Please cite this article in press as: Rao et al., A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping, Cell (2014), http://dx.doi.org/10.1016/j.cell.2014.11.021

Cell

A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping

Suhas S.P. Rao,[1,2,3,4,10] Miriam H. Huntley,[1,2,3,4,5,10] Neva C. Durand,[1,2,3,4] Elena K. Stamenova,[1,2,3,4] Ivan D. Bochkov,[1,2,3] James T. Robinson,[1,4] Adrian L. Sanborn,[1,2,3,6] Ido Machol,[1,2,3] Arina D. Omer,[1,2,3] Eric S. Lander,[4,7,8,*] and Erez Lieberman Aiden[1,2,3,4,9,*]

[1]The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA
[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
[3]Department of Computer Science, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA
[5]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[6]Department of Computer Science, Stanford University, Stanford, CA 94305, USA
[7]Department of Biology, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA
[8]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[9]Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA
[10]Co-first author
*Correspondence: lander@broadinstitute.org (E.S.L.), erez@erez.com (E.L.A.)
http://dx.doi.org/10.1016/j.cell.2014.11.021

SUMMARY

We use in situ Hi-C to probe the 3D architecture of genomes, constructing haploid and diploid maps of nine cell types. The densest, in human lymphoblastoid cells, contains 4.9 billion contacts, achieving 1 kb resolution. We find that genomes are partitioned into contact domains (median length, 185 kb), which are associated with distinct patterns of histone marks and segregate into six subcompartments. We identify ~10,000 loops. These loops frequently link promoters and enhancers, correlate with gene activation, and show conservation across cell types and species. Loop anchors typically occur at domain boundaries and bind CTCF. CTCF sites at loop anchors occur predominantly (>90%) in a convergent orientation, with the asymmetric motifs "facing" one another. The inactive X chromosome splits into two massive domains and contains large loops anchored at CTCF-binding repeats.

INTRODUCTION

The spatial organization of the human genome is known to play an important role in the transcriptional control of genes (Cremer and Cremer, 2001; Sexton et al., 2007; Bickmore, 2013). Yet important questions remain, like how distal regulatory elements, such as enhancers, affect promoters, and how insulators can abrogate these effects (Banerji et al., 1981; Blackwood and Kadonaga, 1998; Gaszner and Felsenfeld, 2006). Both phenomena are thought to involve the formation of protein-mediated "loops" that bring pairs of genomic sites that lie far apart along the linear genome into proximity (Schleif, 1992).

Various methods have emerged to assess the 3D architecture of the nucleus. In one seminal study, the binding of a protein to sites at opposite ends of a restriction fragment created a loop, which was detectable because it promoted the formation of DNA circles in the presence of ligase. Removal of the protein or either of its binding sites disrupted the loop, eliminating this "cyclization enhancement" (Mukherjee et al., 1988). Subsequent adaptations of cyclization enhancement made it possible to analyze chromatin folding in vivo, including nuclear ligation assay (Cullen et al., 1993) and chromosome conformation capture (Dekker et al., 2002), which analyze contacts made by a single locus, extensions such as 5C for examining several loci simultaneously (Dostie et al., 2006), and methods such as ChIA-PET for examining all loci bound by a specific protein (Fullwood et al., 2009).

To interrogate all loci at once, we developed Hi-C, which combines DNA proximity ligation with high-throughput sequencing in a genome-wide fashion (Lieberman-Aiden et al., 2009). We used Hi-C to demonstrate that the genome is partitioned into numerous domains that fall into two distinct compartments. Subsequent analyses have suggested the presence of smaller domains and have led to the important proposal that compartments are partitioned into condensed structures ~1 Mb in size, dubbed "topologically associated domains" (TADs) (Dixon et al., 2012; Nora et al., 2012). In principle, Hi-C could also be used to detect loops across the entire genome. To achieve this, however, extremely large data sets and rigorous computational methods are needed. Recent efforts have suggested that this is an increasingly plausible goal (Sexton et al., 2012; Jin et al., 2013).

Here, we report the results of an effort to comprehensively map chromatin contacts genome-wide, using in situ Hi-C, in which DNA-DNA proximity ligation is performed in intact nuclei. The protocol facilitates the generation of much denser Hi-C maps. The maps reported here comprise over 5 Tb of sequence

data recording over 15 billion distinct contacts, an order of magnitude larger than all published Hi-C data sets combined. Using these maps, we are able to clearly discern domain structure, compartmentalization, and thousands of chromatin loops. In addition to haploid maps, we were also able to create diploid maps analyzing each chromosomal homolog separately. The maps provide a picture of genomic architecture with resolution down to 1 kb.

RESULTS

In Situ Hi-C Methodology and Maps

Our in situ Hi-C protocol combines our original Hi-C protocol (here called dilution Hi-C) with nuclear ligation assay (Cullen et al., 1993), in which DNA is digested using a restriction enzyme, DNA-DNA proximity ligation is performed in intact nuclei, and the resulting ligation junctions are quantified. Our in situ Hi-C protocol involves crosslinking cells with formaldehyde, permeabilizing them with nuclei intact, digesting DNA with a suitable 4-cutter restriction enzyme (such as MboI), filling the 5′-overhangs while incorporating a biotinylated nucleotide, ligating the resulting blunt-end fragments, shearing the DNA, capturing the biotinylated ligation junctions with streptavidin beads, and analyzing the resulting fragments with paired-end sequencing (Figure 1A). This protocol resembles a recently published single-cell Hi-C protocol (Nagano et al., 2013), which also performed DNA-DNA proximity ligation inside nuclei to study nuclear architecture in individual cells. Our updated protocol has three major advantages over dilution Hi-C. First, in situ ligation reduces the frequency of spurious contacts due to random ligation in dilute solution—as evidenced by a lower frequency of junctions between mitochondrial and nuclear DNA in the captured fragments and by the higher frequency of random ligations observed when the supernatant is sequenced (Extended Experimental Procedures available online). This is consistent with a recent study showing that ligation junctions formed in solution are far less meaningful (Gavrilov et al., 2013). Second, the protocol is faster, requiring 3 days instead of 7 (Extended Experimental Procedures). Third, it enables higher resolution and more efficient cutting of chromatinized DNA, for instance, through the use of a 4-cutter rather than a 6-cutter (Data S1, I).

A Hi-C map is a list of DNA-DNA contacts produced by a Hi-C experiment. By partitioning the linear genome into "loci" of fixed size (e.g., bins of 1 Mb or 1 kb), the Hi-C map can be represented as a "contact matrix" M, where the entry $M_{i,j}$ is the number of contacts observed between locus L_i and locus L_j. (A "contact" is a read pair that remains after we exclude reads that are duplicates, that correspond to unligated fragments, or that do not align uniquely to the genome.) The contact matrix can be visualized as a heatmap, whose entries we call "pixels." An "interval" refers to a set of consecutive loci; the contacts between two intervals thus form a "rectangle" or "square" in the contact matrix. We define the "matrix resolution" of a Hi-C map as the locus size used to construct a particular contact matrix and the "map resolution" as the smallest locus size such that 80% of loci have at least 1,000 contacts. The map resolution is meant to reflect the finest scale at which one can reliably discern local features.

Contact Maps Spanning Nine Cell Lines Containing over 15 Billion Contacts

We constructed in situ Hi-C maps of nine cell lines in human and mouse (Table S1). Whereas our original Hi-C experiments had a map resolution of 1 Mb, these maps have a resolution of 1 kb or 5 kb. Our largest map, in human GM12878 B-lymphoblastoid cells, contains 4.9 billion pairwise contacts and has a map resolution of 950 bp ("kilobase resolution") (Table S2). We also generated eight in situ Hi-C maps at 5 kb resolution, using cell lines representing all human germ layers (IMR90, HMEC, NHEK, K562, HUVEC, HeLa, and KBM7) as well as mouse B-lymphoblasts (CH12-LX) (Table S1). Each map contains between 395 M and 1.1 B contacts.

When we used our original dilution Hi-C protocol to generate maps of GM12878, IMR90, HMEC, NHEK, HUVEC, and CH12-LX, we found that, as expected, in situ Hi-C maps were superior at high resolutions, but closely resembled dilution Hi-C at lower resolutions. For instance, our dilution map of GM12878 (3.2 billion contacts) correlated highly with our in situ map at 500, 50, and 25 kb resolutions ($R > 0.96$, 0.90, and 0.87, respectively) (Data S1, I; Figure S1).

We also performed 112 supplementary Hi-C experiments using three different protocols (in situ Hi-C, dilution Hi-C, and Tethered Conformation Capture) while varying a wide array of conditions such as extent of crosslinking, restriction enzyme, ligation volume/time, and biotinylated nucleotide. These include several in situ Hi-C experiments in which the formaldehyde crosslinking step was omitted, which demonstrate that the structural features we observe cannot be due to the crosslinking procedure. In total, 201 independent Hi-C experiments were successfully performed, many of which are presented in Data S1 and S2.

To account for nonuniformities in coverage due to the number of restriction sites at a locus or the accessibility of those sites to cutting (Lieberman-Aiden et al., 2009; Yaffe and Tanay, 2011) we use a matrix-balancing algorithm due to Knight and Ruiz (2012) (Extended Experimental Procedures).

Adequate tools for visualization of these large data sets are essential. We have therefore created the "Juicebox" visualization system that enables users to explore contact matrices, zoom in and out, compare Hi-C matrices to 1D tracks, superimpose all features reported in this paper onto the data, and contrast different Hi-C maps. All contact data and feature sets reported here can be explored interactively via Juicebox at http://www.aidenlab.org/juicebox/.

The Genome Is Partitioned into Small Domains Whose Median Length Is 185 kb

We began by probing the 3D partitioning of the genome. In our earlier experiments at 1 Mb map resolution (Lieberman-Aiden et al., 2009), we saw large squares of enhanced contact frequency tiling the diagonal of the contact matrices. These squares partitioned the genome into 5–20 Mb intervals, which we call "megadomains."

We also found that individual 1 Mb loci could be assigned to one of two long-range contact patterns, which we called compartments A and B, with loci in the same compartment showing more frequent interaction. Megadomains—and the associated squares along the diagonal—arise when all of the 1 Mb loci in

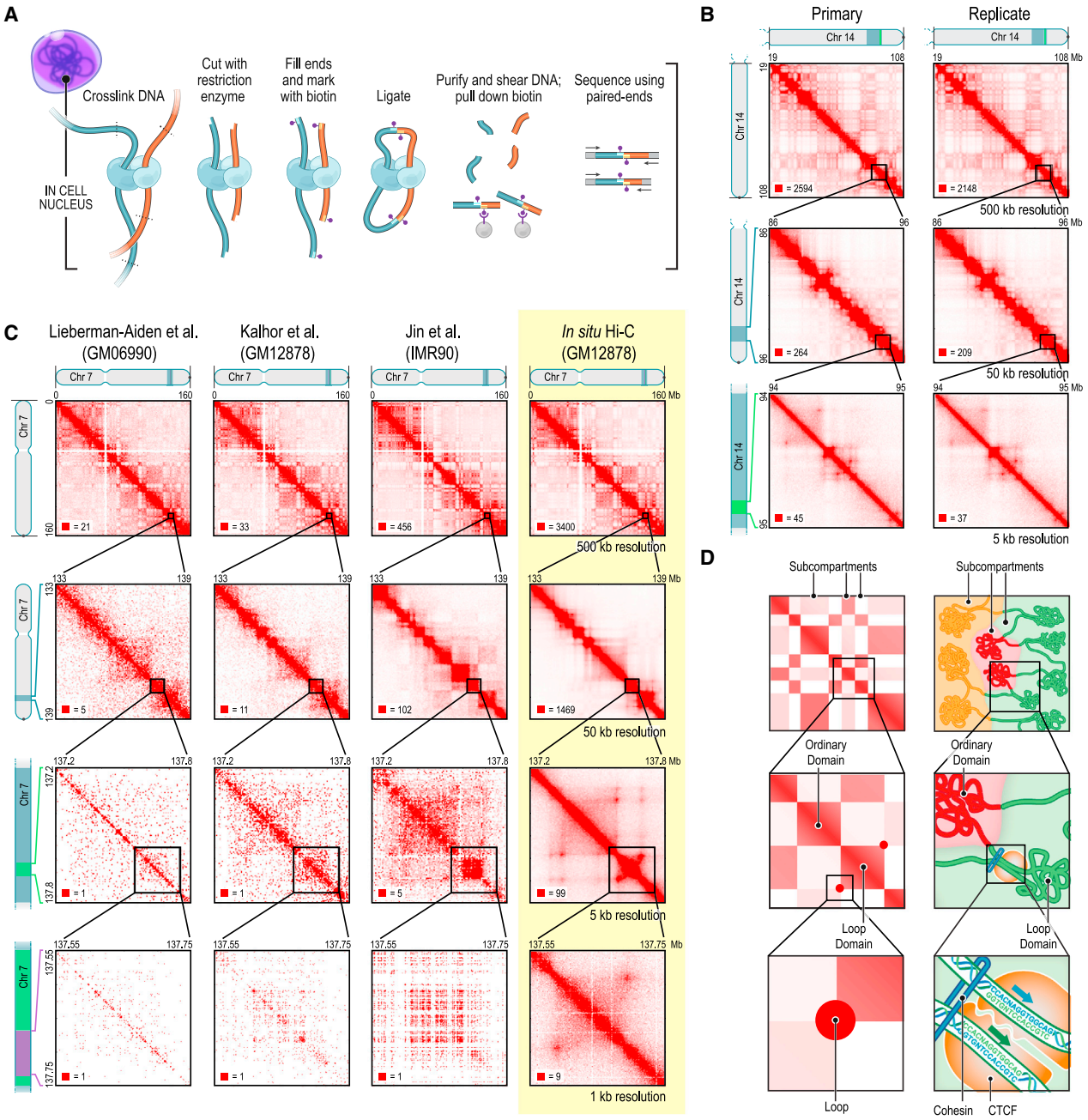

Figure 1. We Used In Situ Hi-C to Map over 15 Billion Chromatin Contacts across Nine Cell Types in Human and Mouse, Achieving 1 kb Resolution in Human Lymphoblastoid Cells

(A) During in situ Hi-C, DNA-DNA proximity ligation is performed in intact nuclei.

(B) Contact matrices from chromosome 14: the whole chromosome, at 500 kb resolution (top); 86–96 Mb/50 kb resolution (middle); 94–95 Mb/5 kb resolution (bottom). Left: GM12878, primary experiment; Right: biological replicate. The 1D regions corresponding to a contact matrix are indicated in the diagrams above and at left. The intensity of each pixel represents the normalized number of contacts between a pair of loci. Maximum intensity is indicated in the lower left of each panel.

(C) We compare our map of chromosome 7 in GM12878 (last column) to earlier Hi-C maps: Lieberman-Aiden et al. (2009), Kalhor et al. (2012), and Jin et al. (2013).

(D) Overview of features revealed by our Hi-C maps. Top: the long-range contact pattern of a locus (left) indicates its nuclear neighborhood (right). We detect at least six subcompartments, each bearing a distinctive pattern of epigenetic features. Middle: squares of enhanced contact frequency along the diagonal (left) indicate the presence of small domains of condensed chromatin, whose median length is 185 kb (right). Bottom: peaks in the contact map (left) indicate the presence of loops (right). These loops tend to lie at domain boundaries and bind CTCF in a convergent orientation.

See also Figure S1, Data S1, I–II, and Tables S1 and S2.

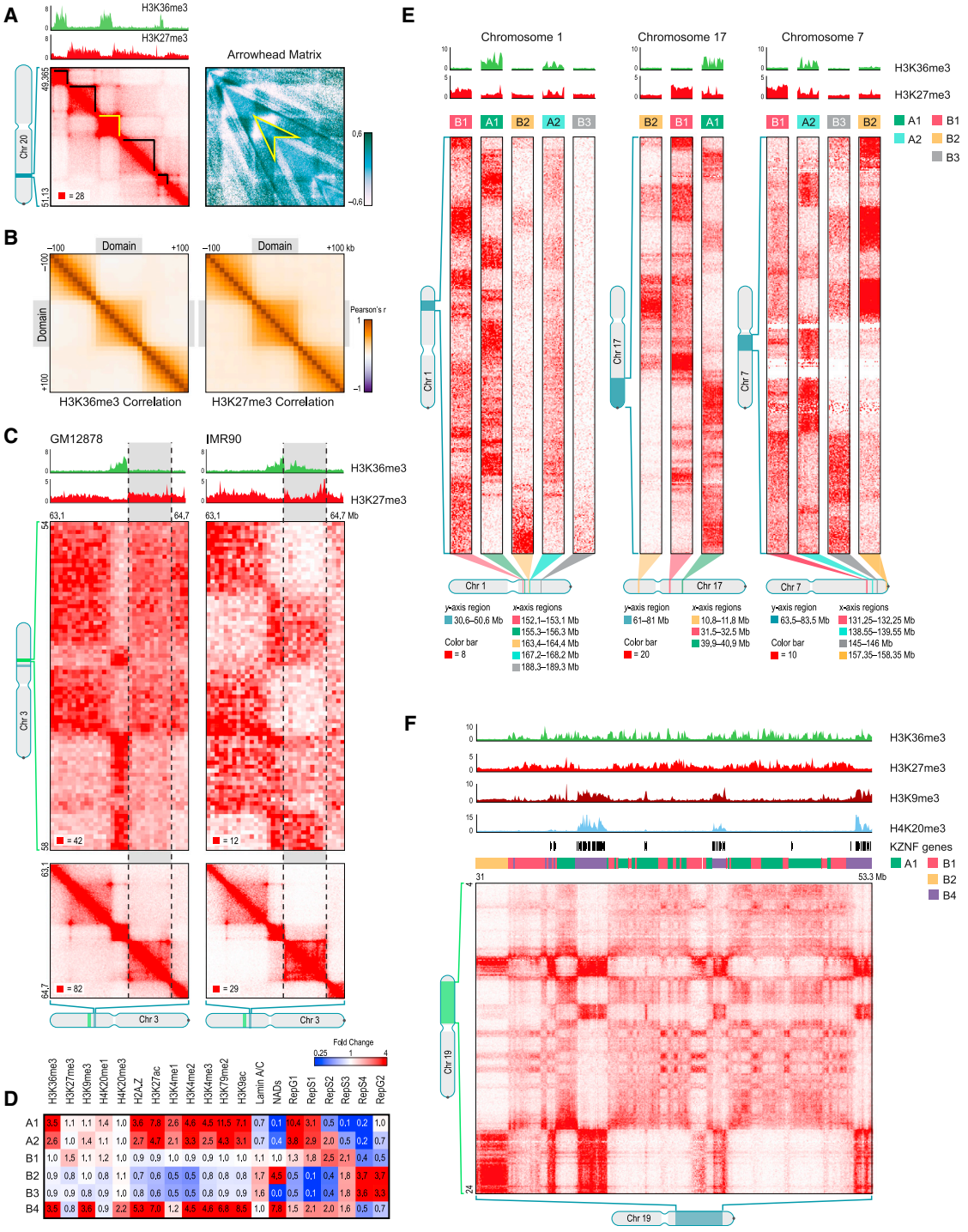

Figure 2. The Genome Is Partitioned into Contact Domains that Segregate into Nuclear Subcompartments Corresponding to Different Patterns of Histone Modifications

(A) We annotate thousands of domains across the genome (left, black highlight). To do so, we define an arrowhead matrix A (right) such that $A_{i,i+d} = (M^*_{i,i-d} - M^*_{i,i+d})/(M^*_{i,i-d} + M^*_{i,i+d})$, where M^* is the normalized contact matrix. This transformation replaces domains with an arrowhead-shaped motif pointing toward the domain's upper-left corner (example in yellow); we identify these arrowheads using dynamic programming. See Experimental Procedures.

(B) Pearson correlation matrices of the histone mark signal between pairs of loci inside and within 100 kb of a domain. Left: H3K36me3; Right: H3K27me3.

(C) Conserved contact domains on chromosome 3 in GM12878 (left) and IMR90 (right). In GM12878, the highlighted domain (gray) is enriched for H3K27me3 and depleted for H3K36me3. In IMR90, the situation is reversed. Marks at flanking domains are the same in both: the domain to the left is enriched for H3K36me3 and the domain to the right is enriched for H3K27me3. The flanking domains have long-range contact patterns that differ from one another and are preserved in both

(legend continued on next page)

an interval exhibit the same genome-wide contact pattern. Compartment A is highly enriched for open chromatin; compartment B is enriched for closed chromatin (Lieberman-Aiden et al., 2009; Kalhor et al., 2012; Sexton et al., 2012).

In our new, higher resolution maps (200- to 1,000-fold more contacts), we observe many small squares of enhanced contact frequency that tile the diagonal of each contact matrix (Figure 2A). We used the Arrowhead algorithm (see Experimental Procedures) to annotate these contact domains genome-wide. The observed domains ranged in size from 40 kb to 3 Mb (median size 185 kb). As with megadomains, there is an abrupt drop in contact frequency (33%) for pairs of loci on opposite sides of the domain boundary (Figure S2G). Contact domains are often preserved across cell types (Figures S3A and S3B).

The presence of smaller domains in Hi-C maps is consistent with several other recent studies (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). We explore the relationship between the domains we annotate and those annotated in prior studies in the Discussion.

Contact Domains Exhibit Consistent Histone Marks Whose Changes Are Associated with Changes in Long-Range Contact Pattern

Loci within a contact domain show correlated histone modifications for eight different factors (H3K36me3, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K79me2, and H4K20me1) based on data from the ENCODE project in GM12878 cells (ENCODE Project Consortium, 2012). By contrast, loci at comparable distance but residing in different domains showed much less correlation in chromatin state (Figures 2B, S2I, and S2K; Extended Experimental Procedures). Strikingly, changes in a domain's chromatin state are often accompanied by changes in the long-range contact pattern of domain loci (i.e., the pattern of contacts between loci in the domain and other loci genome-wide), indicating that changes in chromatin pattern are accompanied by shifts in a domain's nuclear neighborhood (Figures 2C and S3C–S3E; Extended Experimental Procedures). This observation is consistent with microscopy studies associating changes in gene expression with changes in nuclear localization (Finlan et al., 2008).

There Are at Least Six Nuclear Subcompartments with Distinct Patterns of Histone Modifications

Next, we partitioned loci into categories based on long-range contact patterns alone, using four independent approaches: manual annotation and three unsupervised clustering algorithms (HMM, K-means, Hierarchical). All gave similar results (Figure S4B; Extended Experimental Procedures). We then investigated the biological meaning of these categories.

When we analyzed the data at low matrix resolution (1 Mb), we reproduced our earlier finding of two compartments (A and B). At high resolution (25 kb), we found evidence for at least five "subcompartments" defined by their long-range interaction patterns, both within and between chromosomes. These findings expand on earlier reports suggesting three compartments in human cells (Yaffe and Tanay, 2011). We found that the median length of an interval lying completely within a subcompartment is 300 kb. Although the subcompartments are defined solely based on their Hi-C interaction patterns, they exhibit distinct genomic and epigenomic content.

Two of the five interaction patterns are correlated with loci in compartment A (Figure S4E). We label the loci exhibiting these patterns as belonging to subcompartments A1 and A2. Both A1 and A2 are gene dense, have highly expressed genes, harbor activating chromatin marks such as H3K36me3, H3K79me2, H3K27ac, and H3K4me1 and are depleted at the nuclear lamina and at nucleolus-associated domains (NADs) (Figures 2D, 2E, and S4I; Table S3). While both A1 and A2 exhibit early replication times, A1 finishes replicating at the beginning of S phase, whereas A2 continues replicating into the middle of S phase. A2 is more strongly associated with the presence of H3K9me3 than A1, has lower GC content, and contains longer genes (2.4-fold).

The other three interaction patterns (labeled B1, B2, and B3) are correlated with loci in compartment B (Figure S4E) and show very different properties. Subcompartment B1 correlates positively with H3K27me3 and negatively with H3K36me3, suggestive of facultative heterochromatin (Figures 2D and 2E). Replication of this subcompartment peaks during the middle of S phase. Subcompartments B2 and B3 tend to lack all of the above-noted marks and do not replicate until the end of S phase (see Figure 2D). Subcompartment B2 includes 62% of pericentromeric heterochromatin (3.8-fold enrichment) and is enriched at the nuclear lamina (1.8-fold) and at NADs (4.6-fold). Subcompartment B3 is enriched at the nuclear lamina (1.6-fold), but strongly depleted at NADs (76-fold).

Upon closer visual examination, we noticed the presence of a sixth pattern on chromosome 19 (Figure 2F). Our genome-wide clustering algorithm missed this pattern because it spans only 11 Mb, or 0.3% of the genome. When we repeated the algorithm on chromosome 19 alone, the additional pattern was detected. Because this sixth pattern correlates with the Compartment B pattern, we labeled it B4. Subcompartment B4 comprises a handful of regions, each of which contains many KRAB-ZNF superfamily genes. (B4 contains 130 of the 278 KRAB-ZNF genes in the genome, a 65-fold enrichment). As noted in previous studies (Vogel et al., 2006; Hahn et al., 2011), these regions exhibit a highly distinctive chromatin pattern, with strong enrichment for

cell types. In IMR90, the highlighted domain is marked by H3K36me3 and its long-range contact pattern matches the similarly-marked domain on the left. In GM12878, it is decorated with H3K27me3, and the long-range pattern switches, matching the similarly-marked domain to the right. Diagonal submatrices, 10 kb resolution; long-range interaction matrices, 50 kb resolution.

(D) Each of the six long-range contact patterns we observe exhibits a distinct epigenetic profile (data sources are listed in Table S3). Each subcompartment also has a visually distinctive contact pattern.

(E) Each example shows part of the long-range contact patterns for several nearby genomic intervals lying in different subcompartments.

(F) A large contiguous region on chromosome 19 contains intervals in subcompartments A1, B1, B2, and B4.

See also Figures S2, S3, and S4 and Data S1, III–IV.

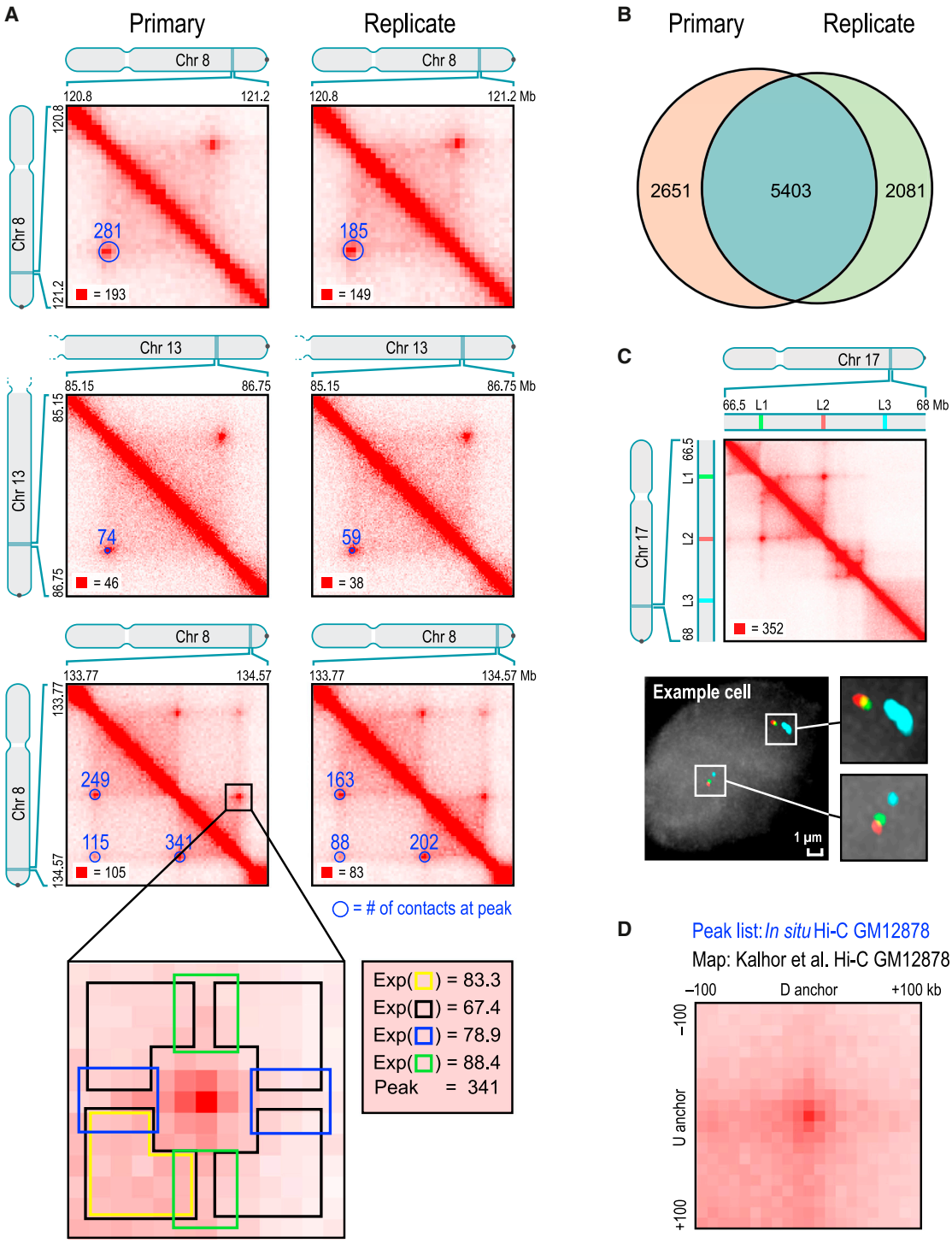

Figure 3. We Identify Thousands of Chromatin Loops Genome-wide Using a Local Background Model

(A) We identify peaks by detecting pixels that are enriched with respect to four local neighborhoods (blowout): horizontal (blue), vertical (green), lower-left (yellow), and donut (black). These "peak" pixels indicate the presence of a loop and are marked with blue circles (radius = 20 kb) in the lower-left of each heatmap. The number of raw contacts at each peak is indicated. Left: primary GM12878 map; Right: replicate; annotations are completely independent. All contact matrices in this and subsequent figures are 10 kb resolution unless noted.

(B) Overlap in peak annotations between replicates.

(C) Top: location of 3D-FISH probes used to verify a peak in the chromosome 17 contact map. Bottom: example cell.

(legend continued on next page)

both activating chromatin marks, such as H3K36me3, and heterochromatin-associated marks, such as H3K9me3 and H4K20me3.

Approximately 10,000 Peaks Mark the Position of Chromatin Loops

We next sought to identify the positions of chromatin loops by using an algorithm to search for pairs of loci that show significantly closer proximity with one another than with the loci lying between them (Figure 3A). Such pairs correspond to pixels with higher contact frequency than typical pixels in their neighborhood. We refer to these pixels as "peaks" in the Hi-C contact matrix and to the corresponding pair of loci as "peak loci." Peaks reflect the presence of chromatin loops, with the peak loci being the anchor points of the chromatin loop. (Because contact frequencies vary across the genome, we define peak pixels relative to the local background. We note that some papers [Sanyal et al., 2012; Jin et al., 2013] have sought to define peaks relative to a genome-wide average. This choice is problematic because, for example, many pixels within a domain may be reported as peaks despite showing no locally distinctive proximity; see Discussion.)

Our algorithm detected 9,448 peaks in the in situ Hi-C map for GM12878 at 5 kb matrix resolution. These peaks are associated with a total of 12,903 distinct peak loci (some peak loci are associated with more than one peak). The vast majority of peaks (98%) reflected loops between loci that are <2 Mb apart.

These findings were reproducible across all of our high-resolution Hi-C maps. Examining the primary and replicate maps separately, we found 8,054 peaks in the former and 7,484 peaks in the latter, with 5,403 in both lists (see Figures 3A and 3B; Data S1, V; Table S4). The differences were almost always the result of our conservative peak-calling criteria (Extended Experimental Procedures). We also called peaks using our GM12878 dilution Hi-C experiment. Because the map is sparser and thus noisier, we called only 3,073 peaks. Nonetheless, 65% of these peaks were also present in the list of peaks from our in situ Hi-C data set, again reflecting high interreplicate reproducibility.

To independently confirm that peak loci are closer than neighboring locus pairs, we performed 3D-FISH (Beliveau et al., 2012) on four loops (Table S5). In each case, we compared two peak loci, *L1* and *L2*, with a control locus, *L3*, that lies an equal genomic distance away from *L2* but on the opposite side (Figures 3C and S5B). In all cases, the 3D-distance between *L1* and *L2* was consistently shorter than the 3D-distance between *L2* and *L3* (Extended Experimental Procedures).

We also confirmed that our list of peaks was consistent with previously published Hi-C maps. Although earlier maps contained too few contacts to reliably call individual peaks, we developed a method called Aggregate Peak Analysis (APA) that compares the aggregate enrichment of our peak set in these low-resolution maps to the enrichment seen when our peak set is translated in any direction (Experimental Procedures). APA showed strong consistency between our loop calls and all six previously published Hi-C experiments in lymphoblastoid cell lines (Lieberman-Aiden et al., 2009; Kalhor et al., 2012) (Figure 3D; Data S2, I.E; Table S6).

Finally, we demonstrated that the peaks observed were robust to particular protocol conditions by performing APA on our GM12878 dilution Hi-C map and on our 112 supplemental Hi-C experiments exploring a wide range of protocol variants. Enrichment was seen in every experiment. Notably, these include five experiments (HIC043-HIC047; Table S1) in which the Hi-C protocol was performed without crosslinking, demonstrating that the peaks observed in our experiments cannot be byproducts of the formaldehyde-crosslinking procedure.

Conservation of Peaks among Human Cell Lines and across Evolution

We also identified peaks in the other seven human cell lines (Table S1). Because these maps contain fewer contacts, sensitivity is reduced, and fewer peaks are observed (ranging from 2,634 to 8,040). APA confirmed that these peak calls were consistent with the dilution Hi-C maps reported here (in IMR90, HMEC, HUVEC, and NHEK), as well as with all previously published Hi-C maps in these cell types (Lieberman-Aiden et al., 2009; Dixon et al., 2012; Jin et al., 2013) (Data S2, I.F).

We found that peaks were often conserved across cell types (Figure 4A): between 55% and 75% of the peaks found in any given cell type were also found in GM12878 (Figure S5D).

Next, we compared peaks across species. In CH12-LX mouse B-lymphoblasts, we identified 2,927 high-confidence contact domains and 3,331 peaks. When we examined orthologous regions in GM12878, we found that 50% of peaks and 45% of domains called in mouse were also called in humans. This suggests substantial conservation of 3D genome structure across the mammals (Figures 4B–4E).

Loops Anchored at a Promoter Are Associated with Enhancers and Increased Gene Activation

Various lines of evidence indicate that many of the observed loops are associated with gene regulation.

First, our peaks frequently have a known promoter at one peak locus (as annotated by ENCODE's ChromHMM) (Hoffman et al., 2013) and a known enhancer at the other (Figure 5A). For instance, 2,854 of the 9,448 peaks in our GM12878 map bring together known promoters and known enhancers (30% versus 7% expected by chance). The peaks include classic promoter-enhancer loops, such as at *MYC* (chr8:128.35–128.75 Mb, in HMEC) and alpha-globin (chr16:0.15–0.22 Mb, in K562). Second, genes whose promoters are associated with a loop are much more highly expressed than genes whose promoters are not associated with a loop (6-fold).

Third, the presence of cell type-specific peaks is associated with changes in expression. When we examined RNA sequencing (RNA-seq) data produced by ENCODE, we found

(D) APA plot shows the aggregate signal from the 9,448 GM12878 loops we report by summing submatrices surrounding each peak in a low-resolution GM12878 Hi-C map due to Kalhor et al. (2012). Although individual peaks cannot be seen in the Kalhor et al. (2012) data (that contains 42 M contacts), the peak at the center of the APA plot indicates that the aggregate signal from our peak set as a whole can be clearly discerned using their data set.
See also Figure S5, Data S1, V. and Data S2,I, and Tables S4, S5, and S6.

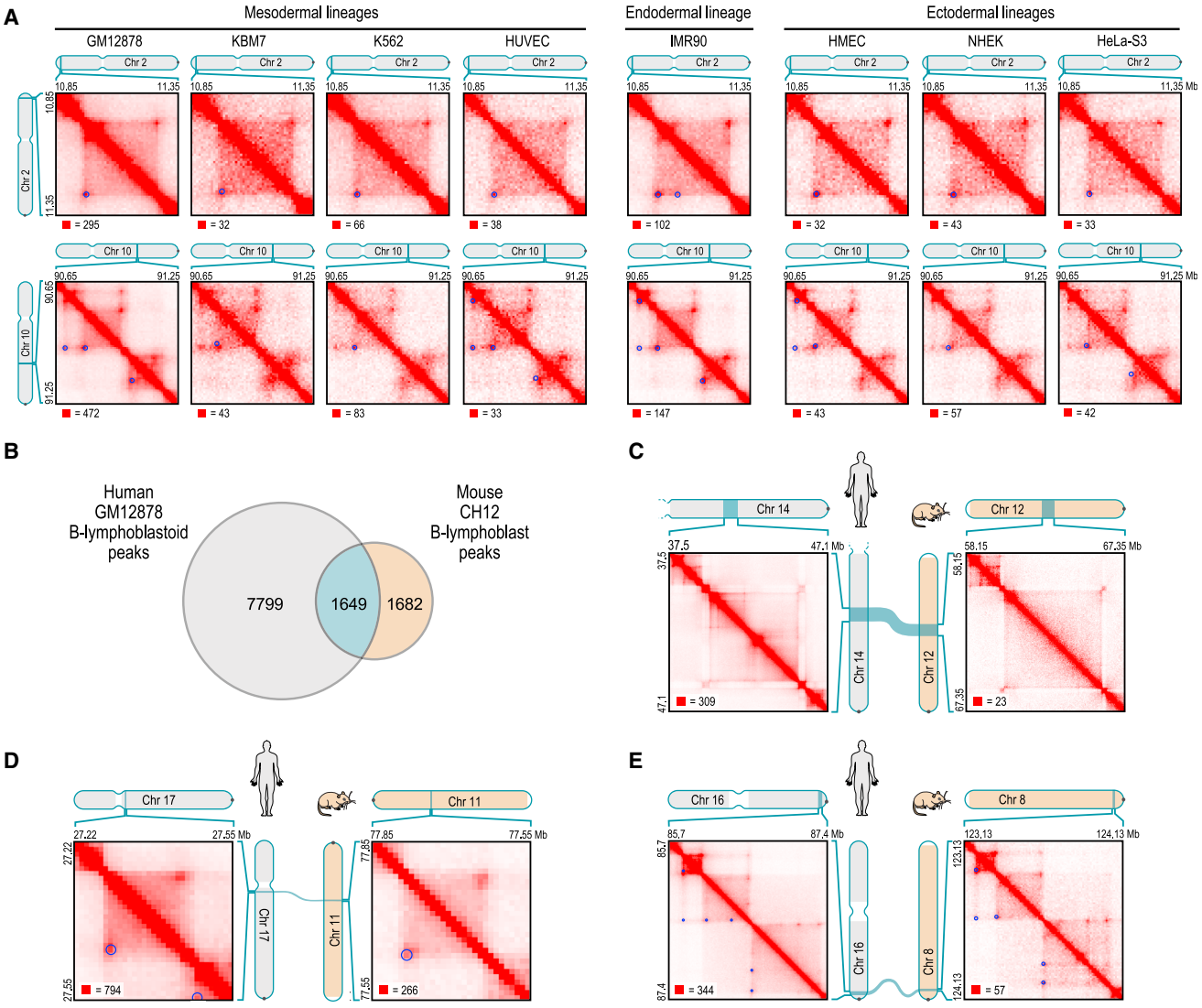

Figure 4. Loops Are Often Preserved across Cell Types and from Human to Mouse

(A) Examples of peak and domain preservation across cell types. Annotated peaks are circled in blue. All annotations are completely independent.

(B) Of the 3,331 loops we annotate in mouse CH12-LX, 1,649 (50%) are orthologous to loops in human GM12878.

(C–E) Conservation of 3D structure in synteny blocks. The contact matrices in (C) are shown at 25 kb resolution. (D) and (E) are shown at 10 kb resolution.

that the appearance of a loop in a cell type was frequently accompanied by the activation of a gene whose promoter overlapped one of the peak loci. For example, a cell-type-specific loop is anchored at the promoter of the gene encoding L-selectin (*SELL*), which is expressed in GM12878 (where the loop is present), but not in IMR90 (where the loop is absent, Figure 5B). Genome-wide, we observed 557 loops in GM12878 that were clearly absent in IMR90. The corresponding peak loci overlapped the promoters of 43 genes that were markedly upregulated (>50-fold) in GM12878, but of only one gene that was markedly upregulated in IMR90. Conversely, we found 510 loops in IMR90 that were clearly absent in GM12878. The corresponding peak loci overlapped the promoters of 94 genes that were markedly upregulated in IMR90, but of only three genes that were markedly upregulated in GM12878. When we compared GM12878 to the five other human cell types for which ENCODE RNA-seq data were available, the results were very similar (Figure 5C; Table S7).

Occasionally, gene activation is accompanied by the emergence of a cell-type-specific network of peaks. Figure 5D illustrates the case of *ADAMTS1*, which encodes a protein involved in fibroblast migration. The gene is expressed in IMR90, where its promoter is involved in six loops. In GM12878, it is not expressed, and the promoter is involved in only two loops. Many of the IMR90 peak loci form transitive peaks with one another (see discussion of "transitivity" below), suggesting that the *ADAMTS1* promoter and the six distal sites may all be located at a single spatial hub.

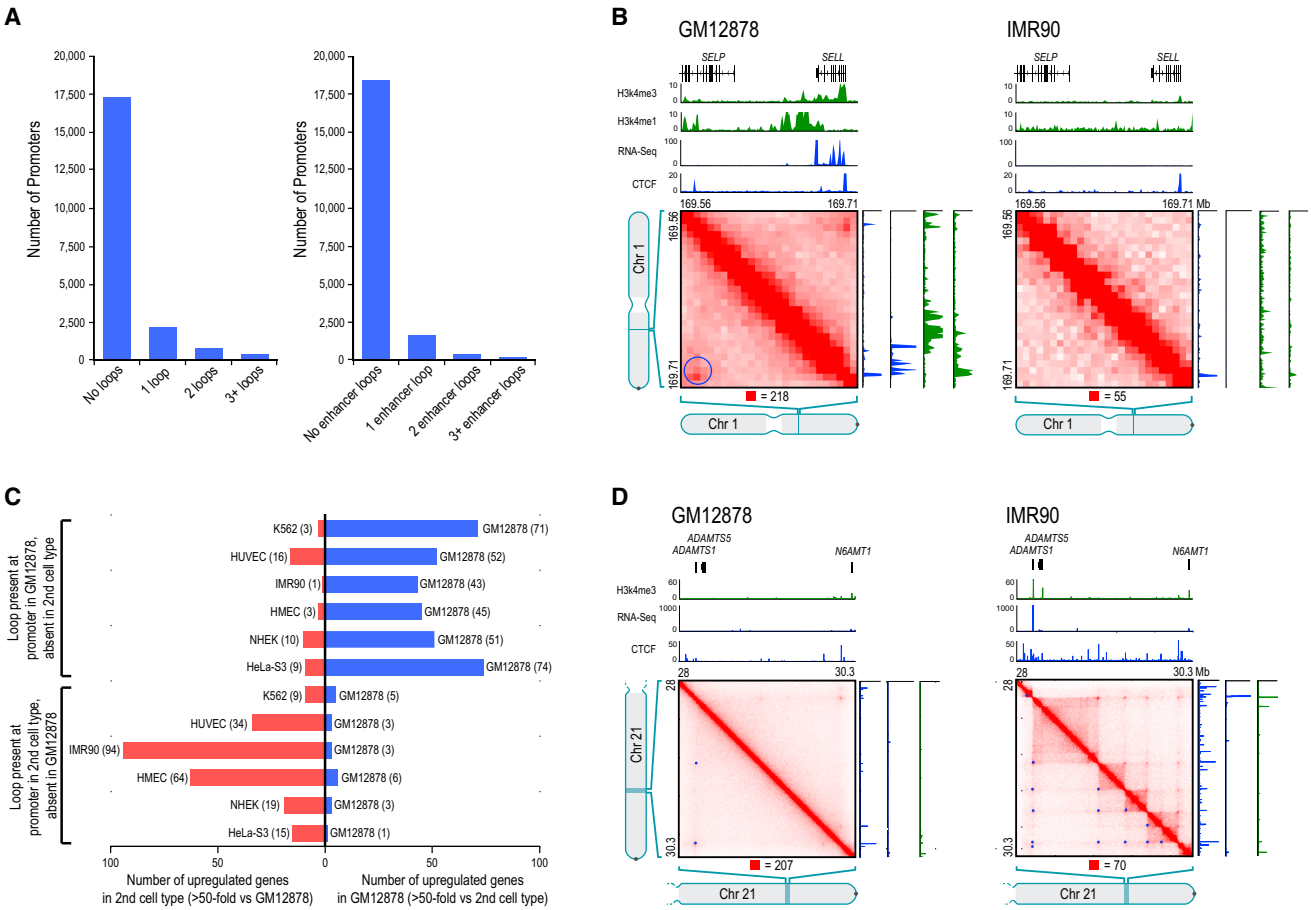

Figure 5. Loops between Promoters and Enhancers Are Strongly Associated with Gene Activation

(A) Histogram showing loop count at promoters (left); restricted to loops where the distal peak locus contains an enhancer (right).

(B) Left: a loop in GM12878, with one anchor at the *SELL* promoter and the other at a distal enhancer. The gene is on. Right: the loop is absent in IMR90, where the gene is off.

(C) Genes whose promoters participate in a loop in GM12878 but not in a second cell type are frequently upregulated in GM12878 and vice versa.

(D) Left: two loops in GM12878 are anchored at the promoter of the inactive *ADAMTS1* gene. Right: a series of loops and domains appear, along with transitive looping. *ADAMTS1* is on.

See also Data S1, VI and Table S7.

These observations are consistent with the classic model in which looping between a promoter and enhancer activates a target gene (Tolhuis et al., 2002; Amano et al., 2009; Ahmadiyeh et al., 2010).

Loops Frequently Demarcate the Boundaries of Contact Domains

A large fraction of peaks (38%) coincide with the corners of a contact domain—that is, the peak loci are located at domain boundaries (Figures 6A and S6). Conversely, a large fraction of domains (39%) had peaks in their corner. Moreover, the appearance of a loop is usually (in 65% of cases) associated with the appearance of a domain demarcated by the loop. Because this configuration is so common, we use the term "loop domain" to refer to contact domains whose endpoints form a chromatin loop.

In some cases, adjacent loop domains (bounded by peak loci *L1-L2* and *L2-L3*, respectively) exhibit transitivity—that is, *L1* and *L3* also correspond to a peak. This may indicate that the three loci simultaneously colocate at a single spatial position. However, many peaks do not exhibit transitivity, suggesting that the corresponding loci do not colocate. Figure 6B shows a region on chromosome 4 exhibiting both configurations.

We also found that overlapping loops are strongly disfavored: pairs of loops *L1-L3* and *L2-L4* (where *L1*, *L2*, *L3* and *L4* occur consecutively in the genome) are found 4-fold less often than expected under a random model (Extended Experimental Procedures).

The Vast Majority of Loops Are Associated with Pairs of CTCF Motifs in a Convergent Orientation

We next wondered whether peaks are associated with specific proteins. We examined the results of 86 chromatin immunoprecipitation sequencing (ChIP-seq) experiments performed by ENCODE in GM12878. We found that the vast majority of peak

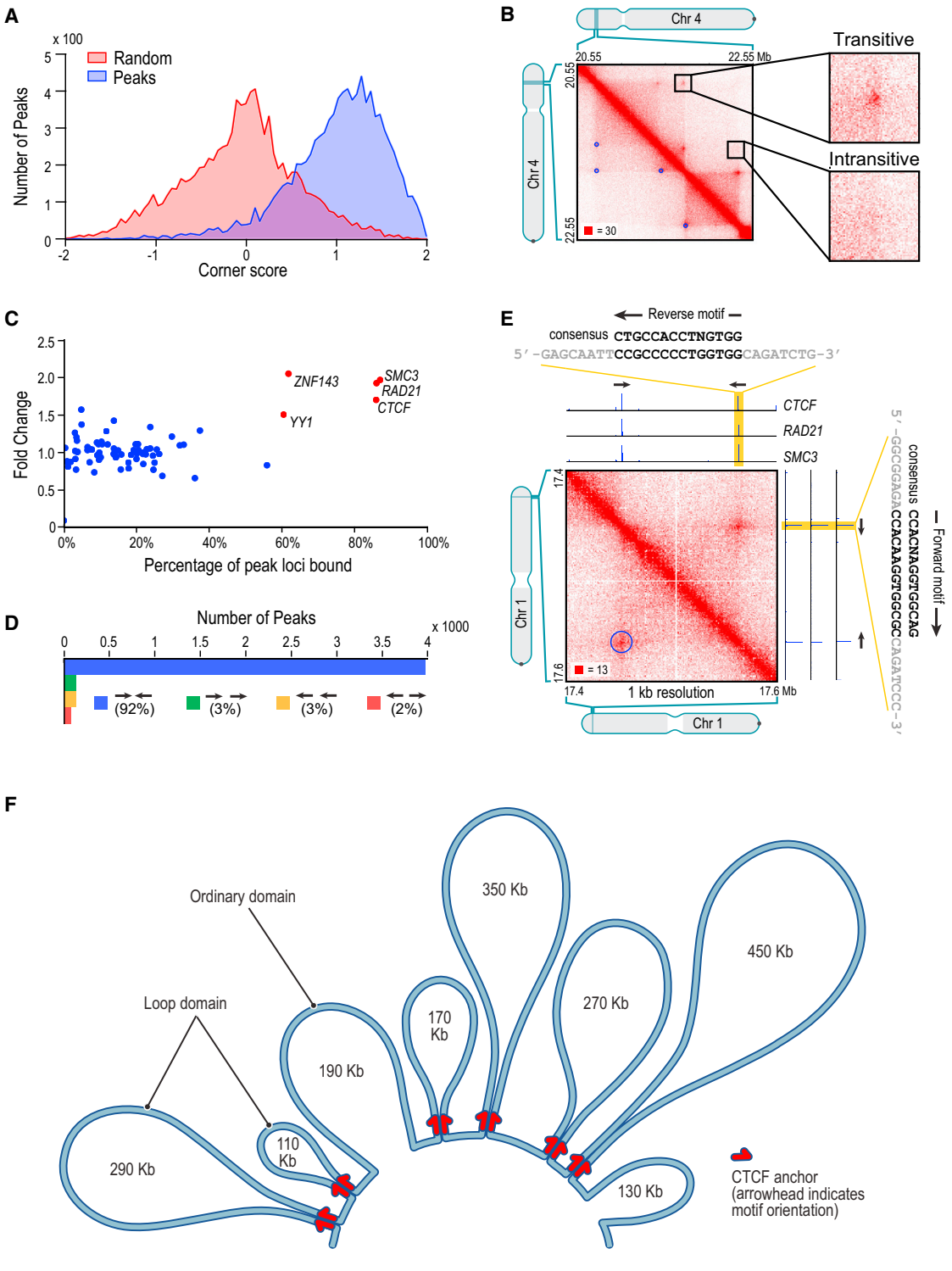

Figure 6. Many Loops Demarcate Contact Domains; The Vast Majority of Loops Are Anchored at a Pair of Convergent CTCF/RAD21/SMC3 Binding Sites

(A) Histograms of corner scores for peak pixels versus random pixels with an identical distance distribution.

(B) Contact matrix for chr4:20.55 Mb–22.55 Mb in GM12878, showing examples of transitive and intransitive looping behavior.

(C) Percent of peak loci bound versus fold enrichment for 76 DNA-binding proteins.

(D) The pairs of CTCF motifs that anchor a loop are nearly all found in the convergent orientation.

(legend continued on next page)

loci are bound by the insulator protein CTCF (86%) and the cohesin subunits RAD21 (86%) and SMC3 (87%) (Figure 6C). This is consistent with numerous reports, using a variety of experimental modalities, that suggest a role for CTCF and cohesin in mediating DNA loops (Splinter et al., 2006; Hou et al., 2008; Phillips and Corces, 2009). Because many of our loops demarcate domains, this observation is also consistent with studies suggesting that CTCF delimits structural and regulatory domains (Xie et al., 2007; Cuddapah et al., 2009; Dixon et al., 2012).

We found that most peak loci encompass a unique DNA site containing a CTCF-binding motif, to which all three proteins (CTCF, SMC3, and RAD21) were bound (5-fold enrichment). We were thus able to associate most of the peak loci (6,991 of 12,903, or 54%) with a specific CTCF-motif "anchor."

The consensus DNA sequence for CTCF-binding sites is typically written as 5′-CCACNAGGTGGCAG-3′. Because the sequence is not palindromic, each CTCF motif has an orientation; we designate the consensus motif above as the "forward" orientation. Thus, a pair of CTCF sites on the same chromosome can have four possible orientations: (1) same direction on one strand, (2) same direction on the other strand, (3) convergent on opposite strands, and (4) divergent on opposite strands.

If CTCF sites were randomly oriented, one would expect all four orientations to occur equally often. But when we examined the 4,322 peaks in GM12878 where the two corresponding peak loci each contained a single CTCF-binding motif, we found that the vast majority (92%) of motif pairs are convergent (Figures 6D and 6E). Overall, the presence, at pairs of peak loci, of bound CTCF sites in the convergent orientation was enriched 102-fold over random expectation (Extended Experimental Procedures). The convergent orientation was overwhelmingly more frequent than the divergent orientation, despite the fact that divergent motifs also lie on opposing strands: in GM12878, the counts were 3,971-78 (51-fold enrichment, convergent versus divergent); in IMR90, 1,456-5 (291-fold); in HMEC, 968-11 (88-fold); in K562, 723-2 (362-fold); in HUVEC, 671-4 (168-fold); in HeLa, 301-3 (100-fold); in NHEK, 556-9 (62-fold); and in CH12-LX, 625-8 (78-fold). This pattern suggests that a pair of CTCF sites in the convergent orientation is required for the formation of a loop.

The observation that looped CTCF sites occur in the convergent orientation also allows us to analyze peak loci containing multiple CTCF-bound motifs to predict which motif instance plays a role in a given loop. In this way, we can associate nearly two-thirds of peak loci (8,175 of 12,903, or 63.4%) with a single CTCF-binding motif.

The specific orientation of CTCF sites at observed peaks provides evidence that our peak calls are biologically correct. Because randomly chosen CTCF pairs would exhibit each of the four orientations with equal probability, the near-perfect association between our loop calls and the convergent orientation could not occur by chance ($p < 10^{-1,900}$, binomial distribution).

In addition, the presence of CTCF and RAD21 sites at many of our peaks provides an opportunity to compare our results to three recent ChIA-PET experiments reported by the ENCODE Consortium (in GM12878 and K562) in which ligation junctions bound to CTCF (or RAD21) were isolated and analyzed. We found strong concordance with our results in all three cases (Li et al., 2012; Heidari et al., 2014) (Extended Experimental Procedures).

The CTCF-Binding Exapted SINEB2 Repeat in Mouse Shows Preferential Orientation with Respect to Loops

In mouse, we found that 7% of peak anchors lie within SINEB2 repeat elements containing a CTCF motif, which has been exapted to be functional. (The spread of CTCF binding via retrotransposition of this element, which contains a CTCF motif in its consensus sequence, has been documented in prior studies [Bourque et al., 2008; Schmidt et al., 2012].) The CTCF motifs at peak anchors in SINEB2 elements show the same strong bias toward convergent orientation seen throughout the genome (89% are oriented toward the opposing loop anchor versus 94% genome-wide). The orientation of these CTCF motifs is aligned with the orientation of the SINEB2 consensus sequence in 97% of cases. This suggests that exaptation of a CTCF in a SINEB2 element is more likely when the orientation of the inserted SINEB2 is compatible with local loop structure.

Diploid Hi-C Maps Reveal Homolog-Specific Features, Including Imprinting-Specific Loops and Massive Domains and Loops on the Inactive X Chromosome

Because many of our reads overlap SNPs, it is possible to use GM12878 phasing data (McKenna et al., 2010; 1000 Genomes Project Consortium et al., 2012) to assign contacts to specific chromosomal homologs (Figure 7A; Table S8). Using these assignments, we constructed a "diploid" Hi-C map of GM12878 comprising both maternal (238 M contacts) and paternal (240 M) maps.

For autosomes, the maternal and paternal homologs exhibit very similar inter- and intrachromosomal contact profiles (Pearson's $R > 0.998$). One interchromosomal difference was notable: an elevated contact frequency between the paternal homologs of chromosome 6 and 11 that is consistent with an unbalanced translocation fusing chr11q:73.5 Mb and all distal loci (a stretch of over 60 Mb) to the telomere of chromosome 6p (Figures 7B and S7B). The signal intensity suggests that the translocation is present in between 1.2% and 5.6% of our cells (Extended Experimental Procedures). We tested this prediction by karyotyping 100 GM12878 cells using Giemsa staining and found three abnormal chromosomes, each showing the predicted

(E) A peak on chromosome 1 and corresponding ChIP-seq tracks. Both peak loci contain a single site bound by CTCF, RAD21, and SMC3. The CTCF motifs at the anchors exhibit a convergent orientation.

(F) A schematic rendering of a 2.1 Mb region on chromosome 20 (48.78–50.88 Mb). Eight domains tile the region, ranging in size from 110 kb to 450 kb; 95% of the region is contained inside a domain (contour lengths are shown to scale). Six of the eight domains are demarcated by loops between convergent CTCF-binding sites located at the domain boundaries. The other two domains are not demarcated by loops. The motif orientation is indicated by the direction of the arrow. Note that not every CTCF-binding site is shown.

See also Figure S6.

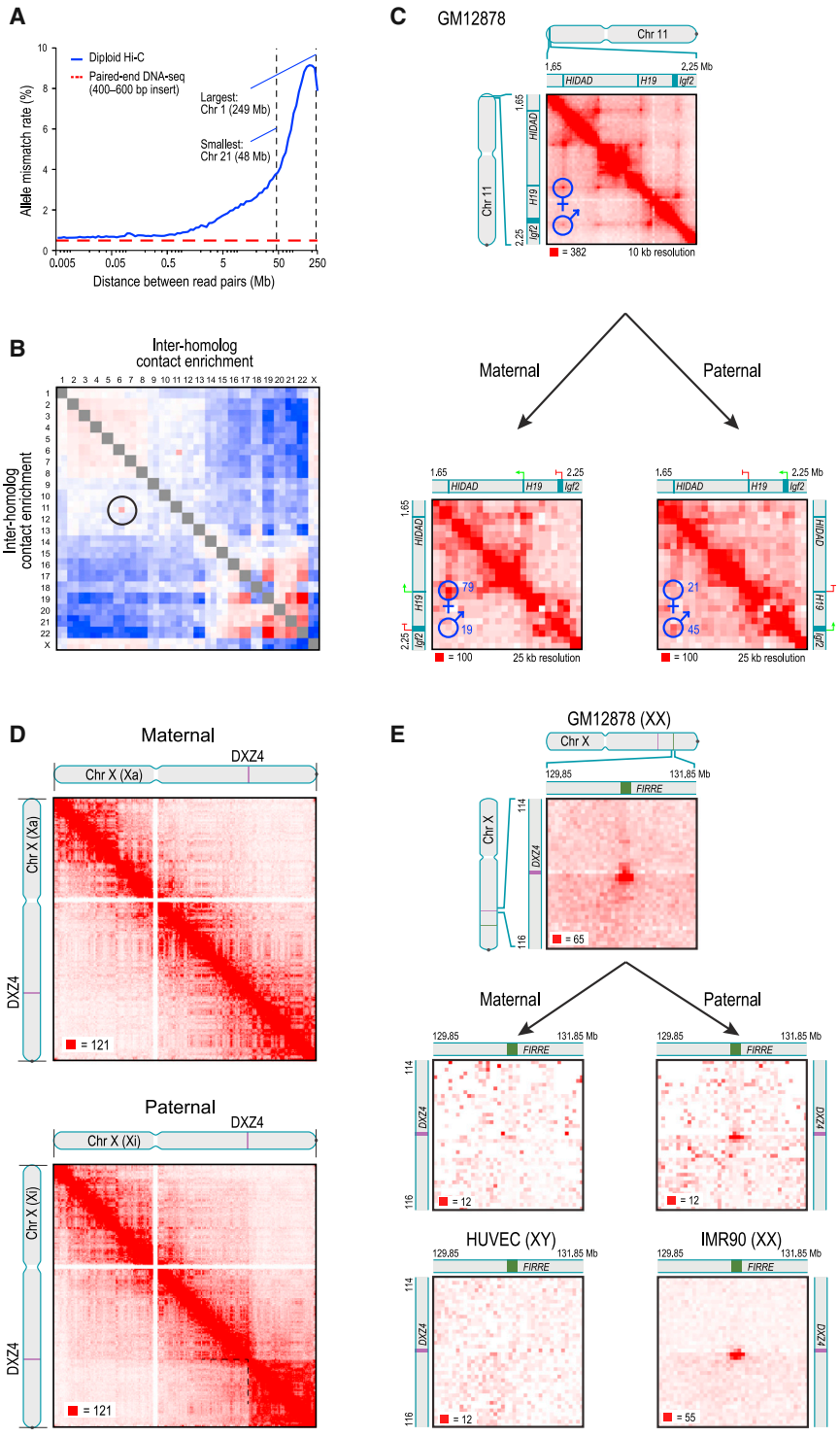

Figure 7. Diploid Hi-C Maps Reveal Superdomains and Superloops Anchored at CTCF-Binding Tandem Repeats on the Inactive X Chromosome

(A) The frequency of mismatch (maternal-paternal) in SNP allele assignment versus distance between two paired read alignments. Intrachromosomal read pairs are overwhelmingly intramolecular.

(B) Preferential interactions between homologs. Left/top is maternal; right/bottom is paternal. The aberrant contact frequency between 6/paternal and 11/paternal (circle) reveals a translocation.

(C) Top: in our unphased Hi-C map of GM12878, we observe two loops joining both the promoter of the maternally-expressed *H19* and the promoter of the paternally-expressed *Igf2* to a distal locus, HIDAD. Using diploid Hi-C maps, we phase these loops: the HIDAD-*H19* loop is present only on the maternal homolog (left) and the HIDAD-*Igf2* loop is present only on the paternal homolog (right).

(D) The inactive (paternal) copy of chromosome X (bottom) is partitioned into two massive "superdomains" not seen in the active (maternal) copy (top). *DXZ4* lies at the boundary. Contact matrices are shown at 500 kb resolution.

(E) The "superloop" between *FIRRE* and *DXZ4* is present in the unphased GM12878 map (top), in the paternal GM12878 map (middle right), and in the map of the female cell line IMR90 (bottom right); it is absent from the maternal GM12878 map (middle left) and the map of the male HUVEC cell line (bottom left). Contact matrices are shown at 50 kb resolution.

See also Figure S7 and Table S8.

translocation, der(6)t(6,11)(pter;q) (Figures S7C–S7F). The Hi-C data reveal that the translocation involves the paternal homologs, which cannot be determined with ordinary cytogenetic methods.

We also observed differences in loop structure between homologous autosomes at some imprinted loci. For instance, the *H19/Igf2* locus on chromosome 11 is a well-characterized case of genomic imprinting. In our unphased maps, we clearly see two loops from a single distal locus at 1.72 Mb (that binds CTCF in the forward orientation) to loci located near the promoters of both *H19* and *Igf2* (both of which bind CTCF in the reverse orientation, i.e., the above consensus motif lies on the opposite strand; see Figure 7C). We refer to this distal locus as the *H19/Igf2* Distal Anchor Domain (HIDAD). Our diploid maps reveal that the loop to the *H19* region is present on the maternal chromosome (from which *H19* is expressed), but the loop to the *Igf2* region is absent or greatly attenuated. The opposite pattern is found on the paternal chromosome (from which *Igf2* is expressed).

Pronounced differences were seen on the diploid intrachromosomal maps of chromosome X. The paternal X chromosome, which is usually inactive in GM12878, is partitioned into two massive domains (0–115 Mb and 115–155.3 Mb). These "superdomains" are not seen in the active, maternal X (Figure 7D). When we examined the unphased maps of chromosome X for the karyotypically normal female cell lines in our study (GM12878, IMR90, HMEC,

NHEK), the superdomains on X were evident, although the signal was attenuated due to the superposition of signals from active and inactive X chromosomes. When we examined the male HUVEC cell line and the haploid KBM7 cell line, we saw no evidence of superdomains (Figure S7G).

Interestingly, the boundary between the superdomains (ChrX: 115 Mb ± 500 kb) lies near the macrosatellite repeat *DXZ4* (ChrX: 114,867,433–114,919,088) near the middle of Xq. *DXZ4* is a CpG-rich tandem repeat that is conserved across primates and monkeys and encodes a long noncoding RNA. In males and on the active X, *DXZ4* is heterochromatic, hypermethylated and does not bind CTCF. On the inactive X, *DXZ4* is euchromatic, hypomethylated, and binds CTCF. *DXZ4* has been hypothesized to play a role in reorganizing chromatin during X inactivation (Chadwick, 2008).

There were also significant differences in loop structure between the chromosome X homologs. We observed 27 large "superloops," each spanning between 7 and 74 Mb, present only on the inactive X chromosome in the diploid map (Figure 7E). The superloops were also seen in all four unphased maps from karyotypically normal XX cells, but were absent in unphased maps from X0 and XY cells (Figure S7I). Two of the superloops (chrX:56.8 Mb-*DXZ4* and *DXZ4*-130.9 Mb) were reported previously in a locus-specific study (Horakova et al., 2012).

Like the peak loci of most other loops, nearly all the superloop anchors bind CTCF (23 of 24). The six anchor regions most frequently associated with superloops are large (up to 200 kb). Four of these anchor regions contain whole long noncoding RNA (lncRNA) genes: *loc550643*, *XIST*, *DXZ4*, and *FIRRE*. Three (*loc550643*, *DXZ4*, and *FIRRE*) contain CTCF-binding tandem repeats that only bind CTCF on the inactive homolog.

DISCUSSION

Using the in situ Hi-C protocol, we probed genomic architecture with high resolution; in the case of GM12878 lymphoblastoid cells, better than 1 kb. We observe the presence of contact domains that were too small (median length = 185 kb) to be seen in previous maps. Loci within a domain interact frequently with one another, have similar patterns of chromatin modifications, and exhibit similar long-range contact patterns. Domains tend to be conserved across cell types and between human and mouse. When the pattern of chromatin modifications associated with a domain changes, the domain's long-range contact pattern also changes. Domains exhibit at least six distinct patterns of long-range contacts (subcompartments), which subdivide the two compartments that we previously reported based on low resolution data. The subcompartments are each associated with distinct chromatin patterns. It is possible that the chromatin patterns play a role in bringing about the long-range contact patterns, or vice versa.

Our data also make it possible to create a genome-wide catalog of chromatin loops. We identified loops by looking for pairs of loci that have significantly more contacts with one another than they do with other nearby loci. In our densest map (GM12878), we observe 9,448 loops.

The loops reported here have many interesting properties. Most loops are short (<2 Mb) and strongly conserved across cell types and between human and mouse. Promoter-enhancer loops are common and associated with gene activation. Loops tend not to overlap; they often demarcate contact domains, and may establish them. CTCF and the cohesin subunits RAD21 and SMC3 associate with loops; each of these proteins is found at over 86% of loop anchors.

The most striking property of loops is that the pair of CTCF motifs present at the loop anchors occurs in a convergent orientation in >90% of cases (versus 25% expected by chance). The importance of motif orientation between loci that are separated by, on average, 360 kb is surprising and must bear on the mechanism by which CTCF and cohesin form loops, which seems likely to involve CTCF dimerization. Experiments in which the presence or orientation of CTCF sites is altered may enable the engineering of loops, domains, and other chromatin structures.

It is interesting to compare our results to those seen in previous reports. The contact domains we observe are similar in size to the "physical domains" that have been reported in Hi-C maps of *Drosophila* (Sexton et al., 2012) and to the "topologically constrained domains" (mean length: 220 kb) whose existence was demonstrated in the 1970s and 1980s in structural studies of human chromatin (Cook and Brazell, 1975; Vogelstein et al., 1980; Zehnbauer and Vogelstein, 1985). On the other hand, the domains we observe are much smaller than the TADs (1 Mb) (Dixon et al., 2012) that have been reported in humans and mice on the basis of lower-resolution contact maps. This is because detecting TADs involves detection of domain boundaries. With higher resolution data, it is possible to detect additional boundaries beyond those seen in previous maps. Interestingly, nearly all the boundaries we observe are associated with either a subcompartment transition (that occur approximately every 300 kb), or a loop (that occur approximately every 200 kb); and many are associated with both.

Our annotation identifies many fewer loops than were reported in several recent high-throughput studies, despite the fact that we have more data. The key reason is that we call peaks only when a pair of loci shows elevated contact frequency relative to the local background—that is, when the peak pixel is enriched as compared to other pixels in its neighborhood. In contrast, prior studies have defined peaks by comparing the contact frequency at a pixel to the genome-wide average (Sanyal et al., 2012; Jin et al., 2013). This latter definition is problematic because many pixels within a domain can be annotated as peaks despite showing no local increase in contact frequency. Papers using the latter definition imply the existence of more than 100,000 loops (1,187 loops were reported in 1% of the genome [Sanyal et al., 2012]) or even more than 1 million loops (reported in a genome-wide Hi-C study [Jin et al., 2013]). The vast majority of the loops annotated by these papers show no enrichment relative to the local background when examined one-by-one and no enrichment with respect to any published Hi-C data set when analyzed using APA (see Extended Experimental Procedures; Figure S8; Data S2). This suggests that these peak annotations may correspond to pairs of loci that lie in the same domain or compartment, but rarely correspond to loops.

We created diploid Hi-C maps by using polymorphisms to assign contacts to distinct chromosomal homologs. We found that the inactive X chromosome is partitioned into two large

superdomains whose boundary lies near the locus of the lncRNA *DXZ4*. We also detect a network of long-range superloops, the strongest of which are anchored at locations containing lncRNA genes (*loc550643*, *XIST*, *DXZ4*, and *FIRRE*). With the exception of *XIST*, all of these lncRNAs contain CTCF-binding tandem repeats that bind CTCF only on the inactive X.

In our original report on Hi-C, we observed that Hi-C maps can be used to study physical models of genome folding, and we proposed a fractal globule model for genome folding at the megabase scale. The kilobase-scale maps reported here allow the physical properties of genome folding to be probed at much higher resolution. We will report such studies elsewhere.

Just as loops bring distant DNA loci into close spatial proximity, we find that they bring disparate aspects of DNA biology—domains, compartments, chromatin marks, and genetic regulation—into close conceptual proximity. As our understanding of the physical connections between DNA loci continues to improve, our understanding of the relationships between these broader phenomena will deepen.

EXPERIMENTAL PROCEDURES

In Situ Hi-C Protocol

All cell lines were cultured following the manufacturer's recommendations. Two to five million cells were crosslinked with 1% formaldehyde for 10 min at room temperature. Nuclei were permeabilized. DNA was digested with 100 units of MboI, and the ends of restriction fragments were labeled using biotinylated nucleotides and ligated in a small volume. After reversal of crosslinks, ligated DNA was purified and sheared to a length of ~400 bp, at which point ligation junctions were pulled down with streptavidin beads and prepped for Illumina sequencing. Dilution Hi-C was performed as in Lieberman-Aiden et al. (2009).

3D-FISH

3D DNA-FISH was performed as in Beliveau et al. (2012) with minor modifications.

Hi-C Data Pipeline

All sequence data were produced using Illumina paired-end sequencing. We processed data using a custom pipeline that was optimized for parallel computation on a cluster. The pipeline uses BWA (Li and Durbin, 2010) to map each read end separately to the b37 or mm9 reference genomes; removes duplicate and near-duplicate reads; removes reads that map to the same fragment; and filters the remaining reads based on mapping quality score. Contact matrices were generated at base pair delimited resolutions of 2.5 Mb, 1 Mb, 500 kb, 250 kb, 100 kb, 50 kb, 25 kb, 10 kb, and 5 kb, as well as fragment-delimited resolutions of 500 f, 200 f, 100 f, 50 f, 20 f, 5 f, 2 f, and 1 f. For our largest maps, we also generated a 1 kb contact matrix. Normalized contact matrices are produced at all resolutions using Knight and Ruiz (2012).

Annotation of Domains: Arrowhead

To annotate domains, we apply an "arrowhead" transformation, defined as $A_{i,i+d} = (M^*_{i,i-d} - M^*_{i,i+d})/(M^*_{i,i-d} + M^*_{i,i+d})$. M^* denotes the normalized contact matrix (see Figures S2A–S2F). This is equivalent to calculating a matrix equal to -1*(observed/expected $-$ 1), where the expected model controls for local background and distance from the diagonal in the simplest possible way: the "expected" value at $i,i + d$ is simply the mean of the observed values at $i,i - d$ and $i,i + d$. $A_{i,i+d}$ will be strongly positive if locus $i - d$ is inside a domain and locus $i + d$ is not. If the reverse is true, $A_{i,i+d}$ will be strongly negative. If the loci are both inside or both outside a domain, $A_{i,i+d}$ will be close to zero. Consequently, if there is a domain at $[a,b]$, we find that A takes on very negative values inside a triangle whose vertices lie at $[a,a]$, $[a,b]$, and $[(a + b)/2,b]$ and very positive values inside a triangle whose vertices lie at $[(a + b)/2,b]$, $[b,b]$, and $[b,2b - a]$. The size and positioning of these triangles creates the arrowhead-shaped feature that replaces each domain in M^*. A "corner score" matrix, indicating each pixel's likelihood of lying at the corner of a domain, is efficiently calculated from the arrowhead matrix using dynamic programming.

Assigning Loci to Subcompartments

To cluster loci based on long-range contact patterns, we constructed a 100 kb resolution interchromosomal contact matrix such that loci from odd chromosomes appeared on the rows, and loci from even chromosomes appeared on the columns. (Intrachromosomal data and data involving chromosome X were excluded.) We cluster this matrix using the Python package *scikit*. For subcompartment B4, the 100 kb interchromosomal matrix for chromosome 19 was constructed and clustered separately, using the same procedure.

Annotation of Peaks: HiCCUPS

Our peak-calling algorithm examines each pixel in a Hi-C contact matrix and compares the number of contacts in the pixel to the number of contacts in a series of regions surrounding the pixel. The algorithm thus identifies "enriched pixels" $M^*_{i,j}$ where the contact frequency is higher than expected and where this enrichment is not the result of a larger structural feature. For instance, we rule out the possibility that the enrichment of pixel $M^*_{i,j}$ is the result of L_i and L_j lying in the same domain by comparing the pixel's contact count to an expected model derived by examining the "lower-left" neighborhood. (The "lower-left" neighborhood samples pixels $M_{i',j'}$ where $i \leq i' \leq j' \leq j$; if a pixel is in a domain, these pixels will necessarily be in the same domain.) We require that the pixel being tested contain at least 50% more contacts than expected based on the lower-left neighborhood and the enrichment be statistically significant after correcting for multiple hypothesis testing (False Discovery Rate < 10%). The same criteria are applied to three other neighborhoods. Thus, to be labeled an enriched pixel, a pixel must be significantly enriched relative to four neighborhoods: (1) pixels to its lower-left, (2) pixels to its left and right, (3) pixels above and below, and (4) a donut surrounding the pixel of interest (Figure 3A). The resulting enriched pixels tend to form contiguous interaction regions comprising 5–20 pixels each. We define the "peak pixel" (or simply the "peak") to be the pixel in an interaction region with the most contacts.

Because of the enormous number of pixels that must be examined, this calculation requires weeks of central processing unit (CPU) time to execute. (For instance, at a matrix resolution of 5 kb, the algorithm must be run on 20 billion pixels.) To accelerate it, we created a highly parallelized implementation using general-purpose graphical processing units resulting in a 200-fold speedup.

Aggregate Peak Analysis

We perform APA on 10 kb resolution contact matrices. To measure the aggregate enrichment of a set of putative peaks in a contact matrix, we plot the sum of a series of submatrices derived from that contact matrix. Each of these submatrices is a 210 kb × 210 kb square centered at a single putative peak in the upper triangle of the contact matrix. The resulting APA plot displays the total number of contacts that lie within the entire putative peak set at the center of the matrix; the entry immediately to the right of center corresponds to the total number of contacts in the pixel set obtained by shifting the peak set 10 kb to the right; the entry two positions above center corresponds to an upward shift of 20 kb and so on. Focal enrichment across the peak set in aggregate manifests as larger values at the center of the APA plot. The APA plots shown only include peaks whose loci are at least 300 kb apart.

ACCESSION NUMBERS

The Gene Expression Omnibus (GEO) accession number for the data sets reported in this paper is GSE63525. The dbGaP accession number for the HeLa data reported in this paper is phs000640.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, eight figures, two data files, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.11.021.

AUTHOR CONTRIBUTIONS

E.L.A. conceived this project. S.S.P.R., M.H.H., E.K.S., and E.L.A. designed experiments. S.S.P.R., E.K.S., I.D.B., A.D.O., and M.H.H. performed Hi-C experiments. E.K.S. and I.D.B. performed 3D-FISH experiments. N.C.D. built the computational pipeline for Hi-C data. N.C.D. and J.T.R. built the visualization system for Hi-C data. S.S.P.R., M.H.H., N.C.D., A.L.S., I.M., E.S.L., and E.L.A. analyzed data. S.S.P.R., M.H.H., N.C.D., E.S.L., and E.L.A. prepared the manuscript.

ACKNOWLEDGMENTS

This paper is dedicated to the memory of Aharon Lieberman. Our work was supported by an NSF Graduate Research Fellowship (DGE0946799 and DGE1144152) to M.H.H., an NIH New Innovator Award (OD008540-01), an NSF Physics Frontier Center (PHY-1427654, Center for Theoretical Biological Physics), an NHGRI CEGS (HG006193), NVIDIA, IBM, Google, a CPRIT Scholar Award (R1304), a McNair Medical Institute Scholar Award, and the President's Early Career Award in Science and Engineering to E.L.A., and an NHGRI grant (HG003067) to E.S.L. We thank Leslie Gaffney, Lauren Solomon, and Bang Wong for assistance with figures; BCM's Integrated Microscopy Core, Michael Mancini, Justin Demmerle, Wendy Salmon, Fabio Stossi, Radhika Dandekar, Sanjay Krishna, and especially Asha Multani for microscopy assistance; Aharon Lieberman, Aviva Presser Aiden, Nicholas Christakis, James Lupski, José Onuchic, Mitchell Guttman, Andreas Gnirke, Louise Williams, Chad Nusbaum, John Bohannon, Olga Dudchenko, and the Aiden laboratory for discussions; and Robbyn Issner and Broad's ENCODE group for several cell lines. The Center for Genome Architecture is grateful to Janice, Robert, and Cary McNair for support. A provisional patent covering in situ Hi-C and related methods has been filed. All sequence data reported in this paper that were not derived from HeLa cells have been deposited at GEO (http://www.ncbi.nlm.nih.gov/geo/) (GSE63525). Some of the genome sequences described in this research were derived from a HeLa cell line. Henrietta Lacks, and the HeLa cell line that was established from her tumor cells without her knowledge or consent in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. The HeLa data generated from this research were submitted to the database of Genotypes and Phenotypes (dbGaP) as a substudy under accession number phs000640.

Received: October 12, 2014
Revised: November 5, 2014
Accepted: November 13, 2014
Published: December 11, 2014

REFERENCES

Ahmadiyeh, N., Pomerantz, M.M., Grisanzio, C., Herman, P., Jia, L., Almendro, V., He, H.H., Brown, M., Liu, X.S., Davis, M., et al. (2010). 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proc. Natl. Acad. Sci. USA *107*, 9742–9746.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. Dev. Cell *16*, 47–57.

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.

Beliveau, B.J., Joyce, E.F., Apostolopoulos, N., Yilmaz, F., Fonseka, C.Y., McCole, R.B., Chang, Y., Li, J.B., Senaratne, T.N., Williams, B.R., et al. (2012). Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc. Natl. Acad. Sci. USA *109*, 21301–21306.

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

Blackwood, E.M., and Kadonaga, J.T. (1998). Going the distance: a current view of enhancer action. Science *281*, 60–63.

Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L., Srinivasan, K.G., Chew, J.-L.L., Ruan, Y., Wei, C.-L.L., Ng, H.H., and Liu, E.T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. *18*, 1752–1762.

Chadwick, B.P. (2008). DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. Genome Res. *18*, 1259–1269.

Cook, P.R., and Brazell, I.A. (1975). Supercoils in human DNA. J. Cell Sci. *19*, 261–279.

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

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.

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

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

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.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

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

Finlan, L.E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J.R., and Bickmore, W.A. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS Genet. *4*, e1000039.

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.

Gaszner, M., and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nat. Rev. Genet. *7*, 703–713.

Gavrilov, A.A., Gushchanskaya, E.S., Strelkova, O., Zhironkina, O., Kireev, I.I., Iarovaia, O.V., and Razin, S.V. (2013). Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. Nucleic Acids Res. *41*, 3563–3575.

Hahn, M.A., Wu, X., Li, A.X., Hahn, T., and Pfeifer, G.P. (2011). Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks. PLoS ONE *6*, e18844.

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. Published online September 16, 2014. http://dx.doi.org/10.1101/gr.176586.114.

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., et al. (2013). Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. *41*, 827–841.

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

Hou, C., Zhao, H., Tanimoto, K., and Dean, A. (2008). CTCF-dependent enhancer-blocking by alternative chromatin loop formation. Proc. Natl. Acad. Sci. USA *105*, 20398–20403.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.

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

Knight, P., and Ruiz, D. (2012). A fast algorithm for matrix balancing. IMA J. Numer. Anal. Published online October 26, 2012. http://dx.doi.org/10.1093/imanum/drs019.

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

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

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

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

Mukherjee, S., Erickson, H., and Bastia, D. (1988). Enhancer-origin interaction in plasmid R6K involves a DNA loop mediated by initiator protein. Cell *52*, 375–383.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

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.

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194–1211.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Schleif, R. (1992). DNA looping. Annu. Rev. Biochem. *61*, 199–223.

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.

Sexton, T., Schober, H., Fraser, P., and Gasser, S.M. (2007). Gene regulation through nuclear organization. Nat. Struct. Mol. Biol. *14*, 1049–1055.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N., and de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. *20*, 2349–2354.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol. Cell *10*, 1453–1465.

Vogel, M.J., Guelen, L., de Wit, E., Peric-Hupkes, D., Lodén, M., Talhout, W., Feenstra, M., Abbas, B., Classen, A.K., and van Steensel, B. (2006). Human heterochromatin proteins form large domains containing KRAB-ZNF genes. Genome Res. *16*, 1493–1504.

Vogelstein, B., Pardoll, D.M., and Coffey, D.S. (1980). Supercoiled loops and eucaryotic DNA replicaton. Cell *22*, 79–85.

Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M., and Lander, E.S. (2007). Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc. Natl. Acad. Sci. USA *104*, 7145–7150.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. *43*, 1059–1065.

Zehnbauer, B.A., and Vogelstein, B. (1985). Supercoiled loops and the organization of replication and transcription in eukaryotes. BioEssays *2*, 52–54.