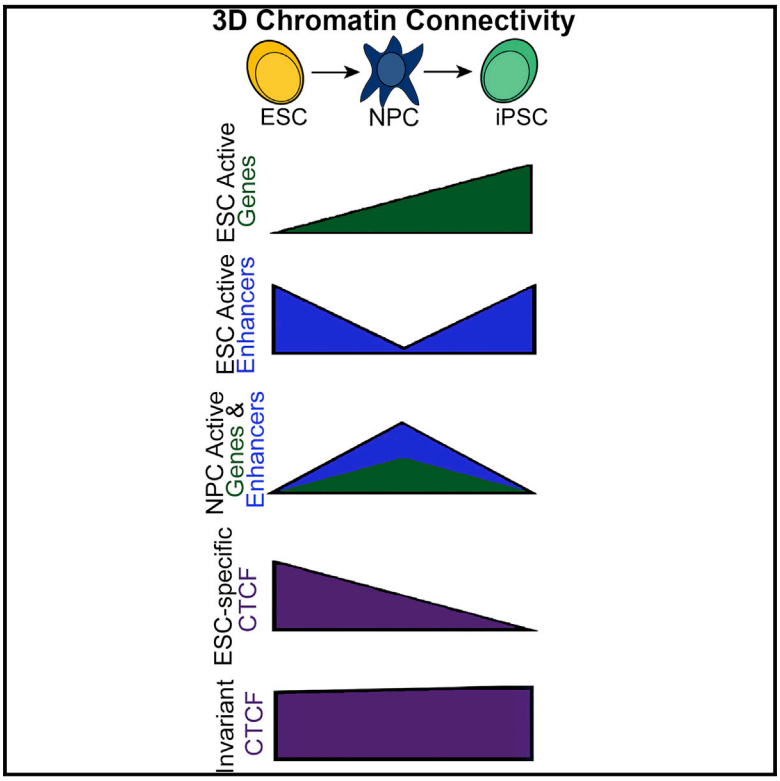
Article

Cell Stem Cell

Local Genome Topology Can Exhibit an Incompletely Rewired 3D-Folding State during Somatic Cell Reprogramming

Graphical Abstract



Authors

Jonathan A. Beagan, Thomas G. Gilgenast, Jesi Kim, ..., Victor G. Corces, Job Dekker, Jennifer E. Phillips-Cremins

Correspondence

jcremins@seas.upenn.edu

In Brief

Phillips-Cremins and colleagues report high-resolution chromatin folding maps in primary NPCs and NPC-derived induced pluripotent stem cells. They find that iPSC genomes can exhibit an imperfectly rewired 3D-folding state linked to poorly reprogrammed, ESC-specific CTCF occupancy and inaccurately reprogrammed gene expression levels. 2i/LIF conditions can fully restore distinct topological hallmarks of pluripotency.

Highlights

- 3D genome architecture is markedly reconfigured during reprogramming
- Some pluripotency genes engage in persistent, NPC-like interactions in iPSCs that break apart in 2i
- ESC-specific interactions that do not reconnect in iPSCs exhibit decreased CTCF binding
- Imperfectly rewired iPSC genome topology is linked to inaccurately reprogrammed expression

Accession Numbers

GSE68582

Beagan et al., 2016, Cell Stem Cell *18*, 611–624
May 5, 2016
http://dx.doi.org/10.1016/j.stem.2016.04.004

CellPress

Cell Stem Cell
Article

Local Genome Topology Can Exhibit an Incompletely Rewired 3D-Folding State during Somatic Cell Reprogramming

Jonathan A. Beagan,[1] Thomas G. Gilgenast,[1] Jesi Kim,[1] Zachary Plona,[1] Heidi K. Norton,[1] Gui Hu,[1] Sarah C. Hsu,[2] Emily J. Shields,[2] Xiaowen Lyu,[3] Effie Apostolou,[5,6] Konrad Hochedlinger,[5] Victor G. Corces,[3] Job Dekker,[4] and Jennifer E. Phillips-Cremins[1,2,*]

[1]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Epigenetics Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Department of Biology, Emory University, Atlanta, GA 30322, USA
[4]Howard Hughes Medical Institute, Program in Systems Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA
[5]Massachusetts General Hospital Cancer Center and Center for Regenerative Medicine, Boston, MA 02114, USA
[6]Present address: Meyer Cancer Center and Department of Medicine, Weill Cornell Medicine, New York, NY 10021, USA
*Correspondence: jcremins@seas.upenn.edu
http://dx.doi.org/10.1016/j.stem.2016.04.004

SUMMARY

Pluripotent genomes are folded in a topological hierarchy that reorganizes during differentiation. The extent to which chromatin architecture is reconfigured during somatic cell reprogramming is poorly understood. Here we integrate fine-resolution architecture maps with epigenetic marks and gene expression in embryonic stem cells (ESCs), neural progenitor cells (NPCs), and NPC-derived induced pluripotent stem cells (iPSCs). We find that most pluripotency genes reconnect to target enhancers during reprogramming. Unexpectedly, some NPC interactions around pluripotency genes persist in our iPSC clone. Pluripotency genes engaged in both "fully-reprogrammed" and "persistent-NPC" interactions exhibit over/undershooting of target expression levels in iPSCs. Additionally, we identify a subset of "poorly reprogrammed" interactions that do not reconnect in iPSCs and display only partially recovered, ESC-specific CTCF occupancy. 2i/LIF can abrogate persistent-NPC interactions, recover poorly reprogrammed interactions, reinstate CTCF occupancy, and restore expression levels. Our results demonstrate that iPSC genomes can exhibit imperfectly rewired 3D-folding linked to inaccurately reprogrammed gene expression.

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 further partitioned into a nested series of Megabase (Mb)-sized topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012) and smaller sub-domains (termed subTADs) (Phillips-Cremins et al., 2013; Rao et al., 2014). TADs/subTADs vary widely in size (i.e., 40 kilobase [kb] to 3 Mb) and are characterized by highly interacting chromatin fragments demarcated by boundaries of abruptly decreased contact frequency. Long-range looping interactions connect distal genomic loci within and between TADs/subTADs (Jin et al., 2013; Phillips-Cremins et al., 2013; Rao et al., 2014; Sanyal et al., 2012). 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: (1) "A" compartments enriched for open chromatin, highly expressed genes, and early replication timing and (2) "B" compartments enriched for closed chromatin, late replication timing, and co-localization with the nuclear periphery (Dixon et al., 2015; Lieberman-Aiden et al., 2009; Pope et al., 2014; Rao et al., 2014). 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 cell (ESC) differentiation. Up to 25% of compartments in human ESCs switch their A/B orientation upon differentiation (Dixon et al., 2015). 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 (Dixon et al., 2015). Similarly, lamina associated domains are dynamically altered during ESC differentiation (Peric-Hupkes et al., 2010). For example, the *Oct4*, *Nanog*, and *Klf4* genes relocate to the nuclear periphery in parallel with their loss of transcriptional activity as ESCs differentiate to astrocytes. TADs are largely invariant across cell types and often maintain their boundaries irrespective of the expression of their resident genes (Dixon et al., 2012). By contrast, long-range looping interactions within and between subTADs are highly dynamic during ESC differentiation (Phillips-Cremins et al., 2013; Zhang et al., 2013b). Pluripotency genes connect to their target enhancers through long-range interactions and disruption of these interactions leads to a marked decrease in gene expression (Apostolou et al., 2013; Kagey et al., 2010).

Thus, data are 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 ESCs in development, somatic cells can also be reprogrammed in the reverse direction to induced pluripotent stem cells (iPSCs) via the ectopic expression of key transcription factors (Takahashi and Yamanaka, 2006). Since the initial pioneering discovery, many population-based and single-cell genomics studies have explored the molecular underpinnings of transcription factor-mediated reprogramming (Hanna et al., 2009; Koche et al., 2011; Rais et al., 2013; Soufi et al., 2012). Recent efforts have uncovered changes in transcription, cell surface markers, and classic epigenetic modifications during intermediate stages in the reprogramming process (Buganim et al., 2012; Lujan et al., 2015; Polo et al., 2012). Although there is some evidence of epigenetic traces from the somatic cell of origin (Bock et al., 2011; Kim et al., 2010; Polo et al., 2010), the emerging model is that ESC-like epigenetic and transcriptional states can be generally reset under proper reprogramming conditions (Stadtfeld et al., 2010).

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 pluripotency genes such as *Nanog* and/or *Oct4* and target enhancers can be reset during reprogramming and precede reactivation of the involved genes (Apostolou et al., 2013; de Wit et al., 2013; Denholtz et al., 2013; Wei et al., 2013; Zhang et al., 2013a). 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 kb resolution chromatin architecture maps in primary neural progenitor cells (NPCs), iPSCs derived from primary NPCs, and pluripotent ESCs. 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 iPSCs. In many cases, pluripotency genes re-engage in fully reprogrammed interactions with their target ESC-specific enhancers. Unexpectedly, we also observe NPC interactions around key pluripotency genes (e.g., *Sox2* and *Klf4*) that remain persistently tethered in our iPSC clone. Pluripotency genes engaged in "persistent NPC-like" interactions can exhibit over/undershooting of gene expression levels in iPSCs, despite the fact that they may have also re-established contact with their target ESC-specific enhancer(s). We also uncover a subset of "poorly reprogrammed" interactions that break apart during differentiation and do not fully reconnect in our iPSC clone. Many poorly reprogrammed interactions exhibit ESC-specific CTCF occupancy that is lost during differentiation and only partially recovered in iPSCs. Importantly, 2i/LIF conditions can (1) abrogate persistent NPC-like interactions, (2) recover poorly reprogrammed interactions, (3) reinstate inadequately reprogrammed CTCF occupancy, and (4) restore precise gene expression levels.

RESULTS

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, iPSCs derived from primary NPCs, and ESCs (Figure 1A). To achieve a comparable genetic background to our pluripotency model (V6.5 ESCs; 129/SvJae × C57BL/6), we selected a previously published iPSC clone derived from primary NPCs isolated from neonatal brains of Sox2-GFP indicator mice (mixed 129/SvJae × C57BL/6 genetic background) (Eminli et al., 2008; Stadtfeld et al., 2008). Hochedlinger and colleagues generated this iPSC clone via the transduction of primary Sox2-GFP NPCs with doxycycline-inducible lentiviral vectors encoding Oct4, Klf4, and c-Myc. Importantly, this iPSC clone was extensively characterized for its pluripotent properties as assessed by (1) expression of endogenous pluripotency markers (Oct4, Sox2, and Nanog), (2) demethylation of *Oct4* and *Nanog* promoters, (3) transgene-independent self-renewal, (4) in vivo teratoma formation of all three germ layers, and (5) generation of chimeric mice (Eminli et al., 2008). Our three cellular states enable a detailed analysis of how chromatin unfolds/refolds between NPCs and iPSCs and also facilitate the comparison of genome topology between ESCs/iPSCs 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 (Dostie et al., 2006). 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 (Dekker et al., 2013). We used a tiled/alternating primer design around *Nanog*, *Sox2*, *Klf4*, *Oct4*, *Nestin*, and *Olig1–Olig2* (described in detail in Phillips-Cremins et al., 2013). Our 5C primer design scheme enabled the creation of ~4–12 kb resolution architecture maps for all loci combined across three cellular states with fewer than 30 million reads per replicate (Table S1). The power in this approach is that it focuses on elucidating fine-scale architecture changes at the sub-Mb scale within TADs (Figure 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 (described in detail in the Supplemental Experimental Procedures). Briefly, raw data (Figure S1A) 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 (Figure S1B). 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 (Figure S1C; Phillips-Cremins et al., 2013). We then applied a blocked binning/smoothing algorithm to attenuate spatial noise in 5C data (Figure S1D). Our "Relative Contact Frequency" heatmaps revealed striking topological patterns that are dynamic across cellular states and are unique to each genomic region (Figure 1C).

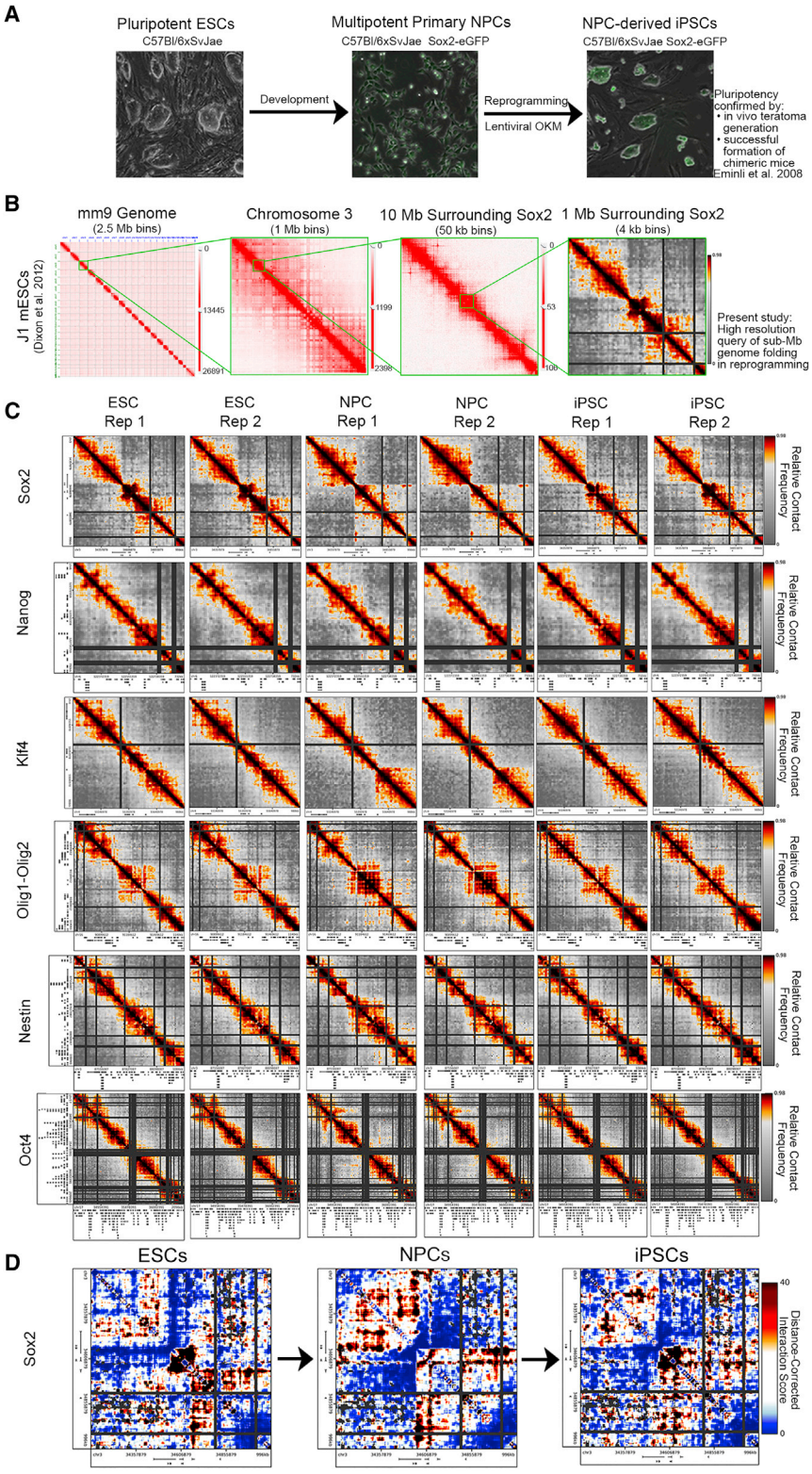


Figure 1. High-Resolution Architecture Maps Reveal Marked Chromatin Reconfiguration during Somatic Cell Reprogramming

(A) Phase contrast images of the reprogramming model system.

(B) Genome-wide ESC Hi-C data (Dixon et al., 2012) 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 (gray) 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 ESCs, NPCs, and iPSCs. Color bars range from low (blue) to high (red/black) interaction scores.

To further resolve the underlying architectural signal, we corrected for the known distance-dependence background in 5C data (Sanyal et al., 2012) (Figures S1E–S1G). Consistent with recent reports (Rao et al., 2014), 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/underestimated by a global background model (Figure S1G). Our "Distance-Corrected Interaction Score" heatmaps showed striking changes in topological features among NPCs, iPSCs, and ESCs (Figure 1D, Figures S1E and S1F), with high consistency between replicates and marked differences among biological conditions (Table S2). A systematic comparative analysis at each stage in the pipeline confirmed that we have reduced known biases in 5C data (Figures S1A–S1I and S2A–S2G).

iPSC 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 (Phillips-Cremins et al., 2013), we observed marked changes in chromatin architecture between ESCs and NPCs. Importantly, we also noticed a striking architectural reconfiguration between NPCs and NPC-derived iPSCs (Figures 1C and 1D). At many loci, iPSC genome folding recapitulated the patterns seen in V6.5 ESCs. However, we also noticed several intriguing cases where iPSC topology retained remnants of the folding patterns from NPCs (Figure 1D).

To further explore the possibility that genome folding might be mis-wired during reprogramming, we conducted principal

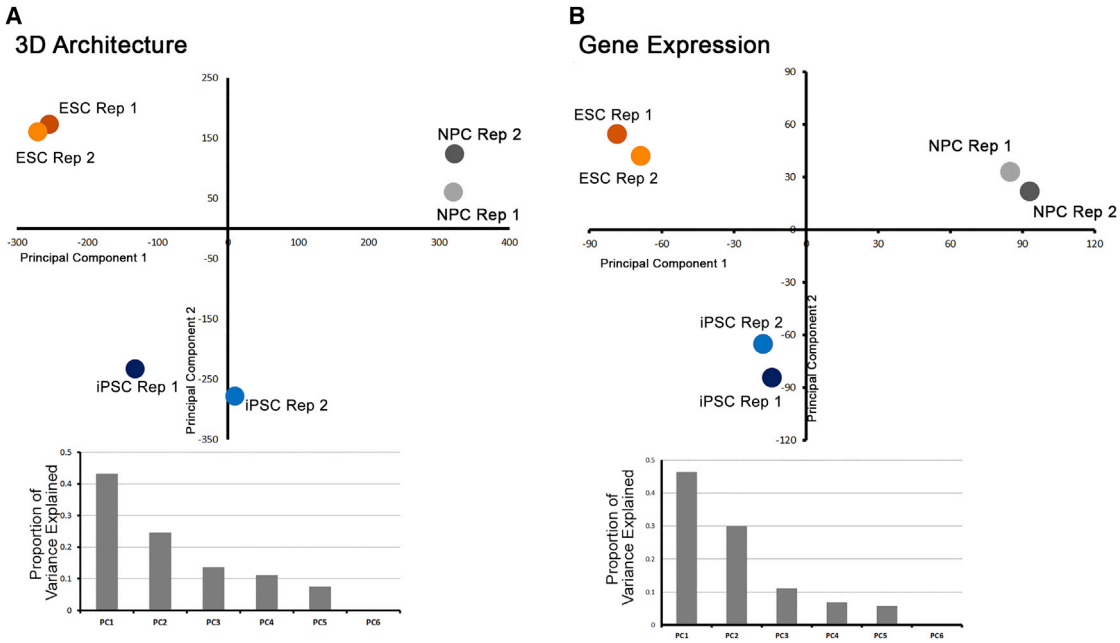


Figure 2. iPSC Genomes Can Exhibit Intermediate Folding and Expression Patterns between Somatic and Pluripotent Stem Cell States

Principal component analysis of (A) Distance-Corrected Interaction Frequency data and (B) normalized RNA-seq data for ESC, NPC, and iPSC replicates. (A and B) Principal components 1 and 2 are scattered and the proportion of variance explained by each principal component is plotted below each scatterplot.

component analysis on our "Distance-Corrected Interaction Frequency" data across all replicates and cellular states. Interestingly, we observed that genome topology in our iPSC clone exhibited folding patterns that were intermediate between NPCs and the pluripotent stem cell state (Figure 2A). To explore the functional significance of potential intermediate iPSC folding patterns, we queried the transcriptome of all three cellular states using RNA-seq. Consistent with our 3D observations, global gene expression profiles in our iPSC clone were also parsed as intermediate between ESCs and NPCs (Figure 2B). Together, these results support the possibility that genome architecture of some iPSC clones might be imperfectly wired within TADs during reprogramming.

Dynamic 3D 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 (Figure S3A, Supplemental Experimental Procedures). We then converted the p value from our fitted models into an interaction score (−10*log2(p values)) that is comparable within and between experiments and allows 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 (Figures 3A–3D). 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 (Figures 3E–3H, Figures S3B and S3C, Supplemental Experimental Procedures). Upon application of our classification scheme, we uncovered several dynamic interaction classes among ESC, NPC, and iPSC cellular states (Figures 3I and 3J), including: (1) 537 interactions present in ESCs, lost in NPCs, and reacquired upon reprogramming (purple class) (Figure 3K); (2) 3,004 interactions present only in ESCs and not reprogrammed (red class) (Figure 3L); (3) 5,043 interactions absent in ESCs, acquired upon differentiation, and lost in iPSCs (green class) (Figure S3D); (4) 1,708 interactions present only in iPSCs (orange class) (Figure S3E); (5) 148 interactions that are high in ESCs and NPCs and not present in iPSCs (gold class) (Figure S3F); and (6) 282 interactions absent in ESCs, acquired in NPCs, and residually connected in iPSCs (blue class) (Figure S3G). Notably, we found that the sensitive detection of these interaction classes, particularly those that distinguish iPSCs from ESCs, was contingent upon the resolution and read depth afforded by the 5C approach (Figures S3H and S3I).

Importantly, we observed that the majority of high-count pixels were spatially adjacent to each other in our Distance-Corrected Interaction Score heatmaps and appear to form larger clusters of enriched 3D contact (Figures 3K–3L and 3N, Figures S3D–S3G). 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) (Figure 3M). Our clustering approach is similar to the methodology employed by Aiden and colleagues for high-resolution Hi-C data (Rao et al., 2014). We emphasize two important points regarding the 3D interaction classes called in this study: (1) the interactions represent both specific looping contacts and

subTAD boundaries that are dynamic across three cellular states and (2) 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.

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 (1) fine-scale chromatin folding, (2) gene expression, (3) histone modifications characteristic of cell-type-specific regulatory elements, and (4) binding profiles of the architectural protein CTCF (Tables S1, S3, and S4).

We first investigated the interactions that were present in ESCs, lost in NPCs, and reconnected during reprogramming (ESC-iPSC; purple class) (Figure 4A). 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 ESCs (Figure 4B). Upon differentiation, the *Sox2*-pluripotent enhancer interaction disassembled in parallel with loss of H3K27ac signal and then subsequently reassembled in iPSCs (Figures 4B and 4C). We also identified ESC-iPSC (purple class) interactions between the *Oct4/Pou5f1* gene and a putative enhancer element ~20 kb upstream marked by ESC-specific H3K4me1/H3K27ac (Figure 4D). As expected given the pluripotent properties of our iPSC clone, the *Oct4*-enhancer interaction breaks apart in NPCs and reconnects again in iPSCs (Figures 4D and 4E). We next quantitatively assessed the enrichment of a wide range of genomic elements in the ESC-iPSC class of successfully reprogrammed 3D interactions. Consistent with previous reports (Apostolou et al., 2013) and our qualitative observations, pluripotency genes and putative ESC-specific enhancers were significantly enriched at the base of ESC-iPSC interactions (Figure 4F). Together, these results indicate that pluripotency genes can form long-range connections with ESC-specific enhancer elements and that these interactions can reprogram in iPSCs.

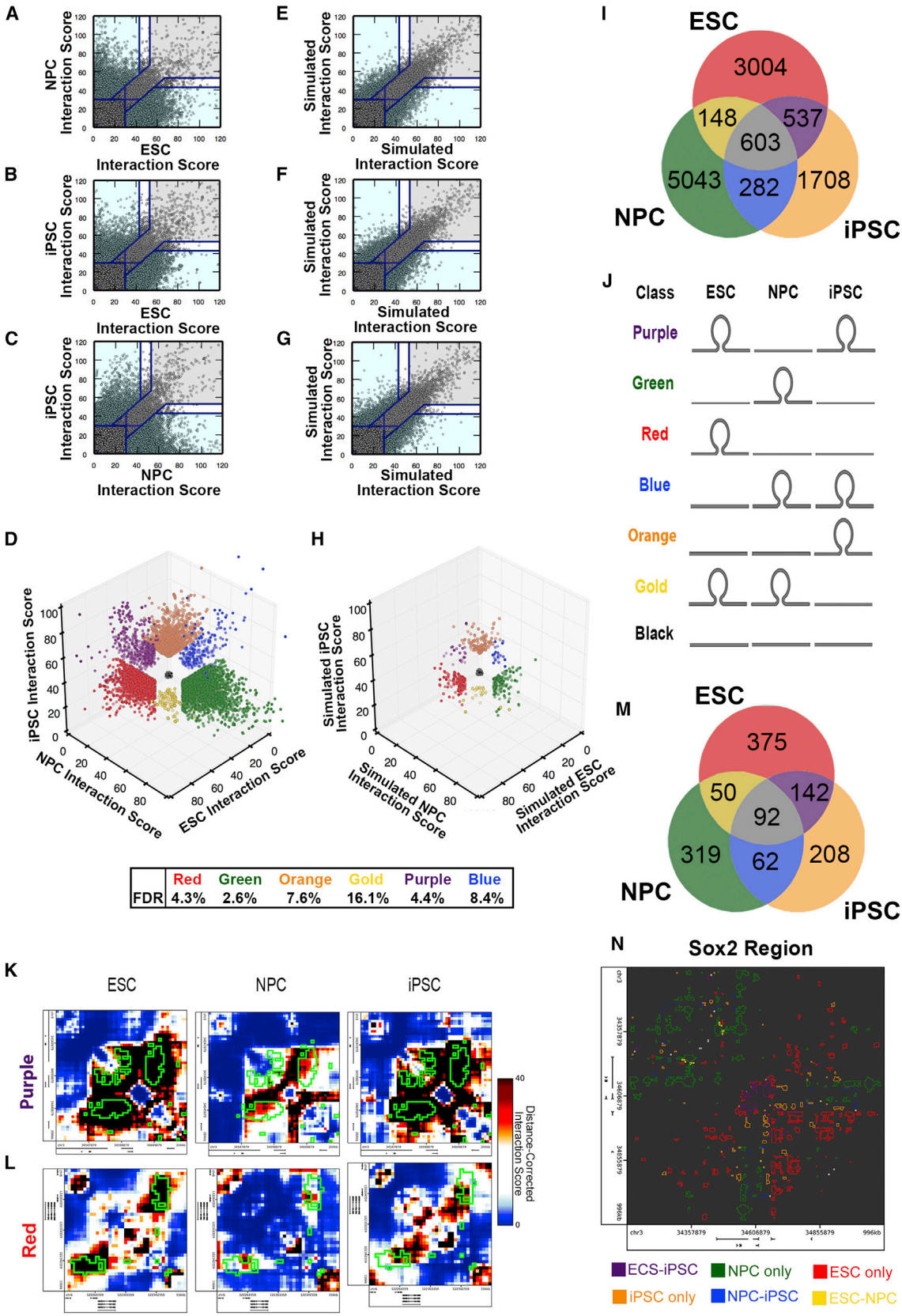
To explore the functional significance of fully reprogrammed interactions, we next conducted genome-wide RNA-seq analysis in ESCs, NPCs, and iPSCs. We examined *Oct4* and *Sox2* gene expression after normalization among libraries to account for any potential batch effects and differences in sequencing depth (Figures S4A–S4D; Table S3, Table S5, and Table S6). Unexpectedly, despite reconnection with target pluripotent enhancers, *Sox2* expression was markedly lower than target ESC expression levels (Figure 4G), whereas *Oct4* expression was more than 2-fold higher than target ESC expression levels (Figure 4H). 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 iPSCs.

Some Pluripotency Genes Reconfigure into New NPC Interactions that Remain Persistent in iPSCs

We next sought to understand larger-scale chromatin folding patterns around *Sox2* (Figure 5A). We hypothesized that chromatin architecture dynamics surrounding the short-range enhancer-promoter interaction might impact the incompletely reprogrammed *Sox2* expression in our iPSC clone. Unexpectedly, we observed that *Sox2* is also engaged in NPC-iPSC (blue class) interactions, which are classified by (1) absence of connection in ESCs, (2) acquisition of connection in NPCs, and (3) residual tethering in iPSCs (Figures 5A and 5B). 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 iPSCs (blue box) (Figure 5A). We observed a similar phenomenon at the *Klf4* locus, where the *Klf4* gene is highly expressed in ESCs and interacts with a putative ESC-specific enhancer element marked by ESC-specific H3K4me1/H3K27ac ~75 kb upstream of the gene (Figures S5A–S5D). In NPCs, *Klf4* disconnects from its pluripotent enhancer and engages with a downstream NPC-specific enhancer (Figures S5E and S5F). In iPSCs, *Klf4* retains its interaction with the NPC-specific enhancer (blue box) while also partially re-tethering to its target pluripotent enhancer (purple box) (Figure S5F).

We hypothesized that the dual tethering of *Sox2/Klf4* genes to their target ESC-specific pluripotent enhancers and their decommissioned NPC-specific enhancers might lead to inaccurate reprogramming of proper expression levels in our iPSC clone. As a first step toward testing this hypothesis, we cultured our iPSC clone under 2i/LIF conditions to promote a naive, ground state of pluripotency and ensure morphological/phenotypic uniformity across the population (Marks et al., 2012; Ying et al., 2008). Strikingly, we noticed that 2i/LIF culture of iPSCs resulted in (1) loss of the *Sox2*- or *Klf4*-NPC enhancer (blue class) interactions, (2) a further amplification in strength of the *Sox2*- or *Klf4*-pluripotent enhancer (purple class) interactions, and (3) a fine-tuning of *Sox2* or *Klf4* expression to ESC levels (Figures 5A and 5C–5D, Figures S5E and S5F). These results indicate that 2i/LIF conditions are capable of untethering persistent somatic cell chromatin architecture in a population of iPSCs and restoring inaccurately reprogrammed gene expression to levels equivalent to those found in V6.5 ESCs. Future causative studies will be necessary to further dissect the link among architectural persistence, naive versus primed pluripotency, and precise gene expression levels during reprogramming.

We then set out to further understand the mechanistic basis of NPC-iPSC (blue class) interactions. Quantitative enrichment analysis revealed three key genomic annotations enriched at the base of NPC-iPSC contacts: (1) ESC-specific genes, (2) NPC-specific CTCF, and (3) constitutive CTCF (Figure 5E). We then computed "sided" enrichments by accounting for the presence/absence of genomic annotations in both anchoring loci at the base of the NPC-iPSC interactions (see schematic, Figure 5F). Consistent with our qualitative observations, ESC-specific genes most significantly contact NPC-specific enhancers when located at the base of NPC-iPSC interactions (Figure 5F).



	Red	Green	Orange	Gold	Purple	Blue
FDR	4.3%	2.6%	7.6%	16.1%	4.4%	8.4%

(legend on next page)

We note that *Sox2* and *Klf4* are classified as ESC-specific genes in our study due to their markedly increased expression in ESCs versus 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 ESCs and NPCs, but regulated by different enhancers in the two cellular states, might be particularly susceptible to inappropriate tethering to off-lineage enhancers in iPSCs.

Our quantitative enrichment analyses also indicated that ESC-specific genes formed significant 3D connections with NPC-specific and constitutively bound CTCF sites (Figures 5E and 5F). Consistent with this quantitative result, we noticed a constitutively bound CTCF site at the base of the *Sox2* NPC-specific enhancer (Figure 5A) and an NPC-specific CTCF site at the base of the *Klf4* NPC-specific enhancer (Figure S5F), suggesting that CTCF might work together with enhancers to facilitate 3D 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 ESC-specific enhancers (Figure 5A, blue and red stars, respectively) at the *Sox2* locus. We found that the NPC-specific enhancer remains constitutively bound by CTCF in ESC, NPC, iPCS, ESC+2i, and iPSC+2i conditions (Figure 5G, left). By contrast, the ESC-specific enhancer exhibited high CTCF in ESCs, loss of binding in NPCs, sustained low CTCF occupancy in iPSCs, and subsequent restoration of occupancy in 2i/LIF (Figure 5G, right).

Intriguingly, CTCF binding patterns correlate with the changes in chromatin architecture around *Sox2*. In ESCs, the constitutive CTCF site interacts with the ESC-specific CTCF site, resulting in spatial co-localization of the ESC- and NPC-specific enhancers (Figure 5A, red box). Loss of CTCF binding at the ESC-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-iPSC (blue class) interaction with the *Sox2* gene (Figure 5A, blue box). We posit that the *Sox2*-NPC-enhancer interaction remains tethered in iPSCs because CTCF does not fully rebind to the ESC-specific enhancer (Figure 5G, right). In support of this idea, 2i/LIF leads to (1) reacquisition of CTCF binding at the ESC-specific enhancer, (2) reconnection of the interaction between both ESC-specific and NPC-specific enhancers, and (3) abrogation of the *Sox2*-NPC-specific enhancer interaction. These observations are consistent with a working model in which "persistent-NPC" interactions can remain in iPSCs 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-iPSC interactions (Figures S6A–S6C). For example, NPC-specific genes and enhancers were primarily enriched in NPC (green class) interactions only, supporting our finding that it is ESC-specific genes, particularly those that remain somewhat active in NPCs, that are redirected into NPC-iPSC contacts. An example illustrating this idea can be found at the *Olig1* and *Olig2* genes that are expressed in an NPC-specific manner and equivalently form NPC (green class) interactions only with a downstream NPC-specific enhancer (Figures S6D and S6E). Expression of *Olig1* and *Olig2* is lost in parallel with loss of the green class 3D interaction. Together, these results support the intriguing possibility that ESC-specific genes that remain partially active in NPCs form new interactions with somatic cell-specific enhancers during differentiation and that these contacts can remain tethered as a form of architectural persistence in iPSCs. Finally, we note that 5C is performed on a population of millions of cells, we cannot distinguish between the possibilities that (1) pluripotency genes simultaneously form both ESC-iPSC and NPC-iPSC contacts in individual cells and (2) pluripotency genes form two different sets of interactions in distinct ESC-like subpopulations.

Pluripotent Interactions that Do Not Reprogram Display Dynamic CTCF Occupancy

Finally, we explored the interactions that are present in ESCs and lost in NPCs but do not reconnect in iPSCs (red group, Figures 6A–6B, Figures S7A and S7B). A noteworthy illustration of these poorly reprogrammed interactions is found at the *Zfp462* gene (highlighted in green, Figure 6A), which interacts with a downstream putative ESC-specific enhancer element in ESCs. *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 ESC-iPSC (purple) group, this gene-enhancer interaction is not reassembled in iPSCs. Similarly, the genes *Mis18a* and *Urb1* form interactions in ESCs that are not reprogrammed (highlighted in yellow and green, respectively; Figure S7A). Together, these

Figure 3. 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) ESCs and NPCs, (B) ESCs and iPSCs, and (C) NPCs and iPSCs. Thresholds are displayed as blue lines. For pairwise plots, cell-type-specific, invariant, and background interactions are represented by blue, gray, 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, ESC only; green, NPC only; orange, iPSC only; gold, ESC-NPC; purple, ESC-iPSC; blue, NPC-iPSC; 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) and (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 and L) Zoomed-in heatmaps of distance-corrected interaction scores for specific (K) ESC-iPSC (purple class) and (L) ESC only (red class) interactions. Classified interaction pixels are outlined in green.

(M) Number of interactions called significant for each 3D classification after clustering directly adjacent 4 kb bins.

(N) Depiction of all interactions called significant in the Sox2 region. Each interaction is outlined by the corresponding classification color.

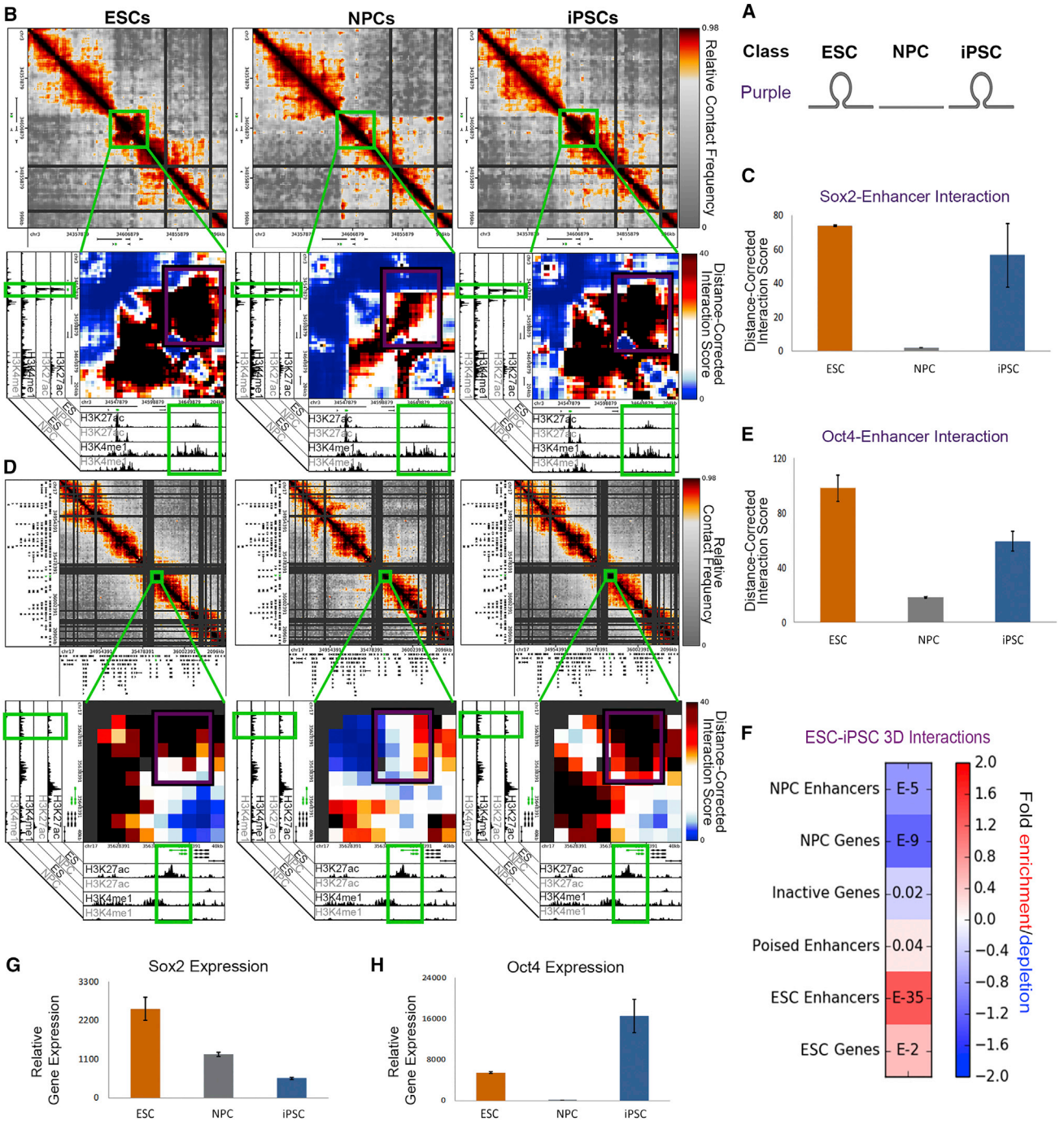


Figure 4. Pluripotency Gene-Enhancer Interactions Can Be Re-established in iPSCs

(A) Schematic illustrating the ESC-iPSC (purple) interaction class.

(B and D) Relative contact frequency heatmaps (top) and zoomed-in distance-corrected interaction score heatmaps (bottom) highlighting key ESC-iPSC interactions (purple class) between (B) *Sox2* and (D) *Oct4* genes and their target enhancers. Heatmaps are overlaid on ChIP-seq tracks of H3K27ac and H3K4me1 in ESCs and NPCs.

(C and E) Distance-corrected interaction score changes at (C) the *Sox2*-enhancer interaction and (E) *Oct4*-enhancer interaction among ESCs, NPCs, and iPSCs. Error bars represent the standard deviation across two 5C replicates.

(F) Fold enrichment of cell-type-specific regulatory elements in ESC-iPSC (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 and H) Normalized gene expression is plotted for (G) *Sox2* and (H) *Oct4* genes. Error bars represent standard deviation across two RNA-seq replicates.

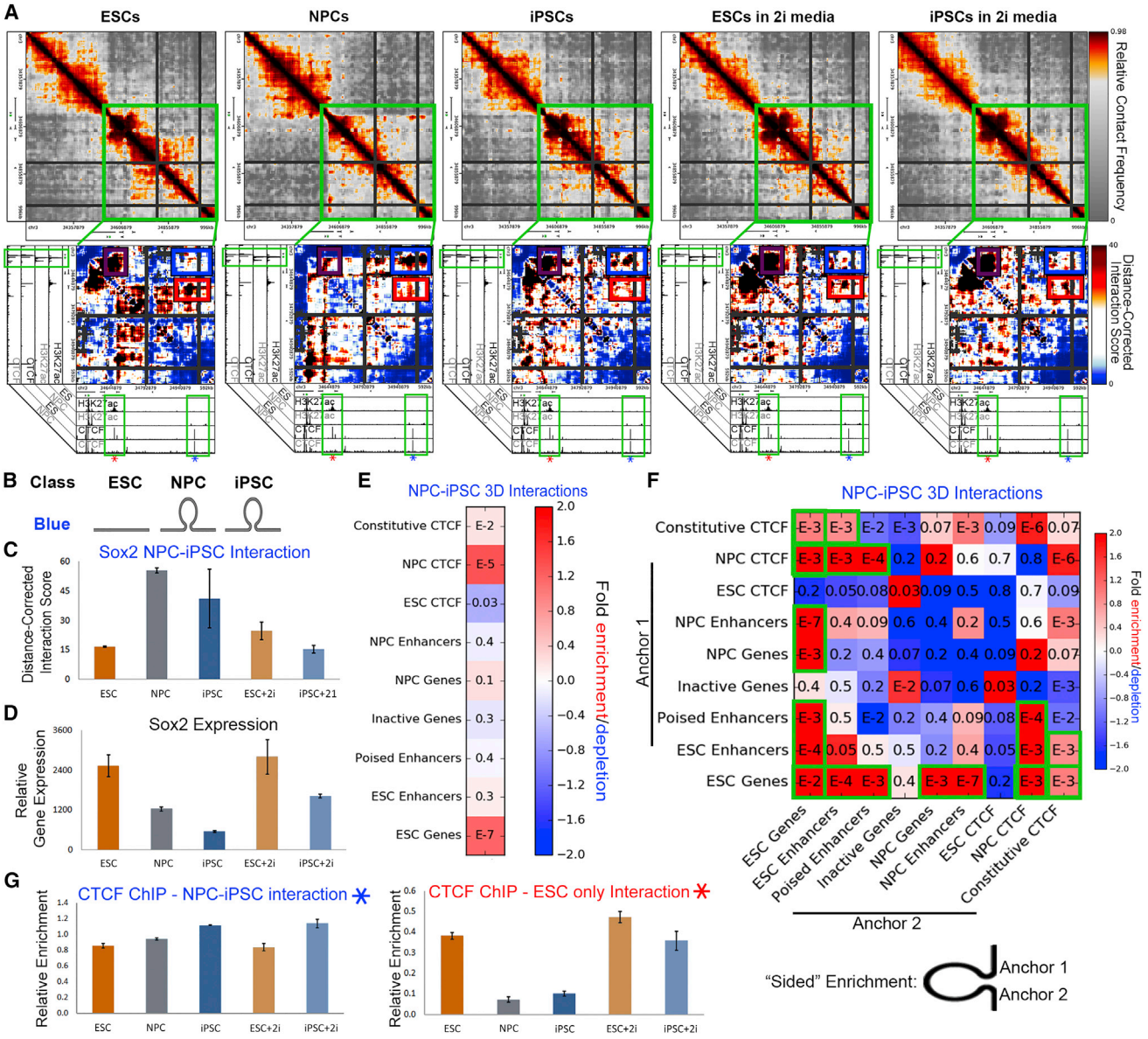


Figure 5. Pluripotency Genes Can Exhibit "Persistent NPC-like" Folding Patterns in iPSCs

(A) Relative contact frequency heatmaps (top) and zoomed-in distance-corrected interaction score heatmaps (bottom) highlighting an NPC-iPSC interaction (blue class) around the *Sox2* gene. Heatmaps are overlaid on ChIP-seq tracks of H3K27ac and CTCF in ESCs and NPCs.

(B) Schematic illustrating the NPC-iPSC (blue) interaction class.

(C) Distance-corrected interaction score changes at an NPC-iPSC interaction around the *Sox2* gene among ESC, NPC, iPSC, ESC+2i, and iPSC+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 RNA-seq replicates.

(E and F) Fold enrichment of cell type-specific regulatory elements in NPC-iPSC (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-iPSC interactions. (F) Enrichment for any given pairwise combination of genomic annotations in the two anchoring bins at the base of NPC-iPSC interactions.

(G) Relative ChIP-qPCR enrichment of CTCF binding at the NPC-iPSC interaction (left, denoted by blue star in A) and ESC only interaction (right, denoted by red star in A). Error bars represent SD across three technical replicates.

genomic loci reveal a class of interactions that are refractory to reprogramming in iPSCs.

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 ESCs, but severely depleted in NPCs, were preferentially located at the base of poorly reprogrammed interactions (green boxes; Figures 6A and S7A). Consistent with this observation, ESC-specific CTCF sites were significantly enriched in ESC only (red class) interactions (Figures 6C and 6D). ChIP-qPCR

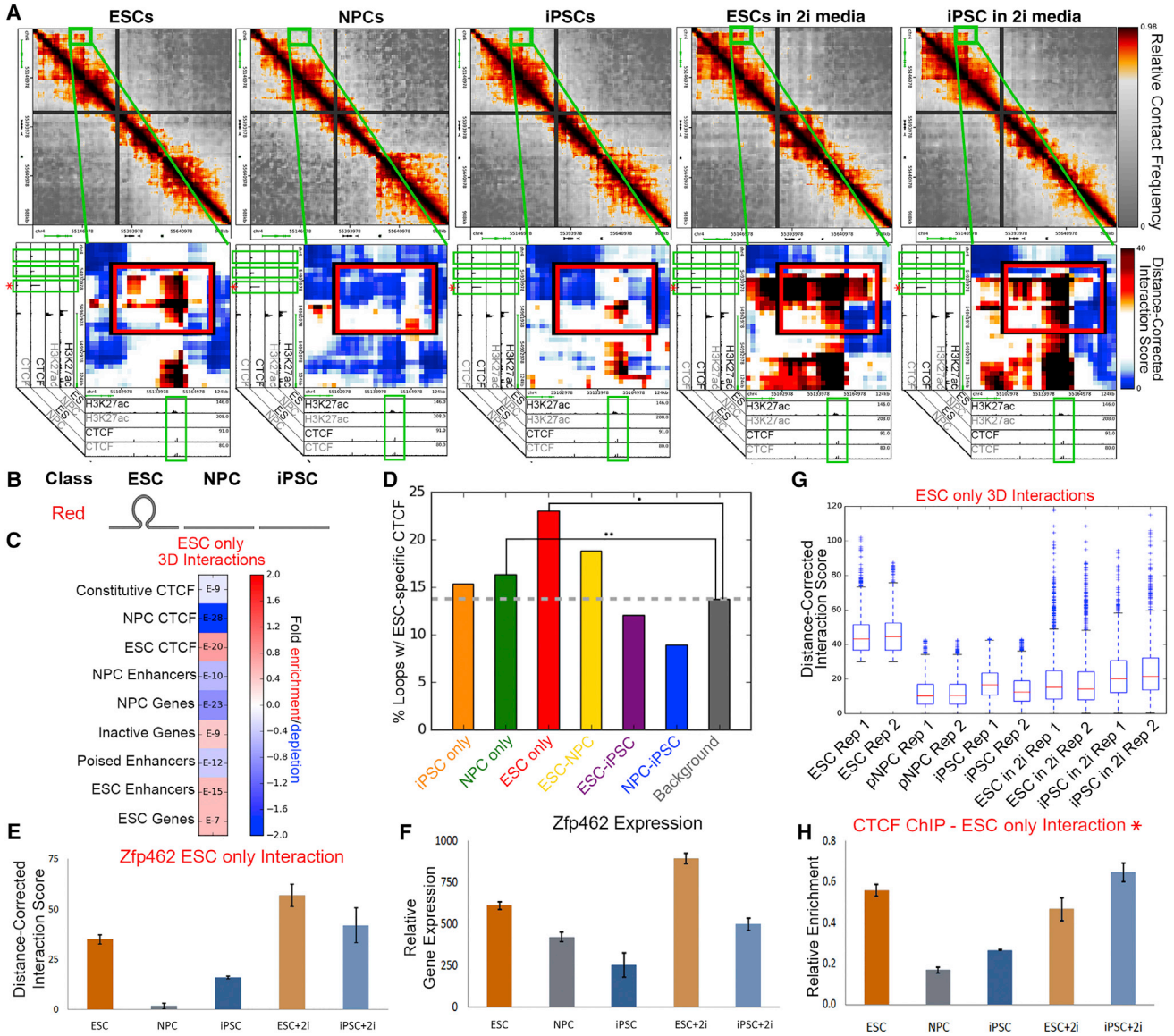


Figure 6. Interactions that Do Not Reprogram Display Poorly Reprogrammed CTCF Occupancy

(A) Relative contact frequency heatmaps (top) and zoomed-in distance-corrected interaction score heatmaps (bottom) highlighting an ESC only (red class) interaction at ESC-specific CTCF binding sites at the *Zfp462* gene (indicated in green). Heatmaps are overlaid on ChIP-seq tracks of H3K27ac and CTCF in ESCs and NPCs.

(B) Schematic illustrating the ESC only (red class) interactions.

(C) Fraction of ESC 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 ESC-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 ESC only interaction around the *Zfp462* gene among ESC, NPC, iPSC, ESC+2i, and iPSC+2i conditions. Error bars represent standard deviation across two 5C replicates.

(F) *Zfp462* gene expression among ESC, NPC, iPSC, ESC+2i, and iPSC+2i conditions. Error bars represent standard deviation across two RNA-seq replicates.

(G) Aggregate distance-corrected interaction score changes among ESC, NPC, iPSC, ESC+2i, and iPSC+2i conditions for loci anchoring red class.

(H) Relative ChIP-qPCR enrichment of CTCF binding at the ESC only interaction (denoted by blue star in A). Error bars represent SD across three technical replicates.

analysis of CTCF occupancy revealed consistent depletion of CTCF in our iPSC clone compared to ESCs (Figures 5G and 6H, Figure S7G). Importantly, culture of our iPSC clone in 2i/LIF media resulted in (1) reacquisition of the red group interactions, (2) re-establishment of CTCF occupancy, and (3) restoration of gene expression levels in iPSCs (Figures 6E–6H, Figures S7C–S7G). Corroborating locus-specific observations, a global analysis of red class interactions demonstrated a marked

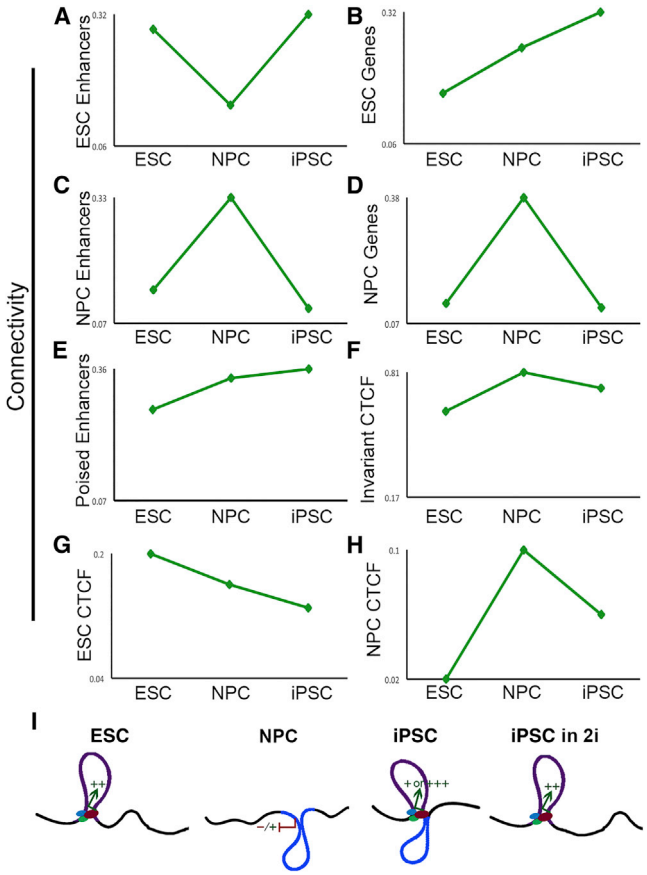


Figure 7. Pluripotency Genes Can Be Hyperconnected in iPSCs

Connectivity of distinct regulatory elements in ESCs, NPCs, and NPC-derived iPSCs. (A) ESC-specific enhancers; (B) ESC-specific genes; (C) NPC-specific enhancers; (D) NPC-specific genes; (E) poised enhancers; (F) invariant CTCF; (G) ESC-specific CTCF; (H) NPC-specific CTCF. (I) Schematic illustrating a model of the "hyperconnectivity" of certain pluripotency genes in our NPC-derived iPSC clone. Key ESC-specific genes (denoted by colored arrows) display the ability to reprogram their connections with ESC-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.

increase in interaction score upon the addition of 2i/LIF media to iPSCs (Figure 6G). 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.

Somatic Elements Are Disconnected and Pluripotent Genes Are Hyperconnected in our iPSC Clone

We hypothesized that distinct types of regulatory elements exhibit differential connectivity patterns as ESCs transition to NPCs and back to iPSCs. To address this hypothesis, we computed a "connectivity" metric for each class of genomic element in each of the three cellular states. ESC-specific enhancers lose their connectivity in NPCs and then reconnect in iPSCs (Figure 7A). Intriguingly, ESC-specific genes become increasingly more connected upon differentiation and subsequent reprogramming (Figure 7B). By contrast, NPC-specific genes/enhancers increase connectivity in NPCs, but then resume ground state ESC-like connectivity in iPSCs (Figures 7C and 7D). Poised enhancers and invariant CTCF sites display minor differences in connectivity across the three cellular states (Figures 7E and 7F), whereas ESC-specific CTCF sites lose their interactions upon differentiation and only partially gain back connectivity in iPSCs (Figure 7G). NPC-specific CTCF sites increase in connectivity in NPCs and then partially resume their disconnected state in iPSCs (Figure 7H).

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 iPSC clone due to persistent cell-of-origin interactions (Figure 7I). 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 ESC-like expression levels in our iPSC clone.

DISCUSSION

Understanding the molecular mechanisms governing somatic cell reprogramming is of paramount importance to our knowledge of cell fate commitment and the use of iPSCs for regenerative medicine applications. Mechanistic studies have primarily focused on profiling gene expression and classic epigenetic modifications at intermediate stages in the reprogramming process (Koche et al., 2011; Polo et al., 2012; Soufi et al., 2012; Stadtfeld et al., 2008). However, the molecular roadblocks that impede the efficiency and timing of epigenome resetting in iPSCs 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 individual loci (e.g., *Nanog* or *Oct4*) reported that pluripotency genes can re-establish long-range connections with their target enhancers in iPSCs (Apostolou et al., 2013; de Wit et al., 2013; Denholtz et al., 2013; Wei et al., 2013; Zhang et al., 2013a). Motivated by the need to understand how chromatin unfolds/refolds more generally in iPSCs, 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 ESC-specific enhancers in ESCs; these interactions break apart in NPCs and then reassemble in iPSCs. Additionally, we find that somatic cell interactions between NPC-specific genes and NPC-specific enhancers generally disconnect in iPSCs. 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-iPSC (blue class) interactions representing persistent chromatin folding patterns from the somatic cell of origin in

iPSCs. Unexpectedly, we find that some key pluripotency genes can form new 3D connections in NPCs that remain tethered in our iPSC clone. For example, *Klf4* and *Sox2* are dually tethered to their target ESC-specific enhancers and their decommissioned NPC-specific enhancers in iPSCs. 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 iPSC 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 ESC line. Notably, we highlight that NPC-specific genes/enhancers form contacts in NPCs that subsequently disassemble in iPSCs, suggesting that somatic genes are not driving the architectural persistence in iPSCs. These results agree with previous studies suggesting that somatic cell gene expression is downregulated during the initiation phase of reprogramming and precedes the reactivation of the pluripotency network (Polo et al., 2012). We favor a model in which reconfiguration of higher-order chromatin topology could be a potential rate-limiting step in the reprogramming process as a result of architectural persistence or incomplete architectural reprogramming (discussed below) blocking the formation of fully reprogrammed iPSCs (Buganim et al., 2012; Tanabe et al., 2013).

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 ESCs (Dixon et al., 2012; Rao et al., 2014; Handoko et al., 2011; Phillips-Cremins et al., 2013). Here we provide a new link between CTCF and reprogramming. We identify a new class of chromatin interactions that are high in ESCs, break apart in NPCs, and are not fully reconfigured in iPSCs. Importantly, we find that these poorly reprogrammed interactions often contain ESC-specific CTCF binding sites that lose occupancy in NPCs and do not reacquire full binding in our iPSC clone. CTCF has largely stable occupancy patterns during development, with 60%–90% of sites remaining bound to the genome between cell types (Kim et al., 2007). 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 (Bell and Felsenfeld, 2000), suggesting a possible link between poorly reprogrammed chromatin contacts and previously reported sources of cell-of-origin epigenetic persistence (Kim et al., 2010; Polo et al., 2010). Indeed, because ESCs cultured in 2i/LIF display global hypomethylation (Ficz et al., 2013; Habibi et al., 2013), 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 iPSCs cultured under optimal conditions (Cahan et al., 2014; Stadtfeld et al., 2010). However, variations in epigenetic profiles among iPSC clones have been attributed to reprogramming method, passage number, genetic background, or laboratory-to-laboratory procedural discrepancies (Bock et al., 2011; Polo et al., 2010). Therefore, we sought to confirm that our observations were truly linked to inefficiencies in the reprogramming of our iPSC clone, and not experimental artifacts due to (1) residual somatic cells in our iPSC population or (2) laboratory-specific culture conditions. Importantly, Hochedlinger and colleagues have extensively characterized the iPSC clone used in this manuscript for its pluripotent properties (Eminli et al., 2008). Additionally, our iPSC 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 iPSC population versus ESCs (Figures S6E–S6G). Thus, we see no evidence of contaminating NPCs in our iPSCs. 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 naive and primed pluripotency) across single cells within our fully reprogrammed iPSC 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, naive pluripotent state with 2i/LIF media abrogates architectural persistence interactions and reinstates poorly reprogrammed interactions. Additionally, although we subjected our iPSCs 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. Noteworthily, these results raise the interesting possibility that an iPSC 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 iPSCs, and a range of iPSC clones derived from multiple somatic cell lineages.

In parallel with this manuscript, de Laat, Graf, and colleagues published a genome-wide analysis of chromatin architecture in iPSCs derived from four independent somatic cell lineages (Krijger et al., 2016). 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, and 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 (Dixon et al., 2012), TAD boundaries remained for the most part consistent among iPSC clones and ESCs. 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 subTAD boundaries and looping interactions than reported in Krijger et al. during the transition among ESC, iPSC, and NPC cellular states. It is noteworthy that when we increase our bin size from 4 kb up to 300 kb (Figure S3H), we can recapitulate the author's high level of correlation between the ESCs and iPSCs (Figure S3I). Krijger et al. and our manuscript 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 iPSCs 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 3D 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 iPSCs 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 catalog of "hotspots" of incomplete architectural reprogramming to address whether specific somatic cell types are more or less resistant to topological changes.

EXPERIMENTAL PROCEDURES

Cell Culture, Differentiation, and Reprogramming

V6.5 ESCs, primary NPCs, and NPC-derived iPSCs were cultured as described in the Supplemental Experimental Procedures. Briefly, ESCs were expanded on Mitomycin-C inactivated MEFs under standard pluripotent conditions and passaged onto feeder-free gelatin-coated plates before fixation. Primary NPCs were isolated from whole brains of P1 129SvJae × C57BL/6, Sox2-eGFP mice and cultured as neurospheres for two passages before adherent culture and fixation. The iPSC clone used in this paper was derived and characterized in Eminli et al. (2008) and expanded/cultured for use in this study to >15 passages with or without 2i/LIF media as described in the Supplemental Experimental Procedures.

Generation and Analysis of 5C Libraries

5C libraries were generated according to standard procedures described in the Supplemental Experimental Procedures. Paired-end reads were aligned to a pseudo-genome consisting of all 5C primers using Bowtie. Interactions were counted when both paired-end reads were uniquely mapped to the 5C primer pseudo-genome. Counts were converted to contact matrices for each genomic region queried, processed, normalized, and modeled as described in the Supplemental Experimental Procedures. Customized algorithms for classification of 5C interactions and the downstream integration of interaction classes with ChIP-seq and RNA-seq data are detailed in the Supplemental Experimental Procedures.

RNA-Seq Library Preparation and Analysis

Cells were lysed with Trizol and total RNA was extracted as detailed in the Supplemental Experimental Procedures. Samples were prepared for sequencing using the Illumina TruSeq Stranded Total RNA Library Prep kit with RiboZero (Illumina RS-122-2202) following the supplier's protocol and sequenced on the Illumina NextSeq500. Libraries were analyzed and corrected for any sequencing depth or batch effect differences with methods described in the Supplemental Experimental Procedures.

ChIP-Seq Analysis and ChIP-qPCR

A summary of published ChIP-seq libraries re-analyzed in this study is provided in Table S4. Reads were aligned to mm9 with Bowtie using default parameters. Only uniquely mapped reads were used for downstream analyses. ChIP-seq peak calling and ChIP-qPCR experiments are detailed in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

The accession number for the data reported in this paper is GEO: GSE68582.

SUPPLEMENTAL INFORMATION

Supplemental Information for this article includes seven figures, seven tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.stem.2016.04.004.

AUTHOR CONTRIBUTIONS

T.G.G., J.K., and Z.P. contributed equally to this work. J.E.P.C, V.G.C., and J.D. conceived the project. J.A.B, G.H., and J.E.P.C designed and performed 3C/5C experiments. K.H. and E.A. provided iPSCs and Sox2-GFP neonates. H.K.N. performed RNA-seq experiments. J.A.B., S.C.H., and X.L. performed ChIP experiments. T.G.G., J.K., Z.P., E.J.S., J.A.B., and J.E.P.C. developed code and analyzed data. J.A.B. and J.E.P.C wrote the paper with input from E.A., K.H., V.G.C., and J.D.

ACKNOWLEDGMENTS

We thank members of the Cremins laboratory for helpful discussions and Michael Duong for cell culture assistance. J.E.P.C. is a New York Stem Cell Foundation Robertson Investigator and an Alfred P. Sloan Foundation Fellow. This work was funded by The New York Stem Cell Foundation (J.E.P.C), the Alfred P. Sloan Foundation (J.E.P.C), the NIH Director's New Innovator Award from the National Institute of Mental Health (1DP2MH11024701; J.E.P.C), 4D Nucleome Common Fund grants (1U01HL12999801 to J.E.P.C; U54DK107980 to J.D.), and an R01 from the National Human Genome Research Institute (R01 HG003143; J.D.). J.D. is an investigator of the Howard Hughes Medical Institute.

Received: June 3, 2015
Revised: December 31, 2015
Accepted: April 15, 2016
Published: May 5, 2016

REFERENCES

Apostolou, E., Ferrari, F., Walsh, R.M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., Stuart, H.T., Polo, J.M., Ohsumi, T.K., Borowsky, M.L., et al. (2013). Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. Cell Stem Cell *12*, 699–712.

Bell, A.C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482–485.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H., et al. (2011). Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. Cell *144*, 439–452.

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell *150*, 1209–1222.

Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. Cell *158*, 903–915.

de Wit, E., Bouwman, B.A., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J., Krijger, P.H., Festuccia, N., Nora, E.P., Welling, M., et al. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature *501*, 227–231.

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. *14*, 390–403.

Denholtz, M., Bonora, G., Chronis, C., Splinter, E., de Laat, W., Ernst, J., Pellegrini, M., and Plath, K. (2013). 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.

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

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

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

Eminli, S., Utikal, J., Arnold, K., Jaenisch, R., and Hochedlinger, K. (2008). Reprogramming of neural progenitor cells into induced pluripotent stem cells in the absence of exogenous Sox2 expression. Stem Cells *26*, 2467–2474.

Ficz, G., Hore, T.A., Santos, F., Lee, H.J., Dean, W., Arand, J., Krueger, F., Oxley, D., Paul, Y.L., Walter, J., et al. (2013). FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. Cell Stem Cell *13*, 351–359.

Habibi, E., Brinkman, A.B., Arand, J., Kroeze, L.I., Kerstens, H.H., Matarese, F., Lepikhov, K., Gut, M., Brun-Heath, I., Hubner, N.C., et al. (2013). Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. Cell Stem Cell *13*, 360–369.

Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. Nat. Genet. *43*, 630–638.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature *462*, 595–601.

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

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H., Ehrlich, L.I., et al. (2010). Epigenetic memory in induced pluripotent stem cells. Nature *467*, 285–290.

Koche, R.P., Smith, Z.D., Adli, M., Gu, H., Ku, M., Gnirke, A., Bernstein, B.E., and Meissner, A. (2011). Reprogramming factor expression initiates widespread targeted chromatin remodeling. Cell Stem Cell *8*, 96–105.

Krijger, P.H., Di Stefano, B., de Wit, E., Limone, F., van Oevelen, C., de Laat, W., and Graf, T. (2016). Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming. Cell Stem Cell *18*, this issue, 597–610.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lujan, E., Zunder, E.R., Ng, Y.H., Goronzy, I.N., Nolan, G.P., and Wernig, M. (2015). Early reprogramming regulators identified by prospective isolation and mass cytometry. Nature *521*, 352–356.

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Stewart, A.F., Smith, A., and Stunnenberg, H.G. (2012). The transcriptional and epigenomic foundations of ground state pluripotency. Cell *149*, 590–604.

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

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., et al. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol. Cell *38*, 603–613.

Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Polo, J.M., Liu, S., Figueroa, M.E., Kulalert, W., Eminli, S., Tan, K.Y., Apostolou, E., Stadtfeld, M., Li, Y., Shioda, T., et al. (2010). Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. Nat. Biotechnol. *28*, 848–855.

Polo, J.M., Anderssen, E., Walsh, R.M., Schwarz, B.A., Nefzger, C.M., Lim, S.M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J., et al. (2012). A molecular roadmap of reprogramming somatic cells into iPS cells. Cell *151*, 1617–1632.

Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S., Canfield, T.K., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402–405.

Rais, Y., Zviran, A., Geula, S., Gafni, O., Chomsky, E., Viukov, S., Mansour, A.A., Caspi, I., Krupalnik, V., Zerbib, M., et al. (2013). Deterministic direct reprogramming of somatic cells to pluripotency. Nature *502*, 65–70.

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

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Soufi, A., Donahue, G., and Zaret, K.S. (2012). Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell *151*, 994–1004.

Stadtfeld, M., Maherali, N., Breault, D.T., and Hochedlinger, K. (2008). Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. Cell Stem Cell *2*, 230–240.

Stadtfeld, M., Apostolou, E., Akutsu, H., Fukuda, A., Follett, P., Natesan, S., Kono, T., Shioda, T., and Hochedlinger, K. (2010). Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. Nature *465*, 175–181.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Tanabe, K., Nakamura, M., Narita, M., Takahashi, K., and Yamanaka, S. (2013). Maturation, not initiation, is the major roadblock during reprogramming toward pluripotency from human fibroblasts. Proc. Natl. Acad. Sci. USA *110*, 12172–12179.

Wei, Z., Gao, F., Kim, S., Yang, H., Lyu, J., An, W., Wang, K., and Lu, W. (2013). Klf4 organizes long-range chromosomal interactions with the oct4 locus in reprogramming and pluripotency. Cell Stem Cell *13*, 36–47.

Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Zhang, H., Jiao, W., Sun, L., Fan, J., Chen, M., Wang, H., Xu, X., Shen, A., Li, T., Niu, B., et al. (2013a). Intrachromosomal looping is required for activation of endogenous pluripotency genes during reprogramming. Cell Stem Cell *13*, 30–35.

Zhang, Y., Wong, C.H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y., Lim, J., Tai, E., Poh, H.M., Wong, E., et al. (2013b). Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature *504*, 306–310.

Cell Stem Cell, Volume *18*

Supplemental Information

Local Genome Topology Can Exhibit an Incompletely Rewired 3D-Folding State during Somatic Cell Reprogramming

Jonathan A. Beagan, Thomas G. Gilgenast, Jesi Kim, Zachary Plona, Heidi K. Norton, Gui Hu, Sarah C. Hsu, Emily J. Shields, Xiaowen Lyu, Effie Apostolou, Konrad Hochedlinger, Victor G. Corces, Job Dekker, and Jennifer E. Phillips-Cremins

Supplemental Materials (Beagan et al.)

Supplemental Figures

Supplemental Figure Captions

Supplemental Tables

Table S3: Summary of paired-end read alignments for RNA-seq libraries, related to Experimental Procedures and Supplemental Experimental Procedures.

Table S4: Summary of external ChIP-seq libraries analyzed in this study, related to Experimental Procedures and Supplemental Experimental Procedures.

Table S5: Genes classified as ES-specific, Related to Figures 4, 5, 6 and Supplemental Experimental Procedures.

Table S6: Genes classified as NPC-specific, Related to Figures 4, 5, 6 and Supplemental Experimental Procedures.

Table S7. Interactions selected for representative interaction score barplots, Related to Figures 4, S5, 5, 6, S7.

Supplemental Experimental Procedures

ES cell culture

Primary Neural Progenitor Cell isolation

iPS cell culture

Culture of pluripotent cells in 2i media

3C template generation and characterization

5C primer design

5C library generation and sequencing

iPS cell transgene integration detection by 5C primers

RNA-seq library preparation

RNA-seq data processing

CTCF binding detection by ChIP-qPCR

5C data processing pipeline

Paired-end read mapping and counting

Low count primer removal

Raw contact matrix visualization

Quantile normalization

Primer correction

Low count fragment-fragment pair removal

Contact matrix binning

Pseudo-fragment level 5C mapping resolution

Identification of bad primer gaps

Distance-dependence normalization

Probabilistic model fitting and distance-corrected interaction scores

GC content bias investigation

Comparison of 5C analysis pipeline to alternative approaches

Principal component analysis

Classification of cell type-specific 3-D interactions

Empirical false discovery rate calculation

Justification of strategy

Model generation – mean parameter estimation

Model generation – estimating the mean-variance relationship

Model generation – variance parameter estimation

Simulations

Monte Carlo, p-value calculation, classification

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

6 Sample vs 10 Sample 5C data processing

Interaction adjacency clustering

ChIP-seq peakcalling

Parsing ES-specific and NPC-specific genes

Parsing ES-specific and NPC-specific enhancers

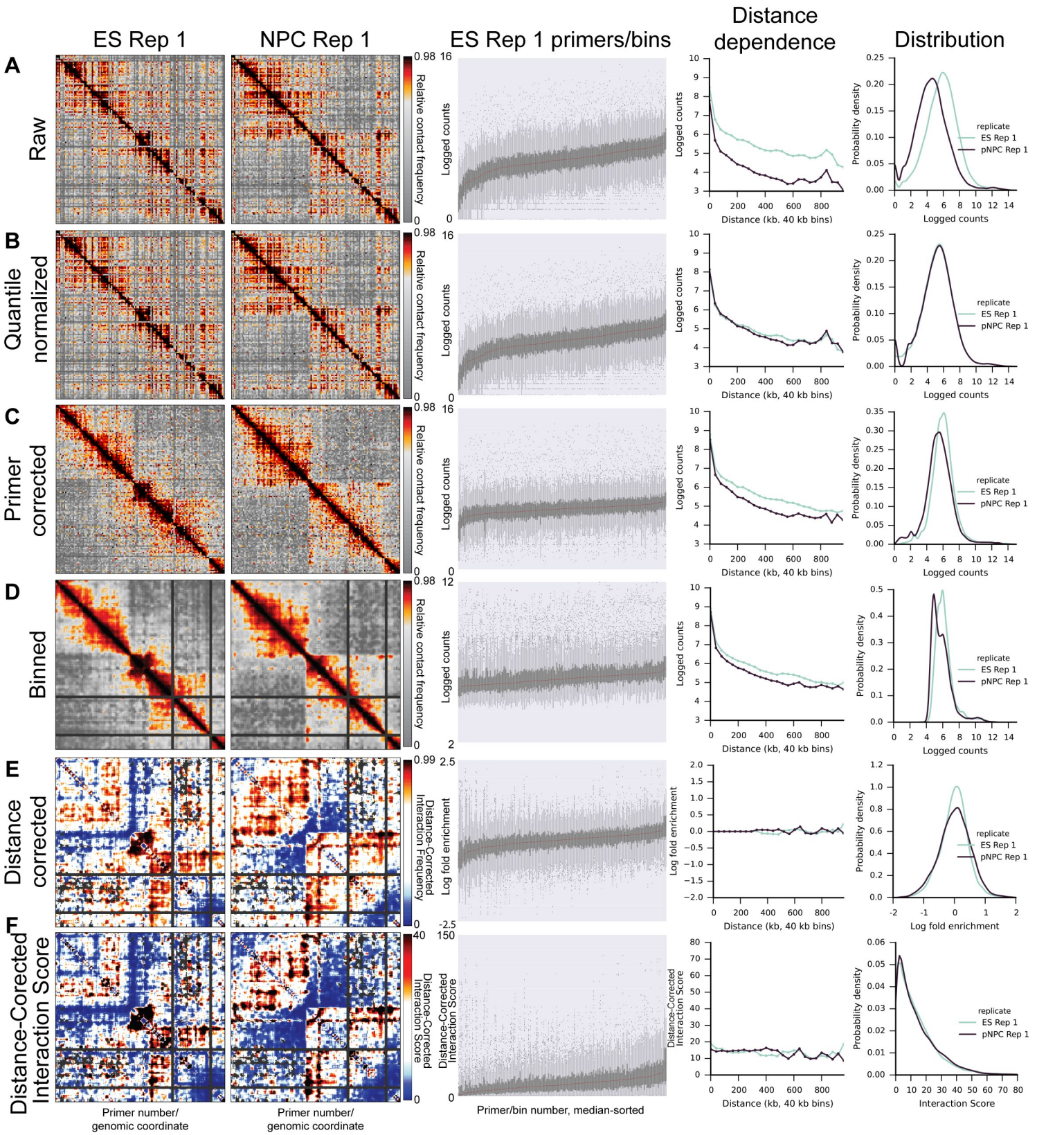
Parsing ES-specific and NPC-specific CTCF sites

Computing enrichments

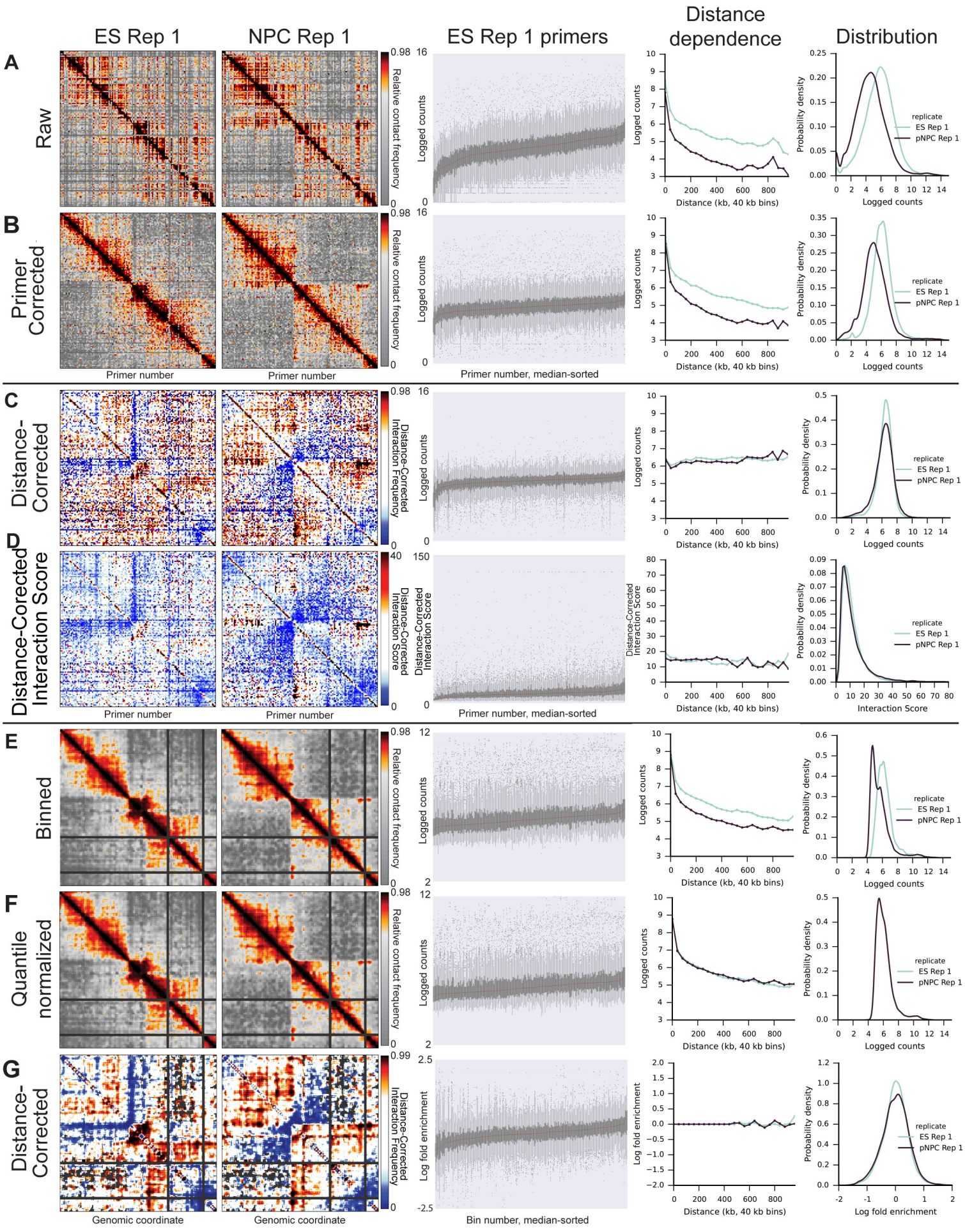
- Annotation intersections
- Computing percentage incidence, fold-enrichment above background, and p-values
- Visualizing enrichments

Computing connectivity

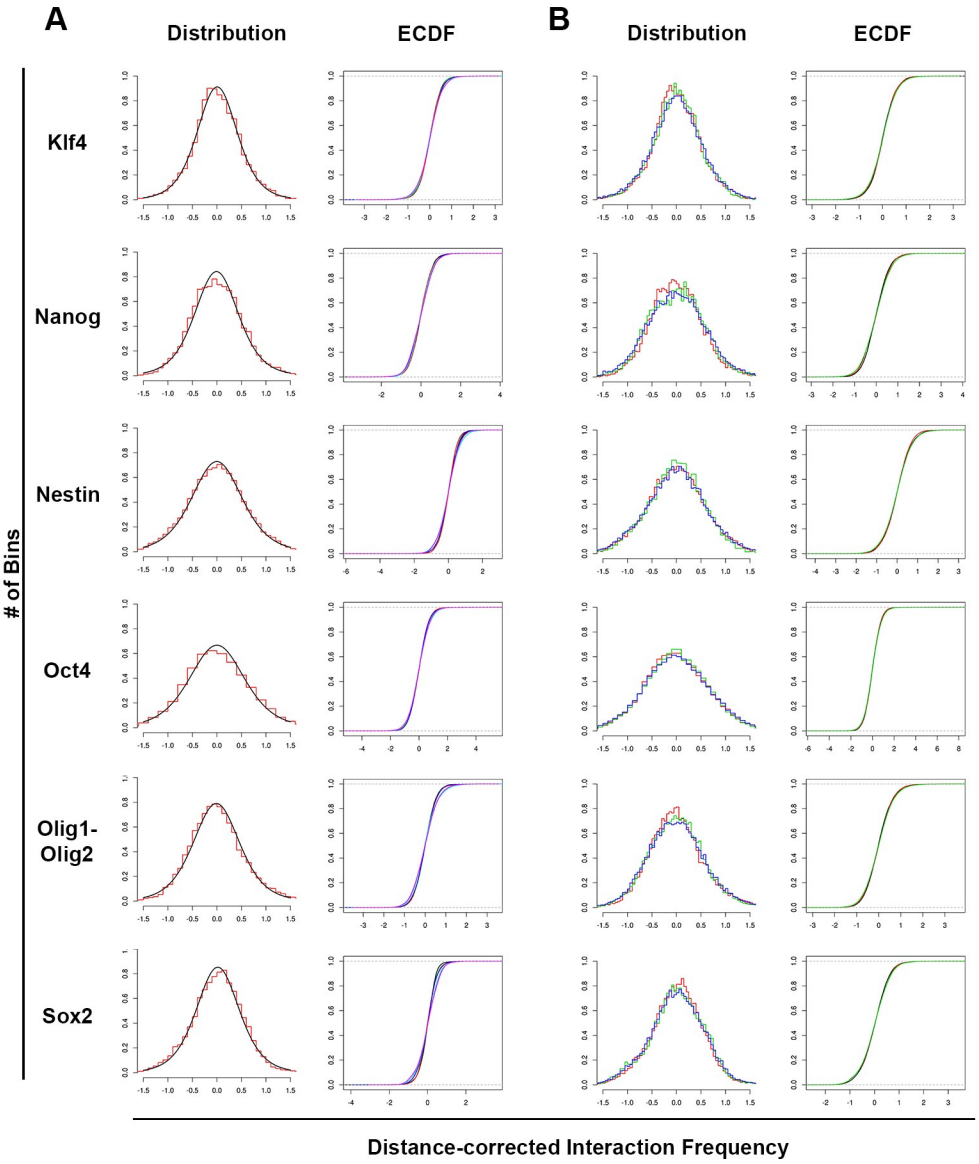
Supplementary References



Beagan et al. 2016 - Figure S1

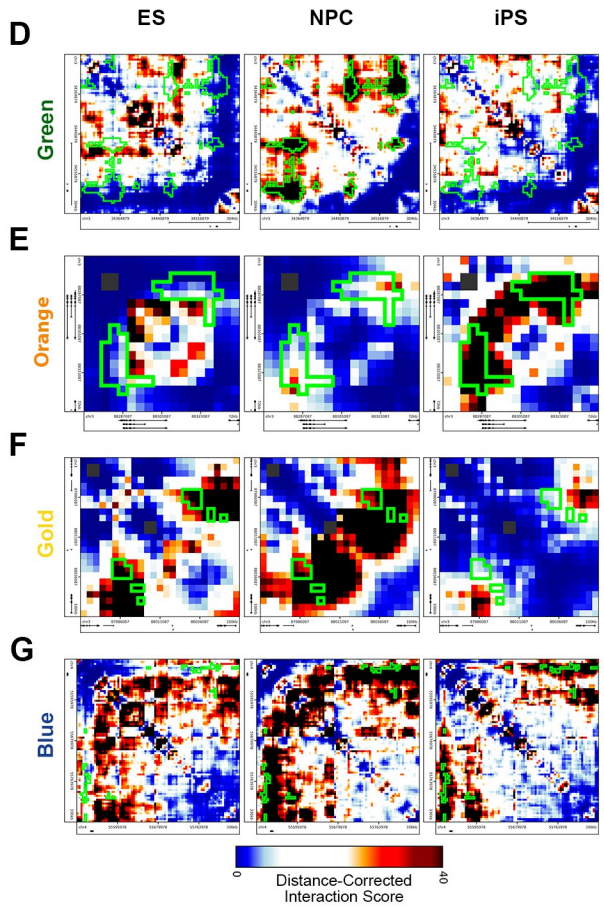


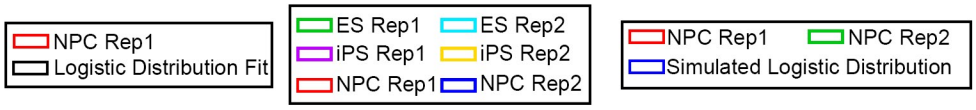
Beagan et al. 2016 - Figure S2

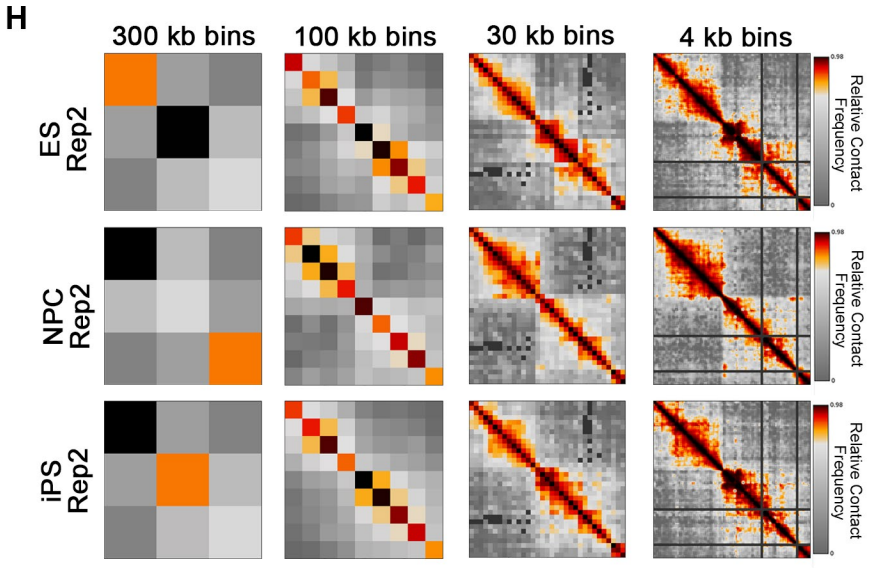


C

FDR	Red	Green	Orange	Gold	Purple	Blue
ES	2.1%	1.2%	3.6%	9.0%	2.5%	4.7%
NPC	3.9%	2.3%	6.9%	16.0%	4.4%	8.4%
iPS	7.1%	4.2%	12.4%	23.2%	6.4%	12.2%







I

Spearman's ρ, 300 kb bins

ES Rep2	1		
NPC Rep2	0.556	1	
iPS Rep2	0.874	0.678	1
	ES Rep2	NPC Rep2	iPS Rep2

Spearman's ρ, 100 kb bins

ES Rep2	1		
NPC Rep2	0.423	1	
iPS Rep2	0.748	0.595	1
	ES Rep2	NPC Rep2	iPS Rep2

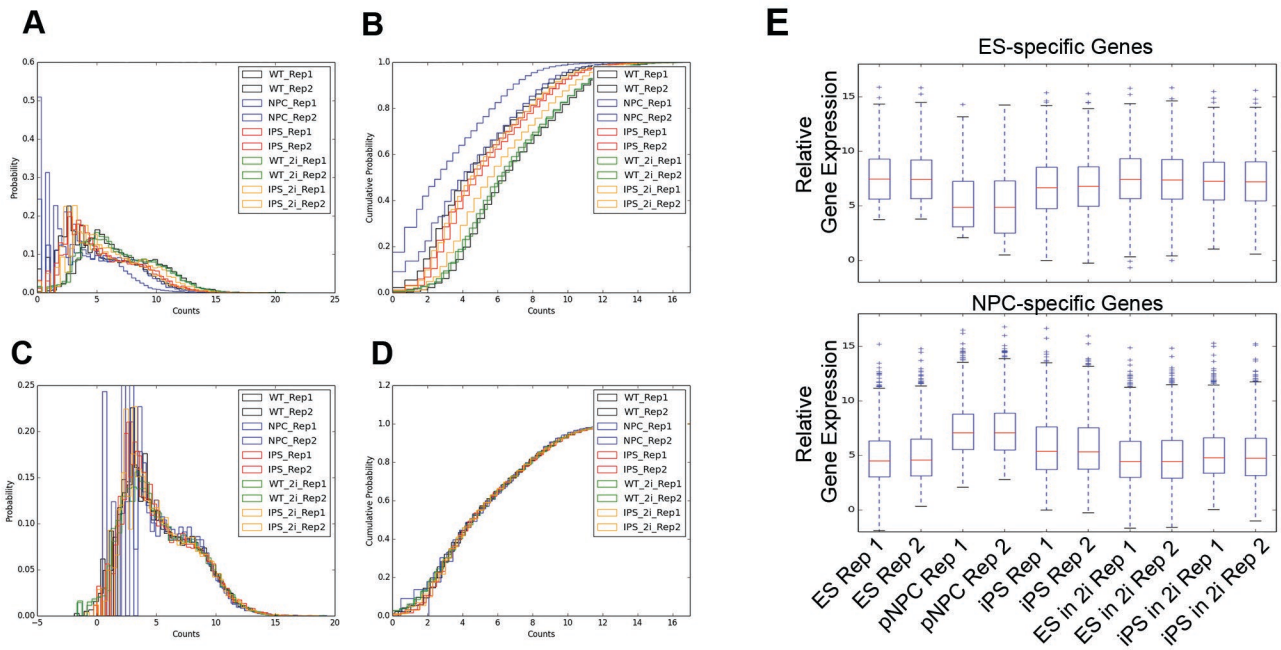
Spearman's ρ, 30 kb bins

ES Rep2	1		
NPC Rep2	0.332	1	
iPS Rep2	0.521	0.526	1
	ES Rep2	NPC Rep2	iPS Rep2

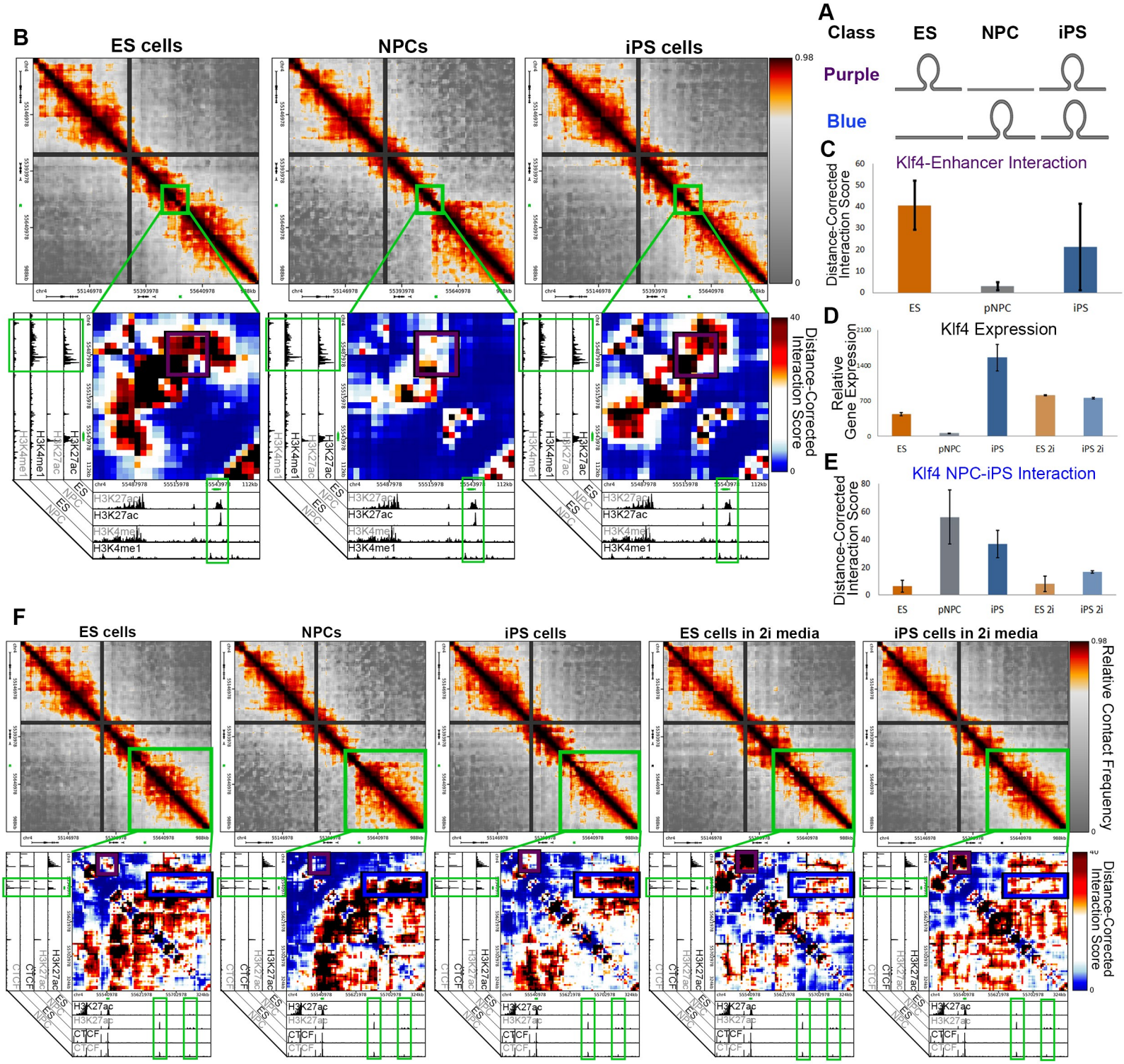
Spearman's ρ, 4 kb bins

ES Rep2	1		
NPC Rep2	0.268	1	
iPS Rep2	0.427	0.493	1
	ES Rep2	NPC Rep2	iPS Rep2

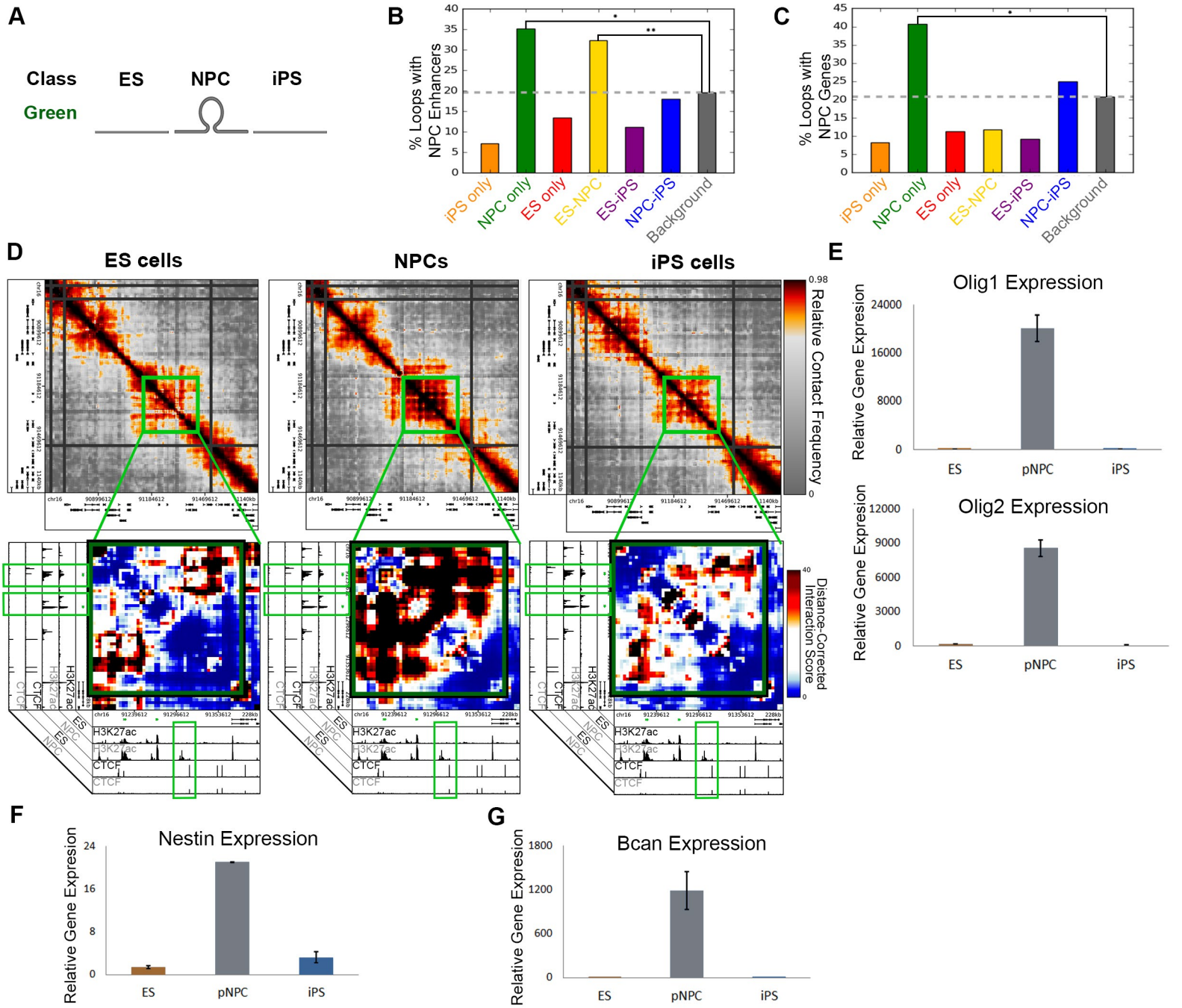
Beagan et al. 2016 - Figure S3



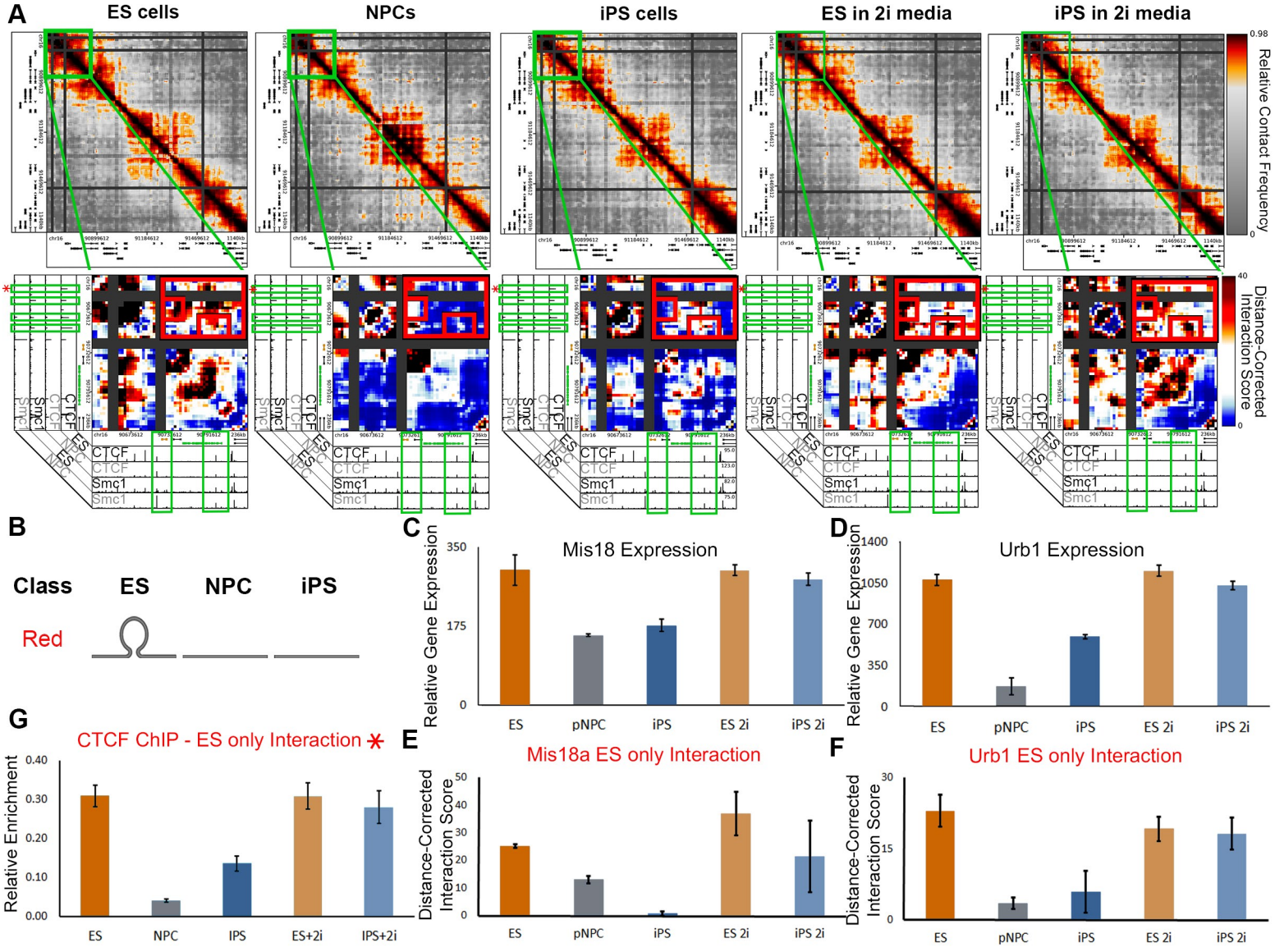
Beagan et al. 2016 - Figure S4



Beagan et al. 2016 - Figure S5



Beagan et al. 2016 - Figure S6



Beagan et al. 2016 - Figure S7

Supplemental Figure Captions

Figure S1 (related to Figure 1). 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 (see **Supplemental Experimental Procedures**).

Figure S2 (related to Figure 1). Progression of 5C data through alternative 5C analysis approaches. (A-D) Grid showing progression of Sox2 region through our previously published analysis pipeline (Phillips-Cremins et al., 2013). From top to bottom: **(A)** raw, **(B)** primer corrected, **(C)** distance-dependence normalized via parametric model described in (Phillips-Cremins et al., 2013) and **(D)** interaction score

computed as -10*$\log_2$(p-value) on p-values computed with compound normal-lognormal distribution fits described in (Phillips-Cremins et al., 2013). 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 S1, where quantile normalization is the first step) **(F)**, and finally distance corrected **(G)**.

Figure S3 (related to Figure 3). 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 **Supplemental Experimental Procedures**). **(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 S4 (related to Figures 2, 4, 5, 6). 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.

Figure S5 (related to Figures 4, 5). 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)**.

Figure S6 (related to Figure 5). 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.

Figure S7 (related to Figure 6). 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.

Supplemental Tables

Table S1: Summary of paired-end read alignments for 5C libraries, related to Experimental Procedures and Supplemental Experimental Procedures.

Library Code	Replicate	Instrument	Lane/Paired End	Total Reads	PE1 Mapped Reads	PE2 Mapped Reads
ES_1	1	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	33678848 33678848 33418892 33418892 33974768 33974768 33399920 33399920	28770023	28505219
ES_2	2	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	31551080 31551080 31299432 31299432 31772324 31772324 31309272 31309272	27875198	27628550
NPC_1	1	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	27804116 27804116 24416680 24416680 13862024 13862024 17389664 17389664	18454027	15832156
NPC_2	2	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1	27793844 27793844 24550324 24550324 13826704 13826704 17240756	18342888	15617223

			L4_P2	17240756		
iPS_1	1	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	23527984 23527984 20602800 20602800 11619608 11619608 14506996 14506996	15039775	13005171
iPS_2	2	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	24074808 24074808 21329464 21329464 11963384 11963384 14902364 14902364	15970612	13768584
ES_2i_1	1	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	22956884 22956884 19862384 19862384 11563004 11563004 14156912 14156912	15065438	12571131
ES_2i_2	2	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	26479112 26479112 23469892 23469892 13319924 13319924 16661424 16661424	17803279	15151910
iPS_2i_1	1	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	21352148 21352148 18236676 18236676 10483824 10483824 13062076 13062076	13147449	11301729
iPS_2i_2	2	Illumina Nextseq	L1_P1	23812716	15400978	12963765

		500	L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	23812716 21226860 21226860 12105132 12105132 15124608 15124608		

Table S2: Spearman's rank correlation coefficients calculated for distance-dependence corrected interaction frequencies of pairs of biological replicates, related to Experimental Procedures.

ES_Rep_1						
ES_Rep_2	0.830632					
NPC_Rep_1	0.280142	0.243655				
NPC_Rep_2	0.27191	0.267666	0.767335196			
iPS_Rep_1	0.548705	0.581172	0.302233915	0.374198865		
iPS_Rep_2	0.44135	0.426666	0.456490393	0.492875434	0.678932815	
	ES_Rep_1	ES_Rep_2	NPC_Rep_1	NPC_Rep_2	iPS_Rep_1	iPS_Rep_2

Table S3: Summary of paired-end read alignments for RNA-seq libraries, related to Experimental Procedures and Supplemental Experimental Procedures.

Library Code	Replicate	Instrument and Number of Lanes	Lane/Paired End	Total Reads	Alignment Summary
ES_1	1	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	63385276 63385276 62210488 62210488 59599184 59599184 59255860 59255860	Aligned pairs: 47708823 of these: 6941410 (14.5%) have multiple alignments 1525220 (3.2%) are discordant alignments 75.6% concordant pair alignment rate.
ES_2	2	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1	49406568 49406568 48742476 48742476 46795280 46795280 46667840	Aligned pairs: 24184676 of these: 6397683 (26.5%) have multiple alignments 5602245 (23.2%) are discordant alignments 38.8% concordant pair alignment rate.

			L4_P2	46667840	
			L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	28202868 28202868 27832228 27832228 27903044 27903044 27359632 27359632	Aligned pairs: 15612304 of these: 4167968 (26.7%) have multiple alignments 3682919 (23.6%) are discordant alignments 42.9% concordant pair alignment rate.
NPC_1	1	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	34176084 34176084 33607532 33607532 33839124 33839124 33064788 33064788	Aligned pairs: 16843964 of these: 5246902 (31.2%) have multiple alignments 4743602 (28.2%) are discordant alignments 35.9% concordant pair alignment rate.
NPC_2	2	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	32832608 32832608 32294456 32294456 32521560 32521560 31787280 31787280	Aligned pairs: 19633261 of these: 5907437 (30.1%) have multiple alignments 2764224 (14.1%) are discordant alignments 52.1% concordant pair alignment rate.
iPS_1	1	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	66486724 66486724 65682424 65682424 62943540 62943540 62797608 62797608	Aligned pairs: 30717628 of these: 8903977 (29.0%) have multiple alignments 7379227 (24.0%) are discordant alignments 36.2% concordant pair alignment rate.

			L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	23466568 23466568 23169000 23169000 23272384 23272384 22782364 22782364	Aligned pairs: 12293064 of these: 3617560 (29.4%) have multiple alignments 2992072 (24.3%) are discordant alignments 40.1% concordant pair alignment rate.
iPS_2	2	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	45551664 45551664 44876400 44876400 43097496 43097496 42875224 42875224	Aligned pairs: 22993950 of these: 6316523 (27.5%) have multiple alignments 4420192 (19.2%) are discordant alignments 42.1% concordant pair alignment rate.
			L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	29625648 29625648 29151848 29151848 29348000 29348000 28673296 28673296	Aligned pairs: 16810920 of these: 4686563 (27.9%) have multiple alignments 3287433 (19.6%) are discordant alignments 46.3% concordant pair alignment rate.
ES_2i_1	1	Illumina Nextseq 500	L1_P1 L1_P2 L2_P1 L2_P2 L3_P1 L3_P2 L4_P1 L4_P2	59127460 59127460 58169908 58169908 55872492 55872492 55774136 55774136	Aligned pairs: 42262509 of these: 6635919 (15.7%) have multiple alignments 2569916 (6.1%) are discordant alignments 69.3% concordant pair alignment rate.
				7370792 7370792 7260416 7260416 7299252 7299252 7149924 7149924	Aligned pairs: 5861538 of these: 950297 (16.2%) have multiple alignments 370120 (6.3%) are discordant alignments 75.5% concordant pair alignment rate.

ES_2i_2	2	Illumina Nextseq 500		41617840 41617840 40991972 40991972 39186452 39186452 39053336 39053336	Aligned pairs: 31055590 of these: 4668288 (15.0%) have multiple alignments 1206653 (3.9%) are discordant alignments 74.2% concordant pair alignment rate.
				12881568 12881568 12701244 12701244 12733836 12733836 12471728 12471728	Aligned pairs: 10705467 of these: 1658953 (15.5%) have multiple alignments 426422 (4.0%) are discordant alignments 81.0% concordant pair alignment rate.
iPS_2i_1	1	Illumina Nextseq 500		43012836 43012836 42257896 42257896 40724336 40724336 40395388 40395388	Aligned pairs: 23098024 of these: 6101858 (26.4%) have multiple alignments 4051057 (17.5%) are discordant alignments 45.8% concordant pair alignment rate.
				7600884 7600884 7470076 7470076 7525056 7525056 7366240 7366240	Aligned pairs: 4591177 of these: 1247317 (27.2%) have multiple alignments 821522 (17.9%) are discordant alignments 50.3% concordant pair alignment rate.
iPS_2i_2	2	Illumina Nextseq 500		45249896 45249896 44710976 44710976 42713772 42713772 42689140 42689140	Aligned pairs: 31164351 of these: 6059475 (19.4%) have multiple alignments 2368280 (7.6%) are discordant alignments 65.7% concordant pair alignment rate.

Table S4: Summary of external ChIP-seq libraries analyzed in this study, related to Experimental Procedures and Supplemental Experimental Procedures.

Antibody	Cell Type	Mapped Test ChIP-Seq reads	Test ChIP Reference	Test Sample GEO ID	Control Samples	Mapped Control ChIP-Seq reads	Control Sample GEO ID
CTCF	mES (159-2)	9,562,677	(Stadler et al., 2011)	GSM747534	mES Whole Cell Extract	10,202,630	GSM747545
CTCF	ES-derived NPC	13,641735	(Phillips-Cremins et al., 2013)	GSM883647	NPC Whole Cell Extract	14,041,323	GSM883648
H3K4me1	mES (V6.5)	5,707,101	(Meissner et al., 2008)	GSM281695	V6.5 Whole Cell Extract	803,601	GSM307625
H3K4me1	ES-derived NPC	4,471,210	(Meissner et al., 2008)	GSM281693	NPC Whole Cell Extract	4,369,951	GSM307617
H3K4me3	mES (V6.5)	6,809,878	(Mikkelsen et al., 2007)	GSM307618	V6.5 Whole Cell Extract	6,008,440	GSM307154, GSM307155
H3K4me3	ES-derived NPC	3,397,613	(Mikkelsen et al., 2007)	GSM307613	NPC Whole Cell Extract	4,369,951	GSM307617
H3K27ac	mES (V6.5)	11,128,384	(Creyghton et al., 2010)	GSM594579 Rep2	V6.5 Whole Cell Extract	14,682,811	GSM307154, GSM307155, GSM594599
H3K27ac	ES-derived NPC	8,831,628	(Creyghton et al., 2010)	GSM594585	NPC Whole Cell Extract	14,041,323	GSM883648

Table S5: Genes classified as ES-specific, Related to Figures 4, 5, 6 and Supplemental Experimental Procedures.

Attached as separate excel spreadsheet.

Table S6: Genes classified as NPC-specific, Related to Figures 4, 5, 6 and Supplemental Experimental Procedures.
Attached as separate excel spreadsheet.

Table S7: Interactions selected for representative interaction score barplots, Related to Figures 4, S5, 5, 6, S7.
Attached as separate excel spreadsheet.

Supplemental Experimental Procedures

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 (Eminli et al., 2008). 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 (Eminli et al., 2008)**.**

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) (Rais et al., 2013). 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 (Dekker et al., 2002; Gheldof et al., 2010; Phillips-Cremins et al., 2013; van Berkum and Dekker, 2009) 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 (Lajoie et al., 2009), as described in detail in (Phillips-Cremins et al., 2013).

5C library generation and sequencing

5C libraries were generated as described previously (Bau et al., 2011; Dostie and Dekker, 2007; Dostie et al., 2006; Phillips-Cremins et al., 2013; van Berkum and Dekker, 2009). 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 (Eminli et al., 2008). 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 (Trapnell et al., 2009) with the parameters: -r 100 --no-coverage-search --library-type fr-firststrand and UCSC gene annotations (**Table S3**). 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 S4**).

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	Chr3:34658834-34659306
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) (Langmead, 2010). 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 (**Table S1**).

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. S1A**). 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 (Bolstad et al., 2003), RNAseq (Bullard et al., 2010) and Hi-C (Dixon et al., 2015) 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. S1A-B**)

while preserving biologically significant architectural features (compare heatmaps in **Fig. S1A-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. S2A+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. S1A-F**).

Primer correction

Consistent with our findings in (Phillips-Cremins et al., 2013), 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. S1A).** 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. S1A**). 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 (Phillips-Cremins et al., 2013). After the primer correction step, we observed a marked attenuation of primer-specific artifacts (heatmaps and boxplots, **Fig. S1C**).

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. S1C+D** and **Fig. S2B+E**). This reduction of spatial noise is concurrent with a tightening of the distribution of counts across this step (compare histograms in **Fig. S1C+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 (Rao et al., 2014). 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 (Phillips-Cremins et al., 2013) 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. S1G**). 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. S1E, S2G**). 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. S1E, 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). Spearman's rank-order correlation coefficients for distance-corrected interaction frequencies of ES, NPC, and iPS replicates can be found in **Table S2**.

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. S3A**). 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:

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

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. S1F**). 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. S1F**) 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. S1F**).

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. S1H**). 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. S1H**). 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 (Jin et al., 2013) 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. S1I**). 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 (Jin et al., 2013).

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. S2A-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 (Phillips-Cremins et al., 2013).

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. S1C and S2B**). 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. S1E and S2C**). 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. S1D-F and S2C-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. S1F and S2D).** 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. S2E-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. S1F**).

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. 3A-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. 3**). 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. The bin numbers of the interactions whose interaction scores are presented in barplots in **Figs. 4, S5, 5, 6, S7** can be found in **Table S7**.

*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. 3, S3**) 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 $\bar{n}_t^{\mathrm{sim}}$ is the average number of bin-bin pairs categorized into category t across all simulations and $\mathrm{FP}_t^{\mathrm{exp}}$ 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:

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

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. 3H-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

A summary of all ChIP-seq data sets re-analyzed in this study is provided in **Table S4**. 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 (found in **Table S5**). NPC-specific genes were further refined by required overlap with high-confidence H3K27ac signal (peaks called at $p < 1 \times 10^{-4}$) in NPCs (found in **Table S6**). 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. A summary of all ChIP-seq datasets utilized can be found in **Table S4**.

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. A summary of all ChIP-seq datasets utilized can be found in **Table S4**.

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. 6D**).

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. 7**), 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.

Supplemental References

Bau, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J., and Marti-Renom, M.A. (2011). The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol *18*, 107-114.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185-193.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics *11*, 94.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A.*, et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A *107*, 21931-21936.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306-1311.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W.*, et al.* (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331-336.

Dostie, J., and Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc *2*, 988-1002.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C.*, et al.* (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res *16*, 1299-1309.

Eminli, S., Utikal, J., Arnold, K., Jaenisch, R., and Hochedlinger, K. (2008). Reprogramming of neural progenitor cells into induced pluripotent stem cells in the absence of exogenous Sox2 expression. Stem Cells *26*, 2467-2474.
Gheldof, N., Smith, E.M., Tabuchi, T.M., Koch, C.M., Dunham, I., Stamatoyannopoulos, J.A., and Dekker, J. (2010). Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. Nucleic Acids Res *38*, 4325-4336.
Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290-294.
Lajoie, B.R., van Berkum, N.L., Sanyal, A., and Dekker, J. (2009). My5C: web tools for chromosome conformation capture studies. Nat Methods *6*, 690-691.
Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics *Chapter 11*, Unit 11 17.
Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B.*, et al.* (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766-770.
Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P.*, et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.
Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y.*, et al.* (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281-1295.
Rais, Y., Zviran, A., Geula, S., Gafni, O., Chomsky, E., Viukov, S., Mansour, A.A., Caspi, I., Krupalnik, V., Zerbib, M.*, et al.* (2013). Deterministic direct reprogramming of somatic cells to pluripotency. Nature *502*, 65-70.
Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S.*, et al.* (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.
Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D.*, et al.* (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490-495.
Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.
van Berkum, N.L., and Dekker, J. (2009). Determining spatial chromatin organization of large genomic regions using 5C technology. Methods Mol Biol *567*, 189-213.