Characterizing the variation in chromosome structure ensembles in the context of the nuclear microenvironment

Priyojit Das[1], Tongye Shen[1,2] and Rachel Patton McCord[1,2,*]

[1] UT-ORNL Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, Tennessee, 37996, USA

[2] Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville, Tennessee, 37996, USA

* To whom correspondence should be addressed. Tel: 865-974-3149; Email: rmccord@utk.edu

ABSTRACT

Inside the nucleus, chromosomes are subjected to direct physical interaction between different components, active forces, and thermal noise, leading to the formation of an ensemble of three-dimensional structures. However, it is still not well understood to what extent and how the structural ensemble varies from one chromosome region or cell-type to another. We designed a statistical analysis technique and applied it to single-cell chromosome imaging data to reveal the heterogeneity of individual chromosome structures. By analyzing the resulting structural landscape, we find that the largest dynamic variation is the overall radius of gyration of the chromatin region, followed by domain reorganization within the region. By comparing different human cell-lines and experimental perturbation data using this statistical analysis technique and a network-based similarity quantification approach, we identify both cell-type and condition-specific features of the structural landscapes. We identify a relationship between epigenetic state and the properties of chromosome structure fluctuation and validate this relationship through polymer simulations. Overall, our study suggests that the types of variation in a chromosome structure ensemble are cell-type as well as region-specific and can be attributed to constraints placed on the structure by factors such as variation in epigenetic state.

INTRODUCTION

In eukaryotic cells, long genomic DNA is packed within a relatively small nucleus, requiring chromosomes to fold into a multi-layered 3D spatial organization (1). At the kilobase length scale, loops form between different chromatin regions (2-4). Some loops are formed by the interaction of cohesin-mediated loop extrusion and boundary factors, a process which can create domains of enriched contacts called Topologically Associated Domains (TADs) at the scale of 100s of kilobases (5-7). Through processes such as phase separation and tethering to nuclear structures the chromatin is further organized at the megabase scale into separate euchromatin and heterochromatin spatial regions, known as A/B compartments (8-12). Finally at the scale of the whole nucleus, chromosomes tend to form their own territories (13-15). Each of these structural features act independently as well as synergistically to contribute to different biological processes, as reported by an ever-increasing number of studies. For

example, chromatin loops and TAD boundary positioning can influence enhancer-promoter interactions and gene regulation (16-18). This multi-layered organization of 3D genome is not static: the chromatin fiber is semiflexible in nature and is also subjected to forces generated by ATP driven molecular machines along with thermal fluctuations (19-21). This leads to both spatial and temporal structural heterogeneity. Measuring this heterogeneity can help reveal detailed folding mechanisms and subpopulations of cells with different structures that mediate different biological outcomes (22).

Several key molecular and imaging techniques to characterize chromosome structure have been developed in recent years (23,24). Chromosome conformation capture (3C) based techniques, including Hi-C, can map the spatial organization of the whole genome in terms of contacts between genomic regions. However, when Hi-C and its derivatives are applied to millions of cells in bulk, it is challenging to study the heterogeneity of chromosome structures from the resulting population averaged measurements. Several polymer- and restraint- based models have been developed to reconstruct ensembles of structures from such population averaged Hi-C data (25-29). However, there are usually multiple possible theoretical ensembles that could fit the population average, so without single cell measurements for comparison, it is difficult to know whether models derived from bulk data capture the variation present in true biological systems. To resolve this issue, researchers have started to develop single-cell counterparts of Hi-C and related techniques (30-32). However, these single cell contact capture techniques often suffer from low coverage, which makes it difficult to characterize the structural heterogeneity within a particular cell-type.

In contrast, oligonucleotide-based FISH imaging techniques have enabled the tracing of chromatin regions at high resolution in thousands of single cells (22,33-37). These techniques can generate distance distribution maps between imaged points along a chromosome that resemble Hi-C contact maps. By analyzing these single-cell structures at different resolutions, recent studies have started to characterize the heterogeneity associated with the structural ensemble. For example, it has been revealed that the inactive chrX exhibits not only canonically expected compact structures but also more chromosome domain inter-mingling as compared to the active chrX (33). Another study revealed that domain-like structure also exists on the inactive chrX at the single-cell level, but such structures are averaged out and not visible in bulk experiments (22). Bintu and co-authors reported the existence of TAD-like structures and sharp domain boundaries at a single cell level (34). Though the boundary positions vary from one cell to another, they show higher preference for structural protein binding sites. By analyzing the single-cell whole chromosome tracing data at a higher resolution, researchers identified a substantial number of TAD-like domains in single cells that cross compartment boundaries observed in bulk data (35).

While initial studies focused on showing that variability in chromosome structures exists, recent work has begun to turn toward the question of how and why this variability arises (27,38,39). For example, transcriptional regulation has been proposed to influence an observed two state transition between closed

and open dumbbell-like conformations of a short chromatin segment spanning a TAD boundary (38). Other groups have analyzed the chromatin tracing data to evaluate the effects of the structural protein cohesin on the dynamics of TAD-like structures. Another study used a folding coordinate to represent the progression of TAD formation These analyses showed that in the absence of structural protein cohesin, a local chromatin region may still exhibit TAD-like structures in individual snapshots, but these structures are less stable due to the entropy gain of the unfolded conformations. (39). This has the effect of converting phase separated globules into more random coil-like state (27).

Although these recent studies have started to shed light on the fundamental mechanisms of local chromatin folding at the single-cell level, it is not clear how the chromosome structure variation differs between cell types or chromosome regions. Here, we focus on quantifying the full structure variation landscape rather than identifying multiple discrete structural states. Support for this view comes from a variety of recent studies. A FISH imaging study in *Caenorhabditis elegans* showed that chrV exhibits four major conformational clusters during early embryonic development, but that the structure ensembles corresponding to different clusters exhibit a broad and continuous distribution rather than a handful of discrete conformational states (36). Similarly, within a particular cell-type, the single-cell structure ensemble can be described as a distribution of conformations around the bulk average structure (mean). We do not know how this fluctuation around an average structure (variance) varies in a cell-type specific manner. This question becomes particularly interesting when the average structural organization looks similar across different cell types or conditions, but there may be different underlying structural ensembles that lead to similar averages. A logical follow up question is what factors might contribute to the extent of structural variation. There are various essential factors such as epigenetic marks, associated structural proteins, and nuclear lamina associations that vary in a cell-type specific manner and contribute to the organization of 3D genome (40-42). Therefore, investigating the role of these different factors in the variation of chromosome structural ensemble is of high importance.

To answer these questions about chromosome structural ensembles, we designed a statistical pipeline to analyze single-cell chromatin tracing distance data. Using this pipeline, we first characterized the landscape of single-cell structures across different cell-types. We then identified cell-type specific differences in the structural ensembles through joint analysis of structural landscapes. To further characterize these differences, we designed two different network-based approaches – one to quantify the degree of continuity across the chromosome structural ensemble (ensemble structure similarity index) and another to quantify the degree of structural variation of local genomic regions across the population (bin similarity index). Using these two approaches, we quantified the ensemble variation and local structural variation across different cell-types and different chromosome regions at different genomic resolutions. We identify overall locus compaction and domain boundary shifting as two primary but independent modes of structural fluctuation that are observed across all length scales. We observe differences in chromosome ensemble structural similarity between regions that have different epigenetic

state profiles, and we are able to demonstrate that interactions governed by epigenetic state can help explain these differences in the type of structure variation exhibited by the chromosome region.

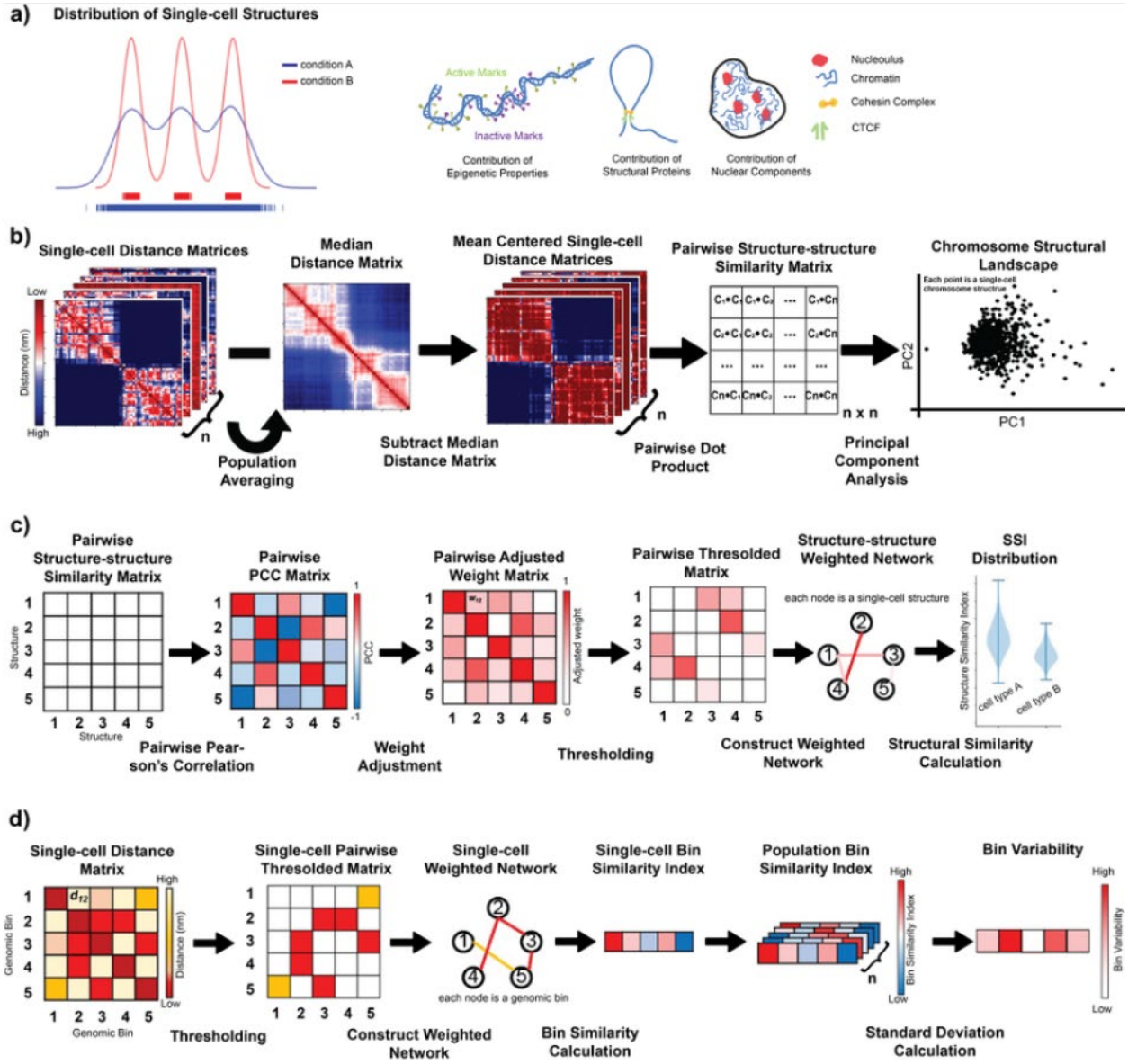

Figure 1. Schematic representation of the statistical pipeline used to quantify conformational diversity present in single-cell chromosome distance matrices. **a)** Cartoon representation of potential differences in chromosome structure landscapes (left). Condition A has a continuous landscape including all structures along a trajectory, while in Condition B, chromosomes fall into 3 major structural clusters. (Right) Conditions and factors that may influence this chromosome structure landscape. **b)** A pairwise similarity score is calculated between each pair of median-centered chromosome distance matrices. PCA is performed on the pairwise structure-structure similarity matrix and transformed data is projected in 2D space using PC1 and PC2 to create the "chromosome structural landscape". **c)** To calculate chromosome structural ensemble SSI, we construct a Pearson's correlation matrix from the pairwise structure-structure similarity matrix. This correlation matrix scaled, thresholded, and then converted into a weighted network from which SSI values can be calculated from each node. The median SSI value is plotted for many resamplings of nodes. **d)** To calculate bin variability, we convert each single-cell distance matrix into a thresholded matrix using a cutoff and then construct a network from each of those matrices with bins as nodes and contacts between bins as edges. We then calculate the bin similarity index for each bin in each single-cell and then take the standard deviation of the bin similarity index across the population as the bin variability.

MATERIALS AND METHODS

The statistical chromosome conformation analysis pipeline used in this work can be broadly divided into two parts. In the first part, the single-cell chromosome imaging data (of different genomic resolutions) is analyzed statistically to calculate pairwise similarity of the structures. Next, a structural similarity quantification approach is applied on that similarity data to quantify the degree of continuity present in the chromosome structural ensemble. In addition, the pairwise similarity data is used to construct a structural landscape to arrange the single-cell chromosome structures according to their individual and bulk structural properties (Fig. 1).

Construction of distance matrices from single-cell chromosome imaging data

For this work, we have used experimental single-cell chromosome imaging data at three different resolutions, each of them reflecting different levels of structural organization of the 3D genome at different genomic length scales. The single-cell chromosome imaging data obtained from Bintu *et al*. (34). has the highest genomic resolution of 30 kb. Here, the authors traced the conformations of multiple ~ 2 – 2.5 Mb spanning chromatin regions of chr21, partitioned into consecutive 30 kb segments in single-cells across different cell-types (e.g., IMR90 – human fetal lung fibroblast cells, K562 – human erythroleukemic cells, A549 – human lung adenocarcinoma cells, and HCT116 – human colon cancer cells) during different physiological and perturbed conditions. The second set of IMR90 chromatin tracing data analyzed here is obtained from Su *et al*. (35), where the single-cell conformations of chr21 non-repetitive region were traced using consecutive 50 kb bins. In addition, that dataset also provides chr2 conformations captured at the same 50 kb resolution, however with an interval of 250 kb between two genomic regions, due to very large size of human chr2. Both the Bintu *et al*. and Su *et al*. data capture chromatin conformations at a comparable resolution (30 and 50 kb respectively), differing only in the size of the genomic region analyzed. Whereas the Bintu *et al*. data primarily represents the structural heterogeneity at the level of specific TADs and sub-TADs, the Su *et al*. data captures the heterogeneity of almost the whole chromosome. Finally, we also analyzed chromosome structural heterogeneity of the whole active and inactive chrX of IMR90, obtained from Wang *et al*. (33). In this dataset, TADs are imaged as the basic subunits along the chromosome, so the resolution of this data is lower (ranging from 440 kb to 7.88 Mb) and reports on the conformations of compartments and the whole chromosome rather than the within-TAD level.

For all these different datasets, the authors reported the single-cell conformations of the chromosomes in a 3D Cartesian coordinate system where each basic genomic region (e.g., 30 or 50 kb genomic region or TAD; from here termed as bin) has X-, Y-, Z- coordinate values. However, in some single cell structures the 3D positions of some of the bins are missing due to technical difficulties. Therefore, to decrease the effect of missing data, we consider only structures with more than ~ 90% measured bins, and the values for missing bins in those structures are assigned by nearest-neighbor interpolation. The 3D positions of chromosomes are also associated with rotational and translational variations which might affect

underlying statistical analysis. To resolve this issue, we convert each 3D structure into a 2D symmetric distance matrix by calculating the pairwise Euclidean distance between all the genomic bins of that chromosome structure. For example, if a particular chromosome has m genomic bins and a total of N copies originating from N single cells, then we will have N symmetric pairwise distance matrices each of size $m \times m$, where d_{ij} represents the Euclidean distance between i^{th} and j^{th} genomic bins. Further, the spatial distance can be converted into contacts by applying a genomic resolution specific thresholding cutoff. However, for our analyses, we use distance matrices instead of contact matrices, unless otherwise explicitly mentioned.

Generate chromosome structural landscapes using statistical analysis technique

We next center all the single-cell matrices around zero by subtracting the median distance matrix from each of them, $h_{ij} = d_{ij} - median(d_{ij})$. Once we obtain the median centered single-cell distance matrices, we calculate the pairwise dot product between each of those single-cell matrices, such as, $s_{\alpha\beta} = \sum_{i,j} h_{ij}^{\alpha}.h_{ij}^{\beta}$, where h^{α} and h^{β} are mean centered distance matrices obtained from single cells ($\alpha, \beta \in \{1, \ldots, \mathrm{N}\}$), respectively. This step calculates the similarity between the structures, where two similar conformations would yield a strong positive value and vice versa. The pairwise dot product similarity information between all possible pairs of the single cells is represented in a symmetric matrix format. Next, we perform PCA on the pairwise similarity matrix and project the single-cell conformations in 2D space using the first two principal components (PCs). This PC projection is termed the chromosome structural landscape.

Quantification of chromosome structural ensemble variation using a network-based similarity index

To quantify the variation associated with the chromosome structural ensemble, we perform a network-based similarity calculation on the pairwise similarity matrix (Fig. 1c). Similar approaches have been used to study cellular network heterogeneity associated with cancer and cellular differentiation (43-45). In this approach, we first calculate Pearson's correlation coefficient (PCC) between each pair of rows of the similarity matrix, where c_{ij} denotes PCC between i^{th} and j^{th} rows of the pairwise similarity matrix. As PCC values ranges between -1 and 1, next we scale the PCC values in the range of 0 and 1, $w_{ij} = 0.5 * (1 + c_{ij})$. Once we have the rescaled pairwise PCC value matrix, we construct a weighted network from that where each node represents a single chromosome structure or conformation and the edges between them denotes rescaled PCC values. However, we only consider the edges having weights above a certain threshold, representing only connections between highly similar structures. This threshold is calculated by taking 50th percentile of all the edge weights of the network. With the constructed network, we calculate the structure similarity index (SSI) of each node S_i, using a formula based on the Shannon-Jayne network entropy (43).

$$S_i = -\frac{1}{\log k_i} \sum_{j \in N(i)} p_{ij} \log p_{ij}$$

where $N(i)$ represents the neighborhood set of the node i and k_i denotes the cardinality of that set. p_{ij} represents the proportionality of the edge weight between nodes i and j compared to all the edges associated with node i.

$$p_{ij} = \frac{w_{ij}}{\sum_{k \in N(i)} w_{ik}}$$

This index S_i is bound between 0 and 1. In our system, the resulting values are compressed near a value of 1, so in order to make comparisons more easily, we rescale all values to an arbitrary unit by first subtracting 0.99 from the values and then by multiplying them with 1000. For our presented analyses, we calculate the median SSI across all nodes in the landscape and call this the ensemble SSI. Specifically, to normalize sample size between conditions, we iteratively resample a subset of nodes from each condition, then plot the distribution of median SSI values from the resampled node sets. Based on the distribution of the median SSI values, we characterize the heterogeneity of the chromosome structural landscape. An SSI distribution with higher mean is considered as having a smooth continuous structural landscape while a distribution with relatively lower mean would have more discrete, disconnected structures. Interpretation of the SSI distribution trends have been discussed in detail in the **Results** section.

Calculating chromosome structure variability for each genomic bin

To characterize the variability of structure at each genomic bin along the chromosome, we use a modified version of the network-based structural similarity calculation approach, discussed in the previous subsection (Fig. 1d). In this step, for each single-cell structure, we first construct a network based on its distance matrix, where each node is a single genomic bin and the edges between them represent the spatial distances. We then threshold the network using an experimentally-derived distance cutoff (330 nm for 30 kb resolution (38) and 500 nm for 50 kb resolution (35)), where edges with spatial distance higher than the cutoff are discarded from the analysis. Next, we calculate the bin similarity index of each of the nodes (bins) in the network using the same formula discussed in the previous section, except here using genomic bins as nodes rather than whole single-cell structures and spatial distances as edge weights rather than structure correlation coefficients. We now have a bin similarity index value for each genomic bin in each single cell structure. To quantify the variability of structural positions occupied by a given bin across all structures, we then calculate the standard deviation of bin similarity index values across all single cell structures for a given bin. We call this measure the bin variability. A higher standard deviation corresponds to a more dynamic or variable genomic bin.

Characterization of chromatin states using epigenetic profiles

In order to characterize chromatin states of the IMR90, K562 and A549 cells, we downloaded a total of eleven histone marks – H3K4Me1, H3K4Me2, H3K4Me3, H3K9Ac, H3K27Ac, H4K20Me1, H3K36Me3, H3K27Me3, H3K9Me3, H2A.Z and DNase for all the three cell types from ENCODE repository (Table S1) (46). Next, we used chromHMM software (47) to annotate genomic regions of desired resolutions into five chromatin states using a combinatorial pattern of the eleven histone marks. Based on the combinatorial histone marks, we defined chromatin states as – Active state 1, Active state 2, Inactive Poised, Repressed Polycomb and Heterochromatin (Fig. S1). Briefly, if an inferred state shows strong enrichments for active marks such as H3K4Me1/2/3, H3K27Ac, H3K9Ac and DNase, we term that state as an Active state. Further, depending on the level of enrichment of those different active marks, we characterize them into Active state 1 and Active state 2. We assign the Heterochromatin label to the state for which we observe strong enrichment of H3K9Me3 and for the Repressive Polycomb assignment, we look for strong H3K27Me3 and a mix of H3K9Me3 and H4K20Me1 marks. And the rest of the states are assigned the Inactive Poised level.

Polymer simulation of chromatin segments based on underlying chromatin states

To test the contribution of epigenetic states to chromosome structure heterogeneity and dynamics, we perform polymer simulations of a specific chromatin segment with the interactions defined based on their epigenetic states. In recent years, polymer simulation-based studies have shown how the compartmental profile or the underlying epigenetic state of the genomic regions can give rise to the population-averaged structural organization of chromatin at different resolutions (9,48,49). Here, we represent a ~ 1.95 Mb spanning chromatin region as a chain of beads using a bead-spring polymer model. The model consists of 1950 beads where beads represent genomic regions of 1 kb. Each bead is assigned a specific chromatin state as described above. In addition to that, we also model loops within the specific chromatin regions based on publicly available high-resolution Hi-C derived chromatin loop sets from Rao *et al*. (50). For each loop, we connect the loop anchors using a harmonic constraint. The potential energy function of the polymer model is described below,

$$\mathrm{U_{ref}}(r) = U_B(r) + U_P(r) + U_L(r) + U_W(r)$$

where $U_B(r)$ corresponds to bond stretching potential, modeled by finite extensible nonlinear elastic potential given by,

$$U_B(r) = -0.5\mathrm{K_S}R_0^2 \ln\left[1 - \left(\frac{r}{R_0}\right)^2\right]$$

where $\mathrm{K_S} = 30.0\frac{\epsilon}{\sigma^2}$ is spring constant and $\mathrm{R_0} = 1.6\sigma$ is the equilibrium bond length.

We use Lennard-Jones potential to model steric repulsion and attraction between the beads of different chromatin state with different parameters. For example, the potential between a bead of chromatin state A and that of B is represented below,

$$U_P(r) = 4\epsilon_{\mathrm{AB}}\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

where ϵ_{AB} denotes interaction strength between bead types A and B (STable 2) and σ refers to the size of each bead.

In case of loops, we use harmonic potential between each of the loop anchors $U_L(r)$, represented by,

$$U_L(r) = K_L(r - R_0^L)^2$$

where $\mathrm{K_L} = 300\frac{\epsilon}{\sigma^2}$ is spring constant and $\mathrm{R}_0^L = 1.2\sigma$ is the equilibrium loop bond length.

Finally, in order to model the repulsive interaction between the chromatin beads and the simulation box boundaries, we use Lennard-Jones potential

$$U_W(r) = 4\epsilon_{\mathrm{W}}\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

where $\epsilon_{\mathrm{W}} = 1.0$ denotes the interaction strength between boundary walls and chromatin beads irrespective of their types.

The LAMMPS molecular dynamics package is used to perform polymer dynamics simulations (51). All the simulations are performed in terms of Lennard-Jones reduced units with τ for time, σ for length and ϵ for energy. To mimic the viscoelastic nature of the nucleoplasm, Langevin dynamics with a damping coefficient of $\gamma = 1.0\tau$ is used to maintain the temperature at $T = 1.0$. We then run the simulation for a total of 1,000,000 steps with a timestep of $dt = 0.000001\tau$. Followed by that, the timestep dt is increased to 0.01τ and a production run of a total 2,000,000 steps is performed and the conformations are saved at intervals of 1,000 steps. We repeated this process for 20 times, each time with different initial conformation and random seeds.

<u>Initial conditions and Equilibration</u> The chromatin fiber is initialized as a self-avoiding worm like chain using the python package PolymerCpp (52) and is placed in a box of size $L = 35\sigma$ with fixed repulsive boundaries. We then allow the polymer to equilibrate inside the box using four consecutive phases. In the first phase, a soft potential is used to remove overlap in the polymer for 400,000 steps with a $dt = 0.01\tau$.

$$U_S(r) = \mathrm{A}\left[1 + cos\left(\frac{\pi \mathrm{r}}{\mathrm{R}_0^S}\right)^6\right]$$

where A is pre-factor, varied gradually from 0 to 100ϵ and $\mathrm{R}_0^S = 2^{1/6}\sigma$.

For the rest three subsequent phases, we switch to potential to Lennard-Jones potential and run the simulation for 400,000, 500,000 and 500,000 steps with timesteps 0.000001τ, 0.0001τ, and 0.01τ respectively.

$$U_G(r) = 4\epsilon_G \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right]$$

where $\epsilon_G = 1.0$ denotes the interaction strength between chromatin beads irrespective of their types.

Loop extrusion simulation of chromatin segments based on underlying RAD21 binding sites

To test the contribution of loop extrusion dynamics to the variability present in chromosome structural ensemble, here we perform loop extrusion polymer simulations of a specific chromatin segment as per the method described in Fudenberg *et al*. (2). For these simulations, we define the domain boundaries based on RAD21 peaks present in the desired chromatin segment. Here, we model a 2 Mb spanning chromatin region as a 2000 beads monomer chain where beads represent genomic regions of 1 kb. All the simulation parameters are kept same between different cell types except the position of the boundary elements (here RAD21 peaks) and their strengths. The strength of boundary elements are defined in terms of boundary permeability parameter in the models which represents the boundary insulation. For each cell-type, these parameter values are obtained by calculating the relative strength of the RAD21 peaks. The RAD21 peaks for the IMR90 and K562 were downloaded from the ENCODE repository (46).

Analysis of nucleolus-associated domain (NAD) mapping data

In order to characterize the nucleolar association of the IMR90 chromosomal regions, we downloaded the NAD mapping data of the proliferating IMR90 cells from Dillinger *et al*. (53). This microarray data contains log2 normalized ratio of nucleolus-associated and genomic DNA at specific probe locations. For our analysis, we bin that log2 normalized ratio data at 10 kb resolution using multiBigwigSummary functionality of the deepTools software (54).

RESULTS

Chromosome structural landscape reveals major features of the structural ensemble

We first apply our statistical analysis technique to the single-cell chromosome structures measured for a 2 Mb region of IMR90 chr21 (28-30 Mb) (Fig. 2a) and produce a chromosome structural landscape (Fig. 2b). We observe that the first major source of variation corresponds well to the compactness of the chromosome structures, which can be quantified using the radius of gyration, R_g (Fig. S2a). We calculate the R_g of the single-cell structures from their experimentally obtained 3D conformations. When we compare the R_g values of the structures with their PC1 values (from here on we refer to the PC1 projection as “PC1”), we observe that the structures with more negative PC1 projections are highly compact with smaller R_g values while structures with more positive PC1 values are less compact with

higher R_g scores. With increasing R_g, the structural divergence also increases, with individual structure points becoming spread further along PC2 and also more spread out from each other.

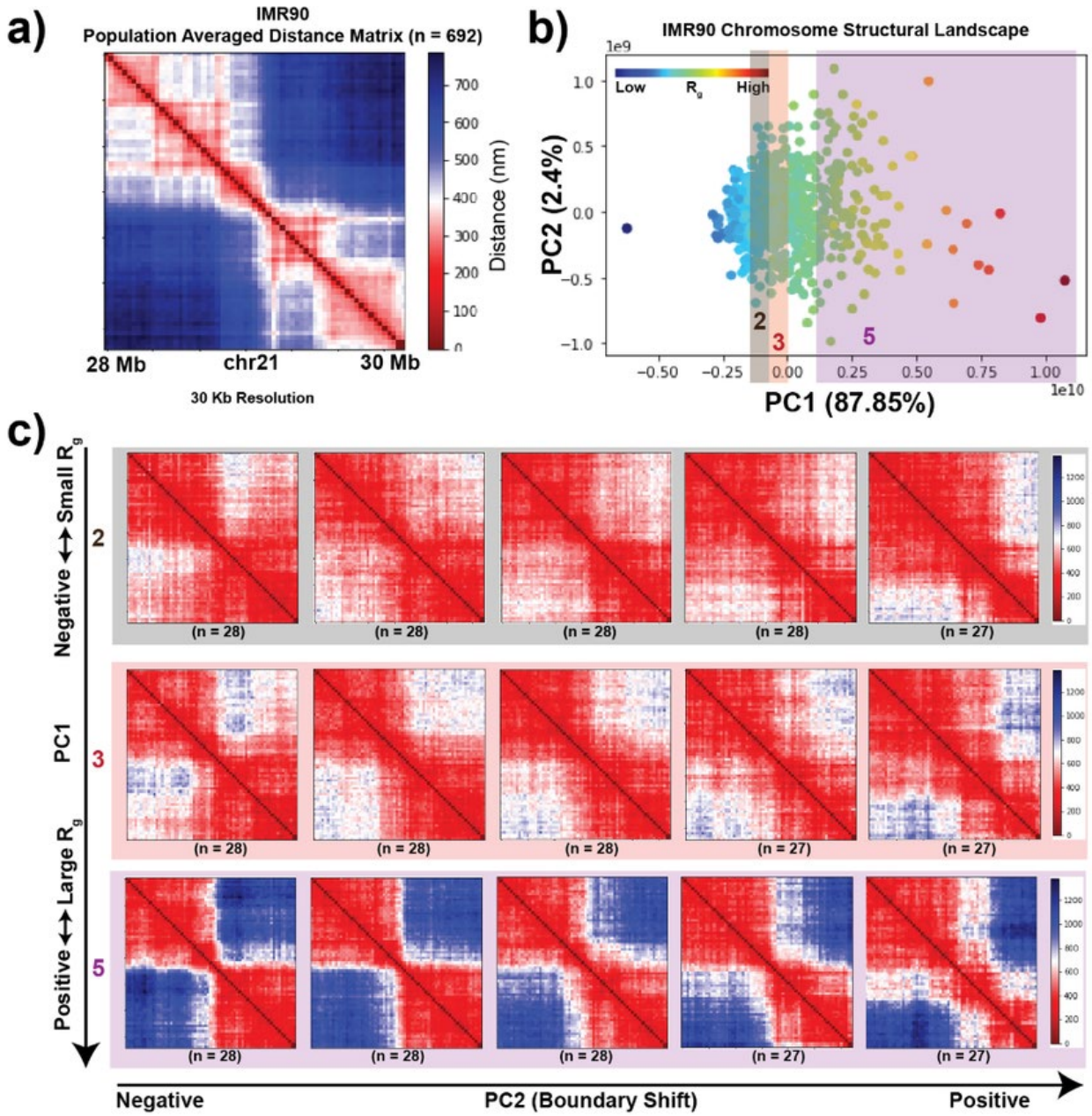

Figure 2. The chromosome structural landscape orders single-cell structures on a continuum based on their radius of gyration followed by domain reorganization. **a)** Population-averaged distance map of IMR90 chr21:28-30 Mb at 30 kb resolution. **b)** Chromosome structural landscape of IMR90 chr21:28-30 Mb chromatin region, generated as described in the **Methods** section. Each point represents a single-cell structure. Points are colored according to the radius of gyration (R_g) of the corresponding structures (dark blue – smaller R_g and dark red – larger R_g). The percent of variation explained by each PC is shown in parentheses. **c)** Structures are first divided into five groups based on the PC1 ordering (which correlates highly with R_g). Structures belonging to groups 2, 3 and 5 are shown. Within each group structures are further divided into five subgroups based on their PC2 ordering. Average structures within each subgroup are shown.

Along PC2, we observe that the structure ordering follows a progressive reorganization of the domains within the region and captures several subclusters of structures (Fig. 2c). As we move from negative to positive PC2, we observe a shifting of a boundary (weakening of one boundary and gaining of another) that leads to the change in the structure of the domain (Fig. S2b). This phenomenon can be seen in different regimes of R_g, showing the existence of domain reorganization within both compact and open structures (Fig. S2b).

When the same approach is applied to different chromosomes and cell-types at different resolutions, we found a consistent ordering of the dynamic modes of the structural landscape (Fig. S2c, Fig. S2d, Fig. S3 and Fig. S4). Although the major dynamic variation remains the R_g ordering for all the different situations, the reorganization of domains along PC2 is dependent on the length of the genomic region as well as the local structural properties.

Ensemble structure similarity index captures cell-type- and chromosome-specific features of the chromosome structural ensemble

To compare the single cell structural heterogeneity between different cell types, we constructed a joint structural landscape using the 28-30 Mb chr21 region data from three cell types – IMR90 (lung fibroblast), K562 (leukemia) and A549 (lung adenocarcinoma) collectively. In this joint structural landscape we can visualize the cell-type-specific variation present in structural ensemble (Fig. 3a). To quantify the differences between cell types, we divide the landscape in 10 equally sized regimes along PC1 (since PC1 correlates with R_g, different PC1 regimes correspond to different R_g values) and calculated the proportion of different cell types within each regime. We observe that K562 shows a higher proportion of chromosome structures in lower R_g regimes and has an overall more compact structural ensemble compared to A549 and IMR90 (Fig. 3a and Fig. 3b).

While the relative R_g shift between cell types is straightforward to observe and quantify, other features of the chromosome structural landscape are more difficult to compare visually. Therefore, we devised a network-based measurement to quantitatively compare the structural landscapes, as discussed in the **Methods** section. According to this approach, a network with similar edge weights will exhibit a higher ensemble SSI value and vice-versa. In terms of the chromosome structural ensemble, a set of structures that smoothly transition across the entire conformational space will have a highly connected structural landscape and thus high ensemble SSI (Fig. 1a, blue line). On the other hand, if structures coalesce around discrete states (Fig. 1a, red line), the discontinuity in the structural ensemble will be higher, and the resulting network will have a larger variation in of edge weights, and a low ensemble SSI score will be produced.

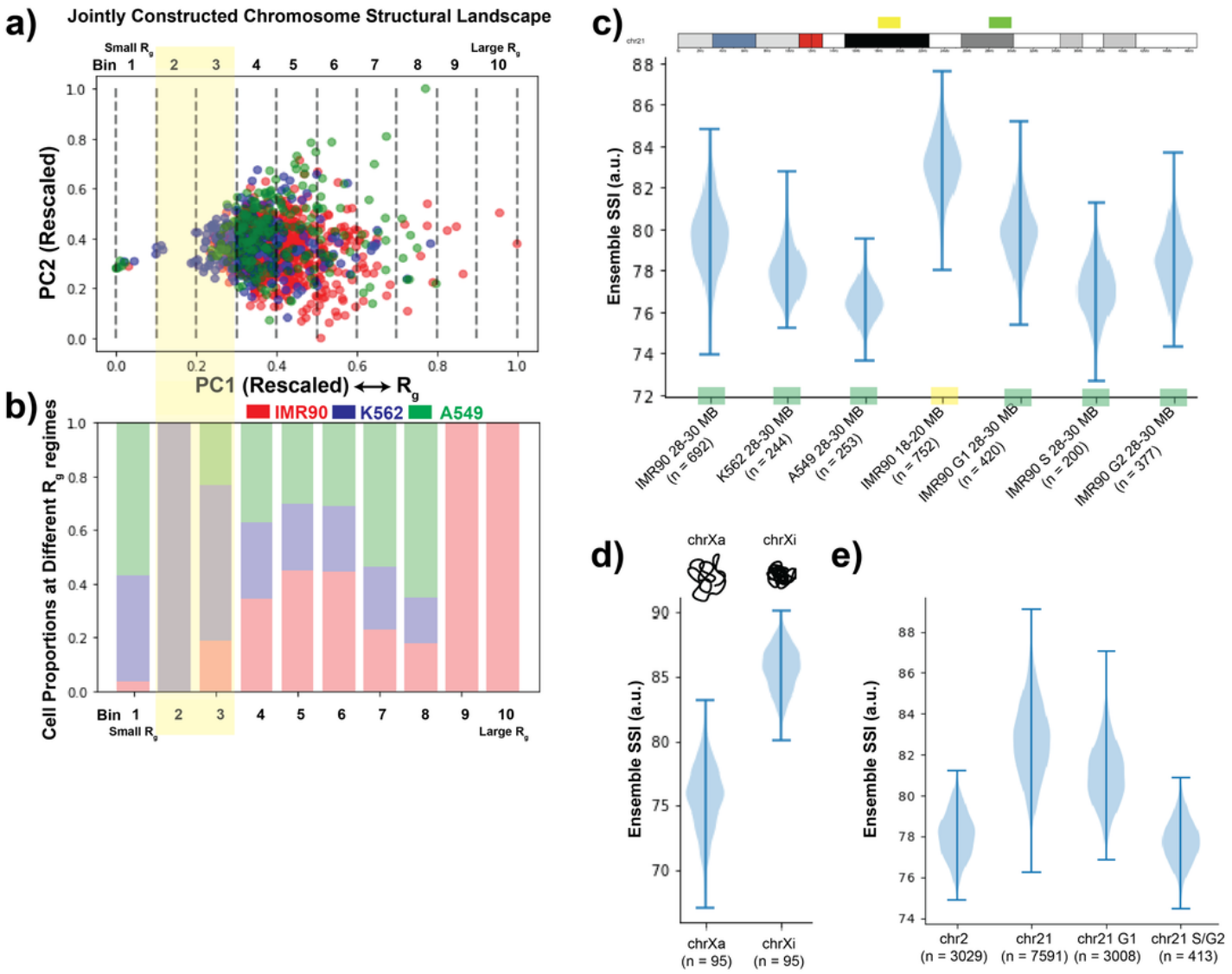

Figure 3. Network based similarity score quantifies cell-type- and chromosome-specific variability present in chromosome structural ensemble at different resolutions. **a)** Jointly constructed chromosome structural landscape of genomic region chr21:28-30 Mb for cell types IMR90, K562 and A549. The only difference between individual and joint landscape is that all the different cell-type structures are considered together while calculating pairwise structure-structure similarity matrix. Points are colored based on the cell types. Further, the landscape is divided into ten equal sized bins (or regions) based on PC1 values, where the bins on the left represent lower R_g regions and bins on the right are with higher R_g. **b)** Relative proportion of IMR90, K562 and A549 single-cell structures present in each bin represented in a). **c)** Chromosome structural ensemble SSI distributions for different 2 Mb chromosome regions from different cell-types and conditions, obtained from Bintu *et al*. (34). The SSI value of the structural ensemble is calculated using a network entropy based approach. For each structural ensemble (size of ensemble indicated as n), 200 single-cell structures are selected randomly (except IMR90 S 28-30 Mb where we select 150 structures for resampling), and the median of the SSI values of the nodes are calculated. This step is repeated 1000 times and values are plotted as a distribution. **d)** Chromosome structural ensemble SSI distributions for active and inactive IMR90 chrX at TAD resolution, obtained from Wang *et al*. (33). 50 single-cell structures are selected randomly each time for resampling. **e)** Chromosome structural ensemble SSI distributions for different IMR90 chromosomes and conditions at higher resolution, obtained from Su *et al.* (35). Here, for resampling 300 single-cell structures are selected randomly each time.

We first applied the network-based similarity quantification approach to structures of 2 Mb regions of chr21 (30 Kb resolution) in different cell-types and conditions obtained from Bintu *et al*. (34). There is variation in the sample size (i.e., the number of conformations) obtained for different experiments, so we relied on a resampling approach to make the comparison consistent. Based on this approach, when we compared the 28-30 Mb genomic region from three different cell-types analyzed above, we find that IMR90 has the highest ensemble SSI, followed by K562 and then A549 with the lowest ensemble SSI (Fig. 3c). Thus, this region in IMR90 cells has the highest continuity across different structures while the corresponding distribution in A549 cells exists as a relatively discrete set of structures.

To evaluate our interpretation of the ensemble SSI result in an orthogonal fashion, we performed hierarchical clustering of the 3D structures obtained from IMR90 and K562 cells based on a structure similarity metric, Q, as represented in Cheng *et al*. (Fig. S5). At the same threshold, both cell types showed the same number of structure clusters. However, in IMR90, the vast majority of structures fall into one of these clusters, corresponding to our high ensemble SSI value (Fig.S5a and b). In K562, the structures are more evenly split into the different clusters, corresponding to our low ensemble SSI value (Fig. S5c and d). Thus, our ensemble SSI metric indeed captures the degree of continuity within the structural ensemble. By using ensemble SSI rather than a hierarchical clustering approach, we avoid the need to arbitrarily select a cutoff or number of clusters. We further assessed the dependence of the ensemble SSI score on the number of cells measured (Fig. S6). We observed that with an increase in the number of cells, the difference in ensemble SSI between cell-types follows the same trend, but the absolute difference becomes smaller (Fig. S6b). It is important to note that as we increase the number of cells, we are including more low-quality measurements (cells with more missing data) and this may also contribute to a broader ensemble SSI distribution. As a final technical test of our metric, we repeated the ensemble SSI analysis after removing extremely compact or large structures from each cell type to see whether those extreme conformations are disproportionately affecting the results (Fig. S7). We found that those extreme structures actually have little influence on the ensemble SSI calculation. Overall, we conclude that our ensemble SSI calculation is robust.

We next compared two different genomic regions of IMR90 chr21 using the ensemble SSI distribution. We found that although both are from same cell-type and same chromosome, the 28-30 Mb region has lower ensemble SSI values compared to 18-20 Mb region (Fig. 3c). This difference indicates that the latter region has a higher degree of continuity in its structural ensemble. This 18-20 Mb region lies entirely in the B compartment (putative inactive, gene poor regions) while the 28-30 Mb region contains a majority of A compartment regions mixed with B compartment regions. To further investigate the potential connection between compact, inactive states and the continuity of the structural landscape, we compared the ensemble structural similarity of the active vs. inactive chrX structural ensembles (440 Kb to 7.8 Mb) (Fig. 3d). We found that chrXi has higher ensemble SSI values compared to its active counterpart. This suggests that the chromosome which has a highly compact global structure has a more continuous

structural landscape compared to the more specific and discrete states of the open and active structure. (Fig. S8a, Fig. S8b and Fig. S8c) (55).

When we analyzed the structural variation of the entire chr2 and chr21 q-arm at high resolution (35), we observed that chr21 has a higher ensemble SSI compared to chr2 (Fig. 3e). To test whether the difference in length between chr2 and chr21 contributed to differences in the ensemble SSI values, we calculated the ensemble SSI of regions of different lengths selected from chr2. We observed that as the length increases the ensemble similarity decreases (Fig. S8d). This effect likely stems from the increased number of possible conformations that a longer polymer can adopt, which corresponds to a sparser coverage of conformational space. We also found that ensemble similarity is lower in S/G2 phase compared to G1 phase for chr21. This result is consistent with the ensemble SSI values of 28-30 Mb IMR90 chr21 region structural ensembles from different cell-cycle stages, where the G1 phase has highest ensemble SSI followed by S and G2 phases, as shown in Fig. 3a. One major difference between the chromosomes in the phases of the cell cycle is the addition of tethers between sister chromatids in S and G2 compared to G1 (56). Increased constraints on the chromosome structure could then lead to more distinct structures and lower structural ensemble similarity.

Contribution of epigenetic properties to the cell-type-specific variation of chromosome structural ensemble

Building on the observation that regions with different A/B spatial compartment identity had different structural landscape SSI, we further examined the influence of local epigenetic state on the resulting structural landscape. Recent studies have shown the association of epigenetic states with the ensemble genome organization (25,26,57). Figure 4a shows chromatin states along the 28-30 Mb genomic region in the three different cell types – IMR90, K562 and A549 (detailed steps in the **Methods** section). We observe that this region in IMR90 mostly has active marks while in K562 the region is mixed between repressive polycomb and heterochromatin marks as well as active marks. The A549 region shows a mixture of active, repressive, and poised marks. By qualitatively comparing this chromatin state data with the ensemble SSI distributions for these three cell types, we observe that the cell type with the more uniform epigenetic state (IMR90) has the highest ensemble SSI while cell types with mixed epigenetic states exhibit lower SSI.

In addition to the overall dynamic properties of the region, we sought to quantitatively compare local genomic bin epigenetic state with the degree of structure fluctuation experienced by that bin. We first calculated the bin variability for each cell type (**Methods**) which represents the structure variability of a particular genomic region across the population of cells (Fig. 4b). We do not see a direct bin-by-bin correlation between epigenetic state and bin variability. Instead, we observe that a higher proportion of active states within the region tends to correspond to higher bin variability across the region while a larger proportion of inactive states, even if focused in one part of the locus, corresponds to restricted bin variability across the whole region. In this comparison, it is important to remember that we are comparing

a population averaged epigenetic state with single cell structural fluctuation. It is possible that additional relationships could become evident if single cell variation in epigenetic state were considered.

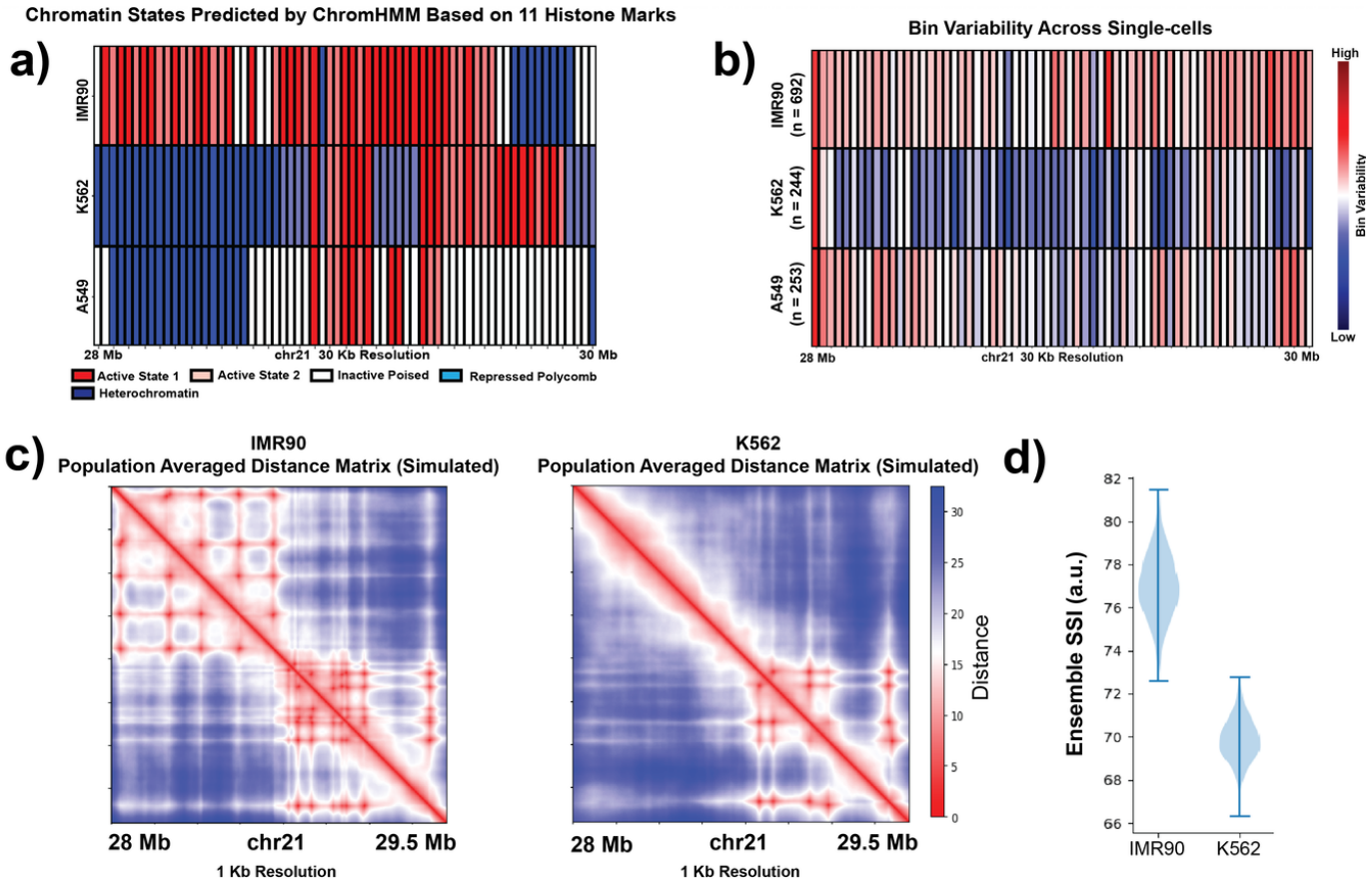

Figure 4. Epigenetic properties (histone modifications) contribute to the cell-type-specific variation in the chromosome structural ensemble. **a)** Chromatin states of chr21:28-30 Mb genomic region from three different cell-types – IMR90, K562 and A549 at 30 kb resolution. The chromatin states are inferred based on 11 different histone marks for each cell-type by ChromHMM tool. **b)** Bin variability of the chr21:28-30 Mb region from IMR90, K562 and A549, where a higher value represents higher degree of structural variation of that genomic position across the population and vice versa. **c)** Population-averaged distance maps of IMR90 and K562 chr21:28-29.5 Mb genomic region from polymer simulations (see Methods) at 1 kb resolution. **d)** Chromosome structural ensemble SSI distributions for chr21:28-29.5 Mb from simulated IMR90 and K562. For resampling, 500 single-cell structures are selected randomly each time.

To test our hypothesis that epigenetic state differences contribute to cell-type-specific variation in the chromosome structural ensemble, we performed a set of polymer simulations of the desired genomic region of 28-29.5 Mb at a resolution of 1 kb, as described in the **Methods** section. We generated the models for IMR90 and K562 separately by assigning interactions between the polymer beads based on their specific chromatin states (Fig. S9a). Once we obtained the 3D conformations from the simulations, we constructed pairwise Euclidean distance matrices for each conformation and further quantified the structural ensemble using our structural similarity approach. For the resampling purpose, we combined all the snapshots from 20 independent simulations and then selected 500 conformations randomly from this

set 1000 times. When we compared the simulated IMR90 ensemble SSI distribution with simulated K562 data, we found that the K562 structural ensemble shows a lower similarity, which is the same trend observed in our analysis of the Bintu *et al*. imaging data (Fig. 3c and Fig. 4d). Since the only difference between cell types in these simulations was the encoded epigenetic state of the polymer beads, this result corroborates the idea that the epigenetic state is one factor that can contribute to the variation in chromosome structural ensemble.

Analysis of the genomic regions extracted from the whole chromosome imaging data reiterates the effect of epigenetic properties of structure heterogeneity

We further evaluated the contribution of epigenetic factors to structural heterogeneity by analyzing single-cell structures of six genomic regions, each 2 Mb in length, obtained from IMR90 chr21 whole q-arm imaging data (Fig. 5a) (35). For simplicity, the regions will be denoted as S1 to S6 based on their starting genomic coordinates. The regions S1 (chr21:18-20 Mb) and S2 (chr21:28-30 Mb) were selected in order to match the two IMR90 genomic regions data analyzed in Bintu *et al*. The three regions S3, S5 and S6 were selected based on their epigenetic state similarity with epigenetic state of the chr21:28-30 Mb region in K562 cells (Fig. S9b). To evaluate epigenetic state similarity, we designed a sliding window-based approach where we move a window of size 40 (2 Mb genomic region with 50 kb resolution) over the whole q-arm of chr21 and for each window compare the IMR90 epigenetic state with the K562 28-30 Mb region epigenetic state using binary comparison (Fig. S9c). However, as the resolution of the Bintu *et al*. data (65 bins per 2 Mb region) is different from the Su *et al*. data (40 bins per 2 Mb region), we take the least common multiple of the two bin numbers and expanded each data to 250 bins by repeating each bin element the required number of times. Here, a higher comparison value shows higher similarity with the K562 epigenetic state. Among the three regions, S3 has the highest score, S6 has the second highest and S5 has the third highest one (Fig. S9c). The region S4 was selected randomly.

When we quantified the variation in the structural ensemble of these six regions, we first observed that the S1 and S2 regions have ensemble SSI distributions that are consistent with the results from the same regions in the Bintu *et al*. data (Fig. 5b). Next, we found that the S3 region has the lowest ensemble SSI distribution among all the six regions which is also consistent with the low ensemble SSI of the Bintu *et al*. K562 region. In addition, the ordering of S6 and S5 are also consistent with their similarity with K562 epigenetic state. Finally, the S4 region, which has a very low similarity score with K562 state, exhibits a very high ensemble SSI distribution. Again, this analysis supports the idea that epigenetic patterns contribute to the degree of variation present in the chromosome structural ensemble.

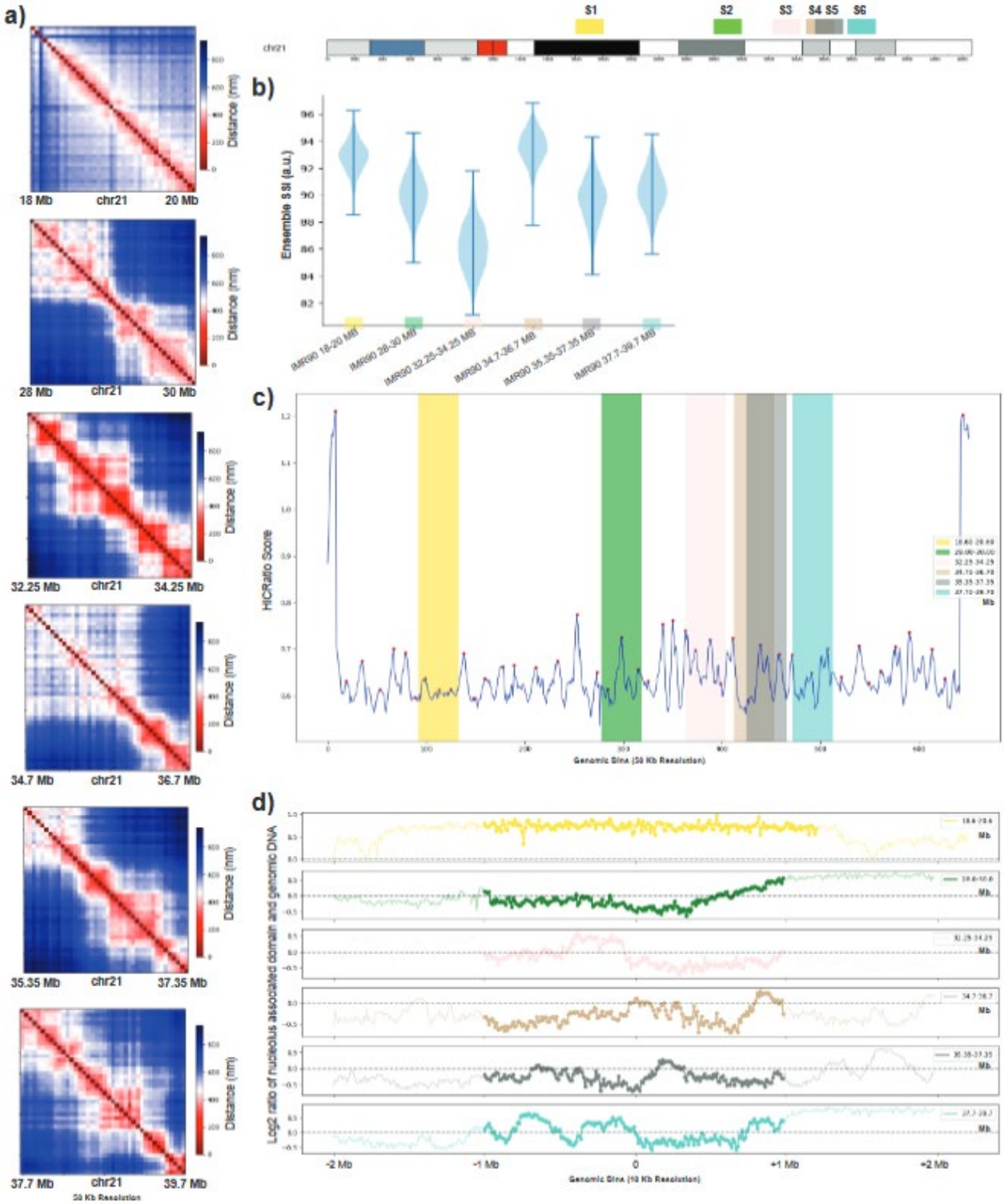

Figure 5. Evaluation of structural variation of regions with different epigenetic, genomic, and structural properties. **a)** Population-averaged distance maps of 6 different genomic regions of IMR90 chr21 at 50 kb resolution obtained from Su *et al.* (35). Locations of these 6 regions are indicated on the top right chromosome ideogram and labeled with colors used throughout the figure. **b)** Chromosome structural ensemble SSI distributions for the same 6 regions of IMR90 chr21. For resampling, 300 single-cell structures are selected randomly each time. **c)** HiCRatio score of each genomic bin of chr21 q-arm at 50 kb resolution (calculated with a 300 kb window size). The red points represent the boundary positions inferred by HiCRatio. **d)** Nucleolus association level of the same 6 regions of IMR90 chr21 (53). For each genomic region, the 1 Mb upstream and downstream regions are also shown. A value greater than 0 represents higher levels of nucleolus association and vice versa. Colors in each panel correspond to chromosome locations indicated above panel b.

In addition to the epigenetic states, we also called domain boundaries along the whole q-arm of chr21 using the HiCRatio approach (with a 300 kb window size) (58). Since this boundary calling software is designed for contact data, we first converted the single-cell chromosome structural distance data into contact data using a cutoff of 500 nm, as reported here. We found that regions with a higher number of strong domain boundaries exhibit lower ensemble similarity and vice versa (Fig. 5c). For example, the S3 region, which has the lowest ensemble SSI, has three strong domain boundaries and one weak. On the other hand, regions like S1 and S4 which exhibit higher ensemble SSIs have fewer or weaker domain boundaries.

Finally, we asked whether chromosome association with the nucleolus could contribute to the chromosome structural ensemble SSI (Fig. 5d). The NAD data source and processing steps are described in the **Methods** section. We observe that S1 and S4, the regions with the highest structural ensemble similarity, are the most uniform in their NAD state: S1 is completely associated with the nucleolus, while S4 has almost no association with the nucleolus. As seen with epigenetic state, this suggests that uniformity of association leads to a higher continuity level in the structural ensemble. In contrast, S3 has the lowest ensemble SSI and also alternating levels of nucleolus association in the first half of the genomic region, followed by an unconstrained region. This variation in constraints across the region likely leads to specific distinct structures and thus a lower ensemble similarity.

Loop extrusion alone does not reproduce experimentally detected chromosome structure variation patterns

Above, we observed that more domain boundaries in a chromosome region associated with a less continuous structural landscape and lower ensemble SSI. Boundaries detected by approaches like HiCRatio can result from cohesin-mediated loop extrusion that is periodically blocked by factors such as CTCF (59), as well as from the borders between A and B compartment regions of chromatin. We next therefore asked whether the loop extrusion process, blocked by different numbers of boundaries, would be expected to generate the patterns of structure variation we observe. We simulated loop extrusion using the chr21:28-30 Mb segment using experimental cohesin binding data from K562 and IMR90 (see Methods). This region has fewer boundaries in K562 (likely due to the portion of the region that is heterochromatic in this cell type). The structure landscapes resulting from a single simulation run show that the cell type with more loop extrusion boundaries (IMR90) would be predicted to have a lower ensemble SSI than the cell type with fewer boundaries if only loop extrusion were at work (Fig. S10a and S10b). Once we start the simulation from several different initial conditions, this difference is diminished (Fig. S10b) due to variation in structure between different runs. While this result matches our expectations that more boundaries associate with lower ensemble SSI, it does not agree with the experimental observations from these cell types, in which IMR90 actually has the higher ensemble SSI. This suggests that epigenetic state-specific interactions must be considered along with loop extrusion to explain the chromosome variability observed.

The effect of cohesin depletion on structural heterogeneity depends on the underlying epigenetic state of the chromosome region

To further examine the effect of cohesin on the chromosome structural landscape, we analyzed chr21 single-cell structural data from HCT116 cells in wildtype (WT) and cohesin depleted (+Auxin) conditions (34). This analysis included two different genomic regions (chr21:28-30 Mb and chr21:34-37 Mb) with different underlying structural properties. The jointly constructed chromosome structural landscape for both genomic regions shows that along PC1 the structures are again ordered based on their R_g (Fig. S10c and S10d). Thus, the chromosome structural landscape approach orders single-cell structures based on their $R_g$irrespective of cell-type, genomic regions, and perturbation condition. The WT and cohesin-depleted structural landscapes from the first genomic region mostly overlap with a slight bias for the +Auxin structures in the higher R_g regime (Fig. 6c). This suggests that removal of cohesin makes the structures less compact to some extent, as previously observed (60).

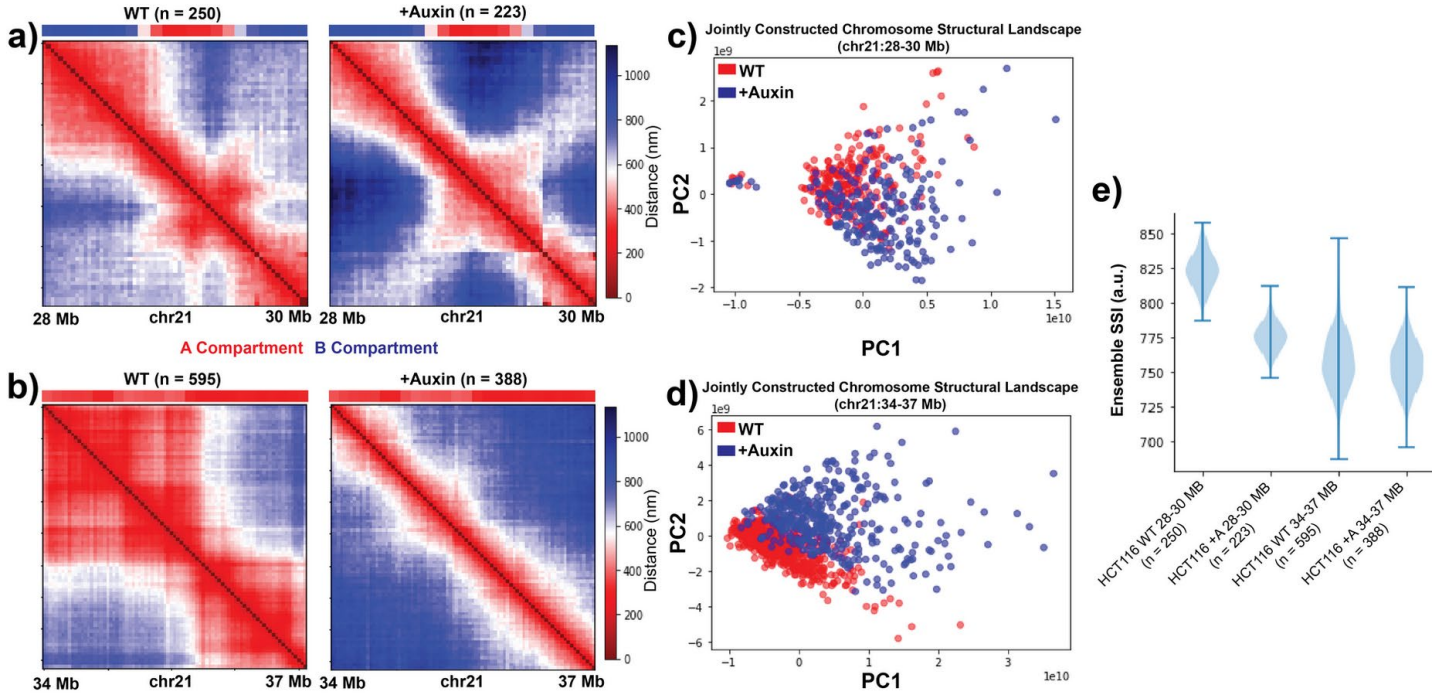

Figure 6. Cohesin depletion alters chromosome structural heterogeneity depending on underlying epigenetic state constraints. **a,b)** Population-averaged distance maps of HCT116 chr21:28-30 Mb (a) or chr21:34-37 Mb (b) at 30 kb resolution for wild-type (WT) and cohesin-depleted (+Auxin) conditions based on data obtained from Bintu *et al*. (34). Compartment status from Hi-C data at these locations shown above each matrix. **c,d)** Jointly constructed chromosome structural landscapes of chr21:28-30 Mb (c) or chr21:34-37 Mb (d) of HCT116 for WT and +Auxin conditions. **e)** Chromosome structural ensemble SSI distributions for different local chromosome regions of HCT116 for WT and +Auxin conditions. 200 single-cell structures are selected randomly each time for resampling.

On the other hand, for the second region, the jointly constructed landscape shows a shift upwards along PC2 along with the shift towards the large R_g regime for cohesin-depleted structures (Fig. 6d). In addition,

we found that after cohesin removal, both genomic regions exhibit more structural variation at the genomic bin level according to BSI standard deviation (Fig. S10e and Fig. S10f). When we examine the structural ensemble SSI of these two regions, we see apparently conflicting effects of cohesin removal: the ensemble SSI is decreased in the 28-30 Mb region but mostly unchanged in the 34-37 Mb region (Fig 6e). Note that the initial ensemble SSI of these regions in the WT follows the expected pattern for regions of different lengths. We observed that this divergent ensemble SSI effect corresponds to a difference in the underlying compartment state of the regions. The 28-30 Mb region is divided into A and B compartment segments (which in earlier examples tends to correspond to lower SSI) (Fig. 6a) while the 34-37 Mb region is all in the A compartment with no compartment boundaries (which in earlier examples tends to correspond to higher SSI) (Fig. 6b).

DISCUSSION

As the number and quality of single cell chromosome structural datasets continues to increase, a diverse toolbox of methods to capture properties of chromosome structure variation is needed. Our statistical analysis of single cell chromosome structures allows us to characterize both the major modes of chromosome structure fluctuation and the properties of the structure fluctuation landscape. While some approaches focus on classifying single cell structures into discrete clusters, other recent work has shown that chromosome structures can exist along a continuum (22,27,36,38,39). Our analysis does not require a decision to be made about how many different clusters of structures exist, but instead can capture the overall types of structure variations and can measure the degree of continuity between the structures.

For all observed length scales of chromosome structures, from local TAD scale to whole chromosome arms, we find that the primary type of single cell structure variation is the overall compaction of the region. For each structure assayed, there exist some cells in which the region is very compact and others in which it is much more extended. This overall size fluctuation is true of both active and inactive regions, though inactive regions and the inactive X chromosome show a shift toward smaller sizes overall. However, we find that this overall region compaction is largely independent of a secondary fluctuation mode: the reorganization of domains. No matter how compact or extended the overall region is, we observe a similar progressive reorganization of the domain boundaries as a secondary type of structure variation.

By quantifying the network-based similarity properties of the structural ensemble, we can measure how discrete vs. continuous the structures adopted by a certain chromosome region are. Combining all of our observations of ensemble SSI in different regions and conditions, we propose that the ensemble SSI of a structural ensemble is influenced by the epigenetic state constraints put on the region by its epigenetic state. These constraints may take the form of domain boundaries, where different parts of the chromosome regions are constrained in self-interacting patterns, or by associations with nuclear structures such as the nucleolus. Overall, constraints on the chromosome structure and boundaries between different types of regions decrease structural ensemble SSI and result in more discrete

structures while regions with more uniform linear chromatin states exhibit a more continuous structural landscape with higher SSI.

It might seem likely that the active or inactive chromatin status of a region would be the primary contribution of epigenetics to the structural fluctuation of a region. That is, one might expect an inactive region to have high structural ensemble SSI and an active region low ensemble SSI or vice versa. However, we find that this is not the case. Instead, the uniformity of the epigenetic state appears to be a key factor, such that a uniformly active region shares structural landscape properties with a uniformly inactive region (both have high SSI) while a region divided between alternating states behaves differently (low SSI). This leads us to the idea that alternating epigenetic states place constraints on the possible conformations and thus create a more disjointed, rather than continuous, structural ensemble.

The idea that epigenetic state-mediated constraints are key to shaping dynamic structural landscapes is echoed again by the results we observe in loop extrusion simulations and cohesin depletion data. Previous work has shown that compartmentalization and loop extrusion can act in opposition to each other (61). When cohesin is removed, we see that a region consisting of multiple epigenetic states can form subdomains and is thus more likely to adopt more discrete structures, resulting in a decrease in ensemble SSI compared to the state where cohesin is active. When cohesin is active, there is a chance of the entire region being pulled into the same loop, creating a more continuous landscape of possible structures (higher SSI). In contrast, when a region exists in a uniform epigenetic state, we see that the removal of cohesin does not result in a reduction of SSI. As before, cohesin removal is expected to reduce the active movement of chromatin, but in this case, the region is entirely in the same compartment, so there are no additional constraints that would lead to the formation of discrete states in the cohesin-depleted structural landscape. Our observations are consistent with the idea that cohesin loop extrusion acts to increase chromosome conformation dynamics, promoting distal contacts and countering the tendency of domains to settle into more discrete sub-structures (61,62). Our loop extrusion simulations suggest that in a uniform epigenetic state, the region with more boundaries would adopt more discrete structures and thus have less landscape continuity. But, loop extrusion alone, without consideration of interactions between similar chromatin states, cannot reproduce the differences in structure variability we see between different cell types.

The analysis of cohesin depletion results also reiterate the value of considering both the structural landscape and the ensemble SSI of that landscape. The structural landscape of a region can change (adopting a different range of radius of gyration or domain structures) while the continuity (SSI) stays the same. Alternately, the ensemble SSI can change while the structural landscape is overall conserved, indicating cases where the range and types of structures a region can adopt are the same, but the region is more often in certain discrete structures within that range rather than continuously sampling the range.

AVAILABILITY

Example analysis code is provided on the ChromosomeStructuralEnsembleCharacterization GitHub repository (https://github.com/rpmccordlab/ChromosomeStructuralEnsembleCharacterization).

SUPPLEMENTARY DATA

Supplementary Data are available online.

ACKNOWLEDGEMENT

The authors would like to thank Dr. Rebeca San Martin for valuable suggestions and insightful discussions in the development of the research project. We thank Dr. Tian Hong for sharing of computational GPU resources.

FUNDING

This work was supported by the National Institutes of Health [NIGMS grant R35GM133557 to R.P.M]. Funding for open access charge: National Institutes of Health/ R35GM133557.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. McCord, R.P., Kaplan, N. and Giorgetti, L. (2020) Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Molecular cell*, **77**, 688-708.
2. Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. and Mirny, L.A. (2016) Formation of chromosomal domains by loop extrusion. *Cell reports*, **15**, 2038-2049.
3. Nichols, M.H. and Corces, V.G. (2015) A CTCF code for 3D genome architecture. *Cell*, **162**, 703-705.
4. Sanborn, A.L., Rao, S.S., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A. and Li, J. (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, **112**, E6456-E6465.

5. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**, 458-472.
6. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J. and Sedat, J. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381-385.
7. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376-380.
8. Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J. and Dorschner, M.O. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, **326**, 289-293.
9. Falk, M., Feodorova, Y., Naumova, N., Imakaev, M., Lajoie, B.R., Leonhardt, H., Joffe, B., Dekker, J., Fudenberg, G. and Solovei, I. (2019) Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature*, **570**, 395-399.
10. Sanulli, S., Justin, N., Teissandier, A., Ancelin, K., Portoso, M., Caron, M., Michaud, A., Lombard, B., da Rocha, S.T. and Offer, J. (2015) Jarid2 methylation via the PRC2 complex regulates H3K27me3 deposition during cell differentiation. *Molecular cell*, **57**, 769-783.
11. Wang, L., Gao, Y., Zheng, X., Liu, C., Dong, S., Li, R., Zhang, G., Wei, Y., Qu, H. and Li, Y. (2019) Histone modifications regulate chromatin compartmentalization by contributing to a phase separation mechanism. *Molecular cell*, **76**, 646-659. e646.
12. Strom, A.R., Emelyanov, A.V., Mir, M., Fyodorov, D.V., Darzacq, X. and Karpen, G.H. (2017) Phase separation drives heterochromatin domain formation. *Nature*, **547**, 241-245.
13. Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C. and Speicher, M.R. (2005) Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, **3**, e157.
14. Cremer, T. and Cremer, C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, **2**, 292-301.
15. Stack, S.M., Brown, D.B. and Dewey, W. (1977) Visualization of interphase chromosomes. *Journal of cell science*, **26**, 281-299.
16. Kadauke, S. and Blobel, G.A. (2009) Chromatin loops in gene regulation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, **1789**, 17-25.
17. Braccioli, L. and de Wit, E. (2019) CTCF: a Swiss-army knife for genome organization and transcription regulation. *Essays in biochemistry*, **63**, 157-165.
18. Zuin, J., Roth, G., Zhan, Y., Cramard, J., Redolfi, J., Piskadlo, E., Mach, P., Kryzhanovska, M., Tihanyi, G., Kohler, H. *et al.* (2021) Nonlinear control of transcription through enhancer-promoter interactions. *bioRxiv*, 2021.2004.2022.440891.
19. Clapier, C.R., Iwasa, J., Cairns, B.R. and Peterson, C.L. (2017) Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature reviews Molecular cell biology*, **18**, 407-422.
20. Kim, Y. and Yu, H. (2020) Shaping of the 3D genome by the ATPase machine cohesin. *Experimental & Molecular Medicine*, **52**, 1891-1897.
21. Kim, J.M., Visanpattanasin, P., Jou, V., Liu, S., Tang, X., Zheng, Q., Li, K.Y., Snedeker, J., Lavis, L.D. and Lionnet, T. (2021) Single-molecule imaging of chromatin remodelers reveals role of ATPase in promoting fast kinetics of target search and dissociation from chromatin. *bioRxiv*.
22. Takei, Y., Zheng, S., Yun, J., Shah, S., Pierson, N., White, J., Schindler, S., Tischbirek, C.H., Yuan, G.-C. and Cai, L. (2021) Single-cell nuclear architecture across cell types in the mouse brain. *Science*, eabj1966.

23. Kempfer, R. and Pombo, A. (2020) Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics*, **21**, 207-226.
24. Jerkovic, I. and Cavalli, G. (2021) Understanding 3D genome organization by multidisciplinary methods. *Nature Reviews Molecular Cell Biology*, 1-18.
25. Di Pierro, M., Cheng, R.R., Aiden, E.L., Wolynes, P.G. and Onuchic, J.N. (2017) De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proceedings of the National Academy of Sciences*, **114**, 12126-12131.
26. Di Pierro, M., Potoyan, D.A., Wolynes, P.G. and Onuchic, J.N. (2018) Anomalous diffusion, spatial coherence, and viscoelasticity from the energy landscape of human chromosomes. *Proceedings of the National Academy of Sciences*, **115**, 7753-7758.
27. Conte, M., Fiorillo, L., Bianco, S., Chiariello, A.M., Esposito, A. and Nicodemi, M. (2020) Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation. *Nature communications*, **11**, 1-13.
28. Wlasnowolski, M., Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., Ruan, Y. and Plewczynski, D. (2020) 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic Acids Research*, **48**, W170-W176.
29. Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., Zhou, Y., Li, H., Zhou, X.J. and Le Gros, M.A. (2016) Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences*, **113**, E1663-E1672.
30. Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z. and Shendure, J. (2017) Massively multiplex single-cell Hi-C. *Nature methods*, **14**, 263-266.
31. Tan, L., Ma, W., Wu, H., Zheng, Y., Xing, D., Chen, R., Li, X., Daley, N., Deisseroth, K. and Xie, X.S. (2021) Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell*, **184**, 741-758. e717.
32. Arrastia, M.V., Jachowicz, J.W., Ollikainen, N., Curtis, M.S., Lai, C., Quinodoz, S.A., Selck, D.A., Ismagilov, R.F. and Guttman, M. (2021) Single-cell measurement of higher-order 3D genome organization with scSPRITE. *Nature Biotechnology*, 1-10.
33. Wang, S., Su, J.-H., Beliveau, B.J., Bintu, B., Moffitt, J.R., Wu, C.-t. and Zhuang, X. (2016) Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, **353**, 598-602.
34. Bintu, B., Mateo, L.J., Su, J.-H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N. and Zhuang, X. (2018) Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, **362**, eaau1783.
35. Su, J.-H., Zheng, P., Kinrot, S.S., Bintu, B. and Zhuang, X. (2020) Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell*, **182**, 1641-1659. e1626.
36. Sawh, A.N., Shafer, M.E., Su, J.-H., Zhuang, X., Wang, S. and Mango, S.E. (2020) Lamina-dependent stretching and unconventional chromosome compartments in early C. elegans embryos. *Molecular cell*, **78**, 96-111. e116.
37. Mateo, L.J., Murphy, S.E., Hafner, A., Cinquini, I.S., Walker, C.A. and Boettiger, A.N. (2019) Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature*, **568**, 49-54.
38. Cheng, R.R., Contessoto, V.G., Aiden, E.L., Wolynes, P.G., Di Pierro, M. and Onuchic, J.N. (2020) Exploring chromosomal structural heterogeneity across multiple cell lines. *Elife*, **9**, e60312.
39. Xie, W.J., Qi, Y. and Zhang, B. (2020) Characterizing chromatin folding coordinate and landscape with deep learning. *PLoS computational biology*, **16**, e1008262.
40. Zhu, C., Zhang, Y., Li, Y.E., Lucero, J., Behrens, M.M. and Ren, B. (2021) Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nature Methods*, **18**, 283-292.

41. Chen, H., Tian, Y., Shu, W., Bo, X. and Wang, S. (2012) Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PloS one*, **7**, e41374.
42. Bersaglieri, C., Kresoja-Rakic, J., Gupta, S., Bär, D., Kuzyakiv, R. and Santoro, R. (2020) Genome-wide maps of nucleolus interactions reveal distinct layers of repressive chromatin domains. *bioRxiv*, 2020.2011.2017.386797.
43. Cheng, F., Liu, C., Shen, B. and Zhao, Z. (2016) Investigating cellular network heterogeneity and modularity in cancer: a network entropy and unbalanced motif approach. *BMC systems biology*, **10**, 301-311.
44. Banerji, C.R., Miranda-Saavedra, D., Severini, S., Widschwendter, M., Enver, T., Zhou, J.X. and Teschendorff, A.E. (2013) Cellular network entropy as the energy potential in Waddington's differentiation landscape. *Scientific reports*, **3**, 1-7.
45. Menichetti, G., Bianconi, G., Castellani, G., Giampieri, E. and Remondini, D. (2015) Multiscale characterization of ageing and cancer progression by a novel network entropy measure. *Molecular BioSystems*, **11**, 1824-1831.
46. Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K. and Narayanan, A.K. (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic acids research*, **46**, D794-D801.
47. Ernst, J. and Kellis, M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, **9**, 215-216.
48. Chiang, M., Michieletto, D., Brackley, C.A., Rattanavirotkul, N., Mohammed, H., Marenduzzo, D. and Chandra, T. (2019) Polymer modeling predicts chromosome reorganization in senescence. *Cell reports*, **28**, 3212-3223. e3216.
49. Shi, G., Liu, L., Hyeon, C. and Thirumalai, D. (2018) Interphase human chromosome exhibits out of equilibrium glassy dynamics. *Nature communications*, **9**, 1-13.
50. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D. and Lander, E.S. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665-1680.
51. Plimpton, S. (1995) Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics*, **117**, 1-19.
52. Stefko, M. and Douglass, K.M. (2017).
53. Dillinger, S., Straub, T. and Nemeth, A. (2017) Nucleolus association of chromosomal domains is largely maintained in cellular senescence despite massive nuclear reorganisation. *PLoS One*, **12**, e0178821.
54. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A. and Manke, T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, **42**, W187-W191.
55. Cheng, Y., Liu, M., Hu, M. and Wang, S. (2021) TAD-like single-cell domain structures exist on both active and inactive X chromosomes and persist under epigenetic perturbations. *bioRxiv*, 2021.2005.2012.443887.
56. Mitter, M., Gasser, C., Takacs, Z., Langer, C.C., Tang, W., Jessberger, G., Beales, C.T., Neuner, E., Ameres, S.L. and Peters, J.-M. (2020) Conformation of sister chromatids in the replicated human genome. *Nature*, **586**, 139-144.
57. Szabo, Q., Bantignies, F. and Cavalli, G. (2019) Principles of genome folding into topologically associating domains. *Science advances*, **5**, eaaw1668.
58. Lazaris, C., Kelly, S., Ntziachristos, P., Aifantis, I. and Tsirigos, A. (2017) HiC-bench: comprehensive and reproducible Hi-C data analysis designed for parameter exploration and benchmarking. *BMC genomics*, **18**, 1-16.

59. Davidson, I.F. and Peters, J.-M. (2021) Genome folding through loop extrusion by SMC complexes. *Nature Reviews Molecular Cell Biology*, **22**, 445-464.
60. Hao, X., Parmar, J.J., Lelandais, B., Aristov, A., Ouyang, W., Weber, C. and Zimmer, C. (2021) Super-resolution visualization and modeling of human chromosomal regions reveals cohesin-dependent loop structures. *Genome biology*, **22**, 1-20.
61. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. and Mirny, L.A. (2018) Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proceedings of the National Academy of Sciences*, **115**, E6697-E6706.
62. Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C.H. and Mirny, L. (2017) Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, **551**, 51-56.

FIGURE LEGENDS

Figure 1. Schematic representation of the statistical pipeline used to quantify conformational diversity present in single-cell chromosome distance matrices. **a)** Cartoon representation of potential differences in chromosome structure landscapes (left). Condition A has a continuous landscape including all structures along a trajectory, while in Condition B, chromosomes fall into 3 major structural clusters. (Right) Conditions and factors that may influence this chromosome structure landscape. **b)** A pairwise similarity score is calculated between each pair of median-centered chromosome distance matrices. PCA is performed on the pairwise structure-structure similarity matrix and transformed data is projected in 2D space using PC1 and PC2 to create the “chromosome structural landscape”. **c)** To calculate chromosome structural ensemble SSI, we construct a Pearson’s correlation matrix from the pairwise structure-structure similarity matrix. This correlation matrix scaled, thresholded, and then converted into a weighted network from which SSI values can be calculated from each node. The median SSI value is plotted for many resamplings of nodes. **d)** To calculate bin variability, we convert each single-cell distance matrix into a thresholded matrix using a cutoff and then construct a network from each of those matrices with bins as nodes and contacts between bins as edges. We then calculate the bin similarity index for each bin in each single-cell and then take the standard deviation of the bin similarity index across the population as the bin variability.

Figure 2. The chromosome structural landscape orders single-cell structures on a continuum based on their radius of gyration followed by domain reorganization. **a)** Population-averaged distance map of IMR90 chr21:28-30 Mb at 30 kb resolution. **b)** Chromosome structural landscape of IMR90 chr21:28-30 Mb chromatin region, generated as described in the **Methods** section. Each point represents a single-cell structure. Points are colored according to the radius of gyration (R_g) of the corresponding structures (dark blue – smaller R_g and dark red – larger R_g). The percent of variation explained by each PC is shown in parentheses. **c)** Structures are first divided into five groups based on the PC1 ordering (which correlates highly with R_g). Structures belonging to groups 2, 3 and 5 are shown. Within each group structures are further divided into five subgroups based on their PC2 ordering. Average structures within each subgroup are shown.

Figure 3. Network based similarity score quantifies cell-type- and chromosome-specific variability present in chromosome structural ensemble at different resolutions. **a)** Jointly constructed chromosome structural landscape of genomic region chr21:28-30 Mb for cell types IMR90, K562 and A549. The only difference between individual and joint landscape is that all the different cell-type structures are considered together while calculating pairwise structure-structure similarity matrix. Points are colored based on the cell types. Further, the landscape is divided into ten equal sized bins (or regions) based on PC1 values, where the bins on the left represent lower R_g regions and bins on the right are with higher R_g. **b)** Relative proportion of IMR90, K562 and A549 single-cell structures present in each bin represented in a). **c)** Chromosome structural ensemble SSI distributions for different 2 Mb chromosome regions from different cell-types and conditions, obtained from Bintu *et al*. (34). The SSI value of the structural ensemble is calculated using a network entropy based approach. For each structural ensemble (size of ensemble indicated as n), 200 single-cell structures are selected randomly (except IMR90 S 28-30 Mb where we select 150 structures for resampling), and the median of the SSI values of the nodes are calculated. This step is repeated 1000 times and values are plotted as a distribution. **d)** Chromosome structural ensemble SSI distributions for active and inactive IMR90 chrX at TAD resolution, obtained from Wang *et al*. (33). 50 single-cell structures are selected randomly each time for resampling. **e)** Chromosome structural ensemble SSI distributions for different IMR90 chromosomes and conditions at higher resolution, obtained from Su *et al.* (35). Here, for resampling 300 single-cell structures are selected randomly each time.

Figure 4. Epigenetic properties (histone modifications) contribute to the cell-type-specific variation in the chromosome structural ensemble. **a)** Chromatin states of chr21:28-30 Mb genomic region from three different cell-types – IMR90, K562 and A549 at 30 kb resolution. The chromatin states are inferred based on 11 different histone marks for each cell-type by ChromHMM tool. **b)** Bin variability of the chr21:28-30 Mb region from IMR90, K562 and A549, where a higher value represents higher degree of structural variation of that genomic position across the population and vice versa. **c)** Population-averaged distance maps of IMR90 and K562 chr21:28-29.5 Mb genomic region from polymer simulations (see Methods) at 1 kb resolution. **d)** Chromosome structural ensemble SSI distributions for chr21:28-29.5 Mb from simulated IMR90 and K562. For resampling, 500 single-cell structures are selected randomly each time.

Figure 5. Evaluation of structural variation of regions with different epigenetic, genomic, and structural properties. **a)** Population-averaged distance maps of 6 different genomic regions of IMR90 chr21 at 50 kb resolution obtained from Su *et al.* (35). Locations of these 6 regions are indicated on the top right chromosome ideogram and labeled with colors used throughout the figure. **b)** Chromosome structural ensemble SSI distributions for the same 6 regions of IMR90 chr21. For resampling, 300 single-cell structures are selected randomly each time. **c)** HiCRatio score of each genomic bin of chr21 q-arm at 50 kb resolution (calculated with a 300 kb window size). The red points represent the boundary positions inferred by HiCRatio. **d)** Nucleolus association level of the same 6 regions of IMR90 chr21 (53). For each genomic region, the 1 Mb upstream and downstream regions are also shown. A value greater than 0 represents higher levels of nucleolus association and vice versa. Colors in each panel correspond to chromosome locations indicated above panel b.

Figure 6. Cohesin depletion alters chromosome structural heterogeneity depending on underlying epigenetic state constraints. **a,b)** Population-averaged distance maps of HCT116 chr21:28-30 Mb (a) or chr21:34-37 Mb (b) at 30 kb resolution for wild-type (WT) and cohesin-depleted (+Auxin) conditions based on data obtained from Bintu *et al*. (34). Compartment status from Hi-C data at these locations shown above each matrix. **c,d)** Jointly constructed chromosome structural landscapes of chr21:28-30 Mb (c) or chr21:34-37 Mb (d) of HCT116 for WT and +Auxin conditions. **e)** Chromosome structural ensemble SSI distributions for different local chromosome regions of HCT116 for WT and +Auxin conditions. 200 single-cell structures are selected randomly each time for resampling.