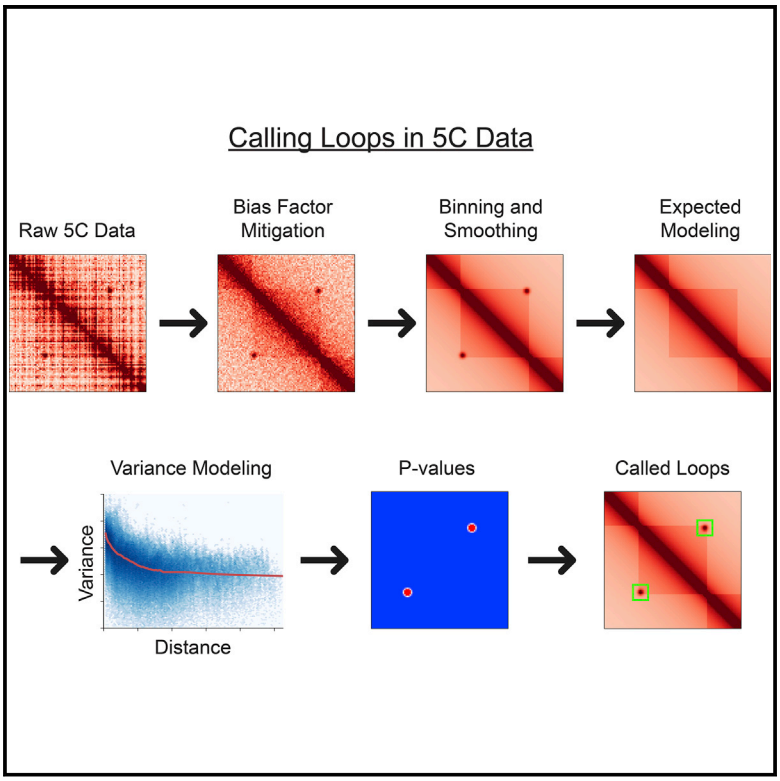
Article

Cell Systems

Systematic Evaluation of Statistical Methods for Identifying Looping Interactions in 5C Data

Graphical Abstract



Highlights

- Identifying loops in 5C data involves modeling and correcting a host of technical biases
- Final loop calls depend on the statistical methods applied at each analysis step
- We survey a range of existing and new analysis approaches and propose our own fine-tuned pipeline

Authors

Thomas G. Gilgenast,
Jennifer E. Phillips-Cremins

Correspondence

jcremins@seas.upenn.edu

In Brief

5C promises to elucidate the three-dimensional structure of specific regions of the genome at high resolution and low cost, but the results may be sensitive to key choices during data analysis. We survey a range of a existing and new computational methods and propose our own fine-tuned 5C analysis pipeline.

Gilgenast & Phillips-Cremins, 2019, Cell Systems *8*, 1–15
March 27, 2019
https://doi.org/10.1016/j.cels.2019.02.006

CellPress

Please cite this article in press as: Gilgenast and Phillips-Cremins, Systematic Evaluation of Statistical Methods for Identifying Looping Interactions in 5C Data, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.02.006

Cell Systems
Article

Systematic Evaluation of Statistical Methods for Identifying Looping Interactions in 5C Data

Thomas G. Gilgenast[1] and Jennifer E. Phillips-Cremins[1,2,3,4,*]
[1]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Lead Contact
*Correspondence: jcremins@seas.upenn.edu
https://doi.org/10.1016/j.cels.2019.02.006

SUMMARY

Chromosome-Conformation-Capture-Carbon-Copy (5C) is a molecular technology based on proximity ligation that enables high-resolution and high-coverage inquiry of long-range looping interactions. Computational pipelines for analyzing 5C data involve a series of interdependent normalization procedures and statistical methods that markedly influence downstream biological results. A detailed analysis of the trade-offs inherent to all stages of 5C data analysis has not been reported. Here, we provide a comparative assessment of method performance at each step in the 5C analysis pipeline, including sequencing depth and library complexity correction, bias mitigation, spatial noise reduction, distance-dependent expected and variance estimation, statistical modeling, and loop detection. We discuss methodological advantages and disadvantages at each step and provide a full suite of algorithms, lib5C, to allow investigators to test the range of approaches on their own 5C data. Principles learned from our comparative analyses can be applied to protein-independent proximity ligation-based data, including Hi-C, 4C, and Capture-C.

INTRODUCTION

Higher-order chromatin folding in the three-dimensional nucleus is critically linked to genome function, including transcription (Deng and Blobel, 2014), replication (Rhind and Gilbert, 2013), recombination (Jhunjhunwala et al., 2009), and X chromosome inactivation (Nora et al., 2012). Molecular methodologies based on proximity ligation and deep sequencing have revealed that genomes are arranged into a hierarchy of complex configurations (Dixon et al., 2012; Lieberman-Aiden et al., 2009). One unique folding feature is the spatial juxtaposition of linearly distant genomic loci into long-range contacts termed looping interactions. More than 10,000 looping interactions have been identified genome-wide in ultra-high-resolution genome architecture maps in human cell lines (Rao et al., 2014). Efforts are currently underway to identify loops across a range of species, cell types, and genetic perturbations (Dekker et al., 2017). As genome-wide loop-resolution maps become widely available across a range of biological conditions, the field will transition to perturbation studies required to dissect the organizing principles and mechanistic roles of specific classes of long-range interactions.

Chromosome-conformation-capture-carbon-copy (5C) is a leading technique for mapping genome folding (Dostie et al., 2006). 5C adds a PCR-based hybrid capture step to the classic proximity ligation procedure, chromosome-conformation-capture (3C), to amplify only ligation junctions across contiguous regions spanning a subset of the genome. The promise of 5C is that genome contacts across several Mb-sized genomic regions may be identified at restriction fragment-level resolution without the high cost of genome-wide Hi-C if the various technical biases, spatial noise, and statistical variance are modeled appropriately. Specifically, 5C requires only 10–30 million reads for fragment-level (250–4,000 kb) resolution chromatin contact maps, whereas Hi-C requires 1–6 billion reads to obtain genome-wide maps at a similar resolution in mammalian systems (Rao et al., 2014). 5C-based technologies continue to evolve and mature, with cutting-edge approaches based on double-alternating primer designs enabling dramatically improved resolution and data quality (Hnisz et al., 2016; Kim et al., 2018). Thus, 5C has a key strength in allowing researchers to create high-resolution chromatin folding maps at specific genomic region(s) across hundreds of biological conditions and perturbations at a fraction of the cost of Hi-C.

The extent to which looping interactions differ among cell types is currently unknown, in part because the methodologies for identifying loops in proximity ligation-based sequencing data vary widely across studies and can dramatically influence the results. A systematic comparison of methods for processing, normalizing, and modeling 5C data with the goal of detecting loops has not been conducted. Moreover, no gold-standard set of algorithms for loop detection in 5C data has been reported. Here, we provide a suite of algorithms, lib5C, for direct systematic comparison of multiple methods at each stage in the 5C analysis pipeline, including (1) sequencing depth and library complexity correction, (2) bias mitigation, (3) contact matrix binning, (4) distance-dependent expected signal and variance estimation, and (5) statistical modeling for the goal of loop detection (Figure S1). We compare and contrast the strengths and weaknesses of each method and make recommendations for approaches that yield high-confidence looping interactions. Together, our described approaches and freely available lib5C

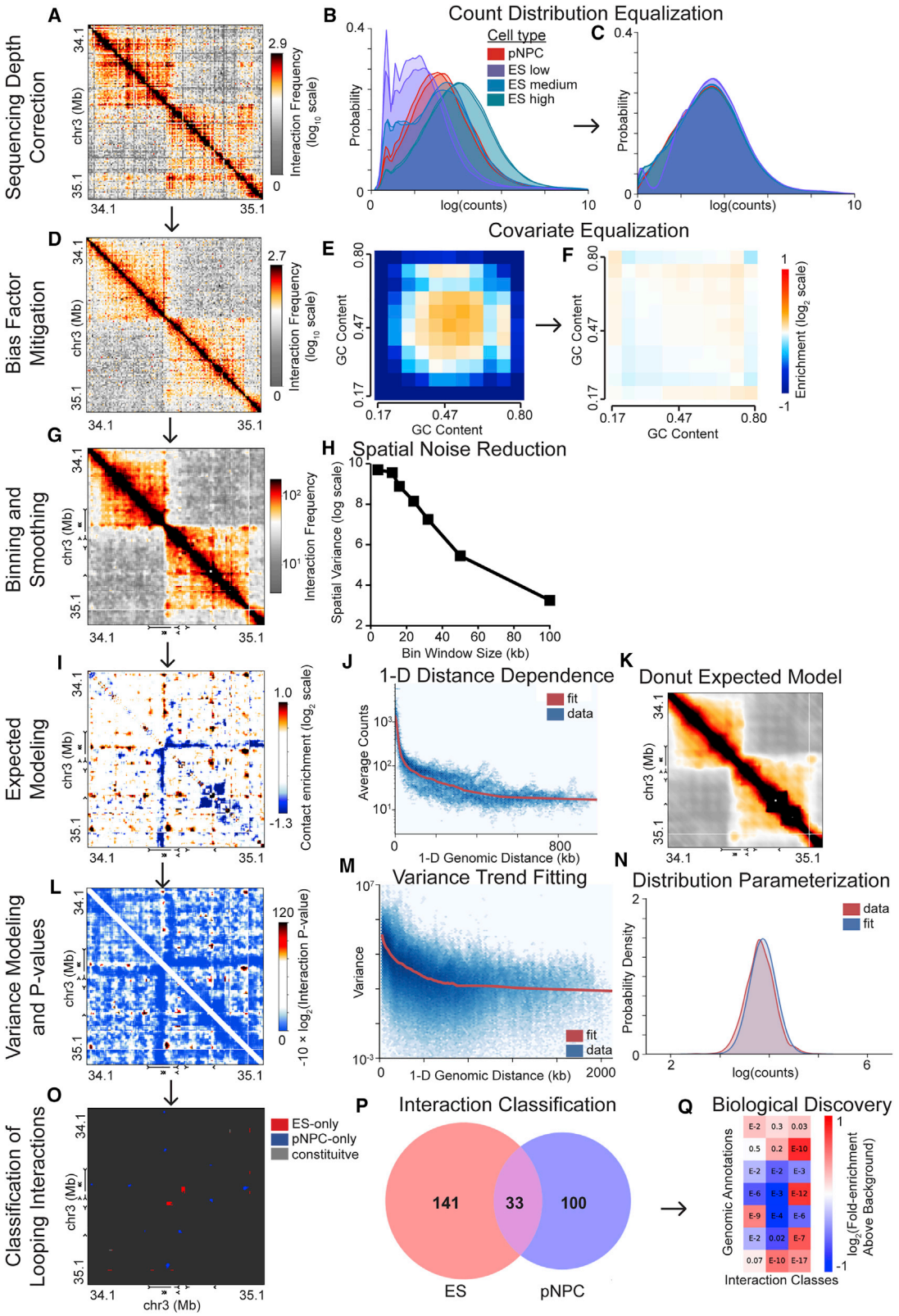


Figure 1. Schematic of 5C Analysis Pipeline

Flowchart illustrating key steps in 5C analysis.

(A) Interaction frequency heatmap of the Sox2 region in primary neural progenitor cells (pNPCs) after sequencing depth correction.

(legend continued on next page)

tools (https://bitbucket.org/creminslab/lib5c/) allow for the sensitive and specific detection of looping interactions from 5C data.

RESULTS

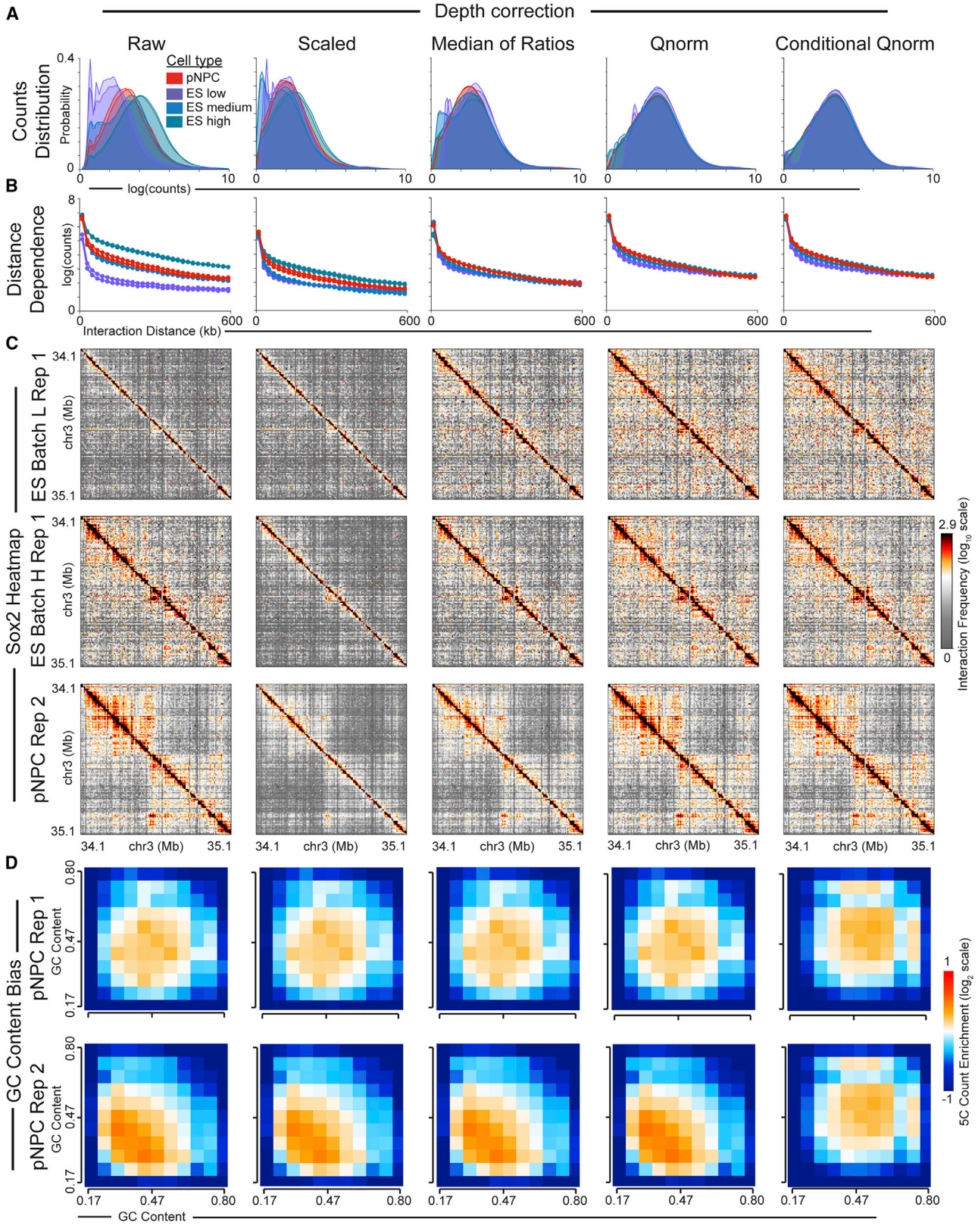
Protein-independent proximity ligation-based techniques (e.g., Hi-C, 5C, 4C, Capture-C) assay many types of non-biological signal along with bona fide looping interactions (Yaffe and Tanay, 2011). Biases vary in their type and severity depending on the method and require modeling and correction prior to biological interpretation. To quantitatively assess biases and compare and contrast the various analysis techniques available, we reanalyzed published 5C data comparing the chromatin interaction patterns of murine v6.5 embryonic stem (ES) cells to those of primary neural progenitor cells (pNPCs) isolated from whole brains of P1 129SvJae x C57/BL6, Sox2-eGFP mice (Beagan et al., 2016). Each condition in this dataset includes two biological replicates created from independent cultures of the source cells. These published 5C data relied on the use of a single alternating primer design (Phillips-Cremins et al., 2013); therefore, they contained significantly more spatial noise and bias than the more recent 5C libraries generated from double alternating designs (Hnisz et al., 2016; Kim et al., 2018). We strategically focused on the analysis of older, single alternating 5C to ensure that our algorithms were robust to low-quality data. All principles reported in this manuscript were robust across low-quality single-alternating and high-quality double-alternating 5C primer designs.

In the specific case of 5C, possible artifacts or confounding signal include (1) sequencing depth and library complexity differences due to technical artifacts and/or batch effects (Figures 1A–1C), (2) biases caused by the intrinsic properties of the restriction fragments queried by the assay (including their length and guanine-cytosine [GC] content) (Figures 1D–1F), (3) spatial noise due to 5C primer design and library complexity (Figures 1G and 1H), and (4) the expected background signal at each length scale, which varies as a function of spatial genomic distance and region-specific topologically associating domain (TAD) and subTAD structure (Figures 1I–1K). Upon correction of these features, it is also critical to understand the distance-variance relationship (DVR) and parameterize an appropriate statistical model to assign p values to each possible interaction (Figures 1L–1N). Finally, we call loops as clusters of highly significant pixels and explore the downstream enrichment of traditional, one-dimensional epigenetic marks (Figures 1O–1Q). By rigorously modeling the data, investigators can distinguish loops from other genome folding patterns as well as background signal and technical noise, thus providing the opportunity for the discovery of biological mechanisms governing long-range interactions.

Similar to all genomics assays (Daley and Smith, 2013; Marinov et al., 2014; Sims et al., 2014), 5C libraries can exhibit large differences in complexity and sequencing depth due to technical variation in ligation and fixation efficiency among experimenters, reagents, and protocols. We observed that technical 5C replicates from the same biological condition can show a high degree of variability in their raw counts distribution (Raw; Figure 2A), distance-dependent expected counts distribution (Raw; Figure 2B), spatial complexity (Raw; Figure 2C), and relationship between the raw interaction counts and GC content of the fragments hybridizing to 5C primers (Raw; Figure 2D). Differences in count distributions, distance-dependent expected curves, and spatial noise trend more with technical batch compared with biological conditions (Raw; Figures 2A–2C), suggesting that they are driven by library complexity and sequencing depth rather than biologically driven effects. Thus, raw 5C counts exhibit biases that require correction prior to biological interpretation.

To compare looping interactions between biological conditions, it is essential to correct for differences in library complexity and sequencing depth. We find that correcting raw counts by a scalar value of total sequencing reads is insufficient to correct for replicate-to-replicate differences in the shape of the raw counts distribution and distance-dependent expected curve (Scaled; Figures 2A–2C). By contrast, correction via the median-of-ratios scaling technique (Anders and Huber, 2010) or quantile normalization can more rigorously normalize the raw counts distributions and distance-dependent expected curves (median-of-ratios and qnorm; Figures 2A–2C). We find that equalizing distributions among replicates of the same condition does not sufficiently equalize GC content bias profiles. Indeed, the relationship between raw interaction signal and the GC content of the ligation junctions remains widely variable between replicates even after quantile normalization or median-of-ratios

(B and C) 5C count distributions (B) before and (C) after sequencing depth correction.
(D) Interaction frequency heatmap of the Sox2 region in pNPCs after bias factor mitigation.
(E and F) GC content bias profile (E) before and (F) after bias mitigation.
(G) Interaction frequency heatmap of the Sox2 region in pNPCs after binning and smoothing.
(H) Spatial variance of the binned contact matrix as a function of the width of the bin window size used during binning.
(I) Heatmap of the Sox2 region in pNPCs showing enrichment (red) and depletion (blue) of contacts relative to the donut expected model.
(J) Illustration of a one-dimensional distance dependence model, which describes the average interaction frequency as a function of linear genomic separation.
(K) Donut expected model interaction frequency heatmap for the Sox2 region in pNPCs after distance-dependence modeling and donut correction.
(L) Heatmap of the Sox2 region in pNPCs showing interaction p values on a scale of $-10 \times \log_2$(right-tail p value).
(M) Relationship between genomic interaction distance and variance of contact frequency across replicates.
(N) Lognormal distribution parameterized using the distance-variance relationship (blue), overlaid with observed data near an expected value of 50 (red).
(O) Heatmap of the Sox2 region showing classified interactions.
(P) Venn diagram showing the numbers of ES-specific, pNPC-specific, and constitutive interactions.
(Q) Enrichment heatmap showing relationships between various classes of significant interactions (rows) and various genomic annotations (columns). Certain classes of interactions may be found to be enriched for certain genomic annotations (red squares on the heatmap) while other combinations may not be enriched (blue and white squares), suggesting possible biological mechanisms or effects of chromatin looping interactions.
See also Figure S1.



(legend on next page)

correction (Figure 2D). These data indicate that most published methods for correcting sequencing depth and library complexity differences in proximity ligation data are insufficient to account for intra-condition technical variation among 5C replicates.

To account for the strong replicate-specific effect of GC content bias on raw interaction count, we developed and applied a variation of the conditional quantile normalization method proposed by Hansen and colleagues for RNA sequencing (RNA-seq) (Hansen et al., 2012). Specifically, we stratified all pairs of restriction fragments by the GC content of the portion of the DNA sequence homologous to the 5C primers. We conditionally quantile normalized ligation junctions in each GC content stratum across all 5C replicates from all biological conditions. Our conditional quantile normalization procedure fully corrected 5C libraries for replicate-specific distributional differences (conditional qnorm; Figures 2A–2C) and GC content bias profiles (conditional qnorm; Figure 2D) without any distortion to the underlying heatmaps (conditional qnorm; Figure 2C). Altogether, our new conditional quantile normalization method offers robust correction for sequencing depth and library complexity differences between technical replicates without negatively affecting the underlying condition-specific genome folding patterns.

Although GC content profiles have been equalized between samples after conditional quantile normalization (Figure 2D), each individual sample still exhibits strong fragment-dependent GC content biases that must then be corrected prior to the detection of looping interactions (Figure 3). Hi-C ligation junctions are also known to exhibit read count biases linked to the intrinsic properties of the fragments (e.g., GC content, fragment length, and mappability) (Jin et al., 2013; Yaffe and Tanay, 2011). The effects of intrinsic biases are not localized to particular pairs of interacting fragments, such as those engaged in looping interactions, but instead increase or decrease the raw interaction counts for all ligation partners of the fragment in question. Thus, intrinsic biases are made manifest as "lines" of under- or over-enriched counts spanning a significant proportion of the raw fragment-fragment contact matrices. Visual inspection confirmed that bias "lines" also exist in 5C data (Raw; Figure 3A, blue arrows). We observed this phenomenon more quantitatively as a wide dynamic range (over 80-fold difference in medians) of interaction count profiles among the restriction fragments (Raw; Figure S2A). We also quantified the presence of lines in the heatmaps by computing the sample variance of the row sums of the contact matrix (Raw; Figures S2C and S2D). An important consequence of fragment bias is that it can obscure biological signal due to looping events, as evidenced by zoom-in heatmaps of two previously reported looping interactions (Beagan et al., 2016) (Raw, Figure 3B). Together, these data highlight that intrinsic fragment biases should be corrected before calling significant biological interactions in 5C data.

Many approaches to correcting intrinsic fragment artifacts in Hi-C data have been reported, but the performance of the correction methods (i.e., explicit bias factor modeling, Knight-Ruiz or ICED matrix balancing, and Express matrix balancing) on 5C data has not been systematically assessed. Moreover, bias correction is complicated by the fact that the "lines" observed on the heatmaps can be caused by loop extrusion via cohesin (Fudenberg et al., 2016; Sanborn et al., 2015), which is difficult to disentangle from the technical bias. We first explored our 5C data for biases due to primer GC content and fragment length previously reported in Hi-C experiments (Jin et al., 2013; Yaffe and Tanay, 2011). Consistent with previous reports, we observed a strong under-representation of detected ligation junctions between fragments with extreme GC content (Raw; Figure 3C). We also observed a trend toward more frequent detection of ligation junctions between larger fragments (Raw; Figure S2B). Finally, we stratified fragments according to the normalized ChIP-seq signal in a 4 kb interval centered on the fragment midpoint and found that ligation junctions tend to exhibit stronger interaction frequency when the architectural protein CTCF exhibits high occupancy in both fragments (Raw; Figure 3D). These results are consistent with previous findings that CTCF anchors the base of looping interactions genome-wide (Beagan et al., 2017; Li et al., 2010; Rao et al., 2014; Sanyal et al., 2012; Tang et al., 2015). Thus, both candidate bias factors and epigenetic marks are covariates that may contribute to the 5C interaction counts.

We next surveyed a variety of methods for attenuating GC content and fragment length biases while keeping the known biological link between CTCF and interaction strength intact. We first explicitly modeled and corrected the conditional quantile normalized counts for GC content and restriction fragment length biases (detailed in STAR Methods). This approach is conceptually similar to seminal approaches in which biases were explicitly modeled and corrected in Hi-C data (Jin et al., 2013; Yaffe and Tanay, 2011). Although GC content and restriction fragment length bias effects were almost completely attenuated after explicit modeling and correction (Explicit; Figures 3B and S2B), the "lines" in the heatmaps remained, and individual fragments still showed a wide dynamic range of counts (Explicit; Figures 3A and S2A). The core assumptions behind explicit modeling approaches are that (1) all the intrinsic factors contributing to technical bias are known and (2) their influence on detected ligation junction counts can be modeled reasonably well with tractable functions. Thus, the explicit GC-content and restriction fragment length modeling results to date suggest that there are still unknown 5C bias factors or that current models

Figure 2. 5C Sequencing Depth Correction Methods

(A) Count distributions across replicates from various cell types before and after the application of sequencing depth correction procedures. Replicates from pNPCs are shown in red while replicates from ES cells are shown in various shades of blue, according to relative sequencing depth (low, medium, or high).

(B) Similar comparison for distance-dependence curves.

(C) Fragment-level contact frequency heatmaps of the region around the Sox2 gene. The noisy nature of the fragment-level 5C data prior to binning is due to the single alternating 5C primer design that does not query every fragment.

(D) Heatmaps showing GC content bias profiles of selected replicates from the same cell type. The color indicates the average relative enrichment of ligation detection events as a function of the GC content of the primer designed to one of the participating fragments (x axis) and that of the other fragment (y axis). Differences in bias profiles between two replicates of the same biological condition suggest the presence of uncorrected technical biases.

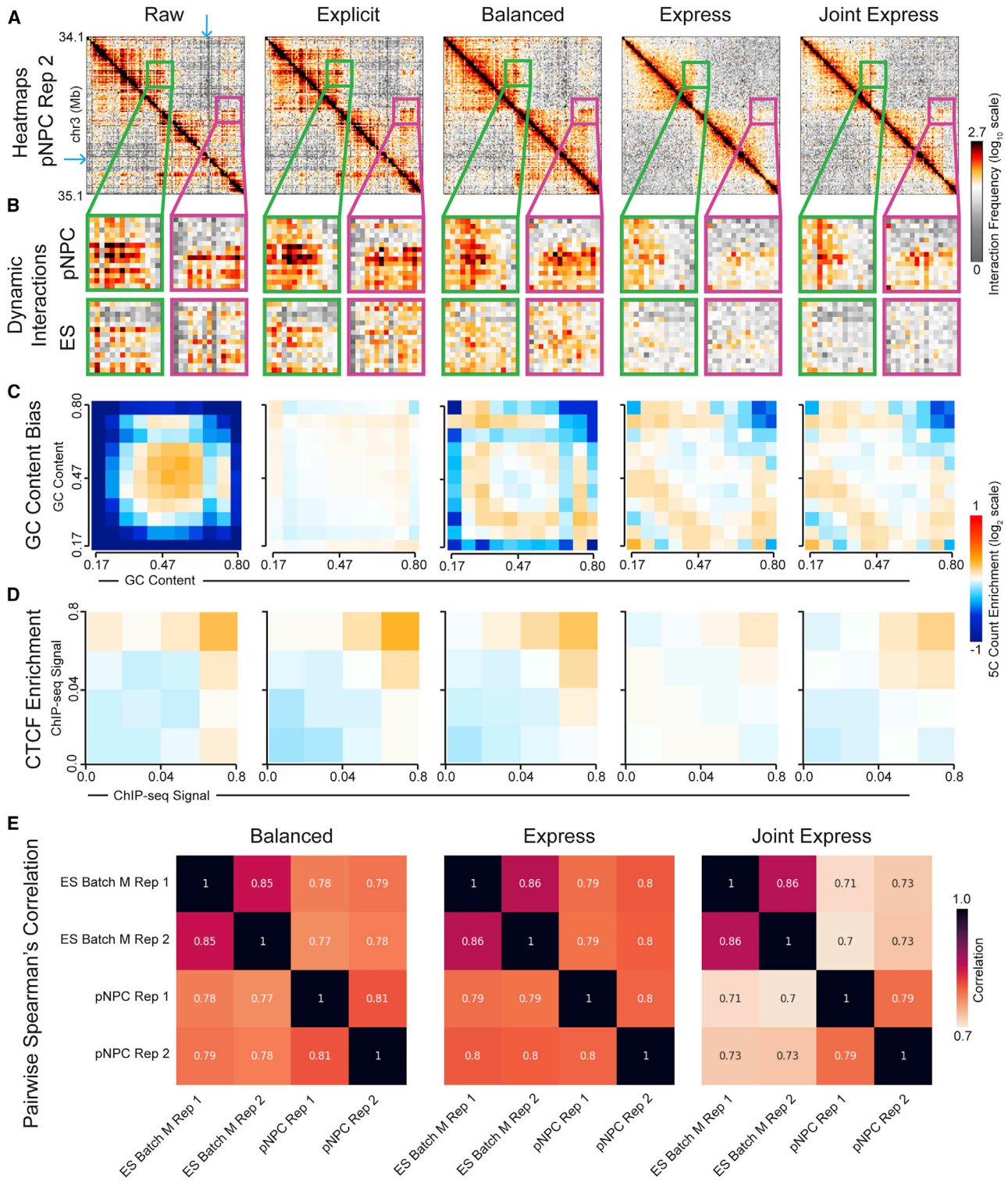


Figure 3. 5C Bias Factor Mitigation Procedures

(A) Fragment-level contact frequency heatmaps of the region around the Sox2 gene in pNPCs. Blue arrows highlight under-represented primers in raw data.

(B) Zoom-in views around previously identified pNPC-specific interactions between the Sox2 gene and NPC-specific enhancers. The upper row shows the interaction profile in pNPCs, while the lower row shows the same window in ES cells.

(C) Heatmaps showing GC content bias profiles before and after bias factor mitigation. The color indicates the average relative enrichment of ligation detection events as a function of the GC content of the primer designed to one of the participating fragments (x axis) and that of the other fragment (y axis).

(legend continued on next page)

make inaccurate assumptions regarding the mathematical relationships among known bias factors.

We next compared explicit modeling to matrix balancing algorithms that implicitly correct for biases without defining their specific sources (Imakaev et al., 2012; Knight and Ruiz, 2013; Rao et al., 2014; Sauria et al., 2015). Matrix balancing algorithms have been effectively applied to Hi-C data (Crane et al., 2015; Imakaev et al., 2012; Rao et al., 2014) and depend on the assumptions that (1) all fragments throughout the genome have ''equal visibility'' (i.e., equal propensity for detection via a proximity ligation assay), and (2) the intrinsic fragment-specific biases can be represented as a single scalar value for each fragment that interacts multiplicatively with the intrinsic biases of its ligation partners. An open unanswered question is whether these assumptions apply to 5C data given that the genomic regions are relatively small (1–10 Mb) and that the biases may follow nonlinear relationships. We first applied ICED matrix balancing to the conditional quantile normalized 5C counts. We observed that lines in the heatmaps are strongly attenuated while preserving looping interactions (Balanced; Figures 3A, 3B, and S2A). Notably, the smooth heatmaps were achieved despite minimal effect on the residual fragment length bias and only partial reduction in GC content bias (Balanced; Figures 3C and S2B). We also investigated the Express matrix balancing algorithm, which additionally incorporates information from a distance-dependent expected model in the computation of the bias factors (Sauria et al., 2015). We found that Express provided the most complete correction of GC content bias, fragment length bias, and lines in heatmaps, but it is overly harsh in attenuating real biological looping interactions (Express; Figures 3A–3C, S2A, and S2B). Consistent with these results, Express resulted in marked attenuation of the known enrichment of CTCF at the base of the strongest interacting fragment-fragment ligation junctions (Figure 3D). These observations support the hypothesis that at least some of the signal contributing to ''lines'' observed in the raw heatmaps may be biologically meaningful rather than simple technical bias. Together, these data indicate that currently available methods fall short of the goal of minimizing fragment biases while preserving bona fide looping interactions.

We created a variant of matrix balancing called Joint Express to account for Express' over-smoothing while retaining its ability to attenuate bias factors. The canonical Express algorithm computes a unique bias vector for each replicate. Our new Joint Express variant computes a single bias vector by integrating information from all replicates (see STAR Methods). Joint Express relies on the assumptions that (1) variation in technical bias is negligible among the conditional quantile normalized replicates, and (2) any observed bias vector differences correspond to biologically meaningful differences in cell-type-specific interaction frequency (i.e., extrusion lines, loops). We see that Joint Express preserves known biological looping interactions and smooths lines in heatmaps to a degree similar to ICED (Joint Express; Figures 3A, 3B, and S2A) while providing slight improvements in the smoothing of GC content and fragment length biases (Joint Express; Figures 3C and S2B). We also observe that Joint Express improves CTCF enrichment at high interaction frequency ligation junctions compared to Express (Figure 3D). Finally, counts processed using the Joint Express algorithm retained a greater degree of cell-type-specific signal, as evidenced by weaker inter-cell type correlations than those processed by canonical Express or ICED (Figure 3E). Altogether, these data indicate that traditional matrix balancing by ICED and our new variant on the Express algorithm, Joint Express, represent the highest-performing methods for removing technical fragment-specific biases from 5C data while preserving biologically important interactions.

Hi-C data are typically binned using non-overlapping windows of a pre-determined width and summing the detected ligation events within each window. The counts in a specific pixel from a binned 5C matrix can be interpreted as the relative interaction frequency between the genomic segments represented by the two anchoring bins across a population of cells. The 5C data re-analyzed here are particularly susceptible to spatial noise because of the use of a single alternating primer design that queries only a subset of all possible ligation junctions. Thus, instead of using non-overlapping bins tiling the genomic region of interest, we employ a sliding window strategy where the bin step size between successive evaluations of the window's interaction frequency is smaller than the bin window size. Region-wide heatmaps binned with a bin window size of 4 kb and bin step size of 4 kb (corresponding to no overlap between adjacent windows) exhibit a high degree of spatial noise, including many segments where no data are available (Figure 4A). By visually inspecting zoomed-in heatmaps, we found discontinuous and noisy interaction signal at a known loop connecting the Sox2 gene with a putative NPC-specific enhancer marked by NPC-specific H3K27ac (Figure 4B). We increased the bin window size to 16 kb while keeping the same 4 kb bin step size and observed markedly reduced spatial noise and clear emergence of punctate loops. Noteworthy, large (100 kb+) bin window sizes markedly attenuate spatial noise and highlight TAD/subTAD structure but at the expense of severe over-smoothing (and in many cases complete loss) of looping interactions (Figure S3B). Consistent with these findings, increasing the bin window size results in a decreased number of bin-bin pairs with strong interactions (Figure 4C) and an inverse relationship with the spatial variance, or noise, observed in the binned contact matrix (Figure 4D; STAR Methods). Together, these data suggest that by adjusting the bin window size, investigators can strike a balance between spatial noise in the 5C contact matrices and the ability to detect fine-grained architectural features at high resolution.

The binned interaction counts (now corrected for sequencing depth, library complexity, intrinsic fragment-specific biases, and spatial noise) exhibit a genomic distance-dependent

(D) Similar to (C), but stratifying fragment-fragment ligation junctions according to CTCF ChIP-seq signal over the fragment (computed as the average number of reads per million mapping to a 4 kb window centered at the fragment midpoint). In this case, enrichment for contacts between CTCF-rich fragments (more orange or red in the upper right of the bias heatmap) is expected, consistent with CTCF's role as an architectural protein.

(E) Tables showing pairwise Spearman's correlation coefficients between replicates after ICED matrix balancing (left), after Express normalization (center), and after Joint Express normalization (right).

See also Figure S2.

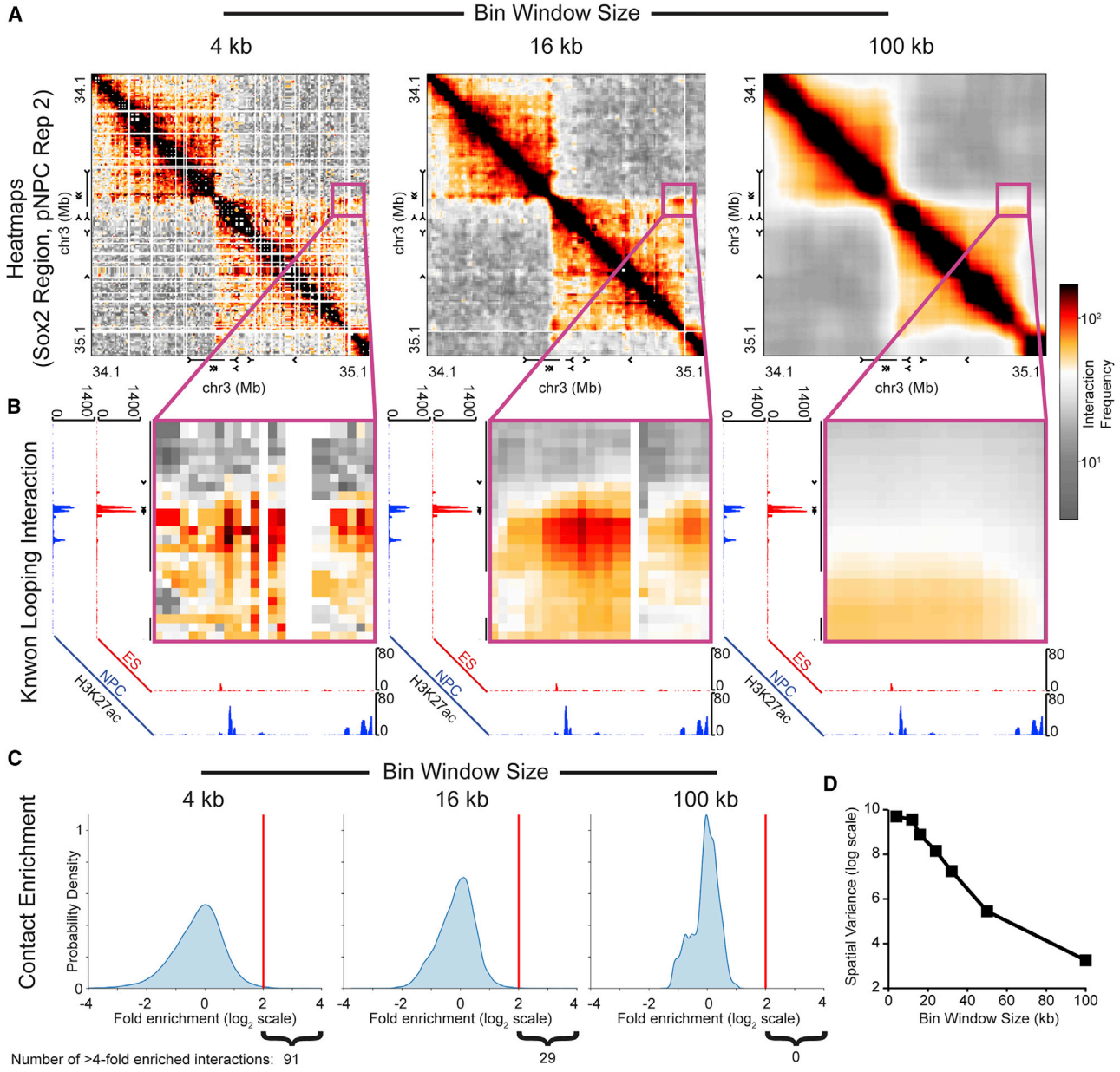


Figure 4. Effects of Binning and Smoothing on 5C Contact Matrices

(A) Contact frequency heatmaps of the Sox2 region in pNPCs, binned at 4 kb bin step size with a variety of bin window sizes (from left to right: 4 kb, 16 kb, 100 kb).

(B) Zoom-in view around previously identified interaction between the Sox2 gene and an NPC-specific enhancer. ChIP-seq tracks show H3K27ac signal in ES cells and NPCs.

(C) Distributions of the log-fold enrichment of smoothed values over an empirical one-dimensional distance-dependent expected value across the Sox2 region. The curly braces indicate the number of bin-bin pairs with smoothed values greater than four times the expected value for their interaction distance. Absence of points showing pixel-wise contact enrichment suggests that fine scale features such as looping interactions may have been smoothed away during binning.

(D) Spatial variance of the contact matrix plotted as a function of bin window size.

See also Figure S3.

interaction signal (Lieberman-Aiden et al., 2009) that must be modeled before detecting loops (Rao et al., 2014). This so-called distance-dependent background is made manifest as a strong band of high interaction counts along the diagonal of genome folding heatmaps (Figure 5A). We first modeled the changes in expected interaction frequency between two loci as a function of the linear genomic distance (Figure 5B; STAR Methods). We compared the average number of empirically observed counts at each distance scale to a locally weighted scatterplot smoothing (LOWESS) distance dependence model and found them to be similar, with the regional empirical model becoming noisy at longer interaction distances where there are fewer

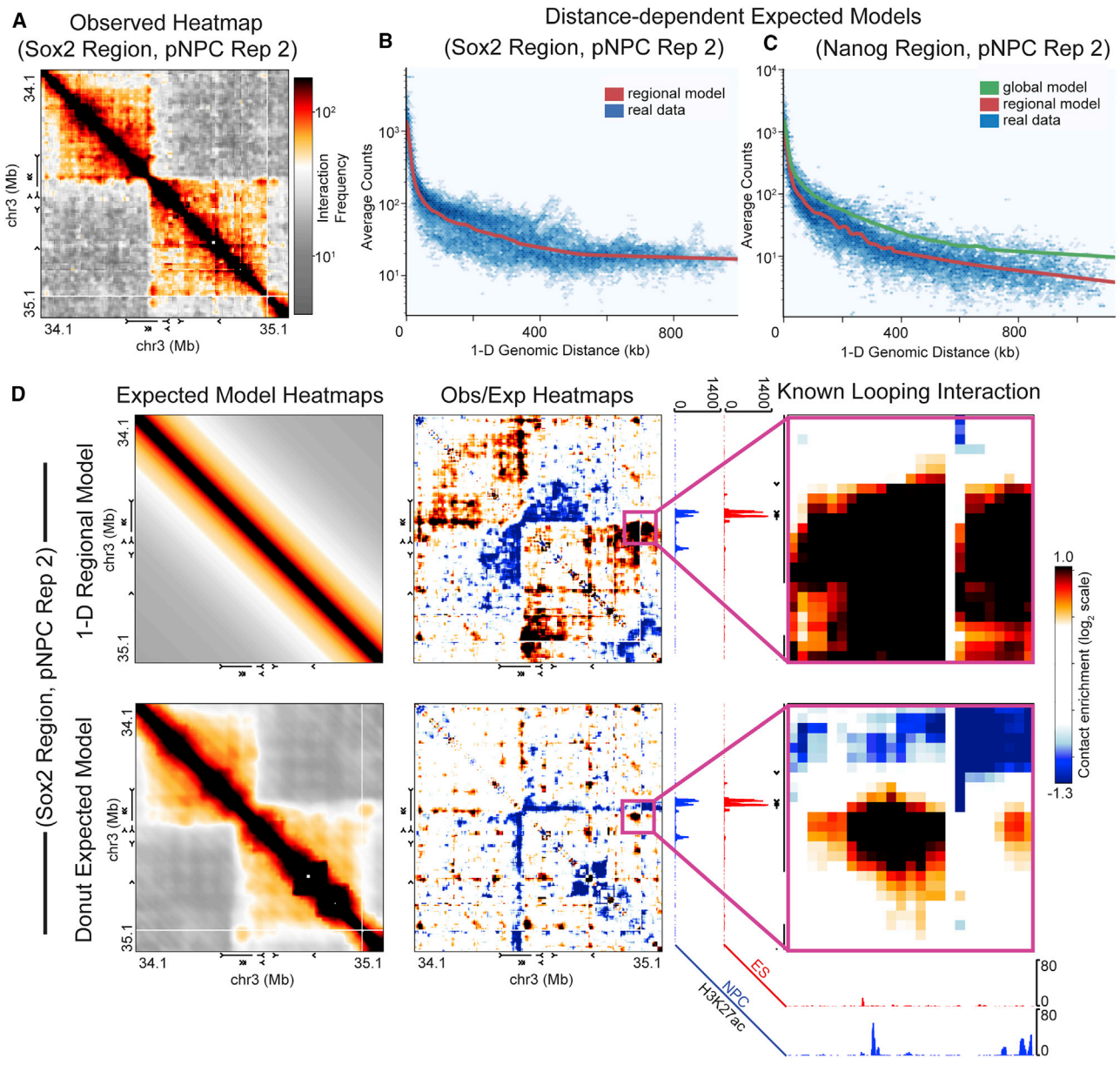


Figure 5. Strategies for Modeling the Expected 5C Counts at Each Genomic Distance Scale

(A) Heatmap visualization of the smoothed contact matrix for the Sox2 region in pNPC Rep 2.

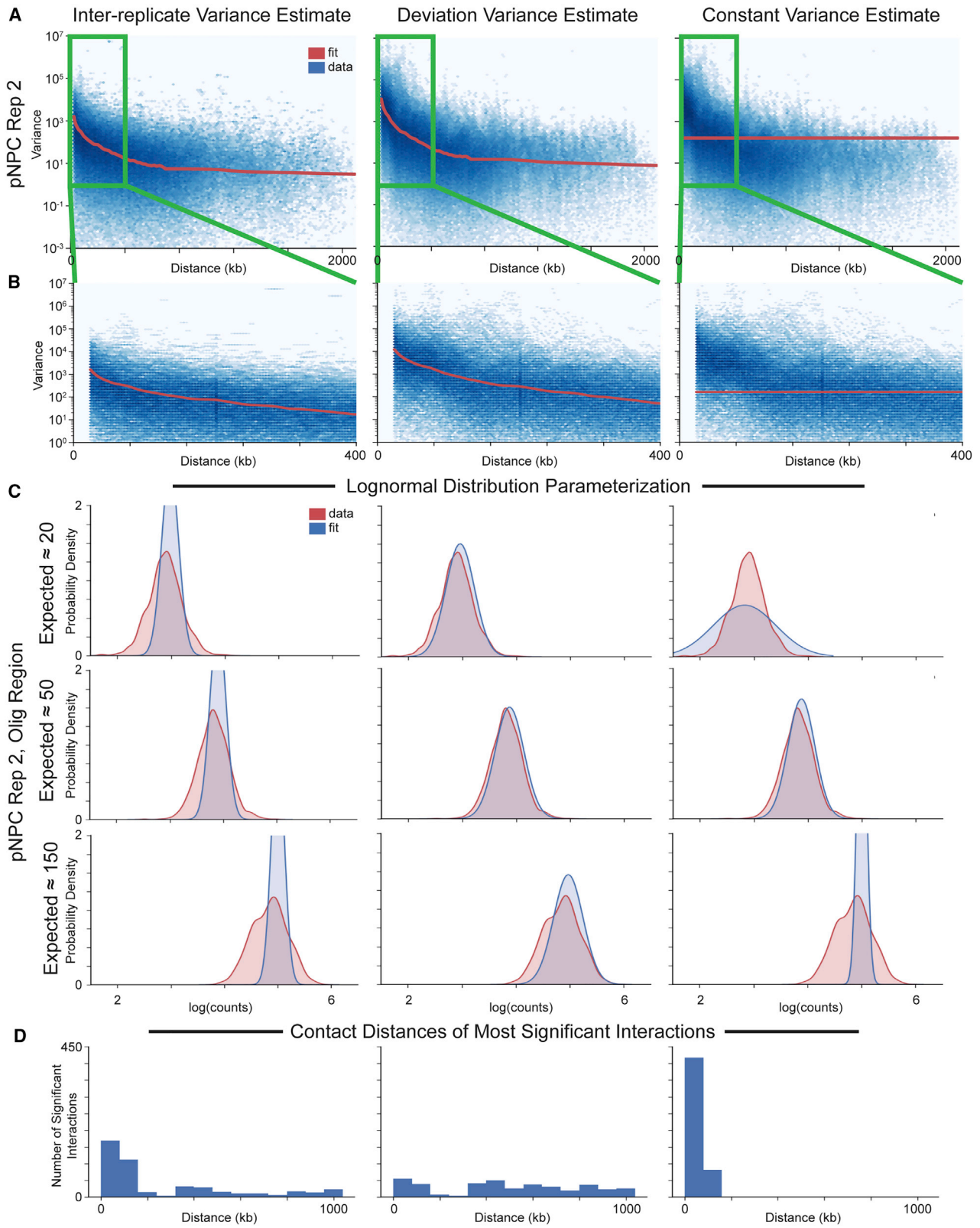
(B) Illustration of expected modeling procedure. The density of entries in the smoothed contact matrix (pixels in the contact frequency heatmap in A) is shown in blue hexagonal bins. Distance-dependent expected models attempt to fit a function (red curve) through the data (blue points).

(C) Smoothed contact matrix entries from the Nanog region in pNPC Rep 2 (blue points) compared to two different log-counts lowess distance-dependent expected models: one fitted only to the data from the Nanog region (red curve), and one fitted to data from all 5C regions in the dataset (green curve). The region-specific model follows the observed data more closely.

(D) The rows compare a log-counts lowess one-dimensional distance-dependent expected model (top row) to a donut expected model (bottom row). The left column shows a contact frequency heatmap visualization of the two models at the Sox2 region. The center column shows log-fold enrichments of the smoothed contact matrix entries over the expected models. The right column shows a zoom-in view around a previously identified interaction between the Sox2 gene and an NPC-specific enhancer. ChIP-seq tracks show H3K27ac signal in ES cells and NPCs. Use of the donut expected model recovers a more punctate looping interaction in the corner of this contact domain.

See also Figure S4.

bin-bin pairs (Figure S4A; STAR Methods). A critical decision in modeling the distance-dependent expected interaction frequency is whether a single, global fit to all the data or a 5C region-specific fit should be applied. We empirically observed that different 5C regions exhibit dramatically different distance-dependent counts relationships (Figure S4B). For example, at



(legend on next page)

the Nanog locus, the global expected model (estimated using a log-counts lowess fit, detailed in STAR Methods) overestimates the average counts at every distance scale (Figure 5C). Thus, the regional outperforms the global distance-dependence expected model for correcting the 5C diagonal prior to loop calling.

In addition to the distance-dependence relationship, the local TAD/subTAD architecture strongly influences contact frequencies observed in 5C datasets. Two loci in the same TAD/subTAD tend to interact more frequently than a pair of loci that span a domain boundary, even when these pairs of loci are separated by the same linear genomic distance (Dixon et al., 2012). It is essential to include TAD/subTAD structure in the expectation when detecting loops. To model the expected counts due to chromatin domains, we applied the donut expected filter proposed by Aiden and colleagues (Rao et al., 2014). The donut filter provides a model of the distance-dependent background and local domain structure for each region without having to know the location of TADs/subTADs *a priori* (Figure S4C; STAR Methods). We observed punctate looping interactions upon correction of the binned 5C counts for the donut expected, whereas relying exclusively on a one-dimensional distance model tends to lead to smearing of punctate looping pixels across the corners of TADs/subTADs (Figure 5C). Overall, we see that the donut provides a more rigorous and accurate expected model for calling looping interactions in 5C data than the one-dimensional distance dependence relationship.

We next investigated methods for estimating the variance under the null model that any given pixel is not engaged in a bona fide looping interaction. More specifically, our null hypothesis is that the observed interaction frequency at a given pixel is not significantly higher than the expected value at that pixel. To parameterize a distribution of possible counts values under the null distribution for each pixel, it is necessary to understand the relationship between the mean and the variance. In a recent kilobase-resolution Hi-C dataset (Rao et al., 2014), loops were called by applying a Poisson distribution to the raw interaction counts for each bin-bin pair. The Poisson distribution was not applicable to our 5C data because we converted raw counts from a discrete to a continuous random variable due to the normalization steps and use of geometric mean binning. Therefore, we sought alternative approaches for variance estimation.

We first computed the sample variance across two biological replicates at each pixel under a lognormal model to obtain an independent variance estimate for each pixel (STAR Methods). Individual variance estimates based only on two replicates exhibited a high degree of noise. However, we observed a strong relationship between genomic interaction distance (or equivalently, expected value, which trends strongly with distance) and variance (left panel; Figures 6A and 6B). We therefore fitted a curve to the trend between variance and genomic distance using LOWESS to obtain a DVR. We then used the estimated variance values from the inter-replicate DVR to parameterize the lognormal distribution across the full range of possible expected values. When we compared the parameterized distributions to the empirical distributions of observed values with similar expected values, we found that our inter-replicate DVR was consistently underestimating the variance (left panels; Figure 6C). Therefore, we reasoned that there must be an additional contributor to the variance in our 5C data beyond the inter-replicate variance.

We next computed an intra-replicate variance using deviations between the observed and expected values within each replicate (detailed in STAR Methods). The resulting variance estimates also trended strongly with genomic distance (middle panels; Figures 6A and 6B), suggesting that the DVR is a general property of 5C data and is not specific to particular methods of variance estimation. Lognormal distributions parameterized using this deviation-based DVR provided better fits to the empirical distribution of observed values (middle panels; Figure 6C). We also fit a constant variance model to the intra-replicate variation data (right panels, Figures 6A and 6B). The constant, distance-invariant intra-replicate estimates systematically underestimates the variance at short distances (high expected values) and overestimates it at long distances (low expected values) (right panels; Figure 6C). Thus, among the options we pursued in this manuscript, an intra-replicate DVR provided the closest resemblance to the underlying data.

To better understand our three variance estimates, we quantified the significance of each bin-bin pair by computing right-tail p values for each observed value using a lognormal distribution parameterized using the expected value from the donut expected model and the variance value from the DVR (detailed in STAR Methods). We visualized the genomic length scale for the top 500 most significant bin-bin pairs (Figure 6D). The constant intra-replicate variance estimate resulted in loops biased to interaction distances of 0–160 kb (with over 80% lying within 80 kb), consistent with the observation that this approach underestimates 5C signal variation at short interaction distances (right panel, Figure 6D). The inter-replicate DVR resulted in loops biased to short length scales (left panel, Figure 6D). By contrast, the intra-replicate deviation DVR resulted in a uniform distribution of interaction distances, suggesting it performs well at minimizing bias toward loops at specific length scales (middle panel; Figure 6D). Overall, these results reveal the presence of a strong relationship between genomic interaction distance and variance in 5C data, which can be modeled to identify interactions across all distance scales in an unbiased manner.

To assess the potential biological relevance of the detected loops, we clustered pixels into significant looping interactions and classified them by their cell-type-specific or -invariant

Figure 6. Strategies for Modeling the Distance-Variance Relationship in 5C Data

(A) The blue hexagonal bins show the joint distribution of genomic interaction distance and variance for every bin-bin pair. In the left column ("Inter-replicate Variance"), the variance is estimated using the variance across two biological replicates. In the middle and right columns ("Deviation Variance" and "Constant Variance"), the variance is estimated based on deviations between the observed and expected values. The red curves show variance trends fitted to the data.

(B) Same data as in (A), but focusing on shorter interaction distances, where variance changes most rapidly.

(C) Lognormal distributions (blue) parameterized using the variance estimates from the fitted curves above, overlaid with the empirical distribution of observed values near similar expected values (red). The three rows highlight different expected values and therefore different distance scales.

(D) Histograms showing the distribution of genomic interaction distances for the top 500 most significant bin-bin pairs according to each variance model.

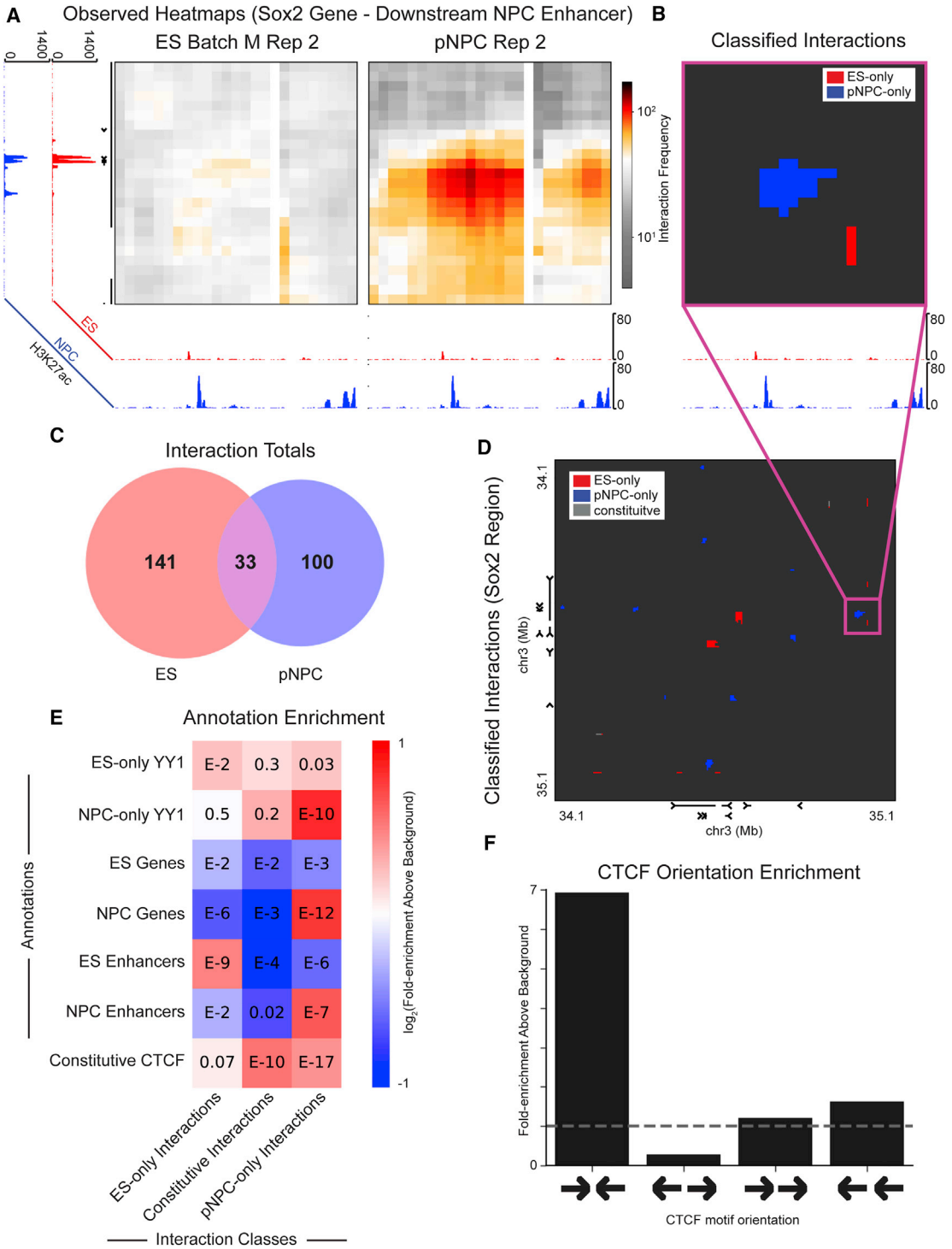


Figure 7. Characterization of Statistically Significant Interactions for Enrichment of Known Epigenetic Marks

(A) Zoom-in view around a previously identified interaction between the Sox2 gene and an NPC-specific enhancer. ChIP-seq tracks show H3K27ac signal in ES cells and NPCs.

(B) Same view as in (A) but colored according to cell-type specificity of significant interactions.

(C) Venn diagram illustrating total numbers of interactions called in each cell type.

(D) Whole-region view of classified interactions throughout the Sox2 region.

(legend continued on next page)

properties. We performed this analysis using the pipeline of (1) conditional quantile normalized raw counts, (2) joint express matrix balancing, (3) binning and smoothing with a 4 kb bin step size and 16 kb bin window size, (4) modeling the expected value using the donut expected, and (5) modeling variance using the intra-replicate deviation-based DVR. Starting from matrices of p values for each replicate, we performed Benjamini-Hochberg false discovery rate (FDR) control at 10% FDR. Because multiple adjacent significant pixels often represent one underlying biological interaction, the remaining pixels that passed multiple testing correction were clustered into groups of adjacent pixels. Clusters with less than three significant pixels and clusters only present in one of the two biological replicates were removed from analysis (STAR Methods). As a representative example, we visualized the observed heatmaps for both ES and pNPC cell types in the vicinity of a previously identified NPC-specific interaction between the Sox2 gene and a downstream NPC-specific enhancer (Figure 7A) and compared it to the clustered loop calls (Figure 7B; additional loci highlighted in Figures S5A–S5D). All together, we identified 141 ES-specific, 100 pNPC-specific, and 33 constitutive interactions (Figure 7C) distributed across the 5C regions (Figure 7D).

We also confirmed that our final loop calls exhibited the expected enrichment for cell-type-specific annotations on the linear Epigenome. Consistent with previously published observations (Beagan et al., 2017; Beagan et al., 2016), ES-specific loops were enriched for putative ES-specific enhancers, constitutive and pNPC-specific loops were enriched for constitutive CTCF, and pNPC-specific loops were enriched for pNPC-specific YY1 and putative pNPC-specific enhancers (Figure 7E). Finally, we evaluated the enrichment of motifs occupied by CTCF in the pNPC condition with different orientations under loops called as significant in the pNPC condition (Figure 7F; ES-specific loops in Figure S5E). We observed a strong enrichment for convergently oriented CTCF motifs anchoring the base of looping interactions previously reported (Rao et al., 2014). Thus, our selected analysis conditions at each stage in the 5C pipeline resulted in loops anchored by known epigenetic modifications and architectural protein-binding sites.

Finally, we compared the loops called by different variations of our 5C pipeline to Hi-C data. We observed that certain variations to the 5C pipeline had a significant impact on the results (e.g., smoothing versus not smoothing; or donut versus the one-dimensional distance-dependent expected) while others seemed to have a less pronounced effect (e.g., reversing the order of the binning and balancing steps; or applying different balancing methods) (Figure S6). Overall, our chosen combination of 5C pipeline analysis conditions resulted in many loops that are readily visible in Hi-C data and strongly enriched for previously reported cell-type-specific annotations on the linear Epigenome.

DISCUSSION

Recovering bona fide, biologically relevant looping interactions from 5C data is a challenging problem, requiring multiple steps to account for sequencing depth and library complexity, primer- and fragment-specific biases, spatial noise, distance-dependent background signal, region- and domain-specific contact frequency effects, and statistical variance estimation. Each of these steps can be addressed with a wide variety of proposed approaches, the appropriate selection of which can often be critical to the success of the loop identification endeavor. Here, we provide a systematic analysis of available methods for analyzing loops in 5C data. We introduce several novel approaches and algorithm variants specifically for 5C data, including: (1) a variant of conditional quantile normalization, (2) the Joint Express balancing algorithm, and (3) a new approach for modeling the variance using a DVR. We systematically assess the performance of each analysis procedure at each stage and create a pipeline that is capable of identifying known reported looping interactions. The looping interactions called with a pipeline of our selected high-performance analysis conditions exhibit strong enrichment of known epigenetic modifications and architectural protein-binding sites at loop anchors and are comparable in many cases to loops observed in Hi-C data. Future application of algorithms discussed in this manuscript to more recently published double alternating 5C libraries will result in loop calls with high concordance with Hi-C data on a targeted subset of genomic loci at a significantly attenuated cost.

Decisions made in the 5C analysis process involve critical trade-offs. Bin size can be decreased to improve matrix resolution, but with insufficient read depth this can increase spatial noise. Fragment-specific biases can be corrected via matrix balancing, but when the assumption of equal visibility does not apply, as is often the case in 5C, this can result in the loss of biologically important structures such as loop extrusion lines. If not appropriately addressed, library complexity, sequencing depth, and batch effect differences across libraries can lead to false-positive classification of bias as a cell-type-specific loop. Without careful modeling of the local domain structure and distance-dependence interactions, a large number of pixels can be mistakenly classified as loops when they are simply interactions due to TADs/subTADs. Thus, a thorough analysis and characterization of computational methodologies at each stage in the process of 5C data processing is essential for high-quality annotation of bona fide looping interactions.

One area we leave open to further exploration is a comprehensive understanding of the differences between 5C and Hi-C datasets. While we offer a brief comparison suggesting that loops visible in ultra-resolution Hi-C datasets can be accurately called from 5C data by our best computational

(E) Heatmap showing log-scale fold-enrichment (a log fold enrichment of zero, indicated by a white color, represents no enrichment above background) of selected genomic annotations (rows) within interaction classes (columns) relative to background bin-bin pairs. The numbers on the heatmap grid represent p values for the enrichment, computed using Fisher's exact test on a two-by-two contingency table (STAR Methods).
(F) Enrichment above background of motif orientations of CTCF sites occupied in NPCs found at the base of significant interactions identified in pNPCs. The most-enriched orientation of occupied CTCF sites is the convergent orientation, consistent with previous reports.
See also Figures S5 and S6.

pipeline, this by no means suggests that the underlying data generated from single alternating primer designs are equivalent. Further improvements to the older 5C data analyzed here have already been published, such as the use of *in situ* ligation and a double alternating primer design, and represent an important, ongoing area of investigation (Hnisz et al., 2016; Kim et al., 2018).

Here, we have highlighted several important trade-offs and lessons to be learned in 5C data analysis, including (1) the benefits of using a local donut expected model to attenuate false-positive loop calls due to local TAD/subTAD structure; (2) the importance of normalizing raw data to correct for batch effects, sequencing depth, and technical library complexity differences before calling loops; (3) the elucidation and preliminary modeling of a DVR in 5C data; and (4) the possibility that the assumption of equal visibility does not apply to 5C data, suggesting that canonical matrix balancing should be used with caution to avoid normalizing bona fide looping interactions. We note that the estimation of a precise and accurate DVR remains an important area for future inquiry. Moreover, the task of classifying differential looping interactions is only discussed briefly here and remains an exciting area for future work. We provide the coding package lib5C to allow investigators to assess the effects of the trade-offs discussed here on their own novel 5C datasets. We expect that the role for specific analysis steps and their parameters in loop identification will remain an important topic of future inquiry given the relative scarcity of universally accepted loop calls and loop calling algorithms in 3C-based data.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - Notation
  - Data Sources and Primer Quality Filtering
  - Sequencing depth and library complexity correction strategies
  - Quantile Normalization
  - Bias Factor Mitigation Strategies
  - Explicit Normalization: Fragment-Level Distance Dependence Model
  - Bias Factor Heatmaps
  - Binning and Smoothing
  - Spatial Noise Quantification
  - Expected Modeling Strategies
  - Variance Modeling Strategies
  - Distribution Parameterization and P-Value Computation
  - Interaction Classification
  - Enrichments
  - Convergency Analysis
  - Hi-C Data Comparison
  - H3K27ac ChIP-seq Track Processing
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2019.02.006.

ACKNOWLEDGMENTS

J.E.P.-C. is a New York Stem Cell Foundation (NYSCF) 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 (1DP2MH11024701; J.E.P.-C.), a 4D Nucleome Common Fund grant (1U01HL12999801; J.E.P.-C.), a joint NSF-NIGMS grant to support research at the interface of the biological and mathematical sciences (1562665; J.E.P.-C.), and an NIH training grant (5T32HL007954-18; T.G.G.).

AUTHOR CONTRIBUTIONS

T.G.G. and J.E.P.-C. designed research, performed research, analyzed data, and wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 11, 2017
Revised: October 19, 2018
Accepted: February 19, 2019
Published: March 20, 2019

WEB RESOURCES

Creminslab / lib5c — Bitbucket, https://bitbucket.org/creminslab/lib5c/

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Beagan, J.A., Duong, M.T., Titus, K.R., Zhou, L.D., Cao, Z.D., Ma, J.J., Lachanski, C.V., Gillis, D.R., and Phillips-Cremins, J.E. (2017). YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res. *27*, 1139–1152.

Beagan, J.A., Gilgenast, T.G., Kim, J., Plona, Z., Norton, H.K., Hu, G., Hsu, S.C., Shields, E.J., Lyu, X.W., Apostolou, E., et al. (2016). Local genome topology can exhibit an incompletely rewired 3D-folding state during somatic cell reprogramming. Cell Stem Cell *18*, 611–624.

Bonev, B., Mendelson Cohen, N.M., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X.L., Lv, X.D., Hugnot, J.P., Tanay, A., et al. (2017). Multiscale 3D genome rewiring during mouse neural development. Cell *171*, 557–572.e24.

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

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. USA *107*, 21931–21936.

Daley, T., and Smith, A.D. (2013). Predicting the molecular complexity of sequencing libraries. Nat. Methods *10*, 325–327.

Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'Shea, C.C., Park, P.J., Ren, B., et al. (2017). The 4D nucleome project. Nature *549*, 219–226.

Deng, W., and Blobel, G.A. (2014). Manipulating nuclear architecture. Curr. Opin. Genet. Dev. *25*, 1–7.

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.

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.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. *3*, 95–98.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of chromosomal domains by loop extrusion. Cell Rep. *15*, 2038–2049.

Hansen, K.D., Irizarry, R.A., and Wu, Z.J. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics *13*, 204–216.

Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods *9*, 999–1003.

Jhunjhunwala, S., van Zelm, M.C., Peak, M.M., and Murre, C. (2009). Chromatin architecture and the generation of antigen receptor diversity. Cell *138*, 435–448.

Jin, F.L., 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.

Kim, J.H., Titus, K.R., Gong, W.F., Beagan, J.A., Cao, Z.D., and Phillips-Cremins, J.E. (2018). 5C-ID: increased resolution chromosome-conformation-capture-carbon copy with in situ 3C and double alternating primer design. Methods *142*, 39–46.

Knight, P.A., and Ruiz, D. (2013). A fast algorithm for matrix balancing. IMA J. Numer. Anal. *33*, 1029–1047.

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

Li, G., Fullwood, M.J., Xu, H., Mulawadi, F.H., Velkov, S., Vega, V., Ariyaratne, P.N., Bin Mohamed, Y., Ooi, H.S., Tennakoon, C., et al. (2010). ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol. *11*, R22.

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.

Marinov, G.K., Kundaje, A., Park, P.J., and Wold, B.J. (2014). Large-scale quality analysis of published ChIP-seq data. G3 (Bethesda) *4*, 209–223.

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.

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.

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

Rhind, N., and Gilbert, D.M. (2013). DNA replication timing. Cold Spring Harb. Perspect. Biol. *5*, a010132.

Sanborn, A.L., Rao, S.S.P., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. USA *112*, E6456–E6465.

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

Sauria, M.E.G., Phillips-Cremins, J.E., Corces, V.G., and Taylor, J. (2015). HiFive: a tool suite for easy and efficient HiC and 5C data analysis. Genome Biol. *16*, 237.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. *15*, 121–132.

Tang, Z.H., Luo, O.J., Li, X.W., Zheng, M.Z., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., et al. (2015). CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell *163*, 1611–1627.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. *43*, 1059–1065.

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

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Bowtie	Langmead et al., 2009	http://bowtie-bio.sourceforge.net/index.shtml
MACS2	Zhang et al., 2008	https://github.com/taoliu/MACS
ICED	Imakaev et al., 2012	https://github.com/hiclib/iced
lib5c	This paper	https://bitbucket.org/creminslab/lib5c
HiC-Pro	Servant et al., 2015	https://github.com/nservant/HiC-Pro
Juicer	Durand et al., 2016	https://github.com/theaidenlab/juicer

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jennifer E. Phillips-Cremins (jcremins@seas.upenn.edu).

METHOD DETAILS

Notation

We will assume we have N_K 5C libraries, indexed with the set K={1,…,N_K}. Each of the 5C libraries queries N_R regions, indexed with the set R={1,…,N_R}. Any given region $r \in R$ is queried by $N_{P,r}$ primers, indexed with the set P_r={1, …,$N_{P,r}$}. Since the 5C assay queries ligations between primers, the set of all primer-primer junctions in a given region $r \in R$ is therefore $J_r = \{\{a, b\} : a \in P_r, b \in P_r\}$, where we note that a junction is identified by an unordered set of the two primers involved in the junction. Not all primer-primer junctions are physically possible, because two primers with the same orientation cannot be ligated together (i.e., 3' to 3' ligations and 5' to 5' ligations are impossible). Therefore, we construct a similar set to represent all primer-primer junctions actually queried by the assay in a given region $r \in R$; this set is $Q_r = \{\{a, b\} : a \in P_r,\ b \in P_r,\ O_{r,a} \neq O_{r,b}\}$, where a and b represent the indices of the two primers in a given junction and $O_{r,a}$ denotes the orientation of primer $a \in P_r$ in region $r \in R$. We will sometimes wish to consider the set of queried primer-primer junctions in a given region $r \in R$ which involve a specific primer $p \in P_r$. We denote this set $Q_{r,p} = \{\{p, b\} : b \in P_r,\ O_{r,p} \neq O_{r,b}\}$.

In exploring intrinsic fragment properties, we construct subsets of the primers according to the GC content of their genome-binding subsequences (i.e., excluding universal tails). We can denote the set of primers in a given region $r \in R$ whose GC content percentage is u as $P_r^u = \{p \in P_r : GC_{r,p} = u\}$ where $GC_{r,p}$ is the GC content percentage of primer $p \in P_r$ in region $r \in R$. We can use these primer subsets to identify primer-primer junctions whose involved primers have specific GC content levels. In particular, we can denote the set of primer-primer junctions in a given region $r \in R$ involving one primer with GC content percentage u and another primer with GC content percentage v as $Q_r^{u,v} = \{\{a,b\} : a \in P_r^u,\ b \in P_r^v,\ O_{r,a} \neq O_{r,b}\}$.

When exploring the linear genomic distance separating fragments, we annotate the midpoint of fragment f that a primer $p \in P_r$ in a given region $r \in R$ is designed to query by $M_{p,r}^f$. We annotate the fragments start coordinate as $S_{p,r}^f$ and its end coordinate as $E_{p,r}^f$. For the purposes of visualizing fragment level contacts, we collect all the 5' oriented primers in each region in to an ordered set $P_r^{5'} = \{p \in P_r : O_{r,p} = 5'\}$ and all the 3' oriented primers in another ordered set $P_r^{3'} = \{p \in P_r : O_{r,p} = 3'\}$. We then constructed a matrix with dimensions $\left|P_r^{5'}\right| \times \left|P_r^{3'}\right|$ whose i,j th entry represents the count of the ligation junction between the i th primer in $P_r^{5'}$ and the j th primer in $P_r^{3'}$. We were then able to visualize this matrix as a contact frequency heatmap.

Data Sources and Primer Quality Filtering

We obtained previously published source data from GEO Series GSE68582 (Beagan et al., 2016; Servant et al., 2015). For each of the samples listed in the table below, we obtained the raw countsfile available as a Supplementary File on that sample. We also obtained information about the 5C primers used for these samples from GSE68582_BED_ES-NPC-iPS-LOCI_mm9.bed.gz, a Supplementary File on the GSE68582 series.

GEO Sample	Replicate Name Used in this Work
GSM1974095	ES Batch M Rep 1
GSM1974096	ES Batch M Rep 2
GSM1974099	pNPC Rep 1
GSM1974100	pNPC Rep 2

Before beginning analysis, we removed any 5C primers which had fewer than 100 total *cis* read counts in any sample, along with those for which less than half of the possible *cis* ligation products were ever detected in any single replicate.

Sequencing depth and library complexity correction strategies

Simple Scalar Normalization

The total number of *cis* contact detections within a given region $r \in R$ in each library $k \in K$ are computed as Equation 1:

$$s_{k,r} = \sum_{q \in Q_r} C_{q,k,r} \qquad \text{(Equation 1)}$$

where $C_{q,k,r}$ is the number of reads counted mapping to primer-primer junction $q \in Q_r$ in region $r \in R$ and library $k \in K$. For each region $r \in R$, we then identify the smallest size factor across the libraries $s_{min,r} = \min_{k \in K} s_{k,r}$ The scaled count, $\widehat{C}_{q,k,r}$, for primer-primer junction $q \in Q_r$ in library $k \in K$ and region $r \in R$ is then computed as Equation 2:

$$\widehat{C}_{q,k,r} = C_{q,k,r} \times \frac{s_{min,r}}{s_{k,r}} \qquad \forall q \in Q_r, k \in K,\ r \in R. \qquad \text{(Equation 2)}$$

Median of Ratios Normalization

The median-of-ratios size correction factor (Anders and Huber, 2010) $s_{k,r}$ for a given region $r \in R$ and library $k \in K$ is computed as Equation 3:

$$s_{k,r} = \underset{q \in Q_r}{\text{median}} \frac{C_{q,k,r}}{\left(\prod_{k' \in K} C_{q,k',r}\right)^{1/N_K}} \qquad \forall k \in K, r \in R. \qquad \text{(Equation 3)}$$

The scaled counts value for primer-primer junction $q \in Q_r$ in region $r \in R$ and library $k \in K$ is then computed as Equation 4:

$$\widehat{C}_{q,k,r} = \frac{C_{q,k,r}}{s_{k,r}} \qquad \forall q \in Q_r, k \in K. \qquad \text{(Equation 4)}$$

Quantile Normalization

To perform quantile normalization, we created a table whose rows represent individual primer-primer ligations $q \in Q_r$ and whose columns represent individual libraries $k \in K$. The count values $C_{q,k,r}$ are entered into the table at the appropriate positions, and the $C_{q,k,r}$ are then sorted column-wise in each k column. Thus, the i th row of the table corresponds to the i th order count value, $\{C_{q,k,r} : q \in Q_r\}_{(i)}$, with the column index k specifying which library this order statistic is taken over. Next, we compute the row-wise average across the columns (i.e. across libraries) of the sorted table (Equation 5).

$$\left\{\widehat{C}_{q,k,r} \ : q \in Q_r\right\}_{(i)} = \frac{1}{N_K} \sum_{k' \in K} \left\{C_{q,k',r} \ : q \in Q_r\right\}_{(i)} \qquad \forall i \in \{1, \ldots, |Q_r|\}. \qquad \text{(Equation 5)}$$

Finally, we unscramble the table back to its original ordering, and read out the normalized count $\widehat{C}_{q,k,r}$ from row q, column k. If there is a tie in library $k \in K$ between a set of tied ranks $T = \{t_1, t_2, \ldots, t_n\}$, it is resolved by identifying the lowest of the tied ranks $t_{min} = \min_{t \in T} t$ and setting the normalized value for every rank in T in library k to the average across libraries at rank t_{min}, or in other words, $\{\widehat{C}_{q,k,r} : q \in Q_r\}_{(t)} = \frac{1}{N_K} \sum_{k' \in K} \{C_{q,k',r} \ : q \in Q_r\}_{(t_{min})}, \quad \forall t \in T.$

Conditional Quantile Normalization

Inspired by the previous application of conditional quantile normalization to RNA-seq data (Hansen et al., 2012), we devised a new conditional quantile normalization scheme for proximity ligation data by performing quantile normalization separately on stratified groups of fragment-fragment junctions with the same GC content properties. The normalized counts values $\widehat{C}$ for the group of fragment-fragment junctions involving one primer with GC content percentage u and another primer with GC content percentage v in a given region $r \in R$ are computed as Equation 6:

$$\left\{\widehat{C}_{q,k,r} : q \in Q_r^{u,v}\right\}_{(i)} = \frac{1}{N_K} \sum_{k' \in K} \left\{C_{q,k',r} : q \in Q_r^{u,v}\right\}_{(i)} \qquad \forall i \in \{1, \ldots, |Q_r|\}, \qquad \text{(Equation 6)}$$

where $\{C_{q,k,r} : q \in Q_r^{u,v}\}_{(i)}$ is the i th smallest non-redundant counts value in region $r \in R$ in library $k \in K$ among primer-primer junctions involving primers with GC content percentages u and v. A separate set $Q_r^{u,v}$ exists for each pair of GC content percentages (u,v) for which at least one queried fragment-fragment junction $q \in Q_r$ consisted of one fragment with GC content percentage u and another with GC content percentage v. If there is a tie in library $k \in K$ between a set of tied ranks $T = \{t_1, t_2, \ldots, t_n\}$, it is resolved by identifying the lowest of the tied ranks $t_{min} = \min_{t \in T} t$ and setting the normalized value for every rank in T in library k to the average across libraries at rank t_{min}, or in other words, $\{\widehat{C}_{q,k,r} : q \in Q_r^{u,v}\}_{(t)} = \frac{1}{N_K} \sum_{k' \in K} \{C_{q,k',r} : q \in Q_r^{u,v}\}_{(t_{min})}, \quad \forall t \in T.$

Bias Factor Mitigation Strategies

Explicit Normalization: Overview

We also developed a method for specifying bias factors explicitly. Previous reports have established that data from proximity ligation-based assays is biased by the GC content and mappability of the sequence near the ligation point as well as the length of the restriction fragments involved in the ligation (Yaffe and Tanay, 2011). In 5C data, mappability is not a prominent contributor to the bias profile; the 5C primers are always sequenced in their entirety and their sequences are unique. Thus, GC content and fragment length are the two major known bias factors that apply to 5C data.

We first constructed models of our bias factors, including (1) the estimated log-scale GC bias factor for a primer-primer ligation given the GC content of the primers involved in the ligation and (2) the estimated length bias factor given the lengths of the fragments involved in the ligation. We annotate the models (which model the degree to which interaction counts for a particular primer-primer ligation junction are over-represented as a function of its properties, as explained in further detail below) as $F_{GC}(GC_{r,a}, GC_{r,b})$ and $F_{len}(L_{r,a},L_{r,b})$, respectively, where $GC_{r,a}$ represents the GC content of primer $a \in P_r$ in region $r \in R$ and $L_{r,a}$ represents the length of the restriction fragment queried by primer $a \in P_r$ in region $r \in R$. Since distance dependence is a strong covariate in 5C data, we also use a model that accounts for the expected value of the primer-primer ligation counts given the genomic distance between the two fragments involved in the ligation. We create one distance dependence model for each library $k \in K$ and for each region $r \in R$, calling it $D_{r,k}\left(\left|M^f_{a,r} - M^f_{b,r}\right|\right)$, where $M^f_{a,r}$ represents the midpoint of the restriction fragment queried by primer $a \in P_r$ and $D_{r,k}\left(\left|M^f_{a,r} - M^f_{b,r}\right|\right)$ represents the expected value of the counts for a particular ligation junction given that the fragments involved in the junction are separated by a midpoint-to-midpoint distance of $\left|M^f_{a,r} - M^f_{b,r}\right|$ base pairs along the linear genome.

GC content and fragment length biases have been discussed as non-independent bias factors by previous reports (Yaffe and Tanay, 2011), though modeling them as fully independent can be challenging (Jin et al., 2013). Here we have chosen to partially address this problem by keeping our bias factor models $F_{GC}(GC_{r,a}, GC_{r,b})$ and $F_{len}(L_{r,a}, L_{r,b})$ nominally independent, but fitting them using an iterative procedure. The intuition for our iterative fitting procedure is that at each iteration step, we choose one bias model (GC or length) and fit it de novo (i.e., ignoring any parameters fitted for this model in earlier iterations) to data that have been adjusted by the distance dependence model as well as the latest version of the other bias model (see model Equations 10, 11, 12, 13, 14, 15, and 16 below). We then switch to the other bias model and repeat this process until the model parameters converge (we chose to declare convergence when, for each model, the relative change in predicted bias from the previous version was within 1×10^{-4}). Specifically, if we choose the GC model for the first iteration, we re-fit a new GC model to the data adjusted for distance dependence as well as the latest version of the length bias model, then we re-fit a new length model to the data adjusted for distance dependence as well as the new GC model we just computed, then fit another GC model, and so on. After we have finished fitting $F_{GC}(GC_{r,a}, GC_{r,b})$ and $F_{len}(L_{r,a}, L_{r,b})$ via this process, the bias-corrected counts, $X_{q,k,r}$, for primer-primer ligation $q=\{a,b\} \in Q_r$ in library $k \in K$ and region $r \in R$ will be Equation 7:

$$X_{q,k,r} = \exp\left[\log\left(\widehat{C}_{q,k,r} + 1\right) - F_{GC}(GC_{r,a},\ GC_{r,b}) - F_{len}(L_{r,a}, L_{r,b})\right] - 1, \quad \text{(Equation 7)}$$

where the details of the GC content and fragment length models $F_{GC}(GC_{r,a}, GC_{r,b})$ and $F_{len}(L_{r,a}, L_{r,b})$ will be explained in detail Equations 10, 11, 12, 13, 14, 15, and 16 below.

Explicit Normalization: Fragment-Level Distance Dependence Model

To obtain a simple fragment-level distance dependence model, we performed a linear regression of the logged interaction distances (in units of base pairs) against the logged counts values within a given library $k \in K$ and region $r \in R$ (Equation 8):

$$\widehat{m}_{r,k}, \widehat{b}_{r,k} = \underset{m_{r,k},\, b_{r,k}}{\operatorname{argmin}} \sum_{q=\{a,b\} \in Q_r} \left(m_{r,k}\left(\log\left(\left|M^f_{a,r} - M^f_{b,r}\right| + 1\right)\right) + b_{r,k} - \log\left(\widehat{C}_{q,k,r} + 1\right)\right)^2. \quad \text{(Equation 8)}$$

Once optimized, the parameters $\widehat{m}_{r,k}$ and $\widehat{b}_{r,k}$ describe a distance dependence model function for library k and region r (Equation 9):

$$D_{r,k}(x) = \exp\left[\widehat{m}_{r,k} \log(x+1) + \widehat{b}_{r,k}\right] - 1. \quad \text{(Equation 9)}$$

Explicit Normalization: Fragment Length Spline

Ligation events involving restriction fragments of different lengths have been previously shown to exhibit different bias-driven enrichments for detection frequency in Hi-C data, but this relationship does not appear to follow a simple functional form (Yaffe and Tanay, 2011). Moreover, the underlying mechanisms behind fragment length bias may be different in 5C vs. Hi-C assays, due to the replacement of the relatively inefficient blunt end ligation step of Hi-C with a more efficient sticky end ligation step in the 5C protocol. We presupposed that the fragment length bias effect should in theory take the shape of a smooth surface. Therefore, we fitted a bivariate spline to the contact detection enrichment of each ligation product, where the two variables were the respective lengths of the two

primers involved in the ligation. The use of a spline enforced some degree of smoothness in the resulting fitted bias model, without making strong assumptions about the functional form of the true bias curve.

To construct the spline model, we assigned $L_{r,p}$ as the length of the restriction fragment queried by primer $p \in P_r$ in region $r \in R$, T as the desired number of internal knots (we chose $T = 20$) and d as the degree of the B-splines to be fitted (we chose $d = 3$ for cubic B-splines). We selected a sequence of internal knots for spline fitting by simply taking the T-quantiles of the set $L = \{L_{r,p} : p \in P_r, r \in R\}$. We added $d+1$ redundant terminal knots to each end of the internal knot sequence to obtain a final knot sequence $(t_i)_{i=1}^{T+2d+2}$.

We then used least squares optimization to identify optimal spline parameters. We will call the $(T+2d+1)\times(T+2d+1)$ matrix of spline parameters $\mathbf{P}$. Then the optimization procedure finds (Equation 10):

$$\widehat{\mathbf{P}} = \underset{\mathbf{P}}{\operatorname{argmin}} \sum_{r \in R} \sum_{k \in K} \sum_{q=\{a,b\} \in Q_r} \left(\sum_{i=1}^{T+2d+1} \sum_{j=1}^{T+2d+1} B_{i,d}(L_{r,a}) B_{j,d}(L_{r,b}) P_{i,j} - \widehat{\widehat{C}}_{q,k,r} \right)^2, \quad \text{(Equation 10)}$$

where $B_{i,d}$ is a B-spline of degree d at knot position t_i, defined recursively according to Equations 11 and 12:

$$B_{i,0}(x) = \begin{cases} 1 & t_i \le x \le t_i + 1 \\ 0 & \text{otherwise} \end{cases}. \quad \text{(Equation 11)}$$

$$B_{i,d}(x) = \frac{x - t_i}{t_{i+d} - t_i} B_{i,d-1}(x) + \frac{t_{i+d+1} - x}{t_{i+d+1} - t_{i+1}} B_{i+1,d-1}(x), \quad \text{(Equation 12)}$$

where $\widehat{\widehat{C}}_{q,k,r}$ is defined according to Equation 13:

$$\widehat{\widehat{C}}_{q,k,r} = \log\left[\widehat{C}_{q,k,r} + 1\right] - \log\left[D_{r,k}\left(\left|M^f_{a,r} - M^f_{b,r}\right|\right) + 1\right] - F_{GC}(GC_{r,a},\ GC_{r,b}), \quad \text{(Equation 13)}$$

and represents the number of reads counted mapping to primer-primer junction $q = \{a,b\} \in Q_r$ in library $k \in K$ and region $r \in R$ after sequencing depth normalization and normalized for distance dependence as well as the latest GC bias model (in accordance with the iterative fitting procedure discussed above). Once optimized, the parameters $\widehat{\mathbf{P}}$ describe the fragment length bias function (Equation 14):

$$F_{len}(L_{r,a}, L_{r,b}) = \sum_{i=1}^{T+2d+1} \sum_{j=1}^{T+2d+1} B_{i,d}(L_{r,a}) B_{j,d}(L_{r,b}) \widehat{P}_{i,j}. \quad \text{(Equation 14)}$$

Explicit Normalization: GC Content Bias Model

In 5C data, the GC content bias factor takes only a small number of discrete values due to the short length of the genome-binding sequence of the 5C primers. Therefore, to account for GC content bias, instead of a spline model we used a simple empirical average of contact detection enrichments for a ligation product involving fragments whose associated primers had the same GC content. We computed $\widehat{\widehat{C}}_{q,k,r}$ as Equation 15:

$$\widehat{\widehat{C}}_{q,k,r} = \log\left[\widehat{C}_{q,k,r} + 1\right] - \log\left[D_{r,k}\left(\left|M^f_{a,r} - M^f_{b,r}\right|\right) + 1\right] - F_{len}(L_{r,a}, L_{r,b}), \quad \text{(Equation 15)}$$

where $\widehat{\widehat{C}}_{q,k,r}$ represents the number of reads counted mapping to primer-primer junction $q = \{a,b\} \in Q_r$ in library $k \in K$ and region $r \in R$ after sequencing depth normalization and normalized for distance dependence as well as the latest fragment length bias model (in accordance with the iterative fitting procedure discussed above). We then define the GC bias function as an empirical average over the sets of primer-primer junctions with identical GC contents, $Q_r^{u,v}$, across all libraries in K and all regions in R (Equation 16):

$$F_{GC}(u,\ v) = \frac{1}{N_K \times \sum_{r \in R} \left|Q_r^{u,v}\right|} \sum_{r \in R} \sum_{k \in K} \sum_{q \in Q_r^{u,v}} \widehat{\widehat{C}}_{q,k,r}. \quad \text{(Equation 16)}$$

Knight-Ruiz Matrix Balancing

We applied an implementation of the Knight Ruiz matrix balancing algorithm (Knight and Ruiz, 2013) published recently by Aiden and colleagues (Rao et al., 2014). We applied the Knight Ruiz algorithm to each region $r \in R$ and each library $k \in K$ independently. Knight Ruiz is an iterative algorithm that attempts to find an optimal bias vector $\mathbf{b}^{r,k}$ (whose pth element $\mathbf{b}_p^{r,k}$ is the bias factor for primer $p \in P_r$ in region $r \in R$) such that the row sums of the normalized counts matrix are as similar as possible (Equation 17):

$$\sum_{q=\{p,a\} \in Q_{r,p}} \frac{\widehat{C}_{q,k,r}}{b_p^{r,k} \times b_a^{r,k}} \approx S \qquad \forall p \in P_r,\ r \in R,\ k \in K, \quad \text{(Equation 17)}$$

for some arbitrary constant S (analogous to a row sum in a fragment-level contact matrix), where $Q_{r,p}$ represents the set of all queried ligation junctions in region $r \in R$ which involve primer $p \in P_r$. Since all junctions in $Q_{r,p}$ involve primer p, we can write them in the form

$\{p,a\}$ where $a \in P_r$ is any other primer in region r which could be involved in a queried junction with primer p. After optimization, the final normalized fragment-level counts are (Equation 18):

$$X_{q,k} = \frac{\widehat{C}_{q,k,r}}{b_a^{r,k} \times b_b^{r,k}} \qquad \forall q = \{a,b\} \in Q_r, r \in R,\ k \in K. \tag{Equation 18}$$

Express Matrix Balancing

The Express matrix balancing algorithm was first proposed by Taylor and colleagues (Sauria et al., 2015). It is similar to Knight-Ruiz matrix balancing in that it iteratively optimizes a single (log-scale) bias vector $\mathbf{b}^{r,k}$ for each region $r \in R$ and each library $k \in K$, but different in that it takes into account a regional, library-specific distance dependence expected model $\left(D_{r,k}\left(\left|M_{p,r}^f - M_{a,r}^f\right|\right)\right)$. It attempts to make the geometric mean of the ratio of the corrected counts values to their simple distance dependence expected value close across each row as close to 1 as possible (Equation 19):

$$\sqrt[|Q_{r,p}|]{\prod_{q=\{p,a\} \in Q_{r,p}} \frac{\widehat{C}_{q,k,r}}{\exp\left[b_a^{r,k}\right] \times \exp\left[b_p^{r,k}\right]}{D_{r,k}\left(\left|M_{p,r}^f - M_{a,r}^f\right|\right)}} \approx 1. \tag{Equation 19}$$

To normalize the counts values for library $k \in K$ in region $r \in R$, the Express algorithm initializes the bias vector at the zeroth iteration $\mathbf{b}^{r,k,0}=0$ and then follows the update procedure (Equation 20):

$$b_p^{r,k,n+1} = b_p^{r,k,n} + \frac{\sum_{q=\{p,a\} \in Q_{r,p}} \left(\log\left[\widehat{C}_{q,k,r} + 1\right] - \log\left[D_{r,k}\left(\left|M_{p,r}^f - M_{a,r}^f\right|\right) + 1\right] - b_p^{r,k,n} - b_a^{r,k,n}\right)}{\sum_{q=\{p,a\} \in Q_{r,p}} 2}, \tag{Equation 20}$$

where $\boldsymbol{b}^{r,k,n}$ represents the bias vector for library $k \in K$ and region $r \in R$ after n iterations. This iteration is repeated for either 1000 iterations or until the relative change in the residual is between two consecutive iterations is smaller than 1×10^{-4}. After optimization, the final normalized fragment-level counts are (Equation 21):

$$X_{q,k} = \frac{\widehat{C}_{q,k,r}}{\exp\left[b_a^{r,k}\right] \times \exp\left[b_p^{r,k}\right]} \qquad \forall q = \{a,b\} \in Q_r, r \in R, k \in K. \tag{Equation 21}$$

Joint Express Matrix Balancing

We applied a minor modification to the Express algorithm when processing multiple replicates from different biological conditions. To avoid normalizing away condition-specific effects, we constrained the Express algorithm to use one shared bias vector $\mathbf{b}^r$ for each region $r \in R$ across all replicates being analyzed.

The update equation thus became (Equation 22):

$$b_p^{r,n+1} = b_p^{r,n} + \frac{\sum_{k \in K} \sum_{q=\{p,a\} \in Q_{r,p}} \left(\log\left[\widehat{C}_{q,k,r} + 1\right] - \log\left[D_{r,k}\left(\left|M_{p,r}^f - M_{a,r}^f\right|\right) + 1\right] - b_p^{r,n} - b_a^{r,n}\right)}{\sum_{k \in K} \sum_{q=\{p,a\} \in Q_{r,p}} 2}. \tag{Equation 22}$$

After optimization, the final normalized fragment-level counts are (Equation 23):

$$X_{q,k} = \frac{\widehat{C}_{q,k,r}}{\exp\left[b_a^{r}\right] \times \exp\left[b_b^{r}\right]} \qquad \forall q = \{a,b\} \in Q_r, r \in R, k \in K. \tag{Equation 23}$$

Bias Factor Heatmaps

To visualize the quantitative strength of the various bias factors and covariates we considered, we created bias factor heatmaps similar to those used in previous analyses of bias relationships in proximity ligation data (Jin et al., 2013; Yaffe and Tanay, 2011). Previous bias factor heatmaps have often been created using only long-range (e.g., > 3 Mb or *trans*) contacts, and have quantified the enrichment for the detection (read count > 0) of ligation junctions with certain properties. Because 5C datasets generally provide a higher read depth over a much smaller area of the genome when compared to Hi-C datasets, we considered only *cis* contacts (within our 5C regions) and visualized the enrichment of total detected ligations relative to a simple distance dependence background (since our included contacts span a wide dynamic range of distance dependence background strength).

For each bias factor or covariate we considered, we first constructed groups of queried ligation junctions based on the properties of the two fragments involved in the ligation. For GC content, we started with the previously defined groups of the form $Q^{u,v}$, which denotes all the queried ligation junctions between one primer with GC content level u and another with GC content level v. Because relatively few primers were designed with extreme GC content values, we collapsed all GC content levels 20% and below into one level, and all GC content levels 70% and above into another level. For fragment length, we partitioned the primers into 7 subsets of nearly equal sizes by separating them according to the septiles of the lengths of the fragments they were designed to. We then

collected groups of queried ligation junctions based on the fragment length septiles of the two fragments involved in the ligation. For CTCF ChIP-seq signal enrichment, we first obtained CTCF ChIP-seq data from the GEO samples listed in the following table:

GEO Sample	Description
GSM2259907	ES CTCF ChIP-seq
GSM2259908	ES Input ChIP-seq
GSM2259909	NPC CTCF ChIP-seq
GSM2259910	NPC Input ChIP-seq

ChIPseq reads were aligned to mouse genome build mm9 using Bowtie (Langmead et al., 2009) with default parameters. PCR duplicates and reads with more than two reportable alignments were discarded. The mapped and filtered reads were then down-sampled to 7 million reads for each library. MACS2 (Zhang et al., 2008) was then run on these libraries with the -B/–bdg flag, and the resulting pileup bedgraph file was converted to bigwig format with the UCSC Kent source tool bedGraphToBigWig. We then computed the average bigwig signal (representing the ChIP-seq read pileup) over a 4 kb window centered on the midpoint of each fragment. We then partitioned the primers into 4 subsets of nearly equal sizes by separating them according to the quartiles of this average CTCF signal. Finally, we collected groups of queried ligation junctions based on the CTCF signal quartiles of the two fragments involved in the ligation. We repeated this procedure with both ES CTCF and pNPC CTCF ChIP-seq datasets. We used a fixed window size of 4 kb to compute the average CTCF ChIP-seq signal to avoid creating correlations with the fragment length covariate.

After establishing groups of queried ligation junctions with similar bias factor properties, we then computed a fold-enrichment relative to the region- and library-specific distance dependence expected models $D_{r,k}(x)$ mentioned above. For example, the fold change enrichment with respect to GC bias for the group of ligation junctions between primers with respective GC content levels u and v in library k is (Equation 24):

$$FC^{GC,k}_{u,v} = \frac{\text{gmean}\left(\left\{\widehat{C}_{q,k,r} : q \in Q^{u,v}_r,\ r \in R\right\}\right)}{\text{gmean}\left(\left\{D_{r,k}\left(\left|M^f_{a,r} - M^f_{b,r}\right|\right) : q = \{a,b\} \in Q^{u,v}_r,\ r \in R\right\}\right)}, \quad \text{(Equation 24)}$$

where the geometric mean of a set of values S is computed as defined in (Equation 25):

$$gmean(S) = \sqrt[|S|]{\prod_{s \in S} s}. \quad \text{(Equation 25)}$$

Binning and Smoothing

To perform binning and smoothing, we first tiled our 5C regions with bins of a desired bin width (for our primary analysis we chose 4 kb). This creates a set of adjacent bins indexed by the set $B_r = \{1, \ldots, N_{B,r}\}$ for each region $r \in R$, where $N_{B,r}$ is the number of bins in region $r \in R$. Let the genomic coordinate representing the midpoint of a given bin $b \in B_r$ be denoted by $M_{b,r}$, computed as the average of the start and end coordinates of the bin when the bin is represented as a half-open interval (i.e., it contains its start coordinate but stops right before its end coordinate). We defined a collection of sets $BP^f = \{BP^f_{p,r} = \{S^f_{p,r},\ S^f_{p,r} + 1, \ldots, E^f_{p,r} - 1\} : p \in P_r, r \in R\}$, where $BP^f_{p,r}$ is the set of base pairs covered by the fragment queried by primer $p \in P_r$ in region $r \in R$, which range across the half-open interval spanning from that fragment's start coordinate $S^f_{p,r}$ to its end coordinate $E^f_{p,r}$. Similarly, we define $BP^w = \left\{BP^w_{i,r} = \left\{M_{i,r} - \frac{w}{2}, M_{i,r} - \frac{w}{2} + 1, \ldots, M_{i,r} + \frac{w}{2} - 1\right\} : i \in B_r, r \in R\right\}$, where $BP^w_{i,r}$ is the set of base pairs covered by a smoothing window of width w (for our primary analysis we chose w = 16,000 for a 16 kb smoothing window) centered on the midpoint of bin $i \in B_r$ in region $r \in R$. We then constructed a $N_{B,r} \times N_{B,r}$ matrix of binned counts for each region $r \in R$, by computing the geometric mean of counts values for fragment-fragment junctions which lay within the smoothing window, according to Equation 26:

$$Y^{r,k}_{i,j} = gmean\left(X_{q,k} : q = \{a,b\} \in Q_r, BP^f_{a,r} \cap BP^w_{i,r}, BP^f_{b,r} \cap BP^w_{j,r}\right), \quad \text{(Equation 26)}$$

where $Y^{r,k}_{i,j}$ represents the binned interaction value between the i th and j th bins of region $r \in R$ in library $k \in K$, and the geometric mean of a set of values S is computed as defined in (Equation 27):

$$gmean(S) = \sqrt[|S|]{\prod_{s \in S} s}. \quad \text{(Equation 27)}$$

Spatial Noise Quantification

To quantify the spatial noise in the binned contact matrices, we computed the sample variance of a three-by-three square submatrix centered on each matrix entry, as long as the three-by-three square submatrix did not extend beyond the edges of the full

contact matrix or across its diagonal. Mathematically, the spatial variance of region $r \in R$ for library $k \in K$ can be written as (Equation 28):

$$SV_{r,k} = \frac{1}{\frac{(N_{B,r}-1)\times N_{B,r}}{2} - N_{B,r}} \sum_{i=1}^{N_{B,r}-1} \sum_{j=1}^{i-1} \sum_{a=i-1}^{i+1} \sum_{b=j-1}^{j+1} \left(Y_{a,b}^{r,k} - \overline{Y}_{i,j}^{r,k} \right)^2, \quad \text{(Equation 28)}$$

where $\overline{Y}_{i,j}^{r,k}$ is the sample mean in the three-by-three submatrix around the entry at i,j (Equation 29):

$$\overline{Y}_{i,j}^{r,k} = \frac{1}{9} \sum_{a=i-1}^{i+1} \sum_{b=j-1}^{j+1} Y_{a,b}^{r,k}. \quad \text{(Equation 29)}$$

Expected Modeling Strategies

In this section we will describe one-dimensional region- and library-specific distance dependence models as functions $D_{r,k} : \mathbb{N} \rightarrow \mathbb{R}$ where $D_{r,k}(|i-j|)$ denotes the distance-dependent expected value of the interaction between the i th bin and the j th bin of a given region $r \in R$ in library $k \in K$, given only the fact that this interaction occurs at a distance of $|i-j|$ bin units. These region-specific models will have analogous "global" alternatives which are fitted across all regions in a given library $k \in K$, which will be denoted with $D_k(|i-j|)$. When describing the final expected model, we will not constrain it to be one-dimensional and instead write $E_{i,j}^{r,k}$ for the expected value of the interaction between the i th bin and the j th bin of a given region $r \in R$ in library $k \in K$.

Empirical One-Dimensional Expected Model

We compute the mean of matrix entries which are the same distance from the diagonal (Equation 30):

$$D_{r,k}(|i-j|) = mean\left(\left\{ Y_{a,b}^{r,k} : a-b = i-j \right\} \right). \quad \text{(Equation 30)}$$

This empirical model can also be computed across all regions at once, as shown in (Equation 31):

$$D_k(|i-j|) = mean\left(\left\{ Y_{a,b}^{r,k} : a-b = i-j,\ r \in R \right\} \right). \quad \text{(Equation 31)}$$

In the following more complex one-dimensional expected models, we fall back to this simple empirical expected value for the first 1/3 of distance scales considered due to the challenge of modeling this portion of the distance dependence curve.

Log-Counts Lowess Fit One-Dimensional Expected Model

We perform a lowess regression of $\log(Y_{i,j}^{r,k}+1)$ against $(i-j)$, $\forall i \in B_r$, $\forall j \leq i$, $i-j > \frac{B_r}{3}$, with lowess smoothing fraction 1/3, to obtain a semilog-scale lowess-fitted function function $f_{r,k}(x)$ for each region $r \in R$ and each library $k \in K$. The final distance dependence functions are then (Equation 32):

$$D_{r,k}(x) = \exp[f_{r,k}(x)] - 1. \quad \text{(Equation 32)}$$

We can also fit this same model across all regions (including all contacts at distances $i-j > \frac{\max_{r \in R} B_r}{3}$) to obtain a semilog-scale lowess-fitted function $f_k(x)$ for each library $k \in K$. The final global distance dependence function is then (Equation 33):

$$D_k(x) = \exp[f_k(x)] - 1. \quad \text{(Equation 33)}$$

Since this fitting is performed on the scale of logged counts, the fitted values are not expected values strictly speaking. For our standard pipeline, we therefore use the empirical one-dimensional expected model described above instead. For the regional one-dimensional expected variant pipeline, we use this lowess smoothed model since the regional empirical expected is somewhat noisy (Figure S4A).

Donut Expected Model

We also computed the local donut expected as first reported by Aiden and colleagues (Rao et al., 2014). The local donut expected is a local correction factor by which a simple one-dimensional distance dependence model can be adjusted to adapt to local domain structure in the counts matrix. As proposed in (Rao et al., 2014), the local correction factor can be computed based on a series of different local windows positioned relative to the bin-bin pair whose expected value is being computed. We computed the "donut filter" as well as the "lower left filter" (Figure S4C) and chose the larger of these two results to be the final expected value. The sizes of donut and lower left filters are determined by parameters w and p, which determine the outer and inner radii of the donut window, respectively (Figure S4C). For this paper, we chose w=15, p=5. We also chose to compute the values of the filters using the global expected models $D_k(|i-j|)$ since the local correction factor accounts for differences on an even smaller scale than individual regions of the 5C primer design. The donut filter value for the interaction between the i th bin and the j th bin of a given region $r \in R$ in library $k \in K$ is (Equation 34):

$$DF_{i,j}^{r,k} = D_k(|i-j|) \times \frac{\sum_{a=i-w}^{i+w} \sum_{b=j-w}^{j+w} Y_{a,b}^{r,k} - \sum_{a=i-p}^{i+p} \sum_{b=j-p}^{j+p} Y_{a,b}^{r,k} - \sum_{a=i-w}^{i-p-1} Y_{a,j}^{r,k} - \sum_{a=i+p+1}^{i+w} Y_{a,j}^{r,k} - \sum_{b=j-w}^{j-p-1} Y_{i,b}^{r,k} - \sum_{b=j+p+1}^{j+w} Y_{i,b}^{r,k}}{\sum_{a=i-w}^{i+w} \sum_{b=j-w}^{j+w} D_k(|a-b|) - \sum_{a=i-p}^{i+p} \sum_{b=j-p}^{j+p} D_k(|a-b|) - \sum_{a=i-w}^{i-p-1} D_k(|a-j|) - \sum_{a=i+p+1}^{i+w} D_k(|a-j|) - \sum_{b=j-w}^{j-p-1} D_k(|i-b|) - \sum_{b=j+p+1}^{j+w} D_k(|i-b|)}.$$

(Equation 34)

The lower left filter value for the interaction between the i th bin and the j th bin of a given region $r \in R$ in library $k \in K$ is (Equation 35):

$$LLF_{i,j}^{r,k} = D_k(|i-j|) \times \frac{\sum_{a=i+1}^{i+w}\sum_{b=j-w}^{j-1} Y_{a,b}^{r,k} - \sum_{a=i+1}^{i+p}\sum_{b=j-p}^{j-1} Y_{a,b}^{r,k}}{\sum_{a=i+1}^{i+w}\sum_{b=j-w}^{j-1} D_k(|a-b|) - \sum_{a=i+1}^{i+p}\sum_{b=j-p}^{j-1} D_k(|a-b|)}. \quad \text{(Equation 35)}$$

Our final donut expected value (taking the largest result among the two filters considered) is (Equation 36):

$$E_{i,j}^{r,k} = \max\left[DF_{i,j}^{r,k},\ LLF_{i,j}^{r,k}\right]. \quad \text{(Equation 36)}$$

For expected models that are one-dimensional and do not use the donut correction factor, we write our final estimate of the expected value of the interaction between the i th and j th bins of region $r \in R$ for library $k \in K$ as (Equation 37):

$$E_{i,j}^{r,k} = D_{r,k}(|i-j|), \quad \text{(Equation 37)}$$

or for global expected models (Equation 38):

$$E_{i,j}^{r,k} = D_k(|i-j|). \quad \text{(Equation 38)}$$

Variance Modeling Strategies

Variance Modeling Overview

Our variance modeling strategies proceed in two stages. First, we obtain independent pixel-wise variance estimates. Second, we fit a trend between distance and variance using these pixel-wise estimates. Before we introduce the pixel-wise variance estimation methods, we will first set up our statistical model and explain how the variance in the statistical model may be broken down and understood.

Lognormal Variance Model

We propose the following lognormal null model for reasoning about variance in 5C data (Equation 39)

$$Y_{i,j}^{r,k} \sim Lognormal\left(E_{i,j}^{r,k}, V_{i,j}^{r,k}\right), \quad \text{(Equation 39)}$$

where $Y_{i,j}^{r,k}$ and $E_{i,j}^{r,k}$ are the observed and expected values, respectively, for the interaction between the i th bin and j th bin in region $r \in R$ in library $k \in K$. $V_{i,j}^{r,k}$ represents the variance of the observed value around its expected value. The intuition for this null model is that in the absence of loops, observed values $Y_{i,j}^{r,k}$ should be scattered around the donut expected value $E_{i,j}^{r,k}$ with some unknown variance $V_{i,j}^{r,k}$. In the presence of loops, the observed values $Y_{i,j}^{r,k}$ should be significantly higher than expected under this null model.

The lognormal model in Equation 39 is equivalent to a corresponding normal model (Equation 40)

$$\log Y_{i,j}^{r,k} \sim Normal\left(\mu_{i,j}^{r,k}, \left(\sigma_{i,j}^{r,k}\right)^2\right), \quad \text{(Equation 40)}$$

where $\mu_{i,j}^{r,k}$ and $\left(\sigma_{i,j}^{r,k}\right)^2$ represent the mean and variance, respectively, of a normal distribution that describes the distribution of the logged observed counts $\log Y_{i,j}^{r,k}$. The model in Equation 40 can also be rewritten as (Equation 41)

$$\log Y_{i,j}^{r,k} = \mu_{i,j}^{r,k} + \varepsilon_{i,j}^{r,k}, \quad \text{(Equation 41)}$$

where $\varepsilon_{i,j}^{r,k}$ is a normally distributed error term with mean zero and variance $\left(\sigma_{i,j}^{r,k}\right)^2$ (Equation 42)

$$\varepsilon_{i,j}^{r,k} \sim N\left(0, \left(\sigma_{i,j}^{r,k}\right)^2\right). \quad \text{(Equation 42)}$$

We will perform certain estimation steps in terms of $\left(\sigma_{i,j}^{r,k}\right)^2$ for simplicity. If we estimate $\left(\sigma_{i,j}^{r,k}\right)^2$, then the two parameters we have in-hand, $\left(\sigma_{i,j}^{r,k}\right)^2$ and $E_{i,j}^{r,k}$, describe different transformations of the observed counts $Y_{i,j}^{r,k}$ ($E_{i,j}^{r,k}$ is an expected value for unlogged counts, while $\left(\sigma_{i,j}^{r,k}\right)^2$ is a variance for logged counts), so we cannot use both of these parameters to parameterize one distribution. We therefore convert estimates of $\left(\sigma_{i,j}^{r,k}\right)^2$ back to the scale of $V_{i,j}^{r,k}$ (the scale of the original, unlogged counts values) using the known expected value $E_{i,j}^{r,k}$ when visualizing variance estimates (as in Figures 6A and 6B) or when parameterizing distributions (as in Figure 6C). In order to do this, we will leverage the following relationships, which follow from the properties of the lognormal distribution (Equations 43 and 44):

$$\mu_{i,j}^{r,k} = \log E_{i,j}^{r,k} - \frac{\left(\sigma_{i,j}^{r,k}\right)^2}{2}. \quad \text{(Equation 43)}$$

$$V_{i,j}^{r,k} = \left(\exp\left(\left(\sigma_{i,j}^{r,k}\right)^2\right) - 1\right) \times \exp\left(2 \times \mu_{i,j}^{r,k} + \left(\sigma_{i,j}^{r,k}\right)^2\right). \quad \text{(Equation 44)}$$

Per-Pixel Variance Estimators

Having obtained observed values $Y_{ij}^{r,k}$ from the biased corrected and binned contact matrices, and expected values $E_{ij}^{r,k}$ from the donut expected, we next considered approaches for estimating the variance $V_{ij}^{r,k}$ $\left(\text{or}\left(\sigma_{ij}^{r,k}\right)^2\right)$ to complete the parameterization of our statistical model. We considered two different per-pixel estimators (i.e., they compute an estimate using data from only one pixel) for $\left(\sigma_{ij}^{r,k}\right)^2$: an inter-replicate variance estimate and an intra-replicate deviation-based variance estimate.

Inter-replicate Variance Estimation

To obtain the inter-replicate variance for each pixel, we computed the sample variance of the logged counts across all libraries with the same biological condition (Equations 45 and 46):

$$\widehat{\mu}_{ij}^{r,c} = \frac{1}{|K_c|}\sum_{k\in K_c}\log\left(Y_{ij}^{r,k}+1\right), \quad \text{(Equation 45)}$$

$$\left(\widehat{\sigma}_{ij}^{r,c}\right)^2 = \frac{1}{|K_c|-1}\sum_{k\in K_c}\left(\log\left[Y_{ij}^{r,k}+1\right]-\widehat{\mu}_{ij}^{r,c}\right)^2, \quad \text{(Equation 46)}$$

where $Y_{ij}^{r,k}$ is the observed count value between the i th and j th bins in region $r\in R$, $K_c\subseteq K$ represents the subset of all libraries corresponding to a specific biological condition c, and $\left(\widehat{\sigma}_{ij}^{r,c}\right)^2$ represents the sample variance of the logged counts across all libraries with the same biological condition.

For any individual pixel, the corresponding variance estimate on the scale of the original counts $\widehat{V}_{ij}^{r,k}$ can then be obtained as the variance of the lognormal distribution whose mean is $E_{ij}^{r,k}$ and whose corresponding normal distribution has variance $\left(\widehat{\sigma}_{ij}^{r,c}\right)^2$ (Equations 47 and 48):

$$\widehat{\mu}_{ij}^{r,k} = \log E_{ij}^{r,k} - \frac{\left(\widehat{\sigma}_{ij}^{r,c}\right)^2}{2}. \quad \text{(Equation 47)}$$

$$\widehat{V}_{ij}^{r,k} = \left(\exp\left(\left(\widehat{\sigma}_{ij}^{r,c}\right)^2\right)-1\right)\times\exp\left(2\times\widehat{\mu}_{ij}^{r,k}+\left(\widehat{\sigma}_{ij}^{r,c}\right)^2\right). \quad \text{(Equation 48)}$$

These are the variance values plotted in blue hexbins in the "Inter-replicate Variance" column of Figures 6A and 6B.

Intra-replicate Deviation Variance Estimation

Under the assumptions of our statistical model, $V_{ij}^{r,k}$ quantifies how different the observed values $Y_{ij}^{r,k}$ are from the expected values $E_{ij}^{r,k}$. To obtain intra-replicate variance estimates for each pixel, we considered directly fitting the noise term in our model. Starting from our normal statistical model (Equation 41), we estimated the variance of the noise term as the average squared residual between what we expect given the donut model and what we actually observe (Equation 49):

$$\left(\sigma_{ij}^{r,k}\right)^2 = Var\left[\varepsilon_{ij}^{r,k}\right] = E\left[\left(\log Y_{ij}^{r,k}-\mu_{ij}^{r,k}\right)^2\right], \quad \text{(Equation 49)}$$

where **Var**[X] represents the variance of a random variable X and **E**[X] represents the expected value of a random variable X. In practice, $\mu_{ij}^{r,k}$, the mean of the normal distribution corresponding to the lognormal distribution we are effectively fitting, depends on $\left(\sigma_{ij}^{r,k}\right)^2$ via Equation 43, so we use $\log[E_{ij}^{r,k}+1]$ as an upward-biased estimate of the expected value (an unbiased estimate could theoretically be obtained by logging the geometric mean expected value instead of $E_{ij}^{r,k}$). However, this should still lead to estimates of the variance that are conservative on average, since the average squared deviation can only get larger as we move the $\log[E_{ij}^{r,k}+1]$ term away from its unbiased estimator. Later, when evaluating statistical significance, we will not make this assumption.

Plugging in $\log[E_{ij}^{r,k}+1]$ for $\mu_{ij}^{r,k}$ and computing the estimate for only one pixel in one replicate, we obtain our per-pixel intra-replicate deviation-based variance estimator (Equation 50):

$$\left(\widehat{\sigma}_{ij}^{r,k}\right)^2 = \left(\log\left[Y_{ij}^{r,k}+1\right]-\log\left[E_{ij}^{r,k}+1\right]\right)^2. \quad \text{(Equation 50)}$$

For any individual pixel, the corresponding variance estimate on the scale of the original counts $\widehat{V}_{ij}^{r,k}$ can then be obtained as the variance of the lognormal distribution whose mean is $E_{ij}^{r,k}$ and whose corresponding normal distribution has variance $\left(\widehat{\sigma}_{ij}^{r,k}\right)^2$ (Equations 51 and 52):

$$\widehat{\mu}_{ij}^{r,k} = \log E_{ij}^{r,k} - \frac{\left(\widehat{\sigma}_{ij}^{r,k}\right)^2}{2}. \quad \text{(Equation 51)}$$

$$\widehat{\widehat{V}}_{i,j}^{r,k} = \left(\exp\left(\left(\widehat{\widehat{\sigma}}_{i,j}^{r,k}\right)^2\right) - 1\right) \times \exp\left(2 \times \widehat{\widehat{\mu}}_{i,j}^{r,k} + \left(\widehat{\widehat{\sigma}}_{i,j}^{r,k}\right)^2\right). \quad \text{(Equation 52)}$$

These are the variance values plotted in blue hexbins in the "Deviation Variance" and "Constant Variance" columns of Figures 6A and 6B.

Variance Relationship Fitting

Our per-pixel variance estimates are computed with a small number of observations (one under the intra-replicate deviation variance model, or one for each library $k \in K_c$ under the inter-replicate variance model). Therefore, a more accurate model of the relationship between the mean and the variance can be obtained by combining information across bin-bin pairs. We observed a strong relationship between the per-pixel variance estimates and genomic interaction distance (Figure 6A). Therefore, we used lowess to the trend between the per-pixel variance estimates $(\widehat{\sigma}_{i,j}^{r,k})^2$ and the genomic interaction distance $|i-j|$ (which is closely correlated with expected value $E_{i,j}^{r,k}$). We include data from all regions in one global fit. Together, this approach yields a library-specific (but not region-specific) Distance-Variance Relationship (DVR) function (Equation 53):

$$\left(\widehat{\widehat{\sigma}}_{i,j}^{r,k}\right)^2 = f_k(|i-j|), \quad \text{(Equation 53)}$$

where $(\widehat{\widehat{\sigma}}_{i,j}^{r,k})^2$ represents the smoothed variance parameter estimate (i.e., our best estimate for the variance of the normally distributed error term $\varepsilon_{i,j}^{r,k}$ in the statistical model in Equation 41) for the interaction between the i th and j th bins in region $r \in R$ in library $k \in K$.

We note that there are potential sources of bias in our estimation of $(\widehat{\widehat{\sigma}}_{i,j}^{r,k})^2$. First, when using deviation-based per-pixel variance estimates, our estimates of $(\widehat{\widehat{\sigma}}_{i,j}^{r,k})^2$ are biased upwards by the presence of true positive looping pixels in the set of all points to which the trend is fitted, which have larger deviations from their expected values than the true null pixels (since they are engaged in real loops). Second, since the sampling distribution of the sample variance is right-skewed, the lowess-fitted variance estimates are biased downward due to the residual-based weighting used in the lowess procedure to reduce the effect of outliers.

When parameterizing distributions (as in Figure 6C), we apply the relationships in Equations 43 and 44 to obtain (Equations 54 and 55):

$$\widehat{\widehat{\mu}}_{i,j}^{r,k} = \log E_{i,j}^{r,k} - \frac{\left(\widehat{\widehat{\sigma}}_{i,j}^{r,k}\right)^2}{2}, \quad \text{(Equation 54)}$$

$$\widehat{\widehat{V}}_{i,j}^{r,k} = \left(\exp\left(\left(\widehat{\widehat{\sigma}}_{i,j}^{r,k}\right)^2\right) - 1\right) \times \exp\left(2 \times \widehat{\widehat{\mu}}_{i,j}^{r,k} + \left(\widehat{\widehat{\sigma}}_{i,j}^{r,k}\right)^2\right), \quad \text{(Equation 55)}$$

where $\widehat{\widehat{V}}_{i,j}^{r,k}$ is the variance of the lognormal distribution describing unlogged observed values $Y_{i,j}^{r,k}$, fitted using the DVR function from Equation 53.

Finally, in order to visualize the fitted variance estimates $(\widehat{\widehat{\sigma}}_{i,j}^{r,k})^2$ as a function of genomic distance, we used our fitted variance estimates to estimate a variance for each distance scale $|i-j|$, again applying the relationships in Equations 43 and 44. However, we could not directly plug $E_{i,j}^{r,k}$ into Equation 43 because there are many different values of $E_{i,j}^{r,k}$ at each distance scale $|i-j|$; therefore, we replaced $E_{i,j}^{r,k}$ with an average value for that distance scale (computed by performing a global log-counts lowess fit between the donut expected values $E_{i,j}^{r,k}$ and their genomic interaction distances $|i-j|$ across all regions $r \in R$). This allowed us to obtain (Equations 56–58):

$$\left(\widehat{\widehat{\sigma}}_{|i-j|}^{k}\right)^2 = f_k(|i-j|), \quad \text{(Equation 56)}$$

$$\widehat{\widehat{\mu}}_{|i-j|}^{k} = \log(D_k(|i-j|)) - \frac{\left(\widehat{\widehat{\sigma}}_{|i-j|}^{k}\right)^2}{2}, \quad \text{(Equation 57)}$$

$$\widehat{\widehat{V}}_{|i-j|}^{k} = \left(\exp\left(\left(\widehat{\widehat{\sigma}}_{|i-j|}^{k}\right)^2\right) - 1\right) \times \exp\left(2 \times \widehat{\widehat{\mu}}_{|i-j|}^{k} + \left(\widehat{\widehat{\sigma}}_{|i-j|}^{k}\right)^2\right), \quad \text{(Equation 58)}$$

where $D_k(x)$ represents an average expected value at distance x as mentioned above, and $\widehat{\widehat{V}}_{|i-j|}^{k}$ represents an averaged smoothed variance estimate for interactions at distance $|i-j|$ in library $k \in K$. We then visualized the trend between genomic distance $|i-j|$ and

variance $\widehat{\widehat{V}}^{k}_{|i-j|}$ as the red curves in the "Inter-replicate Variance" and "Deviation Variance" columns of Figures 6A and 6B (fitting the DVR to $(\widehat{\overline{\sigma}}^{r,c}_{i,j})^2$ and $(\widehat{\overline{\sigma}}^{r,k}_{i,j})^2$, respectively).

Constant Variance Fitting

For comparison to the lowess-fitted variance trends, we fitted a single constant value to the mean value of the intra-replicate deviation-based variance estimates on the scale of the original counts $\widehat{V}^{r,k}_{i,j}$, using the same residual-based reweighting implemented by the lowess procedure. This is equivalent to performing a lowess fit of $\widehat{V}^{r,k}_{i,j}$ against non-correlated random noise but has the advantage of being deterministic. This constant variance value was then plotted as the red curve in the "Constant Variance" column of Figures 6A and 6B.

Relationship between Inter-replicate and Intra-replicate Deviation Variance

We observed that the inter-replicate variance model appeared to underestimate the variance in the data compared to the intra-replicate deviation variance model (Figures 6A–6C). While the full investigation of the relationship between these approaches remains an interesting area for future work, we propose a few speculative hypotheses below.

We speculate that the inter-replicate variance model may implicitly assume that the donut expected value $E^{r,k}_{i,j}$ is an unbiased, noiseless estimator for the mean observed value across replicates $\overline{Y}^{r}_{i,j}$ at all null (non-looping) pixels. If it is not, then pixels where $\overline{Y}^{r}_{i,j}>E^{r,k}_{i,j}$ may be called significant even when they are not actually looping. These pixels are looping in the sense that $\overline{Y}^{r}_{i,j}>E^{r,k}_{i,j}$, but they are not looping in the sense that the elevation of $\overline{Y}^{r}_{i,j}$ over $E^{r,k}_{i,j}$ is driven by downward bias or noise in $E^{r,k}_{i,j}$ at this pixel rather than by biological elevation of $\overline{Y}^{r}_{i,j}$. This may happen quite often if the inter-replicate variance $\overline{V}^{r,k}_{i,j}$ is low compared to the size of the bias or noise in $E^{r,k}_{i,j}$. We suspect that this effect may drive underestimation of the variance under the inter-replicate variance model. We speculate that the intra-replicate deviation variance model has the potential to include biases and noise in $E^{r,k}_{i,j}$ in its variance estimate since it takes $E^{r,k}_{i,j}$ into account in the estimation.

We also speculate that the intra-replicate deviation-based variance model may capture some components of the inter-replicate variance that are distributed across pixels rather than experiments, such as randomness arising from the sampling of detected ligation junctions. Other components of the inter-replicate variance, such as variation due to biological variability or batch effects across experiments, are presumably not included in the intra-replicate deviation-based variance model.

Distribution Parameterization and P-Value Computation

To judge the statistical significance of each observed interaction value $Y^{r,k}_{i,j}$, we parameterized a lognormal distribution with mean $E^{r,k}_{i,j}$ and variance $\widehat{\widehat{V}}^{r,k}_{i,j}$ for each pixel. These are the blue parameterized distributions shown in Figure 6C. The statistical model in Equation 39 represents our null hypothesis, while our alternative hypothesis (corresponding to looping interactions) is that the observed counts are greater than we expect under this model. We therefore compute right-tail P-values using this fitted distribution according to Equation 59

$$P^{r,k}_{i,j} = P\left(X \geq Y^{r,k}_{i,j}\right); \quad X \sim Lognormal\left(E^{r,k}_{i,j},\ \widehat{\widehat{V}}^{r,k}_{i,j}\right). \tag{Equation 59}$$

Interaction Classification

To classify cell type-specific interactions, we first applied Benjamini-Hochberg (BH) multiple testing correction to the right-tail P-values, controlling the false discovery rate (FDR) at 10%. To reduce the impact of noise on our interaction calls, we excluded interactions with interaction distances shorter than 24 kb from consideration, as well as interactions which lay in connected components with 3 or fewer members (i.e., clusters of pixels smaller than four pixels in size). We also discarded interactions which were only significant in one replicate of a given condition and did not reproduce in the other replicate. We then overlapped concordant significant interactions across conditions, calling points which were significant in both conditions "constitutive" and those points which were significant in only one condition "ES-only" or "pNPC-only", according to which of the two conditions they were significant. We note that this additional stringency reduces our effective FDR below its nominal value after BH correction. Additionally, we selected a background threshold P-value b (we chose b=0.6) and created a background class of points which had P-values above this threshold in all libraries analyzed (Equation 60):

$$B = \left\{(i,j,r) : P^{r,k}_{i,j} > b \ \ \forall \ \ k \in K\right\}. \tag{Equation 60}$$

Enrichments

To compute enrichments of genomic annotations with our classified interactions, created the following two-by-two contingency table for each combination of interaction class and genomic annotation tested:

	In Interaction Class	In Background Class
Intersects annotation	*a*	*b*
Does not intersect annotation	*c*	*d*

We considered an interaction as intersecting an annotation if the annotation lay either within one of the two bins involved in the interaction, or one bin away from either of these bins. We computed a fold-enrichment for each combination of interaction class and genomic annotation as $\frac{a/c}{b/d}$, and computed a P-value by applying Fisher's exact test to the two-by-two contingency table. One-dimensional genomic annotations were obtained from Tables S5–S18 of Beagan et al. (2017).

Convergency Analysis

To perform the CTCF convergency analysis, we first obtained a list of CTCF motifs present in the mm9 reference genome, with associated orientation information. We then intersected this list with CTCF ChIP-seq peaks in the cell type under consideration (either ES or pNPC) to get a list of occupied motifs. Next, we intersected this list of occupied motifs with the interactions which were found to be significant in the cell type under consideration, considering each of the two bins involved in the interaction separately. If either anchor contained no occupied CTCF sites, we excluded this interaction from the analysis. If either anchor contained occupied CTCF sites with a mixture of motif orientations, we also excluded this interaction from the analysis. If both anchors contained occupied CTCF sites with a unique motif orientation, we noted the relative orientations of the two motifs counted this as an intersection in cell *a* of a two-by-two contingency table specific to that relative orientation *o*:

	Interaction Significant in Cell Type	Interaction in Background Class
Has relative orientation *o*	*a*	*b*
Does not have relative orientation *o*	*c*	*d*

We then computed the fold enrichment of each relative orientation *o* as $\frac{a/c}{b/d}$, taking the entries from the contingency table specific to relative orientation *o*.

Hi-C Data Comparison

To compare the results of analysis to Hi-C data, we obtained the following datasets from (Bonev et al., 2017):

GEO Sample	Cell Type
GSM2533818	ES
GSM2533819	ES
GSM2533820	ES
GSM2533821	ES
GSM2533822	NPC
GSM2533823	NPC
GSM2533824	NPC
GSM2533825	NPC

Replicates from the same cell type were combined. Raw reads were mapped using HiC-Pro (Servant et al., 2015) using default parameters. Contact matrices were assembled at 10kb resolution and subsequently balanced using the Knight-Ruiz algorithm implemented in Juicer (Durand et al., 2016).

H3K27ac ChIP-seq Track Processing

The H3K27ac ChIP-seq tracks visualized in Figures 4, 5, 7, and S5 were processed as follows. First, the following datasets were obtained from Creyghton et al. (2010):

GEO Sample	Description
GSM594579	ES H3K27ac ChIP-seq
GSM594585	NPC H3K27ac ChIP-seq

Reads were aligned to mouse genome build mm9 using Bowtie (Langmead et al., 2009) with default parameters. PCR duplicates and reads with more than two reportable alignments were discarded. The mapped and filtered reads were then downsampled to 7 million reads for each library. MACS2 (Zhang et al., 2008) was then run on these libraries with the -B/–bdg flag, and the resulting pileup bedgraph file was converted to bigwig format with the UCSC Kent source tool bedGraphToBigWig. Finally, the pileup distributions were quantile normalized to create the final bigwig files used for visualization.

QUANTIFICATION AND STATISTICAL ANALYSIS

The line plots in Figure 2B show the mean count value of groups of primer-primer ligation junctions within an interaction distance bin (one of 30 evenly-spaced bins between 0 kb and 600 kb inclusive). In the boxplots in Figure S2A, the red lines indicate the median, the boxes indicate the IQR, and the whiskers extend to 2.5 times the IQR. The p-values shown on Figure 7D are computed using Fisher's exact test on a two-by-two contingency table:

	In Interaction Class	In Background Class
Intersects annotation	*a*	*b*
Does not intersect annotation	*c*	*d*

and the fold-enrichments shown (on a $\log_2$ scale) are computed as $\frac{a/c}{b/d}$.

DATA AND SOFTWARE AVAILABILITY

The software package developed to perform the analyses in this paper is available as a Python package with full usage instructions on Bitbucket at https://bitbucket.org/creminslab/lib5c/ The accompanying README file is also included within the Supplemental Information as Data S1.