High-Resolution Single-Cell Models of Ensemble Chromatin Structures during *Drosophila* Embryogenesis from Population Hi-C

Qiu Sun[†1], Alan Perez-Rathke[†2], Daniel M. Czajkowsky[3], Zhifeng Shao[*3] and Jie Liang[*2]

[1]*Shanghai Center for System Biomedicine , Shanghai Jiao Tong University, Shanghai 200240, China*

[2]*Department of Bioengineering, University of Illinois at Chicago, SEO, MC-063, Chicago, IL 60607-7052, USA*

[3]*School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*

[†]These two authors contribute equally.

[*]Corresponding authors: `zfshao@sjtu.edu.cn` to ZS and `jliang@uic.edu` to JL

Abstract

We describe a computational method for reconstruction of large 3D ensemble of single-cell chromatin conformations from population Hi-C. Our method identifies specific interactions constituting a small fraction (5–6% in *Drosophila*) of measured Hi-C frequencies. Yet, surprisingly, they are sufficient to drive chromatin folding, giving rise to much of observed Hi-C topological features. Our models reveal dramatic changes in chromatin compaction across three developmental stages of *Drosophila*, with >50% of single cells likely maintaining TAD-like structures, possibly even in early embryos when no population TADs are discernable. Our models reveal details on how varying single-cell domain boundaries become fixated throughout development, with strong preferred-positioning at binding sites of insulator complexes during and after midblastula transition. Overall, our method can be used to re-interpret population Hi-C by generating ensembles of single-cell chromatin conformations at high resolution, facilitating better understanding of genome organization.

Introduction

Understanding the principles of genome organization is essential for gaining insight into fundamental biological processes such as gene expression and DNA replication [1–4]. Studies based on chromosome conformation capture (3C) and related techniques (4C, 5C and Hi-C) have demonstrated a wealth of structural features of genome folding [5–8]. Among these, topologically associating domains (TADs) are important structural units, within which chromatin exhibits heightened self-interacting probabilities [9–11]. They are prominently found in Hi-C studies of many cell types [10] and across different species [11]. In addition, insulator complexes are found to be preferentially located at boundaries of TADs in multiple species [11–13]. Despite abundant high-resolution information provided by Hi-C studies, important questions such as how chromatin conformations of TADs appear at the single-cell level cannot be resolved due to the intrinsic population-averaged nature of these studies.

Recent single-cell Hi-C studies showed that many structural features defined by population Hi-C such as TADs and compartments vary among individual cells [14–17]. 3D fluorescent *in situ* hybridization (FISH) and super-resolution imaging technologies further revealed varied details of TAD-like structures in individual nuclei [18–21]. A recent study with S2R+ cells, which are derived from late embryos of *Drosophila*, revealed the well-organized nature of repressed TADs in single cells [20]. At nanometer-scale resolution, single-cell TAD-like structures with sharp boundaries are found to be widely distributed along the human genome [19]. Despite rapid progress, there are limitations of these single-cell approaches. Although single-cell Hi-C captures chromatin contacts and can

characterize cell-to-cell variability, it is difficult to obtain detailed chromatin structures at high resolution due to the highly sparse nature of single-cell Hi-C data [22]. Super-resolution imaging studies can provide fine-resolution information on single-cell genome folding, but are restricted to a moderate number of cells.

Our understanding of important aspects of chromatin organization is still limited. While population Hi-C studies can identify a large amount of chromatin contacts, it is unclear to what degree they are obligated for formation of structures such as TADs as observed in Hi-C studies [23]. It is also unclear that among spatial neighboring relationships identified in single-cell studies, which reflect functional association and which are due to random collision under effects such as volume confinement [24–26]. An interesting question is whether a minimum set of driver interactions of biological importance exist that are sufficient for chromatin folding [27–29].

Furthermore, as most single-cell studies focus on late embryos or mature cell lines, how spatial chromatin structures such as TADs arise during embryogenesis in single cells and in population, and how such structures evolve remains elusive. In addition, we lack a unified and consistent model that simultaneously: 1) accounts for the ensemble behavior of chromatin as exhibited in Hi-C measurements, 2) quantifies the heterogeneity of chromatin structures in the cell population, and 3) provides detailed information on chromatin architectures in single cells.

In this study, we describe a computational strategy that enables construction of large ensembles of single-cell 3D chromatin conformations from population Hi-C measurements. It complements current single-cell techniques as it can be

applied to more abundantly available population Hi-C data, while not restricted in either genome coverage nor the number of single cells.

Our method first identifies a set of critical chromatin interactions from deep-sequenced population Hi-C data, which are termed *specific interactions*. While constituting only 5–6% of the Hi-C measurements in *Drosophila*, extensive simulations on regions of varying genomic lengths of three *Drosophila* cell types at different developmental stages and on whole chromosome show that these interactions are sufficient to drive chromatin folding and can reproduce much of topological features as observed in Hi-C measurements.

The reconstructed chromatin ensembles reveal novel insight into chromatin organization, including dramatic changes in chromatin compaction during embryogenesis, as well as stage-dependent changes of a functional unit of three-body interactions. In addition, the overall heterogeneity of chromatin structures of the underlying cell populations can be quantified from these ensemble models.

At the single-cell level, our analysis of cells from embryos at cycles 9–13, stages 5–8, and S2R+ shows that a large portion (60%) of S2R+ cells maintain TAD-like structures in a 1 Mb region, similar to the findings in earlier super-resolution imaging studies [20, 21]. Surprisingly, our results suggest that this may also hold in early-stage embryos at cycles 9–13, where $> 50\%$ cells may have at least one TAD-like structure, although no concordant TADs can be seen in population-averaged Hi-C. Our method further suggests that at cycles 9–13, the boundaries of TAD-like structures may vary significantly, but become much more fixated at later stages, with strong preference for binding sites of insulator complexes formed during the midblastula transition.

Overall, results presented in this study demonstrate that our method can be used to re-interpret population-based Hi-C measurement. Through transformation of population Hi-C into single-cell chromatin conformations at high resolution, our method can facilitate discovery of novel biological insight and better understanding of genome organization.

Results

Specific interactions constitute a small fraction of Hi-C contacts

Hi-C measurements may contain many non-specific and likely spurious contacts, as a large number of them result from random collisions of chromatin fibers confined inside the nuclear volume [24, 28]. We employ a physical null model independent of the Hi-C data for normalization. Our null model is that of a large ensemble of self-avoiding random 3D polymer chains of chromatin fibers in nuclear confinement (Figure 1a). Extending the technique of C-SAC [25], we use the fractal Monte Carlo sampling method (under review, see Supplementary Methods) to construct null models for three *Drosophila* cell types at different developmental stages, namely, embryos at cycles 9–13, at stages 5–8, and S2R+ derived from late embryos.

For each cell type, the null model is an ensemble of 2.5×10^5 chromatin chains of 4 Mb length at 2 kb resolution, each consisting of 2,000 beads. The volume constraining the chromatin polymers is proportional to the nuclear volume of

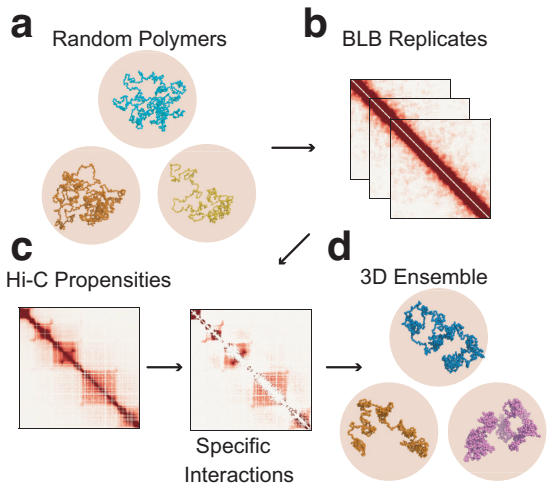


Figure 1. Overview of our chromatin modeling strategy. **(a)** We first construct ensemble of random chromatin polymers as our physical null model. We then **(b)** bootstrap the polymer ensemble to compute statistical significance of each Hi-C interaction. After removal of non-specific interactions, **(c)** we construct an ensemble of 3D chromatin structures constrained by specific interactions we identified under a sequential Bayesian inference framework.

each cell type (see Methods).

We estimate the random contact probabilities of pairs of loci by counting the frequency of 3D conformations in which the spatial distance of the corresponding pair of beads is within a Euclidean threshold of 80 nm [27, 28]. We evaluate the statistical significance of each Hi-C contact by bootstrapping the corresponding random ensemble (Figure 1b, see Methods). Hi-C interactions with BH-FDR [30] adjusted p-values below a threshold of 0.01 are then identified as specific interactions (Figure 1c).

As our null model can be used to identify specific interactions within 4 Mb genomic distance, we slide a window of 4 Mb along the entire genome and identify all such specific interactions genome-wide (See Supplementary Methods). Overall, there are 2.28×10^6 out of 42.39×10^6 (5.4%), 2.03×10^6 out of 40.07×10^6 (5.1%),

and 2.20×10^6 out of 34.98×10^6 (6.3%) Hi-C interactions that are found to be specific for embryos at cycles 9–13, stages 5–8, and S2R+, respectively.

Specific interactions capture important long-range interactions

Specific interactions contain rich information. This can be clearly seen in the heat maps of a polycomb-repressed region shown in Figure 2a. Despite the fact that only ∼5% of Hi-C interactions are retained, key structural patterns such as the progressive formation of TADs and their finer structures are all present (Figure 2a, Supplementary Figure S2a for other regions).

To further evaluate the specific interactions identified, we ask whether long-range interactions with known biological functions are captured. We consider gene *Bsg25A*, which is transcribed during the minor wave of zygotic genome activation (ZGA) [31, 32]. It is known to form a long-range interaction with gene *slam* in early embryos [33]. This interaction is indeed captured as seen in the virtual 4C plot (Figure 2b, lower half), which appears to have captured all relevant long-range interactions, while containing much less noise (Figure 2b, upper half). Another known long-range loop interaction in late embryos is between gene *Scyl* and *chrb* [33, 34]. This is also identified as a specific interaction (Supplementary Figure S2b).

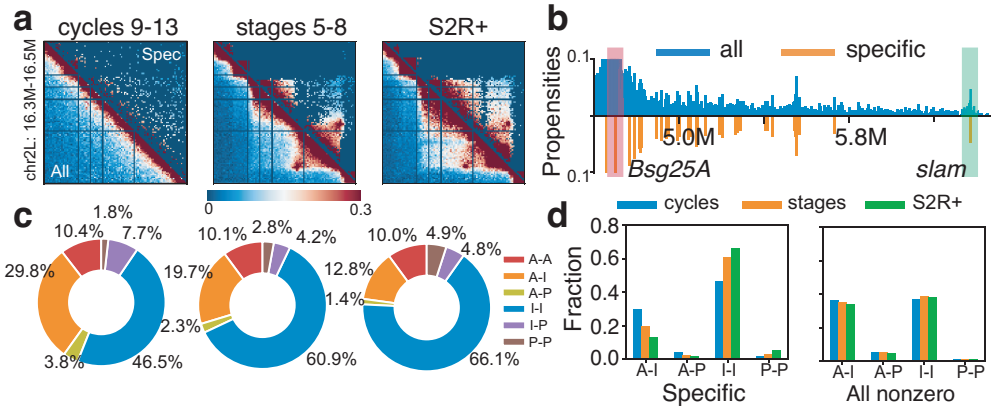


Figure 2. Overview of specific Hi-C interactions. **(a)** Heat maps of specific interactions identified in a polycomb-repressed region (chr2L: 16.3–16.5 Mb) of three cell types at different developmental stages. Lower-left triangles represent all Hi-C interactions, upper-right triangles represent the identified specific interactions. Cell types from top to bottom are embryos at cycles 9-13, embryos at stages 5–8, and S2R+, respectively. **(b)** A virtual 4C plot of the distribution of specific interactions in a 1.86 Mb region of embryos at cycles 9-13. Red bar represents the anchor which contains the gene *Bsg25A*, green bar represents a specific interaction that targets gene *slam*. **(c)** Pie charts of percentages of different types of specific interactions in the three cell types. A: active, I: inactive, P: polycomb-repressed. I-I interactions increase from 46.5% to 60.9% and then to 66.1%, P-P increases from 1.8% to 2.8% and then to 4.9%, while A-I decreases from 29.8% to 19.7%, then to 12.8%, and A-P decreases from 3.8% to 2.3%, then to 1.4%. **(d)** Percentages of four interaction types of specific interactions (left) and all nonzero Hi-C interactions (right). These four types are A-I, A-P, I-I and P-P, respectively.

Clear biological pattern emerging from identified specific interactions

The predicted specific interactions also offer clearer biological patterns of genomic contacts, which would otherwise be obscure. This is illustrated by the nature of specific interactions. Genomic regions of *Drosophila* can be classified broadly into four chromatin states [13, 35] using clustering based on signals of 15 histone modifications and other biomarkers (see Supplementary Methods). These states are: Active (A), Inactive (I), Polycomb-repressed (P), and Undetermined (U) (Supplementary Figure S2c). The global interaction type distribution of spe-

cific interactions exhibits overall an increasing number of Inactive-Inactive (I-I) and Polycomb-Polycomb (P-P) contacts in later embryos of stages of 5–8 and S2R+, while Active-Inactive (A-I) and Active-Polycomb (A-P) interactions are found to have steadily declined (Figure 2c). In contrast, the global distribution of interaction types of all nonzero Hi-C interactions exhibit no clear pattern: there are only small and random variations among cells at different stages of embryogenesis, regardless the interaction types (Figure 2d).

Furthermore, density curves of specific interactions at different genomic distances exhibit increased TAD-level ($\leq$ 400 kb) contact frequencies in embryos at stages 5–8 and S2R+ (Supplementary Figure S2d, top). In contrast, no such pattern can be seen in density curves of all Hi-C interactions (Supplementary Figure S2d, bottom). That is, specific interactions exhibit clearer pattern more consistent with the observations that TADs are established progressively during *Drosophila* embryogenesis [33, 36].

Specific interactions are sufficient to drive chromatin folding in *Drosophila*

To assess the roles of specific interactions, we ask whether they can drive chromatins to fold into conformations as measured in Hi-C studies. To ensure our conclusion is general, we examine 10 genomic regions of varying lengths (200 kb–2 Mb, Supplementary Table S1).

We construct 3D ensembles of single-chain chromatin conformations at high resolution of 2 kb using constraints of specific interactions. The conformations are

generated through a novel approach under the framework of sequential Bayesian inference (Figure 1d, see also Supplementary Methods and Figure S1).

For each region, we construct a proper ensemble of 50,000 single-chain conformations. To estimate contact probability, we take the minimalistic assumption that DNA fragments in close proximity are available for Hi-C ligations. The contact probabilities are then taken as the proportions of 3D conformations in which the spatial distance of the specific pair of loci of interests is within an euclidean distance threshold (80 nm following ref [27, 28]). We then aggregate single-chain conformations to obtain the simulated Hi-C contact maps (Figure 3a, Supplementary Figure S3a).

Simulated Hi-C contact maps using only specific interactions consistently exhibit strong similarities to measured Hi-C contact maps across the 10 regions (Pearson correlation coefficients $r = 0.91$–0.98, distance-adjusted correlation coefficients $r' = 0.56$–0.81 [37], Figure 3c, Supplementary Figure S3b). In addition, the log-log scaling curve of simulated contact probabilities at different genomic distances and the corresponding curve from Hi-C largely match each other (Figure 3b).

We then construct an ensemble of 3D chromatin conformations for the whole chromosome X of S2R+ cells at 5 kb resolution, using 4,485 beads and only 6.1% of the Hi-C contact pairs that are specific (Figure 3f). Figure 3g shows two examples of conformations of chromosome X. Again, the simulated and measured Hi-C contact maps are very similar ($r = 0.94$ and $r' = 0.64$).

Overall, these results show that using only 5–6% of measured Hi-C contacts

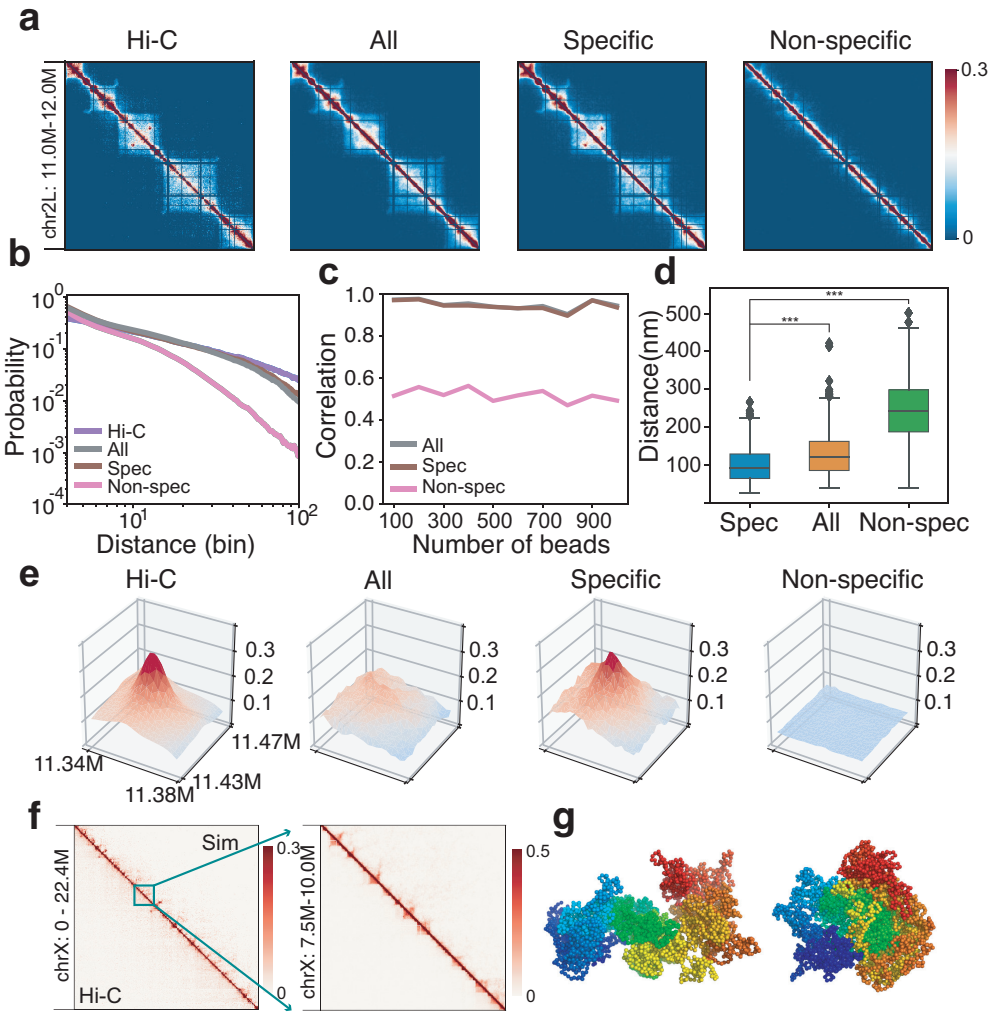


Figure 3. Specific interactions are sufficient to drive chromatin folding in *Drosophila*. **(a)** Illustration of simulation results for a 1 Mb region (chr2L: 11.0–12.0 Mb) of S2R+ cells at 2 kb resolution. Heat maps from left to right represent Hi-C propensities, simulated contact probabilities using all, specific, and non-specific interactions, respectively. **(b)** log-log scaling curves of contact probabilities with genomic distances (bin) derived from the original Hi-C data, simulated ensembles using specific, all, and non-specific interactions. **(c)** Pearson correlation coefficients of the simulated contact probabilities and Hi-C propensities in 10 regions of different genomic lengths. The number of beads ranges from 100 to 1,000. **(d)** Distance distributions of the two anchors of the loop shown in **(a)**. Loop anchors correspond to bead No. 170 and bead No. 190. *** represents two-sided Wilcoxon rank sum test p-value ≤ 0.001. **(e)** Height maps of contact probability of the loop interaction inside the polycomb domain shown in **(a)**. **(f)** Constructed 3D conformations of chromosome X using specific interactions at 5 kb resolution. (Left) Simulated Hi-C heat map, with measured Hi-C propensities at the bottom and simulated contact probabilities at the top. (Right) Zoomed-in heat map of a 2.5 Mb region. Pearson correlation coefficient r is 0.94, distance-adjusted correlation r' is 0.64. **(g)** Visualization of two examples of 3D conformations of chromosome X using PyMOL.

that are predicted as specific interactions, we can consistently reproduce experimental Hi-C contact maps across different regions with high accuracy at high resolution. This strong similarity is maintained at the whole chromosome level. Furthermore, even a smaller fraction of specific interactions may appear in each single cell of the ensemble: only 3.5–6.7% occur in the 1 Mb region shown in Figure 3a.

Our results demonstrate that predicted specific interactions are sufficient to drive chromatin folding in *Drosophila*.

3D loop structures recovered with improved clarity from specific interactions

To further ascertain the roles of specific interactions, we compare simulated ensembles of chromatin chains for the 10 regions using only specific interactions, using all Hi-C interactions, and using only non-specific interactions of the same number of contact pairs as the specific interactions (Figure 3a). Simulated heat maps of contact probability using all and specific interactions exhibit strong similarities to the corresponding heat maps of measured Hi-C ($r = 0.92 - 0.98$ and $0.91 - 0.98$, respectively). In contrast, simulated Hi-C heat maps using non-specific interactions fail to capture much of the structural features observed in Hi-C maps ($r = 0.48 - 0.58$, Figure 3a and 3c).

Although ensembles generated using all Hi-C constraints and using only specific interactions have similar correlations, the latter can recover structural features such as loops with better clarity (Figure 3a). This is illustrated by the

detailed height maps of contact probability of simulated 3D chromatin chains in a ∼40 kb × 40 kb region where a loop interaction site within a polycomb-repressed domain is located (Figure 3e). The height map of contact probability calculated from the ensemble by specific interactions has much stronger resemblance to the original Hi-C frequency map (left) than that from the ensemble by all Hi-C contacts. The ensemble by non-specific interactions fails to capture this loop structure.

We further calculate the spatial distances between the two anchors of this loop interaction (Figure 3d). 41.5% of conformations from the ensemble by specific interactions have a spatial distance less than the ligation threshold of 80 nm. In contrast, the percentage is only 20.0% and 1.7% for ensembles generated using all Hi-C and non-specific interactions, respectively. For a control pair of loci in the same region and of the same genomic distance but without looping interaction, the fractions of conformations with the loci within the ligation threshold are indistinguishable between ensembles by specific and by all interactions (11.1% and 12.4%, Supplementary Figure S3c).

These results show that chromatin ensembles reconstructed by specific interactions have better structural clarity in defining loop interactions, without compromise in detection specificity.

Heterogeneity of single-cell chromatin conformations

Previous studies showed that chromatin organization is established progressively during *Drosophila* embryogenesis [33,36]. A number of long-range interactions are

found to exist during the minor wave of zygotic transcription at cycles 9-13 [31]. This is followed by the establishment of TADs and compartments during the midblastula transition (MBT) [38], which are clearly discernable in Hi-C profiles of embryos at stages 5–8 and S2R+ during and after MBT (Figure 2a). However, these findings are all based on two-dimensional analysis of population Hi-C, and their structural implications in chromatin folding remain elusive.

We set out to examine 3D structural changes of chromatin during *Drosophila* embryogenesis. Specifically, we study a polycomb-repressed region and an active region of the same length in cells at different developmental stages, namely, embryos at cycles 9–13, stages 5–8, and S2R+. Overall, simulated heat maps of both regions show strong similarities to the corresponding heat maps of measured Hi-C (Supplementary Figure S4a for the repressed region and S4b for the active region, r=0.95–0.98, r'=0.71–0.76).

As our simulated ensembles contain 5.0×10^4 properly-sampled single-cell chromatin conformations (Figure 3g), we are able to quantify the heterogeneity of the cell population, which would not be possible for chromatin models based on a single consensus structure such as those in [39–41]. Using hierarchical clustering [42], we group single-cell chromatin conformations of all three cell types into 5 clusters (Figure 4a and 4b, see Supplementary Methods). We observe strong heterogeneity among different clusters of 3D single-cell conformations of the polycomb-repressed region in early embryos at cycles 9–13, with heterogeneity reduced somewhat at stages 5–8. In S2R+, two clusters dominate, each account for 41.1% and 55.7%, respectively. The largest cluster (C5) is the same for cycles 9–13 (31.7%) and S2R+ (55.7%), but a different cluster (C4) transiently become the largest cluster

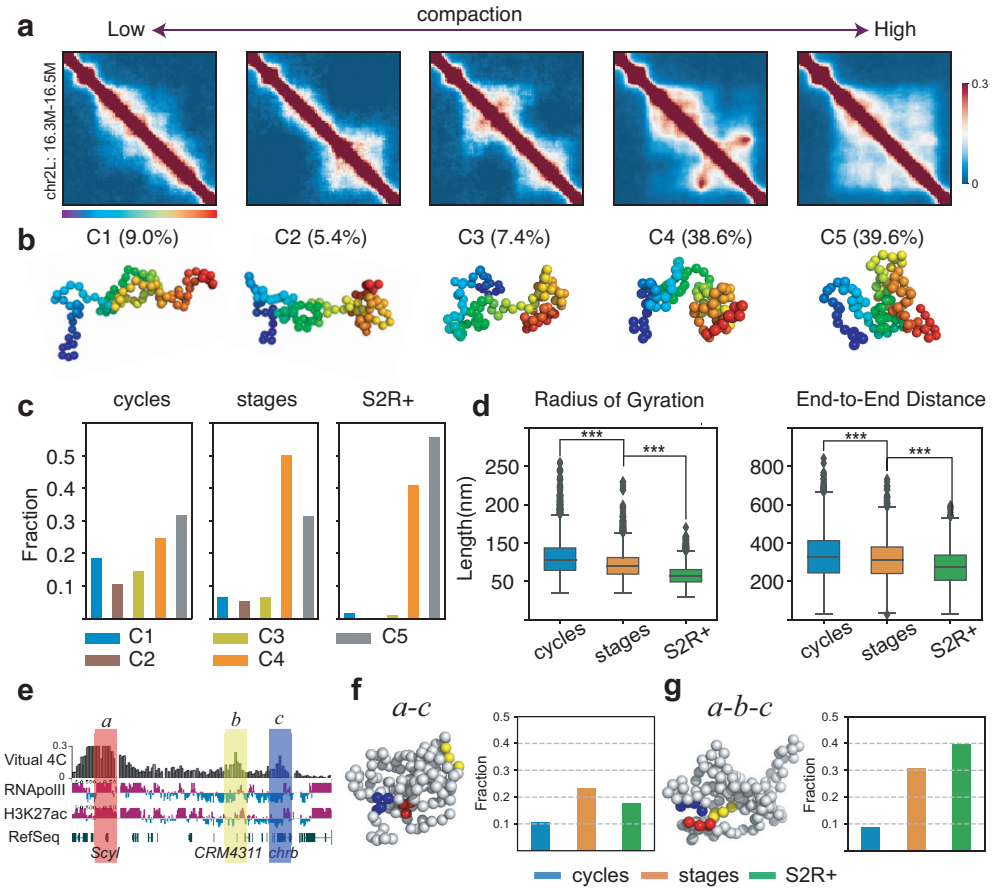


Figure 4. Heterogeneity in single-cell chromatin conformations, chromatin compaction, and dynamic changes of a three-body interaction unit during *Drosophila* embryogenesis. **(a)** Hierarchical clustering of 3D chromatin configurations of all three cell types in the polycomb-repressed region. Aggregated contact heat maps of the 5 clusters are ordered based on their averaged radius of gyration (R_g). **(b)** Representative conformations of the 5 clusters shown in **(a)**. Proportions are labeled on top of the conformations. **(c)** Proportions of the 5 clusters in each cell type. **(d)** Distributions of Radius of gyration (top) and End-to-End distance (bottom) of the region from **(a)**, *** represents two-sided Wilcoxon rank sum test p-value < 0.001. **(e)** A virtual 4C plot of the region that contains a three-body interaction among two promoters and a putative enhancer, with the anchor shown in red bar. Gene *Scyl* is located within this anchor region. Two target regions are shown in the blue bar and yellow bar. Gene *chrb* is localized in the blue region, a putative enhancer overlapped with a ChIP-defined cis-regulatory module (CRM) CRM4311 is in the yellow region. Tracks of RNA polymerase II, H3K27ac, and RefSeq genes are shown at the bottom. **(f)** Left figure shows an example of a two-body interaction between the promoters of gene *Scyl* and *chrb* which are labeled in **(e)**. Right barplot shows the fractions of this two-body interaction in three different cell types. **(g)** Left figure shows an example of a three-body interaction among the two promoters and the enhancer labeled in **(d)**. Right barplot shows the fractions of this three-body interaction among all *Scyl*-*chrb* interactions in the three cell types.

(50.1%) at stages 5–8 (Figure 4c). The active region shows a similar rich pattern of heterogeneity among single-cell chromatin conformations.

Increasing chromatin compactness of polycomb-repressed regions in late embryos

We observed dramatic changes of chromatin compaction in both active and repressed regions during embryogenesis. This is illustrated for the repressed region in Figure 4c, where later embryos have higher proportions of clusters (C4 and C5) with more compact chromatin conformations. The dramatic compaction is also demonstrated in the overall distributions of radius of gyration (R_g) [43] and end-to-end distance of the 3D ensembles in both the repressed (Figure 4d) and the active (Supplementary Figure S4c) region during embryogenesis.

We then couple chromatin chains from the active and the repressed region into one-to-one random pairs and calculate the ratios of R_g of active *vs.* repressed chromatin chain for each cell type. We found 78.4%, 85.1%, and 91.7% pairs have a larger R_g in the active region than in the repressed region in embryos at cycles 9–13, at stages 5–8, and in S2R+ cells, respectively. These results reflect progressively more shrinkage in space occupied by the chromatin in the repressed region compared to the active region in later embryos (Supplementary Figure S4d).

Overall, these results show our models enable detection of important 3D structural changes of chromatin during embryogenesis, which would not be possible by direct examination of experimentally measured Hi-C contacts.

Dynamic changes of a unit of functional three-body interactions during embryogenesis

Consistent with the observed compaction, a functional unit of three-body interactions is found to experience dynamic changes during embryogenesis, exemplifying the compaction. Gene *Scyl* and *chrb* are involved in head involution and many other biological functions [34]. There exists strong three-body interactions among the promoters of these two genes and a putative enhancer in this region [44] (Figure 4e).

We construct 3D ensembles of single-chain conformations of this region for the three cell types. The distribution of spatial distances between *chrb* and *Scyl* and the other two control regions derived from our model is highly consistent with DNA FISH measurements (Supplementary Figure S4g) [44]. Our simulations further reveal that embryos at stages 5–8 have higher *Scyl*–*chrb* (Figure 4f) and *Scyl*–enhancer (Supplementary Figure S4e) contact frequencies when compared to cells at the two other stages, while *chrb* and the enhancer contact more frequently in S2R+ cells (Supplementary Figure S4f). Despite a lower proportion of total *Scyl*-*chrb* interactions, we find an increased propensity of forming a three-body contact with the putative enhancer among these *Scyl*-*chrb* interactions in S2R+ cells (Figure 4g).

These results indicate that details of chromatin 3D structures in observed chromatin compaction can reveal important structural insight into promoter-enhancer interactions.

A large portion of single cells likely maintain TAD-like structures in early embryos

TADs are one of the most prevalent structural units of genome organization emerging from Hi-C studies [9–11]. There are clear TAD structures in Hi-C heat maps of embryos at stages 5–8 and S2R+ cells, but no obvious TADs in early embryos at cycles 9–13 (Figure 2a and 4a). However, Hi-C heat maps are reflection of ensemble-averaged properties. This naturally leads to the question whether TAD-like structures exist in single cells [22]. Specifically, we ask if there are TAD-like structures in early embryonic cells, and whether single-cell TAD structures are different among cells at different developmental stages.

We examine a 1 Mb region at chromosome 2L, from which several TADs can be found in Hi-C maps of late embryos (Figure 5a and 5d, Supplementary Figure S5a). The simulated contact maps from the aggregation of 5.0×10^4 single-cell chromatin conformations are highly similar to the corresponding measured Hi-C maps ($r = 0.97$, 0.95, and 0.95 for embryos at cycles 9–13, stages 5–8, and S2R+, respectively).

To characterize chromatin structures of single cells, we calculate the spatial distance between each pair of beads and generate a spatial-distance map for each single-cell chromatin conformation, following reference [19] (Figure 5c and 5f, Supplementary Figure S5c). There are clear TAD-like structures in the spatial-distance maps of many single-cell conformations of S2R+ (Figure 5f), consistent with studies of super-resolution imaging [19–21]. Surprisingly, we also detect strong presence of TAD-like structures in many single-cells of early embryos at

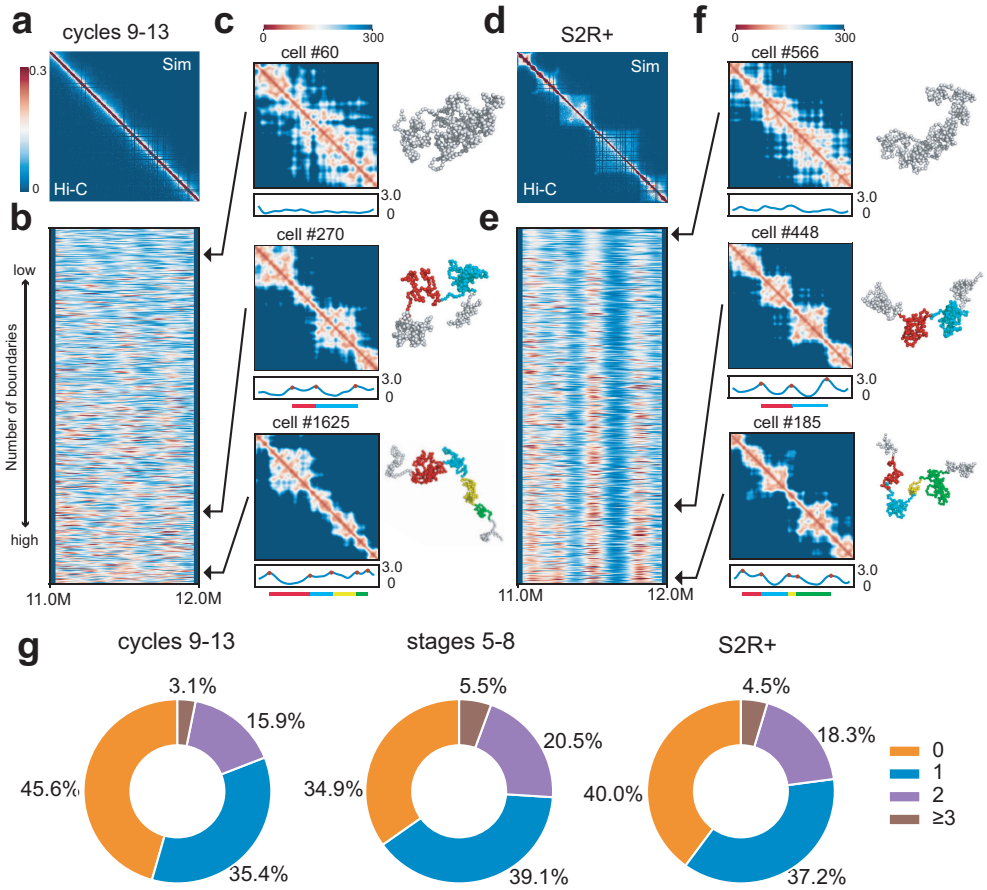


Figure 5. TAD-like structures in single cells during *Drosophila* embryogenesis. Simulated heat maps of the region (chr2L: 11.0–12.0 Mb) in embryos at cycles 9–13 **(a)** and S2R+ **(d)** are shown. Lower left triangles represent the experimental Hi-C propensities, upper right triangles represent the simulated contact probabilities. Resolution is 2 kb. Boundary strength profiles of 5,000 conformations in embryos at cycles 9–13 **(b)** and S2R+ **(e)** are shown below the combo heat maps. They are ordered by the number of domain boundaries. Three representatives of single-cell spatial-distance heat maps (on the left) and corresponding visualizations of conformations (on the right) in embryos at cycles 9-13 **(c)** and S2R+ **(f)** are shown. These single-cell conformations have different numbers of domain boundaries. The number of boundaries from top to bottom is 0, 3, and 5, respectively. Arrows indicate their positions in the boundary strength profiles in **(b)**. Boundary strength curves are drawn under the spatial-distance maps, with red dots representing the local maxima identified as domain boundaries. Bars in different colors under the boundary strength curves represent different domains identified in that conformation, which are also labeled in 3D visualizations. Proportions of single-cell conformations with different numbers of TADs are shown in **(g)** for each cell type.

cycles 9–13 (Figure 5c), even though no TADs can be seen in the ensemble Hi-C heat maps (Figure 5a).

We then examine the domain boundaries in single-cell conformations. We calculate the boundary strength at each genomic position using a spatial-distance ratio (see methods) and generate the boundary-strength curves (see blue curves in Figure 5c and 5f, Supplementary Figure S5c). Local maxima above a threshold of 2.2 in the curves are selected to be the single-cell domain boundaries (in red dots). Consecutive regions with BH-FDR adjusted p-values below 0.05 between pairs of adjacent boundaries are then identified as TAD-like structures (see Supplementary Methods). We randomly select 5,000 3D conformations from the ensemble of each cell type and identify their domain boundaries and TAD-like structures.

A large portion of single cells contain sharp domain boundaries in early embryos. More than half of single cells (54.4%) in embryos at cycles 9–13 possess at least one TAD-like structure. This is only slightly below that of S2R+, which has 60.0% single cells with at least one TAD-like structure (Figure 5g). These results suggest single-cell TAD-like structures may exist universally during *Drosophila* embryogenesis, even at the early developmental stage, where no clear TAD structures can be seen in ensemble Hi-C studies [33].

There is strong variability in boundary positions (Figure 5b and 5e, Supplementary Figure S5b) as well as domain sizes (Supplementary Figure S5d) among individual cells of early embryos at cycles 9-13. These variabilities in early embryos likely lead to the overall loss of concordant TAD structures in the ensemble-averaged Hi-C heat maps.

Predicted single-cell domain boundaries exhibit insulator-binding preference since MBT

In mammalian genomes, Hi-C studies showed that a large portion of TADs are found to be demarcated by insulator complexes of CCCTC binding factor (CTCF) and cohesin [11, 12]. A recent single-cell imaging study revealed that domain boundaries in individual mammalian cells can occur at any genomic positions, but preferentially at CTCF/cohesin binding sites [19]. To assess whether our predicted domain boundaries prefer insulator binding sites in single cells of *Drosophila*, we calculate the boundary probabilities along the linear genomic positions of the same region in Figure 5a for the three cell types. Results show that while there is no clear position preference for boundaries in early embryos at cycles 9–13, strong preference for specific genomic positions appear in embryos during and after the MBT, at stages 5–8 and S2R+ cells (Figure 6a). Despite the strong preference, there is a non-zero probability for any genomic position to be at a domain boundary in late embryos, which is consistent with previous findings of mammalian cells [19].

We then ask whether these preferred boundary positions appearing in late embryos correlate with binding sites of insulator proteins. However, we find no enrichment of boundary probabilities at the binding peaks of the two insulators of CTCF and cohesin (Figure 6b). This is consistent with earlier reports of *Drosophila* Hi-C studies [13,45]. We then examine binding sites of other insulator complexes with extensive presence in *Drosophila* [13, 46]. We find that domain boundaries in single cells of later embryos are highly enriched at the binding peaks of insulator complexes BEAF-32/CP190 and BEAF-32/Chromator (Figure 6a

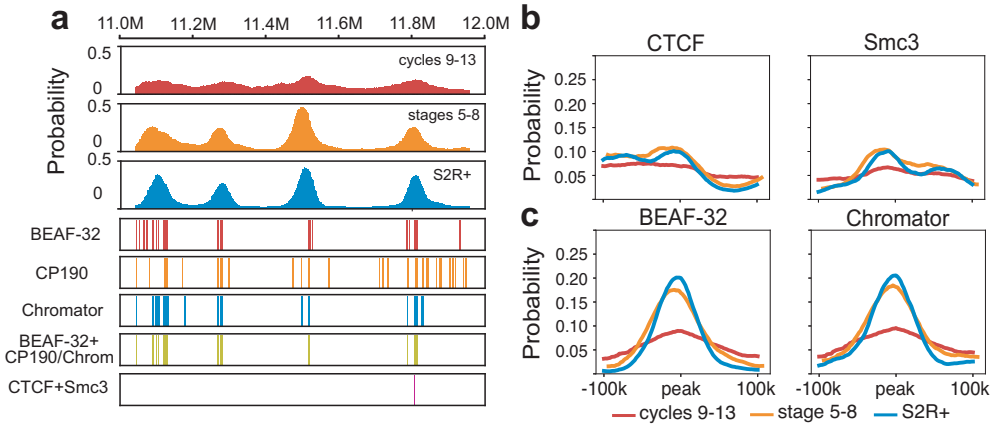


Figure 6. Domain boundaries are preferentially localized at insulator binding sites since MBT. **(a)** Distributions of domain boundary probabilities along genomic positions of the same region in Figure 5A (chr2L: 11.0–12.0 Mb) of three cell types are shown in the top three rows. Tracks in the 4-*th* to the 6-*th* row indicate the binding sites of 3 different insulator proteins of BEAF-32, CP190, and Chromator, respectively. The track in the 7-*th* row indicates the intersection of the binding sites between BEAF-32 and CP190 or between BEAF-32 and Chromator, the last track represents the intersection of the binding sites between CTCF and Smc3 (cohesin subunit). All tracks are from S2-DRSC. **(b)** Enrichment curves of the averaged domain boundary probabilities at the binding peaks of CTCF and Smc3. **(c)** Enrichment curves of the averaged domain boundary probabilities at the binding peaks of BEAF-32 and Chromator.

and 6c). Single-cell boundary probabilities are also enriched at binding peaks of ZW5 (Supplementary Figure S6).

These findings reveal the presence of insulator-binding preference of predicted domain boundaries in *Drosophila* single cells during and after MBT, suggesting that a conserved pattern of genome folding in single cells may exist between mammals and insects.

Discussion

Hi-C measurements of cell populations have provided a wealth of information on chromatin structures at high resolution [11, 13]. However, they do not provide specific details on chromatin structures in individual cells. Although single-cell Hi-C and super-resolution imaging studies can provide such details and can characterize the conserved and the varying features of chromatin structures among different cells [14–21], they are either highly sparse with limited coverage and resolution, or are restricted in number of cells, therefore do not reach the same high resolution or high throughput as that of population Hi-C.

The computational method described in this study can bridge the gap between high-resolution population Hi-C studies and fine-detailed single-cell 3D structures of chromatin. Our results on *Drosophila* show that population Hi-C data contains detailed information on structures of chromatins in individual cells, and such information can be effectively uncovered.

Our method also helps to answer an important question. While a large number

of chromatin contacts are identified in Hi-C studies, it is unclear which ones are required for chromatin structure formation and for biological functions. It is also unclear whether all contacts identified in single-cell studies are obligated for TAD formation or biological functions. It is well known that a large portion of Hi-C interactions are due to experimental biases and random collisions of chromatin fibers in a defined nucleus [24,25,28]. Existing computational methods for normalizing Hi-C data do not readily distinguish important interactions that are specific from background non-specific interactions: Some focus on eliminating experimental biases such as GC content and mapping qualities [47,48], others rely on null models constructed from the Hi-C data itself for removal of background noises [49,50,53], and none explicitly model polymer effects of random collision of chromatin fiber. As our method is based on construction of random self-avoiding 3D polymers in nuclear confinement (Figure 1), important specific interactions can be effectively identified.

Our results show that these specific interactions can be used to effectively reproduce the full patterns of Hi-C interactions, with excellent Pearson correlation ($r \sim 0.95$, Figure 3a and 3c). In contrast, ensembles generated using non-specific Hi-C contacts fail to capture much of the overall structures of chromatin folding.

By removing background noises due to random collision, clear biological patterns such as increasing Inactive-Inactive and Polycomb-Polycomb genomic interactions during embryogenesis emerge with clarity (Figure 2c and 2d). Furthermore, long-range interactions known to be biologically important are accurately uncovered (Figure 2b, Supplementary Figure S2b), and physical structures of loop interactions with fine details are gained (Figure 3d and 3e).

Our results also shed some light on the relationship of genome structure and function, as well as the origin of topological features such as TADs. As the size of the set of specific interactions is rather small, constituting only about 5–6% of Hi-C contacts in *Drosophila*, it is probable that there may exist a minimum set of interactions, likely to be functionally important, which are sufficient to drive chromosome folding and give rise to much of the topological features observed in Hi-C. That is, features such as TADs may arise naturally from a small set of functionally important chromatin interactions sufficient to drive chromatin folding. Furthermore, such driver interactions are likely to be contained within the set of 5–6% interactions identified, although it is possible that the actual number of driver interactions may be even smaller. Our results are consistent with a recent study probing the relationship of genome structure-function, where it was found functional interactions may play important roles in shaping genome structures [54].

Our method also addresses a long-standing challenge in 3D genome studies. As population-based Hi-C studies offer no directly interpretable information on the heterogeneity of 3D chromatin conformations of the cell population, and single-cell studies currently are limited to a small or a moderate number of cells, there is an overall lack of quantitative understanding on the heterogeneous nature of 3D chromatin structures in cell population [51, 52]. With highly efficient algorithm capable of deep sampling, our method can generate large proper ensembles of single-cell 3D conformations of loci and chromosomes. With the ability to converts 2D high-resolution population Hi-C heat maps into *ca.* 5.0×10^4 single-cell chromatin conformations at high resolution, detailed analysis on the specifics of individual chromatin of the ensemble can be carried out, from which the overall

heterogeneity of chromatins of the cell population can be quantified (Fig 4c).

Our method can facilitate identification of important changes in genome folding during *Drosophila* embryogenesis. These include dynamic changes of a functional unit of three-body interactions involving promoters of *Scyl* and *chrb* and a putative enhancer [44] (Fig 4d-f), and the observation of increasing chromatin compaction in later embryos, consistent with an earlier finding of stepwise formation of TADs and compartments during the embryo development [33, 36].

With detailed analysis of domain structures of each single cell of a large population, we are able to detect clear TAD-like structures with sharp domain boundaries in individual S2R+ cells, consistent with earlier imaging studies [19–21]. A surprising finding is that a large portion ($> 50\%$) of single cells may possess TAD-like structures in early embryos of cycles 9–13 (Figure 5a and 5c), where no clear TAD structures can be seen in previous ensemble Hi-C studies [33, 36]. This is also consistent with a previous chromatin modeling study, where it was found domain-like substructures arise from nuclear volume confinement [25], suggesting that finite volume promotes formation of generic TAD-like structures in early embryo single cells. Our findings demonstrate that *Drosophila* chromatin is topologically organized to varying extent, and the universal nature of the presence of TAD-like structures in single-cells during embryogenesis.

Our model of single cell chromatin structures also reveal characteristics of domain boundaries. Similar to mammalian single cells where CTCF and cohesin are preferentially localized at the domain boundaries [19], we detect a stronger insulator-binding preference in *Drosophila* single cells of later embryos during and after the midblastula transition. The single-cell domain boundaries are preferred

at insulator binding sites of BEAF-32/CP190 or BEAF-32/Chromator (Figure 6a), the same insulator complexes found at the TAD borders in ensemble Hi-C analysis [13, 46]. This insulator-binding preference is found in both mammalian and *Drosophila* single cells, suggesting a conserved characteristic of genome folding at the single-cell level.

Overall, our computational strategy is capable of identifying a small set of critical specific interactions from population Hi-C, which are sufficient to give rise to much of Hi-C observed topological features. These specific interactions can be used to construct large ensembles of high-resolution single-cell chromatin conformations. Our method can facilitate discovery of biological patterns at both single-cell and population levels. As demonstrated in the analysis of *Drosophila*, our method can provide new perspectives on genome 3D structural changes during important biological processes such as embryogenesis, enabling discoveries that would not be possible with traditional Hi-C analysis. Our method is robust and can be used for arbitrary loci and for full chromosomes. With quality Hi-C data becoming more abundant, applications of our method can aid in overall understanding of the mechanisms of genome folding and can help to decipher the structure-function relationship of genomes.

Methods

Hi-C propensities

Hi-C data are obtained from the GEO database (embryos at cycles 9-13 and stages 5-8 from GSE103625, S2R+ from GSE101317) and are mapped to the dm3 reference genome following [13]. Hi-C contact matrices generated at 2/5 kb resolution are normalized using ICE [48]. Assuming neighboring regions always form Hi-C ligations [27], Hi-C propensities are calculated as $p_{\text{obs}}(i,\ j) = \frac{C(i,\ j)}{\mathbb{E}_{diag}(1)}$, where $C(i,\ j)$ is the Hi-C contact frequency of loci i and j, $\mathbb{E}_{\text{diag}}(1)$ is the averaged contact frequency of Hi-C pairs $(i,\ j)$ with $|i-j| = 1$ bin.

Model parameters and contact model

We model random chromatin fibers as self-avoiding polymer chains consisting of beads, each represents a 2 kb or 5 kb genomic region, the same as the resolution of the Hi-C matrices. We assume beads-on-string [55] chromatin fiber has a mass density of 165 bp/11 nm [56], thus the bead diameter is roughly 25 nm. The spherical volume within which the polymer chains are confined is proportional to the nuclear volume of each cell type. We choose 292 μm^3 as the volume for S2R+ [57], 335 μm^3 for embryos at cycles 9–13 [58] and 524 μm^3 for embryos at stages 5–8 [58]. The genome size is approximately 700 Mb for tetraploid S2R+ and 350 Mb for diploid embryos at cycles 9-13 and stages 5-8 [59, 60].

We assume regions that are in close proximity are available for Hi-C ligation.

Contact probabilities are then calculated as the proportions of single-cell conformations satisfying the distance requirement, namely, the distances of the pair of loci of interests are within a threshold, which is the longest distance for ligation.

Constructing physical null model of chromatin chains

We generate random chromatin polymer chains using a novel Monte Carlo approach (see Supplementary Methods) and construct an ensemble of 2×10^5 random polymer chains within a defined space for each type of cell (Figure 1a). They are used as our null model to estimate contact probabilities p_{null} of random collisions which lead to non-specific Hi-C interactions. p_{null} were defined as

$$p_{\text{null}}\left(i,\ j\right) = \frac{\sum_{k=1}^{N}\left[\mathbb{I}^{(k)}\left(i,\ j\right)w^{(k)}\right]}{\sum_{k=1}^{N}w^{(k)}} \quad ,$$

where $\mathbb{I}^{(k)}\left(i,j\right)$ is an indicator function of 1 if the distance between i and j in the k-th chain is $< d_c$, with $d_c = 80$ nm [27, 28], N is the total number of polymer chains, and $w^{(k)}$ is the importance weight of the k-th chain for bias-correction due to deviations of the sampling distribution from the target uniform distribution in our null model.

Identification of specific interactions

For each ensemble of random polymer chains of a cell type, we assign a statistical p-value to each pair of loci based on the percentage of random contact probabilities in bootstrap replicates that exceed the relative Hi-C propensity. We use

Bag of Little Bootstrap (BLB) [61] (see Supplementary Methods) to generate a total of 5,000 bootstrap ensembles (Figure 1b). Although each BLB ensemble contains only a small subset ($\sim$ 1,300 polymer chains, or $\leq 1\%$) of the original ensemble, the average physical properties of the BLB ensembles reflect that of the original polymer ensemble [61]. For each BLB ensemble, $p_{\text{null}}(i,j)$ is calculated as described above. After quantile normalization of $p_{\text{null}}(i,j)$ and $p_{\text{obs}}(i,j)$, we assign a p-value to each pair of loci (i,j) according to the percentage of $p_{\text{null}}(i,j)$ that exceed $p_{\text{obs}}(i,j)$,

$$p_{ij} = \frac{\sum_{k=1}^{M} \mathbb{I}\left[p_{\text{null}}^{(k)}(i,\ j) < p_{\text{obs}}(i,\ j)\right]}{M} \quad ,$$

where $\mathbb{I}(\cdot)$ is a indicator function of 1 if the specified condition is satisfied. M is the total number of bootstrap ensembles. Here we have $M = 5000$. Hi-C interactions with BH-FDR [30] adjusted p-values <0.01 are chosen to be the specific interactions (Figure 1c).

Deep sampling of ensembles of chromatin structures using Hi-C constraints

We generate 3D chromatin structures under a sequential Bayesian inference framework using constraints of specific, non-specific or all Hi-C interactions. To generate an ensemble E from the Hi-C data, our goal is to maximize the probability $P(E|H)$, where H represents Hi-C propensities selected as modeling constraints, and E consists of chromatin polymers $X^{(1)}, X^{(2)}, \cdots, X^{(N)}$. By Bayes' rule [62], $P(E|H) = \frac{P(H|E)P(E)}{P(H)}$, with $P(H)$ being a constant. We generate chromatin

polymers sequentially:

$$
\begin{aligned}
P\left(E|H\right) &= P\left(E_1|H_1\right) P\left(E_2|H_2\right) \cdots P\left(E_n|H_n\right) \\
&= \prod_{t=1}^{n} P\left(E_t|H_t\right) \quad \propto \quad \prod_{t=1}^{n} P\left(H_t|E_t\right) P\left(E_t\right) \ ,
\end{aligned}
$$

where n is the length of each chain, H_t the selected Hi-C propensities $p_{\text{obs}}\left(x_1, x_t\right)$, $p_{\text{obs}}\left(x_2, x_t\right), \cdots,\ p_{\text{obs}}\left(x_{t-1}, x_t\right)$. E_t is the intermediate ensemble $X_t^{(1)}, X_t^{(2)}, \cdots, X_t^{(N)}$ at the step t, each chain consists of $(t-1)$ beads that are previously placed and a newly generated bead $x_t^{(k)}$.

$P\left(H_t|E_t\right)$ evaluates the similarities between simulated contact probabilities derived from chromatin conformations and Hi-C propensities. We model this term through a Poisson distribution, which is robust and can decrease the influence of the dominance of large contact counts [63]. $P\left(E_t\right)$ is reversely proportional to the number of all possible valid intermediate ensembles E_t given the previously constructed E_{t-1}. We apply an iterative optimization strategy to find the best polymer ensemble that maximize $P\left(E_t|H_t\right)$ at each growing step (See Supplementary Methods).

Identification of single-cell domain boundaries

We adopt a method similar to [19] to define the domain boundaries in single cells. Details can be found in Supplementary Methods.

Author contributions

QS, ZS and JL conceived and designed the study. QS designed and implemented the sequential Bayesian inference model. APR designed and implemented the Fractal Monte Carlo for null model. QS carried out computation and analysis with assistance from APR. DMC participates in data analysis. JL and ZS supervised the overall study. QS and JL wrote the manuscript with assistance from APR, DMC and ZS. All authors read and approved the final manuscript.

Acknowledgements

We thank Ni Jian for assistance in figure editing and Gamze Gürsoy for useful suggestions.

Funding

This work is supported by grants NIH R21AI126308, R21AI137865, R01CA204962, and R35GM127084 (to JL); National Key Research and Development Program of China grant 2018YFC1003501, CAS-18H100000104 and NSFC-81627801 (to ZS).

Data availability

Hi-C data are downloaded from GEO database (embryos at cycles 9-13 and stages 5-8 from GSE103625, S2R+ from GSE101317). ChIP-chip datasets for clustering are downloaded from modENCODE [64] database with IDs listed in Supplementary Table S2.

Code availability

Source code for null model chromatin folding by fractal Monte Carlo is available via git repository at `https://bitbucket.org/aperezrathke/chr-folder`. Source code for Sequential Bayesian inference framework is available via `https://github.com/qiusun0215/sBIF`.

References

[1] Jonathan Flint, Karen Thomas, Gos Micklem, Helen Raynham, Kevin Clark, Norman A Doggett, Andrew Andrew, and Douglas R Higgs. The relationship between chromosome structure and function at a human telomeric region. *Nature genetics*, 15(3):252, 1997.

[2] Geoff Fudenberg, Gad Getz, Matthew Meyerson, and Leonid A Mirny. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nature biotechnology*, 29(12):1109, 2011.

[3] Tom Sexton and Giacomo Cavalli. The role of chromosome domains in shaping the functional genome. *Cell*, 160(6):1049–1059, 2015.

[4] Benjamin D Pope, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L Vera, Yanli Wang, R Scott Hansen, Theresa K Canfield, et al. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402, 2014.

[5] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *science*, 295(5558):1306–1311, 2002.

[6] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo De Wit, Bas Van Steensel, and Wouter De Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature genetics*, 38(11):1348, 2006.

[7] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, et al. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10):1299–1309, 2006.

[8] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[9] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381, 2012.

[10] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376, 2012.

[11] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[12] Chin-Tong Ong and Victor G Corces. Ctcf: an architectural protein bridging genome topology and function. *Nature reviews Genetics*, 15(4):234, 2014.

[13] Qi Wang, Qiu Sun, Daniel M Czajkowsky, and Zhifeng Shao. Sub-kb hi-c in d. melanogaster reveals conserved characteristics of tads between insect and mammalian cells. *Nature communications*, 9(1):188, 2018.

[14] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59, 2013.

[15] Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661):61, 2017.

[16] Ilya M Flyamer, Johanna Gassler, Maxim Imakaev, Hugo B Brandão, Sergey V Ulianov, Nezar Abdennur, Sergey V Razin, Leonid A Mirny, and Kikuë Tachibana-Konwalski. Single-nucleus hi-c reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648):110, 2017.

[17] Tim J Stevens, David Lando, Srinjan Basu, Liam P Atkinson, Yang Cao, Steven F Lee, Martin Leeb, Kai J Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, et al. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, 544(7648):59, 2017.

[18] Siyuan Wang, Jun-Han Su, Brian J Beliveau, Bogdan Bintu, Jeffrey R Moffitt, Chao-ting Wu, and Xiaowei Zhuang. Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, 353(6299):598–602, 2016.

[19] Bogdan Bintu, Leslie J Mateo, Jun-Han Su, Nicholas A Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N Boettiger, and Xiaowei Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413):eaau1783, 2018.

[20] Quentin Szabo, Daniel Jost, Jia-Ming Chang, Diego I Cattoni, Giorgio L Papadopoulos, Boyan Bonev, Tom Sexton, Julian Gurgo, Caroline Jacquier, Marcelo Nollmann, et al. Tads are 3d structural units of higher-order chromosome organization in drosophila. *Science advances*, 4(2):eaar8082, 2018.

[21] Leslie J Mateo, Sedona E Murphy, Antonina Hafner, Isaac S Cinquini, Carly A Walker, and Alistair N Boettiger. Visualizing dna folding and rna in embryos at single-cell resolution. *Nature*, 568(7750):49, 2019.

[22] Jesse R Dixon, David U Gorkin, and Bing Ren. Chromatin domains: the unit of chromosome organization. *Molecular cell*, 62(5):668–680, 2016.

[23] Bin Zhang and Peter G Wolynes. Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences*, 112(19):6062–6067, 2015.

[24] Andrew S Belmont. Large-scale chromatin organization: the good, the surprising, and the still perplexing. *Current opinion in cell biology*, 26:69–78, 2014.

[25] Gamze Gürsoy, Yun Xu, Amy L Kenter, and Jie Liang. Spatial confinement is a major determinant of the folding landscape of human chromosomes. *Nucleic acids research*, 42(13):8223–8230, 2014.

[26] Nicholas Allen Kinney, Igor V Sharakhov, and Alexey V Onufriev. Chromosome–nuclear envelope attachments affect interphase chromosome territories and entanglement. *Epigenetics & chromatin*, 11(1):3, 2018.

[27] Luca Giorgetti, Rafael Galupa, Elphège P Nora, Tristan Piolot, France Lam, Job Dekker, Guido Tiana, and Edith Heard. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157(4):950–963, 2014.

[28] Gamze Gürsoy, Yun Xu, Amy L Kenter, and Jie Liang. Computational construction of 3d chromatin ensembles and prediction of functional interactions

of alpha-globin locus from 5c data. *Nucleic acids research*, 45(20):11547–11558, 2017.

[29] Michele Di Pierro, Ryan R Cheng, Erez Lieberman Aiden, Peter G Wolynes, and José N Onuchic. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences*, 114(46):12126–12131, 2017.

[30] Yoav Benjamini, Daniel Yekutieli, et al. The control of the false discovery rate in multiple testing under dependency. *The annals of statistics*, 29(4):1165–1188, 2001.

[31] Wolfgang Mayer, Alain Niveleau, Jörn Walter, Reinald Fundele, and Thomas Haaf. Embryogenesis: demethylation of the zygotic paternal genome. *Nature*, 403(6769):501, 2000.

[32] Miler T Lee, Ashley R Bonneau, and Antonio J Giraldez. Zygotic genome activation during the maternal-to-zygotic transition. *Annual review of cell and developmental biology*, 30:581–613, 2014.

[33] Yuki Ogiyama, Bernd Schuettengruber, Giorgio L Papadopoulos, Jia-Ming Chang, and Giacomo Cavalli. Polycomb-dependent chromatin looping contributes to gene silencing during drosophila development. *Molecular cell*, 71(1):73–88, 2018.

[34] Anne Scuderi, Karl Simin, Sandra G Kazuko, James E Metherall, and Anthea Letsou. scylla and charybde, homologues of the human apoptotic gene rtp801, are required for head involution in drosophila. *Developmental biology*, 291(1):110–122, 2006.

[35] Peter V Kharchenko, Artyom A Alekseyenko, Yuri B Schwartz, Aki Minoda, Nicole C Riddle, Jason Ernst, Peter J Sabo, Erica Larschan, Andrey A Gorchakov, Tingting Gu, et al. Comprehensive analysis of the chromatin landscape in drosophila melanogaster. *Nature*, 471(7339):480, 2011.

[36] Clemens B Hug, Alexis G Grimaldi, Kai Kruse, and Juan M Vaquerizas. Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell*, 169(2):216–228, 2017.

[37] Simona Bianco, Darío G Lupiáñez, Andrea M Chiariello, Carlo Annunziatella, Katerina Kraft, Robert Schöpflin, Lars Wittler, Guillaume Andrey, Martin Vingron, Ana Pombo, et al. Polymer physics predicts the effects of structural variants on chromatin architecture. *Nature genetics*, 50(5):662, 2018.

[38] Jeffrey A Farrell and Patrick H O'Farrell. From egg to gastrula: how the cell cycle is remodeled during the drosophila mid-blastula transition. *Annual review of genetics*, 48:269–294, 2014.

[39] Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, C Anthony Blau, and William S Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363, 2010.

[40] Chenchen Zou, Yuping Zhang, and Zhengqing Ouyang. Hsa: integrating multi-track hi-c data for genome-scale reconstruction of 3d chromatin structure. *Genome biology*, 17(1):40, 2016.

[41] Lila Rieber and Shaun Mahony. minimds: 3d structural inference from high-resolution hi-c data. *Bioinformatics*, 33(14):i261–i266, 2017.

[42] Ellen M Voorhees. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management*, 22(6):465–476, 1986.

[43] Marshall Fixman. Radius of gyration of polymer chains. *The Journal of Chemical Physics*, 36(2):306–310, 1962.

[44] Yad Ghavi-Helm, Felix A Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber, and Eileen EM Furlong. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, 512(7512):96, 2014.

[45] Sergey V Ulianov, Ekaterina E Khrameeva, Alexey A Gavrilov, Ilya M Flyamer, Pavel Kos, Elena A Mikhaleva, Aleksey A Penin, Maria D Logacheva, Maxim V Imakaev, Alexander Chertovich, et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome research*, 26(1):70–84, 2016.

[46] Jutta Vogelmann, Antoine Le Gall, Stephanie Dejardin, Frederic Allemand, Adrien Gamot, Gilles Labesse, Olivier Cuvier, Nicolas Negre, Martin Cohen-Gonsaud, Emmanuel Margeat, et al. Chromatin insulator factors involved in long-range dna interactions and their role in the folding of the drosophila genome. *PLoS genetics*, 10(8):e1004544, 2014.

[47] Eitan Yaffe and Amos Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43(11):1059, 2011.

[48] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999, 2012.

[49] Ferhat Ay, Timothy L Bailey, and William Stafford Noble. Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome research*, 24(6):999–1011, 2014.

[50] Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3):458–472, 2012.

[51] Lei Liu, Guang Shi, D Thirumalai, and Changbong Hyeon. Chain organization of human interphase chromosome determines the spatiotemporal dynamics of chromatin loci. *PLoS computational biology*, 14(12):e1006617, 2018.

[52] Daniel A Beard and Tamar Schlick. Computational modeling predicts the structure and dynamics of chromatin fiber. *Structure*, 9(2):105–114, 2001.

[53] Fulai Jin, Yan Li, Jesse R Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D Schmitt, Celso A Espinoza, and Bing Ren.

A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290, 2013.

[54] Yad Ghavi-Helm, Aleksander Jankowski, Sascha Meiers, Rebecca R Viales, Jan O Korbel, and Eileen EM Furlong. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature genetics*, page 1, 2019.

[55] JT Finch and A Klug. Solenoidal model for superstructure in chromatin. *Proceedings of the National Academy of Sciences*, 73(6):1897–1901, 1976.

[56] Kensal E Van Holde. *Chromatin*. Springer Science & Business Media, 2012.

[57] Jeffrey M Verboon, Hector Rincon-Arano, Timothy R Werwie, Jeffrey J Delrow, David Scalzo, Vivek Nandakumar, Mark Groudine, and Susan M Parkhurst. Wash interacts with lamin and affects global nuclear organization. *Current Biology*, 25(6):804–810, 2015.

[58] Annely Brandt, Fani Papagiannouli, Nicole Wagner, Michaela Wilsch-Bräuninger, Martina Braun, Eileen E Furlong, Silke Loserth, Christian Wenzl, Fanny Pilot, Nina Vogt, et al. Developmental control of nuclear size and shape by kugelkern and kurzkern. *Current biology*, 16(6):543–552, 2006.

[59] Hangnoh Lee, C Joel McManus, Dong-Yeon Cho, Matthew Eaton, Fioranna Renda, Maria Patrizia Somma, Lucy Cherbas, Gemma May, Sara Powell, Dayu Zhang, et al. Dna copy number evolution in drosophila cell lines. *Genome biology*, 15(8):R70, 2014.

[60] Lisa L Ellis, Wen Huang, Andrew M Quinn, Astha Ahuja, Ben Alfrejd, Francisco E Gomez, Carl E Hjelmen, Kristi L Moore, Trudy FC Mackay, J Spencer Johnston, et al. Intrapopulation genome size variation in d. melanogaster reflects life history variation and plasticity. *PLoS genetics*, 10(7):e1004522, 2014.

[61] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

[62] Herbert Robbins. An empirical bayes approach to statistics. *Herbert Robbins Selected Papers*, pages 41–47, 1956.

[63] Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. A statistical approach for inferring the 3d structure of the genome. *Bioinformatics*, 30(12):i26–i33, 2014.

[64] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, et al. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22(9):1813–1831, 2012.

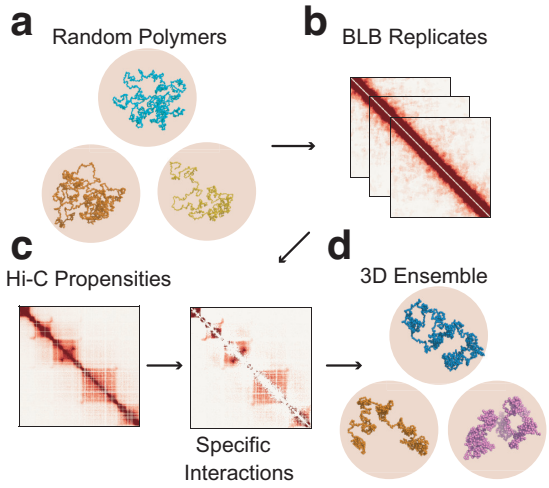


Figure 1. Overview of our chromatin modeling strategy. **(a)** We first construct ensemble of random chromatin polymers as our physical null model. We then **(b)** bootstrap the polymer ensemble to compute statistical significance of each Hi-C interaction. After removal of non-specific interactions, **(c)** we construct an ensemble of 3D chromatin structures constrained by specific interactions we identified under a sequential Bayesian inference framework.

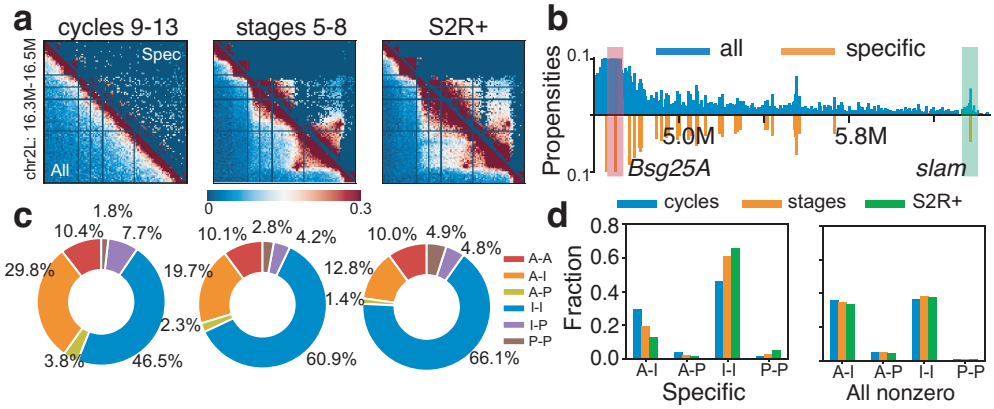


Figure 2. Overview of specific Hi-C interactions. **(a)** Heat maps of specific interactions identified in a polycomb-repressed region (chr2L: 16.3–16.5 Mb) of three cell types at different developmental stages. Lower-left triangles represent all Hi-C interactions, upper-right triangles represent the identified specific interactions. Cell types from top to bottom are embryos at cycles 9-13, embryos at stages 5–8, and S2R+, respectively. **(b)** A virtual 4C plot of the distribution of specific interactions in a 1.86 Mb region of embryos at cycles 9-13. Red bar represents the anchor which contains the gene *Bsg25A*, green bar represents a specific interaction that targets gene *slam*. **(c)** Pie charts of percentages of different types of specific interactions in the three cell types. A: active, I: inactive, P: polycomb-repressed. I-I interactions increase from 46.5% to 60.9% and then to 66.1%, P-P increases from 1.8% to 2.8% and then to 4.9%, while A-I decreases from 29.8% to 19.7%, then to 12.8%, and A-P decreases from 3.8% to 2.3%, then to 1.4%. **(d)** Percentages of four interaction types of specific interactions (left) and all nonzero Hi-C interactions (right). These four types are A-I, A-P, I-I and P-P, respectively.

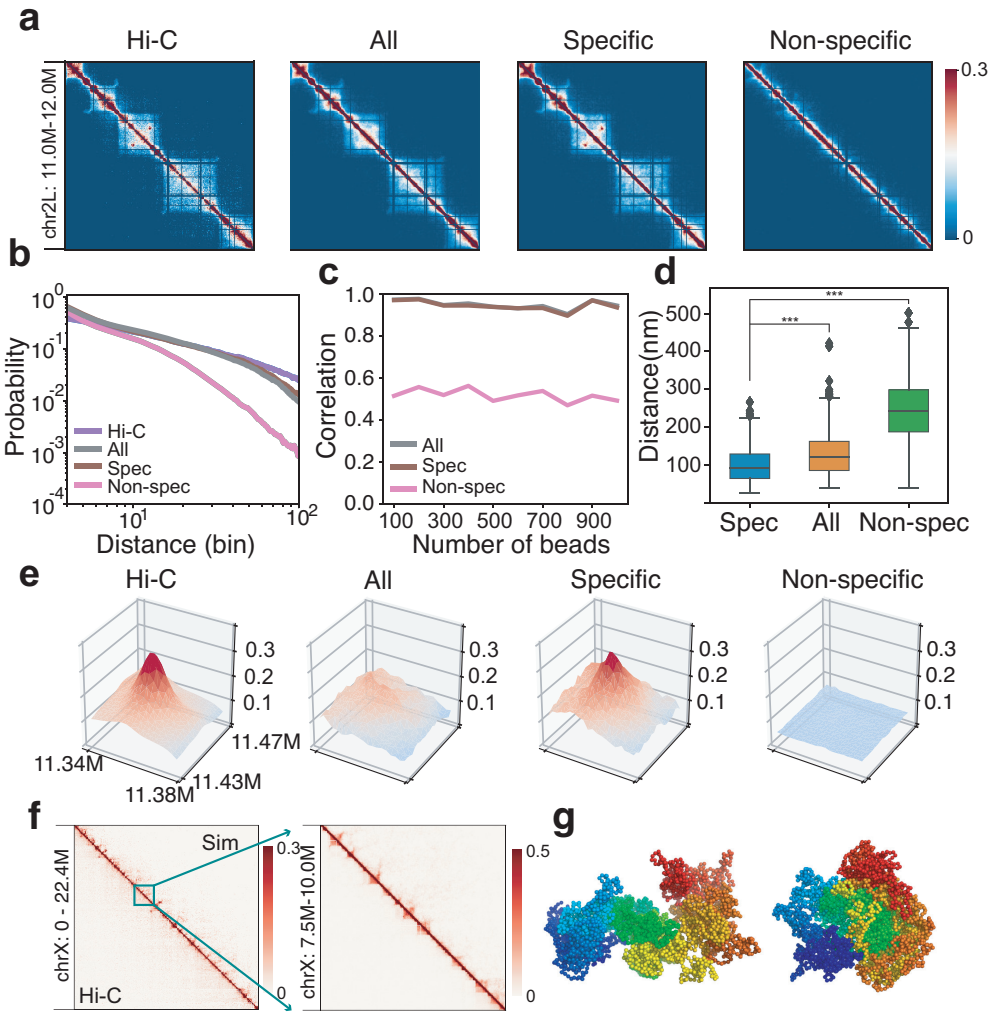


Figure 3. Specific interactions are sufficient to drive chromatin folding in *Drosophila*. **(a)** Illustration of simulation results for a 1 Mb region (chr2L: 11.0–12.0 Mb) of S2R+ cells at 2 kb resolution. Heat maps from left to right represent Hi-C propensities, simulated contact probabilities using all, specific, and non-specific interactions, respectively. **(b)** log-log scaling curves of contact probabilities with genomic distances (bin) derived from the original Hi-C data, simulated ensembles using specific, all, and non-specific interactions. **(c)** Pearson correlation coefficients of the simulated contact probabilities and Hi-C propensities in 10 regions of different genomic lengths. The number of beads ranges from 100 to 1,000. **(d)** Distance distributions of the two anchors of the loop shown in **(a)**. Loop anchors correspond to bead No. 170 and bead No. 190. *** represents two-sided Wilcoxon rank sum test p-value ≤ 0.001. **(e)** Height maps of contact probability of the loop interaction inside the polycomb domain shown in **(a)**. **(f)** Constructed 3D conformations of chromosome X using specific interactions at 5 kb resolution. (Left) Simulated Hi-C heat map, with measured Hi-C propensities at the bottom and simulated contact probabilities at the top. (Right) Zoomed-in heat map of a 2.5 Mb region. Pearson correlation coefficient r is 0.94, distance-adjusted correlation r' is 0.64. **(g)** Visualization of two examples of 3D conformations of chromosome X using PyMOL.

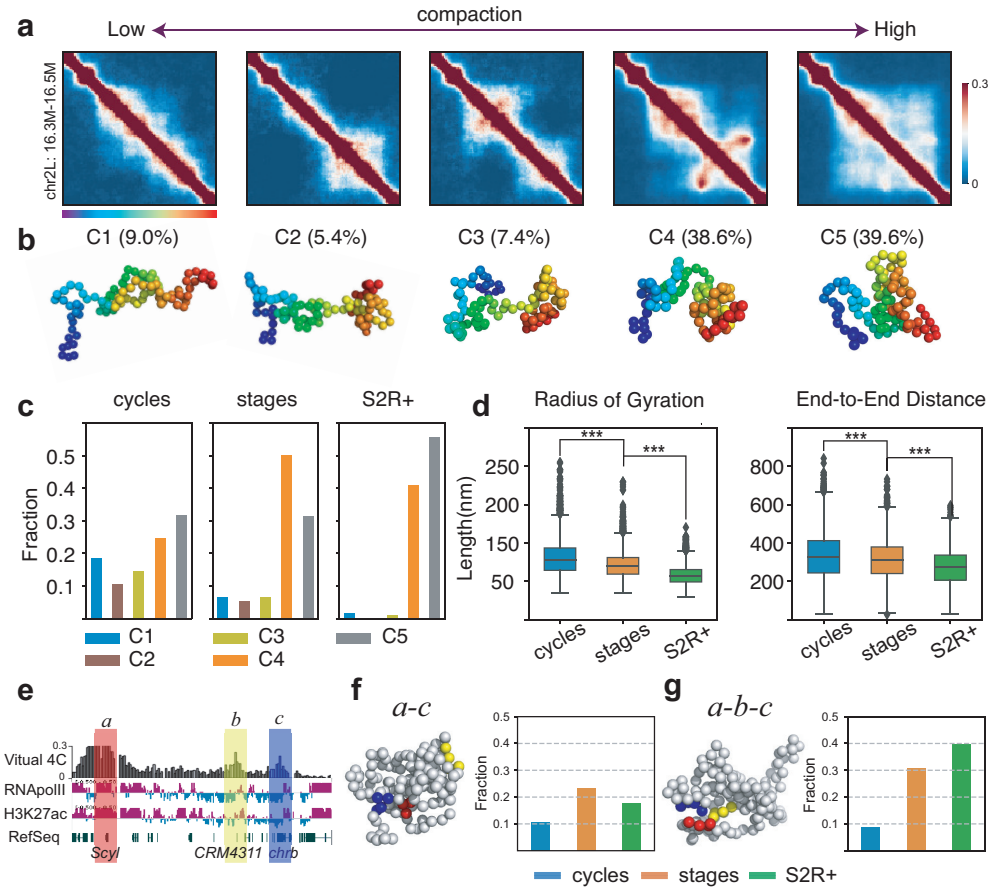


Figure 4. Heterogeneity in single-cell chromatin conformations, chromatin compaction, and dynamic changes of a three-body interaction unit during *Drosophila* embryogenesis. **(a)** Hierarchical clustering of 3D chromatin configurations of all three cell types in the polycomb-repressed region. Aggregated contact heat maps of the 5 clusters are ordered based on their averaged radius of gyration (R_g). **(b)** Representative conformations of the 5 clusters shown in **(a)**. Proportions are labeled on top of the conformations. **(c)** Proportions of the 5 clusters in each cell type. **(d)** Distributions of Radius of gyration (top) and End-to-End distance (bottom) of the region from **(a)**, *** represents two-sided Wilcoxon rank sum test p-value < 0.001. **(e)** A virtual 4C plot of the region that contains a three-body interaction among two promoters and a putative enhancer, with the anchor shown in red bar. Gene *Scyl* is located within this anchor region. Two target regions are shown in the blue bar and yellow bar. Gene *chrb* is localized in the blue region, a putative enhancer overlapped with a ChIP-defined cis-regulatory module (CRM) CRM4311 is in the yellow region. Tracks of RNA polymerase II, H3K27ac, and RefSeq genes are shown at the bottom. **(f)** Left figure shows an example of a two-body interaction between the promoters of gene *Scyl* and *chrb* which are labeled in **(e)**. Right barplot shows the fractions of this two-body interaction in three different cell types. **(g)** Left figure shows an example of a three-body interaction among the two promoters and the enhancer labeled in **(d)**. Right barplot shows the fractions of this three-body interaction among all *Scyl*-*chrb* interactions in the three cell types.

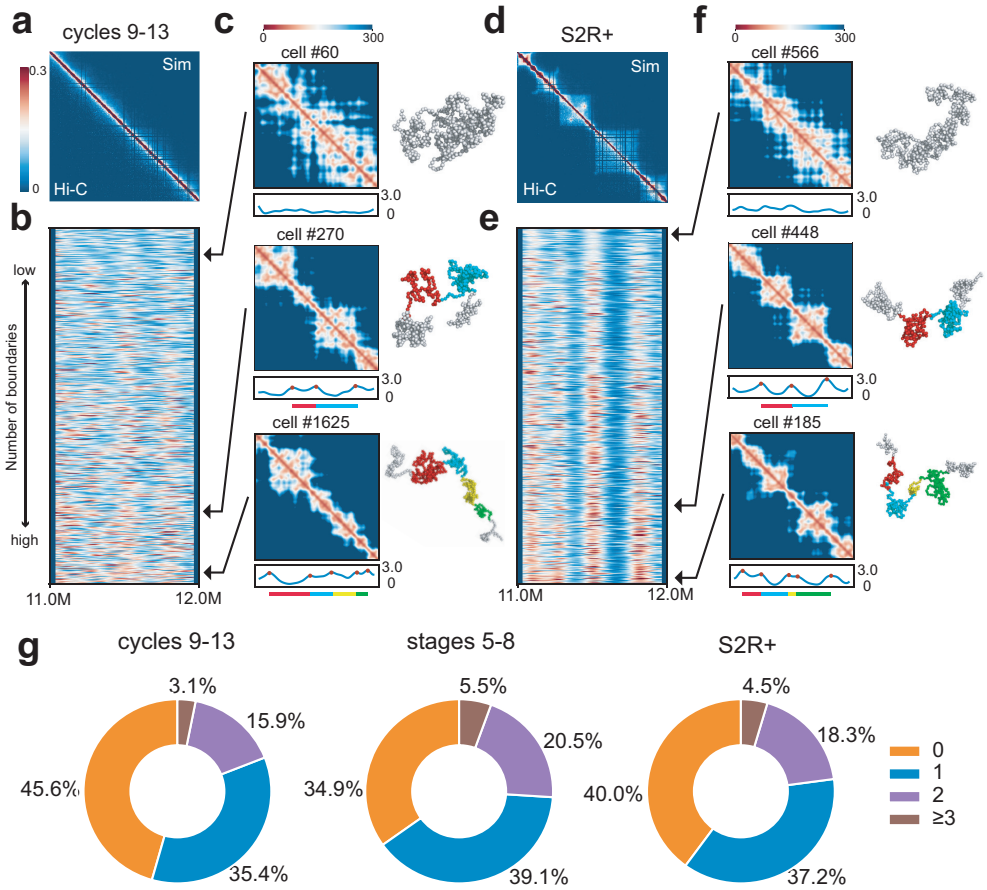


Figure 5. TAD-like structures in single cells during *Drosophila* embryogenesis. Simulated heat maps of the region (chr2L: 11.0–12.0 Mb) in embryos at cycles 9–13 **(a)** and S2R+ **(d)** are shown. Lower left triangles represent the experimental Hi-C propensities, upper right triangles represent the simulated contact probabilities. Resolution is 2 kb. Boundary strength profiles of 5,000 conformations in embryos at cycles 9–13 **(b)** and S2R+ **(e)** are shown below the combo heat maps. They are ordered by the number of domain boundaries. Three representatives of single-cell spatial-distance heat maps (on the left) and corresponding visualizations of conformations (on the right) in embryos at cycles 9-13 **(c)** and S2R+ **(f)** are shown. These single-cell conformations have different numbers of domain boundaries. The number of boundaries from top to bottom is 0, 3, and 5, respectively. Arrows indicate their positions in the boundary strength profiles in **(b)**. Boundary strength curves are drawn under the spatial-distance maps, with red dots representing the local maxima identified as domain boundaries. Bars in different colors under the boundary strength curves represent different domains identified in that conformation, which are also labeled in 3D visualizations. Proportions of single-cell conformations with different numbers of TADs are shown in **(g)** for each cell type.

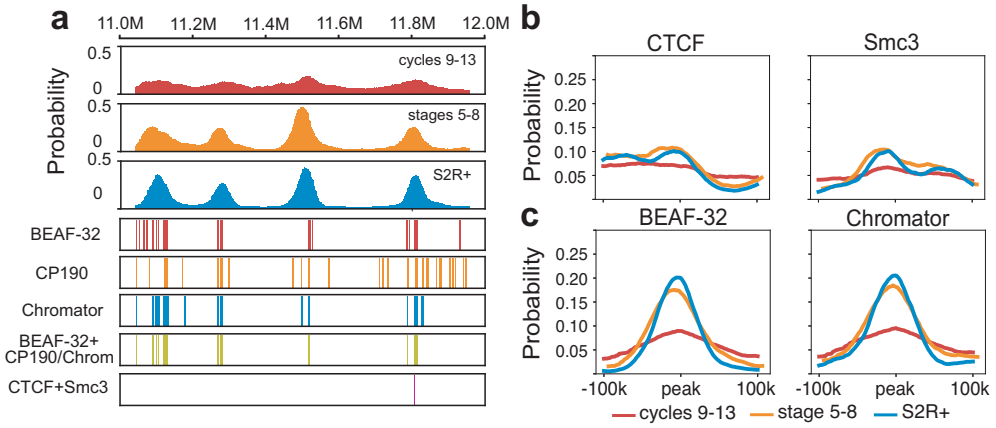
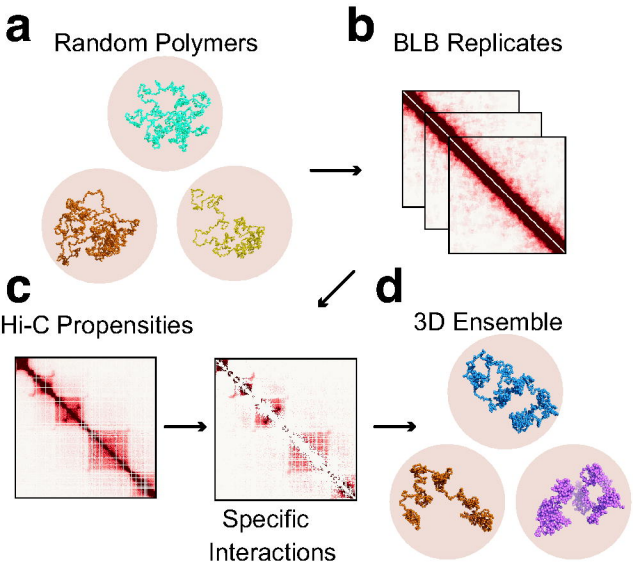
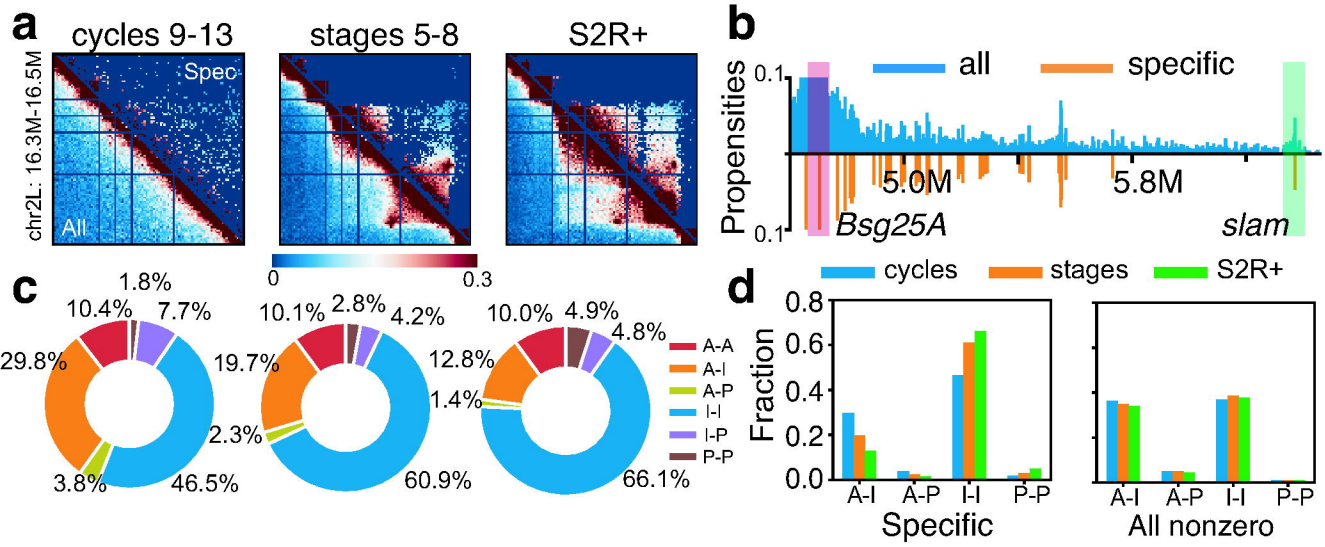
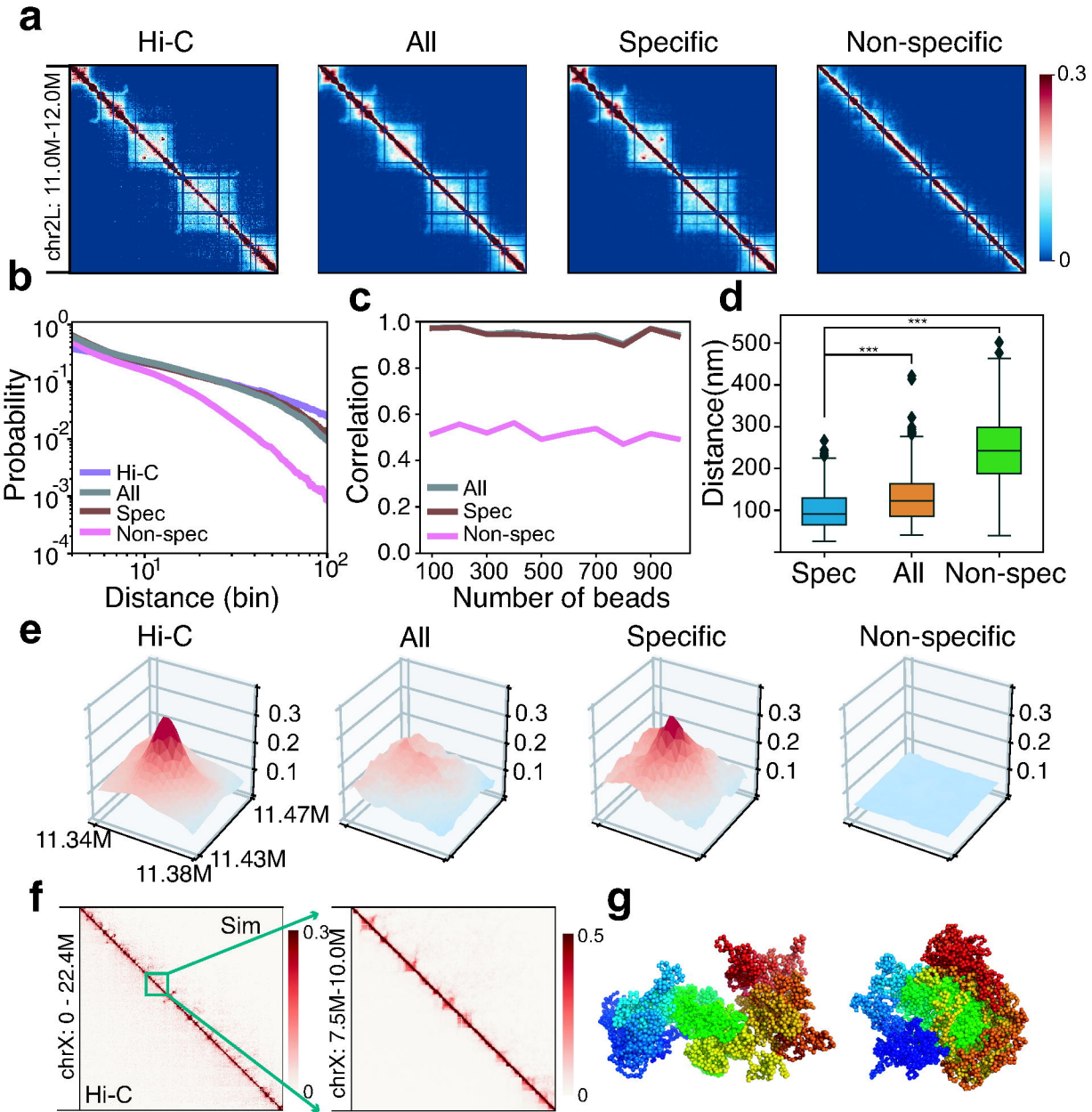
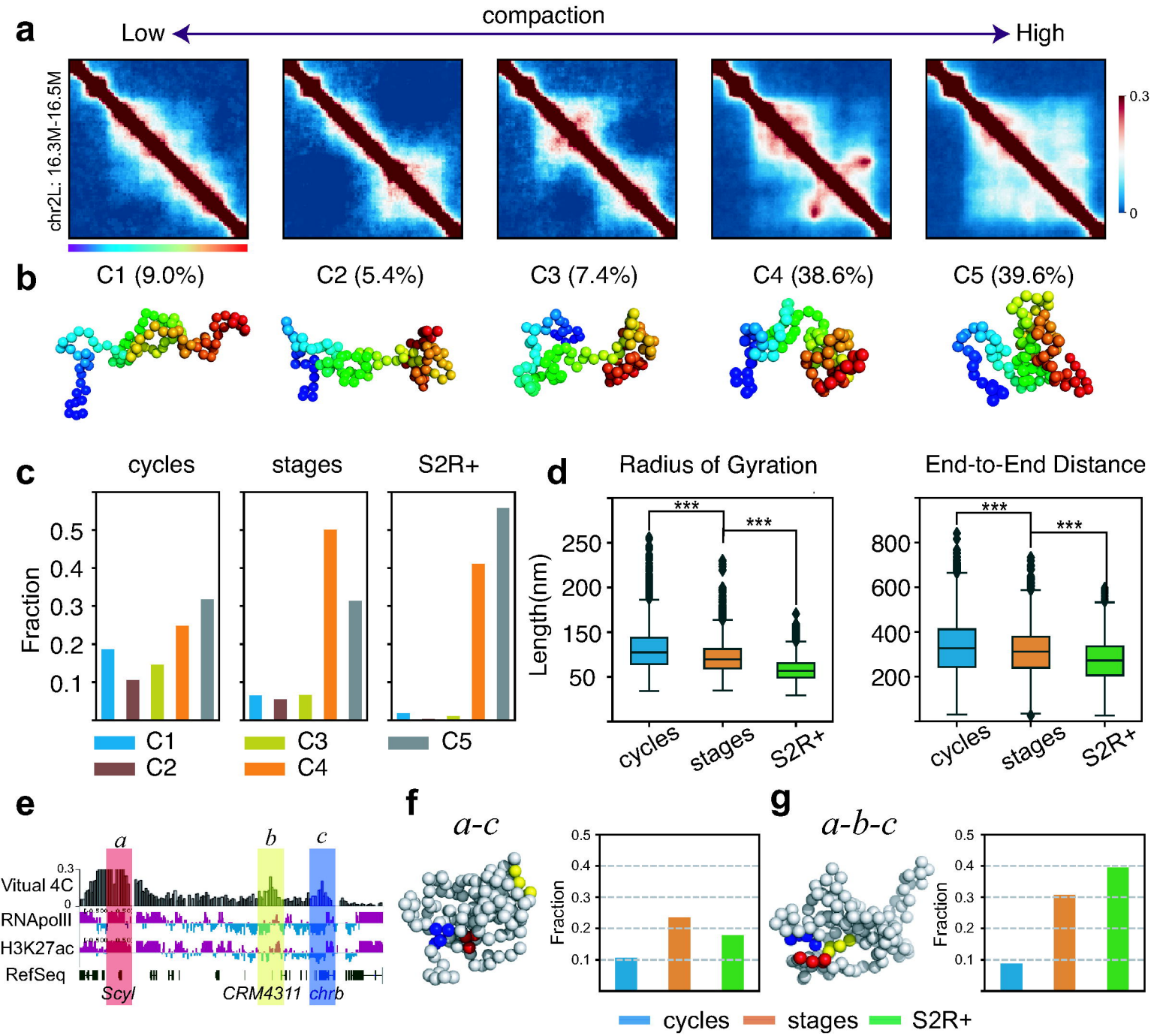


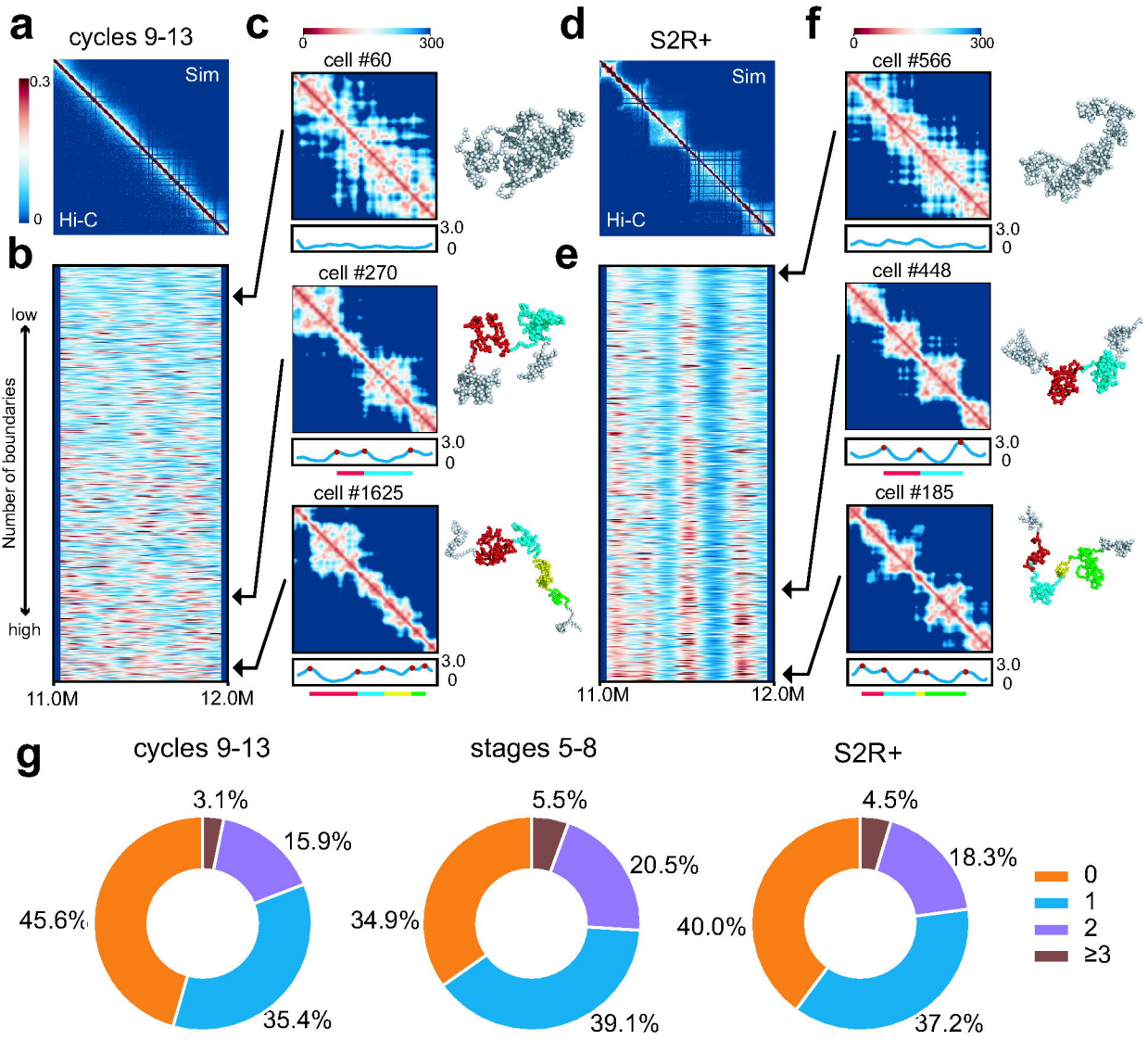
Figure 6. Domain boundaries are preferentially localized at insulator binding sites since MBT. **(a)** Distributions of domain boundary probabilities along genomic positions of the same region in Figure 5A (chr2L: 11.0–12.0 Mb) of three cell types are shown in the top three rows. Tracks in the 4-*th* to the 6-*th* row indicate the binding sites of 3 different insulator proteins of BEAF-32, CP190, and Chromator, respectively. The track in the 7-*th* row indicates the intersection of the binding sites between BEAF-32 and CP190 or between BEAF-32 and Chromator, the last track represents the intersection of the binding sites between CTCF and Smc3 (cohesin subunit). All tracks are from S2-DRSC. **(b)** Enrichment curves of the averaged domain boundary probabilities at the binding peaks of CTCF and Smc3. **(c)** Enrichment curves of the averaged domain boundary probabilities at the binding peaks of BEAF-32 and Chromator.

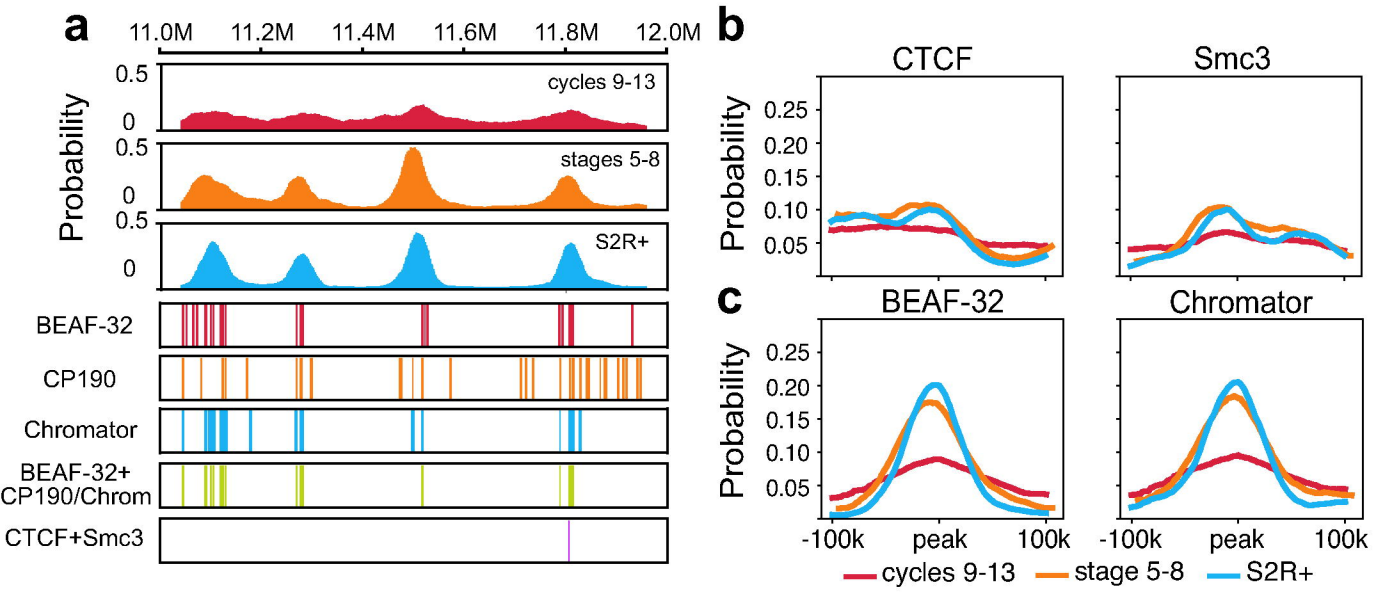

a
Random Polymers
b
BLB Replicates
c
Hi-C Propensities
d
3D Ensemble
Specific
Interactions


a
cycles 9-13
stages 5-8
S2R+
chr2L: 16.3M-16.5M
Spec
All
0
0.3
b
all
specific
Propensities
0.1
0.1
5.0M
5.8M
Bsg25A
slam
c
1.8%
10.4%
7.7%
29.8%
3.8%
46.5%
10.1%
2.8%
4.2%
19.7%
2.3%
60.9%
10.0%
4.9%
4.8%
12.8%
1.4%
66.1%
A-A
A-I
A-P
I-I
I-P
P-P
d
cycles
stages
S2R+
Fraction
0.8
0.6
0.4
0.2
0.0
A-I
A-P
I-I
P-P
Specific
All nonzero


a
Hi-C
All
Specific
Non-specific
chr2L: 11.0M-12.0M
0.3
0
b
Probability
Distance (bin)
Hi-C
All
Spec
Non-spec
c
Correlation
Number of beads
All
Spec
Non-spec
d
Distance(nm)
Spec
All
Non-spec
e
Hi-C
All
Specific
Non-specific
11.34M
11.38M
11.43M
11.47M
f
Sim
Hi-C
chrX: 0 - 22.4M
chrX: 7.5M-10.0M
g


a
Low
compaction
High
chr2L: 16.3M-16.5M
0.3
0
b
C1 (9.0%)
C2 (5.4%)
C3 (7.4%)
C4 (38.6%)
C5 (39.6%)
c
cycles
stages
S2R+
Fraction
0.5
0.4
0.3
0.2
0.1
C1
C2
C3
C4
C5
d
Radius of Gyration
End-to-End Distance
Length(nm)
250
200
150
50
800
600
400
200
***
cycles
stages
S2R+
e
a
b
c
Vitual 4C
RNApolII
H3K27ac
RefSeq
Scyl
CRM4311
chrb
f
a-c
Fraction
g
a-b-c
cycles
stages
S2R+


a
cycles 9-13
Sim
Hi-C
0.3
0
b
low
Number of boundaries
high
11.0M
12.0M
c
0
300
cell #60
3.0
0
cell #270
cell #1625
d
S2R+
e
f
cell #566
cell #448
cell #185
g
cycles 9-13
3.1%
15.9%
45.6%
35.4%
stages 5-8
5.5%
20.5%
34.9%
39.1%
S2R+
4.5%
18.3%
40.0%
37.2%
0
1
2
≥3


a
11.0M
11.2M
11.4M
11.6M
11.8M
12.0M
Probability
0.5
0
cycles 9-13
stages 5-8
S2R+
BEAF-32
CP190
Chromator
BEAF-32+
CP190/Chrom
CTCF+Smc3
b
CTCF
Smc3
0.25
0.20
0.15
0.10
0.05
c
BEAF-32
Chromator
-100k
peak
100k
cycles 9-13
stage 5-8
S2R+