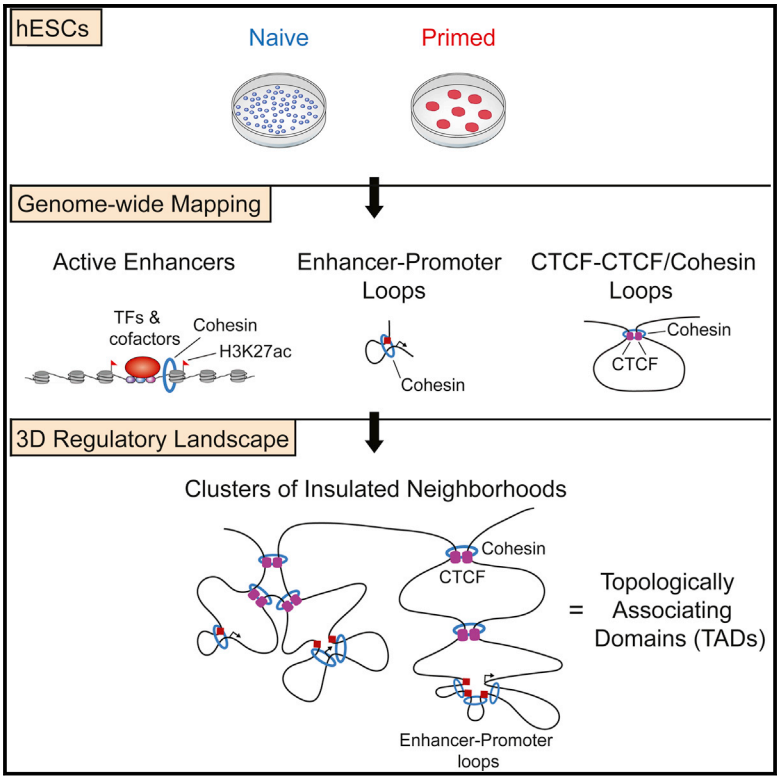
Resource

Cell Stem Cell

3D Chromosome Regulatory Landscape of Human Pluripotent Cells

Graphical Abstract



Authors

Xiong Ji, Daniel B. Dadon, Benjamin E. Powell, ..., Tom Misteli, Rudolf Jaenisch, Richard A. Young

Correspondence

jaenisch@wi.mit.edu (R.J.), young@wi.mit.edu (R.A.Y.)

In Brief

Ji et al. map the chromosome organizational structures that underlie gene regulation in human naive and primed pluripotent cells. Their framework of cohesin-associated CTCF loops, and the cohesin-associated enhancer-promoter loops within them, provides a reference map for future interrogation of regulatory interactions.

Highlights

- ChIA-PET analysis maps enhancers and insulators into looped domains
- Cohesin-associated loops organize topologically associating domains (TADs)
- Regulatory changes during cell state transitions take place within TADs
- The conserved anchors of CTCF-CTCF loops are frequently mutated in cancer

Accession Numbers

GSE69647

Ji et al., 2016, Cell Stem Cell *18*, 262–275
February 4, 2016 ©2016 Elsevier Inc.
http://dx.doi.org/10.1016/j.stem.2015.11.007

CellPress

Cell Stem Cell
Resource

3D Chromosome Regulatory Landscape of Human Pluripotent Cells

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

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA
[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]National Cancer Institute (NCI), NIH, Bethesda, MD 20892, USA
[5]High Throughput Imaging Facility (HiTIF), NCI, NIH, Bethesda, MD 20892, USA
[6]Co-first author
*Correspondence: jaenisch@wi.mit.edu (R.J.), young@wi.mit.edu (R.A.Y.)
http://dx.doi.org/10.1016/j.stem.2015.11.007

SUMMARY

In this study, we describe the 3D chromosome regulatory 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 associating domains (TADs) that are largely preserved during the 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 therefore provide a foundation for future interrogation of the relationships between chromosome structure and gene control in development and disease.

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; Dowen 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-sized 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., 2012, 2015; 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; Gómez-Díaz and Corces, 2014; Merkenschlager and Odom, 2013). CTCF is an 11 zinc-finger protein that binds CCCTC 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 (Dowen et al., 2013, 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; Dowen 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 (Dowen et al., 2014). Insulated neighborhood structures have been described for key pluripotency genes in murine embryonic stem cells (mESCs) (Dowen 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 (hESCs) (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 toward 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 hESCs, 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 TADs in naive and primed hESCs. 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 sequencing (RNA-seq) datasets (Supplemental 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 hESCs

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 hESCs, we identified regions occupied by cohesin and then identified putative enhancers and insulators (Figure 1B; Table S2). Enhancers were identified by generating chromatin immunoprecipitation sequencing (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 hESCs indicate 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 hESCs, 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 chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) data for cohesin in both naive and primed hESCs. The ChIA-PET technique was used because it yields genome-wide, high-resolution (~4kb) 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 thus can identify those interactions, and it also can 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 for both naive and primed hESCs (Table S3). The respective replicates showed a high degree of correlation (Pearson's $r > 0.98$, Figures S1D and S1E), so replicate data were pooled and processed as described in

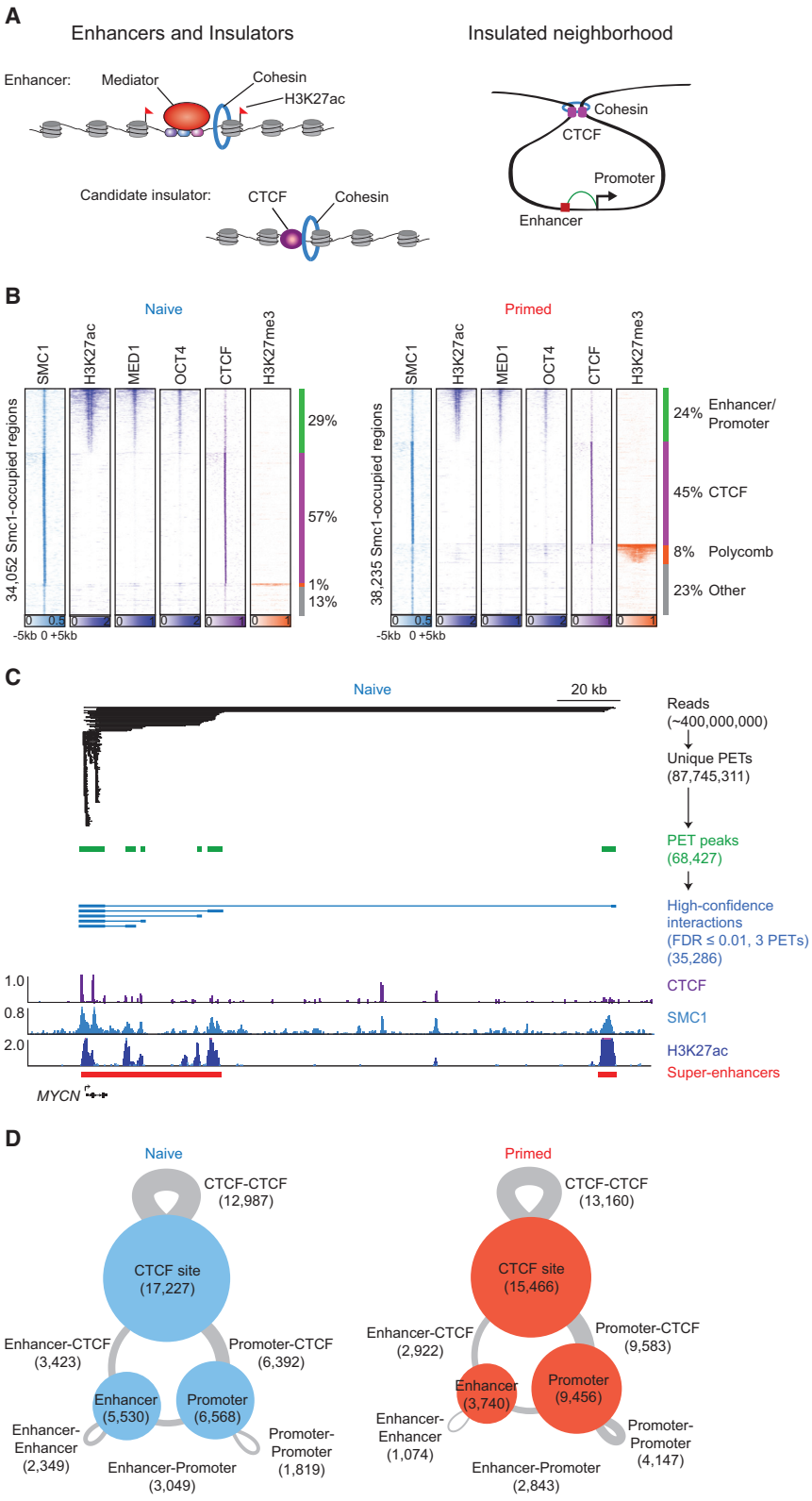


Figure 1. Components of the 3D Regulatory Landscape

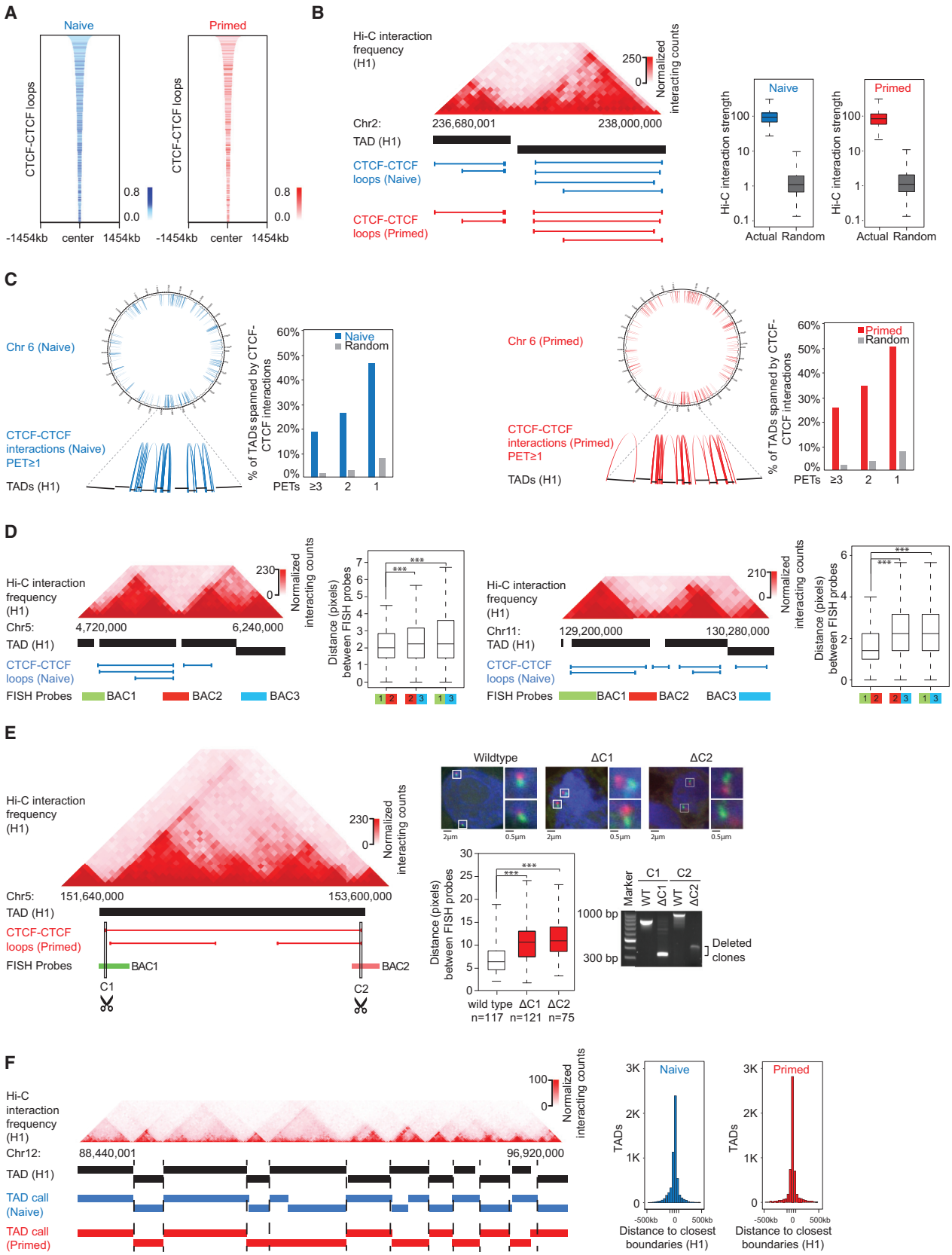
(A) Enhancers and insulators (left). Enhancers are occupied by transcription factors, mediator, and cohesin, and their associated nucleosomes are marked by H3K27ac. Candidate insulators are occupied by CTCF and cohesin. A model shows insulated neighborhoods formed by cohesin-associated CTCF-CTCF interactions,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

(B) Heatmap representation of ChIP-seq data for H3K27ac, MED1, OCT4, CTCF, and H3K27me3 at SMC1-occupied regions in naive (left) and primed (right) hESCs. Read density is displayed within a 10-kb window and color-scale intensities are shown in reads per million mapped reads per base pair. Cohesin occupies three classes of sites as follows: 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 the Supplemental 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) and primed (right) hESCs. CTCF-binding sites, enhancers, and promoters involved in cohesin-associated interactions are indicated as circles, and the size of circles corresponds to the number of sites. The interactions between two regions are indicated as gray lines, and the size of lines corresponds to the number of interactions.

See also Figure S1 and Tables S1, S2, and S3.



(legend on next page)

the Supplemental 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 Dowen et al. (2014) analysis pipeline (Supplemental 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 also were 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 to >800 kb and contained 0–24 protein-coding genes, with a median of 200 kb and one protein-coding gene per loop. Similar numbers were obtained for the primed hESCs (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 (Dowen et al., 2014; Rao et al., 2014). A comparison of the CTCF-CTCF loops identified here with TADs identified previously in H1 hESCs (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 correlated 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 were 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 fluorescence in situ hybridization (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 were 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 showed that loss of either binding site caused a separation between the two boundary regions that were 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

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 heatmap 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 heatmap (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 boxplot plotted in log scale; 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 PET) 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, and $\geq$3 PETs) were used. Random shuffling of TAD locations (100 iterations) served 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 in (B). The green, red, and blue bars indicate the location of BAC probes used for DNA FISH at each locus. Boxplots of minimal normalized distances between pairs of loci generated from >1,500 FISH probe spots per condition are displayed with the corresponding probe pairs labeled below. Significance was determined using the Mann-Whitney test (***$p < 10^{-28}$). Images were obtained using a 40× 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 in (B). The green and red bars indicate the location of BAC probes used for DNA FISH. The scissor-marked regions (C1 and C2) were deleted by CRISPR-mediated deletion. Examples of two-color DNA FISH images are shown (right), and the quantification of distance between green and red probes is displayed with a bar graph below. Significance was determined using the Mann-Whitney test (***$p < 10^{-13}$). Images were obtained using a 100× objective; n = number of alleles quantified for each sample. The genotyping PCR data also are displayed (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 is shown (left). A global analysis indicates that the cohesin ChIA-PET- and Hi-C-derived TAD boundaries are close (right).

See also Figure S2 and Table S3.

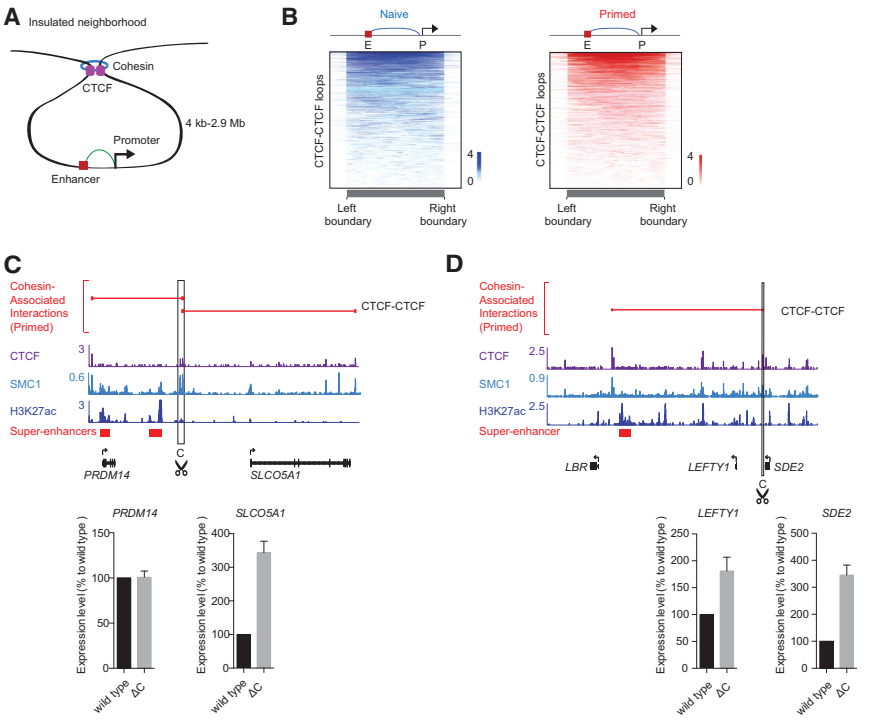


Figure 3. Putative Insulated Neighborhoods in hESCs

(A) Schematic of insulated neighborhood is shown. (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. Gene expression was assayed with at least three technical replicates and is displayed as the mean with SD.

See also Figure S3.

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 mESC TADs using previously generated mESC 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., 2012, 2015; 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 we found a high degree of overlap in predicted loops (Figures S2E–S2H). 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 mESCs 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 (Dowen 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 originated within a CTCF-CTCF loop almost invariably ended within the boundaries of the loop in naive and primed hESCs (Figure S3A). Furthermore, we found that enhancers generally interacted 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 it 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 (Figures 3C, 3D, and 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 mESCs and hESCs 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 mESCs (Dowen et al., 2014) have counterparts in the hESCs studied here, and these CTCF-CTCF loops contain human homologs 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 previously have 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 here we show

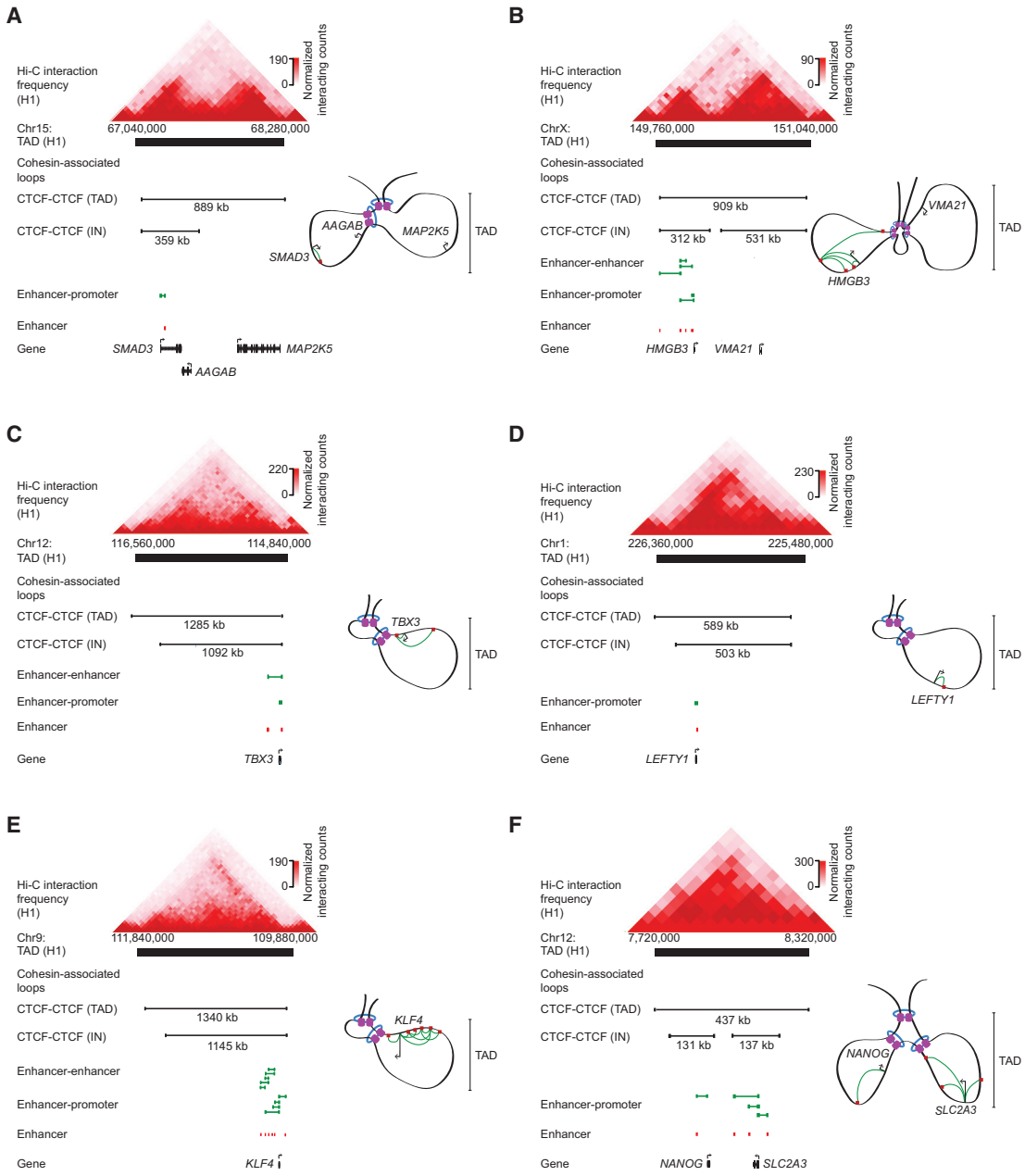


Figure 4. 3D Regulatory Structures of TADs Containing Key Pluripotency Genes

(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) are 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 (Supplemental Experimental Procedures), and a subset of genes present in these loops is shown for simplicity.

See also Figure S4.

a subset that illustrates common themes (Figures 4 and 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 were not saturated and some lower-confidence

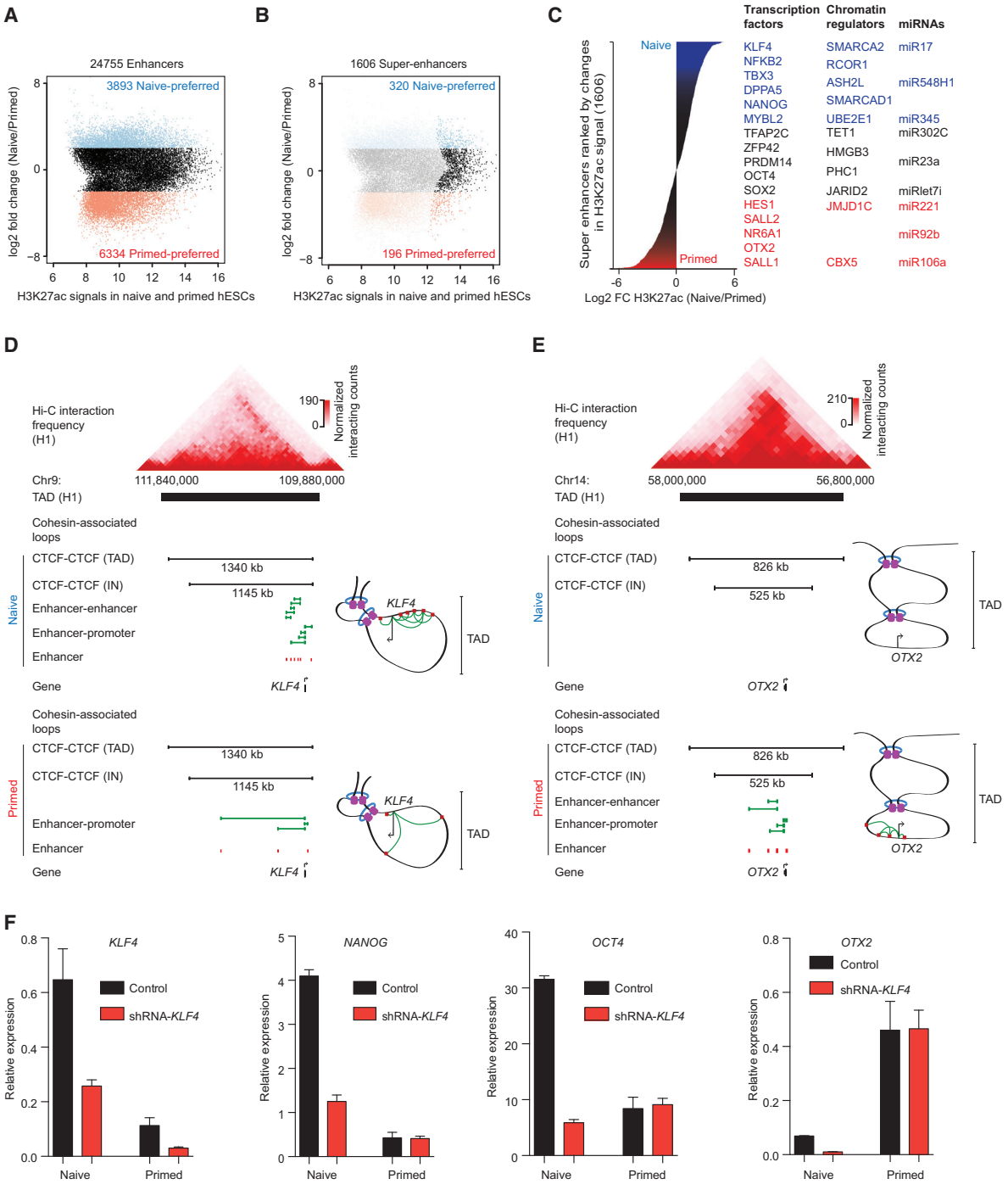


Figure 5. Differential Enhancer Landscape Reveals Key Transcription Factors, Chromatin Regulators, and MicroRNAs in Naive and Primed Pluripotencies

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

(B) Scatterplot comparison shows 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 microRNAs (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.

(legend continued on next page)

data were 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 to 21 Mb and contain 0–768 genes. The median number of CTCF-CTCF insulated neighborhoods that occurred within each TAD was two and these ranged from 4 kb to 2.9 Mb and contained 0–24 protein-coding genes. In this Version 1.0 map of TADs containing insulated neighborhoods, the median neighborhood was 200 kb and contained one gene whose average size was ~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 (Figures 5B and 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 occurred in similar TAD CTCF-CTCF loop structures in naive and primed cells (Figures 5D, 5E, and S5A–S5D). 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 (Figures S5A–S5D). 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 small hairpin RNA (shRNA) in both naive and primed cells, and the results showed 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 hESCs 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 (Figures S5E and 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 reported previously (Cuddapah et al., 2009; Dowen et al., 2014; Heidari et al., 2014; Kim et al., 2007; Phillips-Cremins 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., 2012, 2015; Rao et al., 2014; Vietri Rudan et al., 2015) led us to investigate the extent to which CTCF binding at loop anchors is similarly conserved. Analysis of CTCF-binding sites across ten primates indicated 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 (Figures 6A and S6A). A similar analysis showed that the CTCF DNA sequence motif in hESC loop anchor regions is highly conserved in primates and vertebrates (Figures 6B, 6C, and 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 it 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 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

(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), is shown.

(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. Gene expression was assayed with at least three technical replicates and is displayed as the mean with SD.

See also Figure S5 and Table S4.

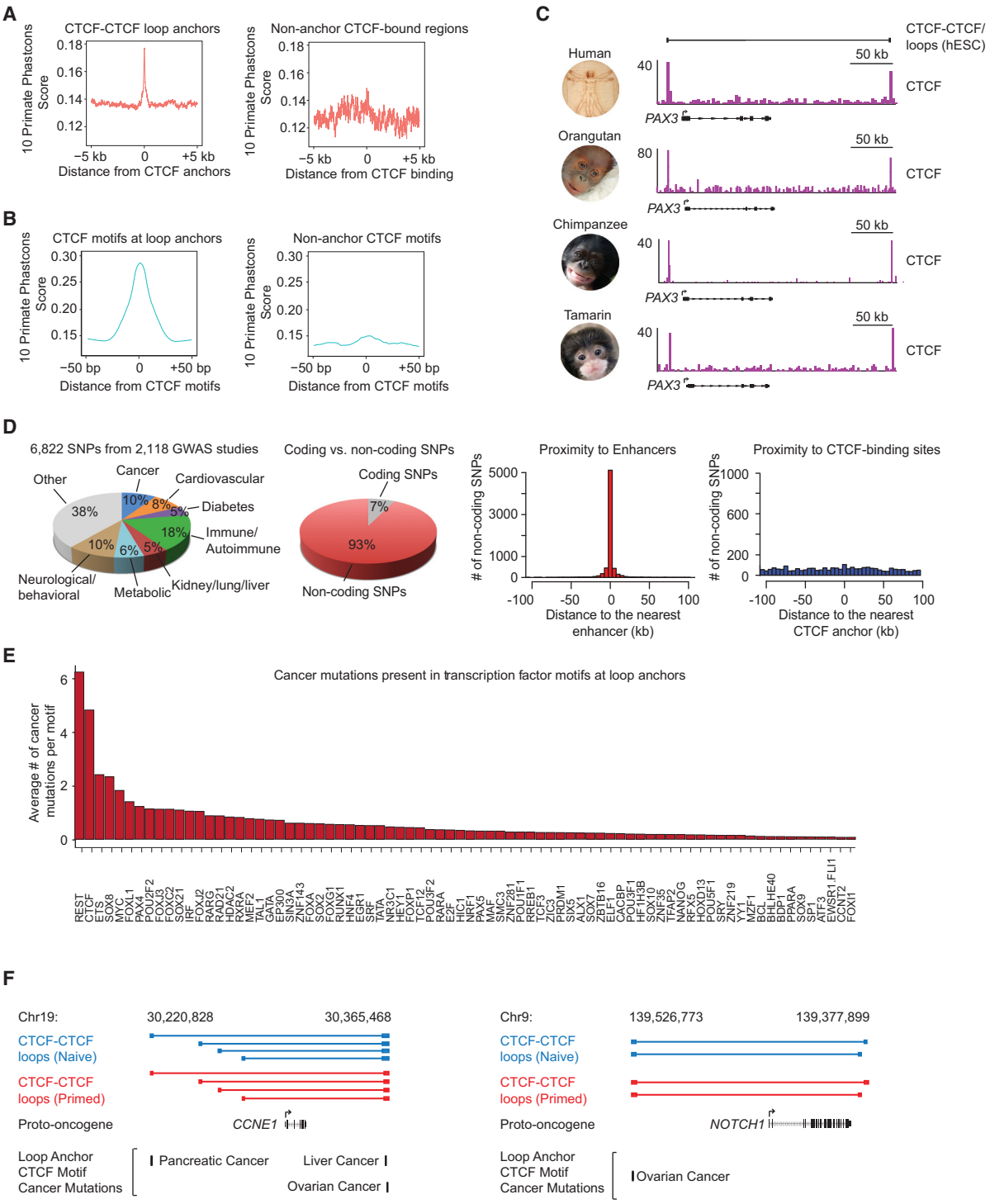


Figure 6. Conservation of 3D Structure and Associations with Disease

(A and B) DNA sequence in anchor regions (A) and the CTCF DNA sequence motif (B) of CTCF-CTCF loops in hESCs are 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 shows 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 (GWASs) and SNP association with enhancer and CTCF anchor regions in hESCs. Pie chart shows 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 is shown (middle left). Location of all noncoding trait-associated SNPs relative to all enhancers identified in 86

(legend continued on next page)

contribute to congenital diseases have been reported (Giorgio et al., 2015; Lupiáñez 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 7,307 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 cancers (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 hESCs in two pluripotent states and new insights into the relationships between chromosome structure and gene regulation. The naive and primed states 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 they 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). The data 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; Dowen 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 thus will 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 hESCs 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 also should prove valuable for identifying and further understanding genetic alterations that disrupt 3D structures and cause disease.

human cell and tissue samples is shown. The 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 is shown (right).

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

(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 ICGC are displayed along with the cancers from which these were sequenced.

See also Figure S6 and Table S5.

EXPERIMENTAL PROCEDURES

Additional information and details regarding this work may be found in the Supplemental Experimental Produces.

Cell Culture

Naive and primed hESCs were cultured as described previously (Theunissen et al., 2014). Detailed information is provided in the Supplemental Experimental Procedures.

ChIP-Seq Library Generation and Sequencing

ChIP was performed as previously described (Ji et al., 2015). Naive or primed hESCs (50 million) 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 Laboratories, A300-793A), and anti-OCT4 (Santa Cruz Biotechnology, sc-8628). For each ChIP, 5 μg antibody and 50 μl protein G Dynabeads (Life Technologies, 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 (Dowen et al., 2014). Naive or primed hESCs (400 million) 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 μg anti-SMC1 antibody (Bethyl Labs, A300-055A) and 250 μl protein G Dynabeads (Life Technologies, 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 × 100 paired-end sequencing using Illumina HiSeq 2000. Details are described in the Supplemental Experimental Procedures.

Data Analysis

Analysis of ChIA-PET data was performed essentially as previously described (Dowen et al., 2014). ChIA-PET interactions and additional ChIA-PET analysis may be found in the Supplemental Experimental Procedures (Tables S3 and S6). The TAD-related analyses were performed as previously described (Dixon et al., 2012). Detailed information is given in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

The accession number for the raw and processed sequencing data reported in this paper is GEO: GSE69647.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.stem.2015.11.007.

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 naive and primed hESCs. 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., and D.B.D. performed genome editing experiments. S.S., G.P., D.B.D., and T.M. designed and performed high-throughput 3D DNA FISH experiments. 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.

ACKNOWLEDGMENTS

We thank Brian J. Abraham and Jill Dowen for data analysis. We thank Tom Volkert at the Whitehead Genome Technology Core for sequencing. We thank Wendy Salmon for assistance with confocal microscopy and Raaji Alagappan, Dongdong Fu, and Tenzin Lungjangwa for preparation of mouse embryonic fibroblasts. This research was in part supported by the Intramural Research Program of the 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.), Ludwig Graduate Fellowship Funds (A.S.W.), NIH 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.

Received: June 22, 2015
Revised: October 21, 2015
Accepted: November 9, 2015
Published: December 10, 2015

REFERENCES

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308.

Bell, A.C., West, A.G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell *98*, 387–396.

Benoist, C., and Chambon, P. (1981). In vivo sequence requirements of the SV40 early promotor region. Nature *290*, 304–310.

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

Calhoun, E.S., Jones, J.B., Ashfaq, R., Adsay, V., Baker, S.J., Valentine, V., Hempen, P.M., Hilgers, W., Yeo, C.J., Hruban, R.H., and Kern, S.E. (2003). BRAF and FBXW7 (CDC4, FBW7, AGO, SEL10) mutations in distinct subsets of pancreatic cancer: potential therapeutic targets. Am. J. Pathol. *163*, 1255–1260.

Chan, Y.-S., Göke, J., Ng, J.-H., Lu, X., Gonzales, K.A.U., Tan, C.-P., Tng, W.-Q., Hong, Z.-Z., Lim, Y.-S., and Ng, H.-H. (2013). Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. Cell Stem Cell *13*, 663–675.

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K., and Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. *19*, 24–32.

de Graaf, C.A., and van Steensel, B. (2013). Chromatin organization: form to function. Curr. Opin. Genet. Dev. *23*, 185–190.

de Laat, W., and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. Nature *502*, 499–506.

De Los Angeles, A., Ferrari, F., Xi, R., Fujiwara, Y., Benvenisty, N., Deng, H., Hochedlinger, K., Jaenisch, R., Lee, S., Leitch, H.G., et al. (2015). Hallmarks of pluripotency. Nature *525*, 469–478.

Dekker, J. (2014). Two ways to fold the genome during the cell cycle: insights obtained with chromosome conformation capture. Epigenetics Chromatin *7*, 25.

DeMare, L.E., Leng, J., Cotney, J., Reilly, S.K., Yin, J., Sarro, R., and Noonan, J.P. (2013). The genomic landscape of cohesin-associated chromatin interactions. Genome Res. *23*, 1224–1234.

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

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Dorsett, D., and Merkenschlager, M. (2013). Cohesin at active genes: a unifying theme for cohesin and gene expression from model organisms to humans. Curr. Opin. Cell Biol. *25*, 327–333.

Dowen, J.M., Bilodeau, S., Orlando, D.A., Hübner, M.R., Abraham, B.J., Spector, D.L., and Young, R.A. (2013). Multiple structural maintenance of

chromosome complexes at transcriptional regulatory elements. Stem Cell Reports *1*, 371–378.

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

Doyle, B., Fudenberg, G., Imakaev, M., and Mirny, L.A. (2014). Chromatin loops as allosteric modulators of enhancer-promoter interactions. PLoS Comput. Biol. *10*, e1003867.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al.; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Etemadmoghadam, D., Weir, B.A., Au-Yeung, G., Alsop, K., Mitchell, G., George, J., Davis, S., D'Andrea, A.D., Simpson, K., Hahn, W.C., and Bowtell, D.D.; Australian Ovarian Cancer Study Group (2013). Synthetic lethality between CCNE1 amplification and loss of BRCA1. Proc. Natl. Acad. Sci. USA *110*, 19489–19494.

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

Gafni, O., Weinberger, L., Mansour, A.A., Manor, Y.S., Chomsky, E., Ben-Yosef, D., Kalma, Y., Viukov, S., Maza, I., Zviran, A., et al. (2013). Derivation of novel human ground state naive pluripotent stem cells. Nature *504*, 282–286.

Geyer, P.K., and Corces, V.G. (1992). DNA position-specific repression of transcription by a Drosophila zinc finger protein. Genes Dev. *6*, 1865–1873.

Giorgio, E., Robyr, D., Spielmann, M., Ferrero, E., Di Gregorio, E., Imperiale, D., Vaula, G., Stamoulis, G., Santoni, F., Atzori, C., et al. (2015). A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). Hum. Mol. Genet. *24*, 3143–3154.

Gómez-Díaz, E., and Corces, V.G. (2014). Architectural proteins: regulators of 3D genome organization in cell fate. Trends Cell Biol. *24*, 703–711.

Gorkin, D.U., Leung, D., and Ren, B. (2014). The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell *14*, 762–775.

Gruber, S., Haering, C.H., and Nasmyth, K. (2003). Chromosomal cohesin forms a ring. Cell *112*, 765–777.

Gruss, P., Dhar, R., and Khoury, G. (1981). Simian virus 40 tandem repeated sequences as an element of the early promoter. Proc. Natl. Acad. Sci. USA *78*, 943–947.

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

Hackett, J.A., and Surani, M.A. (2014). Regulatory principles of pluripotency: from the ground state up. Cell Stem Cell *15*, 416–430.

Hadjur, S., Williams, L.M., Ryan, N.K., Cobb, B.S., Sexton, T., Fraser, P., Fisher, A.G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature *460*, 410–413.

Haering, C.H., Löwe, J., Hochwagen, A., and Nasmyth, K. (2002). Molecular architecture of SMC proteins and the yeast cohesin complex. Mol. Cell *9*, 773–788.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. Nat. Genet. *43*, 630–638.

Heidari, N., Phanstiel, D.H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M.Q., and Snyder, M.P. (2014). Genome-wide map of regulatory interactions in the human genome. Genome Res. *24*, 1905–1917.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol. *16*, 144–154.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Huang, K., Maruyama, T., and Fan, G. (2014). The naive state of human pluripotent stem cells: a synthesis of stem cell and preimplantation embryo transcriptome analyses. Cell Stem Cell *15*, 410–415.

Ji, X., Dadon, D.B., Abraham, B.J., Lee, T.I., Jaenisch, R., Bradner, J.E., and Young, R.A. (2015). Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. Proc. Natl. Acad. Sci. USA *112*, 3841–3846.

Jung, Y.J., Lee, K.H., Choi, D.W., Han, C.J., Jeong, S.H., Kim, K.C., Oh, J.W., Park, T.K., and Kim, C.M. (2001). Reciprocal expressions of cyclin E and cyclin D1 in hepatocellular carcinoma. Cancer Lett. *168*, 57–63.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A.E., Ristolainen, H., Hänninen, U.A., Cajuso, T., et al. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. Nat. Genet. *47*, 818–821.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Lee, T.I., and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237–1251.

Levine, M., Cattoglio, C., and Tjian, R. (2014). Looping back to leap forward: transcription enters a new era. Cell *157*, 13–25.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Martello, G., and Smith, A. (2014). The nature of embryonic stem cells. Annu. Rev. Cell Dev. Biol. *30*, 647–675.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Merkenschlager, M., and Odom, D.T. (2013). CTCF and cohesin: linking gene regulatory elements with their targets. Cell *152*, 1285–1297.

Narendra, V., Rocha, P.P., An, D., Raviram, R., Skok, J.A., Mazzoni, E.O., and Reinberg, D. (2015). Transcription. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science *347*, 1017–1021.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., et al. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell *132*, 422–433.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Plank, J.L., and Dean, A. (2014). Enhancer function: mechanistic and genome-wide insights come together. Mol. Cell *55*, 5–14.

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

Rose, S.L., Kunnimalaiyaan, M., Drenzek, J., and Seiler, N. (2010). Notch 1 signaling is active in ovarian cancer. Gynecol. Oncol. *117*, 130–133.

Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A., and Krumm, A. (2008). CTCF physically links cohesin to chromatin. Proc. Natl. Acad. Sci. USA *105*, 8309–8314.

Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Gonçalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P., and Odom, D.T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell *148*, 335–348.

Schwalie, P.C., Ward, M.C., Cain, C.E., Faure, A.J., Gilad, Y., Odom, D.T., and Flicek, P. (2013). Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. Genome Biol. *14*, R148.

Seitan, V.C., Faure, A.J., Zhan, Y., McCord, R.P., Lajoie, B.R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A.G., et al. (2013). Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Res. *23*, 2066–2077.

Smith, E., and Shilatifard, A. (2014). Enhancer biology and enhanceropathies. Nat. Struct. Mol. Biol. *21*, 210–219.

Sofueva, S., Yaffe, E., Chan, W.C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S.M., Schroth, G.P., Tanay, A., and Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. EMBO J. *32*, 3119–3129.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *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. (2014). Resetting transcription factor control circuitry toward ground-state pluripotency in human. Cell *158*, 1254–1269.

Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L., et al. (2014). Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. Cell Stem Cell *15*, 471–487.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. *10*, 1297–1309.

Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., et al. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. *22*, 1680–1688.

Ware, C.B., Nelson, A.M., Mecham, B., Hesson, J., Zhou, W., Jonlin, E.C., Jimenez-Caliani, A.J., Deng, X., Cavanaugh, C., Cook, S., et al. (2014). Derivation of naive human embryonic stem cells. Proc. Natl. Acad. Sci. USA *111*, 4484–4489.

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature *451*, 796–801.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature *518*, 556–559.

Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., et al. (2011). International Cancer Genome Consortium Data Portal-a one-stop shop for cancer genomics data (Database-the Journal of Biological Databases and Curation).

Zuin, J., Dixon, J.R., van der Reijden, M.I.J.A., Ye, Z., Kolovos, P., Brouwer, R.W.W., van de Corput, M.P.C., van de Werken, H.J.G., Knoch, T.A., van IJcken, W.F.J., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. USA *111*, 996–1001.

Cell Stem Cell

Supplemental Information

3D Chromosome Regulatory Landscape of Human Pluripotent Cells

Xiong Ji, Daniel B. Dadon, Benjamin E. Powell, Zi Peng Fan, Diego Borges-Rivera, Sigal Shachar, Abraham S. Weintraub, Denes Hnisz, Gianluca Pegoraro, Tong Ihn Lee, Tom Misteli, Rudolf Jaenisch, and Richard A. Young

Figure S1

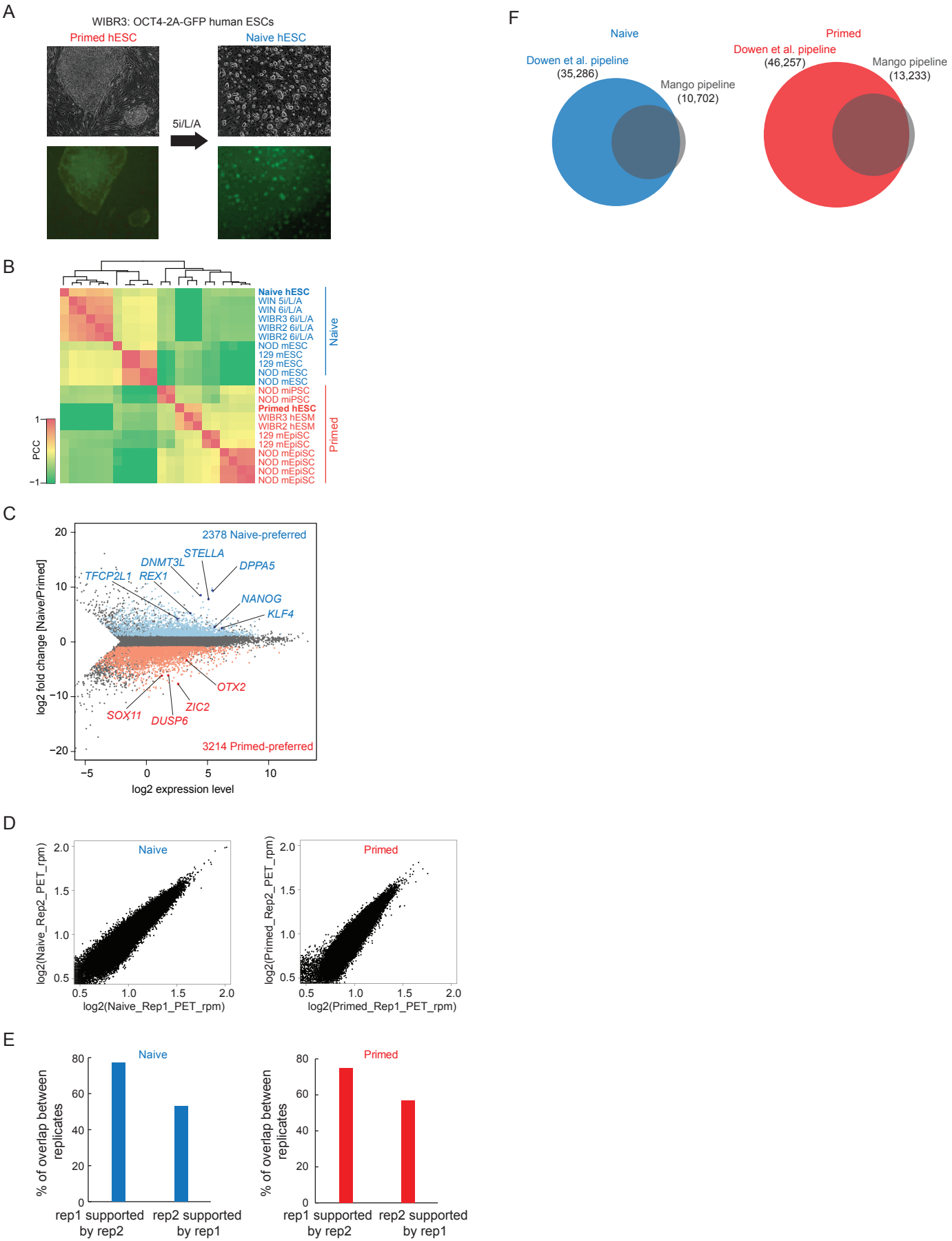

A
WIBR3: OCT4-2A-GFP human ESCs
Primed hESC
Naive hESC
5i/L/A
B
Naive hESC
WIN 5i/L/A
WIN 6i/L/A
WIBR3 6i/L/A
WIBR2 6i/L/A
WIBR2 6i/L/A
NOD mESC
129 mESC
129 mESC
NOD mESC
NOD mESC
NOD miPSC
NOD miPSC
Primed hESC
WIBR3 hESM
WIBR2 hESM
129 mEpiSC
129 mEpiSC
NOD mEpiSC
NOD mEpiSC
NOD mEpiSC
NOD mEpiSC
Naive
Primed
PCC
1
−1
C
2378 Naive-preferred
3214 Primed-preferred
STELLA
DNMT3L
DPPA5
TFCP2L1
REX1
NANOG
KLF4
OTX2
ZIC2
SOX11
DUSP6
log2 fold change [Naive/Primed]
log2 expression level
D
Naive
Primed
log2(Naive_Rep2_PET_rpm)
log2(Naive_Rep1_PET_rpm)
log2(Primed_Rep2_PET_rpm)
log2(Primed_Rep1_PET_rpm)
E
Naive
Primed
% of overlap between replicates
rep1 supported by rep2
rep2 supported by rep1
F
Naive
Dowen et al. pipeline
(35,286)
Mango pipeline
(10,702)
Primed
Dowen et al. pipeline
(46,257)
Mango pipeline
(13,233)

Figure S2

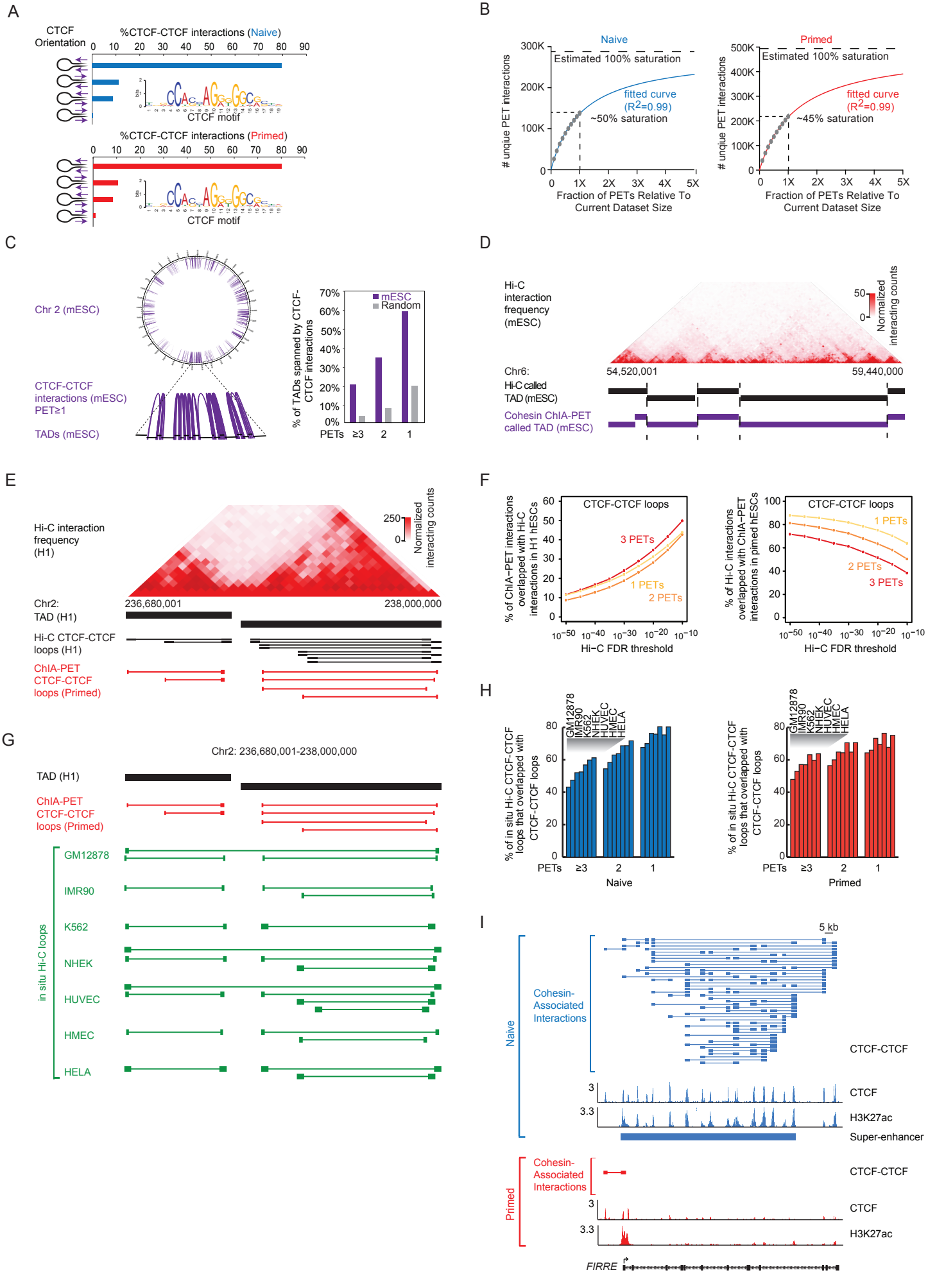

A
CTCF Orientation
%CTCF-CTCF interactions (Naive)
CTCF motif
%CTCF-CTCF interactions (Primed)
CTCF motif
B
Naive
Estimated 100% saturation
fitted curve (R^2=0.99)
~50% saturation
unqiue PET interactions
Fraction of PETs Relative To Current Dataset Size
Primed
Estimated 100% saturation
fitted curve (R^2=0.99)
~45% saturation
unqiue PET interactions
Fraction of PETs Relative To Current Dataset Size
C
Chr 2 (mESC)
CTCF-CTCF interactions (mESC) PET≥1
TADs (mESC)
% of TADs spanned by CTCF-CTCF interactions
mESC
Random
PETs ≥3 2 1
D
Hi-C interaction frequency (mESC)
Normalized interacting counts
Chr6:
54,520,001
59,440,000
Hi-C called TAD (mESC)
Cohesin ChIA-PET called TAD (mESC)
E
Hi-C interaction frequency (H1)
Normalized interacting counts
Chr2:
236,680,001
238,000,000
TAD (H1)
Hi-C CTCF-CTCF loops (H1)
ChIA-PET CTCF-CTCF loops (Primed)
F
% of ChIA-PET interactions overlapped with Hi-C interactions in H1 hESCs
CTCF-CTCF loops
3 PETs
1 PETs
2 PETs
Hi-C FDR threshold
% of Hi-C interactions overlapped with ChIA-PET interactions in pimed hESCs
CTCF-CTCF loops
1 PETs
2 PETs
3 PETs
Hi-C FDR threshold
G
Chr2: 236,680,001-238,000,000
TAD (H1)
ChIA-PET CTCF-CTCF loops (Primed)
in situ Hi-C loops
GM12878
IMR90
K562
NHEK
HUVEC
HMEC
HELA
H
% of in situ Hi-C CTCF-CTCF loops that overlapped with CTCF-CTCF loops
GM12878
IMR90
K562
NHEK
HUVEC
HMEC
HELA
PETs ≥3 2 1
Naive
% of in situ Hi-C CTCF-CTCF loops that overlapped with CTCF-CTCF loops
Primed
I
5 kb
Naive
Cohesin-Associated Interactions
CTCF-CTCF
CTCF
H3K27ac
Super-enhancer
Primed
Cohesin-Associated Interactions
CTCF-CTCF
CTCF
H3K27ac
FIRRE

Figure S3

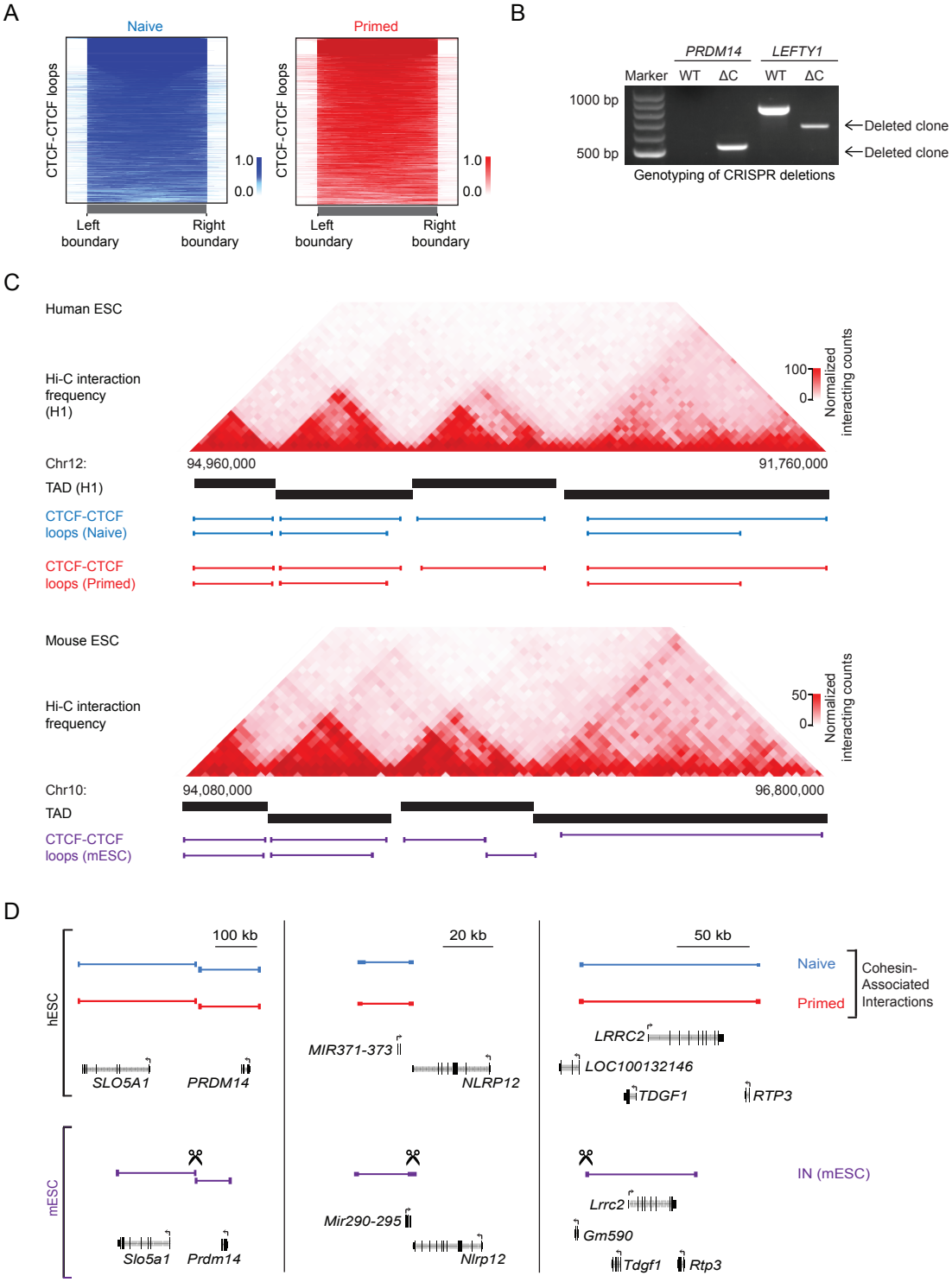

A
Naive
Primed
CTCF-CTCF loops
1.0
0.0
Left boundary
Right boundary
B
PRDM14
LEFTY1
Marker
WT
ΔC
WT
ΔC
1000 bp
500 bp
Deleted clone
Deleted clone
Genotyping of CRISPR deletions
C
Human ESC
Hi-C interaction frequency (H1)
100
0
Normalized interacting counts
Chr12:
94,960,000
91,760,000
TAD (H1)
CTCF-CTCF loops (Naive)
CTCF-CTCF loops (Primed)
Mouse ESC
Hi-C interaction frequency
50
0
Normalized interacting counts
Chr10:
94,080,000
96,800,000
TAD
CTCF-CTCF loops (mESC)
D
hESC
mESC
100 kb
20 kb
50 kb
SLO5A1
PRDM14
MIR371-373
NLRP12
LRRC2
LOC100132146
TDGF1
RTP3
Slo5a1
Prdm14
Mir290-295
Nlrp12
Lrrc2
Gm590
Tdgf1
Rtp3
Naive
Primed
Cohesin-Associated Interactions
IN (mESC)

Figure S4

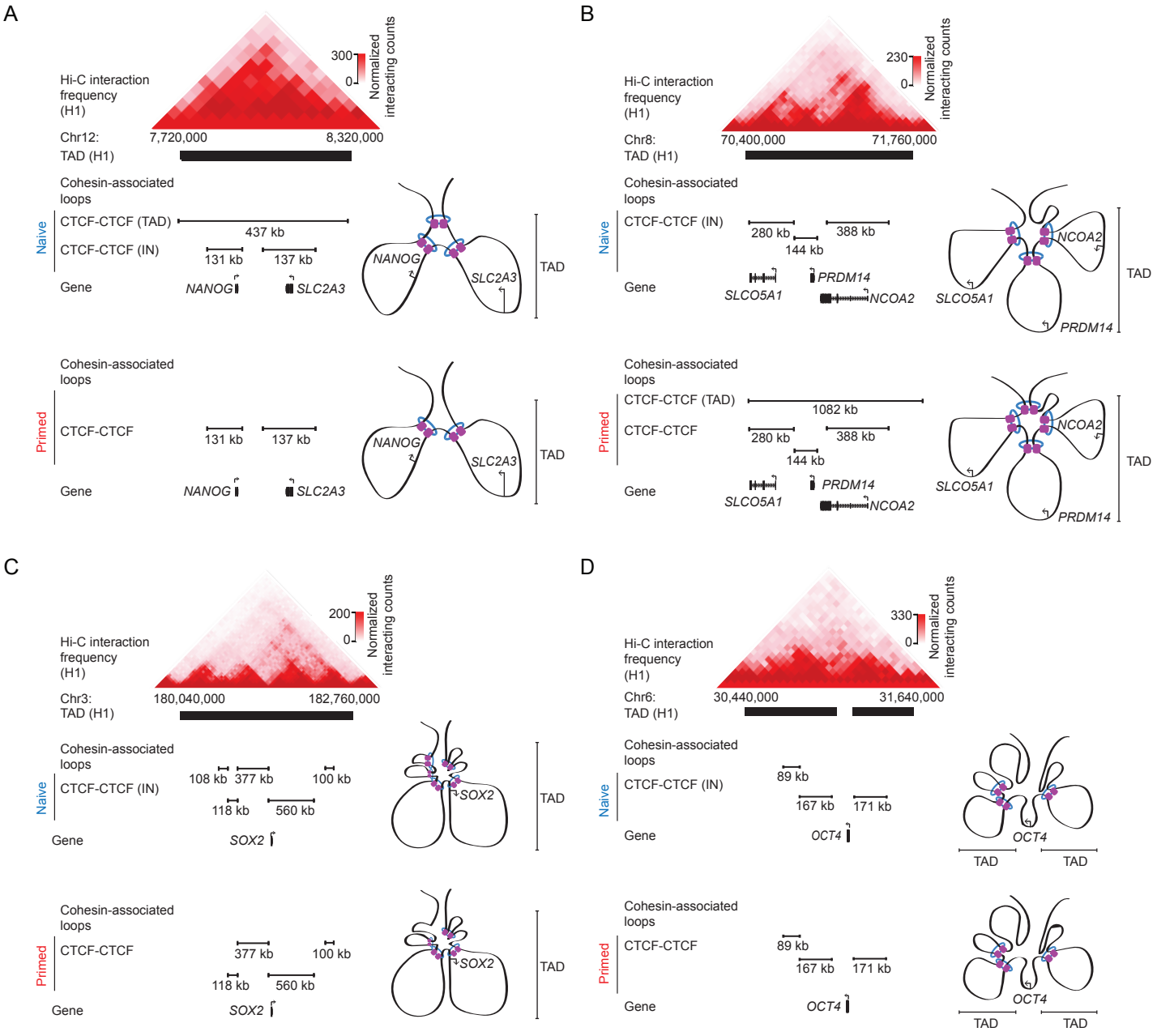

A
Hi-C interaction frequency (H1)
300
0
Normalized interacting counts
Chr12:
7,720,000
8,320,000
TAD (H1)
Cohesin-associated loops
Naive
CTCF-CTCF (TAD)
437 kb
CTCF-CTCF (IN)
131 kb
137 kb
Gene
NANOG
SLC2A3
NANOG
SLC2A3
TAD
Primed
CTCF-CTCF
131 kb
137 kb
Gene
NANOG
SLC2A3
NANOG
SLC2A3
TAD
B
Hi-C interaction frequency (H1)
230
0
Normalized interacting counts
Chr8:
70,400,000
71,760,000
TAD (H1)
Cohesin-associated loops
Naive
CTCF-CTCF (IN)
280 kb
144 kb
388 kb
Gene
SLCO5A1
PRDM14
NCOA2
SLCO5A1
NCOA2
PRDM14
TAD
Primed
CTCF-CTCF (TAD)
1082 kb
CTCF-CTCF
280 kb
144 kb
388 kb
Gene
SLCO5A1
PRDM14
NCOA2
SLCO5A1
NCOA2
PRDM14
TAD
C
Hi-C interaction frequency (H1)
200
0
Normalized interacting counts
Chr3:
180,040,000
182,760,000
TAD (H1)
Cohesin-associated loops
Naive
CTCF-CTCF (IN)
108 kb
377 kb
100 kb
118 kb
560 kb
Gene
SOX2
SOX2
TAD
Primed
CTCF-CTCF
377 kb
100 kb
118 kb
560 kb
Gene
SOX2
SOX2
TAD
D
Hi-C interaction frequency (H1)
330
0
Normalized interacting counts
Chr6:
30,440,000
31,640,000
TAD (H1)
Cohesin-associated loops
Naive
CTCF-CTCF (IN)
89 kb
167 kb
171 kb
Gene
OCT4
OCT4
TAD
TAD
Primed
CTCF-CTCF
89 kb
167 kb
171 kb
Gene
OCT4
OCT4
TAD
TAD

Figure S5

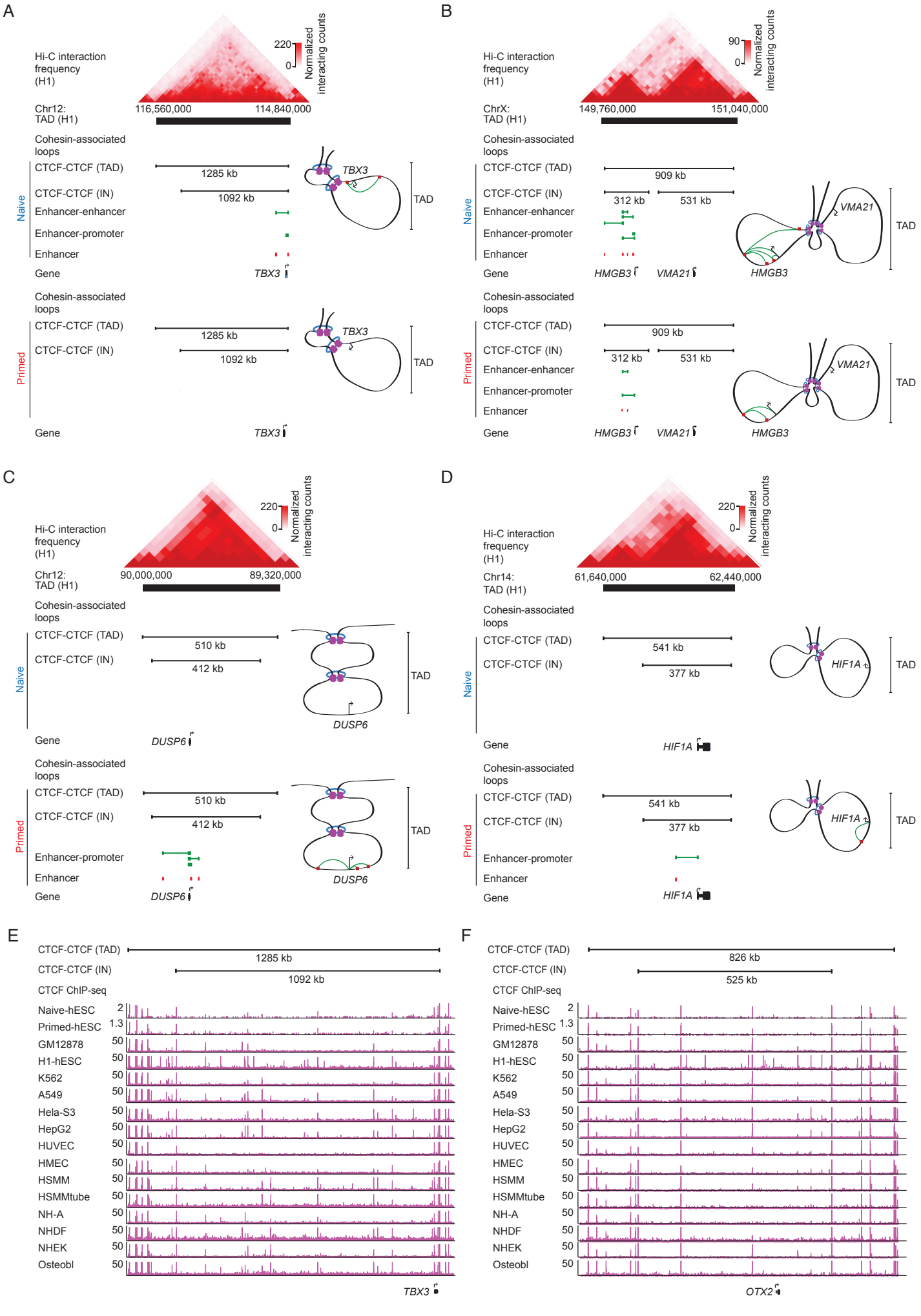
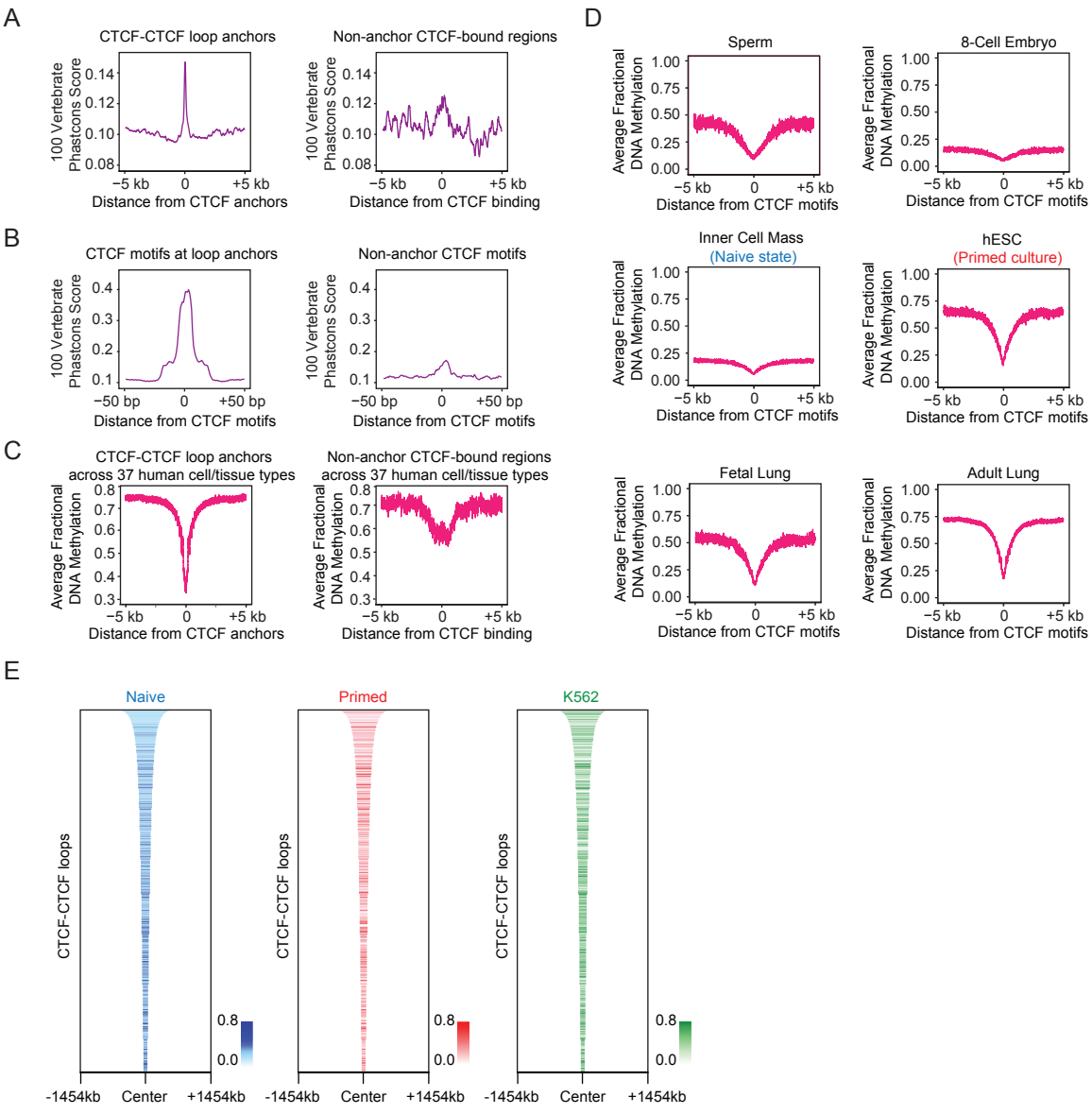

A
Hi-C interaction frequency (H1)
220
0
Normalized interacting counts
Chr12: TAD (H1)
116,560,000
114,840,000
Cohesin-associated loops
Naive
CTCF-CTCF (TAD)
1285 kb
CTCF-CTCF (IN)
1092 kb
Enhancer-enhancer
Enhancer-promoter
Enhancer
Gene
TBX3
TAD
Primed
B
90
ChrX: TAD (H1)
149,760,000
151,040,000
909 kb
312 kb
531 kb
HMGB3
VMA21
C
90,000,000
89,320,000
510 kb
412 kb
DUSP6
D
61,640,000
62,440,000
541 kb
377 kb
HIF1A
E
826 kb
525 kb
CTCF ChIP-seq
Naive-hESC
Primed-hESC
GM12878
H1-hESC
K562
A549
Hela-S3
HepG2
HUVEC
HMEC
HSMM
HSMMtube
NH-A
NHDF
NHEK
Osteobl
OTX2

Figure S6


A
CTCF-CTCF loop anchors
Non-anchor CTCF-bound regions
100 Vertebrate Phastcons Score
Distance from CTCF anchors
Distance from CTCF binding
B
CTCF motifs at loop anchors
Non-anchor CTCF motifs
Distance from CTCF motifs
C
CTCF-CTCF loop anchors across 37 human cell/tissue types
Non-anchor CTCF-bound regions across 37 human cell/tissue types
Average Fractional DNA Methylation
D
Sperm
8-Cell Embryo
Inner Cell Mass (Naive state)
hESC (Primed culture)
Fetal Lung
Adult Lung
E
Naive
Primed
K562
CTCF-CTCF loops
-1454kb
Center
+1454kb

Supplemental Figure Legends

Figure S1. Human ESCs, expression analysis and ChIA-PET data

(A) Phase and fluorescence images of primed hESCs (endogenous OCT4-2A-GFP) and emerging naive colonies induced by treating these primed hESCs with 5i/L/A medium for 10 days. 40x magnification.
(B) Cross-species hierarchical clustering of expression datasets from naive and primed pluripotent cells in both mouse and human highlights the similarity of our datasets to the existing datasets for these cell states in human and mouse samples.
(C) Comparison between the transcriptomes of naive and primed hESCs reveals common and differentially expressed genes.
(D) Correlation analysis for two replicates of cohesin ChIA-PET dataset were displayed by scatter plot.
(E) Percentage of cohesin ChIA-PET interactions that overlap in replicates in naive and primed hESCs.
(F) The overlap of ChIA-PET interactions called by the Dowen et al., 2014 pipeline and the Mango pipeline in naive and primed hESCs.
Related to Figure 1

Figure S2. Cohesin-associated interactions are largely responsible for the organization of TADs

(A) CTCF motif orientation analysis of CTCF-CTCF loops. The percentage of each type of CTCF motif orientation is shown in a bar graph.
(B) Saturation analysis for the cohesin ChIA-PET datasets in naive (left panel) and primed (right panel) hESCs.
(C) CTCF-CTCF loops span many TADs identified using Hi-C data in mESCs. Chromosome 2 is displayed as a circos plot in mESCs, with a zoomed in region below. CTCF-CTCF loops (≥1 PETs) are indicated as purple arcs. The bar graphs show percentages of TADs spanned by CTCF-CTCF loops when various confidence thresholds (1, 2, ≥3 PETs) were used. As a background control, we used random shuffling of TAD locations (100 iterations).
(D) Cohesin ChIA-PET data can be used to discover TADs in mESCs. A comparison of TADs derived with the same algorithm from Hi-C data in mESCs (Dixon et al., 2012) and cohesin ChIA-PET data (mESCs) for a portion of chromosome 6.
(E) Comparison of CTCF-CTCF loops from H1 hESC Hi-C dataset to primed hESC ChIA-PET CTCF-CTCF loops at a locus on chromosome 2. Normalized Hi-C interaction frequencies in H1 hESCs are displayed as a two dimensional heatmap. CTCF-CTCF loops derived from H1 hESC Hi-C dataset are colored in black, ChIA-PET CTCF-CTCF loops are colored in red.
(F) The percent of ChIA-PET CTCF-CTCF loops in primed hESCs present in the Hi-C CTCF loops in H1 hESCs (or vice versa) are plotted as a function of significance thresholds (false discovery rate–FDR) for calling Hi-C interactions displayed as line plots for ChIA-PET CTCF-CTCF loops when various thresholds (1, 2, ≥3 PETs) were used.
(G) Comparison of primed hESC ChIA-PET CTCF-CTCF loops to CTCF-CTCF interactions (green lines) derived from in situ Hi-C in GM12878, IMR90, K562, NHEK, HUVEC, HMEC, and HeLa cells (Rao et al., 2014).
(H) Bar plot indicating the percent of in situ Hi-C CTCF-CTCF loops that overlap with ChIA-PET CTCF-CTCF loops in naive and primed hESCs when various thresholds (1, 2, ≥3 PETs) were used.
(I) Cohesin-associated interactions at the *FIRRE* locus are shown. The cohesin-associated interactions are shown as blue lines (naive) and red lines (primed). The

ChIP-seq binding for CTCF and H3K27ac are shown. The blue bar indicates a super-enhancer in naive hESCs.
Related to Figure 2

Figure S3. Cohesin ChIA-PET interactions
(A) Heatmap showing that cohesin ChIA-PET interactions occur predominantly within CTCF-CTCF loops that define putative insulated neighborhoods in hESCs. See the section entitled "Heatmap Representation of High-confidence ChIA-PET Interactions" for details. The color bar indicates normalized high-confidence interactions per loop.
(B) Genotyping for CRISPR-mediated deletions of anchors of CTCF-CTCF loops constraining the super-enhancer associated genes *PRDM14* and *LEFTY1*.
(C) CTCF-CTCF loops tend to be preserved in syntenic regions of human and mouse ESCs. Heatmaps of Hi-C interaction frequencies in H1 hESCs (upper panel) or mESCs (lower panel) are displayed to illustrate a syntenic region (human chr12: 91,760,000-94,960,000, mouse chr10: 94,080,000-96,800,000). Shared CTCF-CTCF loops are indicated as blue lines (naive hESCs) and red lines (primed hESCs). Mouse CTCF-CTCF loops are shown below in purple.
(D) Multiple loops forming insulated neighborhoods (IN) in mESCs whose CTCF boundaries were previously shown to be necessary for insulator function are preserved in human ESCs. The scissor-marked regions were deleted by CRISPR/Cas9 editing in mESCs, which caused local mis-regulation of gene expression (Dowen et al., 2014).
Related to Figure 3

Figure S4. 3D structures of TADs containing key pluripotency genes in naive and primed hESCs
(A-D) Schematics of 3D structure for TADs containing *NANOG*, *PRDM14*, *SOX2* and *OCT4* in naive and primed 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 and insulated neighborhood spanning CTCF 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. These schematics represent one potential conformation of TADs, but because the underlying data originates with a population of cells, additional conformations are possible.
Related to Figure 4

Figure S5. Differential regulated genes occur in 3D regulatory structures of TADs in naive and primed hESCs
(A-D) Schematics of 3D structure for TADs containing *TBX3*, *HMGB3*, *DUSP6* and *HIF1A* in naive and primed 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. These schematics represent one potential conformation of TADs, but because the underlying data originates with a population of cells, additional conformations are possible.
(E) CTCF binding to the TAD and putative insulated neighborhood (IN) anchor sites is preserved in a broad spectrum of human cell types in the domain containing *TBX3*.
(F) CTCF binding to the TAD and putative insulated neighborhood (IN) anchor sites is preserved in a broad spectrum of human cell types in the domain containing *OTX2*.
Related to Figure 5

Figure S6. Conservation of hESC CTCF loop anchors
(A) DNA sequence in anchor regions of CTCF-CTCF loops in hESCs is more conserved in vertebrates than DNA sequence in hESC regions bound by CTCF that do not serve as loop anchors.
(B) The CTCF sequence motif at sites used to anchor DNA loops in hESCs is more conserved in vertebrates than that motif at sites that do not serve as loop anchors in hESCs.
(C) Anchor regions of hESC CTCF-CTCF loops are hypomethylated relative to regions bound by CTCF that do not serve as anchors.
(D) DNA hypomethylation at CTCF-CTCF loop anchors is constitutive throughout the life cycle of humans.
(E) CTCF loops are largely preserved between normal cells (hESCs) and cancer cells (K562). The 9,344 CTCF-CTCF loops that define putative insulated neighborhoods in naive hESCs were ranked by size and shown. The color bar indicates normalized PET-signal at these CTCF-CTCF loops.
Related to Figure 6

Tables
Table S1. RNA-seq gene expression in naive and primed hESCs. Related to Figure 1.
Table S2. SMC1 ChIA-PET, H3K27ac ChIP-seq, CTCF ChIP-seq peaks for hESCs. Related to Figure 1.
Table S3. High confidence SMC1 ChIA-PET interactions for naive and primed hESCs. Related to Figure 1, 2.
Table S4. Differential Super-enhancers and differential CTCF-CTCF loops between naive and primed hESCs. Related to Figure 5
Table S5. Cancer mutations identified at CTCF motifs within CTCF-CTCF loop anchors. Related to Figure 6
Table S6. Mango-called high confidence SMC1 ChIA-PET interactions for naive and primed hESCs. Related to the experimental procedures.

Sequences used in this study

Gene	**Sequence**	**Application**
PRDM14 sgRNA 1	GTGACACTGTGCAGACCACT	sgRNA target sequence
PRDM14 sgRNA 2	ATAAGAAGGGTGGCCGGGCG	sgRNA target sequence
LEFTY1 sgRNA 1	AAGGTGGGTCTCACAGGATT	sgRNA target sequence
LEFTY1 sgRNA 2	GAAATAGGTAACCTTTTTAA	sgRNA target sequence
TAD L sgRNA 1	GGGGAGGTGCTCCGTACTTC	sgRNA target sequence
TAD L sgRNA 2	AAACAGCTGACAACATCGAA	sgRNA target sequence
TAD R sgRNA 1	GAGCCATCCGGTGGTAGATT	sgRNA target sequence
TAD R sgRNA 2	CAGAGTTGGTGACTCCGTAA	sgRNA target sequence
PRDM14 F	CCTGACATCTCAGTGCACGT	Genotype PCR

PRDM14 R	CCTTGCTCTATCGCCCAGTC	Genotype PCR
LEFTY1 F	AGCGGAAAACAACAGCAAAT	Genotype PCR
LEFTY1 R	GCAACTGAAGTGAGTGCATGA	Genotype PCR
TAD L F	TGGCACTAGATATTTGAGAGAAATTG	Genotype PCR
TAD L R	TCTTCCAGGTTCAACGCTCT	Genotype PCR
TAD R F	CAAGTCCTGGGTTCTCATCC	Genotype PCR
TAD R R	TTGAGATCCCAGGAGTGAGG	Genotype PCR
GAPDH F	CGAGATCCCTCCAAAATCAA	RT-qPCR
GAPDH R	ATCCACAGTCTTCTGGGTGG	RT-qPCR
PRDM14 F	ACACGCCTTTCCCGTCCTA	RT-qPCR
PRDM14 R	GGGCAGATCGTAGAGAGGCT	RT-qPCR
SLCO5A1 F	ACCTCAGCAAAACCTTCTCGG	RT-qPCR
SLCO5A1 R	GAGACCATTAACGCCTGGATG	RT-qPCR
LEFTY1 F	TGATCGTCAGCATCAAGGAG	RT-qPCR
LEFTY1 R	GAGCACAGAGCATTTGTCCA	RT-qPCR
SDE2 F	AGGATTCCGTCCTCAAAGGT	RT-qPCR
SDE2 R	TGGACCCTTCTGCAGTCTCT	RT-qPCR
KLF4 F	GATGGGGTCTGTGACTGGAT	RT-qPCR
KLF4 R	CCCCCAACTCACGGATATAA	RT-qPCR
NANOG F	GCAGAAGGCCTCAGCACCTA	RT-qPCR
NANOG R	AGGTTCCCAGTCGGGTTCA	RT-qPCR
OCT4 F	GCTCGAGAAGGATGTGGTCC	RT-qPCR
OCT4 R	CGTTGTGCATAGTCGCTGCT	RT-qPCR
OTX2 F	CAAAGTGAGACCTGCCAAAAAGA	RT-qPCR
OTX2 R	TGGACAAGGGATCTGACAGTG	RT-qPCR
BAC1 Probe	RP11-487J21	3D DNA FISH
BAC2 Probe	RP11-137I4	3D DNA FISH

Extended Experimental Procedures:

Cell Culture

Primed and naive hESCs were cultured as previously described (Theunissen et al., 2014). Primed hESCs were maintained on mitomycin C-inactivated MEF feeder layers and passaged every 7-10 days. When passaging primed hESCs, clumps of cells were partially dissociated with collagenase type IV (GIBCO, 17104-019), and then subjected to two sedimentation steps in stationary 50 ml tubes for 10 min at room temperature in primed hESC medium to remove single cells. Primed hESC medium (500 ml) consisted of 400 ml of Dulbecco's Modified Eagle Medium: Nutrient Mixture F-12 (DMEM/F12, Invitrogen, 11320), 75 ml Fetal Bovine Serum (FBS, Hyclone, SH30071.03HI), 25 ml KnockOut™ Serum Replacement (KSR, Invitrogen, 10828-028), supplemented with 1

mM glutamine (Invitrogen, 25030-024), 1% nonessential amino acids (Invitrogen, 11140-050), penicillin-streptomycin (Invitrogen, 15140-122), 0.1 mM β-mercaptoethanol (Sigma, M6250-100ML), and 4 ng/ml FGF2 (R&D systems, 233-FB-025).

For the induction of naive hESCs, primed hESCs were cultured for 24 hr in the primed hESC medium described above, further supplemented with 10 μM ROCK inhibitor Y-27632 (Stemgent, 04-0012). Colonies were then trypsinized to form a single cell suspension and cells were plated onto a MEF feeder layer in the primed hESC medium + ROCK inhibitor described above. 24 hr later, the medium was switched to 5i/L/A naive hESC medium. The 5i/L/A naive hESC medium (500 ml) used for induction and maintenance of naive hESCs was made up of 240 ml DMEM/F12, 240 ml Neurobasal (Invitrogen, 21103), 5 ml N2 supplement (Invitrogen, 17502048) and 10 ml B27 supplement (Invitrogen, 17504044), supplemented with 10 μg recombinant human LIF (purified in-lab from *E. coli*), 1 mM glutamine, 1% nonessential amino acids, 0.1 mM β-mercaptoethanol, penicillin-streptomycin, 50 μg/ml BSA (Sigma, A4737-25G), and the following small molecules and cytokines: 1 μM PD0325901 (Stemgent, 04-0006), 1 μM IM-12 (Enzo, BML-WN102-0005), 0.5 μM SB590885 (R&D systems, 2650/10), 1 μM WH-4-023 (A Chemtek) 10 μM Y-27632 (Stemgent, 04-0012), and 10 ng/ml Activin A (Peprotech, 120-14). Following an initial wave of widespread cell death, dome-shaped naive hESC colonies appeared within 10 days and could be expanded and maintained in 5i/L/A naive hESC medium.

Naive hESCs were maintained on mitomycin C-inactivated MEF feeder cells and passaged every 5-7 days. The naive hESCs were passaged by dissociating cells with accutase (GIBCO, A1110501), and then centrifuging cells at 1000 rpm for 5 minutes at room temperature in neutralization medium (DMEM supplemented with 10% FBS, 1 mM glutamine, 1% nonessential amino acids, penicillin-streptomycin, and 0.1 mM β-mercaptoethanol). To harvest cells for downstream experiments, primed and naive hESCs were trypsinized and subsequently pre-plated on gelatin-coated dishes to deplete MEF feeder cells. All cell culture experiments were performed under physiological oxygen conditions (5% O_2, 3% CO_2).

Genome Editing

The CRISPR/Cas9 system was used to create hESCs with CTCF site deletions. For each experiment two target-specific oligonucleotides (sgRNA) flanking the proposed deletion were cloned into plasmids carrying a codon-optimized version of Cas9 (pX330, Addgene: 42230) that had been further engineered with either a GFP or mCherry fluorescent reporter. WIBR3 primed hESCs were cultured in 10 μM ROCK inhibitor (Stemgent; Y-27632) 24 hr prior to electroporation. Two confluent six well plates of cells were harvested using 0.25% trypsin/EDTA (Invitrogen) and resuspended in phosphate buffered saline (PBS). Cells were electroporated with 20 μg of pX330-sgRNA-GFP and 20 μg pX330-sgRNA-mCherry targeting up and downstream of the intended CTCF site deletion. Cells were subsequently plated in MEF feeder layers in primed hESC medium supplemented with 10 μM ROCK inhibitor. 48 hr post electroporation, cells were harvested using 0.25% trypsin/EDTA and double positive GFP+/mCherry+ cells were isolated by Fluorescent Activated Cell Sorting (FACS). After sorting, GFP+/mCherry+ cells were plated on MEF feeder layers in primed hESC medium supplemented with 10 μM ROCK inhibitor. 8-12 days later individual colonies were picked, expanded and genotyped by PCR.

shRNA Knockdown

VSVG coated lentiviruses were generated in HEK-293 cells. Viral containing supernatant was collected 48 and 72 hr post-transfection. Viral supernatant was filtered through a 0.45 µm filter. 24 hr prior to infection primed human ESCs were treated with 10 µM ROCK inhibitor. On the day of infection naive and primed human ESCs were single cell disassociated with Accutase and 0.25% trypsin respectively. Cells were then resuspended in lentiviral supernatant w/polybrene in ultra-low attachment plates and spun in a centrifuge at 2000 rpm. This spin infection was conducted for 1.5 hr. Cells were then replated on DR4 MEF feeder layer (primed cells were supplemented w/ ROCKi). Medium was changed after 20 hr. 48 hr post-infection medium was supplemented with puromycin (0.5 µg/ml) to select for proviral integration, and doxycycline (2 µg/ml) to induce expression of the shRNA. Seven days later, RNA was extracted and gene expression level was measured by RT-qPCR.

Gene Expression Analysis

RNA was isolated using Trizol reagent (Invitrogen, 15596-026), and reverse transcribed using oligo-dT primers and SuperScript III reverse transcriptase (Invitrogen, 18080044) according to the manufacturer`s instructions. Quantitative real-time PCR was performed on a 7000 ABI Detection System with FAST SYBR Green Master Mix (Applied Biosystems, 4309155). Gene expression was normalized to GAPDH.

3D DNA FISH

3D DNA FISH was performed as previously described (Bolland et al., 2013). Briefly, cells were attached to slides using a Cytospin at 500 rpm, 3 min, then fixed with 4% paraformaldehyde (PFA) for 10 min at room temperature, then quenched in 0.1 M Tris-HCl, pH 7.4 for 10 min at room temperature. Next, cells were permeabilized in 0.1% saponin/0.1% Triton X-100 in PBS for 10 min at room temperature. Slides were then washed twice in PBS for 5 min at room temperature and subsequently incubated for at least 20 min in 20% glycerol/ PBS at room temperature. Slides were freeze/ thawed in liquid nitrogen three times, and washed twice in PBS for 5 min at room temperature. Slides were incubated in 0.1 M HCl for 30 min at room temperature, washed in PBS for 5 min at room temperature and permeabilized in 0.5% saponin/0.5% Triton X-100/PBS for 30 min at room temperature. This was followed by two washes in PBS for 5 min at room temperature before equilibration in 50% formamide/2x SSC for at least 10 min at room temperature. BAC probes (Empire Genomics) were pipetted onto a coverslip. FISH slides were air dried and heated to 78 °C for precisely 2 min on a hot plate. Coverslip w/probe was mounted in slides and sealed with rubber cement. Slides were incubated overnight at 37 °C in a dark humidified chamber. The next day, rubber cement was removed and slides were placed in 2x SSC until coverslips detached. Slides were washed in 50% formamide/2x SSC for 15 min at 45 °C, washed in 0.2x SSC for 15 min at 63 °C, washed in 2x SSC for 5 min at 45 °C and washed in 2x SSC for 5 min at room temperature. Subsequently slides were washed in PBS for 5 min at room temperature and stained with DAPI (5 µg/ml in 2x SSC) for 2 min at room temperature. Finally, slides were destained in PBS for 5 min at room temperature and coverslips were mounted on slides. Imaging was carried out by confocal microscopy in the Whitehead Keck imaging facility. Spatial distance between FISH probes was quantified using ImageJ (FIJI).

BACs

High-Throughput 3D DNA FISH BAC clones were purchased from BACPAC (CHORI) and were used to generate fluorescently labeled probes as described (Shachar et al., Cell 2015). Probes were as follows: RP11-261P1 (chr5_TAD1), RP11-810B19

(chr5_TAD2), RP11-1029M14 (chr5_equidist_con), RP11-258M5 (chr11_TAD1), RP11-52J19 (chr11_TAD2), RP11-2L5 (chr11_equidist_con).

High-Throughput 3D DNA FISH

High-throughput 3D DNA FISH in naive human embryonic stem cells was done as previously described in (Shachar et al., 2015). Briefly, 10^5 cells per well were plated in 96-well plates, fixed in 4% PFA in PBS for 15 min at room temperature, permeabilized and denatured as described (Shachar et al., 2015). A mix containing 300 ng of each fluorescently labeled probe was ethanol precipitated and re-suspended in 25 µl of hybridization buffer. Cells were denatured with probe mix at 90 °C for 8 min and left to hybridize overnight. Images were acquired using a Perkin Elmer Opera automated imaging system at >100 randomly sampled fields in multiple wells using a 40X water objective. Image analysis was carried out as described in (Burman et al., 2015). Briefly, FISH spot coordinates were detected in individual nuclei that contained an equal number of spots in each channel. The minimal distance in pixels between each combination of spot pairs (TAD border 1, TSD border 2, control) was calculated using Acapella 2.0 (PerkinElmer) and R (http://www.R-project.org/). For statistical analysis, the Mann-Whitney test was used to compare the closest distance between TAD borders to an equidistant control locus.

RNA-seq

RNA-seq was performed for naive and primed hESCs. 6 million cells were used for each RNA extraction. Total RNA was purified using the mirVana™ miRNA Isolation Kit (Life Technologies, AM1560) following the manufacturer's instructions. 1 µg of total RNA was used for the RNA-seq library construction. A technical replicate was performed for both naive and primed hESCs. Polyadenylated RNA-seq libraries were prepared using the TruSeq Stranded mRNA Library Prep Kit (Illumina, RS-122-2101). The RNA-seq libraries were sequenced on the Illumina HiSeq 2000.

RNA-seq Expression Analysis

RNA-seq alignment and quantification were performed using the TopHat and Cufflinks software tools. RNA-seq reads were first aligned to the human genome (build hg19, GRCh37) using Tophat v2.0.13 (Trapnell et al., 2009) with the parameters: --solexa-quals --no-novel-juncs and using RefSeq gene annotations. The expression levels of RefSeq transcripts were calculated using Cufflinks v2.2.1 (Trapnell et al., 2010). Differentially expressed transcripts were then identified, again using Cufflinks v2.2.1. When multiple transcripts had the same gene name, only the transcript with the highest expression level was kept for further consideration. A gene was considered differentially expressed if it met the following criteria: 1) absolute log2 fold-change ≥ 1 between the mean expression in the two conditions; 2) false discovery rate q-value ≤ 0.05.

Three lines of evidence suggested that the RNA-seq datasets were high-quality: 1) ~80% of all reads in all libraries mapped to RefSeq transcript models (hg19), as expected for sequencing of RNA; 2) ~90% of all reads in all libraries mapped to known RefSeq genes (~83% mapped to the exons and ~7% mapped to the introns), as expected for sequencing of poly-A RNA-enriched samples; 3) the replicates of either naive or primed RNA-seq datasets had a Pearson correlation coefficient of expression levels of 0.98 or greater across all RefSeq transcripts.

Cross-Species Gene Expression Analysis

Cross-species gene expression analysis was performed as previously described (Theunissen et al., 2014). For a given gene, the mean expression value for that gene across all human samples was first calculated. Then for each human sample, the expression of that gene in that sample was divided by the mean expression value. The normalization was repeated for all mouse samples. After normalization, all pairwise comparisons of datasets, both intra- and inter-species, were performed using Pearson correlation coefficients (PCCs). The average linkage hierarchical clustering of the Pearson correlation was shown in the heatmap.

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 µg of antibody and 50 µl 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 Library Generation and Sequencing

ChIA-PET was performed using a modified version of a previously described protocol (Dowen 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 µg anti-SMC1 antibody (Bethyl Labs, A300-055A) and 250 µl 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 ligation was performed in a large volume to encourage ligation of ends that are in close spatial proximity to each other, ideally from fragments that are co-localized via their interaction with cohesin-bound regions and immunoprecipitation of cohesin. The ChIP-DNA with beads were washed once with TE buffer, then incubated in 1x T4 DNA polymerase buffer (NEBuffer 2.1, New England Biolabs, B7202S), with 7.2 µl T4 DNA polymerase (New England Biolabs, M0203S) and 7 µl of 10 mM dNTPs (Life Technologies, 18427013) in 700 µl total volume at 37 °C for 40 min. The beads were then washed three times with ChIA-PET wash buffer (10 mM Tris-HCl pH7.5, 1 mM EDTA, 500 mM NaCl). The beads were incubated with 1x NEB buffer 2 (New England Biolabs, B7002S) containing 7 µl Klenow fragment (3'-5' exo$^-$) (New England Biolabs, M0212S) and 7 µl 10 mM dATP (New England Biolabs, N0440S) in 700 µl total volume at 37 °C for 50 min. The beads were then washed three times with ChIA-PET wash buffer. The beads were then incubated with 1x T4 DNA ligase buffer with 1 mM ATP (New England Biolabs, B0202S) containing 42 µl T4 DNA ligase (Life Technologies, 46300018) and 4 µl bridge linker (200 ng/µl including Forward: /5Phos/CGCGATATC/iBiodT/TATCTGACT; Reverse: /5Phos/GTCAGATAAGATATCGCGT) in 14 ml total volume at 16 °C for 22 hr. The beads were then washed three times with ChIA-PET wash buffer. The beads were then incubated with 1x lambda exonuclease buffer (New England Biolabs, M0262S) containing 6 µl lambda exonuclease (New England Biolabs, M0262S), and 6 µl exonuclease I (New England Biolabs, M0293S) in 700 µl total volume at 37 °C for 1 hr.

DNA elution and crosslink reversal were simultaneously performed by incubating the beads at 55 °C overnight. 10 µl of proteinase K (Life Technologies, AM2546) was included during the overnight incubation. The DNA was then purified by phenol-chloroform extraction and ethanol precipitation.

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 proximity ligation products were tagmented with Tn5 Transposase (5 µl Tn5 transposase (Illumina, FC-121-1030) for 50 ng DNA) at 55 °C for 5 min, then at 10 °C for 10 min. DNA was purified using a Zymo column (VWR, 100554-654) following the manufacturer's instructions. Biotin-labeled DNA was then further affinity purified with M280 streptavidin beads (50 µl for each library, Life Technologies, 11205D), followed by washing five times with 2x SSC/0.5% SDS and then two times with 1x B&W buffer (5 mM Tris-HCl pH7.5, 0.5 mM EDTA, 1 M NaCl). The buffer was discarded and the beads were gently resuspended in 30 µl EB buffer (QIAGEN). 10 µl of the bead slurry was used for PCR amplification. PCR amplification was performed using the Nextera DNA Sample Preparation Kit (Illumina, FC-121-1031) for 10-12 cycles. The DNA was selected for the size range of 300-500 bp and was purified by gel extraction. The ChIA-PET library was subjected to 100 x 100 paired-end sequencing using Illumina HiSeq 2000.

ChIP-seq Data Analysis

All ChIP-Seq datasets were aligned to the human genome (build hg19, GRCh37) using Bowtie (version 0.12.2) (Langmead et al., 2009) with the parameters -k 1 -m 1 -n 2. We used the MACS peak finding algorithm, version 1.4.2 (Zhang et al., 2008) to identify regions of ChIP-seq enrichment over input DNA control with the parameters "--no-model --keep-dup=1". A p-value threshold for enrichment of 1e-09 was used for H3K27ac, H3K27me3 (Theunissen et al., 2014), MED1 and OCT4 datasets, while a p-value of 1e-07 was used for the CTCF dataset. UCSC Genome Browser (Kent et al., 2002) tracks were generated using the MACS wiggle file output option with parameters "-w –S –space=50". All gene-centric analyses in human ESCs were performed using human (build hg19, GRCh37) RefSeq annotations downloaded from the UCSC genome browser (genome.ucsc.edu).

ChIA-PET Data Processing

All ChIA-PET datasets were processed with a method adapted from a previously published computational pipeline (Dowen et al., 2014; Li et al., 2010). The output of paired-end sequencing is a set of reads, where each read is identified by a read id and consists of two mates that represent sequence from the ends of a DNA fragment. The raw sequences of each mate of each read were analyzed for the presence of the PET linker barcodes and trimmed using Cutadapt with the parameters "-m 17 -a forward=ACGCGATATCTTATCTGACT -a reverse=AGTCAGATAAGATATCGCGT --overlap 10" (Martin, 2011) specifically, we searched for a stretch of at least 10 bp that matched the linker sequence. Once this sequence was identified, the linker sequence and all sequence immediately 3' to this sequence was removed. After removal of linker and 3' sequence, only sequences of at least 17 bp in length were retained. For downstream analysis, all mates from all reads where at least one mate contained the linker sequence were used. Sequences of mates were separately mapped to the hg19 human genome using Bowtie with the parameters "-k 1 -m 1 -v 2 -p 4 --best --strata" (Langmead et al., 2009). These criteria retained only the uniquely mapped mates, with at most two base pair mismatches, for further analysis. Aligned mates were paired using

their respective read ids and now considered PETs (paired-end tags). PETs were filtered for redundancy: PETs with identical genomic coordinates and strand information at both ends were collapsed into a single PET. The PETs were further categorized into intrachromosomal PETs, where the two ends of a PET were on the same chromosome, and interchromosomal PETs, where the two ends were on different chromosomes. The sequences from the ends of all PETs were then analyzed for localized enrichment across the genome using MACS 1.4.2 (Zhang et al., 2008) with the parameters "-p 1e-09 -no-lambda –no-model --keep- dup=2". Regions identified with MACS were considered PET peaks.

To identify long-range chromatin interactions, we first removed intra-chromosomal PETs of length < 4 kb because these PETs are suspected to originate from self-ligation of DNA ends from a single chromatin fragment in the ChIA-PET procedure (Dowen et al., 2014). We next identified PETs that overlapped with PET peaks at both ends by at least 1 bp. Operationally, these PETs were defined as putative interactions. Applying a statistical model based upon the hypergeometric distribution identified high-confidence interactions, representing high-confidence physical linking between the PET peaks. To do this, for each PET peak, we calculated a) the total number of PETs that overlap with the peak and b) the number of PETs that overlap with the peak and also connect to another peak. A hypergeometric distribution was used to determine the probability of seeing at least the observed number of PETs linking the two PET peaks. The correction p-values were calculated using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control for multiple hypothesis testing. Operationally, the pairs of interacting sites with three independent PETs and an FDR ≤ 0.01 were defined as high-confidence interactions in the SMC1 ChIA-PET merged dataset and with two independent PETs in the individual SMC1 ChIA-PET replicates. Previously published RAD21 (cohesin) ChIA-PET datasets in K562 (Heidari et al., 2014) were downloaded from ENCODE (https://www.encodeproject.org/experiments/ENCSR000FDB/) and were re-processed exactly as described (Dowen et al., 2014; Li et al., 2010).

Additional ChIA-PET interaction analysis

Several additional analyses were conducted to characterize the sets of interactions identified in this work and improve our confidence in these calls. Interactions were compared to those called by a second analysis pipeline to demonstrate robust identification of interactions. Interactions were analyzed for the presence of expected DNA sequence motifs at their ends. Interactions were analyzed for the presence of expected regulatory elements at their ends. Finally, interactions were compared to interactions detected by Hi-C, a second experimental method.

Interactions were compared to interactions called using a second analysis pipeline (called Mango (Phanstiel et al., 2015)). Different analysis pipelines incorporate different biases and assumptions in identifying interactions, and depend on selection of parameters and thresholds, and are thus likely to yield different results. Regardless, one expectation is that a large number of interactions should be identified robustly if bona fide interactions are being found. Thus, comparison of the outputs of the two pipelines provides some measure of the robustness of the identification of interactions. The Mango software (version 1.0.2) was downloaded (https://github.com/dphansti/mango), installed, and initially run with default settings. For input, the same data (sequencing reads and genomic regions called enriched for signal) used for the Dowen et al. pipeline were applied to the Mango pipeline. 10,702 and 13,233 ChIA-PET interactions were called in naive and primed hESCs, respectively (Table S6) when using the Mango

pipeline. These interactions were then compared to their counterparts derived with the Dowen et al. pipeline. Interactions were considered shared if each end of an interaction identified with the Mango pipeline overlapped by at least one base pair with the respective ends of an interaction identified with the Dowen et al. pipeline. 86% of the 10,702 interactions in naive hESC were identified in both pipelines. 91% of the 13,233 interactions in primed hESC were identified in both pipelines (Figure S1F). The observation of robust identification of at least a subset of the data using different analysis pipelines generally increases our confidence that a large set of bona fide interactions is being identified.

Interactions were analyzed for the presence of expected DNA sequence motifs at their ends. The subset of cohesin-associated, CTCF-CTCF loops are expected to have convergently oriented CTCF DNA-binding sequence motifs underlying each of the two CTCF binding regions that comprise the ends of the loops. Briefly, the location and orientation of the CTCF motifs at CTCF ChIP-seq peaks were identified using the FIMO software package 4 (Grant et al., 2011; Matys et al., 2006) and searching with the canonical CTCF motif from the Jaspar motif database (ID. MA0139.1). The orientation of CTCF motifs at pairs of CTCF ChIP-seq peaks was next determined. For simplicity, we focused on those CTCF-CTCF loops where CTCF peaks could be unambiguously assigned to a CTCF motif. All pairs of CTCF motifs at the two ends of CTCF-CTCF ChIA-PET interactions were classified into one of the four possible classes of motif orientations: a convergent orientation (forward-reverse), a divergent orientation (reverse-forward), the same direction on the forward strand (forward-forward) or the same direction on the reverse strand (reverse-reverse). Additional details can be found in the section titled “CTCF Motif Orientation Analysis at CTCF-CTCF Loops”**.** Approximately 80% of the interactions identified here have CTCF sequence motifs in the expected orientation (convergent) at the ends of the interactions (Figure S2A). The observation that CTCF-CTCF loops identified here display the expected orientation of CTCF sequence motifs at their ends increases our confidence that bona fide interactions are being identified.

Interactions were analyzed for the presence of expected regulatory elements at their ends. Cohesin-associated loops are expected to have ends associated with CTCF sites, enhancers and promoters. For this analysis, CTCF sites and enhancers were identified using ChIP-seq data for CTCF and the histone modification H3K27ac, respectively. Briefly, ChIP-Seq datasets were aligned to the human genome to identify regions of ChIP-seq enrichment over input DNA control. A p-value threshold for enrichment of 1e-07 was used for CTCF, while a p-value of 1e-09 was used for the H3K27ac dataset. H3K27ac-enriched regions were further filtered for those that were at least 2 kb away from a RefSeq transcription start site to identify the set of enhancer regions. Promoters were defined as the region +/- 2 kb around RefSeq transcription start sites. Additional details can be found in the section titled “ChIP-seq Data Analysis”. 75-85% of the interactions identified here in naive or primed hESCs had ends that overlapped with CTCF sites, enhancers or promoters (greater than 1 bp overlap). The large fraction of interactions identified with ends overlapping biologically relevant genomic features increases our confidence that bona fide interactions are being identified.

Interactions were compared to interactions detected by Hi-C, a second experimental method to detect interactions. ChIA-PET detects interactions occurring between sites associated with a specific protein, while Hi-C detects interactions more generally. Thus, the set of ChIA-PET interactions is expected to overlap with the set of Hi-C interactions

A previously published Hi-C dataset from H1 hES cells was downloaded and used to derive a set of interactions (Dixon et al., 2015) Briefly, the raw fragment contact and bias matrices at 40 kb resolution were first obtained using the python hiclib library. The Fit-Hi-C tool (Ay et al., 2014) was then used to call high-confidence DNA interactions (FDR .05). Additional details can be found in the section titled “Calling Hi-C Interactions”. Given the 40 kb bin size, the minimum distance of interactions from the Hi-C data was effectively 80 kb. Thus for comparisons, we compared the Hi-C interactions to ChIA-PET interactions that were 80 kb or greater in length. 74% of the interactions identified using the cohesin ChIA-PET data were also identified in the Hi-C data. The large fraction of the ChIA-PET interactions identified in Hi-C data increases our confidence that bona fide interactions are being identified with the cohesin ChIA-PET data.

Assignment of Interactions to Regulatory Elements

We assigned the PET peaks of interactions to different regulatory elements, including promoters (+/- 2 kb of the Refseq TSS), active enhancers (H3K27ac enriched regions falling outside of promoter regions that are defined as +/- 2 kb of the Refseq TSS), and CTCF ChIP-seq binding sites. Operationally, an interaction was defined as associated with the regulatory element if one of the two PET peaks of the interaction overlapped with the regulatory element by at least 1 base pair. CTCF-CTCF loops were defined as high confidence ChIA-PET loops with CTCF ChIP-seq peaks at both ends of the interactions.

Identification of CTCF-CTCF Loops that Define Putative Insulated Neighborhoods

All CTCF-CTCF loops may potentially form insulated neighborhoods. For this paper putative insulated neighborhoods were defined by incorporating some evidence for loop insulation by measure of directionality index as described below. Briefly, CTCF-CTCF loops were evaluated for putative insulating function by examining the directionality of reads proximal to loop boundaries. One expectation for a loop with insulating function is that, at a loop boundary, interactions originating just upstream of the boundary connect to a distal point located further upstream while interactions originating just downstream of the boundary connect to a distal point located further downstream. Boundaries satisfying these criteria thus have implied functionality in terms of constraining interactions. Adjacent pairs of boundaries satisfying these criteria would thus be candidates for demonstrating insulating function. ChIA-PET interaction directionality preferences were calculated using a method adapted from Hi-C computational analysis (Mizuguchi et al., 2014). Briefly, each chromosome (autosomes and X chromosome) was divided into non-overlapping 40 kb bins. Each intra-chromosomal ChIA-PET interaction (either below or above 4 kb) was then mapped to the matrix comprised of all pairwise combinations of bins. Each end of a ChIA-PET interaction contributed signal to its respective bin, thus generating a matrix of interaction frequencies between bins. ChIA-PET directional preference scores were next calculated from these interaction frequency matrices as the log2 ratio of upstream to downstream contact frequencies for each region i at distances below 400 kb:

$$Di = log_2\left(\sum_{j=-10}^{j=0} C_{i,i+j} \Big/ \sum_{j=0}^{j=10} C_{i,i+j}\right),$$

in which C is the ChIA-PET interaction frequency matrix.

Putative insulated neighborhoods were operationally defined as intra-chromosomal CTCF-CTCF interactions where each end of the interaction displayed a change in directional preference. This type of change in interaction preference between upstream and downstream genomic regions was previously used to computationally define topologically associating domains (Dixon et al., 2012; Nora et al., 2012). To improve the robustness of calculating interaction preferences at the CTCF-occupied peaks at CTCF-CTCF interactions, we calculated the average interaction preference at two neighboring bins in the proximity of the CTCF-occupied peaks. Specifically, we first identified the genomic bins where the two ends of CTCF-CTCF interactions were located. For each of the 5' CTCF-occupied PET peaks of these CTCF-CTCF interactions, we selected two bins: one located where the 5' PET peak was located and the other in the immediately neighboring bin in the 3' direction. For each of the 3' CTCF-occupied PET peaks at CTCF-CTCF interactions, we also selected two bins: one located where the 3' PET peak was located and the other in the immediately neighboring bin in the 5' direction. We then filtered for CTCF-CTCF interactions whose mean of interaction directional preference between the two bins at their 5' PET peak was positive (indicating downstream preferences) and mean of interaction directional preference between two bins at their 3' PET peak was negative (indicating upstream preferences). Since the ChIA-PET interaction frequency matrix was calculated using 40 kb bins, this method allowed us to detect putative insulated neighborhoods greater than 80 kb.

TAD schematic construction

For schematics of TAD structures, we show TAD-spanning loops with at least one PET read. We show those putative insulated neighborhoods that pass the directionality index criteria described above. All non-overlapping putative insulated neighborhoods are shown. When overlapping putative insulated neighborhoods are possible, the loop with the most PET reads supporting the interaction was selected for display. When comparing structures encompassing genes with cell type preferred enhancers in naive versus primed hESCs, structures were first identified in the cell type with the cell type preferred enhancer. The second cell type was then examined for the presence of corresponding structures with evidence for CTCF binding. As a default, when enhancer signals were similar, naive hESC structures were first identified and the primed hESCs were then examined for the corresponding structure. For simplicity, a subset of genes is displayed with their associated enhancers. Enhancers were defined as stitched H3K27ac MACS peaks (using the ROSE algorithm). The loop with the highest PET reads supporting each enhancer-promoter or enhancer-enhancer interaction was shown (using PET ≥ 2).

ChIA-PET Interaction Heatmap at Insulated Neighborhoods

Cohesin ChIA-PET interactions were displayed to examine the similarity of neighborhoods between naive hESCs, primed hESCs, and K562. Insulated neighborhoods for naive hESCs were centered and size-normalized. ChIA-PET PET signal (number of uniquely mapped PETs per million uniquely mapped PETs) was then displayed. For comparison, the ChIA-PET signal from primed hESCs and K562 for the regions with the same coordinates was displayed.

CTCF Motif Orientation Analysis at CTCF-CTCF Loops

The location and orientation of the CTCF motifs at CTCF ChIP-seq peaks were identified using the FIMO software package with a default p value threshold of 10^-4 (Grant et al., 2011; Matys et al., 2006). In the analysis, the canonical CTCF motif from the Jaspar motif database (ID. MA0139.1) was used. The orientation of CTCF motifs at pairs of

CTCF ChIP-seq peaks was next determined. For simplicity, we focused on those CTCF-CTCF loops where CTCF peaks could be unambiguously assigned to a CTCF motif: each end overlapped a single CTCF ChIP-seq peak by at least 1 base pair and only a single CTCF motif was at the peak. All pairs of CTCF motifs at the two ends of CTCF-CTCF ChIA-PET interactions were classified into one of the four possible classes of motif orientations: a convergent orientation (forward-reverse), a divergent orientation (reverse-forward), the same direction on the forward strand (forward-forward) or the same direction on the reverse strand (reverse-reverse).

Hi-C Interaction Heatmap

To generate a matrix of Hi-C interaction frequencies mapped to a more recent build of the human genome, previously published Hi-C datasets in H1 hESCs (Dixon et al., 2015) were first downloaded from GEO (www.ncbi.nlm. nih.gov/geo/; accession GSM1267196 and GSM1267197). The raw reads from these datasets were mapped to the human genome build hg19 and filtered as previously described (Imakaev et al., 2012). Corrected contact probability matrices at 40 kb resolution were obtained using the python hiclib library (https://bitbucket.org/mirnylab/hiclib).

Super-Enhancers in hESCs

Super-enhancers were identified in naive or primed hESCs using ROSE (https://bitbucket.org/young_computation/rose). This code is an implementation of the method used in (Hnisz et al., 2013; Loven et al., 2013). Briefly, regions enriched for H3K27ac signal were identified using MACS. These regions were stitched together if they were within 12.5 kb of each other and enriched regions entirely contained within +/- 2 kb from a TSS were excluded from stitching. Stitched regions were ranked by H3K27ac signal therein. ROSE identified a point at which the two classes of enhancers were separable. Those stitched enhancers falling above this threshold were considered super-enhancers.

SMC1 binding Enrichment Heatmap

The heatmaps show the average ChIP-seq or ChIA-PET read density (r.p.m./bp) of different factors at SMC1 occupied regions. Individual ChIP datasets were processed separately and peaks of enriched signal were identified as described above. For SMC1, the genome was binned into 50 bp bins and read density of signal is shown for the 10 kb region representing +/- 5 kb from the center of each SMC1-enriched region. Similar read density of signal is shown for each other factor at the corresponding regions shown for the SMC1 dataset.

Heatmap Representation of High-confidence ChIA-PET Interactions

ChIA-PET interaction signals relative to the boundaries of CTCF-CTCF loops were mapped in a distance-normalized fashion. For each CTCF-CTCF loop, we demarcated three regions: loop, upstream, and downstream. For the loop region, the region was divided into 50 equally sized bins. For the upstream region, we selected a region extending upstream of the loop itself. The upstream region's length was set at 20% of the length of the corresponding loop. The upstream region was then divided into 10 equally sized bins. Similarly, for the downstream region, we selected a region extending downstream from the loop for a distance corresponding to 20% of the length of the loop itself, and divided the region into 10 equally sized bins.

To see whether interactions originating within the loop were generally confined within the loop, we first filtered high-confidence interactions in two ways. We required high-

confidence interactions to have at least one end in the interrogated region. This removed interactions where both endpoints of the interaction were anchored outside of the region of interest. We removed interactions that had one end at a domain border PET peak and the other end outside of the domain. This removed interactions that originated at a border and had no end within the domain as we did not consider them to be originating within the domain.

The density of the genomic space covered by ChIA-PET interactions in each bin was next calculated as the number of interactions per bin. Interactions within CTCF-CTCF loops were considered. The density of ChIA-PET interactions was row-normalized to the row maximum for each domain and the normalized frequency was displayed. Interactions connecting enhancers and promoters were considered and displayed. The density of ChIA-PET interactions was row-normalized to the row maximum for each domain and the normalized frequency was displayed.

Differential H3K27ac Signal at Enhancer Clusters Between Naive and Primed hESCs

Enhancer clusters were generated to compare enhancer regions between naive and primed hESCs. We first identified the sets of enhancer clusters in naive and primed hESCs using ROSE (https://bitbucket.org/young_computation/rose). Briefly, regions enriched for H3K27ac signal were identified using MACS. These regions were stitched together if they were within 12.5 kb of each other and enriched regions entirely contained within +/- 2 kb from a TSS were excluded from stitching. Enhancer cluster regions from naive and primed hESCs that overlapped by 1 bp were then merged together to form a representative region that spans the combined genomic region. A total of 24,755 enhancer cluster regions were identified. For each region, the read density in reads per million per base pair (r.p.m./bp) from the replicate data (2 replicate H3K27ac ChIP-seq datasets in naive hESCs and 2 replicate H3K27ac ChIP-seq datasets in primed hESCs) was calculated, and from this the relative read count of each region was obtained by multiplying read density by the length of the region. The edgeR package was used to model technical variation due to noise among duplicate data sets and the biological variation due to differences in signal between naive and primed hESCs (Robinson et al., 2010). Sequencing depth and upper- quartile techniques were used to normalize all 4 datasets together before common and tagwise dispersions were estimated. The statistical significance of differences between naive and primed hESCs was next calculated using an exact test and resulting p values were subjected to Benjamini–Hochberg multiple testing correction (FDR). The final regions with differential H3K27ac signal were required to have the absolute log2 fold change of normalized H3K27ac signal greater or equal to 2 and FDR less or equal to 0.05.

Fold Change of H3K27ac Signal at Super-Enhancer Clusters

In order to quantify the signal changes of super-enhancers between naive and primed hESCs, H3K27ac ChIP-Seq signal was calculated at the set of all enhancer cluster regions considered as super-enhancers in at least one condition. Sequencing depth and upper-quartile techniques were used to normalize the H3K27ac ChIP-Seq signal at these super-enhancer clusters using normalization factors derived from the total 24,755 enhancer cluster regions described above. The log2 fold change of normalized H3K27ac signal was displayed.

Saturation Analysis of ChIA-PET Library

To determine the degree of saturation within our ChIA-PET library, we modeled the number of sampled putative interactions, which were defined as PETs that overlapped with two PET peaks at both ends by at least 1 bp, as a function of sequencing depth by a two parameter logistic growth model. Intrachromosomal PETs were subsampled at varying depths, and the number of unique putative interactions that they occupied were counted. Model fitting using non-linear least-squares regression suggested that we sampled approximately 45~50 % of the available intrachromosomal PET space.

Calling Hi-C Interactions

The Fit-Hi-C tool (Ay et al., 2014) was used to call high-confidence DNA interactions from Hi-C datasets in H1 hESCs (Dixon et al., 2015). The raw fragment contact and bias matrices at 40 kb resolution were first obtained using the python hiclib library. The Fit-Hi-C was then used to call high-confidence DNA interactions using the raw fragment contact and bias matrices at 40 kb resolution with the parameters: -L 50,000 –U 5,000,000 –b 200 –p 1 --quiet. The CTCF-CTCF Hi-C interactions were identified by filtering for those Hi-C interactions that have CTCF ChIP-seq peaks within the 40 kb bins at the both ends of the interactions. A Hi-C CTCF-CTCF interaction was classified as "overlapped with a ChIA-PET CTCF-CTCF interaction" if both ends of the Hi-C CTCF-CTCF interaction overlapped with the ends of a ChIA-PET CTCF-CTCF interaction by at least 1 bp. The percentages of Hi-C CTCF-CTCF interactions that overlapped with ChIA-PET CTCF-CTCF interactions (or vice versa) were displayed as line plots.

Comparisons to In situ Hi-C CTCF-CTCF Interactions

Previously published in situ Hi-C CTCF-CTCF interactions with CTCF DNA motifs in 7 different cell types (Rao et al., 2014) were first downloaded from GEO (www.ncbi.nlm.nih.gov/geo/; accession GSE63525). They were next compared to the ChIA-PET CTCF-CTCF interactions in naive and primed hESCs. We tested how often these in situ Hi-C CTCF-CTCF interactions overlapped with the CTCF-CTCF interactions in naive and primed hESCs. Since the in situ Hi-C CTCF-CTCF interactions were identified by requiring CTCF ChIP-seq peaks within a +/- 15 kb window at the both ends of the interactions in the publication (Rao et al., 2014), an in situ Hi-C CTCF-CTCF interaction was classified as "overlapped with a ChIA-PET CTCF-CTCF interaction" if both ends of the in situ Hi-C CTCF-CTCF interaction overlapped with the ends of a ChIA-PET CTCF-CTCF interaction within a +/- 15kb window. The percentages of in situ Hi-C CTCF-CTCF interactions that overlapped with ChIA-PET CTCF-CTCF interactions were displayed as bar plots.

Calling High-Confidence Cell-Type-Specific CTCF-CTCF Interactions in Naive And Primed hESCs

To identify the naive-specific or primed-specific CTCF-CTCF interactions, we took advantage of the strong signal at the PET peaks to increase the confidence to interpret the ChIA-PET interaction data. This was because the PET counts or the ChIP-seq read counts at PET peaks were frequently an order magnitude higher than the PET count for the number of PETs spanning high-confidence interactions allowing for better dynamic range. Briefly, we applied a negative binomial statistical model from the edgeR package to identify differentially occupied CTCF peaks between naive and primed hESCs using ChIP-seq data (FDR 0.01 and absolute log2 fold change >= 2) and overlaid these differential ChIP-seq CTCF regions to the CTCF-CTCF ChIA-PET interactions from naive and primed hESCs.

To identify differentially occupied CTCF peaks between naive and primed hESCs, CTCF ChIP-seq peaks from naive and primed hESCs that overlapped by 1 bp were then merged together to form a representative region that spans the combined genomic region. For each region, the read density in reads per million per base pair (r.p.m./bp) from the replicate data (2 replicate CTCF ChIP-seq datasets in naive hESCs and 2 replicate CTCF ChIP-seq datasets in primed hESCs) was calculated, and from this the relative read count of each region was obtained by multiplying read density by the length of the region. The edgeR package was used to model technical variation due to noise among duplicate data sets and the biological variation due to differences in signal between naive and primed hESCs (Robinson et al., 2010). Sequencing depth and upper-quartile techniques were used to normalize all 4 datasets together before common and tagwise dispersions were estimated. The statistical significance of differences between naive and primed hESCs was next calculated using an exact test and resulting p values were subjected to Benjamini–Hochberg multiple testing correction (FDR). The final regions with differential CTCF signal were required to have the absolute log2 fold change of normalized CTCF signal greater or equal to 2 and FDR less or equal to 0.05. This analysis resulted in 313 naive-specific CTCF peaks and 75 primed-specific CTCF peaks.

We next identified the CTCF-CTCF interactions that were associated with these preferentially occupied CTCF peaks in naive and primed hESCs by requiring at least one end of the CTCF-CTCF interactions overlapped with the preferentially occupied CTCF peaks. To obtain the high-confidence cell-type-specific CTCF-CTCF interactions, we also required the naive-specific CTCF-CTCF interactions that overlapped naive-specific CTCF peaks to have zero PETs in primed hESCs, and primed-specific CTCF-CTCF interactions that overlapped primed-specific CTCF peaks to have zero PETs in naive hESCs. This resulted in only 125 naive-specific CTCF-CTCF interactions and 28 primed-specific CTCF-CTCF interactions.

Topologically Associating Domain (TAD) Calling

TADs were determined from interaction matrices using the method and code previously described in (Dixon et al., 2012). For cohesin ChIA-PET-based TADs, ChIA-PET interactions were used to generate interaction matrices by binning the genome into 40 kb bins and counting the number of PETs connecting any two bins. For H1 hESC Hi-C based TADs, H1 hESC Hi-C data previously generated in (Dixon et al., 2015), was realigned, binned into 40 kb bins, and normalized to generate a Hi-C interaction matrix. Parameters from Dixon et al. were retained (an interaction window of 2 Mb and 40 kb for binning interactions). For human samples, the human reference genome (build hg19, GRCh37) was used and for mouse samples, the mm9 mouse reference genome was used.

Hi-C vs ChIA-PET Interaction Comparison

Hi-C data was examined to see if the Hi-C data supported predicted ChIA-PET interactions. To do this, H1 hESC Hi-C data was first processed to create an interaction matrix as described above. The subset of the Hi-C interaction matrix that could be directly compared to the available ChIA-PET data was then selected. The interaction scores from the Hi-C matrix were then plotted as a box plot. For comparison, a random distribution of Hi-C interactions was generated and also plotted.

TAD Spanning Loops: Percentage and Visualization

TADs derived from Hi-C data from H1 hESCs were examined for the presence of CTCF-CTCF loops that spanned the entire TAD. TADs and Hi-C interactions were derived as described above. For each TAD, we queried if there was at least one CTCF-CTCF loop that connected the upstream and downstream boundaries of the TAD. For this analysis, each boundary was extended by 40 kb both upstream and downstream. A loop was considered spanning if one end was found in the upstream boundary and the other end was found in the downstream boundary. We examined TADs for the number of spanning loops that connected the two boundaries; the percentages of TADs with 1, 2 or 3 spanning loops were reported. For comparison, the analysis was repeated using a set of randomized, shuffled TADs. For the shuffled set, we used the set of H1 Hi-C based TADs but shuffled the chromosome and start site coordinates. Visualization of spanning loops was done using the CRAN-Circlize package (http://cran.r-project.org/web/packages/circlize/index.html).

TAD Boundary Overlap

To compare the consistency of TADs called using either Hi-C or ChIA-PET data, we asked if boundaries of Hi-C based TAD were frequently co-localized with boundaries of ChIA-PET based TAD calls. To do this, we examined the overlap of the boundaries of TADs called using ChIA-PET data and Hi-C data. For each boundary, we measured the distance of each Hi-C called TAD boundary to the nearest ChIA-PET called TAD boundary. The distribution of distances was then plotted in a histogram.

Conservation and Disease Analysis

We examined whether the ends of CTCF-CTCF loops overlapped with genomic regions of high sequence conservation or genomic regions associated with disease-causing mutations. We began by identifying the CTCF motifs (as described above) that were within the anchor sites of high confidence CTCF-CTCF ChIA-PET interactions. We considered two sets of regions, the first being CTCF-CTCF anchor sites and the second being CTCF motif sites that are bound by CTCF and within loop anchor sites. For conservation analysis, the 10 kb of sequence around the midpoint of each CTCF-CTCF anchor site (+/- 5 kb) was used. For each region, for each base pair, the PhastCons score was determined using a 10 way primate multiple alignment (Pollard et al., 2010). This created a vector of PhastCons scores for each region. The vectors for all regions were then averaged and plotted. For association with cancer mutations, the regions described above were overlapped with the coordinates of simple somatic mutations present in cancer from the International Cancer Genome Consortium (ICGC) database (Zhang et al., 2011). For each base pair, the base pair was scored for presence of a mutation. This created a vector of mutation occurrences for each region. The vectors for all regions were then summed and plotted. For association with disease mutations, the regions described above were overlapped with the coordinates of GWAS SNPs (Welter et al., 2014). GWAS SNPs that fell within these regions were reported. All of the analyses were repeated for CTCF motifs. Here, the sequences analyzed included the motif itself plus 200 bp of sequence upstream and downstream.

GWAS Catalog Parsing and Distance Distribution

The NHGRI Genome-Wide Association Study (GWAS) database containing SNPs significantly associated with human traits was downloaded 6/19/2015 and parsed as described in (Hnisz et al., 2013). Briefly, trait-associated SNPs with dbSNP identifiers were reproducibly associated with a trait in two independent studies. SNPs were assigned a genomic position using dbSNP build 142. SNPs falling inside RefSeq coding exons were discarded. The distance distribution of trait-associated, noncoding SNPs to

the nearest border of a region in the union of 86 enhancer sets defined in (Hnisz et al., 2013) were shown. The distance distribution of trait-associated noncoding SNPs to the nearest border of a CTCF anchor in the union of the naive and primed anchor sites were shown. SNPs within these regions were assigned to the 0 bin.

Fractional Methylation Analysis at CTCF Sites

We examined the methylation dynamics of CTCF motifs within CTCF-CTCF high-confidence loop anchor sites throughout early embryonic development by using reduced representation bisulfite (RRBS) sequencing from human preimplantation embryos (Smith et al., 2014). The CTCF motifs were flanked by 4991 bp to create a 10 kb region around each CTCF motif. Each 10 kb region was then overlapped with the fractional methylation RRBS data, this generated a set of genomic locations that fell within these 10 kb regions and had fractional methylation values assigned to them. Each of the fractional methylation values was then added to the correct location within the vector of 10000 values relative to the CTCF motif. This created one 10000 value average fractional methylation vector for five of the samples in Figure S6D (sperm, 8-cell embryo, inner cell mass, hESC primed, and fetal lung). This vector was smoothed over using a 10 bp window and plotted as a line plot.

A similar analysis was performed using whole genome bisulfite sequencing from (http://egg2.wustl.edu/roadmap/web_portal/processed_data.html#MethylData). We first averaged the fractional methylation across 37 cell/tissue types. This generated one average fractional methylation value for each base pair in the genome. These values were overlapped as described above with the same 10 kb regions and plotted in the same manner. Also, the fractional methylation plot for the adult lung sample in Figure S6D was generated in the same way as described for the other five samples but used data from WUSTL.

Transcription Factor Motif and Mutation Analysis within CTCF-CTCF Loop Anchors

We determined the average number of mutations found in occurrences of transcription factor motifs that occur in anchor regions. We first downloaded a set of motif instances from (Kheradpour et al., 2013) consisting of sequence motifs, their assignment to transcription factors and their chromosomal location. We next filtered for those motif instances within anchor regions. For each member of the resulting set of motif instances, we counted how many cancer mutations overlapped the motif instance. The counts for all instances assigned to a given factor were summed and divided by the number of instances assigned to that factor. The simple somatic mutations present in cancer were described in the International Cancer Genome Consortium database (Zhang et al., 2011).

ICGC Simple Somatic Mutations within the Loop Anchors

A VCF file of simple somatic mutations was downloaded from the International Cancer Genome Consortium database (Zhang et al., 2011). The file was filtered for mutation calls generated from projects that are not under embargo (project key PACA-AU, PACA-CA, PRAD-CA, BLCA-CN, GACA-CN, LICA-FR, EOPC-DE, MALY-DE, PBCA-DE, LINC-JP, LIRI-JP, LUSC-KR, CLLE-ES, BRCA-UK, CMDI-UK, PRAD-UK, and OV-AU). We first identified all CTCF-CTCF loops for which there are simple mutations overlapping with CTCF motifs within the loop anchors by at least 1 bp for CTCF-CTCF loops in naive and primed hESCs. We next examined the genes that were contained within these CTCF-CTCF loops by requiring that their TSSs be within the loop. The gene

symbols were cross-referenced to Refseq genes, cancer census genes, proto-oncogenes, and tumor suppressor genes. Cancer census genes (Version 73) were downloaded from the COSMIC database (www.cancer.sanger.ac.uk/cosmic). Proto-oncogenes are operationally defined as genes from the cancer census gene list whose mutations result in a dominant phenotype (Bishop et al., 1991). Tumor suppressor genes were downloaded from the TSGene Tumor suppressor gene database (http://bioinfo.mc.vanderbilt.edu/TSGene/Human_716_TSGs.txt) (Zhao et al., 2012).

REFERENCES

Ay, F., Bailey, T.L., and Noble, W.S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res *24*, 999-1011.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological *57*, 289-300.

Bolland, D.J., King, M.R., Reik, W., Corcoran, A.E., and Krueger, C. (2013). Robust 3D DNA FISH using directly labeled probes. J Vis Exp.

Burman, B., Zhang, Z.Z., Pegoraro, G., Lieb, J.D., and Misteli, T. (2015). Histone modifications predispose genome regions to breakage and translocation. Genes Dev *29*, 1393-1402.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W.*, et al.* (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331-336.

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

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K.*, et al.* (2014). Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. Cell *159*, 374-387.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.

Heidari, N., Phanstiel, D.H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M.Q., and Snyder, M.P. (2014). Genome-wide map of regulatory interactions in the human genome. Genome Res *24*, 1905-1917.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-Enhancers in the Control of Cell Identity and Disease. Cell *155*, 934-947.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods *9*, 999-1003.

Ji, X., Dadon, D.B., Abraham, B.J., Lee, T.I., Jaenisch, R., Bradner, J.E., and Young, R.A. (2015). 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 *112*, 3841-3846.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Research *12*, 996-1006.

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res *23*, 800-811.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology *10*.
Li, G., Fullwood, M.J., Xu, H., Mulawadi, F.H., Velkov, S., Vega, V., Ariyaratne, P.N., Bin Mohamed, Y., Ooi, H.-S., Tennakoon, C.*, et al.* (2010). ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biology *11*.
Loven, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320-334.
Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal *17*, 1-10.
Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K.*, et al.* (2006). TRANSFAC (R) and its module TRANSCompel (R): transcriptional gene regulation in eukaryotes. Nucleic Acids Research *34*, D108-110.
Mizuguchi, T., Fudenberg, G., Mehta, S., Belton, J.-M., Taneja, N., Folco, H.D., FitzGerald, P., Dekker, J., Mirny, L., Barrowman, J.*, et al.* (2014). Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. Nature *516*, 432-435.
Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J.*, et al.* (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381-385.
Phanstiel, D.H., Boyle, A.P., Heidari, N., and Snyder, M.P. (2015). Mango: a bias-correcting ChIA-PET analysis pipeline. Bioinformatics *31*, 3092-3098.
Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research *20*, 110-121.
Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S.*, et al.* (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665-1680.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.
Shachar, S., Voss, T.C., Pegoraro, G., Sciascia, N., and Misteli, T. (2015). Identification of Gene Positioning Factors Using High-Throughput Imaging Mapping. Cell *162*, 911-923.
Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L.*, et al.* (2014). Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. Cell Stem Cell *15*, 471-487.
Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology *28*, 511-515.
Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L.*, et al.* (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research *42*, D1001-1006.

Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B.*, et al.* (2011). International Cancer Genome Consortium Data Portal-a one-stop shop for cancer genomics data. Database-the Journal of Biological Databases and Curation.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology *9*.