**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

**Michael H. Nichols** **Date**

**Fundamental Principles of 3D Genomic Organization**

By

Michael H. Nichols
Doctor of Philosophy

Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

---

**Victor Corces, Ph.D.**
Advisor

---

**Jeremy Boss, Ph.D.**
Committee Member

---

**Roger Deal, Ph.D.**
Committee Member

---

**David Katz, Ph.D.**
Committee Member

---

**William Kelly, Ph.D.**
Committee Member

Accepted:

---

**Lisa A. Tedesco, Ph.D.**
Dean of the James T. Laney School of Graduate Studies

---

Date

# Fundamental Principles of 3D Genomic Organization

**By**

**Michael H. Nichols**

**B.S., Emory University, 2013**

**Advisor: Victor Corces, PhD**

An abstract of
A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in
Graduate Division of Biological and Biomedical Science
Genetics & Molecular Biology
2020

## Abstract

Fundamental Principles of 3D Genomic Organization

By Michael Holden Nichols

The three-dimensional (3D) conformation of chromatin in the nucleus is an elusive but essential aspect of genomic regulation. Only with recent advances in techniques such as Hi-C has it been possible to assess this structure at the sequence level across the entire genome. A variety of architectural patterns are observed in these conformational assays. Here we present insights into the fundamental principles that give rise to these structural phenomena. Two independent processes can explain the majority of the conformational features of the genome. Extrusion of DNA loops by Structural Maintenance of Chromosomes (SMC) complexes form stable CTCF loops and associated topological domains. We present the theoretical logic of this model and a mechanistic explanation for how SMC complexes may extrude loops. Separately, chromatin segments associate preferentially with other regions with similar chromatin features. We show this agglomeration directly corresponds to various epigenetic features and extends beyond canonical binary segregation of transcriptionally active and inactive chromatin. Together these processes organize the genome, playing essential roles in transcriptional regulation and likely other aspects of genome function.

# Fundamental Principles of 3D Genomic Organization

**By**

**Michael H. Nichols**

**B.S, Emory University, 2013**

**Advisor: Victor Corces, PhD**

A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in
Graduate Division of Biological and Biomedical Science
Genetics & Molecular Biology
2020

**Acknowledgements**

This work would not have been possible without the efforts of many members of the Corces Lab past and present. Foremost, Victor Corces who guided my work with great patience and insight and was always willing to indulge my ideas. I also give special thanks my co-authors Kevin Van Bortle and Jordan Rowley whose mentorship and friendship have been essential to my success and whose contributions were critical to the publications contained in this work. For their training and teaching, their discussions and scientific insights, and their friendship I thank Chintong Ong, Naomi Takenaka, Chenhaun Xu, Li Li, Yoon-hee Jung, Axel Poulet, and Bri Bixler. I thank my thesis committee, Jeremy Boss, Roger Deal, David Katz, WIlliam Kelly, and former member Paula Vertino for their encouragement, feedback, and challenges.

I am thankful for the endless support of my family and friends who have always been there to listen and help, and for the unconditional love from Cristina, who has helped me every step of the way to make my graduate years enjoyable.

# Table of Contents

## Tables



## Figures

# Chapter 1: Introduction

## The chromosome as a folded polymer

The three-dimensional folding of polymers is one of the fundamental bases of all known biology. The folding of amino acid chains produces proteins responsible for nearly every biological process. Likewise, the ribonucleic acid (RNA) polymers that encode these proteins also fold to regulate their own function and even to produce enzymatic activity of their own. With recent advances in technique and theory, it is now becoming clear that the folding of deoxyribonucleic acid (DNA) also plays an essential role in its function. Investigating and understanding this organization is a uniquely difficult problem due to the DNA polymer's large size, the highly stochastic and labile nature of this organization, and the myriad of proteins and RNAs that interact with the DNA to regulate folding.

Chromosomes are massive molecules composed of hundreds of millions of DNA base pairs chained together to form a polymer with a contour length measured in centimeters. These polymers are highly flexible with a persistence length of only about 100 base pairs (Gross et al. 2011). Together these features make chromosomes the most structurally complex molecules in biology with many possible folded configurations. Additionally, chromosomes do not each fold only among themselves but intermingle in the nucleus. Human nuclei have diameters on the order of 10 micrometers, meaning the dozens of centimeter-long chromosomes of a diploid nucleus are necessarily extensively folded (Sun, Shen, and Yokota 2000). The sheer enormity of DNA folding poses unique methodological obstacles to its understanding.

Fluorescence in-situ hybridization (FISH) allows for the visualization of the nuclear localization of regions of the genome according to sequence. Whole chromosome FISH has revealed that the organization at the chromosome scale is highly stochastic (Meaburn and Misteli 2007). A given chromosome tends to occupy a given region of the nucleus termed a chromosome territory, but this position varies from cell to cell. These territories are not sharply delineated, with one chromosome's territory blending into those of its neighbors (Meaburn and Misteli 2007). While some non-random features exist at this level of organization, such as a chromosome's distance from the nuclear periphery, these correlations tend to be weak (Meaburn and Misteli 2007).

These findings evince another major obstacle in understanding DNA polymer folding. DNA polymers lack any specific configuration favorable enough to overcome entropy. This intrinsic disorder causes stochastic variation to dominate large-scale genomic organization. This phenomenon is also found in some protein and RNA structures that lack a single stable conformation. As a result, the organization of the genome is highly dynamic from cell to cell and even moment to moment. However, entropy is not the only force at work on the DNA polymer. The DNA is organized by several distinct nuclear phenomena. This makes describing the folding of a DNA polymer more complex than identifying a few optimal conformational states. Rather, DNA folding can only be understood as an energy landscape where polymers in constant thermodynamic flux shift between local minima. This probabilistic nature confounds traditional models of polymer folding.

The dynamic forces driving this organization have only recently begun to be identified and represent another layer of complexity in understanding the folding of DNA polymers. Unlike proteins and single-stranded RNA, double-stranded DNA polymers do not strongly self-interact by themselves. The organization present in DNA is thus not an intrinsic property of the DNA

nucleotide sequence but rather is mediated by DNA binding proteins and their interactions with each other and other pieces of DNA. In addition, many RNAs associate with chromatin and likely play organizational roles as well. The DNA nucleotide sequence still governs organization, but only indirectly by the nature of the proteins with which each sequence interacts. This complicates the process of understanding genomic organization as the nucleotide sequence alone is not sufficient to explain the folding of the DNA. This is not a concept unknown in protein or RNA folding where chaperones play a large role in directing folding, but these ancillary molecules represent the primary forces at work in DNA folding. These “epigenetic” features are dynamic and change between cell types, respond to environmental conditions and change over the course of the cell cycle. The genomic organization is therefore equally dynamic. Understanding the organizational patterns at work in the nucleus thus requires an understanding of the multitude of DNA-interacting proteins that constitute the epigenome.

These features of chromosome folding, its size, its stochasticity, and its protein-driven nature make understanding this incredibly complex phenomenon profoundly difficult. However, just as the structure of a protein or RNA polymer determines its function, there is mounting evidence that the structure of DNA polymers and the organization of the nucleus are key to transcriptional regulation and other nuclear functions. With this driving motivation in recent years, we have begun to elucidate the fundamental forces governing genomic organization.

## Features of genomic architecture

Chromosomes fold across such large length scales that distinct forces play roles in different size regimes. DNA is most directly folded by histone octamers that wrap roughly 146 base pairs around them in approximately 1.5 turns (Andrews and Luger 2011). While this immediate

organization is rather straightforward these integrated proteins play a large role in functional regulation and organization at larger scales. The positioning and spacing of these histones alone are significant features of chromatin organization and regulate the binding of other proteins to the DNA (Struhl and Segal 2013). These histones also function as a canvas for a palette of post-translational modifications (PTMs) that correlate with and are thought to play important roles in numerous genomic processes. This histone code represents the most foundational level of chromosome organization upon which higher levels are built. In addition to histones, the chromatin is festooned with thousands of different proteins. Most of these DNA binding proteins are thought to function as transcription factors that affect transcriptional activity (Spitz and Furlong 2012). While some of these transcription factors bend the DNA with their binding their effects on DNA-polymer folding are poorly understood with only several exceptions. Models of transcription factor function generally require them to physically interact with the transcription machinery at the site of transcription activation, the promoter. However, many transcription factors bind DNA sequences called enhancers, which can be located very far away along the length of the chromosome. It follows that transcription factors rely on the folding of the DNA into a loop between the enhancer and promoter in order to play their proximal role. To what extent transcription factors drive the formation of these loops is not well understood. This enhancer-promoter looping is a key mechanism by which genomic folding is thought to regulate genomic function (Nolis et al. 2009). This basic understanding of DNA folding comes from the direct implications of the proteins that bind to it, but to understand folding at a larger scale requires different methods.

The first glimpses into large-scale nuclear organization came from microscopic analysis of the nucleus using various staining and fluorescence techniques. These analyses revealed that the nucleoplasm was not uniform but rather contained numerous examples of discrete regions with different contents (Zimber, Nguyen, and Gespach 2004). These nuclear bodies lack membranes

and so represent agglomerations of specific subsets of nucleoplasm components due to mutual attractions. Subsequent studies have shown that many of these nuclear bodies are responsible for specific functions such as mRNA splicing (Galganski, Urbanek, and Krzyzosiak 2017). This has led to a general model in which various nuclear functions occur in spatially segregated regions of the nucleus. It is thought that this organization improves the efficiency of these processes by concentrating the necessary components with their targets, but also provides a means of regulating function by the inclusion and exclusion of targets. Some of these membraneless nuclear bodies contain chromatin such as the nucleolus where ribosomal RNA genes congregate along with RNA Polymerase 1 to transcribe ribosomal RNAs (Iarovaia et al. 2019). This has led to the natural hypothesis that functionally related regions of the genome may colocalize together with their intended machinery. An example is transcription, which in some cases is thought to occur in “transcription factories” where genes colocalize with RNA polymerases based on the visualization of focal sites of transcription (Jackson et al. 1993). This concept of functional regulation through spatial segregation into biomolecular condensates appears to be a fundamental feature of genomic organization.

Our understanding of genomic organization at the level of specific genomic loci has largely been informed by a technique known as chromatin conformation capture. This simple method measures the frequency that genomic loci are near each other by chopping the chromosomes into small pieces and then rejoining them together into novel chimeric sequences (Dekker et al. 2002). Because DNA is extensively folded, the rejoining process does not repair the chromosome into its original sequence but rather into a novel order based on the spatial proximity of each piece in the nucleus. In this way, chimeric pieces of DNA composed of sequences potentially far apart along a chromosome or even from separate chromosomes are generated and the frequency of any given chimeric sequence is a function of the frequency with which those two loci were spatially proximal in the cell population. By quantifying the abundance

of a specific chimeric sequence, we can determine how frequently a given pair of genomic loci interact. By sequencing all of the resulting chimeric sequences we can see the interaction frequencies of the entire genome at the sequence level. This latter technique is known as Hi-C and produces a two-dimensional matrix composed of counts of the chimeric sequences originating from each pair of genomic loci (Lieberman-Aiden et al. 2009). As genomic organization is highly stochastic these counts represent the relative frequency these genomic regions are in proximity with one another.

The strongest feature of these Hi-C maps is an expected characteristic of a predominantly disordered polymer. Interaction frequency decays with genomic distance. Loci close together on a polymer interact more frequently than loci farther away according to a power law decay (Mirny 2011). This means sites within ten thousand base pairs of each other interact orders of magnitude more frequently than sites hundreds of thousands of base pairs apart. Several other significant features are found in these Hi-C maps but none compare to distance decay, indicating that even the most enriched interactions are still stochastic and only occur in a fraction of cells or a fraction of the time.

The most significant deviation from the uniform distance decay is a division of the chromosome into types that interact more frequently with other chromatin of the same type. Alternating genomic regions of each type along the chromosome result in a checkerboard-like pattern of increased and decreased interactions in the Hi-C map (Figure 1.1, top). Examining the chromatin features associated with each type revealed that transcriptional activity was largely associated with only one type making the other type transcriptionally inactive. This analysis was first performed in the lymphoblastoid cell line GM12878 and the two types of chromatin were called A and B, for transcriptionally active and inactive, respectively (Figure 1.1, red and blue) (Lieberman-Aiden et al. 2009). This basic pattern is found throughout human cell types and in

every Eukaryote for which Hi-C maps exist (Rowley et al. 2017). One explanation for this pattern of interaction biases among chromatin types is that they are spatially segregated into discrete compartments in the genome. The physical characteristics of these compartments such as their size, their count, and the precision of their segregation, are still unclear. These biomeolcular condensates likely exist along a gradient from large, discrete, and stable formations, such as the larger nuclear bodies to small, fuzzy, and transient formations that would more closely match the appearance of punctate transcription factories but may be even smaller. Where along this gradient each compartment type falls is likely a function of the strength with which the proteins and RNA of each chromatin type self-attract. Stronger attractive forces would more effectively overcome entropy to produce more stable agglomerations while weaker forces would more easily fall apart.

Initial studies of this phenomenon necessarily used a large resolution of 1 megabase leaving their precise relationship with transcription and other chromatin features unclear as these features vary on the scale of kilobases (Lieberman-Aiden et al. 2009). In *Drosophila*, where due to its smaller genome size higher resolution Hi-C maps can be produced, the compartment types directly correspond to transcription such that genomic organization can be reproduced using only measures of transcription or its associated histone modifications (Rowley et al. 2017). It is tempting to extend these results to humans, however, the *Drosophila* genome is extremely gene dense and is almost entirely occupied by histone modification associated with transcriptional activity (H3K27ac) or inactivity (H3K27me3) (Rowley et al. 2017). The human genome, on the other hand, is far more gene sparse, with large inactive stretches of the chromosome that are often not strongly marked by chromatin features associated with transcriptional inactivity. The chromatin features associated with each compartment differ between studies of different cell types, suggesting the two-compartment model is likely an oversimplification. Closer and higher resolution examinations have refined the concept to show

that compartmental interaction patterns are better described by more than just two types of chromatin (Rao et al. 2014).

Another feature of Hi-C maps are contiguous regions of the chromosome that preferentially self-interact called topologically-associating domains (TADs). Naturally, compartmentalization creates TADs as each contiguous region of the same chromatin type will interact more frequently with itself than with neighboring regions of different types. These are called compartmental domains and are responsible for TADs found in *Drosophila* (Rowley et al. 2017). However, compartmental domains are not sufficient to explain all TADs seen in human nuclei, although they are responsible for many. A separate phenomenon is at work in our nuclei. Conspicuously strong interactions are found between the binding sites of the DNA-binding protein CTCF (Rao et al. 2014). This protein's homolog in *Drosophila*, dCTCF, had long been known as an insulator reported to inhibit enhancer-promoter interactions when placed between the two elements (Gerasimova et al. 2007). These strong interactions or loops between CTCF sites are, however, not found in *Drosophila* (Rowley et al. 2017). It is tempting to suppose these loops are formed by stable homodimerization between CTCF proteins. A model by which stochastic motion brings two CTCF sites into contact and they bind tightly to each other could create loops, however this model is actually incompatible with a remarkably curious feature of CTCF looping. CTCF's binding motif is non-palindromic, meaning each site is oriented along the chromosome. This orientation plays an immense and unexpected role in loop formation as the vast majority of CTCF loops form between CTCF sites in a convergent orientation with respect to each other (Figure 1.1, grey arrows). In contrast, looping interactions only rarely form between CTCF sites in a divergent orientation. This orientation bias has strong implications for the mechanism by which CTCF loops form, as any mechanism relying on stochastic motion cannot explain it. Instead, this observation has led to the proposal of a loop extrusion mechanism for CTCF loop formation (Nichols and Corces 2015). In this model CTCF sites form

loops due to the action of a loop extruding complexes for which CTCF sites act as oriented borders.

The loop extrusion model posits the existence of a DNA motor that extrudes loops between CTCF sites. However, at the time, no existing DNA motor fit the required characteristics nor associated strongly with CTCF sites. Attention turned to the Structural Maintenance of Chromosomes (SMC) family of protein complexes. These ring-shaped complexes were capable of hydrolyzing ATP and one member, cohesin, strongly colocalized with CTCF binding sites in the genome (Figure 1.1, green rings). Another member, condensin, was known to play a key role in mitotic condensation of the genome, which had independently been theorized to also utilize a loop extrusion mechanism (Alipour and Marko 2012). This strong circumstantial evidence led to tests of the putative roles of cohesin and condensin in loop extrusion processes and to in vitro experiments testing their ability to act as DNA motors and extrude loops. Both protein complexes are now known to be indispensable for their respective looping functions and to extrude DNA loops in vitro confirming a key prediction of the extrusion model of CTCF loop formation (Rao et al. 2017; Green et al. 2012; Ganji et al. 2018; Davidson et al. 2019). While CTCF oriented loops are absent in *Drosophila*, loop extrusion is likely a conserved function of the SMC complexes, as SMC homologs extrude circular chromosomes even in prokaryotes (Wang et al. 2015). Moreover, cohesin strongly affects genomic organization even in the absence of CTCF sites by its constant extrusion action, which condenses the DNA and increases the frequency of short range interactions (Fudenberg et al. 2017). The mechanism by which these complexes move along the DNA to extrude loops is still a matter of debate without a fully satisfactory model.

By constraining the loops extruded by cohesin, CTCF sites act as unidirectional barriers and change the topology of the chromatin. The chromatin between two convergently oriented CTCF

sites self-interacts more than average. This is another mechanism by which TADs are formed and are known as contact domains. Contact domains form by an independent process but their borders frequently correlate with compartmental domains suggesting a functional role for these borders. These domains may play an important role in regulating distal enhancer-promoter interactions. By increasing interaction frequencies within domains, the frequency of enhancer-promoter loops would be increased and, correspondingly, enhancer-promoter loops between domains would be diminished (Hnisz, Day, and Young 2016). Loop extrusion may even be playing a direct role in enhancer-promoter looping since, as the cohesin complex tracks over the domain, it will bring distal regions directly together (Schoenfelder and Fraser 2019). The preferential localization of cohesin to promoter regions further hints towards this possibility.

While we are only beginning to understand the 3D architecture of the genome, we have already identified at least two key independent processes. Both these processes, compartmentalization and loop extrusion, influence the interaction frequencies of genomic loci with each other and so likely play a role in enhancer-promoter interactions. However, as assessed by Hi-C, even strong interaction biases created by these processes still fall far short of ensuring or prohibiting any given interaction. Thus, it seems unlikely that a modest increase or decrease in enhancer-promoter contact frequencies is the sole function of compartmental or contact domains. We understand only the most basic consequences of these organizational principles and lack a detailed mechanistic understanding of how compartments segregate or how loops are extruded.

# Figures

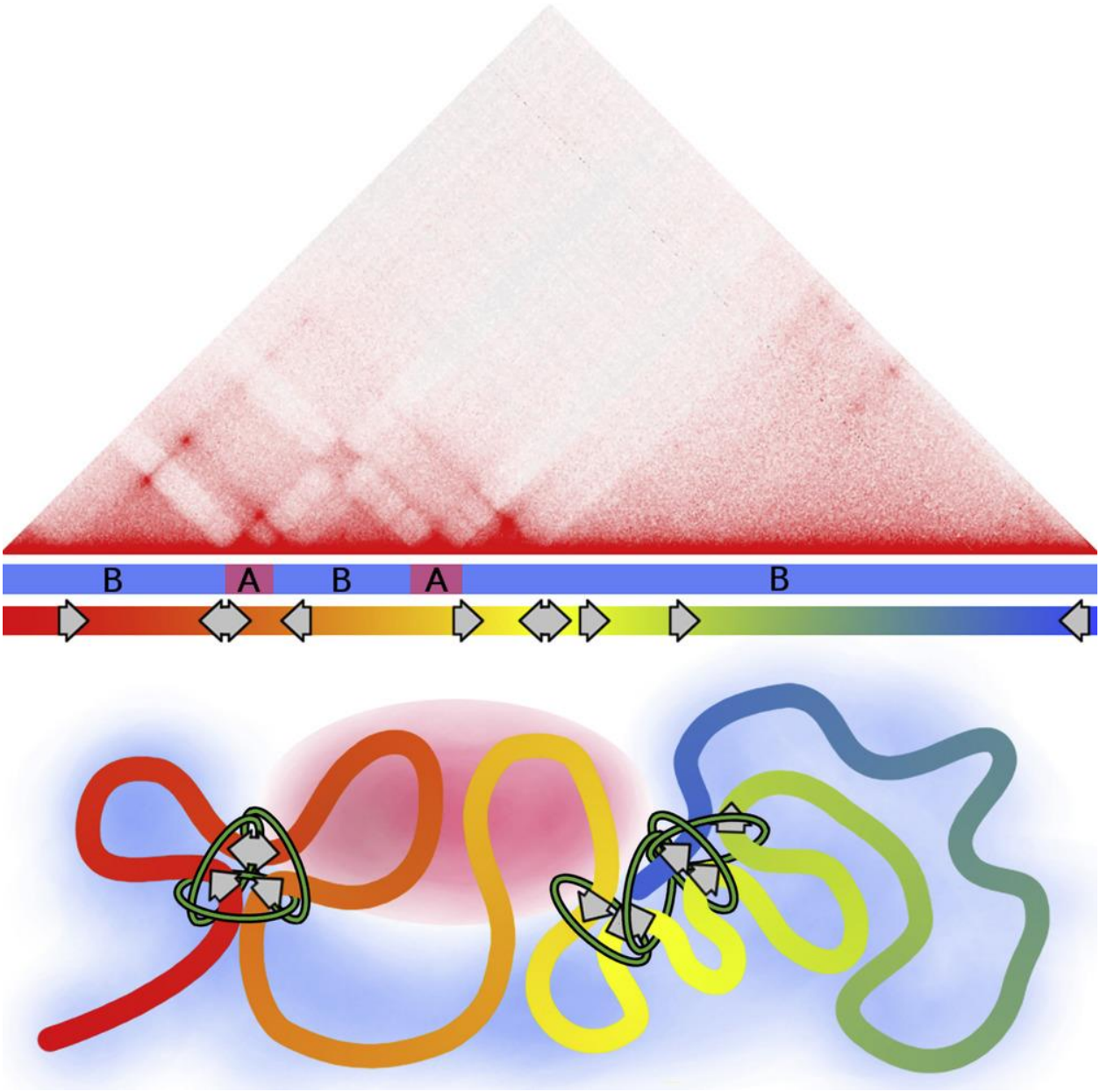



Figure 1.1. Compartments and CTCF loops organize the genome

Human Hi-C map (top) is organized by A (red) and B (blue) compartments (second from top) and oriented CTCF sites (grey arrows) which form loops with cohesin rings (green rings). Together these processes fold the genome (rainbow line).

# References

Alipour, Elnaz, and John F. Marko. 2012. “Self-Organization of Domain Structures by DNA-Loop-Extruding Enzymes.” *Nucleic Acids Research* 40 (22): 11202–12. https://doi.org/10.1093/nar/gks925.

Andrews, Andrew J., and Karolin Luger. 2011. “Nucleosome Structure(s) and Stability: Variations on a Theme.” *Annual Review of Biophysics* 40 (1): 99–117. https://doi.org/10.1146/annurev-biophys-042910-155329.

Davidson, Iain F., Benedikt Bauer, Daniela Goetz, Wen Tang, Gordana Wutz, and Jan-Michael Peters. 2019. “DNA Loop Extrusion by Human Cohesin.” *Science* 366 (6471): 1338–45. https://doi.org/10.1126/science.aaz3418.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. “Capturing Chromosome Conformation.” *Science (New York, N.Y.)* 295 (5558): 1306–11. https://doi.org/10.1126/science.1067799.

Fudenberg, Geoffrey, Nezar Abdennur, Maxim Imakaev, Anton Goloborodko, and Leonid A. Mirny. 2017. “Emerging Evidence of Chromosome Folding by Loop Extrusion.” *Cold Spring Harbor Symposia on Quantitative Biology* 82: 45–55. https://doi.org/10.1101/sqb.2017.82.034710.

Galganski, Lukasz, Martyna O. Urbanek, and Wlodzimierz J. Krzyzosiak. 2017. “Nuclear Speckles: Molecular Organization, Biological Function and Role in Disease.” *Nucleic Acids Research* 45 (18): 10350–68. https://doi.org/10.1093/nar/gkx759.

Ganji, Mahipal, Indra A. Shaltiel, Shveta Bisht, Eugene Kim, Ana Kalichava, Christian H. Haering, and Cees Dekker. 2018. “Real-Time Imaging of DNA Loop Extrusion by Condensin.” *Science* 360 (6384): 102–5. https://doi.org/10.1126/science.aar7831.

Gerasimova, Tatiana I., Elissa P. Lei, Ashley M. Bushey, and Victor G. Corces. 2007. “Coordinated Control of DCTCF and Gypsy Chromatin Insulators in Drosophila.” *Molecular Cell* 28 (5): 761–72. https://doi.org/10.1016/j.molcel.2007.09.024.

Green, Lydia C., Paul Kalitsis, Tsz M. Chang, Miri Cipetic, Ji Hun Kim, Owen Marshall, Lynne Turnbull, et al. 2012. “Contrasting Roles of Condensin I and Condensin II in Mitotic Chromosome Formation.” *Journal of Cell Science* 125 (6): 1591–1604. https://doi.org/10.1242/jcs.097790.

Gross, Peter, Niels Laurens, Lene B. Oddershede, Ulrich Bockelmann, Erwin J. G. Peterman, and Gijs J. L. Wuite. 2011. “Quantifying How DNA Stretches, Melts and Changes Twist under Tension.” *Nature Physics* 7 (9): 731–36. https://doi.org/10.1038/nphys2002.

Hnisz, Denes, Daniel S. Day, and Richard A. Young. 2016. “Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control.” *Cell* 167 (5): 1188–1200. https://doi.org/10.1016/j.cell.2016.10.024.

Iarovaia, Olga V., Elizaveta P. Minina, Eugene V. Sheval, Daria Onichtchouk, Svetlana Dokudovskaya, Sergey V. Razin, and Yegor S. Vassetzky. 2019. “Nucleolus: A Central Hub for Nuclear Functions.” *Trends in Cell Biology* 29 (8): 647–59. https://doi.org/10.1016/j.tcb.2019.04.003.

Jackson, D A, A B Hassan, R J Errington, and P R Cook. 1993. “Visualization of Focal Sites of Transcription within Human Nuclei.” *The EMBO Journal* 12 (3): 1059–65.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

Meaburn, Karen J., and Tom Misteli. 2007. “Cell Biology: Chromosome Territories.” *Nature* 445 (7126): 379–781. https://doi.org/10.1038/445379a.

Mirny, Leonid A. 2011. “The Fractal Globule as a Model of Chromatin Architecture in the Cell.” *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 19 (1): 37–51. https://doi.org/10.1007/s10577-010-9177-0.

Nichols, Michael H., and Victor G. Corces. 2015. “A CTCF Code for 3D Genome Architecture.” *Cell* 162 (4): 703–5. https://doi.org/10.1016/j.cell.2015.07.053.

Nolis, Ilias K., Daniel J. McKay, Eva Mantouvalou, Stavros Lomvardas, Menie Merika, and Dimitris Thanos. 2009. “Transcription Factors Mediate Long-Range Enhancer–Promoter Interactions.” *Proceedings of the National Academy of Sciences* 106 (48): 20222–27. https://doi.org/10.1073/pnas.0902454106.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7): 1665–80. https://doi.org/10.1016/j.cell.2014.11.021.

Rao, Suhas S.P., Su-Chen Huang, Brian Glenn St Hilaire, Jesse M. Engreitz, Elizabeth M. Perez, Kyong-Rim Kieffer-Kwon, Adrian L. Sanborn, et al. 2017. “Cohesin Loss Eliminates All Loop Domains.” *Cell* 171 (2): 305-320.e24. https://doi.org/10.1016/j.cell.2017.09.026.

Rowley, M. Jordan, Michael H. Nichols, Xiaowen Lyu, Masami Ando-Kuri, I. Sarahi M. Rivera, Karen Hermetz, Ping Wang, Yijun Ruan, and Victor G. Corces. 2017. “Evolutionarily Conserved Principles Predict 3D Chromatin Organization.” *Molecular Cell* 67 (5): 837-852.e7. https://doi.org/10.1016/j.molcel.2017.07.022.

Schoenfelder, Stefan, and Peter Fraser. 2019. “Long-Range Enhancer–Promoter Contacts in Gene Expression Control.” *Nature Reviews Genetics* 20 (8): 437–55. https://doi.org/10.1038/s41576-019-0128-0.

Spitz, François, and Eileen E. M. Furlong. 2012. “Transcription Factors: From Enhancer Binding to Developmental Control.” *Nature Reviews Genetics* 13 (9): 613–26. https://doi.org/10.1038/nrg3207.

Struhl, Kevin, and Eran Segal. 2013. “Determinants of Nucleosome Positioning.” *Nature Structural & Molecular Biology* 20 (3): 267–73. https://doi.org/10.1038/nsmb.2506.

Sun, Hui Bin, Jin Shen, and Hiroki Yokota. 2000. “Size-Dependent Positioning of Human Chromosomes in Interphase Nuclei.” *Biophysical Journal* 79 (1): 184–90. https://doi.org/10.1016/S0006-3495(00)76282-5.

Wang, Xindan, Tung B.K. Le, Bryan R. Lajoie, Job Dekker, Michael T. Laub, and David Z. Rudner. 2015. “Condensin Promotes the Juxtaposition of DNA Flanking Its Loading Site in Bacillus Subtilis.” *Genes & Development* 29 (15): 1661–75. https://doi.org/10.1101/gad.265876.115.

Zimber, Amazia, Quang-Dé Nguyen, and Christian Gespach. 2004. “Nuclear Bodies and Compartments: Functional Roles and Cellular Signalling in Health and Disease.” *Cellular Signalling* 16 (10): 1085–1104. https://doi.org/10.1016/j.cellsig.2004.03.020.

# Chapter 2: A CTCF code for 3D genome architecture

Michael H. Nichols and Victor G. Corces

Previously published in: Cell. 2015 Aug 13;162(4):703-5. doi:10.1016/j.cell.2015.07.053.

## Summary

The architectural protein CTCF plays a complex role in decoding the functional output of the genome. In this issue of Cell Guo et al. show that the orientation of a CTCF site restricts its choice of interacting partner, thus creating a code that predicts the three-dimensional organization of the genome.

## Main Text

CTCF is a DNA-binding protein known to play a variety of roles in the regulation of transcription by forming loops in which distant elements of the genome are brought into spatial proximity within the nucleus (Ong and Corces, 2014). The formation of these loops is believed to involve homodimerization of the CTCF protein bound to their bases. By mediating contacts between distant sequences, CTCF regulates enhancer-promoter interactions throughout the genome and appears to play a key role in the formation of Topologically Associating Domains (TADs) (Nora et al., 2012). Analysis of genome-wide interaction data obtained by Hi-C suggests that CTCF-mediated contacts occur much more frequently when the binding sites for this protein are present in the forward and reverse orientations (Rao et al., 2014). Interactions between binding sites arranged in the same orientation still occur, although less frequently, and interactions between CTCF sites in a divergent orientation rarely take place. In this issue, Guo et al. (Guo et

al., 2015) carry out a detailed functional analysis of the role of CTCF binding site orientation in the regulation of enhancer-promoter choice underlying stochastic expression of specific protocadherin isoforms.

The mouse protocadherin genes are arranged in three different clusters, named *Pcdha*, *b* and *g*, located in two different sub-TADs. Each Pcdh protein isoform is encoded by an RNA arising from alternative splicing between a series of alternative constant and variable exons. Each variable exon contains an upstream promoter, and transcription from a specific promoter requires interaction with downstream enhancers via DNA looping. Each variable exon and enhancers are associated with specific CTCF sites. Guo et al noticed that the CTCF binding sites that form loops between promoters and enhancers are arranged in a convergent orientation. Using the CRISPR-Cas9 genome editing system they created inversions of key CTCF binding sites, switching their orientation. The authors then use 4C to show that the inverted CTCF binding sites now have an inverted interaction bias. This confirms the causal relationship between DNA binding site orientation and the direction of looping. Furthermore, the change in looping directionality is accompanied by changes in transcription, indicating a functional role for the CTCF mediated interactions in regulating gene expression.

The authors then expand their investigation to the entire genome using published CTCF ChIA-PET data. They find the same orientation bias in interactions between CTCF sites as previously shown seen using Hi-C. The authors use these data to show that TAD boundaries are enriched in CTCF sites arranged in divergent orientations. This means that CTCF sites at the borders of TADs will tend to loop towards the interior of each TAD. While it was known that CTCF binding sites were enriched in TAD boundaries in specific orientations (Vietri Rudan et al., 2015), this finding helps explain why only a subset of CTCF sites in the genome are able to form these boundaries, and reinforces the functional relevance of these sites to the formation of TADs.

These observations solidify what now appears to be one of the underlying principles by which the orientation of the DNA sequence in CTCF binding sites shapes 3D genome organization. However, this new finding raises a series of questions as to the mechanisms underlying the specificity of interactions between CTCF sites in the genome. CTCF binding sites in divergent and convergent orientations are molecularly identical and impossible to distinguish outside of the larger context of the DNA molecule. Figure 2.1A shows two theoretical CTCF mediated loops. The only difference between the two loops is which side of the CTCF sites the looped-out DNA is on. Despite this, the loop depicted on top occurs much more frequently than the loop depicted below. This means that the mechanism by which CTCF forms loops must be aware of this context and be capable of discriminating between CTCF sites in convergent and divergent orientations. A simplistic model of loop formation that relies on random collisions in the nuclear space between CTCF bound to DNA in different orientations to form interactions (Rao et al., 2014) is incompatible with the observations, as it could not be aware of the relative positions or orientations of the CTCF binding sites.

One potential explanation for the directionality in loop formation is that the bias is created by the binding of CTCF to its recognition site, which causes a ninety degree bending in the DNA, resulting in the formation of an unusual structure that could be interpreted as a loop (MacPherson and Sadowski, 2010). Several potential processes could then contribute to the expansion of the initial loop (Figure 2.1B). rather than in the interacting proteins. If CTCF binding sites have an intrinsic bias for interacting with the DNA on one side of the binding site more frequently than the other side it would explain the bias seen in loop formation. Unpublished analysis of cohesin ChIA-PET data from our lab supports this hypothesis by showing that CTCF sites show an orientation bias in all cohesin mediated loops and not just CTCF-CTCF interactions. The directional bias of CTCF sites can be explained by results from EMSA assays, which show that CTCF can bend DNA to form a DNA structure hypothesized to

be a loop on one side of the CTCF binding site (Macpherson and Sadowski, 2010). As this DNA structure is formed in the same orientation as the bias in looping it seems likely that the two phenomena are causally linked. Figure 2.1B depicts model in which a loop is initiated by CTCF binding and recruitment of cohesin and expanded in size by pulling DNA into the loop. Since the CTCF end of the loop is anchored in place the loop expands unidirectionally to bring DNA in the direction of the loop into proximity with the CTCF site. CTCF is known to interact with the cohesin complex, which has two ATPase domains whose function is currently unclear but could potentially be involved in extruding the loop (Alipour and Marko, 2012; Strick et al., 2004). Transcriptional activity could also contribute to the movement of the DNA through the loop if the transcription complex bound to cohesin at the CTCF site is unable to move through and remains stationary (Lengronne et al., 2004). The same result could be attained if CTCF is bound to mRNAs as they are transcribed (Kung et al., 2015). The observed frequency of interactions between CTCF sites with the same orientation is relatively low (Guo et al., 2015), perhaps due to the exact positioning of the CTCF proteins as they pass through the loop and collide with each other or cohesin in anti-parallel orientations, which may not favor homodimerization. The directionality imposed by this loop extrusion model would result in a CTCF site interacting more frequently with the DNA on one side of it, explaining why divergent CTCF sites interact very infrequently (Guo et al., 2015; Rao et al., 2014). Finally, the requirement for the extrusion of the looped DNA increases the likelihood of collisions between CTCF sites and Mediator complexes or gene promoters (Figure 2.1B), imposing a directionality on these interactions.

# Figures

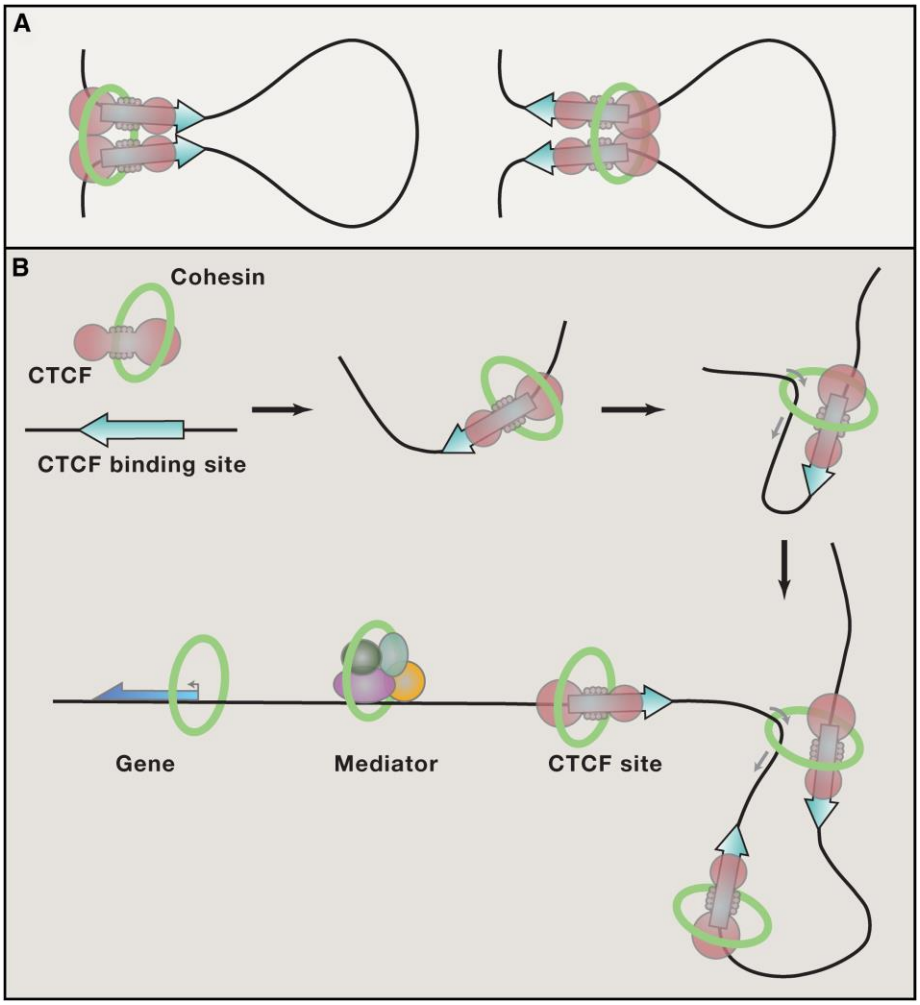



Figure 2.1. Model of Orientation Biased CTCF Looping

(A) CTCF mediated loops in convergent and divergent orientations only differ in how they are connected by the DNA. The top loop occurs much more frequently than the bottom loop suggesting the mechanism of loop formation must be able to distinguish the two cases.

(B) A loop-extrusion model would explain the orientation bias seen in CTCF mediated looping. CTCF bends DNA and could be capable of forming a loop on one side of its binding site only due to the manner in which the DNA is bent. This loop could then be expanded in one direction causing the CTCF site to contact other DNA elements such as other CTCF sites, Mediator complexes, and gene promoters more frequently in one orientation.

# References

Alipour, E., and Marko, J.F. (2012). Self-organization of domain structures by DNA-loop-extruding enzymes. Nucleic acids research *40*, 11202-11212.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y.*, et al.* (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell.

Kung, J.T., Kesner, B., An, J.Y., Ahn, J.Y., Cifuentes-Rojas, C., Colognori, D., Jeon, Y., Szanto, A., del Rosario, B.C., Pinter, S.F.*, et al.* (2015). Locus-specific targeting to the X chromosome revealed by the RNA interactome of CTCF. Molecular cell *57*, 361-375.

Lengronne, A., Katou, Y., Mori, S., Yokobayashi, S., Kelly, G.P., Itoh, T., Watanabe, Y., Shirahige, K., and Uhlmann, F. (2004). Cohesin relocation from sites of chromosomal loading to places of convergent transcription. Nature *430*, 573-578.

MacPherson, M.J., and Sadowski, P.D. (2010). The CTCF insulator protein forms an unusual DNA structure. BMC molecular biology *11*, 101.

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

Ong, C.T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nature reviews Genetics *15*, 234-246.

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

Strick, T.R., Kawaguchi, T., and Hirano, T. (2004). Real-time detection of single-molecule DNA compaction by condensin I. Current biology : CB *14*, 874-880.

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

# Chapter 3: A tethered inchworm model of SMC DNA translocation

Michael H. Nichols and Victor G. Corces

Previously published in: Nat Struct Mol Biol. 2018 Oct;25(10):906-910. doi: 10.1038/s41594-018-0135-4. Epub 2018 Sep 24.

## Abstract

The DNA loop extrusion model is a provocative new concept explaining the formation of chromatin loops, which revolutionizes our understanding of genome organization. Central to this model is the Structural Maintenance of Chromosomes (SMC) protein family that is now being ascribed a new function as a DNA motor. In this Perspective we review and reinterpret the current knowledge of SMC structure and function and propose a novel mechanism for SMC motor activity.

## Main Text

The spatial organization of DNA in the nucleus is critical to its function. A fundamental component of this organization involves DNA “loops”, physical point-to-point interactions between DNA sequences located far apart on the chromosome. These loops are key to chromatin condensation during mitosis and also regulate enhancer-promoter interactions during interphase. Several recent findings have led to the ‘DNA extrusion model’ of loop formation (Box 1). First theorized to explain how mitotic chromatin condensation might proceed without forming knots, the model also elegantly explains the observed CTCF motif orientation bias discussed in Box 1(ref. [1-6]). The loop extrusion model posits that DNA loops begin as small pinches of the DNA molecule with each side held by one end of a proposed extrusion complex (Figure 3.1a). As the extrusion complex reels in DNA, the loop is progressively enlarged (Figure 3.1b). A

stable loop is formed when the complex stops extruding (Figure 3.1c). This relatively straightforward model is a radical departure from previous thinking, and while it explains several puzzles it poses perhaps more.

An important participant in loop extrusion is the highly conserved SMC family of proteins. SMC complexes assemble into large rings thought to encircle DNA strands . Entrapping DNA entirely within a protein complex leads to topological binding that will only be released by an opening of the protein complex. This renders the binding immune to disruption of the protein-DNA contacts and leads to exceptionally long residency times, while also permitting the free sliding of the SMC ring along the DNA.

The SMC complex condensin is known to organize DNA during mitosis. Processive expansion of initially small loops ensures that loop compaction occurs in order and only within a chromosome, precluding the formation of knots[3]. The SMC complex cohesin colocalizes with CTCF and is required for the formation of CTCF loops in interphase. Degradation of cohesin results in a complete loss of CTCF loops, while its stabilization via degradation of the cohesin release factor WAPL leads to additional loops[7,8,9]. This excessive looping condenses interphase chromatin into dense, mitotic-like “vermicelli” chromosomes. Importantly, this observation suggests that interphase loop formation by cohesin and mitotic condensation by condensin are fundamentally related processes. The SMC family also includes structurally similar members in Prokarya, where bacterial condensin juxtaposes the arms of replicating chromosomes in a manner reminiscent of loop extrusion[10]. It is thus likely that SMC complexes are part of an ancient mechanism of moving and organizing DNA via loop extrusion that has been repurposed to many ends over evolutionary time.

The Missing Motor

The loop extrusion model offers an attractive explanation for the reversible and orderly formation of DNA loops within chromosomes, but mechanistic details remain unknown, including how the proposed extrusion complex responsible for initiating and expanding DNA loops would work. The ability of the SMC family to bind DNA topologically recommends a model in which cohesin and condensin rings hold DNA loop ends, but the formation of loops up to millions of bases in size requires a motor: a mechanism by which DNA is pulled into the SMC loop. Numerous explanations have been proposed as to how loop extrusion is powered, including hitching rides with known DNA motors, such as RNA polymerase, pushing by DNA supercoiling, and passive diffusion along gradients of SMC complexes[11,12,13,14]. While each of these processes may be playing some role, there is now direct experimental evidence that the SMC complex condensin is capable of ATP-dependent unidirectional movement on a DNA substrate *in vitro.* In addition, strong circumstantial evidence suggests that loop extrusion is ATP dependent *in vivo*[15,16]. Condensin attached to DNA curtains was detected moving unidirectionally over a DNA molecule at ~60 base pairs per second[17]. Interestingly, on a relaxed single-tethered DNA curtain, condensin compacts DNA through loop formation, but on the taut DNA of a double-tethered curtain, condensin translocates. This demonstrates that condensin can move along the DNA without forming an intramolecular loop. A subsequent experiment using Sytox Orange staining observed the extrusion of a loop on relaxed DNA at speeds up to ~1500 base pairs per second[18]. Importantly this study revealed that the SMC complex can extrude loops as a single complex, and that this extrusion is unidirectional in nature, with DNA being reeled into the loop from only one direction.

It is now clear that condensin is an ATP-powered DNA motor, but similar experiments performed with the cohesin complex have not detected motor activity[19,20,21]. Cohesin and condensin have remarkably similar architectures, and both have been independently

hypothesized to form loops in DNA via loop extrusion. While it is possible that the intrinsic motor activity of cohesin has been replaced with an external process, it is also possible that *in vitro* assays are missing some critical component or post-translation modification. The extensive literature on the structure and function of the SMC complexes does not offer an immediate explanation for how these machines function as motors and a novel mechanism of active DNA translocation is required.

## The Head, the Hinge, and the HAWKs

The core components of cohesin and condensin complexes are the SMC proteins (Figure 3.2a). These proteins have a complex structure with two globular domains, the head and the hinge, separated by long ~45 nm antiparallel coiled-coils. Pairs of SMC proteins heterodimerize at their hinges. SMC1 and SMC3 form the core of cohesin and SMC2 and SMC4 form condensin. The globular head domains contain ABC-type nucleotide binding domains that are thought to mediate dimerization between the two head domains of a complex. Each complex thereby cooperatively binds 2 ATP molecules. A third SMC component, kleisin, interacts with both head domains of the complex, linking them and forming a tripartite ring (Figure 3.2a). Kleisins are largely disordered peptide chains, much longer than is required to bind the two head domains. Kleisins are further bound by various members of a family of proteins that have come to be known as HEAT-repeat proteins associated with kleisins or HAWKs (Figure 3.2a). This family is rich in HEAT-repeat domains consisting of pairs of antiparallel alpha-helices linked together by just a few amino acids. Found in many proteins throughout the cell, HEAT-repeats are remarkable for their conformational flexibility. These structures adopt a horseshoe-like configuration capable of stretching and scrunching[22]. The kleisins of cohesin and condensin

each interact with a number of these HAWKs, which regulate loading, unloading, and likely the motor activity of these complexes.

For a single SMC ring to achieve unidirectional movement, it must possess two means of interacting with the DNA simultaneously: one that will act as a stationary anchor and another that will produce movement along the DNA. The SMC complex has two reported mechanisms of binding DNA, the hinge domains on one end of the molecule and the kleisin and HAWK subcomplex on the other. The hinge domains of cohesin and condensin have high affinity for single-stranded DNA and some affinity for double-stranded DNA[23]. How the hinge interacts with DNA is still uncertain, but some evidence points to a positively charged groove formed by the inner-side of the hinge and the nearby coiled-coils[24]. DNA binding by the hinge has been shown to catalyze ATP hydrolysis by the head domains, and disruption of the hinge can disrupt the function of the entire complex[18]. The SMC hinge is a critical component of the complex that is likely key to the mechanochemical cycle driving SMC movement. In condensin, the kleisin and HAWK subcomplex forms a positively-charged pocket that wraps around the DNA fiber in what is described as a “safety belt” binding mechanism[25]. This creates a topological engagement that holds DNA in a sequence-independent manner. While this specific DNA-binding conformation has only been directly observed in the Brn1-Ycg1 kleisin-HAWK complex of *S. cerevisiae*, many HAWKs have DNA binding affinity. Structural similarities between the kleisin-HAWK subcomplexes that form part of cohesin suggest this may be a conserved mechanism of DNA binding. The kleisins of cohesin and condensin bind to at least two HAWK components simultaneously potentially forming multiple DNA binding pockets in each complex. These subcomplexes could bind to the same molecule of DNA or possibly hold two separate molecules of DNA together.

## The Anchor and the Motor

Even with an understanding of how SMC complexes might engage DNA, it is not immediately obvious which end of the SMC complex would remain stationary and which end would move along the DNA. It has been proposed that the kleisin-HAWK topological binding pocket may serve as the anchor[26], which would leave the comparatively simple hinge domain to serve as a motor. In one proposed model the SMC arms and hinge act as a DNA pump[26,27]. In this model, DNA loops are loaded into the ring formed by the SMC arms and ATP driven conformational changes close the ring, driving the loop into a smaller chamber formed by the kleisin and SMC heads where it combines with a larger loop. This model posits a topological binding of the DNA by the SMC-kleisin ring. However, a recent study of cohesin suggests that SMC rings incapable of topologically binding the DNA are still capable of extrusion[28]. Another potential model for hinge-mediated motor activity might be ATP-driven dissociation of the hinge leading to a walking mechanism. However, studies of the DNA binding capabilities of the hinge monomers have seen little to no independent DNA binding ability[23].

Alternatively, the hinge domain could serve as the anchor, while the kleisin and HAWK subcomplexes move along the DNA. Several features of the structure support this model. The topological engagement of the kleisin-HAWK binding domain would allow movement of the DNA through the groove without release. Indeed, the loose nature of the DNA binding pocket results in low binding affinity for short DNA fragments, suggesting they could slide out of the groove[25]. Additionally, the kleisin and HAWK components appear uniquely suited for large conformational changes. The kleisin-HAWK DNA binding domain is not conformationally frozen, with different configurations of the HAWK and DNA observed in different crystals[25]. Between the terminal domains of kleisins, the protein is mostly unstructured, and much longer than would seem necessary to connect the two head domains together, suggesting there may exist some “slack” in this tether. The HEAT-repeats of the HAWKs are found in many other proteins, where they

are known to stretch and compress in response to mechanical force[22]. Indeed, HEAT-repeats can be thought of as springs capable of stretching and contracting while storing and releasing potential energy[29]. Cryo-EM analysis of the HAWK protein Scc2 revealed a high degree of conformational flexibility with an estimated capacity to stretch lengthwise up to ~11 nm[30]. Taken together, the kleisin-HAWK DNA binding domain would appear to be capable of undergoing large conformational changes and sliding along the DNA. We therefore propose that the kleisin-HAWK subcomplexes represent the mobile DNA binding domain.

A model for SMC complex translocation on DNA must be compatible with both eukaryotic and prokaryotic SMC members. Prokaryotic SMC complexes lack HAWK proteins. Instead their kleisins are bound by much smaller Kite proteins that nevertheless appear to have functional similarities to the HAWKs[31]. Kite proteins are composed of two Winged-Helix Domains (WHD) connected by an intrinsically disordered linker. Each WHD binds to the kleisin creating the potential for two topological DNA binding grooves. Indeed, the eukaryotic Kites of the SMC5/6 complex have recently been found to bind DNA[32]. The disordered linker would permit the orientations of the WHDs to change dramatically allowing for folding and opening that could mimic the conformational flexibility of the HAWK proteins[33]. That the unrelated Kite and HAWK families share distinctive functional characteristics suggests that they might play a conserved role as flexible DNA binding components of the SMC complexes.

## Kinetics

The step rate and step size of the extrusion process are important criteria for evaluating potential models of SMC motors. Unfortunately, the existing estimates of SMC motor kinetics are rough and ambiguous. The speed at which the SMC complex moves depends on the rate at which it steps and the size of its steps. If the SMC heads function similarly to related ABC-type domains, then each ATPase cycle most likely corresponds to the hydrolysis of 1 or 2 molecules

of ATP. In the presence of DNA, condensin hydrolyzes ATP at a rate of ~2 ATP per second[17]. However, this bulk rate represents a mixture of condensin molecules in various states: actively extruding complexes, DNA-bound but stationary, non-extruding complexes, and non-DNA bound complexes. Therefore, the rate of hydrolysis of an actively extruding complex could be significantly higher than this average rate. The most unambiguous observation of the extrusion speed of condensin shows a single condensin extruding up to ~1,500 bp or ~500 nm per second[18]. However, the rate of extrusion displays a strong dependence on the tension on the DNA fiber and slows to a more modest ~600 bp per second rate at physiological tensions of ~0.4 pN. Whether this reduction in speed is a result of changes in step sizes, step rates, or the proportion of productive steps, will have important implications for the mechanism of the SMC motor. Importantly, the experiments discussed above were performed on naked DNA lacking nucleosomes. ATP-independent diffusion of cohesin on DNA is significantly impeded by the presence of nucleosomes[19]. Additionally, the force generated by condensin extrusion, estimated at ~1 pN, would be insufficient to evict the histone octamer[34]. This suggests that SMC complexes likely possess the ability to actively translocate past nucleosomes on chromatin.

Step size can be directly measured by experiments using magnetic tweezers, which precisely detect the compaction of DNA with high temporal resolution. Several magnetic tweezer experiments using condensin and cohesin from *S. cerevisiae* as well as condensin I from *X. laevis* have demonstrated DNA compaction on naked DNA occurring in highly variable steps larger than 100 nm in size[35,36,37]. Such large step sizes are incompatible with models that limit themselves to the ~50 nm length of SMC complexes. However, there is evidence to suggest these steps represent a mechanism of compaction distinct from extrusion. Similar large DNA compaction steps are observed for budding yeast condensin in the absence of ATP; these have been demonstrated to be distinct from smaller, co-occurring steps[34]. Two separate DNA compaction mechanisms have been reported for bacterial SMC complexes as well[38]. Most likely

these large steps represent some form of loop capture distinct from extrusion. Both cohesin and condensin have demonstrated some capability to form inter-complex interactions that could explain these large compaction steps. Further studies will be needed to distinguish between these processes and to establish the kinetics of the SMC motors.

## The Tethered Inchworm Model

While the kleisin-HAWK DNA bound subcomplex is in principle capable of accommodating large conformational changes, it must be the ATP-hydrolyzing head domains that provide the motive force. The ABC-type ATPase domains located in the SMC head domains form 2 ATP binding sites when engaged. ABC-type domains are thought to have a conserved mechanism of action where ATP binding and hydrolysis correspond to head engagement and disengagement[39]. ATP-mediated head engagement is accompanied by a conformational shift, often a rotation, of the interface between the two domains to accommodate the nucleotides. Commonly this rotation is propagated into adjacent domains to perform mechanical work. Crystal structures of SMC heads reveal that ATP-bound forms are rotated ~30 degrees in relation to their unbound form[40]. This rotation dramatically increases the angle between the coiled-coil arms as they exit the head domains. Driving the coiled-coil arms apart likely forces them to bend, widening the ring and propagating this steric strain all the way to the hinge domain (Figure 3.2b). It has been proposed that this tension is relieved by ATP hydrolysis followed by disengagement and separation of the head domains[40]. In this way, ATP binding and hydrolysis could force the head domains apart, using the arms as force-amplifying levers (Figure 3.2c). No structural data for this open conformation exists, however AFM images of SMC dimers often show large, >50 nm distances between the head domains[41].

Taking into consideration the conformational flexibility present in the kleisin-HAWK subcomplexes linking the two head domains, it is possible the kleisin might remain bound to

both heads as they are pulled apart[42]. The disordered structure of kleisin could straighten and unfurl to accommodate this motion. In doing so, this could stretch the HAWK subunits bound at multiple points to the kleisins. If the kleisin-HAWK subcomplexes are topologically engaged with the DNA molecule, then this movement could be permitted by sliding the proteins along the DNA. Together these conformational changes would spread the SMC complex along the DNA. These motions could generate productive unidirectional movement if they were coordinated with changes in DNA binding affinity in the kleisin-HAWK subcomplexes. If in the closed configuration two kleisin-HAWK binding domains had differing affinities for DNA, then the side more weakly bound would preferentially move upon head separation. This would cause the less tightly bound HAWK to slide forward along the DNA (Figure 3.2c). A subsequent closing motion would pull the lagging end of the complex forward, assuming that the stretching of the kleisin-HAWK subcomplexes reversed the DNA binding affinities of the proteins (Figure 3.2d). The HEAT-repeats of the HAWKs act as springs, storing potential energy in their conformational changes. This energy might help drive the lagging step by pulling the head domains back together. Dimerization of the reunited SMC heads would complete a mechanochemical cycle in which ATP binding and hydrolysis powers net unidirectional movement along the DNA. This model, in which opening of the SMC ring pushes the leading end forward along the DNA and subsequent closing pulls the lagging end up, is akin to an inchworm motor. The interesting topology of the DNA-bound complex leads us to suggest the more descriptive term “tethered inchworm” for this model of SMC locomotion.

The tethered inchworm model is a general framework lacking in specifics and leaves several important questions unanswered. A wide range of step sizes would be compatible with this model due to the extremely flexible nature of each component. Step sizes upwards of ~50 nm could be accommodated by eukaryotic kleisins but will ultimately depend on the separation driven by ATP-binding and hydrolysis, which is likely smaller. A related question is how SMC

complexes navigate obstacles such as nucleosomes. While the DNA binding grooves of the kleisin-HAWK subcomplexes are not large enough to permit ~11 nm sized nucleosomes, it is conceivable that HAWK-kleisin dissociation during the walking cycle would allow SMC complexes to step over nucleosomes. It is also unclear in which direction the SMC complex moves, which would be determined by the order of changes in binding affinity of the kleisin-HAWK subcomplexes. Nevertheless, this putative model may begin to explain the known regulatory roles of various HAWKs on SMC function. Chromatin-bound cohesin consists of both mobile and immobile fractions[43]. Depleting the cohesin HAWK PDS5, rather than stopping extrusion, results in enhanced extrusion and condensation of the genome, suggesting that PDS5 may function as a component of immobile cohesin complexes[9]. PDS5 competes for its kleisin binding site with the HAWK NIPBL, whose depletion results in a loss of loops and extrusion[44,45]. PDS5 and NIPBL may represent static and mobile HAWK components, respectively, which compete to turn the cohesin motor off and on[46]. The tethered inchworm model is highly speculative, but our new perception of SMC complexes as loop extruding motors requires a bold reimagining of previous knowledge. Our proposal that the HAWK proteins are conformationally flexible and dynamic DNA binding elements is conjecture that is required to create a functional model of motor activity. Further study of the enigmatic HAWK family will be needed to evaluate this proposition. Our relatively better understanding of the core SMC proteins is unable to account for the motor activity of the SMC complex. Thus, understanding the functions of kleisin and HAWK proteins, namely whether and how they bind to DNA, what conformational changes they undergo during the ATP-hydrolysis cycle, and what roles different subunits play in regulating the complexes, will likely prove key to elucidating the motor function of SMC complexes. Future work on the structures and kinetics of SMC complexes will refine our understanding of this fascinating protein family responsible for DNA organization across all domains of life.

# Figures

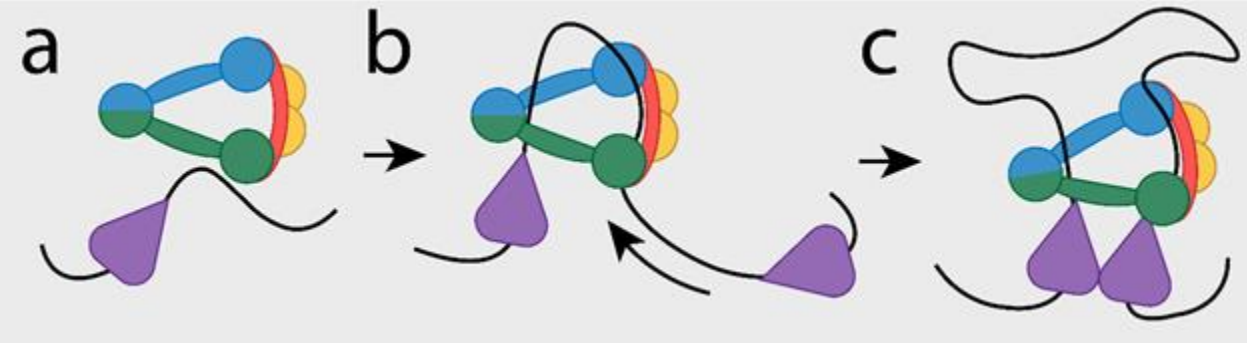



Figure 3.1. DNA loop extrusion model

a. Cohesin complexes load onto the DNA, either randomly or at specific sites, such as CTCF binding sites (purple). b. The cohesin complex reels in DNA, translocating over the DNA and expanding the loop. c. Cohesin complexes stop extruding when they meet a properly oriented CTCF site, leading to a loop between convergently-oriented CTCF anchors.

Box 1. Loop extrusion has been independently proposed to explain the formation of numerous types of DNA loops, but the recent surge in interest is due to the ability of this model to explain the curious phenomenon of motif-oriented CTCF looping. CTCF loops are thought to be formed by two CTCF proteins bound to separate motifs on a chromosome. These loops are a clear and prominent feature of how the genome is organized. The asymmetric CTCF binding motif has an orientation that plays a fundamental role in the formation of these loops. As revealed by chromatin conformation capture assays, CTCF sites interact with each other significantly more when arranged in a convergent orientation. In agreement with this, CTCF loops form predominantly between CTCF sites oriented towards each other47. Conversely, CTCF sites oriented away from each other only rarely form loops. This finding has fundamental implications for the mechanism of loop formation. A simplistic model of loop formation via stabilization of

stochastic collisions taking place in the three-dimensional space cannot account for this orientation bias. Rather, the loop formation mechanism must account for the orientation context of CTCF sites up to millions of base pairs apart. Loop extrusion solves this conundrum by having loops begin as small bends in the DNA that are progressively expanded (Figure 3.1). A loop extruder is theorized to expand the loop by translocating along the DNA, reeling the chromatin into the loop. The orientation bias of CTCF sites can then be explained by orientation-dependent interactions of CTCF with the extrusion machinery.

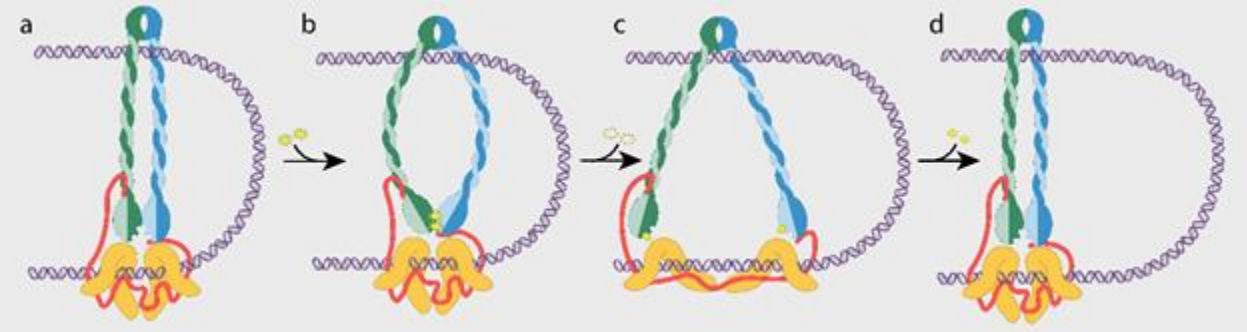



Figure 3.2 The Tethered Inchworm Model. a. The SMC complex is composed of two SMC proteins (green and blue) which dimerize at the hinge (top) and at the head domain (bottom). Tethering the two heads together is a kleisin (red) further bound by HAWK proteins (orange). The SMC complex forms a small loop in the DNA (purple) by binding with both the hinge dimer and the kleisin-HAWK subcomplexes. b. Binding of 2 ATP molecules (yellow) by the ATPase head domains (green and blue) induces a conformational rotation of each head. This movement forces the coiled-coil arms apart, bending them, and propagating the strain to the hinge domains. c. ATP hydrolysis causes dissociation of the head domains and opening of the SMC

arms. In this model, the leading HAWK would slide forward along the DNA due to its weaker affinity for DNA. The kleisin would straighten and unfurl to accommodate this movement and in doing so pull on the HAWKs, stretching these spring-like proteins. d. In the extended configuration the DNA binding affinities of the HAWKs then reverse causing the lagging HAWK to catch up as the head domains reunite, completing a mechanochemical cycle that has enlarged the DNA loop.

## References

1. Riggs, A. D. DNA methylation and late replication probably aid cell memory, and type I DNA reeling could aid chromosome folding and enhancer function. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 326, 285–297 (1990).

2. Nasmyth, K. Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annu. Rev. Genet.* 35, 673–745 (2001).

3. Alipour, E. & Marko, J. F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res.* 40, 11202–11212 (2012).

4. Nichols, M. H. & Corces, V. G. A CTCF Code for 3D Genome Architecture. *Cell* 162, 703–705 (2015).

5. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *PNAS* 112, E6456–E6465 (2015).

6. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* 15, 2038–2049 (2016).

7. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* 171, 305-320.e24 (2017).

8. Haarhuis, J. H. I. *et al.* The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* 169, 693-707.e14 (2017).

9. Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO Journal* 36, 3573–3599 (2017).

10. Wang, X., Brandão, H. B., Le, T. B. K., Laub, M. T. & Rudner, D. Z. Bacillus subtilis SMC complexes juxtapose chromosome arms as they travel from origin to terminus. *Science* 355, 524–527 (2017).

11. Barrington, C., Finn, R. & Hadjur, S. Cohesin biology meets the loop extrusion model. *Chromosome Res* 25, 51–60 (2017).

12. Racko, D., Benedetti, F., Dorier, J. & Stasiak, A. Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes. *Nucleic Acids Research* 46, 1648–1660 (2018).

13. Yamamoto, T. & Schiessel, H. Osmotic mechanism of the loop extrusion process. *Phys. Rev. E* 96, 030402 (2017).

14. Brackley, C. A. *et al.* Extrusion without a motor: a new take on the loop extrusion model of genome organization. *Nucleus* 9, 95–103 (2018).

15. Vian, L. *et al.* The Energetics and Physiological Impact of Cohesin Extrusion. *Cell* 173, 1165-1178.e20 (2018).

16. Wang, X. *et al.* In vivo evidence for ATPase-dependent DNA translocation by the Bacillus subtilis SMC condensin complex. *Molecular Cell* In Press,

17. Terakawa, T. *et al.* The condensin complex is a mechanochemical motor that translocates along DNA. 6 (2018).

18. Ganji, M. *et al.* Real-time imaging of DNA loop extrusion by condensin. *Science* eaar7831 (2018). doi:10.1126/science.aar7831

This study directly imaged real-time unidirectional loop extrusion by single condensin complexes. This represents the strongest evidence to date that SMC complexes are DNA motors and provides important insights into the mechanism by which extrusion occurs.

19. Stigler, J., Çamdere, G., Koshland, D. E. & Greene, E. C. Single-Molecule Imaging Reveals a Collapsed Conformational State for DNA-Bound Cohesin. *Cell Rep* 15, 988–98 (2016).

20. Kanke, M., Tahara, E., Veld, P. J. H. in't & Nishiyama, T. Cohesin acetylation and Wapl-Pds5 oppositely regulate translocation of cohesin along DNA. *The EMBO Journal* 35, 2686–2698

21. Davidson, I. F. *et al.* Rapid movement and transcriptional re-localization of human cohesin on DNA. *EMBO J.* 35, 2671–2685 (2016).

22. Yoshimura, S. H. & Hirano, T. HEAT repeats - versatile arrays of amphiphilic helices working in crowded environments? *J. Cell. Sci.* 129, 3963–3970 (2016).

23. Hirano, M. & Hirano, T. Opening closed arms: long-distance activation of SMC ATPase by hinge-DNA interactions. *Mol Cell* 21, 175–86 (2006).

24. Chiu, A., Revenkova, E. & Jessberger, R. DNA interaction and dimerization of eukaryotic SMC hinge domains. *J Biol Chem* 279, 26233–42 (2004).

25. Kschonsak, M. *et al.* Structural Basis for a Safety-Belt Mechanism That Anchors Condensin to Chromosomes. *Cell* 171, 588-600.e24 (2017).

This study identified the ability and mechanism by which a kleisin-HAWK subcomplex binds DNA. The topological and labile nature of this interaction is, we believe, illustrative of all kleisin-HAWK-DNA interactions and is integral to the tethered inchworm model.

26. Diebold-Durand, M.-L. *et al.* Structure of Full-Length SMC and Rearrangements Required for Chromosome Organization. *Molecular Cell* 67, 334-347.e5 (2017).

27. Marko, J. F., Rios, P. D. L., Barducci, A. & Gruber, S. DNA-segment-capture model for loop extrusion by structural maintenance of chromosome (SMC) protein complexes. *bioRxiv* 325373 (2018). doi:10.1101/325373

28. Srinivasan, M. *et al.* The Cohesin Ring Uses Its Hinge to Organize DNA Using Non-topological as well as Topological Mechanisms. *Cell* 173, 1508-1519.e18 (2018).

29. Kappel, C., Zachariae, U., Dölker, N. & Grubmüller, H. An Unusual Hydrophobic Core Confers Extreme Flexibility to HEAT Repeat Proteins. *Biophys J* 99, 1596–1603 (2010).

30. Chao, W. C. H. *et al.* Structure of the cohesin loader Scc2. *Nature Communications* 8, 13952 (2017).

31. Wells, J. N., Gligoris, T. G., Nasmyth, K. A. & Marsh, J. A. Evolution of condensin and cohesin complexes driven by replacement of Kite by Hawk proteins. *Curr Biol* 27, R17–R18 (2017).

32. Zabrady, K. *et al.* Chromatin association of the SMC5/6 complex is dependent on binding of its NSE3 subunit to DNA. *Nucleic Acids Res.* 44, 1064–1079 (2016).

33. Kamada, K., Miyata, M. & Hirano, T. Molecular Basis of SMC ATPase Activation: Role of Internal Structural Changes of the Regulatory Subcomplex ScpAB. *Structure* 21, 581–594 (2013).

34. Keenholtz, R. A. *et al.* Oligomerization and ATP stimulate condensin-mediated DNA compaction. *Scientific Reports* 7, 14279 (2017).

35. Eeftens, J. M. *et al.* Real-time detection of condensin-driven DNA compaction reveals a multistep binding mechanism. *EMBO J* 36, 3448–3457 (2017).

36. Strick, T. R., Kawaguchi, T. & Hirano, T. Real-time detection of single-molecule DNA compaction by condensin I. *Curr Biol* 14, 874–80 (2004).

37. Sun, M., Nishino, T. & Marko, J. F. The SMC1-SMC3 cohesin heterodimer structures DNA through supercoiling-dependent loop formation. *Nucleic Acids Res* 41, 6149–6160 (2013).

38. Kim, H. & Loparo, J. J. Multistep assembly of DNA condensation clusters by SMC. *Nat Commun* 7, (2016).

39. Hopfner, K. P. & Tainer, J. A. Rad50/SMC proteins and ABC transporters: unifying concepts from high-resolution structures. *Curr Opin Struct Biol* 13, 249–55 (2003).

40. Kamada, K., Su'etsugu, M., Takada, H., Miyata, M. & Hirano, T. Overall Shapes of the SMC-ScpAB Complex Are Determined by Balance between Constraint and Relaxation of Its Structural Parts. *Structure* 25, 603-616.e4 (2017).

This study crystallized ATP-bound condensin heads and showed that they undergo a large conformational shift. The authors propose a model of ATP binding and hydrolysis driving apart the SMC heads that is key to the tethered inchworm model.

41. Eeftens, J. M. *et al.* Condensin Smc2-Smc4 Dimers Are Flexible and Dynamic. *Cell Rep* 14, 1813–8 (2016).

42. Nasmyth, K. & Haering, C. H. The Structure and Function of Smc and Kleisin Complexes. *Annual Review of Biochemistry* 74, 595–648 (2005).

43. Gerlich, D., Koch, B., Dupeux, F., Peters, J. M. & Ellenberg, J. Live-cell imaging reveals a stable cohesin-chromatin interaction after but not before DNA replication. *Curr Biol* 16, 1571–8 (2006).

44. Kikuchi, S., Borek, D. M., Otwinowski, Z., Tomchick, D. R. & Yu, H. Crystal structure of the cohesin loader Scc2 and insight into cohesinopathy. *Proc. Natl. Acad. Sci. U.S.A.* 113, 12444–12449 (2016).

45. Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* 551, 51–56 (2017).

46. Petela, N. *et al.* Multiple interactions between Scc1 and Scc2 activate cohesin's DNA dependent ATPase and replace Pds5 during loading. *bioRxiv* 205914 (2017). doi:10.1101/205914

47. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 159, 1665–1680 (2014).

Acknowledgments

Work in the authors' lab is supported by U.S. Public Health Service Award R01 GM035463 from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## Competing Interests

The authors declare no competing interests.

# Chapter 4: Evolutionarily conserved principles predict 3D chromatin organization

M Jordan Rowley*, Michael H Nichols*, Xiaowen Lyu , Masami Ando-Kuri , I Sarahi M Rivera , Karen Hermetz , Ping Wang , Yijun Ruan , Victor G Corces

*These authors contributed equally to this work.

Previously published in: Rowley MJ, Nichols MH, Lyu X, et al. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Mol Cell*. 2017;67(5):837-852.e7. doi:10.1016/j.molcel.2017.07.022

## Summary

TADs, CTCF loop domains, and A/B compartments have been identified as important structural and functional components of 3D chromatin organization, yet the relationship between these features is not well understood. Using high-resolution Hi-C and HiChIP we show that *Drosophila* chromatin is organized into domains we term compartmental domains that correspond precisely with A/B compartments at high resolution. We find that transcriptional state is a major predictor of Hi-C contact maps in several eukaryotes tested, including *C. elegans* and *A. thaliana*. Architectural proteins insulate compartmental domains by reducing interaction frequencies between neighboring regions in *Drosophila*, but CTCF loops do not play a distinct role in this organism. In mammals, compartmental domains exist alongside CTCF loop domains to form topological domains. The results suggest that compartmental domains are responsible for domain structure in all eukaryotes, with CTCF playing an important role in domain formation in mammals.

# Introduction

The development of Hi-C has enabled the examination of the 3D chromatin conformation of an entire genome. The first Hi-C analyses of mammalian genomes provided low resolution (ca. 1 Mb) contact maps revealing a plaid pattern of interactions representing active A and inactive B compartments (Lieberman-Aiden et al., 2009). Subsequent higher resolution Hi-C experiments (ca. 50 kb) identified topologically associating domains (TADs), which are contiguous segments of the genome that preferentially interact whithin themselves over neighboring regions (reviewed in Rowley and Corces, 2016). TADs in mammals have an average size between 200 kb and 1 Mb and were originally described as related to, but independent of, compartments (Dixon et al., 2012). Using high resolution (ca. 1 kb) data, Lieberman Aiden and collaborators defined contact domains smaller in size than TADs (Rao et al., 2014). Borders of a subset of these smaller contact domains were found to interact preferentially over the rest of the domain creating a “peak” or more intense spot in the Hi-C contact map (Rao et al., 2014). These Hi-C peaks correlate with the presence of the architectural protein CTCF, suggesting that many of these contact domains are CTCF loops (Rao et al., 2014). Strikingly, the orientation of the CTCF motif appears to determine the direction in which CTCF sites will form loops, with convergently oriented CTCF motifs highly enriched at the anchors of CTCF loops (Guo et al., 2015; Rao et al., 2014). Contact domain boundaries often correspond to CTCF loop anchors, but some do not, suggesting that principles other than CTCF-mediated interactions may also govern the establishment of contact domains (Rao et al., 2014).

TADs have also been identified in *Drosophila*, but the low resolution of Hi-C data in early studies has limited the precision with which these domains can be mapped and identified (Hou et al., 2012; Sexton et al., 2012). Broadly, TAD borders defined at 10 kb resolution were reported to be enriched in clusters of architectural protein binding sites (APBSs) (reviewed in Rowley and Corces, 2016). APBSs are often associated with promoters of highly expressed genes,

suggesting a possible relationship between transcription and TAD border formation (Hou et al., 2012; Van Bortle et al., 2014). Several studies have found that boundaries/inter-TAD regions correlate with active chromatin (El-Sharnouby et al., 2017; Hug et al., 2017; Ulianov et al., 2016). However, whether active regions exhibit their own structure or are simply boundaries between TADs is a matter of debate due to the low resolution of currently available Hi-C datasets. Patterns of 3D chromatin organization identified in mammals and *Drosophila* have been found to be applicable to other model organisms. Contact domains of varying size have been found in *S. pombe, S. cerevisiae, C. elegans, and A. thaliana* (reviewed in Rowley and Corces, 2016). These organisms have no known CTCF homologs yet they can form distinct domains reminiscent of those seen in humans. The mechanisms responsible for the establishment of contact domains in these organisms are not known, and it is unclear whether conserved processes are involved in the formation of domains of different sizes and strengths across the evolutionary tree.

High resolution (ca. 250 bp) Here we show that high resolution (ca. 250 bp) Hi-C data in *D. melanogaster* suggest the existence of domains, which we term compartmental domains, smaller in size than the TADs defined originally. Distinct from mammals, we find no evidence of looping mediated by CTCF or other architectural proteins between borders of these domains. Using HiChIP and ChIA-PET for histone modifications and RNA Polymerase II (RNAPII), we find that domains are a direct result of the establishment of A/B compartments defined by the chromatin state of their interior rather than by a border element. This principle also applies to other eukaryotic organisms. Furthermore, we show that mammalian chromosome organization is established via a combination of compartmental domains and point-to-point CTCF interactions, leading to the formation of distinct but often overlapping domains. We conclude that compartmental domains represent the primary mechanism underlying 3D chromatin organization in eukaryotes but that architectural proteins, especially CTCF, are responsible for

additional point-to-point interactions that establish the complex 3D architecture of the mammalian nucleus.

## Results

### Compartmental Domains Are the Main Feature of *Drosophila* Chromatin Organization

Studies of *Drosophila* 3D chromatin organization have identified TADs that are smaller than typical mammalian TADs (Sexton et al., 2012). To gain further insights into the principles controlling the establishment of 3D chromatin organization in *D. melanogaster,* we combined Hi-C datasets acquired in Kc167 cells to obtain nearly a billion uniquely mapped reads (Cubeñas-Potts et al., 2016). In comparison to the ultra-high resolution Hi-C dataset in humans (Rao et al., 2014), this is equivalent to 12-fold higher contacts at short distances (<10kb) (Figure 4.S1A). The high resolution Hi-C map exhibits a clear checkerboard pattern reminiscent of A/B compartments originally found in humans at 1 Mb resolution (Lieberman-Aiden et al., 2009), but is evident in *Drosophila* in 10 kb resolution Pearson correlation maps (Figure 4.1A). To classify these compartments we used a principal component analysis (eigenvector decomposition) of the Pearson correlation matrix (Lieberman-Aiden et al., 2009) at 10 kb resolution (Figure 4.1A, right panel). In mammals, A (active) compartments have high levels of transcriptional activity, chromatin accessibility, and active histone modifications. To test if this is also the case in *Drosophila*, we performed Fast-ATAC-seq (Corces et al., 2016) and examined GRO-seq data. We find that A compartments have higher transcription and chromatin accessibility than B compartments (Figure 4.1B, 4.S1BC). Next, we performed ChIP-seq for seven different histone modifications/variants, including H3K36me3, H4K16ac, H4K20me1, H3K9me3, ubiquitinated H2B (H2Bub), H3.3, and H2A.Z. We also examined previously published ChIP-seq data for H3K27ac, H3K27me3, H3K4me1, H3K4me3, and H3K9me2. We found that the eigenvector closely follows the switch between active and inactive histone modifications (Figure 4.1B). We

tested the relative levels of histone modifications across the compartments and found that the two compartments generally partition active from inactive chromatin (Figure 4.1C, 4.S1D) which is similar to the partitioning of A and B compartments observed in mammals (Lieberman-Aiden et al., 2009).

Upon examination of compartments in *Drosophila,* we noticed several locations with visibly evident compartment switches in the Hi-C heatmap that are unidentified by the standard algorithm (Figure 4.S1E), and thus sought an alternate method to better characterize these fine-scale compartments. Since H3K27ac and H3K27me3 show the most pronounced distinction between A and B compartments (Figure 4.1C, 4.S1F), we performed HiChIP (Mumbach et al., 2016) using antibodies for these two histone modifications (Table 4.S1-4.S2). We chose these histone modifications not only because of their close correspondence to A and B compartments, but because of their prevalence in the *Drosophila* genome, such that nearly every 1 kb bin has either H3K27ac or H3K27me3 (Figure 4.1B, 4.S1G). H3K27me3 is absent at H3K27ac peaks, is highly enriched at Pc-repressed loci, and shows an intermediate level of enrichment in the rest of the genome (Figure 4.1B), a feature that has also been reported by others (El-Sharnouby et al., 2017). We found that HiChIP for H3K27ac or H3K27me3 effectively enriched for A or B compartments respectively (Figure 4.1D, 4.S1H). We next classified compartments at 10 kb resolution using the ratio of interactions from H3K27ac HiChIP versus H3K27me3 HiChIP datasets, and found that the result closely matches the Hi-C eigenvector obtained from principal component analysis. However, the compartment calls obtained using HiChIP data allow the discovery of small compartments that were previously undetected by the Hi-C eigenvector (Figure 4.S1E). Because we found that either H3K27ac or H3K27me3 occupy most of the genome, we then tested how well the HiChIP contact maps recapitulate the full Hi-C data. We combined reads obtained from H3K27ac and from H3K27me3 HiChIP into a single contact map and found a 98.9% correlation with Hi-C data (Figure 4.S1IJ). Altogether this indicates that

HiChIP for these two histone modifications, when combined, can recapitulate Hi-C data, but when used separately can accurately capture compartmental interactions.

Compartments were originally identified in humans at 1 Mb resolution (Lieberman-Aiden et al., 2009) which has led to the notion that compartments are structures encompassing large swaths of the genome. In *Drosophila,* however, we have identified small compartments at 10 kb resolution, indicating that compartments are actually fine-scale features of chromatin organization. We further tested the scalability of compartments by calling compartments at 1 kb resolution. This provided an overall good correspondence between calls at 1 kb and 10 kb resolution, although 1 kb resolution calls afford better identification of some small compartments (Figure 4.S1K). This indicates that compartments represent small, discrete, and scalable interactions that occur between loci with correlated chromatin and transcriptional activity states. We will refer to these domains as compartmental domains in the rest of the manuscript.

*Drosophila* Domain Organization is not a Result of CTCF Looping

High resolution Hi-C contact maps in mammals have shown the presence of strong point-to-point interactions, manifested as bright spots in Hi-C heatmaps, that correspond to CTCF loops at contact domain corners (Rao et al., 2014) (Figure 4.2A). High resolution Hi-C contact maps in *Drosophila* also show the presence of what appear to be similar spots that seem to correspond to interactions between borders of domains (Figure 4.2B). However, we find that the signal corresponding to these interactions is not punctate; instead, it extends beyond the corners of individual domains (blue arrowheads in Figure 4.2B left; see also the magnified view in the right panel). This signal in fact corresponds to compartmental interactions between small flanking domains (Figure 4.2B right). Detection of these domains requires very high resolution Hi-C maps, explaining why previous studies have misidentified these domains as TAD borders and their interactions as loops formed by interactions between boundaries of TADs. Visualization of

these domains in *Drosophila* also requires heatmaps at a smaller genomic scale than in humans due to their differences in size (Figure 4.2AB). Similar to CTCF loops found in human cells, we also found 458 interaction peaks in *Drosophila* enriched in various architectural proteins, but unlike in humans, we did not see an enrichment of CTCF at the anchors of these loops (Cubeñas-Potts et al., 2016; Rao et al., 2014). Importantly, these interaction peaks do not occur at domain corners (Figure 4.S2A). Altogether, these data indicate that domains in *Drosophila* are likely not the result of the establishment of point-to-point interactions by CTCF or other architectural proteins.

In human cells, interaction peaks at some domain corners occur between convergently oriented CTCF sites (Rao et al., 2014). We thus examined *Drosophila* Hi-C data to determine whether the orientation of the CTCF binding motif influences contact domain structure without the need for strongly stabilized boundary associated CTCF loops. We found that only 28% of domains have CTCF within 3 kb of each border. Of those that have CTCF, there is no evidence for motif orientation preference, in contrast to CTCF borders in human cells (Figure 4.S2B). Additionally, the relationship between human CTCF motif orientation and the interaction preference can be visualized at bound CTCF motifs where Hi-C interactions preferentially occur in the same direction as the motif orientation. In humans, right facing CTCF sites preferentially interact with other genomic sequences to the right along a chromosome (Figure 4.2C red) and left facing CTCF sites interact to the left (Figure 4.2C blue). We performed this same analysis in *Drosophila* to test if interactions at CTCF bound motifs follow the same rule. In contrast to humans, *Drosophila* CTCF sites show no directional preference when interacting with other sites along the chromosome (Figure 4.2C bottom). Overall, this indicates that *Drosophila* domains can form without stabilized point-to-point border interactions between CTCF sites, and that *Drosophila's* CTCF differs fundamentally in its function from the human homolog.

Gene Mini-Domains Underlie *Drosophila* Chromatin Organization

Sequences located between large domains appear to be small active domains (Figure 4.2B and 4.S1E). To explore this further, we examined published TAD calls and found that small domains have been consistently misclassified by previous studies due to the low resolution of the Hi-C maps available. For example, TAD calls at low resolution in *Drosophila* frequently labeled small domains as TAD borders (Hou et al., 2012; Sexton et al., 2012) (Figure 4.S2C). Other attempts at domain calling at low resolution labeled many of these domains as inter-TAD regions (Ulianov et al., 2016). More recently, TAD borders identified in nuclear cycle 14 staged embryos correlated with RNAPII (Hug et al., 2017) correspond in fact to small domains and RNAPII is not present at borders between TADs but it is present throughout every active compartmental domain (Figure 4.S2DE). Thus, we find that borders are not defined by transcriptionally active regions/RNAPII binding as was previously suggested (Hug et al., 2017; Ulianov et al., 2016), but rather by the segregation between active and inactive regions that form compartmental domains, suggesting that this is the prevalent mechanism of domain formation in *Drosophila* (Figure 4.S2F). We therefore refer to these domain structures along the diagonal as compartmental domains as described above because they coincide with the A/B compartments defined by Principal Component Analysis.

Small transcriptionally active domains interact to the exclusion of the larger silent or intergenic regions of the genome in a compartmental manner. We tested whether these interactions are associated directly with transcriptional elongation by performing HiChIP with an antibody for RNA Polymerase II phosphorylated on serine 2 (RNAPIISer2ph) (Table S3). We found that the small active compartments found by Hi-C are highly enriched in RNAPIISer2ph HiChIP signal (Figure 4.2D, 4.S3AB). Closer examination of these data indicates the presence of even smaller domains comprised of individual genes (Figure 4.S3C top right). Because an enrichment of interactions is seen within the gene body we call these structures gene mini-domains. To further confirm these findings, we also performed ChIA-PET for RNAPII and found similar gene mini-

domains (Figure 4.2E top right, 4.S3D-F). Hi-C also shows the presence of domains that coincide precisely with a single actively transcribed gene (Figure 4.2E, 4.S3C-E; see panels below the diagonal). Because we found that active compartments are composed of RNAPII interactions in gene mini-domains, we propose that interactions within and between A compartmental domains are composed of gene-to-gene interactions. We took genes at each expression level (no expression and lowest to highest quartiles of GRO-seq signal) and found that gene-to-gene interactions in A compartments correlate with expression (Figure 4.2F). These observations suggest that active compartmental domains are created in a hierarchical manner by gene mini-domains and gene-to-gene interactions.

The correlation between transcription, compartmental interactions, and domain formation suggests that transcriptional activity may be a good measure of domain structure in *Drosophila*. To test this, we used a hidden Markov model (HMM) to classify the genome into active and inactive states based on GRO-seq levels. We find that borders between domains observed using Hi-C form precisely at transcription switches (Figure 4.S3G). We overlaid the GRO-seq transcriptional states on the Hi-C contact map and find a precise correlation with Hi-C contact domains at 1 kb resolution (Figure 4.2G). This indicates that domains are not formed by some feature of borders, but by the segregation between transcriptional states of neighboring domains. Domains identified by this method are similar in size to compartmental domains identified by high resolution Hi-C (Figure 4.S3H). The small size of domains in *Drosophila* would cause them to appear as one or two bins along the diagonal in the 20 kb resolution matrix that was originally used to identify TADs, which may account for the inaccurate border identification mentioned above. Altogether these data indicate that transcriptional or chromatin state plays a prominent role in 3D chromatin organization at the gene level in *D. melanogaster.* Additionally, compartments are not multi-megabase features of chromatin organization, but are composed of gene-to-gene interactions. Perhaps most surprisingly, compartments and domains do not

represent separate features of 3D chromatin organization in *Drosophila,* as is generally thought to be the case in mammals. Rather, the formation of compartments is responsible for the establishment of all domains in the *Drosophila* genome.

RNAPII Occupancy Inside Domains Affects *Drosophila* Chromatin Organization

Since transcriptional state and domain organization are highly correlated, we tested whether inhibition of transcription affects formation of compartmental domains. Triptolide inhibits transcription initiation and heat shock results in widespread repression of transcription in *Drosophila* (Li et al., 2015). Hi-C heatmaps at 10 kb resolution from triptolide-treated cells display decreased signal inside compartmental domains (Figure 4.3AB). The decrease in domain architecture appears more pronounced in cells subjected to heat shock than triptolide treatment, although both result in transcription silencing of most or all genes (Figure 4.3AB). We therefore examined the levels of RNAPII after each treatment and found that heat shock results in a more pronounced decrease of RNAPII levels than triptolide treatment, consistent with its more substantial effect on compartmental domain interactions (Figure 4.S4AB). Active domains showed a greater decrease in interaction frequency than inactive domains (Figure 4.3C). Triptolide treatment also results in an increase in A-B and B-B contacts, but a decrease in A-A contacts, especially at triptolide sensitive domains (Figure 4.3F). When the activity state of A domains decreases to more closely resemble the activity of B domains, segregation and domain structure of both A and B compartments is reduced. We then examined active domains with a >= 2 fold change in RNAPII ChIP-seq signal across the domain, which we term triptolide sensitive domains. Upon treatment, these domains showed a greater decrease in Hi-C signal than other active domains (Figure 4.3DE), suggesting that RNAPII level is an important factor influencing domain architecture.

Treatment of *Drosophila* embryos during the zygotic genome activation stage with triptolide has been recently shown to affect the structure of domains observed by Hi-C (Hug et al., 2017). We compared the extent of reduction in domain structure observed in nc14 embryos with our data in Kc167 cells. Kc167 cells were treated with 10 μM triptolide for 3 hr while nc8-nc14 embryos were treated with 1.8 μM triptolide for roughly 1.5 hr (Hug et al., 2017; Li et al., 2015). We find that nc14 embryos display a smaller decrease in domain structure than Kc167 cells under these conditions (Figure 4.S4CD). We then examined results from RNAPII ChIP-seq experiments performed in each of the two conditions and found that the extended triptolide treatment in Kc167 cells had a greater effect on RNAPII binding than in nc14 embryos (Figure 4.S4EF). The 3 hr treatment with 10 mM triptolide of Kc167 cells resulted in at least a two-fold change in about 69% of RNAPII peaks, while treatment with 1.8 mM triptolide of nc14 embryos affected only about 29% of RNAPII peaks. Therefore, the greater decrease in domain structure observed in Kc167 cells correlates with a larger reduction in RNAPII occupancy, supporting the conclusion that transcription or RNAPII and/or its associated factors are important for the establishment of compartmental domains in *Drosophila*. The effect of triptolide treatment on chromatin organization correlates with its effect on RNAPII occupancy, although it is possible that triptolide treatment alters more than just RNAPII. To test whether triptolide affects transcription factor occupancy at non-promoter sites, we performed ATAC-seq in triptolide-treated cells and examined non-TSS (± 100 bp) associated subnucleosomal size fragments. We did not see loss of ATAC-seq signal in triptolide sensitive domains (Figure 4.3G, 4.S4G). This implicates RNAPII and associated proteins, rather than factors binding at distal regulatory sequences, as having a prominent role in domain organization.

## Architectural Proteins Act as Insulators in Domain Segregation

It was previously reported that TAD boundaries defined with low resolution Hi-C data were enriched in active chromatin and APBSs (Hou et al., 2012; Sexton et al., 2012; Van Bortle et al.,

2014). This conclusion may be influenced by the imprecise TAD boundary calls obtained using low resolution Hi-C data. To further examine the role of architectural proteins in chromatin organization, we performed HiChIP for CP190. HiChIP for this protein resembles that of RNAPIISer2ph, with most interactions occurring in active compartmental domains (Figure 4.4A, 4.S3IJ). Architectural protein occupancy is closely correlated with transcription (Figure 4.S3K), making it difficult to interpret the significance of this observation. In order to distinguish the relative roles of APBS occupancy and transcriptional state we examined APBSs ranked either by architectural protein occupancy or by transcriptional activity, and used the directionality index as an indicator of border formation (Dixon et al., 2012). APBS occupancy and transcriptional activity both correlate with negative to positive Hi-C directionality switches indicative of domain borders (Figure 4.4B). We next grouped APBSs by their presence near highly or lowly transcribed genes and examined Hi-C directionality. We find that highly transcribed genes have negative to positive changes in Hi-C directionality (i.e. domain borders) regardless of APBS occupancy levels (Figure 4.4C). Conversely, APBSs distant from active gene promoters do not show a distinct change in Hi-C directionality, even when at high occupancy (Figure 4.4D). To more directly test domain border organization at APBSs, we plotted the median Hi-C signal around high occupancy APBSs that are distant from transcribed regions. The results suggest that APBSs by themselves do not form strong domain borders when compared to compartmental interactions (Figure 4.S3L). However, this does not preclude the possibility that APBSs play a role in conjunction with transcription.

Although non-TSS associated APBSs do not show a pronounced correlation with compartmental domain border formation (Figure 4.4D), these proteins are known to insulate enhancer-promoter interactions in transgenic assays (Van Bortle and Corces, 2013). To test the effect of APBSs on interactions between genes, we categorized highly expressed genes located in A compartmental domains (Figure 4.2F far right) by the number of architectural proteins

separating pairs of genes. We found that highly expressed genes interact less frequently with each other if they are separated by high occupancy APBSs (Figure 4.4E). We also examined the effect of APBS occupancy at immediately neighboring active genes. We found that gene neighbors separated by more architectural proteins have lower interaction frequencies between them (Figure 4.4F). Finally, we tested the effects of APBS occupancy on interactions between A compartmental domains and find that distance matched A-A compartmental interactions separated by high occupancy APBSs are lower than those separated by low occupancy APBSs (Figure 4.S3M). These observations suggest that transcription can explain much of chromatin organization based on the clustering of active transcriptional states, but that APBSs, commensurate to the number of proteins present, modulate these interactions.

Gene Expression and the Establishment of Contact Domains in other Eukaryotes

Due to the strong link between transcriptional state and domain organization observed in *Drosophila*, we asked whether we could simulate Hi-C contact domains using transcriptional activity data without any information from 3D chromatin architecture. The simulation creates a pseudo-Hi-C interaction map where the interaction frequency in each bin of the matrix is generated using one-dimensional genomic data (i.e. GRO-seq) to test the ability of one dimensional features to recapitulate the real Hi-C data (see STAR Methods). Using GRO-seq, we set the simulated interaction frequency between any two 5 kb segments proportional to the correlation between the activity scores of the two segments. The result is a simulated interaction map that uses only GRO-seq data to predict Hi-C data (Figure 4.5A bottom right). We found that contact maps simulated by GRO-seq alone could capture domains and compartments with high accuracy (Figure 4.5A). Our simulation assumed that all active genes at the same distance will interact with the same frequency. However, results described above suggest that APBSs can exert an insulation effect between highly expressed genes and active compartments (Figure 4.4). We thus asked whether insulation by architectural proteins could explain some features of

Hi-C contact maps that transcriptional state alone cannot. To simulate this, the interaction frequency between each pair of genomic segments is decreased slightly for each architectural protein ChIP-seq peak bound between them. Simulations using APBS insulation alone recreate the large domains though miss the separation of small active domains into A compartments (Figure 4.5B). We then created a third simulation that combines both the principle of transcriptional state segregation and an interaction decay by APBS insulation. When these two components are combined, we see remarkable recapitulation of actual Hi-C data at 1 kb and 5 kb resolutions (Figure 4.5C, 4.S5A-C). We find that GRO-seq based simulations correlate well with actual Hi-C maps, though APBS occupancy combined with GRO-seq improved the accuracy (Figure 4.5D). Indeed the majority of contact bin interactions in the simulation are within 2-fold of the actual Hi-C data at a range of distances (Figure 4.S5D). The accuracy of the GRO-seq plus APBS simulation at high resolution suggests that transcriptional state in combination with ABPS insulation may explain the compartmental domain structures observed by Hi-C. We next asked how this principle contributes to coarser resolution structures, such as previously identified TADs. When the high-resolution simulation is viewed at 25 kb resolution, it recapitulates previously identified TADs, suggesting that TADs are composed of compartmental domains that are binned together and viewed at a coarser resolution (Figure 4.5E).

The high correlation between the experimental results and the computer simulations suggests that segregation of domains based on transcriptional state can explain a large part of chromatin organization in *Drosophila.* We then postulated that the genomes of other organisms may be organized by these same fundamental principles. According to our hypothesis, domain sizes may vary between organisms depending on the lengths of contiguous active and inactive genomic regions. This may explain why large topological domains are not easily observed in gene dense organisms (Rowley and Corces, 2016). For example, *Arabidopsis thaliana* has a genome size similar to that of *Drosophila melanogaster*, but the two differ drastically in gene

content and gene activity profiles. To compare the distribution of transcriptional states between *Arabidopsis* and *Drosophila*, we plotted transcription levels along a 1 Mb region and saw the existence of large non-transcribed regions in *Drosophila* (Figure 4.5F) but constant transcription levels in *Arabidopsis* (Figure 4.5G). In agreement, *Arabidopsis* Hi-C interaction maps do not show large contact domains at most locations in the genome, a result predicted by our computer simulation (Figure 4.5H). However, when we specifically search for large inactive genomic regions, we then observe large domains that align well with blocks of silenced regions separated by small transcribed regions (Figure 4.5I actual). These compartmental domains are captured by the computer simulation (Figure 4.5I simulated, S5E) indicating that transcriptional states play a critical role in domain formation in *Arabidopsis*, and this principle represents an evolutionarily conserved mechanism controlling 3D chromatin organization.

To further test the correlation between 3D genome organization and gene expression throughout eukaryotes, we examined Hi-C contacts from the protist *P. falciparum*, the fungus *N. crassa*, and the animal *C. elegans*. We searched for large regions with different transcriptional states and found that, in each case, contact domain boundaries appear at transcriptionally inactive-active switches, a feature that is recapitulated in the computer simulation (Figures 4.5J-L, 4.S5F-H). We propose that the differences seen in contact domain sizes between eukaryotic organisms are not due to different principles governing chromatin architecture, but are primarily a result of the size of contiguous active and silenced regions, in combination with the resolution of the Hi-C experiments performed. Furthermore, our ability to simulate Hi-C data at such high resolution based solely on transcription information indicates that transcription is a major contributor to 3D chromatin architecture in many eukaryotes.

Compartmental Domains are Small Structures Underlying TADs in Humans Cells

Results described above suggest that compartments are small fine-scale structures in *Drosophila* and, therefore, we hypothesized that compartments may be also fine-scale structures in human cells. To test this hypothesis, we examined Hi-C data in GM12878 cells for evidence of fine-scale compartmentalization. Hi-C data viewed at 1 Mb resolution depicts large compartments as previously identified (Figure 4.S6A left). We compared this to the 100 kb compartments of Rao et al. 2014 and found that 1 Mb compartments are composed of smaller, alternating A and B compartments. The A/B identity of the 1 Mb compartments merely reflects the proportion of smaller A and B compartments that constitute them (Figure 4.S6A right). This suggests that compartments defined at 1 Mb resolution are the result of coarse binning of interaction maps.

Due to the importance of resolution in proper identification of compartments, we asked whether 100 kb compartments could be resolved into even smaller compartments and whether compartmental domains exist in human cells as they do in *Drosophila*. Figure 4.6A shows a typical example of fine-scale compartmental interactions in GM12878 cells. The central active region (black arrowhead) does not interact with neighboring silenced sequences, even within the same CTCF loop (black circle), but interacts preferentially with other nearby active regions, even when located outside of the CTCF loop (green arrowhead). This fine-scale compartmentalization can be better appreciated in the local Pearson correlation matrix (Figure 4.6A right), but it is not detected by compartment calls at resolutions as low as 100 kb (Figure 4.6A). We therefore sought to call fine-scale compartments in human cells by refining compartment calls at 5 kb resolution. Because compartments were already identified at 100 kb resolution in GM12878 cells (Rao et al., 2014), in lieu of using unsupervised learning methods, we classified 5 kb bins as A or B by their propensity to interact with other A or B regions. First we tested this method of compartment refinement utilizing *Drosophila* data and found that the A-B index matches well with the eigenvector and 1 kb HiChIP compartments (Figure 4.S6B). Next

we used the A-B index to refine compartment calls in human GM12878 cells to detect fine-scale compartments as shown in Figure 4.6A. Comparison with GRO-seq data suggests that these 5 kb-resolution compartments correlate with the transcriptional state of genes, similar to what we saw in *Drosophila* (Figure 4.6A, 4.S6C) and what is generally known about compartments (Lieberman-Aiden et al., 2009). These results support the idea that compartments in human cells are fine-scale structures rather than large Mb-sized regions.

Since we find compartmental domains in human cells, we then explored the relationship between these domains and previously identified TADs (Moore et al., 2015). We examined these TAD calls and found that they identify low resolution domains (Figure 4.S6D). When we examine these structures at different intensity scales we find underlying subdomains (Figure 4.6B, 4.S6D). We noticed that these often correspond to compartment switches inside TADs (Figure 4.6B), suggesting that compartmental domains can occur at scales smaller than TADs in mammalian cells. This also indicates that TADs called at low-resolution are composed of compartmental domains (compare Figure 4.6C and 4.S6D). We examined the prevalence of compartment switches occurring within TADs and find that ~71% of TADs contain more than one compartmental switch (Figure 4.S6E).

TADs have been predominately identified at 40 kb resolution in human cells (Dixon et al., 2012; Moore et al., 2015) and they do not appear to correspond to the compartmental domains seen at higher resolution (Figure 4.6B). To further explore this issue, we called TADs in GM12878 cells utilizing the directionality index (Dixon et al., 2012) and the 1 kb resolution contact map (Rao et al., 2014). These TAD calls better define the underlying domain structures (Figure 4.6C). We noted that CTCF loops often coincide with compartmental switches (Figure 4.6C) and questioned whether CTCF or the underlying compartmental switch determines the formation of boundaries between domains. To test this, we selected CTCF loop anchors located at least 50 kb away from a compartmental domain switch and examined the boundary score around these

sites. We found that these CTCF loops still form boundaries (Figure 4.S6F). Interestingly, not all domains show the presence of a loop at the domain corner and correspond instead with the compartmental pattern (Figure 4.6C). To confirm the existence of compartmental domains in the human genome, we examined compartmental switches that were at least 50 kb away from a CTCF loop anchor. These compartmental switches correspond well with the formation of domain boundaries without the need of a CTCF loop (Figure 4.6D). We then determined the proportion of TAD borders that can be explained by compartments, CTCF loops, or both. We found that CTCF loops can explain many TADs, but a large portion of borders occur at compartmental switches (Figure 4.S6G). Additionally, as we noted above (Figure 4.6C) many TAD borders correspond to both a compartmental switch and a CTCF loop anchor, suggesting a correlation between the two (Figure 4.S6G). It should be noted that we found 1,939 TAD borders (23%) that do not correspond to CTCF loop anchors or to compartmental switches and it is unclear which features contribute to the formation of these borders. Altogether these observations suggest that TADs defined based on a directionality index are composed of CTCF loops and/or fine-scale compartments. While CTCF is an important player in controlling 3D chromatin organization in mammalian cells, compartmentalization by transcriptional states likely plays a similarly important role.

## CTCF and Compartments Organize Chromatin into Domains in Human Cells

Results described above suggest that compartmental domains often represent structures smaller than traditionally defined TADs in human cells and in other eukaryotes. This indicates a conserved principle of chromatin organization by the segregation of active and inactive transcription. To further understand the relationship between transcription and known features of 3D chromatin organization such as CTCF loops, we classified the genome into transcriptionally active and inactive segments by their GRO-seq signal using a hidden Markov model in GM12878 cells. We find that transcriptionally active regions form domains with a

structure distinct from that of CTCF loops i.e. lack of an intense signal spot at the corner of the domain (Rao et al., 2014), which is similar to that of domains found in *Drosophila* (Figures 4.7A and 4.2B). As an example, Figure 4.7B shows a region of chromosome 5 containing a domain formed by multiple interactions among transcribed regions. In addition, a CTCF loop is formed between two CTCF sites present inside and outside of this domain (Figure 4.7B circle). The borders of this domain do not correspond to CTCF motifs in convergent orientation, but instead correspond to switches in transcriptional activity (Figure 4.7B GRO-seq). Therefore, since some contact domains can be explained by transcription rather than by the formation of loops between CTCF sites, we hypothesized that these domains should be sensitive to changes in transcription. We tested this hypothesis by finding regions with differential transcription between cell types. In one example, transcription of the PBX1 gene occurs in IMR90, K562, NHEK, and HeLa cells and each has a corresponding domain structure separating this site from the neighboring inactive regions (Figure 4.7C). IMR90 appears to have the strongest expression and correspondingly shows the strongest compartmental domain pattern. Additionally, transcription is lost in GM12878 cells, which correlates with a loss of the compartmental domain (Figure 4.7C). In a second example, transcription occurs in GM12878 and a compartmental domain is formed, while both the domain and transcription are lost in the other cell types (Figure 4.S7A). We tested the validity of these observations genome-wide by taking the median distance normalized interaction signal around regions that are transcribed in IMR90 but not in GM12878 cells. We found that differentially transcribed regions show distinct differential contact domains between the two cell types (Figure 4.7D top). We also tested regions transcribed in GM12878 but not in IMR90 and found that differentially transcribed regions in GM12878 form contact domains structures that are not present in IMR90 (Figures 4.7D bottom).

The finding that compartmental domains are distinct from CTCF loops predicts that long stretches of the genome that lack transcription, such as gene deserts, should display only CTCF

loops. We examined Hi-C data from GM12878 cells and found that gene deserts contain CTCF loops and their corresponding loop domains. However, domain segregation in gene deserts does not appear as strong as in neighboring regions that have both CTCF loops and compartmental domains (Figure 4.7E and 4.S7B).

Motivated by the apparent applicability of the fine-scale compartmentalization model to the human genome we then examined whether chromatin contact maps obtained from Hi-C experiments could be predicted using computer simulations as in other eukaryotes. Figure 4.7E shows an example of the Hi-C contact map in a region containing CTCF loops, predicted transcriptional domains, and evident compartmental interactions. First we recreated features of CTCF loops by creating a simulated Hi-C interaction map where the intensity of the CTCF loop is used to create the Hi-C peak, line of interactions from CTCF anchors, and the enriched interactions comprising the underlying domain (see STAR Methods). This map reproduced some small domains well, but could not account for interactions larger than the CTCF loops themselves (Figure 4.7F). Next, we modeled Hi-C contacts based solely on the correlation of GRO-seq signal, as we previously did for *Drosophila* and other eukaryotes. Simulations using only transcription information produce compartmental domains that match many fine-scale compartments and domain-like structures observed in Hi-C heatmaps, but miss CTCF loop domains. This is particularly evident in large inactive regions of the genome (Figure 4.7G). We then combined CTCF and transcription based simulations to produce a map in which both CTCF and transcription contributed independently to contact signals. The accuracy of the resulting map indicates that both transcription and CTCF looping are important components of chromatin architecture in human cells (Figure 4.7H). Overall, these results suggest that the fine-scale compartmentalization principle underlying *Drosophila* contact domain formation is also operational in human cell nuclei, but that CTCF loops and their resulting domains are not conserved features between the two organisms. Changes in transcriptional state can explain the

establishment of compartments and compartmental domains, whereas CTCF-mediated loops account for the rest of the contact domains observed in human cells. Therefore, transcriptional activity is a major predictor of chromatin organization throughout Eukarya, with CTCF playing a prominent role in mammals.

## Discussion

Results presented here suggest that compartments and contact domains not mediated by CTCF loops are structurally and functionally equivalent, and arise from the segregation of the genome into active and silent regions. These compartmental domains likely represent a basic and ancient form of 3D chromatin organization in eukaryotes. In this model of nuclear architecture, actively transcribed genes form mini-domains that interact more frequently with other active genes. Clusters of active genes without large transcriptionally silent spaces between them form larger, multi-gene domains. Domains of similar transcriptional activity interact to form the characteristic plaid pattern of compartments. Thus, the compartmentalization of the genome by transcriptional state is responsible for the formation of both long-range compartments and local compartmental domains. This appears to be the main mechanisms of 3D organization for organisms that lack architectural proteins such as CTCF. *Drosophila* lacks motif-oriented CTCF looping, which is likely key to its function in mammals, and compartmental domains explain most visible chromatin organization observed by Hi-C. In humans, these compartmental domains exist alongside CTCF loops to constitute structures previously defined as TADs. In *Drosophila*, a large effort has gone into identifying components of TAD borders. A major problem with this approach is that the results depend on an often-inaccurate border identification due to the low resolution of the Hi-C data employed in the analyses, and ignores features within the domain. This has led to the conclusion that TAD borders are enriched for architectural proteins, active chromatin, or transcription/RNAPII (Hou et al., 2012; Hug et al., 2017; Sexton et al., 2012; Ulianov et al., 2016; Van Bortle et al., 2014). Results presented here

suggest that regions where these features are enriched represent small domains rather than domain borders.

Our results appear to conflict with current thinking suggesting that TADs are invariant between mammalian cell types (Dixon et al., 2012). However, the degree of variability in TAD calls between cell lines, for example 54% conservation between mESCs and brain cortex in mouse and 65% between hESC and IMR90 cells in humans (Dixon et al., 2012), is in line with differences in transcription and CTCF distribution among different cell types. The essential contribution of transcriptional state to the 3D architecture of the genome is also supported by observations suggesting that TAD organization is altered during the heat shock response (Li et al., 2015). Our results show that alteration of transcription or correlated factors such as RNAPII occupancy, using either inhibitors or heat shock, results in changes of compartmental domains. Furthermore, differential gene expression between multiple cell types results in the formation of distinct gene-level compartmental domains. This supports the idea that TADs, which are in part formed by these compartmental domains, should be different when comparing cell types with distinct transcription patterns. Recently published studies have examined the role of CTCF in the formation of loops and TADs using an auxin-mediated degradation system in mammals (Kubo et al., 2017; Nora et al., 2017). The loss of CTCF domains and maintenance of compartments seen after CTCF degradation fits with our model. Compartmental domains can explain why TAD-like structures can still be seen after CTCF depletion.

Interestingly, compartmental domains are found in representatives across Eukarya and the relative sizes of active and inactive segments can explain the differences in domain sizes found in these organisms. Our findings invite the question of when animal genomes first acquired oriented CTCF loops. One possibility is that an ancient Bilaterian ancestor possessed oriented looping CTCF whose function was later lost in *D. melanogaster* and *C. elegans*. It has been shown that CTCF motifs are oriented in accordance with topological domain borders in both *D.*

*rerio* and *S. purpuratus*, suggesting that CTCF acquired this role early in the Deuterostome ancestor (Gómez-Marín et al., 2015). Although there is a clear correspondence between transcriptional activity, compartments, and domain formation, the question of what establishes and/or maintains compartmental domains remains unclear. It has been recently suggested that TADs are still established after inhibition of transcription in *Drosophila* embryos using low concentrations of triptolide. However, it is possible that transcription of most genes in the genome was not affected under these conditions, since RNAPIISer5ph remains bound to promoter regions under this treatment (Hug et al., 2017). It is also possible that the presence of RNAPII and other associated proteins, rather than transcription itself, is responsible for the establishment of compartmental domains, since compartmental interactions appear to correlate more closely with occupancy of RNAPII at promoter regions. A role for RNAPII and/or associated proteins in the establishment of compartmental domains is also supported by HiChIP and ChIA-PET results, which identify RNAPII-mediated interactions throughout A compartmental domains. This idea is further supported by analysis of Hi-C data in mouse sperm, which is transcriptionally silent but contains RNAPII and active or silent histone modifications, but shows a similar compartmental organization as embryonic stem cells (Jung et al., 2017).

Segregation of the genome into gene-sized active and inactive components explains structural aspects of chromatin organization in all organisms analyzed to date. Proximal gene domains co-associate to form domains that further interact to form compartments. Together with point-to-point interactions mediated by CTCF, these short and long-range interactions give rise to TADs. Altogether, the correlation between transcriptional state and compartmental domains suggests a fundamental and conserved principle of chromatin organization across Eukarya.

Author Contributions

M.J.R., M.H.N, and V.G.C. designed the project and wrote the manuscript. X.L. performed ATAC-seq. X.L., K.H., M.A-K., and I.S.M.R. performed HiChIP experiments. P.W. and Y.R. performed ChIA-PET experiments. M.J.R. and M.H.N. performed all other analyses.

Acknowledgements

We would like to thank the Genomic Services Lab at the HudsonAlpha Institute for Biotechnology, and specially Drs. Angela Jones and Terri Pointer, for their help in performing Illumina sequencing of samples. This work was supported by U.S. Public Health Service Award R01 GM035463 (V.G.C.) and the Ruth L. Kirschstein National Research Service Award F32 GM113570 (M.J.R.) from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

# References

Andersson, R., Refsing Andersen, P., Valen, E., Core, L.J., Bornholdt, J., Boyd, M., Heick Jensen, T., and Sandelin, A. (2014). Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat. Commun. *5*, 5336.

Ay, F., Bunnik, E.M., Varoquaux, N., Bol, S.M., Prudhomme, J., Vert, J.-P., Noble, W.S., and Le Roch, K.G. (2014). Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Res. *24*, 974–988.

Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J., et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat. Genet. *48*, 1193–1203.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

Core, L.J., Waterfall, J.J., Gilchrist, D.A., Fargo, D.C., Kwak, H., Adelman, K., and Lis, J.T. (2012). Defining the status of RNA polymerase at promoters. Cell Rep. *2*, 1025–1035.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. *46*, 1311–1320.

Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J., and Meyer, B.J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature *523*, 240–244.

Cubeñas-Potts, C., Rowley, M.J., Lyu, X., Li, G., Lei, E.P., and Corces, V.G. (2016). Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. Nucleic Acids Res. *45*, 1714–1730.

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

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016b). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst. *3*, 99–101.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016a). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. *3*, 95–98.

El-Sharnouby, S., Fischer, B., Magbanua, J.P., Umans, B., Flower, R., Choo, S.W., Russell, S., and White, R. (2017). Regions of very low H3K27me3 partition the Drosophila genome into topological domains. PloS One *12*, e0172725.

Galazka, J.M., Klocko, A.D., Uesaka, M., Honda, S., Selker, E.U., and Freitag, M. (2016). Neurospora chromosomes are organized by blocks of importin alpha-dependent heterochromatin that are largely independent of H3K9me3. Genome Res. *26*, 1069–1080.

Goh, Y., Fullwood, M.J., Poh, H.M., Peh, S.Q., Ong, C.T., Zhang, J., Ruan, X., and Ruan, Y. (2012). Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for Mapping Chromatin Interactions and Understanding Transcription Regulation. J. Vis. Exp. e3770.

Gómez-Marín, C., Tena, J.J., Acemel, R.D., López-Mayorga, M., Naranjo, S., de la Calle-Mustienes, E., Maeso, I., Beccari, L., Aneas, I., Vielmas, E., et al. (2015). Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc. Natl. Acad. Sci. *112*, 7542–7547.

Gonçalves, A.P., Hall, C., Kowbel, D.J., Glass, N.L., and Videira, A. (2014). CZT-1 is a novel transcription factor controlling cell death and natural drug resistance in Neurospora crassa. G3 Bethesda Md *4*, 1091–1102.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

He, C., Zhang, M.Q., and Wang, X. (2015). MICC: an R package for identifying chromatin interactions from ChIA-PET data. Bioinforma. Oxf. Engl. *31*, 3832–3834.

Hillier, L.W., Reinke, V., Green, P., Hirst, M., Marra, M.A., and Waterston, R.H. (2009). Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. Genome Res. *19*, 657–666.

Hou, C., Li, L., Qin, Z.S., and Corces, V.G. (2012). Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. Mol. Cell *48*, 471–484.

Hug, C.B., Grimaldi, A.G., Kruse, K., and Vaquerizas, J.M. (2017). Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. Cell *169*, 216–228.e19.

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

Jung, Y.H., Sauria, M.E.G., Lyu, X., Cheema, M.S., Ausio, J., Taylor, J., and Corces, V.G. (2017). Chromatin States in Mouse Sperm Correlate with Embryonic and Adult Regulatory Landscapes. Cell Rep. *18*, 1366–1382.

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive variation in chromatin states across humans. Science *342*, 750–752.

Kensche, P.R., Hoeijmakers, W.A.M., Toenhake, C.G., Bras, M., Chappell, L., Berriman, M., and Bártfai, R. (2016). The nucleosome landscape of Plasmodium falciparum reveals chromatin architecture and dynamics of regulatory sequences. Nucleic Acids Res. *44*, 2110–2124.

Kubo, N., Ishii, H., Gorkin, D., Meitinger, F., Xiong, X., Fang, R., Liu, T., Ye, Z., Li, B., Dixon, J., et al. (2017). Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950–953.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Li, L., Lyu, X., Hou, C., Takenaka, N., Nguyen, H.Q., Ong, C.-T., Cubeñas-Potts, C., Hu, M., Lei, E.P., Bosco, G., et al. (2015). Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. Mol. Cell *58*, 216–231.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science *326*, 289–293.

Moore, B.L., Aitken, S., and Semple, C.A. (2015). Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. Genome Biol. *16*, 110.

Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., and Chang, H.Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat. Methods *13*, 919–922.

Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930–944.

Pai, C.-Y., Lei, E.P., Ghosh, D., and Corces, V.G. (2004). The centrosomal protein CP190 is a component of the gypsy chromatin insulator. Mol. Cell *16*, 737–748.

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

Rowley, M.J., and Corces, V.G. (2016). The three-dimensional genome: principles and roles of long-distance interactions. Curr. Opin. Cell Biol. *40*, 8–14.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell *148*, 458–472.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics *15*, 284.

Swaminathan, J., Baxter, E.M., and Corces, V.G. (2005). The role of histone H2Av variant replacement and histone H4 acetylation in the establishment of Drosophila heterochromatin. Genes Dev. *19*, 65–76.

Ulianov, S.V., Khrameeva, E.E., Gavrilov, A.A., Flyamer, I.M., Kos, P., Mikhaleva, E.A., Penin, A.A., Logacheva, M.D., Imakaev, M.V., Chertovich, A., et al. (2016). Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. Genome Res. *26*, 70–84.

Van Bortle, K., and Corces, V.G. (2013). The role of chromatin insulators in nuclear architecture and genome function. Curr. Opin. Genet. Dev. *23*, 212–218.

Van Bortle, K., Ramos, E., Takenaka, N., Yang, J., Wahi, J.E., and Corces, V.G. (2012). Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. Genome Res. *22*, 2176–2187.

Van Bortle, K., Nichols, M.H., Li, L., Ong, C.-T., Takenaka, N., Qin, Z.S., and Corces, V.G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biol. *15*, R82.

Wang, C., Liu, C., Roqueiro, D., Grimm, D., Schwab, R., Becker, C., Lanz, C., and Weigel, D. (2015). Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. Genome Res. *25*, 246–256.

Wirbelauer, C., Bell, O., and Schübeler, D. (2005). Variant histone H3.3 is deposited at sites of nucleosomal displacement throughout transcribed genes while active histone modifications show a promoter-proximal bias. Genes Dev. *19*, 1761–1766.

Yang, J., Sung, E., Donlin-Asp, P.G., and Corces, V.G. (2013). A subset of Drosophila Myc sites remain associated with mitotic chromosomes colocalized with insulator proteins. Nat. Commun. *4*, 1464.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhu, Y., Rowley, M.J., Böhmdorfer, G., and Wierzbicki, A.T. (2013). A SWI/SNF chromatin-remodeling complex acts in noncoding RNA-mediated transcriptional silencing. Mol. Cell *49*, 298–309.

# Figures

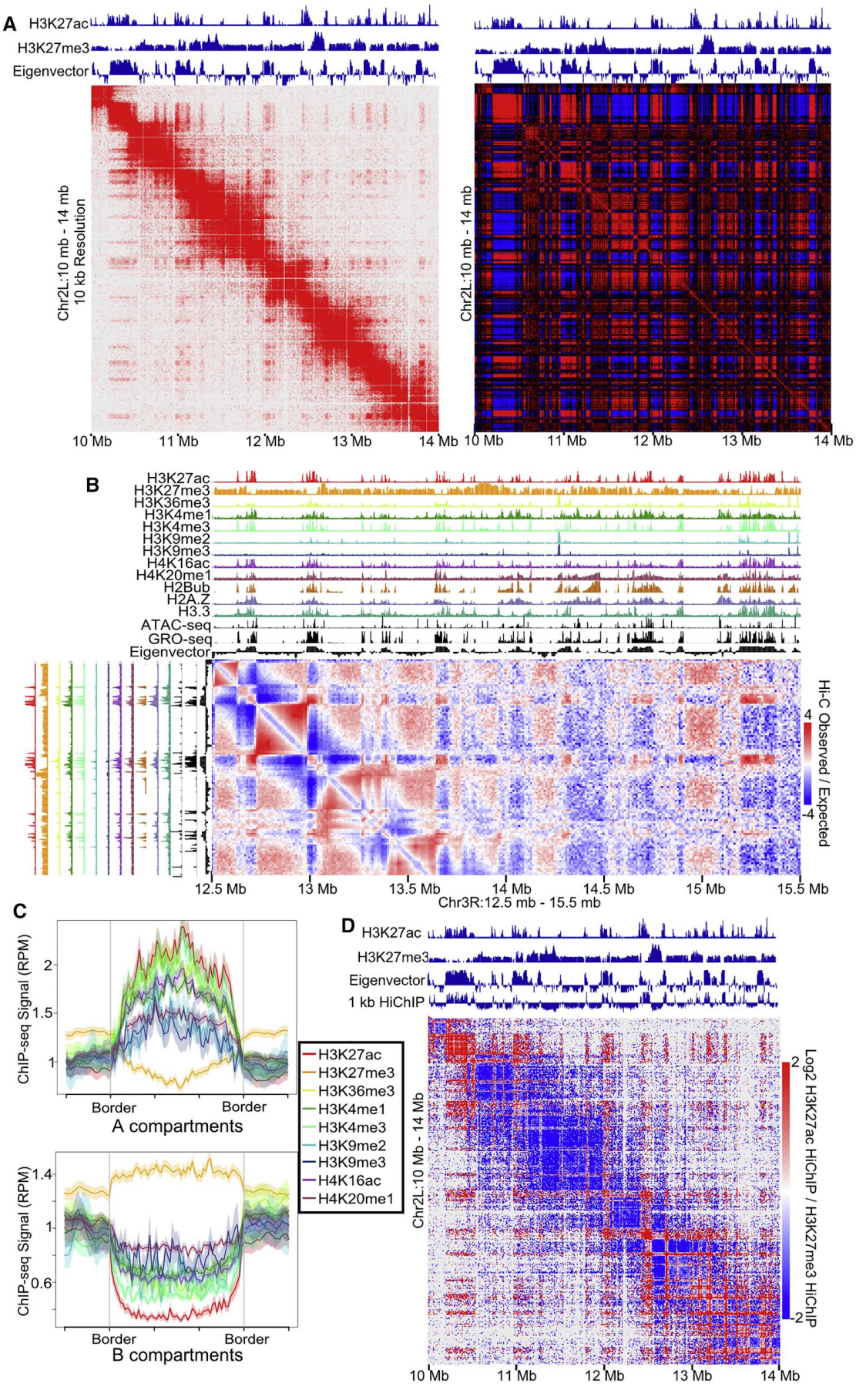

A
H3K27ac
H3K27me3
Eigenvector
Chr2L:10 mb - 14 mb
10 kb Resolution
10 Mb
11 Mb
12 Mb
13 Mb
14 Mb
B
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K9me2
H3K9me3
H4K16ac
H4K20me1
H2Bub
H2A.Z
H3.3
ATAC-seq
GRO-seq
Eigenvector
Hi-C Observed / Expected
4
-4
12.5 Mb
13 Mb
13.5 Mb
14 Mb
14.5 Mb
15 Mb
15.5 Mb
Chr3R:12.5 mb - 15.5 mb
C
ChIP-seq Signal (RPM)
Border
Border
A compartments
B compartments
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K9me2
H3K9me3
H4K16ac
H4K20me1
D
H3K27ac
H3K27me3
Eigenvector
1 kb HiChIP
Chr2L:10 Mb - 14 Mb
Log2 H3K27ac HiChIP / H3K27me3 HiChIP
2
-2

Figure 4.1. *Drosophila* has Fine-Scale Compartments

A. Left: Normalized Hi-C map of Kc167 cells at 10 kb resolution. Right: Pearson Correlation matrix of Hi-C. The eigenvector and H3K27ac and H3K27me3 ChIP-seq are above the Hi-C plot.

B. ChIP-seq for 12 different histone modifications, ATAC-seq, and GRO-seq compared to the Hi-C eigenvector. A slice of the distance normalized Hi-C matrix (observed/expected) is shown corresponding to Chr3R:12.5 Mb – 15.5 Mb (horizontal) and Chr3R:12.5 Mb-13.5 Mb (vertical).

C. Active and inactive chromatin correspond to A and B compartments. Average histone modification profiles over A and B compartments. Color coding of ChIP-seq for histone modifications/variants is indicated.

D. Compartmental interactions defined by HiChIP. Contact map showing differential contacts for H3K27ac vs H3K27me3 HiChIP visualized by Juicebox.

See also Figures 4.S1 and Table 4.S1-S2.

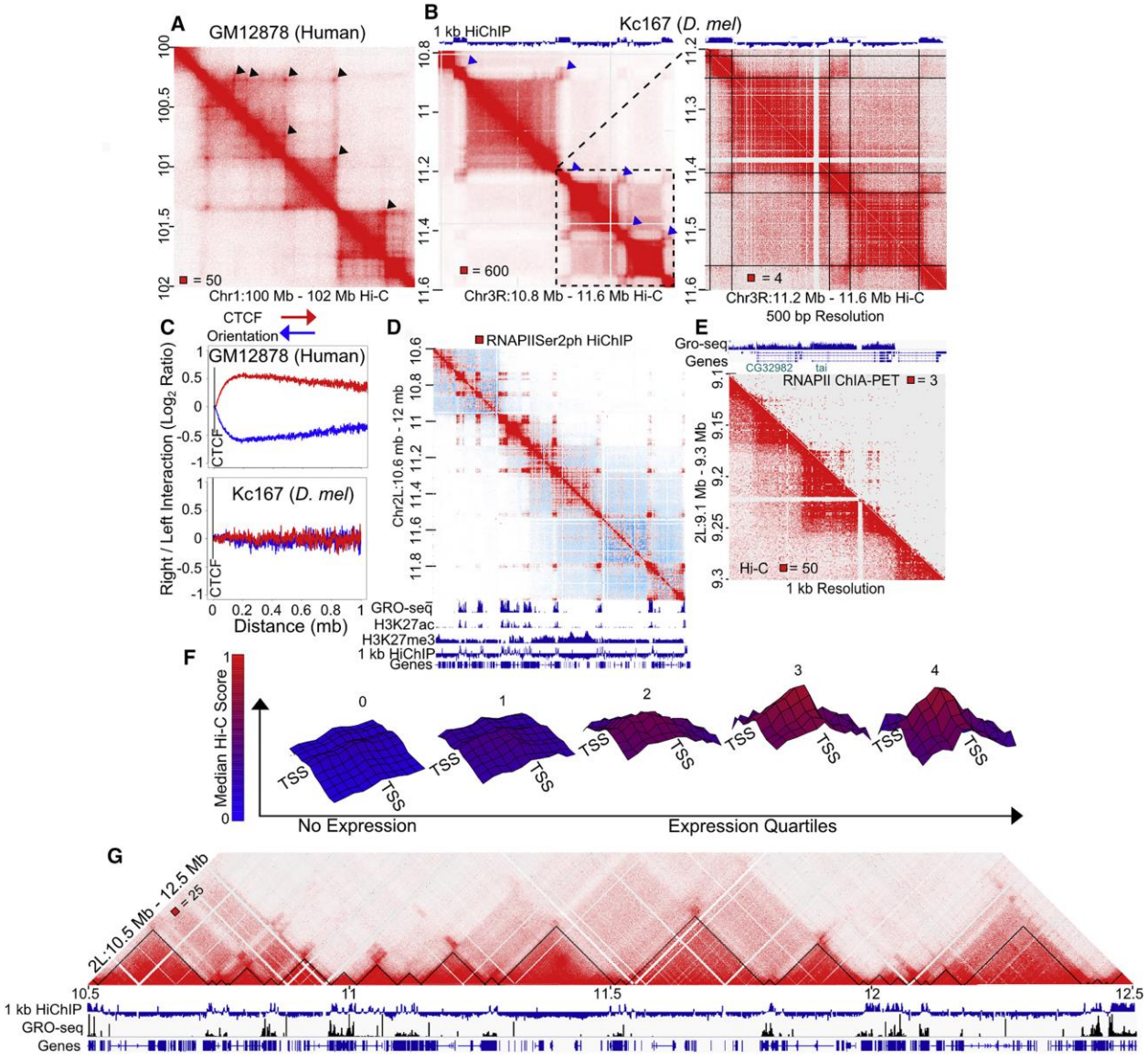



Figure 4.2 Compartments Explain Domain Organization in *Drosophila*

A. Contact domains in human cells show enriched interaction signal between borders (arrowheads). Normalized Hi-C map of GM12878 cells at 5 kb resolution.

B. Contact domains in *Drosophila* do not show enriched interaction signal between borders (arrowheads). Normalized Hi-C map of Kc167 cells at 5 kb resolution (left) and 500 bp

resolution (right). The A/B compartmental interactions computed by H3K27ac vs H3K27me3 HiChIP are shown above. Lines indicate borders.

C. Human CTCF motif orientation has a directional bias, while *Drosophila* does not. Total interactions as $\log_2$ ratio of right/left reads for each distance on right (red) or left (blue) oriented bound CTCF motifs in GM12878 cells (top) or Kc167 cells (bottom).

D. HiChIP for phosphorylated RNAPIISer2 captures active compartments. Raw HiChIP signal for phosphorylated RNAPIISer2 (red) overlaying Hi-C signal (blue). Gene annotations, GRO-seq, H3K27ac, and H3K27me3 ChIP-seq are shown below. 1 kb HiChIP indicates H3K27ac/H3K27me3 HiChIP compartmental interaction preference.

E. Individual genes can form mini-domains. RNAPII ChIA-PET signal in 1 kb bins (top right). Hi-C signal in 1 kb bins (bottom left). GRO-seq and gene annotations are shown above.

F. Distance normalized Hi-C signal at 1 kb resolution is plotted between distinct transcription start sites (TSSs) within the same compartment. Height and color (blue to red) correspond to the relative median observed/expected Hi-C signal. Nodes indicate 1 kb windows from -5 kb to +5 kb surrounding the TSS. Expression level defined by no GRO-seq signal (No Expression) and quartiles of GRO-seq signal. p-value < .05 for each center point (Wilcoxon test compared to no GRO-seq).

G. Transcriptional states correspond to Hi-C domains. Transcriptional state domains identified by GRO-seq (black triangles) overlaying Kc167 Hi-C at 1 kb resolution. GRO-seq and gene annotations are shown below. 1 kb HiChIP indicates compartmental interaction preference.

See also Figure 4.S2-S3 and Table 4.S3.

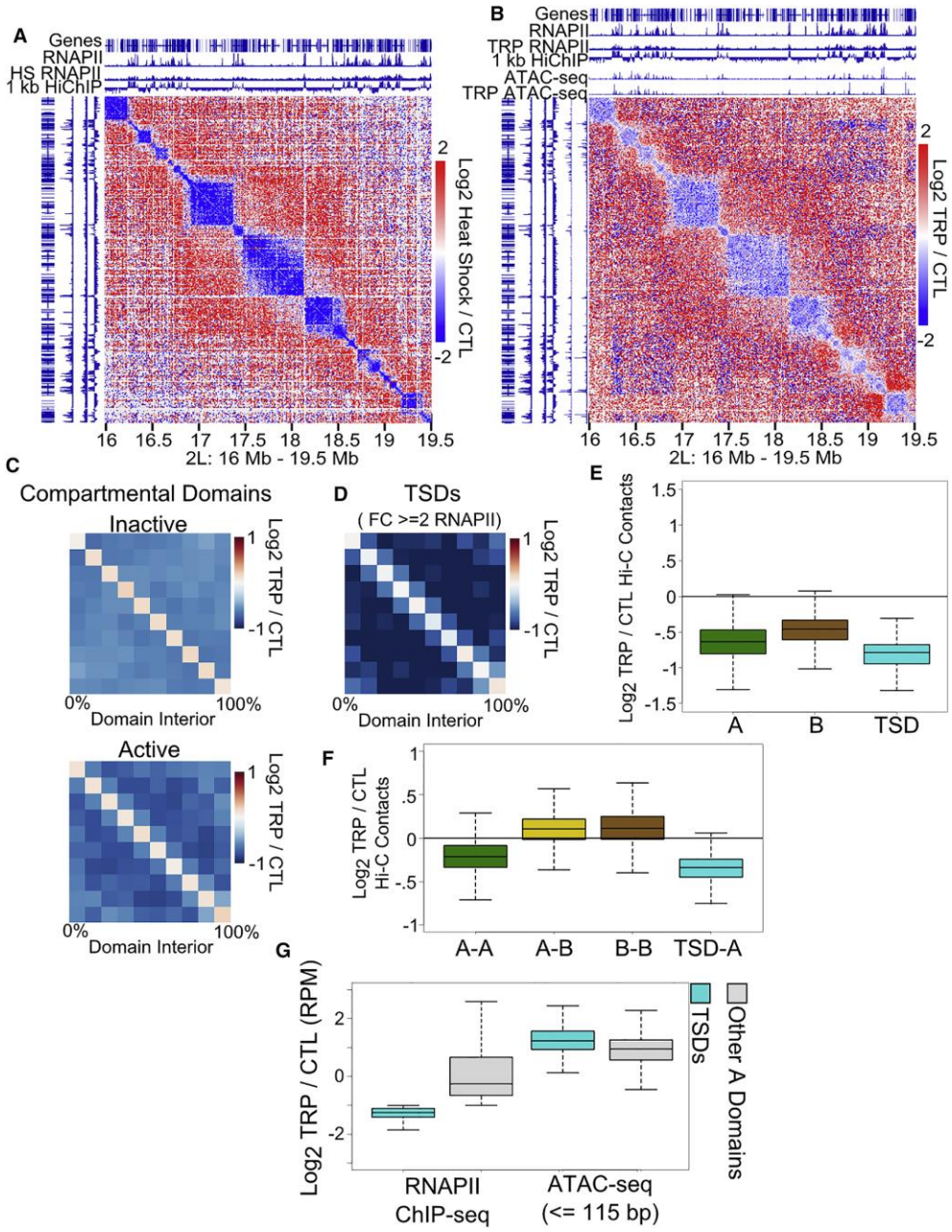



Figure 4.3 RNAPII Depletion Alters *Drosophila* Chromatin Organization

A. Heat shock decreases domain formation. Hi-C heatmap of $\log_2$ ratio of heat shocked to control cells (CTL). Gene annotations, control and heat shocked RNAPII ChIP-seq signal are shown above. 1 kb HiChIP indicates compartmental interaction preference.

B. Inhibition of transcription decreases domain formation. Hi-C heatmap of $log_2$ ratio of triptolide treated (TRP) to control cells (CTL). Gene annotations, control and triptiolide treated RNAPII ChIP-seq signal are shown above. 1 kb HiChIP indicates compartmental interaction preference.

C. Inhibiting transcription decreases contacts in A compartmental domains. Hi-C median metaplot comparing contacts in A and B domains in triptolide treated (TRP) vs control cells (CTL).

D. Hi-C median metaplot A compartmental domains with large decreases in RNAPII after triptolide treatment; i.e. triptolide sensitive domains (TSDs).

E. Decreases in intra-domain contacts in A and B compartments and in triptolide sensitive domains (TSD) after triptolide treatment. Boxes depict median and interquartile range.

F. Ratio of inter-compartmental contact counts in triptolide (TRP) vs control (CTL) treated cells. Boxes depict median and interquartile range.

G. Ratio of RNAPII ChIP-seq or ATAC-seq signal in triptolide sensitive domains (TSDs) or other A compartmental domains (nonTSDs). Boxes depict median and interquartile range.

See also Figure 4.S4.

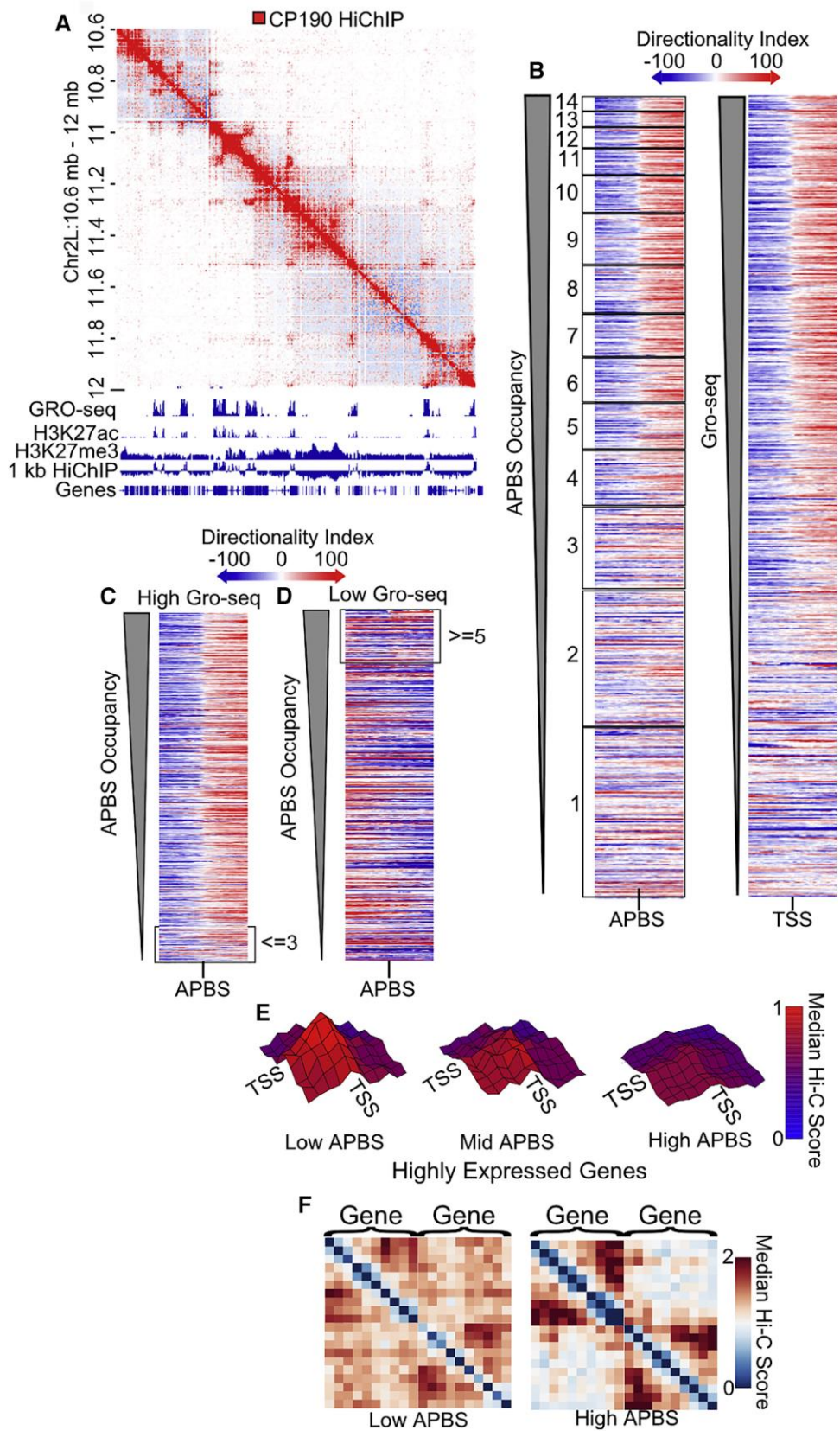

A
CP190 HiChIP
Chr2L:10.6 mb - 12 mb
10.6
10.8
11
11.2
11.4
11.6
11.8
12
GRO-seq
H3K27ac
H3K27me3
1 kb HiChIP
Genes
B
Directionality Index
-100 0 100
APBS Occupancy
Gro-seq
14
13
12
11
10
9
8
7
6
5
4
3
2
1
APBS
TSS
C
High Gro-seq
D
Low Gro-seq
Directionality Index
-100 0 100
APBS Occupancy
>=5
<=3
APBS
APBS
E
TSS
TSS
TSS
TSS
TSS
TSS
Low APBS
Mid APBS
High APBS
Median Hi-C Score
1
0
Highly Expressed Genes
F
Gene
Gene
Gene
Gene
Low APBS
High APBS
Median Hi-C Score
2
0

Figure 4.4 Architectural Proteins Insulate Gene-to-Gene Interactions

A. HiChIP for CP190 captures active compartments. HiChIP signal for CP190 (red) overlaying Hi-C signal (blue). Gene annotations, GRO-seq, H3K27ac, and H3K27me3 ChIP-seq are shown below. 1 kb HiChIP indicates compartmental interaction preference.

B. Heatmaps of Hi-C directionality anchored and ordered by APBS occupancy (left) or GRO-seq signal (right) show switches in directionality (blue to red).

C. Heatmap of APBSs within 250 bp of a highly expressed TSS ordered by APBS occupancy. Low occupancy sites (<= 3 proteins bound) are indicated for comparison with Figure 4.4B.

D. Heatmap of APBSs at least 20 kb away from a highly expressed gene ordered by APBS occupancy. High occupancy sites (>= 5 proteins bound) are indicated for comparison with Figure 4.4B.

E. Distance normalized Hi-C signal at 1 kb resolution is plotted between distinct transcription start sites (TSSs) from the top two GRO-seq quartiles. Low, mid, and high APBSs are defined as the maximum APBS cluster site between genes divided into those containing below 5, 5-8, and above 8 architectural proteins, respectively. Height and color (blue to red) correspond to the relative median observed/expected Hi-C signal. Vertices indicate 1 kb windows from -5 kb to +5 kb surrounding the TSS. p-value < .05 for center point of low APBS compared to high APBS (Wilcoxon text).

F. Neighboring genes are insulated by APBSs. Hi-C metaplot of highly expressed neighboring genes separated by low and high occupancy APBSs.

See also Figure 4.S3 and Table 4.S3.

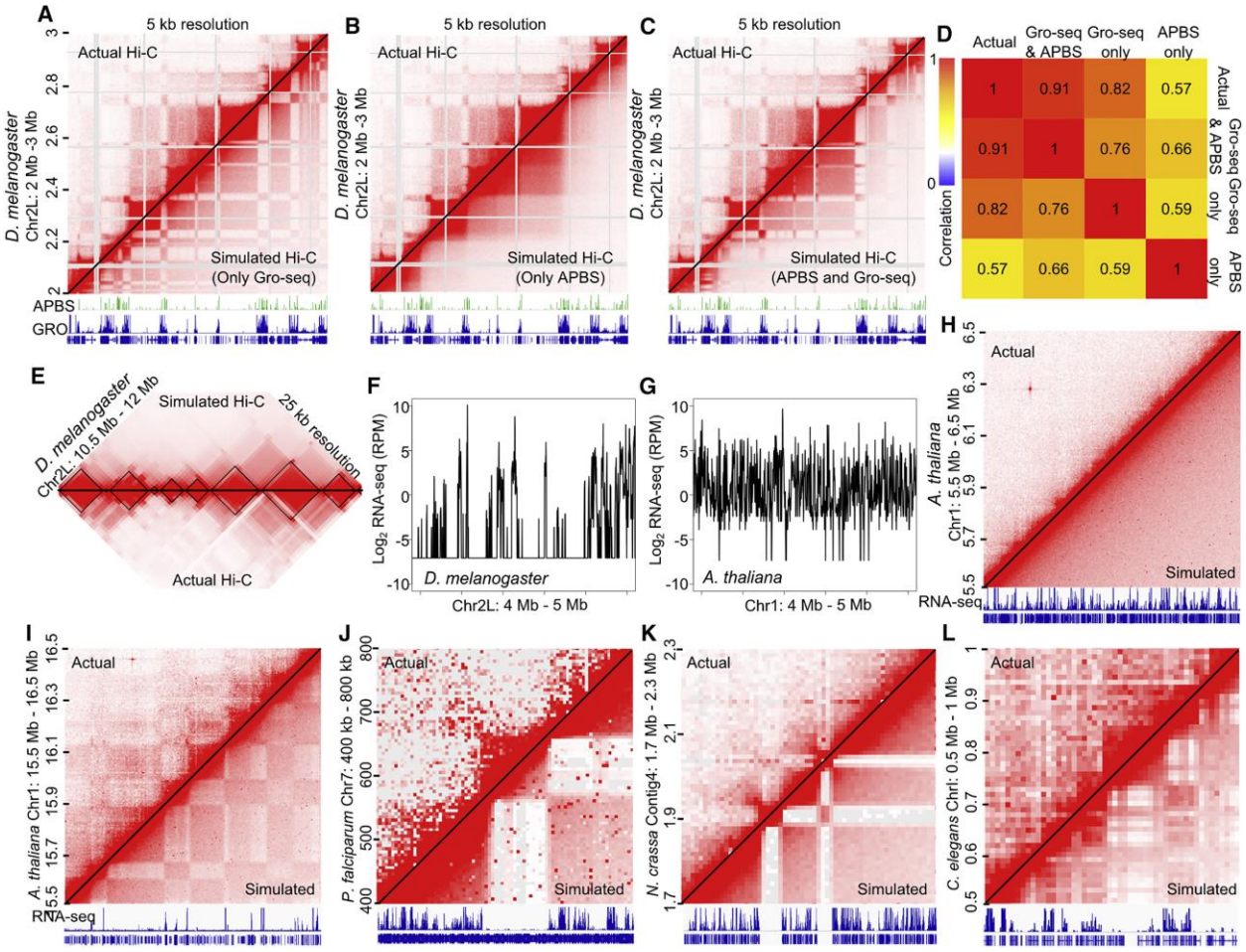



Figure 4.5 Transcriptional States Explain 3D Chromatin Interactions throughout Eukarya

A. Transcription based simulated contact maps predict Hi-C structures. Contact heatmaps at 5 kb resolution using actual Hi-C data (left) and simulated data based on GRO-seq signal only (right). Repetitive/non-mappable regions are shaded grey. Shown below are APBS occupancy counts, GRO-seq, and gene annotations.

B. APBS-based simulated contact maps do not fully explain Hi-C heatmaps. Contact heatmaps at 5 kb resolution using actual Hi-C data (left) and simulated data based on APBS occupancy only (right). Repetitive/non-mappable regions are shaded grey. Shown below are APBS occupancy counts, GRO-seq, gene annotations.

C. GRO-seq and APBS-based simulated contact maps recapitulate domains and compartments in *Drosophila melanogaster*. Contact heatmaps at 5 kb resolution using actual Hi-C data (left) and simulated data based on GRO-seq and APBS occupancy (right). Repetitive/non-mappable regions are shaded grey. Shown below are APBS occupancy counts, GRO-seq, and gene annotations.

D. Spearman correlation of 5 kb bins of actual Hi-C with simulated Hi-C incorporating APBS occupancy, GRO-seq signal, or both.

E. Simulated contacts recapitulate small and large structures. Actual Hi-C (bottom) compared to simulated data (top). TADs are shown in black.

F. *Drosophila* expression varies sharply throughout the genome. $Log_2$ RNA-seq profile of a 1 Mb region in *Drosophila melanogaster.*

G. *Arabidopsis* expression is linearly constant throughout the genome. $Log_2$ RNA-seq profile of a 1 Mb region in *Arabidopsis thaliana*.

H. *Arabidopsis* expression profile contributes to lack of visible domain architecture. Contact heatmaps at 10 kb resolution using actual Hi-C data (left) and simulated data based on RNA-seq data (right). RNA-seq and gene annotations are shown below.

I-L. Large inactive regions form domain structures throughout Eukarya. Contact heatmaps at 10 kb resolution using actual Hi-C data (left) and simulated data based on RNA-seq data (right). RNA-seq and gene annotations are shown below. Sections of the genome with large inactive regions were selected for *A. thaliana* (I), *P. falciparum* (J), *N. crassa* (K), and *C. elegans* (L).

See also Figure 4.S5.

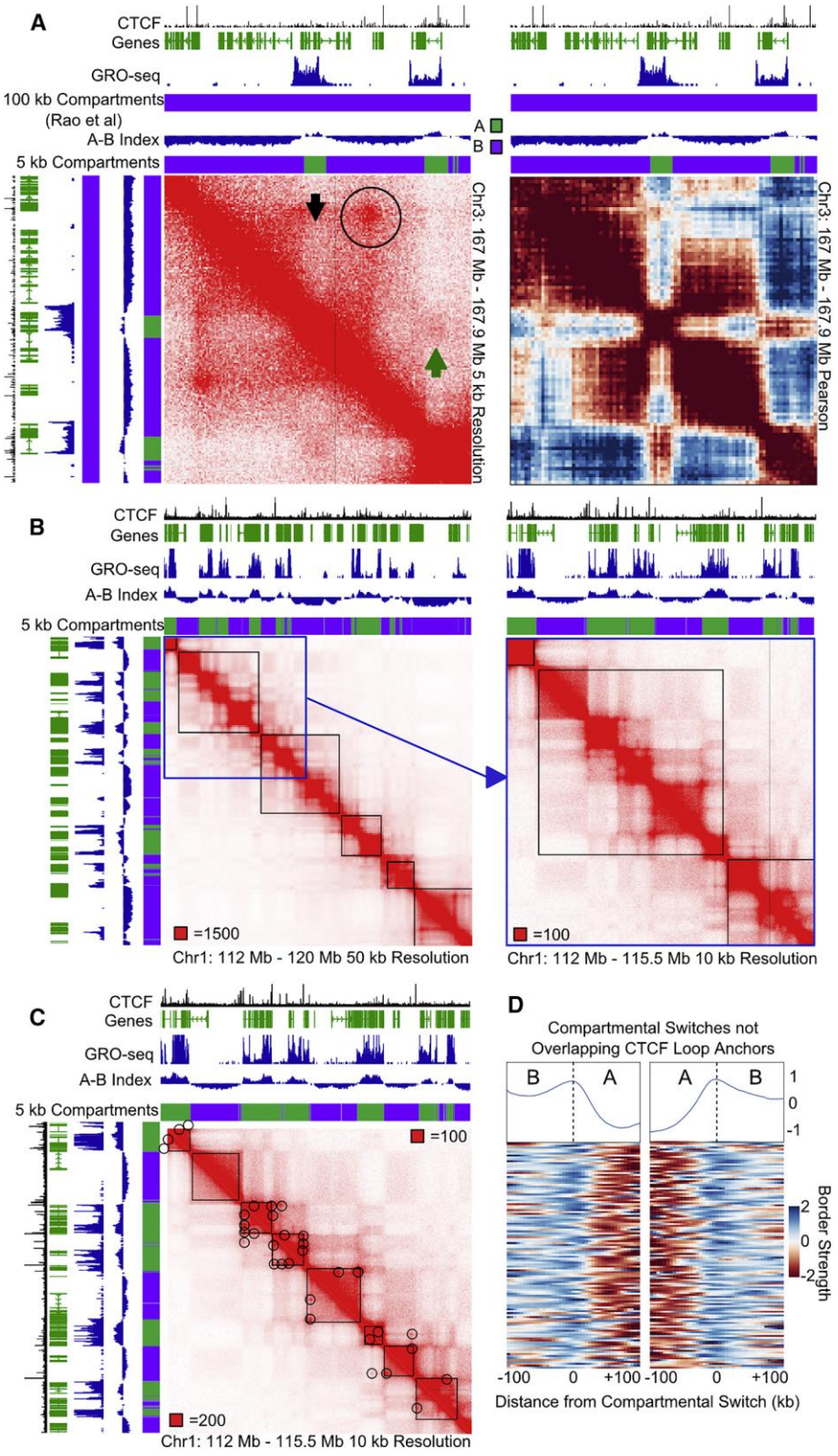



Figure 4.6 Compartments are Fine-scale Structures in Human Cells

A. Compartment identification using an A-B index obtained at 5 kb and previously reported compartments at 100 kb, showing identification of smaller A (green) and B (purple) compartments. Gene annotations are shown above and to the left. Left: Hi-C map at 5 kb resolution. Circle indicates a CTCF loop, black arrow indicates a distinct compartment switch within a CTCF loop, green arrow indicates inter-A compartment interactions. Right: Pearson correlation map showing A and B associations.

B. Compartmentalization subdivides low-resolution TADs. Black squares denote TAD calls at 40 kb(Moore et al., 2015). Blue square denotes area depicted to the right at higher resolution.

C. High resolution TAD calls identify small domains. Black squares denote high resolution TAD calls. Circles denote CTCF loops.

D. Compartments create domains in humans. Boundary score at compartmental switches more than 50 kb from a CTCF loop anchor. The median profile is shown above.

See also Figure 4.S6.

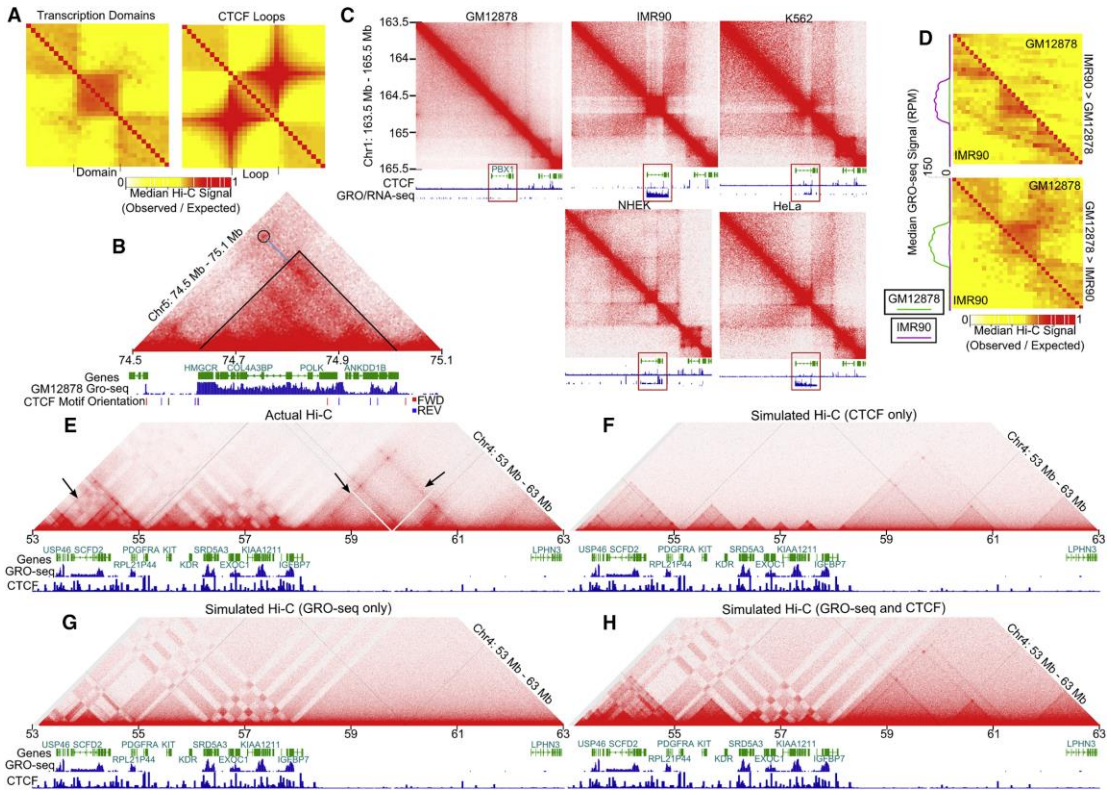



Figure 4.7 Transcriptional States and CTCF Loops Contribute to Formation of Domains in Human Cells

A. Transcriptionally active regions form domains distinct from CTCF loops. Scaled meta-plot of Hi-C interactions at transcriptionally active regions (left) compared to CTCF loops (right).

B. Hi-C heatmap of GM12878 cells at 5 kb resolution. A region where transcriptional activity matches border formation better than CTCF looping (circle) is shown. Blue line indicates CTCF loop anchor. Gene annotations, CTCF forward (red) and reverse (blue) motif orientation, and GRO-seq are shown below.

C. Hi-C heatmap comparing GM12878, IMR90, K562, NHEK, and HeLa cells. Tracks comparing GRO-seq/RNA-seq and CTCF occupancy in each cell line are shown below. Red rectangle indicates differentially expressed region.

D. Transcriptional activity corresponds to domain formation. Scaled meta-plots of distance normalized (observed/expected) Hi-C contacts surrounding transcriptionally active regions in IMR90 that are transcriptionally inactive in GM12878 (top) or vice versa (bottom). Metaplot of GRO-seq signal in GM12878 (green) and IMR90 (pink) for differentially called regions is shown on the left.

E. Transcriptional activity and CTCF looping explains chromatin architecture. Actual Hi-C contact map for a region of chromosome 4. Gene annotations, GRO-seq, and CTCF ChIP-seq signal tracks are shown below. Arrows indicate lines of interactions at CTCF anchors.

F. CTCF looping alone cannot explain chromatin organization. Simulation created using CTCF-loop information only.

G. Transcription alone cannot explain chromatin organization. Simulation created using GRO-seq signal correlation as the probability of two sites interacting.

H. Transcription and CTCF both contribute to chromatin organization. Simulation created using CTCF-loop information as well as GRO-seq signal as a measurement of transcriptional activity. Contacts are a feature of CTCF loops and the correlation in GRO-seq between loci.

See also Figure 4.S7.

Supplemental Figures

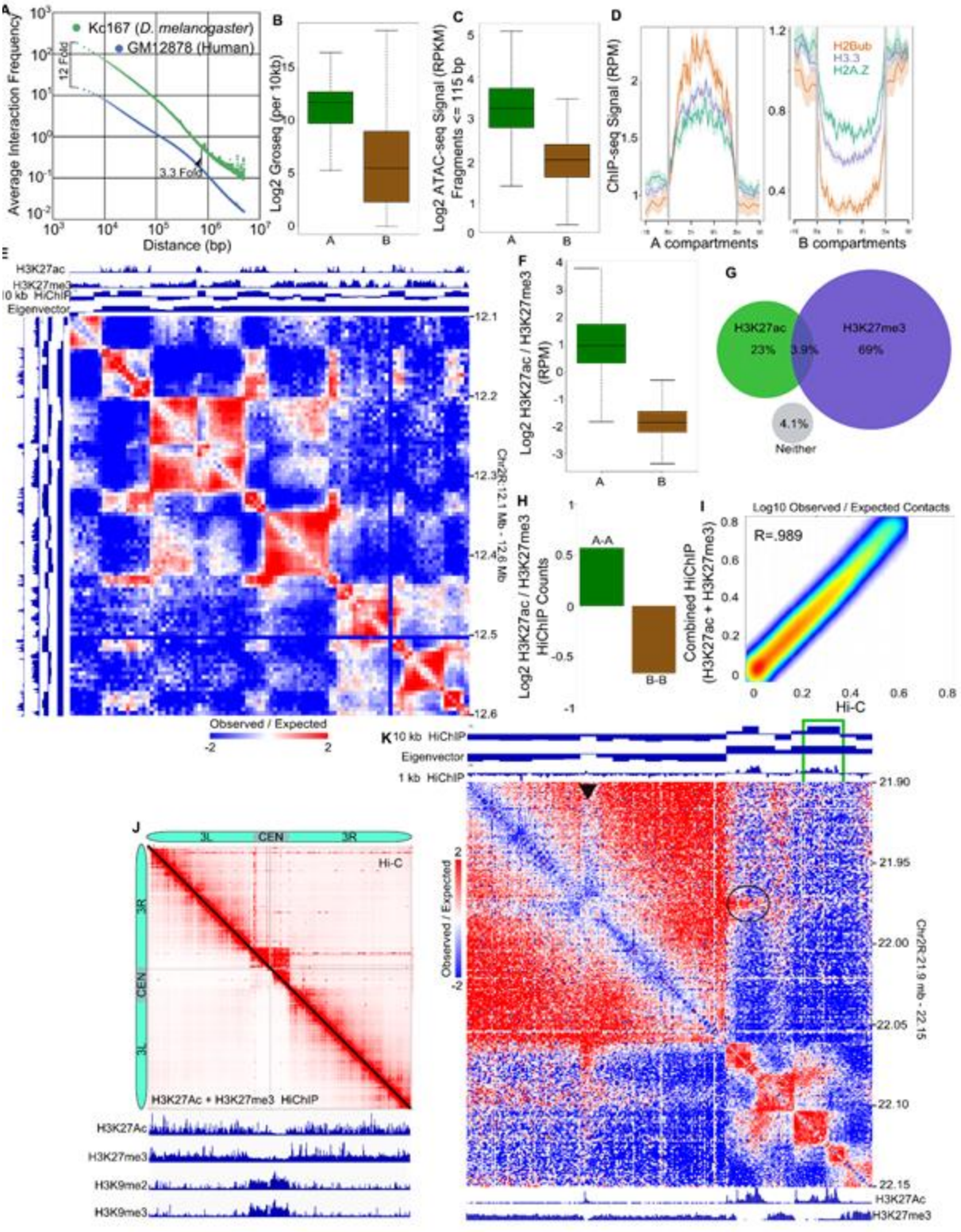



Figure 4.S1 Related to Figure 4.1

A. Comparison of depth between the highest resolution *Drosophila* (green) and human (blue) Hi-C datasets at each distance.

B. A and B compartments have distinct GRO-seq signal. $Log_2$ GRO-seq signal in 10 kb windows on A (green) and B (brown) compartments. Boxes depict median and interquartile range.

C. A and B compartments have distinct transcription factor signal. $Log_2$ ATAC-seq signal corresponding to short protected fragments (<= 115 bp) in A (green) and B (brown) compartments. Boxes depict median and interquartile range.

D. Active and inactive chromatin correspond to A and B compartments. Average histone modification profiles over A and B compartments. Color coding of ChIP-seq for histone modifications/variants is indicated.

E. Hi-C PCA and eigenvector decomposition failed to identify small compartments. Distance normalized Hi-C (observed/expected). Tracks show H3K27ac ChIP-seq, H3K27me3 ChIP-seq, and the Hi-C eigenvector. Track with 10 kb HiChIP depicts the preference of each site to interact in H3K27ac or H3K27me3 HiChIP.

F. H3K27ac and H3K27me3 are good measures of A and B compartments. $Log_2$ H3K27ac/H3K27me3 ChIP-seq signal on A and B compartments. Boxes depict median and interquartile range.

G. Most of the *Drosophila* genome is contains either H3K27ac or H3K27me3. Percentage of 1 kb bins with enriched ChIP-seq signal for either H3K27ac (green), H3K27me3 (purple), or neither (grey).

H. H3K27ac and H3K27me3 capture A and B compartmental interactions, respectively. Interaction counts of inter-A (green) and inter-B (brown) compartments as $\log_2$ ratio between HiChIP for H3K27ac and H3K27me3.

I. H3K27ac and H3K27me3 HiChIP combined recapitulates Hi-C interactions. Correlation between the merged HiChIP data and Hi-C. Distance normalized (observed/expected). Color range white to red indicates low to high density of data points.

J. H3K27ac and H3K27me3 HiChIP combined does not capture centromeric interactions. Hi-C compared to H3K27ac/H3K27me3 combined HiChIP across chromosome 3. ChIP-seq signal for H3K27ac, H3K27me3, H3K9me2, and H3K9me3 are shown below.

K. Compartmentalization occurs at 1 kb resolution. Distance normalized Hi-C (observed/expected). Tracks show H3K27ac ChIP-seq, H3K27me3 ChIP-seq (below), and the Hi-C eigenvector (above). The preference of each bin to interact in H3K27ac or H3K27me3 HiChIP computed at 10 kb and 1 kb resolution is shown. Arrows indicate compartmental switches discovered at 1 kb resolution at a small active site depleted for interactions with the surrounding inactive region. Circle indicates enriched interactions between the small active region and a nearby active region. Green box indicates compartmental switch missed by the eigenvector.

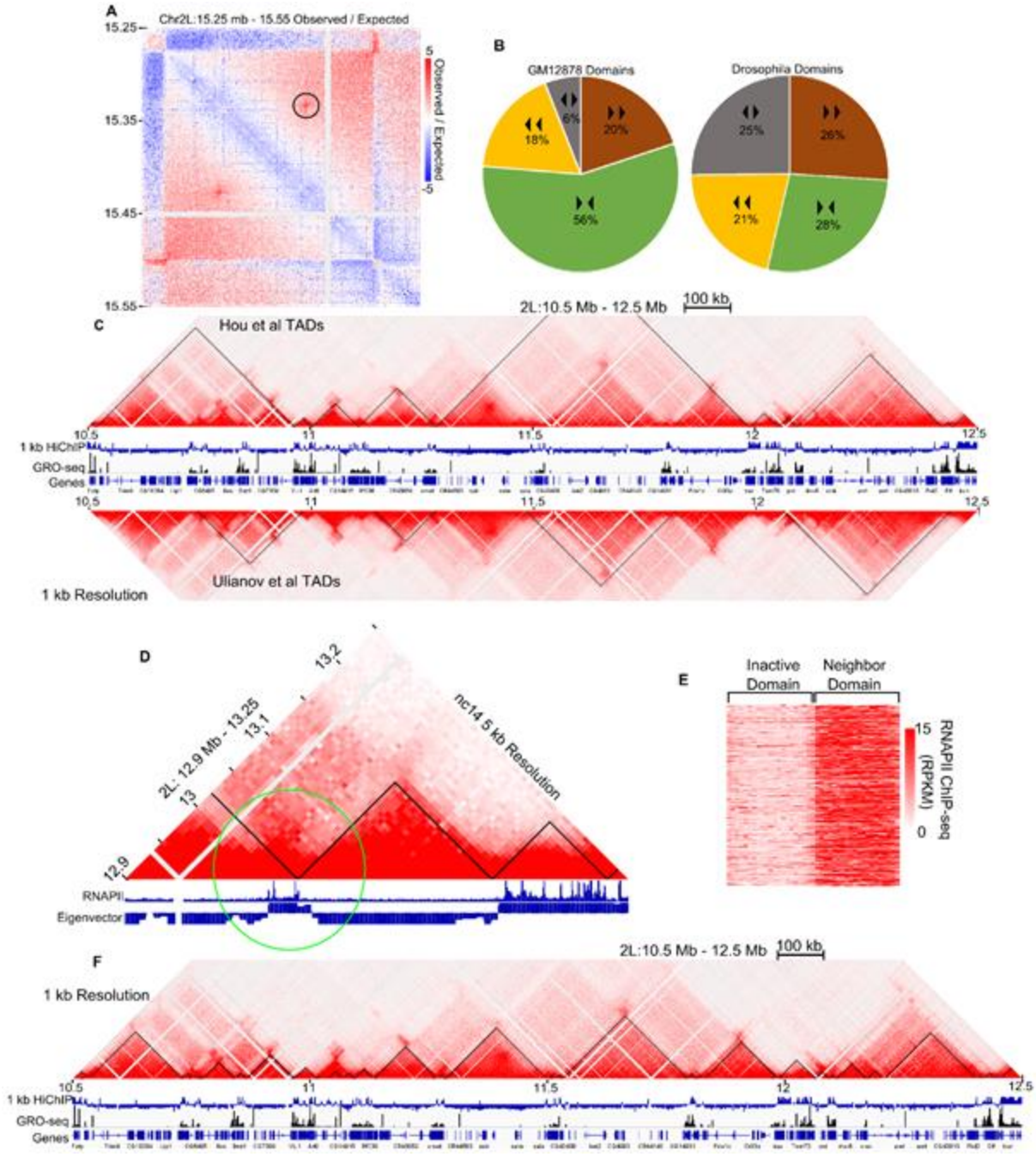



Figure 4.S2 Related to Figure 4.2

A. Drosophila loops do not form domain corners. Distance normalized Hi-C depicting a loop (circled).

B. CTCF orientation does not determine domain corners in Drosophila. Percentage of domains with border associated CTCF motifs in each orientation for GM12878 cells and Drosophila Kc167 cells. Arrows indicate CTCF motif orientations.

C. Hi-C maps with previously identified TAD calls (black) (Hou et al., 2012) shown at 1 kb resolutions. GRO-seq, HiChIP determined compartmental associations, and gene annotations are shown below. TAD calls from an independent source are also shown (bottom) (Ulianov et al., 2016).

D. Hi-C map of nc14 embryos with previously identified TAD calls (black) (Hug et al., 2017) shown at 5 kb resolution. RNAPII ChIP-seq signal and the compartmental associations (eigenvector) are shown below.

E. RNAPII is enriched throughout alternating domains. RNAPII ChIP-seq (RPKM) signal across neighboring domains.

F. Hi-C maps with compartmental domains (black) shown at 1 kb resolution. GRO-seq, HiChIP determined compartmental associations, and gene annotations are shown below.

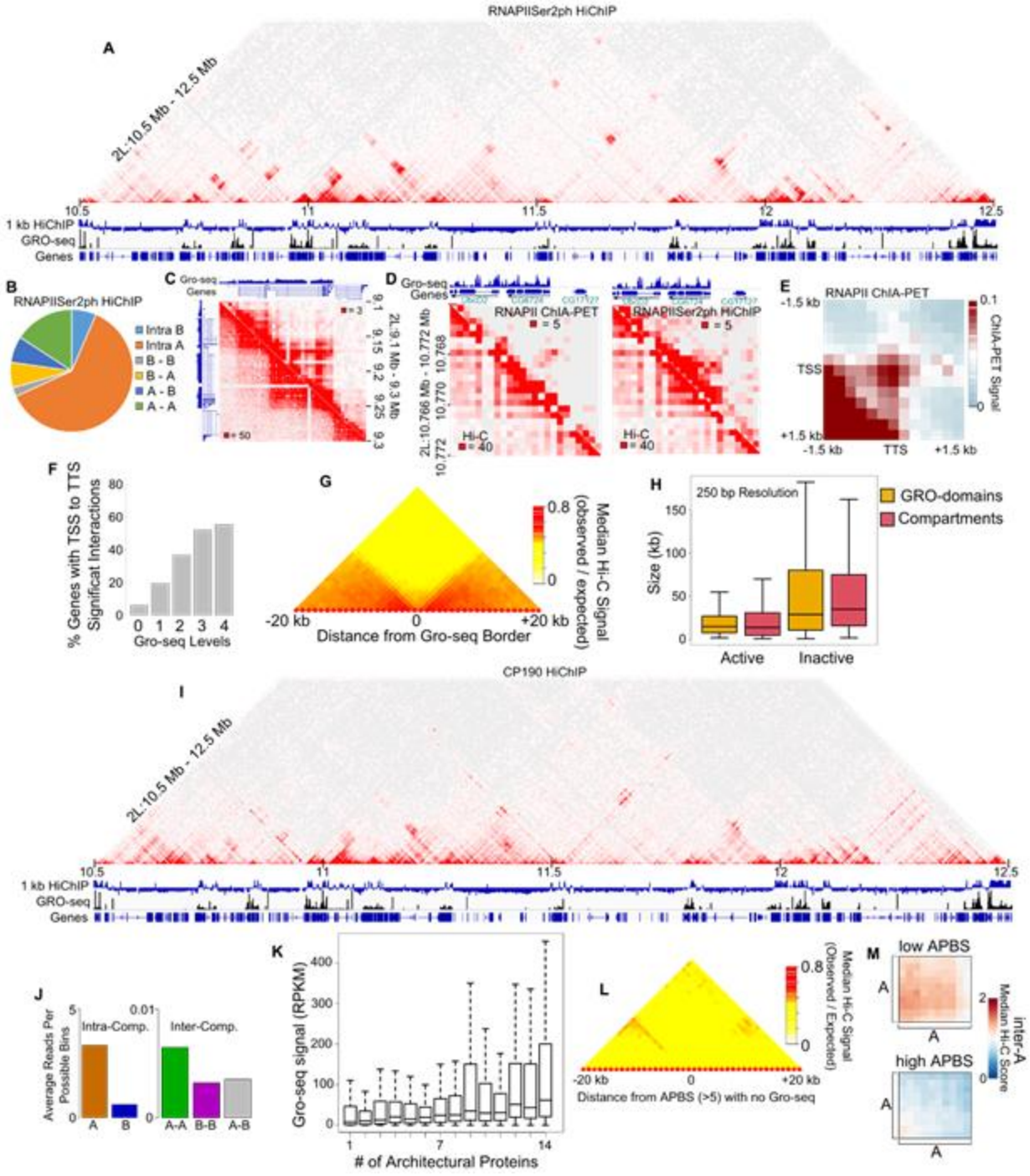



Figure 4.S3 Related to Figure 4.2 and Figure 4.4

A. RNAPII HiChIP heatmap. GRO-seq, HiChIP determined compartmental associations, and gene annotations are shown below.

B. Active compartments are composed of elongating RNAPII. Categorization of contacts from HiChIP for RNAPIISer2ph and the enrichment within A and B compartments.

C. Individual genes can form mini contact domains. RNAPIISer2ph HiChIP (top right) compared to Hi-C signal (bottom left). GRO-seq and gene annotations are shown above and to the left.

D. Individual genes can form mini-domains. RNAPII ChIA-PET (left) and RNAPIISer2ph HiChIP (right) signal compared to Hi-C signal in 250 bp bins. GRO-seq and gene annotations are shown above.

E. RNAPII ChIA-PET displays gene-loops. Median metaplot of RNAPII ChIA-PET signal on genes.

F. Gene loops correspond to expression. Percentage of significant TSS-TTS interactions found by RNAPII ChIA-PET on genes without any GRO-seq signal (0) and in each quartile of expression (1-4).

G. GRO-seq borders correlate with Hi-C domain borders. Median distance normalized (observed/expected) Hi-C signal surrounding identified GRO-seq domain borders.

H. Sizes of *Drosophila* domains. Boxplot of the sizes of compartmental domains (orange) and domains called from GRO-seq data (pink). Boxes depict median and interquartile range.

I. CP190 HiChIP heatmap. GRO-seq, HiChIP determined compartmental associations, and gene annotations are shown below.

J. CP190 interactions are in active A compartments. Categorization of contacts from HiChIP for CP190 and the enrichment within A and B compartments.

K. Transcription corresponds to APBS occupancy. Boxplot of GRO-seq levels categorized by architectural protein occupancy within 250 bp of the TSS. Boxes depict median and interquartile range.

L. High occupancy APBSs do not demarcate Hi-C domain borders. Median distance normalized (observed/expected) Hi-C signal surrounding identified high occupancy (>= 5 protein occupancy) APBSs at least 20 kb away from highly expressed TSSs.

M. A compartments are further decayed by APBS occupancy. Median metaplot of A-A compartmental interactions for those separated by low (left) and high (right) APBSs.

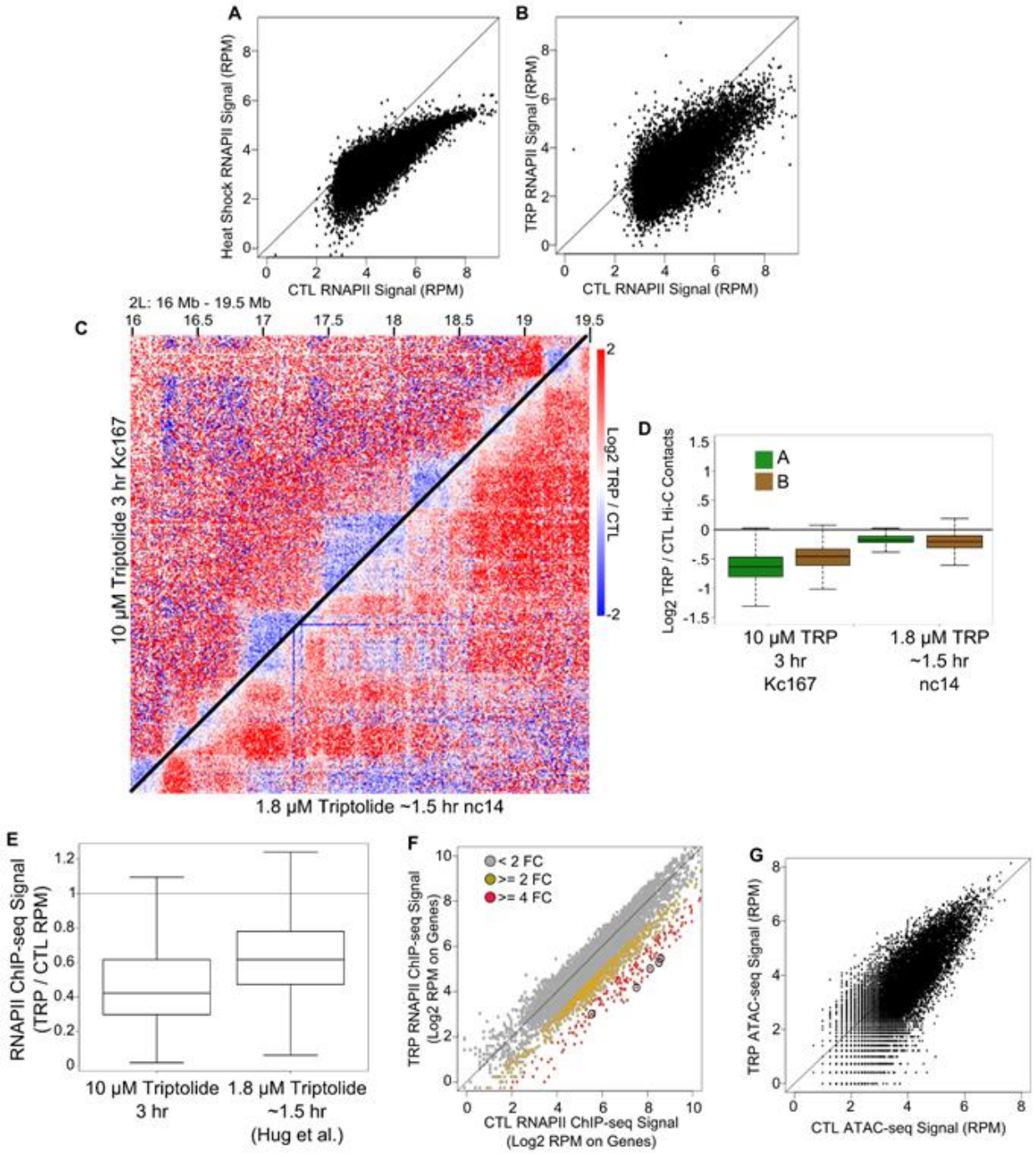



Figure 4.S4 Related to Figure 4.3

A. Effect of heat shock on RNAPII. RNAPII ChIP-seq signal at peaks in control (CTL) vs heat shock.

B. Effect of triptolide on RNAPII. RNAPII ChIP-seq signal at peaks in control (CTL) vs triptolide treatment (TRP).

C. Comparison of high and low triptolide treatment on domain formation. Differential Hi-C signal (triptolide/control) for Kc167 cells with high treatment levels (top left) compared to nc14 with low treatment levels (Hug et al., 2017) (bottom right).

D. Treatment with higher concentration of triptolide for longer times correlates with a greater decrease in domain structure. Intra domain contact differences (triptolide/control) in A and B compartments in KC167 cells vs nc14 embryos(Hug et al., 2017) with different treatment conditions. Boxes depict median and interquartile range.

E. Higher triptolide treatment causes greater changes to RNAPII occupancy at peaks. RNAPII ChIP-seq differential signal (triptolide/control) in Kc167 cells vs nc14 embryos (Hug et al., 2017) with different treatment conditions. Boxes depict median and interquartile range.

F. Low triptolide treatment conditions have little effect on RNAPII within genes. Genes with RNAPII peaks in control were plotted for their total RNAPII signal in the control (CTL) vs triptolide (TRP). Data from Hug et al. were used (Hug et al., 2017). Line indicates slope of 1. Grey, yellow, and red dots represent genes with less than 2, 2-4, and greater than 4-fold decrease in RNAPII ChIP-seq signal respectively. Black circles indicate genes tested by RT-PCR in Hug et al.

G. Effect of triptolide on chromatin accessibility. ATAC-seq signal (fragments <= 115 bp) on peaks in in control (CTL) vs triptolide treatment (TRP).

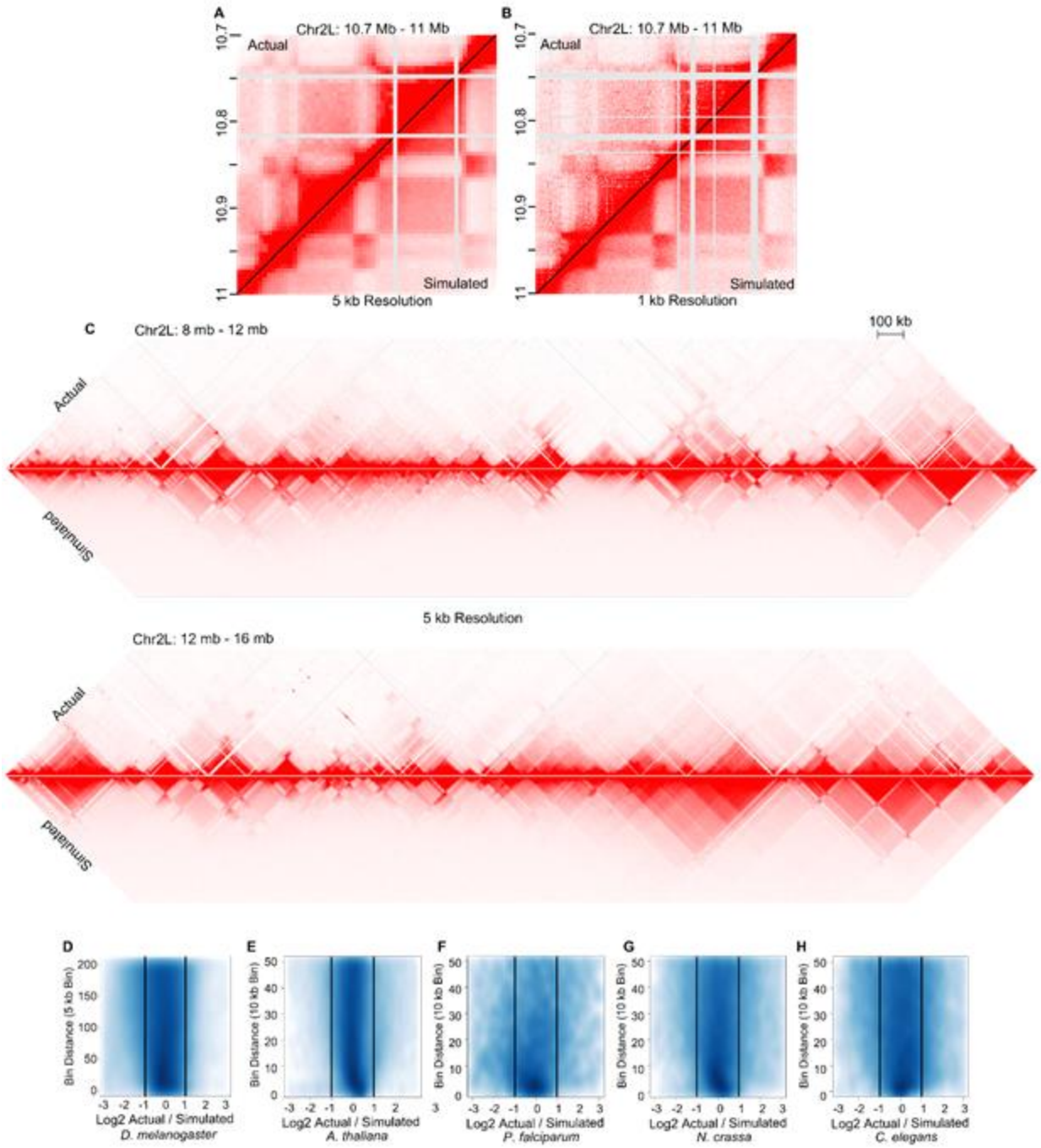



Figure 4.S5 Related to Figure 4.5

A. GRO-seq and APBS based simulated contact maps precisely recapitulate domains and compartments at 5 kb resolution. Contact heatmap at 5 kb resolution using actual Hi-C data

(top) and simulated data based on GRO-seq and APBS occupancy (bottom). 300 kb of chr2L are shown.

B. GRO-seq and APBS based simulated contact maps precisely recapitulate domains and compartments at 1 kb resolution. Contact heatmap at 1 kb resolution using actual Hi-C data (top) and simulated data based on GRO-seq and APBS occupancy (bottom). 300 kb of chr2L are shown.

C. GRO-seq and APBS based simulated contact maps recapitulate domains and compartments. Contact heatmaps at 5 kb resolution using actual Hi-C data (top) and simulated data based on GRO-seq and APBS occupancy (bottom).

D-H. Actual Hi-C compared to simulated data. Density scatter plot showing the ratio of actual Hi-C reads over the simulated reads (x-axis) for each genomic distance (y-axis) for *D. melanogaster* (D), *A. thaliana* (E), *P. falciparum* (F), *N. crassa* (G), *C. elegans* (H). Reads were normalized by distance decay values. Vertical lines indicate a 2-fold change.

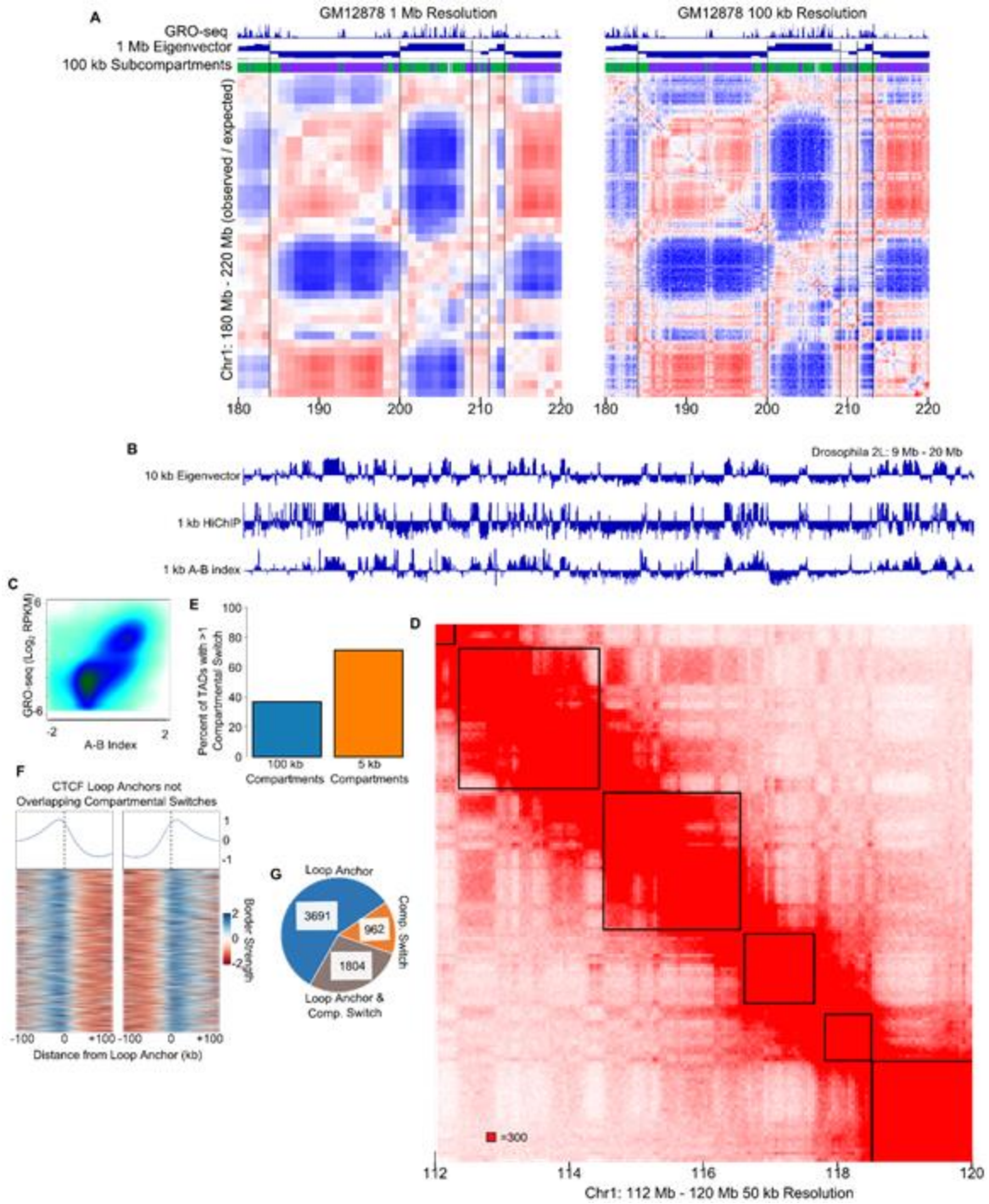



Figure 4.S6 Related to Figure 4.6

A. Large compartments are a result of read binning. Distance normalized Hi-C (observed/expected) in GM12878 cells at 1 Mb (left) and 100 kb (right) resolution. Tracks with 1 Mb eigenvector, 100 kb compartments, and GRO-seq are shown above.

B. Validation of the A/B index compartmental refinement method. *Drosophila* tracks comparing the A-B index to the eigenvector or 1 kb HiChIP compartment calls.

C. Transcriptional activity corresponds to fine-scale compartments. $Log_2$ GRO-seq compared to A-B index. Color gradient indicates density of points.

D. TAD calls at saturation. TAD calls in GM12878 cells at 50 kb resolution.

E. TADs are comprised of compartmental interactions. Percentage of TADs with more than 1 compartment switch within the domain. Previously called (at 100 kb resolution - blue) and newly called (at 5 kb resolution – orange) compartments are shown.

F. CTCF loops correlate with borders. Heatmap displaying the border strength for 100 kb to either side of left and right CTCF loop anchors. Average profiles are shown above.

G. TAD borders correspond to loops and compartments. Number of TAD borders coinciding with loops, compartmental switches, or both. $p < .001$ for loop anchors and compartments via random permutation test.

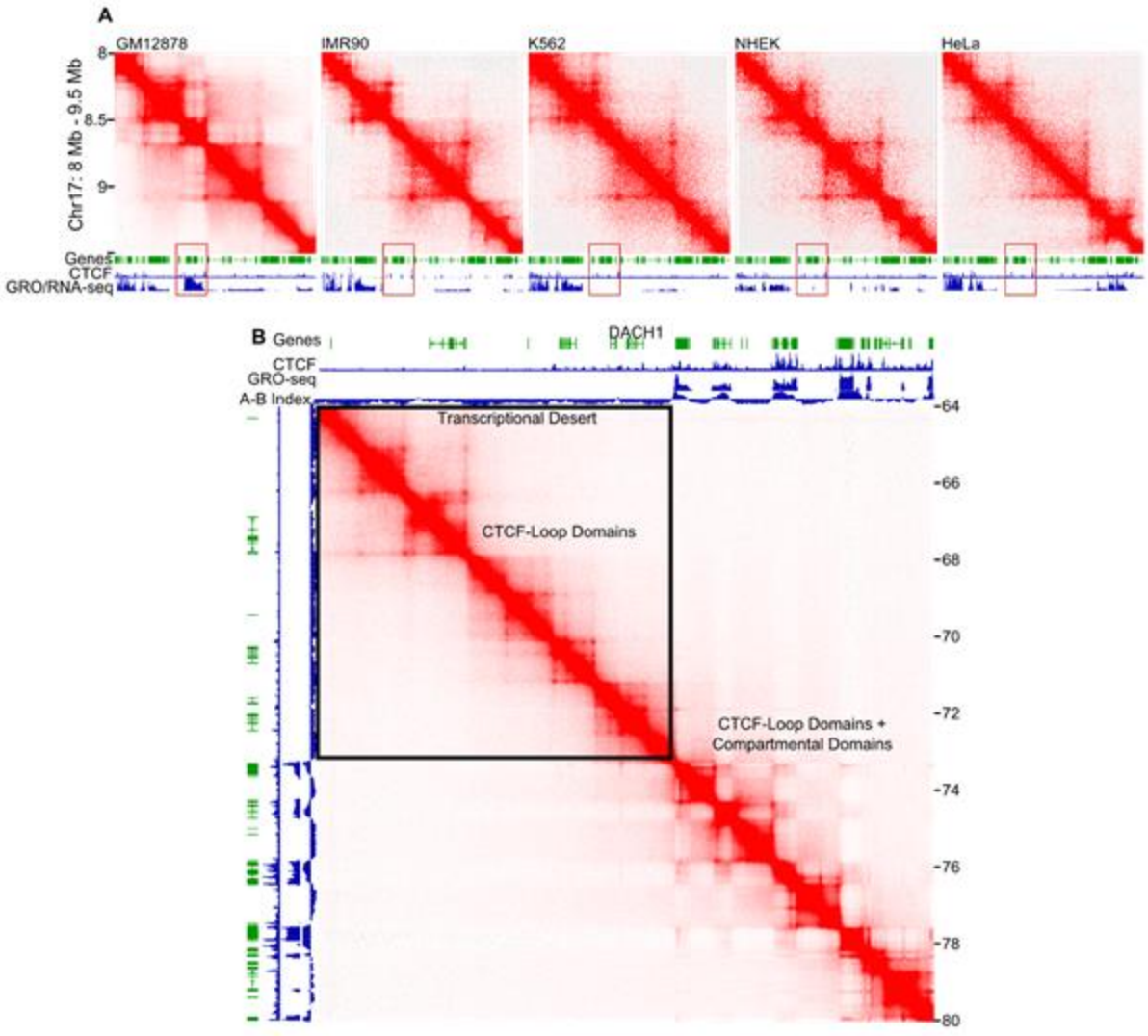



Figure 4.S7 Related to Figure 4.7

A. Transcriptional activity correlates with the formation of domains. Hi-C heatmaps for GM12878, IMR90, K562, NHEK, and HeLa cells. Tracks comparing GRO-seq/RNA-seq and CTCF occupancy between the different cell lines are shown below. Red rectangle indicates differentially expressed region.

B. Compartmental domains underlie CTCF loops. Hi-C heatmap for GM12878 cells displaying a large region with CTCF loops and no compartmental domains (transcriptional desert), as well

as a region with compartmental domains and CTCF loops. Gene annotations, CTCF ChIP-seq, GRO-seq, and fine-scale compartment identification tracks are shown above and to the left.

| | H3K27ac Rep1 HiChIP | H3K27ac Rep2 HiChIP |
|---|---|---|
| Sequenced Read Pairs | 100,225,894 | 22,784,945 |
| Normal Paired | 80,184,199 (80.00%) | 18,663,253 (81.91%) |
| Chimeric Paired | 1,066 (0.00%) | 119 (0.00%) |
| Chimeric Ambiguous | 2,002 (0.00%) | 390 (0.00%) |
| Unmapped | 20,038,627 (19.99%) | 4,121,183 (18.09%) |
| Ligation Motif Present | 31,666,024 (31.59%) | 8,548,749 (37.52%) |
| Alignable (Normal+Chimeric Paired) | 80,185,265 (80.00%) | 18,663,372 (81.91%) |
| Unique Reads | 68,648,039 (68.49%) | 17,945,684 (78.76%) |

| | | |
|---|---|---|
| PCR Duplicates | 11,536,419 (11.51%) | 716,899 (3.15%) |
| Optical Duplicates | 807 (0.00%) | 789 (0.00%) |
| Library Complexity Estimate | 251,239,520 | 236,654,145 |
| Intra-fragment Reads | 889,140 (0.89% / 1.30%) | 1,627,771 (7.14% / 9.07%) |
| Below MAPQ Threshold | 18,799,325 (18.76% / 27.39%) | 4,793,044 (21.04% / 26.71%) |
| Hi-C Contacts | 48,959,574 (48.85% / 71.32%) | 11,524,869 (50.58% / 64.22%) |
| Ligation Motif Present | 12,269,091 (12.24% / 17.87%) | 3,604,694 (15.82% / 20.09%) |
| 3' Bias (Long Range) | 85% - 15% | 78% - 22% |
| Pair Type %(L-I-O-R) | 25% - 25% - 25% - 25% | 25% - 25% - 25% - 25% |
| Inter-chromosomal | 2,632,278 (2.63% / 3.83%) | 786,657 (3.45% / 4.38%) |

| | | |
|---|---|---|
| Intra-chromosomal | 46,327,296 (46.22% / 67.49%) | 10,738,212 (47.13% / 59.84%) |
| Short Range (<20Kb) | 23,349,446 (23.30% / 34.01%) | 5,133,320 (22.53% / 28.60%) |
| Long Range (>20Kb) | 22,975,053 (22.92% / 33.47%) | 5,604,850 (24.60% / 31.23%) |
| Combined Contacts: 60,484,443 | | |

Table S1 Related to Figure 1

HiChIP and ChIA-PET mapping statistics performed in Kc167 cells to the dm6 genome for H3K27ac.

| | H3K27me3 Rep1 HiChIP | H3K27me3 Rep2 HiChIP | H3K27me3 Rep3 HiChIP |
|---|---|---|---|
| Sequenced Read Pairs | 86,473,347 | 22,785,730 | 73,918,791 |

| | | | |
|---|---|---|---|
| Normal Paired | 68,220,334 (78.89%) | 19,171,480 (84.14%) | 54,625,690 (73.90%) |
| Chimeric Paired | 1,421 (0.00%) | 259 (0.00%) | 302 (0.00%) |
| Chimeric Ambiguous | 2,149 (0.00%) | 640 (0.00%) | 1,308 (0.00%) |
| Unmapped | 18,249,443 (21.10%) | 3,613,351 (15.86%) | 19,291,491 (26.10%) |
| Ligation Motif Present | 27,127,062 (31.37%) | 5,362,368 (23.53%) | 31,273,691 (42.31%) |
| Alignable (Normal+Chimeric Paired) | 68,221,755 (78.89%) | 19,171,739 (84.14%) | 54,625,992 (73.90%) |
| Unique Reads | 18,003,439 (20.82%) | 11,579,237 (50.82%) | 46,920,236 (63.48%) |
| PCR Duplicates | 50,217,579 (58.07%) | 7,591,992 (33.32%) | 7,704,177 (10.42%) |
| Optical Duplicates | 737 (0.00%) | 510 (0.00%) | 1,579 (0.00%) |

| | | | |
|---|---|---|---|
| Library Complexity Estimate | 18,462,085 | 17,270,710 | 174,978,194 |
| Intra-fragment Reads | 304,152 (0.35% / 1.69%) | 340,787 (1.50% / 2.94%) | 1,328,032 (1.80% / 2.83%) |
| Below MAPQ Threshold | 6,606,156 (7.64% / 36.69%) | 4,286,041 (18.81% / 37.01%) | 17,587,159 (23.79% / 37.48%) |
| Contacts | 11,093,131 (12.83% / 61.62%) | 6,950,409 (30.51% / 60.04%) | 28,005,045 (37.89% / 59.69%) |
| Ligation Motif Present | 2,951,454 (3.41% / 16.39%) | 1,306,510 (5.73% / 11.28%) | 9,483,361 (12.83% / 20.21%) |
| 3' Bias (Long Range) | 84% - 16% | 74% - 26% | 78% - 22% |
| Pair Type %(L-I-O-R) | 25% - 25% - 25% - 25% | 25% - 25% - 25% - 25% | 25% - 25% - 25% - 25% |
| Inter-chromosomal | 975,417 (1.13% / 5.42%) | 534,655 (2.35% / 4.62%) | 1,839,742 (2.49% / 3.92%) |

| | | | |
|---|---|---|---|
| Intra-chromosomal | 10,117,714 (11.70% / 56.20 %) | 6,417,754 (28.17% / 55.42%) | 26,165,303 (35.40% / 55.77%) |
| Short Range (<20Kb) | 3,935,392 (4.55% / 21.86%) | 3,103,851 (13.62% / 26.81%) | 11,830,360 (16.00% / 25.21%) |
| Long Range (>20Kb) | 6,180,938 (7.15% / 34.33%) | 3,313,884 (14.54% / 28.62%) | 14,334,925 (19.39% / 30.55%) |
| Combined Contacts: 46,048,585 | | | |

Table 4.S2 Related to Figure 4.1

HiChIP and ChIA-PET mapping statistics performed in Kc167 cells to the dm6 genome for H3K27me3.

| Experiment description | RNAPIISer2ph HiChIP Rep 1 | RNAPIISer2ph HiChIP Rep2 | RNAPII ChIA-PET | CP190 HiChIP |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Sequenced Read Pairs | 34,167,551 | 73,194,819 | 33,708,212 | 112,982,203 |
| Normal Paired | 30,167,498 (88.29%) | 62,999,182 (86.07%) | 21,050,572 (62.45%) | 89,231,912 (78.98%) |
| Chimeric Paired | 2,513 (0.01%) | 425 (0.00%) | 3,821,972 (11.34%) | 9,957 (0.01%) |
| Chimeric Ambiguous | 901 (0.00%) | 1,219 (0.00%) | 7,570,272 (22.46%) | 2,803 (0.00%) |
| Unmapped | 3,996,639 (11.70%) | 10,193,993 (13.93%) | 1,265,396 (3.75%) | 23,737,531 (21.01%) |
| Ligation Motif Present | 5,724,083 (16.75%) | 17,071,721 (23.32%) | 88,205 (0.26%) | 23,426,133 (20.73%) |
| Alignable (Normal+Chimeric Paired) | 30,170,011 (88.30%) | 62,999,607 (86.07%) | 24,872,544 (73.79%) | 89,241,869 (78.99%) |
| Unique Reads | 23,764,884 (69.55%) | 57,498,996 (78.56%) | 14,769,195 (43.81%) | 62,817,133 (55.60%) |

| | | | | |
|---|---|---|---|---|
| PCR Duplicates | 6,403,207 (18.74%) | 5,498,867 (7.51%) | 10,098,920 (29.96%) | 26,424,013 (23.39%) |
| Optical Duplicates | 1,920 (0.01%) | 1,744 (0.00%) | 4,429 (0.01%) | 723 (0.00%) |
| Library Complexity Estimate | 60,608,219 | 339,548,156 | 21,598,913 | 119,190,245 |
| Intra-fragment Reads | 1,320,369 (3.86% / 5.56%) | 3,619,547 (4.95% / 6.29%) | 7,539,437 (22.37% / 51.05%) | 9,756,188 (8.64% / 15.53%) |
| Below MAPQ Threshold | 6,546,934 (19.16% / 27.55%) | 13,203,639 (18.04% / 22.96%) | 1,431,119 (4.25% / 9.69%) | 15,674,333 (13.87% / 24.95%) |
| Contacts | 15,897,581 (46.53% / 66.90%) | 40,675,810 (55.57% / 70.74%) | 5,798,639 (17.20% / 39.26%) | 37,386,612 (33.09% / 59.52%) |
| Ligation Motif Present | 2,148,740 (6.29% / 9.04%) | 7,427,401 (10.15% / 12.92%) | 35,778 (0.11% / 0.24%) | 6,873,662 (6.08% / 10.94%) |

| | | | | |
|---|---|---|---|---|
| 3' Bias (Long Range) | 80% - 20% | 73% - 27% | 51% - 49% | 79% - 21% |
| Pair Type %(L-I-O-R) | 25% - 25% - 25% - 25% | 25% - 25% - 25% - 25% | 25% - 25% - 25% - 25% | 25% - 25% - 25% - 25% |
| Inter-chromosomal | 1,027,932 (3.01% / 4.33%) | 1,332,949 (1.82% / 2.32%) | 281,654 (0.84% / 1.91%) | 950,358 (0.84% / 1.51%) |
| Intra-chromosomal | 14,869,649 (43.52% / 62.57%) | 39,342,861 (53.75% / 68.42%) | 5,516,985 (16.37% / 37.35%) | 36,436,254 (32.25% / 58.00%) |
| Short Range (<20Kb) | 8,921,240 (26.11% / 37.54%) | 24,659,290 (33.69% / 42.89%) | 4,893,337 (14.52% / 33.13%) | 26,353,410 (23.33% / 41.95%) |
| Long Range (>20Kb) | 5,935,988 (17.37% / 24.98%) | 14,683,333 (20.06% / 25.54%) | 623,626 (1.85% / 4.22%) | 10,074,135 (8.92% / 16.04%) |
| Combined Contacts: 56,573,391 | | | | |

Table 4.S3 Related to Figure 4.2 and Figure 4.4

HiChIP and ChIA-PET mapping statistics performed in Kc167 cells to the dm6 genome for RNAPIISer2ph, CP190, and RNAPII.

# Methods

## Contact for Reagent and Resource Sharing

Requests for further information or reagents should be directed to the corresponding author, Victor Corces, Email: vgcorces@gmail.com, Phone: 404-727-4250, Fax: 404-727-2880.

## Experimental Model and Subject Details

Kc167 cells (female embryonic) were obtained from the Drosophila Genomics Resource Center (DGRC) and grown at 25°C in Hyclone SFX insect culture media (GE Healthcare).

## Method Details

### *Hi-C, ChIA-PET, and HiChIP Library Preparation and Processing*

Hi-C heatmaps were Knight-Ruiz (KR) normalized and visualized by Juicer and Juicebox (Durand et al., 2016b, 2016a). Resolution estimate was calculated exactly as described (Rao et al., 2014). Hi-C meta-plots were created using custom scripts; scores were set from zero to one

equaling the highest and lowest values within a plot or across plots in a set. All Hi-C datasets from other organisms were reprocessed and normalized using the Juicer pipeline.

ChIA-PET libraries were prepared as previously described (Goh et al., 2012). HiChIP libraries for CP190 and RNA Polymerase II phosphorylated in serine 2 were prepared as described with minor modifications (Mumbach et al., 2016). 100 x 10^6 Kc167 cells at 80% confluency were crosslinked in 1% formaldehyde for 10 min at room temperature, after which cells were incubated in 0.2 M glycine for 5 min to stop the reaction. Cells were pelleted and resuspended in 500 µl cold Hi-C lysis buffer (10 mM Tris-HCl pH8, 10 mM NaCl, 0.2% Igepal CA-630, and 1x Protease Inhibitor (Roche 11873580001) and incubated on ice for 1 h. Nuclei were pelleted at 2500 rcf for 5 min at 4°C, resuspended in 100 µl 0.5% SDS, and incubated for 5 min at 65°C. We then added 290 µl of $H_2O$ and 50 µl of 10% Triton X-100, incubated samples for 15 min at 37°C. 50 µl of 10x *DpnII* buffer and 200 u of *DpnII* (NEB R0543) were added and samples were digested overnight at 37°C with rotation.

After digestion, samples were incubated at 65°C for 20 min to inactivate *DpnII*, and each was divided into two reactions and allowed to cool to room temperature. Biotin fill-in was done with 22.5 µl of water, 1.5 µl each of 10 mM dTTP, dATP, and dGTP, 15 µl of 1 mM biotin-16-dCTP (Jena Bioscience JBS-NU-809-BIO16), and 8 µl of 5 u/µl DNA polymerase I Large (Klenow) fragment (NEB M210). This reaction was placed at 37°C for 1.5 h, after which samples were ligated for 4 h at room temperature with addition of 663 µl $H_2O$, 120 µl 10x NEB T4 DNA Ligase buffer, 100 µl 10% Triton X-100, 12 µl 10 mg/ml BSA, and 5 µl 400 u/µl T4 DNA Ligase (NEB M0202).

Following ligation, nuclei were pelleted and resuspended in 200 µl cold Nuclei Lysis Buffer (50 mM Tris-HCl pH 9, 10 mM EDTA, 1% SDS, and 1x Protease Inhibitors) with incubation on ice for 20 min. After incubation we added 100 µl cold IP Dilution Buffer (0.01% SDS, 1.1% Trition X-

100, 1.2 mM EDTA, 16.7 Tris-HCl pH 8, 16.7 mM NaCl, and 1x Protease Inhibitors) and sonicated to approximately 250 bp fragments. Cell debris was pelleted and the supernatant was transferred into a new 1.5 ml tube for immunoprecipitation.

Each sample was precleared before immunoprecipitation by taking 10 µl Protein A and 10 µl Protein G magnetic beads, washing 3x in 0.5% BSA in 1x PBS, followed by incubation with 10 µl pre-immune rabbit serum in 500 µl 0.5% BSA/PBS for 4 h at 4°C with rotation. Afterward beads were washed with 1 ml 0.5% BSA/PBS for 2 min at room temperature, followed by 2 washes in 1 ml IP Dilution Buffer, and resuspension in 300 µl cold IP Dilution Buffer. Beads with each antibody were also prepared the same way.

Chromatin was diluted 5-fold with cold IP Dilution Buffer and incubated with pre-clear beads for 1-2 h at 4°C with rotation. The unbound portion was then transferred to antibody-coated beads and incubated overnight at 4°C with rotation. After IP, samples were washed 3x with low Low Salt Buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA pH 8, 20 mM Tris-HCl pH 8, 150 mM NaCl), 2x with High Salt Buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA pH 8, 20 mM Tris-HCl pH 8, 500 mM NaCl), 2x with LiCl Buffer (10 mM Tris-HCl pH 8, 1 mM EDTA, 0.25 M LiCl, 1% Igepal CA-630, 1% DOC), and 1x with TE buffer.

DNA was eluted 2x using 150 µl freshly prepared IP elution buffer (0.1 M NaHCO3, 1% SDS) for 10 min at room temperature, followed by 5 min at 37°C and transferring to a new tube, combining eluates. For crosslink reversal we added 20 µl 5 M NaCl, 8 µl 0.5 M EDTA and 16 µl 1 M Tris-HCl pH8, incubating 1.5 h at 68°C. Afterwards we added 8 µl proteinase K and incubated at 50°C for 2 h. After allowing samples to reach room temperature, we precipitated DNA in ethanol with Sodium Acetate, resuspending in 300 µl 10 mM Tris-Cl pH 8.5.

To enrich for ligation events we prepared Streptavidin beads by washing in 400 µl TWB (5mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween 20) and resuspending in 300 µl of 2x

Binding Buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl). Beads were added to the sample and incubated at room temperature for 15 min with rotation. Samples were then washed 2x in TWB and the standard Hi-C library preparation was followed (Rao et al., 2014).

Sequenced reads were mapped to the *Drosophila* dm6 genome, further processed to remove duplicates and self-ligations using the Juicer pipeline, and visualized using Juicebox(Durand et al., 2016a, 2016b). Statistics for each library can be found in Tables 4.S1-S3. The overlap of HiChIP and compartments was computed by the sum of reads divided by the total number of possible bins in each category. Significant interactions were calculated using MICC (He et al., 2015).

*Domains and Compartments*

Identification of *Drosophila* TADs and domains has been described previously (Cubeñas-Potts et al., 2016; Hou et al., 2012; Ulianov et al., 2016) as were GM12878 TADs and smaller contact domains (Moore et al., 2015; Rao et al., 2014). Hi-C directionality index (DI) was calculated as previously described (Dixon et al., 2012) using the equation:

To compensate for the smaller genome and smaller domain structures seen in *D. melanogaster,* we calculated A and B using interactions more than 5 kb but less than 100 kb from each 250 bp bin throughout the genome. Directionality index based domains were called following a hidden Markov model.

*Drosophila* compartments were identified from the eigenvector computation and Pearson correlation matrices as previously reported (Lieberman-Aiden et al., 2009) using Juicebox. Profiles of different histone modifications across compartments were calculated using ngsplot (Shen et al., 2014).

To calculate the correlation between Hi-C and histone modification HiChIP, samples were read normalized by random picking and H3K27ac and H3K27me3 were combined using Juicebox. Distance normalized interaction signals (observed/expected) within each 10 kb bin were then compared to Hi-C and tested by a Pearson correlation. Compartments mapped using HiChIP were identified by computing the preferential contacts of each row in the matrix with H3K27ac or H3K27me3 such that each bin was given a value of $\log_2$(H3K27ac/H3K27me3) contacts. Compartments were then identified from this relative association by a hidden Markov model. Differences in intra-domain and compartmental interactions after triptolide were calculated by the sum of 1 kb resolution interactions more than 2 kb apart.

Human compartments were called by creating a 5 kb by 125 kb matrix and measuring the median $\log_2$ distance normalized interaction score with previously defined lower resolution A and B compartments (Rao et al., 2014). An A-B index was then created by subtracting the A and B scores. This index represents the comparative likelihood of a sequence interacting with A or B. 5 kb bins with positive values (more association with A) were called as high-resolution A compartments, while 5 kb bins with negative values (more association with B) were called as high-resolution B compartments. Overlap of CTCF loops and compartmental switches with TAD borders was calculated for each border with a feature within 40 kb of the border and p-values were calculated by permutation test.

Transcriptional state domains were obtained using a hidden Markov model of GRO-seq data binned at 1 kb resolution (Core et al., 2012; Kwak et al., 2013). This utilized a Gaussian distribution to classify each 1 kb bin as an active or inactive state. Transcriptional domains were determined as regions without transcriptional state switches and regions less than 2 kb were merged into the neighboring domains. Differential active domains between GM12878 and IMR90 cells were identified as those with average signal across the region greater than 1 RPKM in one cell type but less than 0.5 RPKM in the other.

*ChIP-seq Datasets*

Architectural protein binding sites were individually identified by MACS (Zhang et al., 2008). A 200 bp region around the summit was used to combine peaks from all ChIP-seq data-sets. Unique peaks were kept, and overlapping regions were merged placing the center point as the new summit. A second filter was then used to determine occupancy such that RPM normalized read counts were three-fold higher than IgG on the combined peak list. Individual architectural proteins used for APBS occupancy were BEAF32, CAPH2, Chromator, CP190, CTCF, DREF, Fs(1)h-L, L3mbt, Mod(mdg4), Nup98, Rad21, SuHw, TFIIIC, and Z4. Overlap with promoters was determined if the 200 bp region lay within 50 bp of the TSS.

ChIP-seq libraries for histone modifications were prepared and processed in Kc167 cells as previously described (Cubeñas-Potts et al., 2016) and included ChIP-seq for H3K36me3, H3K9me3, H4K16ac, H4K20me1, H2Bub, H3.3, and H2A.Z. ChIP-seq for H3.3 was done in a Kc167 line expressing V5-tagged H3.3 (Wirbelauer et al., 2005).

To calculate the fraction of the *Drosophila* genome bound by H3K27me3 and/or H3K27ac we used input normalized signal levels at H3K27ac peaks to estimate the background signal of H3K27me3. This was done by dividing the genome into 1 kb bins and counting RPM normalized reads in each ChIP-seq and input dataset. The threshold above which most H3K27ac peaks contained H3K27ac but not H3K27me3, and where non-peaks contained the reverse, was used.

*ATAC-seq*

Kc167 cells grown to exponential stage were treated with DMSO or triptolide as previously described (Li et al., 2015). 200,000 ctrl and treated cells were collected and processed using the Fast-ATAC protocol (Corces et al., 2016). Briefly, cell pellets were resuspened in 50 µl Tn5

transposase mixture (0.01% digitonin for permeabilizing cell membrane, 2.5 µl Tn5, 25 µl TD buffer), and incubated at 30°C for 20 min with occasional shaking. After reaction, cells were cooled on ice and DNA was purified using the Minelute Kit (Qiagen). 25 µl of eluted DNA were used for real time PCR amplification in the reaction mixture (2x KAPA HiFi mix and 1.25 µM indexed primers) using the following conditions: 72°C for 5 min; 98°C for 30 sec; and 10-11 cycles at 98°C for 10 sec, 63°C for 30 sec, and 72°C for 1 min. Reads were trimmed of adapters, mapped to the *Drosophila* dm6 genome, deduplicated, and separated into short (<=115 bp) and long fragments (180-247 bp) to obtain transcription factor and nucleosome profiles, respectively. Peaks were identified using MACS2.

*RNA-seq and GRO-seq Analysis*

Transcriptional quartiles were taken by counting reads within the first 100 bp of genes, and removing genes with no reads as a separate set to reduce repetitive biases. Distance normalized Hi-C contacts at 1 kb resolution were calculated surrounding the TSS and TTS or the TSS of another gene. The median of each bin was then taken and plotted as a heatmap or a 3D surface plot using the Lattice wireframe R package.

*HiC Simulations*

*Simulation in Drosophila Cells:*

The *Drosophila* simulated Hi-C matrices were created without any knowledge of three-dimensional contact or domain structure, using only information from GRO-seq and APBS occupancy (ChIP-seq). Simulated contacts between two bins relied on their correlation in transcriptional activity. We noted from Hi-C contact maps that active compartmental interactions were generally stronger than inactive compartmental interactions, thus actively correlating bin scores were increased. These contacts were then reduced based on their distance and the

number of architectural proteins lying between them. To recapitulate the noise of Hi-C data, we added matrix blurring and randomly added contacts following a Poisson and gamma distribution. Simulated contact maps without APBS incorporation (GRO-seq alone) were created by an equal decay rate across bins in lieu of APBS insulation. Simulated contacts without transcriptional activity (APBS alone) were created by replacing all transcriptional activity with null values. Simulated contact heatmaps for *A. thaliana*, *P. falciparum*, *N. crassa*, and *C. elegans* were done solely with transcriptional information.

Simulated read counts for each 1kb interaction bin in the two-dimensional Hi-C matrix were generated using a model that incorporates GRO-seq data and Architectural Protein Binding Sites (APBSs) produced from ChIP-seq peaks of individual architectural proteins. Empirical cutoffs for highly active regions (ac – active cutoff) and inactive regions (ic – inactive cutoff) were determined (1000 and 100 reads per kb respectively) and $\log_{10}$ read counts were taken as the respective value between the two and converted to a probability value with the formula: 1 – ((grocount – ic) / (ac – ic)). This maps all possible read counts to values between 0 and 1, with 0 being active and 1 being inactive.

For each pair of bins the transcriptional activity values determined above were used to create a correlation value using the following formula. The formula computes the similarity or correlation C between the two values and thus will be 0 when one bin is active and the other inactive, but 1 when both bins are active or both bins are inactive. $A_x$ and $A_y$ represent individual GRO-seq bin values calculated above.

$$C = 1 - (A_x - A_y)^2$$

A second step increases the score of bins where both anchors have some activity, doubling the score in the case where both anchors are fully active.

$C' = C * (min(A_x,A_y) + 1)$

APBSs were used to determine the insulation between 2 bins by tallying the number of ChIP-seq peaks of each protein in all the bins between any two anchors. B is the number of APBS peaks in each 1kb section of the genome. I is equal to the total number of APBS ChIP peaks between the interacting bins.

$I_{x,y} = B_i$

Each APBS peak is treated as equally important by the simulation. Ten APBS peaks in a single bin will have the same total effect on insulation as ten peaks spread across multiple bins. A constant, $\alpha$, was chosen to reflect the insulation of each bound protein. The simulations use a value of 0.982. This constant is raised to the power of the total number of intervening architectural proteins to calculate an insulation score, K, between 0 and 1.

$K_{x,y} = \alpha^{Ix,y}$

The insulation score is used to modify the correlation score, causing a distance decay, which is sharper or more gradual depending on the density of the architectural proteins. $\beta$ is a constant that is modified by the correlation and insulation scores of each x, y pair in the matrix. The simulations use a value of 40 for $\beta$.

$M_{x,y} = \beta * K_{x,y} * C'$

An additional distance-dependent factor was added to each interacting bin. The closer the two anchors the larger the value added to represent the distance decay seen in most Hi-C data. This decay follows the power law frequently observed in Hi-C datasets with a decay rate of -0.8. A constant, $\theta$, was set equal to 300 to reproduce the large number of reads near the diagonal of the Hi-C matrix.

$M'_{x,y} = M_{x,y} + \theta * (y - x + 1)^{-0.8}$

To more closely represent the realities of Hi-C data the matrix was blurred. Each bin was averaged with its surrounding bins in the matrix in order to smooth the data. A window size, w, represents the width of the blurring and was set to 3. The averaging step was carried out twice.

$M''_{x,y} = ( M'_{i,j} ) / w^2$

To achieve a source of random fluctuation in the data, a Poisson and gamma distribution were used to add random values to each bin. A gamma distribution with shape of 0.02 and scale of 10 added values to each bin creating a minority of bins with much higher than average values and a poisson distribution was used to randomize all values slightly.

$M'''_{x,y} = Poisson(M'_{x,y} + Gamma(0.02, 10))$

Simulated contact maps without APBS incorporation (GRO-seq alone) were created by an equal decay rate across bins in lieu of APBS insulation. Simulated contacts without transcriptional activity (APBS alone) were created by replacing all transcriptional activity with null values.

Correlations between actual and simulated Hi-C contact maps were done at 5 kb resolution. Distance normalized interactions crossing over, but not landing within the bin, were counted and then normalized by the sum of the interaction counts in each set. These scores were used to create Spearman correlation values. Separately, the smoothed scatter plot was created by taking each distance normalized signal between bins at 5 kb resolution comparing actual to simulated counts.

*Simulation in Human Cells*:

To simulate the human genome Hi-C map at 5 kb resolution we generated the CTCF and transcription regulated components independently and overlaid them for the 54-75 MB region on Chromosome 4 along with a genomic background function.

For the transcriptional segregation component of the combined model, as well as for the stand-alone model, the transcription level of each 5 kb bin was determined by Gro-seq data from GM12878 (GSM1480326) and was mapped to values between 0 and 1 in the same way as other simulations resulting in a correlation score *C*. A second step again increased the score of bins where both anchors had some activity, doubling the score in the case where both anchors are fully active giving *C'*. In lieu of APBS insulation a constant power law decay with the exponent -0.7 was used to decrease interaction by genomic distance. B was set to 50. The transcriptional component of the simulation at a bin is thus described by the following equation where the bin of the upstream anchors is *u* and the downstream anchor is *d*.

$$M_{x,y} = \beta * (d - u + 1)^{-0.7} * C'$$

To complete the transcriptional segregation model the genomic background function was added with θ set equal to 100.

$$M'_{x,y} = M_{x,y} + \theta * (y - x + 1)^{-1}$$

To generate the CTCF mediated component of the simulation, CTCF loops in the 54-75 MB section of Chromosome 4 were annotated manually as computational methods were unable to completely annotate CTCF loops in the region. We approximate the effects of each CTCF loop on the simulation by three patterns: increasing score in all bins between the two anchors, strong lines from each CTCF anchor in the orientation of its interacting partner, and a peak of interactions at the intersection of the two anchors.

The strength of the domains and the lines is modified by the distance between the two CTCF anchors divided by a constant larger than the largest distance between CTCF loops, 800.

$D = 0.2 * (d - u) / 800$

*D* Is thus a constant between 0 and 0.2 correlated with the distance of the CTCF loop. *D* weakens the strength of long range loops in relation to short range loops. All interaction bins within the domain bounded by the CTCF anchors are scored by the following function:

$M_{x,y} = \beta * (y - x + 1)^{-0.7 - D}$

To recreate the lines extending from the diagonal of the matrix to the CTCF loop we use *L* to represent the width of the line, thicker near the diagonal and tapering towards the CTCF loop defined by:

$L = 100 * (y - x + 1)^{-0.4}$

Any values of *L* smaller than 2 are replaced by 2. Each x,y bin within *L* distance of the line is scored by the following equation where *K* is the distance between the bin and the center of the line.

$M_{x,y} = \theta * (y - x + 1)^{-0.6} - D * (K+1)^{-0.2}$

To produce a peak of interactions at the CTCF loop every x,y bin within 10 bins of the center of the peak, u,d is scored as below. First an expected value *E* is computed:

$E = \theta * (y - x + 1)^{-1}$

*O* corresponds to the observed/expected value of the peak of the loop and is used to calculate the final value of the bin below.

$M_{x,y} = E + E * O * (|d - y| + |u - x| + 1)^{-1.5}$

These three features produce the CTCF component of the Hi-C simulations. Where they overlap, the feature that produces the maximum score is used.

Lastly a genomic background function is added to account for uniform genomic background.

$M'_{x,y} = M_{x,y} + \theta * (y - x + 1)^{-1}$

The matrix is then convolved with a Gaussian kernel of size 20 to simulate blurring due to linear proximity. A level of randomized ligations are then added to account for technical effects using a combination of Gamma and Poisson distributions to produce the final matrix.

$M''_{x,y} = Poisson(Max(0,(M'_{x,y} + Gamma(0.02,4) - Gamma(0.02,4))))$

Quantification and Statistical Analysis

Significant differences at center points between interaction metaplots were performed using a Wilcoxon signed-rank test as described in the figure legends. Significance was determined at $p < .05$.

Data and Software Availability

*HinfI* and *DpnII* Hi-C datasets for Kc167 cells have been deposited in the Gene Expression Omnibus (GEO) under the ID code GSE80702. ATAC-seq, ChIA-PET, HiChIP, and ChIP-seq data are available under the ID code GSE89244.

# Chapter 5: Dynamic compartmentalization formed by conserved forces

Michael H. Nichols and Victor G. Corces

Manuscript in preparation.

## Abstract

Chromatin is organized in the nucleus into compartmental domains defined as sequences containing proteins capable of mediating interactions among themselves. While these self-interacting contact domains are one of the most prominent features of genomic organization at the chromosome scale, we lack a nuanced understanding of the different types of compartmental domains present in chromosomes and a mechanistic understanding of the forces responsible for their formation. In this study, we compared different cell types to identify distinct paradigms of compartmental domain formation in human tissues. We identified and quantified compartmental forces correlated with histone modifications characteristic of transcriptional activity as well as previously underappreciated roles for compartmental domains correlated with the presence of H3K9me3, H3K27me3, or none of these histone modifications. We present a simple computer simulation model capable of simulating compartmental organization based on the biochemical characteristics of independent chromatin features. This model allows for dissection and quantification of chromatin features correlated with compartmental organization. Using this computational model, we show that the underlying forces responsible for compartmental domain formation in human cells are conserved and that the diverse compartmentalization patterns seen across cells are due to differences in chromatin features. We extend these findings to *Drosophila* to suggest that the same fundamental forces

are at work beyond humans. These results offer mechanistic insights into the fundamental forces driving genomic compartmentalization.

## Introduction

The highly organized nature of the eukaryotic nucleus has been evident since experiments using immunofluorescence microscopy to determine the subnuclear distribution of various proteins and histone modifications showed the existence of various types of nuclear bodies. These nuclear locations, where proteins with the same functional properties accumulate, have been described more recently as biomolecular condensates created as a consequence of liquid-liquid phase separation due to the presence of high concentrations of multivalent proteins bound to DNA and RNA, dividing the nucleoplasm into functionally distinct compartments (Banani et al. 2017). Some of these nuclear bodies appear to be involved in RNA-processing or sequestration, but others, such as the nucleolus, contain chromatin. These bodies represent distinct nuclear environments that regulate exposure of the DNA to various proteins of the nucleoplasm and are therefore essential to controlling the activity of genes. For example, active genes are present in hubs termed transcription factories where transcribed genes aggregate together with the transcriptional machinery (Jackson et al. 1993). Features of chromatin that are associated with transcriptional silencing also cluster with each other. Polycomb bodies form from the agglomeration of PRC1 and PRC2 protein complexes that epigenetically silence genes, in part by the trimethylation of H3K27 (Pirrotta and Li 2012). Additionally, transcriptionally silenced pericentric heterochromatin colocalizes within the nucleus to form chromocenters in some cells, strongly enriched for HP1a and H3K9me3 (Wang et al. 2019). Several studies have now shown the ability of several chromatin components to drive liquid-liquid phase separation in vitro and in vivo, H3K9me3 and HP1 together produce heterochromatin compartmentalization, as well as

the intrinsically-disordered regions found in PRC1, RNA Polymerase II and transcription factors (Wang et al. 2019; Plys et al. 2019; Ladouceur et al. 2020; Boijja et al. 2018).

With the advent of Hi-C it has become possible to query the organization of the entire genome at the sequence level simultaneously (Lieberman-Aiden et al. 2009). Hi-C identifies all interactions in the genome after fixation with formaldehyde. The precise mechanism of formaldehyde cross-linking is not well understood, but it may involve amino acid side chains modified with imine groups that dimerize in order to form a cross-linked product (Tayri-WIlk, 2020). This reaction is thought to be slow and requires the two reacting groups to be close and stationary relative to each other. This mechanism implies that proteins and DNA need to intimately interact with each other in order to be cross-linked by formaldehyde and that mere proximity in the nuclear space is not sufficient to observe interactions in Hi-C data, a conclusion that guides interpretation of Hi-C contact information. The resulting contact frequency maps prominently display self-associating domains formed by short-range interactions among contiguous segments of the genome and can be visualized as “triangles” present at the diagonal of Hi-C heatmaps. Classically, these contact domains are called compartments and Topologically Associating Domains (TADs). The difference between these two types of domains is not functional but rather refers to the computational approach used to define them. Compartments are defined by Principal Component Analysis (PCA), normally using Hi-C data binned at 0.5-1.0 Mb resolution and, as a consequence, are normally considered to be larger than 1 Mb in size. Compartments can contain sequences in an active (A) or silenced (B) transcriptional state and they interact with other compartments in the same state to give the plaid pattern observed in Hi-C heatmap. As a consequence, the term “compartment” is used to refer to both the self-interacting contact domains present at the diagonal as well as the ensemble of all the inter-domain interactions among all the domains in the same transcriptional state. To avoid confusion, we will use the term “compartmental domains” to refer to self-

interacting contact domains present at the diagonal of Hi-C heatmaps and “compartment” to refer to all interactions among compartmental domains in the same transcriptional state. Different from compartmental domains, TADs are defined using algorithms that detect switches in the directionality of interactions. Analysis of Hi-C data at 1 kb resolution indicates that TADs actually correspond to two different types of domains--CTCF loops and compartmental domains (Rowley and Corces 2018). CTCF loops are formed by the interruption of cohesin extrusion due to the presence of convergent CTCF-bound sites. CTCF loops can be visualized in Hi-C heatmaps by strong punctate signals at the summit of the domain, whereas compartmental domains lack this signal. CTCF loops disappear from Hi-C heatmaps obtained in cells depleted of CTCF whereas compartmental domains remain. Furthermore, compartmental domains are present in regions of the genome containing sequences in the same active or inactive transcriptional state, and can be identified by PCA using 5-10 kb bin sizes. Neighboring regions in separate compartmental domains interact less frequently and represent a compartmental switch or border. In this way, the compartmentalization of the genome creates both local compartmental domains and distant compartmental interactions.

As described above, compartmental domains can be captured by PCA of the Pearson correlation maps of each chromosome. The first principal component (PC1) or eigenvector captures the dimension with the highest variance. Using this vector the genome can be divided into two classes. These classes generally differentiate between transcriptionally active and inactive genomic regions, and so are called A and B respectively. This categorization performs well across mammalian cell types and thus the compartmentalization of the genome is generally thought of as binary. However, epigenetic information suggests that the transcriptional state of the genome is more complex, and that the two-state classification is an oversimplification.

While PCA remains the standard and most common method for calling compartmental domains, some analyses have more closely examined compartmentalization using more sophisticated techniques. The use of a Hidden Markov Model to cluster interchromosomal interactions resolved 6 subtypes of compartmental domains in GM12878 cells, of which two are enriched in different active and four contain inactive chromatin features, including post-translational histone modifications, replication timing, and measures of nucleolar and lamin association (Rao et al. 2014)(Rao 2014). Previous work in *Drosophila* has shown that the majority of compartmental organization can be recapitulated using a measure of transcriptional activity indicating a direct correlation between chromatin state and compartmental domains (Rowley et al. 2017). Other studies have also successfully used polymer simulations to reproduce the compartmental organization of the genome using chromatin-defined states. PCA-derived compartment calls in diverse cell lines and tissues invariably find A/B compartmentalization patterns, but the epigenetic features enriched in those A/B patterns can differ between cell-types. Notably, several studies have found the heterochromatin associated histone modification H3K9me3 strongly enriched in B compartments (Falk et al. 2019). However, this modification was only found enriched in a single subcompartment (B4) in GM12878 cells, which was predominantly found only on chromosome 19 (Rao et al. 2014). Additionally, the binary A/B compartmentalization of the genome is far simpler than what would be predicted from microscopic analyses of nuclei where a large variety of biomolecular condensates composed of different epigenetic features have been observed.

Here we examine two cell types, GM12878 and HCT116, with divergent compartmental definitions in an attempt to better understand the different patterns of compartmental domains seen between different human cells and investigate the potential role of the forces responsible for this aspect of genomic organization, with the goal of reconciling observations derived from Hi-C analyses and microscopy-based studies. The results suggest a consistent model of

genomic organization and offers insights into the mechanistic underpinnings of 3D genomic compartmentalization.

## Results

### Dynamic compartmentalization across human cells

To better understand the mechanisms underlying the formation of compartmental domains and their compartmental interactions, we compared high-resolution Hi-C datasets from two cell types – the lymphoblastoid GM12878 cell line and colorectal carcinoma HCT116 cells. We took the Pearson correlation of the distance-normalized interaction maps in order to display the correlation of interactions of each bin with each other bin. This method can be used to visualize compartmental domains and their interactions because genomic regions in the same compartment will have highly correlated interaction frequencies and will have a high score in the correlation map. Figures 5.1A and 5.1B show the Pearson correlation maps for chromosome 4 of GM12878 and HCT-116, respectively. This chromosome shows very different organizations between the two cell lines. Both possess clear compartmental domains along the diagonal of the map and compartmental interaction patterns as seen by the plaid pattern away from the diagonal. However, the locations and strength of these compartmental domains and the distal compartmental interactions between these domains are very different. We sought to explore whether differences in the epigenetic profiles of the chromosomes of these cell types could explain their distinct compartmentalization patterns. We compared the distribution of H3K27ac, which is correlated with transcriptional activity, H3K27me3, which is correlated with transcriptional silencing, and H3K9me3, which is also correlated with transcriptionally inactive sequences, to the Pearson correlation maps in these cells. We also performed PCA and show

the first principal component (PC1) in both cell lines (Figures 5.1A,B). Remarkably, GM12878 and HCT-116 show very different distributions of H3K9me3. In HCT-116 cells, large H3K9me3-rich domains correlate with prominent compartmental domains that are strongly correlated with each other in their Hi-C interaction frequencies. GM12878, on the other hand, lacks these large H3K9me3-rich domains and correspondingly lacks the prominent compartmental domains associated with them.

To quantify these observations, we used PC1 to call A and B compartments in GM12878 and HCT-116 cells, and measured the relative enrichments of chromatin features on their chromosomes (Figure 5.1C). In GM12878 cells, A/B compartmentalization strongly follows transcriptional activity/inactivity, with histone modifications associated with active transcription such as H3K27ac, H3K4me, and H3K36me3 all enriched in the A compartment and depleted in the B compartment of chromosome 4, while modifications associated with silenced chromatin such as H3K27me3 show an inverse pattern. In HCT116 cells, however, A compartments in chromosome 4 are not strongly enriched for histone modifications associated with transcriptional activity, including Gro-seq, which is a direct measure of transcription (Figure 5.1C and Supplemental Figure 5.1A). Instead, binary compartmental delineation using PC1 divides the chromosome into a transcriptionally inactive H3K9me3-rich portion and the remainder, which consists of both transcriptionally active as well as inactive regions. In contrast, H3K9me3 in chromosome 4 of GM12878 is enriched in the A compartment and presents very differently on the chromosome as sporadic peaks rather than contiguously enriched domains.

Given that the B compartment in chromosome 4 of HCT-116 cells is depleted of transcriptionally active sequences, we asked why its corresponding A compartment is not strongly enriched for transcribed sequences. A simple explanation for this phenomenon is that transcriptionally inactive regions of HCT-116 not containing H3K9me3 correlate more closely in their interaction

frequencies with transcriptionally active regions. Thus, the A compartment in these cells as defined by PCA is composed of a conglomeration of all H3K9me3-poor chromatin, both transcriptionally active and inactive, leading to only a mild enrichment for active marks.

To examine more closely the mechanisms underlying the formation of compartmental domains we focused on a 65-95 Mb region on chromosome 4 and used a resolution of 25 kb to call compartments using PCA (Figures 5.1E and 5.1F). In HCT-116 cells, this region contains instances of all 4 clusters found in chromosome 4. Compartmental domains present in the A compartment defined by PCA in GM12878 cells are highly enriched in H3K27ac with respect to those in the B compartment, whereas H3K27me3 and H3K9me3 are similarly enriched in both A and B compartments (Figures 5.1E and 5.1G). However, in HCT-116 cells there is a clear enrichment of both H3K9me3 and H3K27me3 in the B compartment corresponding to compartmental domains and interactions absent in GM12878 cells (Figures 5.1F and 5.1G). These findings, showing differential enrichment of active and repressive histone modifications in the A and B compartments in different cell lines, are surprising, since it is generally assumed that the A compartment contains transcriptionally active genes and the B compartment is enriched in silenced sequences. This suggests that the canonical binary classification of A and B compartments is insufficient to represent the properties and mechanisms by which compartmental domains form in these cells.

Conserved principles underlie dynamic compartmentalization

To further explore the complex compartmentalization logic observed in GM12878 and HTC-116 cells, which appears to follow different rules in the two cell lines and cannot be explained by a simple binary division of PC1, we employed the unsupervised k-means clustering algorithm to identify compartmental clusters in both cell types. The primarily binary A/B organization of

chromosome 4 in GM12878 cells obtained by PCA can be reproduced by two clusters obtained via unsupervised k-means clustering (Figures 5.1A and 5.1E). However, four clusters are required to produce a meaningful classification of chromosome 4 that correlates with the Hi-C heatmap in HCT-116 cells (Figures 5.1B and 5.1F). As expected, one of these four clusters corresponds directly to H3K9me3-rich regions whereas a second one correlates strongly with H3K27ac. Surprisingly, the two other clusters are both transcriptionally inactive with distinct chromatin features, one highly enriched for H3K27me3, and the last lacking all three histone modifications (Figures 1D and 5.1H). Only H3K9me2 is enriched in this compartment (Supplemental Figure 5.1A). Therefore, chromosome 4 in HCT-116 appears to have four distinct compartmental domains. Interactions among each type give rise to the complex plaid pattern in the Hi-C heatmap, forming compartment A (transcriptionally active), B (H3K27me3-rich), C (H3K9me3-rich), and D (enriched in H3K9me2 but depleted of standard active and silencing histone modifications). We note that the enrichments of epigenetic features in chromosome 4 A and D clusters of HCT-116 cells correspond well to the A and B compartments of chromosome 4 in GM12878 cells, with the exception of H3K9me3 enrichment in the A compartment of GM12878 (Figures 5.1C and 5.1D).

We then sought to understand why H3K9me3 is enriched in different compartmental domains in HCT-116 versus GM12878 cells. The distribution of H3K9me3 in these two cell types is very different, with GM12878 chromosomes typically having narrow peaks of signal whereas HCT-116 chromatin tends to have large consistently-enriched plateaus. At least two possible hypotheses could explain these different distribution patterns. One possibility is that H3K9me3 regions compartmentalize differently in the two cell types due to different nuclear environments determined by cell identity and physiology. A second explanation is that H3K9me3 regions compartmentalize differently due to distinct distributions of this histone modification as a consequence of transcriptional differences between the two cell types. Analysis of chromosome

19 offers an opportunity to distinguish between these two possibilities, since this chromosome contains large domains of H3K9me3 in both GM12878 and HCT-116 cells. Chromosome 19 of GM12878 cells shows a strong correlation between the presence of H3K9me3 and the formation of strong compartmental domains (Figure 5.1I), as was seen on chromosome 4 of HCT-116 cells, and these H3K9me3 compartmental domains are similar in chromosome 19 of both cell types (Figure 5.1J). The similarity between H3K9me3 domains in chromosome 19 of GM12878 and HCT-116 cells suggests that the nuclear environment is not responsible for the differences observed in other chromosomes. Strikingly the histone modification profiles and the A/B compartments defined by PC1 also closely match between the two cell lines (Figure 5.1K). k-means clustering was then performed in both cell-types with 3 clusters as chromosome 19 lacks large regions devoid of any signals, which would fall into the D cluster. The resulting cluster calls closely match each other (Figure 5.1L). The high correspondence in k-means cluster definitions is mirrored by the relative signal enrichments in each cluster. In contrast to chromosome 4, the compartments called by PCA and by k-means showed similar enrichments for various histone modifications in chromosome 19 (Figure 5.1L and Supplemental Figure 5.1C). The alignment of chromosome 19 in both - histone modification profiles and compartmental organization - in HCT-116 and GM12878 fits a model where these histone modifications, or a chromatin feature correlated to them, drive chromosomal compartmentalization, and that the underlying forces driving compartmentalization are consistent between these cell types. We suggest that the diverging compartments seen in chromosome 4 are the result of distinct chromatin profiles and that the resulting conflicting histone modification enrichments for the A and B compartments of chromosome 4 between these cell lines do not reflect differences in the underlying chemistry driving compartmentalization. We suggest that the enrichment of H3K9me3 in the active compartment of GM12878 chromosome 4 can be explained as the inability of small narrow peaks of H3K9me3 to drive compartmentalization against entropic mixing. Their proximity to other active

compartmentalizing features may further inhibit their self-segregation. Alternatively, since these short regions containing H3K9me3 are much smaller than the 25 kb bins used to perform the clustering analyses, it is possible that compartmental calls performed at a resolution higher than the size of these H3K9me3 regions would place these sequences in the same compartment as the large H3K9me3-containing large blocks. This possibility highlights the potential biases introduced by the resolution used in the analyses when interpreting Hi-C data.

We considered whether either of these cell lines perhaps exemplify an outlier unrepresentative of normal human tissues and thus we investigated H3K9me3 distribution in numerous immortalized and primary human cells. We found that neither H3K9me3 pattern observed in GM12878 and HCT-116 cells is unrepresented in other cell types or in primary tissues but rather exemplifies two ends of a continuum of H3K9me3 present across human cells (Supplemental Figure 5.1 D,E). These findings reveal that H3K9me3 domains are highly dynamic and represent significant changes in compartmentalization patterns across development.

Compartmental domains correlate directly with chromatin features

With the understanding that binary models are insufficient to represent human compartmentalization and that the forces responsible for the formation of compartmental domains appear to correlate closely with chromatin features, particularly transcriptional activity and H3K9me3, we sought to visualize and quantify the forces driving compartmentalization. To this end, we sorted the Pearson correlation map of chromosome 14 from HCT-116 cells by various features (Figure 5.2A). These assemblies are produced by reordering the rows and columns of the correlation matrix so that instead of being placed in their natural order bins are arranged by increasing signal of the chosen feature. This allows us to compare the ability of a particular sequence to form a specific type of compartmental domain with its epigenetic

features. This approach also provides a means of visualizing the frequency of interactions within and among compartmental domains or “compartmentalization strength”. Sorting by PC1 values we observe a clear segregation of chromosome 14 into three compartments A, B, and C (Figure 5.2B). Surprisingly, this indicates that PC1, while typically used to delineate binary compartments, could be used to call more compartmental domain types with different thresholds. Chromosome 14, like chromosome 19, largely lacks regions that would fall into the D compartment. Sorting the chromosome by H3K9me3 reproduces the compartmentalization of the H3K9me3 rich C compartment as well as the first principal component itself, however, it was unable to distinguish between the A and B compartments (Figure 5.2C). Sorting by H3K27ac as a marker of transcriptional activity also results in compartmental segregation, although this is less precise (Figure 5.2D). This is likely a function of H3K27ac signal lacking the consistency and continuity of H3K9me3 domains within transcriptionally active regions. Sorting by H3K27me3 organized only the regions most enriched for H3K27me3 and was unable to organize the rest of the chromosome (Figure 5.2E). Together these results show that covalent histone modifications strongly correlate with and are predictive of compartmental organization.

An unusual feature of chromosome 14 is the existence of a trimodal distribution of H3K9me3 (Supplemental Figure 5.2A). While most chromosomes in HCT-116 cells possess a bimodal distribution of H3K9me3 signal leading to H3K9me3 rich and poor regions, chromosomes 13 and 14 have distinct strong and weak H3K9me3 domains (Figure 5.2A). Comparing their correlations in the sorting of chromosome 14 by H3K9me3 shows that weak-H3K9me3 and strong-H3K9me3 regions correlate better with regions of similar strength. This phenomenon, along with the varying correlation strength seen in the heatmaps obtained by sorting various features, suggests that compartmentalization is more accurately thought of as a quantitative rather than a categorical feature of the genome and that classification of chromatin into

categories is a simplification that neglects the effects of varying epigenetic signal strengths in the formation of compartmental domains.

Sorting by a single signal does not result in perfect compartmentalization because multiple, independent forces drive the process of compartmental domain formation and long-range interactions among domains of the same type. The most significant of these forces may be those involved in interactions between sequences containing H3K9me3 and proteins and histone modifications associated with active transcription. We favor a model of compartmentalization in which the genome is organized by multiple independent forces directly correlated with chromatin features that attract and repel each other.

Independent contributions of chromatin features can reproduce compartmental organization

We next sought to test our hypothesis that formation of compartmental domains and establishment of long-range interactions among domains of the same type is driven by the independent contributions of chromatin features. To approach this question, we created a machine learning model to reproduce Hi-C interaction maps using epigenetic features. Results described above suggest that the presence of H3K27ac as an indicator of transcriptionally active regions, H3K27me3 and H3K9me3 as indicators of different types of silenced sequences, or the absence of these three modifications can account for all possible compartmental domains present in chromosomes of human cells. To enable comparison across cell types and experiments we first binned these three epigenetic signals into quantiles at 100 kb resolution. For each normalized signal, an algorithm then learned using a Maximum Likelihood Estimation approach an attraction-repulsion relationship for each pair of quantiles. This attraction-repulsion mapping effectively represents the average enrichment or depletion between all bins with the corresponding level of signal. The estimated contact frequency in the simulated map is then

derived by the simple addition of the estimated effect of each signal and multiplied by a distance-dependent constant representing the average interaction frequency at each genomic distance (Figure 5.3A). This relatively straightforward model, which is only capable of representing the independent attractive and repulsive forces of each chromatin feature, tests to what extent this framework is capable of recapitulating the 3D organization of the genome represented by the compartmental domains and their interactions.

The averaged attraction-repulsion relationships learned from every chromosome of HCT-116 are shown (Figure 5.3B). The learned relationship between the levels of a specific histone modifications and the interaction frequency of the corresponding sequence is similar to that observed experimentally described above (Figures 5.2C-2E). Genomic regions high in a given histone modification show increased interactions with other regions high in that same mark. The model learns and predicts that pairs of regions in which one is high in such a signal and the other low will not be attracted and have reduced interaction frequency. The minimal model using three histone modifications, H3K27ac, H3K27me3, and H3K9me3, is able to recapitulate most aspects of 3D genome organization while remaining easy to interpret. All three signals show a degree of attraction between the highest quantile bins, as seen by the enrichment in the bottom right of the attraction-repulsion maps, as well as repulsion between the highest and lowest quantiles as seen by the depletion in the upper right and bottom left corners (Figure 5.3B). Importantly the strength and nature of these maps differed significantly indicating the forces driving compartmentalization differ for each feature. H3K9me3 maps show strong attraction amongst the most enriched quarter of the genome, which strongly repels the rest of the genome equally. This shows there exists a single critical threshold of H3K9me3 density and quantity and reflects the generally bimodal distribution of this modification in HCT-116. H3K27ac, on the other hand, shows the greatest attraction between the highest quantiles with a more gradual reduction in attraction with reduced signal. H3K27me3 primarily shows only attraction between

the highest quantiles of signal and otherwise contributes little to the organization of the genome. When interpreting this information, it is important to consider that the distributions of these three histone modifications are inter-related, and that regions of the genome lacking one of the modifications may contain one of the other two.

Simulations were generated at 100 kb resolution using the average of the attraction-repulsion maps learned from every chromosome except the one being simulated. A comparison of observed and simulated maps reveals close agreement on the majority of large compartmental features (Figure 5.3C,D). We quantified the accuracy of the model using the Pearson correlations between the observed and simulated maps after dividing by the average distance. Due to the power-law decay of interaction frequency with respect to distance in Hi-C maps any simulation which accurately reproduces this decay will have a high correlation. As this would not represent the ability of the model to reproduce compartmental organization, we normalized for distance to eliminate the natural correlation driven by the accurate representation of the distance decay. Simulations of chromosomes using the average maps derived from all other chromosomes varied in correlation by chromosome, but generally performed well with correlations in the range 0.5-0.7. Using this same methodology to compare biological replicates of a Hi-C experiment resulted in similar ranges of correlation scores across the chromosomes (Supplemental Figure 5.3C). The fact that Hi-C maps can be predicted by only modeling the attraction and repulsion of chromatin features against themselves suggests a direct role in these features, or a chromatin component correlated with these features, in compartmentalizing the nucleus.

Conserved forces give rise to diverse genomic organizations

Given the accuracy of our model in reproducing the 3D organization of HCT-116 cells, we then applied the same model to GM12878 cells. These cells have different distribution of H3K9me3, and their 3D genome organization as seen in PCA enrichment analysis is also different (Figure 5.1C). We found that, with the exception of H3K9me3, the attraction-repulsion maps learned in GM12878 were strikingly similar to those learned in HCT-116 (Figure 5.4A). The incongruence of H3K9me3 attraction-repulsion maps between these cell types was expected as the distribution of this signal is very different between most of their chromosomes. These maps are also able to reproduce the Hi-C interaction maps of other chromosomes of GM12878 as we have shown in HCT-116 (Figure 5.4B, Supplemental Figure 5.4A,B). However, simulating chromosome 19, which is unique in GM12878 due to its large H3K9me3 domains was less successful when using the attraction-repulsion maps learned from the rest of the chromosomes, with the simulation failing to capture the role of H3K9me3 in compartmentalization (Figure 5.4C). Pearson correlations between the real and simulated maps were substantially lower than the simulations of HCT-116, most likely due to the absence of the strong organizing feature of H3K9me3 (Supplemental Figure 5.4C). Nevertheless, the similarity between the attraction-repulsion maps of chromosomes from these two cell types suggests that attraction-repulsion maps learned in one cell type can accurately model the organization of another cell-type. If true, this would indicate that the fundamental forces underlying compartmentalization are largely conserved across human cell types and that dynamic compartmentalization is a consequence of differences in the distribution of histone modifications and their associated proteins.

We then used the attraction-repulsion maps learned from HCT-116 chromosomes to simulate the 3D organization of GM12878 chromatin, and found a high correspondence between the observed and simulated maps (Figure 5.4D, Supplemental Figure 5.4D,E). This is remarkable given that the Hi-C interaction maps of most chromosomes between these cell types are very different. This ability for compartmental forces learned from one cell type to successfully predict

compartmentalization in another indicates that the same underlying forces directly correlated with chromatin features are largely conserved between these two cell types. This is further support that the radical differences between the respective Hi-C maps of these cell lines are a consequence of different chromatin features, primarily the presence and absence of large H3K9me3-rich domains.

As the simulation of chromosome 19 of GM12878 cells using attraction-repulsion maps learned with the rest of the chromosomes was poor, we asked whether chromosome 19 could be better simulated by the attraction-repulsion relationships learned from HCT-116 than from GM12878. The histone modification profiles of chromosome 19 in GM12878 cells, particularly the presence of large H3K9me3 domains, more closely resemble the patterns seen in most chromosomes of HCT-116 cells (Figure 5.1G). Indeed, the simulation of chromosome 19 is significantly more accurate using HCT-116 attraction-repulsion maps (Figure 5.4E). In agreement, the Pearson correlations between the observed and simulated maps of chromosome 19 are higher when simulated with HCT-116 (Supplemental Figure 5.4F). These results again support the idea that the 3D organization of chromosomes is a consequence of the distribution patterns of one-dimensional epigenetic information.

Transcriptional activity and H3K9me3 also compartmentalize the *Drosophila* genome

Given the ability of our model to reproduce the 3D organization of multiple human cell types from a single universal set of attraction-repulsion maps we sought to determine how applicable this model could be outside of humans. Our previous foundational work showed that transcriptional activity was predictive of compartmental organization in multiple representative Eukaryotic genomes (Rowley, 2017). We hypothesized that, just like in humans, the same forces responsible for compartmental domain formation and the establishment of distinct

nuclear compartments via interactions among compartmental domains with the same epigenetic features may be at work in other organisms. Thus, we applied our model to explore whether the distribution of H3K27ac, H3K27me3, and H3K9me3 could reproduce the 3D organization of the *Drosophila* genome for which high-resolution Hi-C data is available.

Hi-C maps from *Drosophila* Kc cells were simulated using the attraction-repulsion model. Due to the higher read count and smaller genome size we were able to simulate the genome at 10 kb resolution. A small modification was made to the model to limit the simulation to less than 2 Mb as *Drosophila* compartmental domains are smaller than those of mammals and long-range interactions among domains decay rapidly beyond this distance. While distinct, the learned attraction-repulsion maps resemble those learned in human cells (Figure 5.5A). The model learned attraction between features with similar histone modifications and repulsion between genomic regions with dissimilar ones. This indicates that, just as in humans, these chromatin features or others strongly correlated with them drive the formation of compartmental domains and their interactions. These observations suggest a universal model of compartmentalization in which the fundamental underlying forces driving this process are conserved between organisms.

As *Drosophila* only has two autosomes of significant size we simulated each chromosome with the attraction-repulsion maps of the other. The resulting simulations largely reproduce the compartmental patterning of these chromosomes (Figure 5.5B, C). Moreover, high resolution maps indicate that the simulation reproduces short range interactions within compartmental domains.

# Discussion

Interphase chromosomes of vertebrates are generally thought to be organized into large domains whose sequences can be in one of two states--active or inactive. Within these large domains are smaller TADs, some of which are flanked by CTCF sites in convergent orientation, and therefore correspond to CTCF loops formed by cohesin extrusion, whereas others lack CTCF at their boundaries and are thus formed by different, unknown mechanisms. Smaller domains termed sub-TADs can also be observed within TADs. Results reported here address the question of what the unit of eukaryotic chromosome organization is and how we can explain its formation using known biochemical and biophysical forces operating in the nucleus. Answers to this question provided by chromosome conformation studies should account for results obtained using microscopy or biochemical approaches. However, results from Hi-C studies seem to contradict well-established concepts in nuclear biology. First, non-transcribed regions of the genome do not simply interact with each other, as one would conclude from the checkerboard pattern observed in Hi-C heatmaps representing long-range interactions among sequences located in the B compartment. Rather, non-transcribed regions in the genome contain either H3K27me3/PCR1/PCR2, H3K9me3/HP1, or lack any characterized histone modifications. Immunofluorescence localization experiments show that Pc-containing regions interact with each other to form Pc bodies, both in Drosophila and mammals (Pirrotta and Li 2012). The same is true for regions containing H3K9me3 and HP1 that form chromocenters and actively transcribed regions to form transcription factories (Jackson et al. 1993). Furthermore, recent results indicate that multivalent proteins present in these three types of genomic regions are able to form biomolecular condensates by liquid-liquid phase separation. Therefore, compartmental domains and their interactions detected by Hi-C must be more complex than is generally assumed and this complexity must reflect existing observations from microscopy and biochemistry.

Results presented here reveal an underappreciated diversity of compartmental domains, the conserved forces underlying their establishment, and the long-range interactions responsible for the formation of nuclear compartments. Multiple independent forces organize the genome into compartments. Transcriptional activity and H3K9me3 correlate with the strongest of these forces. H3K27me3, while weaker, also correlates directly with compartmental patterns. Finally, sequences lacking any of these histone modifications and enriched in H3K9me2 represent a fourth class of sequences that form their own independent compartment. The dynamic nature of these chromatin features across cell types allows for dynamic compartmentalization during cell differentiation.

While both transcriptional activity and H3K9me3 have been previously reported as highly correlated with the formation of nuclear compartments, here we show that each of the biochemical forces associated with the presence of these histone modifications drives compartmentalization independently and can do so in the absence of the other (Lieberman-Aiden et al. 2009; Falk et al. 2019). Previous in-depth Hi-C analyses of GM12878 cells suggested a unique H3K9me3-correlated compartmentalization of chromosome 19 and categorized it as a subcompartment of inactive B chromatin (Rao et al. 2014). Here we show that this compartment, while unique in GM12878, is widespread in other cell lines and is the strongest compartmentalizing force wherever large domains of H3K9me3 are found. On chromosomes where large H3K9me3-rich domains exist the segregation between these domains and the rest of the H3K9me3-poor chromosome represent the strongest feature of the Hi-C maps. As such, using PCA to delineate binary compartments on the Pearson correlation maps divides the genome into H3K9me3-rich and poor, rather than along the expected lines of transcriptional activity. The inclusion of H3K9me3-poor transcriptionally inactive regions into the

A compartment defined by PCA in these cells leads to incompatible and confusing definitions of genomic compartments between cell lines and tissues.

While PCA is a powerful tool to investigate compartments the first principal component (PC1) is canonically reduced to a binary classification of compartments that we have shown is inadequate to represent compartmentalization of the genome. We suggest shifting away from naive unsupervised classification techniques in single cell lines to a categorization informed by the breadth of organizational diversity seen across human samples, which will improve the accuracy and generalizability of future studies. Our findings fit a model of compartmentalization where attraction of similar chromatin states drives interactions between them to the exclusion of other chromatin types. The dependence of H3K9me3 compartmentalization in GM12878 cells on the levels of this modification, where small discrete peaks fail to strongly compartmentalize through most of the genome while large domains seen on chromosome 19 do, suggests the existence of a compartmentalization threshold. The forces driving compartmentalization must overcome the entropy of mixing in the nucleus and, therefore, some minimum quantity of compartmentalizing signal must exist below which the attractive forces at work are insufficient to overcome entropy. We propose that the distinctive patterns of H3K9me3 in GM12878 reflect this threshold where the punctate peaks seen in most of the genome fail to strongly compartmentalize while in the same nuclear environment the larger domains present in chromosome 19 do. As further evidence of the quantity-dependent nature of compartmentalization, several chromosomes in HCT-116 possess both weak and strong H3K9me3 domains, which correspondingly compartmentalize weakly and strongly. These findings suggest that compartmentalization is more accurately represented as a continuum, where each sequence is driven to interact according to the strength, corresponding to the quantity, of chromatin forces driving it, rather than discrete chromatin types.

Our model provides insight into the fundamental mechanisms of genomic organization. That the compartmentalization of the genome can be predicted using just a few histone modifications strongly implies that these chromatin features are either directly or indirectly responsible for the separation of chromatin types in the genome. Our finding that the patterns of attraction and repulsion are largely consistent between cell types with divergent compartments shows that these underlying forces behave consistently and that the fundamental forces shaping chromatin organization are steady between cells. The ability of the same model to reproduce the organization of *Drosophila* chromosomes suggests that the attraction and repulsion of chromatin by the independent contributions of compartmentalizing forces may be a universal driver of compartmentalization across Animalia. Taken together, our results reshape our understanding of human compartments from largely static, binary classes to a highly dynamic and quantitative continuum. The widespread use of canonical A/B compartmentalization is largely a product of the pioneering work done in GM12878, but which does not reflect the epigenetic diversity of human cell types. Informing future analyses of compartments with this understanding and approaching compartmental organization from the perspective of chromatin state driven attraction and repulsion will allow for reproducible and comparable definitions of compartments across human tissues and beyond to other organisms.

# Figures

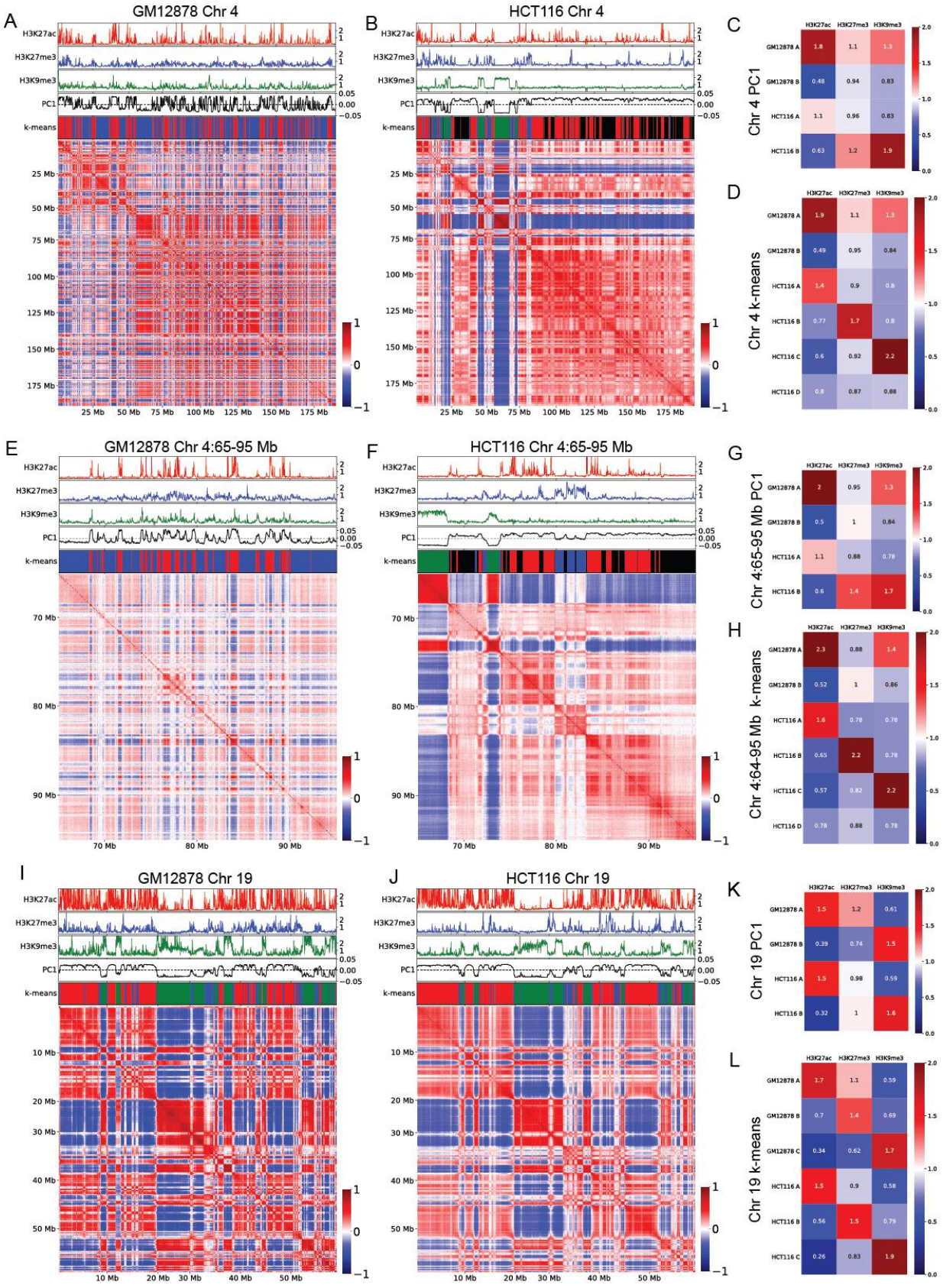

A
GM12878 Chr 4
B
HCT116 Chr 4
C
Chr 4 PC1
D
Chr 4 k-means
E
GM12878 Chr 4:65-95 Mb
F
HCT116 Chr 4:65-95 Mb
G
Chr 4:65-95 Mb PC1
H
Chr 4:64-95 Mb k-means
I
GM12878 Chr 19
J
HCT116 Chr 19
K
Chr 19 PC1
L
Chr 19 k-means
H3K27ac
H3K27me3
H3K9me3
PC1
k-means

Figure 5.1. Divergent compartmentalization between GM12878 and HCT-116

Pearson correlations of distance normalized Hi-C interaction frequency map of various regions in GM12878 and HCT-116. On top of each Hi-C map from top to bottom: fold-change over control shown above for H3K27ac (red), H3K27me3 (blue), H3K9me3 (green); PC1 (black), and k-means cluster classifications A (red), B (blue), C (green), D (black).

A) GM12878's chromosome 4.

B) HCT-116's chromosome 4.

C) Fold-enrichment of each histone modification within each compartment defined by PCA chromosome 4.

D) Fold-enrichment of each histone modification within each compartment defined by k-means clustering on chromosome 4.

E) GM12878's chromosome 4 65-95Mb.

F) HCT-116's chromosome 4 65-95Mb.

G) Fold-enrichment of each histone modification within each compartment defined by PCA on chromosome 4 65-95Mb.

H) Fold-enrichment of each histone modification within each compartment defined by k-means clustering on chromosome 4 65-95Mb.

I) GM12878's chromosome 19.

J) HCT-116's chromosome 19.

K) Fold-enrichment of each histone modification within each compartment defined by PCA on chromosome 19.

L) Fold-enrichment of each histone modification within each compartment defined by k-means clustering on chromosome 19.

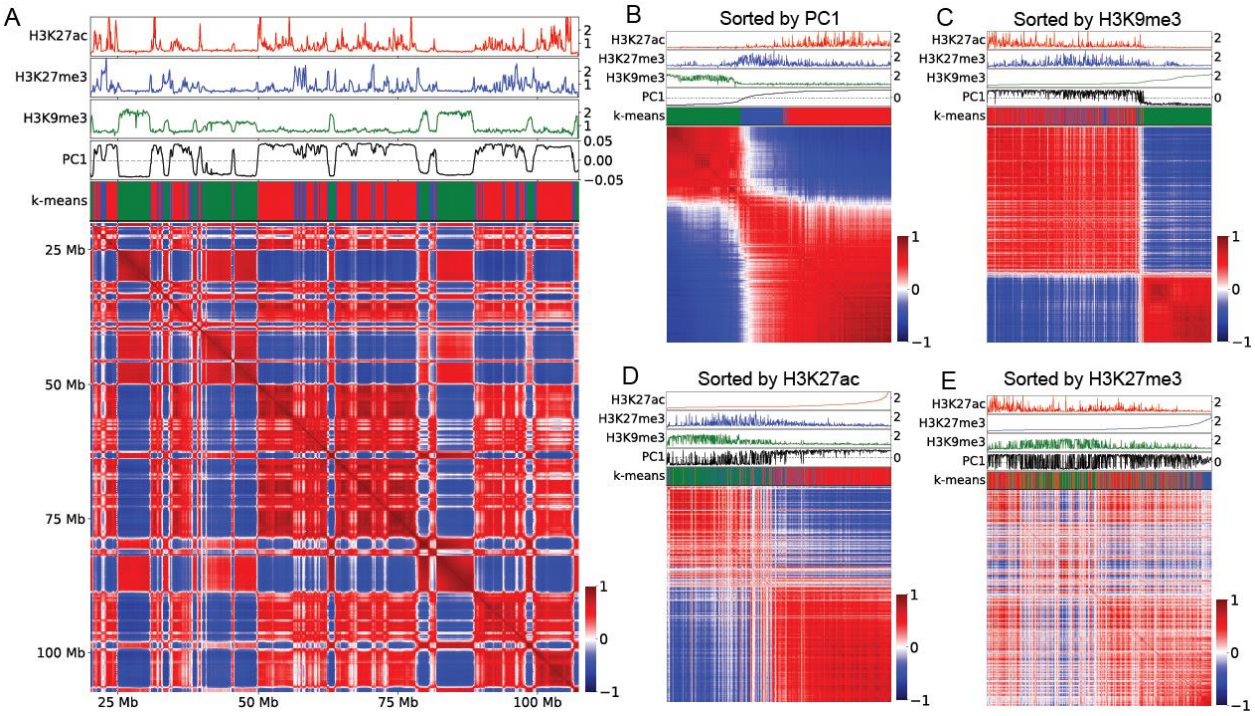



Figure 5.2. Chromosome sortings of HCT-116 chromosome 14.

Pearson correlations of distance normalized Hi-C interaction frequency map of HCT-116's chromosome 14. On top of each Hi-C map from top to bottom: fold-change over control shown above for H3K27ac (red), H3K27me3 (blue), H3K9me3 (green); PC1 (black), and k-means cluster classifications A (red), B (blue), C (green).

A) Chromosome 14 in its natural order.

B) Chromosome 14 sorted according to PC1 from lowest to highest.

C) Chromosome 14 sorted according to H3K9me3 from lowest to highest.

D) Chromosome 14 sorted according to H3K27ac from lowest to highest.

E) Chromosome 14 sorted according to H3K27me3 from lowest to highest.

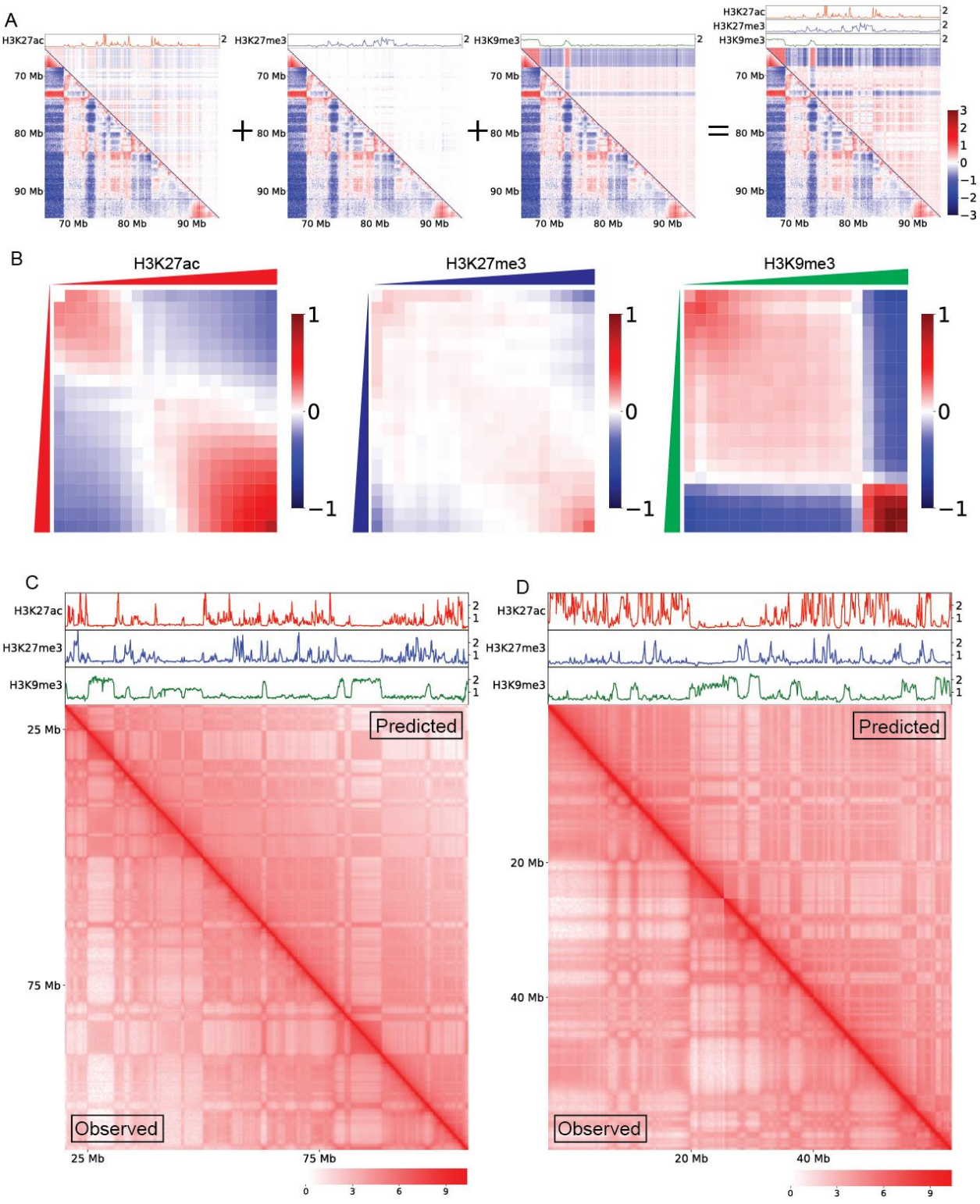



Figure 5.3. Histone modifications can predict compartmentalization using learned attraction-repulsion relationships.

A) Log(observed/expected) of Hi-C interaction maps of HCT-116 chromosome 4 65-95Mb. . Bottom left triangles are observed Hi-C interactions maps while upper right triangles are the simulation using only the components shown above as tracks. From left to right H3K27ac (red), H3K9me3 (green), H3K27me3 (blue), and all three combined.

B) Average of attraction-repulsion relationship map learned by Maximum Likelihood Estimation from every chromosome of HCT-116 for H3K27ac.

C) Average of attraction-repulsion relationship map learned by Maximum Likelihood Estimation from every chromosome of HCT-116 for H3K27me3.

D) Average of attraction-repulsion relationship maps learned by Maximum Likelihood Estimation from every chromosome of HCT-116 for H3K9me3.

E) Comparison of HCT-116 chromosome 14 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated.

F) Comparison of HCT-116 chromosome 19 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated.

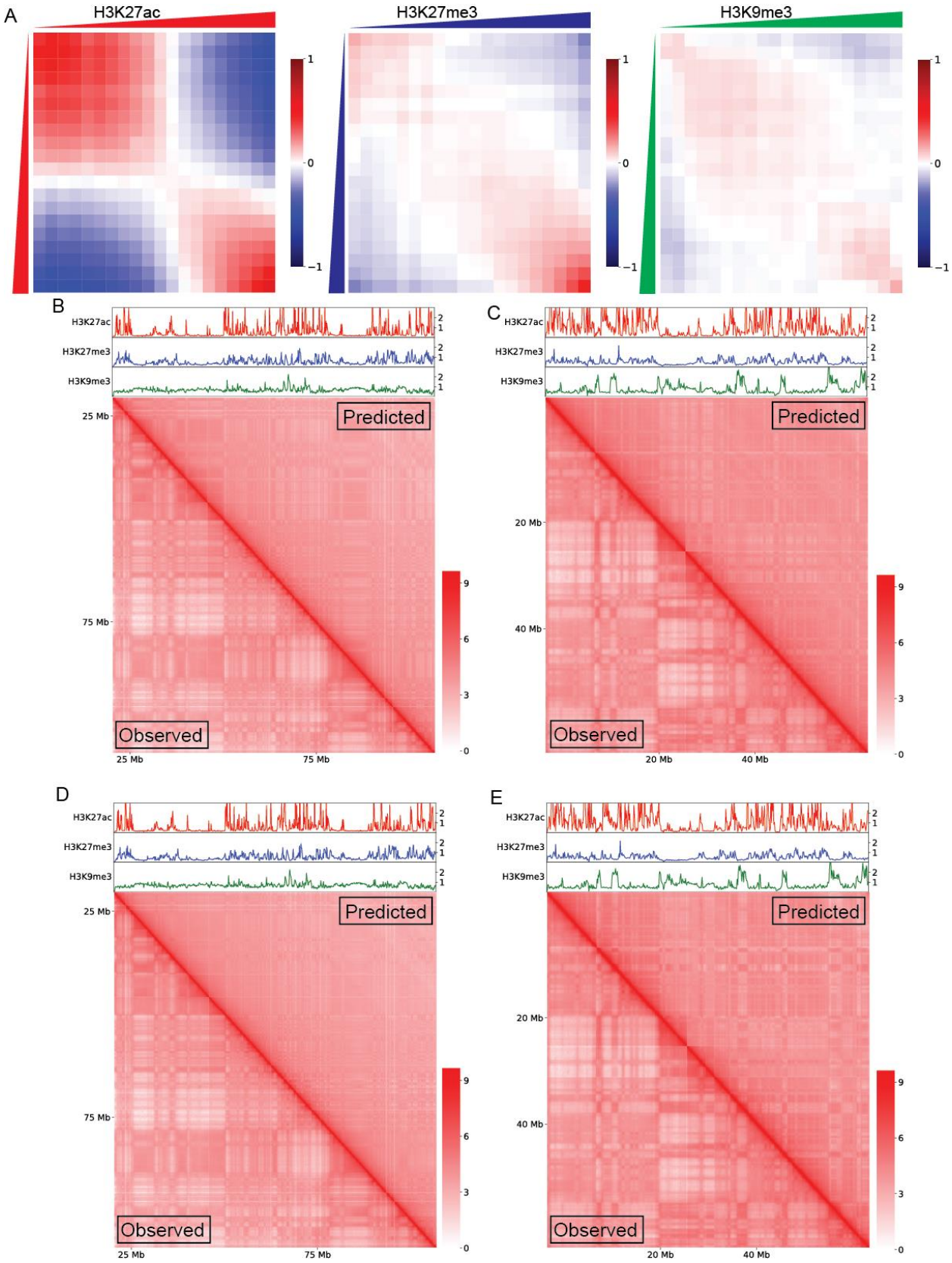



Figure 5.4. Attraction-repulsion relationships are consistent across cell types.

A) Average of attraction-repulsion relationship map learned by Maximum Likelihood Estimation from every chromosome of GM12878 for H3K27ac.

B) Average of attraction-repulsion relationship map learned by Maximum Likelihood Estimation from every chromosome of GM12878 for H3K27me3.

C) Average of attraction-repulsion relationship map learned by Maximum Likelihood Estimation from every chromosome of GM12878 for H3K9me3.

D) Comparison of GM12878 chromosome 14 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated using .

E) Comparison of GM12878 chromosome 19 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated using attraction-repulsion maps learned from GM12878.

F) Comparison of GM12878 chromosome 14 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated using attraction-repulsion maps learned from HCT-116.

G) Comparison of GM12878 chromosome 19 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated using attraction-repulsion maps learned from HCT-116.

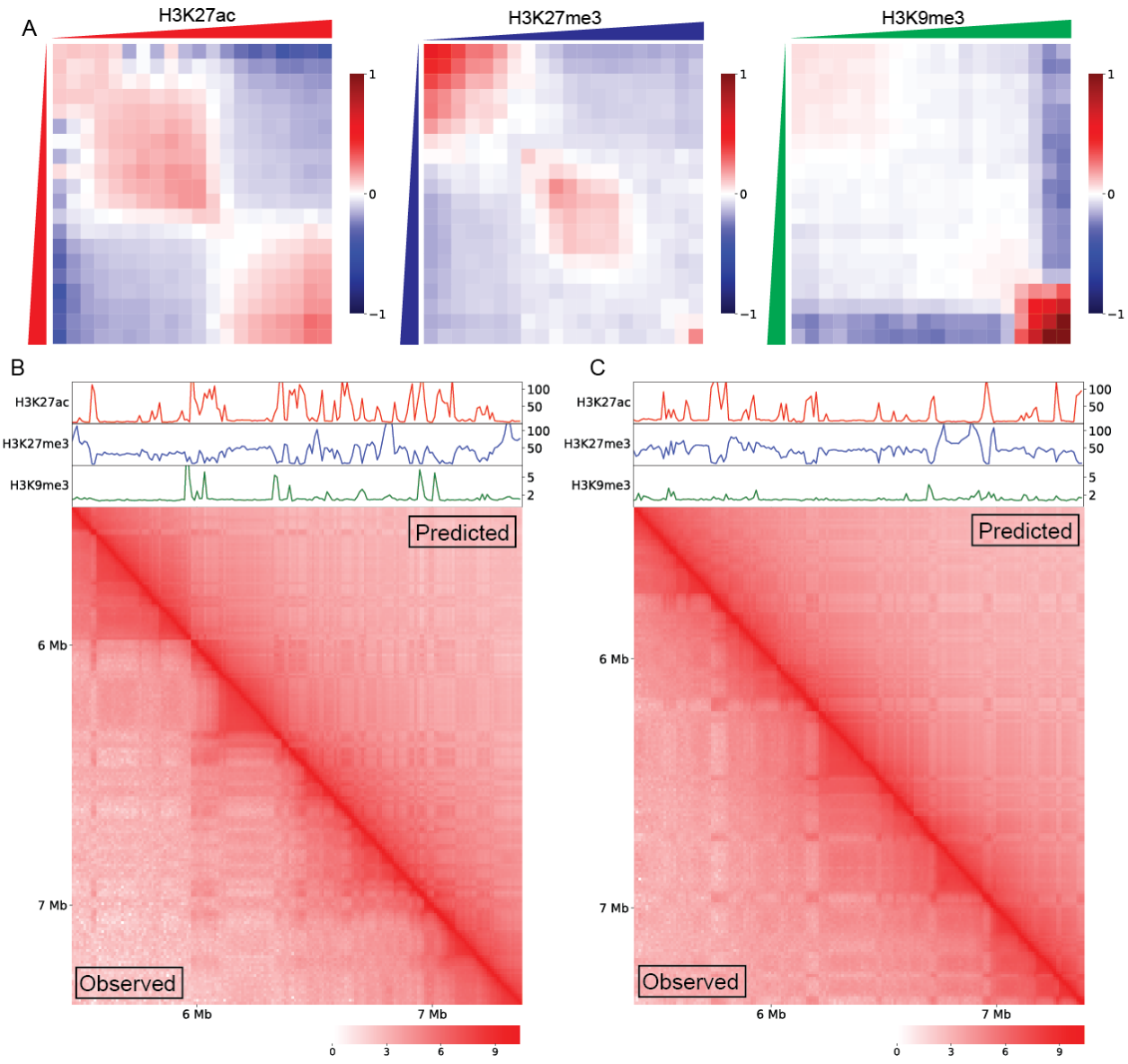



Figure 5.5. Attraction-repulsion relationships explain compartmentalization in Drosophila

A) Average of attraction-repulsion relationship map learned by Maximum Likelihood Estimation from chromosomes 2 and 3 of Kc167 for H3K27ac.

B) Average of attraction-repulsion relationship map learned by Maximum Likelihood Estimation from chromosomes 2 and 3 of Kc167 for H3K27me3.

C) Average of attraction-repulsion relationship map learned by Maximum Likelihood Estimation from chromosomes 2 and 3 of Kc167 for H3K9me3.

D) Comparison of Kc167 chromosome 2 5-6.8 Mb logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated using attraction-repulsion maps learned from Kc167 chromosome 3.

E) Comparison of Kc167 chromosome 3 5-6.8 Mb logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated using attraction-repulsion maps learned from Kc167 chromosome 2.

Supplemental Figures

A

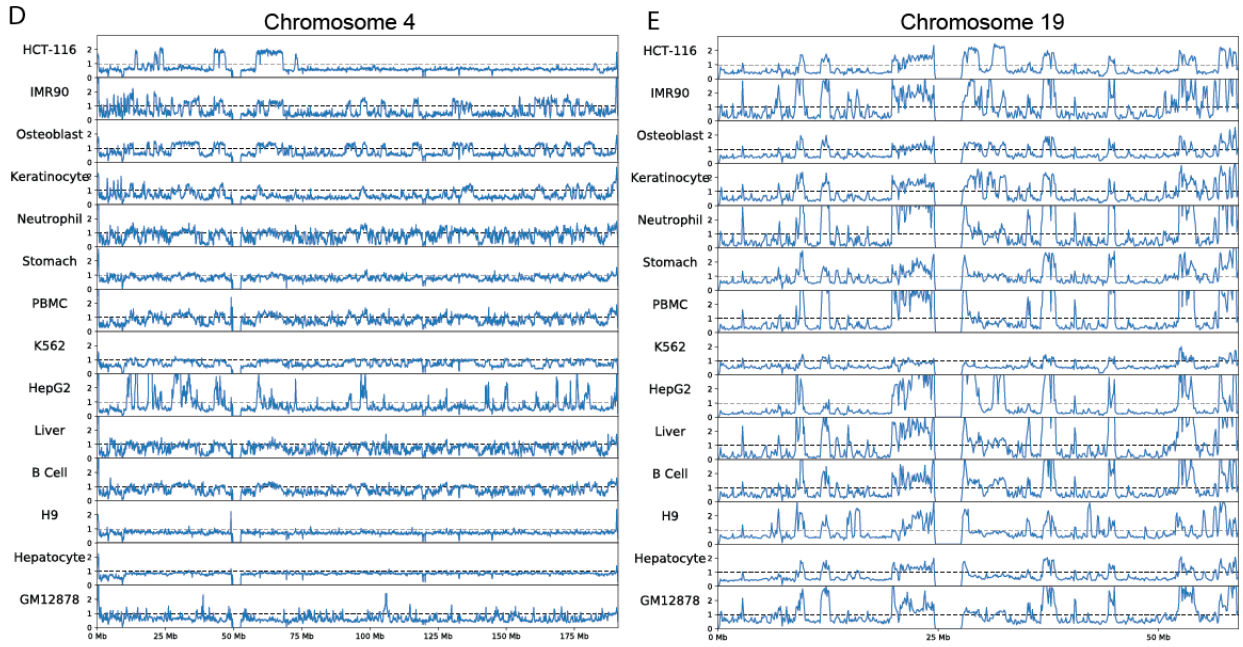
| | Gro | H3K36me3 | H3K4me1 | H3K4me2 | H3K4me3 | H3K27ac | H3K79me2 | H4K20me1 | H3K27me3 | H3K9me2 | H3K9me3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HCT116 Chr4 PCA A | 1.2 | 1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1 | 0.96 | 1 | 0.83 |
| HCT116 Chr4 PCA B | 0.12 | 0.95 | 0.72 | 0.62 | 0.66 | 0.63 | 0.67 | 0.85 | 1.2 | 0.76 | 1.9 |
| GM12878 Chr4 PCA A | 2.3 | 1.6 | 1.5 | 1.7 | 2 | 1.8 | 2.1 | 1.2 | 1.1 | | 1.3 |
| GM12878 Chr4 PCA B | 0.19 | 0.64 | 0.67 | 0.55 | 0.39 | 0.48 | 0.33 | 0.9 | 0.94 | | 0.83 |
| HCT116 Chr4 K-means A | 2.1 | 1.2 | 1.3 | 1.4 | 1.6 | 1.4 | 1.5 | 1.1 | 0.9 | 0.96 | 0.8 |
| HCT116 Chr4 K-means B | 0.84 | 0.86 | 0.83 | 0.81 | 0.79 | 0.77 | 0.81 | 1 | 1.7 | 0.93 | 0.8 |
| HCT116 Chr4 K-means C | 0.086 | 0.99 | 0.7 | 0.58 | 0.63 | 0.6 | 0.67 | 0.78 | 0.92 | 0.7 | 2.2 |
| HCT116 Chr4 K-means D | 0.088 | 0.86 | 0.81 | 0.71 | 0.57 | 0.8 | 0.65 | 0.99 | 0.87 | 1.2 | 0.88 |
| GM12878 Chr4 K-means A | 2.4 | 1.6 | 1.6 | 1.8 | 2.1 | 1.9 | 2.2 | 1.2 | 1.1 | | 1.3 |
| GM12878 Chr4 K-means B | 0.2 | 0.64 | 0.67 | 0.55 | 0.39 | 0.49 | 0.33 | 0.91 | 0.95 | | 0.84 |

B

| | Gro | H3K36me3 | H3K4me1 | H3K4me2 | H3K4me3 | H3K27ac | H3K79me2 | H4K20me1 | H3K27me3 | H3K9me2 | H3K9me3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HCT116 Chr4:65-95Mb PCA A | 1.3 | 1 | 1.1 | 1.1 | 1.2 | 1.1 | 1.1 | 1 | 0.88 | 1.1 | 0.78 |
| HCT116 Chr4:65-95Mb PCA B | 0.021 | 0.88 | 0.71 | 0.56 | 0.52 | 0.6 | 0.63 | 0.91 | 1.4 | 0.81 | 1.7 |
| GM12878 Chr4:65-95Mb PCA A | 2.6 | 1.7 | 1.7 | 1.9 | 2.3 | 2 | 2.4 | 1.1 | 0.95 | | 1.3 |
| GM12878 Chr4:65-95Mb PCA B | 0.25 | 0.69 | 0.68 | 0.55 | 0.36 | 0.5 | 0.34 | 0.95 | 1 | | 0.84 |
| HCT116 Chr4:65-95Mb K-means A | 2.3 | 1.3 | 1.4 | 1.7 | 1.9 | 1.6 | 1.7 | 1.1 | 0.78 | 1 | 0.78 |
| HCT116 Chr4:65-95Mb K-means B | 0.074 | 0.75 | 0.73 | 0.64 | 0.52 | 0.65 | 0.62 | 1.1 | 2.2 | 0.98 | 0.78 |
| HCT116 Chr4:65-95Mb K-means C | 0.0035 | 0.96 | 0.7 | 0.54 | 0.54 | 0.57 | 0.64 | 0.76 | 0.82 | 0.7 | 2.2 |
| HCT116 Chr4:65-95Mb K-means D | 0.55 | 0.82 | 0.83 | 0.74 | 0.58 | 0.78 | 0.65 | 0.94 | 0.88 | 1.1 | 0.78 |
| GM12878 Chr4:65-95Mb K-means A | 3 | 1.8 | 1.8 | 2.1 | 2.7 | 2.3 | 2.8 | 1.1 | 0.88 | | 1.4 |
| GM12878 Chr4:65-95Mb K-means B | 0.26 | 0.7 | 0.71 | 0.58 | 0.37 | 0.52 | 0.35 | 0.95 | 1 | | 0.86 |

C

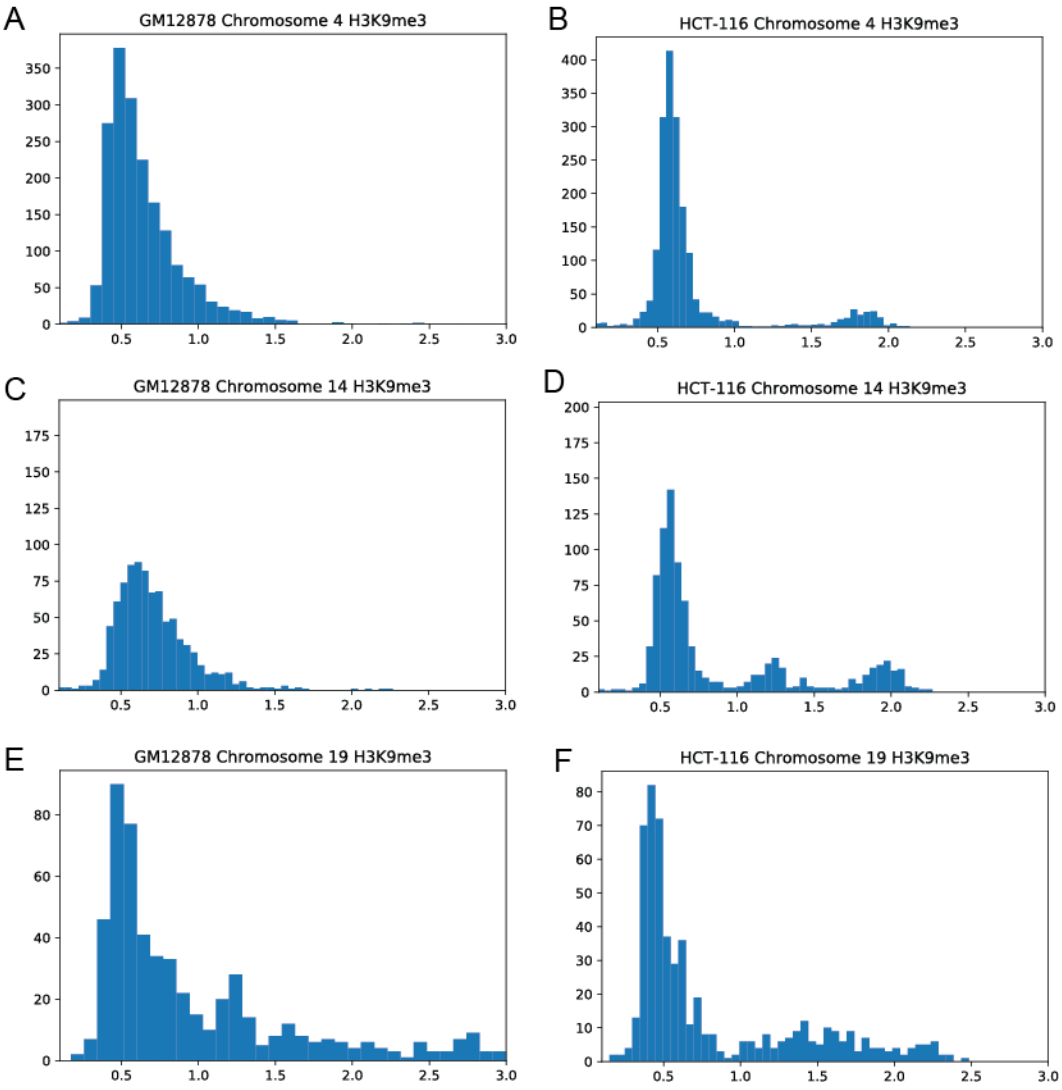
| | Gro | H3K36me3 | H3K4me1 | H3K4me2 | H3K4me3 | H3K27ac | H3K79me2 | H4K20me1 | H3K27me3 | H3K9me2 | H3K9me3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HCT116 Chr19 PCA A | 1.5 | 1.1 | 1.4 | 1.4 | 1.4 | 1.5 | 1.3 | 1.2 | 0.98 | 1 | 0.59 |
| HCT116 Chr19 PCA B | 0.24 | 0.84 | 0.44 | 0.37 | 0.37 | 0.32 | 0.51 | 0.72 | 1 | 1 | 1.6 |
| GM12878 Chr19 PCA A | 1.3 | 1.3 | 1.4 | 1.4 | 1.4 | 1.5 | 1.5 | 1.3 | 1.2 | | 0.61 |
| GM12878 Chr19 PCA B | 0.55 | 0.59 | 0.48 | 0.46 | 0.43 | 0.39 | 0.38 | 0.64 | 0.74 | | 1.5 |
| HCT116 Chr19 K-means A | 1.6 | 1.1 | 1.4 | 1.5 | 1.5 | 1.5 | 1.4 | 1.2 | 0.9 | 0.97 | 0.58 |
| HCT116 Chr19 K-means B | 0.53 | 0.88 | 0.71 | 0.64 | 0.65 | 0.56 | 0.77 | 1 | 1.5 | 1.2 | 0.79 |
| HCT116 Chr19 K-means C | 0.15 | 0.81 | 0.38 | 0.29 | 0.28 | 0.26 | 0.4 | 0.61 | 0.83 | 0.91 | 1.9 |
| GM12878 Chr19 K-means A | 1.5 | 1.4 | 1.5 | 1.6 | 1.7 | 1.7 | 1.7 | 1.3 | 1.1 | | 0.59 |
| GM12878 Chr19 K-means B | 0.7 | 0.82 | 0.88 | 0.77 | 0.64 | 0.7 | 0.7 | 1.1 | 1.4 | | 0.69 |
| GM12878 Chr19 K-means C | 0.52 | 0.54 | 0.4 | 0.4 | 0.38 | 0.34 | 0.31 | 0.55 | 0.62 | | 1.7 |




Supplemental Figure 5.1. A) Fold-enrichment of each chromatin feature within each compartment defined by PCA or k-means clustering on chromosome 4.

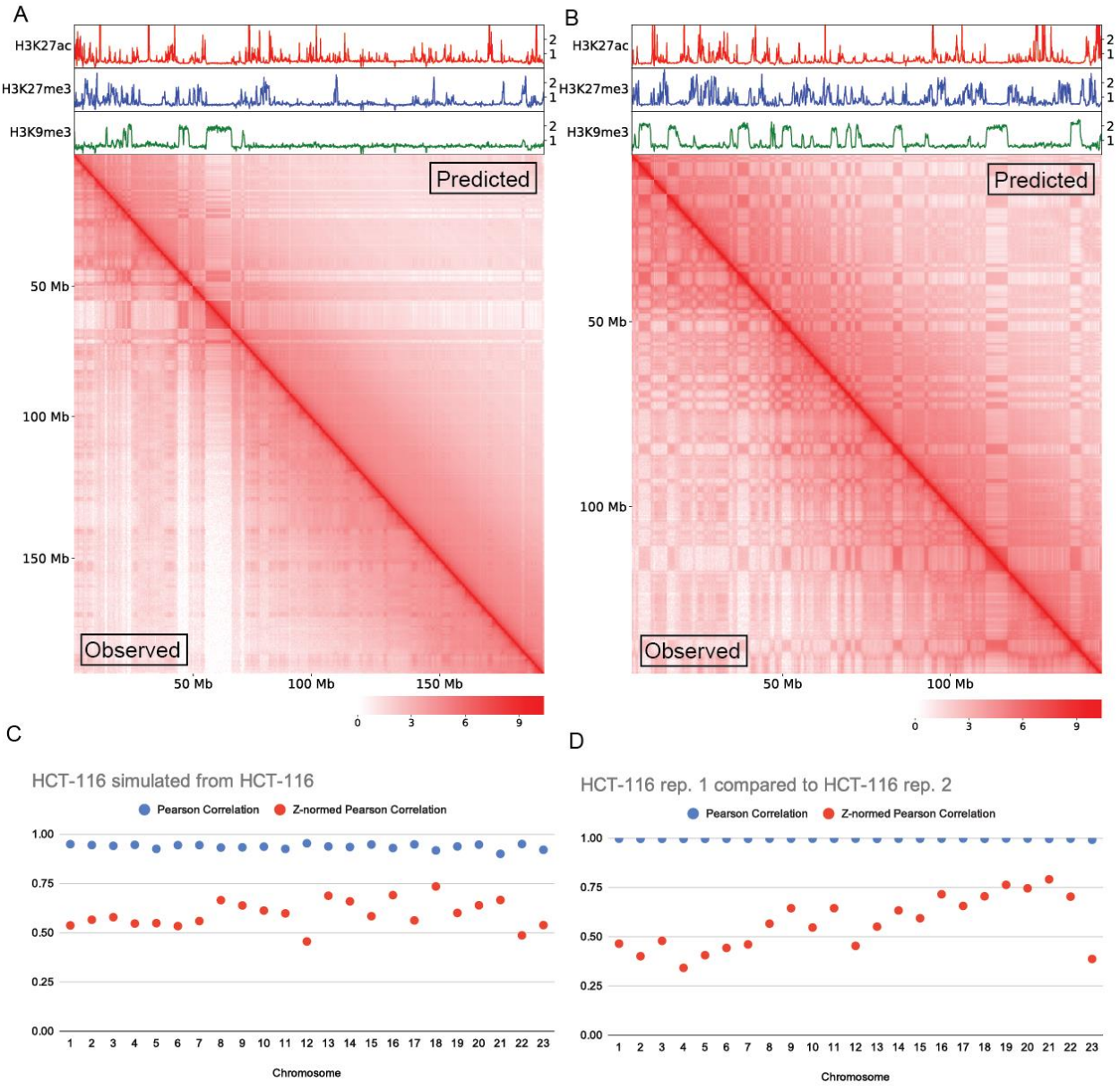
B) Fold-enrichment of each chromatin feature within each compartment defined by PCA or k-means clustering on chromosome 4 65-95Mb. C) A Fold-enrichment of each chromatin feature within each compartment defined by PCA or k-means clustering on chromosome 19. D)

H3k9me3 fold-change over control tracks for a variety of human cell lines and tissues on chromosome 4. E) H3k9me3 fold-change over control tracks for a variety of human cell lines and tissues on chromosome 19.



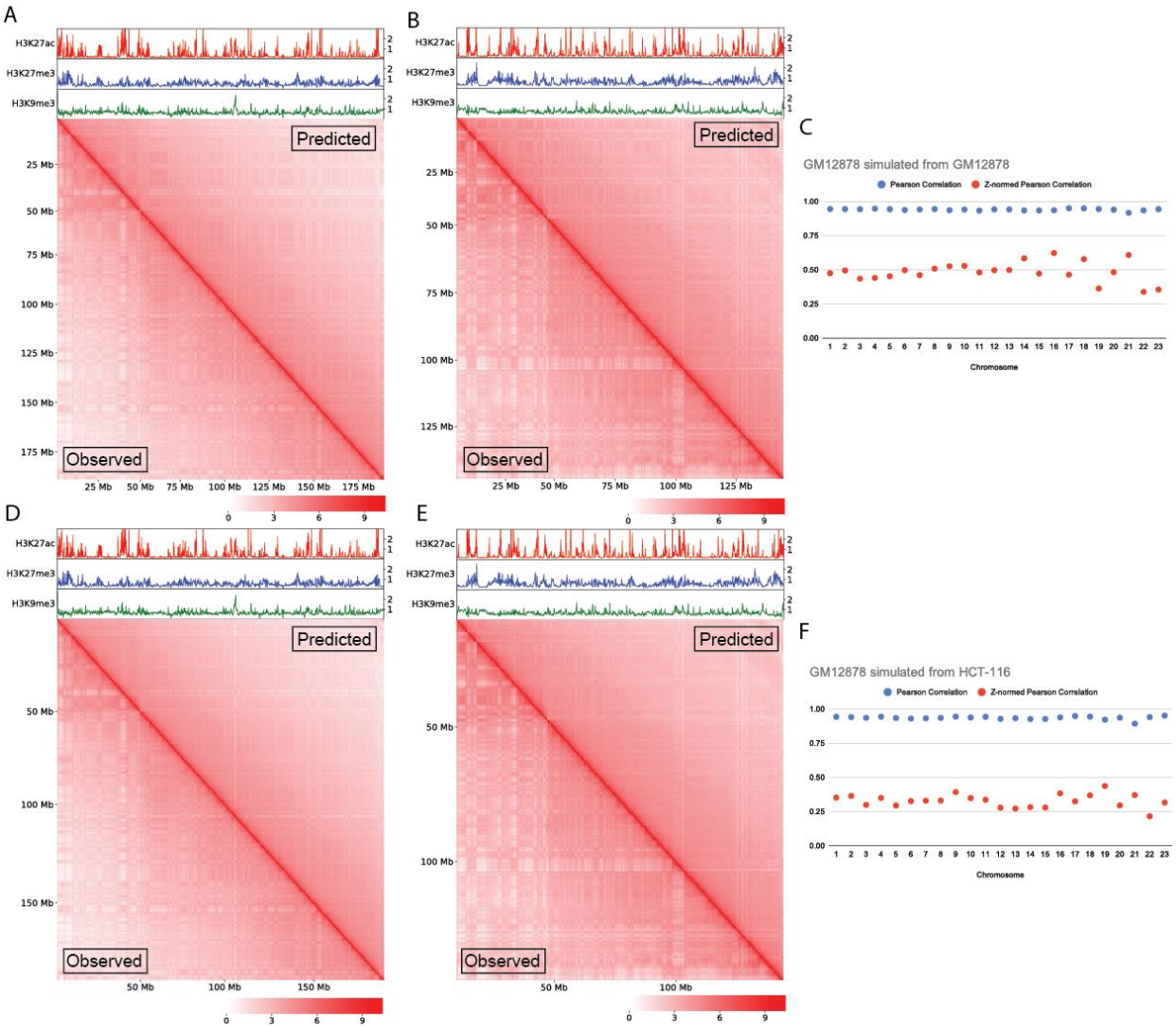

Supplemental Figure 2. A) 100 kb bin histogram of the distribution of H3K9me3 on chromosome 4 in GM12878. B) 100 kb bin histogram of the distribution of H3K9me3 on chromosome 4 in HCT-116. C) 100 kb bin histogram of the distribution of H3K9me3 on chromosome 14 in GM12878. D) 100 kb bin histogram of the distribution of H3K9me3 on chromosome 14 in HCT-116. E) 100 kb bin histogram of the distribution of H3K9me3 on

chromosome 9 in GM12878. F) 100 kb bin histogram of the distribution of H3K9me3 on chromosome 9 in HCT-116.




Supplemental Figure 5.3. A) Comparison of HCT-116 chromosome 4 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated. B)

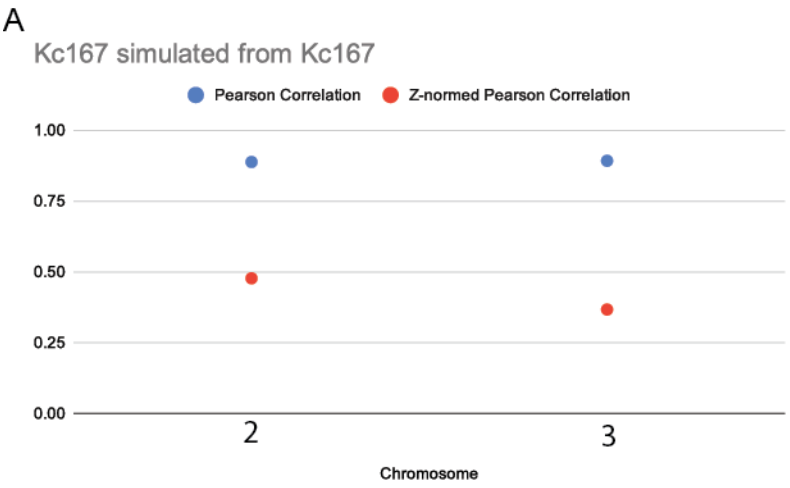
Comparison of HCT-116 chromosome 8 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated. C) Pearson correlation values comparing the observed and simulated maps of HCT-116 for each chromosome. D) Pearson correlation values comparing the two HCT-116 Hi-C replicates.




Supplemental Figure 5.4. A) Comparison of GM12878 chromosome 4 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated. B) Comparison of GM12878 chromosome 8 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated. C) Pearson correlation values comparing

the observed and simulated maps of GM12878 for each chromosome. D) Comparison of GM12878 chromosome 4 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated using attraction-repulsion maps learned from HCT-116. E) Comparison of GM12878 chromosome 8 logged Hi-C interaction maps. The bottom left triangle is observed and the upper right triangle is simulated using attraction-repulsion maps learned from HCT-116.

F) Pearson correlation values comparing the observed and simulated maps of GM12878 using attraction-repulsion maps learned from HCT-116 for each chromosome.




Supplemental Figure 5.5. Pearson correlation values comparing the observed and simulated maps of Drosophila Kc cells for each chromosome.

## Methods

ChIP-seq quantile normalization:

We used fold-change over control with both replicates combined were used in a bigwig format. 100 kb bins with no reads mapped in any ChIP-seq were excluded from analysis. We

normalized the remaining bins into twenty discrete quantiles according to their signal compared genome-wide.

Hi-C Quality Control:

Hi-C maps generated from reads with quality score >Q30 were used. Genomic bins were removed from the maps and all subsequent analyses according to several criteria calculated on each chromosome. Bins which were removed from ChIP-seq quantiles were also removed from the Hi-C. Bins that had a total read sum greater than 3 standard deviations above or less than 3 standard deviations below the average bin read sum were dropped. Bins with non-zero interactions with bins with non-zero interactions greater than 3 standard deviations above or less than 3 standard deviations below the average bin were dropped.

Hi-C Normalization and Pearson correlation:

Hi-C maps were balanced using Knight-Ruiz normalization. For some techniques such as Pearson correlation analysis the Hi-C maps needed to be distance normalized which was done by dividing each interaction by the average of all interactions at that distance. This produces an observed/expected value for all interaction bins. Pearson correlations of Hi-C maps were then generated from these distance normalized matrices.

Hi-C Principal Component Analysis and Compartment calls:

Principal Component Analysis was performed on the chromosomal Pearson correlation maps. The first principal component (PC1) is defined as the eigenvector with the largest eigenvalue. All bins with positive values in PC1 were assigned to one compartment while all negative values were assigned to the other. The compartment with the largest enrichment for Gro-seq signal was defined as the A compartment and the other B.

Hi-C k-means clustering:

To dissect this complex compartmentalization we employed the unsupervised k-means clustering algorithm to identify clusters in both cell types. We used MiniBatchKMeans from scikit-learn's machine learning library with varying numbers for k depending on the features of the given chromosome. Clustering was performed on the Pearson correlation maps of each chromosome separately and clusters were identified as A,B,C, and D by their enrichments for H3K27ac, H3K27me3, H3K9me3, and H3K9me2 respectively.

Compartmentalization by Independent Forces to Simulate Interaction Maps

To enable comparison across cell types and experiments we first binned all epigenetic signals into quantiles at a 100 kb resolution. For each normalized signal, we then learned, using a Maximum Likelihood Estimation approach, an attraction-repulsion relationship for each pair of quantiles. This attraction-repulsion mapping effectively represents the average enrichment or depletion between all bins with the corresponding level of signal. The model then predicts the number of reads at each bin by summing the attraction-repulsion scores for each signal and multiplying by a constant distance factor to account for the power law decay of genomic interactions.

We model the compartmentalization of the genome as the independent contributions of individual 1D epigenetic signals and use a machine learning method Maximum Likelihood Estimation (MLE) to learn the relationship between the signals and interaction frequencies in the Hi-C map. In this way we hope to quantify the nuclear forces driving compartmentalization. Our model treats each epigenetic signal as an independent but additive effect on Hi-C interaction frequency according to the equation:

$$E_{ij} = 1 + H3K27ac_{ij} + H3K27me3_{ij} + H3K9me3_{ij}$$

Where the expected interaction frequency E between any two genomic loci i and j is the sum of each signal's weight effect. Using MLE we find the optimal values of each signal's weights such that the expected result E is as close to the observed value as possible. This produces a trained model that can reproduce the 3D organization of the genome from 1D epigenetic signals by quantifying the expected contributions of each correlated compartmentalizing force.

We use the Maximum Likelihood Estimation approach to optimize the values of a vector $\boldsymbol{\beta}$ where each entry in $\boldsymbol{\beta}$ corresponds to an entry in the attraction-repulsion maps such that for all possible pairs of quantiles for each of the three chromatin signals H3K27ac, H3K27me3, and H3K9me3, there is a corresponding weight in $\boldsymbol{\beta}$. We then construct a sparse matrix X where each row corresponds to an interaction bin in the flattened Hi-C matrix and each column a weight in $\boldsymbol{\beta}$. Each row in X is zeros except in the 3 columns corresponding to the entries in $\boldsymbol{\beta}$ that describe the pair of genomic bins' signal quantiles. If we take an observed Hi-C map which has been normalized for distance by dividing by the expected value at each distance, which we term y, the model's approximation of the normalized observed map is then:

$$y = \vec{1} + X\,\beta$$

Treating the normalized interaction frequency as a normal distribution we derive a likelihood function the log of which we will maximize to optimize the weights of $\boldsymbol{\beta}$ in a Maximum Likelihood Estimation approach:

$$\beta = \ arg\ \max_{\beta} - \sum_{i=1}^{N} (\mathrm{y}_i - \vec{1} + X_i\beta)^2$$

Where N is the total number of unique interacting bins in the linearized Hi-C matrix excluding each genomic bin's interactions with itself. To optimize the parameters of $\boldsymbol{\beta}$ we iteratively solve starting with initial values of 0 for all weights in $\boldsymbol{\beta}$. We use the Newton-Raphson method to update the weights by gradient descent. With each iteration we account for one final feature of the Hi-C matrix. Due to the colocalization of compartmentalizing features along the chromosome the frequency of intra-compartmental interactions is enriched at short ranges. Two genomic bins within a megabase are more likely to be enriched for compartmental interactions than two bins many megabases apart. This bias leads to aberrant distance normalization, and so the distance normalization is updated after each iteration to account for the average compartmental enrichment the model predicts at each distance. The distance normalization is divided by the average enrichment so that it more accurately reflects the true effect of distance on interaction frequency.

Drosophila simulation

The Drosophila simulation works identically to the simulation in humans except for a limit on the maximum size of bins that are considered for the analysis. We excluded all interaction bins further than 2 megabases apart as the compartmental signal beyond this distance was substantially weaker.

Simulation Pearson Correlation Analysis

Simulations were generated at 100 kb resolution using the average of the attraction-repulsion maps learned from every chromosome except the one being simulated. We quantified the accuracy of the model using the Pearson correlations between the observed and simulated maps after dividing by the average distance. Due to the power law decay of interaction frequency with respect to distance in Hi-C maps any simulation which accurately reproduces this decay will have an extremely high correlation. As this would not represent the model's

capacity to reproduce compartmental organization we normalized for distance to eliminate the natural correlation driven by the accurate representation of the distance decay.

Author Contributions

MHN and VGC designed the project and wrote the manuscript. MHN performed all data analyses and wrote the machine learning algorithms.

Acknowledgements

Special thanks to Anthony Christodoulou and Andrea Trevino for their insight into machine learning. This work was supported by U.S. Public Health Service Award R01 GM035463 from the National Institutes of Health. MHN was supported by NIH T32 GM008490. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of Interests

The authors declare no competing interests.

# References

Banani, Salman F., Hyun O. Lee, Anthony A. Hyman, and Michael K. Rosen. 2017. "Biomolecular Condensates: Organizers of Cellular Biochemistry." Nature Reviews Molecular Cell Biology 18 (5): 285–98. https://doi.org/10.1038/nrm.2017.7.

Boijja, Ann, Isaac A. Klein, Benjamin R. Sabari, Alessandra Dall'Agnese, Eliot L. Coffey, Alicia V. Zamudio, Charles H. Li, et al. 2018. "Transcription Factors Activate Genes through

the Phase Separation Capacity of Their Activation Domains." Cell 175 (7): 1842-1855.e16. https://doi.org/10.1016/j.cell.2018.10.042.

Falk, Martin, Yana Feodorova, Natalia Naumova, Maxim Imakaev, Bryan R. Lajoie, Heinrich Leonhardt, Boris Joffe, et al. 2019. "Heterochromatin Drives Compartmentalization of Inverted and Conventional Nuclei." Nature 570 (7761): 395–99. https://doi.org/10.1038/s41586-019-1275-3.

Jackson, D A, A B Hassan, R J Errington, and P R Cook. 1993. "Visualization of Focal Sites of Transcription within Human Nuclei." The EMBO Journal 12 (3): 1059–65.

Ladouceur, Anne-Marie, Baljyot Singh Parmar, Stefan Biedzinski, James Wall, S. Graydon Tope, David Cohn, Albright Kim, Nicolas Soubry, Rodrigo Reyes-Lamothe, and Stephanie C. Weber. 2020. "Clusters of Bacterial RNA Polymerase Are Biomolecular Condensates That Assemble through Liquid–Liquid Phase Separation." Proceedings of the National Academy of Sciences 117 (31): 18540–49. https://doi.org/10.1073/pnas.2005019117.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." Science 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

Pirrotta, Vincenzo, and Hua-Bing Li. 2012. "A View of Nuclear Polycomb Bodies." Current Opinion in Genetics & Development 22 (2): 101–9. https://doi.org/10.1016/j.gde.2011.11.004.

Plys, Aaron J., Christopher P. Davis, Jongmin Kim, Gizem Rizki, Madeline M. Keenen, Sharon K. Marr, and Robert E. Kingston. 2019. "Phase Separation of Polycomb-Repressive Complex 1 Is Governed by a Charged Disordered Region of CBX2." Genes & Development 33 (13–14): 799–813. https://doi.org/10.1101/gad.326488.119.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” Cell 159 (7): 1665–80. https://doi.org/10.1016/j.cell.2014.11.021.

Rowley, M. Jordan, and Victor G. Corces. 2018. “Organizational Principles of 3D Genome Architecture.” Nature Reviews. Genetics 19 (12): 789–800. https://doi.org/10.1038/s41576-018-0060-8.

Rowley, M. Jordan, Michael H. Nichols, Xiaowen Lyu, Masami Ando-Kuri, I. Sarahi M. Rivera, Karen Hermetz, Ping Wang, Yijun Ruan, and Victor G. Corces. 2017. “Evolutionarily Conserved Principles Predict 3D Chromatin Organization.” Molecular Cell 67 (5): 837-852.e7. https://doi.org/10.1016/j.molcel.2017.07.022.

Wang, Liang, Yifei Gao, Xiangdong Zheng, Cuifang Liu, Shuangshuang Dong, Ru Li, Guanwei Zhang, et al. 2019. “Histone Modifications Regulate Chromatin Compartmentalization by Contributing to a Phase Separation Mechanism.” Molecular Cell 76 (4): 646-659.e6. https://doi.org/10.1016/j.molcel.2019.08.019.

# Chapter 6: Discussion

## Conclusions

The folded structures of the cardinal biological polymers DNA, RNA, and proteins are essential to their function. DNA, as the largest and most disordered of these polymers, poses considerable obstacles to the analysis of its structure. Despite this, we have elucidated at least two conserved independent processes that influence DNA folding in the genome (Rowley et al. 2017). The first, a loop extrusion process by which loops are enlarged by the DNA motor activity of members of the SMC family (Nichols and Corces 2015; 2018). Condensin compacts chromosomes by extruding loops during mitosis whereas cohesin constantly extrudes loops along chromosomes during interphase. In vertebrates, CTCF appears to act as a unidirectional border to cohesin complexes and thereby forms loops between its oriented binding sites (Nichols and Corces 2015). Beyond these most basic concepts, we know little about this process. The DNA motor activity of these complexes has been confirmed experimentally but we only have vague models, such as the tethered inchworm, for the mechanism by which these complexes extrude DNA (Nichols and Corces 2018). Moreover, in the case of cohesin, we lack compelling theories for why interphase loop extrusion occurs. This energy-intensive process must play an essential role in the nucleus to justify its existence. CTCF and enhancer-promoter loops may be one mechanism by which loop extrusion serves to regulate transcription, but thus far perturbations of these systems have had only minor effects on transcription in cell lines, despite leading to cell senescence and death (Rao et al. 2017; Nora et al. 2017). CTCF binding sites and associated loops are frequently cell-type specific, pointing towards a role for these loops in transcriptional regulation.

The second independent principle driving the organization of the genome is compartmentalization. The basic concept of functionally related nuclear components colocalizing was seen in the first observations of membraneless nuclear bodies. Advanced techniques that can determine the localization of chromatin have extended this model to encompass colocalization of functionally related chromatin. We now know that compartmentalization of a genomic locus directly corresponds to the chromatin features present there (Rowley et al. 2017). However, just as with loop extrusion, we still know relatively little about the mechanisms that give rise to these structures or their functional implications. While the canonical, binary A/B compartmentalization correlated with transcriptional activity can be found throughout human tissues and indeed all eukaryotic genomes it is becoming increasingly clear that this is only one major axis of chromatin segregation. Transcriptionally inactive regions rich in H3K9me3, H3K27me3, or neither histone modification possess distinct compartmentalization patterns, suggesting distinct transcriptionally inactive compartments. The purpose of this segregation is as yet unclear, however, that these patterns are highly dynamic across human tissues suggests they play an important role in transcriptional regulation and differentiation (Rowley et al. 2017).

## Future directions

The ability to capture the conformation of all of the chromatin in a population of cells has opened the door to the structural analysis of the DNA polymer. The most significant interaction patterns in these stochastic conformations have now been identified. We can explain the large majority of the structural organization of the genome as the function of loop extrusion and/or compartmentalization. But while we understand roughly what these patterns are, we lack a mechanistic understanding of how these forces work and why they are conserved features of

genomes. Future work must investigate genomic organization from these two ends. A finer understanding of the mechanistic details will surely shed light on the downstream roles this organization plays. Vice versa, theoretical breakthroughs on the purpose of the organization will produce testable hypotheses regarding the mechanism.

Answering the outstanding mechanistic questions pertaining to loop extrusion will require structural and functional dissections of these complexes. Paramount to understanding this process is determining how ATP hydrolysis driven conformational changes produce directed movement on the DNA polymer to extrude loops. Secondary questions involve the regulatory processes that constrain extrusion, including CTCF's unidirectional barrier function. To better understand the consequences of loop extrusion and CTCF barriers, more careful perturbations must be undertaken to quantify the effect of these processes on enhancer-promoter looping and transcriptional regulation. Introducing human CTCF barriers to genomes lacking CTCF loops also presents a novel opportunity to see the effects of this organization.

An equally important approach will be understanding the role of cohesin's interphase loop extrusion and CTCF oriented loops in genomic regulation. One popular theoretical model is that CTCF loops form insulated neighborhoods that promote enhancer-promoter interactions within the loop and inhibit interactions across the loop (Hnisz et al., 2016). Several findings at specific genomic loci support this model (Lupiáñez et al., 2015; Williamson et al., 2019). However, from Hi-C analysis we know that the quantitative enrichment of interaction frequency inside and outside of loops is relatively minor. Moreover, genome-wide perturbation of looping has only shown relatively minor effects on transcriptional levels genome wide suggesting extrusion is not indispensable for mediating enhancer-promoter interactions (Nora et al., 2017; Rao et al., 2017; Vian et al., 2018). Where CTCF loops do strongly influence interaction frequencies, is at the loops themselves. CTCF binding sites are enriched in promoter regions and may serve to form

direct loops to enhancers. CTCF sites are found in numerous important genomic regions with distinct spatial arrangement such as the Hox, Protocadherin, and Immunoglobulin loci where they appear to serve specific and distinct roles (Ba et al., 2020; Guo et al., 2012; Heger et al., 2012). Understanding CTCF's specific roles in these distinctly tractable loci will be a key means of understanding its general function genome wide. These loci also provide opportunities to understand the evolutionary origins of CTCF looping and thus potentially its function. The Hox locus in *Drosophila* has conserved CTCF binding sites like those found in the mammalian loci, however *Drosophila* CTCF lacks an interaction orientation bias. CTCF's role as insulator and architectural factor in the absence of the formation of oriented loops is an important missing piece to understanding this protein and genomic organization at large.

Significant theoretical work has led to several models of compartmentalization. In vitro experiments have demonstrated the ability of various nucleoplasm and chromatin components to self-segregate. This evidence is still circumstantial as we lack efficient methods to perturb nuclear compartmentalization. The depletion of key compartmental components or disruption by other means will illuminate the physical nature of these biomolecular condensates. Additionally, techniques that allow for the assessment of chromatin conformation simultaneously with chromatin constituents such as HiChIP and SPRITE, provide an avenue to determine the important epigenetic components of these compartments (Mumbach et al. 2016; Quinodoz et al. 2018).

Theoretical justification for why compartmentalization is such a prominent feature of nuclear organization appears relatively straightforward but has been difficult to experimentally test. Compartmental segregation as a mechanism for improving the efficiency of genomic functions by increasing local concentrations of needed components as well as limiting those functions spatially as a means of regulation is conceptually attractive. However, we lack sophisticated

methods to perturb compartmentalization in the nucleus without significantly disrupting normal function. One surprising avenue of investigation into nuclear compartmentalization of chromatin could be elucidating the principles underlying compartmentalization in prokaryotes where circular chromosomes in the cytoplasm nonetheless show compartmental interaction patterns by Hi-C (Le et al., 2013). Dissecting the organization of these organisms that lack both nuclear envelopes and histones may provide key insights into the origins of compartmental organization and the mechanisms by which is it achieved (Heger et al., 2012).

As we learn more about the mechanisms and functions of genomic architecture the importance of DNA structure will become clearer. That these organizational principles are simultaneously highly evolutionarily conserved, and yet significantly differ in presentation between human tissues, indicates they play an essential role in the genome that we do not yet fully understand.

## References

Ba, Z., Lou, J., Ye, A.Y., Dai, H.-Q., Dring, E.W., Lin, S.G., Jain, S., Kyritsis, N., Kieffer-Kwon, K.-R., Casellas, R., Alt, F.W., 2020. CTCF orchestrates long-range cohesin-driven V(D)J recombinational scanning. Nature 1–6. https://doi.org/10.1038/s41586-020-2578-0

Guo, Y., Monahan, K., Wu, H., Gertz, J., Varley, K.E., Li, W., Myers, R.M., Maniatis, T., Wu, Q., 2012. CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice. PNAS 109, 21081–21086. https://doi.org/10.1073/pnas.1219280110

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E., Wiehe, T., 2012. The chromatin insulator CTCF and the emergence of metazoan diversity. PNAS 109, 17507–17512. https://doi.org/10.1073/pnas.1111941109

Hnisz, D., Day, D.S., Young, R.A., 2016. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. Cell 167, 1188–1200. https://doi.org/10.1016/j.cell.2016.10.024

Le, T.B.K., Imakaev, M.V., Mirny, L.A., Laub, M.T., 2013. High-resolution mapping of the spatial organization of a bacterial chromosome. Science 342, 731–734. https://doi.org/10.1126/science.1242059

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S.A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., Mundlos, S., 2015. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. Cell 161, 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004

Mumbach, Maxwell R., Adam J. Rubin, Ryan A. Flynn, Chao Dai, Paul A. Khavari, William J. Greenleaf, and Howard Y. Chang. 2016. “HiChIP: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture.” Nature Methods 13 (11): 919–22. https://doi.org/10.1038/nmeth.3999.

Nichols, Michael H., and Victor G. Corces. 2015. “A CTCF Code for 3D Genome Architecture.” Cell 162 (4): 703–5. https://doi.org/10.1016/j.cell.2015.07.053.

———. 2018. “A Tethered-Inchworm Model of SMC DNA Translocation.” Nature Structural & Molecular Biology 25 (10): 906–10. https://doi.org/10.1038/s41594-018-0135-4.

Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., Bruneau, B.G., 2017. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell 169, 930-944.e22. https://doi.org/10.1016/j.cell.2017.05.004

Quinodoz, Sofia A., Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M. Lai, et al. 2018. “Higher-Order Inter-Chromosomal Hubs Shape 3D

Genome Organization in the Nucleus.” Cell 174 (3): 744-757.e24. https://doi.org/10.1016/j.cell.2018.05.024.

Rao, S.S.P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.-R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., Huang, X., Shamim, M.S., Shin, J., Turner, D., Ye, Z., Omer, A.D., Robinson, J.T., Schlick, T., Bernstein, B.E., Casellas, R., Lander, E.S., Aiden, E.L., 2017. Cohesin Loss Eliminates All Loop Domains. Cell 171, 305-320.e24. https://doi.org/10.1016/j.cell.2017.09.026

Rowley, M. Jordan, Michael H. Nichols, Xiaowen Lyu, Masami Ando-Kuri, I. Sarahi M. Rivera, Karen Hermetz, Ping Wang, Yijun Ruan, and Victor G. Corces. 2017. “Evolutionarily Conserved Principles Predict 3D Chromatin Organization.” Molecular Cell 67 (5): 837-852.e7. https://doi.org/10.1016/j.molcel.2017.07.022.

Vian, L., Pękowska, A., Rao, S.S.P., Kieffer-Kwon, K.-R., Jung, S., Baranello, L., Huang, S.-C., El Khattabi, L., Dose, M., Pruett, N., Sanborn, A.L., Canela, A., Maman, Y., Oksanen, A., Resch, W., Li, X., Lee, B., Kovalchuk, A.L., Tang, Z., Nelson, S., Di Pierro, M., Cheng, R.R., Machol, I., St Hilaire, B.G., Durand, N.C., Shamim, M.S., Stamenova, E.K., Onuchic, J.N., Ruan, Y., Nussenzweig, A., Levens, D., Aiden, E.L., Casellas, R., 2018. The Energetics and Physiological Impact of Cohesin Extrusion. Cell 173, 1165-1178.e20. https://doi.org/10.1016/j.cell.2018.03.072

Williamson, I., Kane, L., Devenney, P.S., Flyamer, I.M., Anderson, E., Kilanowski, F., Hill, R.E., Bickmore, W.A., Lettice, L.A., 2019. Developmentally regulated Shh expression is robust to TAD perturbations. Development 146. https://doi.org/10.1242/dev.179523