



Published in final edited form as:

Nature. ; 485(7398): 376–380. doi:10.1038/nature11082.

Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions

Jesse R. Dixon^{1,3,4}, Siddarth Selvaraj^{1,5}, Feng Yue¹, Audrey Kim¹, Yan Li¹, Yin Shen¹, Ming Hu⁶, Jun S. Liu⁶, and Bing Ren^{1,2,*}

¹Ludwig Institute for Cancer Research

²University of California, San Diego School of Medicine, Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, 9500 Gilman Drive, La Jolla, CA 92093

³Medical Scientist Training Program, University of California, San Diego, La Jolla CA 92093

⁴Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla CA 92093

⁵Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla CA 92093

⁶Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138

Abstract

The spatial organization of the genome is intimately linked to its biological function, yet our understanding of higher order genomic structure is coarse, fragmented and incomplete. In the nucleus of eukaryotic cells, interphase chromosomes occupy distinct chromosome territories (CT), and numerous models have been proposed for how chromosomes fold within CTs¹. These models, however, provide only few mechanistic details about the relationship between higher order chromatin structure and genome function. Recent advances in genomic technologies have led to rapid revolutions in the study of 3D genome organization. In particular, Hi-C has been introduced as a method for identifying higher order chromatin interactions genome wide². In the present study, we investigated the 3D organization of the human and mouse genomes in embryonic stem cells and terminally differentiated cell types at unprecedented resolution. We identify large, megabase-sized local chromatin interaction domains, which we term “topological domains”, as a pervasive structural feature of the genome organization. These domains correlate with regions of the genome that constrain the spread of heterochromatin. The domains are stable across different cell types and highly conserved across species, suggesting that topological domains are an

Users may view, print, copy, download and text and data- mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: http://www.nature.com/authors/editorial_policies/license.html#terms

*To whom correspondence should be addressed: biren@ucsd.edu.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature

Author Information: All Hi-C data described in this study have been deposited to GEO under the accession number GSE35156. We have developed a web based Java tool to visualize the high resolution Hi-C data at a genomic region of interest that is available at <http://chromosome.sdsc.edu/mouse/hi-c/database.html>. Reprints and permissions information is available at www.nature.com/reprints.

The authors declare no competing financial interests.

Author Contributions: JD and BR designed the studies. JD, AK, YL, YS conducted the Hi-C experiments; JD, SS and FY carried out the data analysis; JL and MH provided insight for analysis; FY built the supporting website; JD and BR prepared the manuscript.

inherent property of mammalian genomes. Lastly, we find that the boundaries of topological domains are enriched for the insulator binding protein CTCF, housekeeping genes, tRNAs, and SINE retrotransposons, suggesting that these factors may play a role in establishing the topological domain structure of the genome.

To study chromatin structure in mammalian cells, we performed the Hi-C experiment² in mouse embryonic stem cells (mESCs), human embryonic stem cells (hESCs), and human IMR90 fibroblasts. Together with Hi-C data for the mouse cortex generated in a separate study³, we analyzed over 1.7 billion read pairs of Hi-C data corresponding to pluripotent and differentiated cells (Supplemental Table 1). We normalized the Hi-C interactions for biases in the data (Supplemental Figure 1 and 2)⁴. To validate the quality of our Hi-C data, we compared the data with previous 5C, 3C, and FISH results⁵⁻⁷. Our IMR90 Hi-C data shows a high degree of similarity when compared to a previously generated 5C dataset from lung fibroblasts (Supplementary Figure 4). In addition, our mESC Hi-C data correctly recovered a previously described cell-type specific interaction at the *Phc1* gene⁶ (Supplementary Figure 5). Furthermore, the Hi-C interaction frequencies in mESCs are well-correlated with the mean spatial distance separating six loci as measured by 2D-FISH⁷ (Supplemental Figure 6), demonstrating that the normalized Hi-C data can accurately reproduce the expected nuclear distance using an independent method. These results demonstrate that our Hi-C data is of high quality and accurately captures the higher order chromatin structures in mammalian cells.

We next visualized 2D-interaction matrices using a variety of bin sizes to identify interaction patterns revealed as a result of our high sequencing depth (Supplemental Figure 7). We noticed that at bin sizes less than 100kb, highly self-interacting regions begin to emerge (Figure 1a, Supplemental Figure 7, seen as “triangles” on the heatmap). These regions, which we term “topological domains,” are bounded by narrow segments where the chromatin interactions appear to end abruptly. We hypothesized that these abrupt transitions may represent boundary regions in the genome that separate topological domains.

To systematically identify all such topological domains in the genome, we devised a simple statistic termed the “directionality index” (DI) to quantify the degree of upstream or downstream interaction bias for a genomic region, which varies considerably at the periphery of the topological domains (Figure 1b, see supplemental methods for details). The DI was reproducible (Supplemental Table 2) and pervasive, with 52 % of the genome having a DI that was not expected by random chance (Figure 1c, FDR = 1%). We then used a Hidden Markov model (HMM) based on the DI to identify biased “states” and therefore infer the locations of topological domains in the genome (Figure 1a, see supplemental methods for details). The domains defined by HMM were reproducible between replicates (Supplemental Figure 8). Therefore, we combined the data from the HindIII replicates and identified 2,200 topological domains in mESCs with a median size of 880kb that occupy ~91% of the genome (Supplemental Figure 9). As expected, the frequency of intra-domain interactions is higher than inter-domain interactions (Figure 1d,e). Similarly, FISH probes⁷ in the same topological domain (Figure 1f) are closer in nuclear space than probes in different topological domains (Figure 1g), despite similar genomic distances between probe

pairs (Figure 1h,i). These findings are best explained by a model of the organization of genomic DNA into spatial modules linked by short chromatin segments. We define the genomic regions between topological domains as either “topological boundary regions” or “unorganized chromatin”, depending on their sizes (Supplemental Figure 9).

We next investigated the relationship between the topological domains and the transcriptional control process. The *HoxA* locus is separated into two compartments by an experimentally validated insulator^{5,8,9}, which we observed corresponds to a topological domain boundary in both mouse (Figure 1a) and human (Figure 2a). Therefore, we hypothesized that the boundaries of the topological domains might correspond to insulator or barrier elements.

Many known insulator or barrier elements are bound by the zinc-finger containing protein CTCF^{10–12}. We see a strong enrichment of CTCF at the topological boundary regions (Figure 2b, Supplemental Figure 10), indicating that topological boundary regions share this feature of classical insulators. A classical boundary element is also known to stop the spread of heterochromatin. Therefore, we examined the distribution of the heterochromatin mark H3K9me3 in humans and mice in relation to the topological domains^{13,14}. Indeed, we observe a clear segregation of H3K9me3 at the boundary regions that occurs predominately in differentiated cells (Figure 2d,e, Supplemental Figure 11). Since the boundaries we analyzed in Figure 2d are present in both pluripotent cells and their differentiated progeny, the topological domains and boundaries appear to “pre-mark” the end points of heterochromatic spreading. Therefore, the domains do not appear to be a consequence of the formation of heterochromatin. Taken together, the above observations strongly suggest that the topological domain boundaries correlate with regions of the genome displaying classical insulator and barrier element activity, thus revealing a potential link between the topological domains and transcriptional control in the mammalian genome.

We compared the topological domains with previously described domain-like organizations of the genome, specifically with the A and B compartments described by Lieberman-Aiden et al.,² with Lamina-Associated Domains (LADs)^{11,15}, replication time zones,^{16,17} and Large Organized Chromatin K9-modification (LOCK) domains¹⁸. In all cases, we can see that topological domains are related to, but independent from, each of these previously described domain-like structures (Supplemental Figures 12–15). Notably, a subset of the domain boundaries we identify appear to mark the transition between either LAD and non-LAD regions of the genome (Figure 2f, Supplemental Figure 12), the A and B compartments (Supplemental Figure 13, 14), and early and late replicating chromatin (Supplemental Figure 14). Lastly, we can also confirm the previously reported similarities between the A and B compartments and early and late replication time zone (Supplemental Figure 16)¹⁷.

We next compared the locations of topological boundaries identified in both replicates of mESCs and cortex, or between both replicates of hESCs and IMR90 cells. In both human and mouse, the majority of the boundary regions are shared between cell types (Figure 3a, Supplemental Figure 17a), suggesting that the overall domain structure between cell types is largely unchanged. At the boundaries called in only one cell type, we noticed that trend of upstream and downstream bias in the DI is still readily apparent and highly reproducible

between replicates (Supplemental figure 17b,c). We cannot determine if the differences in domain calls between cell types is due to noise in the data or due to biological phenomena, such as a change in the strength of the boundary region between cell types¹⁹. Regardless, these results suggest that the domain boundaries are largely invariant between cell types. Lastly, only a small fraction of the boundaries show clear differences between two cell types, suggesting that a relatively rare subset of boundaries may actually differ between cell types (Supplemental Figure 18).

The stability of the domains between cell types is surprising given previous evidence showing cell type specific chromatin interactions and conformations^{6,8}. To reconcile these results, we identified cell-type specific chromatin interactions between mouse ES cell and mouse cortex. We identified 9,888 dynamic interacting regions in the mouse genome based on 20kb binning using a binomial test with an empirical false discover rate of < 1% based on random permutation of the replicate data. These dynamic interacting regions are enriched for differentially expressed genes, (Figure 3b–d, Supplemental Figure 19, Supplemental Table 5). In fact, 20% of all genes that undergo a 4-fold change in gene expression are found at dynamic interacting loci. This is likely an underestimate, because by binning the genome at 20kb, any dynamic regulatory interaction less than 20kb will be missed. Lastly, > 96% of dynamic interacting regions occur in the same domain (Figure 3e). Therefore, we favor a model where the domain organization is stable between cell types, but the regions within each domain may be dynamic, potentially taking part in cell-type specific regulatory events.

The stability of the domains between cell types prompted us to investigate if the domain structure is also conserved across evolution. To address this, we compared the domain boundaries between mouse ES cells and human ES cells using the UCSC liftover tool. The majority of boundaries appear to be shared across evolution (53.8% of human boundaries are boundaries in mouse and 75.9% of mouse boundaries are boundaries in humans, compared to 21.0% and 29.0% at random, p -value $< 2.2 \times 10^{-16}$, Fisher's Exact Test) (Figure 3f). The syntenic regions in mouse and human in particular share a high degree of similarity in their higher order chromatin structure (Figure 3g,h), indicating that there is conservation of genomic structure beyond the primary sequence of DNA.

We explored what factors may contribute to the formation of topological boundary regions in the genome. While most topological boundaries are enriched for the binding of CTCF, only 15% of CTCF binding sites are located within boundary regions (Figure 2c). Thus, CTCF binding alone is insufficient to demarcate domain boundaries. We reasoned that additional factors might be associated with topological boundary regions. By examining the enrichment of a variety of histone modifications, chromatin binding proteins, transcription factors, around topological boundary regions in mESC, we observed that factors associated with active promoters and gene bodies are enriched at boundaries in both mouse and humans (Figure 4a and Supplemental Figures 20–23)^{20,21}. In contrast, non-promoter associated marks, such as H3K4me1 (associated with enhancers) and H3K9me3, were not enriched or were specifically depleted at boundary regions (Figure 4a). Furthermore, transcription start sites (TSS) and global run on sequencing (GRO-Seq)²² signal were also enriched around topological boundaries (Figure 4a). We found that “housekeeping genes” were particularly strongly enriched near topological boundary regions (Figure 4b–d, See Supplemental Table

7 for complete GO terms enrichment). Additionally, the tRNA genes, which have the potential to function as boundary elements^{23,24}, are also enriched at boundaries (p-value < 0.05, Fisher's exact test (Figure 4b). These results suggest that high levels of transcription activity may also contribute to boundary formation. In support of this, we can see examples of dynamic changes in H3K4me3 at or near some cell-type specific boundaries that are cell type-specific (Supplemental Figure 24). Indeed, boundaries associated with both CTCF and a housekeeping gene account for nearly a third of all topological boundaries in the genome (Figure 4e, Supplemental Figure 24)

Lastly, we analyzed the enrichment of repeat classes around boundary elements. We observed Alu/B1 and B2 SINE elements in mouse and Alu SINE elements in humans are enriched at boundary regions (Figure 4a, Supplemental Figures 24,25). In light of recent reports indicating that a SINE B2 element functions as a boundary in mice²⁵, and SINE element retrotransposition may alter CTCF binding sites during evolution²⁶, we believe this contributes to a growing body of evidence suggesting a role for SINE elements in the organization of the genome.

In summary, we show that the mammalian chromosomes are segmented into megabase-sized topological domains, consistent with some previous models of the higher order chromatin structure^{1,27,28}. Such spatial organization appears to be a general property of the genome: it is pervasive throughout the genome, stable across different cell types and highly conserved between mice and humans.

We have identified multiple factors that are associated with the boundary regions separating topological domains, including the insulator binding factor CTCF, housekeeping genes, SINE elements. The association of housekeeping genes with boundary regions extends previous studies in yeast, insects and lower vertebrates and suggests that non-CTCF factors may be also involved in insulator/barrier functions in mammalian cells²⁹.

The topological domains we identified are well conserved between mice and humans. This suggests that the sequence elements and mechanisms that are responsible for establishing higher order structures in the genome may be relatively ancient in evolution. A similar partitioning of the genome into physical domains has also been observed in *Drosophila* embryos³⁰ and in high-resolution studies of the X-inactivation center in mice (termed Topologically Associated Domains or TADs)³¹, suggesting that topological domains may be a fundamental organizing principle of metazoan genomes.

Method Summary

Cell Culture and Hi-C Experiments

J1 mouse embryonic stem cells were grown on gamma-irradiated mouse embryonic fibroblasts cells under standard conditions (85% High Glucose DMEM, 15% HyClone FBS, 0.1mM non-essential amino acids, 0.1mM β -mercaptoethanol, 1mM Glutamine, LIF 500U/mL, +P/S). Before harvesting for Hi-C, J1 mESCs were passaged onto feeder free 0.2% gelatin coated plates for at least 2 passages to rid the culture of feeder cells. H1 Human embryonic stem cells and IMR90 fibroblasts were grown as previously described¹⁴.

Harvesting the cells for Hi-C was performed as previously described, with the only modification being that the adherent cell cultures were dissociated with trypsin prior to fixation.

Sequencing and Mapping of Data

Hi-C analysis and paired end libraries were prepared as previously described² and sequenced on the Illumina Hi-Seq2000 platform. Reads were mapped to reference human (hg18) or mouse genomes (mm9), and non-mapping reads and PCR duplicates were removed. 2-dimensional heat-maps were generated as previously described².

Data Analysis

For detailed descriptions of the data analysis, including descriptions of the directionality index, hidden Markov models, dynamic interactions identification, and boundary overlap between cells and across species, see supplemental methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful for the valuable comments from and discussions with Drs. Zhaohui Qin (Emory University), Arshad Desai (LICR/UCSD), and members of the Ren lab during the course of the present study. We also thank Drs. Wendy Bickmore and Ragnhild Eskeland for sharing the FISH data generated in mouse ES cells. This work was supported by funding from the Ludwig Institute for Cancer Research, California Institute for Regenerative Medicine (CIRM, RN2-00905-1) (to B.R.), and NIH (B.R. R01GH003991). JD is funded by a pre-doctoral training grant from CIRM. YS is supported by a postdoctoral fellowship from the Rett Syndrome Research Foundation.

References

1. Cremer T, Cremer M. Chromosome territories. *Cold Spring Harb Perspect Biol.* 2:a003889. [PubMed: 20300217]
2. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 2009; 326:289–93. [PubMed: 19815776]
3. Shen Y, et al. A Map of cis-Regulatory Sequences in the Mouse Genome. 2012 in submission.
4. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet.* 43:1059–65. [PubMed: 22001755]
5. Wang KC, et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature.* 472:120–4. [PubMed: 21423168]
6. Kagey MH, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature.* 467:430–5. [PubMed: 20720539]
7. Eskeland R, et al. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol Cell.* 38:452–64. [PubMed: 20471950]
8. Noordermeer D, et al. The dynamic architecture of Hox gene clusters. *Science.* 334:222–5. [PubMed: 21998387]
9. Kim YJ, Cecchini KR, Kim TH. Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. *Proc Natl Acad Sci U S A.* 108:7391–6. [PubMed: 21502535]
10. Phillips JE, Corces VG. CTCF: master weaver of the genome. *Cell.* 2009; 137:1194–211. [PubMed: 19563753]

11. Guelen L, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*. 2008; 453:948–51. [PubMed: 18463634]
12. Handoko L, et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet*. 43:630–8. [PubMed: 21685913]
13. Xie W, et al. Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome. *Cell*. 148:816–31. [PubMed: 22341451]
14. Hawkins RD, et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell*. 6:479–91. [PubMed: 20452322]
15. Peric-Hupkes D, et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell*. 38:603–13. [PubMed: 20513434]
16. Hiratani I, et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res*. 20:155–69. [PubMed: 19952138]
17. Ryba T, et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res*. 20:761–70. [PubMed: 20430782]
18. Wen B, Wu H, Shinkai Y, Irizarry RA, Feinberg AP. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat Genet*. 2009; 41:246–50. [PubMed: 19151716]
19. Scott KC, Taubman AD, Geyer PK. Enhancer blocking by the *Drosophila* gypsy insulator depends upon insulator anatomy and enhancer strength. *Genetics*. 1999; 153:787–98. [PubMed: 10511558]
20. Bilodeau S, Kagey MH, Frampton GM, Rahl PB, Young RA. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev*. 2009; 23:2484–9. [PubMed: 19884255]
21. Marson A, et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*. 2008; 134:521–33. [PubMed: 18692474]
22. Min IM, et al. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev*. 25:742–54. [PubMed: 21460038]
23. Donze D, Kamakaka RT. RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in *Saccharomyces cerevisiae*. *EMBO J*. 2001; 20:520–31. [PubMed: 11157758]
24. Ebersole T, et al. tRNA genes protect a reporter gene from epigenetic silencing in mouse cells. *Cell Cycle*. 10:2779–91. [PubMed: 21822054]
25. Lunyak VV, et al. Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science*. 2007; 317:248–51. [PubMed: 17626886]
26. Schmidt D, et al. Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell*.
27. Jhunjhunwala S, et al. The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell*. 2008; 133:265–79. [PubMed: 18423198]
28. Capelson M, Corces VG. Boundary elements and nuclear organization. *Biol Cell*. 2004; 96:617–29. [PubMed: 15519696]
29. Amouyal M. Gene insulation. Part I: natural strategies in yeast and *Drosophila*. *Biochem Cell Biol*. 88:875–84. [PubMed: 21102650]
30. Sexton T, et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*. 148:458–72. [PubMed: 22265598]
31. Nora EP, et al. Spatial partitioning of the regulatory landscape of the X-inactivation center. 2012 In submission.

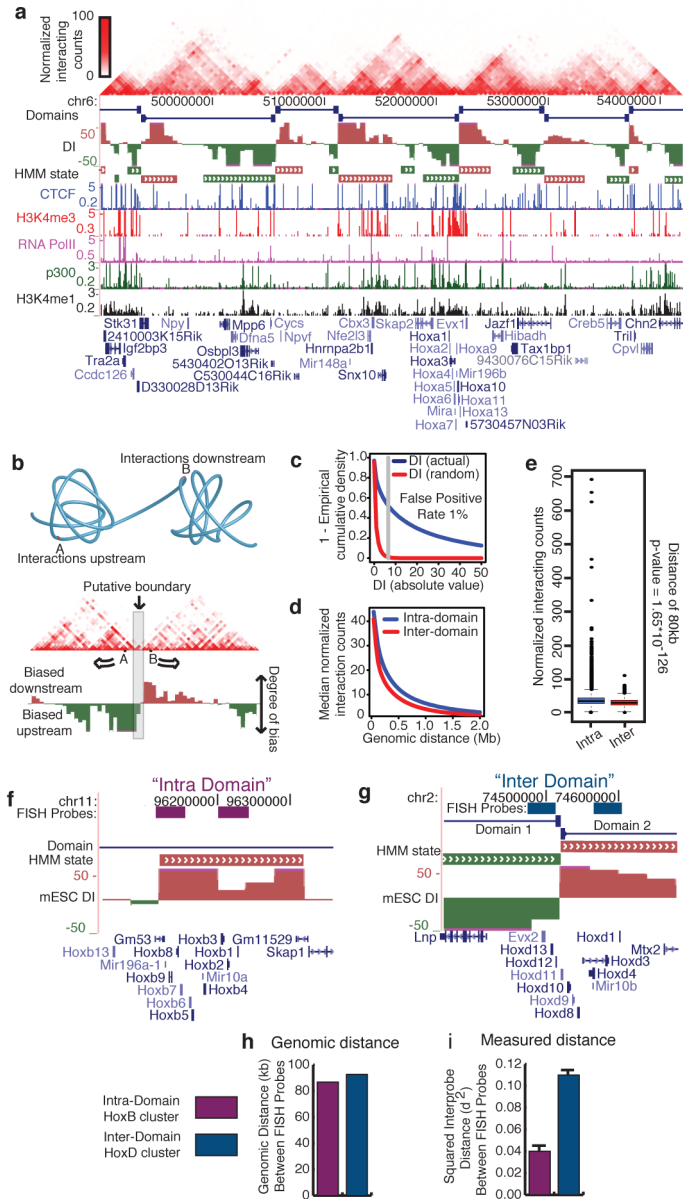


Figure 1. Topological Domains in the Mouse ES cell Genome
 a, Normalized Hi-C interaction frequencies displayed as a 2D heatmap overlaid on ChIP-Seq data (from ref. 3), DI, HMM Bias State Calls, and domains. For both DI and HMM State calls, downstream bias (red) and upstream bias (green) are indicated. b, Schematic illustrating topological domains and resulting directional bias. c, Distribution of the DI (absolute value, in blue) compared to random (red). d, Mean interaction frequencies at all genomic distances between 40kb to 2Mb. Above 40kb, the intra-versus inter-domain interaction frequencies are significantly different ($p < 0.005$, Wilcoxon test). e, Box plot of all interaction frequencies at 80kb distance. Intra-domain interactions are enriched for high-frequency interactions. f–i, Diagram of “Intra-domain” (f) and “Inter-domain” FISH probes (g) and the genomic distance between pairs (h). i, Bar chart of the squared interprobe

distance (from ref. 7) FISH probe pairs. Error bars indicate standard error (n = 100 for each probe pair).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

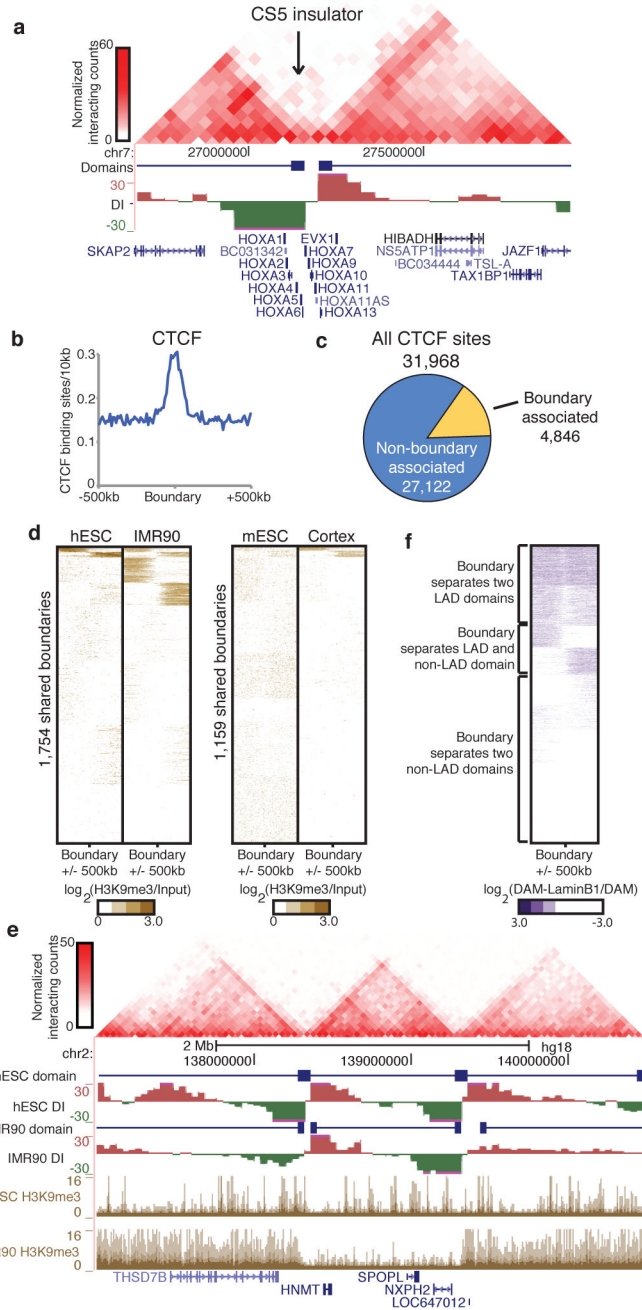


Figure 2. Topological Boundaries Demonstrate Classical Insulator or Barrier Elements Features
 a, 2D heatmap surrounding the HoxA locus and CS5 insulator in IMR90 cells. b, Enrichment of CTCF at boundary regions. c, The portion of CTCF binding sites that are considered “associated” with a boundary (within +/- 20kb window is used as the expected uncertainty due to 40kb binning). d, Heat maps of H3K9me3 at boundary sites in human and mouse. e, UCSC Genome Browser shot showing heterochromatin spreading in the human ES cells and IMR90 cells. The 2D heat map shows the interaction frequency in hES cells. f, Heat map of LADs (from ref. 15) surrounding the boundary regions.

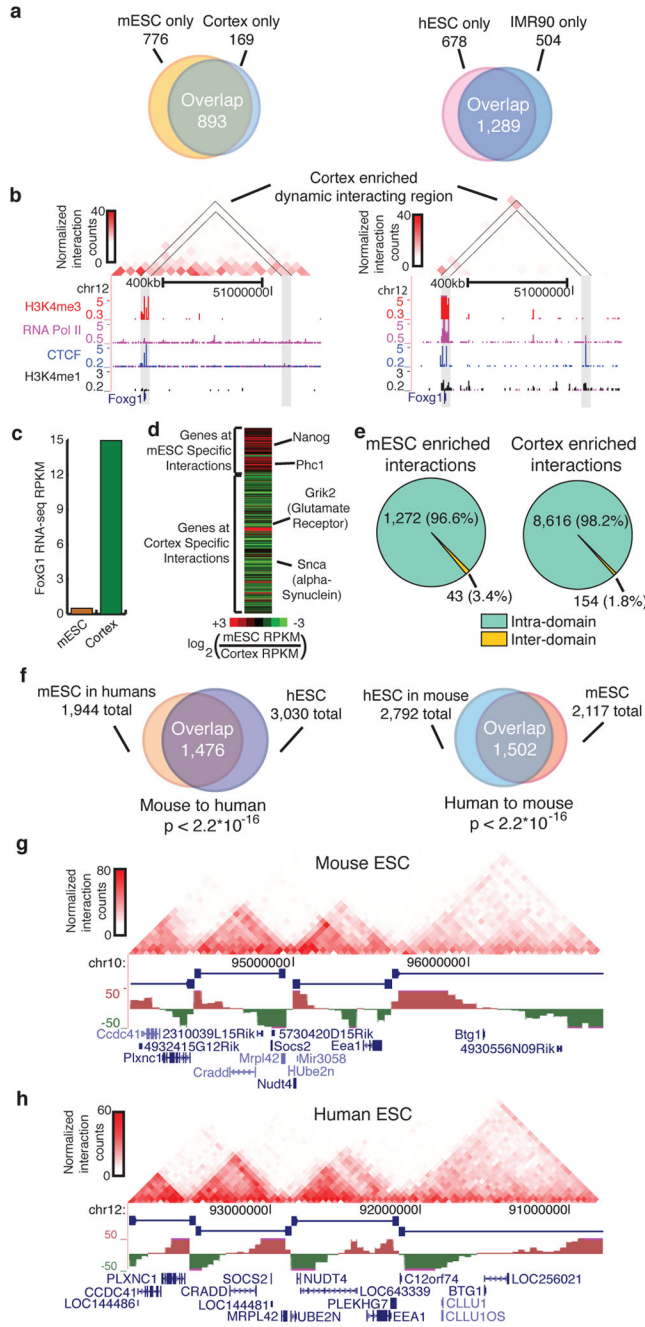


Figure 3. Boundaries are shared across cell types and conserved in evolution

a, Overlap of boundaries between cell types. b, Genome browser shot of a cortex enriched dynamic interacting region that overlaps with the Foxg1 gene. c, Foxg1 expression in mouse ES cells and cortex as measured by RNA-seq. d, Heat map of the gene expression ratio between mouse ES cell and cortex of genes at dynamic interactions. e, Pie chart of inter- and intra-domain dynamic interactions. f, Overlap of boundaries between syntenic mouse and human sequences (p -value $< 2.2 \cdot 10^{-16}$ compared to random, Fisher's exact test). g and h, Genome browser shots showing domain structure over a syntenic region in the mouse (g)

and human ES cells (h). Note: the region in humans has been inverted from its normal UCSC coordinates for proper display purposes.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

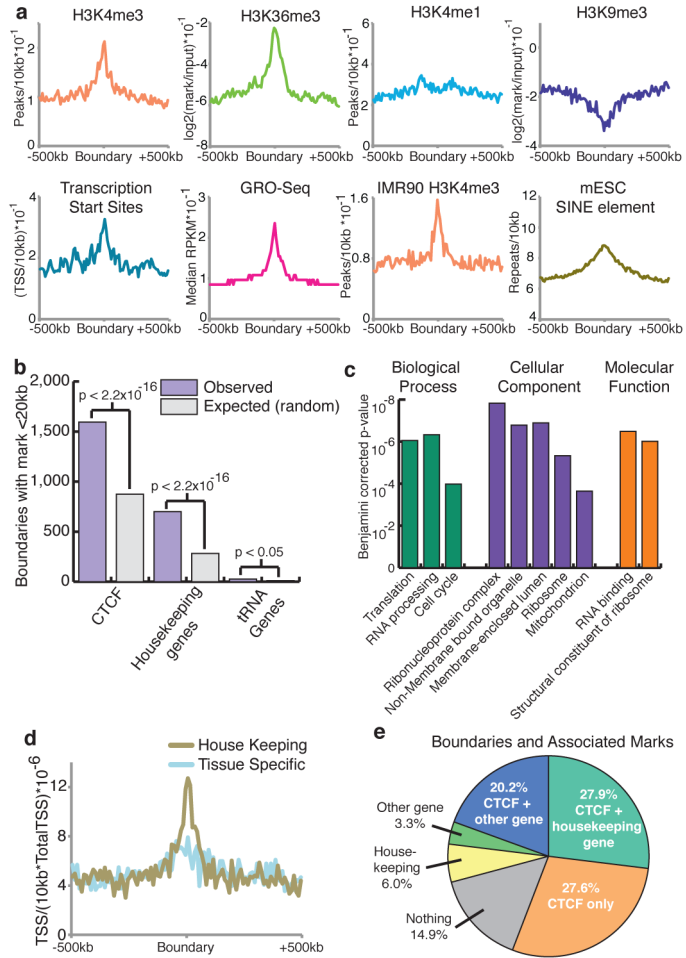


Figure 4. Boundary regions are enriched for housekeeping genes

a, Chromatin modifications, TSS, GRO-Seq, and SINE elements surrounding boundary regions in mESCs or IMR90. b, Boundaries associated with a CTCF binding site, housekeeping gene, or tRNA gene (purple) compared to expected at random (grey). c, Gene Ontology p-value chart. d, Enrichment of housekeeping genes (gold) and tissue specific genes (blue) as defined by on Shannon entropy scores near boundaries normalized for the number of genes in each class (TSS/10kb/total TSS). e, Percentage of boundaries with a given mark within 20kb of the boundaries.