HHS Public Access

Author manuscript

Cell. Author manuscript; available in PMC 2016 December 17.

Published in final edited form as:

Cell. 2015 December 17; 163(7): 1611–1627. doi:10.1016/j.cell.2015.11.024.

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 & 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 Cedex13, France [9]Centre National de la Recherche Scientifique and Muséum National d'Histoire Naturelle, 57 Rue Cuvier, 75231 Paris Cedex 05, France [10]Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA [11]College of Informatics, Huazhong Agricultural University, Wuhan, Hubei 430070, China

Summary

Spatial genome organization and its effect on transcription remains a fundamental question. We applied an advanced ChIA-PET strategy to comprehensively map higher-order chromosome folding and specific chromatin interactions mediated by CTCF and RNAPII with haplotype specificity and nucleotide resolution in different human cell lineages. We find that CTCF/cohesin-

*Correspondence: Yijun.Ruan@jax.org.

[12]Co-first authors

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 O.J.L performed data analysis and interpretation; M.P. and L.M.S. analyzed lampbrush chromosome; 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., G.L.

More details are available in Extended Experimental Procedures.

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 towards 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 meters 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 3-dimensional (3D) structures of the human genome include 3D-FISH nuclear imaging and 3D genome mapping. 3D-FISH (fluorescence *in situ* hybridization) 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 ChIA-PET and Hi-C (Fullwood et al., 2009; Lieberman-Aiden et al., 2009). ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag sequencing) was designed to detect genome-wide chromatin interactions mediated by specific protein factors, whereas Hi-C (High-throughput Chromosome Conformation Capture) 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 ChIP (chromatin immunoprecipitation), ChIA-PET is unique in detecting protein factor mediated chromatin interactions and is capable of generating high-resolution (~100bp) 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, CTCF (CCCTC-binding zinc finger protein) and RNAPII (RNA polymerase II) 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. Herein, 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 uncovered detailed chromatin topology that provide the framework for regulating transcriptional activity.

Results

I. 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 (2x150bp). 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 (Figure 1A, 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 reflects higher-order topological proximity, similar to Hi-C data (Figure S1B-E). Therefore, in theory, the multifaceted ChIA-PET data is 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 demonstrate 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; Extended 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.

II. 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 (Figure S2B-E; Extended 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-Seq data of subunits RAD21 and SMC3. The vast majority (99%) of the CTCF interactions had cohesion 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) (Figure 2A right; Figure S2F; Table S2; Extended 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 (Figure 2A, 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; Extended 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 (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/cohesion-mediated loops often interconnect and continuously cover large chromatin segments (Figure 2C). Based on connectivity and contact frequency (Figure S3A; Extended Experimental Procedures), we identified 2,267 CTCF-mediated chromatin contact domains (CCDs) in GM12878 cells (Figure 2D, S3B). Comparison to Hi-C data showed that CTCF ChIA-PET and non-selective Hi-C were highly concordant in detecting chromatin domain structures (Figure 2C-D, S3C-D), 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 (Figure 2E, S3E). Many CCD boundaries contained multiple CTCF-binding peaks with inward-facing motifs (Figure 2C inserts; Figure S3F), suggesting a possible double-knot-tie mechanics for tightening domain end structures (Figure S3G-I; Extended Experimental Procedure). In contrast, tandem loops were evenly distributed within the CCD space (Figure S3I), and showed high consistency in motif orientation (Figure 2F, 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 8 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.

III. 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-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 (Figure S4B-C; 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 (Extended Experimental Procedures). Unlike the loop regions, the anchors were enriched with active epigenomic markers, RNAPII occupancy and the presence of TSS (transcription start site) (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 towards 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 (Figure 3C-D, S4D), while there was no such notable difference for the anchors of convergent loops. It implies that the head anchor in 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) (Extended Experimental Procedures). We examined the anchor- and loop-genes based on their expression profiles in 56 different human tissues. Gene expression breadth analysis (Extended 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 (Figure 3F right; Figure S4E-F), 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; Extended 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 towards 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 Figure S4G-H.

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.

IV. 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 (Figure 4A, S5A-B; Extended 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 (Figure 4B-C; Extended Experimental Procedures).

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 (13Mb) 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* (23Mb) 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 (Figure 3C-E). 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 (Figure 5D, 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 (Extended 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 8 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 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 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 (209kb) with this CTCF-SNP (rs12936231) in CEU having high pairwise correlation (D'>0.7), and the haplotype comprising these 7 risk alleles was frequently (0.422) observed in the CEU population (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 6 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.

V. 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; Extended 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; Extended 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 (Figure 6C-D; 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 (300kb apart) were approximately 3-fold and 2.5-fold higher, respectively, than the matched maternal allele. In addition, the distal enhancer exhibited more than 3-fold higher paternal-biased binding by 3 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.

VI. 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 (Figure S7A-C) (Szalaj et al., In preparation), 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 (Figure S7D-E; Extended Experimental Procedures). To gain further specificity at the level of individual chromosomes, we modeled Chr1 at different resolutions, and observed a 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; Extended 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; Extended 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; Extended 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.

Lastly, to test if 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; Extended Experimental Procedure).

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 (which probably require more energy and are more stable once established) are engaged to define the CCD/TAD structures and boundaries, whereas the tandem loops (which 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 near future.

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

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 (2X150bp) by HiSeq2500 (Extended Experimental Procedures).

ChIA-PET data processing

Short-read ChIA-PET data was obtained from ENCODE (Table S6) 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. All data has been submitted to the NCBI GEO database (GEO accession: GSE72816).

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 analysed 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 63x 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

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 CZ 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. The HeLa genome sequence described/used in this research was derived from a HeLa cell line (http://www.ncbi.nlm.nih.gov/gap). This study was reviewed by the NIH HeLa Genome Data Access Working Group.

References

Bickmore WA. The spatial organization of the human genome. Annu Rev Genomics Hum Genet. 2013; 14:67–84. [PubMed: 23875797]

Boyle S, Rodesch MJ, Halvensleben HA, Jeddeloh JA, Bickmore WA. Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosome Res. 2011; 19:901–909. [PubMed: 22006037]

Cremer M, Grasser F, Lanctot C, Muller S, Neusser M, Zinner R, Solovei I, Cremer T. Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes. Methods Mol Biol. 2008; 463:205–239. [PubMed: 18951171]

Cullen KE, Kladde MP, Seyfred MA. Interaction between transcription regulatory regions of prolactin chromatin. Science. 1993; 261:203–206. [PubMed: 8327891]

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485:376–380. [PubMed: 22495300]

Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K, et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell. 2014; 159:374–387. [PubMed: 25303531]

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009; 462:58–64. [PubMed: 19890323]

Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y, et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell. 2015; 162:900–910. [PubMed: 26276636]

Hall MH, Magalska A, Malinowska M, Ruszczycki B, Czaban I, Patel S, Ambrozek-Latecka M, Zolocinska E, Broszkiewicz H, Parobczak K, et al. Localization and regulation of PML bodies in the adult mouse brain. Brain Struct Funct. 2015 DOI: 10.1007/s00429-015-1053-4.

He Q, Johnston J, Zeitlinger J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat Biotechnol. 2015; 33:395–401. [PubMed: 25751057]

Horakova AH, Moseley SC, McLaughlin CR, Tremblay DC, Chadwick BP. The macrosatellite DXZ4 mediates CTCF-dependent long-range intrachromosomal interactions on the human inactive X chromosome. Hum Mol Genet. 2012; 21:4367–4377. [PubMed: 22791747]

Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol. 2012; 30:90–98. [PubMed: 22198700]

Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen CA, Lin S, Lin Y, Qiu Y, et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature. 2015; 518:350–354. [PubMed: 25693566]

Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, Ariyaratne PN, Mohamed YB, Ooi HS, Tennakoon C, et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol. 2010; 11:R22. [PubMed: 20181287]

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012; 148:84–98. [PubMed: 22265404]

Liang Z, Zickler D, Prentiss M, Chang FS, Witz G, Maeshima K, Kleckner N. Chromosomes Progress to Metaphase in Multiple Discrete Steps via Global Compaction/Expansion Cycles. Cell. 2015; 161:1124–1137. [PubMed: 26000485]

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–293. [PubMed: 19815776]

McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. Science. 2010; 328:235–239. [PubMed: 20299549]

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–777. [PubMed: 20220756]

Morgan GT. Lampbrush chromosomes and associated bodies: new insights into principles of nuclear structure and function. Chromosome Res. 2002; 10:177–200. [PubMed: 12067208]

Nativio R, Sparago A, Ito Y, Weksberg R, Riccio A, Murrell A. Disruption of genomic neighbourhood at the imprinted IGF2-H19 locus in Beckwith-Wiedemann syndrome and Silver-Russell syndrome. Hum Mol Genet. 2011; 20:1363–1374. [PubMed: 21282187]

Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nat Rev Genet. 2014; 15:234–246. [PubMed: 24614316]

Penrad-Mobayed M, Kanhoush R, Perrin C. Tips and tricks for preparing lampbrush chromosome spreads from Xenopus tropicalis oocytes. Methods. 2010; 51:37–44. [PubMed: 20085818]

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159:1665–1680. [PubMed: 25497547]

Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011; 147:1408–1419. [PubMed: 22153082]

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

Selvaraj S, J RD, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat Biotechnol. 2013; 31:1111–1118. [PubMed: 24185094]

Sims RJ 3rd, Mandal SS, Reinberg D. Recent highlights of RNA-polymerase-II-mediated transcription. Curr Opin Cell Biol. 2004; 16:263–271. [PubMed: 15145350]

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, et al. Population genomics of human gene expression. Nat Genet. 2007; 39:1217–1224. [PubMed: 17873874]

Szalaj P, Tang Z, Michalski P, Pietal M, Luo OJ, Ruan Y, Plewczynski D. 3D-NOME: an integrated 3-Dimensional NucleOme Modeling Engine for data-driven simulation of spatial genome organization. In preparation.

Verlaan DJ, Berlivet S, Hunninghake GM, Madore AM, Lariviere M, Moussette S, Grundberg E, Kwan T, Ouimet M, Ge B, et al. Allele-specific chromatin remodeling in the ZPBP2/GSDMB/

ORMDL3 locus associated with the risk of asthma and autoimmune disease. Am J Hum Genet. 2009; 85:377–393. [PubMed: 19732864]

Zhang Y, Wong CH, Birnbaum RY, Li G, Favaro R, Ngan CY, Lim J, Tai E, Poh HM, Wong E, et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature. 2013; 504:306–310. [PubMed: 24213634]

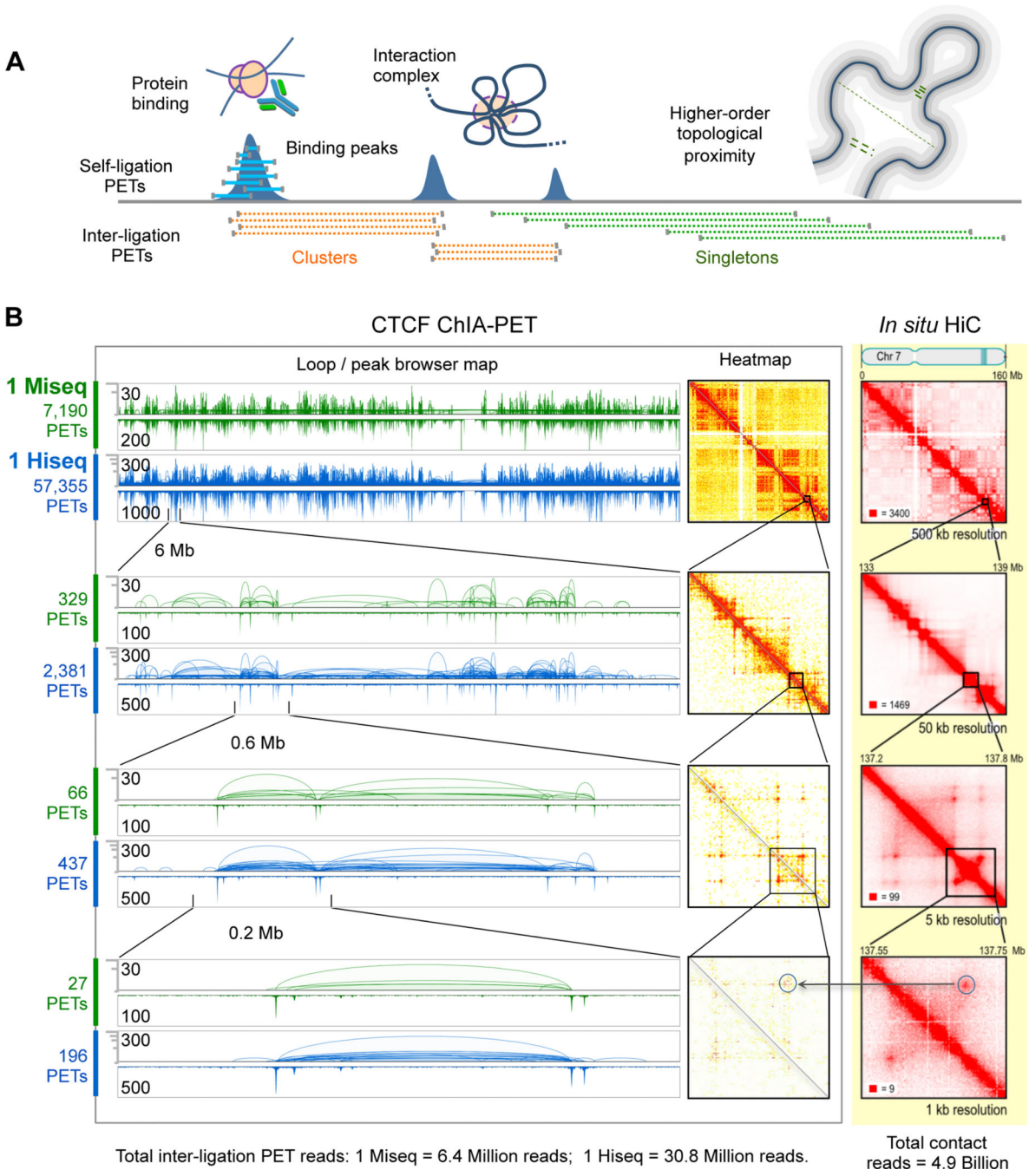


Figure 1. Characteristics of ChIA-PET data for 3D genome mapping

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 & 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.

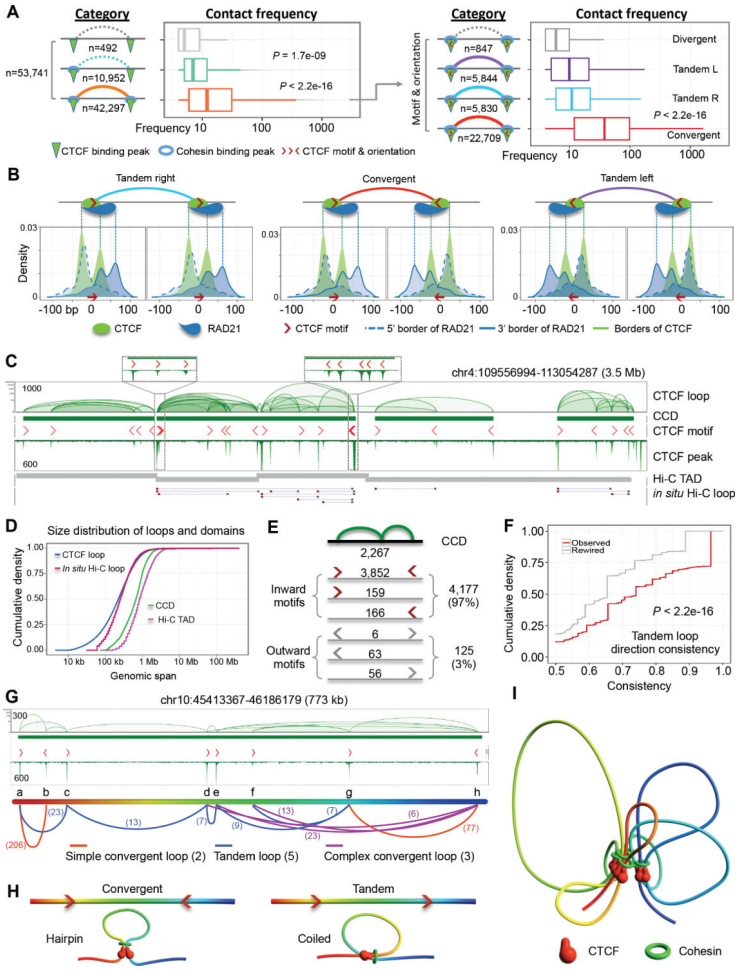


Figure 2. CTCF-defined chromatin looping topology

A. Characterization of CTCF-mediated loops in relation to cohesin binding and CTCF-motif orientation. See also Figure S2E-G.

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. Insert: 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 Figure S3C-D.

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 (grey) 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 2 simple-convergent (not cross other loop anchor, red), 5 tandem (blue) and 3 complex-convergent (cross other loop anchor, purple) loops from 8 anchors (a-h). The motifs in the 5 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 Extended Experimental Procedures.

See also Figure S2 and S3.

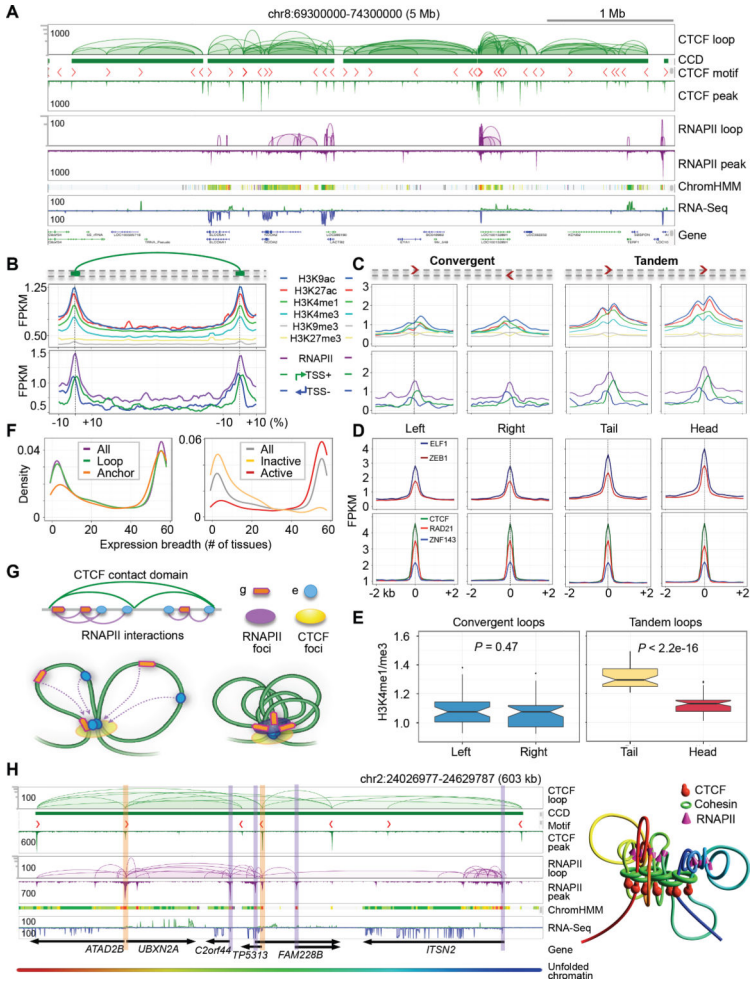


Figure 3. Relationship between CTCF/cohesin-mediated chromatin structure and RNAPII-associated transcriptional function

A. Browser view of a 5Mb genomic segment with 4 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 Extended 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 (±2kb). 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.

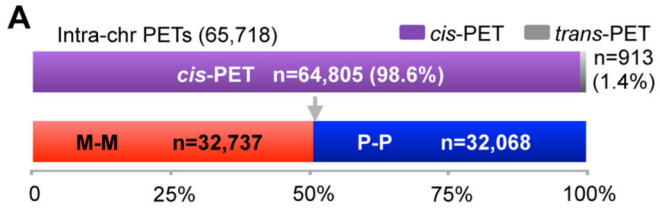
F. Expression breadth (number of tissues a gene is expressed in) of CTCF anchor-genes and loop-genes in GM12878 (left). Anchor-genes (yellow) are significantly less represented as

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 (grey) 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 towards 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 and Extended Experimental Procedures.



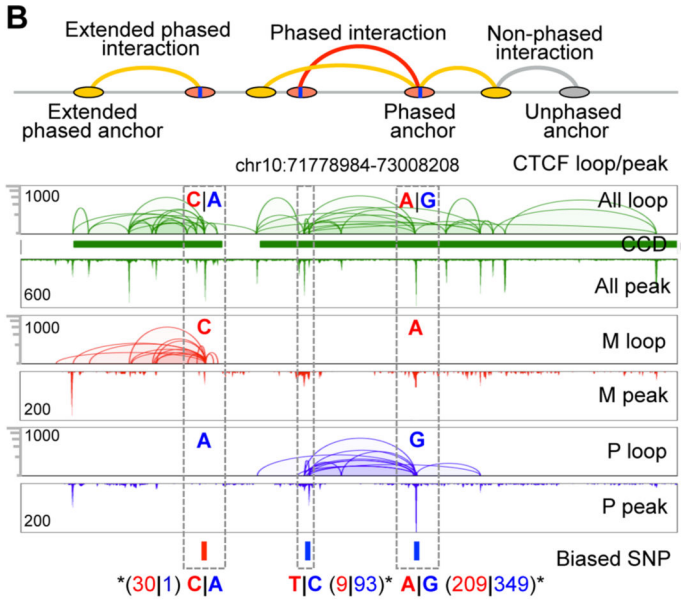
C

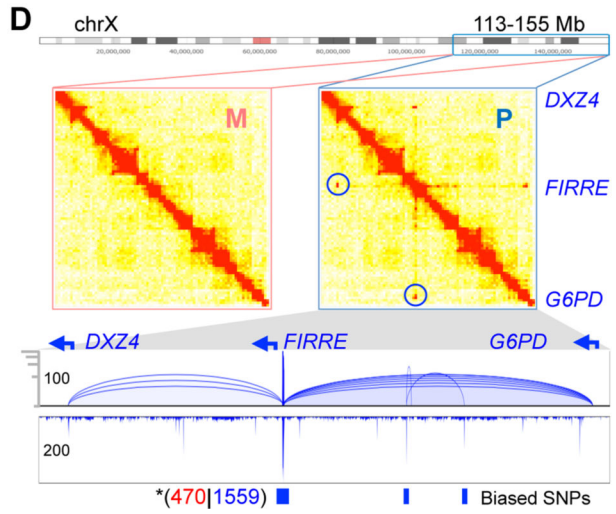
Haplotype-biased anchors

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

Haplotype-biased interactions and affected genes

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





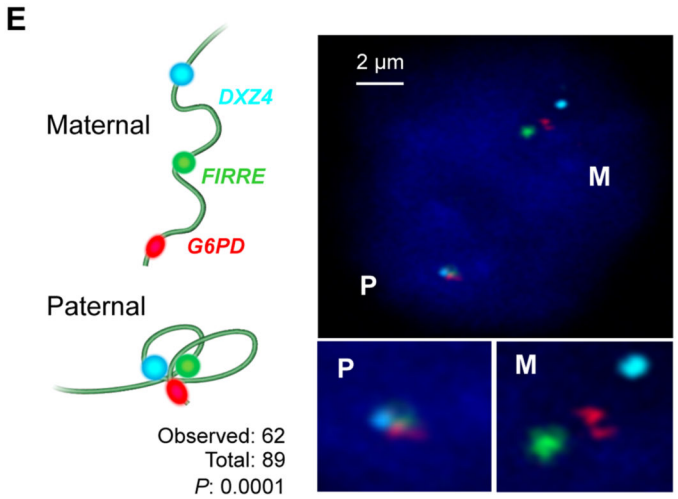


Figure 4. Haplotype mapping of chromatin interaction

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” (grey). 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 3 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 Figure S7F-G.

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 Figure S5.

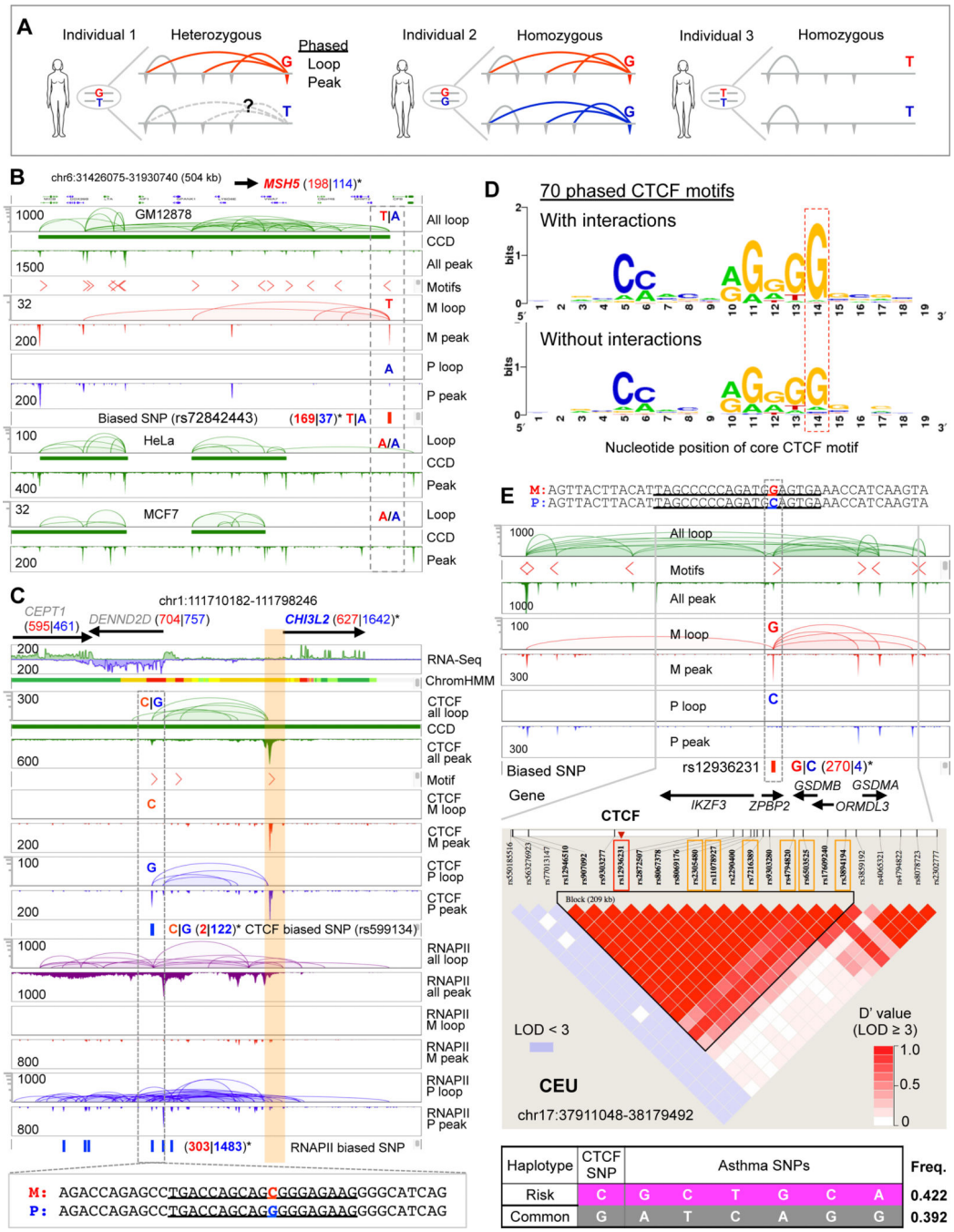


Figure 5. SNPs altering allelic CTCF chromatin interaction and the functional implication

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 6 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 ≥ 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.

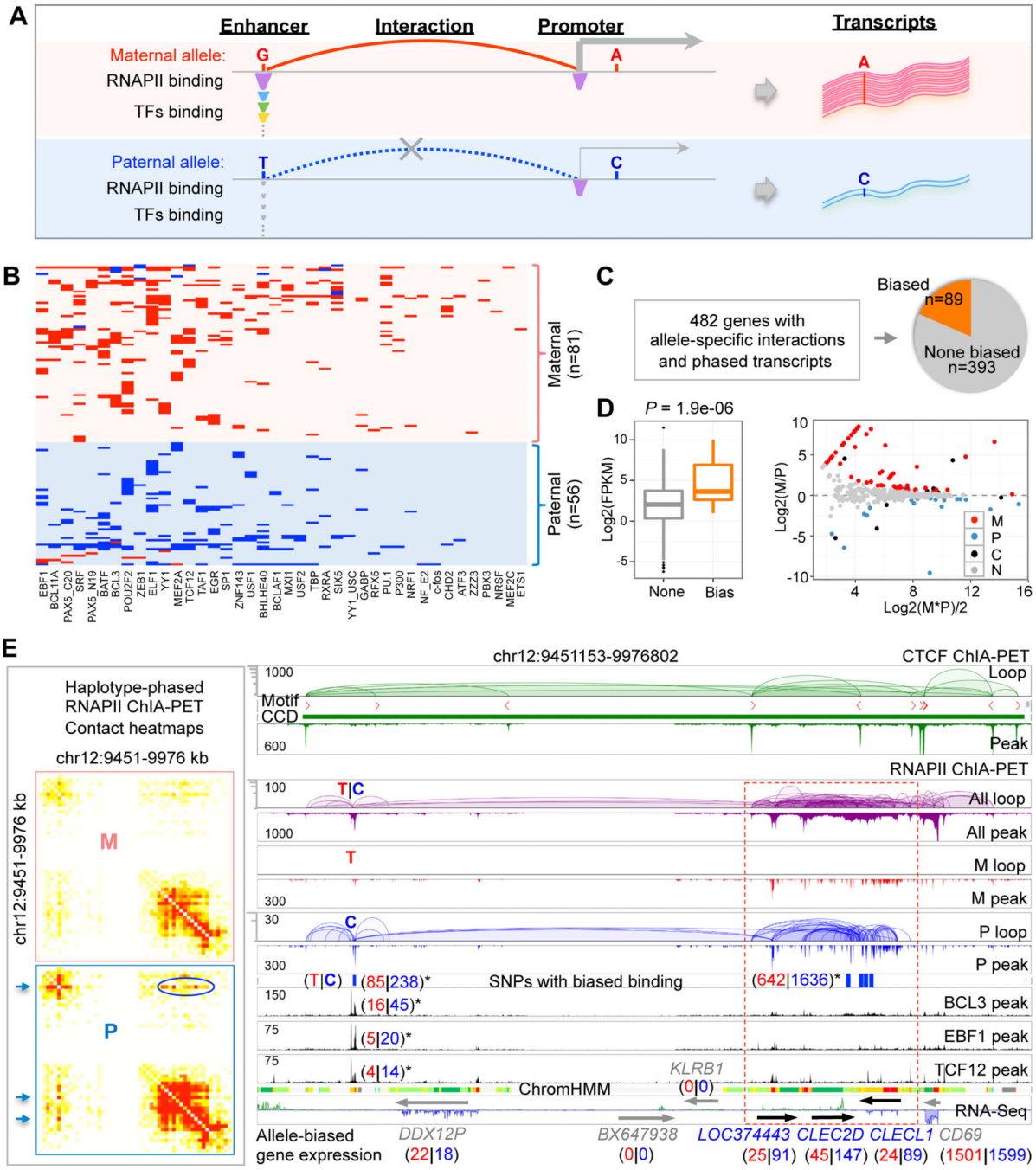


Figure 6. Allele-biased chromatin interactions mediated by RNAPII

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 (89) genes showed significant allele-biased gene expression

D. Left: Boxplot of the expression levels of genes with (red box) and without (grey 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, grey, 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 3 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.

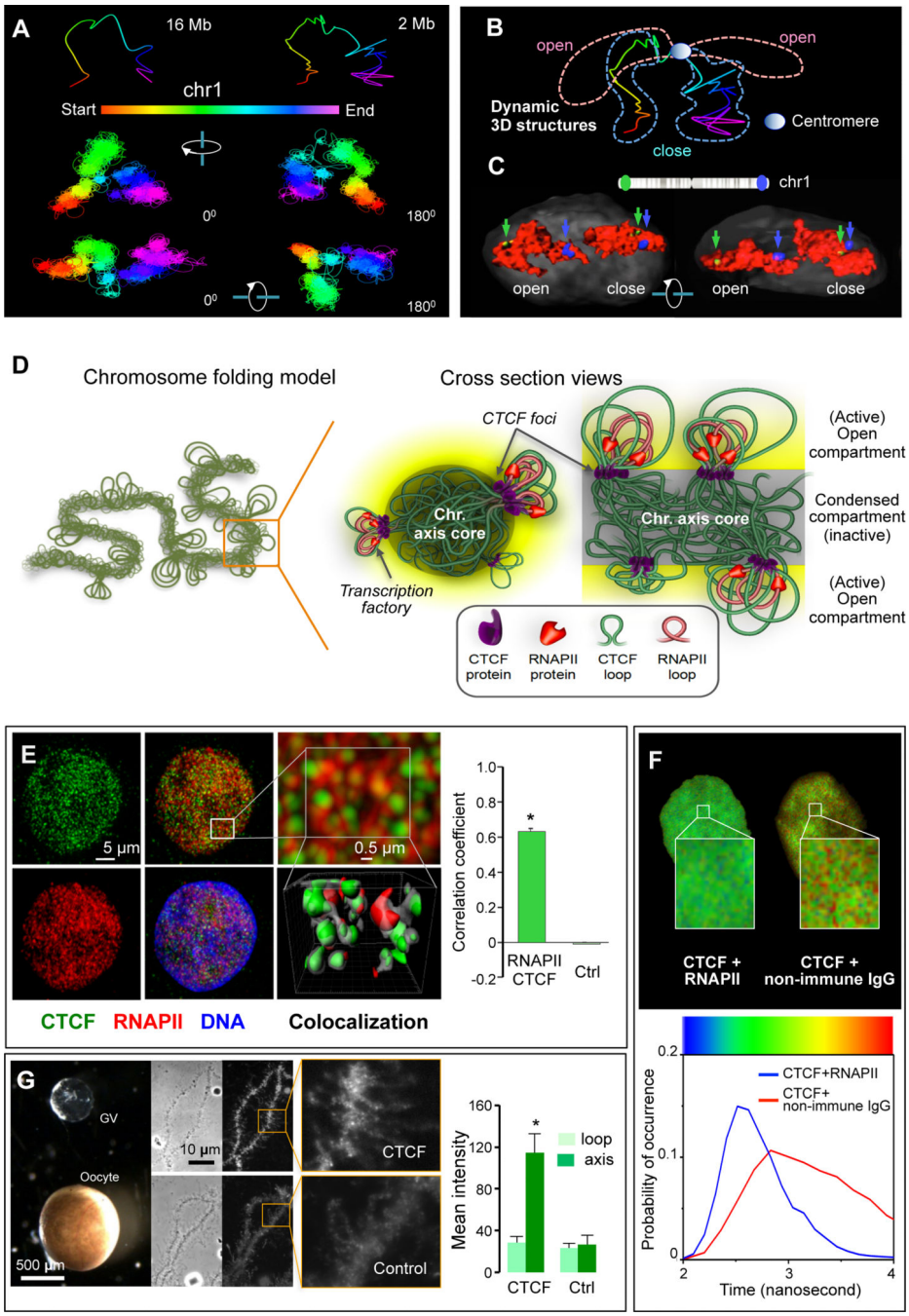


Figure 7. Chromatin model of CTCF foci and RNAPII transcription factories

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

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 immunostatined images. Data are shown as mean with s.e.m. *: $P < 0.001$.

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 ≤ 10 nm) is revealed by the shortening of the lifetime (nanoseconds) 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 s.e.m.

See also Figure S7, Data S2 and Extended Experimental Procedures.