixing method (Neher et al., 2009) using estimated spectra. The iterative algorithm was initialized by fluorescence spectra packed at the wavelength of the reference spectra (http://www.fluorophores.tugraz.at/). The algorithm previously implemented (Neher et al., 2009) and available as ImageJ plugin Poisson NMF was employed when necessary. The excitation wavelengths as well as range of spectral channels were selected according to Neher and Neher (2004). Resulted image-stacks were deconvolved and segmented as described above. The spectral unmixing was calibrated based on human metaphase spreads hybridized with 24XCyte MetaSystems kit exactly according to the manufacture's protocol. The 3D reconstructions of chromosome 1 and all-chromosome territories were obtained with Imaris (Bitplane) and Amira (FEI) respectively. See Figure 7C for individual chromosome (Chr1) result and Figure S7E for all chromosomes (i.e. whole nucleome) result.

Co-localization detection of CTCF and RNAPII by immunostaining and SIM super-resolution microscopy

The immunofluorescence staining was performed according to our routine procedures (Hall et al., 2015). The GM12878 cells were fixed in suspension, with 4% paraformaldehyde in PBS, and spun down onto a glass slide using a cytocentrifuge. Then the cells were blocked and permeabilized with 5% NDS in 0.2% Triton X-100 in PBS (RT, 1 h) and incubated with the mixture of the primary antibodies: mouse anti-RNA Pol II CTD repeat (YSPTSPS) antibody (Abcam) and rabbit anti-CTCF (Cell Signaling) antibody both diluted in 1:100 in PBS containg 5% NGS and 0.2% Triton X-100 in PBS (4° C, overnight). Then the cells were washed three times and incubated sequentially with the appropriate species-specific secondary donkey antibodies conjugated to Alexa488 or to Cy3 (both from Jackon Immunoresearch), diluted 1:200 and 1:300, respectively. To counterstain chromatin, Hoechst 33342 was used. The non-immune IgGs (Abcam), coupled to the same fluorophores was used as a control for immunoreaction specificity. The specimens were analysed in the Core Facility for Laser Techniques at the Institute of Experimental and Clinical Medicine in Warsaw, Poland using a Zeiss ELYRA PS.1 super-resolution system equipped in Plan Apochromat 100x/1.4 Oil immersion objective and the following lasers: diode laser 405nm, and solid state lasers emitting at 488, 561, or 642nm. Series of z-stacks were acquired with a step of 85 nm. The sampling density of the obtained images was 25 nm per pixel. The images were acquired using two cooled EM CCDs at 5-fold rotation of the diffraction grid. The super-resolved images were generated using Zeiss Zen 2012 black edition software. The colocalization analysis was performed using Fiji software (http://fiji.sc/Fiji) module Coloc 2 and the plugin Jacop (http://rsb.info.nih.gov/ij/plugins/track/jacop.html). The quantitative analysis of colocalization degree between CTCF and RNAPII rendered the Spearman correlation coefficient to be 0.64 ± 0.016 (n=21 cells). Spearman coefficient may vary from -1 (exclusion) through 0 (no colocalization) to 1 (perfect colocalization) (Bolte and Cordelieres, 2006). To control for colocalization significance we performed the same analysis using randomized images generated from each cell image; in randomized images the Spearman coefficient occurred to be close to 0 (-0.003 ± 0.002), meaning that CTCF-RNAPII colocalization is highly significant. Thus, there is a very close overlap between these two proteins at the subdiffraction level. See later section and Figure 7E for results.

Co-localization detection of CTCF and RNAPII by FRET-FLIM

Förster resonance energy transfer (FRET) analysis was performed in GM12878 cells following immunocytochemistry performed as described above, with mouse anti-RNA Pol II antibody and rabbit anti-CTCF antibody, and with secondary donkey anti-mouse antibody conjugated to Cy3, serving as an energy acceptor, and with donkey anti-rabbit antibody conjugated to Alexa488 serving as a donor. A non-immune antibody coupled to Cy3 served as a negative control for energy transfer. The measurement of fluorescence lifetime of the donor was performed on a Picoquant PicoHarp 300 Time-Correlated Single Photon Counting (TCSPC) system attached to Leica Sp8 confocal microscope, using 63x oil immersion objective (NA 1.4). The donor fluorescence (Alexa488) was excited by 488 nm pulse of 80Mhz WLL laser (NKT Photonics). Fluorescence intensity was collected in the wavelength band from 500 nm to 550 nm to avoid acceptor fluorescence. Typical fluorescence decays were fitted with the resulting sum of two exponentials interactively convolved with the instrument response function. The mean fluorescence lifetimes were calculated as the mean values of the fit function. Lifetime analysis was performed using SymPhoTime (Picoquant, Germany) software. Two-dimensional map (512x512 pixels) of the mean lifetime value was generated for the given experimental condition. The FRET experiments were repeated independently three times. See later section and Figure 7F for results.

Immunostaining of lampbrush chromosomes using CTCF antibody

Newts of the species *Pleurodeles waltl* were raised at 20°C in techniplast aquariums, enlightened for a photoperiod of 12 hr. Experiments were conducted in accordance with European guidelines for Care and Use of Laboratories Animals. Ovarian biopsies were performed on adult female newts that were anesthetized in 0.1% MS222 (Amino-benzoic Acid Ethyl, Fluka). Germinal vesicles (nuclei) from stage V-VI oocytes were manually isolated in 75mmol/L KCl, 25mmol/L NaCl, 0.01mmol/L MgCl2 and 0.01mmol/L CaCL2, pH7.2 and Lampbrush chromosomes were prepared as previously described (Penrad-Mobayed et al., 2010). Nuclear spread preparations were fixed for 30 min at 4°C in phosphate buffer saline (PBS) containing 2% paraformaldehyde and 1 mM $MgCl_2$, washed 3X 10 minutes with PBS and blocked for 10 min with Horse serum at 10% dilution. They were then incubated with primary antibody (CTCF, Novus biologicals) at 1:100 dilution for 1 hr at room temperature, washed 3X 10 minutes with PBS, incubated with secondary antibody (Alexa 488 goat anti-rabbit IgG, Invitrogen Corp., Carlsbad, CA) at 1:1000 dilution for 1 hr and then washed again in PBS. Preparations were post-stained with Hoechst 33342 (1/1,000; Invitrogen). Standard transmitted light and fluorescence microscopy was performed at the Imaging facility of the Institut Jacques Monod. Images were captured using a camera (CoolSnap HQ, Photometrics) driven by the software Metamorph6 (Universal imaging). A DM IRBE microscope (Leica Microsystems, Wetzlar, Germany), equipped with a piezoelectric Objective Scanner P-721 PIFOC (Physik Instrumente, Karlsruhe/Palmbach, Germany) placed at the base of a 100X PlanApo numerical aperture (NA) 1.4 PH3 objective and a Coolsnap HQ interline charge-coupled device (CCD) camera (Photometrics, Tucson, USA) were used. Phase contrast images were acquired and for imaging Hoechst and Alexa488, two cube filter set were respectively used XF113-2 and XF100-2 (Omega Optical, Brattleboro, USA). The acquisition software (MetaMorph; Universal Imaging, Downingtown, PA) was set to acquire two images at each Z-step (0.3 µm). ImageJ software (version: 1.48v, http://imagej.nih.gov/ij/) was used for imunofluorescence signal quantification. The

fluorescence signal was measured on the axes and the loops (10 different localization for each) for 6 lampbrush chromosomes per conditions (IgG control antibody and CTCF antibody). Area of interest was selected with a rectangle. The background was subtracted by measuring the signal for the same area in a position without chromosome. Results are indicated as means ± s.e.m. Statistical analysis was done with student T test. The results are presented in Figure 7G.

Data processing and analysis

ChIA-PET data processing pipeline and PET reads classification

ChIA-PET v1 (short-read) sequence data for K562 and MCF7 was obtained from ENCODE website (ENCODE Project Consortium, 2012), and was processed as described in Li et al. (2012).

ChIA-PET v2 (long-read) sequence data processing was performed similar to Li et al. (2012) with modifications. Briefly, pair-end read (PET) sequences were scanned for the bridge linker sequence and only PETs with the bridge linker were used for downstream processing. After trimming the linkers, the sequences flanking the linker were mapped to the human reference genome (hg19) using bwa-mem (Li and Durbin, 2010) and only uniquely aligned (MAPQ ≥ 30) PETs were retained. PCR duplicates were removed using the MarkDuplicates tool of the Picard Tools library (http://broadinstitute.github.io/picard/).

Sequence data processing statistics for both ChIA-PET v1 and v2 libraries are summaries in Table S1. The data source from ENCODE is summarized in Table S5.

Each PET was categorized as either a self-ligation PET (two ends of the same DNA fragment) or inter-ligation PET (two ends from two different DNA fragments in the same chromatin complex) by evaluating the genomic span between the two ends of a PET. PETs with a genomic span less than 8 kb are classified as self-ligation PETs and are used as a proxy for ChIP fragments since they are derived in a manner analogous to ChIP-Seq mapping for protein binding sites. PETs with a genomic span greater than 8 kb are classified as inter-ligation PETs and represent the long-range interactions of interest. To accurately represent the frequency of interaction between two loci and to define the interacting regions, both ends of inter-ligation PETs were extended by 500 bp along the reference genome, and PETs overlapping at both ends (with extension) were clustered together as one PET cluster.

The number of PETs in a PET cluster reflects the frequency of interaction between two genomic regions. In this study, the uniquely mapped and non-redundant PETs from all replicates of GM12878 CTCF and RNAPII libraries were combined for PET cluster generation, respectively. The combined GM12878 CTCF and RNAPII ChIA-PET libraries were both deeply sequenced (Table S1). We therefore set the PET count cutoff for PET clusters as 4 for GM12878.

We observed that a lot of anchors of distinct PET clusters were located within the same protein factor binding peak. It is clear that these binding peaks reflect the real chromatin interaction loci in the nucleus. In order to streamline the PET clusters data structure, we

collapsed the individual anchors of all PET clusters with 500 bp extensions to generate merged anchors. We then used the merged anchors to further cluster raw PET clusters. See Figure S2B for schematic illustration. Throughout the text, the merged PET clusters are referred to as interactions or connections. Un-clustered individual inter-ligation PETs and PETs in the clusters below the PET cutoff are referred as singletons.

Identification of binding peaks from ChIA-PET data

All uniquely mapped and non-redundant reads including self-ligation and inter-ligation were used to pileup the protein factor binding coverage along the chromosomes for visualization. Also, all of these reads were applied to the MACS pipeline (version 1.4.2) (Zhang et al., 2008) for protein factor binding peak identification with default parameters.

Contact frequency heatmap generation from ChIA-PET and Hi-C data

The normalized inter-chromosomal contacts between all pairs of chromosomes were calculated as previously described (Lieberman-Aiden et al., 2009). The expected number of inter-chromosomal interactions for each chromosome pair *i,j* was calculated by multiplying the fraction of inter-chromosomal PETs connected to chromosome *i* with the fraction of inter-chromosomal PETs connected to chromosome *j* and then multiplying by the total number of inter-chromosomal PETs. Normalized contact frequencies were calculated by taking the actual number of inter-chromosomal PETs between chromosome *i* and *j* and then dividing by the expected value.

For the intra-chromosomal contacts, each individual chromosome was divided into 1 Mb windows and PETs (excluding self-ligation PETs) were binned according to the location of both ends to produce a contact frequency matrix. The contact frequency matrix for Hi-C data was generated in a similar manner. Then the contact frequency matrix undergoes internal quartile normalization. Briefly, the raw contact frequency values less than Q_1-1.5×IQR are elevated to the Q_1-1.5×IQR value, and the raw contact frequency values greater than the Q_3+1.5×IQR value are reduced to the Q_3+1.5×IQR value, where Q_1 is the 25^{th} percentile, Q_3 is the 75^{th} percentile and IQR is the interquartile range. This normalization process does not dramatically transform the raw data; instead, it fits the outlier data into a narrower and more natural range for the ease of heatmap visualization.

Hi-C data for the GM12878 cell line was obtained from Selvaraj et al. (2013).

Contact frequency heatmap between ChIA-PET and Hi-C display very similar patterns

The self-ligation PET data (derived from chromatin fragment self-circularization) reflects ChIP-enriched DNA fragments, which are used to call protein-binding sites same as ChIP-PET and ChIP-Seq data (Johnson et al., 2007; Wei et al., 2006). The inter-ligation PET data are derived from ligation between different chromatin fragments tethered together in the same chromatin complex. By mapping coordinates of PET reads to a reference genome (hg19), inter-ligation PET data can be classified into clustered PETs and singleton PETs. Clustered PETs were inferred as enriched interactions between discrete chromatin loci associated with the protein binding sites. As reported previously, enriched interaction data has been intensively analyzed for specific chromatin interactions associated with protein factors of interest (Dowen et al., 2014; Fullwood et

al., 2009; Li et al., 2012). However, in fact the vast majority (~70%-~90%) of inter-ligation data from ChIA-PET are singleton PETs, according to the clustering scheme applied (Table S1). We suspected that inter-ligation singleton PET data from ChIA-PET are the non-ChIP-enriched component similar to Hi-C interaction data. In principle, both methods employ nuclear crosslinking, chromatin fragmentation, proximity ligation and paired-end sequencing to map long-range chromatin interactions. The major difference is that ChIA-PET utilizes ChIP to enrich for chromatin interactions associated with specific protein factors, whereas Hi-C does not. In our previous analyses, we focused exclusively on the ChIP-enriched interactions. Inspired by Hi-C data analysis using large bin size (from Mb down to tens of kb) for mapping interaction data (Dixon et al., 2012; Lieberman-Aiden et al., 2009), we considered that singleton ChIA-PET data might share similar features. Indeed, using a similar analysis approach, the ChIA-PET singleton PET profile closely resembled Hi-C data (Figure S1B). To further test this, we processed all uniquely mapped ChIA-PET (CTCF and RNAPII) data from GM12878 cells, using Hi-C-style contact frequency heatmaps, and compared with Hi-C data (Selvaraj et al., 2013). As shown, the overall chromatin interaction profiles at inter-chromosomal, intra-chromosomal, and segmental (Mb) levels were very similar (**ρ**=0.75 to 0.83, Figure S1C-E), suggesting that ChIA-PET data captures higher-order topological structures similar to Hi-C. As expected, ChIA-PET datasets also provided detailed, high-resolution protein binding and chromatin interaction associated with CTCF and RNAPII (Figure S1E).

Reproducibility assessment of ChIA-PET data

Here, we assess the reproducibility of the long-read ChIA-PET at the whole genome and individual chromosome level. At the whole genome level, contact frequencies between all chromosomes (Chr1-22 and X) are calculated by using all inter-chromosomal PETs sequenced from a ChIA-PET library. Then the contact frequency matrix undergoes normalization as described in the previous section. The normalized contact frequencies between replicates are compared by calculating the Spearman's correlation coefficient.

Similarly, at the individual chromosome level, the contact frequencies between all loci within in an individual chromosome are calculated by using the intra-chromosomal PETs (excluding self-ligation) mapped to the particular chromosome at 1 Mb resolution. Then, the internal quartile-normalized contact frequency matrices for the same chromosome from replicates are compared by calculating the Spearman's correlation coefficient.

These reproducibility test procedures were performed systematically between all GM12878 replicates of CTCF and RNAPII, respectively. From the results summarized in Data S1, I, we conclude long-read ChIA-PET data are highly reproducible.

Processing of ENCODE identified CTCF and cohesin binding peaks

The uniform ChIP-Seq peak calls for CTCF, SMC3 and RAD21 in GM12878, HeLa, K562 and MCF7 were downloaded from the ENCODE website (Table S5). For GM12878, the uniform peaks from all four CTCF replicates were pooled and the 4-way consensus regions were extracted and used as the ENCODE identified CTCF-binding peaks. For GM12878, the ENCODE uniform CTCF peaks were compared with the CTCF-binding peaks derived from ChIA-PET. The results indicate 74% of the ChIA-PET derived CTCF-binding peaks directly overlap with the ENCODE identified CTCF-binding peaks (Figure S2D).

Similarly, all GM12878 cell SMC3 and RAD21 uniform peaks from ENCODE were pooled, and the consensus regions were used as the ENCODE identified cohesin-binding peaks. The ENCODE CTCF and cohesin consensus binding peaks for HeLa, K562 and MCF7 were produced in the same way as GM12878. All available CTCF, SMC3 and RAD21 replicates from ENCODE were used in this part of analysis.

CTCF-binding motifs at the loop anchors

For GM12878, the consensus regions between ENCODE CTCF and cohesin peaks were identified, and then the nucleotide sequences in these consensus regions were extracted and applied to STORM (Schones et al., 2007) for CTCF-binding motif identification (CTCF position weight matrix downloaded from JASPAR database http://jaspar.genereg.net). Only the STORM returned motif match with positive score is retained for further analysis. In total, 22,653 candidate CTCF-binding motifs were identified in the 23,932 CTCF and cohesin consensus peak regions, and these motifs were used to determine the directionality of the CTCF loop anchors. For comparison purpose, we also identified 10,947 CTCF-binding motifs in the 11,487 CTCF peak regions without cohesin co-occupancy. This suggests it is equally likely to identify CTCF-binding motifs in the CTCF peak regions with or without cohesin binding. For simplicity, only the 22,653 CTCF motifs located within the CTCF and cohesin co-binding peaks were considered in all the analyses of this study.

CTCF interaction loop directionality assignment

Here, we describe the process of assigning the directionality of the CTCF interaction loops based on the CTCF-binding motif orientation. First, we overlapped the 22,653 CTCF-binding motifs with the 42,297 CTCF interaction loop anchors. The results showed that 38,201 (90%) of these CTCF loops have both anchors overlapping with at least one CTCF motifs. Of these 38,201 interactions, 31,412 have both anchors only overlapping with one CTCF-binding motif, and 6,789 have at least one anchor overlapping with more than one CTCF motifs. Next, if the strand orientations of the motifs in a loop anchor showed contradicting pattern, then such loops were filtered out (n=2,971). As a result, 35,230 CTCF interactions with both anchors having unique motif strand orientation were retained for further characterization. Finally, the directionality/orientation of these 35,230 CTCF loops were assigned based the unique CTCF motif orientation pattern within their anchor regions. The results showed that 22,709 (64%) CTCF loops were of convergent motif orientation at the two anchors, hence, designated as convergent CTCF loops; 5,844 (17%) with tandem left orientation; 5,830 (17%) with tandem right orientation and 847 (2%) with divergent orientation (Figures 2A and S2F). We also discovered tandem loops are relatively smaller than convergent loops, but with a very similar overall size distribution pattern, which was clearly different from the divergent loops (very small) (Figure S2G).

ChIP-Seq identified CTCF, RAD21 and SMC3 binding summit analysis

To investigate the binding position relationship between CTCF and cohesin at CTCF-binding motifs, we downloaded the narrow binding peaks of CTCF, RAD21 and SMC3 in GM12878 from ENCODE, and calculated the distance from the peak summits (i.e. highest base pair coverage position) to the nearest CTCF-binding motif center. The calculated distances were plotted as density plots according to the CTCF motif orientation and the categories of CTCF loops in which the motif is involved (Data S1, II).

ChIP-nexus data processing for protein factor binding footprint identification

ChIP-nexus sequencing raw reads (including the fixed and random barcodes) were first reduced for PCR redundancy, and then the fixed and random barcodes (read positions 1-9) were trimmed. The remaining sequences were aligned to human genome assembly hg19 by bwa-mem (Li and Durbin, 2010) and only uniquely aligned (MAPQ ≥ 30) PETs were retained. The sequence alignment results were used as input to MACE (v1.2) (Wang et al., 2014) to identify RAD21 and SMC3 occupancy borders (footprints).

In addition, we downloaded the CTCF ChIP-Exo data from Rhee and Pugh (2011) and passed it to bwa-mem and MACE for CTCF occupancy border (footprint) detection.

For the detected CTCF, RAD21 and SMC3 binding footprints, we first identified the footprints with CTCF-binding motif mapped inside, and then calculated the distance from the 5' and 3' border to the corresponding CTCF motif according to the motif sequence orientation. The calculated distances were plotted as density plots according to the motif orientation and the categories of CTCF loops in which the motif is involved. The results are presented in Figure 2B and Data S1, II.

Structural analyses related to CTCF-mediated chromatin contact domains

a. Identification of CTCF-mediated chromatin contact domains

In this section, we describe the procedures of identifying CTCF-mediated chromatin contact domains (CCD) by using the CTCF and cohesin peak-supported CTCF interaction loops in GM12878, HeLa, K562 and MCF7 (Table S2). Firstly, based on the continuous connectivity and coverage of individual CTCF loops, each chromosome is partitioned in to a number of CTCF loops covered candidate domain regions. For example, in GM12878, the 42,297 CTCF loops are clustered into 1,689 candidate domains in the whole genome (excluding ChrY), each with continuous CTCF loop coverage. The rest of the genome with no CTCF loop coverage is considered as gap regions. Secondly, to further refine the contact domains, we calculated the aggregate CTCF loop coverage along all chromosomes at base-pair resolution and identified the regions with very low loop coverage (lower than the 5th percentile). These low coverage regions were subtracted from the candidate domains. For GM12878, the subtraction of the low coverage regions from the 1,689 candidate domains yielded 2,308 contact domains. Finally, the resulted domain regions from the last step were filtered by genomic span. The contact domains with genomic size smaller than 10 kb are excluded from the downstream analyses. For GM12878, HeLa, K562 and MCF7, there are 2,267, 3,071, 2,385 and 3,317 CTCF-mediated chromatin contact domains defined, respectively.

In contrast, the complementary genomic regions (after subtracting hg19 reference genome assembly gaps) to the CCDs are defined as gaps. For GM12878, HeLa, K562 and MCF7, there are 2,429, 2,975, 2,464 and 3,266 gap regions defined, respectively. See Figure S3A for an example of the CTCF-mediated chromatin contact domain and domain gap structure in GM12878.

b. CTCF motif orientation characterization at the CTCF-mediated chromatin contact domain boundaries

Here, we briefly describe how the motif orientation is defined at the CCD boundaries in GM12878. We first identified the loop anchors located at the boundaries of each CCD, i.e. one anchor at the left boundary and one anchor at the right boundary. Next, we examined the CTCF motif orientation pattern within the anchors at the boundaries. When an anchor at the boundary overlaps with more than one CTCF motifs, the motif orientation is only considered if a unique motif orientation direction can be found amongst these motifs; otherwise, the motif orientation at this anchor site is considered as un-identifiable. In total, 759 CTCF anchors at the domain boundaries were identified with multiple overlapping CTCF motifs, and the motif orientation pattern was uniquely identifiable. If an anchor at the boundary does not overlap with any CTCF motifs, then the motif orientation at this boundary anchor site is also considered as un-identifiable. For the 2,267 CCDs in GM12878, 4,302 of the 4,534 (2,267×2) domain boundaries had identifiable CTCF motif orientation, of which 97% (4,177) were pointing towards the inside (inward) of the corresponding CCD (Figure 2E). When the domain boundary motif orientations are considered pair-wisely, 84.9% (1,926) of the domains were of convergent motif orientation pattern (Figure S3E).

c. Relative position of the CTCF loop anchors overlapping with multiple CTCF-binding peaks and motifs

In total, we identified 1,222 CTCF loop anchors overlapping with multiple CTCF-binding peaks and motifs in GM12878. As previously discussed, the CTCF binding motifs used in this study were identified within the CTCF and cohesin co-binding peaks. Therefore, a CTCF anchor overlapping with multiple motifs automatically means it overlaps with multiple CTCF-binding peaks. Within the 1,222 CTCF loop anchors with multiple CTCF-binding motifs, 759 (62%) had identifiable unique motif orientation pattern (i.e. non-contradicting motif orientation). We then calculated the position of these 759 anchors relative to the CCD in which they were located. The results showed the CTCF anchors with multiple CTCF-binding peaks and motifs were significantly more likely to be identified at the two boundary regions of the CCDs (Figure S3F).

d. Identification and characterization of simple-convergent CTCF loops

The convergent loops can be further segregated into simple- and complex-convergent loops. A convergent loop is considered as a simple-convergent if it does not entirely contain any other loops (convergent nor tandem) and anchors, otherwise, it is considered as complex-convergent. In total, 5,094 simple convergent loops were identified in GM12878. Generally, simple convergent loops are smaller than tandem and complex convergent loops (Figure S3H).

Next, we examined the position of the simple convergent loops in the CCDs. For each CCD containing simple convergent loop(s), the relative position of the simple-convergent loop center inside the CCD were calculated and summarized as a density curve (Figure S3I, red line). As a control, the relative position of the tandem loop center in the CCD was also calculated (Figure S3I, blue line). In comparison, the simple-convergent loops are likely to be identified at the boundaries of the CCDs, whereas, the tandem loops are evenly distributed within the CCDs.

e. CTCF-mediated chromatin contact domain member loop structure

For GM12878, we characterized directionality of the member loops of each CCD. A loop is considered as member loop of a CCD if it is entirely located with this CCD. Only the

convergent (simple and complex) and tandem (left and right) CTCF loops were used in this analysis, and the detailed numbers of member loops in CCD are summarized as a histogram in Figure S3G.

f. Orientation consistency of tandem loops within individual CTCF-mediated chromatin contact domains

For each CCD covering at least 2 CTCF loops with tandem motif orientation at the two anchors, we calculated the tandem loop direction consistency as:

$$consistency = \frac{\max(n_{left}, n_{right})}{n_{left}+n_{right}},$$

where n_{left} is the number of loops with motifs at anchors both pointing to the left, and n_{right} is the number of loops with motifs both pointing to the right. Hence, for each CCD, the theoretical range of its consistency score is [0.5, 1.0], with score of 0.5 indicating equal number of tandem loops with motifs pointing to the left and right (i.e. no directional bias), and score of 1.0 indicating all tandem loops are of identical directionality (i.e. total directional bias). We also randomly shuffled (rewired) the tandem loop direction and recalculated the consistency score for each CCD with at least two tandem loops. The results suggested that the actual observed direction consistency scores for the CCDs are significantly higher than the rewired data, and the tandem loops within the same CCD have strong tendency of having the same directionality (Figure 2F and Figure S3J).

Comparison of chromatin structure defined by ChIA-PET and Hi-C data

We comprehensively compared the CCDs with the chromatin organizational units derived from Hi-C studies in GM12878. The size range of CTCF loops defined chromatin contact domains is in close concordance with the topologically associated domains (TADs) in Hi-C analysis (Dixon et al., 2012), except Hi-C-defined TADs were generally larger than CCDs (Figure S3C), which probably reflects the difference in detection resolution between Hi-C (bin size 40-100 kb) and ChIA-PET (CTCF-binding site 100-500bp). Moreover, higher-resolution *in situ* Hi-C data more closely matched ChIA-PET defined CCDs (Figure 2C) and individual loops (Figures 2D, S3C and S3D), particularly in the upper range. These observations suggest that CTCF-targeted ChIA-PET and non-selective Hi-C achieved high degree of agreement in detecting sub-megabase chromatin domain structures. Detailed comparison procedures are described as follows:

a. Comparison between ChIA-PET defined CTCF-mediated chromatin contact domains and Hi-C defined topological associated domains

The Hi-C defined topological associated domains (TAD) data of IMR90 cells were downloaded from the NCBI GEO database (GEO: GSE35156) and compared with the GM12878 CTCF ChIA-PET defined CCDs. We tried to overlap the Hi-C defined TADs with the CTCF ChIA-PET defined CCDs, and also calculated the distance from the left and right boundaries of TAD to the closest CCD left and right boundaries, respectively. The results showed that most of TADs contain 1 or 2 CCDs inside their boundaries, and TAD boundaries are flanking outside the boundaries of CCD (Figure S3C, lower). The TAD mapping coordinates were converted from hg18 to hg19 by using the liftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

b. Comparison between CTCF interaction loops and *in situ* Hi-C defined loops from Rao et al., 2014

The interaction loops data from both the primary (Rep 1) and secondary (Rep 2) *in situ* Hi-C replicates of GM12878 were downloaded from the NCBI GEO database (GEO: GSE63525), and compared with the 42,297 GM12878 CTCF ChIA-PET loops with CTCF and cohesin peak support. An *in situ* Hi-C defined interaction loop is considered as matched if it overlaps with CTCF ChIA-PET interaction loop at both anchors with 10 kb extension. The extension window of 10 kb was used as most of the *in situ* Hi-C defined loops in the two replicates were of anchor size of 10 kb. Other extension window sizes were also tried, but did not significantly change the results (data not shown). The comparison results show that 83.5% and 83.3% of the *in situ* Hi-C interaction loops are matched by GM12878 CTCF ChIA-PET interactions, respectively (Figure S3D). In addition, we calculated the distance from the left and right boundaries of *in situ* Hi-C loops to the closest left and right boundaries of CTCF interaction loops. The distance distributions indicate *in situ* Hi-C boundaries are within close distance to the CTCF loop boundaries (Figure S3C, upper).

c. Comparison between CTCF ChIA-PET identified interactions and *in situ* Hi-C defined loops in respect to CTCF-binding motif orientation

In the previous section, we compared the CTCF-mediated loops defined in this study to the *in situ* Hi-C (Rao et al., 2014) defined loops from two replicates. Here, we compare the *in situ* Hi-C defined loops generated from the two replicates combined with our CTCF ChIA-PET identified interaction in GM12878. Rao et al. (2014) identified 9,448 chromatin interaction loops in GM12878 from the combined dataset, of which ~86% were associated with CTCF and cohesin binding. Within the 9,448 loops, only 2,857 (30%) had the two corresponding interaction anchors containing a unique CTCF-binding motif. The rest 70% *in situ* Hi-C identified loops were not assigned with unique motif orientation likely due to the large anchor size containing multiple CTCF motifs that could have contradicting orientations. Of the 2,857 loops with unique CTCF motifs, they found 2,574 (90%) are of 'inward-facing' (convergent) motif orientation, and only 273 (9.6%) loops are of tandem orientation (Table S3). In contrast, our CTCF ChIA-PET data identified 42,297 CTCF-mediated loops all having CTCF and cohesin occupancy at both of the interacting anchors, and 35,230 (83%) of these loops have both anchors with uniquely identifiable CTCF motif pattern, of which 22,709 (64.5%) are convergent, 11,674 are tandem (33.1%) and 847 are divergent (2.4%) (Table S3). When we classified the CTCF ChIA-PET identified loops into categories according to motif orientation, we included those loops with anchors overlapping with multiple motifs as long as the motif orientations were not contradicting. Further comparison showed that 81% of the Rao et al., 2014 reported loops could be recapitulated by our CTCF loops. Furthermore, within the 2,857 *in situ* Hi-C identified loops with unique CTCF motif patterns (2574 convergent, 273 tandem), 98% (2,533) of the convergent loops and 88% (239) of the tandem loops are also captured by our CTCF ChIA-PET identified loops (Table S3). Therefore, we believe the chromatin interaction detection by our CTCF ChIA-PET and the subsequent CTCF motif assignment to the loops are reliable. We then tried to use our CTCF ChIA-PET identified interactions to determine the loop orientation of the 70% *in situ* Hi-C identified loops without assigned CTCF motif pattern. The results showed that 4,886 (69.5%) of these loops could be determined for CTCF motif pattern at the two interacting anchors, with 3,651 (80%) and 860 (19%) in convergent and tandem orientation, respectively.

d. Validity test of the CTCF ChIA-PET identified interactions in respect to CTCF-binding motif orientation

To test if tandem loops identified in our study were false positives or by random chance, we address this question at two levels. First, if the CTCF interaction orientations were by random chance at the whole genome scale, then the 4 classes of orientations (convergent, tandem left, tandem right, and divergent) would be each of 25% probability. But the observed rates for each category are otherwise (Table S3). Particularly, the two categories of tandem loops are significantly below the expected (16.6% and 16.5%). Second, considering that each CCD is a constrained structural unit, we then test the random probability of CTCF loop patterns using a rewiring approach within each CCD: motifs located in each CCDs are randomly paired to create rewired loops. The permutation tests showed that the categorical allocations of CTCF-mediated loops with distinct motif orientation significantly differ from random expectation ($P < 0.0001$, chi-squre test) (Table S3).

RNAPII interaction, active chromatin and gene expression in the CTCF-mediated chromatin contact domains and domain gaps

To functionally characterize CTCF-mediated chromatin topology in GM12878, we analyzed: 1) transcriptional chromatin interactions identified by RNAPII ChIA-PET, 2) the chromatin state of CTCF-defined structures, using histone modification data by ChIP-Seq (ENCODE Project Consortium, 2012) (Table S5), and 3) gene expression output by RNA-Seq (Figure 3A). About 77% (2.22 Gb) of the reference genome is covered by CCDs, with the remaining 23% covered by gap regions (Figure S4A). Most of the large gaps are in repetitive sequence regions where ChIA-PET data cannot be mapped. Some small gaps are genuine, while others are possible artifact (i.e. due to coverage). Within the CTCF-defined genome landscape, 88% of RNAPII-associated chromatin loops (0.67 Gb) were located within CTCF-loop domains, with only 12% (0.09 Gb) falling in gap regions (Figure S4A). The same was observed for active chromatin regions, defined by histone modifications, and actively transcribed regions, by RNA-Seq (Figure S4A). Collectively, this indicates that most transcriptional activity occurs within CTCF-looped chromatin structures. The detailed analytical procedures of these comparisons are described below.

For GM12878, we aggregated the genome space covered by CCDs, which covered 77% (2.2 Gb) of the human genome excluding the assembly gaps (based on the hg19 genome assembly). Similarly, we also aggregated the genome space covered by RNAPII interactions; active chromatin (defined by ChromHMM, see section below) and regions with RNA-Seq read coverage ≥10 (i.e. actively transcribed). We characterized the CTCF-mediated chromatin contact domains and domain gaps by intersecting them with the RNAPII interaction loops, active chromatin and actively transcribed regions. The results showed that 1) comparing to the CTCF-mediated chromatin contact domain coverage, the RNAPII interactions, active chromatin and actively transcribed regions covered much smaller genome space; 2) majority of the RNAPII loops covered regions, active chromatin regions and actively transcribed regions were located within the CTCF-mediated chromatin contact domains, which suggests most of functional section of the human genome is positioned within CTCF interaction loops (Figure S4A).

Chromatin state characterization by ChromHMM

Chromatin state information derived from the ChromHMM method (Ernst and Kellis, 2012) was extensively used in this study to determine the active or inactive status of the analyzed genome regions. For GM12878 and K562, the ChromHMM data was

downloaded from ENCODE (ENCODE Project Consortium, 2012). For HeLa and MCF7, eight histone modification marks (H3K4me1, H3K4me2, H3K4me3, H4K20me1, H3K27ac, H3K27me3, H3K36me3 and H3K9ac) and CTCF ChIP-Seq data mapping results were inputted to the ChromHMM program to compute the chromatin state of HeLa and MCF7 genome, respectively. The ChromHMM was run with the default parameter settings. By default, the ChromHMM method segments the genome into distinct states with color-code as: Active promoter – Bright Red, Weak promoter – Light Red, Inactive/poised promoter – Purple, Strong enhancer – Orange, Weak/poised enhancer – Yellow, Insulator – Blue, Transcription transition/elongation – Dark Green, Weak transcribed – Light Green, Polycomb-repressed – Gray and Heterochromatin/Repetitive/Copy number variation regions – Light Gray.

RNAPII interaction complexes

a. Refinement of RNAPII interactions

To increase the fidelity of the RNAPII interaction loops identified in GM12878, HeLa, K562 and MCF7, we used gene promoter, TES (transcription end site) and enhancer information to filter the RNAPII loops. Only the RNAPII interaction loops with both anchors overlapping any of promoter, TES or enhancer were kept for further analyses. The promoters were defined as the ±2 kb regions surrounding the TSS (transcription start site). The TSS and TES coordinates were adopted from GENCODE (version 14) (Derrien et al., 2012). The cell-type specific enhancer locations were adopted from the ChromHMM data, and both strong and weak/poised enhancers were used. The numbers of RNAPII interaction loops before and after the refinement in GM12878, HeLa, K562 and MCF7 are summarized in Table S2. RNAPII-associated loops were generally much smaller than CTCF loops (Figure S4B), likely because many RNAPII loops are gene-centric.

b. Identification of single-gene and multi-gene RNAPII interaction complex models

The single-gene and multi-gene RNAPII interaction complexes were identified as previously described in Li et al., 2012. Briefly, the refined RNAPII interaction loops in each cell line were collapsed based on the connectivity of overlapping anchors with other loops to form complex interaction models. In addition, the identified interaction complexes can be further classified as single-gene (SG) model if only one promoter is involved, or multi-gene (MG) model if multiple promoters are involved (note: a promoter in the SG or MG model could involve multiple genes, i.e. bi-promoter). The numbers of SG and MG models defined in GM12878, HeLa, K562 and MCF7 are summarized in Table S2.

Characterization of the CTCF interaction anchor and loop regions

Here, we describe the approach of using histone modification ChIP-Seq, RNAPII ChIP-Seq and TSS location data to characterize the CTCF interaction anchor and loop regions as aggregation density plots. We first identified the genomic segments between every two consecutive loop anchors within the same CCD. By doing so, the repetitive usage of the loop regions are avoided, and it also ensures the domain gap regions are not included. Next, the defined genomic regions between every two consecutive loop anchor were extended by 10% from the left and right most boundaries, respectively. Then, each of these extended genomic segments was split into 120 equal-sized bins, and the

histone modification ChIP-Seq, RNAPII ChIP-Seq and TSS densities in each bin were calculated. Finally, the densities across all bins with the same position index were averaged and plotted as a curve for each of the considered dataset (Figure 3B). The averaged density was normalized by the dataset sequencing depth and the window/bin size accordingly (i.e. FPKM). For TSS, the density at each position index was normalized by the total number of TSSs and the window/bin size.

All histone modification and RNAPII ChIP-Seq data used here were downloaded from ENCODE (Table S5). TSS definition was extracted from GENCODE (version 14) (Derrien et al., 2012).

Transcription activity and directionality related to convergent and tandem loops

Here, we describe the approach of using histone modification ChIP-Seq, transcription factor ChIP-Seq and TSS location data to characterize the CTCF interaction loop anchors as aggregate density plots, according to the directionality of the CTCF anchors. First, the individual CTCF interaction loop was separated into 3 categories according to the CTCF motif orientation (convergent, tandem right and tandem left) at the anchors. Next, for each head and tail anchor, the ±2 kb regions from the anchor centers are extracted and split into 100 bp windows. Then the histone modification ChIP-Seq, RNAPII ChIP-Seq and TSS densities in each 100 bp window are calculated and normalized by data size and bin size (i.e. FPKM). Finally, the densities across all 100 bp windows aligned at the same position relative to the anchor center are averaged and plotted as a curve for each of the considered dataset (Figures 3C and 3D). For the ChIP-Seq data, the densities are normalized according to the sequencing depth. Loops with paired motifs in tandem leftward and rightward orientations are combined for data plotting, the X-axis coordinates of tandem left data are flipped to fit the tandem right orientation as shown (Figures 3C and 3D). Within each motif orientation category, the head or tail anchor is considered only once if it is shared by more than one CTCF loop.

Gene positioning in CTCF-mediated chromatin loop structures

Using GENCODE-characterized genes (n=51,001) as a reference (Derrien et al., 2012), there are 45,892 of genes found associated within CCD structures in GM12878 cells. We used the ±5 kb window from the TSSs as the promoter regions of human genes. Close to one-fifth (n=8,664) of the 45,892 genes had promoters overlapped with CTCF interaction anchors, and the rest (n=37,228) are scattered within the CTCF-mediated loops.

CTCF interaction anchor-centric organization of transcription complexes

Here, we discuss the relationship between CTCF interaction anchors and the RNAPII interaction mediated gene complexes in GM12878. We first examined the relationship between CTCF interaction anchors and transcriptional regulatory elements (promoters and enhancers). The CTCF anchors considered here are the 21,777 non-redundant anchors from which the 42,297 cohesin supported CTCF loops originated. By using the GENCODE TSS definition, we used the ±5 kb window from the TSSs as the promoter regions of human genes. Then, a gene promoter is considered as active if connected by

a RNAPII interaction anchor or overlapped with an RNAPII-binding peak. The GM12878 uniform H3K27ac and H3K4me1 ChIP-Seq peaks were downloaded from ENCODE (Table S5), and the consensus regions of the H3K27ac and H3K4me1 peaks were used as enhancers. For the 21,777 CTCF anchors, 6,259 directly overlapped with gene promoters in 1,801 CCDs, of which 3,599 overlapped with active gene promoters. In addition, 4,944 CTCF anchors directly overlapped with enhancers in 1,329 CCDs, and the rest (10,574) of CTCF anchors were not overlapping with promoters or enhancers. We also tried to examine the relationship between the RNAPII defined gene complexes with CTCF anchors. In GM12878, 1,141 SG and 1,592 MG gene complexes were identified as previously described (Table S2). Most of the genes associated with RNAPII-binding are located in MG complexes rather than SG (11,723 vs. 1,350). Then we superimposed the CTCF defined chromatin structures on top of the SG and MG complexes, to test whether the genes located in the CCDs were associated with the CTCF anchors (Table S2). For the genes in MG complexes, 10,789 had the promoters either directly overlapped with CTCF anchors or indirectly tethered to the distal CTCF anchors by a CTCF anchor overlapping promoter and/or enhancer through RNAPII mediated interaction. Similarly, 544 genes in the SG complexes were either directly (226) or indirectly (318) connected to the CTCF anchors. Collectively, these results further proved transcriptional regulatory elements (promoters and enhancers) are enriched at CTCF interaction anchors, and the RNAPII mediated interaction complexes are very likely to have the associated gene promoters docked at the CTCF interaction anchors presumably for stabilization purpose.

Gene expression breadth analysis of CTCF anchor-genes and CTCF loop-genes

The normalized RNA-Seq data for 56 human tissues was downloaded from EBI expression atlas (http://www.ebi.ac.uk/gxa/experiments/E-MTAB-3358). In each tissue, genes with FPKM value more than 1 were considered as expressed in corresponding tissue. Gene expression breadth is defined as the number of tissues a gene is expressed in. The p-value is calculated using the nonparametric Kolmogorov-Smirnov test.

Consistent results were observed in the three tested cell lines (Figures 3F S4E and S4F), suggesting that CTCF/cohesin-defined chromatin structures are highly conserved during cell lineage differentiation, and further indicating that CTCF/cohesin-interaction anchors are selectively enriched for constitutively expressed genes.

GM12878 ChIA-PET data phasing analysis

a. Individual PET phasing

The GM12878 cell line is derived from the 1000 Genome Project individual NA12878, whose genome-wide SNP phasing information is available (ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/2.8/hg19/). By using this SNP phasing information (excluding small indels) for GM12878, we assign the CTCF and RNAPII ChIA-PET reads with MAPQ ≥ 30 to the maternal or paternal haplotype depending on the nucleotide overlapping with the phased SNPs. If the sequenced nucleotide in a read at the phased SNP position is the same as the maternal allele, then this read is assigned to the maternal haplotype. Similarly, if the sequenced nucleotide in a read at the phased SNP position is the same as the paternal allele, then this read is assigned to the paternal haplotype. This analysis is done independently for the two ends of all inter-ligation PETs.

We denote the haplotype of individual end of PET sequences as M (maternal), P (paternal) and N (not determined). Therefore, a PET would have the possible haplotype at the two ends as M-M, P-P, M-N, P-N, M-P and N-N. By using the intra-chromosomal phased inter-ligation PETs, we calculated the fractions of cross-allele (M-P) PETs over all phased PETs as a function of genomic span. The results suggest the cross-allele frequency is well controlled within 2.5 Mb distance (Figure S5A). Furthermore, we also compared the genomic span of phased PETs in all categories, and the results show that most of M-P PETs are of strikingly large genomic span (≥2.5 Mb) for both CTCF and RNAPII data (Figure S5B). Based on these observations, we conclude that intra-chromosomal interactions within 2.5 Mb should all take place within a homolog chromosome. These results confirmed the concept of chromosome territory established decades ago by DNA-FISH experiments (Cremer and Cremer, 2001), and suggested that ChIA-PET mapping data is accurate in mapping haplotype chromatin interactions. Also, based on this notion, we identified 2.24 million phased PETs with the haplotype at the two ends as M-M, P-P, M-N and P-N from GM12878 CTCF and RNAPII ChIA-PET datasets. Similar to non-phased PETs (i.e., N-N), the majority (maternal 89%, paternal 86%) of phased PETs were singletons, which are useful for studying higher-order haplotype-specific chromosomal folding.

b. Interaction anchor, loop phasing and haplotype assignment

Based on the observations and conclusion mentioned above, we extended the phasing interactions to the loops (i.e. PET clusters) with at least one anchor overlapping with heterozygous phased SNP. Before this was done, we first identified the haplotype specificity of all interaction anchors. The haplotype assignment of the anchors were done as follows:

1) Identify all phased SNPs with biased protein factor binding coverage. The maternal and paternal allele frequencies of individual phased SNPs were computed and tested for allele bias by using Binomial test. The SNPs with Benjamini-Hochberg adjusted P values (i.e. FDR) ≤ 0.1 were considered as SNPs with significantly biased protein factor binding coverage.
2) Haplotype specificity of the interaction anchors overlapping with biased SNPs was assigned to the allele according to the direction of the biased SNP. If an anchor overlaps with multiple biased SNPs and the bias directions of these biased SNPs are consistent, then the haplotype specificity of this anchor is assigned accordingly; otherwise, the haplotype of this anchor is unassigned. In addition, if an anchor overlaps with multiple phased SNPs and the SNP with the highest binding coverage showed no allelic bias, then the haplotype bias of such anchor is also undetermined.

The above procedures were practiced in the CTCF and RNAPII ChIA-PET datasets independently.

Next, the interaction loops with at least one end originating from the haplotype-biased anchors were assigned to the allele according to the bias direction of the anchors (Figure 4B). For CTCF interaction loops, 36 were identified with both ends originating from haplotype-biased CTCF anchors; such interactions are referred as *Phased* interactions. In addition, 1,692 CTCF interactions were identified with one end originating from haplotype-biased CTCF anchors; such interactions are referred as *Extended Phased* Interactions. All interaction with genomic span greater than 2.5 Mb were excluded except for the *DXZ4-FIRRE-G6PD* region. Despite spanning ultra-long genomic distance, the haplotype-biased CTCF interactions connecting *DXZ4-FIRRE* (16

Mb) and *FIRRE-G6PD* (23 Mb) were retained for analysis, as strong and repetitively occurring CTCF interactions were detected at these loci.

Similarly, 78 and 1,244 *Phased* and *Extended Phased* RNAPII interactions were also identified, respectively. The *Phased* and *Extended Phased* interactions are all considered as haplotype-biased interactions after assigning their haplotype specificity according to the bias direction of the corresponding anchors. The detailed haplotype-breakdown of the haplotype-biased interactions is summarized in Figure 4C. In addition, for the RNAPII haplotype-specific interactions, the potential target genes were also identified if the gene's promoter overlapped with anchors of the allele specific RNAPII loops (Figure 4C).

The haplotype-specific anchors and interactions appeared to be relatively well balanced between the two factors, as well as between the two haplotypes in autosomes, except the RNAPII-associated anchors and interactions on ChrX due to the imprinting of the paternal homolog (Rozowsky et al., 2011). Consequently, genes associated with haplotype-biased RNAPII interactions identified in autosomes were in equal numbers, whereas ChrX-specific haplotype-biased genes were largely maternal (n=123).

Evaluating the SNP disruption effect in CTCF-binding motifs

In GM12878, 70 CTCF motifs were identified directly overlapping with phased heterozygous SNPs and are also located within phased CTCF anchors with haplotype bias. We extracted the motif nucleotide sequences from both the paternal and maternal alleles at these 70 motif locations. Each of these sequence pair from the two parental alleles only differs by one nucleotide due to the SNP variation. The SNP variation could be located at any position within the CTCF motif. As we already know the haplotype-bias direction at these 70 loci, we collected the sequences from the alleles retaining CTCF interactions into one group (i.e. *with interaction* group), and the sequences from the opposite allele into another group (i.e. *without interaction* group). Subsequently, we used Weblogo package (http://weblogo.berkeley.edu/logo.cgi) to construct the consensus motifs from these two groups, and the results showed the motif derived from the *with interaction* group had a more conserved G nucleotide at the 14^{th} position than the *without interaction* group. The motif logos reflecting the nucleotide frequency at each position are shown in Figure 5D. We systematically assessed the disease association of disrupted CTCF-mediated interactions by extensively examining the linkage disequilibrium (LD) between the 70 SNPs residing in CTCF core motif (i.e. CTCF-SNPs) and GWAS identified disease associated SNPs (EBI GWAS catalog: https://www.ebi.ac.uk/gwas/) in 11 populations available from "1000 genomes" (http://www.1000genomes.org; Phase 3 data). The result shows 32 of the 70 CTCF-SNPs were documented in dbSNP, and 8 showed LD (D'>0.5 and LOD>3) (Slatkin, 2008) with disease associated SNPs in the tested populations (Data S1, IV; Table S4)

GM12878 RNA-Seq data phasing analysis

GM12878 RNA-Seq sequence alignment data from ENCODE was downloaded and applied to the ASECounter pipeline within the GATK software package for allele-specific expression analysis (McKenna et al., 2010). The analysis was done by using the default parameters. The phased SNPs with RNA-Seq coverage of at least 10 were tested by Binomial test for allele bias. Only SNPs with FDR ≤ 0.1 were considered as biased SNPs and retained for further analysis. Within the target genes by haplotype-specific RNAPII

interactions, the ones with at least one SNPs having haplotype bias in RNA-Seq data inside the gene body is considered as allele-specifically expressed gene.

Among the 705 genes associated with allelic-biased RNAPII interactions (Figure 4C), 482 possess phased SNP, thus, are testable for allele-specific gene expression. Although 393 of the tested genes showed indistinguishable expression between the two alleles mostly due to low-level expression, we identified 89 genes that were transcribed in an allele-specific manner (Figure 6C). The vast majority (n=79, 89%) of them displayed the same haplotype-specificity in transcription as the chromatin interactions haplotype-specificity mediated by RNAPII (Figure 6D, Table S5).

GM12878 TF ChIP-Seq data phasing analysis

GM12878 TF ChIP-Seq sequence alignment data listed in factorbook (http://www.factorbook.org/) was downloaded from ENCODE and passed to a custom python script (utilizing the pysam library v0.8.1) to calculate the GM12878 phased SNP base coverage. For each TF ChIP-Seq dataset, the phased SNPs with read count coverage of at least 10 were further tested for significance of allele bias using Binomial test. The threshold of haplotype bias for each test SNP is $P < 0.05$. The replicates for each TF were merged together for analysis.

3D genome structure modeling and Visualization

The combined RNAPII and CTCF ChIA-PET data is used for 3D modeling through entire nucleome to individual chromosome at multi-scale resolutions. For phasing modeling of chromosome X, the unphased singletons and clusters on chromosome X are equally assigned to maternal and paternal copies, respectively. The unphased singletons and clusters combined with phased singletons and clusters are used for individual chromosome X modeling at different resolutions. Detailed information on the 3D modeling will be in a separate publication. The key steps and computational algorithms are briefly described as follow.

a. Hierarchical tree represents nucleome structures at different resolution levels

To take the advantage of multi-scale features of ChIA-PET data, the nucleome was represented as a tree structure with different levels corresponding to different resolutions. At the top level, a root node represents the entire nucleome. This root node consists of 23 nodes to represent 23 chromosomes for human genome (chromosome Y is excluded). Under each chromosome node there are three additional levels to represent the chromosome with a series of beads connected by springs at increasing resolutions. A parent-child relationship exists between nodes at consecutive levels, where lower resolution nodes are defined-by and contain a distinct subset of higher resolution nodes.

We used a conventional beads-on-a-string polymer model to represent individual chromosomes, where each beads denoted a specific genomic region with well-defined start and end genomic coordinates. According to the ChIA-PET data feature, we used the following terminologies to describe the beads on chromosome at different levels. At segment level, the beads denote the segments of chromosome with approximately 2 Mb span. At anchor level, the beads denote the interaction anchors identified by ChIA-PET.

At sub-anchor level, the beads represent the regions between two anchors of ChIA-PET interaction clusters.

b. Chromosome segmentations according to ChIA-PET interaction blocks

As described above, the segments are with approximately 2 Mb span. In details, each chromosome is not simply split into uniform sizes, but is segmented based on the ChIA-PET interaction blocks. We consider such chromosome partitioning approach is more consistent with the ChIA-PET interaction block and closely reflect the topologically association domain in the nucleus. We used a bottom-up iteratively clustering algorithm to define segments on chromosome. Initially, all interaction blocks are treated as separate segments. At each step, the algorithm will merge two neighboring segments into an aggregated segment based on their genomic size and the genomic distance between them. The algorithm will stop to merge segment when the aggregated segment attaining to the predefined size.

c. Simulated annealing for structure reconstruction

The general procedure for generating a structure is the same for each level. Energy is defined as a function of the node positions, and Monte Carlo simulated annealing is used to find a structure that minimizes the energy. In general the energy can be written as

$$E(\{\vec{r}_i\}) = \alpha E_{polymer}(\{\vec{r}_i\}) + \beta E_{data}(\{r_{ij}\},\{d_{ij}\}),$$

where the first term $E_{polymer}(\{\vec{r}_i\})$ includes standard polymer interactions such as stretching and bending energies, and the second term $E_{data}(\{r_{ij}\},\{d_{ij}\})$ includes all additional interactions imposed by the experimental data. The experimental data is used to define a preferred distance, d_{ij}, between each interacting pair of nodes i and j, and the energy is a function of these preferred distances and the actual distances, r_{ij}, between each pair of nodes. The exact energy function and the method of calculating preferred distances are different for each level. At low resolutions we use singleton data to generate contact frequency heatmaps and convert the contact frequencies to preferred distances. At high-resolution levels we convert PET interaction frequencies to a preferred distance between pairs of interacting anchors. We work in a top-down approach, where lower resolution structures are constructed first and are used to inform the structures at higher resolutions.

At each resolution, the Monte Carlo simulated annealing proceeds in the conventional fashion. At each step a random bead is chosen and shifted by a vector drawn at random from a sphere of a specified radius. The new energy is calculated and the move is accepted if $E_{new} \leq E_{old}$. If $E_{new} > E_{old}$, then the move is accepted with probability $p = \exp\left(-\frac{1}{T}\frac{E_{new}}{E_{old}}\right)$, where T is analogous to the temperature. The "temperature" is initialized to $T_{init} > 0$, and is reduced after each step, $T_{new} = \kappa T_{old}$, for some $\kappa < 1$. The simulation is checked every $N_{milestone}$ steps, and the simulation is stopped when the energy decrease since the last milestone is below a user defined threshold.

d. Modeling constraints at sub-anchor level

At the sub-anchor level we consider several factors contributing to the energy in order to properly generate the loop structures between interaction anchors.

Because all polymer models predict a power law relationship between arc length and physical size, we firstly use $d_{i,i+1} = N_{i,i+1}^{\alpha}$ to ensure the physical size of a loop scales

with its genomic span, where $N_{i,i+1}$is the genomic distance between sub-anchors i and $i+1$. These preferred distances contribute a term $E_{dist} = \sum_i \left(r_{i,i+1} - d_{i,i+1}\right)^2$ to the total energy. We then include a bending energy, which prevents excessive curvature. This energy is defined as $E_{bend} = \frac{1}{2}\sum_i \left(1 - \hat{v}_{i-1,i} \cdot \hat{v}_{i,i+1}\right)$, where $\hat{v}_{i,i+1}$ is the unit vector pointing from sub-anchor i to sub-anchor $i+1$.

These energy terms can be used to model smooth, circular loops passing through the fixed anchor beads, but they do not account for singleton interactions between sub-anchors in the loops. To determine the effect of the singletons on loop shapes we build two heatmaps. At first, we use the singletons in the interaction block to construct a sub-anchor heatmap. This heatmap is not directly used to compute preferred distances because it contains many null entries, which are simply a consequence of the sparseness of interaction data at extremely high resolutions. To impute these missing values in the sub-anchor heatmap, we construct several structures using just the distance and bending energies as described above (Figure S7A). For each structure we construct a heatmap using the distance between each pair of loci, and then these heatmaps are averaged to produce a consensus distance heatmap. Each entry in the distance map is then decreased in proportion to the corresponding entry in the singleton sub-anchor heatmap, and these reduced distances are used to define the third energy term, $E_{heat} = \sum_{ij} \left(r_{ij} - d_{ij}\right)^2$. The examples showing the sub-anchor heatmaps affecting loop shapes are presented in Figure S7B.

We also introduce the CTCF motif orientations into our modeling algorithm. The CTCF motif orientations in 3D structure are defined as a unit vector tangent to the chromatin curve at the locations of the motifs. The vector points either “along” the fiber (from the 5’ to 3’ direction) or in the opposite direction, depending on the motif strand orientations. We assume a pair of interacting anchors with CTCF motifs will preferentially align with their tangent vectors pointing in the same spatial direction (Figure S7C). To account for this interaction we include a fourth energy term based on the orientation of interacting anchors, $E_{orn} = \sum_{(i,j)\in P} \left(1 - \hat{o}_i \cdot \hat{o}_j\right)$, where $\hat{o}_i$ is the orientation of the anchor i and P is a set of pairs of interactions in the current interaction block (Figure S7C).

Combining these terms we arrive at $E_{subanchor} = w_{dist}E_{dist} + w_{bend}E_{bend} + w_{orn}E_{orn} + w_{heat}E_{heat}$, where w_{dist}, w_{bend}, w_{orn}, and w_{heat} are weights assigned to particular energy terms. A schematic representation of the models with different energy terms is shown in Figure S7C.

e. 3D model computational visualization

Chromosome structures were visualized using either a custom desktop program written in C++ using OpenGL (https://www.opengl.org) and the Qt library (http://www.qt.io/developers/), or a custom web application written in Javascript using WebGL (https://www.khronos.org/webgl/) and the Three.js library (http://threejs.org). All visualization and simulation software is freely available through the links provided in (Szalaj et al., In preparation).

Chromosome structure dynamics revealed by 3D DNA-FISH

To gain further specificity at the level of individual chromosomes, we studied chromosome 1 at different resolutions, and observed a putative structural framework,

whereby its two chromosomal arms bend and extend in the same direction (Figure 7A). Since our mapping data was generated from millions of cells, the predicted 3D model was likely to reflect an average representative of ensemble structures. Therefore, it is possible that the chromosome 1 ensemble structure is plastic and fluidly adopting positions between “open” and “closed” conformations (Figure 7B). To validate this, we analyzed the chromosome territory by 3D DNA-FISH. For orientation, we included two positioning probes (green and red) at sub-telomeric regions. Among the 50 nuclei examined, we observed a broad spectrum of arm conformations, ranging from widely open to completely closed. In one confocal microscopy imaged nucleus, two representative configurations (i.e., “open” and “closed”) were observed (Figure 7C), lending evidence for dynamic conformation states for chromosome 1 arm structures.

SIM super-resolution microscopy and FRET-FLIM analysis for CTCF and RNAPII nuclear co-localization

As shown in Figure 7E, CTCF forms distinctive foci in nuclei. Similarly, RNAPII also displayed distinctive foci in the nuclei, and resemble the previous standard views of transcription factory (Cook, 2010). Superimposed image showed that CTCF and RNAPII foci are extensively co-localized in the nucleus at the sub-diffraction level. Statistical co-localization analysis showed high correlation between the CTCF and RNAPII signals comparing to random control (Figure 7E right). We also performed Förster resonance energy transfer (FRET)–based assay using fluorescence lifetime imaging microscopy (FLIM) to verify the proximal events of RNAPII and CTCF in nuclei. This technique permits determination of the spatial proximity between donor and acceptor molecules within several nanometers, a distance sufficiently close for molecular interactions to occur (Chen et al., 2003). In this assay, the occurrence of FRET (i.e., co-localization) between the donor and acceptor is revealed by shortening of the donor fluorescence lifetime. As shown in the color coded maps of fluorescence lifetimes (Figure 7F), the mean fluorescence lifetime of the donor fluorophore (Alexa488) labeled CTCF protein occurred to be considerably shorter in the presence of acceptor (Cy3) labeled RNAPII than the control. The quantitative analysis of fluorescence lifetime distribution calculated over entire nuclei further confirms a shift towards shorter values (P=0.047, t-test) of fluorescence lifetime and was found for 45% of the nuclei area (as indicated by non-overlapping region in lifetime distribution).

Supplemental references

Bolte, S., and Cordelieres, F.P. (2006). A guided tour into subcellular colocalization analysis in light microscopy. J Microsc *224*, 213-232.

Chen, Y., Mills, J.D., and Periasamy, A. (2003). Protein localization in living cells and tissues using FRET and FLIM. Differentiation *71*, 528-541.

Cook, P.R. (2010). A model for all genomes: the role of transcription factories. J Mol Biol *395*, 1-10.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet *2*, 292-301.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G.*, et al.* (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res *22*, 1775-1789.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat Methods *9*, 215-216.

Goh, Y., Fullwood, M.J., Poh, H.M., Peh, S.Q., Ong, C.T., Zhang, J., Ruan, X., and Ruan, Y. (2012). Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation. J Vis Exp *62*.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589-595.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297-1303.

Neher, R., and Neher, E. (2004). Optimizing imaging parameters for the separation of multiple labels in a fluorescence image. J Microsc *213*, 46-62.

Neher, R.A., Mitkovski, M., Kirchhoff, F., Neher, E., Theis, F.J., and Zeug, A. (2009). Blind source separation techniques for the decomposition of multiply labeled fluorescence images. Biophys J *96*, 3791-3800.

Schones, D.E., Smith, A.D., and Zhang, M.Q. (2007). Statistical significance of cis-regulatory modules. BMC Bioinformatics *8*, 19.

Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Reviews Genet *9*, 477-485.

Walczak, A., Szczepankiewicz, A.A., Ruszczycki, B., Magalska, A., Zamlynska, K., Dzwonek, J., Wilczek, E., Zybura-Broda, K., Rylski, M., Malinowska, M.*, et al.* (2013). Novel higher-order epigenetic regulation of the Bdnf gene upon seizures. J Neurosci *33*, 2507-2511.

Wang, L., Chen, J., Wang, C., Uuskula-Reimand, L., Chen, K., Medina-Rivera, A., Young, E.J., Zimmermann, M.T., Yan, H., Sun, Z.*, et al.* (2014). MACE: model based analysis of ChIP-exo. Nucleic Acids Res *42*, e156.

Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z.*, et al.* (2006). A global map of p53 transcription-factor binding sites in the human genome. Cell *124*, 207-219.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Table S1. Summary of ChIA-PET libraries, Related to Figure 1

Cell line	Factor	Method[a]	Replicate	Uniquely mapped PETs	Self-ligation PETs	Singleton PET	Clustered PETs	PET clusters	PET count cutoff	Clustered Reads	PET clusters	PET count cutoff	Data source
GM12878	CTCF	ChIA-PET v2	Combined	51,103,494	13,293,289	35,052,570	2,757,635	581,978	2	1,721,227	93,409	4	This study
		ChIA-PET v2	Replicate 1	20,295,087	5,200,045	14,283,578	811,464	173,979	2	569,927	39,747	4	
		ChIA-PET v2	Replicate 2	30,808,407	8,093,244	21,341,436	1,373,727	363,930	2	802,951	53,455	4	
Hela	CTCF	ChIA-PET v2	Replicate 1	18,949,457	2,010,987	15,564,931	1,373,539	572,660	2	338,161	54,971	3	This study
K562	CTCF	ChIA-PET v1	Replicate 1	5,628,606	2,096,114	3,450,038	82,454	22,595	2				ENCODE
MCF7	CTCF	ChIA-PET v1	Replicate 1	18,216,265	4,396,328	13,754,144	65,793	20,145	2				ENCODE
		ChIA-PET v1	Replicate 2	28,743,594	6,150,290	22,388,155	205,149	50,506	2				
GM12878	RNAPII	ChIA-PET v2	Combined	124,338,317	39,112,010	53,409,826	31,816,481	15,352,183	2	885,964	113,591	4	This study
		ChIA-PET v2	Replicate 1	27,898,521	10,526,704	16,306,082	1,065,735	496,320	2	59,482	8,640	4	
		ChIA-PET v2	Replicate 2	48,441,329	15,192,808	32,497,129	751,392	313,565	2	137,178	16,316	4	
		ChIA-PET v2	Replicate 3	47,998,467	13,433,133	33,352,078	1,213,256	545,123	2	134,097	16,166	4	
Hela	RNAPII	ChIA-PET v2	Replicate 1	34,053,675	19,440,372	13,896,436	716,867	347,751	2				This study
K562	RNAPII	ChIA-PET v1	Replicate 1	27,905,188	8,239,985	19,126,043	539,160	153,848	2				Li et al., 2012
MCF7 (control)	RNAPII	ChIA-PET v1	Replicate 1	28,329,512	7,911,694	20,233,585	184,233	55,811	2				Li et al., 2012
MCF7 (E2 treatment)	RNAPII	ChIA-PET v1	Replicate 1	29,169,402	8,246,466	20,614,421	308,515	87,110	2				Li et al., 2012
MCF7 (control)	ER	ChIA-PET v1	Replicate 1	8,013,160	1,525,713	6,486,174	1,273	431	2				Fullwood et al., 2009
MCF7 (E2 treatment)	ER	ChIA-PET v1	Replicate 1	1,850,383	558,588	1,272,340	19,455	3,605	2				Fullwood et al., 2009

[a]: ChIA-PET v1 and v2 indicate the short- and long-reads ChIA-PET methods, respectively.

Table S2. Summary of CTCF and RNAPII interactions and defined complex structures, Related to Figure 2 and 3

A. Summary of CTCF defined chromatin structures

	GM12878	HeLa	K562	MCF7
CTCF loops	80,157	40,312	21,133	49,772
CTCF loops with CTCF & Cohesin peaks	42,297	24,502	9,693	16,748
CTCF loop domains (CCDs)	2,267	3,071	2,385	3,317
CTCF domain covered genome (Mb)	2,224	1,577	883	1,064

B. Summary of RNAPII mediated interaction structures

	GM12878	HeLa	K562	MCF7
RNAPII loops	73,349	70,080	107,008	79,934
Refined RNAPII loops	50,457	13,655	64,572	50,212
RNAPII single-gene complexes	1,141	1,248	1,414	1,390
RNAPII multi-gene complexes	1,592	852	1,463	1,847

C. Summary of CTCF anchor centric RNAPII interaction complexes and involved genes

Category of RNAPII interaction complex	**RNAPII chromatin complex**		**Genes involved**					
	Number	Connected to CTCF anchors	Total genes	CTCF-anchor genes	CTCF-loop genes connected to CTCF anchor genes	CTCF-loop genes connected to CTCF anchor enhancer	CTCF-loop genes connected to CTCF anchor gene/enhancer	CTCF-loop genes not connected with CTCF anchors
Multi-gene	**1,592**	**1,227**	**11,723**	**3,558**	**500**	**952**	**5,779**	**934**
Single-gene	**1,141**	**524**	**1,350**	**226**	**N/A**	**318**	**N/A**	**806**
Basal-promoter	4,009	1,014	4,587	1,176	N/A	N/A	N/A	3,411

Table S3. Summary of chromatin loops according to CTCF motif orientation, Related to Figure 2

	Convergent	**Tandem Left**	**Tandem Right**	**Divergent**	**Total**
Expected in random	25%	25%	25%	25%	
GM12878 CTCF ChIA-PET data Observed *	22,709 (64.5%)	5,844 (16.6%)	5,830 (16.5%)	847 (2.4%)	35,230
GM12878 CTCF ChIA-PET data Rewired (CTCF motif pairs in CCD)#	13,219±89.1 (38.5%)	7,933.0±77.7 (23.1%)	8,335.7±79.2 (24.2%)	4,886.8±64.6 (14.2%)	34,375.4±64.1
in situ Hi-C (Rao et al., 2014)	2,574 (90.1%)	137 (4.8%)	136 (4.8%)	10 (0.3%)	2,857
in situ Hi-C loops recapitulated by CTCF ChIA-PET	2,533	119	120	9	2,781

*: $P < 0.0001$ compared to Rewired, Chi-square test
#: 1000 iterations

Table S4. Summary of GWAS SNPs that are associated with CTCF SNPs, Related to Figure 5

Number of Populations	CTCF SNPs	GWAS SNPs	Traits 1	Risk allele 1	Traits 2	Risk allele 2	Traits 3	Risk allele 3
8	rs1976938	rs10876432	Bone mineral density (spine)	A				
3	rs10821010	rs10992471	Visceral adipose tissue/subcutaneous adipose tissue ratio	G				
11	rs12936231	rs11078927	Asthma	?				
2	rs12741252	rs11119805	Stearic acid (18:0) plasma levels	A				
5	rs7146599	rs12436436	Bipolar disorder	C				
8	rs12936231	rs12946510	Inflammatory bowel disease	T				
4	rs41518444	rs13124827	Hypospadias	C				
1	rs7160073	rs17102423	besity-related traits	A				
1	rs12936231	rs17609240	Hematological parameters	G				
7	rs1976938	rs2016266	Bone mineral density	A	Bone mineral density (spine)	G		
4	rs1976938	rs2272306	Obesity-related traits	A				
11	rs12936231	rs2290400	Type 1 diabetes	?				
3	rs12936231	rs2302777	Multiple myeloma (hyperdiploidy)	A				
11	rs12936231	rs2305480	Asthma (childhood onset)	G	Ulcerative colitis	A		
11	rs12936231	rs2872507	Type 1 diabetes autoantibodies	A	Ulcerative colitis	A	Crohn's disease	A
6	rs12936231	rs3859192	White blood cell count	A				
10	rs12936231	rs3894194	Asthma (childhood onset)	A				
9	rs2256964	rs4302748	Platelet counts	?				
3	rs7160073	rs4466998	Mean corpuscular volume	A				
4	rs12936231	rs4794820	Asthma	?				
1	rs12936231	rs4794822	White blood cell count	T	White blood cell types	T	Neutrophil count	C
10	rs12936231	rs6503525	Asthma	C				
7	rs7160073	rs7148590	Blood trace element (Zn levels)	?				
4	rs7160073	rs7155454	Red blood cell traits	A				
8	rs12936231	rs7212938	Asthma and hay fever	G				
11	rs12936231	rs7216389	Asthma	T				
4	rs7146599	rs8010715	IgG glycosylation	C				
11	rs12936231	rs8067378	Cervical cancer	G	Ulcerative colitis	A		
11	rs12936231	rs8069176	Fractional exhaled nitric oxide (childhood)	A				
6	rs12936231	rs8078723	C-reactive protein and white blood cell count	C	White blood cell count	T		
9	rs12936231	rs907092	Primary biliary cirrhosis	A				
11	rs12936231	rs9303277	Systemic lupus erythematosus and Systemic sclerosis trols	?	Primary biliary cirrhosis	T		
4	rs12936231	rs9303280	Self-reported allergy	T				
5	rs2256964	rs9648428	Obesity-related traits	A				
4	rs10821010	rs9969804	Height	A				