ELUCIDATING THE REGULATORY ROLE OF 3D GENOME FOLDING DURING NEURAL DIFFERENTIATION AND SYNAPTIC ACTIVATION

Jonathan A. Beagan

A DISSERTATION
in
Bioengineering

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

_____
Dr. Jennifer E. Phillips-Cremins
Assistant Professor of Bioengineering
University of Pennsylvania

Graduate Group Chairperson

_____
Dr. Yale Cohen
Professor of Otorhinolaryngology
University of Pennsylvania

Dissertation Committee

Dr. Robert L. Mauck, Professor of Orthopaedic Surgery, University of Pennsylvania

Dr. Zhaolan Zhou, Associate Professor of Genetics, University of Pennsylvania

Dr. Jason D. Shepherd, Associate Professor of Neurobiology and Anatomy, University of Utah

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Jenn, for her unwavering, unflinching, full-forced support of me during my graduate training. Jenn's overflowing energy, commitment to excellence, and unmatched work ethic are contagious, and much of the productivity I have generated during my PhD can be attributed to her constant encouragement. From day one she entrusted me with an enormous amount of responsibility for which I am truly grateful. And perhaps most importantly, she always knew how to effectively motivate me – with Bill Belichick quotes and Tom Brady memes.

I also consider myself truly lucky to have the thesis committee that I did. The lab of Dr. Rob Mauck was one of the focal points that drew me to Penn Bioengineering, and although I couldn't have predicted being drawn away to the allure of 3-D epigenomics, the time I spent working with Rob and his amazing lab was a joy and showed me what the standard was for producing high quality science during graduate school. Dr. Zhaolan Zhou taught me my first ever semester of epigenetics (and can still recall what I presented on in his class, which is mind blowing). Since that point he has been the most positive, encouraging professor each time he has seen me present my work, and I will be forever appreciative of the times he dissolved my anxiety before a talk with a well-placed joke. And despite interacting entirely remotely, Dr. Jason Shepherd has managed brilliantly to be a patient and effective collaborator and mentor that has guided my less-than-well-prepared foray into the world of memory and cognition research. Thanks to you all.

None of this work would have been possible without my rock star colleagues in the Cremins lab. In the early days of the lab when I knew nothing (or at last that's how it felt),

Heidi Norton and Thomas Gilgenast became two of my closest collaborators and friends. Heidi taught me how to culture NPCs and perform RNA-seq (not to mention officiated my wedding and provided 10,000 additional acts of friendship), and anything I possess that remotely resembles coding skills came from studying and breaking Thomas's code and before finally asking him for help. Jesi Kim displayed unmatched dedication in getting Chapter 4 of this thesis published across many late nights in lab of blasting Disney movies, Friends episodes, and taking shelter under desks in moments of peak stress. Then came an immense amount of #postdocwisdom from Drs. Mayuri Rege, Ji Hun Kim, and Zoltan Simandi. To the friendos Linda Zhou and Lindsey Fernandez, thank you for the friendship, collaboration, and midnight sparkling cider toasts needed to maintain a sufficient level of decorum during late nights in lab. A set of remarkably talented undergraduates, Michael Duong and Kelly Feng, were so generous with their time and detail oriented during key experiments (iPS fixation, YY1 knockdown, activity response time course) which I do not think would have succeeded without their presence. Finally, thank you to Jacqueline Valeri, Shawn Srolovitz, Harvey Huang, Sunny Chen, Zach Plona, Gui Hu, Kate Titus, James Sun, Jingjing Ma, Caroline Lachanski, Daniel Gillis, Daniel Emerson, Michael Guo, Harshini Chandrashekar, Chunmin Ge, Ravi Boya, and Wanfeng Gong for the support, jokes, questions and critiques during our time together in lab.

Thank you to my parents for the love, support, and endless supply of running shoes that kept me in motion. Thank you for being sounding boards, cheerleaders, and for finding a surplus of strange concerts that kept us ironically entertained. And most importantly thank you to Becca, who over the course of this work went from girlfriend to fiancé to wife. You put up with more chaos than was reasonable and did it with grace and love.

Thank you for the times when you brought me dinner in lab, shrugged off my egregious lateness, and kept me centered, all while becoming the only *real* doctor in the family. Your empathy for others continues to inspire me.

ABSTRACT

ELUCIDATING THE REGULATORY ROLE OF 3D GENOME FOLDING DURING NEURAL DIFFERENTIATION AND SYNAPTIC ACTIVATION

Jonathan A. Beagan

Jennifer E. Phillips-Cremins

The causal link between the three-dimensional conformation of the genome and spatiotemporal control of gene regulation has long been studied in the form of enhancer-promoter interactions. Only recently have advances in molecular biology and next generation sequencing allowed higher-order chromatin folding to be queried genome-wide at ultra-high-resolution. In this thesis we leverage Chromosome Conformation Capture Carbon Copy (5C) along with RNA-seq and ChIP-seq to elucidate how the genome is reconfigured during neural development, cellular reprogramming, and synaptic activation. We observe that the first step in neural differentiation is accompanied by a bulk decommissioning of nearly half of the architectural protein CTCF's binding sites in the pluripotent genome, a trend which continues throughout terminal neuronal differentiation and results in the dissolution of many chromatin loops present in embryonic stem cells (ESCs). We identify another zinc finger protein, Yin Yang 1 (YY1), at the base of looping interactions between neural progenitor cell (NPC) specific genes and enhancers; siRNA knockdown of YY1 specifically disrupts interactions between key NPC enhancers and their target genes. Additionally, we find that many of the CTCF sites that are decommissioned during neural lineage commitment are not efficiently restored during cellular

reprogramming of NPCs to induced pluripotent stem cells (iPSCs). CTCF sites that do not successfully regain binding in iPSCs underlie incompletely reprogrammed chromatin architecture, resulting in an iPSC genome folding and transcriptional signature that resembles an intermediate state between ESCs and NPCs. Culture in 2i media conditions restores the CTCF binding, genome folding, and gene expression of iPSCs to patterns resembling those of ESCs. Finally, we find that a large subset of chromatin loops surrounding select neuronal activity response genes (ARGs) are induced *de novo* during cortical neuron activation. We observe a striking correlation between the number, length, and kinetics of loops an ARG forms and how much time that ARG takes to be upregulated in response to neuronal activity. Additionally, we find that common single nucleotide variants (SNVs) associated with Autism Spectrum Disorder connect activity-inducible enhancers to upregulated genes, whereas Schizophrenia SNVs anchor pre-existing loops connecting activity-decommissioned enhancers to activity-downregulated genes. Altogether this work begins to elucidate how the 3-D genome orchestrates cellular state and function decisions during mammalian brain development from the earliest neural lineage commitment through the refinement of connections between terminally differentiated neurons.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

A critical unanswered question in genome biology is how the tremendous diversity of neuronal subtypes and synaptic connections are established during development and maturation of the mammalian brain (**Figure 1.1**). Transcriptional signatures unique to each

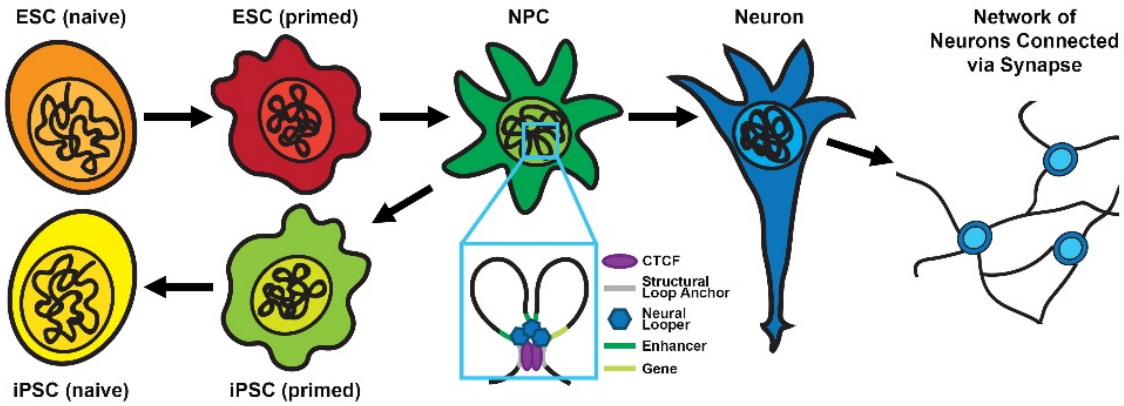


Figure 1.1. The link between 3D genome folding, neural differentiation and synaptic plasticity in human brain development and neurological disorders is unknown.

cell type must be intricately regulated in space and time to orchestrate the exquisite cellular and synaptic diversity of the adult brain[1]. A growing consensus is that numerous epigenetic modifications across the genome function together to create the 'Epigenome' - a molecular barcode on top of the linear DNA sequence that distinguishes one phenotype from another[2]. Genome-wide mapping studies have made great progress in elucidating the spatial distribution of epigenetic marks on the linear DNA polymer and how such marks differ among cell types. Nevertheless, there is still a large gap in our knowledge of how a wide range of epigenetic marks are spatiotemporally regulated to control the formation and cooperation of the extensive cellular heterogeneity in the developing brain. Understanding

the mechanisms governing differentiation and synaptogenesis in the healthy brain will shed new light into how these processes go awry in neurological disorders.

Mammalian genomes are folded into sophisticated configurations that both shape, and are shaped by, a diverse range of cellular functions[3]. Recent advances in molecular and computational technologies have enabled the query of higher-order chromatin architecture at unprecedented resolution and scale[4-6]. The emerging model from these studies is that the mammalian genome is folded into a complex hierarchy of highly self-interacting Megabase (Mb)-scale structures termed topologically associated domains (TADs), nested subTADs and long-range looping interactions [7-10] (reviewed in **Chapter 2**). The highest resolution maps to date have enabled the detection of tens of thousands of long-range looping interactions genome-wide [10, 11]. Loops connected by the architectural protein CTCF are thought to create TADs/subTADs that demarcate the search space of enhancers for their target promoters [12-15]. Enhancers loop to promoters via architectural proteins such as mediator and cohesin to govern spatiotemporally regulated transcription [16-19]. Initial studies have shown that long-range interactions can markedly reconfigure in development, disease, and in response to genetic perturbations [11, 14, 16, 17, 20-26].

Chapter 3 of this thesis begins with the simple question of how CTCF is reconfigured between pluripotent embryonic stem cells (ESCs) and multipotent neural progenitor cells (NPCs). Surprisingly we find a large number of CTCF binding sites that are lost during differentiation which is accompanied by a loss of looping at those sites. However, we find that within larger looping domains formed by constitutively bound CTCF, NPC-specific loops arise to connect NPC enhancers to their target genes such as *Nestin*, *Olig1-2*, and *Sox2*. These NPC-specific enhancer-promoter loops are often

mediated by the transcription factor YY1, knockdown of which disrupts the loops and alters gene expression. Thus, we implicate YY1 as an important regulator of early neural lineage commitment.

Chapter 4 of this thesis then tests how the 3-D genome is reconfigured when NPCs are reverted back to pluripotency through somatic cell reprogramming. We find that the CTCF sites that were lost initially during differentiation are often not efficiently restored, resulting in genome folding patterns that both retain signatures of the NPC state and exhibit pluripotency-specific enhancer-promoter interactions. Culture of iPS cells in 2i media conditions restored pluripotency-like CTCF binding, genome folding patterns, and expression of key pluripotency genes.

Finally, in Chapter 5 we investigate how the 3-D genome organizes the process of synaptogenesis. Neurons form an interconnected network in the mammalian brain. Synaptic connections among neurons allow the mammalian CNS to process and store information. An emerging body of evidence suggests that past synaptic activity of a neuron influences how that neuron operates within its neuronal networks in the future by regulating cellular properties such as dendritic outgrowth, synapse maturation, synapse elimination, and synaptic plasticity (reviewed in[27]). A critical component of this feedback pathway is an upregulation of hundreds of activity response genes (ARGs) rapidly upon neuron depolarization[28-35]; activity response genes such as *fos*[28-32] and *arc*[33-35] are expressed on the order of minutes[36] and are essential for proper long-term learning and memory[37]. A fundamentally important goal toward understanding complex brain functions such as learning and memory is to elucidate the molecular mechanisms governing gene expression changes occurring as a cause or consequence of neuronal activity and synaptic plasticity.

Chapter 5 seeks to begin to answer this question by mapping dynamic genome folding across neuronal activity states. We find that activity response genes loop to a subset of activity-induced enhancers in an activity-dependent manner. Surprisingly the complexity and kinetics of these loops were different depending on whether the gene is expressed in a rapid or delayed manner in response to neuronal activity. Finally, we observe that Schizophrenia and Autism Spectrum Disorder (ASD) genetic variants fall preferentially at the base of chromatin loops with different classes of activity-responsive enhancers. In Chapter 6 I propose future work that is required to establish the causal connection between 3-D epigenome reconfiguration and mammalian synapse formation and function which underlies memory and cognition.

CHAPTER 2: BACKGROUND

A fundamental mystery in genome biology is how three billion base pairs of DNA sequence (~2 meters long) is folded, looped and coiled to fit into a mammalian nucleus that is roughly 5-10 μm in diameter. Mounting evidence suggests that higher-order chromatin structure is linked to spatiotemporal regulation of a wide range of unique cellular functions (e.g. transcription, replication, recombination and repair)[23, 38-41]. Thus, a leading hypothesis is that chromatin packaging cannot be random but must be arranged in precise configurations that are amenable – and perhaps causally linked - to dynamic epigenetic modifications that orchestrate complex phenotypic outcomes.

Rapid progress has been made over the last decade in advancing our understanding of how the genome folds in three-dimensions (3-D)[42]. The emerging picture is that chromatin is arranged in a nested hierarchy of topological features with unique properties at each length scale[7, 9, 10, 38, 43-45]. The recent influx of new insight into genome folding has been primarily driven by advances in molecular and computational sequencing technologies, ultimately enabling scientists to overcome the resolution and throughput limitations of conventional microscopy[4, 46]. In this chapter I introduce the foundational insights into how the genome folds to regulate gene expression and cell identity, upon which my thesis will build.

2.1 How does the genome fold?

Metazoan genomes are folded into a nested series of unique 3-D configurations (illustrated in **Figure 2.1**). At the first level of the packaging hierarchy, the primary DNA

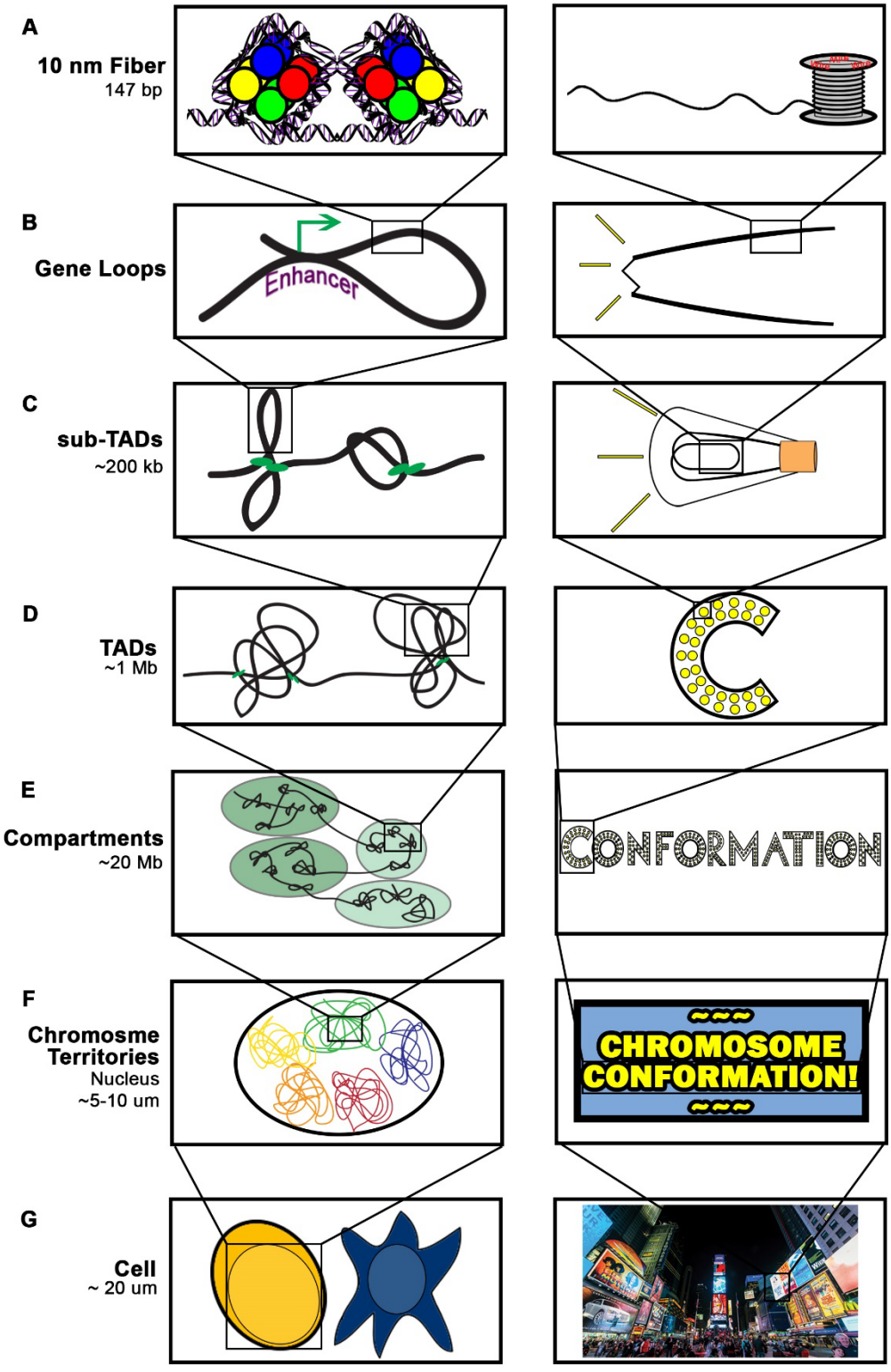


Figure 2.1. Diagram of hierarchical genome organization, ranging through the length scales of: **(A)** DNA wrapped around nucleosomes, **(B)** gene loops, **(C)** sub-TADs, **(D)** TADs, **(E)** compartments, **(F)** chromosome territories, and **(G)** the nucleus.

sequence is wrapped around histones to form nucleosomes that make up the 10 nm chromatin fiber[47] (**Figure 2.1A**). It has long been known that the 10 nm fiber is arranged into an interconnected web of long-range interactions, but the precise details of folding patterns at each length scale have remained unclear[48]. Recent molecular and computational breakthroughs have now enabled a more lucid and precise understanding of several unique chromatin folding configurations[4, 46, 49]. At the second level of packaging, long-range contacts between two non-adjacent loci on the linear genome facilitate the spatial proximity of distal regulatory elements via the "looping out" of intervening, non-contacting sequences[50-54] (**Figure 2.1B**). These so-called 'looping interactions' are hypothesized to in turn serve as the structural foundation that connects the edges of larger architectural folding units generally classified as 'contact domains' (CDs)[9, 10, 38, 42]. CDs are large genomic regions (i.e. numerous genomic loci in series) that have a higher interaction frequency with each other than the surrounding genome sequence – thus creating a domain-like architecture (**Figure 2.1C-D**). CDs are often nested within each other (discussed in detail below) and exhibit a large dynamic range in length scale (i.e. megadomains (5-20 Megabase (Mb)); topologically associating domains (TADs) (200 kilobase (kb) - 3 Mb); sub-TADs (40 kb - 1 Mb))[7-10, 55-57]. Importantly, spatial proximity among two or more TADs or smaller sub-TADs can create higher-order 'compartments' or 'sub-compartments', respectively (**Figure 2.1E**)[10, 44, 45]. Compartments are generally hypothesized to represent spatial neighborhoods of co-regulation within the larger 3-D nucleus[58]. Finally, at a highest level of organization in the hierarchy, individual chromosomes occupy distinct territories with respect to the other chromosomes and the nuclear periphery[45, 48] (**Figure 2.1F**). The spatial placement of territories can in turn affect

the genomic loci that are adjacent to the lamina and also physically proximal to key genetic sequences in other territories[59-61].

2.2 Hypothesizing 3-D architecture from 2-D contact density maps

Much of our recent knowledge of genome folding has been provided by newly published genome-wide data generated by Chromosome-Conformation-Capture (3C)-based methodologies. The unique technological details of each methodology have been extensively reviewed elsewhere[4, 5, 46]. Despite key procedural differences, 3C-based methods generally partition the linear DNA sequence into fixed interval bins that can be plotted on both X and Y axes of a contact density map. Contact maps serve as a 2-D grid in which any given pixel ($C_{i,j}$ where i and j are indices of intervals on X and Y axes, respectively) represents the relative interaction frequency between any two fixed interval bins on the genome. The 2-D contact grid is often visualized as a heatmap to reveal patterns of high and low frequency chromatin architecture (**Figure 2.2**).

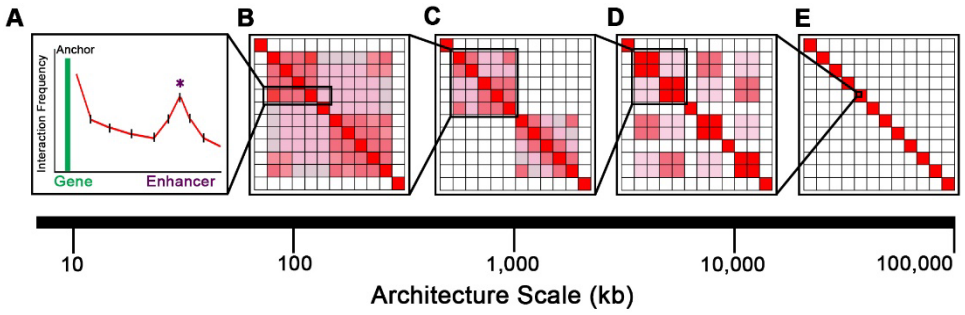


Figure 2.2. Representative heatmaps of chromatin interaction data at different length scales of interactions. Frequency of interaction is depicted on color scale ranging from white (low) to dark red (high). Heatmaps are depicted for the following organizational units: **(A)** gene loops, **(B)** sub-TADs, **(C)** TADS, **(D)** compartments, **(E)** chromosomes.

2.2.1 Looping interactions: single or clustered pixels in kilobase-resolution maps

Looping interactions are identified in proximity ligation data as pairs of genomic loci (generally <10 kb in size) that exhibit a higher contact frequency with each other than with adjacent loci. In 3C-PCR or Circular-3C (4C) data, a looping interaction is identified when a pre-determined anchor sequence exhibits a higher frequency interaction with a specific distal locus than with the intervening genomic sequence (**Figure 2.2A).** For high-throughput 3C methods that query large portions of the genome, such as Hi-C and Chromosome-Conformation-Capture-Carbon-Copy (5C), looping interactions are represented as single or clustered groupings of pixels at key loci in contact density heat maps (**Figure 2.2B**). In practice, it is easiest to discriminate sufficiently between adjacent pixels to reveal underlying looping structure if Hi-C and 5C maps are <10-15 kb resolution (discussed in detail below).

2.2.2 TADs vs. subTADs vs. contact domains: a question of length-scale and resolution

Recent Hi-C maps have uncovered a clear underlying structure to looping interactions and how they intertwine with several higher-order levels of genome organization[10, 38, 57, 62, 63]. For example, in addition to connecting distal regulatory sequences, a leading hypothesis is that looping interactions might form the structural basis for larger architectural folding units termed 'topologically associated domains' (TADs)[7, 8, 24, 55, 64] (**Figure 2.2C**) and their smaller counterparts termed sub-TADs[9] (**Figure 2.2B**). TADs and sub-TADs – recently referred to more generally as contact domains - range in size from 40 kb to 3 Mb[10]. Noteworthy, sub-TADs are often nested within larger TADs[9] (**Figure 2.2B-C**) and both are represented as large squares of elevated interaction

frequency in contact density heat maps. Often, a group of adjacent, high-count pixels are located at the apex of the square, indicating a further enrichment for a loop between genomic segments on opposite edges of the TAD/subTAD – suggesting that looping interactions might form the structural basis for larger CDs. There is also evidence that TADs can then assemble into much larger 5-20 Mb 'Megadomains'[10].

A critical issue to consider when interpreting proximity ligation data is that the underlying structure visualized in heatmaps is dependent upon the resolution at which the experiment is performed and analyzed. The resolution of interaction frequency matrices is often discussed in terms of the number of base pairs, 'n', at which a matrix should be binned (resulting in heatmap pixels with dimensions n x n). For example, Rao et al.[10] define Hi-C mapping resolution as the smallest binning size in which the underlying features of genome folding can be reliably distinguished. Several factors contribute to resolution (e.g. library complexity, genome size, genome coverage, sequencing read depth). As a consequence, published data is available at mapping resolutions ranging from low (~1 Mb-sized pixels) to high (~250-1000 bp pixels).

During the design and analysis of a proximity ligation experiment, it is of utmost importance to select a resolution that is appropriate for the specific length scales of genome folding that will be studied. For example, in genome-wide 3-D folding maps binned at 1-10 Mb resolution, the genome appears organized into a series of large 'Megadomains' (5-20 Mb)[10, 45]. By contrast, matrices binned at 40 kb mapping resolution readily display TADs (median size of 880 kb) tiled along the diagonal of heatmaps[7]. Moreover, higher-resolution maps (~5-10 kb resolution) exhibit sub-TAD structures nested within larger TADs[9]. Ultra-high resolution maps at ~250 bp-5 kb resolution highlight looping

interactions and sub-TADs, but lose sensitivity in resolving larger TAD and Megadomain structures[10]. Thus, it is important to tune experimental and computational parameters to query genome folding features at the length scale of interest to a particular biological question.

2.2.3 Compartments & sub-compartments: spatial neighborhoods of contact domains

A/B compartments were first identified in Mb-resolution maps as ultra-long-range, off-diagonal interactions among a series of two or more TADs with similar chromatin features[45] (**Figure 2.2D**). For example, 'A' compartments are represented by marks characteristic of open chromatin, such as DNAseI hypersensitivity, high gene density, high transcription, and active chromatin marks such as H3K4me3, H3K36me3 and H3K4me1. By contrast, 'B' compartments are represented by marks characteristic of closed chromatin, including: H3K27me3, H3K9me3, association with the nuclear lamina, absence of DNAseI hypersensitivity, low gene density and/or low/silenced transcriptional activity[10, 44, 45, 58]. Recently, the highest resolution Hi-C maps to date have uncovered that sub-TADs within larger TADs can co-localize with other sub-TADs via ultra long-range interactions to create sub-compartments – thus overturning the model that compartments are only formed by a series of TADs[10].

To date, the functional role for Compartments/sub-Compartments remains unknown. A recent Hi-C analysis showed that gene expression is only marginally altered during the switch between compartments during stem cell differentiation, suggesting that the finer-scale temporal and developmental regulation of individual genes is not causally governed by the larger scales of genome architecture[44]. We speculate that compartments

serve as sub-nuclear locations in which large contiguous chromatin regions can spatially co-localize to utilize similar genome machinery. Thus, active or similarly replicating TADs might share one set of chromatin-modifying enzymes (and thus co-localize in A Compartments/sub-Compartments), whereas silenced TADs/sub-TADs might share another very different set of chromatin modifying enzymes (and thus co-localize in B Compartments/sub-Compartments).

2.3 Organizing principles governing looping

The mechanisms regulating the establishment, maintenance and function of 3-D architecture remain scarcely understood – largely because 3-D genome folding studies to date have been descriptive in nature. At the folding level of looping interactions, it remains unknown how specific genomic loci find their precise contact point. Furthermore, it remains unclear whether loops are a cause or consequence of transcriptional activity.

2.3.1 What governs the specificity and directionality of CTCF-mediated interactions?

The most well understood mechanism regulating loop formation involves CCCTC binding factor (CTCF). CTCF contains a highly conserved eleven zinc finger central DNA binding domain embedded within slightly more divergent N- and C-termini[65]. The protein was originally described with in vitro biochemical studies as a 'multivalent factor' due to its ability to bind to a wide range of variant sequences through combinatorial use of different zinc fingers[66]. Since its discovery 25 years ago, the function of this protein has long been shrouded in controversy due to its pleiotropic effects on genome function in vivo. Indeed, CTCF has been specifically linked to transcriptional activation, repression,

splicing, recombination, insulation, and imprinting[67]. Yet the mechanisms by which CTCF performs these distinct functions remain unresolved.

An important question is whether CTCF is a true multivalent factor with the ability to perform many contrasting functions, or if there is a single unifying mechanism that can explain its divergent roles. Recently it has been proposed that CTCF has a conserved role across metazoans as a master architectural protein that orchestrates different categories of chromatin interactions as a function of the combination of zinc fingers engaged with the genome. The "master weaver" hypothesis would predict that all genome regulatory roles linked to CTCF would be secondary effects of its architectural role[67]. Proximity ligation studies support this idea. For example, several seminal 3C-PCR studies have identified CTCF at the base of specific looping interactions and confirmed that knockdown of the protein resulted in decreased/abrogated looping[68-70]. A genome-wide analysis with ChIA-PET (chromatin interaction analysis by paired-end tag) uncovered ~1,500 CTCF-mediated looping interactions in pluripotent cells[71]. More recently, high-resolution (~10 kb resolution) architecture maps surrounding developmentally regulated genes demonstrated that a large subset of looping interactions are unchanged during pluripotent stem cell differentiation and are enriched for CTCF and its binding factor cohesin[9]. Similarly, the highest resolution genome-wide contact maps to date (~1-5 kb resolution) identified ~10,000 looping interactions across the genome. The large majority of interactions identified in this study were anchored by CTCF[10]. Consistent with these results, a genome-wide Hi-C analysis demonstrated that CTCF knockdown in HEK293T cells disrupts looping interactions within subTADS[72]. Together, these data support the idea that CTCF

has a causal role in facilitating looping, but the mechanism by which CTCF works to create long-range interactions remains a significant unaddressed issue.

CTCF's capacity to confer vastly different functions has been attributed to distinct conformations of the protein, which are posited to be governed by differential ZF binding to divergent consensus sequence variants[65]. The 'CTCF code hypothesis' has remained unproven to date, but recent reports have begun to unravel possible links between the underlying genome sequence and 3-D architecture[73-78]. ChIP-seq mapping studies across more than 20 mammalian cell types have uncovered 50,000+ CTCF binding sites. Biochemical and computational studies have identified a 20 bp core consensus sequence that engages with ZF's 4-7 and is essential for occupancy of the protein [65, 79, 80] (**Figure 2.3A**). Moreover, a fraction (10-25%) of consensus sites are flanked by additional secondary motifs that are hypothesized to stabilize CTCF binding[81-83]. Seminal studies exploring the interplay between ZF's and genome sequence relied on in vitro transcribed CTCF mutants and gel shift assays. Recently, Nakahashi et al. used cell lines overexpressing CTCF ZF mutants in combination with ChIP-seq to confirm that the central ZF's 4-7 are essential for binding to the core 20 bp consensus sequence[81]. Importantly, this study also linked ZF's 9-11 association with the upstream motif, thus demonstrating that the protein's orientation can be regulated by the directionality of the consensus sequence. Finally, this study also systematically identified a relationship between CTCF binding affinity and single nucleotide variants within the core consensus sequence, highlighting the critical importance of CTCF occupancy patterns in genetic diversity that might be linked to phenotypic variation and disease susceptibility among individuals.

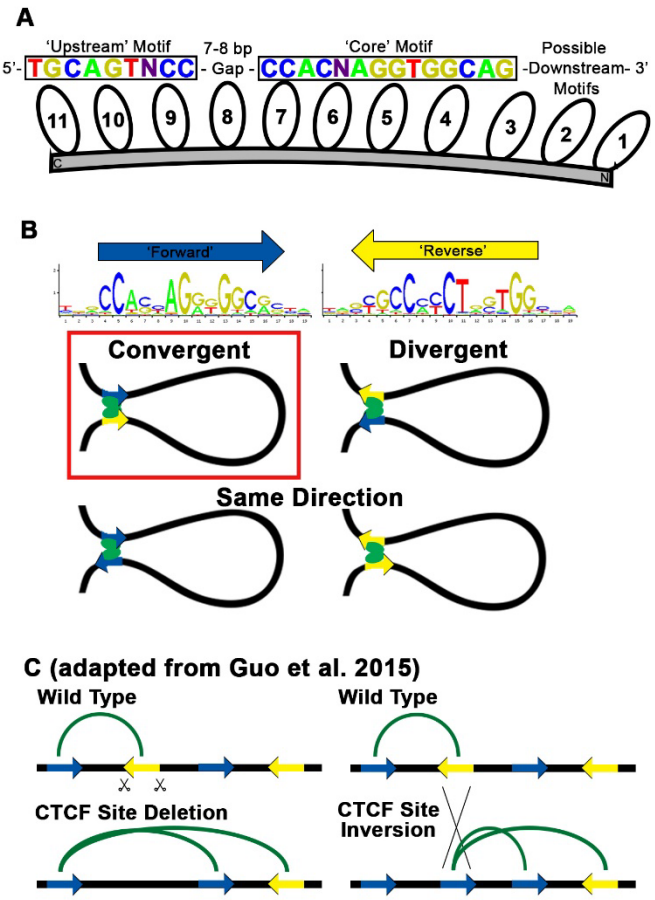


Figure 2.3. CTCF binding orientation influences local chromatin architecture. **(A)** CTCF's eleven zinc-fingers bind distinct DNA sequences within the canonical CTCF binding motifs, resulting in CTCF engaging with the genome in specific directions depending on the underlying sequence. **(B)** The four combinations of 'direction' that two CTCF motifs along the same DNA strand can occupy are displayed. The majority of CTCF motif pairs that form significant three-dimensional interactions are in the 'convergent' orientation. **(C)** Representative diagram of changes in chromatin interactions upon deletion/inversion of CTCF binding sites, based on data presented in Guo et al. 2015. Significant interactions are represented as green arches.

A leading prediction from the 'CTCF code hypothesis' is that alterations in CTCF's conformation due to the underlying genome sequence will ultimately impact the manner in which CTCF organizes higher-order genome folding. Indeed, very recent reports have

provided the first descriptive evidence that the orientation of CTCF binding motifs with respect to the linear genome might be a critical feature governing the specificity/selectivity of long-range interactions among CTCF binding sites[10, 62, 84-86]. Starting with a group of ~10,000 loops identified by ultra-high resolution Hi-C analysis in human cell lines, Rao et al. focused on a subset of ~4,000 loops that (1) exhibit CTCF binding at both loop anchors and (2) the anchoring CTCF sites contain only a single CTCF consensus sequence. In this group of CTCF-mediated loops, >90% contain CTCF binding motifs in a 'convergent' or 'forward-reverse' orientation. Specifically, one anchor at the base of a loop exhibits a 'forward' orientation (the CTCF motif is aligned in the 5'-3' direction along the strand in question) and the second anchor at the base of a loop exhibits either (1) a 5'-3' consensus orientation on the antisense strand or (2) the reverse complement of the consensus in a 5'-3' orientation (**Figure 2.3B**). Similarly, Hadjur and colleagues reported that the directionality of CTCF's consensus sequence is correlated with directionality in looping. More recently, using CTCF ChIA-PET analysis, Tang et al. focused on ~35,000 loops in which CTCF motifs were found at the base of both loop anchors with a unique orientation. In this case, ~65% of interactions showed a convergent consensus orientation while >30% exhibited tandem/same-direction orientation. Intriguingly, the loops with tandem consensus orientation exhibited lower interaction strength than loops with convergent consensus orientation, suggesting that each looping class might have a different function and/or that differential thresholding during the "peak-calling" of looping interactions might influence the results. Independent studies from Guo et al. and de Wit et al. have also reported 10-30% of looping interactions with tandem orientations of the CTCF consensus. Together, these results suggest that the CTCF consensus orientation is an important

contributing factor regulating looping interactions and predict that a convergent orientation is favorable, but not necessarily essential, for a looping interaction to occur.

Guo et al. have recently employed the CRISPR-Cas9 system to explore the causal link between convergent CTCF consensus sequences and looping interactions[84]. Within the protocadherin gene cluster, the authors inverted a specific regulatory sequence containing two CTCF binding sites oriented in the 'reverse' direction in wild type cells (**Figure 2.3C**). The regulatory sequence encodes a putative enhancer element known to loop to upstream alternative promoters in the alpha protocadherin gene cluster. Consistent with the convergent CTCF looping model, CTCF binding sites at alternative alpha promoters contain CTCF sites oriented in the 'forward' direction. Inversion of the protocadherin enhancer element significantly disrupted the convergent promoter-enhancer looping interactions without depleting CTCF binding. Similarly, de Wit et al. also used CRISPR-Cas9 genome editing to demonstrate that deletion of a CTCF binding site can abrogate looping interactions[85]. Intriguingly, re-insertion of CTCF at its endogenous location but in opposite orientation did not fully recover endogenous looping. Thus, these results provide the first evidence that convergent orientation of the CTCF consensus is an important mechanistic feature that causally contributes to chromatin looping.

Additional observations from CRISPR-Cas9 genome editing studies suggest that the CTCF consensus orientation model is likely not the only causally important principle governing looping. Intriguingly, after inversion of the protocadherin enhancer element, new, ectopic loops were formed between CTCF sites in both convergent and same direction orientations[84]. This was not a locus-specific observation, as Guo et al. also observed inverted ectopic loops with both convergent and same direction consensus orientations

upon CTCF site inversion at the β-globin locus. Presumably, if convergent consensus orientation was the sole mechanism governing looping, one would expect it would still apply in the establishment of de novo ectopic loops. Moreover, convergently oriented consensus sites mediate higher strength looping interactions than same direction consensus sites. To explain these conflicting results, we favor a model in which convergent CTCF orientation is favorable, but not essential, for high-affinity looping interactions; additional regulatory mechanisms likely work in concert with CTCF orientation to govern the specificity and downstream transcriptional activation facilitated by CTCF-mediated interactions.

2.4 Loop extrusion is a leading mechanism that governs chromatin domain formation

Significant progress has been made toward understanding the mechanisms that govern the formation of both unnested (**Fig. 2.4A-B**) and nested (**Fig. 2.4C-D**) chromatin domains. Mammalian genomes contain a large number of domains structurally characterized in Hi-C maps by the presence of 'corner dots' -- a punctate group of adjacent pixels with significantly enhanced interaction frequency compared to the surrounding local domain structure (**Fig. 2.4E, Fig. 2.4F**). Corner dot structures are thought to represent long-range looping interactions (schematically drawn in **Fig. 2.4G**) that exhibit a persistently high interaction frequency in a large proportion of cells (i.e. persistent loops). It has been hypothesized that chromatin domains which co-localize with corner dots at their apex represent so-called loop domains. Our own qualitative observation of Hi-C maps in mammalian systems reveals the presence of Mb-scale, unnested loop domains and nested loop domains (**Fig. 2.4E, Fig. 2.4F**).

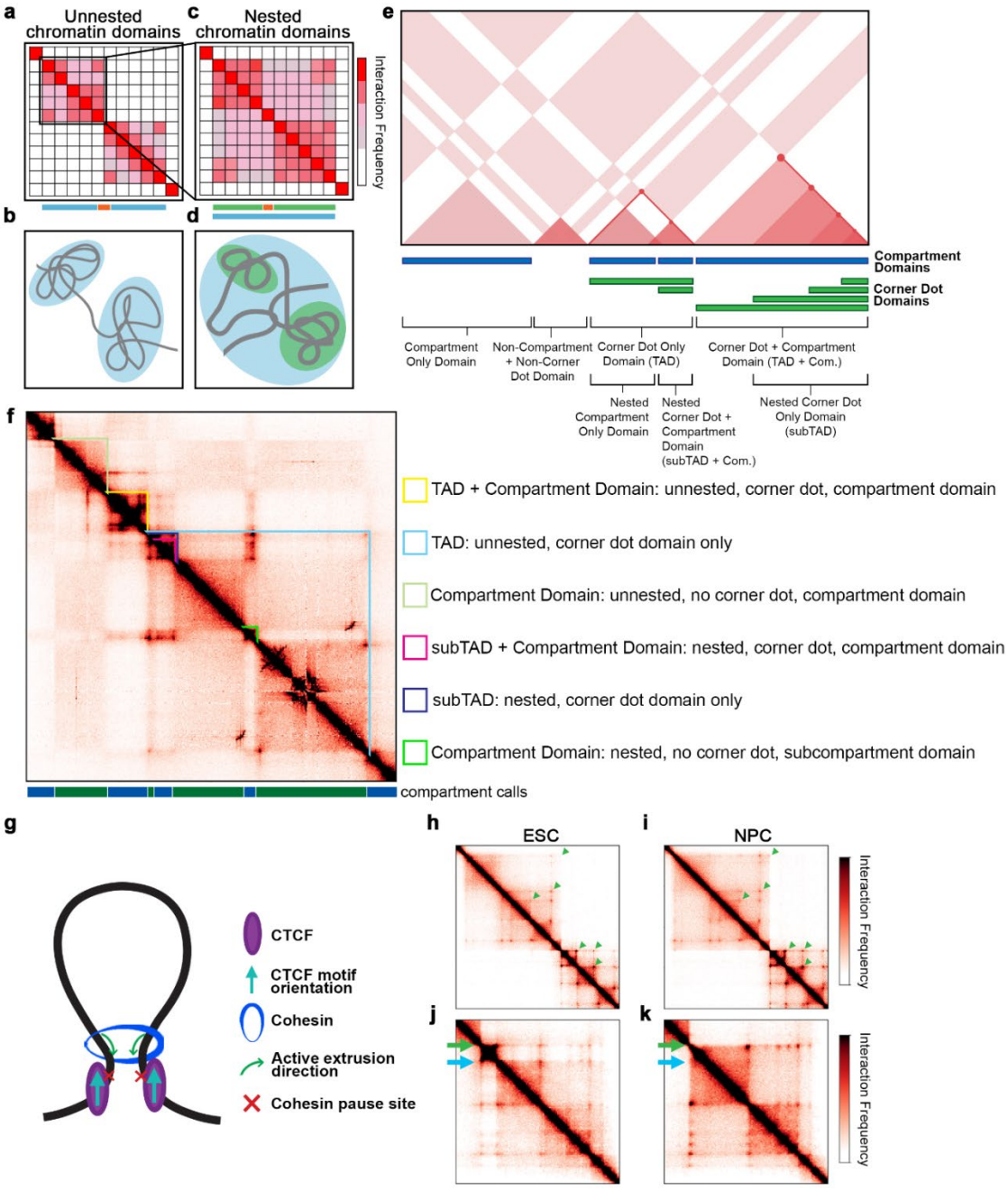


Figure 2.4. The structural features of topologically associating domains. (A-D) Heatmap representations (top) and schematized globular interactions (bottom) of topologically associating domains (TADs, A-B) and nested subTADs (C-D). **(E)** Cartoon representation of different classes of contact domains parsed by their structural features and degree of nesting. **(F)** Identification of contact domains classes from (e) in cortical neuron HiC data from Bonev et al. 2017 binned at 10 kb resolution. **(G)** Cohesin translocation extrudes DNA in an ATP-dependent manner into long-range looping interactions that form the topological basis for

TAD and subTAD loop domains. **(H-K)** Contact frequency heatmaps of high resolution Hi-C data (Bonev et al. 2017) performed on embryonic stem cells (ESC, H+J) and neural progenitor cells (NPC, I+K). **H-I:** Green arrows denote the corners of a subset of the nested chromatin domains evident in this genomic region. **J-K:** Green arrow annotates a high insulation strength, cell type invariant TAD boundary. Blue arrow points to a lower insulation strength, cell type dynamic subTAD boundary.

Recent reports and forthcoming studies by large consortia have identified 10,000-50,000 corner dot structures representing persistent loops in various human cell types[10, 11]. The majority of corner dots are anchored by motifs bound by the architectural protein CTCF[67]. Specifically, 60-90% of all corner dots (estimates vary across studies) with an interpretable CTCF motif in both anchoring fragments display a 'convergent' motif orientation[10,86] (**Fig. 2.4G**). Inversion of CTCF motifs using CRISPR genome editing disrupt the corner dot and the TADs/subTADs demarcated by the dot, demonstrating that convergent CTCF motif orientation is necessary for the formation of loop domains[84, 85, 87]. Moreover, short-term degradation of the CTCF protein results in severe ablation of a large proportion of loop domains[13]. Thus, a significant subset of persistent loops represented by corner dots require binding of the architectural protein CTCF in a convergent orientation on both loop anchors.

A windfall of new data has also recently advanced our understanding of the manner in which the two convergently oriented CTCF binding sites establish and maintain spatial proximity. In principle, the orientation of CTCF motifs should not matter if loop establishment occurs through simple diffusion in the 3-D nucleus. The seminal model of 'loop extrusion' asserts that molecular motors loaded on the genome could track along the DNA sequence, thus 'extruding' the intervening DNA in the process[88, 89]. Compelling

evidence supporting this theory was provided by computational studies indicating that polymer simulations of loop extrusion could recapitulate loop domains from Hi-C maps[87, 90-92]. The authors of these studies predicted the existence of DNA extruding factors.

The mechanisms governing loop extrusion are an intense area of international investigation. It has long been thought that structural maintenance of chromosomes (SMC) complexes, such as cohesin or condensin, could serve as loop anchoring factors, either by stabilizing pre-formed loops or through an active extrusion mechanism. Peaks of enriched cohesin occupancy on DNA identified via ChIP-seq co-localize with CTCF binding sites[79, 93-95], but are slightly shifted to the 3' end of convergently oriented motifs[86, 91]. This was a clue suggesting a tracking mechanism of cohesin-CTCF recruitment. Knock-out of the cohesin release factor WAPL resulted in increased cohesin residence time on the genome, longer looping interactions that cross conventional TAD boundaries, and a marked increase in the number of both TAD and nested subTAD loop domains[96]. Moreover, knock-out of the cohesin loading factors Scc4 and Nipbl, or the Rad21 cohesin subunit, ablated a large fraction of loop domains across multiple mammalian cell types[18, 19, 96]. Direct evidence supporting loop extrusion via SMC complexes came from single molecule imaging studies showing that condensin[97, 98] and cohesin[99-101] can translocate along naked DNA *in vitro* in an ATP-dependent manner. Thus, loop extrusion, in which SMC complexes pass over divergently oriented CTCF motifs and stall at those in convergent orientation (**Fig. 2.4G**) has been proposed as a leading hypothesis for the mechanism of loop domain formation.

We also define a key subgroup of chromatin domains that neither co-localize with corner dots nor register with compartments (**Fig. 2.4E, Fig. 2.4F, 'non-compartment+non-corner dot domains'**). It is important to highlight that, for those

domains formed by extrusion mechanisms, preferential contacts within the domain (i.e. not at corner dots) are hypothesized to be composite signal of active extrusion events (i.e. transient loops)[91]. Therefore, it is possible that non-compartment+non-corner dot domains are mechanistically formed by extrusion. Furthermore, an important area of active exploration is the discovery and dissection of additional extrusion blocking factors. Precise annotation of the suite of diverse proteins that influence extrusion rates across the genome would give credence to the hypothesis that boundaries with unique molecular characteristics can give rise to differential extrusion blocking strength, thus causing corner dots with varying interaction frequency. Thus, alternative mechanisms that could contribute to non-compartment+non-corner dot domains include: (1) loop extrusion against transient boundaries (i.e. highly dynamic boundaries in individual cells), (2) loop extrusion against weak boundaries present in a high proportion of cells, (3) so-called 'exclusion boundaries' in which the boundaries are strong and contribute to extrusion blocking in the TADs/subTADs upstream and downstream of the domain in question, therefore the non-corner dot domain is created as a consequence of placement between two strong TADs, or (4) novel still unknown mechanisms. Future studies to unravel the mechanisms that form transient versus persistent loop domains are of high importance for future inquiry, and the field's progress in understanding the mechanisms governing persistent loop domains is discussed in this Perspective.

2.5 Compartmentalization is a second mechanism that contributes to chromatin domains

A second mechanism that contributes to the establishment or maintenance of

chromatin domains in eukaryotes is compartmentalization. Compartments were initially identified in 1 Mb binned Hi-C heatmaps by their chromosome-wide plaid pattern of ultra-long-range intra-chromosomal and inter-chromosomal contacts[45] (**Fig. 2.4E**). It has been hypothesized that the empirically defined plaid pattern represents the partitioning of the human genome into either A compartments of actively transcribed genes and active histone marks or B compartments with inactive genes and repressive marks[45]. The initial low-resolution Hi-C maps suggested that multiple Mb-scale TADs were nested within a single contiguous segment of an A or B compartment. However, in high-resolution heatmaps it was recently discovered that the mammalian genome was instead partitioned into at least six significantly smaller sub-compartments with various combinations of repressive and active chromatin modifications[10]. Notably, ultra-high-resolution Hi-C maps in flies have uncovered so-called 'compartment domains' - fine-grained compartments that perfectly register with chromatin domains devoid of corner dots[102]. Indeed, the overall quantity of corner dot domains in flies is minimal[102, 103], suggesting that compartmentalization may be the primary driver of chromatin domain formation at least in some non-mammalian eukaryotic organisms.

Together, these high-resolution analyses provide evidence that an intriguing subset of chromatin domains across eukaryotes could be classified as 'compartment domains' due to a perfect alignment between the domain-like structure and compartment coordinates and the absence of a corner dot (**Fig. 2.4E, Fig. 2.4F, 'compartment domain only', 'nested compartment domain only', Definitions Box**). A critical unanswered question is whether loop extrusion occurs in organisms where compartmentalization is the driving chromatin domain mechanism, and, if so, which proteins serve as the extrusion factors, and how the

extrusion blockers work to circumvent the formation of corner dot TADs/subTADs. Ongoing and future work, some of which is discussed below, will shed light on the cause-and-effect relationship of compartment domains in governing transcription, compartment organizing principles, their unknown abundance in unperturbed mammalian genomes, and their interplay/competition with other genome organizing forces.

2.6 What's in a name? Refining the definition of TADs/subTADs as loop extrusion domains mechanistically distinct from compartment domains

One question under intense debate is how to update the historical definitions of TADs/subTADs in light of recent discoveries, most importantly the existence of loop extrusion and the striking competition between compartmentalization and looping mechanisms that underlie the formation of chromatin domains[18, 19, 96, 102]. Indeed, cohesin knock-down results in strengthening of existing compartments and finer-scale compartmentalization upon loss of corner dot TADs/subTADs in mammalian systems[18, 19, 96]. These results suggest that loop extrusion and compartmentalization are distinct and competing forces, thus reinforcing the concept that chromatin domains formed by the two mechanisms need to be uniquely and clearly defined.

Data thus far are consistent with a model in which a subset of both TADs and nested subTADs represent composite signals of loops in the making and thus are loop domains established by dynamic extrusion of SMC complexes blocked by boundaries created by architectural proteins such as CTCF. Importantly, TADs as originally historically discovered are also strongly demarcated by CTCF[7]. Thus, in an effort in this Review to link the definition of TADs to underlying mechanism, we propose to refine the definition

of TADs as unnested, corner dot domains formed mechanistically by persistent loops (**Fig. 2.4, Table 2.1**). TAD loop domains may also be sub-stratified into those that also perfectly correspond to compartments or do not co-localize with compartments (**Fig. 2.4E, Fig. 2.4F, 'TAD only', 'TAD+compartment domain'**). We refine the definition of subTADs as nested, corner dot domains formed mechanistically by persistent loops. subTAD loop domains may also be substratified into those that perfectly correspond to compartments or do not co-localize with compartments (**Fig. 2.4E, Fig. 2.4F, 'nested subTAD only', 'nested subTAD+compartment domain'**). As discussed above, we define the most abstract and poorly understood domain type (i.e. 'non-compartment + non-corner dot domains'), as those that do not correspond to compartments and are not persistent corner dot TADs/subTADs, but could still be created by extrusion blocking from weak boundaries or still unknown mechanisms (**'non-compartment + non-corner dot domains', Fig. 2.4E, Fig. 2.4F**). Evaluating the possible functional or mechanistic difference between loop domains that also co-localize with compartments and loop domains that do not register with compartments is of high importance for future functional and mechanistic dissection.

2.7 TADs, subTADs, and their boundaries can be structurally distinguished by their nested properties

Another currently debated question is whether contact domains are folded

Definitions	Structural Observation	Hypothesized Mechanism
Chromatin Domain	Small triangles of enhanced contact frequency that tile the diagonal of each contact matrix.	
Compartmentalization	Plaid pattern in HiC maps, allowing alternating A/B designations for genomic intervals that display similar plaid patterns.	Co-segregation of chromatin with similar histone marks / bound proteins. Possibly formed in part via phase-separation forces.
Compartment Domain (CD)	A chromatin domain whose boundaries align with inflection points in A/B compartmentalization signal.	
Loop	A point of enriched contacts in HiC heatmaps. Will appear as **dot** (a series of adjacent pixels with enhanced contact frequency with respect to the local chromatin domain structure) if loop occurs in many cells at time of fixation (thus not all loops present across a population manifest structurally as dots due to the transient nature of extrusion).	Active cohesin extrusion of chromatin which is paused by proteins bound to the genome, most notably CTCF.
Loop domain (TAD/subTAD)	A contact domain formed via loop extrusion mechanisms. Often have a dot at their corner (**corner dot domain**), but due to the transient nature of extrusion some do not (**transient loop domain**).	Extrusion pausing manifests in a domain boundary.
Loop and compartment domains are not mutually exclusive. Their overlap and nesting properties enables the identification of 6 distinct chromatin domain classes: **1) TAD + CD**: unnested, loop + compartment domain **2) TAD only**: unnested, loop domain only **3) CD only**: unnested, no looping signatures, compartment domain **4) nested subTAD+CD**: nested, loop, compartment domain **5) nested subTAD only:** nested, loop domain only **6) nested CD only:** nested, no loop, subcompartment domain		

Table 2.1. Chromatin domain definitions box

hierarchically, or if the largest, Mb-scale TADs are simply an artifact of the high spatial noise and low resolution of early Hi-C maps. For example, a recent report in one specific cell type has suggested that the Drosophila genome is partitioned into relatively small compartment domains tiled along the diagonal[102]. An emerging interpretation of new high-resolution Hi-C data is that Drosophila may only have a small number of loop domains[102, 103], and this important structural feature will require confirmation across a range of cell types and functional studies. There is less evidence for nesting in Drosophila than in mammalian systems, suggesting that complex hierarchical domain structures might be less prominent in some organisms. Although more analyses are required to quantitatively resolve the existence of nested domains across species and cell types, it is worth pointing

out that there is strong visual evidence of large TADs and smaller, nested subTADs in the highest resolution Hi-C maps published to date in mouse[104] (**Fig. 2.4H-I, green arrowheads**). Thus, in addition to the classification of TADs as compartment and non-compartment domains loop domains (detailed above), we hypothesize that it is also important to stratify chromatin domains and their boundaries by their nested properties during the design and interpretation of functional and mechanistic experiments (**Fig. 2.1E-F**).

Several lines of evidence support the possibility that nested versus unnested boundaries might have different structural and functional properties. First, Mb-scale TADs are largely cell type-invariant, whereas subTADs exhibit a higher tendency to reconfigure in a cell type-specific manner[7, 9, 105]. In mammalian systems, boundaries on both sides of unnested TADs are conventionally cell type-invariant. Moreover, we observe that one of the subTAD boundaries will often co-localize with TAD boundaries, and in these cases the boundaries are typically invariant across cell types (**Fig. 2.4J-K, green arrow**). By contrast, many subTAD boundaries, often the side truly nested within larger TADs, exhibit cell type-specific structural features (**Fig. 2.4J-K, blue arrow**). Moreover, because long-range interactions occur more frequently over boundaries demarcating nested versus unnested domains, subTAD boundaries exhibit mechanistically weaker insulation than TAD boundaries.

Together, these results leave open the possibility that TAD and subTAD boundaries are regulated by unique organizing principles and might play distinct functional roles. Indeed, we hypothesize that extrusion may assemble both TAD and subTAD corner dot domains, but that the nested, cell type-specific boundaries unique to subTADs might be

governed by different densities or types of architectural proteins than those at unnested, invariant boundaries. Interestingly, recent reports have reported the role for transposable elements in the formation of cell type-specific boundaries[106, 107]. Progress toward testing this hypothesis will be further expedited by computational methods to sensitively and accurately identify the full sweep of domains in ultra-high-resolution Hi-C data. Thus, an important area for future inquiry will be to unravel the structural, functional, and mechanistic differences among boundaries across length scales.

In this Review, we define TADs according to their structural manifestation in the historically first Hi-C maps as Mb-scale continuous genomic intervals (or blocks) in which DNA sequences exhibit significantly higher interaction frequency with other DNA sequences within the block compared to those outside of the block[10]. We also add additional qualifiers: (1) TADs are formed by loop extrusion and contain corner dots indicative of strong extrusion boundaries and persistent loops and (2) TADs, and their respective boundaries, should be at the top level of the domain folding hierarchy and cannot be further nested under larger, on-diagonal corner dot domains (**powder blue corner dot domain, Fig. 2.4F**). We define subTADs as sub-Mb scale corner-dot domains that are nested within larger TADs (**purple corner dot domains, Fig. 2.4F**). subTAD boundaries exhibit weaker long-range contact insulation than those at the top of the folding hierarchy; the molecular basis for this difference and if it is functionally significant remains to be uncovered. Finally, we note that although 'mini-domains' or 'microTADs' have recently been used to describe the smallest scale chromatin blocks encompassing a single gene unit in mammals[108, 109] and flies[102], we currently do not define them in this Review. If further studies illuminate that gene unit domains have corner dots and are created by

loop extrusion, then we suggest to either continue to define them as nested subTADs or re-define them as ultra-nested micro-TADs. However, if future studies indicate that gene unit domains are not formed by loop extrusion, then they should be defined in their future by their mechanism of formation, whether it be by compartmentalization and/or phase separation or a novel organizing principle.

2.8 Chromatin domains and boundaries are clearly present but stochastically detected in single cells

It has long been emphasized that chromatin domains were empirically defined from Hi-C maps – and thus they may only represent an ensemble average interaction frequency across millions of cells. Do domain-like structures indicative of compartment domains, TADs, or subTADs exist in individual eukaryotic nuclei? Seminal single cell Hi-C studies shed initial insight into this question, suggesting that even sparse, low complexity matrices created from individual nuclei were consistent with the possibility that domain-like structures could exist in single cells[57, 110]. Recent super resolution microscopy experiments coupled with Oligopaint probes have enabled the direct visualization of the spatial positioning of thousands of adjacent genomic loci. Consistent with single cell Hi-C, Oligopaint experiments confirmed that genomic loci are spatially grouped into high interaction frequency interaction domains in individual mammalian cells (**Fig. 2.5A**) [111]. Importantly, the most frequently detected boundaries in single cells occurred at the locations predicted by ensemble Hi-C maps[111] (**Fig. 2.5A**). Many wild type single cells also showed random placement of domain-like blocks, which is consistent with the established transient

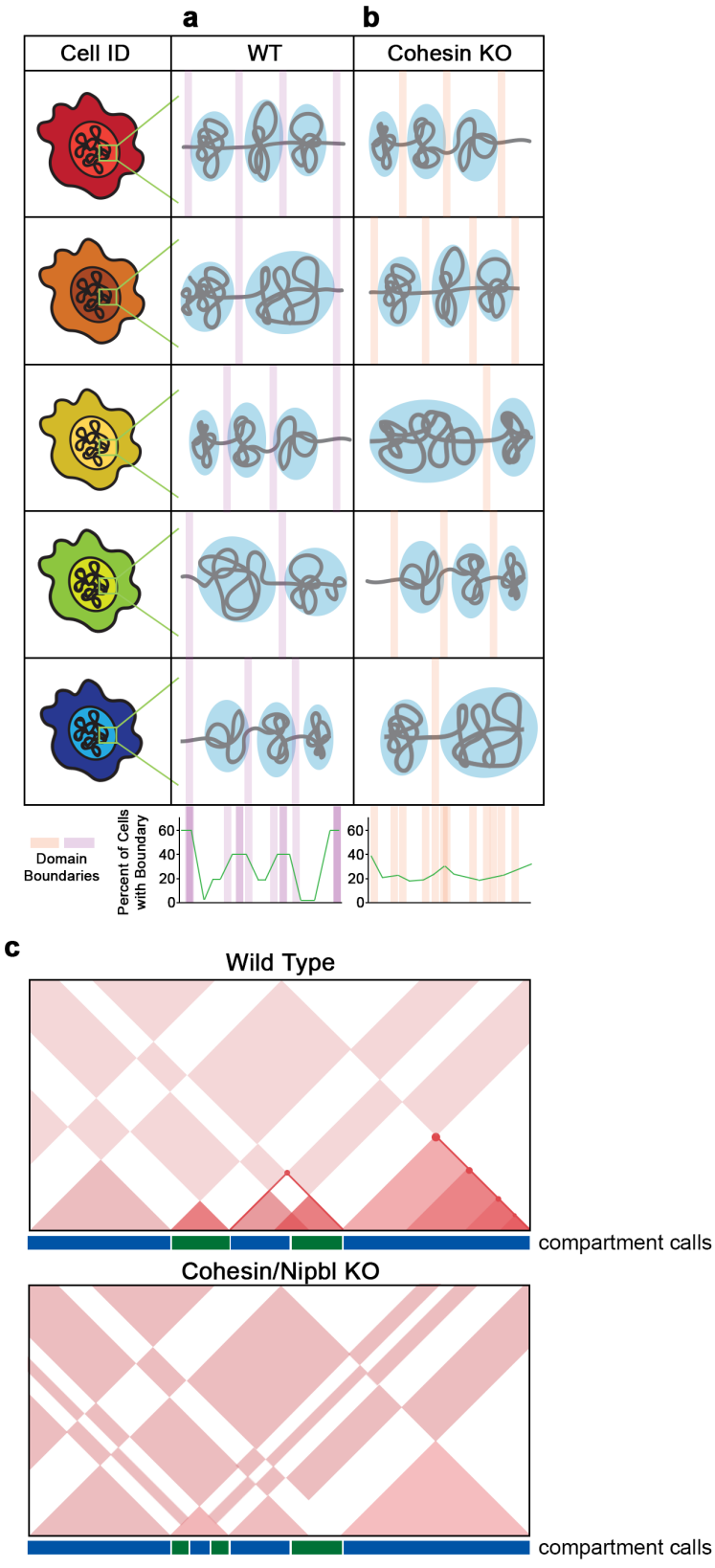

a
b
Cell ID
WT
Cohesin KO
Domain Boundaries
Percent of Cells with Boundary
60
40
20
0
c
Wild Type
compartment calls
Cohesin/Nipbl KO
compartment calls

Figure 2.5. Chromatin domains and their boundaries are present in single cells. (A-B) Cartoon representations of contact domains identified in single cells via high resolution imaging (Bintu et al. 2018). **A:** Wild type cells displayed a biased preference for boundary locations. **B:** Upon cohesin degradation, globular domains still existed but did not display the same boundary preference. **(C)** Representative heatmaps of the effects of cohesin/Nipbl removal on loop and compartment domains, as portrayed in Rao et al. 2017 and Schwarzer et al. 2017.

nature of the extrusion process and would be expected due to imaging a snapshot in time across a populations of individual cells in which extrusion was not synchronized. Indeed, the randomized placement of domain-like blocks in single cells, with preference to strong boundaries observed in ensemble Hi-C data, would be expected given that ensemble Hi-C maps have always shown clear demarcation of TAD blocks as well as low interaction frequencies across boundaries. Overall, Oligopaint imaging studies have attenuated concerns that TADs are only a statistical artifact of Hi-C data by demonstrating that chromatin domains and their boundaries are detectable and tiled across the mammalian genome in single cells. Our own current working hypothesis is that the precise domain demarcations which are strongest in ensemble maps and most frequent in single cell maps might indeed point to the true functional boundary elements. Low frequency demarcation points of blocks in single cells might indeed only represent "loops in the making" and may not be functional boundaries.

One exciting area for future inquiry is to determine if there are structural differences between unnested TADs versus nested subTADs and compartment domains versus TADs/subTADs formed by loop extrusion in single cells. Oligopaint experiments in Drosophila suggest that compartment domains can be readily detected in single cells[112, 113], whereas the individual single cell behavior of loops

and loop domains remains poorly understood at the current time. Surprisingly, chromatin domains in mammalian systems are still distinctly observable in single cells after cohesin depletion, but are distributed across the genome randomly, with a loss in preferential positioning at CTCF sites (**Fig. 2.5B**) [111]. We note that this particular study did not explicitly distinguish among TADs with corner dots, nested subTADs with corner dots, or compartments, so further classification of the precise types of chromatin domains imaged will aid in interpretation of this data. Ensemble Hi-C analyses of genome folding revealed that loop domains are destroyed and that compartment domains are strengthened and become more fine-grained upon knock-down of cohesin (**Fig. 2.5C**)[18, 19, 96]. The ensemble strengthening of compartment domains in cohesin knock-down cells forms the basis for our own working hypothesis that compartment domains would become less random and more synchronized in single cells in a cohesin knock-down imaging experiment. Data from Bintu et al. is in direct opposition to our working model because it shows that domain-like structures remaining after cohesin knock-down are truly random[111], which is not consistent with the compartment domain strengthening from ensemble Hi-C[18, 19, 96]. Thus, the mechanistic and functional nature of chromatin domains that remain in single cells after extrusion disruption remains an important open and unanswered question. It also remains to be seen if the phenomena observed across the ~2 Mb genomic region studied in this first high-resolution imaging study[111] extend genome-wide. Together, these data provide strong evidence that domain-like structures proposed in the early Hi-C studies indeed exist in single cells, and raise new exciting questions regarding whether and how compartment domains undergo random placement in the absence

of cohesin.

2.9 Evidence to date suggests compartments can both instruct and form as a consequence of transcription, potentially via membrane-less organelles

A final leading question covered by this Review is related to the eukaryotic genome's structure-function relationship – does form follow function or does function follow form? Perturbative studies have thus far led to conflicting results, and it is likely that the functional role for chromatin domains is highly specific to the genomic context, developmental timing, and eukaryotic organism in question. We also highlight that genetic dissection of the effect of key architectural features on genome function will be greatly facilitated by first delineating the compartment domains, unnested corner dot TADs, and nested corner dot subTADs. Emerging evidence thus far suggests that compartment and loop domains have strikingly different cause-and-effect relationships with transcription and other genome functions. The functional role of chromatin domains will likely be more difficult to unravel by conflating these structures given their clear mechanistic differences.

It is well established that compartment domains closely correlate with active and repressive chromatin marks, suggesting that there might be functional relationship between compartment domains and transcription[10]. For example, compartments are strongly present on the active X chromosome in mammals and only present on the inactive X at escaper genes with high transcriptional activity[114]. In Drosophila, zygotic genome activation occurs in early development via recruitment of RNA polymerase II to genes at nuclear cycle 13 (nc13) and transcriptional

elongation at nuclear cycle 14 (nc14). Structures that resemble compartment domains[102] form in parallel with transcriptional events, emerging at n13 and strengthening at nc14[115]. In early mouse development, compartments are absent or only weakly present in the zygote and form in parallel with, or subsequent to, zygotic genome activation at approximately the two cell stage[116, 117].

Despite the correlation in developmental timing between compartmentalization and gene expression activation, the possibility that their functional link is more nuanced was recently raised in mammalian systems with a genome-wide, 40 kb-resolution Hi-C study examining A/B compartment switching as embryonic stem cells differentiate along multiple lineages[44]. The authors observed that a large proportion of genes were not upregulated or downregulated during the A-to-B or B-to-A compartment shift, respectively, during differentiation. Similarly, despite slight genome-wide shifts in expression levels, a very large proportion of genes in A-to-B and B-to-A compartment shifts during reprogramming did not commensurately change their expression level[118]. Moreover, in an independent study, only ~10% of the genes upregulated during the reprogramming of B cells to iPS cells undergo a B-to-A compartment switch[119]; ~20% and ~70% remain in stable B and A compartments, respectively, indicating that in the majority of cases transcriptional changes are not accompanied by compartment structure changes. Finally, large numbers of genes fall into the categories of compartment changes that precede, delay or occur in parallel with expression changes during T cell lineage commitment[120]. Thus, a major insight from these studies in aggregate is that compartmentalization across multiple cell fate transitions cannot deterministically regulate gene expression, despite strong correlation with active

genes and chromatin marks (A compartments) and repressive chromatin marks (B compartments).

Understanding how compartment domains form will ultimately enable researchers to conduct gain-of-structure and loss-of-structure studies, and thus further evaluate their cause-and-effect relationship with transcription and other genome functions such as replication[40, 121]. For example, elegant studies have been performed to assess the functional role for global spatial positioning with respect to the nuclear periphery and internal nuclear bodies on gene expression levels[122-124]. Although the mechanisms of compartment domain assembly are not yet definitively known, an emerging idea is that phase separation[125, 126] of multivalent transcription factors into nuclear bodies could create membrane-less organelles with high local concentrations of activating or repressive biomolecules. Phase separated nuclear bodies might be responsible for segregating genomic segments with similar chromatin features into A or B compartments[127], and this spatial proximity within the global nucleus might prevent extensive inappropriate inter-compartment contacts. For example, punctate bodies of RNA polymerase, known as transcription factories, might be membrane-less organelles that contribute to at least a subset of A compartments observed in Hi-C maps[128]. In Drosophila, transcription disruption via chemical inhibition of RNA polymerase II initiation or heat shock resulted in mild but not full disruption of compartment domains as assessed by Hi-C[102, 115]. The degree of disassociation of RNA polymerase II from the genome correlated with the extent of compartment domain disruption[102]. Importantly, in early mouse development, compartments still qualitatively appear to form after chemical disruption of transcription initiation during the timing of zygotic genome activation,

but the degree to which RNA polymerase II genome occupancy and compartment strength was affected by the treatment remains unclear from these studies[116, 117].

We also emphasize that we cannot yet rule out the possibility that compartment structure, and the consequent nuclear clustering of active/inactive genomic segments, might passively influence or actively instruct transcription. Compartmentalization might aide in transcriptional regulation by facilitating an increased local concentration of biomolecules needed for gene activation/repression. Indeed, seminal studies indicate that Drosophila insulator proteins such as CTCF can form punctate nuclear bodies that visually resemble membrane-less organelles[129]. Recently, two specific domain-like structures in Drosophila which resemble fine-scale compartment domains via single cell imaging were topologically disrupted upon deletion of a 4 kb genomic segment containing CTCF[113]. Therefore, we speculate that in forthcoming studies a subset of the many insulator proteins in Drosophila might be revealed to function in collaboration with transcription to aid in the establishment or maintenance of phase-separated compartments.

Overall, the early evidence toward the question of A/B compartment's structure-function relationship are thus far consistent with the possibilities that (1) compartments might occur as a consequence of transcription rather than the cause, (2) transcription and compartmentalization might be uncoupled in many genomic locations, or (3) compartments and/or nuclear periphery localization might instructively contribute to gene expression levels in some cases. Gain- and loss-of-structure studies via compartment engineering will be critical to further dissect whether and how compartment domains might form via transcription and phase separation of nuclear factors into membrane-less organelles. In mammals, B compartments strongly

correlate with lamina associated domains (LADs)[130], therefore new technologies such as CRISPR-GO which allow tethering of specific genomic segments to the nuclear periphery and subnuclear bodies will be highly useful in determining if compartments functionally contribute to transcription[131]. Elegant recent work revealed strong evidence that at least a proportion of mammalian genes undergo activation when moved away from LADs at the nuclear periphery[132]. Moreover, further insight into how A/B compartment domains relate to other genome folding features, such as LADs, nuclear bodies, TADs, subTADs, and loops (currently out of the scope of the current manuscript, but reviewed elsewhere[133]) will continue to facilitate mechanistic and structural stratification that will enable precise dissection of the genome's structure-function relationship.

2.10 Initial causal evidence for CTCF as an enhancer-constraining insulator when forming the boundaries of contact domains

As new features of chromatin architecture emerge, the role for CTCF at multiple layers in the folding hierarchy also complicate the simple model for CTCF as a "looping facilitator". It has been well-documented that CTCF binding sites are highly enriched at the boundaries of TADs and sub-TADs[7, 9, 41, 55, 64, 72], suggesting that the protein could serve traditional enhancer-blocking insulation roles to prevent looping across specific genomic locations. For the purposes of this Review, we define enhancer-blocking (EB) insulators as sequences that block communication between adjacent regulatory elements in a position-dependent manner in ectopic transgene systems. Although extensive insight into potential insulation mechanisms have been gained through the use of ectopic transgene systems, our

knowledge of the function of EB insulation in endogenous mammalian systems remains sparse.

Genome editing approaches have provided new global and functional evidence for EB insulation in the endogenous mammalian genome. Young and colleagues recently used ChIA-PET to identify cohesin-mediated interactions across the genome in pluripotent cells[12]. First, consistent with results from previous studies, Dowen et al. found that the vast majority of cohesin-mediated interactions connect enhancers, promoters and CTCF binding sites. Second, Dowen et al. discovered that cohesin-mediated interactions between super enhancers and developmentally regulated genes are often nested within much larger structures (so-called super-enhancer domains (SDs)) created by CTCF/cohesin-mediated looping interactions. These findings are consistent with the previously reported model of nested, hierarchical looping of smaller, developmentally regulated Mediator/cohesin sites within larger CTCF/cohesin-mediated structures[9]. Third, Dowen et al. reported that the majority of interactions within SDs do not typically cross over the larger CTCF/cohesin loops and that chromatin modifications characteristic of super enhancers do not cross over SD boundaries. Together, these results suggest the CTCF/cohesin-mediated looping interactions form the structural basis for sub-TADs and might serve as classically defined EB insulators at the sub-Mb level of the genome folding hierarchy.

To directly test the EB insulation activity of CTCF/cohesin loops around SDs, Dowen et al. used CRISPR/Cas9 genome editing to delete CTCF binding sites at the boundaries of five SDs[12]. Intriguingly, the change in expression of the gene closest to the deleted CTCF site (but outside of the SD) followed one of two patterns, depending on whether a CTCF site remained between the enhancer and gene. In the first case, a SD

boundary with a single CTCF binding site, expression of the adjacent gene was increased after deletion of the CTCF site (**Figure 2.6A**). Importantly, the consequent increase in gene expression appeared to occur only if no additional CTCF sites were located between the boundary and the adjacent gene. In the second case, in which a SD boundary has a single CTCF binding site, but several additional CTCF sites also exist between the boundary and the adjacent gene, the expression of the first adjacent gene outside of the boundary does not change upon CTCF deletion (**Figure 2.6B**). Finally, for the case in which a SD boundary has dual CTCF binding sites, the first adjacent gene outside of the boundary exhibits a marked increase in expression upon deletion of both CTCF sites (**Figure 2.6C**). These results would predict that CTCF deletion would release the super enhancer to aberrantly loop to and activate off-target genes outside of the domain. The current data

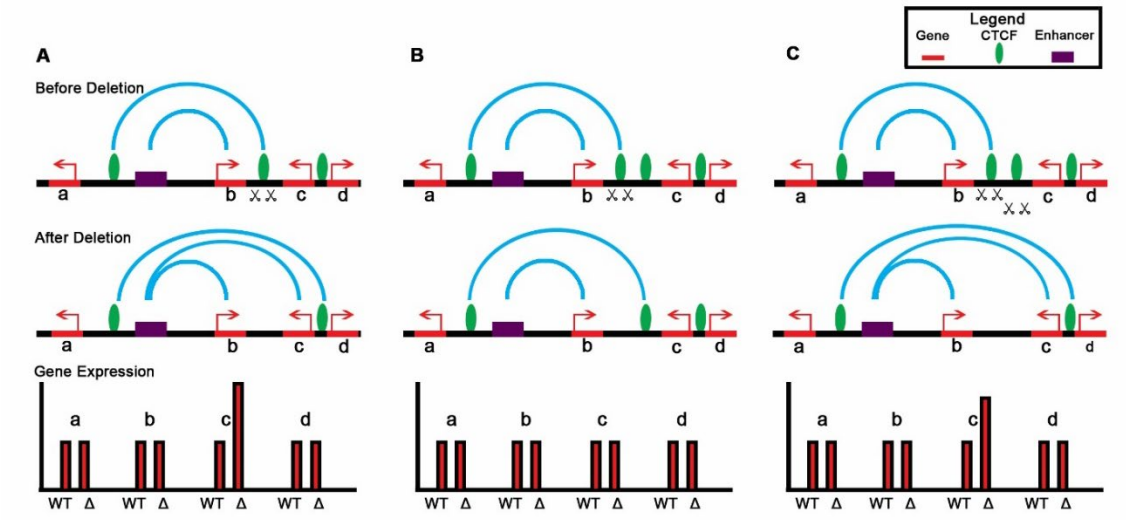


Figure 2.6. Model of sub-TAD gene regulation. (A) CTCF binding site deletion leads to inappropriate enhancer-to-gene interactions, resulting in gene upregulation. **(B)** When two CTCF binding sites appear between the queried enhancer and nearest gene, deletion of a single CTCF site does not affect gene expression. **(C)** When both CTCF binding sites are deleted, the off-target gene is upregulated. Adapted from Dowen et al. 2014.

further supports the idea that additional CTCF sites in between the SD boundary and the adjacent gene could further protect the super enhancer from aberrant long-range gene activation. Together, these results support a model in which the subset of CTCF-mediated looping interactions that create the structural foundation of sub-TADs around super enhancers can function as bona fide endogenous EB insulators.

2.11 Loop domains exhibit a markedly different cause-and-effect relationship with genome function compared to compartment domains

Evidence to date indicates that TADs/subTADs exhibit a distinct functional connection to gene regulation compared to A/B compartments. We discuss data supporting three emerging mechanisms by which loop domains might influence transcription: (1) direct, strong contact of enhancers and promoters via persistent loops (i.e. the enhancer and promoter are at the anchors of the corner dot domains and co-localize with extrusion boundaries) (**Fig. 2.7A**), (2) weak contact of enhancers and promoters via transient extrusion of SMC complexes across the loop domain (i.e. the enhancer and promoter are within a loop domain but not co-localized by a boundary so extrusion factors pass over the elements transiently) (**Fig. 2.7B**), and (3) developmental miswiring of enhancers to non-target promoters outside of the TAD/subTAD after genetic destruction of loop domain boundaries (**Fig. 2.7C**). The majority of the seminal works thus far have used the historically identified TAD definition of a Mb-scale chromatin domain[7], so it is thus far unclear which class of chromatin domain was genetically dissected in each study. For the sake of this Review, we make the assumption that the large Mb-scale domains identified

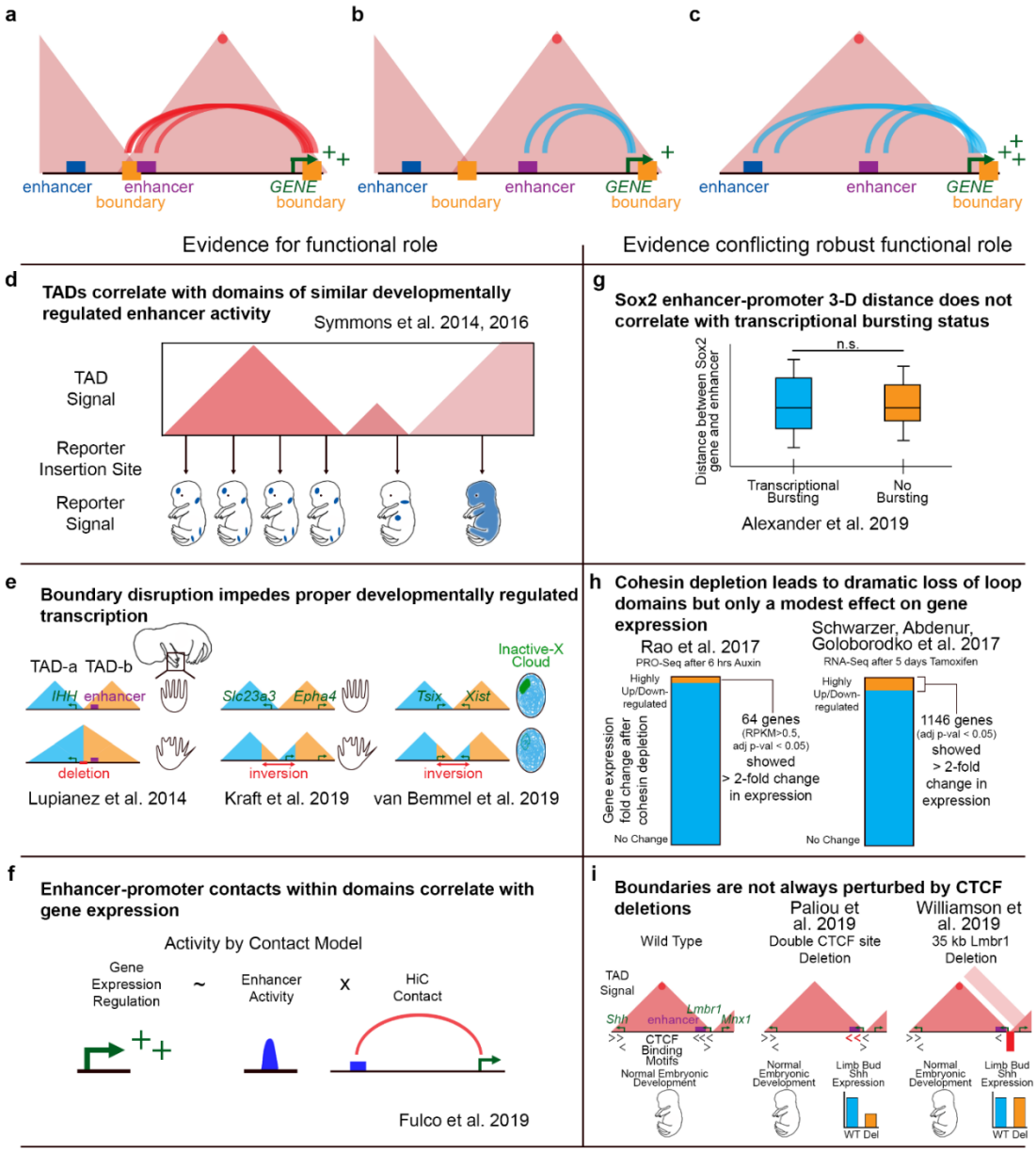


Figure 2.7. Evidence for and against TADs as a critical functional intermediary in the regulation of genes by developmentally active enhancers. **a-c:** Schematics of three emerging mechanisms by which loop domains can influence transcription**. (A)** direct, strong contact of enhancers and promoters via persistent loops (red arcs) at the corners of domains, **(B)** transient, weak contact of enhancers and promoters via transient loop extrusion (blue arcs) across the loop domain, **(C)** developmental miswiring of enhancers to non-target promoters outside of the TAD/subTAD after genetic destruction of loop domain boundaries. **(D)** Representation of the activity readout of a reporter assay upon random integration in genomic loci, from Symmons et al. 2014, 2016. **(E)** Three published

examples of boundary disruption/inversion leading to developmental issues. **(F)** Depiction of a model of long-range transcriptional regulation in which an enhancers regulatory contribution trends with its activity signature and HiC contact frequency with the target gene (Fulco et al. 2019). **(G)** Schematized boxplot of measured enhancer to *Sox2* promoter distances in actively expressing (left) and inactive (right) cells (Alexander et al. 2019). **(H)** Representation of the relatively modest transcriptional changes observed upon cohesin/Nipbl depletion observed in Rao et al. 2017 and Schwarzer et al. 2017. **(I)** Cartoon of unencumbered development that was observed upon perturbation of a TAD boundary opposing the *Shh* gene (Williamson et al. 2019).

in Dixon et al. in mammalian cells represent loop domains[7], but future studies will further test and build upon our assumptions. For the interpretation of future studies, it will be of high importance to delineate corner loop TADs versus nested corner loop subTADs versus compartment domains prior to the genetic dissection of the functional role of these topological features.

First, several elegant genetic perturbation studies over the last ten years have together allowed a model to gain traction in which TADs create insulated neighborhoods that demarcate the enhancer search space for target genes (**Fig. 2.7C**). Importantly, random insertion of an ectopic transgene sensor across the mouse genome showed enhancer activation patterns during embryonic development that correlate with some large Mb-scale TADs[134] (**Fig. 2.7D**). Across numerous studies, it has been demonstrated that genetic disruption of specific TAD boundaries (via experimental intervention or disease) causes ectopic inter-domain contacts between enhancers and non-target promoters and consequent aberrant gene expression[12, 14, 15, 20, 22, 135-138] (**Fig. 2.7E**). Most notably, the studies which focused on model systems connected to key developmentally regulated biological phenomena (e.g. X chromosome inactivation, mammalian limb development, motor

neuron differentiation) have shown a convincing link between TAD boundary disruption, ectopic enhancer-promoter interactions, and alteration of gene expression levels[14, 15, 20, 22, 135, 136, 138]. Moreover, boundary disruptions have also been reported as strongly correlated to pathologically altered gene expression in human cancers[15, 20], neurological disorders[139], rare congenital disorders[137], and diseases of limb development[21, 22]. In these early reports, miswiring of enhancer-promoter interactions across the disrupted boundary has been proposed as the mechanism for pathologically altered gene expression. Thus, evidence continues to grow to support the model that boundaries created by TADs function generally to ensure proper spatio-temporal regulation of gene expression by topologically confining enhancers to their target promoters in the appropriate developmental time window[140].

In addition to the architectural role of corner loop TADs/subTADs in preventing developmental miswiring of enhancer-promoter interactions, corner loop domains also can directly connect enhancers to promoters via CTCF-dependent and -independent mechanisms[9, 39, 86] (**Fig. 2.7A-B**). Spatial proximity can be achieved during the extrusion process (1) when both the enhancer and promoter are placed within the same loop domain and transiently come into contact due to the movement of the extrusion factor (so-called transient loops) (**Fig. 2.7B**) or (2) when the enhancer and promoter anchor the boundaries of a corner loop domain where extrusion factors stall against boundaries and form so-called persistent loops (**Fig. 2.7A**). The direct role for enhancer-promoter contacts in gene expression is only at the early stages of the perturbative studies essential to dissect the cause and effect roles of transient versus persistent loops. A recent high-throughput CRISPRi screen recruited dCas9-KRAB and guide RNAs to thousands of putative

non-coding regulatory elements[141]. The authors found that the multiplicative contribution of interaction frequency and enhancer activity together serve as the best predictor of gene expression levels ('ABC model', **Fig. 2.7F**). Noteworthy, the genomic distance-dependent background interaction frequency (i.e. the diagonal on Hi-C heatmaps) was as predictive of gene expression levels in the 'ABC model' as the observed interaction frequency at every bin-bin pair. Persistent corner loops did not provide any clear additional predictive power, at least for the regions queried by this paper in this specific screen[141]. Imaging studies have also provided evidence that enhancers are spatially proximal to their target promoters in single cells with high expression of the gene[113, 142]. However it is not yet known if the contacts imaged in these studies are persistent or transient loops. Moreover, in some cases enhancers might activate their distal targets without proximity[143] (**Fig. 2.7G**), but the genome-wide extent of this finding has yet to be shown. Finally, forced looping experiments result in upregulation in gene expression upon gain of an engineered long-range connection, but the effects of enhancer proximity on gene expression can sometimes be modest[144, 145]. Together, these early data highlight that enhancer-promoter spatial co-localization can contribute to gene expression levels, however there is a great need systematically dissect the functional role for transient and persistent loops across genomic contexts in governing transcription.

In addition to locus-specific studies, investigators have also assessed gene expression changes globally upon ablation of corner loop TADs/subTADs. Specifically, after depletion of CTCF with an auxin-mediated degron, thousands of loop domains across the genome were disrupted while compartments were unaffected[13]. Moreover, acute

degradation of subunits of the cohesin complex destroyed the majority of loop domains and led to stronger partitioning of the genome into compartment domains[18, 19]. Despite the severe global ablation of corner loop domains, these studies have surprisingly modest effects on transcription on short time scales. CTCF depletion for 24 hours resulted in only 370 differentially expressed genes in mouse embryonic stem cells. After only 6 hours of complete cohesin degradation, only 146 genes showed a 1.75 fold change in expression, and only 2 showed a 5-fold change in expression (**Fig. 2.7H**). The lack of notable gene expression changes despite widespread loop domain dissolution was even more notable because the authors used Pro-seq for nascent transcript detection[18]. Cohesin depletion over a longer 5 day time frame resulted in more than 1000 dysregulated genes, but this higher number is likely due to secondary effects that occur with long-term perturbation studies[19].

It remains of high interest to determine if all enhancer-promoter interactions were abolished with cohesin knockdown (for example, those in non-compartment+non-corner dot domains or compartment domains) or if only those connected via strong corner dot TADs/subTADs were abolished. Moreover, for each gene the functional effect of loop domain disruption may only be made manifest in the specific developmental lineage where nearby enhancers are active and the topological features are relevant; in each of these studies only a single cell type and developmental stage was queried. Indeed, after cohesin removal from mature macrophages, gene expression was preferentially altered upon inflammatory signaling induction, suggesting the effects of cohesin removal may be especially evident upon induction of a new gene expression program[146]. Finally, we note a very recent study demonstrated that Pol II elongation can reduce cohesin binding and disrupt CTCF/cohesin-mediated loops, indicating that transcription can also affect

TADs/subTADs[147]. Recent data also demonstrates that chemical inhibition of transcriptional elongation can compromise TAD boundary strength[148]. Overall, in the case of loop domains, the limited data we have thus far indicates that loops can influence function, albeit to a modest degree in some cases, and genome function in the form of transcription can also influence looping structure.

Beyond our three general models for the functional role of loop domains, the challenging work to assess the link between each individual boundary and developmentally regulated transcription is now in its early stages. Data has recently accumulated providing a nuanced view for the role of specific boundaries in specific genomic contexts in regulating locus-specific gene expression. For example, several studies have genetically dissected topological features at the Sonic Hedgehog (Shh) locus in mouse limb development, which is particularly important for the topic of this Review because a clear corner loop domain connects the *Shh* gene to its target ZRS enhancer. In one study, specific deletions of a specific CTCF site or a 35 kb region encompassing the boundary next to ZRS resulted in minimal disruption of *Shh* expression and no clear developmental defects (**Fig. 2.7I**)[149]. Importantly, structural maps show that the contact domain, including the corner dot connecting ZRS to *Shh*, remains intact with these two deletions, with a minor degree of inter-TAD interactions between ZRS and the adjacent domain (**Fig. 2.7I**). Thus, further genetic perturbations which fully abolish the corner loop connecting ZRS-*Shh* are of high interest toward understanding the role for boundary ablation in *Shh* expression. In an independent study, two CTCF sites at the ZRS boundary were both deleted, including an additional CTCF site not included in the 35 kb deletion from the other study[150]. Deletion of both CTCF sites led to disruption of the corner loop domain and a 50% reduction in *Shh*

levels. Thus, these results reinforce that boundaries consist of multiple protein binding sites, and that ablation of TAD structure often requires multiple deletions to overcome redundancies that preserve important chromatin topological features[151].

Our working model is that chromatin interactions between a gene's enhancer and promoter must be severely abolished (such as by switching the enhancer into a completely different domain) before an effect on gene expression becomes evident at the precisely important developmental timing. This model was built in part by a recent systematic dissection of genome structure-function at the *Sox9-Kcnj2* locus in mouse[138]. The authors show that the boundary demarcating the TADs around *Sox9* and *Kcnj2* is only ablated upon homozygous disruption of all occupied CTCF sites at the boundary and within adjacent domains, highlighting the remarkable redundancy of architectural protein binding sites governing TAD structural integrity. Importantly, despite complete fusion of both TADs, only minor alterations on *Sox9* and *Kcnj2* expression were observed, and there were no apparent phenotypic consequences. *Sox9* and *Kcnj2* could still contact their target enhancers, presumably because cohesin-based loop extrusion still occurs, thus suggesting that developmentally important enhancers-promoter contacts can occur even when their search space is not delimited by TAD boundaries. Another important lesson from this study was acquired through the author's careful analysis of structure and gene expression after a series of genome inversions and insertions. By contrast to the TAD fusion results, the inversion of the boundary or the aberrant placement of the boundary led to gained/lost contacts of *Sox9* and *Kcnj2* with enhancers, thus leading to pronounced effects on gene expression and severe developmental phenotypes. Together, these results teach us that, at least at this locus, ectopic placement of boundaries can break wild type enhancer-promoter

interactions and redirect enhancers to new target genes, thus leading to severe gene expression changes that give rise to pathologic phenotypes. Simply removing a boundary element is not sufficient to modify endogenous enhancer-promoter contacts because it does not sufficiently abolish the endogenous interactions.

Given that genetic inversions at boundaries have a more pronounced effect on gene expression than genetic perturbation of boundary strength, one might hypothesize that severe chromosome rearrangements might have the strongest genome-wide effect on transcription. A recent important study created high resolution maps of genome folding in the case of a Drosophila species with highly rearranged balancer chromosomes[152]. The authors show that extensive genome-wide deletions, duplications, and inversions in Drosophila can markedly shuffle chromatin domain placement, but that this leads only to a minor alteration in gene expression. As evidence continues to accumulate regarding whether and how extrusion occurs in Drosophila and whether or not domain-like structures in Drosophila are compartment domains, it will be critical to determine if the modest effect of domain-like structures on transcription in certain fly species is due to their status as strictly compartment domains. Another critical point is that balancer chromosomes have been selected for their ability to allow animal viability, therefore, it would be interesting to determine how severe chromosomal rearrangements in cases where there are visible phenotypes would affect gene expression. Beyond these exciting questions for future work, it remains important to emphasize that a lesson from this work is that not all genes might be regulated through long-range spatial contacts.

Many of the hypotheses proposed here remain to be rigorously tested. One emerging principle is that distinguishing compartment domains from loop domains,

and careful quantitation of their nested and cell type-specific properties, will be essential to obtain clear insight into the functionality of chromatin domains and their boundaries. Forthcoming studies pairing population- and single cell-based data will account for the strengths and weaknesses of both approaches and will likely yield new insight into the genome's structure-function relationship. Although early studies in the 3-D genome folding field focused on cell lines, emerging studies across model organisms, early developmental stages, time points across the cell cycle, genetic perturbations, and in human disease models will continue to build our understanding of how transcription and other genome functions shape and are shaped by the 3-D genome.

2.12. The functionality of mammalian looping interaction classes

High-resolution 3C-based studies have confirmed and extended the long-hypothesized connection between long-range enhancer promoter interactions and transcriptional activation [9, 23, 153, 154]. Ectopic induction of an enhancer-promoter loop can induce expression from an inactive globin gene [144, 155]. Roughly 10,000 loops have been identified across the human genome with the highest-resolution genome architecture maps to date [10]. Notably, fewer than one third of the loops in a transformed cancer cell line connect enhancers and promoters [10], suggesting that loops with different functional roles and organizing principles might exist. The diversity of looping classes has further been illuminated by the publication of dynamic 3D genome folding changes across developmental conditions and reprogramming [9, 153, 156, 157]. A critically important question in genome biology is whether there are different classes of loops and if they differ in the

organizing principles governing their formation and decommissioning in development and disease. While the mechanisms of ctcf/cohesin mediated loop extrusion are being elucidated, it remains unclear the extent to which these findings hold true across all classes of loops, especially those that do not involve CTCF. Beginning to answer this fundamental question starts with first accurately identifying the different classes of loops.

Perhaps the most well understood class of loops are 'developmentally invariant' or 'constitutive' interactions that play a structural role by anchoring the base of TADs and nested 'sub-TADs' [7, 9, 10, 12]. Constitutive loops are anchored by constitutive occupancy of CTCF and cohesin and often correspond to the boundaries of TADs [10]. A leading model is that constitutive looping interactions provide a framework that constrains the search space of developmentally regulated enhancers for target genes. Two recent studies have tested the functional role for structural looping interactions forming domain boundaries by CRISPR editing the CTCF binding sites. Mutation of the CTCF consensus led to the breakdown of a subTAD boundary, leading to the escape of developmentally regulated superenhancer into an adjacent domain, thereby ectopically upregulating an off-target gene [12]. Moreover, Hnisz et al. reproduced genetic mutations/deletions observed in cancer at CTCF motifs under sub-TAD boundaries surrounding oncogenes using CRISPR-Cas9 editing. Mutating these sites had the effect of ablating CTCF occupancy and the subTAD boundary, allowing for the invasion of an enhancer into the protected domain and upregulating the oncogene [20]. Consistent with these results, several additional studies have reported ectopic enhancer activation of genes in adjacent domains upon boundary disruption in disease [22, 158]. Thus, structural loops anchored by constitutively-bound, convergent CTCF work to establish genomic contact domains, thereby constraining

developmentally-regulated enhancers to interact specifically with their target genes (**Fig. 2.8A**).

Recently, Beagan et al. found that CTCF occupancy was significantly depleted during the transition from naïve pluripotency to neural progenitor cells. Consequently, short-range, developmentally regulated loops between pluripotency genes and enhancers bound by dynamic CTCF/cohesin or Mediator/cohesin were decommissioned, while invariant loops remained intact and created a larger structural framework [153]. Additional recent studies hint that CTCF loop decommissioning may continue through terminal differentiation (at least in some lineages) by finding (i) CTCF expression gradually decreases throughout mammalian brain development [37] and (ii) loops that are lost across the differentiation of monocytic precursors to mature macrophages are enriched for CTCF [157]. Together these results suggest that CTCF binding site inactivation, or 'pruning', may be a mechanism of deactivating structural and/or regulatory loops that had the potential to be activated in other lineages but were no longer necessary, while also increasing the 'search-spaces' of lineage-relevant enhancers [153] (**Fig. 2.8B**). Importantly, investigation of 3D genome folding during cellular reprogramming suggests that loss of CTCF occupancy leading to loop deactivation may be an epigenetic decision that is difficult to reverse during cellular reprogramming or drug treatment [156]. It is well established that methylation of the CTCF consensus sequence disrupts CTCF binding, however, it is less clear whether and how CTCF reengages with the genome upon DNA demethylation. Knock out of DNA methyltransferases DNMT1 and DNMT3B in human cells reactivated only a small fraction (3,237 out of >40,000) CTCF sites that are occupied in other cell types [159]. Together, these results suggest that CTCF pruning may be a mechanism of reinforcing lineage decisions

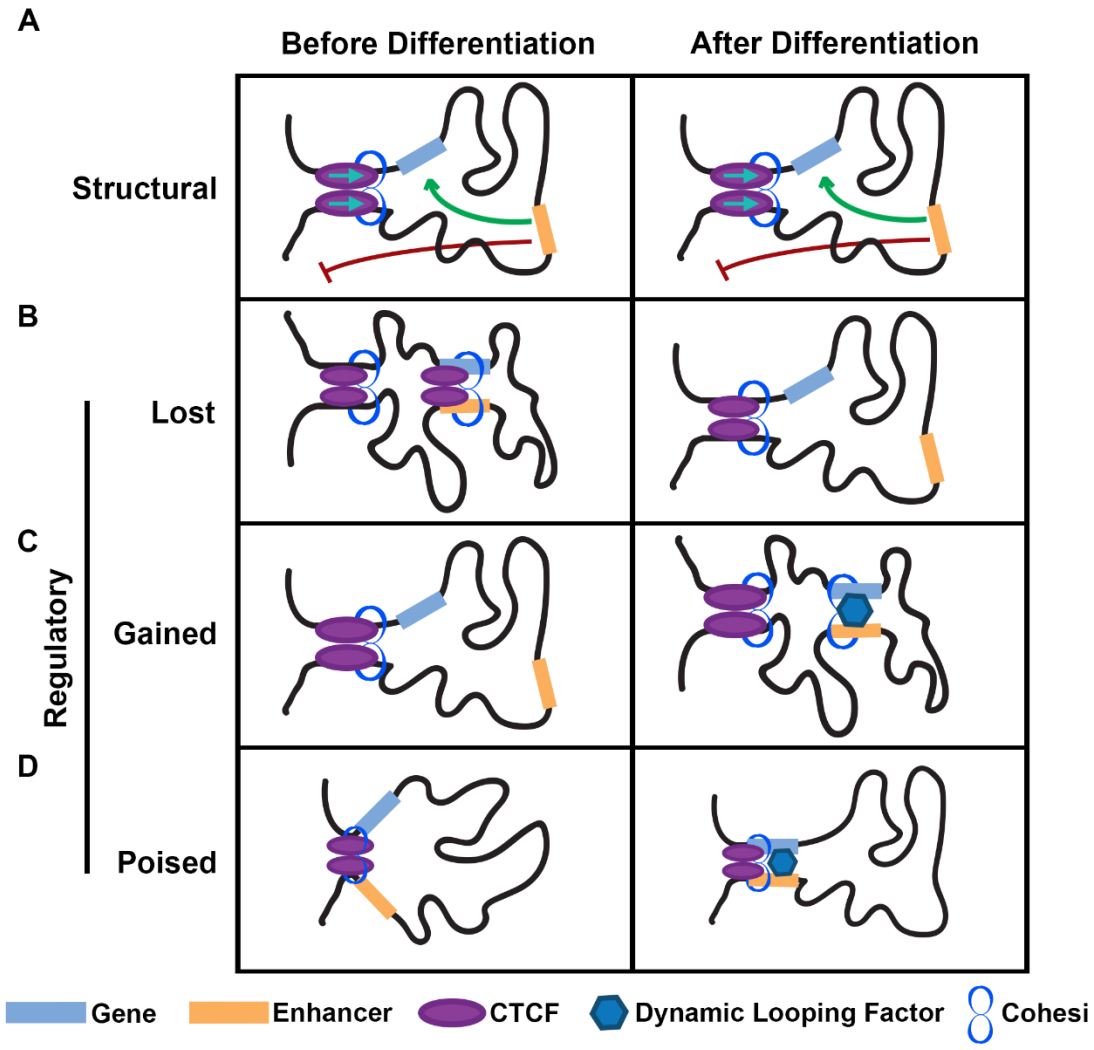


Figure 2.8. Long-range chromatin looping interactions can be divided into classes based on developmental dynamics and underlying mechanism. (A-D) Depictions of theorized looping dynamics across looping classes before (left) and after (right) differentiation, based primarily on recent reports investigating neural lineage commitment (Beagan et al. 2017) and terminal macrophage differentiation (Phanstiel et al. 2017). **(A)** Structural loops are bound by constitutive CTCF and constrain enhancers to interacting only with genes in the same insulated neighborhood. **(B)** During certain differentiation steps, specific CTCF binding sites are inactivated, thereby decommissioning the loop that was connected before differentiation. **(C)** Loops gained de novo during differentiation form within structural loops and are often anchored by proteins other than CTCF, such as YY1. **(D)** Some 'poised' loops are pre-established by CTCF early in development, before the genes/enhancers at the base of the loop are activated via the binding of additional factors.

by eliminating unneeded topological signatures from other lineages. It is important to note that Beagan et al. did not observe a decrease in CTCF expression and genome occupancy in all cellular lineages [153]. Thus, the extent to which CTCF-mediated loop decommissioning pervades mammalian lineage development remains an open question.

Another class of dynamic looping interaction are those that arise de novo upon changes in cellular state. Developmentally regulated loops often connect cell type specific enhancers and promoters [9, 10, 23, 39, 153, 156, 157]. In embryonic stem cells, dynamic loops connecting developmentally regulated pluripotency genes to their target enhancers were anchored by mediator and cohesin, but depleted for CTCF [9, 39]. More recently, a wide range of lineage specific transcription factors can be observed anchoring the base of loops, but it is unclear to what extent they are necessary and sufficient for loop formation. Indeed, architectural proteins such as CTCF should have the capability of connecting looping interactions in the absence of any clear recruitment of activating chromatin marks. Recently, Mehra et al. expressed truncated versions of YY1 without its activation domain in YY1-deleted splenic cells. Importantly, the truncated YY1 protein was sufficient to rescue chromatin loops at the Igh locus [160], suggesting transcriptional/enhancer activation is not necessary for YY1's looping function. It remains an exciting line of inquiry to dissect whether cohesin-mediated loop extrusion plays a similar role in the formation of these loops by interacting with and being stalled by YY1 at YY1-mediated looping sites. Data are consistent with a model that YY1 might be a developmentally regulated architectural protein connecting lineage-specific enhancers and their target genes (**Fig. 2.8C**).

Finally, recent developmental looping studies have not only focused on loops that are gained/lost but also regulatory loops that are constant across differentiation but connect

enhancers that are only active after differentiation. Termed 'activated' [157] or 'poised' loops, such interactions are often enriched for BOTH CTCF and dynamic looping factors such as YY1 and AP-1 [153, 157] (**Fig. 2.8D**). It has been hypothesized that CTCF at these sites may form smaller 'seed' interactions early in development, which can act as an interaction scaffold for the larger loop that is formed upon the binding of dynamic looping factors and enhancer activation [86, 153]. It remains an important question the extent to which proteins found at these poised loops, such as YY1 and Ldb1, can heterodimerize with CTCF, and what role that may play in loop formation [161, 162]. It should be noted that 'gained' and 'poised' loops may often act together to regulate target genes through the formation of 'enhancer hubs' [157]. Thus, while the roles of some architectural proteins and looping classes can be clearly parsed, the extent to which they act together to regulate the 3D epigenome is still unknown and appears to be a promising next frontier.

High-resolution chromatin architecture assays now allow us to readily identify looping interactions and classify their presence/absence across different stages of development. Understanding the organizing principles governing different looping classes will provide insight into the regulatory processes of lineage specification and how they go awry in disease. A detailed understanding of the diverse functionality of distinct looping classes and their underlying mechanisms is therefore significant toward the development of therapeutic strategies to correct malformed chromatin architectures in human disease.

CHAPTER 3: YY1 AND CTCF ORCHESTRATE A 3-D CHROMATIC LOOPING SWITCH DURING EARLY NEURAL LINEAGE COMMITMENT

3.1 Introduction

The spatial organization of the genome within the three-dimensional nucleus is dynamic during development and linked to spatiotemporal regulation of gene expression. Recent advances in proximity-ligation and deep sequencing technologies have enabled the interrogation of genome organization at a genome-wide scale and nucleosome resolution [4, 163]. Within individual chromosomes, open chromatin and active genes tend to spatially cluster into 'A' compartments, while closed, inactive chromatin spatially segregates into 'B' compartments [10, 45]. Although compartments undergo marked reorganization during cell fate transitions, the restructuring only modestly correlates with changes in gene expression, suggesting that transcription is not deterministically regulated at the compartment level [44]. Within compartments, the mammalian genome is partitioned into Megabase (Mb)-sized topologically associating domains (TADs) that are largely invariant across cell types [7, 8]. TAD structural integrity is critical for proper gene expression; perturbation of TAD boundaries leads to ectopic enhancer looping and aberrant activation of non-target genes [12, 14, 15, 20, 22]. Finally, at the sub-Mb scale within TADs, two classes of highly dynamic architectural features exist: (i) small-scale contact domains termed sub-TADs [9, 10, 12] and (ii) loops [10]. Looping interactions and subTADs often link genes to developmentally regulated enhancers and are markedly reorganized between cellular states

[9, 12, 23, 24, 53]. Thus, the emerging model is that mammalian genomes are arranged into a nested hierarchy of unique structural features, of which the finer, sub-Mb scale configurations within TADs are critical for the proper activation and inactivation of genes during development.

CCCTC-binding factor (CTCF) is a ubiquitously expressed zinc finger protein implicated in the regulation of a wide range of genome functions including transcription, insulation, splicing, replication, recombination and repair [164]. A leading hypothesis is that CTCF's diverse regulatory roles can be explained by a unifying mechanism in which it functions as an architectural protein to connect higher-order chromatin configurations [67]. CTCF is found at the base of looping interactions and knockdown of the protein abrogates chromatin connections [68, 69, 71, 165]. In a recent high-resolution, genome-wide proximity ligation study, approximately 10,000 looping interactions were reported in human cells. Importantly, of the subset of loops bound by CTCF with clear consensus sequences, 92% were anchored by consensus sequences pointed toward each other in a convergent orientation [10, 62]. CTCF-mediated interactions can be disrupted by mutation, inversion and/or deletion of the CTCF motif, indicating that consensus orientation is a critical contributing factor in loop establishment and/or maintenance [84, 85, 87]. CTCF is also enriched at the boundaries of TADs [7, 8] and deletion or inversion of these motifs can perturb domain boundaries and disrupt nearby gene expression [12, 14, 15, 22, 84, 85, 87]. Together, these data indicate the CTCF is an architectural protein that functions in an orientation-dependent manner to organize mammalian genomes across several length scales.

Genome-wide CTCF occupancy patterns have been mapped across more than 100 mammalian cell types [73, 75, 78, 159, 166]. Early studies comparing ChIP-seq signal between two

or three cell types reported that CTCF binding was largely invariant, with 65-90% of ~35,000 binding sites detected in all cellular states queried [73, 75]. More recent studies comparing CTCF occupancy across 40 cell lines showed a range of 35,000 – 75,000 binding sites per cellular state, with a total of ~110,000 possible unique genomic locations [159]. Notably, at most 20% of possible unique sites were classified as constitutive when comparing 40+ cellular states, indicating that CTCF binding is more dynamic during development than previously reported [78, 159]. Thus, it is critically important to understand the dynamic patterns of CTCF binding and whether/how they are causally linked to chromatin architecture and gene expression during cellular state transitions in development.

Recent genetic studies have confirmed that CTCF is essential for proper spatiotemporal gene expression in the developing mammalian brain. Conditional knockdown of CTCF at early, embryonic stages of mouse development triggered marked apoptosis of primary neural progenitor cells (NPCs), premature neurogenesis and disruption of tissue architecture [167]. Moreover, CTCF knockout in postmitotic cortical and hippocampal neurons [168] or the hippocampus more broadly [37] resulted in defects in gene expression, synaptic connectivity and learning and memory behavior. Finally, CTCF binding is also required for the differential expression of protocadherin (*Pcdh)* isoforms that enable branching neurites to self-recognize [169]. Together, these studies indicate that CTCF plays an essential role in early neural development and highlight the importance of unraveling the currently unknown mechanisms linking occupancy with genome architecture and expression in the brain.

Here we set out to understand the dynamic CTCF occupancy landscape and how it is linked to the restructuring of fine-scale chromatin architecture at the earliest stages of the establishment of neuronal gene expression programs. We used well-established cellular models of early neural lineage commitment: (i) mouse embryonic stem (ES) cells cultured in Gsk3/MEK inhibitors ('2i' conditions) representing a state of naïve pluripotency from the earliest stages of a pre-implantation embryo [170]; (ii) ES cells cultured in serum/LIF representing a slightly more mature state of pluripotency with increased poising of developmentally regulated genes [171]; and (iii) primary multipotent neural progenitor cells (NPCs) representing the earliest departure from pluripotency and commitment to lineages in the mammalian brain [172]. We uncover several new organizing principles governing higher-order chromatin folding during neural lineage commitment. Our observations support a model in which looping interactions connecting developmentally regulated enhancers to genes undergo an architectural protein switch from CTCF to YY1 early in neural development; YY1-anchored looping interactions arise de novo in NPCs within a larger topological framework connected by constitutively bound CTCF.

3.2 Results

3.2.1 CTCF engagement with the genome decreases during neural development

To investigate CTCF dynamics during the earliest stages of neural development, we performed ChIP-seq in NPCs derived from neonatal mouse brains as well as embryonic stem (ES) cells cultured under both 2i/LIF (2i) and serum/LIF (serum) conditions. The three cellular states were chosen to capture the initial establishment of neural gene expression programs and to benchmark the changes against a presumably less dramatic

transition between naïve and primed/mature pluripotency. Equivalent genetic backgrounds were achieved by utilizing v6.5 ES cells (C57Bl6 x 129SvJae) and NPCs from mice maintained on a mixed C57Bl6/129SvJae background [173]. We first noticed that the number of CTCF binding sites decreased in a stepwise manner during the transition from naïve pluripotency to multipotency, with the sharpest drop in binding sites between ES serum and NPC conditions (**Fig. 3.1A**). To further explore dynamic CTCF during neural development, we utilized available ENCODE CTCF ChIP-seq data sets from the mouse [174]. Consistent with trends in our cellular models, published ENCODE CTCF ChIP-seq peaks also showed a global decrease in adult brain regions (cortex, cerebellum, olfactory) compared to the E14.5 brain tissue (**Fig. 3.1B**). Notably, when we investigated other developmental lineages, we found that CTCF binding can display the opposite trend, in some cases increasing between the embryonic and adult stages (**Fig. 3.2A, B**). Thus, while CTCF occupancy appears to decrease during the transition from pluripotency to early neuronal lineage commitment, it is not a pervasive trend across all developmental lineages.

To gain insight into why NPCs have a unique pattern of decreased CTCF occupancy during early neuronal lineage commitment, we next conducted an analysis of CTCF gene expression and protein levels. We observed a general accordance between the ChIP-seq and RNA-seq results in our cellular models: CTCF gene expression decreased between the pluripotent stem cell states and multipotent NPCs (**Fig. 3.1C**) and also between the embryonic mouse brain and mature adult brain regions (**Fig. 3.1D**). Moreover, Western blot analysis of CTCF protein levels showed a similar decrease in NPCs compared to pluripotent ES cells (**Fig. 3.2C**). Corroborating our results, while this manuscript was under review an independent study also reported a decrease in CTCF protein levels in whole

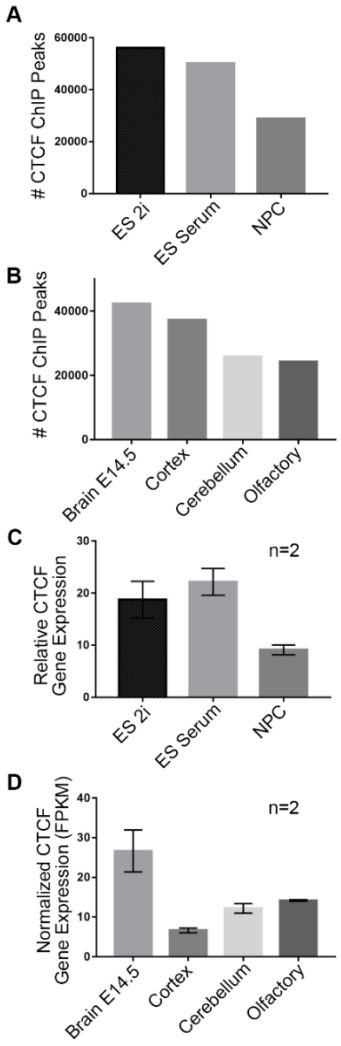


Figure 3.1. CTCF binding and expression decrease during neural development. (A) Number of CTCF ChIP-seq peaks called across the ES 2i, ES serum and NPC cellular states**. (B)** Number of CTCF ChIP-seq peaks across several mouse ENCODE brain tissues (Shen et al. 2012). **(C)** Relative CTCF gene expression across 3 developmental cell types (error bars represent 1 s.d. from mean). **(D)** Normalized CTCF gene expression (FPKM) across mouse ENCODE brain tissues (error bars represent 1 s.d. from mean) [174].

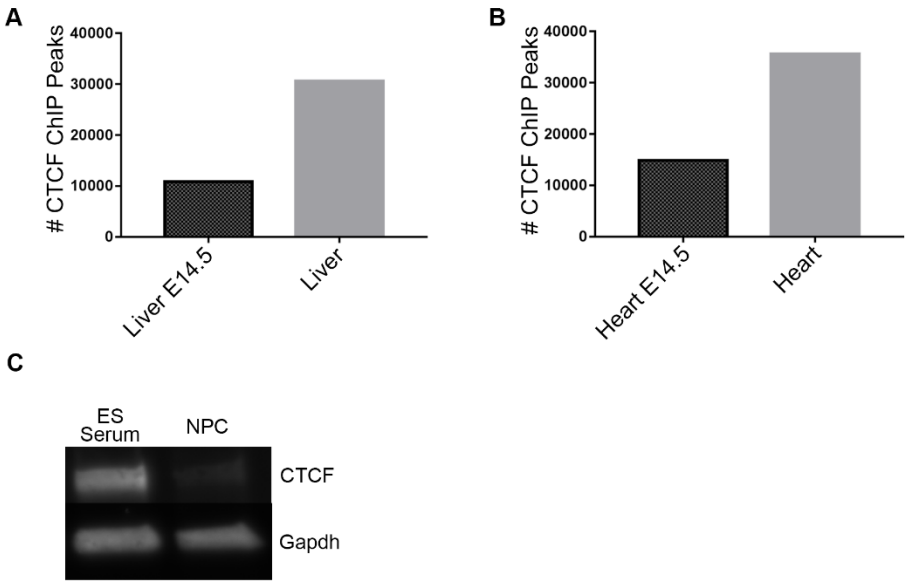


Figure 3.2. Genome-wide CTCF occupancy may not decrease during lineage commitment in some tissues. (A) Number of CTCF binding sites across mouse ENCODE embryonic and adult liver tissues. **(B)** Number of CTCF binding sites across mouse ENCODE embryonic and adult heart tissues. All ChIPseq data from (Shen et al. 2012). **(C)** Western blot analysis querying CTCF and Gapdh protein levels in ES cells in serum-LIF media and NPCs.

mouse brains during the transition from E15 to postnatal week 1 (Sams et al. 2016). Sams et al. also identified differential CTCF levels across neurons, astrocyte, and oligodendrocytes from the hippocampus, highlighting that we cannot rule out the possibility that heterogeneity in cells derived from ENCODE tissues may contribute to the aggregate decrease in CTCF levels. Our NPC cultures exhibited a highly consistent morphology throughout the population and > 90% were Sox2 positive (data not shown), suggesting that our NPC preparations were substantially less heterogenous than brain tissue lysates. Our data indicate that CTCF gene and protein expression levels decrease in the transition from pluripotency to multipotent neural progenitor cells in parallel with a global decrease in the number of genome-wide CTCF binding sites.

3.2.2 CTCF occupancy in NPCs is largely pre-established in the pluripotent state

To better understand the CTCF sites that are dynamic among our cellular states, we parsed CTCF peaks present in the ES 2i, ES serum and NPC conditions into classes based on their cell-type specific occupancy (**Fig. 3.3A, Appendix I Methods**). We identified 56,138, 50,185 and 28,860 binding sites in ES 2i, ES serum and NPCs, respectively, with a total of 60,688 unique, non-redundant sites across all three cell types. We found that approximately 44% of CTCF sites (n=26,435) displayed constant occupancy across our three cell types of interest and were thus classified as 'constitutive'. We also explored several classes of dynamically occupied CTCF sites, including: (i) the '2i only' class present in naïve pluripotency conditions and lost in the transition to a more primed/mature pluripotent cellular state, (ii) the '2i+serum' class present across pluripotency conditions and lost in NPCs and (iii) the 'NPC only' class arising only upon the departure from pluripotency. We confirmed the validity of our parsing scheme for our four CTCF classes of interest by plotting the composite ChIP-seq signal for all three cell types centered on the midpoint genomic location of a given class (**Fig. 3.3B, Fig. 3.4**). The ChIP-seq pileup plots indicate that constitutive CTCF binding sites display markedly higher occupancy signal than sites that are dynamically altered upon changes in cellular state. These results confirm and extend recent reports suggesting that there is a larger class of dynamically occupied CTCF sites than previously appreciated [159, 174].

We next sought to understand dynamic CTCF occupancy patterns in the naïve to mature pluripotency transition and the mature pluripotency to multipotency transition. At the outset of our analysis, we hypothesized that CTCF binding may decrease severely

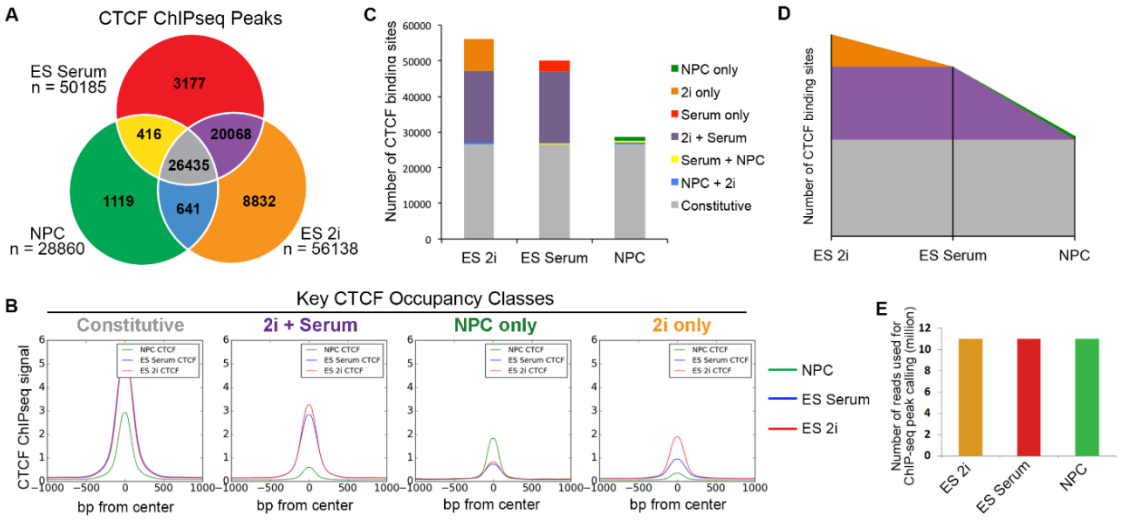


Figure 3.3. Sites bound by CTCF in NPCs are predominantly pre-existing from earlier stages of development. (A) Classification of CTCF binding sites parsed between three developmental cell states. **(B)** Composite CTCF ChIP-seq signal in NPCs (green), ES serum (blue) and ES 2i (red) centered around the peaks of Constitutive, 2i+Serum, NPC only and 2i only CTCF classes. **(C)** Stacked barplot representing the distribution of CTCF binding classes across ES cells in 2i, ES cells in serum, and NPCs. **(D)** Theorized landscape plot depiction of constitutive and dynamic CTCF during the early time points of development. Colors represent same CTCF classes as presented in (C). **(E)** Library read depth is comparable across conditions. After redundant read removal and downsampling, 11 million reads were utilized for the CTCF ChIP-seq analysis of each cell type.

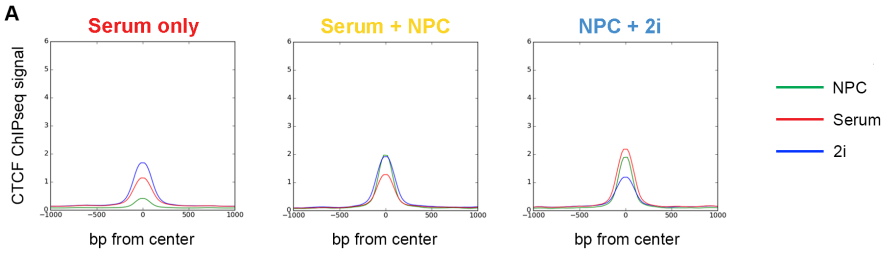


Figure 3.4. CTCF binding strength in additional CTCF occupancy classes. (A) CTCF ChIPseq signal in NPCs (green), ES serum (red) and ES 2i (blue) centered at parsed CTCF peaks (serum only, serum+NPC and 2i+NPC occupancy classes).

between 2i and serum conditions due to the known hypomethylated state of naïve pluripotent stem cells [175-177], but we observed only a relatively minor reduction in CTCF occupancy between ES 2i and ES serum ('2i only' class, n=8,832). By contrast, we noticed that the number of CTCF sites lost between ES serum and NPCs nearly matched that of the constitutive class ('2i+Serum' class, n=20,068), suggesting that the transition from pluripotency to multipotent progenitor cells represents a critical developmental window in establishing the neural CTCF landscape. Importantly, the number of 'NPC only' CTCF sites that arose during differentiation was relatively small (n=1,119), indicating that the vast majority of CTCF peaks called in NPCs were already present in the pluripotent cell types (**Fig. 3.3C**). Our results suggest that the CTCF occupancy landscape in NPCs does not result from a marked reshuffling and/or extensive de novo acquisition of new CTCF binding sites. Rather, a large proportion of CTCF sites are pre-established at least as early in development as naïve pluripotency and selectively lost in early neural lineage commitment (**Fig. 3.3D**, orange and purple classes). ChIP-seq experiments were conducted in the same batch, sequenced on the same flow cell and downsampled to the same read depth to attenuate technical artifacts that might influence our observed results (**Fig. 3.3E**).

3.2.3 The 3D genome is reconfigured during early neural development

CTCF has a well-established role in connecting long-range looping interactions [84, 85]. Given the large number of CTCF peaks that are dynamic across development, we sought to investigate how chromatin folding is altered as a function of occupancy during each cell fate transition. We generated fine-scale chromatin architecture maps (~4-12 kilobase (kb) matrix resolution) across > 7 Mb of the mouse genome surrounding key developmentally

regulated genes with Chromosome-Conformation-Capture-Carbon-Copy (5C) and high-throughput sequencing. In a previous study focused on chromatin folding during somatic cell reprogramming, we generated 5C libraries (n=2 biological replicates) in ES serum, ES 2i and NPC conditions [156]. Here, we elected to begin our 3-D analyses and mechanistic exploration with the raw reads from our published 5C libraries because they were genetically- and culture condition-matched to the pellets used to generate our RNA-seq and CTCF ChIP-seq libraries.

Building on the foundation of our previously published 5C analysis pipeline [156], we further developed and applied a new set of computational methods to better resolve punctate looping interactions present within each cell type. We normalized the intrinsic biases in 5C data, corrected for library complexity and sequencing depth differences and attenuated spatial noise via a 16 kb blocked smoothing window. The resultant 'Relative Interaction Frequency' data binned at 4 kb matrix resolution exhibited high reproducibility between biological replicates (**Fig. 3.5A**, **Fig. 3.6A**). Additionally, our 5C data showed strong biological concordance with published Hi-C data from the murine cortex [7] across a 1 Mb region surround the Sox2 gene NPC (**Fig. 3.6B**).

Looping interactions can be detected in 5C heatmaps as concentrated points of high interaction frequency compared to the surrounding local background [10]. Although one universal distance-dependence expected model could be computed on 5C data, we have found that application of a global expected often leads to over- or under-estimation of looping strength. To compute a local expected interaction frequency, we applied the 'donut' and 'lower-left' background filters (**Fig. 3.5B**, blue and green outlines, respectively) recently proposed by Aiden and colleagues [10]. Local background filters

capture the more nuanced aspects of the distance-dependence expected interaction frequency and the TAD/subTAD domain structure (**Fig. 3.5C,D**, **Figs. 3.7, 3.8**). To take a conservative approach for loop detection, we corrected our 5C counts with the maximum of the two filters (**Fig. 3.5E, Fig. 3.9**). We next modeled the 'expected-corrected interaction frequency' data as a continuous random variable with a logistic distribution (**Fig. 3.10**). The resultant p-values for each pixel were converted to an 'Interaction Score' (IS = -10*$\log_2$(p-value)), allowing for systematic comparison of looping signal within and between 5C regions and experiments. Punctate looping structures were readily apparent in the uncorrected and interaction score heatmaps (**Fig. 3.5A, 3.5F**).

As a critical first step toward understanding the relationship between CTCF occupancy and 3-D chromatin architecture changes, we computationally parsed looping interactions into sub-classes based on their interaction score in each cell type (**Fig. 3.5G**). Pixels in which both replicates of each biological condition similarly passed or failed each threshold (**Fig. 3.11A**) were classified into one of seven looping classes (**Figs. 3.5G, H**). Thresholds were chosen so that our top five largest dynamic looping classes achieved an empirical false discovery rate less than 15% (**Fig. 3.11B-D, Appendix I Methods**). Consistent with previous reports [10], we noticed that pixels of the same looping class were often adjacent to each other and therefore could be clustered together into a contiguous architectural feature. Altogether, we identified several classes of 3-D interactions (**Fig. 3.5I**) and elected to focus our analysis on three main groups: (i) 141 loops present in all 3 cell types ('Constitutive', grey class), (ii) 46 loops present in both ES 2i and ES serum but lost in NPCs ('2i+Serum', purple class) and (iii) 75 loops specific to the NPC state ('NPC only', green class). We confirmed that our looping class interaction scores trended across

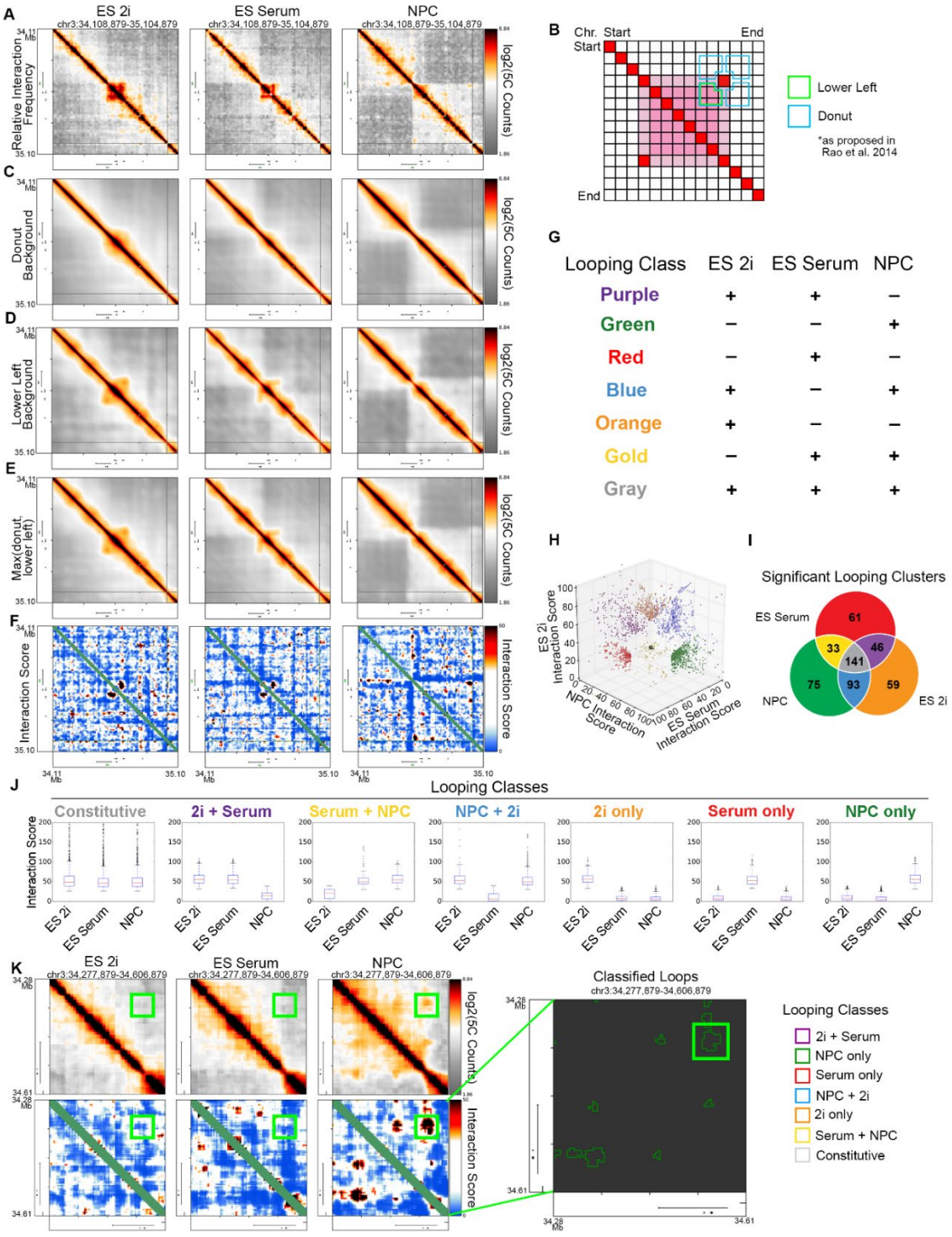


Figure 3.5. Dynamic Classes of 3D Interactions Arise during Neural Lineage Commitment. (A) Heatmaps displaying the relative chromatin contact frequency in a 1 Mb region surrounding the *Sox2* gene in ES 2i, ES serum and NPCs. Color bars range

from low (grey) to high (red/black). **(B)** Schematic depiction of donut (blue) and lower left (green) expected background models. **(C-E)** Expected background heatmaps for the region surrounding the *Sox2* gene. **(C)** Donut filter, **(D)** Lower left filter and **(E)** Maximum value of donut and lower left filters. **(F)** Interaction score heatmaps at the Sox2 locus. Color bar ranges from low (blue) to high (red/black). **(G)** Schematic of looping classes parsed by their dynamic behavior across three cellular states. **(H)** Scatter plot of 5C interaction scores for each pixel classified as part of a looping interaction across the ES 2i, ES serum and NPC states. **(I)** Number of significant looping clusters in each dynamic 3-D interaction class. **(J)** Boxplots representing interaction scores across each cell type for the pixels classified into each looping class. **(K)** Visualization of a *Sox2*-pluripotency enhancer interaction in relative interaction frequency heatmaps (top left row), interaction score heatmaps (bottom left row) and classified loop clusters (right).

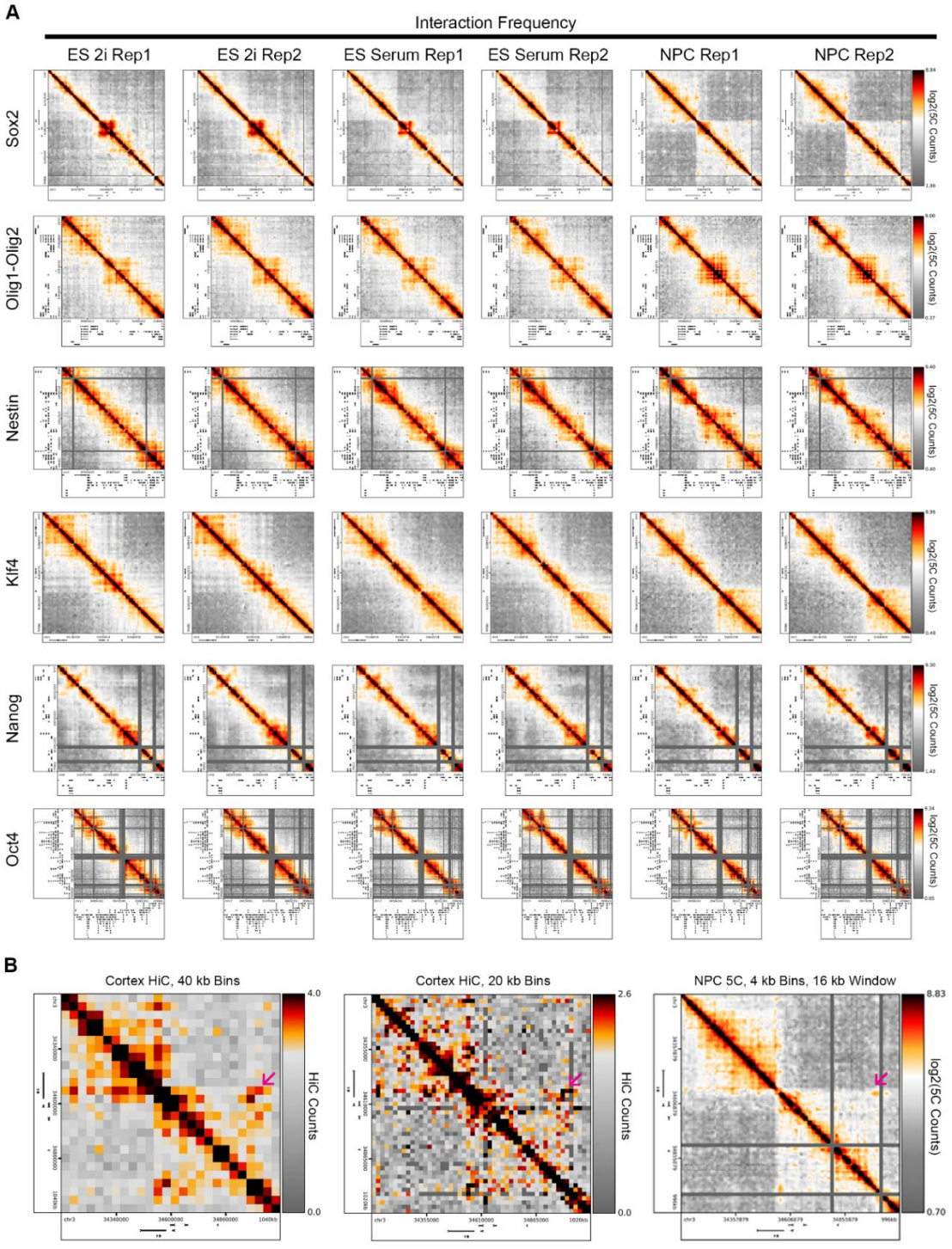


Figure 3.6. Relative interaction frequency heatmaps at key developmental loci. (*A*) Relative interaction frequency heatmaps of 1 Mb surrounding several developmental genes (rows; *Sox2*, *Olig1-Olig2*, *Nestin*, *Klf4*, *Nanog*, *Oct4*) in replicates of ES 2i, ES serum and NPCs (columns). **(*B*)** Comparison of mouse cortex HiC heatmaps at 40 kb

(left) and 20 kb bins (middle) with our 4 kb binned pNPC 5C heatmaps in a 1 Mb region surrounding the Sox2 gene. HiC data from [7].

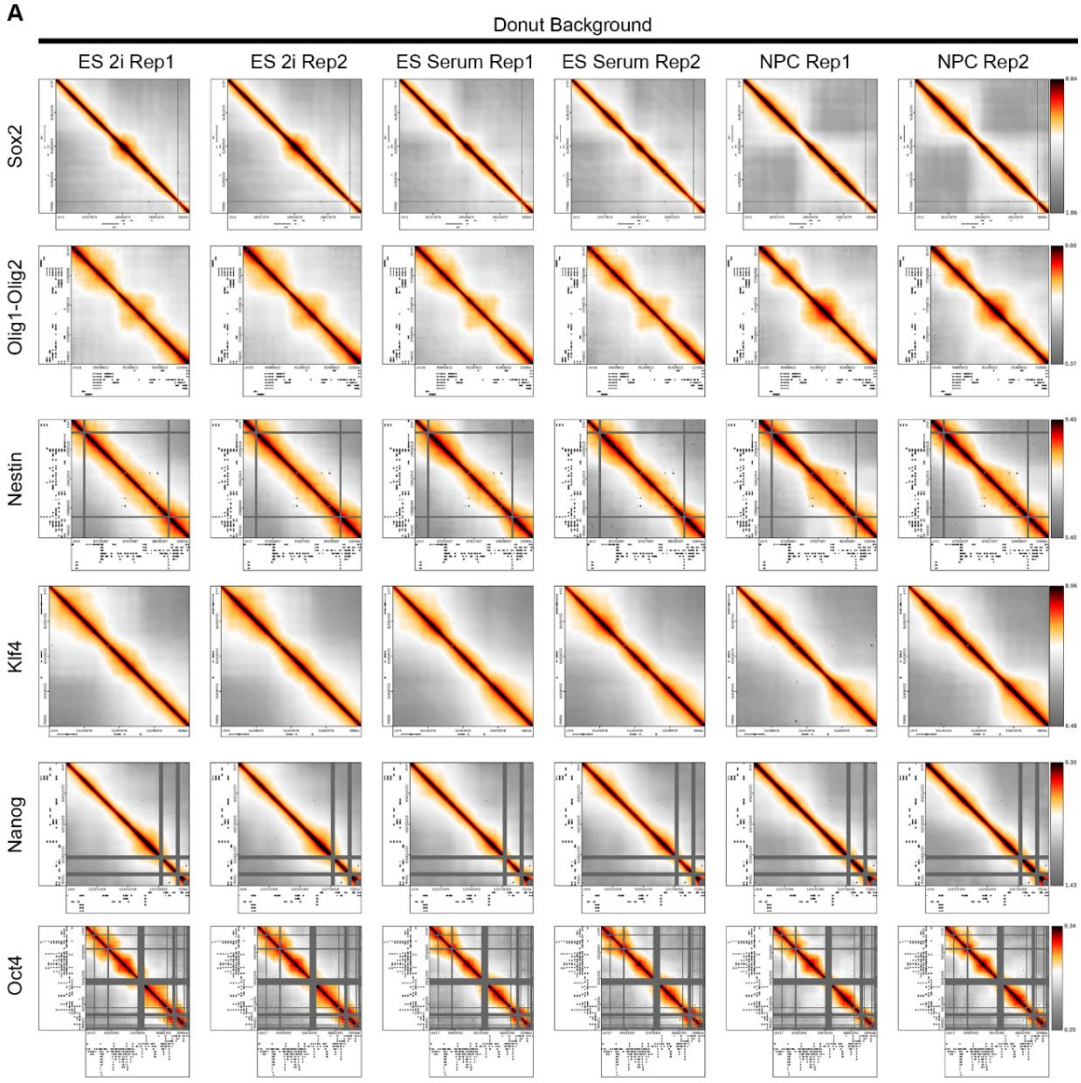


Figure 3.7. Donut expected background model heatmaps at key developmental loci. (*A*) Donut expected background model heatmaps of 1-2 Mb surrounding several developmental genes (rows; *Sox2*, *Olig1-Olig2*, *Nestin*, *Klf4*, *Nanog*, *Oct4*) in replicates of ES 2i, ES serum and NPCs (columns).

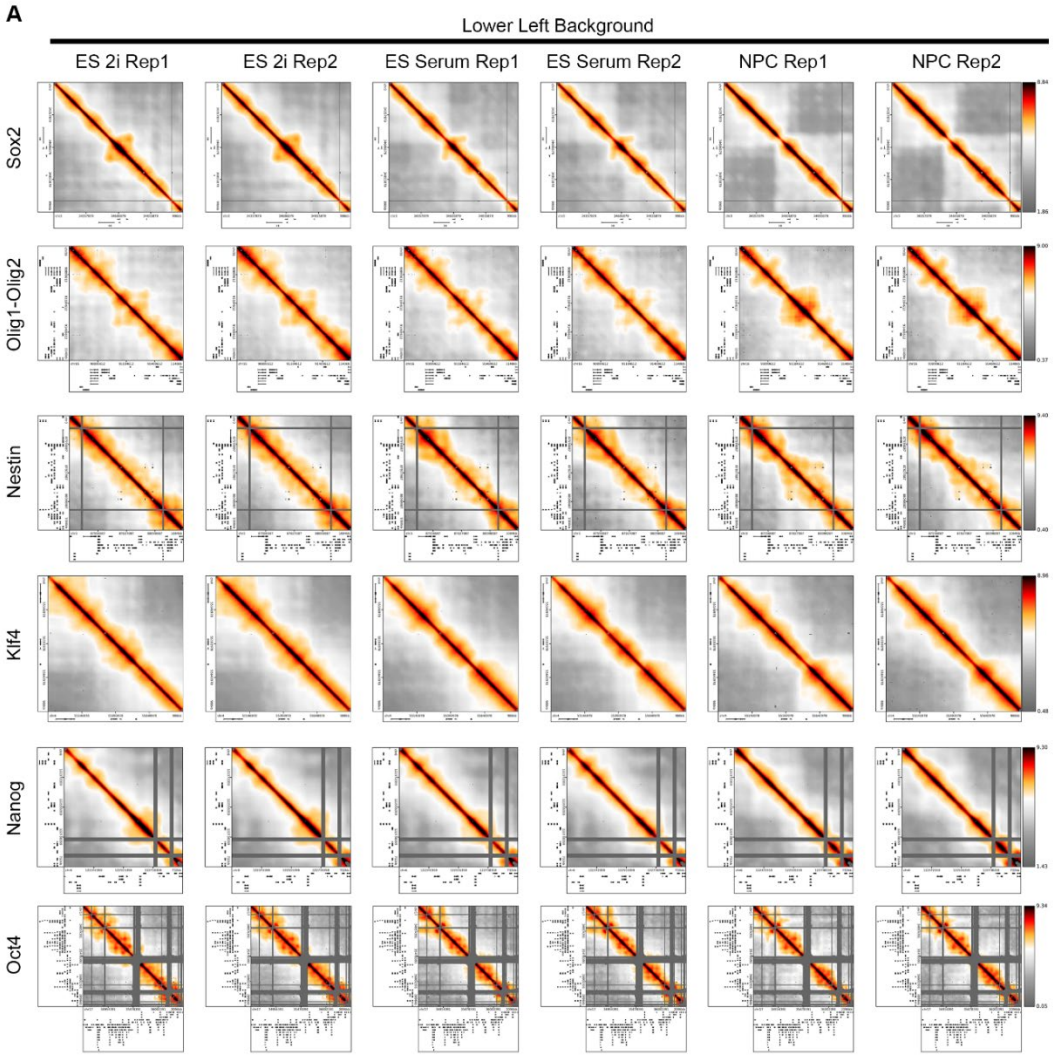


Figure 3.8. Lower left expected background model heatmaps at key developmental loci. (*A*) Lower left expected background model heatmaps of 1-2 Mb surrounding several developmental genes (rows; *Sox2*, *Olig1-Olig2*, *Nestin*, *Klf4*, *Nanog*, *Oct4*) in replicates of ES 2i, ES serum and NPCs (columns).

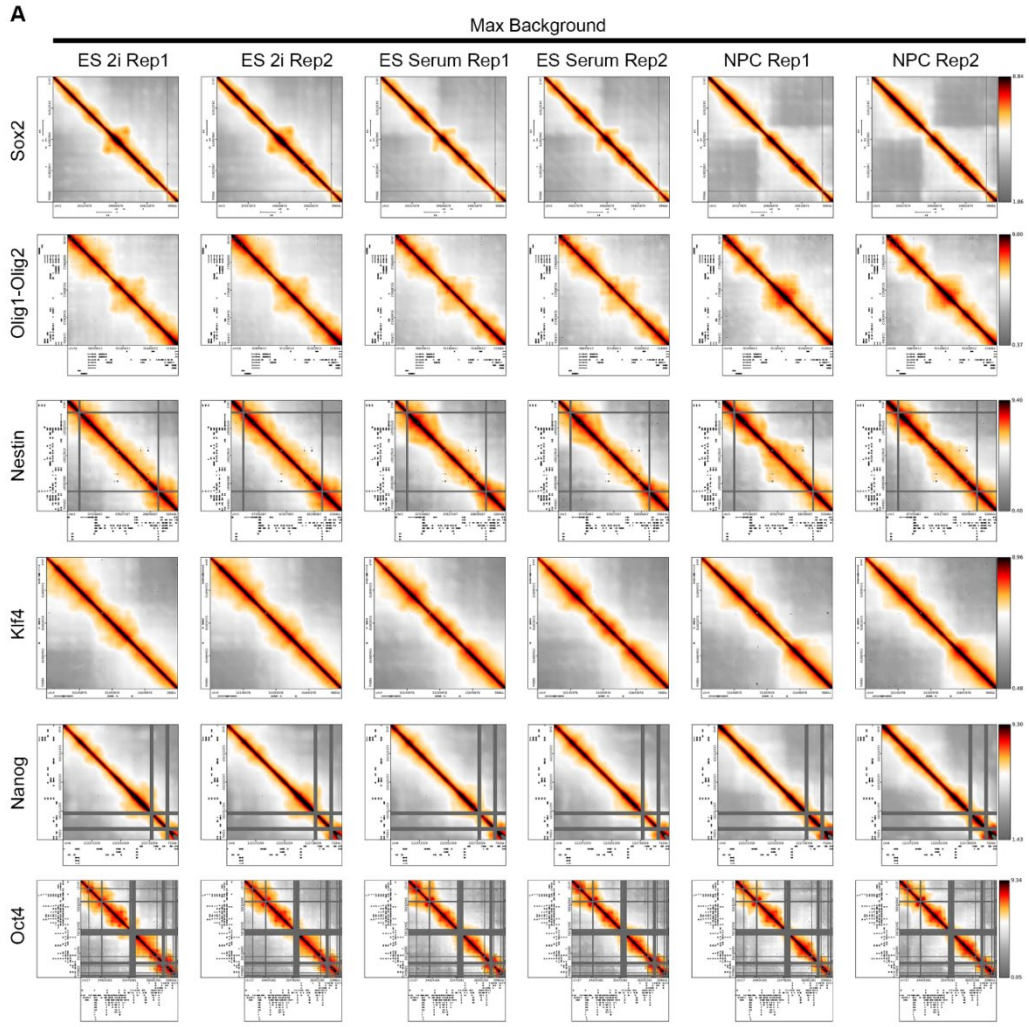


Figure 3.9. Max(donut, lower left) expected background model heatmaps at key developmental loci. (A) Max(Donut, Lower Left) expected background model heatmaps of 1-2 Mb surrounding several developmental genes (rows; *Sox2*, *Olig1-Olig2*, *Nestin*, *Klf4*, *Nanog*, *Oct4*) in replicates of ES 2i, ES serum and NPCs (columns).

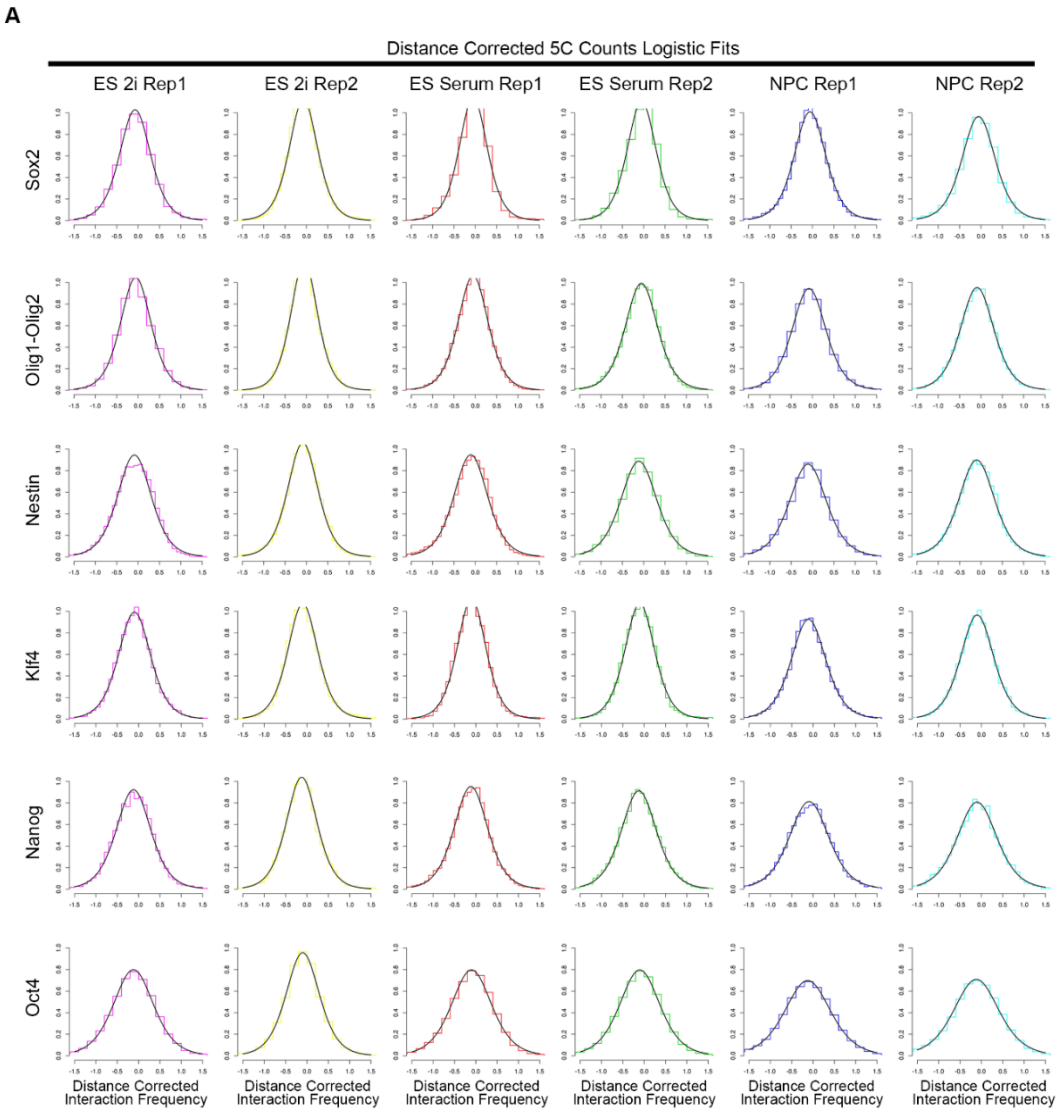
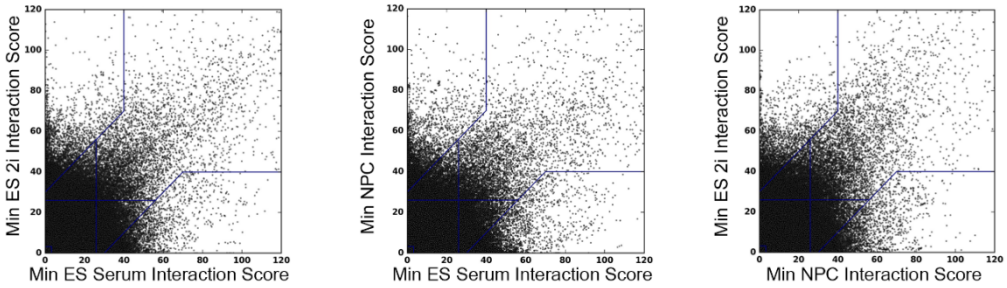


Figure 3.10. Distance-corrected 5C counts fit with the logistic distribution. (*A*) Histograms of distance corrected 5C counts overlaid by logistic distributions fit independently for each region and replicate.

A



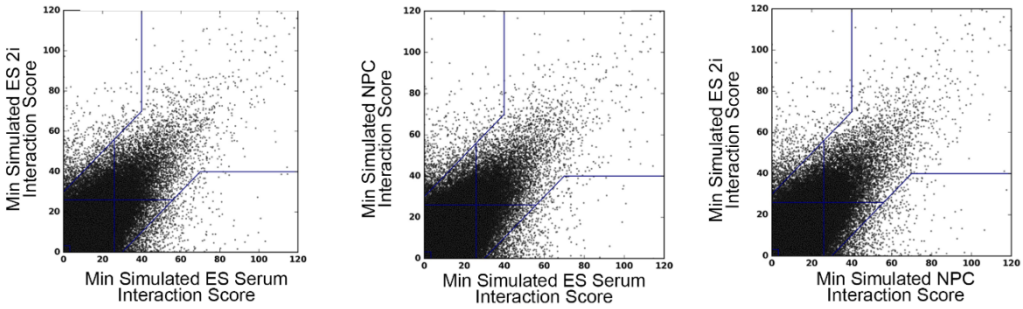
B

Cell Type	Correlation Calculation	Background Normalized Counts Correlations
ES 2i	Pearson's	0.72764936
	Spearman's	0.689918988
ES Serum	Pearson's	0.745865553
	Spearman's	0.726672268
NPC	Pearson's	0.637227701
	Spearman's	0.606833588

		Simulated NPC Reps		Simulated ES Serum Reps	
Looping Class	Real Data Loop Count	Simulated Data Loop Count*	FDR	Simulated Data Loop Count*	FDR
2i + Serum	374	44.181	11.8131	38.137	10.19706
NPC + 2i	530	44.114	8.323396	37.853	7.142075
Serum + NPC	104	43.911	42.22212	38.152	36.68462
NPC only	712	2.384	0.334831	1.6	0.224719
2i only	451	2.238	0.496231	1.583	0.350998
Serum only	347	2.271	0.654467	1.577	0.454467

*Reported as the average number of classified interactions across 1000 simulations

C



D

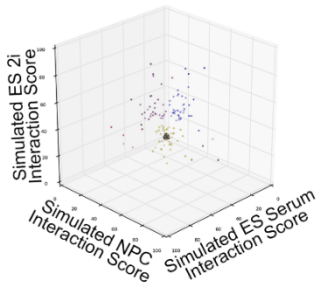


Figure 3.11. Thresholding Interaction Scores to Achieve Reasonable False Discovery Rates. (*A*) 2D Scatterplot of the minimum interaction scores across the two replicates of each cell type for all bin-bin pairs. Blue lines show applied thresholds. **(*B*)** Tables of expected-corrected interaction frequency correlations (left) and real 5C data pixel counts within looping classes compared to simulated pixel count and false discovery rate (FDR) within looping classes of simulated ES serum and NPC replicates (right). **(*C*)** 2D scatterplot of the minimum interaction scores across the two replicates of each simulated cell type for all bin-bin pairs. Blue lines denote applied thresholds. **(*D*)** 3D scatterplot of the classified interactions from the first NPC simulation.

the three cellular states in a manner that was commensurate with their intended classification (**Fig. 3.5J**). Visual inspection of the data confirmed that gold-standard looping interactions, such as the 'NPC only' interaction between the Sox2 gene and an upstream regulatory element were accurately detected, clustered and classified (**Fig. 3.5K**, colored green). Finally, we assessed the orientation of the CTCF motifs at loop anchors and confirmed that loops in all three cell types were highly enriched for convergently oriented motifs compared with divergent or tandem orientations (**Fig. 3.12A-B**). Altogether, our analysis pipeline allowed us to accurately identify and visualize looping interactions critical to each cellular state within our 5C regions.

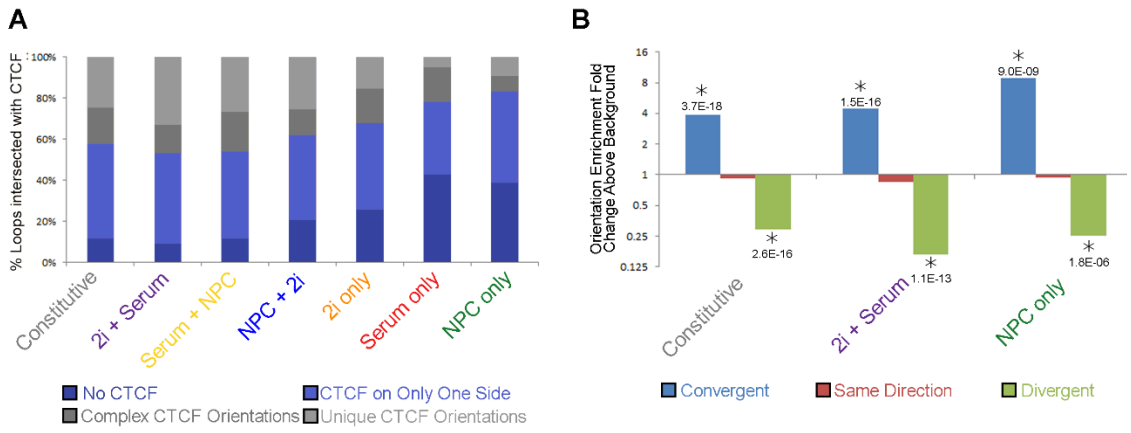


Figure 3.12. CTCF anchoring classified interactions are preferentially oriented in a 'convergent' manner. (A) Stacked barplot characterizing the presence of CTCF in each looping class. Classifications were: no CTCF in loop (dark blue), CTCF found on only 1 side of the looping interaction (light blue), complex CTCF orientations such as conflicting CTCF orientations at the same peak or on the same side of a loop (dark grey) and unique CTCF orientations anchoring both sides of each loop (light grey). **(B)** Fold change enrichment of pairs CTCF motifs in specific orientations across the two sides of interactions present in constitutive, Serum+2i, and NPC-only loops compared to background levels. P-values calculated using Fisher's exact test.

3.2.4 CTCF binding correlates with loss of 3D interactions during the departure from pluripotency

We next investigated the relationship between the significant loss of CTCF binding between the pluripotent and multipotent states and coincident architectural rearrangements. *Sox2* forms a pluripotency-specific loop with a putative ES-specific enhancer ~120 kb downstream that is essential for proper expression of the gene in ES cells [178] (**Fig. 3.13A-B**, magenta arrowhead). We identified several '2i+Serum' CTCF sites at the putative ES-specific enhancer (**Fig. 3.13B**, green boxes on x-axis). During the departure from pluripotency, CTCF binding is lost and the looping interaction connecting the *Sox2* gene to the putative ES-specific enhancer concurrently breaks apart (**Fig. 3.13B-C, Fig. 3.14A-B,** red arrow). *Nanog* displays a similar behavior: in ES 2i and ES serum, the gene interacts with a putative ES-specific enhancer element ~80 kb downstream that is essential for proper expression of the gene [179]. Several '2i+Serum' CTCF sites anchor the '2i+Serum' looping interaction connecting *Nanog* and its putative enhancer (data not shown). In concordance with these locus-specific examples, '2i+Serum' looping interactions across our 5C regions were enriched with '2i+Serum' CTCF sites (**Fig. 3.13D**). Together, these data suggest that the loss of CTCF occupancy at key looping interactions during the departure from pluripotency is accompanied by a decrease in looping strength.

We questioned some conflicting observations: although the loss of CTCF often coincides with the loss of a looping interaction (**Fig. 3.13D**) and NPCs have substantially fewer CTCF peaks than the pluripotent states (**Fig. 3.3C**), NPCs have roughly the same number of looping interactions as ES serum/ES 2i in the genomic regions covered by our 5C primers (**Fig. 3.5J**). Notably, when we explored the percentage of key looping classes

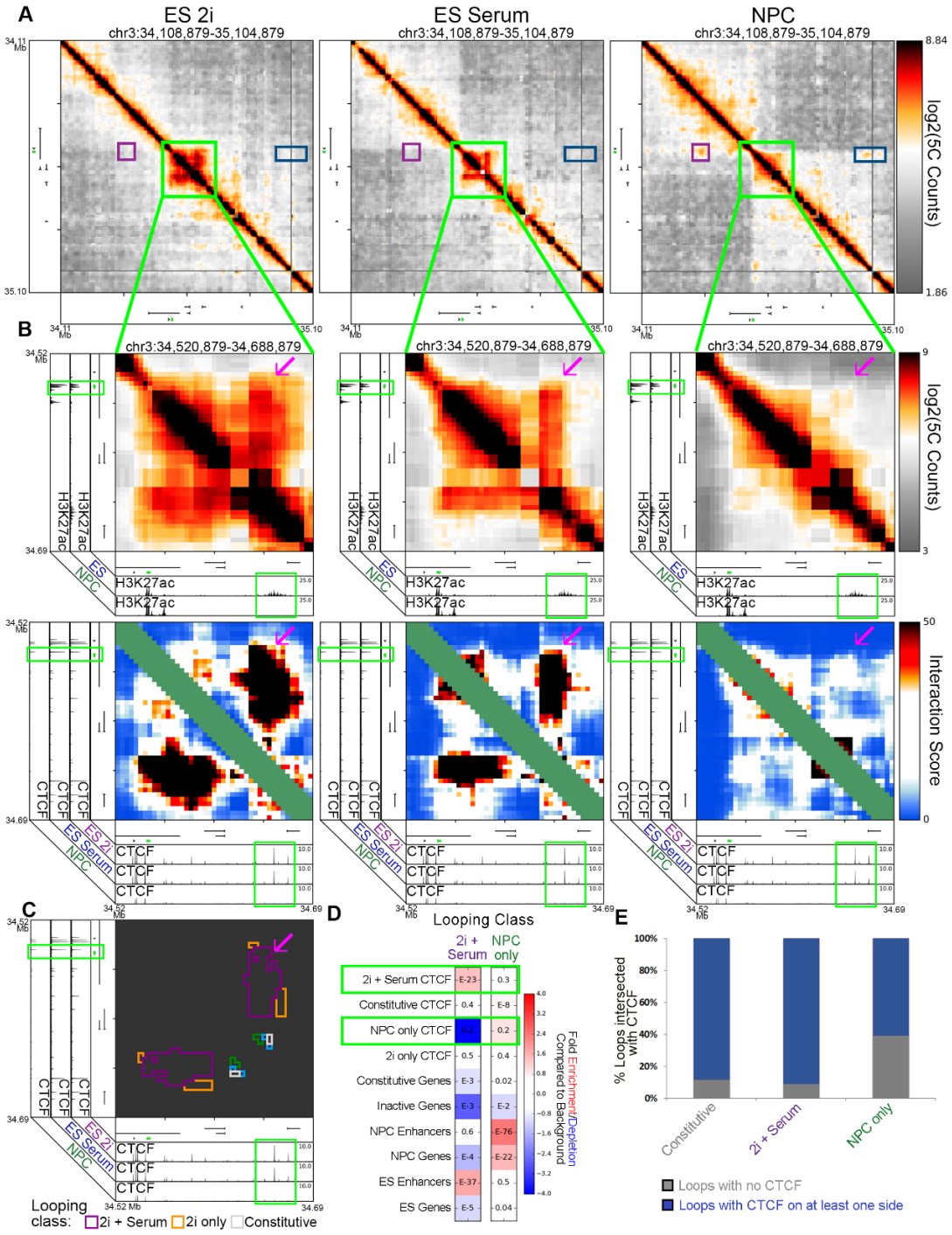


Figure 3.13. Pluripotency interactions that disengage in multipotent NPCs display reduced CTCF occupancy. (A) Global view of relative interaction frequency heatmaps of 1 Mb surrounding the *Sox2* gene. **(B)** Zoom in highlighting a strong pluripotency-specific looping interaction between *Sox2* and an ES-specific enhancer. CTCF binds at both loop anchors (note green boxes). Heatmaps include relative interaction frequency (top row) and background corrected interaction score (bottom row). *Sox2* gene is colored

green. **(C)** Classified interaction clusters are plotted above relevant ChIP-seq tracks. **(D)** Fold enrichment/depletion of chromatin features in 2i+serum and NPC only looping interaction classes compared to presence in background interactions. P-values included in each entry are calculated using Fisher's exact test. **(E)** Stacked barplot contrasting the proportion of loops connected by CTCF in one or both anchoring fragments versus not anchored by CTCF.

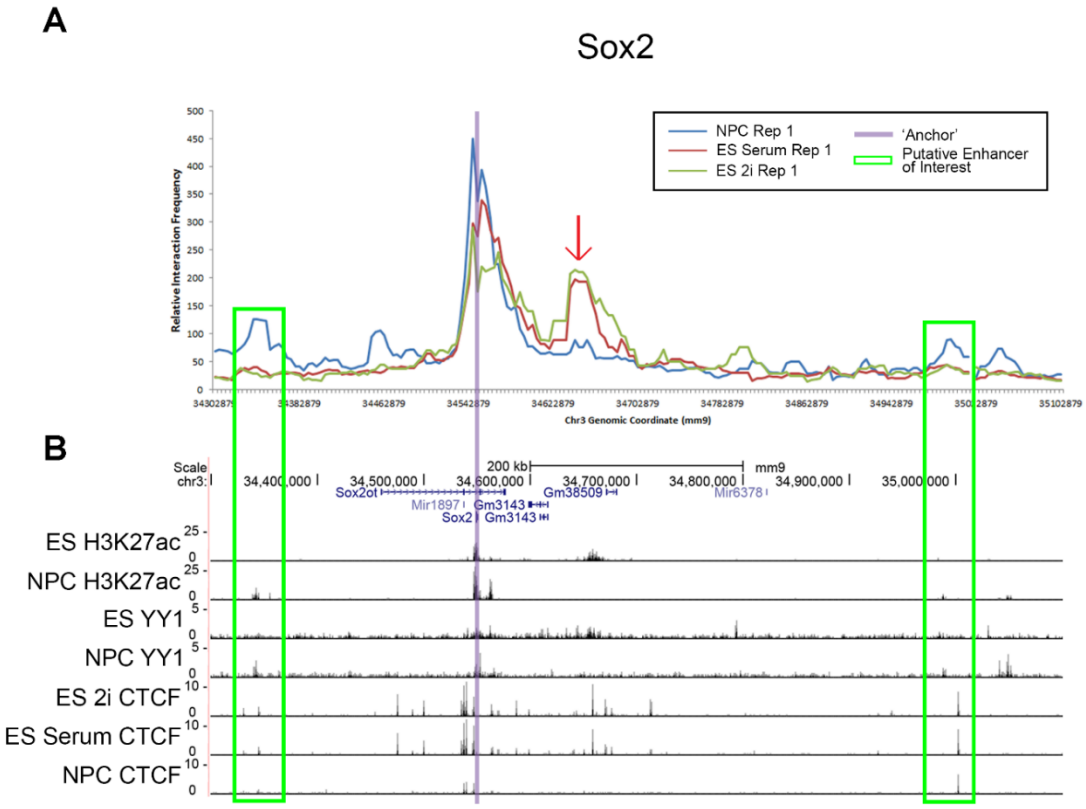


Figure 3.14. Summary of crucial interactions made by the *Sox2* gene. (*A*) Relative interaction frequencies for the interactions between the 5C bin containing the *Sox2* gene (highlighted in purple) and surrounding bins are plotted for the first ES 2i, ES Serum, and NPC replicates. Putative enhancer elements of interest are highlighted in green box(es). **(*B*)** UCSC genome browser tracks are displayed for the same locus as in (A), displaying the H3K27ac, YY1, and CTCF ChIP-seq data utilized in this study.

anchored by CTCF, we found that almost 40% of 'NPC only' interactions were not anchored by CTCF binding, whereas only <5% and <10% of 'constitutive' and '2i + serum'

interactions lacked CTCF binding, respectively (black bars, **Fig. 3.13E**). Moreover, 'NPC only' looping interactions across our 5C regions were not enriched for 'NPC only' CTCF binding (**Fig. 3.13D**). These results suggested that CTCF might not be the critical architectural protein connecting developmentally regulated looping interactions that arise de novo in differentiated NPCs.

3.2.5 YY1 binding is enriched at looping interactions connecting NPC-specific genes and distal regulatory elements

We posited that an additional class of looping proteins might connect developmentally regulated 'NPC only' interactions. We searched for the presence of candidate architectural proteins at the base of 3D interactions between genes critical to the NPC phenotype and their putative target enhancers. Since its discovery, the *Nes* gene has been a widely referenced marker of proliferating NPCs [180, 181]. Therefore, we began our search by investigating published NPC ChIP-seq libraries for their signal at the 'NPC only' long-range interactions between *Nes* and *Bcan* and a putative NPC-specific enhancer roughly 200 kb downstream of the genes (**Fig. 3.15A,B**, magenta arrowhead**, Figs. 3.16A-B**, green box). As expected, *Nes* and *Bcan* expression markedly increased in NPCs in concert with the increase in 3D contact with the putative enhancer element (**Fig. 3.15C**). Interestingly, we observed strong occupancy of the zinc-finger protein Yin Yang 1 (YY1) at the putative NPC-specific enhancer (**Figs. 3.15B+D, Fig. 3.16B**, green box x-axis). Moreover, globally across all our 5C loops, we observed that YY1 was strongly enriched

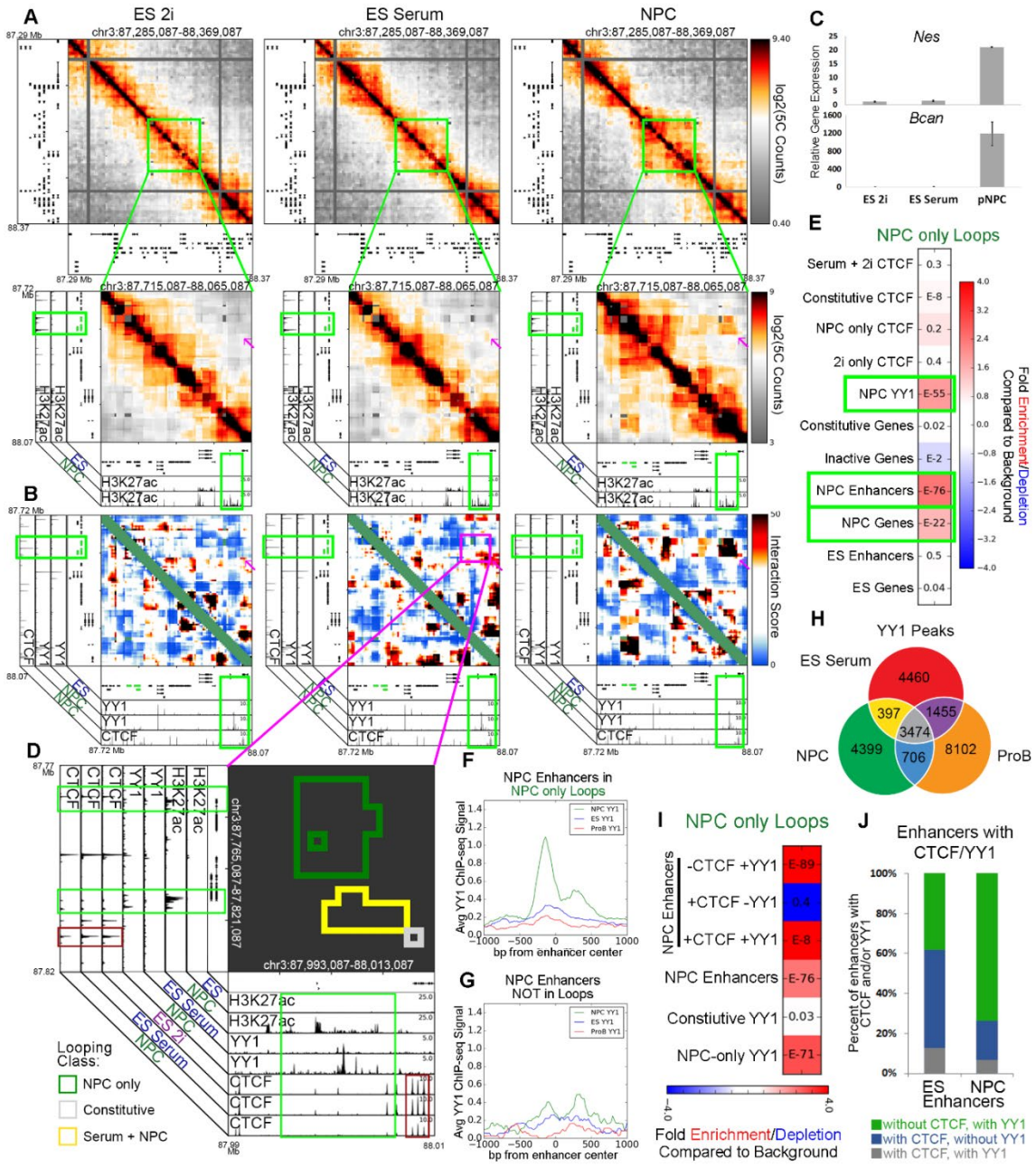


Figure 3.15. YY1 is enriched at NPC-specific enhancers that form developmentally regulated loops. (A) Relative interaction frequency heatmaps of the global view of 1 Mb surrounding *Nes* (top row), and zoom in of 400 kb surrounding nestin with putative NPC enhancer annotations (bottom row, blue bars). *Nes* (upstream) and *Bcan* (downstream) genes are colored green. **(B)** Zoom-in interaction score heatmaps of the nestin/bcan genes interacting with a downstream putative NPC enhancer. Heatmaps are overlaid with ChIP-seq tracks of CTCF in NPCs and YY1 in ES serum and NPCs. The *Nes*(upstream) and *Bcan* (downstream) genes are colored green. **(C)** Relative gene expression of *Nes* and *Bcan* across ES 2i, ES serum, and NPC cellular states. **(D)**

Interaction cluster outlines of the loop boxed in magenta in (B). Plot is overlaid with ChIP-seq tracks of H3K27ac, YY1, and CTCF in the ES 2i, ES serum and NPC conditions. Cluster outline classifications include NPC only (green), serum+NPC (yellow), and constitutive (grey). **(E)** Fold enrichment/depletion of the presence of chromatin features in NPC-only interaction class compared to presence in background. P-values are computed with Fisher's Exact test and listed in each entry. **(F-G)** YY1 ChIP-seq signal in NPCs (green) and ES serum (blue), and ProB cells (red), centered at: **(F)** putative NPC enhancers at the base of NPC only loops, **(G)** NPC enhancers that do not fall at the base of any looping interactions. **(H)** YY1 binding sites parsed by their occupancy across ES cells, NPCs, and ProB cells. **(I)** Fold enrichment/depletion of YY1 peak classes and NPC enhancers parsed based on the presence/absence of CTCF/YY1 in NPC-only loops compared to their presence background interactions. **(J)** Stacked barplot of the breakdown of ES and NPC enhancers that are bound with confidence by a combination of CTCF and/or YY1.

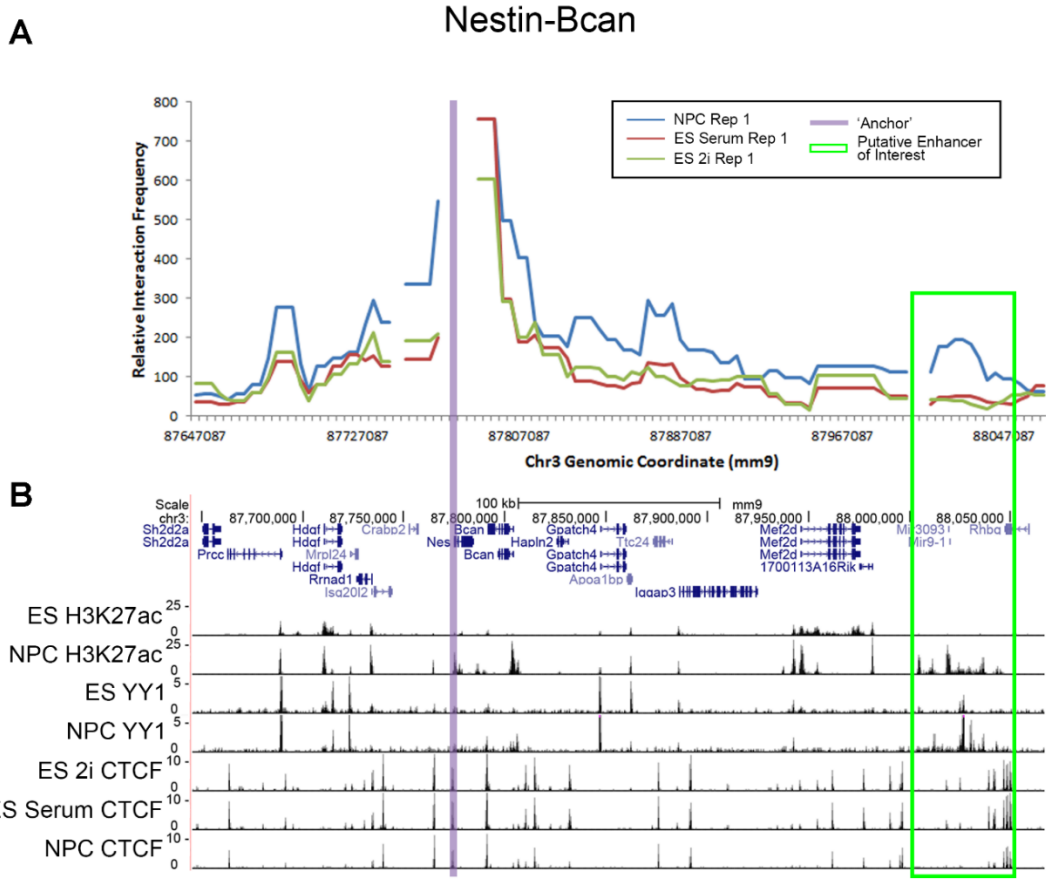


Figure 3.16. Summary of crucial interactions made by the *Nestin* gene. (*A*) Relative interaction frequencies for the interactions between the 5C bin containing the *Nestin* gene (highlighted in purple) and surrounding bins are plotted for the first ES 2i,

ES Serum, and NPC replicates. Putative enhancer elements of interest are highlighted in green box(es). **(*B*)** UCSC genome browser tracks are displayed for the same locus as in (A), displaying the H3K27ac, YY1, and CTCF ChIP-seq data utilized in this study.

in 'NPC only' 3D interactions compared to background non-loops (**Fig. 3.15E**). These data demonstrate that YY1 binding is enriched at 'NPC only' looping interactions.

To better understand the role for YY1 in NPC looping, we parsed our putative genome-wide NPC-specific enhancers into those that engage in 'NPC only' loops (**Fig. 3.15F**) and those that do not participate in long-range interactions (**Fig. 3.15G**). We found strong YY1 signal at NPC-specific enhancers engaged in 'NPC only' looping interactions and negligible YY1 binding at NPC-specific enhancers that do not loop. Similarly, YY1 signal was also enriched at NPC-specific and constitutively expressed genes in looping interactions compared to non-loops (**Figs. 3.17A-C**). These data suggest that YY1 is present at NPC regulatory elements engaged in 3D interactions and support our working hypothesis that YY1 might serve as an architectural protein to connect NPC-specific genes and enhancers.

We next set out to understand YY1 occupancy across cellular states and its co-binding with respect to CTCF. In the case of CTCF, 47%, 53% and 92% of classified binding sites were constitutive (n=26,435) in the ES 2i, ES serum and NPC cellular conditions, respectively (**Fig. 3.3A**). By contrast, a markedly lower proportion of classified YY1 sites were constitutive among ES serum, NPCs and primary pro-B cells (36%, 39% and 25%, respectively; n=3,474), indicating that YY1 might exhibit more cell type specific binding than CTCF (**Fig. 3.15H**). To understand if YY1 co-localizes with CTCF, we explored pileup plots of average ChIP signal over the different classes of dynamic YY1 binding sites (**Figs. 3.17F-H**). CTCF signal was negligible at all classes of YY1 binding,

suggesting that YY1 and CTCF do not, for the most-part, directly co-localize (**Figs. 3.17F-G**). We also observed a striking overlap of YY1 with H3K27ac signal (**Fig. 3.17H**), which prompted us to query the overlap of cell type-specific genes and regulatory elements with CTCF and YY1 and how this impacts looping. Importantly, NPC-specific enhancers were strongly enriched in 'NPC only' looping interactions when bound by YY1 without CTCF, but not when bound by CTCF without YY1 (**Fig. 3.15I**). Similarly, constitutive and NPC-specific genes were also significantly enriched in 'constitutive' and 'NPC only' looping interactions, respectively, when bound by YY1 without CTCF (**Figs. 3.17K, M**). Together, these data indicate that NPC-specific enhancers, constitutive genes and NPC-specific genes can engage in strong 3-D interactions in NPCs when bound by YY1 in the absence of CTCF.

In contrast to the role for YY1 at NPC regulatory elements, the role for YY1 at ES-specific genes and enhancers was less clear. YY1 occupancy signal was low and diffuse across putative ES-specific regulatory elements and did not show a clear preference between those engaged in loops vs. non-loops (**Figs. 3.17D-E**). By focusing on distal cell type-specific regulatory elements that overlap binding sites of CTCF, YY1, or both (and not considering those bound by neither), we observed that the majority of architectural protein-bound NPC-specific enhancers were bound by YY1 without CTCF, whereas the majority of ES-specific enhancers were bound by CTCF without YY1 (**Fig. 3.15J**). Additionally, '2i + serum' looping interactions were enriched for ES-specific enhancers regardless of CTCF and YY1 occupancy, whereas ES-specific genes were only enriched in '2i + serum' looping interactions when bound by CTCF without YY1 (**Figs. 3.17J, L**). It is not clear to what extent our observed differences between ES and NPC YY1 are due

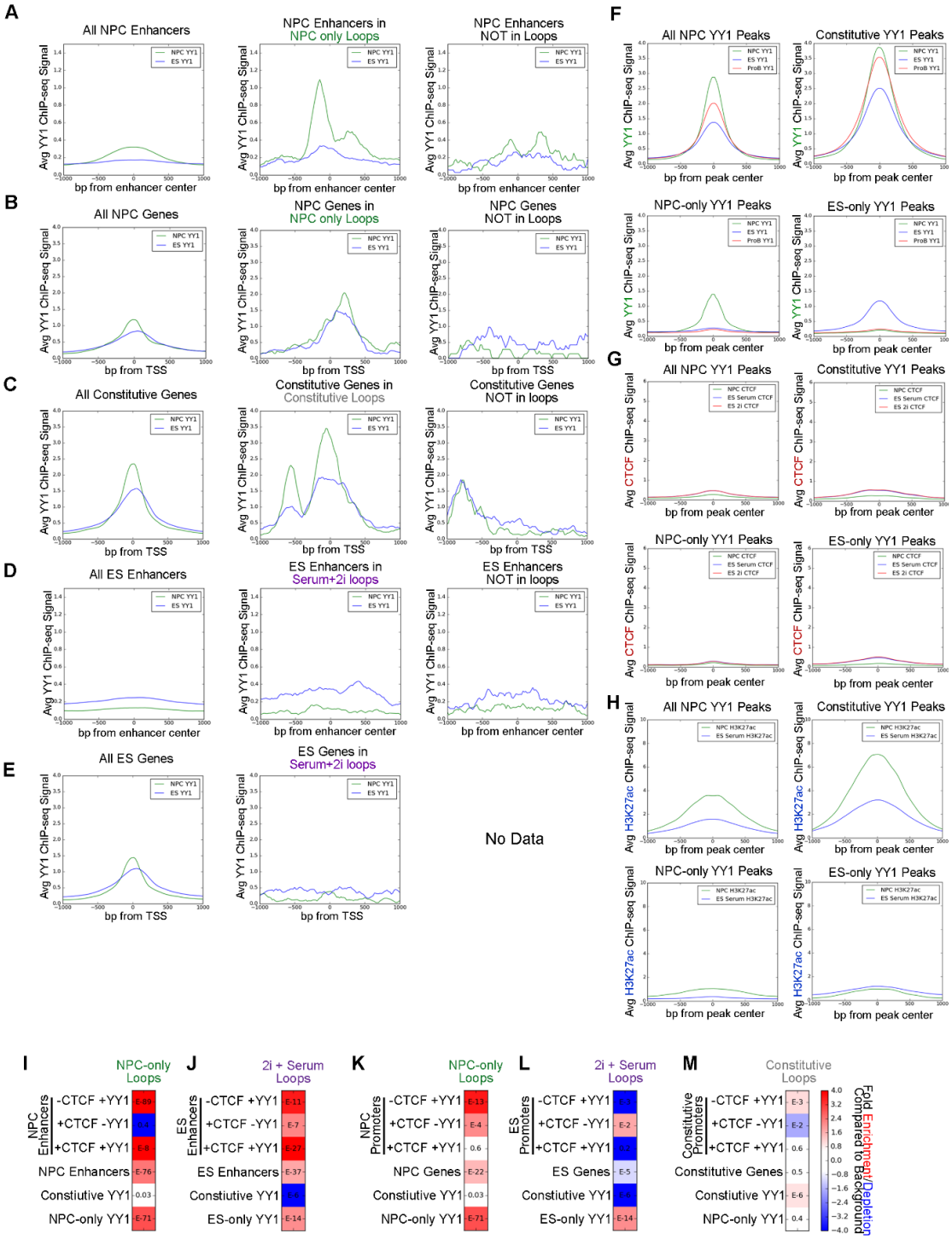


Figure 3.17. YY1 is enriched across genomic annotations 'active' in NPCs in looping interactions. (A-E) Pileups of YY1 ChIP-seq signal at (A) NPC enhancers, (B) NPC genes, (C) constitutive genes, (D) ES enhancers, and (E) ES genes for the total set of each annotation (left), the subset of each annotation found at the base of the

loops of the relevant class (middle), and the subset of each annotation not involved in any looping interaction (right). **(F)** Pileups of YY1 ChIP-seq signal at (top left) all YY1 peaks called in NPCs, (top right) YY1 peaks present in ES cells, NPCs, and ProB cells, (bottom left) NPC-specific YY1 peaks, and (bottom right) ES-specific YY1 peaks. **(G)** Pileups of CTCF ChIP-seq signal across the same set of YY1 peaks as presented in (K). **(H)** Pileups of H3K27ac ChIP-seq signal across the set of YY1 peaks listed above. **(I-M)** Fold enrichment/depletion of the parsed chromatin regulatory elements from (A-E) in the relevant looping class compared to background interactions. P-values are computed with Fisher's Exact test and listed in each entry.

to: (1) an increased reliance by ES cells on CTCF as the primary architectural protein, (2) different ChIP methods between the ES and NPC YY1 datasets or (3) a different regulatory role for YY1 in the two cell types. Thus, although our data indicate that YY1 might be important for developmentally regulated looping in somatic cells, we cannot conclusively define or rule out any role for YY1 in mediating loops in ES cells.

3.2.6 YY1-mediated developmentally regulated looping interactions are often nested within a larger framework mediated by constitutive CTCF

While exploring the gene-enhancer interaction formed by the *Nes* and *Bcan* genes, we noticed a constitutive interaction at the outer corner of the larger 'NPC only' and 'Serum + NPC' looping interaction cluster (**Fig. 3.15D**). At the base of this constitutive interaction, we identified convergently oriented constitutive CTCF sites (**Fig. 3.15D**, lower red boxes on both axes, consensus orientation not shown). We hypothesized that a subset of constitutive CTCF sites might form loops that create a pre-existing topological framework within which critical, developmentally dynamic chromatin interactions form [24, 86, 134]. Constitutive topological frameworks may be critical for proper gene expression because they create insulated neighborhoods around co-regulated genes and enhancers that

will interact during subsequent differentiation steps [12]. Consistent with this idea, constitutive CTCF was the most significantly enriched chromatin mark underlying constitutive interactions (**Fig. 3.18A**). Importantly, NPC genes were also slightly enriched in constitutive interactions (**Fig. 3.18A**), corroborating our observation that NPC gene-enhancer loops connected by YY1 often appear adjacent to and nested within constitutive looping events.

We sought additional examples of punctate 'constitutive' interactions adjacent to 'NPC only' interactions. The *Olig1* and *Olig2* genes encode bHLH transcription factors involved in differentiation along the oligodendrocyte lineage [182]. In NPCs, putative enhancer(s) marked by NPC-specific H3K27ac connect to both *Olig1* and *Olig2* in a rosette-like structure and these genes show markedly increased expression in NPCs compared to ES 2i/ES serum (**Fig. 3.18B-D,** magenta arrows). We observed significant NPC YY1 signal at all NPC-specific genes and enhancers at the *Olig1/2* locus (**Fig. 3.18E, Figs. 3.19 A-D**). Similar to the *Nes* locus, we observed two constitutive interactions anchored by constitutive CTCF sites in a convergent orientation adjacent to the NPC specific interactions formed by the *Olig1* and *Olig2* genes (**Fig. 3.18F**, red boxes/green arrows). Similarly, the *Sox2* gene also forms a long-range 'NPC only' interaction with a putative NPC-specific enhancer marked by H3K27ac (**Fig. 3.18G,** magenta arrow, **Figs. 3.14A-B**, second green box). YY1 is detected at both *Sox2* and the putative NPC-specific enhancer (**Fig. 3.18H-I,** upper green boxes). Again, the NPC only interaction exists adjacent to and nested within a punctate constitutive interaction anchored by convergent CTCF (**Fig. 3.18I,** lower red boxes, consensus orientation not shown). Together, these results support a working model of 3D genome folding in which developmentally regulated

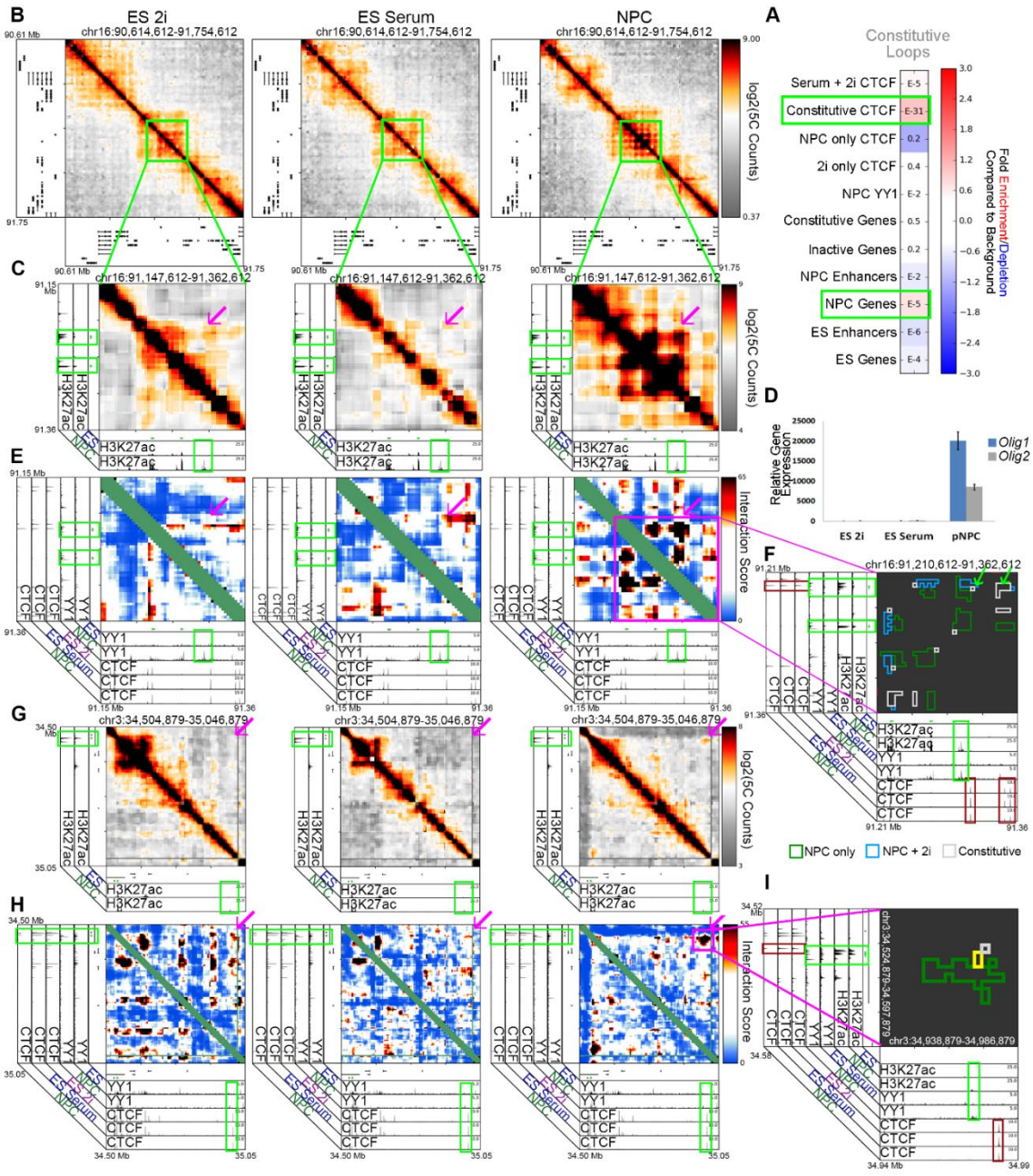


Figure 3.18. YY1 connects neural regulatory elements nested within and adjacent to a framework of constitutive CTCF-mediated interactions. (A) Fold enrichment/depletion of chromatin regulatory elements in the constitutive looping class compared to background interactions. P-values are computed with Fisher's Exact test and listed in each entry. **(B-C)** Relative interaction frequency heatmaps of **(B)** ~1Mb region and **(C)** ~200kb region surrounding the *Olig1* and *Olig2* genes in ES 2i, ES serum and NPCs. Heatmaps in (C) are overlaid with ChIP-seq tracks of H3K27ac in ES serum cells and NPCs. **(D)** Relative gene expression of *Olig1* and *Olig2 genes* across the ES 2i, ES serum, and NPC cellular states. **(E)** Zoom-in interaction score heatmaps of looping

interactions between the *Olig1* and *Olig2* genes and surrounding putative NPC enhancers (green boxes). **(F)** Zoom-in cluster map of classified looping interactions at *Olig2* and *Olig1* with NPC-only (green), serum+NPC (yellow) and constitutive class looping interactions (grey). **(G-I)** Heatmaps and cluster map at different length scales around the *Sox2* gene in ES 2i, ES serum and NPCs. Zoom-in heatmaps of relative interaction frequencies **(G)** and background corrected interaction scores **(H)** across ~500 kb downstream of *Sox2*. Relative interaction frequency heatmaps are overlaid H3K27ac tracks. Interaction score heatmaps are overlaid with ChIP-seq tracks of YY1 and CTCF across cell types. *Sox2* gene is colored green. **(I)** Zoom-in classified cluster map of a ~100 kb window around a *Sox2*-enhancer interaction with NPC-only (green), serum+NPC (yellow) and constitutive classified looping interactions (grey), overlaid on ChIP-seq tracks.

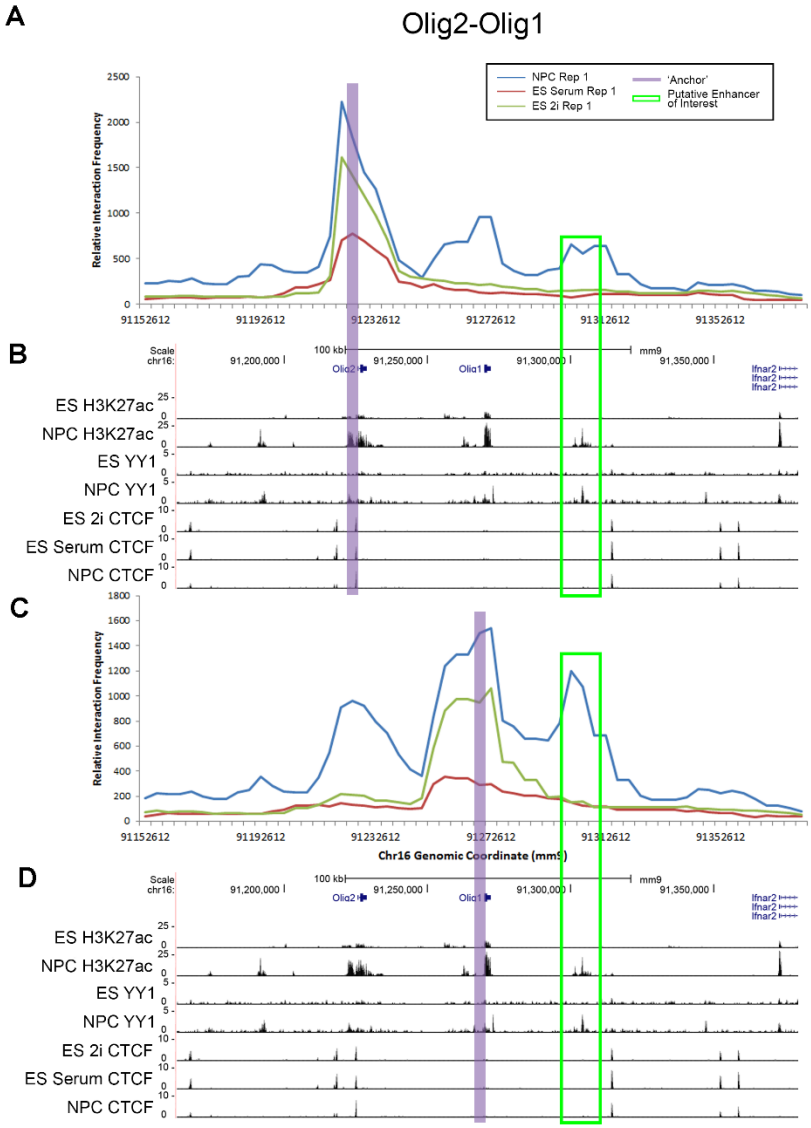


Figure 3.19. Summary of crucial interactions made by the *Olig1/Olig2* genes. **(A,C)** Relative interaction frequencies for the interactions between the 5C bin containing the Olig2 gene (A) and Olig1 gene (C) (highlighted in purple) and surrounding bins are plotted for the first ES 2i, ES Serum, and NPC replicates. Putative enhancer elements of interest are highlighted in green box(es). **(B,D)** UCSC genome browser tracks are displayed for the same loci as in (A,C), displaying the H3K27ac, YY1, and CTCF ChIP-seq data utilized in this study.

genes such as *Nes*, *Olig1*, *Olig2* and *Sox2* form de novo connections to their target enhancers via YY1 within a larger topological framework pre-existing from naïve pluripotency and connected by constitutive CTCF.

3.2.7 YY1 knockdown results in the loss of key NPC enhancer to gene looping interactions

Finally, to better understand the role for YY1 in fine-scale chromatin architecture, we knocked down YY1 in NPCs and assessed changes in looping. We performed YY1 knock down using an siRNA pool purchased from Dharmacon to target multiple sites along the YY1 transcript. Transfection of the YY1-targeting siRNA pool produced a >50% decrease in YY1 expression and protein levels compared to a control non-targeting pool condition (**Fig. 3.20A-B**). Reduction in YY1 levels resulted in a striking loss of interaction frequency between the upstream putative NPC-specific enhancer and the *Sox2* gene (**Figs. 3.5K, 3.14A-B** (second green box)**, 3.20C**) and a decrease in *Sox2* expression (**Figs. 3.20D**). We also observed loop ablation upon YY1 knockout at interactions between the *Klf4* gene and a downstream putative NPC-enhancer (**Fig. 3.21A-B, 3.22A-B,** second green box) and at the *Zfp462* gene (**Fig. 3.21C-D**). Due to technical issues related to poor library complexity across all conditions in this batch of experiments, we were unable to obtain

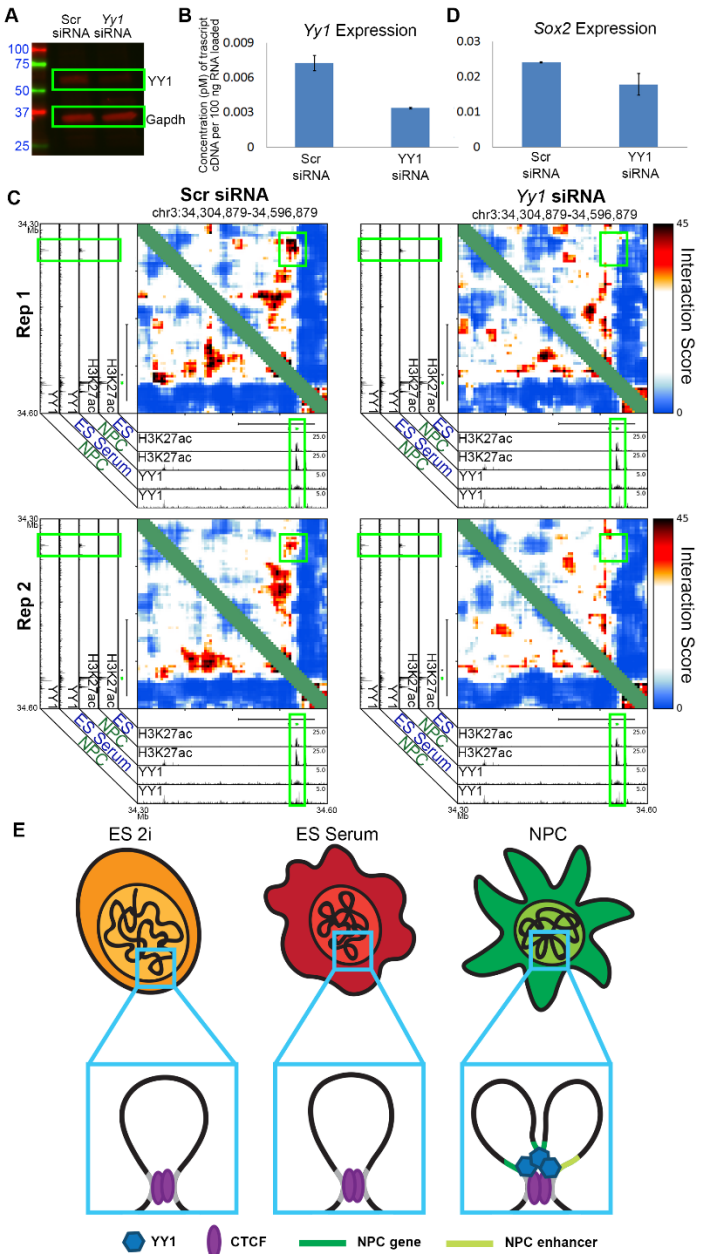


Figure 3.20. YY1-mediated developmentally regulated looping interactions form within a constitutive framework demarcated by CTCF. (A) Western blot analysis querying YY1 and Gapdh protein levels in NPCs exposed to non-targeting control and YY1-targeting siRNA. **(B)** Gene-expression quantified by qPCR of the *YY1* gene in NPCs exposed to control and YY1-targeting siRNA. **(C)** Zoom-in interaction score heatmaps of a loop between the *Sox2* gene and an upstream enhancer (originally presented in Fig. 3K) in NPCs exposed to non-targeting control siRNA (left) and an siRNA targeting YY1 (right). **(D)** Gene-expression quantified by qPCR of the *Sox2* gene in NPCs exposed to control and YY1-targeting siRNA. **(E)** Schematic depicting a CTCF-

mediated constitutive interaction, present across all early stages of neural lineage commitment, and a YY1-mediated gene-enhancer interaction, present only in NPCs.

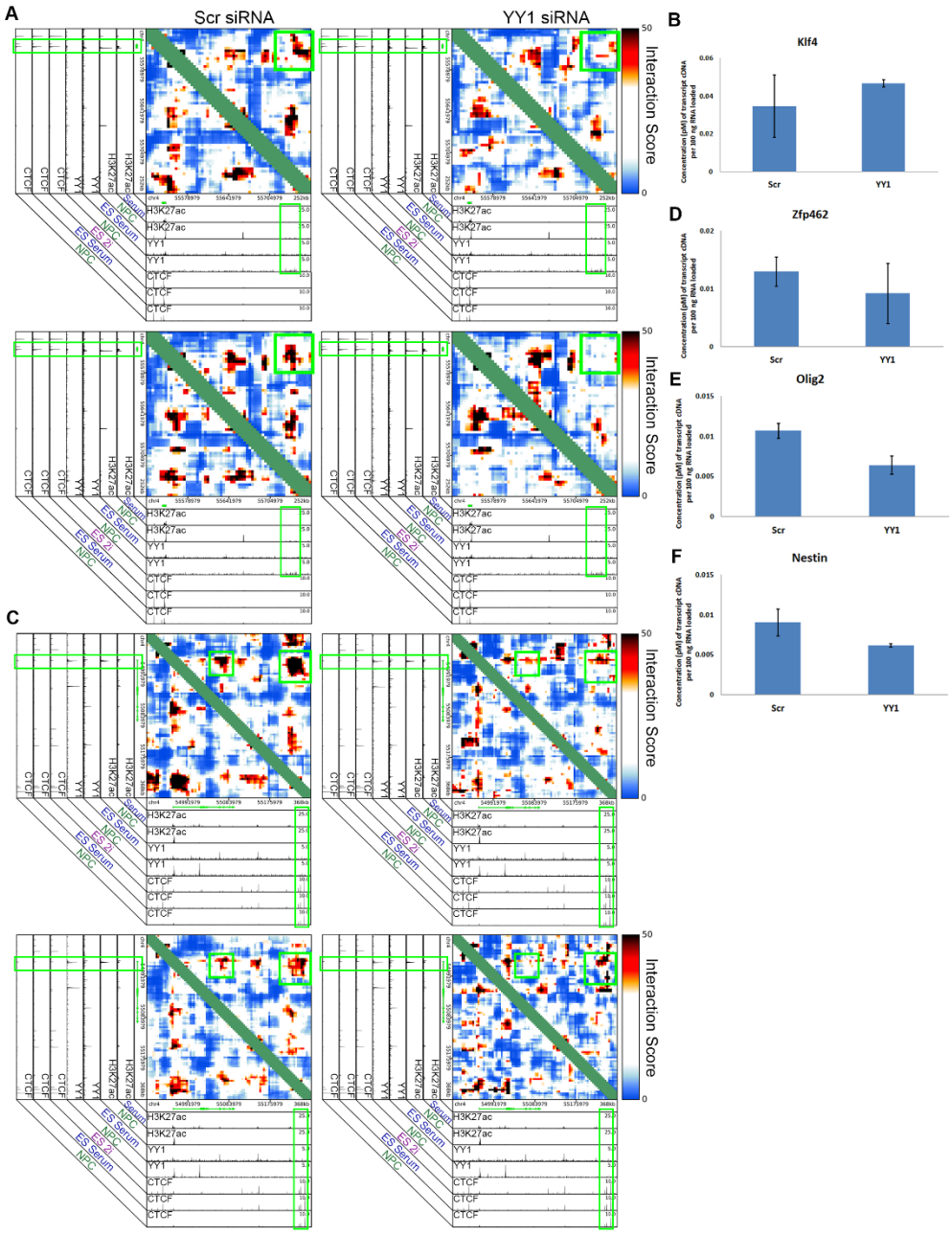


Figure 3.21. 3D looping interactions at the Klf4 and Zfp462 loci are disrupted upon YY1 knockdown. (*A*) Zoom-in interaction score heatmaps of a loop between the *Klf4* gene and downstream enhancer(s) in NPCs exposed to non-targeting control siRNA (left) and an siRNA targeting YY1 (right). **(B)** Gene-expression quantified by qPCR of the *Klf4* gene in NPCs exposed to non-targeting and YY1-targeting siRNA. **(C)** Zoom-in interaction score heatmaps of a loop between the *Zfp462* gene and downstream enhancer(s) in NPCs exposed to non-targeting control siRNA (left) and an siRNA

targeting YY1 (right). **(D-F)** Gene-expression quantified by qPCR of the *Zfp462* (D), *Olig2* (E), and *Nestin* (F) genes in NPCs exposed to non-targeting and YY1-targeting siRNA.

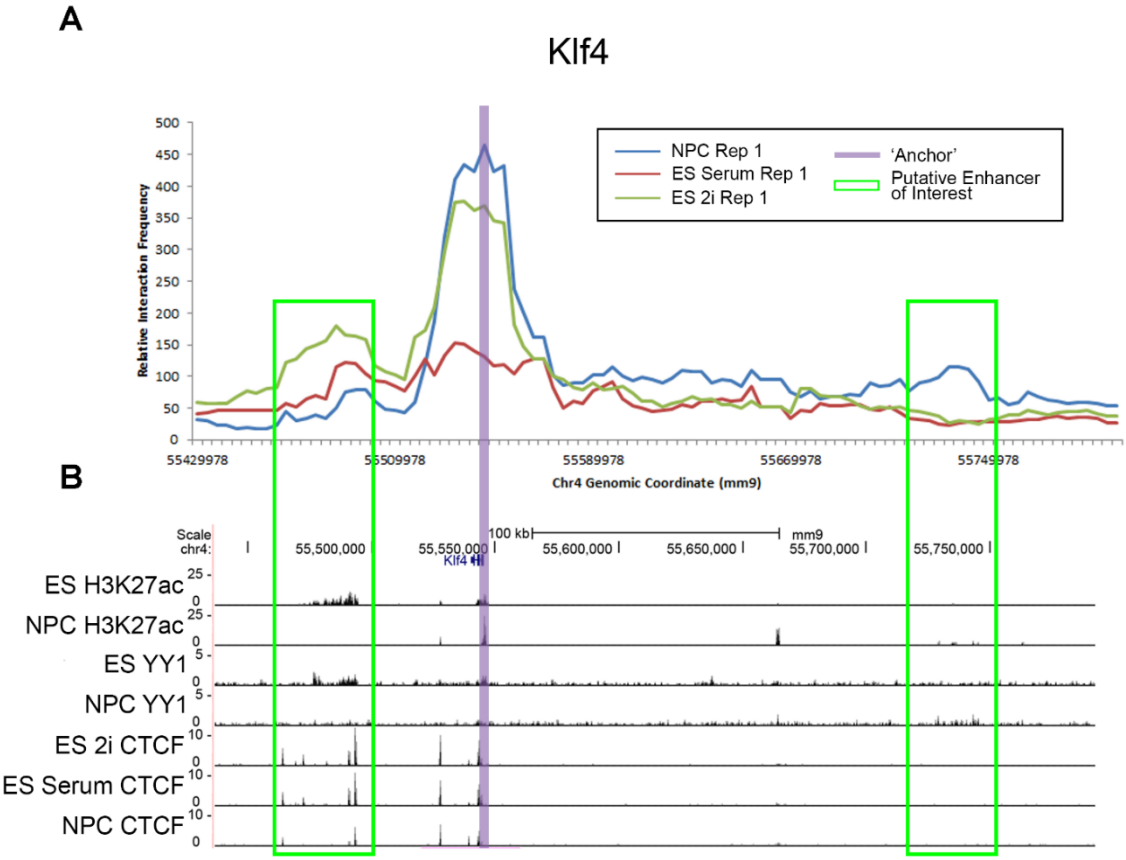


Figure 3.22. Summary of crucial interactions made by the *Klf4* gene. (*A*) Relative interaction frequencies for the interactions between the 5C bin containing the *Klf4* gene (highlighted in purple) and surrounding bins are plotted for the first ES 2i, ES Serum, and NPC replicates. Putative enhancer elements of interest are highlighted in green box(es). **(*B*)** UCSC genome browser tracks are displayed for the same locus as in (A), displaying the H3K27ac, YY1, and CTCF ChIP-seq data utilized in this study.

high complexity 5C maps at *Olig1, Olig2* and *Nes* regions. However, upon YY1 knockdown we observed a striking reduction in the expression of these genes, suggesting that the enhancer-promoter loops that *Nes*, *Olig1 and Olig2* engage in might be disrupted by YY1 knock down (**Fig. 3.21E-F**). Together, these results support our working

hypothesis that YY1 is critical for the formation developmentally regulated looping interactions in NPCs.

3.3 Discussion

CTCF is ubiquitously expressed across cell types and developmental stages and has a well-established role in connecting higher-order genome architecture. Here we seek to shed light on the dynamic CTCF binding landscape and how it is linked to the reconfiguration of chromatin architecture during the earliest stages of the establishment of neuronal expression programs. We present evidence for several organizing principles governing 3D genome folding during early brain development. First, we find that CTCF occupancy is predominantly lost in the transition from ES cells to multipotent NPCs, suggesting that the CTCF occupancy landscape might be saturated in naïve pluripotency and regulated primarily through selective pruning of CTCF binding sites. Second, reduced CTCF occupancy is correlated with the loss of chromatin interactions between ES-specific genes and enhancers, indicating that loss of CTCF binding is a critical step during the decommissioning of pluripotency gene expression programs. Third, we did not observe a strong correlation between CTCF occupancy and NPC-specific interactions. Rather, we detected high levels of occupancy of the zinc finger protein YY1 at NPC-specific genes and enhancers when engaged in NPC-specific 3D interactions and negligible YY1 levels when these regulatory elements did not interact. Upon knockdown of YY1 in NPCs, many 3D interactions break apart, suggesting that YY1 may serve as an architectural protein connecting developmentally regulated genes and enhancers in NPCs. Finally, we found that key YY1-mediated NPC-specific looping interactions occur adjacent to and nested

within punctate constitutive looping interactions anchored by convergently oriented, constitutively bound CTCF. Our data support a model in which YY1-anchored looping interactions arise de novo in NPCs within a larger topological framework established prior to or during naïve pluripotency and connected by constitutively bound CTCF (**Fig. 3.20E**).

Seminal genome-wide CTCF occupancy studies based on 2-3 cell types initially suggested that the CTCF binding landscape remains largely unchanged across mammalian lineages [73, 75]. A more recent comparison of CTCF occupancy across 40 cell types revealed that at least 80% of CTCF sites are dynamic across cellular states [159]. Here we find that CTCF occupancy is highest in the naïve pluripotent stem cell state and globally decreases in parallel with its expression during the commitment to multipotent NPCs. A large cohort of ~8,000 and ~20,000 CTCF sites are lost during the transition from ES 2i to ES serum and ES serum to NPCs, respectively. By contrast, we only observe a small group of ~1200 CTCF sites that are acquired de novo in NPCs, suggesting that the vast majority of the CTCF sites occupied in NPCs were pre-existing from earlier stages in development. We speculate that one hallmark of the initial establishment of the neuronal lineage is a wave of CTCF occupancy loss to remove residual topological configurations required for pluripotency-specific gene expression and off-target lineages that will not be expressed in brain cell types. In the future, additional studies across non-neuronal lineages will also be important to determine how widely our model of CTCF pruning in neural development applies, as our initial analyses of ENCODE data indicated that CTCF occupancy does not always decrease during development across all lineages.

DNA methylation is a critical regulator of neural lineage commitment [183] and CTCF binding [80]. Recent reports suggest that the largest re-arrangement of DNA methylation

during neural development occurs during upon the departure from pluripotency [172, 184]. The transition from ES cells to early NPCs is associated with a large increase in DNA methylation [172]. Importantly, a large proportion of genomic loci that are methylated in NPCs maintain the mark through the duration of neural development [172, 184]. Because CTCF binding is anti-correlated with DNA methylation, we posit that a notable proportion of the large class (n~20,000) of '2i + serum' CTCF sites might be might be methylated during the initial establishment of the neural lineage and subsequently remain methylated and unbound through terminal differentiation in the developing/maturing brain. In support of this hypothesis, we observe the highest levels of CTCF occupancy in our naïve pluripotency cellular state (ES 2i) in which cells have consistently been found to exhibit an extreme state of hypomethylation across the genome [175-177]. The large-scale shifts in DNA methylation and CTCF binding during the transition from ES cells to NPCs suggest that elucidating the CTCF landscape in the progenitor state of development is critically important for understanding the CTCF sites and 3-D topological configurations available for binding across terminally differentiated lineages in the brain.

Although CTCF is the best understood protein-mediated mechanism for connecting 3D chromatin interactions, we hypothesized that additional architectural proteins might exist to connect the 3-D genome. Here, unexpectedly, we found that CTCF was not significantly enriched in NPC-specific loops in our 5C regions. Rather, we observed high levels of the zinc finger protein YY1 at NPC-specific genes and enhancers when engaged in 3-D looping interactions and negligible/low YY1 occupancy when these regulatory elements were not connected. Several key NPC-specific enhancer-gene looping interactions were ablated upon YY1 knock down. YY1 is an intriguing architectural protein

candidate: (i) it is strongly enriched in genome-wide looping interactions in human cell lines [10], (ii) it is necessary for the formation of specific 3D interactions in B cells [185], (iii) it can connect a long-range interaction in B cells in the absence of its transcriptional activation domain [160] and (iv) it is required for proper neural development [186]. Biochemical studies have indicated that the zinc fingers of YY1 may interact with the N-terminus of CTCF [161], suggesting that YY1 could function via homodimerization or heterodimerization mechanisms to connect the genome. Overall, our data are consistent with a model in which YY1 serves a key role in development as a dynamic architectural protein connecting lineage-specific genes and enhancers. Future studies should aim to elucidate the mechanisms by which YY1 connects long-range chromatin interactions and the extent to which YY1 functions as an architectural protein in non-neural lineages. It will also be important to rule out the possibility that YY1's critical role in looping is not due to indirect effects on chromatin activity.

A key finding of this manuscript is that YY1-mediated looping interactions in NPCs are nested within larger constitutive interactions anchored by constitutively occupied CTCF sites. A leading hypothesis is that subTAD/TAD boundaries, anchored by constitutive CTCF, might constrain developmentally regulated enhancers from aberrantly looping to off-target genes [12, 15, 20, 22, 134]. We and others have previously reported that pluripotency genes connect to enhancers in smaller looping interactions nested within larger constitutive structures [9, 12]. Here, our results confirm and extend this model to suggest that CTCF-mediated constitutive interactions might also might serve to pre-mark genomic locations of connections between somatic developmentally regulated gene-enhancer interactions through punctate, constitutive 'seed' interactions. In agreement with this idea,

Ruan and colleagues reported evidence that CTCF-mediated looping interactions might function to coordinate nearby interactions involving RNA Pol II [86]. Furthermore, a recent genomics analysis showed that up to 30% of YY1 sites bind at locations directly adjacent to CTCF and might work together to cooperatively influence occupancy [187]. Thus we posit that architectural proteins such as YY1 might cooperatively build upon a constitutive CTCF architectural 'seed' scaffold to connect nearby developmentally regulated genes and enhancers. Future work teasing out the causal interplay between architectural seeds, CTCF and additional architectural proteins will shed light on the fundamental mechanisms governing proper spatiotemporal regulation of gene expression during development.

3.4 Acknowledgements

Thank you to co-authors Michael Duong, Katelyn Titus, Linda Zhou, Zhendong Cao, Jingjing Ma, Caroline Lachanski, and Daniel Gillis for their critical contributions to this chapter. Thank you to all members of the Cremins lab for helpful discussions, in particular Jesi Kim and Thomas Gilgenast for computational guidance.

CHAPTER 4: LOCAL GENOME TOPOLOGY CAN EXHIBIT AN INCOMPLETELY REWIRED 3D-FOLDING STATE DURING SOMATIC CELL REPROGRAMMING

4.1 Introduction

Mammalian genomes are folded in a hierarchy of architectural configurations that are intricately linked to cellular function. Individual chromosomes are arranged in distinct territories and then are further partitioned into a nested series of Megabase (Mb)-sized topologically associating domains (TADs) [7, 8] and smaller sub-domains (sub-TADs) [9, 10]. TADs/subTADs vary widely in size (i.e. 40 kb - 3 Mb) and are characterized by highly self-associating chromatin fragments demarcated by boundaries of abruptly decreased interaction frequency. Long-range looping interactions connect distal genomic loci within and between TADs/subTADs [9, 10, 23, 38]. Single TADs, or a series of successive TAD/subTADs, in turn congregate into spatially proximal, higher-order clusters termed 'A/B compartments'. Compartments generally fall into two classes: (i) 'A' compartments enriched for open chromatin, highly expressed genes and early replication timing and (ii) 'B' compartments enriched for closed chromatin, late replication timing and co-localization with the nuclear periphery [10, 40, 44, 45]. The organizing principles governing genome folding at each length scale remain poorly understood.

Recent high-throughput genomics studies have shed new light on the dynamic nature of chromatin folding during embryonic stem (ES) cell differentiation. Up to 25% of compartments in human ES cells switch their A/B orientation upon differentiation [44].

Compartments that switch between A and B configurations display a modest, but correlated alteration in expression of only a small number of genes, suggesting that compartmental switching does not deterministically regulate cell type-specific gene expression [44]. Similarly, lamina associated domains are dynamically altered during ES cell differentiation [188]. For example, the *Oct4*, *Nanog* and *Klf4* genes relocate to the nuclear periphery in parallel with their loss of transcriptional activity as ES cells differentiate to astrocytes. TADs are largely invariant across cell types and often maintain their boundaries irrespective of the expression of their resident genes [7]. By contrast, long-range looping interactions within and between sub-TADs are highly dynamic during ES cell differentiation [9, 189]. Pluripotency genes connect to their target enhancers through long-range interactions and disruption of these interactions leads to a marked decrease in gene expression [39, 190]. Thus, data is so far consistent with a model in which chromatin interactions at the sub-Mb scale (within TADs) are key effectors in the spatiotemporal regulation of gene expression during development.

In addition to the forward progression of ES cells in development, somatic cells can also be reprogrammed in the reverse direction to induced pluripotent stem (iPS) cells via the ectopic expression of key transcription factors [191]. Since the initial pioneering discovery, many population-based and single cell genomics studies have explored the molecular underpinnings of transcription factor-mediated reprogramming [154, 192-194]. Recent efforts have uncovered changes in transcription, cell surface markers and classic epigenetic modifications during intermediate stages in the reprogramming process [195-197]. Although there is some evidence of epigenetic traces from the somatic cell of origin [198-200],

the emerging model is that ES-like epigenetic and transcriptional states can be generally reset under proper reprogramming conditions [201].

The role for chromatin topology in the acquisition of pluripotency during reprogramming has not yet been elucidated. Recent studies have suggested that specific long-range interactions between *Nanog* and/or *Oct4* and target enhancers can be reset during reprogramming and precede re-activation of the involved genes [190, 202-205]. Beyond these initial locus-specific studies, it remains unknown whether the somatic cell genome unfolds/refolds at the sub-Mb scale within TADs and how chromatin topology is linked to gene expression changes during reprogramming. Here we report a detailed analysis of local chromatin folding changes during somatic cell reprogramming. We created ~4-12 kilobase (kb) resolution chromatin architecture maps in primary neural progenitor cells (NPCs), iPS cells derived from primary NPCs and pluripotent ES cells. We employed Chromosome-Conformation-Capture-Carbon-Copy (5C) to query fine-scale architectural changes in Mb-sized regions around key developmentally regulated genes. We find that chromatin folding is markedly reconfigured within TADs during the transition from primary NPCs to iPS cells. In many cases, pluripotency genes re-engage in fully reprogrammed interactions with their target ES-specific enhancers. Unexpectedly, we also observe NPC interactions around key pluripotency genes (e.g. *Sox2, Klf4*) that remain persistently tethered in our iPS clone. Pluripotency genes engaged in 'persistent NPC-like' interactions can exhibit over/under-shooting of gene expression levels in iPS, despite the fact that they may have also re-established contact with their target ES-specific enhancer(s). We also uncover a subset of 'poorly reprogrammed' interactions that break apart during differentiation and do not fully reconnect in our iPS clone. Many 'poorly reprogrammed' interactions exhibit ES-specific

CTCF occupancy that is lost during differentiation and only partially recovered in iPS cells. Importantly, 2i/LIF conditions can (i) abrogate 'persistent NPC-like' interactions, (ii) recover 'poorly reprogrammed' interactions, (iii) re-instate inadequately reprogrammed CTCF occupancy and (iv) restore precise gene expression levels.

4.2 Results

4.2.1 Chromatin folding markedly reconfigures at the sub-Mb scale during reprogramming

To investigate changes in 3D chromatin topology during somatic cell reprogramming, we first generated ~4-12 kb-resolution chromatin architecture maps in primary NPCs, iPS cells derived from primary NPCs and ES cells (**Fig. 4.1A**). To achieve a comparable genetic background to our pluripotency model (V6.5 ES cells; 129/SvJae x C57BL/6), we selected a previously published iPS clone derived from primary NPCs isolated from neonatal brains of Sox2-green fluorescent protein (Sox2-GFP) indicator mice (mixed 129/SvJae x C57BL/6 genetic background) [173, 206]. Hochedlinger and colleagues generated this iPS clone via the transduction of primary Sox2-GFP NPCs with doxycycline-inducible lentiviral vectors encoding Oct4, Klf4 and c-Myc. Importantly, this iPS clone was extensively characterized for its pluripotent properties as assessed by (i) expression of endogenous pluripotency markers (Oct4, Sox2, Nanog), (ii) demethylation of Oct4 and Nanog promoters, (iii) transgene-independent self renewal, (iv) in vivo teratoma formation of all three germ layers and (v) generation of chimeric mice [206]**.** Our three cellular states enable a detailed analysis of how chromatin unfolds/refolds between

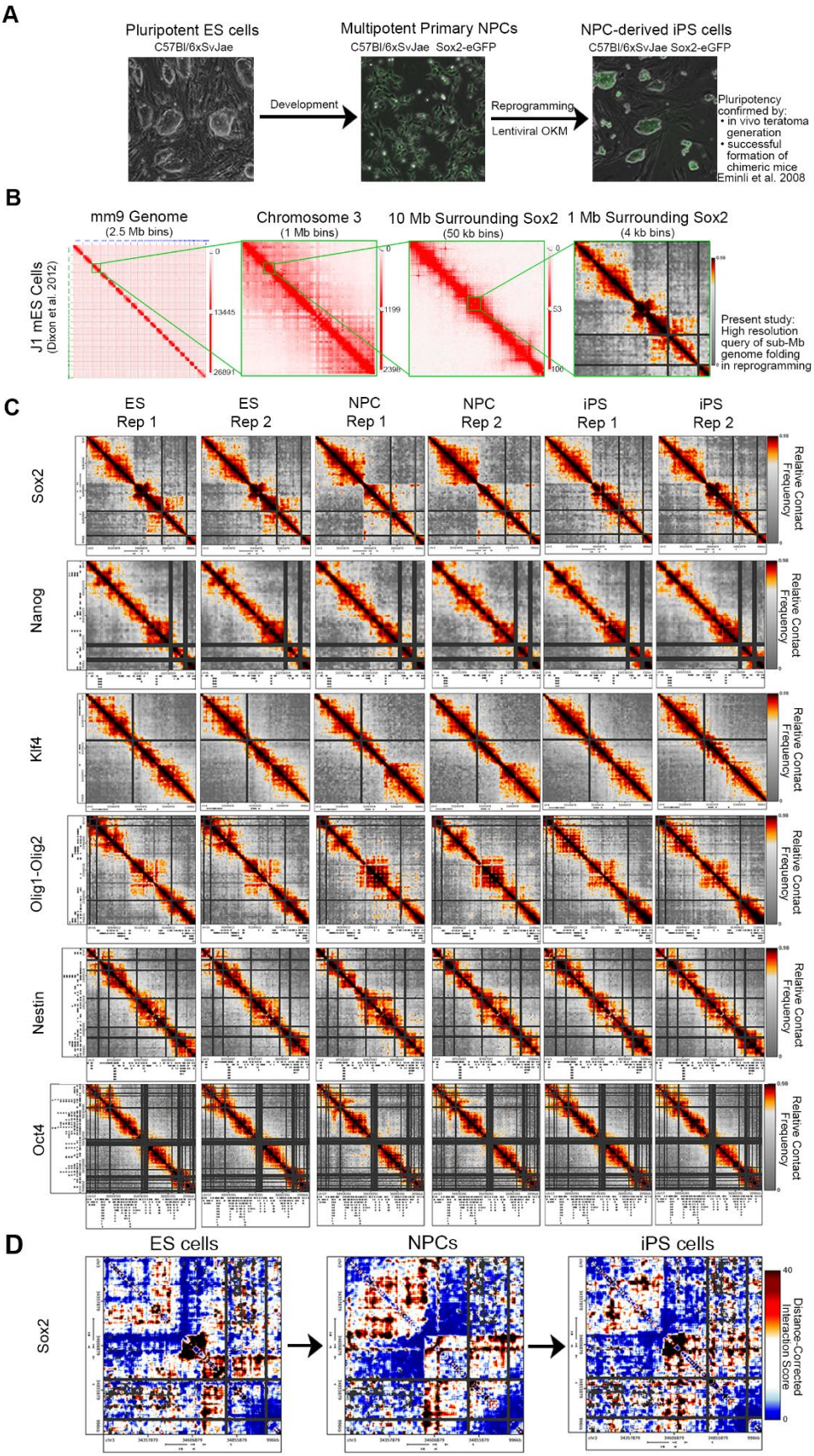

A
Pluripotent ES cells
C57Bl/6xSvJae
Development
Multipotent Primary NPCs
C57Bl/6xSvJae Sox2-eGFP
Reprogramming
Lentiviral OKM
NPC-derived iPS cells
C57Bl/6xSvJae Sox2-eGFP
Pluripotency confirmed by:
• in vivo teratoma generation
• successful formation of chimeric mice
Eminli et al. 2008
B
mm9 Genome
(2.5 Mb bins)
Chromosome 3
(1 Mb bins)
10 Mb Surrounding Sox2
(50 kb bins)
1 Mb Surrounding Sox2
(4 kb bins)
J1 mES Cells
(Dixon et al. 2012)
Present study:
High resolution query of sub-Mb genome folding in reprogramming
C
ES Rep 1
ES Rep 2
NPC Rep 1
NPC Rep 2
iPS Rep 1
iPS Rep 2
Sox2
Nanog
Klf4
Olig1-Olig2
Nestin
Oct4
Relative Contact Frequency
D
ES cells
NPCs
iPS cells
Sox2
Distance-Corrected Interaction Score

Figure 4.1. High-resolution architecture maps reveal marked chromatin reconfiguration during somatic cell reprogramming. (A) Phase contrast images of the reprogramming model system. **(B)** Genome-wide ES cell Hi-C data [7] at different bin sizes illustrating chromosome territories, A/B compartments and TADs. Images made with the Juicebox tool (http://www.aidenlab.org/juicebox/). The 4-12 kb resolution heatmaps from the present study query fine scale genome folding at the sub-Mb scale within TADs. **(C)** Relative contact frequency heatmaps are displayed for all biological replicates and regions queried. Color bars range from low (grey) to high (red/black) interaction frequencies. **(D)** Distance-corrected interaction score heatmaps for a select region around the *Sox2* gene illustrating the presence of dynamic chromatin architecture among ES, NPC and iPS cells. Color bars range from low (blue) to high (red/black) interaction scores.

NPCs and iPS cells and also facilitate the comparison of genome topology between ES/iPS of comparable genetic background.

We employed 5C and high-throughput sequencing to create fine-scale chromatin architecture maps spanning > 7 Mb of the mouse genome within a set of TADs [43]**.** 5C combines Chromosome-Conformation-Capture (3C) with a primer-based hybrid capture step to facilitate cost-effective detection of sub-Mb scale interactions in Mb-sized loci of interest [5]. We used a tiled/alternating primer design around *Nanog, Sox2, Klf4, Oct4, Nestin*, and *Olig1-Olig2* (described in detail [9]). Our 5C primer design scheme enabled the creation of ~4-12 kb resolution architecture maps for all loci combined across three cellular states with less than 30 million reads per replicate. The power in this approach is that it focuses on elucidating fine scale architecture changes at the sub-Mb scale within TADs (**Fig. 4.1B**).

We first visualized 5C data with contact frequency heatmaps. To resolve underlying topological features, we developed an analysis pipeline to correct for known biases in 5C data and to normalize samples within and between biological replicates. Briefly, raw data

(**Fig. 4.2A**) were quantile normalized to bring the dynamic range of all samples onto equivalent scales and to account for technical differences in sequencing depth and library complexity (**Fig. 4.2B**). To account for differences in primer efficiency that lead to non-uniformities in coverage across genomic regions, we applied our previously published primer correction algorithm to quantile-normalized data (**Fig. 4.2C**, [9]). We then applied a blocked binning/smoothing algorithm to attenuate spatial noise in 5C data (**Fig. 4.2D**). Our 'Relative Contact Frequency' heatmaps revealed striking topological patterns that are dynamic across cellular states and unique to each genomic region (**Fig. 4.1C**).

To further resolve the underlying architectural signal, we corrected for the known distance-dependence background in 5C data [23] (**Fig. 4.2E-G**). Consistent with recent reports [10], we found that a local distance-dependence model computed independently for each region would more precisely account for locus-specific differences in chromatin folding that are often over/under-estimated by a global background model (**Fig. 4.2G**). Our 'Distance-Corrected Interaction Score' heatmaps showed striking changes in topological features among NPCs, iPS and ES cells (**Fig. 4.1D, Fig. 4.2E-F**) with high consistency between replicates and marked differences among biological conditions. A systematic comparative analysis at each stage in the pipeline confirmed that we have reduced known biases in 5C data (**Figs. 4.2A-I, 4.3A-G**).

4.2.2 iPS genomes can exhibit imperfectly rewired folding patterns

We next explored fine-scale chromatin folding features within TADs by visually inspecting our heatmaps. Consistent with our previous work, we observed marked changes

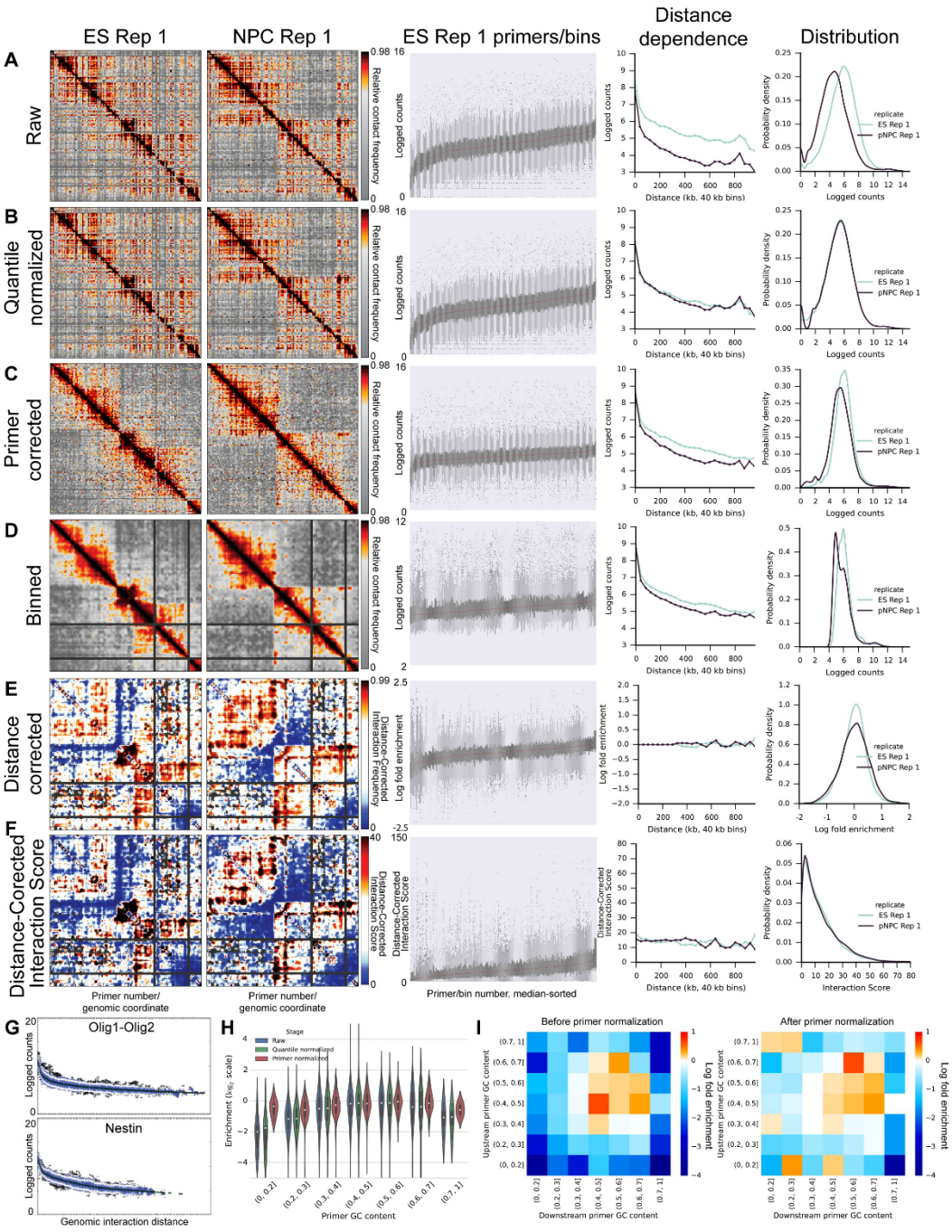


Figure 4.2. Progression of 5C data through analysis pipeline. (A-F) Grid showing progression of Sox2 region through data processing steps. From top to bottom: **(A)** raw, **(B)** quantile normalized, **(C)** primer corrected, **(D)** binned (4 kb bins; 20 kb smoothing window), **(E)** distance-dependence corrected and **(F)** interaction score computed as -10*$\log_2$(p-value) on p-values computed from the distance-dependence corrected data after logistic distribution modeling parameterized for each genomic region. From left to right: (i) contact probability heatmaps for ES Rep1 and NPC Rep1, (ii) boxplots of counts for each primer/bin in the Sox2 region in order of increasing median, (iii) background

distance-dependence interaction frequency, showing the mean of the counts at distance scales binned every 40 kb, (iv) kernel density estimates of the counts probability density. **(G)** Boxplots of 'Relative contact frequency' values at 4 kb intervals across the genomic coordinates queried for each 5C region. Plots for the Olig1-Olig2 and Nestin regions of ES Rep 1 are shown. **(H)** Violin plots showing the distribution of log fold enrichment of total cis primer counts over the mean of cis primer counts (x-axis) as a function of each primer's GC content (y-axis). Data for ES Rep 1 is shown at raw, quantile normalization and primer correction stages in the analysis pipeline. **(I)** Heatmaps comparing GC content bias in ES Rep1 in pairwise fragment-to-fragment contacts before and after primer correction. Fold enrichment is computed within each two-sided GC bin as the sum of the counts for all cis primer-primer pairs falling in the GC content range of the bin divided by the expected number of counts for a bin with that many primer-primer pairs in it.

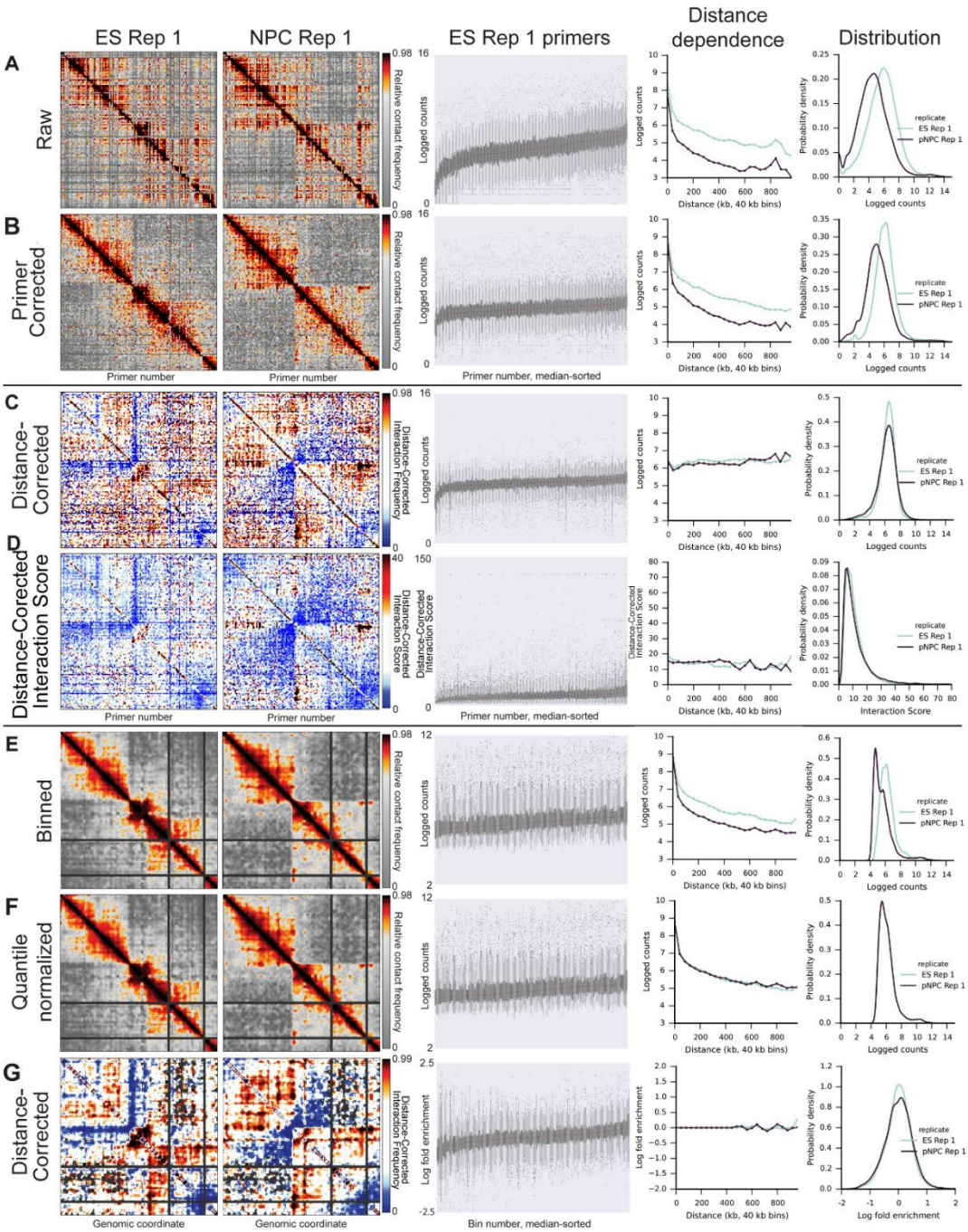


Figure 4.3. Progression of 5C data through alternative 5C analysis approaches. (A-D) Grid showing progression of Sox2 region through our previously published analysis pipeline [9]. From top to bottom: **(A)** raw, **(B)** primer corrected, **(C)** distance-dependence normalized via parametric model described in [9] and **(D)** interaction score computed as -10*$\log_2$(p-value) on p-values computed with compound normal-lognormal distribution fits described in [9]. From left to right: (i) contact probability heatmaps for ES Rep1 and NPC Rep1, (ii) boxplots of counts for each primer/bin in the Sox2 region in order of

increasing median, (iii) distance dependence curves, showing the mean of the counts at distance scales binned every 40 kb, (iv) kernel density estimates of the counts probability density. **(E-G)** Grid showing downstream effects of alternative placement of quantile normalization step within the main 5C analysis pipeline. Primer normalized data shown in **(B)** were binned **(E)**, then quantile normalized (in contrast to Figure 4.2, where quantile normalization is the first step) **(F)**, and finally distance corrected **(G)**.

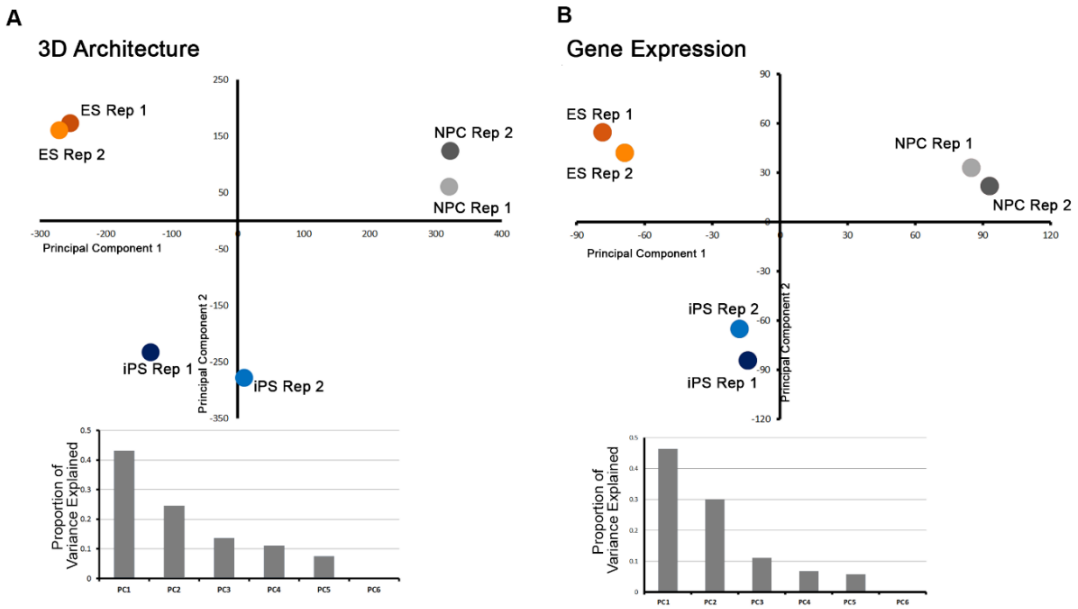


Figure 4.4. iPS genomes can exhibit intermediate folding and expression patterns between somatic and pluripotent stem cell states. Principal component analysis of **(A)** distance-corrected interaction frequency data and **(B)** normalized RNAseq data for ES, NPC and iPS replicates. **(A, B)** Principal components 1 and 2 are scattered and the proportion of variance explained by each principal component is plotted below each scatterplot.

in chromatin architecture between ES cells and NPCs. Importantly, we also noticed a striking architectural reconfiguration between NPCs and NPC-derived iPS cells (**Fig. 4.1C-D**). At many loci, iPS genome folding recapitulates the patterns seen in V6.5 ES cells. However, we also noticed several intriguing cases where iPS topology retained remnants of the folding patterns from NPCs (**Fig. 4.1D**).

To further explore the possibility that genome folding might be mis-wired during reprogramming, we conducted principal component analysis on our 'Distance-Corrected Interaction Frequency' data across all replicates and cellular states. Interestingly, we observed that genome topology in our iPS clone exhibited folding patterns that were intermediate between NPCs and the pluripotent stem cell state (**Fig. 4.4A**). To explore the functional significance of potential intermediate iPS folding patterns, we queried the transcriptome of all three cellular states using RNAseq. Consistent with our 3D observations, global gene expression profiles in our iPS clone were also parsed as intermediate between ES cells and NPCs (**Fig. 4.4B**). Together, these results support the possibility that genome architecture of some iPS clones might be imperfectly wired within TADs during reprogramming.

4.2.3 Dynamic 3-D interaction classes during cell fate transitions

To identify high-confidence, long-range interactions across all developmentally regulated loci, we fit our 'Distance-Corrected Interaction Frequency' data with a logistic distribution with location/scale parameters computed independently for each region (**Fig. 4.5A, Appendix II Methods**). We then converted the p-values from our fitted models into an interaction score (-10*log2(p-value)) that is comparable within and between experiments and allows for the robust detection of interactions that are significant above the expected background signal.

We next employed a thresholding strategy to classify 3D interactions by their dynamic contact frequencies across the three cellular states (**Fig. 4.6A-D**). To minimize false positives, we required that interaction scores cross the threshold boundaries in both

replicates for a given biological condition. Moreover, we iteratively defined thresholds to achieve an empirical False Discovery Rate (eFDR) of < 10% when applied to simulated 5C replicates (**Figs. 4.6E-H, 4.5B+C**, **Appendix II Methods**). Upon application of our classification scheme, we uncovered several dynamic interaction classes among ES, NPC and iPS cellular states (**Figs. 4.6I-J**), including: (i) 537 interactions present in ES cells, lost in NPCs and reacquired upon reprogramming (purple class) (**Fig. 4.6K**), (ii) 3004 interactions present only in ES cells and not reprogrammed (red class) (**Fig. 4.6L**), (iii) 5043 interactions absent in ES cells, acquired upon differentiation and lost in iPS cells (green class) (**Fig. 4.5D**), (iv) 1708 interactions present only in iPS cells (orange class) (**Fig. 4.5E**), (v) 148 interactions that are high in ES cells and NPCs and not present in iPS (gold class) (**Fig. 4.5F**) and (vi) 282 interactions absent in ES cells, acquired in NPCs and residually connected in iPS cells (blue class) (**Fig. 4.5G**). Noteworthy, we found that the sensitive detection of these interaction classes, particularly those that distinguish iPS from ES cells, was contingent upon the resolution and read depth afforded by the 5C approach (**Figs. 4.5H-I**). Importantly, we note that the majority of high-count pixels were spatially adjacent each other in our 'Distance-corrected Interaction Score' heatmaps and appear to form larger clusters of enriched 3-D contact (**Fig. 4.6K-L, 4.6N, 4.5D-G**). To ensure that our approach was not inflating the number of significant interactions, we clustered adjacent pixels that were similarly classified, resulting in a total of only 1,248 unique interactions across three cellular states in our 5C regions (~7.5 Mb) (**Fig. 4.6M**). Our clustering approach is similar to the methodology employed by Aiden and colleagues for high-resolution Hi-C data [10]. We emphasize two important points regarding the

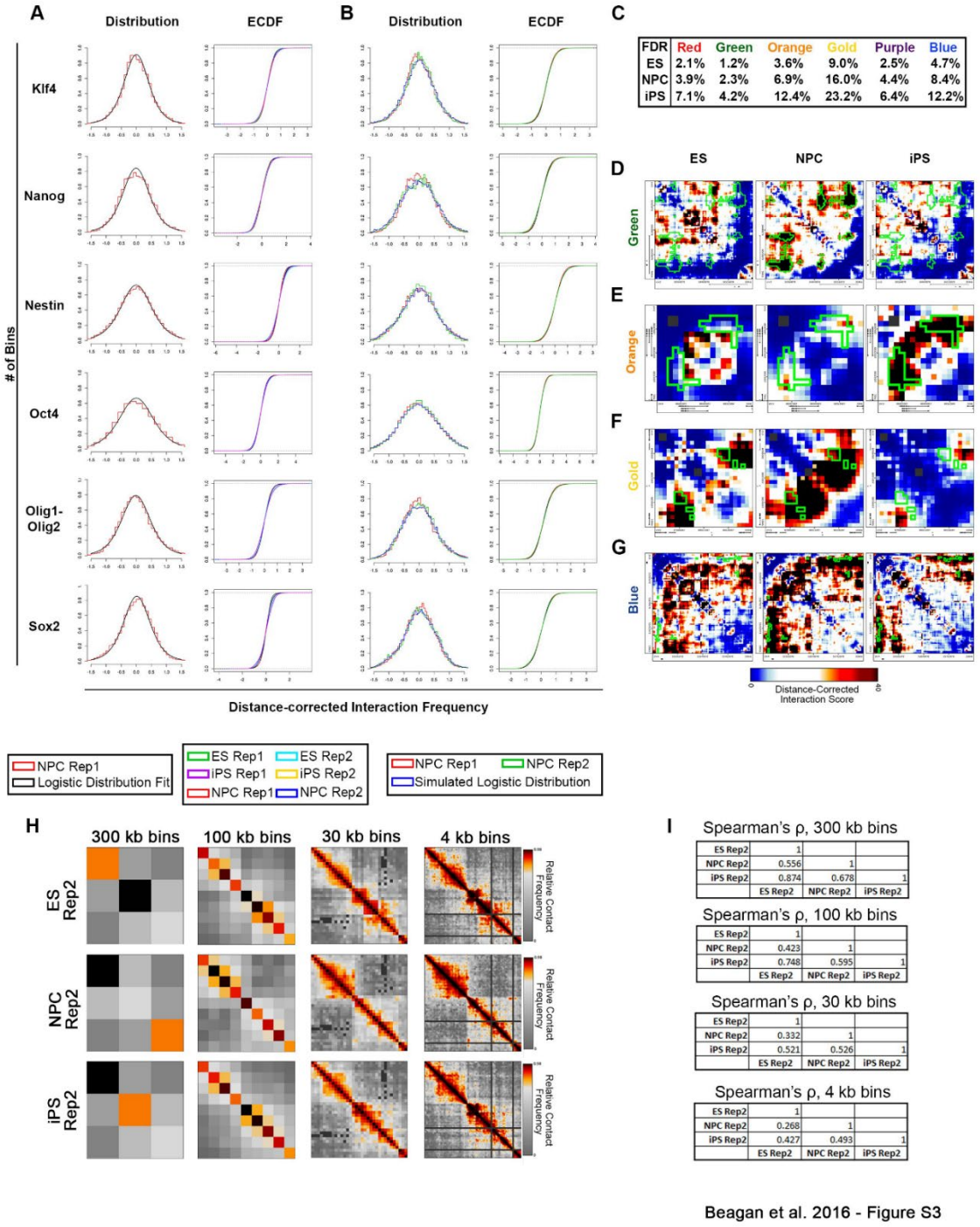


Figure 4.5. Methodology for identification of significant 3-D interaction classes. (A-B) Histograms and empirical cumulative distribution functions (ECDF) of distance-corrected interaction frequency values. **(A)** Distributions of NPC Rep 1 (red) superimposed upon a logistic distribution fit with location/scale parameters computed for each region and biological replicate (black). Juxtaposition of models illustrates that

our distance-corrected data can be modeled with logistic fits. **(B)** Distributions of the two NPC replicates (red and green) plotted alongside the simulated data distribution (blue). Simulated data closely approximate 5C data, supporting their utility in computing empirical False Discovery Rates. **(C)** Empirical false discovery rates computed from simulated data reported for each classification. FDRs vary slightly depending on which cell-type replicates are used to model parameters of the simulations (see **Appendix II Methods**). **(D-G)** Zoomed-in contact density maps for specific **(D)** NPC only interactions (green class), **(E)** iPS only interactions (orange class), **(F)** ES-NPC interactions (yellow class), and **(G)** NPC-iPS interactions (blue class). Classified interaction pixels are outlined in green for each interaction class. **(H)** 5C primer-primer counts data are binned with decreasing bin sizes and displayed as contact density heatmaps. From left to right, heatmaps are shown for bin sizes of 300 kb, 100 kb, 30 kb and finally the 4 kb with a 20 kb smoothing window used in this study. **(I)** Spearman's rank correlation coefficient was calculated using the distance-corrected interaction frequency data of replicates displayed in **(H)** at each bin size.

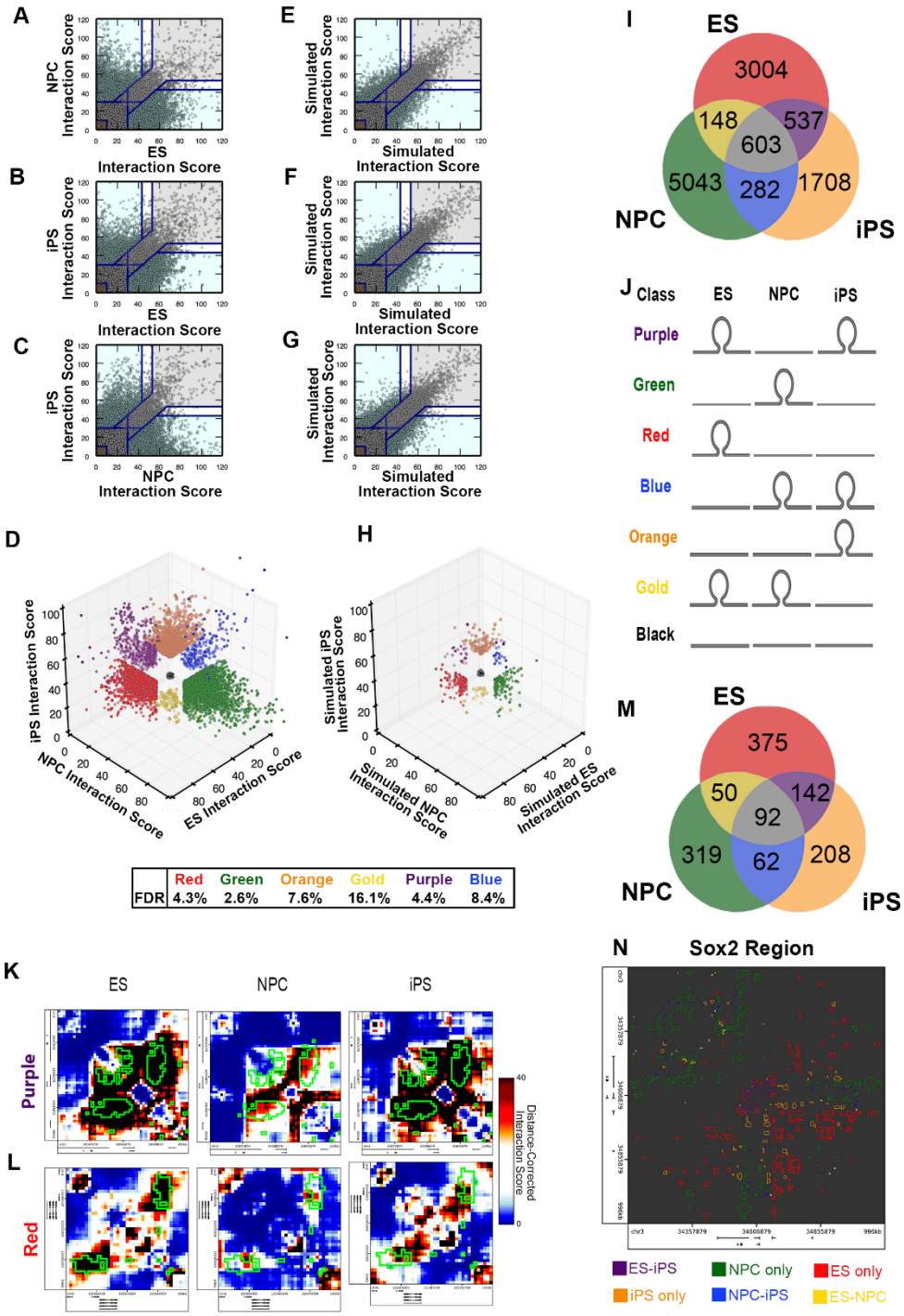


Figure 4.6. Genome architecture can be classified into several distinct dynamic groups during cell fate transitions. (A-C) Scatterplot comparison of distance-corrected interaction scores between **(A)** ES cells and NPCs, **(B)** ES and iPS cells and **(C)** NPCs and iPS cells. Thresholds are displayed as blue lines. For pairwise plots, cell type-specific, invariant and background interactions are represented by blue, grey and brown colored shading, respectively. **(D)** 3D scatterplot of distance-corrected interaction scores for cellular states in which both replicates cross the thresholds displayed in **(A-C)**.

Interaction classes are indicated by color (red, ES only; green, NPC only; orange, iPS only; gold, ES-NPC; purple, ES-iPS; blue, NPC-iPS; black, Background). Empirical false discovery rates computed from simulated data in **(E-G)** are reported for each classification. **(E-G)** Scatterplots of distance-corrected interaction scores from simulated replicates. Empirical false discovery rates were computed based on the number of interactions that cross pre-established thresholds in the simulated data versus the real data. **(H)** 3D scatterplot of distance-corrected interaction scores for simulated libraries that cross the thresholds displayed in **(A-C, E-G)**. **(I)** Number of interactions called significant in each cell-type specific interaction class. **(J)** Schematic illustrating the 3D interaction behavior for each interaction class. **(K-L)** Zoomed-in heatmaps of distance-corrected interaction scores for specific **(K)** ES-iPS (purple class) and **(L)** ES only (red class) interactions. Classified interaction pixels are outlined in green. **(M)** Number of interactions called significant for each 3-D classification after clustering directly adjacent 4 kb bins. **(N)** Depiction of all interactions called as significant in the Sox2 region. Each interaction is outlined by the corresponding classification color.

3-D interaction classes called in this study: (i) the interactions represent both specific looping contacts and subTAD boundaries that are dynamic across three cellular states and (ii) rather than a traditional peak calling approach in just one cell type, we are reporting seven classes of long-range interactions called across three cellular states with a focus on the regions of the genome that are most likely to undergo dynamic restructuring during the reprogramming process. Overall, these results indicate that chromatin architecture is highly dynamic during cell fate transitions, with unique folding classes emerging during the reprogramming process.

4.2.4 Pluripotency genes form interactions that can successfully reprogram

We next set out to explore the biological relevance of our dynamic interaction classes. We utilized a series of integrative computational approaches to elucidate the underlying relationships among: (i) fine-scale chromatin folding, (ii) gene expression, (iii)

histone modifications characteristic of cell type-specific regulatory elements and (iv) binding profiles of the architectural protein CTCF.

We first investigated the interactions that were present in ES cells, lost in NPCs and reconnected during reprogramming (ES-iPS; purple class) (**Fig. 4.7A**). We noticed that the *Sox2* gene formed a strong 3D interaction with a pluripotent enhancer element ~120 kb downstream marked by a large domain of H3K4me1/H3K27ac in ES cells (**Fig. 4.7B**). Upon differentiation, the *Sox2*-pluripotent enhancer interaction disassembled in parallel with loss of H3K27ac signal and then subsequently reassembled in iPS cells (**Fig. 4.7B,C**). We also identified ES-iPS (purple class) interactions between the *Oct4/Pou5f1* gene and a putative enhancer element ~20 kb upstream marked by ES-specific H3K4me1/H3K27ac (**Fig. 4.7D**). As expected given the pluripotent properties of our iPS clone, the *Oct4*-enhancer interaction breaks apart in NPCs and reconnects again in iPS cells (**Fig. 4.7D,E**). We next quantitatively assessed the enrichment of a wide range of genomic elements in the ES-iPS class of successfully reprogrammed 3D interactions. Consistent with previous reports [190, 202-205] and our qualitative observations, pluripotency genes and putative ES-specific enhancers were significantly enriched at the base of ES-iPS interactions (**Fig. 4.7F**). Together, these results indicate that pluripotency genes can form long-range connections with ES-specific enhancer elements and that these interactions can reprogram in iPS cells.

To explore the functional significance of fully reprogrammed interactions, we next conducted genome-wide RNA-seq analysis in ES, NPCs and iPS cells. We examined *Oct4* and *Sox2* gene expression after normalization among libraries to account for any potential

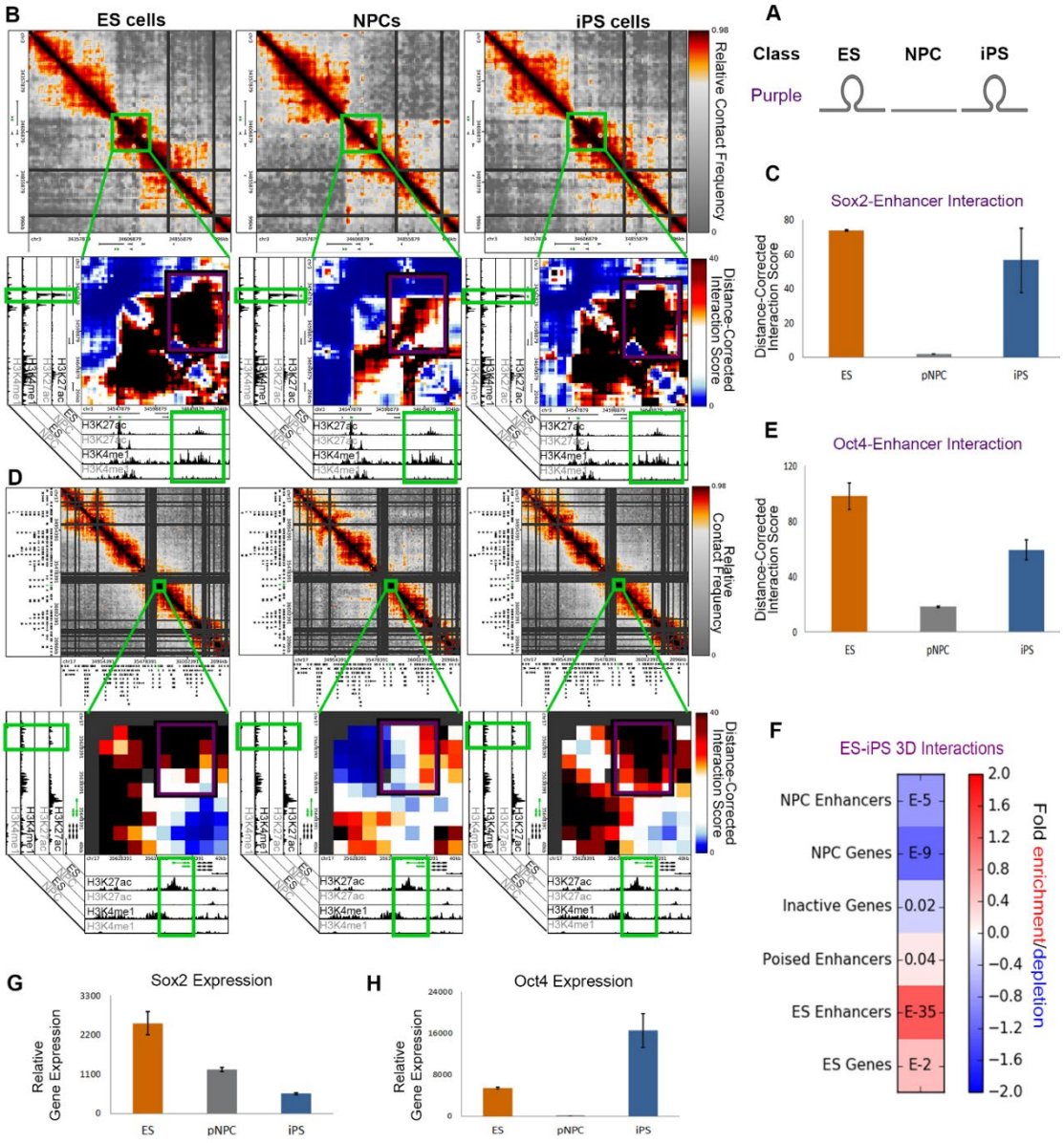


Figure 4.7. Pluripotency gene-enhancer interactions can be re-established in iPS cells. (A) Schematic illustrating the ES-iPS (purple) interaction class. **(B,D)** Relative contact frequency heatmaps (top) and zoomed-in distance-corrected interaction score heatmaps (bottom) highlighting key ES-iPS interactions (purple class) between **(B)** *Sox2* and **(D)** *Oct4* genes and their target enhancers. Heatmaps are overlaid on ChIPseq tracks of H3K27ac and H3K4me1 in ES cells and NPCs. **(C+E)** Distance-corrected interaction score changes at **(C)** the *Sox2*-enhancer interaction and **(E)** *Oct4*-enhancer interaction among ES, NPC and iPS cells. Error bars represent the standard deviation across two 5C replicates. **(F)** Fold enrichment of cell type-specific regulatory elements in ES-iPS (purple class) interactions compared to the enrichment expected by chance across the genome. Color bar represents fold change enrichment over background (blue, depletion;

red, enrichment). P-values are computed with Fisher's Exact test and listed in each bin. **(G-H)** Normalized gene expression is plotted for **(G)** *Sox2* and **(H)** *Oct4* genes. Error bars represent standard deviation across two RNAseq replicates.

batch effects and differences in sequencing depth (**Fig. 4.8A-D**). Unexpectedly, despite reconnection with target pluripotent enhancers, *Sox2* expression was markedly lower than target ES cell expression levels (**Fig. 4.7G**), whereas *Oct4* expression was more than 2-fold higher than target ES cell expression levels (**Fig. 4.7H**). Our observations highlight the importance of further understanding the relationship between genome folding and expression, and led us to question if more global architectural connections around these pluripotent enhancer-promoter interactions could be linked to inaccurately reprogrammed gene expression levels in iPS cells.

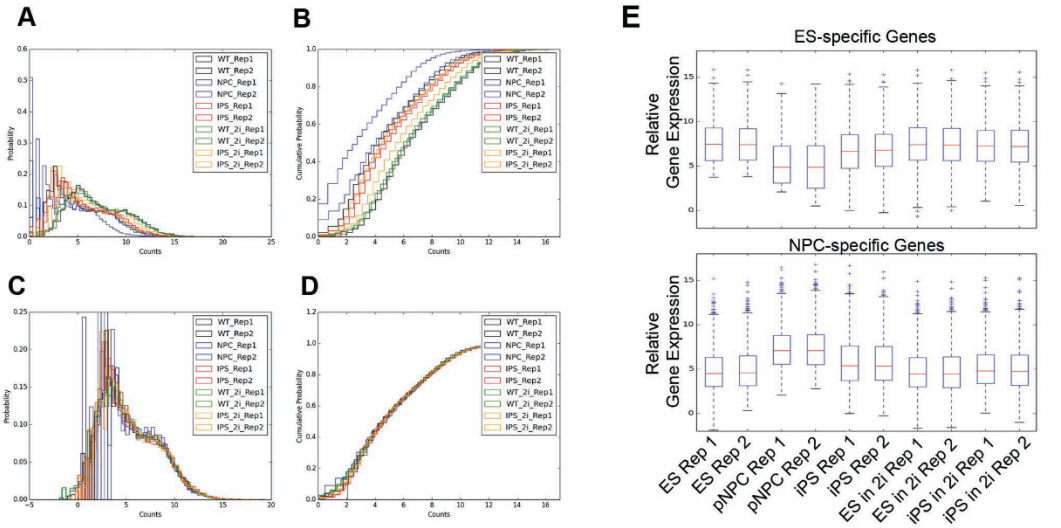


Figure 4.8. RNA-seq library normalization and quality control. (A,C) Frequency histograms of read counts across all genes for each RNA-seq library before **(A)** and after **(C)** normalization**. (B,D)** Cumulative distributions of read counts across all genes for each RNA-seq library before **(B)** and after **(D)** normalization. **(E)** Boxplots of the logged normalized counts of genes parsed as ES-specific or NPC-specific for each replicate.

4.2.5 Some pluripotency genes reconfigure into new NPC interactions that remain persistent in iPS

We next sought to understand larger-scale chromatin folding patterns around *Sox2* (**Fig. 4.9A**). We hypothesized that chromatin architecture dynamics surrounding the short-range enhancer-promoter interaction might impact the incompletely reprogrammed *Sox2* expression in our iPS clone. Unexpectedly, we observed that *Sox2* is also engaged in NPC-iPS (blue class) interactions classified by (i) absence in ES cells, (ii) acquisition in NPCs and (iii) residual tethering in iPS cells (**Fig. 4.9A-B**). In NPCs, the *Sox2*-pluripotent enhancer interaction breaks apart and the gene forms long-range contacts with two distal NPC-specific enhancers marked by NPC-specific H3K27ac/H3K4me1. Intriguingly, although the *Sox2*-pluripotent enhancer interaction is reassembled (purple box), the gene also remains partially tethered to the NPC-specific enhancer in iPS cells (blue box) (**Fig. 4.9A**). We observed a similar phenomenon at the *Klf4* locus, where the *Klf4* gene is highly expressed in ES cells and interacts with a putative ES-specific enhancer element marked by ES-specific H3K4me1/H3K27ac ~75 kb upstream of the gene (**Fig. 4.10A-D**). In NPCs, *Klf4* disconnects from its pluripotent enhancer and engages with a downstream NPC-specific enhancer (**Fig. 4.10E-F**). In iPS cells, *Klf4* retains its interaction with the NPC-specific enhancer (blue box) while also partially re-tethering to its target pluripotent enhancer (purple box) (**Fig. 4.10F**).

We hypothesized that the dual tethering of *Sox2/Klf4* genes to their target ES-specific pluripotent enhancers and their decommissioned NPC-specific enhancers might lead to inaccurate reprogramming of proper expression levels in our iPS clone. As a first step toward testing this hypothesis, we cultured our iPS clone under 2i/LIF conditions to

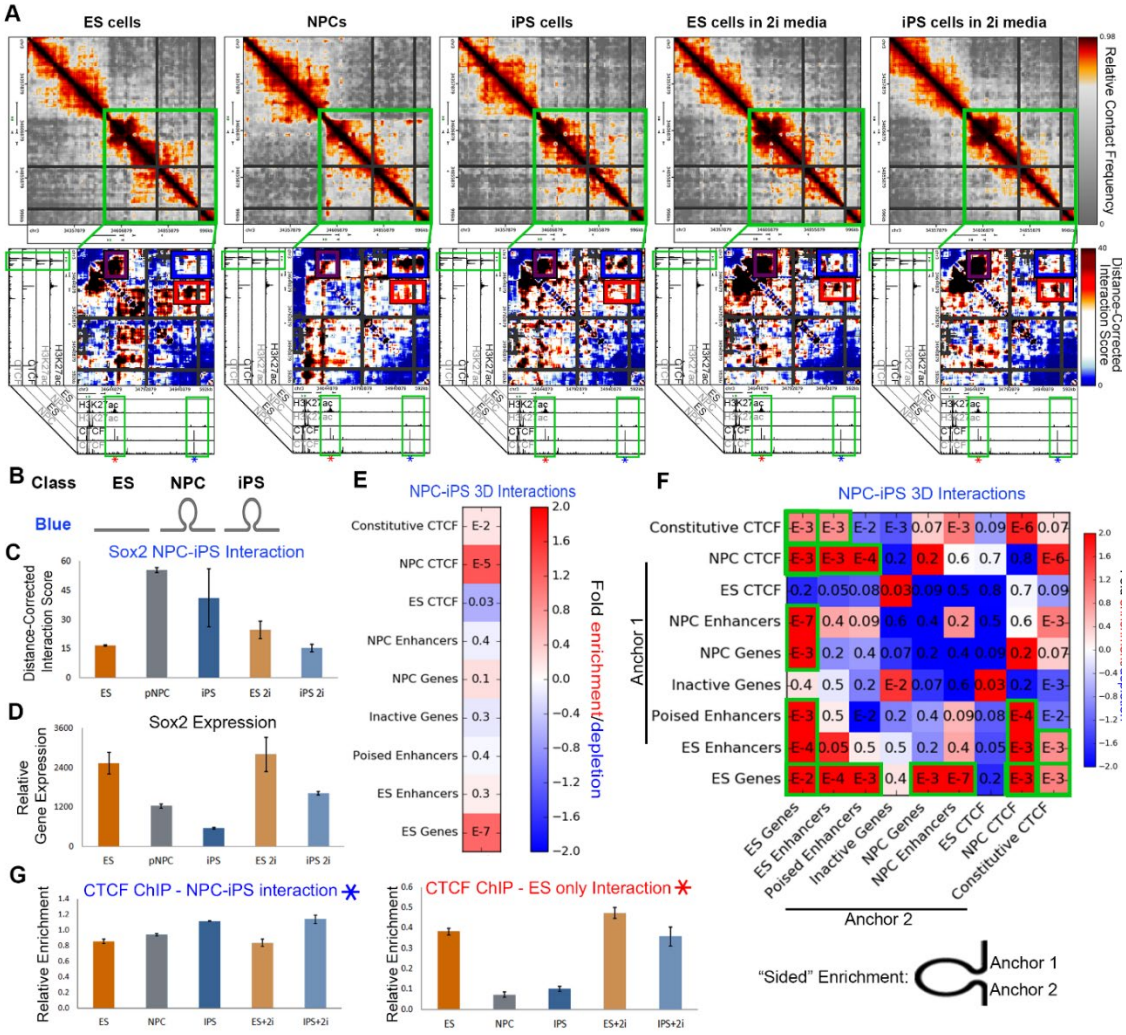


Figure 4.9. Pluripotency genes can exhibit 'persistent-NPC-like' folding patterns in iPS cells. (A) Relative contact frequency heatmaps (top) and zoomed-in distance-corrected interaction score heatmaps (bottom) highlighting an NPC-iPS interaction (blue class) around the *Sox2* gene. Heatmaps are overlaid on ChIPseq tracks of H3K27ac and CTCF in ES cells and NPCs. **(B)** Schematic illustrating the NPC-iPS (blue) interaction class. **(C)** Distance-corrected interaction score changes at an NPC-iPS interaction around the *Sox2* gene among ES, NPC, iPS, ES+2i and iPS+2i conditions. Error bars represent standard deviation across two 5C replicates. **(D)** Normalized expression for the *Sox2* gene. Error bars represent standard deviation across two RNAseq replicates. **(E, F)** Fold enrichment of cell type-specific regulatory elements in NPC-iPS (blue class) interactions compared to the enrichment expected by chance across the genome. P-values are computed with Fisher's Exact test and listed in each bin. **(E)** Enrichment for any given genomic annotation at the base of NPC-iPS interactions. **(F)** Enrichment for any given pairwise combination of genomic annotations in the two anchoring bins at the base of NPC-iPS interactions. **(G)** Relative ChIP-qPCR enrichment of CTCF binding at the

NPC-iPS interaction (left, denoted by blue star in **(A)**) and ES only interaction (right, denoted by red star in **(A)**).

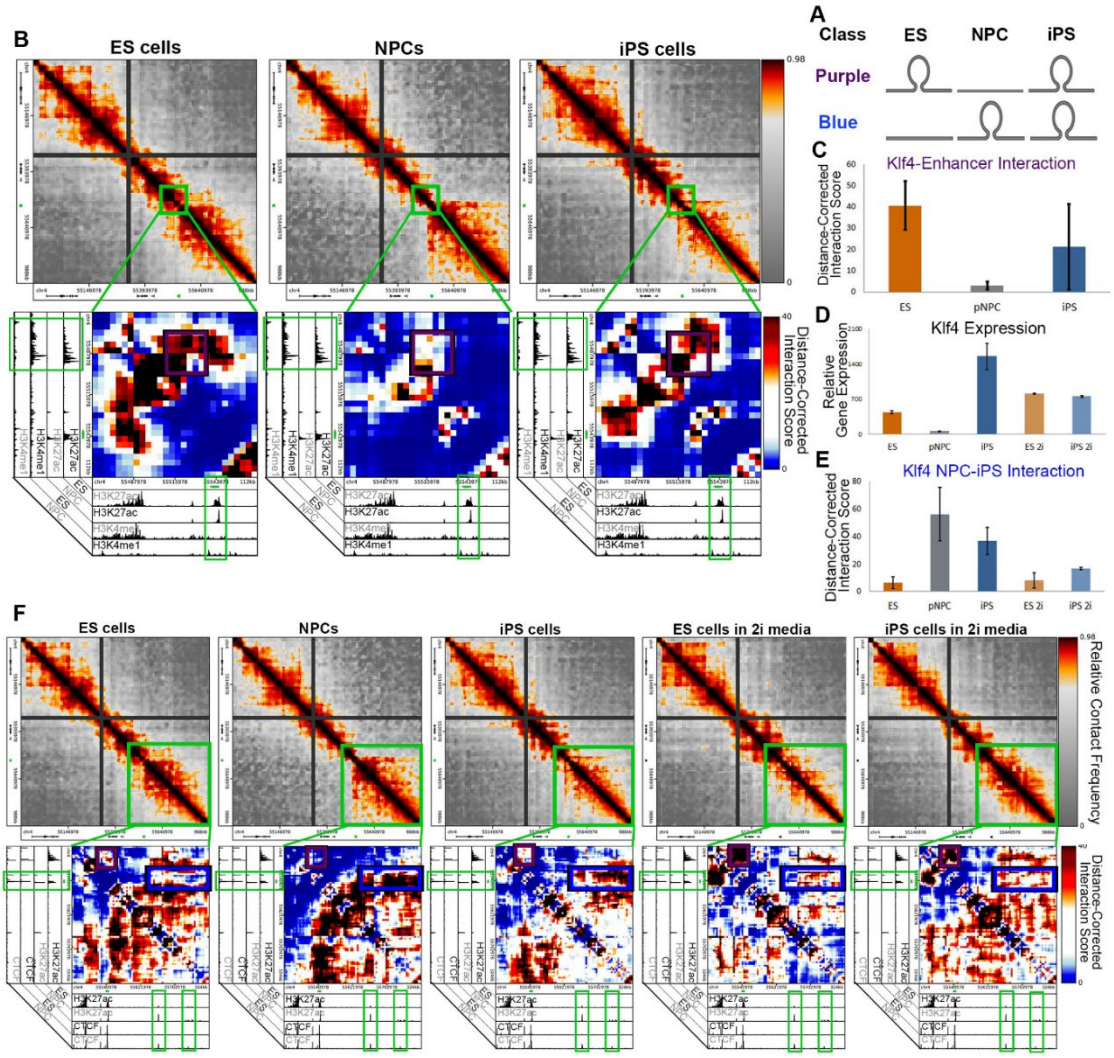


Figure 4.10. The *Klf4* gene engages in both ES-iPS (purple class) and NPC-iPS (blue class) 3-D interactions. (A) Schematic illustrating the ES-iPS (purple) and NPC-iPS (blue) interaction classes. **(B)** Contact frequency heatmaps (top) and zoomed-in heatmaps of distance-corrected interaction scores (bottom) highlighting a key interaction between *Klf4* and an upstream enhancer. Interaction score heatmaps are overlaid on ChIP-seq tracks of H3K27ac and H3K4me1 in ES cells and NPCs. **(C)** Distance-corrected interaction score changes among ES, NPC and iPS cells at the *Klf4*-enhancer ES-iPS (purple class) interaction. Error bars represent standard deviation across two replicates. **(D)** Normalized gene expression for the *Klf4* gene is plotted for ES, NPC and iPS cells, as well as ES and IPS cells cultured in 2i media. Error bars represent standard deviation across two replicates. **(E)** Distance-corrected interaction score changes at an NPC-iPS interaction around the *Klf4* gene among ES, NPC and iPS cells. Error bars

represent standard deviation across two replicates. **(F)** Contact frequency heatmaps (top) and zoomed-in heatmaps of distance-corrected interaction scores (bottom) highlighting the NPC-iPS interaction between the *Klf4* gene and a downstream NPC-specific enhancer. Plotted similar to **(B)**.

promote a naïve, ground state of pluripotency and ensure morphological/phenotypic uniformity across the population [170, 207]. Strikingly, we noticed that 2i/LIF culture of iPS cells resulted in (i) loss of the *Sox2*- or *Klf4*-NPC enhancer (blue class) interactions, (ii) a further amplification in strength of the *Sox2*- or *Klf4*-pluripotent enhancer (purple class) interactions and (iii) a fine-tuning of *Sox2* or *Klf4* expression to ES levels (**Fig. 4.9A, 4.9C-D, 4.10E+F**)**.** These results indicate that 2i/LIF conditions are capable of untethering persistent somatic cell chromatin architecture in a population of iPS cells and restoring inaccurately reprogrammed gene expression to levels equivalent to those found in V6.5 ES cells. Future causative studies will be necessary to further dissect the link among architectural persistence, naïve vs. primed pluripotency and precise gene expression levels during reprogramming.

We then set out to further understand the mechanistic basis of NPC-iPS (blue class) interactions. Quantitative enrichment analysis revealed three key genomic annotations enriched at the base of NPC-iPS contacts: (i) ES-specific genes, (ii) NPC-specific CTCF and (iii) constitutive CTCF (**Fig. 4.9E**). We then computed 'sided' enrichments by accounting for the presence/absence of genomic annotations in both anchoring loci at the base of the NPC-iPS interactions (see schematic, **Fig. 4.9F**)**.** Consistent with our qualitative observations, ES-specific genes most significantly contact NPC-specific enhancers when located at the base of NPC-iPS interactions (**Fig. 4.9F**). We note that *Sox2* and *Klf4* are classified as ES-specific genes in our study due to their markedly increased expression in

ES cells vs. NPCs. However, both genes are still expressed at levels at least 8-fold higher than background in NPCs. Together, these results led us to hypothesize that genes with developmental roles in both ES cells and NPCs, but regulated by different enhancers in the two cellular states, might be particularly susceptible to inappropriate tethering to off-lineage enhancers in iPS cells.

Our quantitative enrichment analyses also indicated that ES-specific genes formed significant 3-D connections with NPC-specific and constitutively bound CTCF sites (**Figs. 4.9E-F**). Consistent with this quantitative result, we noticed a constitutively bound CTCF site at the base of the Sox2 NPC-specific enhancer (**Fig. 4.9A**) and an NPC-specific CTCF site at the base of the Klf4 NPC-specific enhancer (**Fig. 4.10F**), suggesting that CTCF might work together with enhancers to facilitate 3-D connections to the correct target gene(s). To understand how CTCF binding might be altered during reprogramming, we performed CTCF ChIP-qPCR across all five of our cellular states. We queried CTCF occupancy levels in the NPC-specific and ES-specific enhancers (**Fig. 4.9A**, blue and red stars, respectively) at the *Sox2* locus. We found that the NPC-specific enhancer remains constitutively bound by CTCF in ES, NPC, iPS, ES+2i and iPS+2i conditions (**Fig. 4.9G**, left). By contrast, the ES-specific enhancer exhibited high CTCF in ES cells, loss of binding in NPCs, sustained low CTCF occupancy in iPS cells and subsequent restoration of occupancy in 2i/LIF (**Fig. 4.9G**, right).

Intriguingly, CTCF binding patterns correlate with the changes in chromatin architecture around Sox2. In ES cells, the constitutive CTCF site interacts with the ES-specific CTCF site, resulting in spatial co-localization of the ES- and NPC-specific enhancers (**Fig. 4.9A**, red box). Loss of CTCF binding at the ES-specific enhancer

correlates with disconnection of the enhancer-enhancer interaction in NPCs. In parallel, the constitutive CTCF site at the NPC-specific enhancer forms a strong NPC-iPS (blue class) interaction with the *Sox2* gene (**Fig. 4.9A**, blue box). We posit that the *Sox2*-NPC-enhancer interaction remains tethered in iPS cells because CTCF does not fully rebind to the ES-specific enhancer (**Fig. 4.9G**, right). In support of this idea, 2i/LIF leads to (i) reacquisition of CTCF binding at the ES-specific enhancer, (ii) reconnection of the interaction between both ES-specific and NPC-specific enhancers and (iii) abrogation of the *Sox2*-NPC-specific enhancer interaction. These observations are consistent with a working model in which 'persistent-NPC' interactions can remain in iPS cells when some developmentally regulated genes are tethered to NPC-specific enhancers, possibly at constitutive or NPC-specific CTCF sites.

We highlight that somatic cell-specific elements were not specifically enriched in NPC-iPS interactions (**Fig. 4.11A-C**). For example, NPC-specific genes and enhancers were primarily enriched in NPC only (green class) interactions, supporting our finding that it is ES-specific genes, particularly those that remain somewhat active in NPCs, that are redirected into NPC-iPS contacts. An example illustrating this idea can be found at the *Olig1/Olig2* genes that are expressed in an NPC-specific manner and equivalently form NPC only (green class) interactions with a downstream NPC-specific enhancer (**Fig. 4.11D-E**). Expression of *Olig1/2* is lost in parallel with loss of the green class 3-D interaction. Together, these results support the intriguing possibility that ES-specific genes that remain partially active in NPCs form new interactions with somatic cell-specific

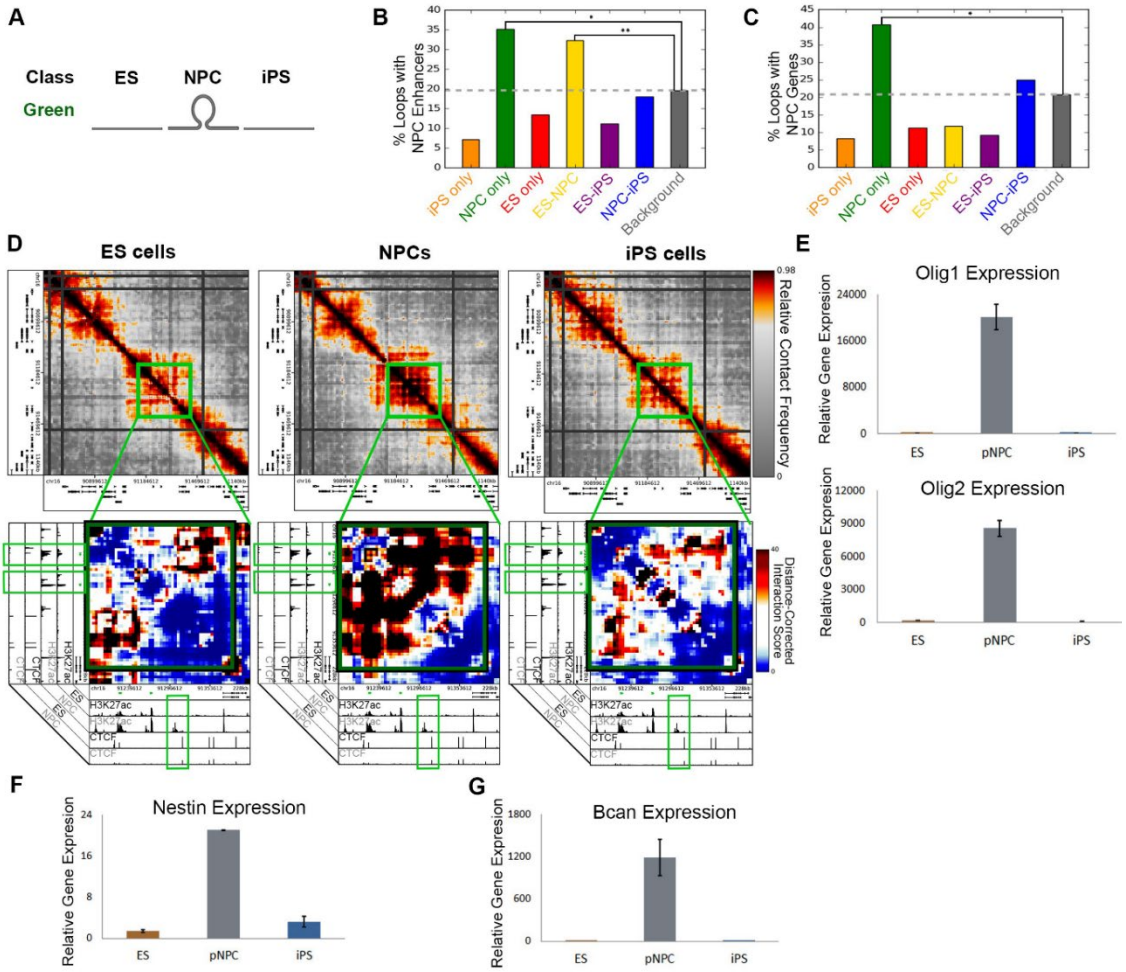


Figure 4.11. NPC-specific genes and enhancers are enriched in NPC only (green class) interactions. (A) Schematic illustrating the NPC only (green) interaction class. **(B)** Bar plot displaying the fraction of each looping class containing NPC-specific enhancers compared to the expected background fraction. Fisher's Exact test: *, P= 3.55182e-58; **, P= 0.00063607. **(C)** Bar plot displaying the fraction of each looping class containing NPC-specific genes compared to the expected background fraction. Fisher's Exact test: *, P= 1.20143e-86. **(D)** Zoomed-in heatmaps of distance-corrected interaction scores highlighting key interactions between the *Olig1* and *Olig2* genes and nearby NPC-active enhancers. Distance-corrected interaction score heatmaps are overlaid on ChIP-seq tracks of H3K27ac and CTCF in ES cells and NPCs. **(E-G)** Normalized gene expression for the *Olig1* and *Olig2* **(E)**, *Nestin* **(F)** and *Bcan* **(G)** genes are plotted for ES, NPC and iPS cells. Error bars represent standard deviation across two replicates.

enhancers during differentiation and that these contacts can remain tethered as a form of architectural persistence in iPS cells. Noteworthy, because 5C is performed on a population of millions of cells, we cannot distinguish between the possibilities that (i) pluripotency genes simultaneously form both ES-iPS and NPC-iPS contacts in individual cells or (ii) pluripotency genes form two different sets of interactions in distinct ES-like subpopulations.

4.2.6 Pluripotent interactions that do not reprogram display dynamic CTCF occupancy

Finally, we explored the interactions that are present in ES cells and lost in NPCs, but do not reconnect in iPS cells (red group, **Figs. 4.12A-B, 4.13A-B**). A notable illustration of these 'poorly reprogrammed' interactions is found at the *Zfp462* gene (highlighted in green, **Fig. 4.12A**), which interacts with a downstream putative ES-specific enhancer element in ES cells. *Zfp462* expression is reduced in NPCs in parallel with loss of H3K27ac at the putative downstream enhancer and loss of the interaction. By contrast to the previously discussed ES-iPS (purple) group, this gene-enhancer interaction is not reassembled in iPS. Similarly, the genes *Mis18a* and *Urb1* form interactions in ES cells that are not reprogrammed (highlighted in yellow and green, respectively; **Fig. 4.13A**). Together, these genomic loci reveal a class of interactions that are refractory to reprogramming in iPS cells.

To investigate the mechanistic basis for poorly reprogrammed (red class) interactions, we again looked for possible dynamic CTCF binding. We noticed that genomic loci where CTCF is bound in ES cells, but severely depleted in NPCs, were preferentially located at the base of poorly reprogrammed interactions (green boxes; **Figs.**

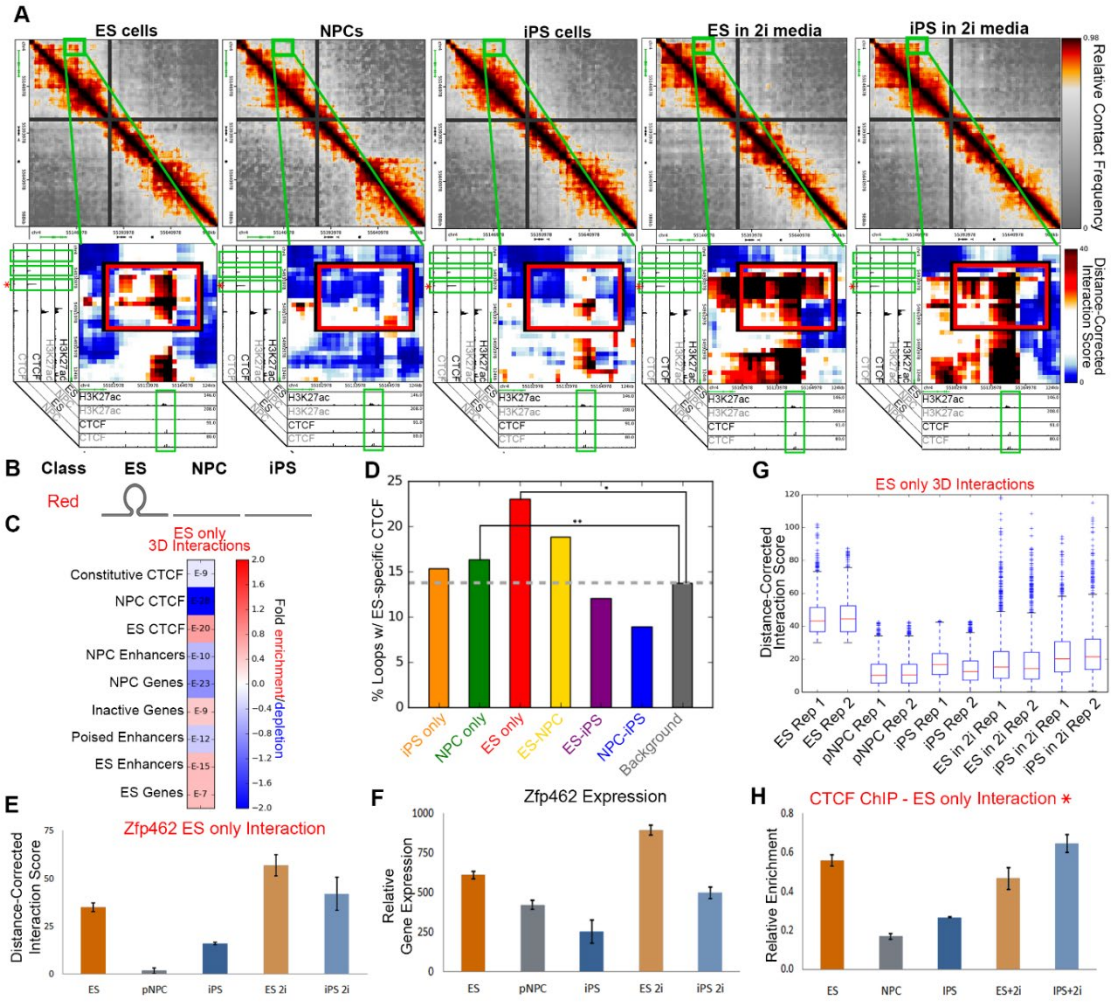


Figure 4.12. Interactions that do not reprogram display poorly reprogrammed CTCF occupancy. (A) Relative contact frequency heatmaps (top) and zoomed-in distance-corrected interaction score heatmaps (bottom) highlighting an ES only (red class) interaction at ES-specific CTCF binding sites at the *Zfp462* gene (indicated in green). Heatmaps are overlaid on ChIPseq tracks of H3K27ac and CTCF in ES cells and NPCs. **(B)** Schematic illustrating the ES only (red class) interactions. **(C)** Fraction of ES only (red class) interactions enriched with distinct cell type-specific regulatory elements compared to the expected enrichment in background. P-values are computed with Fisher's Exact test and listed in each bin. **(D)** Bar plot displaying the fraction of each interaction class containing ES-specific CTCF binding sites compared to the expected background fraction. Fisher's Exact test: *, P= 2.06016e-21; **, P= 0.000541696. **(E)** Distance-corrected interaction score changes at an ES only interaction around the Zfp462 gene among ES, NPC, iPS, ES+2i and iPS+2i conditions. Error bars represent standard deviation across two 5C replicates. **(F)** Zfp462 gene expression among ES, NPC, iPS, ES+2i and iPS+2i conditions. Error bars represent standard deviation across two RNAseq replicates. **(G)** Aggregate distance-corrected interaction score changes among

ES, NPC, iPS, ES+2i and iPS+2i conditions for genes anchoring red class. **(H)** Relative ChIP-qPCR enrichment of CTCF binding at the ES only interaction (denoted by blue star in **(A)**).

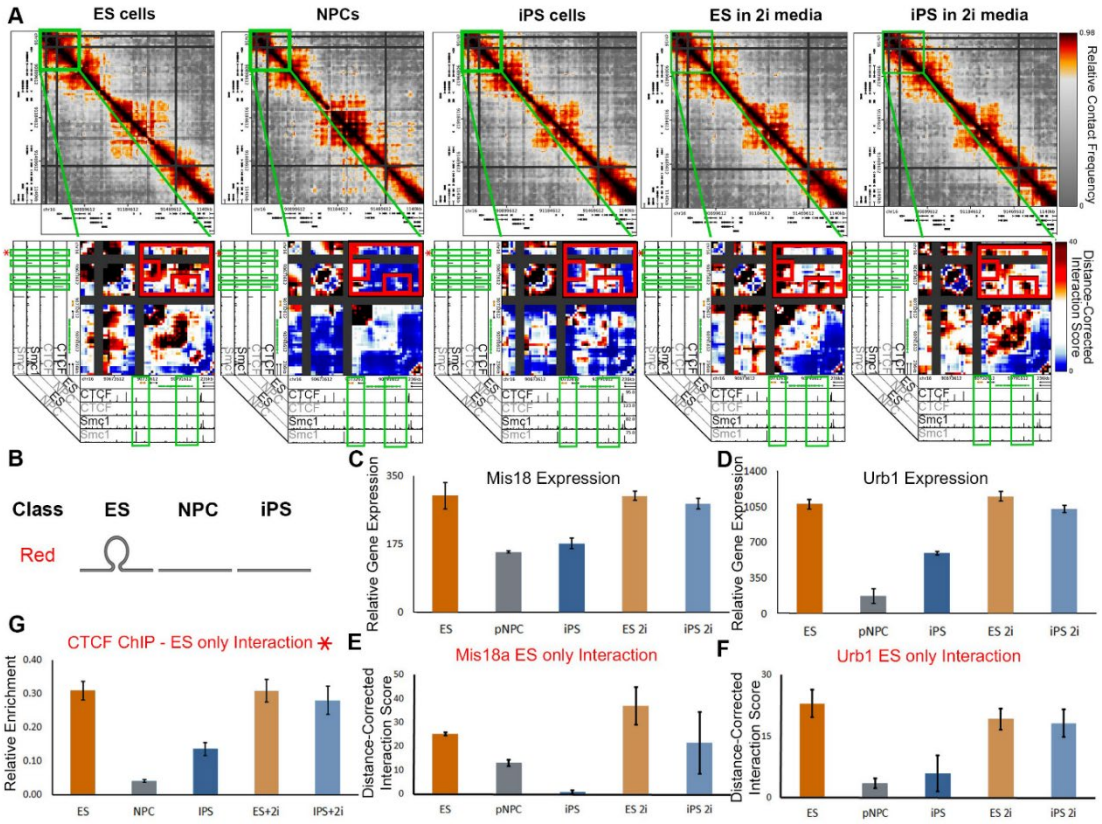


Figure 4.13. The *Mis18* and *Urb1* genes engage in ES only (red class) 3-D interactions linked to inaccurately reprogrammed, ES-specific CTCF binding. (A) Contact frequency heatmaps (top) and zoomed-in heatmaps of distance-corrected interaction scores (bottom) highlighting ES only interactions surrounding the *Mis18a* and *Urb1* genes. Interaction score heatmaps are overlaid on ChIP-seq tracks of CTCF and Smc1 in ES cells and NPCs. **(B)** Schematic illustrating the ES only (red) class of looping interactions. **(C-D)** Normalized gene expression for the *Mis18a* **(C)** and *Urb1* **(D)** genes are plotted for ES, NPC, iPS cells and ES/iPS cells cultured in 2i media. Error bars represent standard deviation across two replicates. **(E-F)** Distance-corrected interaction score changes at *Mis18a* **(E)** and *Urb1* **(F)** ES-only interactions highlighted on heatmaps with small red boxes in **(A)**. Error bars represent standard deviation across two replicates. **(G)** Relative ChIP-qPCR enrichment of CTCF binding at the ES only interaction displayed in **(A).** CTCF site queried is denoted by red star in **(A).** Error bars represent SD across three technical replicates.

4.12A, 4.13A). Consistent with this observation, ES-specific CTCF sites were significantly enriched in ES only (red class) interactions (**Fig. 4.12C+D**). ChIP-qPCR analysis of CTCF occupancy revealed consistent depletion of CTCF in our iPS clone compared to ES cells (**Fig. 4.9G, 4.12H, 4.13G**). Importantly, culture of our iPS clone in 2i/LIF media resulted in (i) reacquisition of the red group interactions, (ii) re-establishment of CTCF occupancy and (iii) restoration of gene expression levels in iPS (**Figs. 4.12E-H, 4.13C-G**). Corroborating locus-specific observations, a global analysis of red class interactions demonstrated a marked increase in interaction score upon addition of 2i/LIF media to iPS cells (**Fig. 4.12G**). On the basis of these results, we posit that the loss of CTCF binding at critical developmentally regulated loci can be inefficiently restored during a cell-fate transition like somatic cell reprogramming.

4.2.7 Somatic elements are disconnected and pluripotent genes hyperconnected in our iPS clone

We hypothesized that distinct types of regulatory elements exhibit differential connectivity patterns as ES cells transition to NPCs and back to iPS cells. To address this hypothesis, we computed a 'connectivity' metric for each class of genomic element in each of the three cellular states. ES-specific enhancers lose their connectivity in NPCs and then reconnect in iPS cells (**Fig. 4.14A**). Intriguingly, ES-specific genes become increasingly more connected upon differentiation and subsequent reprogramming (**Fig. 4.14B**). By contrast, NPC-specific genes/enhancers increase connectivity in NPCs, but then resume

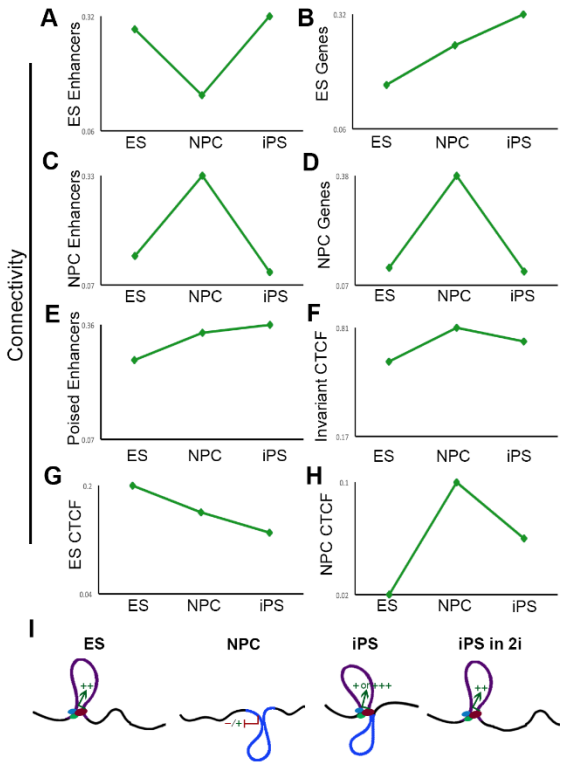


Figure 4.14. Pluripotency genes can be hyperconnected in iPS cells. Connectivity of distinct regulatory elements in ES cells, ES-derived NPCs and NPC-derived iPS cells. **(A)** ES-specific enhancers; **(B)** ES-specific genes; **(C)** NPC-specific enhancers; **(D)** NPC-specific genes; **(E)** Poised enhancers; **(F)** Invariant CTCF; **(G)** ES-specific CTCF; **(H)** NPC-specific CTCF. **(I)** Schematic illustrating a model of the 'hyper-connectivity' of certain pluripotency genes in our NPC-derived iPS clone. Key ES-specific genes (denoted by colored arrows) display the ability to reprogram their connections with ES-specific enhancers (denoted by green/blue 'transcription factor' binding sites) and retain remnants of their somatic connections. This intermediate architectural state correlates with inaccurate reprogramming of gene expression levels (represented by colored +/-) and can be fully restored upon culture in 2i/LIF media.

ground state ES-like connectivity in iPS (**Fig. 4.14C-D**). Poised enhancers and invariant CTCF sites display minor differences in connectivity across the three cellular states (**Figs. 4.14E+F**), whereas ES-specific CTCF sites lose their interactions upon differentiation and only partially gain back connectivity in iPS (**Fig. 4.14G**). NPC-specific CTCF sites

increase in connectivity in NPCs and then partially resume their disconnected state in iPS cells (**Fig. 4.14H**).

Overall, our results support a model in which somatic cell regulatory elements reconfigure to a ground connectivity state during reprogramming, whereas pluripotency genes (particularly those that retain a low level of activity in NPCs) can be 'hyperconnected' in our iPS clone due to persistent cell-of-origin interactions (**Fig. 4.14**). We hypothesize that 'persistent-NPC' and 'poorly reprogrammed' interactions contribute to inaccurate reprogramming of gene expression levels. Consistent with this idea, 2i/LIF can erase 'persistent-NPC' interactions, restore 'poorly reprogrammed' interactions and re-establish precise ES-like expression levels in our iPS clone.

4.3 Discussion

Understanding the molecular mechanism(s) governing somatic cell reprogramming is of paramount importance to our knowledge of cell fate commitment and the use of iPS cells for regenerative medicine applications. Mechanistic studies have primarily focused on profiling gene expression and classic epigenetic modifications at intermediate stages in the reprogramming process [173, 193, 194, 197]. However, the molecular roadblocks that impede the efficiency and timing of epigenome resetting in iPS cells are just beginning to emerge. Here we examine a unique aspect of reprogramming: the higher-order folding of chromatin in the 3D nucleus. We demonstrate that iPS genome architecture at the sub-Mb scale within TADs can be imperfectly rewired during transcription factor-mediated reprogramming.

Recent studies focusing on a single locus (e.g. *Nanog*, *Oct4)* reported that pluripotency genes can re-establish long-range connections with their target enhancers in iPS cells [190, 202-205]. Motivated by the need to understand how chromatin unfolds/refolds more generally in iPS, we created high-resolution maps of chromatin architecture in Mb-sized regions around developmentally regulated genes. Consistent with previous reports, we observe that many pluripotency genes interact with ES-specific enhancers in ES cells; these interactions break apart in NPCs and then reassemble in iPS cells. Additionally, we find that somatic cell interactions between NPC-specific genes and NPC-specific enhancers generally disconnect in iPS cells. Thus, our data confirm and extend several known locus-specific principles of genome folding during reprogramming.

We also uncover new classes of chromatin interactions that do not behave in the expected manner. We identified a small subset of NPC-iPS (blue class) interactions representing persistent chromatin folding patterns from the somatic cell of origin in iPS cells. Unexpectedly, we find that some key pluripotency genes can form new 3-D connections in NPCs that remain tethered in our iPS clone. For example, *Klf4* and *Sox2* are dually tethered to their target ES-specific enhancers and their decommissioned NPC-specific enhancers in iPS cells. We posit that this rare, but intriguing form of 'architectural persistence' might be causally linked to inaccurate reprogramming of target gene expression levels in certain iPS clones. In support of this working model, we find that 2i/LIF conditions are capable of untethering persistent somatic cell chromatin architecture and restoring the inaccurately reprogrammed expression to levels equivalent to those found in a genetically comparable ES cell line. Noteworthy, NPC-specific genes/enhancers form contacts in NPCs that subsequently disassemble in iPS, suggesting that somatic genes are

not driving the architectural persistence in iPS cells. These results agree with previous studies suggesting that somatic cell gene expression is downregulated during the initiation phase of reprogramming and precedes the re-activation of the pluripotency network [197]. We favor a model in which reconfiguration of higher-order chromatin topology could be a potential rate-limiting step in the reprogramming process and that architectural persistence or incomplete architectural reprogramming (discussed below) can block the formation of fully reprogrammed iPS cells [195, 208].

CTCF is a key player in the organization of the 3D genome and anchors the base of a large number of long-range interactions in ES cells [7, 9, 10, 39, 209]Here we provide a new link between CTCF and reprogramming. We identify a new class of chromatin interactions that are high in ES cells, break apart in NPCs and are not fully reconfigured in iPS cells. Importantly, we find that these 'poorly reprogrammed' interactions often contain ES-specific CTCF binding sites that lose occupancy in NPCs and do not re-acquire full binding in our iPS clone. CTCF has largely stable occupancy patterns during development, with 60-90% of sites remaining bound to the genome between cell types [73]. Thus, we speculate a model in which CTCF binding is difficult to lose during differentiation, but once occupancy is abolished it is inefficiently re-established during reprogramming. Importantly, DNA methylation is refractory to CTCF binding [210], suggesting a possible link between poorly reprogrammed chromatin contacts and previously reported sources of cell of origin epigenetic persistence [198, 199]. Indeed, because ES cells cultured in 2i/LIF display global hypomethylation [175, 176], we speculate that the interplay between CTCF and dynamic DNA methylation might serve as a mechanism underlying our observation that 2i/LIF media can fully restore CTCF occupancy and 'poorly reprogrammed' interactions.

Epigenetic and transcriptional signatures are generally reset in fully reprogrammed iPS cells cultured under optimal conditions [201, 211, 212]. However, variations in epigenetic profiles among iPS clones have been attributed to reprogramming method, passage number, genetic background or lab-to-lab procedural discrepancies [199, 200]. Therefore, we sought to confirm that our observations were truly linked to inefficiencies in the reprogramming of our iPS clone, and not experimental artifacts due to (i) residual somatic cells in our iPS population or (ii) lab-specific culture conditions. Importantly, Hochedlinger and colleagues have extensively characterized the iPS clone used in this manuscript for its pluripotent properties [206]. Additionally, our iPS clone was cultured to > 15 passages in serum+LIF-containing growth conditions not amenable to NPC proliferation/survival. Finally, known NPC markers are not upregulated in our iPS population vs. ES cells (**Fig. S6E-G**). Thus, we see no evidence of contaminating NPCs in our iPS cells. Although somatic cells are absent, we cannot rule out the possibility that there could be a gradient of pluripotent properties (e.g. a continuum between naïve and primed pluripotency) across single cells within our fully reprogrammed iPS clonal population. Because we are conducting population-based assays, we would detect all interactions that exist across the different pluripotent states. Consistent with this possibility, we see that conversion of the population to a uniform, naïve pluripotent state with 2i/LIF media abrogates "architectural persistence" interactions and re-instates "poorly reprogrammed" interactions. Additionally, although we subjected our iPS cells with or without 2i/LIF to the same number of passages (p > 15), we cannot rule out the possibility that further long-term passaging might also resolve any mis-wired chromatin interactions. Noteworthy, these results raise the interesting possibility that an iPS clone capable of creating transgenic mice

might still exhibit some level of architectural heterogeneity that can be fully resolved with 2i/LIF media. Exciting lines of future inquiry will query genome folding in higher passages, alternative reprogramming conditions, tetraploid-complementation verified iPS cells and a range of iPS clones derived from multiple somatic cell lineages.

While Beagan et al. was under review, de Laat, Graf and colleagues published a genome-wide analysis of chromatin architecture in iPS cells derived from four independent somatic cell lineages [213]. The authors take a top-down approach in which they generate genome-wide, albeit low resolution, Hi-C maps suited to query higher-order levels of genome organization (i.e. A/B compartments, TADs, nuclear positioning of TADs). Importantly, they demonstrate that A/B compartments are largely reset during reprogramming. Moreover, consistent with the leading idea that TADs are largely invariant among cell types [7], TAD boundaries remained for the most part consistent among iPS clones and ES cells. At the level of sub-Mb scale genome folding, however, the design of the two studies is such that different findings arise. Here we take a bottom-up approach in which we create high-resolution, high-complexity maps focused on fine-scale chromatin folding dynamics within TADs around developmentally regulated genes. Given the sensitivity and statistical power afforded by the 5C assay, it is not surprising that we detect a larger number of dynamic looping interactions and subTAD boundaries than reported in Krijger et al. during the transition among ES, iPS and NPC cellular states. Noteworthy, when we increase our bin size from 4 kb up to 300 kb (**Fig. S3H**), we can recapitulate the author's high level of correlation between the ES and iPS cells (**Fig. S3I**). Krijger et al. and Beagan et al. offer complementary viewpoints into genome architecture dynamics across a wide range of length scales and resolutions during reprogramming. Together, the findings

from these studies are consistent with our working hypothesis that architectural changes causally linked to developmentally relevant alterations in gene expression occur within TADs at the sub-Mb scale.

Overall, we present high-coverage, fine scale maps of chromatin folding within TADs in iPS cells and use our maps to uncover several new organizing principles for genome folding during reprogramming. We find that different cell type-specific regulatory elements exhibit contrasting 3-D connectivity patterns as cells switch fates in forward and reverse directions. A deeper understanding of the role for chromatin folding at each step in the reprogramming process is of critical importance toward the use of iPS cells for disease modeling and regenerative medicine purposes. Future work combining high- and low-resolution mapping approaches will provide a comprehensive view of genome folding across length scales and cellular states to create a catalogue of "hotspots" of incomplete architectural reprogramming and address whether specific somatic cell types are more or less resistant to topological changes.

4.4 Acknowledgements

Thank you to co-authors Thomas Gilgenast, Jesi Kim, Zachary Plona, Heidi Norton, Gui Hu, Sarah Hsu, Emily Shields, Xiaowen Lyu and Drs. Effie Apostolou, Konrad Hochedlinger, Victor Corces, and Job Dekker for their critical contributions to this chapter. Thank you to all members of the Cremins lab for helpful discussions, in particular Michael Duong for cell culture assistance.

CHAPTER 5: 3-D GENOME FOLDING COMPLEXITY AND KINETICS DISTINGUISH EXPRESSION TIMING OF KEY NEURONAL ACTIVITY RESPONSE GENES

5.1 Introduction

Neurons have the remarkable ability to receive, transmit, and store information via a dynamic synaptic network. Experience-dependent neuronal activity regulates synaptic features such as dendritic outgrowth, maturation, elimination, and synaptic plasticity[27]. Neural activity governs synaptic structure and function via the upregulation of hundreds of activity-dependent genes[214]. Rapid-response IEGs (rIEGs), including *c-fos*[28-32] and *Arc/Arg3.1*[33-35], are expressed on the order of minutes upon neuronal activation and are essential for long-term learning and memory. Secondary response genes (SRGs) are induced on the order of hours and require *de novo* protein synthesis[215, 216]. More recently, a class of activity-induced delayed-response IEGs (dIEGs) with transcription kinetics intermediate between rIEGs and SRGs was reported[36]. Cis-acting enhancers – e.g. Synaptic Activity Responsive Elements (SAREs) – have been identified using epigenetic signatures characteristic of cis-regulatory activity and verified using reporter transgenes[217-221]. However, the precise genomic elements determining the differential temporal expression of each specific rIEG, dIEG, and SRG remain elusive, in part because SAREs are distributed across the genome in introns and non-coding regions and their specific target genes are generally unknown.

Chromosome-Conformation-Capture (3C) techniques have recently been used to reveal that the mammalian genome is folded into a hierarchy of structurally and functionally distinct architectural signatures, including chromosome territories[222, 223], A/B compartments[10, 45, 224], topologically associating domains (TADs)[7, 8], nested subTADs[9, 10], and long-range looping interactions[10]. The highest resolution maps to date have enabled the detection of tens of thousands of loops genome-wide across multiple mammalian cell types[10, 11]. Little is known about 3-D genome dynamics during synaptic plasticity, due in part to the paucity of high-resolution architecture maps across a time course of neural activity. In a cerebellum-dependent motor learning task, *in vivo* cohesin deletion in granule neurons disrupted the tactile startle response, suggesting that cohesin-dependent loops might be required for learning[225]. The authors also observed, using H3K4me3-specific PLAC-seq, that 40 minutes of optical stimulation of granule neurons involved in the tactile startle response resulted in a small number of enhancer-promoter interactions with altered contacts[225]. Given the limited understanding of these processes, there is great need for studies that investigate how activity-dependent enhancers are temporally integrated within the nucleus via long-range loops to regulate gene expression during a wide range of neuronal activity paradigms.

Here, we set out to elucidate the extent to which long-range chromatin loops are altered during short- and long-term changes in neural activity and to analyze the dynamic interplay between the 3-D genome and the linear epigenome during the activity-dependent transcriptional response. We create high-resolution genome folding maps in > 12 Megabases (Mb) around key IEGs, SRGs, and synaptic genes using Chromosome-Conformation-Capture-Carbon-Copy (5C-seq) and a double alternating primer design. The

5C-seq approach enabled us to achieve high complexity, fine-scale architecture maps to explore genome folding dynamics without bias toward a particular chromatin feature across seven acute or chronic time points of neural activity inhibition and activation. We demonstrate that activity-inducible enhancers engage in either pre-existing or de novo loops connected to genes that exhibit a 1.3- to 24-fold activity-dependent increase in expression, respectively. We observe that H3K27ac signal at distal looped enhancers, but not nearest enhancers, is a strong predictor of activity-dependent target gene expression. Using both 5C and genome-wide Hi-C data, we demonstrate that rapid-response IEGs (rIEGs) *Arc* and *Fos* connect to target enhancers via singular short-range loops that occur de novo upon activation, whereas delayed-response IEGs (dIEGs) and SRGs connect to multiple activity-inducible enhancers via a complex network of pre-existing and *de novo* loops. Due to our multiple, acute time points, we uncovered that *Fos* and *Arc* short-range loops form within 20 minutes post-stimulation, prior to maximum mRNA levels. By contrast, *Bdnf* long-range loops connect on a later time scale of 60-360 minutes, indicating that looping dynamics might be linked to transcription kinetics. We also identify a subclass of pre-existing loops anchored by enhancers decommissioned upon chronic, 24 hours of neural activation. Unexpectedly, we find that common SNVs linked to schizophrenia anchor pre-existing loops connecting activity-decommissioned enhancers to activity-downregulated genes, whereas autism-associated SNVs connect activity-inducible enhancers to upregulated genes. Together, our data links 3D genome architectural complexity to transcriptional kinetics and uncovers distinct architectural motifs associated with neuropsychiatric disorders.

5.2 Results

We first created high-resolution maps of higher-order chromatin architecture after 24 hours of pharmacologically induced low or high activity in primary cultured mouse cortical neurons. We employed an established *in vitro* model system[226] in which murine cortical neurons were cultured for 15 days *in vitro* and then treated for 24 hours with either 10 μM bicuculline (Bic)[227], which increases neuronal firing by blocking GABA (γ-amino butyric acid)-mediated inhibition, or 1 μM tetrodotoxin (TTX)[228], a sodium channel blocker that inhibits neuronal firing (**Fig. 5.1A, Fig. 5.2**). Chronic pharmacological induction of activity results in multiple forms of synaptic plasticity, including homeostatic changes in AMPA-type glutamate neurotransmitter receptor levels at synapses[229]. Our model system allowed us to interrogate the transcriptional, epigenomic, and architectural features of the mammalian genome in non-dividing, terminally differentiated cortical neurons across inactive (TTX-mediated activity inhibition), moderately active (Untreat), and highly active (Bic-mediated increased activity) states.

We used 5C-seq[43] and a double alternating primer design[230] to create high-resolution maps of genome folding in 12.2 Megabases (Mb) surrounding the rIEGs *Arc* and *c-fos*, dIEG/SRG *Bdnf*, synaptic scaffold genes *Neurexin-1* (*Nrxn1*) and *Neuroligin-3* (*Nlgn3*), and the synaptic vesicle gene *Synaptotagmin-1* (*Syt1*) for a total of N=157 unique transcripts (**Fig. 5.1, Fig. 5.3**). Our genome-wide RNA-seq data confirmed that *Arc*, *c-fos*, and *Bdnf* were upregulated ~10-100 fold in Bic vs. TTX conditions, whereas *Nrxn1, Nlgn3,* and *Syt1* were unchanged (**Fig. 5.1B**). As expected, under the Untreat (basal activity) condition we observed an intermediate level of *Arc*, *c-fos*, and *Bdnf* expression between Bic (high activity) and TTX (inactive) conditions (**Fig. 5.2B-C).** To confirm data quality,

we compared the highest resolution Hi-C maps published to date in mouse embryonic stem (ES) cells, neural progenitor cells (NPCs), and *in vitro* differentiated cortical neurons[104] (**Fig. 5.1C, Fig. 5.3A**) to our 5C maps (**Fig. 5.1D, Fig. 5.3B**). Visual inspection confirmed that 5C maps achieved similar library complexity and minimal spatial noise as the gold-standard Hi-C data. 5C maps from our mature primary cortical neurons and published Hi-C maps from ES-derived cortical neurons were highly correlated and exhibited similar loops (**Fig. 5.3**). Consistent with previous reports[16], we observed a marked restructuring of the 3-D genome during the transition from ES cells to NPCs, whereas the global architectural landscape is highly similar between NPCs and neurons (**Fig. 5.1C-D, Fig. 5.3**). Loops and overall contact frequency surrounding synaptic genes Synaptotagmin-1 and Neurexin-1 appeared especially specific to cortical neurons across both HiC and 5C datasets (**Fig. 5.1C, 5.3, 5.4**). We confirmed that our 5C data correlates more strongly with Hi-C from cortical neurons than NPC or ES cell Hi-C data (**Fig. 5.5A**). Moreover, we confirmed high reproducibility of loops across n=4 5C replicates taken across two independent batches of neuronal cultures (**Fig. 5.5B-C, Fig. 5.6**). Thus, we have created high complexity, ultra-high-resolution maps of genome folding across three neuronal activity states.

We next set out to quantify the extent that loops are altered across different activity states. We normalized the intrinsic biases in 5C data, binned maps to 4 kb matrix resolution, and applied our previously published modeling approaches to identify loops with statistically significant interaction frequency above the local distance-dependence and TAD/subTAD background[16, 156, 231, 232] (**Fig. 5.7a, Appendix III Methods**). We formulated a statistical method, 3DeFDR[233], to stratify loops into invariant and cell type-specific

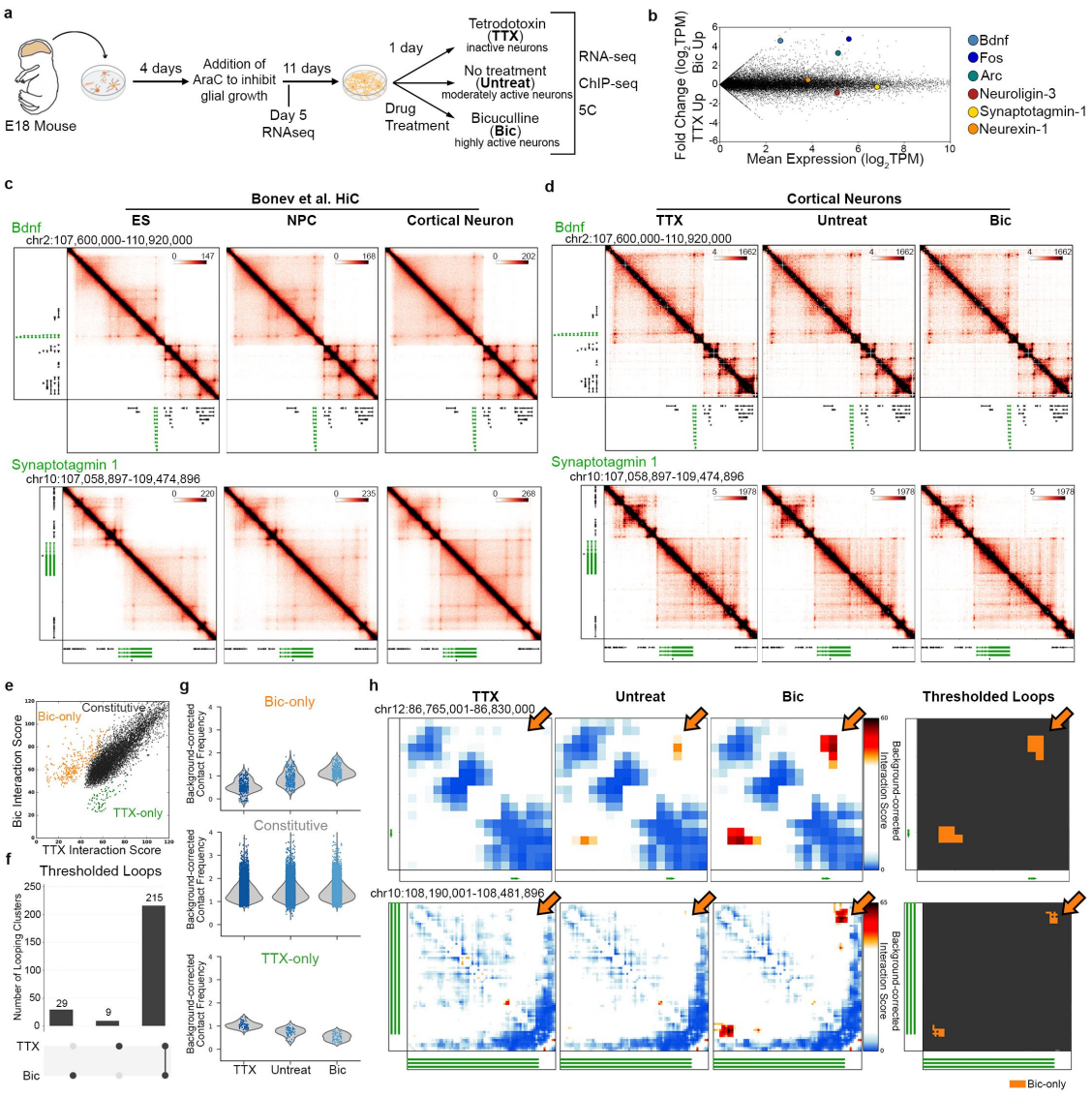


Figure 5.1. Identification of dynamic and invariant looping interactions across neuronal activity states. (A) Primary cultured cortical neuron preparation used to interrogate 3-D genome changes during low, basal or high neuronal activity states. **(B)** RNA-seq data in bicuculline (Bic) and tetrodotoxin (TTX) conditions with selected genes highlighted in colored dots. **(C)** Interaction frequency heatmaps of 1-3 Mb regions surrounding *Bdnf* and *Synaptotagmin-1* genes (labeled in green) across embryonic stem (ES) cells, neural progenitor cells (NPCs), and cortical neurons (CNs) (data analyzed from Bonev et al, 2017). **(D)** Interaction frequency heatmaps of the regions presented in (c) across TTX-treated, untreated, and Bic-treated DIV16 cortical neurons. **(E)** Scatterplot of the interaction scores of thresholded loops in TTX and Bic conditions. **(F)** Activity inhibited (TTX-only), Activity induced (Bic-only), and Activity Invariant (constitutive) loops after thresholding (Appendix III Methods). **(G)** Interaction scores across the TTX,

Untreat, and Bic conditions for each looping class. **(H)** Interaction score heatmaps and thresholded loops demonstrating activity-induced (Bic-only) loops created by *c-Fos* (top) and the *Synaptotagmin-1* TSS (bottom).

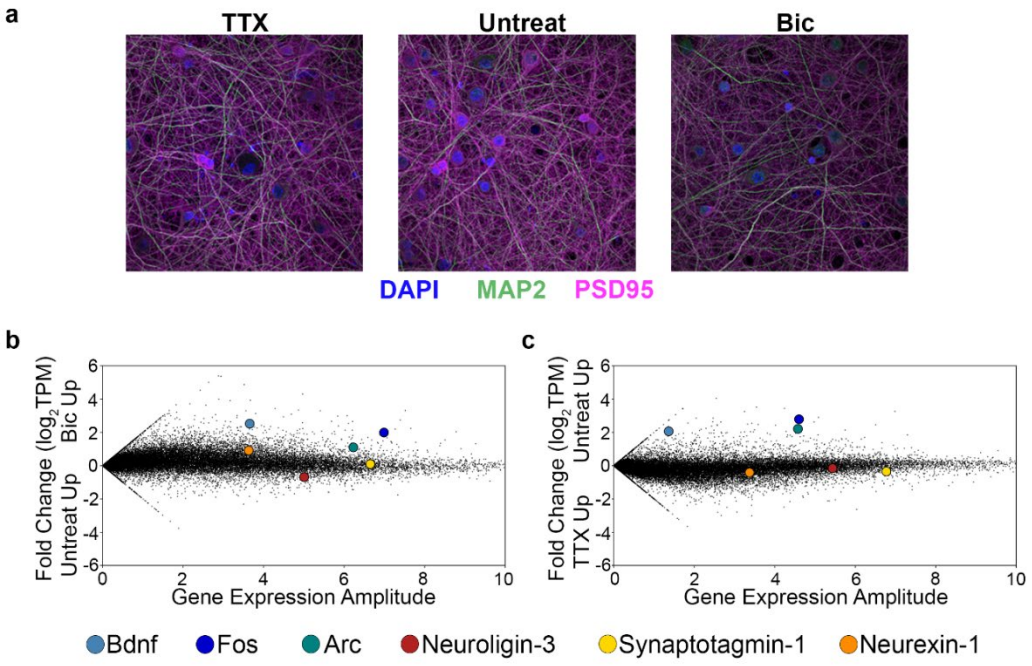


Figure 5.2. Maintenance of neuronal phenotype across neural activity states. (A) Representative immunofluorescence images of DAPI (blue), MAP2 (green), PSD95 (magenta) signal across conditions. **(B-C)** Fold change vs amplitude plots of RNA-seq data comparing the Bic vs Untreat conditions (B) and TTX vs Untreat conditions (C).

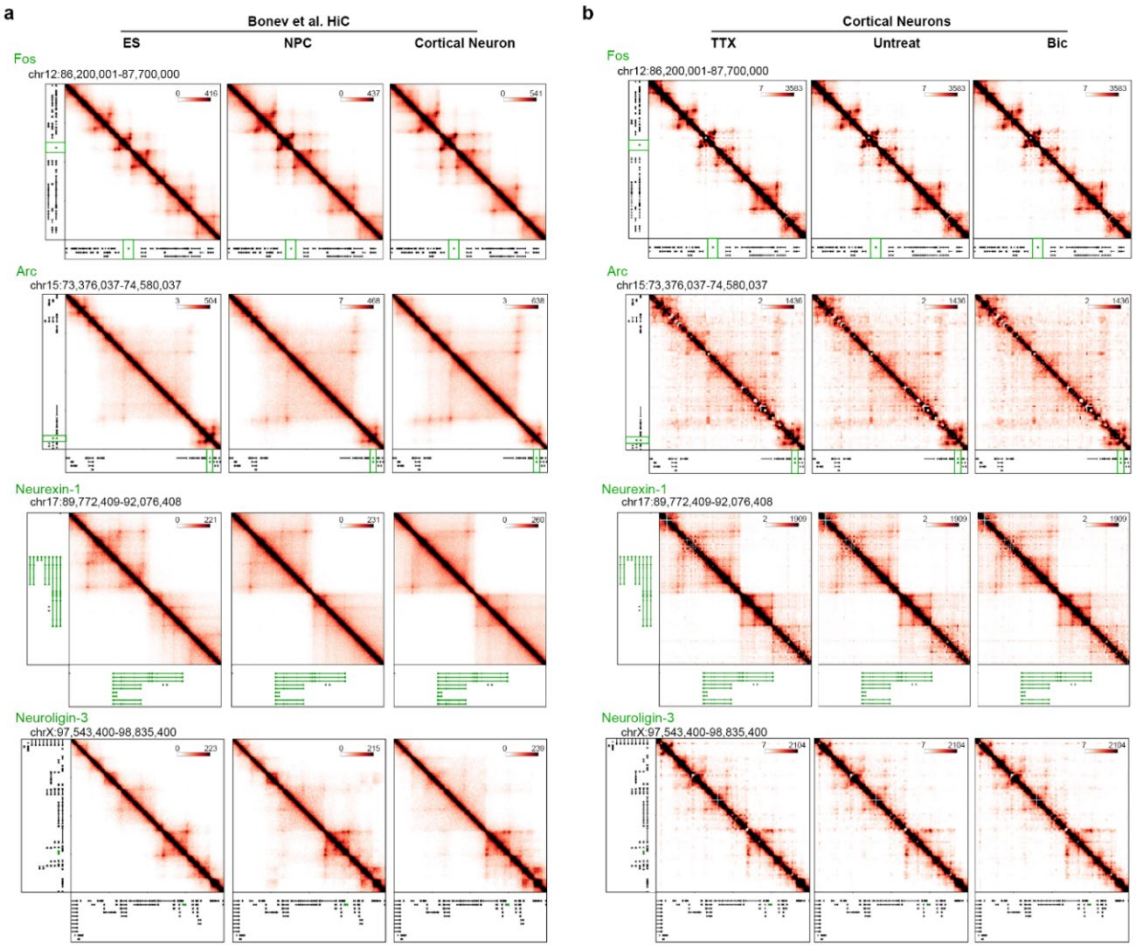


Figure 5.3. Mapping genome folding across neural activity states. (A) Interaction frequency heatmaps of 1-3 Mb regions surrounding the Fos, Arc, Neurexin-1, and Neuroligin-3 genes (labeled in green) across embryonic stem (ES) cells, neural progenitor cells (NPCs), and cortical neurons (CNs) (data analyzed from Bonev+ 2017). **(B)** Interaction frequency heatmaps of the regions presented in (A) across tetrodotoxin-treated (TTX), untreated, and bicuculline-treated (Bic) DIV16 cortical neurons.

classes by thresholding on differences in modeled interaction strength across inactive and highly active neurons (**Fig. 5.1E, Appendix III Methods**). Thresholds were iteratively adjusted to a target empirical false discovery rate computed between real and simulated 5C maps, resulting in the sensitive detection of 215 activity-invariant, 29 activity-induced, and 9 activity-decommissioned interactions within the 12.2 Mb of the genome queried (**Fig.**

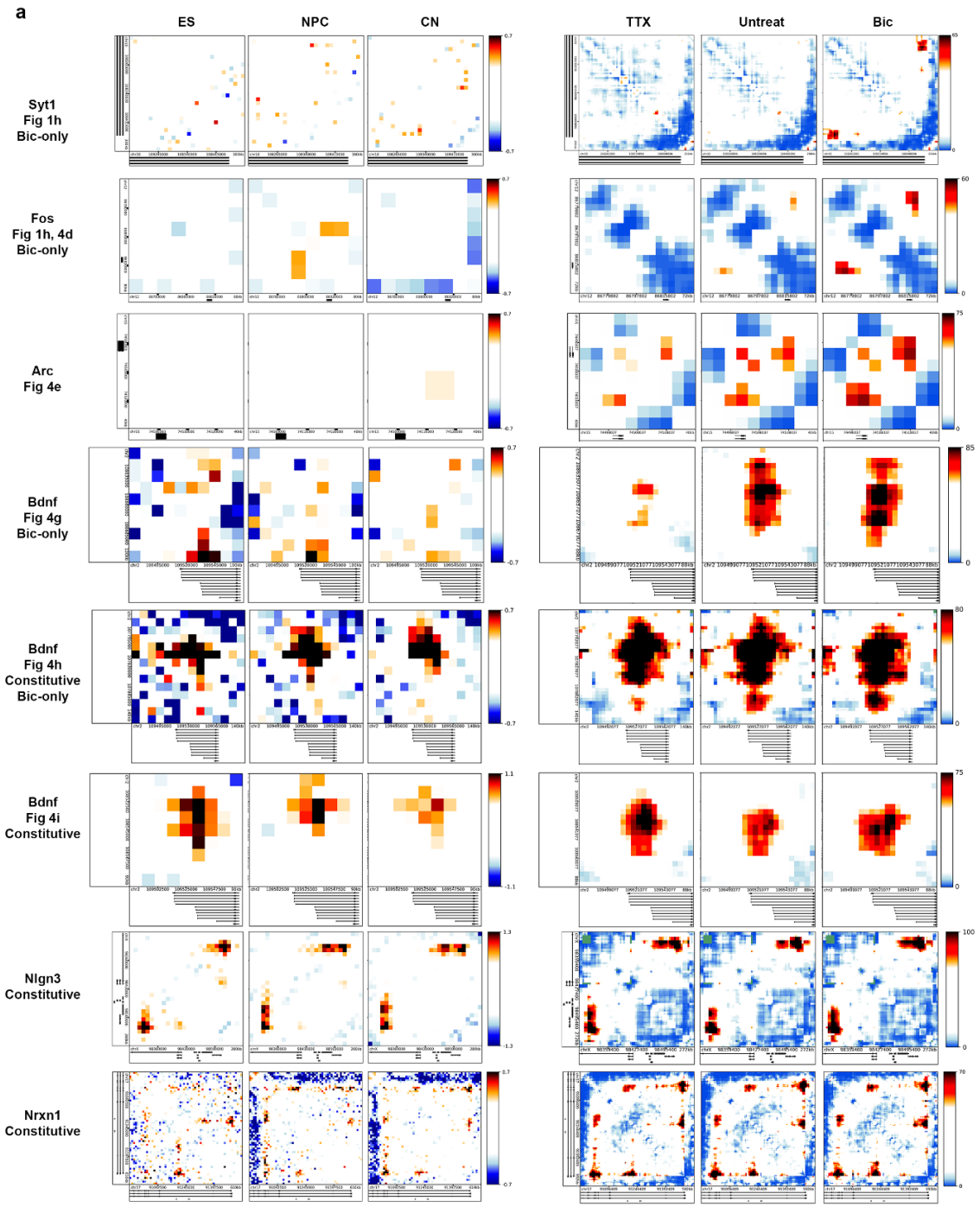


Figure 5.4. Activity-induced loops are not present earlier in cortical neuron differentiation. (A) Zoom-in heatmaps of critical loops presented throughout the paper. From left to right the columns are Obs/Exp heatmaps of HiC (Bonev et al.) data from 1) embryonic stem (ES) cells, 2) neural progenitor cells (NPC), 3) cortical neurons (CN), followed by 5C interaction score heatmaps across the 4) TTX, 5) untreated, and 6) BIC treated conditions. Genes of interest in each zoom window, Figure panels where same loop is further analyzed, and loop classification are listed on left.

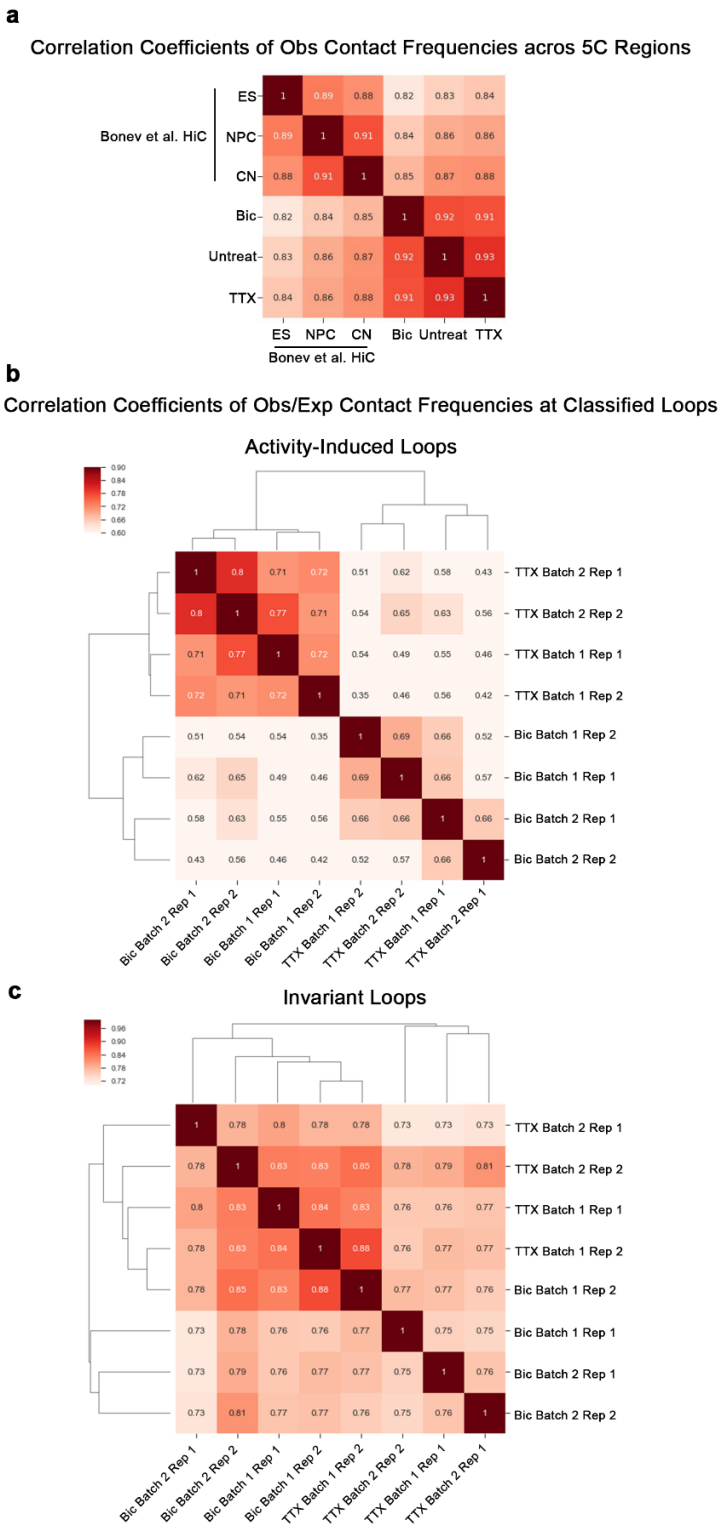


Figure 5.5. 5C data correlates most strongly with cortical neuron HiC, clusters by condition. (A) Spearman's correlation coefficients of comparisons between Bonev et al. HiC data (ES, NPC, CN) and 5C data. Regions of interest were extracted from raw HiC data; HiC and 5C counts were then binned to equivalent 10kb bins, quantile normalized

together and ICE matrix-balanced prior to correlation computation. **(B-C)** Pearson's correlation coefficients of background-normalized contact frequencies ("observed/expected") at activity-induced loops (B) and activity-invariant loops (C) across each pair of replicates. Replicates were then hierarchically clustered based on correlation results.

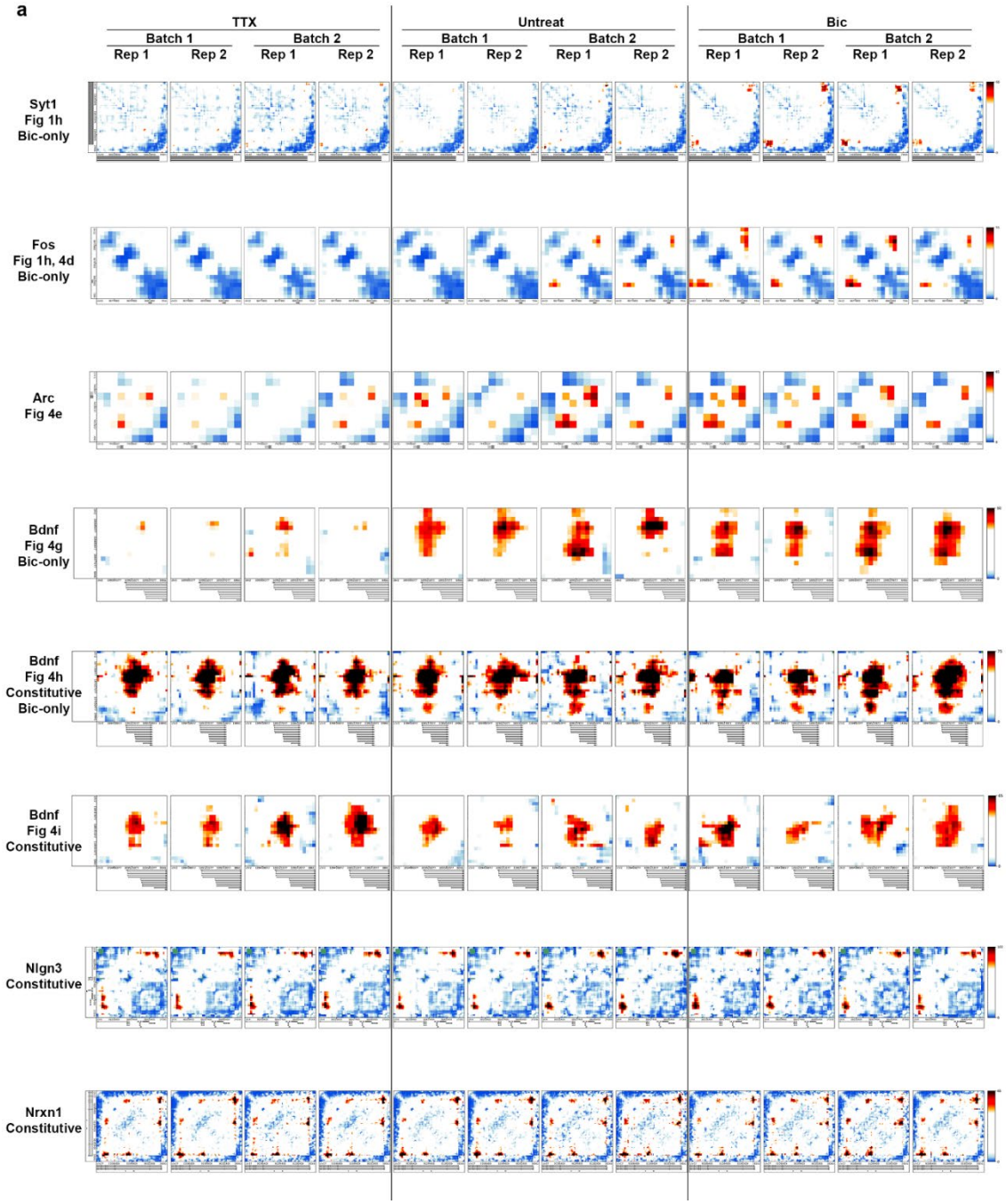


Figure 5.6. Activity-induced and activity-invariant loops are reproducible across condition replicates. (A) Zoom-in interaction score heatmaps from each of the 12 5C replicates generated for critical loops presented throughout the paper. Genes of interest

in each zoom window, Figure panels where same loop is further analyzed, and loop classification are listed on left.

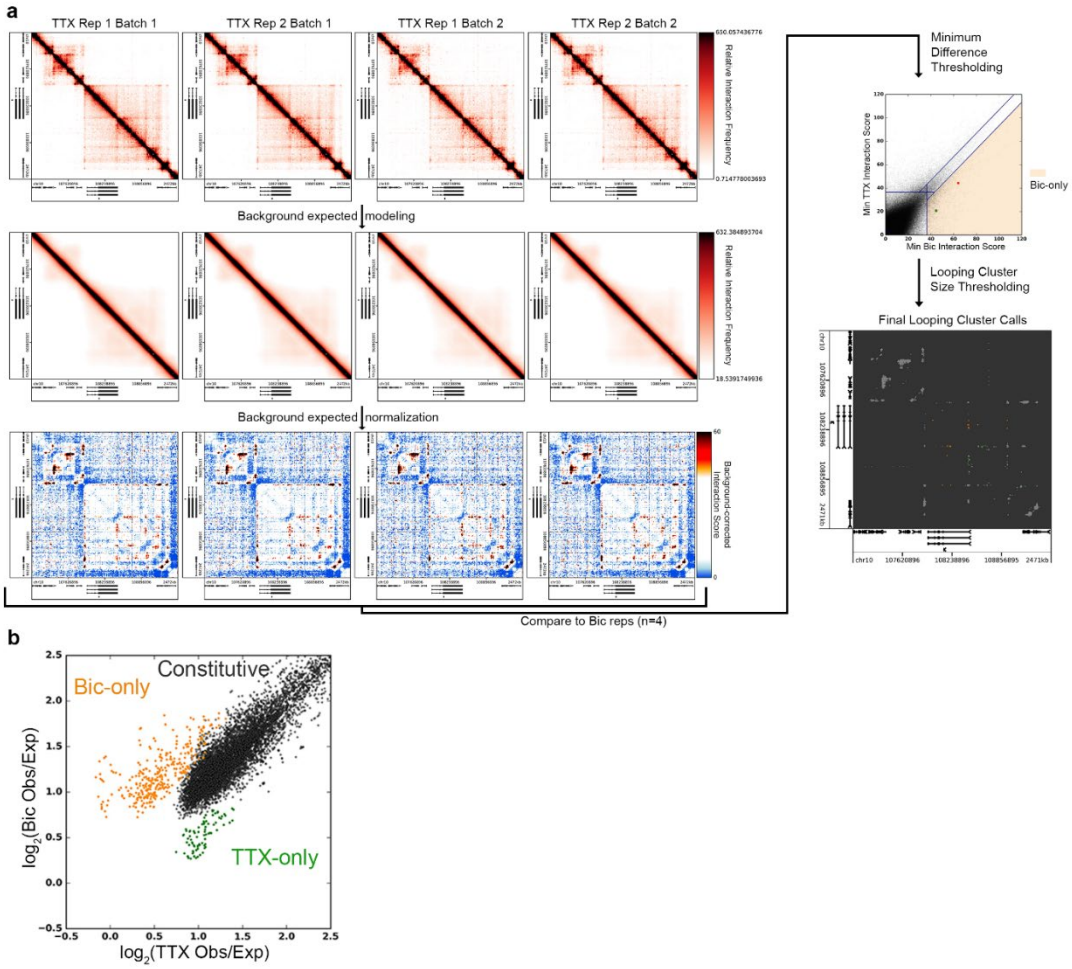


Figure 5.7. Identifying dynamic looping across neural activity states. (A) Diagram of 5C processing pipeline used to call significant constitutive and dynamic loops (bottom right) starting from 5C interaction frequency counts for all pairs of 4 kb genomic bins within queried regions across 4 replicates (from two litter/culture batches) of each condition (top left). First the local domain background signal is quantified using a donut expected model (Rao+ 2014) and removed from the interaction frequency signal. Probabilistic modeling converts these expected-normalized interaction frequencies to an "interaction score" (bottom left). For a bin-bin pair to be classified as looping, its interaction score must fall above a given "significance threshold". For a looping bin-bin pair to be classified as "Bic-only" the minimum interaction score of the Bic replicates must exceed the maximum interaction score of the four TTX replicates by a given "difference threshold" (Supplemental Methods). Looping pixels not classified as Bic- or TTX-only are classified as constitutive (top right). Bin-bin pairs of the same class are then grouped into clusters if they are directly adjacent; clusters below a selected size

threshold are removed from looping classification (bottom right). See Methods for more details. **(B)** Scatterplot of the background-normalized contact frequency ("Observed/Expected") counts of looping-classified pixels in TTX and Bic conditions.

5.1F, Fig. 5.7B). We observed that activity-invariant loops exhibited high interaction frequencies across Untreat, TTX, and Bic conditions (**Fig. 5.1G**). Importantly, activity-induced and activity-decommissioned loops showed 2-3-fold up- or down-regulation in interaction frequency, respectively, but were still lower in overall looping strength than the activity-invariant contacts (**Fig. 5.1G**). We confirmed that an enhancer-promoter loop at the *c-fos* rIEG previously reported as activity-dependent via 3C-PCR[234] was classified here as an activity-induced loop (**Fig. 5.1H, top**) and that additional activity-induced loops occurred across our 5C regions (**Fig. 5.1H, bottom**). These data highlight that both activity-invariant and -dynamic loops encompass IEGs and synaptic genes.

We wondered if the looping landscape and its relationship to activity-dependent enhancers could shed light on the regulation of activity-dependent gene expression. Because the histone mark H3 lysine 27 acetylation (H3K27ac) correlates with enhancer and promoter activity, we conducted H3K27ac ChIP-seq to identify changes in putative non-coding enhancer elements genome-wide in neural activity states. We noticed a strong correlation between activity-dependent changes in promoter H3K27ac signal and gene expression (**Fig. 5.8A**). By contrast, the total sum interaction frequency by each gene was not correlated with gene expression (**Fig. 5.8B**). Next, using thresholded loops (**Fig. 5.1F**), we then applied an adapted ABC model[141], to identify the single loop/enhancer for each gene that displayed the maximum value of (loop strength x enhancer H3K27ac signal) (**Fig. 5.8C, Appendix III Methods**). Importantly, by testing only the thresholded loop with the

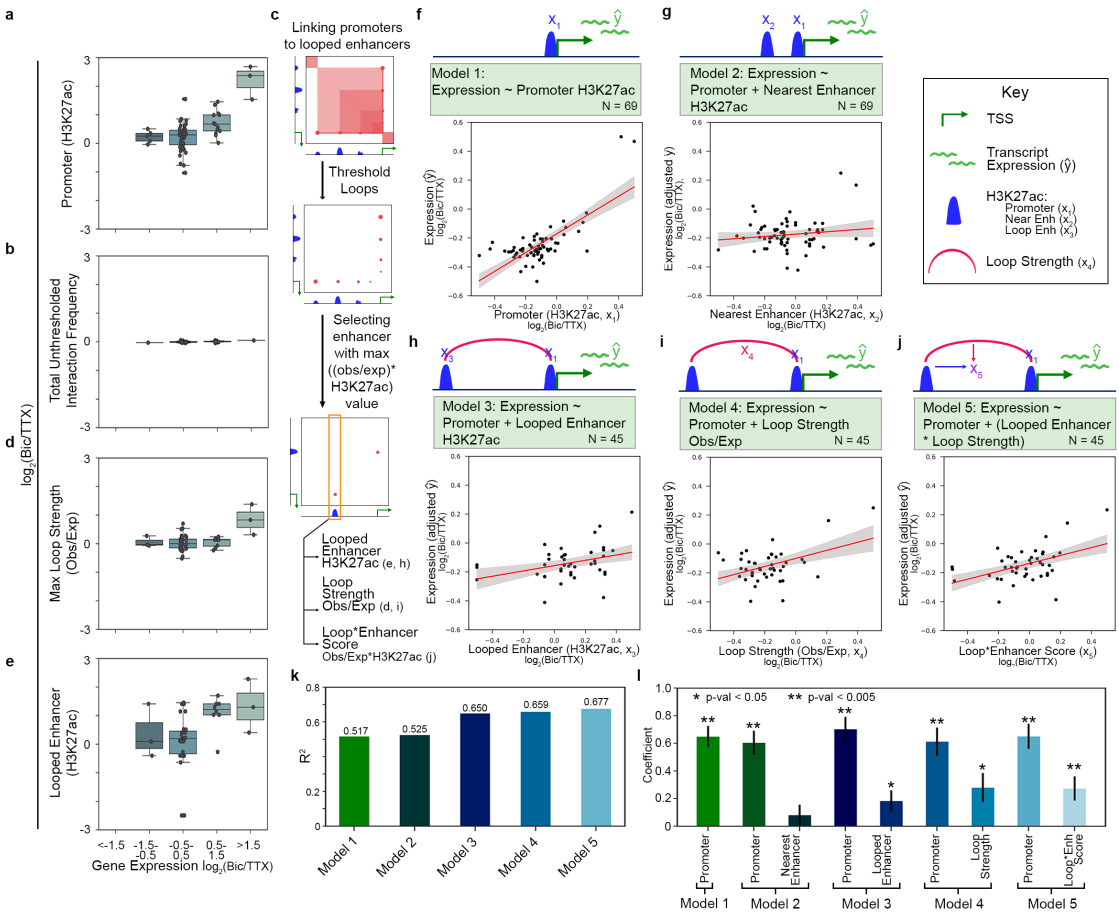


Figure 5.8. Activity-induced enhancers connected to distal target genes via looping interactions predict activity response expression. (A-B) Boxplots of the promoter acetylation (A) and total interaction frequency (B) fold changes of genes grouped by expression fold change. **(C)** Schematic representation of algorithm used to pair each gene with a single loop/enhancer that offered the highest predictive value. Only genes that formed such a loop (N = 45) were queried in the following models (D,E,H,I). **(D-E)** Boxplots of the loop strength (D) and looped enhancer acetylation (E) after loops and enhancers are matched to genes using schema presented in (C). **(F-G)** Cartoon representations and scatter plots of the two 'null' models of Bic/TTX gene expression fold change: (F) promoter acetylation alone (model 1), (G) promoter acetylation plus the acetylation of the nearest enhancer within 200 kb of the TSS (model 2). Expression fold change is plotted on the y-axis while acetylation fold change (of promoter in (F) and nearest enhancer in (G)) is plotted on the x-axis. The expression fold change in (G) has been adjusted to remove the values predicted by the promoter activity term in the model. Values have been min/max scaled to allow cross-model comparison. **(H-J)** Cartoon representations and scatter plots of loop-containing models, plotted in the same manner as (G). **(K)** R^2 values for each of the three models. **(L)** Barplot of explanatory variable

coefficients from models 1-5. t-statistic p-values and standard errors represented via stars and error bars, respectively.

highest (loop x enhancer) score for each gene, we observed a strong increase in interaction strength at the most strongly activity-upregulated genes (**Fig. 5.8D**). Moreover, when using the same enhancer-gene pairing schema (**Fig. 5.8C**), H3K27ac signal was consistently increased at distal putative enhancers linked via looping to activity-upregulated target genes (**Fig. 5.8E**). Together, these data indicate that signal strength of epigenetic marks at distal regulatory elements and the interaction frequency of their long-range loops correlate with activity-dependent gene expression.

It is poorly understood which activity-dependent enhancers regulate specific target genes. The best studied examples of activity-dependent enhancers, those at the *c-fos* and *Arc*[217, 218, 234], are relatively close (≤ 40 kb) to the promoters of these genes, however in many cases nearest enhancers are insufficient to explain transcriptional regulation. We built a predictive model of activity-dependent gene expression (**Appendix III Methods**). Promoter H3K27ac alone explained only 51.7% of the variance in gene expression upon neuronal activation in our 5C regions (**Fig. 5.8F,K-L**). By adding the covariate of H3K27ac signal at the nearest enhancer, we only marginally increased model performance (**Fig. 5.8G,K-L**). We then built a third model with covariates of activity-dependent H3K27ac at (i) promoters and (ii) only distal enhancers engaged in maximum ABC-thresholded loops with their target genes (**Fig. 5.8C, Appendix III Methods**). Our third 'long-range enhancer model' markedly increased the variance of activity-dependent expression explained from 51.7% to 65% (**Fig. 5.8H, 5.8K-L).** Surprisingly, models using the strength of the loop (**Fig. 5.8I**) or the value of (loop strength x enhancer H3K27ac) between the selected

enhancer and promoter (**Fig. 5.8J**) as predictors performed similarly well (**Fig. 5.8I-L**), suggesting that further work is required to determine how loop strength alterations contribute to gene expression levels[141]. These trends remained consistent when we analyzed the promoter and nearest enhancer models for genes that only form long-range loops (**Fig. 5.9**). Together, these data indicate that long-range loops can provide significant improvement in the prediction of activity-dependent expression by connecting specific distal enhancers to their target genes.

a

Spearman's Correlation Coefficient

N = 69	Expression	Promoter	Nearest Enhancer
Expression		0.506138107	0.312387516
Promoter	0.506138107		0.439159027
Nearest Enhancer	0.312387516	0.439159027	

N = 45	Expression	Promoter	Nearest Enhancer	Looped Enhancer	Loop Strength	Looped Enhancer * Loop Strength
Expression		0.47483531	0.459054266	0.575249814	0.122409988	0.530223674
Promoter	0.47483531		0.516658491	0.241706596	0.241987022	0.275060119
Nearest Enhancer	0.459054266	0.516658491		0.386551861	-0.065190171	0.328521547
Looped Enhancer	0.575249814	0.241706596	0.386551861		0.151403441	0.944274145
Loop Strength	0.122409988	0.241987022	-0.065190171	0.151403441		0.402516966
Looped Enhancer * Loop Strength	0.530223674	0.275060119	0.328521547	0.944274145	0.402516966	

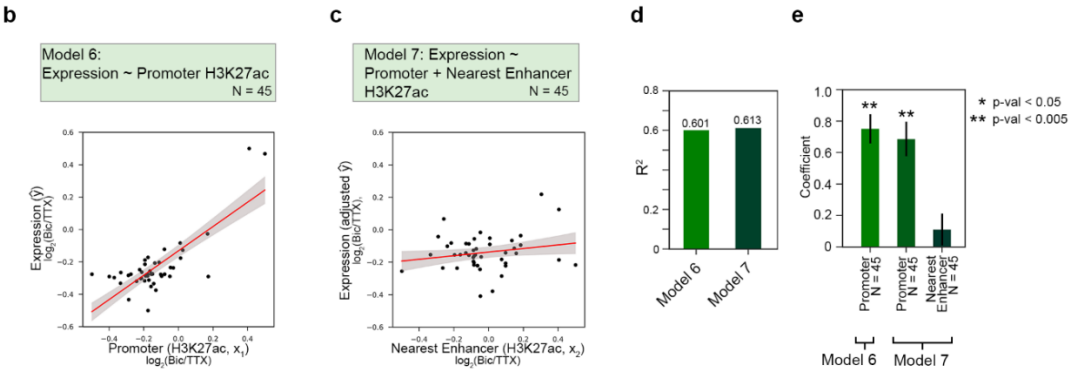


Figure 5.9. Correlation coefficients of modeled regulatory element signals. (A) Spearman's correlation coefficients for terms included in models (Fig. 2f-i). **(B-C)** Results of promoter-only (B) and promoter plus nearest enhancer (c) models for only genes that form loops to classified enhancers within 5C regions. **(D)** R2 values of models presented in (B-C). **(E)** Coefficients of each explanatory variable term in models presented in (B-C).

We next set out to determine the extent to which looping reconfiguration occurred in parallel with activity-dependent enhancer changes or if enhancers were pre-wired to their targets independent of their activation state (**Fig. 5.10A**). We first stratified H3K27ac peaks into activity- invariant (n=14,424), activity-induced (n=6014), and activity-decommissioned (n=5402) putative enhancers (**Fig. 5.10B-C, Appendix III Methods, Fig. 5.11A-C**). We quantified the degree of overlap between our enhancer classes and the anchors of our looping interactions. We identified three major enhancer+loop classes for further exploration: (i) activity-induced loops anchored by activity-induced enhancers (n=11) (Class 1), (ii) activity-invariant loops pre-wired in inactive neurons and anchored by activity-induced enhancers (n=41) (Class 2), and (iii) activity-invariant loops pre-wired in inactive neurons and anchored by activity-decommissioned enhancers which lose their H3K27ac signal upon chronic neuronal activation (n=15) (Class 3) (**Fig. 5.10D-E**). These data reveal a complex long-range cis-regulatory landscape in which enhancer activation does not always correlate with *de novo* loop formation and suggest that diverse loop classes might play unique roles in regulating activity-dependent gene expression.

We next investigated the potential structural and functional properties of our three loop classes. We noticed that activity-induced loops anchored by activity-induced enhancers (Class 1) underwent a 2.2-fold change in interaction frequency after 24 hours Bic treatment (**Fig. 5.10F**). Activity-invariant loops anchored by activity-decommissioned enhancers showed strong and unchanged interaction frequency (Class 3, **Fig. 5.10F**). By contrast, interaction strength further strengthens upon neuronal activation in the case of activity-invariant loops pre-wired to activity-induced enhancers (Class 2, **Fig. 5.10F**). Importantly, although Class 1 loops are a rare occurrence, they corresponded to a 24-fold

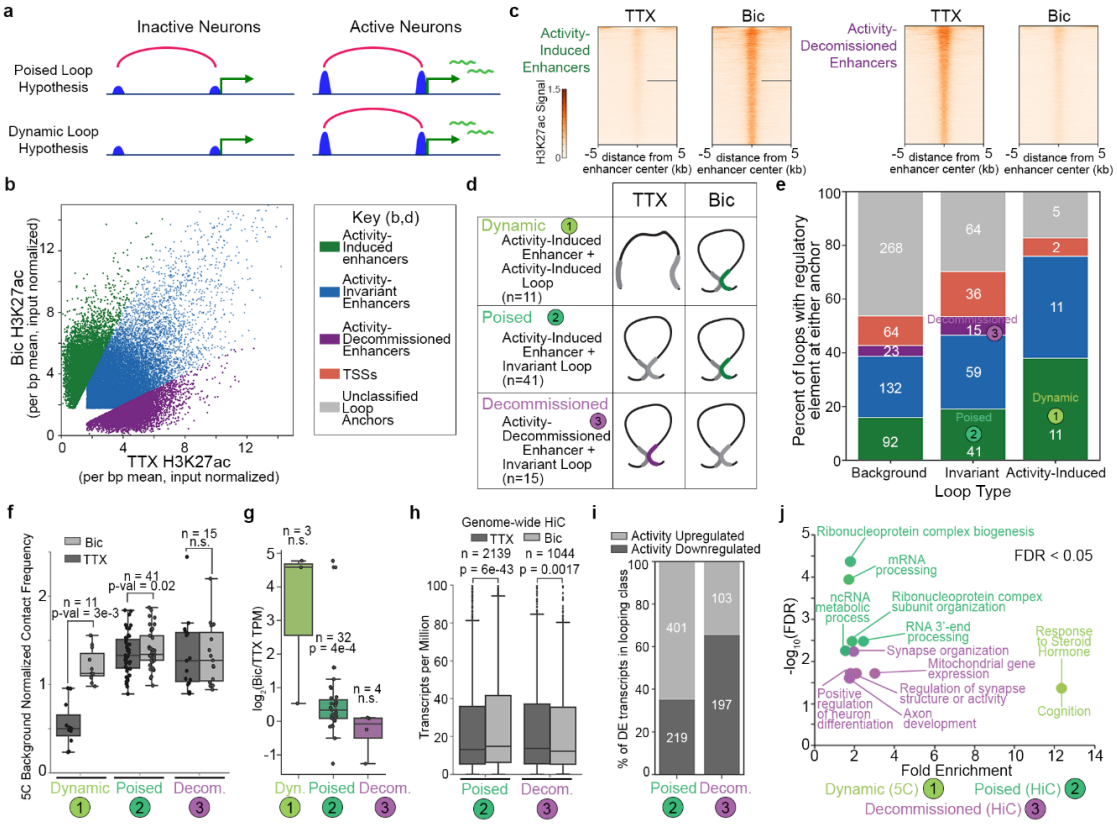


Figure 5.10. Unique topological motifs underlie the activity-dependent transcriptional response. (A) Cartoon representation of hypotheses in which activity-induced enhancers operate to control gene expression via poised (top) or dynamic (bottom) loops. **(B)** Scatterplot of enhancer acetylation across Bic and TTX conditions, thresholded by fold change of input normalized signal and classified into activity-induced (green), activity-invariant (blue), and activity-decommissioned (purple) enhancers. **(C)** Acetylation heatmaps of classified dynamic enhancers. **(D)** Stacked barplot displaying the percent of loops in each looping class with a classified enhancer at either of its anchors. Enhancer class key located to left. Number of loops in each subset depicted on top of bar. Loops could only be assigned to one enhancer class; enhancer class priority order ranges from bottom of barplot (activity-induced enhancers, considered first) to top (TSSs, considered last). **(E)** Cartoon representations of three loop-enhancer classes of top interest from (D). Classified loop anchor colors match those in (B-D). **(F)** Boxplots of background normalized contact frequencies for looping pixels in the five looping classes. P-values presented in **F-H** calculated using two-tailed Wilcoxon signed-rank test. Number of loops in each class listed above boxes. **(G)** Expression fold change (log2(Bic/TTX)) of the transcripts whose promoters intersect each looping class. Number of genes in each class listed above boxes. **(H)** Expression (TPM) of the genes whose promoters fall opposite activity-induced (class 2) and activity-decommissioned (class 3) enhancers in genome-wide cortical neuron loops, original data from Bonev et

al. 2017. Number of genes in each class listed above boxes. **(I)** Percent of differentially expressed genes (parsed using Sleuth[235] Wald test, q-val < 0.05) in each genome-wide looping class that are upregulated in Bic compared to TTX (light grey) or downregulated in Bic compared to TTX (dark grey). **(J)** Gene ontology enrichment of transcripts presented in (G-H). Class 1 genes are from 5C regions only (g), class 2,3 genes were parsed using the genome-wide analyses in (h). Only the top 5 terms for classes 2 could be shown, see Fig. 5.13 for remaining terms at FDR < 0.05.

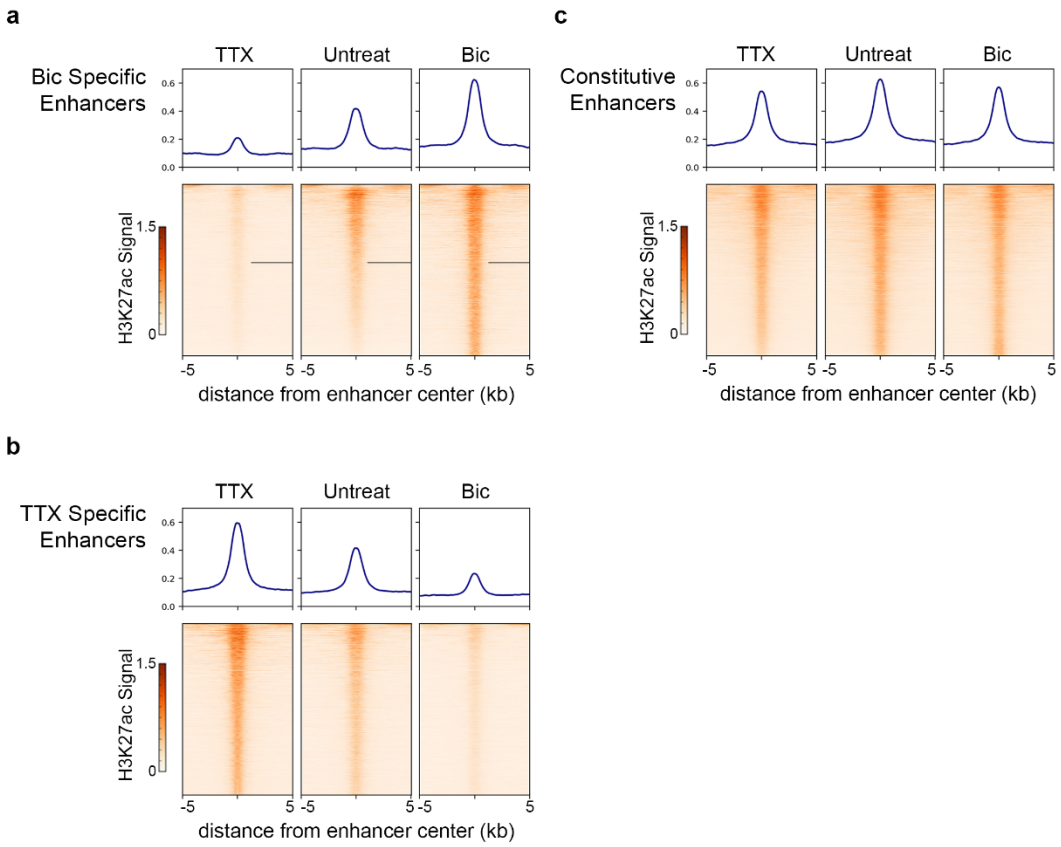


Figure 5.11. Parsing activity dependent enhancers. (A-C) Acetylation heatmaps, pileups of classified activity-induced (A), activity-decommissioned (B), invariant (C) enhancers.

increase in activity-induced expression (**Fig. 5.10G, 5.12A**). Comparatively more genes engaged in Class 2 loops but on average displayed a modest 1.3-fold increase in expression in active neurons (**Fig. 5.10G, 5.12A).** These results suggest that, in our 5C regions, activity-induced loops are rare and connect to genes with large activity-dependent increases

in expression, whereas pre-existing loops exist in larger numbers but correlate with only minor gene expression changes.

To extend our findings genome-wide, we assessed the link between activity-invariant loop Classes 2 and 3 and gene expression using the high-resolution Hi-C maps published in primary cortical neurons[104] and our activity-dependent RNA-seq and ChIP-seq data (**Fig. 5.10H, Fig. 5.12B-C).** We applied established published methods[10] to identify 24,937 loops in cortical neurons (**Fig. 5.12B-C)** and stratify them into Class 2 (n=4,764) and Class 3 (n=3,259) groups (**Appendix III Methods).** Consistent with 5C loops, genes connected to activity-induced enhancers via activity-invariant loops (Class 2) displayed a modest but significant upregulation in expression upon neuronal activation when we queried genome-wide loops (**Fig. 5.10H, 5.12D**). By contrast, genes looped to activity-decommissioned enhancers via activity-invariant loops (Class 3) genome-wide exhibited a slight reduction in expression upon neural activation (**Fig. 5.10H**). The majority of differentially expressed genes in Class 2 versus Class 3 loops were upregulated and downregulated, respectively, due to activity (**Fig. 5.10I**). Together, our data reveal that the genes connected to activity-induced enhancers via rare *de novo* loops show the largest effect size in activity-dependent expression. Genes can also exhibit modest but notable up- or down-regulation when connected via pre-wired, activity-invariant loops to activity-induced (Class 2) or activity-decommissioned (Class 3) enhancers, respectively. Pre-existing Class and Class 3 loops are markedly more abundant in number compared to Class 1.

We explored the ontology of the long-range target genes anchoring each looping class. Class 1 loops connect *c-Fos*, *Bdnf,* and *Tmed10* to activity-inducible enhancers,

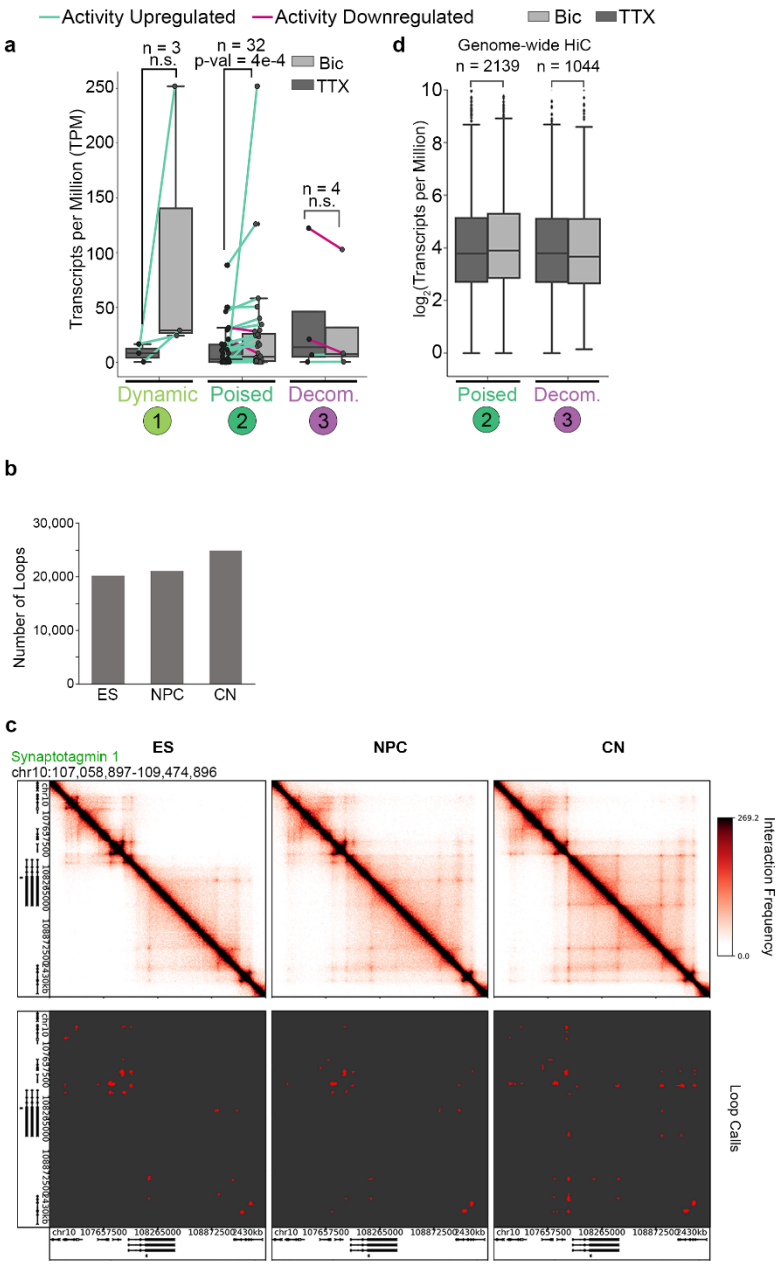
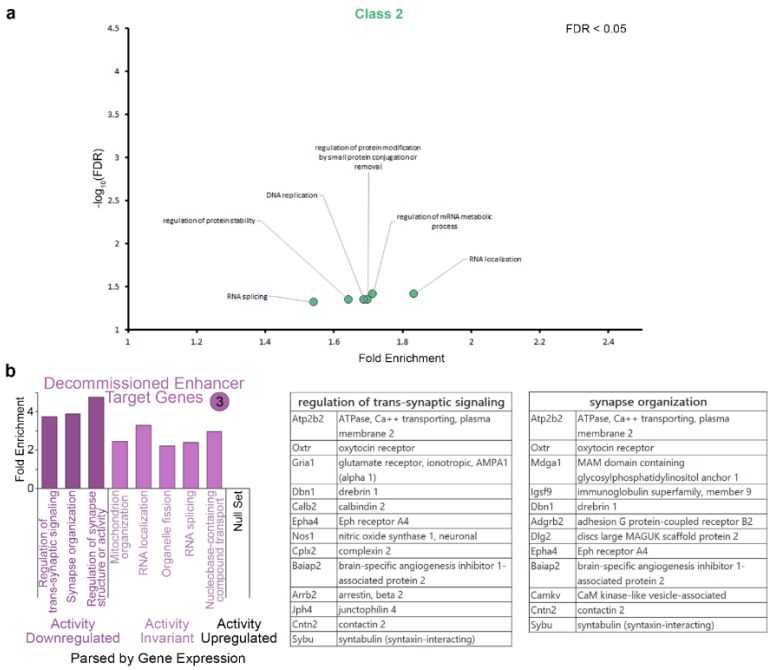


Figure 5.12. Murine HiC (Bonev et al., 2017) loop calls. (A) Expression (TPM) of the transcripts whose promoters intersect each looping class. **(B)** Number of loops called in HiC data obtained from embryonic stem cells (ES), neural progenitor cells (NPCs), and cortical neurons (CN) (Bonev et al. 2017). **(D)** Interaction frequency heatmaps (top) and thresholded loop calls (bottom) for a ~2.5 Mb region surrounding the *Synaptotagmin1* gene. **(D)** Expression (log_2(TPM)) of the genes whose promoters fall opposite activity-induced (class 2) and activity-decommissioned (class 3) enhancers in genome-wide cortical neuron loops, original data from Bonev et al. 2017. Number of genes in each class listed above boxes.

suggesting that the rapid upregulation of IEGs involves *de novo* loop and *de novo* enhancer induction during neural activation (**Fig. 5.10J**). Class 2 pre-existing loops connect genes involved in several general cellular functions such as RNA processing to activity-induced enhancers, whereas Class 3 pre-existing loops anchored by activity-decommissioned enhancers connect genes linked to synaptic organization and the regulation of synaptic activity (**Fig. 5.10J, Fig. 5.13A**). We were intrigued by the placement of synaptic genes in Class 3 loops given that they connect to enhancers that are turned off during chronic (24 hour) high activity levels. Thus, we further stratified genes connected in Class 3 loops by those undergoing a (i) 1.5-fold downregulation, (ii) 1.5-fold upregulation, and (iii) remaining unchanged upon neural activity (**Appendix III Methods**). We found that the cohort of genes undergoing decreased expression in Class 3 loops were predominantly genes involved in synapse organization and signaling, including *Gria1,* the main AMPA receptor subunit (**Fig. 5.10J, Fig. 5.13B**). These results open up the possibility for future inquiry into a potential mechanistic role for Class 3 loops and activity-decommissioned enhancers in facilitating homeostatic plasticity during chronic high neural activity. Together, these data support our working hypothesis that both activity-induced loops connecting activity-induced enhancers (Class 1) and activity-invariant loops connecting activity-decommissioned enhancers (Class 3) play a role in synaptic plasticity. It is well established that rIEGs such as *c-fos* and *Arc* are activated on the order of seconds to minutes in a translation-independent manner, whereas dIEGs/SRGs such as *Bdnf* are activated on the order of minutes to hours[214]. Consistent with this idea, we re-analyzed a recently published RNA-seq time course during pharmacological neuronal activation[36] and found maximum activation of *c-fos* and *Arc* by 60 minutes, whereas maximum *Bdnf*



regulation of trans-synaptic signaling	
Atp2b2	ATPase, Ca++ transporting, plasma membrane 2
Oxtr	oxytocin receptor
Gria1	glutamate receptor, ionotropic, AMPA1 (alpha 1)
Dbn1	drebrin 1
Calb2	calbindin 2
Epha4	Eph receptor A4
Nos1	nitric oxide synthase 1, neuronal
Cplx2	complexin 2
Baiap2	brain-specific angiogenesis inhibitor 1-associated protein 2
Arrb2	arrestin, beta 2
Jph4	junctophilin 4
Cntn2	contactin 2
Sybu	syntabulin (syntaxin-interacting)

synapse organization	
Atp2b2	ATPase, Ca++ transporting, plasma membrane 2
Oxtr	oxytocin receptor
Mdga1	MAM domain containing glycosylphosphatidylinositol anchor 1
Igsf9	immunoglobulin superfamily, member 9
Dbn1	drebrin 1
Adgrb2	adhesion G protein-coupled receptor B2
Dlg2	discs large MAGUK scaffold protein 2
Epha4	Eph receptor A4
Baiap2	brain-specific angiogenesis inhibitor 1-associated protein 2
Camkv	CaM kinase-like vesicle-associated
Cntn2	contactin 2
Sybu	syntabulin (syntaxin-interacting)

Figure 5.13. Significantly enriched gene ontology terms Associated with looping classes 2,3. (A) The remaining gene ontology terms passing the FDR < 0.05 threshold for class 2 (a) which could not be presented in **Figure 5.10**. **(B)** (Left) Gene ontology enrichment ratios for class 3 genes parsed by expression into activity downregulated (Bic/TTX < 2/3), activity invariant (5/6 < Bic/TTX < 6/5), and activity upregulated (Bic/TTX > 3/2) groups. (Right) Genes found in the 'regulation of trans-synaptic signaling' and 'synapse organization' GO terms enriched in activity downregulated class 3 genes.

upregulation occurred at 6 hours (**Fig. 5.14A**). Visual inspection of our 5C heatmaps revealed two unexpected observations linking the kinetics of activity-dependent transcription to looping complexity (**Fig. 5.14B**). First, rIEGs in our 5C regions form simple short-range loops with activity-dependent enhancers, and thus fall nearly exclusively in the Class 1 category. For example, after 24 hours of Bic treatment, *c-Fos* was upregulated more than 100-fold (**Fig. 5.14C**), but we identified only a single 40 kb-sized Class 1 loop with an activity-induced enhancer (**Fig. 5.14D**). Similarly, *Arc* was upregulated more than 12-fold upon neural activation (**Fig. 5.14C**) and also connected in a

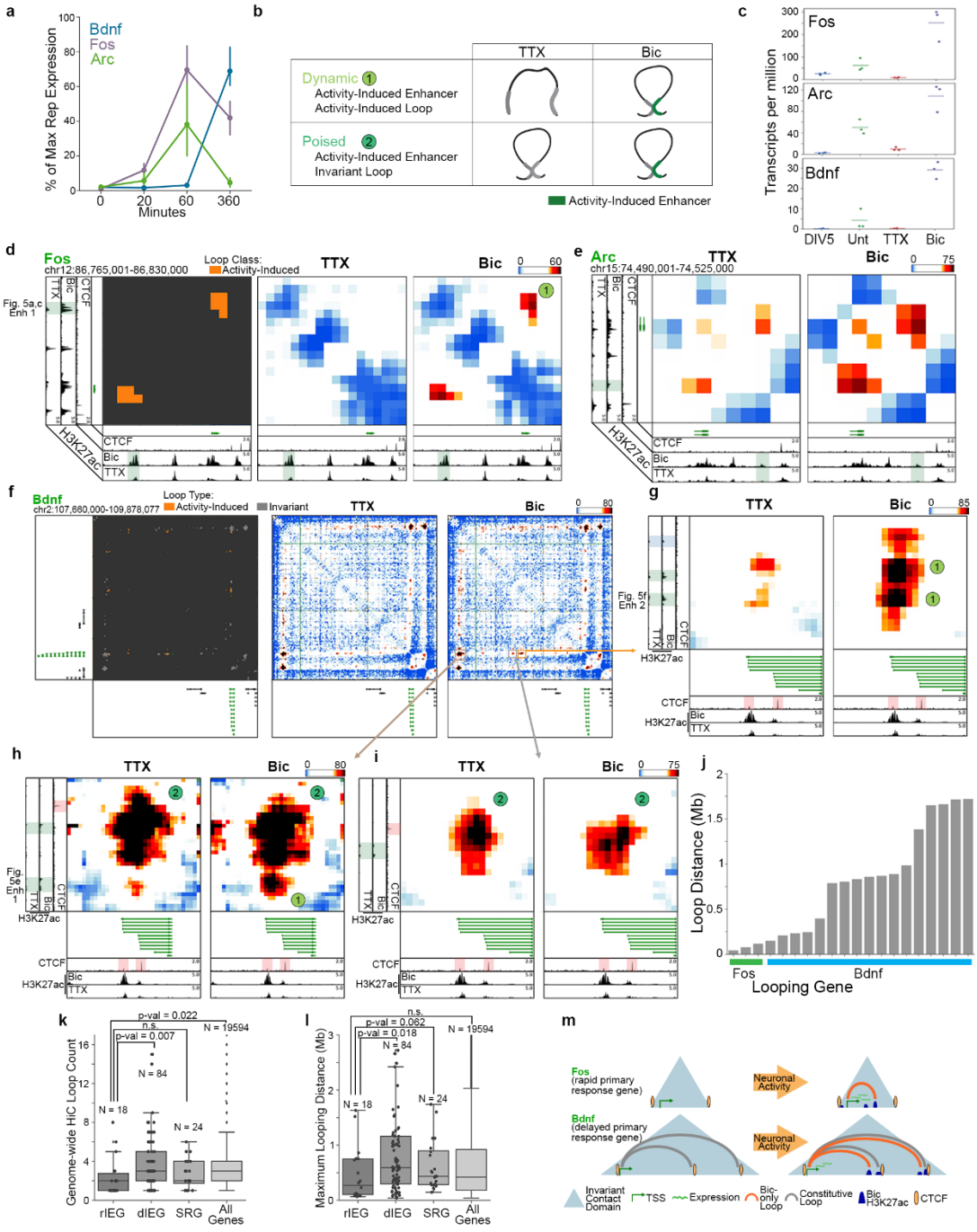


Figure 5.14. Rapid immediate early genes form shorter and less complex loops than secondary response genes. (A) Expression timing of *Bdnf*, *Fos*, and *Arc* following the initiation of cortical neuron stimulation from Tyssowski et al. 2018. **(B)** Cartoon representations of two loop classes identified in Fig. 3. **(C)** Expression (TPM) of the *Arc*, *Bdnf* and *Fos* genes across the 5 days in vitro (DIV5), untreated, TTX, and Bic conditions. **(D)** Loop calls (left), TTX interaction score heatmap (middle) and Bic interaction score heatmap (right) of a ~65 kb region surrounding the *Fos* gene (green). Plotted beneath maps are cortical neuron CTCF (Bonev et al. 2017), Bic H3K27ac, and

TTX H3K27ac tracks. Bic specific enhancer underlying Bic loop highlighted in orange. **(E)** TTX interaction score heatmap (left) and Bic interaction score heatmap (right) of a ~35 kb region surrounding the *Arc* gene (green). **(F)** TTX interaction score heatmap (top), Bic interaction score heatmap (middle), and loop calls (bottom) of a ~2 Mb region surrounding the Bdnf gene (green). Bic loops plotted in orange, and constitutive loops in grey. **(G-I)** Interaction score heatmaps of 3 looping regions highlighted in (F) across TTX (left) and Bic (right) conditions. Plotted beneath maps are cortical neuron CTCF (Bonev et al. 2017), Bic H3K27ac, and TTX H3K27ac tracks. Bic specific enhancers are highlighted in orange and CTCF peaks highlighted in red. **(J)** Genomic distance spanned by each loop formed by the *Fos* (n=3) and *Bdnf* (n=17) genes. **(K-L)** Boxplots overlaid by stripplots of loop count (K) and max looping distance (L) for rapid immediate early genes (rIEGs, as defined as rPRGs in Tyssowski et al. 2018), delayed immediate early genes (dIEGs), secondary response genes (SRGs), and all genes. P-values presented for two-sided Mann Whitney rank tests comparing of rIEGs to other 3 classes. **(M)** Model representation of the distinct looping patterns of the *Bdnf* and *Fos* genes.

singular loop with an activity-induced enhancer (**Fig. 5.14E**). We note that the *Arc* interaction falls below our 30 kb distance threshold and therefore is not formally added to the Class 1 loop list (**Fig. 5.10G-J**). By contrast, *Bdnf* was upregulated 30-fold upon neuronal activation (**Fig. 5.14C**) and connected into a complex network of multiple long-distance Class 1 and Class 2 loops (**Fig. 5.14F-I)**, including: (i) at least two Class 1 activity-induced loops anchored by activity-induced enhancers, but spanning longer distances (840 and 1,700 kb) than those formed with IEGs (**Fig. 5.14G-H**) and (ii) at least two Class 2 activity-invariant loops anchored by activity-induced enhancers (**Fig. 5.14H-I**). The loops formed by *Bdnf* were preferentially located at *Bdnf*'s first promoter, from which we observed the highest level of transcription and strongest upregulation after 24 hours of Bic-induced neuronal activation (**Fig. 5.15**). Loops connected by *Bdnf* were significantly longer than those connected by *c-fos* and *Arc* (**Fig. 5.14J**). These observations provide the basis for our working hypothesis that loop complexity and size contribute to the timing of IEG versus SRG upregulation in response to neuronal activation.

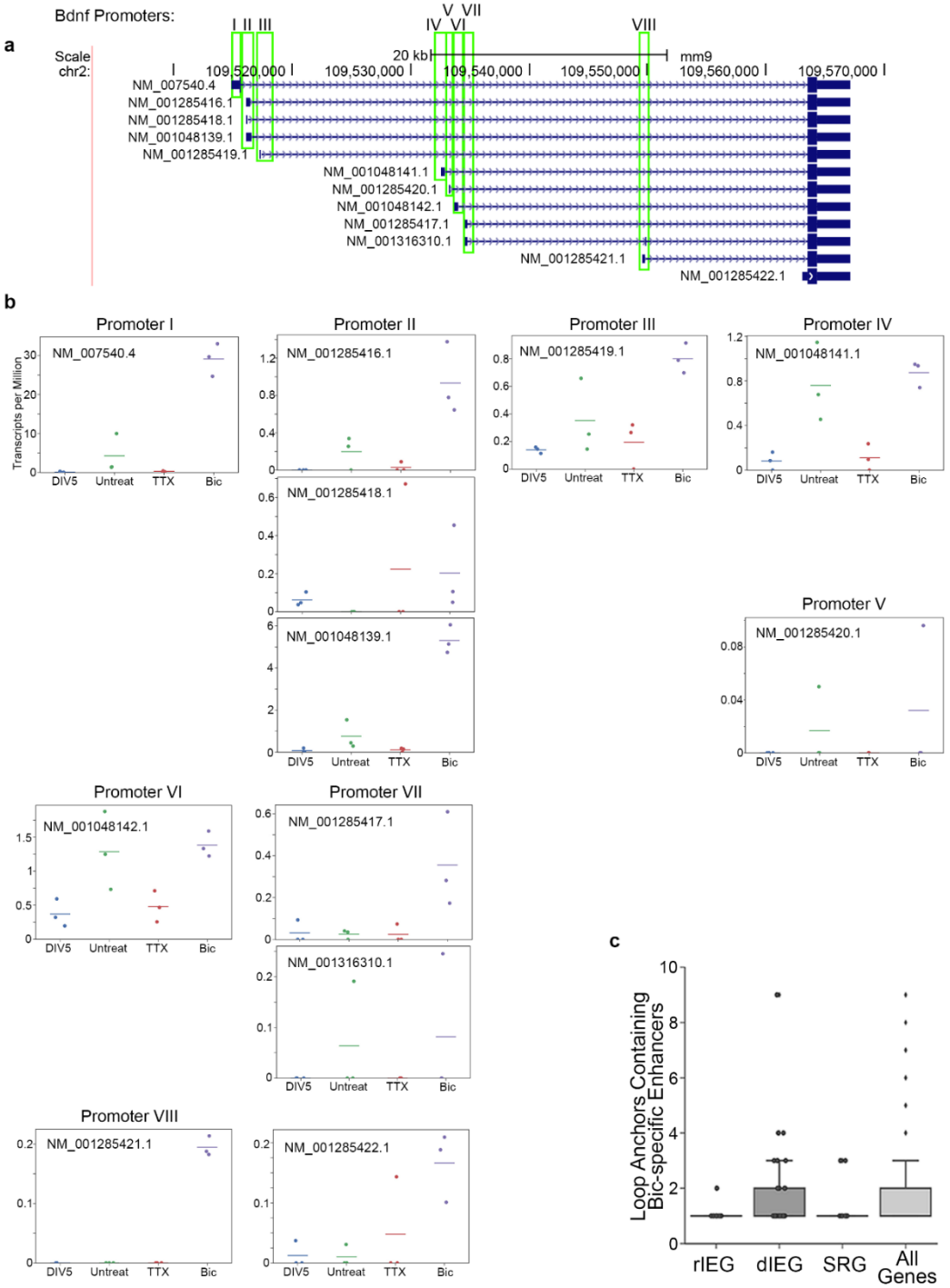


Figure 5.15. Expression of *Bdnf* transcripts. (A) Depiction of the 12 RefSeq transcript isoforms of the *Bdnf* gene, above which we annotate the 8 promoters as in Hong et al., *Neuron*, 2008. **(B)** Expression strip plots of each *Bdnf* isoform, organized in columns by shared promoter. **(C)** Boxplots overlaid by strip plots of count of opposing looping anchors that contain an activity-dependent enhancer for rapid immediate early genes

(rIEGs, as defined as rPRGs in Tyssowski et al. 2018), delayed immediate early genes (dIEGs), secondary response genes (SRGs), and all genes.

We next explored loop complexity genome-wide using published annotations of rIEGs, dIEGs, and SRGs[36] and our 24,937 loops in ES-derived mouse cortical neuron Hi-C (**Fig. 5.12B-C**). Published Hi-C data only represents one untreated activity state, thus we could not assess activity-induced loops (Class 1) genome-wide. Nevertheless, we could integrate our enhancers with cortical neuron Hi-C data to query the complexity of activity-invariant Class 2 loops surrounding known activity-dependent genes genome-wide. Consistent with our locus-specific 5C results, we found that rIEGs form significantly fewer loops (**Fig. 5.14K**), shorter loops (**Fig. 5.14L**), and connect to a lower number of activity-induced putative enhancers (**Fig. 5.15C**) compared to dIEGs and SRGs genome-wide. Together, these data are consistent with our working model in which dIEGs engage in a complex network of long-range regulatory interactions, whereas rIEGs form simple, short-range loops to activity-induced enhancers to facilitate rapid activation independent of new protein synthesis (**Fig. 5.14M**).

The disparate length and number of loops which emerged after chronic neuronal activity at Fos/Arc compared to Bdnf, as well as the differences in the expression timing of the genes, led us to hypothesize that the two sets of loops may display distinct formation kinetics. To explore looping dynamics after short term activity induction, we next created 5C architecture maps in an acute time course of 0, 5, 20, 60, and 360 minutes of pharmacologically induced high activity in primary cultured mouse cortical neurons. To normalize baseline activity across different cultures, we pre-silenced our neural preparations via 24 hours of TTX treatment prior to addition of Bic (**Fig. 5.16, Appendix**

III Methods). We found that the Class 1 loops surrounding *c-Fos* and *Arc* achieved peak contact frequency within 20-60 minutes of neuronal activation (**Fig. 5.16A-B**). We also created total RNA-seq libraries at these time points and observed that enhancer-promoter loop strength for IEGs peaked prior to maximum mRNA levels at 60 min (**Fig. 5.16C-D**). Importantly, at early time points *c-Fos* interacted with an additional enhancer (**Fig. 5.16A**, 'Enhancer 2', magenta arrowhead) compared to its 24-hour activity-induced loop (**Fig. 5.14D**, 'Enhancer 1'), suggesting dynamic engagement with differential activity-induced enhancers over short time scales. We next measured enhancer activity dynamics by quantifying the RNA-seq signal that mapped to each enhancer (eRNAs)[219] (**Appendix III Methods**). We verified that our eRNA analysis approach produced activity-dependent dynamic patterns that resembled a previously published activity-induced eRNA data set[219] and our own H3K27ac ChIP-seq (**Fig. 5.17**). The enhancers that loop to both *c-Fos* and *Arc* peak in activity 20 minutes post neuronal activity, exhibiting lower activity at all other time points (**Fig. 5.16C-D**). Altogether, our data suggests that Class 1 activity-induced enhancers and loops connect rapidly and prior to maximum IEG levels. While we have not determined the full extent to which loops causally drive gene expression, our observation that the rapid activation of enhancer-promoter loops is concordant with the earliest signatures of activity-dependent gene upregulation, and prior to maximum expression levels, supports the assertion that the two are linked.

To test our hypothesis that looping dynamics contribute to the relatively delayed timing of SRG expression (**Fig. 5.14K-N**), we quantified interaction frequency, enhancer activity, and mRNA levels for the Class 1 loops formed by dIEG/SRG *Bdnf* (**Fig. 5.14G-H**). Consistent with our hypothesis, *Bdnf* Class 1 loops did not display looping signals until

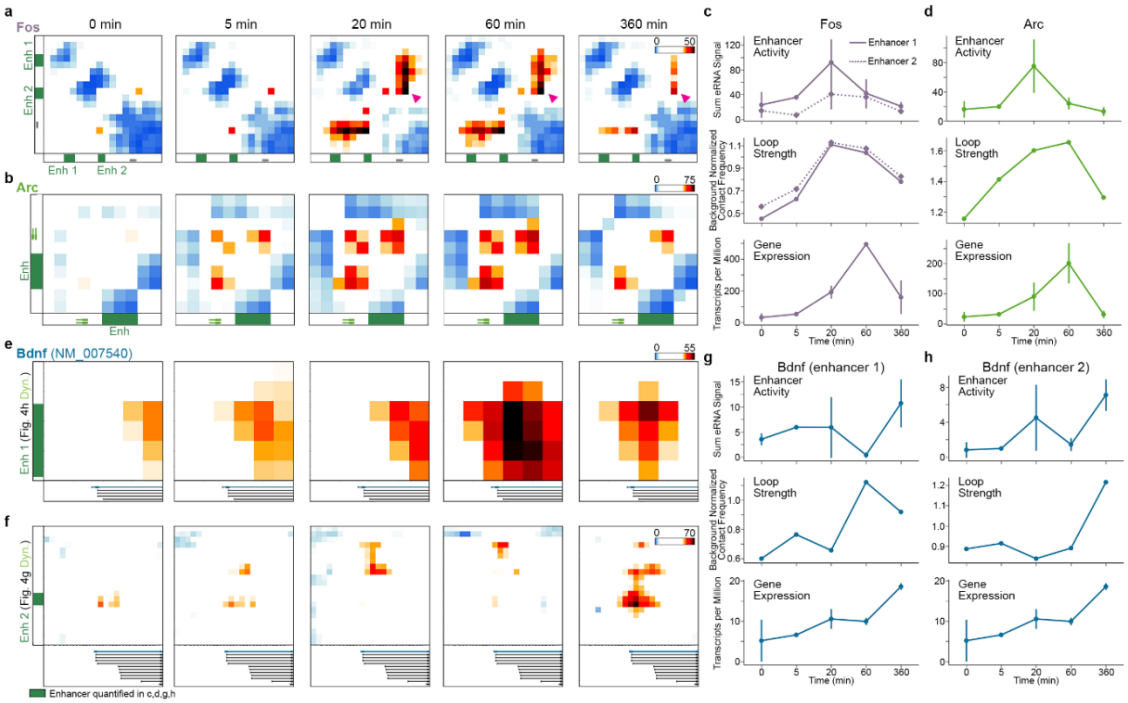


Figure 5.16. Activity-induced loops form before and persist after peak expression of rapid IEGs. (A-B) Interaction score heatmaps surrounding *Fos* (A) and *Arc* (B) across 6 hours of Bic treatment (preceded by 24 hours of TTX silencing). Heatmap coordinates are identical to Figs. 4d (*Fos*) and 4e (*Arc*). Enhancers quantified in (c,d) represented by green boxes. Magenta arrowhead denotes *Fos* loop present only at early time points. **(C-D)** Quantifications of *Fos* (C) and *Arc* (D) enhancer activity (top, quantified by eRNA signal), loop strength (middle, observed/expected 5C counts), and gene expression (bottom, transcripts per million) across the activation time course. **(E-F)** Interaction score heatmaps of activity-induced loops formed by the first *Bdnf* promoter. Heatmap coordinates in (F), "enhancer 2", match those in Fig. 5.14G. Heatmap coordinates in (E), "enhancer 1", represent a zoomed subset of Fig. 5.14H to highlight activity-induced loop. Enhancers quantified in (g,h) represented by green boxes. **(G-H)** Quantifications of *Bdnf* enhancer 1 (G) and enhancer 2 (H) activity (top) and loop strength (middle), coupled with the expression (bottom) of the *Bdnf* isoform with the strongest expression (see Fig. 5.15).

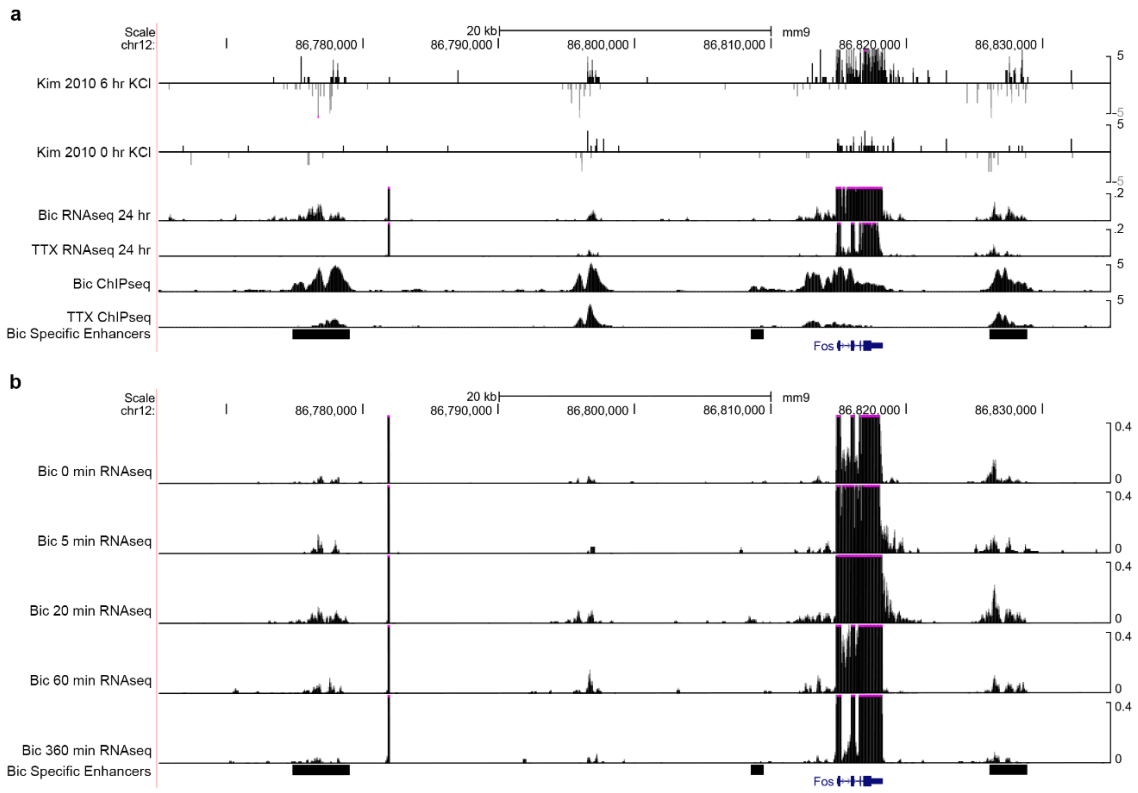


Figure 5.17. Verification of the eRNA signature captures enhancer activity dynamics. (A) Genome browser view of ~50 kb window surrounding the *Fos* gene. Rows from top to bottom present: 1) RNA signal in active neurons from Kim et al. 2010, 2) RNA signal in inactive neurons from Kim et al. 2010, 3) RNA signal from neurons in the Bic condition, 4) RNA signal from neurons in the TTX condition, 5) H3K27ac ChIP-seq signal from neurons in the Bic condition, 6) H3K27ac ChIP-seq signal from neurons in the TTX condition. **(B)** RNA-seq signatures at enhancers near *Fos* across 0, 5, 20, 60, and 360 minutes of acute neuron activation.

60 (enhancer 1, **Fig. 5.14H, 5.16E,G**) or 360 minutes (enhancer 2, **Fig. 5.14G, 5.16F,H**) after activity induction. *Bdnf* enhancers and expression were upregulated in parallel with loops and did not reach maximum signal in our time course until 360 minutes of stimulated activity (**Fig. 5.16G-H**). Thus, *Bdnf* loop and enhancer dynamics are significantly delayed in comparison to *c-Fos* and *Arc* loop dynamics, corroborating our model that looping

structure and dynamics contribute to the delayed expression of SRGs in response to neuronal activation.

Finally, we set out to elucidate whether the long-range 3-D regulatory landscape might give insight into how activity-dependent gene expression could be affected in neuropsychiatric disorders[236]. We investigated the link between our loop classes and common SNVs statistically associated with schizophrenia[237] and autism spectrum disorder (ASD)[238] via genome-wide association studies (GWAS). More than 90% of disease-associated SNVs are localized in non-coding regions with unknown target genes[239], and this has hindered mechanistic understanding of how SNVs might disrupt transcription to cause pathological phenotypes. We identified 24,544 unique loops from published Hi-C data created in human brain tissue derived from the germinal zone and cortical plate[240] (**Fig. 5.18**). We lifted our activity-dependent enhancer classes to the human genome and classified 4,098 Class 2 and 3,822 Class 3 loops from human brain tissue (**Fig. 5.19A**, **Appendix III Methods**). We then assessed if common SNVs for two major neuropsychiatric disease states were enriched in a specific looping class compared to background SNVs matched by the size of the linkage disequilibrium (LD) block, minor allele frequency, distance to nearest gene, and gene density (**Appendix III Methods**)[241]. We found that non-coding SNVs associated with schizophrenia[237] ($P < 5 \times 10^{-8}$) co-localize at Class 3 loops anchored by activity-decommissioned enhancers, whereas ASD-associated SNVs colocalize with Class 2 loops anchored by activity-inducible enhancers[238] ($P < 10{-}4$) (**Fig. 5.19B, Appendix III Methods**). We cross-validated this result using an independent statistical test, LD score regression[242], to quantify the enrichment of heritability for the two diseases within the looping classes. Our LD Score regression analysis confirmed a stronger

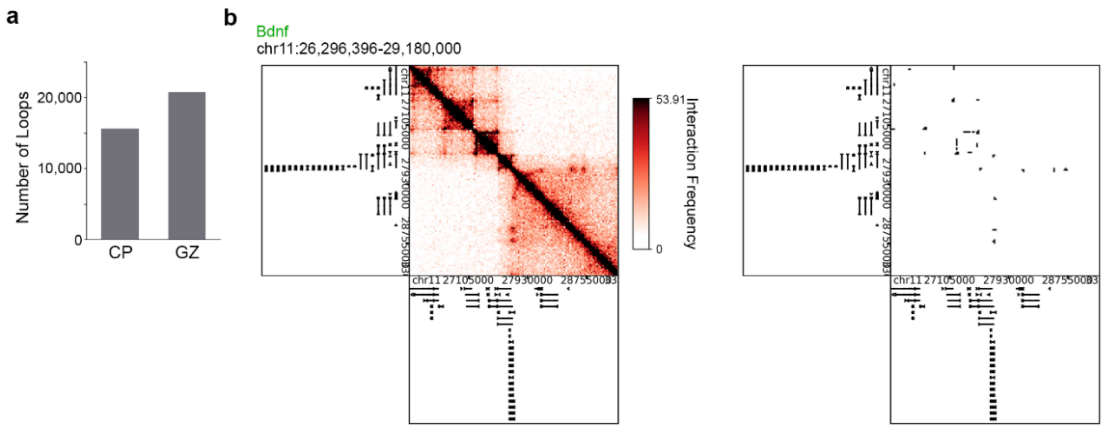


Figure 5.18. Human HiC (Won et al., 2016) loop calls. (A) Number of loops called in HiC data obtained from human fetal cortical plate (CP) and germinal zone (GZ) tissue (Won et al. 2016). **(B)** Interaction frequency heatmap (left) and thresholded loop calls (right) of the 2.5 Mb region surrounding the *Bdnf* gene in human cortical plate (CP) fetal tissue.

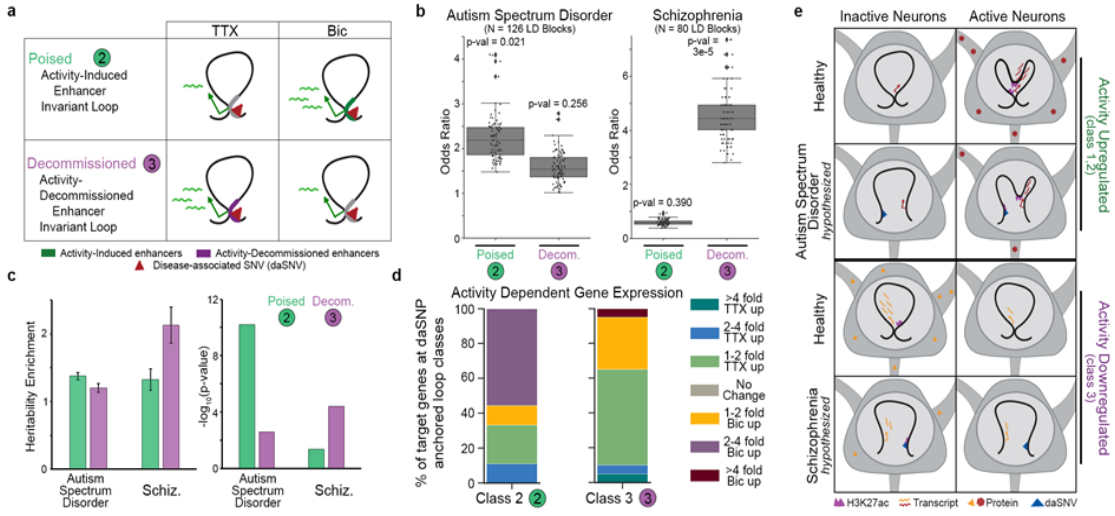


Figure 5.19. Neurodevelopmental disease-associated genomic variants display disease-specific enrichment for activity-induced and -decommissioned enhancer loop anchors. (A) Schematic of Class 2 and Class 3 loop classes computed from human brain tissue Hi-C data reported in Won et al 2016 (Supplemental Methods). **(B)** Odds ratios representing the enrichment of schizophrenia-[237] and ASD-associated[238] common SNVs at the enhancer-containing anchor of each looping class compared to linkage disequilibrium size- and minor allele frequency-matched background SNVs (N=100 sets of background SNVs). tagSNPs which overlap coding regions or could not be matched to background LD blocks were removed prior to analysis. Median Fisher's exact p-values across 100 background sets are included. **(C)** Disease-associated heritability enrichment in each looping class (left) and associated p-values (right), calculated using

LD score regression[242] and summary statistics from ASD and Schizophrenia GWAS studies used in (b). **(D)** Activity-dependent transcription at disease-associated SNV anchored human looping classes plotted as the percent of genes connected to disease-associated SNVs in Class 2 and Class 3 loops that fell within each expression stratum. Expression of the mouse homologs of human genes was used to stratify genes. **(E)** Schematic of our working model of topological regulation in the neuronal activity response. (Row 1) Activity upregulated genes are targeted by activity-induced enhancers in activity-induced (class 1) and activity-invariant (class 2) loops. (Row 2) Autism spectrum disorder SNVs at the base of class 2 loops may disrupt looped enhancer regulation of target gene expression in active neurons. (Row 3) Activity-decommissioned enhancers interact with target genes in invariant looping interactions (class 3). (Row 4) Disruption of looped enhancer function by genome variants associated with schizophrenia at the base of class 3 loops may lead to altered transcriptional control in inactive neurons.

enrichment of ASD-associated heritability in Class 2 loop anchors compared to Class 3, while heritability for Schizophrenia displayed the opposite trend (**Fig. 5.19C**).

We next annotated the genes that contain promoters co-localized to the opposite side loops anchored by disease-associated SNVs (daSNVs). This allowed us to generate a list of long-range candidate genes associated with neural activity for future functional dissection of the effect of common daSNVs. Disease-associated Class 2 loops connect activity-inducible enhancers to target genes that are preferentially *upregulated* upon neural activation (**Fig. 5.19D**). We identified intriguing candidate genes for further future functional inquiry in this set, including *Foxp1*, which has previously been found to regulate brain development and synaptic plasticity[243], displays remarkable mouse to human conservation of local genome architecture, and interacts with several ASD-associated SNVs (**Fig. 5.20A**) and activity-induced enhancers (**Fig. 5.20B-C**). By contrast, disease-associated Class 3 loops connect activity decommissioned enhancers to activity

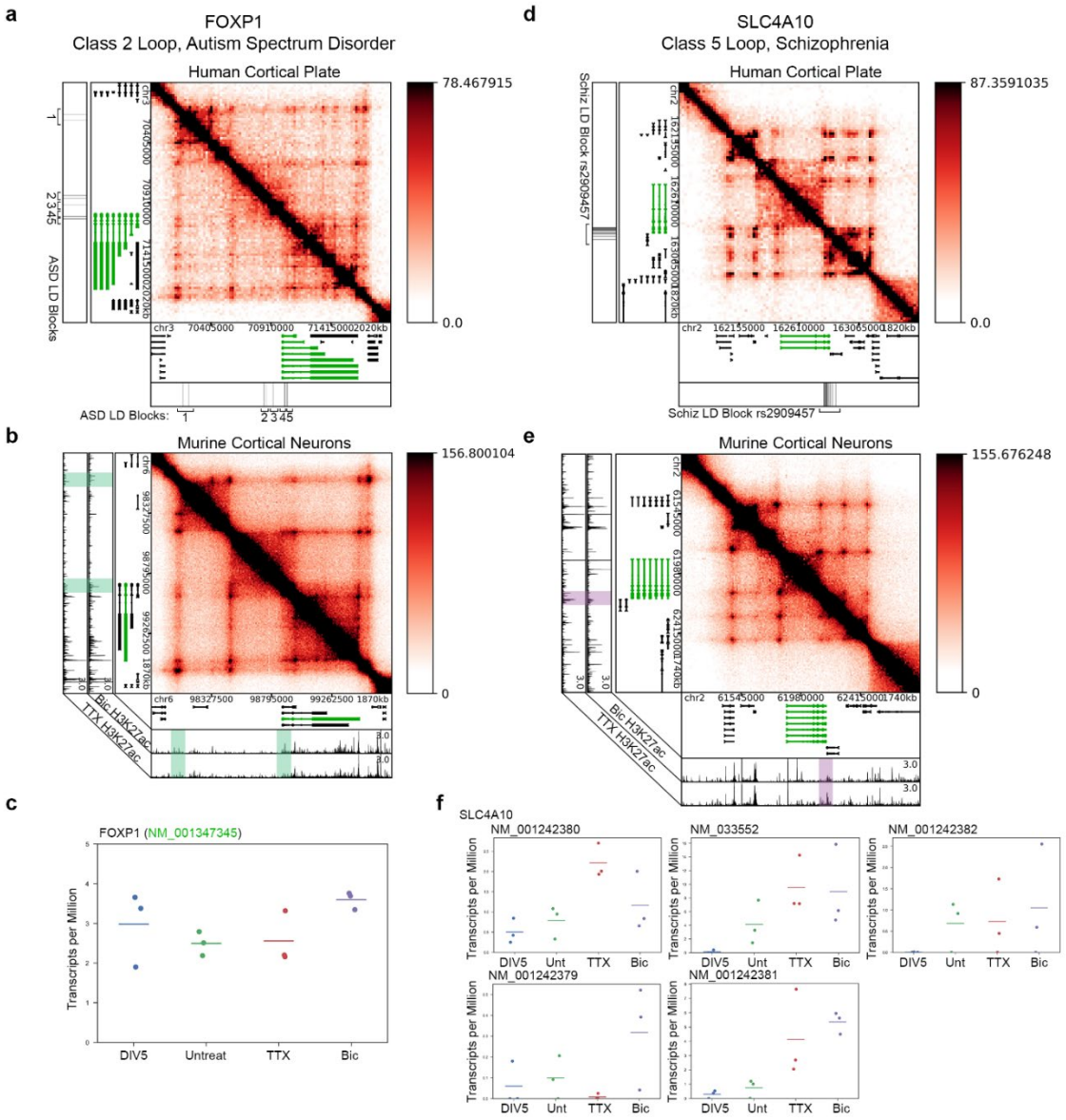


Figure 5.20. ***Foxp1*** **and** ***Slc4a10*** **fall opposite disease-Associated variants in conserved classified loops. (A-C)** Human (A) and mouse (B) interaction frequency heatmaps of a 2 Mb region surrounding the *Foxp1* gene. The expression of the looping Foxp1 isoform labeled in green in (B) is plotted in (C). **(D-F)** Human (A) and mouse (B) interaction frequency heatmaps of a <2 Mb region surrounding the *Slc4a10* gene (green), followed by expression of its 5 expressed isoforms (C).

downregulated target genes (**Fig. 5.19D**). One such gene is *Slc4a1051*[244], which loops downstream to a Schizophrenia-associated SNV (**Fig. 5.20F**) that overlaps a region of

activity-decommissioned H3K27ac (**Fig. 5.20G-H**). Thus, loops anchored by daSNVs predict long-range target genes that change expression in the same direction as their connected activity-dependent enhancers.

In conclusion, our data show that pre-wired and *de novo* loops anchored by activity-inducible enhancers connect to target genes exhibiting activity-dependent upregulation (**Fig. 5.19E, top row**). Conversely, invariant loops anchored by activity-decommissioned enhancers connect to genes that are downregulated upon neuronal activity (**Fig. 5.19E, third row**). Future functional dissection will be required to test the model that (i) common ASD daSNVs will disrupt activity-dependent enhancers or the structure of Class 2 loops, leading to pathologically altered activity-induced target genes (**Fig. 5.19E, second row**) and (ii) common Schizophrenia daSNVs will alter activity-dependent enhancer decommissioning or the structure of Class 3 loops, leading to pathological alterations in the normal activity-dependent downregulation of target genes (**Fig. 5.19E, bottom row**). These data reveal that specific common SNVs associated with neuropsychiatric diseases co-localize with loops anchoring distinct activity-dependent enhancer classes, and these loop classes can connect non-coding daSNVs to unique target genes.

5.3 Discussion

Experience- and activity-dependent gene expression is crucial for sculpting the brain during development and for normal cognition. Here, we show that neuronal activity results in dynamic changes in the 3-D genome that may lead to precise temporal control of activity-dependent gene expression over short and long time scales. We created high-resolution genome folding maps in 12.2 Megabases around IEGs and synaptic genes (total

of N=157 unique transcripts) after multiple time points of acute and chronic exposure to pharmacological agents that activate or inhibit neural activity. We find that >10% of loops in our 5C regions are induced *de novo* during cortical neuron activation. Our identification of numerous activity-induced loops is surprising given that we have previously observed that loops are markedly reconfigured during the developmental transition of ES cells to NPCs, but remain highly similar in the NPC to neuron transition[16]. We observed that most activity-induced loops connecting IEGs to activity-induced enhancers are relatively short-range, and therefore high read depth 5C with a double alternating design was particularly suited for their detection[230, 232]. Future studies focused on genome-wide detection of Class 1 architectural features will require extremely high-resolution maps using Micro-C[109] or high read depth Hi-C created with restriction enzymes that cut four bp restriction sites.

Using chronic (24 hour) neuronal activation and inhibition conditions, we demonstrate that activity-inducible enhancers engage in either *de novo* (Class 1) or pre-existing (Class 2) loops. Class 1 and Class 2 loops connect to genes exhibiting a 24- and 1.3-fold activity-dependent increase in expression, respectively. Our 5C and genome-wide Hi-C results support a working model in which poised/pre-existing loops connected to target genes in advance of activity-induced enhancer activation are abundant in availability but exhibit a modest effect on gene expression. Moreover, our 5C results suggest that loop formation stimulated by activity in parallel with enhancer induction are relatively rare and exhibit a markedly higher effect on activity-dependent upregulation of distal target genes. The quantitative effect of these two looping classes on activity-dependent gene expression levels will be more precisely estimated in the future with genome-wide Hi-C and more diverse activity-induction conditions.

A long-standing question in the transcription field is to what degree enhancer activation and/or looping strength are linked to gene expression. We used our loops and linear epigenetic data in chronic activity inhibition and induction conditions to create simple predictive models of activity-dependent expression changes. We find that H3K27ac signal at distal looped enhancers is a markedly better predictor of activity-dependent target gene expression than nearest enhancers. Additionally, changes in looping strength were only observed in the highest fold-change stratum of activity-dependent gene expression. The ability of our predictive models to explain the variance of activity-dependent gene expression was achieved by building on a critical advance in the functional genomics field. Engreitz et al. published the "Activity-by-Contact" (ABC) model, in which the multiplication of enhancer activity and 3-D interaction frequency was the best predictor of enhancer-target gene pairs[141]. We used the ABC approach to choose a specific enhancer linked to each gene in our model, and this allowed us to prioritize and identify the looped enhancers that most significantly contributed to activity-dependent gene expression. Together, these data suggest that enhancer-target gene prediction would be facilitated by the use of chromatin architecture maps, instead of relying on the enhancer that is closest on the linear genome.

An important area of active research in neurobiology is focused on elucidating the molecular mechanisms by which the differential kinetics of IEGs and SRGs are regulated. Here, we unexpectedly observed that rapid-response IEGs *Arc* and *Fos* connect to enhancers via singular short-range loops that occur *de novo* upon activation. By contrast, we observed that dIEGs/SRGs such as *Bdnf* connect to multiple activity-inducible enhancers via a complex network of invariant and *de novo* loops. We furthered this model

by demonstrating with genome wide Hi-C data that rIEGs form fewer, shorter loops compared to more complex looping architectures formed by dIEGs/SRGs genome-wide. Consistent with our observations, Yamada et al. reported, using H3K4me3 PLAC-seq, that the delayed IEG *Nr4a3* engages in multiple long-range contacts after neuronal stimulation[225]. These observations inspired our working hypothesis that looping complexity and distance are contributing factors to the timing of IEG/SRG activity-induced expression (**Figure 5.14M**). To critically assess this model, we induced acute pharmacological activation of neuronal activity and gathered looping, epigenetic, and transcription data across multiple short time points. We observed striking differences in loop and enhancer induction kinetics for rIEGs vs. dIEGs/SRGs in our 5C regions. For example, the activity of the enhancers and loops surrounding *Fos* and *Arc* peak in signal strength roughly 20 minutes after the induction of neuronal activity, prior to maximum mRNA levels. In contrast, *Bdnf* loops and enhancers gain strength in parallel with mRNA levels over a longer time of sustained activity (360 minutes). We note that, in our study, *Bdnf* is primarily transcribed from its first promoter in response to Bicuculline. However, transcripts initiated from *Bdnf*'s fourth promoter were highly expressed in previous studies using KCl for activation[245], which raises the exciting possibility that different mechanisms of neuronal activation might engage different loops and enhancers. Finally, we also note that *Fos* engages in different short-range loops at 5 minutes versus 20 minutes versus 24 hours of neural activation, shifting interaction strength from a nearby enhancer to one more distal, suggesting that rapid activity-induced enhancer switching via alternative looping might be a mechanistic aspect of rapid IEG upregulation. Together, these data suggest that the 3-D epigenome regulates activity-dependent gene expression across vast (>1 Mb)

genomic distances to ultimately control IEG and SRG expression levels with tight temporal precision.

Finally, the exploration of the link between looping and common SNVs associated with neuropsychiatric disorders is a critical area of inquiry. It is well-established that the large majority of SNVs associated with neuropsychiatric disorders via genome-wide association studies (GWAS)[236, 246-249] are localized to non-coding elements distal from genes[239, 250]. An increased understanding of how activity-dependent enhancers co-localize with disease-associated SNVs and connect over vast distances to distal target genes would provide critical new insight into the molecular mechanisms governing disease pathogenesis. Here, we identify a unique set of loops that are pre-existing before stimulation but anchored by enhancers that decrease in activity during chronic activation conditions. We speculate that enhancer decommissioning may be an epigenetic mechanism involved in homeostatic plasticity. Consistent with this hypothesis, we found that specific genes involved in homeostatic plasticity, such as *Gria1,* are connected in Class 3 loops to activity-decommissioned enhancers and downregulated during chronic high activity. We find that Schizophrenia SNVs are anchored in Class 3 loops and connected to downregulated genes upon synaptic activity. By contrast, Autism SNVs are anchored in Class 2 loops to activity-inducible enhancers and connected to activity-upregulated target genes. These results are striking as they suggest that non-coding SNVs may have very different effects in neuropsychiatric disorders depending on the class of loops that they anchor (**Figure 5.19E**). Moreover, the co-localization of Schizophrenia SNVs with Class 3 loops suggests that defects in enhancer decommissioning might contribute to synaptic plasticity defects in neuropsychiatric diseases[251]. Future work to build human activity-

dependent loop and enhancer maps and dissect their functionality with genome editing will continue to refine our observations of distinct activity-dependent architectural features associated with neuropsychiatric disorders.

5.4 Acknowledgements

Thank you to co-authors Dr. Elissa Pastuzyn, Lindsey Fernandez, Dr. Michael Guo, Kelly Feng, Katelyn Titus, Harshini Chandrashekar and Dr. Jason Shepherd for their critical contributions to this chapter. Thank you to all members of the Cremins lab for helpful discussions.

CHAPTER 6: SUMMARY AND FUTURE DIRECTIONS

6.1 Summary

The three-dimensional conformation of the genome is directly linked to spatiotemporal control of gene regulation during mammalian cellular development[252] (reviewed in **Chapter 2**). Recent technological advances combining Chromosome-Conformation-Capture(3C)-based technologies with deep sequencing have enabled dramatic advances in the throughput and length-scale of regulatory connection mapping[163]. The overall objective of this thesis was to apply these cutting-edge approaches to study the mechanisms governing the restructuring of fine-scale chromatin architecture across critical stages of mammalian central nervous system (CNS) development. My initial hypothesis was that dynamic chromatin loops within topologically associating domains (TADs) connect epigenomic regulatory features that have critical roles in mammalian brain development and neurodevelopmental disease. This hypothesis originated from preliminary data and previously published works showing that (i) large-scale TADs are predominantly invariant across cell types and anchored by invariant binding of the architectural protein CTCF[7]; (ii) dynamic interactions of cell type specific enhancers occur within TADs[9]; (iii) knockout of CTCF at early[167] and late[37, 168] stages of neurogenesis resulted in disruption of neural progenitor cell (NPC) division, tissue architecture, and synaptic connections; (iv) genome folding within TADs is noticeably dynamic across embryonic stem cells (ESCs), NPCs, and induced pluripotent stem cells (iPSCs); (v) CTCF binding is dynamic across some models of neural development. Through the development and/or use of *in vitro* cellular models, 5C, HiC, RNA-seq, ChIP-seq and a suite of

computational tools, this work has begun to shed light on the dynamic three-dimensional genome folding landscape which regulates stages of mammalian brain development.

The departure from pluripotency and commitment to the neural lineage is a critical cellular state decision point during mammalian brain development. In Chapter 3 we found that this cellular state transition was accompanied by a dramatic decrease in CTCF expression, protein levels, and number of genome binding sites. At the conclusion of this work it remained a critical unknown whether this trend continued or reversed as development progressed into terminally differentiated neurons. Through re-analysis of recently published data[104] we have now confirmed that this trend continues in murine cortical neurons (**Figure 6.1**); the number of CTCF binding sites in cortical neurons are notably decreased even beyond the NPC level (**Figure 6.1A**). The decrease in CTCF sites correlates with an increase in overall loop length in cortical neurons (**Figure 6.1B**), which is particularly evident when comparing loops specific to cortical neurons ('CN-only') compared to those present in other cell types (**Figure 6.1C**). The anti-correlation between CTCF site number and loop length led us to hypothesize that dynamic CTCF sites that formed a boundary of smaller, ESC-specific domains were being 'pruned', allowing loop extrusion to continue unimpeded for longer genomic distances leading to longer, neural-specific loops, as we had previously observed at Sox2 (**Figures 3.13, 3.14, 3.18G-I**). Indeed, greater than 60% of CN-only loops (filtered for those >200 kb) shared a looping anchor with a shorter ES-only or ES-NPC loop (**Figure 6.1D**). At these loci the majority of both the shared looping anchors and CN-specific looping anchors contained CTCF peaks which were bound in all 3 cell types (**Figure 6.1E**, 'constitutive' CTCF peak class). Conversely, the ES-specific looping anchors consistently contained CTCF peaks that were

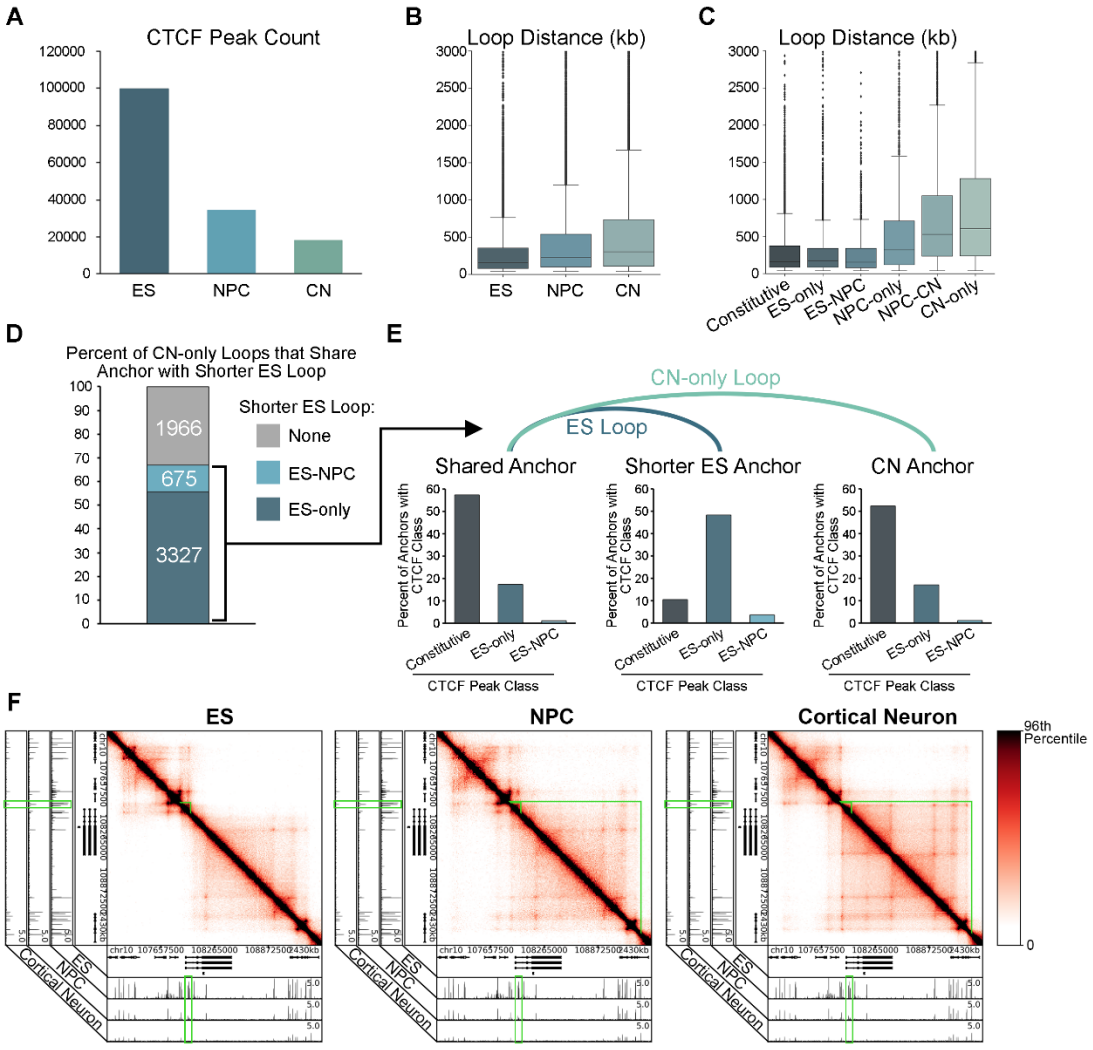


Figure 6.1. Loop length increases across neuronal differentiation due to CTCF site pruning. (A) Number of CTCF peaks called (p-value < 1e-4) across embryonic stem (ES) cells, neural progenitor cells (NPCs), and cortical neurons (CN). Data analyzed from Bonev et al. 2017. **(B)** Boxplots of loop distance for each cell type. **(C)** Boxplots of loop distance for each loop class, parsed by cell types in which the loop was called significant. **(D)** Percent of CN-only loops (> 200 kb in length) that share an anchor with a shorter ES loop. **(E)** Percent of each loop anchor that contains each class of CTCF peak, parsed by cell-type presence. **(F)** Relative interaction frequency heatmaps surrounding the *Synaptotagmin-1* gene. CTCF tracks for each cell type plotted below heatmaps. Green boxes highlight constitutive CTCF site (y-axis), ES-specific CTCF site (x-axis), ES-specific loop (small overlaid on heatmap), and CN-specific loops (large overlaid on heatmap).

deactivated during early neural lineage commitment (**Figure 6.1E**, 'ES-only' CTCF peak class). Altogether this data supports a model of dynamic genome folding during neuronal differentiation similar to what we observe at the *Synaptotagmin-1* gene (**Figure 6.1F**): (1) the abundance of CTCF in ES cells results in many small contact domains (small green box on heatmap), (2) during neuronal development, CTCF sites at specific domain boundaries are decommissioned (green box, x-axis), (3) smaller ES-specific contact domains dissolve as ES-specific CTCF sites are decommissioned, allowing loop extrusion to proceed farther and connect constitutively-bound CTCF sites (green box, y-axis) in neural-specific loops (large green box overlaid on heatmap). In this way CTCF site pruning is a critical process in establishing the chromatin landscape that is necessary for proper mammalian neural development.

CTCF site pruning has far reaching implications for how the neuronal chromatin landscape functions to regulate genes in a proper spatiotemporal manner. At the highest level, the increase in neural loop/domain size allows a large number of NPC-specific (Chapter 3) and neuronal activity-induced (Chapter 5) enhancers to regulate their target genes over vast genomic distances; it is presumably safe to assume this model extends to most subsets of enhancers that operate along the neural lineage. Due to the decrease in CTCF binding in neural progenitor cells, we found NPC-specific enhancers were increasingly reliant on YY1 to operate as an architectural protein[17] to connect them to their target genes (Chapter 3). Similarly, the activity-induced enhancers we identified in response to neuronal activity did not exhibit CTCF binding (**Figure 5.14**). Thus, our working model suggests that neural cell types rely on an additional suite of architectural

proteins to connect enhancers to their target genes more than pluripotent stem cells and perhaps more than other developmental lineages.

Furthermore, in Chapter 4 we found that the CTCF sites that were pruned during neural lineage commitment were not completely restored in NPC-derived induced pluripotent stem (iPS) cells. Improperly reprogrammed CTCF sites in iPS cells fell at the base of incompletely reprogrammed genome architecture, which correlated with disrupted expression of key pluripotency genes (**Figures 4.9, 4.10, 4.12, 4.13**). Culturing iPS cells in 2i media conditions was sufficient to restore CTCF binding at pruned sites, resulting in genome folding and gene expression profiles that more closely resembled those of mouse embryonic stem cells. Altogether this work further confirmed that precise regulation of CTCF levels is necessary for the establishment of cellular gene expression programs and re-establishing target CTCF levels can act as a roadblock during cellular reprogramming[195].

Finally, in Chapter 5 we investigated the manner in which neuronal activity-induced enhancers leverage the 3-D genome to regulate activity response genes. Although we found that activity-induced enhancers are often poised near their target in invariant looping interactions, those target genes were on average only modestly upregulated in active neurons (**Figure 5.10**). Genes that were robustly upregulated (*Fos, Bdnf, Arc*), in addition to forming such poised loops, also dynamically looped to enhancers in an activity-dependent manner (**Figure 5.14**). Surprisingly the kinetics and complexities of these loops differed when comparing rapid response genes (*Fos, Arc*) to delayed response genes (*Bdnf*); *Bdnf* forms many more loops which span vast genomic distances (> 1 Mb) and form slower in response to activation than the dynamic loops that *Fos* and *Arc* form. Finally, we investigated the enrichment of heritability of neurodevelopmental diseases at the base

human looping interactions that contained activity-induced or activity-decommissioned enhancers. Due to the well-established function of activity-induced and activity response genes in regulating the proper synapse formation underlying memory and cognition, our hypothesis was that neurodevelopmental disease heritability would be enriched at loops with activity-induced enhancers. While this was the case for Autism Spectrum Disorder (ASD), we were surprised to identify a strong enrichment for the heritability of Schizophrenia at activity-decommissioned enhancers instead (**Figure 5.19**). Altogether this chapter links architectural complexity to transcriptional kinetics and reveal the rapid time scale with which the 3-D genome folds during synaptic plasticity.

In combination the results in this thesis reveal that the neural genome landscape has a very distinct folding signature. A decrease in CTCF binding and expression of the cohesin unloading complex WAPL[104] establish very large contact domains and increase the distances over which developmentally dynamic enhancers can loop to their target genes. Indeed a general theme throughout this work is that identifying the focused puncta of dynamic chromatin architecture within larger contact domains enables focused identification of enhancers and genes that regulate a particular neural cellular state. One principal goal of this approach is to identify particular epigenomic features linked to genetic diseases and connect those features and genetic variants to target genes, which could in turn be targets for therapeutic intervention. The conclusion of this work (**Figure 5.19**) used Schizophrenia and ASD genetic variants to show that insight is indeed gained by analyzing these variants in the context of human brain chromatin loops and activity-dynamic enhancers. At the conclusion of this work it is my strong belief that further

investigation into the 3-D epigenomic bases of neurodevelopmental diseases will lead to critical, fundamental insights into our understanding and treatment of these diseases.

6.2 Limitations and Future Directions

It is important to highlight one key limitation of these results as a whole: a lack of precise genome-editing experiments that directly test the causal influence of the loops and enhancers identified. This is especially important in the case of testing the function if activity-decommissioned enhancers in the pathogenesis of Schizophrenia. Chapter 5 is the first work to my knowledge that has identified activity-decommissioned enhancers as potentially critical to proper brain development, meaning these enhancers have never been studied in depth. Thus, an important next step is to use CRISPR-Cas9 genome editing and/or epigenetic editors like the CRISPRi system to perturb activity-decommissioned enhancer activity and loop presence in developing, *inactive* neurons and look for resulting changes in gene expression. A thorough investigation will connect epigenetic and transcriptional perturbations to alterations in how synapses form and function, thus leading to changes in properties of neuronal networks and behavioral changes in an animal model. Because our initial findings suggest these enhancers and loops play a role in homeostatic scaling of synaptic strength (**Figure 5.10**) and may be dysregulated in Schizophrenia (**Figure 5.19**), I propose that such follow-up studies have a high probability of revealing foundational insights into how the 3-D epigenome directs mammalian brain development, memory and cognition.

A second tantalizing observation that requires further exploration resides within the deactivation kinetics of the loops that *Fos* and *Arc* form (**Figure 5.14**). We note that the

enhancers in these loops peak in activity at 20 minutes post neuronal activation, which is also when loop strength peaks and prior to max mRNA levels. First, it is important to test the dynamics of nascent transcription in this system using a method like PRO-seq[253], because it remains possible that nascent transcription also peaks at 20 minutes post activation but mRNA continues to accumulate between the 20 and 60 minute timepoints. Additionally we observe that while enhancer activity returns to near baseline by 60 minutes post stimulation, loop strength remains high at the same timepoint and even remains elevated above baseline at 360 minutes post stimulation. This raises the exciting possibility that slow loop decommissioning kinetics may retain an epigenetic 'memory' of past activation events so that neurons are primed for a stronger/faster upregulation of activity response genes upon subsequent stimulation events. To test this hypothesis, I propose an experimental paradigm in which after 360 minutes of activation, activation cues (Bicuculline, KCl and/or Bdnf) are temporarily removed and TTX is added to inactivate the culture on the time scale of minutes to hours. During this inactivation time, enhancer, loop and gene expression inactivation kinetics should be mapped, with the goal of identifying a time point at which enhancer activity and gene expression have fully returned to baseline but residual loop strength remains. If this time point exists, neurons should then be re-stimulated to test the hypothesis that the higher residual loop strength primers the enhancer to activate *Fos* and/or *Arc* on a more rapid timescale or to a higher peak expression value. The results of these experiments have the potential to implicate chromatin loop activation and deactivation as a tool for each neuron to record past activation events within the nucleus and thus integrate the effects of multiple activation

events across time. Such a finding would have large-scale implications for our understanding of the molecular underpinnings of human cognition and memory.

APPENDIX I: METHODS ASSOCIATED CHAPTER 3

Embryonic Stem (ES) Cell Culture

V6.5 ES cells from Novus Biologicals (NBP1-41162) were cultured as previously described [156] under standard pluripotent (serum/LIF) conditions on Mitomycin-C inactivated MEFs. To generate the 2i/LIF condition, ES cells were transitioned to serum-free media containing 3 uM CHIR99021 (Axon Medchem #1386), and 1 uM PD0325901 (Axon Medchem #1408) (as described in [156]) and propagated for 2 passages on feeder cells. Before fixation, both ES cell conditions were passaged onto 0.1% gelatin coated plates to remove the feeder layer, and fixed at ~60% confluency. Thus, the 2i/LIF ES cells were cultured for 3 passages under 2i/LIF conditions before fixation.

Primary Neural Progenitor Cell (NPC) Culture

Neural progenitor cells were cultured as previously described [156]. Briefly, NPCs were cultured as neurospheres for two passages to purify the population of non-adherent NPCs. Neurospheres were then dissociated and passaged onto Poly-D-Lysine Hydrobromide (100 ug/mL, Sigma P7280), and laminin (15 ug/mL, Corning 354232) coated plates, and fixed next day.

CTCF ChIPseq

Chromatin immunoprecipitation was performed as previously described [156]. Libraries were prepared for sequencing using the NEBNext Ultra Library Prep Kit (NEB #E7370) and following the manufacturer's protocol for ChIP-seq library preparation. No size selection step was performed following adapter ligation. The libraries were amplified over 18 PCR

cycles using NEBNext Multiplex Oligos for Illumina (NEB #E7335). The final ChIP libraries were eluted in 30 uL 0.1x TE from the Agencourt AMPure XP beads, at which point we confirmed the library contained DNA fragments ranging from 250 to 1200 bp, including the adapters, by running a High-Sensitivity DNA assay on an Agilent Bioanalyzer. The concentration of these libraries was assayed via the KAPA Illumina Library Quantification Kit (#KK4835), diluted to equivalent concentrations and pooled, and finally sequenced with 75-cyles per paired-end on the Illumina NextSeq500.

ChIP-seq peakcalling

Published ChIP-seq data was downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/) and reanalyzed according to. Reads were aligned to mouse genome build mm9 using Bowtie with default parameters [254]. Reads were considered if they had two or fewer reportable alignments. To facilitate the comparison of ChIPseq libraries across cell types, the mapped reads were filtered to remove optical and PCR duplicates and then downsampled to equivalent read numbers across cellular states. The CTCF ChIP libraries for ES 2i, ES serum and pNPC were downsampled to 11 MM reads and the whole cell extract libraries were downsampled to 15 MM reads. For YY1 ChIPseq libraries, the ES serum, ProB, and pNPC samples and inputs were downsampled to just over 7 MM reads. The H3K27ac ChIP libraries for ES serum and pNPC were downsampled to 7 MM reads and the whole cell extract libraries were downsampled to 7 MM reads. Peaks were identified using Model-based Analysis for ChIP Sequencing v2.0 (MACS2) [255]. For CTCF ChIPseq, default parameters were used with a p-value cutoff of $p < 1E-8$. For YY1, we modified the parameters to facilitate accurate detection of broad peaks (--broad --broad-

cutoff 1E-4 -p 1E-8). For histone modification H3K27ac ChIPseq, the same broad peak calling approach was utilized.

Parsing Cell Type-Specific CTCF Occupancy Sites

CTCF ChIP-seq peaks ($p < 1x10^{-8}$) were utilized to parse CTCF sites into cell type specific occupancy classes with Galaxy. 'ES 2i only' CTCF peaks were defined as CTCF sites that were present in ES cells under 2i/LIF conditions and the absence of CTCF in ES cells in serum/LIF conditions and in NPCs. This class was generated using Galaxy to subtract ES serum and NPC CTCF peaks ($p < 1x10^{-8}$) from ES 2i CTCF peaks. Similarly, 'ES serum only' CTCF was defined by the presence of CTCF in serum/LIF ES cells and the absence of CTCF in ES cells in 2i/LIF conditions and in NPCs; 'NPC only' CTCF was defined by the presence of CTCF in NPCs and the absence of CTCF in ES cells in 2i/LIF and serum/LIF condition. '2i+serum' CTCF was defined by the presence of CTCF in ES cells in 2i/LIF and serum/LIF conditions and the absence of CTCF in NPCs. This class was generated via the intersection of ES 2i CTCF sites with ES serum CTCF sites, followed by the subtraction of NPC CTCF sites. 'Serum+NPC' and '2i+NPC' CTCF sites were similarly parsed. Finally, 'Constitutive' CTCF was defined by the presence of CTCF in ES cells in 2i/LIF and serum/LIF and in NPCs.

siRNA Knockdown of YY1 in pNPCs

pNPCs were cultured as described above. After two passages in suspension, cells were seeded adherently at a density of 20,000 cells/cm2. In order to allow cells to reach a critical density before the start of transfection, 40 hours were allowed to pass between seeding and

the application of siRNA. The following siRNA pools were purchased from Dharmacon: YY1 (# L-050273-00-0005), Non-targeting Pool (# D-001810-10-05). Cells were transfected with a final concentration of 20 nM siRNA. RNAimax (Lifetech #13778-075) was used as a transfection reagent at 1/3 the recommended concentration (2.5 uL per well of 6 well plate, 14.5 uL per 10 cm dish). Reagents were prepared in Optimem according to RNAimax manufacturer's instructions and then added dropwise to culture well/dish. Transfection continued for 78 hours, with media and transfection reagents replaced at hours 24 and 48 after start of transfection. After 78 hours, cells were harvested for RT-qPCR, Western blot, and 3C/5C.

In situ 3C

pNPCs subjected to siRNA transfection were fixed with formaldehyde for 3C as previously described [156]. 4 million cells were utilized per replicate and subjected to an *in situ* 3C protocol adapted from [10]. Cell pellets were resuspended in lysis buffer consisting of 10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 0.2% Igepal CA630 and 1x protease inhibitor and incubated with frequent agitation on ice for 20 minutes. Nuclei were washed twice with 1.2X NEBuffer. SDS was added to a final concentration of 0.3% and the homogenate was incubated for 1 hr at 37°C. SDS treatment was inactivated by the addition of 20% Triton X-100 to a final concentration of 1.8% and incubation at 37°C for 1 hr. Chromatin was digested with HindIII (300U) overnight at 37°C then 65°C for 30 minutes. Chromatin was then ligated upon the addition of ligation buffer components at final concentrations of: 0.83% Triton X-100, 1X BSA, 1mM ATP, 50mM Tris-HCl, 50mM NaCl, 10mM $MgCl_2$, 1mM DTT and 15 uL of T4 DNA ligase (Invitrogen). The ligation reaction occurred at

16°C for 4 hours and then at room temperature for 30 minutes. Finally, samples were Proteinase-K digested, RNase treated, phenol-chloroform extracted with ethanol precipitation and resuspended in 1X TE buffer. 600 ng of 3C template was utilized for 5C as described in [9, 156].

Gene expression quantification via RT-qPCR

RNA isolation was done using the mirVana miRNA isolation kit (Lifetech #AM1560), following manufacturers protocol for total RNA isolation. Cells were lysed in mirVana supplied lysis buffer and stored temporarily at -20°C until all samples were collected. Volume of lysate utilized in organic extraction was adjusted to contain the lysate from 500,000 cells. Manufacturer's protocol was then followed precisely. cDNA was prepared for each sample using the SuperScript First-Strand Synthesis System (Lifetech #11904-018) according to manufacturer's specifications. 100 ng of RNA, quantified via Qubit, was loaded into each reaction. The following primers were designed to query relevant gene expression:

YY1: F: CACGCTAAAGCCAAAAACAACC ; R: ATTCCCAATCACACTCCTGAAG

Sox2: F: GCACATGAACGGCTGGAGCAACG ; R: TGCTGCGAGTAGGACATGCTGTAGG

Olig2: F: GCAGCGAGCACCTCAAATC ; R: GATGGGCGACTAGACACCAG

Nestin: F: AGGCCACTGAAAAGTTCCAG ; R: TAAGGGACATCTTGAGGTGTGC

Zfp462: F: CAAAGCCCATGCTGGTGAAC; R: TTTGCCATGGACCTTGAGGG

Klf4: F: AGACCAGATGCAGTCACAAGTC ; R: TTTTGCCACAGCCTGCATAG

Standard curves for each primer set above were generated by quantifying the product of a conventional PCR reaction and serially diluting the amplicon to create 200 – 0.0002 pM standards. qPCR reactions were performed on the Applied Biosystems StepOnePlus system using the Power SybrGreen PCR Master Mix (Applied Biosystems #4364659). For each qPCR reaction, primers were added to a final concentration of 400 nM and 1 uL of each standard and sample cDNA was loaded. The resulting CT values of the standards were used to generate a standard curve and calculate the concentration of transcript cDNA per 100 ng of RNA loaded into the first strand reaction.

Western blotting

Cells for each condition were washed with ice-cold PBS and lysed in RIPA buffer (Sigma R0278, ~100 uL per 1 million cells). Cells in RIPA were scraped off of the dish and rotated for 30 min at 4°C. Samples were then spun for 20 minutes at 12,000 rpm and 4°C, after which the supernatant was stored at -20°C until further use and the pellet was discarded. Total protein content was estimated by BCA assay (Thermo scientific #23227) in order to target equal total protein loading. Sample to be loaded was then diluted in 4X Laemmli buffer (BioRad #161-0747) and 2-mercaptoethanol (final concentration 355 mM). Samples were run through a BioRad TGX 4-15% gel (#456-8084) and transferred to an LF-PVDF membrane using the BioRad TransBlot Turbo transfer system. After transfer, membranes were washed twice with TBS, then blocked for 1 hour in 3% BSA in TBS at room temperature. The membrane was incubated with primary antibodies (CTCF=Cell Signaling #3418 at 1:200, YY1= Santa Cruz #sc-1703 at 1:50, Gapdh=Cell Signaling #2118 at 1:1000) in 3% BSA in TBS/T overnight at 4°C under constant agitation, then at room

temperature for 10 minutes. 3 washes in TBS/T were performed before incubation with secondary antibody (anti-Rabbit Dylight 650, abcam #ab96894) in 3% BSA in TBS/T at room temperature for 1 hour. Finally, blots were imaged on the ChemiDoc MP Imaging system after 3 washes in TBS/T.

5C data analysis

Technical note on preliminary processing of two analysis groups

Two sets of 5C data, group1 and group2, were processed independently for this study. Group1 represents a re-analysis of raw reads from previously published 5C experiments [156] and consists of ES 2i (n=2 replicates), ES Serum (n=2 replicates) and pNPC (n=2 replicates) conditions. Group2 5C libraries were generated in the present study and consist of YY1 siRNA treated pNPCs (n=2 replicates) and scrambled siRNA treated pNPCs (n=2 replicates). These 5C replicates were sequenced on the Illumina NextSeq 500 with 37 bp paired-end reads and then aligned to a pseudo-genome of the 5C primer set using Bowtie with default parameters [254]. To be considered a count for downstream processing, reads were required to: (i) have only one unique alignment, (ii) have both paired-ends map to the pseudo-genome, (iii) represent an interaction between one forward and one reverse primer. Before downstream analyses, mapped 5C reads were trimmed of entire primers if the total counts sum of that primer was less than 10 or the primer was visually identified as low quality. Group1 data were high quality/high complexity. Preliminary analysis of Group2 revealed a high level of spatial noise likely due to technical artifacts caused by suboptimal ligation for these particular libraries. Although we provide sequencing reads for all our queried 5C regions for Group2, we only publish downstream processing and analysis in

Group2 for the Sox2 and Klf4 regions, as these were highest complexity regions resembling our high quality NPC maps obtained from Group1. Thus, for Group2 data sets, the 5C primers for all regions other than Sox2 and Klf4 were removed before assembling primer-primer junction counts files. Group1 5C libraries were processed separately from Group2 5C libraries. 5C libraries were analyzed as detailed below. Custom scripts for all of the analysis steps are provided as supplemental material for full reproducibility of figures.

Quantile normalization

To account for sequencing depth and technical complexity differences among libraries, 5C replicates were conditionally quantile normalized. Briefly, the GC content of each 5C primer was calculated. Each primer-primer pair could then be assigned a pair of GC content values based on the two constituent primers. Primer-primer pairs with the *same* GC content pair were grouped. Within each group, counts for primer-primer pairs were quantile normalized across replicates as previous described [156]. Counts of the same starting value (i.e. a tie) were assigned the average value of the *lowest* rank in the set of tied counts. Group1 and group2 data were quantile normalized separately.

Primer correction

To account for known primer-specific biases in our 5C data, we applied a modification of the published Express algorithm in which we computed joint bias factors by using counts data from all replicates [256]. Group1 and group2 data were primer corrected separately.

Removal of low confidence primer-primer pairs

Primer-primer pairs were removed from downstream analyses if they did not register at least 10 normalized reads in at least 3 of the replicates (group 1) or if they did not register at least 5 normalized reads (group 2).

Interaction matrix binning

We divided each of our queried regions into adjacent 4 kb bins because 4 kb is roughly the average restriction fragment size after HindIII digestion. Each entry of the binned interaction frequency matrix represents the relative frequency with which two 4 kb bins interact. The relative interaction frequency in each bin was set as the arithmetic mean of the normalized, logged primer-primer pair reads that mapped to within a 16 kb (Group1) or 20 kb (Group2) square smoothing window surrounding the coordinates of the midpoints of the two bins.

Removal of low information content bins

Interaction frequency matrix entries were set to 'NaN' and thus removed from downstream processing if the number of primer-primer pairs within the smoothing window of that matrix entry that were 'NaN' or zero exceeded 80% of the possible primer-primer pairs.

Expected background modeling

To evaluate looping interactions, we employed slight modifications of the donut and lower left background models recently developed by the Aiden group [10]. This approach requires a global distance-dependence model, which we generated by first computing the arithmetic

mean of the interaction frequency matrix entries that represent interactions of equivalent genomic distance. For the shortest 1/3 of interaction distances queried we used the empirical mean as the distance-dependent expected; for the remaining interaction distances we calculated a lowess fit to the empirical means and utilized each fit value as the distance-dependent expected. Global expected values were 'corrected' for local background interaction frequencies through the use of donut and lower left background filters specific to each entry in the binned interaction frequency matrix. The 'Donut' correction was applied according to (1):

$$E_{ij}^{d} = \frac{D_F(i,j)}{D_E(i,j)} \times E_{ij} \quad (1)$$

where E_{ij} is the global distance-dependence expected interaction frequency of bins i and j, and $D_F(i,j)$ and $D_E(i,j)$ are evaluations of a function 'D' over the interaction frequency matrix F and the global distance-dependence expected matrix E, respectively. The function 'D' finds the sums of the values falling within the donut window for the entry (i,j) of the matrix of interest (represented here as 'A') with chosen parameters p and w (2):

$$D_A(i,j) = \sum_{x=i-w}^{i+w} \sum_{y=j-w}^{j+w} A_{xy} - \sum_{x=i-p}^{i+p} \sum_{y=j-p}^{j+p} A_{xy} - \sum_{x=i-w}^{i-p-1} A_{xj} - \sum_{x=i+p+1}^{i+w} A_{xj} - \sum_{y=j-w}^{j-p-1} A_{iy} - \sum_{y=j+p+1}^{j+w} A_{iy} \quad (2)$$

The 'Lower Left' correction was applied according to (3):

$$E_{ij}^{ll} = \frac{LL_F(i,j)}{LL_E(i,j)} \times E_{ij}$$

(3)

where the *LL* function for a matrix *A* is defined as in (4) :

$$LL_A(i,j) = \sum_{x=i-w}^{i-1} \sum_{y=j-w}^{j-1} A_{xy} - \sum_{x=i-p}^{i-1} \sum_{y=j-p}^{j-1} A_{xy}$$

(4)

A schematic of the donut and lower left windows defined by these functions is shown in **Fig. 3.5B**. Eqn. 1 generated 'Donut background' matrices (see **Fig. 3.5C**). Eqn. 3 generated 'Lower left background' matrices (see **Fig. 3.5D**).

The parameters *p* and *w* determine the dimensions of the donut/lower left window surrounding each interaction frequency matrix entry as detailed by Aiden and colleagues [10]. p and w are defined as the number of bins between the pixel/entry of interest to the inner (p) and outer (w) edges of the donut window, respectively. Thus, if the donut window is conceptualized as two squares, one larger containing the second smaller square, p = (width of small square – 1) / 2, w = (width of large square – 1) / 2 (**Fig. 3.5B**). By applying guidelines from *Rao et al.* that *p* should have a distance of 20-25 kb, we set *p* equal to 5 bins of size 4 kb. Similarly, we iterated through values of *w*, ranging from the minimum allowed by the formula (*p+2=6*) to 20 and selected *w*=15.

To capture the most stringent local background model represented within the Donut and Lower Left background models, for each matrix entry we calculated the maximum of the two models and entered this into a new 'Donut/LL Max' background matrix (see **Fig. 3.5E**). If a matrix entry was non-existent ('NaN') in one background model but not both, the available real background value was utilized. Moreover, to avoid propagating expected

values in which we had low confidence, we set the corrected expected matrix entry to 'NaN' and excluded the bin-bin interaction from further analysis if greater than 80% of all possible values within the corresponding donut or lower left window were non-existent.

Probabilistic modeling

As previously described [156], we modeled the background-corrected interaction frequencies as a continuous random variable using the logistic distribution. Using the R fitdistr() function, we parametrized the fit independently for each region and replicate, and computed right-tail p-values. Finally, we computed 'background-corrected interaction scores' with the equation:

$$\mathrm{IS}_{i,j} = -10 \times \log_2\left(p_{i,j}\right)$$

where $p_{i,j}$ is the logistic p-value for a given entry in the background-corrected interaction frequency matrix. Background-corrected interaction score matrices were plotted as heatmaps to visualize 3D chromatin interactions that were enriched above the local interaction background (**Fig. 3.5F**).

Removal of interactions below distance limit

We identified 20 kb as our lower limit of bin to bin distance at which we could meaningfully identify 3D interactions; distance-corrected interaction p-value and distance-corrected interaction score entries for bins that were less than 20 kb apart were also set to 'NaN' and excluded from further analysis.

Thresholding interaction scores into cell-type specific interaction classifications

Each background-corrected interaction score matrix entry was subjected to a series of thresholds to classify each into a set of classifications based their value in each cell type (**Fig. 3.5G**, similar to strategy pursued in Beagan et al. 2016). Both replicates of each cell type were required to pass each threshold in order for an interaction (matrix entry) to be classified into a specific class. Refer to **Fig. 3.11A** for visualization of the thresholds discussed below. Matrix entries with interaction scores ≤ 3.22 (p-value of 0.8) across all six replicates were classified as 'background' interactions. If an entry had interaction score from each cell type less than 25.99 (p-value of 0.165, referred to as the 'significance threshold'), it was not classified into any interaction class. Otherwise, if both replicates from at least one cell type cleared the significance threshold, that entry could be classified as either (i) constitutive, (ii) present in two cell types but not the third (i.e. Serum+2i, Serum+NPC, NPC+2i), (iii) specific to one cell type (Serum-only, 2i-only, NPC-only). As in Supplemental Fig. 8A, this is simplified by first considering pairwise combinations of the cell type interaction scores; in this step, assuming the significance threshold has been passed in at least one of the cell types, an entry can be classified as either 'present only in cell type A', 'present only in cell type B', or 'present in both'. Interactions were 'present in both' if: (i) both replicates for each cell type had an interaction score greater than or equal to 40 (p-value of 0.0625, referred to as the 'constitutive threshold'), or (ii) if all four replicates under consideration cleared the significance threshold and the differences between all pairs of the four interaction scores were less than 30.2 (referred to as the 'difference threshold'). Otherwise in these two-way comparisons, entries were considered only in cell type A or B if in the 'present' cell type their interactions scores passed the significance threshold and the difference threshold when compared to the other cell type.

Finally, the results of these two-way comparisons were stitched together such that matrix entries were parsed as 'constitutive' if always classified as 'present in both' of the cell types queried, present in two cell types (Serum+2i, Serum+NPC, NPC+2i) if classified as 'present in both' when comparing the two named cell types but classified as 'present only in' each of these cell types when compared to the third un-named cell type, or cell type specific (Serum-only, 2i-only, NPC-only) if classified as 'present only in' the named cell type across both comparisons with the other two cell types. **Fig. 3.5H** displays the three-way scatterplot for these classes.

Clustering and Cluster Trimming

Similarly classified interactions that were spatially adjacent were grouped into interaction clusters as previously described [10]. Briefly, for a given classified interaction, if it existed next to an already identified cluster, the interaction was added to that cluster; if not, a new cluster was assigned to that interaction. After iterating through all classified interactions, adjacent clusters of the same classification were merged.

Interaction clustering enabled us to threshold our data based on interaction size in addition to interaction score. For each interaction cluster, the number of individual interaction matrix entries within that cluster and any clusters directly adjacent (of any classification) was tallied. If the individual interaction sum across itself and all directly adjacent clusters was not greater than 2, that cluster was removed as a low confidence cluster. The process of iterating through all clusters was repeated until no clusters were trimmed. The thresholding, clustering, and trimming methods produced our significant interaction cluster calls (**Fig. 3.5I**).

Empirical false discover rates

Six simulated 5C replicates were generated for each of our three cellular conditions as described in detail previously [156]. The 6 simulated background-corrected interaction frequency replicates were then passed through the same processing stages as the real background-corrected interaction frequency replicates (see above). Because the replicates were simulated to be from the same cell type, any interaction that is classified as a dynamic looping category was considered a false positive. Simulations of six biological replicates of the same condition were performed 1000 times and the average number of interactions that were classified for each cell type across the 1000 simulations were reported (**Fig. 3.11B**). One simulation round of six 5C library simulations of the NPC condition was chosen as representative in **Figs. 3.11C-D**.

Parsing Cell-Type Specific YY1

YY1 ChIP-seq datasets (NPC = [257], ES = [258], ProB = [185]) were downsampled together and peak-called with the MACS2 broad-peak caller using a diffuse p-value of 1e-8 and a broad cutoff of 1e-4 (see above). The subsequent broad peaks were parsed into cell type specific occupancy classes using Galaxy. ES serum only YY1 was defined by the presence of YY1 in serum/LIF ES cells and the absence of YY1 in NPCs and ProB cells (subtraction of NPC YY1 and ProB YY1 peaks from ES serum YY1). NPC only and ProB only peaks were parsed similarly. Constitutive YY1 was defined by the presence of YY1 in ES cells in serum, NPCs and ProB cells (intersection of the ES serum YY1 with NPC YY1 and ProB

YY1). A two-way class such as 'NPC and ProB, not ES' was parsed via the intersection of NPC and ProB peaks and the subtraction of ES peaks.

Parsing ES and NPC Enhancers

ES enhancers were defined as overlap between H3K27ac peaks and H3K4me1 peaks in ES cells in serum/LIF and absence H3K27ac in NPCs. This was calculated via the intersection of ES serum H3K27ac ($p < 1x10^{-8}$) with ES serum H3K4me1 ($p < 1x10^{-4}$, from [156]) followed by subtraction of low-confidence NPC H3K27ac ($p < 1x10^{-2}$). Similarly, NPC only enhancers were defined by overlap between H3K27ac peaks in NPCs and H3K4me1 peaks in NPCs and absence H3K27ac in ES cells in serum. To ensure exclusion of all genes from enhancer calls, we required that all parsed ES and NPC enhancers were not within 2 kb of a transcription start site (TSS).

Gene expression and Gene Annotation

Normalized, log2 gene expression counts were utilized from [156]. Genes were required to have a normalized, log2 expression count of at least 4 across both replicates of the cell type in which they were being considered active. Genes for which all pairwise replicate comparisons of ES serum expression with NPC expression displayed at least a 1.8 fold upregulation in ES cells were then intersected with H3K27ac ($p < 1 \times 10^{-8}$ in ES in serum); the resulting annotations were classified as 'ES-specific genes'. Similarly, genes with at least a 1.8 fold upregulation in NPCs compared to ES cells in serum across all replicates were then intersected with NPC H3K27ac ($p < 1 \times 10^{-8}$) and classified as 'NPC-specific genes'. Active genes across both cell types that exhibited than a 1.8 fold difference with

respect to each other were intersected with H3K27ac from both cell types and classified as 'Constitutively expressed'. Genes with normalized, log2 expression counts less than 2.5 across both cell types were classified as 'Inactive'.

Computing the enrichments of genomic annotations within interaction classes

Enrichments of annotations within interaction classes were calculated and visualized as previously described in detail [156].

CTCF Intersection with Consensus Motif and Directionality Enrichment Calculation

The CTCF position weight matrix was selected from the JASPAR core 2014 vertebrates motifs library. The position and orientation of the motif in the mm9 mouse genome were determined with PWM Tools (http://ccg.vital-it.ch/pwmtools/pwmscan.php). We then intersected our called CTCF peaks with this orientation file to assign orientations to each CTCF peak.

First, we parsed CTCF peaks with forward and reverse consensus motif orientations. We then identified the 4 kb bins intersecting with directionally oriented annotations. To take into account our 16 kb 5C smoothing window, we also considered a bin to contain an annotation if an adjacent bin on either side of the bin in question contained the annotation. Next, for each classified interaction, we determined whether the bins at the base of that interaction contained (i) no CTCF, (ii) CTCF on only one side, (iii) conflicting CTCF orientations over a single peak or in a single bin or (iv) unique CTCF orientations within both bins (**Fig. 3.12A**). We next parsed the interactions with unique CTCF orientations on both sides by which motif orientations actually appeared in the two bins:

(i) interactions with the forward motif orientation in its upstream bin and the reverse orientation in its downstream bin were classified as 'Convergent'; (ii) interactions with the same orientation on both sides, i.e. both forward or both reverse, were classified as 'Same Direction' or 'Tandem'; finally (iii) an interaction was considered 'Divergent' if only reverse motif(s) were present on the upstream side of the interaction and forward motif(s) present on the downstream side. This analysis was performed on the 'constitutive', '2i+Serum', and 'NPC-only' interaction classes. The enrichment above background for each of these orientations in each interaction class was also calculated as described above (see '*Computing the enrichments of genomic annotations within interaction classes')*.

APPENDIX II: METHODS ASSOCIATED CHAPTER 4

ES cell culture

V6.5 ES cells (murine; C57Bl/6 x 129SvJae; male) were purchased from Novus Biologicals. ES cells were expanded on Mitomycin-C inactivated MEF feeder layers in media consisting of DMEM, 15% FBS (Hyclone), 10^3 U/mL leukemia inhibitory factor (Millipore), non-essential amino acids (Lifetech), 0.1 mM 2-mercaptoethanol, 4 mM l-glutamine (Lifetech) and penicillin/streptomycin (Lifetech). Prior to fixation, ES cells were passaged onto gelatin-coated, feeder-free plates to remove feeder layer, and fixed at approximately 70% confluence. Cells were grown to ~7e6 cells per 15 cm dish at the time of fixation.

Primary NPC isolation

Neural progenitor cells were isolated from whole brains of newborn 129SvJae x C57/BL6, Sox2-eGFP mice and cultured as neurospheres in Neural Stem Cell media: DMEM/F12 media (Invitrogen 12100-046 and 21700-075) containing 72 mM glucose, 120 mM Sodium Bicarbonate, 5.6 mM Hepes (Sigma H-0887), 27.5 nM Sodium Selenite (Sigma S-9133), 18 nM progesterone (Sigma P0130), 90 ug/mL Apo-transferrin (Sigma T1428), 23 ug/mL insulin (Sigma I6634), 100 uM putrescine (Sigma P-7505), 2 mM L-glutamine (Gibco 25030-081), 1% Pen/Strep (Sigma P0781), 2 ug/mL heparin, 20 ng/mL rhEGF (R&D Systems) and 10 ng/mL rhFGF (R&D systems). Neurospheres were passaged every 3-4 days to prevent the formation of necrotic cores. After two passages, neurospheres were dissociated with Accutase and plated on Poly-D-Lysine Hydrobromide (100 ug/mL, Sigma

P7280), and laminin (15 ug/mL, Corning 354232) coated plates at 60,000 cells/cm^2. Cells were fixed with 1% formaldehyde one day after adherent plating.

iPS cell culture

The iPS cells analyzed in this study were reprogrammed from primary NPCs (pNPCs) as described in [206]. Briefly, pNPCs were transduced with lentiviral vectors to ectopically express Oct4, Klf4, and c-Myc (OKM). iPS cells derived from pNPCs were cultured on irradiated MEFs in medium consisting of Knock-Out DMEM, 15% FBS, Glutamax, non-essential amino acids, penicillin-streptomycin, b-mercaptoethanol and Leukemia Inhibitory Factor (LIF). iPS cells were grown to ~7e6 cells per 15 cm dish at the time of fixation. This iPS clone was extensively characterized for its pluripotent properties as assessed by (i) high expression of endogenous pluripotency markers (Oct4, Sox2, Nanog), (ii) demethylation of Oct4 and Nanog promoters, (iii) in vivo teratoma formation of all three germ layers and (iv) generation of chimeric mice [206].

Culture of pluripotent cells in 2i media

iPS and ES cells were removed from serum-containing media described above and cultured in 2i serum-free media comprised of 500 mL Knock Out DMEM (Life Technologies # 10829-018), 15% Knockout Serum Replacement (Life Technologies #10828), 5 mL N2 supplement (Life Technologies #17502-048), 5 mL B27 Supplement (Life Technologies #17504-044) , 5 mg/mL BSA (Sigma A9418), 1 mM L-Glutamine (Life Technologies # 25030-081), 1% Non-Essential Amino Acids (Millipore #TMS-001-C), 0.1 mM B-Mercaptoethanol (Life Technologies #21985-023), 1% Penicillin-Streptomycin (Sigma

#P0781), 10^3 units/mL LIF (Millipore #ESG1107), 3 uM CHIR99021 (Axon Medchem #1386), and 1 uM PD0325901 (Axon Medchem #1408) [154]. After two passages on feeder cells, ES and iPS cells in 2i media were passaged onto 0.1% gelatin to remove contaminating feeder cells. Cells were grown to ~7e6 cells per 15 cm dish at the time of fixation with 1% formaldehyde before 5C.

3C template generation and characterization

3C templates were produced as previously described [9, 259-261] for ES (n=2), NPC (n=2), iPS (n=2), ES+2i (n=2) and iPS+2i (n=2) pellets. Briefly, cells were fixed in base culture media (serum-free) supplemented with formaldehyde added to a final concentration of 1%. After 10 minute incubation at room temperature, fixation was terminated by adding 2.5M glycine stock to a final concentration of 125 mM glycine. Cross-linking termination was carried out for 5 minutes at room temperature followed by 15 minutes at 4°C. Cells were harvested with silicone scraper and pelleted, washed once with PBS, snap-frozen and stored at -80°C until processing.

Pellets were resuspended in lysis buffer consisting of 10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 0.2% Igepal CA630 and 1x protease inhibitor (Sigma) in sterile water and incubated on ice for 30 minutes. Cells were lysed with a dounce homogenizer and washed with NEB2 buffer. SDS was added to a final concentration of 0.1% and chromatin was solubilized by incubating at 65°C for 10 minutes. Triton X-100 was added to quench the SDS, and HindIII digestion was performed overnight at 37°C. The next day, the HindIII was inactivated and ligation was performed under dilute conditions at 16°C for 2 hours using T4 DNA ligase (Invitrogen) in ligation buffer consisting of 1% Triton X-100,

0.1mg/mL BSA, 1mM ATP, 50mM Tris-HCl, 50mM NaCl, 10mM $MgCl_2$ and 1mM DTT. After ligation, cross-links were reversed via incubation with 63.5µg/mL Proteinase K (Invitrogen) for 4 hours at 65°C, at which point the Proteinase K concentration was doubled and the solution was incubated overnight at 65°C. The 3C template DNA was then purified via a phenol extraction and a subsequent phenol-choloroform extraction before precipitation in ethanol. The resulting DNA pellet was resuspended in TE buffer consisting of 10 mM Tris-HCl (pH 8.0) and 1 mM EDTA (pH 8.0), and again purified by a series of phenol-chloroform extractions and precipitated in ethanol. The resulting DNA pellet was resuspended in TE buffer and treated with 100 ug/mL RNase A for 3 hours at 37°C.

5C primer design

5C primers were designed at HindIII restriction sites using the my5Csuite primer design tools [262], as described in detail in [9].

5C library generation and sequencing

5C libraries were generated as described previously [9, 43, 260, 263, 264]. 600 ng of each 3C template was mixed with final concentration 1 fmol of each 5C primer in 1x NEB4 buffer. Solution was incubated at 55°C for 16 hr to anneal primers to 3C templates. 5C primers annealed to 3C ligation junctions were ligated via the addition of 1x Taq ligase buffer containing 10 U Taq DNA ligase. Solution was mixed by pipetting and incubated for 1 hour at 55°C. Ligated 5C primers were then selectively amplified via the addition of universal forward (T7) and reverse (T3) primers, which anneal to the complementary universal primer tails of the 5C primers. 5C libraries (400 ng per library) were prepared for

sequencing using the NEBNext Ultra DNA Library Prep Kit (NEB # E7370S) and NEBNext Multiplex Oligos for Illumina (NEB # E7335S). After ligation of adapters following manufacturer's protocol, nuclease-free water was added to bring the reaction volume to 100 uL. Fragments of size ~ 220 bp (100 bp 5C product + 120 bp Illumina adapters) were preferentially selected using AgenCourt Ampure XP beads (Beckman Coulter A63881), by first adding 70 uL beads and retaining the supernatant, then adding 25 uL beads, removing the supernatant, and washing and eluting sample from the beads following the manufacturer's protocol. Following adapter ligation and size selection, the libraries with Illumina adapters were amplified with 10 cycles of PCR. The size distribution of the purified libraries were assessed on the Agilent BioAnalyzer using the DNA 1000 kit (Agilent 5067-1505). The resulting 5C libraries were pooled and sequenced with 37-cycles per paired-end on the Illumina NextSeq500.

iPS cell transgene integration detection by 5C primers

This iPS clone was generated via integration of transgenic Oct4, Klf4, and c-Myc genes [206]. Hochedlinger and colleagues demonstrated that this iPS clone exhibits transgene-independent self-renewal potential, which would exclude that these cells still depended on transgenic OKM expression. We note that our 5C approach does not exclude detection of the exogenous *Oct4* and *Klf4* genes (which were likely virally integrated at sites distal to our 5C regions) with 5C primers that directly bind to the Oct4/Klf4 coding sequence. However, short-range, cis interactions represent the majority of the 5C signal, and we do not analyze trans interactions in this study. Thus, we would expect the transgenes to

contribute relatively little to the interaction counts between these genes and other sites within our designed primer set.

RNA-seq library preparation

900,000 cells of each cell type were lysed with Trizol (Life Technologies 15596-026) and snap frozen. Total RNA was extracted and purified using the miRvana miRNA Isolation Kit (Ambion AM 1561) and samples were eluted into 100 uL nuclease free water. All RNA samples had an RNA Integrity Number >9 as assessed by Agilent BioAnalyzer. 50 uL of each RNA sample was treated with 1 uL rDNAse I (Ambion 1906) to remove residual genomic DNA. 350 ng DNAse-treated total RNA was prepared for sequencing using the Illumina TruSeq Stranded Total RNA Library Prep kit with RiboZero (Illumina RS-122-2202) following the supplier's protocol. cDNA libraries with Illumina adapters were amplified with 15 cycles of PCR. Libraries were purified using AgenCourt Ampure XP beads (Beckman Coulter A63881) with two rounds of 1:1 bead:sample selection. The size distributions of the purified cDNA libraries were assessed on the Agilent BioAnalyzer using the DNA 1000 kit (Agilent 5067-1505). Libraries were pooled and sequenced with 75-cyles per paired-end on the Illumina NextSeq500.

RNA-seq data processing

RNA-seq reads were aligned to the mouse genome (build mm9) using the Tophat (Tophat v2.1.0) alignment tool [265] with the parameters: -r 100 --no-coverage-search --library-type fr-firststrand and UCSC gene annotations. Gene level read counts were computed using the htseq-count tool (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html) with

parameters: -m union --stranded=reverse and UCSC gene annotations. For analyses of all 10 samples (ES_Rep1, ES_Rep2, pNPC_Rep1, pNPC_Rep2, iPS_Rep1, iPS_Rep2, ES2i_Rep1, ES2i_Rep2, iPS2i_Rep1, iPS2i_Rep2), genes with more than three counts in at least five libraries were retained, resulting in a total of 11,767 genes analyzed. To account for library-specific differences in sequencing depth, log2-transformed libraries were normalized by read depth of the 75%tile gene. Libraries were assessed for the absence of batch effects before proceeding to downstream biological analyses (**Figure 4.8**).

CTCF binding detection by ChIP-qPCR

Approximately 20 million cells were fixed in serum-free culture media supplemented with formaldehyde added to a final concentration of 1%. After 10 minute incubation at room temperature, fixation was terminated by adding 2.5M glycine stock to a final concentration of 125 mM glycine. Cross-linking termination was carried out for 5 minutes at room temperature followed by 15 minutes at 4°C. Cells were harvested with silicone scraper and pelleted, washed once with PBS, snap-frozen and stored at -80°C until processing.

Cell pellets were thawed for 10 min on ice before use. Nuclei were isolated by resuspending each pellet in 1 mL Cell Lysis Buffer (10 mM Tris pH 8.0, 10 mM NaCl, 0.2% NP-40/Igepal, Protease Inhibitor, PMSF), incubating on ice for 10 min, and spinning to pellet. Nuclei were resuspended in 500 uL Nuclear Lysis Buffer (50 mM Tris pH 8.0, 10 mM EDTA, 1% SDS, Protease Inhibitor, PMSF) and incubated on ice for 20 min. After bringing the samples up to volume by the addition of 300 uL IP Dilution Buffer (20 mM Tris pH 8.0, 2 mM EDTA, 150 mM NaCl, 1% Triston X-100, 0.01% SDS, Protease Inhibitor, PMSF), samples were sonicated for 45 minutes using an Epishear sonicator set

at 100% amplitude, with cycles of 30 seconds on and 30 seconds off. The resulting sheared chromatin was spun down, and the supernatant was transferred to a preclearing solution of 3.7 mL IP Dilution Buffer, 0.5 mL Nuclear Lysis Buffer, 175 uL of Agarose Protein A/G beads, and 50 ug Rabbit IgG, and rotated at 4°C. 35 uL Protein A/G agarose beads were pre-bound with 10 uL anti-CTCF antibody (Millipore #07-729) and incubated for 2 hours during the pre-clear stage. After a two hour pre-clear incubation, the beads were pelleted, and 4.5 mL supernatant was removed. 200 uL was reserved for input control, while the remaining supernatant was transferred to agarose beads pre-bound with antibody and rotated overnight at 4°C. Bound bead complexes were washed once with 1 mL IP Wash Buffer 1 (20 mM Tris pH 8.0, 2 mM EDTA, 50 mM NaCl, 1% Triton X-100, 0.1% SDS), twice with 1 mL High-Salt Buffer (20 mM Tris pH 8.0, 2 mM EDTA, 500 mM NaCl, 1% Triton X-100, 0.01% SDS), once with IP Wash Buffer 2 (10 mM Tris pH 8.0, 1 mM EDTA, 0.25 M LiCl, 1% NP-40/Igepal, 1% Na-deoxycholate), and finally once with 1x TE. Complexes were eluted by twice resuspending bound beads in 110 uL Elution Buffer (100 mM NaHCO3, 1% SDS), pelleting the beads after each elution and transferring 100 uL supernatant to a new tube. Finally, 12 uL of 5M NaCl and 20 ug RNase A were added to both 200 uL IP and input samples and incubated at 65 degrees for 1 hour, followed by the addition of 60 ug of Proteinase K and overnight incubation at 65 degrees. DNA was isolated via phenol-chloroform extraction and ethanol precipitation, and concentration was quantified using Qubit fluorometer.

ChIP libraries were prepared from 3 ng of IP and input DNA using the NEBNext Ultra Library Prep Kit (NEB #E7370) following the manufacturers protocol for preparation of ChIP libraries. After adapter ligation, no size selection step was performed, and ligated

samples were enriched through 18 PCR cycles using NEBNext Multiplex Oligos for Illumina (NEB #E7335). Libraries were eluted in 30 uL 0.1x TE, and a fragment size distribution between 250 and 1200 bp including sequencing adapters was confirmed using a High-Sensitivity assay on a Agilent Bioanalyzer.

Primers were designed to query specific CTCF binding sites:

Figure Panel	**Forward Primer**	**Reverse Primer**	**Genomic Coordinates**
5G (NPC-iPS)	TGTGGTCCTTTGTCCTTCCTG	TGTCACGCATCCTGAATCTTC	Chr3:35002112-35002461
5G (ES only)	AACTCACTAAGTGGCCCGAAG	ACCCCAGCTCCACGAAAATG	Chr3:34658634-34659306
6H	GTGTACAAGCACGCACGTATG	AAAGGGAGGTGCTCAATGGTC	Chr4:54936308-54936574
S7G	TAACCCTCACTGCTTGCGTAG	TGTGTCCTTAGCAGACGTGTC	Chr16:90635525-90635762

Quantitative PCR was performed by loading 1 ng of each sample library into each 20 uL reaction, including 10 uL Power SYBR Green PCR Master Mix (Applied Biosystems # 4367659), and corresponding primers (200 nM final concentration). Reactions were loaded onto an Applied Biosystems StepOnePlus in three replicates and assayed using standard qPCR cycling conditions (95°C for 10 min, followed by 40 cycles of 95°C for 15 sec and 65°C for 1 min). The CT threshold was set at 1900 so as to fall in the middle of the exponential phase for all primers and to capture the CT value for all samples. To facilitate comparison among the five cellular conditions, relative enrichment in CTCF ChIP signal was assessed by normalizing data by a reference control primer representing a constitutively bound CTCF site.

5C data processing pipeline

Paired-end read mapping and counting

5C data were generated with paired-end sequencing (37 bp paired-end reads) on the Illumina NextSeq 500 instrument. The two ends of paired-end (PE) reads were aligned independently to a pseudo-genome consisting of all 5C primers using Bowtie with default parameters (http://bowtie-bio.sourceforge.net/index.shtml) [254]. Only reads with one unique alignment were considered for downstream analyses. Interactions were counted when both paired-end reads could be uniquely mapped to the 5C primer pseudo-genome. Only interactions between forward-reverse primer pairs were tallied as true counts.

Low count primer removal

Primers with fewer than 100 total reads across all possible cis primer ligation partners were excluded from further analysis. Removed primers are listed below:

#track	Start	Stop	Primer ID
chr3	87677389	87683794	5C_326_Nestin_FOR_117:0
chr3	88032708	88035039	5C_326_Nestin_FOR_192:0
chr3	88124897	88125644	5C_326_Nestin_FOR_214:0
chr3	88283586	88286361	5C_326_Nestin_FOR_248:0
chr16	91242594	91247280	5C_325_Olig1-Olig2_FOR_193:0
chr17	35285175	35292115	5C_327_Oct4_FOR_191:0
chr17	36018525	36020858	5C_327_Oct4_FOR_378:0
chr17	36023358	36024542	5C_327_Oct4_FOR_380:0
chr17	36393683	36395722	5C_327_Oct4_FOR_472:0
chr3	34546431	34549386	5C_329_Sox2_REV_154:0

Raw contact matrix visualization

First we designated the restriction fragments to which 5C primers were designed as "queried restriction fragments". Raw contact matrices were generated for each region by

placing the number of counts read for the interaction of the ith queried restriction fragment in the region with the jth queried restriction fragment in the region in the ijth entry of the contact matrix. This created a square, symmetric matrix of contacts with dimensions equal to the number of primers in the region. Because interactions between fragments whose corresponding primers are oriented in the same direction cannot be detected with our 5C primer design, not every entry of this matrix corresponds to a detectable fragment-fragment interaction.

Because approximately half of the entries in this contact matrix represent undetectable fragment-fragment interactions, we visualized raw contact matrices at the fragment level by arranging the forward primers on the x-axis and the reverse primers on the y-axis, in order of primer number, which corresponded directly with the sorted order of genomic coordinates (heatmaps in **Fig. 4.2A**). Thus, the ijth cell of the resulting heatmap represents the number of counts for the interaction of the fragment queried by the jth forward primer with that queried by the ith reverse primer. This heatmap, used only for initial visualization, is therefore asymmetric and not necessarily square.

Quantile normalization

It is essential to account for technical variation among 5C replicates - in particular, batch effects for experiments processed or sequenced on different days - before comparing dynamic architecture between biological conditions. Indeed, we have found that two important factors driving experimental variability between biological replicates are (i) library complexity and (ii) sequencing depth differences between each batch of processed samples. We have found that a simple normalization factor is insufficient to remove bias

due to sequencing depth because the differences in read counts between replicates tend to compound in a nonlinear manner based on the underlying complexity of the library.

Quantile normalization is a rank-based approach that has successfully been used to normalize microarray [266], RNAseq [267] and Hi-C [44] data prior to downstream modeling. Here we also find that quantile normalization is effective at placing different 5C libraries on the same distributional scale (compare distance dependence and histograms in **Fig. 4.2A-B**) while preserving biologically significant architectural features (compare heatmaps in **Fig. 4.2A-B**). We have noticed that quantile normalization is particularly effective on 5C datasets because the strongest signal in the raw data is the distance-dependence background, providing a smooth, ubiquitous rank-order gradient for the comparison of contacts across replicates and conditions. Indeed, we found that our analysis was largely insensitive to the exact placement of the quantile normalization step relative to the other steps. For example, we moved the quantile normalization step to the end of our 5C analysis pipeline (**Fig. 4.3A+B,E-G**) and found that all views of the data show striking similarity to the corresponding stages of our implemented data processing pipeline (**Fig. 4.2A-F**).

Primer correction

Consistent with our findings in [9], we noticed the presence of primer-specific bias in our 5C data. For example, we observed strongly underenriched or overenriched stripes in our raw heatmaps – indicating that entire rows/columns can have increased or decreased counts (heatmaps in **Fig. 4.2A).** Consistent with this observation, the cis interactions for each primer show up to an ~8500-fold variation in mean interaction frequency, suggesting the presence of artifacts independent from the biology that influence the 5C signal (boxplots

in **Fig. 4.2A**). To account for primer-specific artifacts, we applied our previously developed primer correction method that uses stochastic gradient descent to compute primer-effect normalization factors [9]. After the primer correction step, we observed a marked attenuation of primer-specific artifacts (heatmaps and boxplots, **Fig. 4.2C**).

Low count fragment-fragment pair removal

Fragment-fragment pairs with primer-corrected counts below 10 in any replicate were flagged as low outliers with essentially unreliable values and were excluded from further analysis.

Contact matrix binning

We next generated a binned contact frequency matrix by binning each of our queried regions at regular 4 kb intervals (approximately equal to the average cut frequency of our chosen restriction enzyme, HindIII). To assign a value to each element of the binned contact probability matrix, we computed an arithmetic mean of logged counts using a square, 20 kb smoothing window as:

$$b_{i,j} = \frac{\sum_{k,l \ni |m_k - M_i| \leq 10\text{ kb}, |m_l - M_j| \leq 10\text{ kb}} \log_2(n_{k,l} + 1)}{\sum_{k,l \ni |m_k - M_i| \leq 10\text{ kb}, |m_l - M_j| \leq 10\text{ kb}} \mathbf{1}(d_k \neq d_l)}$$

where $b_{i,j}$ is the value assigned to the ijth entry of the binned contact matrix for the region and represents the contact frequency of the ith and jth bins in the region, m_k represents the midpoint of the kth primer in the region, M_i represents the midpoint of the ith bin in the

region, and $n_{k,l}$ represents the number of counts for the interaction of the kth queried fragment in the region with the lth queried fragment in the region after primer normalization. $\mathbf{1}(d_k \neq d_l)$ represents an indicator function that checks whether the kth and lth primer in the region have the same directionality. This ensures that the average is computed only over the possible primer-primer interactions.

If more than 80% of all the fragment-fragment pairs in a bin-bin pair's smoothing window had values that were zero, impossible, or had been previously removed as low outliers, that bin-bin pair was determined to be located in a low-confidence region and was excluded from further analysis. The bin-bin pair removal condition can be represented as:

$$\frac{\sum_{i,j \ni |m_i - M_k| \leq 10\text{ kb}, |m_j - M_l| \leq 10\text{ kb}} \mathbf{1}(n_{i,j} > 0)}{\sum_{i,j \ni |m_i - M_k| \leq 10\text{ kb}, |m_j - M_l| \leq 10\text{ kb}} 1} < 20\%$$

$$\Rightarrow b_{k,l} \text{ excluded from further analysis}$$

We selected the 20 kb smoothing window size and the 4 kb matrix resolution through a process of (1) iteratively testing window sizes and matrix resolutions, (2) visually inspecting the resultant heatmaps and (3) qualitatively comparing heatmaps to classic epigenetic marks. Our final strategy optimally accounted for sampling noise in 5C data while retaining what we term a pseudo-fragment (~12 kb) resolution (discussed in detail below). We chose to assign values to the entries of the binned contact matrix using an average rather than a sum because HindIII has been previously shown to exhibit highly variable restriction site density across the genome. To attenuate the spatial noise present in our fragment-level data, our binning strategy effectively averages counts across a 20 kb

window (compare heatmaps in **Fig. 4.2C+D** and **Fig. 4.3B+E**). This reduction of spatial noise is concurrent with a tightening of the distribution of counts across this step (compare histograms in **Fig. 4.2C+D**).

Pseudo-fragment level 5C mapping resolution

Many definitions of 3C/4C/5C/Hi-C resolution have been reported. Therefore, it is important to clarify our definition of resolution and our strategy for matrix binning. In a recent publication, the so-called "mapping resolution" of a Hi-C contact density map was defined as the smallest locus size such that 80% of the loci have at least 1000 contacts [10]. Importantly, Rao et al. reported the numbers in this definition as the finest scale at which they could reliably discern and distinguish architectural features in a Hi-C heatmap. By contrast to the "mapping resolution" metric, Rao et al. also define an alternative "matrix resolution" metric which is simply the bin size selected by the investigator when constructing a contact density matrix. In our lowest read depth replicate, iPS+2i Rep 1, 97% of the queried fragments have more than 1000 contacts. Thus, if we define our loci as the individual restriction fragments queried by the assay, all our datasets have a mapping resolution equal to the fragment size (~4 kb). We find a 4 kb bin size as the finest scale at which we can discern architectural features in our 5C contact density matrix. On the basis of a strictly "matrix resolution" definition, the resolution of our 5C data would be 4 kb. However, because we use a square 20 kb smoothing function (discussed below), there are hypothetical situations in which we cannot resolve two perfectly punctate features that are within 20 kb of each other. Thus, our "mapping" resolution falls in the range of 4-20 kb.

The design and orientation of 5C primers is another critical factor unique to 5C that must be considered in calculating resolution. Importantly, the true alternating 5C primer design used here and in [9] only queries a subset of possible fragment-fragment interactions. Specifically, forward and reverse primers were tiled in a true alternating manner across our genomic regions. Only forward-reverse (F-R) and reverse-forward (R-F) ligation products can be detected with the ligation-mediated amplification approach. Thus, although we can distinguish most interactions at a ~4 kb resolution, our more generalized resolution due to the alternating primer design is at the level of F-R-F or R-F-R fragment sequences (~12 kb; also the midpoint between our 4-20 kb mapping resolution).

To our knowledge, no Hi-C map has been reported at true single-fragment resolution as even the highest density maps have been binned to 1-5 kb resolution with a 4 bp cutter that cuts approximately every 200-300 bp in the genome. Thus, the highest resolution maps to date still average or sum information from at least 4 (1 kb resolution) but as many as 1000's (1 Mb resolution) of adjacent restriction fragments prior to modeling, parameterization of models, and downstream analyses. The reason for this requisite binning step is that the sampling noise in 5C/Hi-C contact matrices represents a significant barrier in obtaining high-confidence information for the read counts in every bin across the genome. However, a high-confidence understanding of the interaction frequency can be modeled at the expense of losing some resolution by averaging or summing counts from nearby fragment-fragment pairs. Here, we use 5C, which offers key advantages over Hi-C in its ability to obtain high complexity contact density maps with a logistically reasonable sequencing depth. Thus, we have high complexity libraries (i.e. most restriction fragment ligation products have been sampled at an ultra-high count

density). For example, in iPS+2i Rep 1, our lowest-mapping replicate, 80% of our originally queried fragments received >5340 counts. Ultimately, to account for spatial noise, we chose a 20 kb windowing function to yield a search space over an approximately 5x5 grid of primer-primer pairs (F-R-F-R-F or R-F-R-F-R). Overall, we propose that our resolution falls between 4 and 20 kb – with approximately a 12 kb resolution due to the true alternating primer design.

Identification of bad primer gaps

Restriction site density varies widely across the genome. Additionally, it is possible that certain primers fail to produce any counts due to technical error. Finally, many restriction fragments did not receive a primer due to low quality scores, leaving several loci unqueried by the assay. All three factors may affect the distance between one existing "working" primer and the next downstream "working" primer. When this distance is small compared to the smoothing window, the gap will be successfully spanned by multiple unique smoothing windows. When this distance is on a similar scale to the smoothing window, the smoothing window will be too small to reliably smooth across the gap. Within each region, we identified columns of bins that contained no positive counts from any primer ligation. When the length of a run of consecutive missing or zero fragments was greater than half the size of the smoothing window plus one bin, we classified the gap as "unsmoothable." Unsmoothable gaps are marked with dark gray on the heatmaps and excluded from all statistical analyses.

Distance-dependence normalization

To account for the distance-dependence background inherent in 3C-related assays, we computed an empirical expected distance-dependence model (**Fig. 4.2G**). Within each region and replicate, we first grouped the bin-bin pairs according to their interaction distance d, as measured by the number of bins separating the constituent bins in the bin-bin pair. We then computed the mean of the binned interaction frequencies within each group, as follows:

$$\mu_d = \text{mean}_i\left[b_{i,i+d}\right]$$

where μ_d is the mean value at distance d (measured in number of bins of separation), and $\left(b_{i,i+d}\right)_i$ is the sequence of binned contact frequencies for bin-bin pairs at distance d. Since the number of matrix entries included in each average will decrease with increasing distance d, these mean values are statistically weak predictors at long (> 600-700 kb for a 1 Mb region) distance scales. To account for any noise in our empirical distance-dependence estimations, we lowess-smoothed a subset of the empirical expected values in order to obtain a smooth approximation to the empirical expected values. Due to the high number of matrix entries at distances <= 300 kb, we retained the original mean values at short distance scales (<= 300 kb for a 1 Mb region).

We next used our empirical expected model to normalize the binned contact matrices by computing a fold-enrichment of counts relative to the expected (**Figs. 4.2E, 4.3G**). Since the values in our binned contact matrices were already log-transformed, we directly computed a log-scale fold-enrichment as:

$$f_{i,j} = b_{i,j} - \mu_{|i-j|}$$

where $f_{i,j}$, the ijth entry of the distance-normalized contact matrix, represents the log-scale fold-enrichment of interactions between the ith and jth bins in the region, $b_{i,j}$ is the ijth

element of the binned interaction matrix, and $\mu_{|i-j|}$ represents the distance-dependence normalization factor appropriate for a bin-bin pair at distance $d = |i - j|$ within the region under consideration (described above). Distance dependence-normalized counts show no discernable relationship with interaction distance compared to data at earlier stages of the analysis (histograms in **Figs. 4.2E, S2G**).

Noteworthy, the Klf4 region spans two distinct sub-TADs with markedly different interaction frequencies. We divided Klf4 into two separate sub-regions and created independent expected models for sub-region_1 (single block: chr4:54,899,978-55,371,978 x chr4:54,899,978-55,371,978) and sub-region_2 (the union of three blocks: chr4:54,899,978-55,371,978 x chr4:55,371,978-55,887,978, chr4:55,371,978-55,887,978 x chr4:55,371,978-55,887,978 and chr4:55,371,978-55,887,978 x chr4:54,899,978-55,371,978).

Probabilistic model fitting and distance-corrected interaction scores

We modeled our distance-corrected interaction frequency values as a continuous random variable using a logistic distribution parameterized independently for each region and replicate (**Fig. 4.5A**). We fit the logistic distribution by computing region-specific and replicate-specific location (l) and scale (s) parameters with maximum likelihood estimation through the R fitdistr() function. We computed right-tail p-values for every entry of distance-normalized contact matrices via the R plogis() algorithm, the lower.tail=FALSE argument and the below logistic cumulative distribution function:

$$p_{i,j} = 1 - \frac{1}{1 + e^{-(f_{i,j} - l)/s}}$$

where $p_{i,j}$ represents the right-tailed p-value for the relative interaction frequency found in the ijth entry of the distance-normalized contact matrix.

Prior to downstream thresholding/classification of significant 3-D interactions, p-values were transformed into distance-corrected interaction scores with:

$$\mathrm{IS}_{i,j} = -10 \times \log_2(p_{i,j})$$

Our computed distance-corrected interaction score offers a specific metric for identification/detection of significant 3-D interactions that are visually evident but difficult to disentangle from the underlying noise in the raw data (illustrated in heatmaps **Fig. 4.2F**). The highest (red/black) bins in ES and NPC heatmaps show strong cell type-specific correlation with known cell type-specific chromatin marks (heatmaps in **Fig. 4.2F**) while exhibiting strong attenuation of primer effects, absence of distance-dependence background signal and minimal distribution skewing due to technical differences in library complexity (boxplots and histograms in **Fig. 4.2F**).

GC content bias investigation

We assessed the degree of GC content bias in our original data and the degree to which our primer correction step attenuated the bias. First, we grouped restriction fragments into strata according to the GC content of the genome-binding portion of each 5C primer (i.e. the full 5C primer sequence minus the universal T7/T3 tail). We computed the sums of cis interactions for all primers in each strata and plotted each data point as an enrichment over the average cis interaction sum across all primers (**Fig. 4.2H**). A comparison of G-C content bias for each of the first three stages of our analysis pipeline demonstrated that

primers with extreme GC content are relatively depleted for counts in our raw data and that this bias is attenuated after primer correction (**Fig. 4.2H**). The attenuation in primer bias in extreme GC content strata is consistent with the goal of our primer correction scheme to push all primers towards equal visibility.

To further investigate the GC bias relationships in our data, we stratified our primer-primer pairs into a 2-D grid of strata depending on the GC content of the upstream and downstream primer comprising the forward-reverse primer pair. We then visualized the enrichment of counts within each stratum, computed as described by Ren and colleagues [38] as:

$$E_{a,b} = \frac{\sum_{i,j \ni l_a < g_i \leq u_a, l_b < g_j \leq u_b, i > j} c_{i,j}}{\sum_{i,j \ni l_a < g_i \leq u_a, l_b < g_j \leq u_b, i > j} \mu}$$

where $E_{a,b}$ is the enrichment value for the abth stratum in the grid, l_a and u_a are the lower and upper GC content limits, respectively, of the ath stratum, l_b and u_b are the lower and upper GC content limits, respectively, of the bth stratum, g_i is the GC content of the ith primer, $c_{i,j}$ is the number of counts for the interaction of the ith primer with the jth primer, and μ is the mean number of counts across all primer-primer pairs.

We generated GC strata heatmaps for raw and primer corrected data (**Fig. 4.2I**). Although the strata with the most extreme GC contents show less bias after normalization, there was still a noticeable enrichment of counts centered on the 50-60% to 50-60% pairwise GC content range. This result is consistent with previous observations by Ren and colleagues suggesting that there might be a biologically significant enrichment for 3-D

interactions between genomic elements with high GC content levels at distance scales < 2 Mb [38].

Comparison of 5C analysis pipeline to alternative approaches

We compared the results from our current 5C data analysis steps to the results of the corresponding steps in our previously published 5C analysis pipeline (**Fig. 4.3A-D**). In our previous approach, data were not quantile normalized, the distance-dependence background was modeled parametrically with a Weibull distribution, no binning was performed and p-values were computed via modeling single fragment resolution data with a compound normal-lognormal distribution [9].

First, we corrected for primer effects by employing the same primer normalization strategy in our current and original analysis pipelines. The primer correction step attenuated under/over-enriched stripes in the heatmaps, pushing all rows/columns toward equal visibility, independent of whether or not the data were quantile normalized (compare boxplots and heatmaps in **Figs. 4.2C and 4.3B**). Second, our 2016 empirical, region-specific distance-dependence models show improved ability to correct for the short-range distance-dependence relationship than our previous 2013 parametric distance-dependence model (compare heatmaps and distance-dependence curves in **Figs. 4.2E and 4.3C**). Third, our 2016 binning approach at ~12 kb 'pseudo-fragment resolution' (discussed above) offers key improvements in highlighting the true looping signal vs. noise when compared to our 2013 ~4 kb 'single fragment resolution' maps (compare heatmaps in **Figs. 4.2D-F and 4.3C-D**). Finally, our 2016 approach to model distance-corrected interaction frequencies as a continuous random variable with the logistic distribution results in the

clear illumination of underlying looping patterns in distance-corrected interaction score heatmaps. Our previous approach modeling single fragment resolution data with a compound normal-lognormal distribution did allow for the identification of a few of the strongest structural features that change dynamically between cell types. However, distance-corrected interaction score maps from the 2013 pipeline exhibited a much greater degree of spatial noise that obscured many important 3-D interactions (compare heatmaps in **Figs. 4.2F and 4.3D).** Finally, we moved the order of our current pipeline steps - conducting quantile normalization after binning, performing the binning step on unlogged data and logging only for visualization – and the resultant heatmaps showed similar results to our current pipeline steps, suggesting that the biological conclusions are robust to the order at which we conduct our pre-processing steps (**Figs. 4.3E-G**).

Overall, our 5C methods were chosen because they yield highly sensitive and quantitative identification/detection of significant 3-D interactions while exhibiting strong attenuation of primer effects, absence of distance-dependence background signal and minimal distribution skewing due to technical differences in library complexity (**Fig. 4.2F**).

Principal component analysis

Principal component analysis was performed to scatter the six experimental replicates according to their distance-corrected interaction frequencies at each bin-bin pair. The R prcomp() function with active center and scale parameters was used to compute the principal components for our six conditions. We plotted the projection of our six conditions onto the first two principle components as a scatterplot.

Classification of cell type-specific 3-D interactions

To classify cell type-specific 3-D interactions, we generated scatterplots of distance-corrected interaction scores for pairwise combinations of ES cells, NPCs and iPS cells (**Fig. 4.6A-F**). Specifically, for every 4 kb bin, the minimum distance-corrected interaction score between the two replicates for each cell type was plotted to ensure both replicates must fall above any threshold to be considered for classification. Distance-corrected interaction scores ≤ 3.219 in ES cells, NPCs and iPS cells were classified as "background" interactions. Interactions for which all cell types had a distance-corrected interaction score ≤ 30 were not considered in the parsing of any 3-D interaction class.

For each pairwise comparison, distance-corrected interaction scores were classified as: (i) 'present in both cell types', (ii) 'present in cell type 1', (iii) 'present in cell type 2', (iv) 'unable to be differentially assigned with confidence', or (v) a 'background' interaction (i.e. low interaction score) in both cell types (**Fig. 4.6**). Pairwise interaction classifications were then combined to determine differential interactions among all three cell types.

Reproducible distance-corrected interaction scores ≥ 53.219* in cell type 1 *and* cell type 2 were considered 'present in both cell types'. Similarly, if the difference between the minimum interaction scores of both cell types did not exceed 14, the interaction was also classified as 'present in both cell types'. Interactions with differences between the distance-corrected interaction scores of the two cell types greater than 14 that also had interaction scores ≥ 43.219 but < 53.219 in all cell types were removed from consideration because of uncertainty whether to classify them as constitutive or cell-type specific. The remaining interactions (i.e. at least one cell type interaction score > 30, at least one cell type interaction score < 43.219, and the difference between the minimum replicates of the cell

types > 14) were classified as ‘present in cell type 1’ if the interaction score in ‘cell type 1’ was greater and ‘present in cell type 2’ if the interaction score in ‘cell type 2’ was greater.

Pairwise classifications were combined to construct the 3-D interaction categories between the three cell types. Interactions that were considered ‘present in both cell types’ in all pairwise comparisons were parsed into the “constitutive” (grey class) 3-D interaction category. Interactions that were classified as ‘present in both ES and iPS cells’ but were found to be ES- and iPS-specific when comparing these cell types to NPCs were parsed into the “ES-iPS” (purple class) 3-D interaction category. Interactions that were classified as ‘present in ES cells’ when thresholded against both iPS and NPC distance-corrected interaction scores were parsed into the “ES-only” (red class) 3-D interaction category. Similarly, interactions classified as ‘present in both iPS cells and NPCs’ but were found to be iPS- and NPC-specific in comparison with ES cells were parsed into the “NPC-iPS” (blue class) 3-D interaction category. ‘Present in both ES cells and NPCs’ interactions were parsed into the “ES-NPC” (yellow class) 3-D interaction category if the interactions were not present when compared to iPS cells. Finally, interactions classified as ‘present in iPS cells’ when thresholded against both ES cells and NPCs were parsed into the “iPS-only” (orange class) 3-D interaction category, and interactions classified as ‘present in NPCs’ when thresholded against both ES and iPS cells were parsed into the “NPC-only” (green class) 3-D interaction category. We subsequently removed any interaction that was classified but spanned less than 20 kb between the bins involved in the interaction. Additionally, we removed interactions that spanned greater than 400 kb if they did not form an adjacency cluster (See “Interaction Adjacency Clustering” below) of at least 5 pixels.

*Note on thresholds: $53.219 = -10 * \log_2(0.025)$; $43.219 = -10 * \log_2(0.05)$; $30 = -10 * \log_2(0.125)$; $3.219 = -10 * \log_2(0.8)$, thus interaction scores of 53.219, 43.219, 30, and 3.219 correspond to interaction p-values of 0.025, 0.05, 0.125, and 0.8, respectively.

Empirical false discovery rate calculation

Justification of strategy

To compute an empirical false discovery rate (eFDR) for our interaction score thresholds, we employed a strategy in which we simulated 5C experiments consisting of three identical cellular conditions with two replicates each. The motivation/rationale for this strategy was that we wanted to determine how many 3-D interactions would be called by our thresholding/classification scheme (**Figs. 4.5, 4.5**) when comparing three cellular states (n=2 biological replicates each) that have been simulated to contain equivalent 3-D architecture. For example, we simulated ES1_Rep1, ES1_Rep2, ES2_Rep1, ES2_Rep2, ES3_Rep1, and ES3_Rep2, where all six replicates were generated from the same model (modeled based on our experimental ES data, discussed below). After the creation of the simulated replicates, ES1, ES2, and ES3 were treated as the distinct conditions for categorization purposes. By quantifying the number of interactions that we would expect by chance to pass our thresholds (discussed above), we can compute an eFDR for each 3-D interaction class identified when comparing ES vs. NPC vs. iPS cells.

Model generation – mean parameter estimation

First, we generated simulations of 5C data. To generate each of the simulations, we created three independent models, each of which was based on one of three cell type subsets (ES, NPC, iPS) of our experimental data. For each of these three models, we first computed a mean parameter by calculating the mean distance-corrected interaction frequency for that bin-bin pair among the two experimental replicates for the cell type the model was based on. We represent this mathematically as:

$$\mu_{c,s,i,j} = \frac{\sum_{r=1}^{2} f_{c,r,s,i,j}}{2}$$

where $\mu_{c,s,i,j}$ is the mean distance-corrected interaction frequency for the ijth bin-bin pair of the sth region in the model for cell type c and $f_{c,r,s,i,j}$ is the distance-corrected interaction frequency for the ijth bin-bin pair of the sth region in the experimental data for replicate r in cell type c.

Model generation – estimating the mean-variance relationship

Second, to obtain reasonable estimates for variance, we estimated a region-specific mean-variance relationship by performing a linear regression on the scatterplot of mean versus sample standard deviation of the distance-corrected interaction frequency for each bin-bin pair in each region among the two experimental replicates for the cell type being considered. This linear regression allowed us to compute a predicted standard deviation given a mean as:

$$\hat{\sigma}_{c,s,i,j} = m_{c,s}\mu_{c,s,i,j} + b_{c,s}$$

where $\hat{\sigma}_{c,s,i,j}$ is the predicted standard deviation of distance-corrected interaction frequency for the ijth bin-bin pair of the sth region in the model for cell type c, $\mu_{c,s,i,j}$ is the mean distance-corrected interaction frequency for the ijth bin-bin pair of the sth region in the

model for cell type c, and $m_{c,s}$ and $b_{c,s}$ are the slope and y-intercept parameters obtained from the linear regression of mean versus standard deviation for the sth region in the experimental data from cell type c.

Model generation – variance parameter estimation

Third, we used the mean-variance relationship to estimate the standard deviation parameter. We set the simulation standard deviation at each bin-bin pair to a linear combination of the observed standard deviation for that bin-bin pair in the experimental data for that cell type and our predicted standard deviation at that bin-bin pair as follows:

$$\sigma_{c,s,i,j} = \alpha\hat{\sigma}_{c,s,i,j} + \beta\sqrt{\frac{1}{2}\sum_{r=1}^{2}\left(f_{c,r,s,i,j} - \mu_{c,s,i,j}\right)^2}$$

where $\sigma_{c,s,i,j}$ is the final standard deviation parameter for ijth bin-bin pair of the sth region in the model for cell type c, $\sqrt{\frac{1}{2}\sum_{r=1}^{2}\left(f_{c,r,s,i,j} - \mu_{c,s,i,j}\right)^2}$ is the sample standard deviation of the distance-corrected interaction frequencies of the ijth bin-bin pair of the sth region in the experimental data from cell type c (r indexes the replicates), and α and β are constants chosen to ensure that the noise in the data generated by the model closely approximates the noise in the actual experimental data.

Simulations

Fourth, after computing the model parameters $\mu_{c,s,i,j}$ and $\sigma_{c,s,i,j}$, we generated simulated 5C experiments by drawing simulated distance-corrected interaction frequencies from a normal distribution with mean, variance parameters as follows:

$$F_{c,s,i,j} \sim N\left(\mu_{c,s,i,j}, \sigma_{c,s,i,j}\right)$$

where $F_{c,s,i,j}$ is a random variable representing the simulated distance-corrected interaction frequency for the ijth bin-bin pair of the sth region for a simulation of cell type c and $\mu_{c,s,i,j}$ and $\sigma_{c,s,i,j}$ are the mean distance-corrected interaction frequency and the final standard deviation parameter, respectively, for the ijth bin-bin pair of the sth region in the model for cell type c. We chose a normal distribution in accordance with our assumption that the replicate-to-replicate noise for repeated measurement of the same exact bin-bin interaction would be normally distributed.

Monte Carlo, p-value calculation, classification

Fifth, we used the above approach to generate six simulated 5C experiments from the same model, and then applied our logistic fits and our thresholding/classification scheme (described above) to each of the simulations. As in our real 5C data, we modeled the distribution of simulated distance-corrected interaction frequencies with a logistic distribution parameterized independently for each region. Logistic fits were used to assign p-values to every bin-bin pair in the simulation. P-values were converted to interaction scores as described above. The six independently constructed simulations were grouped into three equivalent categories containing two replicates each and subjected to the same thresholding/classification scheme as our experimental data. The number of simulated bin-bin pairs that were categorized into each of our 3-D interaction classes was recorded. This process was repeated 1000 times for each of our three cell types, and the numbers of simulated bin-bin pairs falling into each category were averaged across the 1000 trials and across the three cell types. We confirmed that our simulations fairly recapitulated the noise seen in the experimental data by comparing Spearman's and Pearson's correlation

coefficients as well as histograms and empirical cumulative distribution functions for our simulations to those we observed in our experimental data.

Computing the false discovery rates for each 3-D interaction class

Finally, we computed false discovery rates. Because the six simulated experiments represent simulated biological replicates, any bin-bin pair that was categorized into any category other than constitutive or background represents a false positive. Therefore, we estimated the false positive rate (FPR) for our thresholds for each of the other categories as the number of simulated bin-bin pairs falling into that category divided by the total number of bin-bin pairs in the simulation. Mathematically, this is represented as:

$$\mathrm{FPR}_t^{\mathrm{sim}} = \frac{\bar{n}_t^{\mathrm{sim}}}{N}$$

where $\mathrm{FPR}_t^{\mathrm{sim}}$ is the simulation false positive rate for category t, $\bar{n}_t^{\mathrm{sim}}$ is the average number of bin-bin pairs categorized into category t across all simulations, and N is the total number of bin-bin pairs in each simulation. We then assumed that the FPR for our simulation was a good estimate for the FPR in the categorization of our real experimental data.

$$\mathrm{FPR}_t^{\mathrm{sim}} \approx \mathrm{FPR}_t^{\mathrm{exp}}$$

where $\mathrm{FPR}_t^{\mathrm{sim}}$ is the simulation false positive rate for category t and $\mathrm{FPR}_t^{\mathrm{exp}}$ is the experimental false positive rate for category t. Our real experimental data and our simulations had the same number of bins and therefore the same number of bin-bin pairs to be categorized. Therefore, we estimated that for each category other than background and constitutive, the number of false positives observed in our simulations was equal to the number of false positives in our experimental data.

$$\mathrm{FPR}_t^{\mathrm{sim}} \approx \mathrm{FPR}_t^{\mathrm{exp}} \Rightarrow \bar{n}_t^{\mathrm{sim}} \approx \mathrm{FP}_t^{\mathrm{exp}}$$

where $\bar{n}_t^{\text{sim}}$ is the average number of bin-bin pairs categorized into category t across all simulations and FP_t^{exp} is the experimental number of false positives in category t.

We then estimated the false discovery rate (FDR) in our experimental data by dividing this estimated number of false positives by the total number of bin-bin pairs declared significant in the experimental data. Mathematically, this is represented as:

$$\text{FDR}_t^{\text{exp}} = \frac{\text{FP}_t^{\text{exp}}}{n_t^{\text{exp}}} \approx \frac{\bar{n}_t^{\text{sim}}}{n_t^{\text{exp}}}$$

where n_t^{exp} is the number of bin-bin pairs categorized into category t in the experimental data. Because a different number of bin-bin pairs were declared significant in different categories, we computed different FDRs for different categories (**Fig. 4.6H-I**).

6 sample vs 10 sample 5C data processing

5C data was processed either in a 6 sample batch, which includes only ES, NPC, and iPS replicates, or a 10 sample batch, which includes all 2i replicates in addition to the core 6 samples. Cell-type specific 3D interactions were classified using the '6-sample' group of ES, NPC, and iPS replicates. In instances where heatmaps are displayed for only these three cell types (i.e. Fig. 4, S5B, S6), we use '6-sample' normalized data, whereas when data is displayed for all 5 cell types (i.e. Fig. 5, S5F, 6, S7), we present '10-sample' normalized data.

Interaction adjacency clustering

Spatially adjacent interactions of the same classification were iteratively grouped into clusters in order to quantify the number of interaction clusters present in our data. For a given classified pixel, we queried if that pixel was adjacent to an already identified cluster

– if adjacent, the pixel was appended to that cluster - if not adjacent, the pixel was assigned its own cluster. Clusters of the same classification that were directly adjacent to themselves at the end of the iterative process were merged.

ChIP-seq peakcalling

Data was downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/). Sequences were aligned to NCBI Build 37 (UCSC mm9) using default parameters (-v1 -m1) in Bowtie. Only sequences that mapped uniquely to the genome were used for further analysis. Model-based Analysis for ChIP Sequencing (MACS) was used for peak calling (http://liulab.dfci.harvard.edu/MACS/00README.html). For CTCF ChIP-seq, default parameters were used with a p-value cutoff of $p < 1 \times 10^{-8}$. For histone modification ChIP-seq (e.g. H3K4me1, H3K27ac, H3K4me3), we skipped the model-building step by calling the parameter --no model with at p-value cutoff of either $p < 1 \times 10^{-8}$, $p < 1 \times 10^{-6}$ or $p < 1 \times 10^{-4}$.

Parsing ES-specific and NPC-specific genes

Normalized RNA-seq counts were parsed by fold change between ES cells and NPCs into ES-specific and NPC-specific gene expression categories. Genes that were at least two-fold upregulated in ES cells compared to NPCs were classified as ES-specific, whereas genes that were at least two-fold upregulated in NPCs compared to ES cells were classified as NPC-specific. ES-specific genes were further refined by required overlap with high-confidence H3K27ac signal (peaks called at $p < 1 \times 10^{-6}$) in ES cells. NPC-specific genes were further refined by required overlap with high-confidence H3K27ac signal (peaks

called at $p < 1 \times 10^{-4}$) in NPCs. Inactive genes were parsed by identifying those genes falling within queried 5C regions that did not exhibit H3K27ac signal (peaks called at $p < 1 \times 10^{-2}$) in either ES cells or NPCs.

Parsing ES-specific and NPC-specific enhancers

H3K27ac peaks (ES, $p < 1 \times 10^{-6}$; NPC, $p < 1 \times 10^{-4}$) were merged if they fell within 500 bp end-to-end distance of each other. NPC H3K27ac was peak-called at a lower threshold than the ES H3K27ac after visual observation that there appeared to be a smaller dynamic range of the NPC H3K27ac ChIPseq data between the active and inactive state. ES-specific enhancers were defined by overlap between merged H3K27ac peaks and H3K4me1 peaks ($p < 1 \times 10^{-4}$) in ES cells and the absence H3K27ac in NPCs (defined by subtraction of low-confidence NPC-binding sites for H3K27ac ($p < 1 \times 10^{-2}$)). NPC-specific enhancers were defined by overlap between merged H3K27ac peaks and H3K4me1 peaks ($p < 1 \times 10^{-4}$) in NPCs and the absence H3K27ac in ES cells (defined by subtraction of low-confidence ES-binding sites for H3K27ac ($p < 1 \times 10^{-2}$)). To ensure subtraction of all potential genes, it was required that parsed ES-specific and NPC-specific enhancers did not fall within 2 kb of a transcription start site.

Parsing ES-specific and NPC-specific CTCF sites

ES-specific CTCF was defined by the presence of high-confidence binding sites ($p < 1 \times 10^{-8}$) in ES cells and the absence of CTCF in NPCs (defined by subtraction of low-confidence NPC-binding sites for CTCF ($p < 1 \times 10^{-2}$). NPC-specific CTCF was defined by the presence of high-confidence binding sites ($p < 1 \times 10^{-8}$) in NPCs and the absence of

CTCF in ES cells (defined by subtraction of low-confidence ES-binding sites for CTCF ($p < 1 \times 10^{-2}$)). Constitutive CTCF was defined by the presence of high-confidence binding sites ($p < 1 \times 10^{-8}$) in both cell types.

Computing enrichments

Annotation intersections

For each bin in each of our 5C regions, we identified the genomic elements that overlapped that bin, or the neighboring 2 bins on either side (matching our 20 kb window, see *Contact matrix binning* above); the bin was then considered to 'contain' those genomic elements. Next, to interrogate pairwise connections between distinct genomic elements, we found all the bin-bin pairs whose upstream bin contained the first type of genomic element and whose downstream bin contained the second type of genomic element, or the reverse. For each of these bin-bin pairs, we checked which interaction classification category, if any, they fell into. We recorded the total number of intersections of this interaction class for every pair of types of genomic elements being considered and for every category in our interaction categorization scheme. By considering pairs of genomic elements in this way, we attempted to identify instances of one type of genomic element interacting with another type of genomic element. In our analysis, we included pairs of the same type of genomic elements (e.g., ES-specific genes interacting to ES-specific genes). We also created an artificial type of genomic element (referred to as "wildcard" element) that was present in every bin of every 5C region. Including this "wildcard" genomic element allowed us to query interactions that involved one specified type of genomic element interacting with

any other location, irrespective of what genomic elements were present on the other side (see **Fig. 4.12D**).

Computing percentage incidence, fold-enrichment above background, and p-values

Next, we divided the interaction counts for each pair of genomic element classes in each interaction category by the total number of interactions in that category to obtain the percentage of interactions in that category that involved an interaction between the two types of genomic elements in the pair. We then computed a fold-enrichment for each interaction type's percentage above the background interaction type's percentage. Finally, we computed p-values for the enrichment by applying Fisher's exact test to the contingency table below:

Number of interactions in the selected category involving the two selected annotations	Number of interactions in the background category involving the two selected annotations	Number of interactions in either the selected or the background category involving the two selected annotations
Number of interactions in the selected category not involving the two selected annotations	Number of interactions in the background category not involving the two selected annotations	Number of interactions in the selected or the background category not involving the two selected annotations
Total number of interactions in the selected category	Total number of interactions in the background category	

We used the p-value for the particular tail of the distribution that matched the direction of the enrichment (i.e., the right-tail p-value if the interaction was enriched over background, and the left-tail p-value if the interaction was depleted below background, generally

equivalent to the lesser of the two p-values). P-values were computed using the scipy.stats.fisher_exact function from the scipy Python computational library.

Visualizing enrichments

These enrichment quantification strategies were employed to investigate the intra-regional interactions of a selected annotation on either side of the interaction (via our "wildcard" annotation), and interactions between one selected annotation and another selected annotation falling within each interaction classification. Enrichments were visualized as either bar plots (showing the percentages of interactions between a pair of annotations falling into each of the interaction categories with the height of the different bars) or heat maps (with the color representing the log base 2 fold-enrichment of a certain interaction category above background for the percentage of interactions between a pair of annotations and the text showing the upper bound for the p-value for that enrichment).

Computing connectivity

To compute the 'connectivity' metric for each genomic annotation (**Fig. 4.14**), we first summed the number of significant interactions present in a given cell type that contained that annotation on at least one side of the interaction. A 'connectivity' value was computed by dividing the total number of interactions made by each annotation by the total number of interactions called significant in that cell type. For example, for the "ES enhancers in ES cells" data point, we counted the number significant interactions that intersected an ES enhancer and were categorized as either ES only, ES-iPS, ES-NPC, or constitutive (the

four interaction classes present in ES cells); this sum was then divided by the total number of interactions categorized as ES only, ES-iPS, ES-NPC, or constitutive.

APPENDIX III: METHODS ASSOCIATED CHAPTER 5

Cell Culture

Murine cortical neurons were cultured using a protocol established previously in [226]. Briefly, cortices were dissected from E18 WT C57/BL6 mouse embryos. Cortices were then dissociated in DNase (0.01%; Sigma-Aldrich, St. Louis, MO) and papain (0.067%; Worthington Biochemicals, Lakewood, NJ), then triturated with a fire-polished glass pipette to obtain a single-cell suspension. Cells were pelleted at 1000xg for 4 min, the supernatant removed, and cells resuspended and counted with a TC-20 cell counter (Bio-Rad, Hercules, CA). Neurons were plated in 6-cm dishes (Greiner Bio-One, Monroe, NC) coated with poly-L-lysine (0.2 mg/mL; Sigma-Aldrich) at a density of 200,000 cells/mL. Neurons were initially plated in Neurobasal media containing 5% horse serum (NM5), 2% GlutaMAX, 2% B-27, and 1% penicillin/streptomycin (Thermo Fisher Scientific) in a 37ºC incubator with 5% CO2. On DIV4, neurons were fed via half media exchange with astrocyte-conditioned Neurobasal media containing 1% horse serum (NM1), GlutaMAX, and penicillin/streptomycin, 2% B-27, and 5 μM cytosine β-D-arabinofuranoside (AraC; Sigma-Aldrich). Neurons were fed with astrocyte-conditioned NM1 media every three days thereafter. For chronic activity experiments, neurons were treated for 24 hours with either 1 uM Tetrodotoxin (TTX) or 10 uM Bicuculline (Bic) at DIV15 via addition to the cell culture media or left untreated. For short-term activity induction experiments, samples were subjected to 24 hours of TTX treatment at DIV15 followed by 0, 5, 20, 60, or 360 min of Bic treatment on DIV16. All animal experiments were approved by the Institutional Animal Care and Use Committee of the University of Utah.

ChIP-seq library preparation

At DIV16, neuronal cultures were fixed in 1% formaldehyde for 10 minutes (room temp) via the addition (1:10 vol/vol) of the following fixation solution: 50 mM Hepes-KOH (pH 7.5), 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 11% Formaldehyde. Fixation was quenched via the addition of 2.5 M glycine (1:20 vol/vol) and scraped into pellets of 8 million cells. Each pellet was washed once with cold PBS, flash frozen, and stored at -80ºC. Immunoprecipitation was performed as described previously[153, 156] with slight modifications. Briefly, IP reactions were prepared a day prior to cell lysis by combining 20 uL of protein A and protein G conjugated agarose beads (Invitrogen# 15918-014 and 15920-010, respectively) with 10 uL of anti-H3K27ac antibody (Abcam# ab4729) in 1 mL of cold PBS and rotated overnight. The next day cell pellets were resuspended in 1 mL lysis buffer (10 mM Tris pH 8.0, 10 mM NaCl, 0.2% NP-40/Igepal, Protease Inhibitor, PMSF) and incubated on ice for 10 min. Cells were further lysed with 30 strokes of a dounce homogenizer (pestle A) and then nuclei were pelleted. Nuclei were lysed on ice in 50 mM Tris pH 8.0, 10 mM EDTA, 1% SDS, Protease Inhibitor, PMSF for 20 min. SDS concentration was reduced before sonication by the addition of 300 uL IP Dilution Buffer (20 mM Tris pH 8.0, 2 mM EDTA, 150 mM NaCl, 1% Triston X-100, 0.01% SDS, Protease Inhibitor, PMSF), after which samples were sonicated for 60 minutes (30 seconds on, 30 seconds off cycle, 100% amplitude) using a Qsonica Q800R2 Sonicator. Insoluble fractions were removed via spin, and the supernatant was removed of non-specific binding chromatin via rotation with preclearing solution (3.7 mL IP Dilution Buffer, 0.5 mL Nuclear Lysis Buffer, 175 uL of Agarose Protein A/G beads, and 50 ug Rabbit IgG) for 2 hours at 4°C. Beads were pelleted and 4.7 mL of supernatant was removed. 200 uL of

supernatant was retained as input control (stored at -20°C) while the remaining 4.5 mL was transferred to the beads that had been pre-bound with the H3K27ac antibody overnight; the IP reaction then rotated overnight again at 4°C. Bound bead complexes were washed once with 1 mL IP Wash Buffer 1 (20 mM Tris pH 8.0, 2 mM EDTA, 50 mM NaCl, 1% Triton X-100, 0.1% SDS), twice with 1 mL High-Salt Buffer (20 mM Tris pH 8.0, 2 mM EDTA, 500 mM NaCl, 1% Triton X-100, 0.01% SDS), once with IP Wash Buffer 2 (10 mM Tris pH 8.0, 1 mM EDTA, 0.25 M LiCl, 1% NP-40/Igepal, 1% Na-deoxycholate), and finally twice with 1x TE. Complexes were eluted by twice resuspending bound beads in 110 uL Elution Buffer (100 mM NaHCO3, 1% SDS), pelleting the beads after each elution and transferring 100 uL supernatant to a new tube. Finally, 12 uL of 5M NaCl and 20 ug RNase A were added to both 200 uL IP and input samples and incubated at 65 degrees for 1 hour, followed by the addition of 60 ug of Proteinase K and overnight incubation at 65 degrees. DNA was isolated via phenol-chloroform extraction and ethanol precipitation and concentration was quantified using Qubit fluorometer.

ChIP-seq libraries were prepared for sequencing using the NEBNext Ultra II DNA Library Prep Kit (NEB# E7645S), following manufacturer's protocol with the following user-chosen specifications. 3 ng DNA from all IP and input samples was used as starting material. NEBNext Adaptors were diluted 15x in 10 mM Tris-HCL, pH 8.0 with 10 mM NaCl prior to adaptor ligation. Large DNA fragments were removed via a size selection by adding 15 uL of AMPure XP beads at the first bead addition step and 87 uL of beads at the second bead addition step. Size-selected DNA was amplified using 9 cycles of PCR enrichment. The size-range of the final libraries was confirmed to be between 200-1000 bp using an Agilent Bioanalyzer High Sensitivity DNA test. H3K27ac enrichment was

confirmed prior to sequencing by querying the IP/input qPCR enrichment of primer pairs designed to the Arc, Synaptotagmin-1, and Tcf25 promoter regions. Library concentrations were calculated and normalized using the KAPA Illumina Library Quantification Kit (#KK4835) so that libraries could be equally pooled before sequencing 75 bp single-end reads on the NextSeq500. IP libraries were sequenced to a depth greater than 48 million reads and all input libraries were sequenced to greater than 67 million reads.

ChIP-seq Analysis

H3K27ac ChIP-seq reads were aligned to the mm9 genome using Bowtie[254]. Reads with more than two possible alignments were removed (-m2 flag utilized). IP libraries across the Bic, Untreat, and TTX conditions were downsampled to 38 million reads, while input libraries were downsampled to 44 million reads. Peaks were identified using MACS2[255] with a p-value cutoff parameter of $1x10e^{-8}$ and the broadpeak flag also invoked with a broadpeak cutoff of $1x10e^{-8}$.

Parsing Putative Activity-Dependent Enhancers

H3K27ac peaks (p-value, broadPeak thresholds = $1x10^{-8}$) called in the TTX and Bic conditions were concatenated together and peaks within 2 kb of RefSeq TSS's were removed. The remaining peaks were merged so that peaks within 10 kb of each other were also merged together, thus generating a list of enhancer sites shared across the Bic and TTX conditions. From this master list of enhancer sites, each was parsed into activity-response classes by (i) calculating the average bigwig signal across the enhancer interval using the pybigwig package in both the Bic and TTX IP libraries, (ii) dividing those signal averages

by the average signal in the corresponding input library, (iii) calculating the Bic/TTX fold change of those input-normalized enhancer signals. An enhancer was defined as Bic-specific (activity-induced) if it exhibited a >2 Bic/TTX fold change and its Bic input-normalized signal was in the top 80% of all enhancers; TTX-specific (activity-decommissioned) enhancers were defined in the same manner with the conditions reversed. The remaining enhancer sites were classified as constitutive (activity-invariant) if their Bic and TTX input-normalized signals fell in the top 80% of enhancer signals in both conditions. H3K27ac signal heatmaps for each enhancer class were plotted using the Deeptools package[268].

3C Template Generation

Neuronal cultures were formaldehyde fixed as described for ChIP-seq and stored at -80ºC. For each condition (Bic, Untreat, TTX), *in situ* 3C was performed on 4 replicates (divided evenly across two animal/culture batches) of 4-5 million cells as described previously[10, 230]. Briefly, cells were thawed on ice and resuspended (gently) in 250 uL of lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal CA630) with 50 uL protease inhibitors (Sigma P8340). Cell suspension was incubated on ice for 15 minutes and pelleted. Pelleted nuclei were washed once in lysis buffer (resuspension and spin), then resuspended and incubated in 50 uL of 0.5% SDS at 62ºC for 10 min. SDS was inactivated via the addition of 145 uL H_2O, 25 uL 10% Triton X-100, and incubation at 37ºC for 15 min. Subsequently, chromatin was digested overnight at 37ºC with the addition of 25 uL 10X NEBuffer2 and 100U (5 uL) of HindIII (NEB, R0104S), followed by 20 min incubation at 62ºC to inactivate the HindIII. Chromatin was re-ligated via the addition of 100 uL 10% Triton X-

100, 120 uL NEB T4 DNA Ligation buffer (NEB B0202S), 12 uL 10 mg/mL BSA, 718 uL H_2O, and 2000 U (5 uL) of T4 DNA Ligase (NEB M0202S) and incubation at 16°C for 2 hours (NOTE: This is a deviation from in situ HiC (Rao et al. 2010) in order to promote sticky-end ligation over blunt-end). Following ligation nuclei were pelleted, resuspended in 300 uL of 10 mM Tris-HCl (pH 8.0), 0.5 M NaCl, 1% SDS, plus 25 uL of 20 mg/mL proteinase K (NEB P8107), and incubated at 65°C for 4 hours at which point an additional 25 uL of proteinase K was added and incubated overnight. 3C templates were isolated next day via RNaseA treatment, phenol-chloroform extraction, ethanol precipitation, and Amicon filtration (Millipore MFC5030BKS) (for more details see[153, 156]). Template size distribution and quantity were assessed with a 0.8% agarose gel.

5C Library Preparation

5C primers were designed according to the double-alternating design scheme[20, 139, 145, 230] using the My5C primer design software (http://my5c.umassmed.edu/my5Cprimers/5C.php)[262] with universal "Emulsion" primer tails. Regions were designed to capture TAD structures immediately surrounding the genes of interest (Bdnf, Fos, Arc, Neurexin-1, Neuroligin-3, Synaptotagmin-1) in published mouse cortex HiC data[7]. 5C reactions were carried out as previously described[139, 145, 230]. 600 ng (~200,000 genome copies) of 3C template for each replicate was mixed with 1 fmole of each 5C primer and 0.9 ug of salmon sperm DNA in 1x NEB4 buffer, denatured at 95°C for 5 min, then incubated at 55°c for 16 hours. Primers which had then annealed in adjacent positions were ligated through the addition of 10 U (20 uL) Taq ligase (NEB M0208L) and incubation at 55°C for 1 hour then 75°C for 10 min. Successfully ligated

primer-primer pairs were amplified using primers designed to the universal tails (FOR = CCTCTC TATGGGCAGTCGGTGAT, REV = CTGCCCCGGGTTCCTCATTCTCT) across 30 PCR cycles using Phusion High-Fidelity Polymerase. Presence of a single PCR product at 100 bp was confirmed via agarose gel, then residual DNA <100 bp was removed through AmpureXP bead cleanup at a ratio of 2:1 beads:DNA (vol/vol). 100 ng of the resulting 5C product was prepared for sequencing on the Illumina NextSeq 500 using the NEBNext Ultra DNA Library Prep Kit (NEB E7370) following the manufacturer's instructions with the following parameter selections: during size selection, 70 uL of AMPure beads was added at the first step and 25 at the second step; linkered fragments were amplified using 8 PCR cycles. A single band at 220 bp in each final library was confirmed using an Agilent DNA 1000 Bioanalyzer chip, and library concentration was determined using the KAPA Illumina Library Quantification Kit (#KK4835). Finally, libraries were evenly pooled and sequenced on the Illumina NextSeq 500 using 37 bp paired-end reads to read depths of between 11 and 30 million reads per replicate.

5C Interaction Analysis

The adoption of the double alternating primer scheme and *in situ* 3C significantly improved 5C data quality (see Kim and Titus 2018[230] for more detail) such that some steps of our 5C analysis approach could be changed from those previously utilized[153, 156] to more closely resemble those used for analyzing HiC[10]. Paired-end reads were aligned to the 5C primer pseudo-genome using Bowtie, allowing only reads with one unique alignment to pass filtering. Only reads for which one paired end mapped to a forward/left-forward primer and the other end mapped to a reverse/left-reverse primer were tallied as true counts.

5C is subject to specific biases, such as primer GC content resulting in annealing/PCR biases, that methods such as HiC are not. This manifests in primer-primer pairs with mapped counts that are orders of magnitude higher than the neighboring primer-primer pairs. Such an extreme enrichment of single primer-primer pairs does not resemble the broader distribution of elevated counts, spanning clusters of neighboring primer-primer pairs, that exists at bona fide looping interactions across 5C and HiC data. Therefore, we decided to remove these biased primer-primer pairs before proceeding with interaction analysis. This was done by calculating for each primer-primer pair the median count of itself and the 24 primer-primer pairs nearest to the primer-primer pair in question (i.e. a scipy.ndimage.generic_filter window of size 5 was passed over the primer-primer pair matrix and the median of each window was recorded). If the count of one primer-primer pair was greater than eight-fold higher its neighborhood median then it was flagged as a high spatial outlier and removed. This process was performed for all primer-primer pairs, except for those in the 5C region surrounding the Arc gene for which the 8-fold threshold was found to be too stringent due to low region complexity and a 100-fold threshold was utilized instead.

After high-outlier removal, primer-primer pair counts were quantile normalized across all 12 replicates (4 per condition) as previously described[230, 232]. For plotting purposes quantile normalized counts were merged across replicates via summation, whereas for loop calling analysis all replicates were kept separate. Primer-primer pair counts were then converted to fragment-fragment interaction counts by averaging the primer-primer counts that mapped to each fragment-fragment pair (max of 2 if both a forward/left-forward and a reverse/left-reverse primer were able to be designed to both

fragments and were not trimmed during outlier removal). We then divided our 5C regions into adjacent 4 kb bins and computed the relative interaction frequency of two bins (i,j) by summing the counts of all fragment-fragment interactions for which the coordinates of one of the constituent fragments overlapped (at least partially) a 12 kb window surrounding the center of the 4 kb i^{th} bin and the other constituent fragment overlapped the 12 kb window surrounding the center if the j^{th} bin. Binned count matrices were then matrix balanced using the ICE algorithm[232, 269], at which point we considered each entry (i,j) to represent the 'Relative Interaction Frequency' of the 4 kb bins i and j. Finally, the background contact domain 'expected' signal was calculated using the donut background model as previously described[14] and used to normalize the relative interaction frequency data for the background interaction frequency present at each bin-bin pair. The resulting background-normalized interaction frequency ("observed over expected") counts were fit with a logistic distribution from which p-values were computed for each bin-bin pair and converted into 'Background-corrected Interaction Scores' (interaction score = $-10*\log_2$(p-value)) as previously described[153]. Interaction scores have proven to be informatively comparable across replicates and conditions[9, 153], and as such were used for most visualization analysis and all loop-calling analysis to follow.

Quantitative 5C Loop Identification

We applied the 3DeFDR analysis package[233] to our dataset in order to identify differential interactions across the TTX and Bic conditions (4 replicates of each). Briefly, 3DeFDR identifies differential interactions and empirically estimates a false discovery rate (eFDR) for each identified dynamic looping class. Interactions are only considered for

analysis if the interaction scores of all 8 replicates across both conditions surpassed a 'significance threshold'. Interactions are classified as 'TTX-only' if all 4 interaction scores of the TTX replicates surpassed the interaction scores of the Bic replicates by more than a specified 'difference threshold'. 'Bic-only' interactions are classified in the same manner. Those interactions that pass the significance threshold but are not classified as Bic-only or TTX-only are classified as 'Constitutive'. Finally, significant interactions that pass our thresholds are clustered based on spatial adjacency into 'loops'. Looping clusters smaller than 5 pixels were removed. The 3DeFDR package simulates null replicate sets (i.e. 8 replicates of the same cell type/condition) using on a negative binomial counts generating function parameterized with mean-variance relationships computed from the real data. We compute an empirical FDR (eFDR) for each differential loop class as the total number of significant interactions called in that class on a simulated null replicate set divided by the total number of significant interactions called as that class with the original real replicate set.

We utilized the 'non-adaptive' functionality option of the 3DeFDR analysis package, which sweeps across a wide range of difference threshold and calculates an eFDR for each loop class at each iteration. We generated 250 simulated null replicate sets of 8 replicates based on mean-variance relationships underlying the real TTX replicates. We utilized the default 3DeFDR initialization parameters with the exception of 'bin_properties', which is a tunable parameter that specifies the distance scales over which fragment level interactions are stratified prior to fitting the negative binomial counts generating function to those interactions. We modified 'bin_properties' to capture the full extent of our regional matrices: (1) for close-range interactions (0-150 kb), we stratified

the interactions using fine-grained, 12 kb-sized sliding windows with a 4 kb step, (2) for mid-range interactions (151-600 kb), we stratified the interactions into 24 kb-sized sliding windows with an 8 kb step, and (3) for longer range interactions (601-2500 kb), we stratified the interactions into coarse-grained, 60 kb-sized sliding windows with a 24 kb step. Through this approach we achieved an eFDR of 6.6% for Bic-only (activity-induced) loops utilizing a difference threshold of 6.75, a significance threshold of $-10*\log_2(0.08)$ (i.e. a p-value of 0.08 resulting from the logistic fit to the observed over expected data), and a cluster size threshold of 5.

RNA-seq library preparation

At DIV5 and DIV16, 900,000 neurons were lysed in 1 mL Trizol (Thermo Fisher Scientific 15596026). Lysates were snap frozen and stored at -80ºc until use. Total RNA was then isolated using the mirVana miRNA Isolation Kit (Thermo Fisher Scientific AM1561) according to manufacturer's protocol and eluted from the spin-column using 100 uL nuclease-free water. Samples were DNase treated (Thermo Fisher Scientific AM1906) and tested for quality using an Agilent Bioanalyzer RNA chip. All samples produced an RNA Integrity Number (RIN) greater than 9. To avoid poly-A selection, we utilized the TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold (Illumina RS-122-2301) and prepared each RNA sample for sequencing according to the manufacturer's protocol. cDNA libraries were amplified across 15 PCR cycles followed by AMPure XP Bead clean-up (1:1 bead:solution ratio). Finally, the library sizes were confirmed to be between 200-500 bp using the BioAnalyzer before sequencing 75 bp paired-end reads on the Illumina NextSeq500. To minimize and identify technical variation, three replicates spanning two

culture batches were prepared, pooled, and sequenced to depths of greater than 60 million reads per library.

RNA-seq analysis

RNA-seq reads were mapped to the RefSeq transcriptome (transcriptome fasta downloaded from the UCSC genome browser on July 28, 2017) using Salmon[270]. In accordance with the TruSeq Stranded Total RNA Library Preparation, mapping was done using the -ISR flag. Additionally, 100 bootstraps of transcript quantification were performed. The resulting TPM quantifications for each RefSeq transcript were utilized for all downstream analyses. The Wasabi package (https://github.com/COMBINE-lab/wasabi) was utilized to convert Salmon bootstraps to the format necessary for differential expression analysis by Sleuth[235]. Differentially expressed transcripts were called using the Sleuth wald test, with a q-value threshold of 0.05. For enhancer RNA (eRNA) analysis, RNA-seq reads were mapped to the mm9 genome using STAR version 2.7.1[271] using default settings. Resulting bigwig files were used to quantify RNA signal overlapping each enhancer interval.

Linear Regression Modeling

To assess the relative contributions of cis-regulatory elements to activity response gene expression, for each transcript in our 5C regions we sought to quantify its promoter activity, looping strength, looped enhancer activity, and nearby enhancer activity. Transcripts whose promoter fell within 200kb of the edge of a 5C region were removed due to incomplete/truncated ability to query loops outside the 5C regions. Additionally, if transcripts of the same gene had overlapping promoters (+/- 2kb from TSS), only the

transcript with the highest maximum expression (TPM) across the TTX and Bic RNA-seq replicates was carried forward for further analysis. The promoter activity of each gene was calculated using the PyBigWig package to find the $\log_2$(Bic/TTX) fold change of the sum H3K27ac bigwig signal across the 4 kb promoter (+/- 2kb from TSS) in each condition (**Figure 5.8A,F**).

Each transcript was paired with the enhancer nearest to its TSS along the linear genome. If no enhancers fell within 200kb of the promoter, the transcript was considered to have no 'near enhancer' (only the case for NM_026271). The "activity" of the near enhancers were then also calculated as the $\log_2$(Bic/TTX) fold change of the sum H3K27ac bigwig signals across the enhancer (**Figure 5.8G**). Additionally, the total interaction frequency for each promoter was calculated by summing the observed 5C counts in the Bic and TTX conditions of all 5C bins the promoter overlapped and calculating the $\log_2$(Bic/TTX) fold change (**Figure 5.8B**). Similarly, the promoter of each transcript was intersected with 5C loops so that it could be paired with enhancers that fell at the other anchor of each loop. Often, promoters formed several loops, interacting with multiple enhancers. To select the single enhancer-promoter loop (so that we could accurately compare to the single nearest enhancer) predicted to have the largest regulatory role on the gene in question, we leveraged an adapted 'ABC model' approach originally reported by Engreitz and colleagues[141], selecting the enhancer-promoter loop that had the highest ((H3K27ac signal) * (5C Obs/Exp)) value (**Figure 5.8C**). Only promoters that looped to enhancers were included in calculations of loop strength and looped enhancer signal (**Figure 5.8D-E, H-I**). Notably, the looped enhancer models were more predictive of activity-dependent gene expression than the nearest enhancer and promoter-only models,

and this trend remained whether we used only genes engaged in loops (N=45, **Figure 5.9B-E**) or all genes (N=69, **Figure 5.8D-E, H-I**). 'Loop strength' was then calculated as the log_2(Bic/TTX) fold change of the 5C Obs/Exp counts of the ABC prioritized loop for each gene (**Figure 5.8D,H**). 'Looped enhancer' signal was calculated as the log_2(Bic/TTX) fold change of the sum H3K27ac bigwig signal in each condition at the selected looped enhancer (**Figure 5.8E,I**). Finally, the ((H3K27ac signal) * (5C Obs/Exp)) score itself was used to build a regression model (**Figure 5.8J**). The expression fold change of each transcript was calculated as the log_2(Bic/TTX) fold change of the transcripts per million (TPM) estimate provided by the Salmon quantification algorithm (a pseudocount of 1 was added to the TPM expression counts in each condition before log transformation). Representative boxplots depict: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers.

For linear regression modeling, the vectors of each epigenetic feature described in the prior paragraphs were min-max scaled to a range of -0.5 to 0.5 using the sklearn.preprocessing minmax_scale function so that the calculated coefficients of each model could be compared to each other. The ordinary least squares function of the statsmodels.formula.api package was then used to generate linear regression models from combinations of these epigenetic features as explanatory variables and expression fold change as the response variable. The performances of these models were evaluated by the coefficient (slope) and significance of each term (**Figure 5.8K**) and the percent of the transcriptional variance explained (R^2) of each model (**Figure 5.8L**).

HiC Pre-processing

Mouse[104] and human[240] paired-end reads were aligned to the mm9 and hg19 genomes, respectively, using bowtie2[272] (global parameters:–very-sensitive –L 30 –score-min L,-0.6,-0.2 –end-to-end–reorder; local parameters:–very-sensitive –L 20 –scoremin L,-0.6,-0.2 –end-to-end–reorder) through the HiC-Pro software[273] (Servant et al., 2015). Unmapped reads, non-uniquely mapped reads and PCR duplicates were filtered, uniquely aligned reads were paired, and replicates were merged (Table S1). Cis-contact matrices were assembled by binning paired reads into uniform 20 kb (human) or 10 kb (mouse) bins. After matrix assembly, poorly mapped regions were removed based on the mm9 and hg19 50-mer CRG Alignability tracks from ENCODE. The interactions of 50kb windows that uniquely aligned at a rate below 40% (mouse) and 50% (human) were set to NaN. Due to noticeably lower complexity in the human libraries, rows containing less than seven non-zero pixels within 200kb of the diagonal were completely removed during the human HiC analysis only. Matrices containing the remaining cis-contact counts were balanced using the Juicer implementation of the Knight Ruiz (KR) algorithm with default parameters[274]. The final bias factors were retained for subsequent loop calling (see next section). Balanced matrices were used for plotting (**Figure 5.1C, Figure 5.3**).

HiC Loop Calling

HiC interactions were tested for significance using methods first reported by Aiden and colleagues[10] with some minor alterations. To estimate the local background domain interaction frequency at each locus we utilized the donut expected model approach (described above,[10]) with parameters p=1, w=4 for the 20kb resolution human libraries and p=2, w=6 for the 10kb resolution mouse libraries. For each matrix entry the expected values

were calculated using both the full donut window and just the lower-left region of the donut and the higher of the two was carried forward (i.e. expected = max(donut,lower-left))[153]. However, due to the extremely high on-diagonal counts we found this approach often over-estimated the expected background at short range interactions (less than 100kb). In order to accurately capture short range interactions, we modeled the on-diagonal (less than 100kb) background expected using only the upper-triangle region of the donut footprint. Expected contact matrices were then 'deconvoluted' back to discrete counts using the bias factors generated during KR balancing (see previous section)[10]. Each entry in the cis-contact matrix (pre-balancing) was tested for significance using a poisson distribution parameterized by its corresponding deconvoluted expected value[10]. Resulting p-values were corrected for multiple testing using the Benjamini-Hochberg procedure. In order for an interaction to be called as significantly enriched above background, it was required to pass 3 thresholds: 1) a q-value threshold ($q<0.01$ human, $q<0.025$ mouse); 2) a balanced-count threshold (count>10 human, count>20 mouse); 3) a distance threshold (distance>60 kb human, distance > 40 kb mouse). Matrix entries passing these thresholds were clustered by adjacency into loops; loops made up of fewer than 2 (human) or 3 (mouse) constituent matrix entries (interactions) were removed from further analysis.

Activity-Dependent Loop Classification and Gene Expression Analysis

Both 5C and mouse HiC loops were classified by the presence of enhancers at their anchors into mutually exclusive loop classes. 5C loops (Bic-only, TTX-only, constitutive) were classified using a specific order of intersection: loops were classified as containing a Bic-specific (activity-induced) enhancer (Classes 1+2, **Figures 5.10D,E** green) if a Bic-

specific (activity-induced) enhancer fell at (at least) one of its loop anchors. Of the loops that did not intersect a Bic-specific or constitutive enhancer, the loop was then Class 3 if a TTX-specific (activity-decommissioned) enhancer intersected a loop anchor (Class 3, **Figures 5.10D,E**, purple). If the loop's anchors intersected no enhancers but did intersect a promoter (defined as +/- 2kb surrounding RefSeq TSS's downloaded from UCSC genome browser) it was classified as a 'TSS loop' (**Figure 5.10D,** orange). The remaining loops of each class (Bic-only, TTX-only, constitutive loops) were 'Unclassified' because they did not intersect a queried epigenetic feature. The three classes highlighted in subsequent analyses (**Figures 5.10D-J**) were Bic-specific enhancers in Bic-only loops (Class 1), Bic-specific enhancers in constitutive loops (Class 2), and TTX-specific enhancers in constitutive loops (Class 3). The average observed/expected signal for each looping cluster in each looping class was calculated (**Figure 5.10F**). The promoter (+/- 2kb of TSS) of each RefSeq transcript was then tested for whether it overlapped a loop anchor of each class. If multiple transcripts of the same gene shared (had overlapping) promoters, only the transcript with the maximum expression (TPM) across the Bic and TTX conditions was considered. Additionally, genes were not considered if they fell within 200kb of the edges of our 5C regions because we could not accurately capture their looping profiles. Those transcripts linked to promoters that fell at the base of each loop class were analyzed for Bic/TTX expression upregulation (**Figure 5.10G**) and Class 1 genes were analyzed for their gene ontology (GO) enrichment (**Figure 5.10J**).

Genes at the base of genome-wide mouse cortical neuron (CN) HiC loops (original data from Bonev+ 2017) were similarly classified into mutually exclusive groups based on the enhancers to which they looped (**Figures 5.10H-J**). HiC loops were first classified

based on enhancers that intersected each anchor; Class 2 anchors contain activity-induced enhancers with no activity-decommissioned enhancers, Class 3 anchors contain activity-decommissioned enhancers with no activity-induced enhancers. If an enhancer class overlapped the upstream anchor, the downstream anchor was queried for intersection with promoters. If multiple transcripts of the same gene had promoters that overlapped the same anchor, only the transcript with the highest average expression across the Bic and TTX conditions was considered.

Gene ontology enrichment was performed using WebGestalt[275] (http://www.webgestalt.org/) with the following settings: Organism of interest = mmusculus; Method of interest = overrepresentation enrichment, Functional database = geneontology, biological_process_noRedun. refSeq mRNA IDs were uploaded for each set of classified genes. The genome_protein-coding set was used as the reference set for genome-wide HiC gene classes; all genes that fell within our 5C regions were used as the reference set for 5C gene class enrichment. The enrichment ratios and $-\log_{10}$(BH FDR) values for all GO terms with an FDR < 0.05 were plotted (**Figure 5.10H, Figure 5.13**).

Rapid/Delayed Immediate Early Gene and Secondary Response Gene Analysis

We analyzed rapid primary response genes (rIEG), delayed primary response genes (dIEG), and secondary response genes (SRG) by downloading Supplemental Table 5 from Tyssowski et al. 2018[36]. Genes were removed from each class if their promoter (upstream 10kb from TSS) did not overlap an H3K27ac peak called in the Bic condition or the gene (plus 10kb promoter) did not intersect the anchor of a mouse HiC CN looping interaction. The number of loops each gene (plus 10kb promoter) intersected was recorded (**Figure**

5.14K). Additionally, the distance of each loop was calculated as the difference between the center point of the two anchors (**Figure 5.14L**). For each loop in which an rIEG, dIEG, or SRG gene was at one anchor, the other anchor was tested for an intersection with Bic-specific enhancers. The number of loop anchor paired with each gene that intersected a Bic-specific enhancer were tallied (genes which did not loop to any Bic-specific enhancers were not considered) (**Figure 5.15C**). Representative boxplots depict: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers. Expression timing of Bdnf, Arc, and Fos were calculated using Supplemental Table 2 from Tyssowski et al. 2018[36]. Each count was normalized to the maximum count for that gene across the 4 time points. The mean normalized count at each time point was plotted along with 95% confidence intervals (**Figure 5.14A**).

Disease-Associated GWAS Single Nucleotide Variant (SNV) Enrichment

Common variants associated with neurodevelopmental diseases were analyzed from the following sources:

- Schizophrenia: Schizophrenia Working Group of the Psychiatric Genomics, *Nature*, 2014 [237]
    - P-value $\leq 5 \times 10{-}8$, Table S2 from the referenced paper
- Autism Spectrum Disorder: Autism Spectrum Disorders Working Group of The Psychiatric Genomics, *Mol Autism*, 2017 [238] (European population)
    - P-value < 10-4, Additional File S3 from the referenced paper

Disease-associated SNVs (daSVs) that fell within exons or gene promoters (2 kb upstream of TSS) were discarded from analysis. RsIDs for each disease set were uploaded to

SNPSNAP[241] in order to generate 10,000 matched 'background' SNVs for each daSNV. daSNVs were matched according the 1000Genomes Phase 3 European dataset at an LD distance cut-off of r^2=0.7 and LD buddies at r^2=0.7. daSNVs that could not be background matched using SNPSNAP were discarded. Genome-wide linkage disequilibrium (LD) r^2 values for SNV pairs were downloaded from the SNIPA tool[276]. For each daSNV and background SNV, an LD block was identified as the set of nucleotides for which the SNV in question had an r^2>0.7. Background LD blocks that overlapped each other or a disease-associated block were removed. The size of each LD block, disease and background, was calculated as the number of constituent SNVs. For each daSNV, 5 background SNVs with the same size LD block were selected. If fewer than 5 background LD blocks of the exact same size existed, background LD blocks of size one greater and one smaller than the disease-associated LD block in question were included in the set of 5 size-matched background LD blocks. The size of included background blocks was iteratively increased by one until 5 size-matched background LD blocks could be selected. If fewer than 5 background LD blocks had a size within 10 of the disease-associated block, successful background matching could not occur and the process was stopped. For example, for a daSNV with an LD block of size 75, background SNVs with LD blocks of sizes 65-85 could be matched, with preference given to those of size 75, then 74/76, and so on. Disease-associated SNVs which could not be successfully matched to 5 background LD blocks were removed from further analysis. (Note: For schizophrenia-associated SNVs, the number of size-matched LD blocks was decreased to 4 per daSNV.) If more than 5 background LD blocks were equally able to be matched to a given daSNV, 5 were randomly chosen. Due to this randomness in the algorithm, 100 different sets of

background size-matched SNVs were chosen for each daSNV (note 100 datapoints in **Figure 5.19B**, one per background set).

LD blocks (disease and background) were tested for their presence at loop anchors in the following manner. Loops were called on germinal zone (GZ) and cortical plate (CP) fetal brain tissue HiC data from Won et al. 2016[240] (see HiC processing steps above). CP and GZ loops were then merged to create a master set of 24,544 loops spanning the two brain tissues. Additionally, 25,722 'background loops' were identified as those HiC contact matrix entries which had a p-value > 0.99 and an interaction frequency count > 0 in both CP and GZ datasets. Background loops were confirmed to display the same loop distances and loop sizes as the real loop set. Bic-specific, TTX-specific, and constitutive enhancers were lifted over to the hg19 genome build using the liftOver tool on the UCSC genome browser with default parameters. Fetal brain loops were classified by enhancer presence at its anchor(s) in the same way mouse cortical neuron HiC loops were (see above). Queried LD blocks were then classified based on their presence at loop anchors: if any SNV in the LD block overlapped a loop anchor that was shared by a TTX-specific enhancer and not a Bic-specific enhancer, the LD block was considered a Class 3 variant; if any SNV in the LD block overlapped a loop anchor that was shared by a Bic-specific enhancer and not a TTX-specific enhancer, the LD block was considered a Class 2 variant. LD blocks had to fall at the same anchor as the enhancer to be classified. Finally, those LD blocks that did not overlap a classified loop anchor were tested for their presence at the anchor of a background loop. For each class, enrichment was calculated using Fisher's Exact Test with the following contingency table:

[[disease-associated blocks in loop of class X, background blocks in loop of class X],

[disease-associated blocks in background loops, background blocks in background loops]]. The resulting odds ratios were recorded for each of the 100 background size-matched SNV sets and plotted (**Figure 5.19B**) with the median p-value of the 100 tests.

LD Score Regression

To assess the polygenic enrichments of GWAS datasets listed above within looping classes, we applied LD score regression[242, 277]. LDSR was run on European subset of summary statistics from each GWAS. We used precomputed LD scores based on the European ancestry samples of the 1000 Genomes Project[278, 279] restricted to HapMap3 SNVs and generated partitioned LD scores for each looping class. All default LDSR parameters were used. LDSC version 1.0.0 was used (https://github.com/bulik/ldsc).

We conducted enrichment analyses of the heritability for SNVs located in each looping class. We regressed the $\chi 2$ from the GWAS summary statistics on to looping class-specific LD scores, with baseline scores (original 53 annotation model), regression weights and allele frequencies based on European ancestry 1000 Genome Project data. The enrichment of a looping class was defined as the proportion of SNV heritability in the category divided by the proportion of SNVs in that category; we report enrichment values and statistical significance of this enrichment as p-values (**Figure 5.19C**).

Disease-Associated Gene Expression

For each loop that was found to have a disease associated LD block and classified enhancer at one anchor (see previous section), the other anchor of the same loop was tested for intersection with promoters (+/- 2kb from TSS of human RefSeq database, downloaded

from UCSC genome browser). To identify as many target genes as possible, disease-associated LD blocks that could not be size-matched in the previous section *were* included here because no enrichment against background SNVs was being calculated (however, those SNVs that were not in the 1000Genomes database and therefore could not be assigned LD blocks or matched in SNPSNAP were still excluded, along with all daSNVs that overlapped exons and promoters). Promoters that colocalized on the other side of classified loops are annotated in **Figure 5.19D**. Human gene symbols were matched to mouse homologs using the Jackson labs complete list of human and mouse homologs (http://www.informatics.jax.org/downloads/reports/HOM_MouseHumanSequence.rpt). Mouse homologs of classified genes that fell in loops across from disease-associated LD blocks could then be stratified by their Bic/TTX expression (TPM) fold change and plotted (**Figure 5.19D**).

APPENDIX IV: REFERENCES

1. Amamoto, R. & Arlotta, P. Development-inspired reprogramming of the mammalian central nervous system. *Science* **343**, 1239882 (2014).
2. Telese, F., Gamliel, A., Skowronska-Krawczyk, D., Garcia-Bassets, I. & Rosenfeld, M.G. "Seq-ing" insights into the epigenetics of neuronal gene regulation. *Neuron* **77**, 606-623 (2013).
3. Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787-800 (2007).
4. de Laat, W. & Dekker, J. 3C-based technologies to study the shape of the genome. *Methods* **58**, 189-191 (2012).
5. Dekker, J., Marti-Renom, M.A. & Mirny, L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**, 390-403 (2013).
6. van Steensel, B. & Dekker, J. Genomics tools for unraveling chromosome architecture. *Nat Biotechnol* **28**, 1089-1095 (2010).
7. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).
8. Nora, E.P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385 (2012).
9. Phillips-Cremins, J.E. et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-1295 (2013).
10. Rao, S.S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).
11. Phanstiel, D.H. et al. Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Mol Cell* **67**, 1037-1048 e1036 (2017).
12. Dowen, J.M. et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374-387 (2014).
13. Nora, E.P. et al. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e922 (2017).
14. Narendra, V. et al. Transcription. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017-1021 (2015).
15. Flavahan, W.A. et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110-114 (2016).
16. Beagan, J.A. et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res* (2017).
17. Weintraub, A.S. et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573-1588 e1528 (2017).
18. Rao, S.S.P. et al. Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320

e324 (2017).
19. Schwarzer, W. et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51-56 (2017).
20. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-1458 (2016).
21. Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265-269 (2016).
22. Lupianez, D.G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012-1025 (2015).
23. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113 (2012).
24. Smith, E.M., Lajoie, B.R., Jain, G. & Dekker, J. Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. *Am J Hum Genet* **98**, 185-201 (2016).
25. Javierre, B.M. et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e1319 (2016).
26. Joshi, O. et al. Dynamic Reorganization of Extremely Long-Range Promoter-Promoter Interactions between Two States of Pluripotency. *Cell Stem Cell* **17**, 748-757 (2015).
27. Flavell, S.W. & Greenberg, M.E. Signaling mechanisms linking neuronal activity to gene expression and plasticity of the nervous system. *Annu Rev Neurosci* **31**, 563-590 (2008).
28. Greenberg, M.E. & Ziff, E.B. Stimulation of 3T3 cells induces transcription of the c-fos proto-oncogene. *Nature* **311**, 433-438 (1984).
29. Muller, R., Bravo, R., Burckhardt, J. & Curran, T. Induction of c-fos gene and protein by growth factors precedes activation of c-myc. *Nature* **312**, 716-720 (1984).
30. Curran, T. & Morgan, J.I. Superinduction of c-fos by nerve growth factor in the presence of peripherally active benzodiazepines. *Science* **229**, 1265-1268 (1985).
31. Morgan, J.I., Cohen, D.R., Hempstead, J.L. & Curran, T. Mapping patterns of c-fos expression in the central nervous system after seizure. *Science* **237**, 192-197 (1987).
32. Sagar, S.M., Sharp, F.R. & Curran, T. Expression of c-fos protein in brain: metabolic mapping at the cellular level. *Science* **240**, 1328-1331 (1988).
33. Link, W. et al. Somatodendritic expression of an immediate early gene is regulated by synaptic activity. *Proc Natl Acad Sci U S A* **92**, 5734-5738 (1995).
34. Lyford, G.L. et al. Arc, a growth factor and activity-regulated gene, encodes a novel cytoskeleton-associated protein that is enriched in neuronal dendrites. *Neuron* **14**, 433-445 (1995).

35. Plath, N. et al. Arc/Arg3.1 is essential for the consolidation of synaptic plasticity and memories. *Neuron* **52**, 437-444 (2006).
36. Tyssowski, K.M. et al. Different Neuronal Activity Patterns Induce Different Gene Expression Programs. *Neuron* **98**, 530-546 e511 (2018).
37. Sams, D.S. et al. Neuronal CTCF Is Necessary for Basal and Experience-Dependent Gene Regulation, Memory Formation, and Genomic Structure of BDNF and Arc. *Cell Rep* **17**, 2418-2430 (2016).
38. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294 (2013).
39. Kagey, M.H. et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-435 (2010).
40. Pope, B.D. et al. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402-405 (2014).
41. Sofueva, S. et al. Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J* **32**, 3119-3129 (2013).
42. Gibcus, J.H. & Dekker, J. The hierarchy of the 3D genome. *Mol Cell* **49**, 773-782 (2013).
43. Dostie, J. et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* **16**, 1299-1309 (2006).
44. Dixon, J.R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).
45. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
46. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev* **26**, 11-24 (2012).
47. Wolffe, A.P. & Guschin, D. Review: chromatin structural features and targets that regulate transcription. *J Struct Biol* **129**, 102-122 (2000).
48. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292-301 (2001).
49. Mirny, L.A. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res* **19**, 37-51 (2011).
50. Gondor, A. & Ohlsson, R. Transcription in the loop. *Nat Genet* **38**, 1229-1230 (2006).
51. O'Sullivan, J.M. et al. Gene loops juxtapose promoters and terminators in yeast. *Nat Genet* **36**, 1014-1018 (2004).
52. Sexton, T., Bantignies, F. & Cavalli, G. Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation. *Semin Cell Dev Biol* **20**, 849-855 (2009).
53. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. & de Laat, W. Looping and

interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* **10**, 1453-1465 (2002).
54. Doyle, B., Fudenberg, G., Imakaev, M. & Mirny, L.A. Chromatin loops as allosteric modulators of enhancer-promoter interactions. *PLoS Comput Biol* **10**, e1003867 (2014).
55. Sexton, T. et al. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-472 (2012).
56. Wood, A.M. et al. Regulation of chromatin organization and inducible gene expression by a Drosophila insulator. *Mol Cell* **44**, 29-38 (2011).
57. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013).
58. Fraser, P. & Bickmore, W. Nuclear organization of the genome and the potential for gene regulation. *Nature* **447**, 413-417 (2007).
59. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951 (2008).
60. Zhang, Y. et al. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908-921 (2012).
61. Kind, J. et al. Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell* **163**, 134-147 (2015).
62. Vietri Rudan, M. et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* **10**, 1297-1309 (2015).
63. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**, 1059-1065 (2011).
64. Hou, C., Li, L., Qin, Z.S. & Corces, V.G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell* **48**, 471-484 (2012).
65. Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* **17**, 520-527 (2001).
66. Filippova, G.N. et al. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol* **16**, 2802-2813 (1996).
67. Phillips, J.E. & Corces, V.G. CTCF: master weaver of the genome. *Cell* **137**, 1194-1211 (2009).
68. Hadjur, S. et al. Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature* **460**, 410-413 (2009).
69. Kurukuti, S. et al. CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer

access to Igf2. *Proc Natl Acad Sci U S A* **103**, 10684-10689 (2006).
70. Splinter, E. et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* **20**, 2349-2354 (2006).
71. Handoko, L. et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**, 630-638 (2011).
72. Zuin, J. et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci U S A* **111**, 996-1001 (2014).
73. Kim, T.H. et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231-1245 (2007).
74. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).
75. Cuddapah, S. et al. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19**, 24-32 (2009).
76. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* **36**, 5221-5231 (2008).
77. Chen, H., Tian, Y., Shu, W., Bo, X. & Wang, S. Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PLoS ONE* **7**, e41374 (2012).
78. Wang, H. et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res* **22**, 1680-1688 (2012).
79. Rubio, E.D. et al. CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci U S A* **105**, 8309-8314 (2008).
80. Renda, M. et al. Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *J Biol Chem* **282**, 33336-33345 (2007).
81. Nakahashi, H. et al. A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* **3**, 1678-1689 (2013).
82. Rhee, H.S. & Pugh, B.F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408-1419 (2011).
83. Schmidt, D. et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335-348 (2012).
84. Guo, Y. et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900-910 (2015).
85. de Wit, E. et al. CTCF Binding Polarity Determines Chromatin Looping. *Mol Cell* **60**, 676-684 (2015).
86. Tang, Z. et al. CTCF-Mediated Human 3D Genome Architecture Reveals

Chromatin Topology for Transcription. *Cell* **163**, 1611-1627 (2015).
87. Sanborn, A.L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**, E6456-6465 (2015).
88. Nasmyth, K. Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annu Rev Genet* **35**, 673-745 (2001).
89. 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).
90. Alipour, E. & Marko, J.F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res* **40**, 11202-11212 (2012).
91. Fudenberg, G. et al. Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* **15**, 2038-2049 (2016).
92. Goloborodko, A., Marko, J.F. & Mirny, L.A. Chromosome Compaction by Active Loop Extrusion. *Biophys J* **110**, 2162-2168 (2016).
93. Parelho, V. et al. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422-433 (2008).
94. Stedman, W. et al. Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *EMBO J* **27**, 654-666 (2008).
95. Wendt, K.S. et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796-801 (2008).
96. Haarhuis, J.H.I. et al. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169**, 693-707 e614 (2017).
97. Terakawa, T. et al. The condensin complex is a mechanochemical motor that translocates along DNA. *Science* **358**, 672-676 (2017).
98. Ganji, M. et al. Real-time imaging of DNA loop extrusion by condensin. *Science* **360**, 102-105 (2018).
99. Stigler, J., Camdere, G.O., Koshland, D.E. & Greene, E.C. Single-Molecule Imaging Reveals a Collapsed Conformational State for DNA-Bound Cohesin. *Cell Rep* **15**, 988-998 (2016).
100. Davidson, I.F. et al. Rapid movement and transcriptional re-localization of human cohesin on DNA. *EMBO J* **35**, 2671-2685 (2016).
101. Kanke, M., Tahara, E., Huis In't Veld, P.J. & Nishiyama, T. Cohesin acetylation and Wapl-Pds5 oppositely regulate translocation of cohesin along DNA. *EMBO J* **35**, 2686-2698 (2016).
102. Rowley, M.J. et al. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Mol Cell* **67**, 837-852 e837 (2017).
103. Eagen, K.P., Aiden, E.L. & Kornberg, R.D. Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proc Natl Acad Sci U S A* **114**, 8764-8769 (2017).

104. Bonev, B. et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572 e524 (2017).
105. Norton, H.K. et al. Detecting hierarchical genome folding with network modularity. *Nat Methods* **15**, 119-122 (2018).
106. Kruse, K. et al. Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. *bioRxiv*, 523712 (2019).
107. Zhang, Y. et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet* **51**, 1380-1388 (2019).
108. Hsieh, T.-H.S. et al. Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *bioRxiv*, 638775 (2019).
109. Krietenstein, N. et al. Ultrastructural details of mammalian chromosome architecture. *bioRxiv*, 639922 (2019).
110. Flyamer, I.M. et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110-114 (2017).
111. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362** (2018).
112. Szabo, Q. et al. TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Sci Adv* **4**, eaar8082 (2018).
113. Mateo, L.J. et al. Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* **568**, 49-54 (2019).
114. Giorgetti, L. et al. Structural organization of the inactive X chromosome in the mouse. *Nature* **535**, 575-579 (2016).
115. Hug, C.B., Grimaldi, A.G., Kruse, K. & Vaquerizas, J.M. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell* **169**, 216-228 e219 (2017).
116. Ke, Y. et al. 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell* **170**, 367-381 e320 (2017).
117. Du, Z. et al. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547**, 232-235 (2017).
118. Krijger, P.H. et al. Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming. *Cell Stem Cell* **18**, 597-610 (2016).
119. Stadhouders, R. et al. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nat Genet* **50**, 238-249 (2018).
120. Hu, G. et al. Transformation of Accessible Chromatin and 3D Nucleome Underlies Lineage Commitment of Early T Cells. *Immunity* **48**, 227-242 e228 (2018).
121. Ryba, T. et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res*

20, 761-770 (2010).
122. Therizols, P. et al. Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science* **346**, 1238-1242 (2014).
123. Wijchers, P.J. et al. Cause and Consequence of Tethering a SubTAD to Different Nuclear Compartments. *Mol Cell* **61**, 461-473 (2016).
124. Reddy, K.L., Zullo, J.M., Bertolino, E. & Singh, H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* **452**, 243-247 (2008).
125. Brangwynne, C.P. et al. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**, 1729-1732 (2009).
126. Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K. & Sharp, P.A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13-23 (2017).
127. Falk, M. et al. Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature* **570**, 395-399 (2019).
128. Mitchell, J.A. & Fraser, P. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes Dev* **22**, 20-25 (2008).
129. Gerasimova, T.I., Byrd, K. & Corces, V.G. A chromatin insulator determines the nuclear localization of DNA. *Mol Cell* **6**, 1025-1035 (2000).
130. van Steensel, B. & Belmont, A.S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**, 780-791 (2017).
131. Wang, H. et al. CRISPR-Mediated Programmable 3D Genome Positioning and Nuclear Organization. *Cell* **175**, 1405-1417 e1414 (2018).
132. Leemans, C. et al. Promoter-Intrinsic and Local Chromatin Features Determine Gene Repression in LADs. *Cell* **177**, 852-864 e814 (2019).
133. van Steensel, B. & Furlong, E.E.M. The role of transcription in shaping the spatial organization of the genome. *Nat Rev Mol Cell Biol* **20**, 327-337 (2019).
134. Symmons, O. et al. Functional and topological characteristics of mammalian regulatory domains. *Genome Res* **24**, 390-400 (2014).
135. van Bemmel, J.G. et al. The bipartite TAD organization of the X-inactivation center ensures opposing developmental regulation of Tsix and Xist. *Nat Genet* **51**, 1024-1034 (2019).
136. Kraft, K. et al. Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. *Nat Cell Biol* **21**, 305-310 (2019).
137. Laugsch, M. et al. Modeling the Pathological Long-Range Regulatory Effects of Human Structural Variation with Patient-Specific hiPSCs. *Cell Stem Cell* **24**, 736-752 e712 (2019).
138. Despang, A. et al. Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat Genet* **51**, 1263-1271

(2019).
139. Sun, J.H. et al. Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell* (2018).
140. Norton, H.K. & Phillips-Cremins, J.E. Crossed wires: 3D genome misfolding in human disease. *J Cell Biol* **216**, 3441-3452 (2017).
141. Fulco, C.P. et al. Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations. *bioRxiv*, 529990 (2019).
142. Chen, H. et al. Dynamic interplay between enhancer-promoter topology and gene activity. *Nat Genet* **50**, 1296-1303 (2018).
143. Alexander, J.M. et al. Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. *Elife* **8** (2019).
144. Deng, W. et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244 (2012).
145. Kim, J.H. et al. LADL: light-activated dynamic looping for endogenous gene expression control. *Nat Methods* **16**, 633-639 (2019).
146. Cuartero, S. et al. Control of inducible gene expression links cohesin to hematopoietic progenitor self-renewal and differentiation. *Nat Immunol* **19**, 932-941 (2018).
147. Heinz, S. et al. Transcription Elongation Can Affect Genome 3D Structure. *Cell* **174**, 1522-1536 e1522 (2018).
148. Barutcu, A.R., Blencowe, B.J. & Rinn, J.L. Differential contribution of steady-state RNA and active transcription in chromatin organization. *EMBO Rep* **20**, e48068 (2019).
149. Williamson, I. et al. Developmentally regulated Shh expression is robust to TAD perturbations. *Development* **146** (2019).
150. Paliou, C. et al. Preformed chromatin topology assists transcriptional robustness of Shh during limb development. *Proc Natl Acad Sci U S A* **116**, 12390-12399 (2019).
151. Despang, A. et al. Functional dissection of TADs reveals non-essential and instructive roles in regulating gene expression. *bioRxiv*, 566562 (2019).
152. Ghavi-Helm, Y. et al. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat Genet* **51**, 1272-1282 (2019).
153. Beagan, J.A. et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res* **27**, 1139-1152 (2017).
154. Rais, Y. et al. Deterministic direct reprogramming of somatic cells to pluripotency. *Nature* **502**, 65-70 (2013).
155. Deng, W. et al. Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158**, 849-860 (2014).
156. Beagan, J.A. et al. Local Genome Topology Can Exhibit an Incompletely Rewired 3D-Folding State during Somatic Cell Reprogramming. *Cell Stem Cell* **18**, 611-624 (2016).

157. Phanstiel, D.H. et al. Static And Dynamic DNA Loops Form AP-1 Bound Activation Hubs During Macrophage Development. *bioRxiv* (2017).
158. Flavahan, W.A. et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* (2015).
159. Maurano, M.T. et al. Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Rep* **12**, 1184-1195 (2015).
160. Mehra, P.G., T.; Basu, A.; Jha, V.; Banerjee, A.; Sindhava, V.; Gray, F.; Berry, C.T.; Sen, R.; Atchison, M.L. YY1 controls Eμ-3′RR DNA loop formation and immunoglobulin heavy chain class switch recombination. *Blood Advances* **2016**, 15-20 (2016).
161. Donohoe, M.E., Zhang, L.F., Xu, N., Shi, Y. & Lee, J.T. Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Mol Cell* **25**, 43-56 (2007).
162. Lee, J., Krivega, I., Dale, R.K. & Dean, A. The LDB1 Complex Co-opts CTCF for Erythroid Lineage-Specific Long-Range Enhancer Interactions. *Cell Rep* **19**, 2490-2502 (2017).
163. Denker, A. & de Laat, W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev* **30**, 1357-1382 (2016).
164. Ong, C.T. & Corces, V.G. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* **15**, 234-246 (2014).
165. Splinter, E., de Wit, E., van de Werken, H.J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods* **58**, 221-230 (2012).
166. Chen, X. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117 (2008).
167. Watson, L.A. et al. Dual effect of CTCF loss on neuroprogenitor differentiation and survival. *J Neurosci* **34**, 2860-2870 (2014).
168. Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N. & Yagi, T. CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. *Cell Rep* **2**, 345-357 (2012).
169. Guo, Y. et al. CTCF/cohesin-mediated DNA looping is required for protocadherin alpha promoter choice. *Proc Natl Acad Sci U S A* **109**, 21081-21086 (2012).
170. Ying, Q.L. et al. The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519-523 (2008).
171. Galonska, C., Ziller, M.J., Karnik, R. & Meissner, A. Ground State Conditions Induce Rapid Reorganization of Core Pluripotency Factor Binding before Global Epigenetic Reprogramming. *Cell Stem Cell* **17**, 462-470 (2015).
172. Sharma, A., Klein, S.S., Barboza, L., Lohdi, N. & Toth, M. Principles Governing DNA Methylation during Neuronal Lineage and Subtype Specification. *J Neurosci* **36**, 1711-1722 (2016).
173. Stadtfeld, M., Maherali, N., Breault, D.T. & Hochedlinger, K. Defining molecular

cornerstones during fibroblast to iPS cell reprogramming in mouse. *Cell Stem Cell* **2**, 230-240 (2008).
174. Shen, Y. et al. A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116-120 (2012).
175. Ficz, G. et al. FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* **13**, 351-359 (2013).
176. Habibi, E. et al. Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13**, 360-369 (2013).
177. Leitch, H.G. et al. Naive pluripotency is associated with global DNA hypomethylation. *Nat Struct Mol Biol* **20**, 311-316 (2013).
178. Li, Y. et al. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One* **9**, e114485 (2014).
179. Blinka, S., Reimer, M.H., Jr., Pulakanti, K. & Rao, S. Super-Enhancers at the Nanog Locus Differentially Regulate Neighboring Pluripotency-Associated Genes. *Cell Rep* **17**, 19-28 (2016).
180. Lendahl, U., Zimmerman, L.B. & McKay, R.D. CNS stem cells express a new class of intermediate filament protein. *Cell* **60**, 585-595 (1990).
181. Park, D. et al. Nestin is required for the proper self-renewal of neural stem cells. *Stem Cells* **28**, 2162-2171 (2010).
182. Zhou, Q., Wang, S. & Anderson, D.J. Identification of a novel family of oligodendrocyte lineage-specific basic helix-loop-helix transcription factors. *Neuron* **25**, 331-343 (2000).
183. Meissner, A. et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770 (2008).
184. Ziller, M.J. et al. Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature* **518**, 355-359 (2015).
185. Medvedovic, J. et al. Flexible long-range loops in the VH gene region of the Igh locus facilitate the generation of a diverse antibody repertoire. *Immunity* **39**, 229-244 (2013).
186. He, Y. et al. The transcription factor Yin Yang 1 is essential for oligodendrocyte progenitor differentiation. *Neuron* **55**, 217-230 (2007).
187. Schwalie, P.C. et al. Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome Biol* **14**, R148 (2013).
188. Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell* **38**, 603-613 (2010).
189. Zhang, Y. et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504**, 306-310 (2013).

190. Apostolou, E. et al. Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. *Cell Stem Cell* **12**, 699-712 (2013).
191. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676 (2006).
192. Hanna, J. et al. Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* **462**, 595-601 (2009).
193. Koche, R.P. et al. Reprogramming factor expression initiates widespread targeted chromatin remodeling. *Cell Stem Cell* **8**, 96-105 (2011).
194. Soufi, A., Donahue, G. & Zaret, K.S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151**, 994-1004 (2012).
195. Buganim, Y. et al. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209-1222 (2012).
196. Lujan, E. et al. Early reprogramming regulators identified by prospective isolation and mass cytometry. *Nature* **521**, 352-356 (2015).
197. Polo, J.M. et al. A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* **151**, 1617-1632 (2012).
198. Kim, K. et al. Epigenetic memory in induced pluripotent stem cells. *Nature* **467**, 285-290 (2010).
199. Polo, J.M. et al. Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nat Biotechnol* **28**, 848-855 (2010).
200. Bock, C. et al. Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439-452 (2011).
201. Stadtfeld, M. et al. Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature* **465**, 175-181 (2010).
202. Wei, Z. et al. Klf4 organizes long-range chromosomal interactions with the oct4 locus in reprogramming and pluripotency. *Cell Stem Cell* **13**, 36-47 (2013).
203. Zhang, H. et al. Intrachromosomal looping is required for activation of endogenous pluripotency genes during reprogramming. *Cell Stem Cell* **13**, 30-35 (2013).
204. de Wit, E. et al. The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **501**, 227-231 (2013).
205. Denholtz, M. et al. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. *Cell Stem Cell* **13**, 602-616 (2013).
206. Eminli, S., Utikal, J., Arnold, K., Jaenisch, R. & Hochedlinger, K. Reprogramming of neural progenitor cells into induced pluripotent stem cells in the absence of exogenous Sox2 expression. *Stem Cells* **26**, 2467-2474 (2008).

207. Marks, H. et al. The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**, 590-604 (2012).
208. Tanabe, K., Nakamura, M., Narita, M., Takahashi, K. & Yamanaka, S. Maturation, not initiation, is the major roadblock during reprogramming toward pluripotency from human fibroblasts. *Proc Natl Acad Sci U S A* **110**, 12172-12179 (2013).
209. Phillips-Cremins, J.E. & Corces, V.G. Chromatin insulators: linking genome organization to cellular function. *Mol Cell* **50**, 461-474 (2013).
210. Bell, A.C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**, 482-485 (2000).
211. Apostolou, E. & Hochedlinger, K. Chromatin dynamics during cellular reprogramming. *Nature* **502**, 462-471 (2013).
212. Cahan, P. et al. CellNet: network biology applied to stem cell engineering. *Cell* **158**, 903-915 (2014).
213. Krijger, P.H. et al. Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming. *Cell Stem Cell* (2016).
214. Yap, E.L. & Greenberg, M.E. Activity-Regulated Transcription: Bridging the Gap between Neural Activity and Behavior. *Neuron* **100**, 330-348 (2018).
215. Fowler, T., Sen, R. & Roy, A.L. Regulation of primary response genes. *Mol Cell* **44**, 348-360 (2011).
216. Herschman, H.R. Primary response genes induced by growth factors and tumor promoters. *Annu Rev Biochem* **60**, 281-319 (1991).
217. Kawashima, T. et al. Synaptic activity-responsive element in the Arc/Arg3.1 promoter essential for synapse-to-nucleus signaling in activated neurons. *Proc Natl Acad Sci U S A* **106**, 316-321 (2009).
218. Pintchovski, S.A., Peebles, C.L., Kim, H.J., Verdin, E. & Finkbeiner, S. The serum response factor and a putative novel transcription factor regulate expression of the immediate-early gene Arc/Arg3.1 in neurons. *J Neurosci* **29**, 1525-1537 (2009).
219. Kim, T.K. et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187 (2010).
220. Malik, A.N. et al. Genome-wide identification and characterization of functional neuronal activity-dependent enhancers. *Nat Neurosci* **17**, 1330-1339 (2014).
221. Su, Y. et al. Neuronal activity modifies the chromatin accessibility landscape in the adult brain. *Nat Neurosci* **20**, 476-483 (2017).
222. Lanctot, C., Cheutin, T., Cremer, M., Cavalli, G. & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet* **8**, 104-115 (2007).
223. Schardin, M., Cremer, T., Hager, H.D. & Lang, M. Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. *Hum Genet* **71**, 281-287 (1985).
224. Cremer, T. et al. The 4D nucleome: Evidence for a dynamic nuclear landscape

based on co-aligned active and inactive nuclear compartments. *FEBS Lett* **589**, 2931-2943 (2015).
225. Yamada, T. et al. Sensory experience remodels genome architecture in neural circuit to drive motor learning. *Nature* **569**, 708-713 (2019).
226. Shepherd, J.D. et al. Arc/Arg3.1 mediates homeostatic synaptic scaling of AMPA receptors. *Neuron* **52**, 475-484 (2006).
227. Straughan, D.W., Neal, M.J., Simmonds, M.A., Collins, G.G. & Hill, R.G. Evaluation of bicuculline as a GABA antagonist. *Nature* **233**, 352-354 (1971).
228. Narahashi, T., Moore, J.W. & Scott, W.R. Tetrodotoxin Blockage of Sodium Conductance Increase in Lobster Giant Axons. *J Gen Physiol* **47**, 965-974 (1964).
229. Shepherd, J.D. & Huganir, R.L. The cell biology of synaptic plasticity: AMPA receptor trafficking. *Annu Rev Cell Dev Biol* **23**, 613-643 (2007).
230. Kim, J.H. et al. 5C-ID: Increased resolution Chromosome-Conformation-Capture-Carbon-Copy with in situ 3C and double alternating primer design. *Methods* **142**, 39-46 (2018).
231. Phillips-Cremins, J.E. & Gilgenast, T.G. Systematic evaluation of statistical methods for identifying looping interactions in 5C data. *bioRxiv* (2017).
232. Gilgenast, T.G. & Phillips-Cremins, J.E. Systematic Evaluation of Statistical Methods for Identifying Looping Interactions in 5C Data. *Cell Syst* **8**, 197-211 e113 (2019).
233. Fernandez, L.R., Gilgenast, T.G. & Phillips-Cremins, J.E. 3DeFDR: Identifying cell type-specific looping interactions with empirical false discovery rate guided thresholding. *bioRxiv*, 501056 (2018).
234. Joo, J.Y., Schaukowitch, K., Farbiak, L., Kilaru, G. & Kim, T.K. Stimulus-specific combinatorial functionality of neuronal c-fos enhancers. *Nat Neurosci* **19**, 75-83 (2016).
235. Pimentel, H., Bray, N.L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods* **14**, 687-690 (2017).
236. Ebert, D.H. & Greenberg, M.E. Activity-dependent neuronal signalling and autism spectrum disorder. *Nature* **493**, 327-337 (2013).
237. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).
238. Autism Spectrum Disorders Working Group of The Psychiatric Genomics, C. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol Autism* **8**, 21 (2017).
239. Maurano, M.T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).
240. Won, H. et al. Chromosome conformation elucidates regulatory relationships in

developing human brain. *Nature* **538**, 523-527 (2016).
241. Pers, T.H., Timshel, P. & Hirschhorn, J.N. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* **31**, 418-420 (2015).
242. Finucane, H.K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-1235 (2015).
243. Araujo, D.J. et al. Foxp1 in Forebrain Pyramidal Neurons Controls Gene Expression Required for Spatial Learning and Synaptic Plasticity. *J Neurosci* **37**, 10917-10931 (2017).
244. Sinning, A., Liebmann, L. & Hubner, C.A. Disruption of Slc4a10 augments neuronal excitability and modulates synaptic short-term plasticity. *Front Cell Neurosci* **9**, 223 (2015).
245. Hong, E.J., McCord, A.E. & Greenberg, M.E. A biological function for the neuronal activity-dependent component of Bdnf transcription in the development of cortical inhibition. *Neuron* **60**, 610-624 (2008).
246. Nestler, E.J., Pena, C.J., Kundakovic, M., Mitchell, A. & Akbarian, S. Epigenetic Basis of Mental Illness. *Neuroscientist* **22**, 447-463 (2016).
247. Gandal, M.J., Leppa, V., Won, H., Parikshak, N.N. & Geschwind, D.H. The road to precision psychiatry: translating genetics into disease mechanisms. *Nat Neurosci* **19**, 1397-1407 (2016).
248. Brainstorm, C. et al. Analysis of shared heritability in common disorders of the brain. *Science* **360** (2018).
249. Gandal, M.J. et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693-697 (2018).
250. Xiao, X., Chang, H. & Li, M. Molecular mechanisms underlying noncoding risk variations in psychiatric genetic studies. *Mol Psychiatry* **22**, 497-511 (2017).
251. Luscher, C. & Malenka, R.C. NMDA receptor-dependent long-term potentiation and long-term depression (LTP/LTD). *Cold Spring Harb Perspect Biol* **4** (2012).
252. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-286 (2014).
253. Mahat, D.B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat Protoc* **11**, 1455-1476 (2016).
254. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 17 (2010).
255. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
256. Sauria, M.E., Phillips-Cremins, J.E., Corces, V.G. & Taylor, J. HiFive: a tool suite for easy and efficient HiC and 5C data analysis. *Genome Biol* **16**, 237 (2015).
257. Mendenhall, E.M. et al. GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet* **6**, e1001244 (2010).

258. Sigova, A.A. et al. Transcription factor trapping by RNA in gene regulatory elements. *Science* **350**, 978-981 (2015).
259. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311 (2002).
260. van Berkum, N.L. & Dekker, J. Determining spatial chromatin organization of large genomic regions using 5C technology. *Methods Mol Biol* **567**, 189-213 (2009).
261. Gheldof, N. et al. Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. *Nucleic Acids Res* **38**, 4325-4336 (2010).
262. Lajoie, B.R., van Berkum, N.L., Sanyal, A. & Dekker, J. My5C: web tools for chromosome conformation capture studies. *Nat Methods* **6**, 690-691 (2009).
263. Bau, D. et al. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* **18**, 107-114 (2011).
264. Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. *Nat Protoc* **2**, 988-1002 (2007).
265. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
266. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).
267. Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
268. Ramirez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165 (2016).
269. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**, 999-1003 (2012).
270. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).
271. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
272. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
273. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259 (2015).
274. Durand, N.C. et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98 (2016).
275. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res* **45**, W130-W137 (2017).

276. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. & Kastenmuller, G. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics* **31**, 1334-1336 (2015).
277. Bulik-Sullivan, B.K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015).
278. International HapMap, C. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58 (2010).
279. Genomes Project, C. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).