

## MIT Open Access Articles

### *3D Chromosome Regulatory Landscape of Human Pluripotent Cells*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Ji, Xiong et al. "3D Chromosome Regulatory Landscape of Human Pluripotent Cells." *Cell Stem Cell* 18.2 (2016): 262–275.

**As Published:** <http://dx.doi.org/10.1016/j.stem.2015.11.007>

**Publisher:** Elsevier

**Persistent URL:** <http://hdl.handle.net/1721.1/107115>

**Version:** Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

**Terms of use:** Creative Commons Attribution-Noncommercial-Share Alike





Published in final edited form as:

*Cell Stem Cell*. 2016 February 4; 18(2): 262–275. doi:10.1016/j.stem.2015.11.007.

## 3D chromosome regulatory landscape of human pluripotent cells

Xiong Ji<sup>1,6</sup>, Daniel B. Dadon<sup>1,2,6</sup>, Benjamin E. Powell<sup>1,6</sup>, Zi Peng Fan<sup>1,3,6</sup>, Diego Borges-Rivera<sup>1,2,6</sup>, Sigal Shachar<sup>4</sup>, Abraham S. Weintraub<sup>1,2</sup>, Denes Hnisz<sup>1</sup>, Gianluca Pegoraro<sup>5</sup>, Tong Ihn Lee<sup>1</sup>, Tom Misteli<sup>4</sup>, Rudolf Jaenisch<sup>1,2,\*</sup>, and Richard A. Young<sup>1,2,\*</sup>

<sup>1</sup>Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142

<sup>2</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

<sup>3</sup>Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139

<sup>4</sup>National Cancer Institute, NIH, Bethesda, MD 20892

<sup>5</sup>High Throughput Imaging Facility (HiTIF), National Cancer Institute, NIH, Bethesda, MD 20892

### SUMMARY

In this study, we describe the 3D regulatory chromosome structural landscape of human naive and primed embryonic stem cells. To devise this map, we identified transcriptional enhancers and insulators in these cells and placed them within the context of cohesin-associated CTCF-CTCF loops using cohesin ChIA-PET data. The CTCF-CTCF loops we identified form a chromosomal framework of insulated neighborhoods, which in turn form topologically associated domains (TADs), that are largely preserved during transition between the naive and primed states.

Regulatory changes in enhancer-promoter interactions occur within insulated neighborhoods during cell state transition. The CTCF anchor regions we identified are conserved across species, influence gene expression, and are a frequent site of mutations in cancer cells, underscoring their functional importance in cellular regulation. These 3D regulatory maps of human pluripotent cells

\*Correspondence: Rudolf Jaenisch, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, jaenisch@wi.mit.edu. Richard A. Young, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, young@wi.mit.edu.

<sup>6</sup>Co-first author

### ACCESSION NUMBERS

Raw and processed sequencing data were deposited in GEO under accession number GSE69647

### SUPPLEMENTAL INFORMATION

Supplemental information includes Extended Experimental Procedures, 6 supplementary figures and 6 tables.

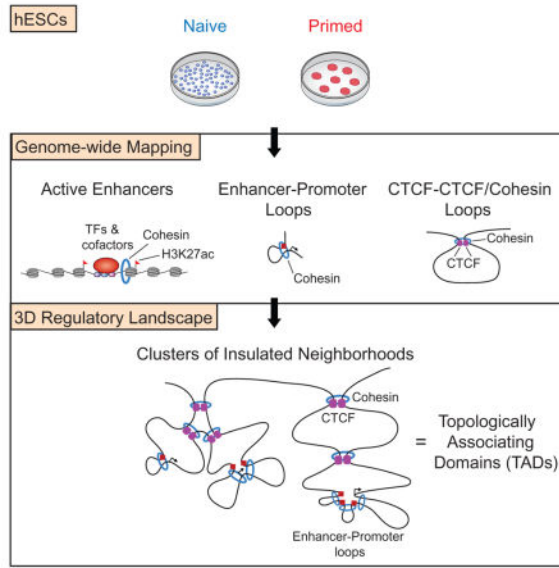
### AUTHOR CONTRIBUTIONS

X.J., D.B.D., B.E.P. and R.A.Y. designed experiments. X.J., D.B.D., Z.P.F., D.B.R., T.I.L. and R.A.Y. designed data analysis. B.E.P., D.B.D. and X.J. generated human naive and primed ESCs. X.J. performed ChIA-PET, ChIP-seq, RNA-seq and 3D DNA FISH. B.E.P. performed shRNA knockdown experiments. B.E.P., X.J., D.B.D. performed genome editing experiments. S.S. and G.P. performed high-throughput 3D DNA FISH. Z.P.F. and D.B.R. performed computational analyses. T.I.L., D.H. and A.S.W. contributed critical comments on the manuscript. X.J., D.B.D., T.I.L., R.J. and R.A.Y. wrote the paper. All authors edited the manuscript.

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

therefore provide a foundation for future interrogation of the relationships between chromosome structure and gene control in development and disease.

**Graphical Abstract**



**INTRODUCTION**

The gene expression programs that establish and maintain specific cell states in humans are controlled by regulatory proteins that bind specific genomic elements (Heinz et al., 2015; Levine et al., 2014; Plank and Dean, 2014; Smith and Shilatifard, 2014; Spitz and Furlong, 2012). Enhancer elements, first described over 30 years ago (Banerji et al., 1981; Benoist and Chambon, 1981; Gruss et al., 1981), are bound by transcription factors and can loop long distances to contact and regulate specific genes. There are approximately 1 million enhancers that have been identified in the human genome (Dunham et al., 2012; Thurman et al., 2012), and the constraints that cause them to operate only on their specific target genes are not fully understood (Zabidi et al., 2015). Insulator elements are bound by CTCF and prevent enhancers from operating across insulator boundaries (Bell et al., 1999; Geyer and Corces, 1992) and recent studies suggest such boundaries function in the context of 3D chromosome structures (Dixon et al., 2012; Downen et al., 2014; Handoko et al., 2011; Heidari et al., 2014; Nora et al., 2012; Phillips-Cremins et al., 2013; Rao et al., 2014). Understanding the control of a cell’s gene expression program thus requires a map of enhancers and insulators in the context of 3D chromosome structure.

The 3D topology of the genome is thought to contribute to the regulation of gene expression by creating constraints that produce regions of active and repressed transcription (Bickmore, 2013; de Graaf and van Steensel, 2013; de Laat and Duboule, 2013). Recent evidence indicates that both active and repressed compartments of chromosomes are partitioned into megabase-size topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012). TADs are regions of chromosomes that show evidence of relatively high DNA

interaction frequencies based on Hi-C chromosome conformation capture data. TADs are largely maintained through development, as TAD boundaries tend to be similar among various cell types (Dixon et al., 2015; Dixon et al., 2012; Phillips-Cremins et al., 2013).

The chromosome-structuring proteins CTCF and cohesin are important for the integrity of TAD boundaries and substructures (Guo et al., 2015; Narendra et al., 2015; Phillips-Cremins et al., 2013; Seitan et al., 2013; Sofueva et al., 2013; Zuin et al., 2014). CTCF and cohesin are essential for early embryogenesis, ubiquitously expressed and retained on their interphase chromatin sites in mitotic chromatin, and are thought to play important roles in epigenetic inheritance (Dorsett and Merkenschlager, 2013; Gomez-Diaz and Corces, 2014; Merkenschlager and Odom, 2013). CTCF is an 11 zinc-finger protein that binds CCTC motifs and can form homodimers, enabling two distal DNA-bound CTCF molecules to loop DNA. Cohesin is loaded at enhancer-promoter loops and occupies these sites and CTCF sites (Downen et al., 2013; Downen et al., 2014; Hadjur et al., 2009; Kagey et al., 2010; Parelho et al., 2008; Rubio et al., 2008; Wendt et al., 2008). Cohesin forms a large ring capable of encircling two DNA molecules (Gruber et al., 2003; Haering et al., 2002) and is thought to facilitate establishment and/or maintenance of enhancer-promoter loops and CTCF-CTCF loops. An emerging model suggests that cohesin-associated CTCF-CTCF loops occur within TADs and that enhancers generally interact with genes that occur within these loops (DeMare et al., 2013; Downen et al., 2014; Doyle et al., 2014; Handoko et al., 2011; Heidari et al., 2014; Phillips-Cremins et al., 2013; Rao et al., 2014). These CTCF-CTCF loops appear to function as insulated neighborhoods for gene regulation because the loss of either of the CTCF sites that close the loop can alter gene regulation within and immediately outside the loop (Downen et al., 2014). Insulated neighborhood structures have been described for key pluripotency genes in murine ESCs (Downen et al., 2014), but the extent to which these structures account for the Hi-C DNA interactions used to define TADs is not clear.

It is assumed that development is controlled by transcriptional and epigenetic regulators that function in the context of various chromosome structures, but we lack a map of such structures in cells representative of early human development. With the recent isolation of naive human embryonic stem cells (Chan et al., 2013; Gafni et al., 2013; Takashima et al., 2014; Theunissen et al., 2014; Ware et al., 2014), it is possible to deduce the 3D regulatory landscape of one of the earliest stages of human development. Naive ESCs represent the ground state of pluripotency and are characterized by a gene expression profile that is similar to that in cleavage embryos (De Los Angeles et al., 2015; Hackett and Surani, 2014; Martello and Smith, 2014). Primed ESCs, while pluripotent, represent a subsequent post-implantation epiblast cell state that has a developmental bias towards the ectoderm (De Los Angeles et al., 2015; Hackett and Surani, 2014; Martello and Smith, 2014). Although these two states of human pluripotency have been characterized by global gene expression analyses, it is not known whether the conversion of the naive to the primed state involves changes in chromatin configuration. Defining the 3D regulatory landscape of these two cell states should prove to be valuable for understanding the transcriptional control of early human development.

To deduce the 3D regulatory landscape of naive and primed human embryonic stem cells, we identified enhancers, insulators and cohesin-associated chromatin interactions. The results show how cohesin-associated CTCF-CTCF loops, and the cohesin-associated enhancer-promoter loops within them, organize Topologically Associating Domains (TADs) in naive and primed human embryonic stem cells. Enhancers interact with specific genes located within the CTCF-CTCF loops, indicating that they function as insulated neighborhoods. The CTCF sites that contribute to these loops are highly conserved and hypomethylated, and loss of these sites occurs frequently in cancer. We thus provide an initial map of the 3D regulatory landscapes of human pluripotent cells and a foundation for further studies of the relationships between chromosome structure and gene control in development and disease.

## RESULTS

To investigate the 3D regulatory landscape of naive hESCs and the isogenic primed hESCs from which the naive cells were derived, we generated populations of both cell states and reinvestigated their morphology and gene expression programs to confirm that they maintained key features previously described for these cells reproducibly (Theunissen et al., 2014). As expected, the colonies of naive hESCs exhibited a dome-shaped morphology and the colonies of primed hESCs had a flat morphology (Figure S1A). The gene expression programs were reinvestigated by generating high-quality RNA-seq datasets (Extended Experimental Procedures). Cross-species clustering confirmed that the naive and primed hESC gene expression datasets were highly similar to those previously established for naive and primed hESCs as well as their murine counterparts (Figure S1B) (Theunissen et al., 2014; Huang et al, 2014). Further analysis of the RNA-seq data confirmed that genes previously noted as preferentially expressed in either naive or primed hESCs were indeed preferentially expressed in these RNA-seq datasets (Figure S1C) (Takashima et al., 2014; Theunissen et al., 2014). A complete list of genes that are preferentially expressed in the naive or primed hESCs can be found in Table S1. These results confirm that the conditions used for growth and maintenance of these isogenic naive and primed human pluripotent states are reproducible (Theunissen et al., 2014).

### Enhancers, insulators and cohesin-associated DNA interactions in human ESCs

Enhancers and insulators are *cis*-regulatory elements that can be identified by the regulatory proteins that occupy them and by the looped structures that are formed by cohesin-associated interactions (Figure 1A). For both naive and primed ESCs, we identified regions occupied by cohesin and then identified putative enhancers and insulators (Figure 1B, Table S2). Enhancers were identified by generating ChIP-seq data for histone H3K27ac and confirmed with ChIP-seq data for the MED1 subunit of Mediator and the OCT4 master transcription factor (Figure 1B, Table S2). Candidate insulators were identified by determining the genome-wide occupancy of CTCF (Figure 1B, Table S2). The data for the naive ESCs indicates that ~29% of cohesin-occupied sites involve active enhancers and promoters and ~57% involve CTCF sites that are not associated with enhancers and promoters (Figure 1B). Similar results were obtained for the primed ESCs, except there was

a substantial fraction of cohesin-occupied regions that were associated with Polycomb modifications (Figure 1B), as noted previously (Theunissen et al., 2014).

To identify cohesin-associated loops, we generated ChIA-PET data for cohesin in both naive and primed hESCs. The ChIA-PET technique was used because it yields genome-wide, high-resolution (~4 kb) interaction data coupled to the location of a specific protein, thus providing potential mechanistic insight into that set of chromatin interactions (Fullwood et al., 2009). We selected cohesin for ChIA-PET because it is a relatively well-studied SMC complex that is loaded at enhancer-promoter loops and can thus identify those interactions, and can also occupy CTCF sites and thus identify those interactions as well.

Biological replicate ChIA-PET datasets for the cohesin subunit SMC1 were generated for both the naive and primed hESCs. We acquired a total of ~400 million reads acquired for both naive and primed hESCs (Table S3). The respective replicates showed a high degree of correlation (Pearson's  $r > 0.98$ , Figure S1D, Figure S1E), so replicate data was pooled and processed as described in Extended Experimental Procedures. The naive hESC dataset contained ~88 million unique paired-end tags (PETs) that identified 35,286 high confidence cohesin-associated intra-chromosomal interactions (Table S3), and the primed hESC dataset contained ~125 million unique PETs that identified 46,257 high confidence cohesin-associated intra-chromosomal interactions (Table S3). The results for the *MYCN* locus in naive hESCs and the effects of filtering for high-confidence interactions are shown in Figure 1C. At this locus, multiple interactions between super-enhancer constituents and the *MYCN* promoter are observed. A summary of a subset of the high confidence interactions identified in naive and primed hESCs based on the Downen et al. analysis pipeline (Downen et al., 2014; Extended Experimental Procedures) is shown in Figure 1D. These high confidence interactions were used for further analyses unless otherwise stated.

### Cohesin-associated loops organize TADs

We first studied the cohesin-associated DNA loops that occur between two CTCF-bound sites in the two hESC conditions and found that the majority (80%) of such loops in naive hESCs were also found in the primed hESCs (Figure 2A). There were 12,987 CTCF-CTCF loops in naive hESCs, encompassing 37% of the genome and 33% of protein-coding genes (Table S3). These CTCF-CTCF loops ranged from 4 kb to >800 kb and contained 0–24 protein-coding genes, with a median of 200 kb and 1 protein-coding gene per loop. Similar numbers were obtained for the primed human ESCs (Table S3). Previous studies have noted that when CTCF homo-dimers form DNA loops, the two CTCF sequence motifs that are bound occur in specific orientations (Rao et al., 2014), and we found that the two occupied sites that contribute to CTCF loops do occur predominantly in the convergent orientation (Figure S2A).

Recent studies have noted a degree of correspondence between CTCF-CTCF loops and TAD structures (Downen et al., 2014; Rao et al., 2014). A comparison of the CTCF-CTCF loops identified here with TADs identified previously in H1 ESCs (primed hESCs) with Hi-C data (Dixon et al., 2015) revealed several especially striking features. The CTCF-CTCF loops almost always occurred within TADs and showed interactions that closely corresponded to the Hi-C interaction heatmap (Figure 2B). Genome-wide analysis indicated that the CTCF-

CTCF loops correlate with the H1 hESC Hi-C signal to a striking degree and much more than would be expected at random (Figure 2B). We found evidence for CTCF-CTCF loops that spanned entire TADs for a large fraction of TADs (Figure 2C). Saturation analysis indicated that the ChIA-PET datasets are approximately 50% complete (Figure S2B), indicating that only a subset of all CTCF-CTCF loops were identified in the high confidence data, so it is possible that most TAD boundaries are defined by CTCF-CTCF loops. For two TADs spanned by CTCF-CTCF loops, 3D DNA FISH was used to show that the boundaries of these TADs were in close proximity compared to their distance to a third, genomically equidistant, locus in naive hESCs (Figure 2D). To determine whether TAD-spanning CTCF-CTCF loops are required for the proximity of TAD boundaries, we first confirmed that the boundaries of a different TAD with CTCF-CTCF spanning loops are in close proximity and then deleted the CTCF binding sites at either end and measured the effects on proximity of the two boundaries using 3D DNA FISH (Figure 2E). The results show that loss of either binding site causes a separation between the two boundary regions that are otherwise close in space. Together, these results suggest that TAD-spanning CTCF-CTCF loops make important contributions to TAD structure.

If cohesin-associated loops play a major role in TAD structure, it should be possible to reconstruct TADs, which were previously derived solely from Hi-C data, by using cohesin ChIA-PET data. The results shown in Figure 2F confirm that the cohesin ChIA-PET data, processed using the same Hidden Markov algorithm used to process Hi-C data, can capture most TAD boundaries derived from Hi-C data in H1 hESCs. This observation led us to determine whether similar results could be obtained for murine ESC TADs using previously generated murine ESC cohesin ChIA-PET data; the results confirmed that CTCF-CTCF loops span a substantial portion of TADs (Figure S2C) and that cohesin ChIA-PET data and Hi-C data produce similar TAD structures (Figure S2D). Because TAD structures are known to be largely preserved in various cell types (Dixon et al., 2015; Dixon et al., 2012; Phillips-Cremins et al., 2013), we compared the primed hESC CTCF-CTCF loops to those predicted from Hi-C datasets from H1 hESCs and in situ Hi-C datasets from seven different somatic cell types (Rao et al., 2014) and found a high degree of overlap in predicted loops (Figure S2E–H). These results are consistent with the idea that cohesin-associated DNA interactions provide much of the underlying structure of TADs.

### CTCF-CTCF loops form insulated neighborhoods

Previous studies in murine ESCs showed that CTCF-CTCF loops containing active pluripotency genes or silent Polycomb-associated genes can function as insulated neighborhoods for gene control (Figure 3A) because DNA interactions between regulatory elements and genes occur within the CTCF-CTCF loops and the loss of either of the CTCF sites can alter gene regulation within or immediately outside the loop (Downen et al., 2014). If the CTCF-CTCF loops identified in hESC function as insulated neighborhoods, we expect that most cohesin-associated interactions with an endpoint inside the loop have their other endpoint within the loop. Indeed, interactions that originate within a CTCF-CTCF loop almost invariably end within the boundaries of the loop in naive and primed hESCs (Figure S3A). Furthermore, we found that enhancers generally interact with a target gene within the

CTCF-CTCF loops (Figure 3B), consistent with the view that the CTCF-CTCF loops constrain enhancer-promoter interactions within the loops.

To determine whether CTCF-CTCF loops are functionally important for normal expression of local genes, we performed CRISPR-mediated deletions of anchors of CTCF-CTCF loops that surround the super-enhancer associated genes *PRDM14* and *LEFTY1*. Deletion of CTCF anchor sites had limited effects on expression of these super-enhancer driven genes that play essential roles in these ESCs, but caused substantial upregulation of genes that occur immediately outside the CTCF-CTCF loop, consistent with the idea that the loop constrains the super-enhancers to function within the loop (Figure 3C, 3D, S3B). These results further demonstrate that the integrity of CTCF-CTCF loops is important for proper expression of local genes.

We examined CTCF-CTCF loops in syntenic regions of human and mouse chromosomes to ascertain whether they are conserved, as has been observed previously for TADs (Dixon et al., 2012). Examination of the CTCF-CTCF loop structures in murine and human ESCs revealed that they are largely preserved in these syntenic regions (Figure S3C). We found that the CTCF boundary elements that were shown to function as insulated neighborhood boundaries in murine ESCs (Downen et al., 2014) have counterparts in the hESCs studied here, and these CTCF-CTCF loops contain human homologues of the murine pluripotency genes (Figure S3D). Thus, cohesin-associated CTCF-CTCF loops are largely preserved between syntenic regions of human and mouse, where conserved boundary CTCF sites have previously been shown to be essential for insulator function in mouse.

### Schematics of TAD structures

We assembled structural schematics of TADs based on CTCF-CTCF loops and enhancer-promoter loops and show a subset that illustrate common themes (Figure 4; Figure S4). CTCF-CTCF loops frequently span TADs, effectively forming one large insulated neighborhood. In some TADs, nested CTCF-CTCF interactions occur such that genes are embedded within two or more independent CTCF-CTCF loops. Cohesin-associated enhancer-promoter interactions essentially always occur within the smallest CTCF-CTCF loop formed within the TAD where the gene occurs, again consistent with the idea that the CTCF-CTCF loops have insulating properties. The CTCF-CTCF loop structures of TADs in naive and primed cells were very similar, although there were some instances where a TAD-spanning loop or an internal loop identified in one of the pluripotent cells was absent in data for the other cell (Figure S4).

The TAD schematics assembled with cohesin-associated loop data likely represent the minimal set of structures, as the cohesin ChIA-PET data is not saturated and some lower-confidence data is not included. In addition, these schematics represent one potential conformation of TADs, but because the underlying data originated from a population of cells, additional conformations are possible. Nonetheless, it is useful to use this to estimate the minimal frequencies of TAD substructures due to CTCF-CTCF loops. TADs range in size from 0.2–21 Mb and contain 0–768 genes. The median number of CTCF-CTCF insulated neighborhoods that occurred within each TAD was 2 and these ranged from 4 kb to 2.9 Mb and contained 0–24 protein-coding genes. In this Version 1.0 map of TAD-containing



insulated neighborhoods, the median neighborhood was 200 kb and contained one gene whose average size was approximately 30 kb.

### Differential gene control in naive and primed hESCs

To gain insights into the differential regulation of genes that contribute to naive and primed states of pluripotency, we compared the enhancer landscapes of the two cell types (Figure 5A). Of the 24,755 active enhancers identified in naive and primed hESCs using H3K27ac ChIP-seq data, 16% showed > 2-fold H3K27ac signal in naive hESCs relative to primed hESCs and 26% showed > 2-fold H3K27ac signal in primed hESCs relative to naive hESCs (Figure 5A). To focus on genes likely to contribute to the differential control of these pluripotent states, we concentrated our analysis on super-enhancers and their associated genes (Figure 5B, 5C; Table S4) because super-enhancers are known to drive expression of key pluripotency genes in mESCs (Whyte et al., 2013).

We found that differentially regulated pluripotency genes generally occur in similar TAD CTCF-CTCF loop structures in naive and primed cells (Figure 5D, 5E; Figure S5A–D). Inspection of 3D chromosome structure at loci for genes that have naive-preferred enhancers and are preferentially expressed in naive hESCs revealed that they share cohesin-associated CTCF-CTCF structures in naive and primed hESCs, as shown for *KLF4* in Figure 5D. *KLF4* has a super-enhancer only in naive cells and is expressed 5-fold higher in naive than primed cells. Similarly, many genes that are preferentially expressed in primed hESCs occur within shared CTCF-CTCF structures, as shown for *OTX2* in Figure 5E. *OTX2* has a super-enhancer only in primed cells and is expressed 10-fold higher in primed cells than in naive cells. The theme of differential expression within the context of similar CTCF-CTCF loops was observed in many additional TADs (Figure S5A–D). Although we could identify 125 naive-specific CTCF-CTCF loops and 28 primed-specific CTCF-CTCF loops (Table S4) some of which showed striking differences in loop structure (Figure S2I), none of these harbored genes with known roles in pluripotency. These results indicate that differential expression of a key set of pluripotency genes generally occurs in the context of preserved structural frameworks composed of CTCF-CTCF loops in naive and primed hESCs.

The differential regulation of certain pluripotency genes might be due to a state-specific function in naive or primed cells. To test this idea, we investigated whether the naive state is more dependent on *KLF4* than the primed state. *KLF4* mRNA was knocked down using shRNA in both naive and primed cells, and the results show that expression of pluripotency markers in naive cells is more dependent on normal levels of *KLF4* mRNA in primed cells (Figure 5F). These results suggest that naive pluripotency in human ESCs is especially dependent on the *KLF4* transcription factor.

The CTCF sites at CTCF-CTCF loop anchors in the hESCs are consistently bound by CTCF in many other human cell types, as exemplified by ChIP-seq data for 16 different cell types at the *TBX3* and *OTX2* loci (Figure S5E, S5F), so these may contribute to similar loop structures in differentiated cells. Similar evidence for consistent binding of CTCF in multiple cell types has been described (Cuddapah et al., 2009; Downen et al., 2014; Heidari et al., 2014; Kim et al., 2007; Phillips-Cremens et al., 2013; Schmidt et al., 2012; Wang et al., 2012). This reinforces the idea that CTCF, together with cohesin, generates similar

chromosomal frameworks in different cells and that transcriptional regulatory elements function within this context to produce cell-type specific gene expression programs.

### Conservation of 3D structure and associations with disease

It has been estimated that approximately 65% of Hi-C derived chromosome structures are static among different cell types and different species (Dixon et al., 2015). The observation that chromosome structures are largely conserved in primates (Dixon et al., 2015; Dixon et al., 2012; Rao et al., 2014; Vietri Rudan et al., 2015) led us to investigate the extent to which CTCF binding is similarly conserved. Analysis of CTCF binding sites across 10 primates indicates that the DNA sequence in anchor regions of CTCF-CTCF loops in hESCs is more conserved in primates and vertebrates than in regions bound by CTCF that do not participate in loops (Figure 6A, S6A). A similar analysis showed that the CTCF DNA sequence motif in hESC loop anchor regions is highly conserved in primates and vertebrates (Figure 6B–C, S6B). CTCF is known to preferentially bind to hypomethylated DNA sequences (Wang et al., 2012), and further analysis revealed that the CTCF binding sequences in hESC loop anchor regions exhibit DNA hypomethylation across 37 human cell/tissue types (Figure S6C). Hypomethylation at CTCF-CTCF loop anchors persists in a broad spectrum of human cells, and is evident even during stages of embryogenesis when DNA is globally hypomethylated (Figure S6D).

The conservation of CTCF-CTCF loop anchor sequences led us to consider whether their variation contributes to various human diseases and syndromes. Analysis of disease-associated single nucleotide polymorphisms (SNPs) showed that they tend to occur in proximity to enhancers, as observed previously (Hnisz et al., 2013; Maurano et al., 2012), but were not enriched in CTCF-CTCF loop anchor regions that lack evidence of local enhancer activity (Figure 6D). Deletions, duplications and inversions that affect TAD structure and contribute to congenital diseases have been reported (Giorgio et al., 2015; Lupianez et al., 2015), but the present results suggest that disease-associated SNPs generally occur much more frequently in enhancers than in CTCF loop anchor regions.

Misregulation of gene expression is a common feature in cancer (Hanahan and Weinberg, 2011; Lee and Young, 2013) and with evidence that proper regulation of gene expression depends on CTCF-CTCF insulated neighborhoods, it is possible that this framework is altered in cancer cells. Indeed, a recent report indicates that CTCF/cohesin-binding sites are frequently mutated in colorectal cancer (Katainen et al., 2015). Analysis of somatic mutations present in the International Cancer Genome Consortium (ICGC) database (Zhang et al., 2011) revealed that 7307 mutations occur in hESC CTCF loop anchors (Table S5), and that the CTCF DNA binding motif present in hESC loop anchor regions is among the most altered human factor binding sequence in cancer cells (Figure 6E). Given the conservation of CTCF loop anchors, and evidence that CTCF-CTCF loops are mostly preserved between hESC and cancer cells (Figure S6E), it was striking to note that the mutations in the ICGC database that occur in CTCF anchor sites are often adjacent to oncogenes and other cancer-associated genes known to be dysregulated in specific cancer cells (Table S5). For example, *CCNE1* overexpression is associated with pancreatic, liver and ovarian cancer (Calhoun et al., 2003; Etemadmoghadam et al., 2013; Jung et al., 2001), and mutations affecting the

anchor CTCF motifs have been documented for the hESC loop containing *CCNE1* (Figure 6F). Similarly, *NOTCH1* overexpression is associated with ovarian cancer (Rose et al., 2010) and mutations affecting an anchor CTCF motif has been documented for the hESC loop containing *NOTCH1* (Figure 6F). These results support the idea that mutations that alter the cohesin-associated CTCF-CTCF loops identified in hESCs may contribute to the misregulation of gene expression that is inherent to the cancer state.

## DISCUSSION

We describe here a first draft of the 3D regulatory landscape of human embryonic stem cells (ESCs) in two pluripotent states and new insights into the relationships between chromosome structure and gene regulation. The naive and primed state of pluripotency represent an *in vitro* correlate of the earliest states of human development and our results are likely relevant for our understanding of epigenetic mechanisms that govern initial cell fate decisions in the embryo. Enhancers and genes generally interact within the context of the CTCF-CTCF loops identified here, and these loops thus form insulated neighborhoods that constrain interactions between regulatory elements and genes. TADs appear to be formed by clusters of CTCF-CTCF loops and the gene regulatory interactions that occur within them. The CTCF sites that contribute to insulated neighborhoods in hESCs are highly conserved in primates, are rarely affected by sequence variation in humans, but are frequently altered in cancer. These initial 3D regulatory maps of human pluripotent cells thus reveal how cohesin-associated CTCF-CTCF and enhancer-promoter loops contribute to the control of key genes and provide a foundation for further studies of development and disease.

Our results suggest that TADs can be considered as nested sets of cohesin-associated CTCF-CTCF loops, as illustrated by the schematics shown in Figure 4. In many cases, the largest CTCF-CTCF loop spans the TAD and additional CTCF-CTCF loops often occur within the TAD. This structure helps explain why enhancers generally control only a limited number of genes despite having an ability to function in either orientation and at long distances, and why only a subset of CTCF-bound sites function as insulators. The pairs of CTCF-bound sites that interact to form a loop can function to produce an insulated neighborhood within which regulatory interactions occur. These results confirm and provide a mechanistic explanation for the hypothesis that TADs provide physical and functional constraints on interactions between regulatory elements and genes (Dekker, 2014; Gorkin et al., 2014). They are also consistent with a growing body of evidence that cohesin-associated CTCF-CTCF loops occur within TADs and that enhancers generally interact with genes that occur within these loops (DeMare et al., 2013; Downen et al., 2014; Handoko et al., 2011; Heidari et al., 2014; Phillips-Cremins et al., 2013; Rao et al., 2014).

The CTCF binding sites that form the loop anchors of insulated neighborhoods in hESCs are highly conserved in primates. These loop anchor CTCF sites are hypomethylated, which may be important for CTCF binding and/or for formation of CTCF-cohesin loop structures. The anchor sites are rarely affected by human sequence variation, but are frequently altered by somatic mutations in cancer. It will thus be important to determine whether cancer cells exploit rearrangement of insulated neighborhoods to facilitate acquisition of their oncogenic gene expression programs.

The naive and primed human ESCs studied here represent the earliest stages of human development that can be cultured. Comparison of these genetically identical naive and primed ESCs revealed that key differences in gene control occur in the context of similar insulated neighborhoods in the two pluripotent cell states. Most of the hESC chromosome structures that occur in these cells are probably retained during differentiation and thus provide a foundation for further understanding transcriptional control of cell identity in a broad spectrum of human cells, where approximately a million regulatory elements have been mapped but most have yet to be physically and functionally linked to genes. These maps of hESC genome structure should also prove valuable for identifying and further understanding genetic alterations that disrupt 3D structures and cause disease.

## EXPERIMENTAL PROCEDURES

### Cell Culture

Human naive and primed ESCs were cultured as described previously (Theunissen et al., 2014). Detailed information is described in the Extended Experimental Procedures.

### ChIP-seq Library Generation and Sequencing

Chromatin immunoprecipitation (ChIP) was performed as previously described (Ji et al., 2015). 50 million naive or primed hESCs were used for each ChIP experiment. The following antibodies were used for ChIP: anti-H3K27ac (Abcam, ab4729), anti-CTCF (Millipore, 07-729), anti-MED1 (Bethyl Labs, A300-793A), anti-OCT4 (Santa Cruz, sc-8628). For each ChIP, 5  $\mu$ g of antibody and 50  $\mu$ l protein G Dynabeads (Life Technology, 10004D) were used. The ChIP-seq libraries were prepared using the TruSeq ChIP Sample Prep Kit (Illumina, IP-202-1012), and sequenced on the Illumina HiSeq 2000.

### ChIA-PET

ChIA-PET was performed using a modified version of a previously described protocol (Downen et al., 2014). 400 million naive or primed hESCs were used for each ChIA-PET library construction. The ChIA-PET libraries were generated in three stages. In the first stage, ChIP was performed using 25  $\mu$ g anti-SMC1 antibody (Bethyl Labs, A300-055A) and 250  $\mu$ l protein G Dynabeads (Life Technology, 10004D). This stage was the same as the experimental procedure described in the ChIP-seq library generation. The second stage was proximity ligation of ChIP-DNA fragments which consists of end blunting and A-tailing to create easily ligated ends, followed by ligation to simultaneously add linker sequences required for later steps and ligate ends of fragments together. The third stage was the tagmentation of ligated products, purification of the tagmented DNA fragments, amplification of the DNA by PCR, size selection and paired-end sequencing. The ChIA-PET library was subjected to 100 $\times$ 100 paired-end sequencing using Illumina HiSeq 2000. Detailed information is described in Extended Experimental Procedures.

### Data Analysis

The ChIA-PET data analyses were performed as previously described (Downen et al., 2014; Phanstiel et al., 2015) (Table S3, S6). The topologically associating domain related analyses

were performed as previously described (Dixon et al., 2012). Detailed information is described in Extended Experimental Procedures.

All the other information is described in Extended Experimental Procedures.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank Brian J. Abraham and Jill Downen for data analysis. We thank Tom Volkert at the Whitehead Genome Technology Core for sequencing. We thank Wendy Salmon for assistance with confocal microscopy, Raaji Alagappan, Dongdong Fu and Tenzin Lungjangwa for preparation of MEFs. This research was in part supported by the Intramural Research Program of the National Institutes of Health (NIH), NCI, Center for Cancer Research. This work was supported by an Erwin Schrödinger Fellowship (J3490) from the Austrian Science Fund (FWF) (to D.H.), National Institutes of Health Grants HG002668 (to R.A.Y.) and HD 045022 (to R.J.) and a grant from the Simons Foundation SFLIFE #286977 (to R.J.). R.J. is a founder of Fate Therapeutics and R.A.Y. is a founder of Syros Pharmaceuticals.

## References

- Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA-sequences. *Cell*. 1981; 27:299–308. [PubMed: 6277502]
- Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*. 1999; 98:387–396. [PubMed: 10458613]
- Benoist C, Chambon P. In vivo sequence requirements of the SV40 early promoter region. *Nature*. 1981; 290:304–310. [PubMed: 6259538]
- Bickmore WA. The Spatial Organization of the Human Genome. *Annual Review of Genomics and Human Genetics*. 2013; 14:67–84.
- Calhoun ES, Jones JB, Ashfaq R, Adsay V, Baker SJ, Valentine V, Hempen PM, Hilgers W, Yeo CJ, Hruban RH, et al. BRAF and FBXW7 (CDC4, FBW7, AGO, SEL10) mutations in distinct subsets of pancreatic cancer: potential therapeutic targets. *The American journal of pathology*. 2003; 163:1255–1260. [PubMed: 14507635]
- Chan YS, Goeke J, Ng JH, Lu X, Gonzales KAU, Tan CP, Tng WQ, Hong ZZ, Lim YS, Ng HH. Induction of a Human Pluripotent State with Distinct Regulatory Circuitry that Resembles Preimplantation Epiblast. *Cell Stem Cell*. 2013; 13:663–675. [PubMed: 24315441]
- Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res*. 2009; 19:24–32. [PubMed: 19056695]
- de Graaf CA, van Steensel B. Chromatin organization: form to function. *Current Opinion in Genetics & Development*. 2013; 23:185–190. [PubMed: 23274160]
- de Laat W, Duboule D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*. 2013; 502:499–506. [PubMed: 24153303]
- De Los Angeles A, Ferrari F, Xi R, Fujiwara Y, Benvenisty N, Deng H, Hochedlinger K, Jaenisch R, Lee S, Leitch HG, et al. Hallmarks of pluripotency. *Nature*. 2015; 525:469–478. [PubMed: 26399828]
- Dekker J. Two ways to fold the genome during the cell cycle: insights obtained with chromosome conformation capture. *Epigenetics & chromatin*. 2014; 7:25. [PubMed: 25435919]
- DeMare LE, Leng J, Cotney J, Reilly SK, Yin J, Sarro R, Noonan JP. The genomic landscape of cohesin-associated chromatin interactions. *Genome Res*. 2013; 23:1224–1234. [PubMed: 23704192]

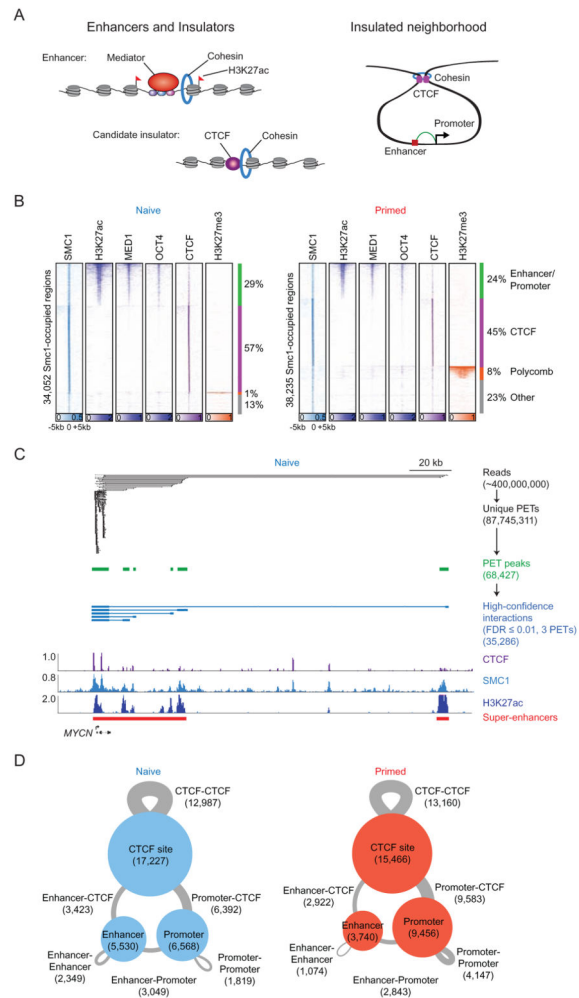
- Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, et al. Chromatin architecture reorganization during stem cell differentiation. *Nature*. 2015; 518:331–336. [PubMed: 25693564]
- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012; 485:376–380. [PubMed: 22495300]
- Dorsett D, Merkenschlager M. Cohesin at active genes: a unifying theme for cohesin and gene expression from model organisms to humans. *Curr Opin Cell Biol*. 2013; 25:327–333. [PubMed: 23465542]
- Dowen JM, Bilodeau S, Orlando DA, Hubner MR, Abraham BJ, Spector DL, Young RA. Multiple structural maintenance of chromosome complexes at transcriptional regulatory elements. *Stem cell reports*. 2013; 1:371–378. [PubMed: 24286025]
- 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]
- Doyle B, Fudenberg G, Imakaev M, Mirny LA. Chromatin loops as allosteric modulators of enhancer-promoter interactions. *PLoS computational biology*. 2014; 10:e1003867. [PubMed: 25340767]
- Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. [PubMed: 22955616]
- Etemadmoghadam D, Weir BA, Au-Yeung G, Alsop K, Mitchell G, George J, Davis S, D'Andrea AD, Simpson K, Hahn WC, et al. Synthetic lethality between CCNE1 amplification and loss of BRCA1. *Proc Natl Acad Sci U S A*. 2013; 110:19489–19494. [PubMed: 24218601]
- 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]
- Gafni O, Weinberger L, Mansour AA, Manor YS, Chomsky E, Ben-Yosef D, Kalma Y, Viukov S, Maza I, Zviran A, et al. Derivation of novel human ground state naive pluripotent stem cells. *Nature*. 2013; 504:282–286. [PubMed: 24172903]
- Geyer PK, Corces VG. DNA position-specific repression of transcription by a drosophila zinc finger protein. *Genes & Development*. 1992; 6:1865–1873. [PubMed: 1327958]
- Giorgio E, Robyr D, Spielmann M, Ferrero E, Di Gregorio E, Imperiale D, Vaula G, Stamoulis G, Santoni F, Atzori C, et al. A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Human molecular genetics*. 2015; 24:3143–3154. [PubMed: 25701871]
- Gomez-Diaz E, Corces VG. Architectural proteins: regulators of 3D genome organization in cell fate. *Trends in Cell Biology*. 2014; 24:703–711. [PubMed: 25218583]
- Gorkin DU, Leung D, Ren B. The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell Stem Cell*. 2014; 14:762–775. [PubMed: 24905166]
- Gruber S, Haering CH, Nasmyth K. Chromosomal cohesin forms a ring. *Cell*. 2003; 112:765–777. [PubMed: 12654244]
- Gruss P, Dhar R, Khoury G. Simian virus-40 tandem repeated sequences as an element of the early promoter. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*. 1981; 78:943–947.
- 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]
- Hackett JA, Surani MA. Regulatory principles of pluripotency: from the ground state up. *Cell Stem Cell*. 2014; 15:416–430. [PubMed: 25280218]
- Hadjur S, Williams LM, Ryan NK, Cobb BS, Sexton T, Fraser P, Fisher AG, Merkenschlager M. Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*. 2009; 460:410–413. [PubMed: 19458616]
- Haering CH, Lowe J, Hochwagen A, Nasmyth K. Molecular architecture of SMC proteins and the yeast cohesin complex. *Mol Cell*. 2002; 9:773–788. [PubMed: 11983169]

- Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011; 144:646–674. [PubMed: 21376230]
- Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F, et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet*. 2011; 43:630–638. [PubMed: 21685913]
- Heidari N, Phanstiel DH, He C, Grubert F, Jahanbani F, Kasowski M, Zhang MQ, Snyder MP. Genome-wide map of regulatory interactions in the human genome. *Genome Res*. 2014; 24:1905–1917. [PubMed: 25228660]
- Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. *Nature reviews Molecular cell biology*. 2015; 16:144–154. [PubMed: 25650801]
- Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, Young RA. Super-enhancers in the control of cell identity and disease. *Cell*. 2013; 155:934–947. [PubMed: 24119843]
- Huang K, Maruyama T, Fan G. The naive state of human pluripotent stem cells: a synthesis of stem cell and preimplantation embryo transcriptome analyses. *Cell Stem Cell*. 2014; 15:410–415. [PubMed: 25280217]
- Ji X, Dadon DB, Abraham BJ, Lee TI, Jaenisch R, Bradner JE, Young RA. Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. *Proceedings of the National Academy of Sciences of the United States of America*. 2015; 112:3841–3846. [PubMed: 25755260]
- Jung YJ, Lee KH, Choi DW, Han CJ, Jeong SH, Kim KC, Oh JW, Park TK, Kim CM. Reciprocal expressions of cyclin E and cyclin D1 in hepatocellular carcinoma. *Cancer letters*. 2001; 168:57–63. [PubMed: 11368878]
- Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*. 2010; 467:430–435. [PubMed: 20720539]
- Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen H, Hanninen UA, Cajuso T, et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet*. 2015 Advance online publication.
- Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanekov VV, Ren B. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*. 2007; 128:1231–1245. [PubMed: 17382889]
- Lee TI, Young RA. Transcriptional Regulation and Its Misregulation in Disease. *Cell*. 2013; 152:1237–1251. [PubMed: 23498934]
- Levine M, Cattoglio C, Tjian R. Looping Back to Leap Forward: Transcription Enters a New Era. *Cell*. 2014; 157:13–25. [PubMed: 24679523]
- Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*. 2015; 161:1012–1025. [PubMed: 25959774]
- Martello G, Smith A. The nature of embryonic stem cells. *Annual review of cell and developmental biology*. 2014; 30:647–675.
- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012; 337:1190–1195. [PubMed: 22955828]
- Merkenschlager M, Odom DT. CTCF and Cohesin: Linking Gene Regulatory Elements with Their Targets. *Cell*. 2013; 152:1285–1297. [PubMed: 23498937]
- Narendra V, Rocha PP, An D, Raviram R, Skok JA, Mazzoni EO, Reinberg D. Transcription. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science*. 2015; 347:1017–1021. [PubMed: 25722416]
- Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. 2012; 485:381–385. [PubMed: 22495304]

- Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, Canzonetta C, Webster Z, Nesterova T, et al. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*. 2008; 132:422–433. [PubMed: 18237772]
- Phanstiel DH, Boyle AP, Heidari N, Snyder MP. Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics*. 2015; 31:3092–3098. [PubMed: 26034063]
- Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong CT, Hookway TA, Guo C, Sun Y, et al. Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell*. 2013; 153:1281–1295. [PubMed: 23706625]
- Plank JL, Dean A. Enhancer function: mechanistic and genome-wide insights come together. *Mol Cell*. 2014; 55:5–14. [PubMed: 24996062]
- Rao SSP, 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]
- Rose SL, Kunnimalaiyaan M, Drenzek J, Seiler N. Notch 1 signaling is active in ovarian cancer. *Gynecologic oncology*. 2010; 117:130–133. [PubMed: 20060575]
- Rubio ED, Reiss DJ, Weicsh PL, Disteché CM, Filippova GN, Baliga NS, Aebersold R, Ranish JA, Krumm A. CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105:8309–8314. [PubMed: 18550811]
- Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*. 2012; 148:335–348. [PubMed: 22244452]
- Seitan VC, Faure AJ, Zhan Y, McCord RP, Lajoie BR, Ing-Simmons E, Lenhard B, Giorgetti L, Heard E, Fisher AG, et al. Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Research*. 2013; 23:2066–2077. [PubMed: 24002784]
- Smith E, Shilatifard A. Enhancer biology and enhanceropathies. *Nat Struct Mol Biol*. 2014; 21:210–219. [PubMed: 24599251]
- Sofueva S, Yaffe E, Chan WC, Georgopoulou D, Vietri Rudan M, Mira-Bontenbal H, Pollard SM, Schroth GP, Tanay A, Hadjur S. Cohesin-mediated interactions organize chromosomal domain architecture. *The EMBO journal*. 2013; 32:3119–3129. [PubMed: 24185899]
- Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*. 2012; 13:613–626.
- Takashima Y, Guo G, Loos R, Nichols J, Ficz G, Krueger F, Oxley D, Santos F, Clarke J, Mansfield W, et al. Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. *Cell*. 2014; 158:1254–1269. [PubMed: 25215486]
- Theunissen TW, Powell BE, Wang H, Mitalipova M, Faddah DA, Reddy J, Fan ZP, Maetzel D, Ganz K, Shi L, et al. Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. *Cell Stem Cell*. 2014; 15:471–487. [PubMed: 25090446]
- Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. The accessible chromatin landscape of the human genome. *Nature*. 2012; 489:75–82. [PubMed: 22955617]
- Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep*. 2015; 10:1297–1309. [PubMed: 25732821]
- Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res*. 2012; 22:1680–1688. [PubMed: 22955980]
- Ware CB, Nelson AM, Mecham B, Hesson J, Zhou W, Jonlin EC, Jimenez-Caliani AJ, Deng X, Cavanaugh C, Cook S, et al. Derivation of naive human embryonic stem cells. *Proc Natl Acad Sci U S A*. 2014; 111:4484–4489. [PubMed: 24623855]
- Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*. 2008; 451:796–801. [PubMed: 18235444]



- Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell*. 2013; 153:307–319. [PubMed: 23582322]
- Zabidi MA, Arnold CD, Scherhuber K, Pagani M, Rath M, Frank O, Stark A. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*. 2015; 518:556–559. [PubMed: 25517091]
- Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B, et al. International Cancer Genome Consortium Data Portal-a one-stop shop for cancer genomics data. *Database-the Journal of Biological Databases and Curation*. 2011
- Zuin J, Dixon JR, van der Reijden MIJA, Ye Z, Kolovos P, Brouwer RWW, van de Corput MPC, van de Werken HJG, Knoch TA, van Ijcken WFJ, et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the United States of America*. 2014; 111:996–1001. [PubMed: 24335803]



**Figure 1. Components of 3D regulatory landscape**

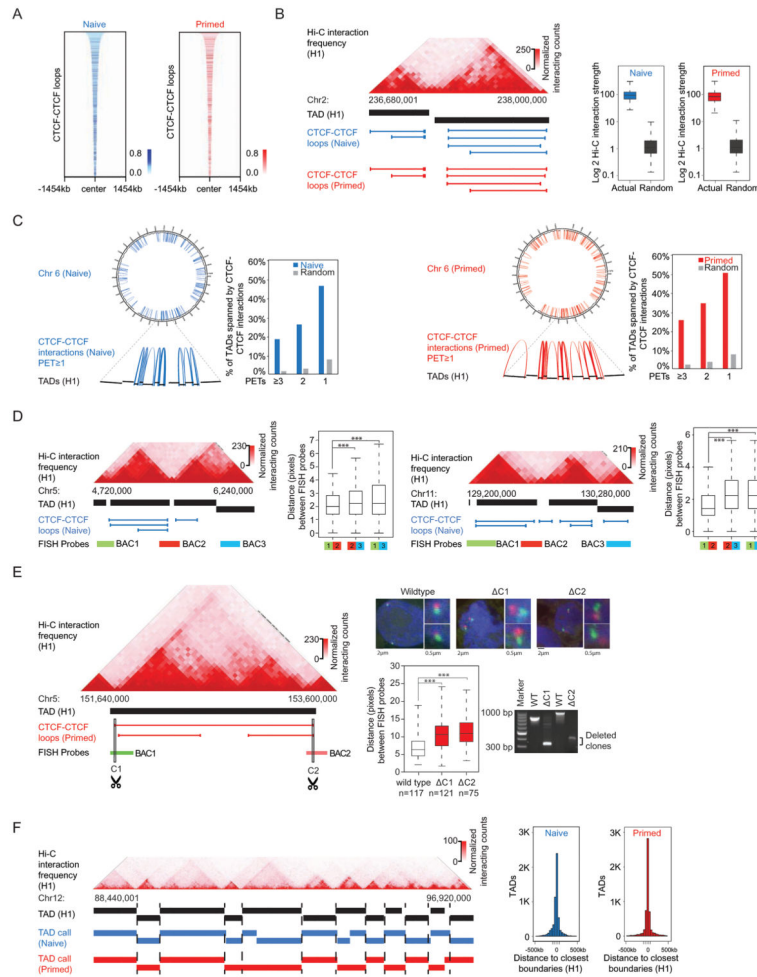
(A) Enhancers and insulators (left panel). Enhancers are occupied by transcription factors, mediator and cohesin, and their associated nucleosomes are marked by H3K27ac. Candidate insulators are occupied by CTCF and cohesin. Model of insulated neighborhoods formed by cohesin-associated CTCF-CTCF interactions, within which enhancers loop to promoters of target genes (right panel).

(B) Heatmap representation of ChIP-seq data for H3K27ac, MED1, OCT4, CTCF and H3K27me3 at SMC1-occupied regions in naive (left panel) and primed (right panel) hESCs. Read density is displayed within a 10 kb window and color scale intensities are shown in rpm/bp. Cohesin occupies three classes of sites: enhancer-promoter sites, Polycomb-occupied sites, and CTCF-occupied sites.

(C) Cohesin (SMC1) ChIA-PET data analysis at the *MYCN* locus in naive hESCs. The algorithm used to identify paired-end tags (PETs) is described in Extended Experimental Procedures. PETs and interactions involving enhancers and promoters within the window are displayed at each step in the analysis pipeline. Binding profiles for CTCF, SMC1 and H3K27ac are displayed at the bottom.

(D) High-confidence cohesin-associated interaction maps in naive (left panel) and primed (right panel) hESCs. CTCF binding sites, enhancers and promoters involved in cohesin-associated interactions are indicated as circles, and the size of circles correspond to the number of sites. The interactions between two regions are indicated as gray lines, and the size of lines correspond to the number of interactions.

See also Figure S1, Table S1, S2, S3



**Figure 2. CTCF-CTCF loops underlie much of TAD structure**

(A) Heatmap of cohesin-associated CTCF-CTCF loops showing that these loops in naive hESCs are largely preserved in primed hESCs. The 9,344 CTCF-CTCF loops that define the putative insulated neighborhoods in naive hESCs were ranked by size and shown. The color bar indicates normalized PET-signal at these CTCF-CTCF loops.

(B) TAD heat map of interaction frequencies and CTCF-CTCF loops that define the putative insulated neighborhoods. Normalized Hi-C interaction frequencies in H1 hESCs are displayed in a two-dimensional heat map (Dixon et al., 2015) with the TADs indicated as black bars. Shared CTCF-CTCF loops are indicated as blue lines (naive) and red lines (primed). A correlation analysis between Hi-C interaction frequency (H1 hESCs) and CTCF-CTCF loops in naive and primed hESCs is displayed to the right in a box plot; randomly generated TADs were used as the background control.

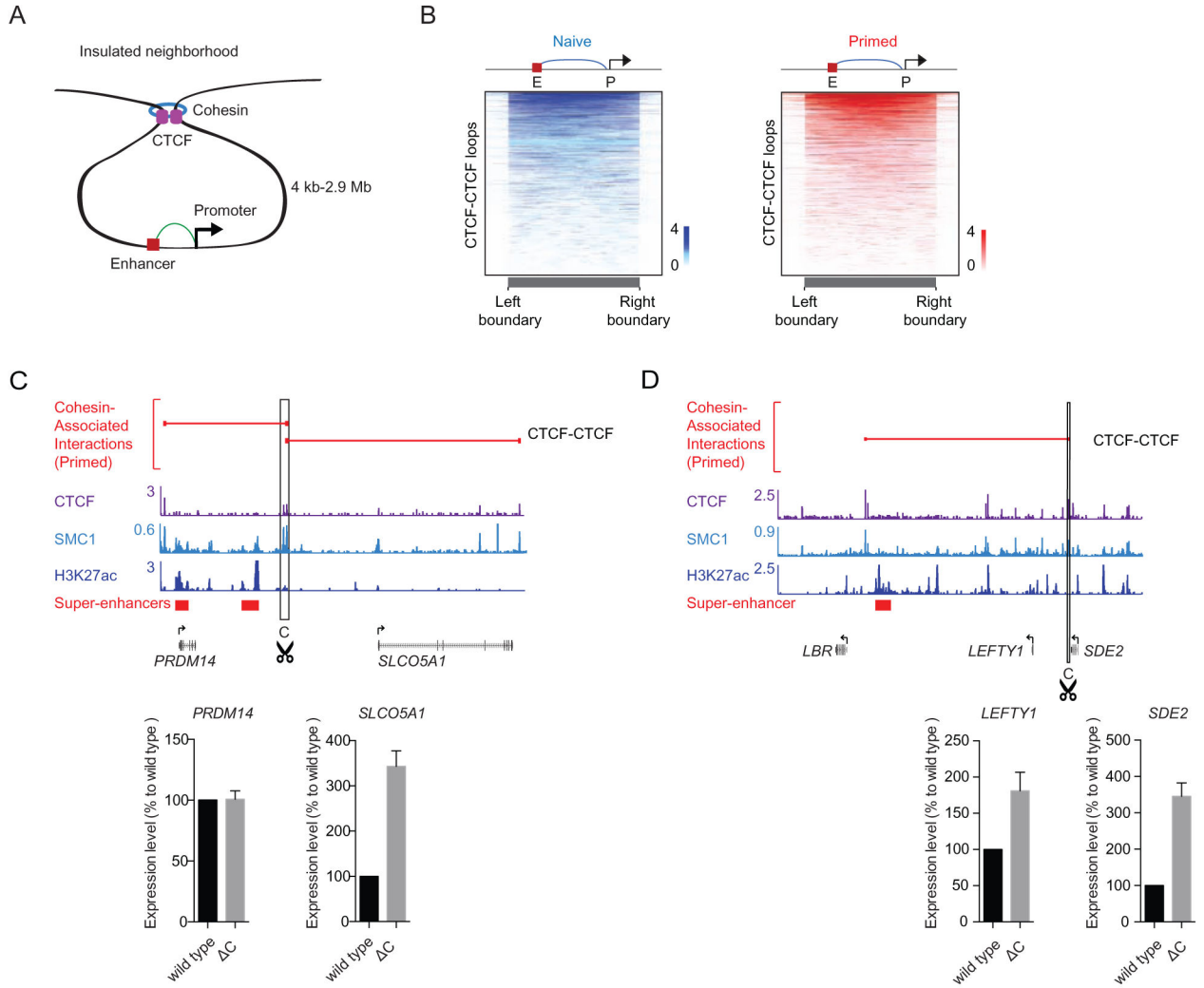
(C) CTCF-CTCF loops span many TADs identified using Hi-C data in H1 hESCs. Chromosome 6 is displayed as a circos plot in both naive and primed hESCs, with zoomed in regions below. CTCF-CTCF loops (  $\geq 1$  PETs) are indicated as blue arcs (naive) and red arcs (primed). The bar graphs show percentages of TADs spanned by CTCF-CTCF loops when various confidence thresholds (1, 2, 3 PETs) were used. Random shuffling of TAD locations (100 iterations) serve as the background control.

(D) Physical distance between TAD borders is shorter than an equidistant control locus. The Hi-C interaction heatmaps, TADs and CTCF-CTCF loops were shown the same as (B). The green, red and blue bars indicate the location of BAC probes used for DNA FISH at each locus. Box plots of minimal normalized distances between pairs of loci generated from >1500 FISH probe spots per condition are displayed with the corresponding probe pairs labeled below. The stars indicate significance using the Mann-Whitney test ( $***P < 10^{-28}$ ). Images were obtained using a 40X objective.

(E) Measurement of DNA proximity by 3D DNA FISH before and after deletions of CTCF binding sites at either end of a TAD-spanning CTCF-CTCF loop. The Hi-C interaction heatmaps, TADs and CTCF-CTCF loops were shown the same as (B). The green and red bars indicate the location of BAC probes used for DNA FISH. The scissor-marked regions (C1, C2) were deleted by CRISPR-mediated deletion. Examples of two color DNA FISH images are shown in the right panel, the quantification of distance between green and red probes are displayed with bar graphs shown below. The stars indicate significance using the Mann-Whitney test ( $***P < 10^{-13}$ ). Images were obtained using a 100X objective; the n indicates the number of alleles quantified for each sample. The genotyping PCR data are displayed at the bottom right.

(F) Cohesin ChIA-PET data can be used to discover TADs. A comparison of TADs derived with the same algorithm from Hi-C data (Dixon et al., 2015) and cohesin ChIA-PET data for a portion of chromosome 12 (left panel). A global analysis indicates that the cohesin ChIA-PET and Hi-C derived TAD boundaries are close (right panel).

See also Figure S2, Table S3



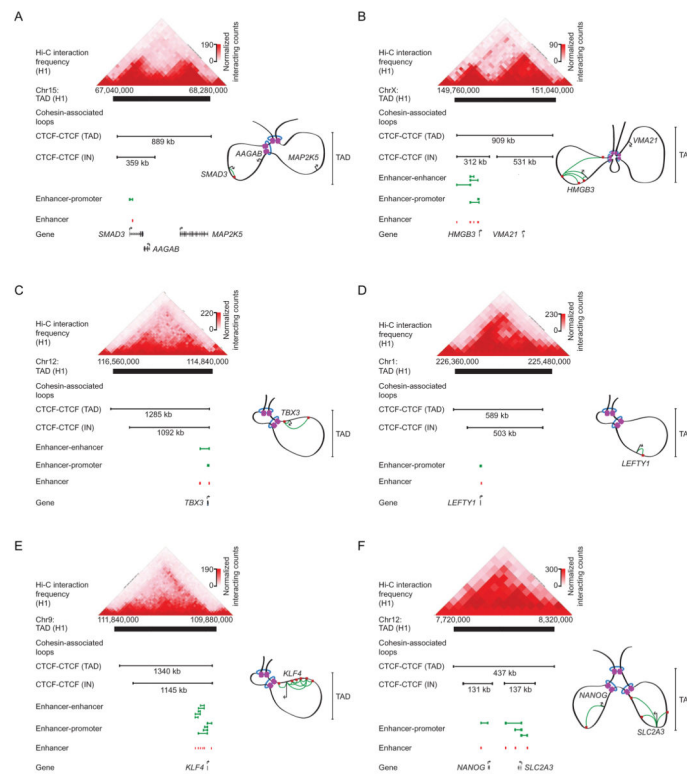
**Figure 3. Putative insulated neighborhoods in hESCs**

(A) Schematic of insulated neighborhood.

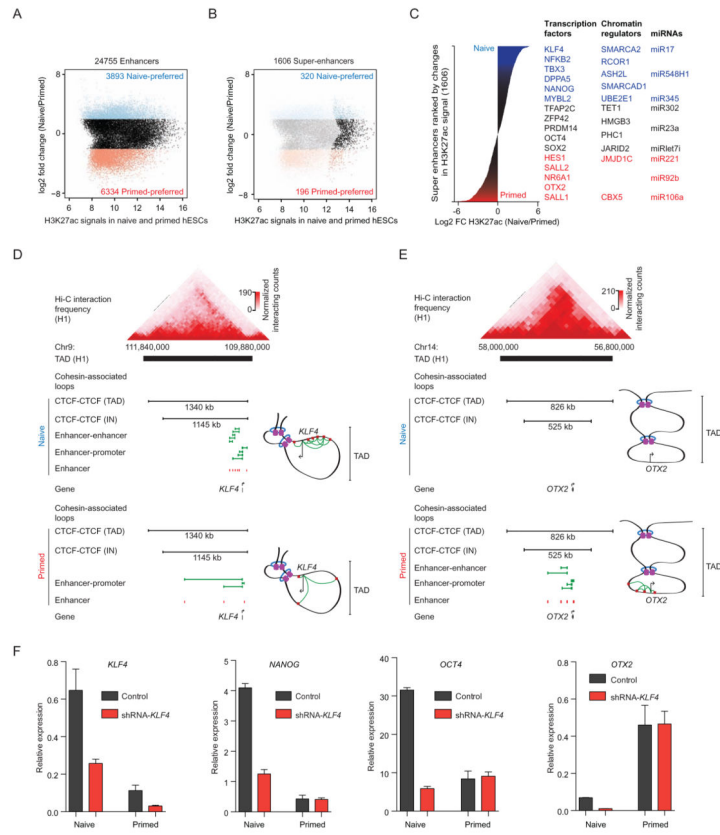
(B) Enhancer-promoter interactions occur predominantly within CTCF-CTCF loops that define putative insulated neighborhoods in hESCs. The color bar indicates the number of enhancer-promoter interactions spanning the genomic location.

(C) and (D) CRISPR-mediated deletion of CTCF sites at two loci (*PRDM14* locus (C) and *LEFTY1* locus (D)). The top of each panel shows a subset of CTCF-CTCF loops depicted as red lines and binding profiles for CTCF, cohesin (SMC1), and H3K27ac in primed hESCs at the respective loci. A subset of genes present in these loops is shown for simplicity. The super-enhancers are indicated as red bars. The bottom of each panel shows RT-qPCR results for the gene expression levels of the indicated genes in wild type and cells with deleted CTCF sites. Error bars were generated from at least three replicates.

See also Figure S3



**Figure 4. 3D regulatory structures of TADs containing key pluripotency genes**  
 (A–F) Schematics of 3D structure for TADs containing *SMAD3*, *HMGB3*, *TBX3*, *LEFTY1*, *KLF4* and *NANOG* in naive hESCs. For each TAD, Hi-C interaction data (Dixon et al., 2015) is shown together with cohesin-associated loop data for TAD-spanning CTCF loops, insulated neighborhood-spanning CTCF loops, enhancer-enhancer loops and enhancer-promoter loops. A subset of CTCF-CTCF loops was selected for display based on a directionality index (Extended Experimental Procedures) and a subset of genes present in these loops is shown for simplicity.  
 See also Figure S4



**Figure 5. Differential enhancer landscape reveals key transcription factors, chromatin regulators and miRNAs in naive and primed pluripotency**

(A) Scatterplot comparison of H3K27ac ChIP-seq peaks used to call enhancers in naive and primed hESCs.

(B) Scatterplot comparison of super-enhancers in naive and primed hESCs.

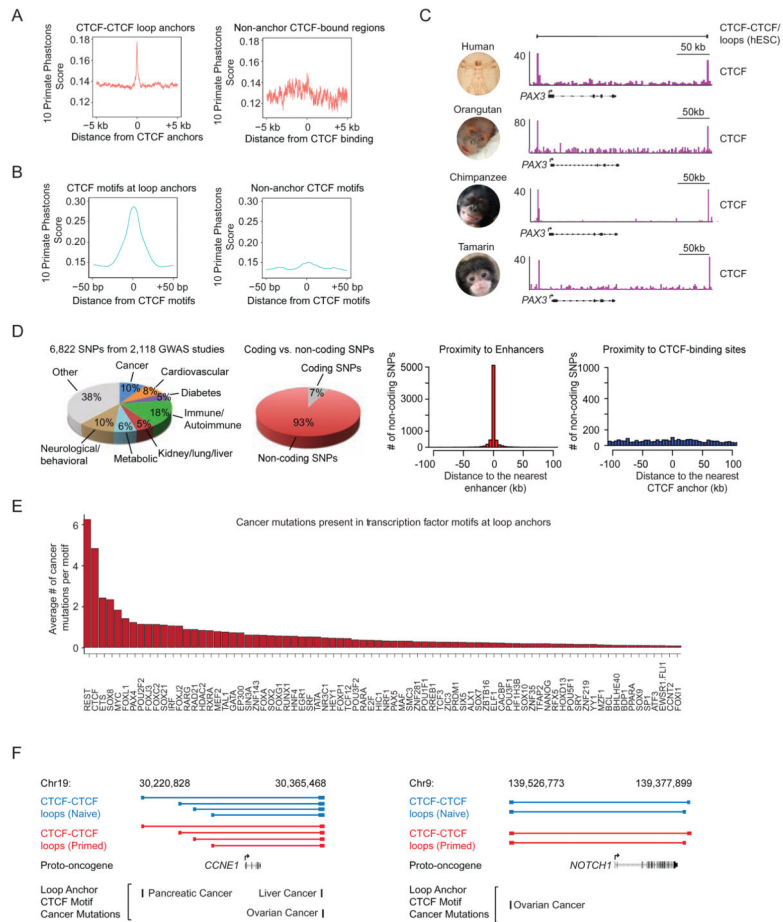
(C) Distribution of differential H3K27ac ChIP-seq signal density across the super-enhancer regions of naive and primed hESCs. Genes encoding key transcription factors, chromatin regulators, and miRNAs associated with super-enhancers are listed.

(D) 3D regulatory structure of a TAD containing *KLF4* in both naive and primed hESCs with Hi-C and cohesin ChIA-PET data as described in Figure 4. The naive and primed cells share TAD and insulated neighborhood structure, but a super-enhancer and cohesin-associated interactions between the super-enhancer and the *KLF4* promoter are readily detected only in naive cells.

(E) 3D regulatory structure of a TAD containing *OTX2* in both naive and primed hESCs with Hi-C, cohesin ChIA-PET and enhancer data as described in (D).

(F) Gene expression analysis after shRNA knockdown of *KLF4* in naive and primed hESCs. The RT-qPCR results were displayed as black (control) and red (shRNA *KLF4*) bar graphs. Error bars were generated from at least three replicates. See also Figure S5, Table S4





**Figure 6. Conservation of 3D structure and associations with disease**

(A) and (B) DNA sequence in anchor regions (A) and the CTCF DNA sequence motif (B) of CTCF-CTCF loops in hESCs is more conserved in primates than DNA sequence in hESC regions bound by CTCF that do not serve as loop anchors.

(C) A CTCF-CTCF loop containing the *PAX3* gene in human and ChIP-seq gene tracks showing conserved binding of CTCF at this locus in Human, Orangutan, Chimpanzee and Tamarin genomes (Schwalie et al., 2013).

(D) Catalog of SNPs linked to phenotypic traits and diseases in genome-wide association studies (GWAS) and SNP association with enhancer and CTCF anchor regions in hESCs. Pie chart showing percentage of SNPs associated with the highlighted classes of traits and diseases (Left). Distribution of trait-associated SNPs in coding and noncoding regions of the genome (Middle Left). Location of all noncoding trait-associated SNPs relative to all enhancers identified in 86 human cell and tissue samples. x axis reflects binned distances of each SNP to the nearest enhancer. SNPs located within enhancers are assigned to the 0 bin (Middle Right). Location of all noncoding trait-associated SNPs relative to CTCF binding sites in loop anchor regions (Right).

(E) Cancer mutations in transcription factor motifs at hESC CTCF-CTCF loop anchors.

(F) Cancer mutations found at CTCF motifs at the anchors of CTCF-CTCF loops in hESCs that contain the proto-oncogenes *CCNE1* and *NOTCH1*. Blue (naive) and red (primed)

CTCF-CTCF loops with mutations within the CTCF motifs in their anchors are displayed above the proto-oncogene contained within these loops. Below, mutations from the International Cancer Genome Consortium are displayed along with the cancers from which these were sequenced.

See also Figure S6, Table S5

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript