The 3D genome organization of *Drosophila melanogaster* through data integration

Qingjiao Li[1,a], Harianto Tjong[1,a], Xiao Li[1], Ke Gong[1], Xianghong Jasmine Zhou[1], Irene Chiolo[1,*], Frank Alber[1,*]

[1]Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA

*Correspondence should be addressed to F.A. (alber@usc.edu) and I.C. (chiolo@usc.edu).

[a]These authors contributed equally.

Abstract

Genome structures are dynamic and non-randomly organized in the nucleus of higher eukaryotes. To maximize the accuracy and coverage of 3D genome structural models, it is important to integrate all available sources of experimental information about a genome's organization. It remains a major challenge to integrate such data from various complementary experimental methods. Here, we present an approach for data integration to determine a population of complete 3D genome structures that are statistically consistent with data from both genome-wide chromosome conformation capture (Hi-C) and lamina-DamID experiments. Our structures resolve the genome at the resolution of topological domains, and reproduce simultaneously both sets of experimental data. Importantly, this framework allows for structural heterogeneity between cells, and hence accounts for the expected plasticity of genome structures. As a case study we choose *Drosophila melanogaster* embryonic cells, for which both data types are available. Our 3D geome structures have strong predictive power for structural features not directly visible in the initial data sets, and reproduce experimental hallmarks of the *D. melanogaster* genome organization from independent and our own imaging experiments. Also they reveal a number of new insights about the genome organization and its functional relevance, including the preferred locations of heterochromatic satellites of differnet chromosomes, and observations about homologous pairing that cannot be directly observed in the original Hi-C or lamina-DamID data. To our knowledge our approach is the first that allows systematic integration of Hi-C and lamina DamID data for complete 3D genome structure calculation, while also explicitly considering genome structural variability.

Keywords: 3D genome structure, higher order genome organization, population-based modeling, data integration, Hi-C, lamina-DamID, homologous pairing, *Drosophila melanogaster*.

Introduction

It has become increasingly clear that a chromosome's three-dimensional organization influences the regulation of gene expression and other genome functions. Early microscopy and biochemical studies showed that chromosomes in higher eukaryotes form distinct territories, which although stochastically organized tend to be located at preferred positions within the nucleus. For example, lamina-DamID experiments have identified specific chromatin domains with a high propensity to be located at the nuclear envelope (NE), confirming the important role of the NE in spatial genome organization and gene regulation in Drosophila, human and mouse [1-3]. Chromosome conformation capture experiments (Hi-C and variants) detect chromatin interactions at genome-wide scale [4-10] and reveal a hierarchical chromosome organization: the chromatin can be segmented into domains, which in turn combine to form subcompartments of functionally related chromatin [10-12]. Topological associated domains (TADs) are defined by observing an increased probability of interaction between chromatin regions in a domain relative to interactions between domains. In addition, it has been shown that the border regions between domains are enriched in specific insulator proteins, such as CTCF and ZNF143 in mammalian cells and BEAF, CTCF and CP190 in Drosophila cells. However, the precision of domain border detection depend to some extent on the sequencing depth as well as algorithmic parameter settings. At increased sequencing depth it is possible to detect reliably individual chromatin loops, which often demarcate contact domains (at ~100kb domain length) [5].

Computational approaches can aid in mapping the global 3D structures of genomes at various scales [13-19]. These are currently divided into data-driven and *de novo*

modeling techniques [20]. Data-driven models use experimental information, often Hi-C data, to generate 3D genome structures that are constrained to be consistent with the data. Data-driven models can be further subdivided into two classes. The first represents the genome as a consensus structure most consistent with the data. This approach often assumes an anti-correlation between the Hi-C contact probability of two chromatin regions and their average spatial distance. In contrast, the second class of methdos explicitly models the large variability of genome structures between isogenic cells (even within a sample of synchronized cells) by creating a population of thousands of model structures, in which the accumulated chromatin contacts in all structures reproduce the observed Hi-C matrix rather than each structure individually. These approaches do not need to assume any functional relationship between contact frequencies and spatial distances.

We introduced a method for population-based modeling to analyze the structure of complete diploid genomes from Hi-C data [6, 21, 22]. Our method uses an iterative, probabilistic optimization framework to deconvolve the Hi-C data into a population of individual structures by inferring cooperative chromatin interactions that are likely to co-occur in the same cells. Our method generates a large number of genome structures whose chromatin contacts in the models are statistically consistent with those from the Hi-C data. Other ensemble-based methods have been introduced and applied to individual chromosomes or chromatin domains [13, 19].

So far, computational models of genome structures have typically relied on just one data type, such as Hi-C, even though a single experimental method cannot capture all aspects of the spatial genome organization. However, data are available from a wide

range of technologies with complementary strengths and limitations. Integrating all these different data types would greatly increase the accuracy and coverage of genome structure models. Moreover, such models would offer a way to cross-validate the consistency of data obtained from complementary technologies. For example, lamina-DamID experiments show a chromatin region's probability to be close to the lamina at the nuclear envelope [23, 24], while Hi-C experiments reveal the probability that two chromatin regions are in spatial proximity. Large-scale 3D fluorescence in situ hybridization (FISH) experiments show the distance between loci directly, and can be used to measure the distribution of distances across a population of cells.

It remains a major challenge to develop hybrid methods that can systematically integrate data from many different technologies to generate structural maps of the genome. In this paper, we present a method for integrating data from Hi-C and lamina-DamID experiments to maximize the accuracy of population-based 3D genome structural models. We apply this approach to model the diploid genome of Drosophila.

Drosophila melanogaster is a popular model organism to study the organization and functional relevance of 3D genome structure, owing to its relative small genome and the availability of many genetic tools. A variety of microscopy-based experiments have already studied the nuclear organization of *D. melanogaster*, and elucidated some regulatory mechanisms [25-29]. For example, the pairing of homologous chromosomes has been observed in the somatic cells of *D. melanogaster* and other dipteran insects [30-33]. This kind of pairing can influence gene expression by forming interactions between regulatory elements on homologous chromosomes, a process called transvection [26, 34]. Although transvection is common in Drosophila, not every gene

region with homologue pairing shows transvection. Therefore, questions remain as to whether somatic homolog pairing has other regulatory roles. In Drosophila, the centromeres tend to cluster and relocate close to the periphery of the nucleolus during interphase [35]. Centromere clustering is also observed in many other organisms, including yeast, mouse and human, and this process is thought to play an important role in determining the overall genome architecture [36, 37].

Over the past ten years, high-throughput genetic and genomic techniques have generated genome-wide maps of histone modifications, transcription factor binding, and chromatin interactions for *D. melanogaster* [1, 7, 8, 38, 39]. Pickersgill *et al.* used lamina-DamID experiments combined with a microarray technique to detect the binding signals of genome-wide chromatin to the lamina matrix [1]. Around 500 genes were detected to interact with the lamina. These genes were transcriptionally silenced and late-replicating. Pickersgill *et al.* then used FISH experiments to confirm that the lamina-targeted loci were more frequently located at the nuclear envelope than other loci. Recently, genome-wide chromatin contacts have been determined for 16-18 hr Drosophila embryos using the Hi-C technique [8]. The euchromatin genome was divided into 1169 physical domains based on Hi-C interaction profiles. These physical domains (which would be referred to as topological associated domain, or TADs, in mammalian cells) were assigned to four functional classes based on their average epigenetic signatures: Null, Active, Polycomb-Group (Pc-G) and HP1/Centromere.

Despite all this work, the global 3D nuclear architecture of the *D. melanogaster* genome is still unknown. Because both Hi-C and lamina-DamID data are available for Drosophila embryonic cells, we use these data to test our integration method. Each diploid genome

structure in our population-based model is defined by the 3D positions of all 1169 TADs. The structures are generated by optimizing a likelihood function, so that the ensemble is statistically consistent with both the experimentally derived contact probabilities between all chromatin domains from Hi-C data and the probability that a given chromatin domain is close to the NE from lamina-DamID data.

We validated our 3D genome models against independent experimental data and known structural features. Our models confirm the formation of distinct chromosome territories, with relatively low rates of intermingling between chromosomes [40, 41]. In addition, our models often show a polarized organization of chromosomes in the nucleus [27, 42, 43]. Analysis of the model population leads to a number of new insights about the nuclear organization of *D. melanogaster* and its functional relevance. For instance, our models reveal the preferred locations of heterochromatin and the nucleolus, which we were able to confirm by 3D FISH experiments. The nucleolus serves as an anchor for chromosomes, and is surrounded by pericentromeric heterochromatin. The distance of pericentromeric heterochromatin regions from the periphery varies by chromosome, with chromosomes 4 and X heterochomatin more peripheral relative to Pericentromeric regions of other chromosomes. Interestingly, the frequency of homologous pairing varies along the chromosomes with the lowest frequencies observed in our models for domains enriched in protein binding sites for Mrg15. These observations support the model that Mrg15 plays a role in the dissociation of homologous chromosome pairs during interphase, as previously suggested [44]. Finally, the structure population suggests that homologous

chromosome pairing plays a functional role in transcriptional activity and DNA replication program.

Results

Population-based genome structure modeling from data integration

Our goal is to determine a population of 3D genome structures for *Drosophila melanogaster* that is consistent with data from Hi-C and lamina-DamID experiments. Suppose $\mathbf{A}$ is a probability matrix derived from Hi-C data, and E is a probability vector derived from lamina-DamID data. The elements of $\mathbf{A}$ describe how frequently a given pair of TADs are in contact with each other in an ensemble of cells, and E describes how frequently a given TAD is in contact with the nuclear envelope (NE). The goal is to generate a population of genome structures $\mathbf{X}$, whose TAD-TAD and TAD-NE contact frequencies are statistically consistent with both $\mathbf{A}$ and E. We formulate the genome structure modeling problem as a maximization of the likelihood $P\left(\mathbf{A}, E \middle| \mathbf{X}\right)$.

More specifically, the structure population is defined as a set of *M* diploid genome structures $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_M\}$, where the *m*-th structure $\mathbf{X}_m$ is a set of 3-dimensional vectors representing the center coordinates of 2*N* domain spheres $\mathbf{X}_m = \left\{\vec{x}_{im} : \vec{x}_{im} \in \Re^3, \quad i = 1,\ 2\ ...,\ 2N\right\}$. *N* is the number of TAD domains, but each domain has two homologous copies with different coordinates. The contact probability matrix $\mathbf{A} = \left(a_{IJ}\right)_{N \times N}$ for *N* domains is derived from the Hi-C data, which do not distinguish between homologous copies. Each element a_{IJ} is the probability that a direct contact between domains *I* and *J* exists in a structure of the population. (Note that the capital

letter indices *I* and *J* refer to domains without distinguishing between their homologous copies, while the lowercase indices *i, i'* and *j, j'* do distinguish between copies.) The contact probability vector $E=\left\{e_I \middle| I=1,2,...,N\right\}$ is derived from the lamina-DamID data, and defines the probability for each TAD to be localized at the nuclear envelope (NE). With known $\mathbf{A}$ and E, we calculate the structure population $\mathbf{X}$ such that the likelihood $P\left(\mathbf{A},E\middle|\mathbf{X}\right)$ is maximized.

The Hi-C and lamina-DamID experiments provide data that is averaged over a large population of cells, so they cannot reveal which contacts co-exist in the same 3D structure. Therefore, both $\mathbf{A}$ and E are interpreted as ensemble averages. To represent information derived from individual cells, we introduce two latent variables $\mathbf{W}$ and $\mathbf{V}$. The "contact indicator tensor" $\mathbf{W}=(w_{ijm})_{2N\times 2N\times M}$ is a binary, 3rd-order tensor. It contains the information missing from the Hi-C data $\mathbf{A}$, namely which domain contacts belong to each of the M structures in the model population and also which homologous chromosome copies are involved ($w_{ijm}=1$ indicates a contact between domain spheres *i* and *j* in structure *m*; $w_{ijm}=0$ otherwise). $\mathbf{W}$ is a detailed expansion of $\mathbf{A}$ into a diploid, single-structure representation of the data. The structure population $\mathbf{X}$ is consistent with $\mathbf{W}$. Therefore, the dependence relationship between these three variables is given as $\mathbf{X}\rightarrow\mathbf{W}\rightarrow\mathbf{A}$. Another latent variable, $\mathbf{V}=(v_{im})_{2N\times M}$, specifies which domain is located near the NE in each structure of the population and also distinguishes between the two homologous TAD copies ($v_{im}=1$ indicates that TAD *i* is located near the NE in structure *m*; $v_{im}=0$ otherwise). The dependence relationship between $\mathbf{X}$, $\mathbf{V}$ and E is

given as $\mathbf{X} \to \mathbf{V} \to E$, because $\mathbf{X}$ is the structure population consistent with $\mathbf{V}$ and $\mathbf{V}$ is a detailed expansion of E at a diploid and single-structure representation of the data.

In addition to the Hi-C and lamina-DamID data, we also consider additional information specific for Drosophila genome organization, e.g. the nuclear volume, an upper bound for homolog chromosome pairing, constraints connecting consecutive domains (including heterochromatin domains) as well as costraints for anchoring centromeres to the nucleolus (see the detailed description in the **Materials and Methods** section).

Thus, the optimization problem is expressed as:

$$\widehat{\mathbf{X}} = \arg\max_{\mathbf{X},\mathbf{W},\mathbf{V}} \log P(\mathbf{A}, E, \mathbf{W}, \mathbf{V} \mid \mathbf{X})$$
$$\text{subject to} \begin{cases} \text{spatial constraint I: nuclear volume constraints} \\ \text{spatial constraint II: excluded volume constraints} \\ \text{spatial constraint III: chromosome pairing upper bound} \\ \text{spatial contraint IV: consecutive domain constraint} \end{cases} \quad (1)$$

The log likelihood can be expanded as

$$\begin{aligned} \log P(\mathbf{A}, E, \mathbf{W}, \mathbf{V} \mid \mathbf{X}) &= \log P(\mathbf{A}, E \mid \mathbf{W}, \mathbf{V}) P(\mathbf{W}, \mathbf{V} \mid \mathbf{X}) \\ &= \log P(\mathbf{A} \mid \mathbf{W}) P(E \mid \mathbf{V}) P(\mathbf{W}, \mathbf{V} \mid \mathbf{X}) \end{aligned} \quad (2)$$

We have developed a variant of the EM method to iteratively optimize the log likelihood [22]. Each iteration consists of two steps (**Fig. 1A**):

- Assignment step (*A-step*): Given the current model $\mathbf{X}^{(i)}$, estimate the latent variables $\mathbf{W}^{(i+1)}$ and $\mathbf{V}^{(i+1)}$ by maximizing the log-likelihood over all possible values of $\mathbf{W}$ and $\mathbf{V}$.

$$\mathbf{W}^{(i+1)}, \mathbf{V}^{(i+1)} = \arg\max_{\mathbf{W},\mathbf{V}} \log P(\mathbf{A} \mid \mathbf{W}) P(E \mid \mathbf{V}) P(\mathbf{W}, \mathbf{V} \mid \mathbf{X}^{(i)}) \quad (3)$$

- Modeling step (*M-step*): Given the current estimated latent variables $\mathbf{W}^{(i+1)}$ and $\mathbf{V}^{(i+1)}$, find the model $\mathbf{X}^{(i+1)}$ that maximizes the log-likelihood function.

$$\mathbf{X}^{(i+1)} = \arg\max_{\mathbf{X}} \log P(\mathbf{A} \mid \mathbf{W}^{(i+1)}) P(E \mid \mathbf{V}^{(i+1)}) P(\mathbf{W}^{(i+1)}, \mathbf{V}^{(i+1)} \mid \mathbf{X}) \quad (4)$$

The detailed implementation of the A-step and M-step are described in **Materials and Methods**. We follow the step-wise optimization strategy described previously [22], and gradually increase the optimization hardness by adding contact constraints at a decreasing contact probability threshold.

A population of Drosophila genome structures at the TAD level

The euchromatin regions of *Drosophila melanogaster* chromosomes 2, 3, 4, and X are partitioned into 1169 TADs, as previously described [8]. The region of pericentromeric heterochromatin of each chromosome arm is spatially clustered and represented by a single domain (**Fig. 1B**) [45-47] (**Materials and Methods**). The nuclear diameter is set to 4 microns. The model also contains a nucleolus, represented by a sphere with a radius 1/6 of the nuclear radius. We estimated the nucleolus volume from our immunofluorescence analysis of Drosophila Kc cells (**Fig. S7A**) (**Materials and Methods**).

By optimizing the likelihood function (Eq. (1)) we generated a population of 10,000 genome structures that accurately reproduces the domain contact probabilities from Hi-C experiments and the probabilities for domains to reside at the nuclear envelope (NE) from lamina-DamID experiments (**Materials and Methods**). For comparison, we also

generated a population of structures using only Hi-C data, referred to hereafter as a control model. To test the reproducibility of our method, we generated a second, independently calculated model. The second model confirms our conclusions (**Figs. 6C** and **S10**).

Validation of the structure population

Reproducing the Hi-C contact probabilities. We first assessed the consistency between the chromatin contact probabilities in our structure population with those observed experimentally. The contact probability of any two domains is defined as the fraction of model genome structures for which the two domains are in physical contact with each other,measured over the entire population (a domain-domain contact is defined by an overlap between their soft sphere contact radius). The domain contact probability matrix in our model shows excellent agreement (high correlation) with the Hi-C data, and also closely reproduces the interaction patterns visible in the matrix. The average column-based Pearson's correlation coefficient (PCC) is 0.9840, and the element-wise PCC is 0.9839 (**Suppl. Table 1**). The correlation coefficients of the intra-chromosome arm contact probabilities range between 0.9795 and 0.9981 over all arms, confirming the excellent visual comparison shown in **Fig. 2A**. The correlation coefficients for inter-arm and inter-chromosome contact probabilities are lower, ranging between 0.1475 and 0.3822 (**Suppl. Table 1**). This relatively weak agreement between the model and the experimental data for inter-arm and inter-chromosome interactions can be explained by the following argument. In the Hi-C data, inter-arm and inter-chromosome interactions are relatively infrequent and unstructured, indicating that contacts between chromosomes are predominantly random. Due to their low occurrence, these

interactions are also less reproducible than intra-arm interactions, especially at low sequencing depth. This reasoning is confirmed by comparing two Hi-C experiments performed with two different restriction enzymes [6, 48]. The differences in contact frequencies between the two experiments are generally much larger for inter-chromosome arm interactions than for intra-chromosome arm interactions.

Another quality measure for our models is how well we can predict the frequencies of chromatin interactions that were not included as constraints in the optimization. In our models, we did not impose constraints for any pair of TADs whose contact probability was lower than a_{ij}=0.06. Very low contact probabilities are expected to contain a higher fraction of experimental noise. Such pairs include ~99.99% of all inter-chromosome and inter-chromosome arm interactions. However, our structure population is capable of predicting the missing data (**Fig. 2B** right panel). Many of the low-frequency contacts are formed as a consequence of imposing more significant interactions (with contact probabilities a_{ij}>0.06), and their correct prediction is a good indicator of the model quality.

Reproducing the lamina-DamID binding frequency. Lamina-DamID experiments identify the probability that a locus is associated with the NE (more precisely, with the lamina protein located at the NE). We first assess the consistency between our structure population and the lamina-DamID experiment (a TAD domain-NE contact is defined when the domain surface is less than 50 nm from the NE). The association probabilities are in excellent agreement, with a Pearson's correlation of 0.95 (**Figs. 2C, 2D** and **S1A**). Recalling that the TADs of Drosophila are divided into four functional classes, we find that TADs in the "Active" class are less frequently in contact with the NE than those

from the other three classes (HP1, PcG, and Null) (**Fig. S1A**). This result agrees with prior observations in the literature that the genes interacting with lamina are usually transcriptionally silent and lack active histone marks [1]. The control population generated using only Hi-C data also shows good (albeit substantially lower) correlations between its NE association probabilities and the lamina-DamID experiments (Pearson's correlation is 0.64, with p-value < 2.2e−16) (**Figs. 2D** and **S1B**). This relatively high correlation value in the control population shows a strong consistency between the Hi-C based models with the independent lamina-DamID data and confirms the generally good quality of our Hi-C based structure modeling.

Agreement with FISH experiments. Our genome structures also predict well the NE association frequencies observed by independent FISH mapping of 11 different genomic loci [1]. The Spearman's rank correlation coefficient between experiment and model is 0.642 for these loci, with a significant p-value = 0.03312 (**Fig. S2A**). The corresponding correlation with the control structure population is substantially lower (Spearman's rank correlation coefficient = 0.376 with p-value = 0.2542) (**Fig. S2B**), demonstrating the benefit of data integration to generate more accurate genome structures.

Presence of chromosome arm territories. Chromosome territories have been observed directly in higher eukaryotes, including mammalian cells [49, 50]. In Drosophila, chromosome territories can be inferred from the fact that Hi-C contact frequencies between chromatin regions in the same chromosome arms are substantially higher than those between chromosome arms [7, 8]. Previous 4C experiments on larval brain tissue confirm the limited nature of interactions between genes on different chromosome arms

[41]. FISH experiments have also suggested chromosome territories in Drosophila [40]. In our models, we analyze the formation of chromosome territories by calculating a territory index (TI), which measures the extent of chromosome mixing [24]. To calculate TI in each structure, first we define the spanning volume of each chromosome, which is the surface convex hull of all its domain positions [24]. TI is then defined as the percentage of all domains occupying the chromosome spanning volume of the target chromosome (**Suppl. Methods C.2**). By definition, the maximum TI value of 1 indicates that the chromosome's spanning volume is exclusively occupied by its own domains, and therefore experiences limited chromosome mixing. When considering domains from homologue chromosome copies, the territorial index ranges between 0.96 and 1.0 for all the chromosome arms (**Fig. S3A** and **Suppl. Table 2**). When separating the homologue chromosomes, however, the TI values range between 0.62 and 1.0 for the larger chromosome arms (**Fig. S3B**), suggesting that homologue chromosome pairs share almost the same territory due to strong homologue pairing.

Residual polarized organization. In a polarized genome organization, each chromosome occupies an elongated territory with the centromere at one nuclear pole and telomeres on the opposite side of the nucleus. Such an organization, called Rabl, typically occurs after mitosis and has been observed in a variety of plants [23], yeast, and both polytene and non-polytene Drosophila nuclei; it is also common in Drosophila embryos [27, 42, 43]. In the majority of our genome structures (67.4%, **Suppl. Methods C.3**), more than half of the chromosomes arms (chr2L, chr2R,chr3L, chr3R and chrX) are organized with their centromeres and telomeres located in opposite nuclear hemispheres (**Figs. S4B, C, and D**). This organization is also apparent when calculating the localization

probabilities of chromosomes, which are highest for the telomeres in a region near the NE opposite to their respective centromeres (**Figs. 3A and B**). Taken together, these results suggest that interphase chromosomes retain some features of Rabl organization.

Nuclear colocalization of Hox gene clusters. In Drosophila, the two PcG-regulated Hox gene clusters (Antennapedia complex and Bithorax complex) tend to co-localize in the head of 10-11 stage embryos [51], despite being separated by 10Mb in sequence on chromosome 3 (**Fig. 1B**). To test their spatial colocalization in our models, we calculate the pairwise spatial distances between the two gene clusters in every structure of the population (**Suppl. Methods C.4**). As a random control, we also calculate the pairwise distances between 30 pairs of gene clusters that only contain repressive TADs and share similar chromatin features in order to mimic the PcG-regulated Hox genes. In this control group each pair of gene clusters contains the same number of repressive domains, and are separated by the same sequence distance, as the pair of Hox gene clusters (**Suppl. Methods C.4**). We found that the Hox gene clusters are colocated in about 4.1% of structures in the population, a substantially higher rate than that observed in the control groups (median value 1.18%). Only 3 pairs of clusters among the 30 control groups are more frequently colocated than the Hox gene clusters (**Fig. S5** bottom panel). One of the three shows interactions between the pericentromeric regions and the Null domains. The other two pairs of gene clusters are brought together by nearby active domains, which form frequent interactions. Interestingly, our model does not impose contact constraints between the Hox gene clusters, because their contact probability was below 0.06 (**Fig. S5** top panel). Therefore, these results support the predictive power of our model.

White gene localizing near pericentromatic heterochromatin. Position-effect variegation (PEV) is a process whereby a euchromatic gene is deactivated through an abnormal juxtaposition with heterochromatin, due to chromosome rearrangements or transpositions. PEV has been intensively studied for the Drosophila *white* gene [52, 53], which is on the distal end of chromosome X and separated by more than 19 Mb from the pericentromeric heterochromatin region (**Fig. 1B**). A chromosome inversion can insert the *white* gene in sequence next to the pericentromeric heterochromatin, which leads to its repression. Hence, such chromosomal rearrangement may be favored if the *white* gene has an increased chance of being in spatial proximity to the heterochromatin. However, technical limitations prevent us from directly measuring contacts between the white gene and heterochromatin with Hi-C experiments. Using our structure population, we can measure how often the *white* gene is located close to the pericentromeric heterochromatin of chromosome X. As a control set, we took the four domains that are located at equivalent sequence distances to the heterochromatin regions on chromosomes 2 and 3.

Interestingly, the spatial distance between the *white* gene and the X chromosome heterochromatin is significantly smaller than the corresponding distances of the control groups (one-tailed Welch's two sample t-test, p-value < 2.2e−16) (**Fig. S6A**). Although it is unlikely for distal loci to come together in 3D, we found that in ~1.3% of structures the *white* gene and the heterochromatin were positioned (within a distance of 200 nm) (**Fig. S6B**). This frequency is nine times larger than the colocalization frequency in the control sets (0.14% of structures). Therefore, our models reveal an increased propensity for the white gene to be located near the pericentromeric heterochromatin, compared to

equivalent sites on other chromosomes. This result suggests that spatial proximity facilitates the occurrence of this translocation in living cells.

Different chromosome domains have different preferred locations in the nucleus

The evidence listed above demonstrates the consistency of our models with experimental data and known properties of the Drosophila genome organization. Next, we describe new findings on the nuclear architecture and its functional significance based on our analysis of the model structure population.

Nucleolus and heterochromatin positioning. The nucleolus is a subnuclear structure linked to the assembly of ribosomal subunits. It is formed by nucleolar organizer chromatin regions (NOR), which contain the ribosomal DNA (rDNA) and are located close to the pericentromeric heterochromatin of chromosome X [45]. Our model allows the nucleolus to freely explore the nuclear space. However, its most likely radial position is between the center and periphery of the nucleus (**Figs. 4A** left panel and **S4A**). The large bodies of heterochromatin of each chromosome often enclose the nucleolus (**Fig. 4A**).

Importantly, we validated this model prediction *in vivo*, using Drosophila Kc cells (**Fig. S7**). Immunofluorescence analysis of nucleoli and pericentromeric heterochromatin confirms that the average distance between the center of the nucleolus and the nuclear periphery is less than half of the nuclear radius (**Fig. S7B**). Interestingly, the nucleolus is positioned close to the nuclear periphery in 68% of cells, and close to the center of the nucleus in the remaining cells, revealing a bimodal distribution (**Figs. S7A, C**). In most cells, pericentromeric heterochromatin partially encloses the nucleolus (**Fig. S7A**).

Interestingly, our model predicts certain location preferences for the heterochromatin of individual chromosomes. The heterochromatin regions of chromosomes 4 and X are usually positioned close to each other (**Fig. 4B and Supp. Methods C.5**), and both are more peripheral in the nucleus than the heterochromatin regions of chromosomes 2 and 3 (**Fig. 4A** right panel). The heterochromatin of chromosome 4 appears to be often positioned between the nucleolus and the NE (**Figs. 4A** right panel and **S4A**). We reason that the metacentric chromosomes 2 and 3 are roughly double the size of the acrocentric chromosome X, and therefore spread out more towards the interior of the nucleus. Notably, we confirmed these predictions using FISH staining of heterochromatic repeated sequences (satellites) in Drosophila cells of larval brains. As shown in **Fig. 4C**, the satellite repeats of chromosomes X and 4 are more often closer to each other than those of chromosomes X and 2, or 2 and 4 (**Fig. 4D** top panel), in agreement with our models (**Fig. 4D** bottom panel). Further, the satellite repeats of chromosomes X and 4 are more often closer to the nuclear periphery than those of chromosome 2 (**Fig. 4E** left panel), which also confirms our findings in the model structure population (**Fig. 4E** right panel). Together, these *in vivo* data support our model, and suggest that the predicted chromosome organization is not limited to embryonic cells.

Localization of all euchromatin domains. When plotting the average radial position for every euchromatic TAD (**Fig. 5A**) we observe that the arms near the pericentromeric heterochromatin regions are preferentially positioned in the nuclear interior, while euchromatic regions at the telomeric ends prefer the periphery. This preference is also seen for chromosome 4, despite its small size.

Euchromatic regions are either active or repressed, and can be divided into 4 classes based on their epigenetic profiles: Null, Active, Polycomb-Group (PcG), and HP1 [8] (**Suppl. Table 3**). The TADs of the Null, Active, and PcG classes have similar average radial positions (**Fig. 5B**). The average radial positions of the HP1 TADs have larger variance. The pericentromeric HP1 TADs (excluding all TADs on chr4) are found near the nuclear interior substantially more often than non-pericentromeric HP1 TADs.

Based on our model structures, we can create localization probability density plots (LPD) for the euchromatic regions of different chromosomes (**Fig. 5C**). The chromosome with the most distinct location preference is number 4, whose euchromatic regions reside very close to the NE. In contrast, a large part of chromosome 3L is located on the side of the NE opposite to chromosome 4 along the central axis, coinciding with the line drawn between the centers of the nucleus and nucleolus (vertical dashed line in **Fig. 5C**). Chromosome 2, on the other hand, prefers to avoid the central axis. The right and left arms have similar location preferences. The location distributions of chromosomes 2 and 3 are qualitatively similar, but chromosome 3 is more likely to be found close to the central axis. Chromosome X resides fairly close to the nucleolus, around the midpoint of the central axis, and is considerably less dispersed than the arms of chromosomes 2 and 3.

Analysis of homologous pairing

Distances between homologous pairs vary along the chromosome. D. melanogaster shows somatic homologous chromosome pairing in interphase nuclei [31-33, 44]. Moreover, the paired chromosomes touch only at a few specific interstitial sites [31]. In

our structures, we define a domain as being paired if the surface-to-surface distance between the two homologs is less than 200nm (**Fig. 6A**). Interestingly, the pairing frequencies of homologous domains show distinct and reproducible variation along the chromosomes (**Fig. 6B** left panel). The active class shows smallest homologous pairing frequency for each chromosome (**Fig. 6B** right panel). During the optimization, all pairs of homologue TAD copies are subject to a generic upper bound constraint, which limits their maximum separation to 4 times the TAD diameter. Even though this constraint is the same for all domains, it is noteworthy that in the optimized structures, certain pairs of homologue TADs consistently have small average separations while others consistently have separations close to the upper bound. Hence, this distance variation is TAD-specific and highly reproducible in independently calculated structure populations (**Fig. 6C**). This effect is an indirect consequence of the genome-wide Hi-C and lamina-DamID constraints imposed on the structures.

The consistency of this pairing behavior raises the question of why certain regions attain higher levels of pairing. One clue is that we find a small but significant correlation between pairing frequency and the location of the TAD in the nucleus. Pearson's correlation between the frequency of pairing and the frequency of being in proximity to the NE is 0.34 (p-value < 2.2e-16) (a TAD-NE contact is defined when its domain surface is is less than 50nm from the NE). We hypothesize that genomic regions that are often positioned near the NE may be more restricted in their movements, which may facilitate homolog pairing. We also investigated whether the local crowdedness around the domains could influence the spatial distances between homologues, and found that

in the majority of structures the local crowdedness is not different between paired domains and unpaired domains (**Supp. Methods C.6**).

Mrg15 is enriched in active domains and depleted in repressive domains. Several proteins have been reported to affect somatic homolog pairing in Drosophila [32, 33, 44]. Among them is Mrg15, which binds to chromatin and recruits CAP-H2 protein to mediate homolog unpairing [44]. Interestingly, we find an anticorrelation between Mrg15 binding enrichment in a domain and a domain's homologous pairing frequency, even though this information is not imposed as an input constraint in our models (**Fig. 6D**). The higher the Mrg15 enrichment signal in a domain, the lower the fraction of paired homologues in the structure population (**Fig. 6D**). Pearson's correlation coefficient between the binned Mrg15 binding signal and the averaged frequency of homologous pairing for each bin is −0.81, with p-value = 7.59e−06 (**Fig. 6D**). In the control model (using only Hi-C data), the Pearson's correlation coefficient between them is -0.697 with p-value = 0.000446. We also divided the domains into three subsets based on their Mrg15 scores. The average pairing frequency for domains enriched with Mrg15 is significantly less than that for domains with lower Mrg15 scores (one-tailed Mann–Whitney U test, p-value < 2.2e-16) (**Fig. S8A**).

Among the four TAD classes, "Active" domains are generally more enriched with Mrg15-binding sites (**Fig. 6E** right panel). Appropriately, we observe that transcriptionally active domains have a lower pairing frequency than the three repressive classes (**Fig. 6E** left panel). The most intuitive explanation is that a loose pairing makes an active domain more accessible to regulatory factors. PcG domains, which are enriched with polycomb group proteins, show higher levels of homologous pairing in our models than the active

domains (one-tailed Welch's two sample t-test, p-value = 2.09e−9). Therefore, our structure population supports the notion that PcG domains form tight pairs to enhance gene silencing (reviewed in ref. [26]).

While active domains generally have low frequencies of homologous pairing, our models also have some clear and reproducible counterexamples of active domains with extremely high frequencies of homologous pairing (the specific TADs with this behavior are reproducible in independently generated structure populations) (**Fig. 6C**). Therefore, we divided the active domains into two subclasses, labeled "active-tight" and "active-loose". Interestingly, domains in the active-loose subclass have significantly higher Mrg15 enrichment than domains in the active-tight subclass (one-tailed Mann-Whitney U test, p-value = 0.03436) (**Fig. S8B**). It is interesting that our model further supports a role for Mrg15 in disrupting homologue pairing, even though the structures were generated without any locus-specific constraints on the separation of homologous domains. Importantly, the anticorrelation between homologue pairing frequency and Mrg15 binding signal further increases when lamina-DamID data is integrated in the model, which indicates that data integration helps generate more accurate genome structures.

Active-tight domains show higher transcriptional efficiency. Interestingly, we found significant functional differences between active-loose and active-tight domain subclasses. Active-tight domains contain more genes (**Fig. 7A**). Surprisingly, the active-tight subclass shows significantly lower binding levels of the TATA-binding protein (TBP) and RNA polymerase II, as well as lower H3K4me2 signals (one-tailed Mann-Whitney U test, p-values are 1.17e−04, 2.95e−03 and 5.19e−04 respectively) (**Fig. 7B**). However,

the gene expression levels in the two subclasses are comparable, despite the significantly smaller amount of bound RNA Pol-II transcription machinery in the active-tight subclass. This observation suggests that homologue pairing of active alleles might improve transcription efficiency even at lower concentration of transcription factors.

Active-tight domains tend to be late-replicating. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) is a biochemical method to identify nucleosome-depleted regions in the genome. It has been shown that these DNA sequences overlap with active regulatory sites and DNaseI hypersensitive sites [54]. Active-loose domains are significantly enriched with in the FAIRE signal compared to domains in the active-tight subclass (**Fig. 7C**). This indicates that chromatin in the active-loose domains is more depleted of nucleosomes, and hence these domains contain a higher density of regulatory chromatin complexes. In Drosophila, the organization of nucleosomes plays an important role in determining origin recognition complex (ORC) binding sites [55]. The difference in FAIRE enrichment leads us to investigate DNA replication timing during interphase for the different classes. The Active domains are generally more enriched with ORC than the other three types, with significant p-values (one-tailed Mann-Whitney U test, p-values are 3.259e−15, 0.001715 and 0.01837 for NULL, HP1 and PCG respectively), indicating that DNA replication is often initiated in the chromatin of the active class (**Fig. S9B**). Strikingly, we discovered that ORC-binding regions are much more frequent in active-loose domains than in active-tight domains (one-tailed Mann-Whitney U test, p-value=1.54e−4) (**Fig. 7D**), supporting the model that chromatin in the active-loose subclass replicates significantly earlier (i.e., in early S-phase as opposed to late S-phase).

Discussion

It has become increasingly clear that a chromosome's folding pattern and nuclear location have far-reaching impacts on the regulation of gene expression and other genome functions. Therefore, a thorough understanding of a genome's function entails detailed knowledge about its spatial organization. A wide range of complementary technologies exists to provide such information. For instance, genome-wide ligation assays provide critical information about chromatin-chromatin interactions, lamina-DamID experiments reveal the propensity of a given locus to be located close to the NE, and 3D imaging technologies can reveal the spatial locations of individual loci in single cells. However, many computational models of genome structures rely on a single data type, such as Hi-C, which limits their accuracy. Integrating complementary data types increases the accuracy and coverage of genome structure models, and also provides a way to cross-validate the consistency of data obtained from complementary technologies. Thus, a major and vital challenge of computational biology is to develop hybrid methods that can systematically integrate data obtained from different technologies to generate structural maps of the nucleome (e.g., as this study integrates Hi-C and lamina-DamID).

In this paper, we present a computational platform that can systematically integrate experimental data obtained from different technologies to map the 3D structures of entire genomes. Our probabilistic approach explicitly models the variability of genome structures between cells by simultaneously deconvolving data from Hi-C and lamina-DamID experiments into a model population of distinct diploid 3D genome structures.

Our models therefore incorporate the stochastic nature of chromosome conformations, and allow a detailed analysis of alternative chromatin structure states.

Our method can be applied to genomes of any organism, including mammalian genomes. As a proof of principle, we mapped the structure of the *D. melanogaster* genome in interphase nuclei. We demonstrated that our method produces an ensemble of genome structures whose chromatin contacts are statistically consistent with Hi-C data while also reproducing the likelihoods of chromatin loci being close to the NE derived from lamina-DamID experiments.

The ensemble of model structures has strong predictive power for structural features not directly visible in the initial data sets. We observed that, in embryonic cells, chromosomes 2 and 3 are often organized with their centromeres and telomeres located in opposite hemispheres of the nucleus. In addition, each chromosome pair occupies a distinct territory in our models. Our structures also predicted correctly a relatively high colocalization probability between the two PcG-regulated Hox gene clusters, even though no contact constraints were imposed between these genes when the model structures were generated.

Due to technical limitations, no Hi-C measurements are available to confirm interactions of heterochromatin regions. However, using our 3D model structures, we can analyze the positions of chromatin loci with respect to heterochromatin regions. For instance, our model shows a high preference for the *white* gene on chromosome X to be positioned close to pericentromeric heterochromatin in comparison to similar gene locations on other chromosomes, thus facilitating the *white* gene's translocation next to

heterochromatin. Our analysis also reveals distinct differences between some chromosomes in terms of heterochromatin localization probabilities. For example, pericentromeric heterochromatin of chromosomes X and 4 are more proximal to each other than to pericentromeric heterochromatin of chromosomes 2 and 3. The preferred euchromatin locations of chromosome 4 are also distinctly different from those of the other chromosomes.

We also make intriguing observations about homologous pairing that cannot be directly observed in the original Hi-C or lamina-DamID data. In our models, the tendency for domains to pair varies a great deal along the chromosome, which confirms the idea that pairing initiates from several distinct loci and spreads to neighboring regions. The observed pairing tendency of the domains is highly reproducible over several independent simulations, and also correlates with distinct functional features of the domains. We investigated why certain domains are more frequently paired than others. Interestingly, there is an anti-correlation between pairing frequency and the enrichment in Mrg15 protein binding, which is known to affect somatic chromosome pairing in Drosophila. This information was not explicitly included in the modeling process. The pairing frequencies of homologous domains also differ between those containing active or repressed chromatin. Active domains generally have a lower frequency of chromosome pairing than repressed domains such as those enriched in the polycomb groups (PcG) of proteins. However, we also identified some active domains that break this pattern, with extremely high rates of chromosome pairing across many independent simulations. Interestingly, when we compare these outlier active domains with the more common type of active domain having low pairing frequencies, the former have

substantially lower levels of Mrg15 binding signals, later DNA replication timing, and lower FAIRE signals. These attributes are similar in other regions with high pairing frequencies.

Homologous pairing has been studied for years, and it has been found to play a large role in gene regulation. Transvection is a phenomenon whereby gene expression is modulated by the physical pairing of homologous loci. A case study showed that more transcripts are produced when both alleles of the gene Ubx are paired than when they are spatially separated [28]. A possible explanation is that each gene copy can be activated by both its own and the other copy's enhancer [26]. Interestingly, when we compare actively transcribed genes in chromatin regions with very high or very low levels of homologous pairing, the former show significantly lower signals in RNAPII and TATA protein binding, but at the same time similar levels of transcripts. This observation indicates that a more efficient transcription of genes occurs when pairing is frequent. Our model also shows that regions with looser homologue pairing initiate replication earlier than regions with tighter homologue pairing.

Conclusions

In this study, we address one of the principal challenges of genome structure analysis: the development of a method that systematically integrates complementary data from different technologies to map the 3D organizations of genomes. Data from a single source, such as Hi-C or lamina-DamID experiment alone, cannot capture all aspects of a genome's organization. Integrating multiple data types is therefore not just beneficial but necessary to enhance the accuracy and coverage of structural models. Furthermore,

the detailed analysis of such structural models is a valuable complement to experimental studies, because it can provide new structural insights. For example, the 3D models can reveal the relative locations of specific chromatin regions in the nucleus which are not immediately visible in the initial data. In the future, genome structure modeling should rely on all available data, including live fluorescence and 3D FISH imaging, as well as Hi-C and lamina-DamID experiments from both large-scale single cell and ensemble technologies. This approach will permit an extremely detailed analysis of the genome's structural features, at high resolution and fully consistent with all experimental findings. Our work is a first step towards this goal, in that it allows the integration of genome-wide Hi-C as well as lamina-DamID data for 3D genome structure analysis, and provides a robust computational framework for integrating structural constraints from other types of experiments.

Materials and Methods

General description

The population-based approach is a probabilistic framework to generate a large number of 3D genome structures (i.e., the structure population) whose chromatin domain contacts are statistically consistent with experimental Hi-C data and other spatial constraints derived from *a priori* knowledge and/or independent data types. Our model is a deconvolution of the ensemble-averaged Hi-C data, and the resulting structures can be considered the most likely representation of the true structure population over a population of cells, given all the available data. Our method distinguishes between interactions involving chromosome homologues, so it can generate structure populations representing entire diploid genomes. Further, because the generated population contains many different structural states, this approach can accommodate all experimentally observed chromatin interactions, including those that would be mutually exclusive for a single structure. Compared to our previous research, which introduce the population-based approach using Hi-C data alone, in this study we also integrate lamina-DamID data to generate an improved structure population.

Chromosome representation

The nuclear architecture of Drosophila cells consists of the nuclear envelope (NE), the nucleolus, and eight individual chromosomes (the diploid pairs chr2, chr3, chr4 and chrX). Chr2 and chr3 each have two arms, labeled 2L-2R and 3L-3R, connected by centromeres (**Fig. 1B**).

Each chromosome contains three main regions: euchromatin, pericentromeric heterochromatin, and a centromere (**Fig. 1B**). Euchromatin regions in chromosome arms 2L, 2R, 3L, 3R, 4 and X are linearly partitioned into a total of 1169 well demarcated physical domains [8], which are represented as spheres in the model [22]. A domain sphere is characterized by two radii: (1) its hard (excluded volume) radius, which is estimated from the DNA sequence length and the nuclear occupancy of the genome; and (2) its soft (contact) radius which is twice the hard radius. A contact between two spheres is defined as an overlap between the spheres' soft radii. This two-radius model allows for the possibility that chromatin can partially loop out of its bulk domain region to form contacts, while establishing a minimum genome occupancy in the nucleus. According to experimental data, the combined hard-core spheres of all euchromatin domains occupy around 12% of the nuclear volume. The total volume of heterochromatin is set to 1/27 of the nuclear volume. This figure is in agreement with estimates from microscopy images ([46] and **Fig. S7A**), which show the heterochromatin cluster to occupy roughly one third of the nuclear diameter. The heterochromatin regions of each chromosome are modeled as spheres occupying volumes proportional to 5.4 : 11.0 : 8.2 : 8.2 : 3.1 : 20.0, according to the chromosome outlines depicted in **Fig. 1B** (these volumes are taken from the data shown in ref. [56]). For every chromosome, the centromere is modeled as a sphere with 5% the volume of its corresponding heterochromatin domain (or sum of two heterochromatin domains for chr2 and chr3).

The nuclear radius is set to 2 microns (μm) as suggested by fluorescence imaging experiments ([35, 46] and **Figs. 4C, S7A**). The nucleolus radius is set to 1/6 of the

nuclear radius (**Fig. S7A**). Centromeres are clustered together and attached to the nucleolus [35]. Pericentromeric heterochromatin of chrX surrounds the rDNA cluster regions, so it lies in close proximity to the nucleolus. (**Suppl. Table 3** lists all domain radii in the model.)

All these units are represented by a total of 2359 spheres (see **Table 1**).

Table 1: Structural units of our *Drosophila melanogaster* genome model

Genome component	Unit quantity	Number of spheres	Description
TAD	1169	2338	Euchromatin TADs
HET	6	12	Heterochromatin clusters on 2L, 2R, 3L, 3R, 4, X
CEN	4	8	Centromeres of chromosomes 2, 3, 4, and X
Nucleolus	1	1	Localization of nucleoli

The outlines of the chromosomes are depicted in **Fig. 1B**. In the next section, we briefly describe the chromosome model and list all of the structural constraints that we imposed while optimizing the population.

Probabilistic platform for data integration

Our method closely follows our recent population-based modeling framework [22]. However, we now generalize this framework to support the integration of lamina-DamID data with Hi-C data. The Hi-C data is contained in the ensemble contact probability

matrix $\mathbf{A}$, and the lamina-DamID data is contained in the ensemble chromatin-NE contact probability vector E.

We aim to generate a structure population $\mathbf{X}$ that maximizes the likelihood $P\left(\mathbf{A},E\middle|\mathbf{X}\right)$. We introduce two latent variables $\mathbf{W}$ and $\mathbf{V}$, which represent features of individual cells that aggregate into the ensemble information $\mathbf{A}$ and E, respectively. $\mathbf{W}=(w_{ijm})_{2N\times 2N\times M}$ is the contact indicator tensor, which contains the missing information in the Hi-C data $\mathbf{A}$: the presence or absence of contacts between all domain homologues, in each structure of the population ($w_{ijm}=1$ indicates a contact between domain spheres *i* and *j* in structure *m*; $w_{ijm}=0$ otherwise). The second latent variable, $\mathbf{V}=(v_{im})_{2N\times M}$, contains information whether each domain homologue is located near the NE, in each structure of the population ($v_{im}=1$ indicates that domain sphere *i* is near the NE in structure *m*; $v_{im}=0$ otherwise). Note that while these latent variables are indexed over domain homologues (lowercase indices *i*, *j*), which are independent spheres in the model, the ensemble datasets $\mathbf{A}$ and E in the formulas below are indexed over haploid domain identities observed in the experimental data (uppercase indices *I*, *J*). The maximum likelihood problem is then formally expressed as Eq. (1) and the expansion form is described as in Eq. (2).

Furthermore, $P\left(\mathbf{W},\mathbf{V}\middle|\mathbf{X}\right)$ can be expanded into a product of every contact indicator probability, i.e. $P\left(\mathbf{W},\mathbf{V}\middle|\mathbf{X}\right)=\prod_{m=1}^{M}\prod_{\substack{i,j=1\\ i\neq j}}^{2N}P\left(w_{ijm}\middle|\vec{x}_{im},\vec{x}_{jm}\right)\prod_{i}^{2N}P\left(v_{im}\middle|\vec{x}_{im}\right)$. Then the term $P\left(\mathbf{A}\middle|\mathbf{W}\right)$ can be expanded as $P\left(\mathbf{A}\middle|\mathbf{W}\right)=\prod_{I,J}P(a_{IJ}\,|\,a'_{IJ})$ where a'_{IJ} is the contact

probability of the domain pair *I* and *J*, $a'_{IJ} = \frac{1}{2M}\sum_{m=1}^{M} \overline{w}_{IJm}$. The projected contact tensor $\overline{\mathbf{W}} = (\overline{w}_{IJm})_{N\times N\times M}$ is derived from $\mathbf{W}$ by aggregating its diploid representation to the haploid counterpart.

Likewise, $P(E|\mathbf{V}) = \prod_I P(e_I \mid e'_I)$, where e'_I is the probability for domain *I* to be near the NE. This is calculated as $e'_I = \frac{1}{2M}\sum_{m=1}^{M} \overline{v}_{Im}$. The term $\overline{v}_{Im}$ is a matrix element of the projected matrix $\overline{\mathbf{V}} = (\overline{v}_{Im})_{N\times M}$, and indicates how many domain *I* representations in structure *m* are near the NE; thus, its possible values are {0, 1, 2} when the diploid representation is projected to the haploid counterpart.

With these probabilistic models, we can maximize the log-likelihood $\log P(\mathbf{A}, E, \mathbf{W}, \mathbf{V} \mid \mathbf{X})$, expressed as follows:

$$\begin{aligned}\log P(\mathbf{A}, E, \mathbf{W}, \mathbf{V}|\mathbf{X}) &= \log P(\mathbf{A}|\mathbf{W}) + \log P(E|\mathbf{V}) + \log P(\mathbf{W}, \mathbf{V}|\mathbf{X}) \\ &= \sum_{\substack{I,J=1 \\ I\neq J}}^{N} \log P(a_{IJ} \mid a'_{IJ}) + \sum_{I=1}^{N} \log P(e_I \mid e'_I) \\ &+ \sum_{m=1}^{M}\sum_{\substack{i,j=1 \\ i\neq j}}^{2N} \log P(w_{ijm}|\vec{x}_{im}, \vec{x}_{jm}) + \sum_{m=1}^{M}\sum_{i=1}^{2N} \log P(v_{im}|\vec{x}_{im})\end{aligned} \quad (5)$$

We assume that a pair of spheres (*i*, *j*) are in contact in structure *m* if and only if their center distance $d_{ijm} = \left\|\vec{x}_{im} - \vec{x}_{jm}\right\|_2$ is between certain lower and upper bounds, $L \le d_{ijm} \le U$. The lower bound is the sum of their hard radii, $L = R_i + R_j$, and the upper bound is the sum of their soft radii, $U = 2(R_i + R_j)$. We modeled the probability of a contact between

two domain spheres *i* and *j* as a variant of the rectified or truncated normal distribution, expressed as follows.

$$P(w_{ijm}=1\,|\,\vec{x}_{im},\vec{x}_{jm})=\begin{cases}1, & L\le\left|\left|\vec{x}_{im}-\vec{x}_{jm}\right|\right|_2\le U\\ \exp\left(-\dfrac{\left(\left|\left|\vec{x}_{im}-\vec{x}_{jm}\right|\right|_2-U\right)^2}{2\sigma_w^2}\right), & \left|\left|\vec{x}_{im}-\vec{x}_{jm}\right|\right|_2>U\end{cases} \quad (6)$$

with very small variance, e.g. $\sigma_w\to 0$.

The probability for a domain to reside near the NE is described as

$$P(v_{im}=1\,|\,\vec{x}_{im})=\begin{cases}1, & \left|\left|\vec{x}_{im}\right|\right|_2\ge\lambda R_{\text{nuc}}\\ \exp\left(-\dfrac{\left(\left|\left|\vec{x}_{im}\right|\right|_2-\lambda R_{\text{nuc}}\right)^2}{2\sigma_v^2}\right), & 0\le\left|\left|\vec{x}_{im}\right|\right|_2\le\lambda R_{\text{nuc}}\end{cases} \quad (7)$$

where $\lambda=0.975$ to ensure that the enforced TAD is at the inside surface of the NE, and likewise $\sigma_v\to 0$.

Additional spatial constraints for the Drosophila genome

In addition to the data from Hi-C and lamina-DamID experiments, we include the following additional information as spatial constraints:

1. *Nuclear volume constraint*: All 2359 spheres are constrained to lie completely inside a sphere with radius R_{nuc}, i.e. $\left|\left|\vec{x}_{im}\right|\right|_2\le R_{\text{nuc}}$. Without loss of generality, we use the origin (0,0,0) as the nuclear center, so $\left|\left|\vec{x}\right|\right|_2$ is the distance from the nuclear center.

2. *Excluded volume constraint*: The model prevents any overlapping between the 2359 spheres, as defined by their hard radius. For every pair of spheres *i* and *j* in every structure *m*, we enforce $\left\|\vec{x}_{im}-\vec{x}_{jm}\right\|_2 \geq (R_{im}+R_{jm})$.

3. *Homologue pairing constraint*: Based on experimental evidences, homologous chromosomes are somatically paired in Drosophila and so both copies of a gene are usually close to each other [30-33]. Therefore, we constraint the distance between 2 homologous domains to be less than an upper bound, which is four times the sum of their radii i.e. $\left\|\vec{x}_{im}-\vec{x}_{i'm}\right\|_2 \leq 4(R_i+R_{i'})$.

4. *Consecutive TAD constraint:* To ensure chromosomal integrity, we apply an upper bound to the distance between two consecutive TAD domains, which is derived from the experimentally determined contact probability a_{ij}. The upper bound distance is $d_{ij}(a_{ij},r_i,r_j)=\left(\frac{7}{a_{ij}}+1\right)^{1/3}\left(r_i+r_j\right)$. Note that $d_{ij}=2(r_i+r_j)$ when $a_{ij}=1$.

5. Additional knowledge-based chromosome integrity *constraints*: The heterochromatic region of a given chromosome or chromosome arm forms a clustered subcompartment, so is represented by a single domain. No Hi-C data are available for the heterochromatic regions. To ensure chromosome integrity, the domains representing heterochromatic regions are always in contact with their adjacent TAD as well as with the centromeric domain. The constraint between the heterochromatin sphere and the adjacent TAD sphere *i* is $\left\|\vec{x}_{Hm}-\vec{x}_{im}\right\|_2 \leq 1.5(R_H+R_i)$. The constraint between the heterochromatin domain and the adjacent centromere sphere is $\left\|\vec{x}_{Hm}-\vec{x}_{Cm}\right\|_2 \leq 1.1(R_H+R_C)$, where $\vec{x}_{Hm}$ and $\vec{x}_{Cm}$ are the centers of the heterochromatin and centromere spheres, and R_H and R_C are the hard radii of the heterochromatin and centeromere spheres. Based on experimental evidence [35], all centromeres are in proximity to the nucleolus. Therefore, we constrain the centromere spheres to be close to the spherical

volume representing the nucleolus, defined as $\left\|\vec{x}_{Nu}-\vec{x}_{C}\right\|_2 \le 1.1(R_{Nu}+R_{C})$ where R_{Nu} is the radius of the nucleolus volume.

Distance threshold method for estimating $\mathbf{W}$ and $\mathbf{V}$

We adopt the distance threshold method introduced elsewhere [22] to estimate the distribution of contacts among the diploid genome across a population of structures. The distance threshold d_{IJ}^{act} for each domain pair (*I*,*J*) is determined based on the empirical distribution of all distances between their homologous copies across all structures of the population. The procedure to determine a distance threshold for estimating an element of the projected contact indicator tensor, $\overline{w}_{IJm}$, is as follows. Let (*I*, *J*) be a domain pair (with homologues *i*, *i'* and *j, j'*) and let their Hi-C contact probability $a_{IJ}>0$. We construct an empirical distribution of the pairwise domain distances between homologous copies of the domain pair (*I*, *J*). When *I* and *J* are domains from the same chromosome, we collect the distances d_{ijm} and $d_{i'j'm}$ in all model structures (*m*=1, 2, …, *M*), forming a set of 2*M* distances. When *I* and *J* are domains from different chromosomes, we collect the smallest 2 distances from the set of all possible distances $\{d_{ijm}, d_{i'jm}, d_{ij'm}, d_{i'j'm}\}$, again for a total set of 2*M* distances. Next, the 2*M* distances are ranked in increasing order. The distance threshold, d_{IJ}^{act}, is defined as the distance value with the $(2M\cdot a_{IJ})$th rank among the 2*M* sorted distances. Once all the distance thresholds are obtained, we populate the tensor $\overline{\mathbf{W}}$ by counting how many of the pooled distances between (*I*,*J*) from structure *m* in the set of 2*M* distances that fall below the corresponding distance threshold. The structure optimization then

assigns contacts to the pairs with shorter distance out of 4 possible pairs between homologue domains, for every w_{ijm}. This procedure maximizes $\log P(\mathbf{A},\mathbf{W}|\mathbf{X})$, which is composed of two items: $\log P(\mathbf{W}|\mathbf{X})$ and $\log P(\mathbf{A}|\mathbf{W})$. This is true for two reasons. (i) It assigns contacts only to domain pairs with short distances, maximizing $\log P(\mathbf{W}|\mathbf{X})$. (ii) It uses the $2a_{IJ}M$ th-quantile of all 2*M* distances as the distance threshold to determine w_{ijm}, which heuristically maximizes the first term $\log P(\mathbf{A}|\mathbf{W}) = \sum_{\substack{I,J=1 \\ I\neq J}}^{N} \log P(a_{IJ} \mid a'_{IJ})$ by making a_{IJ} exactly equal to a'_{IJ}.

We adapted this procedure to estimate the TAD-NE contact matrix $\mathbf{V} = (v_{im})_{2N\times M}$. The distance threshold for every TAD is determined. Again we sort a set of 2*M* distances to the NE related to domain *I* in increasing order, and select the $(2M \cdot e_I)$ th rank as the distance threshold. Once the distance thresholds are obtained, we populate the matrix $\bar{\mathbf{V}} = (\bar{v}_{Im})_{N\times M}$ by counting how many of the pooled distances from each structure *m* in the 2*M* distances are lower or the same as the corresponding distance threshold. Note that there are only three possible values of the matrix element: $\bar{v}_{Im} \equiv \{0,1,2\}$. A value of 2 means that both homologues of TAD have to be located near the NE; a value of 1 means only 1 of the homologues has to be located near the NE; and a value of 0 means that neither homologue is forced to be located near the NE. The optimization step will then assign v_{im} accordingly as either 0 or 1. When $\bar{v}_{Im} = 1$, the ambiguity as to whether $(v_{im} = 1, v_{i'm} = 0)$ or $(v_{im} = 0, v_{i'm} = 1)$ is solved on the fly, during the dynamic optimization of the genome structure, where 1 is favored for shorter distances to the NE.

Optimization

As described elsewhere [22], we used step-wise optimization and the A/M iteration algorithm to generate the structure population. We first generated a population of structures satisfying all Hi-C constraints, then fine-tuned the model structures by gradually including the lamina-DamID constraints. For the Hi-C constraints, we included new contact probabilities in several stages during the optimization, at the lower thresholds Θ = {1, 0.7, 0.4. 0.2, 0.1, 0.07, 0.06}. One or more iterations were performed at every probability level. Contact probabilities less than 0.06 were not used at all. 26 A/M iterations were required to generate a structure population consistent with the Hi-C data. The lamina-DamID data were also included in several stages, at the probability levels Θ = {0.2, 0.1, 0.06}. Ten additional A/M iterations were performed to optimize the structure population with respect to the lamina-DamID data. The optimization was performed using a combination of simulated annealing molecular dynamics and conjugate gradient methods. The algorithm was implemented using the Integrated Modeling Platform (IMP) [57].

Data collection and processing

Our processing methods for Hi-C, lamina-DamID and other epigenetics data are described in the Supplemental Material.

Analysis of the structure population

Our statistical analysis of the structure population and details on all statistical tests are described in the Supplemental Material.

Cell culture and immunofluorescence

Kc cells were maintained at 25°C as logarithmically growing cultures in Schneider's medium (Sigma) + FBS (Gemini), and fixed and stained as previously described [46]. The antibodies used were anti-Fibrillarin (Cytoskeleton, Cat. #AFB01, 1:200) and anti-H3K9me2 (Upstate, Cat. # 07-442, 1:500).

Larval fluorescence in situ hybridization (FISH)

Wild-type w^{1118} flies were raised at 25°C. Brains were dissected from third instar larvae and squashed before fixation, as described in [58]. Fixation and FISH staining were carried out as described in [59], using the following probes: 5'-6-FAM-$(AACAC)_7$ for chromosome 2 satellites, 5'-Cy3-TTTTCCAAATTTCGGTCATCAAATAATCAT for chromosome X satellites (359 bp), and 5'-Cy5-$(AATAT)_6$ for chromosome 4 satellites. FISH probes were purchased from Integrated DNA Technologies, and designed as described in [58].

Imaging and image analysis

All images were captured using a Deltavision fluorescence microscopy system equipped with a CoolsnapHQ2 camera, using 60x and 100x objectives and 10-12 Z stacks with Z-intervals of 0.2-0.4. Images were deconvolved with softWorx software (Applied Precision/GE Healthcare) using the conservative algorithm with five iterations. The distances between signals in 3D volume reconstructions of Kc cells or in individual Z stacks of larval tissues were calculated with softWorx. All distances were normalized to the nuclear diameter of their respective cells. Quantification of FISH signals in larval

brains was limited to cells that displayed clear homologous pairing, defined as proximal or overlapping FISH signals for each probe.

Additional files

Additional file 1: This .docx file contains the supplementary methods.

Additional file 2: This .docx file contains the following supplementary figures: S1- S10. Legends for these figures are presented under each figure.

Additional file 3: This .xls file constains three supplementary spreedsheets, each included as a separate tab: Suppl. Tables 1 (Summary of the Pearson's correlation between contact probability from structure models and Hi-C experiment), Suppl. Table 2 (Summary of chromosomal territory index (TI) for individual arms and pairs of homologous arms.) and Suppl. Table 3 (The sphere size of structural units of model).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

QJ, HT, KG and FA designed the 3D modeling methodology and parameterization with input from IC and XJZ. QJ and HT generated and analyzed the genome structure population, and QJ, HT and FA interpreted the results. XL and IC carried out FISH and immunofluorescence experiments and analyzed the results. QJ, HT, FA, IC and XJZ wrote the manuscript. All authors read and approved the manuscript.

Acknowledgements

The work was supported by the Arnold and Mabel Beckman foundation (BYI program) (to F.A), NIH (U54DK107981-01 to F.A and X.J.Z. and NHLBI MAP-GEN U01HL108634 to X.J.Z), and NSF CAREER (1150287 to F.A.). F.A. is a Pew Scholar in Biomedical Sciences, supported by the Pew Charitable Trusts. This work was also supported by a Mallinckrodt Foundation Award and NIH R01GM117376 to I.C. We thank L. Delabaere for assistance with FISH experiments and for generating some of the FISH probes, and the Chiolo Lab for helpful discussions.

References

1. Pickersgill H, Kalverda B, de Wit E, Talhout W, Fornerod M, van Steensel B: **Characterization of the Drosophila melanogaster genome at the nuclear lamina.** *Nat Genet* 2006, **38:**1005-1014.
2. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, van Steensel B: **Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.** *Nature* 2008, **453:**948-951.
3. Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SW, Solovei I, Brugman W, Graf S, Flicek P, Kerkhoven RM, van Lohuizen M, et al: **Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.** *Mol Cell* 2010, **38:**603-613.
4. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326:**289-293.
5. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell* 2014, **159:**1665-1680.
6. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L: **Genome architectures revealed by tethered chromosome conformation capture and population-based modeling.** *Nat Biotechnol* 2012, **30:**90-98.
7. Hou C, Li L, Qin ZS, Corces VG: **Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains.** *Mol Cell* 2012, **48:**471-484.
8. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the Drosophila genome.** *Cell* 2012, **148:**458-472.
9. Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong CT, Cubenas-Potts C, Hu M, Lei EP, Bosco G, et al: **Widespread Rearrangement of 3D Chromatin Organization Underlies Polycomb-Mediated Stress-Induced Silencing.** *Mol Cell* 2015, **58:**216-231.
10. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions.** *Nature* 2012, **485:**376-380.
11. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ: **TopDom: an efficient and deterministic method for identifying topological domains in genomes.** *Nucleic Acids Res* 2016, **44:**e70.
12. Dixon JR, Gorkin DU, Ren B: **Chromatin Domains: The Unit of Chromosome Organization.** *Mol Cell* 2016, **62:**668-680.
13. Bau D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA: **The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules.** *Nat Struct Mol Biol* 2011, **18:**107-114.
14. Giorgetti L, Galupa R, Nora EP, Piolot T, Lam F, Dekker J, Tiana G, Heard E: **Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription.** *Cell* 2014, **157:**950-963.
15. Hu M, Deng K, Qin Z, Dixon J, Selvaraj S, Fang J, Ren B, Liu JS: **Bayesian inference of spatial organizations of chromosomes.** *PLoS Comput Biol* 2013, **9:**e1002893.
16. Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J: **3D genome reconstruction from chromosomal contacts.** *Nat Methods* 2014, **11:**1141-1143.
17. Segal MR, Xiong H, Capurso D, Vazquez M, Arsuaga J: **Reproducibility of 3D chromatin configuration reconstructions.** *Biostatistics* 2014, **15:**442-456.
18. Varoquaux N, Ay F, Noble WS, Vert JP: **A statistical approach for inferring the 3D structure of the genome.** *Bioinformatics* 2014, **30:**i26-33.

19. Zhang B, Wolynes PG: **Topology, structures, and energy landscapes of human chromosomes.** *Proc Natl Acad Sci U S A* 2015, **112:**6062-6067.
20. Imakaev MV, Fudenberg G, Mirny LA: **Modeling chromosomes: Beyond pretty pictures.** *FEBS Lett* 2015.
21. Gong K, Tjong H, Zhou XJ, Alber F: **Comparative 3D genome structure analysis of the fission and the budding yeast.** *PLoS One* 2015, **10:**e0119672.
22. Tjong H, Li W, Kalhor R, Dai C, Hao S, Gong K, Zhou Y, Li H, Zhou XJ, Le Gros MA, et al: **Population-based 3D genome structure analysis reveals driving forces in spatial genome organization.** *Proc Natl Acad Sci U S A* 2016, **113:**E1663-1672.
23. Cowan CR, Carlton PM, Cande WZ: **The polar arrangement of telomeres in interphase and meiosis. Rabl organization and the bouquet.** *Plant Physiol* 2001, **125:**532-538.
24. Kinney NA, Sharakhov IV, Onufriev AV: **Investigation of the chromosome regions with significant affinity for the nuclear envelope in fruit fly--a model based approach.** *PLoS One* 2014, **9:**e91943.
25. Gemkow MJ, Verveer PJ, Arndt-Jovin DJ: **Homologous association of the Bithorax-Complex during embryogenesis: consequences for transvection in Drosophila melanogaster.** *Development* 1998, **125:**4541-4552.
26. Pirrotta V: **Transvection and chromosomal trans-interaction effects.** *Biochim Biophys Acta* 1999, **1424:**M1-8.
27. Marshall WF, Dernburg AF, Harmon B, Agard DA, Sedat JW: **Specific interactions of chromatin with the nuclear envelope: Positional determination within the nucleus in Drosophila melanogaster.** *Molecular Biology of the Cell* 1996, **7:**825-842.
28. Goldsborough AS, Kornberg TB: **Reduction of transcription by homologue asynapsis in Drosophila imaginal discs.** *Nature* 1996, **381:**807-810.
29. Wang L, Brown JL, Cao R, Zhang Y, Kassis JA, Jones RS: **Hierarchical recruitment of polycomb group silencing complexes.** *Mol Cell* 2004, **14:**637-646.
30. McKee BD: **Homologous pairing and chromosome dynamics in meiosis and mitosis.** *Biochim Biophys Acta* 2004, **1677:**165-180.
31. Fung JC, Marshall WF, Dernburg A, Agard DA, Sedat JW: **Homologous chromosome pairing in Drosophila melanogaster proceeds through multiple independent initiations.** *Journal of Cell Biology* 1998, **141:**5-20.
32. Bateman JR, Larschan E, D'Souza R, Marshall LS, Dempsey KE, Johnson JE, Mellone BG, Kuroda MI: **A Genome-Wide Screen Identifies Genes That Affect Somatic Homolog Pairing in Drosophila.** *G3-Genes Genomes Genetics* 2012, **2:**731-740.
33. Joyce EF, Williams BR, Xie T, Wu CT: **Identification of genes that promote or antagonize somatic homolog pairing using a high-throughput FISH-based screen.** *PLoS Genet* 2012, **8:**e1002667.
34. Mellert DJ, Truman JW: **Transvection is common throughout the Drosophila genome.** *Genetics* 2012, **191:**1129-1141.
35. Padeken J, Mendiburo MJ, Chlamydas S, Schwarz HJ, Kremmer E, Heun P: **The nucleoplasmin homolog NLP mediates centromere clustering and anchoring to the nucleolus.** *Mol Cell* 2013, **50:**236-249.
36. Weierich C, Brero A, Stein S, von Hase J, Cremer C, Cremer T, Solovei I: **Three-dimensional arrangements of centromeres and telomeres in nuclei of human and murine lymphocytes.** *Chromosome Research* 2003, **11:**485-502.
37. Mekhail K, Seebacher J, Gygi SP, Moazed D: **Role for perinuclear chromosome tethering in maintenance of genome stability.** *Nature* 2008, **456:**667-670.

38. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B: **Systematic protein location mapping reveals five principal chromatin types in Drosophila cells.** *Cell* 2010, **143:**212-224.
39. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al: **Comprehensive analysis of the chromatin landscape in Drosophila melanogaster.** *Nature* 2011, **471:**480-485.
40. Dernburg AF, Broman KW, Fung JC, Marshall WF, Philips J, Agard DA, Sedat JW: **Perturbation of nuclear architecture by long-distance chromosome interactions.** *Cell* 1996, **85:**745-759.
41. Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, Nieuwland M, Simonis M, de Laat W, van Lohuizen M, van Steensel B: **Interactions among Polycomb domains are guided by chromosome architecture.** *PLoS Genet* 2011, **7:**e1001343.
42. Hochstrasser M: **Spatial organization of chromosomes in the salivary gland nuclei of Drosophila melanogaster.** *The Journal of Cell Biology* 1986, **102:**112-123.
43. Lowenstein MG, Goddard TD, Sedat JW: **Long-range interphase chromosome organization in Drosophila: a study using color barcoded fluorescence in situ hybridization and structural clustering analysis.** *Mol Biol Cell* 2004, **15:**5678-5692.
44. Smith HF, Roberts MA, Nguyen HQ, Peterson M, Hartl TA, Wang XJ, Klebba JE, Rogers GC, Bosco G: **Maintenance of interphase chromosome compaction and homolog pairing in Drosophila is regulated by the condensin cap-h2 and its partner Mrg15.** *Genetics* 2013, **195:**127-146.
45. Hilliker A: **The genetic analysis of D. melanogaster heterochromatin.** *Cell* 1980, **21:**607-619.
46. Chiolo I, Minoda A, Colmenares SU, Polyzos A, Costes SV, Karpen GH: **Double-strand breaks in heterochromatin move outside of a dynamic HP1a domain to complete recombinational repair.** *Cell* 2011, **144:**732-744.
47. Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, Gorchakov AA, Jaffe JD, Kennedy C, Linder-Basso D, et al: **Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin.** *Genome Res* 2011, **21:**147-163.
48. Yaffe E, Tanay A: **Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture.** *Nat Genet* 2011, **43:**1059-1065.
49. Cremer T, Cremer C: **Chromosome territories, nuclear architecture and gene regulation in mammalian cells.** *Nat Rev Genet* 2001, **2:**292-301.
50. Hochstrasser M: **Three-dimensional organization of Drosophila melanogaster interphase nuclei. I. Tissue-specific aspects of polytene nuclear architecture.** *The Journal of Cell Biology* 1987, **104:**1455-1470.
51. Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, Bonnet J, Tixier V, Mas A, Cavalli G: **Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila.** *Cell* 2011, **144:**214-226.
52. Elgin SC, Reuter G: **Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila.** *Cold Spring Harb Perspect Biol* 2013, **5:**a017780.
53. Muller HJ: **Types of visible variations induced by x-rays in Drosophila.** *Journal of Genetics* 1930, **22:**299-U297.
54. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD: **FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin.** *Genome Res* 2007, **17:**877-885.
55. Eaton ML, Prinz JA, MacAlpine HK, Tretyakov G, Kharchenko PV, MacAlpine DM: **Chromatin signatures of the Drosophila replication program.** *Genome Res* 2011, **21:**164-174.
56. Adams MD: **The Genome Sequence of Drosophila melanogaster.** *Science* 2000, **287:**2185-2195.

57. Russel D, Lasker K, Webb B, Velazquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A: **Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies.** *PLoS Biol* 2012, **10:**e1001244.

58. Larracuente AM, Ferree PM: **Simple method for fluorescence DNA in situ hybridization to squashed chromosomes.** *J Vis Exp* 2015**:**52288.

59. Ryu T, Spatola B, Delabaere L, Bowlin K, Hopp H, Kunitake R, Karpen GH, Chiolo I: **Heterochromatic breaks move to the nuclear periphery to continue recombinational repair.** *Nat Cell Biol* 2015, **17:**1401-1411.

Figures and Legends

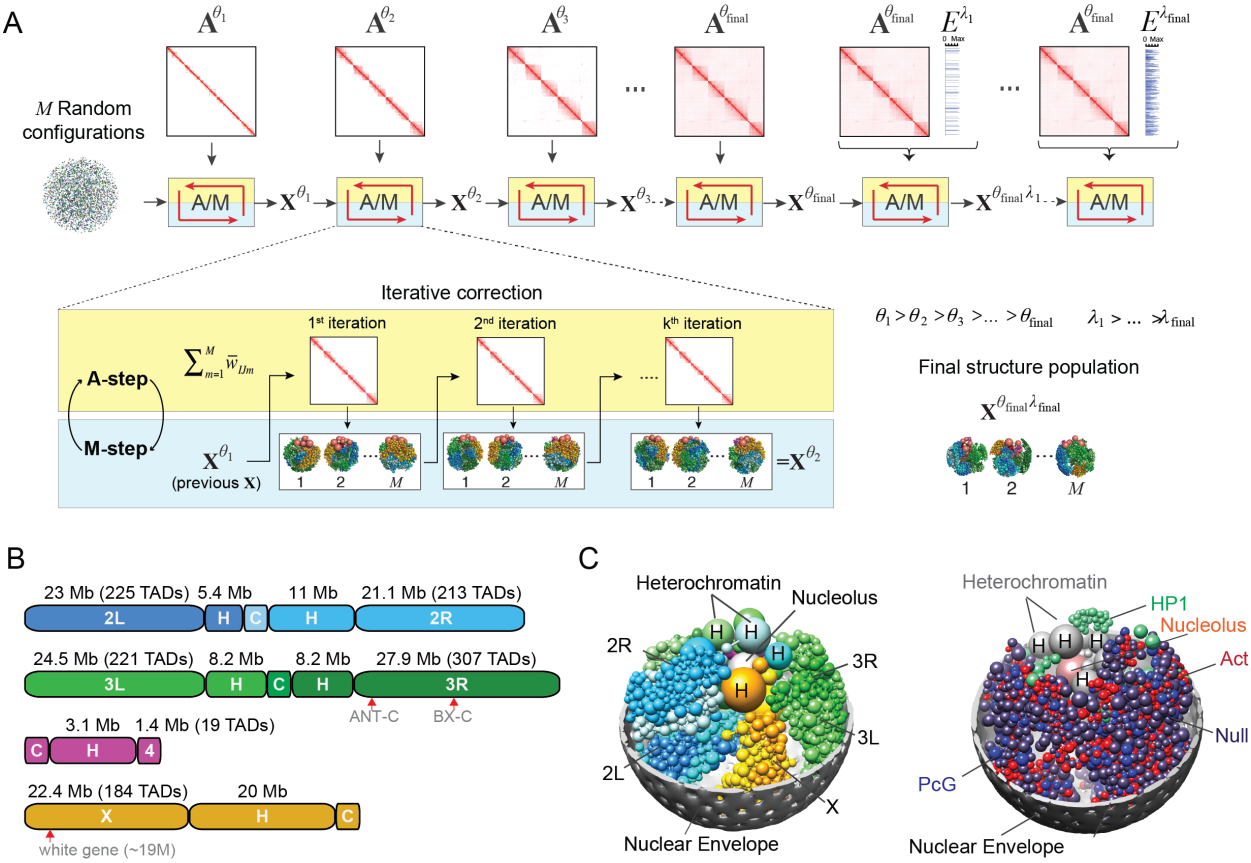

Figure 1. Overview of the population-based genome structure modeling approach and its application to the Drosophila genome. (A) The initial structures are random configurations. Maximum likelihood optimization is achieved through an iterative process with two steps, assignment (A) and modeling (M). We increase the optimization hardness over several stages by including contacts from the Hi-C matrix $\mathbf{A}$ with lower probability thresholds (θ). After the population reproduces the complete Hi-C data, we include the vector E (lamina-DamID), again in stages with decreasing contact probability thresholds (λ). (B) Schematic of the Drosophila genome. The autosome arms are designated 2L, 2R, 3L, 3R, 4 and X. The arms of chr2 and chr3 are connected by centromeres labeled "C". Euchromatic regions are labeled as the arm. The numbers along the top of a genome indicate the length of the section in megabases (Mb), and for euchromatin the number of spheres (TADs) in the structure model is also given. The heterochromatic region of each chromosome arm is labeled "H". The white gene is located ~19M away from the heterochromatin of chrX. Also indicated are the Hox genes: 5 genes of the Antennapedia complex (ANT-C) are located at ~2.3M-2.8M from the heterochromatin of chr3R, and 3 genes of the Bithorax complex (BX-C) are located at ~12.4M-12.7M from the heterochromatin of chr3R. (C) Snapshot of a single structure randomly picked from the final population. (Left panel) The full diploid chromosomes are shown in colors: blue-chr2, green-chr3, magenta-chr4, and orange-chrX. The heterochromatin spheres are larger than the euchromatin domains. The nucleolus is colored in silver. (Right panel) The euchromatin domains are colored to reflect their epigenetic class: red-Active, blue-PcG, green-HP1 and dark-Null. Heterochromatin spheres are grey, and the nucleolus is pink.

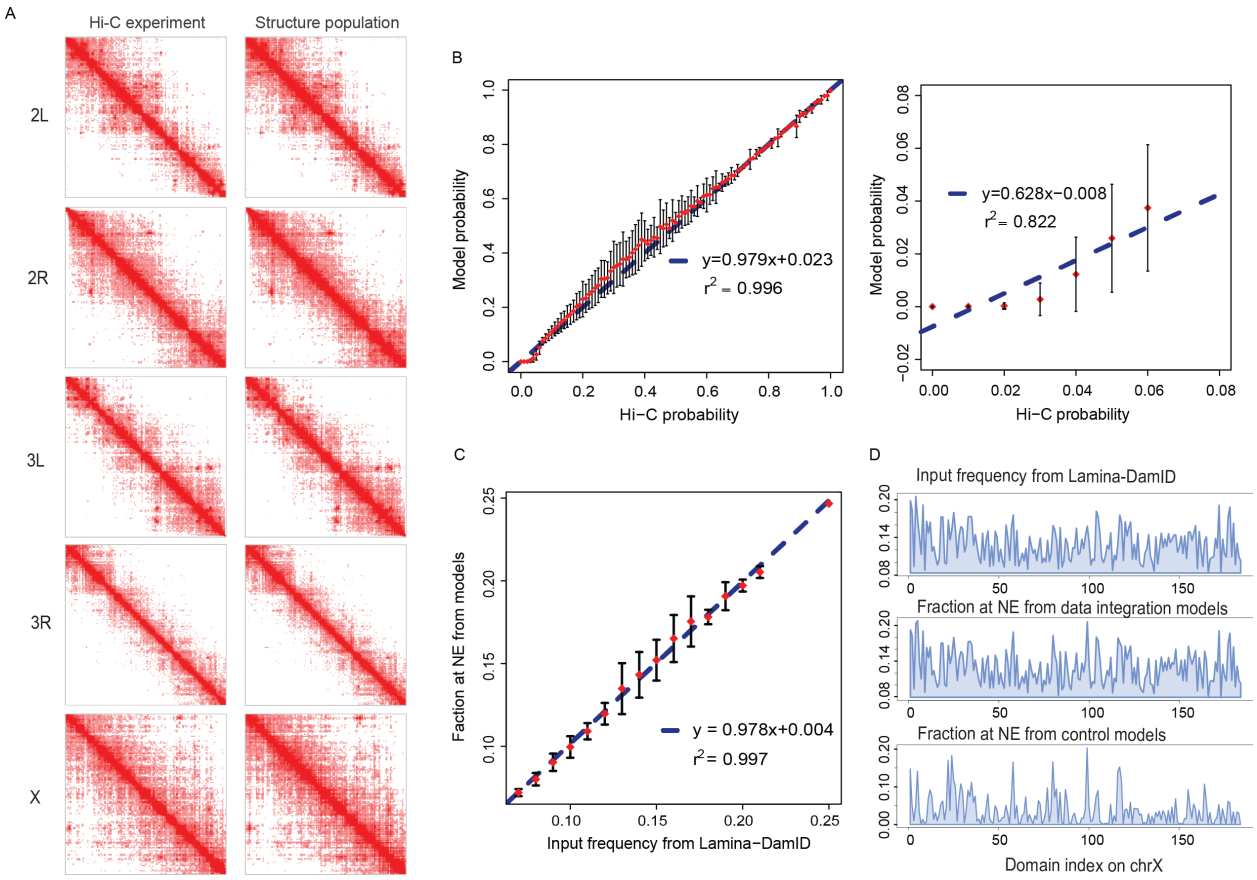

Figure 2. Reproduction of Hi-C and lamina-DamID data. (A) Heat maps of intra-arm contact probabilities from Hi-C experiments (left) and intra-arm contact frequencies from the structure population (right). Their similarity is quantified by element-wise Pearson's correlations, which are 0.9844, 0.9852, 0.9840, 0.9859 and 0.9795 for chr2L, chr2R, chr3L, chr3R and chrX, respectively. The maps only show interactions with probabilities no less than 6%, which are used as constraints in our modeling procedure. (B) Agreement between the experimental data and model contact probabilities. (Left panel) The input Hi-C probabilities are divided into 100 bins, the corresponding model probabilities in one bin are summarized by mean and variance, and then the error bar plot is shown. The blue dot-line is the linear regression line between the average model probabilities of each bin and the mid-point Hi-C probabilities of the bins. Their Pearson's correlation is 0.998 with p-value < 2.2e−16. (Right panel) Close-up of the agreement between experiment and model for contacts with probabilities less than 6%, which are not used as constraints in our modeling procedure. In this range, Pearson's correlation is 0.907 with p-value = 0.004867. (C) The agreement between NE association frequencies from lamina-DamID experiments and the model population. This figure is plotted in the same way as (B). The structure population well reproduces the input frequencies derived from lamina-DamID data, with a Pearson's correlation of 0.95 and p-value < 2.2e−16. (D) Comparison of experimental and model lamina-DamID frequencies on chrX. The top panel shows the input frequencies derived from the lamina-DamID signal, the middle panel shows the fraction of domains located at the NE in the structure population obtained by Hi-C and lamina-DamID data integration, and the bottom panel shows the fractions obtained in our control structure population generated using only Hi-C data.

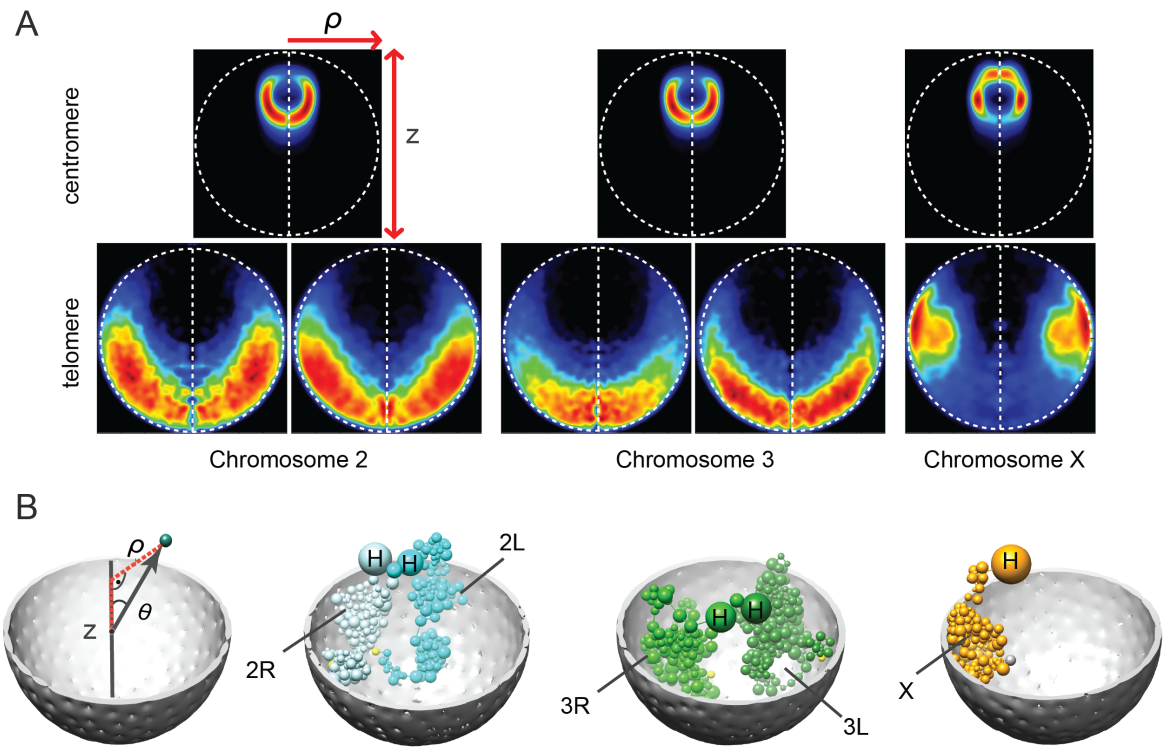

Figure 3. Residual polarized organization. (A) Projected localization probability densities (LPDs) of centromeres and peri-telomeric sequences for all chromosome arms calculated from the structure population. Probability densities are determined with respect to two principle axes of the nuclear architecture. The z-axis connects the center of the nucleolus with the origin at the nuclear center. The radial axis defines the distance of a point from the central z-axis (shown in the left panel in B). The left half of the projected localization density plot mirrors the right half for visual convenience. (B) Illustration of the genome organization for different chromosome arms in one genome structure.

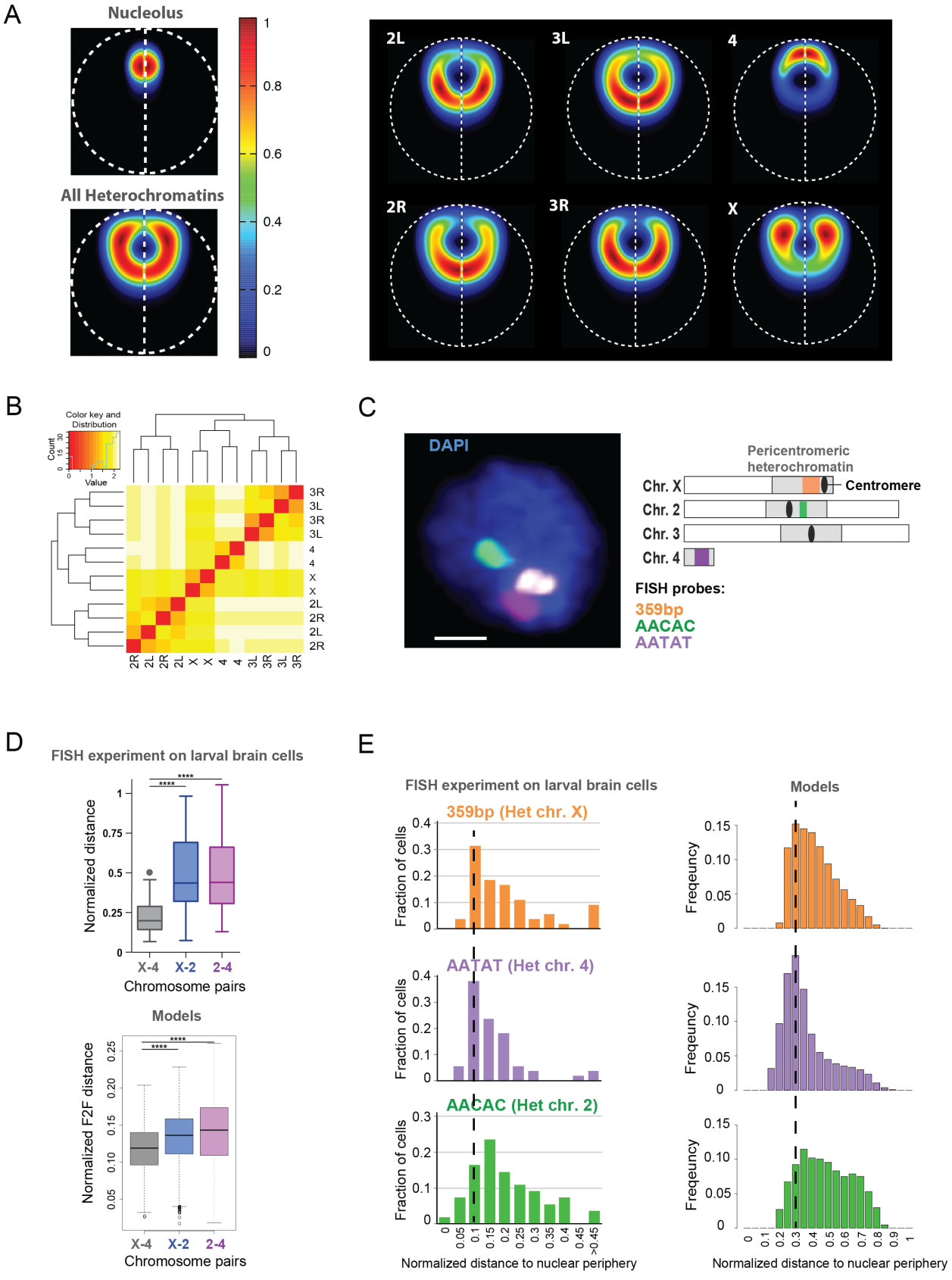

Figure 4. Heterochromatin and nucleolus positions. (A) (Left panel) LPD plots of the nucleolus and all pericentromeric heterochromatin regions in the model. On average, the nucleolus occupies an intermediate position between the center and the periphery, and is surrounded by pericentromeric heterochromatins. (Right panel) LPD plots for pericentromeric heterochromatin of different chromosome arms. They all exhibit different preferred locations. Those of chr4 and chrX are significantly more peripheral than the others. (B) Clustering of pericentromeric heterochromatin regions based on their averaged surface-to-surface distances. Heterochromatin domains of arms from the same chromosome naturally show preferred clustering. Heterochromatin domains from chr4 and chrX are usually closer to each other than to those from other chromosomes. (C) (Left panel) FISH signals in larval brain cells. The image shows the middle Z-stack of a representative nucleus. Scale bar = 1 μm. (Right panel) Schematic of the position of FISH probes used for this study, relative to the pericentromeric regions of each chromosome (chrX,chr2,chr4). (D) (Top panel) The positions of heterochromatic satellites from different chromosomes relative to each other, as measured by FISH experiments on larva brain cells. ****p-value < 0.0001 by paired t-test, N = 55 cells. (Bottom panel) Pairwise distances (surface-to-surface distance normalized by the diameter of the nucleus) between the heterochromatin domains as measured in the model. Similar to the data in vivo, the distance between the heterochromatin domains of chrX and chr4 is significantly smaller than the distance between the other two pairs, according to paired t-tests (p-value < 2.2e−16). (E) (Left panel) Positions of heterochromatic satellites from different chromosomes relative to the nuclear periphery, obtained from FISH experiments on larva brain cells. The heterochromatic satellites on chrX and chr4 are closer to the NE than those of chr2. (Right panel) A similar plot generated from the structure population shows good agreement with the FISH experiments.

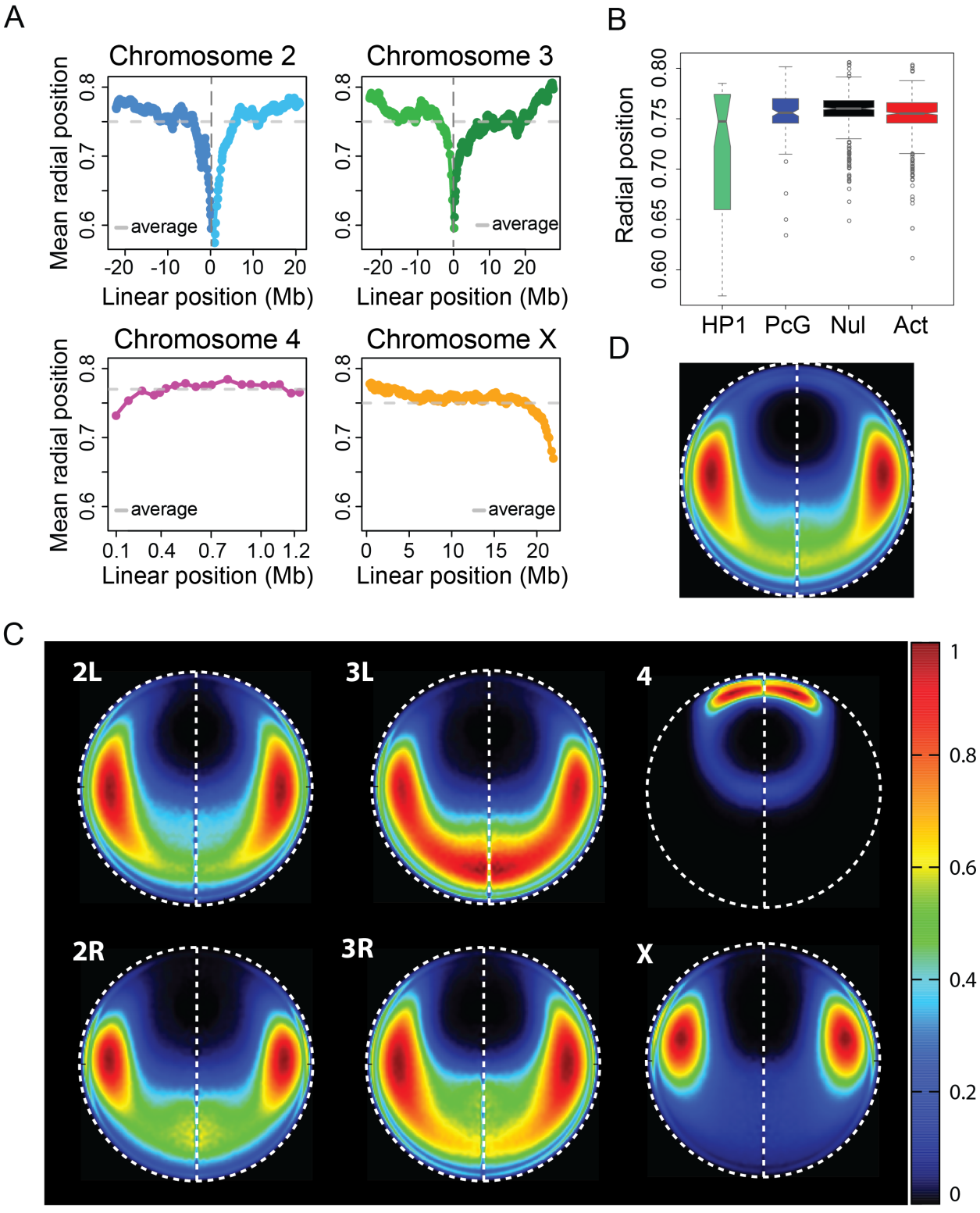

Figure 5. Localization of euchromatin domains in the structure population. (A) The average radial position for each euchromatin domain, plotted by position along its chromosome. The 0 location along the x-axis of chr2 represents the euchromatin region closest to the centromere, with 2L domains on the left and 2R domains on the right. Chr3 domains are plotted with the same coordinate system as chr2. The domains of chr4 are plotted from left to right, while the domains of chrX are plotted from right to left; this convention follows the schematics in Figure 1. Centromeric regions and pericentromeric heterochromatin regions are not shown in this figure. The domains near pericentromeric regions are closer to the nuclear center on average, while the domains near telomeric ends are preferentially close to the nuclear periphery. (B) The average radial positions of each domain, grouped by epigenetic class. (C) LPD plots of all euchromatin domains from each chromosome arm in nuclear space. (D) LPD plot of all euchromatin domains.

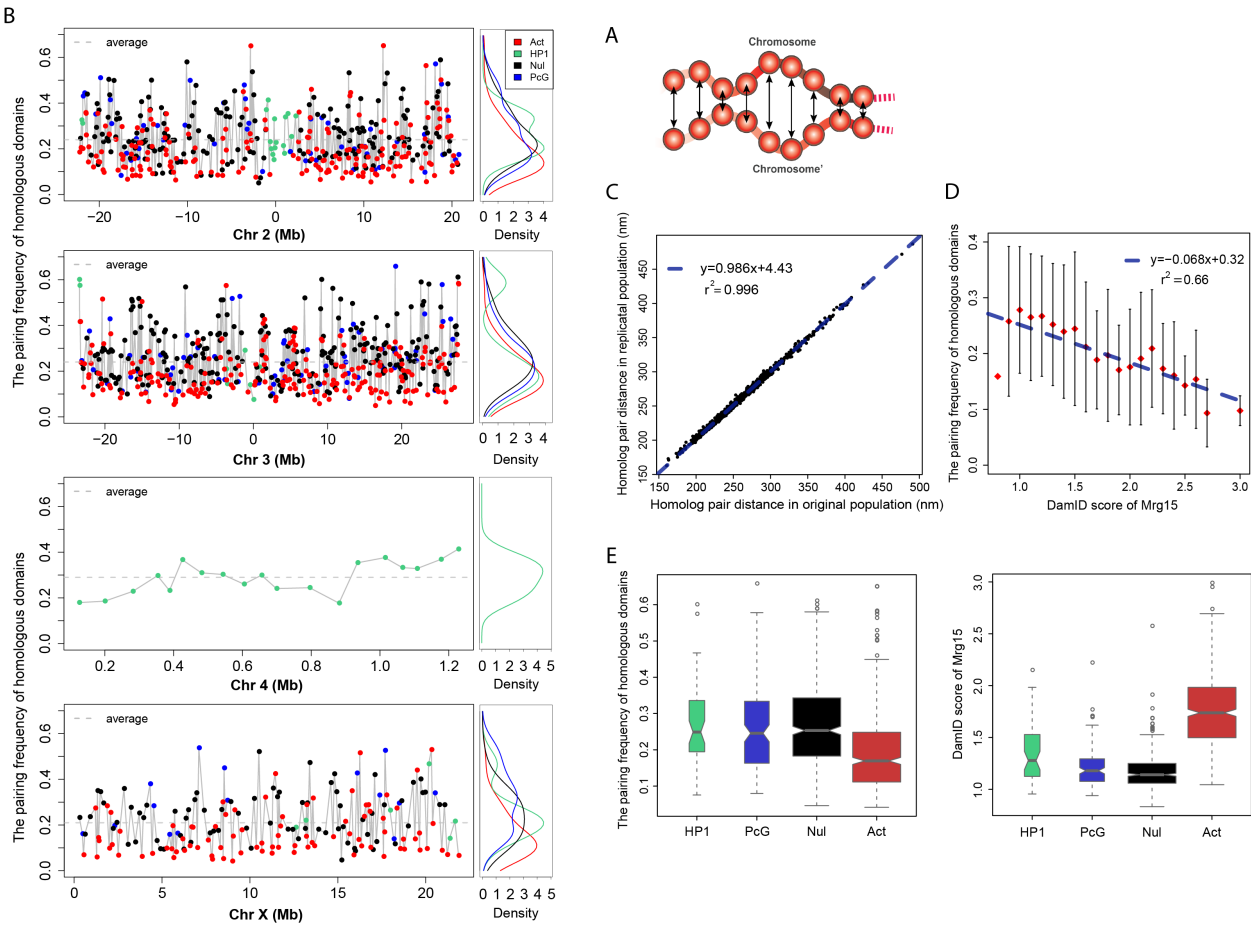

Figure 6. Analysis of homologous pairing. (A) Schematic view of surface-to-surface distances between homologous domains. Different domains exhibit different degrees of homologue pairing. (B) (Left panel) Pairing frequency for each euchromatin domain, plotted by chromosome. We define a domain as being paired in a structure if the surface-to-surface distance between the two homologs is less than 200nm. The X-axes are the same as the plots in Fig. 5A. The domains are colored by their epigenetic classes: Green-HP1, Blue-PcG, Black-Null, and Red-Active. (Right panel) Density plot of the domain pairing frequencies, grouped by epigenetic class. The active class has the smallest mean homologous pairing frequency for each chromosome. (C) Reproducibility of the average homologue distances between two independently generated structure populations. The Pearson's correlation between them is 0.998, with p-value < 2.2e−16. (D) There is a negative correlation between the pairing frequencies of homologous domains and their Mrg15 enrichment. The Mrg15 scores range from 0.8 to 3.0, and are divided into 21 equal bins. The corresponding pairing frequencies from our models in a given Mrg15 bin are summarized as a mean and variance, and the latter is displayed as an error bar. The blue dotted line is the linear regression between the average pairing frequency in each bin and the midpoint Mrg15 enrichment value of the bin. The Pearson's correlation between them is −0.8126852, with p-value = 7.591e−06. (E) (Left panel) Pairing frequencies of homologous domains, grouped by epigenetic class. (Right panel) Enrichments of Mrg15 binding, grouped by epigenetic class. Active domains are generally more enriched with Mrg15, and have lower pairing frequencies, than the other three repressive classes.

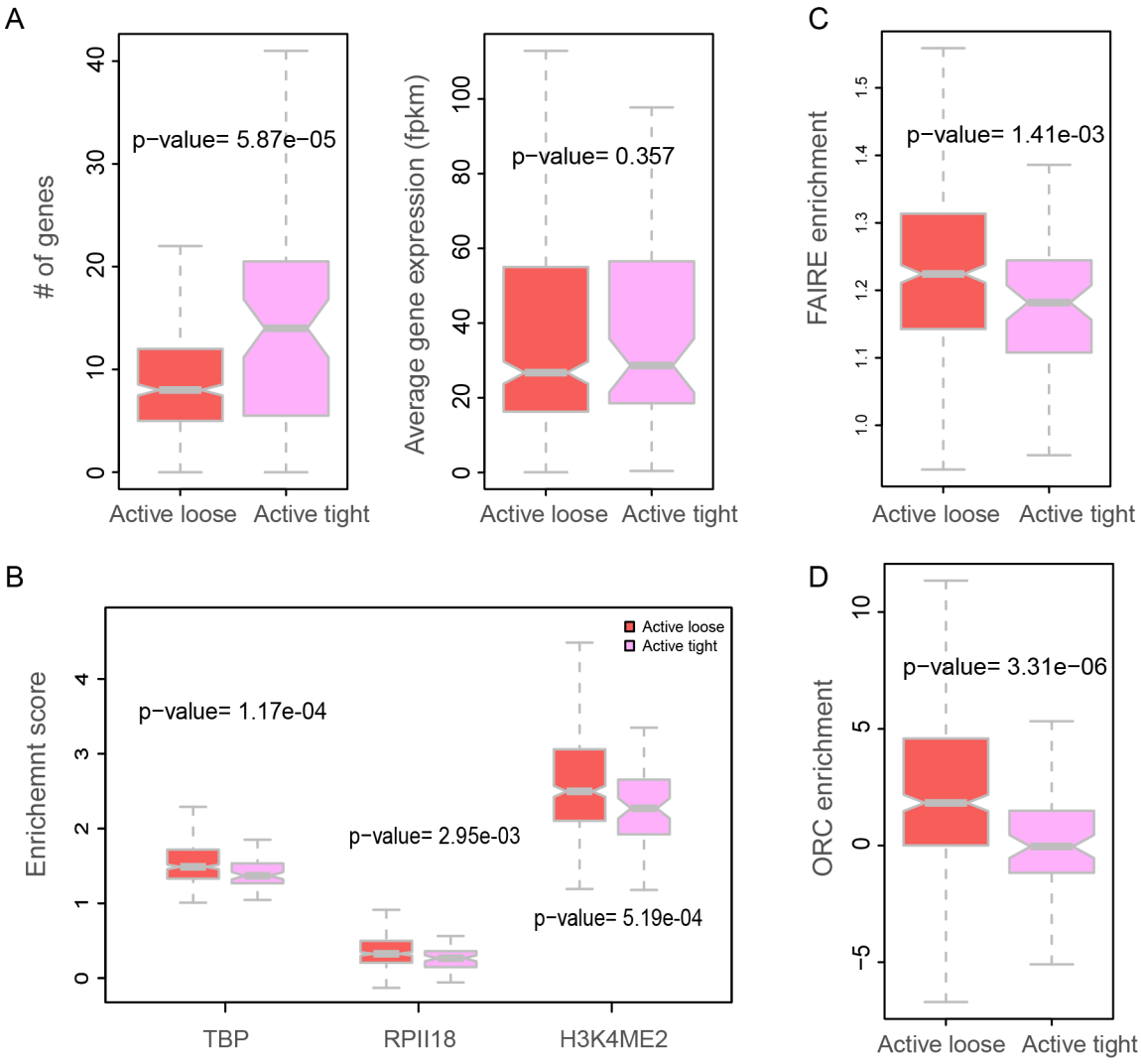

Figure 7. Transcriptional efficiency and DNA replication timing for genes in two sub-classes of the Active domains. (A) Domains in the “active-loose” subclass have lower frequencies of homologue-pairing than those in the “active-tight” subclass (Supp. Methods C.6). The active-tight subclass includes 71 domains, and the active-loose sub-class includes 423 domains. All the statistical tests are performed using one-tailed Mann-Whitney U test. (Left panel) Domains in the active-tight subclass contain significantly more genes than domains in the active-loose subclass. (Right panel) Genes in both sub-classes have similar average expression values. (B) TBP (TATA binding protein), PolII binding signal and H3K4me2 signals are more enriched in domains of the active-loose subclass. (C) FAIRE signal is significantly stronger in domains of the active-loose subclass. (D) ORC is significantly more enriched in domains of the active-loose subclass.

Supplementary methods

A. Hi-C data processing

A.1 Bin-level contact frequency

The sequencing data were downloaded from Gene Expression Omnibus under accession number GSE34453 [1]. We adopted the pipeline developed by Leonid Mirny lab [2] to process the Hi-C data. First, the two sides of each read were mapped to the Drosophila melanogaster genome (assembly dm3) independently using bowtie2 [3] with "very-sensitive" option. We truncated the reads to 20bp, and then remapped the non-mapped and multiple mapped reads by increasing truncation length with 5bp gradually. The truncating step significantly yields more double-sided mapped reads. 216,199,696 uniquely double-sided mapped reads are retained from the original 362,669,793 paired reads (with the mapping ratio at ~60%).

Then, the reads alignments for the artificial or non-informative contacts were filtered out, including self-ligation products, the products without ligation junction and PCR duplicates. After filtering process, 14,481,367 interactions are left for downstream analysis. Third, the valid double-sided alignments were used to construct the genome-wide contact matrix at 40k bin size, resulting in a 3012*3012 matrix. Before correcting the matrix for biases, we performed four types of bin-level filtering which will affect the normalization procedure. We removed the contacts between loci located within the same bin; removed bins with more than 50% are N's in the reference genome; removed

1% of bins with low coverage; and truncated top 0.05% of inter-chromosomal counts which truncated values of the top 0.05% to be the same as highest value among the rest 99.95%. Finally, the iterative correction was performed on the filtered genome-wide contact map to get a normalized map, denoted as $F_K = (f_{ij})_{K*K}$.

A.2 Bin-level contact probability

The contact probability is defined as the probability for observing a given contact in the structure population. We define a threshold value f^{max}, which defines the frequency at which a contact is formed in 100% of the structure population. We assume the contact frequencies that constitute any stable TAD can serve as reference where the interaction could exist in 100% of the cells. First, we register all contact frequencies inside TADs. Then we apply an R (CRAN statistical software) function *boxplot.stats* on this contact frequency set to get the extreme lower whisker of the boxplot and set it as the f^{max}. Namely, in R

$$f^{max} = boxplot.stats\big(f_{ij}(intra\ TAD)\big)\,\$stat[1]$$

The contact probability between bin i and bin j is derived as

$$a_{ij} = \min\{\frac{f_{ij}}{f^{max}}, 1\}$$

We applied the f^{max} calculation method to the normalized frequency matrix to obtain the contact probability at 40 kb-binned matrix.

A.3 Domain-level contact probability

Based on the Hi-C contact frequency map, 1169 physical domains or topological associated domains (TADs) are detected by a quantitative probabilistic approach [1].

The chromatin in our population structure is represented at the level of TADs, therefore the contact probability at domain-level need to be obtained. The domain level contact probability is denoted as $A_N = (\mathrm{a_{IJ}})_{\mathrm{N*N}}$, where $\mathrm{a_{IJ}}$ is the contact probability between domain I and domain J, and N is the total number of domains in the genome. $\mathrm{a_{IJ}}$ is derived from the corresponding contact probability at the bin level. If $\mathrm{b(I)}$ is the set of all bins in domain I, and $\mathrm{b(J)}$ is the set of all bins in domain J, then

$$\mathrm{a_{IJ}} = mean\left\langle top\ 10\%\{a_{ij} |\ i \epsilon\ b(I), j \epsilon\ b(J)\}\right\rangle$$

is the average value of the top 10% ranked contact probabilities in the set of all pairwise combinations between bins in $b(\mathrm{I})$ and $b(\mathrm{J})$.

B. Lamina-DamID data processing

The genome-wide lamina-DamID binding signal is collected from [4]. The binding signal for each TAD ($L = \{l_I \mid I = 1,2,...,N\}$) is calculated using BigWigSummary tool (from USC Genome Browser). The DNA content of the nuclear periphery was measured to be ~12% per nucleus for Kc167 cell line [5]. To reproduce the experimentally observed DNA content at the nuclear periphery, we relate the average lamina-DamID signal of all chromatin regions (measured in an ensemble of cells) to the average domain-NE localization probability ($E = \left\{e_I \middle| I = 1,2,...,N\right\}$) in the structure population so that $\mathrm{mean}(E)=0.12$. The lamina-DamID signal is transferred into a probability for each domain to be close to the NE as $\mathrm{e}_I = \dfrac{0.12}{\mathrm{mean}(L)} l_I$.

C. Analysis of the structure population

C.1 Reproducing lamina-DamID binding frequency

We define a chromatin domain-lamina contact, if the distance between the domain's surface and the NE is less than 50nm. The NE-domain association probability is the fraction of structures in the population, in which the domain is in contact to the NE.

C.2 Chromosome territory index

To quantify how effectively one chromosome excludes other chromosomes from the volume it occupies in the 3D space, we adopted the quantity called chromosome territory. There is no universal definition of chromosome territory, but we follow the definition in a recent publication about the structure modeling of Drosophila polytene chromosomes [6]. Chromosome territory index (TI) is defined as the fraction of domains inside a convex hull that belongs to the chromosome used for its construction.

We first calculate the convex hull for a chromosome arm (with domain number N_{chr}) using the function *delaunayn* in MATLAB (http://www.mathworks.com/help/matlab/ref/delaunayn.html). *T=delaunayn(X)* computes a set of simplices such that no data points of X are contained in any circumspheres of the simplices. The set of simplices forms the Delaunay triangulation. Then, the function *tsearchn* is used to search all the domains inside of the convex hull, and the number of detected domains is denoted as N_{hull}. Finally, the TI is calculated as $\frac{N_{chr}}{N_{hull}}$.

The maximum N_{hull} for each chromosome are the same, which is the total number of domains inside the nucleus (2238 euchromatin TADs in this study). The minimum N_{hull} for each chromosome corresponds to the number of domains that belongs to themselves respectively. Under this definition, the maximum TI is 1, indicating that all domains inside the chromosome spanning volume are exclusively occupied by its own chromosome domains, and therefore shows a strong chromosome territory formation with only limited chromosome mixing. The theoretical minimum values for each chromosome arm are 0.096, 0.091, 0.095, 0.131, 0.008 and 0.0787, and for each pair of homologous arms are 0.192, 0.182, 0.189, 0.263, 0.016 and 0.157. The average TIs for chromosome arms in the structure population are 0.64 (2L), 0.65 (2R), 0.62 (3L), 0.62 (3R), 1.0 (4) and 0.67 (X). The average TI for individual arms is around 60%, suggesting the homolog pairs share territory almost equally. Indeed, the paired arms together possess high territorial index, i.e. 0.97, 0.98, 0.96, 0.98, 1.0 and 0.98 for arms 2L, 2R, 3L, 3R, 4, and X, respectively.

C.3 Residual polarized organization

The polarized (Rabl-like) organization shows that each chromosome occupies an elongated territory, with the centromere in one nuclear hemisphere and telomere in the opposite hemisphere. We investigated the position of each centromere and its corresponding telomere and obtain the number of polarized configuration chromosomes or arms (chr4 are excluded, therefore the number ranges from 0 to 10). To identify the presence of this organization, we measure the angle between each centromere, the nuclear center, and its corresponding telomere. If the cosine of the angle is positive, then centromere and telomere are in the same hemisphere. Otherwise, they occupy

opposite hemispheres, forming polarized organization. If more than half of the chromosome arms (>=6) in one nucleus are in polarized organization, we consider this as a polarized nuclear structure.

C.4 Nuclear colocalizations of Hox gene clusters

The Antennapedia complex contains 5 genes located in 3 consecutive TAD domains, while the Bithorax complex contains 3 genes located in 2 TAD domains with one domain between them. Each control group contained two clusters, one cluster with 3 consecutive repressive domains and the other with 2 repressive domains separated by one domain. The 5 repressive domains were separated by the same linear distances as those in the Hox gene clusters. Because there are no available combinations of PcG TADs with the same genomic distances as the Hox gene clusters, the control data set involved the three types of repressive classes (Null, PcG and HP1 class). In total, we identified 30 combinations that meet the requirements of control groups. The average contact probability of each control group and hox gene clusters are shown in Fig. S5, top panel. All the contact probability of clusters are lower than 6%, which means the constraints are not imposed for those clusters in our models. If the closest surface-to-surface distance between two clusters in one group is less than 200 nm, we consider these clusters colocalized.

C.5 Pericentromeric heterochromatin cluster detection

We calculated all pairwise surface-to-surface distances (normalized by the sum of the radii of the domain spheres) among the 12 heterochromatin spheres, for each structure in the

population, then obtained the average pairwise distances in the matrix. Hierarchical cluster analysis is performed on the average distance matrix by using *hcluster* function in R.

C.6 Homologous pairing

A domain is defined as paired if the surface-to-surface distance between two homologs is less than 200 nm. Then the pairing frequency is defined as how often a domain is paired among the structure population. The domains with frequency higher than Top 3rd quantile are determined to be “tight”, otherwise, to be “loose”. We have 71 “active-tight” domains and 423 “active-loose” domains.

Pairing is not determined by the positon and the neighborhood crowdedness in the nucleus

The variation of the distance along homologous chromosomes and among nuclei raises the question of why certain regions attain higher level of homologous pairing than others in certain nuclei. First, we notice that the linear distance to centromere or to heterochromatin does not influence the extent of pairing of a domain (Fig. 6B), which exclude a possible influence of heterochromatin clustering on the extent of pairing. Next, we tested whether the 3D position of a domain in the nucleus influences pairing. We hypothesize that genomic regions near NE may have less space for movement, thus promoting homologous pairing. Indeed, the Pearson’s correlation between the contact frequency with NE (if the surface distance to NE is less than 50 nm, this domain is defined in contact with NE) and the homologous pairing frequency is 0.34 with p-value < 2.2e-16. Similarly, Pearson’s correlation between the average radial position and the homologous pairing frequency is 0.10 with p-value = 0.00049. Finally, we tested the

hypothesis that the crowdedness of the neighborhood around a domain influences this domain's pairing. The Neighborhood Crowdedness (NC) of a domain is defined as the number of other domains whose surface-to-surface distance to the domain is less than 200 nm. We calculate NC for each pair of homologous domains in individual models and compare the difference between paired and unpaired groups. The domains of paired groups have higher NC than the unpaired in 16.62% of models and lower in 6.84%, the rest of models (76.54%) show no significant difference. This data support the idea that the NC around domains does not significantly influence the pairing of homologous regions.

C.7 Epigenetic analyses

Chromatin domains were classified into four classes based on their epigenetic signatures: Active, Polycomb-Group (PcG), HP1/Centromere and Null [1]. Active domains comprise 42% of the domains with smaller domain size, and they are actively transcribed and characterized by high gene density. PcG domains are bound by PcG proteins and associated with the histone mark H3K27me3. HP1/Centromere domains are bound by the heterochromatin proteins HP1 and Su(var)3-9 and associated with H3K9me2. Null domains are not enriched for any of those marks.

We followed this 4-class annotation in our structure analysis. We also collected the data of histone modifications and binding of chromatin proteins in the study [4] from http://research.nki.nl/vansteensellab/Drosophila_53_chromatin_proteins.htm. Wig files were downloaded, and then transferred into bigwig format. BigwigSummary program (from USC Genome Browser) was used to extract the individual signals for requested

regions (the defined 1169 TADs in this study). The signal was calculated as an average per domain, avoiding the bias of the genomic length.

C.8 Transcription analyses

Gene expression data (embryonic samples collected at 16-18h) were obtained from the modENCODE project [7]. 1169 physical domains covered 12947 genes with available expression data. The number of genes in each domain varies, ranging from 0 to 170, and the average number is 11. The average gene expression values are calculated for each domain. RNA polymerase II binding data for Kc167 cells were also from modENCODE project (accession no. GSE20806, link http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20806). TBP (TATA-binding protein) is acomponent of the basal transcription machinery and the protein binding data are from [4]. BigwigSummary program was used to calculate the average signal for each defined domain.

C.9 DNA replication analyses

Data for ORC-binding regions and early activating replication origins for Kc167 cell line were downloaded from modENCODE (accession no. GSE20889 and GSE17285, respectively). BigwigSummary program was used to calculate the average signal for each defined domain.

C.10 Statistical test

The association test between two paired signals is done by *cor.test* function in R using Pearson's product moment correlation coefficient. For example, the positive correlation

between the frequency near NE derived from our population of structures and the lamina binding signal from lamina-DamID experiment provides a validation for our models; the negative correlation between the frequencies of homologous pairing derived from our population of structures and the Mrg15 protein binding signal from lamina-DamID experiment matches the unpairing function of Mrg15 protein and also validate our models.

The difference test between two sets is done by *wilcox.test* function in R, which performs a Wilcoxon rank sum test (equivalent to the Mann-Whitney test) when the population cannot be assumed to be normally distributed. For example, we found the ORC binding signals are much stronger in the active-loose than in the active-tight domains.

References Supplementary methods

1. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the Drosophila genome.** *Cell* 2012, **148:**458-472.
2. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA: **Iterative correction of Hi-C data reveals hallmarks of chromosome organization.** *Nat Methods* 2012, **9:**999-1003.
3. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.
4. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B: **Systematic protein location mapping reveals five principal chromatin types in Drosophila cells.** *Cell* 2010, **143:**212-224.
5. Pickersgill H, Kalverda B, de Wit E, Talhout W, Fornerod M, van Steensel B: **Characterization of the Drosophila melanogaster genome at the nuclear lamina.** *Nat Genet* 2006, **38:**1005-1014.
6. Kinney NA, Sharakhov IV, Onufriev AV: **Investigation of the chromosome regions with significant affinity for the nuclear envelope in fruit fly--a model based approach.** *PLoS One* 2014, **9:**e91943.
7. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al: **The developmental transcriptome of Drosophila melanogaster.** *Nature* 2011, **471:**473-479.

Supplemental Figures

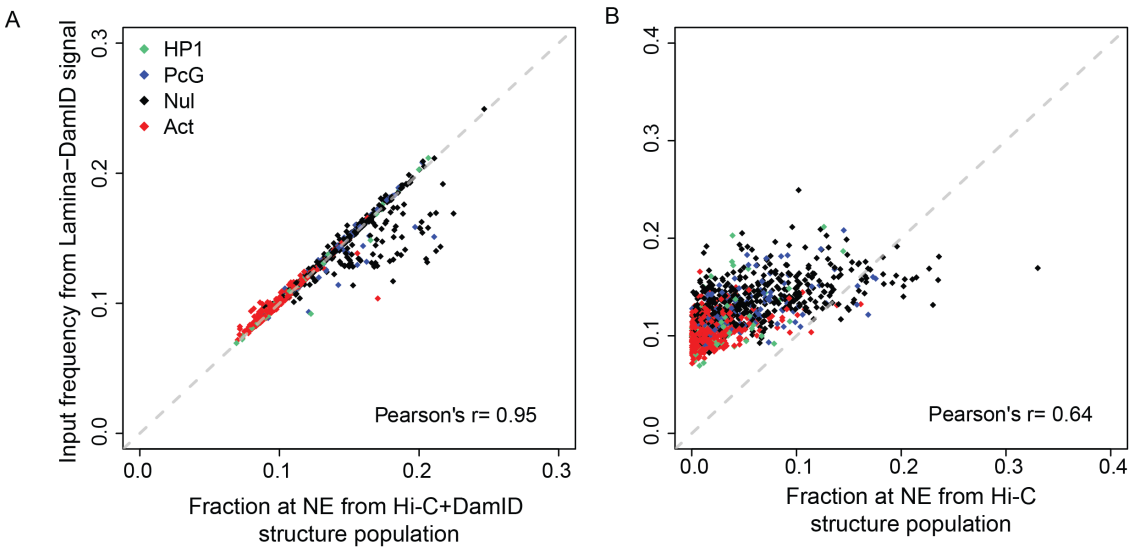

Figure S1. Agreement between the NE-association of euchromatin domains from lamina-DamID experiment and the models. (A) Fraction of domains at NE from population structure generated by data integration of Hi-C and lamina-DamID data well reproduces the input frequency derived from lamina-DamID data with Pearson's correlation coefficient=0.95 and p-value< 2.2e-16. The points are colored according to the epigenetic classes. (B) Fraction of domains at NE from the control model with a structure population generated only from Hi-C data has a good correlation with the frequency derived from lamina-DamID data (Pearson's correlation coefficient = 0.64 with p-value < 2.2e-16).

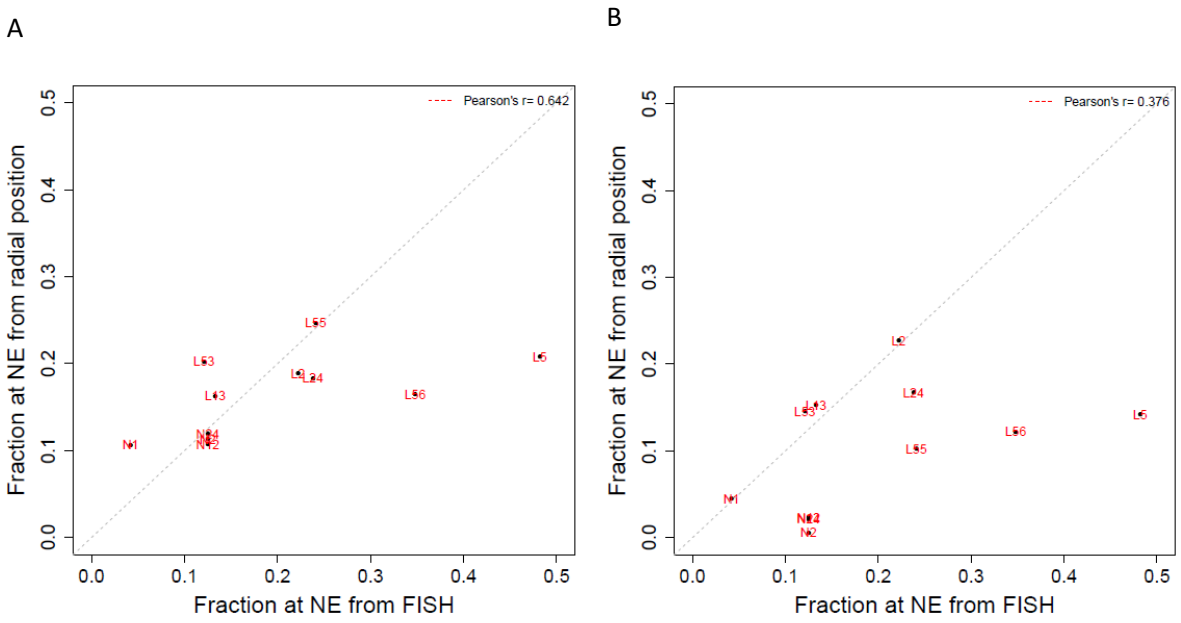

Figure S2. Agreement between the NE-association of individual loci from FISH experiment and the models. (A) Comparison of the NE association frequencies of individual loci from FISH experiment and from the model generated by data integration of Hi-C and lamina-DamID data. The NE association frequencies in the structure population agree well with FISH data for 11 loci (Spearman correlation coefficient=0.642 with p-value 0.03312). (B) Comparison of the NE association frequencies of individual loci from FISH experiment and the control model with a structure population generated only from Hi-C data. (Spearman correlation coefficient = 0.376 and p-value=0.2542).

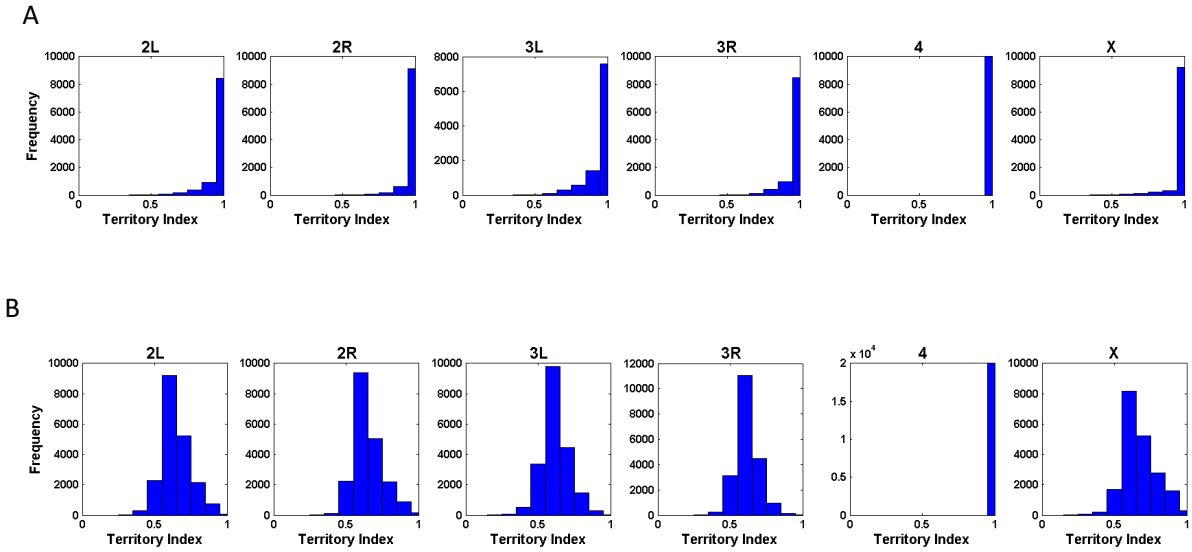

Figure S3. Territory index (TI). (A) TIs for the pairs of homologous chromosome arms. (B) TIs of each chromosome arm considering each homologues chromosome separately.

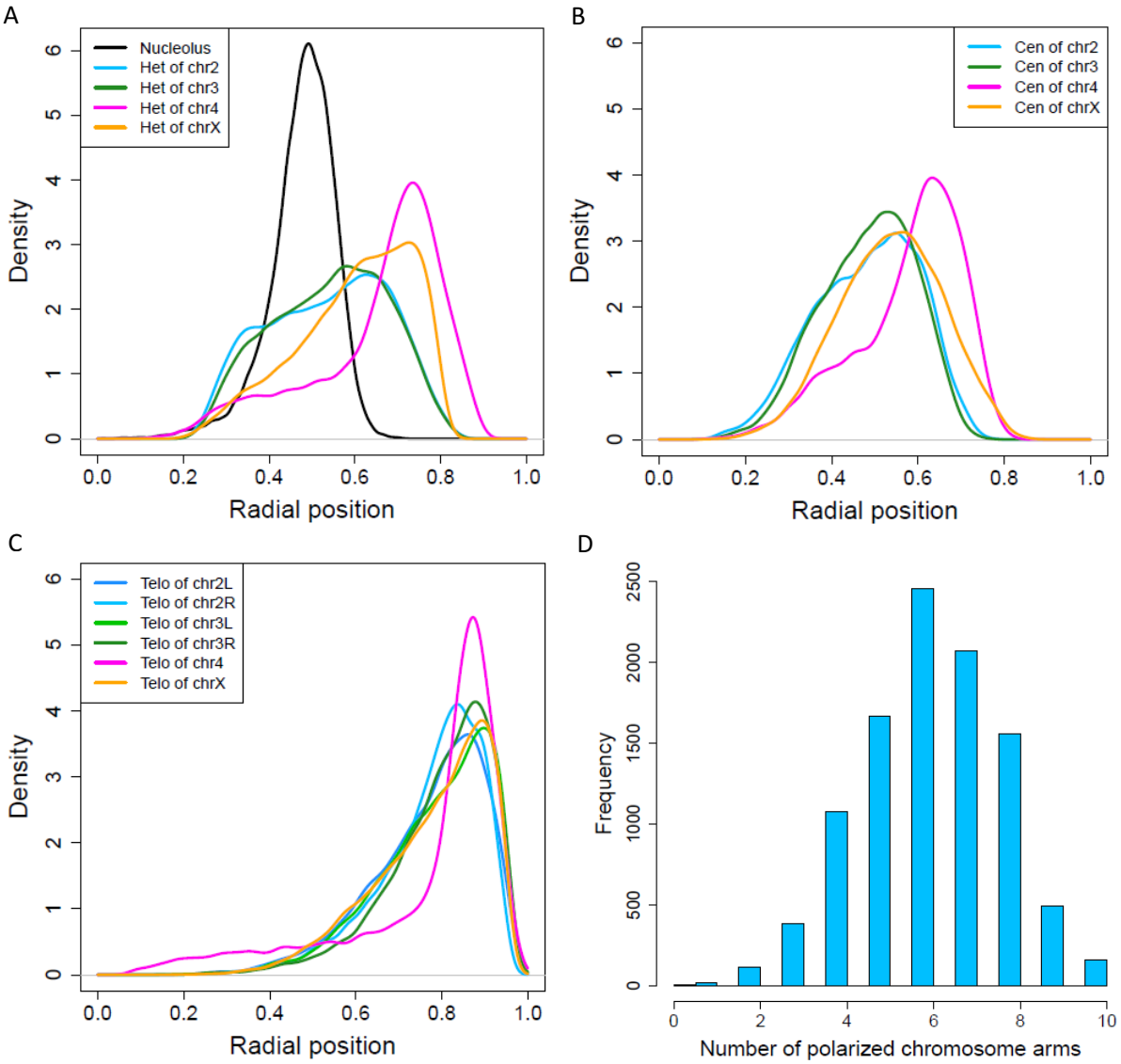

Figure S4. (A) Density plot of radial positions of the nucleolus and heterochromatin regions of different chromosomes. (B) Density plots of radial positions for centromeres. (C) Density plots of radial positions for peri-telomeric sequences. (D) Number of polarized chromosome arms (chr4 is excluded) among the population of structures.

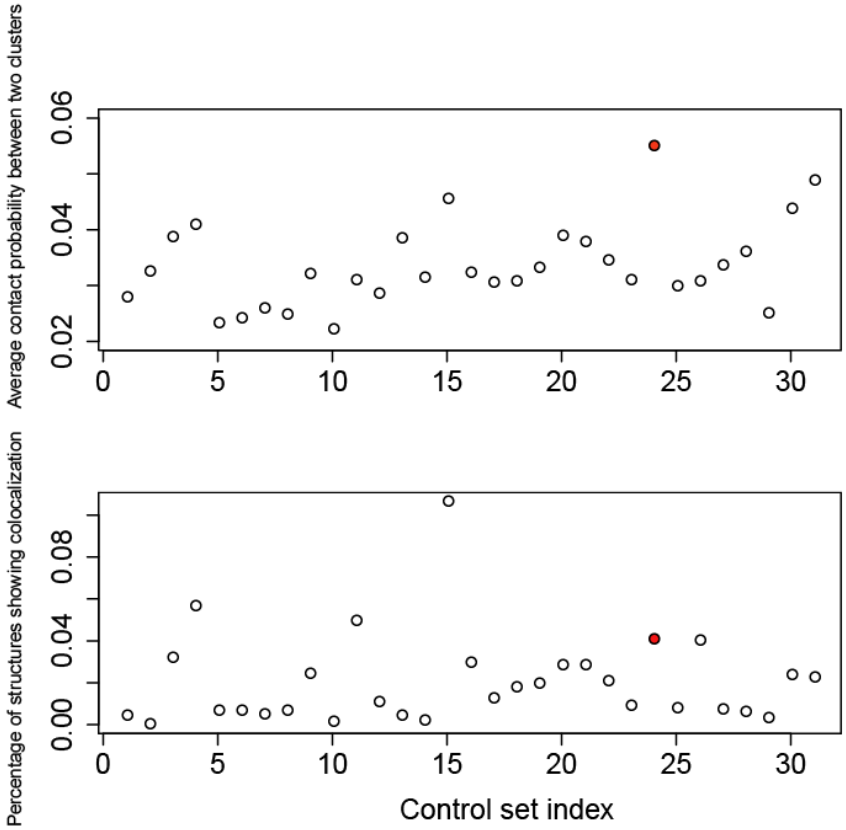

Figure S5. Hox gene clusters are prone to be co-localized comparing to control groups. (Top panel) Hi-C experiment shows the hox gene clusters have higher average contact probability than any other control clusters, and all the average contact probabilities within clusters (both hox gene and control) are smaller than 6%. The point for Hox gene clusters is highlighted by red color. (Bottom panel) The hox gene clusters show contacts in higher percentage of structures in the population compared to all other control clusters except three.

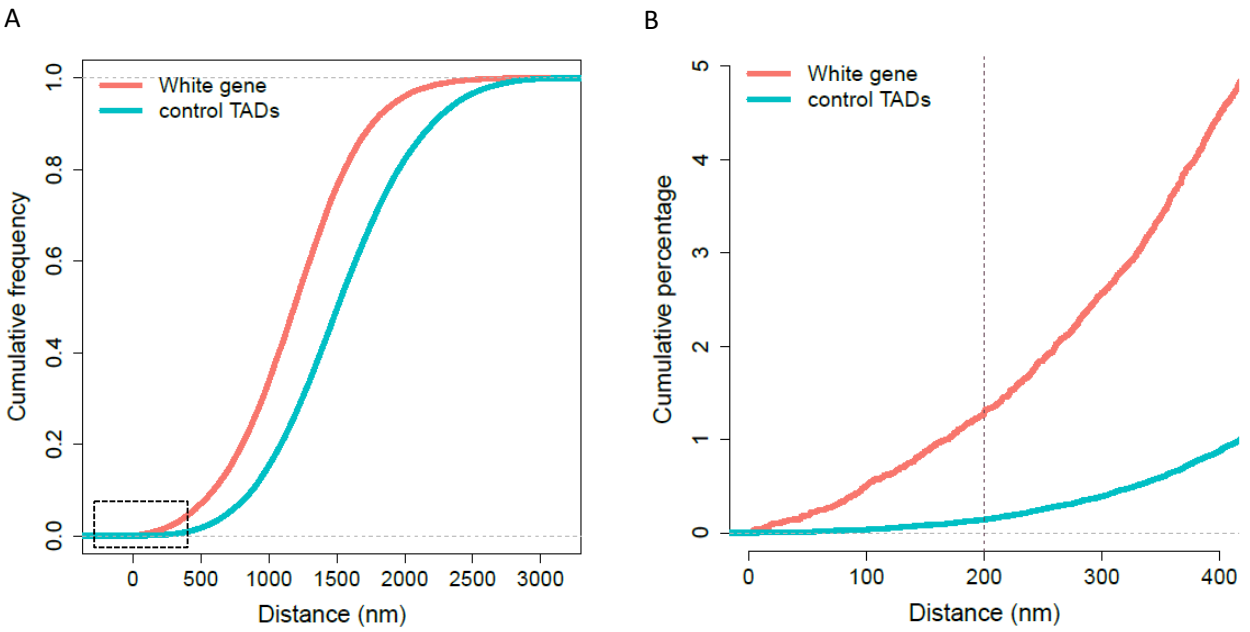

Figure S6. The *white* gene is prone to localize near to the pericentric heterochromatin. (A) Cumulative frequency plots for the distance of the *white* gene to its heterochromatin and of the control TADs to their corresponding heterochromatins. (B) Zoom the plot into the small distance, the *white* gene is 9-fold more frequently located proximal to pericentric heterochromatin (using 200 nm as a threshold), relative to control TADs.

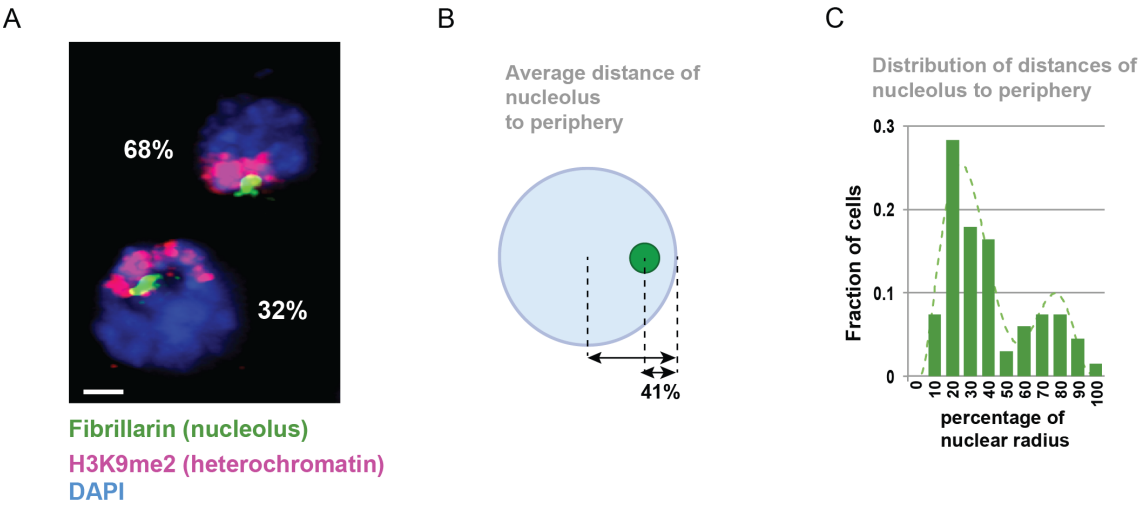

Figure S7. Nucleolus and heterochromatin positions in *Drosophila* Kc cells. (A) Immunofluorescence analysis with anti-Fibrillarin and anti-H3K9me2 antibodies, and DAPI staining for DNA (nuclear staining) shows the position and organization of the heterochromatin domain and the nucleolus in *Drosophila* Kc cells. The image shows a max intensity projection of two representative nuclei. Percentages indicate the population of cells in each configuration, *i.e.* with the nucleolus proximal to the nuclear periphery or more internal. N = 113 cells. Scale bar = 1 µm. (B) Quantification of the distance between the center of the nucleolus and the nuclear periphery shows the average position of the nucleolus relative to the center of the nucleus and the distribution of these distances in the cell population. N = 63 cells. (C) The distribution of the distance between the center of the nucleolus and the nuclear periphery show a bimodal distribution.

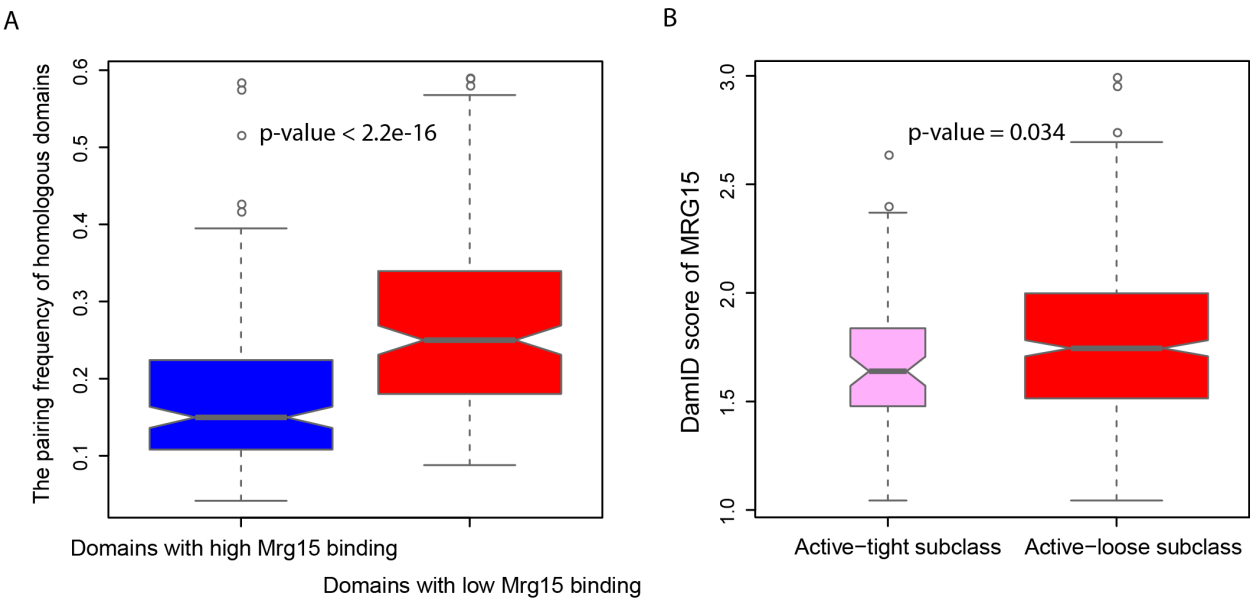

Figure S8. Inverse correlation between Mrg15 binding and homolog pairing frequency (A) Boxplot for the pairing frequency of domains with high and low Mrg15 binding. All domains are divided into 3 subsets based on their Mrg15 binding score. 293 domains are in the subset with high Mrg15 (top 25% binding scores); 293 domains are in the subset with low Mrg15 (bottom 25% binding score). The pairing frequencies for domains enriched with Mrg15 are significantly less than those for domains with low Mrg15 score (one-tailed Mann-Whitney U test, p-value < 2.2e-16). (B) Boxplot of Mrg15 score for the active-tight and active-loose subclasses. Active domains are divided into active-tight and active-loose based on their pairing frequencies (**Suppl. Methods C.6**). Active-tight domains, which have high pairing frequencies in our models, are significantly less enriched with Mrg15 comparing to active-loose ones (one-tailed Mann-Whitney U test, p-value = 0.03436).

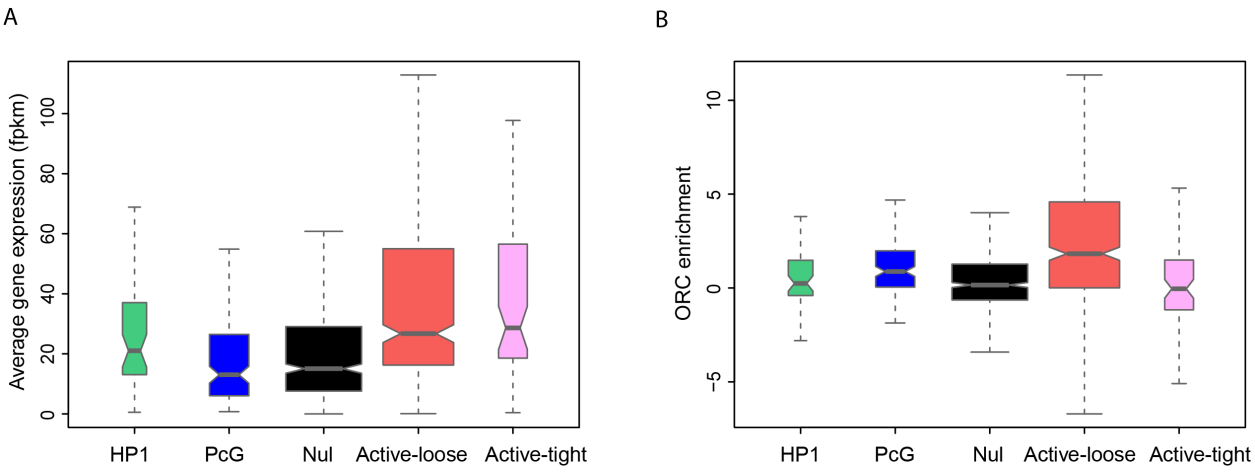

Figure S9. Transcriptional activity and DNA replication for all five classes (A) Domains in both two active subclasses have higher gene expression levels compared to those in the three repressive classes. (B) Domains in the active-loose subclass are more enriched with the replication complex ORC compared to the domains in the three repressive classes and the active-tight subclass. p-values: 1.48e-4, 7.59e-4, <2.2e-16 and 2.78e-5 respectively for HP1, PcG, Null and Active-tight.

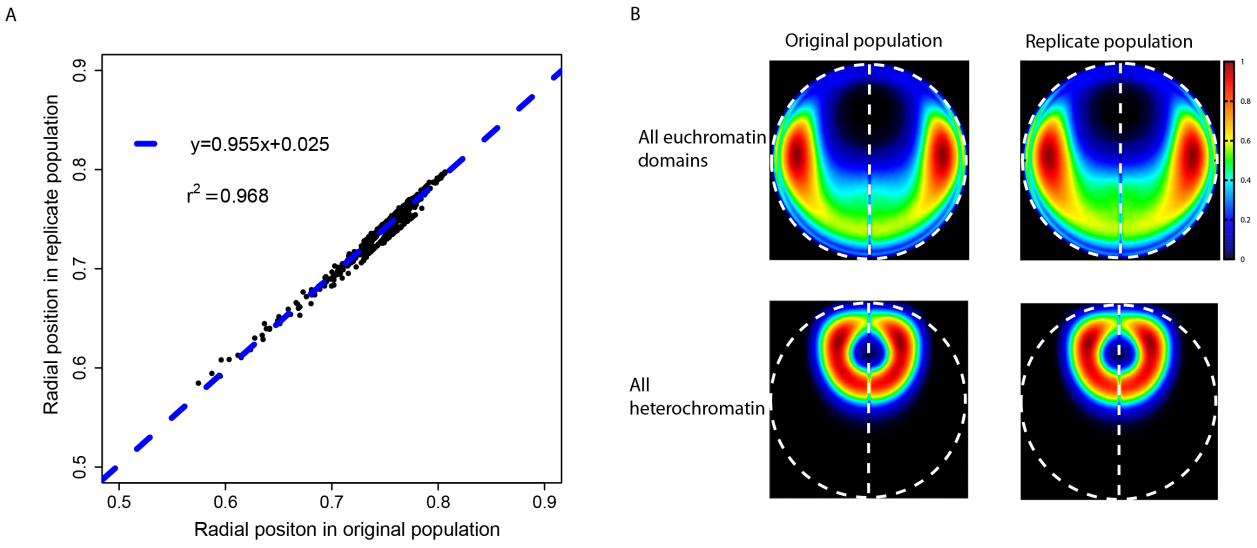

Figure S10. Reproducibility of the modeling method applied to the Drosophila genome (A) Agreement of the average radial positions between two populations of structures. The Pearson's correlation between them is 0.984, with p-value < 2.2e-16. (B) (Top panel) LPD plots of all euchromatin domains for the original population and the replicate population respectively. (Bottom panel) LPD plots of all heterochromatins for two populations of structures show highly consistent results.