UNIVERSITY OF CALIFORNIA, SAN DIEGO

Higher Order Chromatin Architecture in Mammalian Genomes

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biomedical Sciences

by

Jesse Dixon

Committee in charge:

Professor Bing Ren, Chair
Professor Arshad Desai
Professor Ronald Evans
Professor Christopher Glass
Professor Cornelius Murre

2013

The Dissertation of Jesse Dixon is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2013

DEDICATION

I would like to dedicate this dissertation to my family. To my parents, Jack and Claudia, who supported my scientific endeavors from a very early age. To my sister, Sarah, who has looked out for me and served as a role model. And to my beautiful wife, Katie, who, in addition to marrying me, has supported me during my PhD in enumerable ways. I love you all.

EPIGRAPH

You have got to be in the middle of the street if you want to get hit by a bus.

Jack Dixon

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures vii

Acknowledgements xi

Vita xiii

Abstract of the Dissertation xiv

Chapter 1 1

Chapter 2 28

Chapter 3 87

Chapter 4 134

Chapter 5 168

LIST OF FIGURES

Chapter 1

Figure 1. Global features of nuclear organization 16

Figure 2. Outline of proximity based ligation methods 18

Figure 3. Looping interaction at the β-globin locus 20

Chapter 2

Figure 1. Topological domains in the mouse ES cell genome.....29

Figure 2. Topological boundaries demonstrate classical insulator or barrier element features.....29

Figure 3. Boundaries are shared across cell types and conserved in evolution 30

Figure 4. Boundaries regions are enriched for housekeeping genes.....31

Supplementary Figure 1. Raw Hi-C data and restriction enzyme bias 49

Supplementary Figure 2. Normalized Hi-C data shows no restriction enzyme bias 51

Supplementary Figure 3. Pearson correlation between replicates 52

Supplementary Figure 4. Comparison with previous 5C.....54

Supplementary Figure 5. Comparison with previous 3C data 55

Supplementary Figure 6. Hi-C interaction frequency and mean spatial distance.....56

Supplementary Figure 7. Hi-C interaction heat maps at varying bin sizes.....58

Supplementary Figure 8. Overlap of topological domain boundaries between Hi-C replicates 59

Supplementary Figure 9. Size distribution and gene content of topological domains, boundaries and unorganized chromatin 60

Supplementary Figure 10. CTCF enrichment at topological domain boundary regions ...61

Supplementary Figure 11. Average enrichment plot of H3K9me3 surrounding the boundaries 62
Supplementary Figure 12. Comparison of topological domains with lamina associated domains (LADs) 63
Supplementary Figure 13. Comparison of A and B compartments with topological domains in mouse ES cells 64
Supplementary Figure 14. Comparison of topological domains with A and B compartments and replication time zones 66
Supplementary Figure 15. Comparison of topological domains with LOCK domains 67
Supplementary Figure 16. Correlation of A and B compartments and replication time zones in mouse ES cells 68
Supplementary Figure 17. Domains are largely stable between cell types 70
Supplementary Figure 18. Cell type specific domains 72
Supplementary Figure 19. Enrichment of differentially expressed genes at dynamic interaction regions 73
Supplementary Figure 20. Histone modification, chromatin binding protein, and transcription factor enrichment near boundary regions 74
Supplementary Figure 21. Heat maps of boundary enrichment of histone modification, chromatin binding protein, and transcription factor enrichment near boundary regions ... 76
Supplementary Figure 22. Heat maps of boundary enrichment of histone modification, chromatin binding protein, and transcription factor enrichment near boundary regions ... 78
Supplementary Figure 23. Marks enriched at boundaries in each mouse ES cell replicate ..
..... 80

Supplementary Figure 24. Random association of CTCF and housekeeping genes in mESCs 81
Supplementary Figure 25. Repeat content at mouse ES cell boundaries.....82
Supplementary Figure 26. Repeat content at human boundaries.....83
Supplementary Figure 27. Expected intermolecular ligations.....84
Supplementary Figure 28. HMM with mixture of Gaussian output.....85
Chapter 3
Figure 1. Cohesin cleavage reduces long-range interactions within the H19/IGF2 domain 109
Figure 2. Cohesin cleavage reduces interactions within toplogical domains genome-wide.....111
Figure 3. CTCF depletion reduces the function of domain boundaries.....113
Figure 4. Transcriptional changes after cohesin cleavage and CTCF depletion.....115
Supplementary Figure 1. Replacement of endogenous RAD21 by RAD21cv and incorporation in the cohesin complex.....116
Supplementary Figure 2. RAD21 cleavage causes premature loss of sister chromatid cohesin 117
Supplementary Figure 3. Cell cycle distribution of treated and untreated RAD21cv cells.....128
Supplementary Figure 4. Loss of interactions around KCNQ1 after RAD21 cleavage..119
Supplementary Figure 5. Conservation of long-range chromosomal interactions between different tissues.....120

Supplementary Figure 6. Control cells expressing RAD21 wt w/o HRV site do not respond to HRV protease transfection 121
Supplementary Figure 7. Enrichment of cohesin/CTCF sites at boundaries and differential heat map plot after RAD21 cleavage 122
Supplementary Figure 8. Live cell imaging shows preservation of chromatin morphology after RAD21 cleavage 123
Supplementary Figure 9. Cohesin depletion leads to a loss of intra-domain boundary associated interactions 124
Supplementary Figure 10. CTCF RNAi depletes CTCF but does not change the levels of cohesin bound to chromatin 125
Supplementary Figure 11. Changes of long-range interactions around the HOXA and the HOXB locus after RAD21 cleavage 127
Supplementary Figure 12. Empirical cumulative density plot of the distance from the transcription start site of an RAD21 or CTCF regulated gene to the nearest binding site for either SMC3 or CTCF 128
Chapter 4
Figure 1. Reproducibility of Hi-C data 154
Figure 2. Stability of topological domains 156
Figure 3. Alterations in intra-domain interaction frequency between lineages 158
Figure 4. Identification of A/B compartments 160
Figure 5. Inter-domain changes in interaction frequency and the A/B compartments 162
Figure 6. Association between A/B compartment changes and gene expression 163
Figure 7. Identification of genes associated with changes in A/B compartment status 164

ACKNOWLEDGEMENTS

I would like to acknowledge my advisor, Professor Bing Ren, for his guidance and support during my Ph.D. He has been instrumental in the creative and analytical aspects of the work. He supported me undertaking an ambitious project and has helped guide it to exciting new places.

I would also like to acknowledge our collaborators in the laboratory of Kerstin Wendt. Kerstin and her student Jessica Zuin in particular have contributed to a fruitful collaboration that is described in chapter 3 of this dissertation.

I would like to thank several people who have helped me considerably during my Ph.D. in both the analytical and experimental aspects of my work. I have been greatly aided by Siddarth Selvaraj, Dr. Feng Yue, and Dr. Inkyung Jung in analytical aspects of the work. Siddarth has made significant contributions to our efforts to identify and characterize topological domains as well as leading analytical efforts that extend beyond the scope of this dissertation. Feng has been instrumental in our understanding of topological domains and in creating tools that allow for their efficient visualization. As Bing says, Feng "gave us eyes." Inkyung has made important contributions to our ongoing efforts to characterize the changes in higher order chromatin structure during differentiation of human embryonic stem cells. He has contributed to the analysis in this dissertation as well as work that expands beyond its scope.

I would like to thank Audrey Kim, Dr. Yin Shen, and Dr. Yan Li in their help in experimental aspects of this work, some of which is described in this dissertation. Audrey also helped me in many different projects during my Ph.D., many of which are

not included here. Yin has taught me tremendous amounts about the practice of molecular biology, and for this I am grateful.

I would also like to thank Dr. Feng Yue and Dr. Gary Hon for their help in teaching me how to do bioinformatic analysis. When I joined Bing's lab, the extent of my computer expertise was making a web page in 8th grade computer class. The tremendous amount that I feel I have learned since is owed in large part to Feng and Gary being fantastic and patient teachers.

I would also like to thank Dr. Feng Yue and Dr. Andrea Smallwood for critical reading of parts of this dissertation. Andrea also deserves thanks for helping me get started when I first joined the lab and for answering the seemingly endless stream of questions that I had at that time.

Lastly, I would like to thank my committee, Professor Arshad Desai, Professor Ron Evans, Professor Christopher Glass, and Professor Cornelius Murre for their guidance and helpful suggestions throughout my Ph.D.

Chapter 2, in full, is a reprint of the material as it appears in *Nature*, volume 485, May 17 2012. Dixon, Jesse R.; Selvaraj, Siddarth, Selvaraj; Yue, Feng; Kim, Audrey; Li, Yan; Shen, Yin; Hu, Ming; Liu, Jun S.; Ren, Bing. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, has been submitted for publication of the material as it may appear in *Nature*. Zuin, Jessica; Dixon, Jesse R.; van der Reijden, Michael .I.J.A; Kolovos, Petros; Ye, Zhen; Brouwer, Rutger W.W.; van de Corput, Mariette P.C.; van Ijcken, Wilfred F.J.; Grosveld, Frank G.; Ren, Bing; Wendt, Kerstin S. The dissertation author was the co-primary investigator and the co-primary author of this material.

VITA

2006	Bachelor of the Arts, Princeton University
2007	Research Technician, University of Michigan
2013	Doctor of Philosophy, University of California, San Diego
expected 2015	Doctor of Medicine, University of California, San Diego

PUBLICATIONS

Pagliarini DJ, Wiley SE, Kimple ME, **Dixon JR**, Kelly P, Worby CA, Casey PJ, and Dixon JE (2005) Involvement of a mitochondrial phosphatase in the regulation of ATP production and insulin secretion in pancreatic beta cells. Mol. Cell 19 197-207.

Barish, GD, Yu, RT, Karunasiri, M, Ocampo, CB, **Dixon, J**, Benner, C, Dent, AL, Tangirala, RK, Evans, RM (2010) Bcl-6 and NF-kappaB cistromes mediate opposing regulation of the innate immune response. Genes Dev. 24, 2760-2765

Dixon, JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485 (7398) 376-80.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, **Dixon J**, Lee L, Lobanenkov VV, Ren B. (2012) A map of cis-regulatory sequences in the mouse genome. Nature 488 (7409):116-20.

Hu M, Deng K, Qin Z, **Dixon J**, Selvaraj S, Fang J, Ren B, Liu JS. (2013) Bayesian inference of spatial organizations of chromosomes. PLoS Computational Biology 9, e1002893

ABSTRACT OF THE DISSERTATION

Higher Order Chromatin Architecture in Mammalian Genomes

by

Jesse Dixon

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2013

Professor Bing Ren, Chair

A detailed understanding of higher order chromatin structure is critical to understand the mechanisms used by regulatory elements to affect expression of their target genes. Yet, until recently, most methods used to study higher order chromatin structure were laborious and low-throughput. Recent advancements in the methods used to study chromatin structure have allowed for the first time the identification of genome

wide patterns of higher-order chromatin interactions. During my Ph.D, I have used the Hi-C technique to study genome wide patterns of chromatin interactions. I have observed that chromosomes appear to fold into megabase-sized self-interacting structures that we have termed "topological domains." These domains are stable between cell types and conserved in evolution. These domains appear to be separated from each other by boundary elements in the genome, and appear to contain clusters of co-regulated *cis*-acting elements. The structure of the domains is dependent on the DNA-binding factor CTCF and the Cohesin complex. Notably, Cohesin and CTCF appear to have unique roles in regulating domain structure. CTCF appears to affect both intra- and inter-domain interactions, whereas Cohesin appears to primarily affect intra-domain interactions. When we compare topological domain interaction patterns across a variety of embryonic stem cell derived lineages, we observe that the interactions appear to be regulated on a domain-wide scale. Increases in domain-wide interaction frequency within domains correlates with active histone modifications, DNaseI hypersensitivity, and increased gene expression. In addition, we observe a wide-spread re-organization of inter-domain interactions between cell types. This correlates with a re-structuring of the "A" and "B" compartments in the nucleus. The alterations of the A/B compartments appear to modestly correlate with changes in gene expression, most notably for particular subsets of genes. We anticipate that these studies will lay the ground work for future experiments to elucidate the impact of higher order chromatin structures on diverse fields of biology, ranging from human disease to the evolution of genomes.

Chapter 1

Introduction

The sequencing of the human genome in 2001 revealed for first time the composition of our genome and demonstrated that nearly 98% of it did not have the capacity to code for any protein. Since that time, significant efforts have been devoted to elucidating the functions of the non-coding portion of our genome. Technological advances, namely in high-throughput shotgun sequencing technologies, have greatly facilitated this effort. Studies from our lab and others mapping chromatin modifications and transcription factor binding sites have demonstrated that *cis*-regulatory elements, such as enhancers, promoters, and insulators, are pervasive throughout the genome. In addition, by identifying the location of enhancers throughout the genome, it is clear that many exist at great genomic distances from any possible gene they could regulate. These results appear logical in context, as previous research has suggested that mechanistically, enhancers are known to regulate distal genes by being brought in close physical proximity to a target gene by "looping" of the intervening DNA sequences. Likewise, insulators are also proposed to utilize higher-order chromatin structures in order to functionally partition DNA elements. The implication, therefore, is that to truly understand the function of the regulatory elements that appear to litter the 98% percent of our genome that is non-protein coding, we need a better understanding of how chromosomes fold into higher order structures in the nucleus.

Our understanding of higher-order chromatin structure was initially facilitated by studies utilizing light and electron microscopy. Early studies described the nucleus as

being organized into two compartments, heterochromatin and euchromatin (1). The initial distinction was based on electron density, with the more electron dense heterochromatin typically located at the nuclear periphery while the less electron dense euchromatin occupied a greater volume of the nuclear interior. Years of subsequent research has demonstrated that in addition to their physical separation from each other, heterochromatin and euchromatin are also functionally distinct. Euchromatin is typically more "active" and is the site of the majority of gene expression, while heterochromatin is considered "inactive" or repressed. However, euchromatin and heterochromatin are not static, as recent studies have demonstrated that pluripotent stem cells typically have less heterochromatin than their differentiated progeny (2). In addition, while the vast majority of cells follow the pattern of having heterochromatin at the periphery of the nucleus, rod photoreceptors are a clear exception to this rule, with their heterochromatic compartment located in the center of the nucleus (3). The functional implications of this rearrangement are not yet clear. Remarkably, despite this rearrangement of the locations of euchromatin and heterochromatin relative to the nuclear periphery, the two remain spatially distinct compartments in the nucleus.

Heterochromatin and euchromatin describe nuclear compartmentalization on a global level. However, within the nucleus, individual chromosomes also form distinct structures. During interphase, each chromosome occupies a distinct territory, termed a chromosome territory (CT)(4, 5) (Figure 1a). While chromosome territories were originally identified using light microscopy techniques, recent methods, including interphase chromosome painting and more high-throughput techniques (6), have confirmed these observations. Furthermore, CTs occupy non-random positions within

the nucleus, with smaller, more gene dense chromosomes typically located closer to the nuclear center, while larger more gene poor chromosomes are located near the nuclear periphery (7, 8). The mechanisms that lead to CT formation and non-random positioning of chromosomes in the nucleus have yet to be discovered. However, the CT model does make clear predictions, namely, that regulatory elements functioning via long-range interactions would be more likely to act in *cis* than in *trans* simply due to increased spatial proximity.

Our understanding of the structure of chromosome within CTs remains murky, but has been aided by several observations using fluorescence microscopy and *in situ* hybridization (FISH). One property of chromosome folding at the sub-CT scale observed by multiple groups been termed "looping out" (Figure 2b). When a region is "looped out" of its chromosome territory, the locus, in some cases several megabases worth of the genome, moves outside the bulk of its chromosome territory. By using FISH probes labeling a particular region of the genome coupled with chromosome paints, "looping out" has been most clearly observed at regions such as the HoxB locus (9) and the Major Histocompatibility Locus (10). It has been suggested that "looping out" is more common for regions that have undergone transcriptional activation (9) and for gene rich loci (10). Gene density may be the critical factor. For example, gene rich regions localize to the edge of a CT even in the absence of transcription (11). Additionally, under steady state conditions gene expression can be evident in the interior of a CT (12). The observation that genes are located at the periphery of a CT is also supported by the observation that active genes, even those located *in trans*, are more likely to co-localize in the nucleus than would be expected by random chance (13). Given the organization of individual

chromosomes as CTs, this would only be possible if more gene dense regions were located at their periphery.

Recent studies have expanded upon the idea of looping out of genes at the periphery of a chromosome territory. By using oligo-synthesized FISH probes for all exons along mouse chromosome 2, Bickmore and colleagues showed that exons tend to localize *en masse* outside of what is observed as a CT by chromosome painting (14). These authors suggest that "looping out" may therefore be a product of de-condensation of active regions of a chromosome at the periphery of its CT. As a result, chromosome painting is less intense in these regions, leading to the perception that a genomic locus is located distal to the remainder of its CT. This suggests that the term "looping out," which may imply that a single region re-locates outside of a CT, may not be an entirely accurate description of the positioning of gene dense regions at the periphery of a CT.

FISH data has also formed the experimental basis for theoretical models of how chromosomes fold within a CT. Central to these theoretical models is the idea that, at its most basic level, a chromosome is a polymer of nucleotides. These models therefore assume that chromosomes should fold on some level according to physical principles that govern the behavior of polymers in solution. Typically, these polymer physics based models are tested by comparing the theoretical predictions of chromosome folding with experimental data derived from FISH probes tiled along regions of the genome.

The earliest of these models relied on the idea that, free of constraints, a polymer undergoing random fluctuations should follow a random-walk behavior. The prediction from random-walk behavior is that the average spatial separation of two probes should depend on the square-root of the linear genome distance between them (15). When

analyzing the spatial distance between FISH probes separated by genomic distances less than 1-2Mbp, this was indeed the case (15). However, a clear alteration in the scaling of the separation between loci was observed above 2Mbp, suggesting some level of constraint in chromosome structure at that length. The authors concluded from this that a chromosome is composed of 1-2Mbp loops of chromatin that determine the degree of scaling of spatial separation between loci. This model was termed the "Random-walk/giant-loop" model for chromosome organization (16, 17). Alterative models have suggested that instead of single giant loops, "rosettes" containing multiple loops at the same anchor can also account for the observed distances between FISH probes, termed the "multi-loop sub-compartment" model (18, 19), which may have better agreement with experimental data (20).

Central to both the random-walk /giant loop and the multi-loop sub-compartment model is the concept that at genomic distances less than 1-2Mbp, the patterns of spatial separation of loci is markedly different than at distances larger than 1-2Mbps. This phenomenon occurs in both models and in the experimental data, and suggests that some genomic constraints are present at a size of 1-2Mbp that contribute to the overall structure of chromosomes inside chromosome territories. At genomic distances much larger than 1-2Mbps, there can be a "leveling off" in the average spatial distance between two loci as this distance approaches the size of the chromosome territory (21), for which additional polymer-based models have been developed to account for this behavior (22).

Microscopy has limits to what it can reveal about the structure of the nucleus. While it can identify structures and relative nuclear spacing on a single cell level, FISH and electron microscopy can be laborious to perform and is limited in throughput and

resolution. As a result, other molecular biology based methods have been critical in our efforts to understand the principles nuclear organization. One approach that has yielded abundant information is to identify the linear sequences in the genome that correspond to known nuclear structures. This has been particularly fruitful when applied to studying the regions of the genome that associate with the nuclear lamina.

The linear sequences of the genome that associate with the nuclear lamina have been identified using the DamID technique. This approach creates stable cell lines where a protein of interest is fused to bacterial DNA adenine methylatransferase (Dam), resulting in adenine DNA-methylation of GATC sequences near binding sites of the target protein (23). Fusion of the constitutive B-type lamins and DamID mapping has yielded maps of Lamina-Associated Domains (LADs) in organisms ranging from flies to humans (23-25). Lamina-associated regions of the genome tend to be gene poor, A/T-rich, lowly expressed, and are functionally conserved in evolution (25, 26). Alterations in Lamina association appear to correlate with alterations in gene expression during development, with regions that move away from the lamina being activated or poised for later activation (24). However, the role of B-type lamins in gene suppression is unclear, as mouse ES cells and trophectodermal cells completely lacking any B-type lamins appear functionally normal and continue to suppress genes that are normally lamina associated (27). Instead B-type lamins appear to be critical for normal organogenesis. This may be mediated by diverse processes such as spindle orientation, and nuclear and neuronal migration (27). Ultimately, these results point to a role for Lamins as a potential genome organizer in the nucleus.

Studies of DNA replication timing in the genome have also provided insight into genome organization. Years of research has determined that the time at which different regions of the genome replicate their DNA during S-phase is clearly correlated with their positioning in the nucleus. Early replicating regions tend to be located in the interior of the nucleus and late replicating regions tend to be located at the nuclear periphery (28-31) (Figure 1c). High resolution mapping of these "replication domains" using microarrays revealed that early replicating regions are gene dense and generally more euchromatic, while late replicating regions are gene poor, A/T-rich, and contain fewer active genes. The latter are also well-correlated with the previously described lamina-associated domains (25, 32, 33). Furthermore, alterations in timing of DNA replication during development appear to correlate with activation or repression of gene expression (32, 33). Interestingly, early and late replicating domains appear to remain stably isolated from each other throughout the cell cycle, suggesting that these regions represent a fundamental compartmentalization of the nucleus (31, 34).

The similarity of the compartmentalization of the nucleus observed for LADs or replication timing suggests that these are related to fundamental properties of the genome. Notably, both replication timing and lamina-association have been correlated with underlying GC content. Indeed, mammalian genomes are composed of long (>300kbp) regions of either AT or GC rich regions termed "isochores" (35). The correlation of these regions with both replication timing and lamina-association suggests that AT and GC rich regions may also spatially compartmentalized in the nucleus. While the exact implications and evolutionary origins of this sequence compartmentalization

remain unclear (36), these structures are undoubtedly important and should be considered in any large scale study of nuclear genome organization.

While both FISH and microarray based studies have provided a tremendous amount of information about the global structure of the genome in the nucleus, these methods are insufficient to understand local higher-order structural patterns of chromatin. Knowledge of these structures is essential for our ability to elucidate the principles of communication between regulatory elements and their target genes. Fortunately, a remarkable set of "proximity-based ligation" methods have been developed over the last decade to allow the study of both local and genome wide higher-order chromatin structures.

All proximity-based ligation methods rely of simple molecular biology manipulations of chromatin to assess the frequency with which two regions of the genome interact. The first of these methods was termed Chromatin Conformation Capture, or 3C (37). In a 3C experiment, cells or tissues are fixed with formaldehyde to preserve both protein-protein and DNA-protein interactions. Nuclei are isolated and chromatin is digested with a restriction enzyme of choice. Digested free sticky-ends of chromatin are then ligated together to create contiguous fragments of DNA between regions of the genome that were originally in close spatial proximity. In a 3C assay, the frequency of interactions between two candidate loci is assessed using PCR compared to a control template (Figure 2c). 3C can be performed at high-resolution to assess the interaction frequency between loci separated by as little as 20kb, so it is ideal to test for interactions between regulatory elements and proximal target genes. 3C is limited,

however, by the need for *a priori* candidate regions to test in the assay. This limits the potential for *de novo* identification of interacting regions without a prior hypothesis.

The invention of several 3C variants limited the need for candidate regions and increased the throughput of the assay. The first of these approaches, termed 4C, relies on inverse PCR of a 3C library to test for all potential interacting loci with a "bait region" (note: multiple groups have developed "4C" assays, and 4C can alternatively mean "Chromosome Conformation Capture on ChIP" (38) or "Circular Chromosome Conformation Capture" (39)). After inverse PCR, the subsequent 4C library can be assayed by subsequent cloning (40, 41), hybridization to microarrays (38, 39), or, more recently, high-throughput DNA sequencing (42, 43) (Figure 2d). These methods still require one candidate or "bait" locus whose interacting partners are then identified in an unbiased, genome-wide manner. While there may be certain limits in the local resolution of 4C methods (44), this technique provides a powerful method for assessing the genome wide interaction partners of a candidate locus.

An alternative 3C variant developed to assess interaction frequency over a local region is termed Chromosome Conformation Capture Carbon-Copy, or 5C (45). 5C is designed to assess all possible interactions over a given region of the genome, which can range from 100kbp to several megabases. The initial fixation, digestion, and ligation steps are identical to 3C. However, after ligation, an oligonucleotide pool tiling the region of interest undergoes ligation mediated amplification (LMA) using the 3C library as a template. Each oligonucleotide has a universal set of PCR primers, so the entire 5C library can be amplified simultaneously and then sequenced (Figure 2e). 5C provides

high sensitivity and resolution, but, like 3C and 4C, is limited by the need to pick a candidate region of the genome for analysis.

Two recently developed genome-wide methods based on 3C have for the first time allowed a nearly unbiased analysis of chromatin interactions in the genome. The first of these methods introduced was a genome-wide unbiased 3C variant termed Hi-C (6) (Figure 2f). The initial steps of Hi-C are identical to 3C. However, in a Hi-C experiment, before the ligation of digested chromatin fragments, sticky-ends of DNA are filled in with nucleotides, one of which is covalently linked to biotin. After blunt end ligation, DNA is isolated and sheared. Ligated fragments are then specifically precipitated using a streptavidin coated bead, and the bead bound library is then PCR amplified and sequenced. This allows an unbiased assessment of both intra- and inter-chromosomal interactions throughout the genome. Hi-C can be subject to certain experimental biases due to variables such as the choice of restriction enzyme, the mappability of genomic regions, and the GC content of restriction fragment ends which must be accounted for during data analysis (46). Hi-C is also highly dependent on both the sequencing depth obtained for the library and the size of the genome of the organism being studies (47). Due to the current sequencing capacity of most laboratories, this can limit the resolution of the method and therefore the ability to analyze local chromatin interactions in mammalian organisms. However, with the constant improvements in sequencing technologies and the resultant price decreases, this limitation will likely vanish in the coming years.

A second genome-scale 3C variant that has been recently introduced is termed Chromatin Interaction Analysis by Paired-End Tag sequencing, or ChIA-PET (48)

(Figure 2g). ChIA-PET experiments are conceptually similar to Hi-C, yet the initial step is to immunoprecipitate a factor of interest, so that the resulting dataset is a map of interactions between regions of the genome bound by a particular factor. This method has the great potential to identify enhancer-promoter interactions systematically throughout the genome at high-resolution, but it is contingent upon choosing a factor of interest.

The combinations of these proximity-based ligation methods has rapidly expanded our understanding of higher-order chromatin structure throughout the genome. Most notably, these studies have shed light on the mechanism of activation of promoters by distal enhancers through "DNA looping." Initial observations about the mechanism of enhancer function suggested that actual physical linkage between an enhancer and its target promoter was essential for activation (49, 50). This implied that some degree of molecular communication in *cis* between the enhancer and the promoter was required for function. While, numerous models have been proposed over the years to explain how enhancers can activate gene expression from a distance (51, 52), proximity based ligation methods have contributed to a favoring of model of "looping" between the enhancer and promoter.

The "looping" model proposes that when an enhancer activates its target promoter, the two are brought in close physical proximity by forming a "loop" of the intervening DNA. This allows for the direct physical interaction between factors bound at the enhancer and promoter. The concept of a "looping" model stems from model of transcription activation by recruitment (52-54). The general understanding of activation by recruitment is that promoter proximal DNA-binding factors recruit both transcription

activators and co-activators that facilitate recruitment of a stable RNA-polymerase II holoenzyme to the promoter of a gene leading to its transcription. This process occurs by direct protein-protein interactions between the transcription activators and co-activators and RNA-polymerase II. The "looping" model is a logical extension of this concept, supposing that distally located DNA binding factors should also utilize direct physical interactions to facilitate RNA-polymerase II recruitment and activation (52).

Evidence for DNA looping between enhancer and promoters was provided using early versions of proximity based ligation assays (55), but the advent of 3C has contributed to a wealth of support for the looping model. Experiments focusing on several model genes, most notably the β-globin locus, have been particularly fruitful in our understanding of DNA-looping and enhancer promoter interactions and deserve special mention (Figure 3). The β-globin locus consists of five linearly organized, developmentally regulated globin genes that require a distally located locus control region (LCR) for their full activation (56)) (Figure 3a). Early 3C studies showed interactions between the LCR and the globin genes in erythroid cells but not in tissues that lack globin gene expression (57, 58), thus providing some of the most conclusive evidence of enhancer to promoter DNA looping. Subsequent studies have shown that the LCR-globin gene interactions form of a complex set of interactions between the LCR, the globin genes, and other DNaseI hypersensitive sites that has been termed an "active chromatin hub" (Figure 3b) (45, 57-60). Interactions appear to be developmentally regulated (38, 59), thus implicating enhancer to promoter interactions in the regulation of lineage and cell type specific gene activation. Recent experiments where a potential looping factor, Ldb1, is artificially tethered to the promoter of globin genes, forcing

looping and subsequent gene expression, provide some of the strongest evidence to date that looping plays a direct role in transcription activation at the β-globin locus (61).

A critical unresolved question in the field is what factors contribute to DNA-looping and enhancer-promoter interactions? As would be expected by the concept of transcription activation by recruitment, various DNA-binding transcription factors appear to be important for enhancer-promoter DNA looping. Studies of the β-globin locus have revealed that looping between the LCR and the globin genes requires lineage specific DNA-binding transcription factors such as GATA-1 and KLF1/EKLF (62, 63). Additionally, the ubiquitious zinc-finger protein CTCF has also been implicated in regulating interaction at the β-globin locus (60), suggesting that a multitude of factors may contribute to DNA-looping in this region (Figure 3c).

DNA-looping may also have a general role in higher-order organization of the genome. Numerous studies have pointed to the zinc-finger protein CTCF as being central to this organization. CTCF has been shown to be critical for looping interactions at a variety of loci, including the imprinted IGF2/H19 locus and the β-globin locus (60, 64, 65). Interestingly, CTCF extensively co-binds in the genome with the Cohesin complex (66). Cohesin has long been known to be a factor that mediates sister-chromatid cohesion, but recent evidence has also indicated that Cohesin, possibly in conjunction with CTCF, may mediate long-range interaction in the genome and contribute to enhancer-promoter interactions (66-69). CTCF is also prominently known as an insulating factor in the genome. CTCF is required for the activity of the chicken globin insulator (70) and is known to bind to nearly all known mammalian insulators. CTCF's insulating activity and its role in organizing higher-order chromatin structure are likely

intertwined. While these numerous studies have shed light on some of the factors responsible for contributing to DNA-looping and higher order chromatin interactions, these studies typically focused on one or several loci. Therefore, an understanding of how these factors contribute to chromatin structure on a genome-wide level remains unclear.

It is clear that while much is known about the higher order structure of chromatin and folding of chromosomes in the nucleus, there are still substantial gaps that remain. At a large scale in the nucleus, it is clear that chromosomes fold into chromosome territories. Active region of the genome are typically early replicating, gene rich and located in the center of the nucleus, while inactive regions are gene poor, late replicating, and associated with the nuclear lamina. Theoretical polymer-physics based models have indicated that there may be higher order domains in the genome, but the location of and mechanisms that contribute to the formation of these structures is unclear. On a local level, higher order chromatin structure has a major impact on regulation of gene expression through enhancer-promoter looping and chromatin insulation. However, the structural basis for these interactions on a genome wide scale has been lacking, as is an understanding the role of the underlying chromatin state in their formation.

Here I present the work performed by myself and my collaborators during my Ph.D. I have been using the Hi-C technique to characterize in the highest resolution to date the higher order chromatin interaction patterns throughout mammalian genomes. I will present evidence of structures that we term “topological domains” that appear to be stable self-interacting regions of chromatin that may form a basic unit of chromosome

organization. Topological domains are stable between cell types, conserved in evolution, and may have a role in regulating enhancer-promoter communication. I will also present evidence about the role for the factors CTCF and Cohesin in the formation of topological domains. We have noted that the boundaries between topological domains appear to be enriched for binding of the insulator protein CTCF as well as for members of the Cohesin complex. Working with my collaborators, I have shown that CTCF is important for the spatial segregation of topological domains from each other, while CTCF and Cohesin appear to regulate the self-association of topological domains, albeit at different spatial scales. Lastly, I will present data on the alterations in chromatin structure that occur during differentiation of human embryonic stem cells into a variety of lineages. I have observed that the topological domains are stable across diverse lineages. However, the association between domains appears to be cell-type specific, and this inter-domain association appears to modestly correlate with alterations in gene expression. Furthermore, I will present evidence for the role of histone modifications, chromatin state, and DNaseI Hypersensitivity in shaping the higher order landscape of the human genome. With the implication of alterations in genomic structure have a causal role in cancers, as well as the growing evidence that disease associated genetic variants are often found in distal regulatory elements, my hope is that by better understanding higher-order chromatin structures in the nucleus, we can gain a better understanding of how genomic structure ultimately contributes to development and disease.

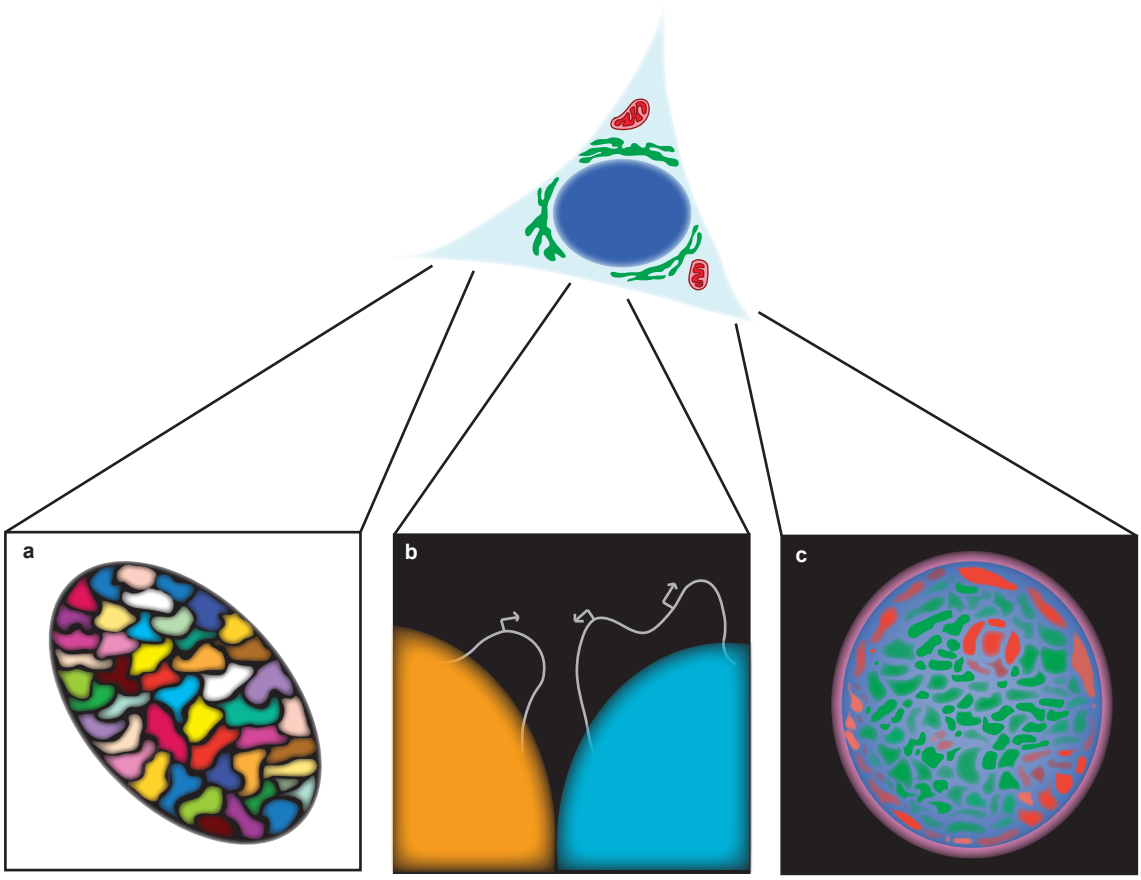


Figure 1. Global features of nuclear organization. a, Organization of chromosomes within the nucleus into chromosome territories. Each pair of homologous chromosomes is represented by a distinct color. Chromosomes separate into unique territories with little overlap between them. Smaller, more gene rich chromosomes are typically found closer to the nuclear interior, while larger, gene poor chromosomes are found closer to the nuclear periphery. b, Diagram of "looping out." Two chromosomes territories (in orange and blue) are shown along with two stretches of the chromosomes that have "looped out." Looping out is more common for gene rich regions of the chromosome. c, Diagram demonstrating how the nucleus is organized into large scale compartments. Late replicating chromatin, shown in red, is enriched at the nuclear periphery and in pericentromeric heterochromatin, while early replicating chromatin is located in the nuclear interior. Late replicating chromatin is also associated with the nuclear lamina (shown as a purple ring at the periphery of the nucleus) ultimately forming so-called "Lamina associated Domains (LADs)."

Figure 2. Outline of Proximity Based Ligation Methods. a, All proximity based ligation methods begin with fixation of nuclei to preserve both protein-protein and DNA-protein interactions of regions of the genome that were located in close spatial proximity. b, In 3C, 4C, 5C, and Hi-C, cross linked chromatin is digested with a restriction enzyme of choice. This yields sticky ends of DNA from fragments of the genome. c, In a 3C experiment, sticky ends of DNA are ligated together and ligation junctions are assessed typically using quantitative PCR. d, In a 4C experiment, after ligation of the sticky ended fragments and isolation of DNA, a second restriction enzyme is performed and the DNA is circularized by a second round of ligation. Inverse PCR is performed and interacting regions are detected either by microarray or by high-throughput sequencing. e, In a 5C experiment, after ligation of sticky ended fragments, primers that tile a given region are ligated using the genomic library as a template. Ligation products are then amplified using PCR and subsequently detected by microarray or sequencing. f, In a Hi-C experiment, prior to ligation, sticky ended DNA fragments are filled in with nucleotides, one of which is covalently linked to biotin. A blunt end ligation is then performed, and ligated DNA molecules are isolated using strepatavidin coated beads, PCR amplified, and sequenced. g, In a ChIA-pet experiment, no initial restriction enzyme digestion is performed. Instead, cross-linked chromatin is immunoprecipitated as in a ChIP experiment. DNA ends are repaired and a biotinylated linker is added. DNA ends are then ligated together and ligation products are isolated using streptavidin coated beads, PCR amplified, and sequenced.

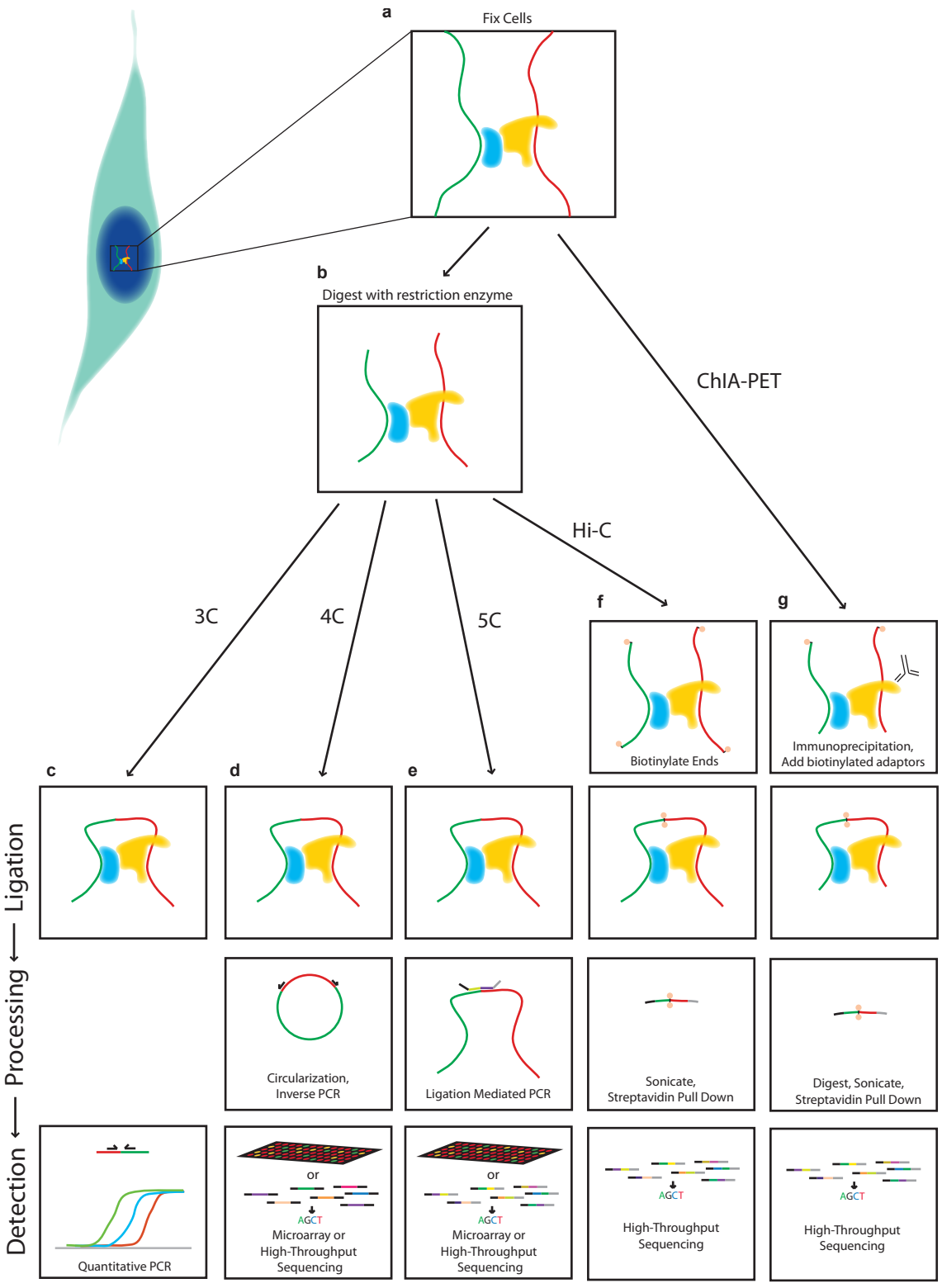
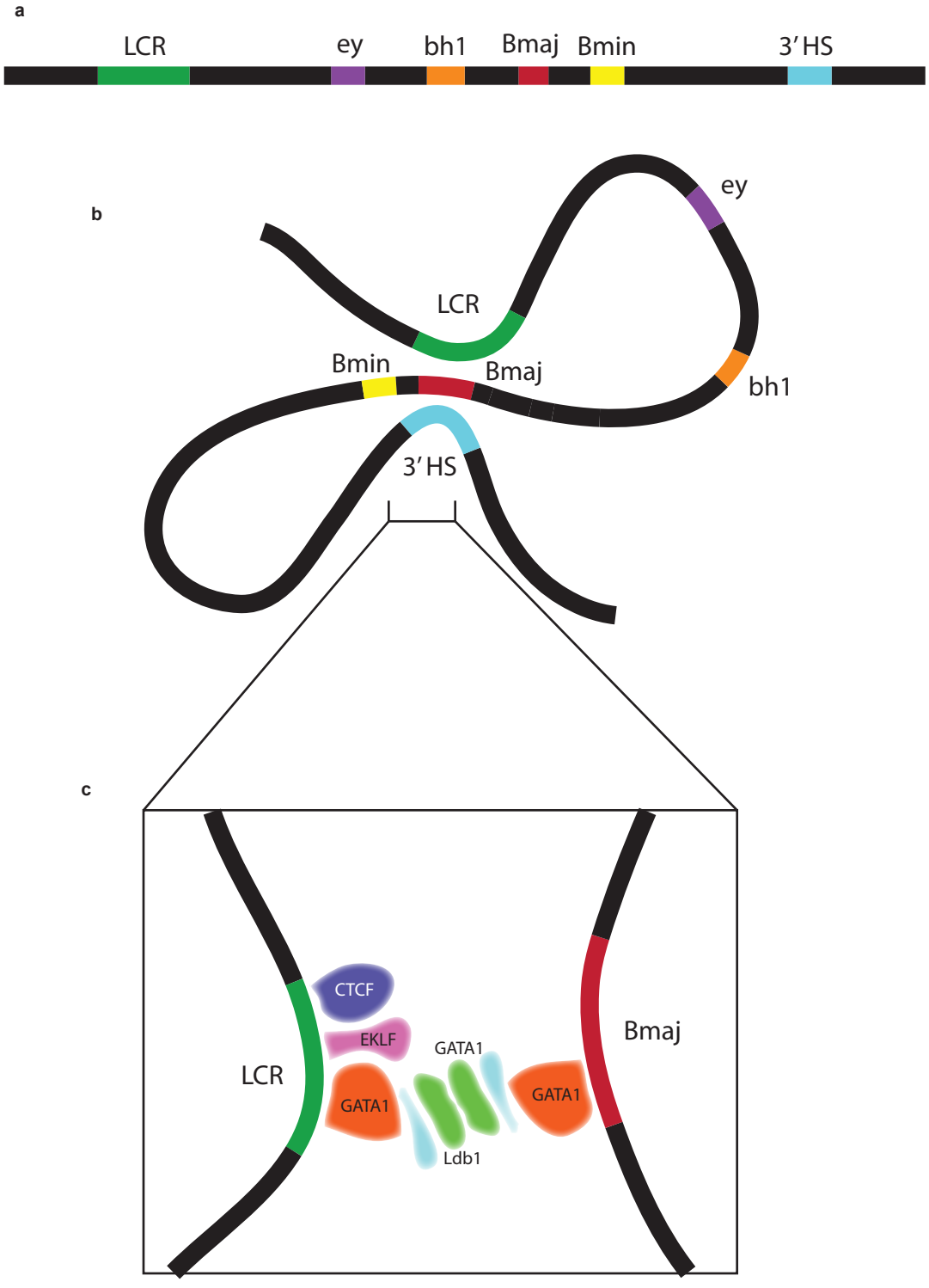

a
Fix Cells
b
Digest with restriction enzyme
ChIA-PET
Hi-C
3C
4C
5C
f
g
Biotinylate Ends
Immunoprecipitation,
Add biotinylated adaptors
c
d
e
Ligation
Processing
Detection
Circularization,
Inverse PCR
Ligation Mediated PCR
Sonicate,
Streptavidin Pull Down
Digest, Sonicate,
Streptavidin Pull Down
Quantitative PCR
or
AGCT
Microarray or
High-Throughput
Sequencing
or
AGCT
Microarray or
High-Throughput
Sequencing
AGCT
High-Throughput
Sequencing
AGCT
High-Throughput
Sequencing

Figure 3. Looping interaction at the β-globin locus. a, Linear structure of the globin locus. The 5' end of the locus contains a locus control region with numerous DNaseI Hypersensitive sites. Five globin genes whose expression is developmentally regulated are linearly organized at the locus followed by a 3` DNaseI hypersensitive site (only 4 globin genes are shown here). b, Diagram showing the configuration of the β-globin locus as an "active chromatin hub" with the LCR and 3' HS sites being brought in close proximity with the β-major and β-minor genes. c, Diagram of factors localized the LCR and the β-major genes. GATA1, ELKF, and CTCF have been shown to be important DNA-binding factors that contribute to looping formation, while LDB1 has been shown to be important for looping and activation of gene expression.


a
LCR
ey
bh1
Bmaj
Bmin
3′HS
b
ey
LCR
Bmin
Bmaj
bh1
3′HS
c
CTCF
EKLF
GATA1
GATA1
GATA1
LCR
Bmaj
Ldb1

References

1. Heitz E. Das Heterochromatin der Moose: I. Jarhbucher fur wissenschaftliche Botanik. 1928;69:762-818.

2. Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell.6(5):479-91. PMCID: 2867844.

3. Solovei I, Kreysing M, Lanctot C, Kosem S, Peichl L, Cremer T, et al. Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. Cell. 2009;137(2):356-68.

4. Stack SM, Brown DB, Dewey WC. Visualization of interphase chromosomes. J Cell Sci. 1977;26:281-99.

5. Zorn C, Cremer C, Cremer T, Zimmer J. Unscheduled DNA synthesis after partial UV irradiation of the cell nucleus. Distribution in interphase and metaphase. Exp Cell Res. 1979;124(1):111-9.

6. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289-93.

7. Sun HB, Shen J, Yokota H. Size-dependent positioning of human chromosomes in interphase nuclei. Biophys J. 2000;79(1):184-90. PMCID: 1300924.

8. Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis JA, Bickmore WA. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Hum Mol Genet. 2001;10(3):211-9.

9. Chambeyron S, Bickmore WA. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. Genes Dev. 2004;18(10):1119-30. PMCID: 415637.

10. Volpi EV, Chevret E, Jones T, Vatcheva R, Williamson J, Beck S, et al. Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. J Cell Sci. 2000;113 (Pt 9):1565-76.

11. Mahy NL, Perry PE, Bickmore WA. Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. J Cell Biol. 2002;159(5):753-63. PMCID: 2173389.

12. Mahy NL, Perry PE, Gilchrist S, Baldock RA, Bickmore WA. Spatial organization of active and inactive genes and noncoding DNA within chromosome territories. J Cell Biol. 2002;157(4):579-89. PMCID: 2173868.

13. Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet. 2004;36(10):1065-71.

14. Boyle S, Rodesch MJ, Halvensleben HA, Jeddeloh JA, Bickmore WA. Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosome Res.19(7):901-9. PMCID: 3210351.

15. van den Engh G, Sachs R, Trask BJ. Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. Science. 1992;257(5075):1410-2.

16. Sachs RK, van den Engh G, Trask B, Yokota H, Hearst JE. A random-walk/giant-loop model for interphase chromosomes. Proc Natl Acad Sci U S A. 1995;92(7):2710-4. PMCID: 42288.

17. Yokota H, van den Engh G, Hearst JE, Sachs RK, Trask BJ. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. J Cell Biol. 1995;130(6):1239-49. PMCID: 2120584.

18. Munkel C, Langowski J. Chromosome structure predicted by a polymer model. Physical Review E. 1998;57(5):5888-96.

19. Munkel C, Eils R, Dietzel S, Zink D, Mehring C, Wedemann G, et al. Compartmentalization of interphase chromosomes observed in simulation and experiment. J Mol Biol. 1999;285(3):1053-65.

20. Jhunjhunwala S, van Zelm MC, Peak MM, Cutchin S, Riblet R, van Dongen JJ, et al. The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. Cell. 2008;133(2):265-79. PMCID: 2771211.

21. Mateos-Langerak J, Bohn M, de Leeuw W, Giromus O, Manders EM, Verschure PJ, et al. Spatially confined folding of chromatin in the interphase nucleus. Proc Natl Acad Sci U S A. 2009;106(10):3812-7. PMCID: 2656162.

22. Bohn M, Heermann DW, van Driel R. Random loop model for long polymers. Phys Rev E Stat Nonlin Soft Matter Phys. 2007;76(5 Pt 1):051805.

23. Pickersgill H, Kalverda B, de Wit E, Talhout W, Fornerod M, van Steensel B. Characterization of the Drosophila melanogaster genome at the nuclear lamina. Nat Genet. 2006;38(9):1005-14.

24. Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SW, Solovei I, Brugman W, et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol Cell.38(4):603-13.

25. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008;453(7197):948-51.

26. Meuleman W, Peric-Hupkes D, Kind J, Beaudry JB, Pagie L, Kellis M, et al. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res.23(2):270-80. PMCID: 3561868.

27. Kim Y, Sharov AA, McDole K, Cheng M, Hao H, Fan CM, et al. Mouse B-type lamins are required for proper organogenesis but not by embryonic stem cells. Science.334(6063):1706-10. PMCID: 3306219.

28. Ma H, Samarabandu J, Devdhar RS, Acharya R, Cheng PC, Meng C, et al. Spatial and temporal dynamics of DNA replication sites in mammalian cells. J Cell Biol. 1998;143(6):1415-25. PMCID: 2132991.

29. Jackson DA, Pombo A. Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. J Cell Biol. 1998;140(6):1285-95. PMCID: 2132671.

30. O'Keefe RT, Henderson SC, Spector DL. Dynamic organization of DNA replication in mammalian cell nuclei: spatially and temporally defined replication of chromosome-specific alpha-satellite DNA sequences. J Cell Biol. 1992;116(5):1095-110. PMCID: 2289349.

31. Dimitrova DS, Gilbert DM. The spatial position and replication timing of chromosomal domains are both established in early G1 phase. Mol Cell. 1999;4(6):983-93.

32. Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, Chang CW, et al. Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol. 2008;6(10):e245. PMCID: 2561079.

33. Hiratani I, Ryba T, Itoh M, Rathjen J, Kulik M, Papp B, et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res.20(2):155-69. PMCID: 2813472.

34. Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res.20(6):761-70. PMCID: 2877573.

35. Bernardi G. The neoselectionist theory of genome evolution. Proc Natl Acad Sci U S A. 2007;104(20):8385-90. PMCID: 1866311.

36. Eyre-Walker A, Hurst LD. The evolution of isochores. Nat Rev Genet. 2001;2(7):549-55.

37. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295(5558):1306-11.

38. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet. 2006;38(11):1348-54.

39. Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet. 2006;38(11):1341-7.

40. Wurtele H, Chartrand P. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. Chromosome Res. 2006;14(5):477-95.

41. Ling JQ, Li T, Hu JF, Vu TH, Chen HL, Qiu XW, et al. CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. Science. 2006;312(5771):269-72.

42. Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJ, Zhu Y, et al. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes Dev.25(13):1371-83. PMCID: 3134081.

43. Stadhouders R, Kolovos P, Brouwer R, Zuin J, van den Heuvel A, Kockx C, et al. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. Nat Protoc.8(3):509-24.

44. van de Werken HJ, Landan G, Holwerda SJ, Hoichman M, Klous P, Chachik R, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods.9(10):969-72.

45. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for

mapping interactions between genomic elements. Genome Res. 2006;16(10):1299-309. PMCID: 1581439.

46. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet.43(11):1059-65.

47. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, et al. A three-dimensional model of the yeast genome. Nature.465(7296):363-7. PMCID: 2874121.

48. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009;462(7269):58-64. PMCID: 2774924.

49. Dunaway M, Droge P. Transactivation of the Xenopus rRNA gene promoter by its enhancer. Nature. 1989;341(6243):657-9.

50. Mueller-Storm HP, Sogo JM, Schaffner W. An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge. Cell. 1989;58(4):767-77.

51. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. Science. 1998;281(5373):60-3.

52. Bulger M, Groudine M. Looping versus linking: toward a model for long-distance gene activation. Genes Dev. 1999;13(19):2465-77.

53. Ptashne M, Gann A. Transcriptional activation by recruitment. Nature. 1997;386(6625):569-77.

54. Ptashne M. Gene regulation by proteins acting nearby and at a distance. Nature. 1986;322(6081):697-701.

55. Cullen KE, Kladde MP, Seyfred MA. Interaction between transcription regulatory regions of prolactin chromatin. Science. 1993;261(5118):203-6.

56. Engel JD, Tanimoto K. Looping, linking, and chromatin activity: new insights into beta-globin locus regulation. Cell. 2000;100(5):499-502.

57. Carter D, Chakalova L, Osborne CS, Dai YF, Fraser P. Long-range chromatin regulatory interactions in vivo. Nat Genet. 2002;32(4):623-6.

58. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell. 2002;10(6):1453-65.

59. Palstra RJ, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W. The beta-globin nuclear compartment in development and erythroid differentiation. Nat Genet. 2003;35(2):190-4.

60. Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. 2006;20(17):2349-54. PMCID: 1560409.

61. Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell.149(6):1233-44. PMCID: 3372860.

62. Drissen R, Palstra RJ, Gillemans N, Splinter E, Grosveld F, Philipsen S, et al. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. Genes Dev. 2004;18(20):2485-90. PMCID: 529536.

63. Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, et al. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Mol Cell. 2005;17(3):453-62.

64. Kurukuti S, Tiwari VK, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, et al. CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. Proc Natl Acad Sci U S A. 2006;103(28):10684-9. PMCID: 1484419.

65. Murrell A, Heeson S, Reik W. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. Nat Genet. 2004;36(8):889-93.

66. Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, et al. Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell. 2008;132(3):422-33.

67. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature. 2008;451(7180):796-801.

68. Hadjur S, Williams LM, Ryan NK, Cobb BS, Sexton T, Fraser P, et al. Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature. 2009;460(7253):410-3. PMCID: 2869028.

69. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature.467(7314):430-5. PMCID: 2953795.

70. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell. 1999;98(3):387-96.

Chapter 2

Reprint of "Topological domains in mammalian genomes identified by analysis of chromatin interactions." Published 17, May 2012 in *Nature*.

Topological domains in mammalian genomes identified by analysis of chromatin interactions

Jesse R. Dixon[1,2,3], Siddarth Selvaraj[1,4], Feng Yue[1], Audrey Kim[1], Yan Li[1], Yin Shen[1], Ming Hu[5], Jun S. Liu[5] & Bing Ren[1,6]

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, and numerous models have been proposed for how chromosomes fold within chromosome territories[1]. 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 advances in the study of three-dimensional genome organization. In particular, Hi-C has been introduced as a method for identifying higher order chromatin interactions genome wide[2]. Here we investigate the three-dimensional 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, indicating that topological domains are an inherent property of mammalian genomes. Finally, we find that the boundaries of topological domains are enriched for the insulator binding protein CTCF, housekeeping genes, transfer RNAs and short interspersed element (SINE) retrotransposons, indicating that these factors may have a role in establishing the topological domain structure of the genome.

To study chromatin structure in mammalian cells, we determined genome-wide chromatin interaction frequencies by performing the Hi-C experiment[2] in mouse embryonic stem (ES) cells, human ES cells, and human IMR90 fibroblasts. Together with Hi-C data for the mouse cortex generated in a separate study (Y. Shen *et al.*, manuscript in preparation), we analysed over 1.7-billion read pairs of Hi-C data corresponding to pluripotent and differentiated cells (Supplementary Table 1). We normalized the Hi-C interactions for biases in the data (Supplementary Figs 1 and 2)[3]. To validate the quality of our Hi-C data, we compared the data with previous chromosome conformation capture (3C), chromosome conformation capture carbon copy (5C), and fluorescence *in situ* hybridization (FISH) results[4–6]. Our IMR90 Hi-C data show a high degree of similarity when compared to a previously generated 5C data set from lung fibroblasts (Supplementary Fig. 4). In addition, our mouse ES cell Hi-C data correctly recovered a previously described cell-type-specific interaction at the *Phc1* gene[5] (Supplementary Fig. 5). Furthermore, the Hi-C interaction frequencies in mouse ES cells are well-correlated with the mean spatial distance separating six loci as measured by two-dimensional FISH[6] (Supplementary Fig. 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 are of high quality and accurately capture the higher order chromatin structures in mammalian cells.

We next visualized two-dimensional interaction matrices using a variety of bin sizes to identify interaction patterns revealed as a result of our high sequencing depth (Supplementary Fig. 7). We noticed that at bin sizes less than 100 kilobases (kb), highly self-interacting regions begin to emerge (Fig. 1a and Supplementary Fig. 7, seen as 'triangles' on the heat map). 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 identify systematically all such topological domains in the genome, we devised a simple statistic termed the directionality index to quantify the degree of upstream or downstream interaction bias for a genomic region, which varies considerably at the periphery of the topological domains (Fig. 1b; see Supplementary Methods for details). The directionality index was reproducible (Supplementary Table 2) and pervasive, with 52% of the genome having a directionality index that was not expected by random chance (Fig. 1c, false discovery rate = 1%). We then used a Hidden Markov model (HMM) based on the directionality index to identify biased 'states' and therefore infer the locations of topological domains in the genome (Fig. 1a; see Supplementary Methods for details). The domains defined by HMM were reproducible between replicates (Supplementary Fig. 8). Therefore, we combined the data from the HindIII replicates and identified 2,200 topological domains in mouse ES cells with a median size of 880 kb that occupy ~91% of the genome (Supplementary Fig. 9). As expected, the frequency of intra-domain interactions is higher than inter-domain interactions (Fig. 1d, e). Similarly, FISH probes[6] in the same topological domain (Fig. 1f) are closer in nuclear space than probes in different topological domains (Fig. 1g), despite similar genomic distances between probe pairs (Fig. 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 (Supplementary Fig. 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[4,7,8], which we observed corresponds to a topological domain boundary in both mouse (Fig. 1a) and human (Fig. 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 (refs 9–11). We see a strong enrichment of CTCF at the topological boundary regions (Fig. 2b and Supplementary Fig. 10), indicating that topological boundary regions

[1]Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, California 92093, USA. [2]Medical Scientist Training Program, University of California, San Diego, La Jolla, California 92093, USA. [3]Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, California 92093, USA. [4]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, California 92093, USA. [5]Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02138, USA. [6]University of California, San Diego School of Medicine, Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, UCSD Moores Cancer Center, 9500 Gilman Drive, La Jolla, California 92093, USA.

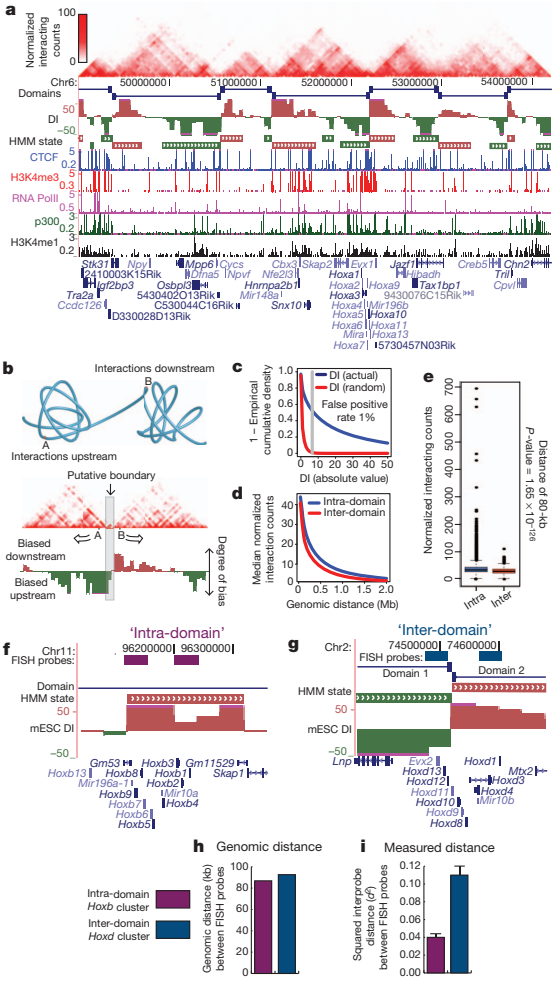


Figure 1 | Topological domains in the mouse ES cell genome. **a**, Normalized Hi-C interaction frequencies displayed as a two-dimensional heat map overlayed on ChIP-seq data (from Y. Shen *et al.*, manuscript in preparation), directionality index (DI), HMM bias state calls, and domains. For both directionality index 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 directionality index (absolute value, in blue) compared to random (red). **d**, Mean interaction frequencies at all genomic distances between 40 kb to 2 Mb. Above 40 kb, the intra- versus inter-domain interaction frequencies are significantly different ($P < 0.005$, Wilcoxon test). **e**, Box plot of all interaction frequencies at 80-kb 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 inter-probe distance (from ref. 6) FISH probe pairs. mESC, mouse ES cell. Error bars indicate standard error ($n = 100$ for each probe pair).

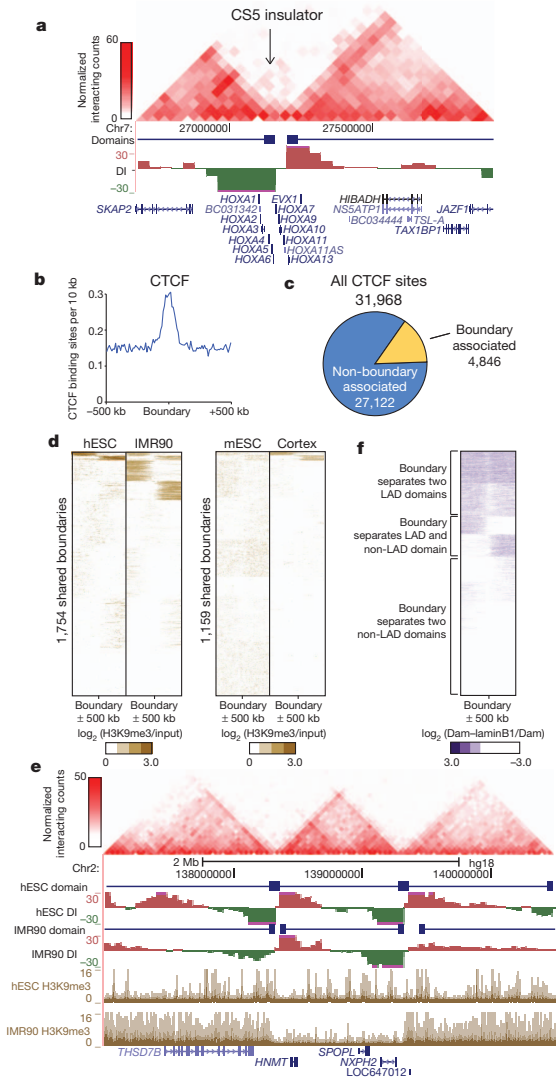


Figure 2 | Topological boundaries demonstrate classical insulator or barrier element features. **a**, Two-dimensional heat map surrounding the *Hoxa* locus and CS5 insulator in IMR90 cells. **b**, Enrichment of CTCF at boundary regions. **c**, The portion of CTCF binding sites that are considered 'associated' with a boundary (within ±20-kb window is used as the expected uncertainty due to 40-kb 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 (hESC) and IMR90 cells. The two-dimensional heat map shows the interaction frequency in human ES cells. **f**, Heat map of LADs (from ref. 14) surrounding the boundary regions. Scale is the $\log_2$ ratio of DNA adenosine methylation (Dam)–lamin B1 fusion over Dam alone (Dam–laminB1/Dam).

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[12,13]. Indeed, we observe a clear segregation of H3K9me3 at the boundary regions that occurs predominately in differentiated cells (Fig. 2d, e and Supplementary Fig. 11). As the boundaries that we analysed in Fig. 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 seem 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 ref. 2, with lamina-associated domains (LADs)[10,14], replication time zones[15,16], and large organized chromatin K9 modification (LOCK) domains[17]. In all cases, we can see that topological domains are related to, but independent from, each of these previously described domain-like structures (Supplementary Figs 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 (Fig. 2f and Supplementary Fig. 12), the A and B compartments (Supplementary Fig. 13, 14), and early and late replicating chromatin (Supplementary Fig. 14). Lastly, we can also confirm the previously reported similarities between the A and B compartments and early and late replication time zone (Supplementary Fig. 16)[16].

We next compared the locations of topological boundaries identified in both replicates of mouse ES cells and cortex, or between both replicates of human ES cells and IMR90 cells. In both human and mouse, most of the boundary regions are shared between cell types (Fig. 3a and Supplementary Fig. 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 directionality index is still readily apparent and highly reproducible between replicates (Supplementary Fig. 17b, c). We cannot determine if the differences in domain calls between cell types is due to noise in the data or to biological phenomena, such as a change in the strength of the boundary region between cell types[18]. Regardless, these results indicate 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 (Supplementary Fig. 18).

The stability of the domains between cell types is surprising given previous evidence showing cell-type-specific chromatin interactions and conformations[5,7]. 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 20-kb binning using a binomial test with an empirical false discovery rate of $<1\%$ based on random permutation of the replicate data. These dynamic interacting regions are enriched for differentially expressed genes (Fig. 3b–d, Supplementary Fig. 19 and Supplementary Table 5). In fact, 20% of all genes that undergo a four-fold change in gene expression are found at dynamic interacting loci. This is probably an underestimate, because by binning the genome at 20 kb, any dynamic regulatory interaction less than 20 kb will be missed. Lastly, $>96\%$ of dynamic interacting regions occur in the same domain (Fig. 3e). Therefore, we favour 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. Most of the 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 $<2.2\times10^{-16}$, Fisher's exact test; Fig. 3f). The syntenic regions in mouse and human in particular share a high degree of similarity in their higher order chromatin structure (Fig. 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. Although most topological boundaries are enriched for the binding of CTCF, only 15% of CTCF

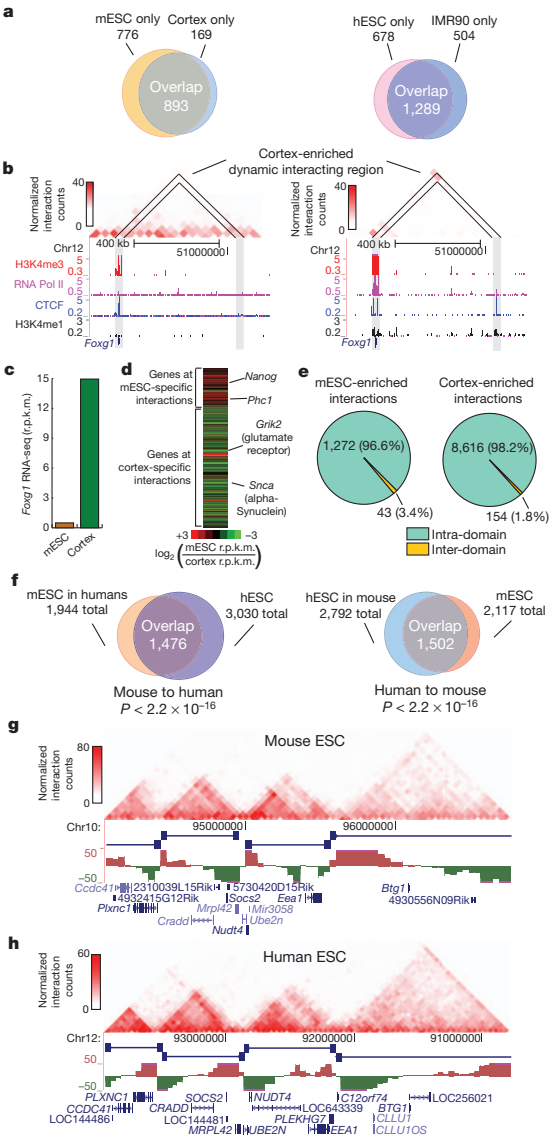


Figure 3 | Boundaries are shared across cell types and conserved in evolution. **a**, Overlap of boundaries between cell types. **b**, Genome browser shot of a cortex enriched dynamic interacting region that overlaps with the *Foxg1* gene. **c**, *Foxg1* expression in reads per kilobase per million reads sequenced (r.p.k.m.) 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 ($P<2.2\times10^{-16}$ compared to random, Fisher's exact test). **g**, **h**, Genome browser shots showing domain structure over a syntenic region in the mouse (**g**) and human (**h**) ES cells. Note: the region in humans has been inverted from its normal UCSC coordinates for proper display purposes.

binding sites are located within boundary regions (Fig. 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 and transcription factors around topological boundary regions in mouse ES cells, we observed that factors associated with active promoters and gene bodies are enriched at boundaries in both mouse and humans (Fig. 4a and Supplementary Figs 20–23)[19,20]. In contrast, non-promoter-associated marks, such as H3K4me1 (associated with enhancers) and H3K9me3, were not enriched or were specifically depleted at boundary regions (Fig. 4a). Furthermore, transcription start sites (TSS) and global run on sequencing (GRO-seq)[21] signal were also enriched around topological boundaries (Fig. 4a). We found that housekeeping genes were particularly strongly enriched near topological boundary regions (Fig. 4b–d; see Supplementary Table 7 for complete GO terms enrichment). Additionally, the tRNA genes, which have the potential to function as boundary elements[22,23], are also enriched at boundaries (P value <0.05, Fisher's exact test; Fig. 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 (Supplementary Fig. 24). Indeed, boundaries associated with both CTCF and a housekeeping gene account for nearly one-third of all topological boundaries in the genome (Fig. 4e and Supplementary Fig. 24).

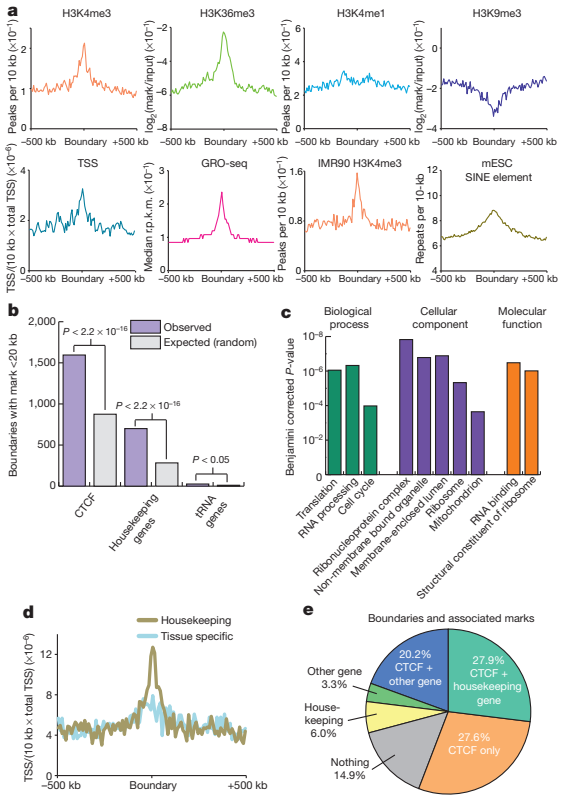


Figure 4 | Boundary regions are enriched for housekeeping genes. **a**, Chromatin modifications, TSS, GRO-seq and SINE elements surrounding boundary regions in mouse ES cells or IMR90 cells. **b**, Boundaries associated with a CTCF binding site, housekeeping gene, or tRNA gene (purple) compared to expected at random (grey). **c**, Gene Ontology P-value chart. **d**, Enrichment of housekeeping genes (gold) and tissue-specific genes (blue) as defined by Shannon entropy scores near boundaries normalized for the number of genes in each class (TSS/10 kb/total TSS). **e**, Percentage of boundaries with a given mark within 20 kb of the boundaries.

Finally, we analysed the enrichment of repeat classes around boundary elements. We observed that Alu/B1 and B2 SINE retrotransposons in mouse and Alu SINE elements in humans are enriched at boundary regions (Fig. 4a and Supplementary Figs 24 and 25). In light of recent reports indicating that a SINE B2 element functions as a boundary in mice[24], and SINE element retrotransposition may alter CTCF binding sites during evolution[25], we believe that this contributes to a growing body of evidence indicating 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[1,26,27]. Such spatial organization seems 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 and SINE elements. The association of housekeeping genes with boundary regions extends previous studies in yeast and insects and suggests that non-CTCF factors may also be involved in insulator/barrier functions in mammalian cells[28].

The topological domains we identified are well conserved between mice and humans. This indicates 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* embryos[29] and in high-resolution studies of the X-inactivation centre in mice (termed topologically associated domains or TADs)[30], indicating that topological domains may be a fundamental organizing principle of metazoan genomes.

METHODS SUMMARY

Cell culture and Hi-C experiments. J1 mouse ES cells were grown on gamma-irradiated mouse embryonic fibroblasts cells under standard conditions (85% high glucose DMEM, 15% HyClone FBS, 0.1 mM non-essential amino acids, 0.1 mM β-mercaptoethanol, 1 mM glutamine, LIF 500 U ml^{-1}, 1× Gibco penicillin/streptomycin). Before collecting for Hi-C, J1 mouse ES cells were passaged onto feeder free 0.2% gelatin-coated plates for at least two passages to rid the culture of feeder cells. H1 human ES cells and IMR90 fibroblasts were grown as previously described[13]. Collecting the cells for Hi-C was performed as previously described, with the only modification being that the adherent cell cultures were dissociated with trypsin before fixation.

Sequencing and mapping of data. Hi-C analysis and paired-end libraries were prepared as previously described[2] 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. Two-dimensional heat maps were generated as previously described[2].

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 Supplementary Methods.

Received 26 September 2011; accepted 27 March 2012.
Published online 11 April 2012.

1. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* **2,** a003889 (2010).
2. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
3. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genet.* **43,** 1059–1065 (2011).
4. Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472,** 120–124 (2011).
5. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467,** 430–435 (2010).
6. Eskeland, R. *et al.* Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol. Cell* **38,** 452–464 (2010).
7. Noordermeer, D. *et al.* The dynamic architecture of Hox gene clusters. *Science* **334,** 222–225 (2011).

8. Kim, Y. J., Cecchini, K. R. & Kim, T. H. Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. *Proc. Natl Acad. Sci. USA* **108,** 7391–7396 (2011).
9. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137,** 1194–1211 (2009).
10. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).
11. Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature Genet.* **43,** 630–638 (2011).
12. Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148,** 816–831 (2012).
13. Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6,** 479–491 (2010).
14. Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* **38,** 603–613 (2010).
15. Hiratani, I. *et al.* Genome-wide dynamics of replication timing revealed by *in vitro* models of mouse embryogenesis. *Genome Res.* **20,** 155–169 (2010).
16. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20,** 761–770 (2010).
17. Wen, B., Wu, H., Shinkai, Y., Irizarry, R. A. & Feinberg, A. P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nature Genet.* **41,** 246–250 (2009).
18. Scott, K. C., Taubman, A. D. & Geyer, P. K. Enhancer blocking by the *Drosophila* gypsy insulator depends upon insulator anatomy and enhancer strength. *Genetics* **153,** 787–798 (1999).
19. Bilodeau, S., Kagey, M. H., Frampton, G. M., Rahl, P. B. & Young, R. A. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev.* **23,** 2484–2489 (2009).
20. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134,** 521–533 (2008).
21. Min, I. M. *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev.* **25,** 742–754 (2011).
22. Donze, D. & Kamakaka, R. T. RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in *Saccharomyces cerevisiae*. *EMBO J.* **20,** 520–531 (2001).
23. Ebersole, T. *et al.* tRNA genes protect a reporter gene from epigenetic silencing in mouse cells. *Cell Cycle* **10,** 2779–2791 (2011).
24. Lunyak, V. V. *et al.* Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* **317,** 248–251 (2007).
25. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell.* **148,** 335–348 (2012).
26. Jhunjhunwala, S. *et al.* The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell* **133,** 265–279 (2008).
27. Capelson, M. & Corces, V. G. Boundary elements and nuclear organization. *Biol. Cell* **96,** 617–629 (2004).
28. Amouyal, M. Gene insulation. Part I: natural strategies in yeast and *Drosophila*. *Biochem. Cell Biol.* **88,** 875–884 (2010).
29. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148,** 458–472 (2012).
30. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* http://dx.doi.org/10.1038/nature11049 (this issue).

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

Acknowledgements We are grateful for the comments from and discussions with Z. Qin, A. Desai and members of the Ren laboratory during the course of the study. We also thank W. Bickmore and R. 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). J.R.D. is funded by a pre-doctoral training grant from CIRM. Y.S. is supported by a postdoctoral fellowship from the Rett Syndrome Research Foundation.

Author Contributions J.R.D. and B.R. designed the studies. J.R.D., A.K., Y.L. and Y.S. conducted the Hi-C experiments; J.R.D., S.S. and F.Y. carried out the data analysis; J.S.L. and M.H. provided insight for analysis; F.Y. built the supporting website; J.R.D. and B.R. prepared the manuscript.

Author Information All Hi-C data described in this study have been deposited in the GEO under 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/. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.R. (biren@ucsd.edu).

SUPPLEMENTARY INFORMATION

doi:10.1038/nature11082

Supplemental Table of Contents

I. Public Datasets analyzed

II. Supplemental Methods

III. Supplemental References

IV. Supplemental Figures

V. Supplemental Tables

I. Public Datasets analyzed

Dataset	Figure	Accession	Reference
Lymphoblastoid Hi-C	Supplemental Figure 7	GSE18199	Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289-93 (2009).[31]
H3K4me3, H3K4me1, H3K27ac, p300, CTCF, ChIP-seq, mESC and cortex RNA-seq	Figures 1-4, Supplemental Figures 5,10,20-23		Shen, Y. et al. A Map of cis-Regulatory Sequences in the Mouse Genome. *in submission* (2012).[32]
Lung Fibroblast 5C	Supplemental Figure 4		Wang, K.C. et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472, 120-4.[33]
Med1, Med12, Smc1, Smc3,	Supplemental Figure 5, 20-22	GSE22557	Kagey, M.H. et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467, 430-5.[34]
mESC 2D-FISH	Figure 1, Supplemental Figure 6		Eskeland, R. et al. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol Cell* 38, 452-64.[35]
Cortex H3K9me3	Figure 2	GSE33722	Xie, W. et al. Base-resolution analysis of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* 148 (4), 816-831.[36]
IMR90 H3K4me3, hESC H3K9me3, IMR90 H3K9me3	Figure 2, 4	SRP000941	Hawkins, R.D. et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* 6, 479-91.[37]
mESC Lamina DAM-id	Figure 2, Supplemental Figure	GSE17051	Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell* 38, 603-13.[38]
mESC Replication Timing	Supplemental Figure 14, 16	GSE18019	Hiratani, I. et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res* 20, 155-69.[39]
H3K9me2 (LOCK) Domain ChIP-Chip	Supplemental Figure 15	GSE13445	Wen, B., Wu, H., Shinkai, Y., Irizarry, R.A. & Feinberg, A.P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat Genet* 41, 246-50 (2009).[40]
mESC H3K27me3, H4K20me3	Supplemental Figure 20-22	GSE12241	Mikkelsen, T.S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553-60 (2007).[41]

mESC H3K36me3, H3K79me2, Oct4, Sox2, Nanog	Figure 4, Supplemental Figure 20-22	GSE11724	Marson, A. et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 134, 521-33 (2008).[42]
mESC H3K9me3	Figure 2, 4	GSE18371	Bilodeau, S., Kagey, M.H., Frampton, G.M., Rahl, P.B. & Young, R.A. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev* 23, 2484-9 (2009).[43]
mESC Jarid2, Jarid1a, Suz12, Ezh2	Supplemental Figure 20-22	GSE18776	Peng, J.C. et al. Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* 139, 1290-302 (2009).[44]
mESC PolII Serine 5, PolII Serine 2, NelfA, Ctr9, Spt5	Supplemental Figure 20-22	GSE20530	Rahl, P.B. et al. c-Myc regulates transcriptional pause release. *Cell* 141, 432-45.[45]
DNase I HS	Supplemental Figure 20-22		Schnetz, M.P. et al. CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression. *PLoS Genet* 6, e1001023.[46]
GRO-Seq	Figure 4	GSE27037	Min, I.M. et al. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* 25, 742-54.[47]
bioGPS database	Figure 4		Lattin, J.E. et al. Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Res* 4, 5 (2008).[48]

II. Supplemental Methods

Mapping

We mapped the data using BWA using default parameters. We consider only uniquely mapping reads (mapping quality > 10). We remove PCR duplicate reads using Picard (http://picard.sourceforge.net).

Interaction Matrices

The interaction matrices were calculated as previously described[31] at bin sizes ranging from 10kb to 1Mb.

Normalization

We normalized the Hi-C data as previously described by Yaffe and Tanay[49]. However, we did not perform linear weight smoothing and BFGS non-linear optimization

and the normalization is still effective at removing restriction enzyme bias (see Supplemental Figures 1 and 2).

Heat Maps and Visualization of Data

To visualize the high-resolution interaction data, we generated 2D heat-maps that were overlaid with publicly available ChIP-Seq data sets visualized in a genome browser (Figure 1a). Interaction frequencies were calculated as above. Interaction frequencies between any two loci can be visualized by identifying the point off the axis where diagonals originating from each locus intersect, in a manner similar to a linkage disequilibrium plot.

The heat maps in Supplementary Figure 4 are made differently. This is to correspond to the method used in (ref. 33) so we can accurately compare the interaction frequencies between our Hi-C data and the published 5C data from Wang et al. The interaction matrix is generated as follows. The 120kb HoxA locus is split into 30 segments using a 30kb sliding window with sliding in 3kb intervals. For each interaction between two 30kb windows *i* and *j*, we identify all possible HindIII cut sites in *i* and *j* and all possible HindIII cut sites interactions between these bins *i* and *j*. The interaction score between two segments of the heatmap is the mean frequency of interactions among all possible HindII cut site combinations between the two bins. The data for the Wang et al. 5C heatmaps was downloaded from the accompanying supplemental data[33].

Estimate of Intermolecular Ligation Rates

We estimated the intermolecular ligation rate between any two loci in the genome by analyzing the number of reads that map from a nuclear chromosome (chr(N)) to the

mitochondrial chromosome (chrM). As random intermolecular interactions will depend on the concentration of molecules in solution, the number of random interactions between the nuclear and mitochondrial chromosomes should be proportional to the amount of nuclear and mitochondrial DNA in solution during the ligation step of the protocol. As the number of mitochondria can vary between cell types, we use an estimated number of mitochondria of 40 based on previous experiments in the literature to test the number of mitochondria in mouse ES cells[50]. The total amount of "interacting space" between the mitochondrial genome and the nuclear genome is the product of the amount of mitochondrial DNA in solution (roughly 16kb/mitochondria * 40 mitochondria/cell) and the size of DNA in solution (roughly 5.1 Gigabases per diploid nucleus). By dividing the total number of chrM to chr(N) reads by this "interacting space," we can get an estimate of the number of reads/kbp^2 for any interaction in the genome. Our estimate suggest that for any two 40kb bins, there would be on average 0.015 reads per bin due to intermolecular ligations in the mouse ES cell HindIII original library and 0.079 reads /40kb interaction in the mouse ES cell replicate library. This is detailed in Supplemental Figure 27.

We would note that there are two potential pitfalls of this method. First, this requires an estimate of the number of mitochondria in a given cell type, which may not be available for any particular cell type of interest and can potentially vary by orders of magnitude. A second potential pitfall is that for the NcoI restriction enzyme, there are no mappable NcoI cut sites in the mitochondrial chromosome. Therefore, this method of analysis is not amenable to all restriction enzymes that could be used in a Hi-C experiment.

Correlation Between Experiments

We calculate the correlation between two experiments as follows: The set of all possible interactions I_{ij} for two experiments *A* and *B* were correlated by comparing each point in interaction matrix I_A from experiment *A* with the same point I_B from experiment *B*. Because the interaction matrix is highly skewed towards proximal interactions, we restricted the correlation to a maximum distance between points *i* and *j* of 50 bins. We use R to calculate the Pearson correlation between the two vectors of all point in I_A and I_B.

Directionality Index, Domain and Boundary Calling

We noted that the regions at the periphery of the topological domains are highly biased in their interaction frequencies. In other words, the most upstream portion of a topological domain is highly biased towards interacting downstream, and the downstream portion of a topological domain is highly biased towards interacting upstream. We reasoned that by identifying such biases in interaction frequency in the genome, we would be able to identify the locations of topological domains and boundaries in the genome.

To determine the directional bias at any given bin in the genome, we developed a Directionality Index (DI) to quantify the degree of upstream or downstream bias of a given bin. The directionality index is calculated in equation 1, where A is the number of reads that map from a given 40kb bin to the upstream 2Mb, B is the number of reads that

map from the same 40kb bin to the downstream 2Mb, and E, the expected number of reads under the null hypothesis, is equal to (A + B)/2.

Eq. 1

$$DI = \left(\frac{B - A}{|B - A|}\right)\left(\frac{(A - E)^2}{E} + \frac{(B - E)^2}{E}\right)$$

The directionality index is based on the chi-squared test statistic, where the null hypothesis is that each bin is equally likely to interact with the regions upstream and downstream of it. Bins that show a directional bias have a directionality index proportional to the degree of bias, with more biased bins having a higher magnitude of directionality index. We use a 40kb bin size and a 2Mb because these parameters maximize the reproducibility of the DI and the domain calls while retaining a sufficiently high resolution to identify domains and boundary regions.

To generate a random directionality index, we randomized the direction either upstream or downstream of every read pair that mapped to a given bin and calculated the directionality index with the randomized directions. Bins with large random directionality indexes are virtually absent by chance, with less than 1% of the absolute value of random DI being greater than 6.57.

We consider the directionality index as an observation and believe that the "true" hidden directionality bias (DB) can be determined using a hidden Markov model (HMM). The HMM assumes that the directionality index observations are following a mixture of Gaussians and then predicts the states as "Upstream Bias", "Downstream Bias" or "No

Bias" (See Supplementary Figure 28 for a mathematical representation of our Hidden Markov Model).

Describing the observed directionality index as Y's $[Y_1,Y_2..Y_n]$, the hidden true directionality biases as Q's $[Q_1,Q_2..Q_n]$ and the mixtures as M's $[M_1,M_2..M_n]$. The probability $P(Y_t|Q_t = i,M_t = m)$ is represented using a mixture of Gaussians for each state *i*. The Conditional probability distribution [CPDs] of Y_t and M_t nodes are,

$P(Y_t = y_t|Q_t = i,M_t = m) = N(y_t;\mu_{i,m},\Sigma_{i,m})$

$P(M_t = m|Q_t = i) = C(i,m)$, where C encodes the mixture weights for each state i.

We used Baum-Welch algorithm [EM] to compute maximum likelihood estimates and the parameter estimates of transition and emission (characterized by mean, covariance and weights). The posterior marginals were then estimated using the Forward-backward algorithm.

For each chromosome, we allowed 1 to 20 mixtures and chose the mixture with best goodness of fit using the AIC criterion, $AIC = 2k - 2\ln(L)$, k is the number of parameters in the model and L being the maximum likelihood estimate. Matlab was used to perform the HMM.

As a post-processing step, we estimated the median posterior probability of a region, defined as a stretch of same state, and believed only in regions having a median posterior marginal probabilities ≥ 0.99 or a region that is at least 80kb long.

Domains and boundaries are then inferred from the results of the HMM state calls throughout the genome. A domain is initiated at the beginning of a single downstream

biased HMM state. The domain is continuous throughout any consecutive downstream biased states. The domain will then end when the last in a series of upstream biased states are reached, with the domain ending at the end of the last HMM upstream biased state. We term the regions in between the topological domains as either "topological boundaries" or "unorganized chromatin." We defined unorganized chromatin to be these regions that are > 400kb, and the topological boundaries to be less than 400kb. We would note that the topological boundaries, though defined as regions less than 400kb, are mostly quite small, with 76.33% being less than 50kb in size (mESC data).

Transcription Factor and Histone Modification Enrichment Analysis

We collected histone modification ChIP-Seq datasets from a variety of publically available databases. For mouse, each dataset was mapped using Bowtie[51] to the NCBI Build 37/mm9 reference genome. For humans, the data was mapped using Bowtie to NCBI Build 36/hg18. Peaks were called using MACS[52]. We performed post-processing of the MACS peaks by filtering out peaks with less than a 2-fold enrichment in signal compared to matched input or less than an absolute difference in RPKM of 1. The peak or binding sites frequency was then calculated for every 10kb bin in the genome. For generating the average peak frequency plots, the mid-point of each boundary region was identified, and peak frequency was calculated in 10kb bins for +/- 500kb from the boundary mid-point. For block like factors (H3K9me3, H3K27me3, H3K36me3, and H3K79me2), we did not use MACS peak calling and each 10kb bin score was simply the log2 ratio of the total ChIP-seq signal over the 10kb window divided by the input signal

of the window. The data were either averaged for the enrichment graphs (Figure 4, Supplementary Figure 20) or were plotted as heatmaps (Figure 2).

For determining which boundaries are associated with a given factor, we considered a boundary to be associated with a factor if there were a binding site called by MACS (for chromatin factors like CTCF) or if there were a locus (for example, the transcription start site of a housekeeping gene) within +/- 20kb of the boundary. The 20kb window is chosen because this reflects the inherent uncertainty in the exact position of the domain calls due to 40kb binning. The analysis shown in the pie chart in Figure 4e is performed as follows: First, boundaries with CTCF were identified. Second, boundaries with housekeeping genes were identified. If a boundary was not associated with a housekeeping gene, yet is associated with a non-housekeeping gene according to entropy scores, that is shown as a "other gene" associated boundary.

For the analysis of the patterns of H3K9me3 and Lamina DamID signal surrounding the boundary regions shown in Figure 2, we used k-means clustering to cluster the data. For Figure 2d, k-means clustering is performed on the hESC and IMR90 data simultaneously. Likewise, the mESC and cortex data were also clustered simultaneously.

GO Terms Enrichment analysis

GO terms enrichment analysis was performed using the DAVID tool. In figure 4, we display only non-redundant GO terms with a Benjamini corrected p-value less than 10^{-3}.

Dynamic Interactions

Differential interactions between mESCs and cortex were modeled as a Binomial distribution. For this analysis we combined the data from two pairs of replicates together (mouse ES cell versus cortex). We performed a binomial test for each possible interaction in the genome up to a distance of 5Mbp. The total number of trials (n) is equal to the sum of the reads in the two mESC replicates plus the sum of the reads in the two cortex replicates that map between two 20kb bins (I_{ij}) at a distance (d) ($n = I_{ij\text{-mESC}} + I_{ij\text{-cortex}}$). The expected ratio (p) of the mESC to cortex read ratio is equal to the ratio of the sums of all reads in the two mESC replicates between bins at distance (d) throughout the genome compared to the sum of the reads total reads between bins at distance d ($p = \Sigma I_{mESC}/n$ at distance d or $p = \Sigma I_{cortex}/n$). Therefore, deviations in the ratio of the number of interactions in mouse ES cells ($I_{ij\text{-mESC}}$) to the number of interactions in cortex ($I_{ij\text{-cortex}}$) will result in a significant p-value. We would note that this method accounts for the differences in sequencing depth between the two libraries by considering the expected ratio (p), which is proportional to the total sequencing depth. To model the extent to which noise or variability could contribute to dynamic interacting regions, we performed the same analysis but randomly permuted the combination of data. Specifically, under random permutation 1, we combine the mouse ES replicate 1 with the cortex replicate 1 and compared this to the combination of mouse ES replicate 2 with cortex replicate 2. For random permutation 2, we combined the mouse ES replicate 1 with the cortex replicate 2 and compared this to the combination of mouse ES replicate 2 with cortex replicate 1. Under a null hypothesis that the mouse ES cell and cortex Hi-C data sets are the same, we would expect a similar number of dynamic interactions when the actual groupings were considered (mESC1+mESC2 vs. cortex1+cortex2) as we would under the

random permutation (mESC1+cortex1 vs mESC2+cortex2 or mESC1+cortex2 vs. mESC2+cortex1). This also allows for an estimate of the number of dynamic interactions that would be observed to due random chance or noise, allowing us to calculate the False Discover Rate (FDR) of identifying dynamic interaction regions (the FDR is equal to the number of observed dynamic interactions in the randomly permutated data divided by the number of observed interaction in the actual data). For the dynamic interaction analysis, we only considered data from Hi-C experiments using the HindIII restriction enzyme to eliminate restriction enzyme effects as a possible confounding factor.

Housekeeping and Tissue Specific Gene Expression

"Housekeeping" and "Tissue Specific" genes were identified based on gene expression data from the bioGPS gene atlas database[48]. Specifically, the normalized probe intensities are used as a measure of absolute gene expression, with gene x being expressed at a level x_i in a given tissue or cell type *i*. The probability of expression p_i in a given cell *i* type is calculated as:

$$p_i = \frac{x_i}{\sum_1^N x}$$

and the entropy score for a given gene x is calculated as:

$$H(x) = -1 * \sum_1^N p_i \log2(p_i)$$

High entropy scores (> 6.12, corresponding to uniform expression in >70/96 tissues) have relatively uniform expression patterns and are considered to be "housekeeping" genes, while low entropy scores (<4.9) have highly variable expression patterns and are considered tissue specific (uniform expression in < 30/96 tissues). We exclude genes with entropy score between 4.9 and 6.12 as these are not well categorized as either "tissue specific" or "housekeeping."

Boundary Correlation Between and Across Cell Types

To correlate the boundaries both between and across cell types, we calculated the Spearman correlation coefficient of the directionality index between two cells. Specifically, if a boundary was called by the HMM in either cell type, we would identify the center of that boundary and correlate a vector of directionality indexes +/- 10 bins from the center of the boundary between two experiments of interest. For random correlation, we randomly selected 20 bins from each of the two cell types and calculated the spearman correlation between the two vectors. We repeated the randomization 10,000 times to achieve the random distribution of spearman correlation coefficients. Boundaries were called as "cell type specific" if the boundary regions was identified by the HMM domain calling in only one cell and lacked a significant correlation in the directionality index between the two cell types.

Boundary Conservation Across Species

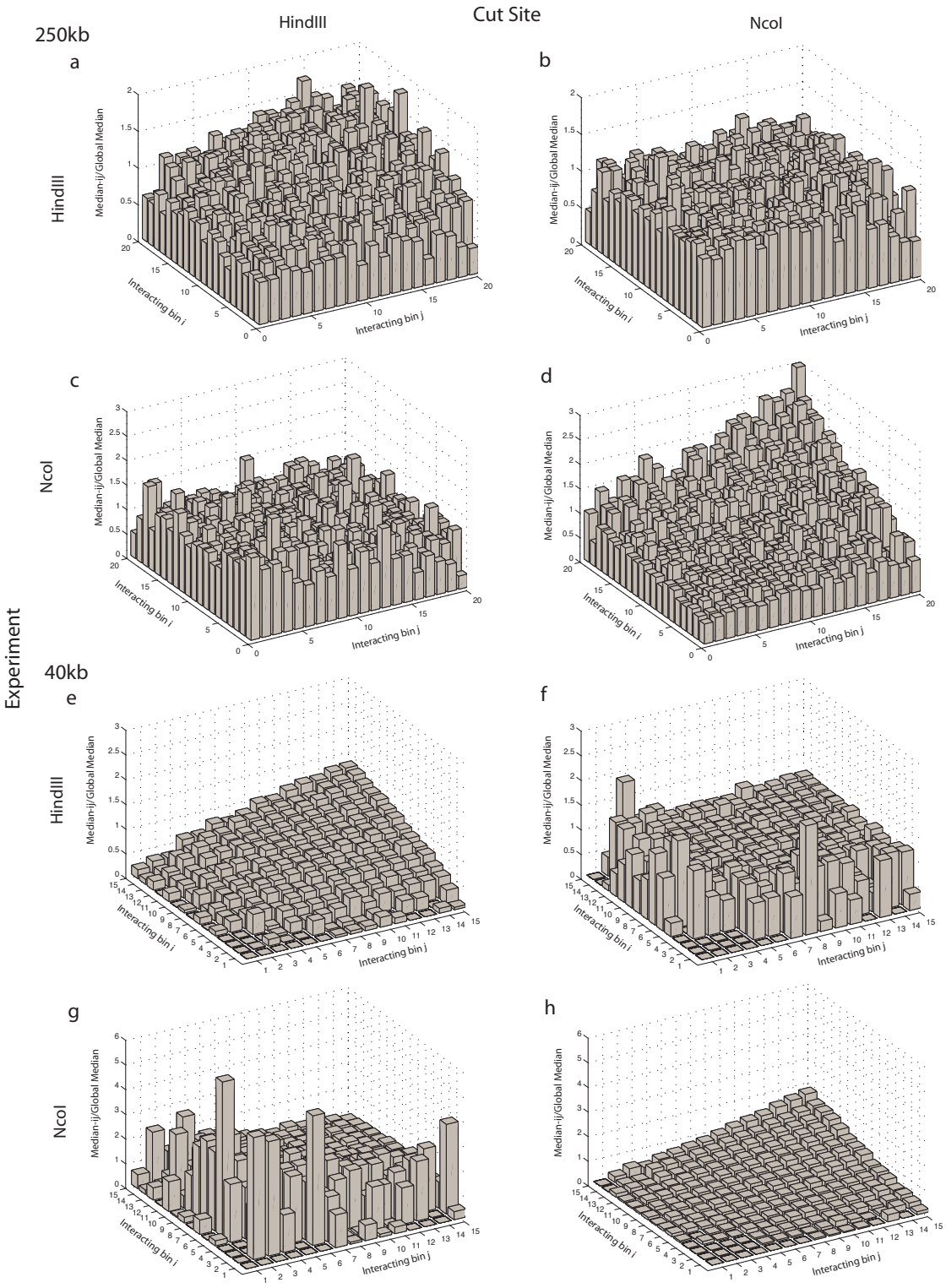
Boundaries were lifted over using the UCSC Liftover tool[53] from species1 to species2 and the overlap between species1to2:species2 and species2to1:species1 were estimated. This overlap was compared with the random boundaries. The random boundaries were constrained on the distribution of boundary lengths and distribution of chromosomal occurrence.

III. Supplemental References

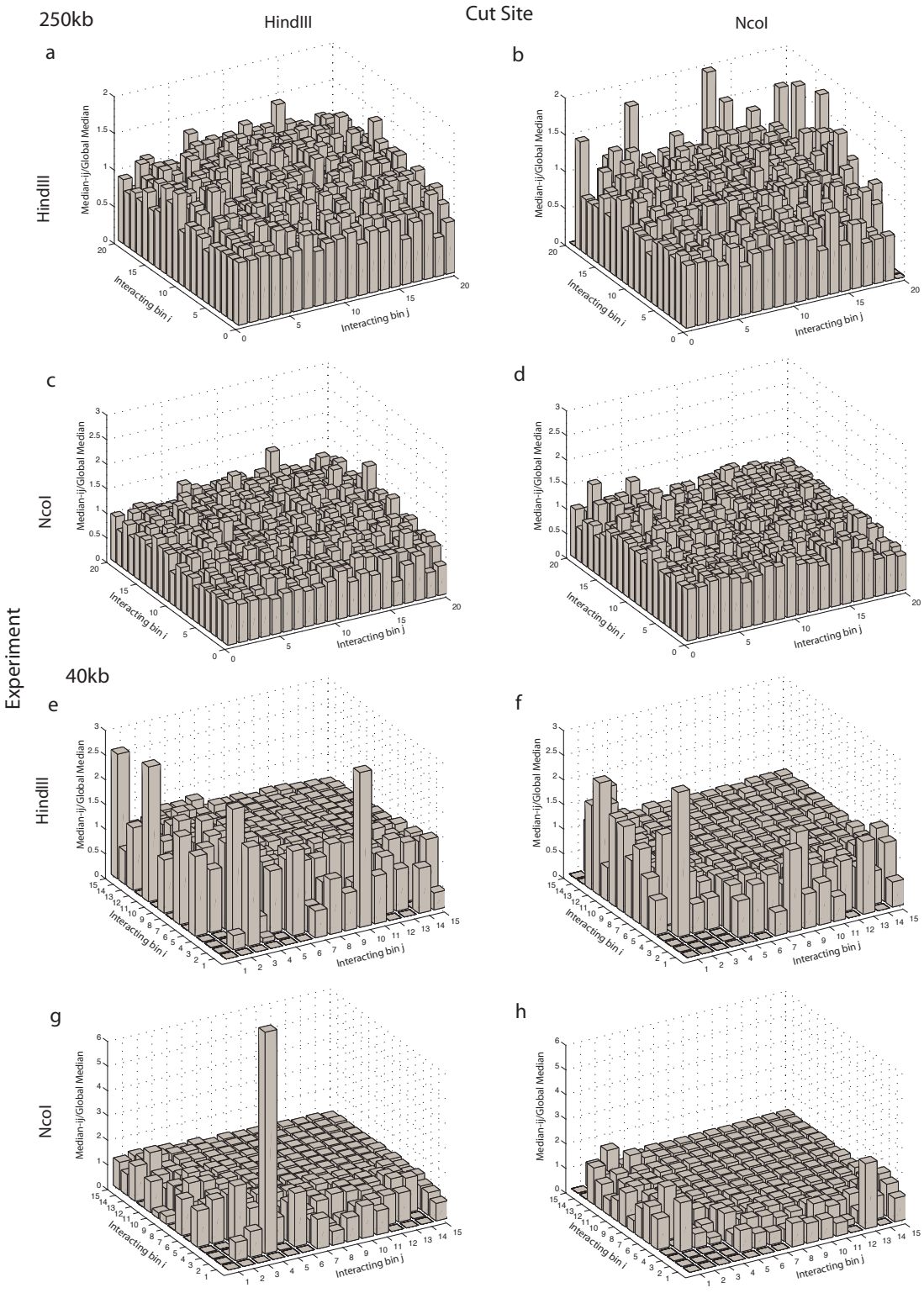
31. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
32. Shen, Y. et al. A Map of cis-Regulatory Sequences in the Mouse Genome. *in submission* (2012).
33. Wang, K.C. et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-4.
34. Kagey, M.H. et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-5.
35. Eskeland, R. et al. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol Cell* **38**, 452-64.
36. Xie, W. et al. Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome. *Cell* **148**, 816-31.
37. Hawkins, R.D. et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479-91.
38. Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell* **38**, 603-13.
39. Hiratani, I. et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res* **20**, 155-69.
40. Wen, B., Wu, H., Shinkai, Y., Irizarry, R.A. & Feinberg, A.P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat Genet* **41**, 246-50 (2009).
41. Mikkelsen, T.S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-60 (2007).
42. Marson, A. et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521-33 (2008).
43. Bilodeau, S., Kagey, M.H., Frampton, G.M., Rahl, P.B. & Young, R.A. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev* **23**, 2484-9 (2009).
44. Peng, J.C. et al. Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* **139**, 1290-302 (2009).
45. Rahl, P.B. et al. c-Myc regulates transcriptional pause release. *Cell* **141**, 432-45.
46. Schnetz, M.P. et al. Genomic distribution of CHD7 on chromatin tracks H3K4 methylation patterns. *Genome Res* **19**, 590-601 (2009).
47. Min, I.M. et al. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* **25**, 742-54.
48. Lattin, J.E. et al. Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Res* **4**, 5 (2008).
49. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**, 1059-65.

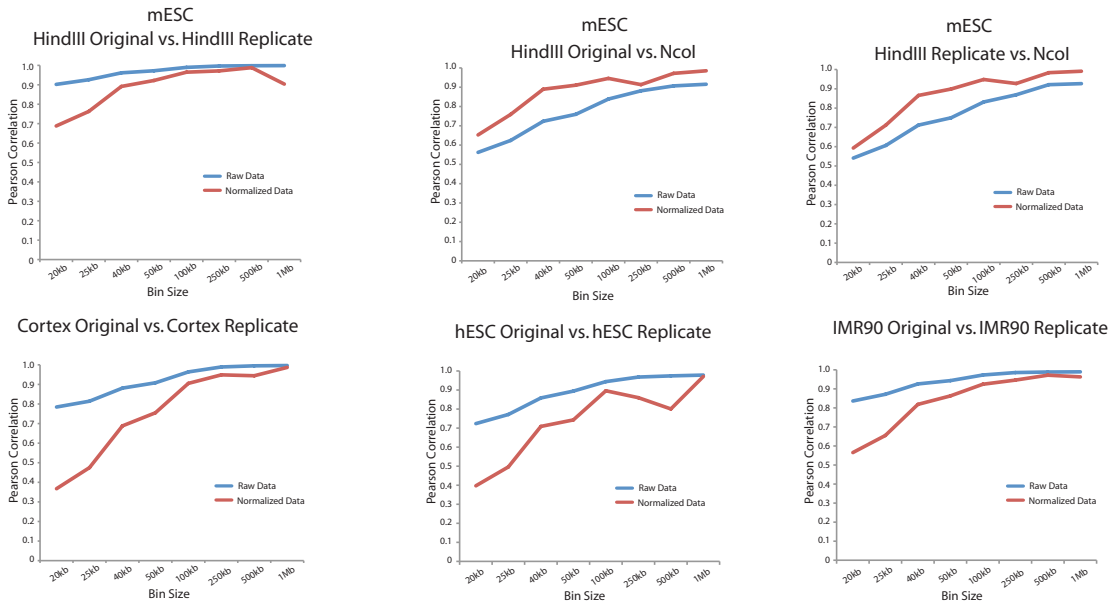
50. Facucho-Oliveira, J.M., Alderson, J., Spikings, E.C., Egginton, S. & St John, J.C. Mitochondrial DNA replication during differentiation of murine embryonic stem cells. *J Cell Sci* **120**, 4025-34 (2007).
51. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
52. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
53. Kent, W.J. et al. The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).

Supplementary Figure 1. Raw Hi-C Data and Restriction Enzyme Bias. a-d, Bias plots showing the correlation between restriction enzyme cut site frequency and Hi-C interaction frequency using a bin size of 250kb at a distance of 1Mb. For a-d, all 250kb bins were grouped into 20 equal sized groups based on increasing restriction enzyme frequency. The two horizontal axes correspond to the restriction enzyme group of each of the two bins, *i* and *j*, involved in an interaction I_{ij}. The vertical axis shows the median of all interactions I_{ij} divided by the global median. Perfectly unbiased data should have all values roughly equal to 1. a, Comparison of HindIII restriction enzyme frequency with HindIII Hi-C data. b, Comparison of NcoI restriction enzyme frequency with HindIII Hi-C data. c, Comparison of HindIII restriction enzyme frequency with NcoI Hi-C data. d, Comparison of NcoI restriction enzyme frequency with NcoI Hi-C data. Note the correlation between the restriction enzyme cut site frequency and the Hi-C interaction frequency is only present when considering the restriction enzyme used in the Hi-C experiment. e-h, Similar to a-d, but using a bin size of 40kb and a distance of 80kb. The horizontal axis in e-h are the number of cut sites/40kb bin.


Cut Site
HindIII
NcoI
Experiment
250kb
40kb
HindIII
NcoI
a
b
c
d
e
f
g
h
Median-ij/Global Median
Interacting bin i
Interacting bin j

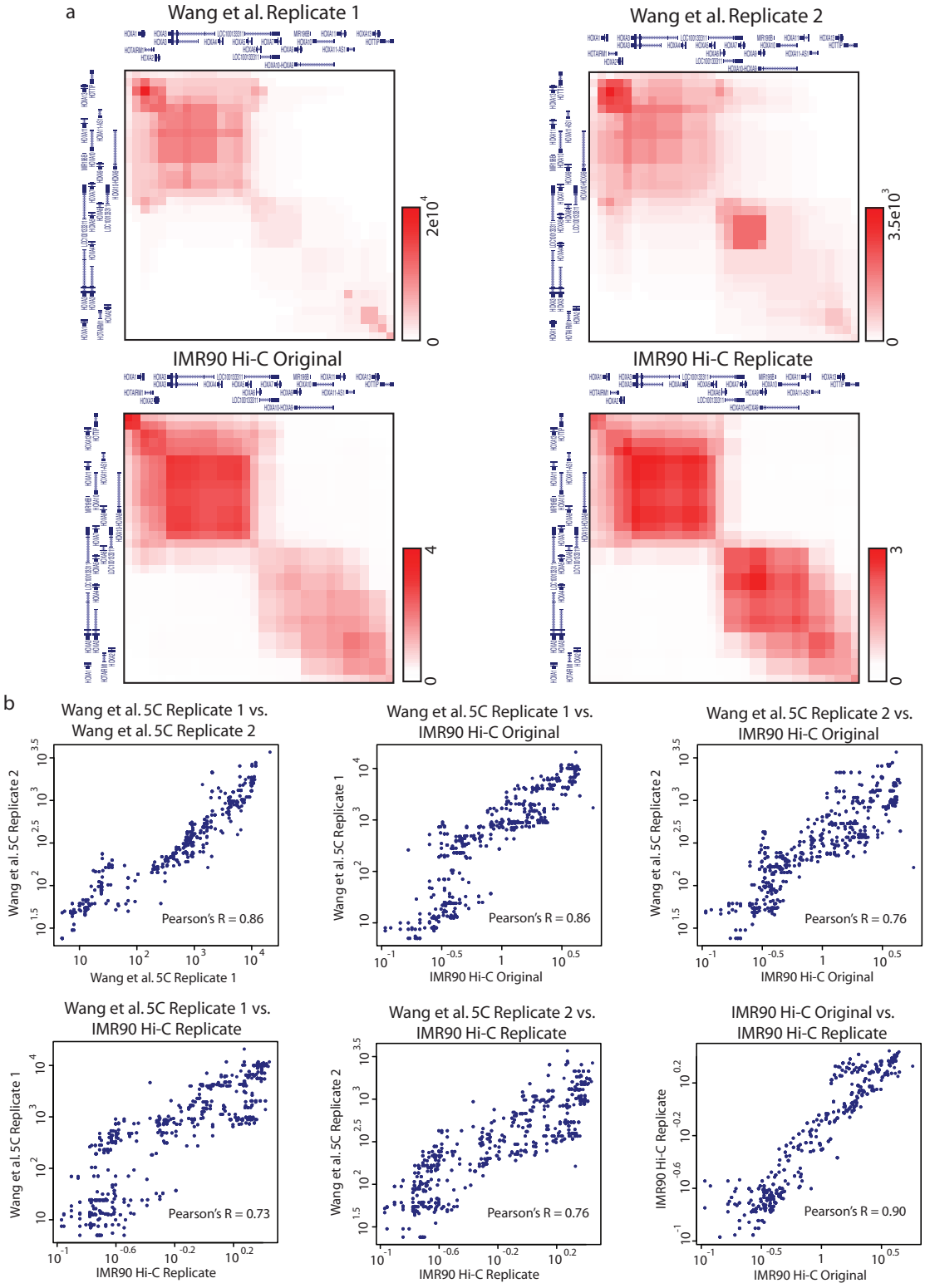
Supplementary Figure 2. Normalized Hi-C data shows no restriction enzyme bias. Identical to Supplementary Figure 1, yet using the normalized Hi-C data. Note that most values are roughly equal to 1, regardless of bin size or restriction enzyme, demonstrating that the restriction enzyme bias has been eliminated with normalization.

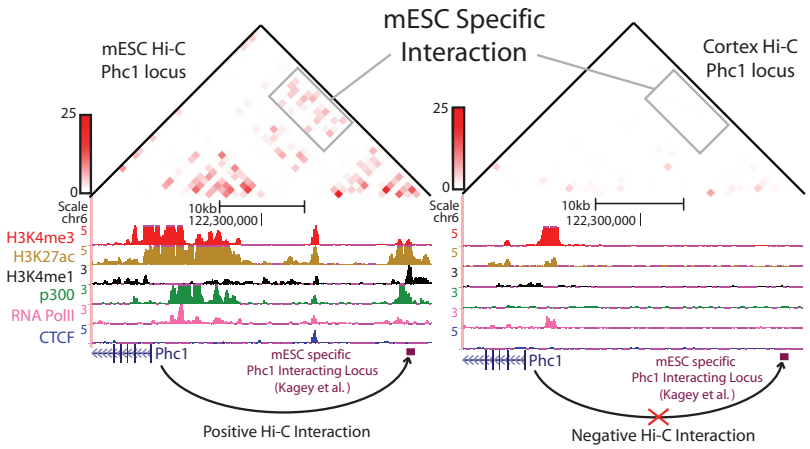

250kb
Cut Site
HindIII
NcoI
Experiment
HindIII
NcoI
40kb
HindIII
NcoI
a
b
c
d
e
f
g
h
Median-ij/Global Median
Interacting bin i
Interacting bin j



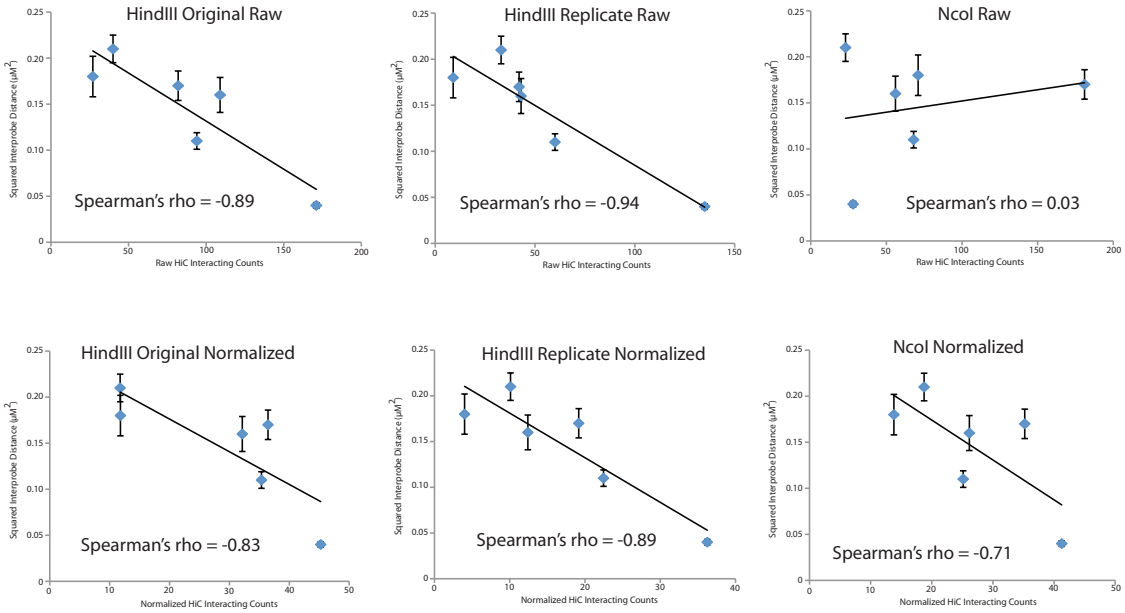
Supplementary Figure 3. Pearson Correlation between replicates. The Pearson correlation was calculated between each Hi-C replicate at varying bin sizes. The non-normalized data are shown in blue. The normalized data are shown in red.

Supplementary Figure 4. Comparison with Previous 5C. a, Heat maps over the HoxA locus of 5C data from lung fibroblasts as reported previously[33] and the IMR90 Hi-C data generated in this report. Visually, there are two separate clusters of interactions in the upper left and lower right portions of the heat map. b, Scatter plots showing the correlations between 5C replicates and Hi-C data. In all cases, the correlation is > 0.73, demonstrating a high degree of correlation between IMR90 Hi-C data and existing 5C data a similar cell type.


a
Wang et al. Replicate 1
2e10^4
0
Wang et al. Replicate 2
3.5e10^3
0
IMR90 Hi-C Original
4
0
IMR90 Hi-C Replicate
3
0
b
Wang et al. 5C Replicate 1 vs. Wang et al. 5C Replicate 2
Wang et al. 5C Replicate 2
Wang et al. 5C Replicate 1
Pearson's R = 0.86
Wang et al. 5C Replicate 1 vs. IMR90 Hi-C Original
Wang et al. 5C Replicate 1
IMR90 Hi-C Original
Pearson's R = 0.86
Wang et al. 5C Replicate 2 vs. IMR90 Hi-C Original
Wang et al. 5C Replicate 2
IMR90 Hi-C Original
Pearson's R = 0.76
Wang et al. 5C Replicate 1 vs. IMR90 Hi-C Replicate
Wang et al. 5C Replicate 1
IMR90 Hi-C Replicate
Pearson's R = 0.73
Wang et al. 5C Replicate 2 vs. IMR90 Hi-C Replicate
Wang et al. 5C Replicate 2
IMR90 Hi-C Replicate
Pearson's R = 0.76
IMR90 Hi-C Original vs. IMR90 Hi-C Replicate
IMR90 Hi-C Replicate
IMR90 Hi-C Original
Pearson's R = 0.90

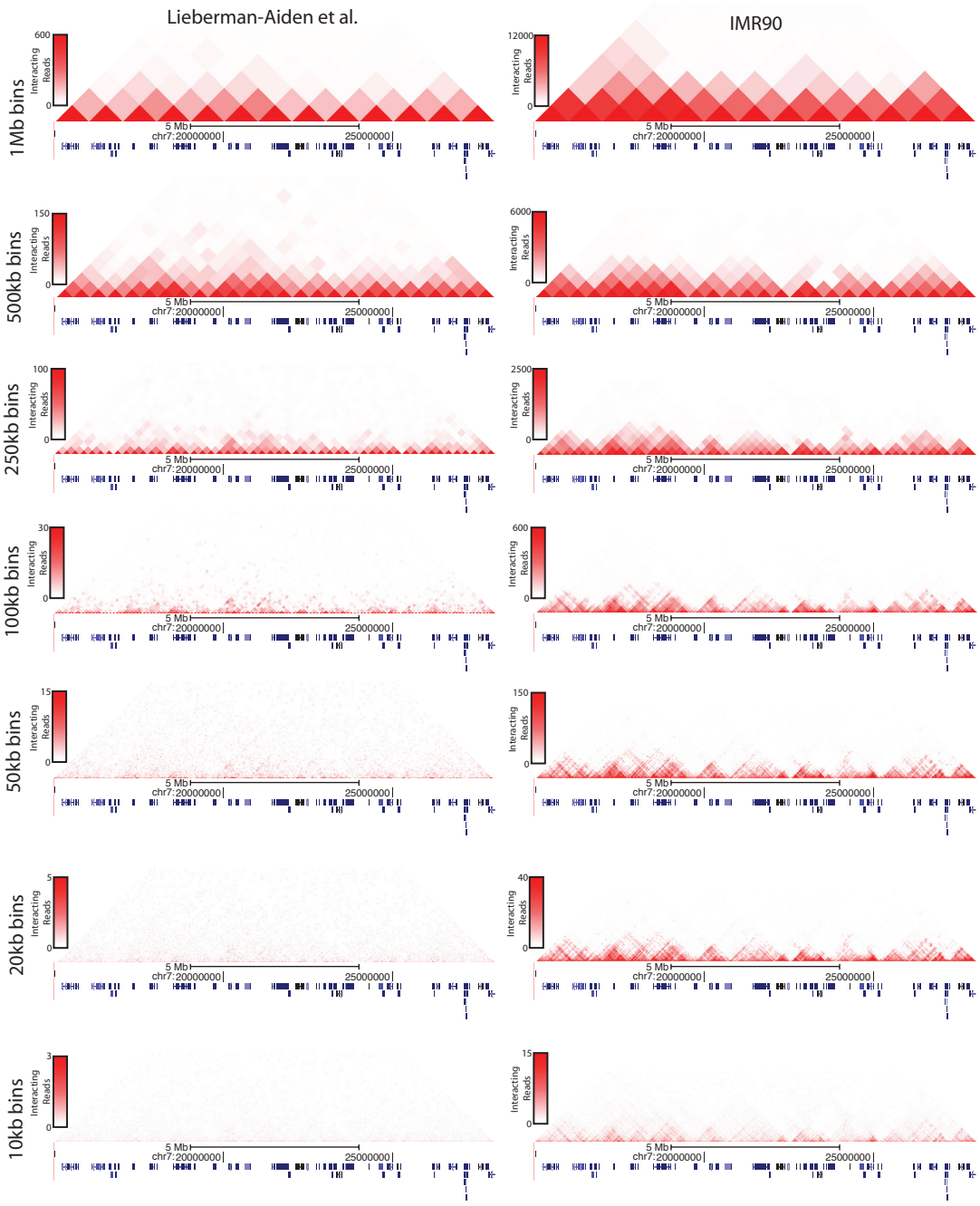


Supplementary Figure 5. Comparison with Previous 3C data. 2D heatmap of Hi-C interactions at the Phc1 locus. The Phc1 promoter was previously shown to interact with a nearby enhancer by 3C, indicated by the arrow and red box. Gray boxes indicate the mESC specific Hi-C interactions.

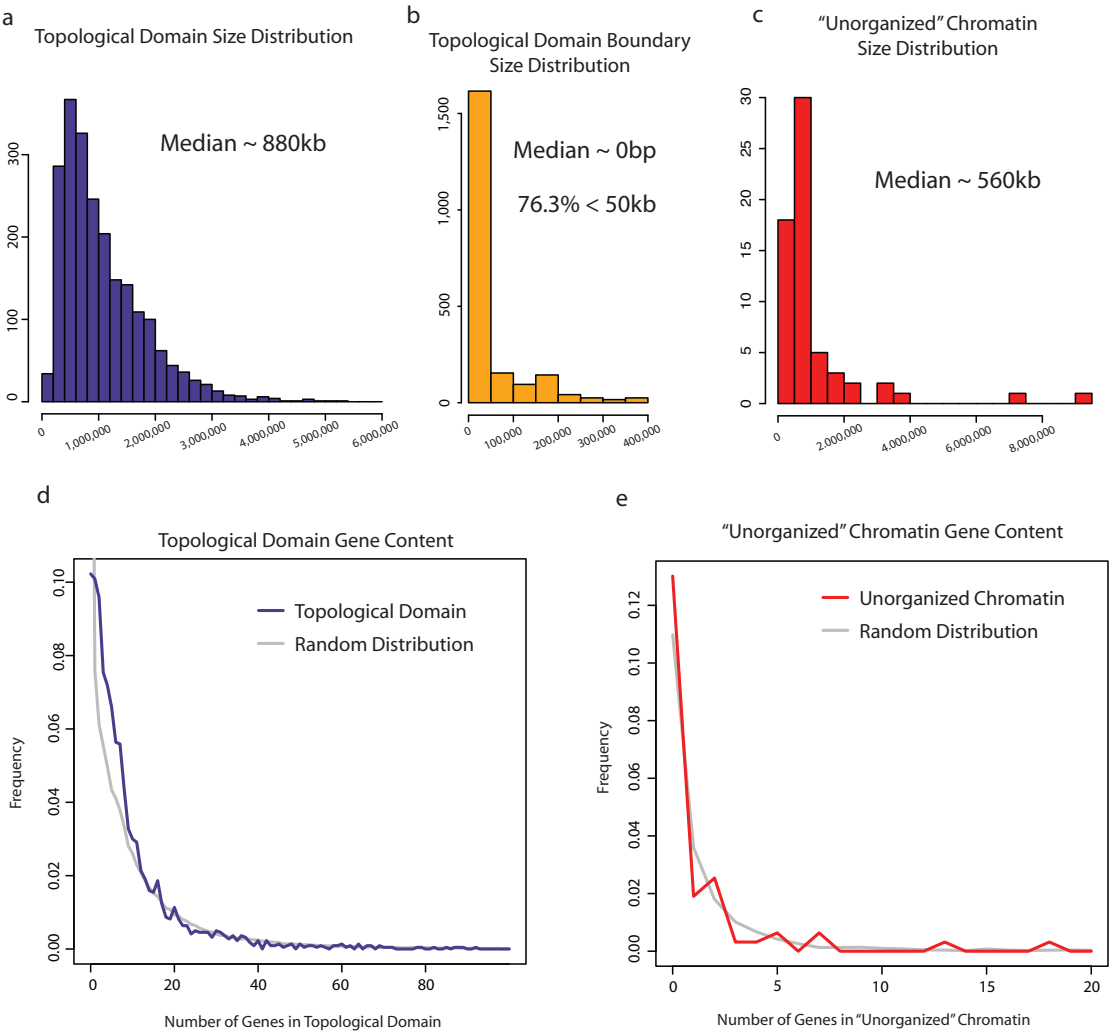


Supplementary Figure 6. Hi-C interaction frequency and mean spatial distance. The raw and normalized Hi-C interaction frequencies were compared with the mean nuclear separation as measured by 2D-FISH between six loci. The 2D-FISH data are from ref. 35.

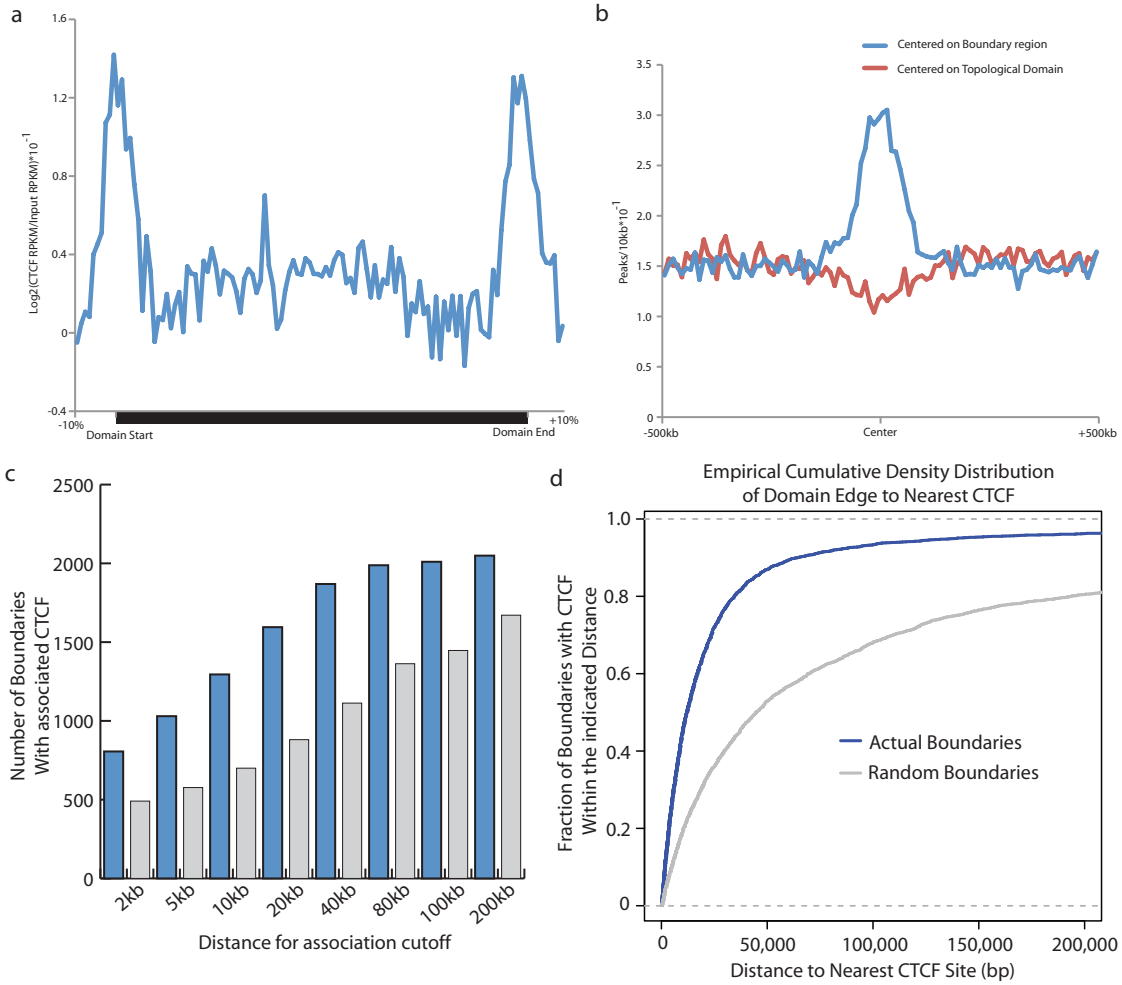
Supplementary Figure 7. Hi-C interaction heat maps at varying bin sizes. Hi-C interaction frequencies are displayed as 2D heatmaps using differing bin sizes over a single locus on chromosome 7. Not the presence of the "triangles" on the heat map at a bin size or resolution of 100kb or less. A comparison with the data from the original Hi-C report is also shown for comparison[31].


Lieberman-Aiden et al.
IMR90
1Mb bins
500kb bins
250kb bins
100kb bins
50kb bins
20kb bins
10kb bins
Interacting Reads
5 Mb
chr7:20000000
25000000

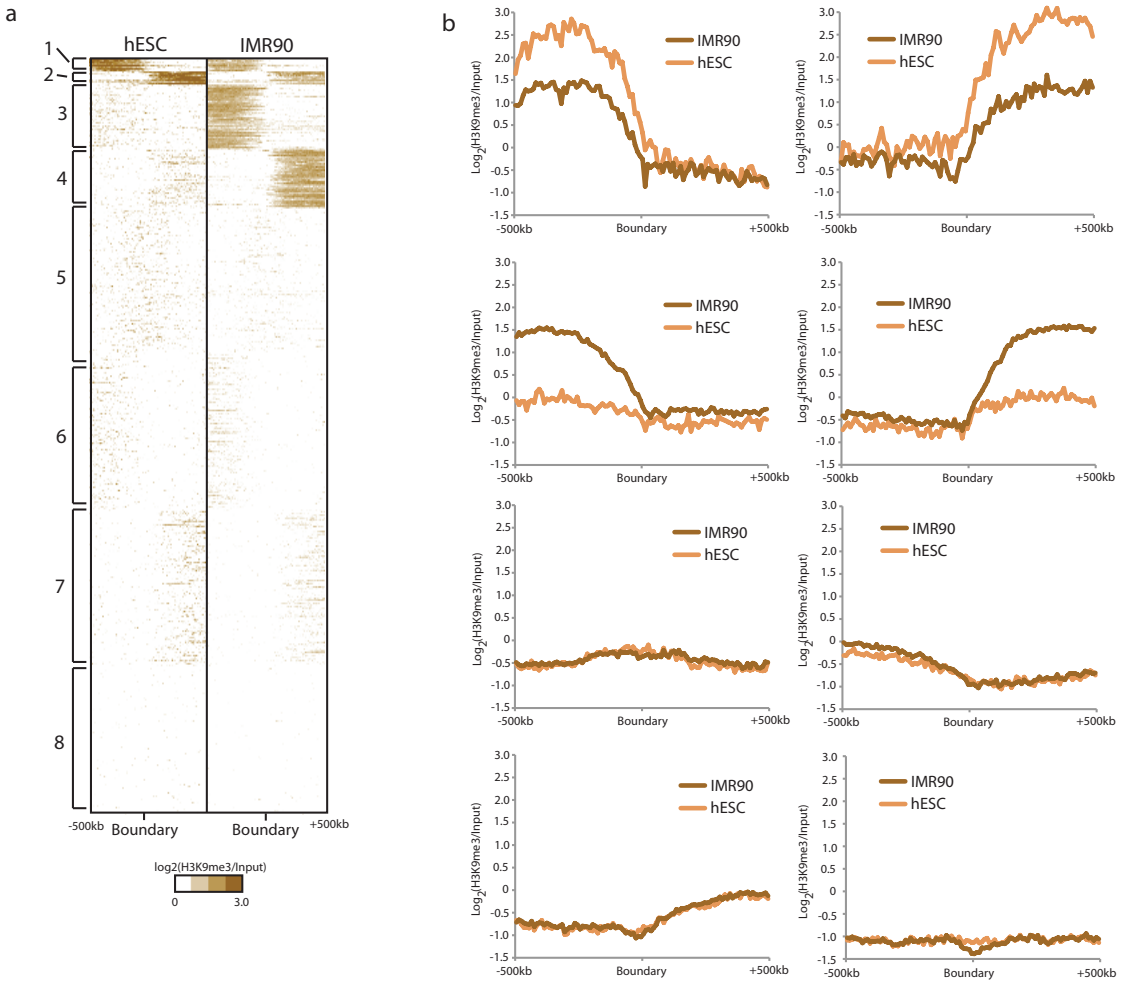
Supplementary Figure 8. Overlap of Topological Domain Boundaries between Hi-C replicates. Venn-diagrams comparing the amount of overlap between the topological domain boundaries called in each pair of Hi-C replicates.



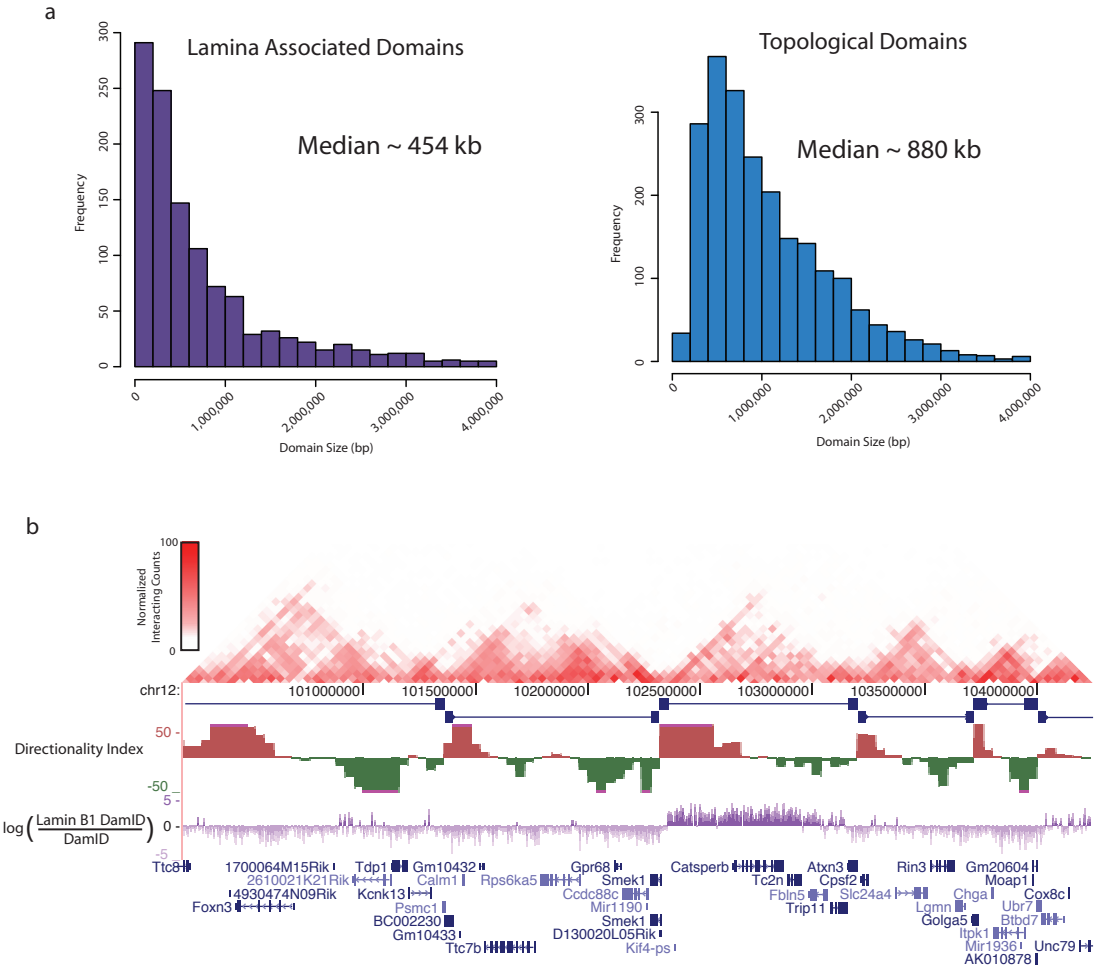
Supplementary Figure 9. Size distribution and gene content of topological domains, boundaries, and unorganized chromatin. a-c, Histograms of the sizes of topological domains (a), topological boundaries (b), and unorganized chromatin (c). d,e, Distribution of the gene content of topological domains and unorganized chromatin. Shown in gray is the gene content for randomly chosen regions of the genome with the same size distribution. Neither topological domains nor unorganized chromatin appear to differ from what is expected at random in terms of the distribution of their gene content.



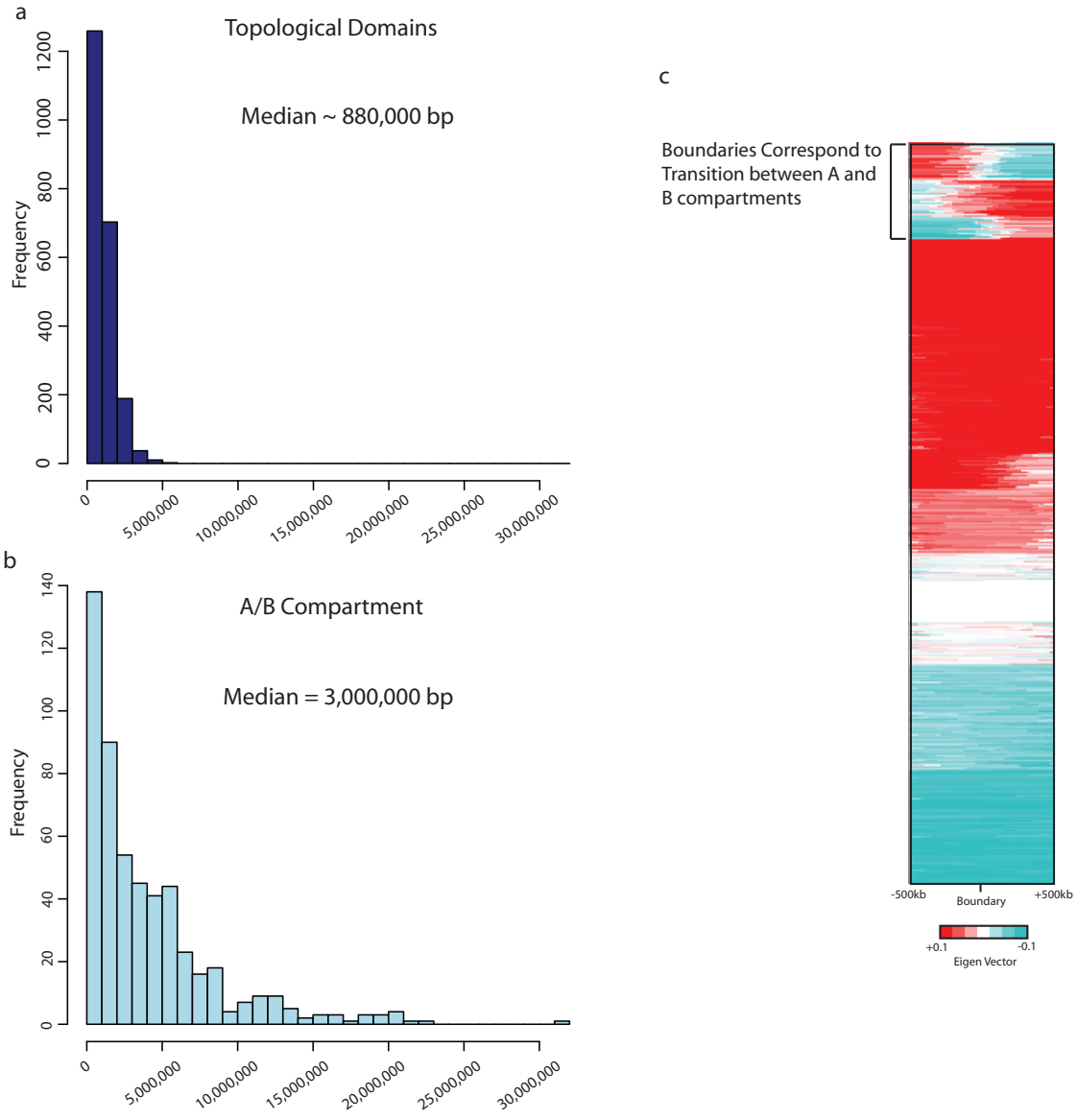
Supplementary Figure 10. CTCF enrichment at topological boundary regions. a, Average enrichment plot of CTCF over topological domains. Each topological domain was divided into 100 equally sized bins (+/- 10 bins from each end of the domain as well). The log2 ratio of CTCF RPKM over Input was calculated for each bin and shown as an average. CTCF appears to be enriched at the edges of each topological domain. b, Average enrichment plot of CTCF as shown in peaks/10kb bin. Shown in blue is the CTCF signal when centered on the topological boundary region. Shown in blue is the CTCF signal when centered on the middle of a topological domain, showing no enrichment. c, The number of boundaries with an associated CTCF site is shown for varying window size cut offs. For each distance *D*, the number of boundaries with a CTCF within +/- *D* are shown in blue. Shown in gray is the number expected at random at the same distance cut-off. d, The empirical cumulative density distribution of the distance between the domain border and the nearest CTCF binding site (in bp). The distance between the actual boundaries and the nearest CTCF site is shown in blue. The distance to randomized boundaries is shown in grey.



Supplementary Figure 11. Average Enrichment Plots of H3K9me3 surrounding the boundaries. a, Identical to Figure 2d in the main text, but labeled with cluster names 1-8 based on k-means clustering. b, The average enrichment plots of H3K9me3 for clusters 1-8 from panel a. Clusters 1-4 show clear enrichment of H3K9me3, and the transition from enriched to depleted H3K9me3 regions coincides with the location of the topological boundaries.

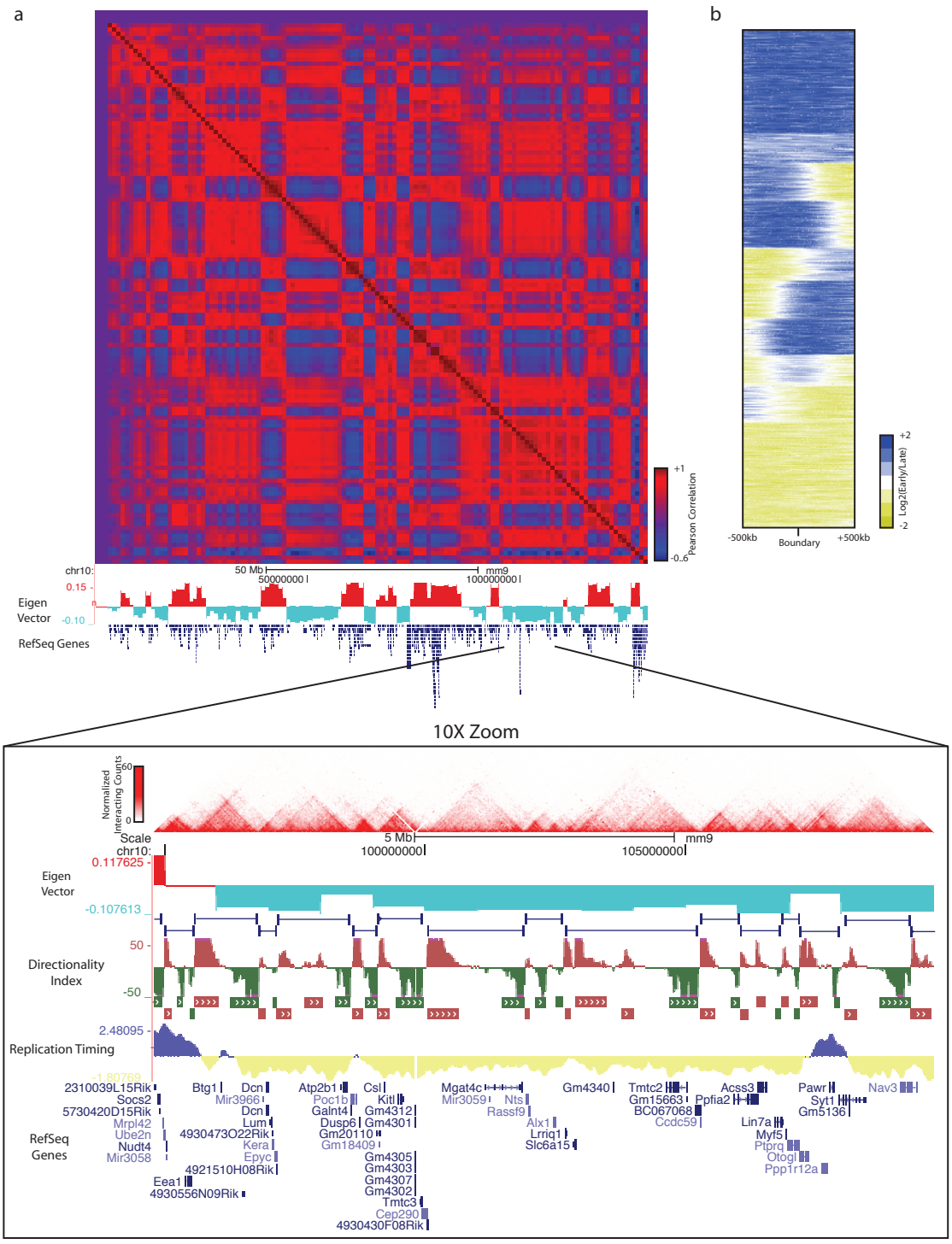


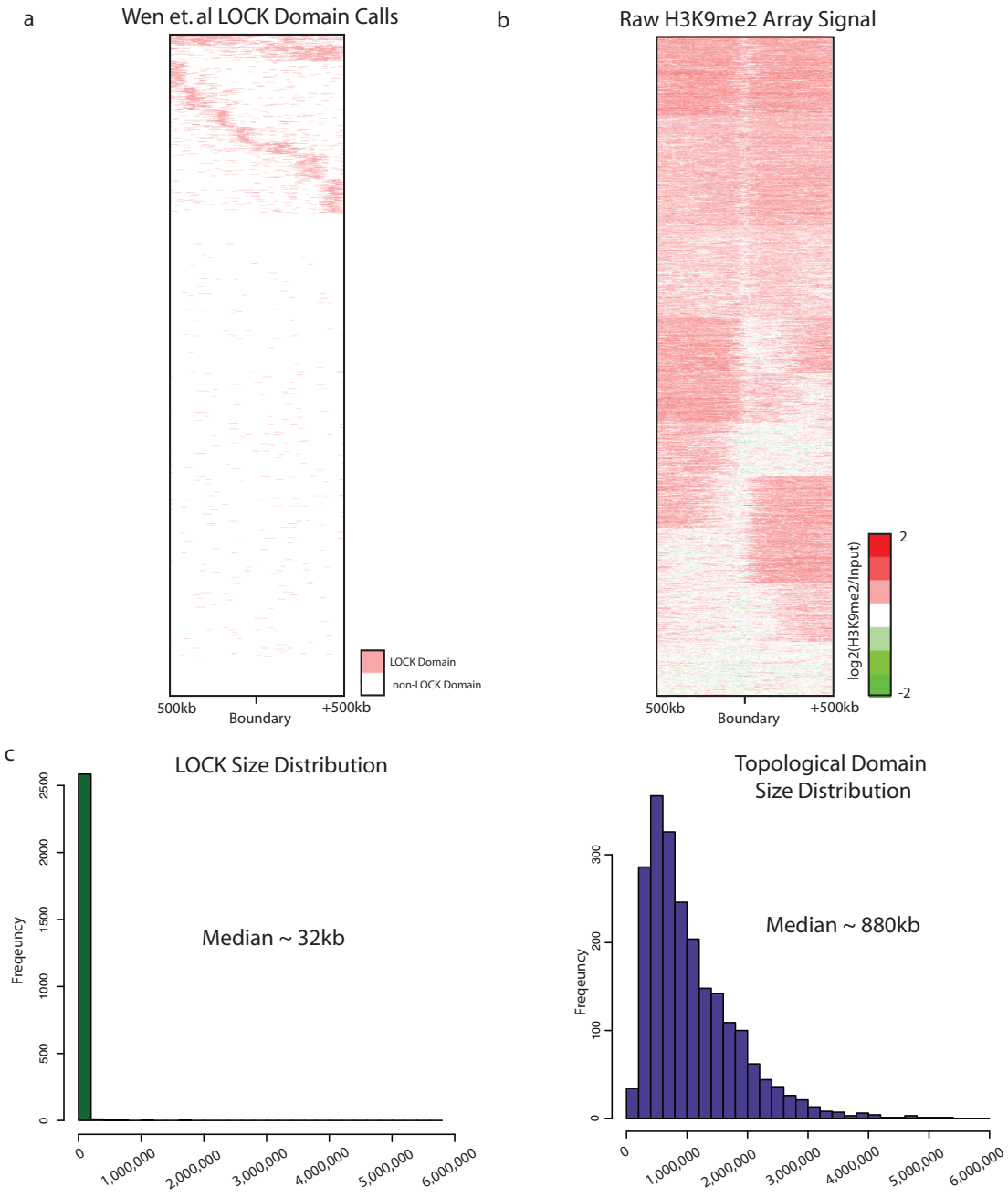
Supplementary Figure 12. Comparison of Topological Domains with Lamina Associated Domains (LADs). a, Histogram showing the size distribution of the topological domains and the LADs. Generally, LADs are smaller in size than topological domains. b, Genome browser shot showing a region on chromosome 12 with multiple topological domains, one of which appears to be entirely lamina-associated, with the remainder are non-lamina associated.



Supplementary Figure 13. Comparison of A and B compartments with topological domains in mouse ES cells. a,b, Histograms showing the size distributions of topological domains (a) and A and B compartments (b). Generally, the A and B compartments are larger than the topological domains. c, Heat map of the Eigen Vector values used to determine the A and B compartments at the topological boundary regions in mouse ES cells. The subset of boundaries that mark the transition between an A and B compartment are marked.

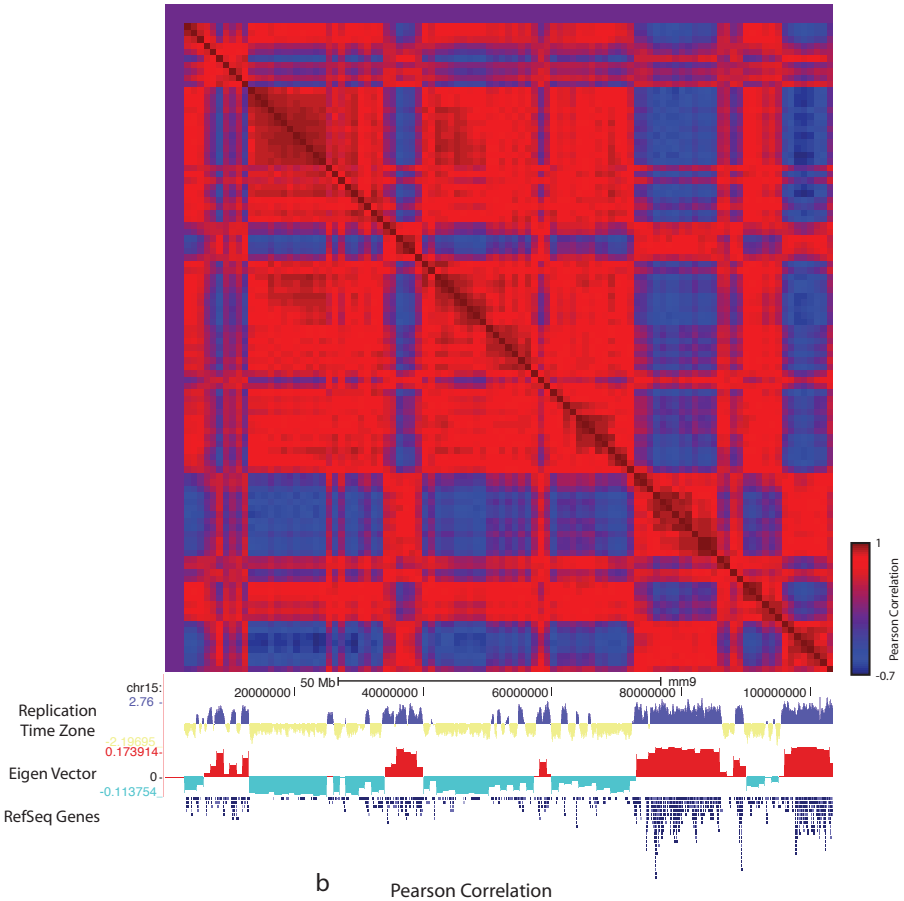
Supplementary Figure 14. Comparison of Topological Domains with A and B compartments and Replication Time Zones. a, Pearson correlation interaction heat map over chromosome 10. Shown in the blow up is a 10X zoom on a region entirely within the "B" compartment with multiple topological domains present in the region. b, Heat map of the replication time zone microarray data (ref. 39), surrounding the topological boundary regions.


a
b
+1
Pearson Correlation
-0.6
+2
Log2(Early/Late)
-2
-500kb
Boundary
+500kb
chr10:
50 Mb
50000000
mm9
100000000
0.15
Eigen
Vector
-0.10
RefSeq Genes
10X Zoom
Normalized
Interacting Counts
60
0
Scale
5 Mb
chr10:
100000000
105000000
mm9
0.117625
Eigen
Vector
-0.107613
50
Directionality
Index
-50
2.48095
Replication Timing
-1.80769
RefSeq
Genes
2310039L15Rik
Socs2
5730420D15Rik
Mrpl42
Ube2n
Nudt4
Mir3058
Btg1
Dcn
Mir3966
Dcn
Lum
4930473O22Rik
Kera
Epyc
4921510H08Rik
Eea1
4930556N09Rik
Atp2b1
Csl
Poc1b
Kitl
Galnt4
Gm4312
Dusp6
Gm4301
Gm20110
Gm18409
Gm4305
Gm4303
Gm4307
Gm4302
Tmtc3
Cep290
4930430F08Rik
Mgat4c
Mir3059
Nts
Rassf9
Alx1
Lrriq1
Slc6a15
Gm4340
Tmtc2
Gm15663
BC067068
Ccdc59
Acss3
Ppfia2
Lin7a
Myf5
Ptprq
Otogl
Ppp1r12a
Pawr
Syt1
Gm5136
Nav3



Supplementary Figure 15. Comparison of Topological Domains with LOCK domains. a,b, Heat maps showing the enrichment of LOCK domains surrounding the topological boundary regions. Shown in (a) are the called LOCK domains[40], displayed as either LOCK in red or non-LOCK in white. Shown in (b) is the raw microarray data. c. Histograms showing the size distribution of LOCK domains and topological domains.

a



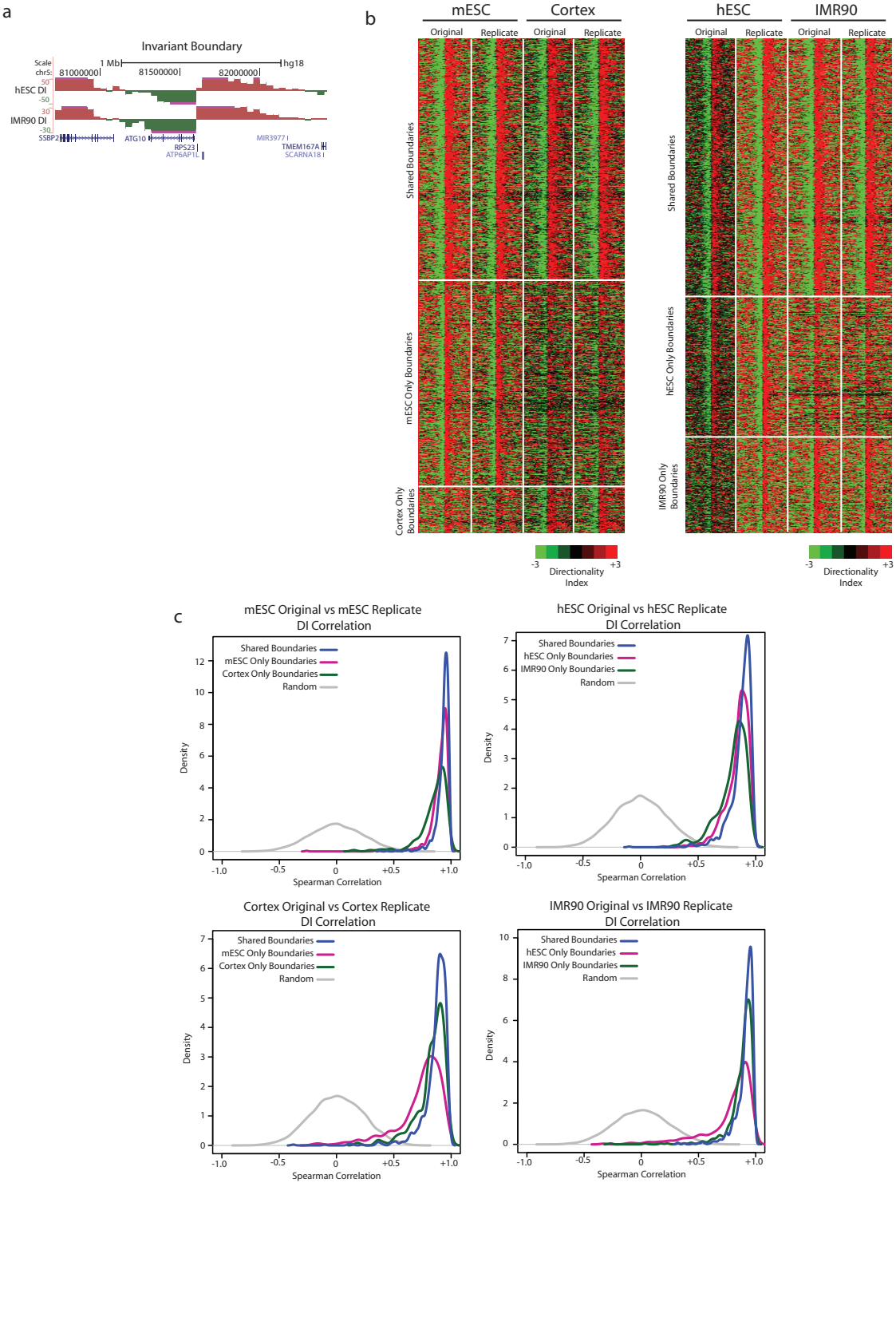
b

Pearson Correlation

chr1	0.8851848
chr2	0.8840393
chr3	0.8435305
chr4	0.8936122
chr5	0.8861238
chr6	0.8416694
chr7	0.867437
chr8	0.8644409
chr9	0.8491173
chr10	0.872831
chr11	0.8923897
chr12	0.8651408
chr13	0.7615307
chr14	0.8289586
chr15	0.9228144
chr16	0.8513164
chr17	0.8556229
chr18	0.8418625
chr19	0.8943993

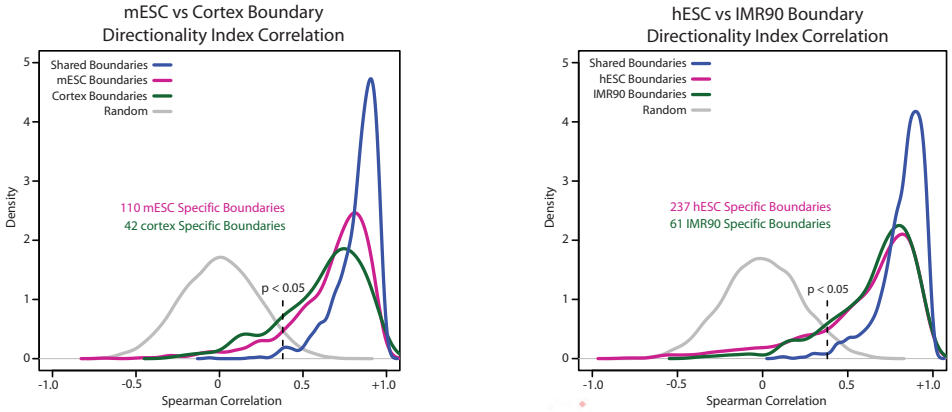
Supplementary Figure 16. Correlation of A and B compartments and replication time zones in mouse ES cells. a, Pearson correlation interaction heat map across chromosome 15. Below the heat map is the genome browser view of the Eigen vector used to determine the A or B compartments and the replication timing microarray data [39] b, Pearson correlation coefficients of the Eigen vector values and the average probe intensity for replication timing data in 1Mb bins over each chromosome.

Supplementary Figure 17. Domains are largely stable between cell types. a, Genome browser shot of an invariant boundary between hESC and IMR90 and the DI surrounding the boundary regions. b, Heat maps showing the directionality index surrounding the topological boundary regions. The heat maps are divided into three regions. Shared boundaries, boundaries called in cell type A and boundaries called in cell type B. c, Density plot of the Spearman correlations between the directionality indexes between Hi-C replicates at the topological boundary regions. Shown in blue are the shared boundaries. Shown in red is the boundaries called in ES cells (human or mouse) and shown in green are boundaries called in differentiated cells (human or mouse). Shown in grey are randomly generated spearman correlations. The replicates are all highly correlated at the boundary regions, regardless of whether the boundaries are called as shared or cell type specific.

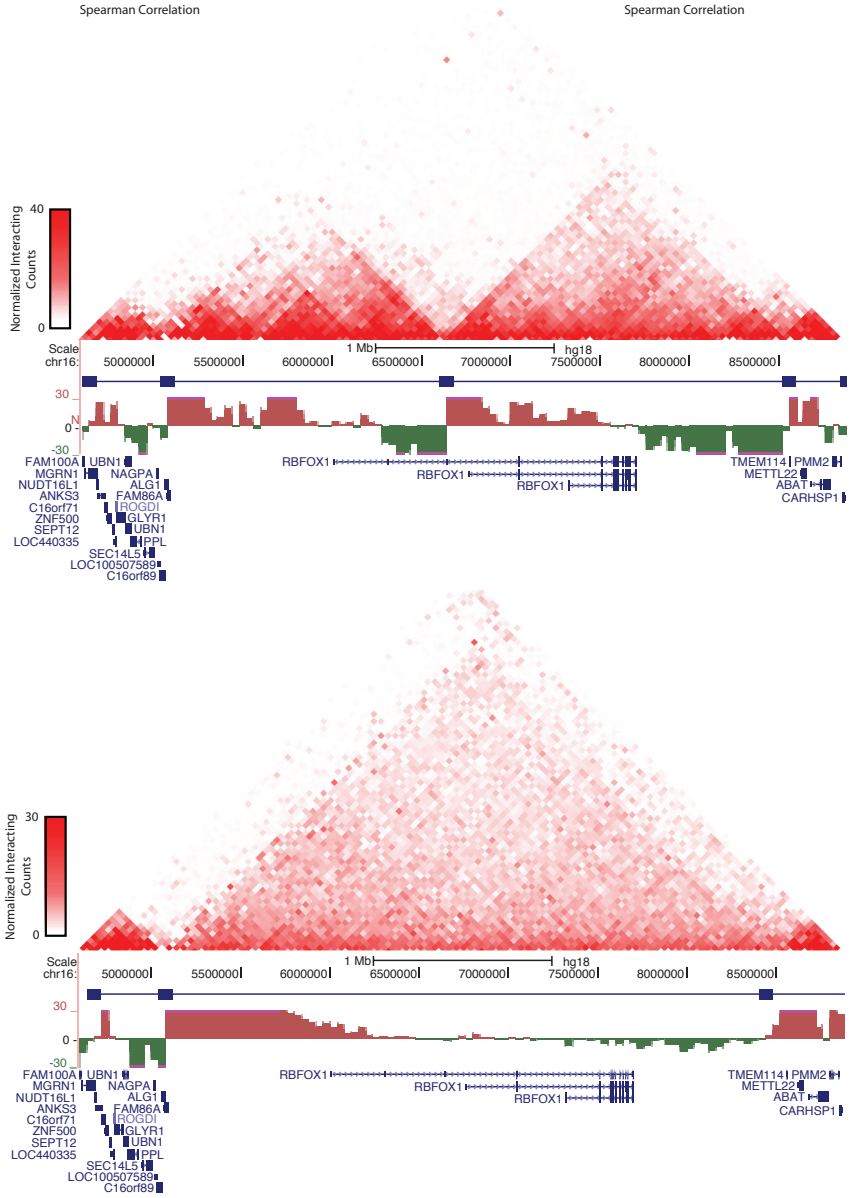

a
Invariant Boundary
Scale
chr5:
1 Mb
hg18
81000000
81500000
82000000
hESC DI
IMR90 DI
SSBP2
ATG10
RPS23
ATP6AP1L
MIR3977
TMEM167A
SCARNA18
b
mESC
Cortex
hESC
IMR90
Original
Replicate
Shared Boundaries
mESC Only Boundaries
Cortex Only Boundaries
hESC Only Boundaries
IMR90 Only Boundaries
-3
Directionality Index
+3
c
mESC Original vs mESC Replicate DI Correlation
hESC Original vs hESC Replicate DI Correlation
Cortex Original vs Cortex Replicate DI Correlation
IMR90 Original vs IMR90 Replicate DI Correlation
Shared Boundaries
mESC Only Boundaries
hESC Only Boundaries
IMR90 Only Boundaries
Cortex Only Boundaries
Random
Density
Spearman Correlation
-1.0
-0.5
0
+0.5
+1.0

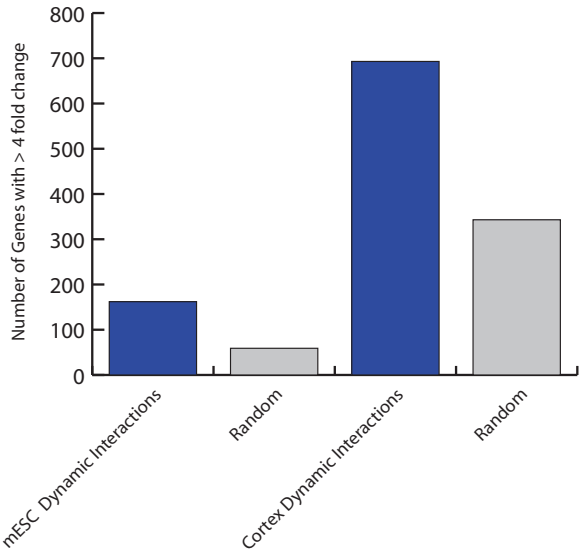
Supplementary Figure 18. Cell type specific domains. a, we determined cell type specific domains between cell types by calculating the spearman correlation coefficient between the DI at each boundary called in a cell types. The DI at most boundaries is still well correlated in different cell types. We call a boundary as cell type specific if the boundary is called by HMM in only one cell type and the spearman correlation of the directionality index is not significant when compared to a random distribution of spearman correlations. A minority of boundaries are actually called as cell types specific. b, A genome browser shot of a cell type specific domain on chromosome 16. The domain is called in hESCs and is not called in IMR90.

a

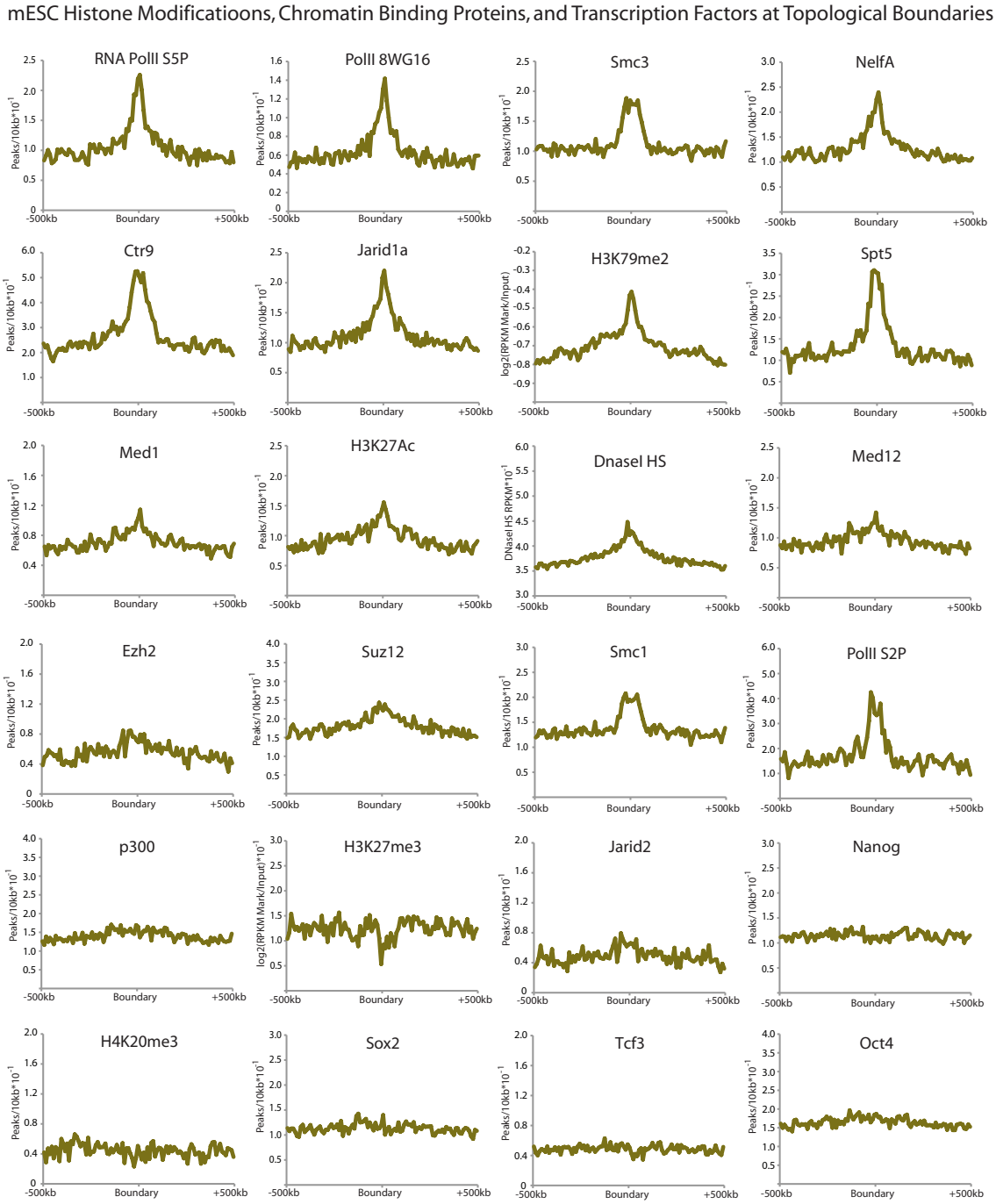

mESC vs Cortex Boundary
Directionality Index Correlation
Shared Boundaries
mESC Boundaries
Cortex Boundaries
Random
110 mESC Specific Boundaries
42 cortex Specific Boundaries
p < 0.05
Density
Spearman Correlation
hESC vs IMR90 Boundary
Directionality Index Correlation
Shared Boundaries
hESC Boundaries
IMR90 Boundaries
Random
237 hESC Specific Boundaries
61 IMR90 Specific Boundaries
p < 0.05
Density
Spearman Correlation

b


Normalized Interacting Counts
Scale
chr16:
1 Mb
hg18
FAM100A
MGRN1
NUDT16L1
ANKS3
C16orf71
ZNF500
SEPT12
LOC440335
UBN1
NAGPA
ALG1
FAM86A
ROGDI
GLYR1
UBN1
PPL
SEC14L5
LOC100507589
C16orf89
RBFOX1
TMEM114
PMM2
METTL22
ABAT
CARHSP1



Supplementary Figure 19. Enrichment of Differentially Expressed genes at dynamic interacting regions. The number of genes with a > 4-fold change in gene expression are that are found in a dynamic interacting region in either mouse ES cell or cortex are shown. Shown in grey is the number of > 4-fold changed gene expected using randomly permuted dynamic interacting regions.

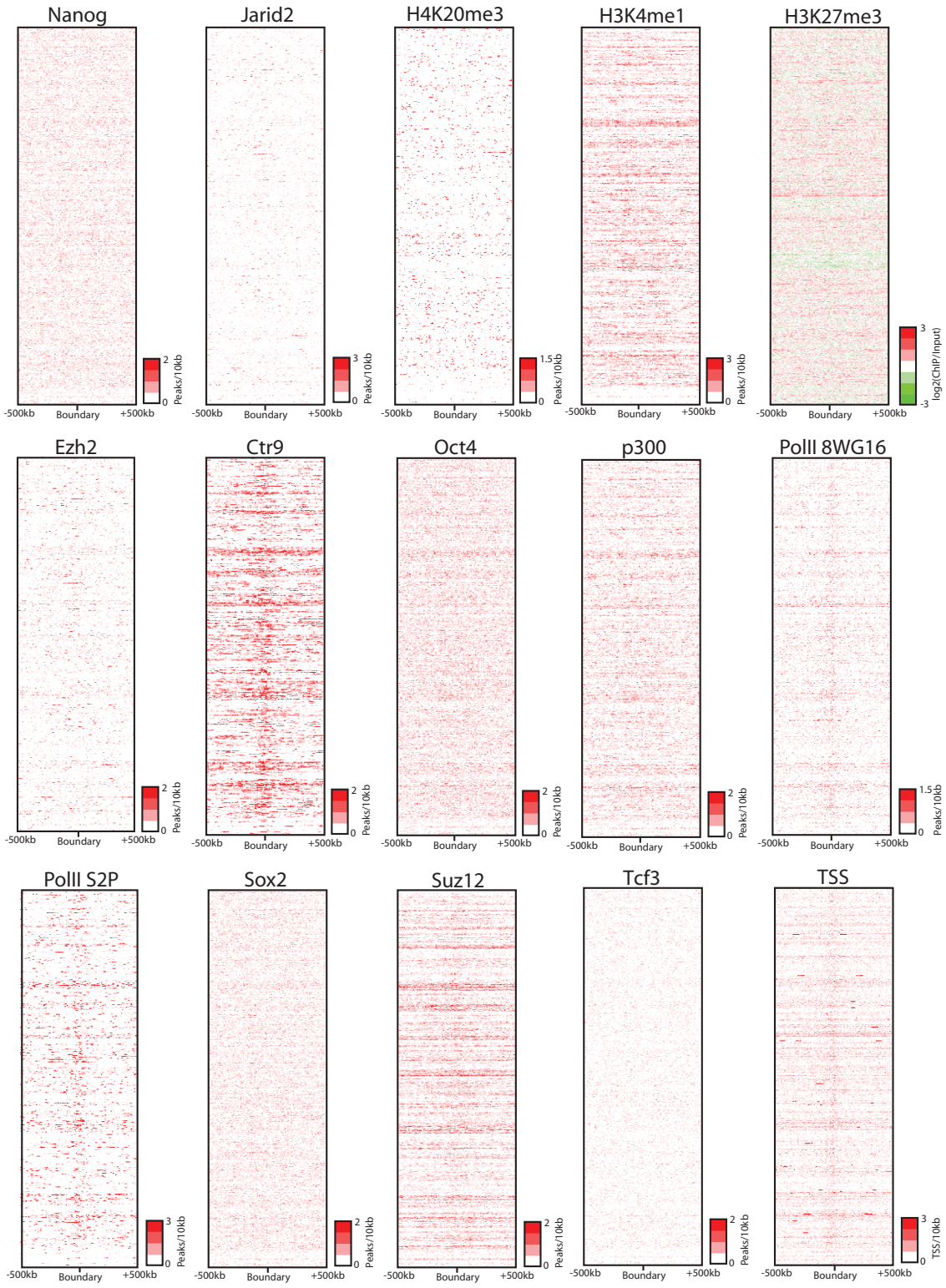


Supplementary Figure 20. Histone modification, chromatin binding protein, and transcription factor enrichment near boundary regions. Average enrichment plots for factors surrounding boundary regions called in mESC. For most marks, the signal is shown as the frequency of peaks or binding sites per 10kb. For "block like" marks, such as H3K27me3 and H3K79me2, the signal shown is the log2(ChIP/Input) over 10kb windows.

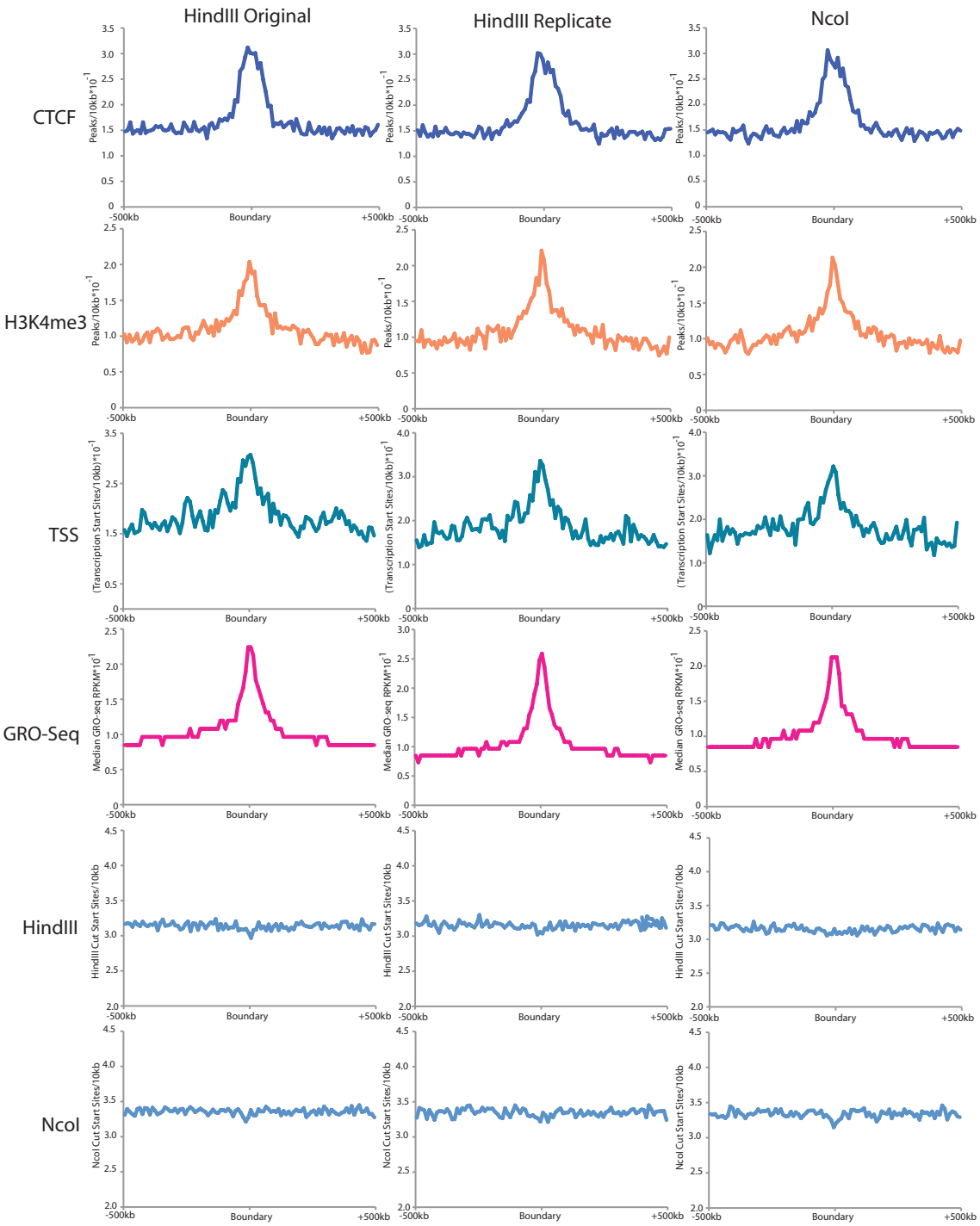
Supplementary Figure 21. Heat maps of boundary enrichment of Histone modification, chromatin binding protein, and transcription factor enrichment near boundary regions. Raw heat maps of each signal at the boundary region of a subset of marks from Supplemental Figure 20.

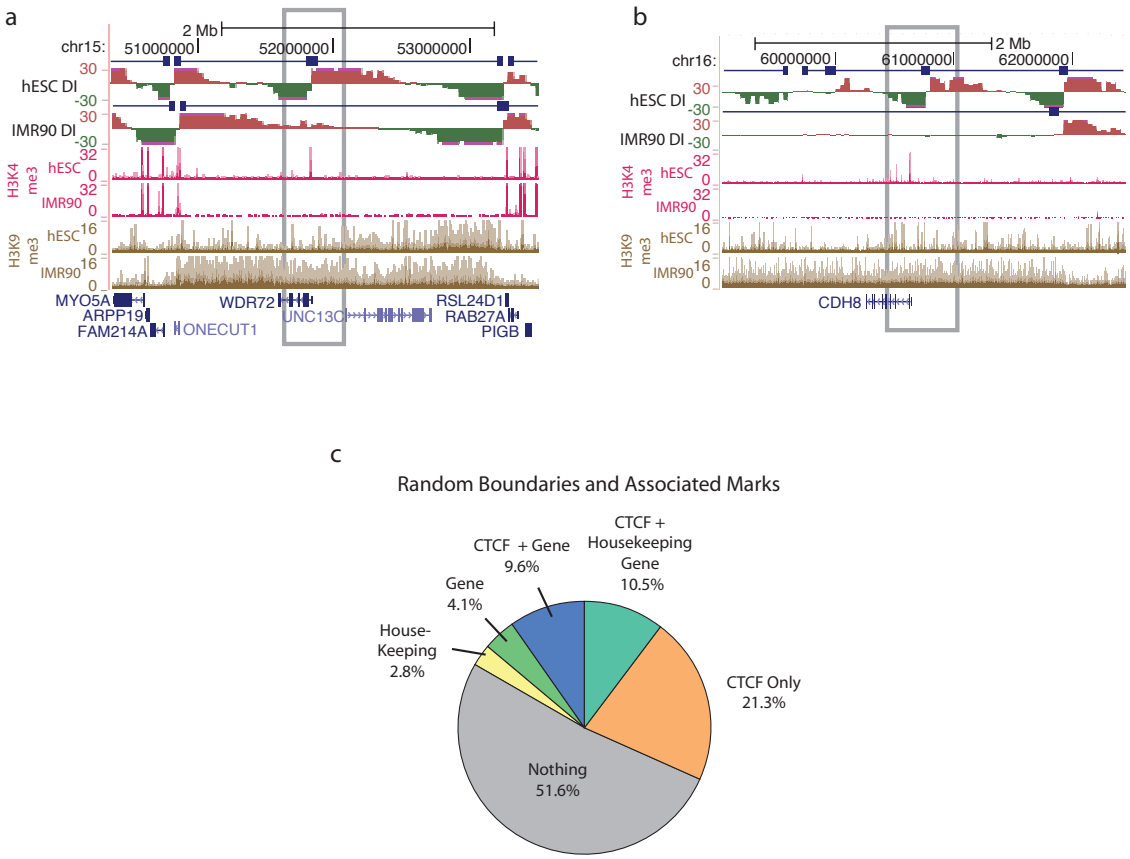

GRO Seq
H3K4me3
H3K36me3
NelfA
H3K79me2
Spt5
CTCF
Jarid1a
Pol II S5P
Smc1
Smc3
DNase I
Med1
Med12
H3K27Ac
-500kb
Boundary
+500kb
RPKM
Peaks/10kb
log2(ChIP/Input)

Supplementary Figure 22. Heat maps of boundary enrichment of Histone modification, chromatin binding protein, and transcription factor enrichment near boundary regions. Raw heat maps of each signal at the boundary region of the remainder of marks from Supplemental Figure 20 not shown in Supplementary Figure 21.

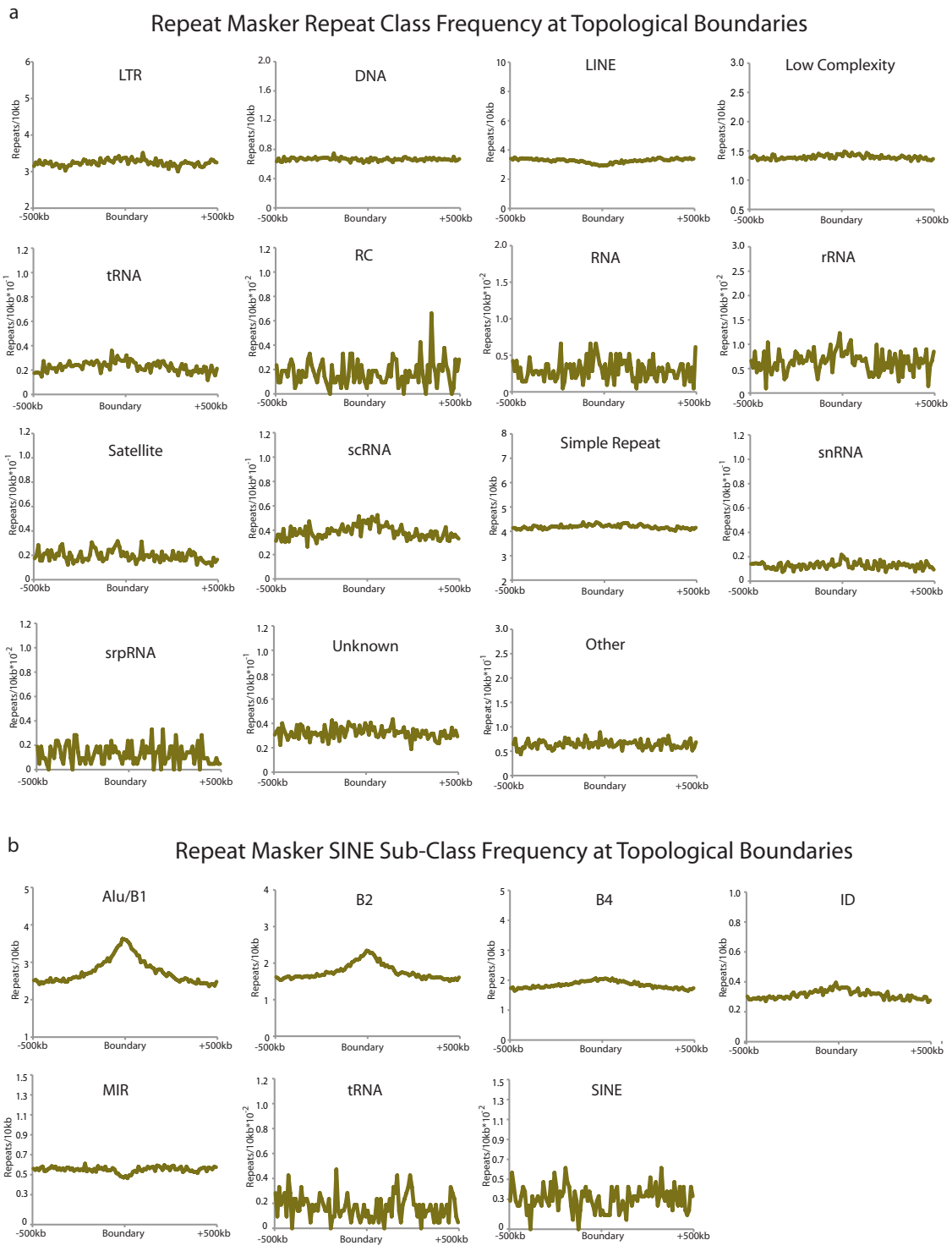

Nanog
Jarid2
H4K20me3
H3K4me1
H3K27me3
Ezh2
Ctr9
Oct4
p300
PolII 8WG16
PolII S2P
Sox2
Suz12
Tcf3
TSS
-500kb
Boundary
+500kb
Peaks/10kb
log2(ChIP/Input)
TSS/10kb

Supplementary Figure 23. Marks enriched at boundaries in each mouse ES cell replicate. The enrichement plots for CTCF, H3K4me3, transcription start sites, and GRO-seq signal were calculated similarly to Supplementary Figure 20 for each of the three mouse ES cell replicates. Also calculated and plotted is the average enrichment of HindIII and NcoI cut sites at the boundary regions.

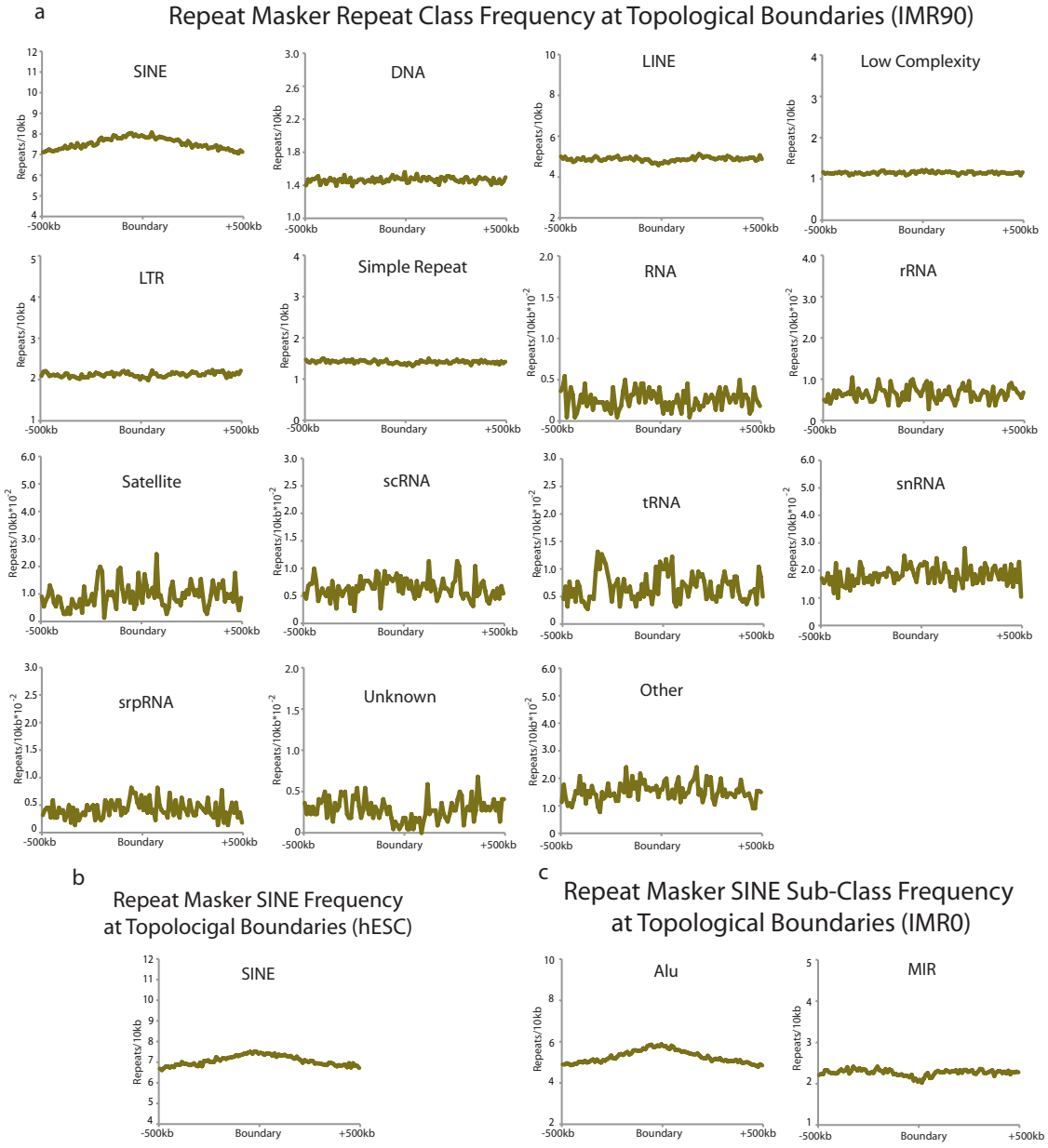

HindIII Original
HindIII Replicate
NcoI
CTCF
H3K4me3
TSS
GRO-Seq
HindIII
NcoI
Peaks/10kb*10-1
(Transcription Start Sites/10kb)*10-1
Median GRO-seq RPKM*10-1
HindIII Cut Start Sites/10kb
NcoI Cut Start Sites/10kb
-500kb
Boundary
+500kb



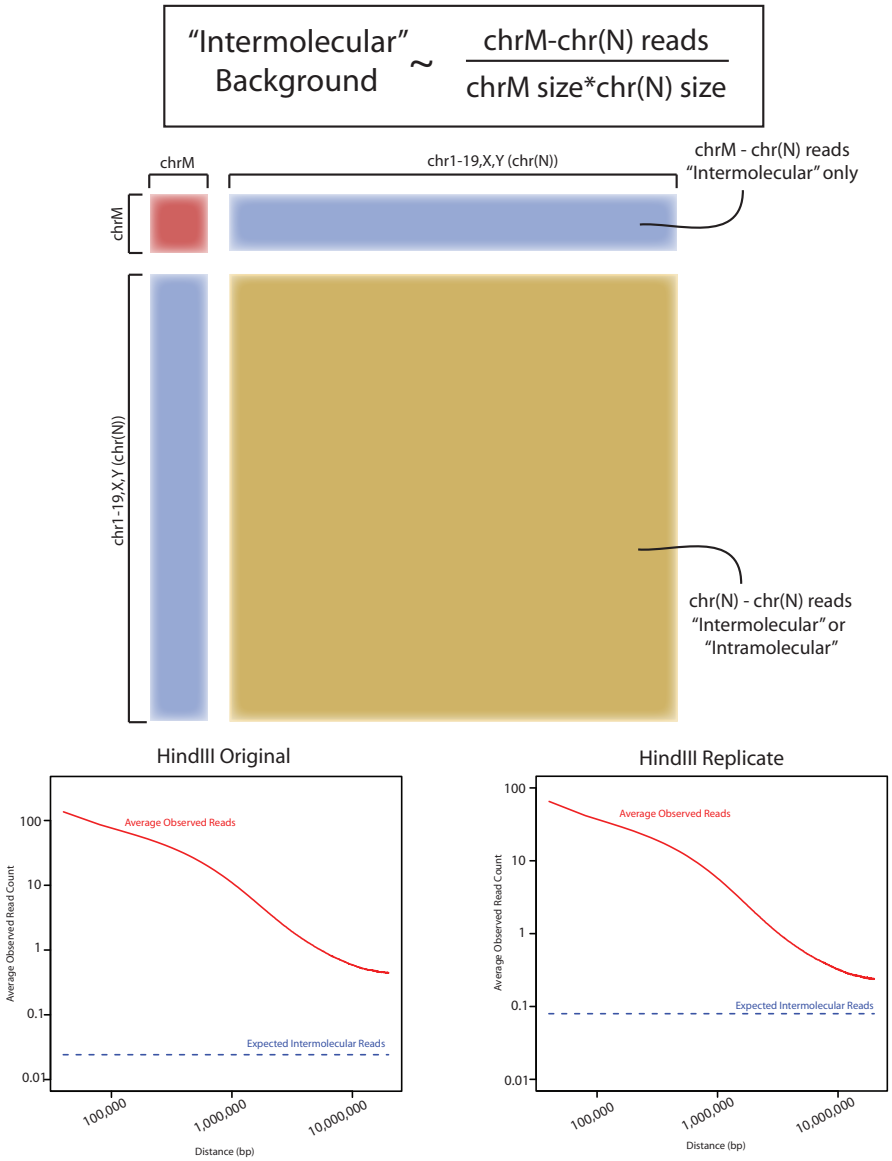
Supplementary Figure 24. Random assocation of CTCF and housekeeping genes in mESCs. a,b, Cell type specific boundaries between hESC and IMR90 that show associated changes in H3K4me3 near the boundary. c, Analogous to Figure 4e, pie chart showing the expected proportion of boundaries associated with CTCF, housekeeping genes, or other genes in mouse ES cells based on randomly generated boundaries.



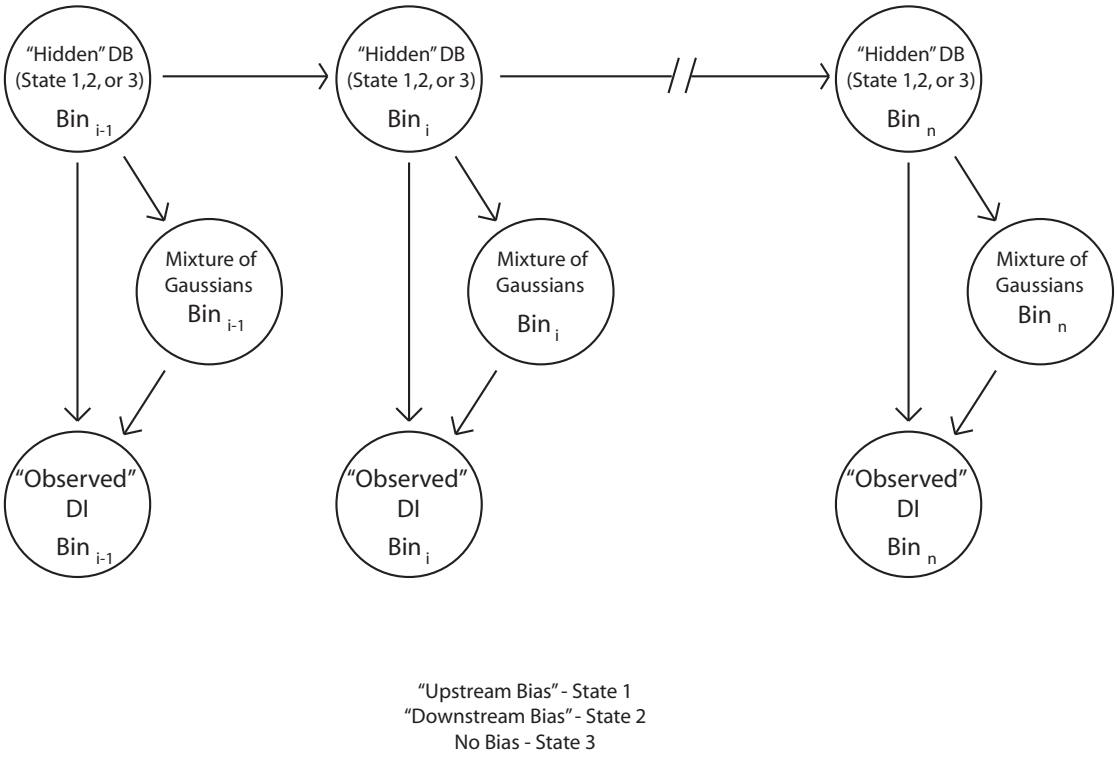
Supplementary Figure 25. Repeat Content at mouse ES cell boundaries. a, The frequency of repeats from UCSC Repeat Masker was calculated near the boundary regions. Only SINE element, shown in Figure 4a, show any enrichment at boundary regions. b, SINE subclass frequency at the topological boundary regions in mouse ES cells using UCSC Repeat Masker.



Supplementary Figure 26. Repeat Content at human boundaries. a, The enrichment of different classes of repeats at the IMR90 boundaries was calculated using the UCSC Repeat Masker data. b, Enrichment of SINE element frequency at boundaries in human ES cells. c, Enrichment of SINE element subclasses at the topological boundary regions in IMR90.



Supplementary Figure 27. Expected Intermolecular Ligations. To model the expected number of interactions between two loci in the genome due to random intermolecular ligation events, we calculated the expected number of reads per kbp^2 between the nuclear and mitochondrial chromosomes. As the nuclear and mitochondrial genomes are in different organelles, these reads can only occur due to random intermolecular ligations. We assume that the expected number of intermolecular reads between any two bins is constant, regardless of whether the two bins are nuclear or mitochondrial. Therefore, the number of intermolecular reads per bin between the nuclear and mitochondrial chromosomes should be equal to the number of intermolecular reads between any two bins both located on the nuclear chromosomes. Also shown is the number of reads at each distance (in red) for 40kb bins along the same chromosome. The number of random intermolecular reads is on average < 2% of what is actually observed for bins on the same chromosome less than 2 Mbp apart.



Supplementary Figure 28. HMM with mixture of Gaussian output. Each 40kb bin *i* along a chromosome having n bins has an observed Directionality Indexes ("Observed" DI) and a hidden Directionality Biases ("Hidden" DB, shown in the figure as states 1, 2, or 3 for simplicity). Assuming that the observed DI's are a mixture of Gaussians, we determine DB state (1, 2 or 3) at bin *i*.

Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in *Nature*, volume 485, May 17 2012. Dixon, Jesse R.; Selvaraj, Siddarth, Selvaraj; Yue, Feng; Kim, Audrey; Li, Yan; Shen, Yin; Hu, Ming; Liu, Jun S.; Ren, Bing. The dissertation author was the primary investigator and author of this paper.

Chapter 3. CTCF and Cohesin differentially affect higher order chromatin structure and gene expression.

Introduction

Recent studies of genome-wide chromatin interactions have revealed that the human genome is partitioned into many self-associating topological domains (1-4). The boundary sequences are enriched for binding sites of CTCF and the cohesin complex, implicating these two factors in the establishment or maintenance of topological domains (5-9). CTCF is a transcription factor that can bind to transcriptional insulator sequences and prevent enhancers from inappropriately activating non-target genes (10, 11), however, the exact mechanism of CTCF's insulator function is not well understood. Several observations have led to the proposal that CTCF might act via recruitment of cohesin to facilitate long-range interactions as "master weaver" of the genome (11). First, CTCF and the cohesin complex, consisting of the core subunits SMC3, SMC1, RAD21 and STAG1/SA1 or STAG2/SA2, were found to colocalize extensively throughout mammalian genomes (8, 12, 13). Second, both factors are involved in mediating long-range interactions (5-7, 14-16). Finally, cohesin was shown to be important for CTCF's chromatin insulation function(8, 12, 13), while CTCF is necessary to recruit cohesin to the shared binding sites but not to chromatin (8).

Results

To understand the contribution of CTCF and cohesin to genome organization, we employed an engineered HEK293T cell line (original cell line derived from human

embryonic kidney) in which we can rapidly remove the cohesin complex from interphase chromosomes by proteolytic cleavage of its RAD21 subunit (17). This cell line contains an episome-based vector that allows doxycycline-inducible expression of siRNA targeting endogenous RAD21 and a RAD21-EGFP variant containing a recognition site for Human rhinovirus 3C (HRV) protease (RAD21cv) (18) (Fig. 1a,b). Three days after doxycycline induction, RAD21cv completely replaces the endogenous RAD21 and is incorporated in the cohesin complex (Supplementary Fig. 1). Subsequent transfection of the cells with a construct expressing HRV protease led to full cleavage of RAD21cv and release of cohesin from chromatin within 24 hours (Fig. 1 c,d,e). Consistently, RAD21cv cells entering mitosis 24 hours after HRV transfection show increased defects in sister chromatid cohesion (Supplementary Fig. 2). Cleavage of RAD21cv by HRV protease (RAD21cv/HRV) does not change the cell cycle distribution compared to transfection with a control protease (RAD21cv/TEV). Nevertheless, we noted that both transfected cell populations have more cells in G2 phase than untreated cells (RAD21cv) (Supplementary Fig. 3). This rapid release of cohesin allows the study of the immediate effect of cohesin loss on chromatin structure, without interfering with cohesin function in cell division.

To test whether removal of cohesin from chromatin affects long range chromatin interactions, we performed 3C-seq (a multiplexed 4C variant) (19) in RAD21cv/TEV and RAD21cv/HRV cells. We examined the interior and the borders of one topological domain at the well characterized chr11p15.5 region comprising H19, Igf2 and other imprinted genes (referred to here as H19/IGF2 domain, Fig. 1f-h). We have previously used this region to establish the role of cohesin in chromatin insulation by CTCF (8). In

RAD21cv/TEV we observed, as reported before (5), that the IGF2 promoter region (VP1) and an intergenic region (VP2) interact strongly. Further contacts persist over a 500kb region until the proximal keratin cluster (KRTA5) marking the domain boundary. Viewpoints placed between H19 and IGF2 (VP2) and upstream of H19 (VP3) confirm these interactions (Fig. 1g). A viewpoint in the neighbouring domain at the centromeric side (VP6) consistently shows interactions until the domain boundary (Supplementary Fig. 4). A viewpoint placed at the telomeric boundary (VP5) shows weak interactions with both domains. The CTSD gene residing in a cohesin-depleted region is remarkably excluded from interactions; although a 3C-seq viewpoint there (VP4) shows some of the interactions detected by the other viewpoints (Fig. 1g). We observed similar interaction profiles in the breast endothelial cell line 1-7HB2 (abbreviated HB2) with normal karyotype, indicating their conservation between cell lines (Supplementary Fig. 5).

Cleavage of RAD21 leads to a global loss of interactions across the entire domain at all viewpoints (Fig. 1g). A control with a cell line lacking the HRV cleavage site in RAD21-EGFP (RAD21wt) did not show altered cohesin binding and long-range interactions after transfection with the cleavage protease (Supplementary Fig. 6). These results strongly support that cohesin is required for the higher order chromatin structure within this domain.

To investigate whether cohesin plays a general role in topological domain organization, we performed Hi-C experiments with control RAD21cv/TEV cells and RAD21cv cells after RAD21 cleavage (RAD21cv/HRV). We obtained greater than 370 million non-redundant uniquely mapping read pairs for both control and RAD21 cleaved cells, split between two replicates for each condition. We normalized the Hi-C interaction

frequencies according to the iterative correction method (20). For each replicate both before and after RAD21 cleavage, we identified the location of topological domains using a previously described algorithm (2). We also performed ChIP-seq for the cohesin subunit SMC3 in control (RAD21cv/TEV) cells to determine the cohesin binding sites in the genome in these cells. Similarly to what had previously been observed for CTCF, cohesin appears to be enriched at the borders or boundaries between domains (Fig. 2a). Notably, only SMC3 sites that co-localize with CTCF show enrichment at boundaries, while CTCF-independent SMC3 sites show no enrichment at boundary regions (Supplementary Fig. 7a). To compare Hi-C interaction frequencies with SMC3 binding, we separated the genome into 40kb interacting bin-pairs and stratified them according to if each bin in the pair is bound by SMC3 ("SMC3 2x"), or if only one bin ("SMC3 1x") or no bins ("None") were bound by SMC3 (Fig. 2b). We observed a higher interaction frequency in control Hi-C experiments between bin-pairs containing SMC3 sites on both ends than when only one or no SMC3 site is present (Fig. 2c), consistent with the notion that cohesin binding could mediate long-range chromatin interaction frequencies genome-wide. Upon cleavage of RAD21, we observed an overall loss in local chromatin interaction frequency primarily occurring at distances up to 2 Mb, with a maximum in the range between 100-200kb (Fig. 2d, Supplementary Fig. 7b). The loss in interaction frequency is highest after RAD21 cleavage when both interacting loci are bound by SMC3 (Fig. 2d inset). Using DNA-FISH (21) we observed a spatial separation of cosmid probes placed in the H19/IGF2 domain (Fig. 2e,f) and in the HOXD domain (Fig. 2g,h) after RAD21 cleavage, consistent with the aforementioned loss of interactions.

We next investigated the effects of cohesin complex destruction on topological domain organization. Surprisingly, the positions of most topological domains do not markedly change upon cleavage of RAD21 (Fig. 2i,j). The “triangle” pattern of topological domains is still readily apparent in the interaction heat maps, and, though we consistently call fewer domains in the RAD21 depleted cells (Fig. 2j), there is a strong overlap in domain boundaries called between control and RAD21 depleted cells. The preservation of topological domains after RAD21 cleavage is consistent with live cell imaging observations of histone H2A-RFP in RAD21cv/TEV and RAD21cv/HRV cells showing no general changes of chromatin morphology after RAD21 cleavage (Supplementary Fig. 8). However, consistent with the previously described general loss in interaction frequency, we also observed a clear reduction in interaction frequency both within and between domains after RAD21 depletion (Supplementary Fig. 9). Interestingly, the degree of depletion in interaction frequency within domains is most marked when one or both interacting bins is associated with a boundary region (Supplementary Fig. 9). Taken together, these results suggest that cohesin contributes to the self-association within topological domains by promoting interactions between regions near the boundaries. However, cohesin depletion does not appear to contribute to the positioning and segregation of neighbouring domains from each other.

To determine CTCF’s role in mediating chromatin interactions and to compare it to the effects of RAD21 cleavage, we performed two replicates of Hi-C experiments for CTCF and control siRNA knockdowns in HEK293T cells (Supplementary Fig. 10a,b). We obtained between 95 and 288 million unique reads for each replicate. Similar to SMC3, CTCF is enriched at the boundaries of topological domains in control cells (Fig.

3a). Likewise, CTCF binding correlates with the strength of Hi-C interaction frequency, where interacting bin-pairs bound by CTCF on each side form stronger interactions compared to regions with only one or no CTCF sites (Fig. 3b). Upon knockdown of CTCF, we observed a loss of interactions within topological domains, but with a different pattern with respect to the distance between interacting loci compared to RAD21 cleavage (Fig. 3c,d). RAD21 depletion appears to most markedly affect interacting loci separated by 100 to 200kb (Fig. 2d), while CTCF knockdown appears to most prominently affect interacting loci separated by less than 100kb (Fig. 3b). This implies that CTCF and cohesin may affect intra-topological domain interaction frequency on different spatial scales (Figs. 3e,f blue line). The more remarkable difference between CTCF and cohesin depletion concerns the interactions between topological domains. RAD21 depletion leads to a loss in interactions within and between domains but primarily in intra-domain interactions (Fig. 3e yellow line). CTCF depletion, on the other hand, leads to a significant gain of interactions between neighbouring domains (Fig. 3f yellow line). This increase in interaction frequency is seen at nearly all distance scales between interacting loci, suggesting that CTCF is necessary to maintain topological domain boundaries throughout the genome. The interactions gained by CTCF depletion could involve delocalised cohesin which now forms “non-specific” interactions, as we have previously shown that cohesin is delocalised but still present on chromatin after CTCF knockdown (8) (Supplementary Fig. 10c). Indeed, we observed that the largest gains of inter-domain interaction frequency after CTCF knockdown occur between bins containing CTCF or cohesin sites (Fig. 3g,h). Altogether our observations suggest that

cohesin and CTCF are both important in shaping genomic structure on the level of topological domains in a non-redundant manner.

The above observations suggest that loss of cohesin or CTCF affects chromatin structure in different ways. Given the intimate relationships between chromatin structure and gene regulation, we would predict that loss of these two factors would differentially affect gene expression. To test this prediction, we performed RNA-sequencing (RNA-seq) in the control (RAD21cv/TEV) and RAD21 depleted (RAD21cv/HRV) cells as well as CTCF RNAi and mock treated cells. In both cases we observed only modest changes in gene expression (Supplementary tables 1-4), consistent with earlier observations (8). We observed 48 and 161 differentially expressed genes (FDR <5%) for RAD21 and CTCF depletion, respectively but very little overlap between these sets (Supplementary table 1, Fig. 4a). Among the genes with reduced expression after RAD21 depletion are several Hox genes (HOXA11AS, HOXA-AS3, HOXB-AS3, HOXB5, HOXC9) (Fig. 4b, Supplementary Fig. 11). We validated the reduced expression of HOXB-AS3, HOXA-AS3 and H19 by RT-PCR and qPCR (Fig. 4c). Hox genes have been shown to be regulated by antisense transcription as well as the topological organization of the locus (22, 23), but have never been reported to depend on cohesin.

Among genes which are differentially expressed after CTCF depletion, we observed a clear enrichment of CTCF binding at their promoters (Fig. 4e), with a median distance from the TSS to the nearest CTCF binding site being only 191bp (Supplementary Fig. 12). In contrast, genes that are differentially expressed after cohesin depletion are not directly bound at their promoter by SMC3 (Fig. 4f), though they are located closer to SMC3 binding sites than would be expected at random (median distance

~4kb, Supplementary Fig. 12). This indicates that altered expression of genes after RAD21 cleavage may be a product of higher order chromatin structural changes. To validate this, we analysed interactions of cohesin-regulated genes with DNase hypersensitive sites as markers for potential distal gene regulatory regions at a restriction fragment level resolution. We observed that cohesin regulated genes lose more interactions with distal DNaseI hypersensitive sites than with non-cohesin regulated Refseq genes (Fig. 4g). These results suggest that cohesin may regulate gene expression by affecting the interaction frequency of genes with distal regulatory elements, while CTCF may directly regulate genes by binding at their promoters.

In summary we show for the first time how cohesin and CTCF contribute to the topological domain architecture of the human genome. We observed a loss of interactions within and also between domains after cohesin cleavage. On the contrary, CTCF depletion reduces the intra-domain interactions at a somewhat shorter distance while leading to a gain of interactions across domain boundaries. This suggests that cohesin is mainly involved in the self-association property of domains while CTCF is important for their spatial segregation (Fig. 4h-j). We hypothesize that CTCF maintains boundaries by determining cohesin localization, and, in the absence of CTCF, cohesin might form ”non-specific” interactions reaching beyond boundaries.

Consistent with these differential contributions to the overall architecture, we observed different sets of genes changing after cohesin removal or CTCF depletion. For CTCF, a direct role for transcription is emerging since promoters of CTCF-dependent genes are often bound by CTCF. Genes differentially expressed after RAD21 cleavage have no cohesin bound to their promoters and might be primarily regulated by distal

regulatory sequences whose interaction with their target promoters is impaired after chromatin structural changes. This is in good agreement with previous observations suggesting a role for cohesin in enhancer function (8, 9, 24, 25). Taken together, these results provide an initial model for understanding the mechanisms of higher-order chromatin organization and its relationship to gene regulation.

Methods summary:

RAD21 cleavage experiments. HEK293T stable cell lines containing episomes coding dfor RAD21cv or RAD21wt were grown for 3 days in presence of doxycycline until the endogenous RAD21 was replaced by the engineered RAD21 versions, transfected with either control protease (TEV) or cleavage protease (HRV) according to the manufacturer's instructions and harvested after 24 hours. Cells were then prepared according to the experimental protocols.

Cleavable HRV RAD21-eGFP construct (RAD21cv): The construct encoding the cleavable RAD21 subunit (RAD21cv) was previously described in (26). Briefly, the first RAD21-separase cleavage site was replaced by one for 3C protease of the human rhinovirus (HRV protease) using a PCR-based mutagenesis. The second cleavage site was unchanged to ensure less cell cytotoxicity. RAD21cv was then cloned in front of an EGFP cassette. The tobacco etch virus protease (TEV protease), which does not recognize the HRV cleavage site is used as control.

siRNA cassette for the endogenous RAD21: For the knock-down of the endogenous RAD21 subunit, the following 3'UTR-directed siRNA were used:

5′-ACUCAGACUUCAGUGUAUA-3′ (Scc1-1),

5′-AGGACAGACUGAUGGGAAA-3′ (Scc1-2).

Episomal system: The vector pRTS-1, described in (27), presents a doxycycline-responsive cassette composed of a bidirectional promoter that drives the expression of

two genes in a coordinated fashion. The cleavable RAD21-eGFP and the siRNA for RAD21 cassettes were cloned under the control of the bidirectional promoter; thus after doxycycline treatment both cassettes are expressed simultaneously.

Generation of HEK293T cell stably containing the episomal constructs: HEK293T cell line was cultured in DMEM supplemented with 0.2mM L-glutamine, 100 units/ml penicillin, 100 mg/ml streptomycin and 10% FCS and was grown at 37°C and 5% CO_2. Transfection of the episome was done by Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions and cells carrying the episome vector were selected by growing in a medium containing 150 µg/mL hygromycin. Single clones were picked and analysed for expression of RAD21cv and RAD21wt constructs and depletion of the endogenous RAD21 three days after induction with 2 µg/ml doxycycline.

RAD21 cleavage experiments: To activate transgene expression, cells were cultured for 3 days in the presence of 2 µg/ml of doxycycline. After 3 days, cells were split to 50% confluency and transfected with TEV or HRV construct using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. Cells were harvested 24 hours after protease transfection.

Spreading of mitotic chromosomes: Cells were treated with nocodazole (Sigma) for 2 hours and fixed with methanol/acetic acid after hypotonic treatment. After spreading of the cells on cover slips the chromosomes were stained with Giemsa.

Live cell imaging of Histone-RFP after RAD21 cleavage: To activate transgene expression, cells were cultured for 3 days in the presence of 2 μg/ml of doxycycline. After 3 days, cells were split to 50% confluency and transfected with TEV or HRV constructs and H2A-RFP using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. Cells were imaged 24 hours after protease transfection using a SpinD1454 Roper/Nikon spinning disk microscope with temperature controller. Cells were imaged using the 60X Objective, 491 nm and 561 nm lasers and 700 ms exposure time. Image stacks were processed with ImageJ and projected in a single plane.

RNAi depletion of CTCF and RAD21: HEK293T cells were seeded in DMEM supplemented with 0.2mM L-glutamine and 10% FCS and transfected with siRNA oligos (Ambion) directed against CTCF and a non-targeting control siRNA using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions.

Cells were harvested 48 hours after transfection according to the respective protocols for Hi-C or western blotting.

The following siRNA oligos were used:

CTCF siRNA	sense	GGAGCCUGCCGUAGAAAUUTT
	antisense	AAUUUCUACGGCAGGUCCTC
Control siRNA	sense	CGUACGCGGAAUACUUCGATT
	antisense	UCGAAGUAUUCCGCGUACGTT

Depletion of RAD21 in HEK293T by siRNA transfection was only performed for the transcript analysis with the same protocol used for CTCF depletion.

RAD21 siRNA sense GGUGAAAAUGGCAUUACGGtt

antisense CCGUAAUGCCAUUUUCACCtt

Fractionation in soluble and chromatin bound proteins: To prepared soluble and chromatin bound fractions cells were harvested and lysed (20mM Tris-HCl pH 7.5, 100mM NaCl, 5mM MgCl2, 2mM CaCl2, 10% glycerol, 0.2% NP-40, 1mM NaF, 0,5mM DTT and protease inhibitors. An aliquot was taken as total lysate and the remaining lysate centrifuged 10 min at 1500 rpm to collect the chromatin pellet. The supernatant was collected as soluble fraction. The pellet was wash 3 times with lysis buffer, resuspended in TBS/T and the chromatin-bound proteins solubilized by sonication and benzonase treatment.

Antibodies: Primary antibodies used were mouse anti-CTCF (BD, for immunoblotting), rabbit anti-CTCF (Millipore, for ChIP), mouse monoclonal anti-EGFP (Sigma, immunoblotting), mouse anti-tubulin (Sigma, immunoblotting) and rabbit anti-TopoII (Millipore, immunoblotting). Polyclonal rabbit antibodies against RAD21 and STAG1/STAG2 (immunoblotting) were a gift from Jan-Michael Peters and described in (28). Rat monoclonal antibodies against SMC1 (immunoblotting) were a gift from Niels Galjart. Polyclonal rabbit antibodies against the human SMC3 (ChIP) were raised against the peptide (C-EMAKDFVEDDTTHG) as described before(28) (Absea, China) and purified using the peptide antigen. Polyclonal rabbit antibodies against EGFP were raised against recombinant EGFP produced in *E. coli* (Absea, China) and purified using the protein antigen.

Immunoprecititation: Immunoprecipitations were performed as described(29) using antibodies against SMC3 and EGFP.

Chromosome conformation capture sequencing (3C-seq) and analysis: Chromosome conformation capture sequencing was performed as previously described in(30, 31). Briefly, cells were crosslinked with 1% (w/v) formaldehyde for 10 minutes and quenched with 120mM glycine. Crosslinked-cells were resuspended in lysis buffer (50mM Tris-HCl pH 8.0, 0.5% NP-40, 50mM NaCL and Complete protease inhibitor (Roche)) and subjected to enzymatic digestion using 400 units of BglII (Roche). Digested chromatin was then diluted and ligated using 5 units of T4 DNA ligase (Promega) under conditions favouring intramolecular ligation events. After reversing the crosslink at 65°C over night, the digested and ligated chromatin was subjected to a second enzymatic digest using NlaIII (New England Biolabs) to produce smaller DNA fragment.. The resulting digested DNA underwent a second ligation using 10 units of T4 DNA ligase (Promega) under conditions favouring self-ligation events that produce circular DNA molecules. The unknown DNA fragment, ligated to the fragment of interest (called viewpoint), was amplified by inverse-PCR using specific primer design in the outer part of the restriction site of the viewpoints, linked with the Illumina adapter sequences. The samples were then single-read sequenced using the Illumina Genome Analyzer II generating 76bp reads. The reads were trimmed to remove the illumina adapter sequences and mapped against human genome (hg18). The reads were extended to 56bp in the 3' direction using the r3C-Seq pipeline (Thongjuea, Stadhouders et al., in preparation ; http://www.bioconductor.org/packages/2.11/bioc/html/r3Cseq.html). Interaction

frequencies were calculated using the number of reads per million (RPM). The data were visualized using a local UCSC mirror browser.

Cell cycle analysis: Cells were fixed with methanol and after RNAse treatment the DNA was stained with propidium iodine. The cells were analyzed with a BD FACS Aria Cell sorter and FlowJo software.

Chromatin immunoprecipitation (ChIP): Chromatin immunoprecipitation was performed as described before (Wendt, 2008). In brief, cells at 70–80% confluence were crosslinked with 1% formaldehyde for 10 minutes and quenched with 125mM glycine. After washing with PBS, cells were resuspended in lysis buffer (50mM Tris-HCl pH 8.0, 1% SDS, 10mM EDTA, 1mM PMSF and Complete protease inhibitor (Roche)) and chromatin was sonicated (Diagenode Bioruptor). After a centrifugation step to remove the debris. the lysate was diluted 1:4 with IP dilution buffer (20mM Tris-HCl pH 8.0, 0.15 M NaCl, 2mM EDTA, 1% TX-100, protease inhibitors) and precleared with Affi-Prep Protein A support beads (BioRad). The respective antibodies were incubated over night with the lysate at 4°C, followed by 2 hours incubation at 4°C with blocked protein A Affiprep beads (Bio-Rad) (blocking solution: 0.1 mg/ml BSA or 0.1 mg/ml fish skin gelatine). The beads were washed with washing buffer I (20mM Tris-HCl pH 8.0, 0.15 M NaCl, 2mM EDTA, 1% TX-100, 0.1% SDS, 1mM PMSF), washing buffer II (20mM Tris-HCl pH 8.0, 0.5 M NaCl, 2mM EDTA, 1% TX-100, 0.1% SDS, 1mM PMSF), washing buffer III (10mM Tris-HCl pH 8.0, 0.25 M LiCl, 1mM EDTA, 0.5% NP-40, 0.5% sodium deoxycholate) and TE-buffer (10mM Tris-HCl pH 8.0, 1mM EDTA). The

beads were eluted twice (25mM Tris-HCl pH 7.5, 5mM EDTA, 0.5% SDS) for 20 minutes at 65°C. The eluates were treated with proteinase K and RNase for 1 hour at 37°C and decrosslinked at 65°C over night. The samples were further purified by phenol-chloroform extraction and ethanol-precipitated. The pellet was dissolved in 50µll TE buffer. CTCF ChIP-seq data in the HEK293T cell line was downloaded from the ENCODE consortium.

ChIP sequencing and peak detection: The ChIP DNA library was prepared according to the Illumina protocol (www.illumina.com). Briefly, 10 ng of ChIPped DNA was end-repaired, ligated to adapters, size selected on gel (200±25 bp range) and PCR amplified using Phusion polymerase as follow: 30sec at 98ºC, 18 cycles of (10sec at 98ºC, 30sec at 65ºC, 30sec at 72ºC), 5min at 72ºC final extension. Cluster generation was performed using the Illumina Cluster Reagents preparation. The libraries were sequenced with the Illumina HiSeq2000 system. Read lengths of 36 bases were obtained. Images were recorded and analyzed by the Illumina Genome Analyzer Pipeline (GAP 1.6.0. and 1.7.0.). The resulting sequences were mapped against Human_UCSChg18 using the Bowtie(32) alignment software, with the following parameters: -v 3 -m 1 --best --strata -S --time -p 8. Unique reads were selected for further analysis. PCR duplicate reads were removed using Picard MarkDuplicates.

ChIP-qPCR: ChIP samples (2µl) were used for 25µl PCR reaction. Analyses by qPCR were performed using Platinum Taq and SYBR Green (Invitrogen) on ABI 9500 cycler. The results were presented as the percentage of input-chromatin that was precipitated.

Transcript analysis by reverse transcription (RT) and qPCR: Total RNA was prepared using TRIzol Reagent (Invitrogen) according to manufacturer's instruction. After chloroform extraction and isopropanol precipitation, pellets were dissolved in DEPC water. cDNA was generated by reverse transcription using oligo(dT)18 primer (Invitrogen), Superscript II Reverse Transcriptase (RT) (Invitrogen) and RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen) according to the manufacturer's instructions. The DNA was then purified using PCR purification Kit (Qiagen) according to the manufacturer's instructions. The amounts of the different transcripts were compared by qPCR using SYBR Green and Platinum Taq Polymerase (Invitrogen) in CFX96 lightcycler (BioRad) and specific primers. The transcripts of the housekeeping gene SNAPIN were used for normalization of the samples.

Probe labelling for Fluorescence in situ hybridization (FISH): Cosmids for the human chromosome 11 (G248P89139F12 and G248P85529B4) and for the human chromosome 2 (G248P85616F7 and G248P80003A6) were obtained from the CHORI library. 500ng cosmid DNA was labelled with Alexa 488-5-dUTP or Alexa 594-5-dUTP (Invitrogen) using the BioPrime Random Prime Labeling kit (Invitrogen). After labelling, probes were purified from unincorporated nucleotides using Sephadex G-50 column elution. Fractions containing the labelled cosmids DNA were pooled, ethanol precipitated and dissolved in hybridization mix containing 50% deionized formamide (Sigma), 2XSSC, 100 mM phosphate buffer pH 7.5, 5x Denhardt solution, 5% dextran sulphate (Sigma) to a final concentration of 500ng/ml.

3D DNA Fluorescence in situ hybridization (3D DNA-FISH): 3D DNA-FISH was performed as described before(33). In brief, cells were grown on 18mm poly-D-lysine (Sigma) coated coverslips (VWR) , fixed with 2% (w/v) formaldehyde/1×PBS for 10 minutes and permeabilized in 1×PBS containing 0.5% (v/v) Triton X-100 (Sigma) and 0.5% Saponin (Sigma) for 10 minutes at room temperature. Next cells were treated with 0.1N HCl for 5 minutes at room temperature. Cosmids probe mixes with a final concentration of 4-10ng/ml each and 50x excess of human Cot I DNA (Sigma) were added to the slides. Probes and cells were denatured simultaneously at 70°C for 2 minutes on a hot plate and hybridized overnight at 37°C in a humidified chamber. After hybridization slides were washed with 2xSCC at 37°C for 30 minutes and one time with 2xSCC at RT for 15 minutes. Coverslips were mounted with Prolong Gold contained DAPI (Invitrogen).

Confocal Laser Scanning Microscopy: All cell samples were imaged using a Leica SP5 confocal laser scanning microscope using the LAS software provided with the instrument. The system was equipped with a 63x plan-apochromat oil NA1.4 DIC objective. The pinhole diameter was set to 1 airy unit. DAPI, Alexa 488 and Alexa 594 fluorochromes were excited with a 405nm diode laser, a 488nm Argon laser and a 594nm laser respectively and detected using a multi-track imaging mode of which the band pass filters were 410-450nm (DAPI), 505-585nm (Alexa 488) and 605-700nm (Alexa 594). 8 bit images with a 512 x 512 pixels frame size and 51x51nm pixel size were acquired with 400Hz scan speed, 2-times line averaging and an optical sectioning of 120nm. The point

spread function was measured using 100nm red and green beads (Thermo Scientific) and the chromatic shift was measured using 500nm TetraSpeck beads (invitrogen).

All confocal images were deconvolved using the Huygens Professional software v4.1.0p8 (SVI) using the measured Point Spread Function and the classical maximum likelihood-estimation algorithm. The background, signal to noise ratios and chromatic shift were corrected during the deconvolution process.

RNA-seq experiments and data analysis: Total RNA was isolated using TRIzol reagent (Invitrogen). PolyA RNA was isolated using Dynal beads mRNA purification kit (Invitrogen), and paired-end libraries were prepared as previously described(34).

Reads were aligned to hg18 using Tophat with the following parameters: -g 1 -p 12 --solexa1.3-quals --library-type fr-firststrand --segment-length 25 --bowtie1. A GTF file for UCSC genes was provided for the initial alignment. Wig files were generated using an in house pipeline. We normalized each wig file using trimmed mean of M normalization(35) using Refseq exons to calculate the scaling factors between experiments.

Read counts for RefSeq genes were calculated using an in house pipeline and differentially expressed genes were called using edgeR. Common and tagwise dispersions were estimated based on all 8 RNA-seq experiments performed (2 replicates for each of 4 condiations: RAD21cv/TEV, RAD21cv/HRV, siRNA CTCF, siRNA Control). Differentially expressed genes were called by an FDR <5%.

Hi-C experiments and data analysis: Hi-C experiments were performed as previously described (1). Reads were aligned as single end reads using bwa with default parameters against hg18 reference genome. Single end reads were filtered for uniquely mapping reads and paired manually using an in house pipeline. Hi-C interaction matrices were generated as previously described (1) and normalized using the iterative correction method either using 40kb bins or at a restriction fragment based level (for Figure 4g) (36). To facilitate comparison of Hi-C interaction frequencies between different experiments, interaction matrices were also normalized for "depth," with the normalized interaction frequency (I_{ij}) between two loci *i* and *j*, being normalized by the sum of all I_{ij} in a given chromosome wide normalized interaction matrix. This is analogous to read-depth based normalization schemes (i.e. RPKM) of other high-throughput sequencing experiments. These normalized interaction matrices serve as the input for generating the directionality index and topological domain calls using previously described methods (37).

To generate the "delta" interaction matrices (Fig. 3c,d), we subtracted the normalized interaction frequency I_{ij} at each locus of an experimental treatment (RAD21cv/HRV or siRNA CTCF) from the control treatment (RAD21cv/TEV or siRNA Control) to generate a new ΔI_{ij} for comparison between experiments.

For analysis of interactions between promoters and distal DNaseI Hypersensitive (DHS) sites, we used DHS sites from the ENCODE consortium and used the UCSC liftover tool to convert these coordinates into hg18. We identified restriction fragments containing a DHS site greater than 5kb away from any RefSeq promoter, and considered all possible interactions between these fragments and restriction fragments containing a

RefSeq promoter that were within 500kb of each other. We computed the fold-change in interaction frequency between the control (RAD21cv/TEV) and RAD21 depleted (RAD21cv/HRV) samples and calculated the fraction of potential promoter-to-DHS interaction that showed a 50% gain or reduction in interaction frequency. Fisher's exact test was used to asses the enrichment of cohesin regulated genes versus all genes for a loss or gain of promoter-to-DHS interactions.

Figure 1. Cohesin cleavage reduces long-range interactions within the H19/IGF2 domain. a, Endogenous RAD21 is replaced with RAD21cv using a doxycycline-inducible bidirectional promoter driving expression of RAD21cv and siRNA targeting the endogenous RAD21 from an episomal construct stably integrated in HEK293T cells. b, Outline of the experiment showing the replacement of RAD21 by RAD21cv and transfection of the protease HRV, which cleaves RAD21cv, and TEV, which does not. c, Time course showing full cleavage of RAD21cv after 24 hours by detecting the C-terminal EGFP-tag in RAD21cv. The shift is consistent with the loss of a 20 kD fragment form the N-terminus of RAD21cv. Note RAD21cv gets more abundant in the lysates due to its release from chromatin after cleavage. d, Fractionation of the lysates from RAD21cv/TEV (TEV), RAD21cv/HRV (HRV) and uninduced cells (-dox) in soluble (Supernatant) and chromatin-bound fraction (Chromatin). Similar levels of endogenous RAD21 (-dox) and RAD21cv (TEV control) are observed bound to chromatin. Blotting for RAD21 shows absence of endogenous RAD21 in TEV and HRV (+dox). Full cleavage of RAD21cv is shown by blotting for EGFP, due to the release from chromatin the cleavage product appears now in the soluble fraction. Blotting with a bispecific antibody against the cohesin subunits STAG1/SA1 (SA1) and SATG2/SA2 (SA2) shows that these subunits are also released from chromatin after HRV cleavage. CTCF binding to chromatin is not affected. e, Chromatin immunoprecipitation (ChIP) with anti-SMC3 from RAD21cv/TEV and RAD21cv/HRV cells and qPCR with primers specific for cohesin binding sites shows a reduced signals in the SMC3 ChIP after RAD21 cleavage. (f.-g.) Topological domains and chromosomal interactions around the H19/IGF2 locus on chromosome 11. f, Topological domains (DC) assigned on the basis of the directional bias (DI) of Hi-C interactions in RAD21cv/TEV (TEV) and RAD21cv/HRV (HRV) cells. g. Chromosomal interactions detected by 3C-seq for five different viewpoints (Vp1-5, marked with black bars) in RAD21cv/TEV cells (blue tracks) and RAD21cv/HRV cells (red tracks). The observed interaction frequency is displayed for each BglII restriction fragment and normalized to reads per million sequenced reads. h, Binding sites for cohesin (SMC3) and CTCF detected by ChIP-sequencing in HEK293T cells. Below the graph the genes annotated by ENSEMBL for this region are shown.

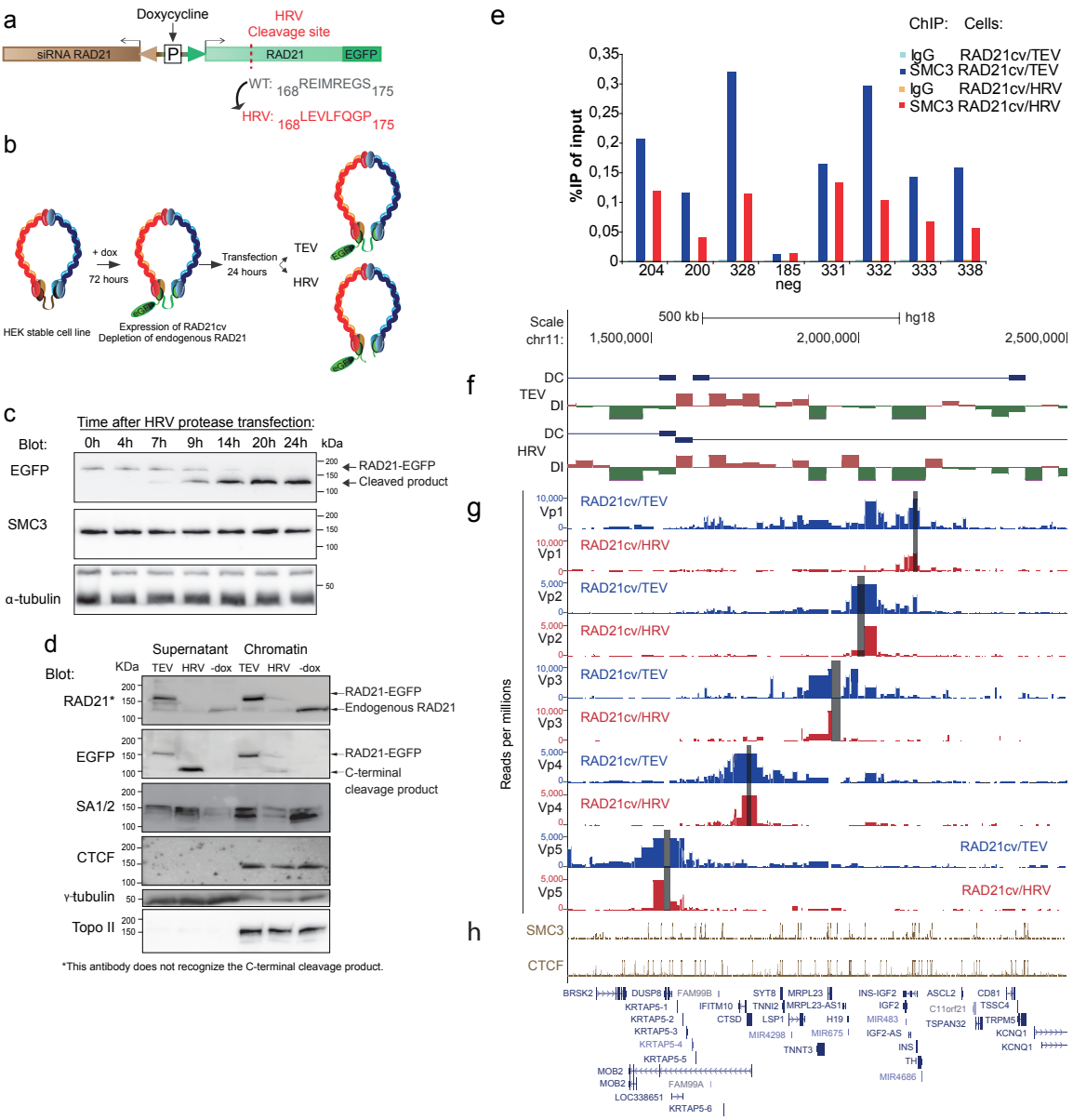

a
Doxycycline
HRV Cleavage site
siRNA RAD21
P
RAD21
EGFP
WT: 168REIMREGS175
HRV: 168LEVLFQGP175
b
+ dox
72 hours
Transfection
24 hours
TEV
HRV
HEK stable cell line
Expression of RAD21cv
Depletion of endogenous RAD21
c
Time after HRV protease transfection:
Blot:
0h 4h 7h 9h 14h 20h 24h kDa
EGFP
SMC3
α-tubulin
RAD21-EGFP
Cleaved product
d
Supernatant
Chromatin
Blot:
KDa
TEV HRV -dox TEV HRV -dox
RAD21*
EGFP
SA1/2
CTCF
γ-tubulin
Topo II
RAD21-EGFP
Endogenous RAD21
RAD21-EGFP
C-terminal cleavage product
*This antibody does not recognize the C-terminal cleavage product.
e
ChIP: Cells:
IgG RAD21cv/TEV
SMC3 RAD21cv/TEV
IgG RAD21cv/HRV
SMC3 RAD21cv/HRV
%IP of input
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
204 200 328 185 331 332 333 338
neg
f
Scale
chr11:
500 kb
hg18
1,500,000
2,000,000
2,500,000
DC
TEV
DI
DC
HRV
DI
g
Reads per millions
Vp1
Vp2
Vp3
Vp4
Vp5
RAD21cv/TEV
RAD21cv/HRV
h
SMC3
CTCF
BRSK2
DUSP8
FAM99B
SYT8
MRPL23
INS-IGF2
ASCL2
CD81
KRTAP5-1
IFITM10
TNNI2
MRPL23-AS1
IGF2
TSSC4
KRTAP5-2
CTSD
LSP1
H19
MIR483
C11orf21
TSPAN32
TRPM5
KRTAP5-3
MIR4298
MIR675
IGF2-AS
KCNQ1
KRTAP5-4
TNNT3
INS
KRTAP5-5
TH
MOB2
MIR4686
FAM99A
LOC338651
KRTAP5-6

Figure 2. Cohesin cleavage reduces interactions within topological domains genome-wide. a, SMC3 binding frequency across topological domains. SMC3 is enriched at the borders of domains. Each domain was split into 100 bins +/- 10 bins upstream and downstream of the domain boundaries. The frequency of SMC3 binding sites per kb was calculated and averaged over all domains. b, Schematic representation of the stratification method of interacting loci based on SMC3 binding. Interacting loci are broken into 3 classes, regions that have at least one SMC3 binding site at each interacting locus (2x), regions that have at least one SMC3 binding site at either interacting locus (1x), and regions that have no SMC3 binding (None). c, Hi-C interaction frequency correlated with cohesin. Comparison of the average normalized interaction frequency between SMC3 2x, 1x and none interacting loci at distances from 40kb to 2Mb. The inset is the fold change of the SMC3 2x and SMC3 1x categories relative to the "None" category. The largest fold change in interaction frequency appears between 100-200kb. d, Cohesin depletion leads to a loss of local interaction frequency. The y-axis shows that average loss of interaction frequency in the RAD21cv/HRV (HRV) cells compared to RAD21cv/TEV (TEV) cells for distances ranging from 40kb to 10Mb. The largest losses occur between interacting loci that are less than 2Mb apart. The inset shows the degree of depletion for the SMC3 2x, 1x and none categories. The SMC3 2x category is most affected by RAD21 depletion, and the maximal degree of depletion appears to occur in the 100-200kb range. e, Position of the cosmid-based FISH probes in the H19/IGF2 domain relative to the topological domain as shown by Hi-C interaction data. A subset of genes in the region is shown. The color (red, green) of the cosmids corresponds to the FISH images. f, DNA-FISH using two cosmid-based probes located in the H19/IGF2 domain in control cells (RAD21cv/TEV, left panel) and after RAD21 cleavage (RAD21cv/HRV, right panel). g, Cosmid-based FISH probes at the topological domain of the HOXD gene locus. Only a subset of genes in the region is shown. The color of the cosmid probes (red, green) corresponds to the FISH images. h, DNA-FISH using the cosmid probes shown in (g) in control cells (RAD21cv/TEV, left panel) and after RAD21 cleavage (RAD21cv/HRV, right panel). The marked FISH signals (white boxes) are shown enlarged at the right side of each panel. Consistent with the loss of interactions observed in the chromatin conformation capturing experiments we observed a separation of the FISH signals after cohesin cleavage. i, The positions of topological domains does not markedly change with RAD21 depletion. Browser shot showing heat maps of interaction frequency in the Control and RAD21 depleted cells. Also shown are the domain calls (DC) and directionality index (DI) over this region. j, Comparison of the topological domain boundary calls between replicates (TEV1/TEV2; HRV1/HRV2) and control (TEV) and RAD21 depleted cells (HRV). The differences between replicates and between control and knockdown experiments are comparable and largely unchanging.

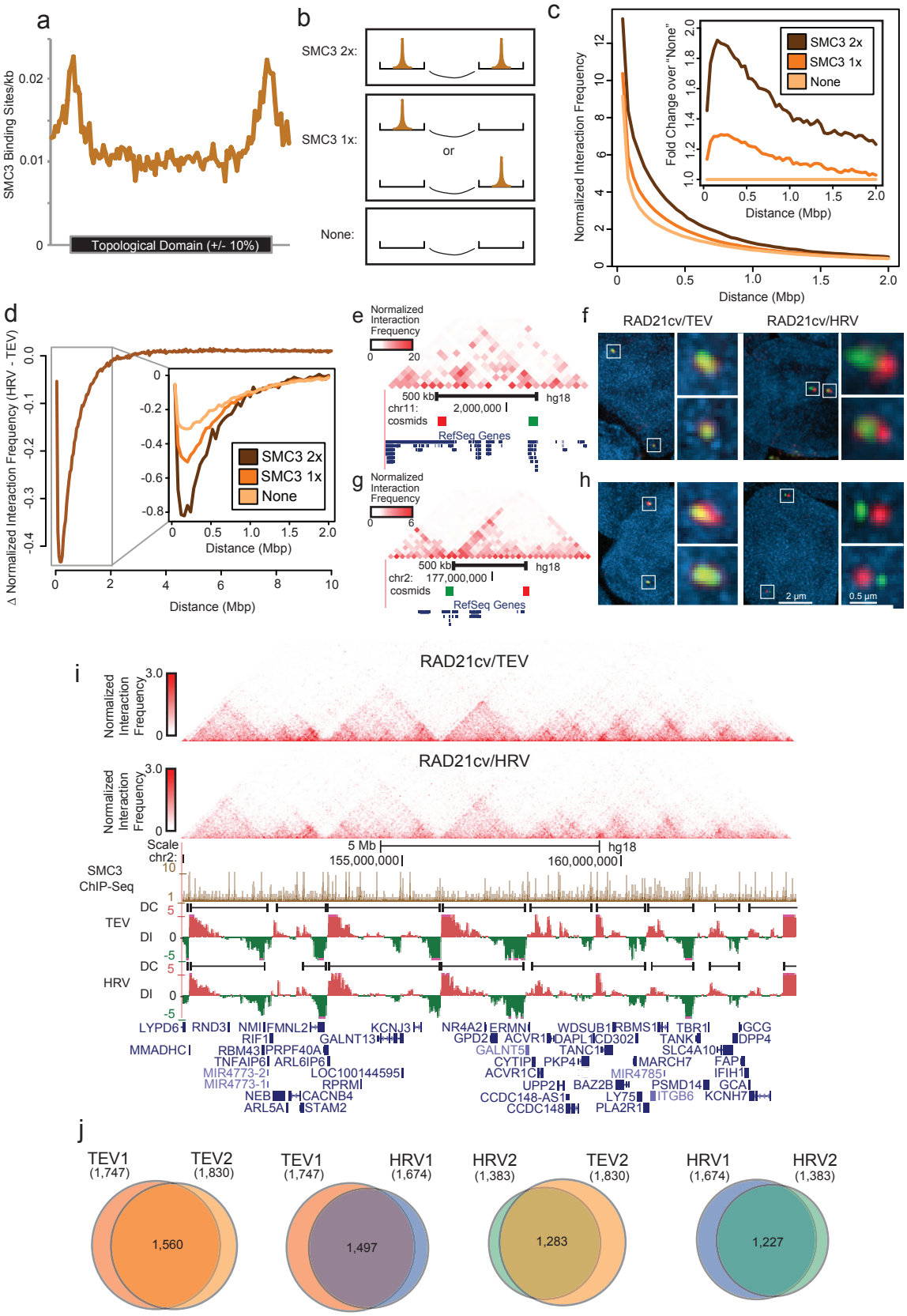

a
SMC3 Binding Sites/kb
0.02
0.01
0
Topological Domain (+/- 10%)
b
SMC3 2x:
SMC3 1x:
or
None:
c
Normalized Interaction Frequency
Distance (Mbp)
Fold Change over "None"
SMC3 2x
SMC3 1x
None
d
Δ Normalized Interaction Frequency (HRV - TEV)
Distance (Mbp)
e
Normalized Interaction Frequency
500 kb
hg18
chr11:
2,000,000
cosmids
RefSeq Genes
f
RAD21cv/TEV
RAD21cv/HRV
g
Normalized Interaction Frequency
chr2:
177,000,000
h
2 μm
0.5 μm
i
RAD21cv/TEV
RAD21cv/HRV
Scale
chr2:
5 Mb
155,000,000
160,000,000
SMC3 ChIP-Seq
TEV
HRV
DC
DI
LYPD6
RND3
NMI
IFNL2
KCNJ3
NR4A2
ERMN
WDSUB1
RBMS1
TBR1
IGCG
RIF1
GALNT13
GPD2
ACVR1
DAPL1
CD302
TANK
DPP4
MMADHC
RBM43
PRPF40A
GALNT5
TANC1
SLC4A10
TNFAIP6
ARL6IP6
CYTIP
PKP4
MARCH7
FAP
MIR4773-2
LOC100144595
ACVR1C
MIR4785
IFIH1
MIR4773-1
RPRM
UPP2
BAZ2B
PSMD14
GCA
NEB
CACNB4
CCDC148-AS1
LY75
ITGB6
KCNH7
ARL5A
STAM2
CCDC148
PLA2R1
j
TEV1 (1,747)
TEV2 (1,830)
1,560
TEV1 (1,747)
HRV1 (1,674)
1,497
HRV2 (1,383)
TEV2 (1,830)
1,283
HRV1 (1,674)
HRV2 (1,383)
1,227

Figure 3. CTCF depletion reduces the function of domain boundaries. a, CTCF binding frequency across topological domains. Similar to Fig. 2a, CTCF is enriched at the borders of topological domains. b, Using a stratification scheme similar to what was described in Fig. 2b, CTCF binding correlates with interaction frequency. The average interaction frequency at each distance was calculated for each category of interaction (CTCF 2x, CTCF 1x, and none). CTCF 2x shows the strongest interactions. c, Heat maps showing changes in interaction frequency between control and RAD21 depleted samples, as well as the actual interaction frequencies and domain calls. In the top heat map, blue color indicates a loss of interaction frequency, and red indicates a gain in interaction frequency. Upon RAD21 depletion there is predominantly a loss of intra-domain interaction frequency. d, Similar to c, but showing the changes in interaction frequency over the same locus after CTCF siRNA. A similar pattern of loss of intra-domain interaction frequency is observed (see blue triangles). However, unlike after RAD21 depletion, CTCF siRNA leads to an increase in inter-domain interaction frequency (see red signal in between blue triangles). e, Quantification of average change in interaction frequency after RAD21 depletion for intra-domain interaction (blue) and inter-domain interactions (yellow). In both cases, RAD21 leads to a loss of interaction frequency. f, Similar to e, but showing the change in interaction frequency after CTCF siRNA for intra- and inter-domain interactions. CTCF siRNA leads to a loss of intra-domain interaction frequency, but unlike RAD21 depletion, CTCF siRNA leads to an increase in inter-domain interaction frequency. g,h, Changes in inter-domain interaction frequency after CTCF siRNA depletion at sites stratified for either CTCF binding (g) or SMC3 binding (h). The loci that show the greatest increase in inter-domain interactions after CTCF siRNA tend to have higher frequencies of CTCF or SMC3 binding.

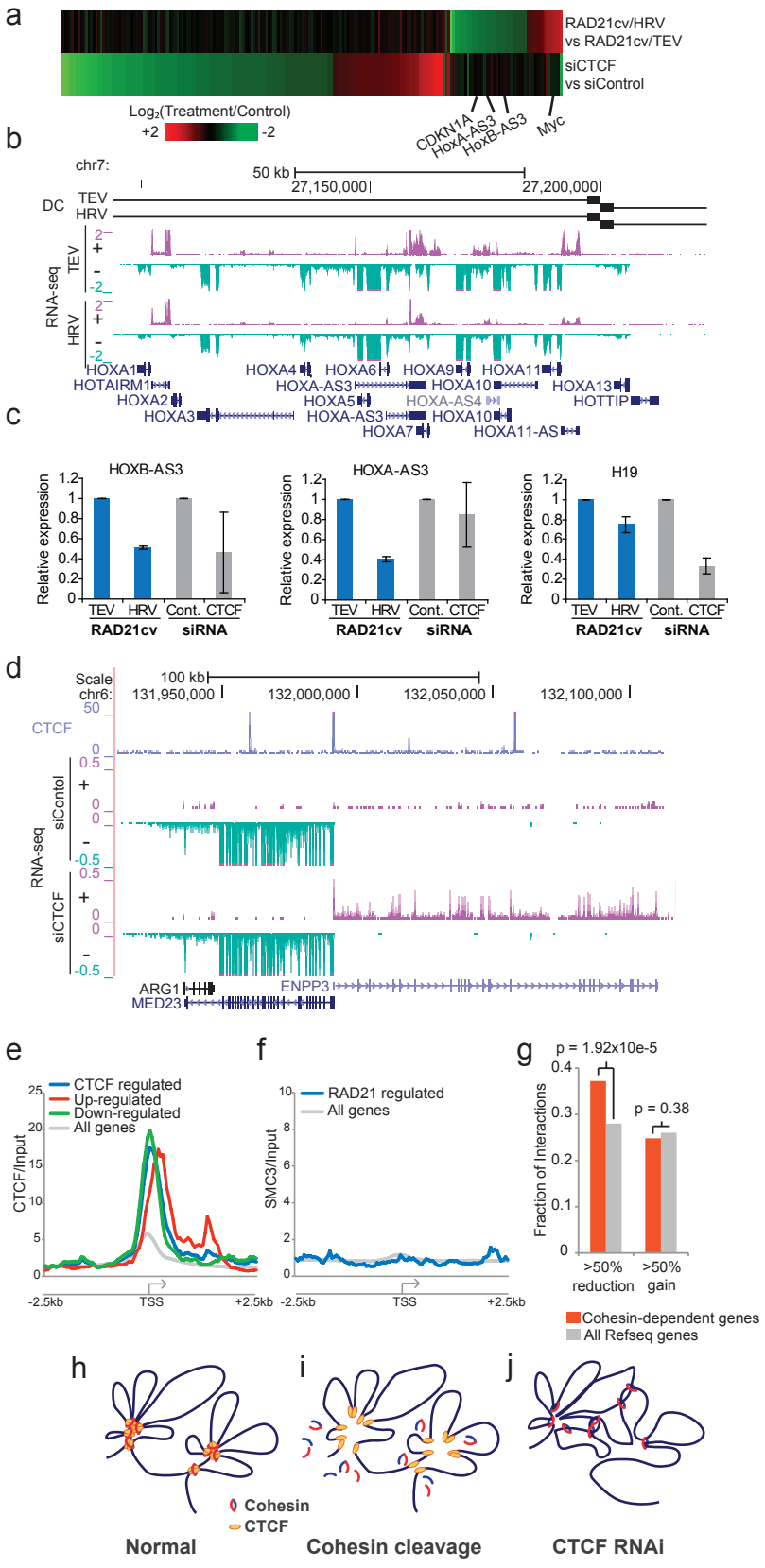
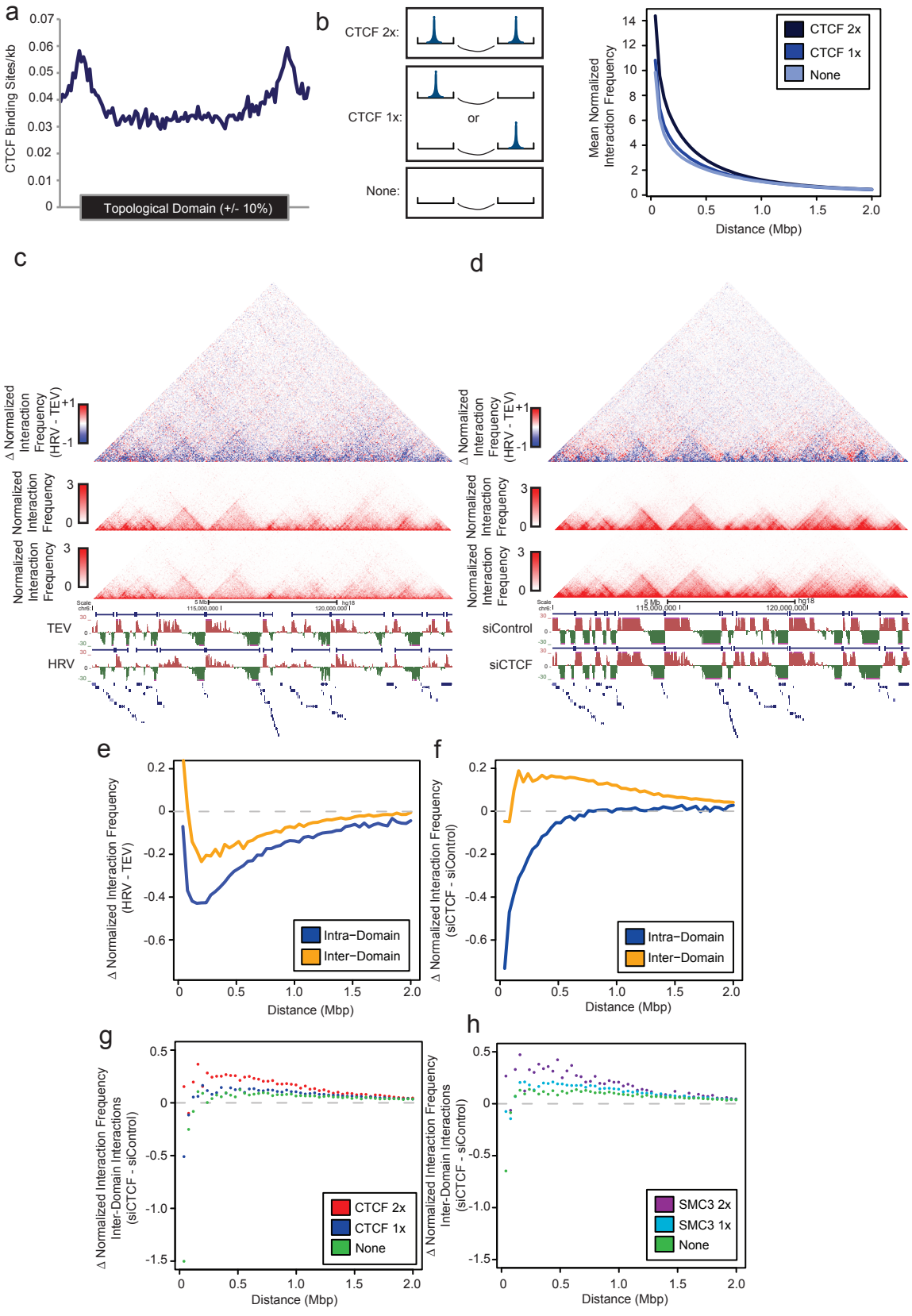
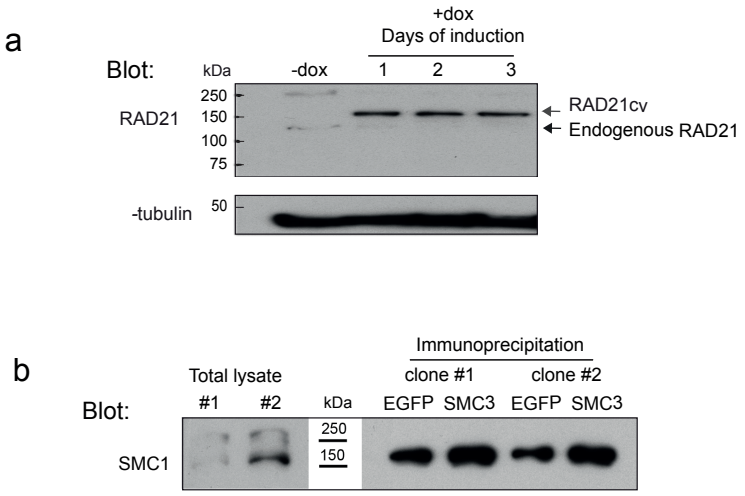
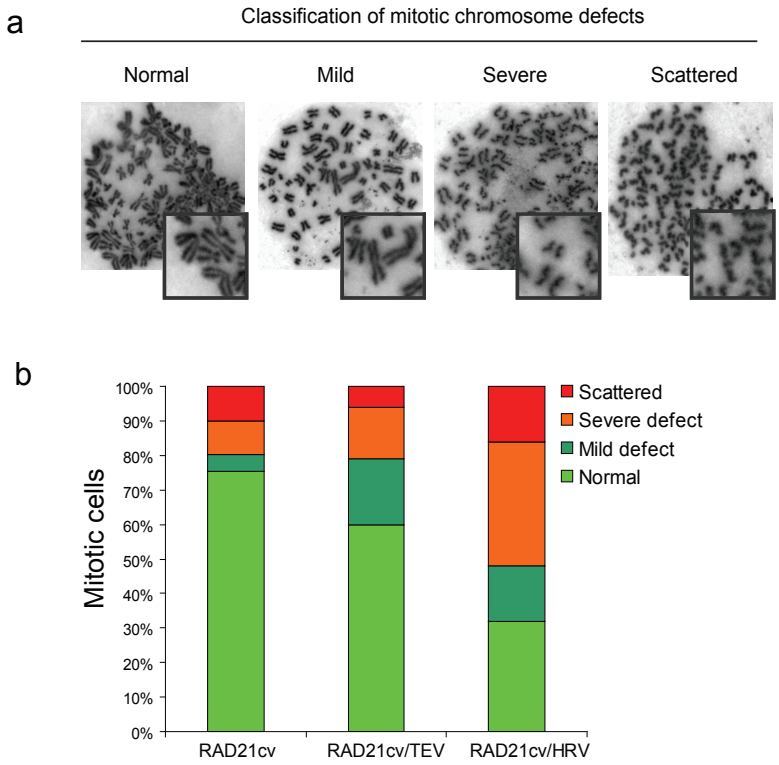

a
RAD21cv/HRV
vs RAD21cv/TEV
siCTCF
vs siControl
Log2(Treatment/Control)
+2
-2
CDKN1A
HoxA-AS3
HoxB-AS3
Myc
b
chr7:
50 kb
27,150,000
27,200,000
DC
TEV
HRV
RNA-seq
HOXA1
HOTAIRM1
HOXA2
HOXA3
HOXA4
HOXA-AS3
HOXA5
HOXA-AS3
HOXA7
HOXA6
HOXA9
HOXA10
HOXA-AS4
HOXA10
HOXA11
HOXA11-AS
HOXA13
HOTTIP
c
HOXB-AS3
HOXA-AS3
H19
Relative expression
TEV
HRV
Cont.
CTCF
RAD21cv
siRNA
d
Scale
chr6:
100 kb
131,950,000
132,000,000
132,050,000
132,100,000
CTCF
siContol
siCTCF
RNA-seq
ARG1
MED23
ENPP3
e
CTCF regulated
Up-regulated
Down-regulated
All genes
CTCF/Input
-2.5kb
TSS
+2.5kb
f
RAD21 regulated
All genes
SMC3/Input
g
p = 1.92x10e-5
p = 0.38
Fraction of Interactions
>50%
reduction
>50%
gain
Cohesin-dependent genes
All Refseq genes
h
i
j
Cohesin
CTCF
Normal
Cohesin cleavage
CTCF RNAi

Figure 4. Transcriptional changes after cohesin cleavage and CTCF depletion. a, Changes in expression levels for differentially expressed genes (FDR <5%) are compared between RAD21 cleavage and CTCF RNAi treatment by ranking first the fold changes from highest to lowest for RAD21 cleavage (upper part) and for CTCF depletion (the lower part). Only very few genes behave similar in both experiments. b, Expression of HOXA genes is altered after RAD21 cleavage. RNA-seq read coverage normalized to total sequenced reads for both strands are shown for RAD21cv/TEV and RAD21cv/HRV cells (+ strand shown in purple,- strand shown in turquoise). Please note the strong reduction of HOXA-AS3 and HOXA11-AS and HOXA7 . c, Reduced expression of HOXB-AS3 and HOXA-AS3 after RAD21 cleavage was confirmed by qPCR. CTCF depletion did not lead to a consistent reduction, as also seen in the analysis of the RNAseq data (Supplementary table 1). Transcription of the H19 noncoding RNA was reduced after CTCF depletion but to a smaller extent by RAD21 cleavage (mean n=3, +/- s.d.). d, Transcription of the ENPP3 gene is increased after CTCF knockdown. RNA-seq read coverage normalized to total sequenced reads for both strands are shown for control siRNA and CTCF siRNA (+ strand shown in purple,- strand shown in turquoise). The gene has CTCF binding sites at the promoter and also intragenic. The upregulation was confirmed by RT-PCR to depend solely on CTCF knockdown (Supplementary fig. 9f). e, The position of CTCF sites was analysed relative to transcription start sites of all genes (grey line) and genes with altered expression after CTCF depletion (blue line). Each line represents that average fold-enrichment of CTCF (RPKM) relative to input (RPKM) over a +/- 2.5 kb window surrounding the promoters of CTCF regulated genes. CTCF is clearly enriched at the TSS of differentially expressed genes. f, Similar to e, except showing the fold-enrichment of SMC3 (RPKM) over input (RPKM) over the promoter of genes altered after RAD21 depletion. SMC3 does not appear to be enriched at the promoter of the genes regulated by cohesin depletion.g, Analysis of change in interaction frequency between restriction fragments containing a promoter and restriction fragments containing a distal DNaseI hypersensitive site (DHS). Shown is the fraction of genes that display a 50% reduction or 50% increase in interaction frequency after RAD21 depletion for either cohesin regulated genes (orange) or all Refseq genes (grey). Cohesin regulated genes are enriched for a loss of interactions with restriction fragments containing distal DHS sites relative to all Refseq genes (Fisher's exact test). h-j, Models describing the different changes of chromosomal interactions after cohesin cleavage (i) and CTCF depletion (j) h, Cohesin and CTCF colocalize and are both necessary to shape long-range interactions. i, RAD21 cleavage releases cohesin from chromatin but does not change CTCF binding. Loss of cohesin leads to reduced interactions within domains. CTCF binding might still influence the topology of the chromatin fibre and maintain domain identity. j, CTCF depletion leads to non-specific cohesin localization which could lead to interactions across domain boundaries normally prevented by CTCF's insulation function.

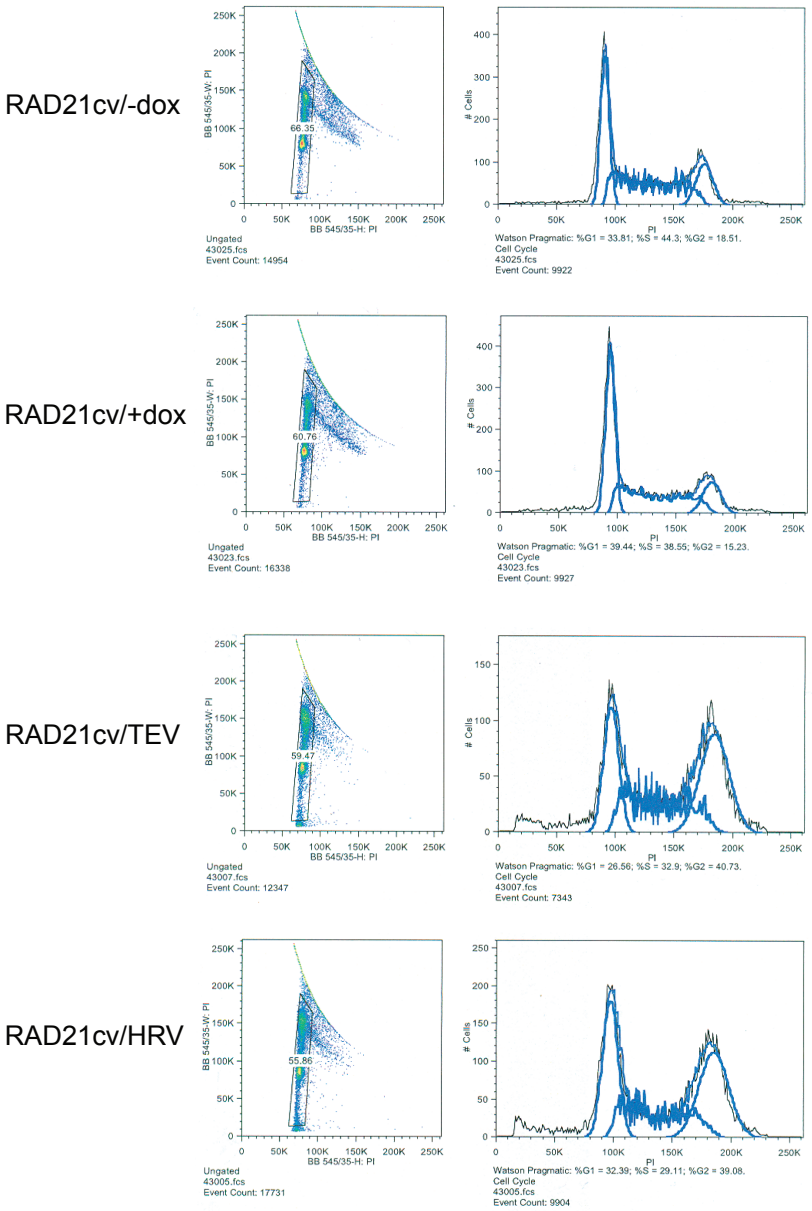

a
CTCF Binding Sites/kb
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Topological Domain (+/- 10%)
b
CTCF 2x:
CTCF 1x:
or
None:
Mean Normalized Interaction Frequency
CTCF 2x
CTCF 1x
None
Distance (Mbp)
c
Δ Normalized Interaction Frequency (HRV - TEV)
Normalized Interaction Frequency
TEV
HRV
d
siControl
siCTCF
e
Δ Normalized Interaction Frequency (HRV - TEV)
Intra−Domain
Inter−Domain
Distance (Mbp)
f
Δ Normalized Interaction Frequency (siCTCF - siControl)
g
Δ Normalized Interaction Frequency Inter-Domain Interactions (siCTCF - siControl)
h
SMC3 2x
SMC3 1x



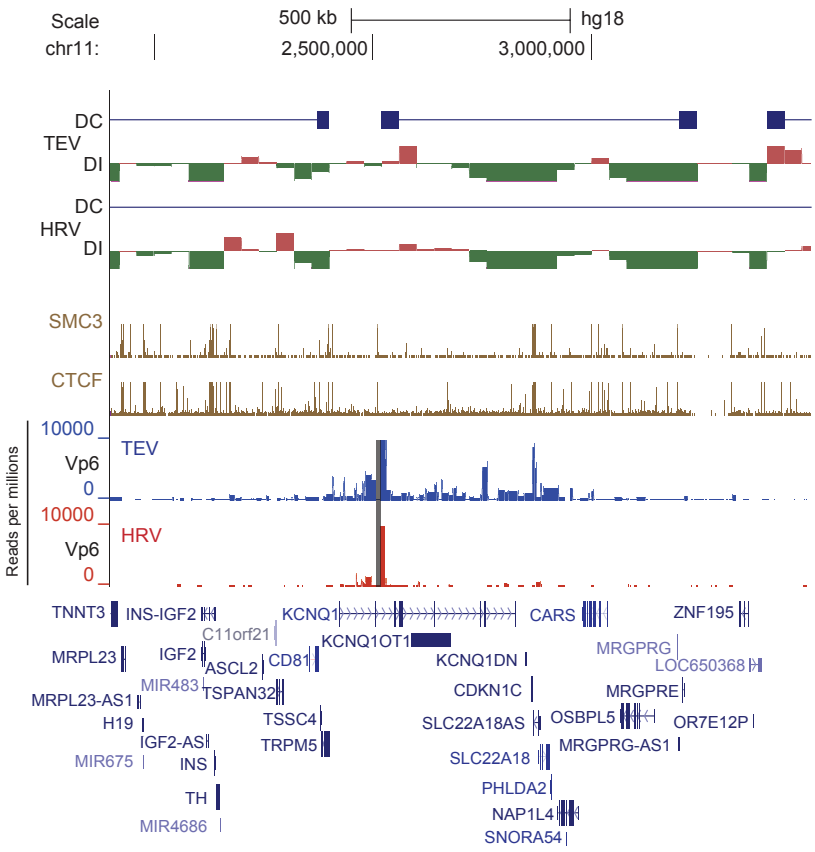
Supplementary Figure 1. Replacement of endogenous RAD21 by RAD21cv and incorporation in the cohesin complex. a, Western blot showing the replacement of endogenous RAD21 by RAD21cv in a time-course until 3 days after doxycycline induction. Note that the chromatin-bound level of RAD21cv is similar to the endogenous RAD21 level (Fig. 1d), although the expression level of RAD21cv is higher than the endogenous level. b, Immunoprecitpitation with anti-EGFP and anti-SMC3 antibodies was performed from two different RAD21cv clones and the Western blot probed with anti-SMC1. The co-precipitation of SMC1 with RAD21cv (EGFP-tag) shows that RAD21cv is incorporated in the cohesin complex.



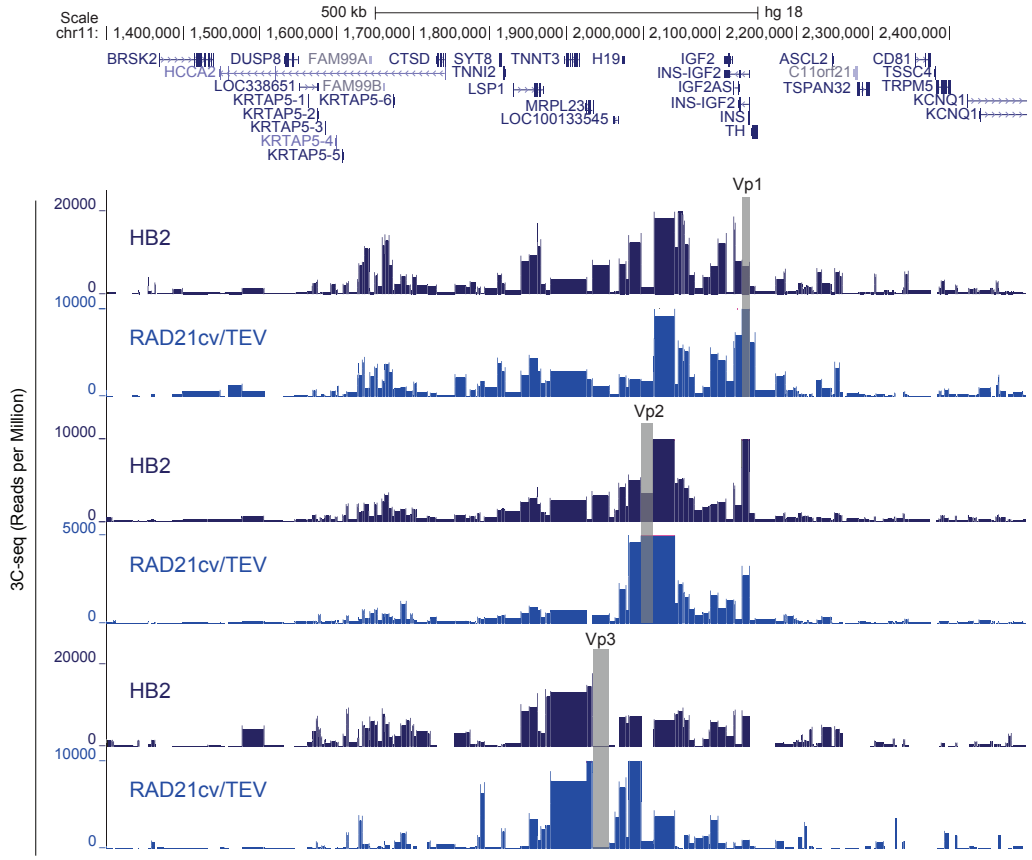
Supplementary Figure 2. RAD21 cleavage causes premature loss of sister chromatid cohesin. Cells were treated for 2 hours with nocodazole and then spreaded to analyse the mitotic cells for sister chromatid cohesion defects. a, Different degrees of sister chromatid cohesion defects used to cluster observed mitotic defects. b, Bar chart displaying the percentage of counted mitotic cells showing different degrees of sister cohesion defects. We analyzed doxycycline induced RAD21cv cells (RAD21cv/+dox), and RAD21cv cells transfected with the different proteases (RAD21cv/TEV and RAD21cv/HRV).



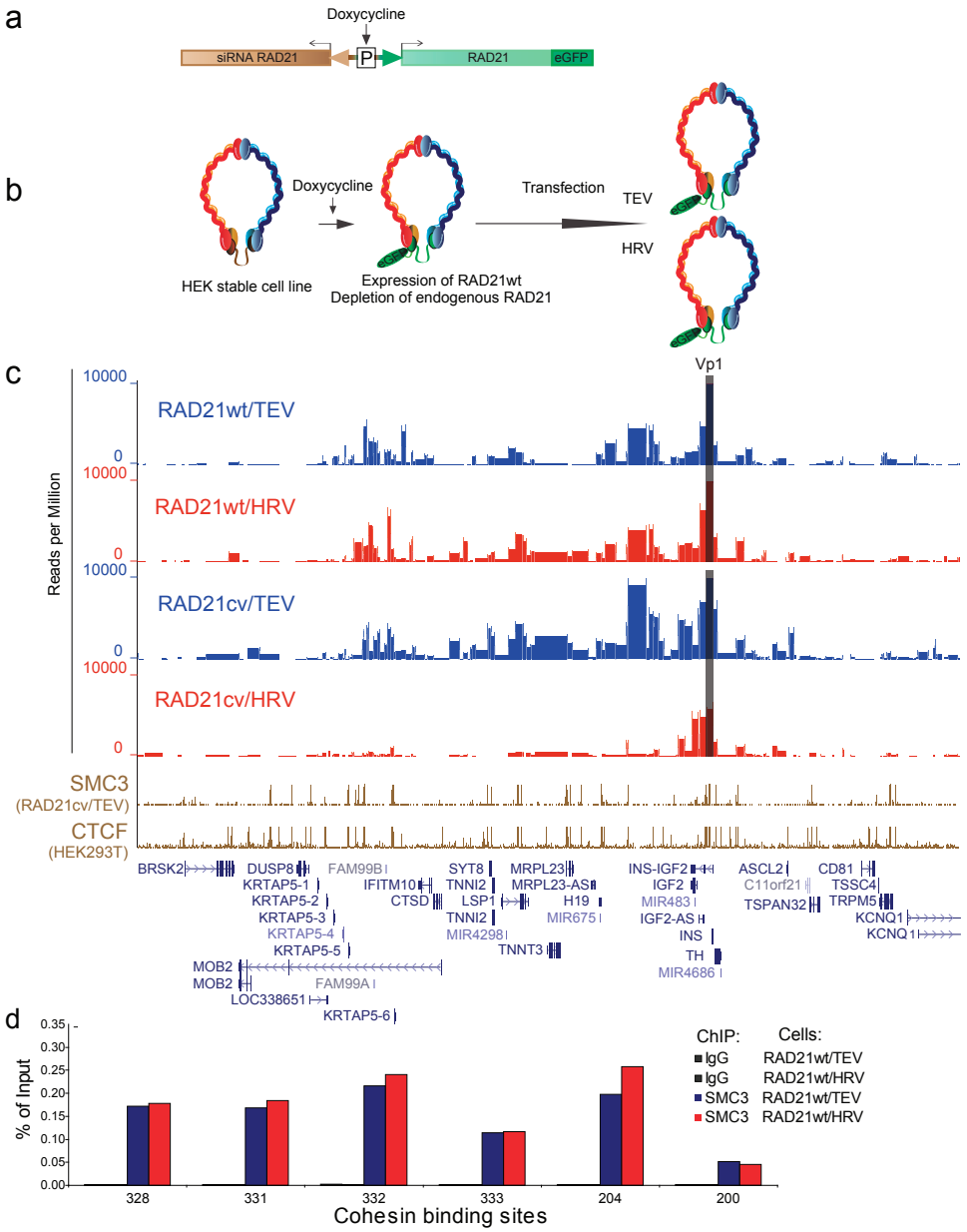
Supplementary Figure 3. Cell cycle distribution of treated and untreated RAD21cv cells. FACS analysis of the cell cycle distribution of the cells used in the RAD21 cleavage experiments: uninduced cells (RAD21cv/-dox), doxycycline induced cells (RAD21cv/+dox) and cells after transfection of the different proteases (RAD21cv/TEV, RAD21cv/HRV).



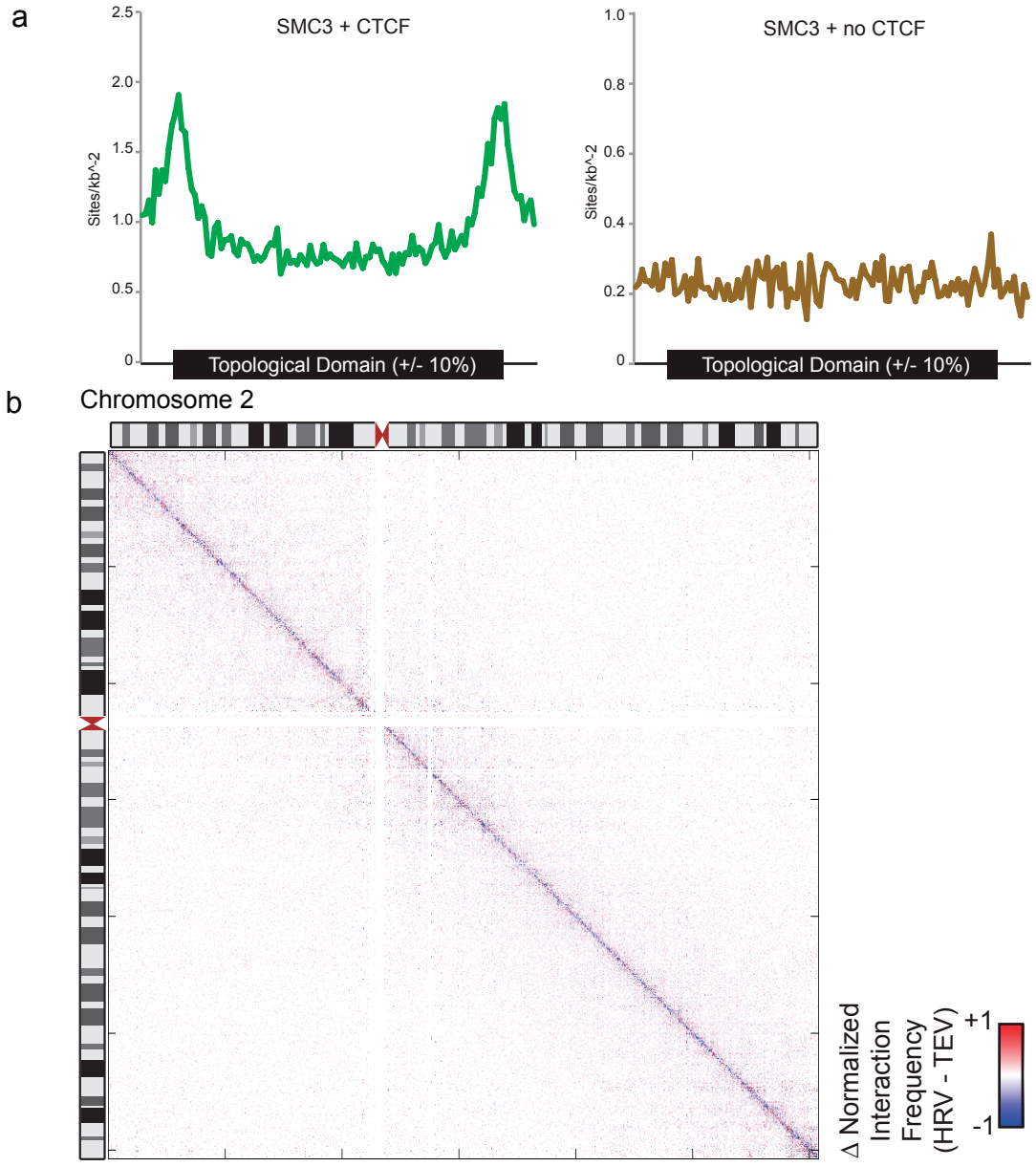
Supplementary Figure 4. Loss of interactions around KCNQ1 after RAD21 cleavage. Domain organization and long range-interactions around a 3C-seq viewpoint in the KCNQ1 gene (Vp6) are shown for RAD21cv/TEV (TEV) and RAD21cv/HRV (HRV) cells. The topological domains in the region are shown using domain calls (DC) and the directionality index (DI). Cohesin binding sites are shown for RAD21cv/TEV cells and CTCF sites for wildtype HEK293 cells. Interactions are detected within the KCNQ1 gene and a region downstream comprising several cohesin/CTCF sites (blue bar graph). RAD21 cleavage leads to an overall loss of interactions (red bar graph). Interactions are displayed as reads per million sequenced reads.



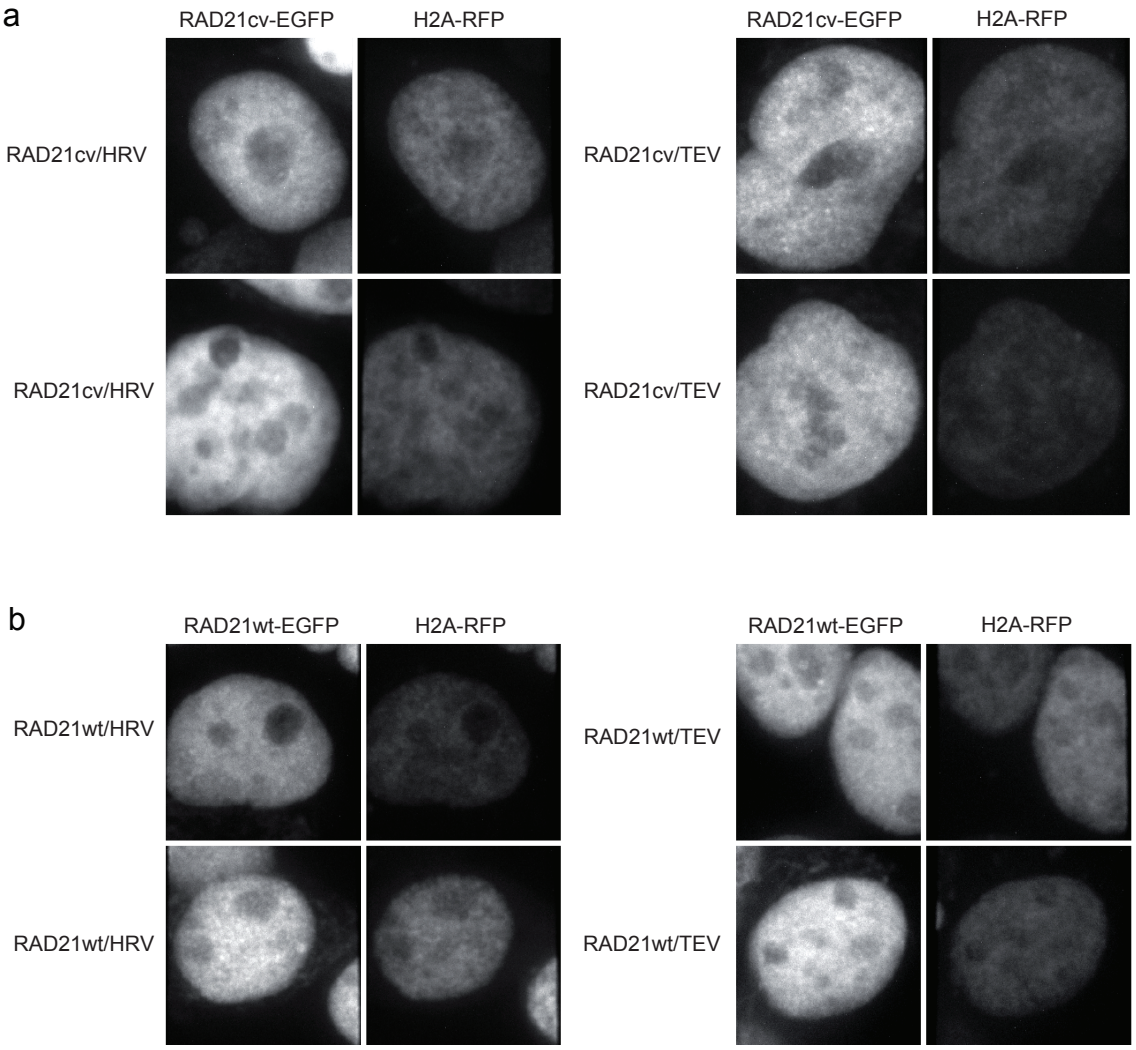
Supplementary Figure 5. Conservation of long-range chromosomal interactions between different tissues. 3C-sequencing was performed for three different viewpoints in a breast endothelial cell line with normal karyotype (HB2, dark blue bar graph) and the control cells for the RAD21 cleavage experiment (RAD21cv/TEV, blue bar graph) using the same protocol. The viewpoint positions are marked with grey bars. All three viewpoints in the IGF2-H19 domain show similar interactions, indicating conservation of chromosomal interactions between different cell types.



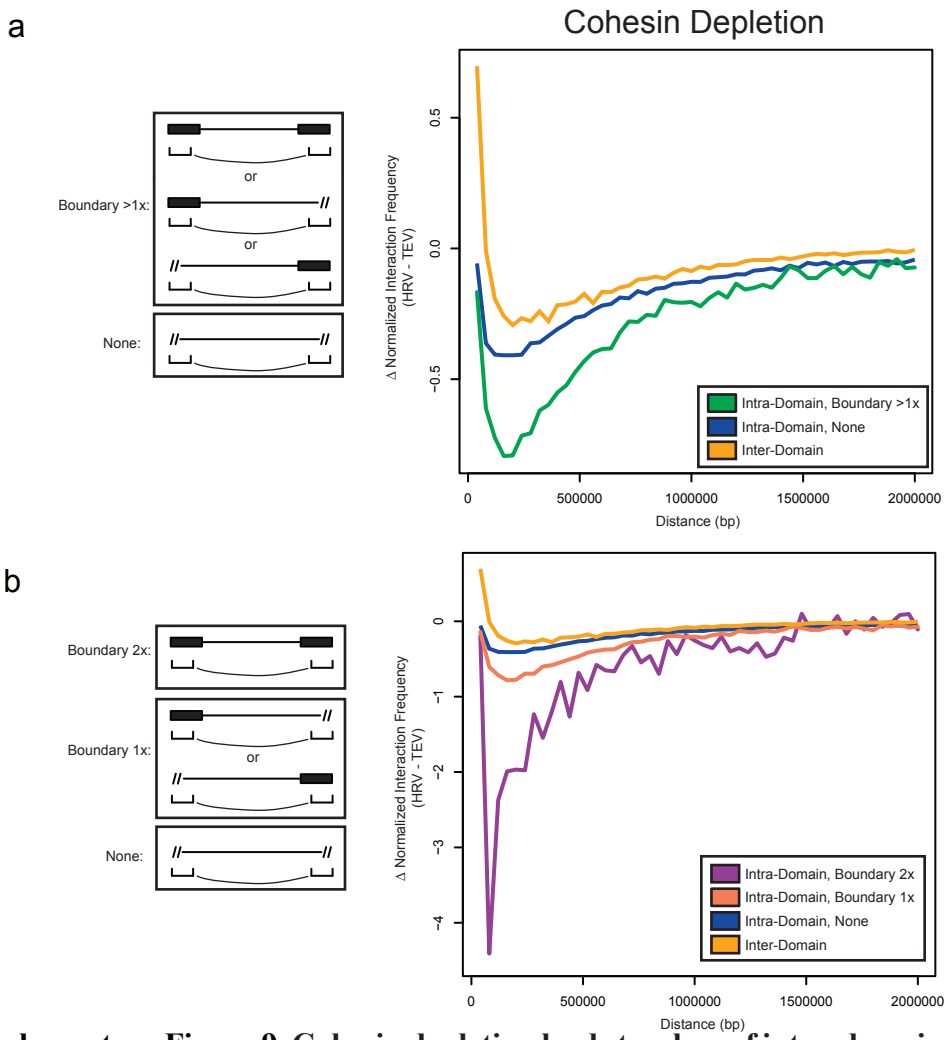
Supplementary Figure 6. Control cells expressing RAD21wt w/o HRV site do not respond to HRV protease transfection. a, Schematic representation of the episomal construct used for the control cell line containing a doxycycline-inducible bidirectional promoter driving expression of RAD21wt and siRNA targeting endogenous RAD21 simultaneously. b, Outline of the experiment. c, Cells expressing protease-insensitive RAD21-EGFP (RAD21wt) do not show altered long-range interactions for viewpoint 1 (VP1) after HRV protease transfection (RAD21wt/HRV, red bar graph), in contrast to cells expressing cleavable RAD21 (RAD21cv/HRV, red bar graph). The respective transfections with control protease are shown as blue bar graphs. d, The ChIP-qPCR assay with several primer corresponding to cohesin sites for RAD21wt/TEV and RAD21wt/HRV shows no effect.



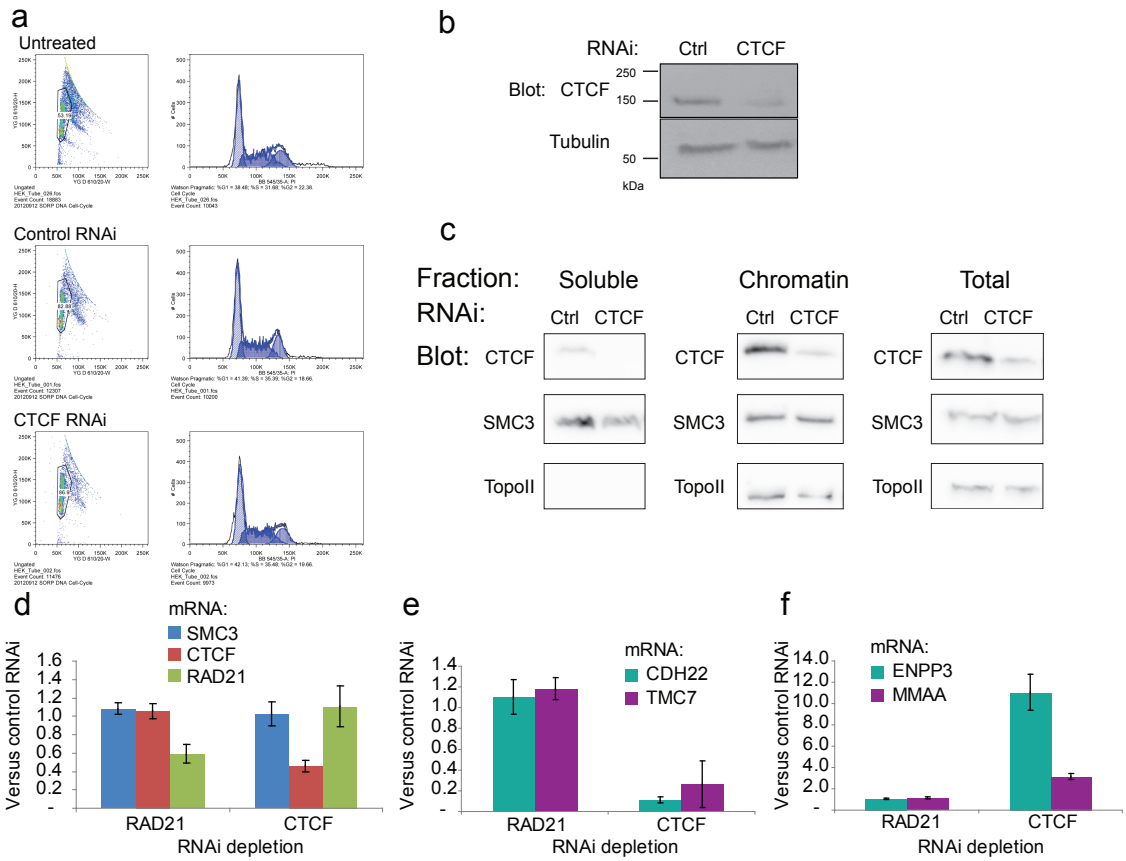
Supplementary Figure 7. Enrichment of cohesin/CTCF sites at boundaries and differential heat map plot after RAD21 cleavage. a, SMC3 is enriched at the borders of domains when colocalizing with CTCF but not without. Each domain was split into 100 bins +/- 10 bins upstream and downstream of the domain boundaries. The frequency of SMC3/CTCF or SMC3 only binding sites per kb was calculated and averaged over all domains. b, Chromosome 2 - heat map of changes in interaction frequency. The heat map shows the changes in interaction frequency (RAD21cv/HRV – RAD21cv/TEV). The largest differences are observed very close to the diagonal, indicating that the most prominent changes in interaction frequency are local (<2Mb), and loss of interactions (blue) is dominating.



Supplementary Figure 8. Live cell imaging shows preservation of chromatin morphology after RAD21 cleavage. Live cell imaging of RAD21cv and RAD21wt cells transfected with HRV and TEV protease and in addition histone H2A-RFP. The EGFP signal of RAD21cv/RAD21wt and the RFP signal of the H2A-RFP construct were imaged using a spinning disc mode. RAD21cv and RAD21wt show a speckled pattern (EGFP) when cells are transfected with TEV. a,The EGFP signal turns into an amorphous pattern when transfected with HRV, consistent with a release of RAD21cv from chromatin. The H2A-RFP patterns do not change visibly between the different protease transfections, indicating that cohesin cleavage does not trigger major changes in chromatin morphology. b,The RAD21wt and H2A-RFP patterns do not change when HRV is transfected, indicating that the localization of RAD21wt does not change.

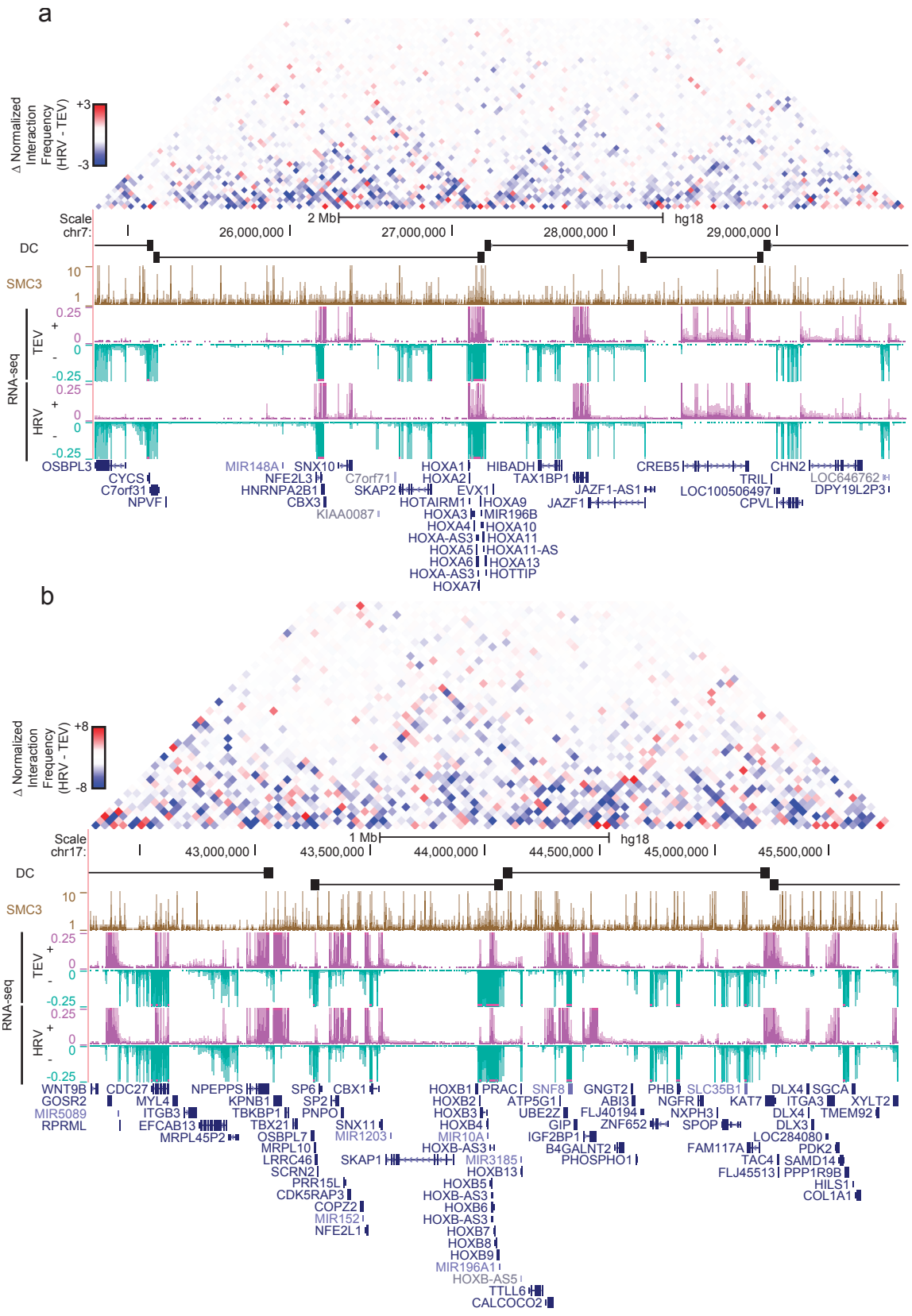


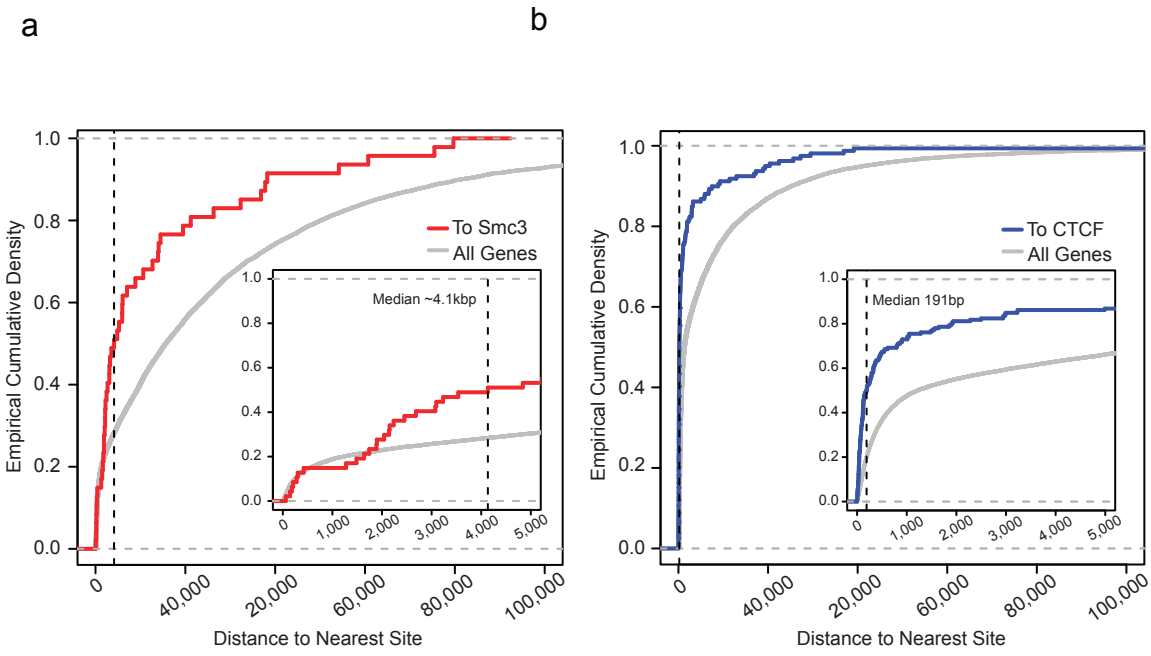
Supplementary Figure 9. Cohesin depletion leads to a loss of intra-domain boundary associated interactions. a, Graphs showing the degree of depletion in interaction frequency between RAD21cv/TEV (TEV) and RAD21cv/HRV (HRV) cells at each distance up to 2Mb. Interacting bin-pairs were stratified either as inter-domain (yellow), non-boundary associated (”None”, blue), or where at least one bin was boundary associated (”Boundary >1x”, green). The loss of interaction frequency is highest when at least one bin is associated with a boundary. b, Similar to panel a, but in this case also showing bin-pairs with exactly one boundary assoicated bin (”Boundary 1x”, orange) or two boundary associated bins (”Boundary 2x”, magenta). There are relatively few “Boundary 2x” bin pairs in the genome, which accounts for the jaggedness of the magenta line. In both cases, a schematic explaining the bin-pair segregation scheme is shown on the left.



Supplementary Figure 10. CTCF RNAi depletes CTCF but does not change the levels of cohesin bound to chromatin. a, FACS analysis of the cell cycle distribution of untreated HEK293T cells and cells treated with CTCF and control RNAi. b, Western blot for CTCF and tubulin as loading control to show the depletion of CTCF by siRNA. c, HEK293T cells were transfected with control and CTCF RNAi and fractionated in total cell extract (total), the soluble pool containing cytoplasm and nucleoplasm (soluble) and the chromatin-bound pool (chromatin) of proteins. The western blot for SMC3 shows equal levels of chromatin-bound proteins after CTCF depletion. The chromatin-bound fraction is marked by blotting for Topoisomerase II (TopoII) after CTCF depletion. This confirms observations from our earlier study (8). d, Transcripts of SMC3, CTCF and RAD21 were analysed after RNAi depletion of RAD21 and CTCF in HEK293T cells. Consistent with the depletion we observe a reduction of RAD21 and CTCF after the respective siRNA treatments. Fold expression compared to the control RNAi is shown (mean n=3, +/- s.d.). e, Two genes (CDH22 - cadherin 22 precursor, TMC7 - transmembrane channel-like 7) found by RNA-seq reduced after CTCF were validated by RNAi by qPCR. Fold expression versus control RNAi is shown (mean n=3, +/- s.d.). f, Validation of two genes (ENPP3 - ectonucleotide pyrophosphatase/phosphodiesterase,MMAA - methylmalonic aciduria type A precursor) found to be up-regulated after CTCF RNAi in our RNA-seq data by qPCR. RAD21 depletions did not have any effect on these genes. (mean n=3, +/- s.d.).

Supplementary figure 11. Changes of long-range interactions around the HOXA and the HOXB locus after RAD21 cleavage. Differences in Hi-C interaction frequency after RAD21 depletion in regions surrounds the HOXA locus (a) and the HOXB locus (b). Heat maps show the difference in interaction frequency between the RAD21cv/HRV and RAD21cv/TEV cells (HRV - TEV). Reduced interactions are shown in blue and increased interactions shown in red. Further we show the location of topological domains uding the domain calls (DC), SMC3 ChIP-seq profiles for RAD21cv/TEV cells and RNA-seq experiments with (RAD21cv/HRV) and without (RAD21cv/TEV) RAD21cleavage. In the RNA-seq tracks, positive stranded reads are shown in purple, while negative stranded reads are shown in turquoise.


a
Δ Normalized Interaction Frequency (HRV - TEV)
+3
-3
Scale
chr7:
2 Mb
hg18
26,000,000
27,000,000
28,000,000
29,000,000
DC
SMC3
10
1
RNA-seq
TEV
HRV
0.25
0
-0.25
OSBPL3
CYCS
C7orf31
NPVF
MIR148A
SNX10
NFE2L3
HNRNPA2B1
CBX3
KIAA0087
C7orf71
SKAP2
HOXA1
HOXA2
EVX1
HOTAIRM1
HOXA3
HOXA4
HOXA-AS3
HOXA5
HOXA6
HOXA-AS3
HOXA7
HOXA9
MIR196B
HOXA10
HOXA11
HOXA11-AS
HOXA13
HOTTIP
HIBADH
TAX1BP1
JAZF1-AS1
JAZF1
CREB5
TRIL
LOC100506497
CPVL
CHN2
LOC646762
DPY19L2P3
b
Δ Normalized Interaction Frequency (HRV - TEV)
+8
-8
Scale
chr17:
1 Mb
hg18
43,000,000
43,500,000
44,000,000
44,500,000
45,000,000
45,500,000
DC
SMC3
RNA-seq
TEV
HRV
WNT9B
GOSR2
MIR5089
RPRML
CDC27
MYL4
ITGB3
EFCAB13
MRPL45P2
NPEPPS
KPNB1
TBKBP1
TBX21
OSBPL7
MRPL10
LRRC46
SCRN2
PRR15L
CDK5RAP3
COPZ2
MIR152
NFE2L1
SP6
SP2
PNPO
SNX11
MIR1203
CBX1
SKAP1
HOXB1
HOXB2
HOXB3
HOXB4
MIR10A
HOXB-AS3
HOXB13
HOXB5
HOXB-AS3
HOXB6
HOXB-AS3
HOXB7
HOXB8
HOXB9
MIR196A1
HOXB-AS5
TTLL6
CALCOCO2
PRAC
SNF8
ATP5G1
UBE2Z
GIP
IGF2BP1
B4GALNT2
PHOSPHO1
MIR3185
GNGT2
ABI3
FLJ40194
ZNF652
PHB
NGFR
NXPH3
SPOP
SLC35B1
KAT7
DLX4
ITGA3
SGCA
XYLT2
TMEM92
DLX3
LOC284080
FAM117A
PDK2
TAC4
SAMD14
FLJ45513
PPP1R9B
HILS1
COL1A1



Supplementary Figure 12. Empirical Cumulative Density plots of the distance from the transcription start site of an RAD21 (a) or CTCF (b) regulated gene to the nearest binding site for either SMC3 (a) or CTCF (b). a. Shown in red is the empirical cumulative densitydistribution of the distance between the TSS of RAD21 regulated genes to the nearest SMC3 binding site. The median distance to the nearest site is ~4.1kb (shown with a vertical dashed line). Shown in grey is the distribution for all RefSeq genes, demonstrating that RAD21 regulated genes typically have an SMC3 binding site closer than would be expected at random. The inset shows the same data but zoomed into a 5kb limit. b. Similar to a, but showing the distance between the TSS of CTCF regulated genes and the nearest CTCF binding site (blue). The median distance for CTCF regulated genes to the nearest CTCF binding site is 191bp, showing that CTCF regulated genes tend to be directly bound at their promoter by CTCF.

References

1. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289-93.

2. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature.advance online publication.

3. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell.148(3):458-72.

4. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation center. in submission. 2012.

5. Nativio R, Wendt KS, Ito Y, Huddleston JE, Uribe-Lewis S, Woodfine K, et al. Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. PLoS Genet. 2009;5(11):e1000739. PMCID: 2776306.

6. Hadjur S, Williams LM, Ryan NK, Cobb BS, Sexton T, Fraser P, et al. Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature. 2009;460(7253):410-3. PMCID: 2869028.

7. Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. 2006;20(17):2349-54. PMCID: 1560409.

8. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature. 2008;451(7180):796-801.

9. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature.467(7314):430-5. PMCID: 2953795.

10. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell. 1999;98(3):387-96.

11. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009;137(7):1194-211. PMCID: 3040116.

12. Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, et al. Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell. 2008;132(3):422-33.

13. Rubio ED, Reiss DJ, Welcsh PL, Disteche CM, Filippova GN, Baliga NS, et al. CTCF physically links cohesin to chromatin. Proc Natl Acad Sci U S A. 2008;105(24):8309-14. PMCID: 2448833.

14. Kurukuti S, Tiwari VK, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, et al. CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. Proc Natl Acad Sci U S A. 2006;103(28):10684-9. PMCID: 1484419.

15. Mishiro T, Ishihara K, Hino S, Tsutsumi S, Aburatani H, Shirahige K, et al. Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. EMBO J. 2009;28(9):1234-45. PMCID: 2683055.

16. Hou C, Dale R, Dean A. Cell type specificity of chromatin organization mediated by CTCF and cohesin. Proc Natl Acad Sci U S A.107(8):3651-6. PMCID: 2840441.

17. Uhlmann F, Lottspeich F, Nasmyth K. Sister-chromatid separation at anaphase onset is promoted by cleavage of the cohesin subunit Scc1. Nature. 1999;400(6739):37-42.

18. Schockel L, Mockel M, Mayer B, Boos D, Stemmann O. Cleavage of cohesin rings coordinates the separation of centrioles and chromatids. Nat Cell Biol.13(8):966-72.

19. Stadhouders R, Kolovos P, Brouwer R, Zuin J, van den Heuvel A, Kockx C, et al. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. Nat Protoc.8(3):509-24.

20. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods.9(10):999-1003.

21. van de Corput MP, de Boer E, Knoch TA, van Cappellen WA, Quintanilla A, Ferrand L, et al. Super-resolution imaging reveals three-dimensional folding dynamics of the beta-globin locus upon gene activation. J Cell Sci.125(Pt 19):4630-9.

22. Alexander T, Nolte C, Krumlauf R. Hox genes and segmentation of the hindbrain and axial skeleton. Annu Rev Cell Dev Biol. 2009;25:431-56.

23. Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D. The dynamic architecture of Hox gene clusters. Science.334(6053):222-5.

24. Rollins RA, Korom M, Aulner N, Martens A, Dorsett D. Drosophila nipped-B protein supports sister chromatid cohesion and opposes the stromalin/Scc3 cohesion factor to facilitate long-range activation of the cut gene. Mol Cell Biol. 2004;24(8):3100-11. PMCID: 381657.

25. Seitan VC, Hao B, Tachibana-Konwalski K, Lavagnolli T, Mira-Bontenbal H, Brown KE, et al. A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. Nature.476(7361):467-71. PMCID: 3179485.

26. Schockel L, Mockel M, Mayer B, Boos D, Stemmann O. Cleavage of cohesin rings coordinates the separation of centrioles and chromatids. Nat Cell Biol. 2011;13(8):966-72.

27. Bornkamm GW, Berens C, Kuklik-Roos C, Bechet JM, Laux G, Bachl J, et al. Stringent doxycycline-dependent control of gene activities using an episomal one-vector system. Nucleic Acids Res. 2005;33(16):e137.

28. Sumara I, Vorlaufer E, Gieffers C, Peters BH, Peters JM. Characterization of vertebrate cohesin complexes and their regulation in prophase. J Cell Biol. 2000;151(4):749-62.

29. Kueng S, Hegemann B, Peters BH, Lipp JJ, Schleiffer A, Mechtler K, et al. Wapl controls the dynamic association of cohesin with chromatin. Cell. 2006;127(5):955-67.

30. Simonis M, Kooren J, de Laat W. An evaluation of 3C-based methods to capture DNA interactions. Nat Methods. 2007;4(11):895-901.

31. Stadhouders R, Brouwer RW, Kolovos P, Zun J, vna den Heuwel A, Kockx C, et al. Multiplexed Chromosome Conformation Capture Sequencing (m3C-Seq) for rapid genome-scale high resolution detection of long-range chromatin interactions. Nature protocols. 2013;in press.

32. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

33. van de Corput MP, de Boer E, Knoch TA, van Cappellen WA, Quintanilla A, Ferrand L, et al. Super-resolution imaging reveals three-dimensional folding dynamics of the beta-globin locus upon gene activation. J Cell Sci. 2012;125(Pt 19):4630-9.

34. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 2009;37(18):e123.

35. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.

36. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012;9(10):999-1003.

37. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376-80.

Acknowledgements

Chapter 3, in full, has been submitted for publication of the material as it may appear in *Nature*. Zuin, Jessica; Dixon, Jesse R.; van der Reijden, Michael .I.J.A; Kolovos, Petros; Ye, Zhen; Brouwer, Rutger W.W.; van de Corput, Mariette P.C.; van Ijcken, Wilfred F.J.; Grosveld, Frank G.; Ren, Bing; Wendt, Kerstin S. The dissertation author was the co-primary investigator and the co-primary author of this material.

Chapter 4

Genome wide changes in higher-order chromatin structure and their relationship with chromatin state and gene expression

Introduction

Higher order chromatin structure is emerging as an important regulator of gene expression. For example, *cis*-acting enhancer elements regulate their target genes by forming DNA loops that allow direct physical interactions between proteins bound at the enhancer and the promoter (1). This enhancer-promoter looping occurs even when the two elements are separated by large linear genomic distances (2). Similarly, insulator elements have been shown to adopt chromatin looping structures as part of their mechanism of insulation (3). Recently, significant advances in high-throughput methods used to identify higher-order chromatin structure have allowed for the genome wide identification of higher-order chromatin structures using a variety of methods (4).

These studies have helped reveal principles of both global and local genome organization. For instance, experiments using the 5C method and the ChIA-pet method showed that genes are often regulated by multiple enhancers forming potentially complex higher-order chromatin structures (5-7). This phenomenon has also been suggested by studies examining the co-occurrence of *cis*-regulatory activity with gene expression across many cell types and tissues (8, 9). These results indicate that the genome is

organized in part into clusters of co-regulated and physically interacting units of genes and their regulatory elements, termed enhancer promoter units (EPUs)(9).

Previous work from our laboratory and others examining higher-order chromatin structure has demonstrated the presence of topological domains as a fundamental unit of genome organization (10-12). These domains appear to be separated by insulator elements in the genome, and the position of topological domains appears to be well correlated with the locations of enhancer-promoter units (9, 10). One of the implications of these studies is that topological domains may contribute to the organization of enhancers and promoters into physically interacting units in the genome.

Despite these recent advances, our understanding of the impact of higher-order chromatin structure on gene expression is incomplete. This in part stems from the fact that most studies of higher order chromatin structure examine single genes or loci in only one or two cell types. As such, we lack datasets that allow for a comprehensive comparison of variability in higher-order chromatin structure between cell types. It is therefore difficult to determine the extent to which gene expression, cellular identity, and chromatin state are determined by higher-order chromatin structure.

To address this issue, we have generated maps of higher-order chromatin interactions using the Hi-C method in H1 human ES cells and four H1-derived lineages. These datasets, as well as previously generated Hi-C libraries from our lab in IMR90 fibroblasts, allow for a comprehensive study of the variability in higher-order chromatin structure between cell types. These cell lines have also been extensively profiled as a part of the NIH Epigenome Roadmap project (13). Therefore, we can compare the variability of higher-order chromatin structure with gene expression, DNaseI

Hypersensitivity, and chromatin state of the underlying genomic regions. This dataset has revealed several interesting observations. We observe that between different cell types, topological domains can undergo concerted, domain-wide alterations in interaction frequency. These alterations appear to be related to variation in DNaseI hypersensitivity, active histone modifications, and gene expression. In addition, we observe a re-organization of inter-domain interactions that alters the A/B compartment status between cell types. These alterations in A/B compartment are modestly correlated with changes in gene expression, and certain subset of genes show particularly strong association with their A/B compartment status. In summary, these results demonstrate that the position of the topological domains are highly stable between cell types, but that both intra- and inter-domain contacts can vary considerably between lineages and correlate with underlying gene expression and chromatin state.

Results

Hi-C in H1 hESCs and H1-derived lineages

In order to better understand the structure of topological domains across diverse cell types, we performed Hi-C experiments in H1 human embryonic stem cells (hESC) as well as four H1-derived lineages using previously established differentiation protocols (14-18). We performed two replicates of Hi-C for H1 hESCs and each of the four H1-derived lineages, including mesendodermal precurser cells (ME), Neural progenitor cells (NPC), mesenchymal stem cells (MSC), and trophoblast cells (TB). For each replicate, we obtained a minimum of 200 million uniquely mapping monoclonal read pairs. The

data was normalized as previously described (19) an analyzed using a bin size of 40kb. Each pair of replicates was highly reproducible (Figure 1).

The utility of this dataset is two-fold. First, this dataset represents the largest panel of cell lines that have been profiled by the Hi-C experiment to date, and therefore should allow a systematic analysis of the variability in higher order chromatin structure across multiple lineages. Second, as a result of our laboratories efforts in the NIH Epigenome Roadmap project, our lab and our collaborators have generated extensive maps of histone modifications, DNaseI hypersensitive sites, and gene expression data from each of these cell types (13). Therefore, we have a unique opportunity with this data to compare the dynamics of higher order chromatin interactions with the underlying chromatin states.

Topological domains positioning is highly stable across H1 and H1-derived lineages

We have previously shown that the topological domains we identified are stable between two cell in both mice and humans. We first sought to determine whether this conclusion held when considering a larger panel of cell types. We identified domains using a previous algorithm (10). By visual inspection, there is a clear similarity in the position of topological domains across all the cells analyzed (Figure 2a). We identified approximately 2,400 domains in ES cells. Using our previous described directionality index (DI), we analyzed the pattern of bias in interaction frequency in the regions flanking the boundaries of these domains to assess their stability between different cell types (10). After performing k-means clustering of the DI surrounding boundary regions, we can observe clear patterns of bias in the interaction frequency surrounding boundaries.

However, no clusters reveal a clear pattern of change in bias between cell types (Figure 2b). This suggests that the interaction bias underlying the topological domain structure is highly stable between cell types, similar to what we previously described.

The remarkable stability of the positioning of topological domains suggests that the features that contribute to domain structure may be highly stable between cell types. The most stable feature of the genome across diverse cell types is the genomic sequence itself. Therefore, we analyzed sequence features of the genome near topological domain boundaries to determine if there are particular features enriched at domain boundaries that could explain their high degree of stability.

The simplest sequence feature of the genome is the regional nucleotide content. Using the domain calls from human ES cells, we noticed a modest enrichment of GC content at the boundaries of topological domains called (Figure 2c). GC content can vary considerably in the genome, either in the form of large isochores or in more focal alterations such as CpG islands associated with genic regions (20, 21). Local variation in di-nucleotide frequency are dependent on the underlying G/C content. Therefore, we determined the observed and expected di-nucleotide frequency given the regional variation in G/C content throughout the genome. We observe a clear enrichment of CpG di-nucleotides relative to what is expected by regional GC content near topological domain boundaries (Figure 2d). This result is not entirely unexpected as we have previously observed an enrichment of marks of active genes near the boundary regions, such as H3K4me3 (Figure 2e).

We have also previously observed the presence of the insulator factor CTCF at the boundaries between topological domains (Figure 2g) (10). As CTCF is a sequence

specific DNA-binding factor, we determined the enrichment of CTCF binding motifs surrounding topological domain boundaries. We saw a clear increase in the frequency of CTCF motifs at domain boundaries, suggesting that this stable sequence feature could be a determinant of the stability of topological domains between cell types (Figure 2f). Despite this, we still observe an increase in CTCF occupancy at boundary regions as measured by the number of CTCF binding sites given the number of underlying motifs in a region (Figure 2h). This either implies that boundaries contain "stronger" motifs than the rest of the genome, or that epigenetic features may contribute to the increased likelihood of CTCF motif occupancy in these regions. In either case, it appears as though there are underlying sequence features of the genome that may contribute in part to the remarkable stability of the positioning of topological domains.

Variability in intra-domain interactions is related to the underlying chromatin state

Though the position of topological domains is highly stable between cell types, we previously observe variability of the interaction frequency within the same domains. With this expanded dataset, we sought to further characterize how interaction frequency changes between cells types. By subtracting the interaction frequency of each bin in the ES cells from the differentiated cell types, we can observe that the prominent pattern of change between cell types is local (Figure 3a). There is some degree of variability in how much each cell type differs from ES cells, with the smallest differences observed for mesendodermal cells and the largest frequency for IMR90 cells. In all cases, the largest differences in interaction frequency typically occur between loci separated less than 1Mbp from each other (Figure 3a).

We can visualize where the changes in interaction frequency occur between ES cells their differentiated progeny by displaying the change in interaction frequency using a heat map. Regions with increased interaction frequency in ES cells are shown in blue, and regions with increased interaction frequency in differentiated cells are shown in yellow (Figure 3b). The changes in interaction frequency between cell types reveal a clear pattern. Notably, there appear to be concerted, domain wide alterations in interaction frequency. For example, if a domain were re-configured between cell types such that a certain portion of the domain lost old-contacts and gained new contacts, we would observe a mixture of blue and yellow within each domain. We can see that within certain domains that this is decidedly not the case (note the blue and yellow triangles in Figure 3b). Instead, there appear to be domain-wide gains and losses in interaction frequency between cell types.

To quantify the degree to which this phenomenon occurs, we computed the average change in interaction frequency for each domain and compared this to a random average domain interaction frequency in which the interactions within a domain were drawn at random from throughout the genome (Figure 3c). We observe that the distribution of the average change in interaction frequency is more broadly distributed than would be expected at random, suggesting that domains tend to alter their interaction frequency in a concerted manner between cell types (Figure 3c). Further, for each domain we can compute the significance for the "concerted" change in interaction frequency by comparing the observed changes in interaction frequency for a given domain with what would be expected from a random "non-concerted" change based on the distribution of changes genome wide between cell types (See methods for details).

We observe that for every differentiated cell type examined at least 25% and at most 68% of domains show a domain wide concerted change in interaction frequency (Figure 3d, FDR = 0.1%). For each cell type, similar numbers of domains are increasing and decreasing in interaction frequency, though the ratio between the two varies depending on the cell type examined (Figure 3d).

The domain-wide alterations in interaction frequency are associated with changes in gene expression and the underlying chromatin state. Notably, genes within domains that increase in interaction frequency tend to be higher expressed, while genes in domains that decrease in interaction frequency tend to be lower expressed (Figure 3e). Likewise, by correlating the domain wide interaction frequency with the domain-wide alterations in various chromatin marks, we observe a strong correlation between changes in domain wide interaction frequency with changes in DNaseI hypersensitivity, CTCF, and active chromatin marks (Figure 3f, see methods for details). This correlation appears most prominent at local distances (Figure 3g, see methods for details), consistent with these regions being the most dynamic between cell types. In summary, we observe concerted, domain-wide changes in interaction frequency upon differentiation of ES cells to differentiated lineages. This alteration in interaction frequency is correlated with the activity of the domain, as increases in the interaction frequency are accompanied by increases in the presence of DHS, active chromatin marks, and increased gene expression. While interaction frequency has been correlated in the past with CTCF or DNaseI Hypersensitivity (22-25), we believe this is the first time of observing the phenomenon of domain wide alteration in interaction frequency throughout the genome.

Dynamic alteration of inter-domain contacts is lineage specific

In the initial report of the Hi-C method, the authors noted that the genome appeared to be segregated into two compartments, which they termed compartment A and compartment B. These two compartments are well correlated with a variety of genomic features, including gene density, activity of histone marks, DNaseI hypersensitivity, gene expression, DNA-replication timing, and lamina association (26, 27). This implied that Hi-C experiments can produce detailed genomic maps of the higher order spatial relationships of the euchromatin/heterochromatin organization of the nucleus. This approach therefore allows for comprehensive mapping of the spatial relationship between genomic loci for the first time.

Numerous reports examining individual loci in the genome have identified a relationship between gene expression and nuclear positioning. For instance, the immunoglobulin locus is repositioned upon activation in mature B-cells, while the globin loci have been shown to co-localize more often than would be expected at random when actively expressed in erythocytes (28, 29). This has led to the idea that alterations in gene expression could be accompanied by alteration of the spatial positioning of genes within the nucleus. While our Hi-C datasets do not give an absolute measure of positioning within the nucleus of a given locus, we can estimate the relative spatial positioning of loci using the A and B compartment information. As we have generated Hi-C datasets now in H1 hESCs and the H1-derived lineages, we sought to determine if alterations in the relative spatial positioning of genes correspond to changes in gene expression in a genome wide manner.

We generated genome wide maps of A and B compartments in H1 hESCs and H1-derived lineages as well as in IMR90 fibroblasts. We used a slight modification of the previously described algorithm that allows for the identification of A and B compartments at a resolution of 40kb. Briefly, the A and B compartments are identified from Hi-C interaction matrices that are normalized relative to the expected interaction frequency given the distance separating the two loci. These normalized matrices are converted to correlation matrices by calculating the Pearson correlation between the distance-normalized interaction frequencies of each 40kb bin in the genome. These correlation matrices are then decomposed using principle component analysis, and the values of the first principle component (PC1) are used to identify if a region is either "A" or "B" (See methods for details). PC1 was highly reproducible between replicates and yielded highly similar A and B compartment locations as the previously described algorithm, albeit at higher resolution (Figure 4a,b). Furthermore, we observe, as previously described, a clear correlation between the A and B compartment status and the chromatin state of a given 40kb bin in the genome, with the A compartment being enriched for active chromatin marks and DNaseI HS sites while the B compartment is generally enriched for the repressive chromatin mark H3K9me3 (Figure 4c).

We compared the PC1 values across different cell types and observed that the PC1 can vary considerably between different cell types. For instance, 27.5% of 40kb bins change from either A to B or B to A upon differentiation of H1 hESCs to MSCs (Figure 5a,b). For most cell types, there is an equal redistribution of regions that change from A to B as B to A. Notably, in the MSC and IMR90 lineages, there is a clear expansion of the total portion of the genome that changes to the B-compartment (Figure

5a,b). This agrees with the previously described expansion of repressive heterochromatic upon differentiation of embryonic stem cells (30). It is unclear if the fact that the expansion of the B-compartment in these lineages is related to function (connective tissue or mesenchymal origin) or perhaps to the degree of differentiation from the embryonic stem cell state (MSCs are the furthest differentiated of the H1-derived lineages, IMR90 is a terminally differentiated cell line).

In total, we identified that 41% of the genome undergoes an alteration of compartment status from A to B or B to A upon differentiation of hESCs (Figure 5d, A compartment is shown in blue, B compartment is shown in yellow). We used K-means clustering to identify patterns of alteration in the A and B compartments across the different cell lines for the 41% of the genome that changes upon differentiation. We observed two patterns from this clustering. First, transitions from B to A compartment upon differentiation appear to be cell type specific with relatively little overlap in the regions of the genome that undergo cell type specific activation (Figure 5d). In terms of the transition from A to B upon differentiation, there appears to be some cell type specific regions that undergo an A to B transition. However, the most prominent A to B transitions represent are shared between the MSC and IMR90 lineages, as previously mentioned.

We noted a striking pattern in terms of where the changes in A and B compartments occured in the genome. When we examined the changes in A/B compartment status across different cell lines, we observed that regions that were undergoing changes in A/B compartment have a clear relationship with the boundaries between topological domains. It appears as though when a region undergoes a transition

from A to B or B to A, the unit of the genome that undergoes this transition is either a single or a series to contiguous topological domains (Figure 5c,e). The regions that undergo a change in compartment status and their stable neighboring regions appear to be separated by topological domain boundaries throughout the genome. This leads us to consider a model in which topological domains are stable units of the genome upon which a more dynamic A and B compartmentalization of the genome is built.

We next examined the relationship between the changes in A/B compartment with changes in gene expression. We noted a general trend in the patterns of gene activation and repression in regions of the genome that change in compartment status. Specifically, regions that change from A to B tend to have genes that are being down-regulated, while regions that change from B to A tend to have genes that are being up-regulated (Figure 5f). These results would agree with previous data that alterations in gene expression have a relationship with alterations in spatial positioning. However, we believe the relationship between relative spatial positioning and gene expression is more nuanced. For instance, the median fold change for genes undergoing either and A to B or B to A compartment shift is at or near zero in all lineages examined, suggesting that most genes that change compartment status are not in fact changing their expression. Indeed, while we can observe genes whose expression patterns are highly correlated with their A and B compartment status (Figure 6a, see OTX2), we also observe neighboring genes whose expression patterns lack any relationship (Figure 6a, see C14orf101, EXOC5, AP5M1).

There are several possibilities that could explain the relationship we observe between gene expression and relative spatial positioning. The first possibility is that there is no relationship between gene expression and spatial positioning, and that

observing a gene such as the OTX2 with a correlation between its expression and A/B compartment status is purely random. We believe that this is not the case for multiple reasons. First, if the relationship between gene expression and spatial positioning were random, we would not expect to see the trend we observe in the fold change of expression of genes undergoing a change in compartment, though this trend is indeed subtle (Figure 5f). Second, we can calculate the Pearson correlation between the RPKM expression value for every gene in the genome and the PC1 for the transcription start site of the gene (Figure 6b). When compared with the distribution of randomly generated Pearson correlations, we observe that the actual Pearson correlation coefficients are enriched for high positive Pearson correlations relative the random control (See methods for details) (Figure 6c). Furthermore, converting the actual Pearson correlation to a rank-based p-value based on the random permutation correlation values we observe that a subset of genes have a high p-value, indicating a non-random relationship between gene expression and spatial positioning (Figure 6d).

Therefore, we favor an interpretation of these results where there exists a relationship between gene expression and spatial positioning, but only for a particular subset of genes in the genome. To identify these genes, we took a two-fold approach. First, we performed linear regression between the RPKM expression values of a gene and the PC1 of its TSS. Second, we performed a random linear regression after permuting the RPKM expression values across each of the lineages examined and calculated a rank-based p-value based on this random distribution. We considered genes with a regression coefficient greater than 125 and a Pearson correlation greater than 0.8 as having a relationship between their nuclear position and gene expression based on the distribution

of the randomized data (Figure 7a,b). We performed GO terms analysis on this subset of genes and observed numerous significant GO terms, including Cell Adhesion, Extracellular Matrix organization, and Regulation of Cell Communication (Figure 7c). It is unclear why this subset of genes has a particularly strong relationship between their expression and spatial positioning. However, the fact that we can identify significantly enriched GO terms suggests that there may be subsets of genes whose regulations may be particularly dependent on their spatial positioning within the nucleus. In summary, changes in the A and B compartment status between cell types that appear to result from alterations in inter-domain interactions. The differences in A and B compartments between cell types appears to be related to gene expression. However, this may only apply for a subset of genes in the genome, and most genes may be impervious to their relative spatial positioning.

Discussion

We have generated maps of higher-order chromatin interaction frequency in human ES cells and their differentiated progeny. These datasets represent the most extensive characterization of higher order chromatin structure and its variation between cell types to date. We have observed that topological domains are highly stable between cell types, but that they differ in their patterns of intra- and inter-domain interactions.

We have observed that within topological domains, there appear to be domain wide alterations in interaction frequency between cell types. An increased interaction frequency within a domain could be due to the domain occupying a smaller volume in the nucleus, or, it could be due to an increase in the dynamics of intra-domain interactions

within a constant volume. Hi-C data does not allow us to distinguish these two possibilities. However, these results do suggest that alterations in local interaction frequency may not be as simple as a one-to-one "looping" between two loci. Instead, global concerted alterations in structure appear to be common, in which many loci are either increasing or decreasing in interaction frequency simultaneously. We believe that these results have important implications with regards to how local regulatory elements may interact differently between cell types, and that understanding local, concerted changes in chromatin interactions may be crucial toward understanding how *cis*-regulatory elements find their targets genes either individually or in combination.

We have also observed re-organization of inter-domain contacts that result in the alteration of the A and B compartments in the genome between cell types. This alteration appears to correlate with modest changes in gene expression that may also be specific to particular subsets of genes. We believe that the information gleaned from analysis of the A and B compartments can provide high resolution, genome wide information as to the relative spatial positioning of genomic loci. Therefore, we believe these results suggest that most genes that alter their relative spatial positioning do not appear to change expression, but that the expression of particular subsets of genes may be highly sensitive to spatially regulated signals.

Lastly, as we have previously described, the position of topological domains in the genome is extremely stable, and we believe this stability suggests an essential function of topological domains as a structural unit of the genome. The regulation of interaction frequency either within or between domains likely plays a more critical role in the establishment of cellular identity.

Methods

Cell Culture, ChIP-Seq, and Hi-C experiments: H1 human embryonic stem cells were cultured under feeder free conditions on matrigel coated plates. Differentiation of ES cells into H1-derived lineages was performed using previously described protocols (14-18). Cells were harvested for Hi-C with 1% formaldehyde for 10 minutes at room temperature. Two biological replicates were performed for each lineage. Hi-C was performed as previously described (26). Hi-C libraries were sequenced using the Illumina Hi-Seq 2000. ChIP-seq experiments for CTCF were performed as previously described (9) and sequenced on the Hi-Seq 2000. All other ChIP-seq experiments are from the NIH Epigenome Roadmap project.

***Hi-C library mapping and data processing*:** For all Hi-C libraries, each read-pair was mapped to the hg18 human genome build independently using BWA with default parameters. Only reads that mapped with a mapping quality >10 were kept as uniquely mapping and PCR duplicate read pairs were removed using Picard MarkDuplicates. We generated raw Hi-C interaction matrices with a bin size of 40kb using an in house pipeline and normalized the matrices as previously described (19). The Pearson correlation between replicates was performed only considering bins separated by less than 2 million base pairs, as the vast majority of the data is found in this region. Topological domains were called as previously described (10).

Sequence Feature Enrichment at Domain Boundaries: Enrichment of various features surrounding topological domain boundaries was performed using a +/- 500kb window surrounding the center of each called domain boundary. The enrichment of each feature within 10kb bins within this window was computed. For GC content, we computed the fraction of G or C nucleotides within each 10kb bin. For CpG content, we computed observed CpG dinucleotide frequency in each 10kb bin relative to the expected frequency of CpG dinucleotides. The expected frequency is simply the frequency of observing a C followed by a G, which is the product of the C-frequency and the G-frequency over a 10kb window.

For the CTCF motif enrichment, we first identified the presence of CTCF motifs as follows. We used HOMER to call *de novo* motifs in the MSC lineage using ChIP-Seq data we have generated. Using the position weight matrix generated from the peak calling, we identified CTCF motifs throughout the genome using STORM with a p-value cut off of 0.8. The frequency of CTCF motifs was calculated for every 10kb bin. CTCF occupancy was measured as the frequency of CTCF peaks in a given 10kb bin relative to the number of CTCF motifs in the same 10kb bin. For the enrichment of both H3K4me3 and CTCF ChIP-seq signal at domain boundaries, we computed the frequency of peak calls for each mark surrounding the boundaries using MACS to generate peak calls.

Changes in Intra-Domain interaction frequency: To compute the change in interaction frequency between cell types, we first merged the Hi-C data between two replicates for each cell type. The merged, normalized interaction matrices were quantile normalized between all lineages to accommodate for differences in frequency strictly due to

sequencing depth. The differences between cell types were computed by simply subtracting the interaction frequency of each bin I_{ij} of ES cells from the differentiated cell types. Domain wide averages in interaction frequency were calculated as the arithmetic average of the difference in interaction frequency between cell types for each potential interaction within a domain using 40kb bins. To compute the value expected for each domain if the changes were not occurring in a concerted manner, in other words, if the domains were randomly changing with both increased and decreased interaction frequencies, we randomly selected interacting bins from the genome and computed the domain wide average while preserving the number of interactions at each distance in a given domain. This randomization was performed 10 times for each domain. Significantly "concerted" domains were assessed by using a Wilcoxan test between the actual and each of the randomized domains. The reported p-value is the average of the log base-10 of each of these p-values. The FDR was calculated using Benjamini correction.

The domain wide correlation in interaction frequency with various histone modifications and chromatin marks was calculated as follows. One vector was constructed from the differences in interaction frequency for each potential interaction in a domain. A second vector was created using the sum of the presence of each chromatin mark in the two interacting bins multiplied by a weight based on the average Hi-C interaction frequency between loci separated by a given genomic distance. The two vectors were used to calculate a Pearson correlation. We also correlated the change in interaction frequency and the presence of chromatin modifications at each distance up to 2Mb. In this case, no distance dependent weight was applied.

Identification of A and B compartments: Identification of A and B compartments was performed conceptually similarly to what has been previously described, though with several modifications. We used the normalized 40kb interaction matrices for each cell type and calculated the expected interaction frequency between two 40kb bins given the distance separating them in the genome. We used a sliding window approach with a bin size of 400kb and a step side of 40kb to generate an observed/expected matrix at a 40kb bin size. The observed frequency was the sum of all observed interaction frequencies of the 40kb bins making up the larger 400kb bin. Likewise, the expected frequency was the sum of the expected frequencies of each of the 40kb bins making up the larger 400kb bin. This value was used to generate the observed over expected matrix. This was then converted to a Pearson correlation matrix and subsequently used for principle component analysis as previously described (26). The first principle component for each chromosome was used to identify regions of the genome as belonging to either the A or B compartment. The direction of the Eigen values is arbitrary, and therefore positive values were set to "A" and negative were set to "B" based on their association with gene density. A switch from A to B or B to A was considered if a given bin changed from a negative to a positive Eigen value (or vice versa) between cell types.

The Pearson correlation and regression between the PC1 values and gene expression over a given gene was performed as follows. The log of the RPKM expression value for each gene was used as the dependent variable for regression and the independent variable was the Eigen value of the bin containing the TSS for the given gene. The random correlation and regression values were computed by randomizing the

expression values for each cell type of a given gene. GO terms analysis was performed using DAVID.

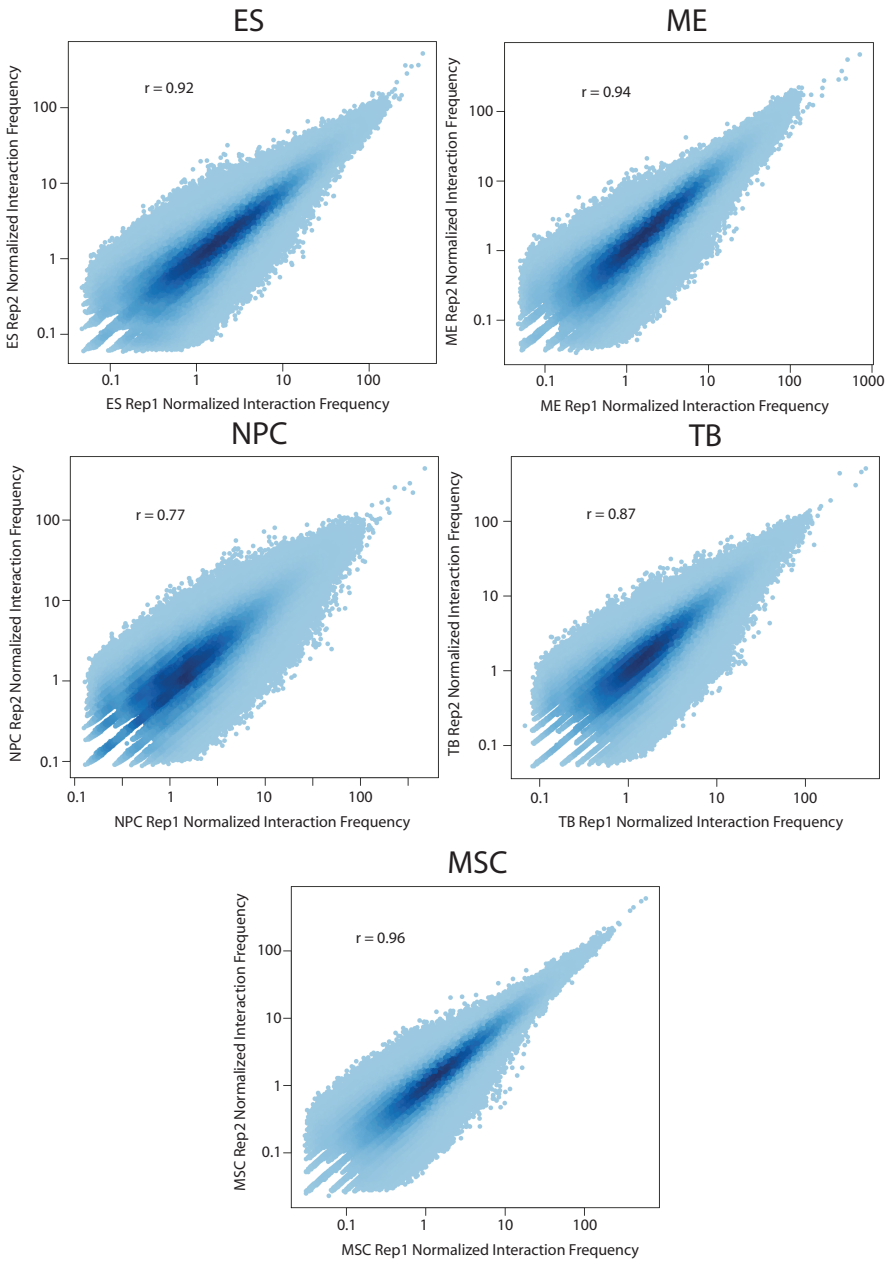


Figure 1. Reproducibility of Hi-C data. Scatter plots showing the correlation between replicates for each for H1 ES cells and H1-derived lineages. The Pearson correlation between replicates is shown in each plot.

Figure 2. Stability of Topological Domains. a, Heat maps showing the normalized interaction frequency in each of the lineages examined. Also shown is the directionality index and the domain calls in ES cells as well as the position of RefSeq genes. b, Directionality index surrounding topological domain boundaries in each lineages. For each cell type, the directionality index in a +/- 1Mb region is shown surrounding the boundaries called in ES cells. Upstream biased regions are shown in yellow, and downstream biased regions are shown in blue. K-means clustering (k = 20) was performed. The pattern of bias is highly similar across cell types. c, The GC content percentage in 10kb bins in a +/- 500kb window surrounding the boundaries in ES cells. d, CpG dinucleotide frequency surrounding domain boundaries. The observed dinucleotide frequency relative to the underlying GC content was calculated in 10kb bins surrounding each boundary in a +/- 500kb window. e, The frequency of H3K4me3 binding sites surround each boundary. f, The frequency of CTCF motifs per 10kb surrounding the boundary. g, The frequency of CTCF binding sites per 10kb surrounding each boundary. h, CTCF occupancy surrounding the boundaries. CTCF occupancy is measured as the number of CTCF binding sites per motif in each 10kb bin.

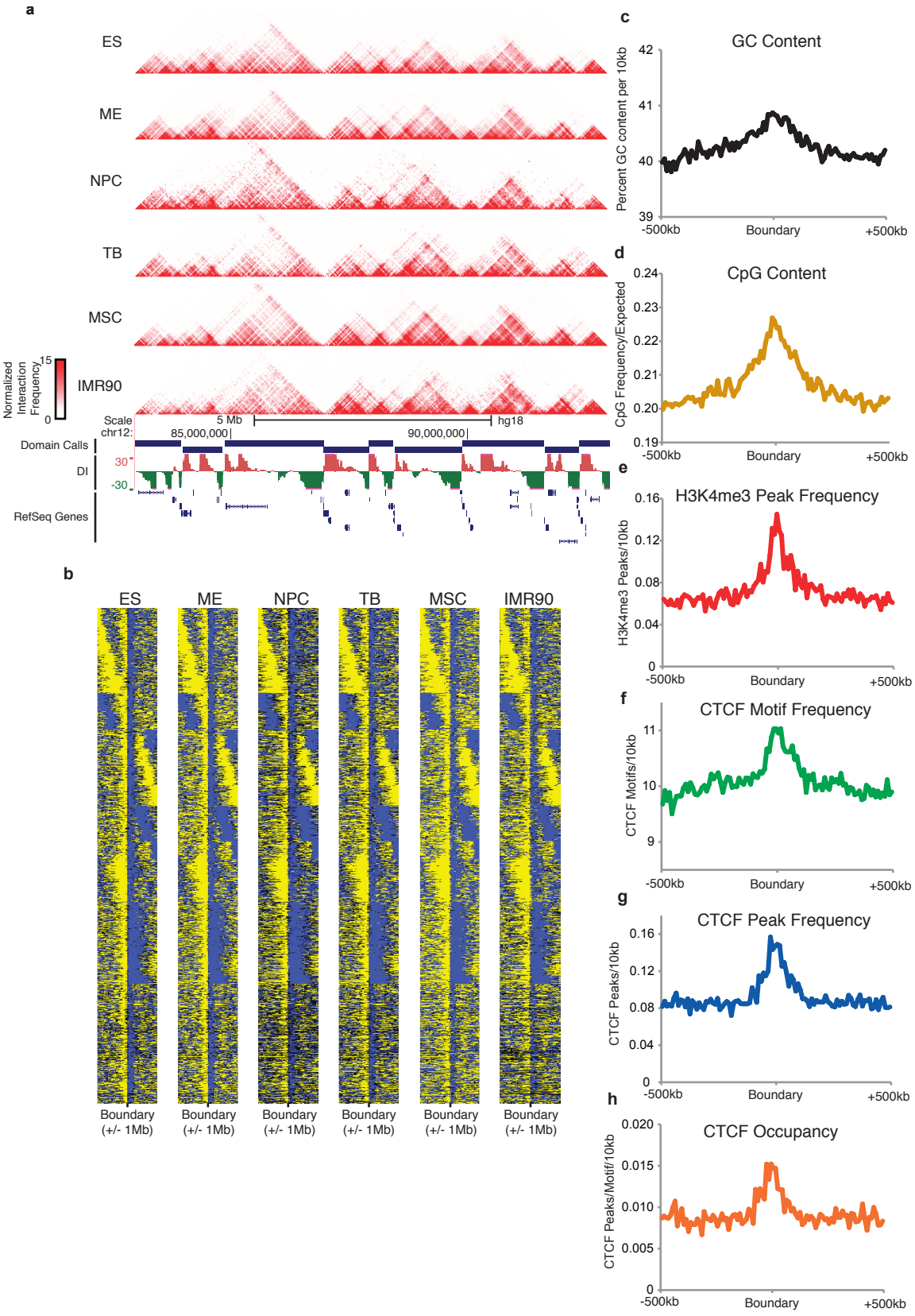

a
ES
ME
NPC
TB
MSC
IMR90
Normalized Interaction Frequency
15
0
Scale
chr12:
5 Mb
hg18
85,000,000
90,000,000
Domain Calls
DI
30
-30
RefSeq Genes
b
ES
ME
NPC
TB
MSC
IMR90
Boundary (+/- 1Mb)
c
GC Content
Percent GC content per 10kb
42
41
40
39
-500kb
Boundary
+500kb
d
CpG Content
CpG Frequency/Expected
0.24
0.23
0.22
0.21
0.20
0.19
e
H3K4me3 Peak Frequency
H3K4me3 Peaks/10kb
0.16
0.12
0.08
0.04
0
f
CTCF Motif Frequency
CTCF Motifs/10kb
11
10
9
g
CTCF Peak Frequency
CTCF Peaks/10kb
h
CTCF Occupancy
CTCF Peaks/Motif/10kb
0.020
0.015
0.010
0.005

Figure 3. Alterations in intra-domain interaction frequency between lineages. a, Plot showing the average absolute value of the change in interaction frequency between cell types (y-axis) versus genomic distance. The most prominent changes observed are less than 1Mb. b, Heat map of the change in interaction frequency between cell types. The difference between MSC and ES cells is shown, as is the DNaseI HS frequency in ES cells and MSC cells, as well as the directionality index in ES cells. c, Distributions of the average domain wide change in interaction frequency for each domain. Shown in grey is the distribution of the change in interaction frequency using randomly generated domain averages. d, Proportion of domains showing a significant concerted change in interaction frequency. Domains with an increase are shown in gold, domains with an average decrease are shown in green, and domains with no significant change are shown in grey (FDR = 0.1%, Wilcoxan test with Benjamini correction). e, Distribution of the fold change in gene expression upon differentiation for each found in domains that have an increased interaction frequency ("+"), a decreased interaction frequency ("-"), or no change in interaction frequency ("o"). f, Distribution of Pearson correlations between various chromatin modifications, CTCF, and DHS with changes in interaction frequency on a domain-wide level. g, Pearson correlation of various chromatin modifications, CTCF, and DHS with changes in interaction frequency based on the distance separating the two interaction loci.

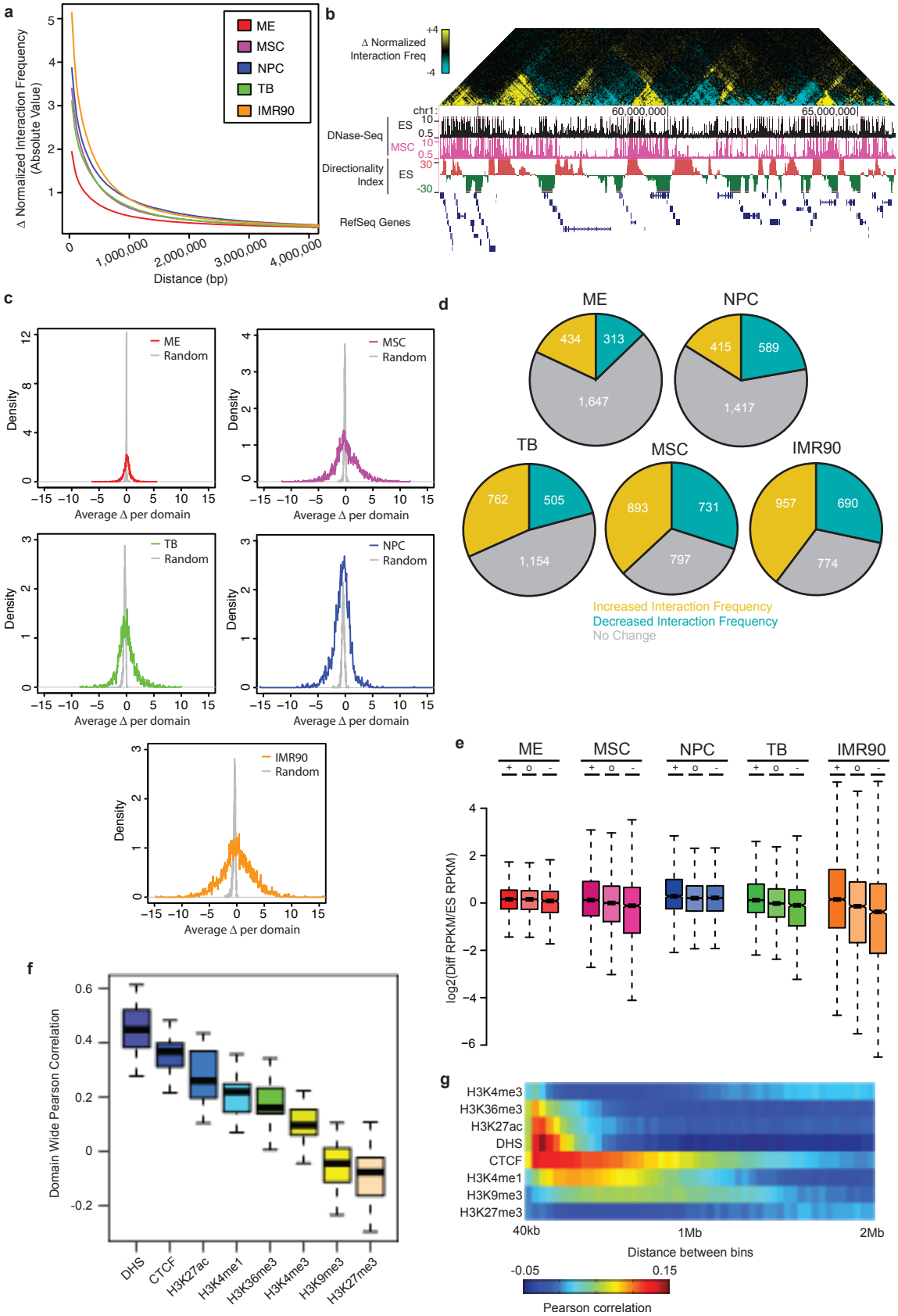

a
Δ Normalized Interaction Frequency (Absolute Value)
ME
MSC
NPC
TB
IMR90
Distance (bp)
b
Δ Normalized Interaction Freq
chr1:
60,000,000
65,000,000
DNase-Seq
ES
MSC
Directionality Index
ES
RefSeq Genes
c
Density
Average Δ per domain
d
ME
NPC
TB
MSC
IMR90
Increased Interaction Frequency
Decreased Interaction Frequency
No Change
e
log2(Diff RPKM/ES RPKM)
f
Domain Wide Pearson Correlation
DHS
CTCF
H3K27ac
H3K4me1
H3K36me3
H3K4me3
H3K9me3
H3K27me3
g
Distance between bins
Pearson correlation

Figure 4. Identification of A/B compartments. a, Browser shots of the A/B compartments using a previously generated approach and using our sliding window approach. b, Correlation of the PC1 values between replicates show that the sliding window approach is highly reproducible between replicates. c, Spearman correlation of the PC1 values with histone modifications and DHS for each cell type shows that active marks are correlated with the PC1 values.

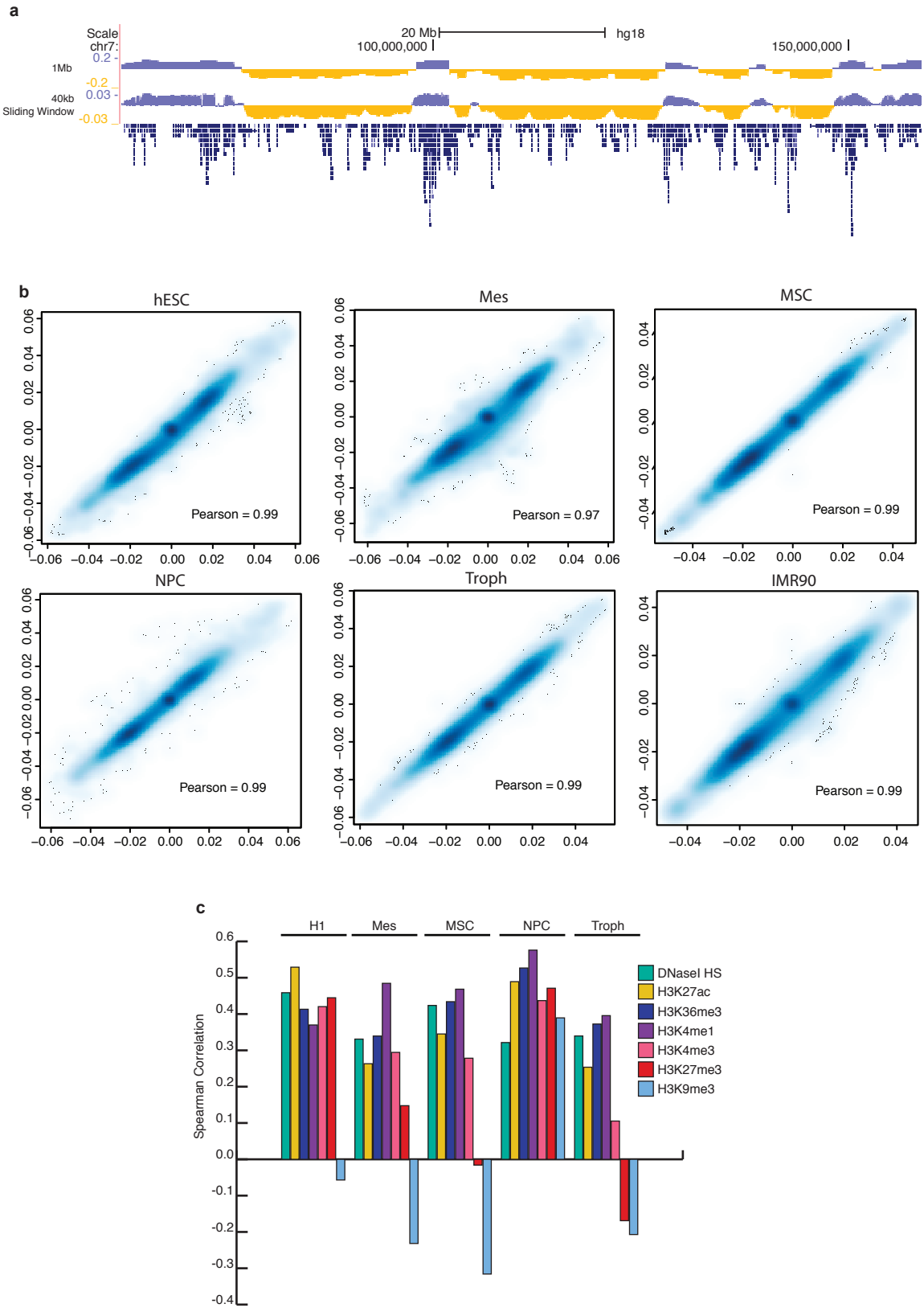
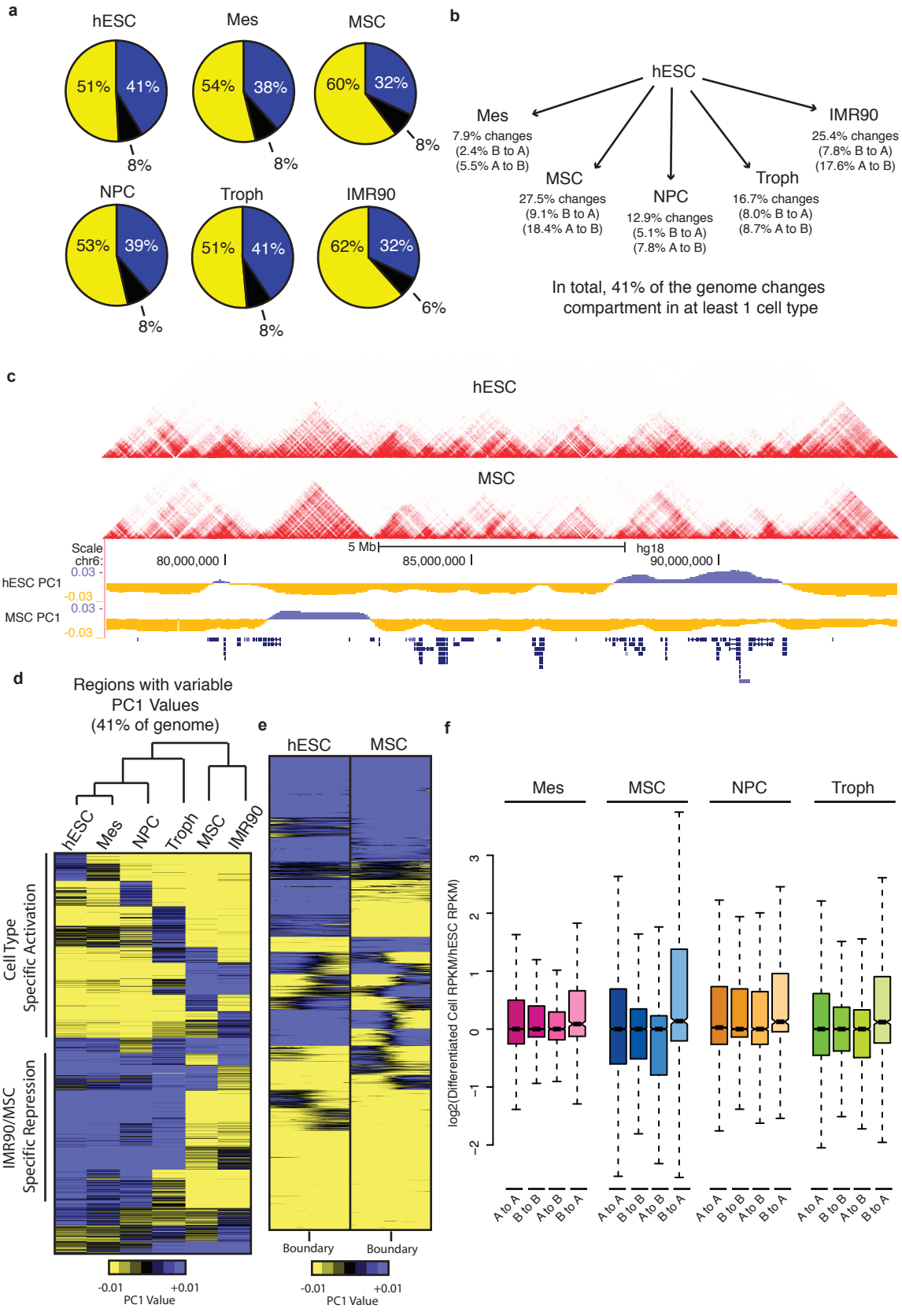

a
Scale
chr7:
1Mb
40kb
Sliding Window
20 Mb
hg18
100,000,000
150,000,000
b
hESC
Mes
MSC
NPC
Troph
IMR90
Pearson = 0.99
Pearson = 0.97
c
H1
Mes
MSC
NPC
Troph
Spearman Correlation
DNaseI HS
H3K27ac
H3K36me3
H3K4me1
H3K4me3
H3K27me3
H3K9me3

Figure 5. Inter-domain changes in interaction frequency and the A/B compartments. a, Proportion of the genome that is either A compartment (blue) or B compartment (yellow) or either (black, mostly unmappable regions) in each lineage. b, Diagram showing the percent of the genome that changes its A or B compartment status upon differentiation of ES cells. c, Heat maps of topological domains and in ES and MSC and the underlying PC1 values over these regions. The regions of the genome that switch compartments tend to occur in units of single or multiple contiguous topological domains. d, K-means clustering of the PC-1 values for each of the 41% of the bins in the genome that change A/B compartment status between cell types. Lineages were clustered hierarchically to show the differences in A/B compartment between lineages. e, Heat map of the PC1 values from ES and MSC surrounding each of the topological domain boundaries in ES cells. Switching between A/B compartments tends to occur at topological domain boundaries. f, Distribution in the fold change in interaction frequency between the differentiated cell types and ES cells for genes that change from A to A, B to B, A to B, and B to A.


a
hESC
51%
41%
8%
Mes
54%
38%
8%
MSC
60%
32%
8%
NPC
53%
39%
8%
Troph
51%
41%
8%
IMR90
62%
32%
6%
b
hESC
Mes
7.9% changes
(2.4% B to A)
(5.5% A to B)
MSC
27.5% changes
(9.1% B to A)
(18.4% A to B)
NPC
12.9% changes
(5.1% B to A)
(7.8% A to B)
Troph
16.7% changes
(8.0% B to A)
(8.7% A to B)
IMR90
25.4% changes
(7.8% B to A)
(17.6% A to B)
In total, 41% of the genome changes
compartment in at least 1 cell type
c
hESC
MSC
Scale
chr6:
80,000,000
5 Mb
85,000,000
hg18
90,000,000
hESC PC1
0.03
-0.03
MSC PC1
0.03
-0.03
d
Regions with variable
PC1 Values
(41% of genome)
hESC
Mes
NPC
Troph
MSC
IMR90
Cell Type
Specific Activation
IMR90/MSC
Specific Repression
-0.01
+0.01
PC1 Value
e
hESC
MSC
Boundary
Boundary
-0.01
+0.01
PC1 Value
f
Mes
MSC
NPC
Troph
log2(Differentiated Cell RPKM/hESC RPKM)
3
2
1
0
−1
−2
A to A
B to B
A to B
B to A

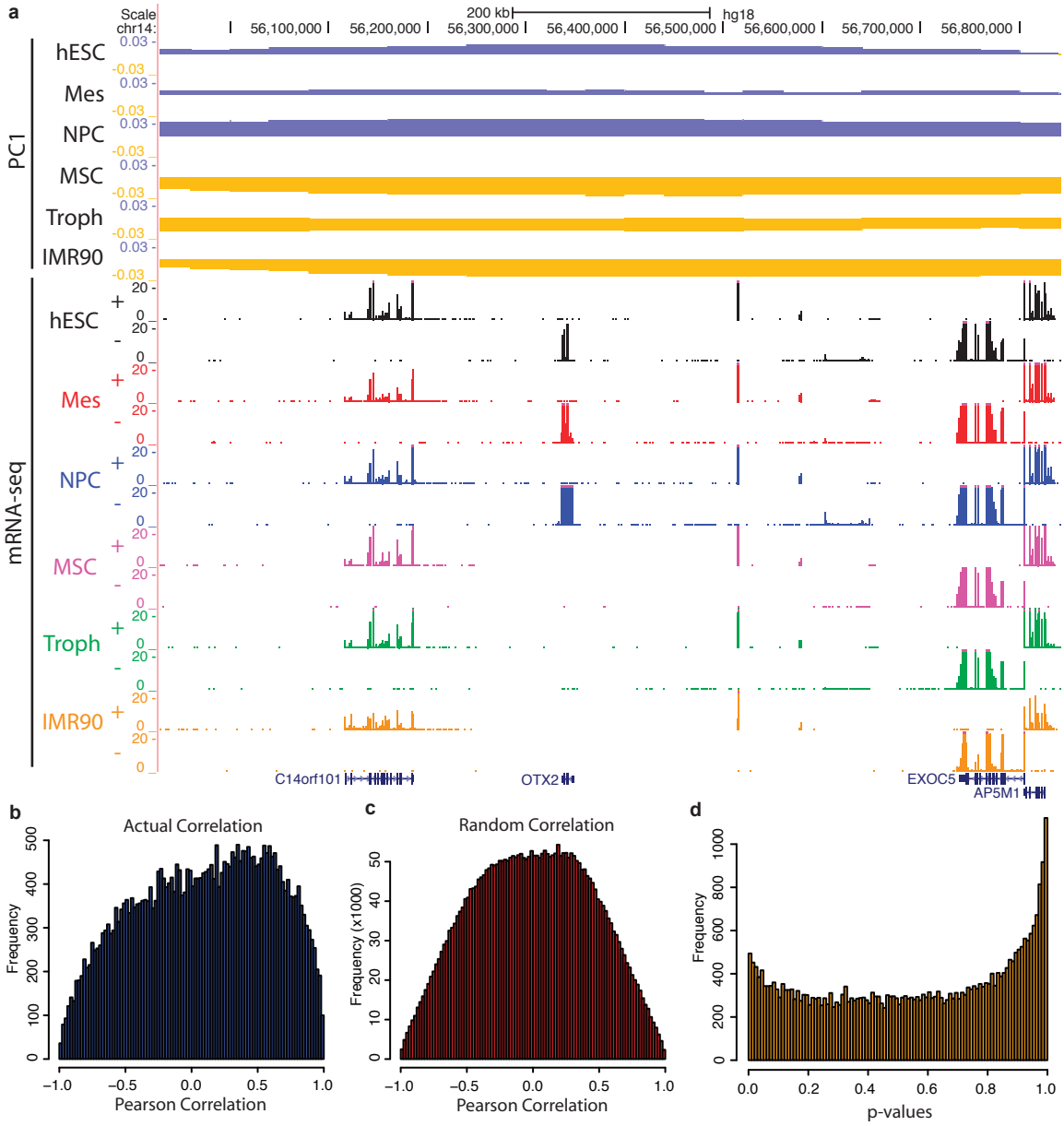
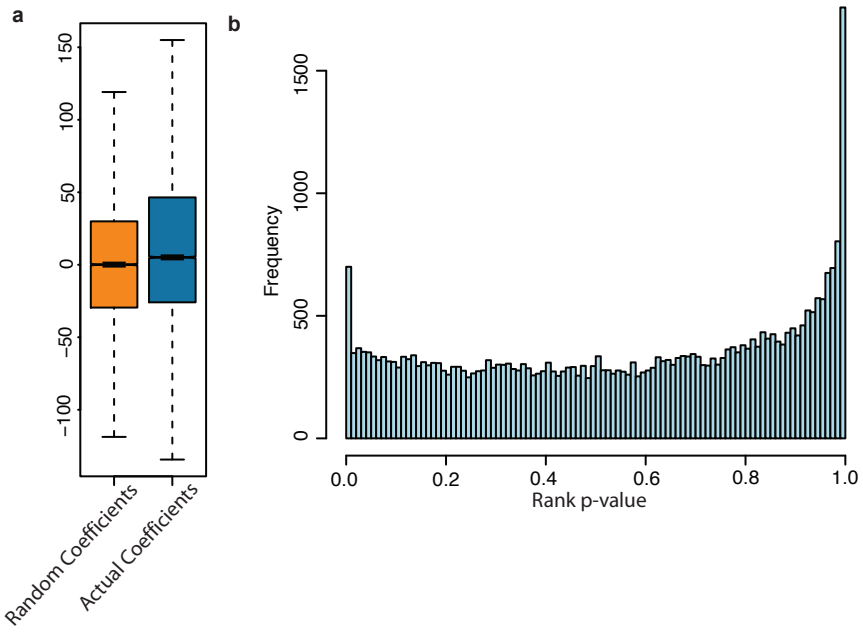


Figure 6. Association between A/B compartment changes and gene expression. a, Browser shot showing the RNA-seq expression values of several genes and the PC1 values of the genomic regions surrounding them. Note the pattern of OTX2 expression matches the PC1 values and the A/B compartment statues. b, Distribution of Pearson correlations between the PC1 values and the log of the RNA-seq RPKM expression values for each gene. c, Distribution of the Pearson correlations between the PC1 values and the log of the RNA-seq RPKM expression values for each gene after randomizing the RPKM values for a given gene among the six lineages. d, Distribution of rank based p-values of the Pearson correlations. The p-value is assigned for each gene after by computing the rank of the actual Pearson correlation with 1000 randomly generated Pearson correlations by shuffling the RNA-seq values of a given gene between each of the lineages.



c

Term	Count	PValue	Benjamini
Cell adhesion	118	1.22E-09	4.64E-06
Biological adhesion	118	1.31E-09	2.49E-06
Positive regulation of cell proliferation	73	4.57E-07	5.82E-04
Regulation of cell proliferation	119	5.16E-07	4.93E-04
Tube development	44	4.87E-06	0.0037115
Cell-substrate adhesion	25	1.36E-05	0.00860783
Circulatory system process	37	3.50E-05	0.01891013
Blood circulation	37	3.50E-05	0.01891013
Regulation of cell adhesion	30	3.64E-05	0.01722545
Morphogenesis of an epithelium	24	6.99E-05	0.02923538
Cell-matrix adhesion	22	8.27E-05	0.03110541
Enzyme linked receptor protein signaling pathway	56	9.75E-05	0.03330223
Actin cytoskeleton organization	41	1.08E-04	0.03376863
Extracellular matrix organization	24	1.13E-04	0.03255033
Positive regulation of cell adhesion	17	1.24E-04	0.03320881
Vasculature development	44	1.31E-04	0.03274353
Positive regulation of signal transduction	49	1.97E-04	0.04606065
Actin filament-based process	42	2.16E-04	0.04744955
Positive regulation of cell communication	53	2.33E-04	0.04816358

Figure 7. Identification of Genes associated with changes in A/B compartment status. a, Distribution of regression coefficients between the PC1 values and the RNA-seq expression values. Least-squares linear regression was performed on the actual data and on data where the RNA-seq values were randomized for a given gene across each of the lineages 1000 times. b, The rank-based p-value of the actual regression coefficient compared to the randomly generated coefficients demonstrates that a subset of genes have a strong relationship between their expression and compartment status. c, GO terms table of genes with a regression coefficient of at least 125 and p-value > 0.8.

References

1. Smallwood A, Ren B. Genome organization and long-range regulation of gene expression by enhancers. Curr Opin Cell Biol.

2. Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, Breedveld GJ, et al. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A. 2002;99(11):7548-53. PMCID: 124279.

3. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009;137(7):1194-211. PMCID: 3040116.

4. Sajan SA, Hawkins RD. Methods for identifying higher-order chromatin structure. Annu Rev Genomics Hum Genet.13:59-82.

5. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature.489(7414):109-13. PMCID: 3555147.

6. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009;462(7269):58-64. PMCID: 2774924.

7. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell.148(1-2):84-98. PMCID: 3339270.

8. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature.473(7345):43-9. PMCID: 3088773.

9. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, et al. A map of the cis-regulatory sequences in the mouse genome. Nature.488(7409):116-20.

10. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature.advance online publication.

11. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation center. in submission. 2012.

12. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell.148(3):458-72.

13. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Pradipta R, et al. Epigenomic Analysis of Multi-lineage Differentiation of Human Embryonic Stem Cells. Cell. 2013;In Press.

14. Xu RH, Chen X, Li DS, Li R, Addicks GC, Glennon C, et al. BMP4 initiates human embryonic stem cell differentiation to trophoblast. Nat Biotechnol. 2002;20(12):1261-4.

15. Yu P, Pan G, Yu J, Thomson JA. FGF2 sustains NANOG and switches the outcome of BMP4-induced human embryonic stem cell differentiation. Cell Stem Cell.8(3):326-34. PMCID: 3052735.

16. Chambers SM, Fasano CA, Papapetrou EP, Tomishima M, Sadelain M, Studer L. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat Biotechnol. 2009;27(3):275-80. PMCID: 2756723.

17. Chen G, Gulbranson DR, Hou Z, Bolin JM, Ruotti V, Probasco MD, et al. Chemically defined conditions for human iPSC derivation and culture. Nat Methods.8(5):424-9. PMCID: 3084903.

18. Vodyanik MA, Yu J, Zhang X, Tian S, Stewart R, Thomson JA, et al. A mesoderm-derived precursor for mesenchymal stem and endothelial cells. Cell Stem Cell.7(6):718-29. PMCID: 3033587.

19. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics.28(23):3131-3. PMCID: 3509491.

20. Eyre-Walker A, Hurst LD. The evolution of isochores. Nat Rev Genet. 2001;2(7):549-55.

21. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.

22. Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. 2006;20(17):2349-54. PMCID: 1560409.

23. Carter D, Chakalova L, Osborne CS, Dai YF, Fraser P. Long-range chromatin regulatory interactions in vivo. Nat Genet. 2002;32(4):623-6.

24. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell. 2002;10(6):1453-65.

25. Palstra RJ, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W. The beta-globin nuclear compartment in development and erythroid differentiation. Nat Genet. 2003;35(2):190-4.

26. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289-93.

27. Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res.20(6):761-70. PMCID: 2877573.

28. Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet. 2004;36(10):1065-71.

29. Kosak ST, Groudine M. The undiscovered country: chromosome territories and the organization of transcription. Dev Cell. 2002;2(6):690-2.

30. Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell.6(5):479-91. PMCID: 2867844.

Chapter 5

Conclusion

Summary

We have undertaken several studies using the Hi-C method to better characterize higher-order chromatin structure in mammalian organisms. First, by performing Hi-C in mouse ES cells, mouse cortex, human ES cells, and human IMR90 fibroblasts, we observed structures that we term topological domains (1). These domains are stable between cell types and are separated by insulator elements in the genome. The domains are well conserved in evolution, which suggests that they may be functional. Other work from our lab characterizing the activity of enhancers and promoters across a multitude of cell types have identified units of co-regulation between enhancers and promoters, termed enhancer-promoter units (EPUs). The location of EPUs appears to be correlated with the location of topological domains throughout the genome (2). Therefore, we speculate that topological domains are related to the organization of enhancers and promoters into higher-order regulatory structures.

These studies raised several questions, notably, what factors are important for topological domain structure? We had observed enrichment of several factors, including CTCF and the Cohesin complex, at the boundaries that separate topological domains. CTCF and Cohesin have both been suggested to play a role in organizing higher-order chromatin structure (3). Therefore, we considered them likely candidate organizers of topological domains. In collaboration with Kerstin Wendt's laboratory, we investigated

the changes in higher-order chromatin structure after depletion of Cohesin or CTCF. Despite their shared binding profiles in the genome, loss of CTCF and Cohesin appears to have different affects on chromatin organization. Within topologoical domains, depletion of CTCF or Cohesin leads to a reduction in interaction frequency, albeit at different spatial scales. Between topological domains, loss of CTCF but not Cohesin leads to an increase in interactions between domains. Remarkably, the genes affected by CTCF and Cohesin loss are nearly entirely different, suggesting that at both the structural and gene regulatory level the two factors are independent of each other.

Lastly, we have also expanded the number of cell types in which we have generated genome wide interaction maps by performing Hi-C in H1 human ES cells and H1-derived cells. By comparing the alterations in chromatin structure to the alterations in the underlying chromatin states and gene expression, we have observed correlations between alterations in chromatin structure and regulatory activity. Notably, we observe that some topological domains will undergo domain-wide alterations in interaction frequency. These changes correlate with changes in DNaseI hypersensitivity, active chromatin marks, and gene expression over the domain. This suggests that interactions may not differ as a one-to-one alteration in looping structure. Instead, alterations in chromatin structure between cell types may occur as concerted, systematic, and complex changes. We also observe alteration in inter-domain contacts that are manifest as changes in the A and B compartments between cell types. These compartment changes correlate with alteration in gene expression, though this may only be relevant for particular subsets of genes. These differences in A/B compartment status also appear to be occurring at the level of one or several contiguous topological domains. Therefore, we

believe that topological domains are the basic unit upon which the A/B compartment structure is built.

Future Perspectives

Technical Challenges

One of the great promises of Hi-C as a method is the ability to interrogate interactions on a genome wide scale at a resolution capable of discerning regulatory element interactions. Though certain technical challenges may make this difficult, the primary limitation to performing Hi-C at such a resolution is cost. With the dramatic rate of reduction in sequencing cost that we are observing, the day when the cost of this experiment is no longer a major limitation is coming. It is worthwhile to note both the methodological and analytical challenges to making genome wide interrogation of regulatory element interactions a reality.

The primary methodological limitation to probing genome-wide interactions at a regulatory element scale is library complexity. In the Hi-C libraries that we have produced, library complexity is the major limiting factor outside of cost in achieving a large number of reads. This can be controlled by simply performing the fewest number of PCR cycles as possible to obtain enough DNA to sequence. Using the Picard MarkDuplicates tool, we can estimate that libraries with poor complexity contain at most 100-200 million unique reads, which is insufficient to discern regulatory-element scale interactions. High complexity libraries, on the contrary, can have up to 2 billion unique reads within a single library. We believe that obtaining such a high number of

sequencing reads for a library will allow for the identification of genome wide regulatory element scale interactions.

Several other technical challenges exist outside of library complexity in performing Hi-C at a regulatory element scale interaction. The initial technical challenge in constructing a Hi-C library is to limit the number of contiguous, non-ligated DNA fragments that "contaminate" the Hi-C library. These sequences are essentially leftover genomic DNA fragments that are selected against but not entirely removed during the protocol. We have observed that failed Hi-C libraries can contain upward of 80% of the library as these short, contiguous fragments, which have also been termed "dangling ends" (4). We do not have systemic evidence to identify the source of these fragments in a Hi-C library, but they are likely due to a combination of several factors, including cross-linking efficiency, star activity of the restriction enzyme, ligation efficiency, and the quality of biotin removal from non-ligated DNA ends. These are factors that are not insurmountable, and we believe that relatively minor technical advances will make it easy for most laboratories to perform Hi-C experiments where the number of these contaminating sequences is limited.

A second technical challenge in library construction in maximizing the ratio of *cis* interacting reads relative to *trans* interacting reads. Alterations in the Hi-C method such as Tethered Conformation Capture (TCC) have been designed to maximize this ratio (5). We have observed that Hi-C libraries can vary considerably in their *cis* to *trans* ratio, varying 1:5 to a 5:1. There are likely numerous factors that contribute to this variability. In terms of technical issues, the primary factor is the number of non-specific inter-molecular ligation events that occur in a library. We have seen that the *cis* to *trans* ratio

can vary somewhat between replicates of the same cell type, so undoubtedly technical variation can contribute to the *cis* to *trans* ratio. However, we have also observed that the largest degrees of variation in the ratio occur between cell types, which may reflect differences in biology. Increased DNA content in the nucleus, potential due to the number of cells in G1 versus G2/S likely contributes to this. Other factors such as the compactness of the nucleus and the degree of "intermingling" among chromosome territories may also play a factor as well. This will likely be an important technical factor to optimize in the future to allow for the generation of the highest quality Hi-C libraries.

Outside of the wet-lab aspects of Hi-C, analytical challenges will also contribute to our ability to utilize Hi-C as a method for probing genome wide regulatory element interactions. One of the most obvious analytical challenges will be determining when two elements are "interacting" or not. Our lab has been developing algorithms not discussed in this dissertation to address such a problem and inevitably others will develop alternative methods as well. However, the primary challenge in this regard is not methodological but is instead philosophical. What does it mean for two elements in the genome to "interact?" We can detect sequencing reads at all distances throughout the genome and between chromosomes in non-random manners. This inevitably means that in a large enough population of cells, nearly every possible interaction that can occur likely will occur, and that the primary distinction is a matter of frequency. Therefore, the central question is, how frequently do two elements need to interact to be considered interacting or not interacting?

Perhaps to answer this question we need to draw on lessons learned from the analysis of gene expression data. In an RNA-seq or microarray analysis of gene

expression, extremely large numbers of transcripts can be detected in a given cell type, in some cases numbering in the tens of thousands of transcripts. The critical information that can be drawn from RNA expression experiments are typically not the absolute level of expression but relative levels of expression. For instance, house-keeping genes likely need to be expressed at much higher levels than non-housekeeping genes. Likewise, low levels of expression of a gene in one cell type may be highly important compared to the absence of expression is different cell type. Analogously, certain higher-order chromatin interactions may be "housekeeping" interactions that are critical for fundamental structures of the genome, such as topological domains. Other interactions may be more variable between cell types and contribute to cellular identity. Therefore, it likely critical that interaction frequency is considered as a relative entity, either compared within a cell type or between different cell types.

With regards to the comparison of relative interaction frequencies between cell types, particularly at a regulatory element resolution, another major challenge is that of multiple testing. If we can detect 20,000 enhancers in a given cell type and we are interesting in detecting their interactions with 20,000 promoters between two cell types, we must perform an extremely large number of statistical tests due to the frequency of regulatory elements and the fact that we are probing the interactions between them. This leads to the problem of correcting for multiple testing, which, when so many interactions are considered, it may be difficult to perform multiple testing corrections that do not remove large numbers of true differences in interactions frequency. This may require even more extreme numbers of sequencing or algorithms that limit the number of tests performed based on prior knowledge.

Overall, the technical challenges to performing high-resolution Hi-C are not insurmountable. Potentially though either modest technical improvements in the methods and analysis, studying regulatory element interactions using Hi-C data will be feasible. Additional methods, such as more high-throughput FISH studies, may also be of great promise in the future to validate and expand upon the conclusions drawn form these studies.

Implications

In the future, I predict there will be many areas that will be impacted by the study of higher order chromatin structure. There are three in particular that I believe are particularly ripe for advancement as a product of studies using Hi-C and related techniques. These areas are gene regulation by enhancers, large-scale evolution of genomes, and cancer biology.

As the above discussion points out, one of the future promises of Hi-C data is to understand interactions between regulatory elements. One interesting avenue for future research is to define the relationship between interaction frequency and functional activation. For instance, if an enhancer and promoter interact frequently, is that sufficient to drive the expression of the target gene? Likely this is too simplistic of a model, and other factors, such as the particular combinations of transcription factors and co-activators, will also be critical for determining activation. This likely will have an impact in terms of which enhancers ultimately activate which promoters. We have seen that the strongest interactions tend to be local, so that enhancers that exist at long distances from their target promoter will likely have some degree of interactions with off target

promoters. Are these non-specific interactions functional, and if not, what confers specificity in the face of a potential highly dynamic interacting landscape?

With regards to evolution, one of the most interesting observations about the positioning of topological domains is that they appear to be quite stable between humans and mice. Other recent reports have also demonstrated the presence of topological domains in *Drosophila*, suggesting that this structure is deeply conserved in evolution (6). Notably, this domain structure appears to be absent in yeast (7). We have shown that Cohesin and CTCF both play a role in organization of topological domains, and it is interesting to note that yeast and *C. elegans* lack a clear CTCF homolog (8). Notably, *C. elegans* may be an exception among nematodes in lacking CTCF. One possibility is that the presence of topological domains co-evolved with genome organizers such as CTCF. Future studies examining domain organization in other organisms will hopefully shed light on this problem.

The genes encoding proteins such as CTCF are one source for the impact of evolution on chromatin structure. Another is the alteration of the genome structure itself. Our genomes are not stable units in evolution, and breaks in the syntenic structure of chromosomes can alter the proximity of genes to neighboring regulatory elements. Likewise, duplications, inversions, deletions, and repeat expansions also will contribute to a re-organization of the linear structure of our genome. What impact these re-organizations have on higher-order chromatin structure is unclear. For instance, if a topological domain is broken during a syntenic break in evolution, what is the resulting structure? Generation of chromatin interaction maps coupled with comparative genomics of related species has the potential to reveal the impact of variation in genome structure

on higher-order chromatin structure and vice versa. Studies have already revealed that expansion of SINE elements appears to have been responsible for the generation of new CTCF binding sites during evolution (9). This is likely just scratching the surface of the alterations in genomic structure impacting alterations in higher-order chromatin structure.

Changes in the structure of the genome does not only happen during evolution. Alterations in the form of translocations, inversions, and copy number duplications occur quite frequently in cancer cells. Notably, recent reports have suggests that chromosomes in cancer cells can undergo large-scale rearrangements in structure termed "chromothripsis" (10). The impact of these re-organizations on higher order chromatin structure is not clear, but it has the potential to contribute to vastly different regulatory landscapes in cancer genomes. For instance, translocations have long been known to create gene fusions that can drive oncogenesis. I speculate that alterations of genome structure in cancer also have the potential to create "domain fusions" that may create new juxtapositions of regulatory elements and target genes. The degree to which such alterations may impact cancer biology and oncogenesis is an area of great promise in the future.

In summary, I believe that the work presented here has made a worthwhile contribution to our understanding of higher-order chromatin structure in mammalian genomes. I hope that in the future, these results may have an impact on diverse areas of biology, including gene expression, evolutionary genetics, and cancer biology. As will likely always be the case in science, it appears as though the combination of new technologies applied to long-standing problems will make for exciting discoveries in the coming years.

References

1. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature.advance online publication.

2. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, et al. A map of the cis-regulatory sequences in the mouse genome. Nature.488(7409):116-20.

3. Merkenschlager M, Odom DT. CTCF and Cohesin: Linking Gene Regulatory Elements with Their Targets. Cell.152(6):1285-97.

4. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods.58(3):268-76.

5. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol.30(1):90-8.

6. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell.148(3):458-72.

7. Wong H, Marie-Nelly H, Herbert S, Carrivain P, Blanc H, Koszul R, et al. A predictive computational model of the dynamic 3D interphase yeast nucleus. Curr Biol.22(20):1881-90.

8. Heger P, Marin B, Schierenberg E. Loss of the insulator protein CTCF during nematode evolution. BMC Mol Biol. 2009;10:84. PMCID: 2749850.

9. Schmidt D, Schwalie PC, Wilson MD, Ballester B, GonÁalves n, Kutter C, et al. Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. Cell.

10. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell.144(1):27-40. PMCID: 3065307.