bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Conformational analysis of chromosome structures reveals vital role of chromosome morphology in gene function

Yuxiang Zhan[1,2,3], Asli Yildirim[1,2], Lorenzo Boninsegna[1,2], Frank Alber[1,2,3*]

[1]Department of Microbiology, Immunology, and Molecular Genetics, University of California Los Angeles, 520 Boyer Hall, Los Angeles, CA 90095, USA

[2]Institute of Quantitative and Computational Biosciences, University of California Los Angeles, 520 Boyer Hall, Los Angeles, CA 90095, USA

[3]Department of Quantitative and Computational Biology, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA

*Correspondence should be addressed to F.A. falber@g.ucla.edu

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

The 3D conformations of chromosomes are highly variant and stochastic between single cells. Recent progress in multiplexed 3D FISH imaging, single cell Hi-C and genome structure modeling allows a closer analysis of the structural variations of chromosomes between cells to infer the functional implications of structural heterogeneity. Here, we introduce a two-step dimensionality reduction method to classify a population of single cell 3D chromosome structures, either from simulation or imaging experiment, into dominant conformational clusters with distinct chromosome morphologies. We found that almost half of all structures for each chromosome can be described by 5-10 dominant chromosome morphologies, which play a fundamental role in establishing conformational variation of chromosomes. These morphologies are conserved in different cell types, but vary in their relative proportion of structures. Chromosome morphologies are distinguished by the presence or absence of characteristic chromosome territory domains, which expose some chromosomal regions to varying nuclear environments in different morphologies, such as nuclear positions and associations to nuclear speckles, lamina, and nucleoli. These observations point to distinct functional variations for the same chromosomal region in different chromosome morphologies. We validated chromosome conformational clusters and their associated subnuclear locations with data from DNA-MERFISH imaging and single cell sci-HiC data. Our method provides an important approach to assess the variation of chromosome structures between cells and link differences in conformational states with distinct gene functions.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

With the advent of single cell super resolution imaging[1,2], multiplexed FISH imaging[3–6], single cell genomics experiments[7–10], and data driven genome modeling[11–22] it is now possible to analyze 3D structures of chromosomes and entire genomes at single cell level. Chromatin loops, topological associated domains (TADs) and patterns of chromatin compartmentalization are readily detected in ensemble averaged Hi-C data[23–26] but are very dynamic in nature and subsequently show large stochastic variations at single cell level[27,28]. For instance, chromatin loops, detected at specific locations in ensemble Hi-C are likely present only in 3 to 6.5% of cells at any given time[29] and TAD domain boundaries are only rarely observed at the ensemble average position but rather stochastically distributed, because of dynamic loop extrusion processes[13,28,30,31]. Thus, detailed analysis of single cell chromosome structures are only meaningful when considering the entirety of structural variability observed in a cell population[32–34].

Unlike chromatin loops and TADs, little research has been conducted on structural variations of long-range interactions and whole chromosome morphologies, specifically to investigate the role of these structural variations on global genome organization and gene function. Recent evidence from multiplexed FISH imaging[3,6,32] and single cell Hi-C experiments[10,35–37] suggests large structural variations of chromosome morphologies between single cells. These structural differences can affect spatial positions of genes within chromosome territories and thus a gene's exposure to functional compartments and nuclear bodies, which have been shown to be of relevance for gene function[38]. For instance, transcriptional activities of individual genes can be heightened in the immediate vicinity of nuclear speckles[39–41]. However, up to this point it remains unclear if the variability in 3D chromosome morphology plays any role in the regulation and cell-to-cell heterogeneity of gene function.

In this study, we address this point by examining the cell-to-cell variability of 3D chromosome morphologies within their nuclear environment and by studying how these structural variations alter the functional microenvironment of genomic regions in the nucleus, as defined by their radial positions, or distances to nuclear speckles and lamina compartments. Because of the stochastic nature of 3D chromosome structures, several important questions emerge. First, can the structures of the same chromosome in different cells be classified into prevailing structural states that define distinct chromosome morphologies? Second, do chromosome morphologies

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of prevailing structural states relate to distinct functional properties of genes in these chromosomes?

To address these questions, we first introduce an approach to classify structural variations of chromosome morphologies in single cells from ensembles of 3D chromosome structures, extracted either from multiplex DNA-MERFISH imaging[3], or structure models generated with our data-driven structure modeling approach[11]. We then study if chromosomal regions are exposed to different nuclear microenvironments in different structural states to detect potential functional variations of genes located in different chromosome morphologies.

Due to the dynamic nature of chromosome structures, their classification based on 3D coordinates is challenging, because some functionally unrelated regions can show large degree of randomness in their relative positions, overshadowing relevant structural relationships between other chromosomal regions. Our approach overcomes this problem and transforms the problem of classifying individual chromosomes structures to a problem of detecting maxima of a density distribution in a reduced 2-dimensional space, where each data point represents a chromosome conformation and the detection of local maxima in the probability density function determines locations of highly occupied clusters of chromosomes with similar 3D conformations. Thus, our approach determines subpopulations of chromosomes with similar 3D morphology. We discovered that a given chromosome can be clustered into around 5 to 10 morphology classes, which are distinguished by the presence or absence of characteristic chromosome territory domains that vary in their relative locations to each other. The boundaries of these territory domains play a fundamental role in establishing conformational variation and their sequence locations are shared across various cell types (GM12878, H1-hESC and HFFc6). We validated the observed chromosome conformational states and chromosome territory domain boundaries with data from multiplex DNA-MERFISH imaging data[3] and single cell sci-HiC experiments[10]. We then discovered that distinct chromosome morphologies (i.e. conformational states) favor certain nuclear locations of some chromosomal regions and thus modulate the functional properties of these genomic regions. For instance, preferences in radial positions, distances to the nearest speckle, nucleolus and lamina differ substantially between the same chromosomal regions in different conformational states. These observations point to functional differences of chromatin in different conformational states, as smaller distances to nuclear speckles are typically associated with increased transcriptional activities of genes. Our observations therefore indicate that chromosome morphologies can play a key role in modulating functional properties of some chromosomal regions, and can, at least partially, be

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

responsible for the cell-to-cell heterogeneity of the expression of some genes. Our method provides an important approach to study chromosome conformational variations and reveal links between conformational states of chromosomal regions and gene functions.

Results

Structure generation.

We first apply our approach for characterizing chromosome morphologies and their functional qualities to an ensemble of diploid 3D genome structures that were generated at 200kb resolution from Hi-C data[11,13,25,26]. Our population-based 3D genome modeling method (Integrative Genome Modeling, Methods)[11,13,42,43] provides us with a large sampling of 10,000 diploid 3D genome structures per cell type, which, as a whole, reproduce Hi-C data and predict with high accuracy other orthogonal experimental data[11,13], namely average radial positions of genomic regions from GPseq experiments[44], mean speckle distances from SON TSA-seq[45], mean distances and contact frequencies to the nuclear periphery from lamin B1 TSA-seq[45] and lamin B1 pA-DamID[46], respectively. Moreover, predicted chromosome structures show good agreement with single cell 3D chromosome conformations from multiplex DNA-MERFISH experiments[3,11,13], and also reproduce with good accuracy speckle and lamina association frequencies of genomic regions from DNA MERFISH[11,13]. We first focus our analysis on genome structure models from lymphoblastoid cells (GM12878) from previously published work[13], human fibroblast (HFFc6)[11] and human embryonic stem cells (H1ESC)), before classifying genome structures from DNA-MERFISH experiments[3].

Approach.

To classify chromosomes based on their morphology, we first extract individual chromosomes of a given type from each of the whole genome structures in the cell population. Both homologous chromosome copies in the diploid genome are selected, resulting in a total of 20,000 chromosome structures for each autosome (**Fig. 1**). For each 3D chromosome structure, we then construct a distance matrix, which then serves as input into our dimension reduction and clustering scheme (**Fig. 1**). We then use a two-step dimension reduction approach to cluster chromosome structures based on their distance matrices into conformational states (Methods). Specifically, each normalized distance matrix is represented as a 2D image. Our two-step process then combines a convolutional autoencoder (consisting of an encoder and a decoder

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

module) with a dimension reduction step using t-distributed stochastic neighbor embedding (t-SNE)[47] (Methods). The encoder module reduces a distance matrix to a latent vector that can reconstruct the original matrix by the decoder module. The method reduces dimensions, while preserving enough information to reconstruct the original image. To construct a convolutional autoencoder we use convolutional layers, max pooling layers and up sampling layers, which is frequently used for image embedding and classification (**Supplementary Fig. 1**) (Methods). t-SNE, a method to separate data points in a reduced data space, is then used to map the latent vectors (generated by the autoencoder) to a lower dimensional space (**Fig. 1**). Finally, we use a kernel density estimation to calculate a probability density function (pdf) that represents the chromosome conformational space in the t-SNE reduced dimensions. The resulting density probability matrix shows a balanced distribution of local maxima separated by deep valleys (**Fig. 1** lower panels), indicating the presence of a number of preferred conformational states per chromosome, which are subsequently identified by a segmentation algorithm (Methods).

We also tested other clustering methods. However, principal component analysis (PCA), multidimensional scaling (MDS)[48], locally linear embedding (LLE)[49], isomap[50] and spectral embedding (SE)[51] methods are unable to determine distinct clusters with chromosomes of similar conformational morphology, while uniform manifold approximation & projection (UMAP)[52] and T-distributed stochastic neighbor embedding (t-SNE) (applied directly to distance matrices alone) produced unbalanced clusters, in which the majority of structures were part of only a single cluster (**Supplementary Fig. 2**). Instead, the balanced distribution of local maxima in the resulting density probability matrix of our two-step clustering approach (**Fig. 1** lower panels) indicates the presence of a number of preferred conformational states per chromosome.

A large fraction of chromosome structures can be clustered into a few conformational states.

We then identify clusters of similar chromosome conformations by determining local maxima in the probability density distribution as cluster centers and identify structures associated to each cluster center by watershed segmentation of the probability density distribution (**Fig. 1** lower left panel, and Methods). Chromosome structures part of the same segmentation are in the same conformational cluster. For chromosome 6 about 40% of all chromosome structures can be clustered into 8 dominant conformational clusters (**Fig. 1**, lower left panel, **Fig. 2A**). The occupancy of each cluster is defined by the number of structures in a cluster divided by the total number of all clustered chromosome conformations. The occupancy among clusters varies (**Fig.**

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2A). For chromosome 6, cluster 4 has the highest occupancy containing ~40% of all the clustered structures, while all other clusters each occupy less than 20% of all clustered structures. Similar results are found also for other chromosomes (**Supplementary Fig. 4**).

Preferred conformational states define distinct chromosome morphologies.

While chromosomes within each cluster share similar morphology, chromosomes between clusters differ in their structures. For instance, when we measure the compactness of a chromosome structure by calculating its radius of gyration, it is apparent that the structures of each cluster vary largely in their shapes (**Fig. 2B**). Chromosomes in cluster 8 show the lowest compaction with an average radius of gyration that is about 50% larger than chromosomes in cluster 4, which show the most compact structures (**Fig. 2B**). Overall, chromosomes in clusters with relatively low compaction, and thus, highest radius of gyration show also the largest variations of their compaction values within the cluster (e.g., clusters 6, 8). Moreover, cluster 4 containing chromosomes with the most compact structures shows the highest occupancy (**Fig. 2AB**).

Next, we quantify the structural similarity between chromosome structures within and between clusters by calculating the Wasserstein distance[53] between their pairwise chromatin distance distributions. Specifically, we measure the difference between distributions of all intra-chromosomal distances calculated from all chromosomes in each cluster. In other words, for a given pair of chromosomal regions i and j, we calculate the distance distribution from all chromosomes in a cluster and assess its similarity with the corresponding distance distribution calculated from the structures in another cluster. The similarity between two such distance distributions is calculated by their Wasserstein distance metric[53]. The combined distance measure between two clusters is then defined as the average of all Wasserstein distances between all intra-chromosomal distance distributions calculated from the chromosomes in the two clusters (**Fig. 2C**). We normalize this measure by the average Wasserstein distance for chromosome structures within the same cluster (Methods). We observe that the average Wasserstein distance is always substantially larger (i.e., ~2-4 fold) for structures in different clusters, showcasing the structural distinction between chromosomes in the different clusters. We also found similar results when assessing clusters with other distance measures, including a Euclidean distance measure and Gaussian dissimilarity[32,54], confirming an overall higher similarity between structures within than between different clusters (Methods) (**Supplementary Fig. 3**).

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Importantly, each cluster shows distinct average contact frequency matrices, calculated from the physical chromatin contacts of all chromosomes in each cluster (**Fig. 2D**, average contact frequency matrix shown in fifth panel from top). These distinct contact patterns confirm the presence of characteristic chromosome morphologies in each cluster. A characteristic feature of these contact patterns is the presence of domain boundaries that divide the chromosome territory into large units, with increased interaction frequencies within, and reduced contact frequencies between territory domains (for instance in clusters 1, 3, 6 and 8 of chromosome 6, **Fig. 2D** (territory domain locations are indicated by green blocks below the contact frequency matrix.)) These territory domains are particularly evident when calculating the average distance matrix for each cluster (i.e., from all intra-chromosomal 3D distances in a chromosome), because territory domains show increased spatial distances from each other, thus are separated spatially from each other in 3D space (see for instance domains 2 in clusters 3 and 6 in **Fig. 2D**). The spatial separation between domains explains the reduced contact frequencies between the territory domain regions. The location and size of territory domains vary between the clusters. On average, chromosome structures contain between 1 and 4 territory domains per studied chromosome (green blocks below in **Fig. 2D, Supplementary Fig. 4DH**). **Figure 2D** also shows representative structures of the chromosome morphology found in each cluster of chromosome 6. The structures show the spatial separation between territory domains, as observed in the average distance matrices. For instance, cluster 6 of chromosome 6 shows a territory domain boundary at around 134 Mb sequence position, which separates the q-arm terminal end of the chromosome into a separate territory domain (domain 2 in cluster 6 (134-171 Mb), **Fig. 2D**). This domain shows relatively increased spatial distances and low chromatin contact frequencies to other chromosomal regions upstream of the territory domain boundary (**Fig. 2D**). In contrast, cluster 3 shows a domain boundary at sequence position 155 Mb, which forms an even smaller domain at the q-terminal end of chromosome 6 (domain 2 of cluster 3 (155-171 Mb), **Fig. 2D**), which is well separated from the bulk of the remaining chromosome (also evident in the representative chromosome structure (lower panel).) Cluster 2 contains a relatively small chromosome territory domain at the p-arm terminal end of chromosome 6 (domain 1 in cluster 2 (0-10 Mb), **Fig. 2D**). Noticeable, the boundaries between territory domains act as hinge regions allowing the relative positions of territory domains to vary in 3D between models, while the territory domain itself appears as structural units (representative structures in **Fig. 2D**). Accordingly, chromosomes in different clusters also show differences in their local chromatin compaction, as measured by the radius of gyration (RG) over a 1 MB window of the chromatin fiber (**Fig. 2D**, third profile panel from top**)**. Local peaks in the

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RG profile are regions with relatively low fiber compactness (which often correspond to TAD boundaries as previously shown[11,13]). These profiles show distinct differences between clusters, noticeable at locations of some territory domain boundaries. To highlight these differences, we calculated the RGRatio as the log ratio of the average RG value for a chromatin region in the cluster and the overall ensemble. For instance, the RGRatio profile of both domain 3 boundaries in cluster 8 show high values, indicating that these boundary regions are decompacted in cluster 8 in comparison to the same region in the overall ensemble (**Fig. 2D**, RG and RGRatio profiles are shown in the third and fourth panel from top.) Also, the boundary that separates the small domain 2 in cluster 3 from the bulk of the chromosome territory shows substantially increased RGRatio, thus shows a substantial decrease in fiber compactness in the cluster in comparison to the compactness in the overall ensemble of chromosomes, therefore allowing domain 2 the freedom to loop away from the bulk of the remaining chromosome territory (see black arrow in RGRatio of cluster 3 in **Fig. 2D**). These observations indicate that local chromatin properties can facilitate the formation of specific chromosome morphologies.

Besides chromosome 6, also other chromosomes show very similar results, with distinct chromosome contact frequency patterns for each cluster. For instance, conformational clusters for chromosome 8 and chromosome 10 are also distinguished by a total of 8-10 different chromosome morphologies with distinct territory domains whose locations vary between individual clusters (**Supplementary Fig. 4**).

Chromosome clusters can be validated by imaging experiments.

We assessed our findings with data from multiplex DNA-MERFISH imaging experiments[3], which traced 3D coordinates of whole diploid genomes in IMR90 fibroblast cells at a step size of ~3Mb. To allow a direct comparison, we down sampled our models to the genomic regions sampled in the experiment and classified the chromosome conformations from DNA MERFISH into clusters (Methods), based on the similarity of their distance matrix with cluster averages in our models (**Fig. 3AB**). For chromosome 6, around 60% of all chromosome structures from DNA MERFISH (about 4,000 structures) can be classified into our predicted chromosome clusters based on their structural similarity (**Fig. 3A-E)**. When we average the distance matrices of all imaged chromosome structures in each cluster, we see that the cluster averages from DNA-MERFISH experiments are almost identical to those calculated from our models (**Fig. 3AB**). Also in experiment, cluster 4, with the most compact chromosome structures, shows the highest occupancy (**Fig. 3F**). Moreover, individual representative single cell chromosome structures from DNA-MERFISH imaging show almost identical chromosome structures and

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

distance matrices to those from our predicted models (**Fig. 3CD**). Thus, DNA-MERFISH imaging confirms the presence of preferred chromosome morphologies and the presence of chromosome territory domains that vary in their locations between the clusters.

Chromosome morphologies show distinct preferences in nuclear locations of chromosomal regions.

We now focus on the nuclear organization of chromosomes in different morphologies. The question we want to address is: does the morphology of chromosome structures relate to specific nuclear locations of chromosomal regions and thus modulate their functional properties? Because we model whole genome structures, we can analyze chromosome structures in their nuclear context. First, we can extract the nuclear radial positions of chromosomal regions and average the radial positions from chromosomes in each conformational cluster (**Fig. 2D**, top profile panel). We assess the differences of the radial (RAD) profile between clusters by calculating the RadRatio, defined as the log ratio between the average radial position of a genomic region (RAD) in a cluster and its value in the whole population of clustered structures (**Fig. 2D**, second profile panel from top). A negative RadRatio value for a genomic region indicates that its average radial position in the cluster is closer to the nuclear interior in the cluster than in the overall population as a whole. We see that RadRatio profiles differ substantially between clusters for chromosome 6 in particular for clusters 1, 3, 5, 6 and 8, which show pronounced peaks in the RadRatio profile, both in positive and negative values. For instance, the RadRatio profiles in clusters 1 and 3 differ substantially across the entire chromosome (**Fig. 2D**). In cluster 1 the genomic location 24-48 Mb (region **I** of cluster 1 in **Fig. 2D**), which includes the MHC gene cluster, is substantially shifted towards interior nuclear positions (i.e. negative RadRatio values) in comparison to the location of the same region **I** in clusters 3, 5 and 6, which show positive RadRatio values, thus more exterior locations than in the population average (p-values of Welch's t-test[55] on average radial position against clusters 3, 5 and 6 = 4.06e-12, 1.02e-03 and 6.17e-09 (Table 1), **Fig. 2D**). Instead, cluster 3 shows a small chromosome territory domain at the q-terminal chromosome end, readily visible in the contact frequency and distance maps (genomic location 155-171 Mb: domain 2/region **III** in cluster 3 in **Fig. 2D**). This territory domain loops towards the nuclear interior in cluster 3, as shown in the RadRatio profile and the representative structures (**Fig. 2D**). In other clusters (e.g., cluster 1 and 2) the same region **III** shows a more exterior nuclear location and does not form a separate domain, but is part of an overall larger chromosome territory domain (see region **III** in

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

clusters 1 and 2, **Fig. 2D**). Another example is the genomic region $\mathbf{II}$ at sequence location 105-114 Mb. It forms a small territory domain in cluster 8 (domain 3, region $\mathbf{II}$ in cluster 8 in **Fig. 2D**), which shows strongly negative RadRatio, and therefore loops towards the nuclear interior. Instead, the same region $\mathbf{II}$ in cluster 6 is part of a larger territory domain, which restricts its looping towards the nuclear interior in comparison to cluster 8 (p-values = 1.04e-09, Welch's t-test on average radial positions in cluster 8 and 6 (Table 2)). In cluster 2 the same region $\mathbf{II}$ is even shifted more towards the nuclear periphery in comparison to the population average (see negative RadRatio for regions $\mathbf{II}$ in cluster 2 in **Fig. 2D**)(p-values = 4.95e-17, Welch's t-test on average radial positions in cluster 8 and 2 (Table 2)). Moreover, in cluster 4 chromosomes are in their most compact conformation, which correlates with a shift of almost the entire chromosomal regions towards the nuclear periphery in comparison to the population average (see positive values in RadRatio profile in **Fig. 2D**) (p-values = 1.69e-34, Welch's t-test on average radial positions in cluster 8 and 4 (Table 2)).

Chromosome morphologies are likely linked to differences in gene functions.

The differences in radial positions of genomic regions could indicate differences in their functional properties. It is known that transcriptionally active chromatin is more likely located towards the nuclear interior, and highly transcribed genes are often found close to nuclear speckles[3,23,56]. It has been previously shown that smaller mean speckle distances of genes correlate with high transcriptional activity[13,39–41]. We therefore examine if chromosome conformations influence speckle associations of genomic regions. We previously showed that our models can predict with good accuracy locations of nuclear speckles in single cell models[11,13] by identifying the geometric centers of highly connected clusters of speckle-associated chromatin in each model. Our models predicted with good accuracy data from SON TSA-seq experiments[45], which measure the mean distances of genomic regions to speckles (0.87 Pearson's correlation coefficient between predicted and experimental data[13]), as well as speckle association frequencies (SAF) of genomic regions from DNA-MERFISH imaging experiments (0.79 Pearson's correlation coefficient between prediction and experiment[13]). When we predicted SAF, SON-TSA-seq and mean speckle distances (SpD) for chromosomes in different clusters, we noticed that all these profiles vary considerably for chromosomes in different clusters (**Fig. 4A**, top 3 panels). For instance, in cluster 8 the more interior nuclear location of chromosomal region $\mathbf{II}$ (genomic location: 105-114 Mb) (**Fig. 4A**) leads to a substantially decreased mean speckle distance (**Fig. 4B)**, as shown by the higher predicted

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SON TSA-seq signals and a ~four times larger SAF value in comparison to the same region in cluster 6 (**Fig. 4A)**, which is buried in a larger chromosome territory (p-value = 2.36e-06, Welch's t-test on average speckle distances between cluster 6 and 8 (**Fig. 4B,** Table 2). We showed previously that a higher SAF, thus smaller mean speckle-distance, generally correlates with higher transcriptional activity[13]). We therefore speculate that in cluster 8 genes in this region (105-114 Mb – region **II** in **Fig. 4A**), if transcriptionally active, are predisposed for higher transcriptional activity than the same regions in cluster 6 (**Fig. 4A**). Moreover, chromosomes in cluster 4 have the most compact structures and almost all chromosomal regions, with the exception of the MHC gene cluster, are predicted to have larger speckle distances and decrease speckle association frequencies (SAF) in comparison to chromosomes in other clusters (**Fig. 3A**). We validated our predicted speckle distances and SAF values using the clustered chromosome conformations taken from DNA-MERFISH imaging[3], which also image speckle locations. These images confirm that the mean speckle distances and SAF profiles vary considerably for chromosomes in different clusters (**Fig. 3B**). For a better comparison, we calculated the SafRatio, defined as the log ratio between the SAF of a genomic region in the cluster and the overall value in the ensemble of all clustered structures. The most compact conformations in cluster 4 shows indeed a more exterior nuclear location and reduced speckle association frequencies for regions across the entire chromosome, except the MHC genes, in comparison to more extended chromosome conformations in all other clusters, confirming our predictions (see reduced SafRatio (Methods), **Fig. 3B**). For instance, cluster 1 shows a distinctly different SafRAtio in comparison to cluster 4 and 7, with certain chromosomal regions showing opposing SafRatio values, indicating that these regions show different speckle distances in different clusters. For instance the p-terminal end of chromosome 6 in cluster 1 (0-24 Mb (indices 0-8) in **Fig. 3AB**) shows substantially increased SAF over the population average in both the predicted and experimental cluster 1, while the same region in cluster 4 and 7 show decreased speckle association frequency. This observation may indicate an increased transcriptional propensity for genes in the region (0-24 Mb (indices 0-8)) in cluster 1. Moreover, region IV (indices 14-20) in cluster 1 shows decreased speckle association frequencies in both experiment and prediction (SafRatio in **Fig. 3AB**), while the same region shows increased speckle association frequency in clusters 2, 3 and 8, in both the experiment and the prediction. Overall, the experimental structures confirm the predicted patterns of increased and reduced speckle associations across the chromosome clusters (see magenta and green shaded blocks in SafRatio of **Fig. 3AB**), even though the experiments were done on IMR90 cells with a different nuclear shape. In summary, our analysis reveals that changes in chromosome

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

conformations can be linked to changes in specific nuclear locations of genomic regions relative to nuclear bodies, which possibly affect their functional properties.

Characteristic features of chromosome territory domain boundaries.

Next, we investigate the characteristic properties that define a territory domain (TD) border (**Fig. 2D**). We noticed that if a territory domain undergoes a major change in nuclear position in a cluster (i.e., RadRatio shows pronounced minima or maxima) its boundaries show lower chromatin compaction to facilitate the passage of the domain to the interior (or exterior) regions of the nucleus (**Fig. 4AC**). The local compaction of the chromatin fiber can be calculated from the radius of gyration (RG) of a chromosomal region (Methods): The log ratio between the RG profile in a cluster over the population average (RgRatio) shows a sharp peak when a TD boundary is present in a cluster, as is shown for cluster 8 (**Fig. 4AC**) (Methods). Thus, these boundary (bd) regions show substantially reduced chromatin compaction in chromosomes of cluster 8 than in the population average (**Fig. 4C**). Interestingly, boundaries of domain 3 (region II) in cluster 8 seem to have an alternative second downstream boundary (bd_2'), which allows the inclusion of a small gene cluster into the domain in some structures (**Fig. 4C**). In contrast, in cluster 4, where there are no domain boundaries present at these positions, chromosomes do not show any decompaction at the corresponding chromosomal regions, on the contrary, they are slightly more compacted with lower RG values than observed in the population average (RgRatio < 0, RGRatio is defined as the log ratio of the RG value for a genomic region in the cluster and its value in the ensemble (Methods) (**Fig. 4C**).

Interestingly, TD boundaries are often located close to transitions between gene poor and gene rich chromosomal regions (**Fig. 4C**). The territory domain itself contains chromatin with histone modifications related to active chromatin (e.g. H3K27ac, H3K4me1 and H3K4me3) (**Fig. 4C, Supplementary Fig. 5B, 6B, 7B**). These domains have positive SON TSA-seq signals of intermediate strength, indicating that these regions can be close to nuclear speckles in some structures. The domain boundary contains genes lacking H3K27ac and other activating histone modifications but instead are marked often with repressive histone modifications (such as H3K27me3), which mark the boundary to gene poor regions outside the domain (**Fig. 4C**). Also, domain boundaries often mark transitions between Hi-C subcompartments, as defined by Rao et al[25]. For instance, most territory domain boundaries in chromosome 6 coincide with boundaries between the A2 and B3 subcompartments. Finally, we also noticed that territory

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

domains are also chromosomal regions with generally high cell-to-cell variability in their radial positions (see δRAD profiles in **Fig. 4C**).

Inter-chromosomal interactions are specific to chromosome clusters.

Chromosome morphologies influence the predisposition of chromosomal regions to form inter-chromosomal interactions. In some morphologies chromosomal regions may be shielded from inter-chromosomal interactions, while the same region in another morphology may be exposed to other chromosomes. To quantify the role of chromosome morphology on inter-chromosomal interactions, we calculated for each cluster the inter-chromosomal proximity matrix, as the frequency of a genomic region to be in spatial proximity with a specific region of other chromosomes.

As expected, the proximity matrices for a given chromosome in different conformational clusters vary considerably (**Fig. 5A**, second panels from the left). To quantify these differences we calculated the IPP profile for chromosome 6, defined as the total number inter-chromosomal proximities for a given chromosomal region averaged over all genome structures in a given cluster (Methods). To compare the IPP profiles between different clusters, we calculated the IppRatio, defined as the ratio of the IPP profiles in a cluster and the whole population of clustered structures (Methods). IppRatios of different clusters vary substantially (**Fig. 5A**, right profile panels). As expected, the IPPRatio of chromosome 6 in cluster 4 shows reduced exposure to inter-chromosomal interactions across the entire chromosome, due to the more compact chromosome structure observed in cluster 4 in comparison to those in other clusters (**Fig. 5A**). Indeed, cluster 4 shows the overall lowest IPPRatio values. Overall, cluster 8 shows the highest IPPRatio. However, there are differences for individual domains in each cluster. For instance, domain 3 in cluster 5 shows higher IPPratio then the corresponding region in cluster 8 (region indicated by green bar in **Fig. 5A**). Interestingly, chromosomes in different clusters favor interactions to different chromosomes (**Fig. 5B**). Chromosome 6 in cluster 8 shows the highest averaged IppRatio and thus substantially increased inter-chromosomal interactions with chromosomes 2, 21 and 8, while chromosome 6 in cluster 5 shows increased interchromosomal interactions with chromosome 20 and 16 instead (**Fig. 5B**). For instance, **Figure 5C** compares the inter-chromosomal proximity matrix between chromosome 6 and chromosome 2 for cluster 8 and cluster 4. Chromosome 6 shows increased inter-chromosomal interactions with chromosome 2 in cluster 8 in comparison to cluster 4 (**Fig. 5C**).

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Chromosome clusters can be validated by single cell Hi-C experiments.

We further assessed our findings by single cell Hi-C (sci-HiC) data of GM12878 cells[10], containing more than 11,000 single cell contact maps, each with relatively low coverage (on average 3,879 contacts per cell in 200kb resolution) (**Fig. 6AB**). To increase the relatively low contact coverage, we applied the scHiCluster method[57], a single cell Hi-C imputation method based on linear convolution and random walk algorithms (**Fig. 6B**). About 8,500 imputed single cell contact matrices of chromosome 6 could then be classified into clusters based on their similarity to the average contact maps of our detected clusters (Methods) (**Fig. 6CD**). The average contact frequency maps for each cluster for both, the imputed and raw sci-HiC maps, show good agreement with those from our models (**Fig. 6AB, Supplementary Fig. 8B**). To ensure that the agreement is not due to overfitting, we generated a control experiment, where contact entries in each sci-HiC contact matrix were randomly rearranged while maintaining its diagonality and the number of overall contacts. Performing the same analysis with randomized sci-HiC matrices, resulted in cluster averages that did not reproduce the contact patterns in our clusters (**Supplementary Fig. 8C**).

Finally, we also used an available data set from single nucleus sn-HiC experiments[36,37] of WTC11 cells (made available by the laboratory Bing Ren), which contained only 188 single cell contact maps (on average 129,499 contacts per cell in 200kb resolution) and were preprocessed by the imputation method Higashi[27]. Despite the small sample size, we found representative single cell maps of chromosome 6, with similarities in their contact patterns to those found in our clusters (**Fig. 6E**). Therefore, single cell/nucleus Hi-C data confirms the presence of predicted chromosome morphologies, even though these data do not distinguish between homologous copies. We assume that when phased single cell data becomes available the already good agreement will be further improved.

Chromosome morphologies are conserved across cell types.

Chromosome morphologies are distinguished by territory domain boundaries, which are found at a few chromosome locations and vary per cluster. To determine if chromosome morphologies and the locations of territory domain boundaries are conserved between different cell types, we applied the same clustering protocol to genome structures generated from Hi-C maps of two additional cell types, the human embryonic cell H1 hESC, and the fibroblast cell HFFc6. **Figure 7** compares the averaged contact patterns of all highly occupied conformational clusters for chromosome 6 between all the three cell types (H1 hESC, HFFc6, GM12878). In all cell types

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

we detected very similar clusters containing chromosome structures with very similar chromosome morphologies and thus similar locations of chromosome territory domain boundaries. The only exceptions are clusters 5 and 7, which were not detected in H1 hESC cells. Thus, preferred chromosome morphologies and chromosome territory domains are largely conserved between these cell types. However, the number of structures in each cluster, i.e. the cluster occupancy, can vary between the cell types.

Discussion

Our manuscript addresses one of the key challenges in genome biology, namely, how to systematically characterize the cell-to-cell variability of whole chromosome structures and analyze the role of structural stochasticity in gene function. Due to the large dynamic variability of genome structures in a population of cells, clustering of whole chromosome and genome structures is very challenging. Traditional tools used in structural biology fail to detect functionally relevant structural similarities in chromosome morphologies, because some functionally unrelated regions can show large degree of randomness in their relative positions, overshadowing relevant structural relationships between other chromosomal regions. Subsequently, so far, little attention has been given to the heterogeneity of long-range chromosome morphologies between cells and the impact of these variations on genome function[33]. Here we present the first large-scale quantitative assessment of the 3D structural variability of whole chromosome morphologies that also considers the context of the nuclear environment. To achieve this goal, we introduce an efficient two-step clustering method that combines convolutional autoencoder with t-SNE dimension reduction to cluster large ensembles of single cell whole 3D chromosome structures, either from genome structure models or multiplex FISH imaging, into dominant conformational clusters with characteristic chromosome morphologies. Importantly, these chromosome morphologies are analyzed in their nuclear context, that is, in relation to the entire single cell diploid genome as well as relative locations to nuclear bodies and compartments. Our findings were validated by independent experiments from multiplex DNA-MERFISH imaging and single cell Hi-C.

We found that almost half of the structures for each chromosome can generally be described by 5-10 dominant chromosome morphologies, which play a fundamental role in establishing conformational variation of chromosome structures. Each morphology cluster is distinguished by a specific combination of 2 to 4 chromosome territory domains that vary between clusters.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Territory domains are consecutive chromosomal regions with enhanced interactions that can be spatially separated from other territory domains. Interestingly, we found that the detected chromosome morphologies and locations of territory domain borders are largely conserved across different cell types and similar conformational clusters are found in GM12878, H1-hESC and HFFc6 cells. However, the relative cluster occupancy can vary between the cell types.

Importantly, our analysis not only discovered dominant chromosome morphologies, but also identified the relationship of chromosome morphologies with specific nuclear microenvironments of chromosomal regions. Specifically, we discovered a link between chromosome morphologies and specific subnuclear properties of chromosomal regions, such as specific preferences in nuclear positions and associations to nuclear speckles, lamina, and nucleoli. Our analysis shows that chromosome morphologies can either enhance or shield the exposure of specific chromosomal regions to the nuclear interior, exterior or nuclear speckles. Subsequently, the radial positions and speckle association frequencies for the same chromosomal region can differ substantially in different chromosome morphologies. It is known that shorter distances to nuclear speckles can enhance gene expression efficacy for some transcriptionally active genes[39–41]. Thus, by modulating the exposure of chromosomal regions to specific nuclear microenvironments, chromosome morphologies could influence chromatin function in single cells. Our work indicates that some chromosomal regions may show functional distinction in different chromosome morphologies, which could contribute to the heterogeneity of gene transcription in single cell RNA-seq experiments[58,59]. This information is crucial to uncover the role of genome structure on regulatory processes of genome function.

Most prominently we observe that some territory domains allow chromosomal regions to undergo relatively large changes in their nuclear position between different morphologies. These chromosomal regions show overall large cell-to-cell variability of their radial positions in the population of cells and are often also part of the active chromatin compartment that are embedded between extended regions of inactive B compartment chromatin. The corresponding territory domain borders are often close to transitions between the A2 and B3 Hi-C subcompartment. These regions show intermediate SON TSA-seq signals, indicative of active chromatin with a larger mean distance to nuclear speckles and lower gene expression levels than active chromatin in the A1 subcompartment, which is speckle associated in almost all cells of a population. Another interesting conclusion is that our results provide a connection between local chromatin structure properties and the presence of chromosome morphologies. For instance, we found that if a territory domain is present in a chromosome morphology, the corresponding domain boundary region shows increased chromatin fiber decompaction in

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

comparison to the whole ensemble average. One could speculate that a decreased occupancy of cohesin (or other chromatin properties) at these locations in some single cells could favor the presence of a chromosome domain border at a specific sequence location, and subsequently could favor the presence of a specific chromosome morphology.

Finally, we also showed that genome structural models can facilitate the classification of multiplex FISH imaging data. Because of the currently relatively low coverage in multiplex FISH chromosome tracing experiments, a direct clustering of chromosome structures from experiments is challenging and can fail to detect clusters. However, as demonstrated here, it is possible to cluster chromosome structures from models generated at higher resolution. The resulting structural clusters can then facilitate the classification of chromosome morphologies in the DNA-MERFISH data. Thus, high resolution modeling of chromosome structures can also play a key role for the structural analysis of low-resolution 3D structures determined by chromosome tracing in multiplex FISH imaging experiments.

Acknowledgements

This work was supported by the National Institutes of Health (grants U54DK107981 and UM1HG011593 to F.A). We thank Nan Hua, Wenyuan Li, Francesco Musella and Ye Wang for useful discussions. We thank Ruochi Zhang and Jian Ma for providing Higashi imputed sn-HiC of 188 cells provided by the Bing Ren lab.

Author contributions

Y.Z. and F.A. designed research. Y.Z. performed all calculations and initial data analysis..Y.Z. and F.A. interpreted results and data analysis. Y.Z. wrote software and documentation. Y.Z., A.Y. and L.B. generated genome structures and structure features. Y.Z. and F.A. wrote the initial draft of the manuscript. Y.Z., F.A., L.B. and A.Y. edited the manuscript. All authors approved the final manuscript. F.A. secured funding.

Competing interests

The Authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

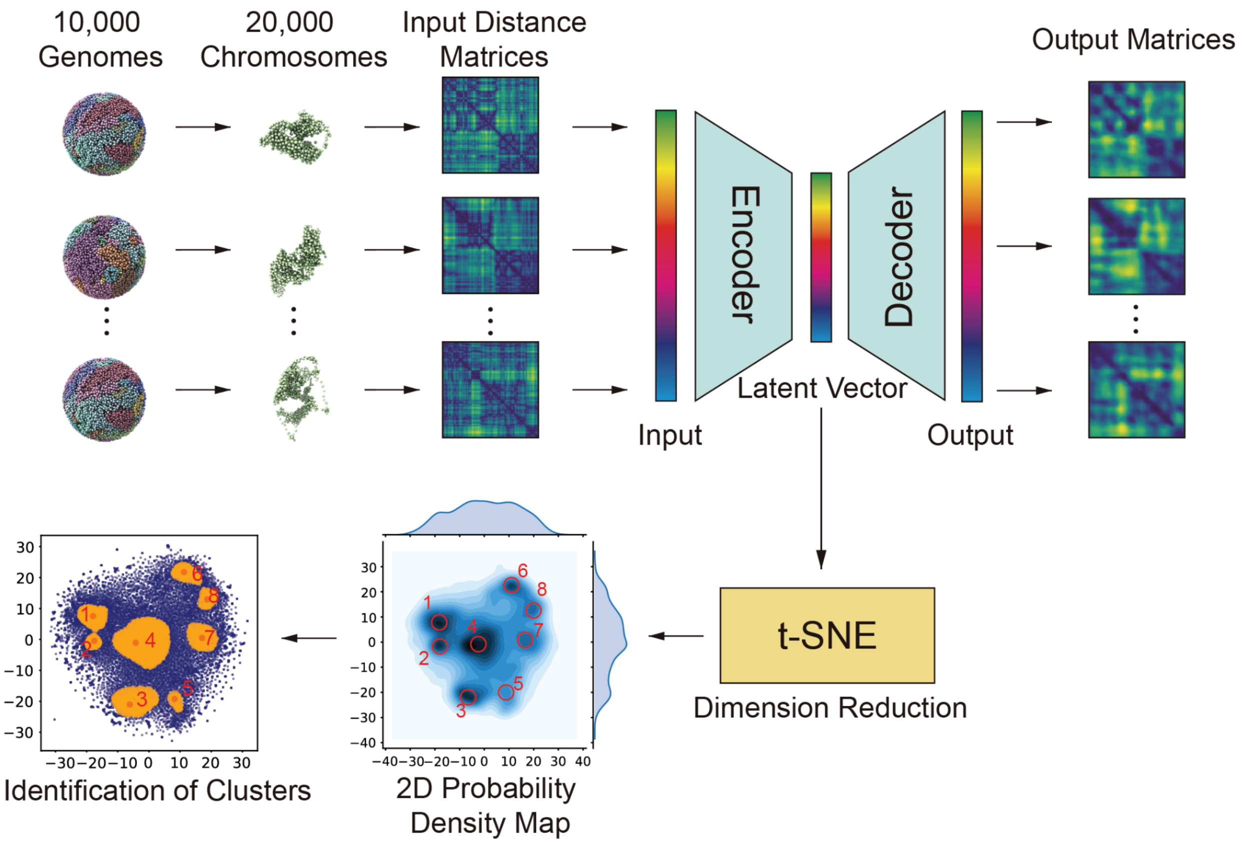
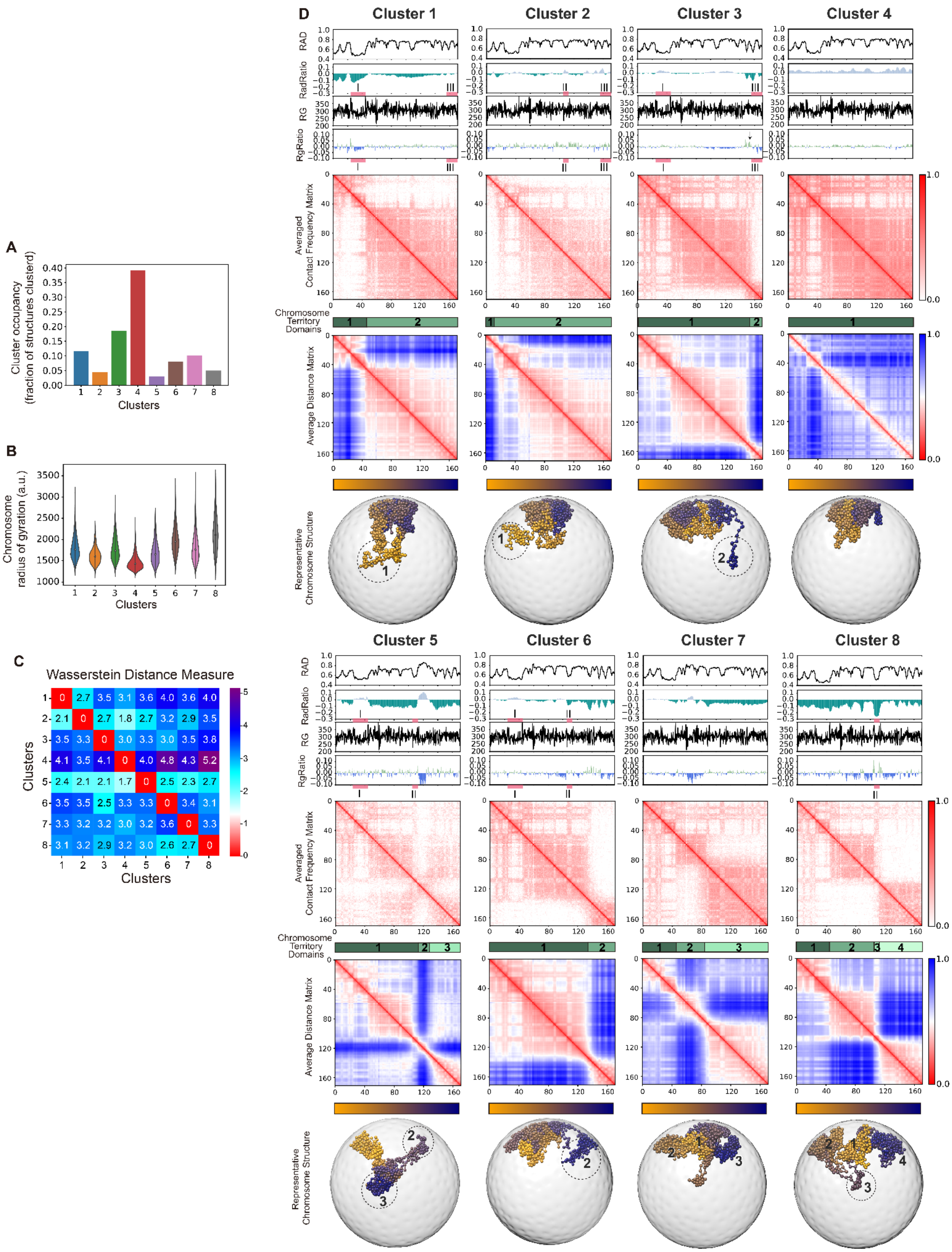


Figure 1: Overview of the two-step dimension reduction Every chromosome structure is represented by an input distance matrix, which is constructed by calculating pairwise Euclidean distances between each pair of loci in the chromosome structure. After preprocessing, the matrix is then used as the input of the autoencoder. After minimizing the loss between input matrices and output matrices, the latent vectors are then embedded by t-SNE[47] to obtain a distribution of all chromosome structures in 2D space. The resulting distribution is further used for peak detection and identification of clusters of chromosome structures (Methods).

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2: Clustering of chromosome 6 structures reveals dominant chromosome morphologies. A, The cluster occupancy of the 8 predicted clusters for chromosome 6. The occupancy is defined by the fraction of structures in each cluster. **B,** The distributions of the radius of gyration of all structures in each cluster (Methods). **C,** Pairwise dissimilarity measure between chromosome structures in 8 clusters. The dissimilarity matrix is calculated by measuring the average Wasserstein distance[53] of all intra-chromosomal distance distributions between two clusters. Each entry represents the log fold ratio between the inter-cluster dissimilarity and the intra-cluster dissimilarity, where positive values indicate the inter-cluster dissimilarity is larger than the intra-cluster dissimilarity. **D,** For each cluster the following information is shown (Methods): (top panel) The average radial position profile (RAD) calculated from all all chromosome structures in the cluster; (second panel from top) RadRatio: the log fold ratio of the average radial position in the cluster with respect to full ensemble average; (third panel from top) RG: the radius of gyration of a 1Mb chromosomal region centered at the target locus (fourth panel from top) RGRatio: the log fold ratio of RG in the cluster with respect to the value in the full ensemble; (fifth panel from top) The average contact frequency matrix calculated from all structures in a cluster; (sixth panel from top) different shade of green indicate the location of chromosome territory domains; (seventh panel from top) The average distance matrix calculated from all chromosome structures in the cluster; (eighth panel from top) Selected example of a chromosome structure in the cluster. Numbers and circles indicate chromosomal regions of the corresponding chromosome territory domains. The color bar indicates the sequence position of each chromosomal region. We also highlighted several specific genomic regions (regions I, II and II) below the RadRatio and RGRatio profiles, which are compared and discussed in the text.

.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

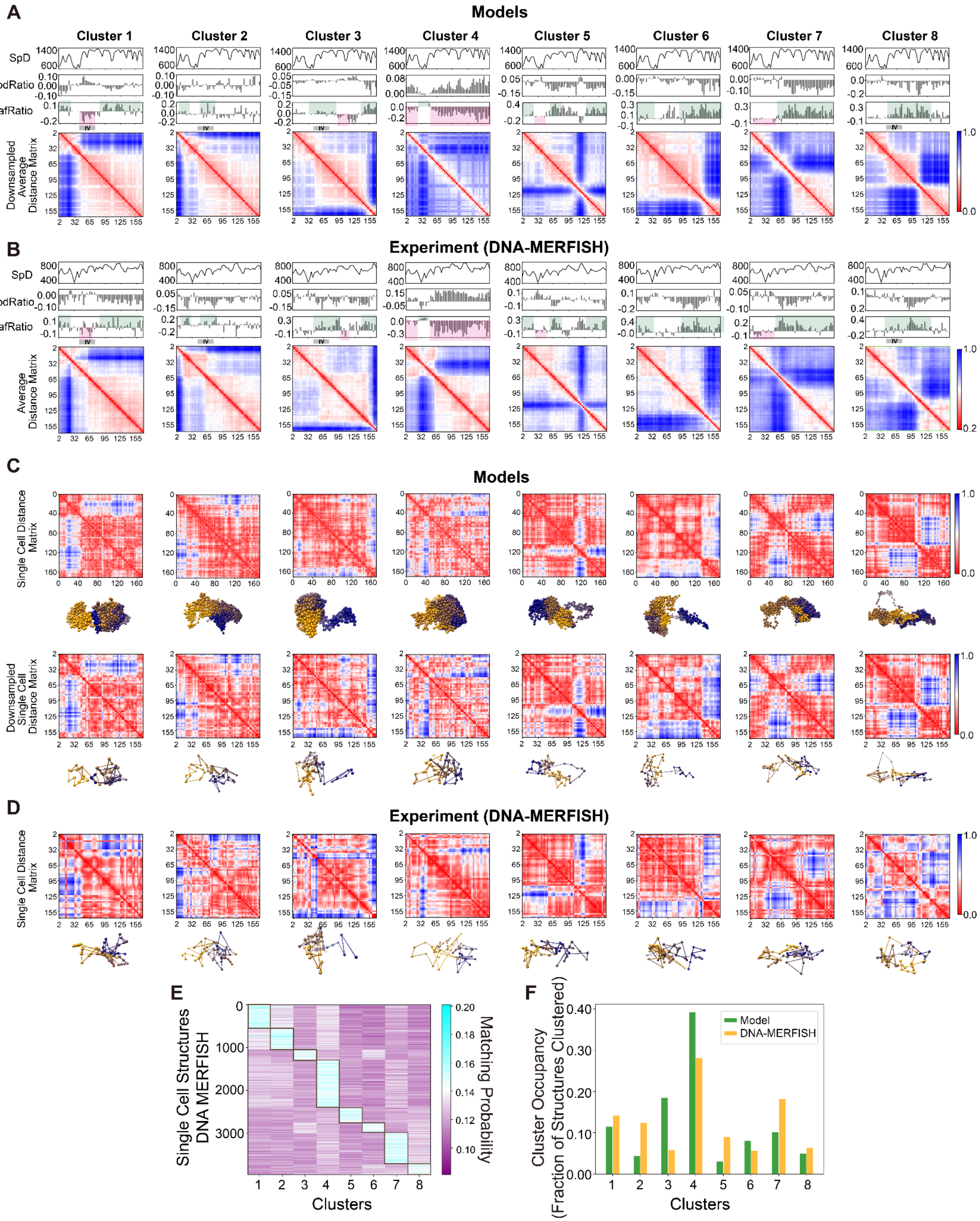


Figure 3: Chromosome clusters can be validated by imaging experiments. A, Average distance matrices from modeled chromosomes (chromosome 6) in each cluster downsampled to the respective coverage as observed in

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DNA-MERFISH experiments[3]. Tick labels of all distance matrices indicate sequence location in Mb. Also shown are several structural features calculated from the genome structures of the cluster ensemble, namely the average distance to the nearest speckle (SpD), the log fold ratio of the average distance to nearest speckle (SpdRatio) in the cluster with respect to full ensemble average and the log fold ratio of the speckle association frequency (SafRatio) in the cluster with respect to the full ensemble average. Green and magenta shaded areas indicate coarse matches of the varied SafRatio in predicted and experimental determined clusters. For comparison with experiment distance matrices were down sampled to the same coverage in the experiment (window size 3Mb). **B,** The corresponding average distance matrices of chromosomes from DNA-MERFISH experiments for each cluster. Shown are also the average distance to the nearest speckle (SpD) in each cluster, the log fold ratio of the average distance to nearest speckle (SpdRatio) in each cluster and the log fold ratio of the speckle association frequency (SafRatio) in each cluster, all measured from DNA-MERFISH imaging[3]. Green and magenta shaded areas indicate coarse matches of the varied SafRatio in predicted and experimental determined clusters. **C,** Selected representative examples of single cell modeled structures and the corresponding downsampled version for chromosome structures for each cluster. Matrices are calculated with window size 200kb. **D,** Average distance matrices and selected representative single cell chromosome structures from DNA-MERFISH experiment for each cluster. In panels A, B, lower panel C and D distance matrices are shown at 3Mb resolution (coverage in DNA MERFISH experiment); in upper panel C the distance matrix is shown at 200 kb resolution. **E,** Matching probabilities indicating the similarities between distance matrices of all classified single cell DNA-MERFISH chromosome structures against all modeled clusters. Note that around 60% of the single cell structures are successfully classified and assigned to one of the modeled clusters. **F,** Comparison of the cluster occupancy between the chromosome conformational clusters observed in our models and corresponding chromosome conformational clusters from DNA-MERFISH experiments.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

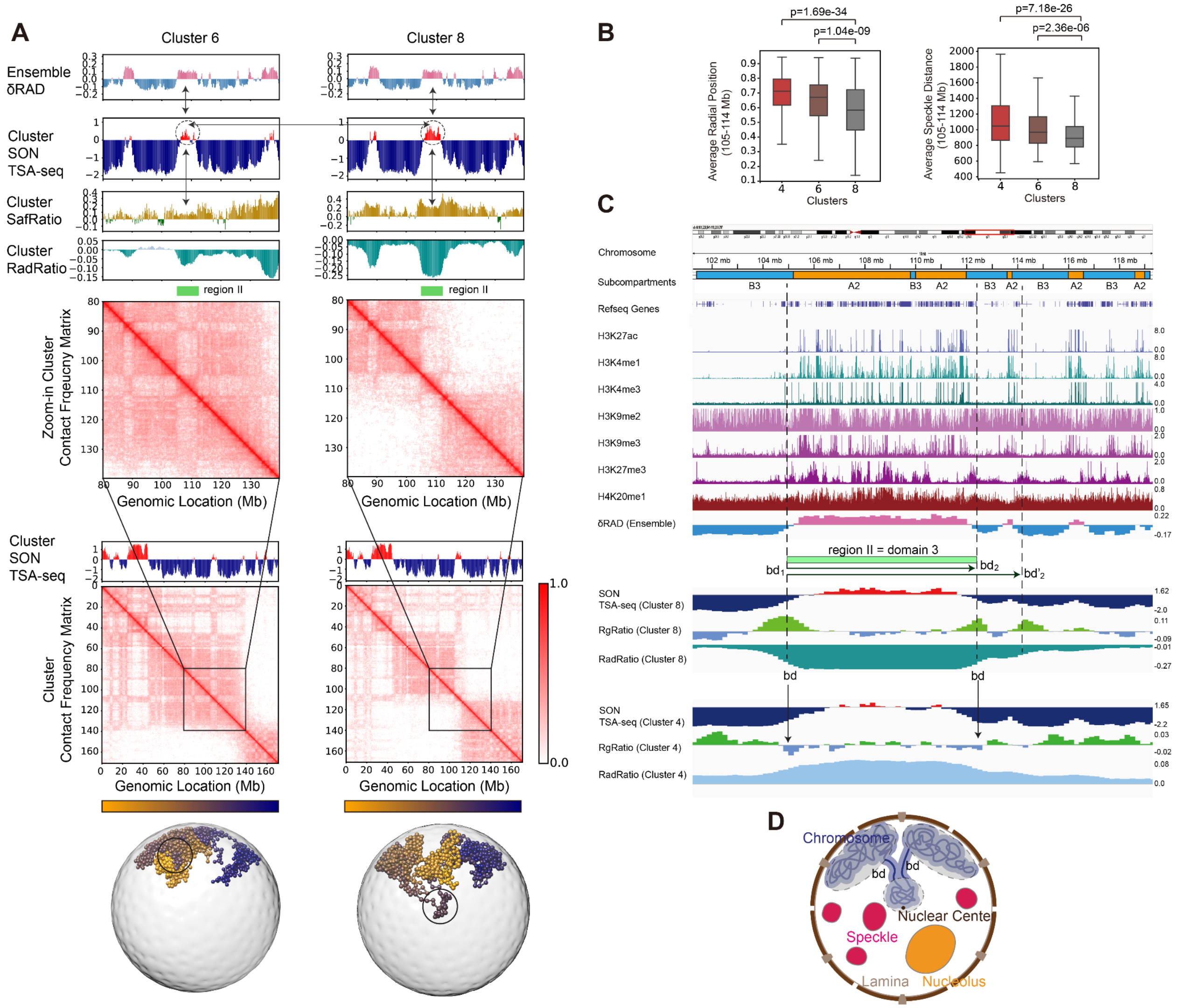


Figure 4: Chromosome morphologies show preferences in nuclear locations A, Contact frequency matrices and 4 profiles of different structural properties for chromosome 6 in clusters 6 and 8. (Top panel) The structural variability (δRAD) of chromosomal regions (Methods). Positive values in red indicate regions of high structural variability in the ensemble of all structures. Negative regions in blue indicate regions with low structural variability in the ensemble. (Second panel from top) SON TSA-seq predicted from genome structures in each cluster. Positive values indicate shorted mean distances to nuclear speckles. (Third panel from top) SafRatio, log ratio of SAF calculated from chromosomes in the cluster over SAF calculated from structures in the whole ensemble (Methods). (Fourth panel from top) RadRatiocalculated from chromosomes in the cluster, RadRatio is defined as in Figure 2 and Methods. (Fifth panel from top) Average contact frequency matrices calculated from structures in the cluster. The first five panels show data for a zoomed-in genomic region in chromosome 6. The sixth and seventh panels from top show the predicted cluster SON TSA-seq and cluster average contact frequency matrices for the full length chromosome 6. The bottom panel shows selected representative structures for chromosome 6 in each cluster. Also indicated is region II (genomic location (105-114 Mb)), which is discussed in the text. **B,** Distributions of the average radial position and average speckle distances for a region II (genomic location: 105-114 Mb) in clusters 4, 6 and 8 as well as the p-values of Welch's t-test[55] between two pairs of clusters. Differences between clusters of the radial positions and average speckle distances are significant. **C,** Characteristic features for a chromosomal region spanning across region II, which forms territory domain 3 in cluster 8 and the same region II in cluster 4. Shown are epigenetic marks and other features for this sequence region. From the top to the bottom, the displayed features are chromosome

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sequence location, Hi-C subcompartments, refseq genes, H3K27ac, H3K4me1, H3K4me3, H3K9me2, H3K9me3, H3K27me3, H4K20me1, and the ensemble structural variability (δRAD) calculated from all structural models. In addition the following features are shown for the same regions calculated from cluster 8 and cluster 4: SON TSA-seq, RgRatio and RadRatio (Definitions as in Figure 2 and 4, Methods). Also shown are lines that indicate the territory domain in cluster 8 and corresponding domain boundaries that overlap with regions of reduced chromatin compaction (RGRatio) (bd1 and bd2). For bd2 two alternative boundaries exist in the cluster (bd2 and bd2'). **D,** Illustration of schematic features of a chromosome morphology with three territory domains. Shown are also nuclear bodies. bd regions indicate domain boundaries that show increased decompaction of the chromatin fiber (i.e., RG) in comparison to the ensemble average, and which allow the territory domain to loop towards the nuclear interior, while other territory domains remain at the periphery.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

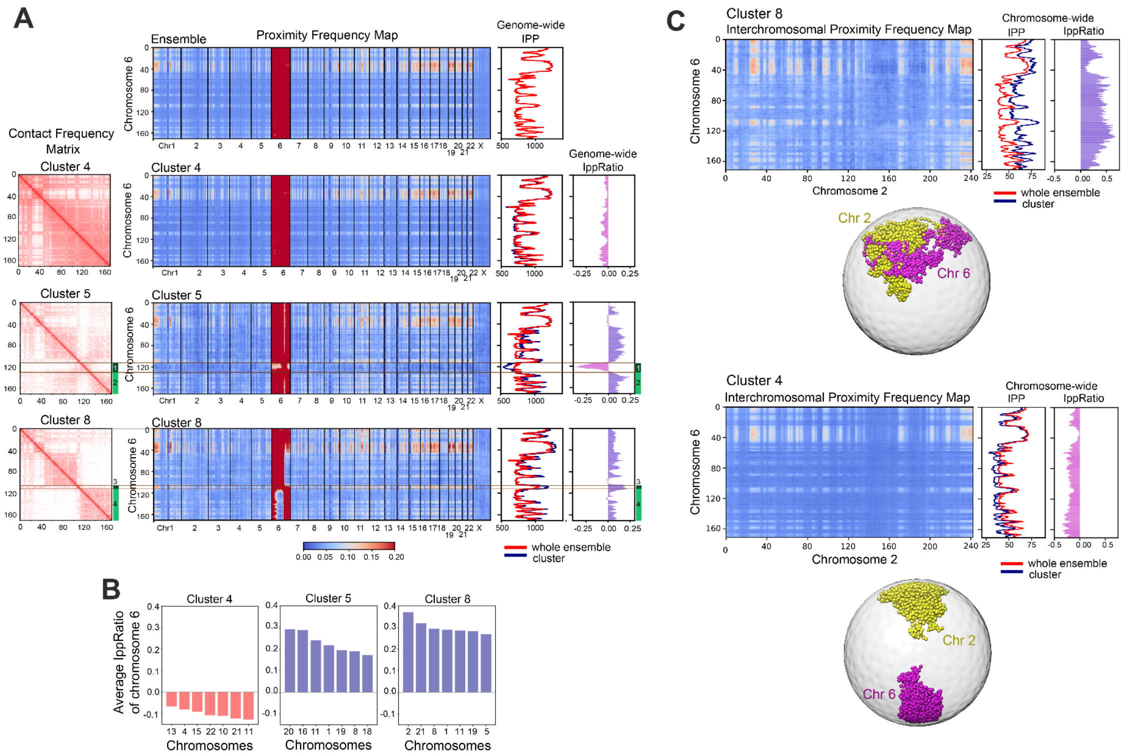


Figure 5: Comparison of inter-chromosomal proximity frequency map and associated features for chromosomes in different clusters. A, (left panels) The average contact frequency matrices of chromosome 6 calculated from all structures in different clusters. Indicated by green boxes are also 4 genomic regions, whose properties are discussed in the text. (Second panels from left) The average proximity frequency matrix between structures of chromosome 6 and structure of all other chromosomes in the genome for different clusters (Methods), (third panels from left) Inter-chromosomal proximity profile (IPP), defined as the total number of inter-chromosomal contact proximities of a genomic region with any other chromosomal region of any chromosome divided by the total number of genome structures in a cluster (Methods). The red line shows the genome-wide IPP profile calculated from the whole ensemble of structures, while the blue line shows the IPP profiles calculated from the structures in each cluster. (Forth panels from left) IppRatio, defined as the log ratio of IPP values in a cluster over the IPP value calculated from the ensemble of all clustered structures. Each row of panels shows these properties calculated from different clusters, namely clusters 4, 5 and 8. **B,** Ranking of the average IppRatios between chromosome 6 and the other chromosomes in different clusters. Only the 7 top ranked chromosomes leading to the highest averaged IPPRatio are shown. **C,** (top left panel) Interchromosomal proximity frequency map between chromosome 6 and chromosome 2 calculated from structures in cluster 8 and (bottom left panel) and structures in cluster 4. (Top middle panel) IPP profile of chromosome 6 considering interactions only to structures of chromosome 2 in cluster 8 and (bottom middle panel) in cluster 4. (Top right panel) IppRatio profiles between chromosomes 6 and 2 in cluster 8 and (Bottom right panel) in cluster 4. Also shown are representative structures of chromosome 6 and 2 in cluster 8 (top) and cluster 4 (bottom).

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

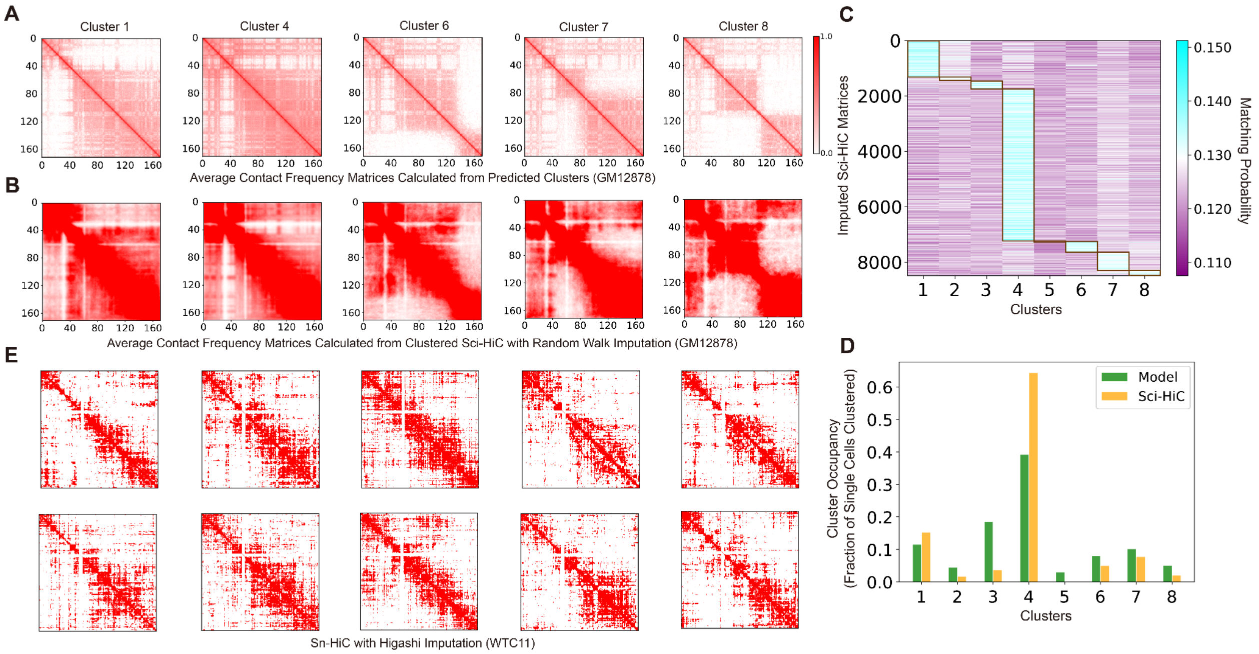


Figure 6: Assessment of predicted clusters by single-cell and single nucleus Hi-C data for chromosome 6. A, The average contact frequency matrices of clusters 1, 4, 6, 7 and 8 calculated for predicted clusters. **B,** Average contact frequency maps calculated from clustered sci-HiC contact maps[10] imputed by convolution and random walk with restart[57]. **C,** Matching probabilities indicating the similarities of all classified sci-HiC contact matrices against all modeled clusters. **D,** Comparison of the cluster occupancy for clusters observed in our models and imputed sci-HiC data. **E,** Selected examples of sn-HiC contact matrices[36,37] imputed by Higashi[27] for different clusters.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

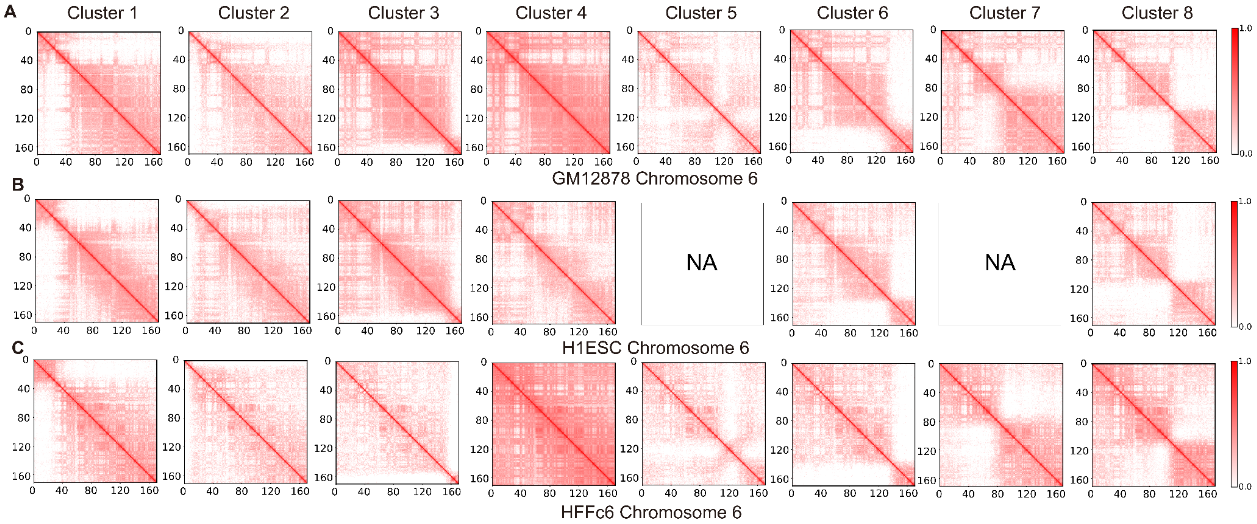


Figure 7: Comparative analysis of predicted clusters of chromosome 6 from genome structures of GM12878, H1-hESC and HFFc6 cells. A, The contact frequency matrices of the 8 predicted clusters of chromosome 6 in GM12878 cells. **B,** The contact frequency matrices of the predicted clusters of chromosome 6 in H1-hESC cells. Clusters 5 and 7 were not observed in H1-hESC cells and are indicated by "N/A". **C,** The contact frequency matrices of the predicted clusters of chromosome 6 in HFFc6 cells.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Clusters	Average Radial Position
2	2.71e-05
3	4.06e-12
4	8.09e-18
5	1.02e-03
6	6.17e-09
7	9.71e-05
8	2.14e-02

Table 1: P-values of the two-sample t-test (Welch's t-test)[55] between cluster 1 and the other clusters of Chr6 on average radial position of region I (24-48 Mb)

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Clusters	Average Radial Position	Average Speckle Distance
1	9.67e-11	4.21e-09
2	4.95e-17	4.43e-09
3	6.95e-22	3.02e-15
4	1.69e-34	7.18e-26
5	5.09e-05	5.37e-03
6	1.04e-09	2.36e-06
7	2.33e-10	1.88e-05

Table 2: P-values of the two-sample t-test (Welch's t-test)[55] between cluster 8 and the other clusters of Chr6 on average radial position and average speckle distance of region II (105-114 Mb)

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Population-based modeling

We used ensemble Hi-C data with the Integrative Genome Modeling (IGM) platform[11] to generate one 10,000 whole genome structure population for HFFc6 (raw Hi-C data from 4DN Data Portal, accession code 4DNES2R6PUEK[60]) and H1-hESC (raw Hi-C data from 4DN Data Portal, accession code 4DNES2M5JIGV[60]) cell lines, and used a previously generated and analyzed 10,000 structure population for GM12878[13].

IGM simulates a population of structures that is compatible with the available ensemble Hi-C data by optimizing the positions of the chromatin regions. Let $\boldsymbol{X}_s = \{\boldsymbol{x}_{1s}, \ldots, \boldsymbol{x}_{Ns}\}$ denote a diploid whole genome structure of N regions, $\boldsymbol{x}_{1s} \in \mathbb{R}^3$ being the Cartesian coordinates of the ith genomic region. A population of structures is defined as a collection of S such structures $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_S\}$. Also, let $\boldsymbol{A} = \left(a_{IJ}\right)_{H\times H}$ denote the Hi-C contact probability matrix, so that $0 \le a_{IJ} \le 1$ indicates the probability that two unphased loci I and J (I, $J \in \{1, \ldots, H\}$) are in contact. In the following, we will denote with (lowercase) i and i' the two copies associated with unphased region I (uppercase). Our genome simulation approach numerically approximates the solution to the optimization problem $\hat{\boldsymbol{X}} = argmax_{\boldsymbol{X}} P(\boldsymbol{A}|\boldsymbol{X})$, where $P(\boldsymbol{A}|\boldsymbol{X})$ is the probability that a population of structures $\boldsymbol{X}$ reproduces the input contact matrix $\boldsymbol{A}$. However, this poses major difficulties: first of all, it is an extremely highly dimensional maximization problem. Second, the input data A does not provide information on which contacts coexist within the same structure in the population and, since it is unphased, does not specify which alleles in the representation (either i or i', j or j') are actually in contact. In order to account for this missing information, we introduce an indicator tensor $\boldsymbol{W} = \left(w_{ijs}\right)_{N\times N\times S}$, $H \le N$, such as $w_{ijs} = 1(0)$ indicates that loci i and j are (not) in contact in diploid structure s-th. It is then essential to jointly optimize both $\boldsymbol{X}$ and $\boldsymbol{W}$ variables, i.e.

$$\hat{\boldsymbol{X}}, \hat{\boldsymbol{W}} = argmax_{\boldsymbol{X}, \boldsymbol{W}} P(\boldsymbol{A}, \boldsymbol{W}|\boldsymbol{X}).$$

We adapted a hard Expectation-Maximization algorithm that uses a series of numeric strategies for efficient and scalable model estimation to tackle such a daunting task. We first initialize the chromatin structures in random territories, and then we start an iterative optimization, where $\boldsymbol{W}$

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and $\boldsymbol{X}$ are alternatively optimized. Each iteration consists of one Assignment step (A-step), where a given subset of contacts from the input Hi-C matrix are optimally allocated across the structures ($\boldsymbol{W}$ is optimized), and a Modeling Step (M-step) where the structure coordinates are optimized using Simulated Annealing Molecular Dynamics and Conjugate Gradient ($\boldsymbol{X}$ is optimized). Additional batches of chromatin contacts are gradually added in each iteration, so as to improve and facilitate overall convergence. Upon convergence, a population $\hat{\boldsymbol{X}}$ of single-cell whole genome structures are available, which are statistically consistent with the input ensemble Hi-C matrix $\boldsymbol{A}$, and also predict a number of orthogonal observables. More details on IGM formulation and implementation can be found in Boninsegna *et al.*[11] and Tjong *et al.*,[43].

Preliminary raw Hi-C datasets preprocessing into a 200k base pair resolution contact probability matrix was accomplished by following the protocol detailed in Yildirim *et al.*[13].

Genome representation

Chromosomes are represented in our models as homopolymer chains of monomers at 200-kb base-pair resolution, so that the full diploid genome is represented with $\mathrm{N}=30{,}332$ monomers for GM12878 and $\mathrm{N}=29{,}838$ for both H1-hESC and HFFc6. Each 200-kb chromatin region is modeled as a sphere of radius around $R_{bead} = 118\,nm$ in all cell lines, so that the ratio of the genome volume to the nuclear volume is 0.4[13,43]. The nuclei for GM12878 and H1-hESC are modeled as spheres of radius $R_{nuc} = 5,000\,nm$[13,43]. The nucleus for HFFc6 is modeled as an ellipsoid of semiaxes $(a, b, c) = (7,840\,nm,\ 6,470\,nm,\ 2,450\,nm)$[11].

Two-step dimension reduction

The basic aim of this study is selecting representative single cell structures from our population and studying their significance. Multiple features of a single cell structure can be extracted and calculated, such as contact matrix and distance matrix, which can be further calculated as a feature vector. Due to the high dimension of the feature vector of our 200kb model, direct classification of these feature vectors is unrealistic. We here introduce the two-step dimension reduction that preserves both efficiency and accuracy during the dimension reduction.

Removal of unrestrained beads

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For each single structure, we remove those beads that are not restrained. All beads remaining in the structure belonging to the “domain” category (not centromeres or telomeres) are considered to construct the distance matrix, while beads belonging to “cen” (centromeres or telomeres) are removed.

Input distance matrix

The distance matrix $\boldsymbol{D}^{(s)} = (d_{ij}^{(s)})$ of chromosomal structure s is calculated by the surface-to-surface distance between bead i and bead j:

$$d_{ij}^{(s)} = \|\mathbf{x}_{is} - \mathbf{x}_{js}\|_2 - 2R_{bead}$$

where $\mathbf{x}_{is}$ and $\mathbf{x}_{js}$ are the 3D coordinates of bead i and bead j in structure s and R_{bead} is the bead radius in our model. $d_{ij}^{(s)}$ is set to be 0 if $i = j$. The matrix is then applied normalization to ensure the maximum entry in the matrix is 1 (dropping the structure superscript s):

$$d_{ij}^{'} = \frac{d_{ij}}{max_{k,l}\, d_{kl}}$$

Due to the size of the layers in the convolutional autoencoder, the input matrix is then resized to multiples of 50 by bilinear interpolation so that the size of the input matrix matches the reconstructed output matrix by the autoencoder.

Convolutional autoencoder

Convolutional autoencoders are frequently used in image classification. In this study, we regard each input distance matrix as an image, which is then regarded as the input of the input layer. The autoencoder consists of an encoder and a decoder, where the input distance matrix is the input of the encoder while the latent matrix is the output. The latent matrix is then used as the input of the decoder to generate the final output. We perform 15 epochs with batch size 200 to train the autoencoder after shuffling the input dataset. The autoencoder is implemented by the python package keras https://github.com/keras-team/keras.

Convolutional layer

A convolutional layer performs convolution operation to an original input over each window and constructs a new output. We use three convolutional layers in the encoder and four convolutional layers in the decoder. In the encoder, the first layer has 16 filters with a kernel with size $(10, 10)$. The next two layers each have 8 filters with a kernel with size $(10, 10)$ and 4

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

filters with a kernel with size $(10,\ 10)$. We use the ReLU activation function to process the output of each convolutional layer:

$$ReLU(x)\ =\ max(0,\ x)$$

In the decoder, the first layer has 4 filters with a kernel with size $(10,\ 10)$. The next two layers each have 8 filters with a kernel with size $(10,\ 10)$ and 16 filters with a kernel with size $(10,\ 10)$. We use the ReLU activation function to process the output of these convolutional layers. To generate the output, the last convolutional layer uses 1 filter with a kernel with size $(10,\ 10)$, we use the sigmoid activation function to ensure that the values in the final output matrix are located between 0 and 1:

$$Sigmoid(x)\ = \frac{1}{1+e^{-x}}$$

For each convolutional layer, we use the stride size $(1,\ 1)$ and the same padding size to ensure the output has the same height and width as the input.

Max pooling layer

A max pooling layer is used to downsample an original input by calculating the maximum value in each window and generate a new value. We use three max pooling layers in the encoder. The first two layers have a pooling window with size $(5,\ 5)$. The last layer has a pooling window with size $(2,\ 2)$. We use the same padding size for each max pooling layer to generate the output. The stride size is the same as the window size for each layer.

Upsampling layer

An upsampling layer is used to up sample an original input by filling each window with the corresponding value. We use three upsampling layers in the decoder. The first layer has a sampling window with size $(2,\ 2)$ and the next two layers have a sampling window with size $(5,\ 5)$.

Latent vector

The latent vector is generated by directly flattening the latent matrix. We then use standard normalization to normalize the whole set of latent vectors to ensure that each dimension $\mathbf{x}_l^{'} = (x_{l1}^{'},\ x_{l2}^{'},\ ..,\ x_{lN}^{'})$ of the set of new vectors has mean 0 and standard deviation 1:

$$x_{li}^{'} = \frac{x_{li}-\overline{x}_l}{\sigma_l}$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where $\overline{x_l}$ and σ_l is the mean and the standard deviation of each dimension $\mathbf{x}_l = (x_{l1}, x_{l2}, ..., x_{lN})$ of the set of original vectors. N is the size of the training set.

Mean squared error

We use the mean squared error (MSE) to calculate the loss between the input matrix $\boldsymbol{M}^{input}$ and the output matrix $\boldsymbol{M}^{output}$, which is the mean Euclidean distance between the input matrix and the output matrix of the training dataset:

$$MSE = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{M}_i^{input} - \boldsymbol{M}_i^{output}\|_2^2$$

where N is the size of the training set.

Optimizer

We use an optimizer that applies the Adadelta algorithm[61] to train the autoencoder. In comparison with other gradient descent methods, this method does not require setting of the learning rate parameter and is relatively more robust.

T-distributed stochastic neighbor embedding

T-distributed stochastic neighbor embedding (t-SNE) is a robust and nonlinear dimension reduction method[47]. By using proper probability distributions $\boldsymbol{P} = (p_{ij})$ and $\boldsymbol{Q} = (q_{ij})$ to measure similarities between data points in both the original space and the lower dimensional space. The method facilitates the embedding by minimizing the Kullback–Leibler divergence[62] between the two distributions by:

$$KL(\boldsymbol{P}||\boldsymbol{Q}) = \sum_{i \neq j} p_{ij} log\frac{p_{ij}}{q_{ij}}$$

We set the dimension of the embedded space to be 2, the perplexity to be 200 and the learning rate to be 1,000.

Principal component analysis

Principal component analysis (PCA) is a frequently used dimension reduction method which computes the principal components of the data. The method uses singular value decomposition (SVD) of the covariance matrix to construct principal components which are then used to find embedded data points in the lower dimensional space. The dimension of the embedded space is set to be 2.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Multidimensional scaling

Classical multidimensional scaling (MDS), which is also known as Principal Coordinates Analysis (PCoA), is another nonlinear dimension reduction[48]. The classical MDS transforms pairwise distances between data points into dissimilarities and minimizes a cost function. We use Euclidean distances as dissimilarity measurement. The dimension of the embedded space is set to be 2.

Locally linear embedding

Unlike PCA which projects data points in a linear way, locally linear embedding (LLE) is a nonlinear dimension reduction technique[49]. The method can be viewed as a collection of local PCA which preserves distances within each local neighborhood graph. The dimension of the embedded space is set to be 2.

Isomap

Isomap, which is also a nonlinear dimension reduction method, is an extended version of MDS[50]. Specifically, the method uses geodesic distances of each local neighborhood graph as similarity measurement before performing MDS. The dimension of the embedded space is set to be 2.

Spectral embedding

Spectral embedding (SE) is another nonlinear dimension reduction method[51]. The method uses eigenvectors of the Laplactian matrix to construct embedded data points in the lower dimensional space. The dimension of the embedded space is set to be 2. All embedding listed above including t-SNE, PCA, MDS, LLE, Isomap and SE are performed by the python package sklearn[63].

UMAP

Uniform Manifold Approximation and Projection (UMAP) uses knowledge of algebraic topology and simplicial complexes to perform dimension reduction[52]. The method is an increasingly frequently used nonlinear dimension reduction method which is often used to compete with t-SNE. UMAP is performed by the python package umap-learn[52]. We set the dimension of the embedded space to be 2 and the learning rate to be 1.0.

Peak detection

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In the 2D embedded conformational space, every data point represents a single structure. The structures that are closed with each other in 2D distance are more likely to have similar conformations. The next step is to sample part of the data points which are representative from the 2D distribution.

Outlier removal

To remove outlier data points, we first calculate a pairwise distance matrix of all data points. We then generate the total distance between each data point and all the other data points by calculating the row sum of each row. The data points that have extreme total distances are removed by the 3-sigma rule. We only select data points whose row sums are within 3-sigma range $(\overline{s_l} - 3\sigma_l, \overline{s_l} + 3\sigma_l)$, where $\overline{s_l}$ and σ_l is the mean and the standard deviation of the row sum vector s_l.

Kernel density estimation

Then we use bivariate kernel density estimation to calculate the probability density function of the distribution. Given a 2D independent and identically distributed sample $\boldsymbol{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_N)$, each point can be we are able to find a density function p so that this set of data points is sampled directly from a distribution with joint probability density function p:

$$p(\mathbf{x}) = \frac{|\mathbf{H}|^{-\frac{1}{2}}}{N} \sum_{i=1}^{N} K(\mathbf{H}^{-\frac{1}{2}}(\mathbf{x} - \mathbf{x}_i))$$

where $\mathbf{x}_i = (x_{i1}, x_{i2})^T$. We choose K to be the gaussian kernel. The bandwidth $\mathbf{H}$ is estimated by Scott's Rule[64]. The resulting 2D density measures how data points are distributed in the conformation space. Each local maximum of the 2D density is defined as a peak, which is a representative conformation.

Grid approximation

We use a $(1000, 1000)$ grid $\boldsymbol{G}$ to approximate probability density function $p(\mathbf{x})$ and generate a 2D density matrix $\boldsymbol{P}$. The grid is constructed with the minimum value and the maximum value of each of the two dimensions x_{i1} and x_{i2}:

$$\boldsymbol{G} = \left\{\left(min_i\, x_{i1}, max_i\, x_{i1}\right), \left(min_i\, x_{i2}, max_i\, x_{i2}\right)\right\}_{1000\times1000}$$

We then calculate the density value by the probability density function $p(x)$ at each grid point and construct the density matrix $\boldsymbol{P} = (p_{ij})$:

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$p_{ij} = p(\boldsymbol{G}(i,\, j))$$

Maximum filter

A local maximum is an entry that is larger than all its 8 neighbors in the 2D density matrix P. To avoid selecting multiple local maxima in a small area, however, we compare each entry with a larger range of its neighborhood. A maximum filter with size $(5,\, 5)$ is applied to matrix P to generate another matrix $\boldsymbol{P}^{'} = (p_{ij}^{'})$. We then use exclusive disjunction (XOR) to generate matrix $\boldsymbol{Q} = (q_{ij})$ by comparing $\boldsymbol{P}$ and $\boldsymbol{P}^{'}$:

$$q_{ij} = I_{p_{ij}=p_{ij}^{'}} \oplus I_{p_{ij}=0}$$

where I_A is the indicator function which equals 1 when A is true. The entries in matrix Q with value 1 are detected as local maxima or peaks.

Cluster analysis

Boundary estimation by watershed

Considering each local maximum, we use a watershed-like approach to simulate the cluster boundary around it. We first create a density gradient with 100 levels ranging from 0 to the largest density in the 2D density matrix P. A set of contour lines which are polygons formed by grid points gradually change (shrink) over the density gradient. The change terminates when there is a contour in the set containing only the target local maximum, which results in our target contour line that contains only the corresponding maximum. All points surrounded by the contour line are then considered as the cluster members of the corresponding maximum. A cluster is not considered if it contains fewer than 100 members.

Contact frequency matrix construction

By selecting a certain number of neighbors around each peak, we are able to construct a contact frequency matrix for each peak. We estimate a path (polygon) surrounding each peak based on density. We then select all points inside the polygon as a cluster that corresponds to the peak. To calculate the contact frequency matrix $\boldsymbol{CM}^{(a)}$ for structure a, we say beads i and j are in contact is structure a (i.e., $cm_{ij}^{(a)} = 1$) if and only if:

$$\|\mathbf{x}_{ia} - \mathbf{x}_{ja}\|_2 \leq 3R_{bead}$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where $\mathbf{x}_{ia}$ and $\mathbf{x}_{ja}$ are the 3D coordinates of bead i or bead j. R_{bead} is the bead radius in our model. The contact frequency matrix for cluster A, $\boldsymbol{CM}^{(A)}$, is calculated by the sum of all contact matrices in the cluster:

$$\boldsymbol{CM}^{(A)} = \sum_{a\epsilon A} \boldsymbol{CM}^{(a)}$$

The contact frequency matrix for all structures that are classified to a cluster is calculated as:

$$\boldsymbol{CM}^{(Ens)} = \sum_{a\epsilon S} \boldsymbol{CM}^{(a)}$$

where S is the set of clustered structures. To enhance off-diagonal contacts, we visualize all contact frequency matrices from the models by applying transformation $log_2(cm_{ij} + 1)$. All color bars shown together with contact frequency matrices in the figures show a ratio with regards to the maximum value.

Average distance matrix construction

Similarly to the construction of a contact frequency matrix, we can also construct an average distance matrix for each cluster. Each entry $dm_{ij}^{(a)}$ of the distance matrix $\boldsymbol{DM}^{(a)}$ for structure a is calculated by the Euclidean distance between bead i and bead j:

$$dm_{ij}^{(a)} = \|\mathbf{x}_{ia} - \mathbf{x}_{ja}\|_2$$

where $\mathbf{x}_{ia}$ and $\mathbf{x}_{ja}$ are the 3D coordinates of beads i and j. $dm_{ij}^{(a)}$ is set to be 0 if the entry is at the diagonal. After min-max normalization of each distance matrix, the average matrix for cluster A (which we will denote as $\boldsymbol{DM}^{(A)}$) is calculated by the average of all matrices in the cluster:

$$\boldsymbol{DM}^{(A)} = \frac{1}{S_A} \sum_{a\epsilon A} \boldsymbol{DM}^{(a)}$$

where S_A is the number of structures in cluster A. All color bars shown together with average distance matrices in the figures show a ratio with regards to the maximum value.

Dissimilarity measurement

Euclidean distance dissimilarity

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We first construct two flattened distance matrices $\boldsymbol{R}^{(a)}$ and $\boldsymbol{R}^{(b)}$ for structure a and structure b. Each matrix contains Euclidean distances between all possible pairs of beads in each structure. The Euclidean distance between these two structures $s_e^{(ab)}$ is further calculated by:

$$s_e^{(ab)} = \|\boldsymbol{R}^{(a)} - \boldsymbol{R}^{(b)}\|_2$$

Then the final Euclidean distance dissimilarity between cluster A and cluster B is the average value of all possible pairs between these two clusters:

$$s_e^{(AB)} = \frac{1}{M} \sum_{a \epsilon A,\ b \epsilon B} s_e^{(ab)}$$

where M is the total number of pairs between the two clusters. To compare inter-cluster dissimilarity and intra-cluster dissimilarity, we normalize $s_e^{(AB)}$ by the intra-cluster dissimilarity of cluster A $s_e^{(AA)}$:

$$rs_e^{(AB)} = log_2 \frac{s_e^{(AB)}}{s_e^{(AA)}}$$

Gaussian dissimilarity

The calculation of Gaussian dissimilarity is adapted from Eastwood and Wolynes[54] and Cheng et al[32], which is an alternative way to compare pairwise distances between two structures. After generating $d_{ij}^{(a)}$ and $d_{ij}^{(b)}$ which are the Euclidean distances between bead i and bead j for both structure a and structure b, the Gaussian dissimilarity $s_g^{(ab)}$ is calculated by:

$$s_g^{(ab)} = 1 - \frac{1}{N} \sum_{i<j} exp(- \frac{(d_{ij}^{(a)} - d_{ij}^{(b)})^2}{2\sigma^2})$$

where the scaling factor $\sigma = 8R_{bead}$ and R_{bead} is the bead radius in our model. N is the total number of pairs of beads in the structure. Similarly, the final Gaussian dissimilarity between cluster A and cluster B is the average value of all possible pairs between these two clusters:

$$s_g^{(AB)} = \frac{1}{M} \sum_{a \epsilon A,\ b \epsilon B} s_g^{(ab)}$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where M is the total number of pairs between the two clusters. To compare inter-cluster dissimilarity with intra-cluster dissimilarity, we normalize $s_g^{(AB)}$ by the intra-cluster dissimilarity of cluster A $s_g^{(AA)}$:

$$rs_g^{(AB)} = log_2 \frac{s_g^{(AB)}}{s_g^{(AA)}}$$

Due to computational complexity, we randomly select 200 structures from each cluster to compute the similarities above.

Wasserstein distance dissimilarity

We calculate both intra-cluster dissimilarity and inter-cluster dissimilarity by distance measurement to compare low intra-cluster dissimilarity with high inter-cluster dissimilarity. The Wasserstein distance $W(u,\ v)$ measures the dissimilarity between two probability distributions u and v by:

$$W(u,\ v) = \int_{-\infty}^{+\infty} |U - V| ds$$

where U and V are the cumulative probability distributions of u and v[53]. To measure dissimilarity between two clusters of structures, for each pair of bead i and bead j, we obtain the 1D probability distributions for the distances between pair i and j in cluster A $d_{ij}^{(A)}$ and the distances between pair i and j in cluster B $d_{ij}^{(B)}$, which are then used to calculate the Wasserstein distance of these two distributions. The final dissimilarity $s_w^{(AB)}$ is obtained by averaging the Wasserstein distances of all possible pairs:

$$s_w^{(AB)} = \frac{1}{N} \sum_{i<j} W(d_{ij}^{(A)},\ d_{ij}^{(B)})$$

where N is the total number of pairs of beads in the structure. For intra-cluster dissimilarity, we randomly sample a subcluster with size $\frac{M}{2}$ and indices i which is then used to calculate the final dissimilarity with its reverse subcluster with indices $(M - i - 1)$, where M is the number of structures in cluster A. For inter-cluster dissimilarity, we directly apply the method above. To compare inter-cluster dissimilarity with intra-cluster dissimilarity, we normalize $s_w^{(AB)}$ by the intra-cluster dissimilarity of cluster A $s_w^{(AA)}$:

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$rs_w^{(AB)} = log_2 \frac{s_w^{(AB)}}{s_w^{(AA)}}$$

Proximity frequency map

The calculation of a proximity frequency map is similar to the calculation of a contact frequency matrix. We select a larger range to visualize inter-chromosomal contact patterns. To calculate the proximity frequency map $\boldsymbol{PM}^{(a)}$ for structure a, we define the i-th bead and the j-th bead in structure a forms a contact (i.e., $pm_{ij}^{(a)} = 1$) if and only if:

$$\|\mathbf{x}_{ia} - \mathbf{x}_{ja}\|_2 \leq R_{soft}$$

where $\mathbf{x}_{ia}$ or $\mathbf{x}_{ja}$ are the 3D coordinates of bead i or bead j. We set $R_{soft} = 2,000\, nm$. The proximity frequency map for cluster A $\boldsymbol{PM}^{(A)}$ is calculated as:

$$\boldsymbol{PM}^{(A)} = \frac{1}{S_A} \sum_{a \epsilon A} \boldsymbol{PM}^{(a)}$$

where S_A is the number of structures in cluster A. The inter-chromosomal parts of each map are shown as the average of both homologous copies, while the intra-chromosomal part is calculated from the target chromosome copy only.

Structural features prediction

Radial position (RAD)

The radial position of a chromatin region i in structure s in a spherical nucleus (as GM12878) is calculated as:

$$r_i^{(s)} = \frac{\|\mathbf{x}_{is}\|_2}{R_{nuc}}$$

where $\mathbf{x}_{is}$ is the the 3D coordinates of bead i in structure s, and R_{nuc} is the nucleus radius which is 5 µm. $r_i^{(s)} = 0$ means the region i is at the nuclear center while $r_i^{(s)} = 1$ means it is located at the nuclear surface. The average radial position (RAD) of cluster A is the average of radial positions of all structures in this cluster:

$$r_i^{(A)} = \frac{1}{S_A} \sum_{a \in A} r_i^{(a)}$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where S_A is the number of structures in cluster A. To compare against the ensemble profile, we use a log ratio comparison to show the difference. The log ratio of cluster A radial position against the ensemble one (RadRatio) is calculated as:

$$rr_i^{(A)} = log_2 \frac{r_i^{(A)}}{r_i^{(Ens)}}$$

where the ensemble radial position is calculated in the same way, but for all structures that are classified to any cluster. Similarly, we can calculate all following structural features with the (Ens) superscript.

Radius of gyration (RG) (i.e., local chromatin fiber decompaction)**.**

The local compaction of the chromatin fiber at the location of a given locus is estimated by the radius of gyration for a 1 Mb region centered at the locus. To estimate the values along an entire chromosome we use a sliding window approach over all chromatin regions in a chromosome. The radius of gyration for a 1 Mb region centered at locus i in structure a, is calculated as:

$$rg_i^{(a)} = \sum_{j=1}^{5} d_j^2$$

where d_j is the distance between the chromatin region j to the center of mass of the 1-Mb region. The average radius of gyration (RG) of cluster A is the average of radial positions of all structures in this cluster:

$$rg_i^{(A)} = \frac{1}{S_A} \sum_{a \in A} rg_i^{(a)}$$

where S_A is the number of structures in cluster A. Similarly, the log ratio of cluster A radius of gyration against the ensemble one (RgRatio) is calculated as:

$$rrg_i^{(A)} = log_2 \frac{rg_i^{(A)}}{rg_i^{(Ens)}}$$

For the overall compactness of the conformation, we use all the beads to calculate the radius of gyration for structure a:

$$rg^{(a)} = \sum_{j=1}^{N} d_j^2$$

where N is the total number of beads in the structure, d_j is the distance between the chromatin region j to the center of mass of the whole structure.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Structural variability (δRAD)

The structural variability (δRAD) of region i in cluster A is calculated as:

$$sv_i^{(A)} = log_2 \frac{\sigma_i^{(A)}}{\overline{\sigma^{(A)}}}$$

where $\sigma_i^{(A)}$ is the standard deviation of the population of radial positions of region i in cluster A and $\overline{\sigma^{(A)}}$ is the mean standard deviation calculated from all regions within the same chromosome of the target region. Positive values ($sv_i^{(A)} > 0$) result from high cell-to-cell variability of radial position, whereas negative values ($sv_i^{(A)} < 0$) indicate low variability. The log ratio of cluster A structural variability against the ensemble one (δRadRatio) is calculated as:

$$rsv_i^{(A)} = log_2 \frac{sv_i^{(A)}}{sv_i^{(Ens)}}$$

Building chromatin interaction networks

A chromatin interaction network (CIN) is calculated for each model and for chromatin in each subcompartment separately as follows: Each vertex represents a 200-kb chromatin region. An edge between two vertices i, j is drawn if the corresponding chromatin regions are in physical contact in the model if the spatial distance $d_{ij} \leq 4R_{bead}$, where R_{bead} is the bead radius in our model.

Identifying spatial partitions by Markov clustering

Spatial partitions of subcompartments as well as regions in A compartment with low and high structural variability are identified by applying Markov Clustering Algorithm (MCL)[65], a graph clustering algorithm, which identifies highly connected subgraphs within a network. MCL clustering is performed for each subcompartment subgraph in each structure by using the MCL tool in the MCL-edge software[65]. The 25% smallest subgraphs (with less than 7 nodes) are discarded from further analysis to focus on highly connected subgraphs. Speckle locations are identified as the geometric center of A1 subgraphs identified by Markov clustering of A1 subgraphs. In each structure, A1 subgraphs are considered with size larger than 3 nodes.

Speckle distance (SpD)

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The speckle distance (SpD) for region i is calculated by measuring the distance between the surface of each chromatin region i to the nearest speckle:

$$sd_i^{(A)} = \frac{1}{S_A} \sum_{a \epsilon A} d_{il}^{(a)}$$

where S_A is the number of structures in cluster A, $d_{il}^{(a)}$ is the distance between the region i and the predicted nearest nuclear body location l. The log ratio of cluster A speckle distance against the ensemble one (SpdRatio) is calculated as:

$$rsd_i^{(A)} = log_2 \frac{sd_i^{(A)}}{sd_i^{(Ens)}}$$

Speckle TSA-seq (SON TSA-seq)

Speckle TSA-seq can be viewed as an average over distances to all speckles. To predict TSA-seq signals for speckle from our models, we use the following equation:

$$sg_i^{(A)} = \frac{1}{S_A} \sum_{a \epsilon A} \sum_{l=1}^{L} e^{-kd_{il}^{(a)}}$$

where S_A is the number of structures in cluster A, L is the number of predicted speckle locations in structure a, $d_{il}^{(a)}$ is the distance between the region i and the predicted nuclear body location l, and k is the estimated decay constant in the TSA-seq experiment[45] which is set to 4 in our calculations. The normalized TSA-seq signal for region i then becomes:

$$ts_i^{(A)} = log_2 \frac{sg_i^{(A)}}{\overline{sg^{(A)}}}$$

where $\overline{sg^{(A)}}$ is the mean signal calculated from all regions in the genome. The predicted speckles are used for distance calculations. The log ratio of cluster A speckle TSA-seq against the ensemble one (SON TSA-seq Ratio) is calculated as:

$$rts_i^{(A)} = log_2 \frac{ts_i^{(A)}}{ts_i^{(Ens)}}$$

Speckle association frequency (SAF)

For a given 200-kb region, the association frequency to the speckle (SAF) is calculated as:

$$saf_i^{(A)} = \frac{n_{d_i < d_t}}{S_A}$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where S_A is the number of structures in cluster A, $n_{d_i<d_t}$ is the number of structures, in which region i have a distance to the speckle smaller than the association threshold d_t. We set d_t to be 1,000 nm for the model and 500 nm for the DNA-MERFISH dataset[3]. For SAF calculation, we use the predicted speckle to calculate distances (see Identifying spatial partitions by Markov clustering), where we calculate distances from the surface of the region to the center-of-mass of the partition. The log ratio of cluster A SAF against the ensemble one (SafRatio) is calculated as:

$$rsaf_i^{(A)} = log_2 \frac{saf_i^{(A)}}{saf_i^{(Ens)}}$$

Inter-chromosomal proximity profile (IPP)

The calculation of inter-chromosomal proximity profile (IPP) is based on the proximity frequency map. For a given 200-kb region, the process is similar to the calculation of speckle association frequency, but we replace the distance to the smallest speckle by the contact with any inter-chromosomal regions which can be chromosome-wide or genome-wide:

$$ipp_i^{(A)} = \frac{n_{d_i \le R_{soft}}}{S_A}$$

where S_A is the number of structures in cluster A, $n_{d_i \le R_{soft}}$ is the total number of contacts, in which region i is within contact range $R_{soft} = 2,000\, nm$ with any target inter-chromosomal regions from the same genome structure. Every IPP is shown as the average of both homologous copies. The log ratio of cluster A IPP against the ensemble one (IppRatio) is calculated as:

$$ripp_i^{(A)} = log_2 \frac{ipp_i^{(A)}}{ipp_i^{(Ens)}}$$

When calculating the average IppRatio of a chromosome, we calculate the mean of chromosome-wide IppRatios of the chromosome.

Histone modification signals and reference genes

We collected histone modification signals including H3K27ac, H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H4K20me1 for GM12878 and H3K9me2 for GM23338 from the ENCODE[66]. The reference genes file for hg38 was downloaded from the UCSC Genome Browser[67]. All

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

related signals and genes together with other structural features are shown by the Integrative Genomics Viewer (IGV)[68].

Cluster assessment with experimental single cell data

Single cell Hi-C assessment

Sci-HiC dataset

We collected multiple sci-HiC datasets of GM12878 from the 4DN data portal (4DNESUE2NSGS)[10]. Each dataset consists of single cell sequencing data of thousands of cells and we collected more than 11,000 single cells in total. A systematic way of massively demultiplexing single cell Hi-C is discussed in Ramani et al[10] which applies combinatorial cellular indexing to chromosome conformation capture. We use the provided pipeline to process all collected sci-HiC datasets. Due to the large number of missing contacts, it is necessary for us to preprocess the datasets to reconstruct missing information. We adapt the preprocessing method from Zhou et al[57]. Given a raw single cell contact matrix $\boldsymbol{M}^{raw} = (m_{ij}^{raw})_{n\times n}$, we construct a new matrix $\boldsymbol{M}^{conv} = (m_{ij}^{conv})_{n\times n}$ by applying convolution with filter $\boldsymbol{F} = (f_{ij})_{(2w+1)\times(2w+1)}$:

$$m_{ij}^{conv} = \sum_{k,\,l} f_{kl} m_{kl}^{raw}$$

For a 200kb matrix, we set $w = 5$. In this step, we integrate the interaction information from the genomic neighbors to impute the interaction at each position. Random walk with restart is then performed to estimate contact probability between every two beads. In order to perform a random walk, a transition matrix $\boldsymbol{M}^{trans} = (m_{ij}^{trans})_{n\times n}$ is calculated based on the contact matrix after convolution. Every entry in the original matrix is normalized by its corresponding row sum:

$$m_{ij}^{trans} = \frac{m_{ij}^{conv}}{\sum_{j'} m_{ij'}^{conv}}$$

We initialize the random walk by an identity matrix $\boldsymbol{R}_0$ so that the contact probability between every two beads is set to be 0. By applying the following recurrence formula, we are able to obtain a resulting matrix after the values converge:

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\boldsymbol{R}_0 = \mathbf{I}$$

$$\boldsymbol{R}_t = (1 - p)\boldsymbol{R}_{t-1}\boldsymbol{M}^{trans} + p\mathbf{I}$$

where $\mathbf{I}$ is the identity matrix and p is the restart probability with 0.5. We define there is a convergence when $\|\boldsymbol{R}_t - \boldsymbol{R}_{t-1}\|_2 \leq 10^{-6}$. Each element in the resulting matrix $\boldsymbol{R}_t$ after convergence indicates the probability of the random walk to reach the j-th node when starting from the i-th node. All contacts with probability larger than the 75th percentile of all probabilities in each row are chosen to convert $\boldsymbol{R}_t$ into a binary matrix $\boldsymbol{M}^{rw}$.

Sci-HiC assessment

For comparison of clusters, a direct way is to compare their contact frequency matrices. We define the difference matrix of cluster A to be:

$$d_{ij}^{(A)} = log_2 \frac{S_A m_{ij}^{(Pop)}}{S_{Pop} m_{ij}^{(A)}}$$

where $\boldsymbol{D}^{(A)} = (d_{ij}^{(A)})$ is the resulting difference matrix. $\boldsymbol{M}^{(Pop)} = (m_{ij}^{(Pop)})$ is the contact frequency matrix for the whole population calculated by contact range 2, while $\boldsymbol{M}^{(A)} = (m_{ij}^{(A)})$ is the contact frequency matrix for the cluster. S_A is the cluster size while S_{Pop} is the population size. Due to the sparsity of single cell Hi-C, we preprocess each raw contact matrix by the preprocessing method above to construct a processed contact matrix. The next step is to assign each contact matrix to the clusters defined by our model. For cluster A, a superiority mask $\boldsymbol{M}_{sup}^{(A)} = (ms_{ij}^{(A)})$ and an inferiority mask $\boldsymbol{M}_{inf}^{(A)} = (mi_{ij}^{(A)})$ are calculated by its difference matrix $\boldsymbol{D}^{(A)}$:

$$ms_{ij}^{(A)} = I_{d_{ij}^{(A)} \geq 5}$$

$$mi_{ij}^{(A)} = I_{d_{ij}^{(A)} \leq -1}$$

where I_A is the indicator function which equals 1 when A is true. For each preprocessed matrix $\boldsymbol{M}^{rw}$ from the sci-HiC population, we define the assessment score as:

$$s^{(A)} = exp\left(\frac{\langle \boldsymbol{M}^{rw}, \boldsymbol{M}_{inf}^{(A)} \rangle_F}{\langle \boldsymbol{E}, \boldsymbol{M}_{inf}^{(A)} \rangle_F} - \frac{\langle \boldsymbol{M}^{rw}, \boldsymbol{M}_{sup}^{(A)} \rangle_F}{\langle \boldsymbol{E}, \boldsymbol{M}_{sup}^{(A)} \rangle_F}\right)$$

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle_F$ is the Frobenius inner product between matrix $\boldsymbol{X}$ and matrix $\boldsymbol{Y}$. $\boldsymbol{E}$ is a matrix of ones. For each contact matrix, we choose the pair of masks that has the largest assessment score with the matrix and assign the matrix to the corresponding cluster. To filter matrices that are different from all clusters, we only classify matrices to a cluster A_1 when $s^{(A_1)} - s^{(A_2)} \geq 0.01$, where A_1 is the cluster with the largest matching score and A_2 is the second largest one. The final matching probability is defined as:

$$p^{(A_1)} = \frac{s^{(A_1)}}{\sum_{k=1}^{K} s^{(A_k)}}$$

where K is the total number of clusters. Similarly, a contact frequency matrix can be generated using all inferred sci-HiC matrices classified to each cluster. To preserve symmetry, we symmetrize each contact frequency matrix by selecting the minimum number of contacts between pairs (i, j) and (j, i).

Sci-HiC control dataset

A control dataset is generated to ensure our assessment procedure is not classifying artifacts and false signals. For each sci-HiC contact matrix, we randomly rearrange all the entries while maintaining its diagonality and the total number of contacts to construct a sudo single cell contact matrix. We apply this process for every sci-HiC matrix and construct a control dataset in the end. The same assessment procedure is then applied to the control dataset.

Sn-HiC dataset

In total, 188 WTC-11 sn-HiC matrices were obtained from the 4DN data portal (4DNESF829JOW and 4DNESJQ4RXY5)[36,37]. Higashi[27] is then used to impute missing contacts from the raw contact matrices. All contacts with probability larger than $0.003 p_{max}$ are chosen to convert each imputed matrix into a binary matrix, where p_{max} is the maximum probability of the imputed matrix.

Imaging assessment

DNA-MERFISH dataset

We process the DNA-MERFISH datasets from Su et al[3] which includes high-resolution coordinates of chromosome 2 from 3,029 copies and low-resolution coordinates of chromosome 6 from 7,336 copies. For chromosome 6, we also process distances of imaged genomic regions to the nearest detected speckle. All datasets are preprocessed by linear interpolation to remove

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

missing values if applicable. We remove copies without valid values in coordinates and in speckle distances.

DNA-MERFISH assessment

The preprocessed DNA-MERFISH coordinates can be then used for assessment. Each single structure is used to calculate a distance matrix $\boldsymbol{DM}$ in the same way stated above. To compare $\boldsymbol{DM}$ with the average distance matrix $\boldsymbol{DM}^{(A)}$ for cluster A, we first downsample $\boldsymbol{DM}^{(A)}$ by selecting the beads that are mapped by the DNA-MERFISH coordinates. Then we flatten both matrices by extracting the upper triangular parts and normalizing them by min-max normalization to generate two distance vectors $\boldsymbol{R}$ and $\boldsymbol{R}^{(A)}$. We define the assessment score as:

$$s^{(A)} = exp(r(\boldsymbol{R},\ \boldsymbol{R}^{(A)}))$$

where $r(\mathbf{x},\ \mathbf{y})$ measures the Pearson's correlation coefficient between vector $\mathbf{x}$ and vector $\mathbf{y}$. For each distance matrix, we choose the average distance matrix that has the largest assessment score with the matrix and assign the matrix to the corresponding cluster. To filter matrices that are different from all clusters, we only classify matrices to a cluster A_1 when $s^{(A_1)} - s^{(A_2)} \geq 0.05$, where A_1 is the cluster with the largest matching score and A_2 is the cluster the second largest one. The final matching probability is defined as:

$$p^{(A_1)} = \frac{s^{(A_1)}}{\sum_{k=1}^{K} s^{(A_k)}}$$

where K is the total number of clusters. Similarly, an average distance matrix can be generated using all DNA-MERFISH distance matrices classified to each cluster.

Data visualization

All chromosome structures are visualized by Chimera[69].

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Nguyen, H. Q. *et al.* 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. *Nat Methods* **17**, 822–832 (2020).
2. Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, eaau1783 (2018).
3. Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **182**, 1641-1659.e26 (2020).
4. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).
5. Payne, A. C. *et al.* In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science* **371**, eaay3446 (2021).
6. Takei, Y. *et al.* Integrated spatial genomics reveals global architecture of single nuclei. *Nature* **590**, 344–350 (2021).
7. Nagano, T. *et al.* Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat Protoc* **10**, 1986–2003 (2015).
8. Tan, L., Xing, D., Chang, C.-H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924–928 (2018).
9. Li, G. *et al.* Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat Methods* **16**, 991–993 (2019).
10. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat Methods* **14**, 263–266 (2017).
11. Boninsegna, L. *et al.* Integrative genome modeling platform reveals essentiality of rare contact events in 3D genome organizations. *Nat Methods* **19**, 938–949 (2022).
12. Boninsegna, L., Yildirim, A., Zhan, Y. & Alber, F. Integrative approaches in genome structure analysis. *Structure* **30**, 24–36 (2022).
13. Yildirim, A. *et al. Population-based structure modeling reveals key roles of nuclear microenviroment in gene functions*. http://biorxiv.org/lookup/doi/10.1101/2021.07.11.451976 (2021) doi:10.1101/2021.07.11.451976.
14. Yildirim, A., Boninsegna, L., Zhan, Y. & Alber, F. Uncovering the Principles of Genome Folding by 3D Chromatin Modeling. *Cold Spring Harb Perspect Biol* a039693 (2021) doi:10.1101/cshperspect.a039693.
15. Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G. & Onuchic, J. N. Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 12168–12173 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

16. Qi, Y. & Zhang, B. Polymer Modeling of Whole-Nucleus Diploid Genome Organization. *Biophysical Journal* **118**, 550a–550a (2020).
17. Mendieta-Esteban, J., Di Stefano, M., Castillo, D., Farabella, I. & Marti-Renom, M. A. 3D reconstruction of genomic regions from sparse interaction data. *NAR Genom Bioinform* **3**, lqab017 (2021).
18. Krietenstein, N. *et al.* Ultrastructural Details of Mammalian Chromosome Architecture. *Mol Cell* **78**, 554-565.e7 (2020).
19. Conte, M. *et al.* Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation. *Nat Commun* **11**, 3289 (2020).
20. Paulsen, J., Gramstad, O. & Collas, P. Manifold Based Optimization for Single-Cell 3D Genome Reconstruction. *PLoS Comput Biol* **11**, e1004396 (2015).
21. Carstens, S., Nilges, M. & Habeck, M. Inferential Structure Determination of Chromosomes from Single-Cell Hi-C Data. *PLoS Comput Biol* **12**, e1005292 (2016).
22. Giorgetti, L. *et al.* Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**, 950–963 (2014).
23. Misteli, T. The Self-Organizing Genome: Principles of Genome Architecture and Function. *Cell* **183**, 28–45 (2020).
24. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat Rev Genet* **21**, 207–226 (2020).
25. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
26. Akgol Oksuz, B. *et al.* Systematic evaluation of chromosome conformation capture assays. *Nat Methods* **18**, 1046–1055 (2021).
27. Zhang, R., Zhou, T. & Ma, J. Multiscale and integrative single-cell Hi-C analysis with Higashi. *Nat Biotechnol* **40**, 254–261 (2022).
28. Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, eaau1783 (2018).
29. Gabriele, M. *et al.* Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging. *Science* **376**, 496–501 (2022).
30. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports* **15**, 2038–2049 (2016).
31. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**, E6456-6465 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

32. Cheng, R. R. *et al.* Exploring chromosomal structural heterogeneity across multiple cell lines. *Elife* **9**, e60312 (2020).
33. Finn, E. H. & Misteli, T. Molecular basis and biological function of variability in spatial genome organization. *Science* **365**, eaaw9498 (2019).
34. Finn, E. H. *et al.* Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. *Cell* **176**, 1502-1515.e10 (2019).
35. Tan, L., Xing, D., Chang, C.-H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924–928 (2018).
36. Dekker, J. *et al.* The 4D nucleome project. *Nature* **549**, 219–226 (2017).
37. Reiff, S. B. *et al.* The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nat Commun* **13**, 2365 (2022).
38. van Steensel, B. & Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**, 780–791 (2017).
39. Kim, J., Han, K. Y., Khanna, N., Ha, T. & Belmont, A. S. Nuclear speckle fusion via long-range directional motion regulates speckle morphology after transcriptional inhibition. *J Cell Sci* **132**, jcs226563 (2019).
40. Kim, J., Venkata, N. C., Hernandez Gonzalez, G. A., Khanna, N. & Belmont, A. S. Gene expression amplification by nuclear speckle association. *J Cell Biol* **219**, e201904046 (2020).
41. Alexander, K. A. *et al.* p53 mediates target gene association with nuclear speckles for amplified RNA expression. *Mol Cell* **81**, 1666-1681.e6 (2021).
42. Hua, N. *et al.* Producing genome structure populations with the dynamic and automated PGS software. *Nat Protoc* **13**, 915–926 (2018).
43. Tjong, H. *et al.* Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc Natl Acad Sci U S A* **113**, E1663-1672 (2016).
44. Girelli, G. *et al.* GPSeq reveals the radial organization of chromatin in the cell nucleus. *Nat Biotechnol* **38**, 1184–1193 (2020).
45. Chen, Y. *et al.* Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J Cell Biol* **217**, 4025–4048 (2018).
46. van Schaik, T., Vos, M., Peric-Hupkes, D., Hn Celie, P. & van Steensel, B. Cell cycle dynamics of lamina-associated DNA. *EMBO Rep* **21**, e50636 (2020).
47. van der Maaten, L. & Hinton, G. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).
48. Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29**,

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

115–129 (1964).

49. Roweis, S. T. & Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **290**, 2323–2326 (2000).
50. Tenenbaum, J. B., Silva, V. de & Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290**, 2319–2323 (2000).
51. von Luxburg, U. A tutorial on spectral clustering. *Stat Comput* **17**, 395–416 (2007).
52. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at http://arxiv.org/abs/1802.03426 (2020).
53. Ramdas, A., Garcia, N. & Cuturi, M. On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests. Preprint at http://arxiv.org/abs/1509.02237 (2015).
54. Eastwood, M. P. & Wolynes, P. G. Role of explicitly cooperative interactions in protein folding funnels: A simulation study. *J. Chem. Phys.* **114**, 4702 (2001).
55. Welch, B. L. THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED. *Biometrika* **34**, 28–35 (1947).
56. Belmont, A. S. Nuclear Compartments: An Incomplete Primer to Nuclear Compartments, Bodies, and Genome Organization Relative to Nuclear Architecture. *Cold Spring Harb Perspect Biol* **14**, a041268 (2022).
57. Zhou, J. *et al.* Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. *Proc Natl Acad Sci U S A* **116**, 14011–14018 (2019).
58. Osorio, D., Yu, X., Yu, P., Serpedin, E. & Cai, J. J. Single-cell RNA sequencing of a European and an African lymphoblastoid cell line. *Sci Data* **6**, 112 (2019).
59. SoRelle, E. D. *et al.* Single-cell RNA-seq reveals transcriptomic heterogeneity mediated by host-pathogen dynamics in lymphoblastoid cell lines. *Elife* **10**, e62586 (2021).
60. Krietenstein, N. *et al.* Ultrastructural Details of Mammalian Chromosome Architecture. *Mol Cell* **78**, 554-565.e7 (2020).
61. Zeiler, M. D. ADADELTA: An Adaptive Learning Rate Method. Preprint at http://arxiv.org/abs/1212.5701 (2012).
62. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Statist.* **22**, 79–86 (1951).
63. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
64. Scott, D. W. *Multivariate density estimation: theory, practice, and visualization*. (Wiley, 2014).
65. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.18.528138; this version posted February 19, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

detection of protein families. *Nucleic Acids Res* **30**, 1575–1584 (2002).

66. Zhang, J. *et al.* An integrative ENCODE resource for cancer genomics. *Nat Commun* **11**, 3696 (2020).
67. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996–1006 (2002).
68. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).
69. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612 (2004).