Published in final edited form as:
Cell. 2014 December 18; 159(7): 1665–1680. doi:10.1016/j.cell.2014.11.021.

A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping

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

[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 Harvard and Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA

[5]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA

[6]Department of Biology, MIT, Cambridge, MA 02139, USA

[7]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

[8]Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA

Summary

We use *in situ* Hi-C to probe the three-dimensional 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-kilobase resolution. We find that genomes are partitioned into local domains, 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.

eToC Blurb

Contact: lander@broadinstitute.org (E.S.L.), erez@erez.com (E.L.A.).
[9]Co-first author.

Author Contributions: Author contributions are as follows: E.L.A. conceived of 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.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.S. and I.M. analyzed data; S.S.P.R., M.H.H., N.C.D., E.S.L., and E.L.A. prepared the manuscript.

Loop anchors typically occur at domain boundaries and bind CTCF in a convergent orientation, with the asymmetric motifs ‘facing’ one another. On the inactive X-chromosome, large imprinted loops are anchored at CTCF-binding repeats. Loops are conserved across cell types and species.

Introduction

The spatial organization of the human genome is known to play an important role in the transcriptional control of genes (Bickmore, 2013; Cremer and Cremer, 2001; Sexton et al., 2007). Yet important questions remain, such as how promoters are affected by distal regulatory elements such as enhancers and how intervening insulator elements can abrogate these effects (Banerji et al., 1981; Blackwood and Kadonaga, 1998; Gaszner and Felsenfeld, 2006). Both phenomena have long been presumed to involve the formation of protein-mediated loops bringing pairs of genomic sites that lie far apart along the linear genome into close physical proximity within the nucleus (Schleif, 1992). Loops joining promoters and enhancers have been suggested to mediate enhancer function by drawing transcription factors close to the genes that they regulate (Pennacchio et al., 2013; Ptashne, 1986), while loops joining insulator elements have been proposed as a mechanism to create segregated chromatin domains, excluding enhancers lying outside the domain (Phillips and Corces, 2009).

The existence of DNA loops was first demonstrated in the 1980s based on studies of operons in prokaryotes and in phage (Schleif, 1992). These early studies convincingly demonstrated that DNA looping played a role in transcription, replication, and recombination, using methods such as differential gel electrophoresis, protein cooperativity, enzymatic protection assays, careful studies of DNA bending and torsion, and, most dramatically, direct visualization of entire loops by electron microscopy (Dunn et al., 1984; Eismann et al., 1987; Griffith et al., 1986; Krämer et al., 1987; Mukherjee et al., 1988; Oehler et al., 1990). In one seminal study, the binding of a protein to sites at opposite ends of a restriction fragment created a loop, thereby promoting 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).

Loops are believed to play a significant role in eukaryotes as well. In mammals, the DNA binding protein CTCF is reported to be strongly associated with DNA loops (Phillips and Corces, 2009). Chromatin immunoprecipitation (ChIP) experiments reveal tens of thousands of CTCF-binding sites across the genome, which tend to occur at a highly specific sequence motif (Kim et al., 2007; Xie et al., 2007). In transgenic assays, the presence of an intervening CTCF-binding site blocks the effects of distal enhancers on gene promoters, and CTCF is often thought to be an insulator protein that delimits regulatory domains. CTCF is capable of forming dimers *in vivo* (Yusufzai et al., 2004), suggesting that it may mediate chromatin looping, possibly by tethering DNA loci to subnuclear structures (Dunn et al., 2003; Yusufzai and Felsenfeld, 2004; Yusufzai et al., 2004). Notably, the behavior of CTCF is not always consistent with an insulator role; in reporter gene assays, its behavior often resembles that of a transcription factor, exhibiting the characteristics of a transcriptional activator (Vostrov, 1997) or repressor (Filippova et al., 1996; Klenova et al., 1993; Köhne et al., 1993; Lobanenkov et al., 1990) depending on the context.

Over the past quarter-century, new methods have been developed to assess the three-dimensional architecture of the cell nucleus *in vivo*. Some of these approaches have been based on direct visualization of DNA loci by means of fluorescent in situ hybridization (FISH) (Amano et al., 2009; Gerasimova et al., 2000). For instance, FISH was used in *Drosophila* cells to visualize a loop formed between adjacent Gypsy insulators, each tethered to the nuclear periphery. When a third Gypsy element was introduced between the original pair, it also became localized to the nuclear periphery, subdividing the structure into two disjoint loops (Gerasimova et al., 2000).

A different family of methods, derived from “cyclization enhancement,” use molecular biology in lieu of imaging. They include nuclear ligation assay (Cullen et al., 1993) and the widely-employed chromosome conformation capture (Dekker et al., 2002). Such approaches have been used to study DNA looping at specific loci, interrogating both specific promoter-enhancer (Spilianakis and Flavell, 2004; Tolhuis et al., 2002) and specific insulator-insulator loops (Hou et al., 2008; Kurukuti et al., 2006; Murrell et al., 2004; Splinter et al., 2006). Technical improvements have allowed the examination of several loci simultaneously (5C, (Dostie et al., 2006)) or all loci bound by a particular protein (CHIA-PET, (Fullwood et al., 2009)). Efforts to annotate loops in a high-throughput fashion using these techniques have reported numerous promoter-enhancer and CTCF-mediated loops.

To interrogate all pairs of 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 (Dixon et al., 2012; Kalhor et al., 2012; Lieberman-Aiden et al., 2009; Sexton et al., 2012). Subsequent analyses have suggested the presence of smaller domains, and have led to the important proposal that compartments are partitioned into condensed structures roughly one megabase in size, dubbed “topologically associated domains” (TADs) or “topological domains” (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 (Ay et al., 2014; Jin et al., 2013; Lin et al., 2012; Sexton et al., 2012).

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 terabases of sequence data recording over 15 billion distinct contacts; they are larger, by an order of magnitude, than all published Hi-C datasets combined. Using these maps, we are able to clearly discern local domain structure, intricate 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 kilobase. They show that the genome is partitioned into domains that are associated with particular patterns of histone marks and that segregate into at least six sub-compartments, distinguished by unique long-range contact patterns. Using the maps, we identify ~10,000 distinct loops across the genome and study their properties, including their strong association with gene activation and their tendency to demarcate domains. Strikingly, the vast majority of loop anchors bind CTCF. Moreover, the two CTCF motifs that occur at the anchors of a loop are found in a convergent orientation – that is, with the asymmetric CTCF motifs 'facing' one another – over 90% of the time. The diploid maps show that the inactive X-chromosome is partitioned into two massive domains, and contains large loops anchored at CTCF-binding repeats.

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 and DNA-DNA proximity ligation is performed in intact nuclei. Our *in situ* Hi-C protocol involves cross-linking 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). The *in situ* Hi-C protocol described above resembles a recently published single-cell Hi-C protocol (Nagano et al., 2013), which also performed DNA-DNA proximity ligation inside nuclei in order 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). 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 much faster, requiring three days instead of seven (Extended Experimental Procedures). Third, it enables higher resolution and more efficient cutting of chromatinized DNA, for instance, through the use of a 4-cutter (MboI) rather than a 6-cutter (typically, HindIII) (Figure S1A).

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 1Mb or 1Kb), 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 do not align uniquely to the genome, that correspond to unligated fragments, or that are duplicates.) The contact matrix can be visualized as a heatmap, whose entries we call "pixels". An "interval" refers to a (one-dimensional) set of consecutive loci; the contacts between two intervals thus form a "rectangle" or "square" in the contact matrix. We define “matrix resolution” as the locus size used to construct a particular contact matrix and “map resolution” as the smallest locus size such that 80% of loci have at least 1000 contacts. The map resolution is meant to reflect the finest scale at which one can reliably discern local features when visually examining the data.

Contact maps spanning 9 cell lines containing over 15 billion contacts

We constructed *in situ* Hi-C maps of 9 cell lines in human and mouse (Table S1). Whereas our original Hi-C experiments had a map resolution of 1Mb, these maps have a resolution of 1Kb or 5Kb. Our largest map, in human GM12878 B-lymphoblastoid cells, aggregates the results of nine biological replicate experiments derived from independent cell cultures. It contains 4.9 billion pairwise contacts and has a map resolution of 950bp (“kilobase resolution”). We used this Hi-C map to construct contact matrices with locus sizes ranging from 2.5Mb to 1Kb.

We also generated eight *in situ* Hi-C maps at 5kb 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 of these maps contains between 395M and 1.1B contacts.

To test reproducibility, we compared our “primary” GM12878 map (2.6 billion contacts from a single culture) to a “replicate” map (2.3 billion contacts aggregated from experiments on eight other samples). The results were strongly correlated both visually and statistically (Pearson’s R>0.998, 0.996, 0.96 and 0.85 at matrix resolutions of 500, 50, 5, and 1Kb; $p<10^{-324}$ in all three cases, bivariate normal distribution) (Figure 1B–D, S1F–K, Extended Experimental Procedures). We compared biological replicates in IMR90, HMEC, K562, KBM7, and CH12-LX with similar results.

To ensure that our results were comparable with those of previous Hi-C experiments, we used our original dilution Hi-C protocol to generate a map of GM12878 with 3.2 billion contacts; the *in situ* and dilution Hi-C showed high reproducibility (R>0.96,0.90,0.87 at 500,50,25Kb; $p<10^{-324}$ in all three cases, bivariate normal distribution). We repeated this procedure in IMR90, HMEC, NHEK, HUVEC, CH12-LX with similar results.

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 crosslinking time, restriction enzyme, ligation volume/time, and biotinylated nucleotide. The experiments demonstrated that our findings were robust to particular experimental conditions (see the sections on loop calling). In total, 201 independent Hi-C experiments were successfully performed.

To identify fine-scale features in Hi-C maps, it is essential to account for non-uniformities in coverage due to the number of restriction sites at a locus or the accessibility of those sites to cutting (Cournac et al., 2012; Hu et al., 2012; Imakaev et al., 2012; Lieberman-Aiden et al., 2009; Yaffe and Tanay, 2011). Either circumstance would increase the number of restriction fragments at the locus available for ligation, and thus the frequency of contacts involving the locus and any other locus. We account for these non-uniformities by normalizing each contact matrix using a matrix-balancing algorithm due to Knight and Ruiz (2012). We also used three other published Hi-C bias-correction methods (Cournac et al., 2012; Imakaev et al., 2012; Lieberman-Aiden et al., 2009); all produced similar results (Extended Experimental Procedures).

The genome is partitioned into small domains

We next sought to use the vastly higher (200- to 1000-fold) map resolution of the present data to re-examine the three-dimensional partitioning of the genome. In our earlier experiments at 1Mb map resolution, we saw large squares of enhanced contact frequency tiling the diagonal of the contact matrices. These squares partitioned the genome into 5–20Mb intervals, which we here call “megadomains.” On opposite sides of a megadomain boundary, the contact frequency between pairs of loci drops sharply. Megadomains are very frequently preserved across cell types.

We also found that individual 1Mb 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 1Mb loci in an interval exhibit the same genome-wide contact pattern (Kalhor et al., 2012; Lieberman-Aiden et al., 2009; Sexton et al., 2012). Compartment A is highly enriched for open chromatin, and correlates strongly with DNaseI accessibility, active genes, and H3K36me3. Compartment B is enriched for closed chromatin.

In our new, higher resolution maps, we observe many small squares of enhanced contact frequency that tile the diagonal of each contact matrix (Figure 2A). We used a dynamic programming algorithm to annotate these domains genome-wide. (Results using a previously published domain-calling algorithm (Dixon et al., 2012) were similar.) The observed domains range in size from 40Kb to 3Mb (median size 185Kb). 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). Domains are often preserved across cell type (50–67% of domains found in an alternate cell type are also found in GM12878) (Figure S2M,N).

The presence of smaller domains in Hi-C maps is consistent with other recent studies (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). The domains we observe are similar in size to the “physical domains” that have been seen in Hi-C maps of *Drosophila*, where extremely high map resolution is possible due to the smaller genome size (100Kb; Sexton et al. (2012)). They are also similar in size to the chromatin state domains (median size: 200kb) that were recently annotated in humans by clustering epigenetic marks (Julienne et al., 2013), and to the “topologically constrained domains” (mean length: 220kb) reported in structural studies of mammalian chromatin dating back to the mid-1970s (Cook and Brazell, 1975; Vogelstein et al., 1980; Zehnbauer and Vogelstein, 1985). The domains are considerably smaller than the Topologically Associated Domains (1Mb; Dixon et al. (2012), Nora et al. (2012)) that have previously been reported in human and mouse on the basis of contact mapping. This accords well with assessments suggesting that extremely dense contact maps are required in order to resolve small domains (Ciabrelli and Cavalli, 2014) (Figure S2R–T).

Domains exhibit consistent histone marks, whose changes are associated with changes in long-range contact pattern

Loci within a domain show correlated chromatin states for eight different histone modifications (H3K36me3, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K79me2, and H4K20me1) based on data from the ENCODE project in GM12878 cells (ENCODE Consortium, 2011; ENCODE Consortium et al., 2012). By contrast, loci at comparable distance but residing in different domains showed much less correlation in chromatin state (Figure 2B, S2I-K). For instance, the correlation between the H3K36me3 signals for two 10Kb loci separated by 50Kb was 0.52 if the loci were in the same contact domain, but only 0.23 if they were in different contact domains. For H3K27me3, the corresponding correlations were 0.59 and 0.19, respectively.

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 (Figure 2C, S2O-Q, 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 sought to characterize the long-range contact patterns in our data. We partitioned loci into categories based on long-range contact patterns alone, using four independent approaches: manual annotation, and three objective clustering algorithms (HMM, K-means, Hierarchical). All gave similar results (Extended Experimental Procedures*,* Figure S3B). We then investigated the biological meaning of these categories.

When we analyzed the data at low matrix resolution (1Mb), we reproduced our earlier finding of two compartments (A and B). At high resolution (25Kb), however, we found strong 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 (Imakaev et al., 2012; Yaffe and Tanay, 2011). We found that the median length of an interval lying completely within a subcompartment is 300Kb. Although the five subcompartments are defined solely based on their Hi-C interaction patterns, they show distinctive properties with respect to both their genomic and epigenomic content.

Two of the five interaction patterns are strongly correlated with loci in compartment A (Figure S3E). 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 envelope and at nucleolus associated domains (NADs). (See Figure

2D,E, S3I.) 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). (Data sources are listed in Table S8.)

The other three interaction patterns (labeled B1, B2, and B3) are strongly correlated with loci in compartment B (Figure S3E), and show very different properties. Subcompartment B1 correlates positively with H3K27me3 and negatively with H3K36me3, suggestive of facultative heterochromatin (Figure 2D,E). Replication of this subcompartment peaks during the middle of S-phase. 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 tends to lack all of the above-noted marks, suggesting ordinary heterochromatin; it is enriched at the nuclear lamina (1.6-fold), but strongly depleted at NADs (76-fold). Subcompartments B2 and B3 do not replicate until the end of S-phase (See Figure 2D).

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 11Mb, 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 (Barski et al., 2007; Hahn et al., 2011; Vogel et al., 2006), these regions exhibit a highly distinctive chromatin pattern, with strong enrichment for both activating chromatin marks, such as H3K36me3, and heterochromatin-associated marks, such as H3K9me3 and H4K20me3.

In principle, the fact that domains lying in the same subcompartment exhibit similar chromatin marks might reflect either that (i) spatial proximity enhances the spread of histone modifications, or (ii) similarity of histone modifications helps bring about spatial proximity.

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 heatmap, 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 (Jin et al., 2013; Li et al., 2012; Sanyal et al., 2012) have sought to define peaks relative to the 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 9448 peaks in the *in situ* Hi-C map for GM12878 at 5kb map 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 less than 2Mb apart.

These findings were extremely reproducible across all of our high-resolution Hi-C maps. Examining the primary and replicate maps separately, we found 8054 peaks in the former and 7484 peaks in the latter, with 5403 in both lists; see figures 3A, 3B and S4A. The differences were almost always the result of our conservative peak-calling criteria. We also called peaks using our GM12878 dilution Hi-C experiment. Because the map is sparser and thus noisier, we called only 3073 peaks. Nonetheless, 65% of these peaks were also present in the list of peaks from our in situ Hi-C dataset, again reflecting good inter-replicate reproducibility.

As an independent confirmation that peak loci have greater physical proximity than neighboring locus pairs, we performed 3D-FISH (Beliveau et al., 2012) on 4 loops. In each case, we compared two peak loci, L1 and L2, with a control locus, L3, that lies an equal distance away from L2 but on the opposite side (Figure 3C, S4B). In all cases, the distance between L1 and L2 was consistently shorter than the distance between L2 and L3. (Peak 1: 32% of looping pairs co-localized, vs. 5% of control pairs; Peak 2: 29% vs. 9%; Peak 3: 25% vs. 9%; Peak 4: 18% vs. 4%. Co-localization was defined as a distance of <0.25μm. See 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 peaks are translated in any direction (See Experimental Procedures). APA showed strong consistency between our loop calls and all six previously published Hi-C datasets for lymphoblastoid cell lines (Kalhor et al., 2012; Lieberman-Aiden et al., 2009) (Figure 3D, Extended Experimental Procedures, Figure S4G, Table S5).

Finally, we demonstrated that the list of peaks was robust to particular protocol conditions by performing APA analysis 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 single experiment. Notably, these include five experiments (HIC043, HIC044, HIC045, HIC046, and HIC047; see 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 six human cell lines (IMR90, HMEC, NHEK, K562, HUVEC, HeLa, and KBM7). Because these maps contain fewer contacts, sensitivity is reduced, and fewer peaks are observed (ranging from 2634 to 8040). Notably, APA analysis showed 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 (the untranslated peak set showed an enrichment above translated control that ranged from 1.51-fold to 1.92-fold, $p<5\times10^{-6}$ for all; z-score) (Dixon et al., 2012; Jin et al., 2013; Lieberman-Aiden et al., 2009) (Figure S4H).

Overall, 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 S5A).

We also compared peaks across species. In CH12-LX mouse B-lymphoblasts, we identified 2927 high-confidence domains and 3331 peaks. We frequently observed a correspondence between orthologous regions in GM12878 and CH12-LX. Overall, 50% of peaks and 45% of domains called in mouse were also called in humans, suggesting substantial conservation of three-dimensional genome structure across the mammals (Figure 4B–E).

Loops anchored at a promoter are associated with enhancers and increased gene activation

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

First, our peaks frequently have a known promoter at one peak locus (as annotated by ENCODE's ChromHMM), and a known enhancer at the other (Figure 5A). For instance, 2854 of the 9448 peaks in our GM12878 map bring together known promoters and known enhancers (30%, vs. 7% expected by chance). These peaks include well-studied promoter-enhancer loops, such as at MYC (chr8:128.35–128.75Mb) and alpha-globin (chr16:0.15–0.22Mb). 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 gene expression. Although peaks are strongly correlated across cell types, there were also many cases in which a peak was present in one cell type but not another. When we examined RNA-Seq data produced by ENCODE (ENCODE Consortium, 2011; ENCODE Consortium et al., 2012), we found 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 instance, we observed 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 (>50-fold difference in RNA level), but of only 3 genes that were markedly upregulated in GM12878 (31-fold depletion). Conversely, we found 557 loops in GM12878 that were clearly absent in IMR90. The corresponding peak loci overlapped the promoters of 43 genes that were markedly upregulated in GM12878, but of only 1 gene that was markedly upregulated in IMR90: a 43-fold depletion. When we compared GM12878 to the five other human cell types for which ENCODE RNA-Seq data was available (all but KBM7), the results were very similar (Figure 5B).

One example of a cell-type specific loop is anchored at the promoter of the SELL gene, which encodes L-selectin, a lymphocyte-specific surface marker that is expressed in GM12878 but not IMR90 (Figure 5C). Gene activation is occasionally 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, suggesting that the ADAMTS1 promoter and the six distal sites may all be spatially co-located.

These observations are consistent with the classic model of promoter-enhancer function, in which looping between a promoter and enhancer activates a target gene (Ahmadiyeh et al., 2010; Amano et al., 2009; Tolhuis et al., 2002). The loop-associated activation of target genes has also been reported on the basis of other recent high-throughput studies examining chromatin contact patterns (Li et al., 2012). Many reports have also described the existence of gene looping, a phenomenon we also observe in our data (O'Sullivan et al., 2004; Tan-Wong et al., 2012). (For instance, in GM1278, we annotate 139 such loops.)

Peaks frequently demarcate the boundaries of domains

A large fraction of peaks (38%) coincide with the corners of a domain – that is, the peak loci are located at domain boundaries (Figure 6A). 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. To our knowledge, this is the first study to report the tendency of loops to delimit structural domains (in the sense of intervals of self-interacting chromatin). Because this configuration is so common, we will use the term "loop domain" to refer to 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. In these situations, the three loci may simultaneously co-locate at a single spatial position. However, many peaks do not exhibit transitivity, suggesting that the loci may not co-locate simultaneously. 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 far less often than expected under a random model (4-fold depleted; Extended Experimental Procedures).

The vast majority of peaks are associated with pairs of CTCF motifs in a convergent orientation

We next wondered whether peaks are associated with specific proteins. We therefore examined the results of 86 ChIP-Seq experiments performed by ENCODE in GM12878 (ENCODE Consortium, 2011; ENCODE Consortium et al., 2012). Strikingly, we found that the vast majority of peak 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 (Hou et al., 2008; Phillips and Corces, 2009; Splinter et al., 2006; Tolhuis et al., 2002). Because many of our loops demarcate domains, this observation is also consistent with studies suggesting that CTCF delimits structural and regulatory domains (Cuddapah et al., 2009; Dixon et al., 2012; Xie et al., 2007).

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 (6991 of 12,903, or 52%) with a specific CTCF-binding site “anchor”. (Interestingly, in mouse, 8% of these anchors lie inside SINEB2 repeats.)

The consensus DNA sequence for CTCF-binding sites is typically written as 5′-CCACNAGGTGGCAG-3′. Because the sequence is not palindromic, each CTCF site 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 4 orientations to occur equally often. But when we examined the 4322 peaks in GM12878 where the two corresponding peak loci each contained a single CTCF-binding motif, we found a stunning result: the vast majority (92%) of motif pairs are convergent (Figure 6D,E). 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). Notably, 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 3971–78 (51-fold enrichment of convergent vs. divergent); in IMR90, 1456–5 (291-fold); in HMEC, 968–11 (88-fold); in K562, 723 to 2 (362-fold); in HUVEC, 671–4 (168-fold); in HeLa, 301–3 (100-fold); in NHEK, 556–9 (62-fold); and in CH12, 625–8 (78-fold). This surprising 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 (8175 of 12,903, or 63.4%) with a single CTCF-binding site.

The specific orientation of CTCF sites at observed peaks provides strong 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 particular orientation could not occur by chance ($p < 10^{-1900}$, 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 (resp. RAD21) were isolated and analyzed. We found strong concordance with our results in all three cases (K562 CTCF: $p<10^{-13311}$; K562 RAD21: $p<10^{-8914}$; GM12878 RAD21: $p<10^{-11860}$; hypergeometric distribution) (Heidari et al., 2014; Li et al., 2012).

Diploid Hi-C maps reveals 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 assign contacts to specific chromosomal homologs. Using GM12878 SNP-phasing data (Gil et al., 2012; McKenna et al., 2010), we found that we could frequently assign reads to either the maternal or paternal homolog (Figure 7A). Using these assignments, we constructed a “diploid” Hi-C map of GM12878 comprising both maternal (238M contacts) and paternal (240M) maps. We studied these maps for differences between homologous chromosomes in contact frequencies, domain structure, and loop structure.

For autosomes, the maternal and paternal homologs exhibit very similar inter- and intra-chromosomal contact profiles (Pearson’s R>.998, $p<10^{-324}$; bivariate normal distribution). 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:73.5Mb and all distal loci (a stretch of over 60Mb) to the telomere of chromosome 6p (Figure 7B, 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 translocation, der(6)t(6,11)(pter;q) (Figure S7C–F). Notably, 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 1.72Mb (which 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 Fig. 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).

Most strikingly, differences were seen on the diploid intra-chromosomal maps of chromosome X. The paternal X chromosome, which is usually inactive in GM12878, is partitioned into two massive domains (0–115Mb and 115–155.3Mb). These “superdomains” are not seen in the active, maternal X (Fig. 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 markedly 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: 115Mb +/− 500Kb) 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 non-coding RNA. In males and on the active X, DXZ4 is heterochromatic, hyper-methylated and does not bind CTCF. On the inactive X, DXZ4 is euchromatic, hypo-methylated, 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 extremely large “superloops,” each spanning between 7 and 74Mb, present only on the inactive X chromosome in the diploid map (Fig. 7E). The superloops were also seen in all 4 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:56.8Mb-DXZ4 and DXZ4-130.9Mb) have been reported previously, and their presence on the inactive X alone has been confirmed using multiple methods (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 very large (up to 200kb). Four of these anchor regions contain whole lncRNA genes: loc550643; XIST; DXZ4; and FIRRE. Three (loc550643, and DXZ4, and FIRRE) contain CTCF-binding tandem repeats that only bind CTCF on the inactive homolog.

DISCUSSION

The *in situ* Hi-C protocol allowed us to probe genomic architecture with extremely high resolution; in the case of GM12878 lymphoblastoid cells, better than 1Kb.

We observe the presence of domains that were too small to be seen in our original Hi-C maps, which had resolution of 1Mb (Lieberman-Aiden et al., 2009). 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.

Strikingly, when the pattern of chromatin modifications associated with a domain changes, the domain’s long-range contact pattern also changes. The domains annotated here exhibit at least six distinct patterns of long-range contacts (subcompartments), which subdivide the two compartments that we had 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.

High-resolution *in situ* Hi-C data makes 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 lymphoblastoid cells, we observe 9448 loops.

The loops reported here have many interesting properties. First, most loops are short (<2Mb). Second, loops are strongly conserved across cell types and between human and mouse. Third, promoter-enhancer loops are common and are strongly associated with gene activation. This finding is, of course, consistent with the classical model for promoter-enhancer function. Fourth, loops often demarcate domains, and may establish them. We refer to such structures as loop domains. Fifth, loops tend not to overlap. Sixth, loops are closely associated with the presence of CTCF and the cohesin subunits RAD21 and SMC3; each of these proteins is found at over 86% of loop anchors. The association of CTCF with many promoter-enhancer loops is particularly unexpected.

The most surprising property of loops is that the pair of CTCF motifs present at the loop anchors occurs in a convergent orientation in >90% of cases (vs. 25% expected by chance). The importance of motif orientation between loci that are separated by, on average, 360Kb is unexpected 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 should shed light on this mechanism. Such experiments may also enable the engineering of loops, domains, and other chromatin structures.

It is interesting to compare our results to those seen in previous reports. The 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: 220kb) 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 (1Mb) (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, or a loop; and many are associated with both.) Our observations are consistent with recent suggestions that smaller domains would become apparent as the resolution of Hi-C maps increased (Ciabrelli and Cavalli, 2014).

Surprisingly, our annotation identifies many fewer loops than were reported in several recent high-throughput studies, despite the fact that we have far more data than any previous study. 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, several previous studies have defined peaks by comparing the contact frequency at a pixel to the genome-wide average (Jin et al., 2013; Li et al., 2012; Sanyal et al., 2012). This latter definition is problematic because many pixels within a domain can be annotated as peaks despite showing no local increase in contact frequency. Previous papers using the latter definition imply the existence of more than 100,000 or even more than 1 million peaks (Extended Experimental Procedures).

We also 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 (Chadwick, 2008). We also detect a network of extremely long-range (7 – 74Mb) “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. We hypothesize that Xi-specific CTCF-binding participates in the formation of these massive chromatin structures.

In our original report on Hi-C, we observed that Hi-C maps can be used to study physical models of genome folding. For example, we noted that our megabase-scale maps were consistent with a fractal globule model. 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 used were cultured following the manufacturer’s recommendations. Two to five million cells were crosslinked with 1% formaldehyde for 10 minutes at room temperature. (This step is optional; in several experiments, crosslinking was omitted. The resulting maps are noisier.) Nuclei were permeabilized. DNA was digested with 100 units of MboI (or DpnII), the ends of restriction fragments were labeled using biotinylated nucleotides, and then ligated in a small volume. After reversal of crosslinks, ligated DNA was purified and sheared to a length of roughly 400 base pairs, at which point ligation junctions were pulled down with streptavidin beads and prepped for high-throughput Illumina sequencing. Dilution Hi-C was performed as in (Lieberman-Aiden et al., 2009). See Extended Experimental Procedures.

3D-FISH

FISH probes were designed using the OligoPaints database. DNA-FISH was performed as described in (Beliveau et al., 2012), with minor modifications. See Extended Experimental Procedures.

Hi-C Data Pipeline

All sequence data was produced using Illumina paired-end sequencing. We processed sequence 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 2.5Mb, 1Mb, 500Kb, 250Kb, 100Kb, 50Kb, 25Kb, 10Kb, and 5Kb, as well as fragment-delimited resolutions of 500f, 200f, 100f, 50f, 20f, 5f, 2f, and 1f. For our largest data sets, the file also contains a 1Kb contact matrix. Normalized contact matrices are produced at all resolutions using (Knight and Ruiz, 2012). See Extended Experimental Procedures.

Annotation of Domains

To annotate domains, we apply a novel “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 Figure S2A–F. This transformation can be thought of as 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 observed value at $i,i-d$ and $i,i+d$. $A_{i,i+d}$ will be strongly positive if and only 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. See Extended Experimental Procedures.

Assigning loci to subcompartments

To cluster loci based on long-range contact patterns, we constructed a 100Kb resolution contact matrix comprising a subset of the interchromosomal contact data. Loci on odd chromosomes appeared on the rows, and loci from the even chromosomes appeared on the columns. (Chromosome X was excluded.) We cluster this matrix using the Python package *scikit*. To generate our annotation of subcompartment B4, the 100kb interchromosomal matrix for chromosome 19 was constructed and clustered separately, using the same procedure. See Extended Experimental Procedures.

Annotation of Peaks

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 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 $i \le i' \le j' \le j$; 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, and that this enrichment be statistically significant after correcting for multiple hypothesis testing (FDR<10%). The same criteria are applied to three other neighborhoods. To be labeled an “enriched pixel,” a pixel must therefore be significantly enriched relative to four neighborhoods: (i) pixels to its lower-left; (ii) pixels to its left and right; (iii) pixels above and below; and (iv) a donut surrounding the pixel of interest (Figure 4A).

Using this approach, we identified numerous enriched pixels across the genome. The 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 largest number of contacts. Because over 10 billion $(10Kb)^2$ pixels must be examined, this calculation requires weeks of CPU time to execute. To accelerate it, we created a highly parallelized implementation using general-purpose graphical processing units, resulting in a 200-fold speedup relative to our initial, CPU-based approach. See Extended Experimental Procedures.

Aggregate Peak Analysis

We perform APA on 10Kb 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 210Kb × 210Kb 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 10Kb to the right; the entry two positions above center corresponds to an upward shift of 20Kb, and so on. Focal enrichment across the peak set in aggregate manifests as larger values at the center of the APA plot. APA analyses only include peaks whose loci are at least 300Kb apart. The “translated control” used in the above text is the mean of the values seen in the lower-left hand corner of the APA plot. See Extended Experimental Procedures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by National Science Foundation Graduate Research Fellowship under Grants DGE0946799 and DGE1144152 (M.H.H.); an NIH New Innovator Award (1DP2OD008540-01), the NHGRI Center for Excellence for Genomic Sciences (P50HG006193), a Cancer Prevention Research Institute of Texas Scholar Award (R1304), a McNair Medical Institute Scholar Award, and the President’s Early Career Award in Science and Engineering, to (E.L.A.). We thank Leslie Gaffney for assistance with the design and preparation of figures; Bang Wong for input on figure design; Justin Demmerle, Wendy Salmon, Michael A. Mancini, Fabio Stossi, Radhika Dandekar, Sanjay Krishna and the Integrated Microscopy Core at the Baylor College of Medicine for assistance with 3D-FISH; Asha S. Multani for assistance with karyotyping; Matt Nicklay for assistance with 3D-FISH image processing; Aharon Lieberman, Aviva Presser Aiden, Brendan Lee, Arthur L. Beaudet, James R. Lupski, José N. Onuchic, Mitchell Guttman, Aviv Regev, Andreas Gnirke, Louise Williams, Chad Nusbaum, John Bohannon, Olga Dudchenko, and all members of the Aiden lab for discussions; and Robbyn Issner and the ENCODE chromatin group at the Broad Institute and Massachusetts General Hospital for providing some cell lines. The Center for Genome Architecture is grateful to Janice, Robert, and Cary McNair for support. Hi-C sequence data reported in this paper has been deposited at the GEO database (www.ncbi.nlm.nih.gov/geo/), accession no. DDD. A provisional patent on the in situ Hi-C method has been filed.

References and Notes

Ahmadiyeh N, Pomerantz MM, Grisanzio C, Herman P, Jia L, Almendro V, He HH, Brown M, Liu XS, Davis M, et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proceedings of the National Academy of Sciences. 2010; 107:9742–9746.

Amano T, Sagai T, Tanabe H, Mizushina Y, Nakazawa H, Shiroishi T. Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. Developmental cell. 2009; 16:47–57. [PubMed: 19097946]

Ay F, Bailey TL, Noble WS. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome research. 2014

Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell. 1981; 27:299–308. [PubMed: 6277502]

Barski A, Cuddapah S, Cui K, Roh TY, Schones D, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. Cell. 2007; 129:823–837. [PubMed: 17512414]

Beliveau B, Joyce E, Apostolopoulos N, Yilmaz F, Fonseka C, McCole R, Chang Y, Li J, Senaratne T, Williams B, et al. Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:21301–21306. [PubMed: 23236188]

Bickmore W. The spatial organization of the human genome. Annual review of genomics and human genetics. 2013; 14:67–84.

Blackwood E, Kadonaga J. Going the distance: a current view of enhancer action. Science (New York, NY). 1998; 281:60–63.

Chadwick B. 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 research. 2008; 18:1259–1269. [PubMed: 18456864]

Ciabrelli F, Cavalli G. Chromatin driven behavior of topologically associating domains. Journal of molecular biology. 2014

Consortium EP. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biology. 2011:9.

Bernstein B, Birney E, Dunham I, Green E, Gunter C, Snyder M. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

Cook PR, Brazell IA. Supercoils in human DNA. Journal of cell science. 1975; 19:261–279. [PubMed: 1202042]

Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. Normalization of a chromosomal contact map. BMC genomics. 2012; 13:436. [PubMed: 22935139]

Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature Reviews: Genetics. 2001; 2:292–301.

Cuddapah S, Jothi R, Schones D, Roh TY, Cui K, Zhao K. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome research. 2009; 19:24–32. [PubMed: 19056695]

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

Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002; 295:1306–1311. [PubMed: 11847345]

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

Dostie J, Richmond T, Arnaout R, Selzer R, Lee W, Honan T, Rubio E, Krumm A, Lamb J, Nusbaum C, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome research. 2006; 16:1299–1309. [PubMed: 16954542]

Dunn K, Zhao H, Davie J. The insulator binding protein CTCF associates with the nuclear matrix. Experimental cell research. 2003; 288:218–223. [PubMed: 12878173]

Dunn T, Hahn S, Ogden S, Schleif R. An operator at -280 base pairs that is required for repression of araBAD operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. Proceedings of the National Academy of Sciences of the United States of America. 1984; 81:5017–5020. [PubMed: 6089170]

Eismann E, von Wilcken-Bergmann B, Müller-Hill B. Specific destruction of the second lac operator decreases repression of the lac operon in Escherichia coli fivefold. Journal of molecular biology. 1987; 195:949–952. [PubMed: 3116268]

Filippova G, Fagerlie S, Klenova E, Myers C, Dehner Y, Goodwin G, Neiman P, Collins S, Lobanenkov V. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. Molecular and cellular biology. 1996; 16:2802–2813. [PubMed: 8649389]

Finlan LE, Sproul D, Thomson I, Boyle S, Kerr E, Perry P, Ylstra B, Chubb JR, Bickmore WA. Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS genetics. 2008:4.

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

Gaszner M, Felsenfeld G. Insulators: exploiting transcriptional and epigenetic mechanisms. Nature Reviews: Genetics. 2006; 7:703–713.

Gavrilov AA, Gushchanskaya ES, Strelkova O, Zhironkina O, Kireev II, Iarovaia OV, Razin SV. Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. Nucleic acids research. 2013; 41:3563–3575. [PubMed: 23396278]

Gerasimova T, Byrd K, Corces V. A chromatin insulator determines the nuclear localization of DNA. Molecular cell. 2000; 6:1025–1035. [PubMed: 11106742]

Gil AM, David MA, Richard MD, Gonçalo RA, David RB, Aravinda C, Andrew GC, Peter D, Evan EE, Paul F, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012:491.

Griffith J, Hochschild A, Ptashne M. DNA loops induced by cooperative binding of lambda repressor. Nature. 1986; 322:750–752. [PubMed: 3748156]

Hahn MA, Wu X, Li AX, Hahn T, Pfeifer GP. Relationship between Gene Body DNA Methylation and Intragenic H3K9me3 and H3K36me3 Chromatin Marks. PLoS One. 2011

Heidari N, Phanstiel DH, He C, Grubert F, Jahanbanian F, Kasowski M, Zhang MQ, Snyder MP. Genome-wide map of regulatory interactions in the human genome. Genome research. 2014

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

Hou C, Zhao H, Tanimoto K, Dean A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:20398–20403. [PubMed: 19074263]

Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Lui JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012; 28:3131–3133. [PubMed: 23023982]

Imakaev M, Fudenberg G, McCord R, Naumova N, Goloborodko A, Lajoie B, Dekker J, Mirny L. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature methods. 2012; 9:999–1003. [PubMed: 22941365]

Jin F, Li Y, Dixon J, Selvaraj S, Ye Z, Lee A, Yen CA, Schmitt A, Espinoza C, Ren B. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013; 503:290–294. [PubMed: 24141950]

Julienne H, Zoufir A, Audit B, Arneodo A. Human Genome Replication Proceeds through Four Chromatin States. PLoS computational biology. 2013

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

Kim T, Abdullaev Z, Smith A, Ching K, Loukinov D, Green R, Zhang M, Lobanenkov V, Ren B. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell. 2007; 128:1231–1245. [PubMed: 17382889]

Klenova E, Nicolas R, Paterson H, Carne A, Heath C, Goodwin G, Neiman P, Lobanenkov V. CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. Molecular and cellular biology. 1993; 13:7612–7624. [PubMed: 8246978]

Knight P, Ruiz D. A fast algorithm for matrix balancing. IMA Journal of Numerical Analysis. 2012

Köhne A, Baniahmad A, Renkawitz R. NeP1: A Ubiquitous Transcription Factor Synergizes with v-ERBA in Transcriptional Silencing. Journal of molecular biology. 1993

Krämer H, Niemöller M, Amouyal M, Revet B, von Wilcken-Bergmann B, Müller-Hill B. lac repressor forms loops with linear DNA carrying two suitably spaced lac operators. EMBO Journal. 1987; 6:1481–1491. [PubMed: 3301328]

Kurukuti S, Tiwari V, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, Lobanenkov V, Reik W, Ohlsson R. CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103:10684–10689. [PubMed: 16815976]

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

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

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

Lin YC, Benner C, Mansson R, Heinz S, Miyazaki K, Miyazaki M, Chandra V, Bossen C, Glass CK, Murre C. Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. Nature immunology. 2012; 13:1196–1204. [PubMed: 23064439]

Lobanenkov V, Nicolas R, Adler V, Paterson H, Klenova E, Polotskaja A, Goodwin G. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5′-flanking sequence of the chicken c-myc gene. Oncogene. 1990; 5:1743–1753. [PubMed: 2284094]

McKenna A, Hanna M, Banks E, Sivachenko A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome …. 2010

Mukherjee S, Erickson H, Bastia D. Enhancer-origin interaction in plasmid R6K involves a DNA loop mediated by initiator protein. Cell. 1988; 52:375–383. [PubMed: 3345564]

Murrell A, Heeson S, Reik W. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. Nature genetics. 2004; 36:889–893. [PubMed: 15273689]

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013; 502:59–64. [PubMed: 24067610]

Nora E, Lajoie B, Schulz E, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum N, Meisig J, Sedat J, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012; 485:381–385. [PubMed: 22495304]

O'Sullivan JM, Tan-Wong SM, Morillon A, Lee B, Coles J, Mellor J, Proudfoot NJ. Gene loops juxtapose promoters and terminators in yeast. Nature genetics. 2004; 36:1014–1018. [PubMed: 15314641]

Oehler S, Eismann E, Krämer H, Müller-Hill B. The three operators of the lac operon cooperate in repression. EMBO Journal. 1990; 9:973–979. [PubMed: 2182324]

Pennacchio L, Bickmore W, Dean… A. Enhancers: five essential questions. Nature Reviews …. 2013

Phillips J, Corces V. CTCF: master weaver of the genome. Cell. 2009; 137:1194–1211. [PubMed: 19563753]

Ptashne M. Gene regulation by proteins acting nearby and at a distance. Nature. 1986; 322:697–701. [PubMed: 3018583]

Sanyal A, Lajoie B, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012; 489:109–113. [PubMed: 22955621]

Schleif R. DNA looping. Annual review of biochemistry. 1992; 61:199–223.

Sexton T, Schober H, Fraser P, Gasser S. Gene regulation through nuclear organization. Nature structural & molecular biology. 2007; 14:1049–1055.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012; 148:458–472. [PubMed: 22265598]

Spilianakis C, Flavell R. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. Nature immunology. 2004; 5:1017–1027. [PubMed: 15378057]

Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, Galjart N, de Laat W. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes & development. 2006; 20:2349–2354. [PubMed: 16951251]

Tan-Wong SM, Zaugg JB, Camblong J, Xu Z, Zhang DW, Mischo HE, Ansari AZ, Luscombe NM, Steinmetz LM, Proudfoot NJ. Gene loops enhance transcriptional directionality. Science (New York, NY). 2012; 338:671–675.

Tolhuis B, Palstra R, Splinter E, Grosveld F, de Laat W. Looping and interaction between hypersensitive sites in the active beta-globin locus. Molecular cell. 2002; 10:1453–1465. [PubMed: 12504019]

Vogel MJ, Guelen L, de Wit E, Hupkes DP. Human heterochromatin proteins form large domains containing KRAB-ZNF genes. Genome …. 2006

Vogelstein B, Pardoll DM, Coffey DS. Supercoiled loops and eucaryotic DNA replicaton. Cell. 1980; 22:79–85. [PubMed: 7428042]

Vostrov AA. The Zinc Finger Protein CTCF Binds to the APBbeta Domain of the Amyloid beta - Protein Precursor Promoter. EVIDENCE FOR A ROLE IN TRANSCRIPTIONAL ACTIVATION. Journal of Biological Chemistry. 1997:272.

Xie X, Mikkelsen T, Gnirke A, Lindblad-Toh K, Kellis M, Lander E. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104:7145–7150. [PubMed: 17442748]

Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature genetics. 2011; 43:1059–1065. [PubMed: 22001755]

Yusufzai T, Felsenfeld G. The 5′-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101:8620–8624. [PubMed: 15169959]

Yusufzai T, Tagami H, Nakatani Y, Felsenfeld G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. Molecular cell. 2004; 13:291–298. [PubMed: 14759373]

Zehnbauer BA, Vogelstein B. Supercoiled loops and the organization of replication and transcription in eukaryotes. BioEssays. 1985

Highlights

- Domains segregate into six subcompartments with distinct patterns of histone marks
- Loop anchors occur at domain boundaries and bind CTCF in a convergent orientation
- Loops correlate with gene activation, and are conserved across cell types and species
- The inactive X-chromosome contains large loops anchored at CTCF-binding repeats

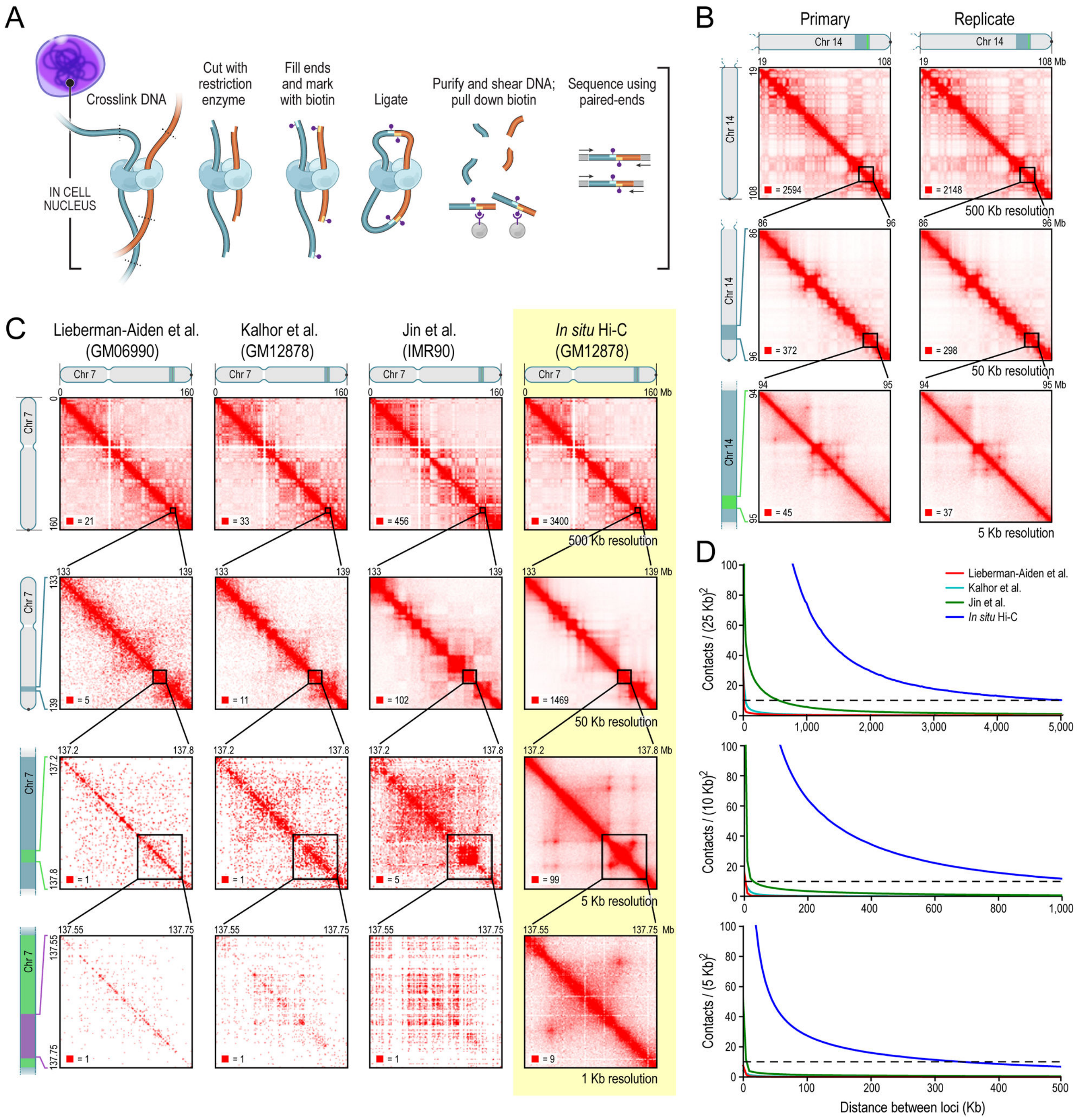
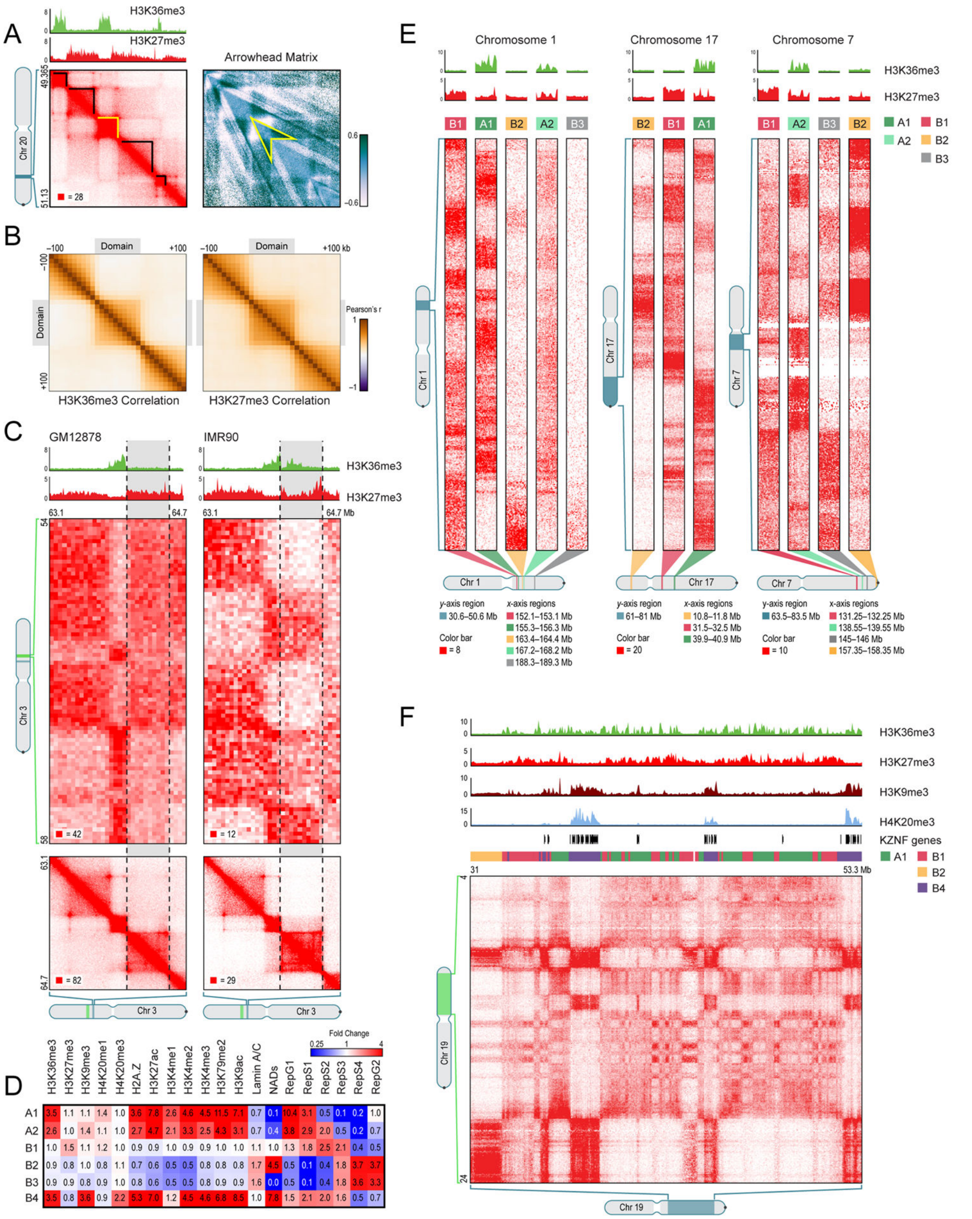


Fig. 1. We used in situ Hi-C to map over 15 billion chromatin contacts across nine cell types in human and mouse, achieving 1 kilobase resolution in human lymphoblastoid cells

(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 500Kb resolution (top); 86–96Mb/50Kb resolution (middle); 94–95Mb/5Kb resolution (bottom). Left: GM12878, primary experiment; Right: 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., Kalhor et al., and Jin et al. **(D)** Mean contacts per pixel vs distance, at various resolutions, compared to published Hi-C experiments (dashed line = 10). See also Data S1, Table S1 and Table S2.



	H3K36me3	H3K27me3	H3K9me3	H4K20me1	H4K20me3	H2A.Z	H3K27ac	H3K4me1	H3K4me2	H3K4me3	H3K79me2	H3K9ac	Lamin A/C	NADs	RepG1	RepS1	RepS2	RepS3	RepS4	RepG2
A1	3.5	1.1	1.1	1.4	1.0	3.6	7.8	2.6	4.6	4.5	11.5	7.1	0.7	0.1	10.4	3.1	0.5	0.1	0.2	1.0
A2	2.6	1.0	1.4	1.1	1.0	2.7	4.7	2.1	3.3	2.5	4.3	3.1	0.7	0.4	3.8	2.9	2.0	0.5	0.2	0.7
B1	1.0	1.5	1.1	1.2	1.0	0.9	0.9	1.0	0.9	0.9	1.0	1.0	1.1	1.0	1.3	1.8	2.5	2.1	0.4	0.5
B2	0.9	0.8	1.0	0.8	1.1	0.7	0.6	0.5	0.5	0.8	0.8	0.8	1.7	4.5	0.5	0.1	0.4	1.8	3.7	3.7
B3	0.9	0.9	0.8	0.9	1.0	0.8	0.6	0.5	0.5	0.8	0.9	0.9	1.6	0.0	0.5	0.1	0.4	1.8	3.6	3.3
B4	3.5	0.8	3.6	0.9	2.2	5.3	7.0	1.2	4.5	4.6	6.8	8.5	1.0	7.8	1.5	2.1	2.0	1.6	0.5	0.7

Fig. 2. The genome is partitioned into 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 towards the domain's upper-left corner (example in yellow). The arrowhead size corresponds to the domain size. Using dynamic programming, this transformation allows us to efficiently compute a "corner score" for each pixel in a Hi-C matrix, indicating the likelihood that the pixel lies at the upper-right corner of a domain. See Experimental Procedures. **(B)** Pearson correlation matrices of the histone mark signal between pairs of loci inside, and within 100Kb of, a domain. *Left:* H3K36me3; *Right:* H3K27me3. **(C)** Conserved 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 which differ from one another and are preserved in both cell types. In IMR90, the central 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, 10Kb resolution; long-range interaction matrices, 50Kb resolution. **(D)** Each of the six long-range contact patterns we observe exhibits a distinct epigenetic profile. All epigenetic data is from ENCODE experiments in GM12878 except nuclear lamin (derived from skin fibroblast cells) and NAD (HeLa). See Table S8. 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 compartments. **(F)** A large contiguous region on chromosome 19 contains intervals in subcompartments A1, B1, B2, and B4. See also Data S2 and Data S3.

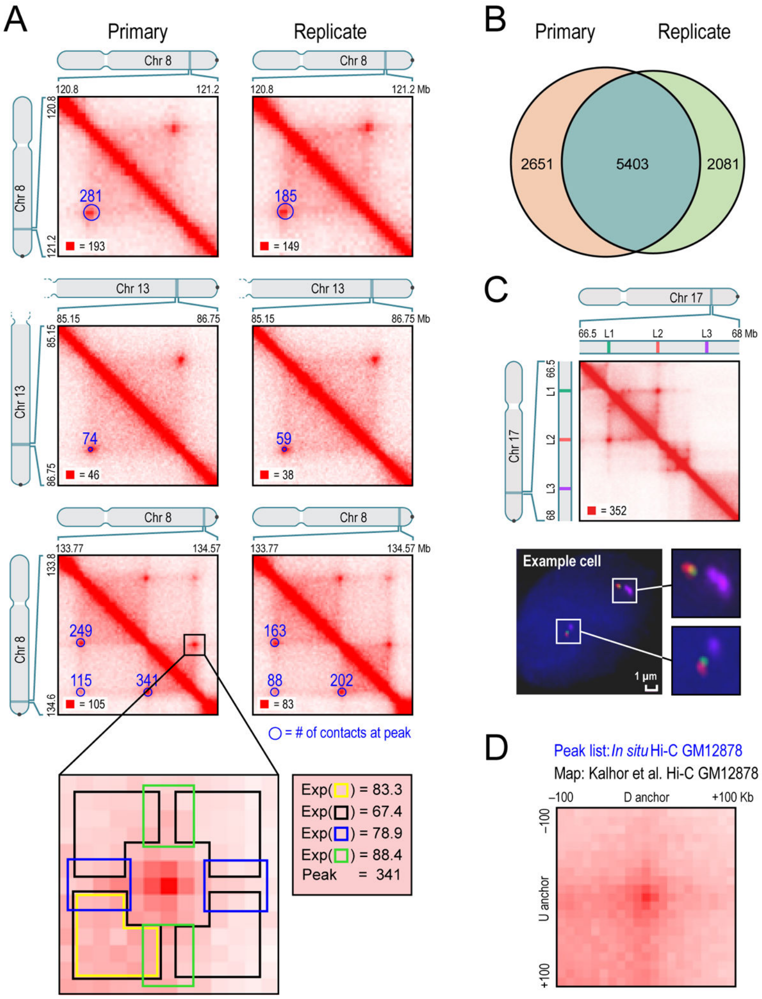


Fig. 3. We identify thousands of chromatin loops genome-wide using a local background model

(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 are marked with blue circles (radius=20Kb) 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 these figures are 10Kb resolution unless noted. **(B)** Overlap between replicates. **(C)** *(Top)* Location of 3D-FISH probes *(Bottom)* Example cell. **(D)** APA plot shows the aggregate signal from the 9948 GM12878 loops we report by summing submatrices surrounding each peak in a low-resolution GM12878 Hi-C map due to Kalhor et al. See also Figure S4, Table S3, Table S4, and Table S5.

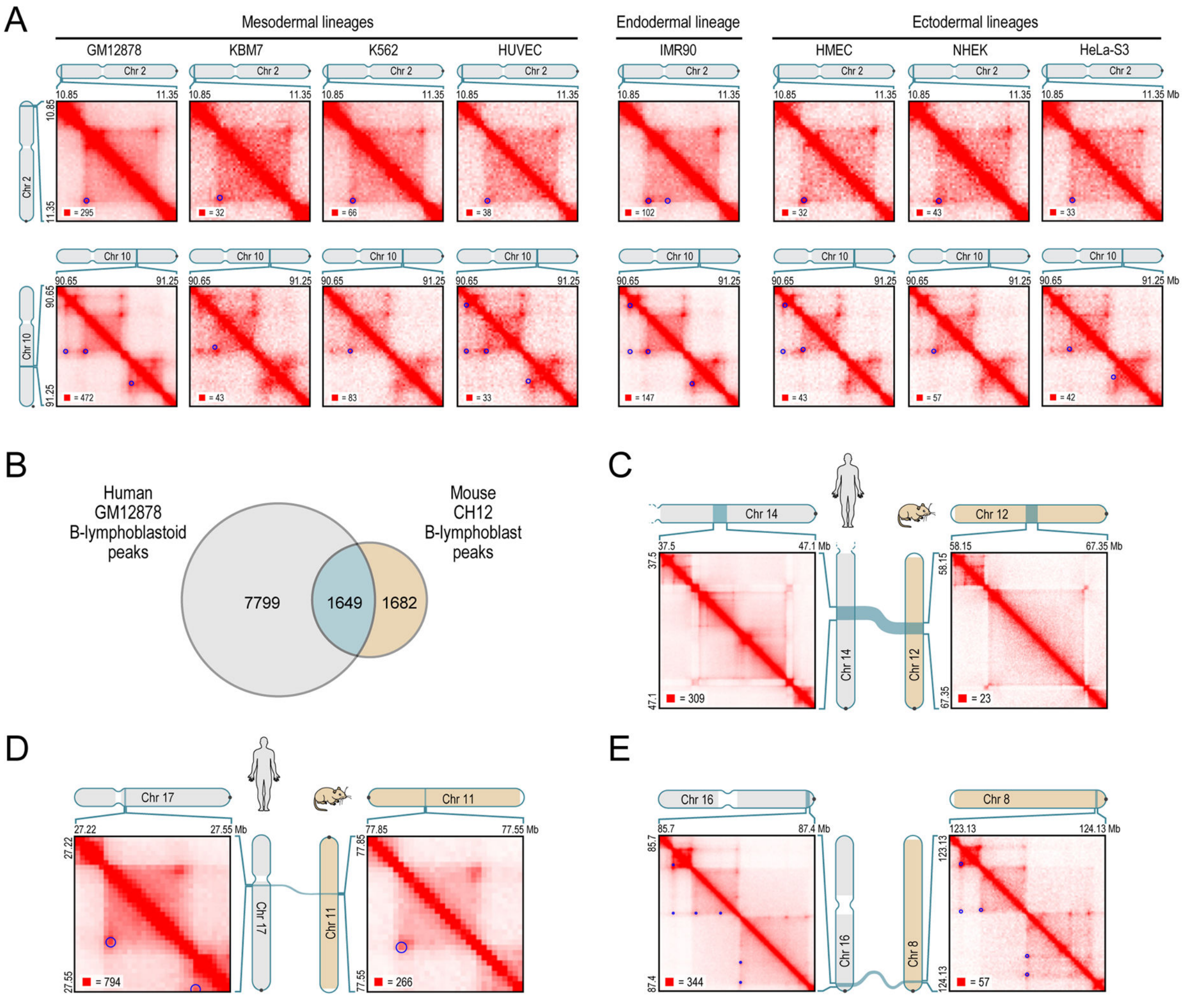


Fig. 4. Loops are often preserved across cell types and from human to mouse

(A) Examples of peak and domain preservation across cell types. Annotated peaks are circled in blue. All annotations are completely independent. **(B)** Of the 3331 loops we annotate in mouse CH12-LX, 1649 (50%) are orthologous to loops in human GM12878. **(C–E)** Conservation of three-dimensional structure in synteny blocks. See also Figure S5.

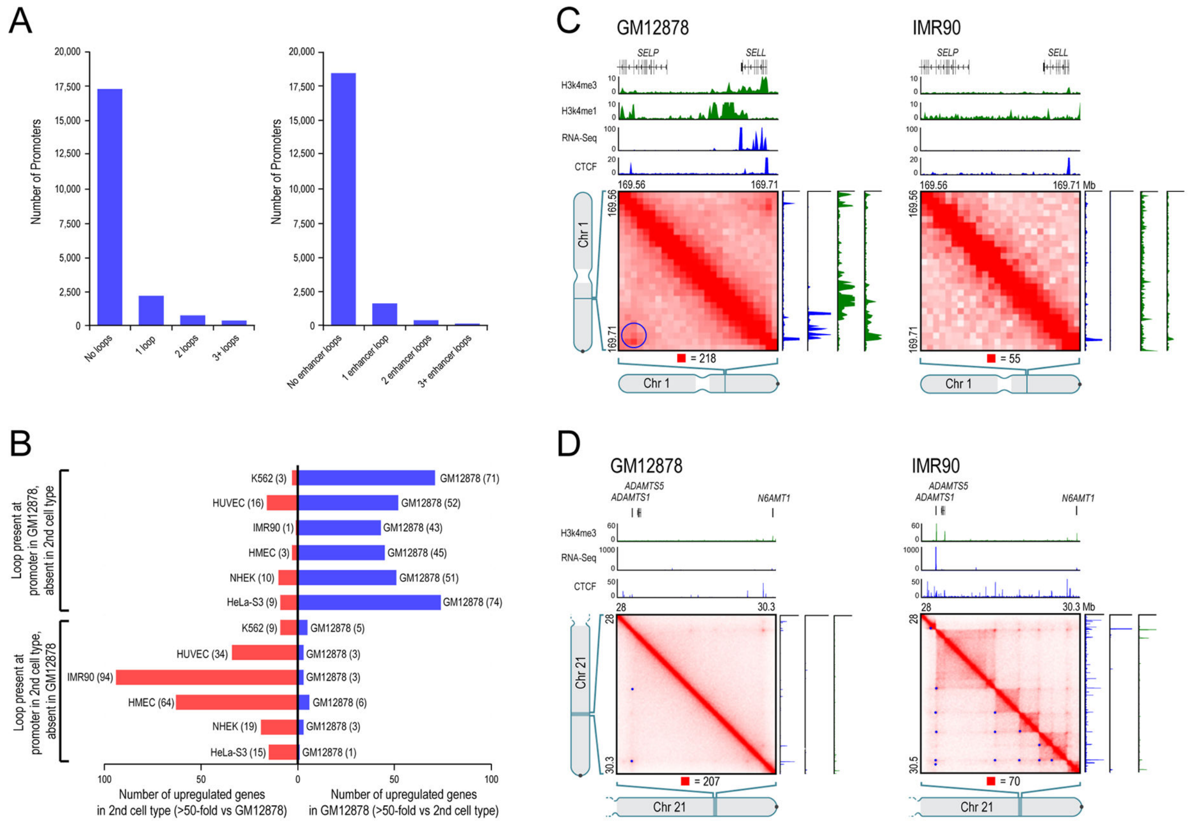


Fig. 5. Loops between promoters and enhancers are strongly associated with gene activation

(A) Histogram showing loop count at promoters (left); restricted to loops where the distal peak locus contains an enhancer (right). **(B)** Genes whose promoters participate in a loop in GM12878 but not in a second cell type are frequently upregulated in GM12878, and vice-versa. **(C)** *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. **(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 evident transitive looping. ADAMTS1 is on. See also Figure S5 and Table S6.

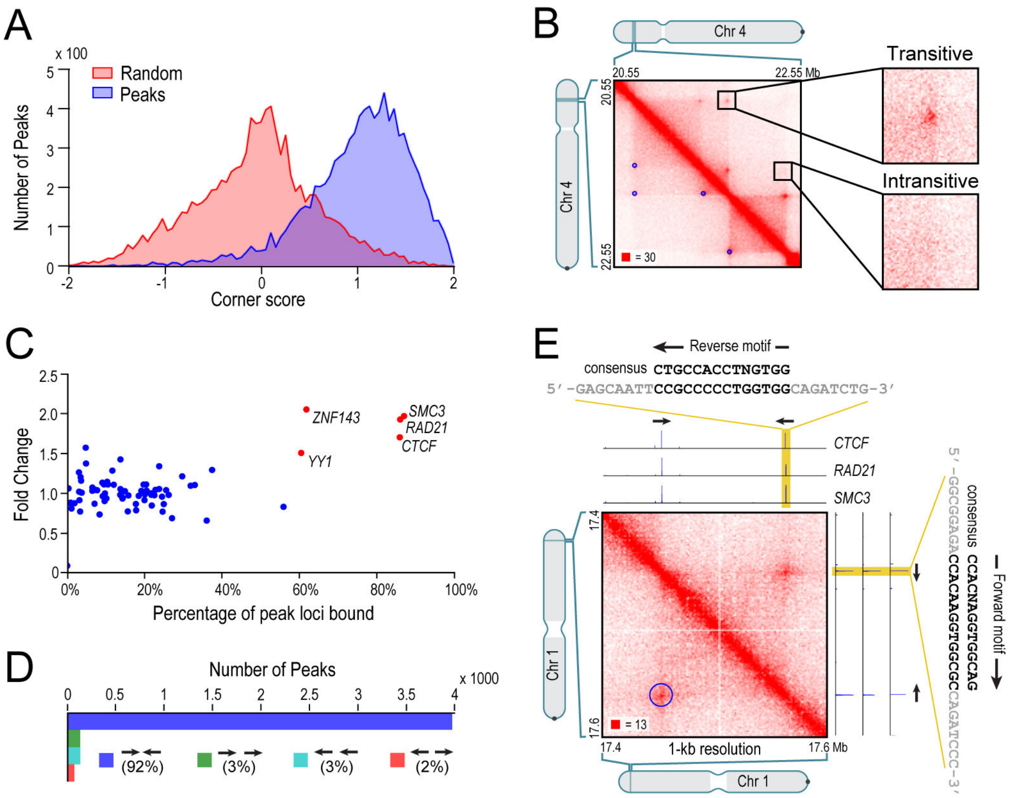


Fig. 6. Many loops demarcate domains; the vast majority of loops are anchored at a pair of convergent CTCF/RAD21/SMC3 binding sites

(A) Histograms of corner score for peak pixels vs. random pixels with an identical distance distribution. **(B)** Contact matrix for chr4:20.55Mb-22.55Mb in GM12878, showing examples of transitive and intransitive looping behavior. **(C)** % of peak loci bound vs. 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. **(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. See also Figure S6.

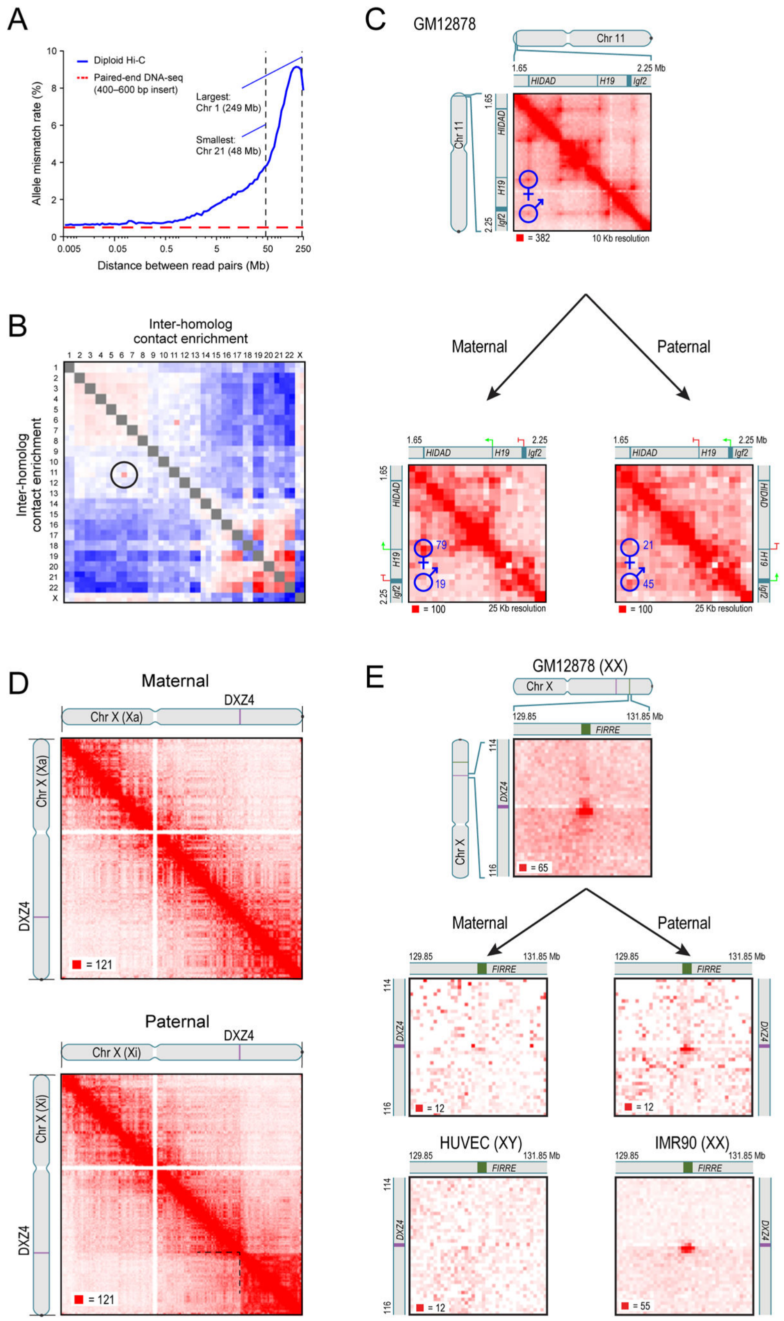


Fig. 7. Diploid Hi-C maps reveal superdomains and superloops anchored at CTCF-binding repeats on the inactive X chromosome

(A) The frequency of mismatch (maternal-paternal) in SNP allele assignment vs 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 6p and 11p (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. **(E)** The “superloop” between FIRRE and DXZ4 is present in the GM12878 haploid 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). See also Figure S7 and Table S7.